LA BUSINESS ANALYTICS DE A A Z SUR LE CAS ORION STAR

GRÉGOIRE DE LASSENCE

"We make a living by what we get, but we make a life by what we give"
Winston Churchill

Table des matières :

Introduction à l'informatique décisionnelle

« L'ignorance est mère de tous les maux » F. Rabelais

Préambule :

Le monde dans lequel nous vivons est de plus en plus complexe. Les technologies de l'information nous génèrent une multitude de données comme jamais auparavant. Le problème n'est donc plus tant d'acquérir une masse de données, mais de l'exploiter. Pour cela il faut collecter de l'information de qualité, la normaliser, la classer, l'agréger, et l'analyser, pour l'exploiter afin d'en extraire la substantifique moelle et donc prendre la bonne décision au bon moment. Dans ce but, il est nécessaire de mettre en place un système d'information particulier, appelé système décisionnel. Ce système doit permettre de présenter de manière simple les chiffres recueillis pour mettre en lumière la conjoncture actuelle et indiquer implicitement la voie à suivre. Un système décisionnel ne remplace pas les systèmes opérationnels qui font fonctionner l'entreprise, mais il vient s'y intégrer, en extrayant des données, afin d'en diffuser la connaissance, de la manière la plus facilement exploitable par les personnes concernées.
Le système opérationnel n'est pas, apriori, modifié par la mise en place du système décisionnel, ce dernier vient le compléter par une exploitation avancée de l'information. Il est donc nécessaire d'ajouter aux systèmes opérationnels qui permettent très bien de gérer l'entreprise au quotidien, un système offrant la capacité d'analyser le passé, le présent et de simuler l'avenir pour anticiper les changements constants de notre société.
Un système décisionnel doit notamment permettre de passer de la simple réactivité à l'anticipation et à la pro-activité. Être proactif, cela veut dire par exemple qu'au lieu d'attendre que le client vous envoie la résiliation de son contrat, étant donné que vous avez déjà modélisé ce comportement, vous pouvez anticiper. Vous lui envoyez donc une proposition devant vous permettre de le fidéliser, avant qu'il n'ait pris cette décision de résiliation. Les organisations ont pour la plupart, mis en place leur système opérationnel pour être efficaces et réactives, elles doivent maintenant, pour être proactives, déployer un système décisionnel.

Au risque d'apporter de la confusion avec un nouvel anglicisme, on parle désormais de Business Analytics, qui peut être vu comme le rassemblement de la Business intelligence avec le Data Mining. La question de savoir si ces deux composants étaient vraiment séparés dans l'esprit des différents acteurs du domaine est tout à fait légitime car telle n'est pas forcément le cas. Faute de définitions partagées par tous, voici donc une expression nouvelle importante : la Business Analytics.
Pour utiliser des mots de notre magnifique langue on peut dire que la Business Analytics intègre les différents éléments de l'informatique décisionnelle classique (outil d'extraction, de qualité, de normalisation des données, de transformation, de chargement, de gestion de bases de données gigantesques, de gestion de bases de données multidimensionnelles, de création d'environnement de rapport pour tous, de génération de rapport à la demande ou de masse, etc.) ; et des outils d'analyses statistiques, des séries temporelles, d'économétrie, de fouille de données, de recherche opérationnelle, etc.
Je n'ai pas trouvé d'expression en français pour traduire avec tout son sens la « Business Analytics », je vous propose donc l' « informatique mathématique décisionnelle ».

Le mot Data Mining est communément traduit par fouille de données, mais souvent distingué des analyses statistiques et de la recherche opérationnelle, alors que dans l'expression « d'informatique mathématique décisionnelle », se retrouvent tous ces éléments intégrés. Statistique, économétrie, Data Mining et recherche opérationnelle peuvent tous être compris dans les mathématiques mais surtout nécessite de vraies bases mathématiques pour être correctement utilisés dans toutes les situations.

La tendance actuelle est de vouloir une recette de cuisine pour résoudre chaque problème. Or les problèmes de demain ne sont par définition pas encore connus.
Par exemple pour le Data Mining, probablement 90%[1] des problématiques d'hier, d'aujourd'hui et de demain peuvent être résolut par un processus automatique performant. Le problème, c'est que la

[1] Ce pourcentage est seulement une approximation grossière.

différence se fait sur les 10%, où les méthodes standards ne fonctionnent pas. C'est là que doit intervenir un mathématicien.

J'utilise le mot mathématicien au sens noble du terme, c'est-à-dire quelqu'un qui a une culture mathématique très large, donc une bonne connaissance des techniques pouvant être mise en place, et non juste un ensemble de recettes magiques. Une telle personne peut alors sentir la philosophie d'un problème et utiliser les bons algorithmes, ou bien comprendre que telle technique qui fonctionnait bien jusqu'à présent ne peut plus être utilisé car le contexte a changé.

Un problème récurrent est que ce profil appelé parfois : « le chercheur fou » est souvent cloisonné dans son laboratoire alors qu'il doit être intégré notamment par les outils, par les plateformes décisionnelles, afin que les organisations ne soit pas seulement réactives, mais bien agile et proactive ; afin qu'elles puissent anticiper et s'adapter en permanence.

Un second problème important est que par une méconnaissance des technologies, ce qui peut paraître contradictoire dans un monde de technologie, les mauvaises techniques sont trop souvent utilisées. On fait un petit test et sans rigueur scientifique, on met en production un assemblage de bricolage, de technologies non éprouvées, jusqu'à ce que cela finisse par devenir catastrophique.

Un troisième problème destructeur de l'informatique décisionnelle est la mauvaise qualité des données. Sans données de qualité, on obtient des analyses, des rapports faux, voir contre-productif. Pour résoudre ce problème, on peut notamment aborder deux éléments technologiques :

1. La normalisation des données avec un outil de qualité de données
2. Une administration rigoureuse permettant une attribution juste des droits, une traçabilité depuis les données sources jusqu'au rapport final. Si dans ce processus, il y a une analyse mathématique, elle doit elle-même être administrée et traçable.

La plupart des plateformes décisionnelles actuelles n'intègrent pas au sens propre, une vraie palette d'algorithmes d'analyses statistiques, Data Mining, économétriques et de recherche opérationnelle, éprouvés. Vous trouvez souvent soit des outils qui, sous le couvert d'une vulgarisation à outrance, vous font croire que vous pouvez utiliser ces fonctions sans en avoir les compétences, et les dérives peuvent être désastreuses ; soit une intégration qui ressemble beaucoup plus au bricolage.
Par déontologie, je ne cite personne quand je ne dis pas que des choses sympathiques, mais vous l'aurez bien comprit, mon point de vue n'est pas des plus objectifs. Oui, travaillant pour SAS depuis 2002, je souhaite partager avec vous mon enthousiasme pour une véritable plateforme de Business Analytics.

Cette définition ne vaut que par ce qu'elle introduit, mais étayons la un peu. Le plus gros problème de la Business Analitycs (Informatique Mathématique Décisionnelle), c'est de faire travailler ensemble des informaticiens avec des analystes et des utilisateurs fonctionnels. Afin de réduire la complexité, dans beaucoup de projet, la composante analytique qui est le vrai démultiplicateur de retour sur investissement, a été négligée, minimisée voir pas du tout prise en compte. Il est parfois possible de faire la même remarque sur la partie informatique où le projet n'a pas été fait selon les règles de l'art (sans véritable administration, cryptage, traçabilité, capacité à monter en puissance, intégration, etc…).
Cette expression : « l'informatique mathématique décisionnelle » a donc pour vocation de présenter à poids égale, ces trois composantes : informatique, mathématique et décisionnelle.

La mise en place d'un système décisionnel permet d'apporter des réponses efficaces à tous les niveaux de l'entreprise ; cet aspect décisionnel est présent dans les organisations depuis de nombreuses années, il revêt l'apparence de rapports et de tableaux de bord. Mais, beaucoup d'entre elles s'aperçoivent que ces simples outils de Reporting ne satisfont pas entièrement leurs attentes. Elles se rendent compte que la mise en place d'un entrepôt de données global, transversal et cohérent, lié à des outils d'analyses, est nécessaire. La Business Intelligence est devenue une priorité pour les directions informatiques, pour ne pas dire la priorité.
Si globalement les projets informatiques ne sont plus en forte croissance après leur explosion de la fin des années 90 et du début des années 2000, les projets d'informatique décisionnelle ont toujours le vent en poupe malgré la crise, pour ne pas dire : grâce à la crise. Il faut noter deux axes majeurs : premièrement la gestion de la relation client et deuxièmement la chasse aux coûts par une meilleure compréhension des mécanismes de création de valeur. En effet, je ne vous apprends rien en vous

disant que l'objectif d'une société privée est de gagner de l'argent, c'est le nerf de la guerre. Pour augmenter la marge, les deux principaux leviers sont d'augmenter le chiffre d'affaires et de réduire les coûts. Nous nous intéresserons donc notamment aux solutions permettant :

- de mieux comprendre le client pour mieux lui vendre sur le long terme,
- et de mieux comprendre l'utilisation de l'argent dépensé pour découvrir des économies intelligentes potentielles.

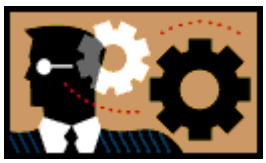
Il est important, dès à présent, de relativiser cet objectif très matérialiste par le fait que si ces outils peuvent être utilisés par des personnes sans scrupules, nous ne nous intéresserons ici qu'à des objectifs satisfaisant des valeurs strictes et une intégrité forte, avec une philosophie de développement durable. Par exemple nous ne nous intéresserons pas à piller le client comme cela peut être le cas, mais bien à développer une relation durable, pérenne : gagnant-gagnant, beaucoup plus rentable sur le long terme.
Si l'on modélise le comportement du client de façon primaire, cela peut développer des relations commerciales uniquement basées sur la performance à court terme ; alors que si l'on modélise finement ce comportement et que l'on intègre la recherche opérationnelle, on obtient une optimisation globale, qui généralement, tend à inciter des rapports commerciaux sur le long terme.

Une crainte fréquemment rencontrée vis-à-vis des technologies de l'information est le fait de se sentir espionné. J'ai souvent eu la réflexion juste après avoir présenté la Business Intelligence : « mais c'est Big Brother » ?
Oui, ces outils peuvent être utilisés à des fins portantes atteintes aux libertés fondamentales et à la vie privée ; ce ne sont que des outils. Néanmoins, avec une politique de transparence forte qui se doit de garantir une véritable intégrité, l'informatique décisionnelle permet souvent d'améliorer significativement la performance des organisations.
De plus, et en France tout particulièrement, des lois très strictes règlementent l'utilisation des bases de données.
Il est important de rappeler à tous les consommateurs que nous sommes, qu'il nous est possible de ne pas prendre une carte de fidélité, et donc de ne pas bénéficier des avantages qu'elle propose ; on peut aussi à tout moment demander de ne plus recevoir d'offres promotionnelles.

En résumé, l'informatique décisionnelle peut, comme tout autre système, être pervertie. Une plate-forme décisionnelle intégrée est nécessaire pour transformer des données légalement utilisables en connaissance afin d'amélioration la performance à court, moyen et long terme, des organisations.

Enfin, cette croissance de l'informatique décisionnelle est particulièrement forte dans le secteur public aussi où des organisations gigantesques doivent fournir des rapports précis sur notamment la manière dont elles utilisent l'argent du contribuable. Par exemple, en France, la LOLF, Loi Organique sur la Loi de Finance, impose quasiment de gérer les établissements publics, comme ceux du privé, avec des objectifs de performance. C'est une révolution qui impose de passer d'une logique de moyen à une logique de résultat, ce qui nécessite des changements gigantesques et notamment en termes de rapport et d'analyse.

Les projets d'informatique décisionnelle sont donc des projets actuels mais surtout d'avenir.



Ce qu'il faut retenir : L'informatique décisionnelle a pour objectif de transformer les données déjà présentes dans et à l'extérieur de l'entreprise, en connaissance, afin de permettre la judicieuse décision au bon moment.

Publics visés

Le défi majeur de l'informatique décisionnelle, bien au-delà de la technologie, est humain. En effet, ce processus de transformation de la donnée en connaissance nécessite de nombreuses compétences qui doivent s'allier pour travailler vers un objectif commun. L'objectif ici, sera de montrer aux différentes parties prenantes, l'ensemble du processus décisionnel, afin que chacun puisse mieux y participer.

Nous nous adressons donc à plusieurs profils d'étudiants, de professeurs ou de professionnels que l'on peut répartir de la façon suivante :

- Les techniciens : Les personnes ayant des compétences informatiques doivent pouvoir prendre en main concrètement la plate-forme décisionnelle SAS, mais aussi d'appréhender le besoin d'ergonomie de l'applicatif final, la nécessité d'intégrer l'analyse de données, et comprendre quelques problématiques[1] métiers[2]. Trop de projets décisionnels sont purement techniques et non fonctionnels ; il est donc important d'élargir le champ des informaticiens par un cas concret, aux besoins analytiques et métiers. Le processus de la chaine du décisionnel sera vu du début à la fin, avec des approfondissements sur les choix techniques pouvant être mis en œuvre. Un projet de Business Analytics est un projet technologique qui nécessite des experts en la matière.
- Les architectes et administrateurs de système décisionnel : Les futurs chefs de projets, architectes ou administrateurs de projets décisionnels y trouveront une vision globale de la plate-forme décisionnelle SAS afin d'en conceptualiser la structure. Parmi les discussions importantes qui seront soulevées, notons la gestion du projet et du changement induit, l'administration, la sécurité, les différentes architectures techniques et fonctionnelles envisageables et l'intégration d'une telle plate-forme dans les systèmes informatiques.
- Les statisticiens : Les analystes, statisticiens ou Data Miner pourrons mieux appréhender les nouvelles opportunités apportées par l'intégration de l'analyse de données dans le processus d'aide à la décision ; les fusions amonts (ETL, administration, structure de Data Warehouse, OLAP, etc.) et avals (Reporting, portail, procédure stockée, etc.). Il est primordial d'analyser les données nombreuses et complexes, mais surtout d'industrialiser leur intégration dans le processus décisionnel.
- Les analystes métiers (de l'anglais Business Analyste) : des personnes ayant des connaissances techniques et fonctionnelles afin de créer le lien entre le Data Warehouse et les métiers. Personnes qui vont assister la maîtrise d'ouvrage, créer des Information Map[3] avec SAS et/ou faire des rapports spécifiques complexes à la demande.
- Les décideurs actuels ou futurs : Il est fondamental d'aider les décideurs présents ou futurs, à mieux appréhender la face cachée de l'iceberg de l'informatique décisionnelle. Cette ouverture vers un monde souvent rendu obscur par un jargon technique spécialisé, truffé de sigles et d'anglicismes, devra leur permettre de conceptualiser la valeur ajoutée d'une plate-forme décisionnelle.
- Tout le monde dans l'entreprise, avec une dédicace toute particulière aux formidables assistantes qui sont souvent les plus grosses utilisatrices directs du décisionnel. Dans « informatique décisionnelle », le mot décideur donne souvent une connotation un peu prétentieuse au sujet. Le terme décideur serait plutôt à prendre au sens large, depuis les cadres ayant à prendre beaucoup de décisions, jusqu'aux employés ayant des rapports à fournir régulièrement. L'objectif est donc de donner à ces utilisateurs non techniciens, une meilleure connaissance de ces systèmes pour qu'ils les utilisent mieux et prennent une part encore plus active dans leur élaboration et leur évolution.

[1] Ici, le mot problématique est bien à prendre dans son acceptation littéraire, c'est-à-dire ensemble de problèmes.

[2] Métier : les métiers génériques de l'entreprise sont la production, les ressources humaines, le marketing, les forces de ventes. Les utilisateurs métiers d'un système décisionnel sont donc des utilisateurs généralement non informaticiens, ayant besoin d'accéder de manière simple à la connaissance contenue dans les systèmes d'information.

[3] Une Information Map de SAS peut être comparée à un Univers de SAP Business Object. Voir le chapitre « Information Map » sur le sujet.

Objectifs

Objectif global : Dans le monde dans lequel nous vivons, les décisions ne peuvent plus être prises uniquement sur des coups de génie. Pour bien gérer une organisation, il est nécessaire de s'appuyer sur des informations de qualité. Les managers d'aujourd'hui et surtout de demain, ont besoin d'analyses pour comprendre le présent et simuler l'avenir.

Les objectifs de conceptualisation d'un projet décisionnel peuvent donc être présentés selon trois axes majeurs :

1. L'axe technique : afin de comprendre l'organisation logique de la chaine de création de valeur de l'informatique décisionnelle.
2. L'axe analytique : on peut regarder cet axe comme « données → informations → connaissances ». L'informatique décisionnelle doit récupérer des données diverses et variées, et si l'on se borne au Reporting, présenter l'information que ces données contiennent. Mais si l'on ajoute l'analytique, à ce moment-là, les rapports ne fournissent plus de l'information, mais de la connaissance.
3. L'axe fonctionnel : l'objectif est de conceptualiser l'apport d'un système décisionnel intégrant des analyses statistiques et du Data Mining pour des problématiques métier. Par affinité personnelle, je prendrais surtout des exemples marketing, mais aussi ressources humaines ; du domaine de la banque ou de la grande distribution, etc.

Le but de l'informatique décisionnelle est de transformer les données de l'entreprise en 'Intelligence'. Il est important de préciser que le mot 'intelligence' est à prendre dans son acceptation anglo-saxonne, c'est-à-dire connaissance de l'entreprise ; connaissance de son fonctionnement, de ses partenaires : clients – fournisseurs, de sa structure : groupe – filiale, de ses produits, de ses processus, de son organisation, de ses ressources humaines, de son histoire, de son présent mais aussi de son futur probable.
Remarquons que le terme Intelligence est souvent galvaudé, mais il est vrai que le Data Mining, peut être considéré comme une partie intelligente de l'informatique décisionnelle et il est certain, que c'est souvent un élément démultiplicateur de retour sur investissement.

Les entreprises sont en perpétuel changement. Les organisations privées ou publiques ont en général comme points communs, des clients, des employés et des partenaires de plus en plus exigeants. Face à ces changements de plus en plus rapides et à cette concurrence de plus en plus forte la simple réactivité ne suffit plus : il faut anticiper. Cette anticipation ne peut être efficace qu'en s'appuyant sur des informations pertinentes, des prévisions justes, voir des simulations et la recherche d'optimum. Mais, dans leurs organisations actuelles, les données sont volatiles, surabondantes, non organisées pour la prise de décision, et souvent éparpillées dans de multiples systèmes hétérogènes. Il devient donc capital de rassembler et d'homogénéiser les données afin de permettre l'analyse des indicateurs nécessaires aux prises de décisions.

L'informatique décisionnelle apporte à l'entreprise l'information élaborée l'aidant à comprendre, à maintenir et à gérer sa compétitivité, à accroître sa part de marché, à fidéliser sa clientèle et à optimiser ses processus et ses coûts.

De manière communément acceptée, le terme Business Intelligence (BI) se traduit en français par informatique décisionnelle (ID) et de manière réciproque.

Un peu d'Histoire

C'est une longue histoire, alors pour ne citer que quelques grandes dates clés, commençons en 1958, le chercheur d'IBM Hans Peter Luhn, a utilisé le terme Business Intelligence qu'il a défini comme « *la capacité de présenter les interrelations entre des faits de telle sorte que cela permette de guider les actions pour atteindre le but espéré.* »
L'informatique décisionnelle est associée au concept de Data Warehouse, ce qui se traduit généralement en français par entrepôt de données. Ce concept a été formalisé en 1992 par Bill Inmon dans l'ouvrage "Developing the Data Warehouse" (Construire l'Entrepôt de Données) de la façon

suivante : « *Un Data Warehouse est une collection de données thématiques, intégrées, non volatiles et historisées pour la prise de décisions.* ».
En 1996, Ralph Kimball, a publié le livre de référence “The Data Warehouse Toolkit ».
Bill Inmon et Ralph Kimball sont souvent considéré comme les pères du Data Warehousing et de la Business Intelligence.

Selon les ouvrages, on lit souvent qu'à la suite des infocentres des années 70 suivent les EIS (Executive Information System) des années 90, avant de parler de l'informatique décisionnelle avec le Data Warehouse et la Business Intelligence.
Le problème est que selon les auteurs ou les sociétés, ces échelons varient. Beaucoup cherchent à revendiquer la paternité d'un concept, d'une méthodologie ou d'une technologie.
Le même constat peut être fait pour le Data Mining qui serait apparu au milieu des années 90, avec par exemple la première conférence de KDD (Knowledge Discovery and Data Mining) en 1995 à Montréal.
Voici donc quelques définitions qui n'ont certainement pas la prétention d'être la vérité absolue mais simplement une proposition critiquable, socle pour la suite.

Voici trois définitions préliminaires :
Infocentre : collection de données pour la prise de décision. Terme utilisé d'après la littérature entre les années 70 et la fin des années 90. On dit souvent que l'infocentre est l'ancêtre du Data Warehouse.

EIS : Executive Information System, terme apparemment utilisé dans les années 90 pour désigner des outils de restitution d'information synthétique, souvent sous forme graphique, essentiellement pour le top management. Contrairement aux outils Business Intelligence actuels, les EIS fournissaient des rapports statiques. Il était quasiment nécessaire de faire un projet informatique dans les règles pour chaque rapport, c'est-à-dire faire un cahier des charges, un de spécification, le développement et le test. Il y avait donc un décalage important entre le moment où le besoin était formulé, et celui de la livraison.

SIAD : Système d'Information d'Aide à la Décision ou bien aussi Système d'Information et Aide à la Décision, ou encore Systèmes d'Information et Analyse Décisionnelle. Les briques de l'analyse de donnée, du Data Mining et de la recherche opérationnelle sont généralement plus mises en avant.

Pour certains, ces termes : Infocentre, EIS voir SIAD sont obsolètes. Souvent, on considère que les infocentres et les EIS étaient statiques, alors que maintenant, une plateforme BI permet l'interactivité. L'utilisateur peut lui-même créer la plupart de ses rapports.

Un entrepôt de données est une base de données :

- Thématiques : c'est-à-dire orienté sujet, métier. Contrairement aux bases de données opérationnelles généralement orientées processus, les bases des entrepôts décisionnels sont modélisées pour répondre facilement à toutes les questions d'utilisateurs non-informaticiens. Il faut donc apporter à l'utilisateur l'information selon sa définition métier. Ce premier point n'est pas vérifié dans le DDS. Voir le chapitre a) DDS
- Intégrées : regroupant généralement des sources hétérogènes de données. Exemple : pour avoir la connaissance métier d'un client, il faut souvent rassembler les informations issues des systèmes opérationnels
    - de gestion des forces de ventes pour connaître ses dernières commandes,
    - de la comptabilité pour savoir s'il a payé sa dernière facture,
    - du service après-vente s'il l'a utilisée,
    - du serveur Web s'il s'est connecté sur le site,
    - du partenaire avec lequel un programme de fidélité commun a été mis en place,
    - etc.

  Un entrepôt de données décisionnelles d'entreprise doit permettre d'avoir une vision unique et transversale de l'information.
- Non volatiles : stables, non modifiables. Contrairement à un système opérationnel modifié après chaque transaction, les informations d'un système décisionnel ne changent quasiment pas.
- Historisées = archivées : datées, afin de conserver un historique. Cela permet les analyses comparatives. Exemple : le solde du compte en banque du client est une variable volatile qui

change à chaque transaction. On va donc retrouver généralement, si la volumétrie le permet, l'information de détail, le plus petit dénominateur commun à différents problèmes. Ayant ce détail daté, il sera alors toujours possible de recalculer un indicateur à un instant t afin de par exemple comparer le solde aujourd'hui à celui de l'année dernière à la même date (YTD[1]).

Tout comme cette définition très technique, dans l'expression « informatique décisionnelle », il y a le mot « informatique » qui donne maladroitement une connotation technique qui peut faire peur. Certes, les projets d'informatique décisionnelle requièrent de nombreuses compétences informatiques, mais il ne faut pas perdre de vue que leur objectif est de mettre en place des solutions d'aide à la décision pour des personnes non informaticiennes. Il faut donc que l'équipe projet comporte des compétences fonctionnelles pour que l'applicatif final soit orienté métier.
D'un autre côté, les projets d'informatique décisionnelle requièrent des informaticiens compétents et rigoureux. Même si les interfaces des utilisateurs sont de plus en plus intuitives, il est consternant de voir des utilisateurs faire leurs rapports sur des pseudos Data Warehouse. Pour faire correctement du décisionnel, il faut un vrai Data Warehouse digne de ce nom, construit de façon rigoureuse par des informaticiens, capables de le maintenir.

Pour les éditeurs de logiciel, les intégrateurs et les SSII, le conseil métier se vendant généralement plus cher, l'aspect fonctionnel devient souvent trop mis en valeur par rapport à l'aspect technique. Les présentations Power Point se développent parfois sur du vent. Rares sont ceux qui maitrisent véritablement les aspects technologiques et fonctionnels. Sans une architecture technique solide, un système décisionnel n'est rien et est voué à mourir.
Un projet décisionnel est à la fois un projet technique et fonctionnel.
Les mots « informatique », « statistique » ou « décisionnel » ; peuvent faire peur. L'objectif est ici d'abord pédagogique, de les réunir.

Pour la partie fonctionnelle, il ne faut pas non plus se focaliser uniquement sur les directions générales. Certes, il est plus flatteur de dire que l'on travaille pour le tableau de bord du directeur général, mais un système décisionnel a aussi très souvent pour but de faciliter le travail de Reporting des assistantes. C'est d'ailleurs souvent le premier retour sur investissement mesurable : la diminution significative du temps consacré au Reporting grâce à l'industrialisation de celui-ci.
L'informatique décisionnelle s'adresse donc à un public très large et hétérogène. Il est important de formaliser les termes du décisionnel afin que tous puissent communiquer normalement. Pour complexifier la chose, les différents protagonistes du décisionnel peuvent avoir des définitions différentes pour un même mot. Les définitions présentées ici n'ont pas la prétention d'être seules et uniques ; simplement de tenter de formaliser un peu ce jargon obscur où chacun définit les termes comme cela l'arrange.

Il faut remarquer que les expressions « entrepôt de données » et « informatique décisionnelle » recouvrent tous les deux, deux notions synecdoques[2] :
- Pour la première, l'entrepôt de données ou Data Warehouse, renvoie à l'infrastructure décisionnelle dans son ensemble, ou seulement au contenant des données décisionnelles, le stockage de l'information décisionnelle.
- Pour la seconde, l'informatique décisionnelle ou Business Intelligence, renvoie à l'ensemble du processus dans sa globalité (extraction – validation - transformation – chargement – stockage – restitution – analyse de données – diffusion de la connaissance – génération de rapport – gestion de la performance - etc.) ou seulement aux applications finales de Reporting et de pilotage, partie émergée de l'iceberg, visible par l'ensemble des utilisateurs.

[1] YTD : Year To Date. Pour votre culture, voici une expression très utilisée que vous connaissez surement déjà.
[2] Une synecdoque est une figure de style qui consiste à donner à un mot un sens plus large ou plus restreint qu'il ne comporte habituellement. Par exemple, étant sur un port, dire « je vois une voile à l'horizon » et équivalent à dire « je vois un bateau à l'horizon ».

Le décisionnel, pourquoi faire ?

«Toutes les défaites du monde se résument en 2 mots : trop tard ! » Général Mac Arthur

« C'est dans le moule de l'action que notre intelligence a été coulée » Henri Bergson

« L'information est le seul avantage compétitif » Jack Welsh – CEO de General Electric

« *Il est toujours sage de regarder en avant, mais il est difficile de regarder plus loin qu'on ne peut voir* » Winston Chuchill

L'informatique décisionnelle doit permettre d'avoir une vision claire, nette et précise du passé, le rétroviseur de l'image ci-dessus, mais aussi de comprendre la situation actuelle et de prévoir, simuler le futur.

- Passé : tous les utilisateurs du système décisionnel doivent pouvoir naviguer dans l'information dont ils ont besoin, depuis le détail, jusqu'au général. Ils doivent pouvoir faire des tableaux croisés, des histogrammes, des camemberts, etc. de manière intuitive pour faire des rapports et les partager. Un système décisionnel a notamment pour fonction d'être la mémoire de l'entreprise. La première fonction du décisionnel est de permettre à n'importe qui dans l'organisation d'accéder à l'information dont il a besoin, tout seul, de manière intuitive ; c'est-à-dire, offrir à tout le monde la possibilité, sans formation, en trois clics de souris, d'avoir une information précise. A la lumière du passé, on découvre souvent implicitement la voie à suivre.
- Présent : il est important d'avoir l'information juste de la situation actuelle, consolidée avec les informations sur le marché et sur les partenaires de l'entreprise, pour savoir où on est, par rapport aux autres, dans quel environnement. L'informatique décisionnelle doit permettre de comparer des données internes avec des données externes de façon fiabilisée.
- Futur : un système décisionnel intégrant des outils d'analyses de données et notamment de Data Mining, voir de recherche opérationnelle (optimisation) doit permettre de simuler le futur en connaissance de l'historique ; par exemple définir le potentiel d'achat d'un client au regard de son passé. Cette vision du futur reste une simulation avec une incertitude plus ou moins importante, néanmoins, elle est d'une très forte valeur ajoutée pour la prise de décision.

Les systèmes décisionnels doivent être une aide aux collaborateurs, leur permettant d'être proactifs sur leurs marchés, c'est-à-dire d'analyser, d'anticiper et de capitaliser sur l'expérience pour décider en fonction d'informations disponibles facilement et rapidement.

On peut comparer « conduire une voiture » à « piloter une activité ». Un véhicule doit :

- avoir un tableau de bord pour connaître la vitesse, les réserves d'essence, présenter des alertes, etc.
- offrir une vision si possible claire de la route, de ce qui se passe devant.

- Des initiatives, des leviers d'action, des commandes pour piloter.

Dans le cadre du management, à défaut d'avoir une vision limpide du futur, des outils de simulation permettent de scénariser ce futur. Comment piloter si l'on ne sait pas où l'on va ? Les outils d'informatique décisionnelle permettent de mettre à la disposition de tous les utilisateurs la connaissance du passé (la mémoire), une définition concise, relative et précise du présent ; et différentes prédictions pour l'aide à la décision. En connaissance de ces informations, il est alors possible de définir une carte routière avec des objectifs à atteindre et de contrôler si l'on est dans la bonne direction.

Les mots « stratégie » et « management » sont à mon avis beaucoup trop utilisés, à tort et à travers et donc galvaudés. On parle de « Corporate Performance Management », de « Business Performance Management » ou bien d' « Enterprise Performance Management » et si l'on recherche les définitions précises, chacun a la sienne mais ce n'est pas forcement intelligibles, précis et rigoureux.
Mais qu'il y a-t-il derrière ces écrans parfois fumeux. Si l'on regarde les différents sites ou si l'on lit les ouvrages traitant du sujet, on y retrouve beaucoup de choses intéressantes, souvent issues du bon sens.

La gestion de la performance implique d'utiliser des méthodologies, pour mesurer le système d'information

Mais un élément primordial dans la définition de la stratégie, souvent oublié, c'est d'associer aux objectifs des initiatives appelées aussi levier. En effet, si un indicateur tourne au rouge dans notre tableau de bord mais que l'on n'y a pas associé de levier d'action, c'est comme être dans une voiture lancée à pleine vitesse où l'on voit le mur arriver mais où l'on n'aurait pas la possibilité d'appuyer sur le frein !

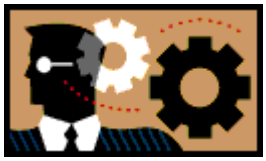
Petit exemple simplifié :
- objectif : Qualité de service après-vente
- responsabilité : Le responsable du service après-vente (comme son nom l'indique)
- KPI (Key Performance Indicator) : délais moyen de réponse
- Initiative : recrutement de ressource - optimisation de l'affectation des appels – capitalisation sur l'expérience – etc.

Si l'indicateur tourne au rouge, il y a un responsable qui doit avoir des moyens pour agir ; c'est de sa responsabilité.

L'indicateur « délai moyen de réponse » seul ne sert à rien s'il est juste là pour l'autosatisfaction du chef et son plaisir de taper sur ceux qui sont en dessous.
Un indicateur est pertinent lorsqu'il est lié à une stratégie et à des moyens pour y parvenir.

Dans un projet, afin de réduire le nombre d'indicateurs aux indicateurs clés de performance (KPI), ceux qui doivent être retenus sont ceux sur lesquels on peut agir. Ce sont les mesures de performance utiles. Les autres sont notamment du « flicage » inutile !

Autre point très important : il faut avoir une définition précise des indicateurs. Par exemple, si l'on parle d'un Turnover en ressources humaines, presque tout le monde comprend le mot ; sauf que dans la réalité, il est rare de trouver des personnes ayant la même définition, la même formule de calcul. Pour certains, le turnover sera le nombre d'employés quittant l'entreprise sur le nombre total d'employés de l'entreprise ; ce nombre total d'employés quittant l'entreprise comprenant ceux qui démissionnent, ceux qui arrivent en fin de CDD, ce qui partent à la retraite ou bien ceux à qui l'on a indiqué la porte de sortie. Pour d'autres, le turnover sera réduit au turnover volontaire et donc ne seront pris en compte que les employés qui démissionnent sur le nombre total d'employés de l'entreprise. Ce ne sont là que des définitions simples du turnover, et vous pouvez en trouver de multiples autres. La morale de ces deux petits exemples est que pour chaque indicateur il faut absolument avoir sa définition, son objectif (pourquoi on l'utilise), sa formule de calcul, qui en est responsable et quels sont les plans d'action à mettre en place si cet indicateur tourne au rouge.



Un outil décisionnel ne doit pas être un simple outil de contrôle mais bien un outil de partage de la connaissance, de communication, de prévision et de simulation pour l'analyse et l'amélioration de la performance des organisations. Les indicateurs doivent être définis précisément et reliés à des actions – des initiatives.

Quelques mots sur le décisionnel

On trouve maintenant des systèmes opérationnels dans presque tous les domaines. Certaines entreprises, étant dans un milieu encore plus concurrentiel et plus complexe, ont déjà installé des outils d'informatique décisionnelle. Néanmoins très peu de grands groupes ont une plateforme décisionnelle d'entreprise intégrée leur permettant d'avoir une vision transversale, intégrant des prévisions, des simulations et de l'optimisation, pour l'analyse et la gestion de la performance. De plus, les changements sont perpétuels et les systèmes décisionnels sont donc en perpétuelle évolution.

Une architecture d'entrepôt de données décisionnelle, intégrée et ouverte, a pour but de fournir une information de qualité aux décideurs de l'entreprise et de faciliter la prise de décision stratégique. Bâtir cet entrepôt pour répondre à des problématiques métiers, est réputé être avant tout un projet technologique complexe. En effet, ce type de projet nécessite des capacités d'intégration fortes compte tenu de l'environnement très souvent hétérogène des systèmes d'information opérationnels et de la complexité des données.
Un système décisionnel intégré permet de collecter l'information de différents systèmes opérationnels afin d'avoir une vision globale unique, sans remettre en cause les systèmes opérationnels actuels. Dans le jeu des fusions - acquisitions, le système décisionnel doit faciliter l'intégration des différentes composantes. On retrouve souvent de multiples systèmes d'information opérationnels à travers les organisations, sociétés, groupes, etc. et un outil de Reporting attaché à chaque système.
Un véritable « entrepôt de données d'entreprise », traduction d' « Enterprise Data Warehouse », doit offrir une vision unique, globale, transversale et du général au détail. Le terme « entreprise » a été rajouté pour insister sur la notion d'unicité et de transversalité, c'est-à-dire, que ce n'est pas un entrepôt qui est la somme de différents sous-ensembles. Dans trop de réunion, les chiffres de deux directions sont différents, ce qui créer des conflits, car il n'y a pas une vision unique et globale.

Un autre point de vue pour discuter de systèmes décisionnels intégrés, est le constat que dans beaucoup de cas on applique un outil de Reporting à chaque application opérationnelle et l'on se retrouve avec autant de système décisionnel que de système opérationnel. Les rapports de base sont bien fournis, mais on est très loin d'avoir une vision globale et transversale.

Un entrepôt de données d'entreprise doit donc impérativement faire passer toutes les informations par un point unique afin de

- les intégrer,
- les consolider,
- les valider,
- en vérifier la qualité,
- et les standardiser.

Techniquement, on utilise généralement un espace dédié que l'on appelle DDS (Detail Data Store). Le DDS est donc un espace stockant une quasi copie de toutes les données issues des différents systèmes opérationnels, internes ou externes. On l'appelle aussi souvent l'ODS (Operational Data Store) mais cela peut être confondu avec l'ODS (Output Delivery System) de SAS.

Remarque : Ce Data Store est généralement modéliser en troisième forme normale[1].
Notons que le DDS est de plus en plus souvent utilisé mais qu'il n'est absolument pas nécessaire dans tous les cas. Il existe de nombreuses théories sur la modélisation des bases de données des

[1] On 'normalise' une base de données afin de la rendre robuste, notamment en supprimant les redondances.

entrepôts de données décisionnels. La qualité fondamentale nécessaire pour faire cette modélisation est le BSP. Sans cette faculté, on complique et on intellectualise trop le problème. Revenons donc à la base, l'objectif de la mise place d'une plateforme décisionnelle intégrée d'entreprise et de permettre à tous les collaborateurs de l'entreprise, pour la plupart non informaticiens, en quelques clics de souris d'accéder à l'information dont ils ont besoin pour analyser les situations, les comprendre et donc améliorer la performance de l'organisation. Et pour couronner le tout, il faut une information unique et de qualité.
Mais qu'est-ce que le BSP ; une valeur trop souvent oubliée au profit d'intérêts personnels. Si tout le monde possédait cette qualité, ce serait fantastique. Le BSP, c'est le « Bon Sens Paysan » !

Informatique décisionnelle : système d'aide à la décision ?

Un système décisionnel est par définition un système d'aide à la décision. Cela ne veut certainement pas dire qu'un tel système prend des décisions à la place d'êtres humains et aussi intelligents soient ces systèmes, ils ne nous remplaceront pas. Par contre, au lieu de prendre des décisions dans le flou, les décisions seront prises en connaissance de cause. Il ne va pas sans dire qu'une telle transformation dans le processus de prise de décision, ne se fait que très rarement sans réticences. Il est donc important d'intégrer un vrai processus de gestion du changement pour amener l'ensemble des décideurs à utiliser la connaissance apportée par le système décisionnel. Dans un monde de plus en plus complexe, il est devenu vital d'utiliser la technologie pour exploiter le déluge d'information à des fins d'analyse de la situation passée, présente, et future, pour prendre la bonne décision au bon moment.

Prenons l'exemple d'une grande banque française qui souhaite modéliser le comportement du client afin de prévenir d'un éventuel départ à la concurrence. Pour ce faire, sur un historique de plusieurs années, un modèle de Data Mining a été construit. Plus précisément, connaissant presque pour chaque client quelques cinq cent cinquante variables comme, le revenu moyen, le fait d'être propriétaire, l'activité moyenne du compte principal, les différents produits bancaires détenus, leur utilisation maximale et moyenne, l'âge du client, sa situation familiale, etc. il est possible de modéliser sans trop se tromper, le fait que les clients ayant toutes ces caractéristiques sont des clients fidèles, alors que ceux dans telle ou telle situation, sont fortement susceptibles de partir à la concurrence. Sur un historique de plusieurs années, on connaît les clients qui sont encore clients et ceux qui ne le sont plus.
Les outils de Data Mining nous permet, par apprentissage sur l'historique, de modéliser le comportement du client, et donc de prévoir sa probabilité, son potentiel de partir à la concurrence.
La gestion du changement s'est déroulée, pour cette banque en plusieurs étapes. Tout d'abord, sur un ensemble d'agences pilotes sélectionnées au hasard, sur l'application opérationnelle des responsables de compte, une cellule a été rajoutée dans leur écran opérationnel. Cette cellule comprenait la note du potentiel du client à partir. Chaque responsable de compte a alors, lorsque qu'il ouvre la fiche d'un client, en plus des informations classiques, cette note du potentiel du client, de clôturer ses comptes.
Après six mois d'une phase de test sur ces agences pilotes, le retour des responsables de compte a été synthétisé de la manière suivante. Pour l'ensemble des clients ayant une note élevée de potentiel de partir à la concurrence :

- Dans environ 30% des cas, fort d'une expérience pouvant aller jusqu'à quarante ans de bon et loyaux services, le banquier n'avait pas besoin de cette note pour savoir qu'il y avait un risque de départ. Il connaît certain profil type.
- Dans environ 30% des cas, cette note élevée était une fausse alerte. Le banquier après avoir questionné longuement le client, n'y a pas relevé de risque réel.
- Dans environ 40% des cas, le banquier a admis que s'il n'avait pas eu cette information, il n'aurait rien fait pour comprendre son client et essayer de le fidéliser, alors qu'il y avait un risque.

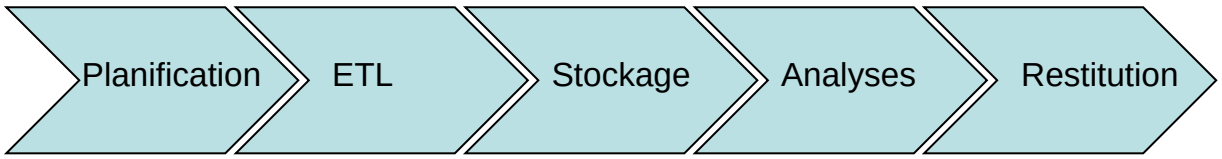
Pour résumer, dans 60% des cas, notre modélisation n'apporte aucune connaissance, voire pire, dans 30% des cas, nous fait perdre du temps. Néanmoins, si l'on compare la performance de ces agences pilotes par rapport à l'ensemble des agences du groupe, les agences pilotes ont divisé par quatre le taux de départ. Sachant que pour ce groupe bancaire, les banquiers ont une prime en fonction du nombre de clients, il est alors possible de gérer le changement en présentant de manière chiffrée, le gain apporté par le décisionnel, notamment sur la prime de fin d'année de chaque collaborateur.

Pour déployer cette information issue du Data Mining à tous les collaborateurs, après ce test, il est possible de leur présenter de la façon suivante : « Voici un outil d'aide à la décision, cela ne remette pas en cause votre connaissance client, votre savoir-faire. Nous ne sommes pas là pour dire que vous faites mal votre travail et que l'on doit vous assister. Grâce aux techniques de Data Mining qui ont pu analyser des millions de clients, nous avons un modèle qui permet de connaitre la probabilité d'un client de partir à la concurrence. Si la note que vous avez pour un client est faible, il semble être fidèle et inversement, si son score[1] de churn[2] est important, il a un fort potentiel de partir à la concurrence.
Nous avons testé ce modèle sur 100 agences pendant 6 mois, et avons obtenu les résultats ci-dessus, mais surtout, si vous faites attention au modèle, vous devriez diviser par 4 votre taux de départ, ce qui, vu vos objectifs, est assez significatif sur votre prime de fin d'année. »

Dans la plupart des projets décisionnels, il est important de procéder par étape et par itération afin de gérer correctement le changement. On définit un pilote, on analyse les retours sur investissement générés, et si les résultats sont suffisamment probants, on déploie sur un périmètre plus large avec des chiffres justificatifs à l'appui.

Définition de l'informatique décisionnelle :

L'informatique décisionnelle permet de transformer la donnée en connaissance.



La chaine du décisionnelle

Nous allons aborder l'informatique décisionnelle en cinq grandes étapes :

1. La panification : Mettre en place une plateforme décisionnelle est généralement un projet complexe, il est donc important de commencer sur des bases solides. Il faut pour cela avoir une méthodologie bien définie.
2. ETL (extraction – transformation – loading) que l'on peut traduire par extraction – transformation et chargement. La chaine du décisionnelle commence par l'intégration, la validation de la qualité des données, la transformation et le chargement de l'entrepôt de données.
3. Une plateforme décisionnelle doit souvent gérer d'importants volumes de données. Les bases de données décisionnelles se mesurent parfois en dizaines, centaines de téraoctets voire en pétaoctets. La performance du stockage de cet entrepôt est cruciale.
4. Analyse : le décisionnel doit permettre de transformer des données en connaissance ce qui nécessite généralement des outils de statistiques, voire de Data Mining, d'économétrie ou de recherche opérationnelle.
5. Restitution : la chaine du décisionnel se termine par la diffusion de la connaissance à travers un portail, un outil permettant à des utilisateurs non informaticiens de faire de façon intuitive en quelques clics des rapports, des analyses et de les partager.

[1] Score : le score se traduit en français par la note.

[2] Churn : C'est le potentiel de départ de client. Le score de churn est la note, souvent la probabilité du potentiel du client de partir à la concurrence.

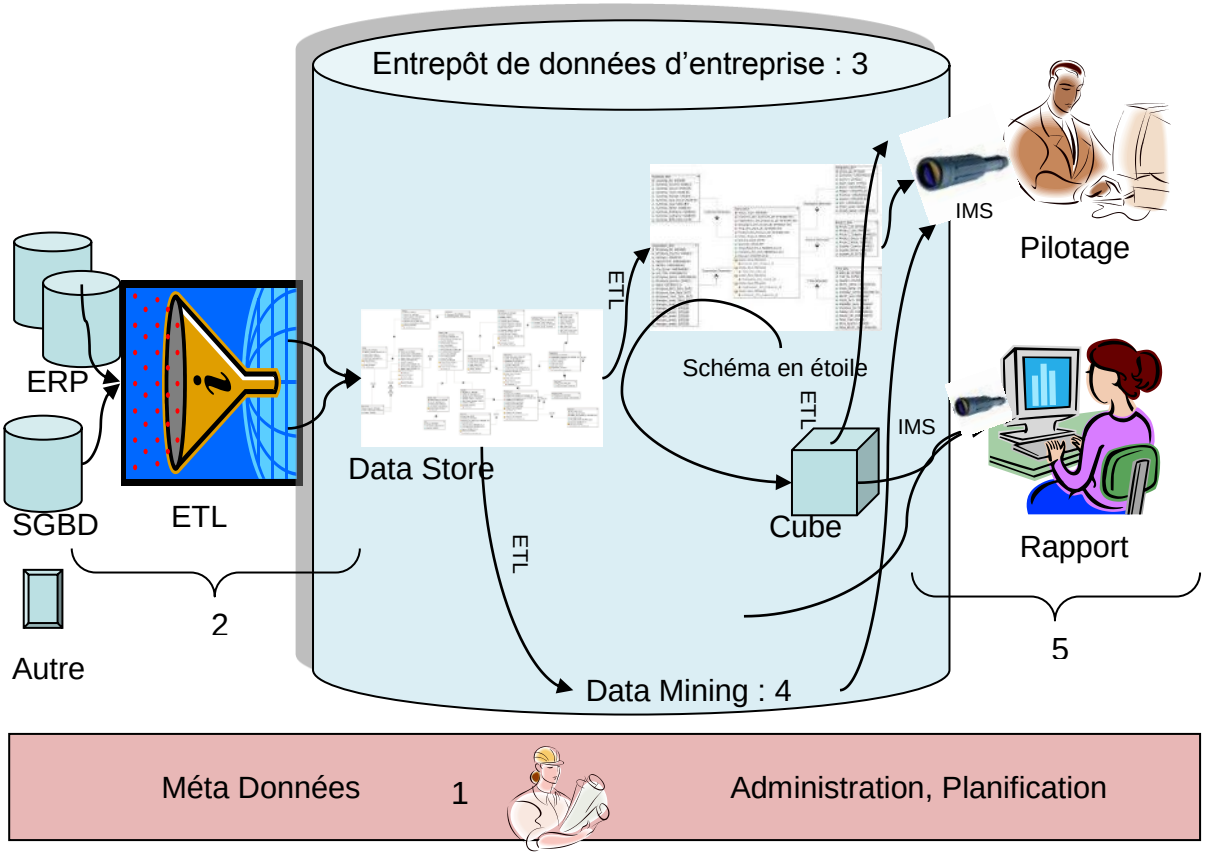


Schéma de principe de l'informatique décisionnelle :

« Un Data Warehouse ne s'achète pas, il se construit » Bill Inmon

Si deux entreprises du même secteur ont généralement des systèmes opérationnels de conception similaire, ces mêmes entreprises ont généralement des systèmes décisionnels différents, proche de leur philosophie respective. Néanmoins ces mêmes entreprises ont généralement des problématiques similaires auxquelles des solutions décisionnelles verticales[1] peuvent répondre.

Description du schéma de principe de l'informatique décisionnelle

1) Métadonnées

Pour mettre en place une plate-forme décisionnelle d'entreprise intégrée, la première étape est donc la planification de ce projet. Un tel projet nécessite une administration solide. Prenons un exemple sensible, les ressources humaines. Un responsable peut voir le salaire des personnes de son équipe mais pas celui de son chef. Ceci nécessite l'élaboration d'une stratégie de sécurité rigoureuse. Autre exemple : on me présente un rapport, mais quelle en est la qualité ? Si l'on a mis en place un vrai système d'administration et de gestion des métadonnées, on a alors un outil de traçabilité de l'information de la source à la cible et vis-versa. Sans cette traçabilité, les rapports, les tableaux de bord ont peu de valeurs. La traçabilité ne garantit pas forcement l'exactitude de l'information fournie, mais permet de vérifier le processus de création et cette traçabilité permet de garantir qu'il n'y a pas eu de falsification. Pour mettre en place cette traçabilité, il faut mettre en place une administration rigoureuse. L'administration est donc une première étape contraignante, qui nécessite beaucoup de

[1] Ici, dans le cadre de ce document, les solutions décisionnelles métiers horizontales font référence aux différents métiers de l'entreprise (i.e. : le marketing, la finance, les ressources humaines, l'informatique, etc.) alors que celles dites verticales font références aux secteurs d'activités (i.e. : la banque, les laboratoires pharmaceutiques, les télécommunications, le secteur public, etc.)

rigueur, mais qui si elle est bien faite, permet que le processus se déroule de la meilleur façon possible. Le socle de base d'une plateforme décisionnelle est une administration bien pensée. Trop de projets décisionnels relèvent plus du bricolage que du professionnalisme.

2) ETL

ETL : Le terme « ETL » est maintenant rentré dans le langage courant, mais pour présenter la démarche, on pourrait dire IVDT:

a) Intégration

Plus que d'extraire de l'information, il faut s'intégrer aux systèmes sources, ce qui implique une communication bidirectionnelle. Le système opérationnel fournit des données brutes, de détaillées, et le système décisionnel doit parfois en retour fournir de la connaissance à forte valeur ajoutée au système opérationnel. Par cette remarque, je souhaite insister sur le fait que le système décisionnel n'est pas supérieur, au-dessus du système opérationnel, mais bien complémentaire. Enfin, s'il n'y a pas de système opérationnel, il ne peut pas y avoir de système décisionnel.
L'intégration entre les deux systèmes peut se faire au niveau de la métadonnée, et non simplement au niveau de la donnée. Plutôt qu'un long discours, prenons le cas d'un projet de mise en place d'un système décisionnelle pour l'amélioration de la performance des ressources humaines. Les problématiques décisionnelles ressources humaines classiques sont la baisse du turnover volontaire, du délai de vacances des postes, l'optimisation du système de rémunération, l'amélioration des compétences, etc. Il existe des solutions horizontales comprenant des squelettes de Data Warehouse, de cube, de modèles de Data Mining, de rapports, de tableaux de bord, etc. ; par exemple : SAS Human Capital Management. Pour cette solution, au lieu de se connecter à toutes les tables du système d'information ressources humaines opérationnel, si celui-ci est par exemple SAP HR, il suffit de connecter la solution SAS HCM à SAP HR pour que l'intégration au niveau des métadonnées se fasse.
Le premier avantage de cette intégration est le gain de temps d'implémentation. En effet, il suffit de définir la racine de SAP HR, et l'ensemble des tables nécessaires au chargement du Data Warehouse, dont la solution SAS HCM fournie le squelette, sont automatiquement reconnues.
Le second avantage porte sur la maintenance du Data Warehouse. Comme l'a aussi dit Bill Inmon « le plus dur dans un Data Warehouse, ce n'est pas de le construire, mais de le maintenir ».

Le fait de travailler au niveau des métadonnées, permet une simplification de l'administration et de la traçabilité, donc normalement, de la maintenance de ce Data Warehouse.

b) Validation

Si l'on souhaite avoir des rapports de qualité, commençons par analyser la qualité des données, il y a-t-il des doublons, des valeurs aberrantes, manquantes. La validation et la consolidation de l'information est une étape complexe cruciale. Elle permet notamment de définir des processus de normalisation ou de standardisation. On parle beaucoup de Quality Master Data Managment (QMDM). SAS a racheté le leader : Data Flux en l'an 2000. Si vous avez des données de mauvaise qualité, vos rapports n'auront aucune valeur, et l'on ne parle même pas de la pertinence des modèles de Data Mining. De plus, pour garantir cette qualité, il faut une administration permettant la traçabilité.

c) Définition de la structure Cible

Définition de la structure Cible en fonction du cahier de spécification détaillé. Trop de Data Warehouse ont été conçu par des techniciens et non des fonctionnelles. Il ne faut pas regarder la structure des bases de données sources et progressivement concevoir le Data Warehouse. Pour concevoir un Data Warehouse qui réponde correctement à des questions métiers, il faut naturellement partir de ces questions métiers pour le concevoir. Nous en reparlerons au chapitre « La planification du *Data Warehouse »*

d) Transformation

Création des processus d'extraction, de normalisation des données sources de transformation et de chargement dans la structure cible. Après d'une part, s'être intégré aux données sources et en avoir

validé leurs contenus, et d'autre part, avoir définir la structure cible en fonction du cahier des spécifications, on peut à présent développer les processus de transformation-normalisation.

Même si le processus développé est d'ETL (Extraction - Transformation - Chargement), la méthodologie la plus couramment employée est de d'abord s'intégrer au système source, puis de définir la cible en fonction du cahier des spécifications et enfin de créer les processus de transformation permettant de charger la cible. Il ne faut pas tenter un processus linéaire qui partirait de la source et arriverait progressivement à la cible, ou bien inversement ; cela et ne permet pas de bien construire un Data Warehouse.
NB : On retrouvera aussi entre les différentes étapes du Data Warehouse des processus ETL.

Un système décisionnel ne doit pas simplement extraire des données des systèmes opérationnels, il doit s'y intégrer, c'est-à-dire qu'il doit y avoir une communication dans les deux sens entre ces applications. Il faut donc extraire la donnée, la métadonnée associée et souvent réinjecter des informations consolidées. Par exemple, on va extraire des données sur le client depuis différentes bases de données comme celles des services commerciales, comptabilité, après-vente, contentieux et autres, transformer, stocker et analyser ces données. La note pour chaque client de son potentiel de départ, issu de ces analyses pourra être ajoutée à la base de donnée opérationnelle.

Les processus d'alimentation d'un entrepôt de données peuvent être décomposés en cinq étapes :

1. Définition des données à extraire

La première étape à considérer dans le processus de mise à disposition de l'information, est l'extraction des différentes données des systèmes opérationnels et éventuellement des données externes.

2. Nettoyage

Les données extraites sont dans un deuxième temps nettoyées et valorisées avant d'être chargées dans le Data Warehouse. Ce processus permet d'en assurer la qualité et la validité.
Ce nettoyage est difficile à définir apriori, l'optimisation du processus d'extraction ne peut souvent se faire qu'après une analyse. On obtient donc le cycle : extraction des données, analyse, donc détection des données les plus pertinentes et celles inutiles, amélioration du processus d'extraction.

3. Définition de l'architecture cible - Indexation

L'ensemble des données et des processus de construction est documenté dans un dictionnaire unique de métadonnées. Ce dictionnaire d'adresses vers des données permet une restitution plus rapide de l'information.
Ces métadonnées (ici synonyme de "données sur les données ") sont techniques (structures et contenu de la base) et fonctionnelles (caractéristiques liées à l'usage fait de l'information et différenciant la donnée brute de l'information finale).

4. Création d'un flux de processus et test

Ayant importé la structure des données sources et ayant défini l'organisation cible, on peut maintenant créer un processus d'extraction – validation – transformation – chargement.
On exécutera une première fois ce processus afin de le valider.

5. L'automatisation des processus d'extraction

Un fois les processus définis, il faut les ordonnancer.

ETL et ELT

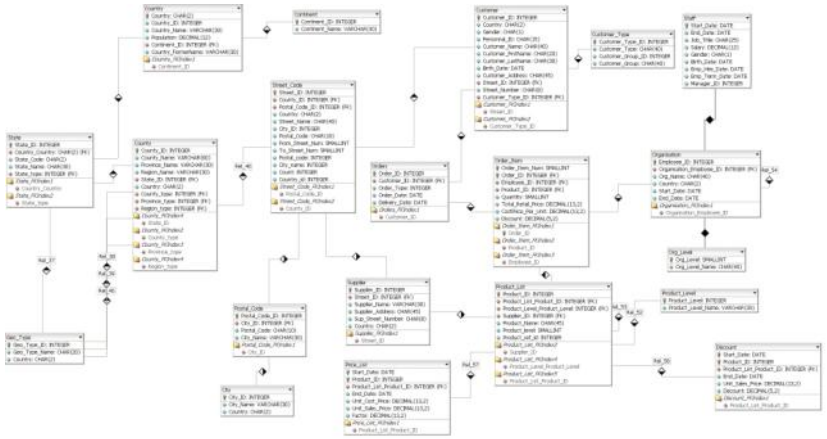
Le sigle ETL est utilisé pour les outils ayant un moteur interne et ELT où l'on a inversé deux lettres est utilisé pour les outils n'ayant pas de moteur interne et se basant sur les SGBD sources et/ou cibles. En fait, les deux approche ayant chacune leurs avantages et leurs inconvénients, on utilise maintenant avec un ETL comme SAS Data Integration Studio les deux approches. C'est-à-dire que parfois, l'outil va extraire les données sources, les transformer dans son moteur puis les charger dans la structure cible ; ou bien, l'ETL va être utilisé en ELT, c'est-à-dire qu'il va envoyer le code, généralement SQL au SGBD qui va l'exécuter.

3) Stockage

Il existe de multiples manières de stockée la donnée dans un data Warehouse. Chacune ayant ses avantages et ses inconvénients. L'administrateur des bases de données décisionnelles pourra notamment choisir entre :

a) DDS

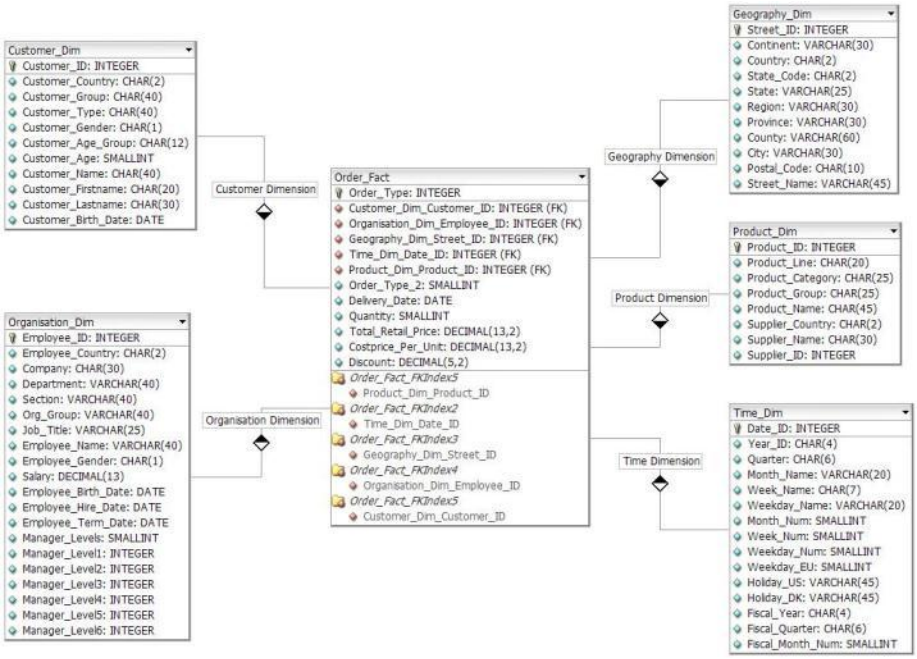
Detail Data Store est de plus en plus un élément permanant du Data Warehouse. On parle parfois aussi d'ODS (Operational Data Store) mais avec la plateforme SAS, cela prête à confusion avec l'ODS (Output Delivery System) de SAS Base. Le DDS est modélisé de manière classique, normalisée. Une grande partie du DDS est une copie des données de production. On y trouve un référentiel (un dictionnaire) unique permettant de standardiser la donnée. S'il a longtemps été qu'un simple élément tampon entre l'opérationnel et le décisionnel, c'est maintenant le point clé unique par lequel passe toute l'information ce qui permet de garantir une information globale, transversale et de qualité. Si pendant longtemps le schéma en étoile a été la référence en termes de conception de Data Warehouse, ce n'est plus forcément le cas. Depuis le livre fondateur de Bill Inmon de 1992 "Developing the Data Warehouse" (Construire l'Entrepôt de Données), on a entendu en boucle, pour concevoir un système opérationnel, il faut normaliser, pour un décisionnel, il faut dé-normaliser avec un schéma en étoile. Le socle de base de la plupart des « Enterprise Data Warehouse » est maintenant une base de données normalisée. Prenons l'exemple d'une multinationale du secteur High-tech pour qui, juste après l'élection de son PDG, celui-ci s'est aperçu que plusieurs indicateurs globaux n'était pas disponibles simplement, car il n'y avait pas un « Enterprise Data Warehouse », mais plus de 600 « Data Warehouse ». Le fait de rajouter le terme « Enterprise » devant l'expression « Data Warehouse », hormis le fait que l'on conçoive aisément que cela soit plus cher ; permet d'insister sur l'intégration de toutes les composantes d'un groupe international. Pour se faire, le travail de conception est souvent pharaonique. Il faut trouver et comprendre toutes les différences entre les différents éléments. Trouver et comprendre les différentes règles de calcul du chiffre d'affaire, entre les différents pays, cartographier la structure des fournisseurs et des clients afin de mieux comprendre les relations de dépendance existantes entre différentes sociétés locales et des groupes internationaux. Pour bien construire un DDS (ou ODS) il faut donc intégrer dans une base normalisé et historisé (pas orienté métier), toutes les informations nécessaires au Enterprise Data Warehouse. L'intégration de la connaissance grâce au DDS permet d'avoir une vision transversale des problématiques de l'entreprise.



b) Les schémas en étoile

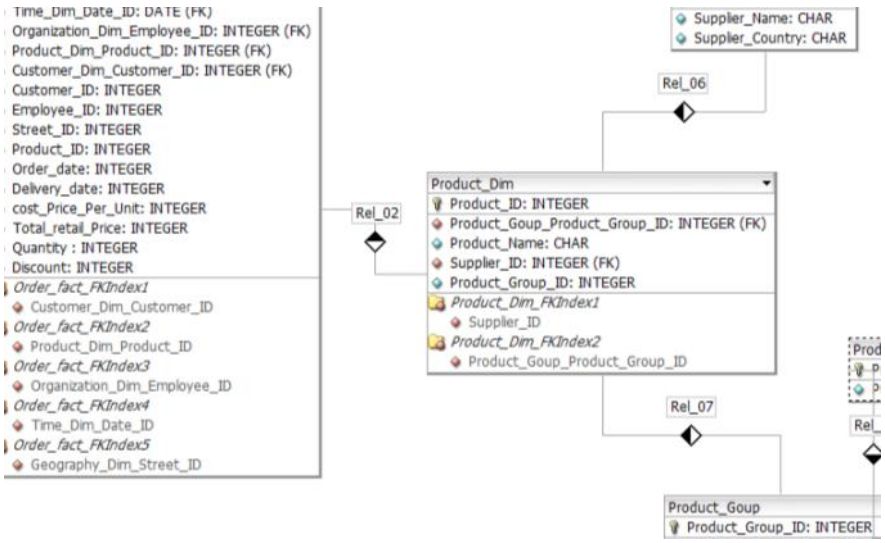
Cette modélisation de base de données classique en informatique décisionnelle est particulièrement utile pour permettre une utilisation facile par les utilisateurs. La table centrale comprend les faits. On l'appelle table de fait. Les tables autours sont les tables de dimensions. Par exemple, pour l'analyse des ventes, le fait principal est la vente, que l'on va analyser selon les dimensions du temps, de l'espace, des produits vendus, des clients ou bien encore de l'organisation des forces de ventes. Les schémas en étoile étaient particulièrement appropriés aux requêteurs des années 90. Les univers de

Business Object équivalent des Informations Map de SAS ou des modèles de Microsoft, les rendent aujourd'hui moins nécessaires.



c) Schémas en flocon

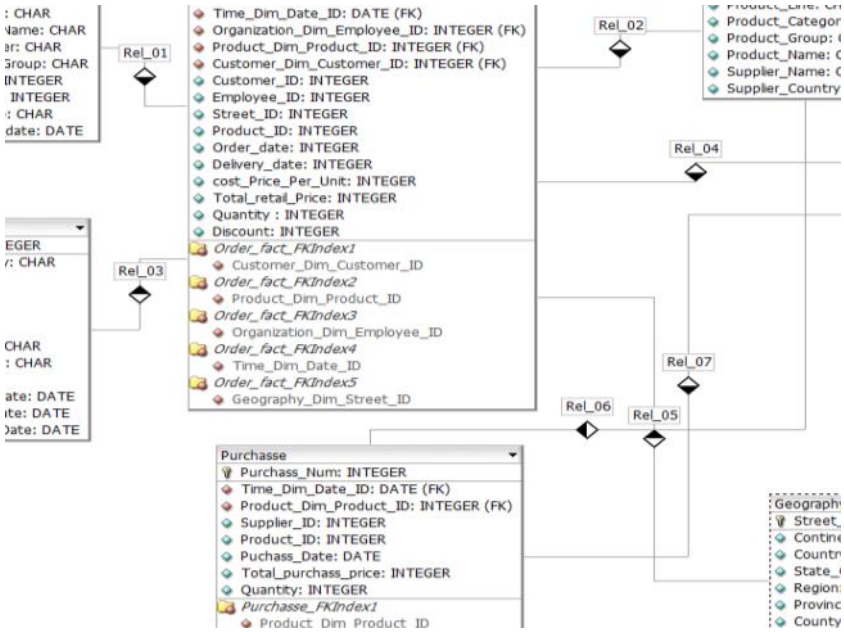
Les schémas en flocon, ressemblent aux schémas en étoile, mais les dimensions ne sont pas dé-normalisées. Concrètement, cela veut dire qu'ils sont moins simple d'utilisation (mais les outils actuels suppriment cet aspect) et moins volumineux. La dé-normalisation permet de simplifier l'utilisation pour des utilisateurs non avertis, mais créer des tables stockant de façon redondante l'information, ce qui implique une augmentation de l'espace disque nécessaire. On les utilise de moins en moins.
Réellement, la plus grosse table d'un schéma en étoile et la table de fait et l'on gagne relativement peu d'espace à dé-normaliser les table de dimension.



d) Schémas en constellation

Les schémas en constellation. On construit généralement un schéma en étoile par problématiques. Par exemple, un schéma en étoile pour l'analyse des ventes, un autre pour analyser les ressources

humaines, un troisième pour l'analyse des promotions, etc. ; tous ces schémas peuvent mutualiser des tables de dimensions communes comme le temps, ou celui des ventes et celui des promotions peuvent avoir la dimension produit en commun. On parle alors de schéma en constellation. Ce qu'il faut retenir, lors de la construction de système décisionnel, il faut prévoir des dimensions qui puissent être utilisées par différentes directions. Il est nécessaire pour cela d'avoir des tables de dimension conformes. Une table de dimension conforme à une clé primaire unique indépendante de tout système. La table de dimension produit peut aussi bien appartenir au schéma en étoile des ventes qu'à celui de la finance.



e) OLAP

Les hypercubes multidimensionnels OLAP, appelé souvent simplement cubes, sont des structures qui permettent de naviguer dans l'information suivant différentes dimensions, du général au détail, et vis-versa. D'un point de vue de l'utilisateur, les outils de navigation dans les cubes sont particulièrement conviviaux. D'un point de vue du gestionnaire des bases de données décisionnelles, ils permettent de supporter de multiples connexions simultanées et de toujours répondre rapidement à toutes les requêtes.

La technologie « OnLine Analytical Processing » (OLAP) permet des performances élevées pour les analyses multidimensionnelles et pour la génération de rapports sur de large volume de donnée. Dans la gestion de grands groupes, il est souvent nécessaire de naviguer dans l'information du général au détail, à travers le temps, etc. Un outil de Reporting basé sur une base multidimensionnelle permet une interaction rapide et intuitive pour naviguer à travers l'information.

Nous définirons donc le sigle OLAP comme étant équivalant à une base de données multidimensionnelle, ou un cube multidimensionnel. La technologie OLAP est donc une technologie de Système de Gestion de Base de Données (SGBD) conçue pour améliorer les performances techniques des problématiques de Reporting.

Offrir une application de Reporting permettant en un clic de passer au niveau d'agrégation supérieure à celui inferieur et vice-versa, sur une base de données relationnelle classique, génère automatiquement des requêtes longues et complexes. Dans un rapport du chiffre d'affaire généré par pays et par groupe de produit, il serait intéressant d'avoir en cliquant sur le pays le détail des régions qui le constituent, en cliquant sur un groupe de produit, d'avoir le détail par produit, de faire des sélections sur des périodes de temps, etc. Si beaucoup de personnes font cette simple manipulation en même temps, une base de données relationnelle classique tomberait.

Une façon très efficace d'avoir des temps de réponse à une requête excellent, est d'avoir calculé au préalable la réponse à la question. Le principe d'une base OLAP est de calculer la réponse aux futures questions lors de sa construction de la base.

Dans le cadre d'une base de données OLAP, toutes les combinaisons possibles de croisement et de sélections pourront être pré-calculés de sorte que lorsqu'un utilisateur fait une requête, la réponse à sa question ayant été calculée, lui est donnée immédiatement.
Calculer toutes les combinaisons possibles d'agrégation afin de répondre à toutes les réponses imaginables, peut très vite faire exploser la taille sur le (ou les) disque(s) de la base de donnée OLAP. Tout le sel de la technologie OLAP est de calculer suffisamment d'agrégats, mais pas trop, pour répondre à l'optimisation du système l'équation :

- Minimisation de l'espace Disque de la base OLAP
- Minimisation du temps de chargement (de calcul) de la base OLAP (problème corrélé à la première équation)
- Minimisation du temps de réponse à des requêtes utilisateurs

Prenons les trois cas :

1. dans le cas d'une base de donnée OLAP « pure », souvent appelée MOLAP (Multidimensionnel OnLine Analytical Processing), si on calcule tous les agrégats possibles, à savoir, toutes les combinaisons possibles de croisement, si la complexité de la base augmente, le temps de chargement et l'espace disque requis explosent, mais le temps de réponse aux requêtes utilisateurs, restent très faibles.
2. Dans le cas d'une base de donnée OLAP sans calcul des agrégats, ce qui revient à une base de données relationnelle, base souvent appelée ROLAP (Relationnel OnLine Analytical Processing), si la complexité de la base augmente, le temps de chargement et l'espace disque restent proportionnel à la base de données en entrée, mais le temps de réponse aux requêtes utilisateur peut exploser. Par exemple, si plusieurs utilisateurs souhaitent connaître en même temps le chiffre d'affaire, sur un groupe de produits, au niveau global, sur une période de plusieurs années, le temps de calcul de cette somme de l'ensemble des transactions sur cette période, peut exploser.
3. Le compromis entre une base MOLAP et une base ROLAP est souvent appelé HOLAP (Hybride OnLine Analytical Processing). Dans le cadre d'une base de données HOLAP, l'idée est de calculer suffisamment d'agrégat pour que les temps de réponse aux principales requêtes soit bon, mais de ne pas calculer trop d'agrégats, afin que la base n'explose pas en terme d'espace disque. La création de cette base nécessite donc de connaître au préalable les requêtes fréquentes, ce qui par définition d'un système décisionnel évolutif, est impossible. L'idée est donc de construire de manière itérative la base de données HOLAP en commençant sur le socle d'une base ROLAP. Au fur et à mesure des mises à jour de la Base OLAP, en fonction des journaux de l'utilisation cette base dans la période précédence, des agrégats supplémentaires pourront être calculés pour les niveaux ayant généré des temps de réponses insatisfaisants. Donc progressivement, en fonction de l'utilisation de la base, des agrégats sont calculés lors des mises à jour afin d'optimiser l'espace disque et les temps de réponse.

On interroge une base relationnelle avec une requête SQL et une base multidimensionnel avec une requête MDX. Le langage de requête MDX a été développé à l'origine par l'éditeur Microsoft et est maintenant le langage standard pour interroger les bases multidimensionnelles. Généralement, des générateurs de requêtes MDX permettent une utilisation conviviale sans jamais taper une ligne d'une telle requête. Toutefois, il y a toujours des cas particuliers dans chaque entreprise nécessitant le développement d'interfaces spécifiques, ou l'ajout de clauses particulières, ce qui peut demander au développeur, du code MDX.

Le sigle OLAP a été défini par Ted Codd, chercheur chez IBM, en 1993, au travers de 12 règles que doit respecter une base de données si elle veut adhérer au concept OLAP.

Les 12 règles de Codd sont :

1. Vue conceptuelle multidimensionnelle
2. Transparence
3. Accessibilité
4. Constance des temps de réponses
5. Architecture Client/serveur
6. Indépendance des dimensions
7. Gestion des matrices creuses
8. Accès multiutilisateurs
9. Pas de restrictions sur les opérations inter et intra dimensions
10. Manipulation des données aisée

11. Simplicité des rapports
12. Nombre illimité de dimensions et nombre illimité d'éléments sur les dimensions

Remarque, Le modèle de base de données relationnelle a été défini par ce Ted Codd en 1970, sur la base de la théorie des ensembles et de l'algèbre relationnelle. Edgar (Ted) Codd (1923 – 2003).
Quelques définitions importantes pour les bases OLAP :

Dimension : une dimension est un axe d'analyse de la base OLAP. Une base OLAP a généralement plusieurs dimensions, comme la dimension temps, la dimension géographie, la dimension produit, la dimension client, la dimension organisation, etc.
Une table classique a deux dimensions, des lignes et des colonnes, un cube multidimensionnel a trois dimensions : ligne, colonne, profondeur ; enfin, un hypercube a plus de trois dimensions. Par abus de langage, on appelle souvent un hypercube par simplement le terme cube.
Il est préférable qu'une base OLAP n'ait pas trop de dimensions. Il devient contreproductif de construire des hypercubes de quinze, vingt ou plus, dimensions. Revenons aux choses simples : pourquoi a-t-on créé des cubes, pour que des utilisateurs non informaticiens puissent simultanément et rapidement naviguer à travers l'information. Donc si l'on arrive à la construction d'hypercubes avec trop de dimensions, il y a de fortes chances que l'on puisse les diviser en deux sous-hypercubes. Aucun utilisateur ne créera un tableau croisé avec dix dimensions.

Un niveau : Chaque dimension a au moins un niveau, mais généralement plusieurs. En anglais : Level
Exemples :
Dans la dimension temps, on retrouve par exemple souvent les niveaux :
- Année
- Trimestre
- Mois
- Semaine
- Jour

Pour la dimension géographique
- Continent
- Pays
- Région
- Département
- Arrondissement
- Canton
- Commune
- Adresse

Pour la dimension client :
- Age
- Date de naissance
- Sexe
- Groupe de client
- Panier moyen
- Profession et Catégorie Sociale

Les niveaux ne sont pas forcément corrélés entre eux

Une hiérarchie. Une hiérarchie est un sous ensemble d'une dimension. Une dimension a au moins une hiérarchie, et peut en avoir plusieurs. Une hiérarchie a au moins un niveau. Une hiérarchie est le mode de navigation à travers l'information, proposé à un groupe d'utilisateur. Lors d'une requête, on ne peut utiliser qu'une hiérarchie par dimension. Attention au terme hiérarchie, il n'y a pas nécessairement de relation hiérarchique entre deux niveaux d'une même hiérarchie, même si le nom le laisse supposer.

Exemple de hiérarchie pour la dimension temps :
- Hiérarchie 1(rapide)
  - Année
  - Mois

- Jour
- Hiérarchie 2 (complète)
  - Année
  - Trimestre
  - Mois
  - Semaine
  - Jour de la semaine
  - Date
  - Tranche horaire

Deux hiérarchies d'une même dimension peuvent avoir des niveaux en commun.

Exemple de hiérarchie pour la dimension client :
- Hiérarchie 1
  - Profession et Catégorie Sociale
  - Groupe de client
  - Sexe
  - Age

Dans cette hiérarchie, il n'y a pas de notion hiérarchique. Lorsque l'utilisateur souhaitera plus de détails pour une PCS donnée donc lorsque qu'il cliquera dessus, il aura l'information par Groupe de client.

Une mesure : est une quantité numérique qui sera analysé. Par exemple, la somme, la moyenne, le maximum du chiffre d'affaire sont des mesures.

NWAY : Dans toutes les directions, donc selon toutes les dimensions. Un cube NWAY est un cube où toutes les agrégations, selon tous les croisements possibles, ont été calculées. Un cube NWAY est donc équivalent à un cube HOLAP.

Drill-through : Naviguer à travers, généralement, à partir d'agrégats.

Drill-down : descendre d'un niveau, aller du général vers le détail. On parle de zoom avant. L'inverse est drill-up : monter d'un niveau, aller du détail au général.

Slicing : extraction d'une tranche d'information d'un cube.

f) Virtuel

Du virtuel, il est possible des créer des outils de Reporting, qui se connectent directement sur les systèmes d'information opérationnels. Concrètement, cela permet d'avoir une information en temps réel et de ne rien stocker physiquement dans le Data Warehouse. Par contre, dans le cadre des requêtes complexes, une telle configuration peut s'avérer catastrophique en termes de performance et peut impliquer une surcharge considérable du système opérationnel. Il ne faut donc utiliser cette possibilité que dans de très rares cas. Concrètement, on créer une Information Map (ou un univers de Business Object) directement sur les données opérationnelles. Un avantage certain est d'offrir l'information en temps réel.

Récapitulatif sur le stockage

"Perfection is spelled P-A-R-A-L-Y-S-I-S" Winston Churchill

Un Data Warehouse peut-être défini de façons très différentes.
Un Data Warehouse d'entreprise est l'ensemble virtuel comprenant l'ensemble des données du système d'information décisionnel d'entreprise, couvrant toute l'activité de l'entreprise et son environnement. Il alimente ainsi tous les processus de décision stratégique de l'entreprise. Littéralement : entrepôt de données, toutes les informations décisionnelles transitent par lui.

Le Data Warehouse peut-être décomposé en Data Mart.

Un Data Mart se concentre sur une problématique départementale ou métier. Littéralement : magasin de données, il stocke les informations spécifiques à un domaine. Si l'on fait la comparaison avec un groupe de grande distribution ayant une fonction achat centralisée, tous les produits passent par l'entrepôt (Warehouse), mais chaque magasin offre une gamme personnalisée, les produits vendus dans le nord de la France, ne sont pas les mêmes qu'en Bretagne, ou dans le sud-ouest, par exemple.
Un " Data Store " ou " Operational Data Store " est une copie des données de production sans modification du modèle des données. Il permet d'exploiter les données de production sans gêner cet environnement. Pour ne pas avoir de confusion avec l'ODS de SAS Base, on parlera de DDS (Detail Data Store).
Le Data Warehouse est l'ensemble logique du DDS s'il y en a un et des différents Data Mart et sous Data Mart.
Remarque : On confond parfois Data Warehouse et DDS.

On trouve de plus en plus dans les entreprises qui ont une culture décisionnelle avancée des entrepôts de données avec une architecture à deux ou trois niveaux combinant ces trois formes complémentaires de stockage des données. Si tout le monde s'accorde pour définir un Data Mart comme un sous ensemble du Data Warehouse, exemple : un Data Warehouse Monde, des Data Mart par continent et des sous Data Mart par pays, ou, autre exemple, un Data Warehouse pour toute l'entreprise et des Data Mart par direction ; sur certain projet la différence de définition du domaine du Data Warehouse et de celui du Data Mart peut être différente entre les différents intervenants (société, intégrateur, éditeur de logiciel).

Ce qu'il faut retenir : peu importe le terme que l'on utilise pour définir l'endroit où l'on stocke toutes les informations à but décisionnel (Data Store, Data Warehouse, Data Mart, infocentre, ou autre) il est important de définir dans les spécifications du projet et dans les métadonnées, l'ensemble défini par chaque mot, son périmètre.

Remarque linguistique : on parle souvent de Data Warehouse pour désigner l'entrepôt de donnée, mais il se peut que sous ce même mot, on prenne en considération l'ensemble du processus décisionnel.

Remarque technique : Revenons sur le Detail Data Store, il n'est pas nécessaire qu'il y en ait un pour un Data Warehouse peu complexe. Par exemple, si le système d'information opérationnel est basé sur une seule base de données, la problématique d'intégration est moindre.

Dans d'autre cas, le Data Store n'est qu'une étape intermédiaire entre le système opérationnel et le Data Warehouse, sa durée de vie est alors du temps du processus ETL : c'est une copie temporaire des données opérationnelles. Mais de plus en plus, le Data Store est une composante du Data Warehouse, pour ne pas dire le socle de base du Data Warehouse. Pendant de nombreuses années, les experts en décisionnel ne juraient que par la modélisation en étoile, aujourd'hui, notamment selon l'un des pères du Data Warehouse : Bill Inmon, les méthodes classique de normalisation, sont remises au goût du jour pour la construction de Data Warehouse. En effet, afin que le Data Warehouse puisse s'élargir à toutes les directions opérationnelles de l'entreprise, intégrer de plus en plus de composante et être toujours performant, un socle modélisé de manière relationnelle et normalisé est très fortement conseillé. Cela n'empêche pas de construire par la suite des Data Mart modélisés en étoile.

Le but du Data Warehouse est de fournir un environnement de stockage optimisé pour les requêtes décisionnelles parfois complexes, accédant à des volumes de données importants, sans pénaliser les temps de réponse. Pour être exploitables, toutes les données vont être organisées, intégrées et enfin stockées pour donner à l'utilisateur une vue globale des informations : une information complète et transversale. Pour ne pas avoir une Connaissance Obsolète Généralement Non Orienté Stratégie, il est important d'ingérer de l'analytique ; nous reverrons ce thème plus tard.

Bill Inmon dans son ouvrage de référence « Using the Data Warehouse » définit le Data Warehouse de la façon suivante : «Le Data Warehouse est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision.»

Comme présenté au début de cette partie, on peut voir cinq types de structure de stockage des données dans un Data Warehouse ou dans un Data Mart :

1. Des fichiers dits « plats ». Un fichier plat est un fichier indépendant de toute base de données ne contenant que du texte. Ils sont très utilisés pour transférer des données d'un système à un autre car ils sont simples, pratiques, mais souvent, on perd les métadonnées associées. La chaîne de la traçabilité est alors perdue, ce qui peut avoir des conséquences importantes pour la qualité des données. Des tables simples avec des lignes et des colonnes, mais sans relation avec d'autres tables ; par exemple, des tables pour répondre à des questions spécifiques ou bien les tables pour faire les analyses statistiques, Data Mining, économétriques, etc.

2. Des tables relationnelles. Elles sont stockées dans un système de gestion de base de donnée (SGBD) qui gèrent notamment leur relation. Les « rollback segment » et la gestion des contraintes d'intégrités ne sont pas des problèmes décisionnels fondamentaux. Les SGBD opérationnels ne sont donc pas forcément les plus adaptés pour le stockage de Data Warehouse complexe, leurs performances sont moins bonnes car elles gèrent des problèmes opérationnels, ce qui les surcharge. Si on peut utiliser un SGBD opérationnel pour un petit Data Warehouse, ce n'est pas forcément le cas pour les Data Warehouse complexes. Il y a deux types de stockage relationnel décisionnel :
   a) les bases relationnelles normalisées du DDS (Detail Data Store ou ODS Operationnal Data Store)
   b) les base dé-normalisées des schémas en constellation.

3. Des structures multidimensionnelles (voir définition ci-dessous OLAP)

4. un ensemble virtuel (voir ci-dessous Map).

Donc, en fonction du besoin, de la quantité de données, des temps de réponses souhaités et acceptables, du budget, etc., différentes formes de stockage sont possibles et il conviendra de choisir le meilleur compromis.

4) Analyse de données

Nous considérerons ici l'analyse de données comme regroupant l'ensemble des techniques de statistiques, d'économétrie, de Data Mining, et de recherche opérationnelle. Mais cette définition n'est pas universelle d'autant plus que les traductions franco-anglaises sont parfois surprenantes. En effet, les mots français : « analytique » et « analyse », ou anglais : « analytics » et « analysis » sont utilisés de façon très diverses. Parfois, « analytics » est utilisé pour désigner des objets graphiques interactifs de tableau de bord comme des jauges, des baromètres, cadrans, tachymètre, compteurs, etc. D'autres personnes utilisent « analytics » comme l'ensemble des solutions décisionnelles ; depuis les solutions de gestion de la performance, en passant par les solutions métiers de gestion de la relation client, de contrôle de gestion, jusqu'aux plateformes de génération de rapports d'analyses statistiques, d'entrepôt de données, d'ETL, etc. Pour finir dans cette salade, certains considèrent que lire un rapport avec un tableau croisé et des graphiques, c'est les analyser, et ils considèrent donc le Reporting comme de l'analytique.
L'analyse de données demande souvent des compétences statistiques avancées. Néanmoins certaines solutions embarquent ces fonctionnalités pré-paramétrées à des cas de figures bien définies, afin d'offrir leur valeur ajoutée à des personnes fonctionnelles. Dans ce cadre-là, des personnes ni informaticiennes, ni statisticiennes, vont faire du Data Mining, de l'optimisation, sans compétences spécifiques.

L'augmentation la plus importante de valeur ajoutée de l'informatique décisionnelle, processus de transformation de l'information en connaissance, est l'analyse de donnée. Depuis le simple tableau croisé, jusqu'aux modèles complexes de Data Mining, l'analyse de données est la clé permettant d'extraire de la donnée, la substantifique connaissance. Beaucoup trop de projets décisionnels sont réduits au simple périmètre du Reporting et finalement, ne répondent pas aux problématiques d'analyse de la performance permettant une gestion éclairée.
Faire des tableaux croisés, de jolis graphiques, des statistiques descriptives, sont effectivement un bon début qu'il est nécessaire de compléter par des analyses plus complexe, depuis comme les analyses statistiques avancées, l'économétrie, le Data Mining ou la recherche opérationnelle.

Quelle sont les points clés pour bien intégrer l'analyse de données :

1) Intégration de données : il faut que l'information soit intégrée et accessible aux bonnes personnes. Il faut donc être capable d'administrer et de gérer le débit et la flexibilité des données dans l'organisation.
2) Avoir la bonne technologie éprouvée en place pour soutenir les analyses d'aide à la décision. Il faut réduire le temps de latence sans pour autant faire une analyse complète (parfaite) en une seule fois.
3) Recruter et développer les compétences dont vous avez besoin. Les ressources humaines sont surement le point le plus critique. Trop d'organisation croit encore que l'on peut devenir expert en analyse de données avec une petite formation. Même avec un esprit de mathématicien large et une culture analytique forte, le sujet est tellement vaste, qu'il faut en permanence apprendre pour ne pas être trop en retard. De plus, ces ressources sont rarement valorisées à leur juste valeur.
4) Demander à votre organisation des décisions basées sur des faits. Demandez à vos employés comment ils savent cela, comment cela va évoluer. Il faut dès à présent discuter de cela.
5) Avoir un processus transparent. La quantité de gens qui n'aime pas les statistiques et telle que le processus de changement est difficile. Il faut donc faire savoir vos succès, communiquer sur ce qui a fonctionné, à quel rythme, en combien de temps.
6) Développez un centre de compétence analytique. Il faut capitaliser sur les succès et les échecs.
7) Changer les mentalités. Si l'on fait tout cela correctement, le nombre de personnes promouvant l'analyse augmente. On rentre dans une organisation apprenante par test et apprentissage.
8) Redéfinissez votre stratégie : souvent. Dans un monde en perpétuel changement, il faut toujours s'adapter.

a) Statistiques

Statistiques descriptives où inférentielles. Les statistiques descriptives décrivent une information et celles inférentielles permettent de modéliser, souvent sur un échantillon, un phénomène connu, puis si certaines règles sont vérifiées, extrapoler le modèle sur une population plus large ou pour laquelle la réalisation de ce phénomène n'a pas encore eu lieu.

b) Recherche opérationnelle

La recherche opérationnelle regroupe notamment les algorithmes d'optimisation. Le cas le plus générique est de maximiser (ou minimiser) une fonction (un gain, une marge, un coût) sous différentes contraintes. La recherche opérationnelle permet de trouver l'optimum de problèmes souvent complexes.

c) Économétrie

En simplifiant, l'économétrie regroupe l'ensemble des techniques statistiques permettant d'analyser des grandeurs économiques, dont l'analyse des séries temporelles.

d) Data Mining

Le Data Mining est un terme, souvent galvaudé, à géométrie variable selon les interlocuteurs. Pour prendre une définition large, c'est un ensemble de techniques permettant d'exploiter des grosses bases de données et d'en extrapoler la substantifique moelle. Les techniques de Data Mining sont issues de techniques de statistiques que l'on a optimisées afin de pouvoir les appliquer à de gigantesques bases de données. Ces techniques peuvent par exemple être parallélisées sur plusieurs machines. Le Data Mining est à cheval sur les statistiques, l'informatique, la recherche opérationnelle, l'intelligence artificielle et les métiers sur lesquelles on applique ces algorithmes

« Le vrai génie réside dans l'aptitude à évaluer l'incertain, le hasardeux, les informations conflictuelles » Winston Churhill

Le *Data Mining* est un ensemble de techniques permettant d'extraire des masses de données importantes, des informations à forte valeur ajoutée. La traduction littérale du terme *Data Mining* en français, est fouille de donnée.

La plupart des techniques de *Data Mining* ne datent pas d'aujourd'hui, elles sont des années soixante et soixante-dix ; certaines ont même été définies dans les années trente. Néanmoins, la recherche dans ce domaine est actuellement encore très prolifique et certains nouveaux algorithmes, des années 1990 et 2000, viennent apporter de nouveaux concepts. Ce qui est nouveau, c'est la quantité des données disponibles, la capacité des machines à les traiter, et surtout, ce qui fait que le *Data Mining* est si populaire, c'est le retour sur investissement qu'il peut engendrer.
En effet, on peut constater que la quantité des données stockées, augmente beaucoup plus vite que la puissance de calcul des machines. D'après la loi de Moore's, le volume de données stockées double à peu près tous les deux ans. On enregistre les patients à l'hôpital, les commandes, les stocks, les transactions bancaires, les réservations pour le train, l'avion, etc. ; une liste exhaustive de la masse d'information hébergée dans les ordinateurs, serait bien impossible à créer. Le volume des données stockées est donc exorbitant, mais parmi ces données, se trouvent une mine d'informations qui ne demande qu'à être exploitées.

Je vous propose de définir le *Data Mining* par l'ensemble des techniques que l'on peut classer dans ce domaine. Je n'ai certainement pas la prétention d'avoir la définition universelle ; la littérature en propose de nombreuses. Les limites de ce domaine étant variables selon les auteurs, l'objectif de cette définition est simplement de préciser ce qu'il faut comprendre par *Data Mining* dans ce document.
On peut classer les techniques de Data Mining en deux grandes familles, les techniques descriptives qui permettent de décrire la situation actuelle, et les techniques prédictives qui, en apprenant sur le passé, simulent l'avenir.

- Les techniques descriptives :
    - Segmentation : Faire des groupes homogènes sans a priori afin de mieux comprendre l'essentiel d'une base de données.
    - Recherche d'associations : par exemple les associations de produit, le but est de trouver dans l'ensemble des transactions d'achats les associations de produits intéressantes du type, les clients qui achètent le produit A, achètent aussi le produit B, il y a tant de clients qui ont acheté ces deux produits, et parmi tous les clients ayant acheté A, tant ont acheté aussi B. Recherche de séquences : la recherche de séquences est identique à la recherche d'associations sauf qu'il y a la notion d'ordre. Par exemple, les clients ayant visité la page A de ce site Web, vont ensuite sur la page B.
- Les techniques prédictives : Les techniques prédictives permettent d'apprendre sur un historique connu, pour appliquer ceci sur le présent et ainsi simuler l'avenir.
    - Régression
    - Arbres de décision
    - Réseaux de neurones
        - Raisonnement à base de cas
- Et parfois, selon les auteurs, d'autres éléments.

Il est important de noter que ces techniques doivent être capables de traiter des très gros volumes de données et qui est important de pouvoir facilement intégrer les segmentations, résultat d'analyse et les *scoring*[1]dans les outils de production.

La modélisation par des algorithmes prédictifs, appelés aussi algorithmes à apprentissage supervisé permet d'apprendre sur quelque chose où l'on connaît la réponse afin de l'appliquer sur autre chose où l'on souhait une simulation.

Le Data Mining est le croisement de plusieurs domaines, dont les statistiques, les bases de données et l'intelligence artificielle. La limite entre ces différents domaines et le *Data Mining* est particulièrement mal définie.

[1] Un score est une note. Les modèles prédictifs de Data Mining permettent de noter un individu sur sa propension à, par exemple, être un bon client. Scorer une base de client, c'est utiliser un modèle mathématique ayant appris sur un historique, pour affecter à chaque client sa note, souvent sa probabilité, à acheter ce produit, ou bien à partir, etc. Faire du scoring, c'est donc apprendre sur un *historique*, pour simuler le futur.

On utilise souvent le Data Mining pour la détection de la fraude, les typologies de clients, la prévision des ventes, l'attrition[1], la fidélisation, les ventes additionnelles et croisées, ou bien l'analyse des tickets de caisse.

Le Data Miner, celui qui fait du Data Mining doit donc connaître le domaine dans lequel il travail, les données et les méthodes d'analyse.

Il important de noter que ce n'est pas parce que vous avez une énorme base de données et un super outil de Data Mining que forcément vous allez trouver des pépites d'or. S'il n'y a rien à trouver, si la base de données est de trop mauvaise qualité, vous ne trouverez rien.

SAS Entreprise Miner ne travaille pas exactement sur des tables SAS classique, mais sur une Data Mining DataBase (DMDB) qui est une copie reformulée de la table originale avec un catalogue de métadonnée afin de permette la parallélisation et l'amélioration de la performance.

Pour moi, la limite entre les statistiques et le Data Mining se trouve essentiellement dans la taille des bases de données traitées ce qui a plusieurs conséquences :

- En statistiques inférentielles, pour pouvoir tirer des conclusions des analyses faites sur un échantillon, il est souvent nécessaire de valider l'indépendance des variables alors qu'en Data Mining, vu que l'on travaille sur des gros volumes de données, le problème n'existe presque plus. Par exemple, pour un million d'individus, l'hypothèse d'indépendance entre deux variables au risque 5%, sera rejetée si le coefficient de corrélation linéaire est supérieur à 0.002 en valeur absolue ; ce qui est presque inutile de vérifier.
- Le Data Mining est surtout empirique, on recherche un bon modèle, on le teste et si cela fonctionne, on l'utilise. L'objectif est beaucoup plus de trouver rapidement une réponse opérationnelle que de faire une analyse théorique.
- Donc, en statistiques, étant donné que l'on travaille sur des petites bases de données, la rigueur est de mise. On répète la méthode « Hypothèse H0, etc. » avec une rigueur d'horloger alors qu'en Data Mining, la loi des grands nombres nous permet d'outrepasser ces problèmes.

Une différence majeure entre une régression de SAS/STAT (par exemple PROC REG, qui peut être lancé depuis SAS Enterprise Guide), et une régression de SAS Enterprise Miner (Par exemple PROC DMREG), réside dans le fait que la régression de SAS Enterprise Miner ne travaille pas sur une table SAS classique mais sur une Data Mining Data Base. En effet, tout processus de Data Mining commence par la procédure DMDB qui prépare l'analyse Data Mining ultérieure par notamment la génération de métadonnées permettant d'accélérer le processus. Presque toutes les procédures de SAS Enterprise Miner peuvent être parallélisées. Si l'on teste la performance des deux outils sur une petite base, sur un mono processeur, en mode Batch pour ne comparer que le temps d'exécution ; SAS Enterprise Miner est légèrement plus lent car il est préparé par la Data Mining Data Base pour la parallélisation ultérieure, ce qui ne sert à rien. Par contre, sur des très gros volumes, sur plusieurs machines multiprocesseurs, la surperformance de SAS Enterprise Miner sur SAS/STAT est incontestable.

Retour d'expérience personnelle : Pour une analyse de Data Mining solide avec deux arbres de décisions pour partager les individus classiques traités par six régressions, des individus atypiques, traités par deux réseaux neuronaux, sur une table environs 300 variables et 100 000 lignes ; il fallait 2 heures 47 minutes de traitement sur un ordinateur simple (un processeur, un disque dur). Le même traitement (2 arbres, 6 régressions, 2 réseaux de neurones), sur la même table (~ 300 colonnes), mais sur tous les clients actifs de la société, c'est-à-dire trois millions au lieu des cent milles précédemment, et sur une machine huit processeurs, 10 Giga de RAM ; 12 minutes de traitement.
Or, 2h47 que multiplient 30 et que divisent 8, sont largement supérieur à 12 minutes.

De plus, SAS Enterprise Miner utilise massivement la recherche opérationnelle. En effet, en travaillant sur des grosses tables avec beaucoup de variables, et parfois en travaillant sur des combinaisons de variables ; la complexité du problème explose. Un bon modèle est un modèle suffisamment complexe mais pas de trop, pour être robuste. La recherche opérationnelle est donc utilisée pour trouver cet

[1] Attrition (ou churn en Anglais) : Le potentiel d'attrition d'un client, est sa capacité à partir. Dans beaucoup de sociétés, des modèles de Data Mining permettent d'en calculer pour chaque client, sa probabilité

optimum par exemple entre des arbres à plusieurs branches de différentes profondeurs, des régressions utilisant plus ou moins de paramètres, etc.

Remarque : si l'on considère qu'une solution de Data Mining doit proposer ces différents algorithmes, et être capable de les appliquer sur des bases de données énormes, ce qui suppose, d'être capable de paralléliser sur plusieurs processeurs, sur plusieurs machines, sur plusieurs OS, en utilisant la recherche opérationnelle pour optimiser rapidement la complexité ; à ma connaissance, il n'y a que SAS. Mais je ne considère pas SAS comme étant le seul outil de Data Mining.
J'ai déjà rencontré un consultant qui en faisant un tableau croisé avec Excel, disait fouiller dans les données et donc se vendait comme expert en Data Mining.

5) Le Reporting

La génération de tableau de bord, est la partie émergée de l'iceberg de l'informatique décisionnelle. C'est la partie vue par la plupart des utilisateurs. Ce sont généralement de jolies interfaces intuitives permettant à un utilisateur lambda, en fonction de ces droits, de consulter des rapports, des tableaux de bord, de les annoter, voire de les créer lui-même. Voici quelques termes :

a) Reporting de masse

Les outils de Reporting de masse permettent aux masses de faire des rapports simplement. On crée pour cela des Informations Map qui offre des objets métiers afin que des utilisateurs fonctionnels puissent en quelques clics créer des rapports et les partager.

Une information Map est créée par un analyste métier (Business Analyst), une personne ayant des compétences techniques afin de comprendre la structure d'un Data Warehouse, et une connaissance métier pour créer des indicateurs (=des mesures) et des catégories (=des axes d'analyses), utilisables par les fonctionnelles. L'outil SAS pour créer les informations Map est SAS Information Map Studio. C'est l'un des produits de la suite BI de SAS. Les points communs avec Business Objet designer sont nombreux. Business Objet designer permet de créer des Univers.

SAS Web Report Studio est un outil que l'on peut qualifier de Reporting de masse, dans le sens où il permet des masses d'utilisateurs de créer des rapports, simplement, réellement quasiment sans formation.

b) Reporting à la demande

Les outils de Reporting Ad-Hoc[1], ou de Reporting spécifique à la demande, permettent à un utilisateur plus ou moins avancé de répondre à un besoin ponctuel. Dans plus de 80% des cas, un utilisateur lambda, sans formation, pourra répondre à ce besoin avec un outil de Reporting de masse comme SAS Web Report Studio. Néanmoins, il est parfois nécessaire qu'un analyste métier, une personne ayant des compétences plus importantes en analyse, fasse une analyse avec un outil plus puissant comme SAS Enterprise Guide. Il reste enfin quelques très rares cas où un développement spécifique peut-être nécessaire, auquel cas on sort de ce cadre du Reporting Ad-hoc pour passer sur un projet.

En résumer, les outils de Reporting Ad-hoc permette de répondre très rapidement à des questions ponctuelles. « Là, maintenant, tout de suite, j'ai besoin de telle information ».

c) Reporting en mode push ou en mode pull

Et oui, encore des anglicismes ; donc en français, on peut traduire cela littéralement en rapports poussés ou tirés ! Les rapports en mode push (poussés) sont des rapports envoyés automatiquement aux utilisateurs. Par exemple, la plateforme décisionnelle envoie tous les matins un e-mail aux manageurs des indicateurs de la veille, le chiffre d'affaire, la marge, la liste des ventes les plus importantes, le classement des meilleurs collaborateurs, etc. Ces rapports peuvent être poussés sur différents canaux comme le mail, le SMS, le MMS, ou bien sur la page d'accueil du portail d'entreprise. Ces rapports peuvent être de simples indicateurs, mais généralement on les présente dans des graphiques interactifs si le canal le permet.

[1] Ad hoc : locution latine dont le sens est : « qui va vers ce vers quoi il doit aller », c'est-à-dire : « formé dans un but précis. Une analyse ad-hoc est une analyse ponctuelle sur demande.

Il existe plusieurs possibilité de rapports dits « tirés », en mode pull. Cela va du rapport déposé dans une espace partagé sécurisé, et donc consultables voire téléchargeable par qui en a les droits. Attention à la confusion ; dans certaines entreprises, on parle de rapport en mode pull pour les rapports Ad-hoc.

Dans le cadre de cette introduction, il est important de dire deux mots sur le portail décisionnel. C'est généralement le point d'accès de tout le monde à des rapports en mode push ou pull, à des applications de Reporting Ad-hoc etc.

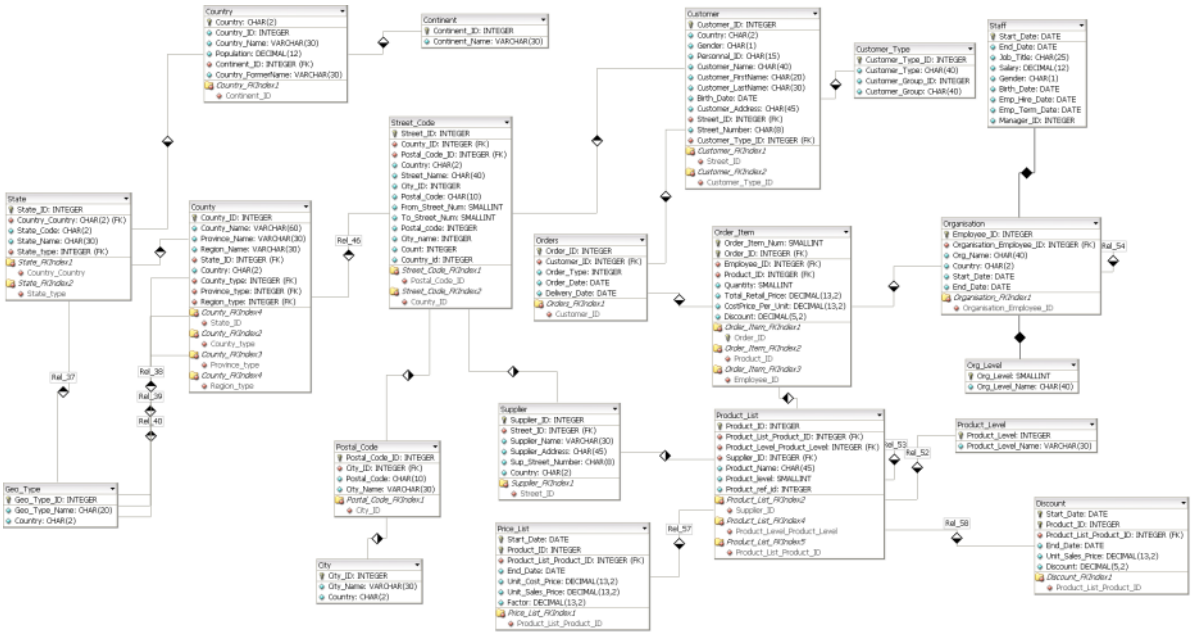
Petit scénario pour résumer : Le directeur marketing reçoit un SMS (Reporting en mode push) où il s'aperçoit que le chiffre d'affaire de la veille n'est pas si important qu'il l'espérait. Il souhaite donc plus de détail sur la marge générée par telle campagne en comparaison de telle autre campagne, et demande à son assistante un rapport qui sera créé en trois clics dans SAS Web Report Studio (Reporting de masse en mode pull). Après l'analyse des graphiques qu'il a reçu dans l'application de son Smartphone, il demande à l'analyste métier de son département une petite étude statistique pour savoir quels sont les facteurs discriminants sur cette campagne. L'analyse réalisée avec SAS Enterprise Guide (Reporting Ad-hoc avancé (« avancé » par opposition à « de masse »)) est transformée en procédure stockée pour être réutilisable de manière simple.

Aussi bien construit que soit le *Data Warehouse*, il est nécessaire que des non informaticiens puissent accéder à l'information qu'il recèle de façon simple, grâce à un outil de Reporting intuitif.
De plus, aussi performantes soient les analyses, si leurs résultats ne sont pas présentés de manière compréhensible par ceux concernés, cela ne sert pas à grand-chose.
La dernière phase du processus décisionnel est donc de transformer l'information contenue dans le Data Warehouse en connaissance exploitable et de la diffuser à qui de droit.

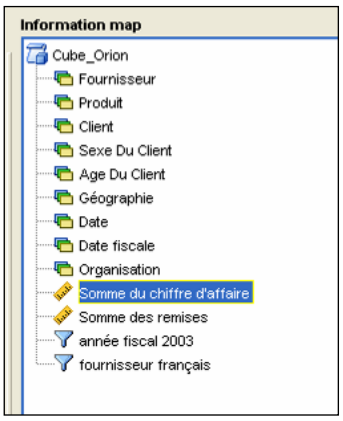
Les Map

Dans la plupart des entreprises, beaucoup d'utilisateurs sont à la recherche d'informations. En général, les applications informatiques apportent les informations nécessaires, tout au moins celles qui ont été demandées et acceptées lors de la conception de l'application qui fournit ces données. Mais très souvent, au cours de la vie de l'application, les besoins des utilisateurs, évoluent, s'affinent ou s'adaptent : bref ils changent parfois dans des proportions importantes. Ces besoins ne sont parfois utiles qu'une seul fois.
Afin de répondre à ces besoins, on peut faire appel aux services informatiques, mais les délais de réponse peuvent être rédhibitoires et l'application arriverait trop tard pour être utile.
Une solution pour éviter ce problème est de donner aux utilisateurs la possibilité de créer eux-mêmes leurs rapports, laissant ainsi à l'informatique le temps de se charger des missions critiques, des infrastructures techniques et fonctionnelles du système d'information.
Néanmoins, les utilisateurs qui non pas les compétences pour comprendre et interpréter la structure complexe des données stockées dans le *Data Warehouse,* pourront intuitivement naviguer dans l'information et créer leur rapport. Le *Data Warehouse* a une structure essentiellement informatique, peu parlante pour les fonctionnels. Afin de pallier ces manques, nous allons donc créer des Map, appelées aussi dans la littérature académique : « concept métier » ou « vues métiers » ; ou bien pour l'éditeur SAS : « *Information Map* », ou bien encore, pour l'éditeur Business Object : « Univers ».

Une Map n'est pas une structure de données en soit, c'est un ensemble de métadonnées décrivant les données du Data Warehouse en terme métier. Une Map permet au consommateur d'information de créer son rapport, elle lui présente la structure du Data Warehouse selon sa vision.



Ce type de présentation de l'information est incompréhensible pour l'utilisateur fonctionnel.



Nous allons donc lui créer une Map dans laquelle les mesures et les catégories ont été définies selon l'environnement de l'utilisateur.

Une *Map* est donc un ensemble de métadonnées comprenant :

- L'endroit où les données sont stockées physiquement.
- Les relations entre les tables.
- Les transformations nécessaires pour convertir l'information selon la vision de l'utilisateur, ou règles métier.

Une *Map* permet donc de générer des requêtes simplement, elle contient toutes les métadonnées permettant la génération automatique de la requête, de façon transparente pour l'utilisateur.

***Deux sortes de** Map*

Il existe deux sortes de *Map* :

- Les *Map* relationnelles, qui génèrent des requêtes SQL interrogeant une ou plusieurs tables relationnelles.
- Les *Map* multidimensionnelles, qui génèrent des requêtes MDX interrogeant un cube multidimensionnel.

Transformation

Deux utilisateurs peuvent avoir une vision différente d'une même information. Prenons quelques cas courant :

La langue : Quelle que soit la langue de stockage de l'information dans le *Data Warehouse*, l'utilisateur final souhaite obtenir l'information dans sa langue. Par exemple, il suffit donc de déclarer que la colonne « customer_age » est « *customer age* » pour les anglophones et « âge du client » pour les francophones.

La date : hormis le fait que la date est stockée dans SAS comme le nombre de jour depuis le 01 janvier 1960, alors qu'elle s'affiche suivant différent formats, 12 mars 2006 ou 12/03/06 par exemple, il peut aussi y avoir quelques transformations spécifiques simples. On peut avoir pour certains départements besoin de la date calendaire, et pour d'autres, de la date fiscale. Il n'y a entre ces deux types de date qu'un simple décalage de quelque mois.

La devise : Pour une société internationale, quelle que soit la devise utilisée pour le stockage, le français souhaitera cette information en euro, le britannique en livre sterling et l'américain en dollar. Un processus ETL peut donc charger tous les jours la table de conversion. La *Map* gèrera alors le fait que le montant, est le montant stocké que multiplie le taux de conversion, affiché avec le symbole adéquat (€, £ ou $).

Une *Map* permet donc de combler le fossé entre la structure technique du *Data Warehouse* et la conception pratique de l'information, de l'utilisateur final.

Une *Map* sera la structure de l'information vue par un utilisateur dans *Web Report Studio* pour la création de rapport et dans le portail pour naviguer dans cette information.

Qui utilise SAS Information Map Studio ?

La personne créant des *Map* doit parfaitement comprendre la structure du *Data Warehouse* et le besoin d'information des utilisateurs, leur univers. C'est donc la personne qui permettra de combler la différence technique entre le *Data Warehouse* et les consommateurs d'informations. Il doit donc avoir des compétences avancées en requête SQL et /ou MDX.

Remarque : Dans certaines entreprises, il y a un *Data Warehouse* et chaque *Map* étant une transformation métier, est appelée *Data Mart*.

Map versus vues
La définition d'une *Map* donnée au début de ce chapitre peut porter à confusion, il est important de bien dissocier une vue et une *Map*.
Une *Map* :

- Décrit le type de rapport qui peut être produit.
- Peut porter sur plusieurs bases de données indépendantes et différentes.
- Cache la structure réelle de la donnée si bien que l'utilisateur du système n'a pas besoin de connaître le SQL pour l'interroger.
- Gérer des catégories et des mesures qui peuvent être classées en dossier.
- Peut contenir des filtres et des indicateurs pré-calculés.
- Peut être étendue à une procédure stockée.

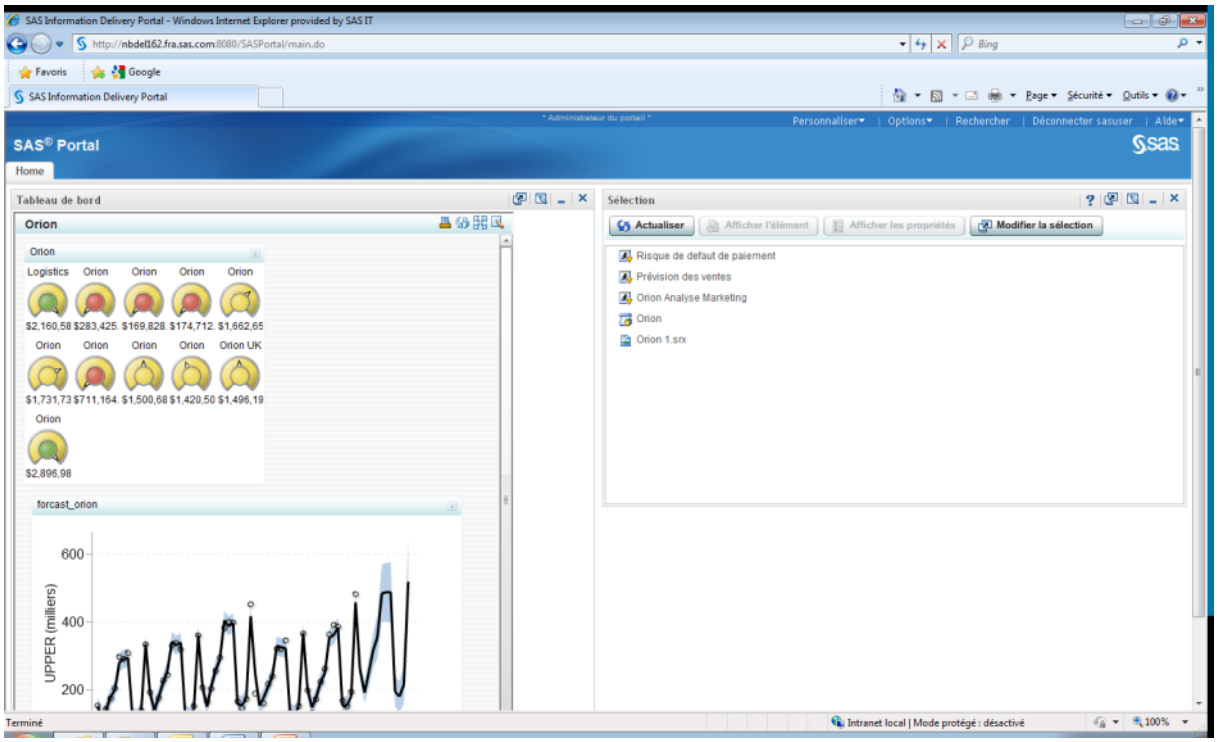
Le Reporting de masse et le Reporting Ad-hoc

Un système décisionnel doit permettre de répondre rapidement aux questions permettant d'extraire de l'ensemble de l'entrepôt de données, la connaissance utile à la prise de décision. La création de Map offre aux utilisateurs une grande latitude pour naviguer dans l'information. De plus, la génération de rapports généraux peut être industrialisée, de telles sortes que tous les ayants droits, reçoivent leur

rapport. On parle de Reporting de masse, on diffuse massivement la connaissance suivant les règles définies pour la construction du Data Warehouse.
SAS Web Report Studio (WRS) est typiquement un outil de Reporting de masse qui permet à des non-initiés, de faire en quelques clics le rapport dont ils ont besoin.
Néanmoins, il y a toujours des besoins de rapports à la demande : « Ad hoc », rapports non prévus lors de la mise en place de la plate-forme décisionnelle.
SAS Enterprise Guide est un outil de Reporting avancé qui va permettre à un utilisateur plus averti de faire des rapports plus complexes pouvant inclure des analyses statistiques avancées et d'industrialiser le processus

Le portail décisionnel et mobile

Le portail est un terme générique pour désigner un site Web qui sert de point d'entrée sur des rapports, des tableaux de bord et des applications décisionnelles. Lorsqu'un utilisateur se connecte, il obtient la plupart de la connaissance dont il a besoin pour prendre ses décisions et agir.
De plus en plus, ces informations sont mises à disposition sur Smartphone, avec des tableaux de bord facilement exploitables sur des écrans de taille réduite. Depuis son Smartphone Android, son iPhone ou son BlackBerry, l'utilisateur peut en fonction de ses droits, consulter, annoter, modifier ou bien créer son rapport ou son tableau de bord.



Fréquence de mise à jour

Le problème de la fréquence de mise à jour est relativement complexe et donne souvent lieu à controverse. Nous allons donc commencer par définir quelques concepts de base.
La fréquence de mise à jour s'applique à différents stades de la chaine du décisionnel :

- mise à jour du DDS. Il peut arriver qu'une partie du DDS soit mis à jour en temps réel. Lors du 'comit' d'une transaction, le DDS est mise à jour. Mais ces cas sont très complexes à mettre en place en place et à gérer, donc très couteux ; pour ne pas dire trop coûteux. La mise à jour en pseudo temps réels, appelée 'just in time' ou 'juste à temps' est de plus en plus fréquente avec des bases mises à jour toutes les demi-heures, toutes les heures, ou par demi-journée par exemple. Attention sur les expressions utilisées ici, elles varient fortement d'un acteur à un autre. On utilise notamment pour cela les CDC (Change Data Capture) qui permettent de regarder et de charger uniquement les nouveaux éléments.
- La mise à jour des tables des schémas en étoile est plus généralement faites toutes les nuits où toutes les semaines, par exemple le dimanche matin.
- La mise à jour des cubes est souvent faite toutes les semaines, voire tous les mois.
- Certaines tables et certains rapports sont mis à jour à la volée. Par exemple, lorsque vous créez, ouvrez un rapport ; une procédure stockée met à jour les données.

En fonction du besoin des utilisateurs ces bases peuvent avoir des fréquences de mise à jour variant du pseudo temps réel à l'année.
De même, certain rapport sont rafraichis à chaque ouverture, d'autre tous les jours, ou toutes les semaines, et ainsi de suite.

Ce n'est que lorsque le projet décisionnelle devient suffisamment complexe que l'on s'aperçoit de la nécessité d'un Ordonnanceur (scheduling en anglais), voir le chapitre ETL.

Un agrégat est une valeur synthétique obtenue par la combinaison ou la somme d'informations de détail. Exemple : la somme du chiffre d'affaire pour une société.
Tout le sel de l'information décisionnelle est de pré-calculer suffisamment d'agrégats mais pas de trop. Voir chapitre Optimisation d'un cube OLAP.
Ces agrégats doivent eux aussi être recalculer à chaque mise à jour.

Structure logique	Normalisé	Dé-normalisé	OLAP	Fichier Plat	Virtuel
Elément du Data Warehouse	DDS (ou ODS)	Schéma en étoile Schéma en constellation	Cube Multi dimensionnel	Table de Data Mining	Information Map directement sur le système opérationnel
Utilisation	Pour avoir un Enterprise Data Warehouse, intégré et historisés toutes les données du groupe, de l'entreprise	Pour des rapports assez fréquent,			A ne jamais utiliser, sauf dans de rares cas confirmant la règle pour accéder à une information simple en temps réel.
Utilisateur	Informaticien averti	Analyste métier	Analyste métier, fonctionnel	Analyste, statisticien, Data Miner	Temps réel.
Fréquence de mise à jour	Pseudo temps réel, toutes les demi-heures, demi- journée, voir tous les jours	Tous les jours, toutes les semaines, voir tous les mois	Tous les jours, toutes les semaines, voir tous les mois	Cela dépend du modèle	

En fonction des Data Warehouse, une ou plusieurs formes logiques de base de données sont utilisées.

Système décisionnel versus système opérationnel

Tout d'abord, avant de rentrer dans le détail des systèmes décisionnels, positionnons-les par rapport aux systèmes opérationnels.
Un système opérationnel est un système informatique permettant de gérer l'information au quotidien. Toutes les grandes entreprises ont mis en place de tel système. Parmi ces systèmes notons les ERP, Enterprise Ressources Planning, ou en français PGI : Progiciel de Gestion Intégré comme par exemple l'éditeur SAP®. De tels systèmes permettent notamment de gérer des flux d'information à travers l'entreprise, de gérer la comptabilité, de générer les factures, les feuilles de paye, la logistique, etc.

Par exemple, si un commercial vend une voiture, un flux d'information va être transmis à la production qui devra assembler cette voiture, au service comptable qui va générer la facture, aux approvisionnements qui devront prendre en compte la demande dans leur commandes, etc. De tels systèmes permettent l'industrialisation de la communication d'informations opérationnelles.
Il est donc normalement possible à tout moment de savoir combien il y a de voiture dans le portefeuille de commande, où en est la commande de ce client, est-ce que les stocks sont suffisants, etc. Un système opérationnel permet l'efficacité au jour le jour.
Par contre, si l'on souhaite faire des rapports, des tableaux de bord présentant une information transversale, si l'on désire faire des prévisions et des simulations, comme par exemple savoir quel client est susceptible de commander un nouveau produit, ou quel autre client pourrait partir à la concurrence, alors ces systèmes opérationnels peuvent difficilement, et souvent pas, répondre à ces problématiques ; c'est alors qu'interviennent les systèmes décisionnels.
Un système décisionnel est un système informatique complémentaire des systèmes opérationnels auxquels il devra s'intégrer.

La mise en place d'un système opérationnel est un projet structurant, on impose aux utilisateurs de nouvelles méthodes de travail : par exemple, avant le commercial faisait un bon de commande sur papier, maintenant il a une interface où il doit obligatoirement remplir un certain nombre de champ. Un des concepts de base des ERP est de définir l'ensemble des processus qui permettent de faire fonctionner l'entreprise, de les normaliser et de les appliquer.

Un système décisionnel doit s'adapter aux méthodes de travail des utilisateurs, par exemple, ils faisaient leurs rapports avec Microsoft® Office Excel, sans changer d'environnement, ils auront un accès à l'entrepôt de données décisionnelles et à des fonctions analytiques avancées.
Si un système opérationnel est piloté par la rationalisation des processus, un système décisionnel est surtout piloté par les métiers[1], la structure du système est gouvernée par leurs besoins.
Un utilisateur peut potentiellement ne pas utiliser les outils décisionnels mais est pour ainsi dire obligé d'utiliser une application opérationnelle. Le corolaire de ce constat est qu'il est parfois nécessaire de prouver la valeur ajoutée, le gain de performance, d'un tel système. Je dirais même volontiers que ce devrait être quasi obligatoire.

Techniquement, un système opérationnel est conçu pour gérer beaucoup de petites transactions par seconde, alors qu'un système décisionnel va gérer peu de transactions mais beaucoup de requêtes complexes.

Selon mon point de vue, la principale différence entre la mise en place d'un système opérationnel et celle d'un décisionnel réside dans le fait que pour un système opérationnel, on commence par l'adapter à l'organisation, puis toute cette organisation doit utiliser ce système alors que pour un projet décisionnel, il faut beaucoup plus s'adapte à la stratégie de l'entreprise, à sa philosophie et aux utilisateurs.

Les organisations ont pour la plupart en place leur système opérationnel pour être efficaces et réactives, elles doivent maintenant pour être proactives, déployer un système décisionnel.

[1] Métier : les métiers génériques de l'entreprise sont la production, les ressources humaines, le marketing, les forces de ventes. Les utilisateurs métier d'un système décisionnel sont donc des utilisateurs généralement non informaticiens, ayant besoin d'accéder de manière simple à la connaissance.

Un progiciel opérationnel est généralement sans analyses pertinentes.

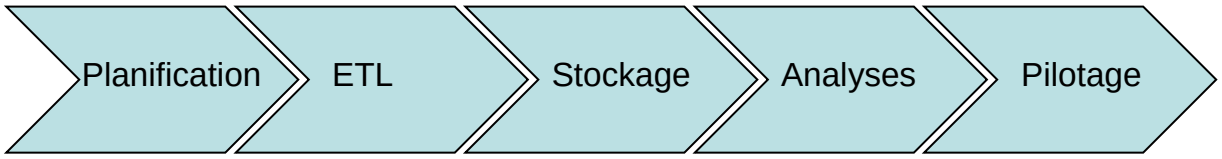
Récapitulatif :

Système opérationnel	**Système décisionnel**
Efficacité	Avantage compétitif
Jour le jour	Stratégie à long terme
Beaucoup de transactions par seconde	Beaucoup de requêtes complexes par seconde
Application statique	Application dynamique
Base de données changeantes	Historique évolutif
Routine	Créativité
Réactivité	Pro-activité
Piloté par les processus	Piloté par les métiers

Le projet décisionnel

Un processus décisionnel peut se décomposer en cinq phases principales :

1. la planification : Méthodologie de mise en place, modélisation du système
2. l'intégration : Intégration aux systèmes opérationnels - extraction - validation de la qualité des données - transformation – chargement des données dans l'entrepôt décisionnel
3. le stockage : stocker l'information à but décisionnel sur le meilleur support
4. l'analyse : extraie de la donnée, le maximum de connaissance – *Data Mining* - optimisation
5. et la diffusion de la connaissance : permettre à des non informaticiens d'accéder simplement à la connaissance.



Nous allons donc maintenant présenter ces cinq étapes.

La planification

Avant de construire un système décisionnel à structure évolutive, il est important d'avoir une vision d'un potentiel ensemble final. Il faut savoir où l'on va et par quel chemin. On peut résumer la méthodologie d'élaboration, de construction, de déploiement et d'élargissement, d'un système décisionnelle, par la phrase « voir grand, et commencer petit ». En effet, il est important de concevoir un système décisionnel qui puisse grandir tout au long de sa vie, donc de prévoir dès le début de sa conception, sa capacité à s'étendre à de nouveau domaine, à intégrer de plus en plus d'information, pour de plus en plus d'utilisateurs. Dans le choix des technologies, cette propension de croissance doit être prise en compte afin de ne pas se retrouver au bout de quelques temps dans une impasse technologique. « Voir grand », c'est prévoir un système qui sera capable à terme, par exemple,

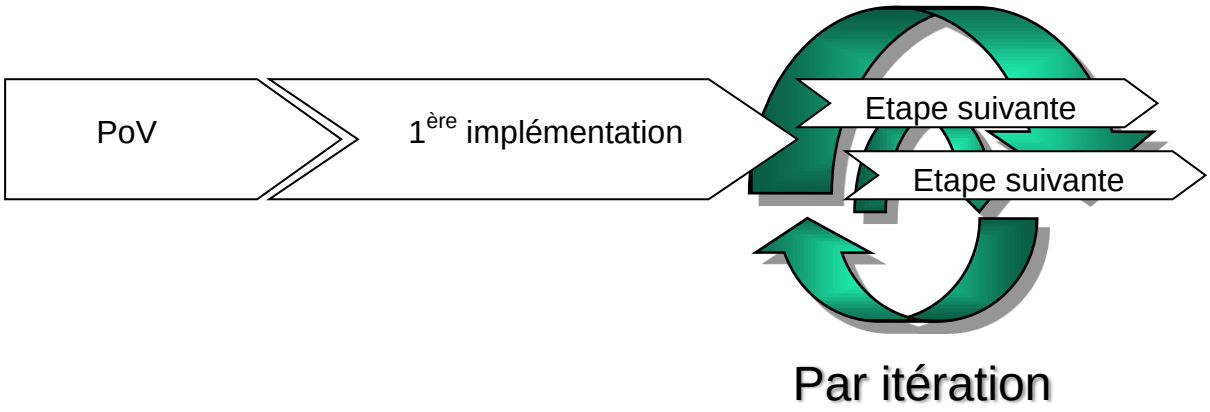
- de prévoir quel client va acheter quel produit,
- de prévoir quel client est susceptible de partir à la concurrence,
- d'avoir une vision globale et multi transversale de l'entreprise, afin d'en optimiser la structure et le comportement,
- d'intégrer de nouvelles informations pour répondre à de nouvelles problématiques
- de s'élargir à de nouveaux utilisateurs
- etc.

Commencer petit, c'est commencer par mettre en place un système qui réponde concrètement à quelques problématiques importantes. La réussite de cette première étape doit être un socle solide permettant de bâtir les étapes suivantes. Pour lancer un projet décisionnel, avant de partir dans la construction d'édifice pharaonique, il faut rapidement en prouver la faisabilité, preuve, loin s'en faut, non évidente.

La méthodologie de mise en place d'un système décisionnelle est donc plutôt une méthodologie itérative incrémentale. C'est-à-dire que l'on va faire plusieurs étapes successives incrémentales.

On peut aborder un projet décisionnel en trois grandes étapes :

1. Le PoV : Proof of Value, prouver la valeur ajoutée de mise en place d'un tel système
2. La première implémentation : Mise en place d'un système sur un périmètre restreint
3. Les étapes suivantes : élargissement itératif du système décisionnel



Le PoV

Généralement, on commence par prouver la valeur de la mise en place d'un système décisionnel par un PoV : *Proof of Value*, un projet au périmètre réduit permettant de commencer par réellement mesurer la valeur du projet. Un PoV est une forme plus aboutie du PoC : *Proof of Concept*, prononcé dans le jargon francophone « poque ». Différents éditeurs de logiciel, intégrateur ou SSII, proposent des PoC voire des PoV. Un PoV se base en fait sur une grosse maquette, utilisant les données de l'entreprise, ce qui va permettre de présenter aux futurs utilisateurs ce à quoi va ressembler l'application et donc d'argumenter sur son concept et sa valeur ajoutée. Un PoV s'accompagne d'une méthodologie et d'indicateur permettant de mesurer a priori, le potentiel de création de valeur d'un système décisionnel.
Si le cahier des charges du futur système décisionnel n'est pas encore fini, voire pas commencé, ce qui arrive souvent ; la réalisation de cette maquette est souvent très utile afin de mieux communiquer entre maîtrise d'œuvre et maîtrise d'ouvrage.
Il n'est pas par définition, strictement nécessaire de passer par cette étape, mais elle est très largement répandue car très utile.

Maîtrise d'œuvre – maîtrise d'ouvrage

Pour les néophytes des projets, la maîtrise d'œuvre est l'équipe qui œuvre, qui travaille, qui développe quelque chose pour la maîtrise d'ouvrage qui le lui a commandé. Pour retenir ces dénominations, quelques rappels linguistiques peuvent être utiles. Il faut se rappeler qu'un chef d'œuvre est la plus belle œuvre d'un artiste, le résultat de son travail alors qu'un ouvrage est simplement le résultat que l'on a devant soi. Œuvrer, c'est travailler ; un ouvrage d'art, c'est quelque chose qui est le résultat d'un art, d'un savoir-faire.
Dans un projet informatique, on retrouve toujours ceux qui commandent une application et qui devront la payer, et ceux qui la développe et la délivre.
Prenons un exemple : Dans le cas de la mise en place d'un système d'information décisionnelle pour la gestion de la relation avec le client, la direction marketing demande à la direction informatique une application répondant à différents critères, cette direction marketing est la maîtrise d'ouvrage. Les collaborateurs de cette direction maîtrisent la gestion de la relation avec le client, mais peu, voire pas du tout, l'informatique. Pour répondre à cette demande de développement, la maîtrise d'œuvre va œuvrer pour comprendre le besoin et développer l'application qui y réponde. Je suis persuadé que l'enjeu majeur de la mise en place d'une plate-forme décisionnelle n'est pas tant technologique qu'humain. Le sujet est suffisamment complexe et stratégique pour que des tensions se créent, et aussi performante soit la plate-forme mise en place, s'il y a des résistances aux changements trop grandes, si bien que si personne n'utilise le décisionnel, l'ensemble des moyens utilisés ne sert à rien.

Il est absolument primordial que l'ensemble des protagonistes prennent le temps de comprendre l'enjeu du décisionnel, les réponses que cela peut apporter et ce qui techniquement est réalisable.

La première implémentation

Suite à la validation de la valeur du projet, de son concept, la phase suivante est fréquemment la mise en place d'une solution sur un périmètre restreint, solution répondant aux problématiques prioritaires de la société. Cette seconde étape est souvent appelé « quickstart », littéralement démarrage rapide. Notons le mot rapide, il ne faut certainement pas oublier la notion du temps. Des dates échelons doivent être mise en place pour chaque étape d'un projet. Il faut rapidement mettre en place une application concrète, en effet trop de projets meurent dans de la réunionnite incapable de trouver un périmètre de démarrage.
Viennent donc après le « PoV », des conceptions, installations, développements, implémentations et déploiement, plus larges, sur des problématiques plus importantes.

Dans le cadre d'entreprises de taille moyennes, la mise en place d'un système décisionnel doit se faire très rapidement. L'ensemble du projet doit souvent être réalisé en moins d'un mois, deux maximum.

La première implémentation est similaire aux suivantes, sauf qu'elle est la première. On n'a généralement qu'une seule chance de faire une première bonne impression. Une attention particulière doit donc être portée sur ce socle, pierre angulaire du futur système décisionnel.

Pour cette première implémentation, il est particulièrement utile de trouver les « pépites », ces questions métiers dont l'implémentation de la réponse n'est pas très compliquée et qui rapporte beaucoup. Vos utilisateurs ont généralement une myriade de questions, il faut sélectionner les deux ou trois dont le retour sur investissement semble le plus important. Ce sera fort utile pour démontrer la valeur ajouté d'un projet décisionnel par la suite.

Dans la cadre d'un grand projet, projet, on pourra développer l'embryon du DDS avec les données nécessaires à la création du premier Data Mart. Dans les étapes suivantes, selon cette méthodologie itérative incrémentale, le DDS sera consolidé et les nouveaux Data Mart complétés.

Les implémentations suivantes

La construction d'un système décisionnel ne doit n'y être monolithique, ni fragmenté.
Une construction monolithique serait de vouloir mettre en place le système décisionnel en une seule fois. Il est très prétentieux de vouloir implémenter d'un seul coup, un système qui réponde à l'ensemble des problématiques d'une société !
Une construction fragmentée est de construire plusieurs systèmes indépendants qui répondent chacun à des problématiques différentes. Cela va l'encontre même d'un système décisionnel, qui est de fédérer l'information afin d'avoir une vision synthétique, globale, et transversale de l'entreprise.

Ce qu'il faut retenir : la mise en place et l'élargissement d'un système décisionnel d'entreprise se fait par itération, chaque itération dépendant de la précédente. On définit un système avec une architecture évolutive vers la résolution d'un ensemble de problématiques très larges, et l'on démarre par la construction rapide d'un projet répondant à quelques besoins les plus importants.

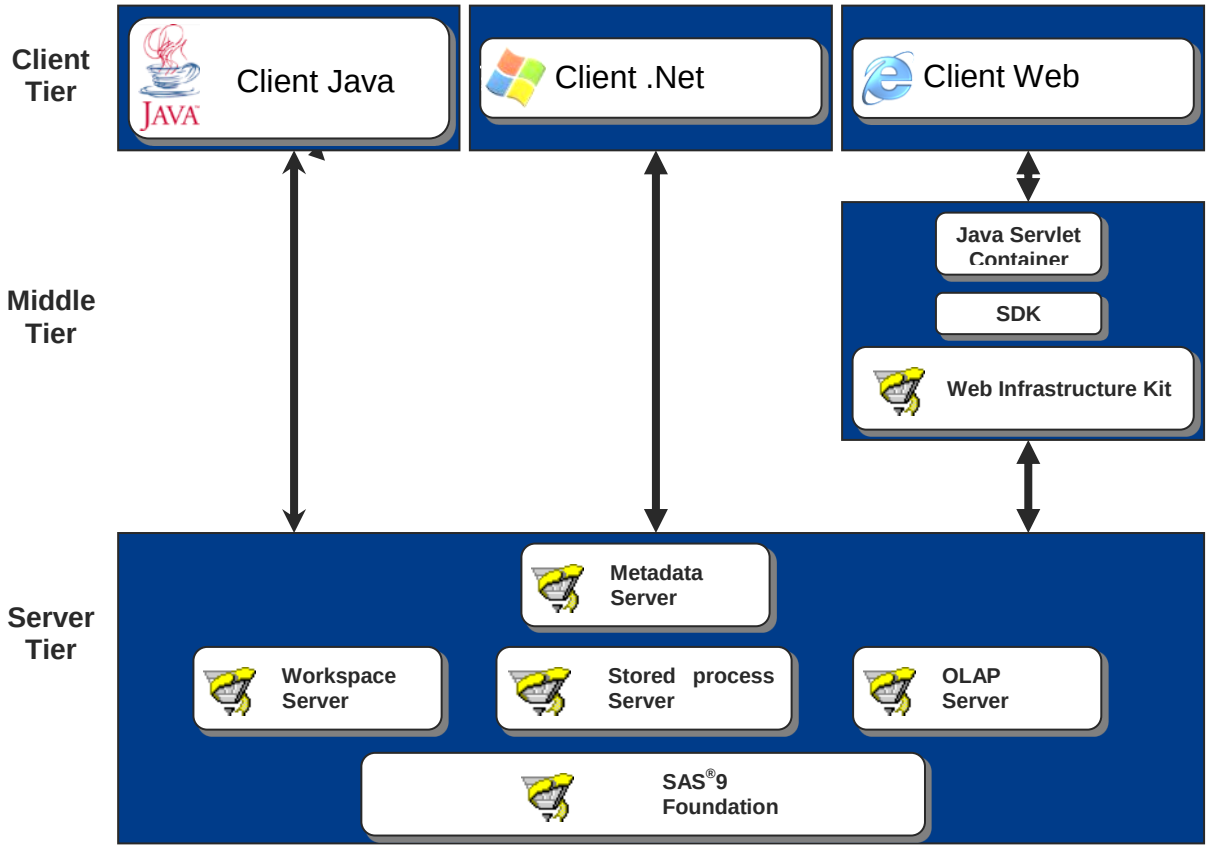
Voici un exemple de méthodologie pour chaque itération, la première et les suivantes :

1. Faisabilité
    a. Identification des compétences des intervenants aux techniques et enjeux du projet
    b. Définition des fonctions et du rôle de chaque intervenant
2. Analyse du Besoin
    a. Définition des besoins
    b. Définition des fonctionnalités clés attendues du système
    c. Analyse, simulation des impacts sur l'organisation
3. Conception
    a. Analyse et conception de l'architecture de l'application fonctionnelle et technique du système.
    b. Création du cahier des charges

4. Construction
    a. Construction et développement des différents composants
5. Validation
    a. Validation que le développement fait répond au cahier des charges
6. Déploiement
    a. Expérimentation pilote
    b. Transfert de compétences vers les utilisateurs
7. Bilan

Bilan du processus de développement, analyse des retours d'expérience positifs et négatifs du projet, analyse de l'utilisation réelle.

Architecture de la plateforme décisionnelle SAS



La plateforme décisionnelle de l'éditeur SAS se décompose en trois tiers principaux :

- Le tiers client : partie sur le poste de l'utilisateur, tiers qui peut être de deux types :
  1. client lourd : il faut installer quelque chose sur le poste de l'utilisateur. Tous les clients de la plateforme décisionnelle sont développés en Java, à quelques exceptions près, notamment SAS Enterprise Guide et SAS Add-in to Microsoft Office. Ces deux composants sont développés avec la technologie .NET de Microsoft. Hormis ces deux produits, les clients lourds de SAS comme la SAS Management Console, SAS Data Integration Studio (ETL), SAS Information Map ou SAS Enterprise Miner, nécessitent un Java Runtime Environment (JRE) ; c'est-à-dire qu'il faut installer un JRE spécifique au système d'exploitation, avant d'installer le client. En d'autres termes, si l'on est sur une machine Windows, il faut installer le JRE de Windows, et les clients que l'on souhaite, pour une machine Linux, il faut le JRE Linux associé et l'on peut installer nos mêmes clients Java. L'intérêt de ce type de technologie est que l'on ne développe qu'une seul fois chaque interface cliente, puis pour les installer dans différents environnements, il suffit d'installer préalablement le bon JRE. Tous ces clients sont extensibles, c'est-à-dire que l'on peut y ajouter des fonctionnalités spécifiques, développer des « Plug-in » particuliers et les déployer sur tous les clients.
  2. Client léger : il ne faut rien installer sur le poste client. Les navigateurs Firefox ou Internet Explorer par exemple sont supposé. Le déploiement de ce type de client est très facile, il n'y a rien à faire, il suffit à l'utilisateur de se connecter à une adresse Web et il a accès à toutes les applications clientes légères.

- Le middle tiers : le tiers intermédiaire est uniquement nécessaire pour les clients légers (interface nécessitant seulement Microsoft Internet Exploreur sur le poste utilisateur). Cette partie comprend toujours les composants :
  1. Un SDK (Software Development Kit ou Servlet Development Kit), en l'occurrence le JDK (Java Development Kit) de Sun. Un SDK est un ensemble d'outils permettant aux développeurs de créer des applications Web. Il est donc possible de développer des applications spécifiques pour tous les clients Web léger de la plateforme SAS.
  2. Un conteneur de servlet Java. Presque toutes les applications légères de la plateforme SAS sont des servlets. Parmi les différentes technologies Java (par exemple : client lourd Java ou Applet Java dans une page Web) les servlets sont des applications que l'on installe uniquement dans un serveur conteneur de servlets Java. Ce conteneur de servlets doit être associé à un JDK, on peut aussi interfacer les servlets à l'aide d'API à des bases de données relationnelles ; ou à des structures hiérarchiques de type répertoire comme par exemple les annuaires LDAP. Le kit de développement AppdevStudio est livré avec les API Java pour piloter de manière complète une plate-forme SAS dans le cadre de servlet (et d'applet). Le conteneur de servlets génère des pages html à la demande, en exécutant la servlet. Sur le marché actuel, on peut citer JBoss (gratuit) ; WebSphere, développé par IBM et WebLogic, développé par BEA (société appartenant à Oracle).
  3. Le WIK (Web Infrastructure Kit) : ensemble des classes Java nécessaires pour connecter la partie intermédiaire avec la partie serveur.

- Le tiers serveur : la partie serveur est l'endroit où est stocké l'entrepôt de données décisionnelles et où s'exécutent les programmes. Ces programmes sont de plusieurs types : processus ETL (Extraction – Transformation – Chargement), programme d'analyse statistique, d'analyse Data Mining, de recherche opérationnelle, de réponse à des requêtes, de génération de rapport, etc.
  - Le serveur de métadonnée est le cœur du système, il permet la gestion de toute la plateforme décisionnelle car tout passe par lui.
  - Le serveur d'espace de travail logique regroupe l'ensemble des serveurs physiques d'espace de travail. Ce serveur logique répartit via un « Object Spawner », les programmes à exécuter, entre les différents serveurs physiques.
  - Le serveur de procédures stockées gère et exécute les procédures stockées. Par exemple, un statisticien crée un programme d'analyse et l'enregistre en procédure stockée. Cette procédure peut être alors exécutée à distance par différents utilisateurs.
  - Le serveur OLAP renvoie la réponse aux requêtes sur les bases de données multidimensionnelles qu'il gère.
  - SAS Foundation : SAS Foundation est l'ensemble des modules SAS, dont SAS/BASE. Dans une architecture où le serveur est sur une machine Windows, SAS/BASE et les modules de SAS Foundation se trouve installés par défaut dans le répertoire C:\Program Files\SASHome\SASFoundation\9.3. Les programmes exécutés par les serveurs d'espace de travail et de procédure stockée sont en fait des programmes SAS, exécutés sur dans une session SAS/BASE.

Explication du processus : prenons l'exemple d'un utilisateur de SAS Web Report Studio, ce client se connecte au serveur intermédiaire en entrant l'adresse Web de SAS Web Report Studio, qui lui demande de s'identifier. Pour établir une connexion, le serveur intermédiaire demande au serveur de métadonnée, le cerveau de la plateforme, si l'utilisateur est connu dans les métadonnées de SAS et si le bon mot de passe est bon.
Une fois connecté, donc reconnu, si notre utilisateur souhaite créer un rapport nécessitant l'utilisation d'informations gérées par différentes bases de données, et nécessitant plusieurs procédures, lors de la conception de ce rapport, seul les métadonnées gérées par le serveur de métadonnées sont manipulées, puis lorsque l'utilisateur demande la génération, un programme SAS (un Job SAS) est créé de tel sort qu'il exploite au maximum la puissance machine disponible.
Le serveur de métadonnées stockant l'information sur les bases de données, les utilisateurs, les machines, les droits associés aux différentes entités, possède toute l'information permettant de créer un programme parallélisé en accès aux bases de données et réparti sur plusieurs machines, le scaling out, répartition parallélisée elle-même sur les processeurs de ces machines, scalling up.
Les procédures sont donc parallélisables par processeur ou CPU (Central Processing Unit, « Unité centrale de traitement »)

Un système décisionnel doit pouvoir monter en charge au fur et à mesure de sa vie. Il est donc primordial d'exploiter au maximum la puissance machine. De plus, l'administration centralisée permet si besoin est, d'ajouter simplement une machine de traitement et une fois référencée dans les métadonnées, cette machine sera utilisée par les algorithmes de répartition de charge.

Qui fait quoi ?

L'un des défis majeurs du décisionnel, est de faire travailler ensemble des collaborateurs aux profils très différents : consommateur d'information, décideur, analyste métier, administrateur, analyste ETL, architecte, etc. Pour chaque profil, une interface adaptée au besoin a été conçu. Dans certain projet, la même personne peut avoir plusieurs fonctions ou rôles.

Utilisateur métier (traduction de Business User). On peut répartir ces utilisateurs selon l'échelle hiérarchique de l'entreprise, depuis le sommet avec des personnes consultant et annotant des rapports, jusqu'en bas, avec des utilisateurs ayant beaucoup de rapports à générer pour leurs supérieurs hiérarchiques. Tous ces utilisateurs non informaticiens ont besoin d'accéder de façon intuitive et rapide à l'information. Ils consultent les rapports depuis un portail Web, sur leurs Smartphones, avec la suite Microsoft Office, voire pour les plus avancés, créent des rapports avec SAS Web Report Studio ou l'Add-in à Microsoft. On les appelle souvent les utilisateurs finaux de la BI dans le sens où ils se trouvent à la fin de la chaine de la BI qui transforme les données en information, voire en connaissance. On parle aussi de consommateurs d'information où bien d'utilisateurs métiers ou fonctionnels. C'est généralement une partie de la maîtrise d'ouvrage.

Architecte. Ayant une longue expérience du sujet, il maîtrise les différentes architectures et méthodologies pouvant être utilisées afin d'arbitrer dans leur choix.

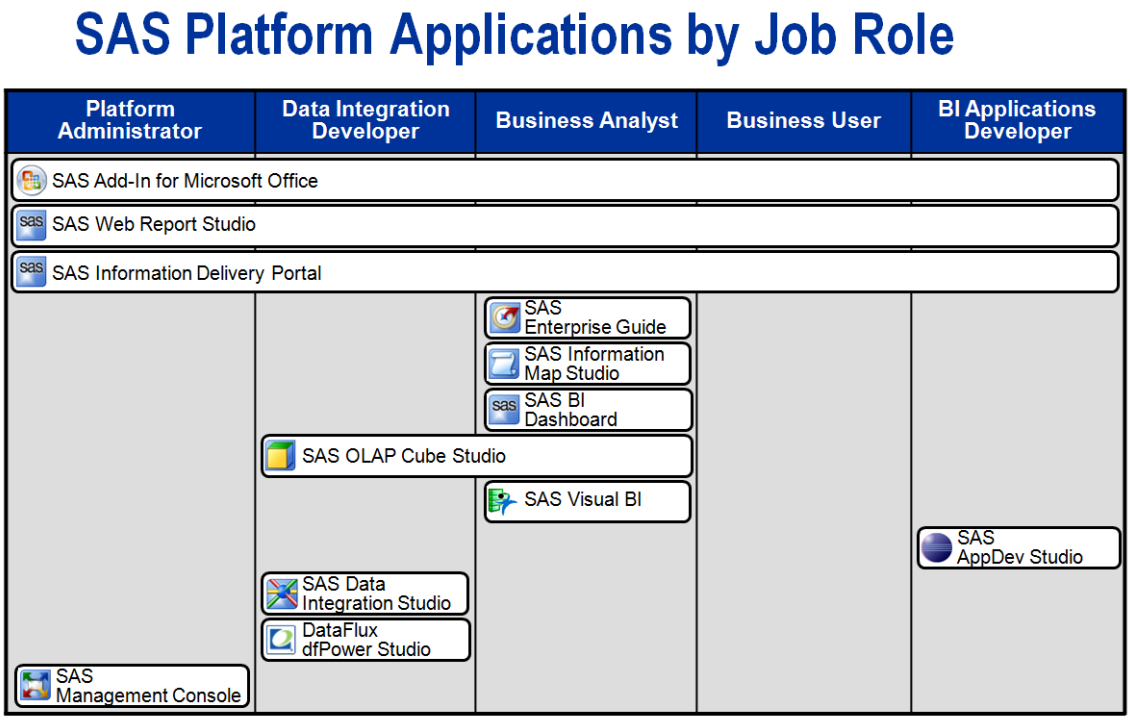
Analyste Métier (traduction de Business Analyst). Ayant des compétences fonctionnelles, métiers et des compétences techniques, l'analyste métier comprenant non seulement la structure des données du Data Warehouse, mais aussi les besoins de l'organisation. Il peut donc aussi bien discuter avec le département informatique qu'avec les utilisateurs finaux. Avec SAS Information Map Studio, il va permettre de présenter un accès simple aux données, même pour des utilisateurs sans aucunes connaissances informatiques. Avec le SAS BI Dashboard, il va présenter les indicateurs de manière percutante. Avec SAS Enterprise Guide, il pourra développer des rapports spécifiques à la demande et les automatiser via les procédures stockées par exemple.

Développeur ETL. Maîtrisant notamment la programmation SAS et le SQL avancé, il construit le Data Warehouse avec SAS Data Integration Studio, valide et normalise la qualité des données avec SAS Data Quality Studio (Data Flux).

Administrateur. Le responsable de l'installation, de la configuration et de la maintenance de la plateforme, il gère les droits avec la SAS Management Console.

Chef de Projet. Connaissant les meilleurs pratiques, il doit assurer l'intégrité et la qualité du projet.

Développeur d'application spécifique. Ayant des compétences avancées en programmation SAS, html et Java, ce développeur en utilisant le produit SAS AppDev Studio, bénéficie de l'environnement de développement Eclipse, pour développer des IHM (Interface Homme Machine) spécifique. La plateforme SAS et conçu pour que l'on puisse étendre les interfaces.



Présentation du cas ‘Orion Star’

Cette étude de cas ‘Orion Star’ partant du système informatique opérationnel pour aboutir à la diffusion de la connaissance est un exemple fictif, propriété de l’éditeur de logiciel SAS, Copyright© 2003 SAS Institute Inc., Cary, NC, USA. Il ne peut être diffusé sans l’accord écrit de SAS Institut.
Nous utiliserons le cas Orion Star comme fil conducteur de cette formation pour voir l’ensemble de l’informatique décisionnelle depuis le système opérationnel jusqu’aux outils de restitutions.

La société : Orion Star

Cette société fictive, présente au niveau mondial, est spécialisée dans la commercialisation d’articles de sport et d’extérieur.
Le siège social, aux États-Unis, gère des filiales en Belgique (depuis 2004), Pays-Bas, Allemagne, Royaume Uni, Danemark, France, Italie, Espagne et Australie.
Les produits sont vendus en magasin, par catalogue et par Internet.
Une carte de fidélité: ‘Orion Star Club’, propose beaucoup d’avantages.
Toutes les informations sont disponibles sur la période allant du premier janvier 2003 au premier janvier 2008. Nous nous considérerons donc virtuellement en date du premier janvier 2008.

Structure de l’organisation :

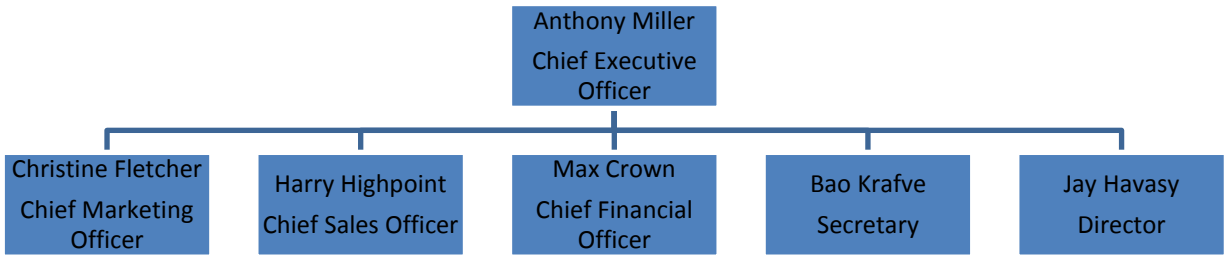
Le siège social qui se trouve aux Etats Unis héberge la majeure partie des fonctions administratives, soit un nombre important d’employés, entre 600 et 800. Le siège social centralise aussi la gestion des stocks, la vente par catalogue, la vente par Internet et l’import – export. Néanmoins, certains employés gèrent aussi ces fonctions depuis les différentes filiales.

Les employés sont enregistrés dans la base de données selon cinq niveaux :

1. Pays
2. Compagnie
3. Département
4. Section
5. Groupe

Les informations complémentaires sur les employés sont notamment :

- Date d’entrée et de départ de l’employé
- Date de début et de fin de contrat (pour certain contrat)
- Adresse
- Sexe
- Salaire
- Responsable hiérarchique



L'offre

La société Orion Star propose environ 5500 références de produits. Certaines ne sont pas vendues dans tous les pays.

Les produits sont organisés selon 4 niveaux:

1. Ligne de produit
2. Catégorie de produit
3. Groupe de produit
4. Produit

Chaque produit a un coût et un prix de vente. Le système informatique gère tous les prix en dollars. En utilisant les dates de début et de fin, ces prix varient en fonction du temps. Cet historique est sauvegardé. Le système gère aussi les remises pour certains produits, à certaines périodes. Les prix sont généralement uniques de par le monde.

Les clients

Les clients d'Orion Star sont repartis à travers le monde, notamment dans les pays où se trouvent des filiales, mais pas uniquement.
Les noms et adresses sont fictifs, même si les villes, régions/comtés et pays, sont réels.
La base de données enregistre environ 90 000 clients, tous ne sont pas actifs.
L'adresse des clients comprend tout ou partie des informations:

- Rue
- Code postal
- Ville
- Région / département / comté
- Etat
- Pays
- Continent

La gestion des adresses est contrôlée par des pointeurs (identifiant de colonnes), ce qui facilite le changement d'adresse.
Les clients sont classés dans des groupes en fonction de leur activité d'achat.

Les commandes

La plupart des commandes de cette étude de cas sont pour des clients détenteurs de la carte de fidélité 'Orion Gold', clients pour lesquels les informations sont enregistrées.
Chaque commande pointe vers le commercial qui a enregistré la vente. Environ 980 000 ligne de commandes sont enregistrées dans cette étude de cas, commandes qui reflètent notamment les saisonnalités.
Chaque commande comprend une ou plusieurs lignes, une ligne par produit.

Les fournisseurs

Chaque produit provient d'un fournisseur qui est basé dans un pays, mais toutes les commandes sont passées par le siège social. Il y a 64 fournisseurs, mais un seul fournisseur par produit.

Les utilisateurs du système décisionnel

Nous allons utiliser différents profils d'utilisateur correspondant à différentes fonctions. Dans certains projets, au moins une personne occupe chaque poste, mais dans bien des cas, une même personne pourra cumuler plusieurs responsabilités. Parmi ces fonctions dont il serait impossible de faire une liste exhaustive tellement chaque entreprise a ses spécificités, notons l'administrateur de la plateforme, le développeur ETL, le gestionnaire des rapports, les utilisateurs métiers, le statisticien, le chef de projet.

La société Orion Star vient d'acquérir une plateforme décisionnelle pour l'amélioration de sa performance. Ce Système devra à terme être utilisé par tous les départements. Il extrait l'information de l'actuel système transactionnel, du système de gestion opérationnel intégré, mais aussi de sources externes.

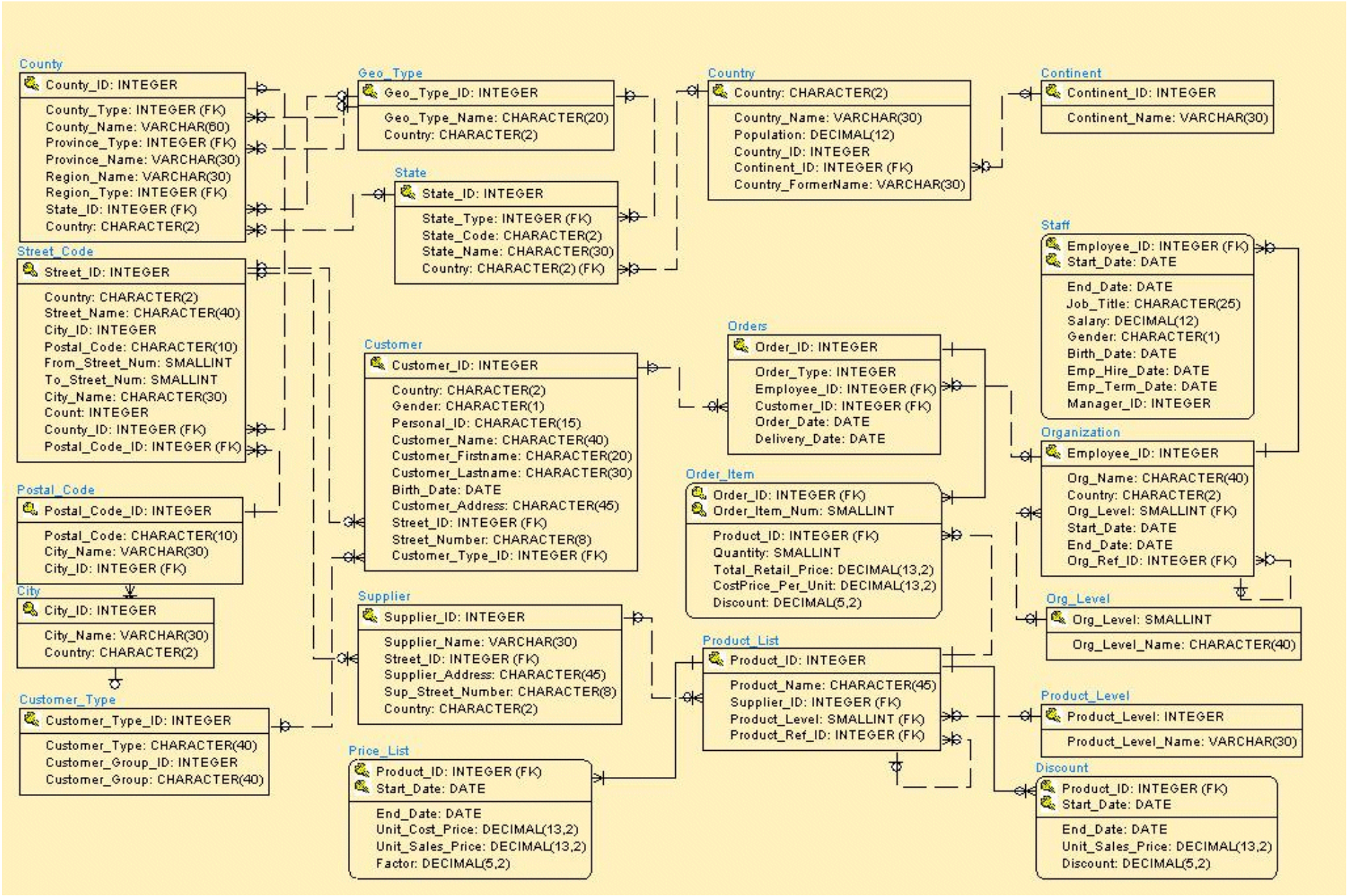
Le département informatique doit donc construire un *Data Warehouse* et des *Data Marts* capables de répondre au besoin de requête, de *Reporting*, d'analyses avancées et de prévisions.

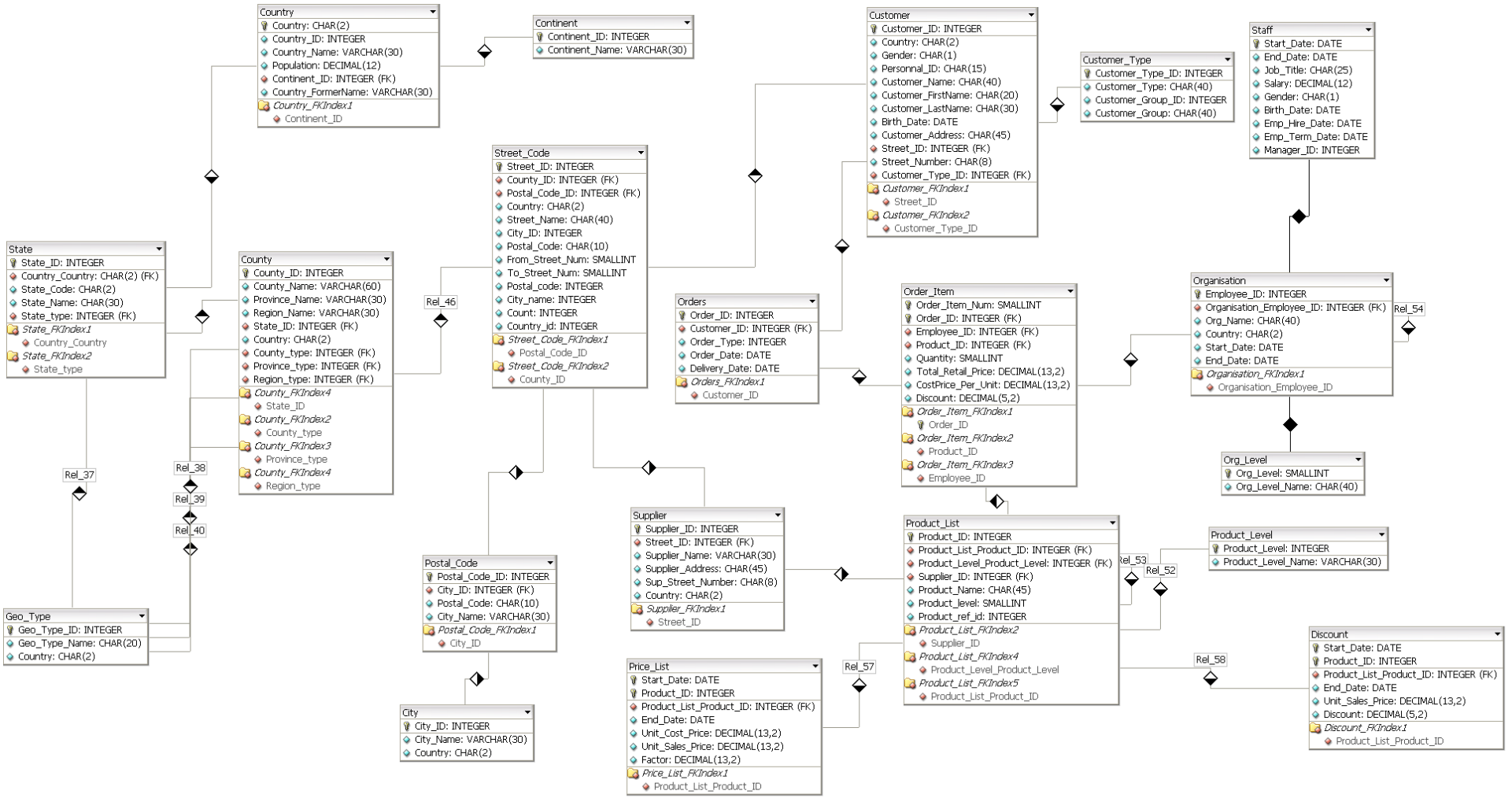
L'équipe projet se constitue notamment de :

Groupe	Numéro Employé	Nom	Fonction	Responsabilité	Pays	Application
Administrateur	120804	Ahmed Zied	Administrateur	Installation, configuration, définition des droits, des bibliothèques, intégration avec les applications existantes	US	SMC
ETL administrateur	120791	Richard Chiselolff	Spécialiste ETL	ETL team Leader, architect – design - DW, manage team members, ordonnancement des tâches	US	SMC ETL
ETL Développeur	120798	Elizabeth Ardskin	ETL Développer	ETL développer, design implémentation processus	US	ETL
Développement d'application	120805	Robert Walker	Report Administrateur	Création, administration et distribution de rapport, de procédures stockées.	US	EG, SMC, IMS, WRS, Portal
Analyste	120727	Donald Marples	Assistant marketing	création et distribution d'analyse avancée et de rapport	US	EG
PDG	120259	Anthony Miller	Chief Executive Officer	Consommateur de rapport	US	Portal WRS
Commerciale	120413	Abderhaman Ferbu	Commerciale	Consommateur de rapport	FR	Portal WRS
Chef de projet		Marcel Dupree	Chef de projet	Chef du projet Orion	US	

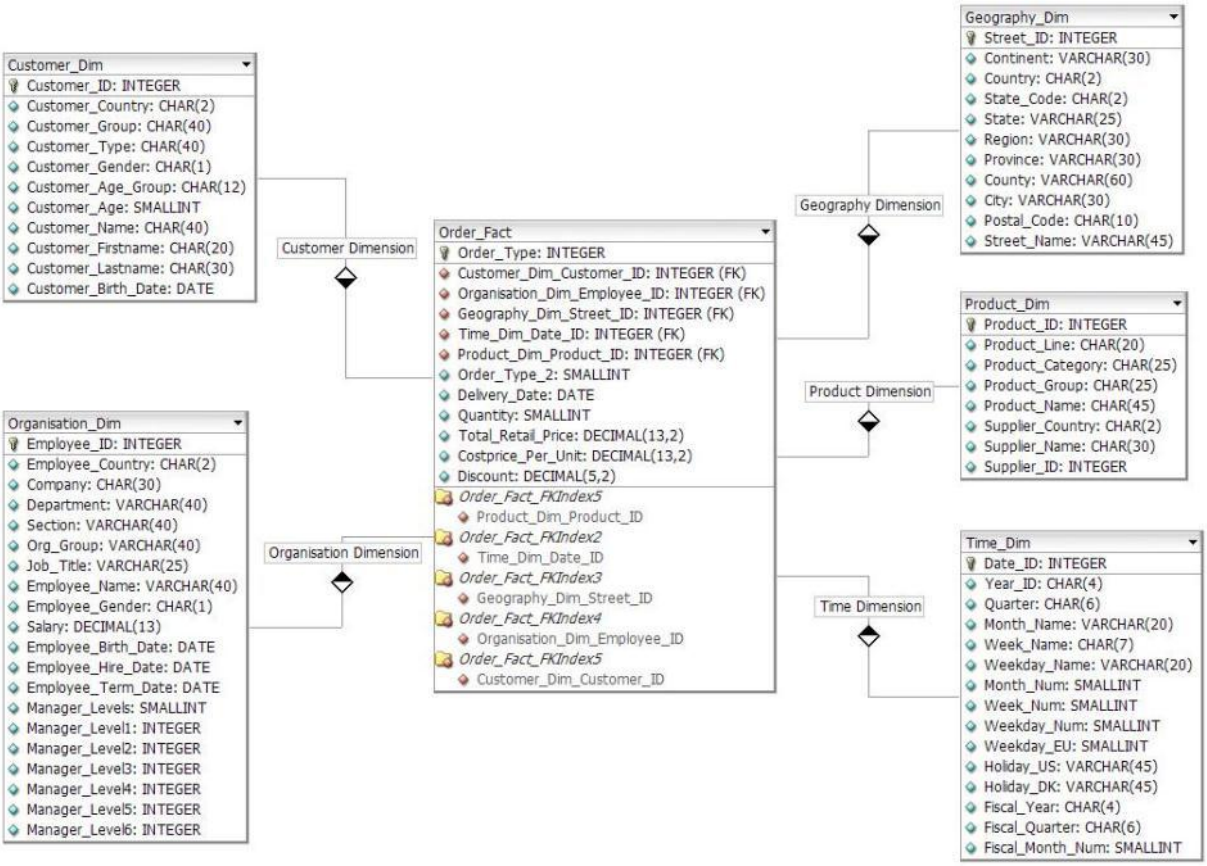
Schéma de la base de données opérationnelle

Les deux schémas ci-dessous sont identiques.


County
County_ID: INTEGER
County_Type: INTEGER (FK)
County_Name: VARCHAR(60)
Province_Type: INTEGER (FK)
Province_Name: VARCHAR(30)
Region_Name: VARCHAR(30)
Region_Type: INTEGER (FK)
State_ID: INTEGER (FK)
Country: CHARACTER(2)
Geo_Type
Geo_Type_ID: INTEGER
Geo_Type_Name: CHARACTER(20)
Country: CHARACTER(2)
State
State_ID: INTEGER
State_Type: INTEGER (FK)
State_Code: CHARACTER(2)
State_Name: CHARACTER(30)
Country: CHARACTER(2) (FK)
Country
Country: CHARACTER(2)
Country_Name: VARCHAR(30)
Population: DECIMAL(12)
Country_ID: INTEGER
Continent_ID: INTEGER (FK)
Country_FormerName: VARCHAR(30)
Continent
Continent_ID: INTEGER
Continent_Name: VARCHAR(30)
Street_Code
Street_ID: INTEGER
Country: CHARACTER(2)
Street_Name: CHARACTER(40)
City_ID: INTEGER
Postal_Code: CHARACTER(10)
From_Street_Num: SMALLINT
To_Street_Num: SMALLINT
City_Name: CHARACTER(30)
Count: INTEGER
County_ID: INTEGER (FK)
Postal_Code_ID: INTEGER (FK)
Customer
Customer_ID: INTEGER
Country: CHARACTER(2)
Gender: CHARACTER(1)
Personal_ID: CHARACTER(15)
Customer_Name: CHARACTER(40)
Customer_Firstname: CHARACTER(20)
Customer_Lastname: CHARACTER(30)
Birth_Date: DATE
Customer_Address: CHARACTER(45)
Street_ID: INTEGER (FK)
Street_Number: CHARACTER(8)
Customer_Type_ID: INTEGER (FK)
Orders
Order_ID: INTEGER
Order_Type: INTEGER
Employee_ID: INTEGER (FK)
Customer_ID: INTEGER (FK)
Order_Date: DATE
Delivery_Date: DATE
Staff
Employee_ID: INTEGER (FK)
Start_Date: DATE
End_Date: DATE
Job_Title: CHARACTER(25)
Salary: DECIMAL(12)
Gender: CHARACTER(1)
Birth_Date: DATE
Emp_Hire_Date: DATE
Emp_Term_Date: DATE
Manager_ID: INTEGER
Organization
Employee_ID: INTEGER
Org_Name: CHARACTER(40)
Country: CHARACTER(2)
Org_Level: SMALLINT (FK)
Start_Date: DATE
End_Date: DATE
Org_Ref_ID: INTEGER (FK)
Org_Level
Org_Level: SMALLINT
Org_Level_Name: CHARACTER(40)
Order_Item
Order_ID: INTEGER (FK)
Order_Item_Num: SMALLINT
Product_ID: INTEGER (FK)
Quantity: SMALLINT
Total_Retail_Price: DECIMAL(13,2)
CostPrice_Per_Unit: DECIMAL(13,2)
Discount: DECIMAL(5,2)
Postal_Code
Postal_Code_ID: INTEGER
Postal_Code: CHARACTER(10)
City_Name: VARCHAR(30)
City_ID: INTEGER (FK)
City
City_ID: INTEGER
City_Name: VARCHAR(30)
Country: CHARACTER(2)
Supplier
Supplier_ID: INTEGER
Supplier_Name: VARCHAR(30)
Street_ID: INTEGER (FK)
Supplier_Address: CHARACTER(45)
Sup_Street_Number: CHARACTER(8)
Country: CHARACTER(2)
Product_List
Product_ID: INTEGER
Product_Name: CHARACTER(45)
Supplier_ID: INTEGER (FK)
Product_Level: SMALLINT (FK)
Product_Ref_ID: INTEGER (FK)
Product_Level
Product_Level: INTEGER
Product_Level_Name: VARCHAR(30)
Customer_Type
Customer_Type_ID: INTEGER
Customer_Type: CHARACTER(40)
Customer_Group_ID: INTEGER
Customer_Group: CHARACTER(40)
Price_List
Product_ID: INTEGER (FK)
Start_Date: DATE
End_Date: DATE
Unit_Cost_Price: DECIMAL(13,2)
Unit_Sales_Price: DECIMAL(13,2)
Factor: DECIMAL(5,2)
Discount
Product_ID: INTEGER (FK)
Start_Date: DATE
End_Date: DATE
Unit_Sales_Price: DECIMAL(13,2)
Discount: DECIMAL(5,2)


Country
Country: CHAR(2)
Country_ID: INTEGER
Country_Name: VARCHAR(30)
Population: DECIMAL(12)
Continent_ID: INTEGER (FK)
Country_FormerName: VARCHAR(30)
Country_FKIndex1
Continent_ID
Continent
Continent_ID: INTEGER
Continent_Name: VARCHAR(30)
Customer
Customer_ID: INTEGER
Country: CHAR(2)
Gender: CHAR(1)
Personnal_ID: CHAR(15)
Customer_Name: CHAR(40)
Customer_FirstName: CHAR(20)
Customer_LastName: CHAR(30)
Birth_Date: DATE
Customer_Address: CHAR(45)
Street_ID: INTEGER (FK)
Street_Number: CHAR(8)
Customer_Type_ID: INTEGER (FK)
Customer_FKIndex1
Street_ID
Customer_FKIndex2
Customer_Type_ID
Customer_Type
Customer_Type_ID: INTEGER
Customer_Type: CHAR(40)
Customer_Group_ID: INTEGER
Customer_Group: CHAR(40)
Staff
Start_Date: DATE
End_Date: DATE
Job_Title: CHAR(25)
Salary: DECIMAL(12)
Gender: CHAR(1)
Birth_Date: DATE
Emp_Hire_Date: DATE
Emp_Term_Date: DATE
Manager_ID: INTEGER
Street_Code
Street_ID: INTEGER
County_ID: INTEGER (FK)
Postal_Code_ID: INTEGER (FK)
Country: CHAR(2)
Street_Name: CHAR(40)
City_ID: INTEGER
Postal_Code: CHAR(10)
From_Street_Num: SMALLINT
To_Street_Num: SMALLINT
Postal_code: INTEGER
City_name: INTEGER
Count: INTEGER
Country_id: INTEGER
Street_Code_FKIndex1
Postal_Code_ID
Street_Code_FKIndex2
County_ID
State
State_ID: INTEGER
Country_Country: CHAR(2) (FK)
State_Code: CHAR(2)
State_Name: CHAR(30)
State_type: INTEGER (FK)
State_FKIndex1
Country_Country
State_FKIndex2
State_type
County
County_ID: INTEGER
County_Name: VARCHAR(60)
Province_Name: VARCHAR(30)
Region_Name: VARCHAR(30)
State_ID: INTEGER (FK)
Country: CHAR(2)
County_type: INTEGER (FK)
Province_type: INTEGER (FK)
Region_type: INTEGER (FK)
County_FKIndex4
State_ID
County_FKIndex2
County_type
County_FKIndex3
Province_type
County_FKIndex4
Region_type
Rel_46
Orders
Order_ID: INTEGER
Customer_ID: INTEGER (FK)
Order_Type: INTEGER
Order_Date: DATE
Delivery_Date: DATE
Orders_FKIndex1
Customer_ID
Order_Item
Order_Item_Num: SMALLINT
Order_ID: INTEGER (FK)
Employee_ID: INTEGER (FK)
Product_ID: INTEGER (FK)
Quantity: SMALLINT
Total_Retail_Price: DECIMAL(13,2)
CostPrice_Per_Unit: DECIMAL(13,2)
Discount: DECIMAL(5,2)
Order_Item_FKIndex1
Order_ID
Order_Item_FKIndex2
Product_ID
Order_Item_FKIndex3
Employee_ID
Organisation
Employee_ID: INTEGER
Organisation_Employee_ID: INTEGER (FK)
Org_Name: CHAR(40)
Country: CHAR(2)
Start_Date: DATE
End_Date: DATE
Organisation_FKIndex1
Organisation_Employee_ID
Rel_54
Org_Level
Org_Level: SMALLINT
Org_Level_Name: CHAR(40)
Rel_37
Rel_38
Rel_39
Rel_40
Geo_Type
Geo_Type_ID: INTEGER
Geo_Type_Name: CHAR(20)
Country: CHAR(2)
Postal_Code
Postal_Code_ID: INTEGER
City_ID: INTEGER (FK)
Postal_Code: CHAR(10)
City_Name: VARCHAR(30)
Postal_Code_FKIndex1
City_ID
City
City_ID: INTEGER
City_Name: VARCHAR(30)
Country: CHAR(2)
Supplier
Supplier_ID: INTEGER
Street_ID: INTEGER (FK)
Supplier_Name: VARCHAR(30)
Supplier_Address: CHAR(45)
Sup_Street_Number: CHAR(8)
Country: CHAR(2)
Supplier_FKIndex1
Street_ID
Price_List
Start_Date: DATE
Product_ID: INTEGER
Product_List_Product_ID: INTEGER (FK)
End_Date: DATE
Unit_Cost_Price: DECIMAL(13,2)
Unit_Sales_Price: DECIMAL(13,2)
Factor: DECIMAL(13,2)
Price_List_FKIndex1
Product_List_Product_ID
Rel_57
Product_List
Product_ID: INTEGER
Product_List_Product_ID: INTEGER (FK)
Product_Level_Product_Level: INTEGER (FK)
Supplier_ID: INTEGER (FK)
Product_Name: CHAR(45)
Product_level: SMALLINT
Product_ref_id: INTEGER
Product_List_FKIndex2
Supplier_ID
Product_List_FKIndex4
Product_Level_Product_Level
Product_List_FKIndex5
Product_List_Product_ID
Rel_53
Rel_52
Rel_58
Product_Level
Product_Level: INTEGER
Product_Level_Name: VARCHAR(30)
Discount
Start_Date: DATE
Product_ID: INTEGER
Product_List_Product_ID: INTEGER (FK)
End_Date: DATE
Unit_Sales_Price: DECIMAL(13,2)
Discount: DECIMAL(5,2)
Discount_FKIndex1
Product_List_Product_ID

Modèle de données du Data Warehouse



Guide de démarrage avec SAS Enterprise Guide 4.3

Introduction

Aaron Levenstein
Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel.

Jean Dion
Les chiffres sont aux analystes ce que les lampadaires sont aux ivrognes : ils fournissent bien plus un appui qu'un éclairage.

George E.P. Box
All models are wrong but some are useful

Paul Valéry
Tout ce qui est simple est faux, tout ce qui ne l'est pas est inexploitable.

Les statistiques vues par les décideurs :

Churchill :
Je ne crois qu'aux statistiques que j'ai moi-même falsifiées.

Jacques Maillot - 1949
Les statistiques ont une particularité majeure : elles ne sont jamais les mêmes selon qu'elles sont avancées par un homme de gauche ou par un homme de droite.

L'objectif ici est la manipulation de données, la création de rapports et la génération de statistiques. La plateforme décisionnelle SAS offre de nombreuses possibilités de manipulation de données, de création de rapport et d'analyse de données. Nous utiliserons dans le cadre de ce travail dirigé uniquement SAS Enterprise Guide®.

Ce guide a aussi pour vocation d'introduire quelques notions de statistiques de base par l'exemple. Il ne se substitue en aucun cas à un cours de statistiques complet. Il a donc pour vocation de simplement mettre le pied à l'étrier. L'étudiant souhaitant aller plus loin pourra trouver les informations nécessaires notamment dans le guide d'utilisation de SAS.
Aucun pré requis n'est nécessaire pour commencer ce travail dirigé.

Afin de compléter l'aide de SAS Enterprise Guide, vous trouverez sur le site suivant un tutorial très pédagogique en anglais pour faire des analyses statistiques.
http://support.sas.com/learn/statlibrary/

Le mot statistique donne parfois à lui tout seul de l'urticaire à certains individus. Le but est ici de prouver, même à de telle personne, que cela ne mord pas et peut souvent être très utile. Nous n'aborderons ici presque pas les statistiques par l'approche théorique classique, mais nous commencerons plutôt par une petite question managériale sur notre cas fil rouge « Orion Star ». Nous déroulerons ensuite simplement l'analyse permettant d'y répondre. Nous insisterons donc sur l'interprétation des résultats et peu sur les fondements théoriques. L'objectif n'est pas de former des spécialistes de l'analyse de données ; il est simplement de se familiariser avec celle-ci. Si à la fin de ces travaux pratiques vous êtes convaincus de la nécessité d'analyser les données, ce but sera atteint.

Proverbe chinois :
J'entends et j'oublie,
Je vois et je retiens,
Je fais et je comprends.

Une simple présentation sur l'importance de l'analyse de données serait trop vite oubliée. Nous allons faire quelques statistiques simples pour vous permettre de comprendre ce que cela peut vous

apporter. L'objectif n'est pas de faire de vous statisticiens. Il est de vous permettre d'améliorer votre communication avec les statisticiens et de vous permettre de faire et d'interpréter quelques statistiques simples.

La prise de décision nécessite une connaissance métier forte. Il me semble qu'il ne faut surtout pas utiliser les statistiques pour se justifier mais plutôt avoir une culture du chiffre qui entraîne naturellement l'intuition. La première étape des statistiques est de présenter clairement des chiffres qui présentent la situation actuelle et indique implicitement la voie à suivre. Ensuite l'analyse de données peut être utilisée pour modéliser un passé connu et en l'appliquant sur le présent, en déduire le futur.

Le nom SAS vient à l'origine de « Statistical Analysis System». Maintenant, c'est un nom et cela ne se prononce pas « S.A.S. » mais « Sas ». Né dans les statistiques, SAS propose des solutions décisionnelles où les statistiques, la modélisation, la simulation ou la recherche opérationnelle notamment, sont des composants démultiplicateurs de retour sur investissement importants. SAS excelle notamment particulièrement dans l'analyse de données pointue, la modélisation rigoureuse ou l'optimisation complexe, sur de gigantesques bases de données, structurées ou non.
Le mot statistique donne parfois à lui tout seul de l'urticaire à certains individus. Le but est ici de prouver, même à de telles personnes, que cela ne mord pas et peut souvent être très utile.
L'objectif n'est pas de faire ici des statisticiens, je ne ferai pas un vrai cours de statistiques, seulement de vous permettre d'améliorer votre communication avec les statisticiens et de vous permettre de faire et d'interpréter quelques statistiques simples.

« Les chiffres sont aux analystes ce que les lampadaires sont aux ivrognes : ils fournissent bien plus un appui qu'un éclairage. » - Jean Dion
Il me semble important de rappeler qu'il ne faut surtout pas utiliser les statistiques pour se justifier mais plutôt avoir une culture du chiffre qui entraîne naturellement l'intuition. La première étape des statistiques est de présenter clairement des chiffres qui présentent la situation actuelle et indique implicitement la voie à suivre. Ensuite l'analyse de données peut être utilisée pour modéliser un passé connu et en l'appliquant sur le présent, en déduire le futur potentiel.

Ceci est particulièrement vrai pour dans le domaine du marketing. On imagine souvent que le marketing c'est du feeling et de la communication, ce qui fait plutôt rêver : Faux ! C'est surtout du travail et de la rigueur, avec une très forte dose de chiffres. Sans analyse, il est très difficile de comprendre le marché et de prendre de bonnes décisions.

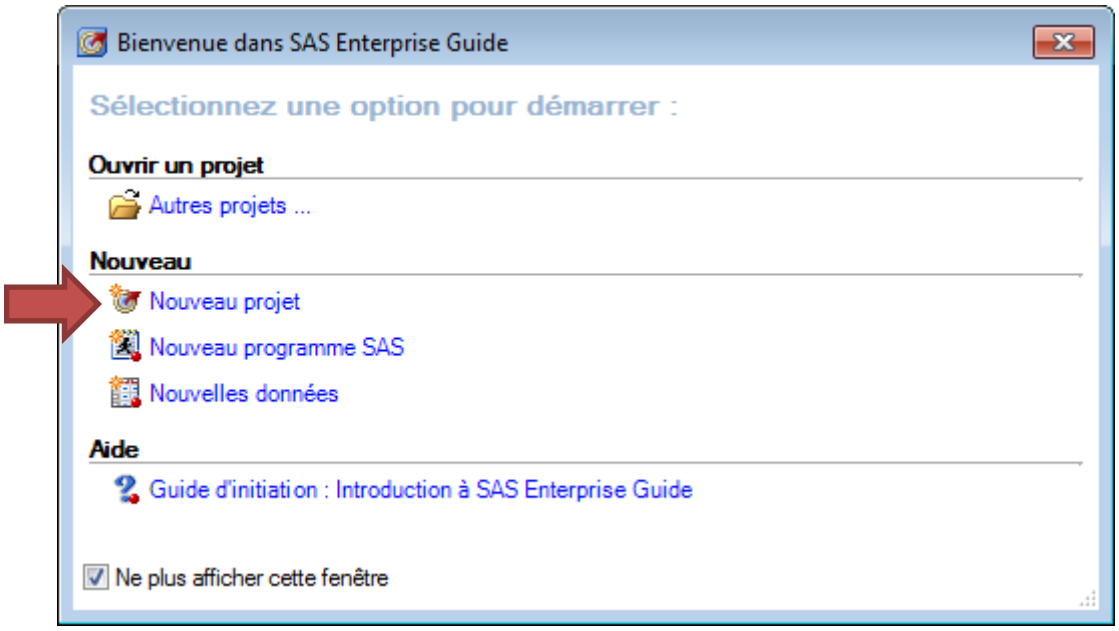
Mon souhait ici est de présenter quelques statistiques simples, un peu de Data Mining et d'optimisation. Néanmoins, chaque sujet pour prétendre former des experts, nécessiterait plusieurs volumes conséquents. Excusez-moi la banalité suivante : il est simple de faire des statistiques simples, et cela peut apporter beaucoup ; et il est compliqué de faire des modèles complexes, mais c'est parfois nécessaire et cela peut rapporter énormément.

Vous allez voir, ce n'est pas méchant et cela peut vous rapporter !

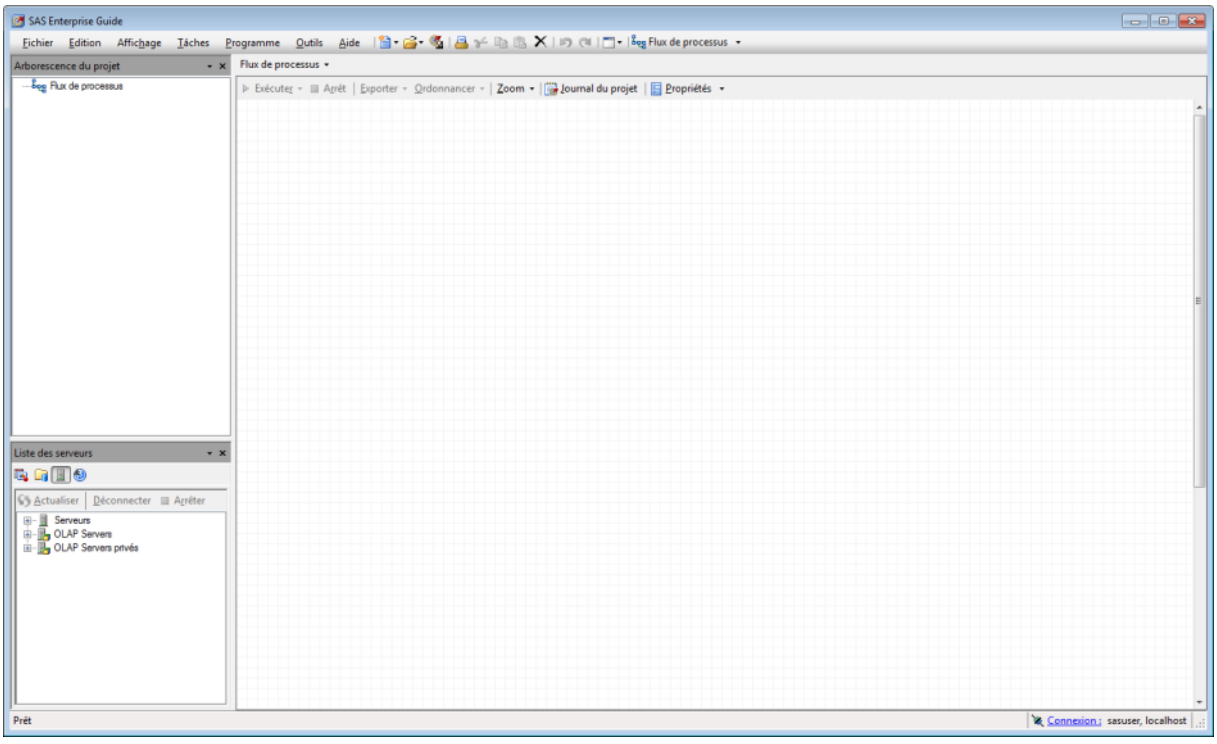
Démarrage avec SAS Enterprise Guide

Ouvrir SAS Enterprise Guide

Ouvrir SAS Enterprise Guide
Depuis Démarrer → Programmes → SAS → Enterprise Guide 4.3

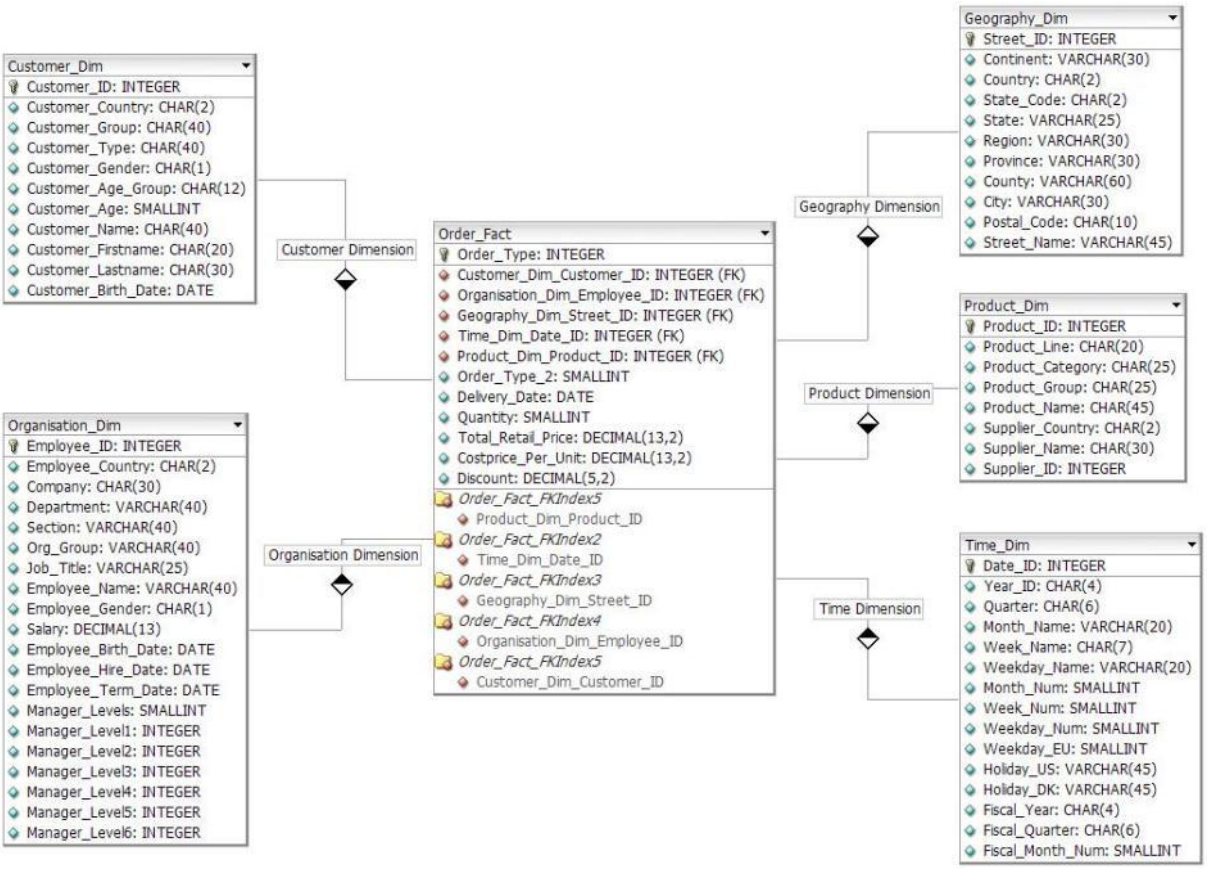


Si cette fenêtre s'ouvre, sélectionner un nouveau projet



Bienvenue dans l'application SAS Enterprise Guide.

Schéma de la base de données



Les 6 tables de l'entrepôt de données « Orion Gold » ci-dessus stockent l'historique des transactions de la société Orion entre le 1er janvier 2003 et le 31 décembre 2007.

La table de fait centrale, **Order_Fact,** stocke les lignes des commandes. Chaque ligne de cette table correspond à une ligne de commande. Une commande peut avoir plusieurs lignes, une par produit. Sur chaque ligne d'une commande, on retrouve

- le numéro du produit vendu (Product_ID)
- la quantité de ce produit vendu dans cette commande (Quantity)
- le prix unitaire (Costprice_per_unit)
- la somme facturée pour ce produit = le chiffre d'affaires (Total_Retail_Price)
    - c'est-à-dire : Total_Retail_Price = prix de vente unitaire x la quantité x la remise
- le numéro du client qui a commandé (Customer_ID)
- le numéro du commercial qui a fait la vente (Employee_ID)
- la date de la commande (Date_ID)
- l'identifiant de l'adresse où habite le client (Street_ID)

Autour de cette table de fait se trouvent 5 tables dites de dimension qui détaillent le produit, le client, son adresse, le commercial et la date de la commande. Par exemple, dans la table Product_Dim, avec le numéro du produit, on a le nom du produit, son groupe, sa catégorie, sa ligne de produit, son fournisseur et le pays de ce fournisseur.

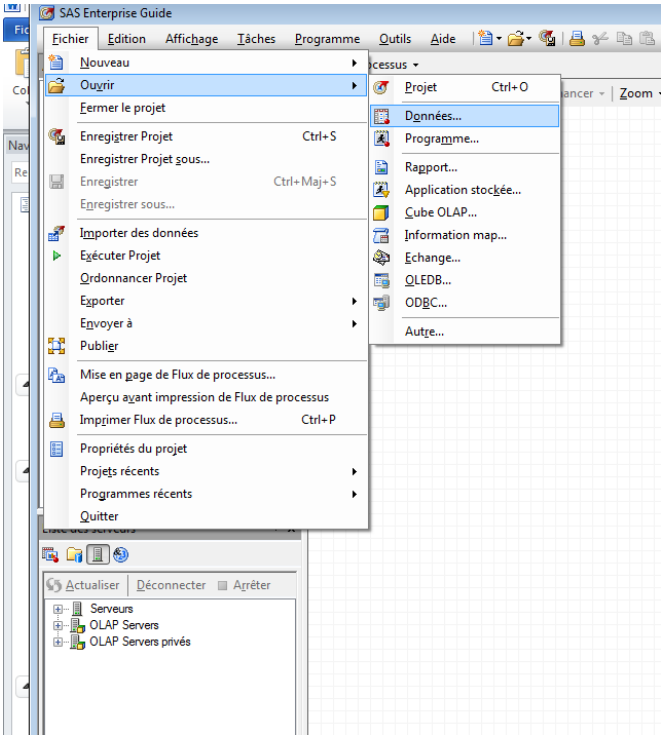
Questions :

Dans le cadre de ce travail dirigé, nous allons répondre notamment aux questions suivantes du directeur marketing :

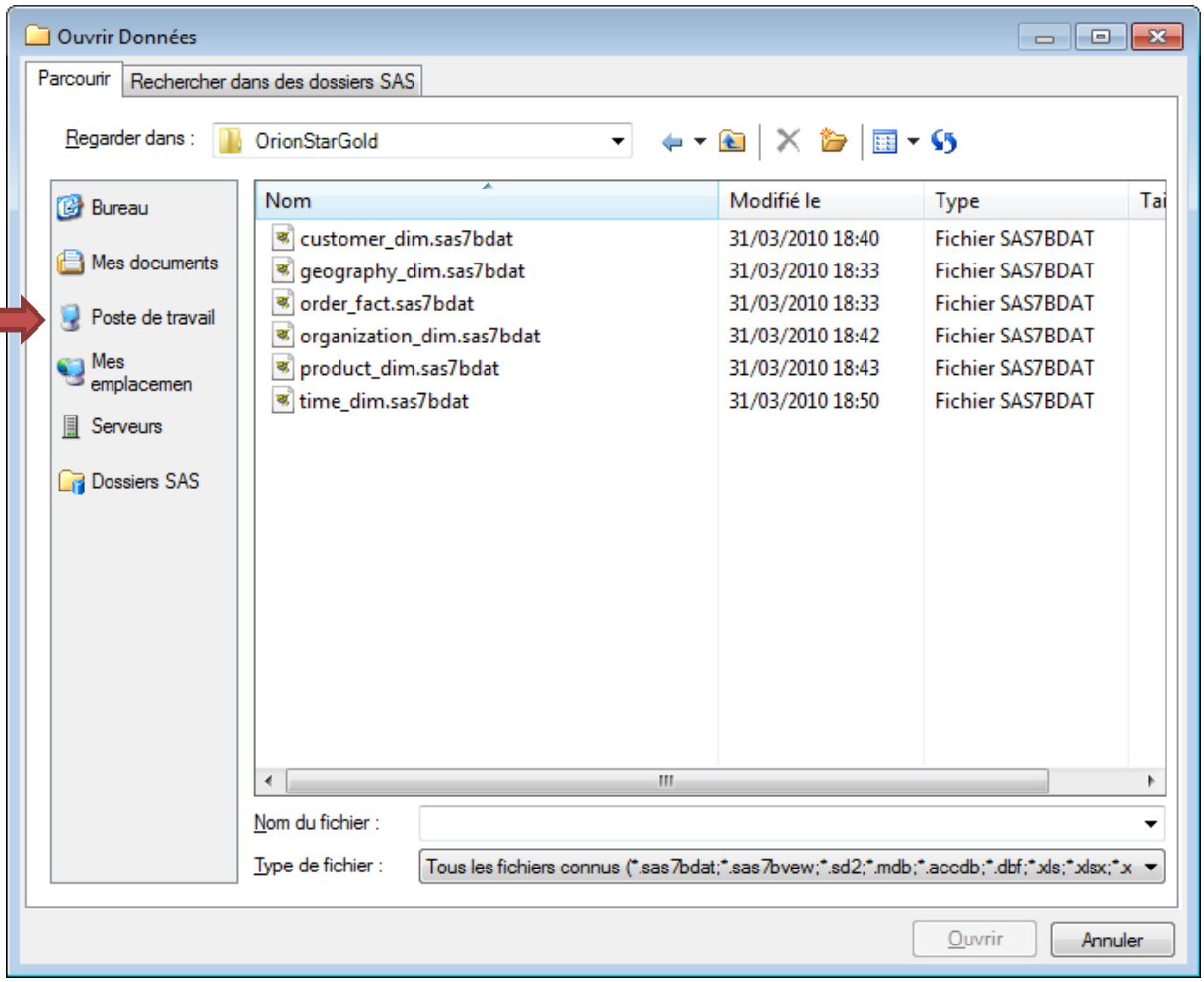
- Quels sont les 20 produits qui se vendent le mieux en termes de chiffre d'affaires ?
- Quelle est la moyenne et l'écart type du chiffre d'affaires ?
- A-t-il indépendance entre le sexe des commerciaux et celui des clients ?
- Quelles sont les variables qui expliquent le mieux l'importance du chiffre d'affaires ?
  A-t-il une différence significative entre la moyenne de la somme du chiffre d'affaires géré par les commerciaux de sexe féminin et celle des commerciaux de sexe masculin ?
- Trouver la relation et son importance, entre le chiffre d'affaires généré par un commercial et son salaire.

Pour commencer, il faut ouvrir les tables fournies avec ce petit guide.

Ouvrir une table :

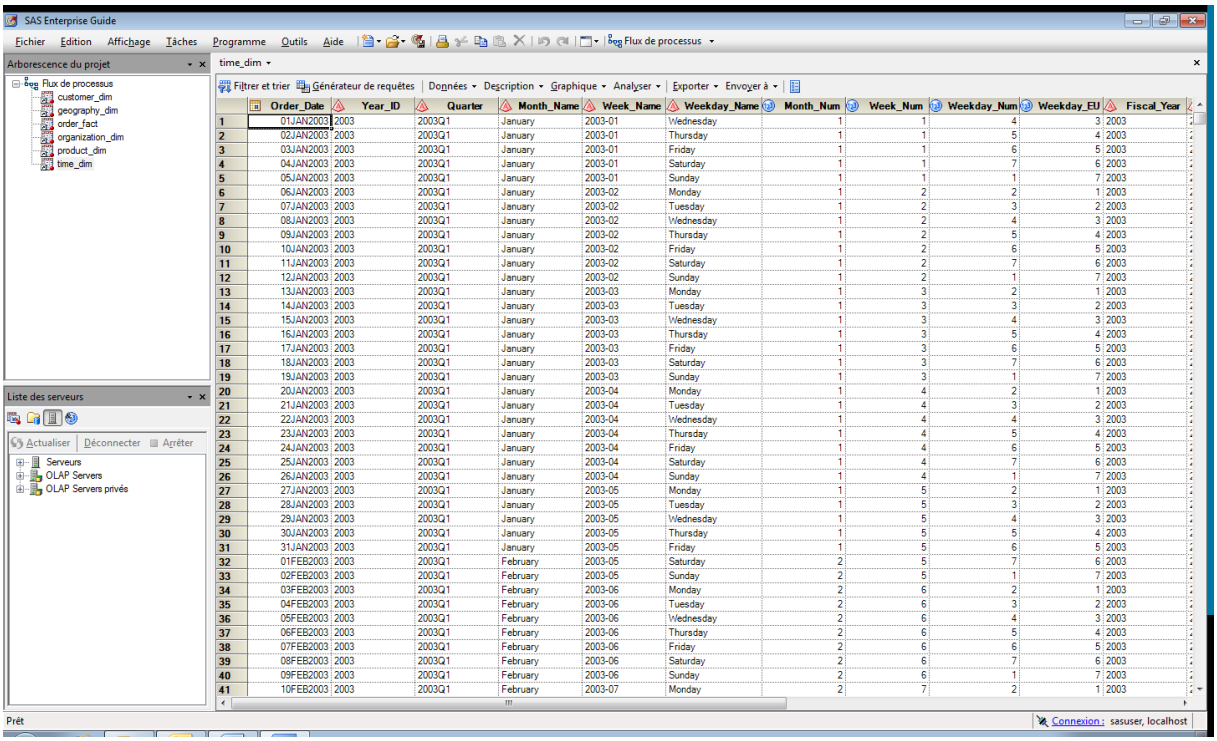


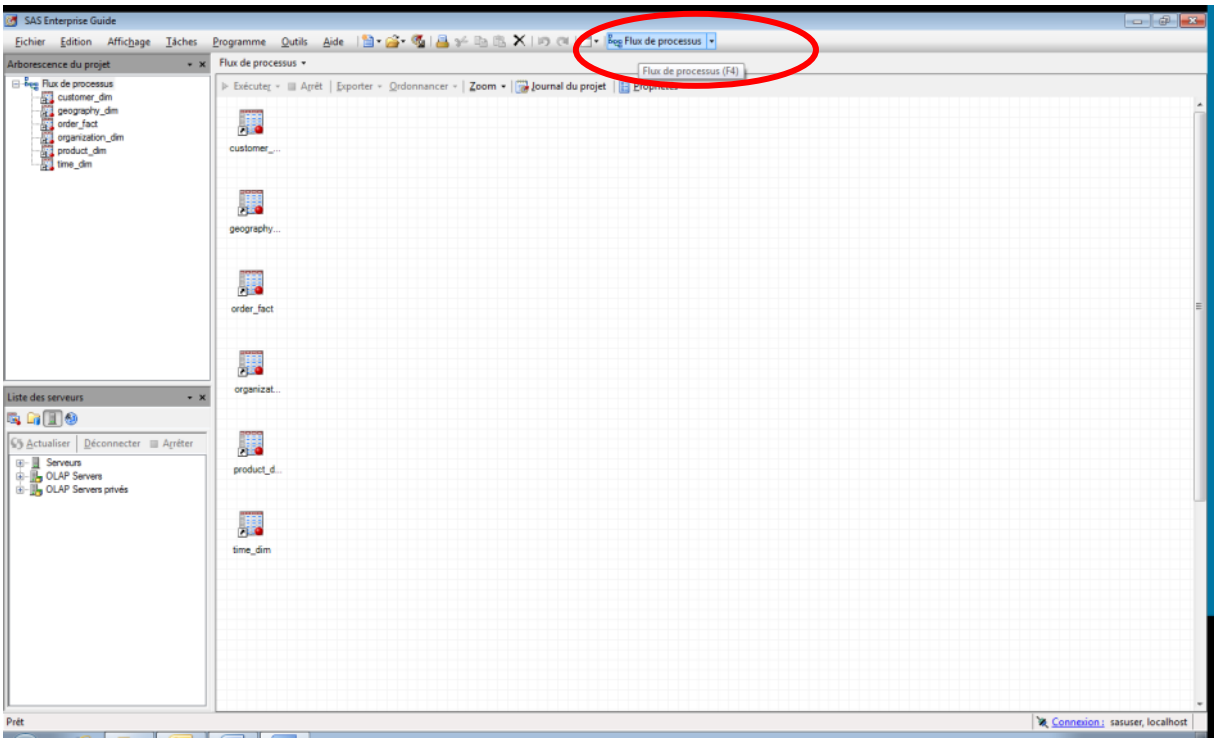
Depuis le menu **Fichier → Ouvrir → Données**



Depuis votre **Ordinateur local**

Aller dans le dossier **Orion_Star_Gold** où se trouvent les données associées à ce support.
Sélectionner toutes les tables (par exemple avec la touche contrôle enfoncée)
Cliquer sur **ouvrir**



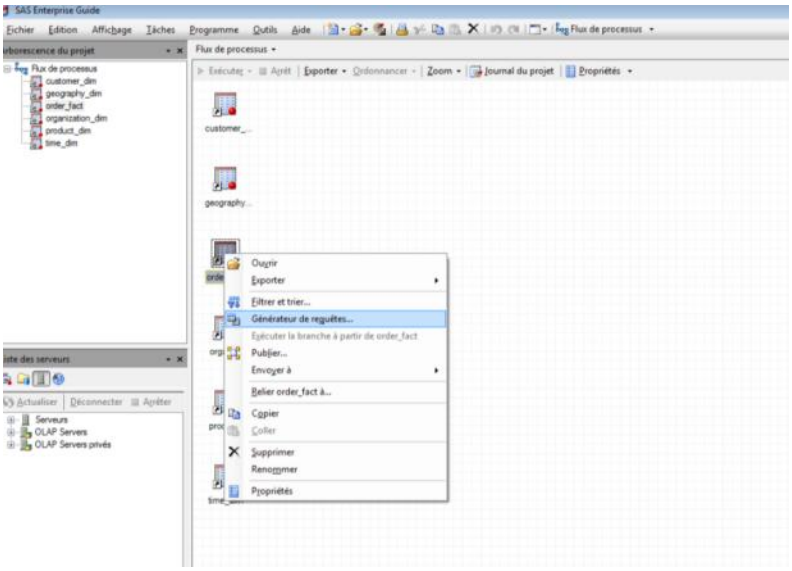


Cliquer sur le bouton **Flux de processus** pour accéder au flux de processus du projet.

Faire une requête

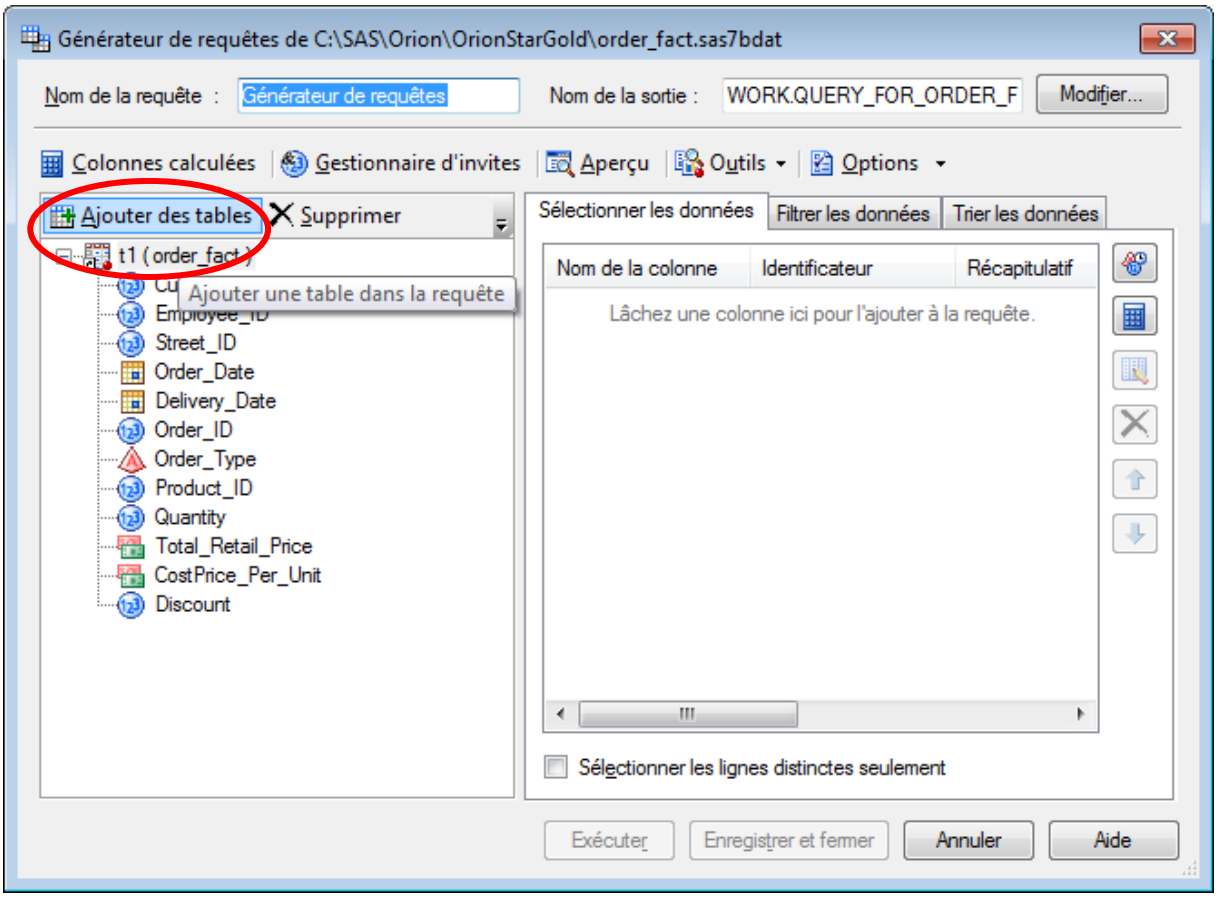
Objectif :

Créer une table de la somme du chiffre d'affaires par nom des produits, triée par la somme décroissante du chiffre d'affaires.

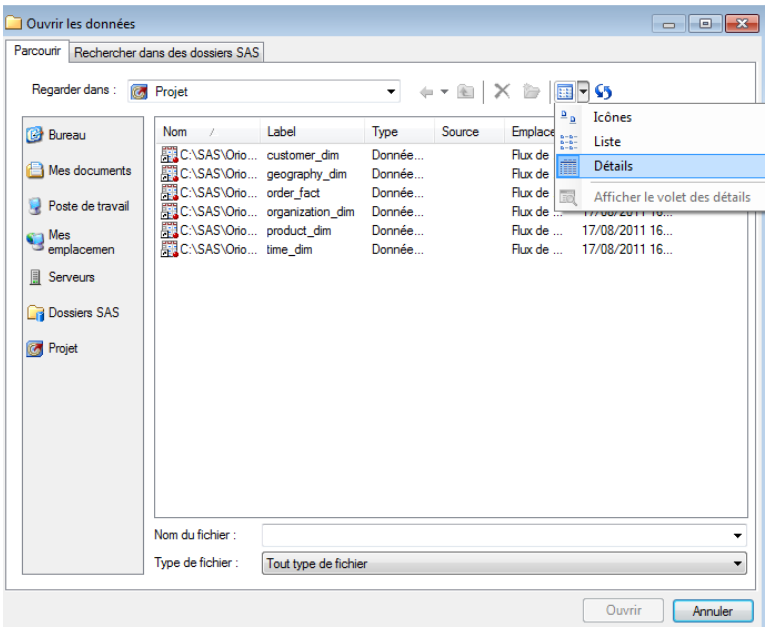


Sélectionner la table en cliquant dessus, soit dans l'explorateur du projet, soit dans le flux de processus ; la table est alors grisée.
Clic-droit de la souris sur cette table sélectionnée, **générateur de requête**.

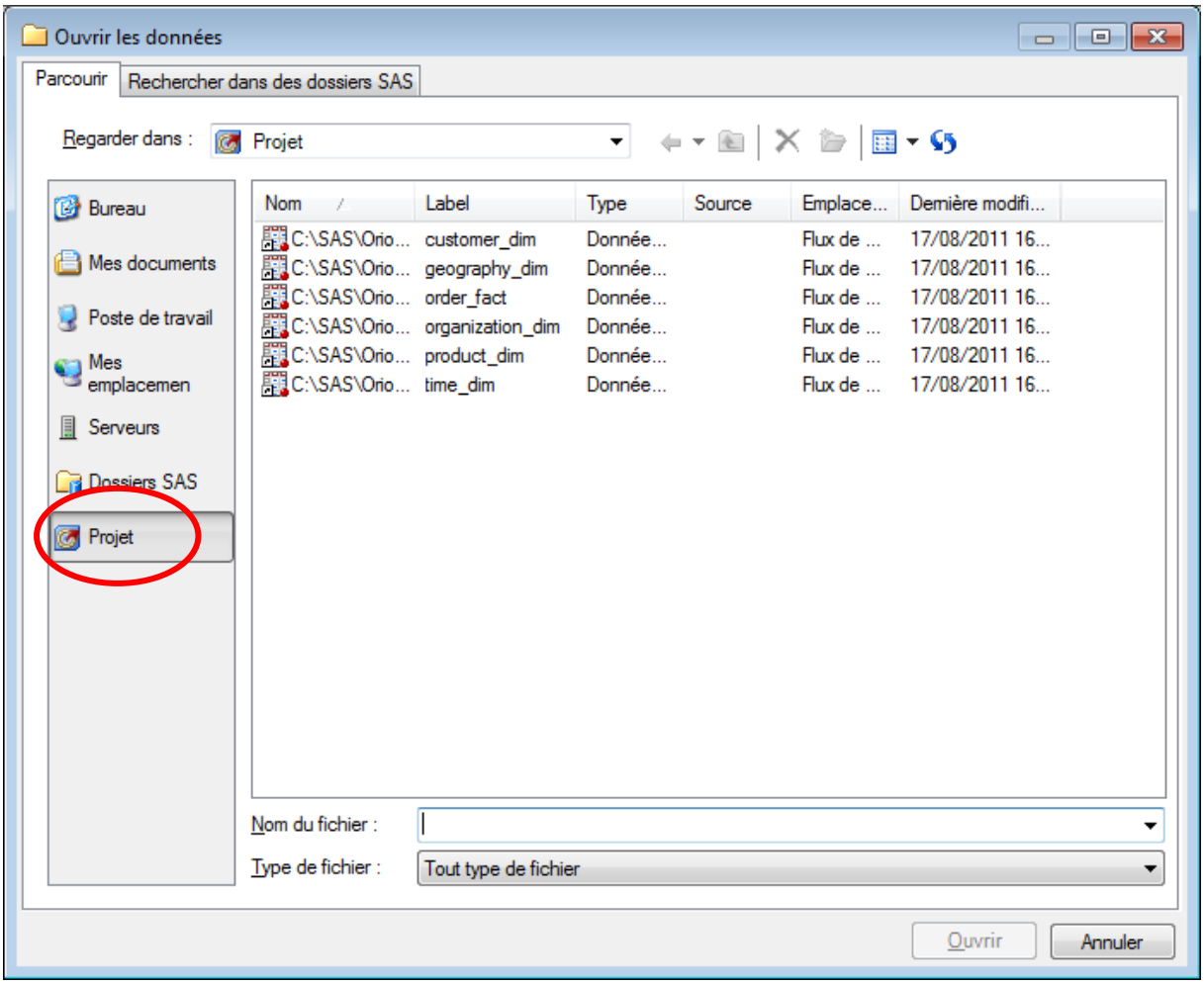
Dans le cadre de notre premier exercice, sélectionner la table **Order_fact** → clic-droit → **générateur de requête**.



Pour ajouter une table, cliquer sur le bouton **Ajouter des tables** en haut à gauche.

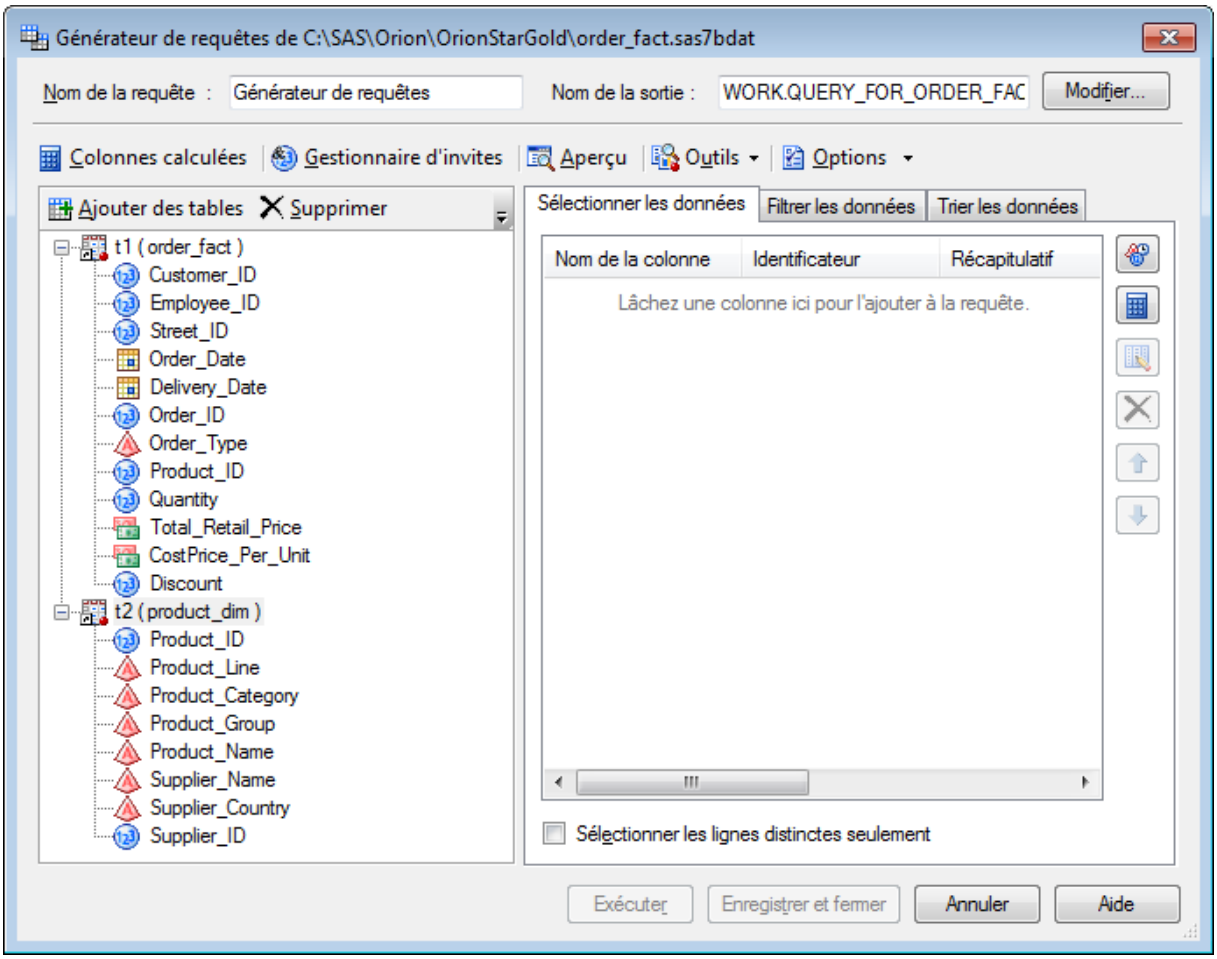


Si besoin, afficher les éléments détaillés

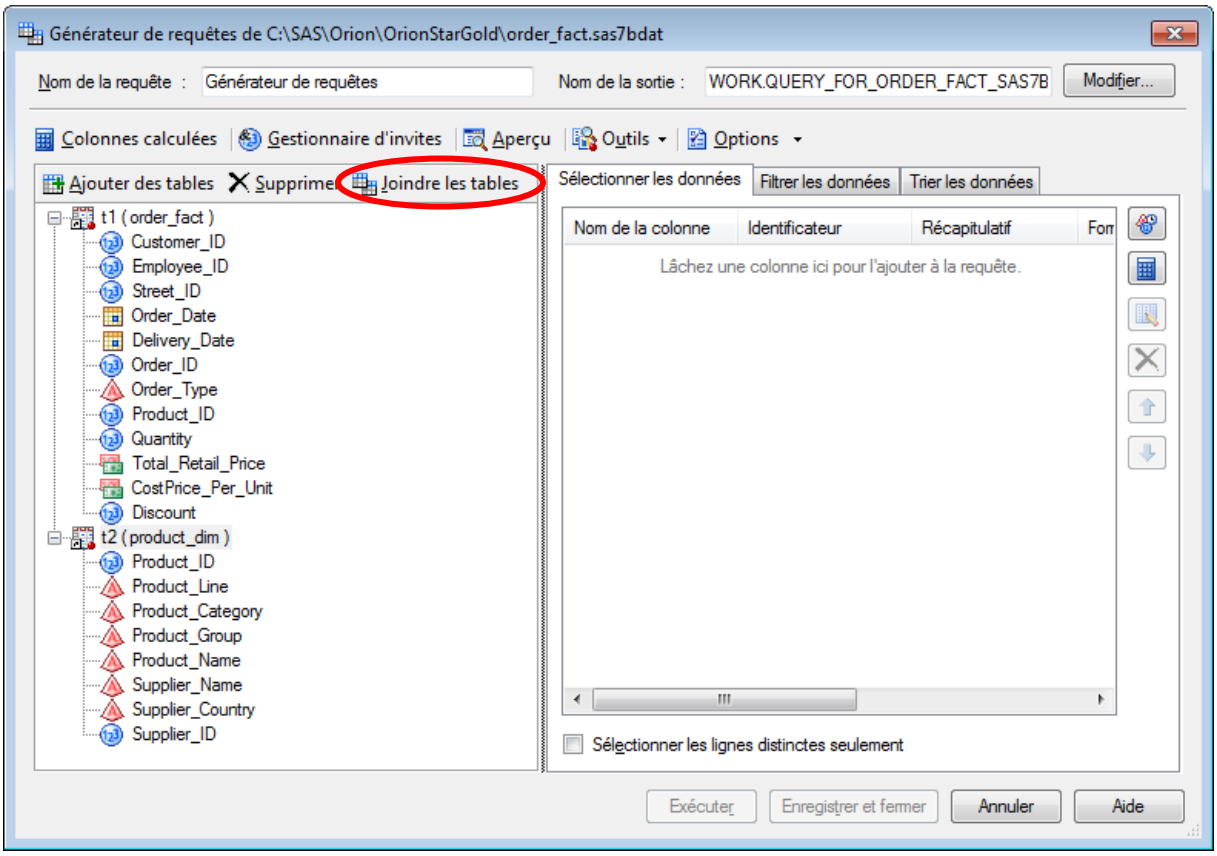


Depuis le **projet**,

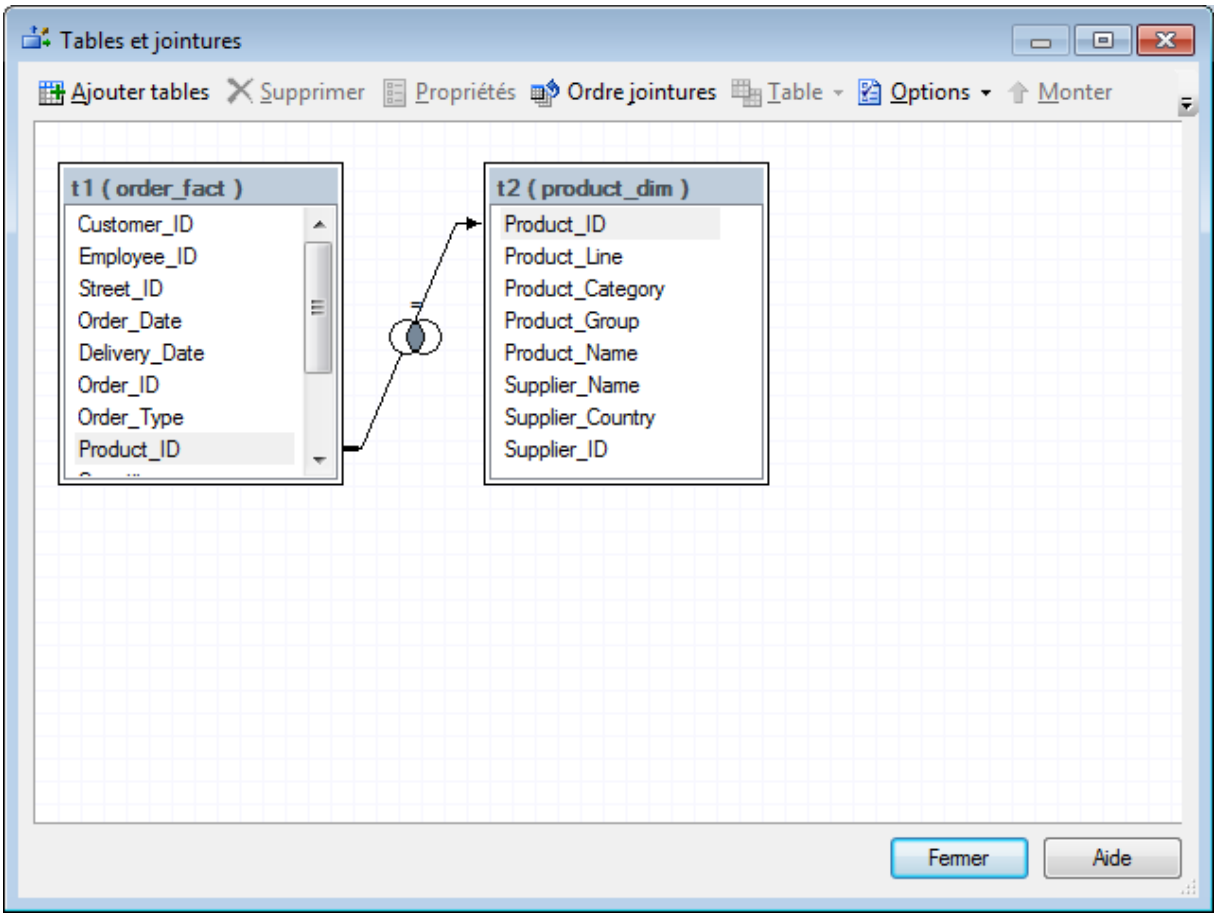
Dans le cadre de notre premier exercice, sélectionner la table **Product_dim** (la table Order_fact étant déjà dans la requête).
Ouvrir



Pour voir la jointure (à titre pédagogique, ce n'est pas nécessaire pour l'exercice)

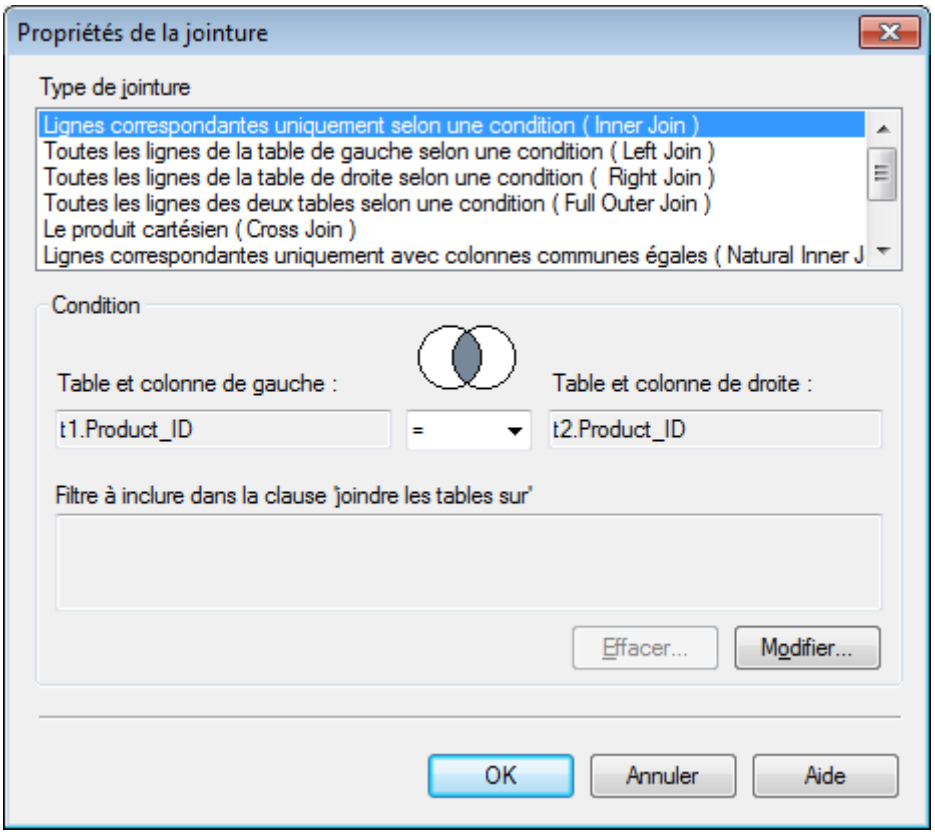


Cliquez sur **joindre les tables**,
NB : si le bouton n'apparait pas, agrandir la fenêtre ou cliquer sur la flèche vers le bas.



La jointure entre les tables **Order_fac** et **Product_dim** et ici une jointure de type « INNER JOIN ».
Si l'on clique dessus on peut sélectionner le type de jointure que l'on souhaite, soit :

Jointure Inner Join



Exemple de jointure « Inner Join »:

A	B
1	a
2	b
3	c
4	d

+

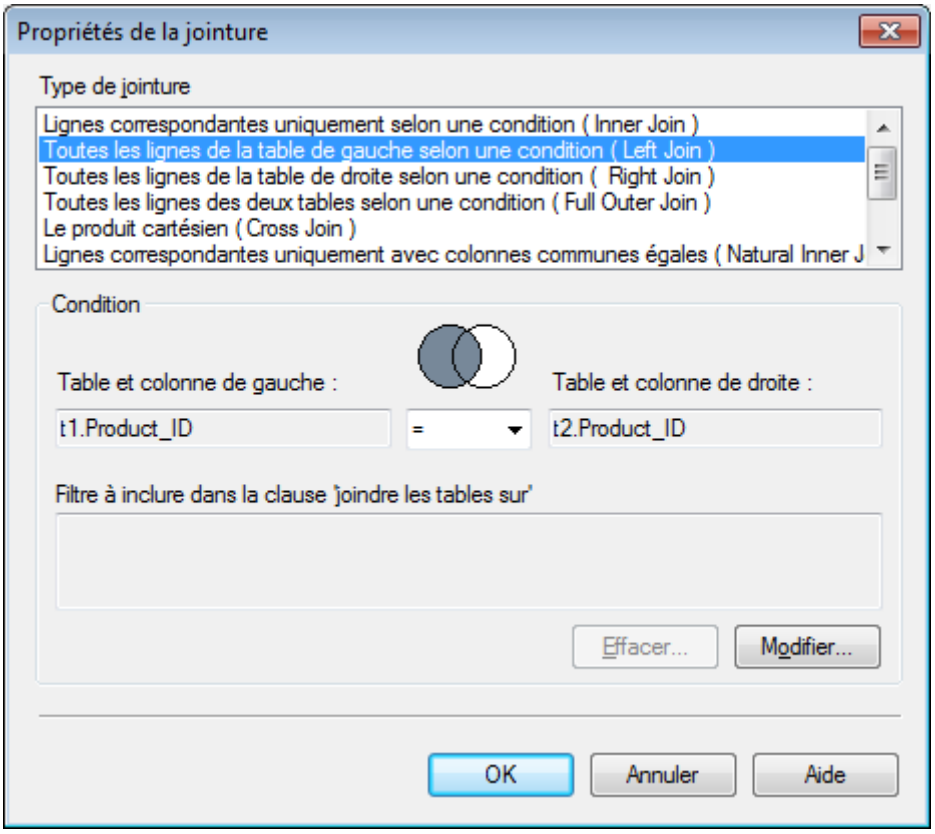
C	D
a	A
b	B
d	C
e	D

=

A	B	C	D
1	a	a	A
2	b	b	B
4	d	d	C

Dans l'exemple ci-dessus la jointure entre les deux tables se fait sur les colonnes B et C, c'est-à-dire que si la valeur de la colonne B est exactement celle de la colonne C de l'autre table, alors les informations de la première table sont ajoutées aux informations de la seconde dans la table de sortie. La valeur c de la colonne B et la valeur e de la colonne C n'ont pas de correspondance. On ne les retrouve pas dans la table de sortie.

Jointure gauche



Exemple de jointure gauche :

A	B
1	a
2	b
3	c
4	d

+

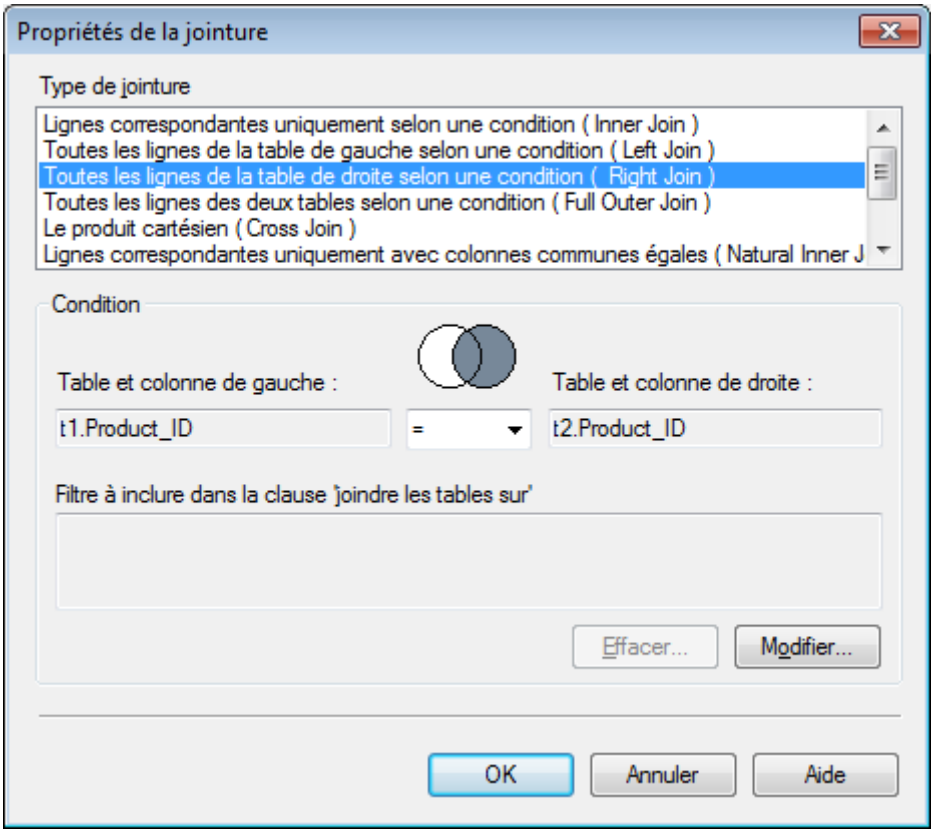
C	D
a	A
b	B
d	C
e	D

=

A	B	C	D
1	a	a	A
2	b	b	B
3	c		
4	d	d	C

On ajoute ici à toutes les colonnes de la table de gauche celles qui correspondent dans la table de droite.

Jointure droite



Exemple de jointure droite :

A	B
1	a
2	b
3	c
4	d

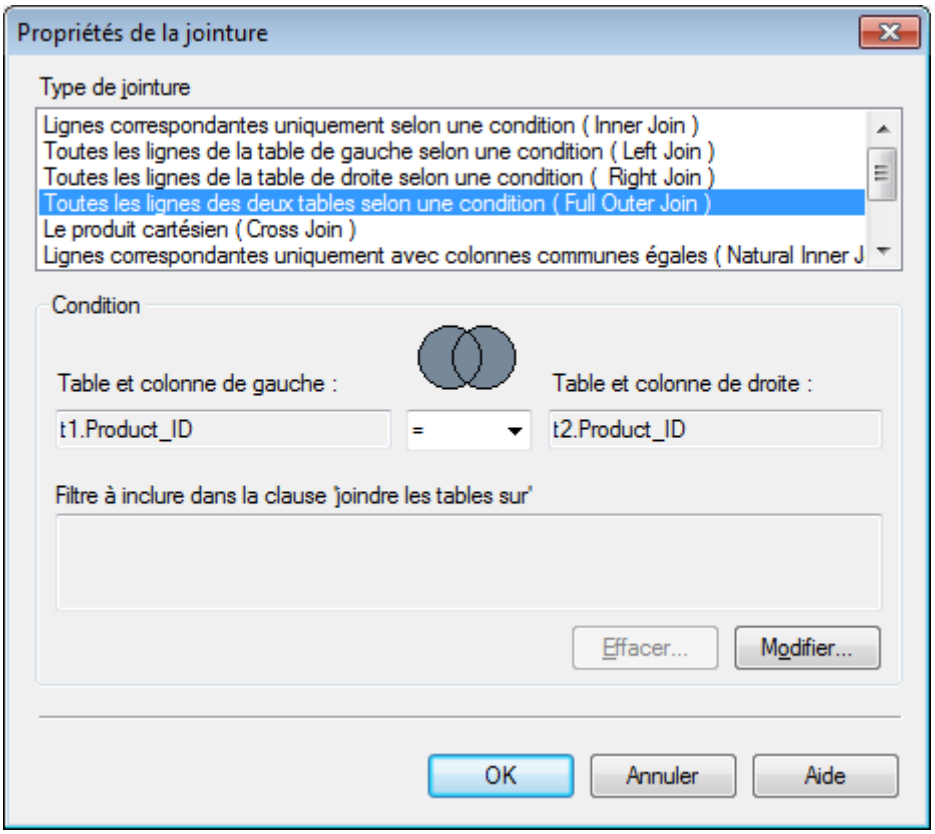
+

C	D
a	A
b	B
d	C
e	D

=

A	B	C	D
1	a	a	A
2	b	b	B
4	d	d	C
		e	D

Jointure complète



Exemple de jointure complète :

A	B
1	a
2	b
3	c
4	d

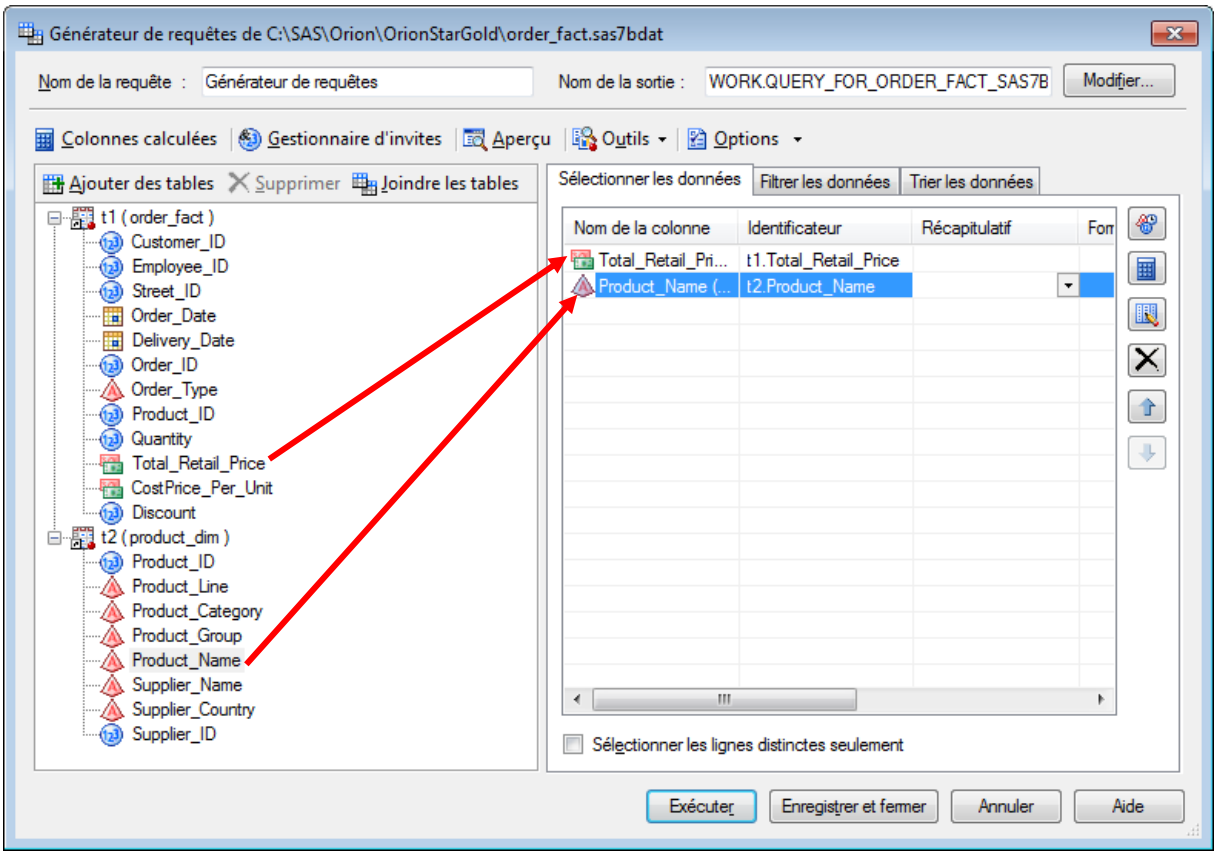
+

C	D
a	A
b	B
d	C
e	D

=

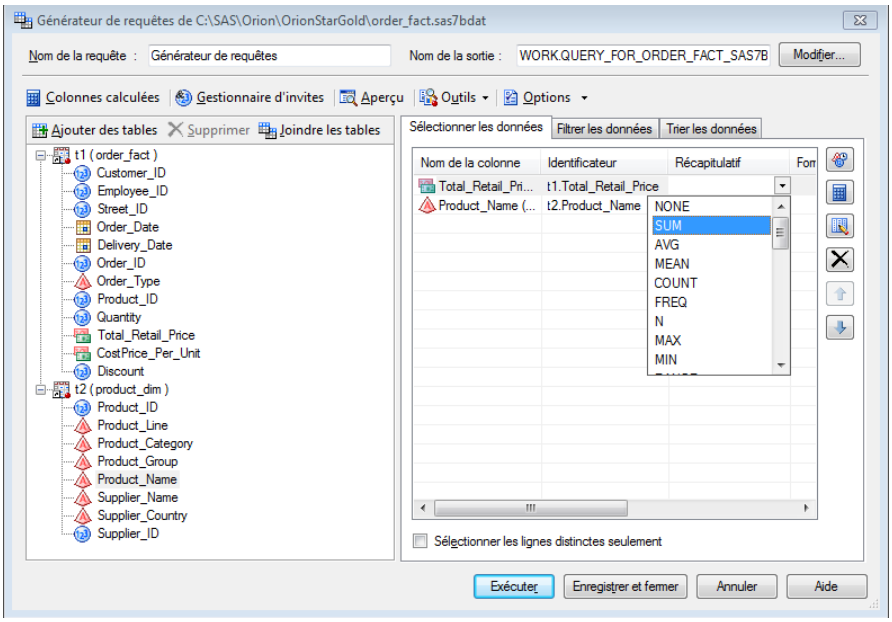
A	B	C	D
1	a	a	A
2	b	b	B
3	c		
4	d	d	C
		e	D

Pour revenir à la fenêtre principale de la création de la requête, fermer la fenêtre de la jointure.

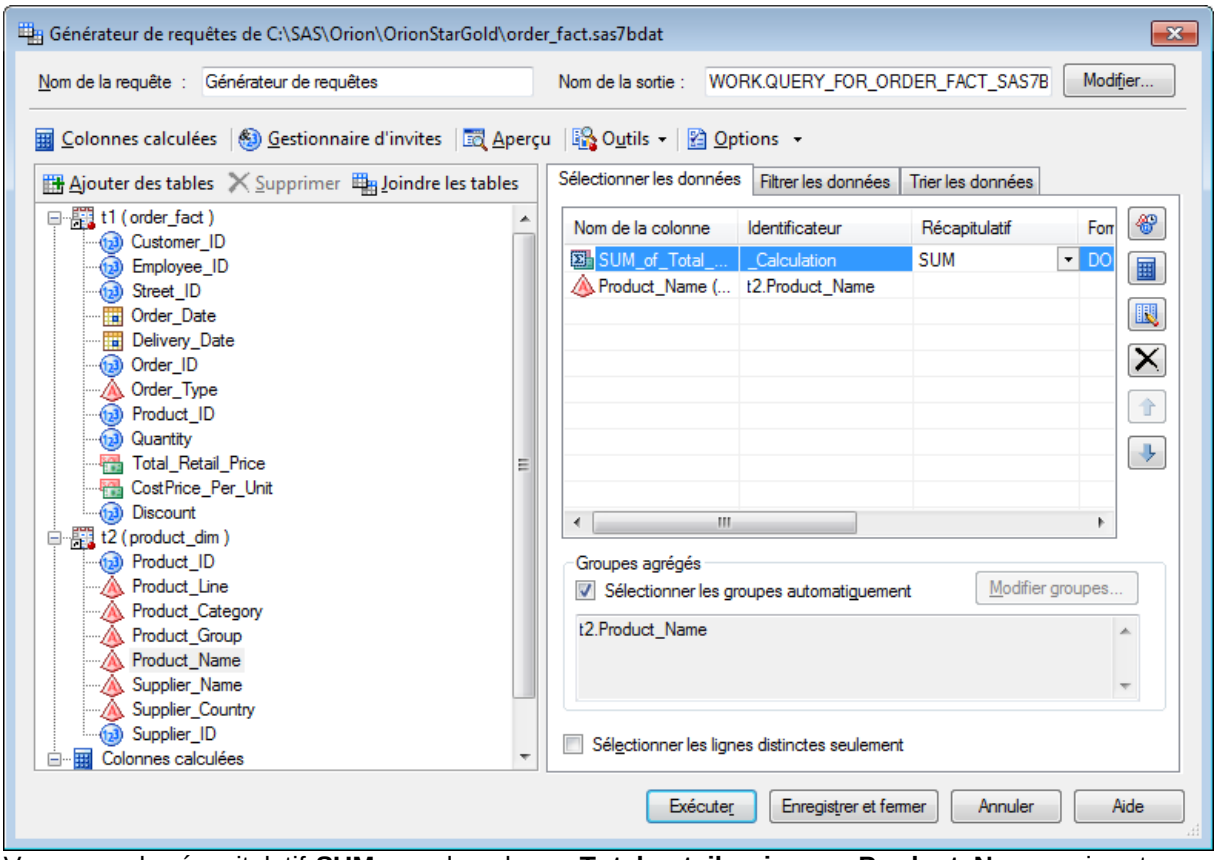


Dans la cadre de notre exercice, nous avons besoin de la somme des ventes au détail par produit. Il faut donc glisser-lâcher les colonnes **Total_Retail_Price** et **Product_Name** dans la fenêtre de droite.

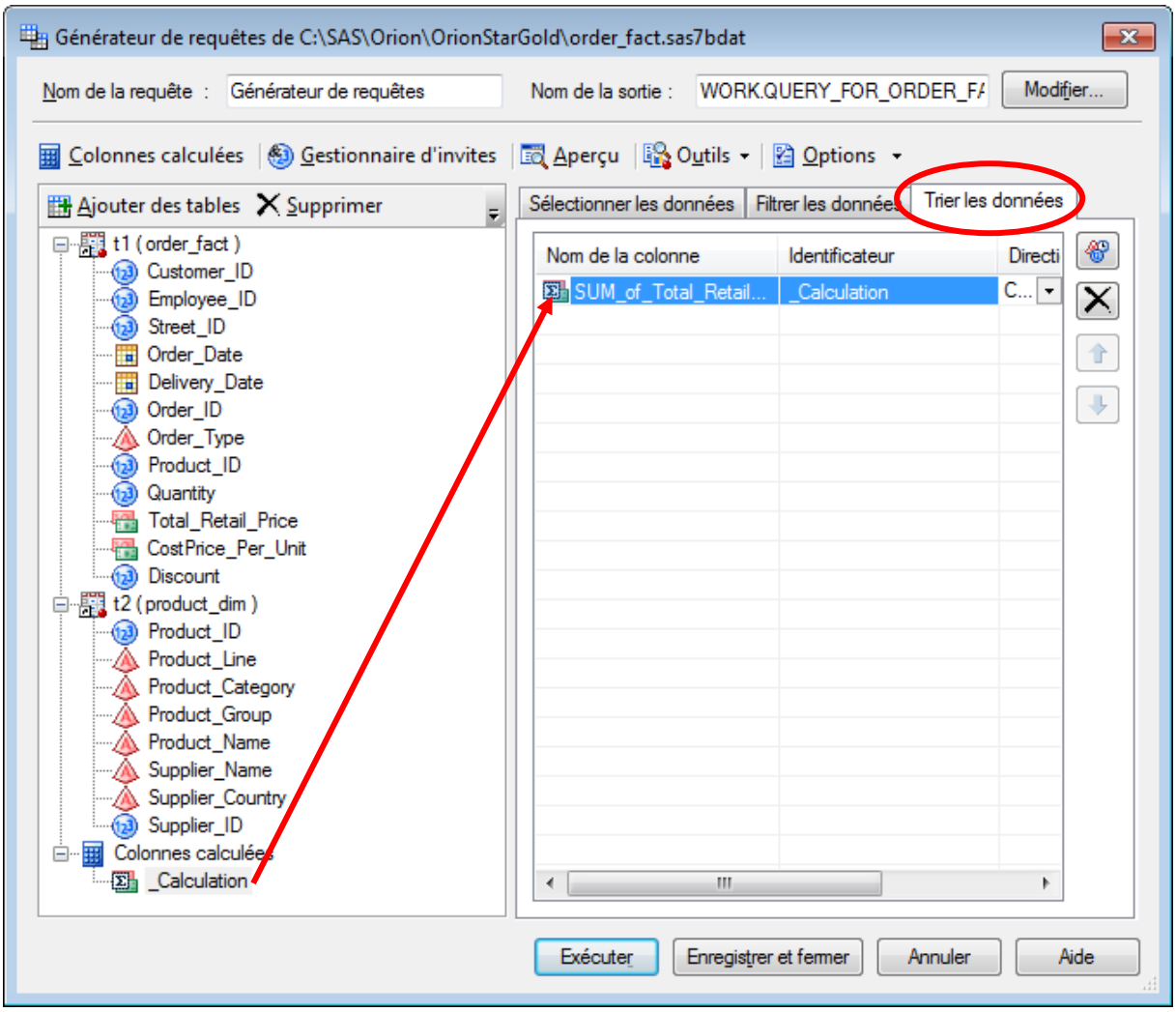
Pour avoir la somme des ventes au détail par produit, il faut sélectionner le récapitulatif de type somme, et le groupe produit.



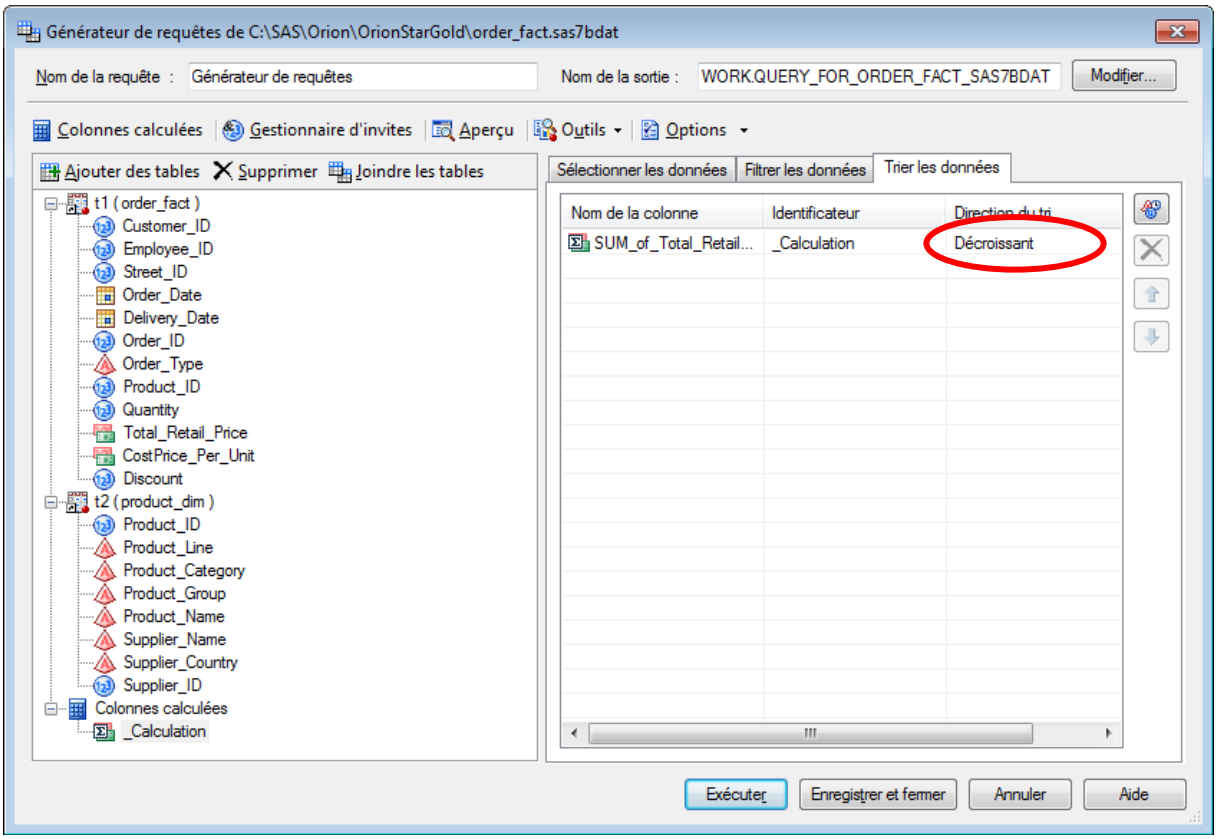
Sur la ligne **Total_Retail_Price**, dans **récapitulatif**, sélectionner **SUM**.



Vous avez le récapitulatif **SUM** pour la colonne **Total_retail_price** par **Product_Name** qui se trouve dans les groupes agrégés.

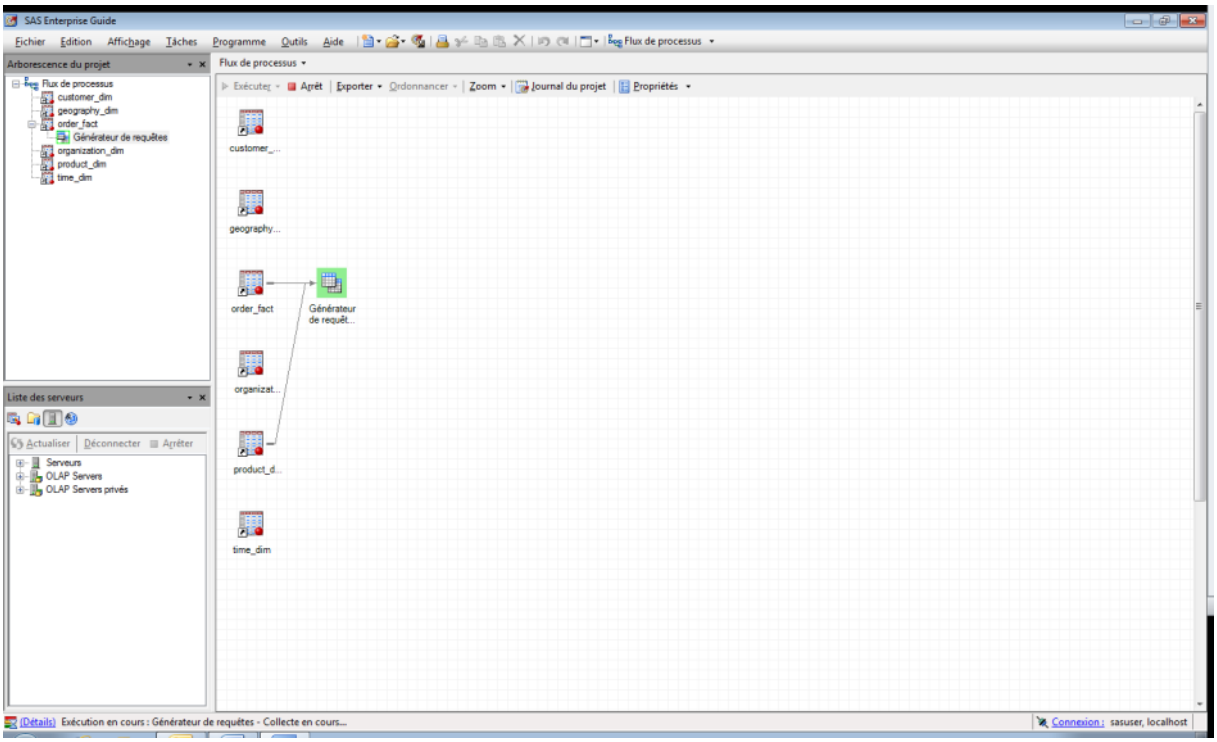


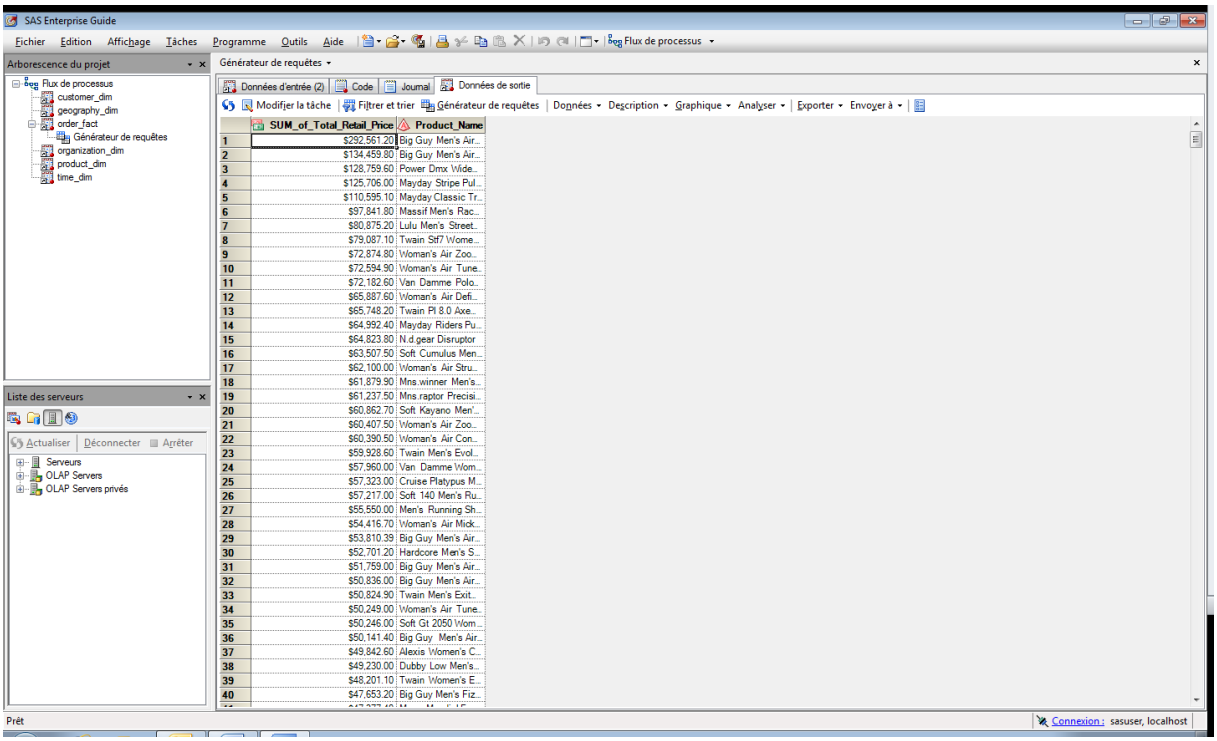
Dans l'onglet **trier les données**, glisser-lâcher la colonne **SUM_OF_Total_Retail_Price** dans la fenêtre de gauche.



Sélectionner un tri **décroissant**,

Cliquer sur le bouton **Exécuter**.





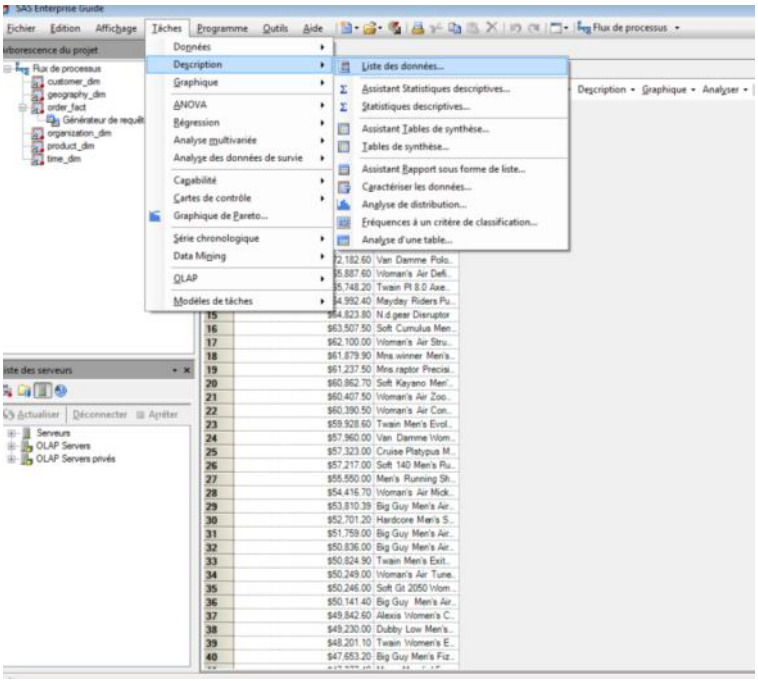
	SUM_of_Total_Retail_Price	Product_Name
1	$292,561.20	Big Guy Men's Air...
2	$134,459.80	Big Guy Men's Air...
3	$128,759.60	Power Dmx Wide...
4	$125,706.00	Mayday Stripe Pul...
5	$110,595.10	Mayday Classic Tr...
6	$97,841.80	Massif Men's Rac...
7	$80,875.20	Lulu Men's Street...
8	$79,087.10	Twain Stf7 Wome...
9	$72,874.80	Woman's Air Zoo...
10	$72,594.90	Woman's Air Tune...
11	$72,182.60	Van Damme Polo...
12	$65,887.60	Woman's Air Defi...
13	$65,748.20	Twain Pl 8.0 Axe...
14	$64,992.40	Mayday Riders Pu...
15	$64,823.80	N.d.gear Disruptor
16	$63,507.50	Soft Cumulus Men...
17	$62,100.00	Woman's Air Stru...
18	$61,879.90	Mns.winner Men's...
19	$61,237.50	Mns.raptor Precisi...
20	$60,862.70	Soft Kayano Men'...
21	$60,407.50	Woman's Air Zoo...
22	$60,390.50	Woman's Air Con...
23	$59,928.60	Twain Men's Evol...
24	$57,960.00	Van Damme Wom...
25	$57,323.00	Cruise Platypus M...
26	$57,217.00	Soft 140 Men's Ru...
27	$55,550.00	Men's Running Sh...
28	$54,416.70	Woman's Air Mick...
29	$53,810.39	Big Guy Men's Air...
30	$52,701.20	Hardcore Men's S...
31	$51,759.00	Big Guy Men's Air...
32	$50,836.00	Big Guy Men's Air...
33	$50,824.90	Twain Men's Exit...
34	$50,249.00	Woman's Air Tune...
35	$50,246.00	Soft Gt 2050 Wom...
36	$50,141.40	Big Guy Men's Air...
37	$49,842.60	Alexis Women's C...
38	$49,230.00	Dubby Low Men's...
39	$48,201.10	Twain Women's E...
40	$47,653.20	Big Guy Men's Fiz...

Nous avons la table de la somme du chiffre d'affaires par produit sur les 5 dernières années, triée par chiffre d'affaires décroissant. Nous allons maintenant créer un rapport simple de cette liste en ne prenant que les 20 premiers produits ayant générés le plus de chiffre d'affaires.

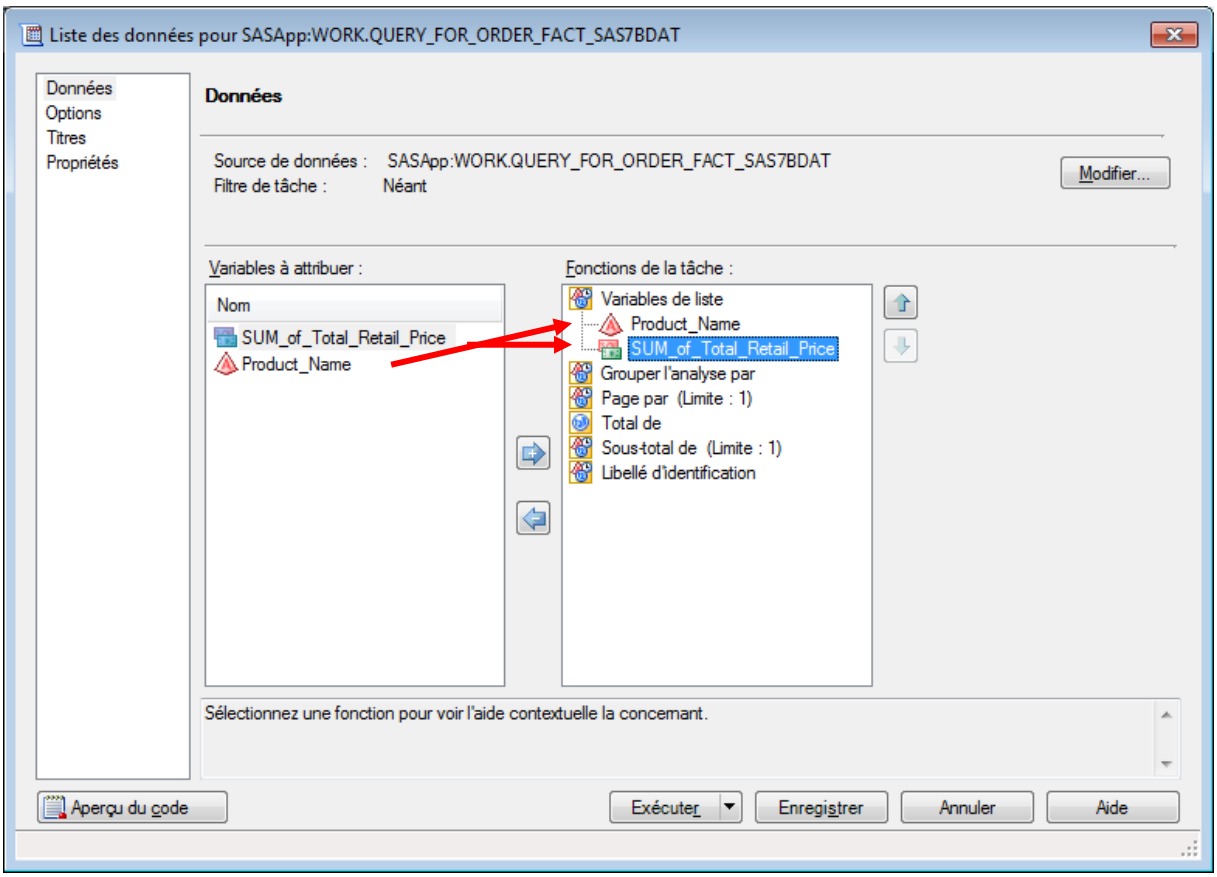
Création d'un rapport d'une liste de données

Objectif :

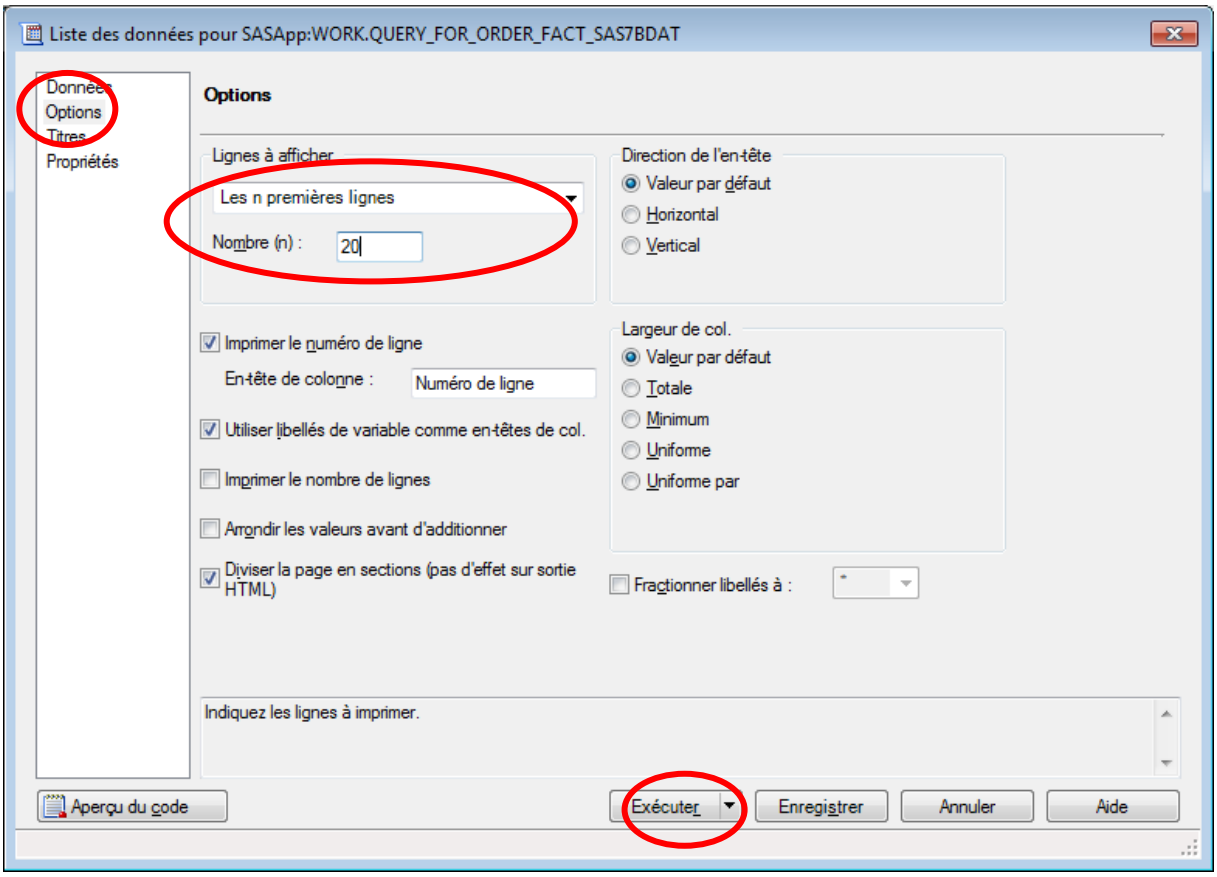
Créer la liste des vingt produits ayant générés les sommes du chiffre d'affaires les plus importantes.



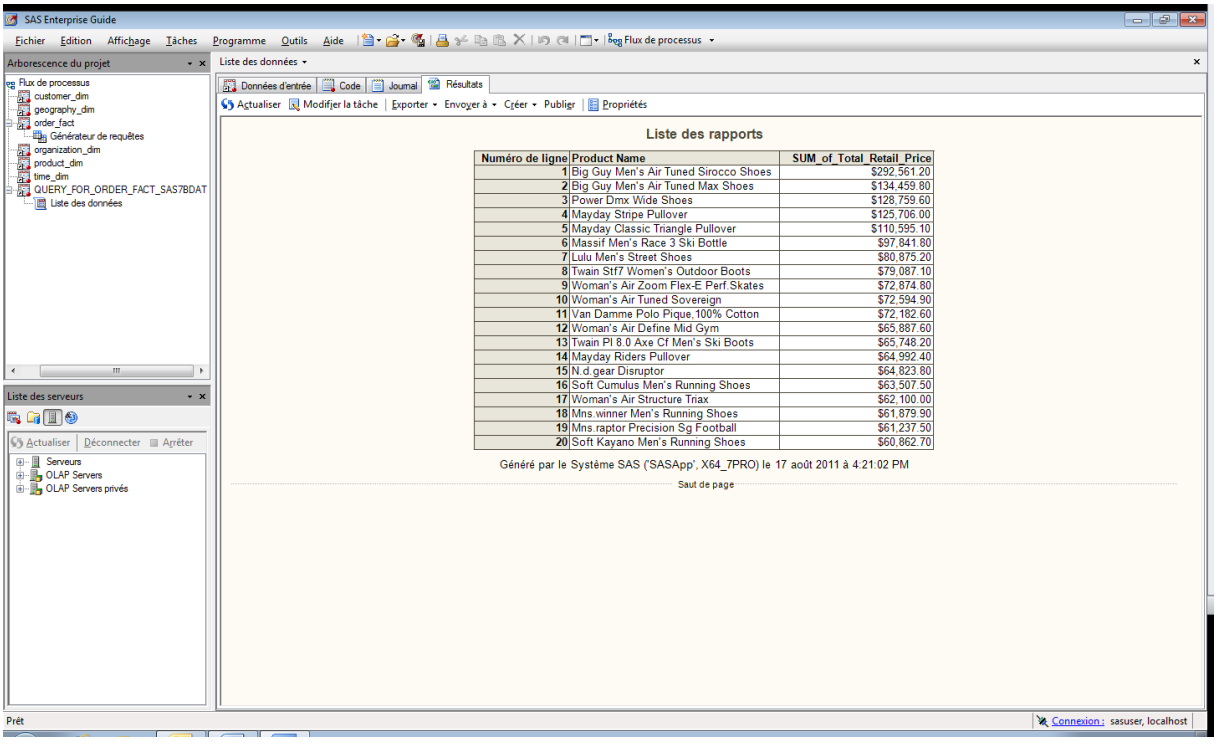
Dans le menu **Tâches** → **Description** → cliquer sur **liste de données**.



Glisser-lâcher les deux colonnes dans les variables de la liste.

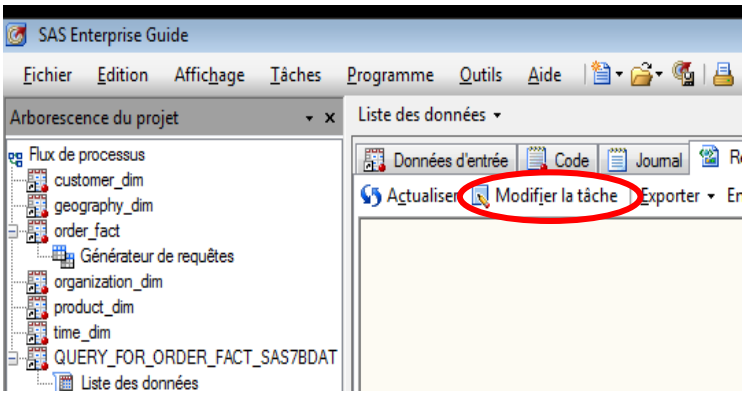


Sélectionner le paramétrage des options dans la fenêtre de gauche.
Sélectionner les n premières lignes, et entrer n= 20.
Exécuter

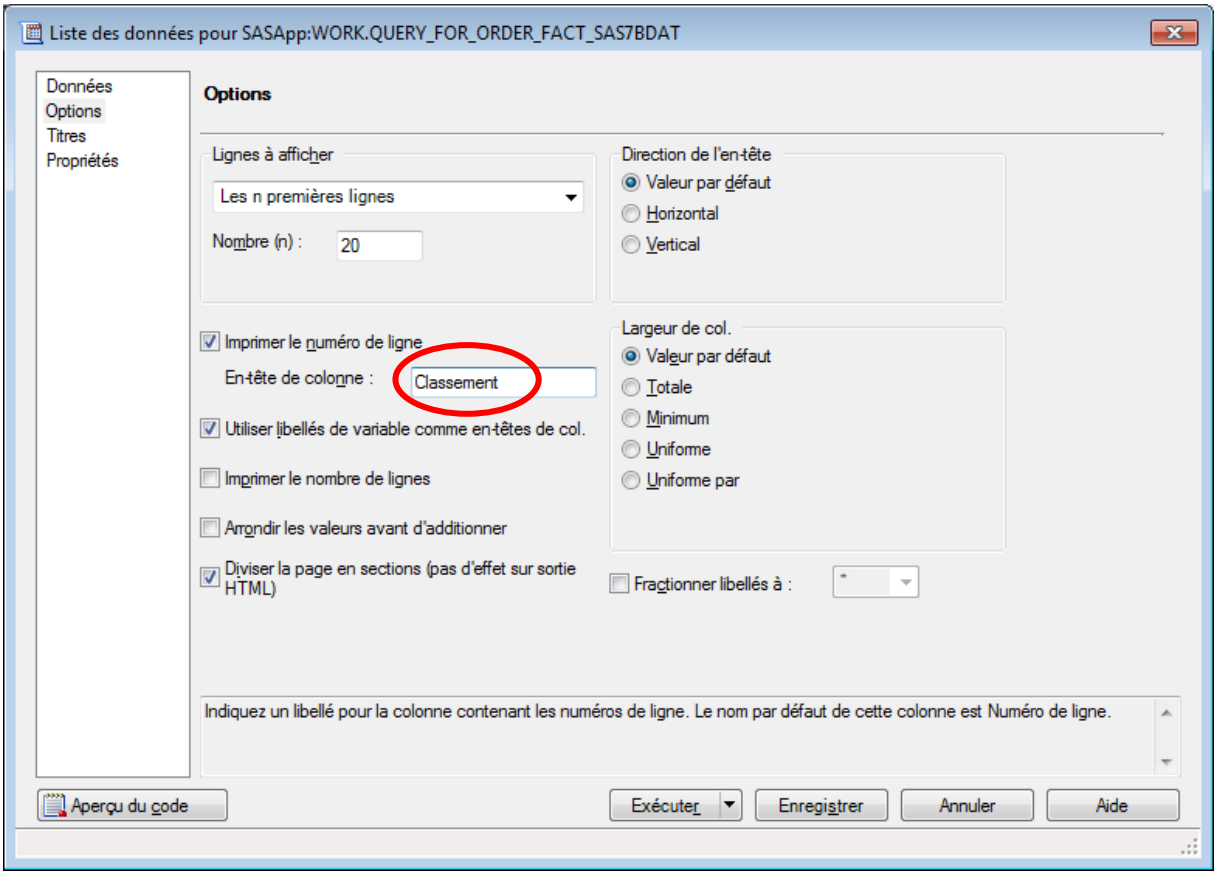


Et voilà notre premier rapport.
Exercice complémentaire :

Vous pouvez changer l'en tête de la colonne du numéro de la ligne.

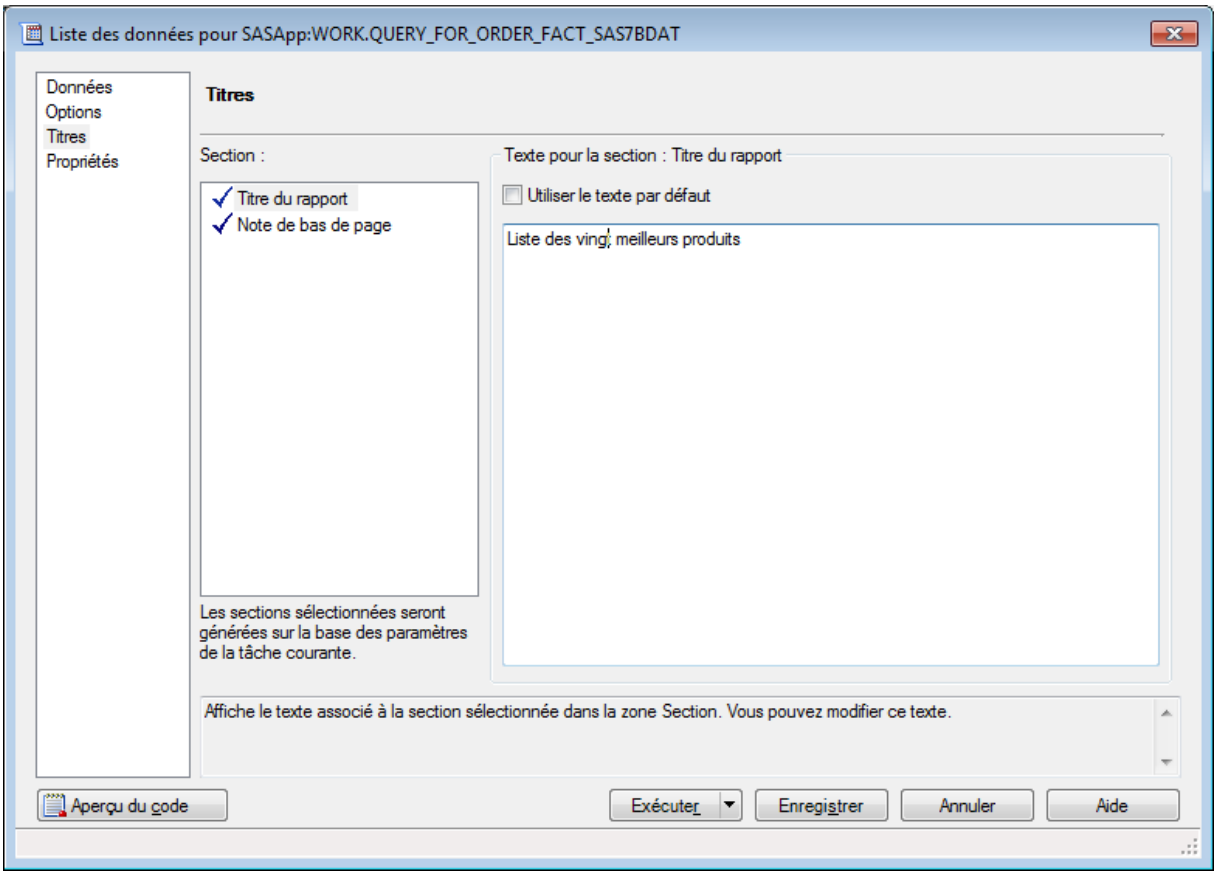


Cliquer sur modifier la tâche

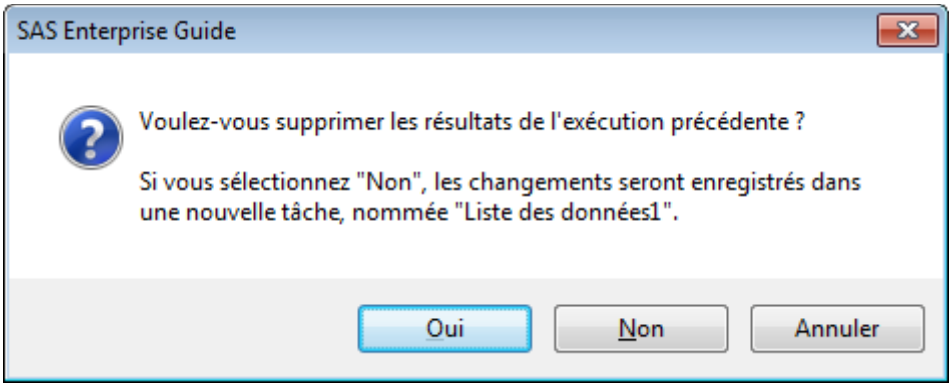


Dans la fenêtre des paramètres des options de la tâche de création de la liste des données. Changer l'entête de la colonne du numéro de ligne.

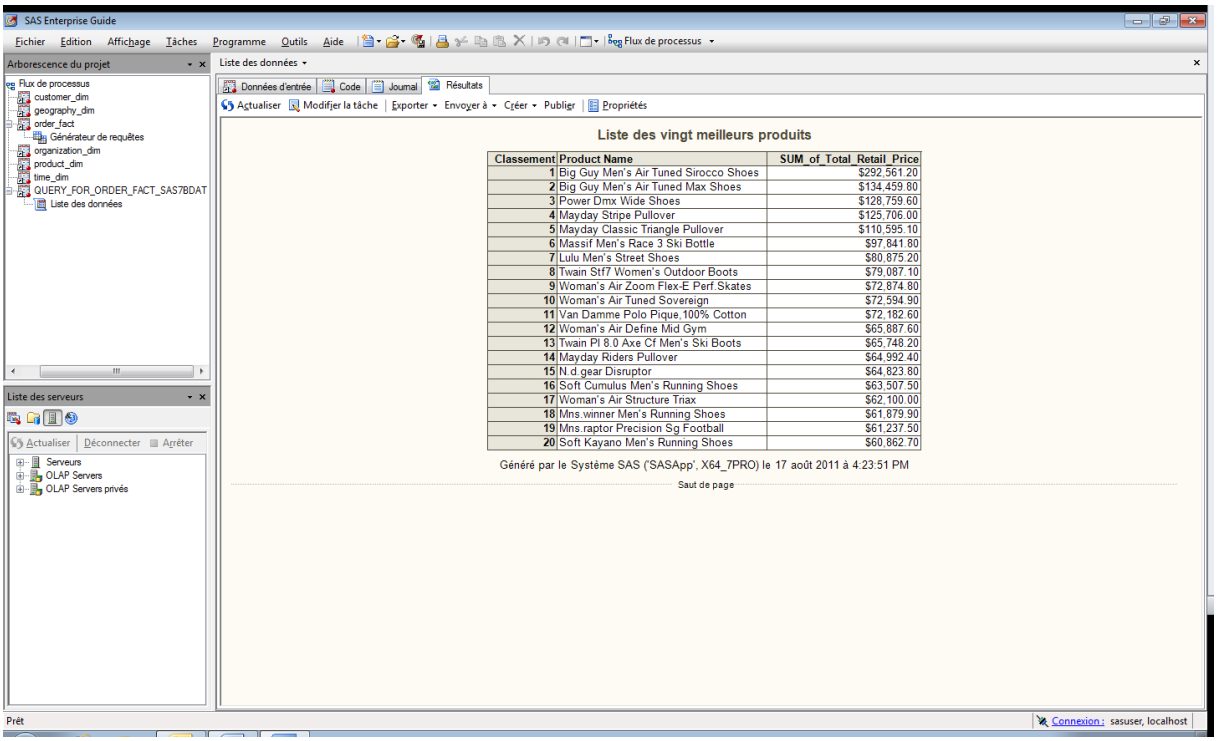
Changer le titre et le pied de page depuis le menu titre de la tâche de création de la liste des données.



Ré-**exécuter** la liste des données

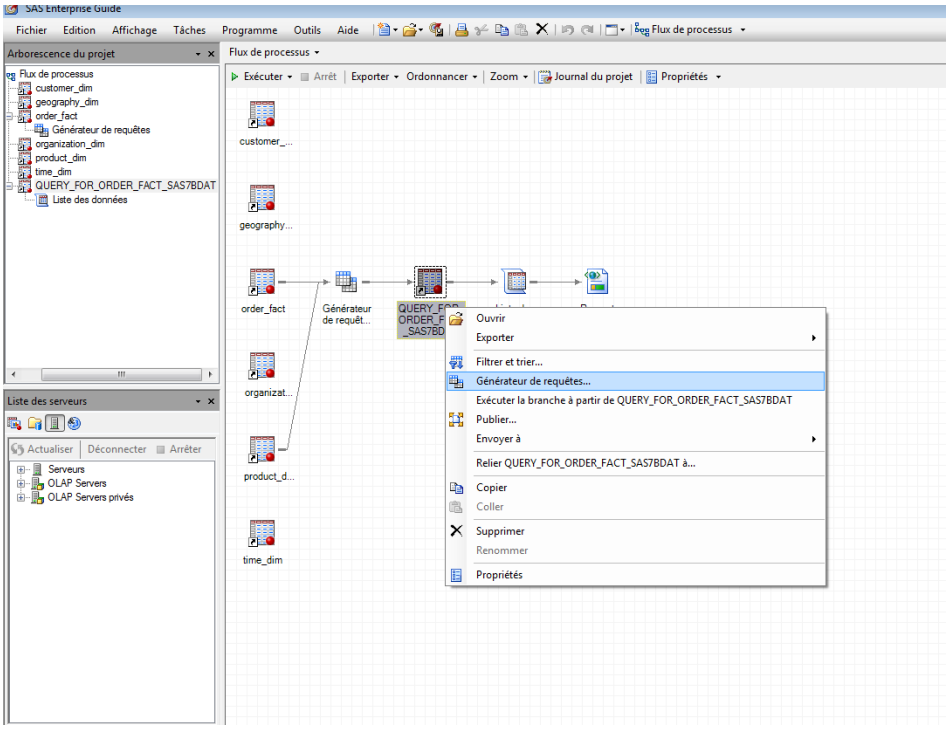


Oui

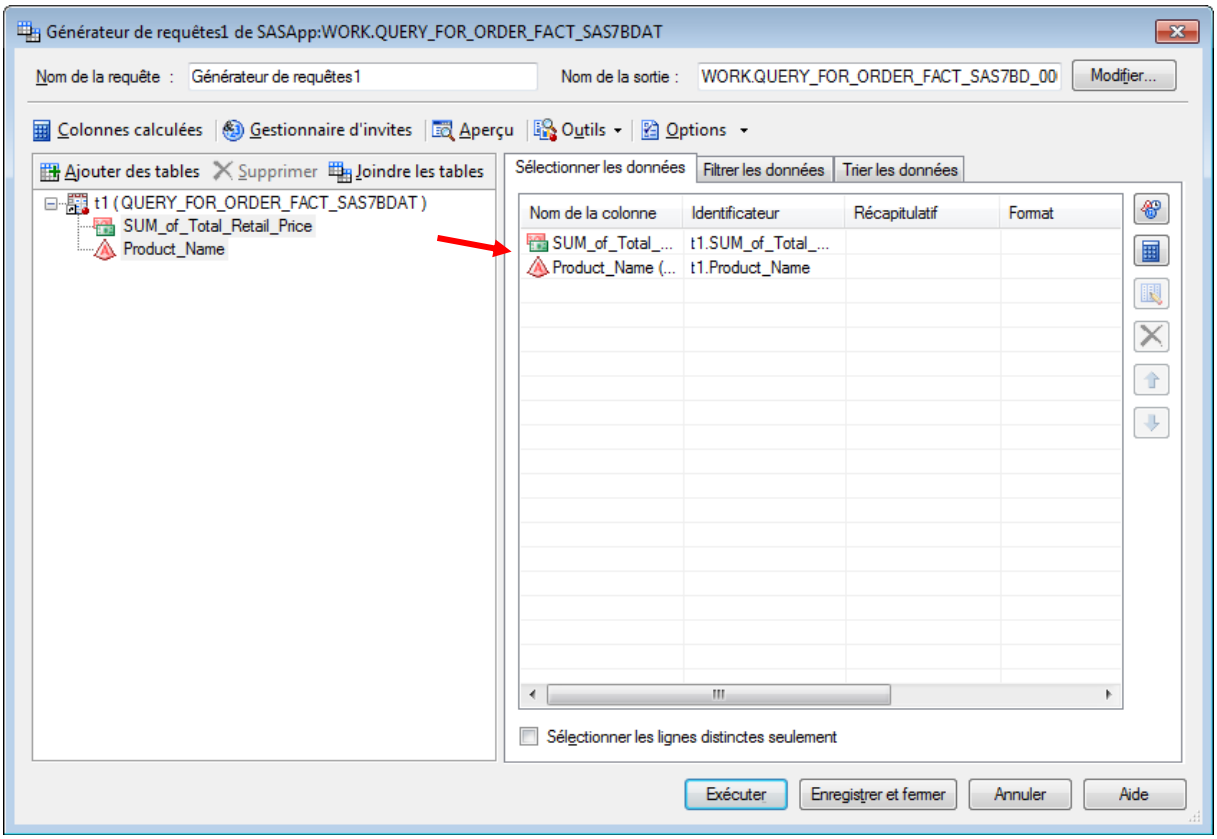


Création d'un Histogramme

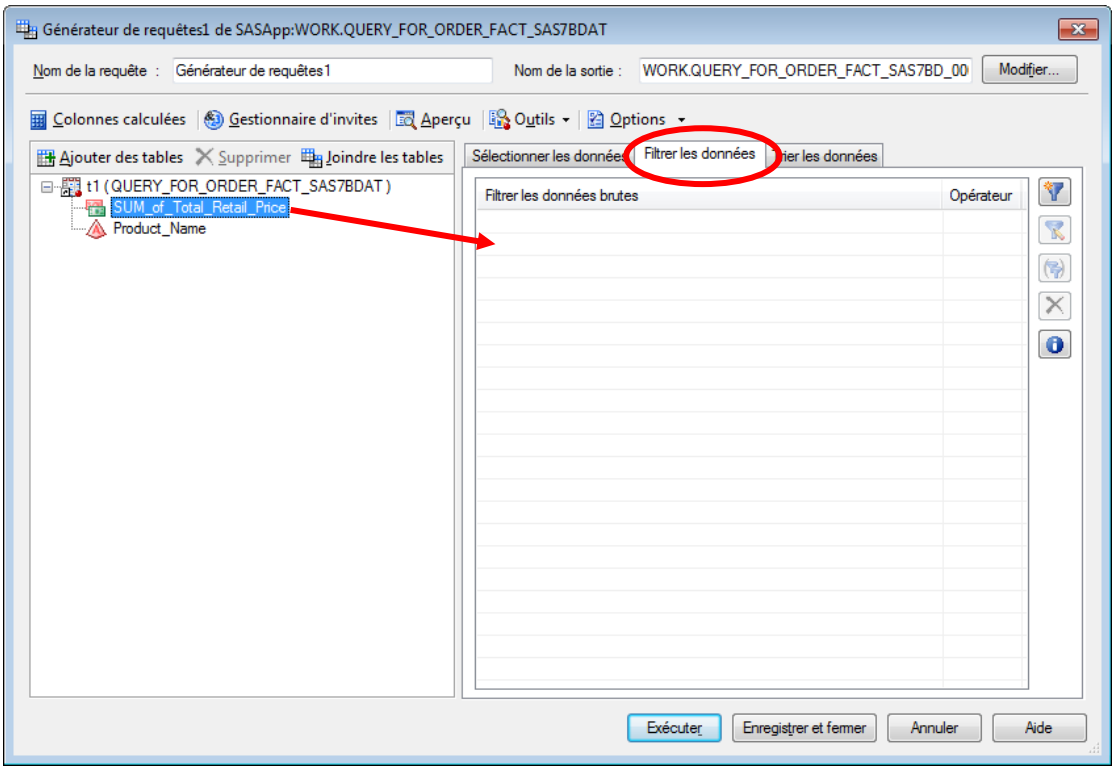
Nous souhaitons maintenant présenter ces mêmes éléments, les produits ayant générés plus de 65 000 $ dans un histogramme.
Pour cela, on peut faire un filtre sur la table de la somme chiffre d'affaires par produit sur ceux ayant générés plus de 65 000 $.



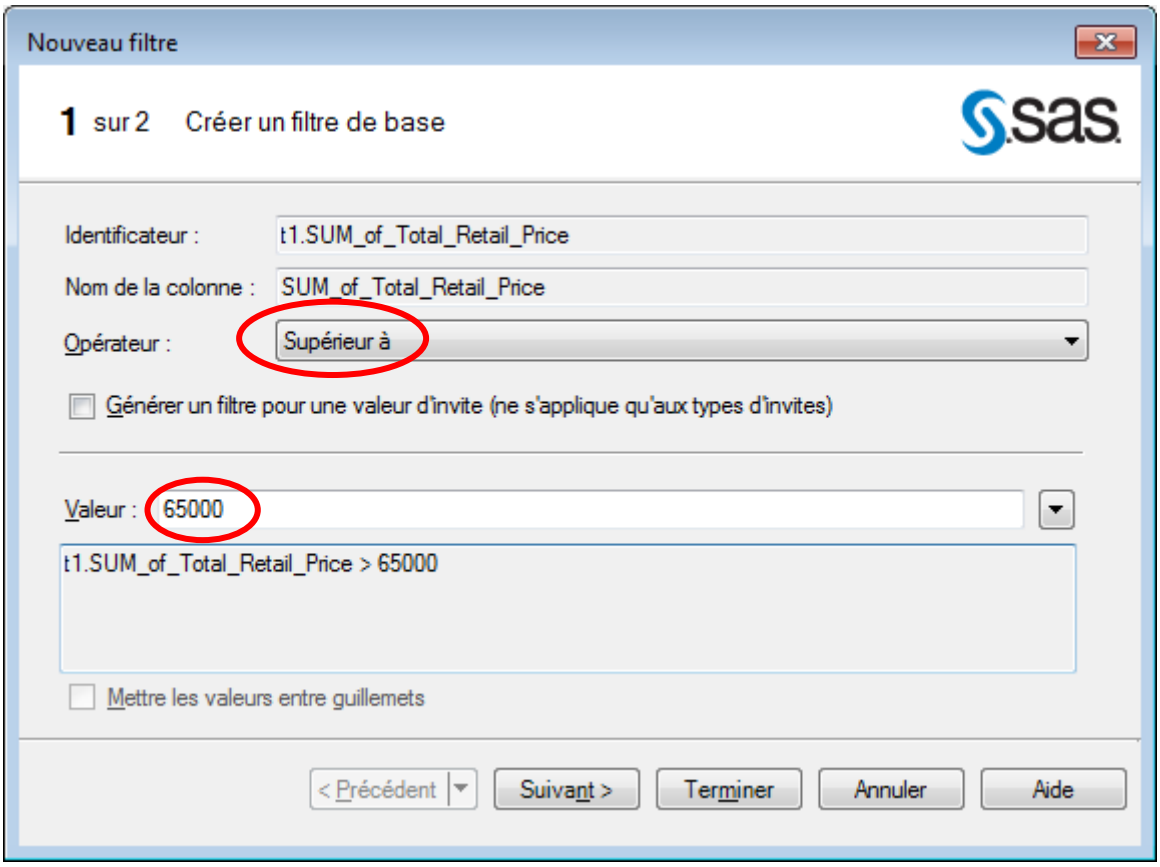
Dans la fenêtre du **flux de processus,** sur la table de la somme chiffre d'affaires par produit, créer une requête.



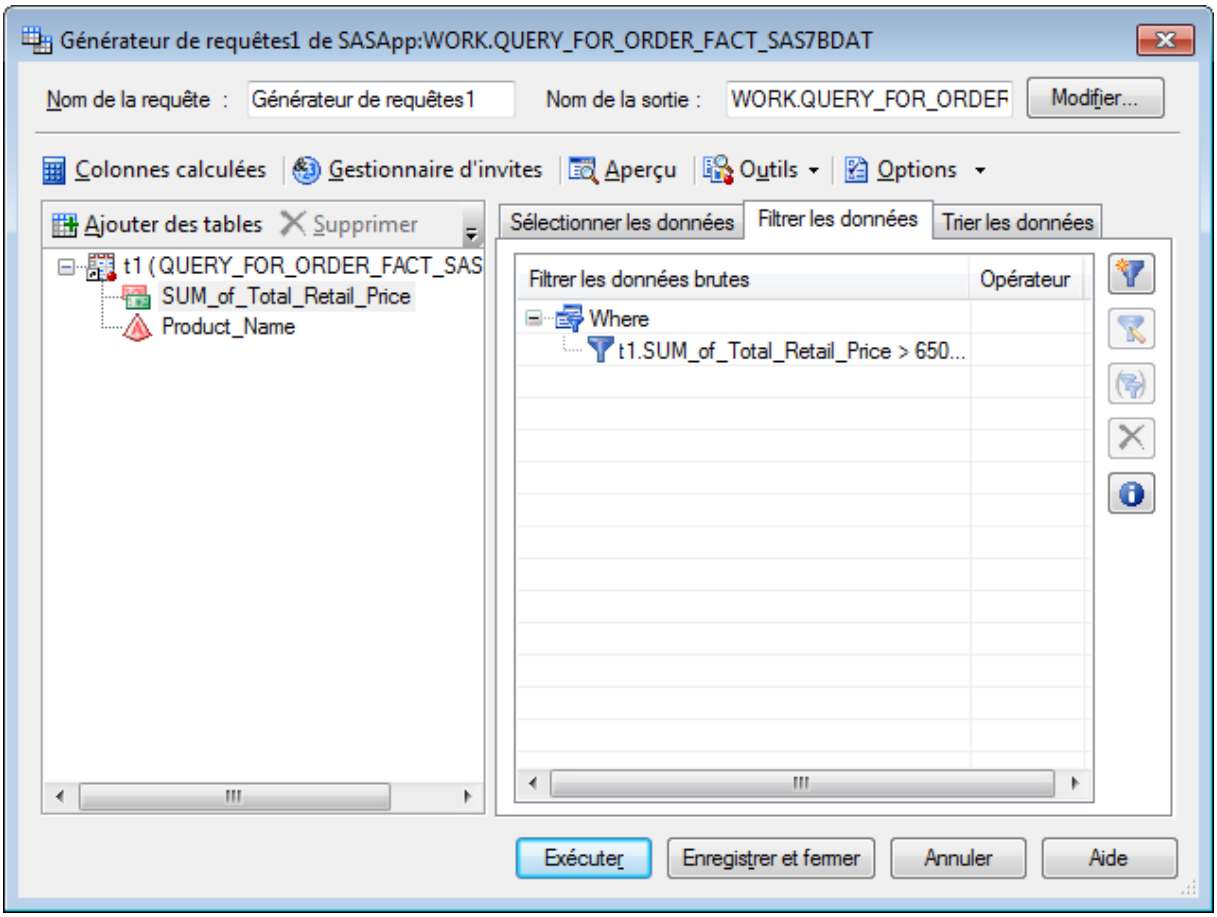
Sélectionner, glisser et lâcher les colonnes de la table source dans les colonnes que vous souhaitez en sortie.



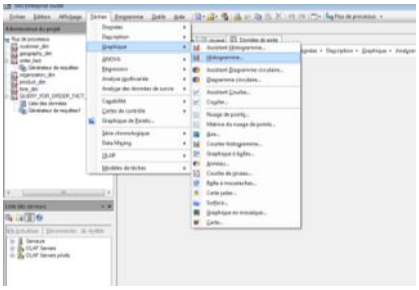
Aller dans l'onglet pour filtrer les données
Sélectionner la colonne SUM_OF_Total_Retail_Price



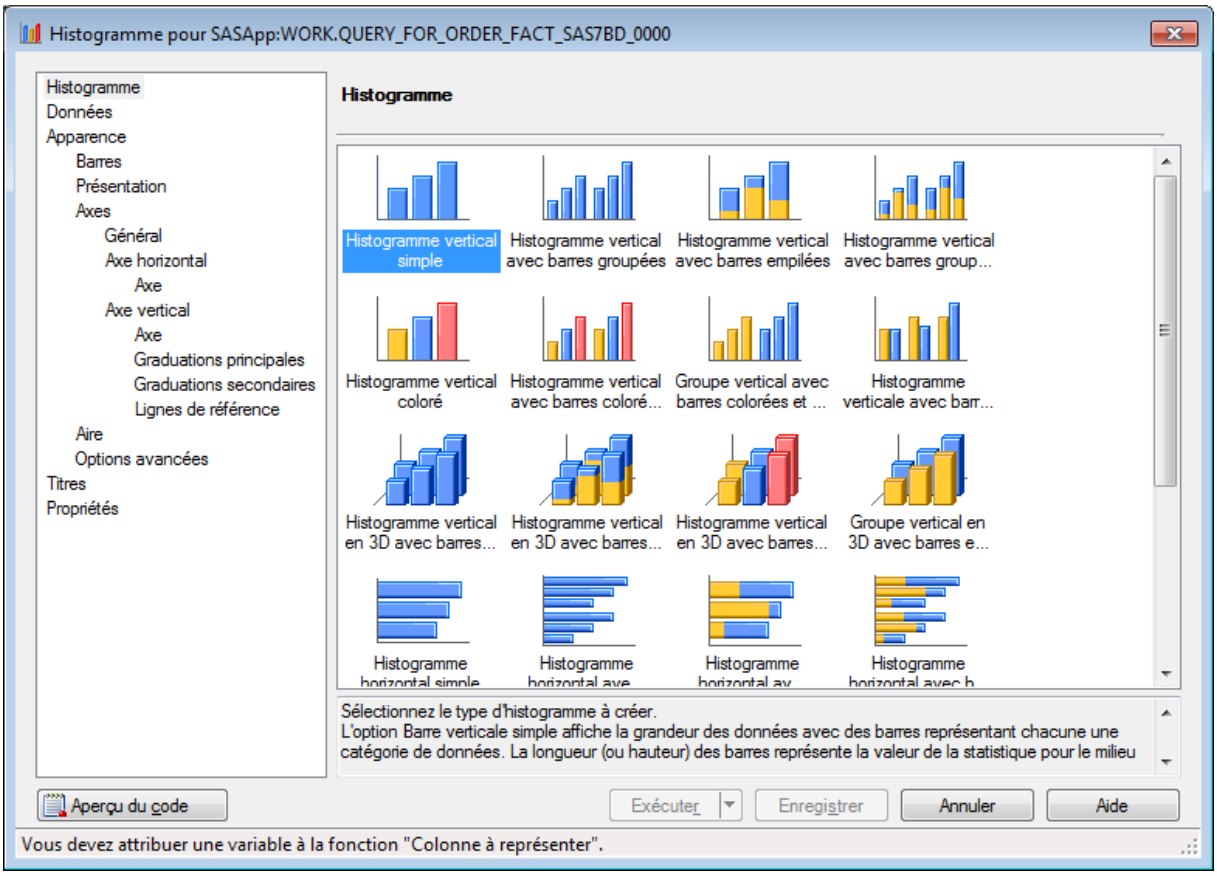
Sélectionner l'opérateur **supérieur à**
Entrer la valeur **65000**
Puis cliquer sur **Terminer**



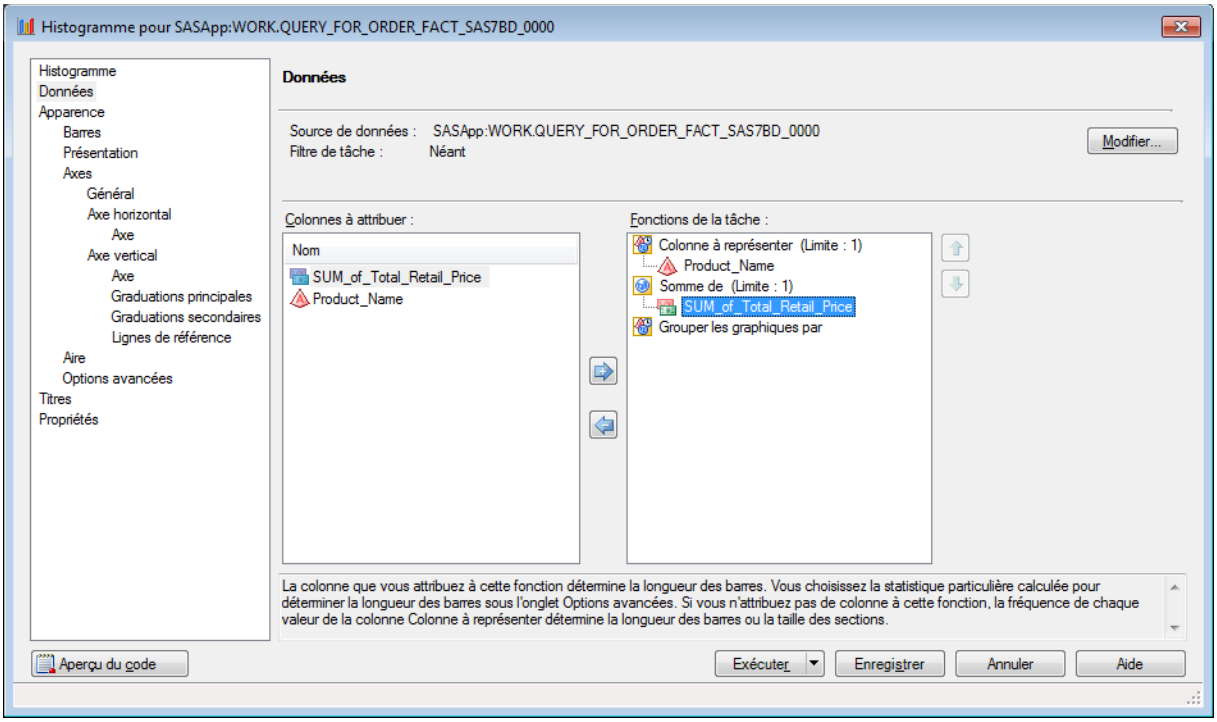
Exécuter



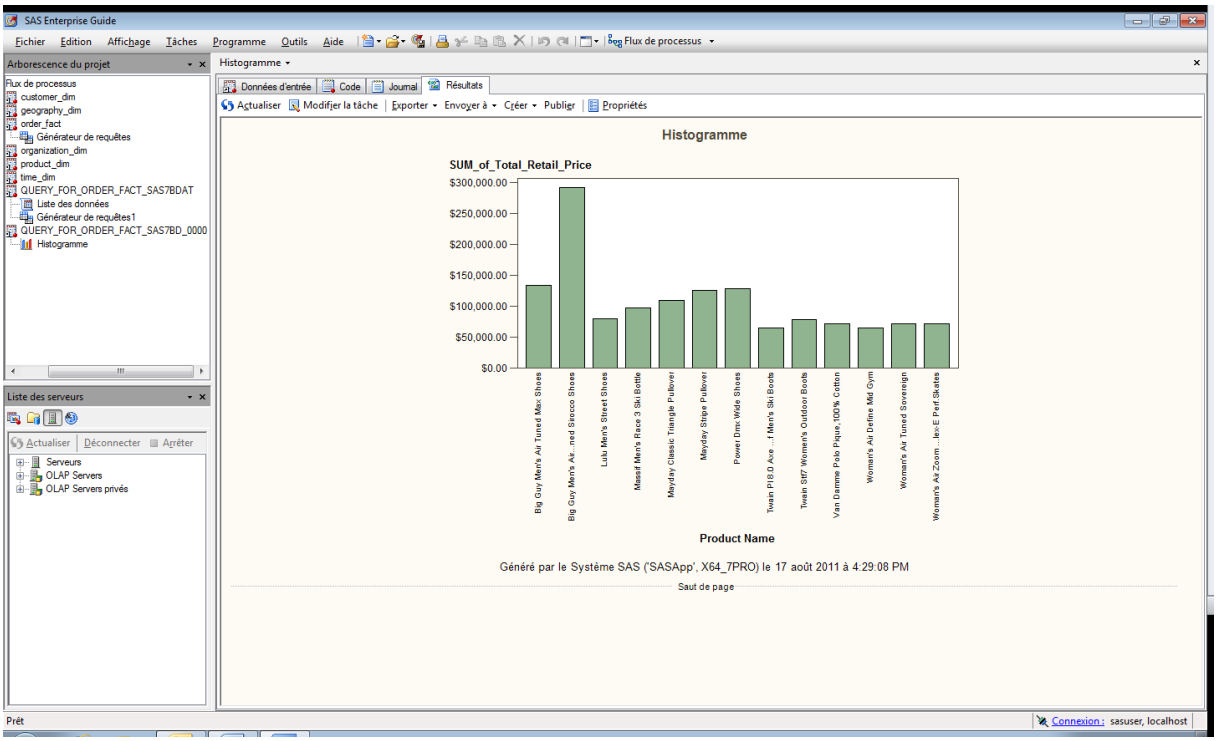
Sélectionner dans le menu des **tâches → Graphique → Histogramme**



Sélectionnez le graphique à barre vertical simple (double cliquer dessus)



Dans **Fonctions de la tâche** fenêtre de gauche,
Glisser-lâcher la colonne **Product_Name** dans la fonction **colonne à représenter**.
Glisser-lâcher la colonne **SUM_OF_Total_Retail_Price** dans la fonction **somme de**.
Cliquez sur **Exécuter.**


SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
Flux de processus
customer_dim
geography_dim
order_fact
Générateur de requêtes
organization_dim
product_dim
time_dim
QUERY_FOR_ORDER_FACT_SAS7BDAT
Liste des données
Générateur de requêtes1
QUERY_FOR_ORDER_FACT_SAS7BD_0000
Histogramme
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Histogramme
Données d'entrée Code Journal Résultats
Actualiser Modifier la tâche Exporter Envoyer à Créer Publier Propriétés
Histogramme
SUM_of_Total_Retail_Price
$300,000.00
$250,000.00
$200,000.00
$150,000.00
$100,000.00
$50,000.00
$0.00
Big Guy Men's Air Tuned Max Shoes
Big Guy Men's Air...ined Sircco Shoes
Lulu Men's Street Shoes
Massf Men's Race 3 Ski Bottle
Mayday Classic Triangle Pullover
Mayday Stripe Pullover
Power Dmc Wide Shoes
Twain P1B.O Ace ...f Men's Ski Boots
Twain Sft7 Women's Outdoor Boots
Van Damme Polo Pique,100% Cotton
Woman's Air Define Mid Gym
Woman's Air Tuned Sovereign
Woman's Air Zoom ...lex-E Perf Skates
Product Name
Généré par le Système SAS ('SASApp', X64_7PRO) le 17 août 2011 à 4:29:08 PM
Saut de page
Prêt
Connexion : sasuser, localhost

Statistiques descriptives

Petit rappel :

Commençons par la statistique descriptive la plus connue et la plus utilisée : la moyenne.

$$Moyenne = \frac{\sum_{i=1}^{n} Yi}{n} = \frac{Y1 + Y2 + Y3 + ... + Yn}{n}$$

Yi est la valeur de y pour l'individu i
n est le nombre d'individu dans la population

La moyenne est souvent notée : $\bar{y}$

Une autre statistique très utilisée est la médiane. C'est la valeur qui sépare l'échantillon en deux parties égales : 50% des observations sont supérieures à la médiane, 50% sont inférieures. Si la base a un nombre impair d'observations, la médiane est la valeur du milieu, si non, c'est la moyenne des deux valeurs du milieu.

L'étendue est la différence entre la valeur maximum et la valeur minimum.

La variance, notée σ^2, mesure la dispersion des données autour de la moyenne.

L'écart type, noté σ, est la racine carrée de la variance.

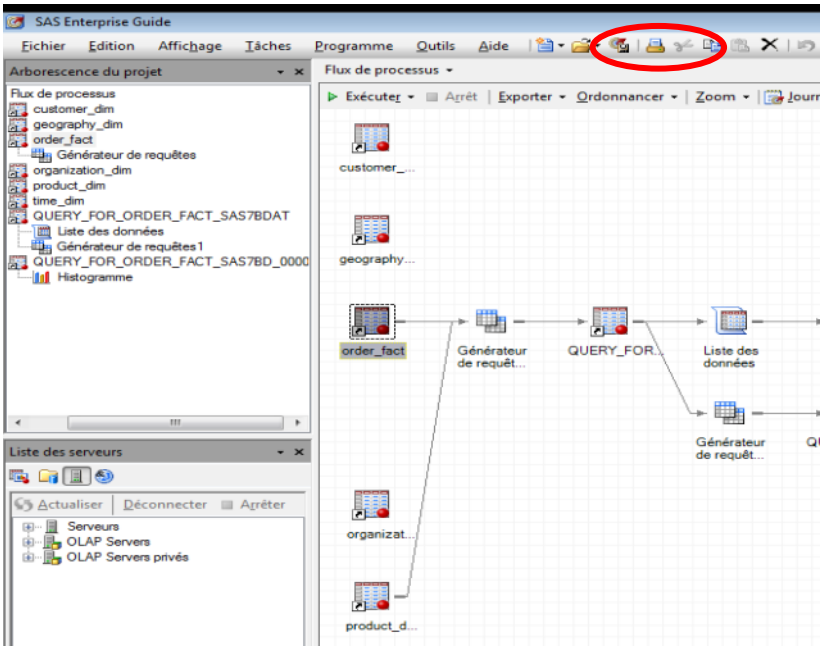
Observation	donnée	écart	Ecart²
1	3	-2	4
2	8	3	9
3	4	-1	1
Somme	15	0	14
Moyenne	5	0	σ^2 =14/3
			$\sigma = \sqrt{14/3} = 2.16$

Nous avons ici trois observations : '3', '8' et '4'
La moyenne de ces trois observations est 5.
La variance est de 14/3
L'écart type et de 2.16

Création d'un rapport de statistique simple :

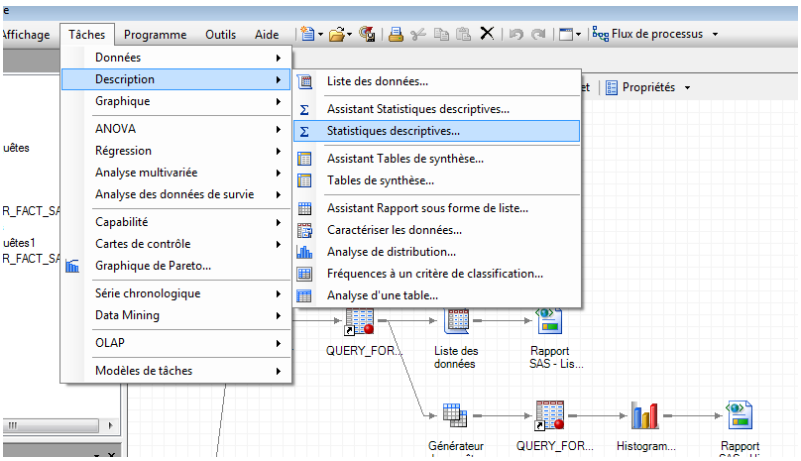
Objectif :

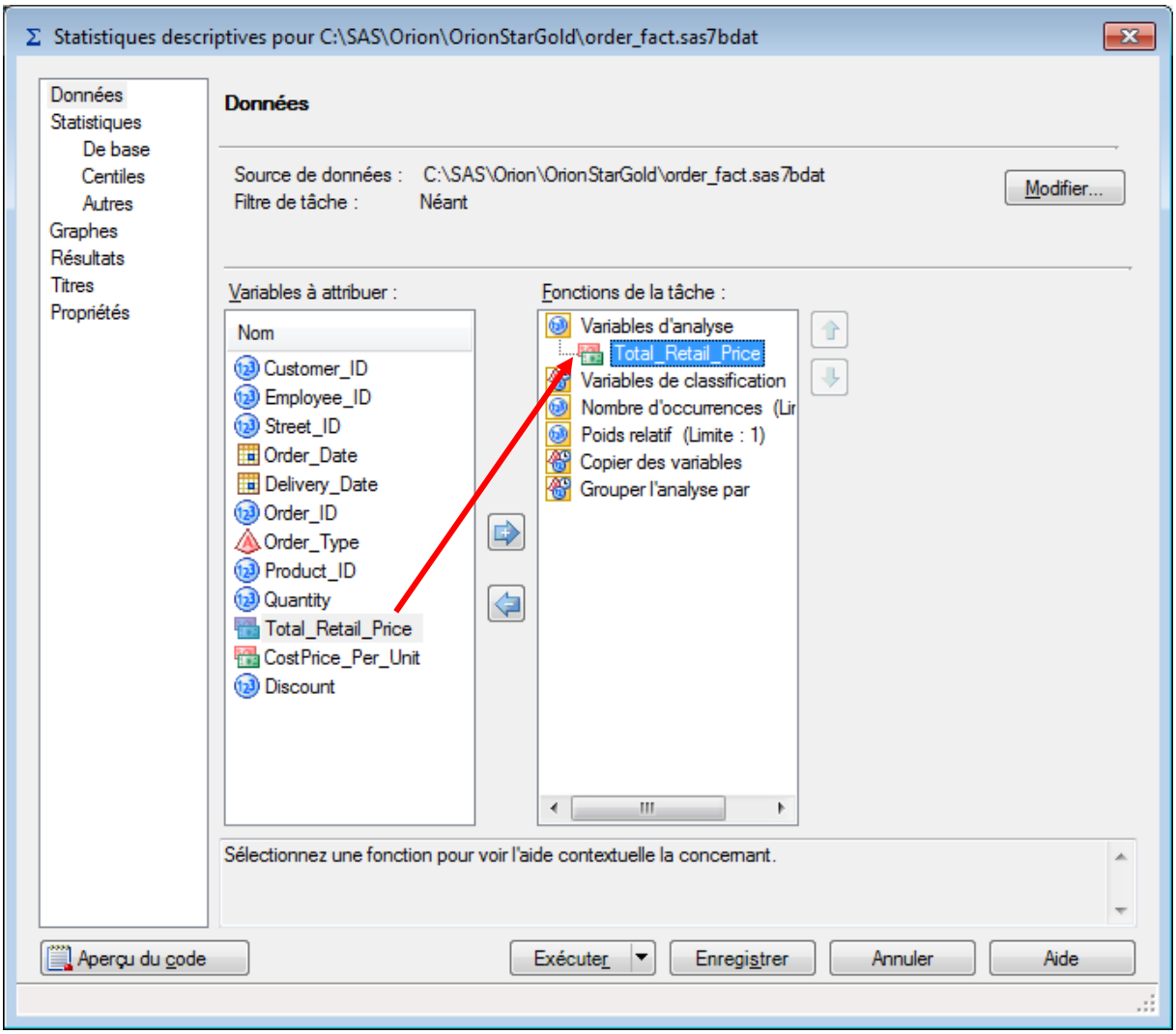
Création d'un rapport avec les statistiques descriptives de base du chiffre d'affaires pour répondre à la question : Quels sont la moyenne et l'écart type du chiffre d'affaires ?



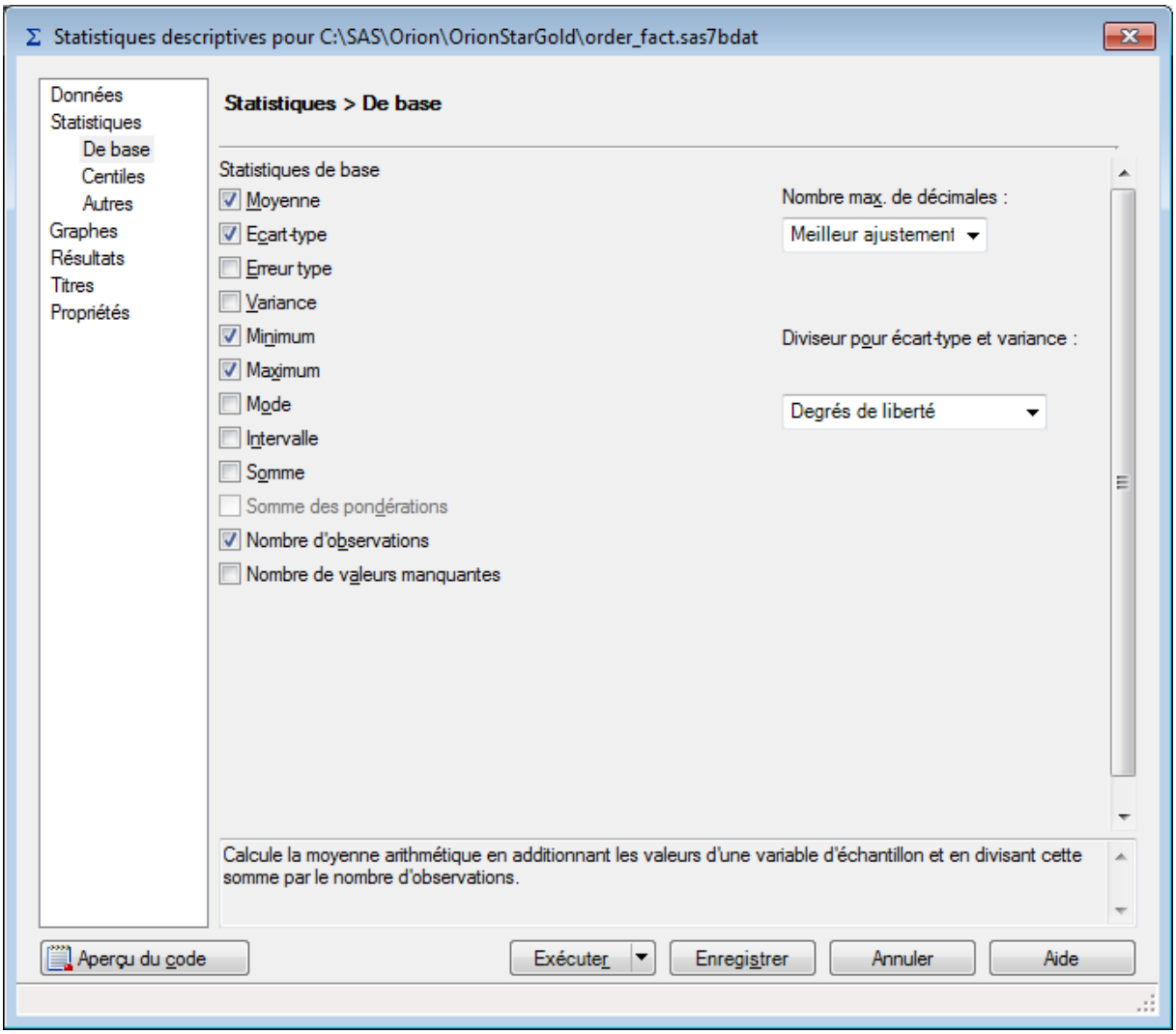
Aller dans la fenêtre de **flux de processus**.

Sélectionner la table **Order_Fact**, (il faut qu'elle soit grisée) puis dans le menu des **tâches** → **Description**, allez dans les **statistiques descriptives.**



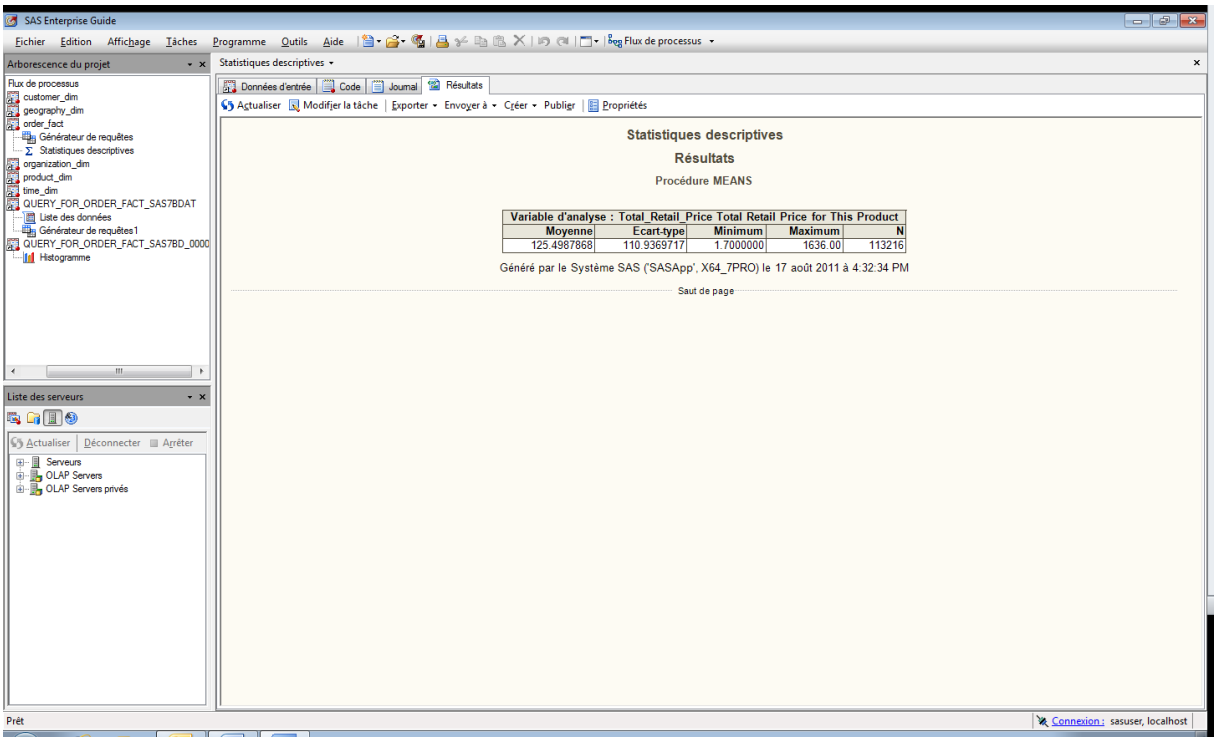


Glisser-lâcher la colonne du chiffre d'affaires dans la fonction **Variable d'analyse**.



En sélectionnant les statistiques de base, vous pouvez sélectionner ou désélectionner les statistiques et de même pour les centiles et les graphes.

Cliquer sur **Exécuter.**


SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
Flux de processus
customer_dim
geography_dim
order_fact
Générateur de requêtes
Statistiques descriptives
organization_dim
product_dim
time_dim
QUERY_FOR_ORDER_FACT_SAS7BDAT
Liste des données
Générateur de requêtes1
QUERY_FOR_ORDER_FACT_SAS7BD_0000
Histogramme
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Statistiques descriptives
Données d'entrée Code Journal Résultats
Actualiser Modifier la tâche Exporter Envoyer à Créer Publier Propriétés
Statistiques descriptives
Résultats
Procédure MEANS
Variable d'analyse : Total_Retail_Price Total Retail Price for This Product
Moyenne Ecart-type Minimum Maximum N
125.4987868 110.9369717 1.7000000 1636.00 113216
Généré par le Système SAS ('SASApp', X64_7PRO) le 17 août 2011 à 4:32:34 PM
Saut de page
Prêt
Connexion : sasuser, localhost

Test d'hypothèse

En statistiques,
L'hypothèse initiale, notée H0 est appelée hypothèse nulle. Dans la plupart des analyses statistiques, on vérifie si cette hypothèse est acceptable ou non. L'hypothèse notée H1 est appelée hypothèse alternative.
Le niveau de significativité est habituellement noté α, c'est la probabilité d'erreur de type 1. Généralement, α est fixé à 1%, 5% voir 10%. Cela correspond à 0.01, 0.05 ou 0.1. Ces analyses ont donc pour but de vérifier ou d'infirmer si l'hypothèse nulle est acceptable avec une probabilité de 99%, 95% ou 90%.

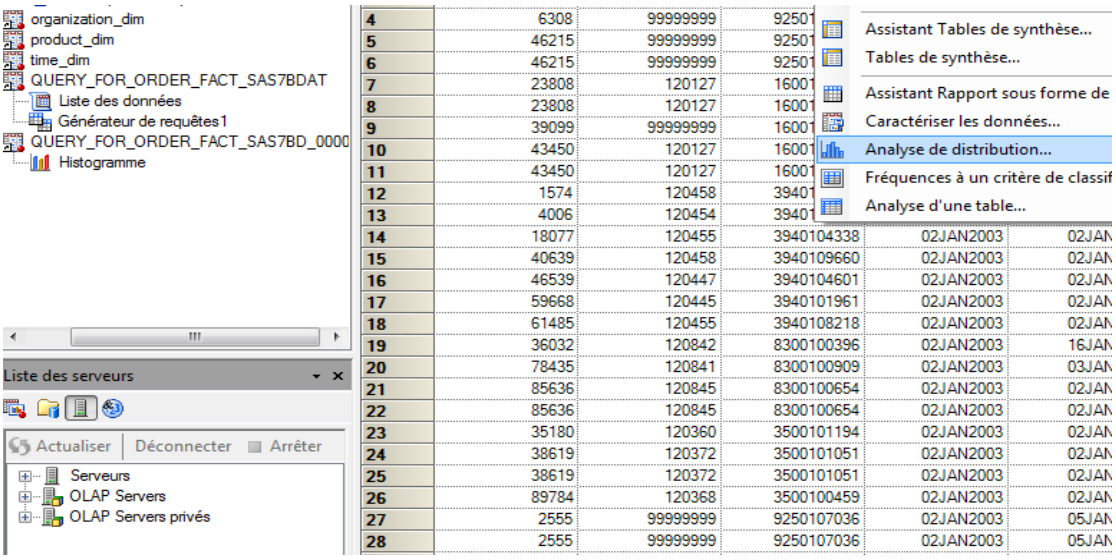
Règle de décision
On rejette l'hypothèse nulle si la p-value est inférieure à α (en d'autres termes, l'évènement est rare)
On ne peut pas rejeter l'hypothèse nulle si la p-value est supérieure à α.

Ce qu'il faut retenir :
p-value < 0.05 → à 95%, on rejette l'hypothèse nulle
p-value > 0.05 → à 95%, on ne rejette pas l'hypothèse nulle

Normalité d'une distribution

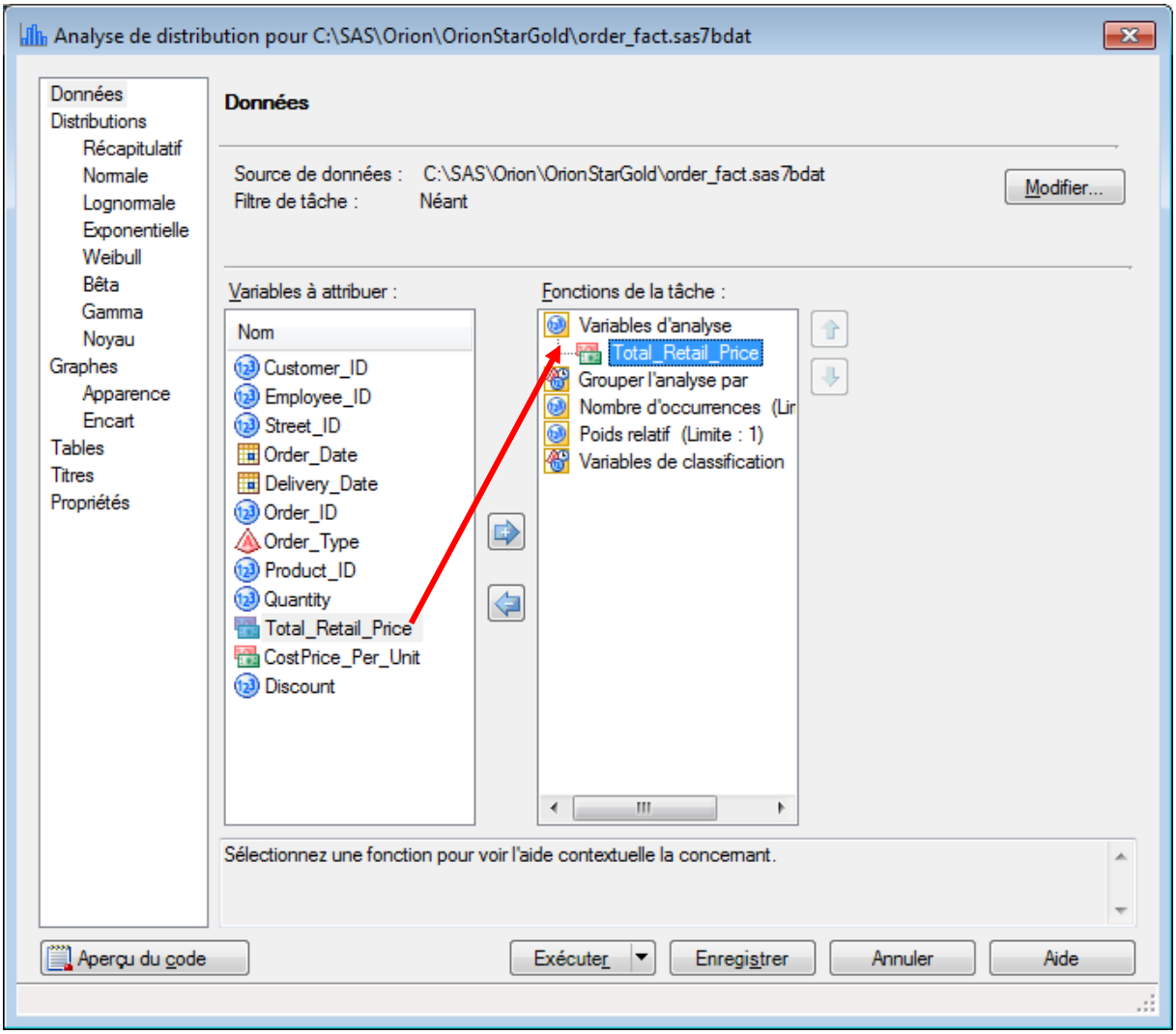
De nombreuses statistiques sont interprétables sous l'hypothèse que les valeurs soient normalement distribuées. Beaucoup d'analyses commencent donc par tester la normalité d'une distribution.

Dans notre cas, nous allons analyser la distribution du chiffre d'affaires.

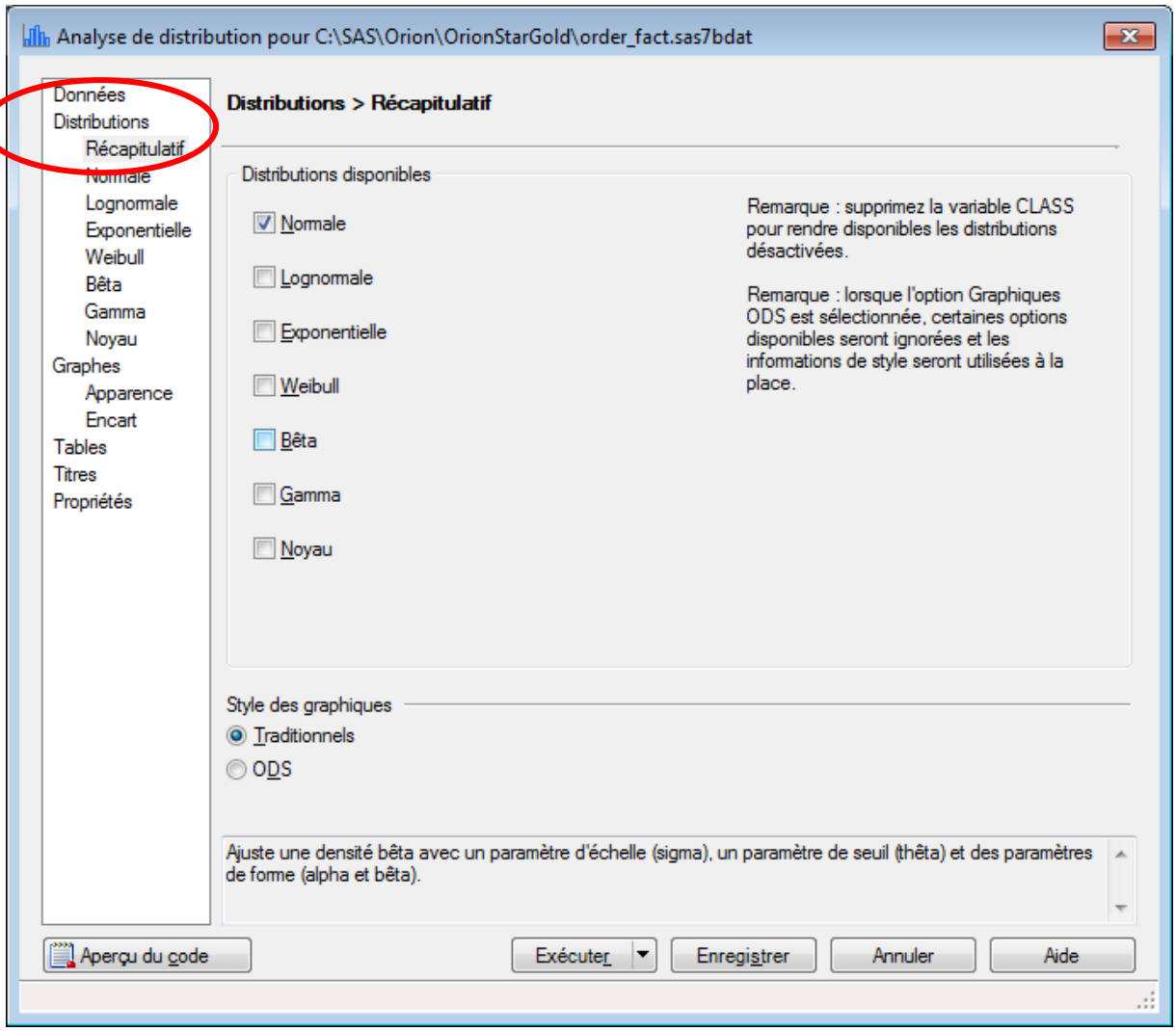


4	6308	99999999	92501		
5	46215	99999999	92501		
6	46215	99999999	92501		
7	23808	120127	16001		
8	23808	120127	16001		
9	39099	99999999	16001		
10	43450	120127	16001		
11	43450	120127	16001		
12	1574	120458	3940		
13	4006	120454	3940		
14	18077	120455	3940104338	02JAN2003	02JAN
15	40639	120458	3940109660	02JAN2003	02JAN
16	46539	120447	3940104601	02JAN2003	02JAN
17	59668	120445	3940101961	02JAN2003	02JAN
18	61485	120455	3940108218	02JAN2003	02JAN
19	36032	120842	8300100396	02JAN2003	16JAN
20	78435	120841	8300100909	02JAN2003	03JAN
21	85636	120845	8300100654	02JAN2003	02JAN
22	85636	120845	8300100654	02JAN2003	02JAN
23	35180	120360	3500101194	02JAN2003	02JAN
24	38619	120372	3500101051	02JAN2003	02JAN
25	38619	120372	3500101051	02JAN2003	02JAN
26	89784	120368	3500100459	02JAN2003	02JAN
27	2555	99999999	9250107036	02JAN2003	05JAN
28	2555	99999999	9250107036	02JAN2003	05JAN

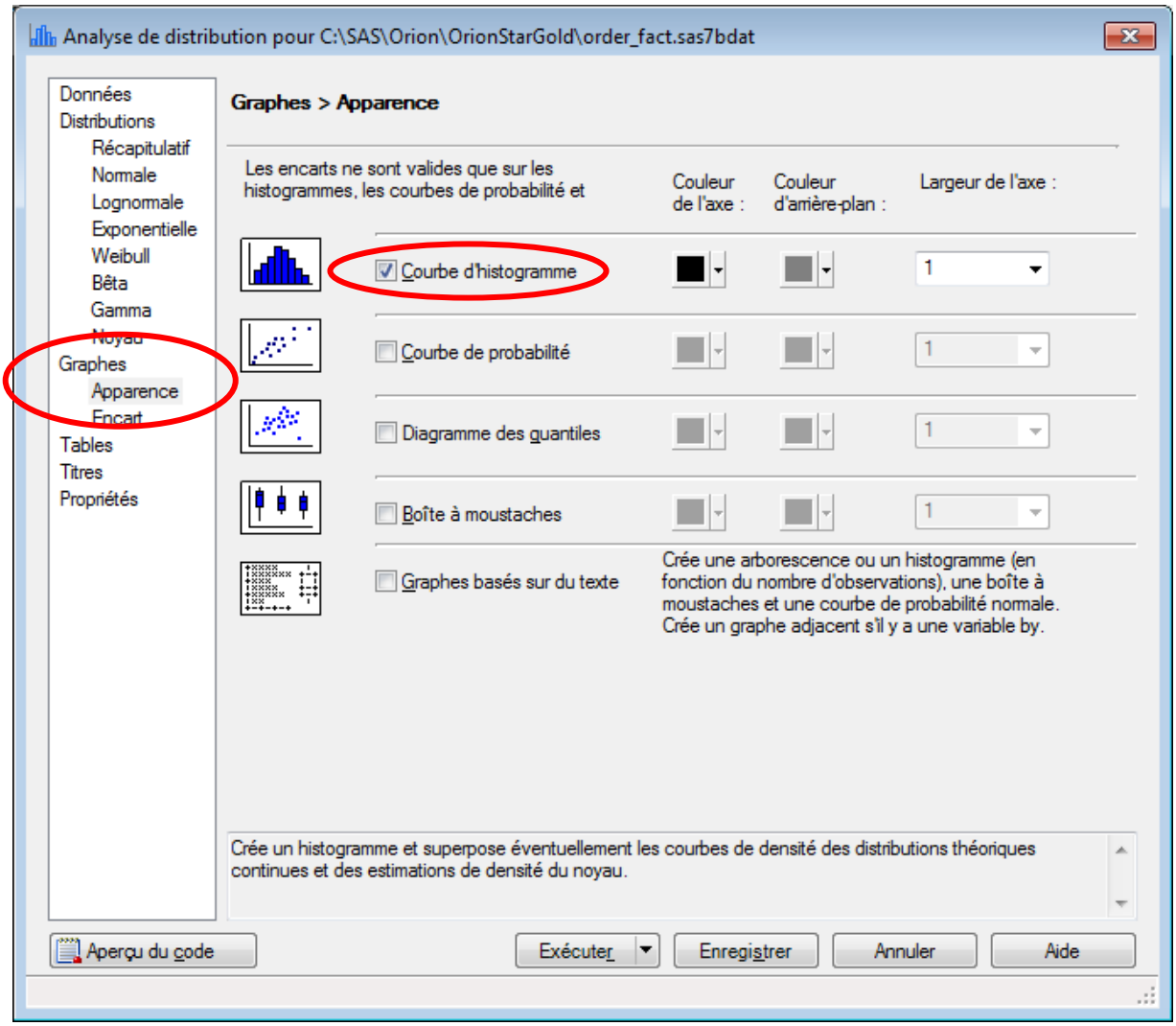
Aller dans la fenêtre de **flux de processus.**
Sélectionner la table **Order_Fact**, puis dans le menu des **tâches → Description** → sélectionner **Analyse de distribution**



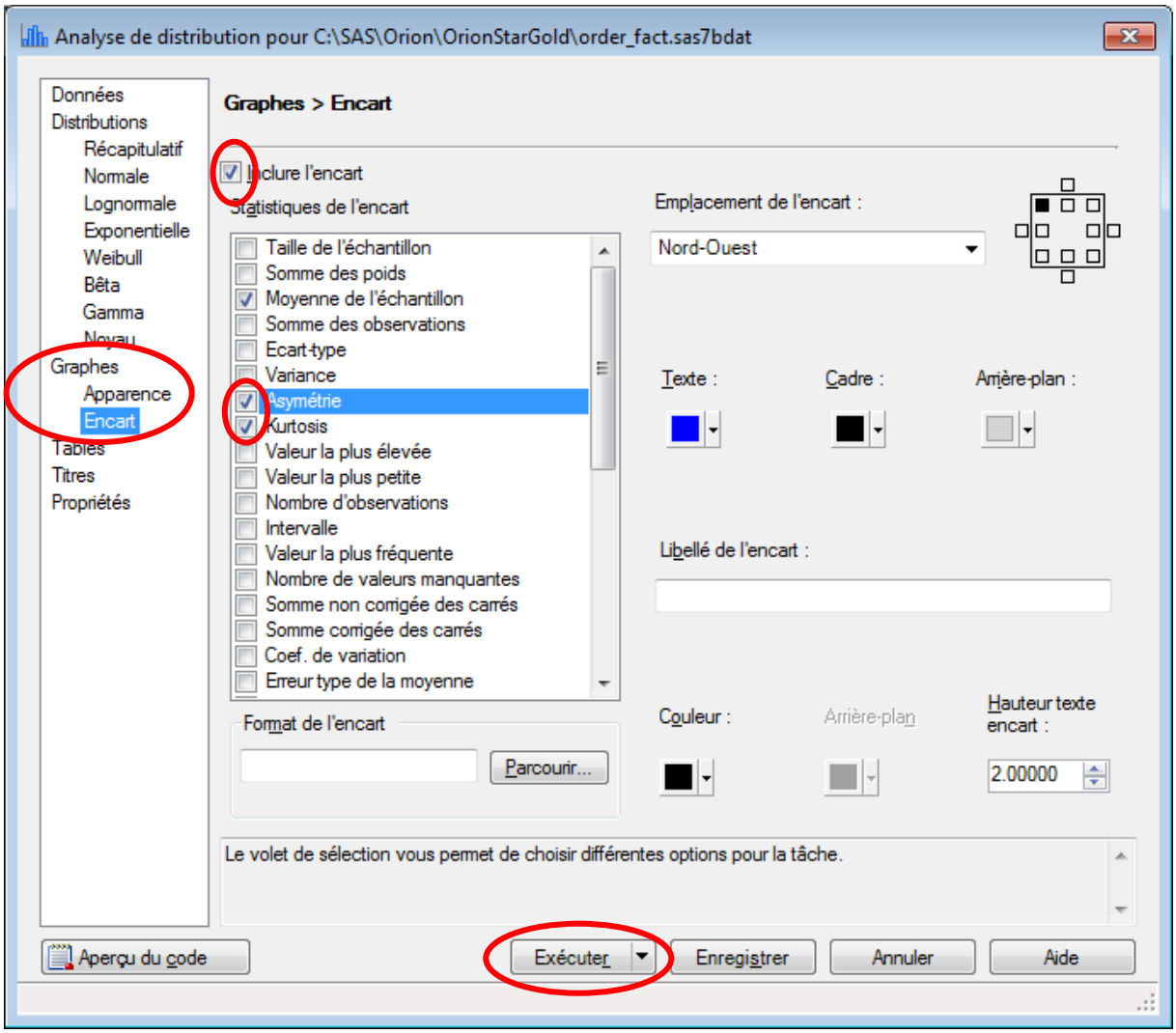
Affecter le chiffre d'affaires comme variable à analyser.



Dans l'onglet Distributions, sélectionner la loi Normale



Dans l'onglet Graphes, sélectionner la courbe d'histogramme,

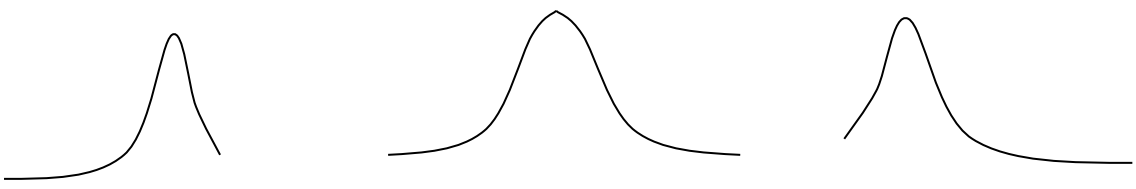


Dans l'onglet Encart, sélectionner le coefficient d'asymétrie (Skewness) et celui d'aplanissement (Kurtosis).

Cliquer sur Exécuter.

Interprétation des deux coefficients de caractéristiques de forme de distribution :

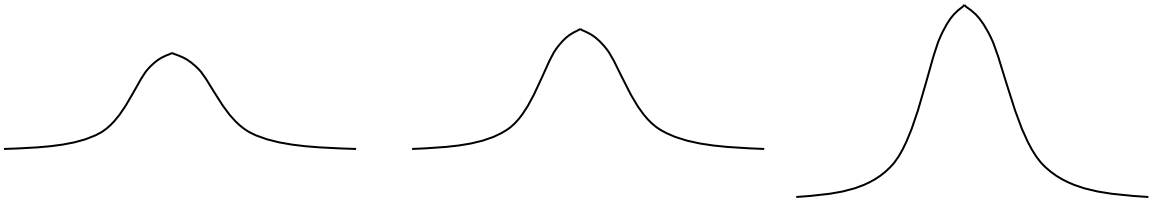
Skewness :



Skewness < 0 Skewness = 0 Skewness > 0

- Si le Skewness est inférieur à zéro, la distribution est asymétrique vers la gauche.
- Si le Skewness est égale à zéro, la distribution est symétrique.
- Si le Skewness est supérieur à zéro, la distribution est asymétrique vers la droite.

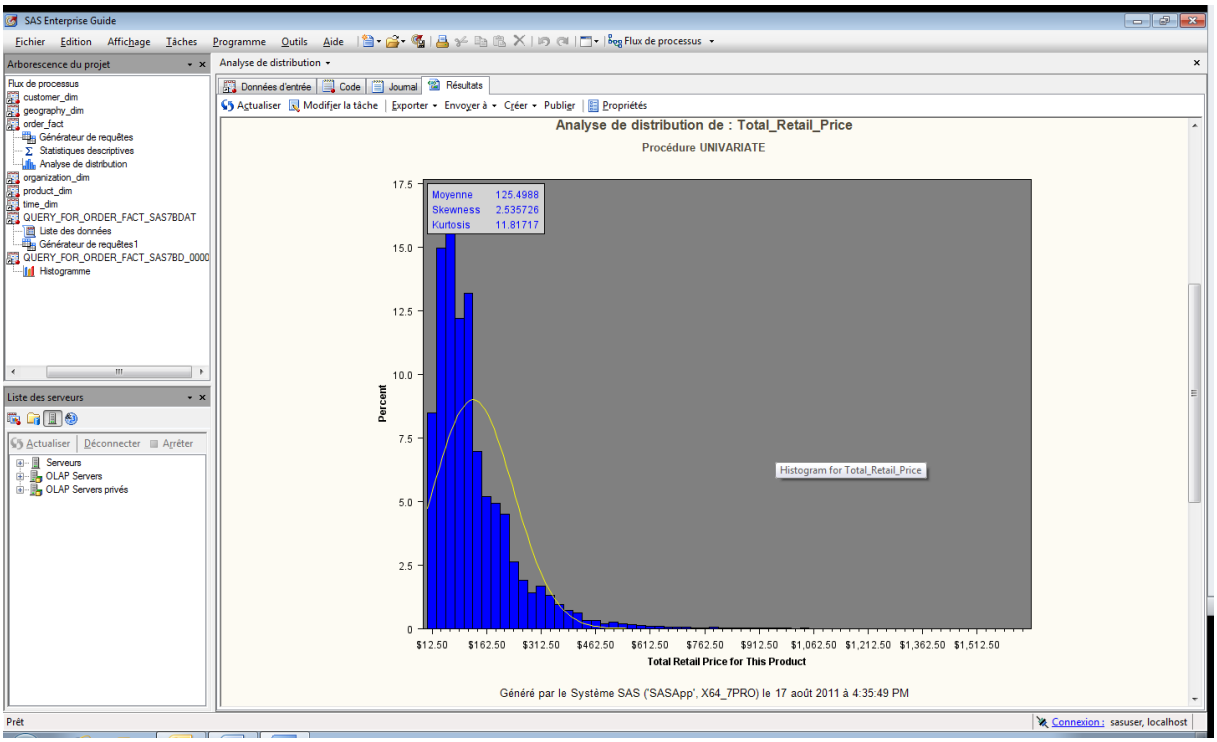
Kurtosis :



Kurtosis < 0 Kurtosis = 0 Kurtosis > 0

- Si le Kurtosis est inférieur à 0, la distribution est plus aplatie que celle de Gauss.
- Si le Kurtosis est égale à 0, la distribution a la même concentration (aplatissement) que celle de Gauss.
- Si le Kurtosis est supérieur à 0, la distribution est plus concentrée que celle de Gauss.

Remarque : Par abus de langage, on parle ici de l'excess Kurtosis qui le Kurtosis -3.



Test du chi²

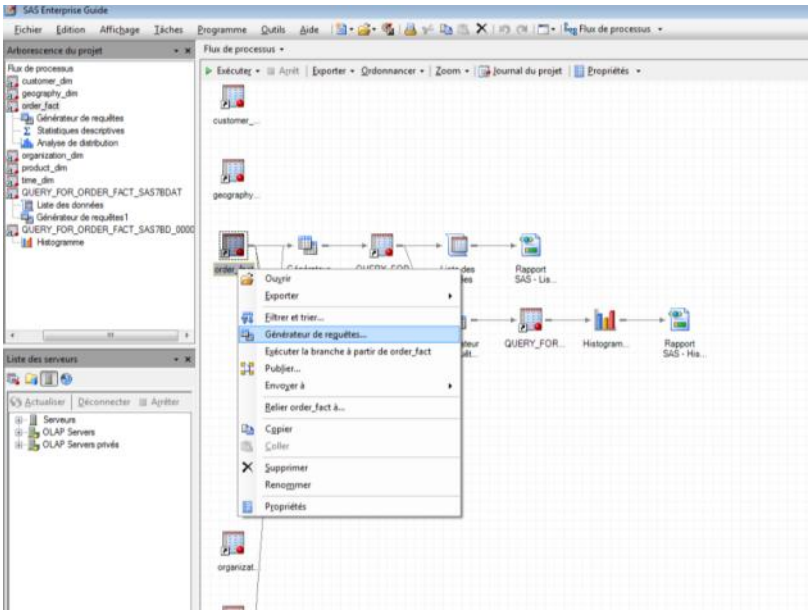
Ce test sert notamment à vérifier la liaison existante entre deux variables qualitatives. Il est **très** utilisé.

ATTENTION: Les effectifs des cellules doivent être supérieurs à 5 pour utiliser les résultats de ce test

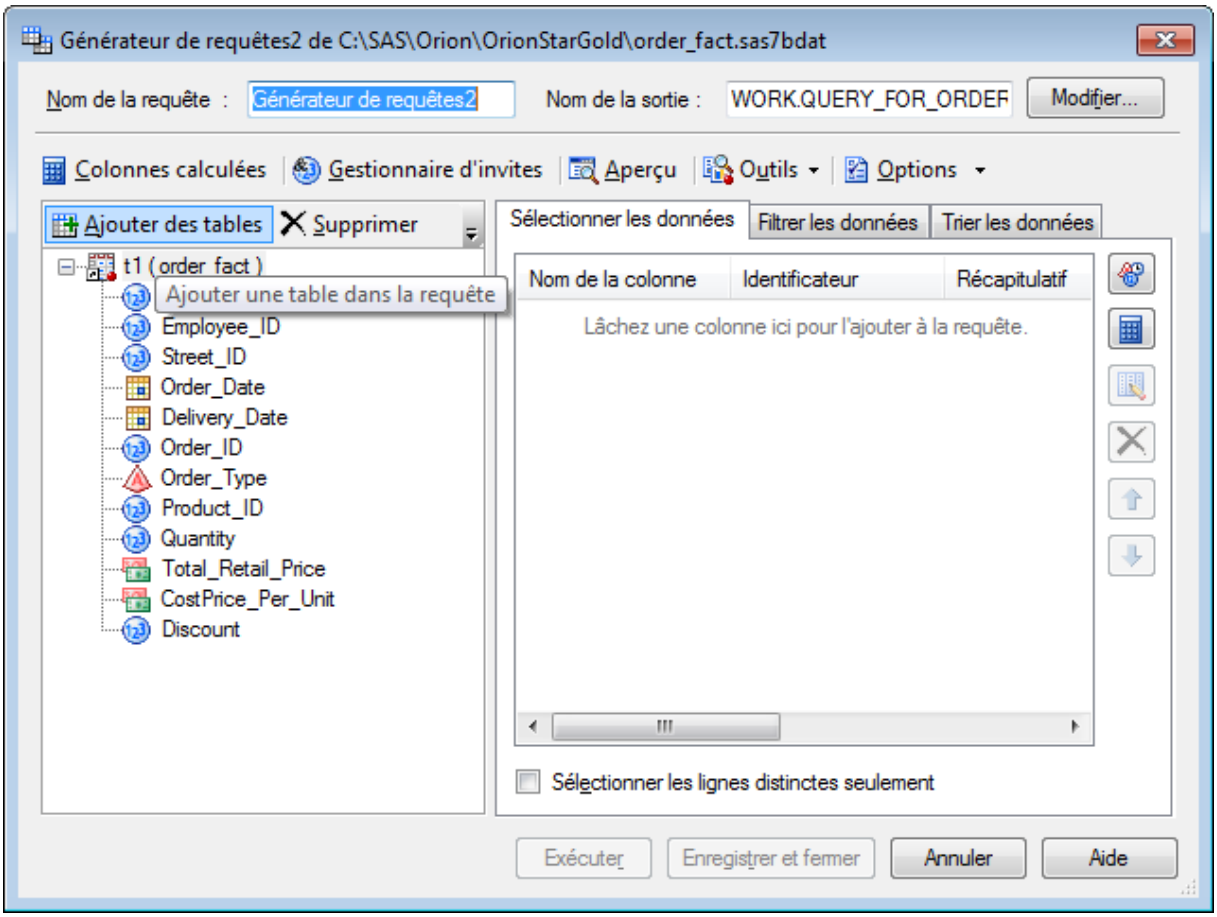
Question :

Y a-t-il indépendance entre le sexe des commerciaux et celui des clients ?

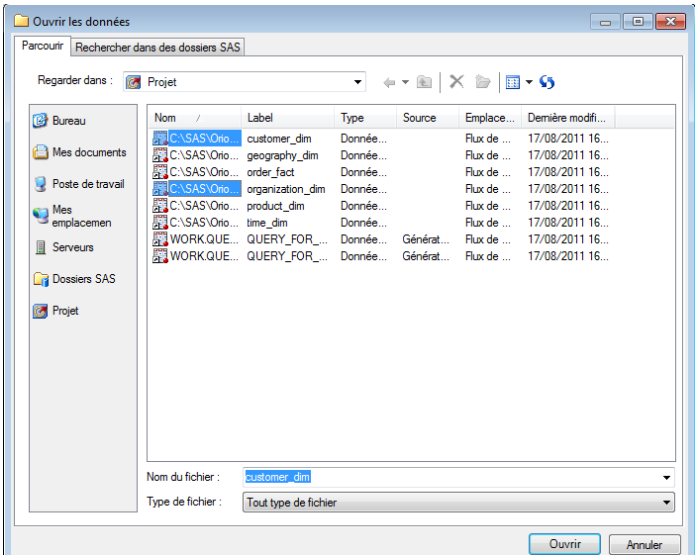
Pour cela, nous allons créer la table de la liste des commandes avec le sexe du client et celui du commercial.
Remarque : Il faut exclure de l'analyse toutes les ventes par Internet ou par catalogue dans lesquelles on ne connait pas le sexe du commercial.



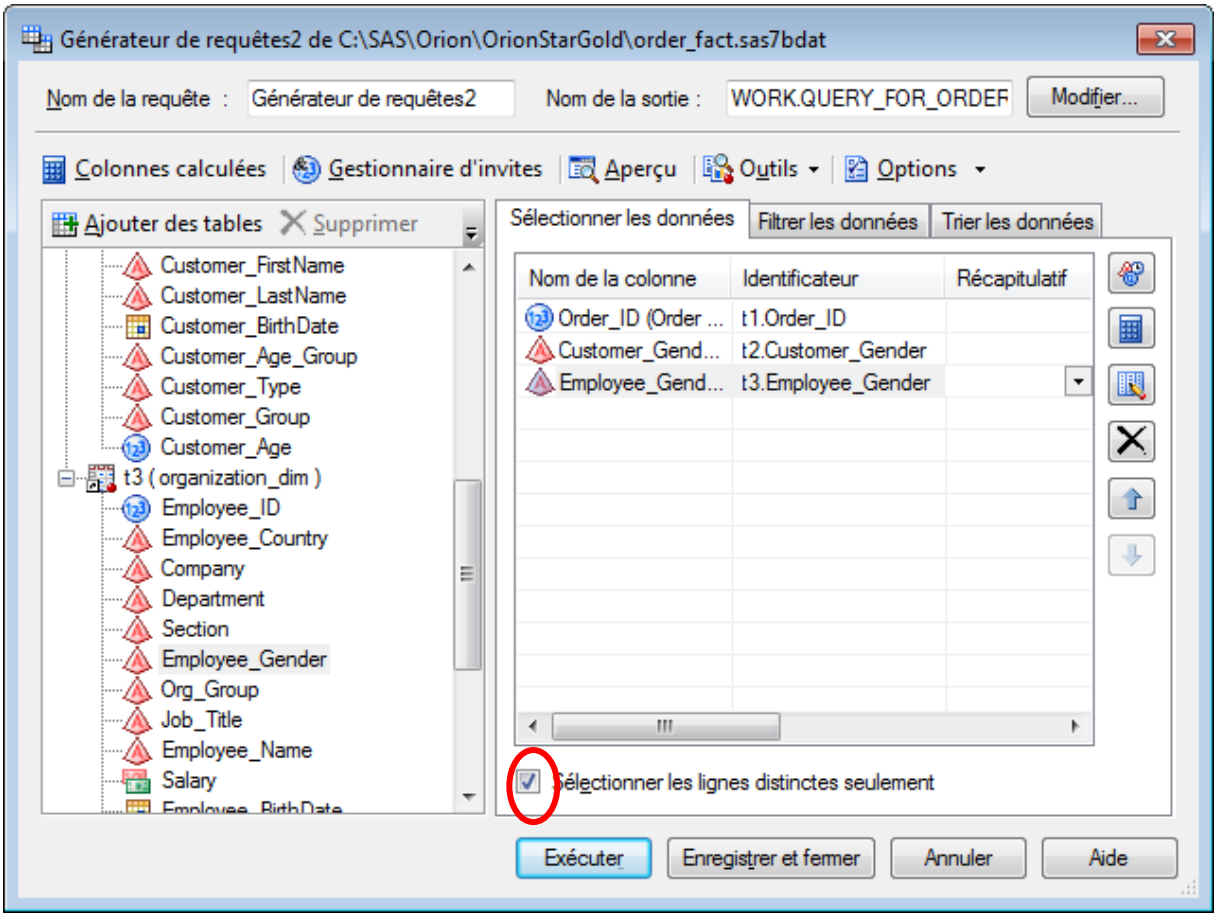
Revenir dans la fenêtre du flux de processus
Créer une requête sur la table **Order_Fact**



Ajouter les tables de dimension client et collaborateur qui se trouvent déjà dans le projet.

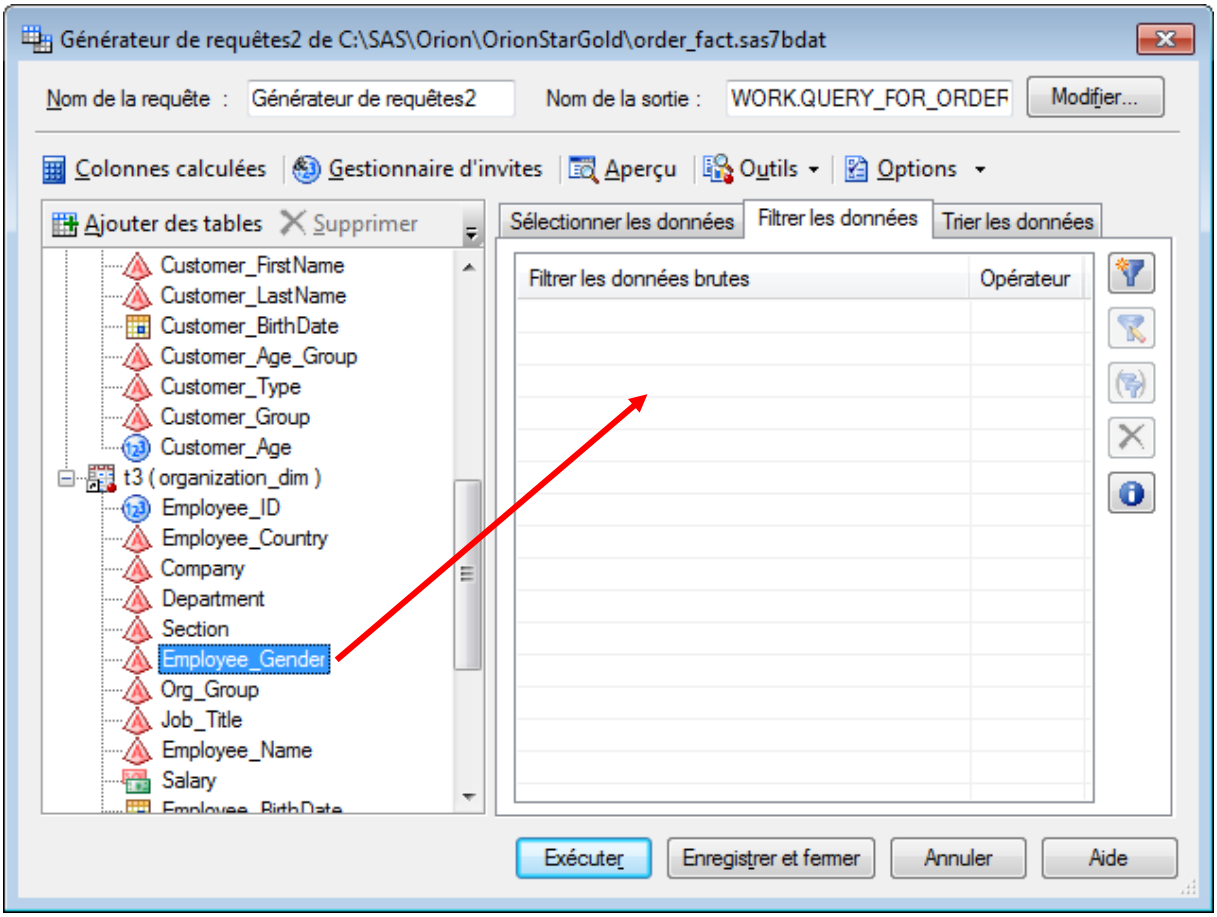


OK



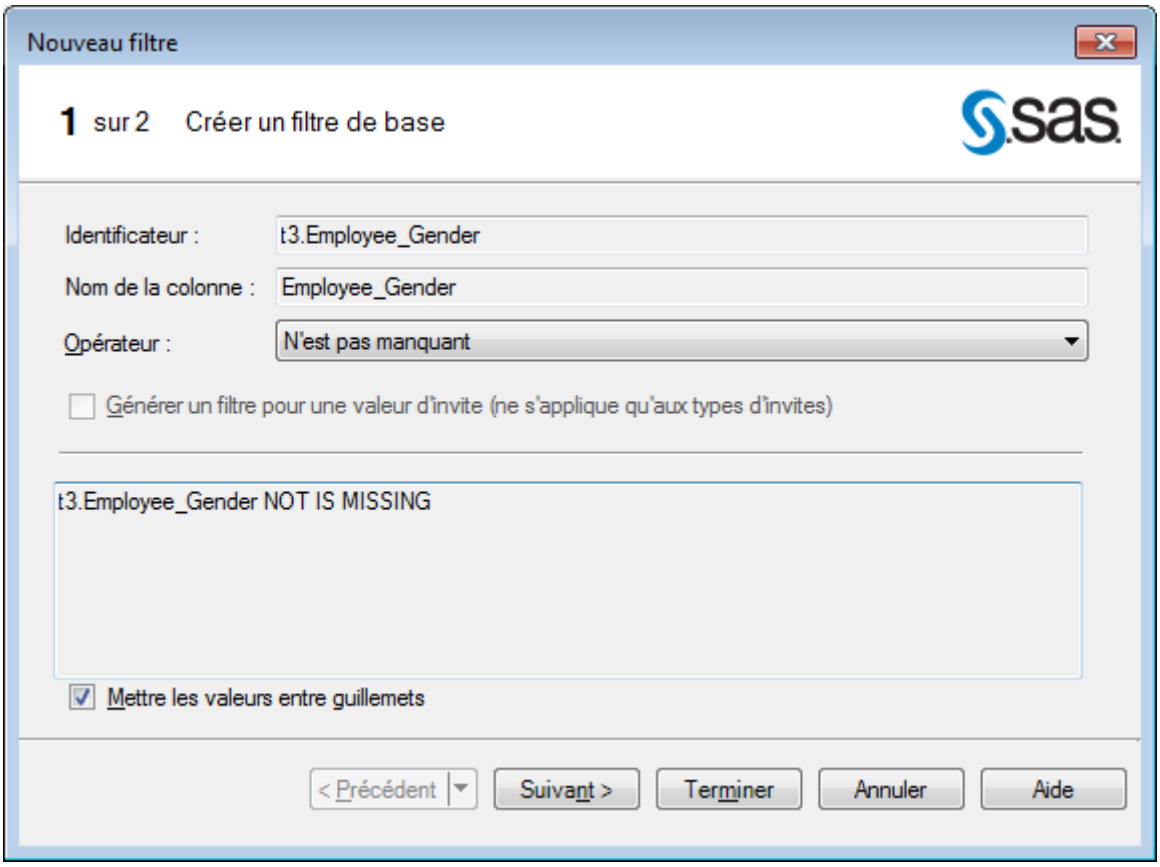
Sélectionner les colonnes **Order_ID (Order_Fact), Customer_Gendre (Customer_Dim)** et **Employee_Gender (Organization_Dim)**

Il peut y avoir plusieurs lignes de commande dans une commande. On souhaite uniquement une ligne par commande. Il faut donc cocher **Sélectionner les lignes distinctes seulement.**

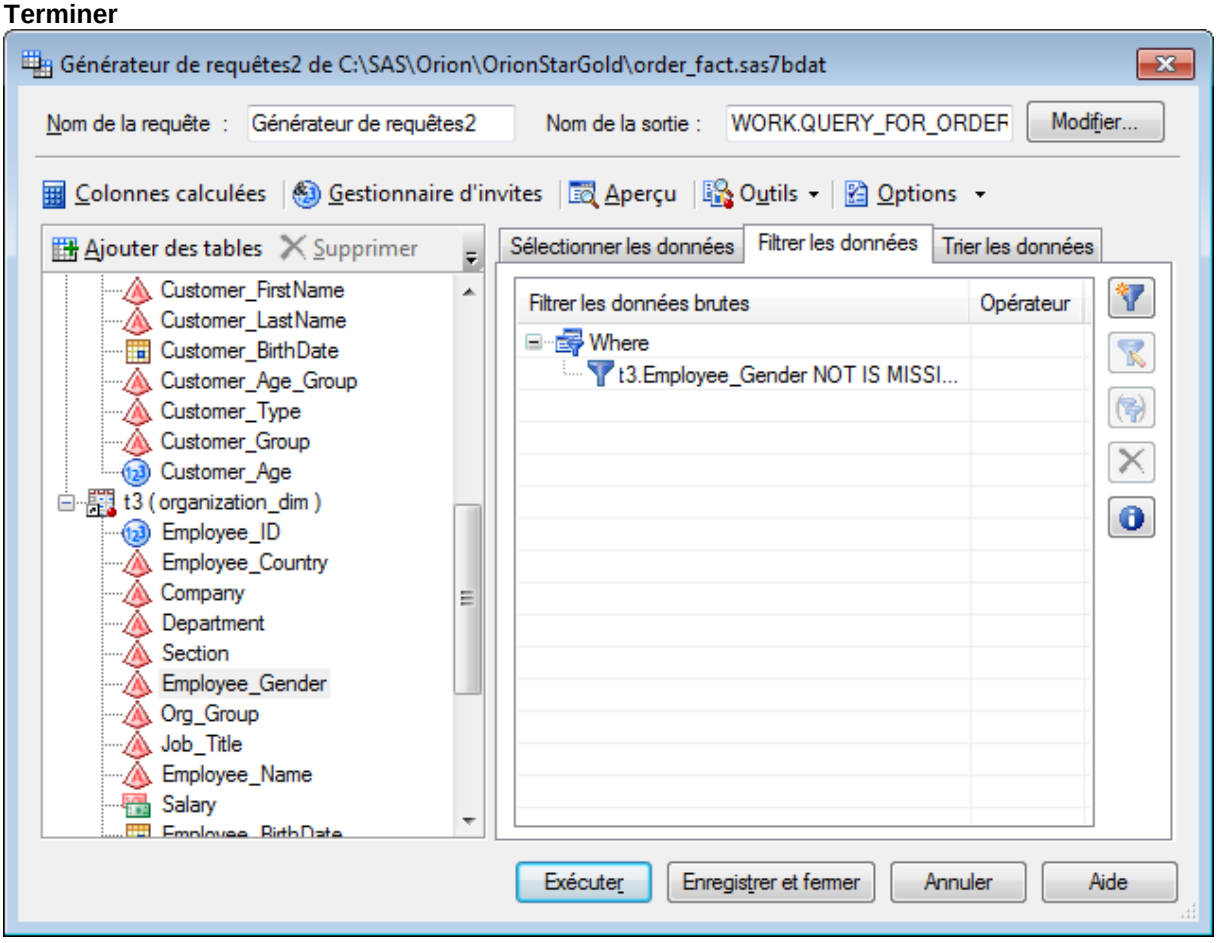


On ne souhaite pas les lignes des commandes par internet, c'est-à-dire celle où le sexe du commercial n'est pas définit.
Il faut donc faire un filtre sur le sexe du commercial de telle sorte que celui-ci ne soit pas manquant.

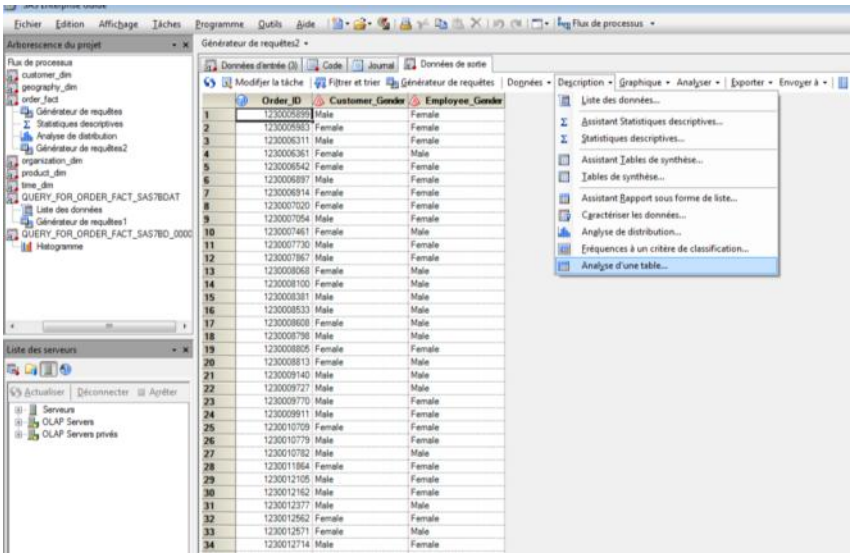
Aller dans l'onglet **Filtrer les données**
Sélectionner la colonne du sexe de l'employé



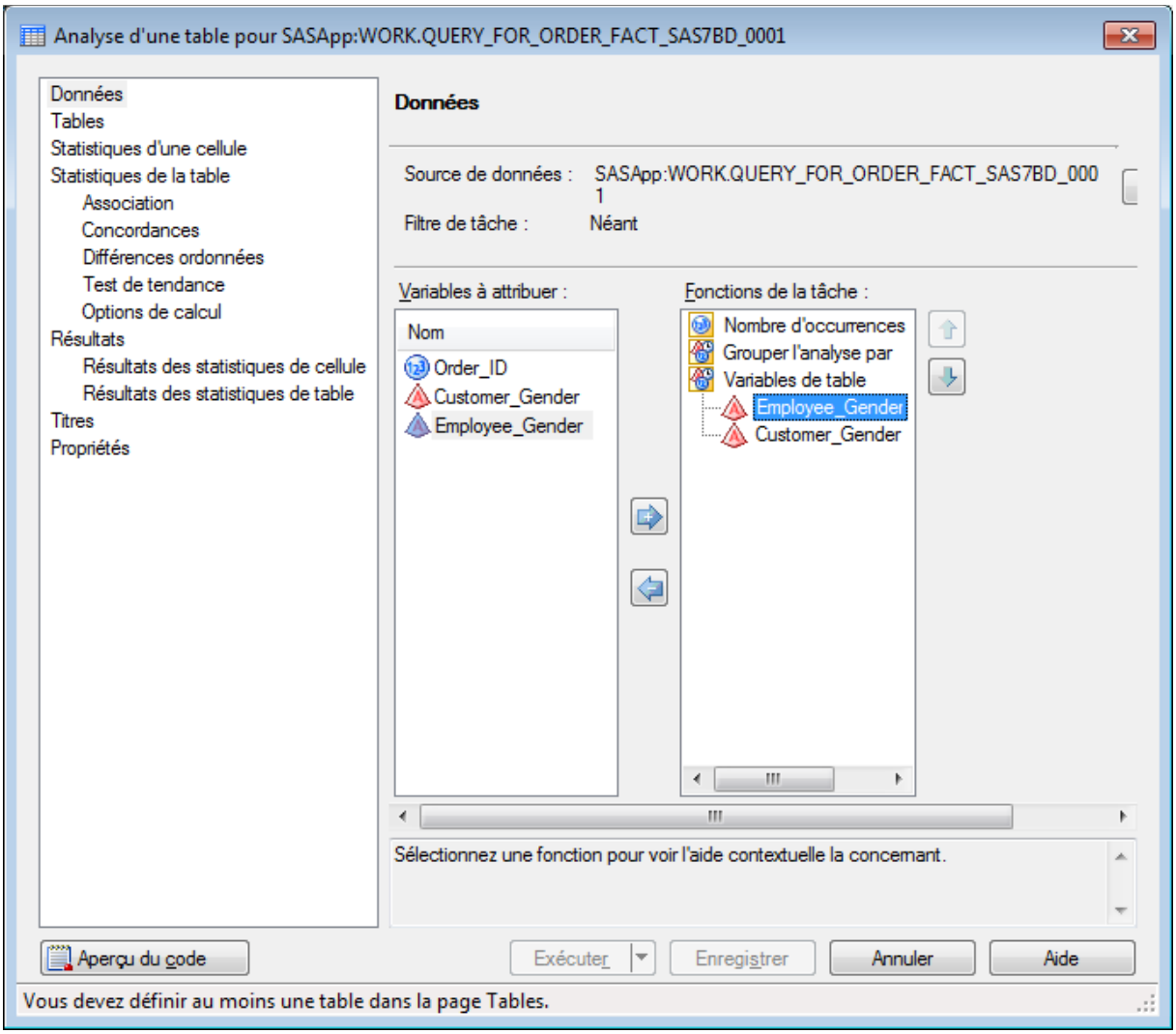
Sélectionner l'opérateur **n'est pas manquant**
Terminer



OK, puis **exécuter** la requête

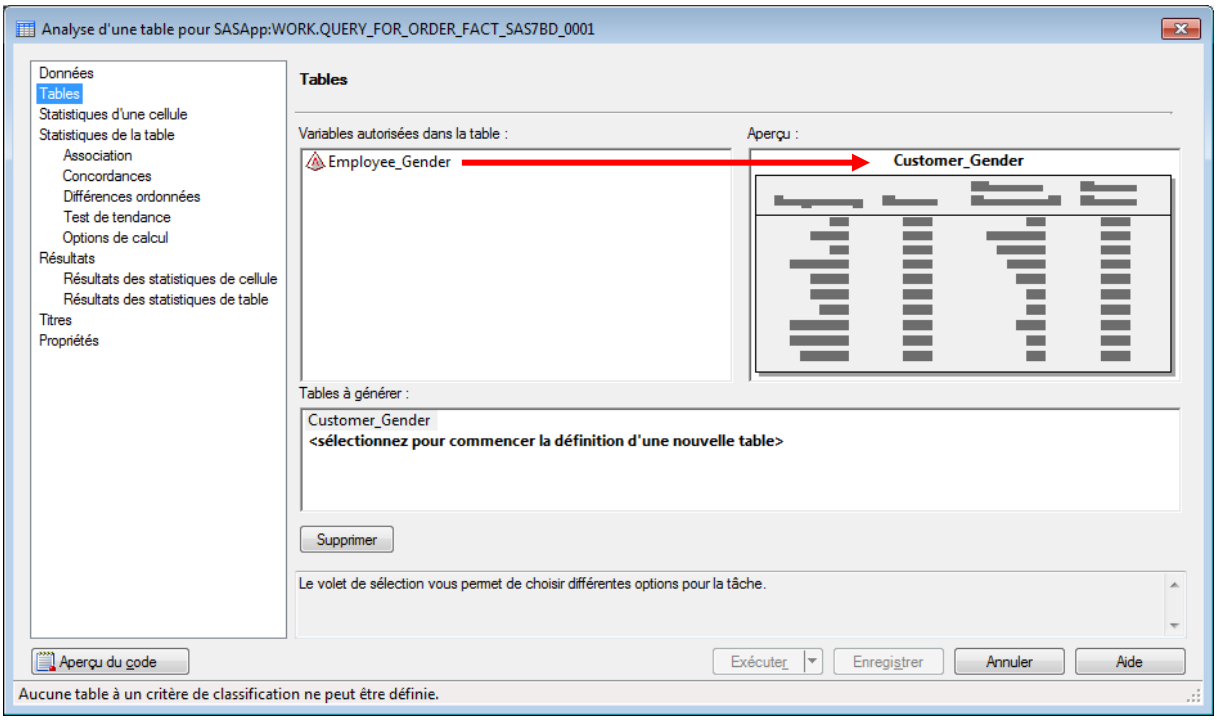


Ayant la liste des commandes, le sexe du client et celui du commercial
Aller dans le menu des **tâches → Description →** pour lancer l'**Analyse d'une table**

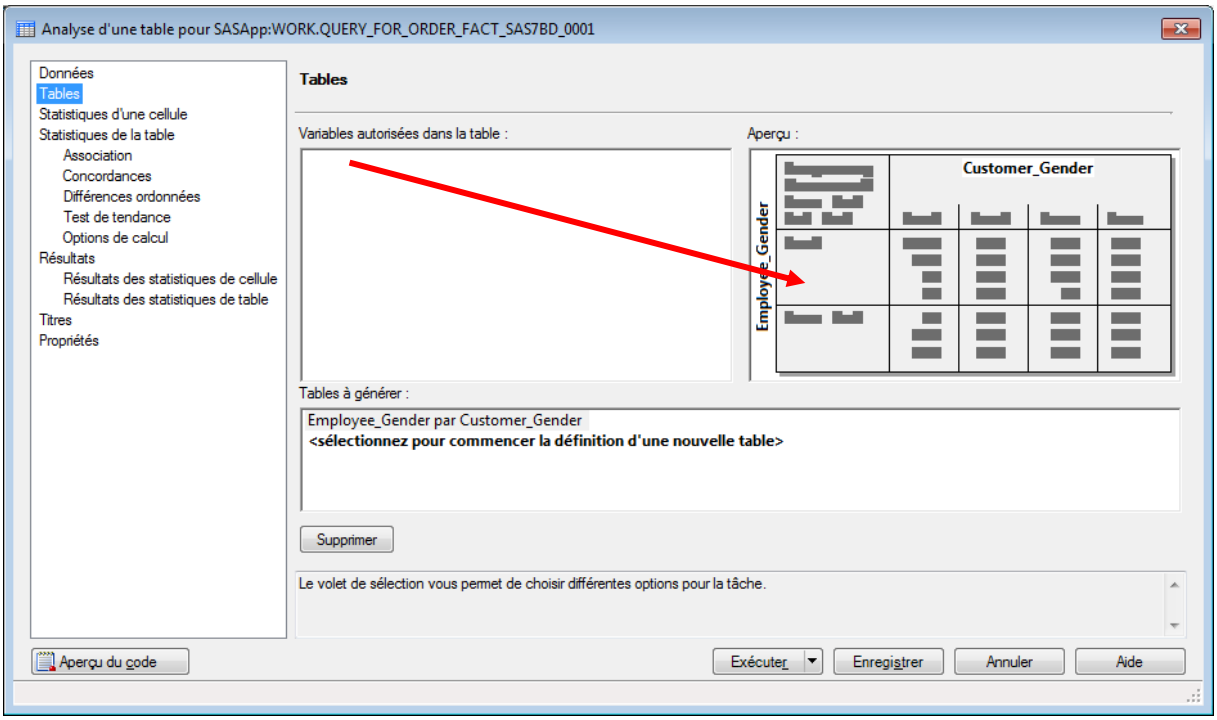


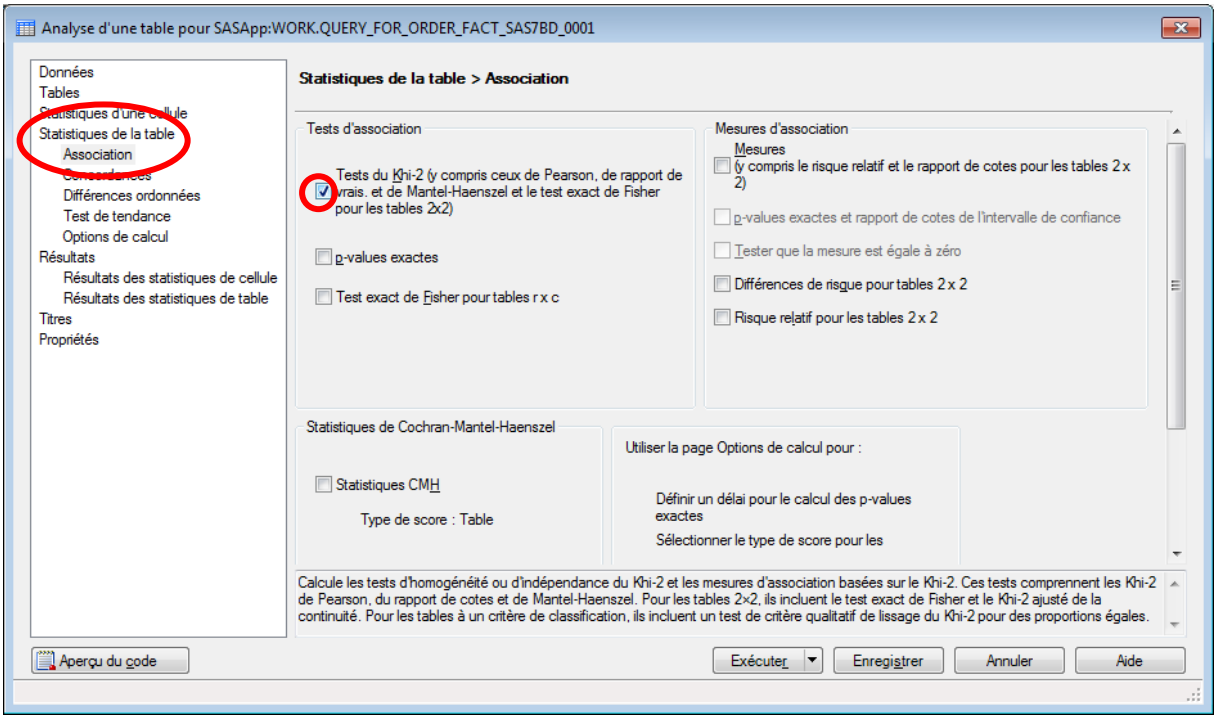
Sélectionner les colonnes du sexe du commercial et celui du client en **variables de table**

Dans le menu de gauche, sélectionner la définition des paramètres des **Tables**
Déplacer le sexe du client dans la zone **aperçu**

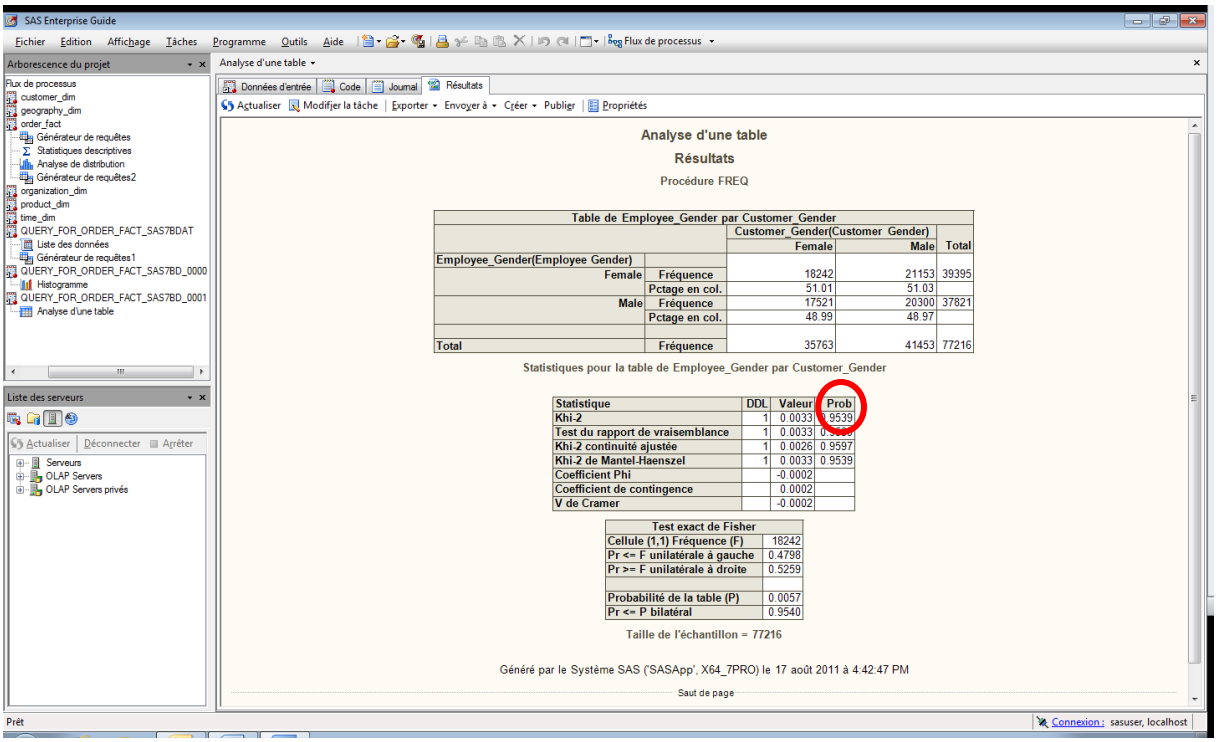


Déplacer le sexe de l'employé dans la zone **aperçu**





Dans le menu de gauche, sélectionner la définition des paramètres d'**Association**
Cocher le **test du Chi-deux**
Exécuter



Le Test du Chi² donne une Proba.=0.9539 très largement supérieur à 0.05. Nous ne pouvons pas rejeter l'hypothèse d'indépendance entre les deux variables au seuil de 5%.
Sur l'ensemble des commandes, il y a donc indépendance entre le sexe du client et celui du commercial.
Ce qu'il faut retenir de l'interprétation du test du khi-2 :
Proba. Inférieur à 0.05 → Les variables sont dépendantes. Il existe une relation.
Proba. Supérieur à 0.05 → Les variables sont indépendantes. Il n'y a pas de relation.

Comparer des moyennes avec les tests t

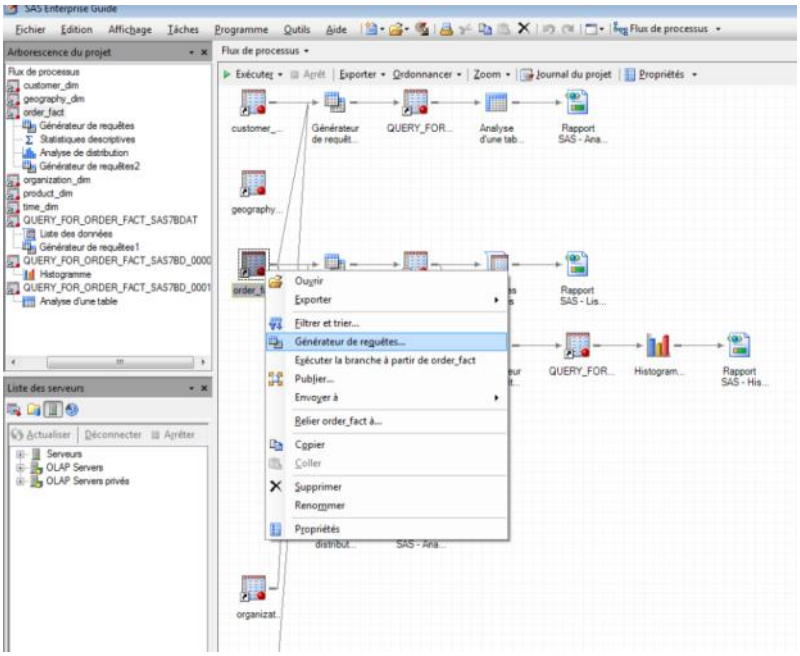
Parmi les analyses de variance, (comment « varient » les variables) dites ANOVA (ANalysis Of VAriance), la première que nous allons voir est le test en T ou « T test » en anglais, ou bien encore appeler le test de Student.

L'objectif est de faire un test en T pour savoir s'il y a une différence significative entre la performance commerciale des hommes et celle des femmes de la société Orion Star.

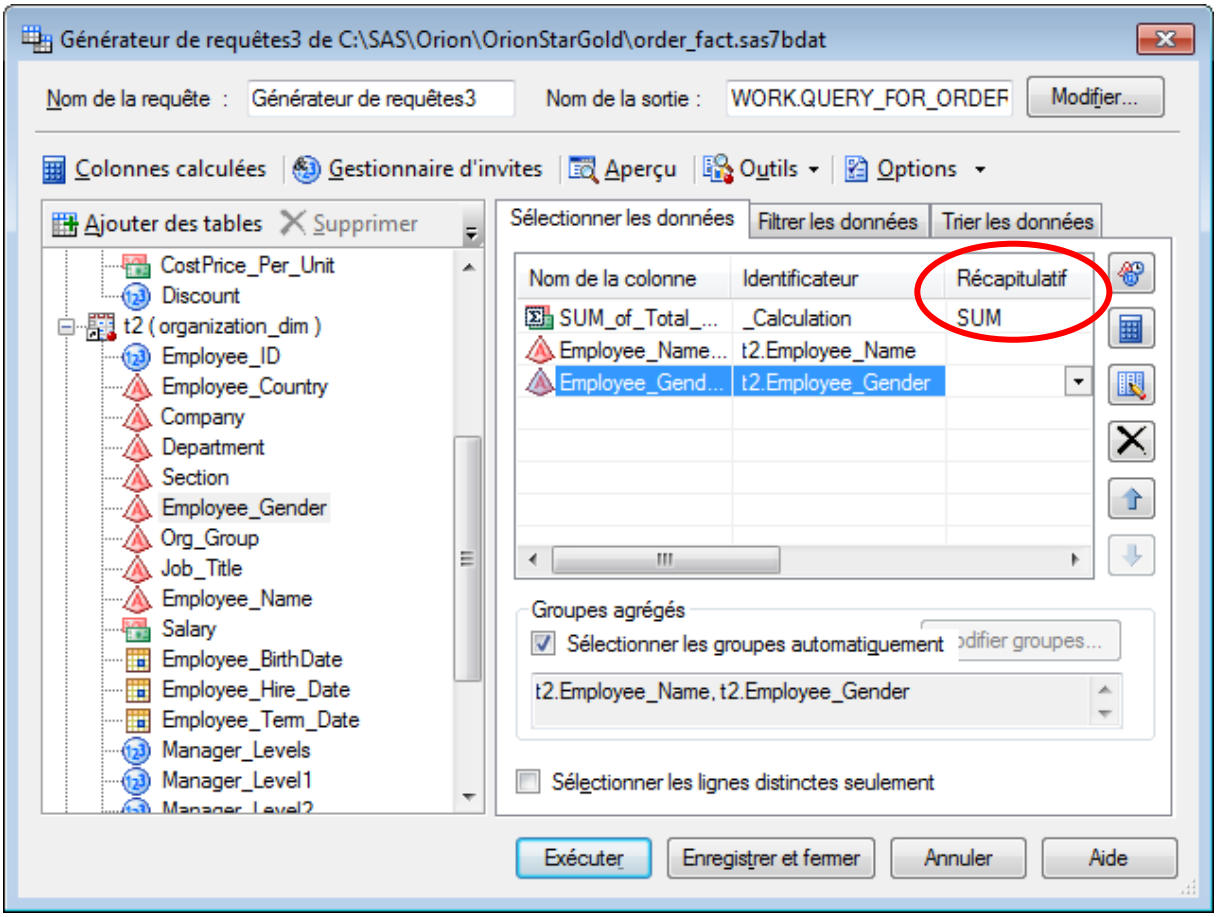
Pour cela, il faut faire une requête sur la table Order_fact, ajouter la table Organization_DIM, calculer la somme du chiffre d'affaires par collaborateur et par sexe du collaborateur.

Un test *t* permet de :

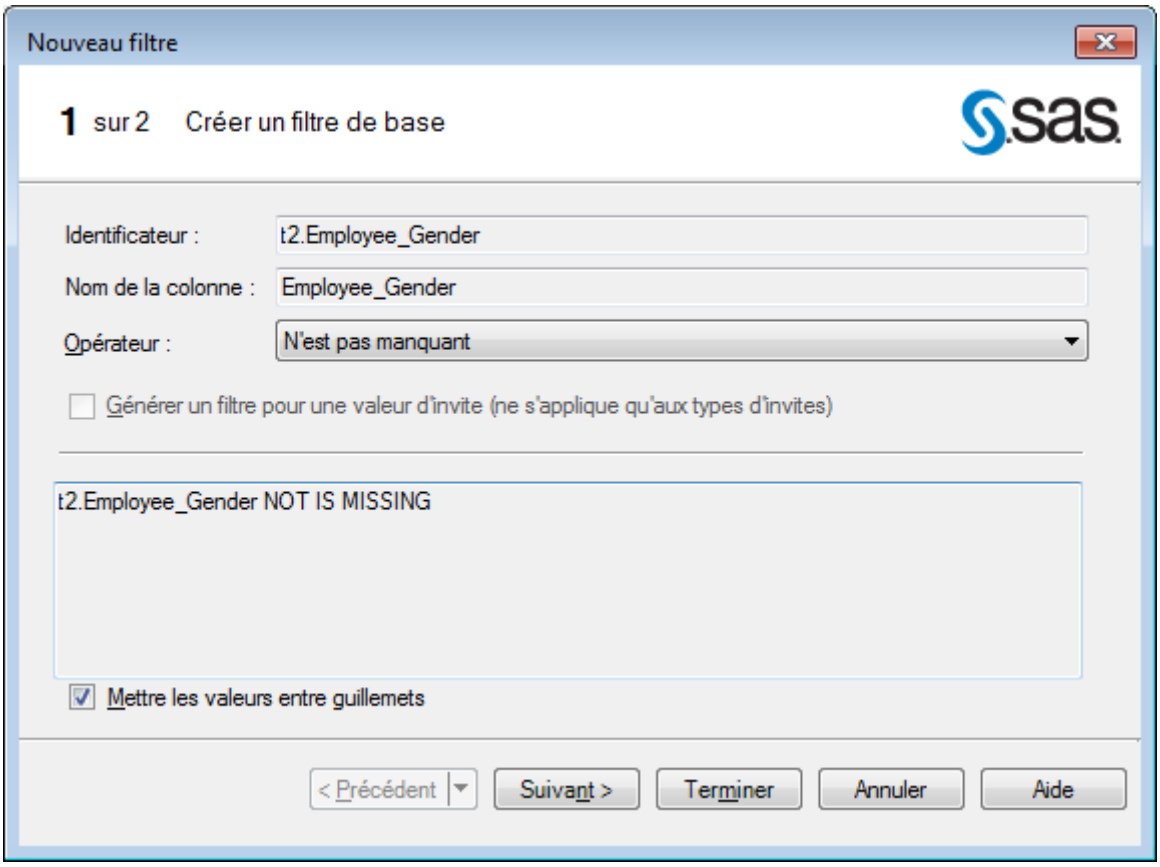
- Comparer une moyenne d'un échantillon à une valeur connue,
- Comparer les moyennes de deux échantillons liés (ou d'échantillons observés à des instants différents),
- Comparer les moyennes de deux échantillons indépendants.



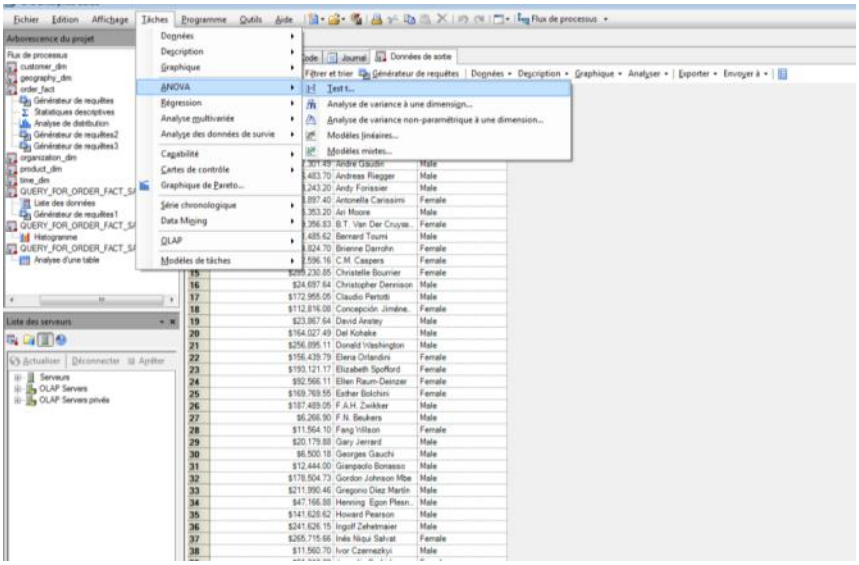
Revenir dans la fenêtre de **flux de processus**.
Créer une requête sur la table **Order_Fact**
Ajouter la table **Organization_Dim**



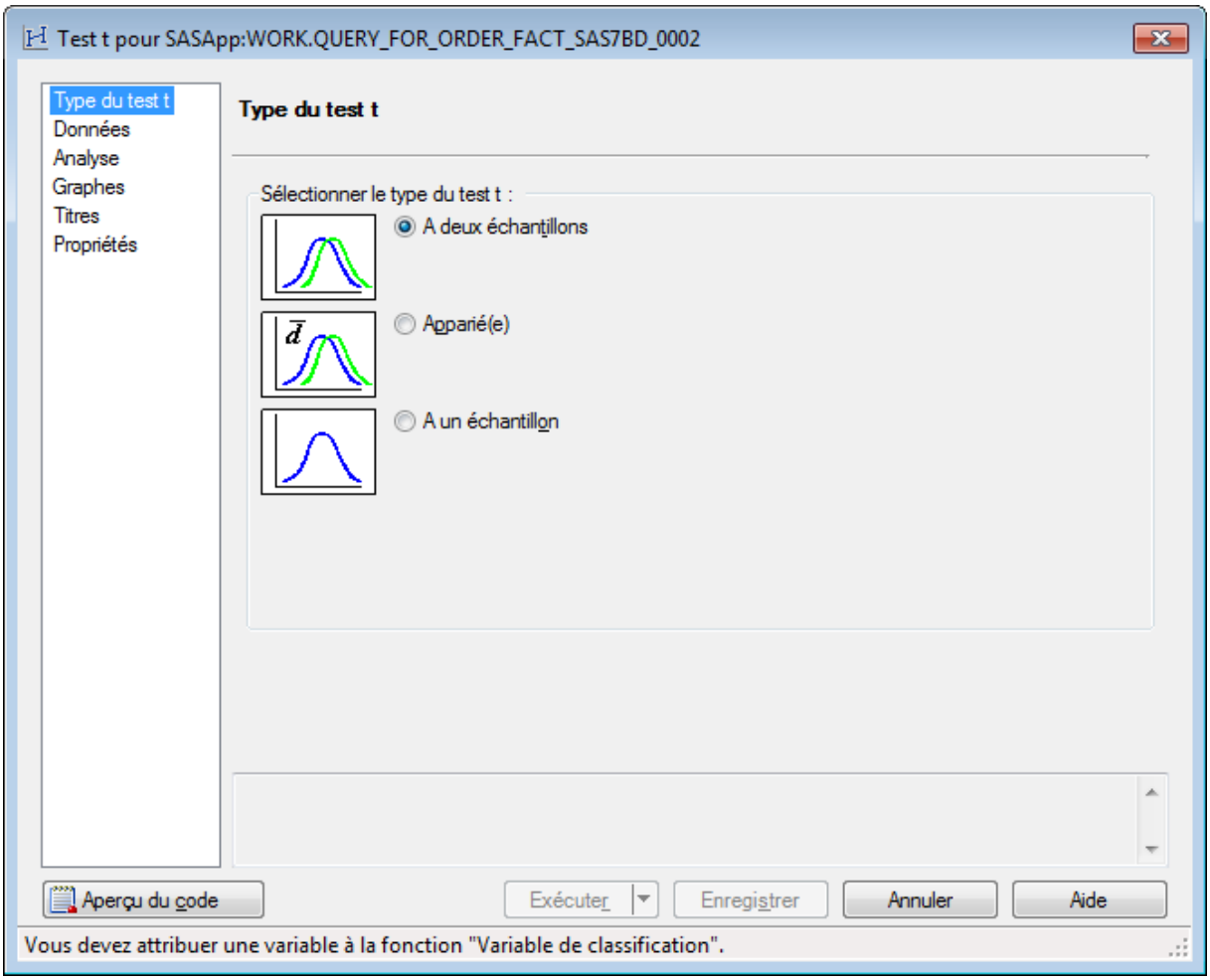
Sélectionner les colonnes du chiffre d'affaires (Total_Retail_Price), du sexe (Employee_Gender) et du nom de l'employé (Employee_Name).
Sélectionner la somme du chiffre d'affaires depuis récapitulatif (Total_Retail_Price devient SUM_of_Total_Retail_Price).
Remarque : il est préférable d'utiliser Employee_ID au lieu de Employee_Name au cas où il y ait des homonymes.



Faire un filtre sur le sexe de l'employé, de telle sorte que celui-ci ne soit pas manquant, ce qui revient à éliminer les ventes faites par internet et par catalogue.
Exécuter la requête.



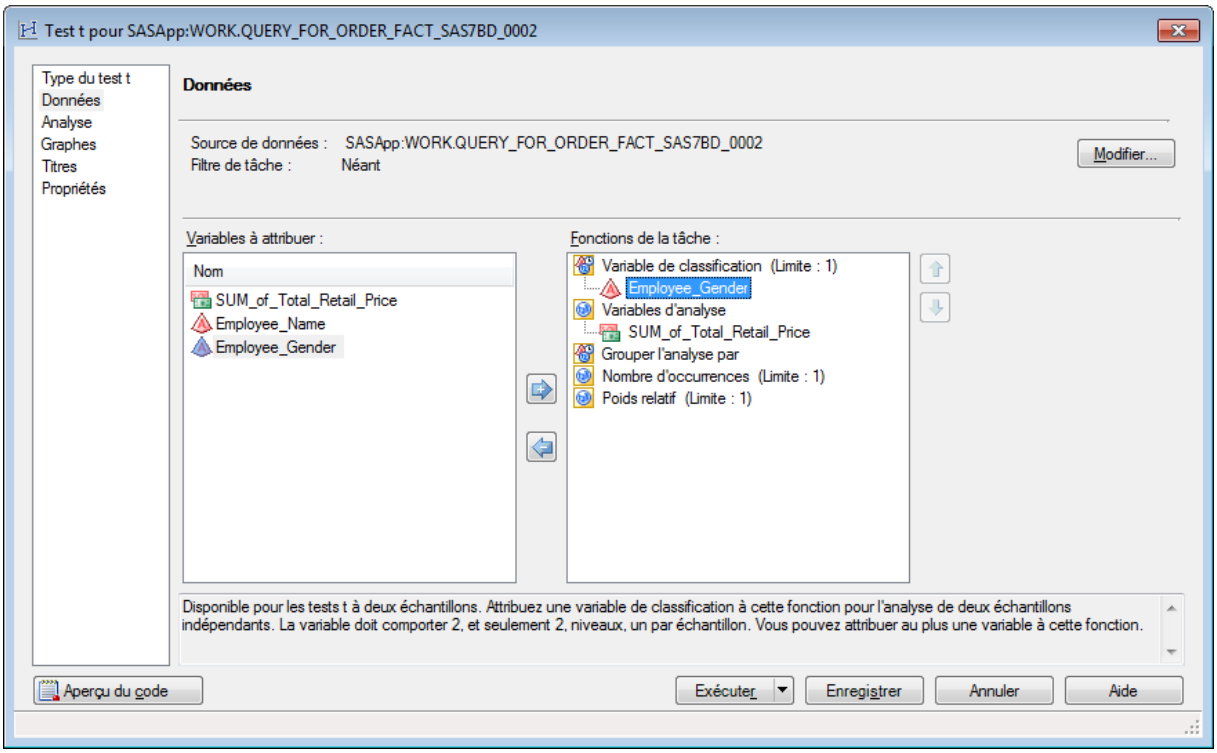
Dans le menu des tâche, parmi les « ANOVA » sélectionner le test en T.



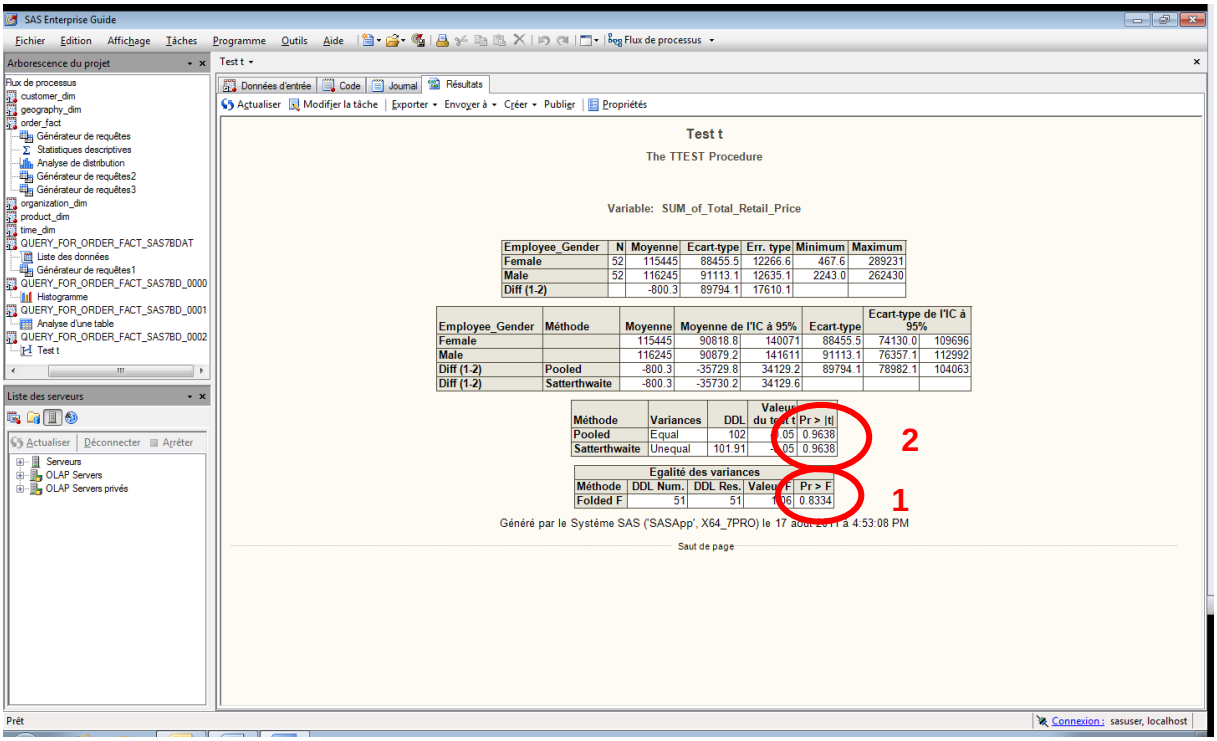
Dans notre cas, nous avons deux échantillons : les hommes et les femmes. (Une colonne de la variable d'analyse et une colonne avec deux modalités pour identifier les deux échantillons)

Si vous souhaitez analyser strictement le même échantillon, mesuré à deux instants ou de deux façons, il faut utiliser le test Apparié. (Indication mnémotechnique : dans « apparié », il y a « paire ») Par exemple, si vous analysez la même population avant et après une campagne, pour savoir si la campagne a eu un effet ou pas. (Une colonne du chiffre d'affaires et une colonne avec deux modalités)
(Deux colonnes de la variable d'analyse suivant les deux modalités)

Si vous souhaitez comparer la moyenne d'un seul échantillon à une valeur précise, c'est le troisième test.



Dans l'onglet **Fonction de la tâche**, affecter l'**Employee_Gender** à **Variable de classification** et le Chiffre d'affaires à **Variables d'analyses**.
Exécuter



1) Dans les résultats, dans Égalité des variances, l'hypothèse nulle est « les variances sont égales ». Si ce test est significatif, on a la preuve que les variances pour chaque groupe sont différentes (en d'autres termes, l'hypothèse d'égalité des variances n'est pas vérifiée). Dans notre exemple, la p-

value (Pr > F = 0.8334) est supérieure au seuil de 0.05 (5%) et l'on peut donc conclure que l'hypothèse d'égalité des variances est vérifiée. Les variances sont égales.

2) Ensuite, dans les résultats intitulés Tests de Student, la p-value à lire est celle pour des variances égales (Equal). Elle est donc supérieure à 0,05 (Pr > F = 0.9638) et l'on peut donc en conclure que la somme du chiffre d'affaires est en moyenne non significativement différente entre les hommes et les femmes.

D'après le premier tableau, même si la moyenne de la somme du chiffre d'affaires de 116 245$ pour les hommes est supérieure à celle de 115 445$ pour les femmes, on ne peut pas conclure qu'il y a une différence significative.

Ce qu'il faut retenir de l'interprétation du test en T pour deux échantillons
On lit d'abord le tableau du même nom pour savoir s'il y a égalité des variances
- Si Pr>F Inférieur à 0.05 → Les variances sont différentes
- Si Pr>F Supérieur à 0.05 → Les variances sont égales

On lit d'abord le tableau du test de Student en fonction de l'égalité ou non des variances
- Si Pr>|t| Inférieur à 0.05 → Les moyennes sont différentes
- Si Pr>|t| Supérieur à 0.05 → Les moyennes sont égales

Exercice : Faire une requête sur la table Order_fact, ajouter la table Customer_DIM, calculer la somme du chiffre d'affaires par Customer_name et par Customer_gender. Faire un test en T pour savoir s'il y a une différence significative entre la moyenne des achats des hommes et celle des femmes.

Anova : Modèle linéaire

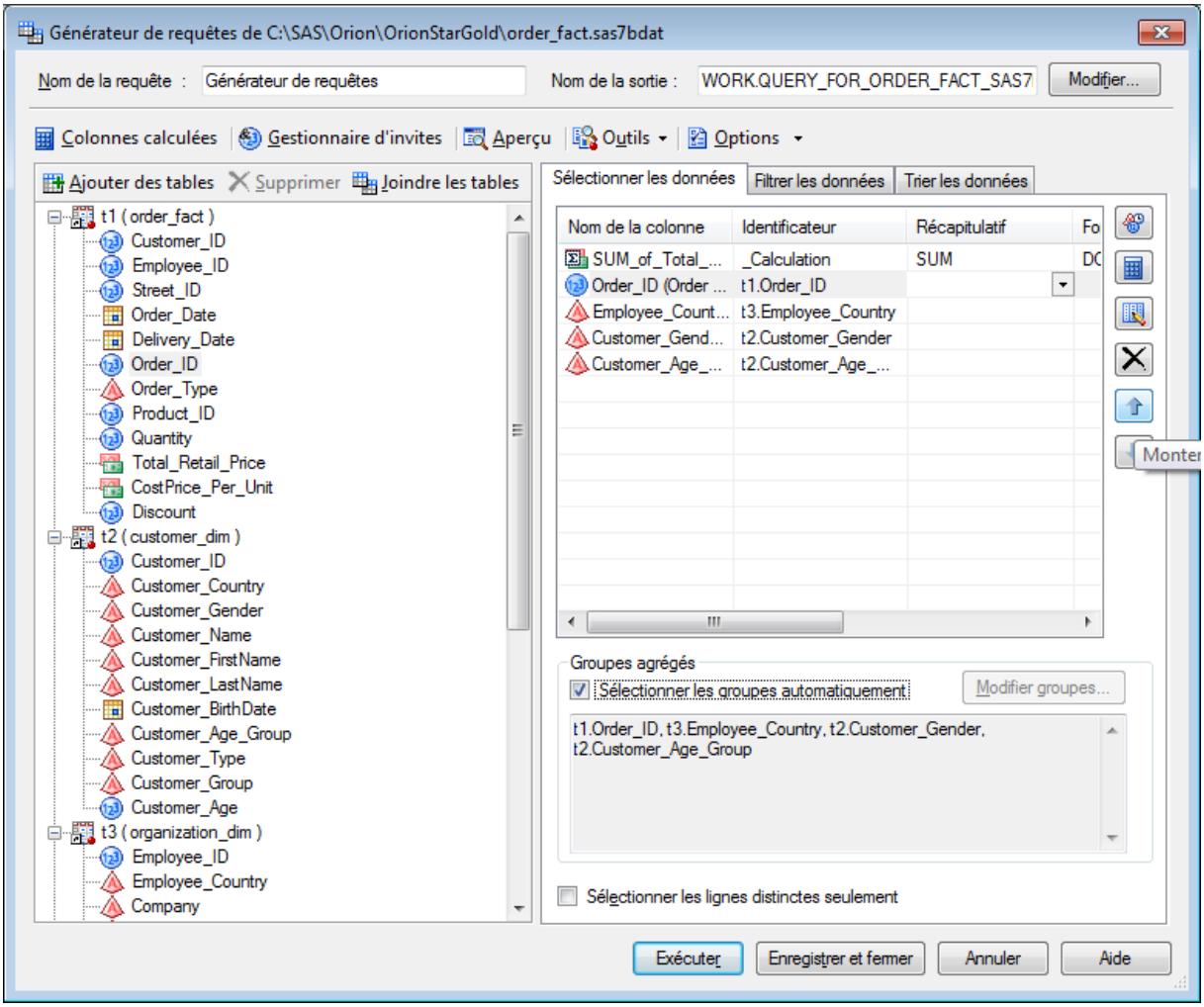
Le test en T permet de comparer deux échantillons.
Pour aller plus loin, nous allons maintenant utiliser un modèle linéaire généralisé de type ANOVA pour savoir en une seule fois s'il y a ou non une relation entre plusieurs variables quantitatives ou qualitatives et un variable dépendante continue.

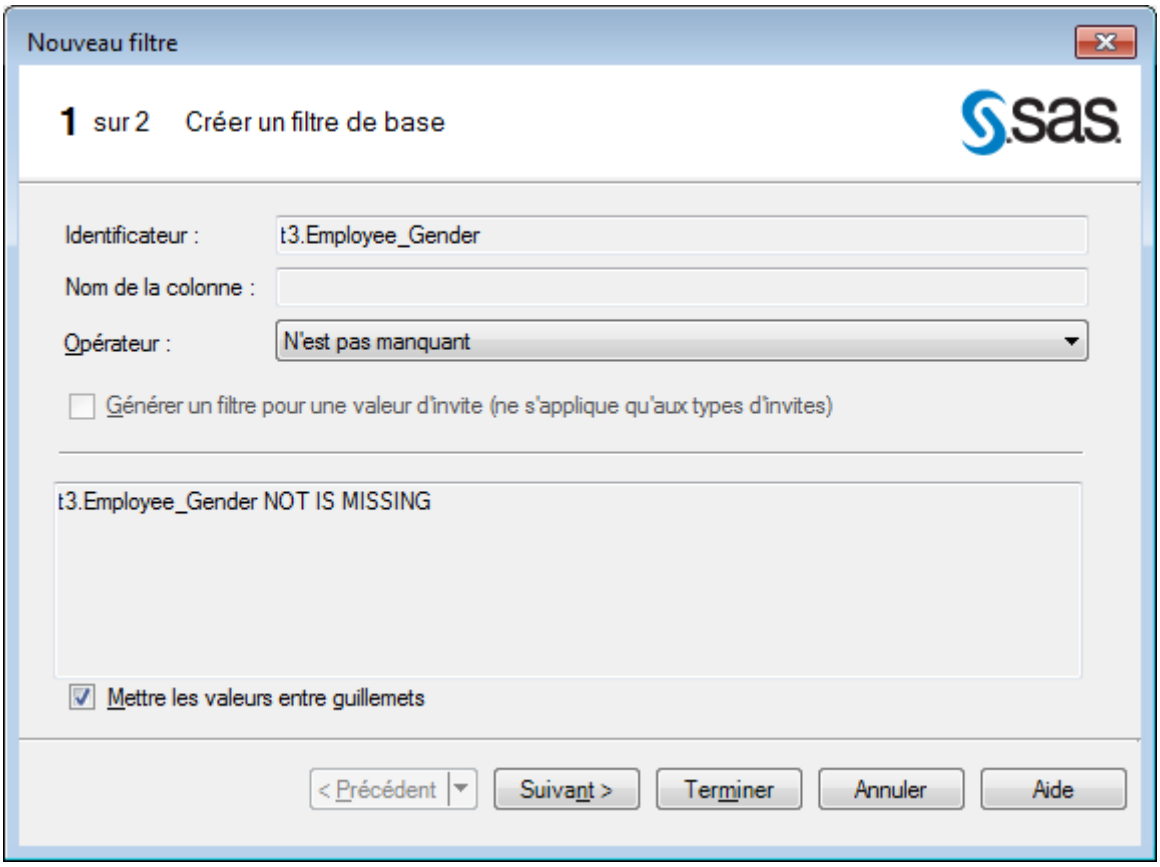
La question est : Quelles sont les variables dont dépend le panier moyen ?

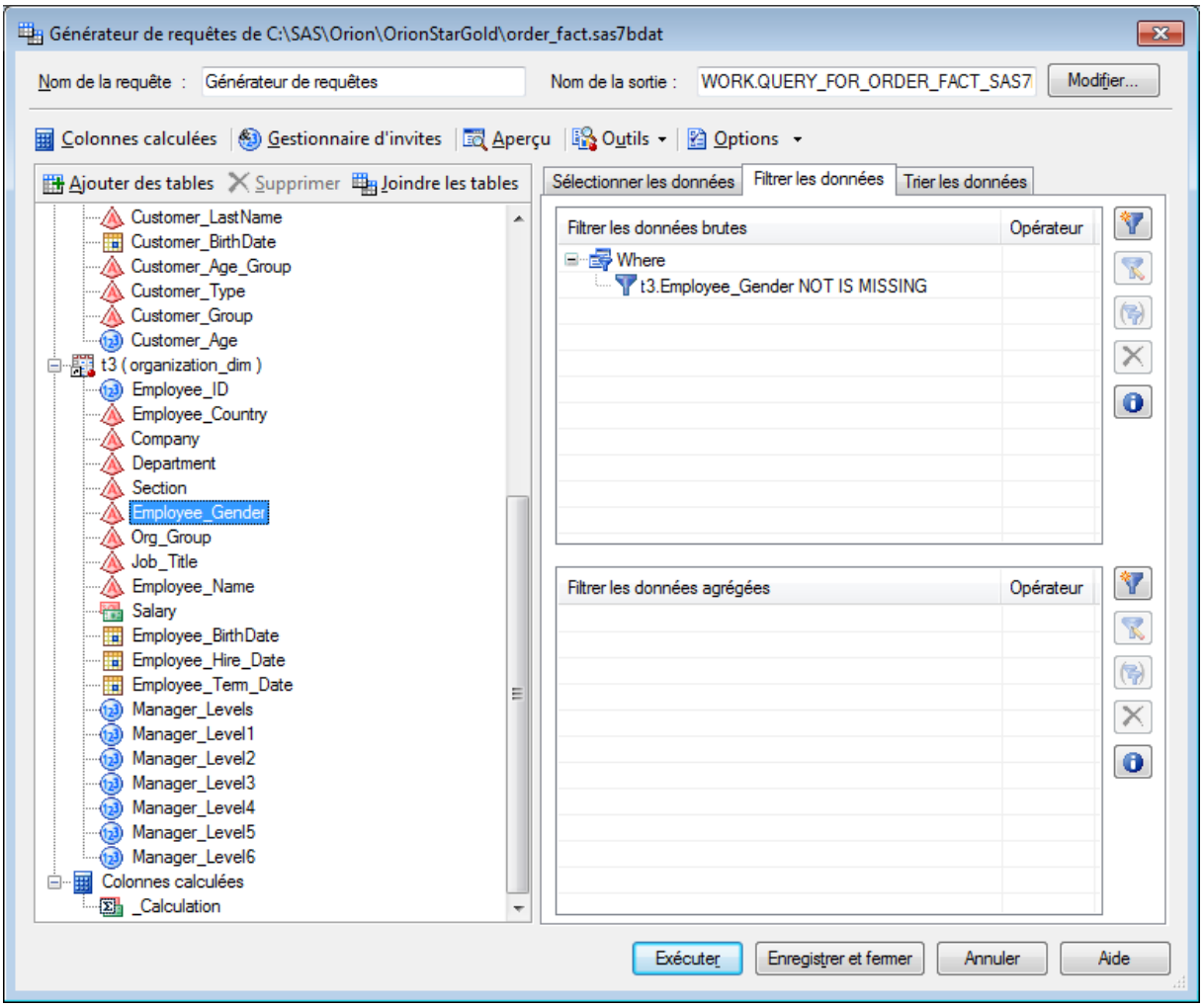
Pour cela, nous allons créer une table avec la somme du chiffre d'affaires par order_ID afin de calculer le panier et par

- Pays de l'employé
- Sexe du client
- Groupe d'âge du client.

Revenir dans la fenêtre de **Flux de processus**
Créer une requête sur la table **Order_fact**
Ajouter les tables **Customer_Dim** et **Organization_Dim**

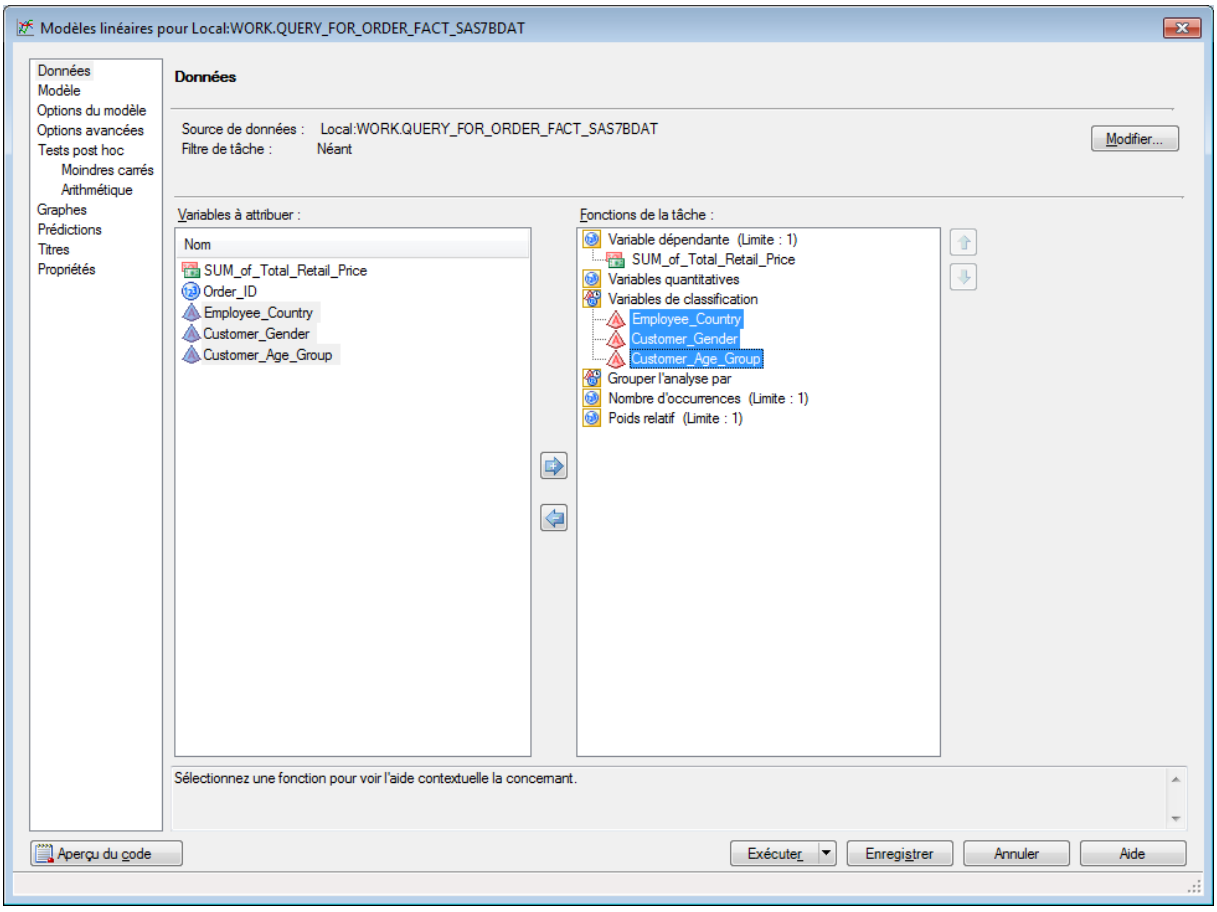


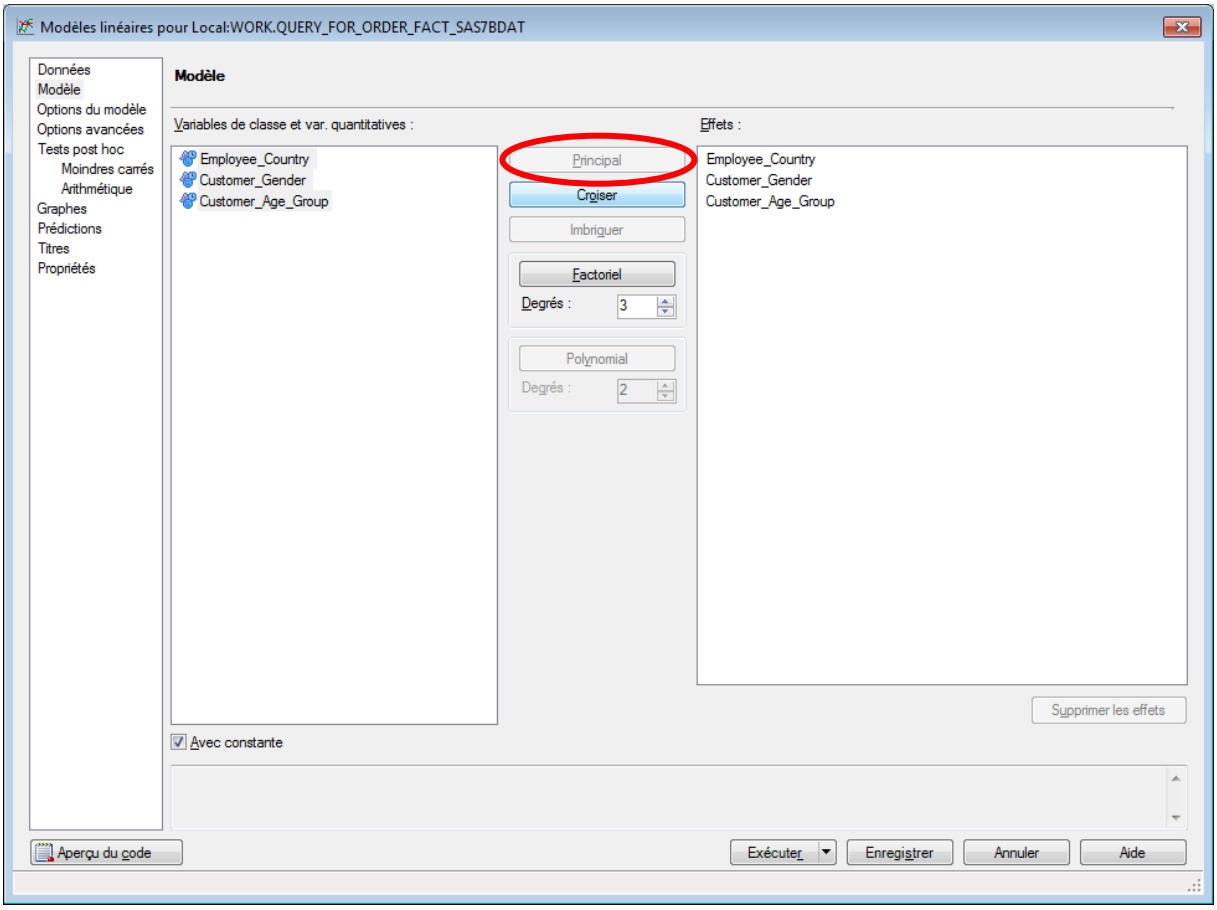

Nouveau filtre
1 sur 2 Créer un filtre de base
sas
Identificateur :
t3.Employee_Gender
Nom de la colonne :
Opérateur :
N'est pas manquant
Générer un filtre pour une valeur d'invite (ne s'applique qu'aux types d'invites)
t3.Employee_Gender NOT IS MISSING
Mettre les valeurs entre guillemets
< Précédent
Suivant >
Terminer
Annuler
Aide



Faire un filtre de tel sorte que les commandes par internet ou catalogue soient exclus. C'est-à-dire, exclure les lignes où le sexe du collaborateur n'est pas renseigné.
Exécuter la requête.
Créer sur la table générée un modèle linéaire (menu **Analyses → ANOVA → Modèle linéaire**)

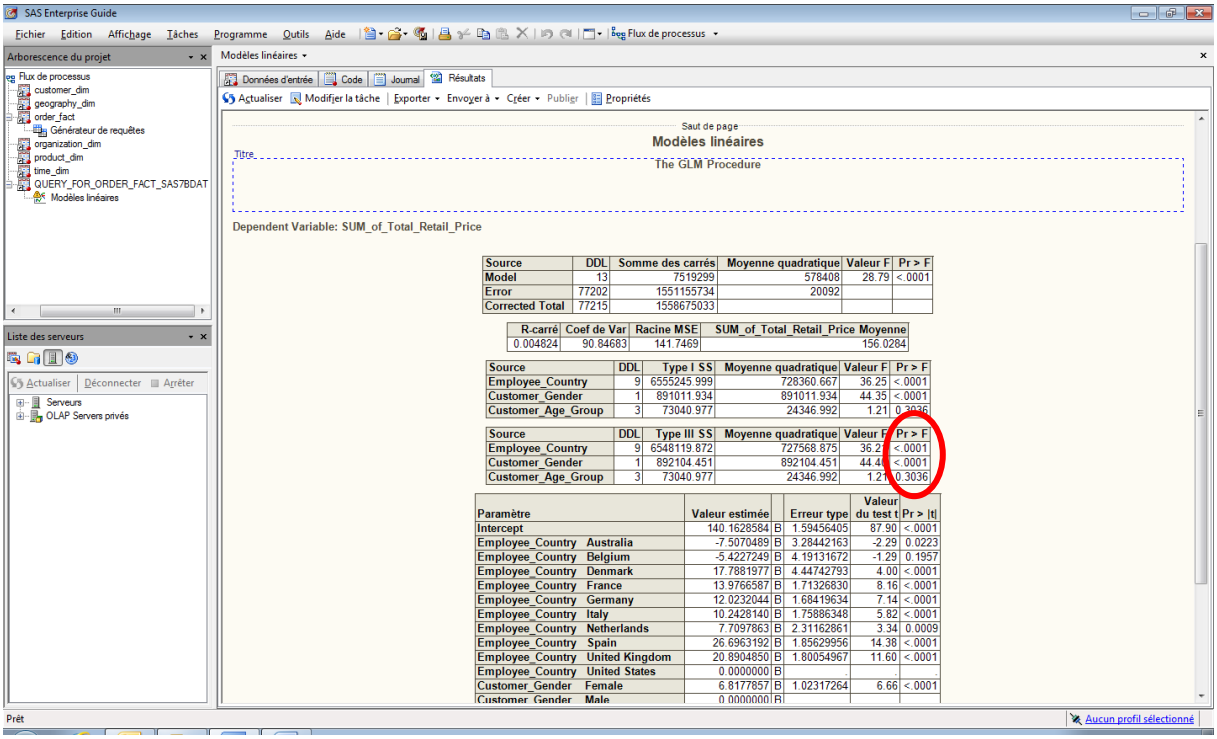
La somme du chiffre d'affaires doit être affectée à la variable dépendante.
Le pays et le sexe de l'employé et le sexe du client sont les variables de classification.


Modèles linéaires pour Local:WORK.QUERY_FOR_ORDER_FACT_SAS7BDAT
Données
Modèle
Options du modèle
Options avancées
Tests post hoc
Moindres carrés
Arithmétique
Graphes
Prédictions
Titres
Propriétés
Données
Source de données : Local:WORK.QUERY_FOR_ORDER_FACT_SAS7BDAT
Filtre de tâche : Néant
Modifier...
Variables à attribuer :
Nom
SUM_of_Total_Retail_Price
Order_ID
Employee_Country
Customer_Gender
Customer_Age_Group
Fonctions de la tâche :
Variable dépendante (Limite : 1)
SUM_of_Total_Retail_Price
Variables quantitatives
Variables de classification
Employee_Country
Customer_Gender
Customer_Age_Group
Grouper l'analyse par
Nombre d'occurrences (Limite : 1)
Poids relatif (Limite : 1)
Sélectionnez une fonction pour voir l'aide contextuelle la concernant.
Aperçu du code
Exécuter
Enregistrer
Annuler
Aide


Modèles linéaires pour Local:WORK.QUERY_FOR_ORDER_FACT_SAS7BDAT
Données
Modèle
Options du modèle
Options avancées
Tests post hoc
Moindres carrés
Arithmétique
Graphes
Prédictions
Titres
Propriétés
Modèle
Variables de classe et var. quantitatives :
Employee_Country
Customer_Gender
Customer_Age_Group
Principal
Croiser
Imbriquer
Factoriel
Degrés : 3
Polynomial
Degrés : 2
Effets :
Employee_Country
Customer_Gender
Customer_Age_Group
Supprimer les effets
Avec constante
Aperçu du code
Exécuter
Enregistrer
Annuler
Aide

Dans la section **Modèle,** (sélectionner Modèle dans la fenêtre de gauche)
Sélectionner toutes les colonnes dans la partie gauche.
Cliquer sur le bouton **Principal.**
Toutes les colonnes se retrouvent une fois dans la partie droite.
Exécuter

Voici les résultats du modèle linéaire :



Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	13	7519299	578408	28.79	<.0001
Error	77202	1551155734	20092		
Corrected Total	77215	1558675033			

R-carré	Coef de Var	Racine MSE	SUM_of_Total_Retail_Price Moyenne
0.004824	90.84683	141.7469	156.0284

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
Employee_Country	9	6555245.999	728360.667	36.25	<.0001
Customer_Gender	1	891011.934	891011.934	44.35	<.0001
Customer_Age_Group	3	73040.977	24346.992	1.21	0.3036

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
Employee_Country	9	6548119.872	727568.875	36.21	<.0001
Customer_Gender	1	892104.451	892104.451	44.40	<.0001
Customer_Age_Group	3	73040.977	24346.992	1.21	0.3036

Paramètre		Valeur estimée		Erreur type	Valeur du test t	Pr > |t|
Intercept		140.1628584	B	1.59456405	87.90	<.0001
Employee_Country	Australia	-7.5070489	B	3.28442163	-2.29	0.0223
Employee_Country	Belgium	-5.4227249	B	4.19131672	-1.29	0.1957
Employee_Country	Denmark	17.7881977	B	4.44742793	4.00	<.0001
Employee_Country	France	13.9766587	B	1.71326830	8.16	<.0001
Employee_Country	Germany	12.0232044	B	1.68419634	7.14	<.0001
Employee_Country	Italy	10.2428140	B	1.75886348	5.82	<.0001
Employee_Country	Netherlands	7.7097863	B	2.31162861	3.34	0.0009
Employee_Country	Spain	26.6963192	B	1.85629956	14.38	<.0001
Employee_Country	United Kingdom	20.8904850	B	1.80054967	11.60	<.0001
Employee_Country	United States	0.0000000	B	.	.	.
Customer_Gender	Female	6.8177857	B	1.02317264	6.66	<.0001
Customer_Gender	Male	0.0000000	B	.	.	.

C’est le quatrième tableau qui nous intéresse, celui pour le Type III SS.
Si Pr > F est plus petite que 0.05, on rejette l’hypothèse nulle.

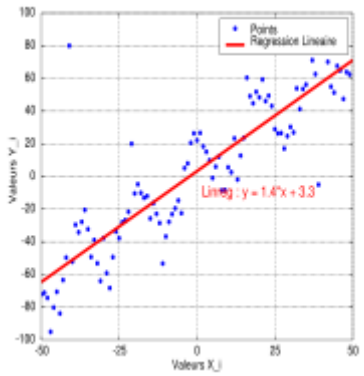
La lecture est donc :
Lorsque les variables Employee_Country, et Customer_Gender changent de modalité, le panier moyen varie.

Régression linaire simple

Petite introduction à la régression linéaire

La régression est un modèle mathématique très utilisé. Si l'on représente des individus par un nuage de point, l'algorithme de la régression recherche l'équation de la droite passant par ces points, ou plutôt au milieu de ces points.

Exemple : on cherche à expliquer Y en fonction de X. tous les individus sont représentés par des points bleus.



La régression linéaire simple est une équation du type Y = a * X + b ou a et b sont des valeurs à déterminer.
La droite de régression est la droite qui passe au milieu du nuage de point.
B est une constante. C'est la valeur de l'ordonnée à l'origine (pour X=0), c'est-à-dire le niveau où la droite de régression coupe l'axe des ordonnées. Dans SAS, elle s'appelle « Intercept ».
A est la pente de la droite.
La régression linéaire consiste à estimer les valeurs a et b.

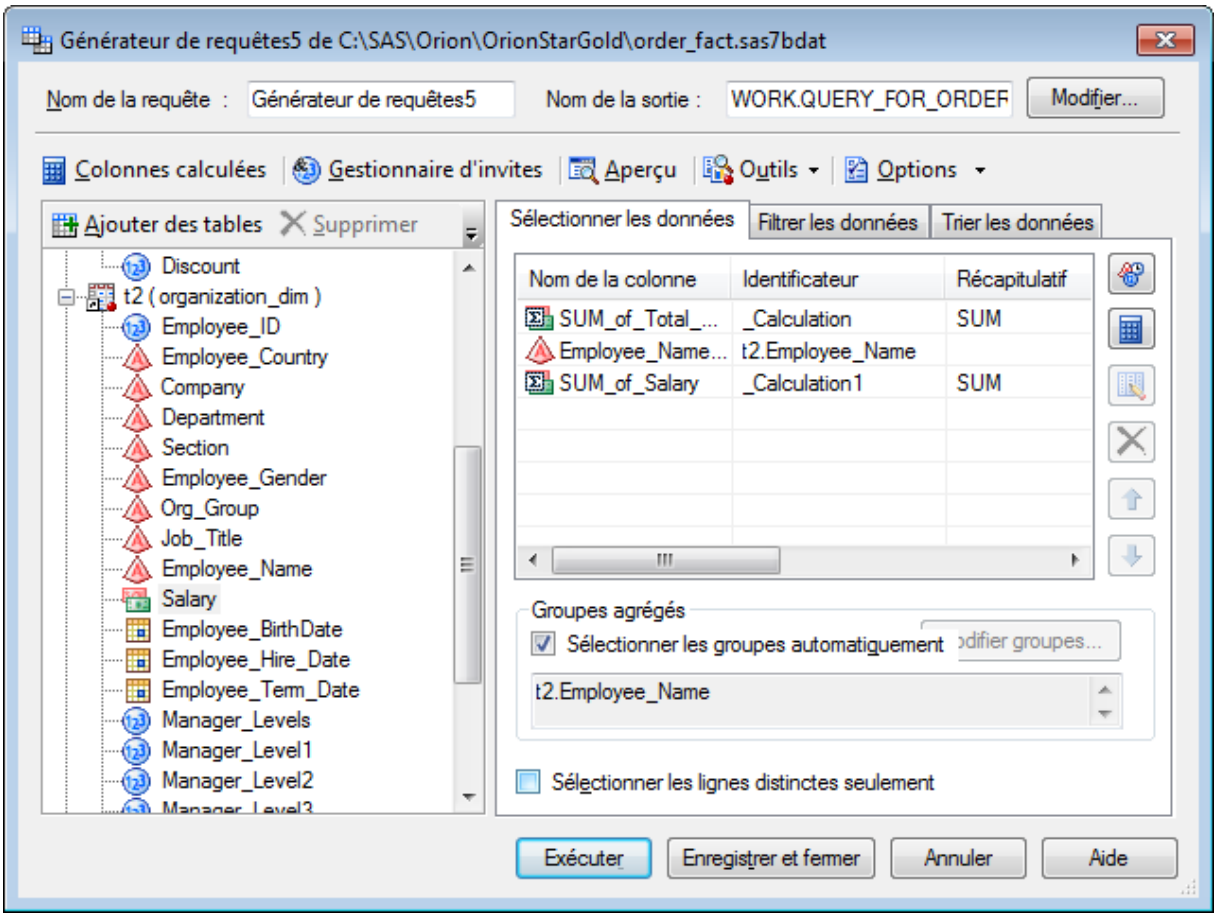
La généralisation à p variables s'appelle la régression linéaire multiple.
Y = a0 + a1·X1 + a2·X2 + … + ap·Xp

La régression linéaire permet d'expliquer une variable continue (quantitative) en fonction d'autres variables.

La question est : Trouver la relation entre le chiffre d'affaires généré par un commercial et son salaire.

L'analyse est donc de regarder a postériori si le salaire du commercial est bien fonction de ses ventes.

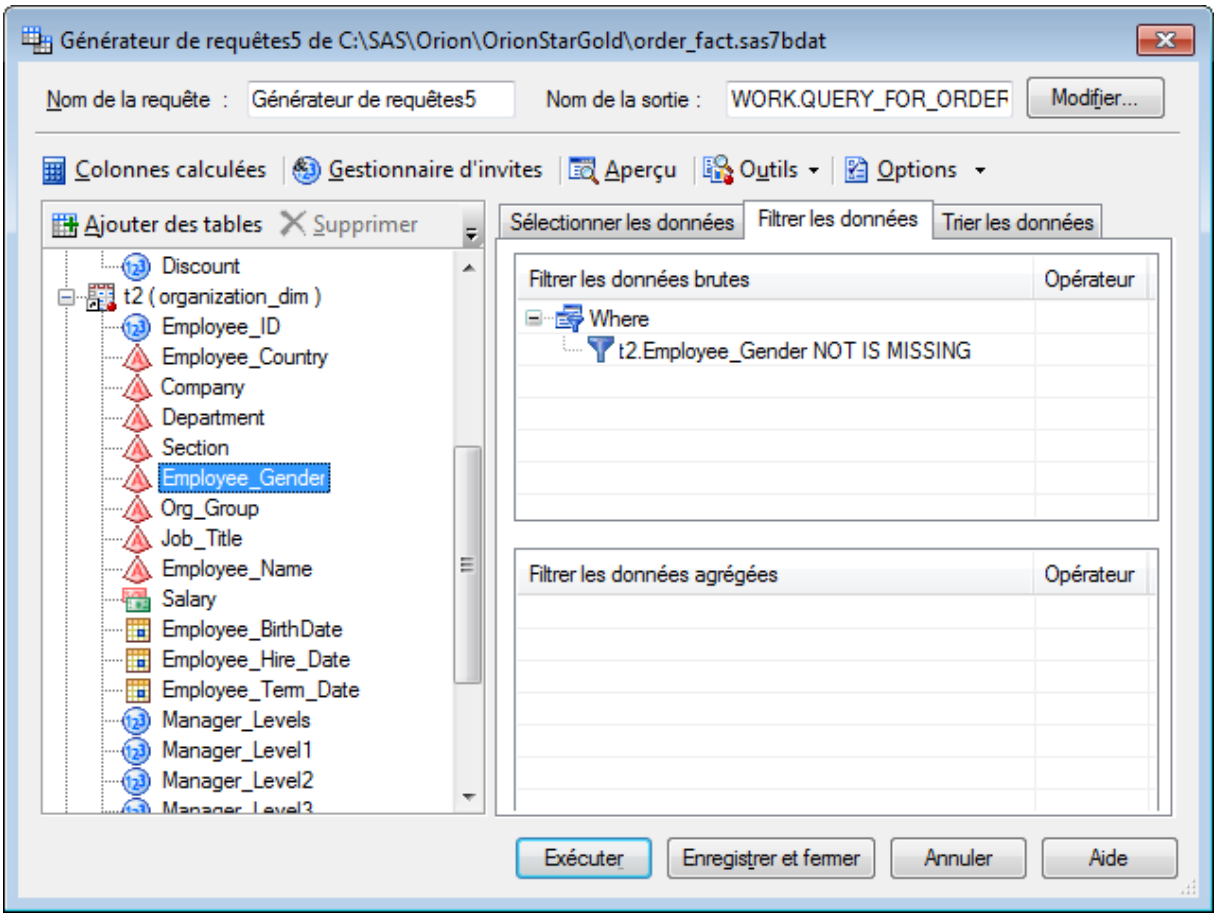
Pour cela,
Depuis la fenêtre de **flux de processus**
Sélectionner la table **Order_fact.**
Clic-droit de la souris sur cette table sélectionnée, **Générateur de requête**.
Ajouter la table **Organization_Dim**.



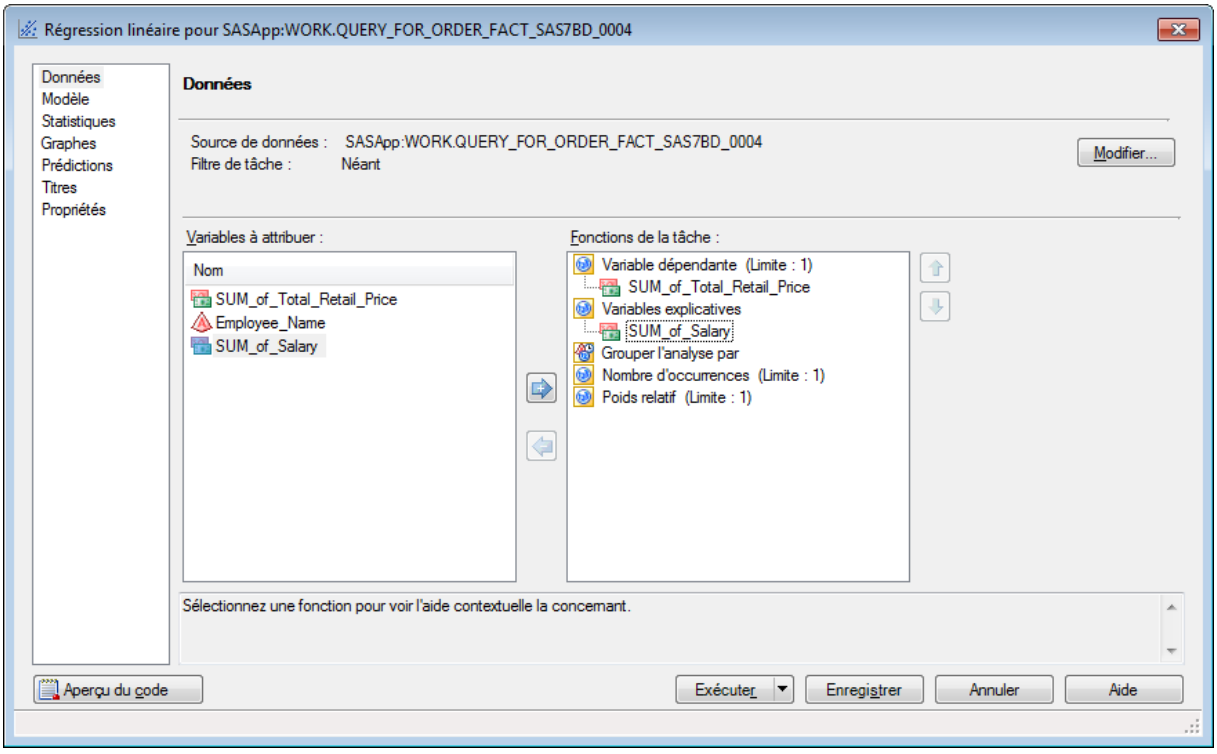
Sélectionner les colonnes :

- **Employee_Name**
- **Total_retail_price** (Prendre la somme dans la colonne Récapitulatif)
- **Employee_salary** (Prendre la somme dans la colonne Récapitulatif)

Dans l'onglet filtrer les données, faire un filtre de tel sorte que le sexe de l'employé ne soit pas manquant.



Exécuter
Dans le menu **tâche → Régression → Modèle linéaire**



Affecter les variables aux fonctions comme ci-dessus.
Exécuter

Résultats de la régression linéaire

Procédure REG
Modèle : Linear_Regression_Model
Variable dépendante : SUM_of_Total_Retail_Price

Nombre d'observations lues	104
Nombre d'observations utilisées	104

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	7.337173E11	7.337173E11	843.51	<.0001
Erreur	102	88723701815	869840214		
Total sommes corrigées	103	8.22441E11			

Root MSE	29493	**R carré**	0.8921
Moyenne dépendante	115845	**R car. ajust.**	0.8911
Coeff Var	25.45905		

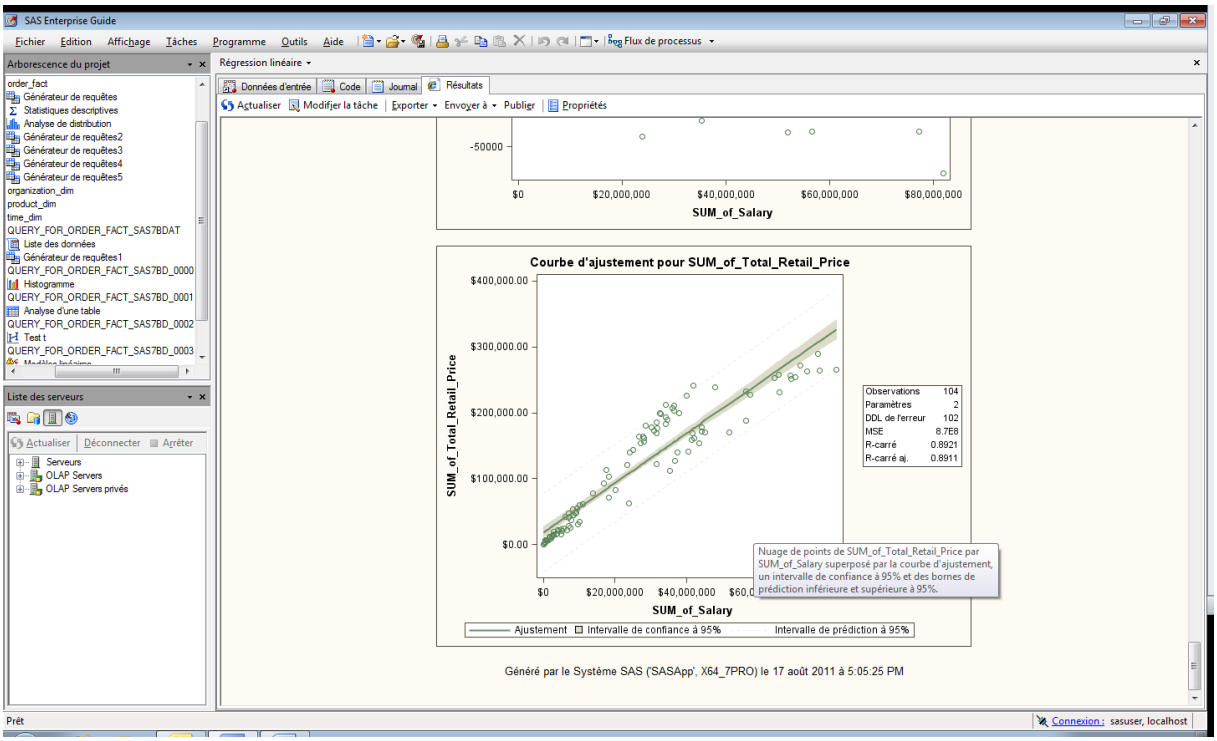
Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	**1**	18471	4427.72539	4.17	<.0001
SUM_of_Salary	**1**	0.00376	0.00012957	29.04	<.0001

Il faut lire dans les résultats :
La valeur du R-Square est de 0.8921indiquant que la variable SUM_OF_Salary explique 89.21% de la variable SUM_OF_Total_Retail_Price.

L'équation de la regression est :
SUM_OF_Total_Retail_Price = 18471 + 0.00376 * SUM_OF_Salary

18471 est la valeur de **Intercept**



Prenons un peu de recul : L'analyse que nous venons de faire est absolument idiote. Elle ne sert à rien.
En effet, en calculant la somme du chiffre d'affaires et la somme du salaire ; par salarié, on a multiplié le salaire par autant de vente qu'a fait ce collaborateur. Donc plus le vendeur fait de vente, plus son chiffre d'affaires est important et la multiplication de son salaire aussi. Cela ne veut rien dire.
Pour faire ce type d'analyse, la table aurait dû par exemple être créée en faisant un filtre sur une année, et en calculant la somme du chiffre d'affaire sur une année donnée ; puis en l'analysant par rapport au salaire, pour chaque collaborateur.

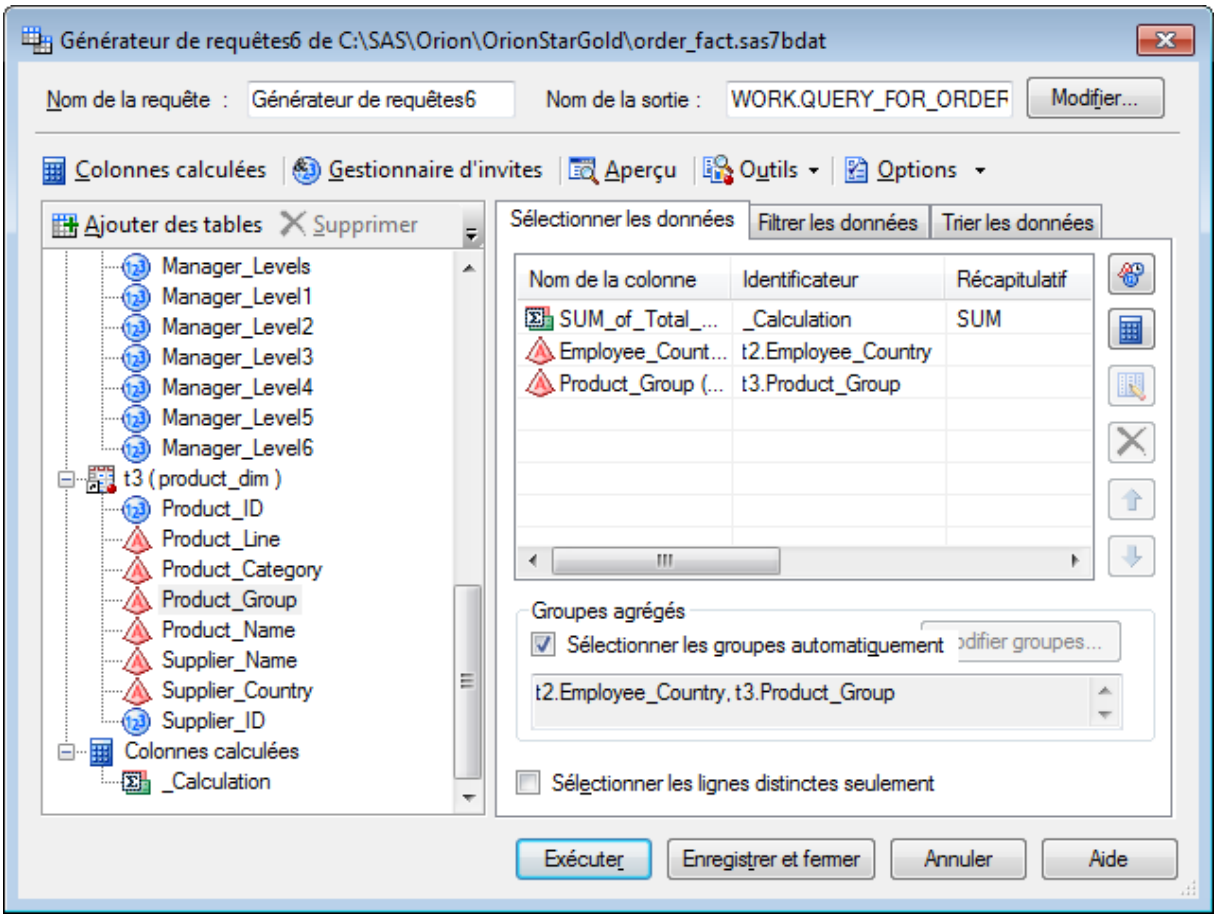
Analyses multivariées

Objectif créer la table de la somme du chiffre d'affaires triée par pays et par groupe de produit.

Pour revenir sur la fenêtre du projet, cliquer sur le bouton **flux de processus**

Clic-droit sur la table **Order_Fact** → **Générateur de requête**

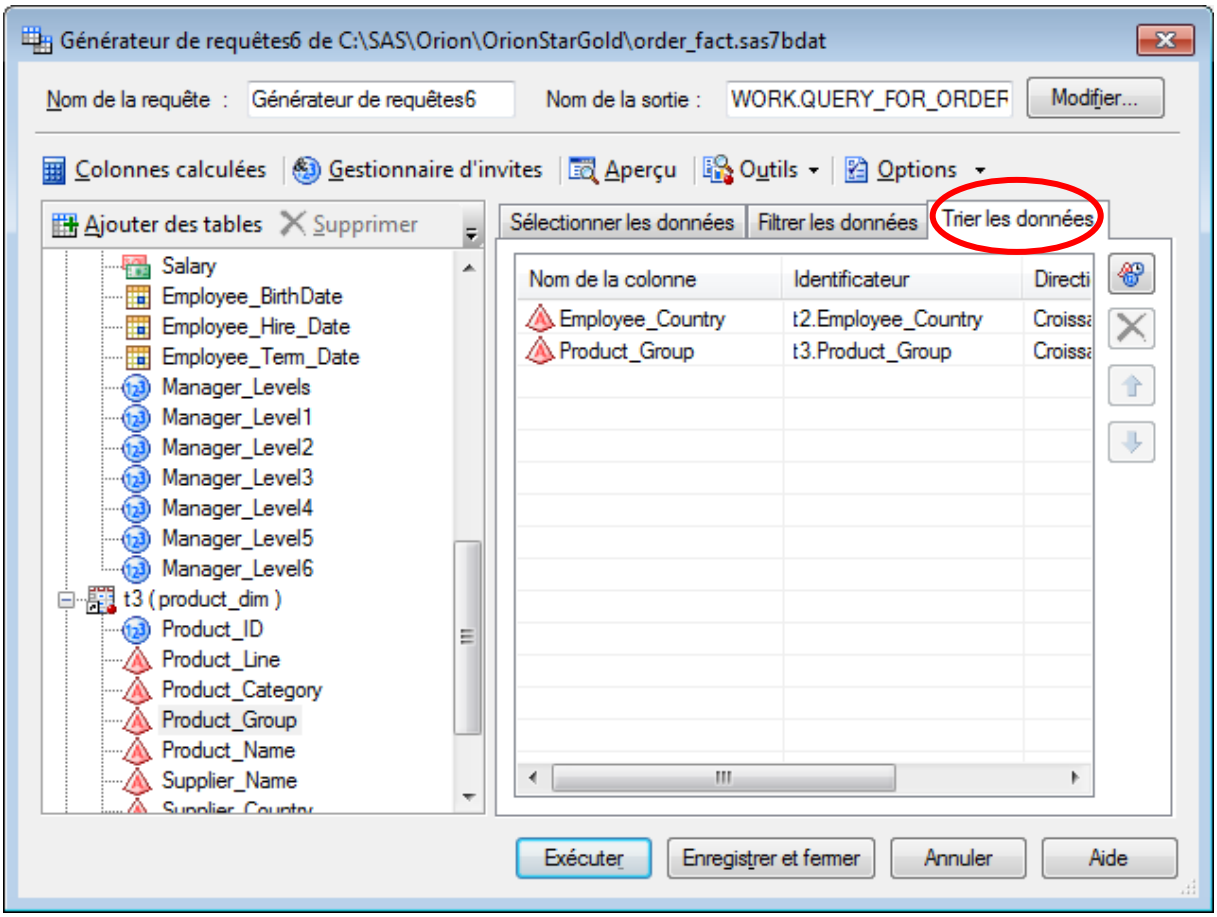
Ajouter les tables **Organization_Dim** et **Product_Dim**.



Sélectionner les trois colonnes suivantes par glisser-lâcher :

Total_retail_price
Employee_country
Product_group

Sélectionner la somme (**SUM**) du chiffre d'affaires.
Filtrer les employées dont le sexe n'est pas manquant.

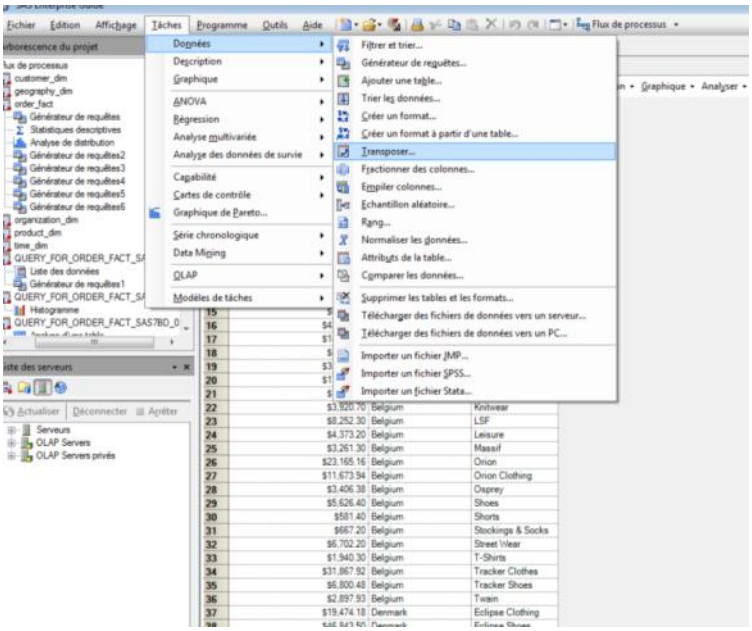


Dans l'onglet trier les données, sélectionner par glisser-lâcher les colonnes :

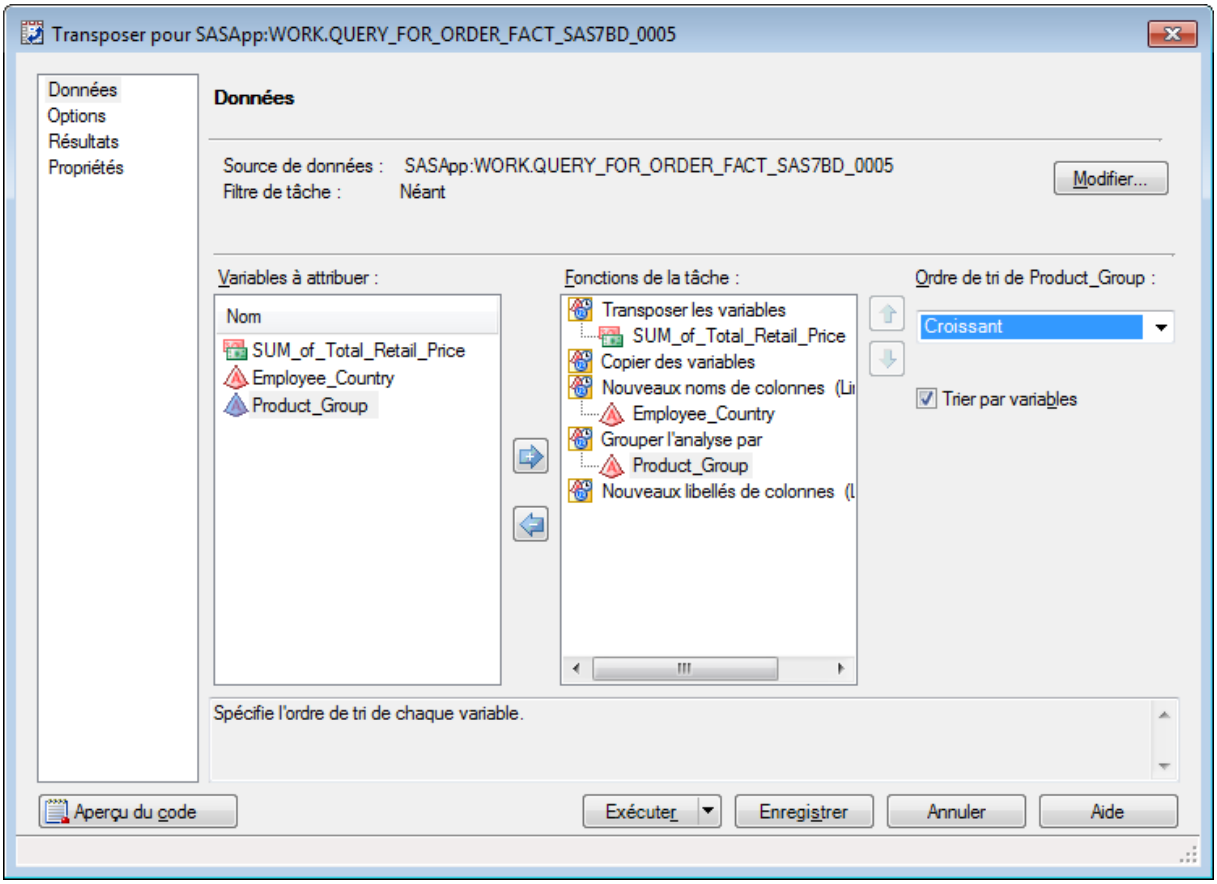
Employee_country
Product_group

Exécuter la requête.

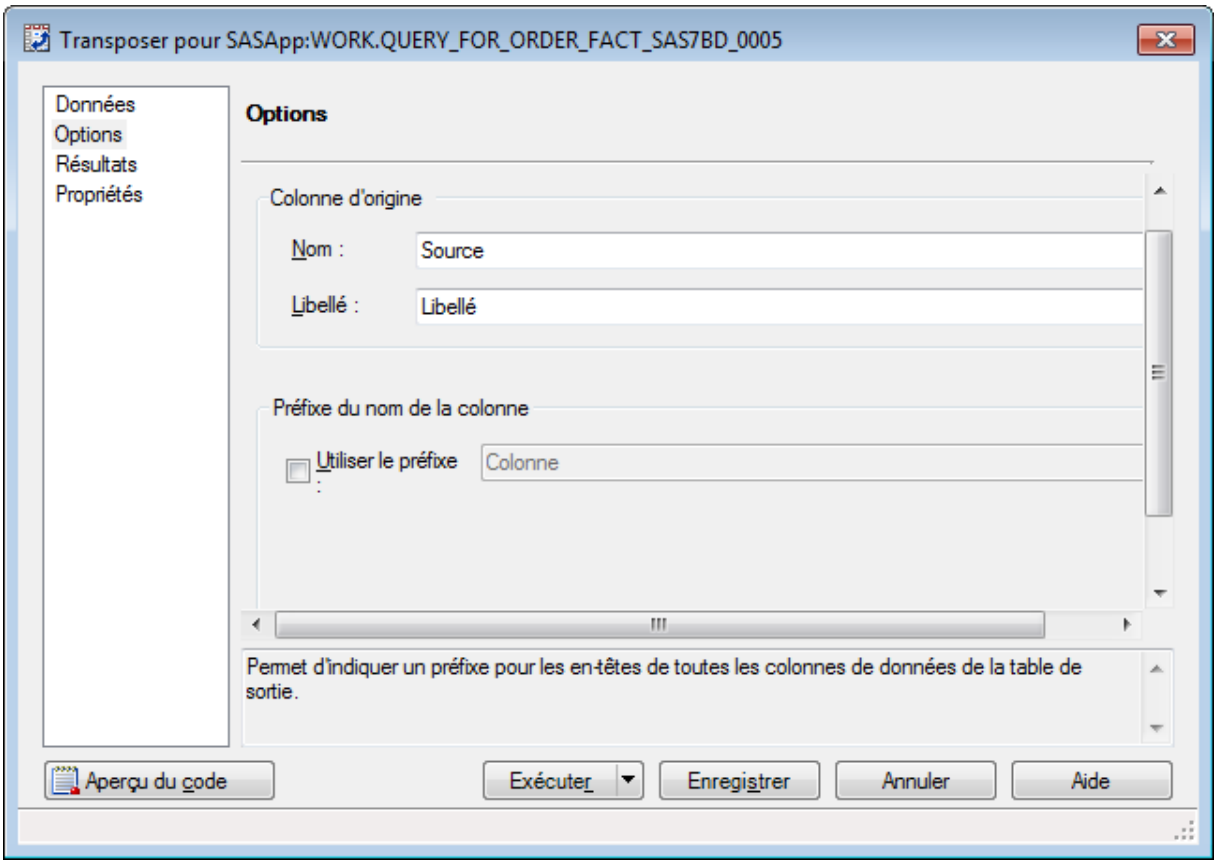
Transposer une colonne



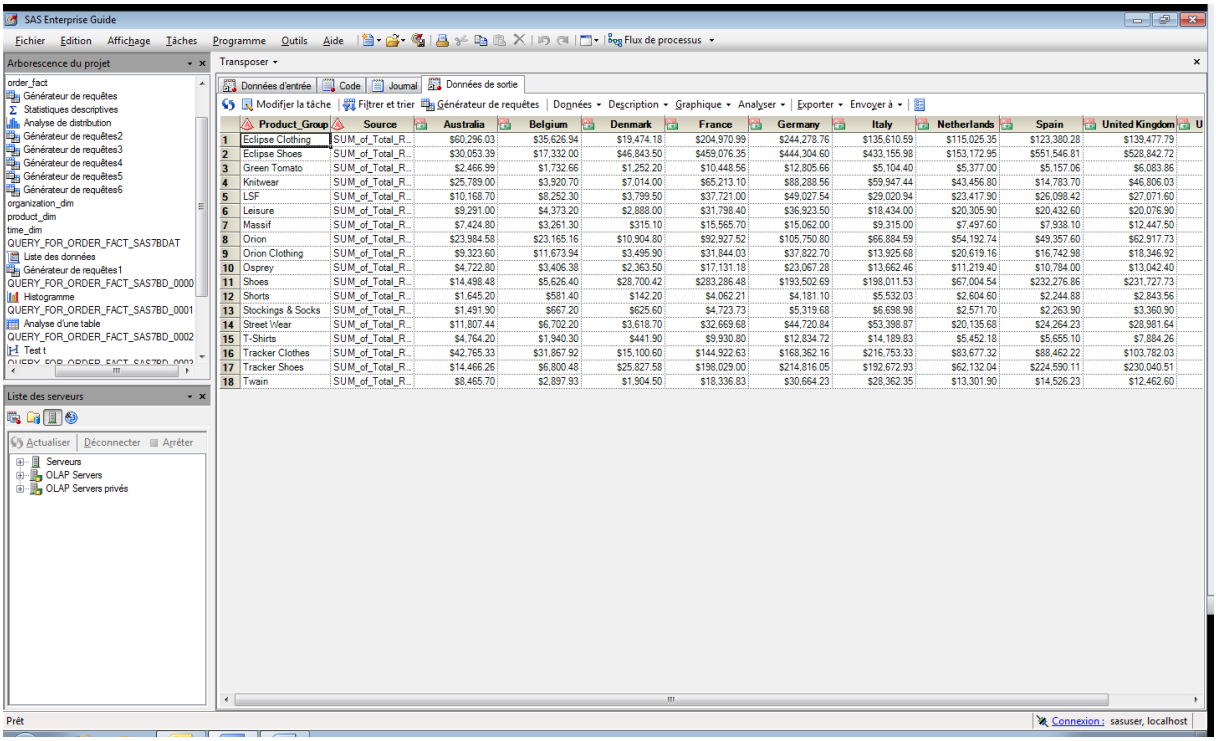
Dans le menu des **tâches → Données → Transposer**



Affecter la colonne **SUM_of_Total_Retail_Price** à la **variable à transposer.**
Affecter la colonne **Employee_country** au **nouveau nom de colonne**.
Affecter la colonne **Product_Group** à **grouper l'analyse par**.

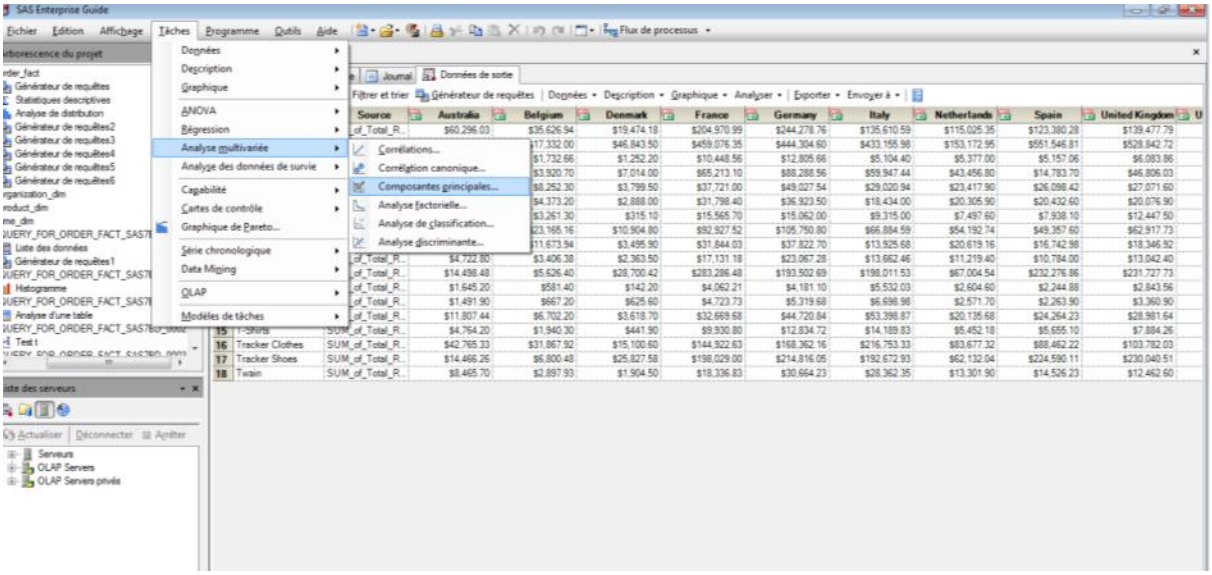


Dans l'onglet Option, désélectionner **utiliser le préfixe**.
Exécuter

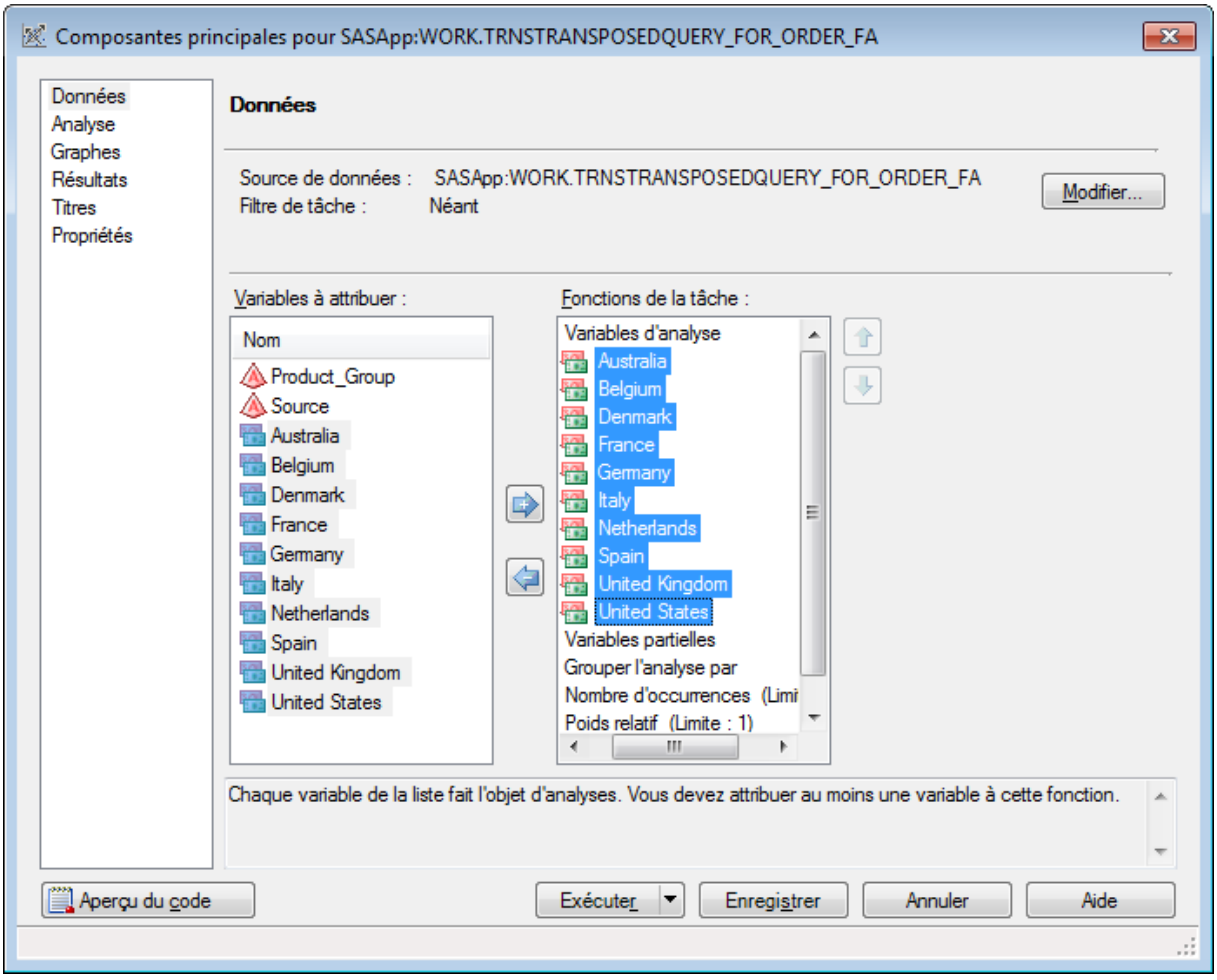


	Product_Group	Source	Australia	Belgium	Denmark	France	Germany	Italy	Netherlands	Spain	United Kingdom
1	Eclipse Clothing	SUM_of_Total_R...	$60,296.03	$35,626.94	$19,474.18	$204,970.99	$244,278.76	$135,610.59	$115,025.35	$123,380.28	$139,477.79
2	Eclipse Shoes	SUM_of_Total_R...	$30,053.39	$17,332.00	$46,843.50	$459,076.35	$444,304.60	$433,155.98	$153,172.95	$551,546.81	$528,842.72
3	Green Tomato	SUM_of_Total_R...	$2,466.99	$1,732.66	$1,252.20	$10,448.56	$12,805.66	$5,104.40	$5,377.00	$5,157.06	$6,083.86
4	Knitwear	SUM_of_Total_R...	$25,789.00	$3,920.70	$7,014.00	$65,213.10	$88,288.56	$59,947.44	$43,456.80	$14,783.70	$46,806.03
5	LSF	SUM_of_Total_R...	$10,168.70	$8,252.30	$3,799.50	$37,721.00	$49,027.54	$29,020.94	$23,417.90	$26,098.42	$27,071.60
6	Leisure	SUM_of_Total_R...	$9,291.00	$4,373.20	$2,888.00	$31,798.40	$36,923.50	$18,434.00	$20,305.90	$20,432.60	$20,076.90
7	Massif	SUM_of_Total_R...	$7,424.80	$3,261.30	$315.10	$15,565.70	$15,062.00	$9,315.00	$7,497.60	$7,938.10	$12,447.50
8	Orion	SUM_of_Total_R...	$23,984.58	$23,165.16	$10,904.80	$92,927.52	$105,750.80	$66,884.59	$54,192.74	$49,357.60	$62,917.73
9	Orion Clothing	SUM_of_Total_R...	$9,323.60	$11,673.94	$3,495.90	$31,844.03	$37,822.70	$13,925.68	$20,619.16	$16,742.98	$18,346.92
10	Osprey	SUM_of_Total_R...	$4,722.80	$3,406.38	$2,363.50	$17,131.18	$23,067.28	$13,662.46	$11,219.40	$10,784.00	$13,042.40
11	Shoes	SUM_of_Total_R...	$14,498.48	$5,626.40	$28,700.42	$283,286.48	$193,502.69	$198,011.53	$67,004.54	$232,276.86	$231,727.73
12	Shorts	SUM_of_Total_R...	$1,645.20	$581.40	$142.20	$4,062.21	$4,181.10	$5,532.03	$2,604.60	$2,244.88	$2,843.56
13	Stockings & Socks	SUM_of_Total_R...	$1,491.90	$667.20	$625.60	$4,723.73	$5,319.68	$6,698.98	$2,571.70	$2,263.90	$3,360.90
14	Street Wear	SUM_of_Total_R...	$11,807.44	$6,702.20	$3,618.70	$32,669.68	$44,720.84	$53,398.87	$20,135.68	$24,264.23	$28,981.64
15	T-Shirts	SUM_of_Total_R...	$4,764.20	$1,940.30	$441.90	$9,930.80	$12,834.72	$14,189.83	$5,452.18	$5,655.10	$7,884.26
16	Tracker Clothes	SUM_of_Total_R...	$42,765.33	$31,867.92	$15,100.60	$144,922.63	$168,362.16	$216,753.33	$83,677.32	$88,462.22	$103,782.03
17	Tracker Shoes	SUM_of_Total_R...	$14,466.26	$6,800.48	$25,827.58	$198,029.00	$214,816.05	$192,672.93	$62,132.04	$224,590.11	$230,040.51
18	Twain	SUM_of_Total_R...	$8,465.70	$2,897.93	$1,904.50	$18,336.83	$30,664.23	$28,362.35	$13,301.90	$14,526.23	$12,462.60

Analyse en Composantes Principales



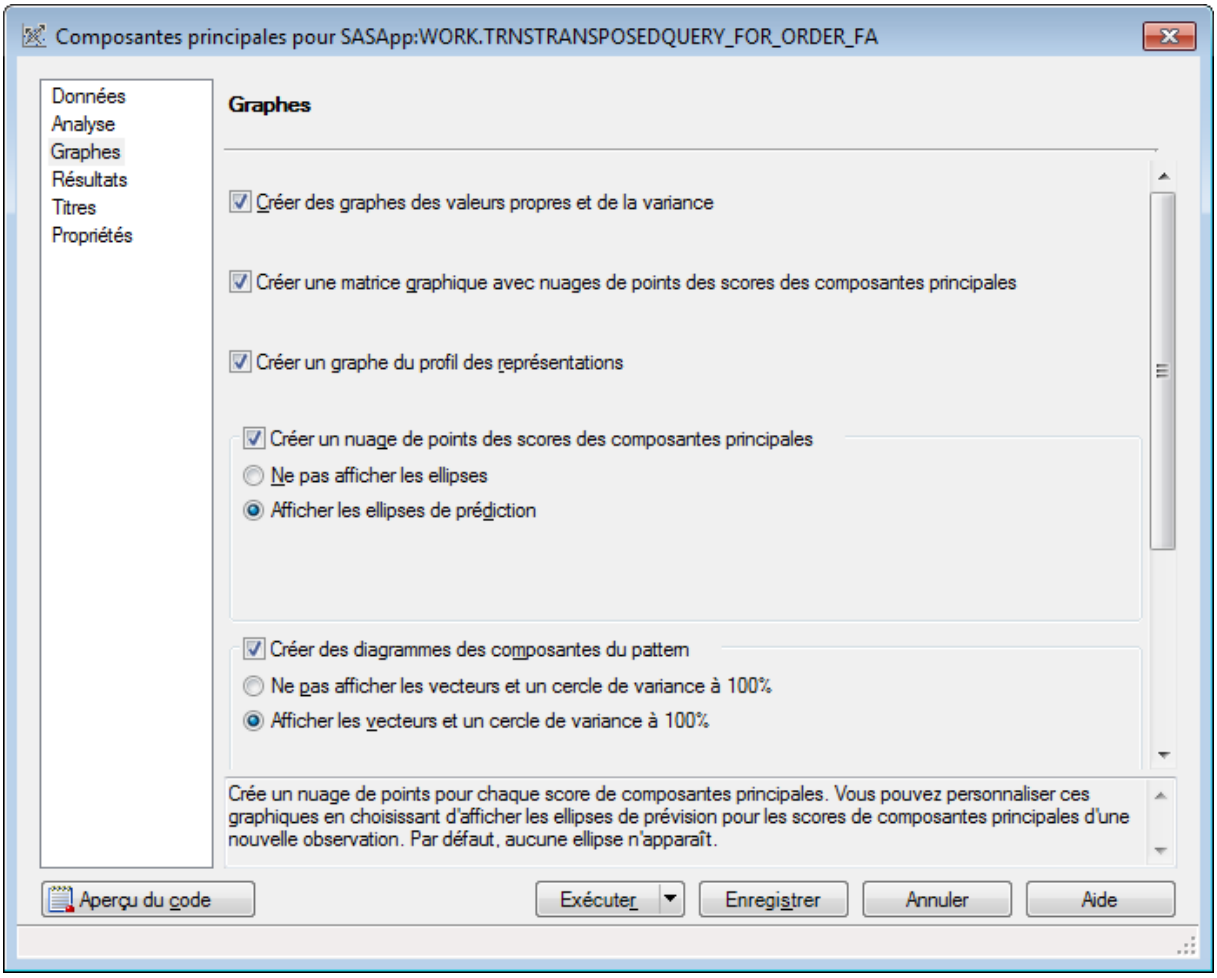
Dans le menu des **tâches →Analyse → Analyse multivariée → Composantes principales**

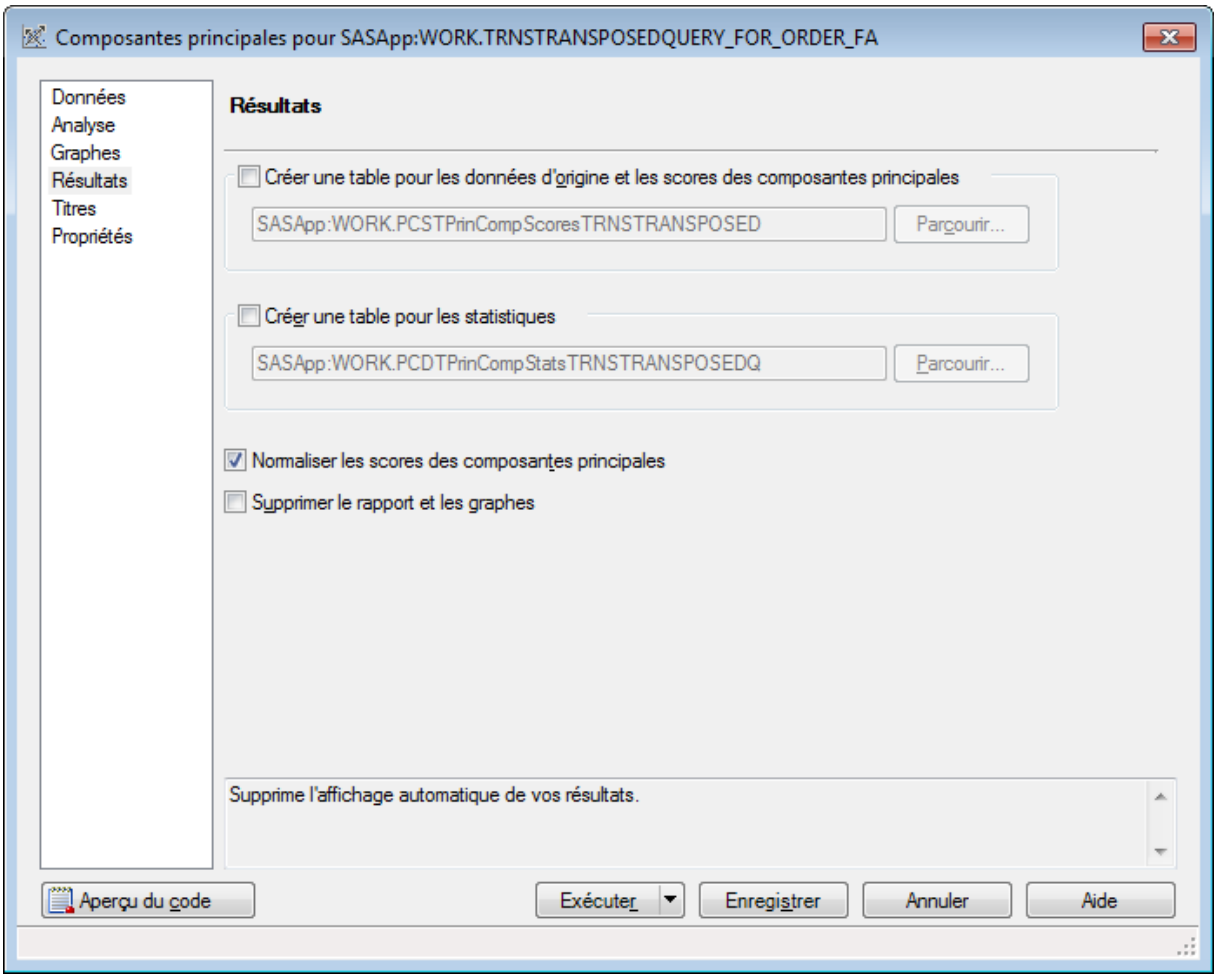


Affecter toutes les variables du chiffre d'affaires par pays aux variables d'analyses.

Dans l'onglet Graphes

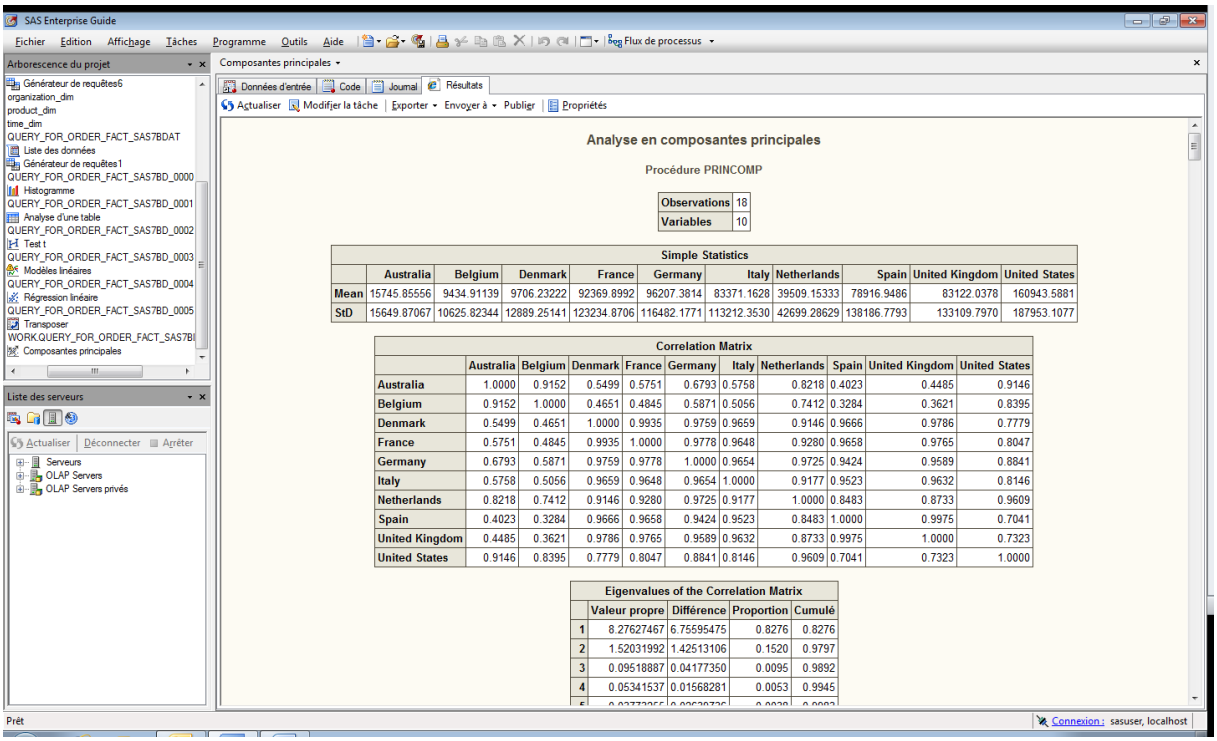
Sélectionner tout





Dans l'onglet des résultats
Sélectionner **Normaliser les scores des composantes principales.**

Exécuter.



Analyse en composantes principales

Procédure PRINCOMP

Observations	18
Variables	10

Simple Statistics

	Australia	Belgium	Denmark	France	Germany	Italy	Netherlands	Spain	United Kingdom	United States
Mean	15745.85556	9434.91139	9706.23222	92369.8992	96207.3814	83371.1628	39509.15333	78916.9486	83122.0378	160943.5881
StD	15649.87067	10625.82344	12889.25141	123234.8706	116482.1771	113212.3530	42699.28629	138186.7793	133109.7970	187953.1077

Correlation Matrix

	Australia	Belgium	Denmark	France	Germany	Italy	Netherlands	Spain	United Kingdom	United States
Australia	1.0000	0.9152	0.5499	0.5751	0.6793	0.5758	0.8218	0.4023	0.4485	0.9146
Belgium	0.9152	1.0000	0.4651	0.4845	0.5871	0.5056	0.7412	0.3284	0.3621	0.8395
Denmark	0.5499	0.4651	1.0000	0.9935	0.9759	0.9659	0.9146	0.9666	0.9786	0.7779
France	0.5751	0.4845	0.9935	1.0000	0.9778	0.9648	0.9280	0.9658	0.9765	0.8047
Germany	0.6793	0.5871	0.9759	0.9778	1.0000	0.9654	0.9725	0.9424	0.9589	0.8841
Italy	0.5758	0.5056	0.9659	0.9648	0.9654	1.0000	0.9177	0.9523	0.9632	0.8146
Netherlands	0.8218	0.7412	0.9146	0.9280	0.9725	0.9177	1.0000	0.8483	0.8733	0.9609
Spain	0.4023	0.3284	0.9666	0.9658	0.9424	0.9523	0.8483	1.0000	0.9975	0.7041
United Kingdom	0.4485	0.3621	0.9786	0.9765	0.9589	0.9632	0.8733	0.9975	1.0000	0.7323
United States	0.9146	0.8395	0.7779	0.8047	0.8841	0.8146	0.9609	0.7041	0.7323	1.0000

Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulé
1	8.27627467	6.75595475	0.8276	0.8276
2	1.52031992	1.42513106	0.1520	0.9797
3	0.09518887	0.04177350	0.0095	0.9892
4	0.05341537	0.01568281	0.0053	0.9945

Conclusion sur les statistiques

Nous avons vu rapidement les statistiques suivantes :

1. Statistiques descriptives de base (moyenne, médiane et écart type) et de distribution (asymétrie et aplatissement).
2. Test du chi² : Nous avons seulement utilisé le test du chi² pour analyser l'indépendance entre deux variables avec quelques modalités chacune.
3. Tests en t : Il y a-t-il une différence significative entre la moyenne de deux échantillons ?
4. Anova : Modèle linéaire : Y a-t-il une relation entre une variable à expliquer et des variables explicative ou de classification ?

Pour plus d'information, voir le site suivant qui est très intuitif : http://support.sas.com/learn/statlibrary/.

ETL

L'objectif de ce chapitre est de présenter la planification d'un projet de Data Warehouse, et la création de flux ETL : extraction, transformation et chargement. Un processus ETL se décompose en cinq phases :

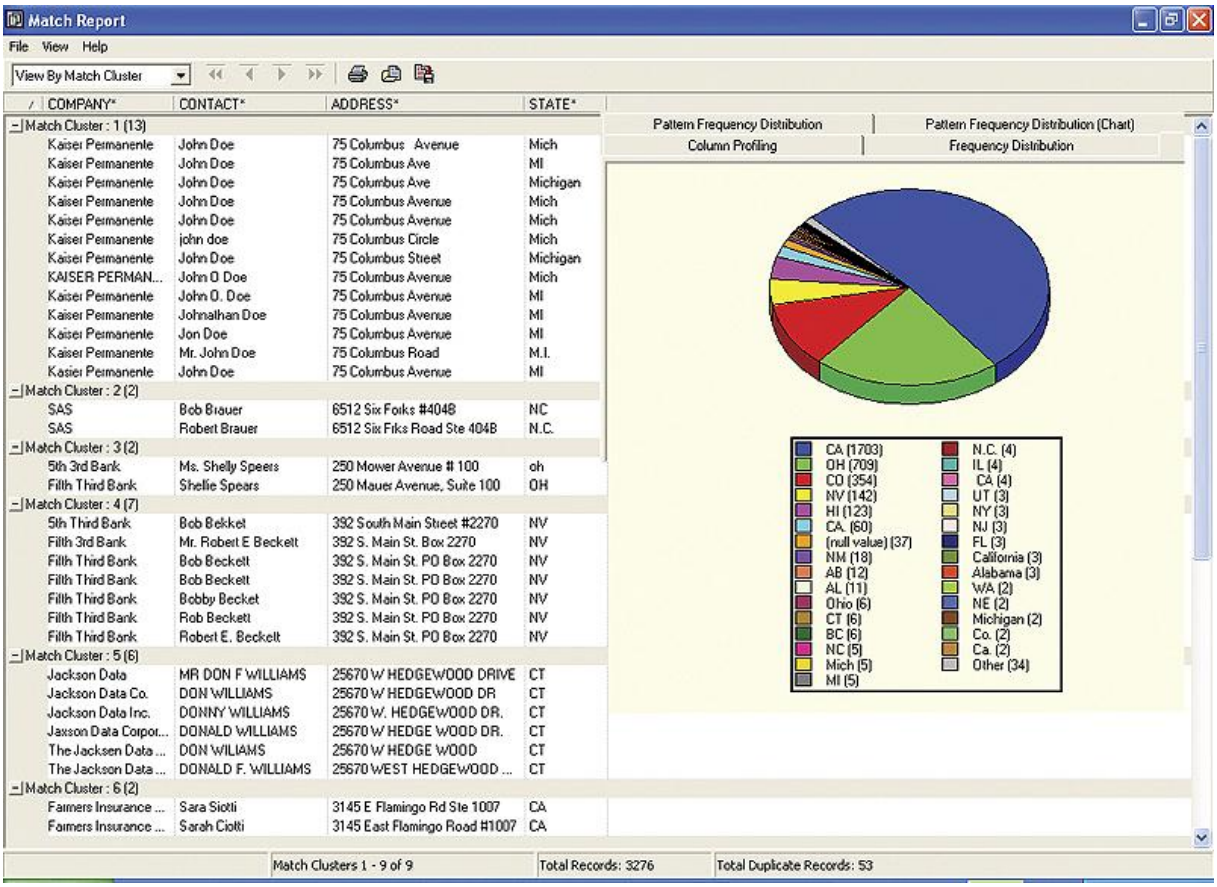
1. Définition des métadonnées sources : description des bases de données opérationnelles.
2. Définition des métadonnées cibles : description des futures bases du Data Warehouse.
3. Validation de la qualité des données
4. Création du processus qui permet de charger la cible en utilisant les données sources.
5. Ordonnancement.

La génération d'un processus ETL est au premier abord un simple problème technique. Mais c'est souvent après quelques années que l'on s'aperçoit de l'importante nécessité d'une méthodologie rigoureuse. Nous allons donc aborder ce chapitre en suivant une méthodologie qui peut paraître lourde, mais qui est nécessaire pour faire évoluer le système décisionnel tout au long de sa, si possible, très longue vie.

Ce chapitre s'adresse tout principalement aux:

Utilisateurs de SAS Data Integration Studio, appelé dans le jargon « ETL'istes ». Personne ayant pour mission de générer les processus d'extraction, transformation et chargement des entrepôts de données. Des connaissances approfondies en informatique et plus particulièrement en base de données et en programmation BASE/SAS sont pré-requises. En effet, comme nous allons le voir dans ce chapitre, SAS Data Integration Studio génère du code SAS/BASE, comprenant du code SAS/MACRO et des procédures SQL. Même si les fonctions de transformations implémentées dans SAS Data Integration Studio couvrent une large gamme, il est très souvent nécessaire de coder certaines transformations. Pour les étudiants se destinant à cette fonction, la certification « SAS Certified Base Programming » est un plus conséquent. La certification pour développeur ETL avec SAS Data Integration Studio est : Certification Data Integration SAS®.
L'une des forces de la plateforme SAS, est que l'on peut toujours revenir à la base, au code SAS.

Problème métier :
Dans un monde où la concurrence est de plus en plus forte, des décisions doivent être prises sur des informations claires, extraites de données nettoyées et intégrées. Depuis les cadres supérieurs jusqu'aux ouvriers, l'information est nécessaire pour prendre des décisions tactiques et stratégiques. Il est prouvé que des milliards se perdent chaque année à cause de données de mauvaise qualité, erronées, etc. Derrière la mauvaise affectation des ressources, ce sont des clients insatisfaits, une érosion de la crédibilité et l'incapacité à prendre la bonne décision au bon moment.
La construction du Data Warehouse est le socle du processus décisionnel. L'expression « garbage in, garbage out » est ici particulièrement vraie ; si vous construisez votre *Data Warehouse* avec des données de mauvaise qualité, vous aurez des rapports, des analyses, des simulations, des prévisions, des optimisations fausses. L'objectif du décisionnel est d'apporter à tous les utilisateurs, une information de qualité ; il est donc nécessaire d'évaluer, d'analyser et de corriger si besoin, l'information, dès le départ, dans le processus ETL.
La société SAS a racheté la société Data Flux en 2000 et a intégré ces fonctions de gestion de la qualité des données.



Voici une copie de l'interface principale DFPower Studio. L'intégration avec SAS Data Integration Studio se fait via la base de connaissance (KB ou Knowledge Base).
Les deux fonctions majeures dans le processus de qualité des données sont d'abord l'analyse puis la normalisation ou standardisation.

Exemple d'entreprise 1 : suite à l'analyse des différents patterns des numéros de téléphone, nous avons constaté qu'il y avait dans cette base de données, 32 façons de saisir le numéro de téléphone pour la France. Voici quelques exemples :

+33 1 60 62 12 19
(+33) 1 60 62 12 19
+33 (0)1 60 62 12 19
01 60 62 12 19
00 33 (0)1 60 62 12 19
00 33 1 60 62 12 19
0160621219
01.60.62.12.19
…

On génère alors un processus de standardisation qui en récoltant les neufs derniers entiers, les placent dans la forme normalisée.

Exemple d'entreprise 2 : Référentiel pour le Data Warehouse d'une société française. Grâce à l'aide précieuse d'un expert, un dictionnaire a été créé, comprenant les différents noms du même produit, avec les dates de début, voire de fin, et les domaines d'utilisation des noms. En effet, il y avait de quoi se perdre et fournir des rapports hétérogènes entre le nom utilisé par l'acheteur qui utilise la dénomination du fournisseur, les différents noms définis par le marketing pour tenter de booster les ventes, et les changements de versions au département R&D et à la production.

Exemple d'entreprise 3 : c'est l'un des exemples les plus courants : le dé-doublonnage d'adresse. Il faut détecter le nom, le prénom, les adresses, le code postal et la ville.

Les noms et prénoms peuvent être inversés ou manquants.
Dans l'adresse, on retrouve par exemple souvent le cas de la rue, du boulevard, de Strasse, de Street, etc. écrient en entier ou abrégé (Bd, boulevard ; r, rue, etc…) avec le numéro avant ou après, avec une virgule ou pas.

Les deux adresses suivantes sont identiques :
"Eric Martin
33 Av Victor Hugo
75000 Paris"

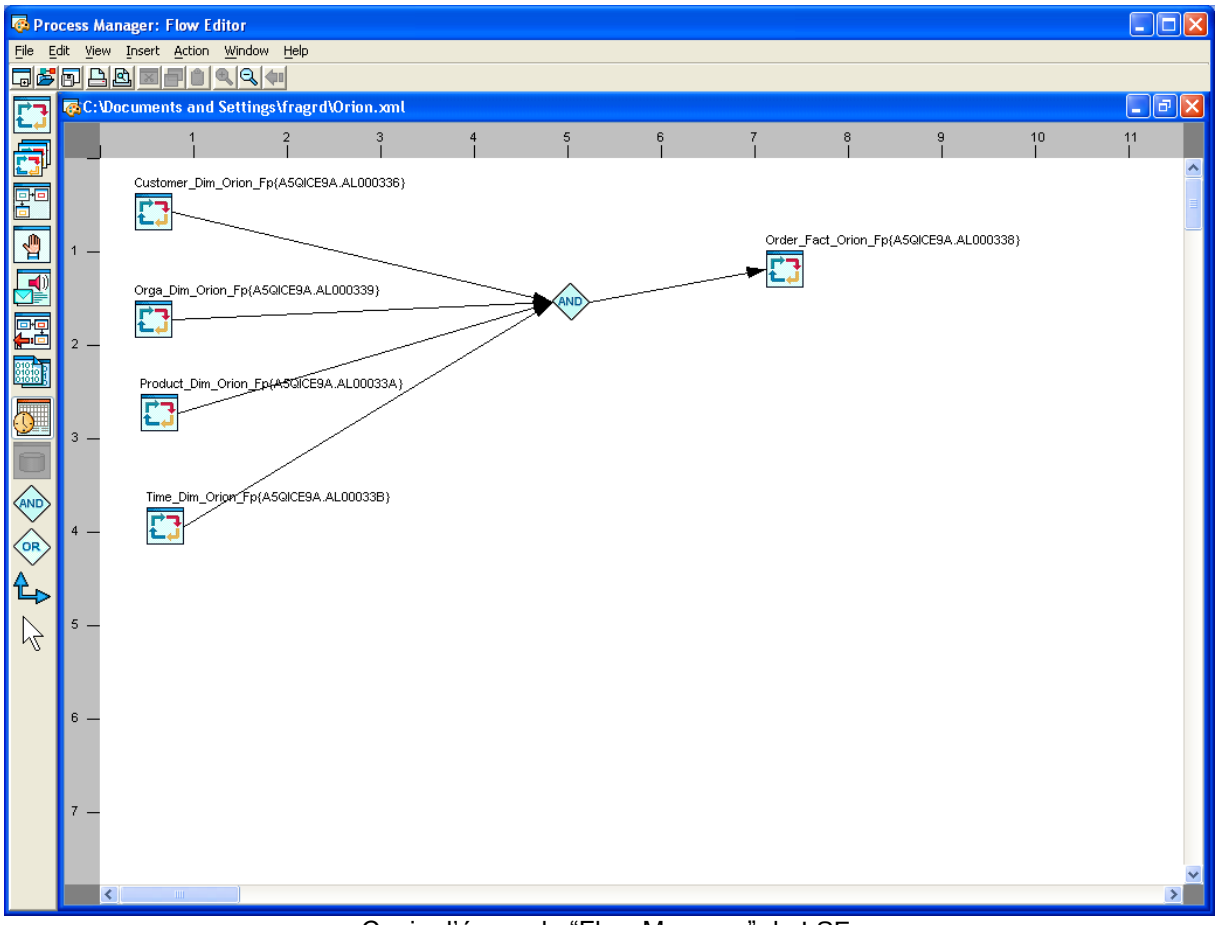
"MARTIN Eric
33, avenue Victor Hugo
75O16 PARIS"

Présentation de l'outil SAS® Data Integration Studio

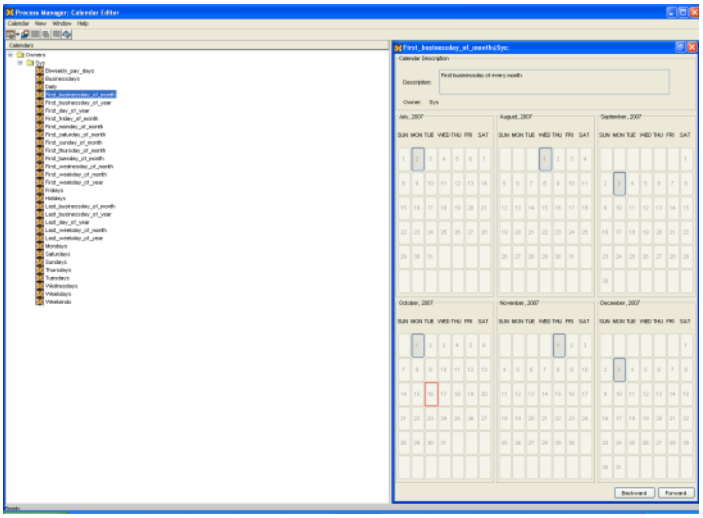
SAS® Data Integration Studio est une interface cliente Java de la plateforme décisionnelle SAS. Elle fait partie des packages « SAS Data Integration Server » et « SAS Enterprise Data Integration Server ».
SAS Data Integration Server comprend notamment SAS/Base, SAS/CONNECT, SAS Data Integration Studio.

SAS Enterprise Data Integration Server comprend notamment SAS/Base, SAS/CONNECT, SAS/SHARE, 2 SAS/Access au choix, SAS Data Integration Studio, SAS Data Quality Server et LSF.

LSF est un Ordonnanceur permettant, comme son nom l'indique, d'ordonnancer des flux de processus.
SAS Data Integration Studio permet de créer des flux de processus, qui une fois créés, vont être déployés pour l'ordonnancement.
On arrive alors dans l'opérationnel du décisionnel, c'est-à-dire, l'industrialisation de processus décisionnels. Dès que quelque chose se passe dans telle base de données, toutes les 10 minutes, les demi-heures, les demi-journées, les jours, les semaines, les mois, les trimestres, les années, l'ordonnanceur doit lancer des processus ETL, des créations et des mises à jour de tables, de cubes, des créations, des envois, par mail, SMS, sur un portail de rapports ; des processus de Data Mining, etc…
La copie d'écran du "Flow Manager" de LSF sur la page suivante montre qu'il faut mettre à jour toutes les tables de dimension avant de pouvoir mettre à jour la table de fait (car celle-ci comprends des clés étrangères sur les clés primaires des tables de dimension).



Copie d'écran du "Flow Manager" de LSF



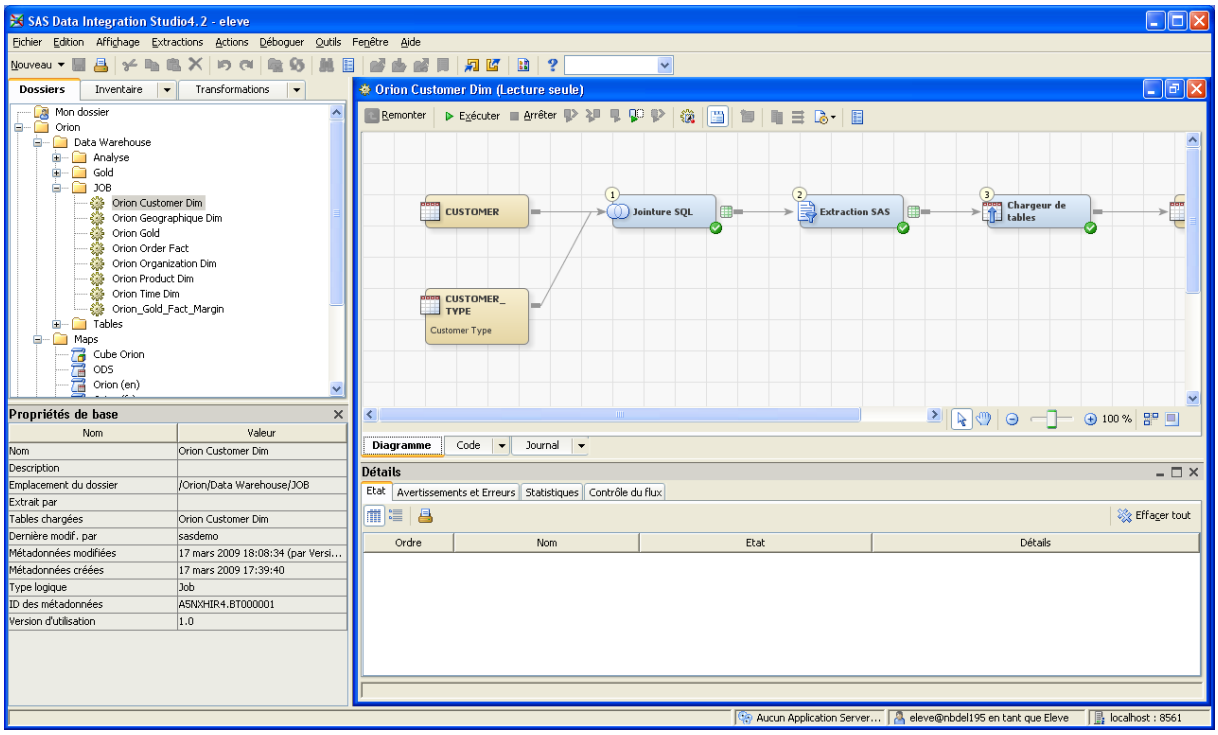
Copie d'écran du « Calendar Editor » de LSF

A propos de SAS® Data Integration Studio, notons que l'outil :

- Support des développements concurrents de développeur ETL, gestion des versions via le « Check-Out/Check-In ». Dans un projet ETL, l'administrateur de la plateforme décisionnelle crée un référentiel de métadonnées dit « référentiel projet » pour chaque développeur ETL. Ceux-ci peuvent alors extraire (Check-out) du référentiel principal, dans leur référentiel dédié, l'élément qu'il souhaite modifier. Les autres utilisateurs ne peuvent pas modifier les éléments extraits par d'autre. Une fois les modifications faites, l'utilisateur réintègre (Check-in) ces éléments dans le référentiel principal, avec la version N+1. Le Check-Out/Check-In permet

donc à plusieurs développeurs ETL de travailler sans conflit sur un même projet et il permet la gestion des versions.

- Permet un développement administré et donc sécurisé, ce qui est nécessaire pour la gestion du changement.
- Permet d'accéder à l'ensemble des systèmes de gestion de bases de données du marché, via les différents SAS/ACCESS to ... (SAS/ACCESS to Oracle, SAS/ACCESS to DB2, SAS/ACCESS to Terradata, etc.)
- Permet l'intégration avec plusieurs ERP du marché comme par exemple SAP. Le terme intégration prend ici toute sa valeur, on n'extrait pas seulement de la donnée, on accède à la métadonnée de l'application de l'entreprise, ce qui permet une véritable communication bidirectionnelle. Mais surtout, le fait d'accéder à la structure globale de l'application opérationnelle, permet dans le cas de solution métier, un raccourcissement considérable du temps de mise en place du projet. Cela fonctionne particulièrement bien avec SAP qui a la réputation d'être très bien structuré (et structurant). Un « SAS/Adapter to . » permet l'intégration avec les différents grands progiciels du marché.
- Support les modèles de donnée du standard CWM. Common Warehouse Model, selon le site web www.cwm.org, CWM est une norme pour les outils décisionnels permettant de transférer des fichiers de métadonnées entre différentes applications décisionnelles de différents éditeurs de logiciel. Par exemple, intégrer une l'application de Reporting Business Object® à la plateforme SAS®.
- Il est possible d'ajouter des plug-in java pour compléter l'ensemble des fonctionnalités, de créer rapidement ses propres outils ETL.
- Permet la construction de Cube (vous devez avoir une licence pour le package « SAS® Intelligence Storage » ou SAS OLAP Serveur)



La planification du *Data Warehouse* est une phase cruciale qui nécessite donc une certaine rigueur. Pour le bon déroulement de celle-ci, je vous propose la liste suivante de points importants qui est la concaténation de différents éléments récupérés à divers endroits. Ce n'est pas une méthodologie à proprement parler, car l'objectif ici est plus d'avoir une tête bien faite que bien pleine. C'est-à-dire que je préfère que vous sentiez la philosophie du décisionnelle plutôt que de vous donner une méthodologie à appliquer bêtement. De plus il n'y a pas de méthodologie parfaite si l'on n'a pas saisi la philosophie du problème pour pouvoir l'adapter. Enfin, les deux livres *Entrepôts de données. Guide pratique de modélisation dimensionnelle, 2ème édition*, et Le *Data Warehouse : Guide de conduite de projet* de Ralph Kimball, aux éditions Eyrolles ; sont toujours d'actualité et couvrent remarquablement bien ces deux thèmes.

1. Définition des besoins.
    a. Création de la liste des questions métiers auxquelles devra répondre le projet décisionnel. Pour élaborer cet inventaire, la maîtrise d'ouvrage doit interroger les différents utilisateurs potentiels pour récolter leur besoin. Pour chaque indicateur, il est très important d'avoir une définition précise de sa formule (règle de calcul), une description de son interprétation et de son utilité. Si vous demandez à quelques personnes autour de vous, la définition du turnover en ressources humaines, presque tous pensent la connaître, ou tout du moins ont une idée sur le sujet, mais l'on arrive très vite à soulever des différences. Parle-t-on du turnover volontaire ou bien du taux de renouvellement global ? Quel sont les contrats pris en compte ? CDD, CDI ? En effet, ce turnover peut être le nombre de salariés en CDI démissionnant, divisé par le nombre moyen de salariés total dans l'entreprise, sur l'année. Ce peut aussi être le nombre de fin de contrat (fins de CDD, démissions, départs à la retraite, licenciements, etc.) sur le nombre de salariés total en début d'année. Suivant les éléments au numérateur et au dénominateur de ce rapport, les résultats et les interprétations peuvent facilement varier. Mais le plus important est peut-être de définir l'utilité de chaque indicateur et donc les moyens d'action potentiels à mettre en œuvre pour corriger le tir si cet indicateur tourne au rouge.
    b. Définition de l'ensemble des dimensions qui devront être prise en compte afin de répondre aux questions précédemment listées. Est-ce que le chiffre d'affaire doit être analysé par produits, par clients, par région, par période de temps. En posant ces questions : quels indicateurs, suivant quels axes ; les différents schémas en étoile ou cubes multidimensionnels se dessinent progressivement.
    c. Définition de qui utilisera ces informations et quand elles seront utilisées, afin de définir le nombre de requêtes simultanées.
    d. Définition de la fréquence de mise à jour des informations, en pseudo-temps réel, tous les jours, toutes les semaines, tous les mois, tous les trimestres, toutes les années.
    e. Définition de la durée de vie des données dans le Data Warehouse, notamment les données de détail ne doivent pas nécessairement rester ad vitam et aeternam dans le Data Warehouse.
    f. Définir les indicateurs clés et leur ordre de priorité pour les mettre en place. Lors de la collecte des indicateurs souhaités, on obtient facilement une liste très longue. La construction d'un Data Warehouse se déroule de manière itérative et incrémentale. L'idée est de sélectionner les indicateurs à implémenter en priorité. Il y a souvent des « pépites », des indicateurs où la maîtrise d'ouvrage se doute fortement que leur implémentation n'est pas très compliqué et peut rapporter beaucoup. C'est pépites permettront de rapidement justifier de la valeur ajoutée d'un Data Warehouse.
2. Faire l'inventaire des données nécessaires à la construction du Data Warehouse.
    a. Lister l'ensemble des systèmes et des bases de données, auxquelles il sera nécessaire de se connecter pour l'étape d'extraction. Récupérer le format de stockage, le format des tables et des colonnes des bases de données amont.
    b. Localiser l'ensemble des informations qui seront nécessaires à la population du Data Warehouse. Récupérer les noms des responsables et des administrateurs pouvant donner accès à ces données.
3. Création des modèles de données du Data Warehouse.
    a. Combinaison de l'ensemble des données opérationnelles dans des schémas optimaux suivant le compromis temps de chargement, temps de réponse et espace

de stockage, pour l'ensemble des requêtes qui interrogeront le Data Warehouse. En fonction du nombre de requêtes simultanées, du temps de réponse souhaité, la forme logique de stockage pourra ainsi être définie. Est-ce qu'un schéma en étoile et plus ou moins approprié par rapport à un cube multidimensionnel de type R-OLAP, M-OLAP ou bien H-OLAP. Est-ce qu'un accès direct au DDS est suffisant.

4. Prendre en compte la possibilité d'étendre le Data Warehouse à d'autres problématiques.
    a. Estimation du besoin de hardware et de software pour l'achat de ces derniers.
    b. Prendre notamment en compte le stockage, les serveurs de traitement et de sauvegardes (comme un système opérationnel dont le fonctionnement est vital pour l'organisation, un système décisionnel doit généralement avoir un système de secours).
5. Planification des flux de mise à jour.
    a. Définition de la fréquence de mise à jour des données.
    b. Planification des flux de processus d'extraction, de transformation de nettoyage et de validation de l'information.
6. Définition des processus de nettoyage et de validation de l'information.
    a. Définition des processus d'extraction et de transformation qui chargeront le Data Warehouse.
    b. Définition des processus qui génèreront de manière automatique des rapports.
    c. Définition des règles d'administrations.
    d. Définition de qui pourra voir quelles informations.
    e. Définition des droits des développeurs.
7. Installation et test de la plateforme décisionnelle. Il faut noter que jusqu'à ce point, rien n'a été fait sur machine ; seul la documentation technique a été écrite.
8. Développement et test des systèmes de sauvegarde et des copies de secours. C'est un phénomène nouveau, jusqu'alors, les systèmes décisionnels n'étaient pas encore considérés comme vitaux. Dorénavant, de plus en plus de systèmes décisionnels bénéficient d'un système de secours. Il faut donc tester la récupération du système dès le départ ; avant de le mettre en place.
9. Création des processus ETL défini précédemment. C'est ce que nous allons faire maintenant.
10. Exécuter d'une première fois chaque processus pour vérifier les temps de réponse, la qualité des données, etc.
11. Ordonnancement des processus. Définition des conditions d'exécution automatique des processus. Tous les jours, toutes les semaines, tous les mois etc.
12. Suivie et maintenance du Data Warehouse.

Mise en œuvre :

Mise en place du Data Warehouse Orion

La première étape pour lancer un projet décisionnel est de faire l'inventaire des questions qui se trouvent sans réponse à travers l'ensemble de l'entreprise. En effet, il est fréquent qu'un projet décisionnel avec un périmètre très restreint, par exemple, uniquement pour la force de ventes, fonctionne bien dans un premier temps. Mais ses extensions n'ayant pas été prévues au départ, plusieurs choix techniques bons pour un petit projet, s'avèrent être aujourd'hui des handicaps majeurs à l'évolution du système.
En moyenne, le volume des données double tous les deux ans. On parle donc souvent de 'scalibility', soit la capacité de monter en puissance d'une plateforme décisionnelle.

Dans le cadre de notre projet Orion, voici les principales questions

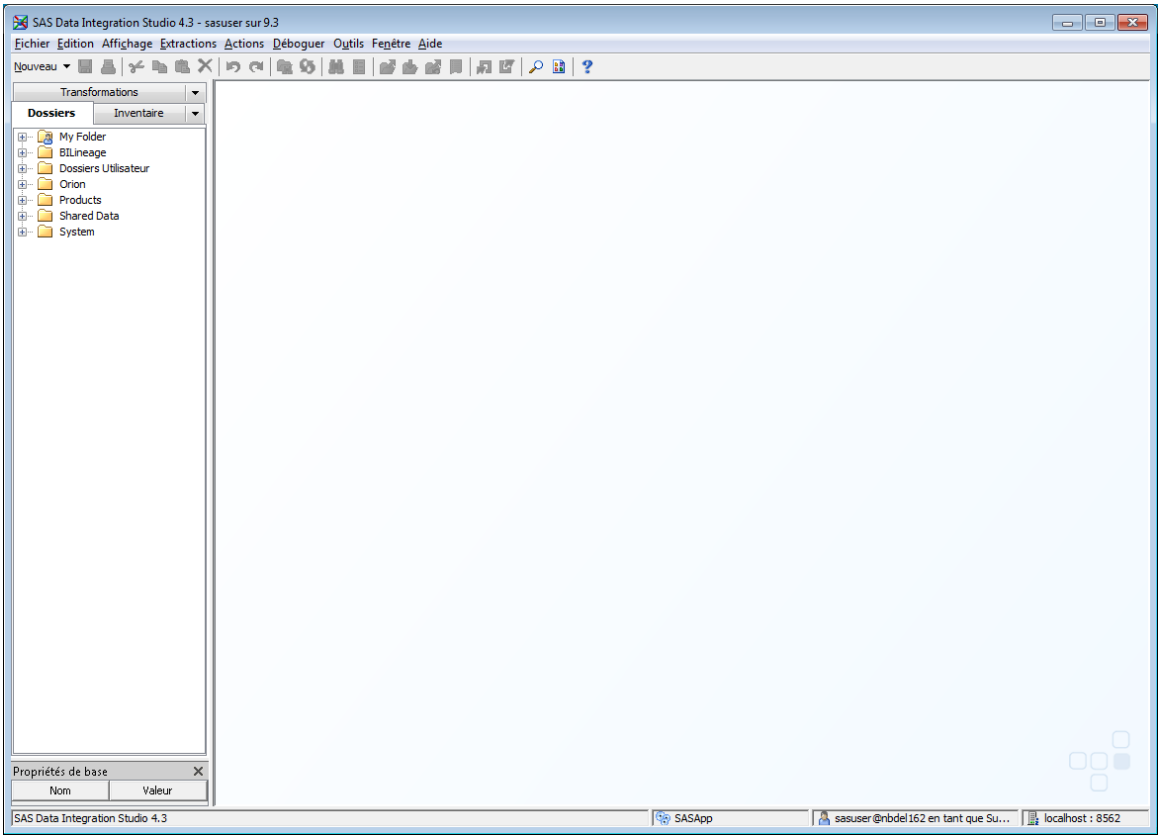
- Quelle est la tendance des ventes pour l'année à venir ?
- Quels sont les 20 commerciaux qui ont fait le plus de vente en 2002 ?
- Quels sont les commerciaux qui performent le mieux par pays, sexe, âge, salaire ?
- Quels sont les 10 produits qui génèrent le plus de marge ?
- Quels sont les produits les moins vendus ?
- Quels sont les produits qui contribuent à moins de 0.05% du CA pour un pays/une année donné(e)? - Est-ce que ces produits peuvent être remisés?
- Quelle est la marge générée par ce produit, ce groupe de produit, catégorie de produit et ligne de produit ?
- Est-ce que la marge dépend de la quantité vendue ?
- Est-ce que les remises font augmenter les ventes ?
- Est-ce que les remises font augmenter la marge ?
- Y a-t-il une relation entre le temps, l'espace et la vente de produit ?

Souvent, lorsque l'on arrive avec son cahier pour faire l'inventaire des questions, si dans un premier temps les collaborateurs se demandent ce qu'est ce nouveau projet informatique qui vient les déranger, on peut très rapidement récolter une multitude de griefs sur les systèmes actuellement mis en place.

Implémentations ETL sur le cas Orion Star

La maîtrise d'ouvrage a élaboré le cahier des charges du projet décisionnel de la société Orion. La maîtrise d'œuvre l'a analysé et a défini le cahier de spécification, qui lui-même a été validé par la maîtrise d'ouvrage.
Maintenant que la conception du Data Warehouse a été réalisée, il faut créer les six processus d'ETL qui vont charger les six tables du schéma en étoile du Data Warehouse Orion Star. Il faut ensuite créer le processus qui sélectionne les données du Data Mart « Orion Gold ». Enfin, le projet nécessite de créer un hyper cube multidimensionnelle.

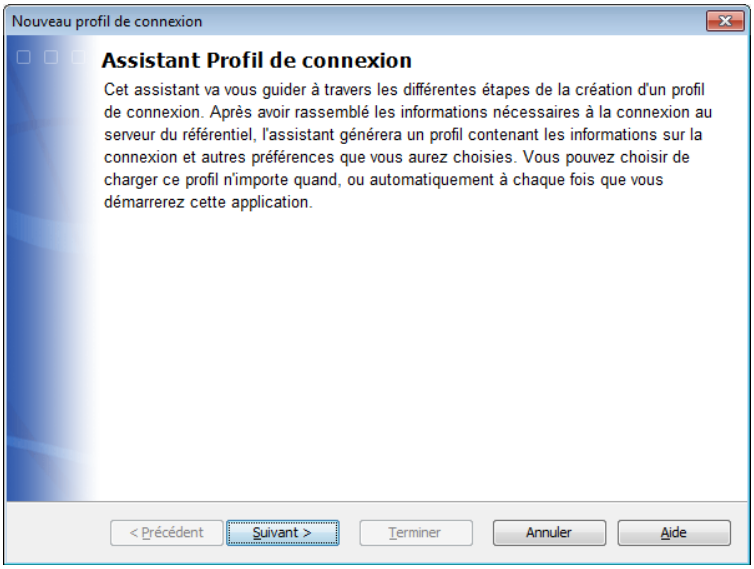
Se connecter à SAS® Data Integration Studio



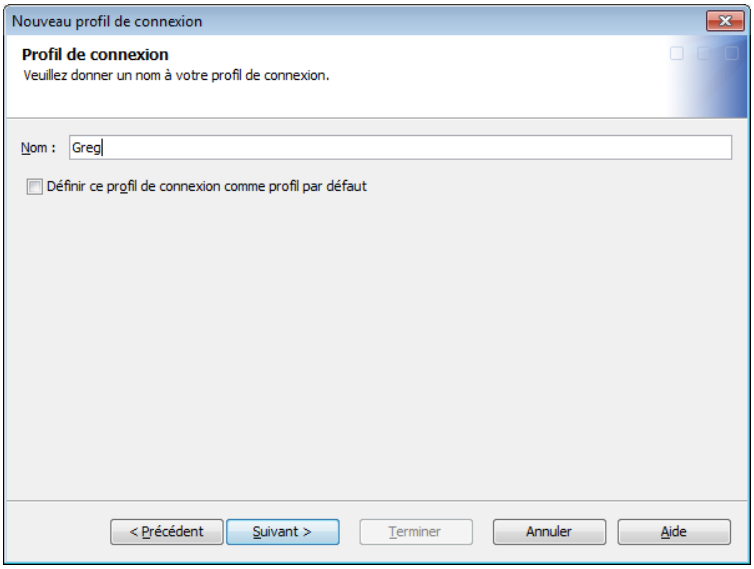
Gestion des profils

Pour se connecter à une interface cliente Java de la plateforme décisionnelle SAS, il est nécessaire de créer un profil. Ce profil contient les informations sur l'utilisateur : nom d'utilisateur et mot de passe, et sur la connexion au serveur : nom du serveur et port de communication.
Il est nécessaire de créer un profil de métadonnées utilisateur pour se connecter aux interfaces comme notamment la console d'administration, SAS® Data Integration Studio ou Information Map Studio®.
Lors de la première connexion à une application cliente Java, l'assistant suivant apparaît.
Nous allons créer un nouveau profil : sélectionner « Créer un Profil de métadonnées » et cliquer sur OK.

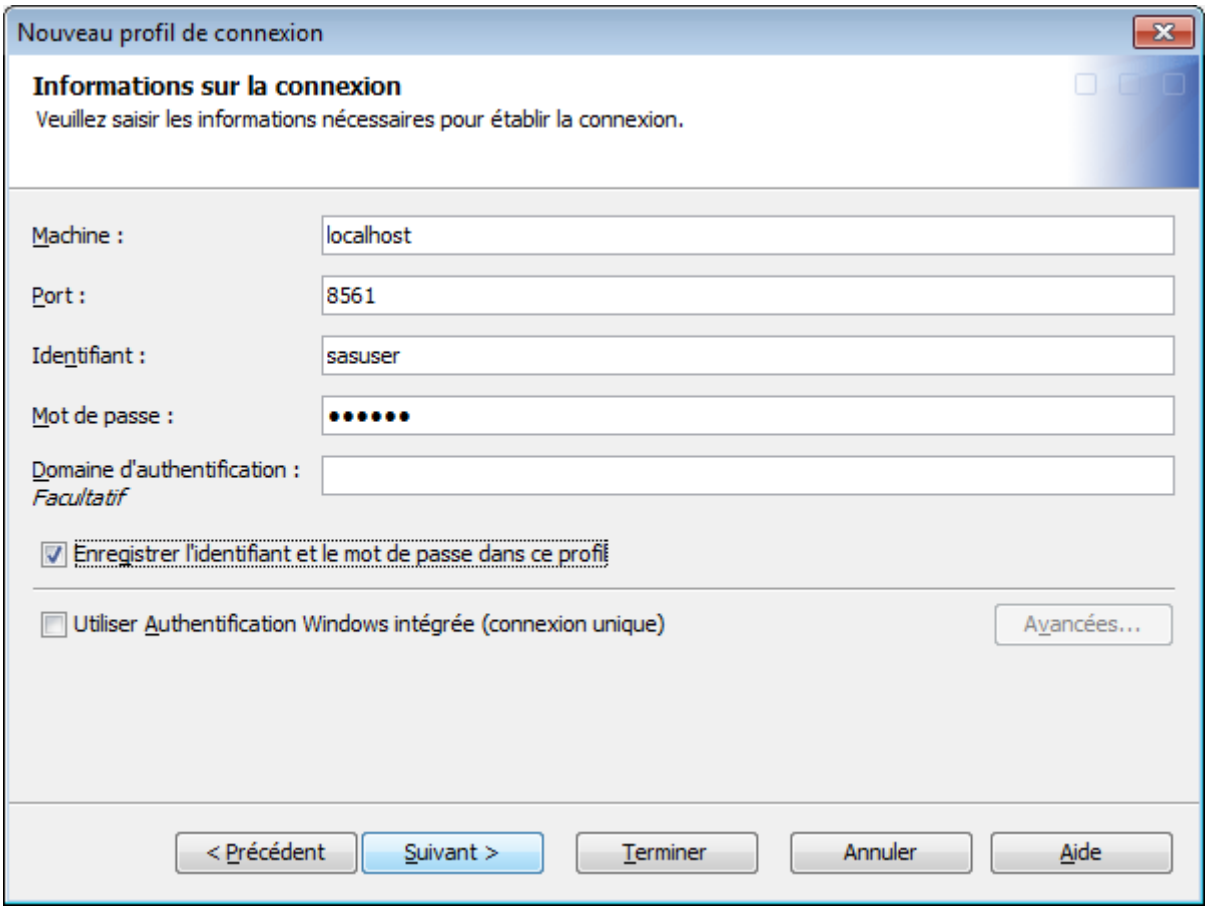
Pour la connexion à SAS Data Integration Studio :
Démarrer → Programmes → SAS → SAS Data Integration Studio



Suivant.

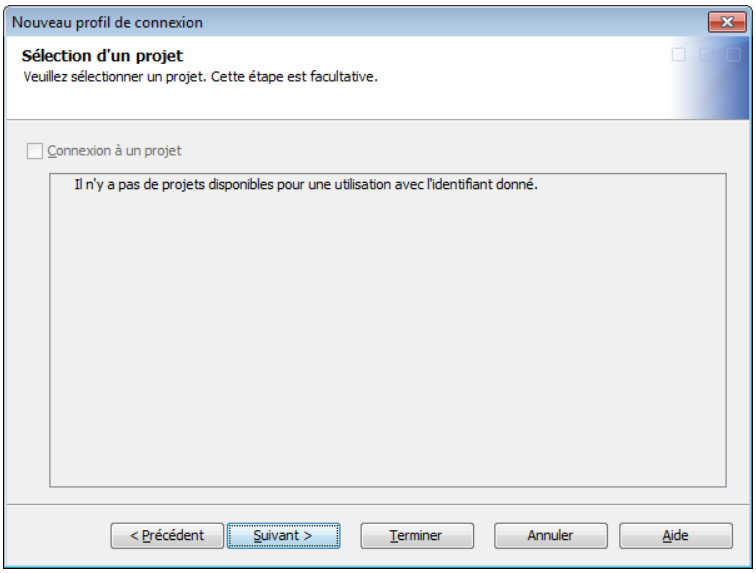


Lui donner un nom
Un utilisateur doit souvent se connecter à plusieurs environnements sur plusieurs machines, ne pas hésiter à donner un nom de profil de métadonnées explicite.

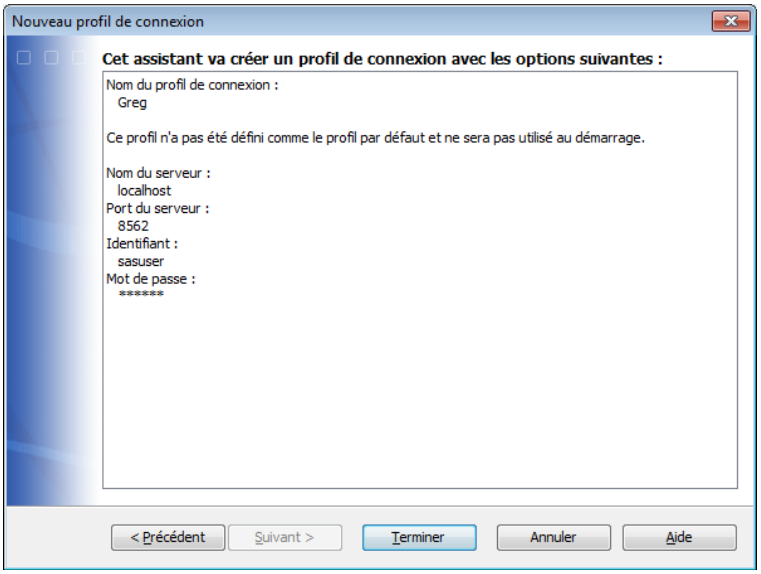


Rentrer le nom de la machine à laquelle vous souhaitez vous connecter.
Lorsque le serveur de métadonnées auquel vous vous connectez se trouve sur la même machine que celle où vous êtes, le nom de machine « localhost » fonctionne. L'adresse IP du serveur fonctionne aussi.
Le port par défaut du serveur de métadonnées est le port 8561 (par défaut).
Entrer votre identifiant et votre mot de passe (eleve – SASpw1) ou parfois (sasdemo – Orion123) (Marcel – Student1)

Cliquer sur suivant.



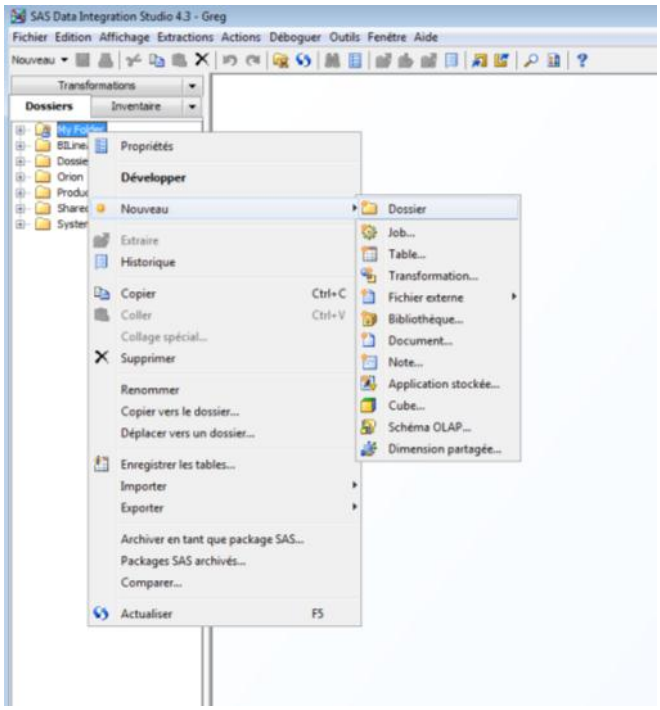
Suivant



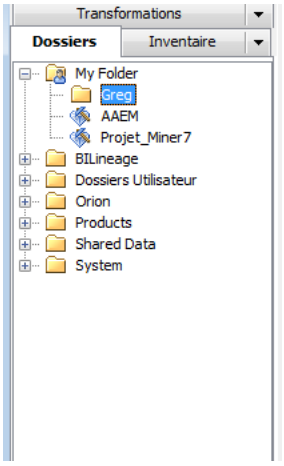
Cliquez sur terminer.

Définition de l’arborescence des métadonnées ETL personnalisées :

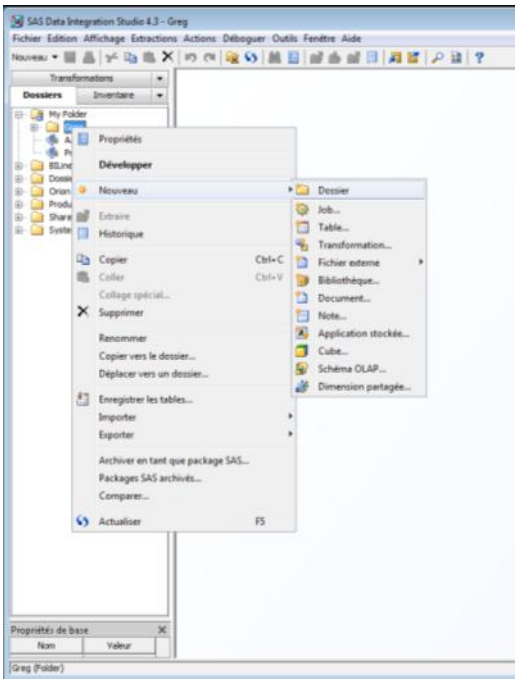
Dans les **dossiers**, Clic-droit sur le dossier **Mon Dossier → Nouveau → Dossier**

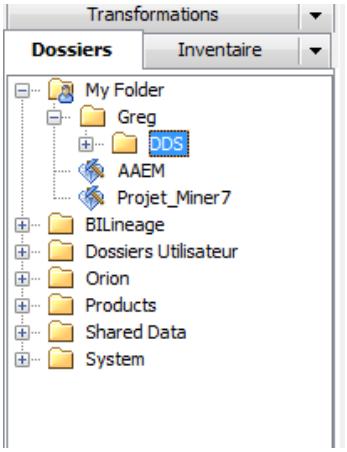


Créer un dossier avec votre **nom**. Tous les objets que vous créerez seront sauvegardés dans ce dossier.

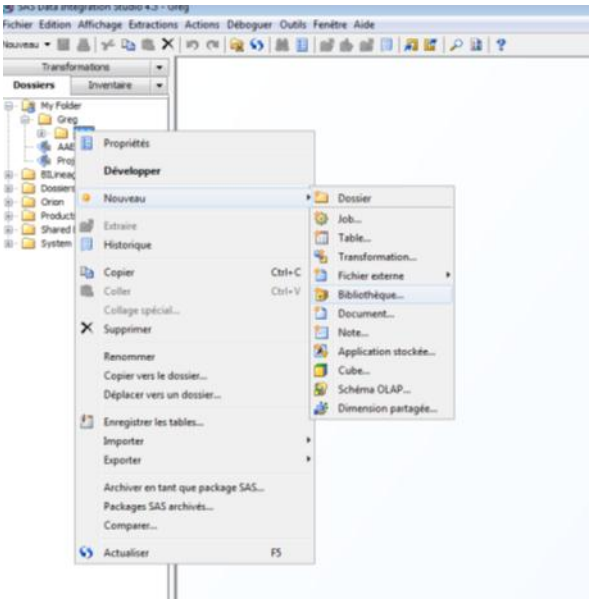


Créer un sous dossier **DDS** (Detail Data Store)

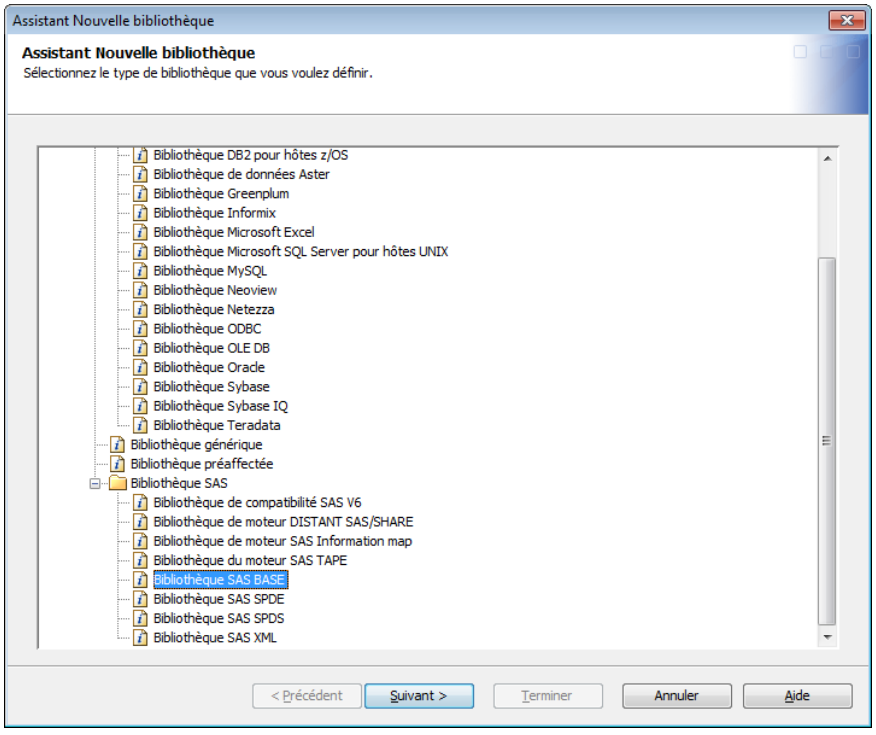




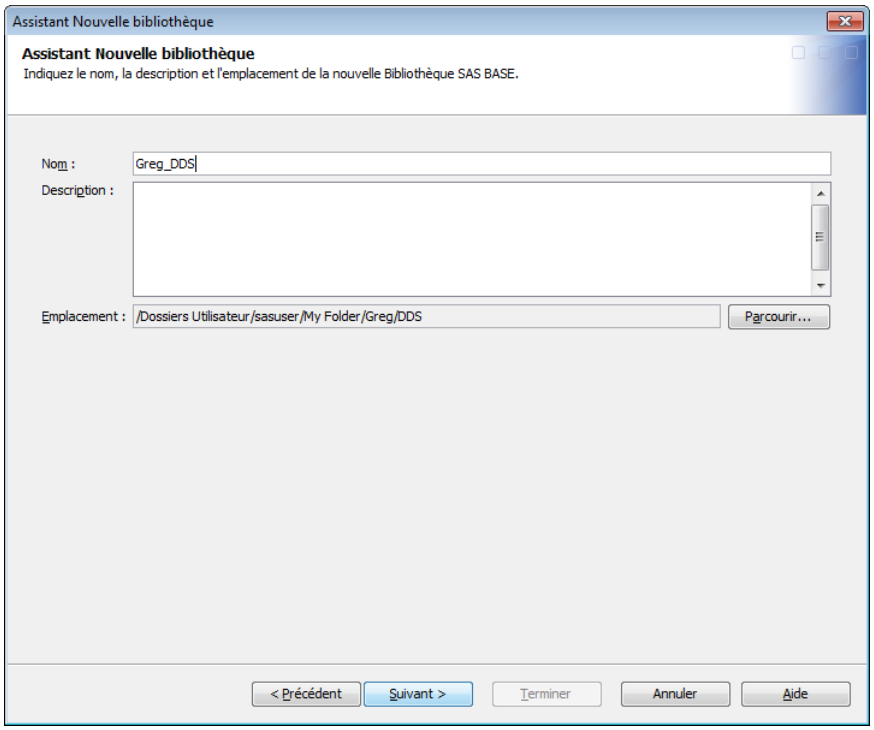
Création de la bibliothèque DDS



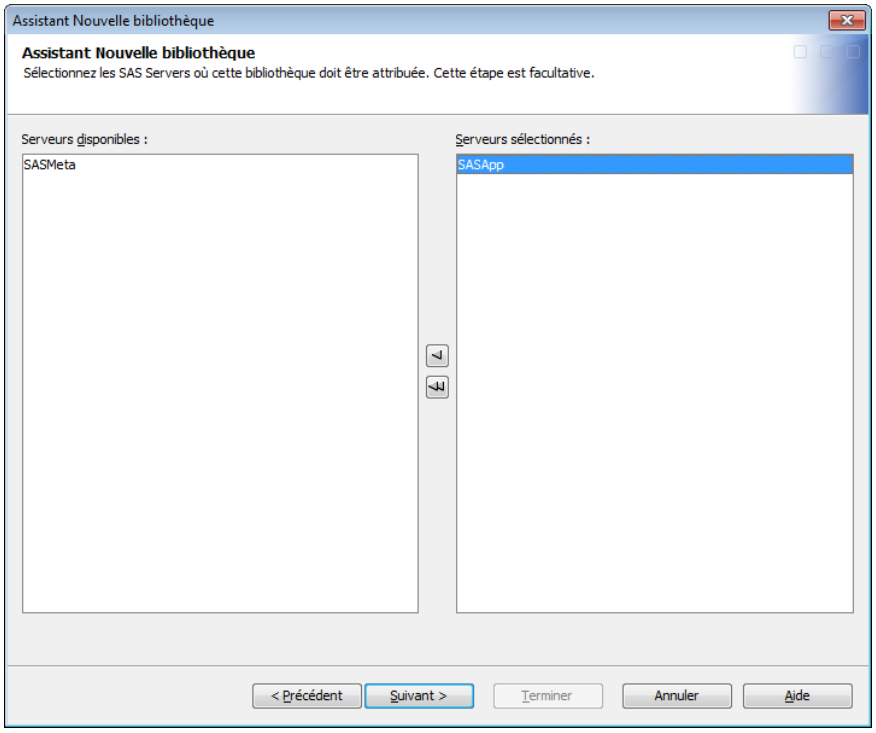
Clic-droit sur le dossier **DDS → Nouveau → Bibliothèque**



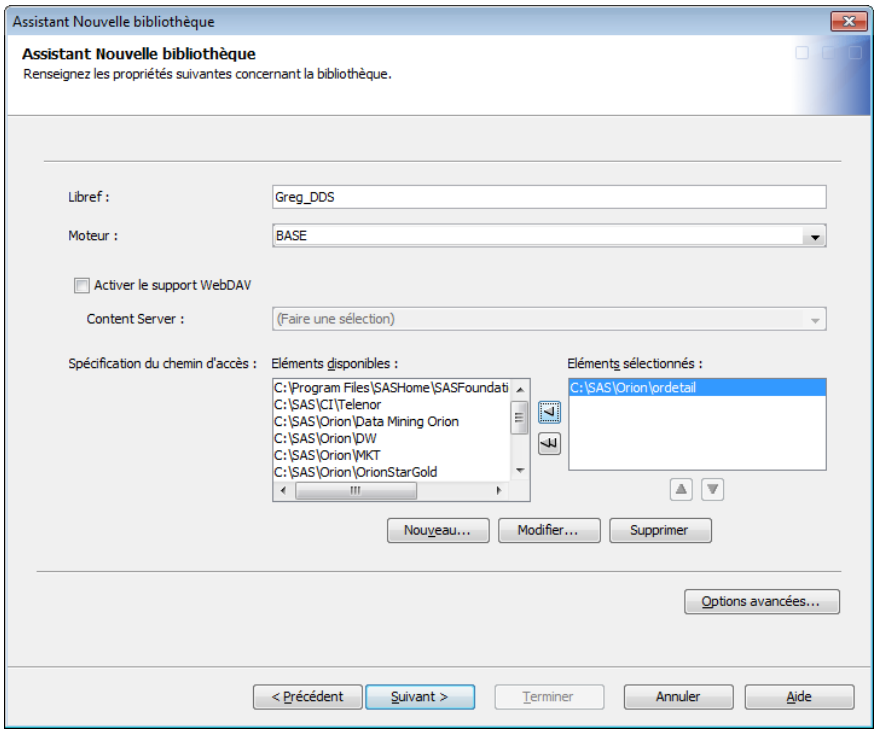
Sélectionner la **Bibliothèque SAS Base**
Suivant



Lui donner un nom : **Votre_mon_DDS**
Suivant



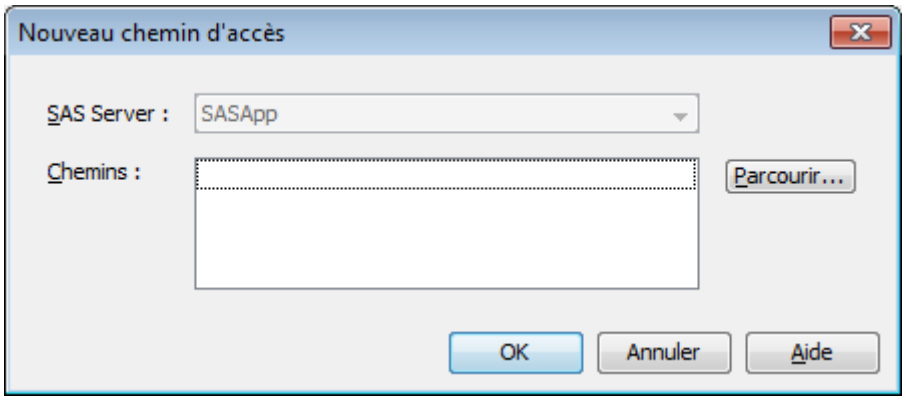
Sélectionner le serveur **SAS App** (ce sera toujours le cas)
Suivant



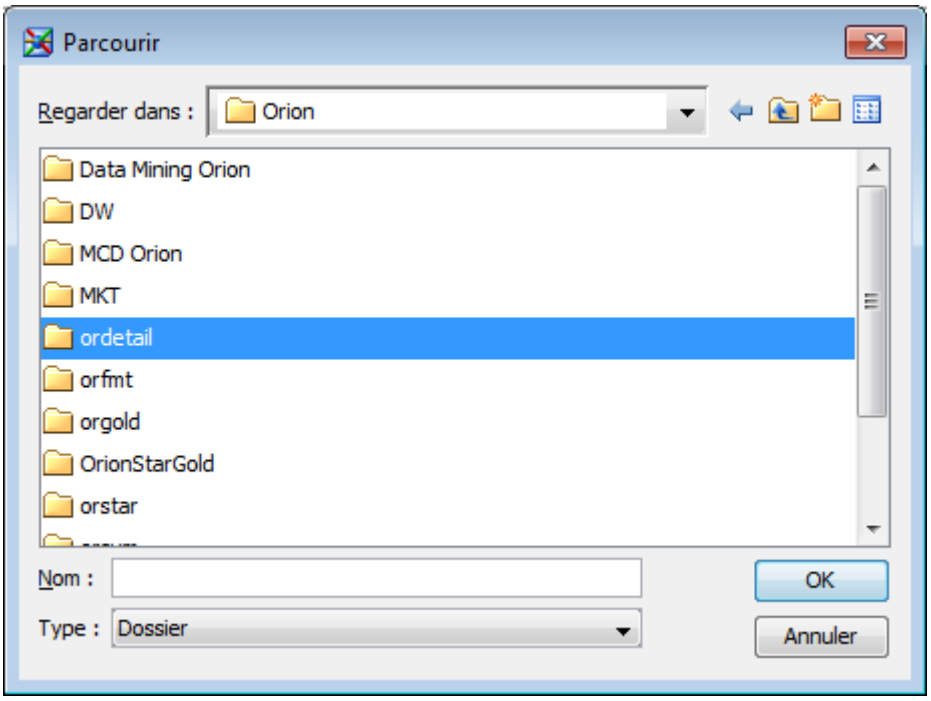
Définir un libref
Le libref est de maximum 8 caractères. Il ne commence pas par un chiffre. Il ne peut être composé que des 26 lettres de l'alphabet, des chiffres ou bien le souligné « _ ».
Le libref peut être **les_premieres_lettres_de_votre_nom_DDS**

Sélectionner le chemin C:\SAS\Orion\ordetail (ou équivalent D:\SAS\Orion\ordetail ou C:\Orion\ordetail ou S:\Workshop\OrionStar\ordetail). Généralement il existe déjà. C'est le dossier sur le serveur où se trouvent les données du DDS.

Si ce chemin n'existe pas, il faut le créer. Sinon, sauter une page
Cliquer sur nouveau pour définir un chemin

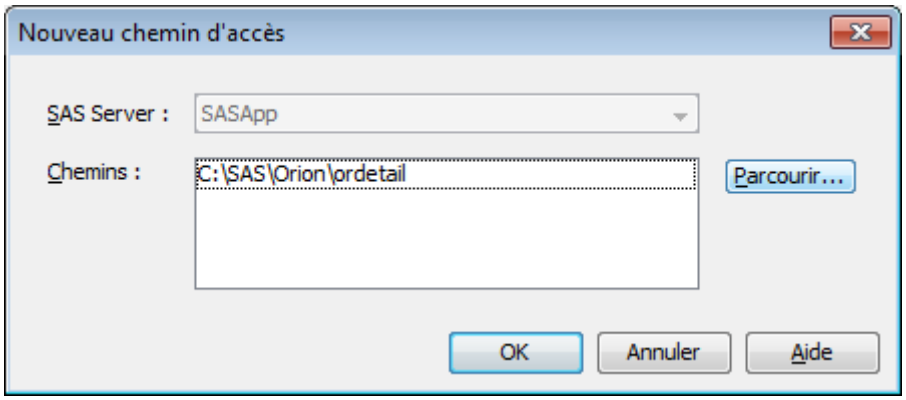


A l'aide du bouton **parcourir,** sélectionner le dossier où se trouvent les données du cas Orion.

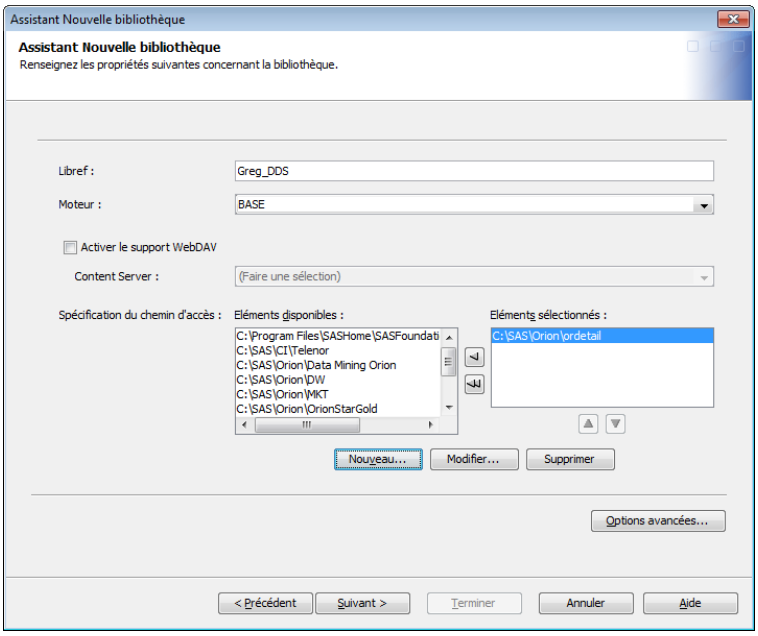


Souvent C:\SAS\Orion\ordetail, voir D:\SAS\Orion\ordetail, ou E:\SAS\Orion\ordetail

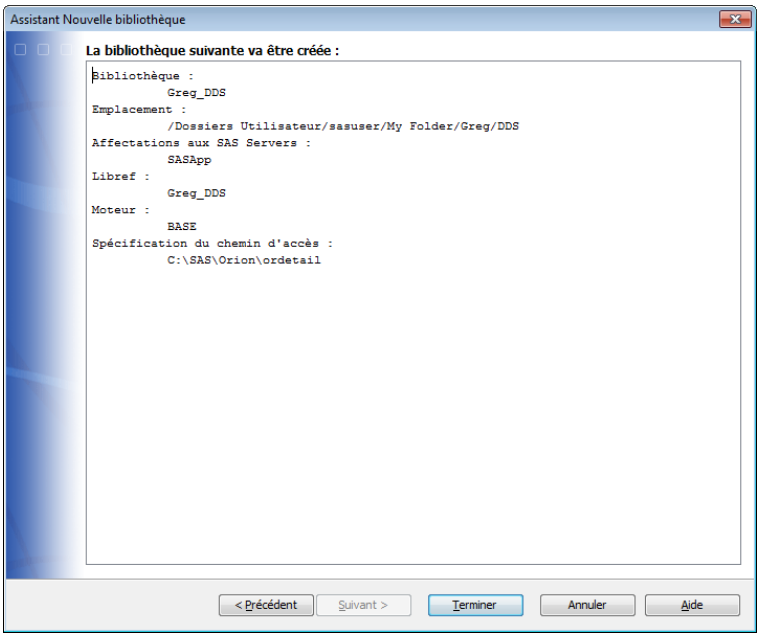
Sélectionner le dossier **ordetail**, ne pas rentrer dedans
OK



OK

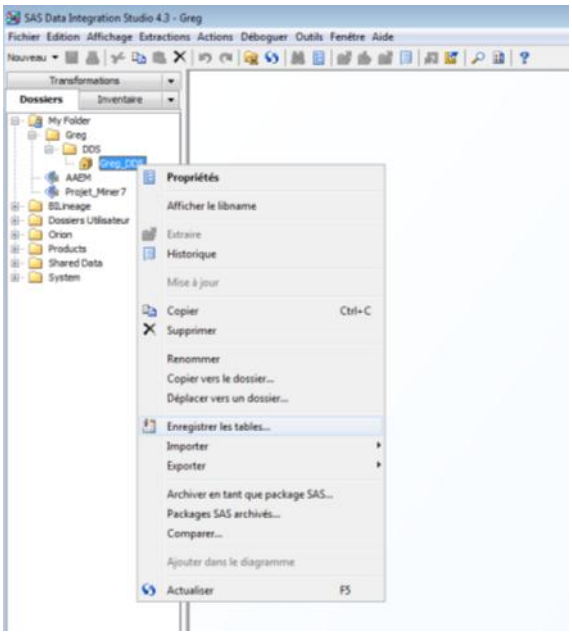


Votre bibliothèque avec **votre_nom_DDS** pointera vers C:\SAS\Orion\ordetail (ou équivalent)
Suivant

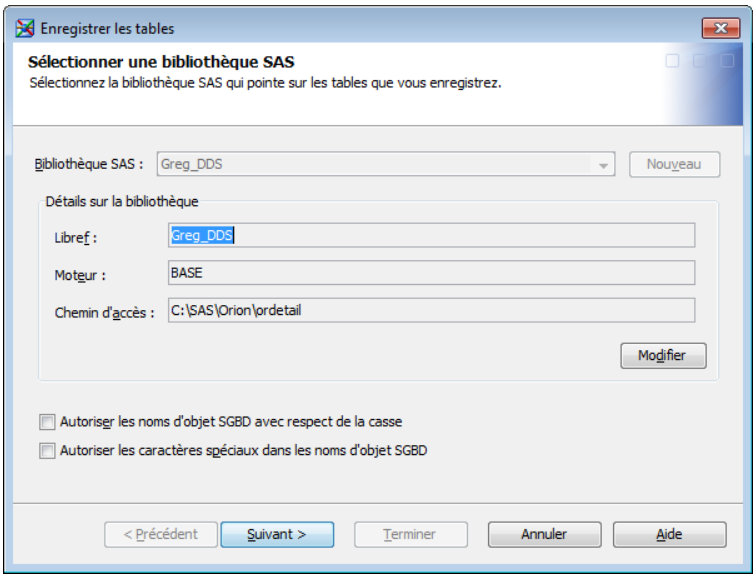


Terminer

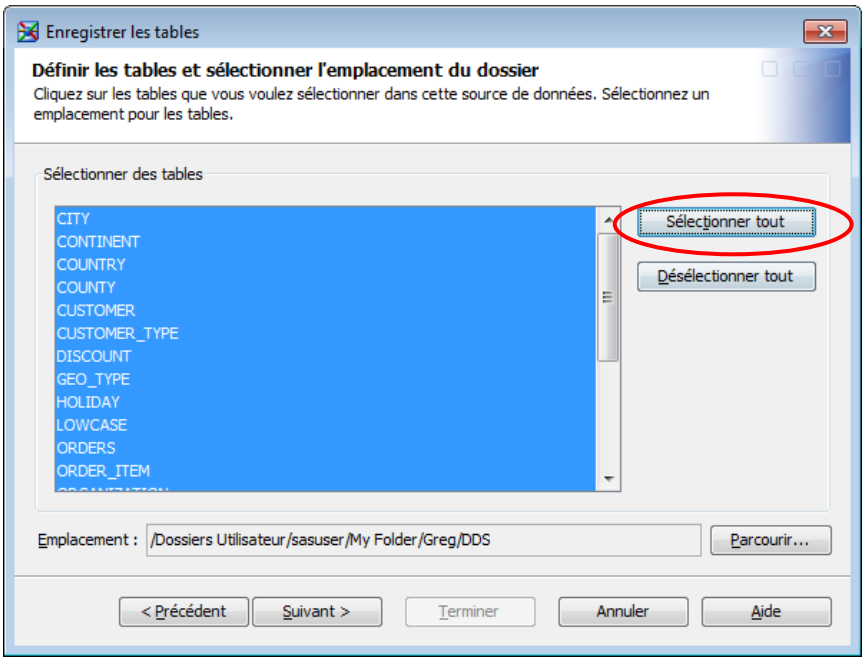
Enregistrement des métadonnées des tables du DDS



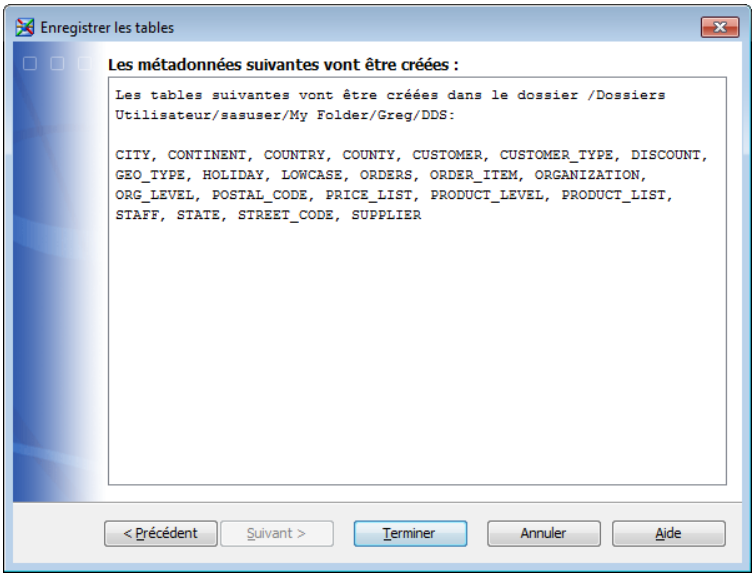
Clic-droit sur la bibliothèque créée → **Enregistrer les tables**



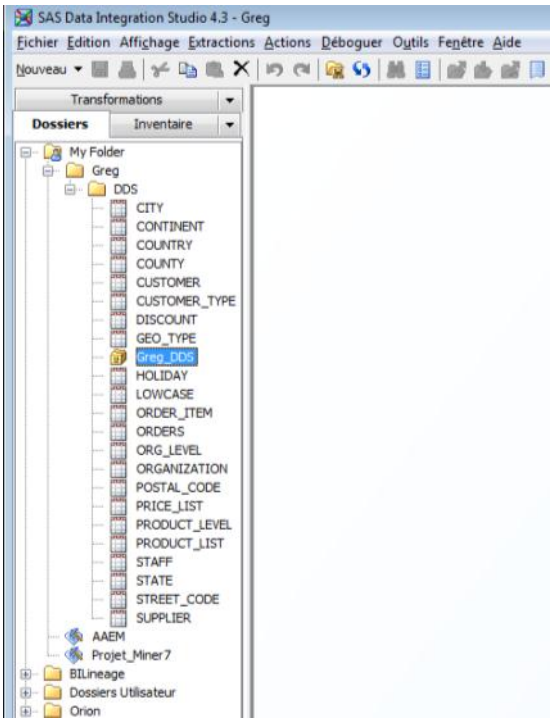
Suivant



Sélectionner toutes les tables
Suivant

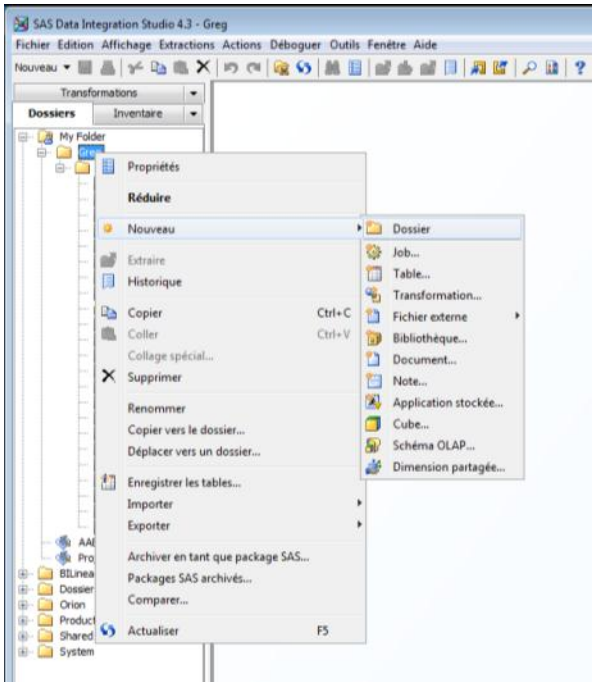


Terminer

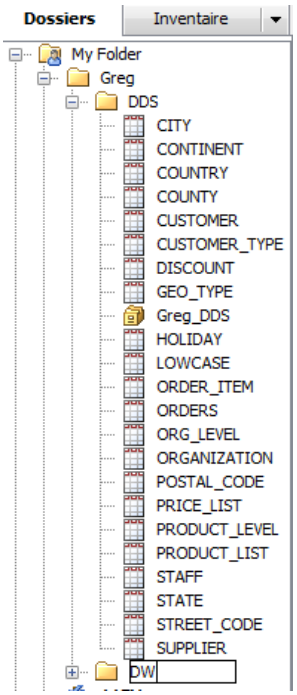


Dans votre dossier, dans DDS, se trouve la bibliothèque et toutes les métadonnées des tables du DDS.

Création de du dossier de votre Data Warehouse

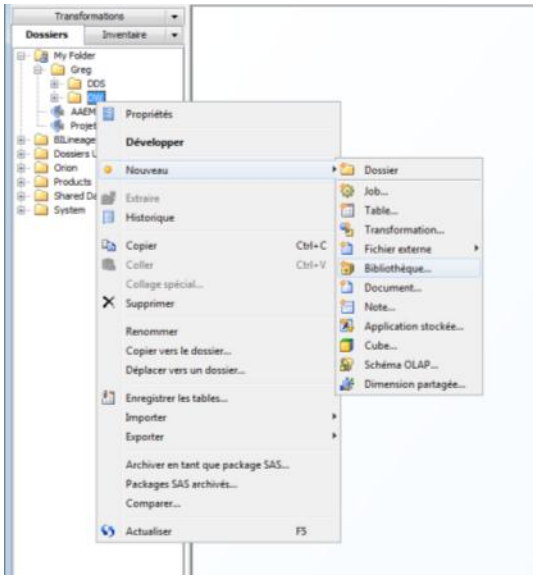


Clic-droit sur votre dossier → Nouveau → Dossier

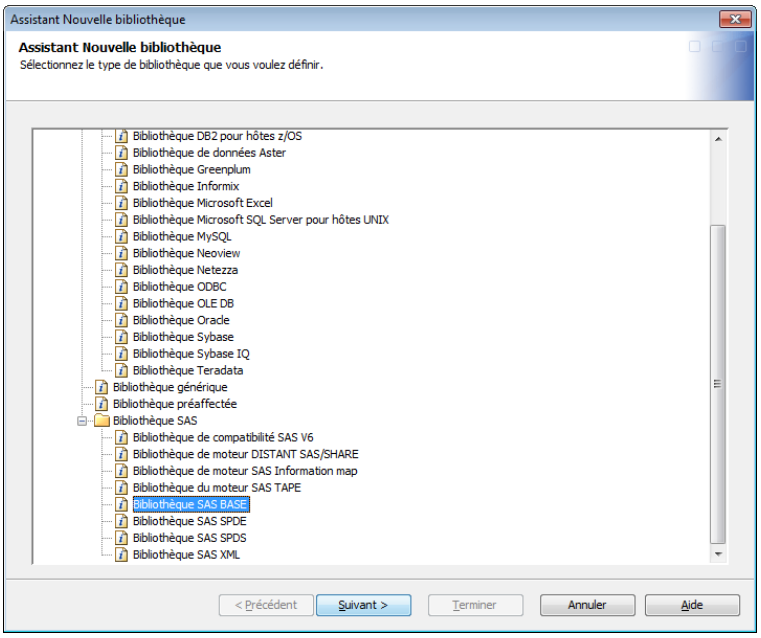


Créer un dossier DW (Data Warehouse)

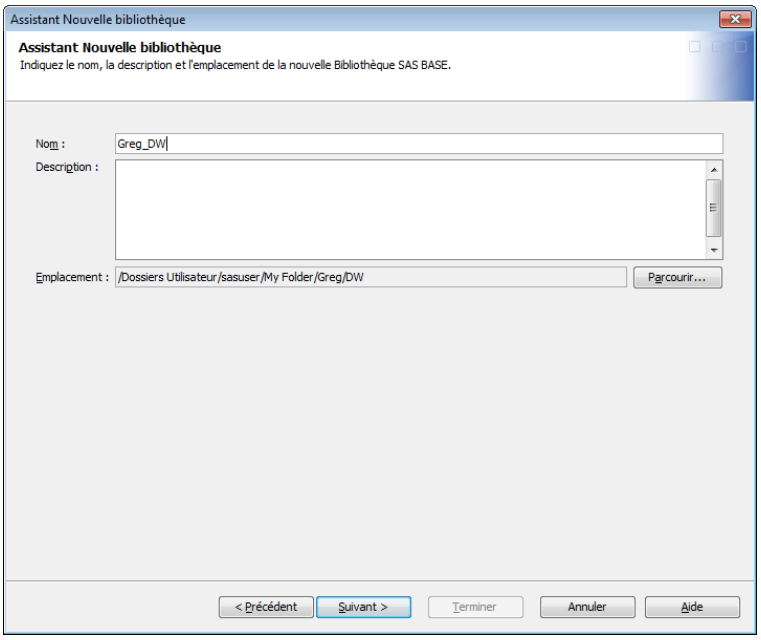
Création de la bibliothèque Orion Data Warehouse.



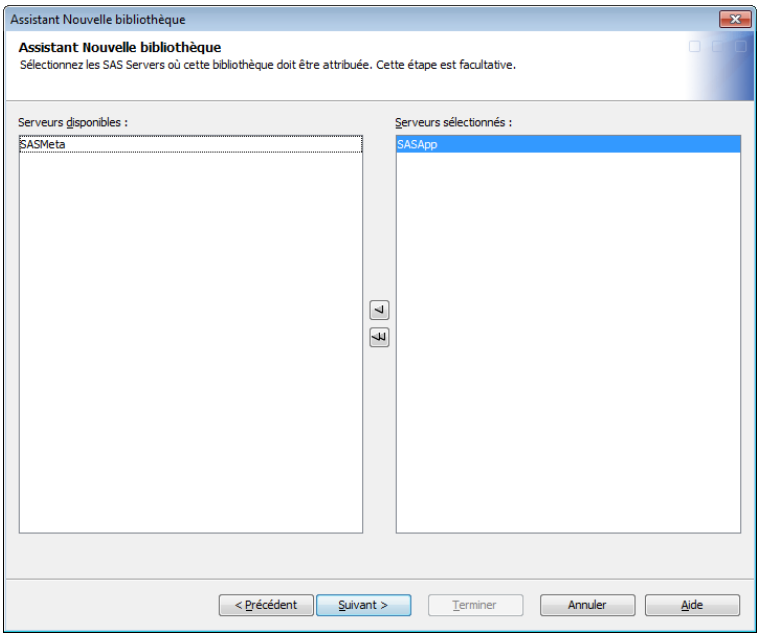
Clic-droit sur votre dossier DW → nouveau → Bibliothèque



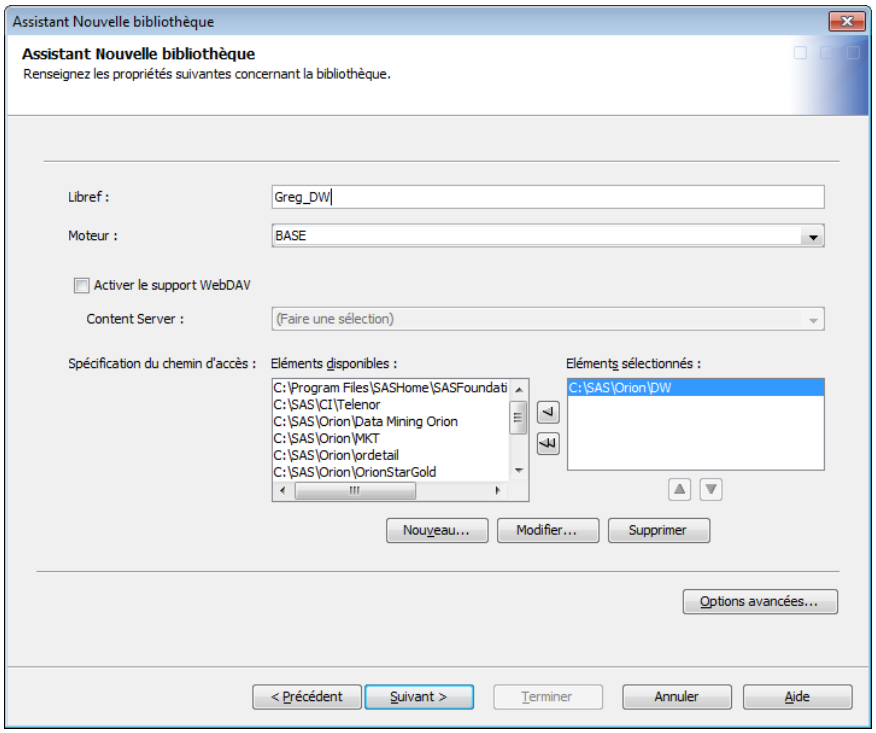
Sélectionner la bibliothèque de type **SAS BASE**
Suivant



Lui donner un nom : **Votre_nom_Orion_Data_Warehouse**
Suivant

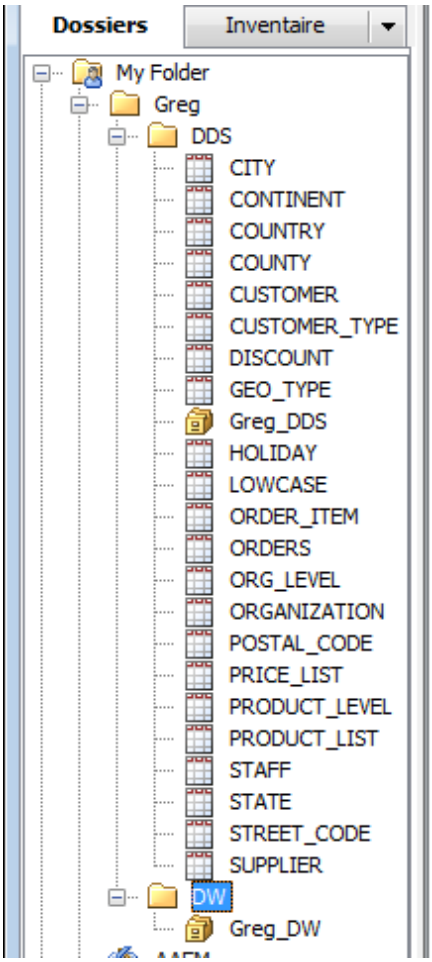


Sélectionner le serveur **SASApp**
Suivant



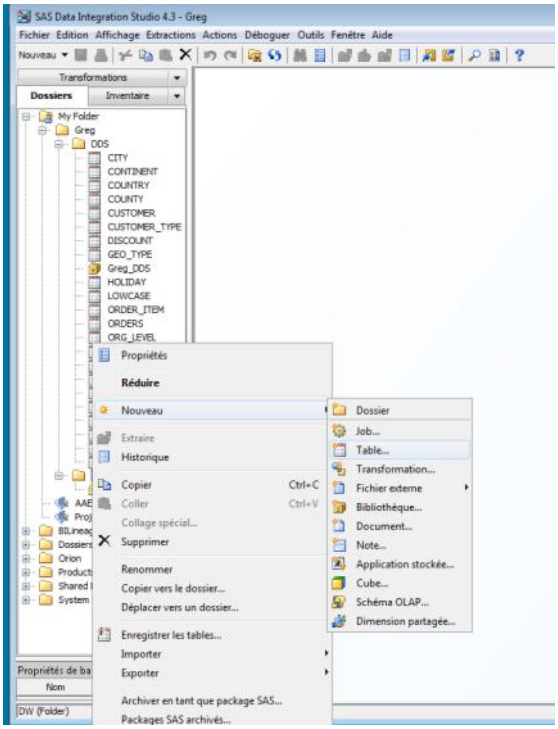
Donner un libref :
Le libref peut être **les_premieres_lettres_de_votre_nom_DW**
Sélectionner le chemin vers le dossier où vous souhaitez stocker physiquement votre Data Warehouse
Souvent **C:\SAS\Orion\DW**, voir D:\SAS\Orion\DW, ou E:\ C:\SAS\Orion\DW_Orion ou S:\Workshop\OrionStar\
Suivant

Terminer

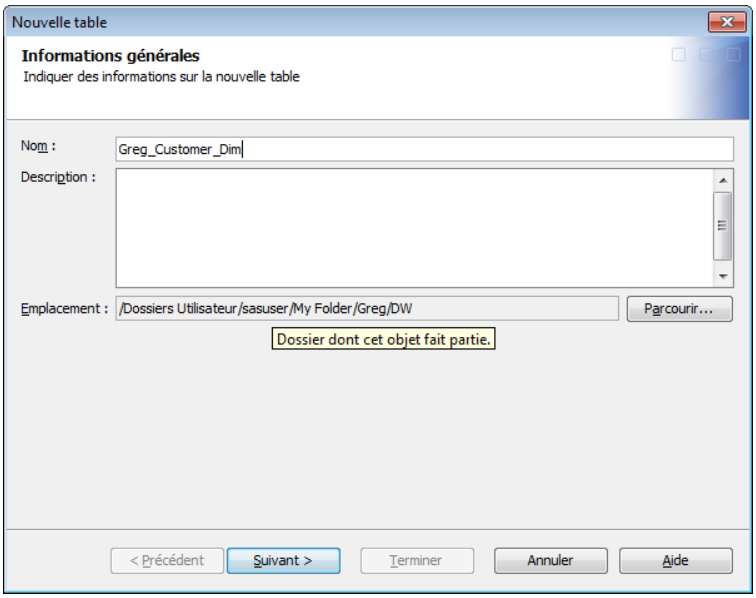


Dans Mon Dossier, il y a un dossier avec votre nom et deux sous-dossiers, l'un du DDS et l'autre du DW. Dans ses sous-dossiers se trouve la bibliothèque associée. Pour le DDS, vous avez les métadonnées de toutes les tables.

Création de la table CUSTOMER_DIM :



Créer une nouvelle table (clic-droit sur votre dossier **DW → Nouveau → Table**)

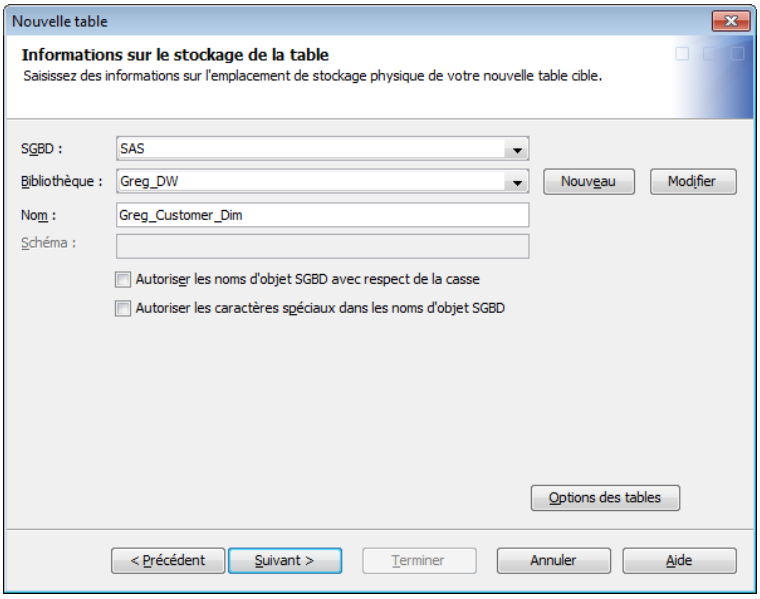


Donner un nom à votre table de dimension client commençant par votre nom. Par exemple Greg_Customer_Dim.
Un nom SAS ne contient pas de caractères spéciaux, d'espace, et ne commence par un chiffre. Seuls sont acceptés les 26 lettres de l'alphabet, les 10 chiffres et le souligné « _ » (sous le 8 d'un clavier azerty).
Par contre, vous pouvez mettre ce que vous souhaitez comme caractère dans la description.

Règle pour ce TP : tous les objets que vous créerez commenceront par votre acronyme, vos initiales ou votre nom.

Parfois, n'hésitez pas à jeter un coup d'œil sur les pages suivant votre état. De petit aller-retour sont souvent débloquant.

Suivant.

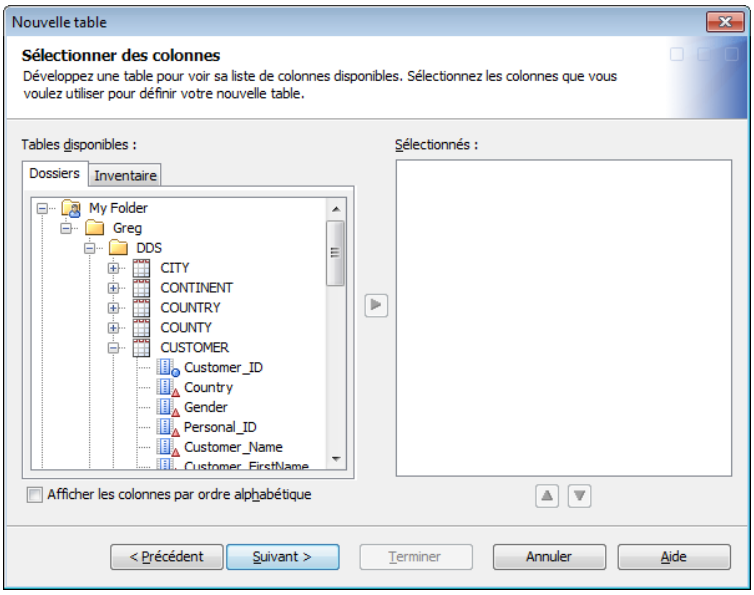


Sélectionner la bibliothèque **Votre_nom_DW.** Toutes les tables que vous créerez dans le cadre de ce TP seront sauvegardées dans la bibliothèque **Votre_nom_DW**

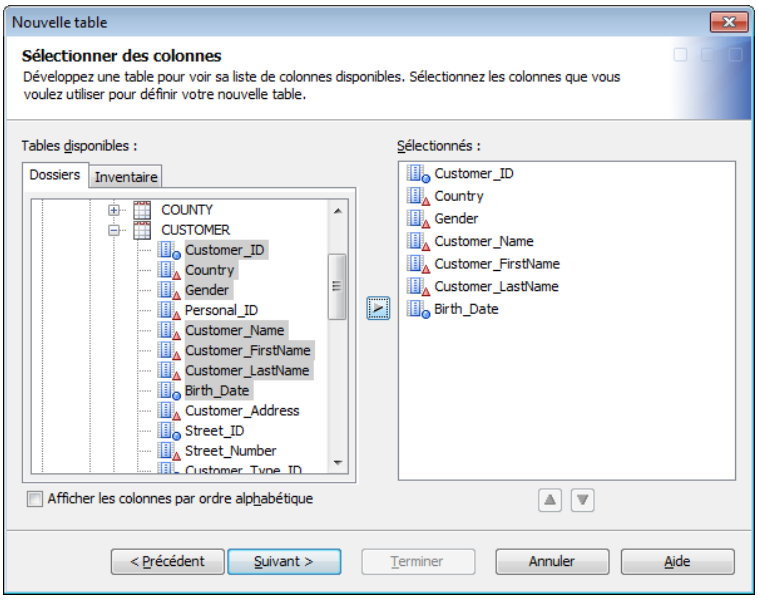
Suivant.

Suite à la définition des spécifications détaillées, en réponse au cahier des charges de la société Orion, la table cible à créer doit avoir les caractéristiques du tableau suivant, certaines des colonnes ont les mêmes caractéristiques que leur colonne mère des tables sources. Afin de simplifier la définition de ces colonnes, on peut utiliser la définition des colonnes sources pour définir celles-ci.

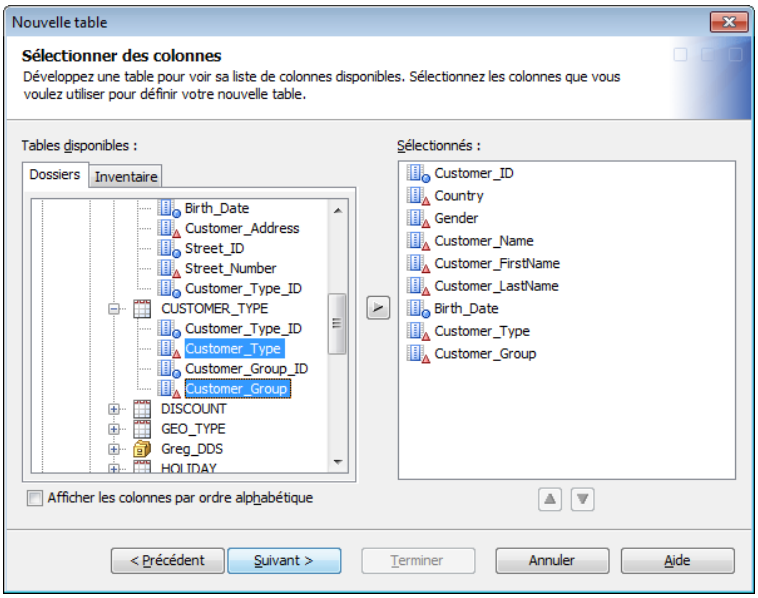
Nom	Longueur	Type	Format	Remarque
Customer_ID	8	Numérique	12.	Identique à la table : Customer
Customer_Country	2	Caractère	$COUNTRY.	Identique à la table : Customer
Customer_Gender	1	Caractère	$GENDER.	Identique à la table : Customer
Customer_Name	40	Caractère	(Néant)	Identique à la table : Customer
Customer_FirstName	20	Caractère	(Néant)	Identique à la table : Customer
Customer_LastName	30	Caractère	(Néant)	Identique à la table : Customer
Customer_BirthDate	4	Numérique	DATE9.	Identique à la table : Customer
Customer_Age_Group	12	Caractère	(Néant)	A créer
Customer_Type	40	Caractère	(Néant)	Identique à la table : Customer_type
Customer_Group	40	Caractère	(Néant)	Identique à la table : Customer_type
Customer_Age	3	Numérique	(Néant)	A créer



Dans le dossier **Mon Dossier → Votre nom → DDS**



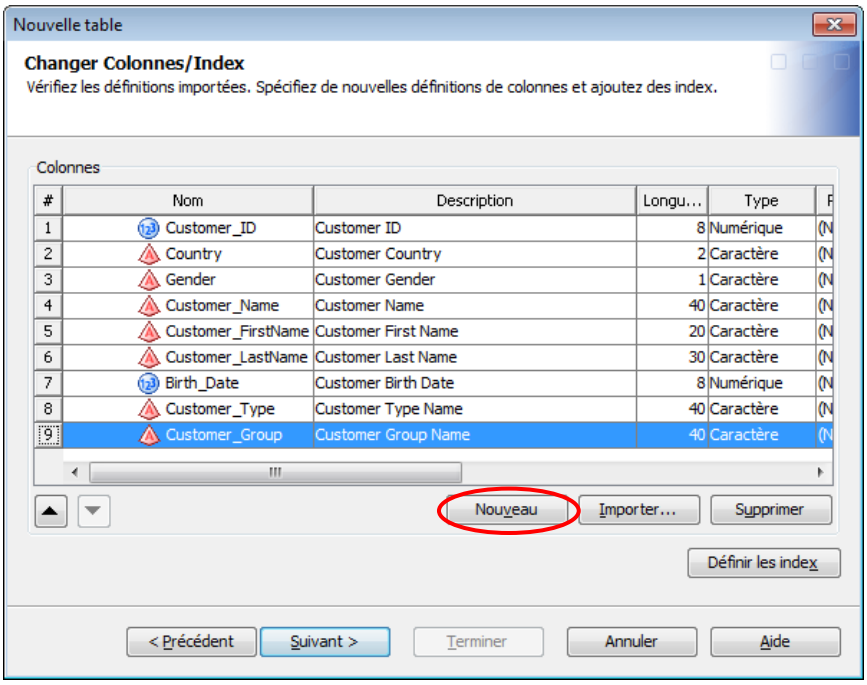
Sélection les collones Customer_ID, Country, Gender, Customer_Name, Customer_FirstName, Customer_LastName et Birth_Date depuis la table Customer.



Sélection les collones Customer_Type et Customer_Group (avec un triangle rouge) depuis la table Customer_type.

Vous avez donc, depuis le répertoire personnalisé où se trouvent les tables de détail d'Orion, sélectionnées dans les tables Customer et Customer_type ; les 9 colonnes identiques dans le système source et dans la cible.

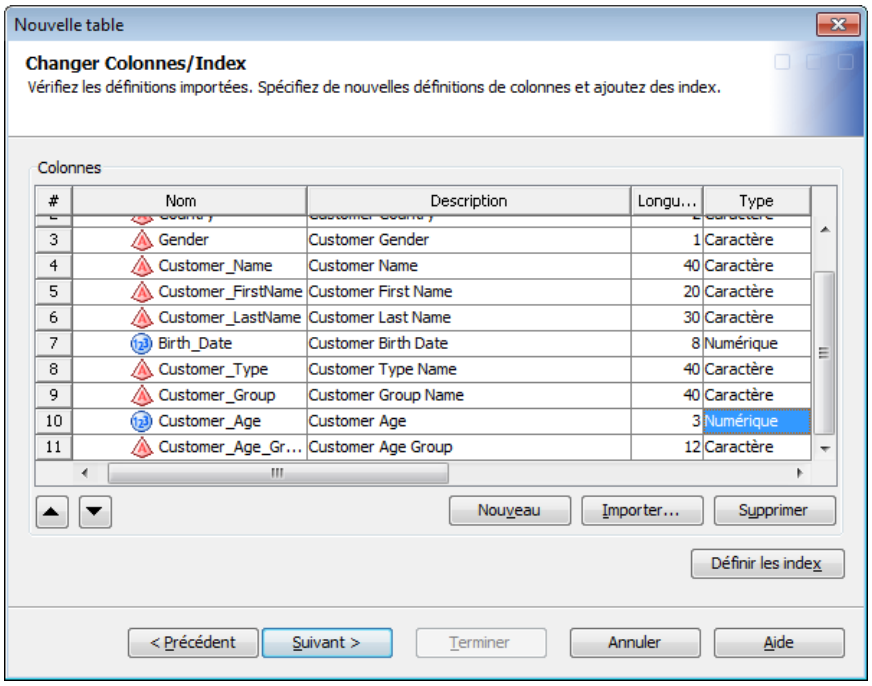
Suivant



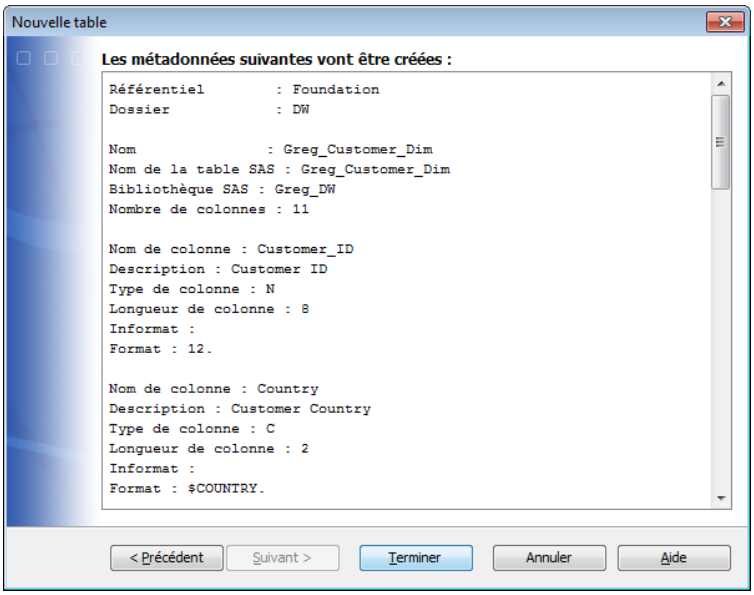
Créer les deux colonnes Customer_Age et Customer_Group_Age suivant le tableau ci-dessous, avec leur nom, description, type, format et longueur. Le nom d'une colonne ne dois pas avoir de caractères spéciaux, d'espaces, et ne commence par un chiffre. Seuls sont acceptés les 26 lettres de l'alphabet, les 10 chiffres et le souligné « _ » (sous le 8 d'un clavier azerty). La longueur maximale est de 32 caractères.
Par contre, vous pouvez mettre ce que vous souhaitez comme caractère dans la description.
Dans notre cas, par exemple, la colonne Customer_Age a un nom conforme est sa description contient un espace : « Customer Age ». Sa longueur est de 3 et son type numérique.

Nom	Longueur	Type	Format
Customer_Age_Group	12	Caractère	(Néant)
Customer_Age	3	Numérique	(Néant)

Nous avons défini les métadonnées des 11 colonnes de la table Customer_Dim.

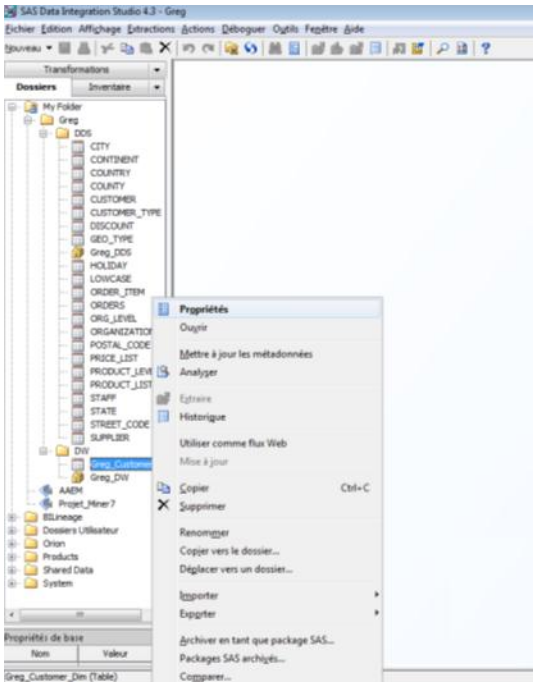


Suivant.

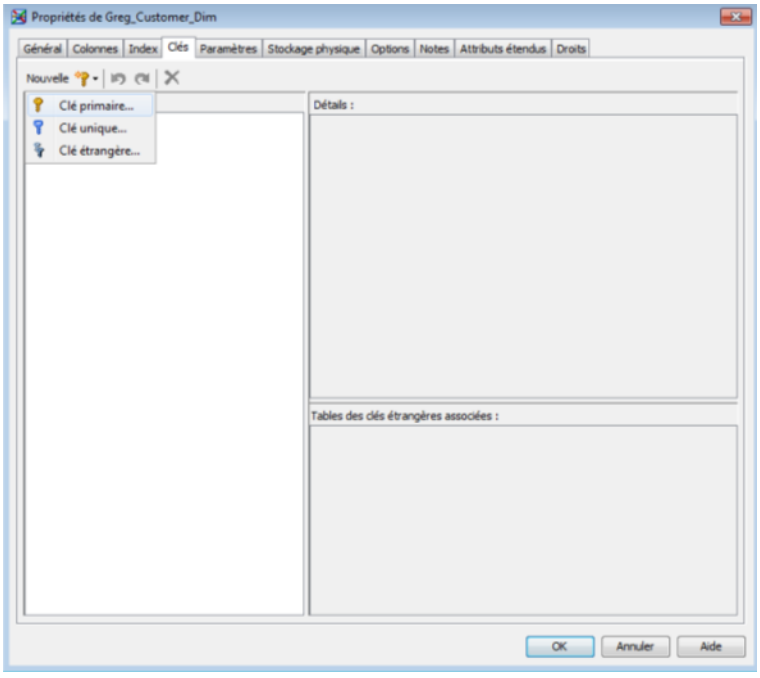


Terminer

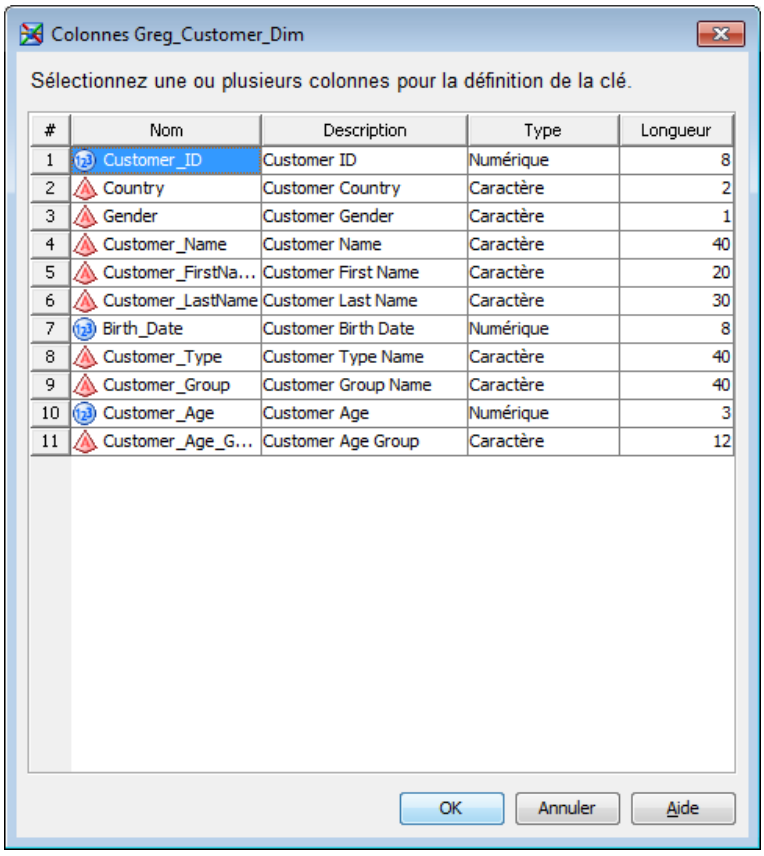
Aller dans les propriétés de la table pour définir la clé primaire et l'index sur Customer_ID.



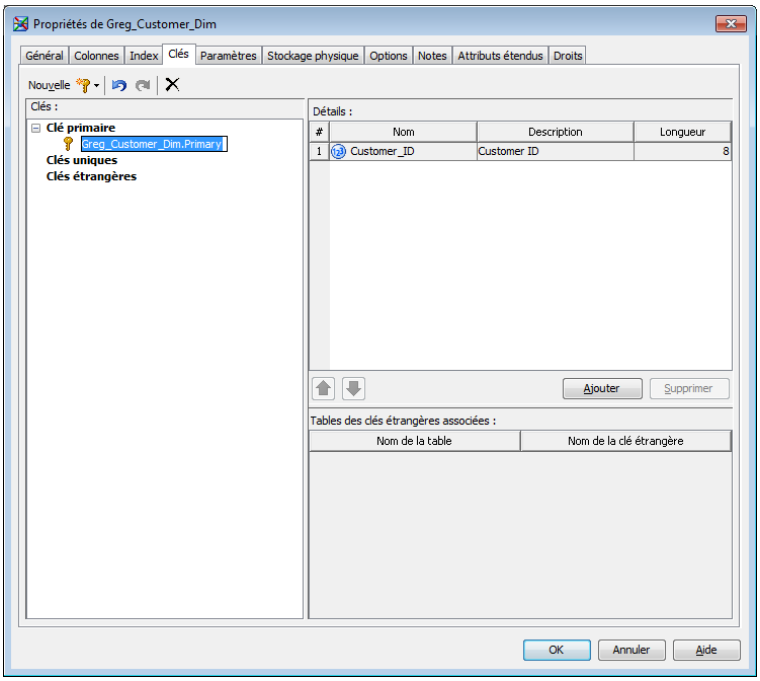
Clic-droit sur la table Votre_Nom_Customer_Dim, puis **Propriété**.



Dans l'onglet **Clés**, cliquer sur le bouton **Nouvelle**, **Clé primaire…**



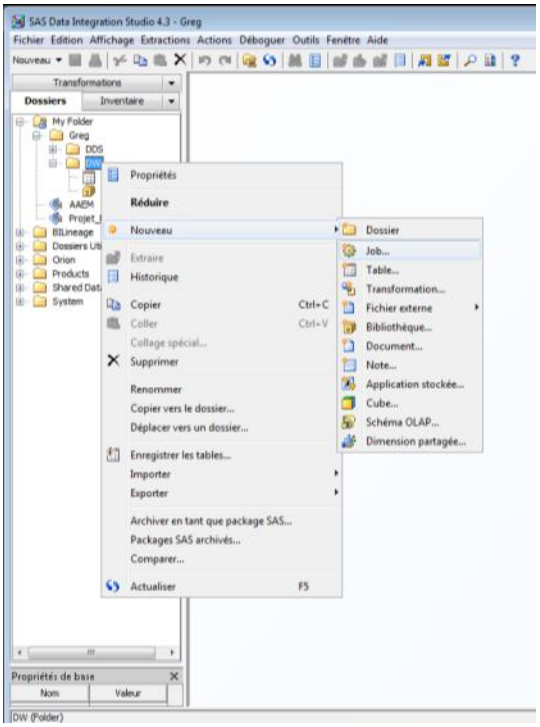
Sélectionner **Customer_ID**



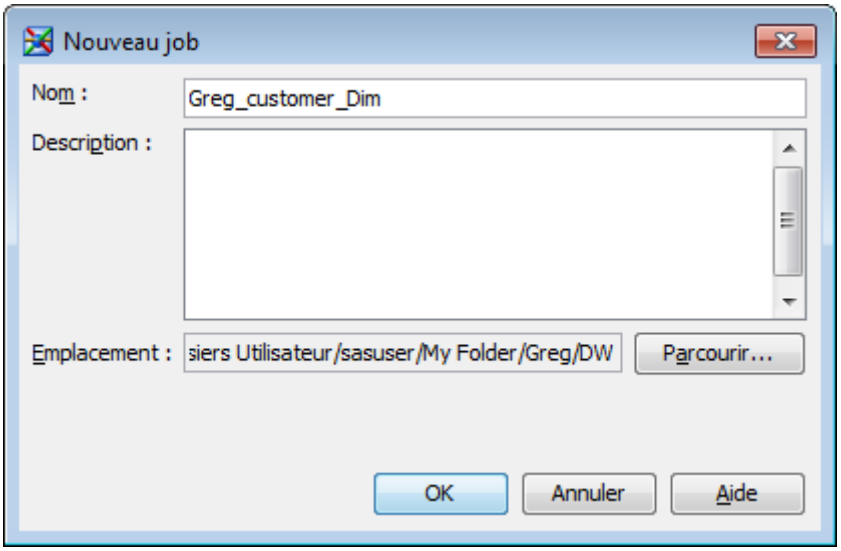
Si vous jetez un coup d'œil à l'onglet Index, vous constaterez qu'un index a automatiquement été défini pour cette clé primaire.
Remarque : nous utilisons ici une clé primaire identique à la clé primaire du système opérationnel. Pour avoir une dimension conforme, on utilise généralement une nouvelle clé, indépendante.

OK

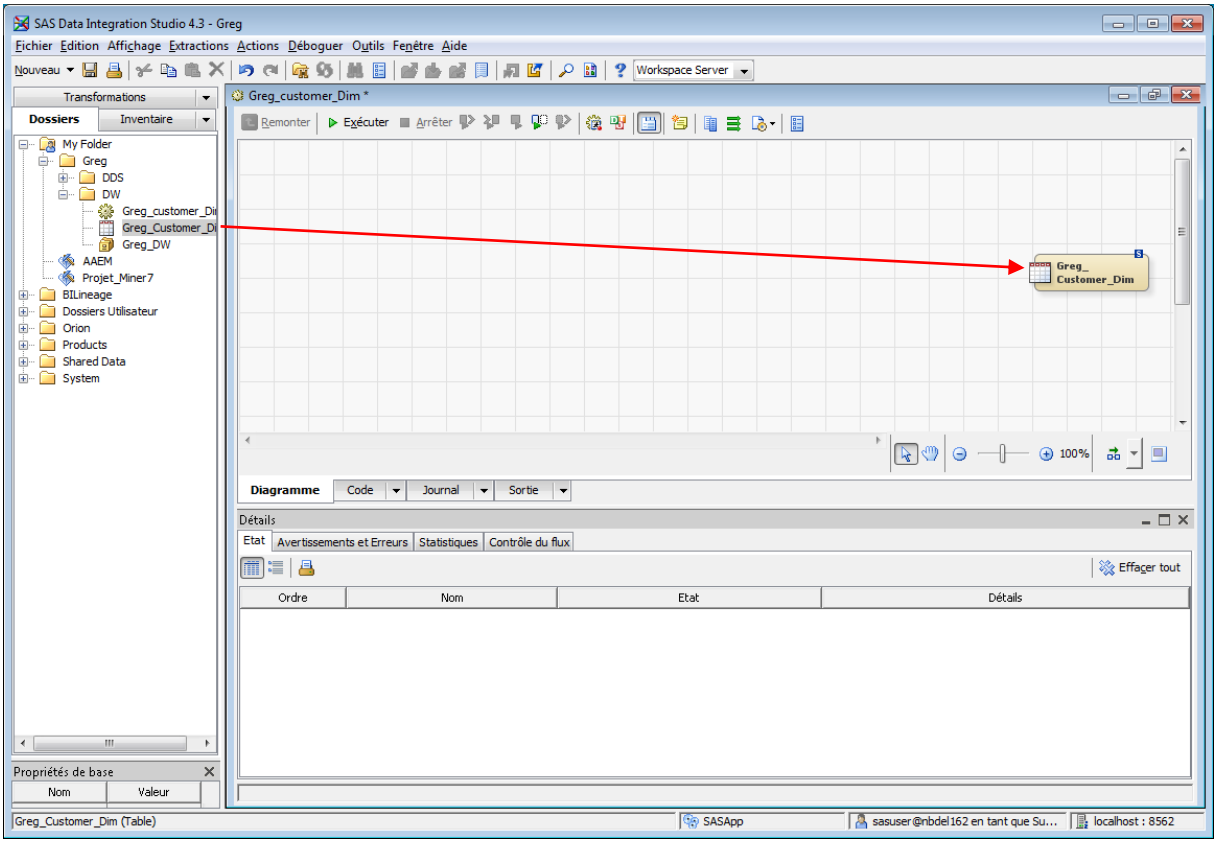
Nous avons défini les métadonnées de la table Customer_DIM du Data Warehouse Orion Star. Ces métadonnées ne sont que le squelette de la table, il faut maintenant le remplir par un processus de transformation qui liera et transformera les données du Data Store et les chargera dans cette structure.



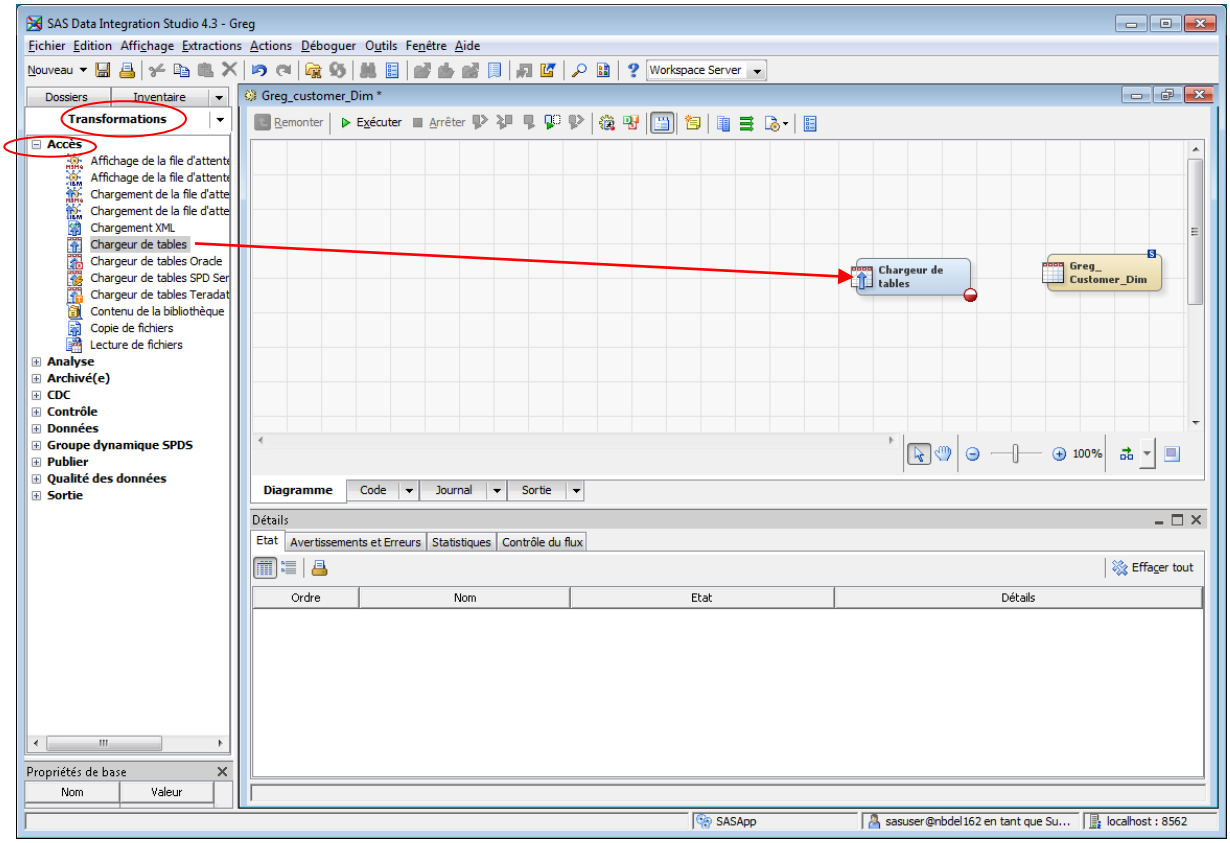
Sur votre dossier, clic-droit → **Nouveau → Job**



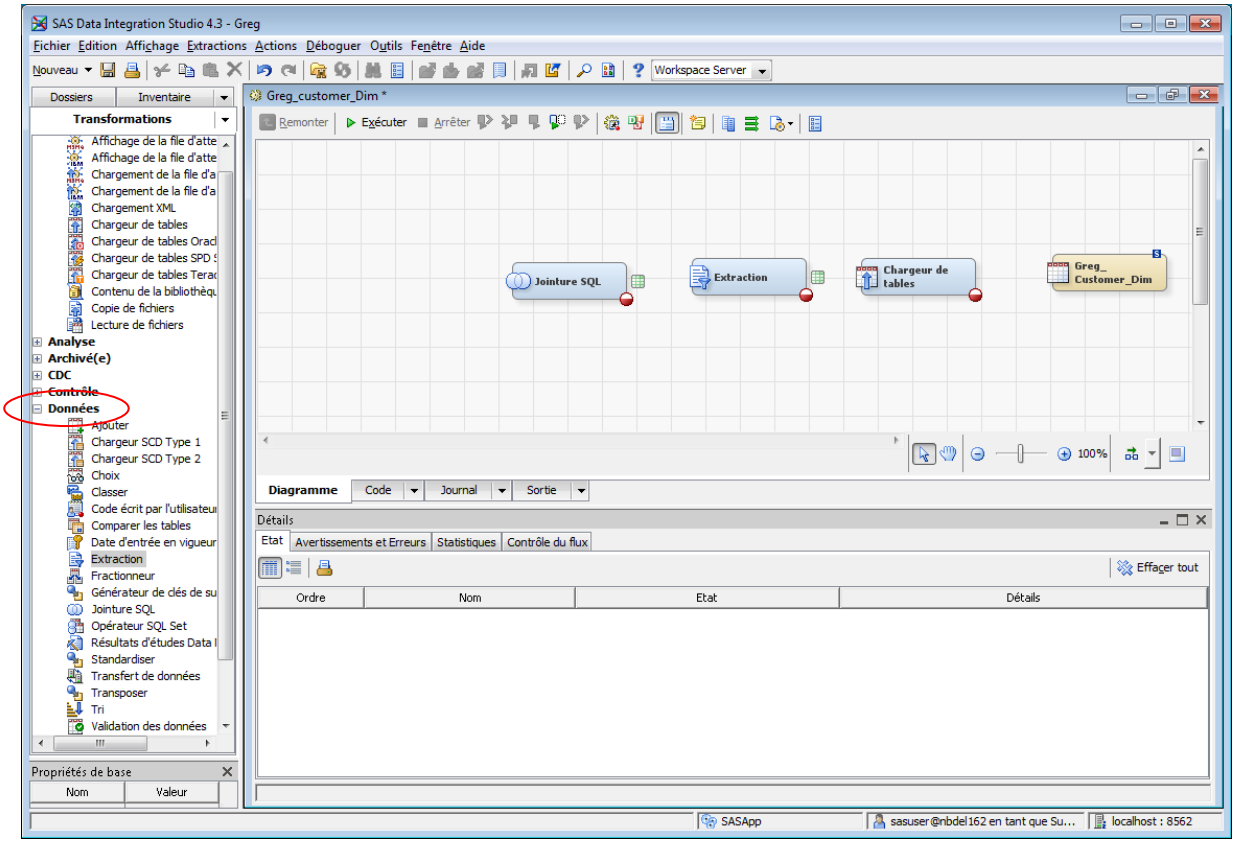
Donner un nom à votre job commençant par votre nom.
OK



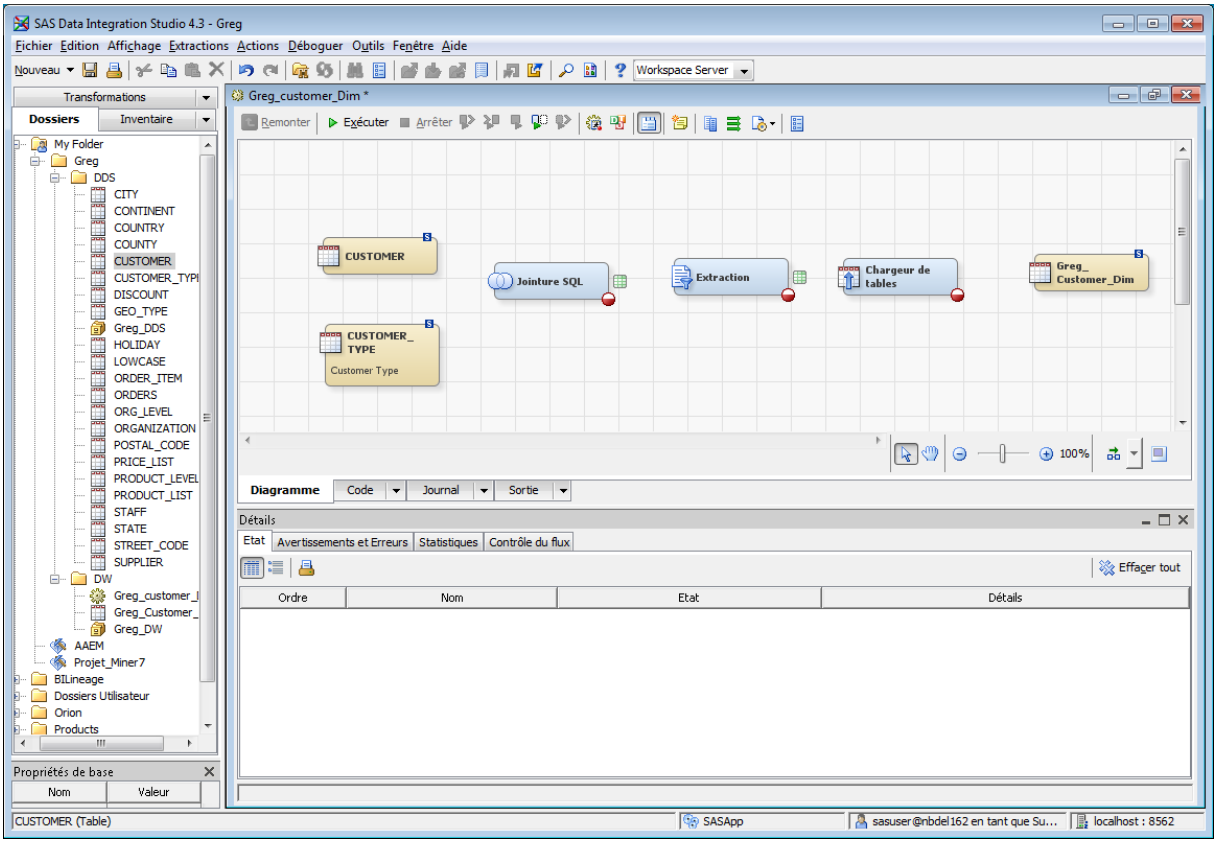
Depuis l'onglet Dossier, sélectionnez votre table et glissez la dans le processus.



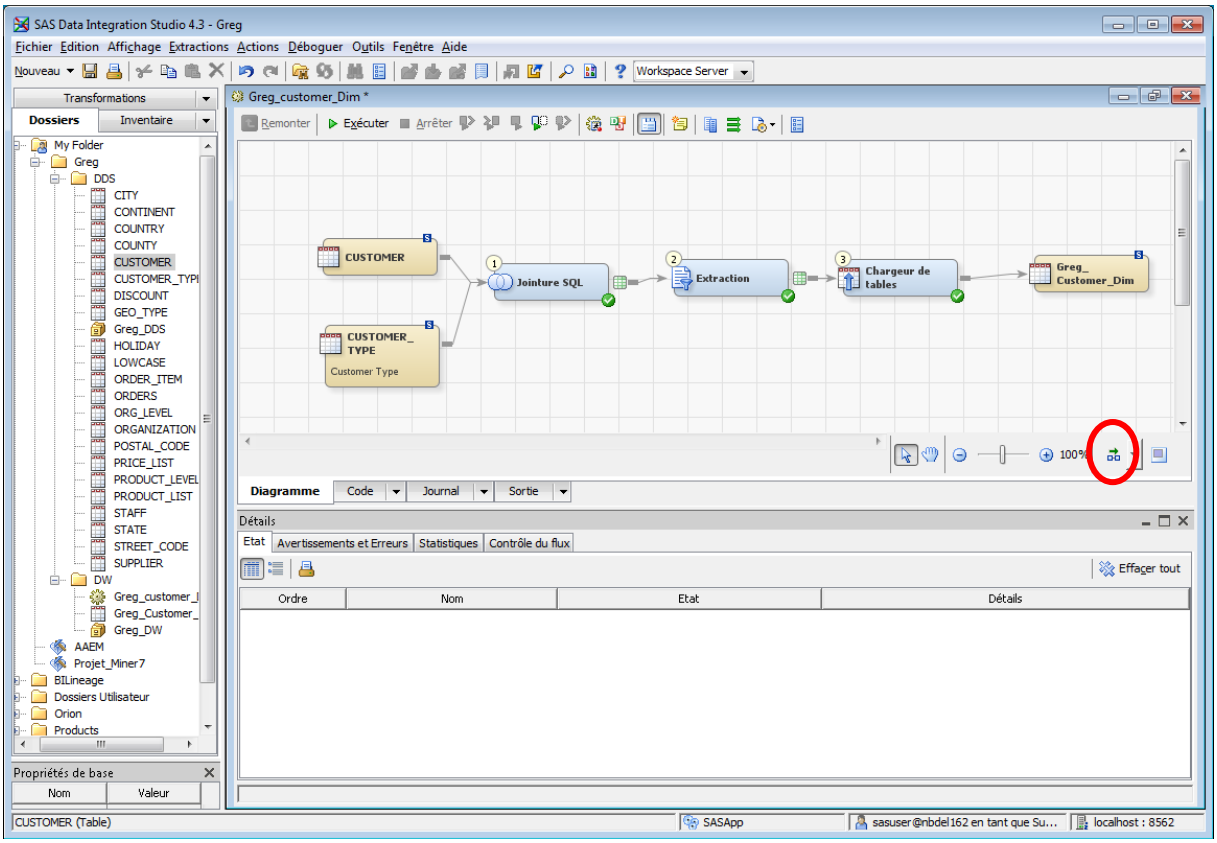
Ajouter le **chargeur de table** depuis le dossier **Accès** se trouvant dans l'onglet **Transformations**



Ajouter l'outil d'**extraction** et la **jointure SQL** depuis le dossier **Données** de l'onglet **Transformation**.



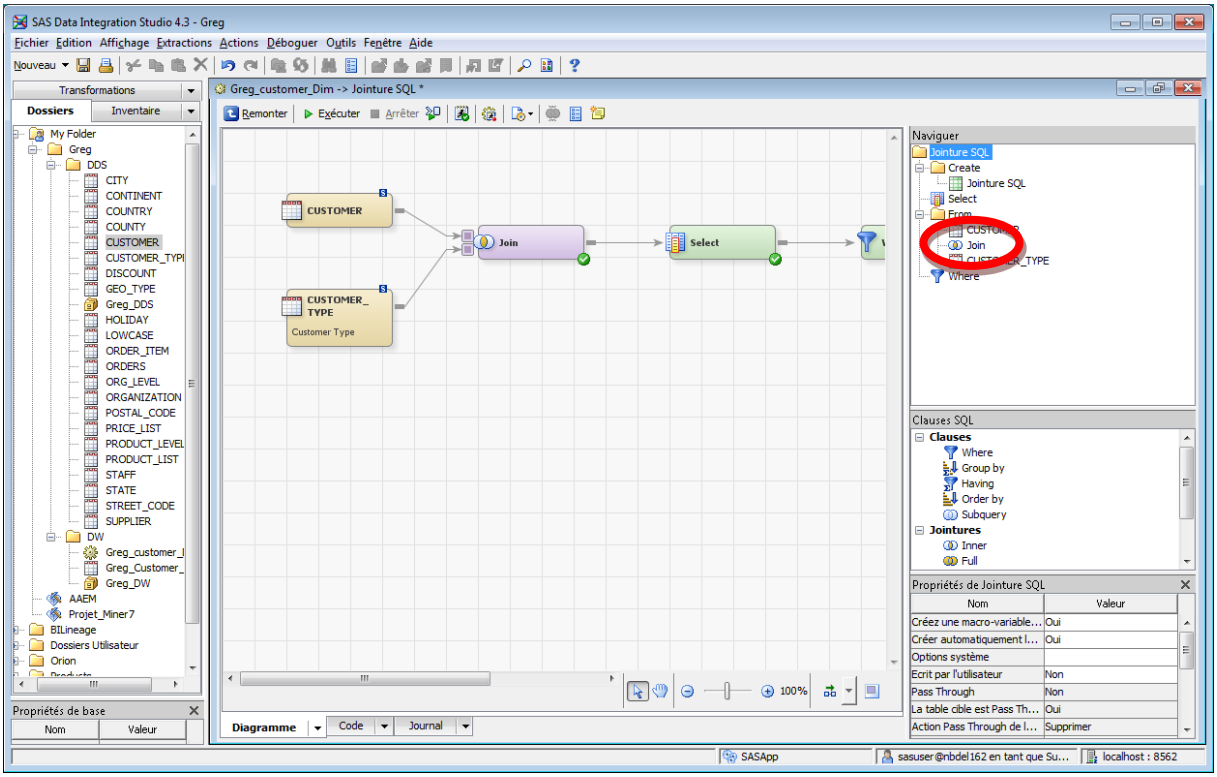
Ajouter les tables **Customer** et **Customer_Type** depuis le dossier **Mon Dossier → votre nom → DDS**, dans l'onglet **Dossier**



Tracer les flèches entre les éléments comme sur le schéma ci-dessus.
Quand la souris est à la fin d'un nœud, est prend l'image d'un crayon. Cliquer, glisser et lâcher sur le nœud suivant.

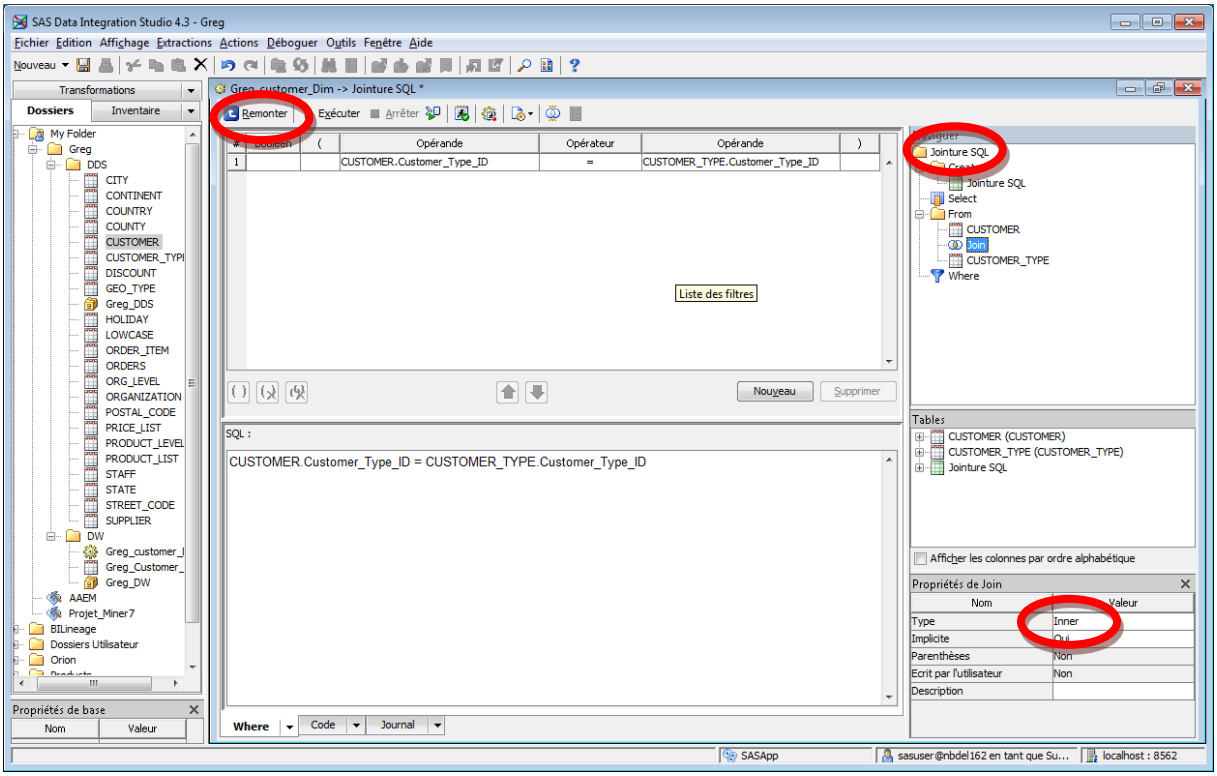
Le bouton en bas à droite du diagramme permet d'aligner les éléments.

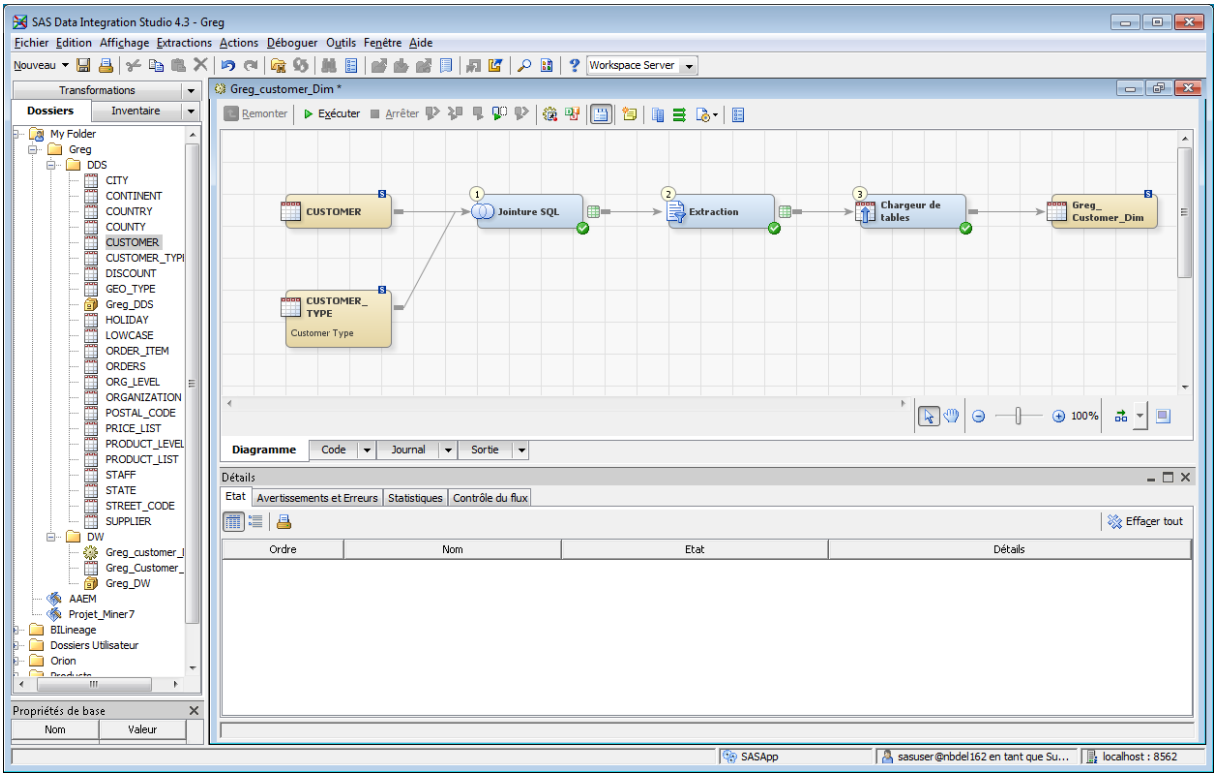
Dans les propriétés de la jointure SQL (double cliquer sur la jointure SQL ou clic-droit → propriétés), dans l'onglet « tables » :



Vérifier que la jointure est bien faite entre les tables Customer_Type et Customer.

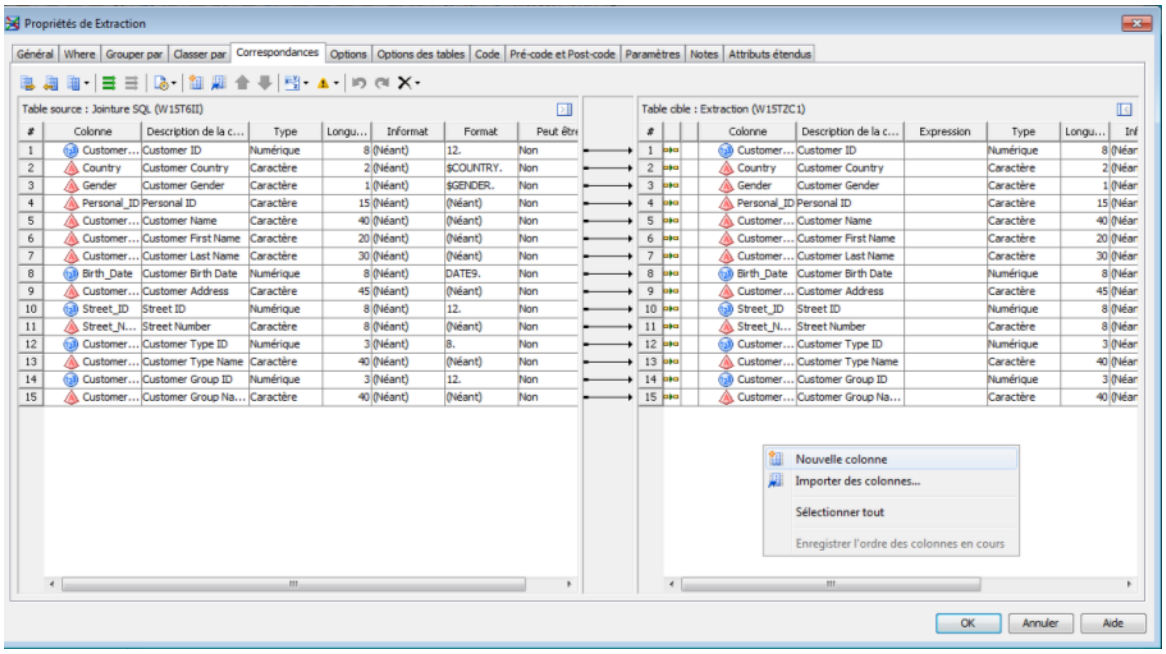
Si vous sélectionnez le « Join », vous verrez que la jointure est de type « inner join » par défaut. Pour revenir à l'assistant des requêtes SQL, cliquer sur **Jointure SQL** en haut à droite.
Cliquez sur le dossier de la jointure SQL (en haut à droite) pour revenir au digramme de la jointure SQL ou cliquez sur remonter (en haut à gauche) pour revenir au digramme général.





Ouvrir les propriétés de l'Extraction. (double cliquer dessus ou clic-droit, puis propriété)

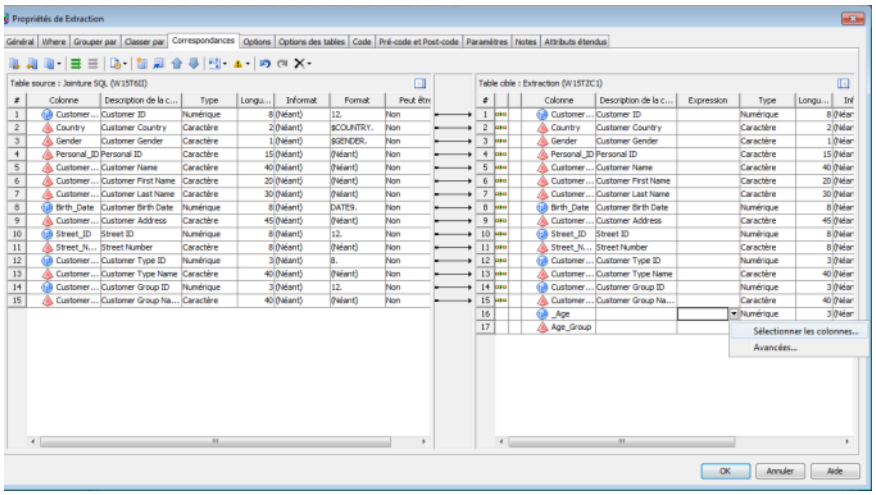
Dans l'onglet des correspondances, créer deux nouvelles colonnes :



Clic-droit → nouvelle colonnes

Nom	Expression	Type	longueur
_AGE	FLOOR(YRDIF(Birth_Date ,TODAY(),'actual'))	Numérique	3
Customer_Age_Groupe	PUT(CALCULATED _AGE ,agegroup.)	Caractère	12

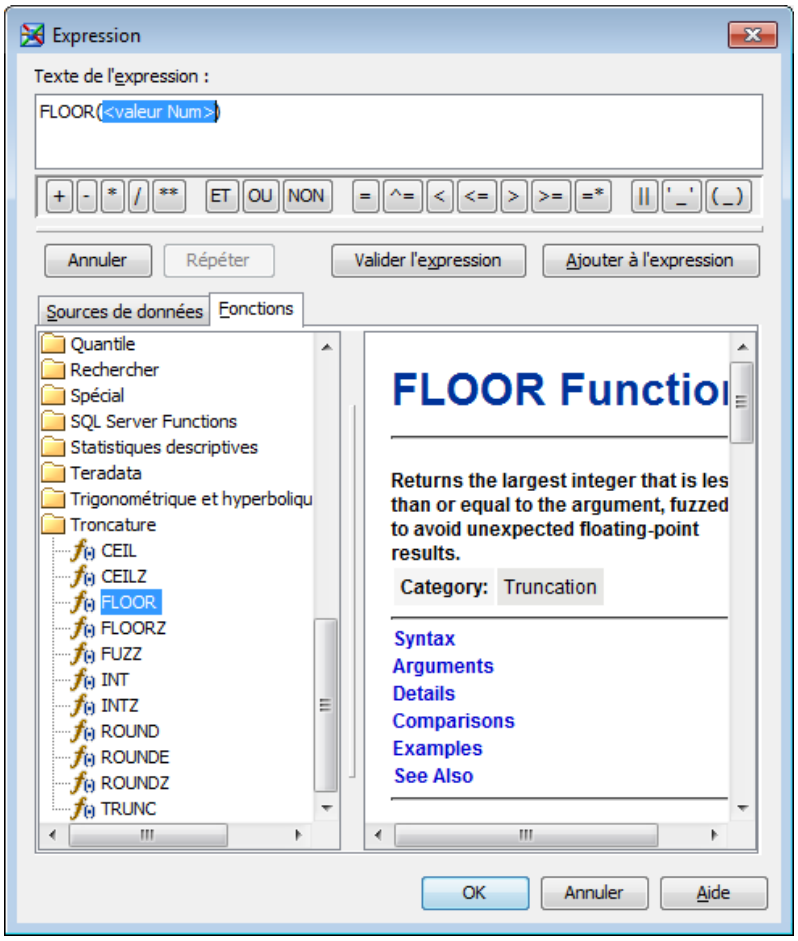
Entrer le nom, le type et la longueur, puis double cliquer dans « expression » et sélectionner « avancées »



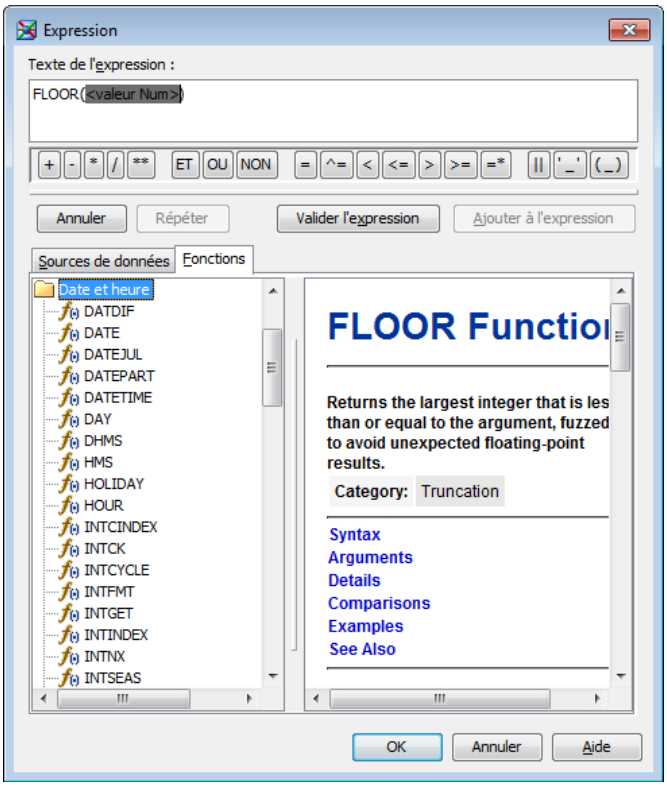
L'expression de l'âge est une troncature de l'argument (floor) d'une différence de date (dans date et heure) sélectionnez YRDIF.
Dans source de données, sélectionnez pour le premier argument la colonne Birth_Date, pour le deuxième argument, dans les fonctions date & heure, sélectionnez la date du jour : Today, pour le deuxième argument, et tapez 'actual' pour le dernier.

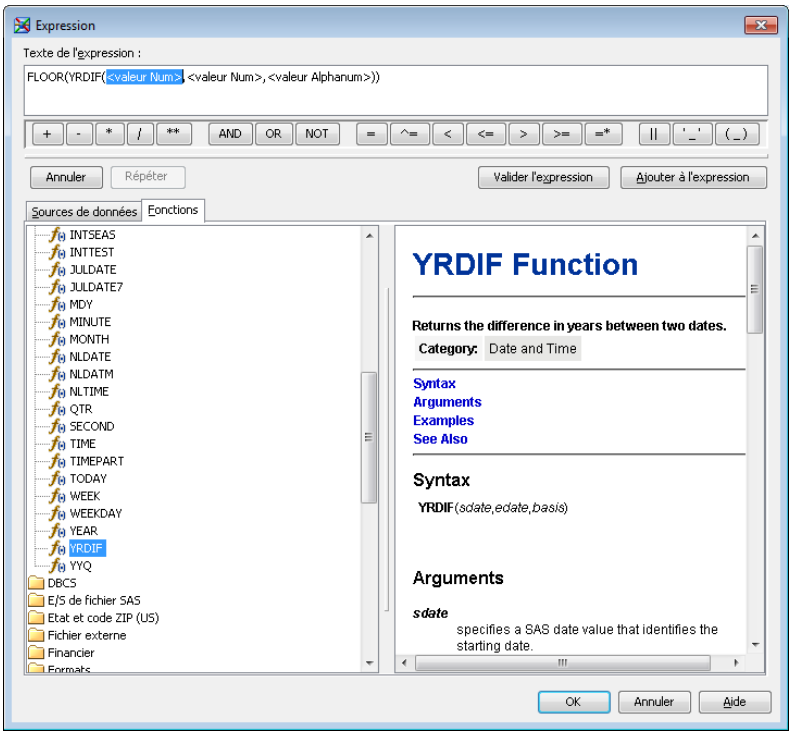
L'expression est donc :
floor(YRDIF(Birth_Date ,today(),'actual'))
Soit :

Dans la colonne Expression, clic-droit → **Avancée**

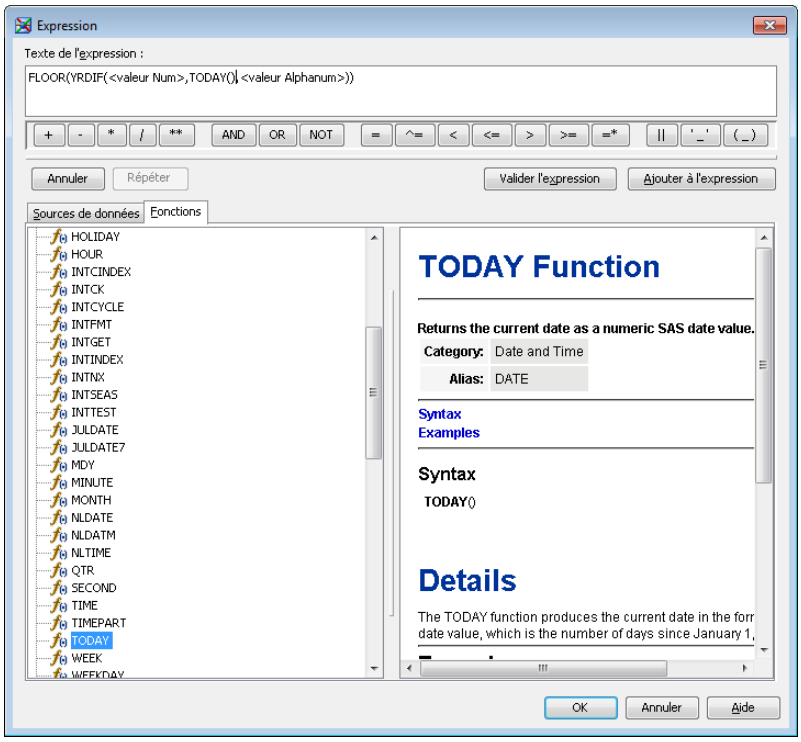


Dans les **Fonctions → Troncature → FLOOR**

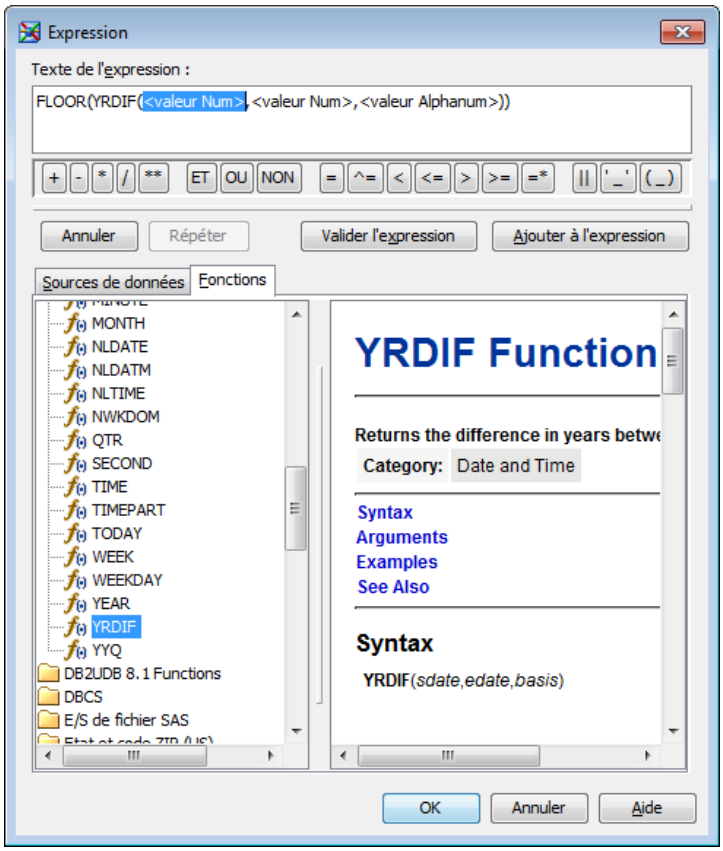


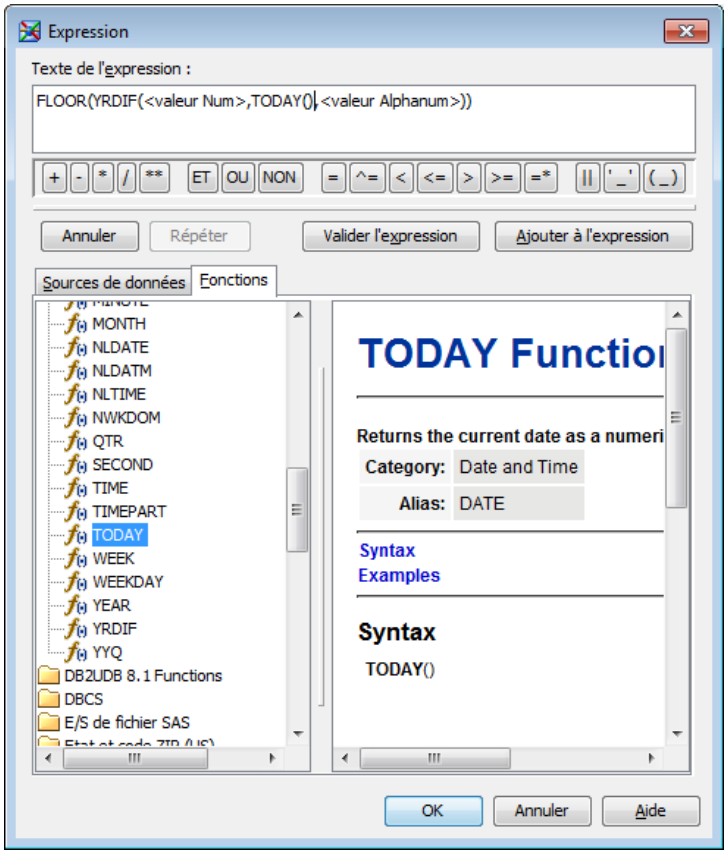


Dans les **Fonctions → Date et heure → YRDIF**

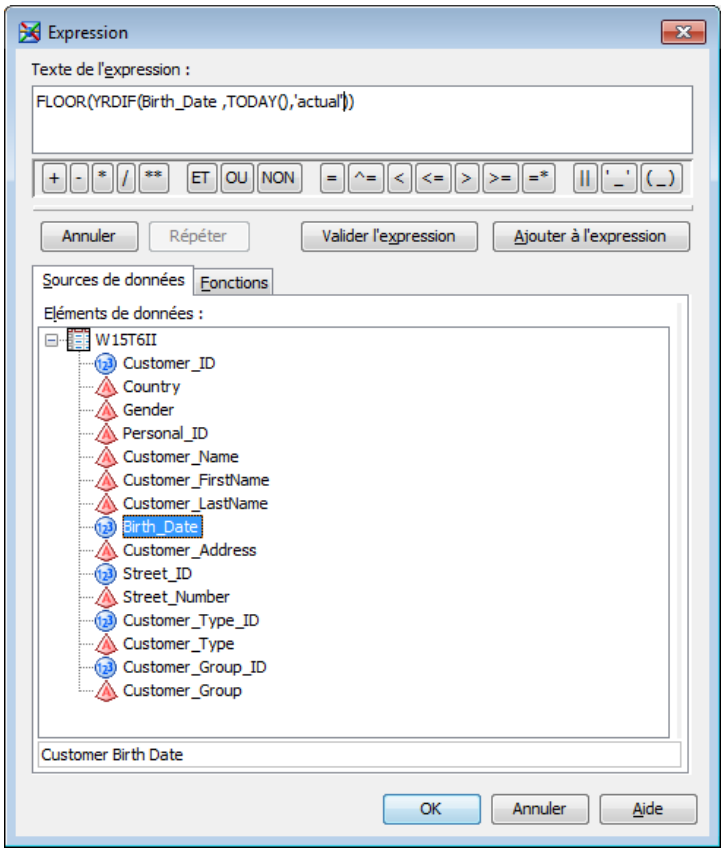


Dans les **Fonctions →Date et heure → TODAY()** (dans le second élément de YRDIF)



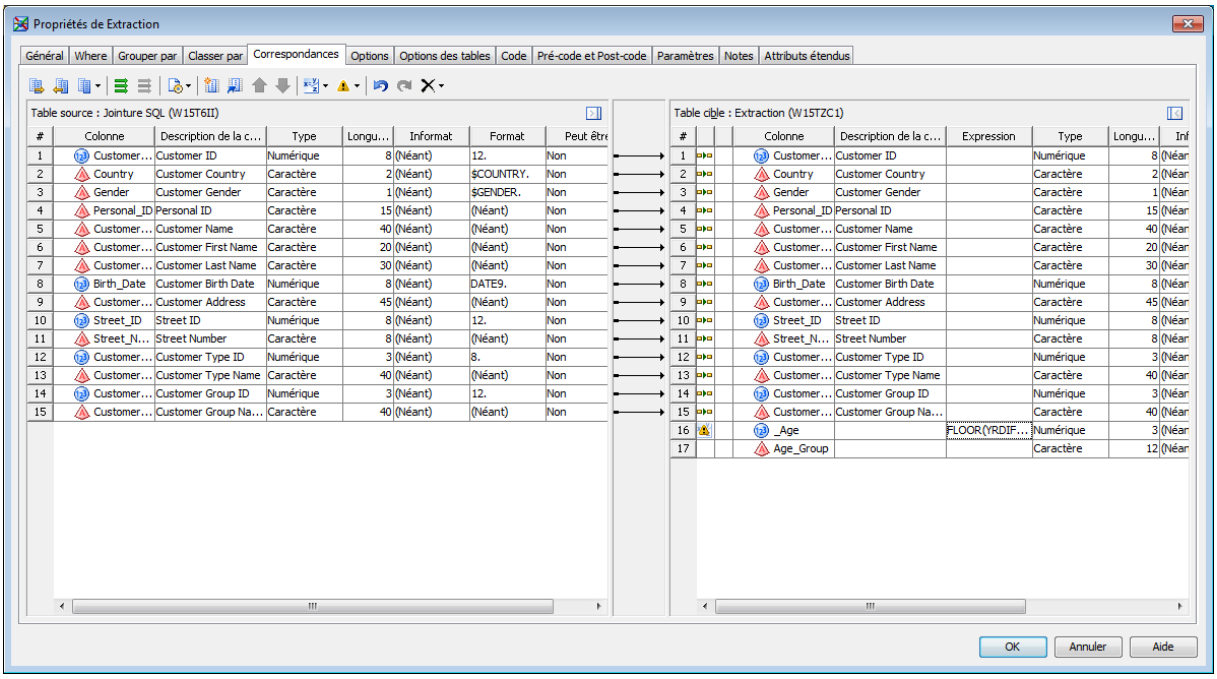


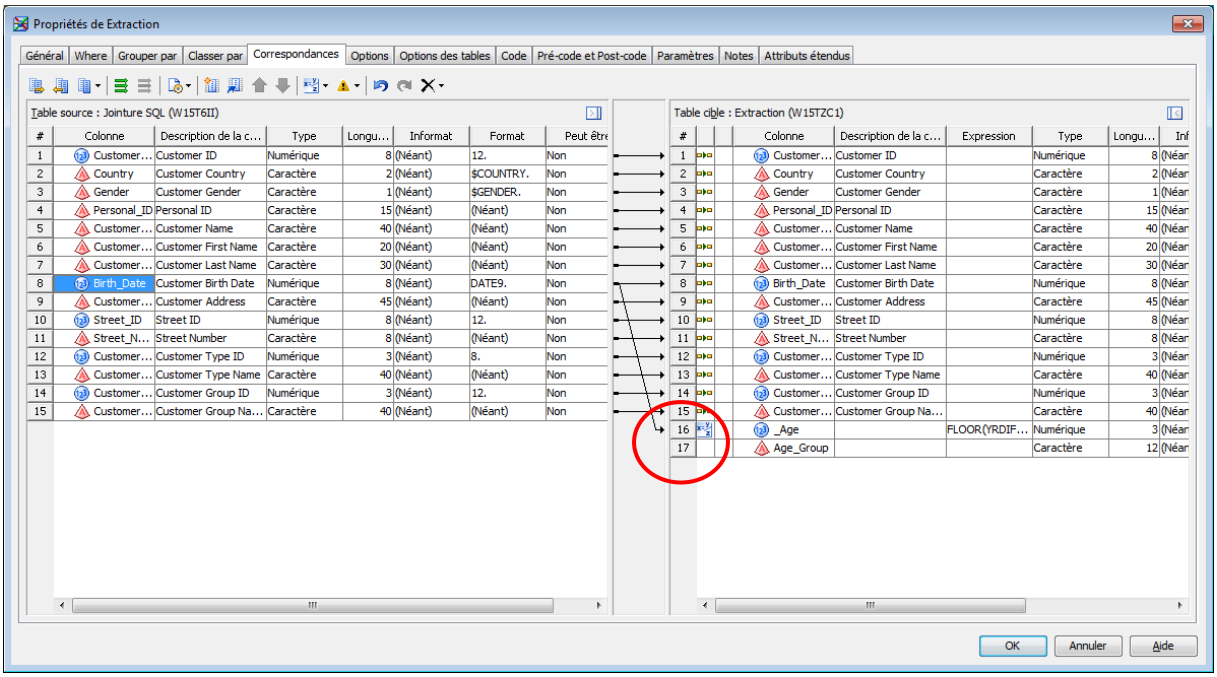
Dans les **Sources de données → Birth_Date** (dans le premier élément de YRDIF)



Dans le dernier élément de YRDIF, taper ‘actual’, en minuscule, entre simple cote.

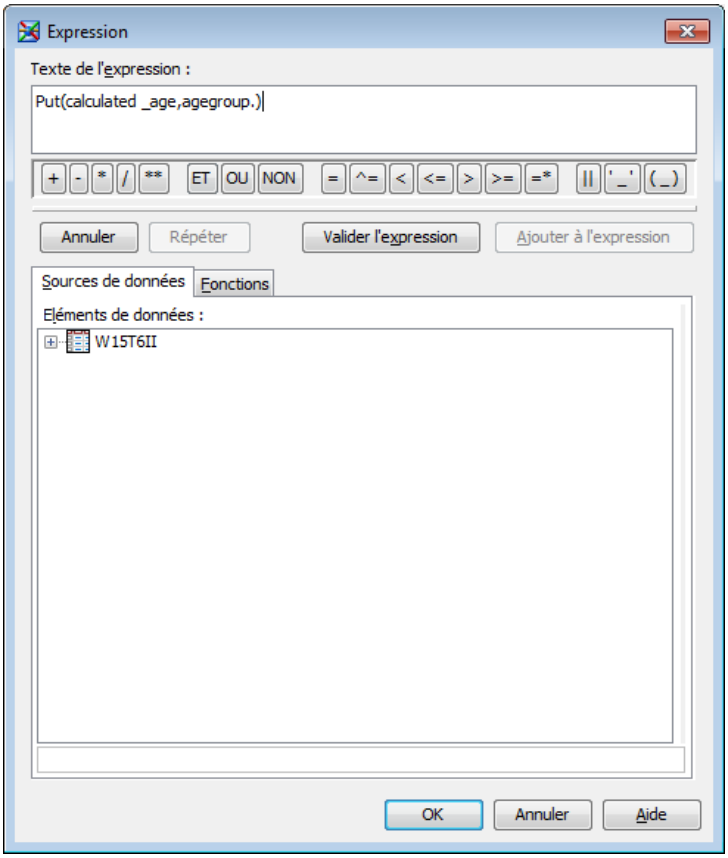
OK.





Pour que l'icône de la colle intermédiaire créé soit valide, il faut créer le lien entre la colonne source Birth_Date et la colonne cible _Age. Pour cela, Clic-droit sur la colonne d'entrée Bith_Date, glisser, lâcher sur la colonne cible _Age. Le symbole devant la ligne de la colonne de l'âge n'est plus un point d'exclamation, mais celui d'une colonne calculée.

Créer l'expression du groupe d'âge :



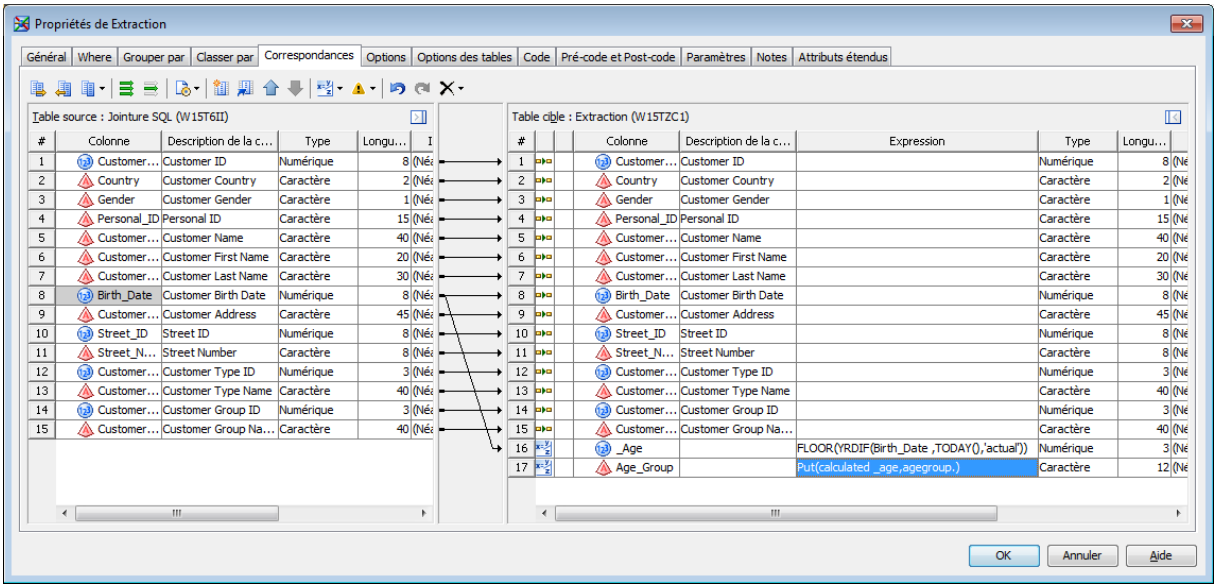
Créer l'expression de Customer_age_groupe : PUT(CALCULATED _AGE ,agegroup.)
Attention : il y a un espace entre « CALCULATED » et « _AGE » et un point après « agegroup. »

« _Age » est le nom de la colonne intermédiaire créé précédemment.
Ce peut être en minuscule ou en majuscule.

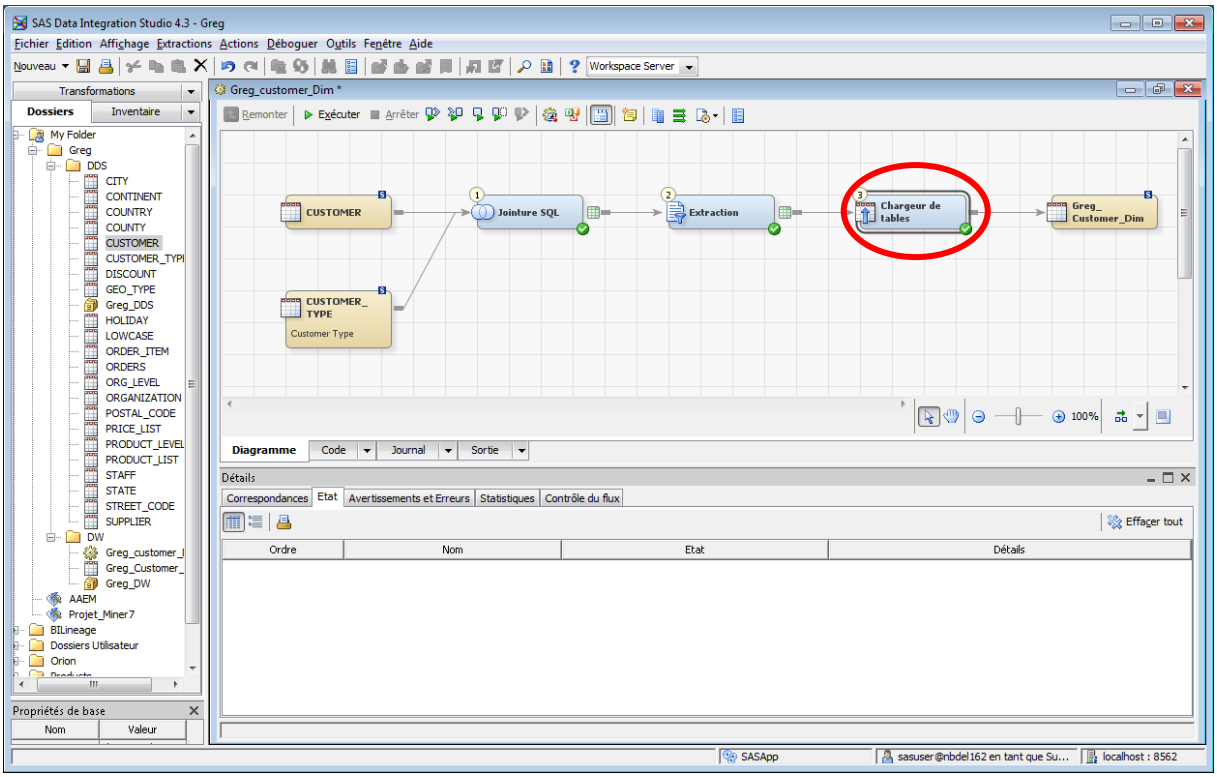
Le format “agegroup.” est un format SAS. Il se trouve dans les formats fournis avec le cas Orion. Il permet la transformation suivante :

- Si l’âge est compris entre 15 et 30, il écrit la chaine de caractère « 15-30 years »
- Si l’âge est compris entre 30 et 45, il écrit la chaine de caractère « 31-45 years »
- Si l’âge est compris entre 45 et 60, il écrit la chaine de caractère « 46-60 years »
- Si l’âge est compris entre 60 et 75, il écrit la chaine de caractère « 61-75 years »
- Si, non, il écrit l’âge.

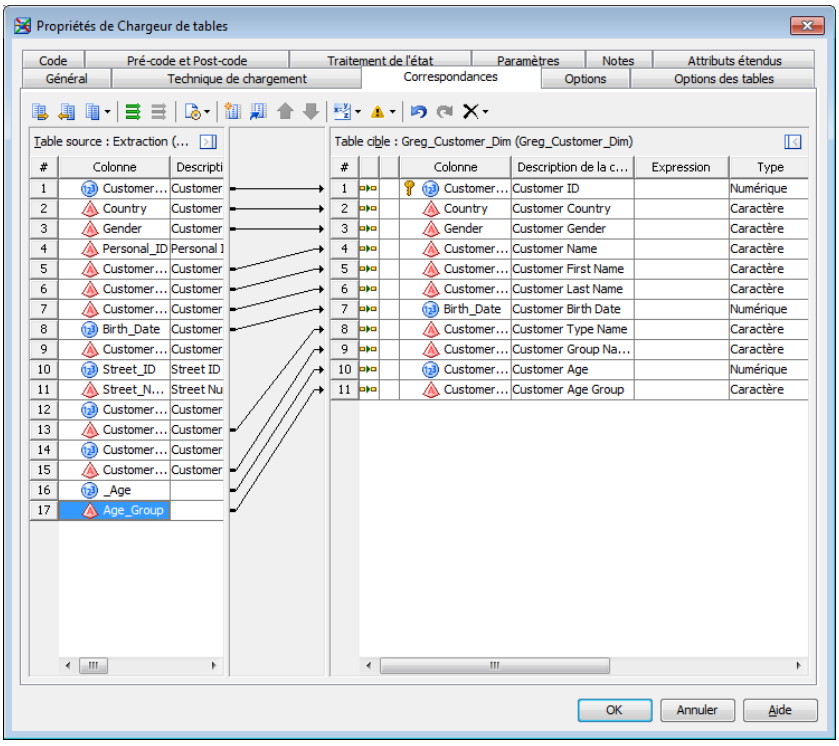
OK.

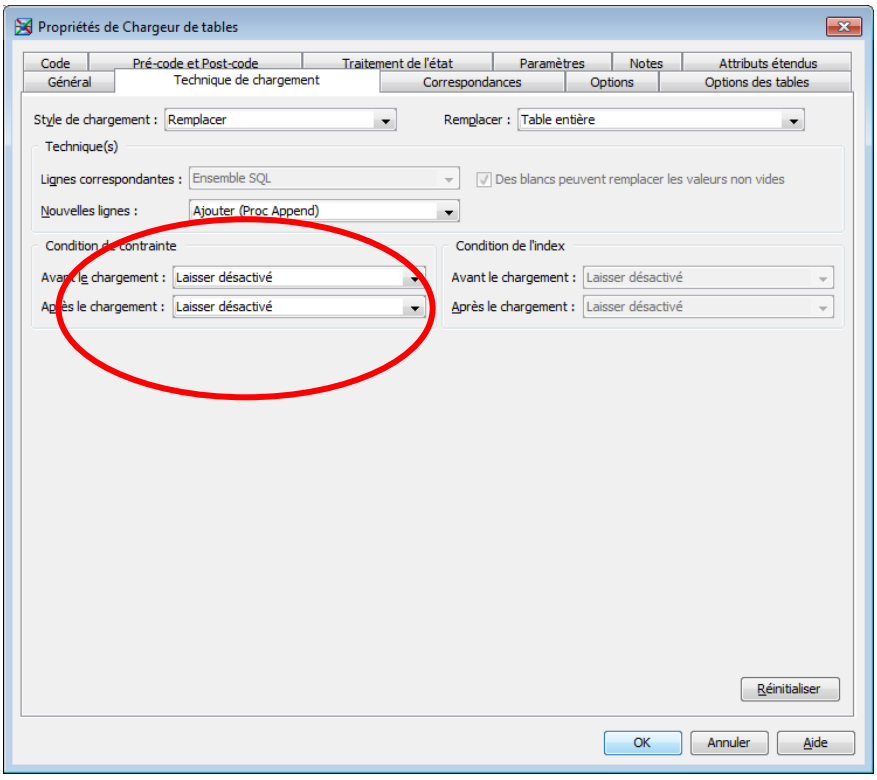


Dans l’outil de chargement, dans les correspondances, mettre à jour les correspondances.
Il faut fermer en cliquant sur **OK** les propriétés de l’extraction où vous avez créé deux colonnes intermédiaires et aller dans l’outil de chargement, dans l’onglet des correspondances, pour les mettre à jour.



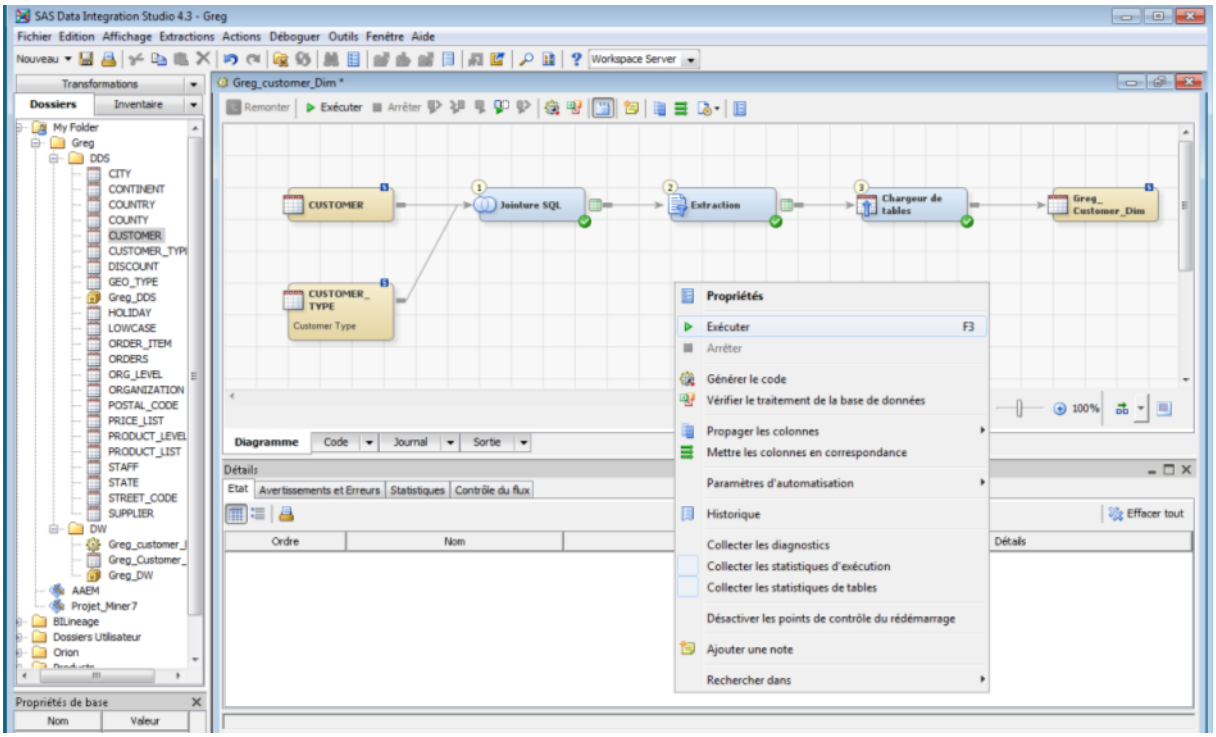
Il faut créer le lien entre les colonnes intermédiaires et les colonnes cibles.





Sélectionner la technique de chargement « Remplacer » et « laisser désactivé » les conditions de contrainte.

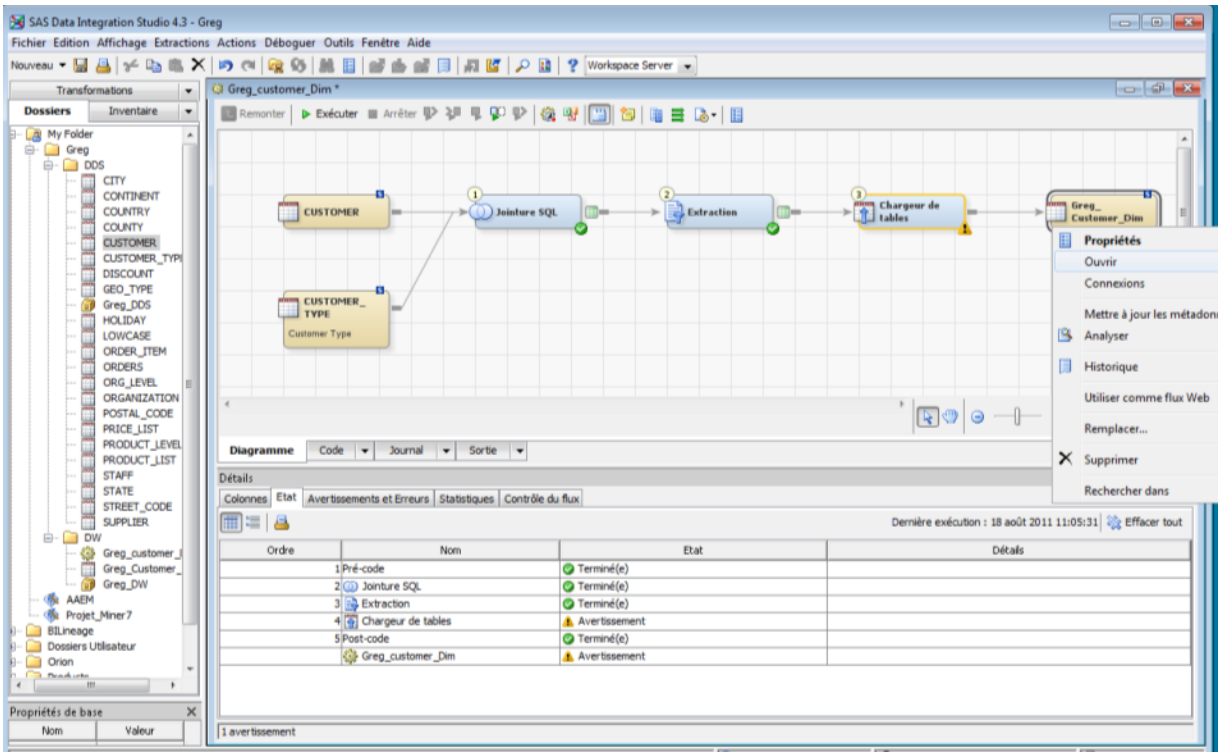
OK

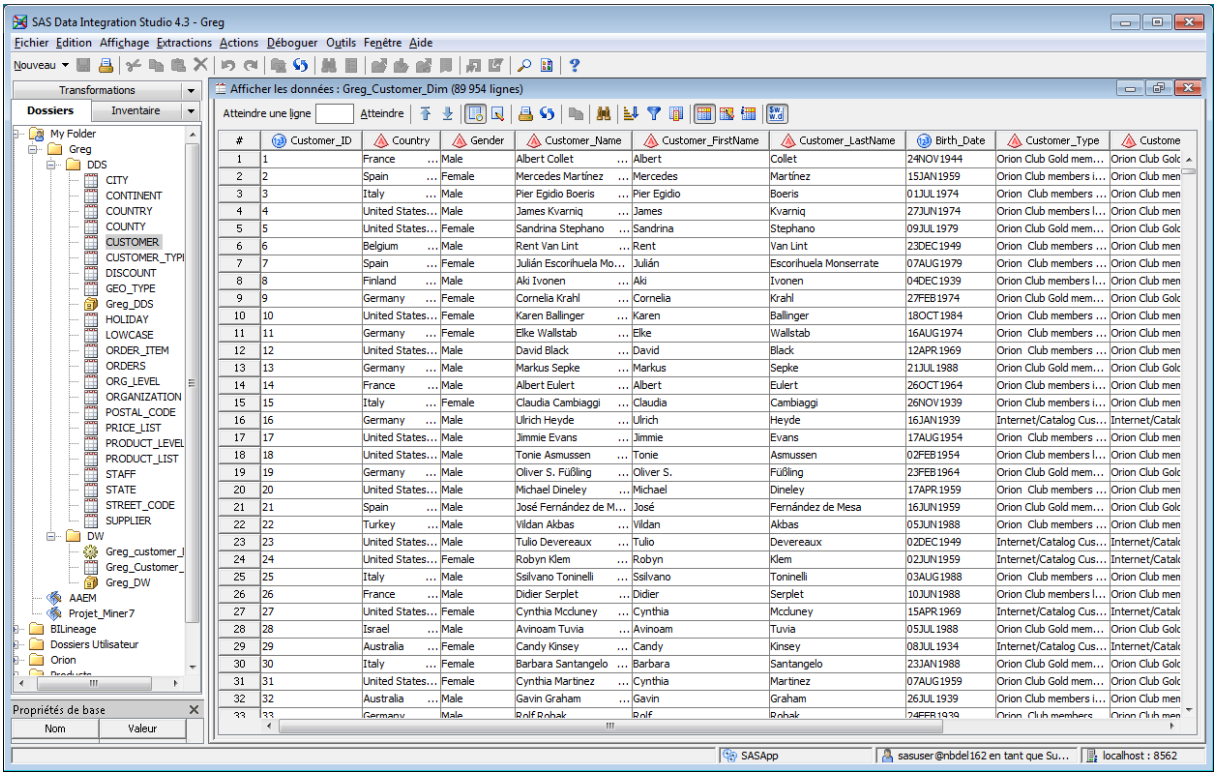


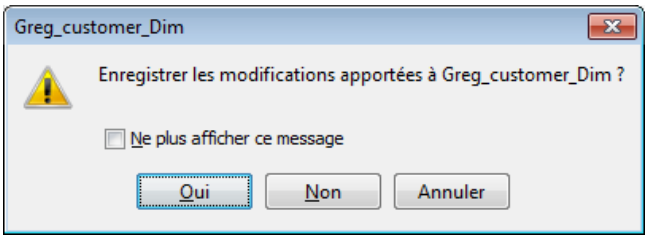
Exécuter le processus

Clic-droit sur le processus → Exécuter.

Si le processus s'est exécuté correctement, allez voir votre table pour vérifier que vous avez bien 89 954 lignes et que toutes les colonnes sont bien chargées.
Pour voir la table, clic-droit sur celle-ci → ouvrir
Dans le cadre de ce TP, on ne s'intéressera pas aux éventuels avertissements.







Fermer le processus de création de la table Votre_Nom_Customer_Dim
Enregistrez les modifications.
Oui.

Création de la table PRODUCT_DIM

Pour créer la table PRODUCT_DIM, il faut les tables Product_List et Supplier, joint par une jointure SQL. Voici la définition des colonnes de cette table.

Nom	Longueur	Type	Format	Remarque
Product_ID	8	Numérique	12.	Identique à la table source
Product_Line	20	Caractère	(Néant)	A créer
Product_Category	25	Caractère	(Néant)	A créer
Product_Group	25	Caractère	(Néant)	A créer
Product_Name	45	Caractère	(Néant)	Identique à la table source
Supplier_Country	2	Caractère	$COUNTRY.	Identique à la table source
Supplier_Name	30	Caractère	(Néant)	Identique à la table source
Supplier_ID	4	Numérique	12.	Identique à la table source

De plus, Product_ID est la clé Primaire.

Les produits sont organisés selon 4 niveaux :

1. Ligne de produit
2. Catégorie de produit
3. Groupe de produit
4. Produit

La table Product_List a été créée en utilisant les formats SAS.
Voici quelques lignes explicatives sur les formats utilisés pour ce processus :
Si vous exécutiez le programme suivant (en ayant modifié le libname) :

```
libname library 'chemin Windows jusqu'au répertoire où se trouvent les formats du cas Orion Star';
proc format library=library fmtlib;
  select proddim;
run;
```

Vous obtiendriez le résultat :

```
„ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ†
,      FORMAT NAME: PRODDIM  LENGTH:   12   NUMBER OF VALUES: 5579       ,
,   MIN LENGTH:   1  MAX LENGTH:  40  DEFAULT LENGTH  12  FUZZ: STD       ,
‡ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ…ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ…ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ‰
,START          ,END            ,LABEL  (VER. V7|V8   08AUG2003:10:05:32),
‡ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ‰
,    210000000000,    210000000000,.                          ,
,    210100000000,    210100000000,210000000000               ,
,    210100100000,    210100100000,210100000000               ,
,    210100100001,    210100100001,210100100000               ,
,    210100100002,    210100100002,210100100000               ,
,    210100100003,    210100100003,210100100000               ,
,    210100100004,    210100100004,210100100000               ,
,    210100100005,    210100100005,210100100000               ,
,    210100100006,    210100100006,210100100000               ,
,    210100100007,    210100100007,210100100000               ,
,    210100100008,    210100100008,210100100000               ,
```

Le format proddim. permet donc :

- A la première application, de passer de l'identifiant du produit à l'identifiant du groupe de produit.
    - A la deuxième application, de passer de l'identifiant du groupe de produit à l'identifiant de la catégorie de produit.
        - A la troisième application, de passer de l'identifiant de la catégorie de produit à l'identifiant de la ligne de produit.

Voici le début du résultat du programme :

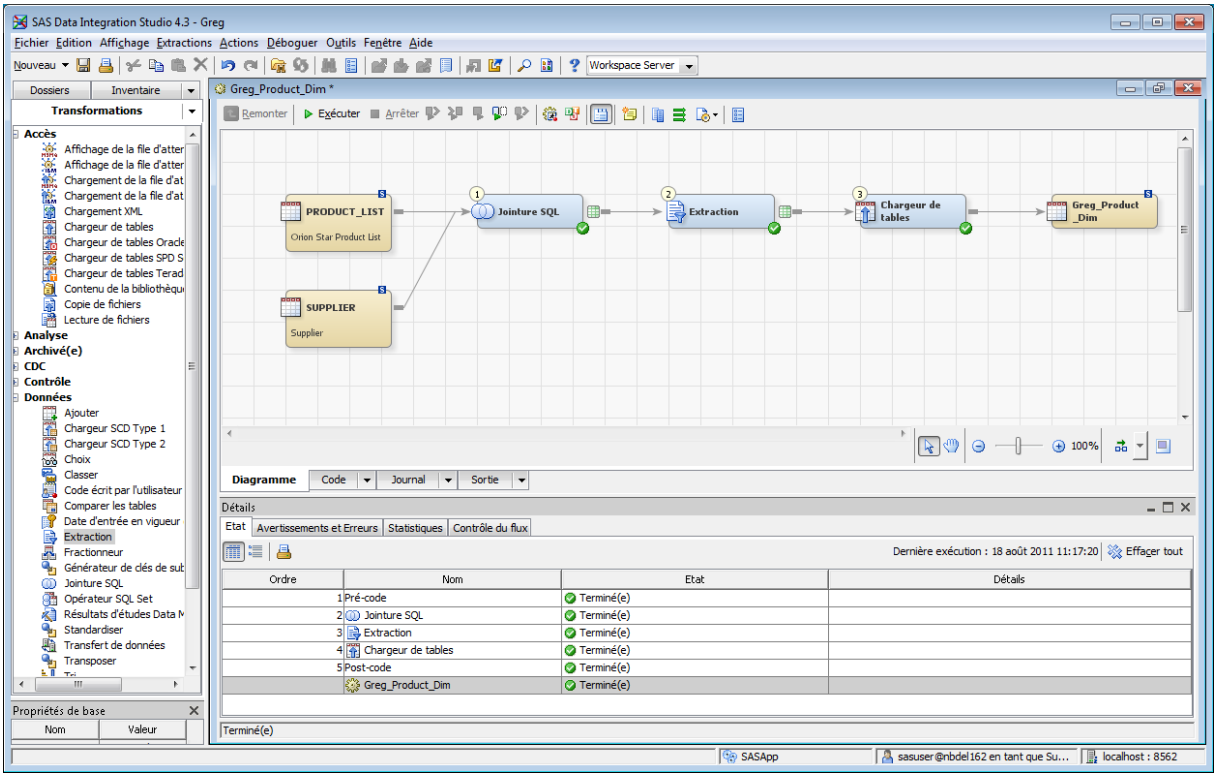
```
libname library 'chemin Windows jusqu'au répertoire où se trouvent les formats du cas Orion Star';
proc format library=library fmtlib;
  select product;
run;
```

```
„ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ†
,      FORMAT NAME: PRODUCT  LENGTH:   45   NUMBER OF VALUES: 5579        ,
,   MIN LENGTH:   1  MAX LENGTH:  45  DEFAULT LENGTH  45  FUZZ: STD       ,
‡ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ…ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ…ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ‰
,START          ,END            ,LABEL  (VER. V7|V8   08AUG2003:10:05:32),
‡ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ‰
,    210000000000,    210000000000,Children                                  ,
,    210100000000,    210100000000,Children Outdoors                          ,
,    210100100000,    210100100000,Outdoor things, Kids                       ,
,    210100100001,    210100100001,Boy's and Girl's Ski Pants with Braces     ,
,    210100100002,    210100100002,Children's Jacket                         ,
,    210100100003,    210100100003,Children's Jacket Sidney                   ,
,    210100100004,    210100100004,Children's Rain Set                       ,
,    210100100005,    210100100005,Children's Rain Suit                      ,
```

Le format product. permet donc

- Avec l'identifiant de la ligne de produit, d'avoir le nom de ligne de produit.
- Avec l'identifiant de la catégorie de produit, d'avoir le nom de la catégorie de produit.
- Avec l'identifiant du groupe de produit, d'avoir le nom du groupe de produit.

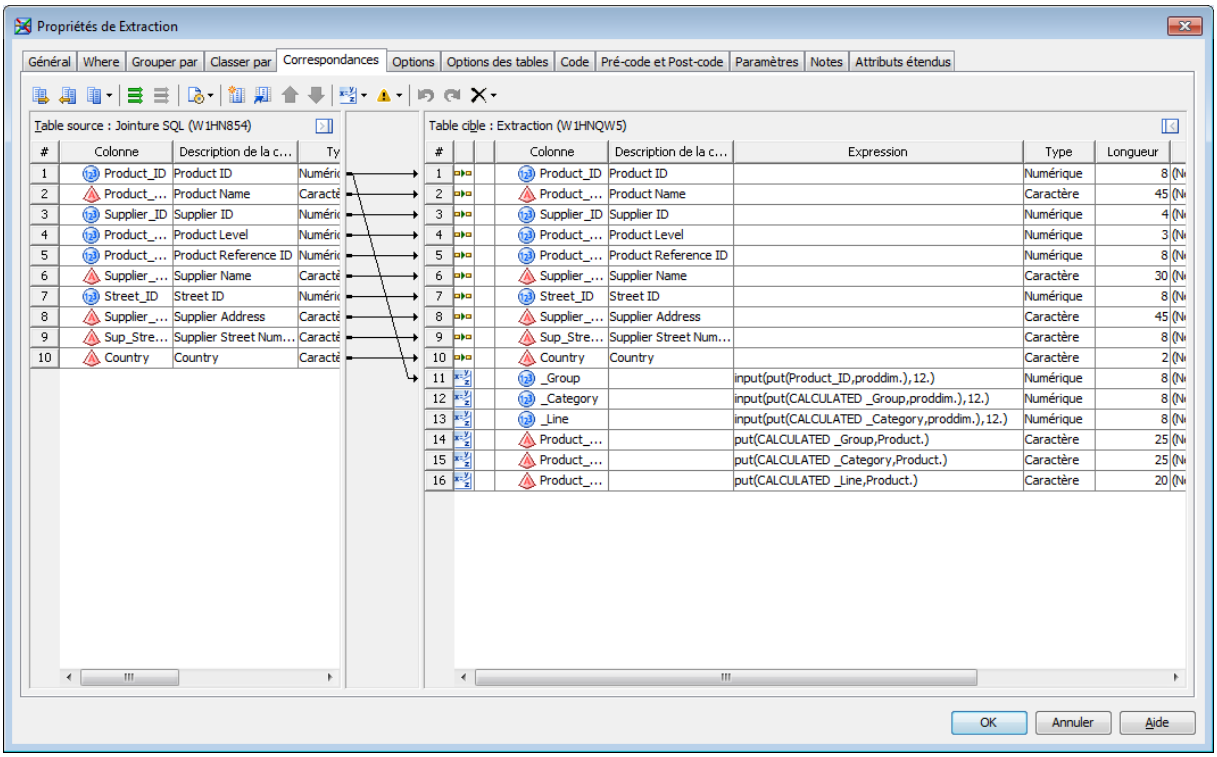
Créer la table Votre_nom_Product_Dim comme vous l'avez fait pour la table Votre_Nom_Customer_Dim et créer le processus suivant :

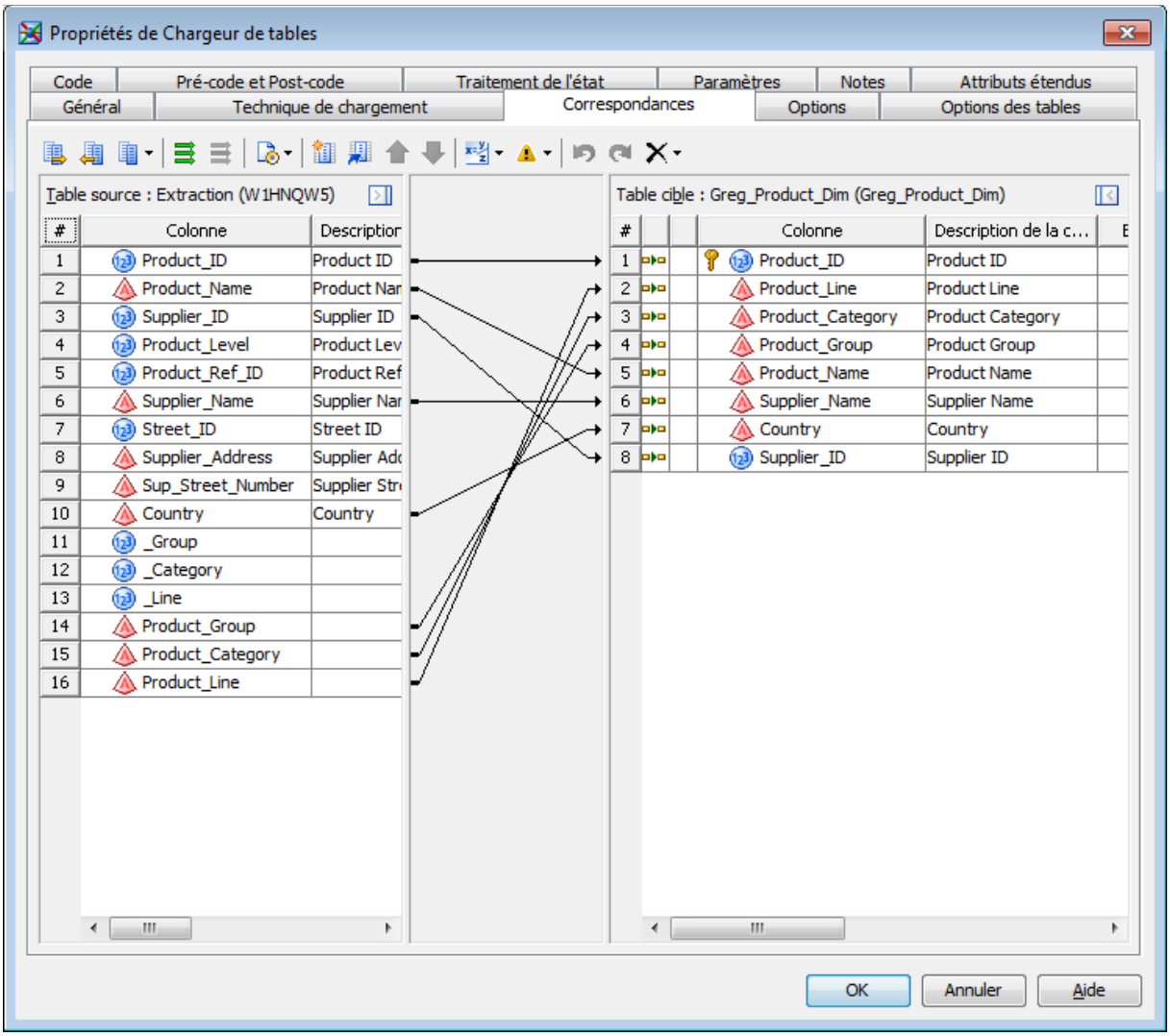


Voici donc le processus ETL de création de la table Product_Dim.

Dans l'outil d'extraction, créer les 6 colonnes suivantes.

Nom	Fonction	Type	longueur
_Group	input(put(Product_ID,proddim.),12.)	Numérique	8
_Category	input(put(CALCULATED _Group,proddim.),12.)	Numérique	8
_Line	input(put(CALCULATED _Category,proddim.),12.)	Numérique	8
Product_Group	put(CALCULATED _Group,Product.)	Caractère	25
Product_Category	put(CALCULATED _Category,Product.)	Caractère	25
Product_Line	put(CALCULATED _Line,Product.)	Caractère	20


Propriétés de Extraction
Général
Where
Grouper par
Classer par
Correspondances
Options
Options des tables
Code
Pré-code et Post-code
Paramètres
Notes
Attributs étendus
Table source : Jointure SQL (W1HN854)
Table cible : Extraction (W1HNQW5)
OK
Annuler
Aide


Propriétés de Chargeur de tables
Code
Pré-code et Post-code
Traitement de l'état
Paramètres
Notes
Attributs étendus
Général
Technique de chargement
Correspondances
Options
Options des tables
Table source : Extraction (W1HNQW5)
Table cible : Greg_Product_Dim (Greg_Product_Dim)
OK
Annuler
Aide

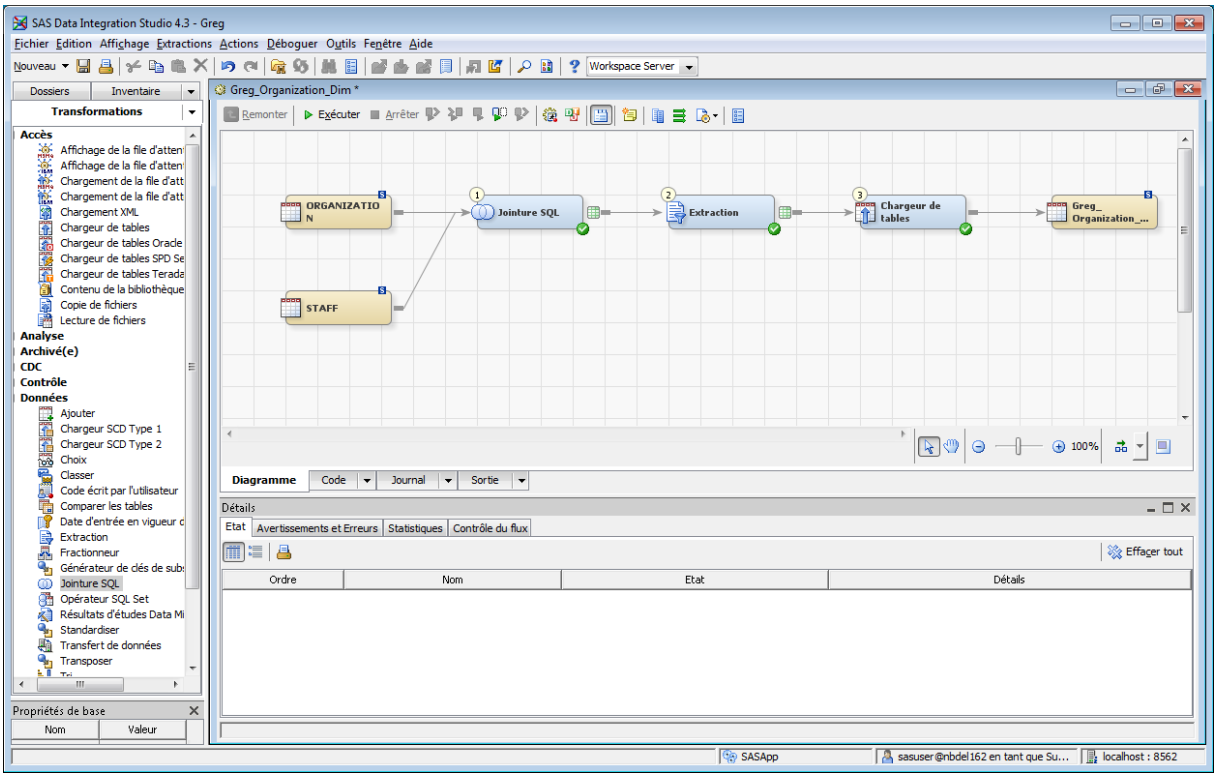
Dans l’outil de chargement, mettre à jour les correspondances.
Exécuter le processus.
Vous devez avoir 5 504 lignes.

Création de la table ORGANIZATION_DIM

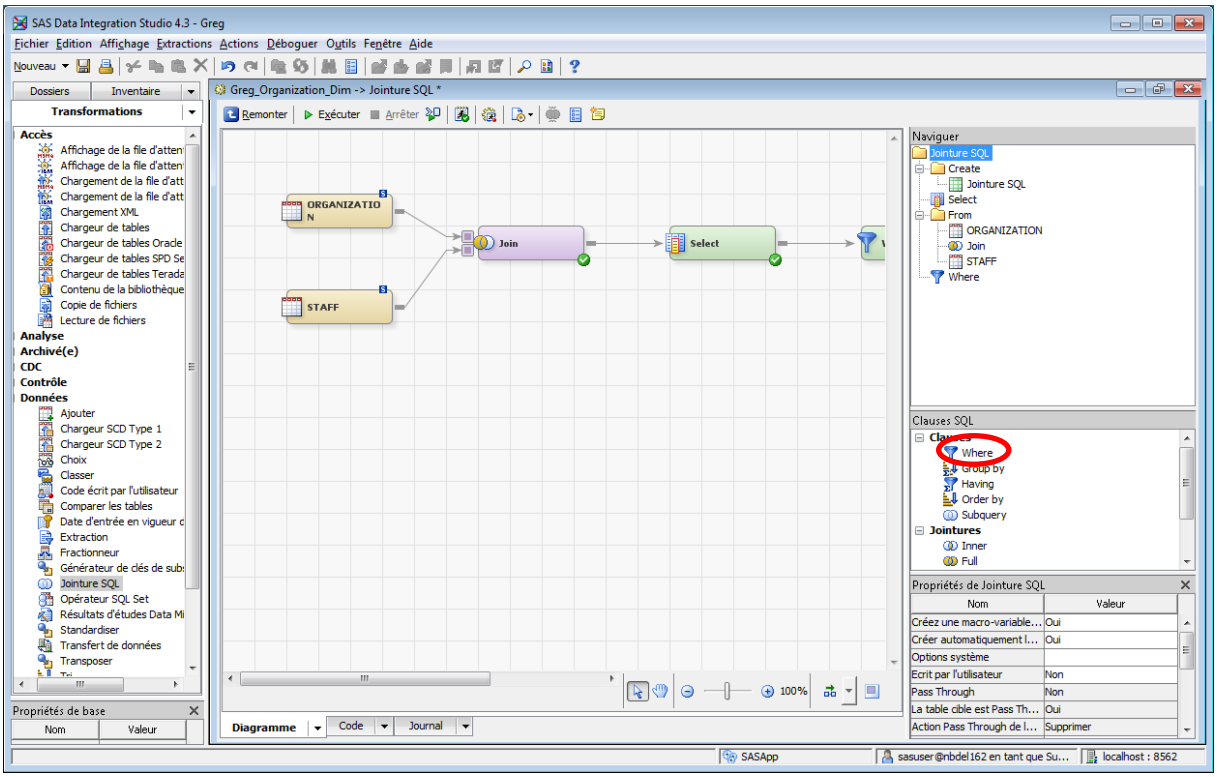
Création de la table ORGANIZATION_DIM avec les informations ci-dessous.
Pour la jointure SQL, il faut faire une jointure droite ou gauche entre les tables STAFF et ORGANIZATION en fonction de la table se trouvant à droite, et mettre une clause de sélection sur le premier niveau de l'organisation.
Il y a 1 048 lignes dans la table Staff, une par employé ayant travaillé, et 1 489 dans la table Organization, une par collaborateur, groupe, section, département ou compagnie. On souhaite 1 049 lignes dans la table de dimension de l'organisation, soit une par collaborateur, plus une pour les ventes faites par internet ou par catalogue. On sélectionne donc uniquement toutes les lignes de la table de l'organisation pour le premier niveau, le niveau collaborateur, soit 1 049.

Nom	Longueur	Type	Format	Remarque
Employee_ID	8	Numérique	12.	Identique à la table source
Employee_Country	2	Caractère	$COUNTRY.	Identique à la table source
Company	30	Caractère	(Néant)	A créer
Department	40	Caractère	(Néant)	A créer
Section	40	Caractère	(Néant)	A créer
Group	40	Caractère	(Néant)	A créer
Job_Title	25	Caractère	(Néant)	Identique à la table source
Employee_Name	40	Caractère	(Néant)	Identique à Org_name
Employee_Gender	1	Caractère	$GENDER.	Identique à la table source
Salary	8	Numérique	DOLLAR13.	Identique à la table source
Employee_BirthDate	8	Numérique	DATE9.	Identique à la table source
Employee_Hire_Date	8	Numérique	DATE9.	Identique à la table source
Employee_Term_Date	8	Numérique	DATE9.	Identique à la table source

Clé Primaire : Employee_ID



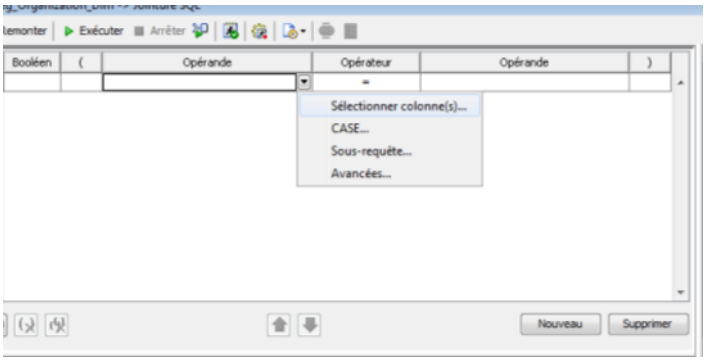
Après avoir ajouté l'extraction, outil qui sera utilisé pour créer de nouvelles colonnes, ajouter l'outil de jointure SQL puis la table Organization et enfin la table Staff.



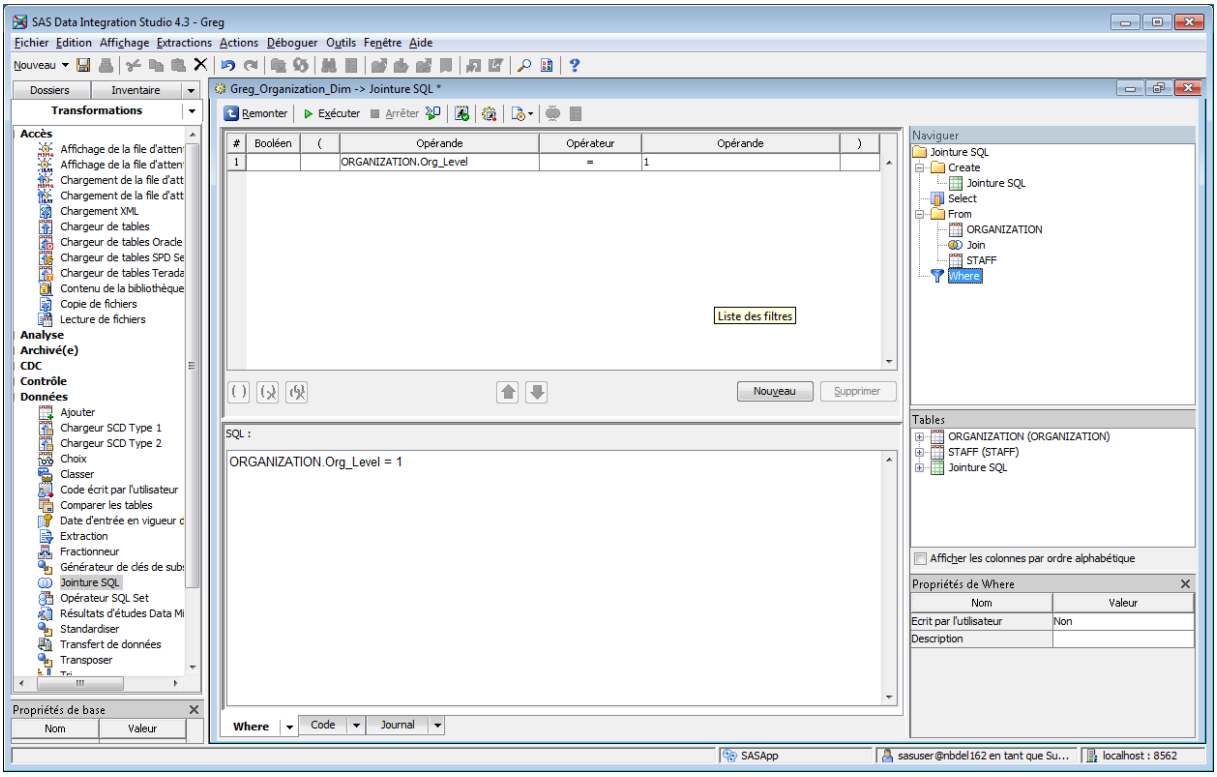
Dans l'assis tant de la jointure SQL, définir une jointure gauche sur la table organization.
Cliquer sur l'icône dans l'assistant
Dans les propriétés, changer en « Left »
Vous souhaitez toutes les lignes de la table organization.

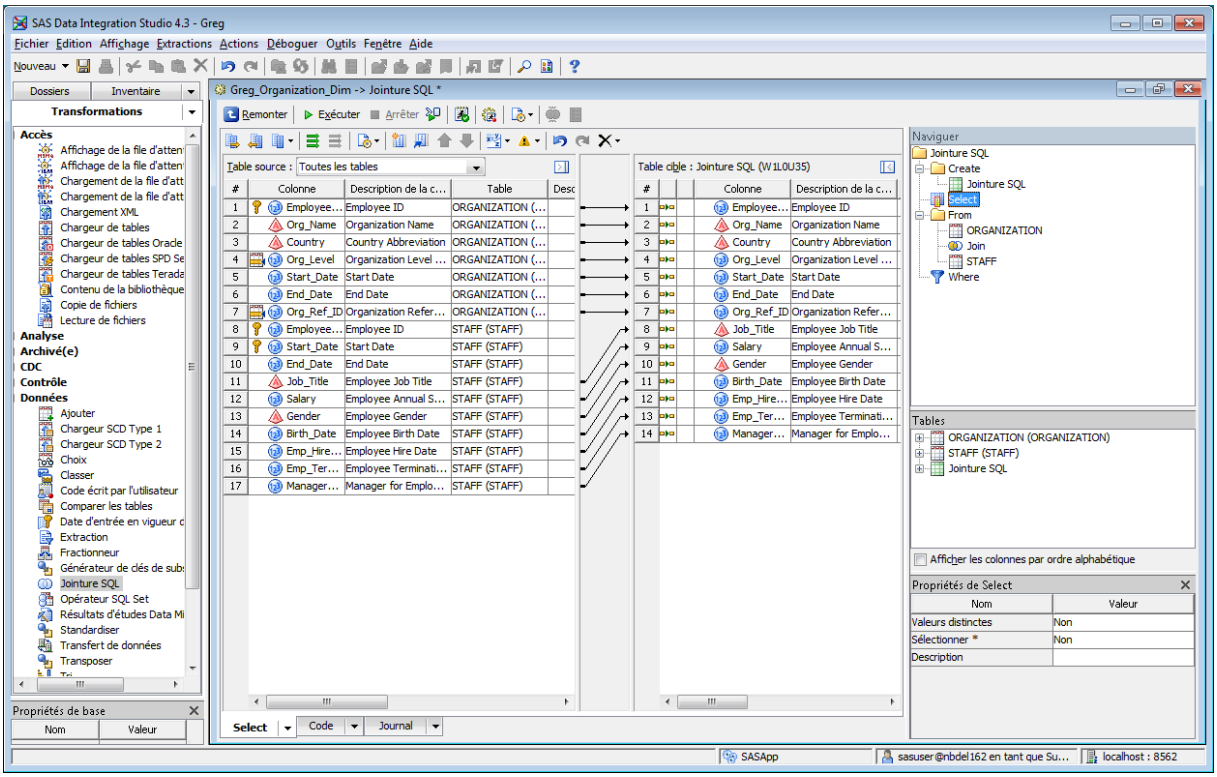
Ajouter une clause Where en double cliquant sur Where
Sélectionner le Where apparu dans l'assistant
Double-cliquer dessus

Cliquer sur Nouveau



Ajouter Org_level à gauche (sélectionner colonne de la tale) et « 1 » à droite.





Dans la source de la requête, on doit retrouver le code suivant :

```
/*****************************************************************************/
proc sql;
  create view work.W5MZZT9A as
    select
      ORGANIZATION.Employee_ID length = 8
                               format = 12.
                               label = "Employee ID",
      ORGANIZATION.Start_Date length = 4
                              format = DATE9.
                              informat = F10.
                              label = "Start Date",
      ORGANIZATION.End_Date length = 5
                            format = DATE9.
                            informat = F10.
                            label = "End Date",
      STAFF.Job_Title length = 25
                      label = "Employee Job Title",
      STAFF.Salary length = 8
                   format = DOLLAR12.
                   label = "Employee Annual Salary",
      STAFF.Gender length = 1
                   format = $GENDER.
                   label = "Employee Gender",
      STAFF.Birth_Date length = 4
                        format = DATE9.
                        label = "Employee Birth Date",
      STAFF.Emp_Hire_Date length = 4
                          format = DATE9.
                          informat = DATE9.
                          label = "Employee Hire Date",
      STAFF.Emp_Term_Date length = 4
                          format = DATE9.
```

```
                    informat = DATE9.
                    label = "Employee Termination Date",
    STAFF.Manager_ID length = 8
                  format = 12.
                  label = "Manager for Employee",
    ORGANIZATION.Org_Name length = 40
                    label = "Organization Name",
    ORGANIZATION.Country length = 2
                   format = $COUNTRY.
                   label = "Country Abbreviation",
    ORGANIZATION.Org_Level length = 3
                    format = 12.
                    label = "Organization Level Number",
    ORGANIZATION.Org_Ref_ID length = 8
                     label = "Organization Reference ID"
  from
    OrODS.ORGANIZATION left join
    OrODS.STAFF
      on
      (
        ORGANIZATION.Employee_ID = STAFF.Employee_ID
      )
  where
    ORGANIZATION.Org_Level = 1
  ;
quit;
/*****************************************************************************/
```

L'idée est d'avoir toutes les lignes de la table de l'organisation pour le premier niveau, et celles qui correspondent de la table Staff.

Comme pour la table PRODUCT_DIM, le format orgdim permet :

- A la première application, de passer de l'identifiant de l'employé à l'identifiant du groupe.
    - A la deuxième application, de passer l'identifiant du groupe à l'identifiant de la section.
        - A la troisième application, de passer de l'identifiant de la section à l'identifiant du département.
            - A la quatrième application, de passer de l'identifiant du département à l'identifiant de la compagnie.

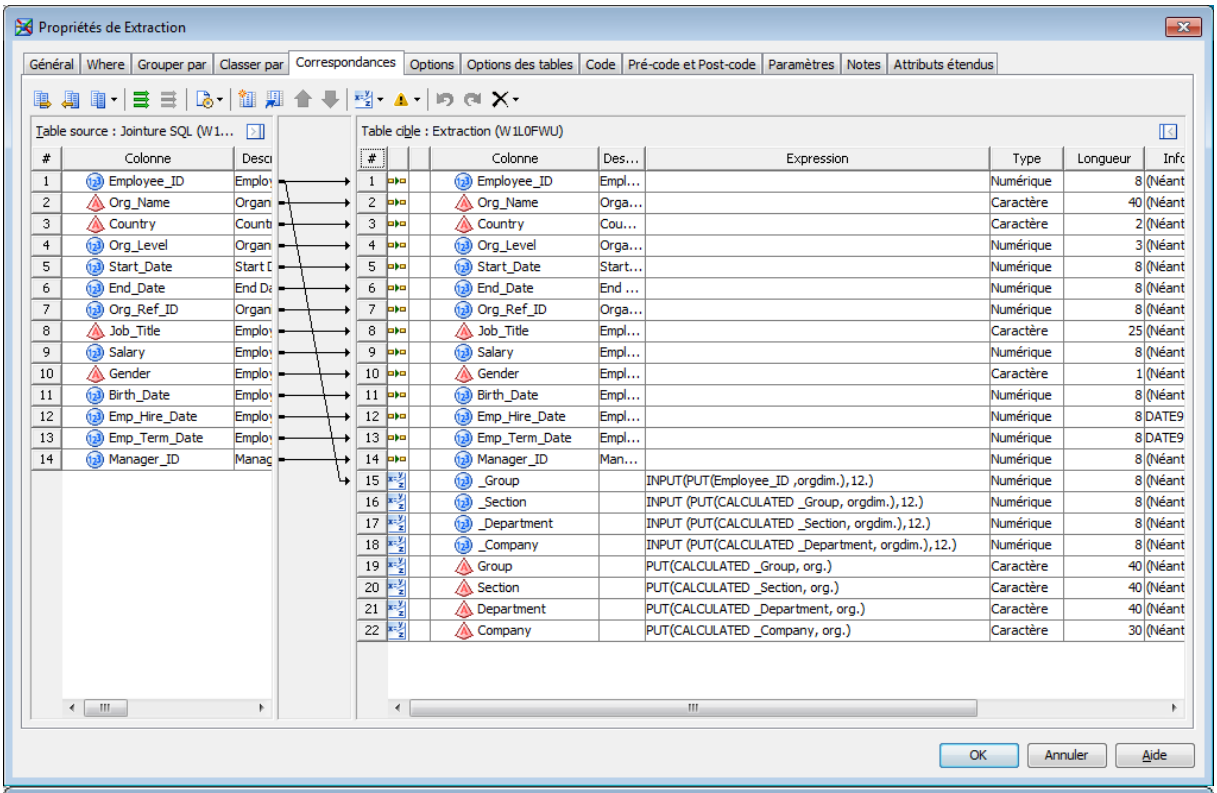
Et le format org permet :

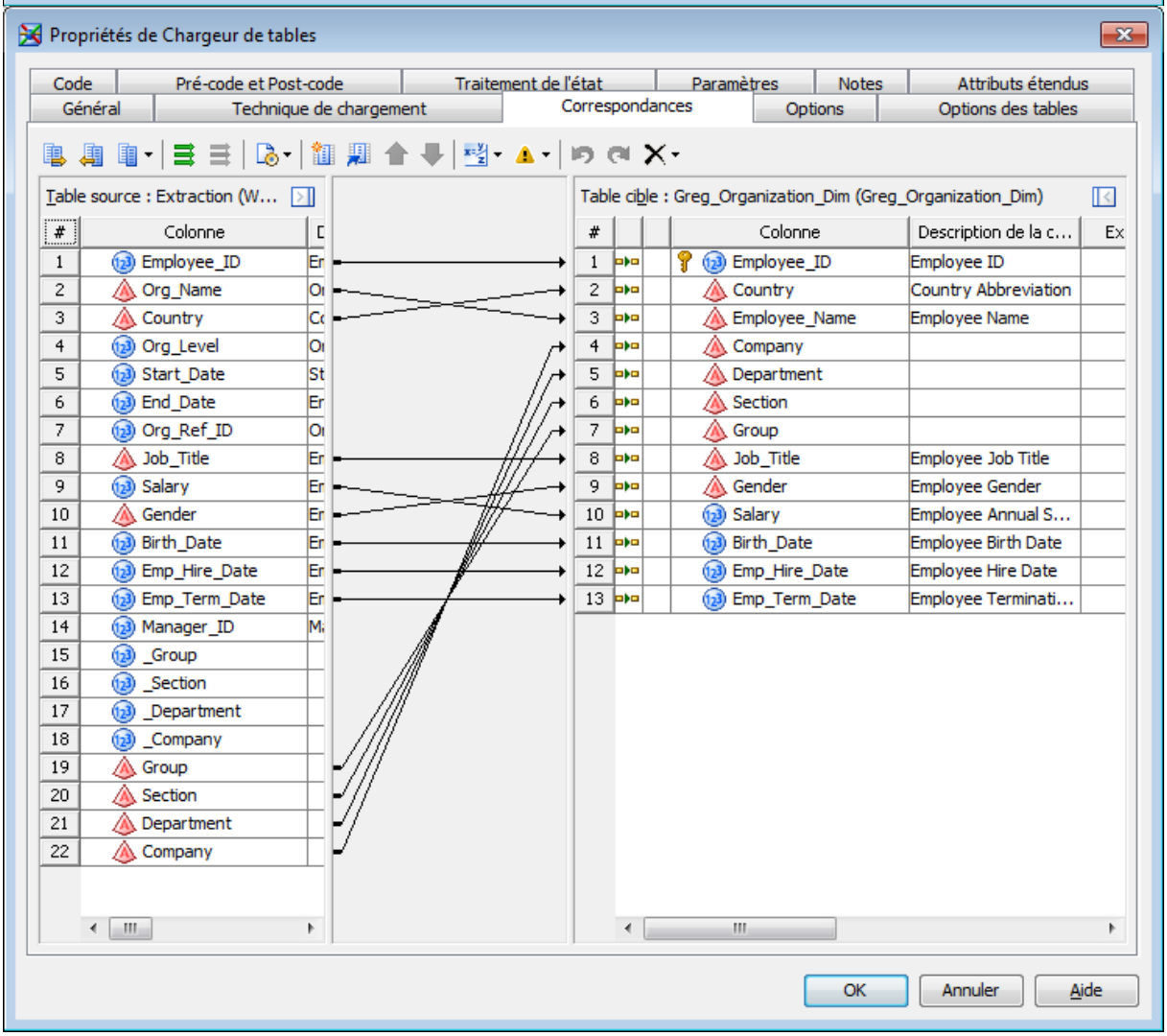
- Avec l'identifiant du groupe, d'avoir le nom du groupe.
- Avec l'identifiant de la section, d'avoir le nom de la catégorie de la section.
- Avec l'identifiant du département, d'avoir le nom du département.
- Avec l'identifiant de la compagnie, d'avoir le nom de la compagnie.

Voici donc les colonnes et les transformations à créer dans l'outil d'extraction.

Nom	Fonction	Type	longueur
_Group	INPUT(PUT(Employee_ID ,orgdim.),12.)	Numérique	8
_Section	INPUT (PUT(CALCULATED _Group, orgdim.),12.)	Numérique	8
_Department	INPUT (PUT(CALCULATED _Section, orgdim.),12.)	Numérique	8
_Company	INPUT (PUT(CALCULATED _Department, orgdim.),12.)	Numérique	8
Group	PUT(CALCULATED _Group, org.)	Caractère	40
Section	PUT(CALCULATED _Section, org.)	Caractère	40
Department	PUT(CALCULATED _Department, org.)	Caractère	40
Company	PUT(CALCULATED _Company, org.)	Caractère	30

Attention : « Department » ou « Company » en anglais !


Propriétés de Extraction
Général
Where
Grouper par
Classer par
Correspondances
Options
Options des tables
Code
Pré-code et Post-code
Paramètres
Notes
Attributs étendus
Table source : Jointure SQL (W1...
Table cible : Extraction (W1L0FWU)
Colonne
Expression
Type
Longueur
OK
Annuler
Aide


Propriétés de Chargeur de tables
Code
Pré-code et Post-code
Traitement de l'état
Paramètres
Notes
Attributs étendus
Général
Technique de chargement
Correspondances
Options
Options des tables
Table source : Extraction (W...
Table cible : Greg_Organization_Dim (Greg_Organization_Dim)
Colonne
Description de la c...
OK
Annuler
Aide

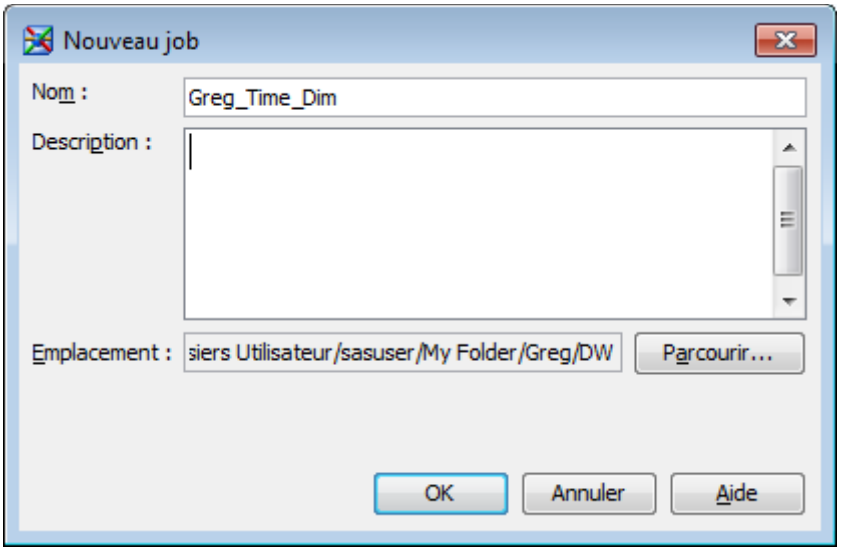
Pour finir, dans le chargeur, il faut ajouter les liens nécessaires dans l’onglet correspondance.

On doit avoir 1049 lignes.

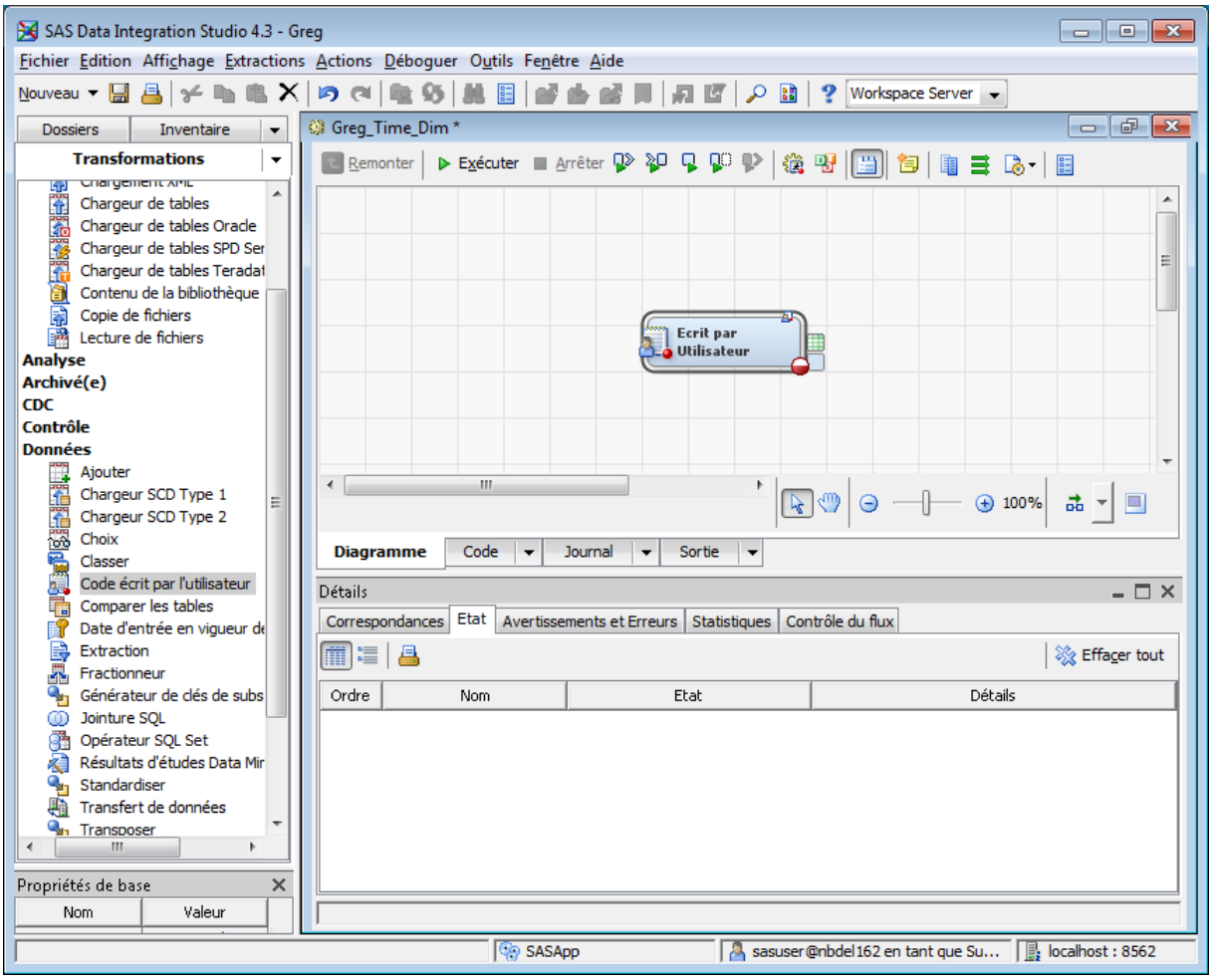
Création de la table TIME_DIM

Nom	Longueur	Type	Format
Date_ID	4	Numérique	DATE9.
Year_ID	4	Caractère	(Néant)
Quarter	6	Caractère	(Néant)
Month_Name	20	Caractère	(Néant)
Week_Name	7	Caractère	(Néant)
Weekday_Name	20	Caractère	(Néant)
Month_Num	3	Numérique	(Néant)
Week_Num	3	Numérique	(Néant)
Weekday_Num	3	Numérique	(Néant)
Weekday_EU	3	Numérique	(Néant)
Fiscal_Year	4	Caractère	(Néant)
Fiscal_Quarter	6	Caractère	(Néant)
Fiscal_Month_Num	3	Numérique	(Néant)
WeekDay_FR	8	Caractère	(Néant)
Month_FR	9	Caractère	(Néant)

Pour créer la table de dimension du temps, une solution simple est d'utiliser une étape de code écrit par l'utilisateur.

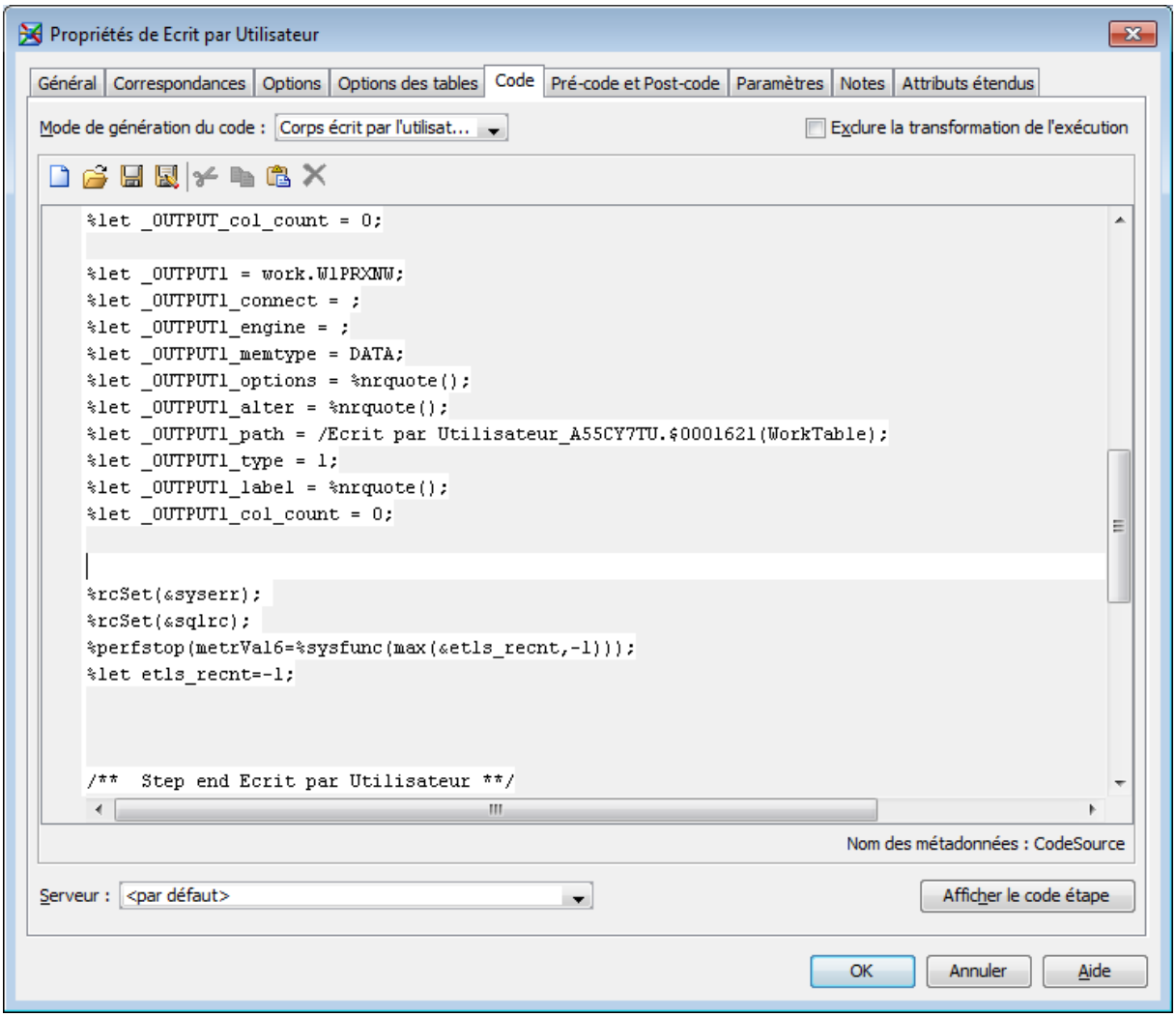


Créer directement un job,



Ajouter l'outil de code écrit par l'utilisateur.

Double cliquer dessus



Dans l'onglet Code,

Dans la ligne blanche
Recopier le code suivant

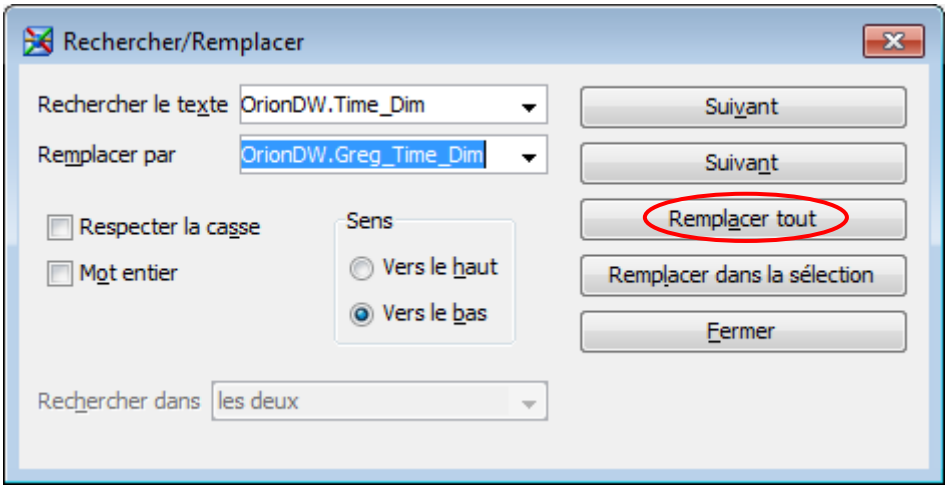
Changer le chemin dans le libname si besoin.

```
/*************************************************************************************************************/


libname OrionDW 'C:\SAS\Orion\DW';

 Proc SQL;
 CREATE TABLE OrionDW.Time_Dim (
      Date_ID            DATE LABEL='Date',
      Weekday_Num         SMALLINT LABEL='US WeekDay Number',
      Weekday_EU          SMALLINT LABEL='European Weekday Number',
      Weekday_Name         VARCHAR(20) LABEL='Weekday Name',
      Week_Num            SMALLINT LABEL='European Week Number',
      Week_Name           CHARACTER(7) LABEL='Week Name in display format (YYYY-WW)',
      Month_Num           SMALLINT LABEL='Month Number',
      Month_Name          VARCHAR(20) LABEL='Month Name',
      Quarter            CHARACTER(6) LABEL='Quarter',
      Year_ID            CHARACTER(4) LABEL='Year',
      Fiscal_Year         CHARACTER(4) LABEL='Fiscal Year',
```

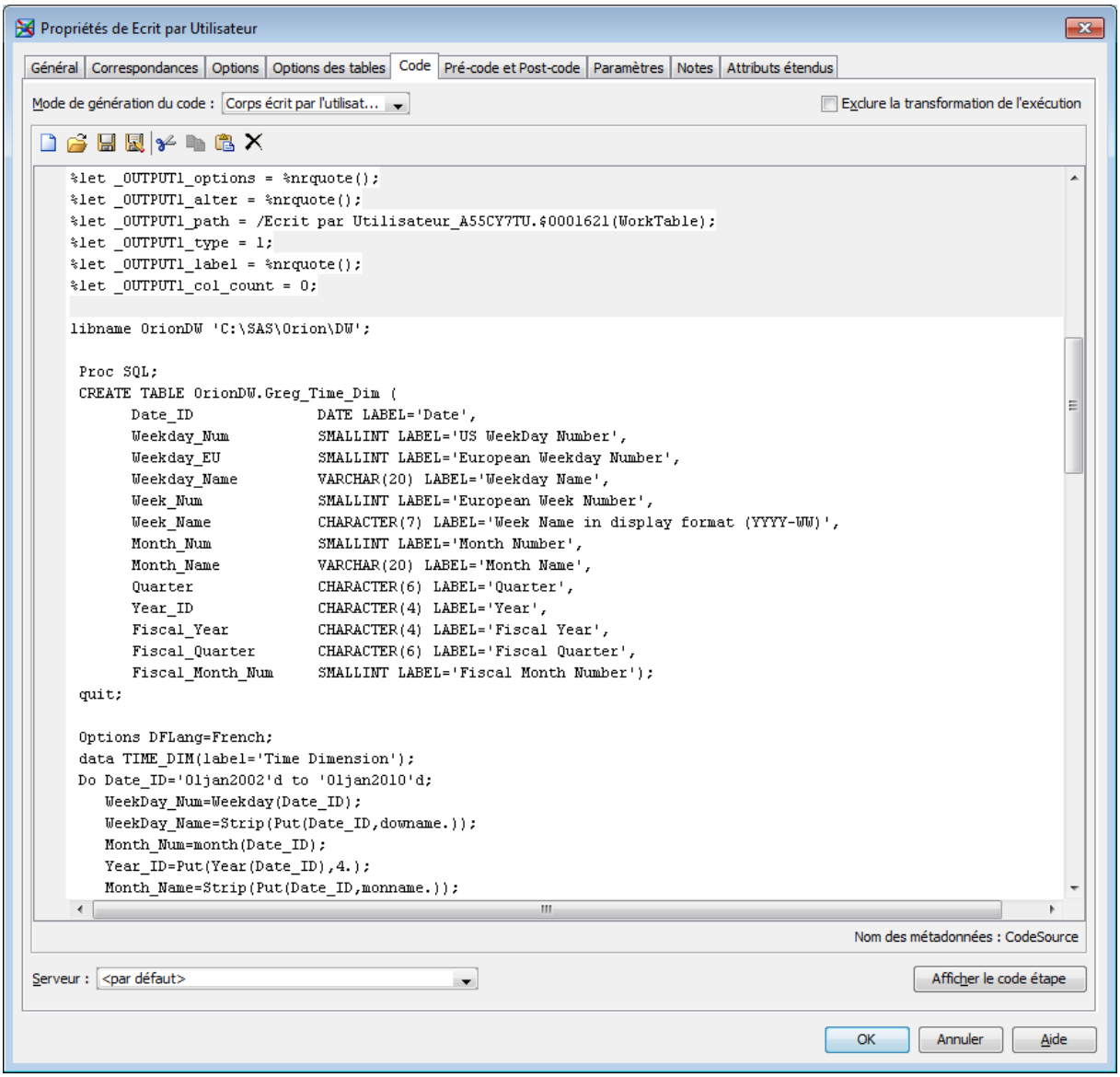
```
    Fiscal_Quarter      CHARACTER(6) LABEL='Fiscal Quarter',
    Fiscal_Month_Num    SMALLINT LABEL='Fiscal Month Number');
quit;

Options DFLang=French;
data TIME_DIM(label='Time Dimension');
Do Date_ID='01jan2002'd to '01jan2010'd;
  WeekDay_Num=Weekday(Date_ID);
  WeekDay_Name=Strip(Put(Date_ID,downame.));
  Month_Num=month(Date_ID);
  Year_ID=Put(Year(Date_ID),4.);
  Month_Name=Strip(Put(Date_ID,monname.));
  Quarter=put(Year_ID,4.)!!'Q'!!put(qtr(Date_ID),1.);
  WeekDay_FR=Strip(Put(Date_ID,EURDFDWN.));
  Month_FR=Strip(Put(Date_ID,EURDFMN.));
        if Month_Num <=11 then Fiscal_Year=Put(Year(Date_ID),4.);
        else Fiscal_Year=Put(Year(Date_ID)+1,4.);
  if Month_Num <=11 then Fiscal_Month_Num =month(Date_ID)+1;
        else Fiscal_Month_Num=1;
        Fiscal_Quarter=Fiscal_Year!!'Q'!!put(ceil(Fiscal_Month_Num/3),1.);
  Output;

end;
run;

Proc format;
Picture We Low-High= '9999-99';
run;

Data OrionDW.Time_Dim(Compress=Yes Label='Time Dimension' Sortedby=Date_ID);
Length Date_ID 4 Year_ID $4 Quarter $6 Month_Name $20 Week_Name $7
Weekday_Name $20 Month_Num Week_Num Weekday_Num Weekday_EU 3
Fiscal_Year $4 Fiscal_Quarter $6 Fiscal_Month_Num 3;
Format Date_ID Date9.;
set OrionDW.TIME_DIM(obs=0) TIME_DIM;
retain Week 1 weekyear 2002;
Weekday_EU=weekday(Date_ID-1);
if Weekday_EU=1 then do;
  Week+1;
  if mdy(12,29,Input(Year_ID,4.)-(month_Num=1))
  <=Date_ID<=mdy(1,4,Input(Year_ID,4.)+(Month_Num=12)) then
  do;
    Week=1;
    WeekYear+1;
  end;
end;
Week_Num=Week;
Week_Name=Put(WeekYear*100+Week_Num,we.);
Drop Week WeekYear ;
run;

Proc SQL;
CREATE INDEX Year_ID        ON OrionDW.TIME_DIM(Year_ID);
CREATE INDEX Quarter        ON OrionDW.TIME_DIM(Quarter);
CREATE INDEX Month_Num      ON OrionDW.TIME_DIM(Month_Num);
CREATE INDEX Fiscal_Year    ON OrionDW.TIME_DIM(Fiscal_Year);
CREATE INDEX Fiscal_Quarter ON OrionDW.TIME_DIM(Fiscal_Quarter);
CREATE UNIQUE INDEX Date_ID ON OrionDW.TIME_DIM(Date_ID);
ALTER TABLE OrionDW.TIME_DIM
   ADD PRIMARY KEY (Date_ID) ;
```

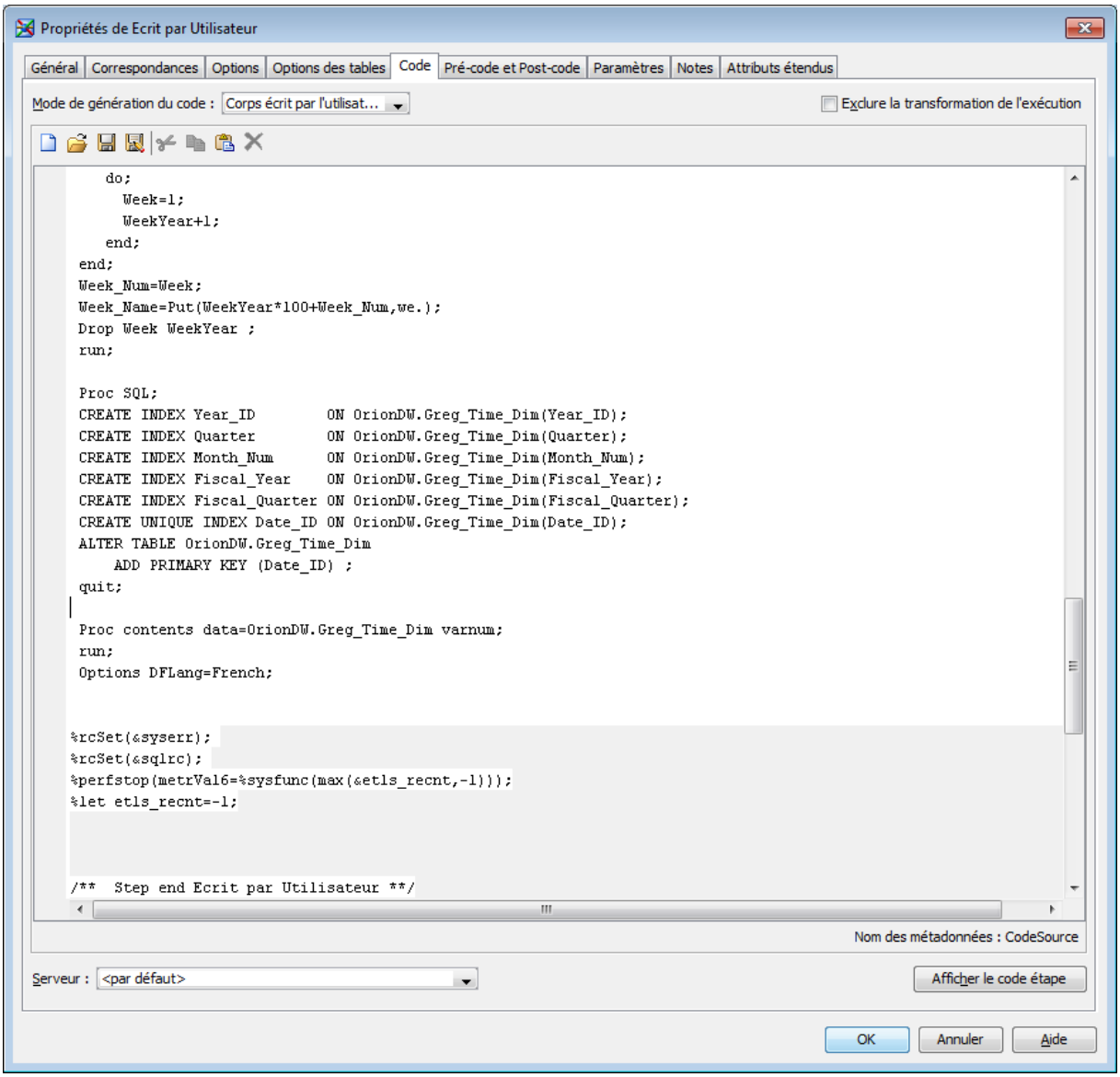
```
quit;

Proc contents data=OrionDW.TIME_DIM varnum;
run;
Options DFLang=French;
/*****************************************************************************************************/
```

Remplacer tous les (Ctrl – F) « OrionDW.Time_Dim » par « OrionDW. votre_nom_Time_Dim »

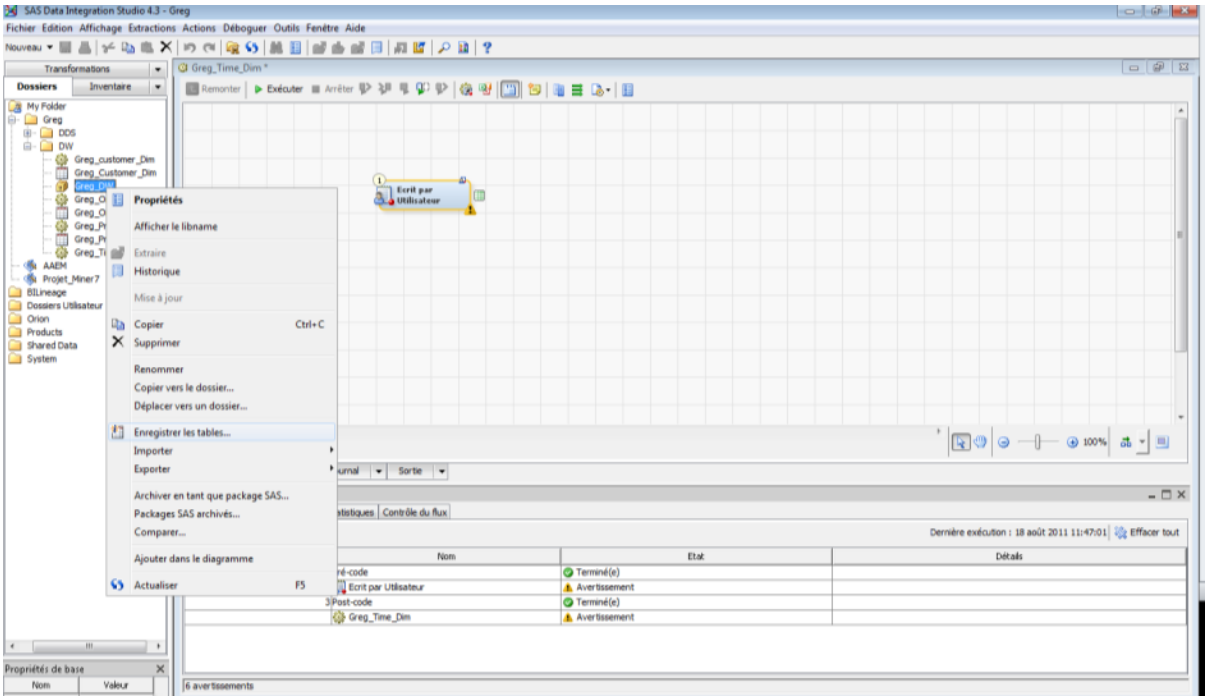


Attention, si vous faites un copier-coller à ne pas copier les numéros de page.

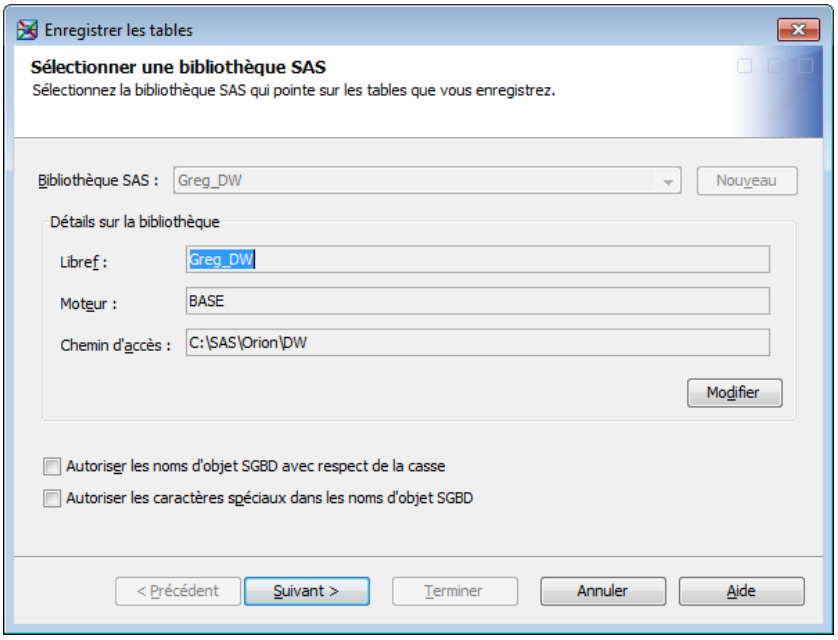

Propriétés de Ecrit par Utilisateur
Général
Correspondances
Options
Options des tables
Code
Pré-code et Post-code
Paramètres
Notes
Attributs étendus
Mode de génération du code :
Corps écrit par l'utilisat...
Exclure la transformation de l'exécution
%let _OUTPUT1_options = %nrquote();
%let _OUTPUT1_alter = %nrquote();
%let _OUTPUT1_path = /Ecrit par Utilisateur_A55CY7TU.$0001621(WorkTable);
%let _OUTPUT1_type = 1;
%let _OUTPUT1_label = %nrquote();
%let _OUTPUT1_col_count = 0;
libname OrionDW 'C:\SAS\Orion\DW';
Proc SQL;
CREATE TABLE OrionDW.Greg_Time_Dim (
Date_ID DATE LABEL='Date',
Weekday_Num SMALLINT LABEL='US WeekDay Number',
Weekday_EU SMALLINT LABEL='European Weekday Number',
Weekday_Name VARCHAR(20) LABEL='Weekday Name',
Week_Num SMALLINT LABEL='European Week Number',
Week_Name CHARACTER(7) LABEL='Week Name in display format (YYYY-WW)',
Month_Num SMALLINT LABEL='Month Number',
Month_Name VARCHAR(20) LABEL='Month Name',
Quarter CHARACTER(6) LABEL='Quarter',
Year_ID CHARACTER(4) LABEL='Year',
Fiscal_Year CHARACTER(4) LABEL='Fiscal Year',
Fiscal_Quarter CHARACTER(6) LABEL='Fiscal Quarter',
Fiscal_Month_Num SMALLINT LABEL='Fiscal Month Number');
quit;
Options DFLang=French;
data TIME_DIM(label='Time Dimension');
Do Date_ID='01jan2002'd to '01jan2010'd;
WeekDay_Num=Weekday(Date_ID);
WeekDay_Name=Strip(Put(Date_ID,downame.));
Month_Num=month(Date_ID);
Year_ID=Put(Year(Date_ID),4.);
Month_Name=Strip(Put(Date_ID,monname.));
Nom des métadonnées : CodeSource
Serveur :
<par défaut>
Afficher le code étape
OK
Annuler
Aide



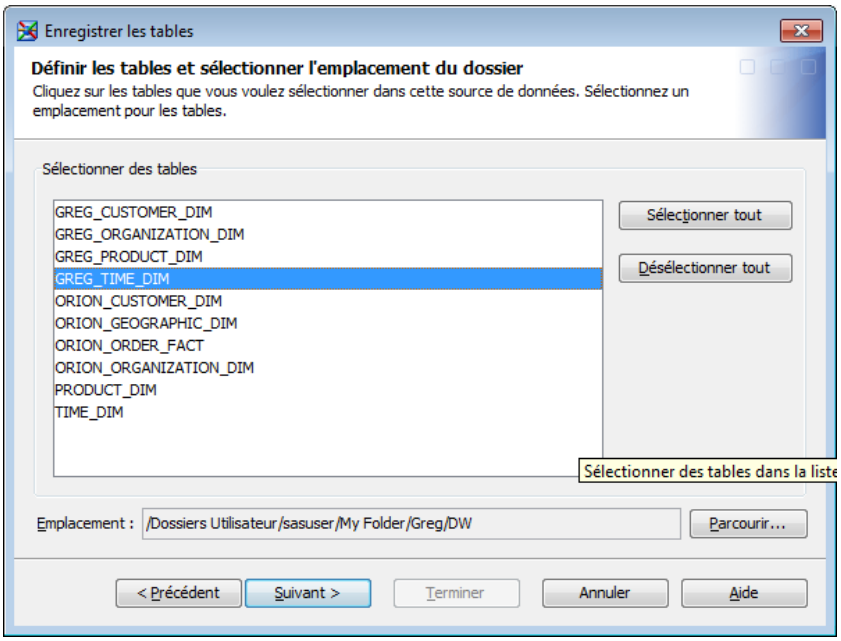
Fermer la fenêtre (**OK**)
Exécuter le processus



Pour importer les métadonnées de la table créer, Clic-droit sur la bibliothèque **Votre_nom_DW** se trouvant dans le répertoire **Mon Dossier → votre nom → DW → Enregistrer les tables**



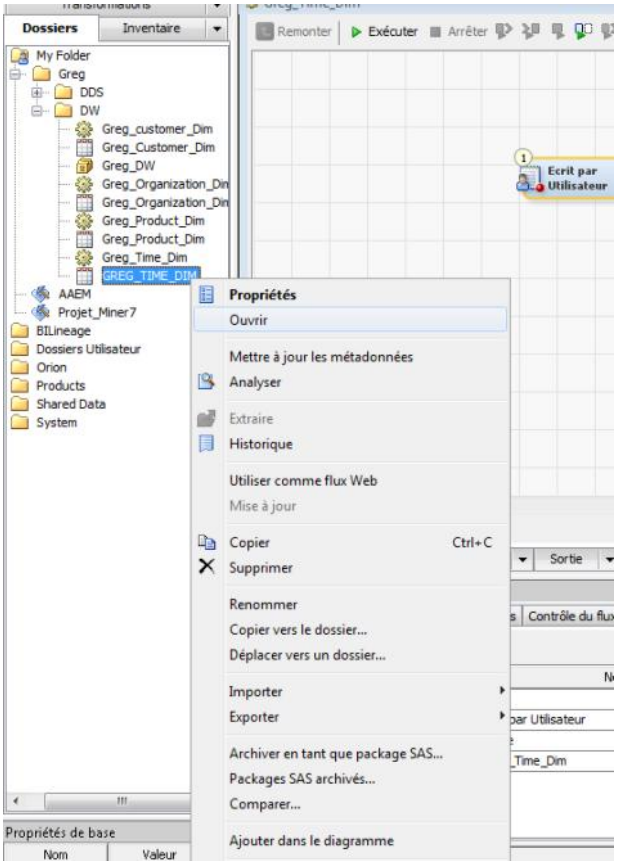
Suivant,

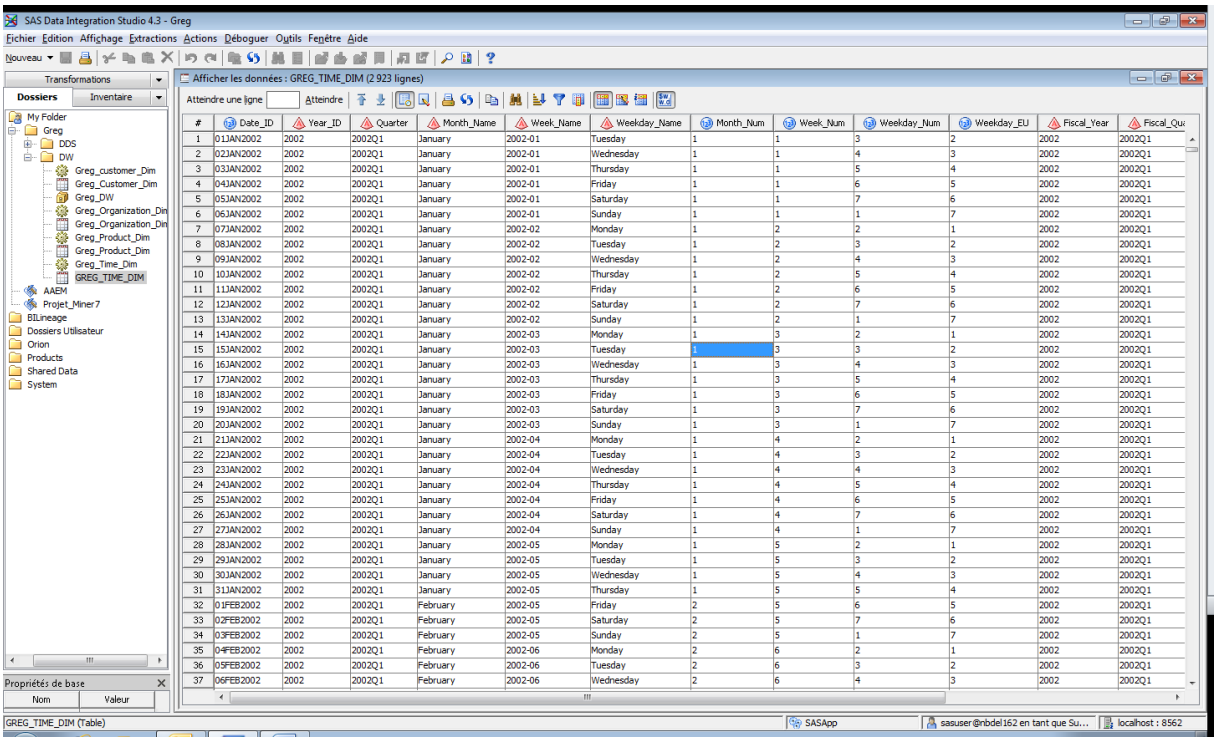


Sélectionner uniquement votre table.

Suivant,
Terminer

Si vous ouvrez la table, Il y a toutes les dates du premier janvier 2002 au premier janvier 2010.





#	Date_ID	Year_ID	Quarter	Month_Name	Week_Name	Weekday_Name	Month_Num	Week_Num	Weekday_Num	Weekday_EU	Fiscal_Year	Fiscal_Qu
1	01JAN2002	2002	2002Q1	January	2002-01	Tuesday	1	1	3	2	2002	2002Q1
2	02JAN2002	2002	2002Q1	January	2002-01	Wednesday	1	1	4	3	2002	2002Q1
3	03JAN2002	2002	2002Q1	January	2002-01	Thursday	1	1	5	4	2002	2002Q1
4	04JAN2002	2002	2002Q1	January	2002-01	Friday	1	1	6	5	2002	2002Q1
5	05JAN2002	2002	2002Q1	January	2002-01	Saturday	1	1	7	6	2002	2002Q1
6	06JAN2002	2002	2002Q1	January	2002-01	Sunday	1	1	1	7	2002	2002Q1
7	07JAN2002	2002	2002Q1	January	2002-02	Monday	1	2	2	1	2002	2002Q1
8	08JAN2002	2002	2002Q1	January	2002-02	Tuesday	1	2	3	2	2002	2002Q1
9	09JAN2002	2002	2002Q1	January	2002-02	Wednesday	1	2	4	3	2002	2002Q1
10	10JAN2002	2002	2002Q1	January	2002-02	Thursday	1	2	5	4	2002	2002Q1
11	11JAN2002	2002	2002Q1	January	2002-02	Friday	1	2	6	5	2002	2002Q1
12	12JAN2002	2002	2002Q1	January	2002-02	Saturday	1	2	7	6	2002	2002Q1
13	13JAN2002	2002	2002Q1	January	2002-02	Sunday	1	2	1	7	2002	2002Q1
14	14JAN2002	2002	2002Q1	January	2002-03	Monday	1	3	2	1	2002	2002Q1
15	15JAN2002	2002	2002Q1	January	2002-03	Tuesday	1	3	3	2	2002	2002Q1
16	16JAN2002	2002	2002Q1	January	2002-03	Wednesday	1	3	4	3	2002	2002Q1
17	17JAN2002	2002	2002Q1	January	2002-03	Thursday	1	3	5	4	2002	2002Q1
18	18JAN2002	2002	2002Q1	January	2002-03	Friday	1	3	6	5	2002	2002Q1
19	19JAN2002	2002	2002Q1	January	2002-03	Saturday	1	3	7	6	2002	2002Q1
20	20JAN2002	2002	2002Q1	January	2002-03	Sunday	1	3	1	7	2002	2002Q1
21	21JAN2002	2002	2002Q1	January	2002-04	Monday	1	4	2	1	2002	2002Q1
22	22JAN2002	2002	2002Q1	January	2002-04	Tuesday	1	4	3	2	2002	2002Q1
23	23JAN2002	2002	2002Q1	January	2002-04	Wednesday	1	4	4	3	2002	2002Q1
24	24JAN2002	2002	2002Q1	January	2002-04	Thursday	1	4	5	4	2002	2002Q1
25	25JAN2002	2002	2002Q1	January	2002-04	Friday	1	4	6	5	2002	2002Q1
26	26JAN2002	2002	2002Q1	January	2002-04	Saturday	1	4	7	6	2002	2002Q1
27	27JAN2002	2002	2002Q1	January	2002-04	Sunday	1	4	1	7	2002	2002Q1
28	28JAN2002	2002	2002Q1	January	2002-05	Monday	1	5	2	1	2002	2002Q1
29	29JAN2002	2002	2002Q1	January	2002-05	Tuesday	1	5	3	2	2002	2002Q1
30	30JAN2002	2002	2002Q1	January	2002-05	Wednesday	1	5	4	3	2002	2002Q1
31	31JAN2002	2002	2002Q1	January	2002-05	Thursday	1	5	5	4	2002	2002Q1
32	01FEB2002	2002	2002Q1	February	2002-05	Friday	2	5	6	5	2002	2002Q1
33	02FEB2002	2002	2002Q1	February	2002-05	Saturday	2	5	7	6	2002	2002Q1
34	03FEB2002	2002	2002Q1	February	2002-05	Sunday	2	5	1	7	2002	2002Q1
35	04FEB2002	2002	2002Q1	February	2002-06	Monday	2	6	2	1	2002	2002Q1
36	05FEB2002	2002	2002Q1	February	2002-06	Tuesday	2	6	3	2	2002	2002Q1
37	06FEB2002	2002	2002Q1	February	2002-06	Wednesday	2	6	4	3	2002	2002Q1

Création de la table GEOGRAPHIC_DIM

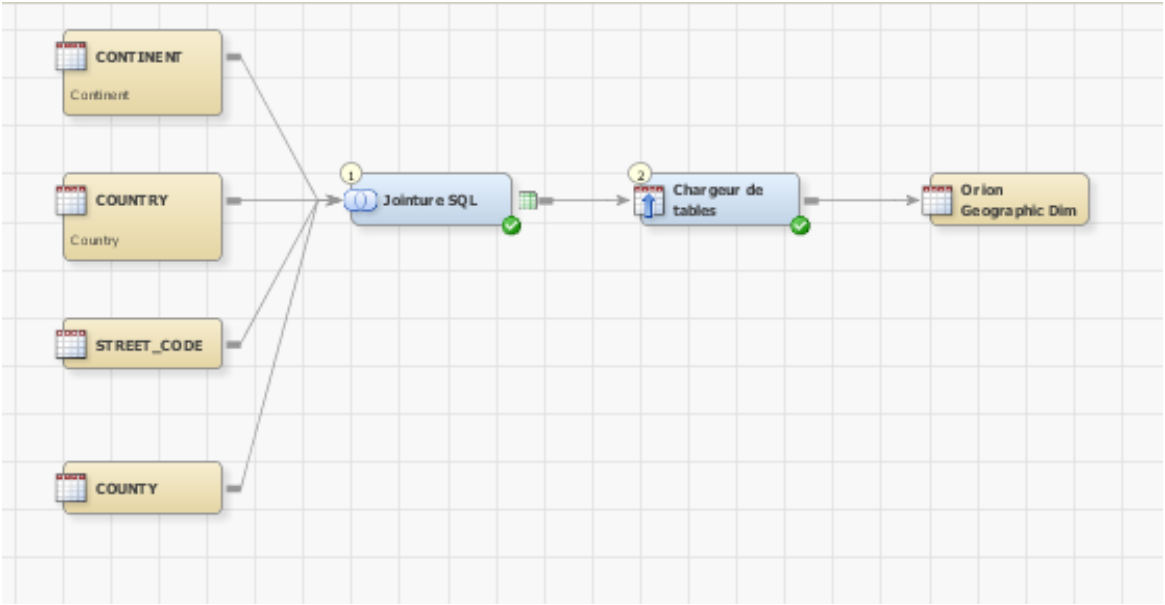
Créer la table de dimension GEOGRAPHIC_DIM selon la définition suivante :

Nom	Longueur	Type	Format
Street_ID	8	Numérique	12.
Continent_Name	30	Caractère	(Néant)
Country_Name	30	Caractère	(Néant)
Region_Name	30	Caractère	(Néant)
Province_Name	30	Caractère	(Néant)
County_Name	60	Caractère	(Néant)
City_Name	30	Caractère	(Néant)
Postal_Code	10	Caractère	(Néant)
Street_Name	40	Caractère	(Néant)

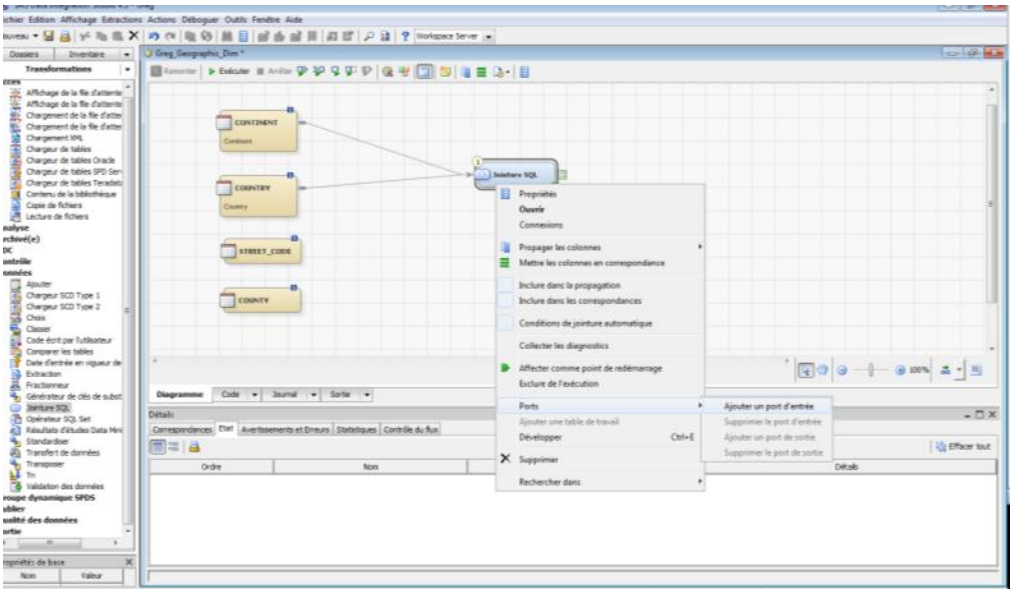
Clé Primaire : Street_ID

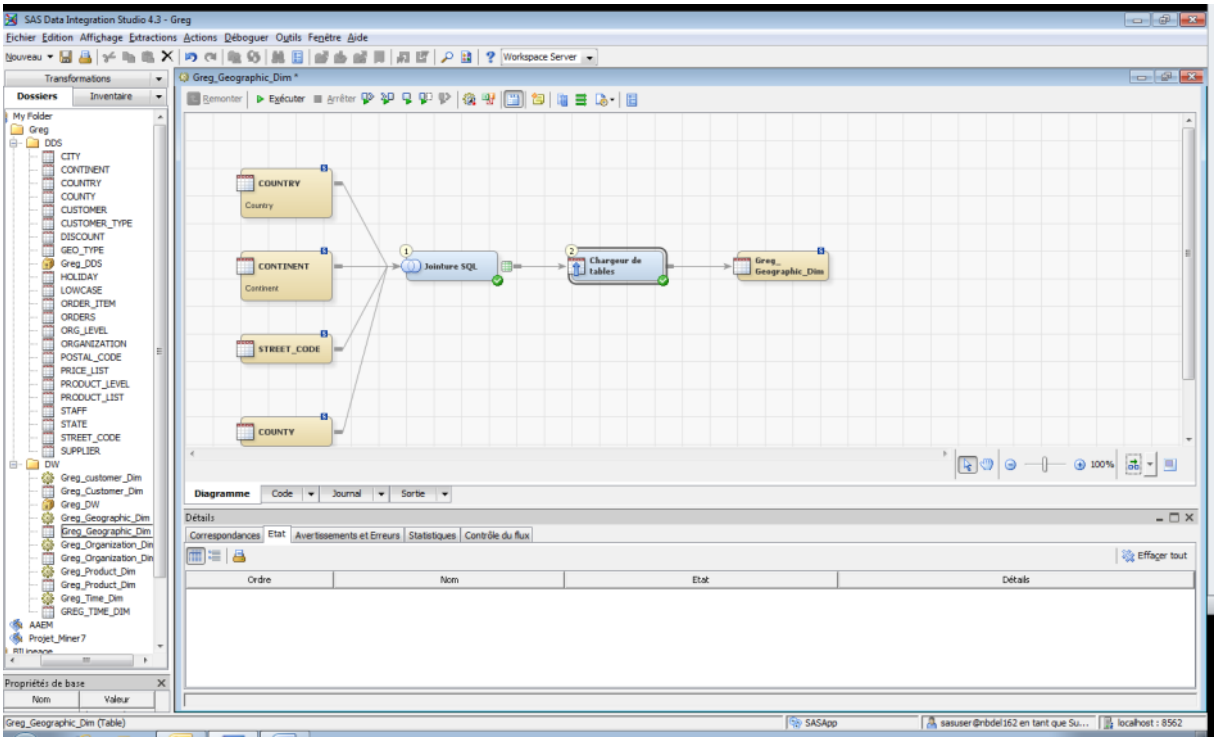
Il faut toutes les lignes de Street code !
Le problème ici est que si vous définissez les liens entre les tables comme dans le MCD, il vous manquera des lignes. Vous devez obtenir 89 807 lignes. Il faut donc toutes les lignes de Street_Code et faire une jointure directement entre Street Code et Country sur la colonne Country.

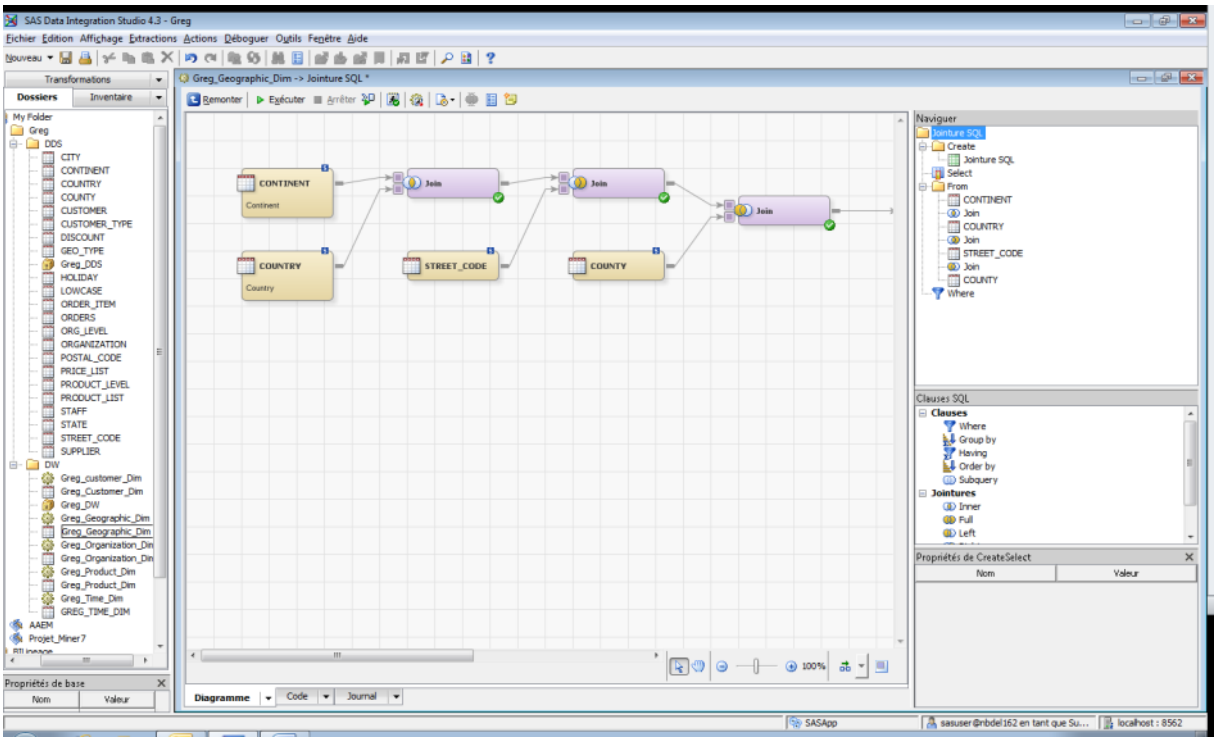
Si vous ne définissez pas correctement les jointures, parmi les erreurs possibles, vous pouvez faire une table dont la taille explose. Si vous n'êtes pas sûr, faites vérifier votre travail.

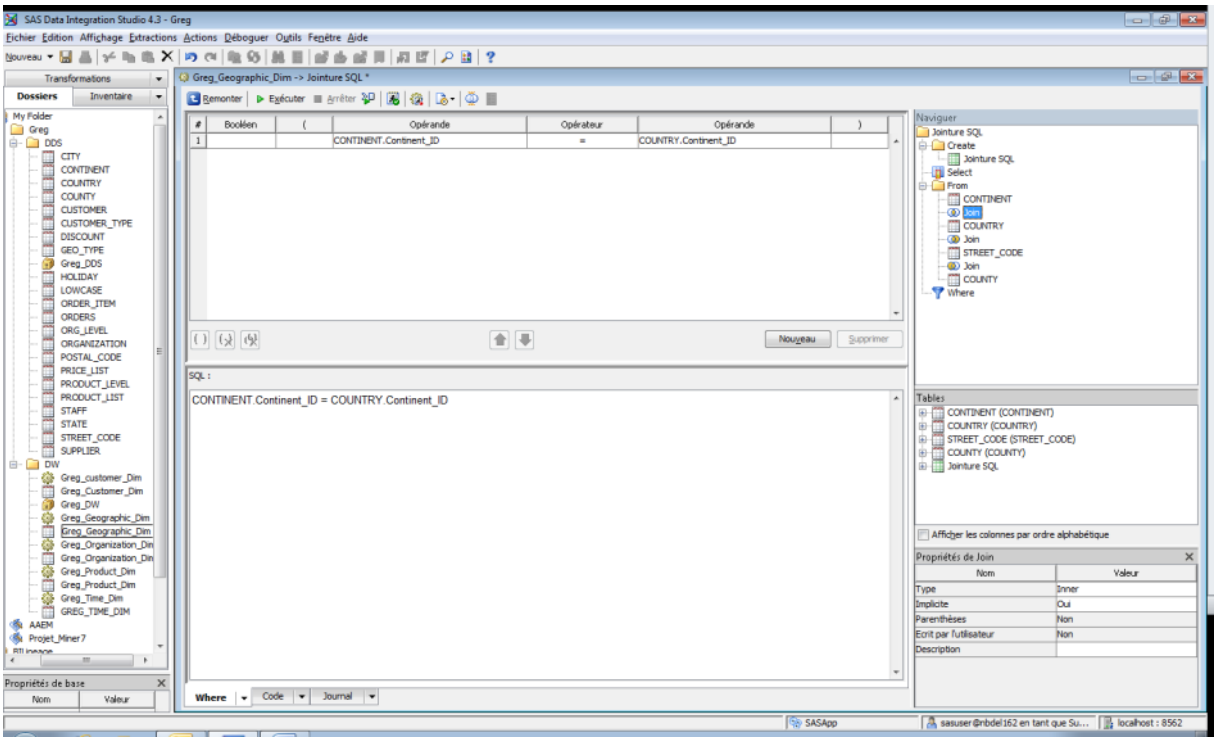


Pour simplifier, il est conseillé d'ajouter **dans l'ordre** les tables Continent, Country, Street_Code et county à la jointure SQL. Il faut faire un clic-droit sur la jointure pour ajouter des ports afin de faire une jointure avec les 4 tables.

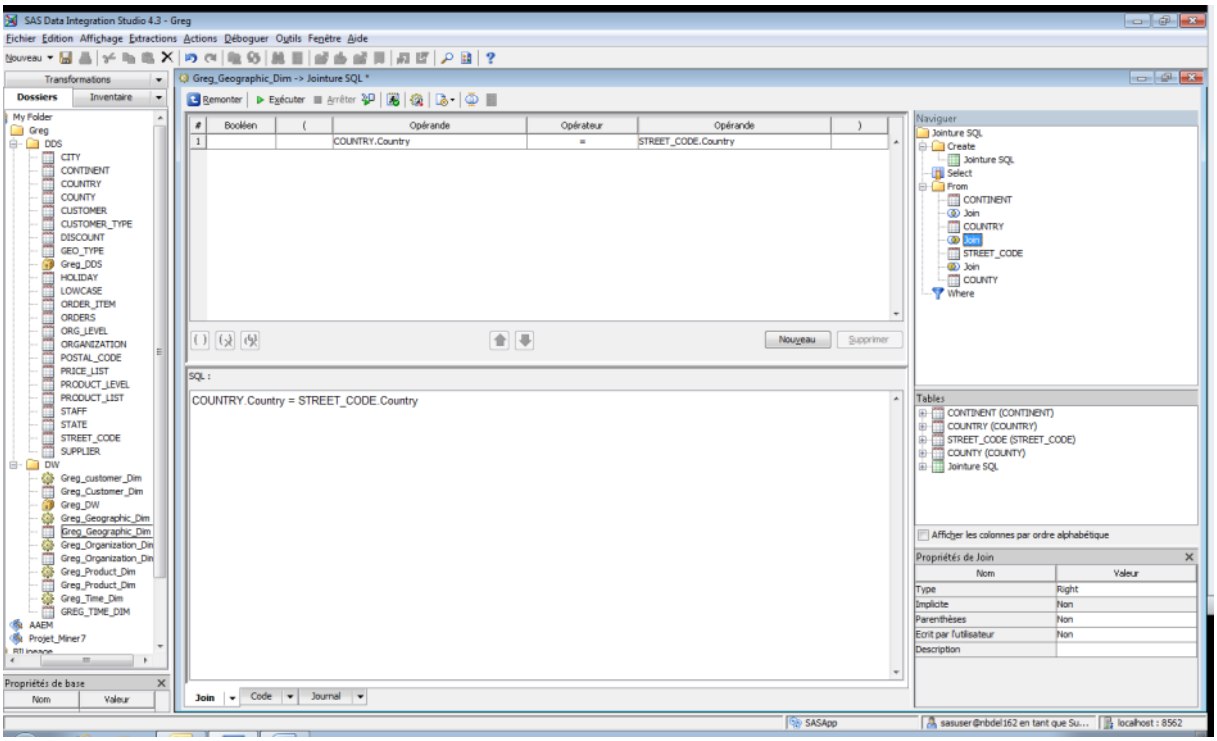



SAS Data Integration Studio 4.3 - Greg
Fichier Edition Affichage Extractions Actions Déboguer Outils Fenêtre Aide
Workspace Server
Greg_Geographic_Dim *
Exécuter
COUNTRY
CONTINENT
STREET_CODE
COUNTY
Jointure SQL
Chargeur de tables
Greg_ Geographic_Dim
Diagramme
Code
Journal
Sortie
Détails
Correspondances
Etat
Avertissements et Erreurs
Statistiques
Contrôle du flux
Effacer tout
Ordre
Nom
Etat
Détails
Dossiers
Inventaire
Transformations
Propriétés de base
Valeur
Greg_Geographic_Dim (Table)
SASApp
localhost : 8562

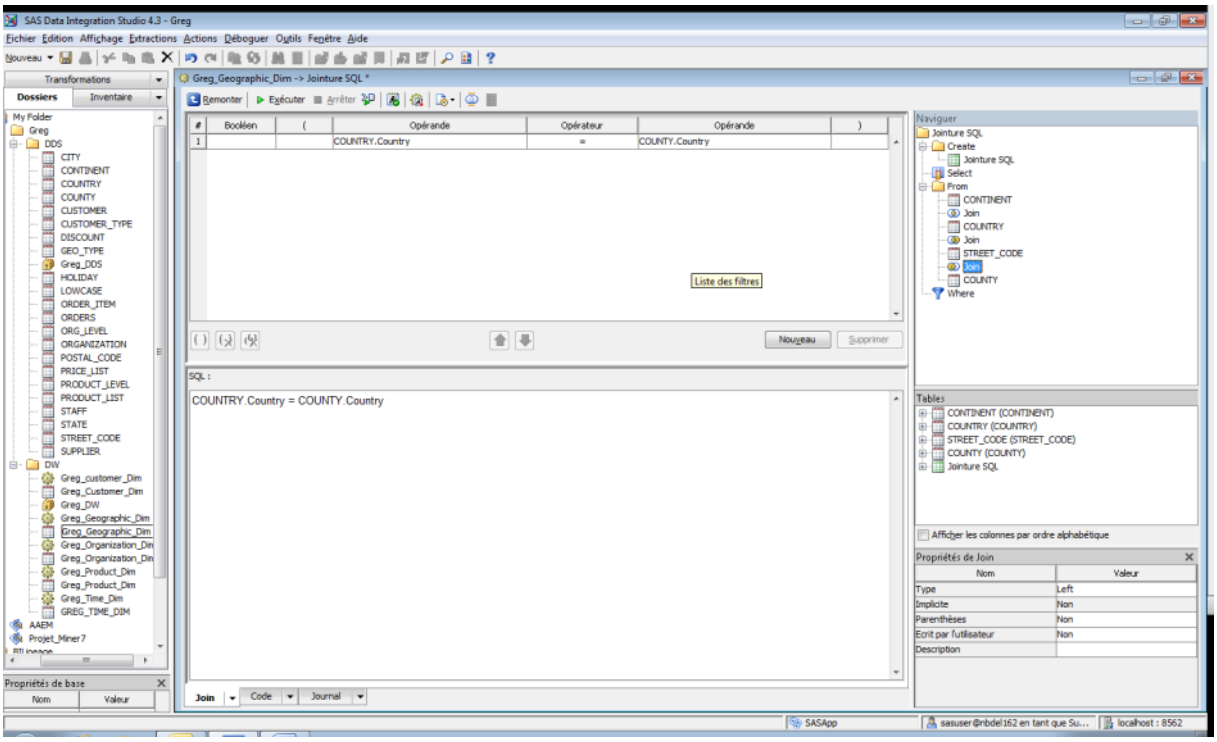




La première jointure entre les tables Continent et Country doit être une jointure Inner-Join sur Continent.continent_ID = Country.continent_ID

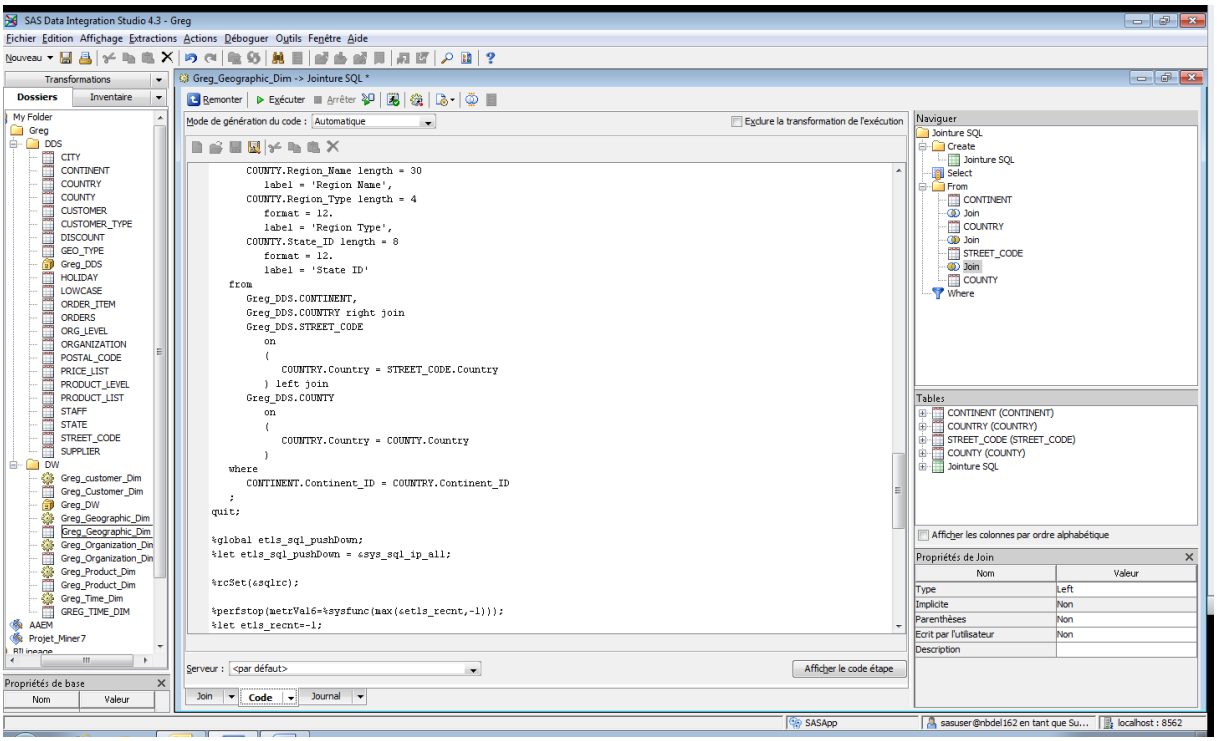


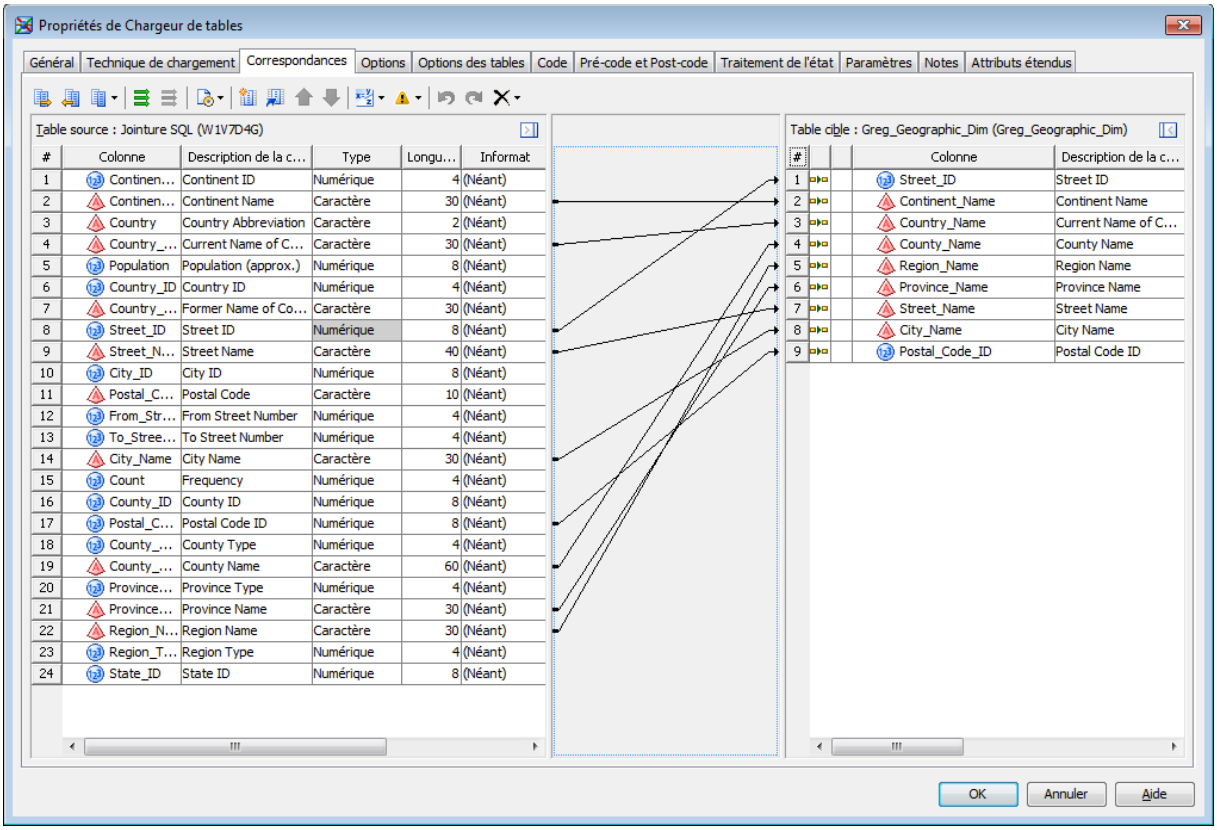
La seconde jointure doit être une jointure droite entre Country et street_code sur Country.Country = street_code.Country.



La troisième jointure doit être une jointure gauche entre street_code et county sur **street_code.county_ID = county.county_ID**

Le code SQL doit être :


SAS Data Integration Studio 4.3 - Greg
Fichier Edition Affichage Extractions Actions Déboguer Outils Fenêtre Aide
Greg_Geographic_Dim -> Jointure SQL *
Remonter Exécuter Arrêter
Mode de génération du code : Automatique
Exclure la transformation de l'exécution
COUNTY.Region_Name length = 30
label = 'Region Name',
COUNTY.Region_Type length = 4
format = 12.
label = 'Region Type',
COUNTY.State_ID length = 8
format = 12.
label = 'State ID'
from
Greg_DDS.CONTINENT,
Greg_DDS.COUNTRY right join
Greg_DDS.STREET_CODE
on
(
COUNTRY.Country = STREET_CODE.Country
) left join
Greg_DDS.COUNTY
on
(
COUNTRY.Country = COUNTY.Country
)
where
CONTINENT.Continent_ID = COUNTRY.Continent_ID
;
quit;
%global etls_sql_pushDown;
%let etls_sql_pushDown = &sys_sql_ip_all;
%rcSet(&sqlrc);
%perfstop(metrVal6=%sysfunc(max(&etls_recnt,-1)));
%let etls_recnt=-1;
Serveur : <par défaut>
Afficher le code étape
Join Code Journal
Naviguer
Jointure SQL
Create
Jointure SQL
Select
From
CONTINENT
Join
COUNTRY
Join
STREET_CODE
Join
COUNTY
Where
Tables
CONTINENT (CONTINENT)
COUNTRY (COUNTRY)
STREET_CODE (STREET_CODE)
COUNTY (COUNTY)
Jointure SQL
Afficher les colonnes par ordre alphabétique
Propriétés de Join
Nom
Valeur
Type
Left
Implicite
Non
Parenthèses
Non
Ecrit par l'utilisateur
Non
Description
Propriétés de base
Nom
Valeur
SASApp
sasuser@nbdel162 en tant que Su...
localhost : 8562


Propriétés de Chargeur de tables
Général
Technique de chargement
Correspondances
Options
Options des tables
Code
Pré-code et Post-code
Traitement de l'état
Paramètres
Notes
Attributs étendus
Table source : Jointure SQL (W1V7D4G)
Colonne Description de la c... Type Longu... Informat
1 Continen... Continent ID Numérique 4 (Néant)
2 Continen... Continent Name Caractère 30 (Néant)
3 Country Country Abbreviation Caractère 2 (Néant)
4 Country_... Current Name of C... Caractère 30 (Néant)
5 Population Population (approx.) Numérique 8 (Néant)
6 Country_ID Country ID Numérique 4 (Néant)
7 Country_... Former Name of Co... Caractère 30 (Néant)
8 Street_ID Street ID Numérique 8 (Néant)
9 Street_N... Street Name Caractère 40 (Néant)
10 City_ID City ID Numérique 8 (Néant)
11 Postal_C... Postal Code Caractère 10 (Néant)
12 From_Str... From Street Number Numérique 4 (Néant)
13 To_Stree... To Street Number Numérique 4 (Néant)
14 City_Name City Name Caractère 30 (Néant)
15 Count Frequency Numérique 4 (Néant)
16 County_ID County ID Numérique 8 (Néant)
17 Postal_C... Postal Code ID Numérique 8 (Néant)
18 County_... County Type Numérique 4 (Néant)
19 County_... County Name Caractère 60 (Néant)
20 Province... Province Type Numérique 4 (Néant)
21 Province... Province Name Caractère 30 (Néant)
22 Region_N... Region Name Caractère 30 (Néant)
23 Region_T... Region Type Numérique 4 (Néant)
24 State_ID State ID Numérique 8 (Néant)
Table cible : Greg_Geographic_Dim (Greg_Geographic_Dim)
Colonne Description de la c...
1 Street_ID Street ID
2 Continent_Name Continent Name
3 Country_Name Current Name of C...
4 County_Name County Name
5 Region_Name Region Name
6 Province_Name Province Name
7 Street_Name Street Name
8 City_Name City Name
9 Postal_Code_ID Postal Code ID
OK
Annuler
Aide

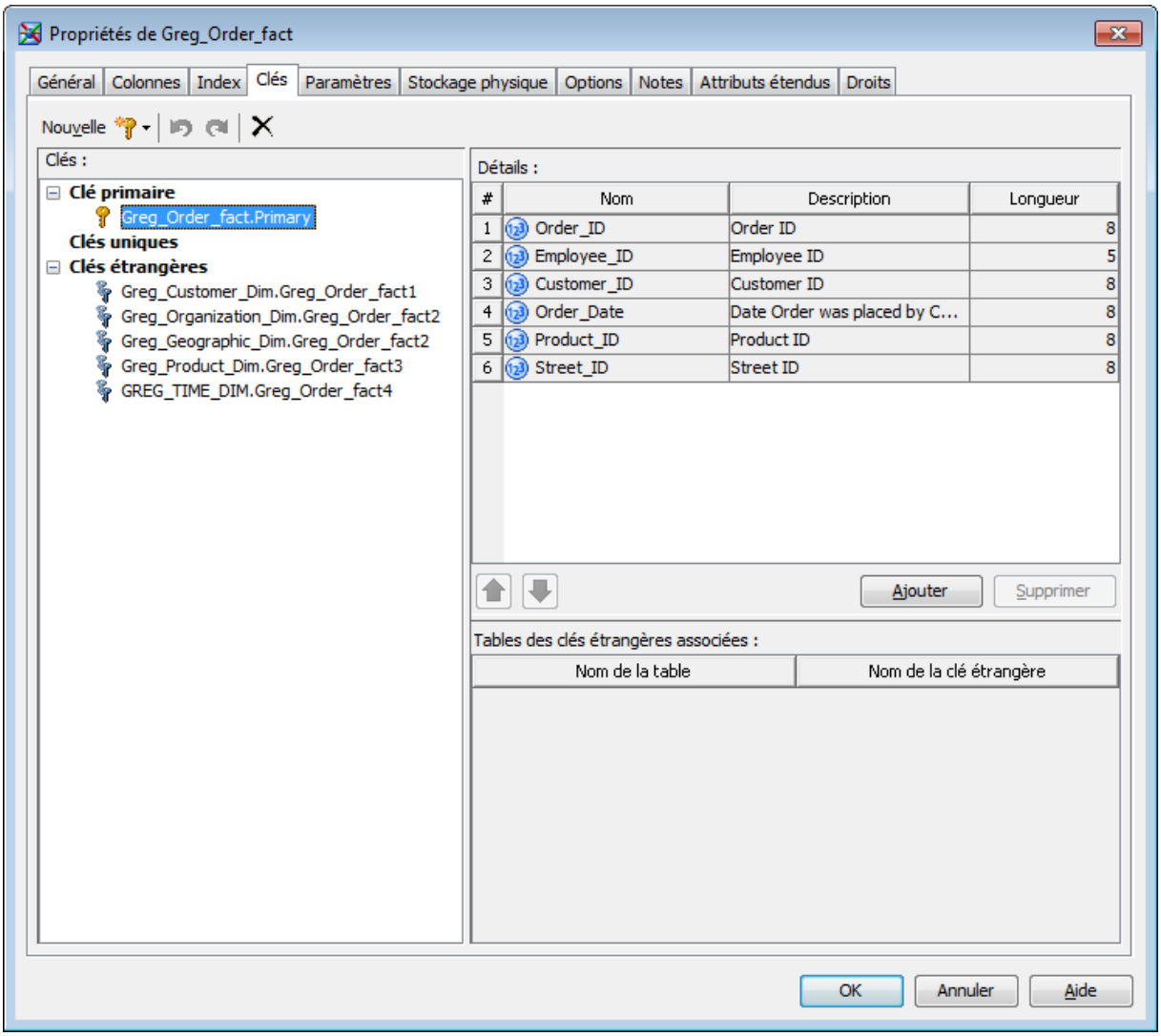
Création de la table ORDER_FACT

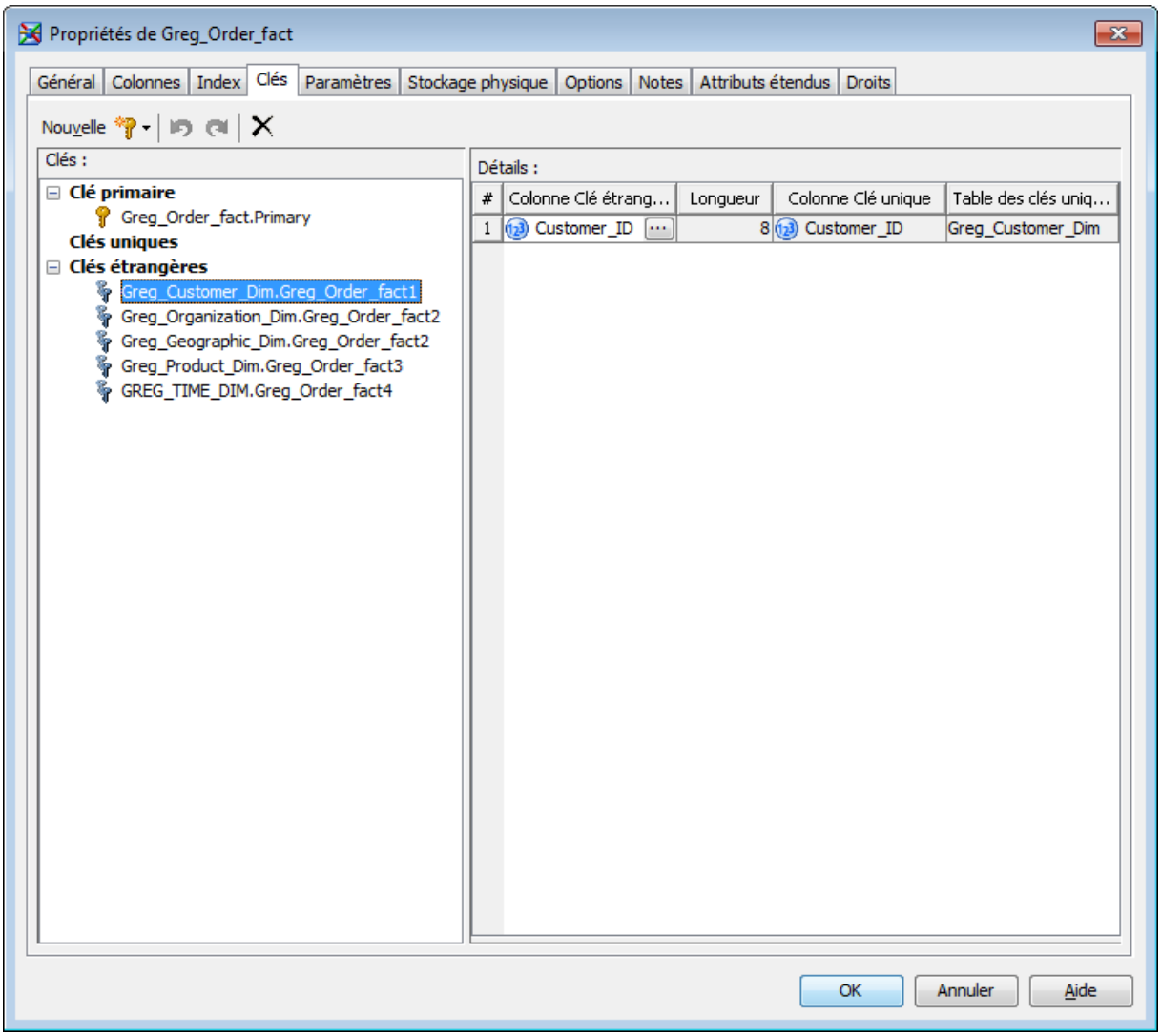
Remarque : on ne peut créer la table Order_fact que lorsque l'on a créé les 5 autres tables de dimension, car cette table contient des clés étrangères sur ces autres tables.

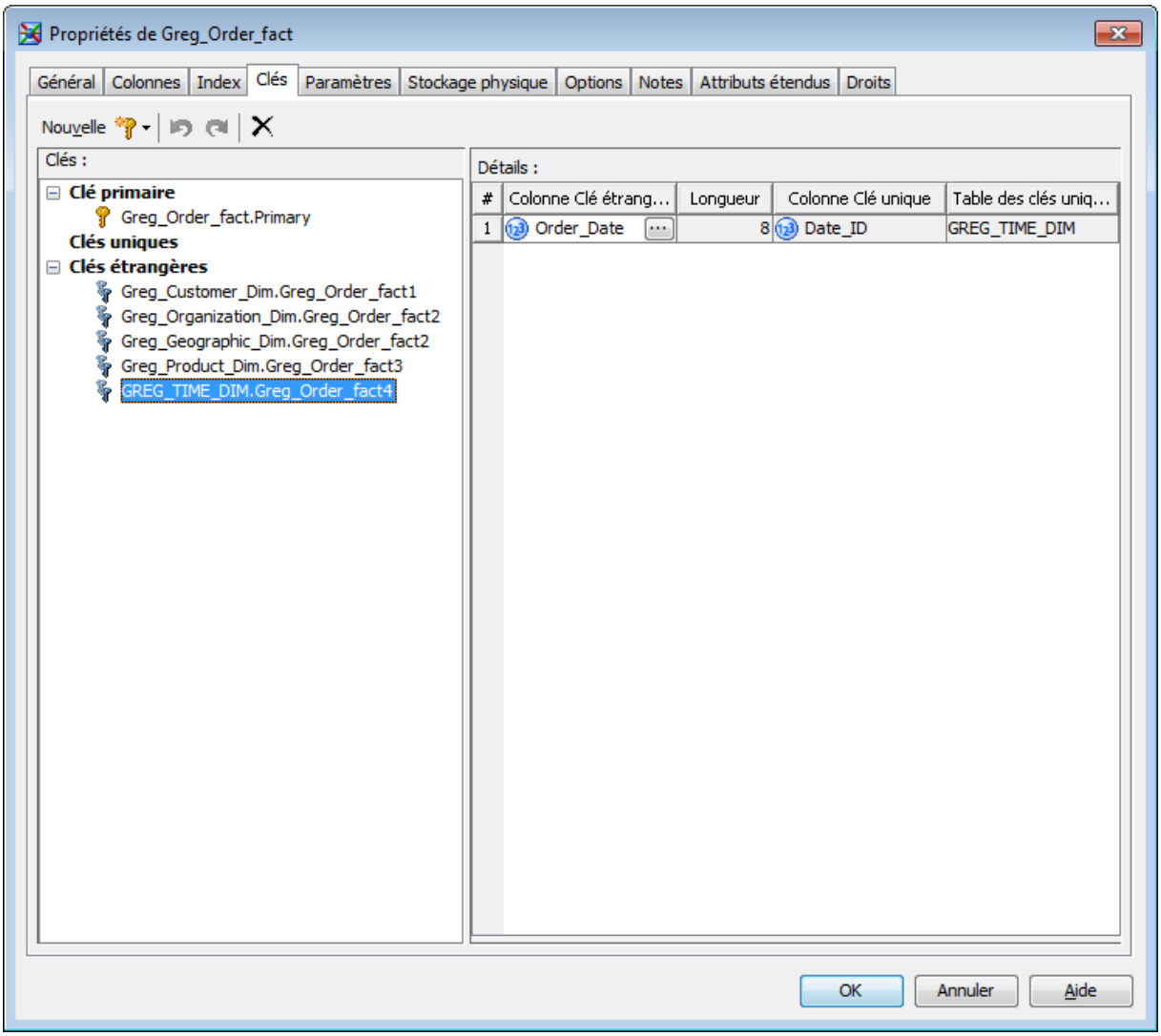
Créer la table ORDER_FACT avec les éléments suivants :

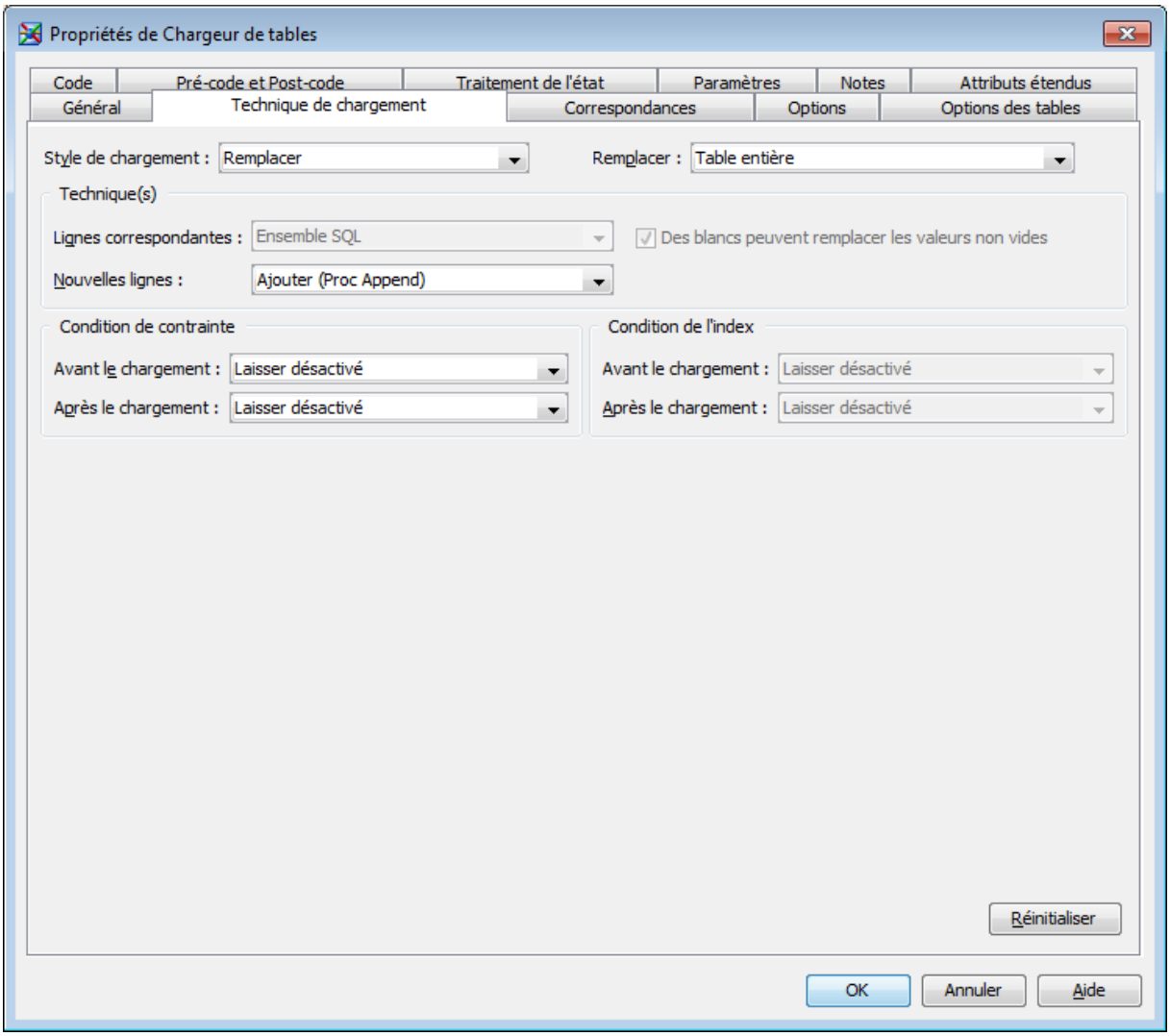
Nom	Longueur	Type	Format	Remarque
Customer_ID	8	Numérique	12.	Identique à la table source
Employee_ID	8	Numérique	12.	Identique à la table source
Street_ID	8	Numérique	12.	Identique à la table source
Order_Date	8	Numérique	DATE9.	Identique à la table source
Delivery_Date	8	Numérique	DATE9.	Identique à la table source
Order_ID	8	Numérique	12.	Identique à la table source
Order_Type	8	Numérique	ORDER_TYPE.	Identique à la table source
Product_ID	8	Numérique	12.	Identique à la table source
Quantity	8	Numérique	(Néant)	Identique à la table source
Total_Retail_Price	8	Numérique	DOLLAR13.2	Identique à la table source
CostPrice_Per_Unit	8	Numérique	DOLLAR13.2	Identique à la table source
Discount	8	Numérique	PERCENT.	Identique à la table source

Créer la clé primaire de la table ORDER_FACT comme ci-dessous :

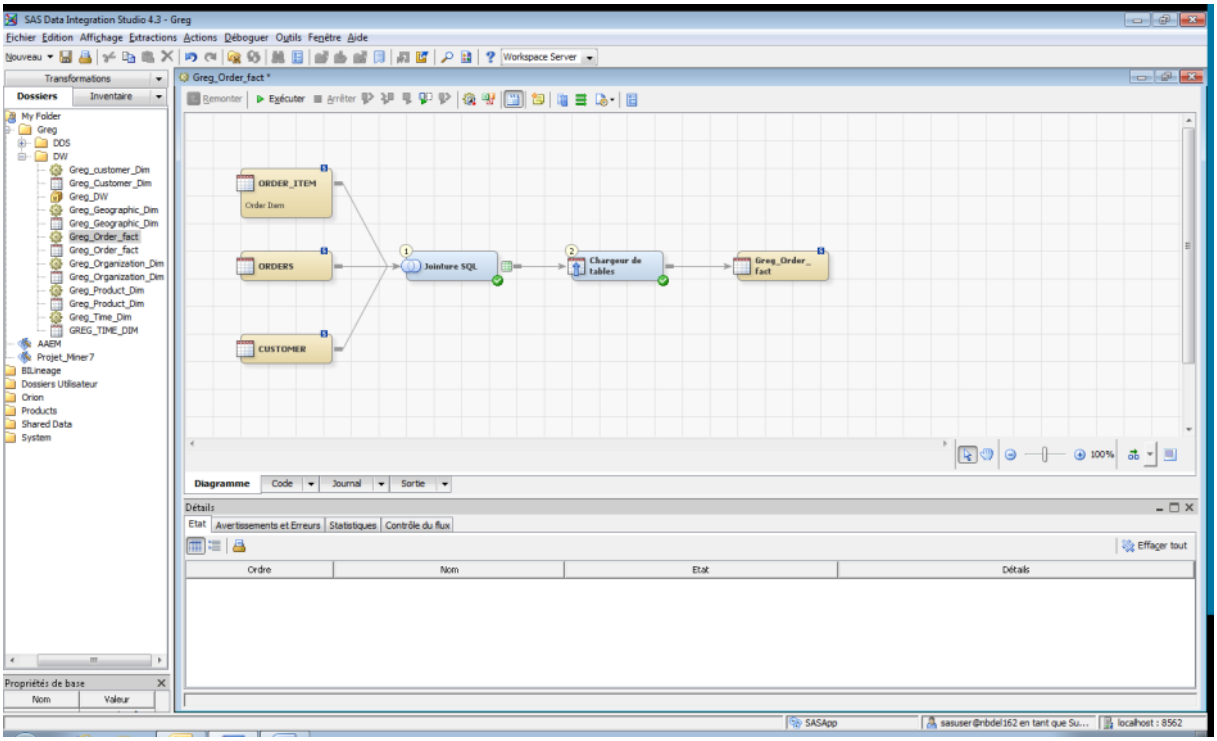

Propriétés de Greg_Order_fact
Général
Colonnes
Index
Clés
Paramètres
Stockage physique
Options
Notes
Attributs étendus
Droits
Nouvelle
Clés :
Clé primaire
Greg_Order_fact.Primary
Clés uniques
Clés étrangères
Greg_Customer_Dim.Greg_Order_fact1
Greg_Organization_Dim.Greg_Order_fact2
Greg_Geographic_Dim.Greg_Order_fact2
Greg_Product_Dim.Greg_Order_fact3
GREG_TIME_DIM.Greg_Order_fact4
Détails :
Nom Description Longueur
1 Order_ID Order ID 8
2 Employee_ID Employee ID 5
3 Customer_ID Customer ID 8
4 Order_Date Date Order was placed by C... 8
5 Product_ID Product ID 8
6 Street_ID Street ID 8
Ajouter
Supprimer
Tables des clés étrangères associées :
Nom de la table
Nom de la clé étrangère
OK
Annuler
Aide


Propriétés de Greg_Order_fact
Général
Colonnes
Index
Clés
Paramètres
Stockage physique
Options
Notes
Attributs étendus
Droits
Nouvelle
Clés :
Clé primaire
Greg_Order_fact.Primary
Clés uniques
Clés étrangères
Greg_Customer_Dim.Greg_Order_fact1
Greg_Organization_Dim.Greg_Order_fact2
Greg_Geographic_Dim.Greg_Order_fact2
Greg_Product_Dim.Greg_Order_fact3
GREG_TIME_DIM.Greg_Order_fact4
Détails :
Colonne Clé étrang... Longueur Colonne Clé unique Table des clés uniq...
1 Customer_ID 8 Customer_ID Greg_Customer_Dim
OK
Annuler
Aide

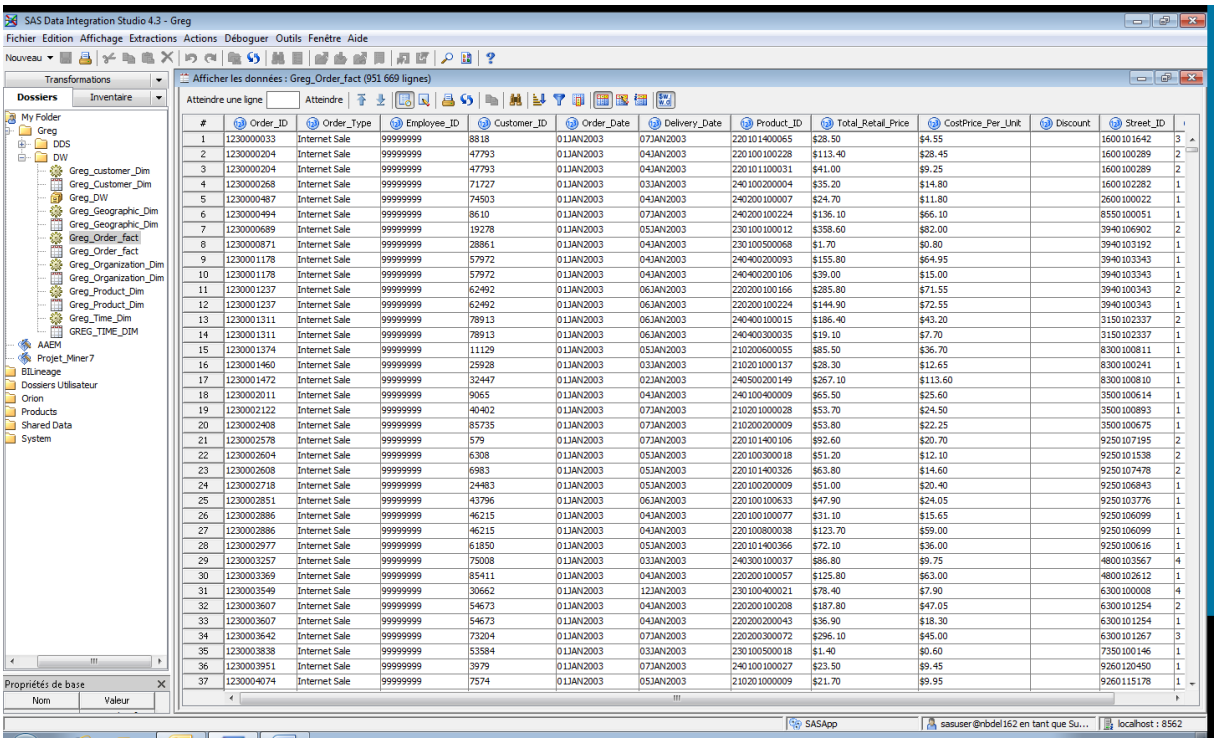

Propriétés de Greg_Order_fact
Général
Colonnes
Index
Clés
Paramètres
Stockage physique
Options
Notes
Attributs étendus
Droits
Nouvelle
Clés :
Clé primaire
Greg_Order_fact.Primary
Clés uniques
Clés étrangères
Greg_Customer_Dim.Greg_Order_fact1
Greg_Organization_Dim.Greg_Order_fact2
Greg_Geographic_Dim.Greg_Order_fact2
Greg_Product_Dim.Greg_Order_fact3
GREG_TIME_DIM.Greg_Order_fact4
Détails :
Colonne Clé étrang... Longueur Colonne Clé unique Table des clés uniq...
1 Order_Date 8 Date_ID GREG_TIME_DIM
OK
Annuler
Aide


Propriétés de Chargeur de tables
Code
Pré-code et Post-code
Traitement de l'état
Paramètres
Notes
Attributs étendus
Général
Technique de chargement
Correspondances
Options
Options des tables
Style de chargement :
Remplacer
Remplacer :
Table entière
Technique(s)
Lignes correspondantes :
Ensemble SQL
Des blancs peuvent remplacer les valeurs non vides
Nouvelles lignes :
Ajouter (Proc Append)
Condition de contrainte
Avant le chargement :
Laisser désactivé
Après le chargement :
Laisser désactivé
Condition de l'index
Avant le chargement :
Laisser désactivé
Après le chargement :
Laisser désactivé
Réinitialiser
OK
Annuler
Aide

Il faut toutes les lignes de la table Order_item.

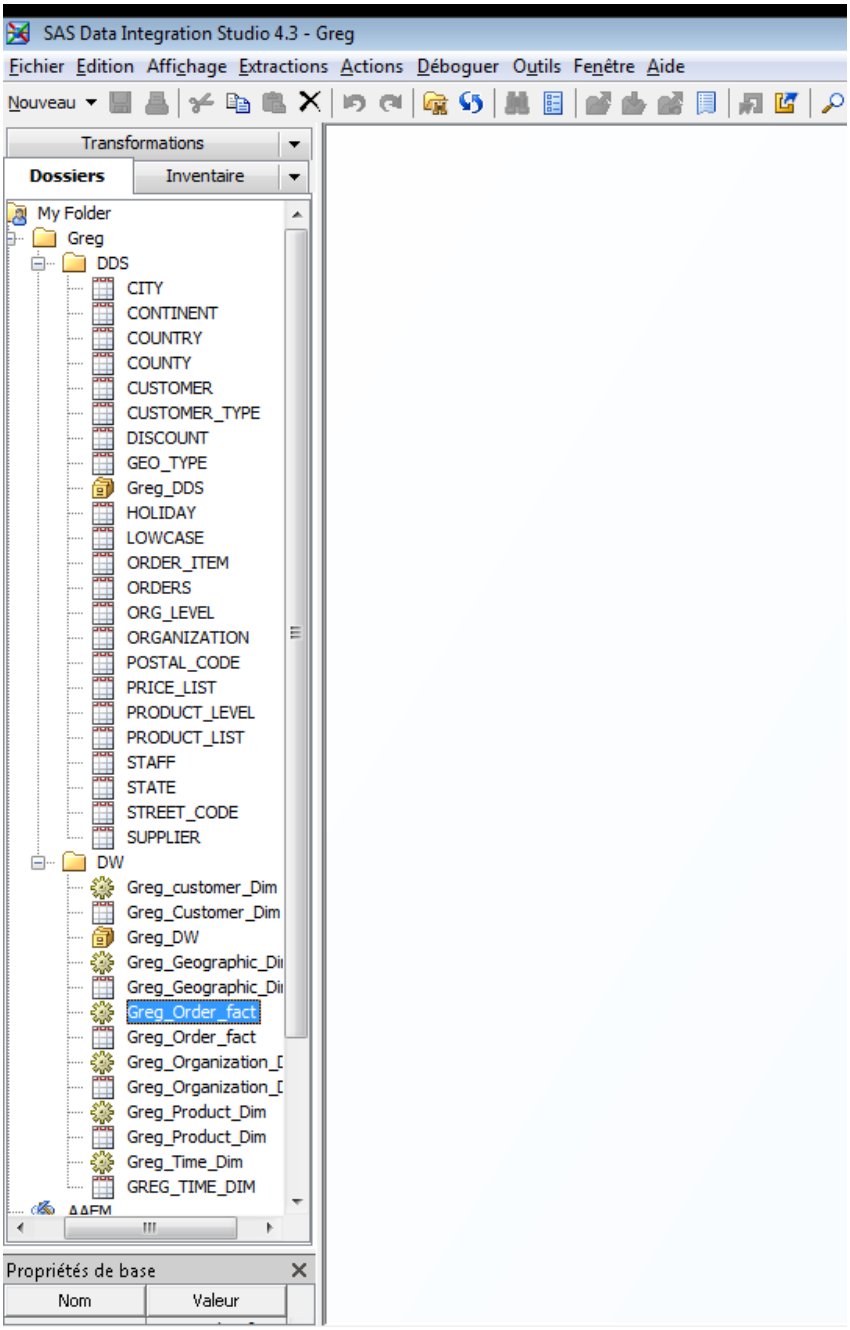


Si vous laissez activée la vérification des contraintes, dans les techniques de chargement, du chargeur, le processus peut être très long.



#	Order_ID	Order_Type	Employee_ID	Customer_ID	Order_Date	Delivery_Date	Product_ID	Total_Retail_Price	CostPrice_Per_Unit	Discount	Street_ID
1	1230000033	Internet Sale	99999999	8818	01JAN2003	07JAN2003	220101400065	$28.50	$4.55		1600101642
2	1230000204	Internet Sale	99999999	47793	01JAN2003	04JAN2003	220100100228	$113.40	$28.45		1600100289
3	1230000204	Internet Sale	99999999	47793	01JAN2003	04JAN2003	220101100031	$41.00	$9.25		1600100289
4	1230000268	Internet Sale	99999999	71727	01JAN2003	03JAN2003	240100200004	$35.20	$14.80		1600102282
5	1230000487	Internet Sale	99999999	74503	01JAN2003	04JAN2003	240200100007	$24.70	$11.80		2600100022
6	1230000494	Internet Sale	99999999	8610	01JAN2003	07JAN2003	240200100224	$136.10	$66.10		8550100051
7	1230000689	Internet Sale	99999999	19278	01JAN2003	05JAN2003	230100100012	$358.60	$82.00		3940106902
8	1230000871	Internet Sale	99999999	28861	01JAN2003	04JAN2003	230100500068	$1.70	$0.80		3940103192
9	1230001178	Internet Sale	99999999	57972	01JAN2003	04JAN2003	240400200093	$155.80	$64.95		3940103343
10	1230001178	Internet Sale	99999999	57972	01JAN2003	04JAN2003	240400200106	$39.00	$15.00		3940103343
11	1230001237	Internet Sale	99999999	62492	01JAN2003	06JAN2003	220200100166	$285.80	$71.55		3940100343
12	1230001237	Internet Sale	99999999	62492	01JAN2003	06JAN2003	220200100224	$144.90	$72.55		3940100343
13	1230001311	Internet Sale	99999999	78913	01JAN2003	06JAN2003	240400100015	$186.40	$43.20		3150102337
14	1230001311	Internet Sale	99999999	78913	01JAN2003	06JAN2003	240400300035	$19.10	$7.70		3150102337
15	1230001374	Internet Sale	99999999	11129	01JAN2003	05JAN2003	210200600055	$85.50	$36.70		8300100811
16	1230001460	Internet Sale	99999999	25928	01JAN2003	03JAN2003	210201000137	$28.30	$12.65		8300100241
17	1230001472	Internet Sale	99999999	32447	01JAN2003	02JAN2003	240500200149	$267.10	$113.60		8300100810
18	1230002011	Internet Sale	99999999	9065	01JAN2003	04JAN2003	240100400009	$65.50	$25.60		3500100614
19	1230002122	Internet Sale	99999999	40402	01JAN2003	07JAN2003	210201000028	$53.70	$24.50		3500100893
20	1230002408	Internet Sale	99999999	85735	01JAN2003	07JAN2003	210200200009	$53.80	$22.25		3500100675
21	1230002578	Internet Sale	99999999	579	01JAN2003	07JAN2003	220101400106	$92.60	$20.70		9250107195
22	1230002604	Internet Sale	99999999	6308	01JAN2003	05JAN2003	220100300018	$51.20	$12.10		9250101538
23	1230002608	Internet Sale	99999999	6983	01JAN2003	05JAN2003	220101400326	$63.80	$14.60		9250107478
24	1230002718	Internet Sale	99999999	24483	01JAN2003	05JAN2003	220100200009	$51.00	$20.40		9250106843
25	1230002851	Internet Sale	99999999	43796	01JAN2003	06JAN2003	220100100633	$47.90	$24.05		9250103776
26	1230002886	Internet Sale	99999999	46215	01JAN2003	04JAN2003	220100100077	$31.10	$15.65		9250106099
27	1230002886	Internet Sale	99999999	46215	01JAN2003	04JAN2003	220100800038	$123.70	$59.00		9250106099
28	1230002977	Internet Sale	99999999	61850	01JAN2003	05JAN2003	220101400366	$72.10	$36.00		9250100616
29	1230003257	Internet Sale	99999999	75008	01JAN2003	03JAN2003	240300100037	$86.80	$9.75		4800103567
30	1230003369	Internet Sale	99999999	85411	01JAN2003	04JAN2003	220200100057	$125.80	$63.00		4800102612
31	1230003549	Internet Sale	99999999	30662	01JAN2003	12JAN2003	230100400021	$78.40	$7.90		6300100008
32	1230003607	Internet Sale	99999999	54673	01JAN2003	04JAN2003	220200100208	$187.80	$47.05		6300101254
33	1230003607	Internet Sale	99999999	54673	01JAN2003	04JAN2003	220200200043	$36.90	$18.30		6300101254
34	1230003642	Internet Sale	99999999	73204	01JAN2003	07JAN2003	220200300072	$296.10	$45.00		6300101267
35	1230003838	Internet Sale	99999999	53584	01JAN2003	03JAN2003	230100500018	$1.40	$0.60		7350100146
36	1230003951	Internet Sale	99999999	3979	01JAN2003	07JAN2003	240100100027	$23.50	$9.45		9260120450
37	1230004074	Internet Sale	99999999	7574	01JAN2003	05JAN2003	210201000009	$21.70	$9.95		9260115178

Résumé sur la partie ETL



Vous devez avoir :

89 954 lignes dans la table de la dimension client
89 807 lignes dans la table de la dimension géographique
1 049 lignes dans la table de la dimension organisation
5 504 lignes dans la table de la dimension produit
2 923 lignes dans la table de la dimension temps
951 669 lignes dans la table de fait

OLAP

Le sigle OLAP veut dire OnLine Analytical Process. Un cube OLAP est une structure multidimensionnelle. En plus des données contenues dans les tables sources, un cube peut contenir des informations agrégées, permettant de naviguer plus rapidement dans les données. On peut créer des cubes dans ETL Data Integration Studio ou dans OLAP Studio. Si vous avez uniquement SAS OLAP serveur sous licence, vous utiliserez SAS OLAP Studio. Si vous avez aussi SAS ETL Server, vous pouvez utiliser les deux.
On peut générer des cubes de type ROLAP, MOLAP ou HOLAP. Voir paragraphe sur l'optimisation des cubes.

Lorsque l'on parle de cube multidimensionnel, il faut voir deux éléments : la partie émergée de l'iceberg et celle cachée.
La partie émergée, celle qui sera vue par les utilisateurs, a quelque chose de magique. Lorsque vous présentez à vos futurs utilisateurs un cube, ils sont souvent fascinés. Vous leur offrez enfin la possibilité de naviguer facilement dans l'information, du général au détail, et vice-versa. Ils peuvent croiser l'information de multiples façons. Voir les chapitres Navigation dans un cube depuis SAS Enterprise Guide et Microsoft Excel.
La partie cachée : la structure du cube d'un point de vue informatique permettant l'optimum en terme de temps de réponse et d'espace disque. Voir le chapitre d'optimisation d'un cube OLAP.

Vocabulaire OLAP

Le sigle OLAP veux dire OnLine Analytical Process, par opposition aux bases de données OLTP (OnLine transactional Process) des SGBD (Système de Gestion de Bases de Données) opérationnelles. Une base OLAP est conçue pour naviguer dans l'information et donc l'analyser alors qu'une base OLTP est conçue pour supporter des transactions.

On parle de cube par opposition à une table qui a deux dimensions, des lignes et des colonnes ; un cube en a trois. La vraie dénomination devrait être hyper-cube multidimensionnel car une structure OLAP a souvent quatre, cinq, six voir sept dimensions. Pour représenter une table, on dessine souvent un tableau. On peut dessiner un cube avec ses trois dimensions, mais il est difficile de représenter un hyper-cube avec quatre ou plus dimensions.
Il est rare qu'un cube excède sept dimensions, ce qui est déjà beaucoup. Lors de la conception, si l'on a plus de cinq dimensions pour un problème donné, on a plutôt tendance à créer deux cubes. En effet, il est rare de trouver un problème avec tant de dimensions et après analyse, on s'aperçoit souvent qu'il vaut mieux diviser ce problème en deux sous-problèmes.
Il faut remarquer l'exception des cubes pour le contrôle de gestion suivant la méthode ABC où l'on crée souvent des cubes avec plus de 25 dimensions. L'Activity Based Costing est une méthode de comptabilité analytique par activités.

Un cube comprend

- Des **dimensions** : les axes d'analyses
- Des **hiérarchies** : les chemins de navigation dans l'information
- Des **niveaux** d'agrégations
- Des **mesures** : une mesure est l'indicateur que l'on analyse dans un cube.

Par exemple, on peut analyser la somme du chiffre d'affaires, une mesure, suivant les dimensions produit, organisation et client. Une hiérarchie de la dimension produit peut être composée des niveaux : Ligne de produit→ Catégorie de produit → Groupe de produit → Produit. Une hiérarchie de la dimension organisation peut être : Pays → Département → Magasin → Section → Groupe → Employé. Une autre hiérarchie de la dimension organisation peut être Magasin → Employé. La dimension client peut avoir une hiérarchie : Groupe d'âge du client → Age du client → Sexe du client.

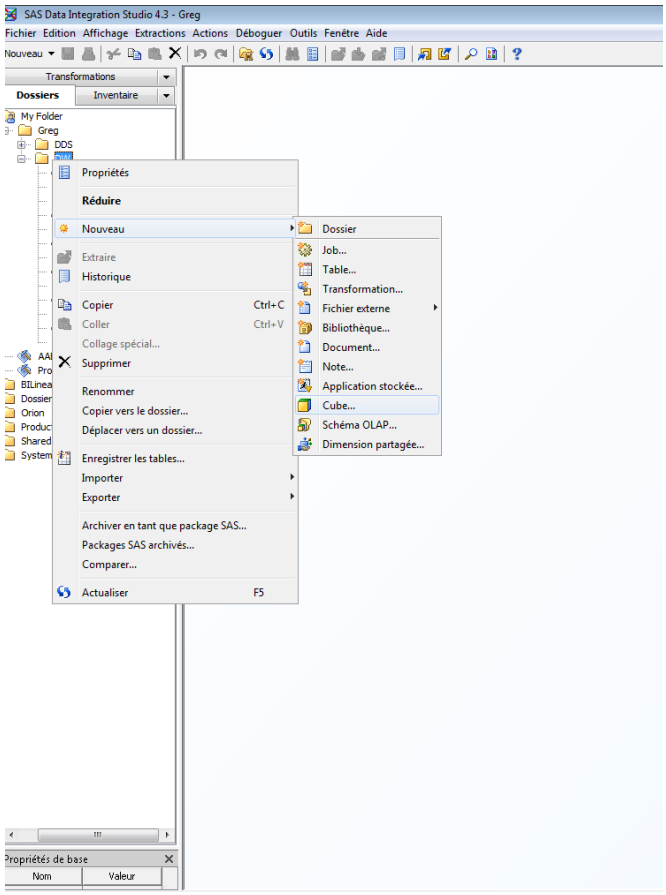
Règles:
Un hyper-cube multidimensionnel, par abus de langage, un cube, peut avoir une ou plusieurs dimensions et une ou plusieurs mesures.
Une dimension peut avoir une ou plusieurs hiérarchies.
Une hiérarchie peut avoir un ou plusieurs niveaux.
Un niveau appartient à au moins une hiérarchie.

Un niveau appartient à une seule dimension.
Une mesure peut être présentée par une ou plusieurs statistiques : la somme, la moyenne, le minimum, le maximum, etc…

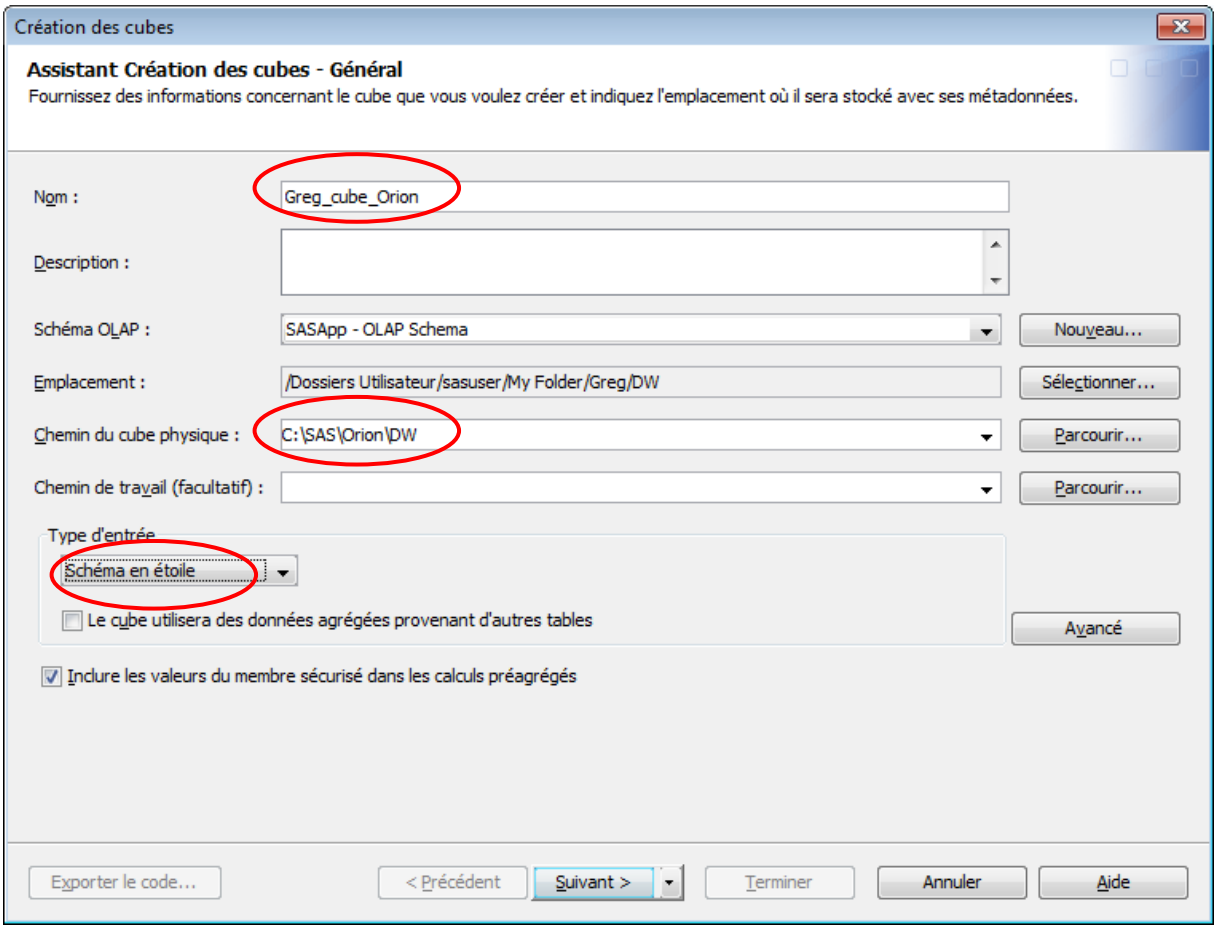
Mise en œuvre, création du cube Orion :

Nous allons créer le cube OLAP sur le schéma en étoile Orion.

Si vous n'avez pas créé les 5 tables de dimensions et la table de fait, vous pouvez créer le cube sur les données Orion Gold. Pour cela, il faut créer une bibliothèque votre_nom_Orion_Gold de type SAS Base, sur le serveur SASApp, de libref vos_initiale_OG pointant sur le dossier C:\SAS\Orion\OrionStarGold (ou dossier équivalent) ; puis enregistrer toutes les tables.



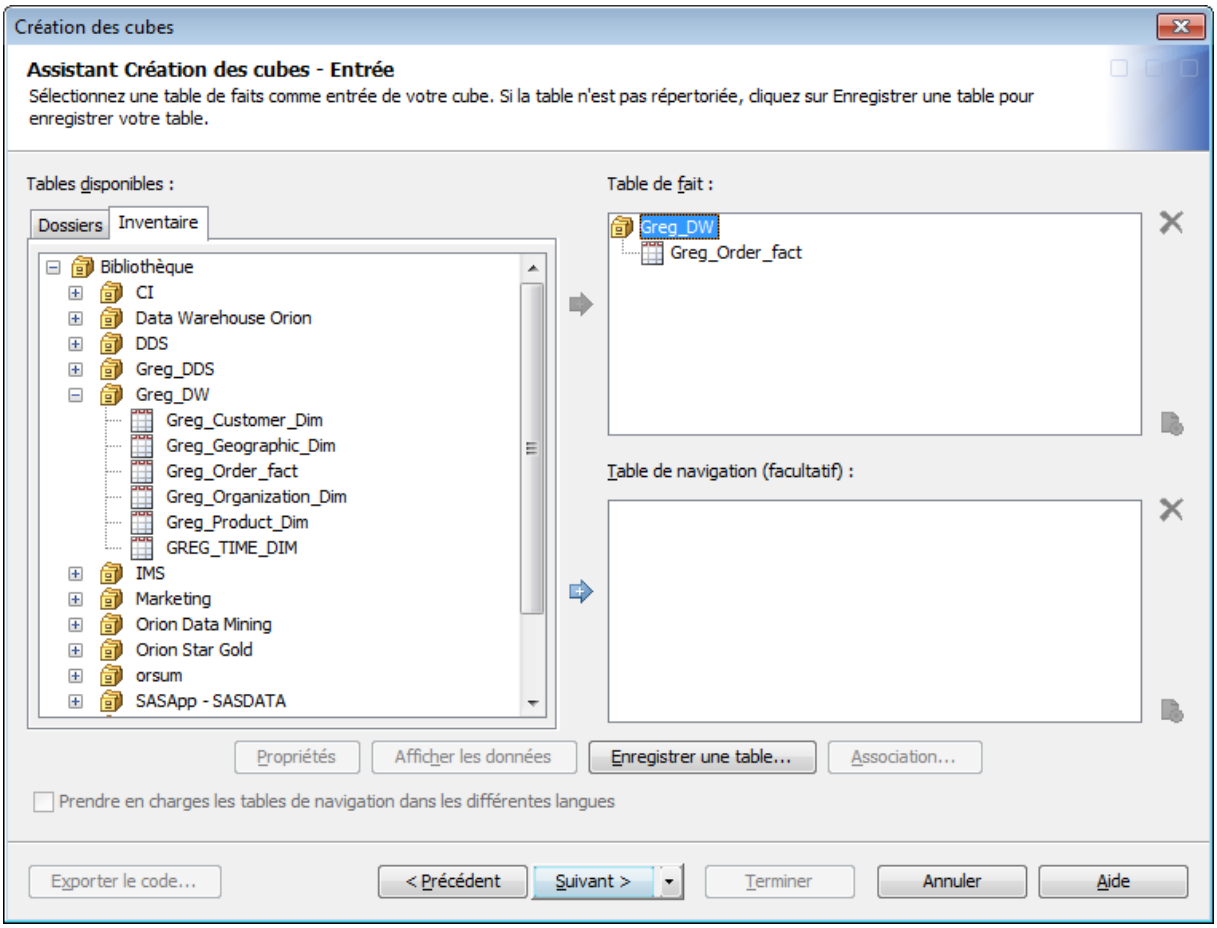
Créer un nouveau cube



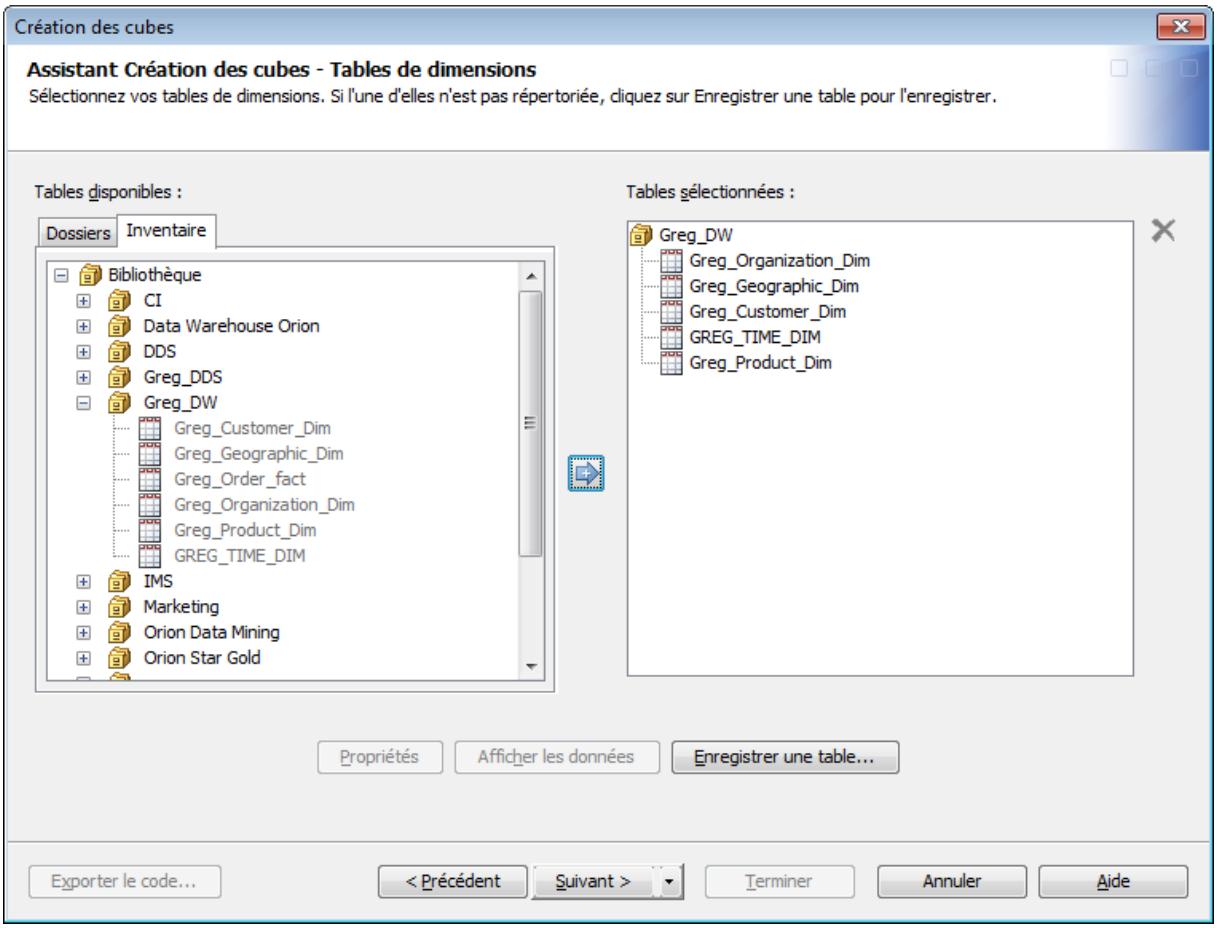
Entrer :

1. le nom du cube (Votre_Nom_Cube_Orion sans espace ou caractère spécial),
2. le chemin du répertoire de stockage physique du cube
3. et l'option « schéma en étoile ».

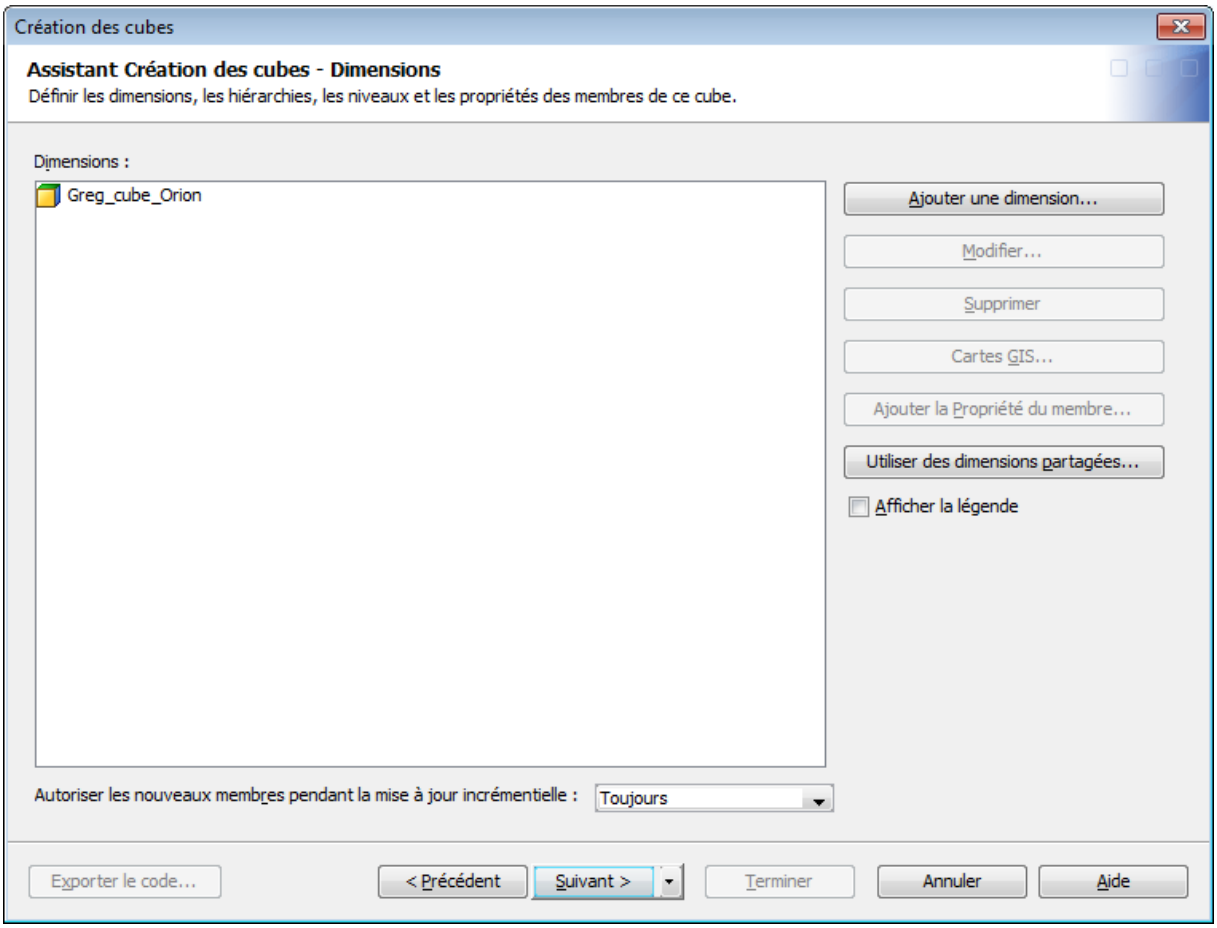
Suivant.



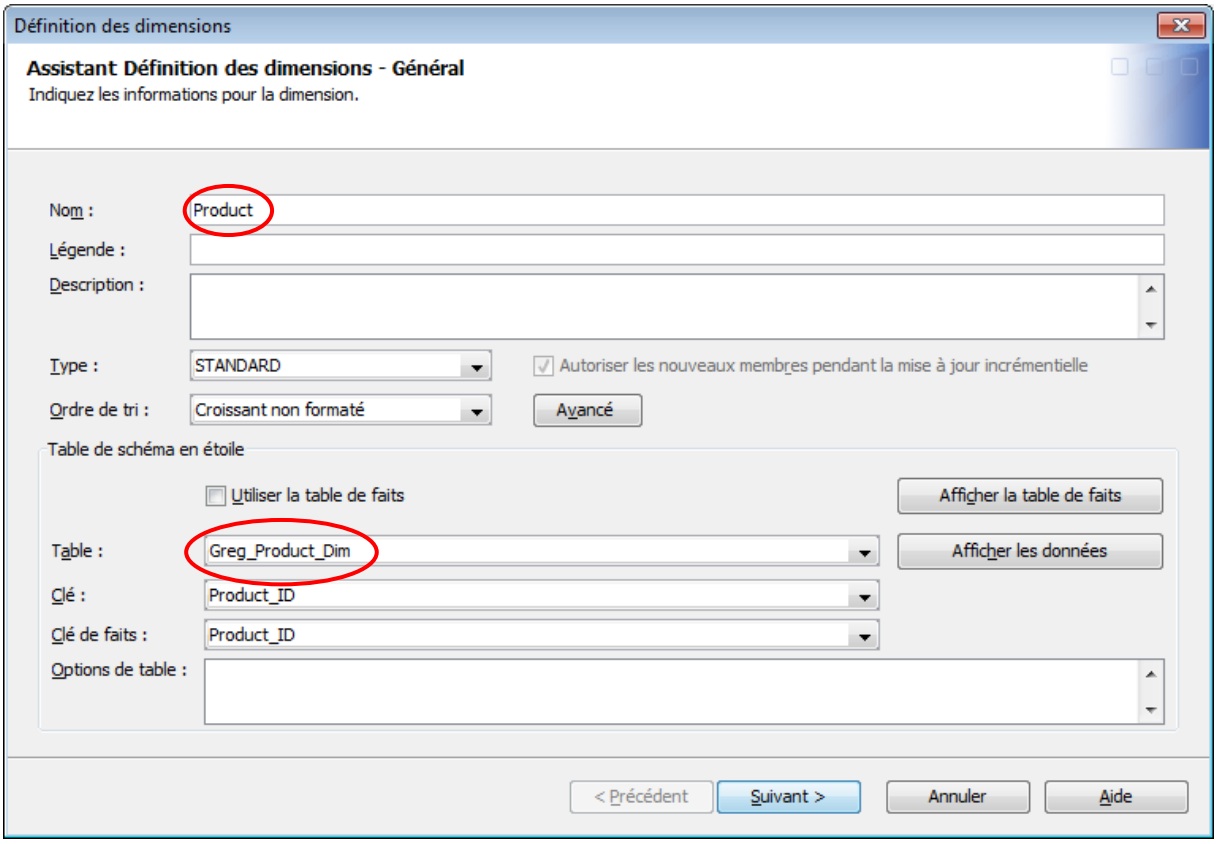
Sélectionner votre table de fait dans votre bibliothèque : votre_nom_Order_Fact (ou bien la table Order_fact de la bibliothèque Orion Gold)
Suivant.



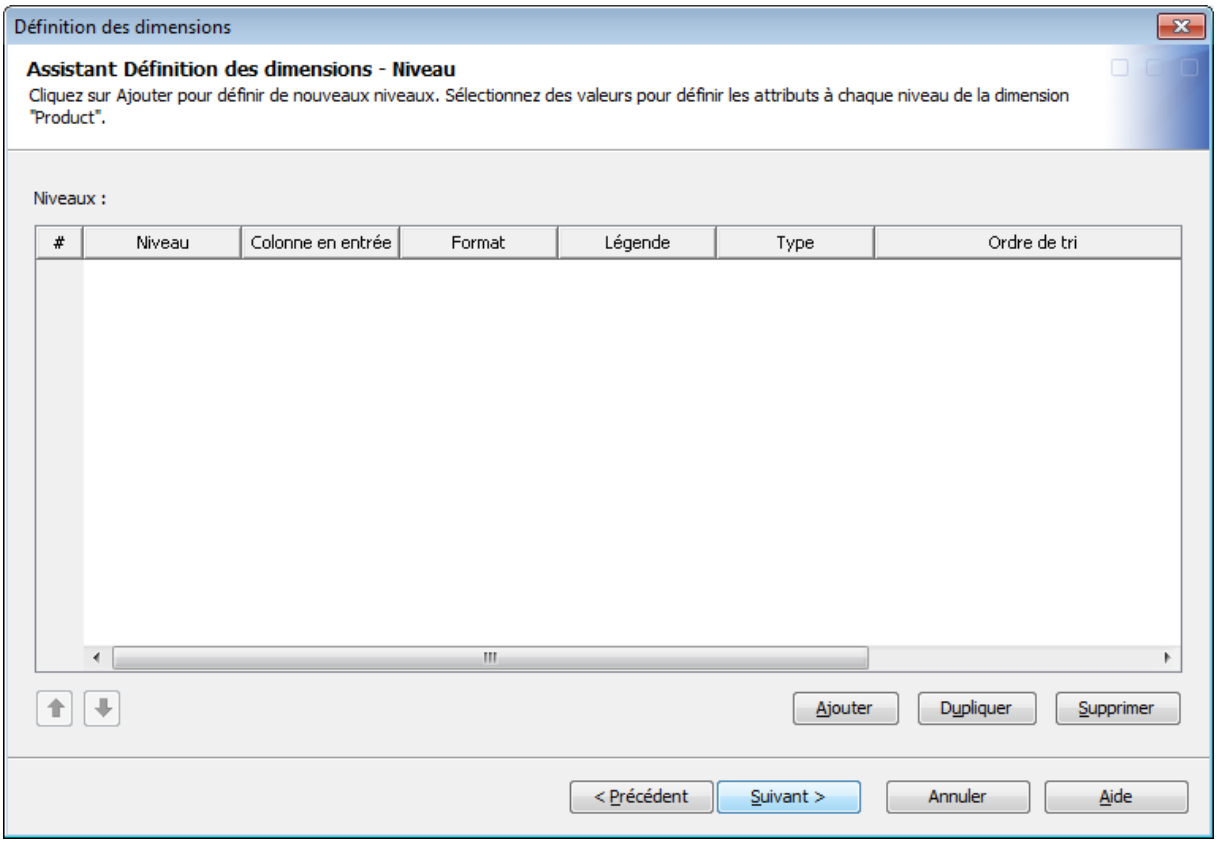
Sélectionner les cinq tables de dimension correspondantes :
Suivant.



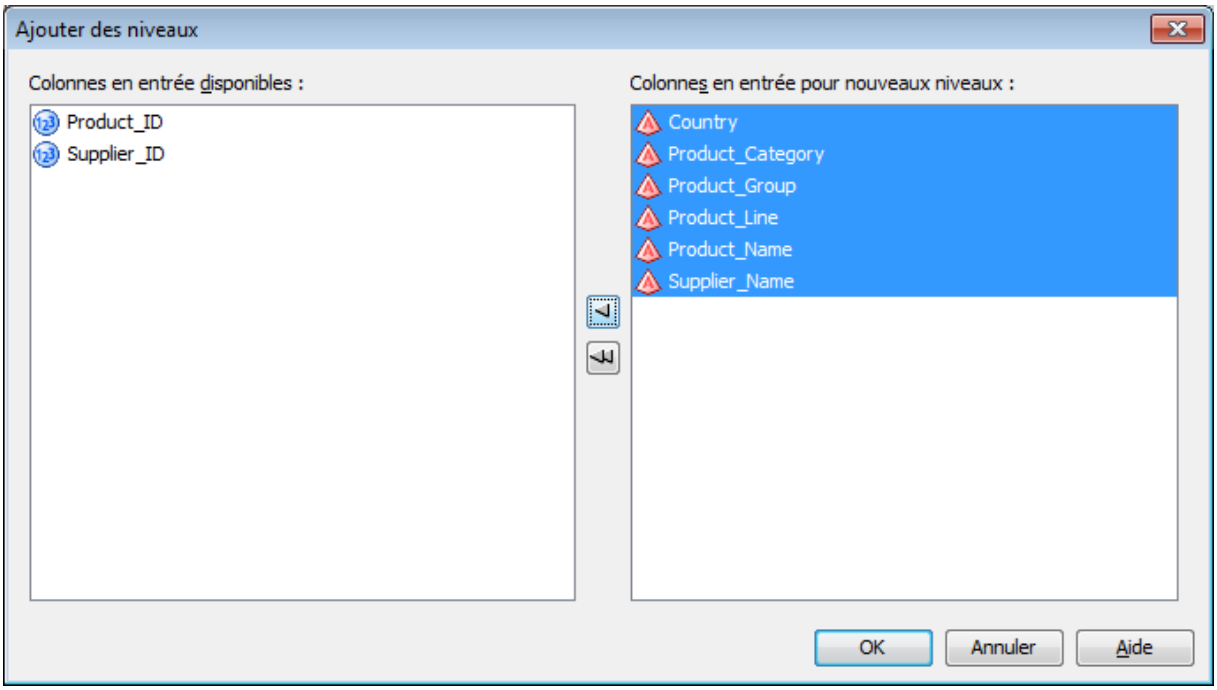
Cliquer sur « Ajouter » pour créer une nouvelle dimension.



Lui donner un nom (pas d'espace, ni de caractère spéciaux)
Vous pouvez mettre ce que vous souhaitez dans la description.
Sélectionner la table de dimension,
Vérifier que la clé de la table de dimension est bien la même que celle de la table de fait.
Suivant

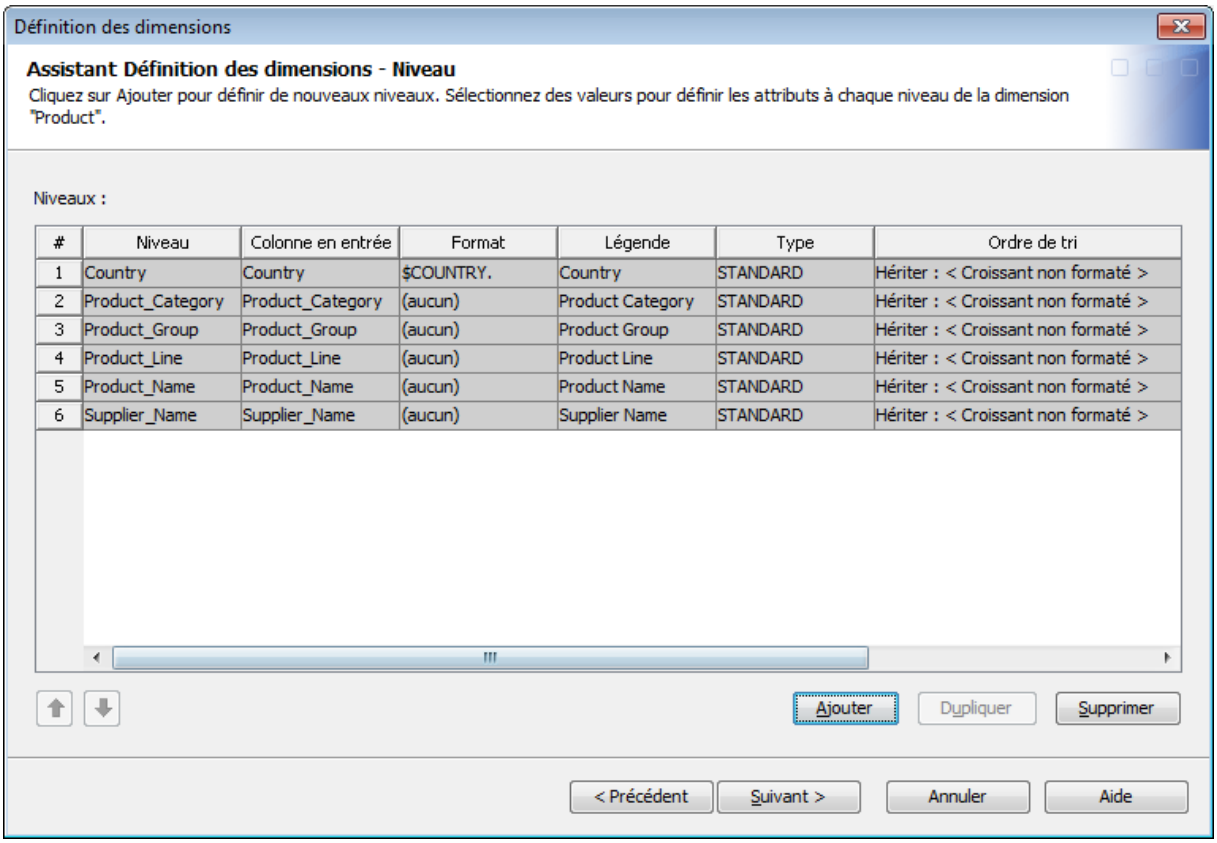


Ajouter

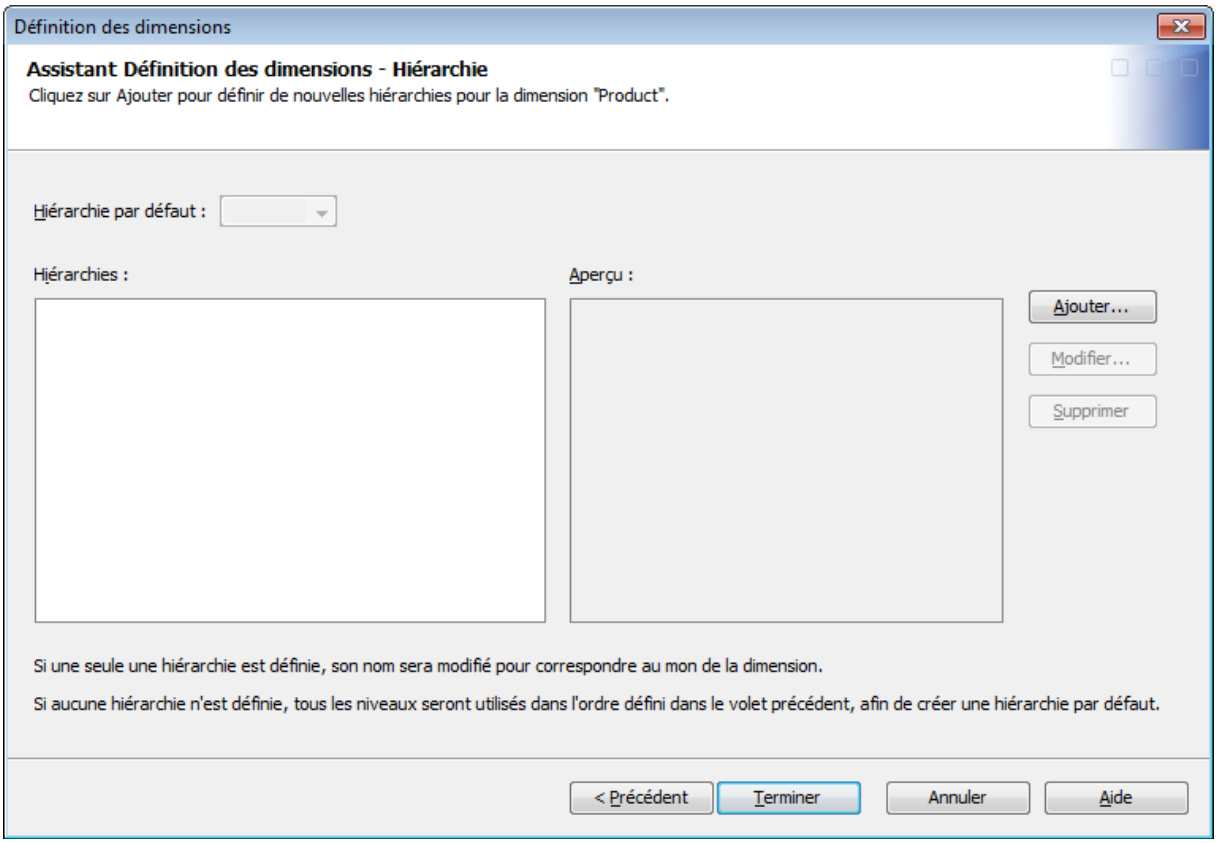


Pour la dimension produit, on souhaite Product_Line, Product_Category, Product_Group, Product_Name, Supplier_Name et Supplier_Country.

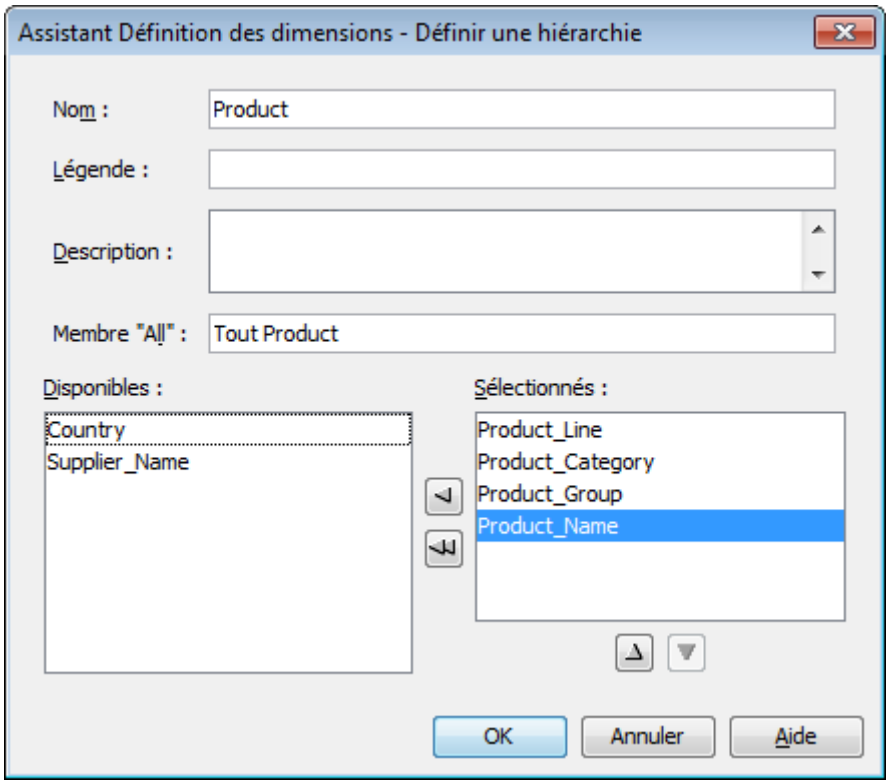
Sélectionner les niveaux que vous souhaitez,



Suivant



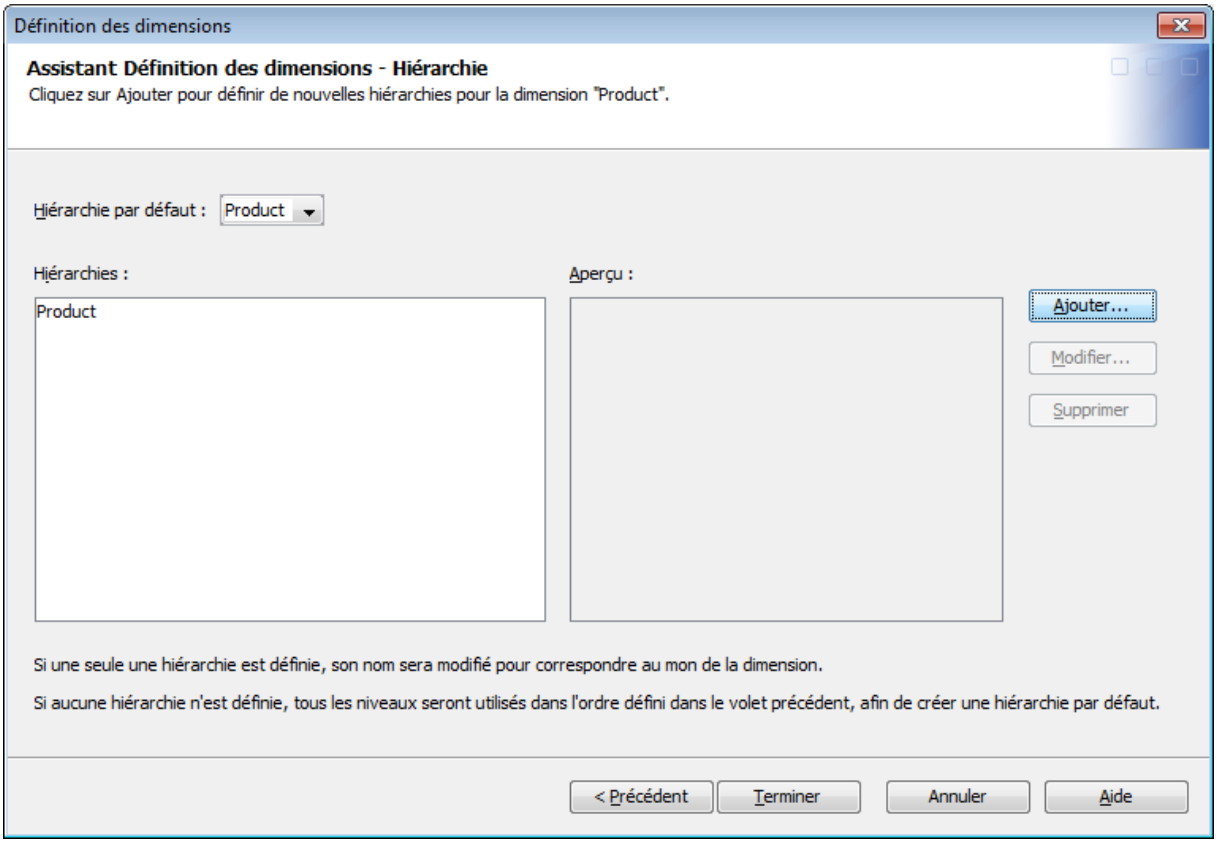
Ajouter



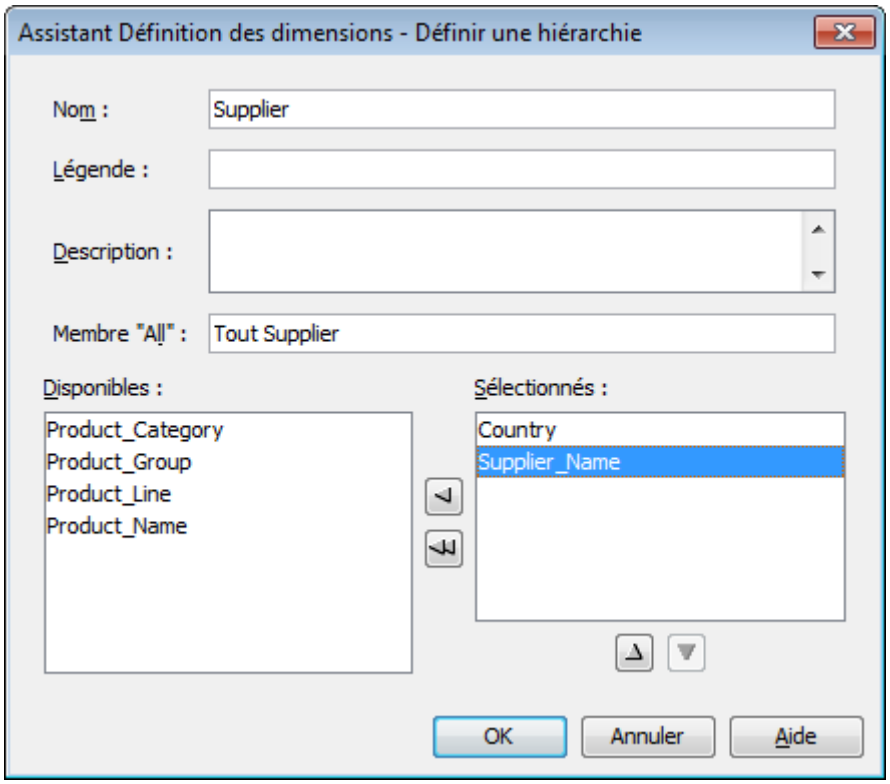
Donner un nom à la hiérarchie,

Sélectionnez dans l'ordre les niveaux que vous souhaitez
Pour la dimension Produit : Product_Line → Product_Category → Product_Group → Product_Name

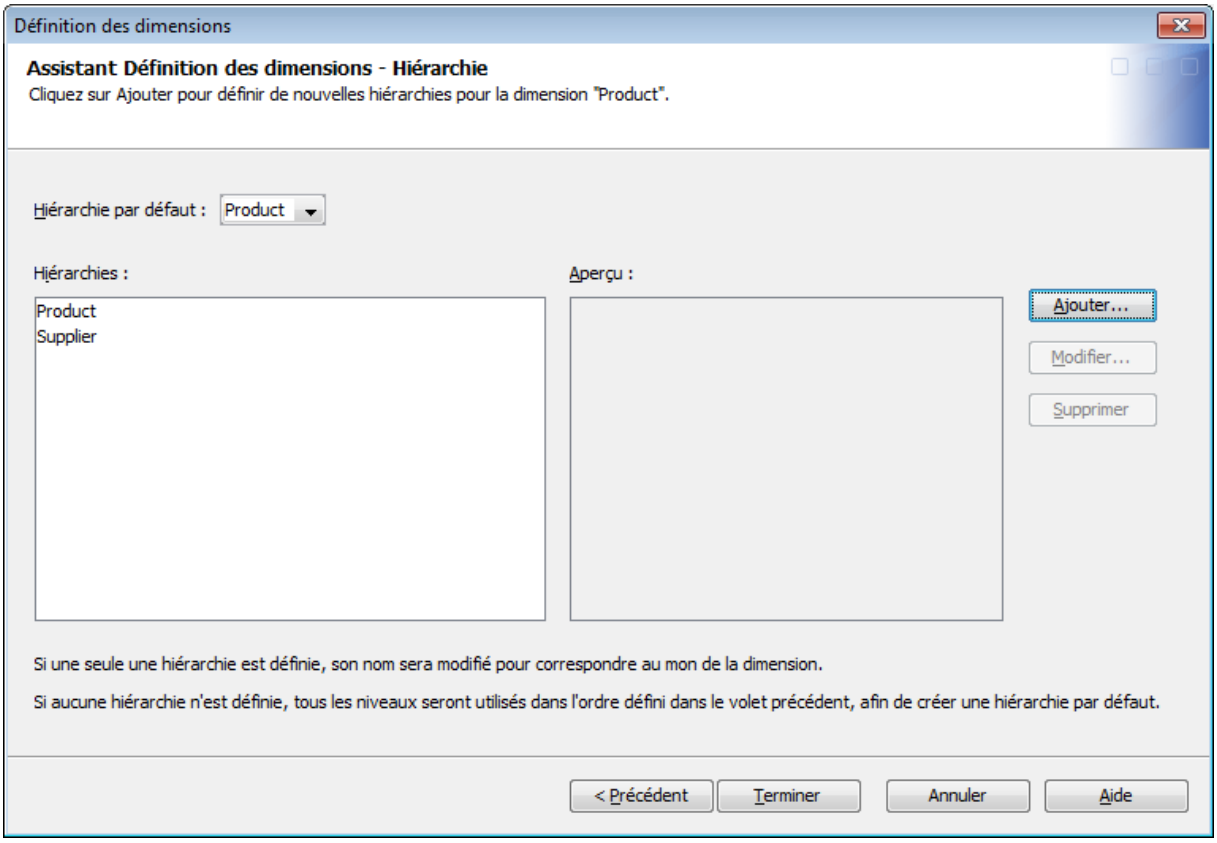
OK



Ajouter la hiérarchie fournisseur suivant la dimension produit.



OK



Suivant la dimension produit, vous avez deux hiérarchies, celle produit et celle fournisseur.

Terminer

Créer les autres dimensions de la même manière, de tel sorte que vous obteniez un cube comme ci-dessous :

Selon la dimension Produit :

Hiérarchie Produit : Ligne → catégorie → groupe → Produit
Hiérarchie Fournisseur : Pays → Fournisseur

Selon la dimension Temps : (jointure parfois sur Date_ID et Order_date)

Hiérarchie Temps : Année → trimestre → mois → date
Hiérarchie Temps Fiscal : Année fiscal → trimestre fiscal → mois fiscal

Selon la dimension Client :

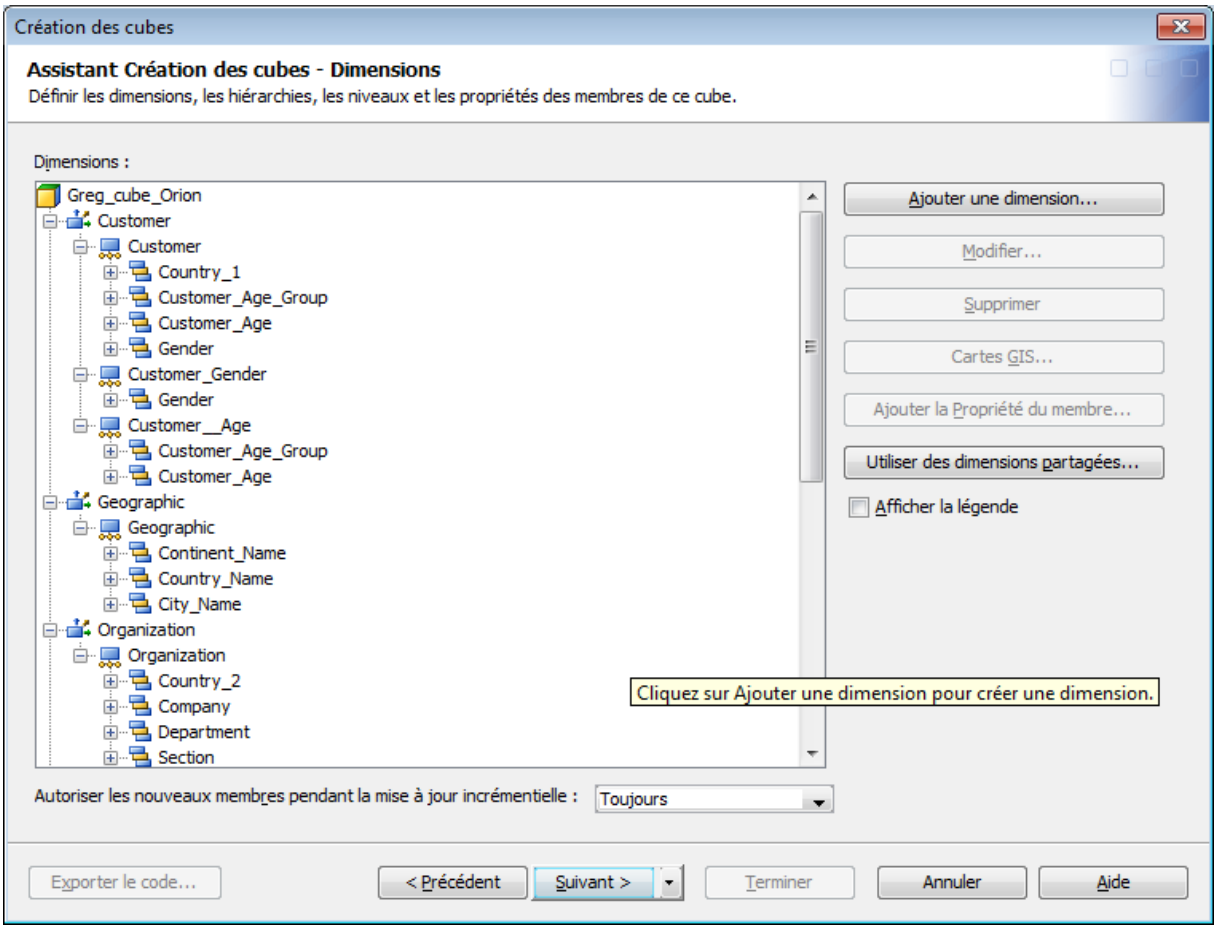
Hiérarchie Client : Pays du client → groupe d'âge → Age → Sexe du client
Hiérarchie : Age du client : Groupe d'âge → Age
Hiérarchie sexe du Client : Sexe du client

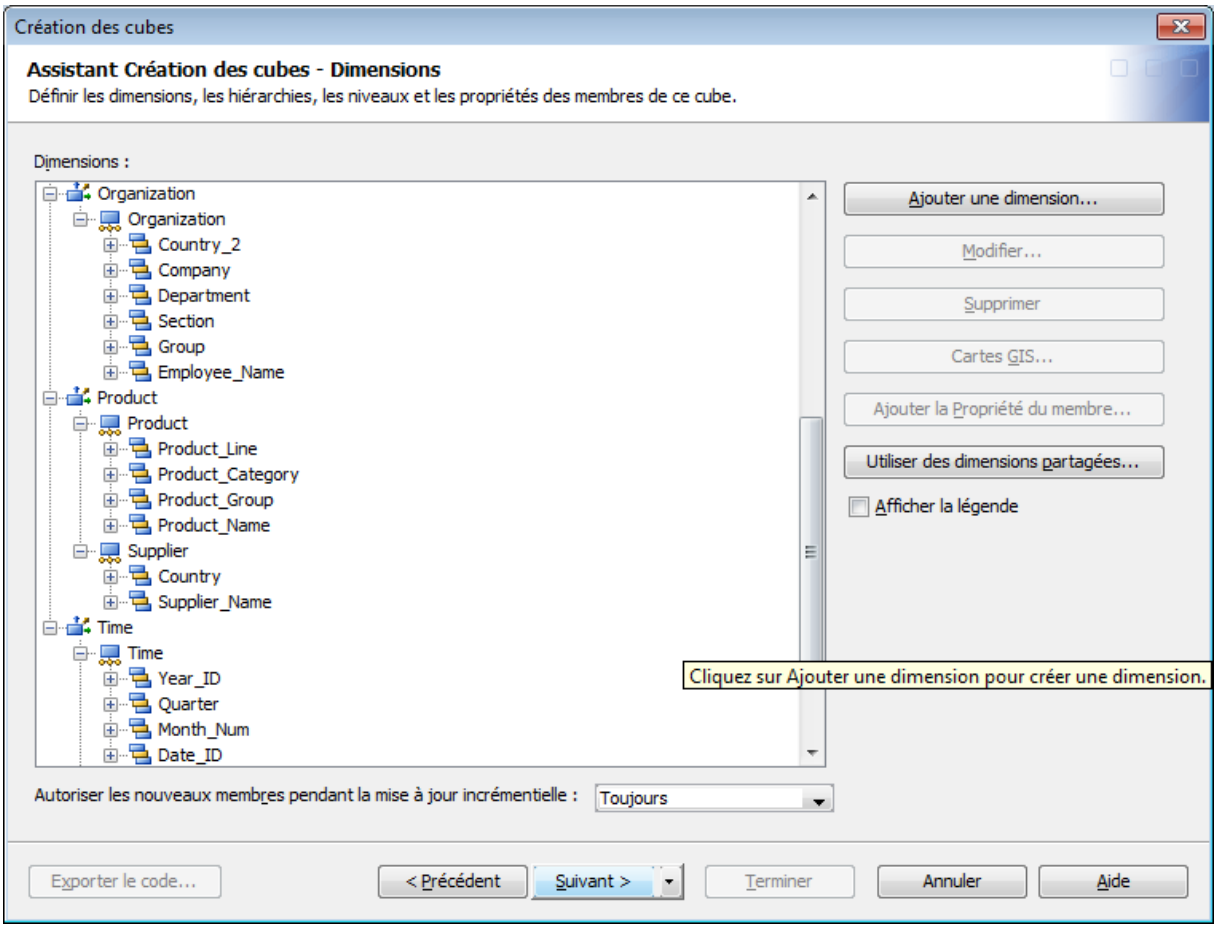
Selon la dimension Organisation :

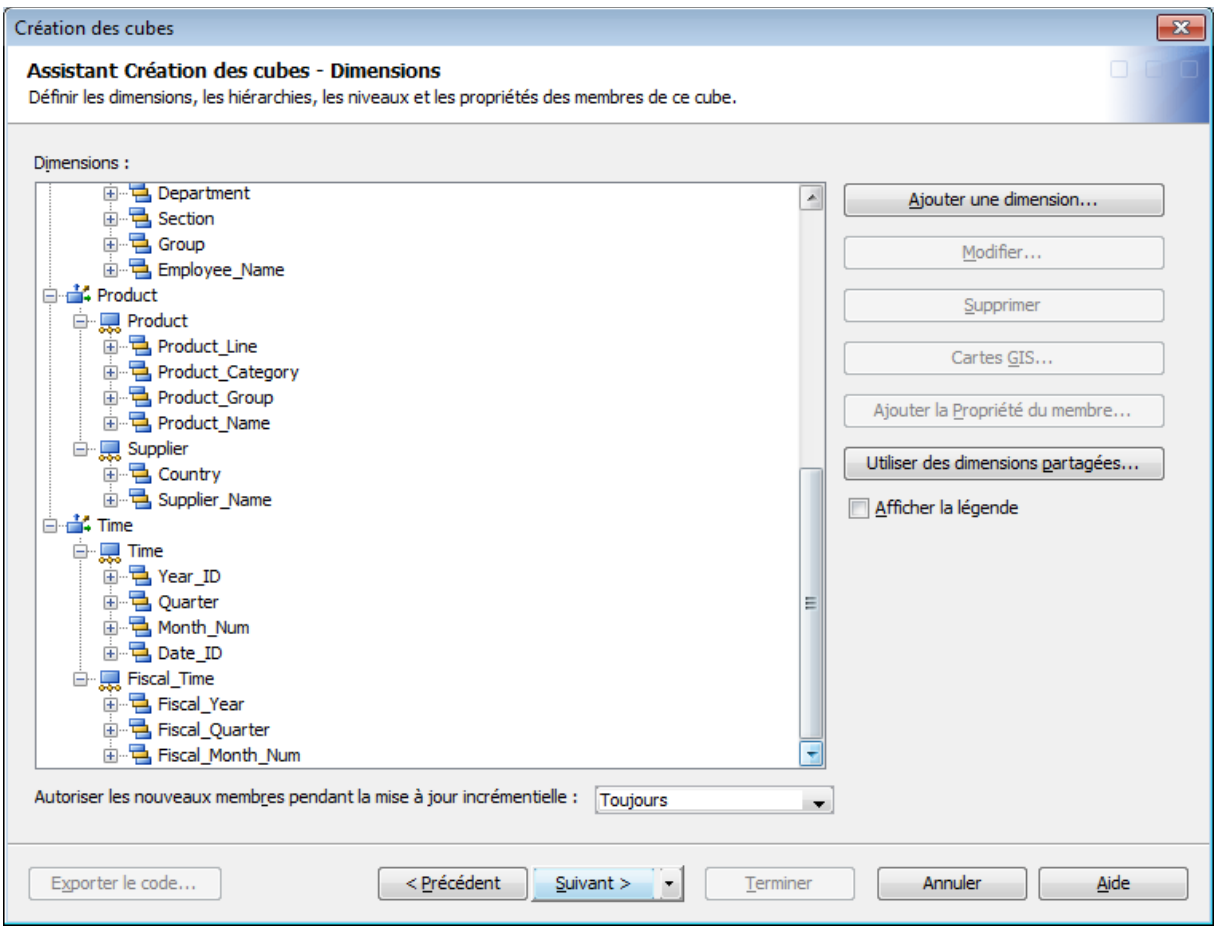
Hiérarchie Organisation : Pays → Compagnie → département → section → Groupe → Collaborateur

Selon la dimension Géographie :

Hiérarchie Géographie: Continent → Pays → Ville

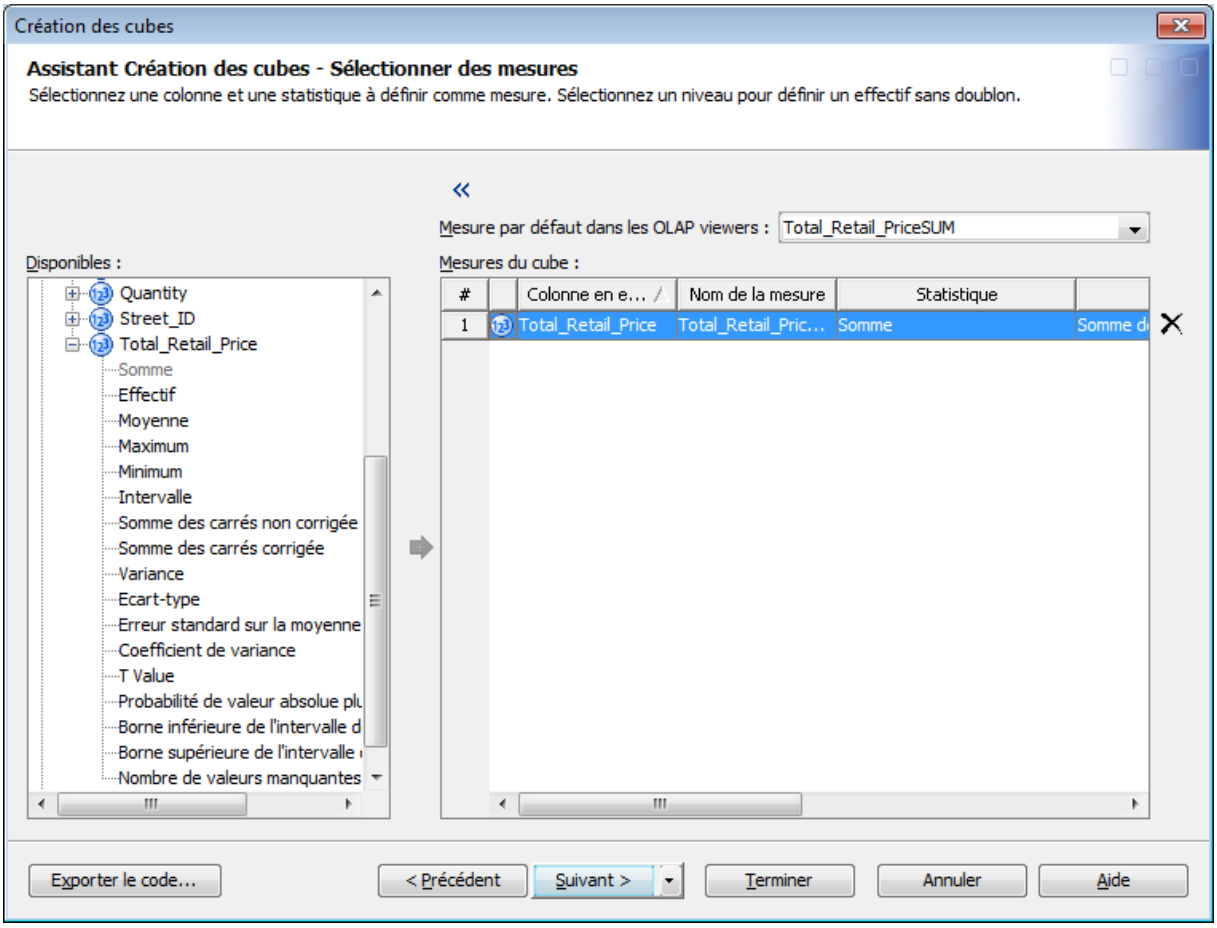



Création des cubes
Assistant Création des cubes - Dimensions
Définir les dimensions, les hiérarchies, les niveaux et les propriétés des membres de ce cube.
Dimensions :
Organization
Organization
Country_2
Company
Department
Section
Group
Employee_Name
Product
Product
Product_Line
Product_Category
Product_Group
Product_Name
Supplier
Country
Supplier_Name
Time
Time
Year_ID
Quarter
Month_Num
Date_ID
Ajouter une dimension...
Modifier...
Supprimer
Cartes GIS...
Ajouter la Propriété du membre...
Utiliser des dimensions partagées...
Afficher la légende
Cliquez sur Ajouter une dimension pour créer une dimension.
Autoriser les nouveaux membres pendant la mise à jour incrémentielle :
Toujours
Exporter le code...
< Précédent
Suivant >
Terminer
Annuler
Aide

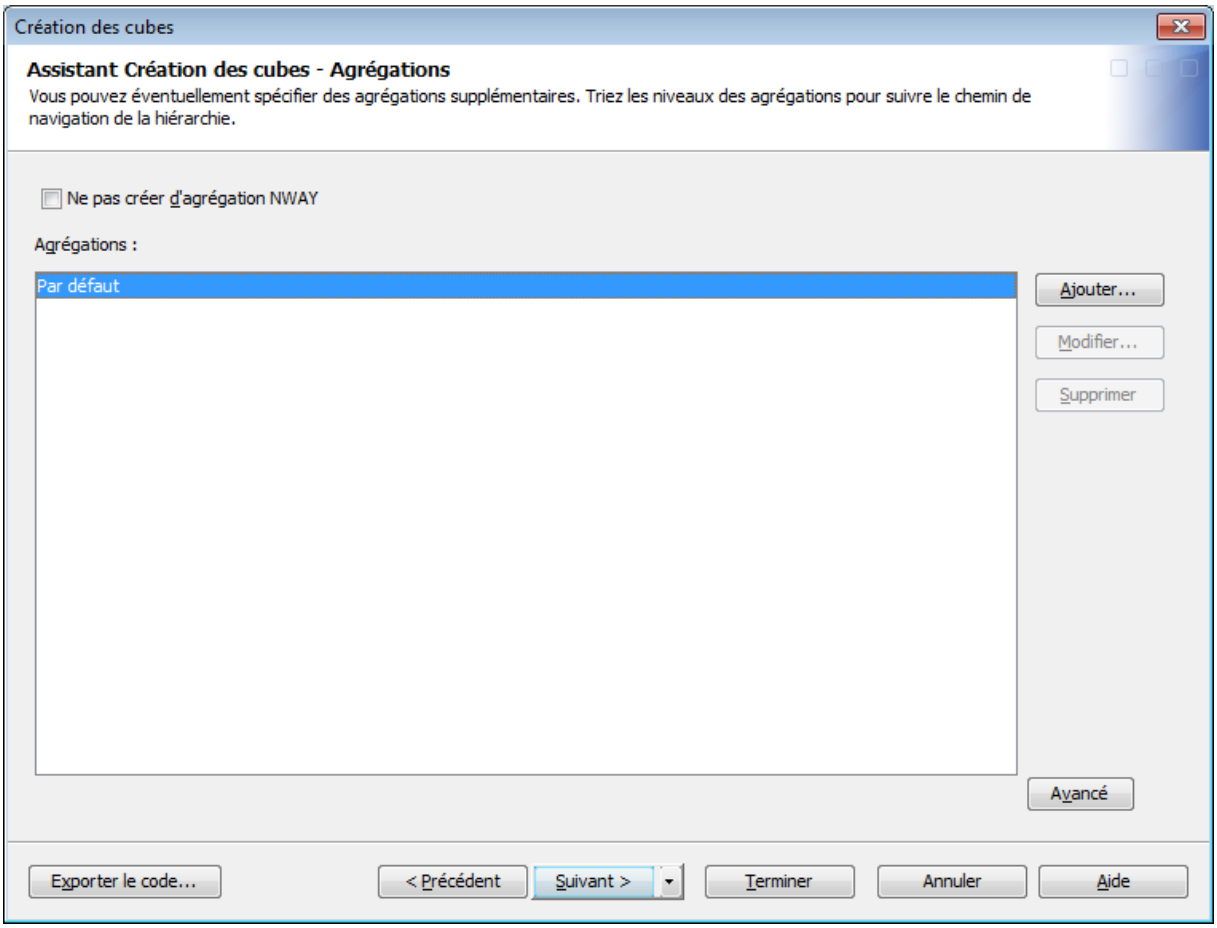

Création des cubes
Assistant Création des cubes - Dimensions
Définir les dimensions, les hiérarchies, les niveaux et les propriétés des membres de ce cube.
Dimensions :
Department
Section
Group
Employee_Name
Product
Product
Product_Line
Product_Category
Product_Group
Product_Name
Supplier
Country
Supplier_Name
Time
Time
Year_ID
Quarter
Month_Num
Date_ID
Fiscal_Time
Fiscal_Year
Fiscal_Quarter
Fiscal_Month_Num
Ajouter une dimension...
Modifier...
Supprimer
Cartes GIS...
Ajouter la Propriété du membre...
Utiliser des dimensions partagées...
Afficher la légende
Autoriser les nouveaux membres pendant la mise à jour incrémentielle :
Toujours
Exporter le code...
< Précédent
Suivant >
Terminer
Annuler
Aide

Après avoir défini vos dimensions, cliquer sur suivant,

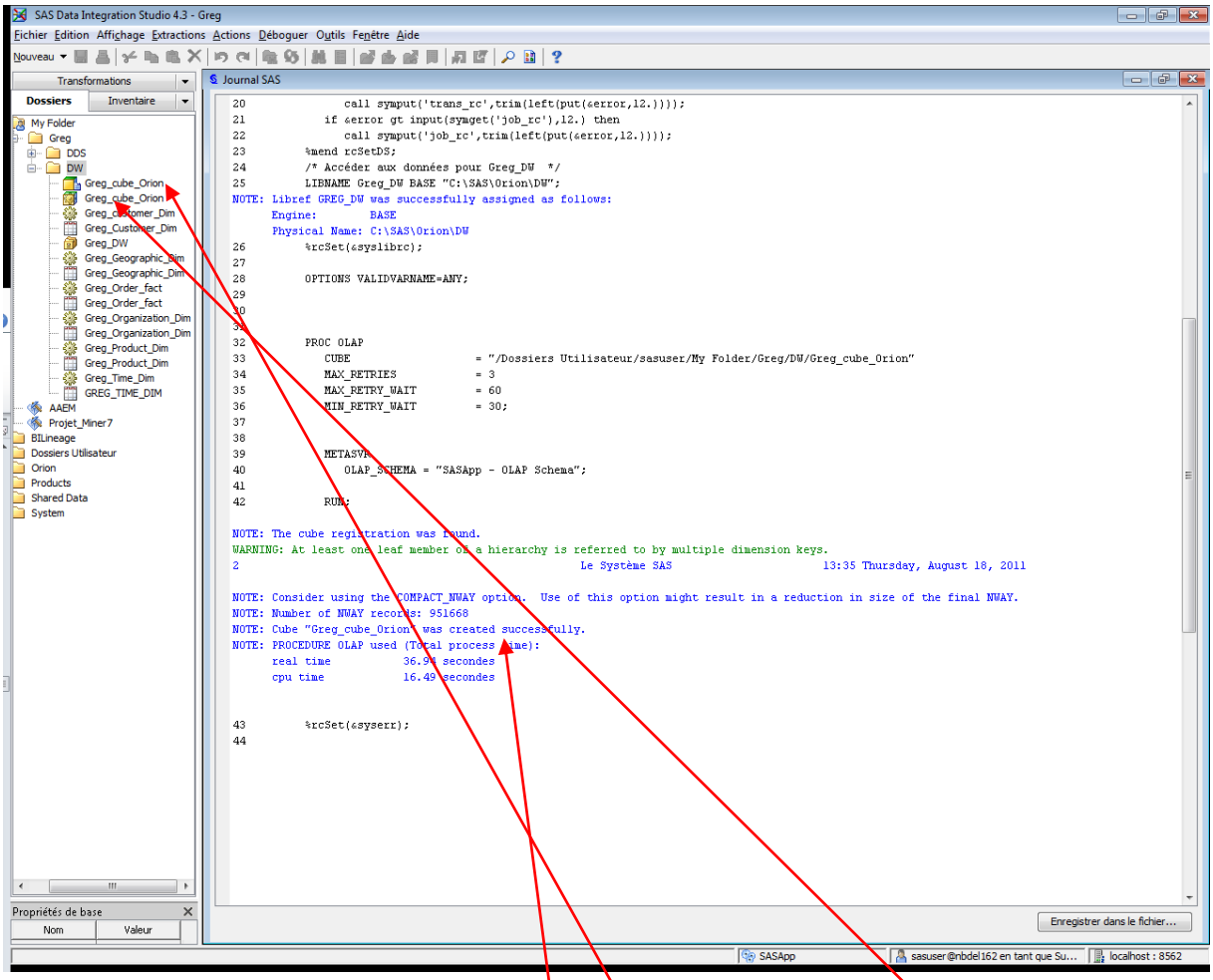
Sélectionner les mesures



Sélectionner la somme du chiffre d'affaire
Suivant



Suivant
Terminer



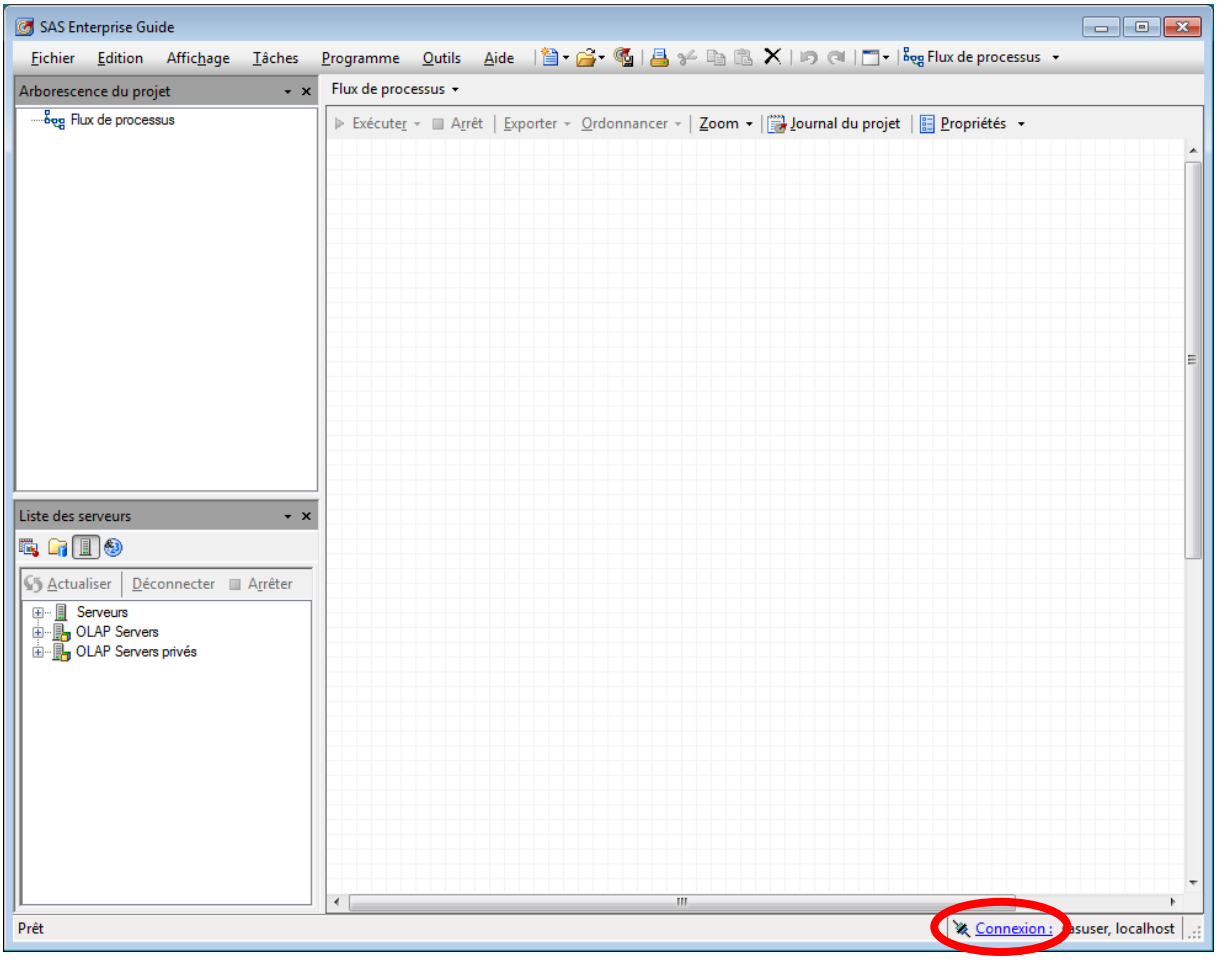
Vous devez obtenir le message que votre cube a été créé « successfully » (avec succès).

Dans l'arborescence des métadonnées, vous obtenez un objet pour le cube et un second avec le même nom et presque le même icône du cube, mais avec une roue dentée de mécanisme devant, pour spécifier que c'est le processus qui génère le cube. Pour ordonnancer le processus, il faudra faire un clic-droit dessus.

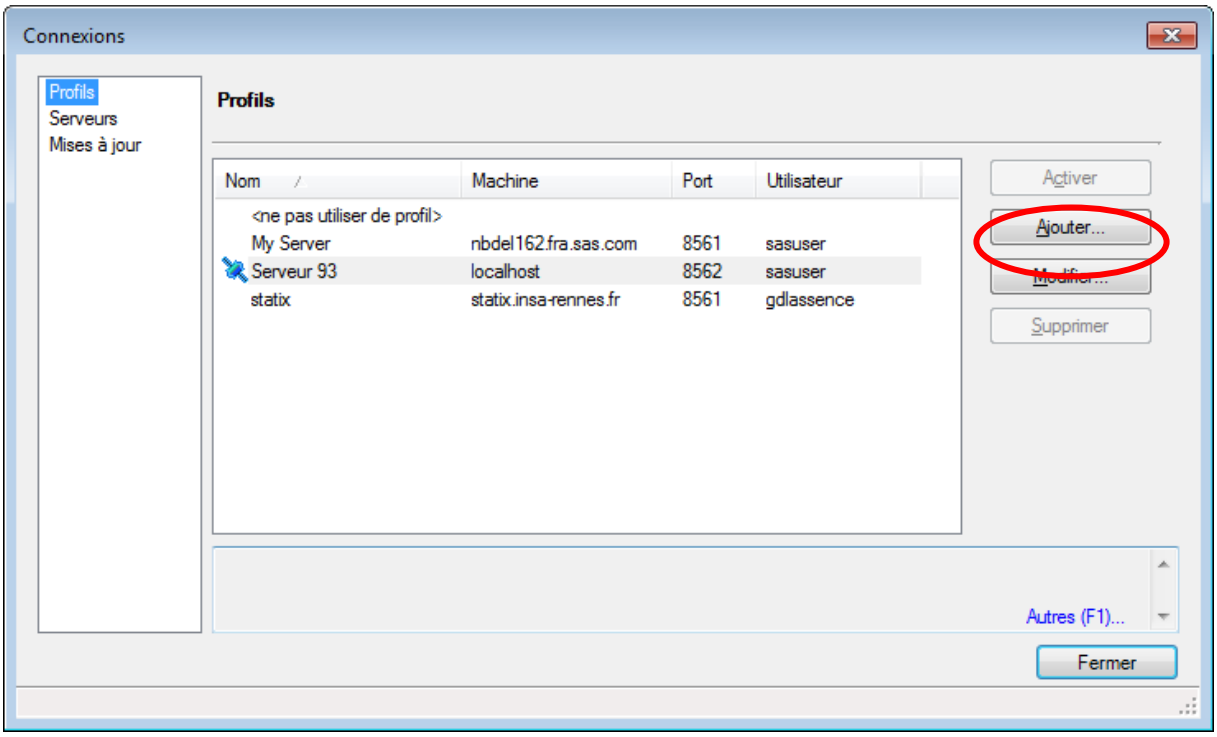
Si votre cube ne s'est pas créer correctement, Clic-droit sur celui-ci → Modifier sa structure.

Navigation dans un cube depuis SAS Enterprise Guide :

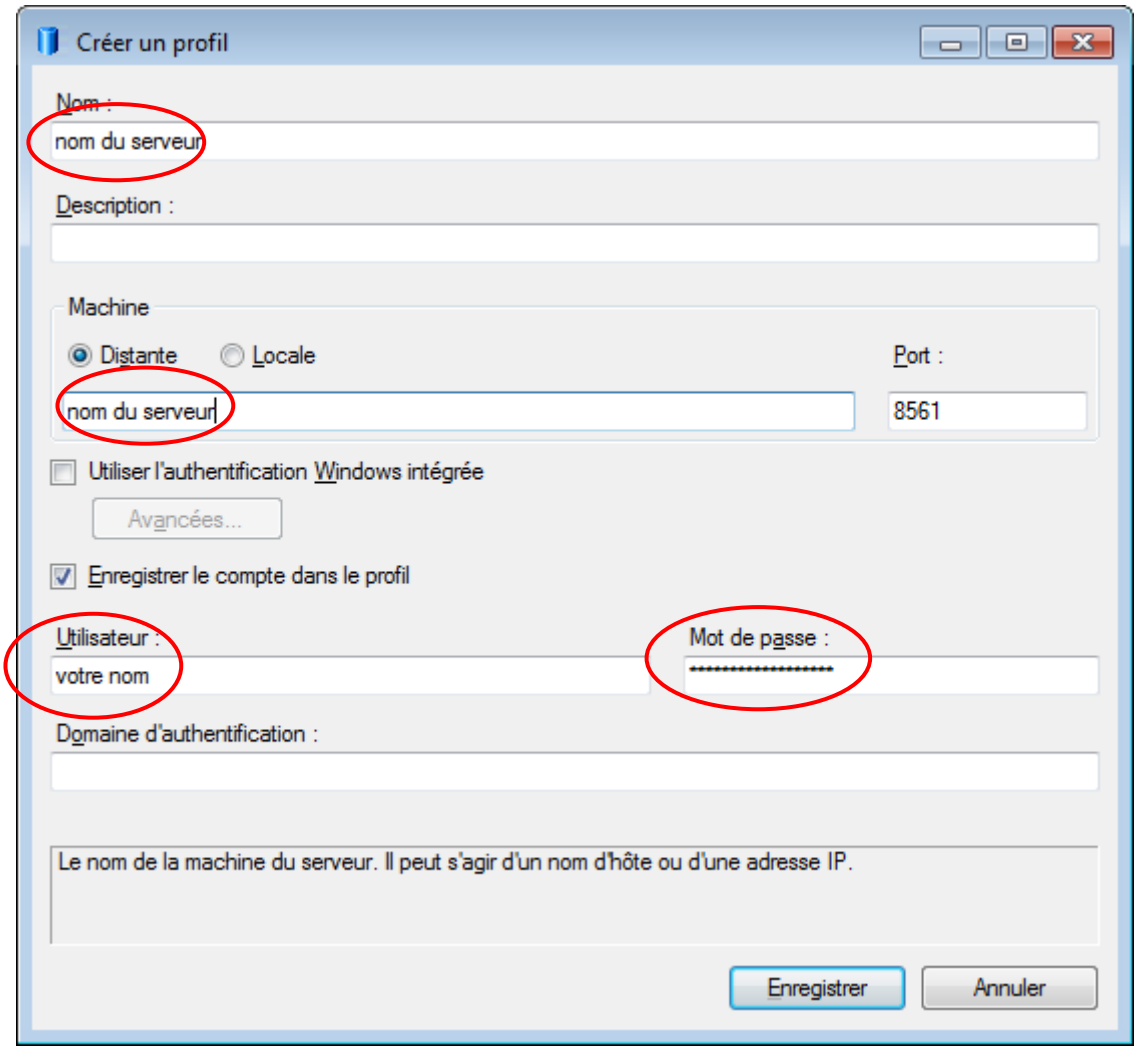
Ouvrir SAS Enterprise Guide
Depuis Démarrer → Programmes → SAS → Enterprise Guide



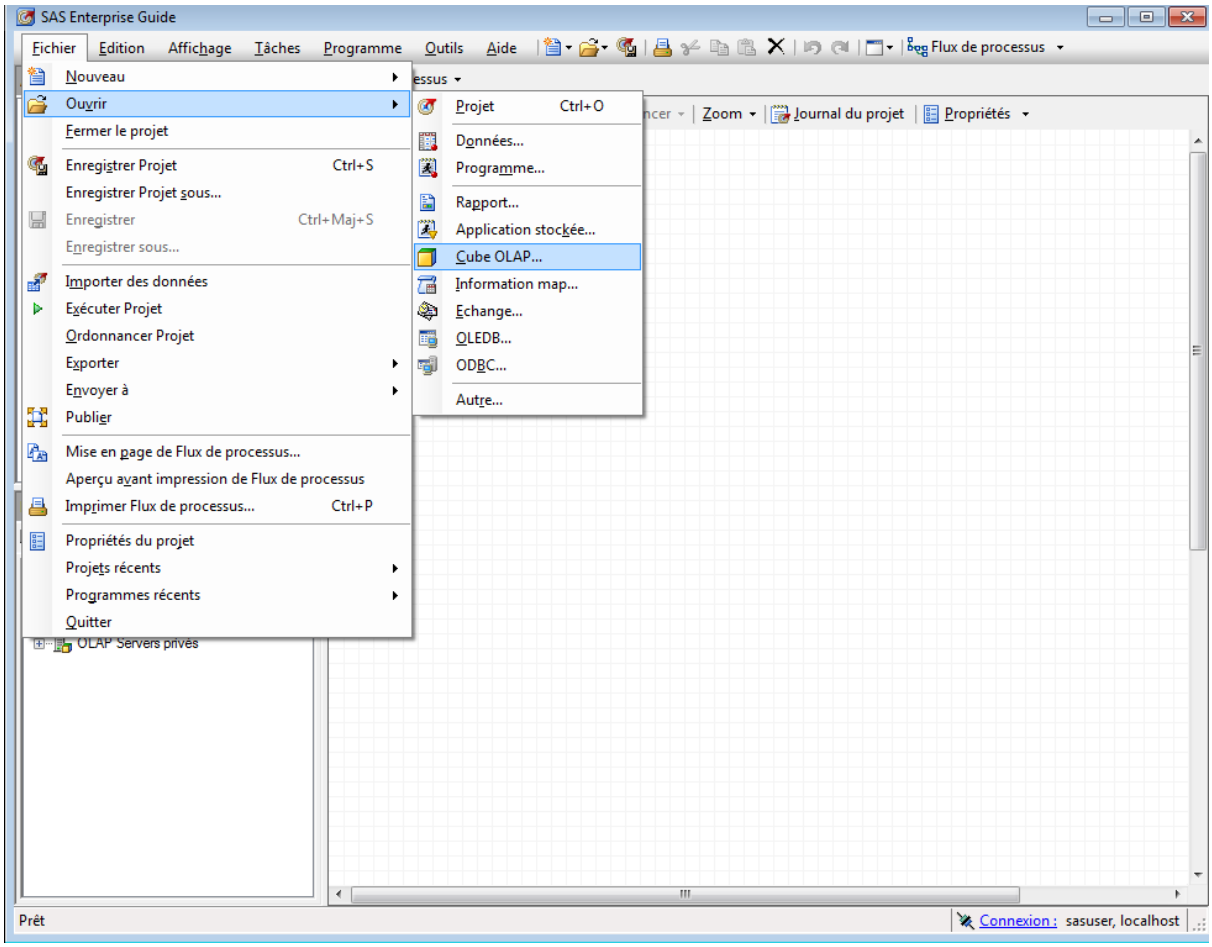
Il faut que vous soyez connecté au serveur de métadonnées. Si vous n'avez pas connexion en bas à droite de la fenêtre d'Enterprise Guide, Cliquer sur « pas de connexion ».



S'il n'y a pas de connexion, ajoutez-en une :

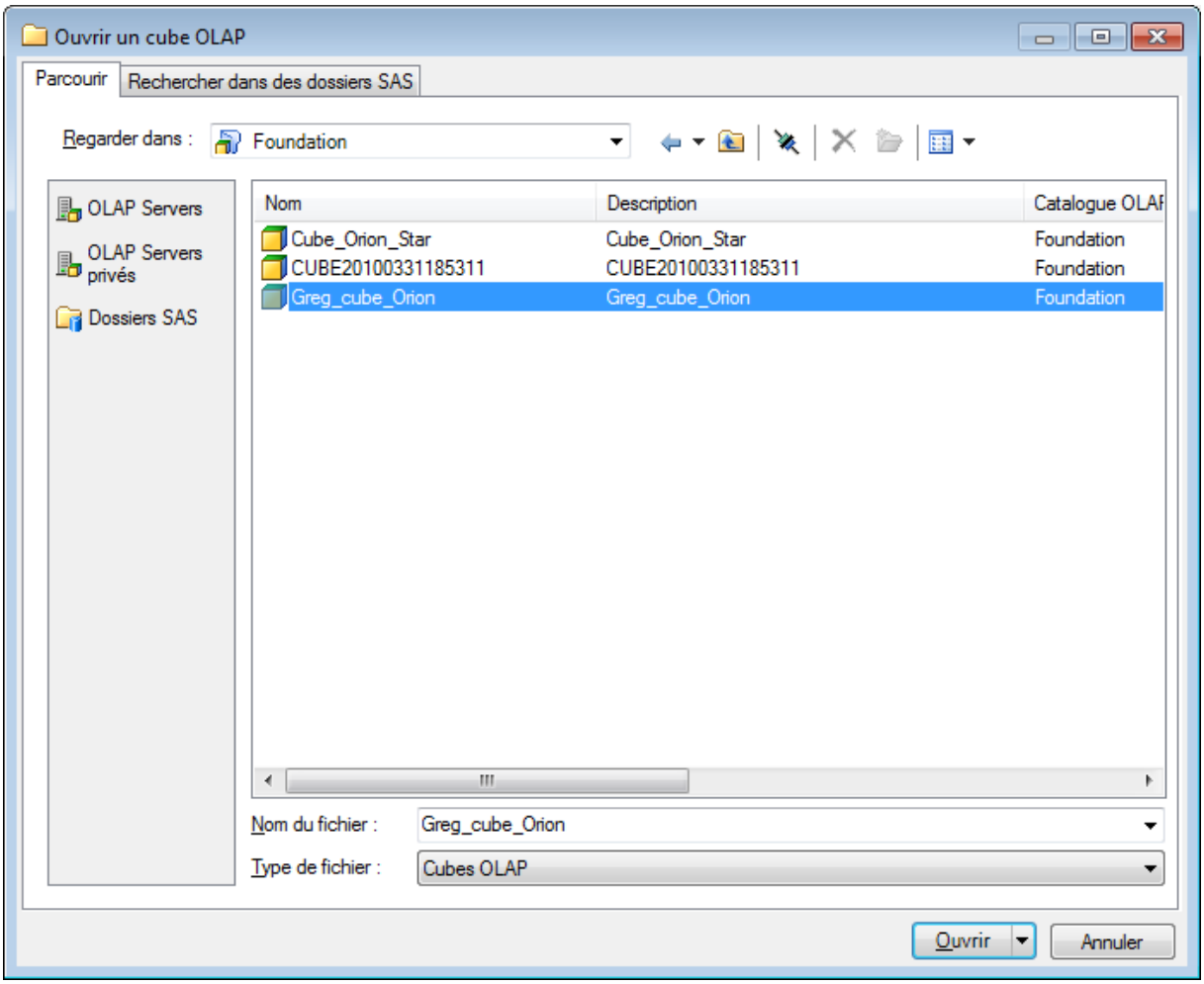


Il faut rentrer un nom de connexion, le nom du serveur, l'utilisateur et son mot de passe. Enregistrer. Activer la connexion si besoin.

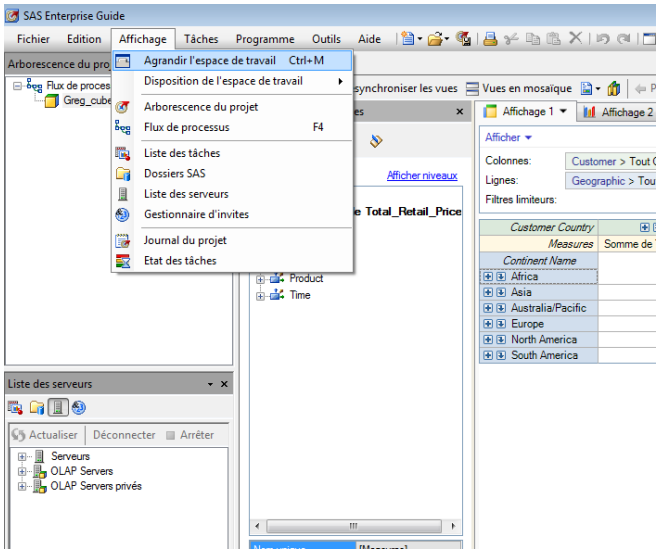


Dans le menu **Fichier → Ouvrir → Cube OLAP**

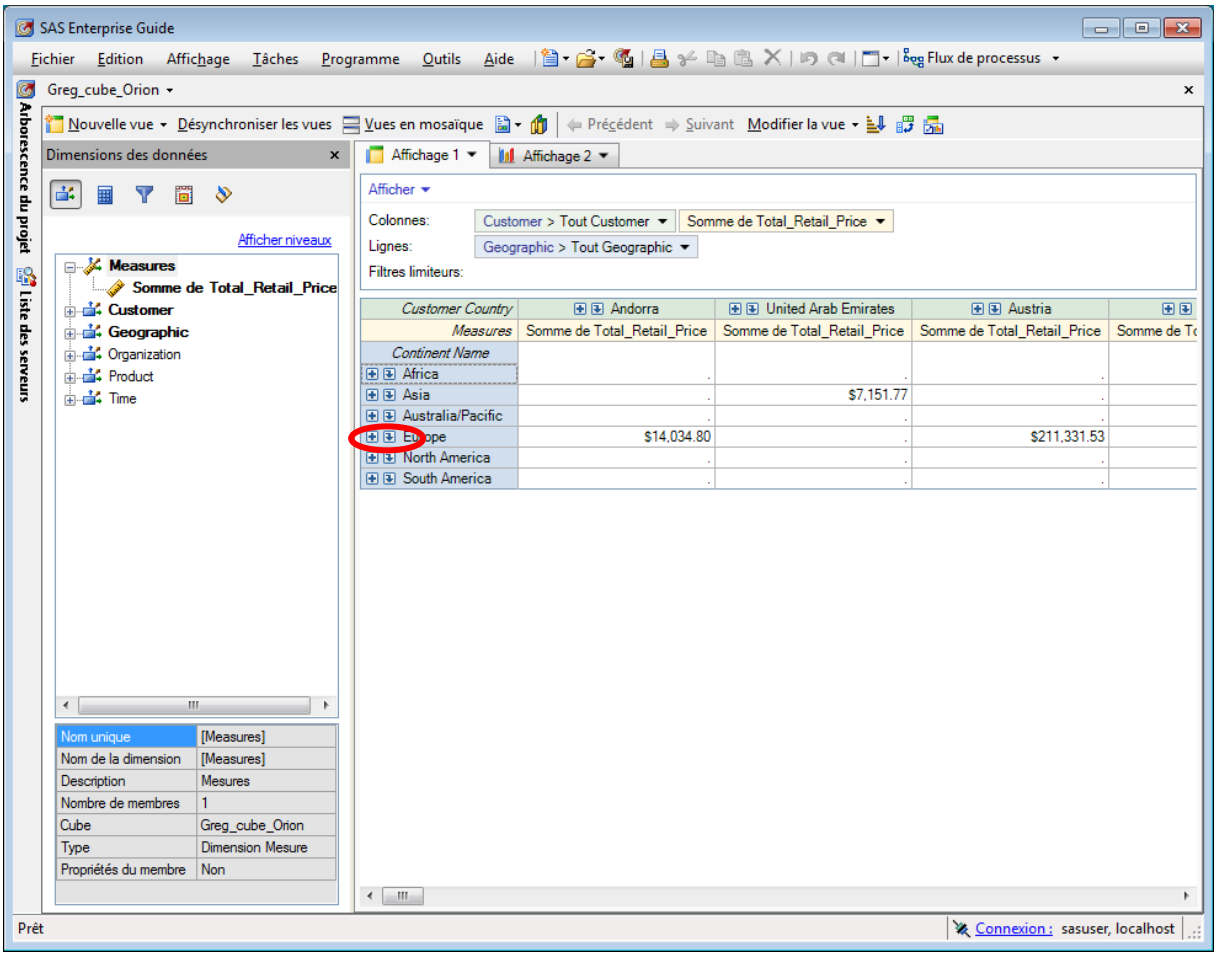
OLAP servers → SASApp → Foundation →



Sélectionner votre cube.
Ouvrir,

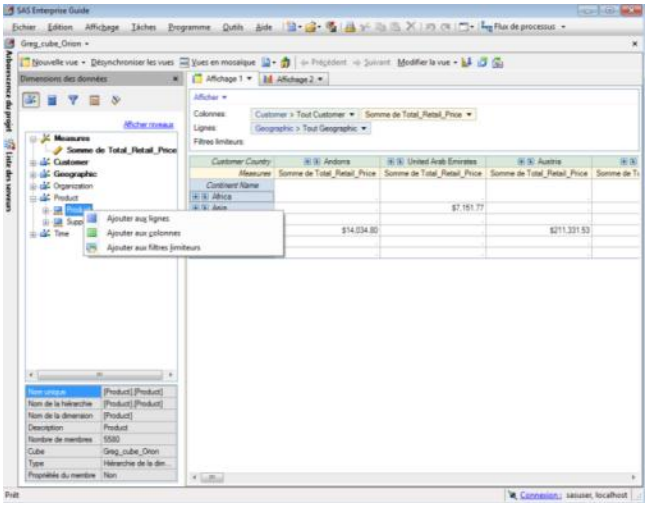


Il peut être agréable de maximiser l'espace de travail, depuis le menu affichage.

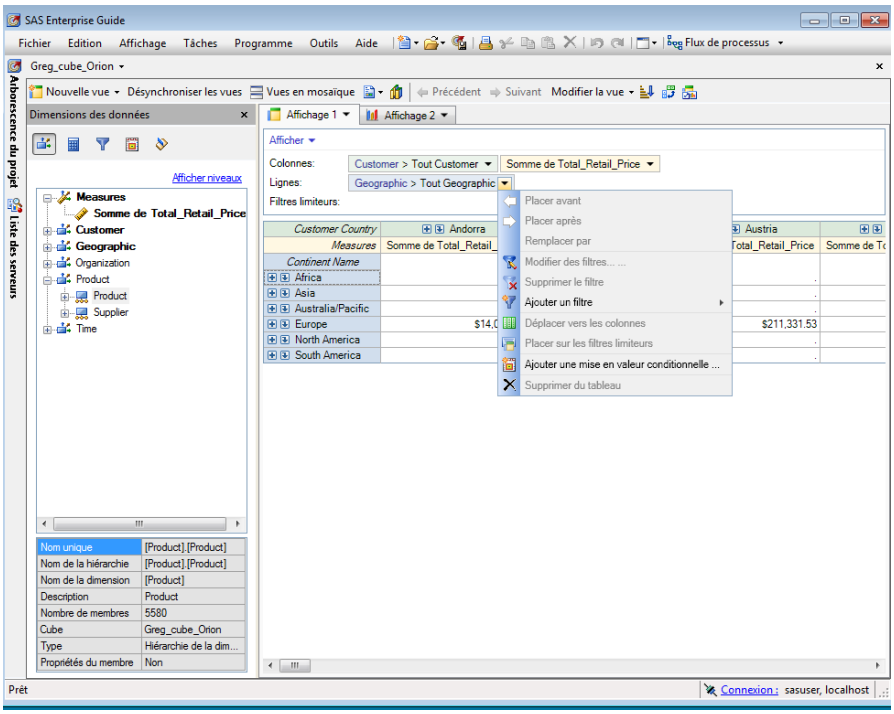


Pour naviguer du général au détail (Drill down), vous avez :

-  la croix permet de voir les niveaux en dessous et laisse les autres niveaux actuels inchangés.
-  la flèche vers le bas permet d'afficher les niveaux en dessous et supprime l'affichage des autres niveaux actuels.

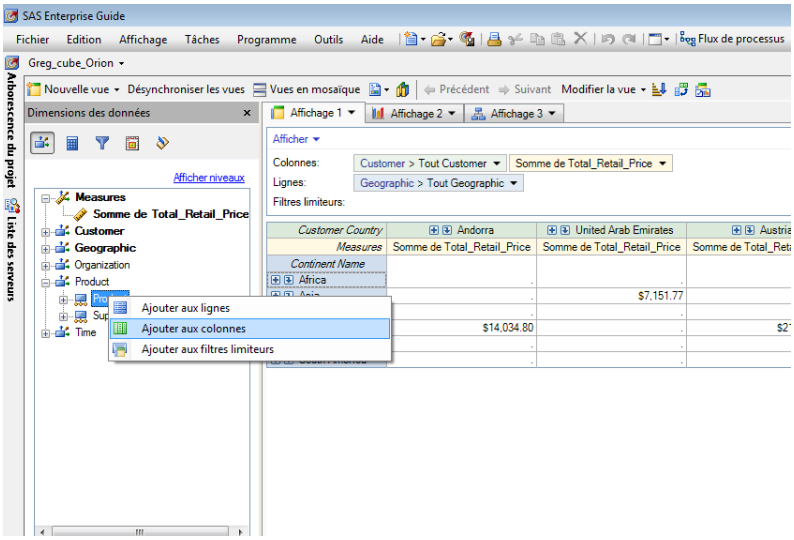


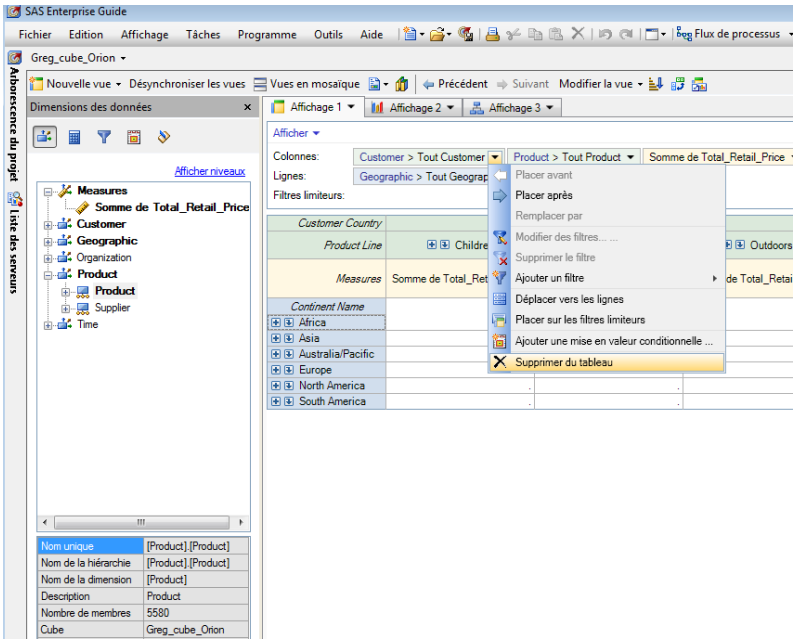
Sélectionner une hiérarchie pour l'ajouter en ligne, en colonne ou l'utiliser en filtre.



Sélectionner un élément en ligne, ou en colonne pour le changer de place, le modifier, etc..

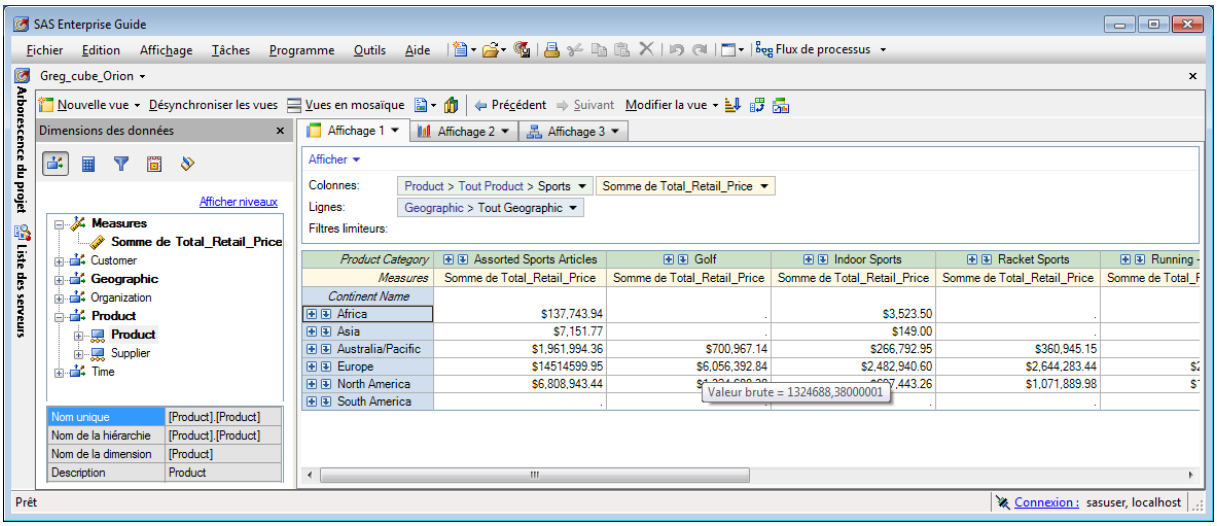
Ajouter les produits en colonne

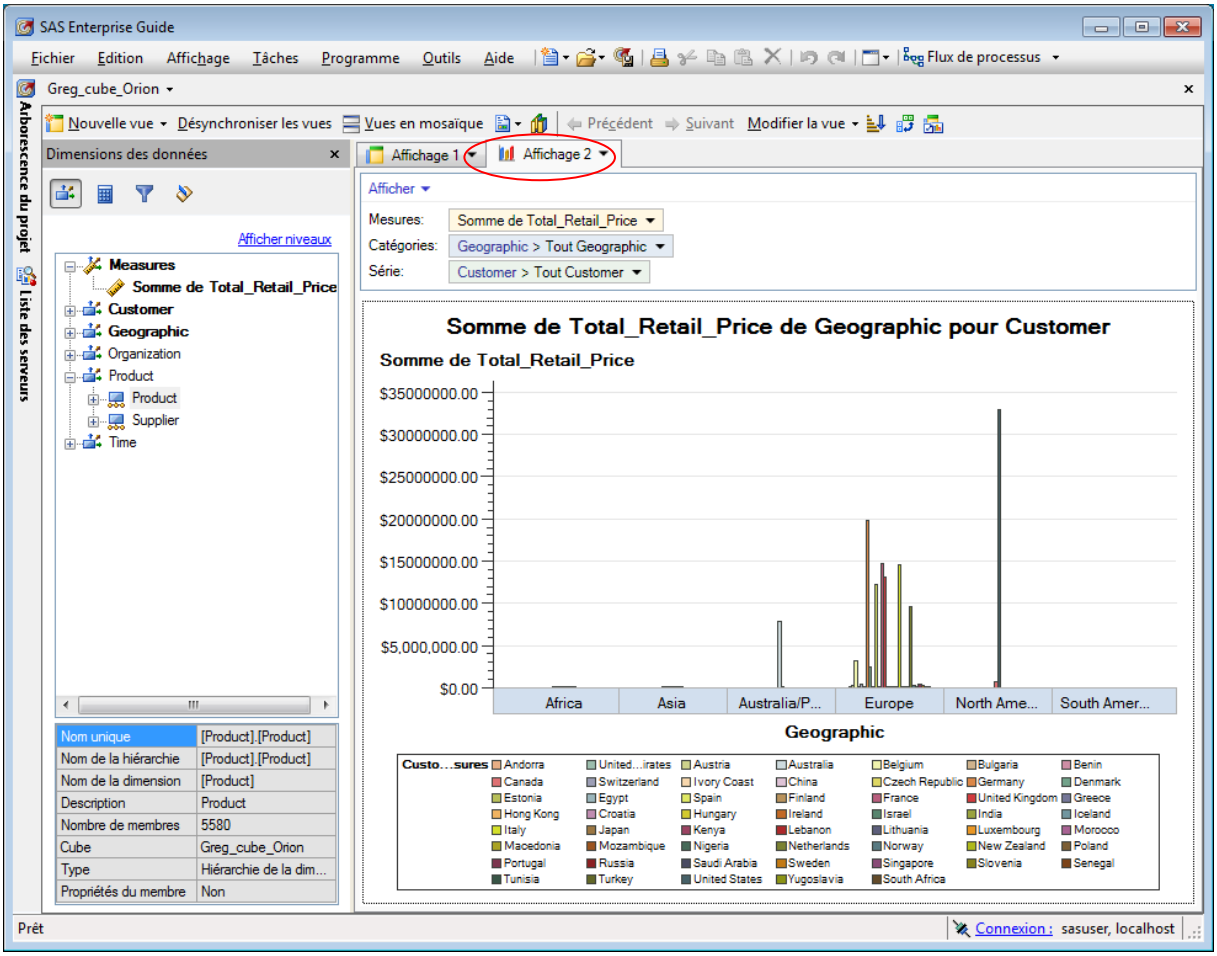




Supprimer la hiérarchie client du tableau

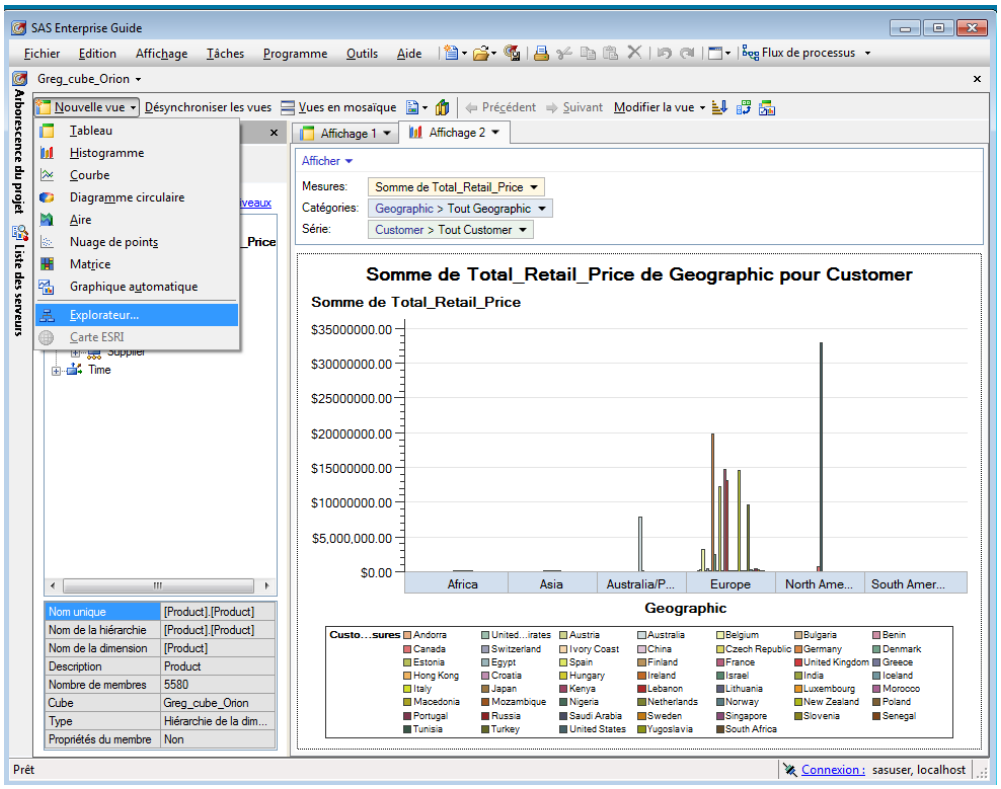
Déplier les articles de sport en cliquant sur la flèche vers le bas

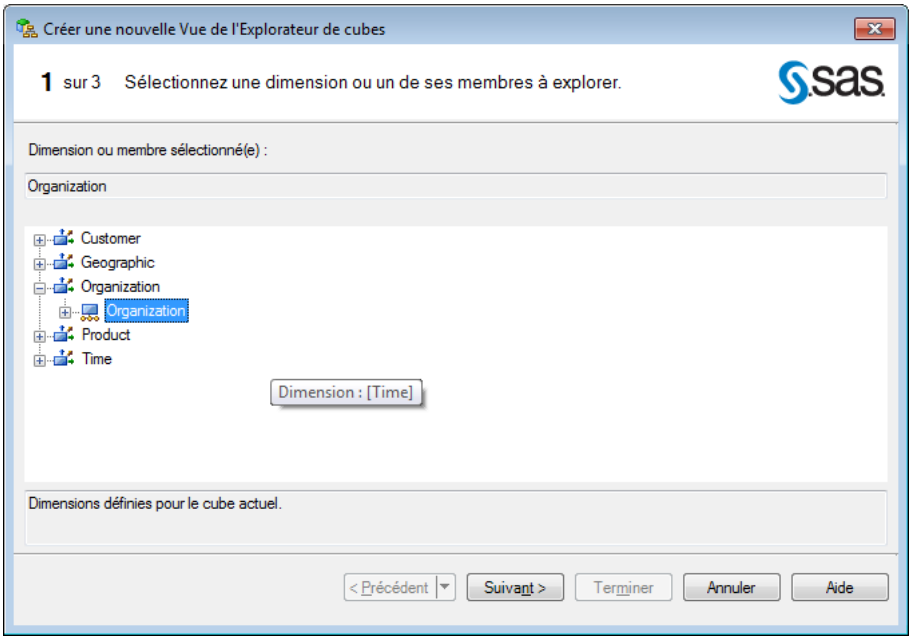




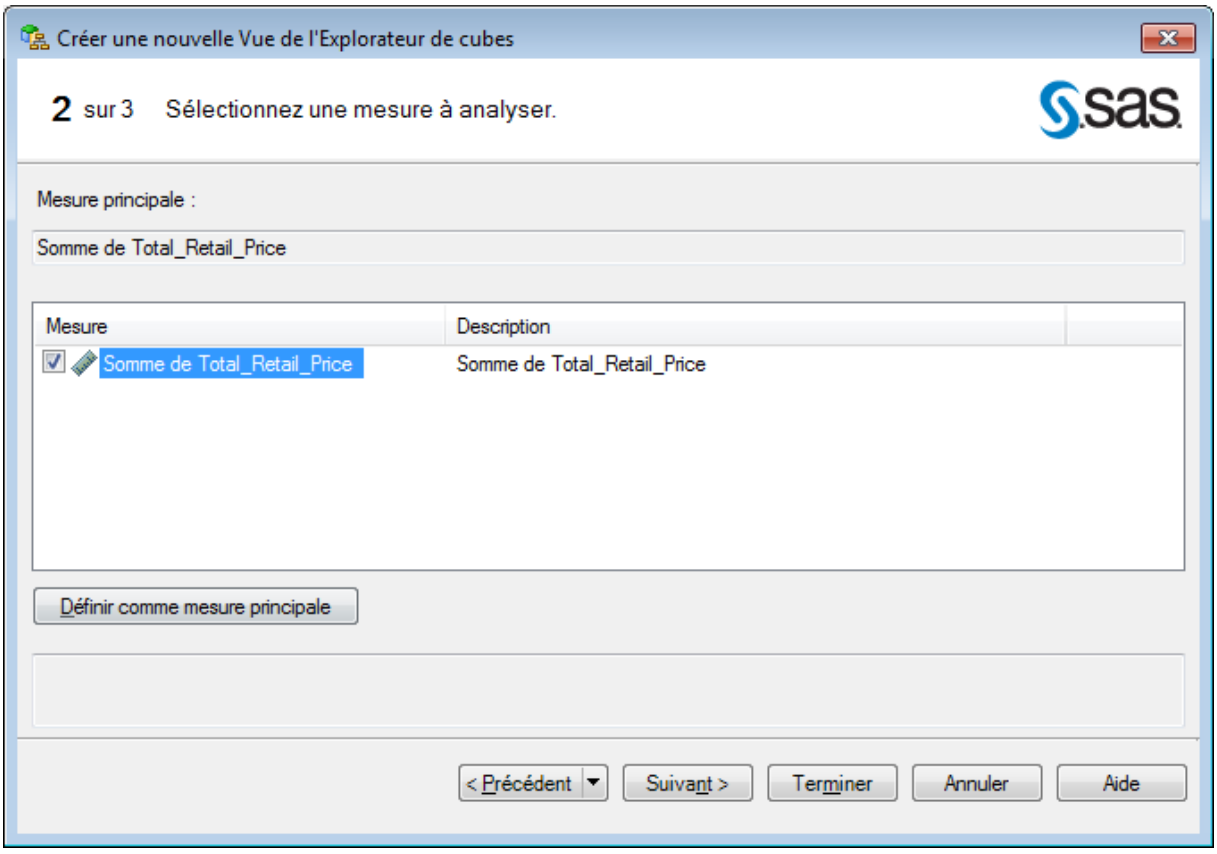
Dans le second onglet d'affichage se trouve par défaut la vue en histogramme.

Pour créer une nouvelle vue, aller dans le menu du même nom.

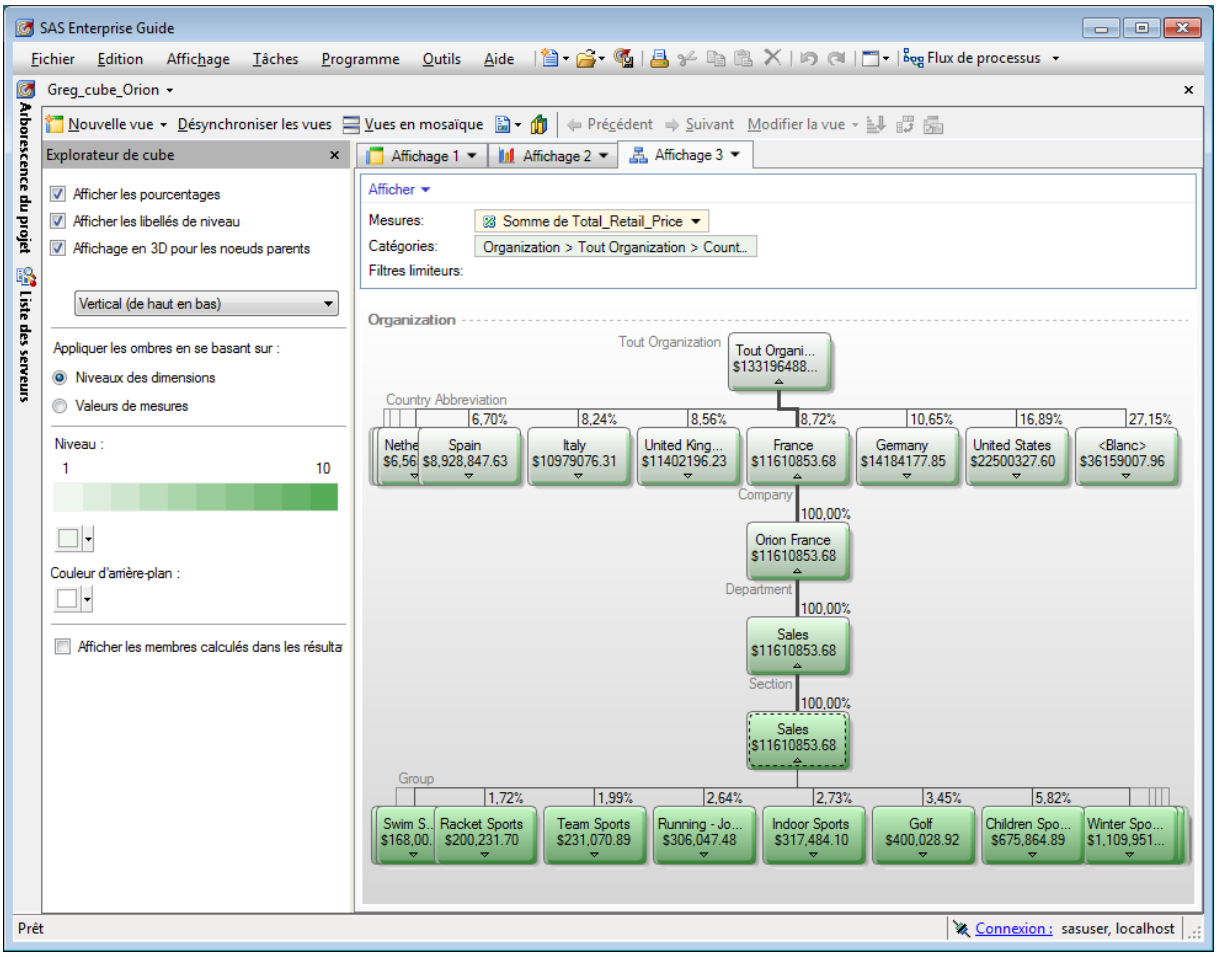




Sélectionner une hiérarchie, par exemple l'organisation,
Suivant, sélectionner la somme du chiffre d'affaires, Suivant, Suivant, Terminer

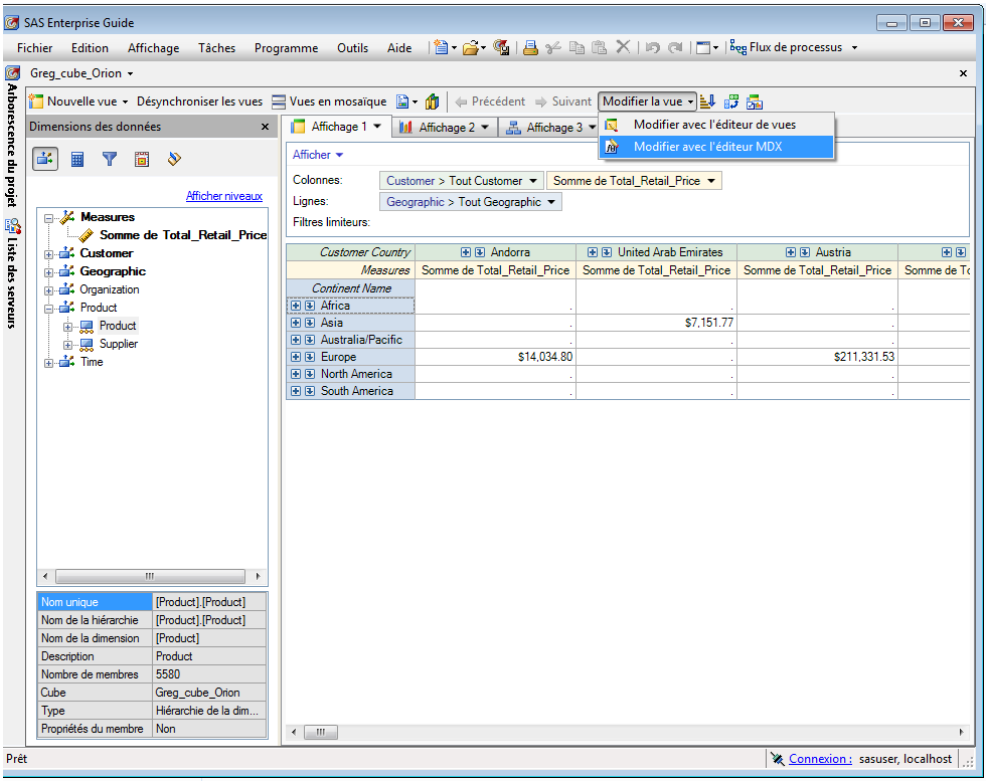

Créer une nouvelle Vue de l'Explorateur de cubes
2 sur 3 Sélectionnez une mesure à analyser.
sas
Mesure principale :
Somme de Total_Retail_Price
Mesure
Description
Somme de Total_Retail_Price
Somme de Total_Retail_Price
Définir comme mesure principale
< Précédent
Suivant >
Terminer
Annuler
Aide

Terminer

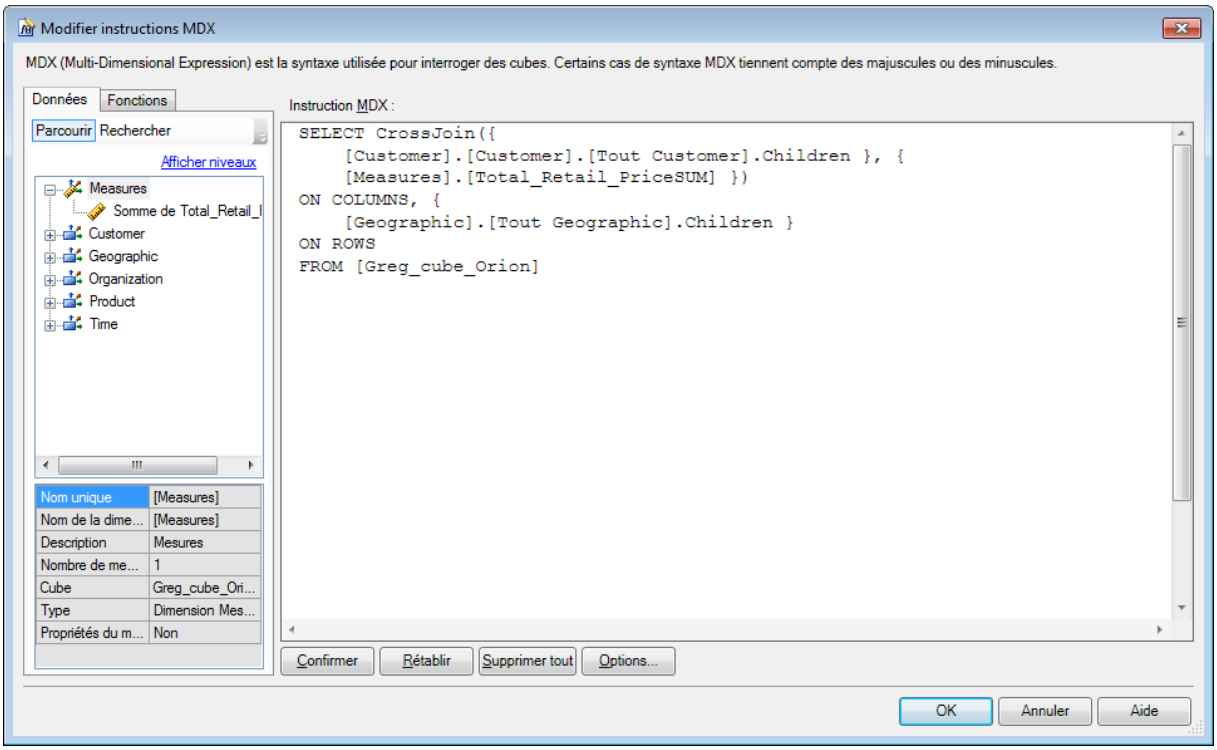


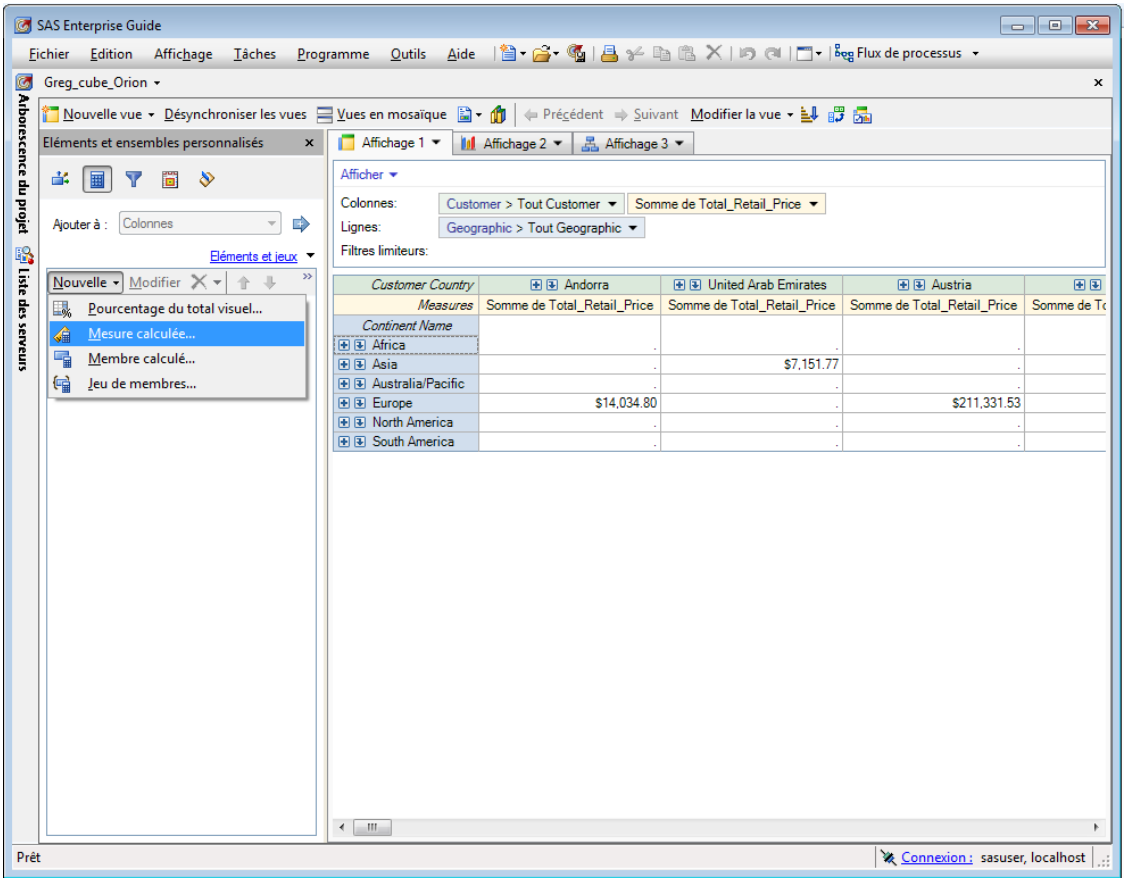
Cliquer sur les éléments pour les déplier ou les replier.

Vous pouvez aussi modifier une vue avec l'éditeur de code MDX :

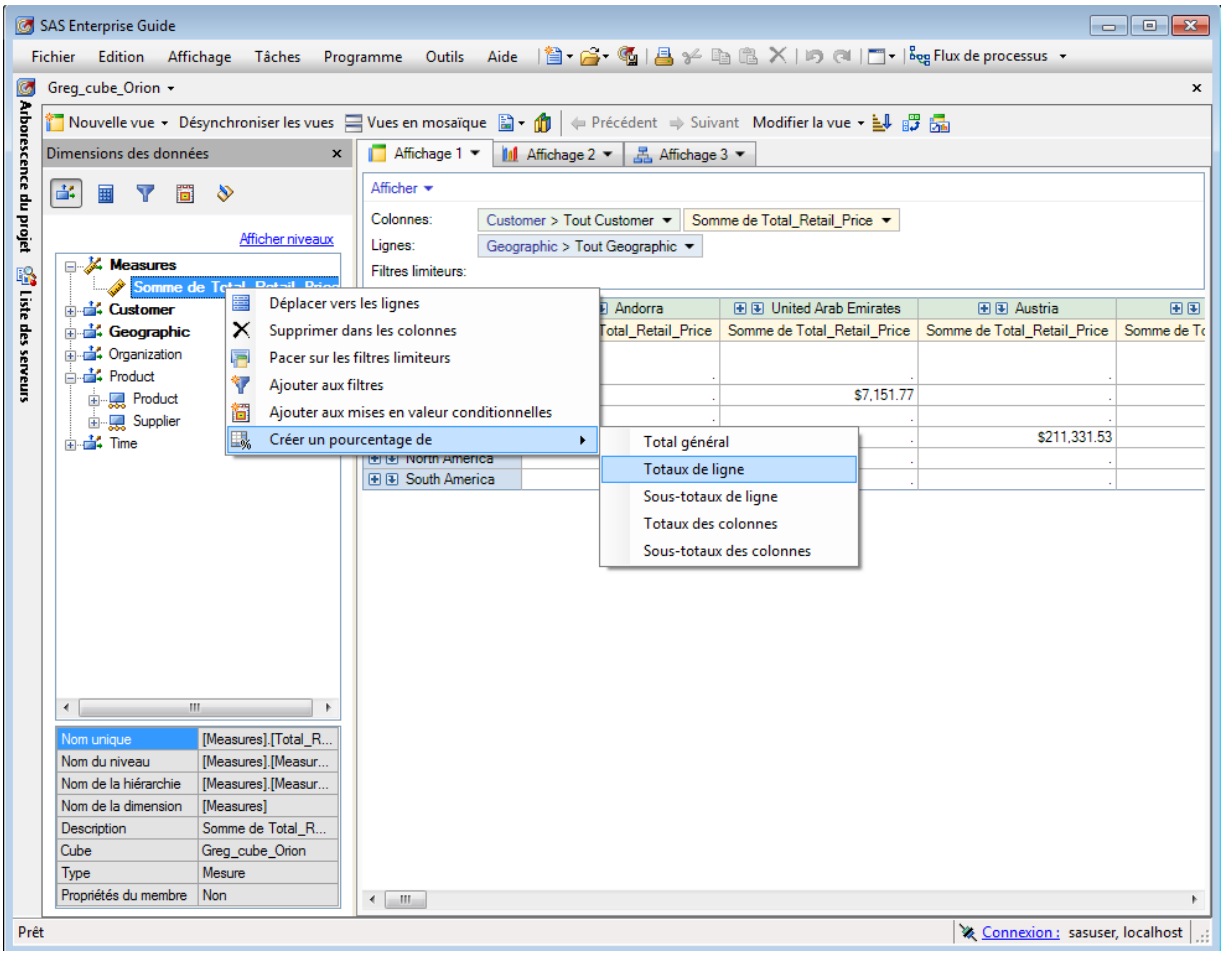


En cliquant sur le bouton **Editeur MDX** vous avez accès au code MDX qui est généré.





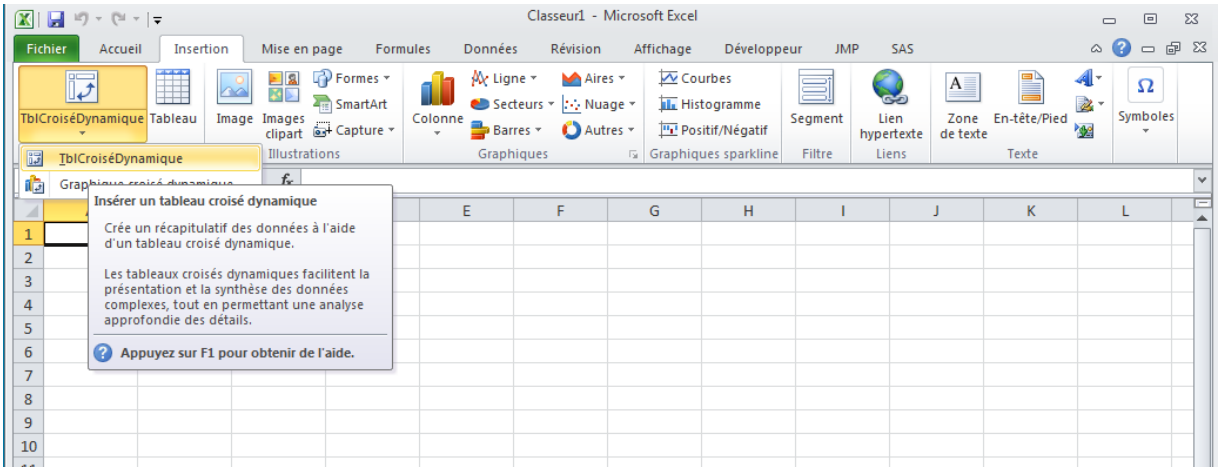
Ou créer des mesures calculées.



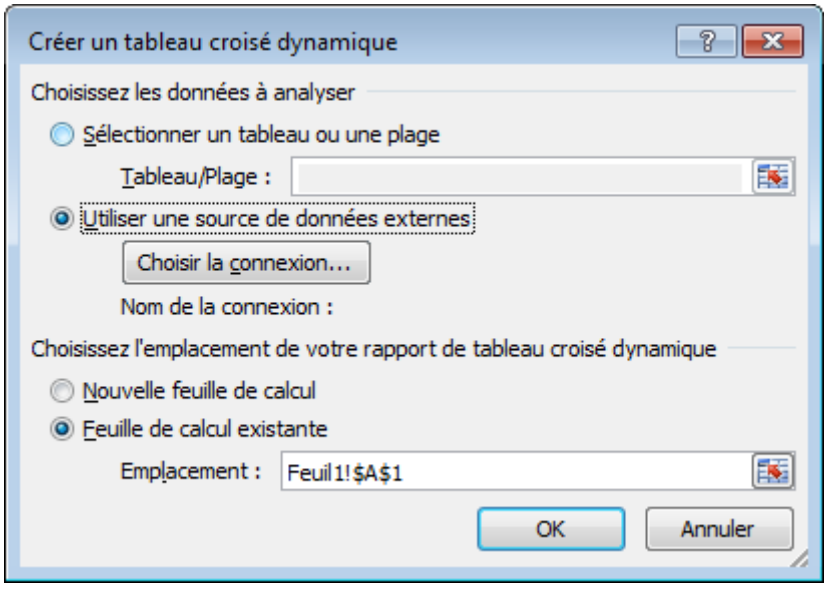
Ajouter des pourcentages en ligne ou en colonne,

Navigation dans un cube depuis Microsoft Excel :

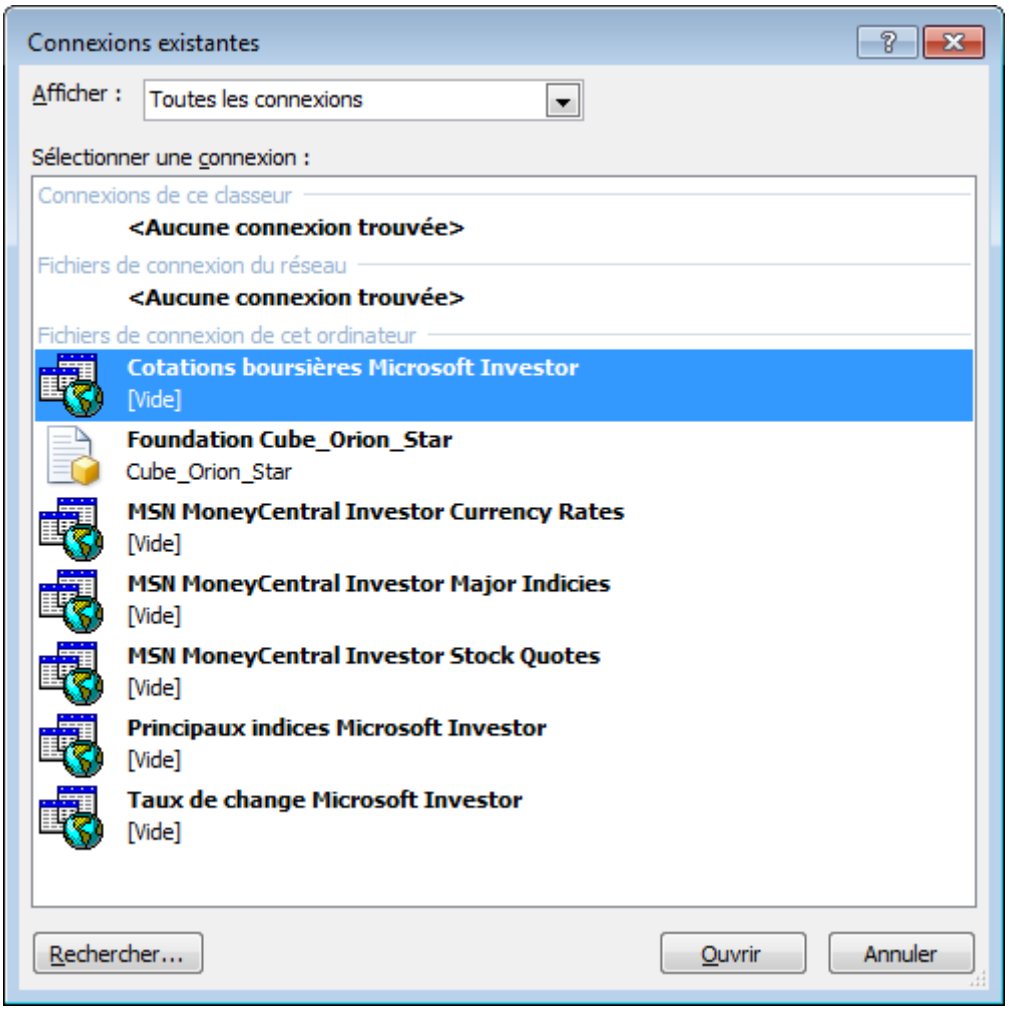
Ouvrir une feuille Excel :



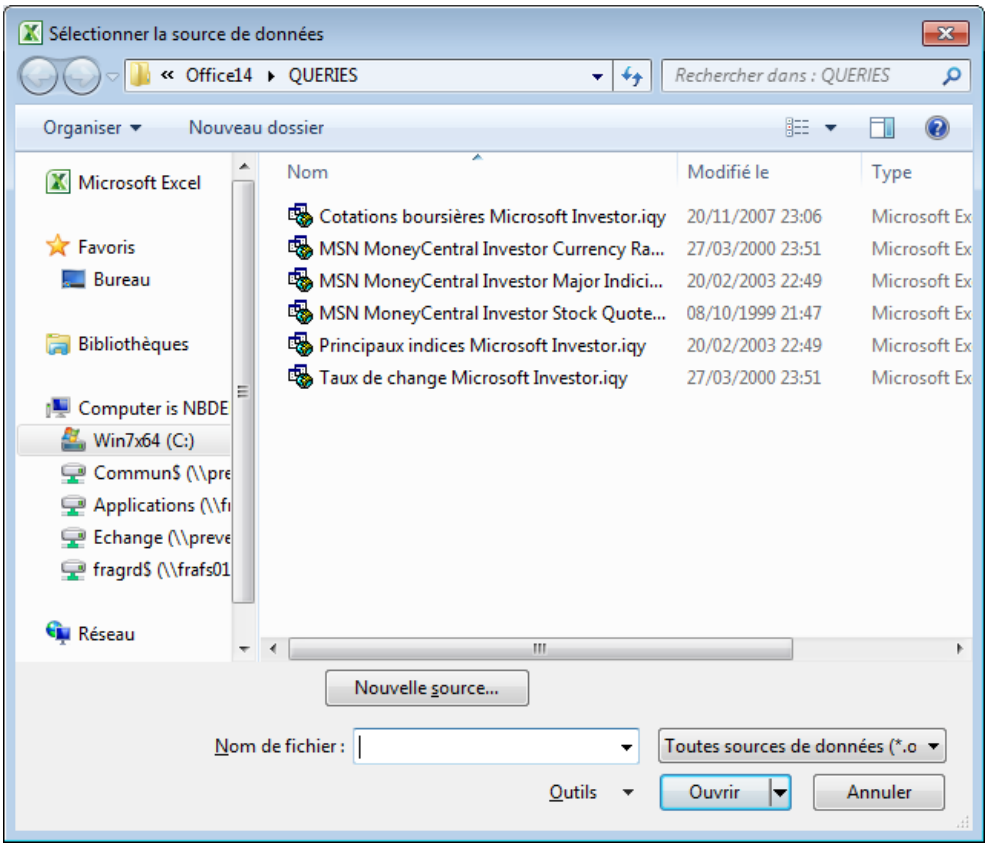
Dans l'onglet Insertion → Ouvrir des données dans tableau croisé dynamique



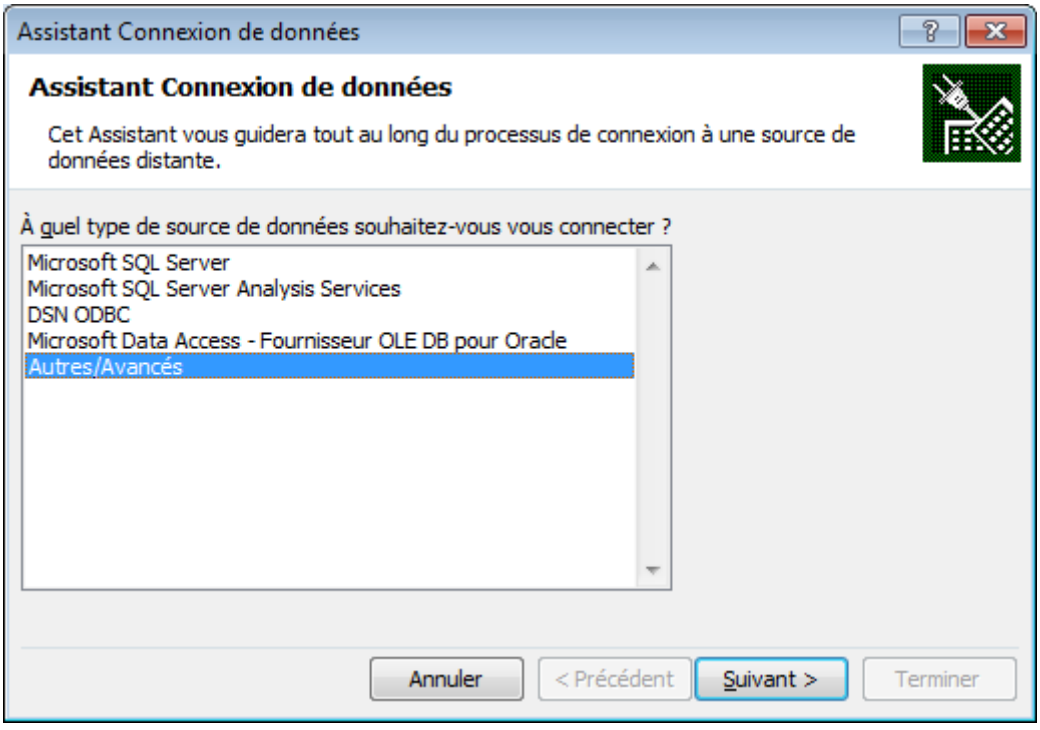
Utiliser une source de données externes → Choisir la connexion


Connexions existantes
Afficher :
Toutes les connexions
Sélectionner une connexion :
Connexions de ce classeur
<Aucune connexion trouvée>
Fichiers de connexion du réseau
<Aucune connexion trouvée>
Fichiers de connexion de cet ordinateur
Cotations boursières Microsoft Investor
[Vide]
Foundation Cube_Orion_Star
Cube_Orion_Star
MSN MoneyCentral Investor Currency Rates
[Vide]
MSN MoneyCentral Investor Major Indicies
[Vide]
MSN MoneyCentral Investor Stock Quotes
[Vide]
Principaux indices Microsoft Investor
[Vide]
Taux de change Microsoft Investor
[Vide]
Rechercher...
Ouvrir
Annuler

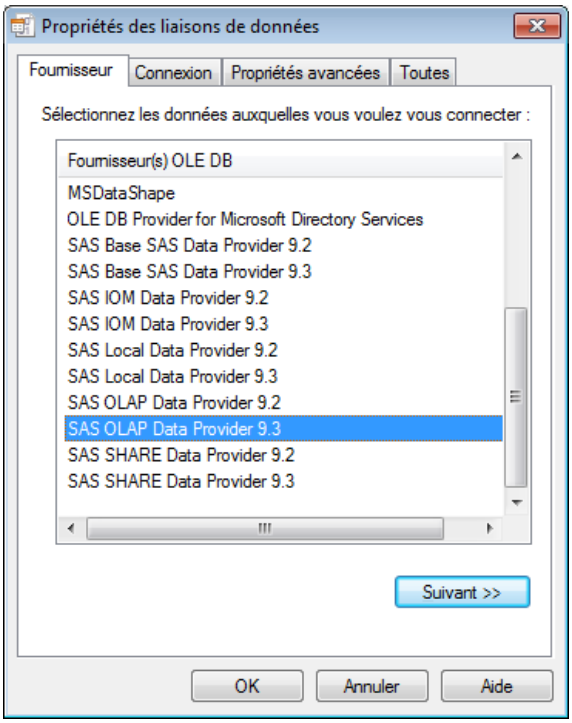
Rechercher



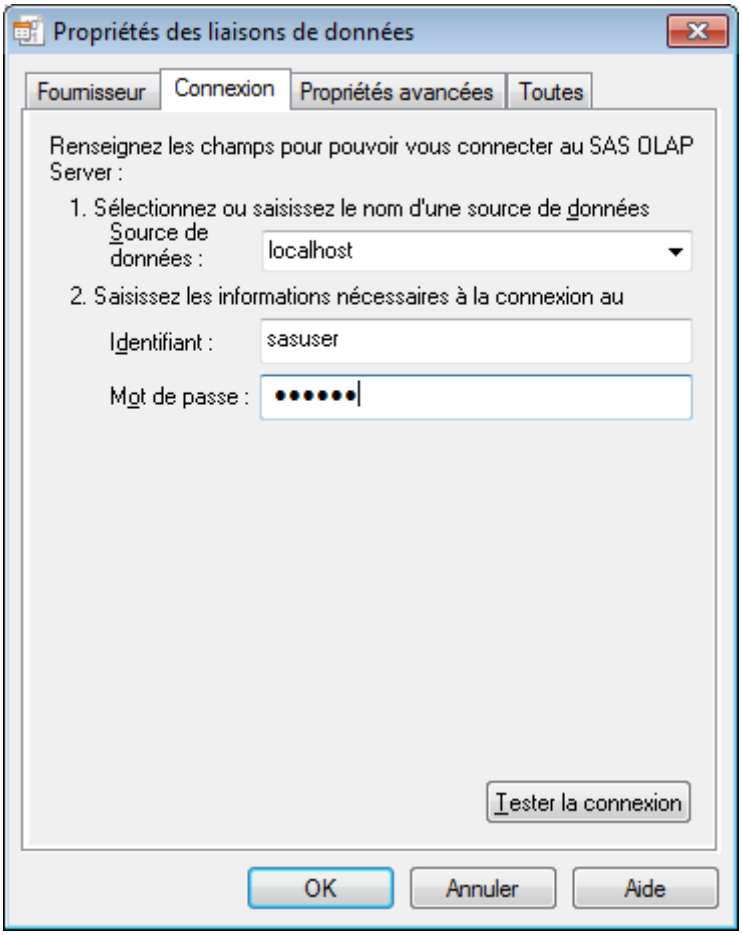
Nouvelle source



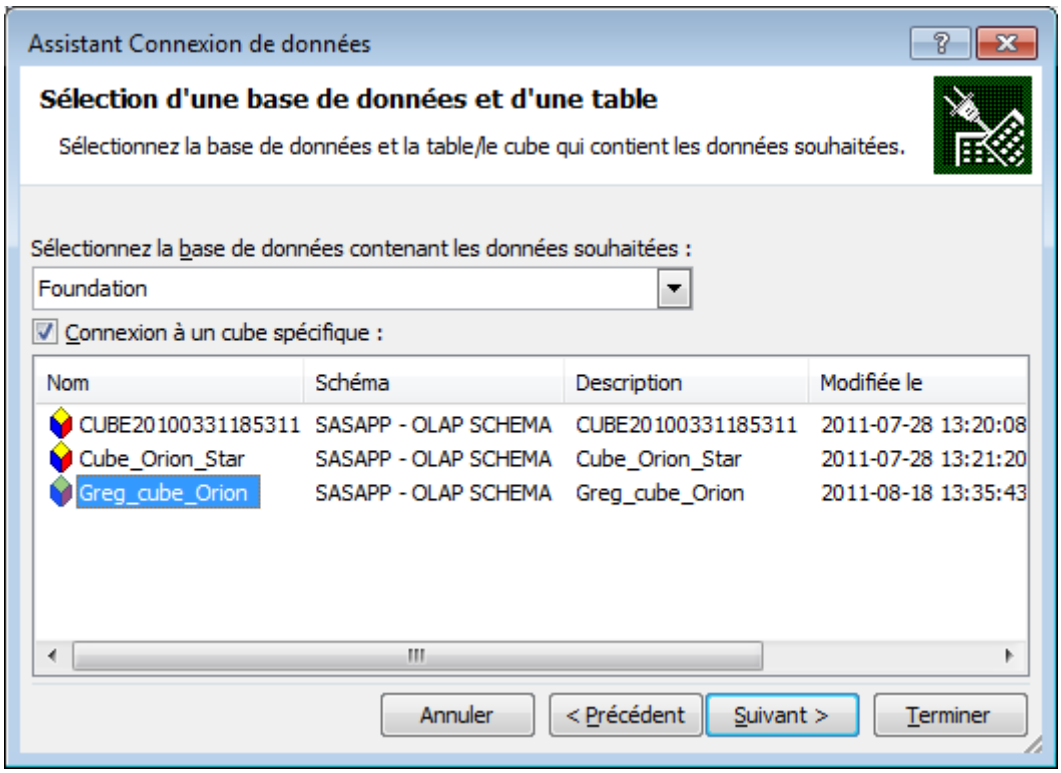
Autre/Avancés
Suivant



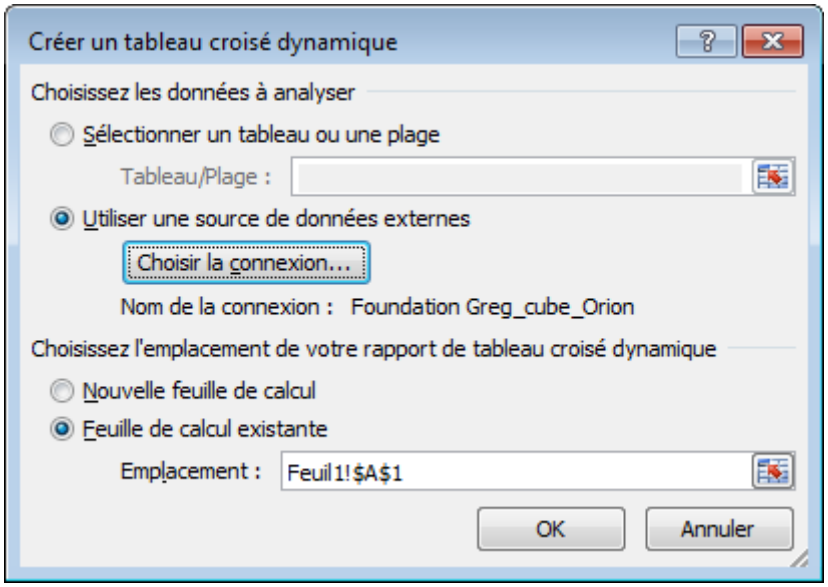
SAS OLAP Data Provider 9.3
Suivant



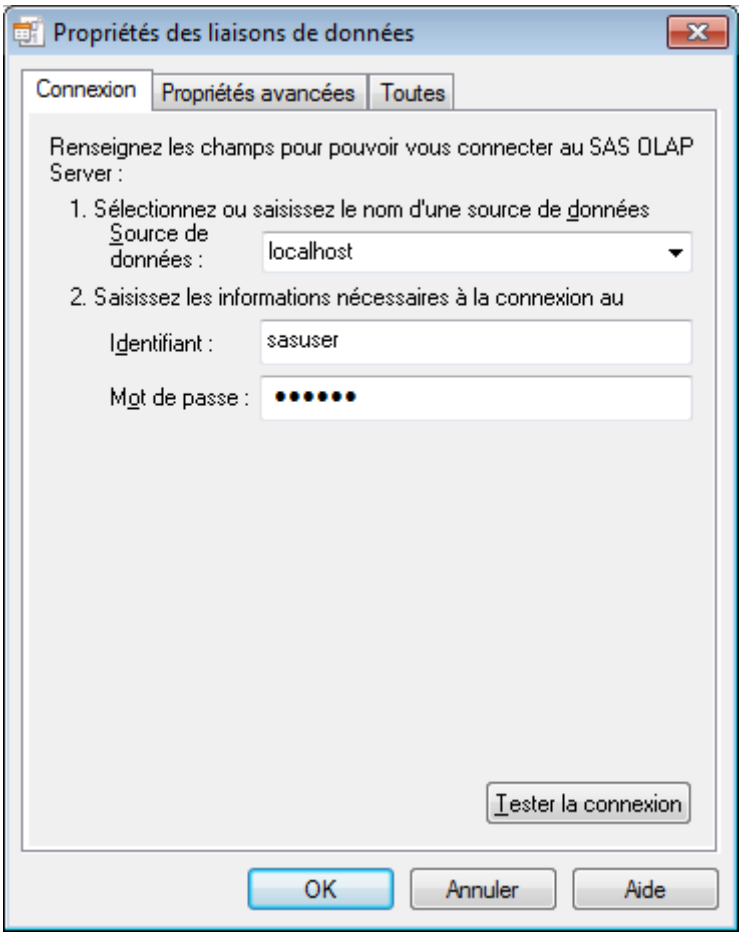
Entrer le nom du serveur, votre nom d'utilisateur et votre mot de passe
OK



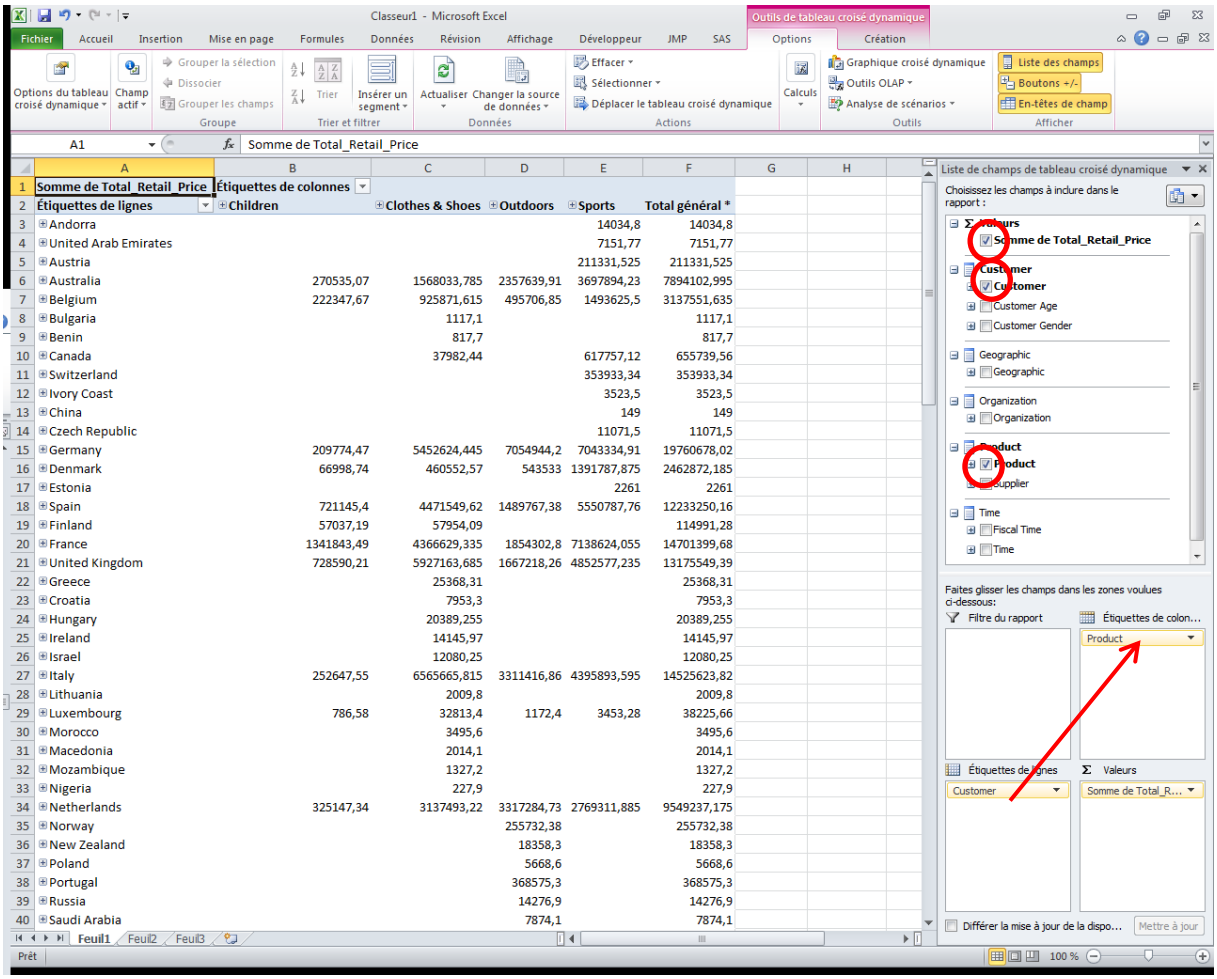
Sélectionner votre cube
Terminer



OK

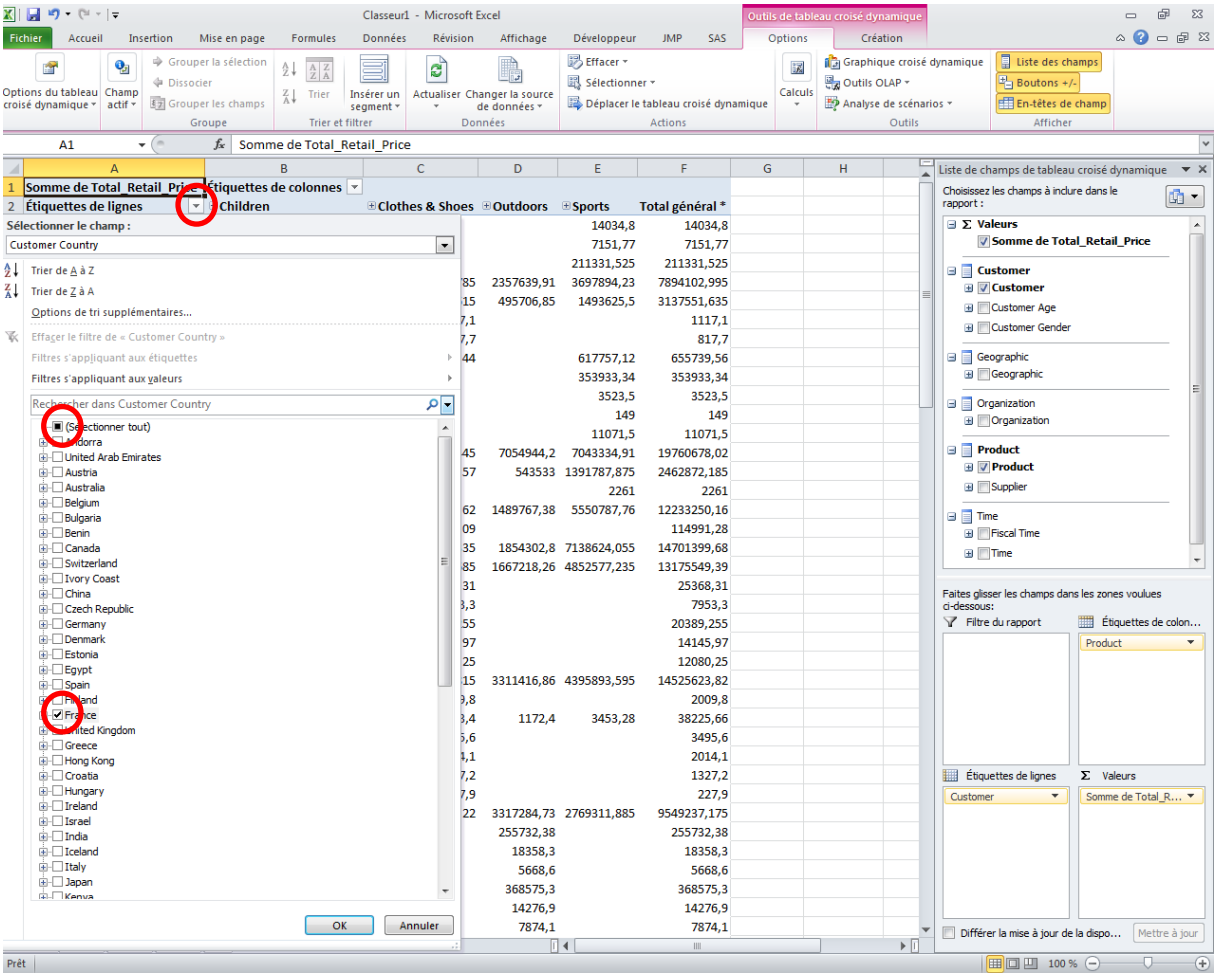


Entrer le mot de passe
OK

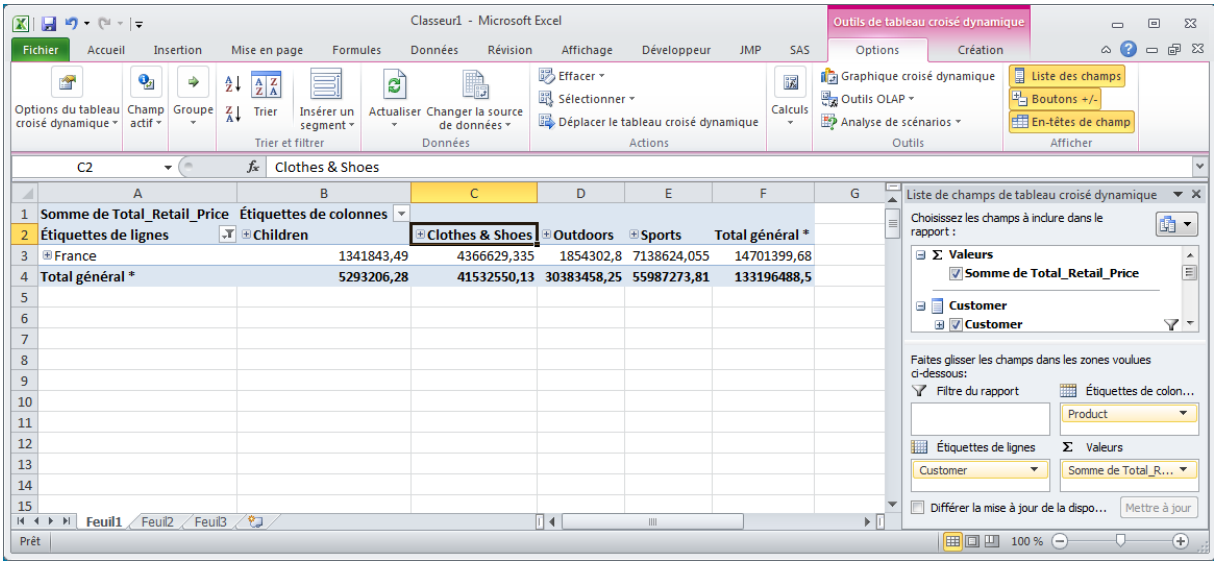


Somme de Total_Retail_Price	Étiquettes de colonnes				
Étiquettes de lignes	Children	Clothes & Shoes	Outdoors	Sports	Total général
Andorra				14034,8	14034,8
United Arab Emirates				7151,77	7151,77
Austria				211331,525	211331,525
Australia	270535,07	1568033,785	2357639,91	3697894,23	7894102,995
Belgium	222347,67	925871,615	495706,85	1493625,5	3137551,635
Bulgaria		1117,1			1117,1
Benin		817,7			817,7
Canada		37982,44		617757,12	655739,56
Switzerland				353933,34	353933,34
Ivory Coast				3523,5	3523,5
China				149	149
Czech Republic				11071,5	11071,5
Germany	209774,47	5452624,445	7054944,2	7043334,91	19760678,02
Denmark	66998,74	460552,57	543533	1391787,875	2462872,185
Estonia				2261	2261
Spain	721145,4	4471549,62	1489767,38	5550787,76	12233250,16
Finland	57037,19	57954,09			114991,28
France	1341843,49	4366629,335	1854302,8	7138624,055	14701399,68
United Kingdom	728590,21	5927163,685	1667218,26	4852577,235	13175549,39
Greece		25368,31			25368,31
Croatia		7953,3			7953,3
Hungary		20389,255			20389,255
Ireland		14145,97			14145,97
Israel		12080,25			12080,25
Italy	252647,55	6565665,815	3311416,86	4395893,595	14525623,82
Lithuania		2009,8			2009,8
Luxembourg	786,58	32813,4	1172,4	3453,28	38225,66
Morocco		3495,6			3495,6
Macedonia		2014,1			2014,1
Mozambique		1327,2			1327,2
Nigeria		227,9			227,9
Netherlands	325147,34	3137493,22	3317284,73	2769311,885	9549237,175
Norway			255732,38		255732,38
New Zealand			18358,3		18358,3
Poland			5668,6		5668,6
Portugal			368575,3		368575,3
Russia			14276,9		14276,9
Saudi Arabia			7874,1		7874,1

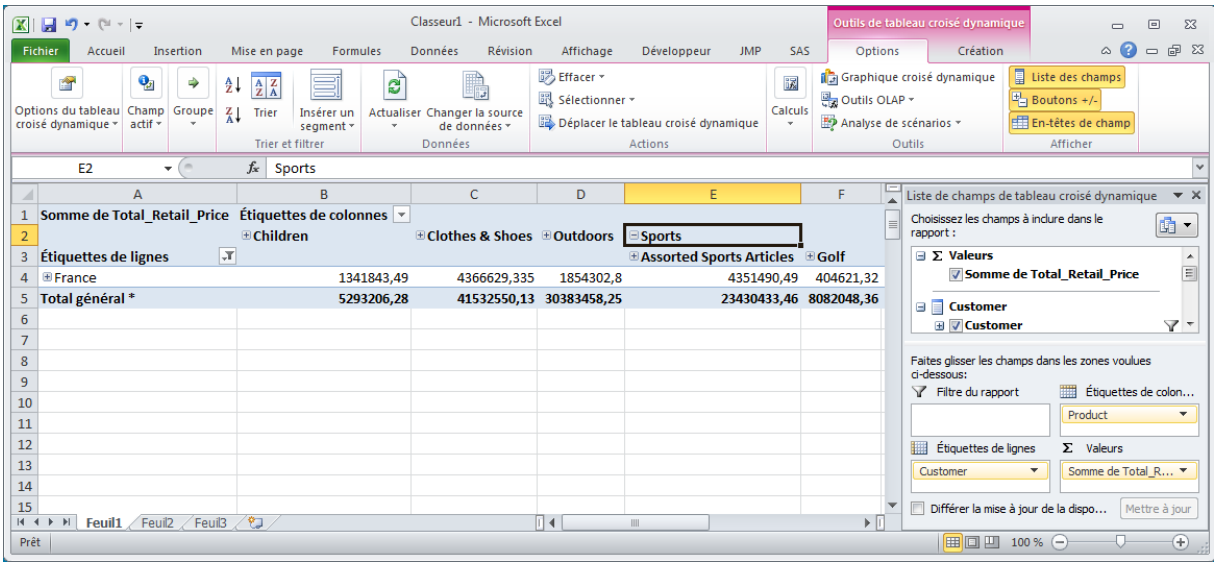
Vous pouvez par exemple sélectionner la somme du chiffre d'affaire par client et par produit
Vous pouvez mettre les produit en colonne



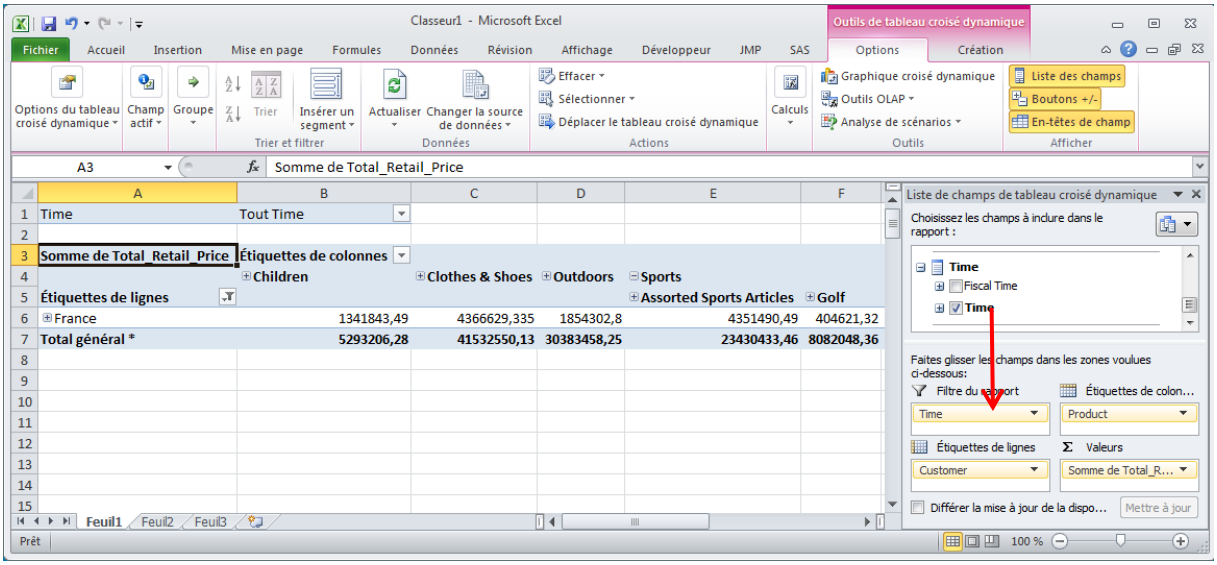
Sur les lignes, vous pouvez tout désélectionner pour ne sélectionner que la France



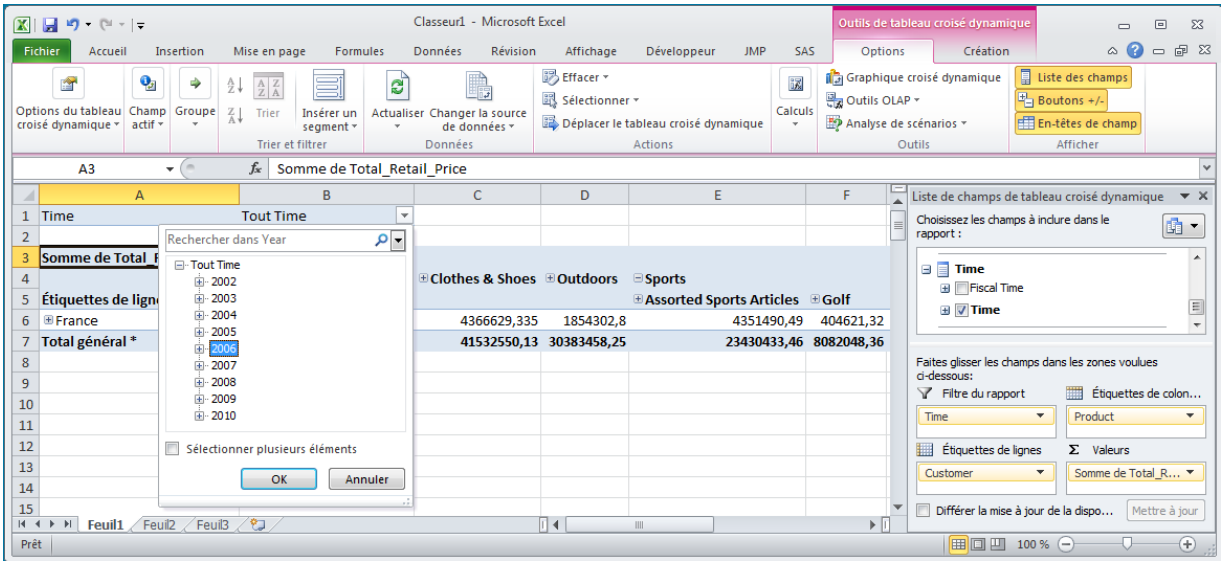
Déplier les articles de sport



Ajouter le temps au filtre



Sélectionner 2006, etc…



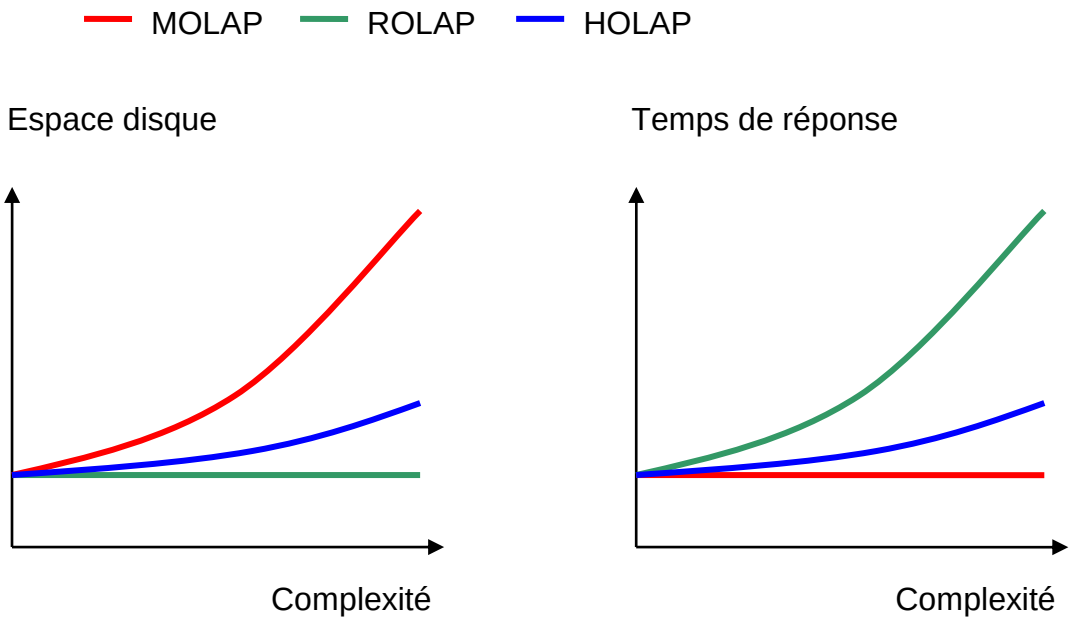
Optimisation d'un cube OLAP

Un bon moyen d'obtenir des temps de réponse performants consiste à «pré-calculer» tous les totaux logiques. Un cube contenant tous ces sous-totaux est un cube MOLAP : Multidimensionnel OLAP. C'est la forme la plus pure d'un cube. Lors de la navigation dans de tels cubes, vu que tous les totaux et sous-totaux ont déjà été calculés, il n'y a pas besoin de les recalculer. Cela permet une navigation très rapide. Par contre un cube MOLAP est très volumineux. Le pré-calcul de tous les agrégats peut prendre beaucoup de place sur le disque. Vu que l'on pré-calcule tous les croisements possibles d'information, cela peut exploser.

Le nombre de toutes les combinaisons possibles de n parmi p est $\subset_n^p = \frac{n!}{p!(n-p)!}$. La fonction factorielle est une fonction qui explose vite.

Un cube ROLAP, Relationnel OLAP n'a pas de structure forcément dédiée. Il permet de naviguer dans l'information comme dans un cube MOLAP mais aucun agrégat n'a été pré-calculé. Donc lorsque vous faites une requête sur ce cube, il faut calculer tous les totaux.

Un cube HOLAP est l'Hybride entre un cube ROLAP et un cube MOLAP. On pré-calcule uniquement les agrégats nécessaires à l'optimum du temps de réponse et de l'espace disque.



Sur les graphiques ci-dessus, la complexité représente la taille du cube en termes de nombre de données, de hiérarchies et de niveaux.

Pour un cube MOLAP :

- L'espace disque nécessaire explose lorsque la complexité augmente.
- Les temps de réponse sont toujours excellents.

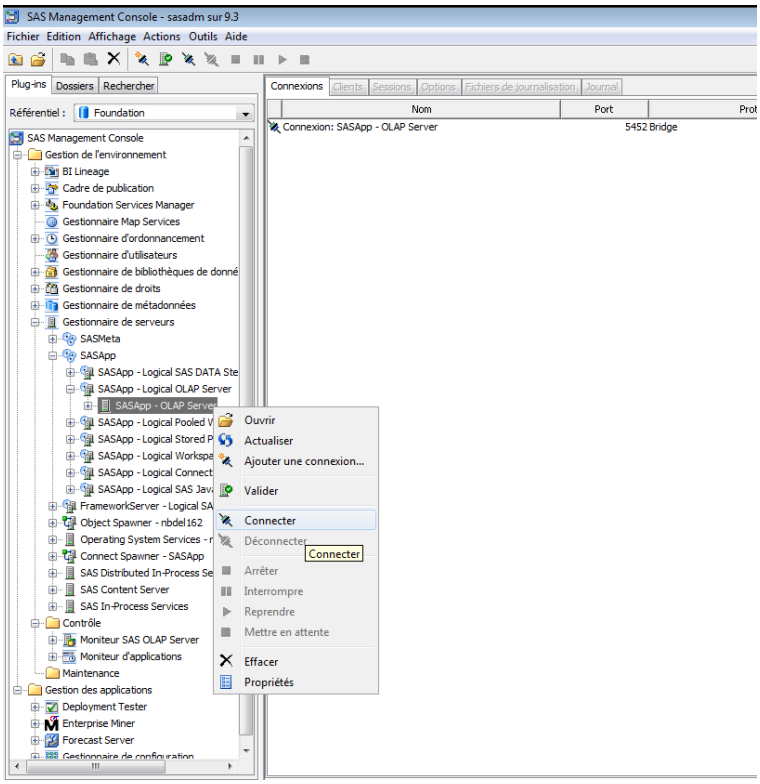
Pour un cube ROLAP :

- L'espace disque nécessaire est toujours réduit.
- Les temps de réponse explosent lorsque la complexité augmente.

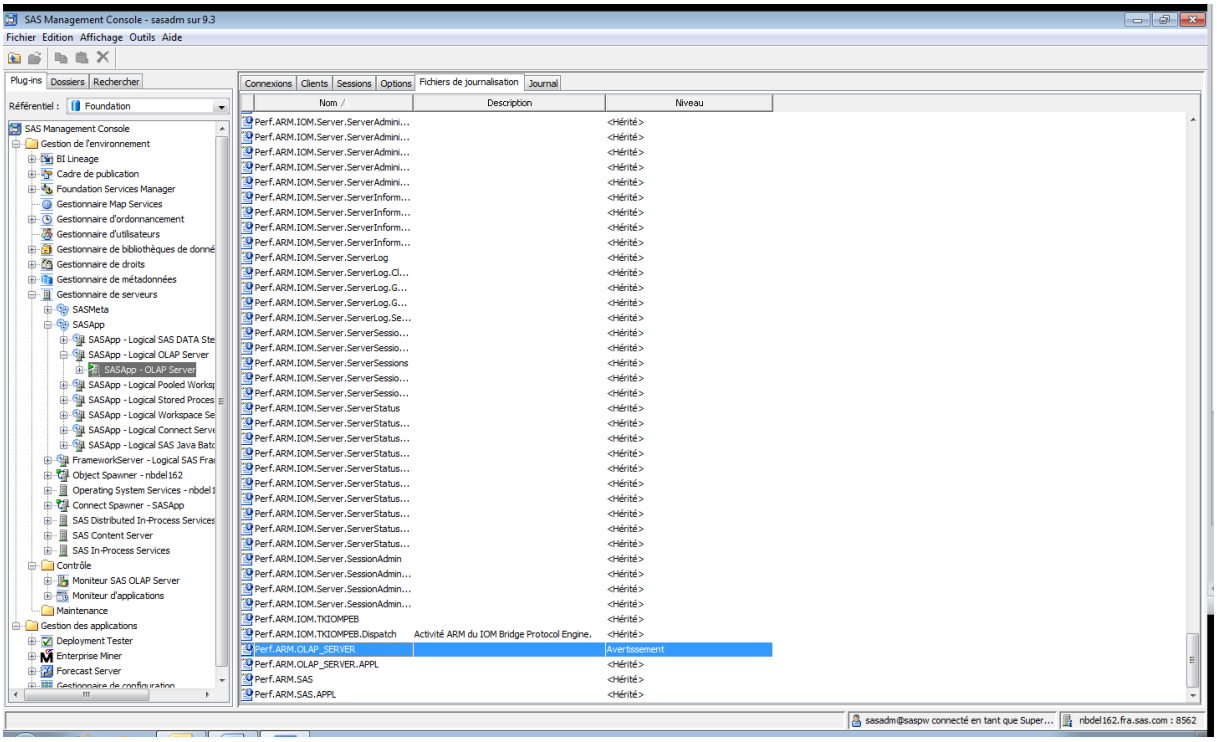
Un cube HOLAP est le compromis entre ces deux cubes. Généralement, pour construire un cube HOLAP, on crée un cube ROLAP et on ajoute des agrégations. Si on ajoute toutes les agrégations possibles, on obtient un cube MOLAP. Un agrégat est le résultat d'un calcul d'une mesure sur les éléments d'un niveau ou d'un croisement de niveau. Par exemple, la somme du chiffre d'affaire (la

mesure) par mois (le niveau) ou bien la moyenne du chiffre d'affaire (la mesure) par mois et par pays (un croisement de niveau).

Pour pouvoir optimiser un cube avec l'analyse des performances AMR, il faut que cette fonction soit activée dans la SAS Management console :

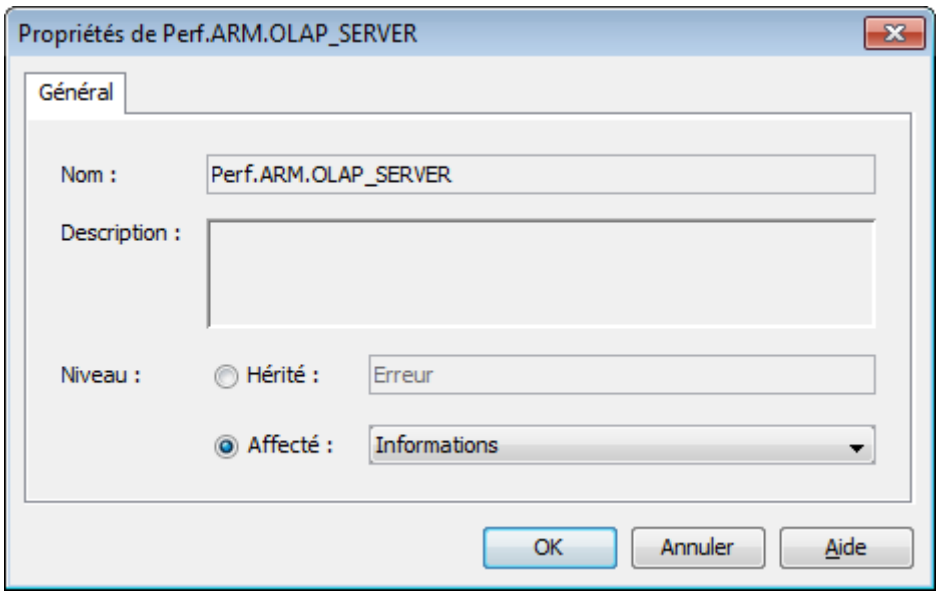


Se connecter dans la dans la SAS Management Console en super administrateur, se connecter au serveur OLAP (Gestionnaire des serveurs → SASApp – Logical OLAP Serveur → SASApp – OLAP Server → Clic-droit → connecter



L'onglet « fichier de journalisation » n'est plus grisé.

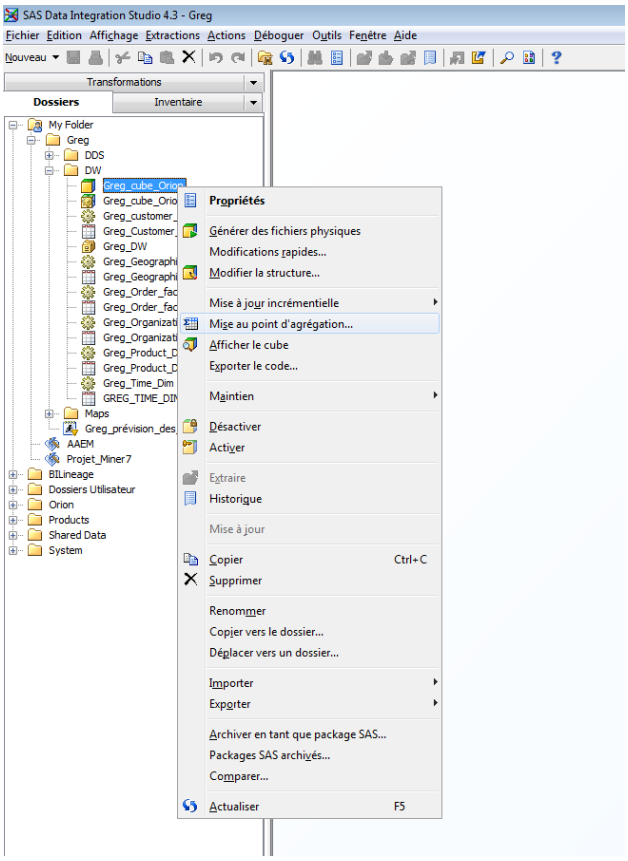
(Trier par nom) Dans cet onglet, sélectionner « Perf.ARM.OLAP_Server » → Clic-droit → propriétés



Sélectionner « Information ».

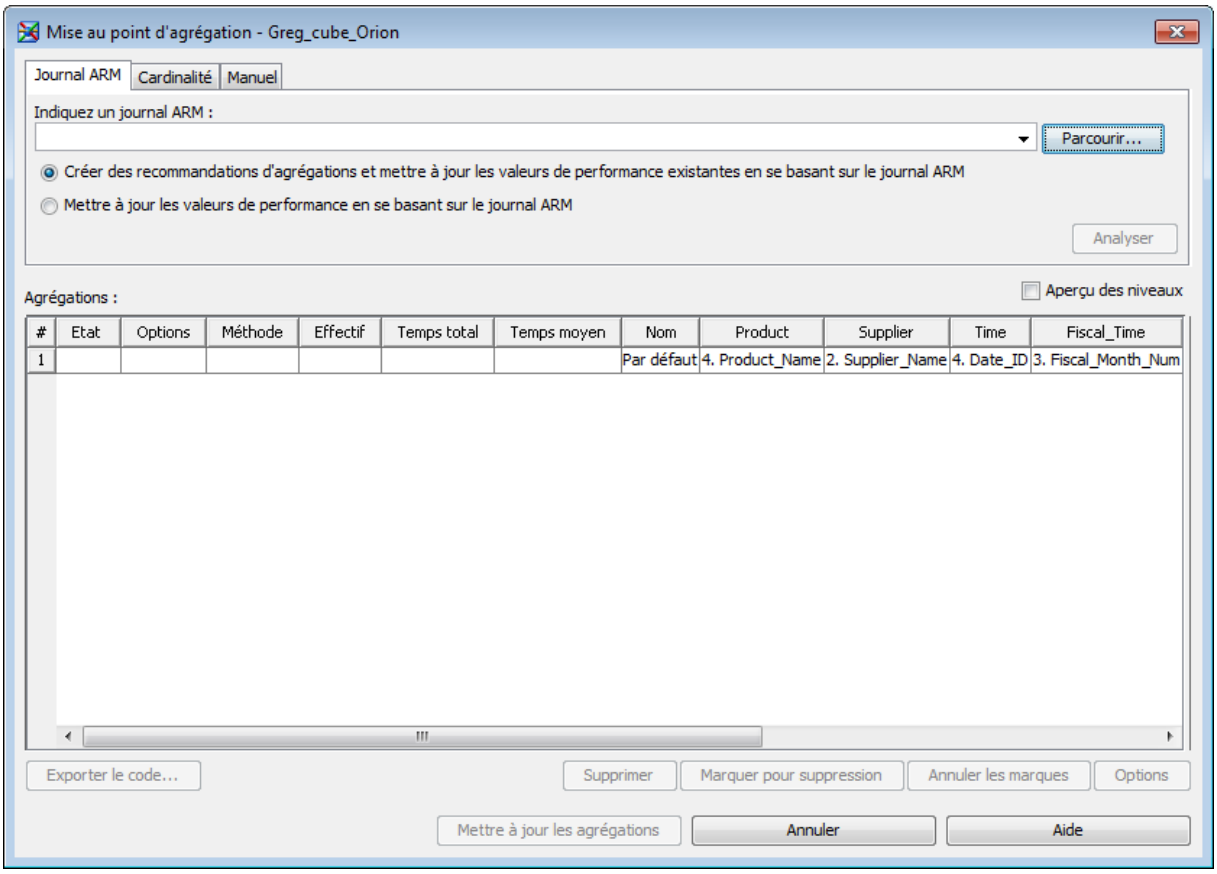
OK

Dans SAS Data Integration Studio, en faisant un clic-droit sur le cube, on peut sélectionner la **Mise au point d'agrégation.**



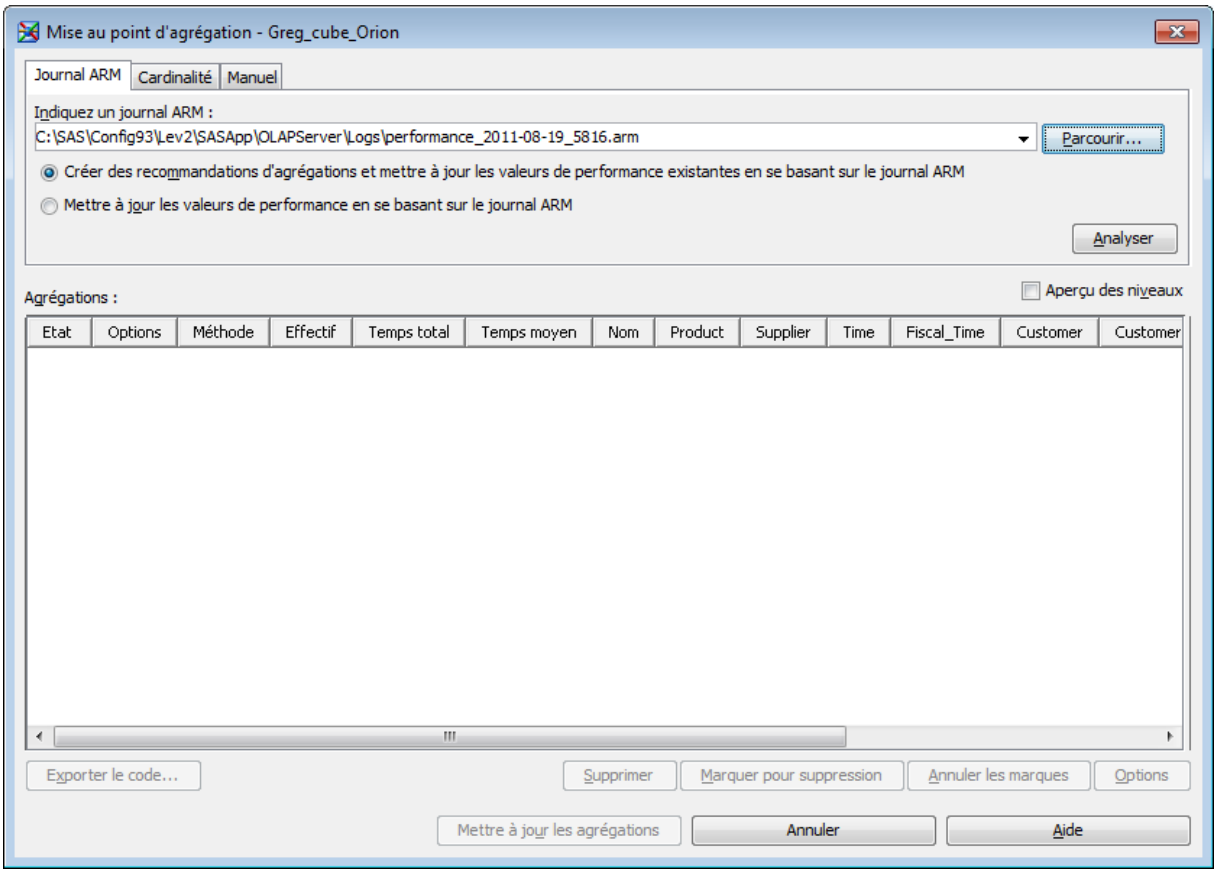
Il existe trois types de mise au point d'agrégation :

1. L'analyse du fichier de journal AMR (Application Response Measurement) permet d'obtenir un rapport sur le temps de réponse des requêtes MDX sur le cube, sur les agrégats par croisements de niveau non déjà calculés. Pour chaque agrégat non déjà calculé, on obtient le nombre de fois où il a été nécessaire de calculer cet agrégat et le temps qu'a nécessité ce calcul. En fonction du nombre de requêtes par croisement de niveaux et des temps de réponse sur ces croisement de niveaux, on calcule les agrégats nécessaires de telle sorte qu'à l'a prochaine utilisation du cube, le temps de réponse sur cet agrégats soit quasi nul, car s'il est déjà calculé, cela fonctionne beaucoup plus vite.
   Il faut pour cela créer un cube ROLAP et naviguer dedans afin de générer des éléments dans le fichier de journal ARM. Fermez la mise au point d'agrégation, modifiez la structure du cube en cochant à la dernière étape (agrégation) « ne pas créer d'agrégation NWAY », régénérer votre cube (il sera alors ROLAP), naviguez dedans avec SAS Enterprise Guide, Microsoft Excel ou une autre interface ; et vous pourrez alors analysez les journaux ARM de votre cube.

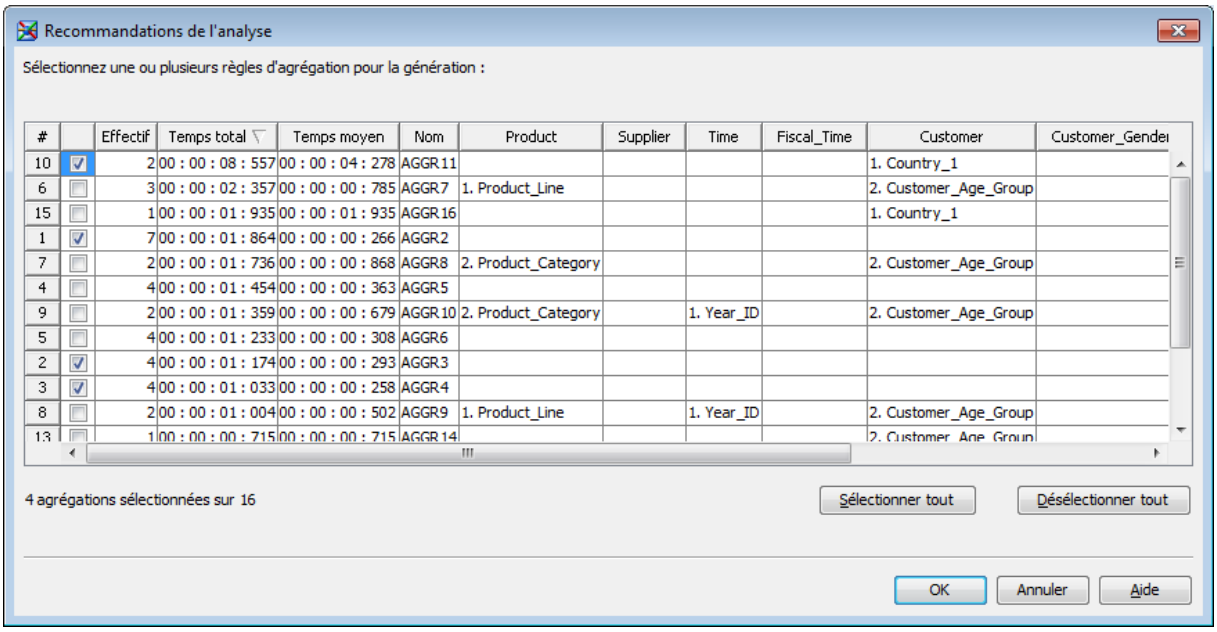


Cliquez sur **Parcourir**

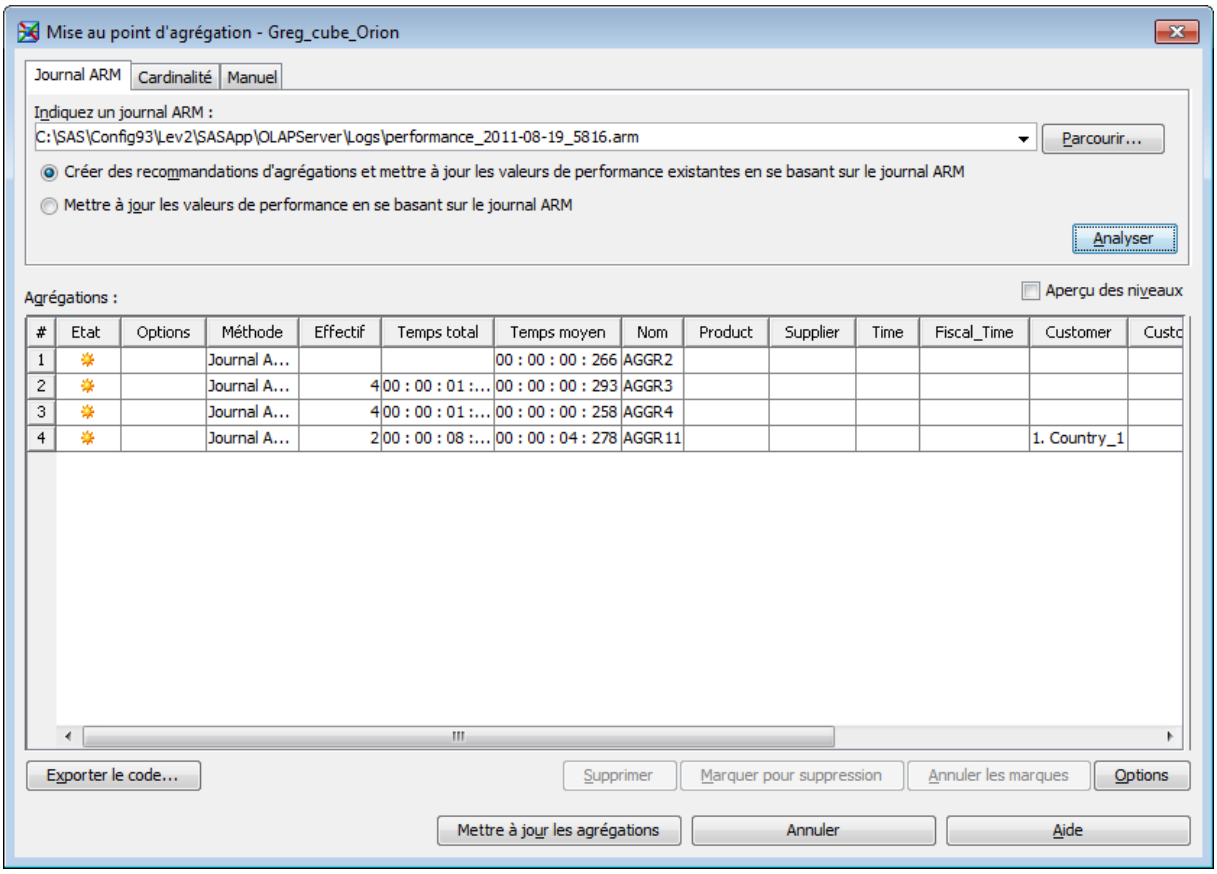
Le fichier ARM se trouve dans un dossier sur le serveur dont le chemin ressemble à : C:\SAS\Config\Lev1\SASApp\OLAPServer\Logs



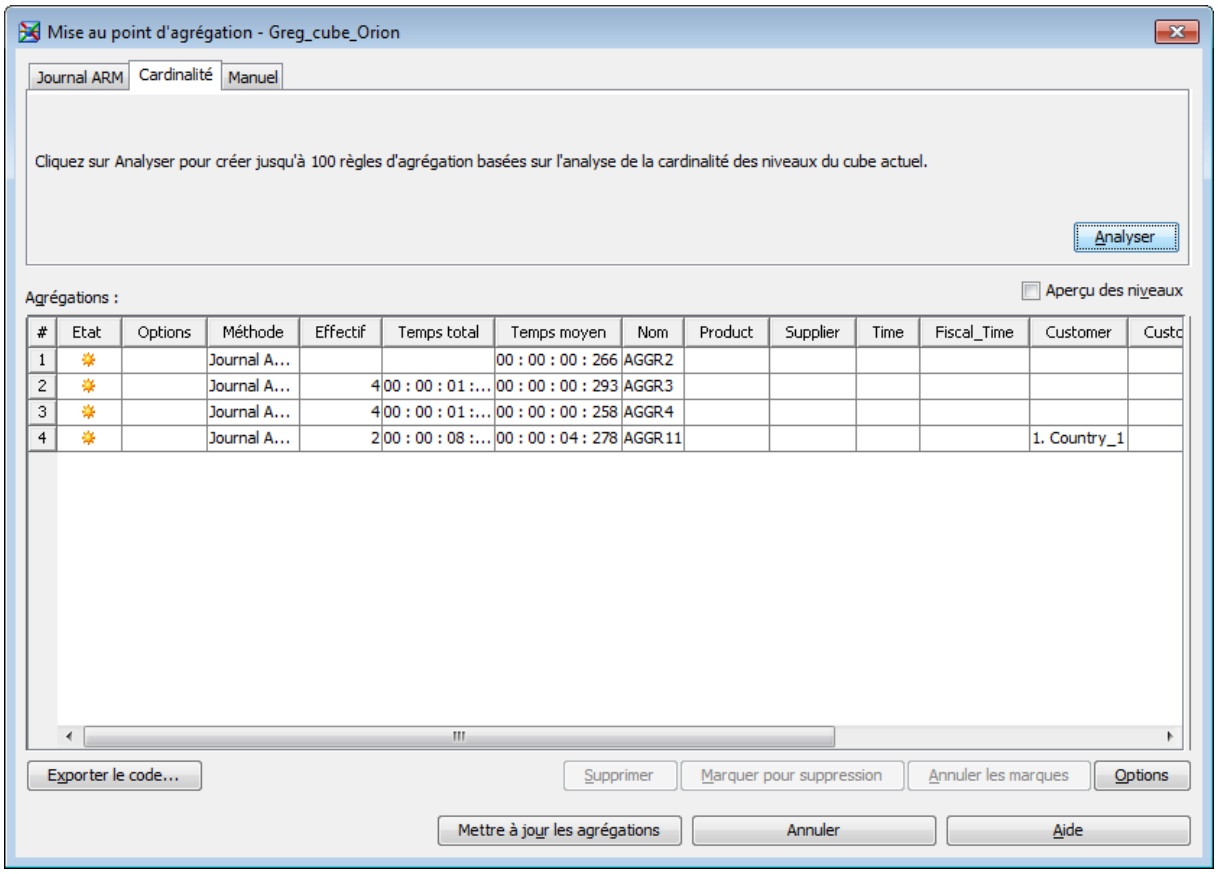
Lorsque vous avez sélectionné votre fichier journal du jour, cliquez sur le bouton Analysez



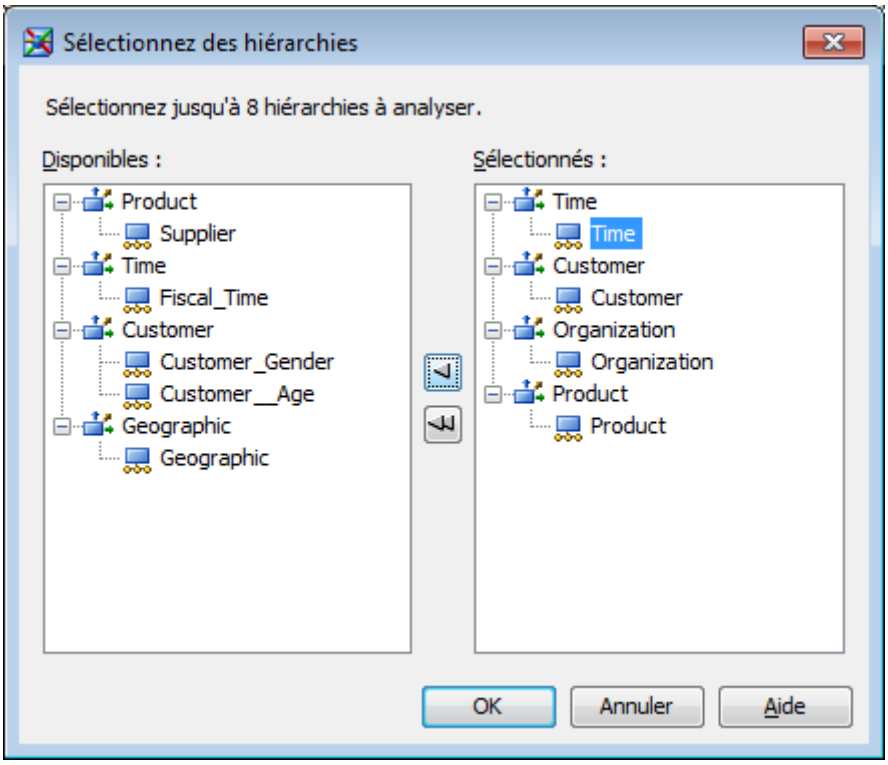
En triant par Effectif, temps total ou temps moyen, vous pouvez sélectionner les agrégats les plus pertinant.
OK



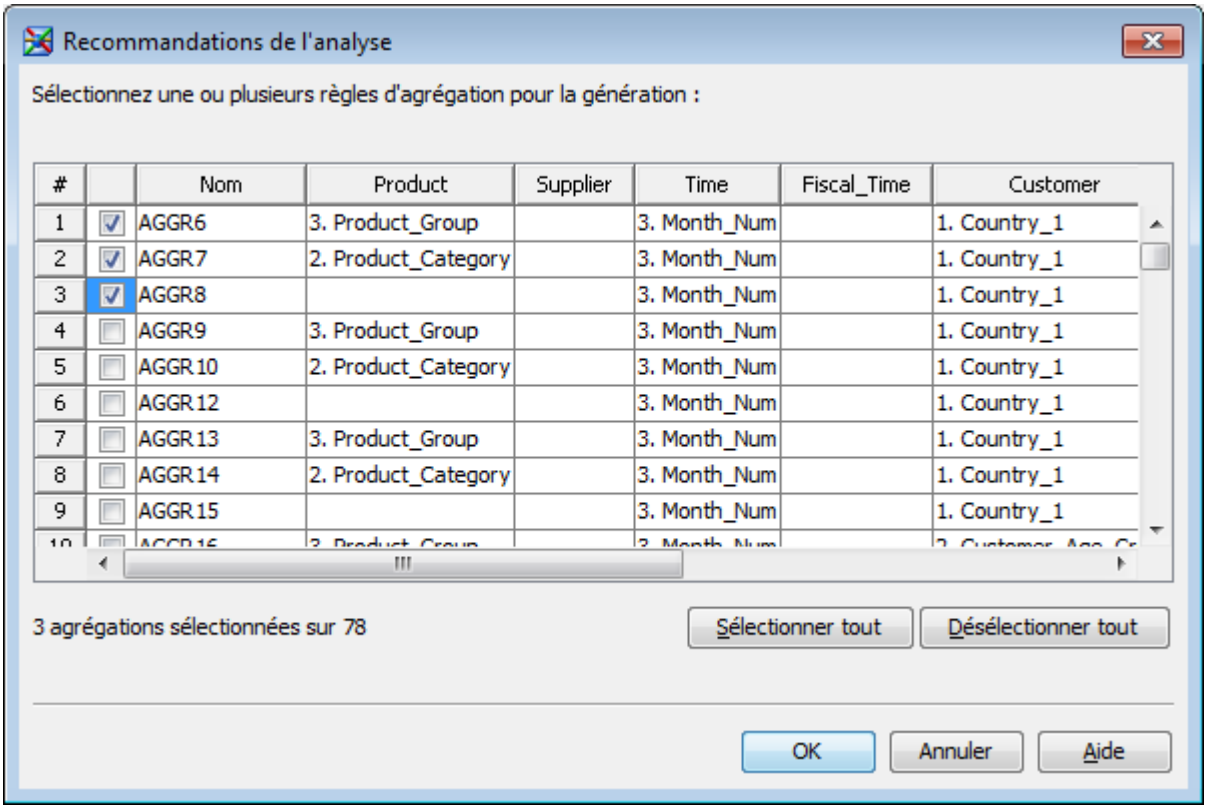
2. La second méthode de mise au point des agrégats est basée sur le calcul de la cardinalité. On utilise souvent cette méthode pour créer un embryon de cube H-OLAP sur les agrégats ayant la cardinalité la plus importante. Par la suite, pour optimiser ce cube, on utilisera la méthode de l'analyse du fichier de log AMR. L'analyse de la cardinalité est basée sur un calcul qui permet de mesurer les zones les plus creuses du cube et donc propose les agrégations ayant le plus fort intérêt à être calculés pour améliorer le compromis H-OALP.
   Dans l'onglet cardinalité, cliquez sur Analyser



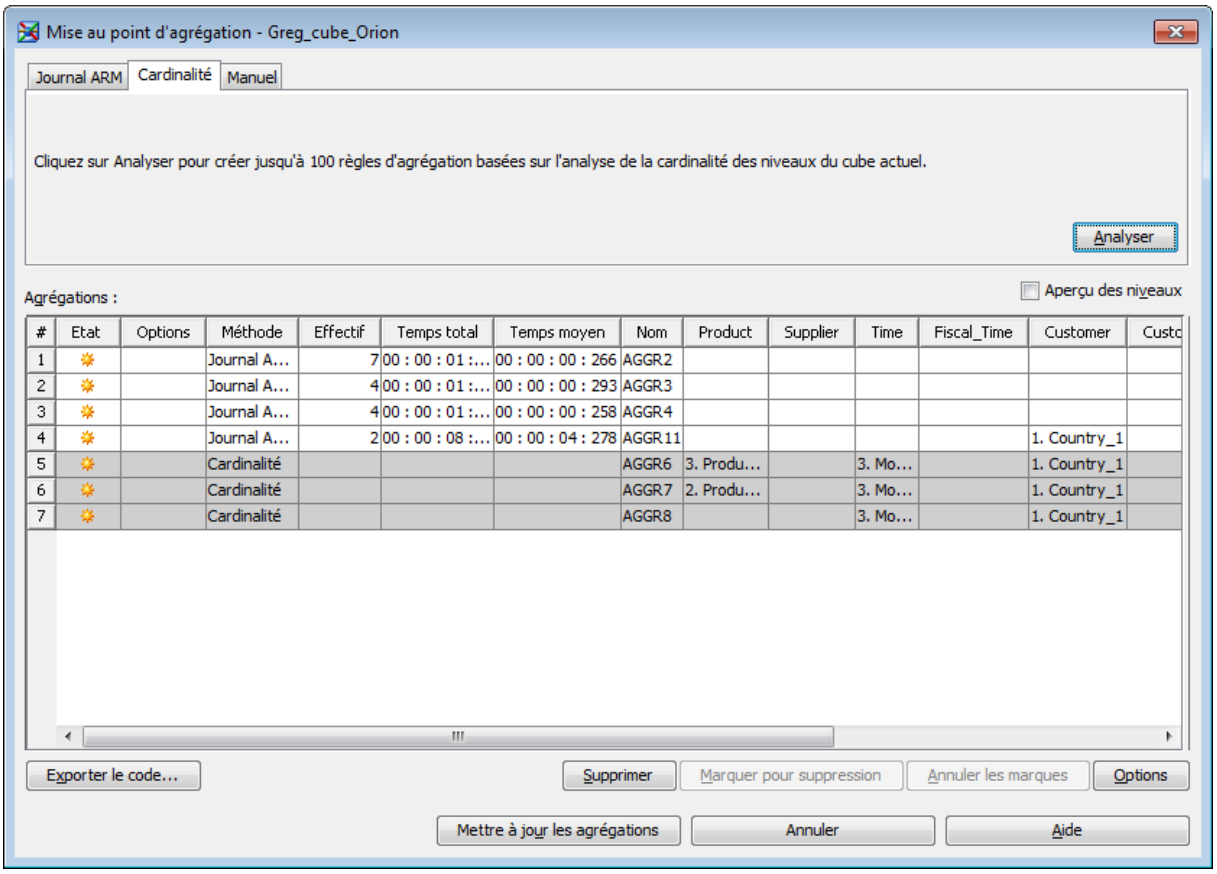
Sélectionner au maximum huit hiérarchies afin pour ce calcul



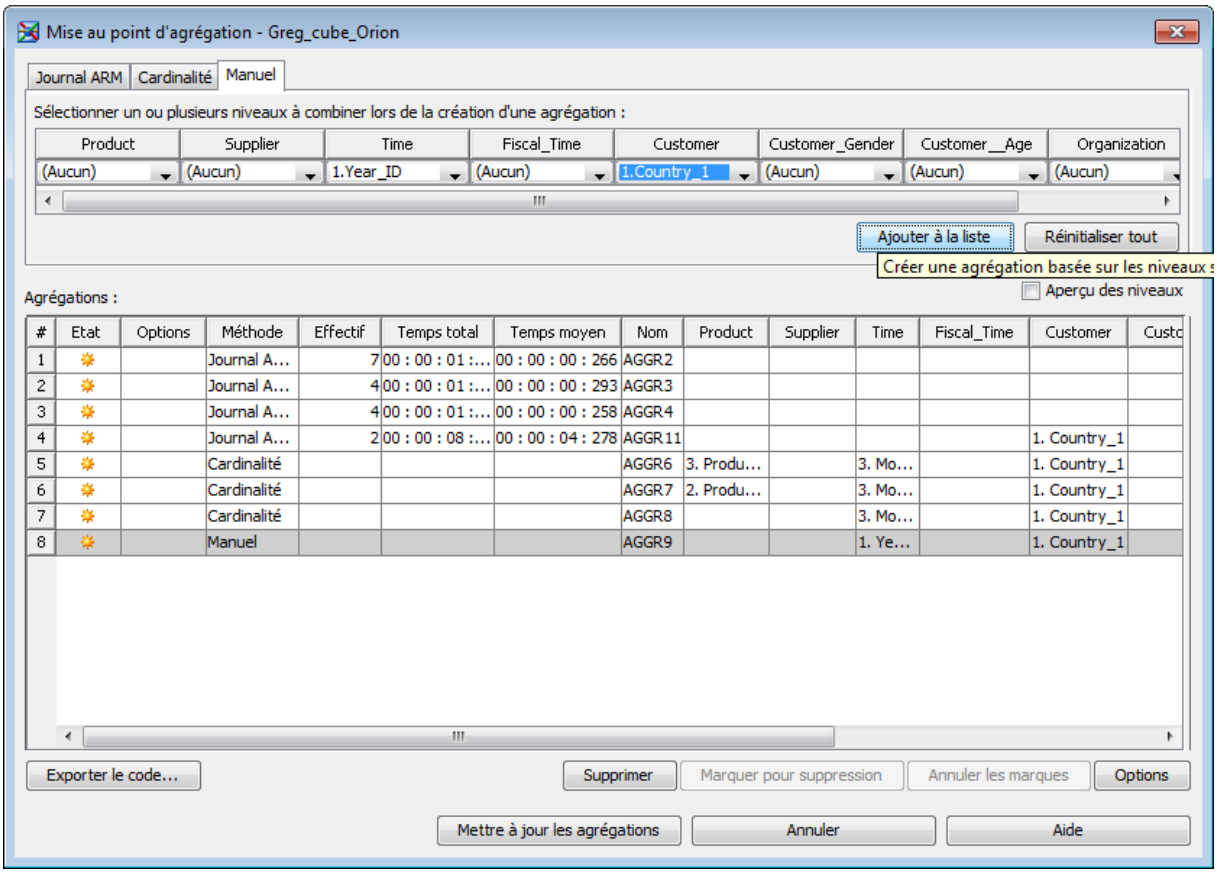
OK



Les première hiérarchies seront celle qui amélioreront le plus la performance du cube, mais aussi celles qui prendrons le plus de place sur le disque (et très largement).
Vous pouvez par exemple sélectionner les premières.
OK



3. La mise au point manuelle. On sélectionne manuellement les niveaux ou les croisements de niveau sur lesquels on souhaite créer des agrégats.

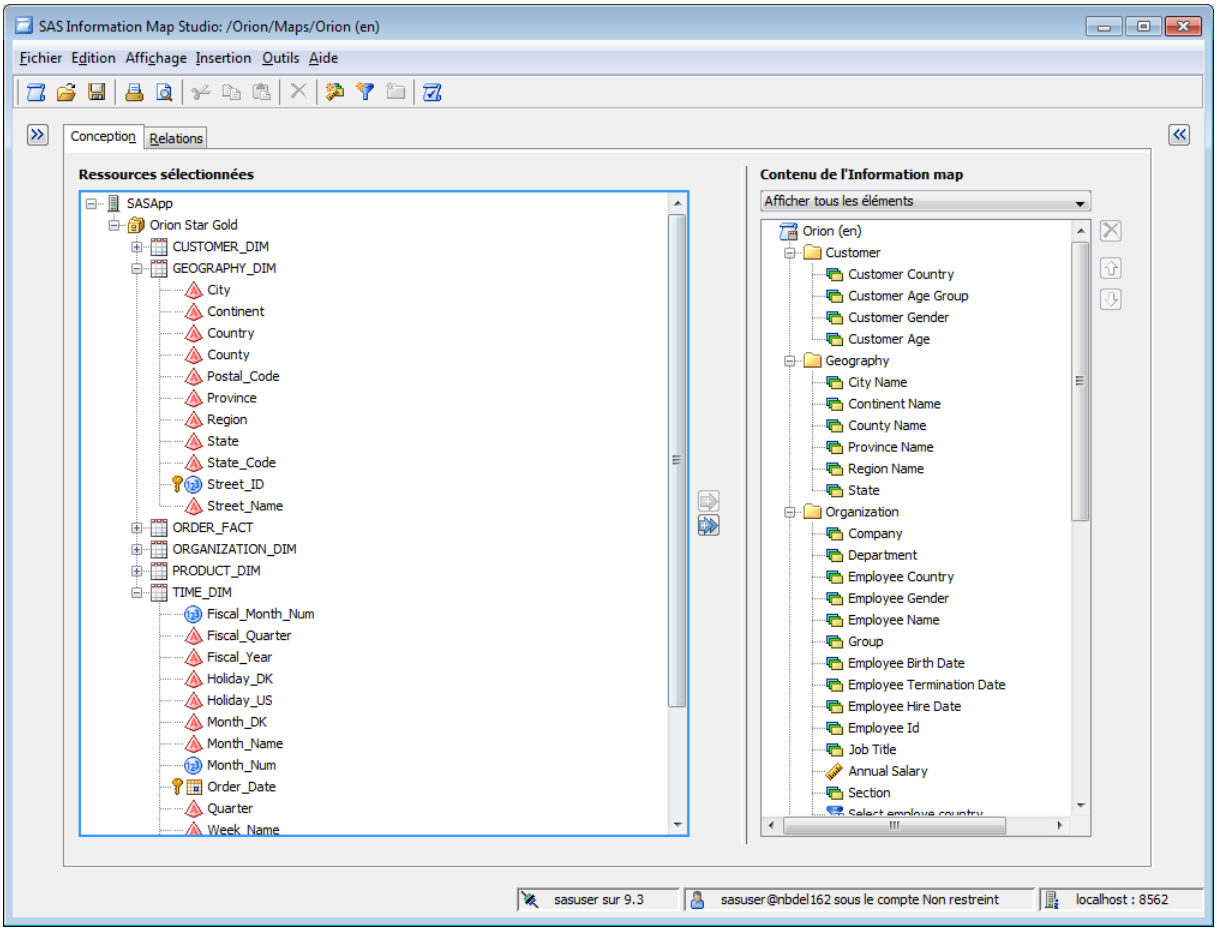


Exercice : Comparer un cube R-OLAP d'un Cube M-OLAP et d'un cube H-OLAP en termes d'espace disque et de temps de réponse.

Information Map

Un des objectifs principaux de l'informatique décisionnelle est de permettre à des utilisateurs non informaticiens de pouvoir créer simplement des rapports. L'informatique décisionnelle, par rapport à son ancêtre, l'EIS (Executive Information Systeme) est dynamique. Un EIS fournissait des rapports statiques. La maîtrise d'ouvrage définissait son cahier des charges, puis la maîtrise d'œuvre développait les rapports souhaités. Les solutions actuelles fournissent la possibilité aux utilisateurs de la maîtrise d'ouvrage de faire eux-mêmes leur rapport sans connaissance particulière de l'informatique. Les systèmes décisionnels actuels sont dynamiques.
Une Information Map de l'éditeur SAS, créée avec l'outil SAS Information Map Studio, ressemble beaucoup à un Univers de l'éditeur Business Object (Racheté par SAP), créé avec l'Outil BO Designer.

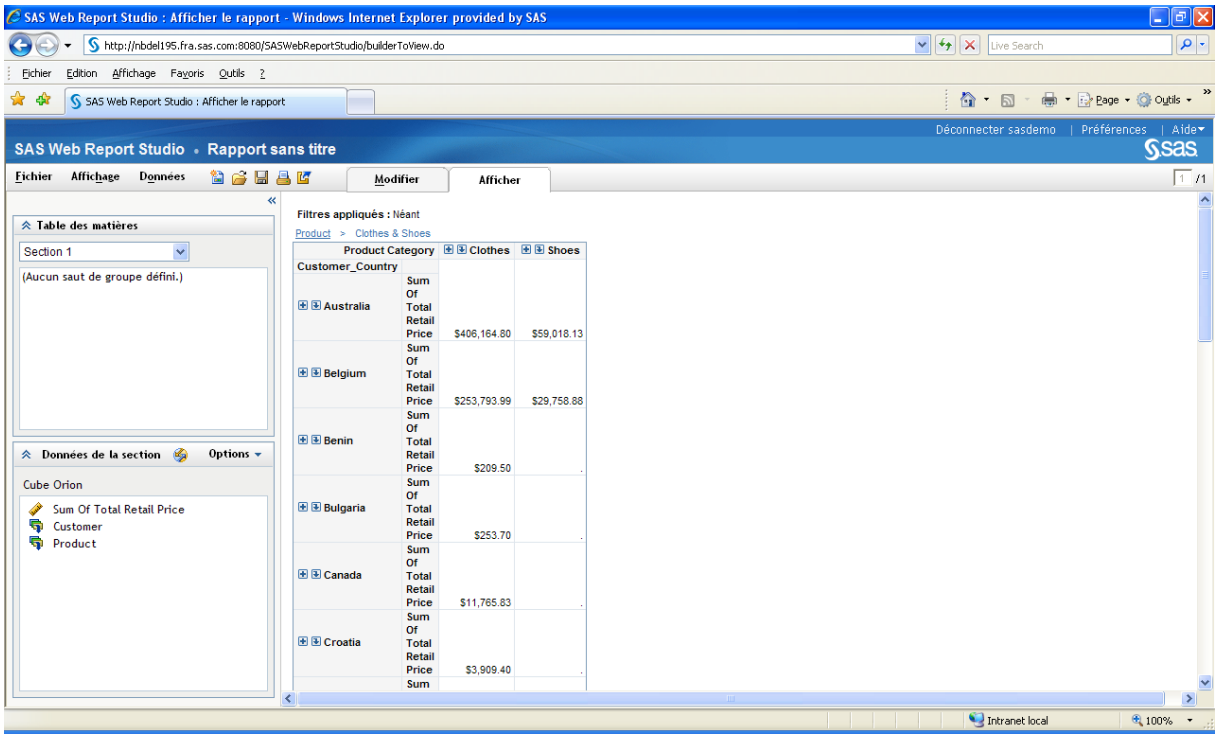
L'interface SAS Information Map Studio



SAS Information Map Studio = IMS.
Pour la création d'une *Map*, il y a deux parties importantes : la structure physique, partie de gauche dans l'espace de création des *Map* ; et la partie virtuelle, conceptuelle, celle que verra l'utilisateur non informaticien de SAS Web Report Studio, partie droite de l'interface : « *information Map* ».

Elément d'une *Map*

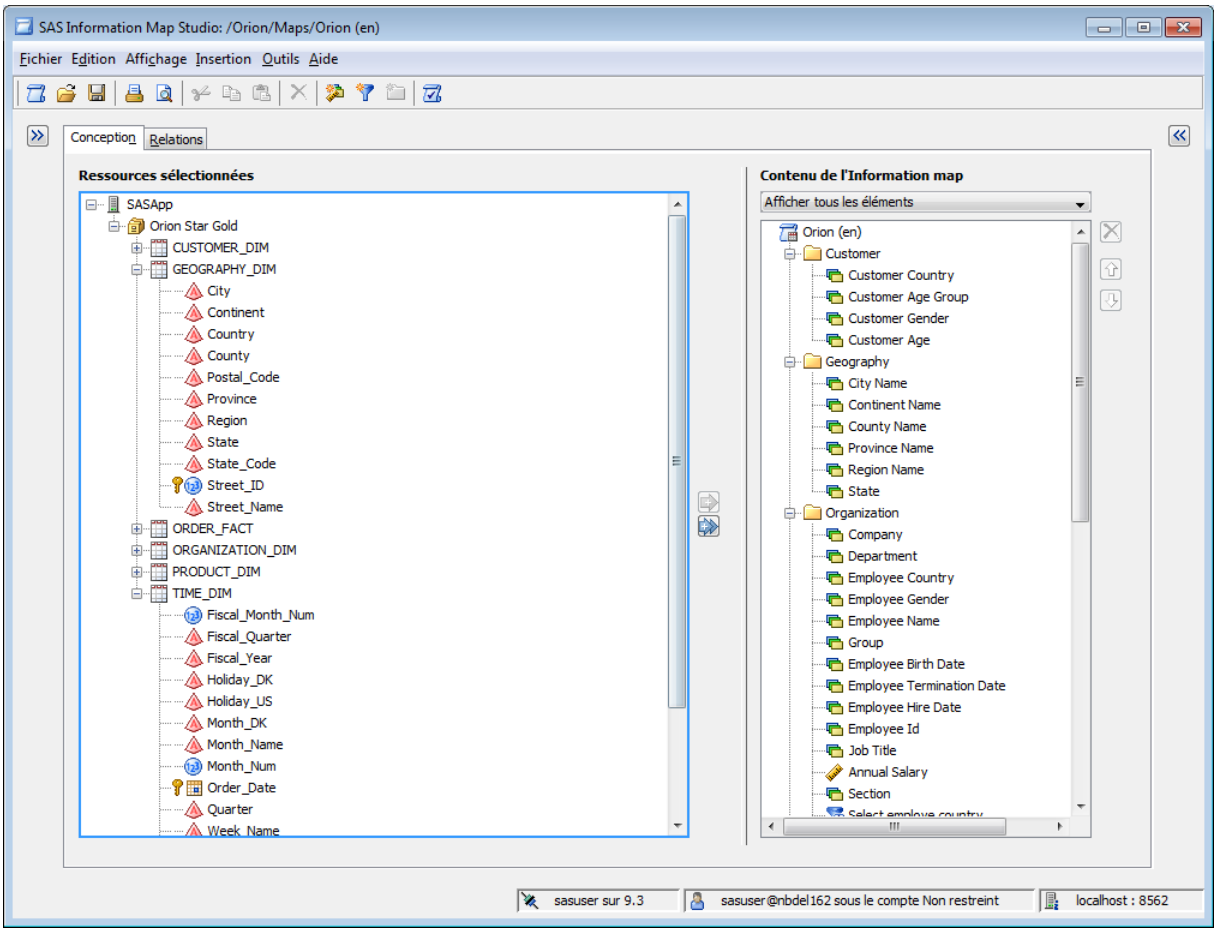
Une *Map* contient deux types d'éléments : des mesures et des catégories. Concrètement, pour l'utilisateur final, les catégories seront les éléments qu'il pourra placer en ligne ou en colonne, et les mesures seront les chiffres qu'il souhaitera exploiter dans le tableau.

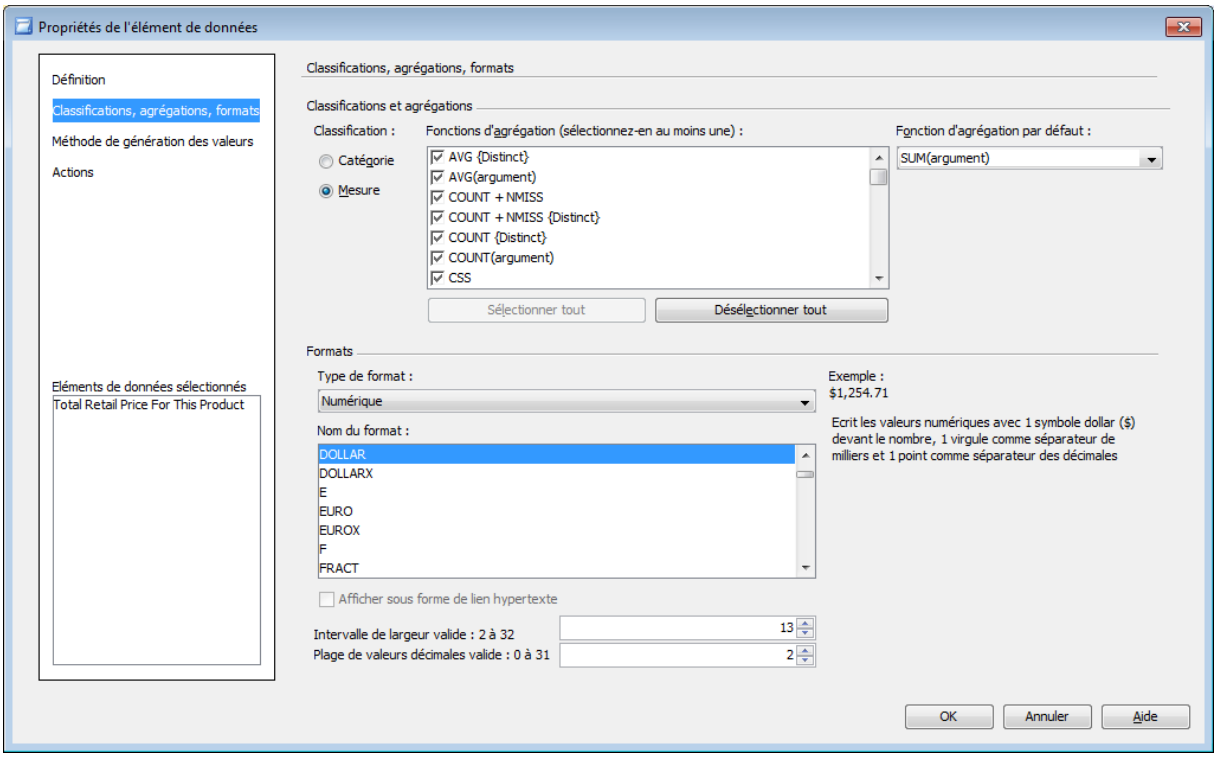


Dans l'exemple ci-dessus, pour cette navigation depuis le portail, on a sélectionné :

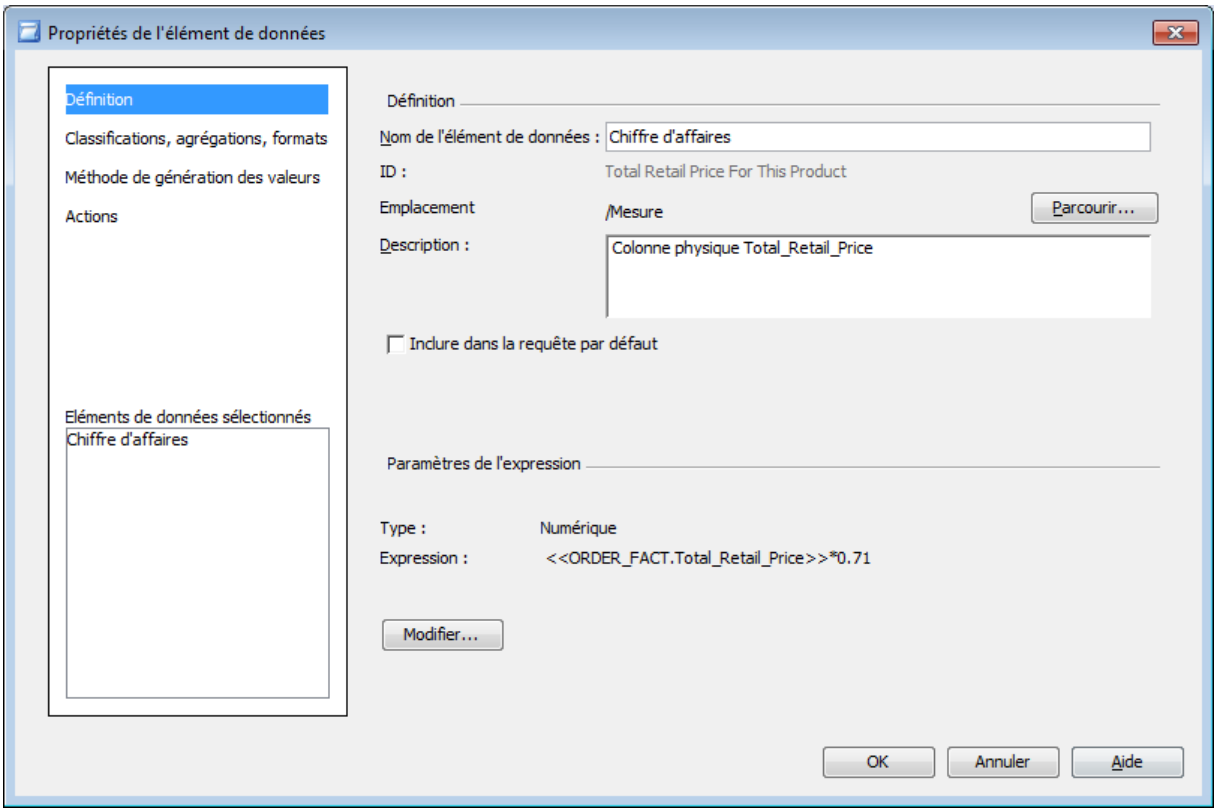
- Une mesure
  - la somme du chiffre d'affaires,
- deux catégories
  - les clients en ligne,
  - les produits en colonne.

Une variable de type caractère sera une catégorie, une variable numérique pourra être de type mesure ou catégorie.
Dans les propriétés de chaque élément, on peut sélectionner le type de l'élément.



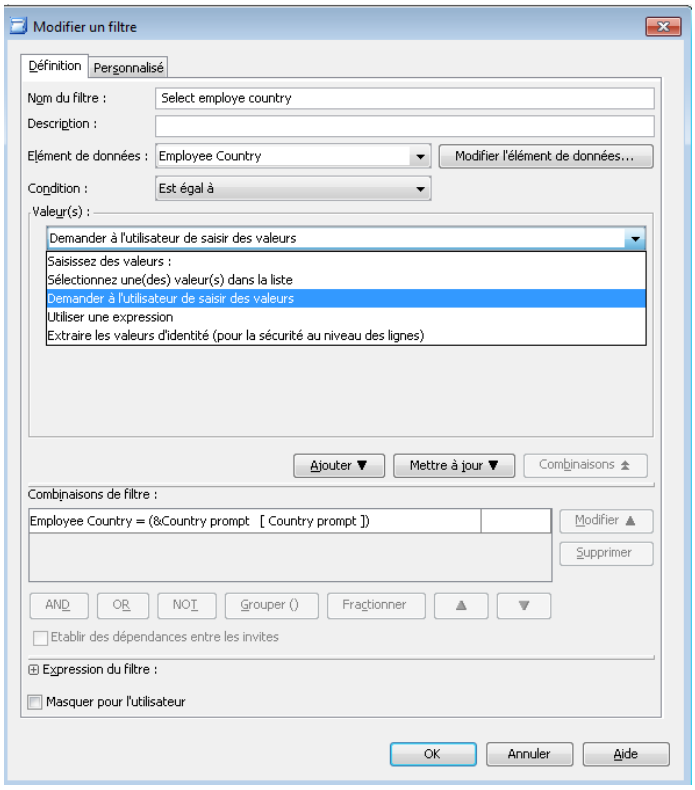


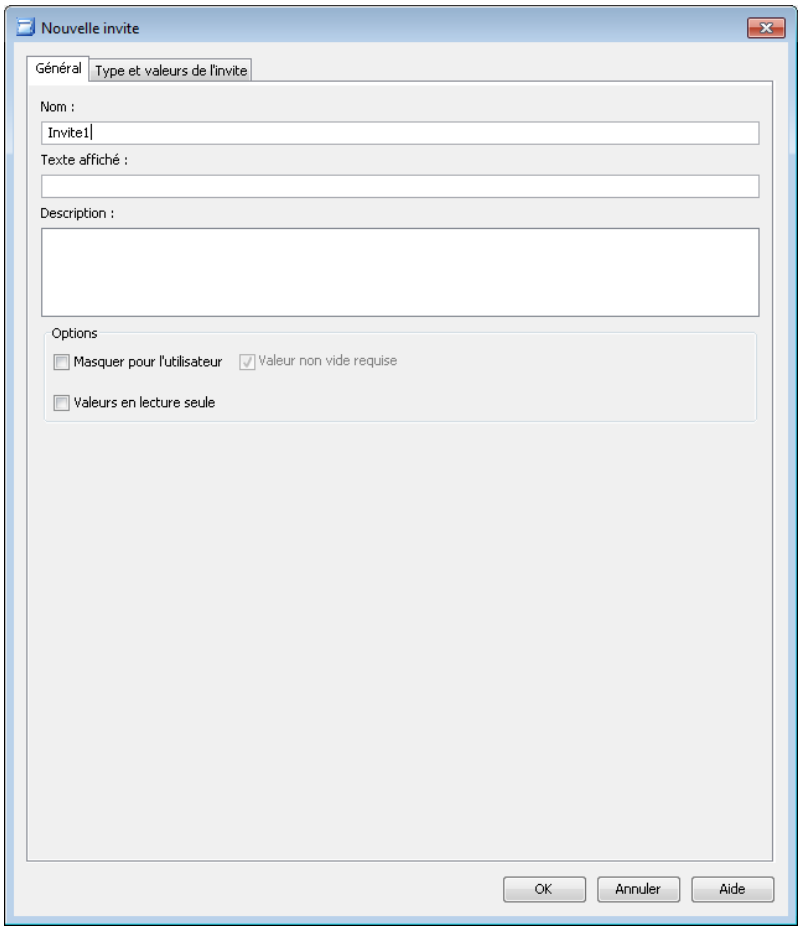
Dans les propriétés d'un élément, on pourra notamment définir aussi le format d'affichage, les agrégations souhaitées, ou modifier la valeur :

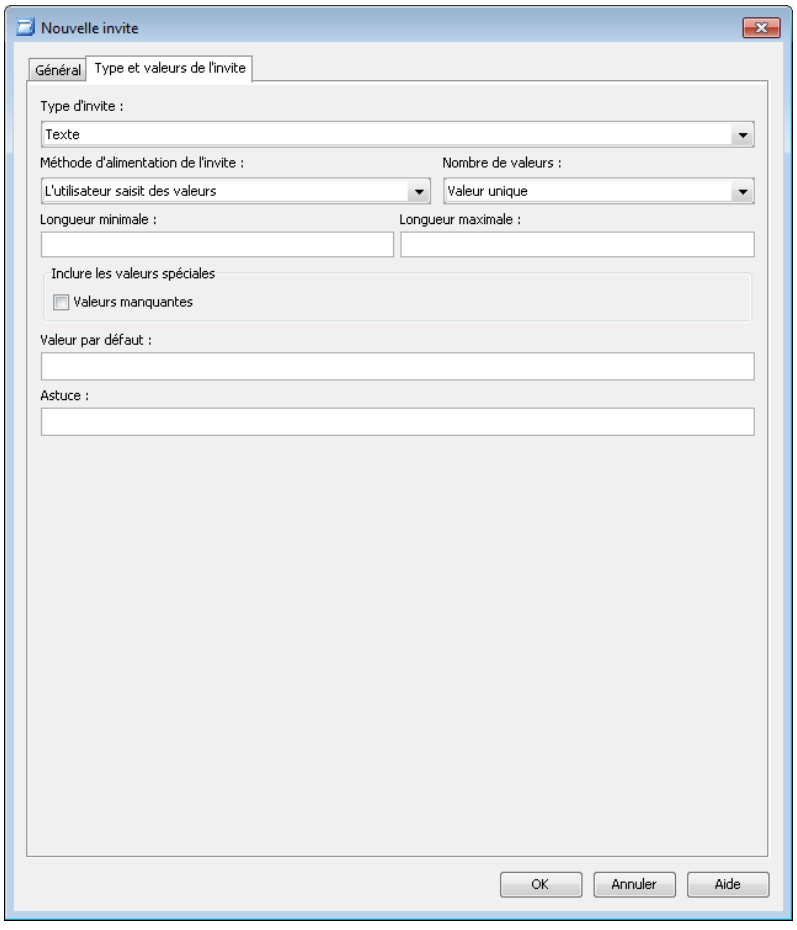


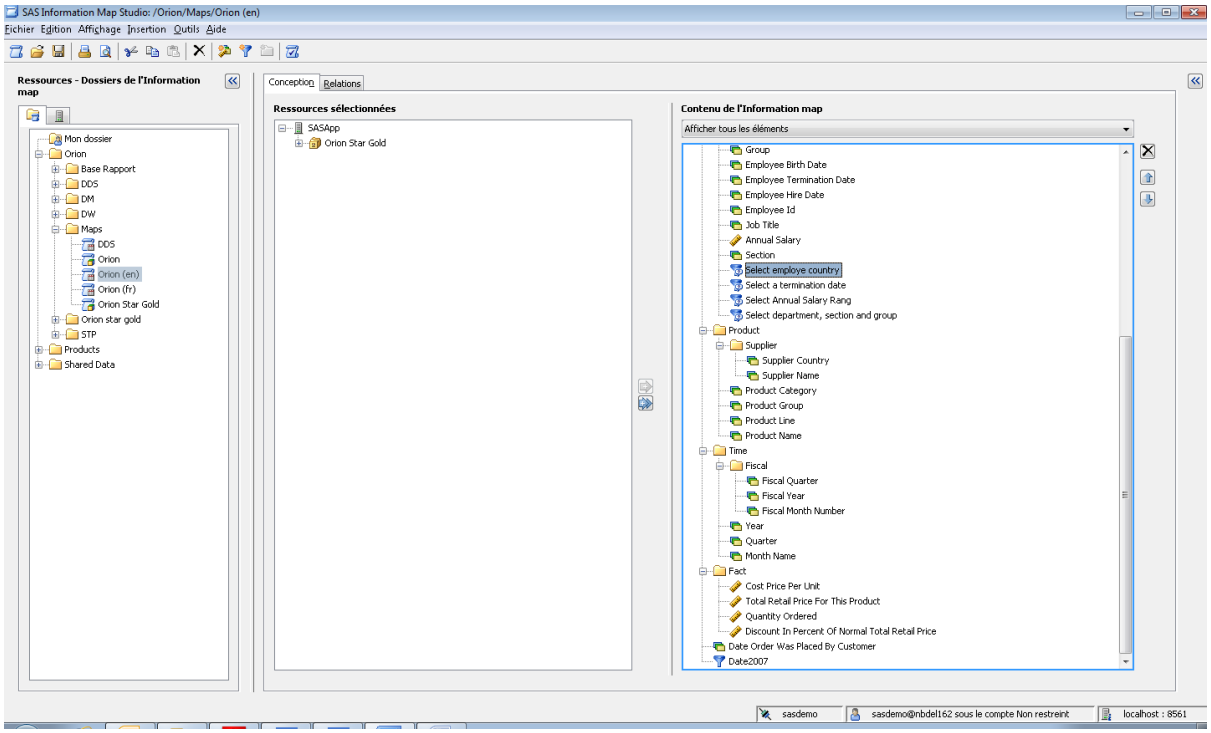
Dans l'exemple ci-dessus, le chiffre d'affaires stocké en dollar dans le *Data Warehouse* est multiplié par 0.71 pour ainsi le convertir en euro pour un utilisateur français.

On peut aussi créer des éléments de type filtre pour que l'utilisateur puisse faire des sélections simplement.




Nouvelle invite
Général
Type et valeurs de l'invite
Nom :
Invite1
Texte affiché :
Description :
Options
Masquer pour l'utilisateur
Valeur non vide requise
Valeurs en lecture seule
OK
Annuler
Aide

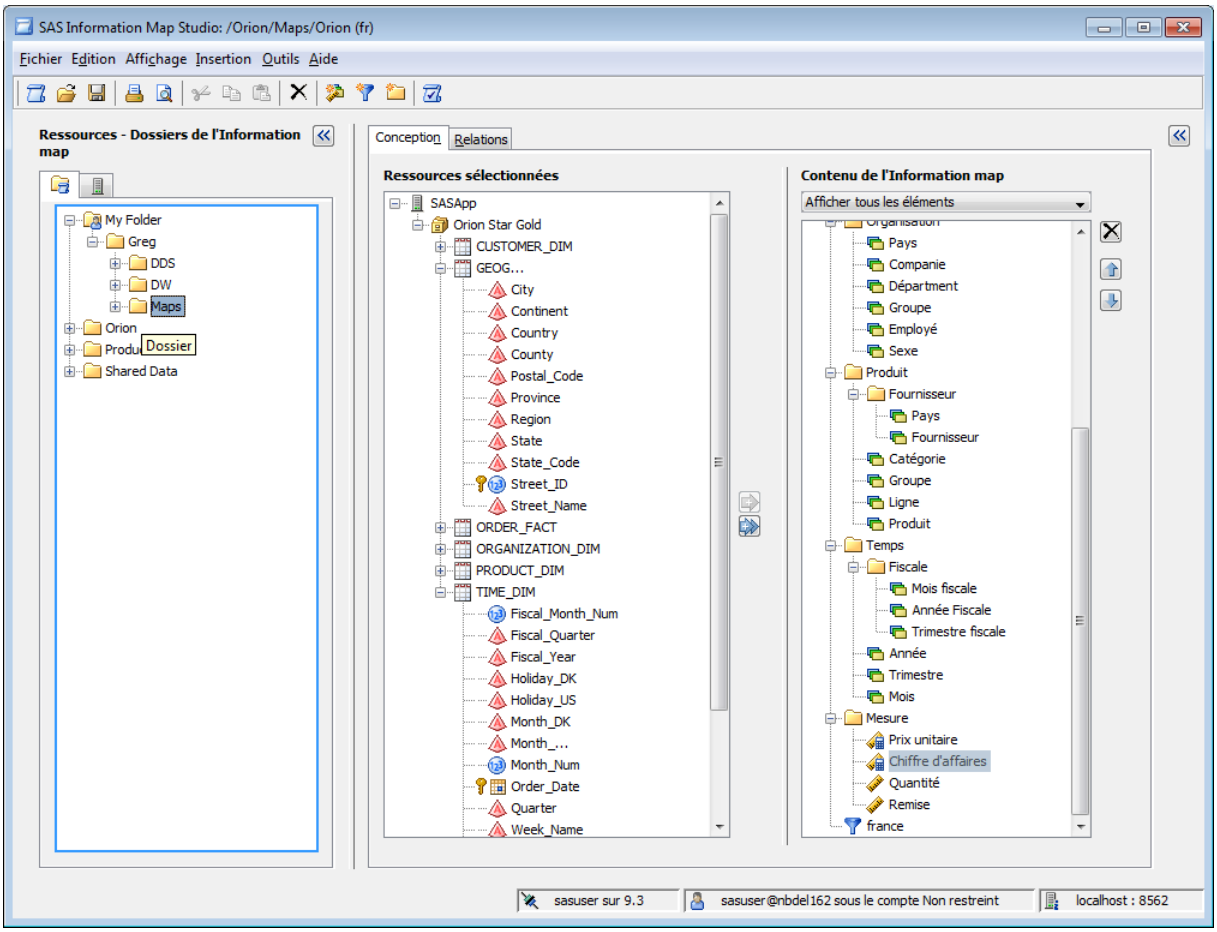

Nouvelle invite
Général
Type et valeurs de l'invite
Type d'invite :
Texte
Méthode d'alimentation de l'invite :
L'utilisateur saisit des valeurs
Nombre de valeurs :
Valeur unique
Longueur minimale :
Longueur maximale :
Inclure les valeurs spéciales
Valeurs manquantes
Valeur par défaut :
Astuce :
OK
Annuler
Aide

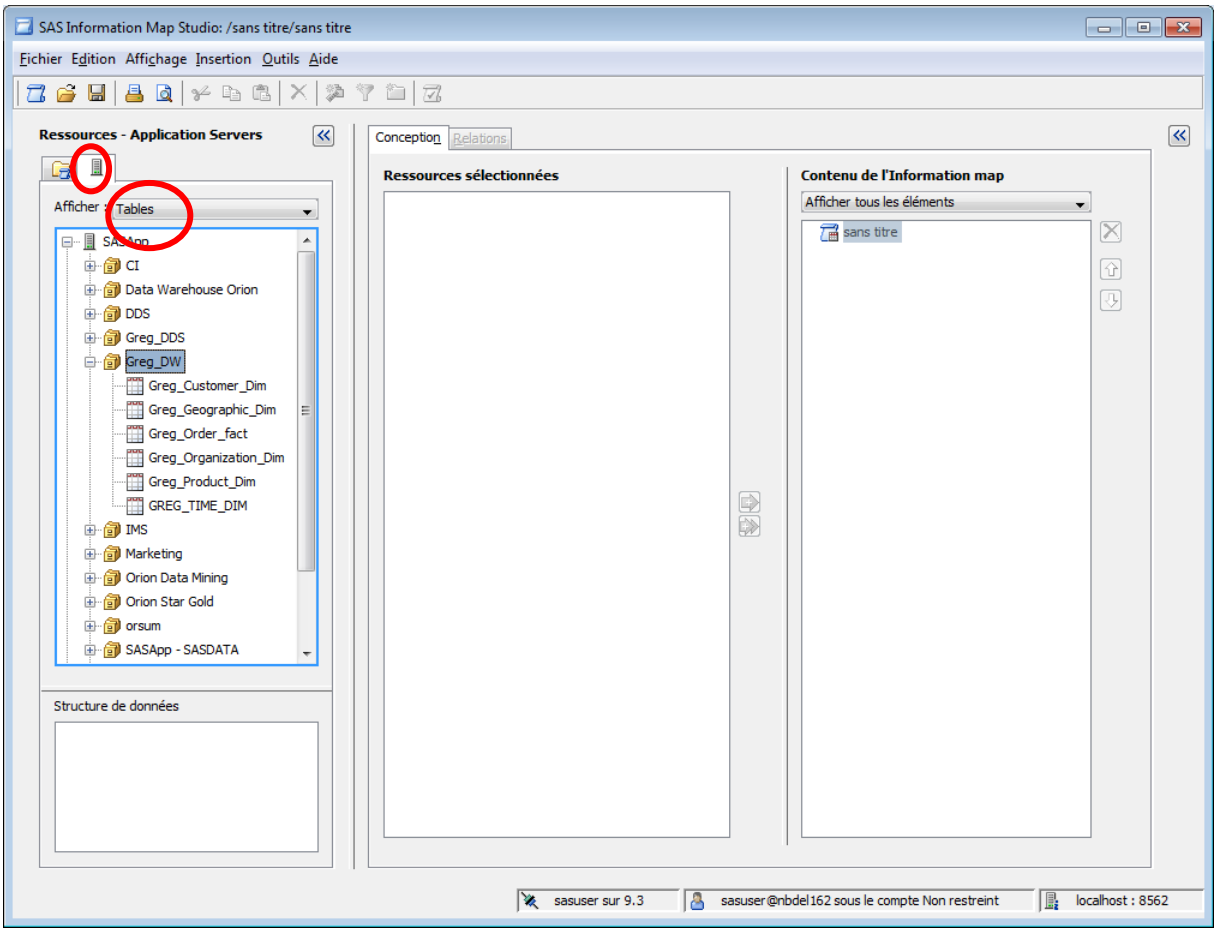


Mise en œuvre sur le cas Orion Star

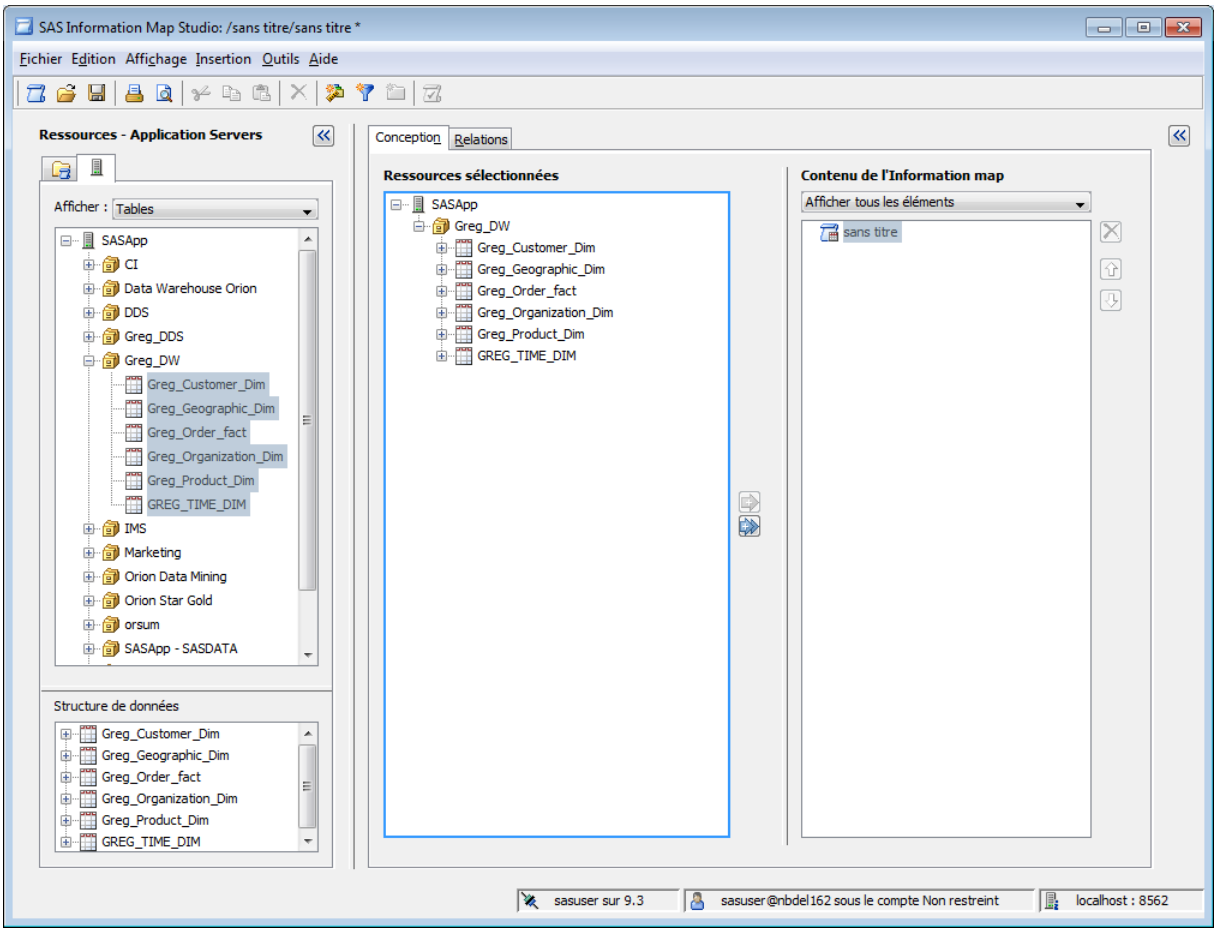
Nous allons, dans ce chapitre créer deux types de *Map*, l'une sur des données relationnelles, l'autre sur un cube multidimensionnel.

Lancer SAS *Information Map Studio*®, en sélectionnant dans Démarrer → Programmes → SAS →, SAS Information Map Studio


SAS Information Map Studio: /Orion/Maps/Orion (fr)
Fichier Edition Affichage Insertion Outils Aide
Ressources - Dossiers de l'Information map
My Folder
Greg
DDS
DW
Maps
Orion
Shared Data
Dossier
Conception
Relations
Ressources sélectionnées
SASApp
Orion Star Gold
CUSTOMER_DIM
GEOG...
City
Continent
Country
County
Postal_Code
Province
Region
State
State_Code
Street_ID
Street_Name
ORDER_FACT
ORGANIZATION_DIM
PRODUCT_DIM
TIME_DIM
Fiscal_Month_Num
Fiscal_Quarter
Fiscal_Year
Holiday_DK
Holiday_US
Month_DK
Month_...
Month_Num
Order_Date
Quarter
Week_Name
Contenu de l'Information map
Afficher tous les éléments
Pays
Companie
Département
Groupe
Employé
Sexe
Produit
Fournisseur
Pays
Fournisseur
Catégorie
Groupe
Ligne
Produit
Temps
Fiscale
Mois fiscale
Année Fiscale
Trimestre fiscale
Année
Trimestre
Mois
Mesure
Prix unitaire
Chiffre d'affaires
Quantité
Remise
france
sasuser sur 9.3
sasuser@nbdel162 sous le compte Non restreint
localhost : 8562

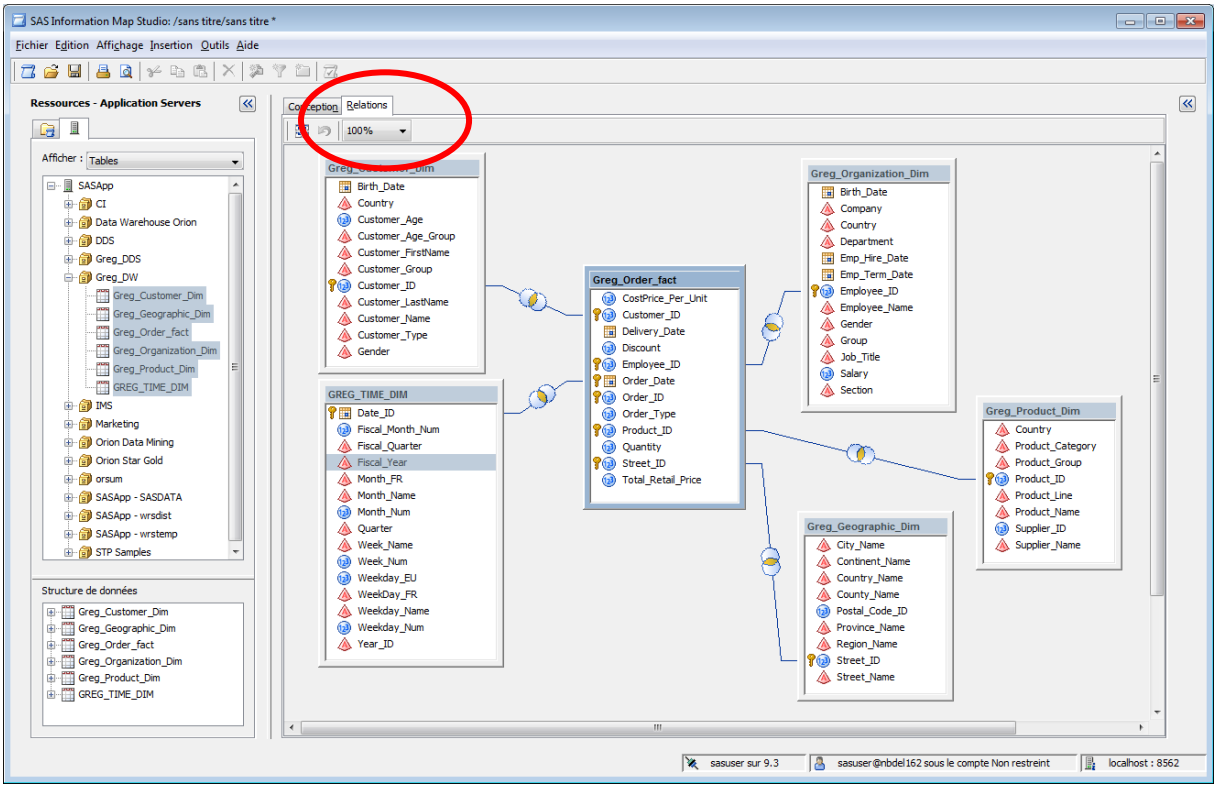


Dans ce serveur App,
Table

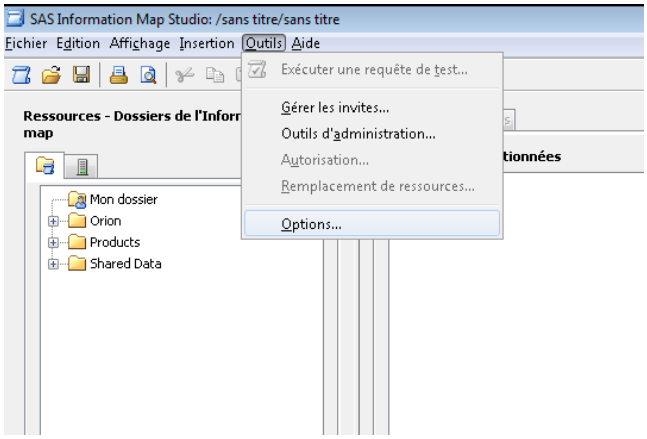


Dans la bibliothèque Orion Gold, sélectionner les 6 tables :

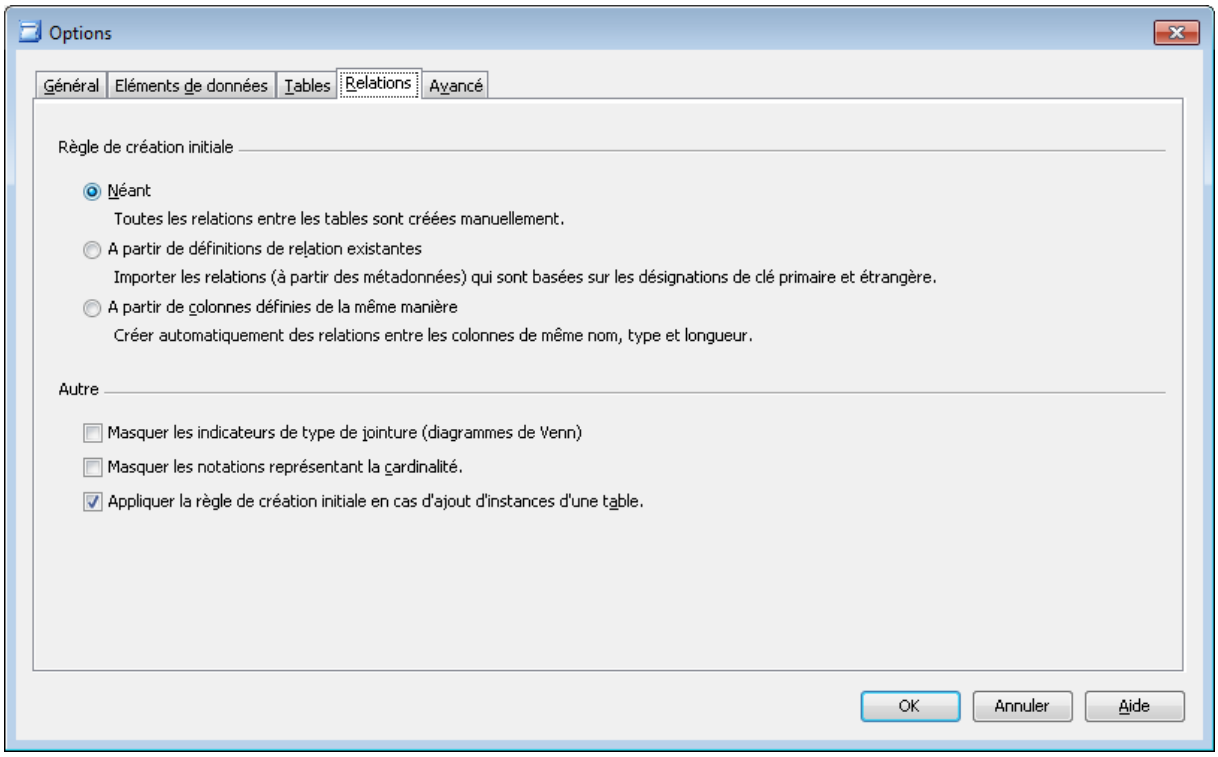
- order_fact
- customer_dim
- geography_dim
- organization_dim
- product_dim
- time_dim



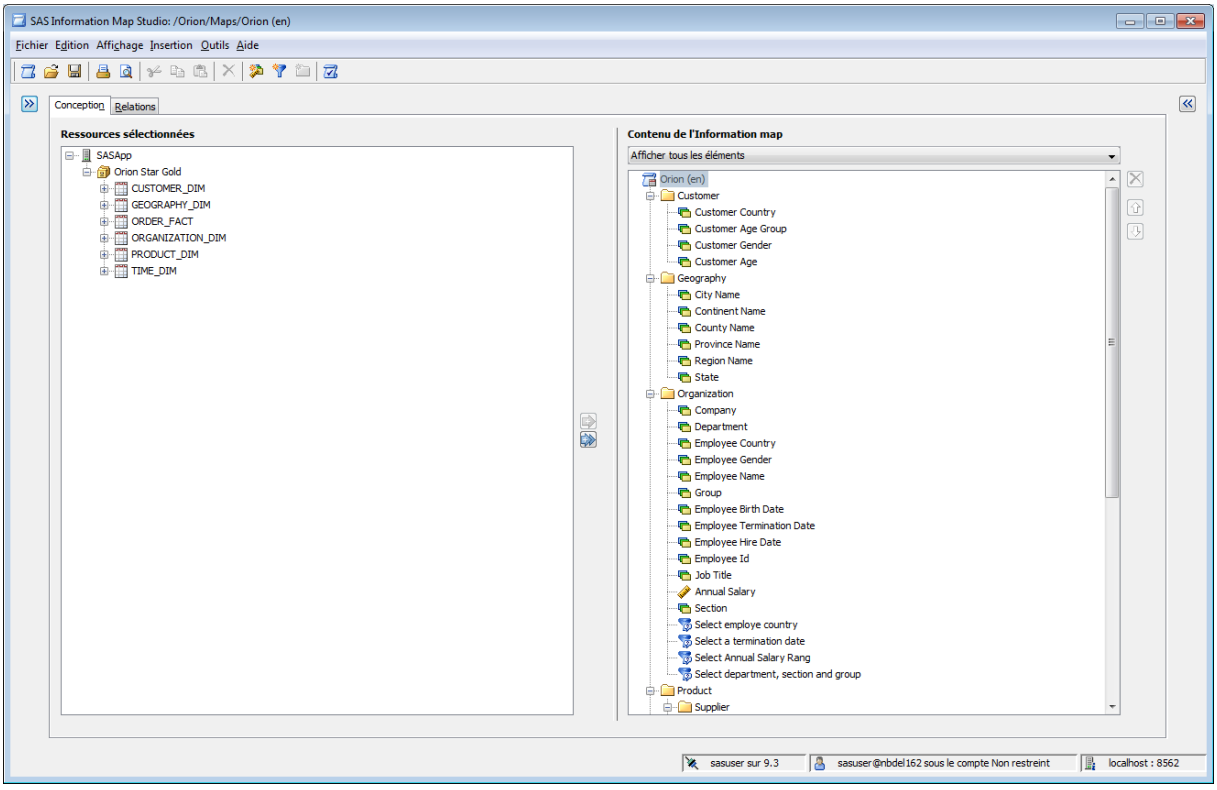
Cliquer sur l'onglet « Relations » pour définir les relations entre les tables.
Créer les 5 jointures sur les clés primaires des tables de dimension, par glisser-lâcher.

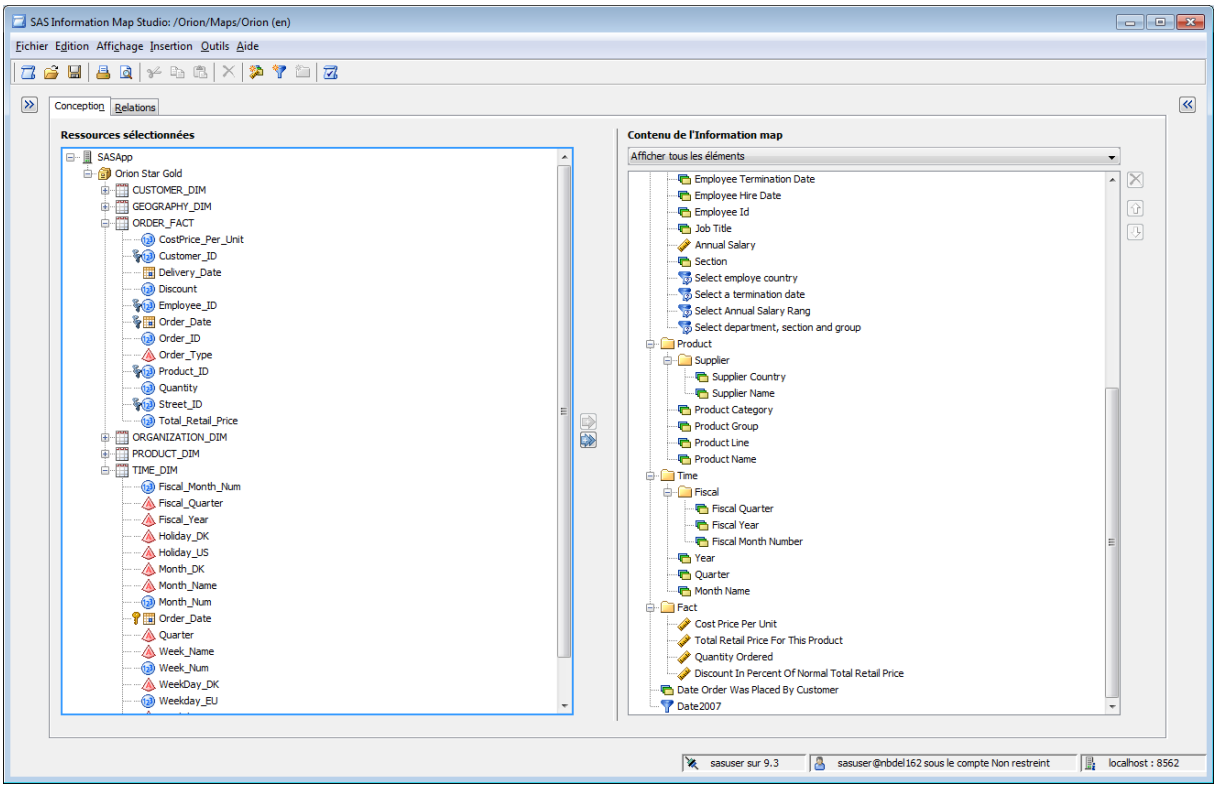


Dans les options, depuis le menu outil, la définition des règles est définie par défaut à néant. Si les métadonnées ont bien été définies dans les processus ETL notamment, il est plus pratique de les utiliser. La reconnaissance des jointures possible par colonne de nom et de type équivalent est à prendre avec précautions.



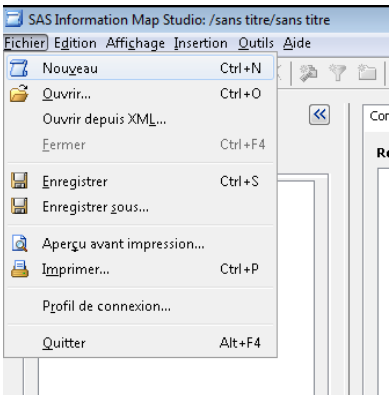
Revenir dans l'onglet présentation pour créer les éléments de la *Map*.
Avec un clic-droit dans la *Map*, créer les répertoires permettant de regrouper les catégories par thème. Ensuite, par glisser-lâcher, prendre les éléments de la partie physique (à gauche) et remplir les dossiers précédemment créés.

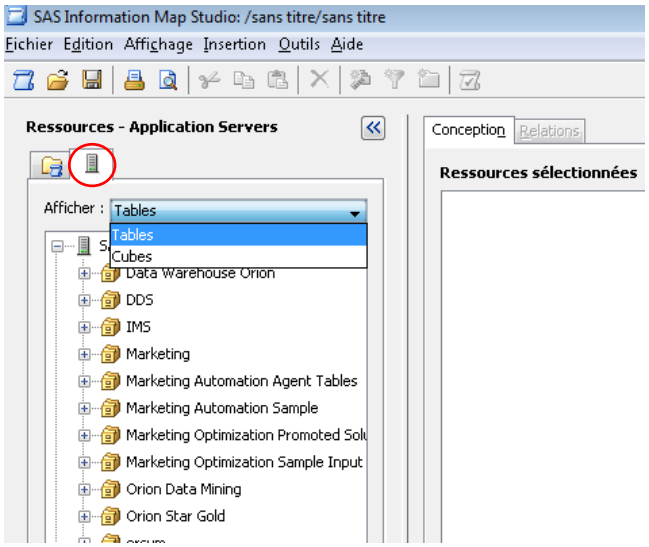




Grâce à un clic-droit sur chaque élément, aller dans les propriétés de l'élément pour changer le nom et la classification.
Le nom de l'élément est celui qui sera vu par l'utilisateur final, il peut comprendre des espaces et des caractères accentués.
Si l'on veut utiliser la variable « âge du client » comme variable de classe, aller dans le deuxième onglet et sélectionner « catégorie ».

Pour créer une nouvelle Map : Fichier option → Nouveau

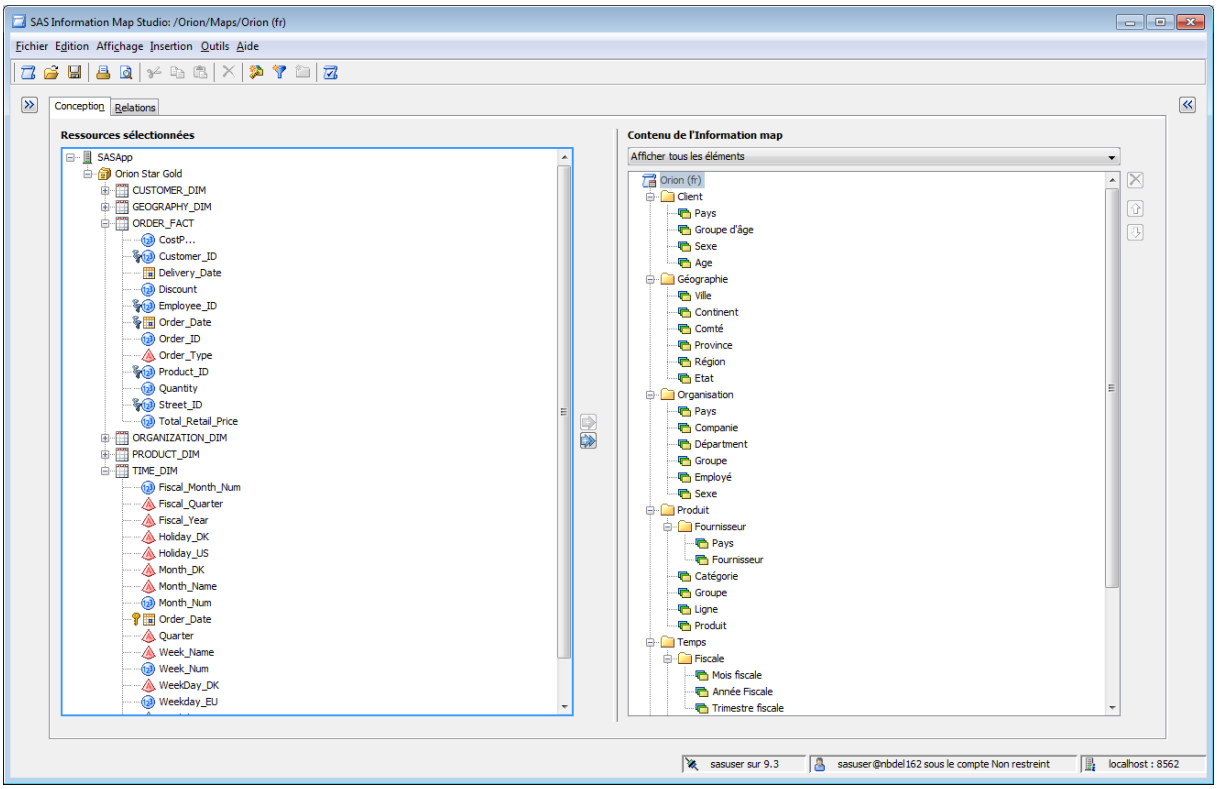


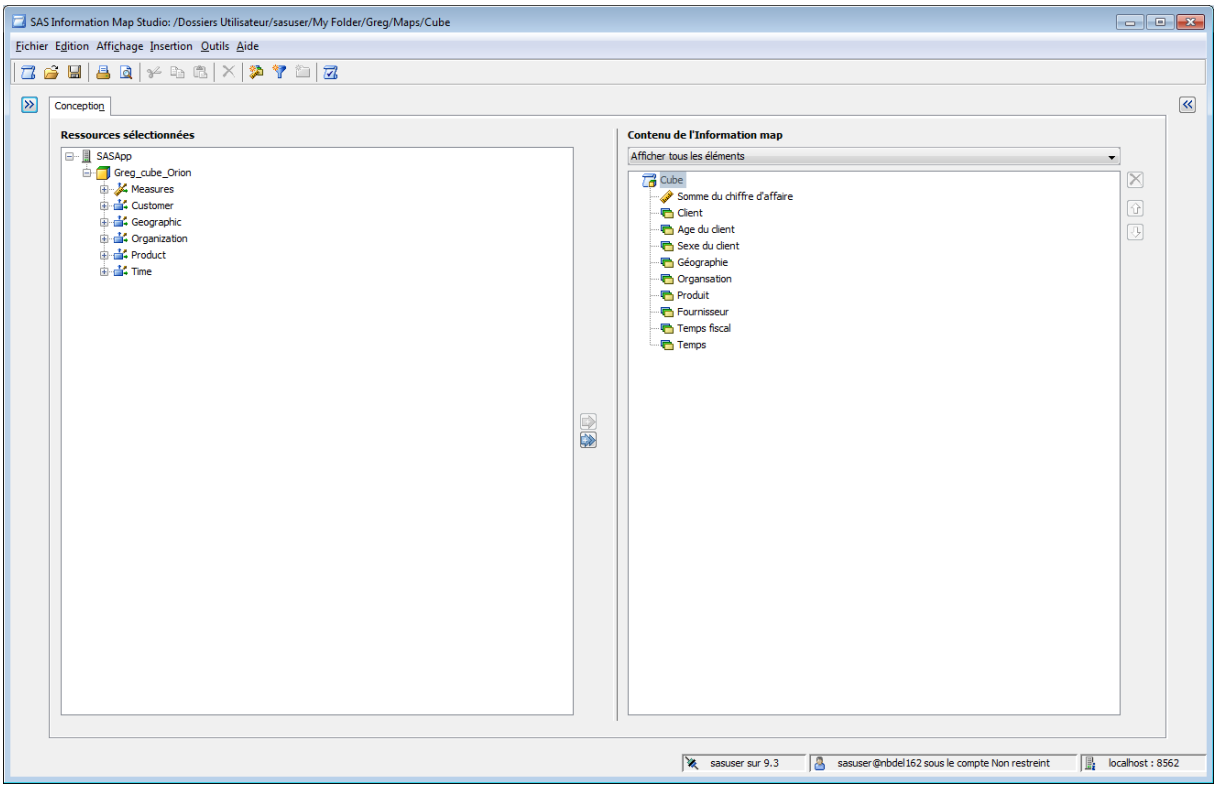


Dans l'onglet des ressources → vous pouvez sélectionner le type de données physiques que vous souhaitez utiliser dans votre Map : soit des tables ou soit des cubes. Une information Map sur des tables génèrera des requêtes SQL. Une information Map sur un cube génèrera des requêtes MDX.

Création des Map :

Il y a quatre Information Map à créer : deux sur le cube et deux sur les tables du schéma en étoile ; croisées avec deux en anglais et deux en français.



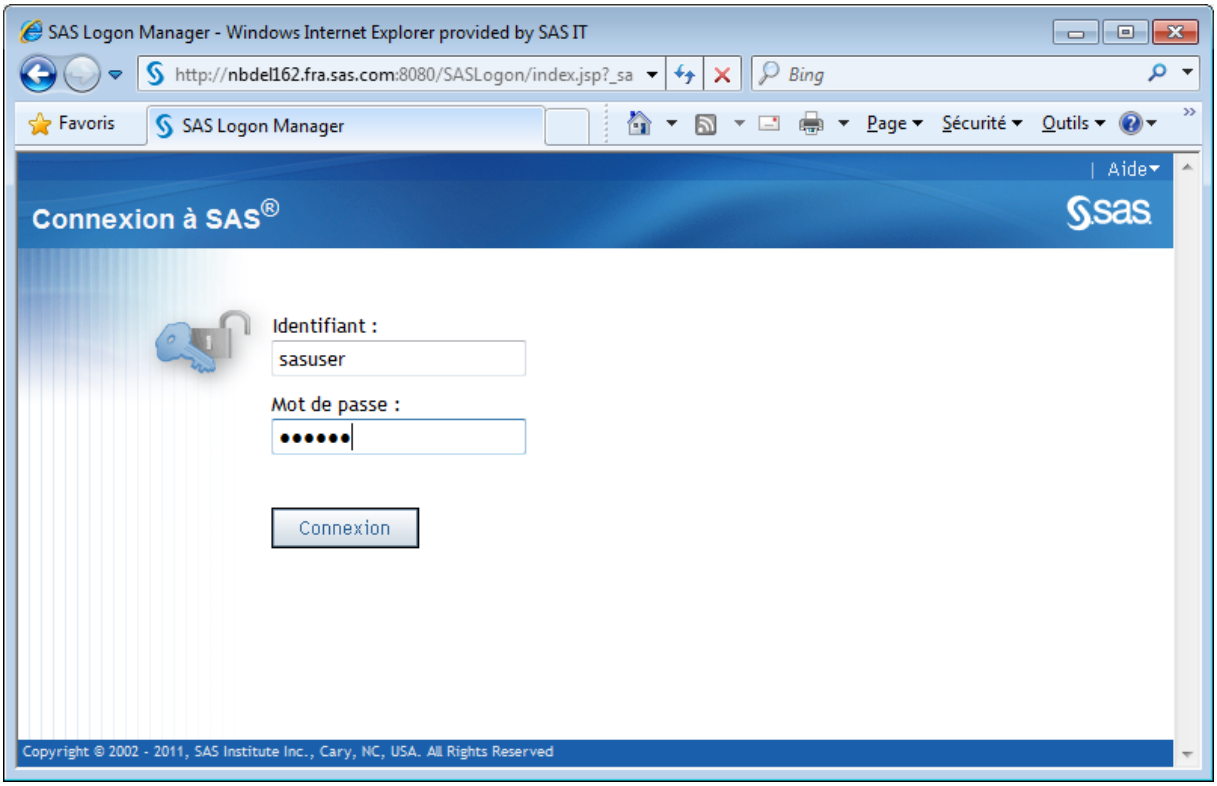

SAS Information Map Studio: /Dossiers Utilisateur/sasuser/My Folder/Greg/Maps/Cube
Fichier Edition Affichage Insertion Outils Aide
Conception
Ressources sélectionnées
SASApp
Greg_cube_Orion
Measures
Customer
Geographic
Organization
Product
Time
Contenu de l'Information map
Afficher tous les éléments
Cube
Somme du chiffre d'affaire
Client
Age du client
Sexe du client
Géographie
Organsation
Produit
Fournisseur
Temps fiscal
Temps
sasuser sur 9.3
sasuser@nbdel162 sous le compte Non restreint
localhost : 8562

Création de rapport avec SAS Web Report Studio

Introduction :

Nous allons dans le cadre de ce TP, créer un rapport avec l'outil de dit de Reporting de masse, SAS Web Report Studio. Un outil de Reporting de masse est par définition un outil permettant aux masses, des utilisateurs ayant généralement peu ou prou de connaissance informatique, de créer simplement, intuitivement leur rapport, de les partager et de les distribuer.
SAS Web Report Studio est un client léger, c'est-à-dire qu'il n'y a rien à installer sur le poste de l'utilisateur. Il suffit qu'il ait le navigateur Web Microsoft Internet Explorer et une connexion au serveur. Cette application est suffisamment intuitive pour qu'avec une simple prise en main, voire pas du tout, un utilisateur non-informaticien puisse créer et partager son rapport. Il n'est donc plus nécessaire d'être un spécialiste pour créer de jolis rapports.

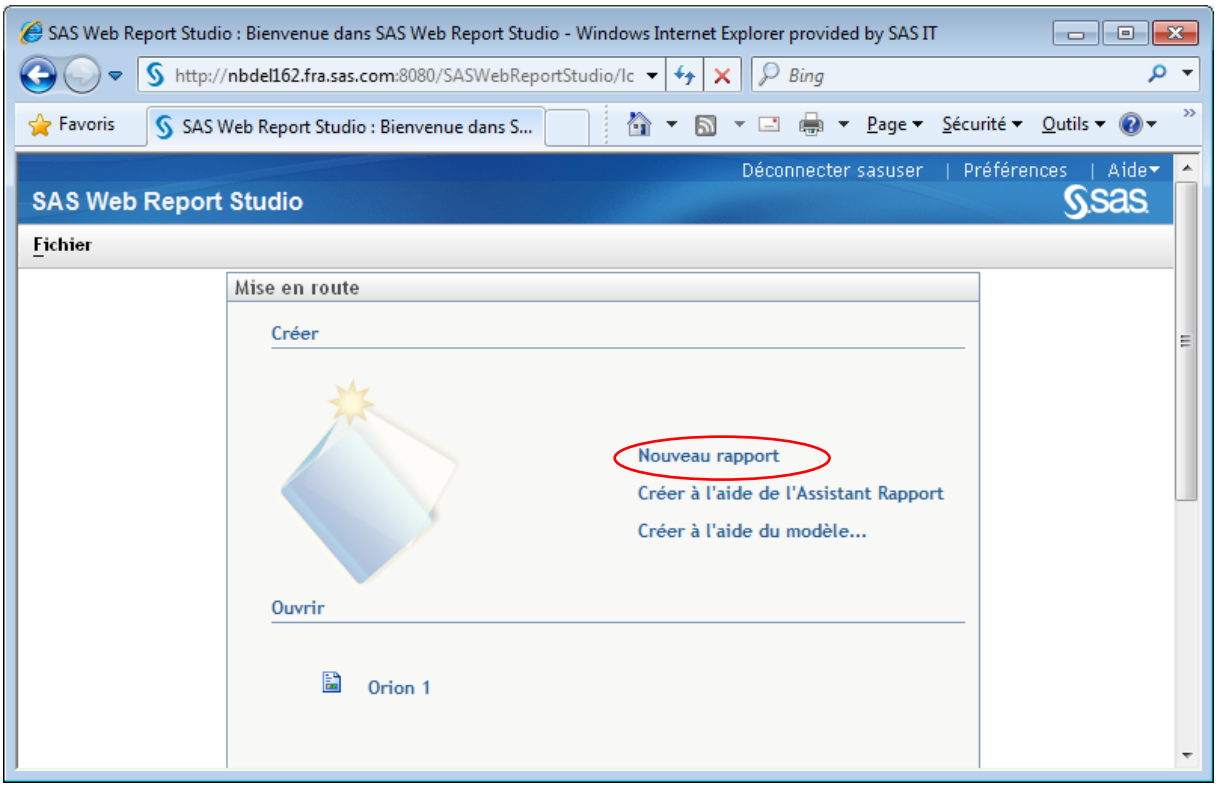
Ouvrir SAS Web Report Studio : http://nonduserveur:8080/SASWebReportStudio



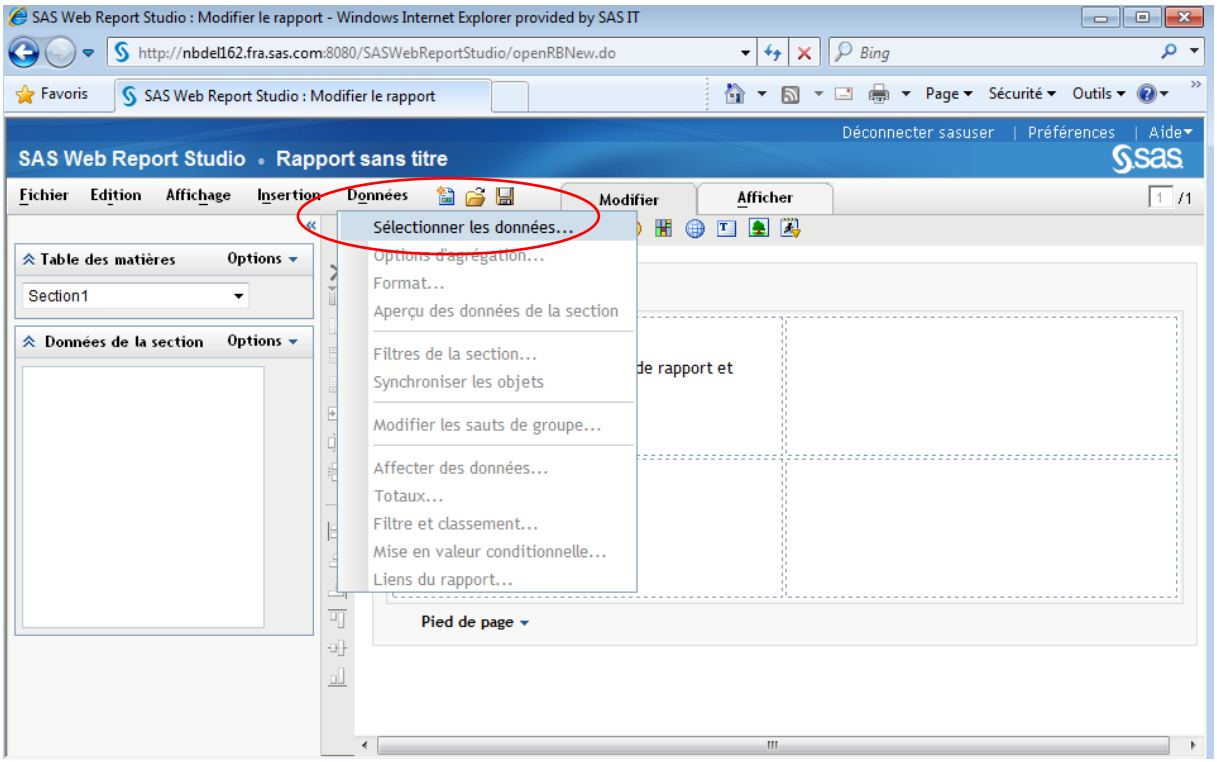
Nous utiliserons aussi notre utilisateur générique « eleve » et son mot de passe = « SASpw1 » (ou parfois sasdemo – SASpw1 ou saspw1ou Orion 123)

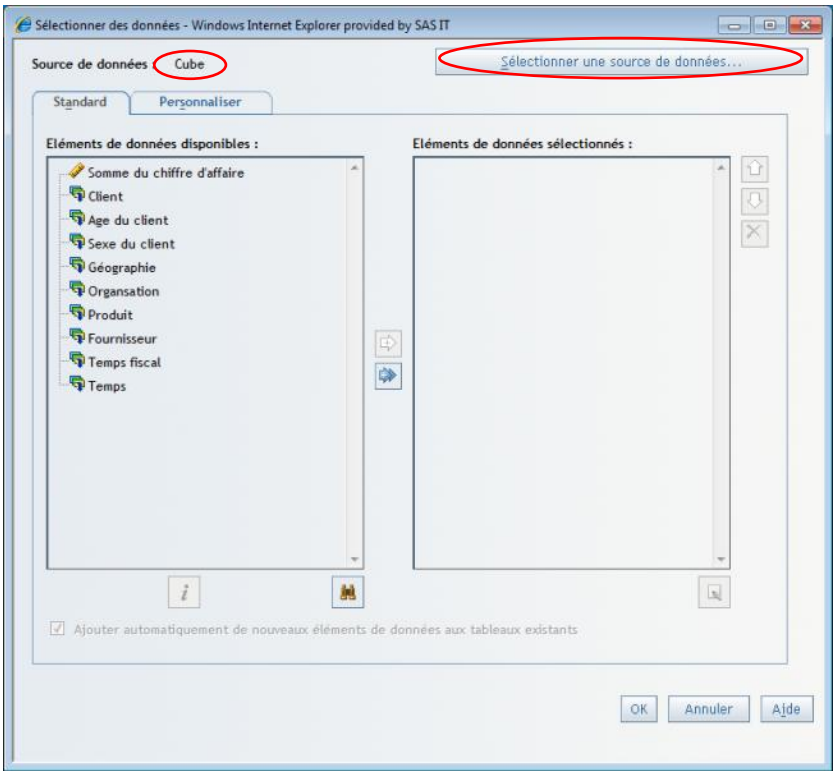
Connexion

Cliquez sur « nouveau rapport »,

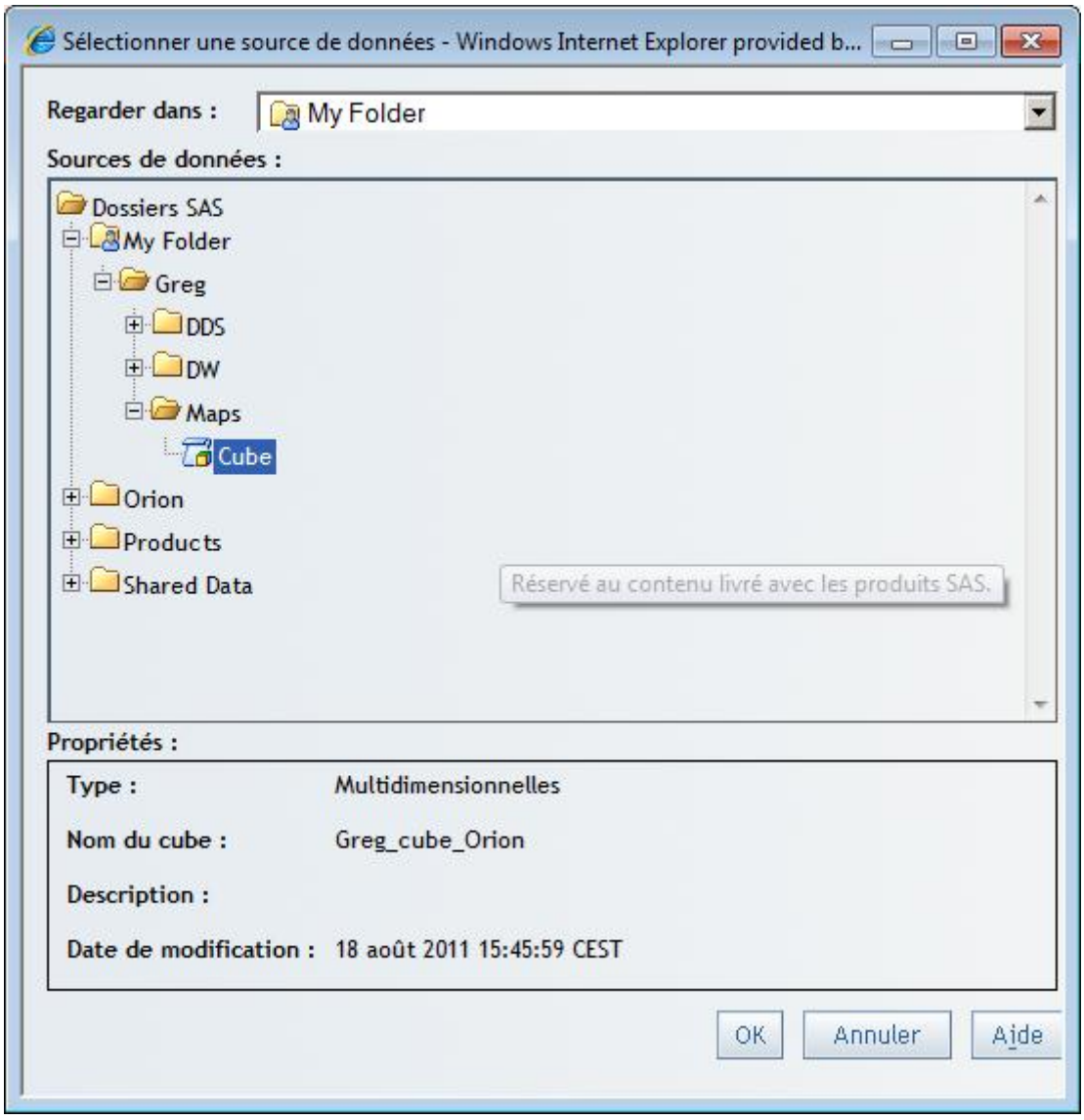


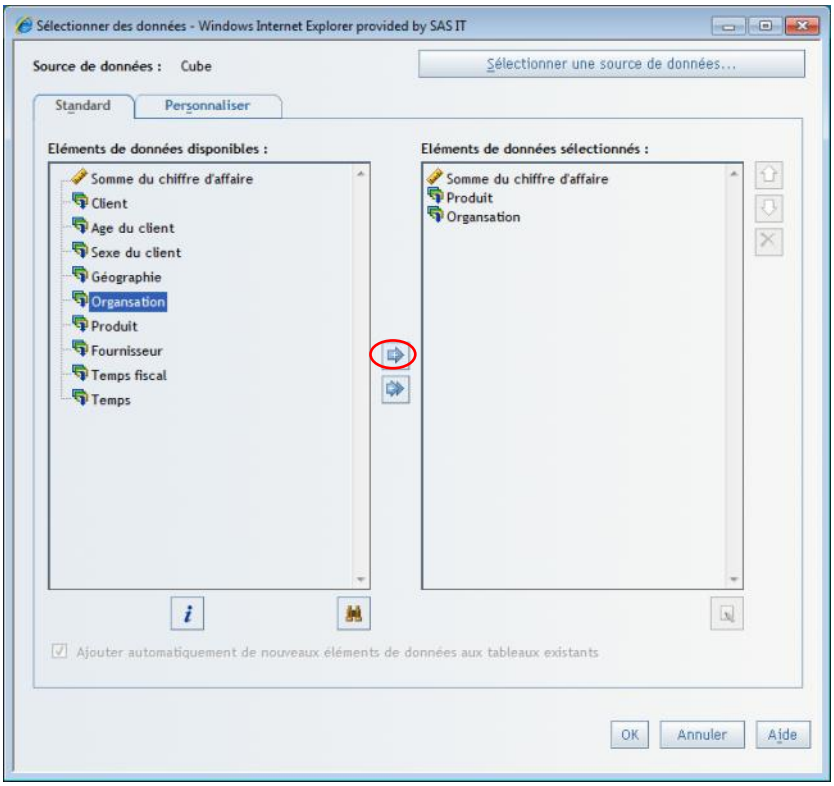
Dans le menu Données, « sélectionner des données… »



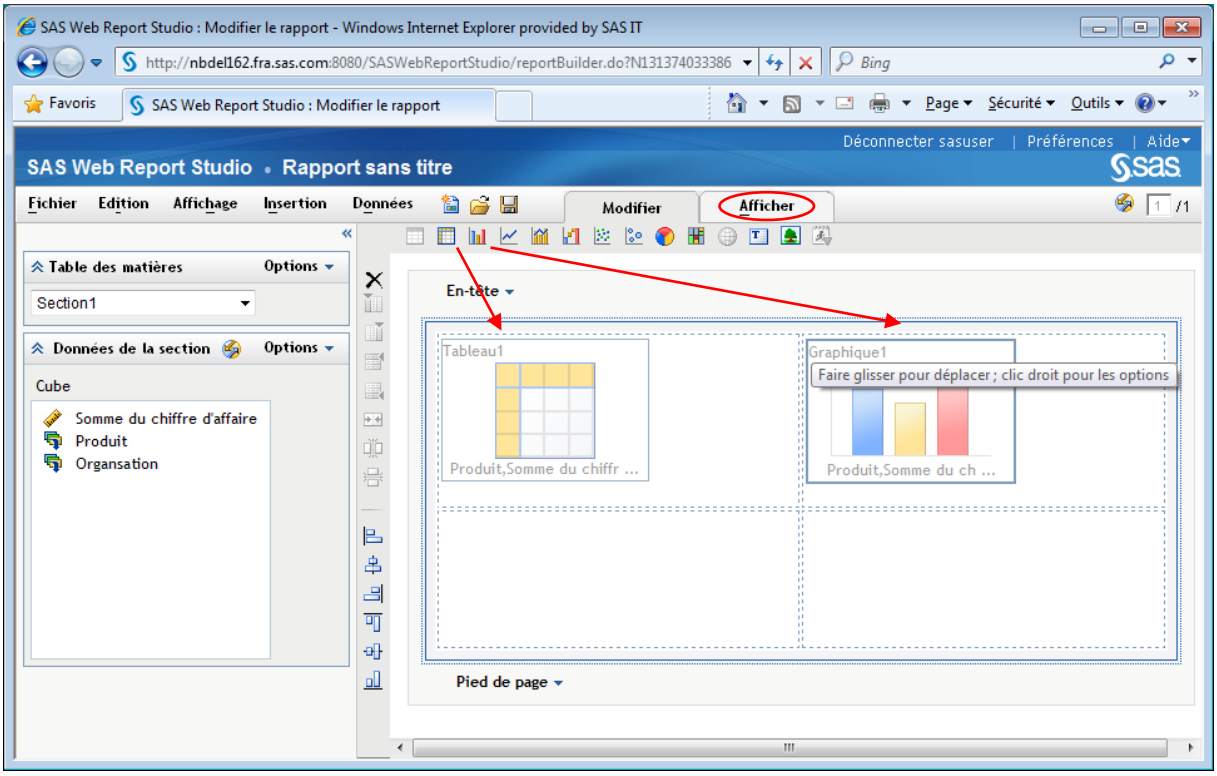


Si ce n'est pas la bonne information Map, cliquez sur le bouton « Changer la source » et sélectionnez votre Information Map.

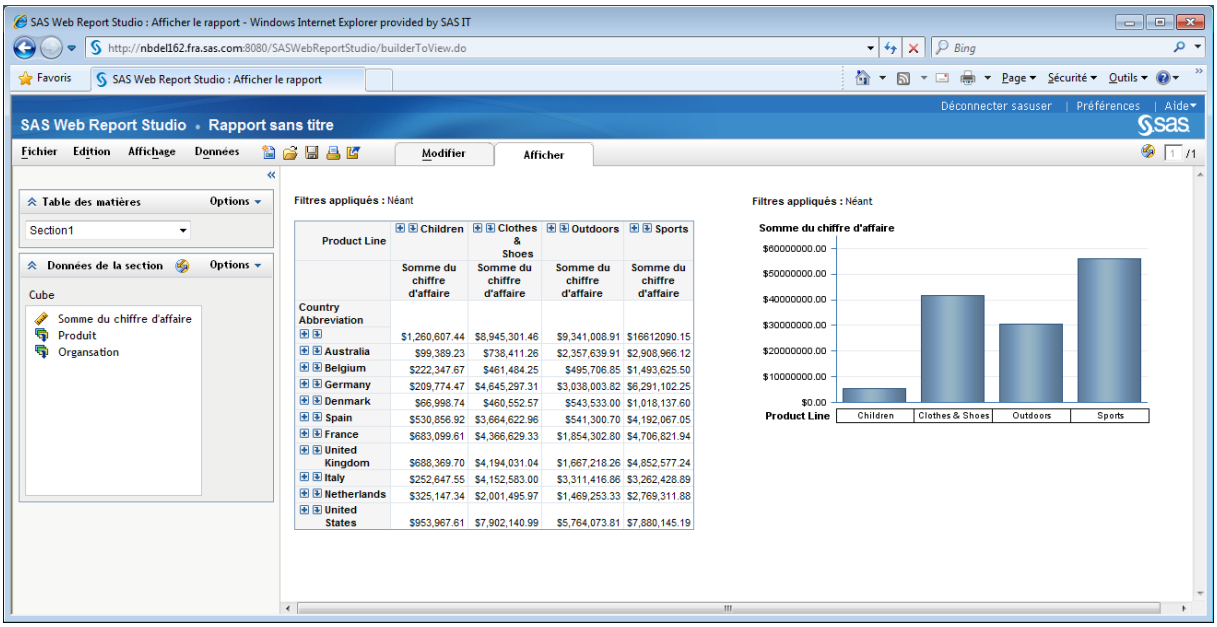

Sélectionner une source de données - Windows Internet Explorer provided b...
Regarder dans :
My Folder
Sources de données :
Dossiers SAS
My Folder
Greg
DDS
DW
Maps
Cube
Orion
Products
Shared Data
Réservé au contenu livré avec les produits SAS.
Propriétés :
Type : Multidimensionnelles
Nom du cube : Greg_cube_Orion
Description :
Date de modification : 18 août 2011 15:45:59 CEST
OK
Annuler
Aide



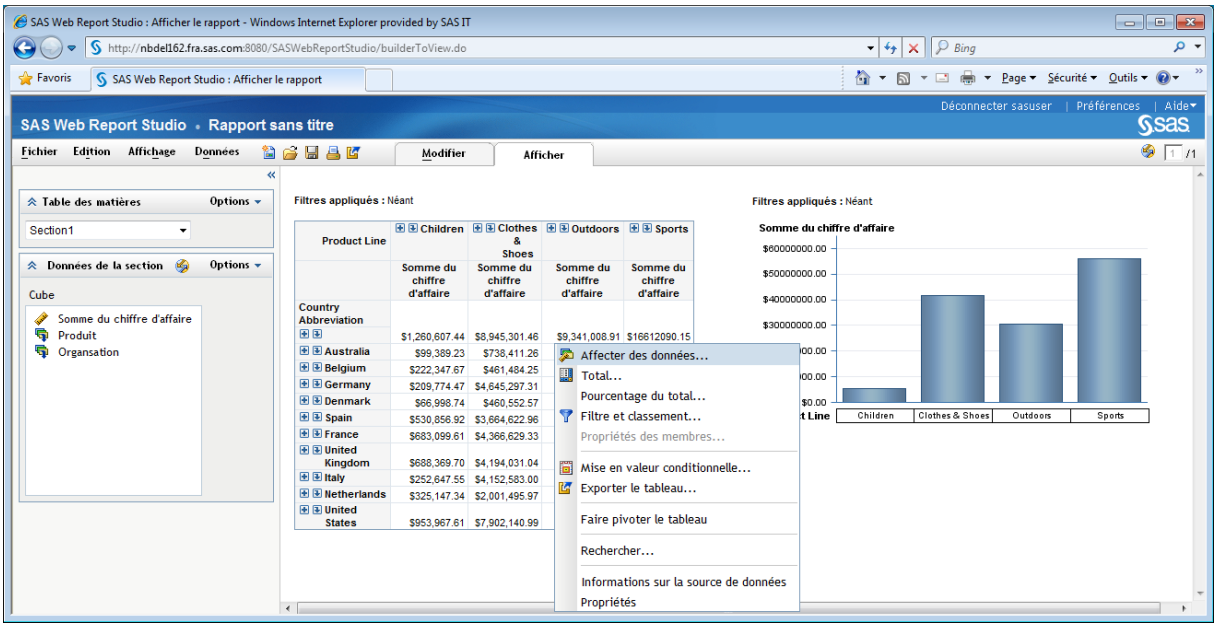
Sélectionner la mesure du chiffre d'affaires et les hiérarchies produit et organisation.
Cliquer sur OK



Ajoutez à votre rapport un tableau croisé et un histogramme par glisser-lâcher et cliquer sur « afficher ».



Voici votre premier rapport avec SAS Web Report Studio.



Faire un clic-droit sur votre tableau pour faire des filtres, des classements, ajouter des pourcentages et des totaux, en ligne ou en colonne, mettre en valeur conditionnellement, etc…

Pour ajouter des entêtes, pieds de page, sections à votre rapport, il faut revenir dans l'onglet « modifier ».

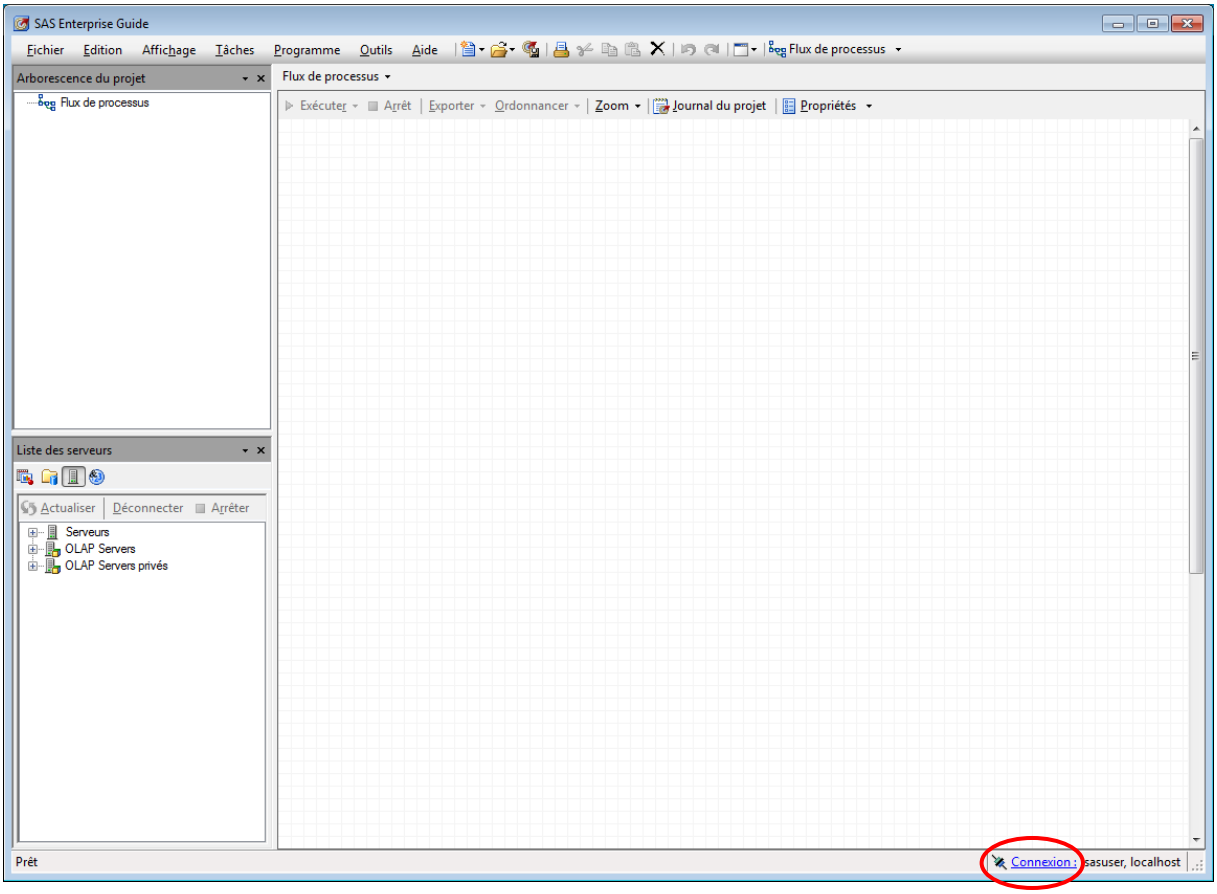
Pour exporter votre rapport, fichier → exporter. Cela vous donnera un fichier zip de votre travail.

Création d'une procédure stockée

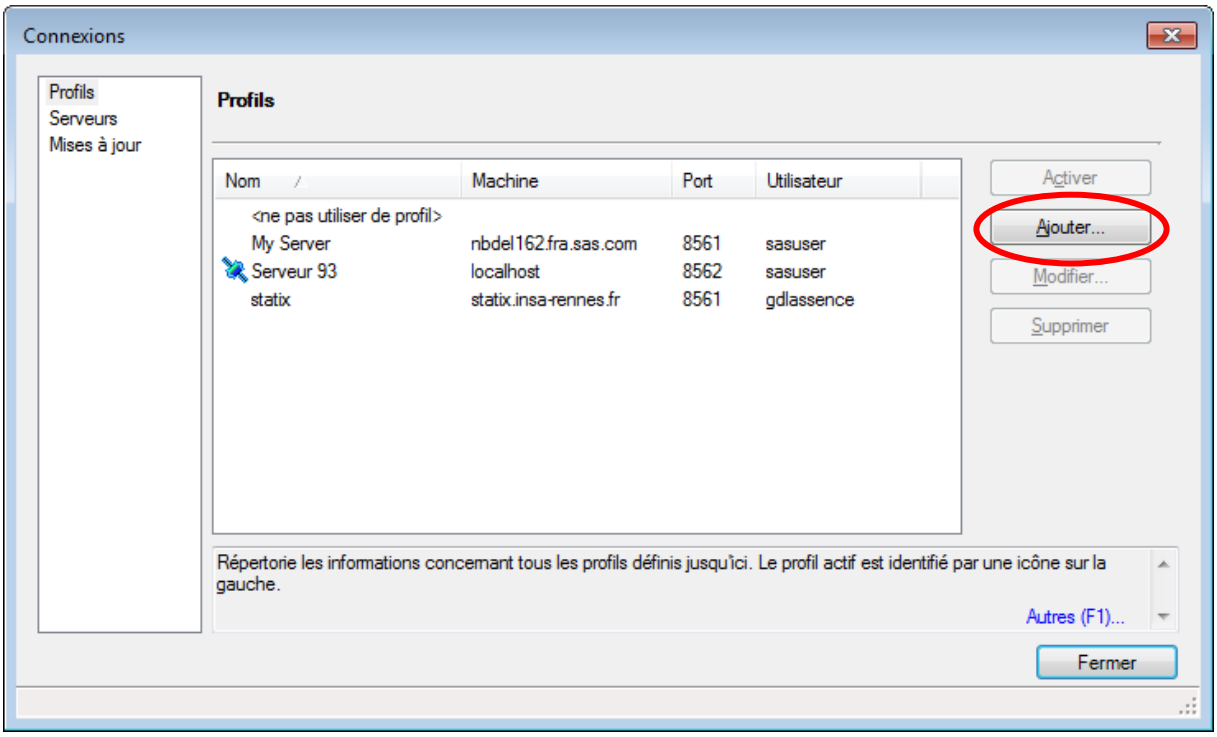
La procédure stockée que nous allons créer est une procédure, qui après avoir proposé à l'utilisateur de sélectionner un paramètre, le pays, fait la prévision du chiffre d'affaires de ce pays.

1. On prend ici les tables de la bibliothèque « Orion Gold » du serveur SASApp.
2. On fait une requête pour avoir la somme du chiffre d'affaires par pays du commercial, par année et par numéro du mois.
3. On tri par année et par mois.
4. On fait un filtre sur un pays.
5. On fait la prévision du chiffre d'affaires.
6. On récupère la liste des pays dans une table.
7. On récupère le code de la requête et celui de la prévision.
8. On les retraite.
9. On crée la procédure stockée.

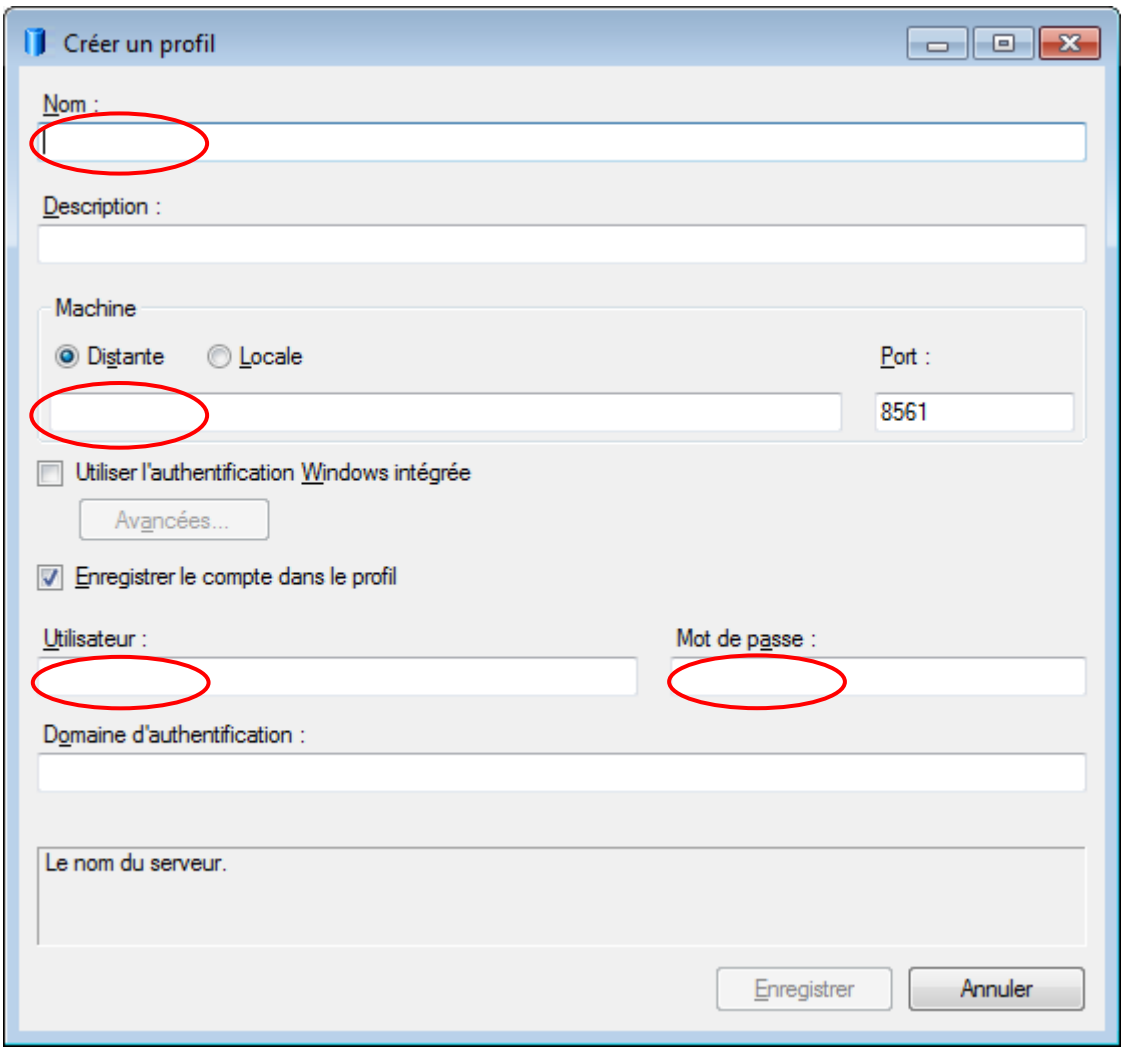
Ouvrir SAS Enterprise Guide
Depuis Démarrer → Programmes → SAS → Enterprise Guide



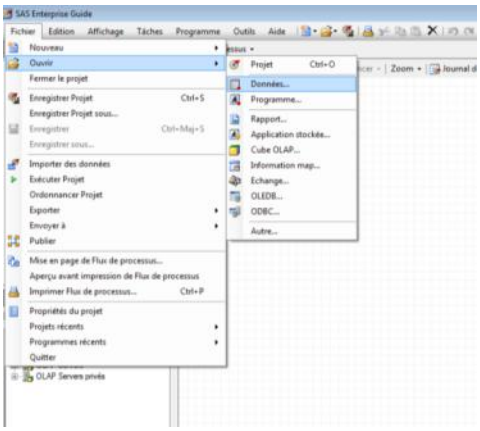
Il faut que vous soyez connecté au serveur de métadonnée. Si vous n'avez pas connexion en bas à droite de la fenêtre d'Enterprise Guide, Cliquer sur « pas de connexion »



S'il n'y a pas de connexion, ajoutez-en une :

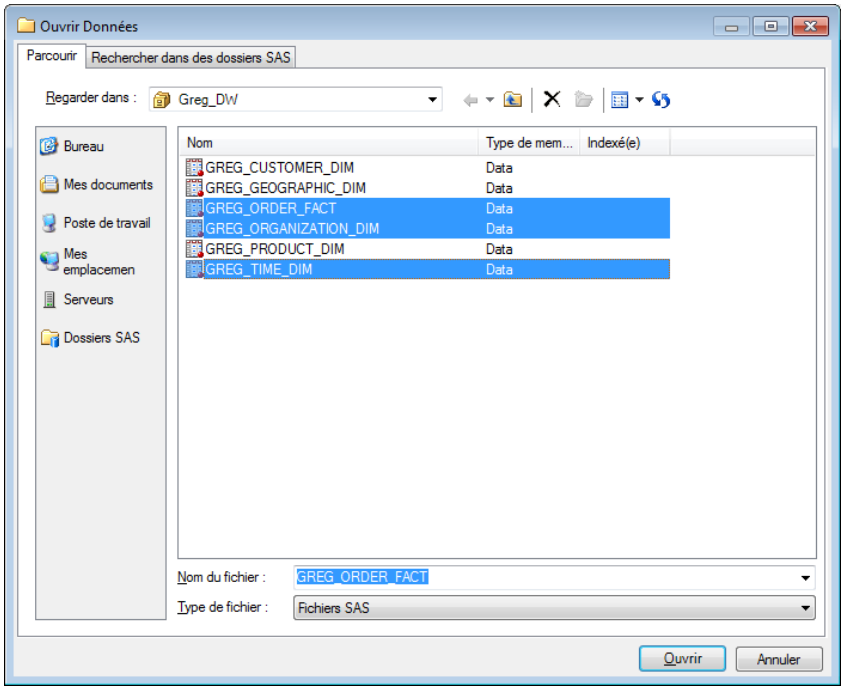


Il faut rentrer un nom de connexion, le nom du serveur, l'utilisateur et son mot de passe. Enregistrer.

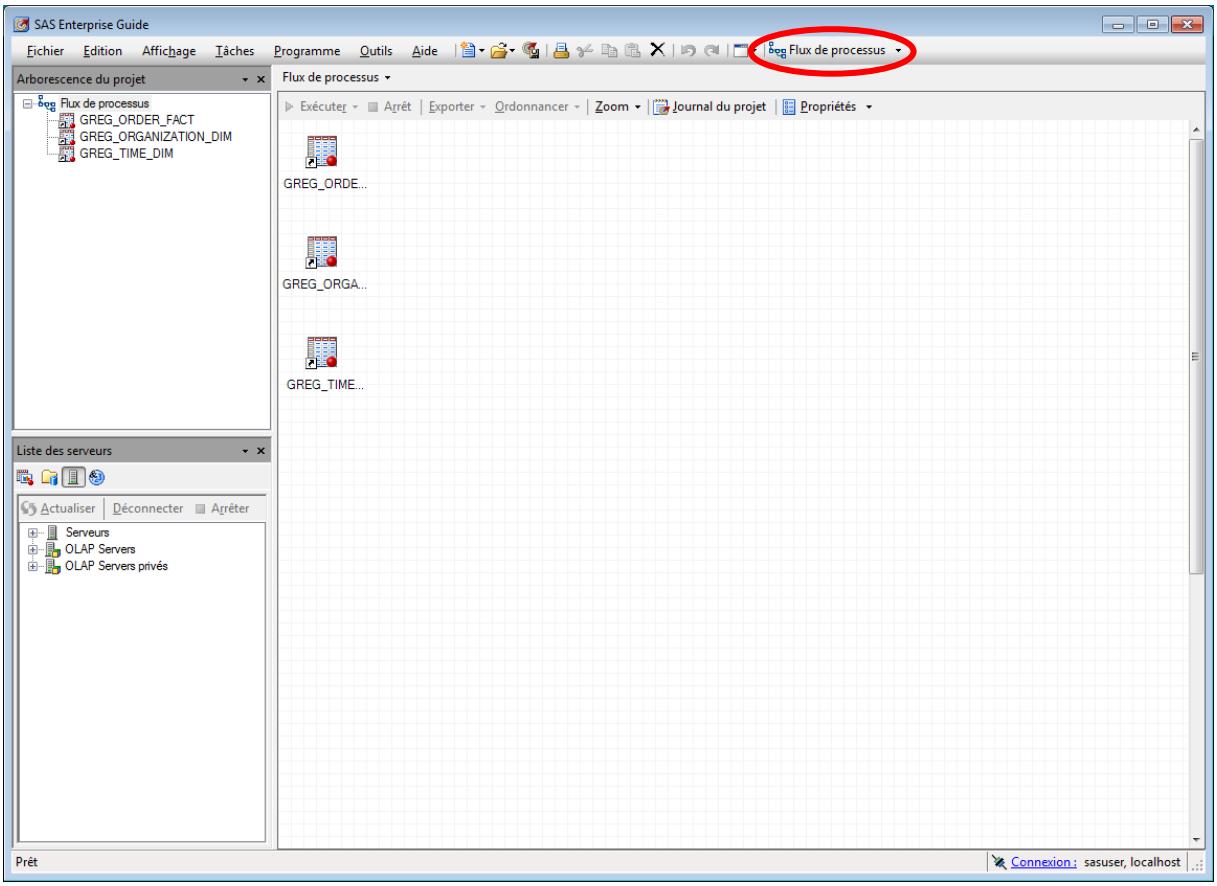


Fichier → Ouvrir → données

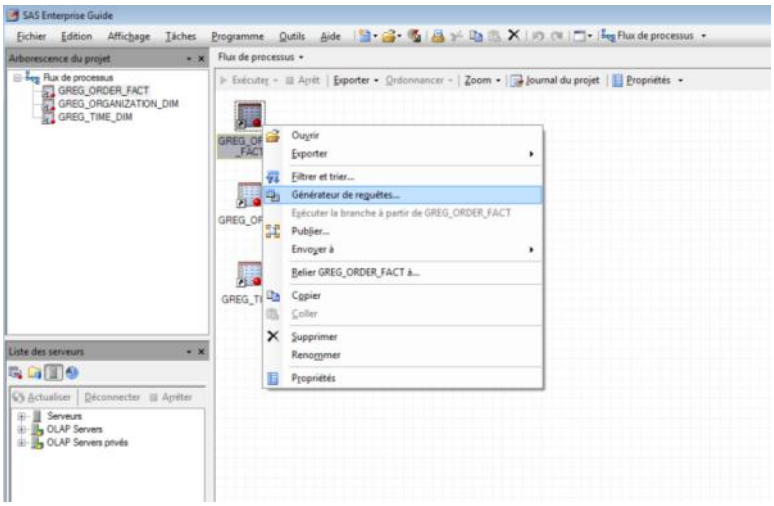
Serveur → SASApp → Votre bibliothèque (ou « Orion Star Gold » si vous n'en avez pas)



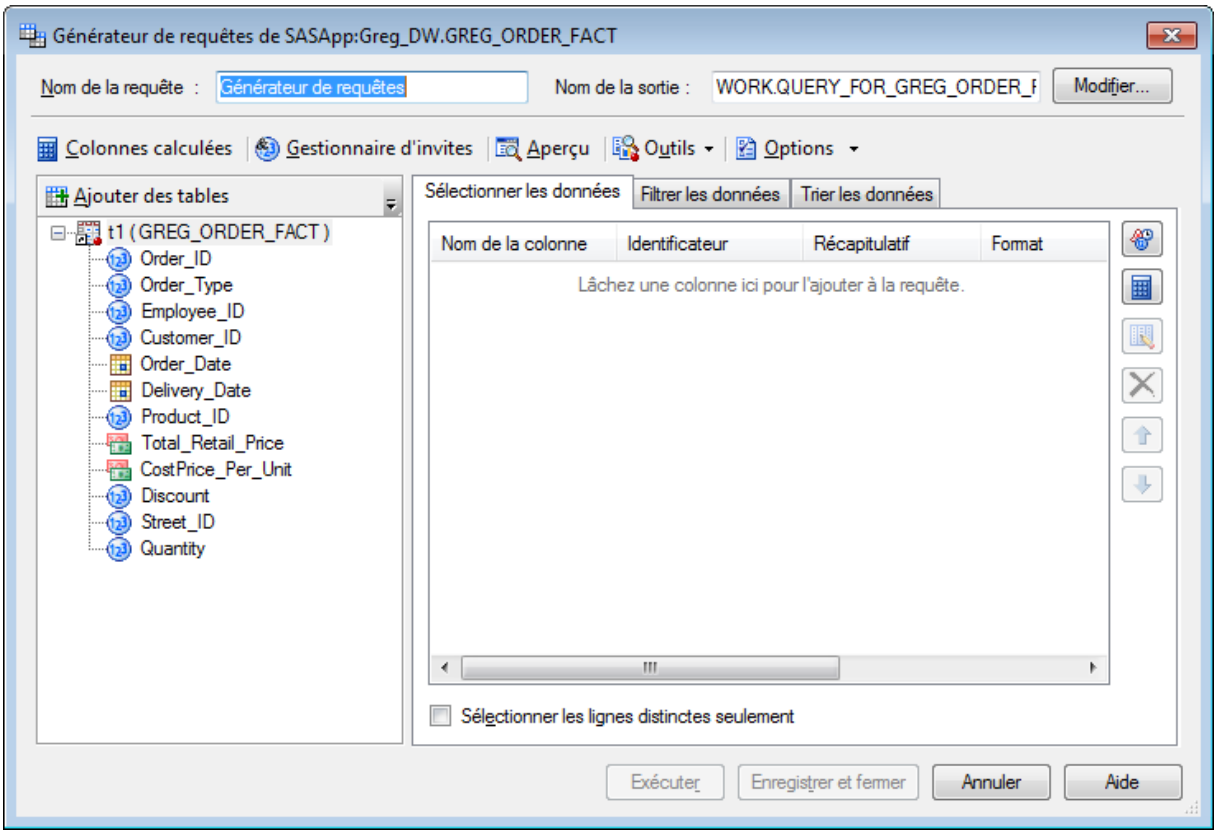
Sélectionner la table de fait et les tables de dimensions de l'organisation et du temps.



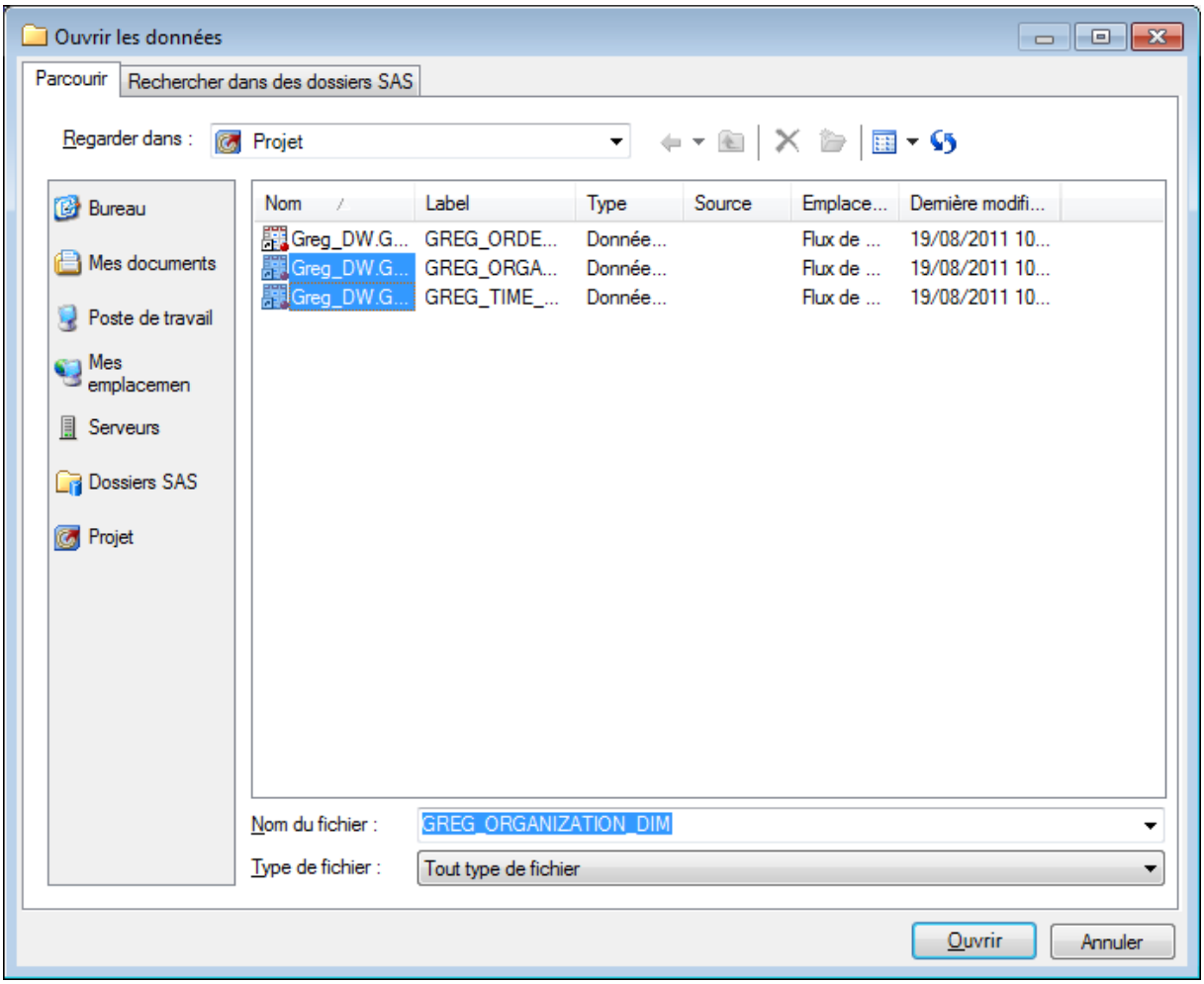
Pour revenir au flux de processus, cliquer sur le bouton du même nom.



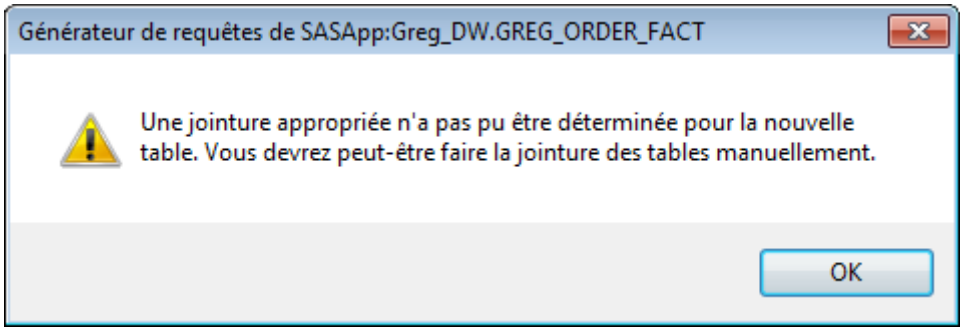
Clic-droit sur la table Order_Fact → sélectionner le générateur de requêtes



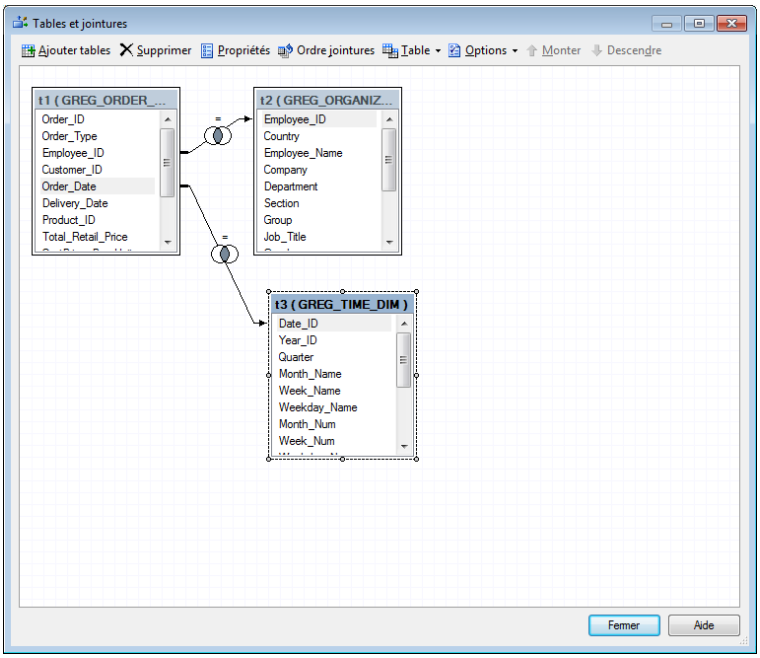
Ajoutez les tables de l'organisation et du temps depuis votre projet,

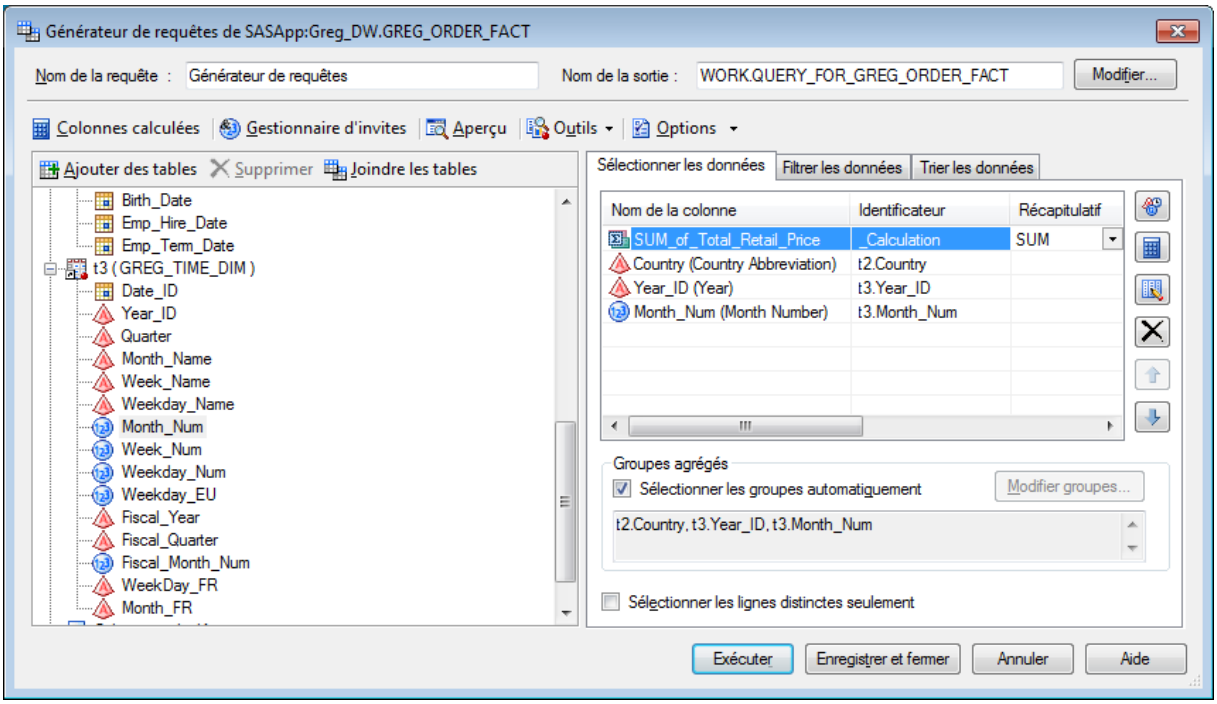


Ouvrir

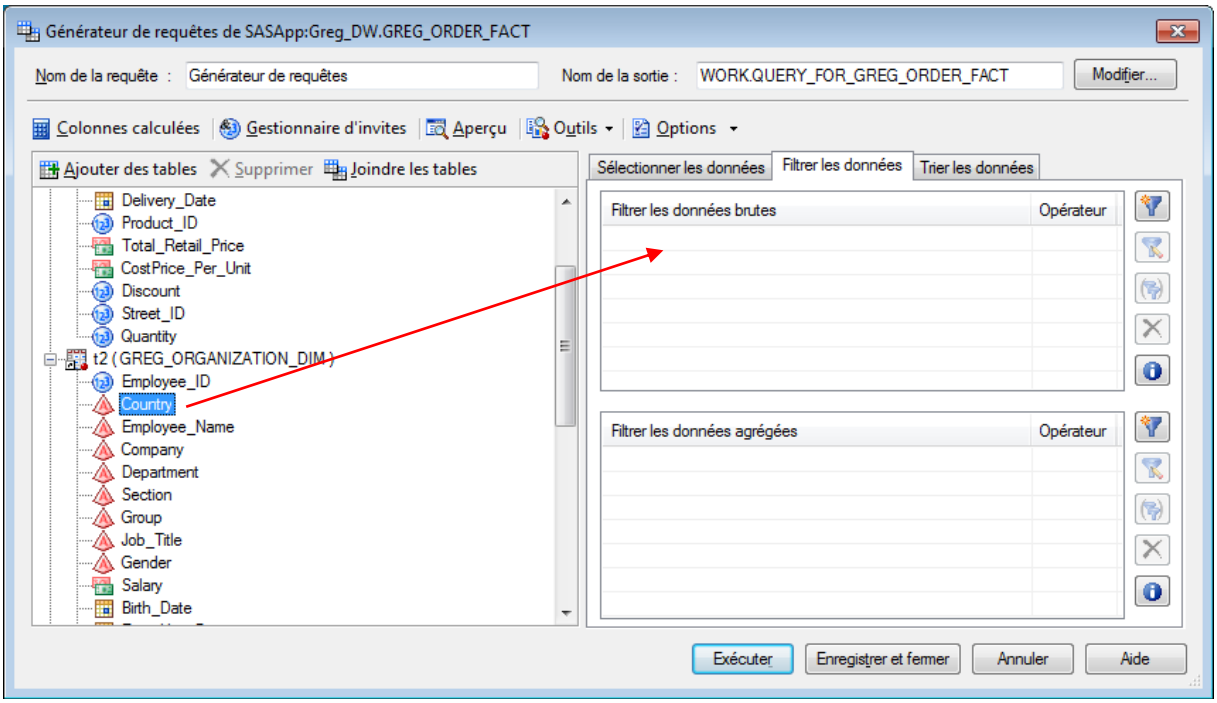


Si vous utilisez vos tables, les jointures par défaut se faisant sur les colonnes de même nom, il faut faire le lien entre la table de fait et celle de la dimension temps à la main.

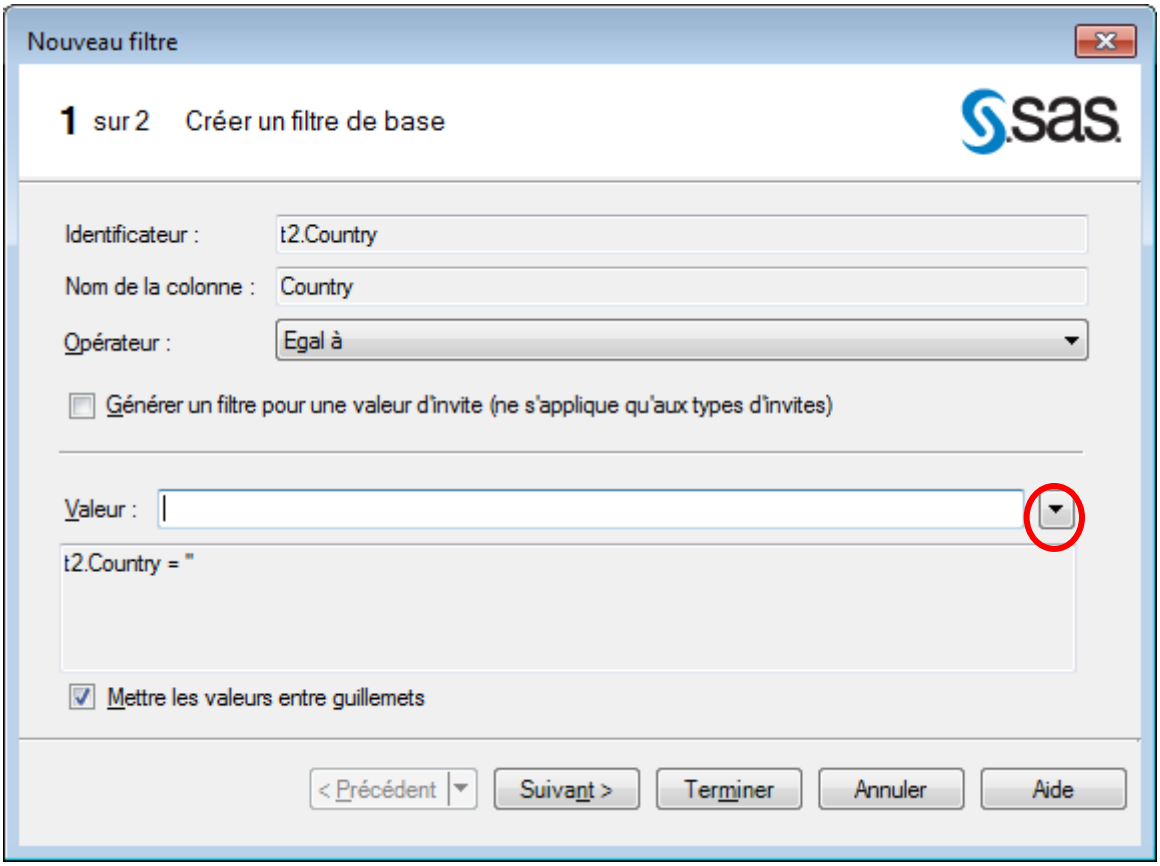


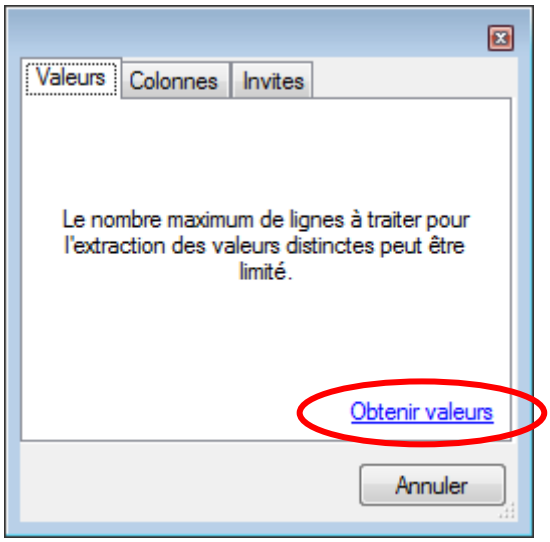


Sélectionner la colonne du chiffre d'affaires (Total_Retail_Price), en prendre la somme
Sélectionner la colonne du pays de l'employé
Sélectionner la colonne de l'année
Sélectionner la colonne du numéro du mois

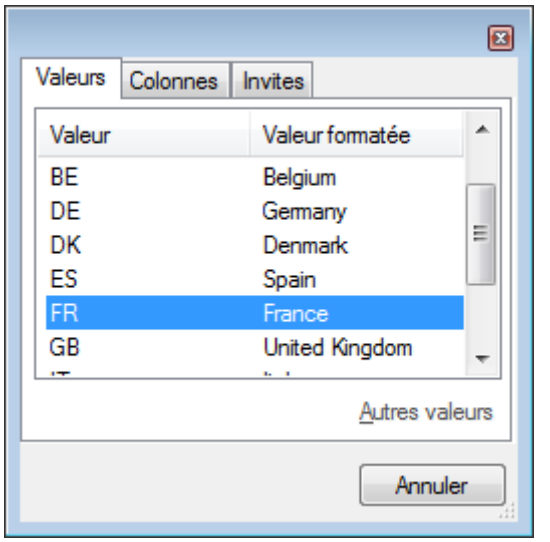


Aller dans l'onglet Filtrer les données
Faire un filtre sur le pays de l'employé

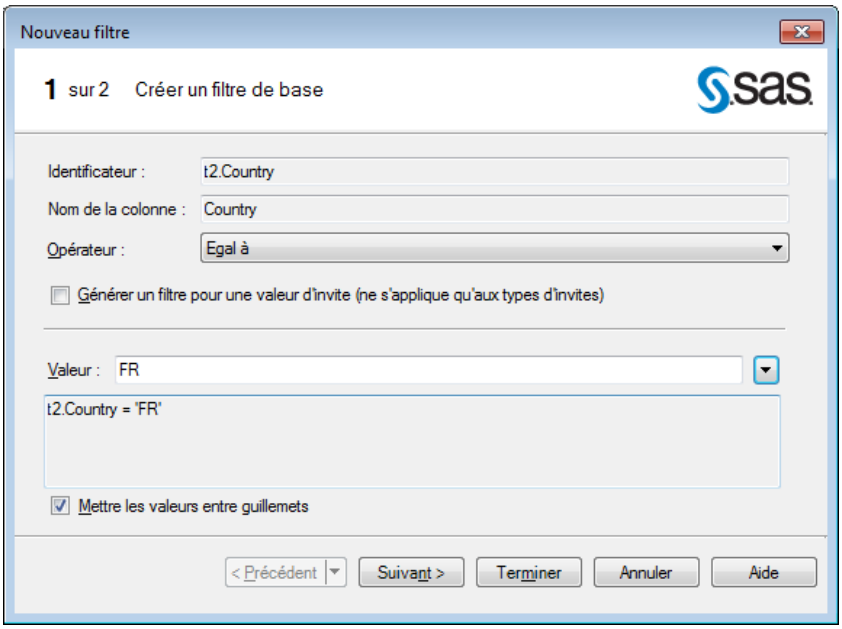




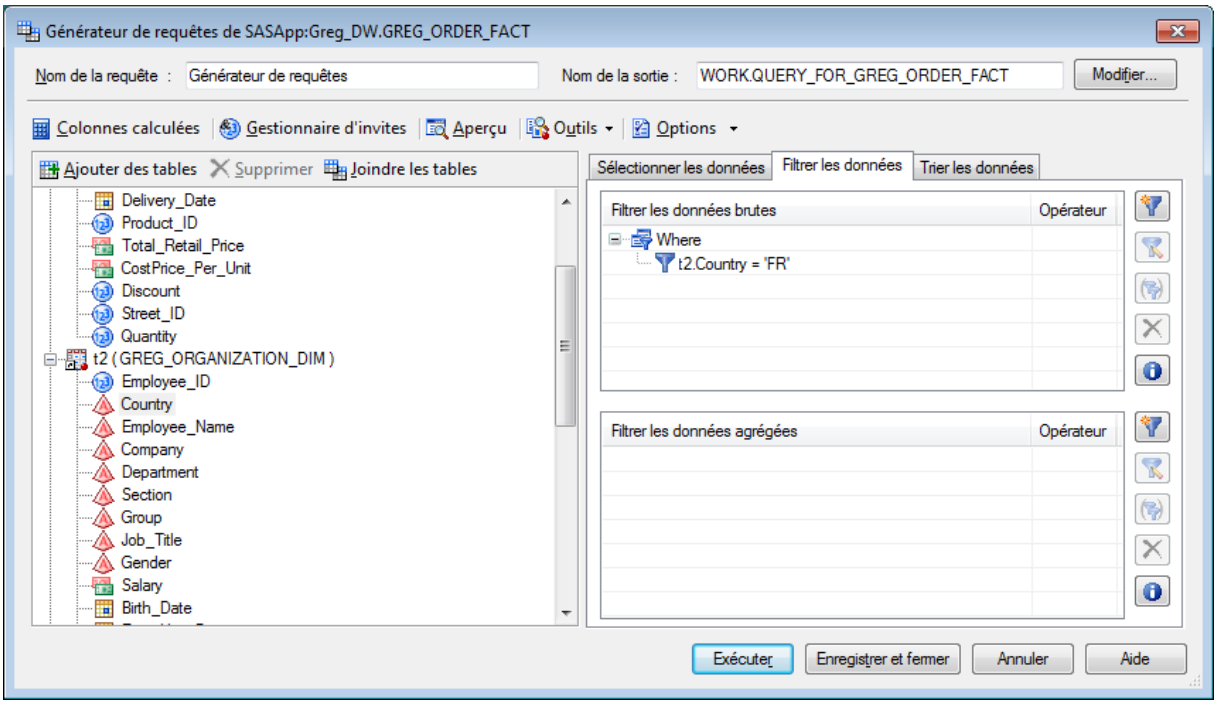
Obtenir les valeurs



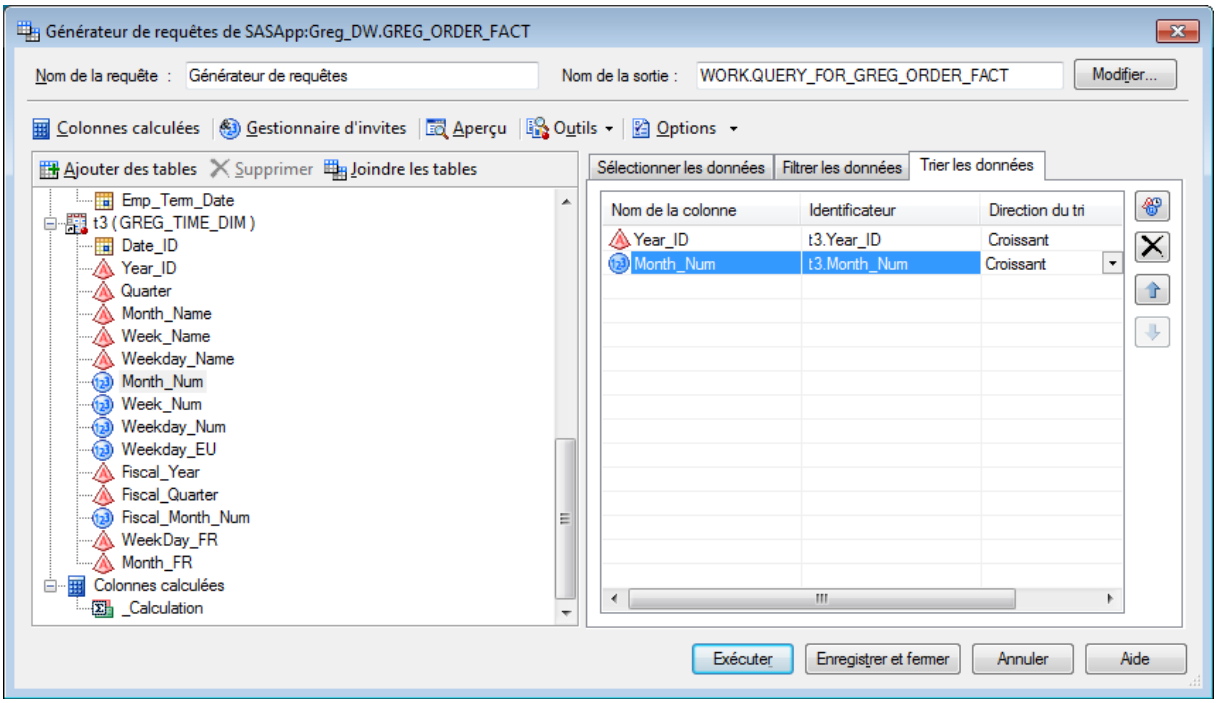
Sélectionner la France (par exemple)



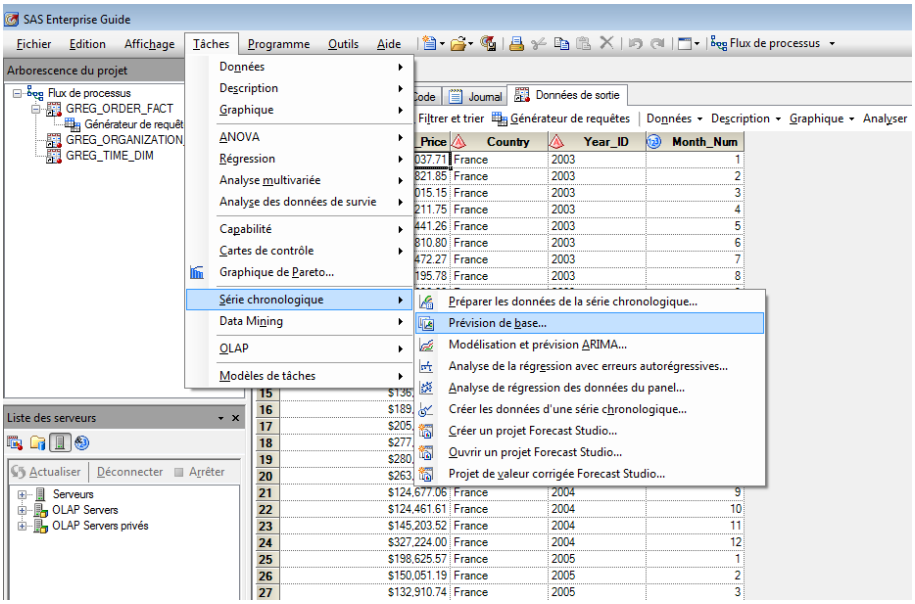
Terminer



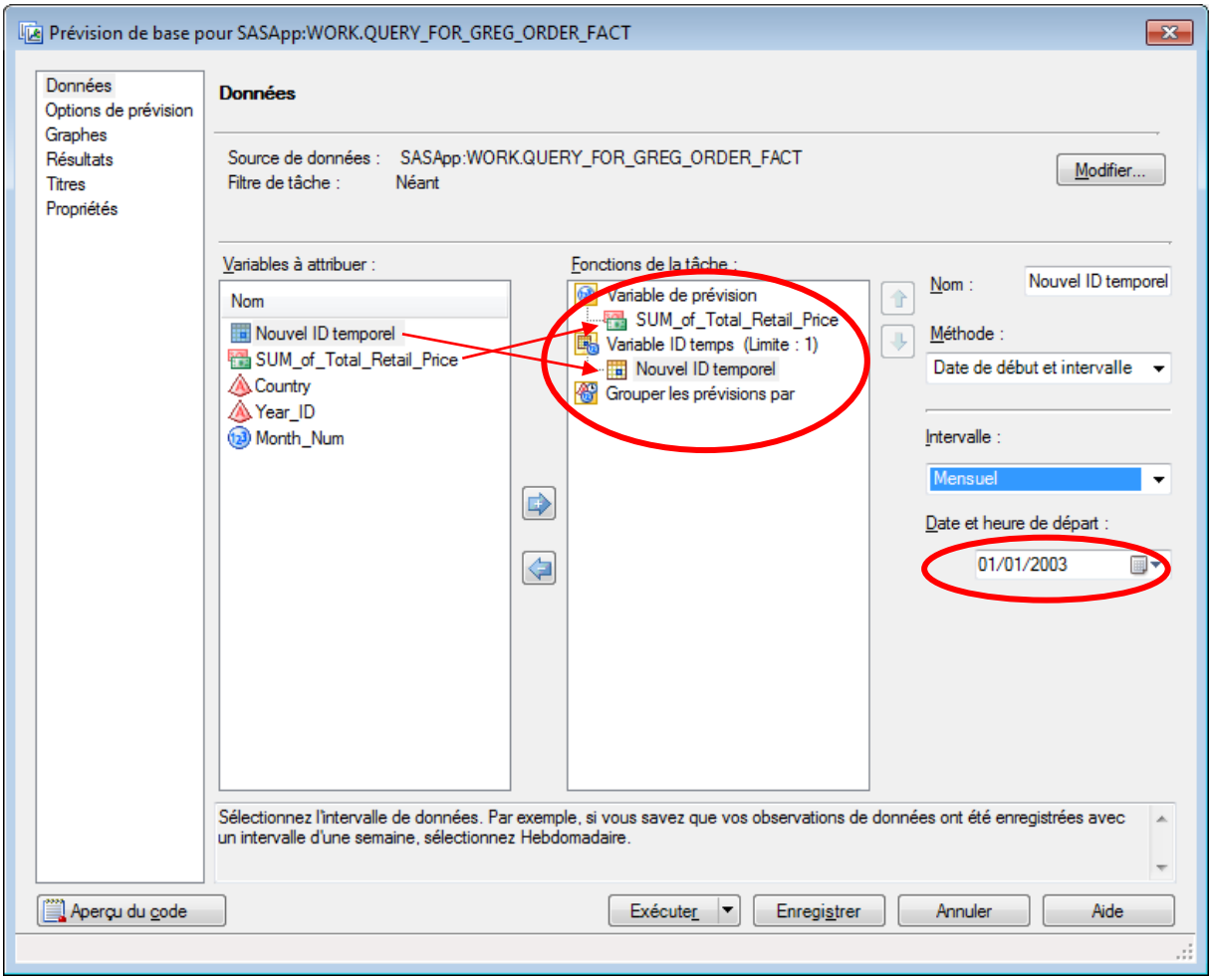
Allez dans l'onglet « trier les données »



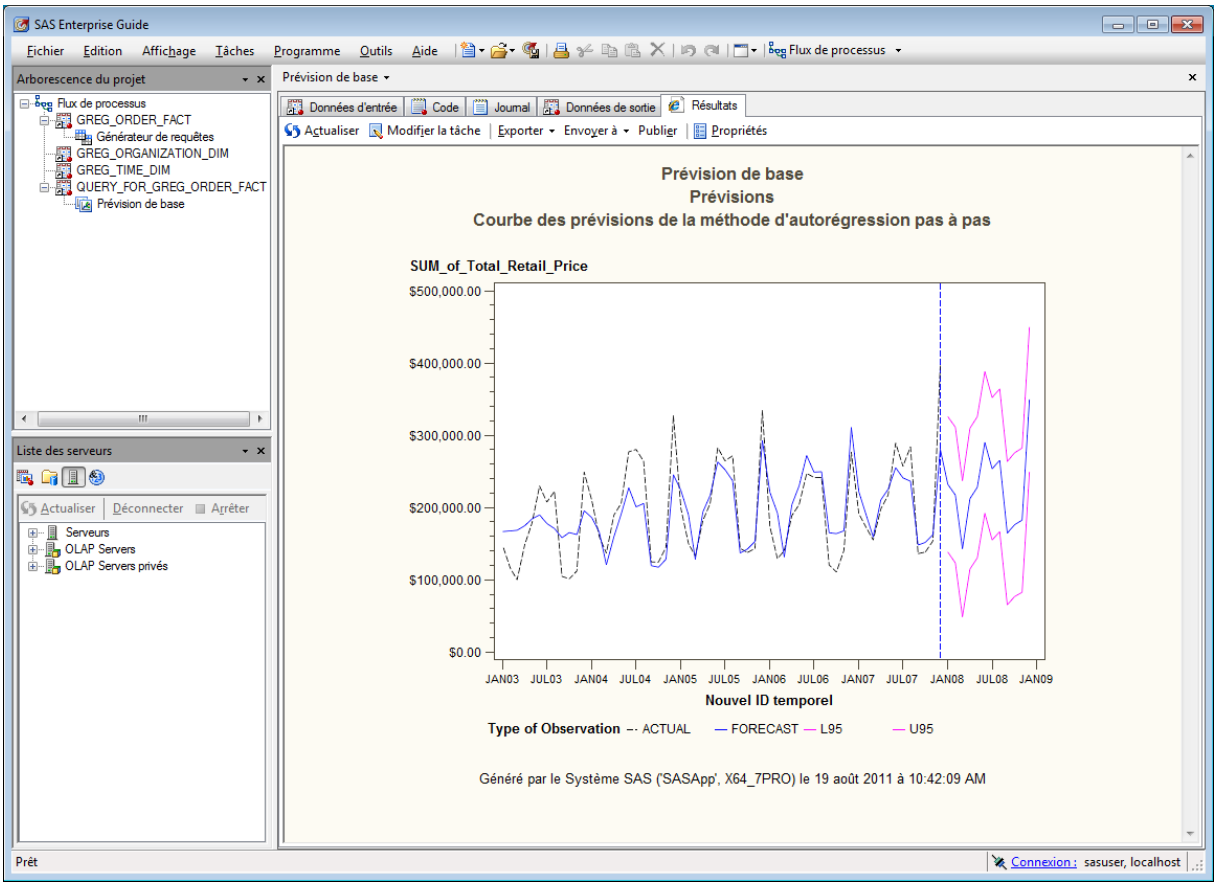
Aller dans l'onglet pour **Trier les données**
Trier par année puis par numéro du mois croissant
Exécuter la requête.



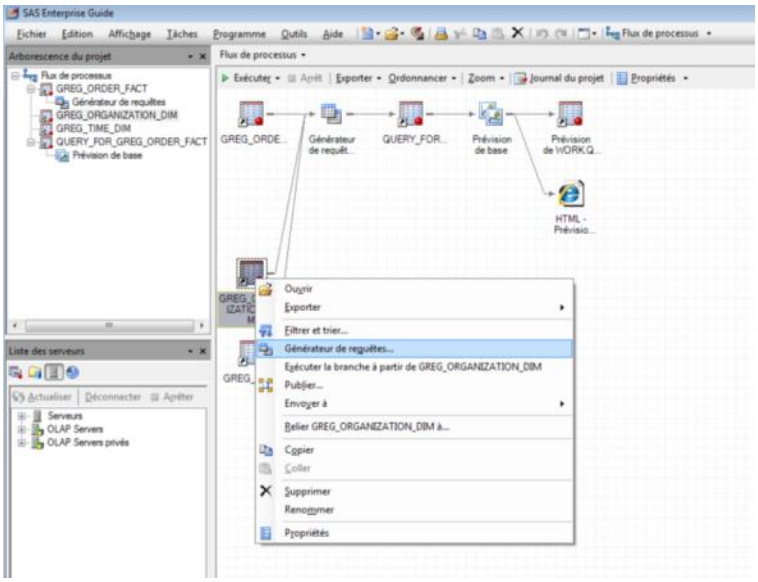
Aller dans le menu des **tâches → Série Chronologique → Prévision de Base**

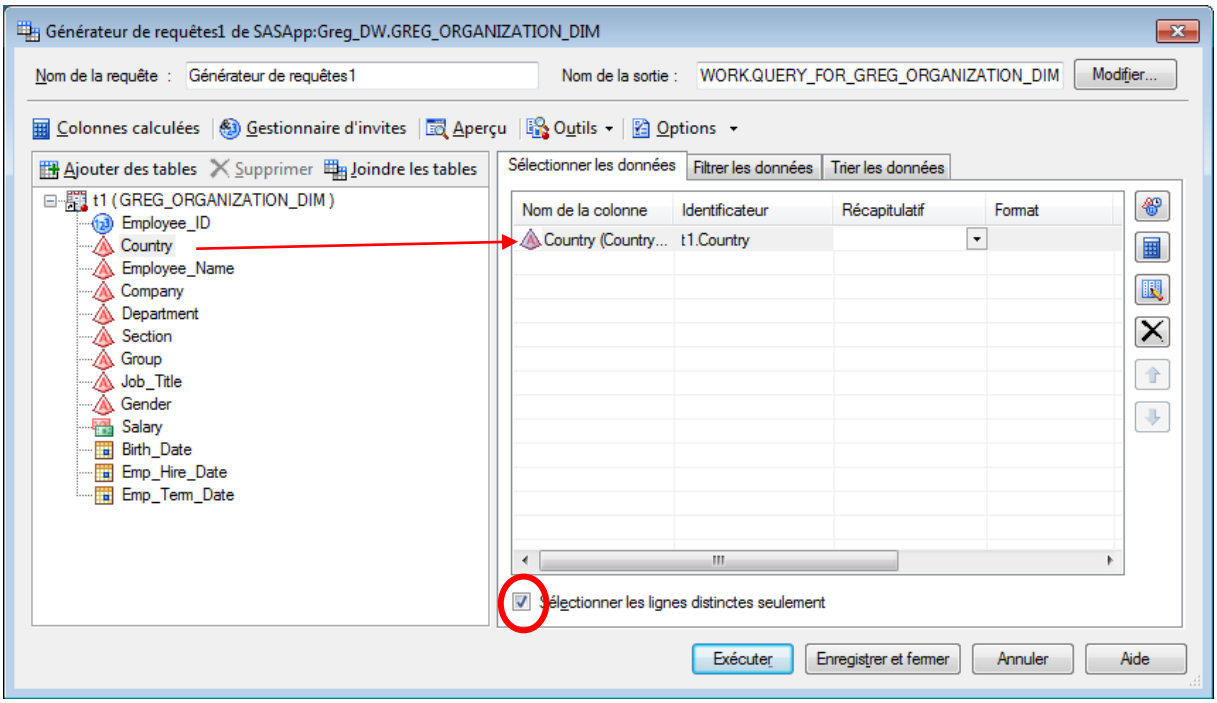


Sélectionner la somme du chiffre d'affaires en variable de prévision
Grouper la prévision par le pays de l'employé
Sélectionner la nouvelle variable du temps en variable d'identifiant du temps.
Entrer la date de départ : **01/01/2003**
Exécuter

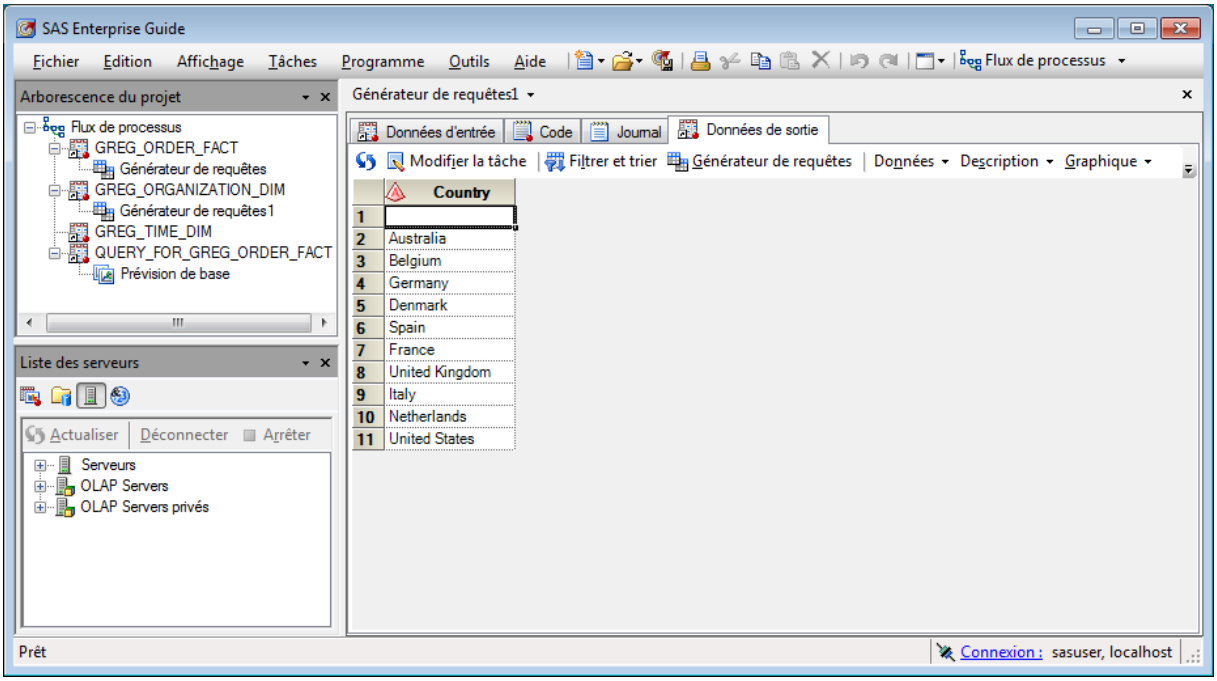


Revenir au flux de processus et faire une requête sur la table de l'organisation afin de sélectionner les différents pays.

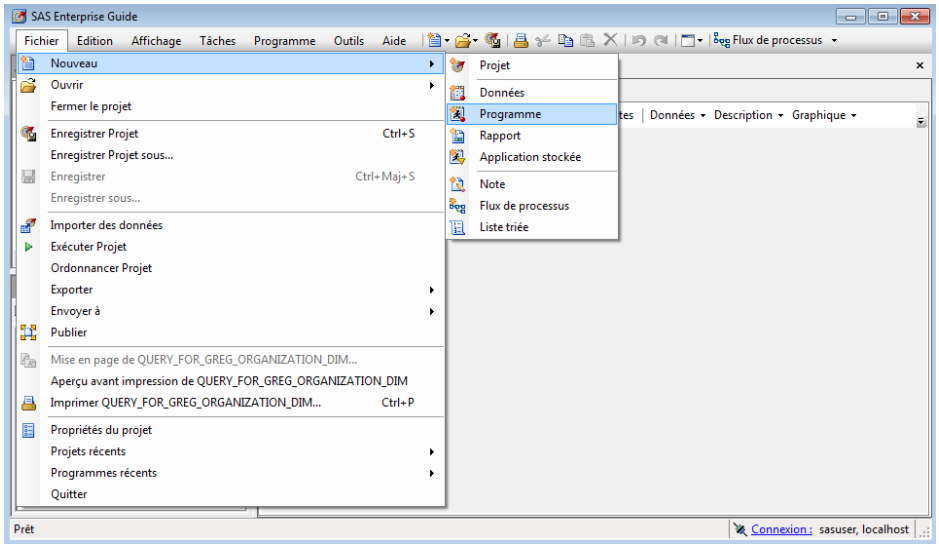


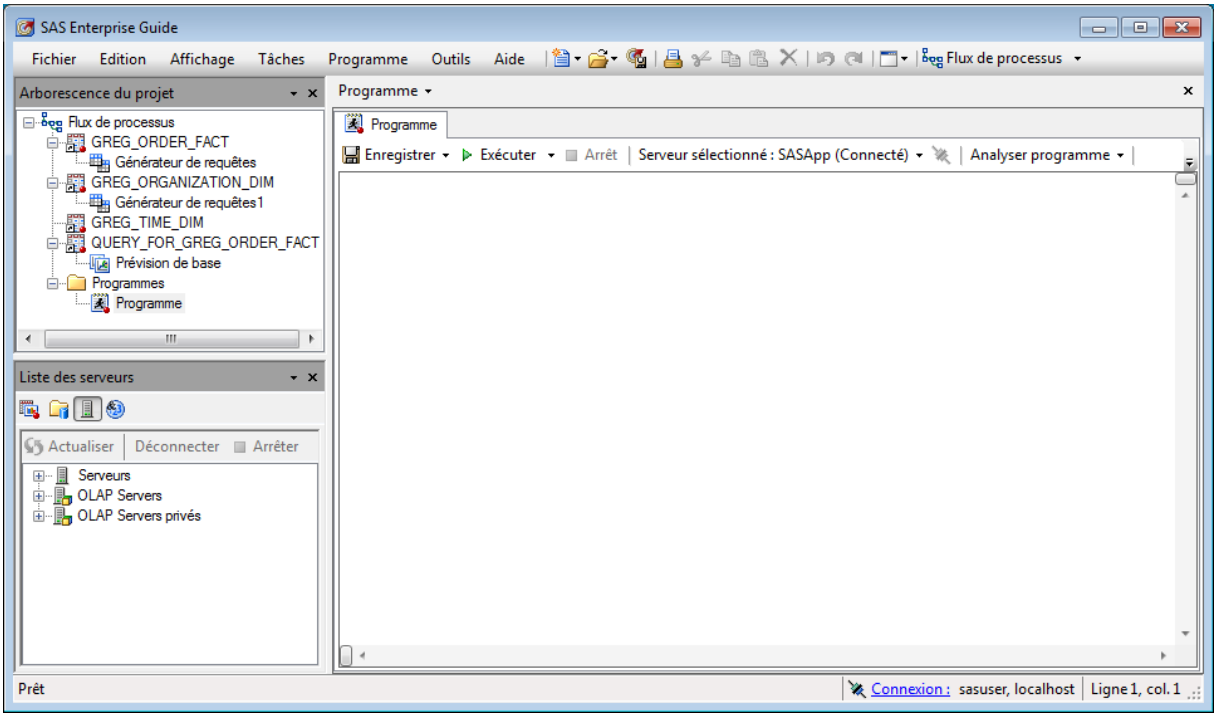


Sélectionner les lignes distinctes du pays de l'employé seulement,
Exécuter.



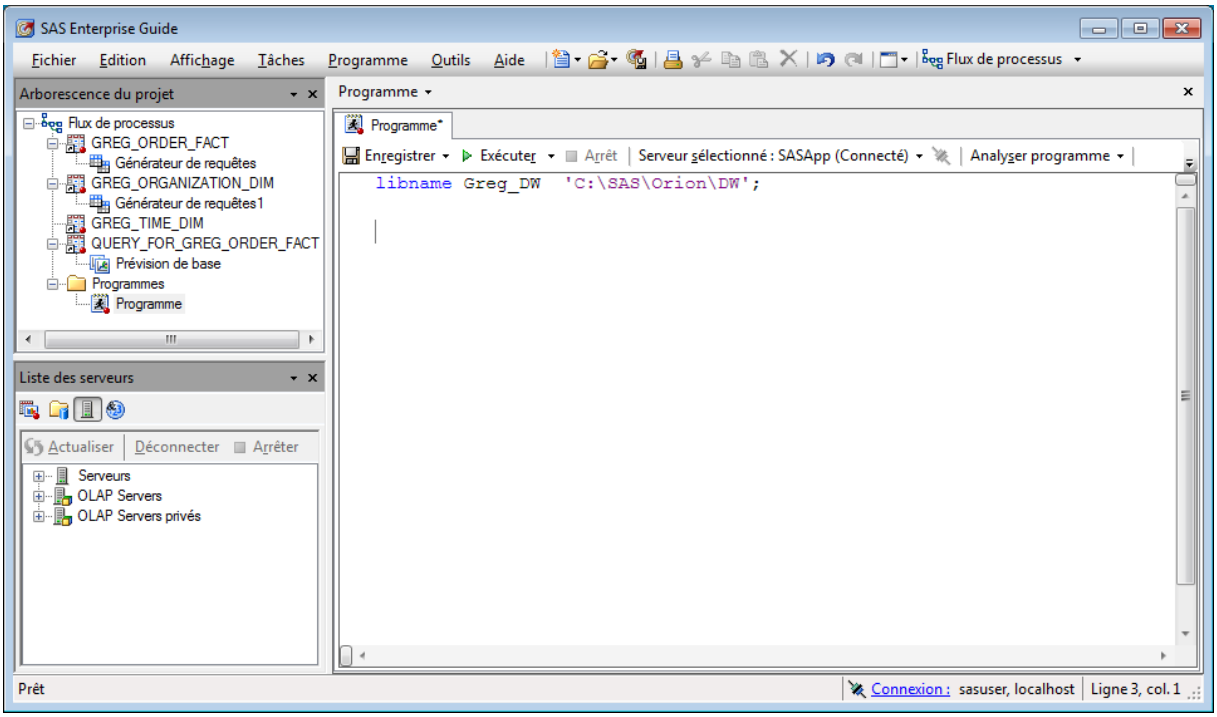
Créer un nouveau code

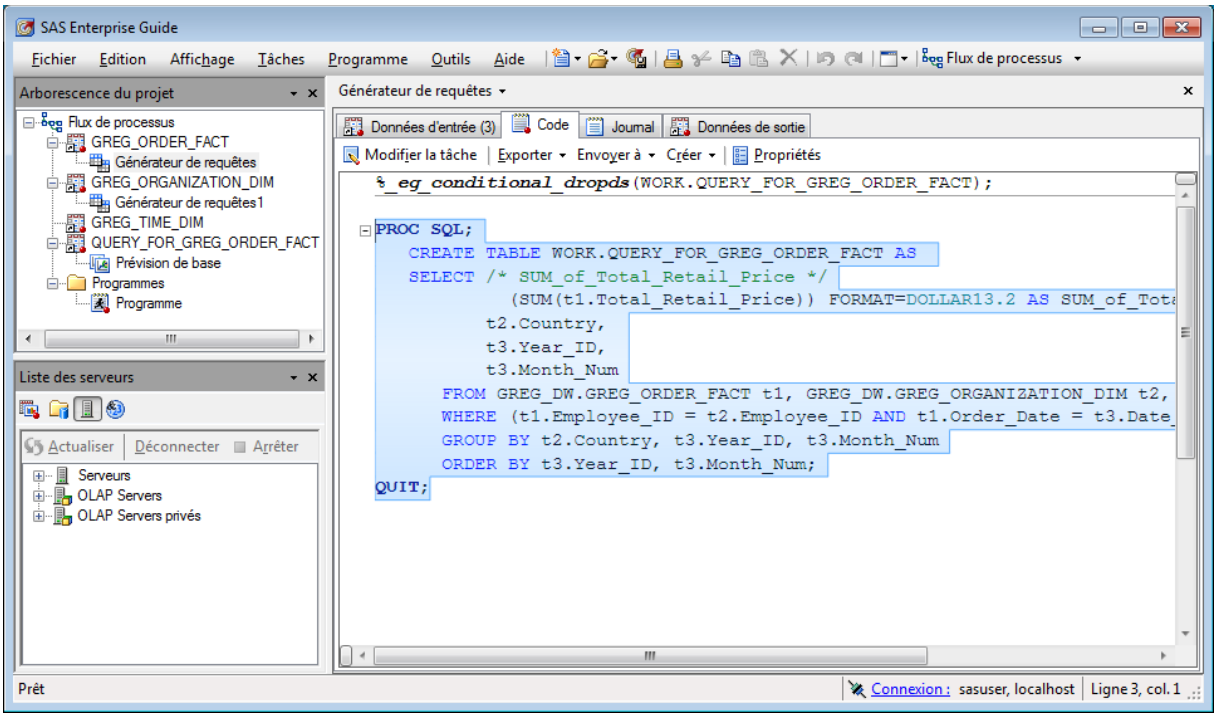




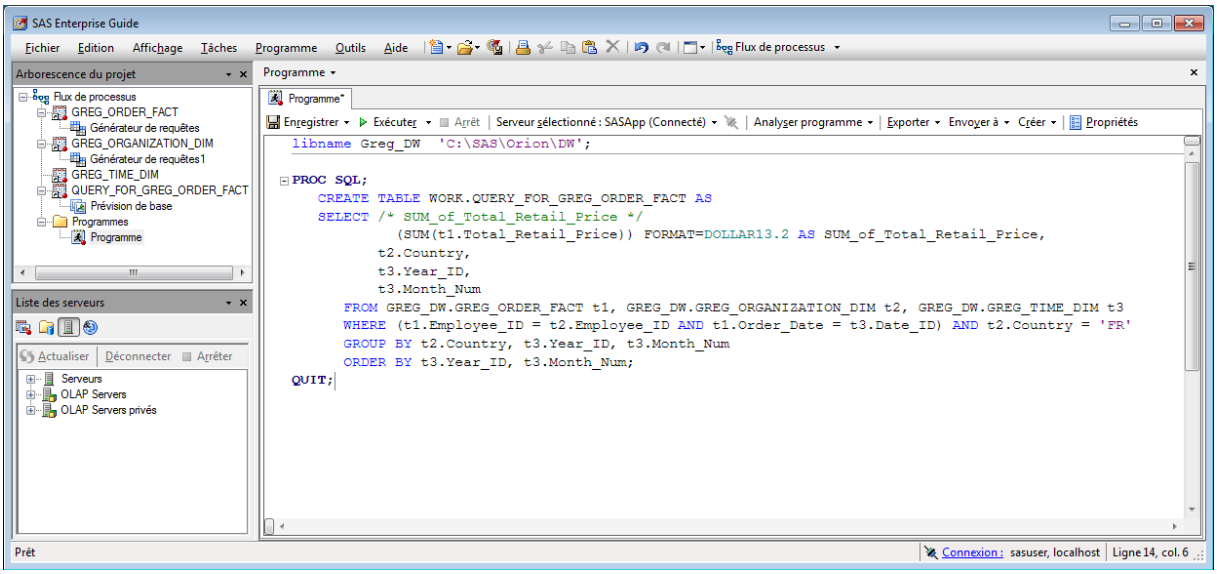
Entrer le libname : libname votrenomdw 'chemin vers là où se trouve les données de la bibliothèque Orion Gold';

Par exemple : libname Greg_DW 'C:\SAS\Orion\DW';

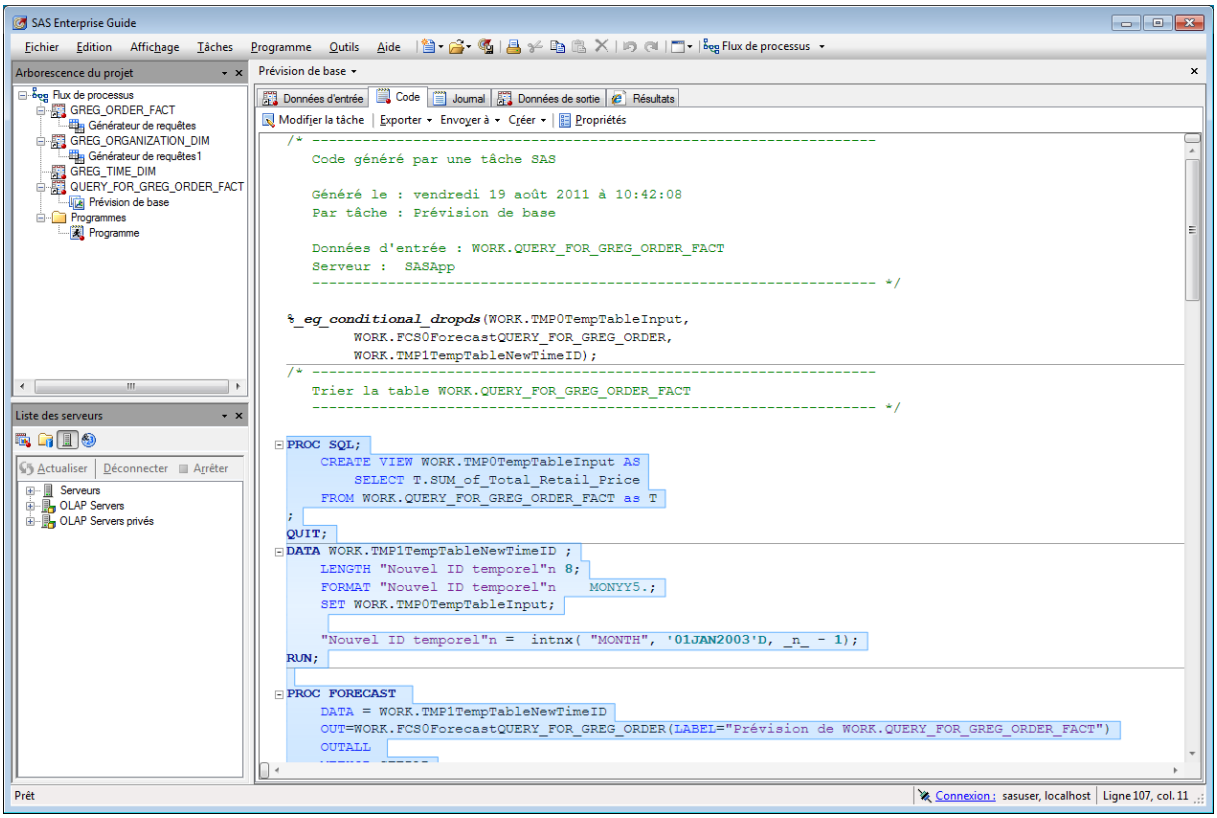


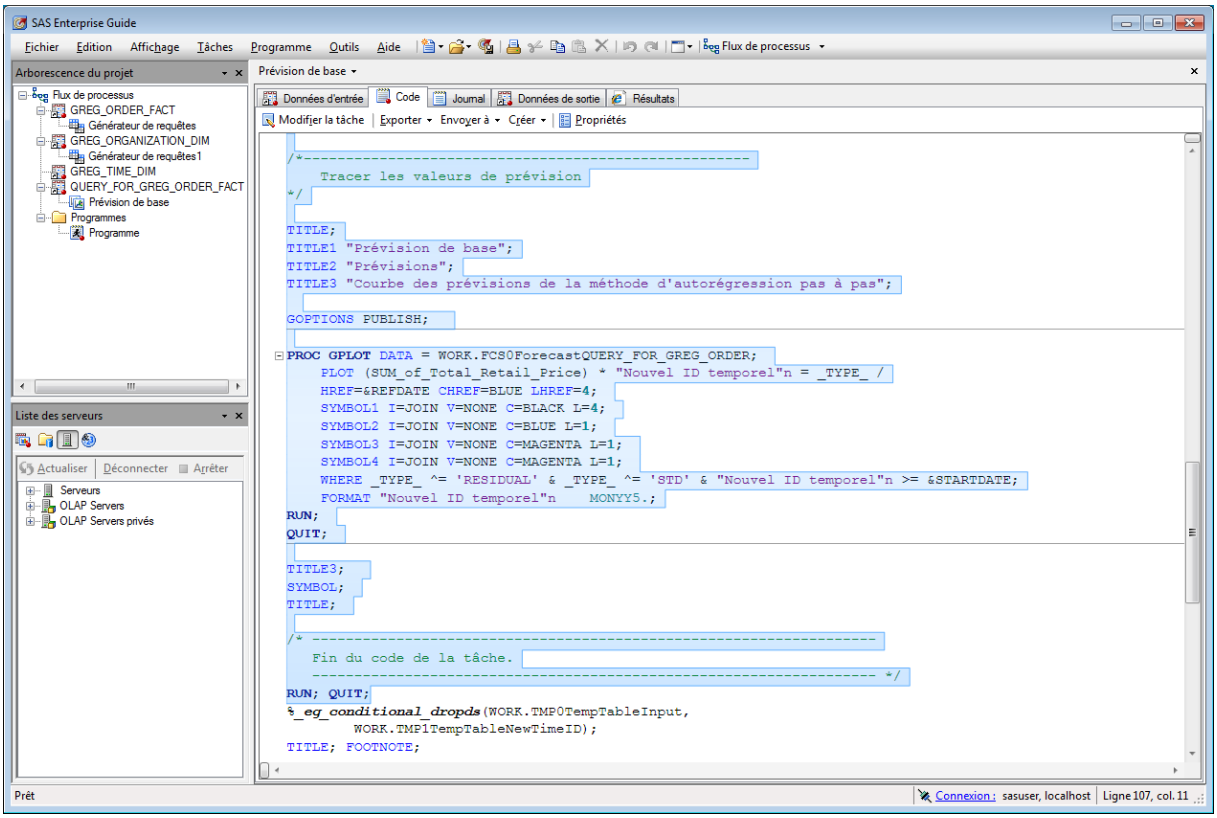


Copier le code de la requête SQL comme ci-dessus et coller le dans votre code.



Copier le code de la série chronologique compris entre les commentaires « Trier la table WORK.QUERY_FOR_ORDER_FACT » et « Fin du code de la tâche ».
Coller le dans votre programme.





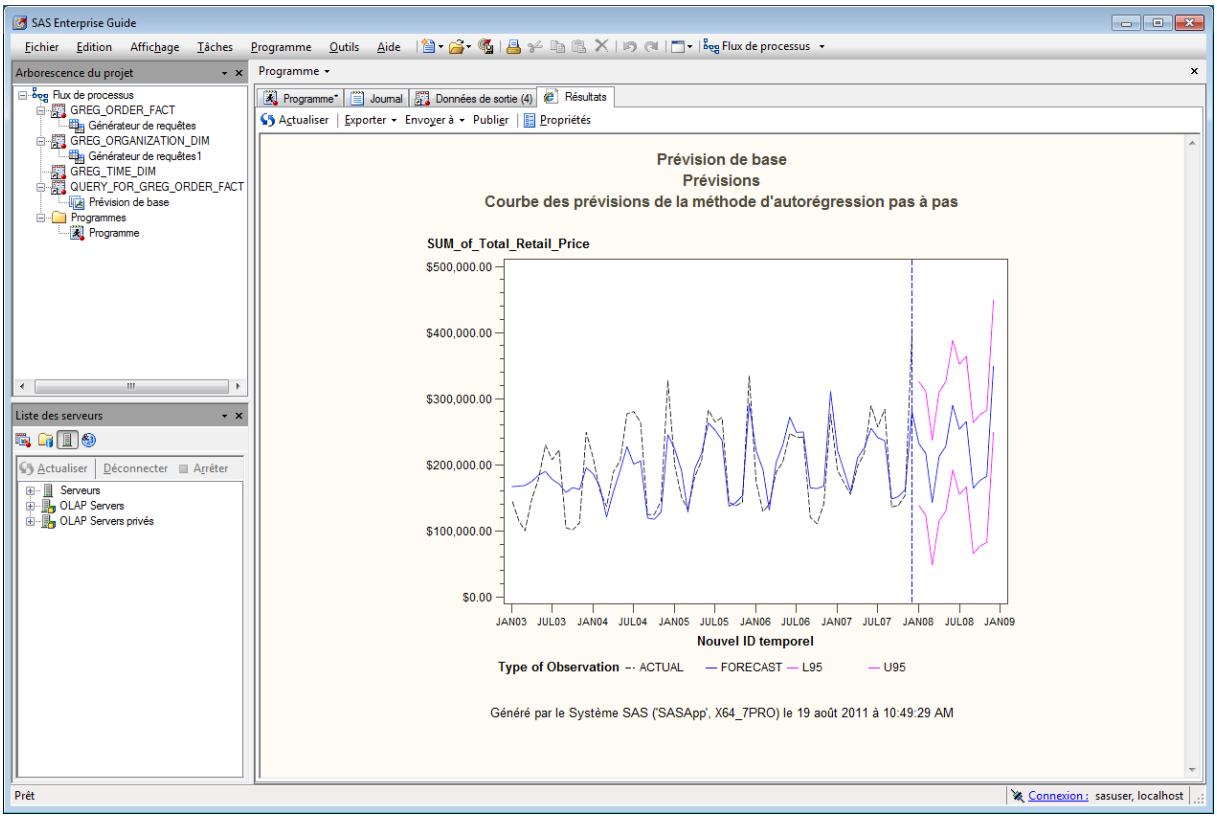
Exemple de code obtenu :

```
/*************************************************************************************/
libname Greg_DW  'C:\SAS\Orion\DW';

PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_GREG_ORDER_FACT AS
   SELECT /* SUM_of_Total_Retail_Price */
            (SUM(t1.Total_Retail_Price)) FORMAT=DOLLAR13.2 AS
SUM_of_Total_Retail_Price,
          t2.Country,
          t3.Year_ID,
          t3.Month_Num
      FROM GREG_DW.GREG_ORDER_FACT t1, GREG_DW.GREG_ORGANIZATION_DIM
t2, GREG_DW.GREG_TIME_DIM t3
      WHERE (t1.Employee_ID = t2.Employee_ID AND t1.Order_Date =
t3.Date_ID) AND t2.Country = 'FR'
      GROUP BY t2.Country, t3.Year_ID, t3.Month_Num
      ORDER BY t3.Year_ID, t3.Month_Num;
QUIT;

PROC SQL;
     CREATE VIEW WORK.TMP0TempTableInput AS
          SELECT T.SUM_of_Total_Retail_Price
     FROM WORK.QUERY_FOR_GREG_ORDER_FACT as T
;
QUIT;
DATA WORK.TMP1TempTableNewTimeID ;
     LENGTH "Nouvel ID temporel"n 8;
     FORMAT "Nouvel ID temporel"n     MONYY5.;
     SET WORK.TMP0TempTableInput;
```

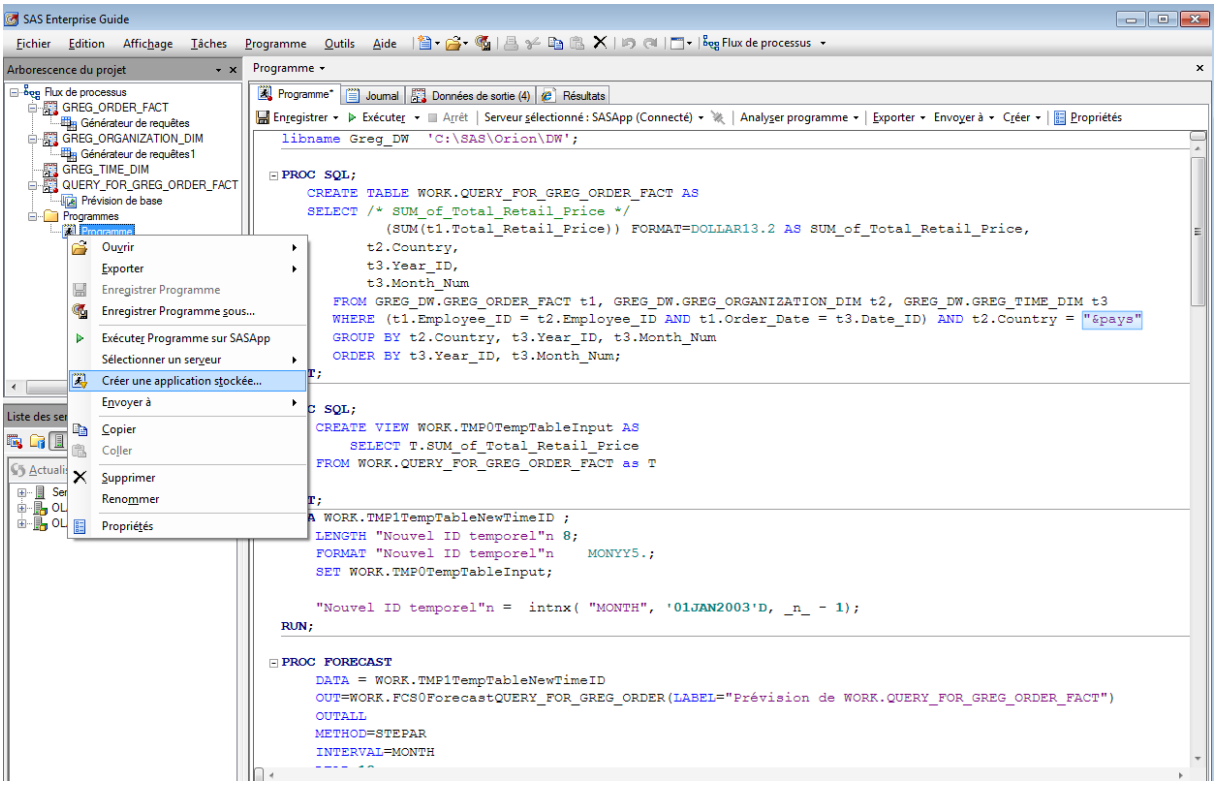
```
    "Nouvel ID temporel"n =  intnx( "MONTH", '01JAN2003'D, _n_ - 1);
RUN;

PROC FORECAST
      DATA = WORK.TMP1TempTableNewTimeID
      OUT=WORK.FCS0ForecastQUERY_FOR_GREG_ORDER(LABEL="Prévision de
WORK.QUERY_FOR_GREG_ORDER_FACT")
      OUTALL
      METHOD=STEPAR
      INTERVAL=MONTH
      LEAD=12
      TREND=2
      ALPHA=0.05
;
      ID "Nouvel ID temporel"n
;
      VAR SUM_of_Total_Retail_Price
;
      FORMAT
            'Nouvel ID temporel'n      MONYY5.;

RUN;TITLE;

TITLE; FOOTNOTE; RUN;
DATA _NULL_;
      DSID = OPEN("WORK.TMP1TempTableNewTimeID");

      NUMOBS = ATTRN(DSID, 'NOBS');
      IF NUMOBS = -1 THEN
            NUMOBS = ATTRN(DSID, 'NLOBSF');
      PLOTSTART = 1;

      DATEVAR = VARNUM(DSID, "Nouvel ID temporel");

      SYSRC = FETCHOBS(DSID, NUMOBS);
      REFVAL = GETVARN(DSID, DATEVAR);
      SYSRC = FETCHOBS(DSID, PLOTSTART);
      STARTVAL = GETVARN(DSID, DATEVAR);

      RC = CLOSE(DSID);

      CALL SYMPUT('REFDATE', REFVAL);
      CALL SYMPUT('STARTDATE', STARTVAL);
RUN;
FOOTNOTE;
FOOTNOTE1 "Généré par le Système SAS (&_SASSERVERNAME, &SYSSCPL)
le %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) à %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";



/*---------------------------------------------------
    Tracer les valeurs de prévision
*/
```

```
TITLE;
TITLE1 "Prévision de base";
TITLE2 "Prévisions";
TITLE3 "Courbe des prévisions de la méthode d'autorégression pas à
pas";

GOPTIONS PUBLISH;

PROC GPLOT DATA = WORK.FCS0ForecastQUERY_FOR_GREG_ORDER;
      PLOT (SUM_of_Total_Retail_Price) * "Nouvel ID temporel"n =
_TYPE_  /
      HREF=&REFDATE CHREF=BLUE LHREF=4;
      SYMBOL1 I=JOIN V=NONE C=BLACK L=4;
      SYMBOL2 I=JOIN V=NONE C=BLUE L=1;
      SYMBOL3 I=JOIN V=NONE C=MAGENTA L=1;
      SYMBOL4 I=JOIN V=NONE C=MAGENTA L=1;
      WHERE _TYPE_ ^= 'RESIDUAL' & _TYPE_ ^= 'STD' & "Nouvel ID
temporel"n >= &STARTDATE;
      FORMAT "Nouvel ID temporel"n    MONYY5.;
RUN;
QUIT;

TITLE3;
SYMBOL;
TITLE;

/* ------------------------------------------------------------------
--
   Fin du code de la tâche.
   ------------------------------------------------------------------
-- */
RUN; QUIT;

/*****************************************************************************/
```

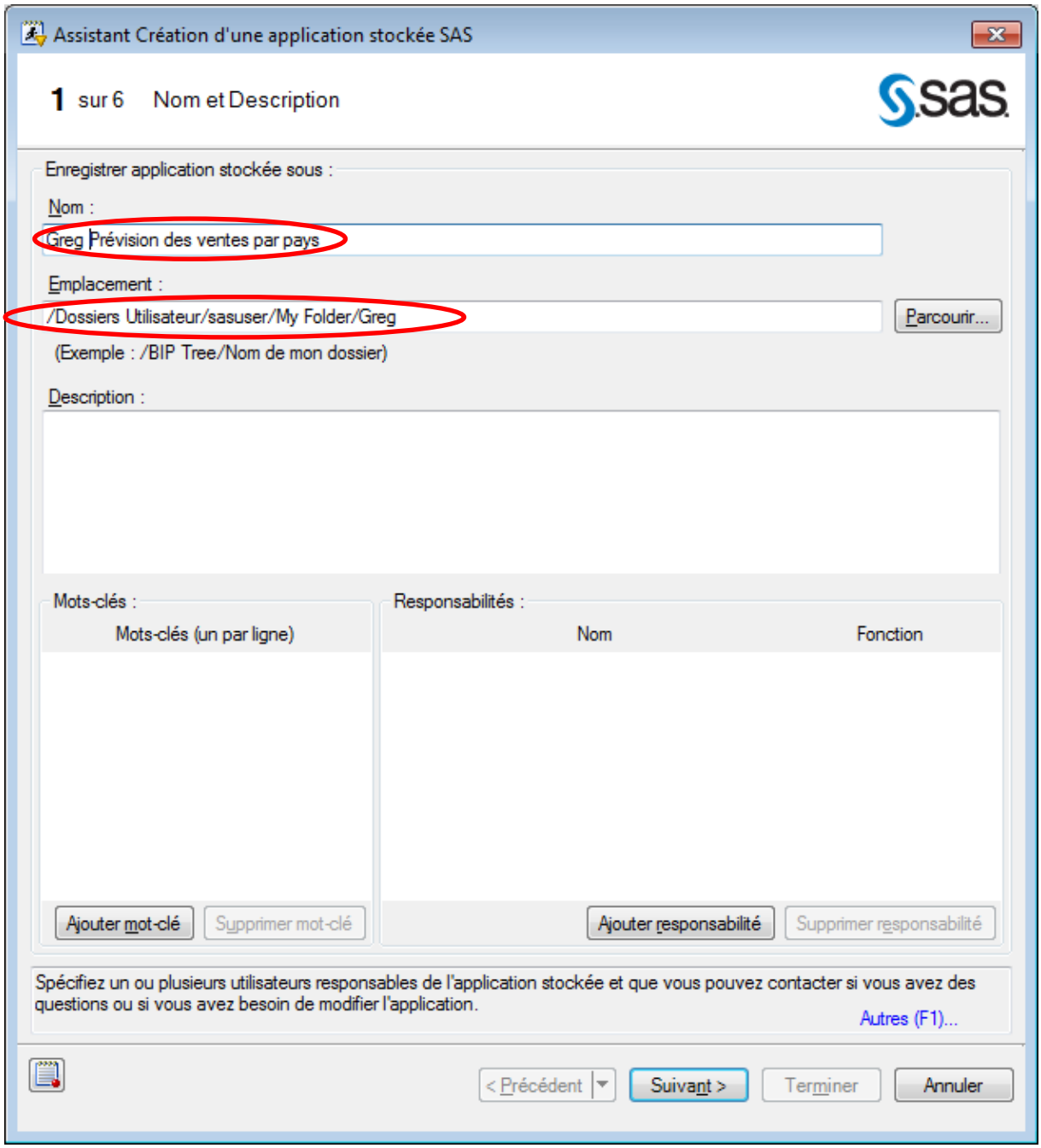
Exécuter le code pour voir s'il fonctionne.



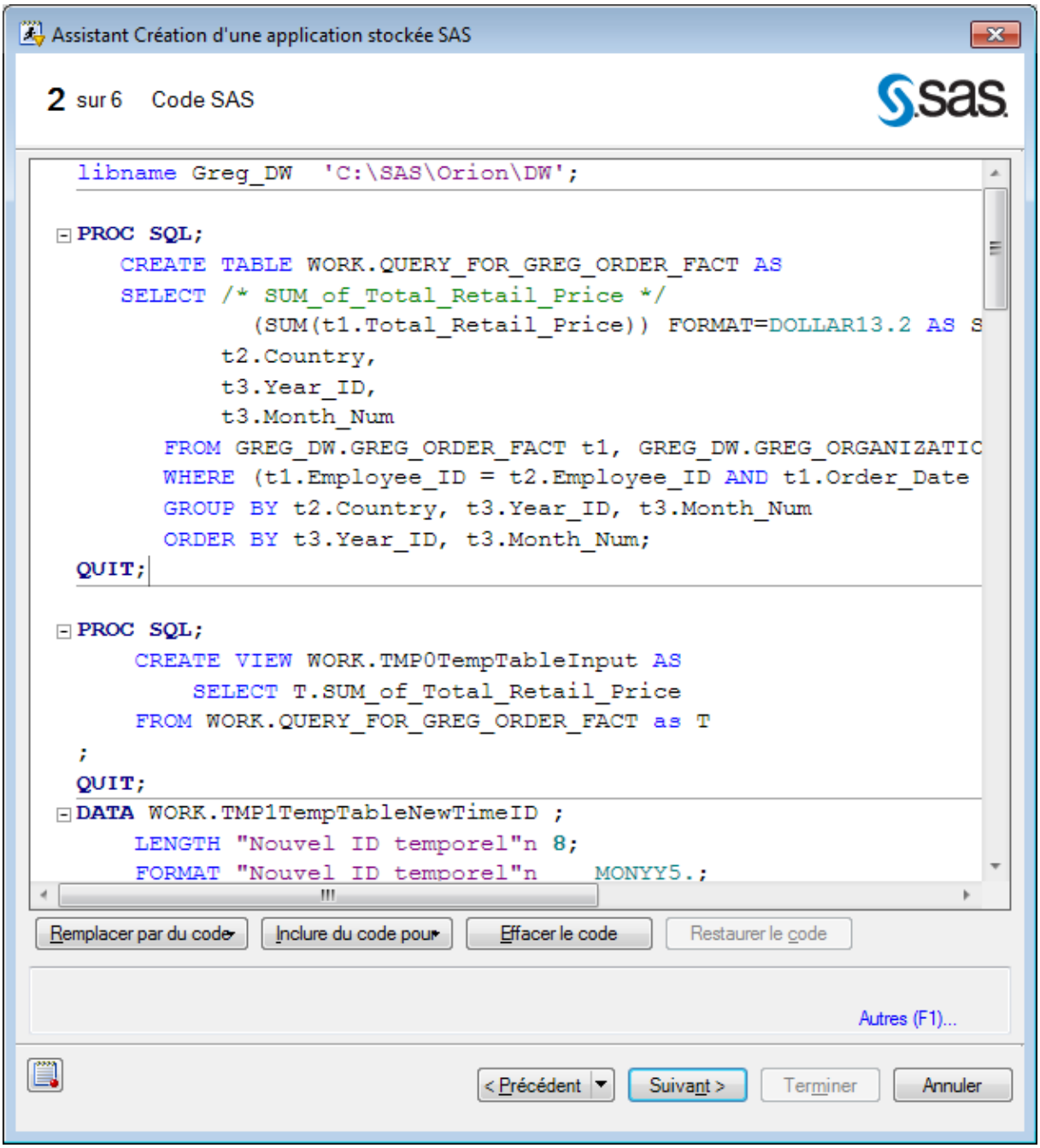
Remplacer dans le code 'France' par "&pays".
Attention : double quotte (du ‘3’ d’un clavier azerty).



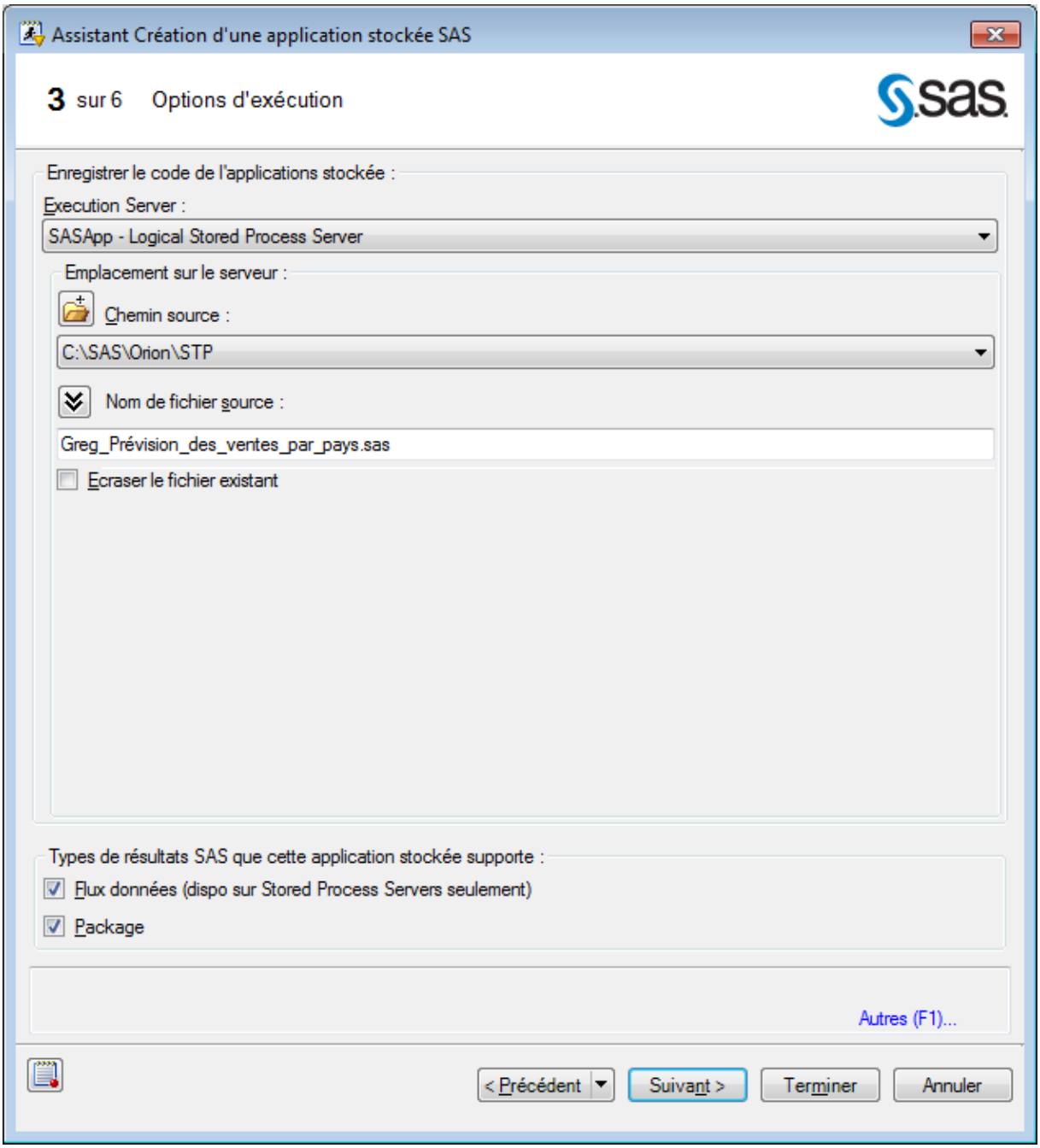
Clic-droit sur le programme → créer une application stockée.



Lui donner un nom
Si c'est la première procédure stockée, il faut entrer l'emplacement, soit un dossier dans votre sous-dossier
Suivant

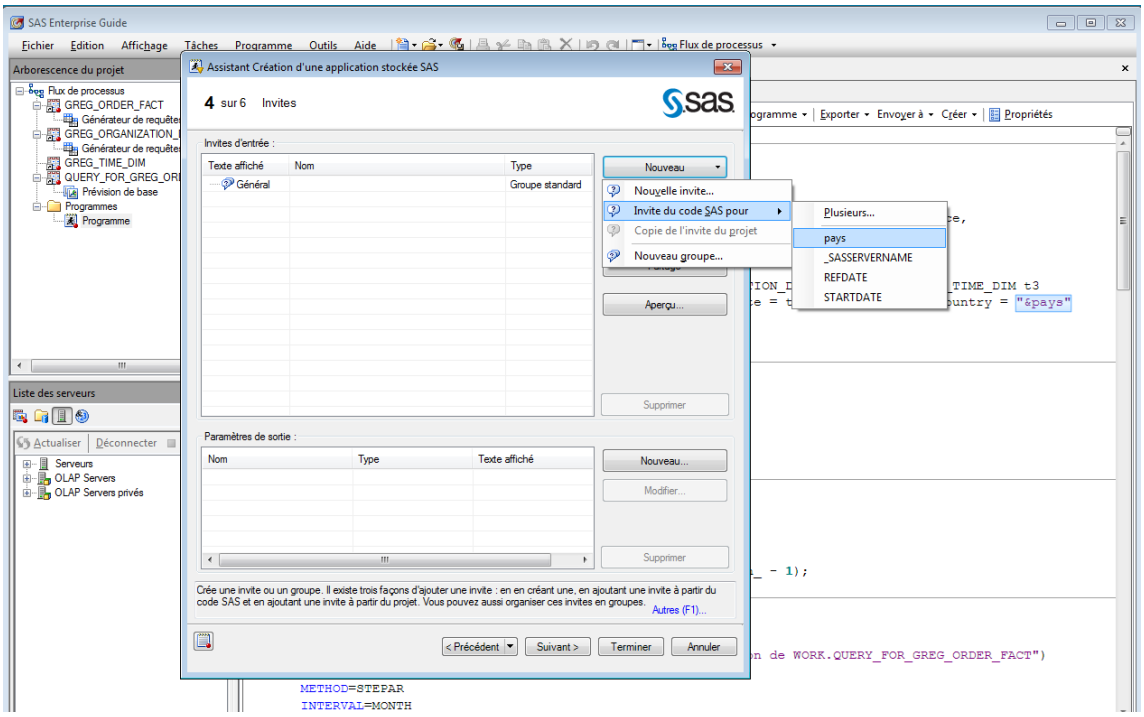

Assistant Création d'une application stockée SAS
2 sur 6 Code SAS
sas
libname Greg_DW 'C:\SAS\Orion\DW';
PROC SQL;
CREATE TABLE WORK.QUERY_FOR_GREG_ORDER_FACT AS
SELECT /* SUM_of_Total_Retail_Price */
(SUM(t1.Total_Retail_Price)) FORMAT=DOLLAR13.2 AS S
t2.Country,
t3.Year_ID,
t3.Month_Num
FROM GREG_DW.GREG_ORDER_FACT t1, GREG_DW.GREG_ORGANIZATIC
WHERE (t1.Employee_ID = t2.Employee_ID AND t1.Order_Date
GROUP BY t2.Country, t3.Year_ID, t3.Month_Num
ORDER BY t3.Year_ID, t3.Month_Num;
QUIT;
PROC SQL;
CREATE VIEW WORK.TMP0TempTableInput AS
SELECT T.SUM_of_Total_Retail_Price
FROM WORK.QUERY_FOR_GREG_ORDER_FACT as T
;
QUIT;
DATA WORK.TMP1TempTableNewTimeID ;
LENGTH "Nouvel ID temporel"n 8;
FORMAT "Nouvel ID temporel"n MONYY5.;
Remplacer par du code
Inclure du code pour
Effacer le code
Restaurer le code
Autres (F1)...
< Précédent
Suivant >
Terminer
Annuler

Suivant

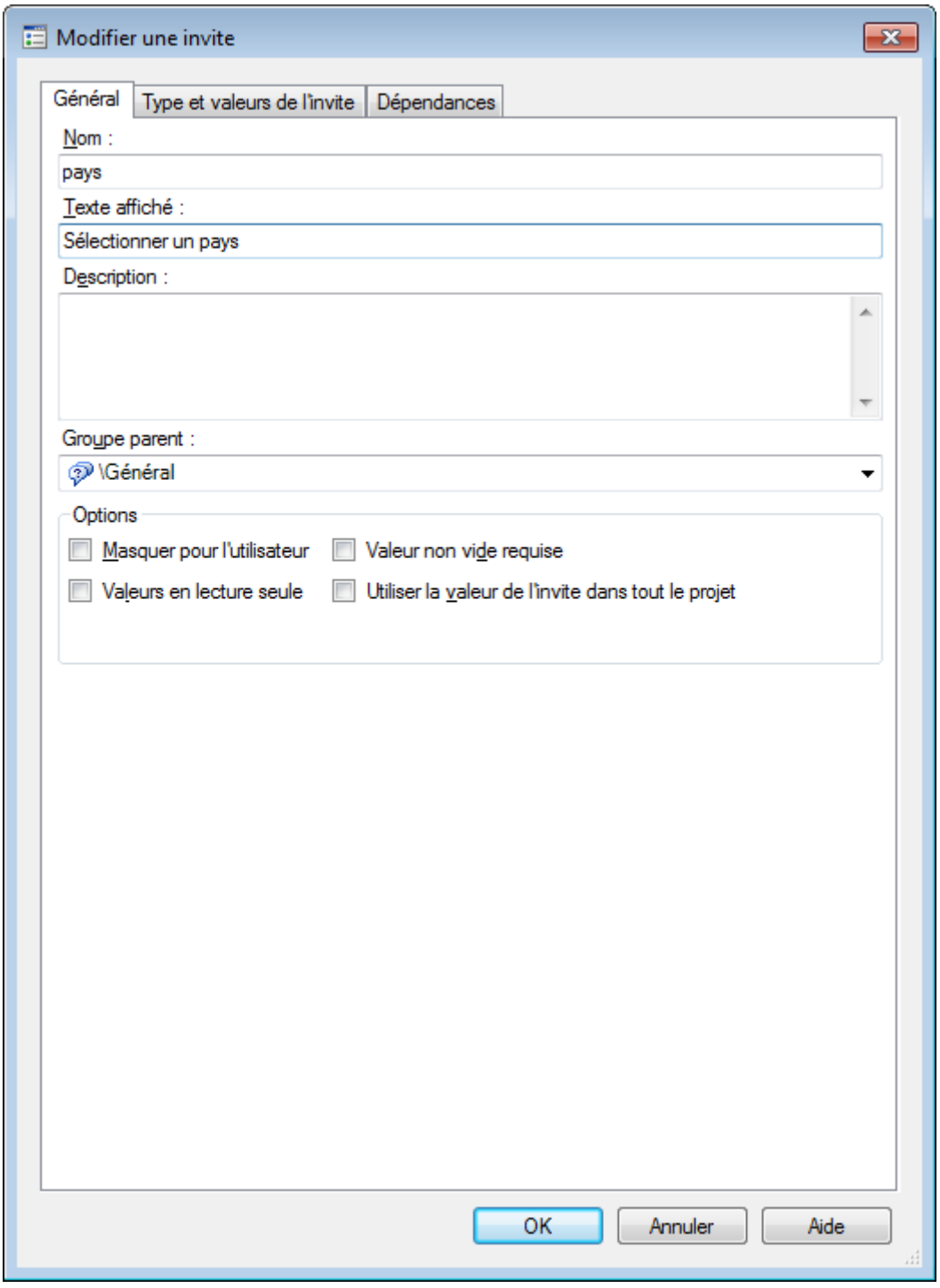


Si c'est la première procédure stockée que vous faites, il faut parfois sélectionner le serveur d'exécution : le SASApp Logical Stored Process Server

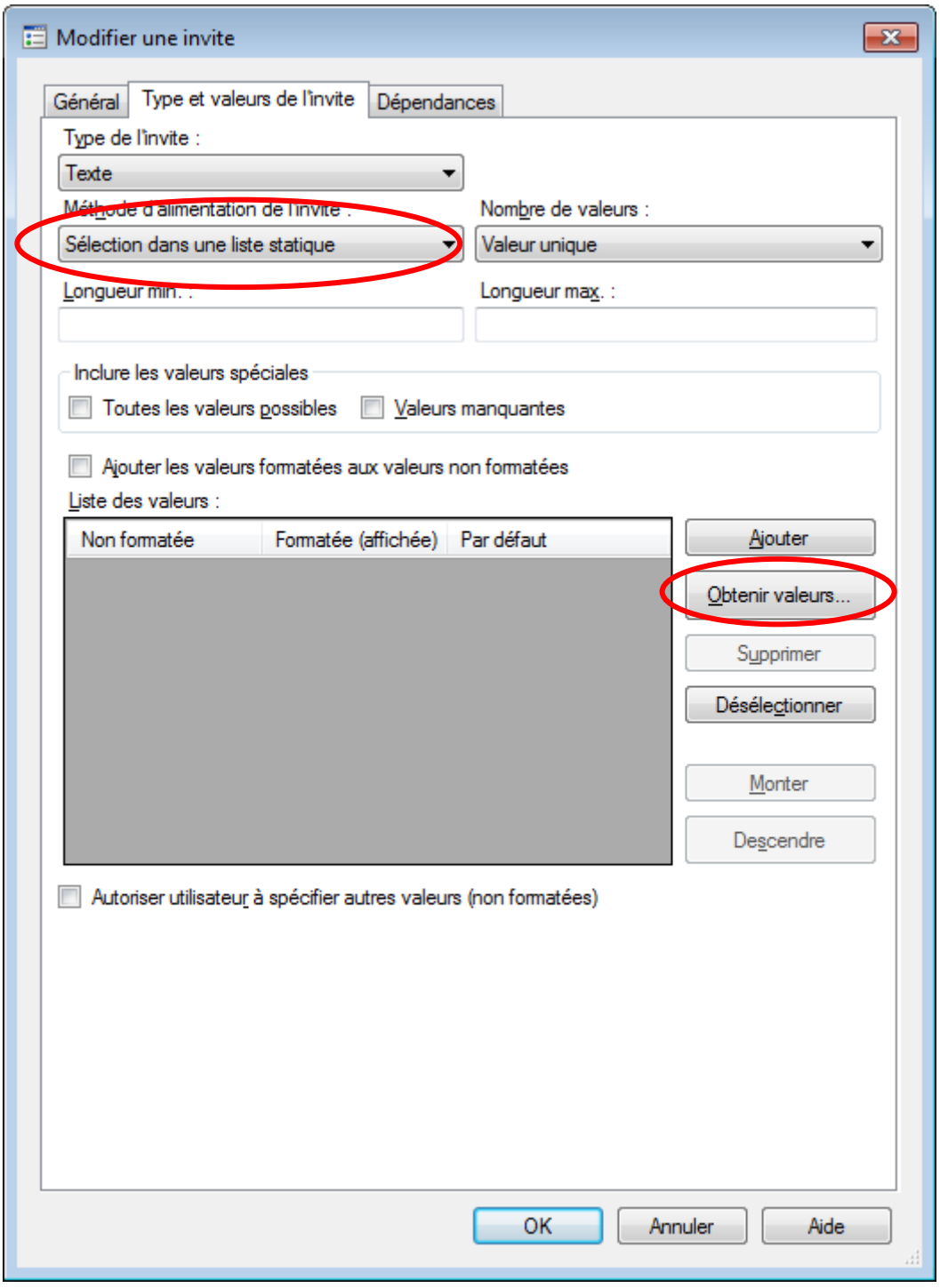
Sélectionner un nom chemin sources sur le serveur où vous avez tous les droits.
Suivant



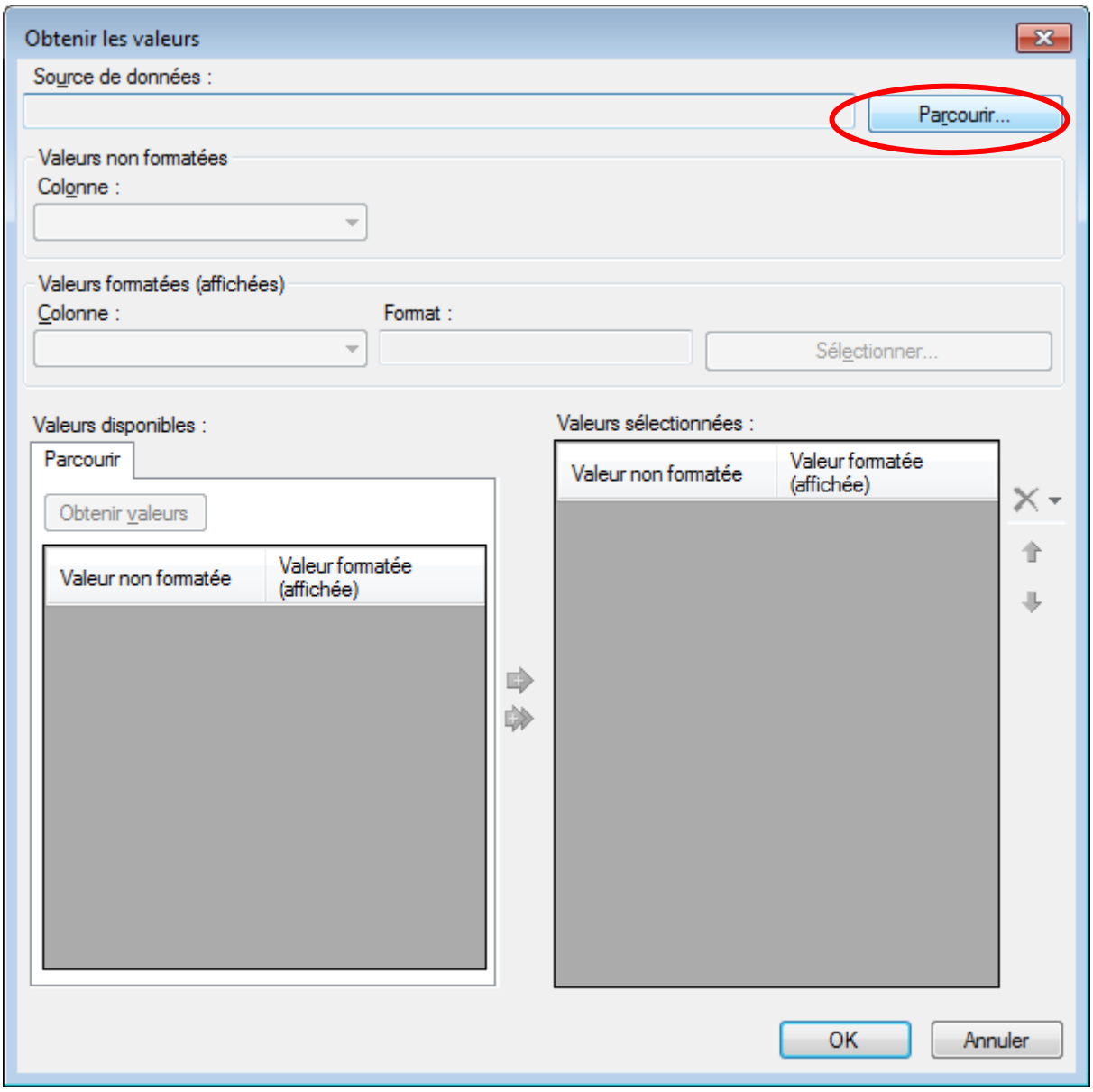
Ajouter une **Invite du code SAS pour** la variable macro **Pays**



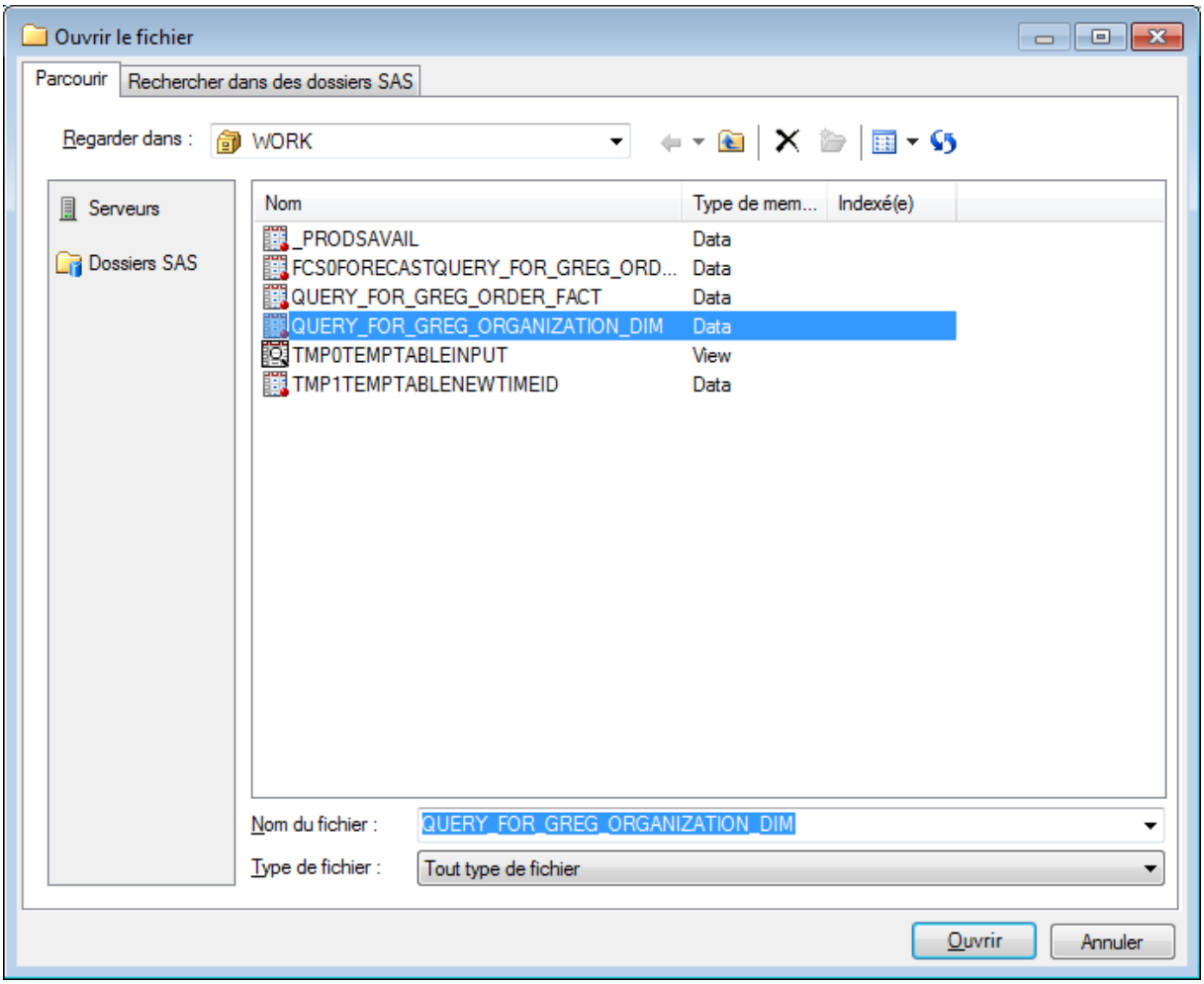
Entrer le texte d'invite à afficher à l'utilisateur



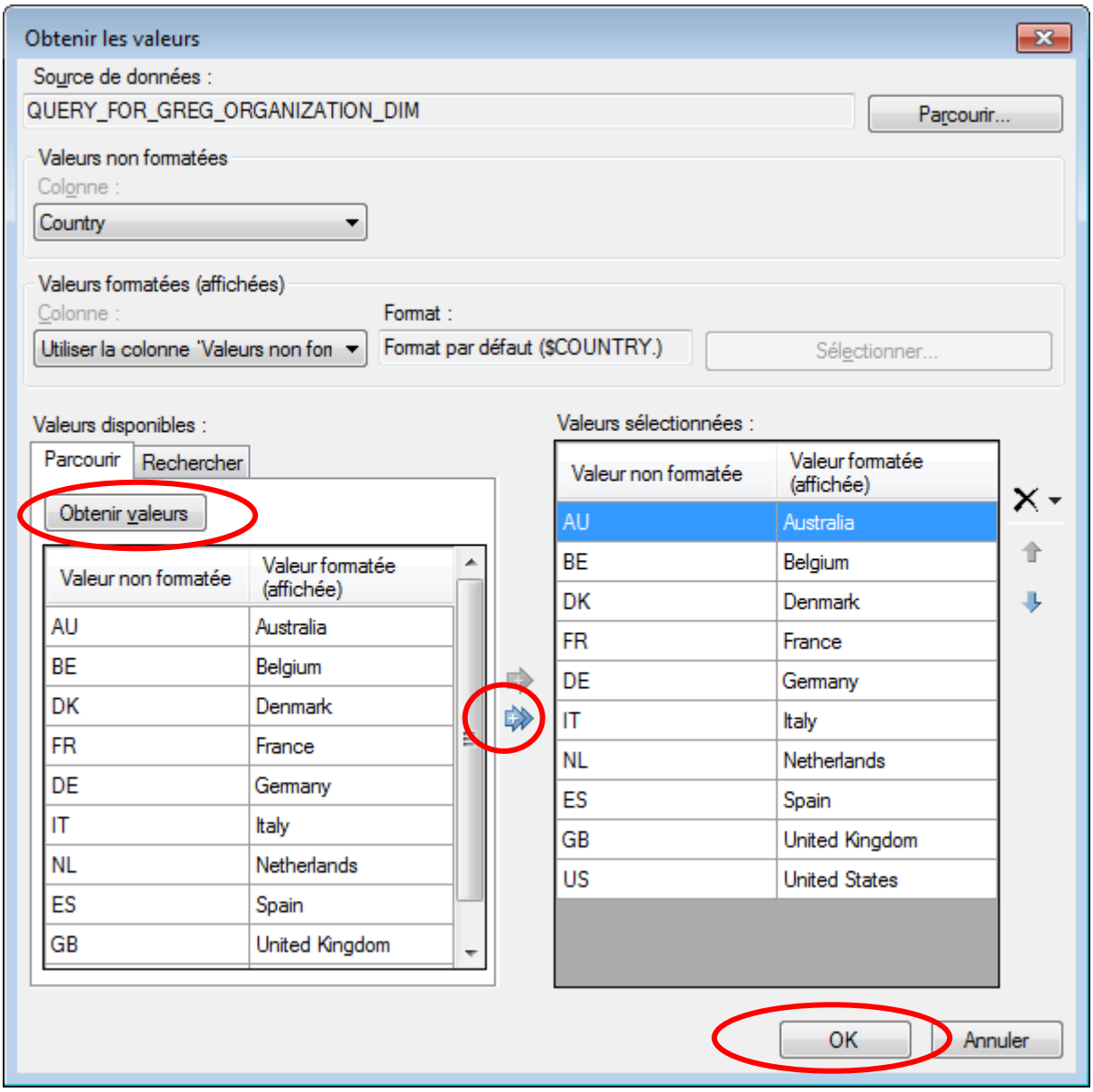
Sélectionner les valeurs dans une liste statique
Obtenir valeurs


Obtenir les valeurs
Source de données :
Parcourir...
Valeurs non formatées
Colonne :
Valeurs formatées (affichées)
Colonne :
Format :
Sélectionner...
Valeurs disponibles :
Parcourir
Obtenir valeurs
Valeur non formatée
Valeur formatée (affichée)
Valeurs sélectionnées :
Valeur non formatée
Valeur formatée (affichée)
OK
Annuler

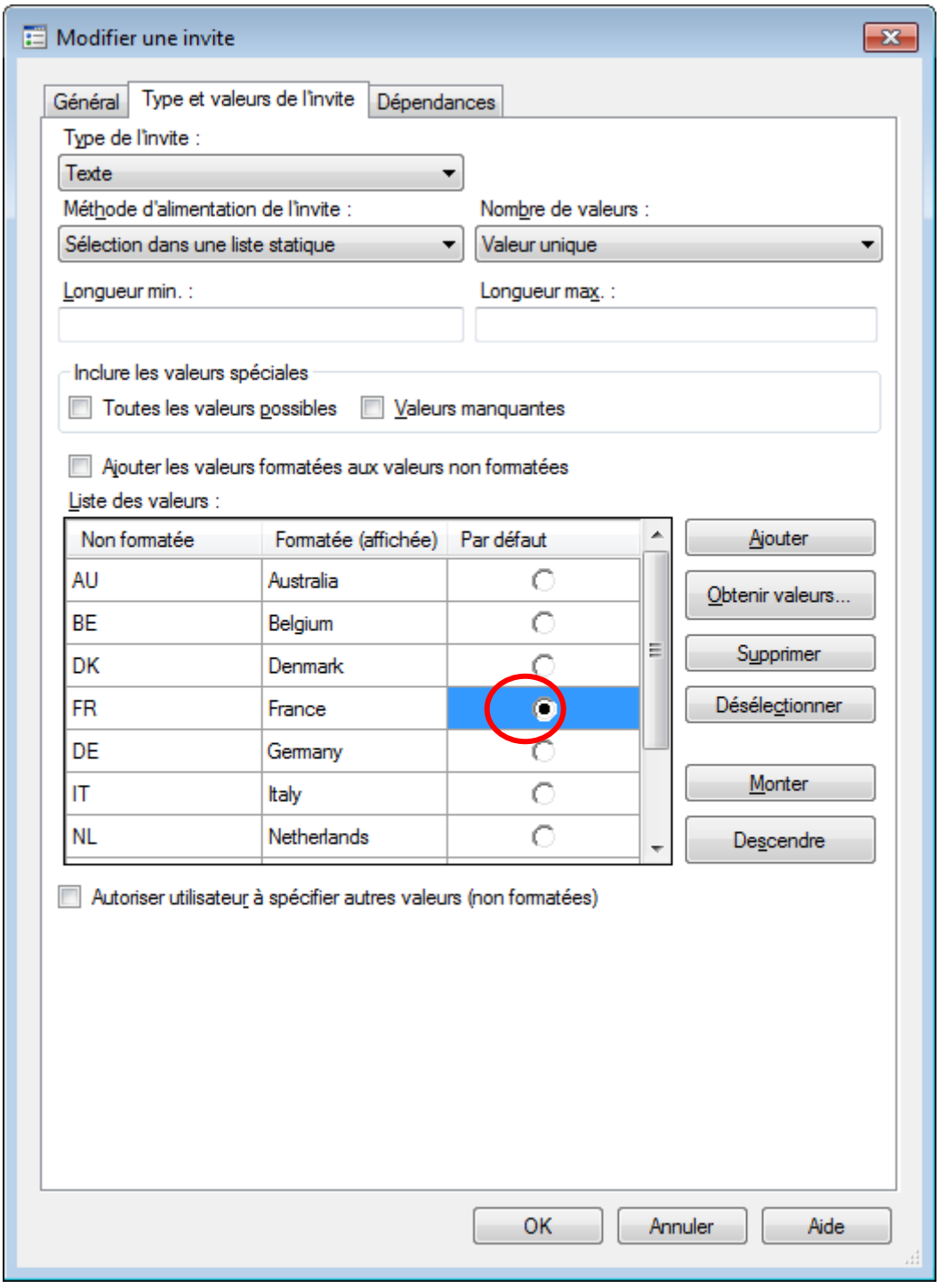
Parcourir



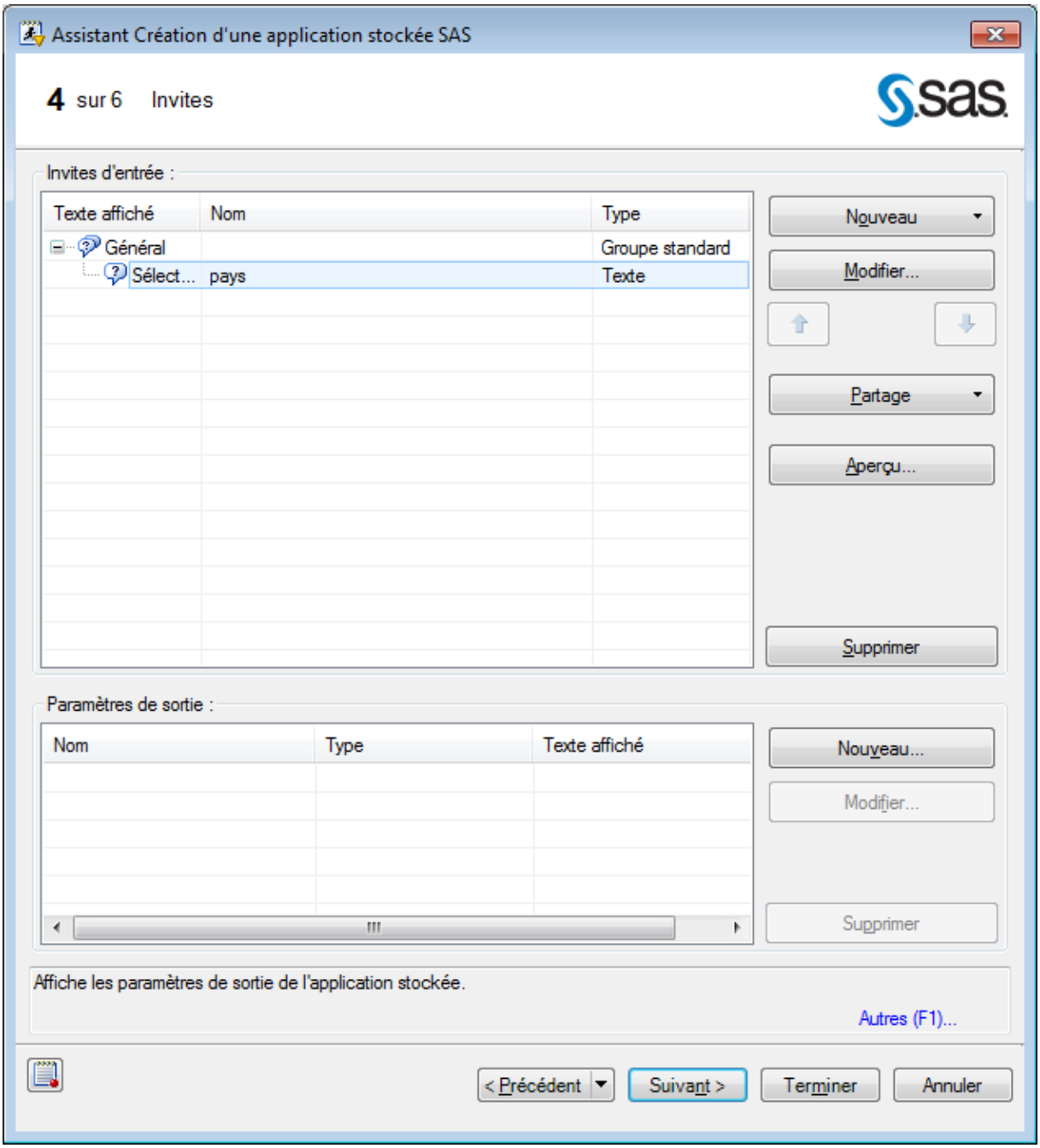
Sur le serveur SASApp, dans les bibliothèques, dans la Work → la table Query_for_organization_dim contient la liste des pays que l'on souhaite.



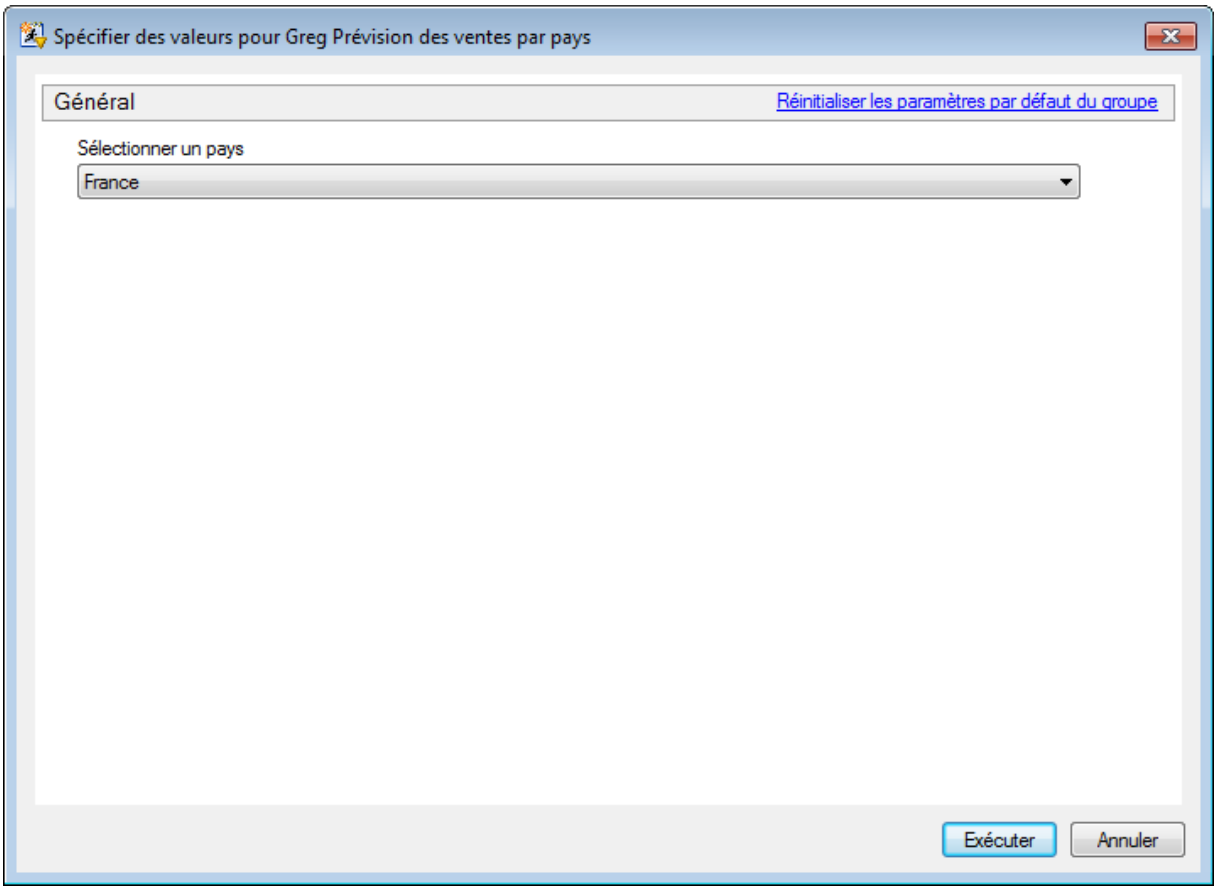
Obtenir les valeurs
Sélectionner tout avec la double flèche
OK


Modifier une invite
Général
Type et valeurs de l'invite
Dépendances
Type de l'invite :
Texte
Méthode d'alimentation de l'invite :
Sélection dans une liste statique
Nombre de valeurs :
Valeur unique
Longueur min. :
Longueur max. :
Inclure les valeurs spéciales
Toutes les valeurs possibles
Valeurs manquantes
Ajouter les valeurs formatées aux valeurs non formatées
Liste des valeurs :
Non formatée
Formatée (affichée)
Par défaut
AU
Australia
BE
Belgium
DK
Denmark
FR
France
DE
Germany
IT
Italy
NL
Netherlands
Ajouter
Obtenir valeurs...
Supprimer
Désélectionner
Monter
Descendre
Autoriser utilisateur à spécifier autres valeurs (non formatées)
OK
Annuler
Aide

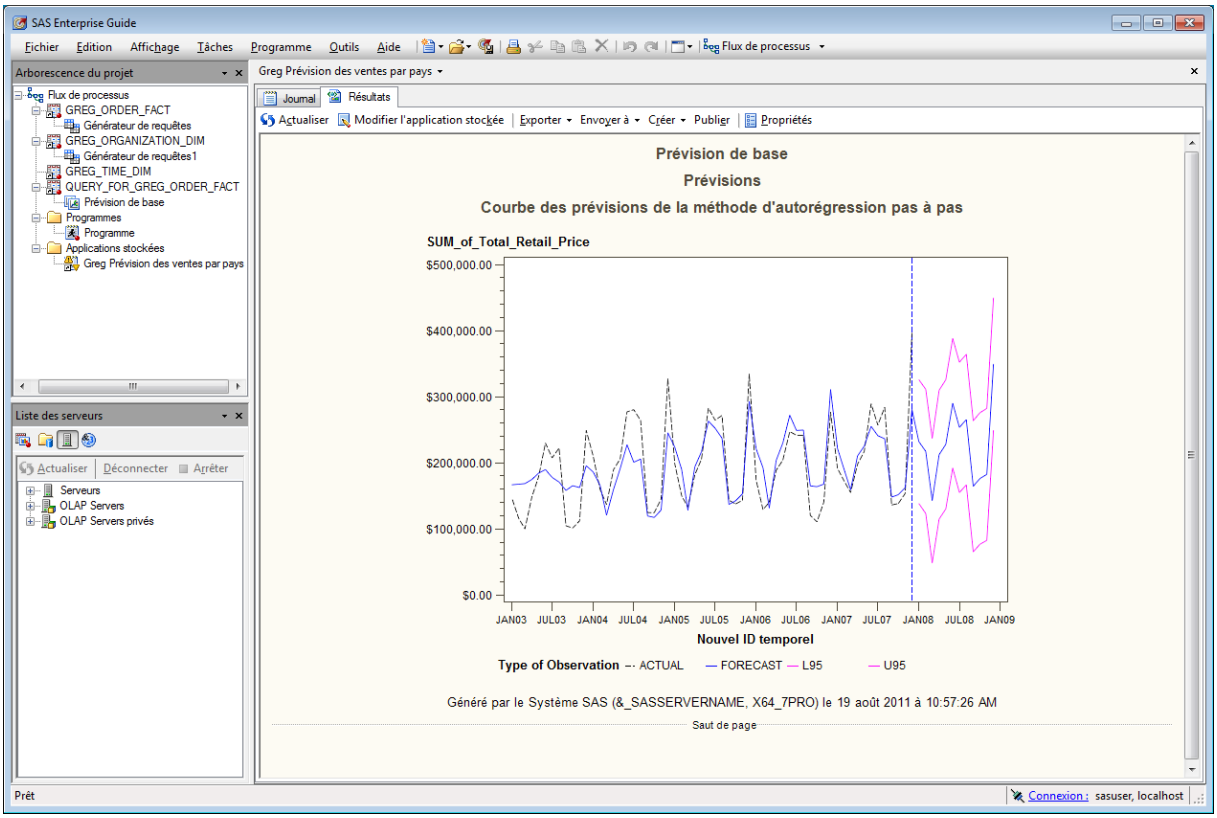
OK



Suivant
Terminer


Spécifier des valeurs pour Greg Prévision des ventes par pays
Général
Réinitialiser les paramètres par défaut du groupe
Sélectionner un pays
France
Exécuter
Annuler

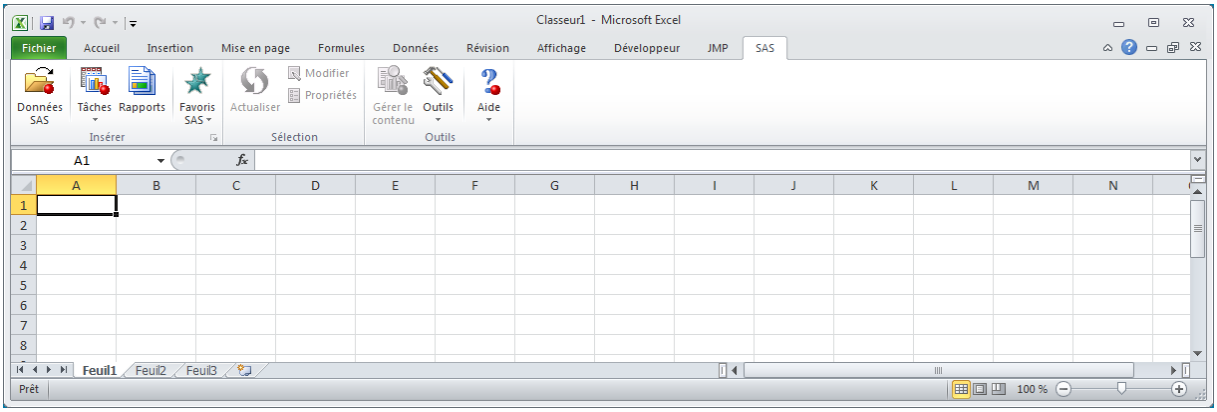
Exécuter


SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
Flux de processus
GREG_ORDER_FACT
Générateur de requêtes
GREG_ORGANIZATION_DIM
Générateur de requêtes1
GREG_TIME_DIM
QUERY_FOR_GREG_ORDER_FACT
Prévision de base
Programmes
Programme
Applications stockées
Greg Prévision des ventes par pays
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Greg Prévision des ventes par pays
Journal
Résultats
Actualiser Modifier l'application stockée Exporter Envoyer à Créer Publier Propriétés
Prévision de base
Prévisions
Courbe des prévisions de la méthode d'autorégression pas à pas
SUM_of_Total_Retail_Price
$500,000.00
$400,000.00
$300,000.00
$200,000.00
$100,000.00
$0.00
JAN03 JUL03 JAN04 JUL04 JAN05 JUL05 JAN06 JUL06 JAN07 JUL07 JAN08 JUL08 JAN09
Nouvel ID temporel
Type of Observation ACTUAL FORECAST L95 U95
Généré par le Système SAS (&_SASSERVERNAME, X64_7PRO) le 19 août 2011 à 10:57:26 AM
Saut de page
Prêt
Connexion : sasuser, localhost

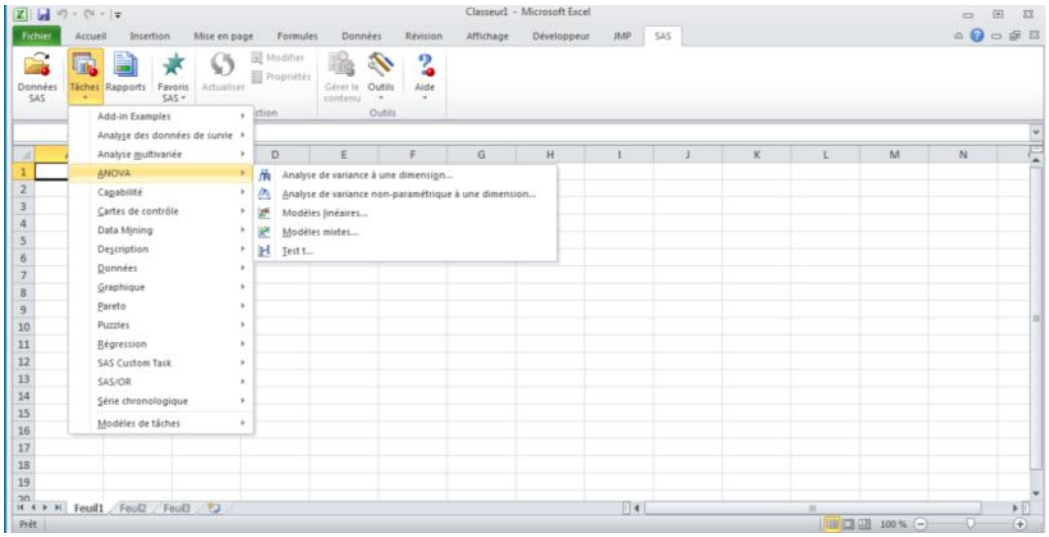
Utilisation de l'application stockée depuis d'Add-in Office :

L'add-in SAS pour Microsoft Office permet d'accéder à la puissance de SAS depuis Microsoft Word, Excel, ou Power Point.

Voici le bandeau que vous aurez par exemple dans Microsoft Excel



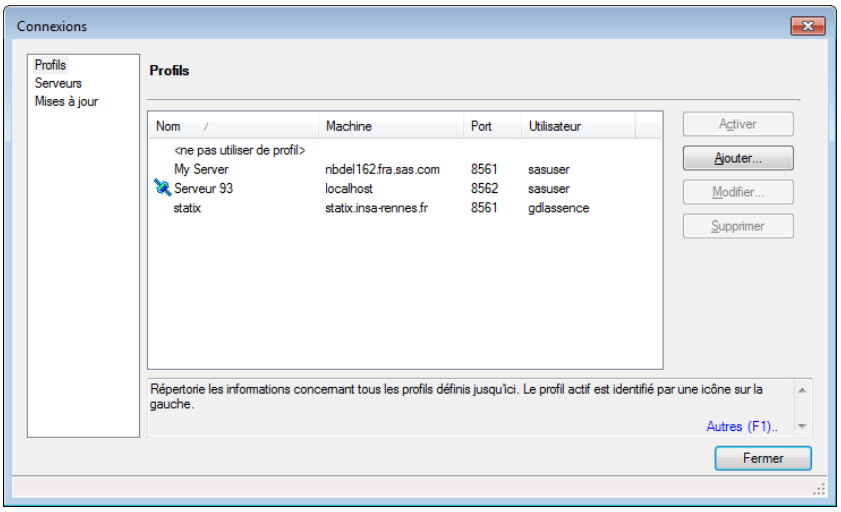
Depuis Excel, vous pouvez ouvrir des bases de données gigantesques depuis des serveurs SAS, et par défaut, seules 500 lignes seront affichées. Il est simple de modifier ce paramètre par défaut. Vous pouvez accéder à des tables ou à des cubes de façon tout à fait transparente. Vous avez aussi accès à des tâches de manipulation de données, d'analyses statistiques descriptives et inférentielles, de Data Mining, etc.



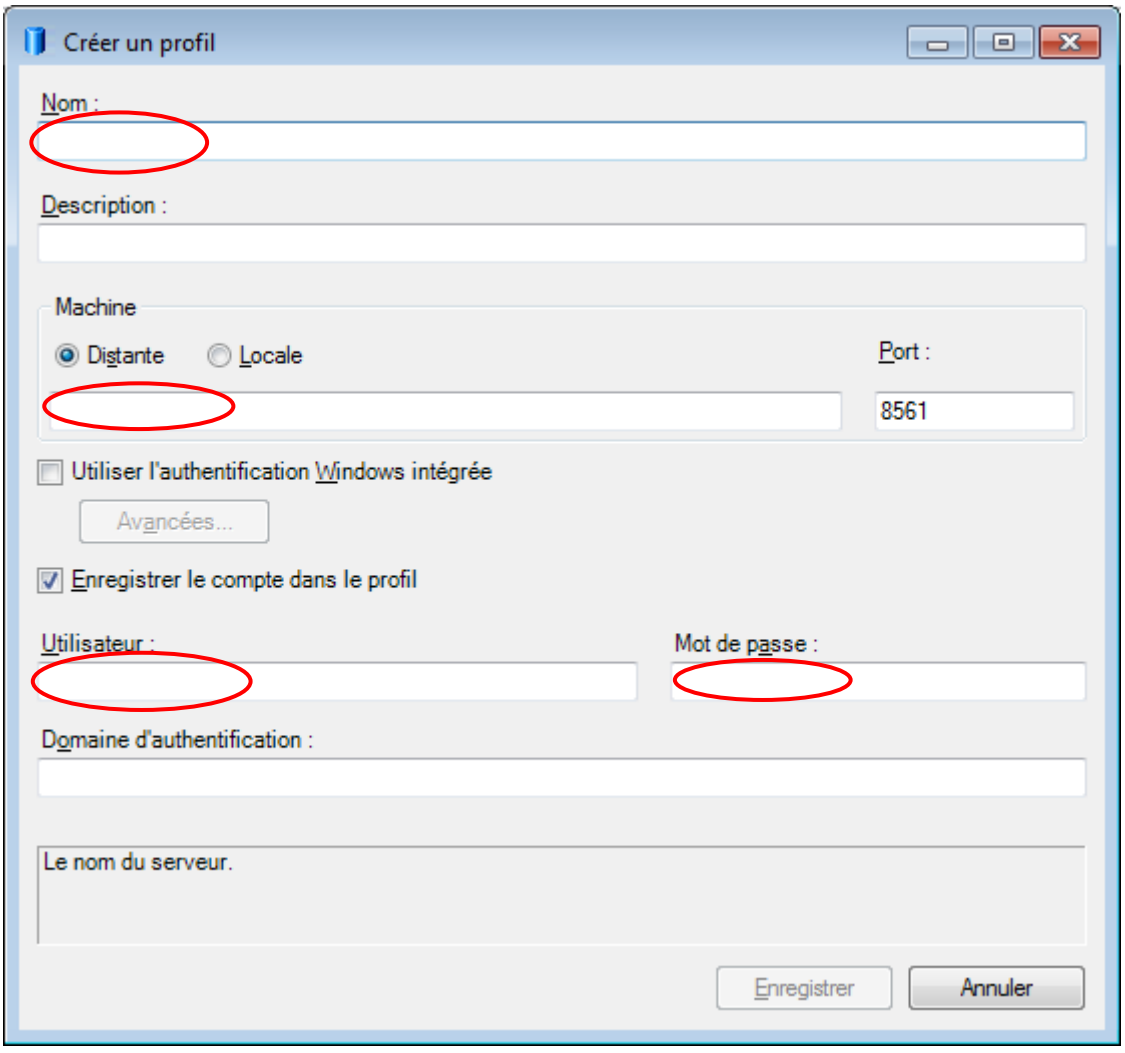
Si l'Add-in Office est installé, aller dans Microsoft Word, ou Excel, ou Power Point,

Remarque, il peut être nécessaire de rentrer les paramètres de connexion avant de lancer cet Add-in.

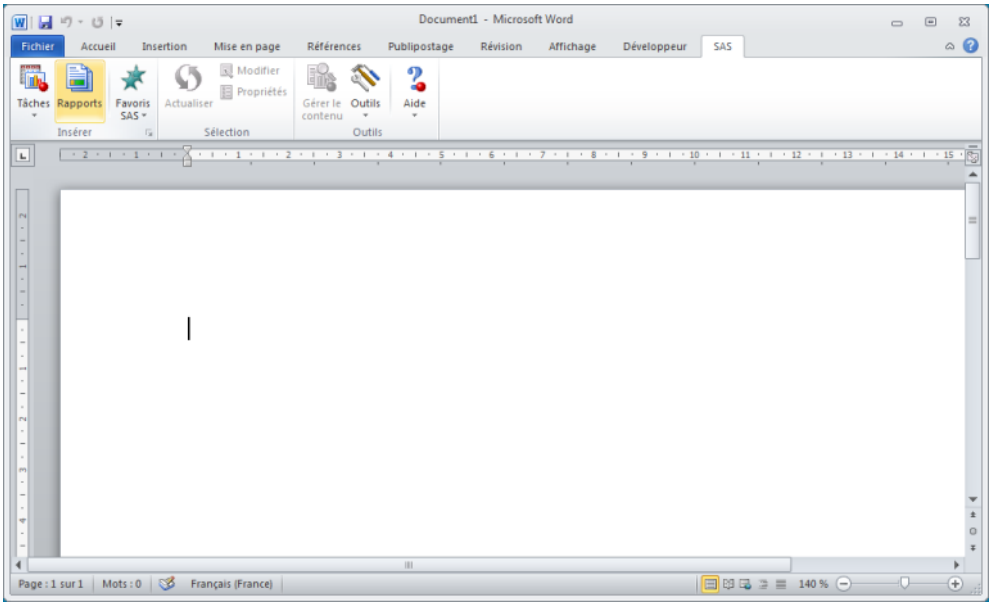
Dans l'onglet SAS de Microsoft Word → outil → connexion



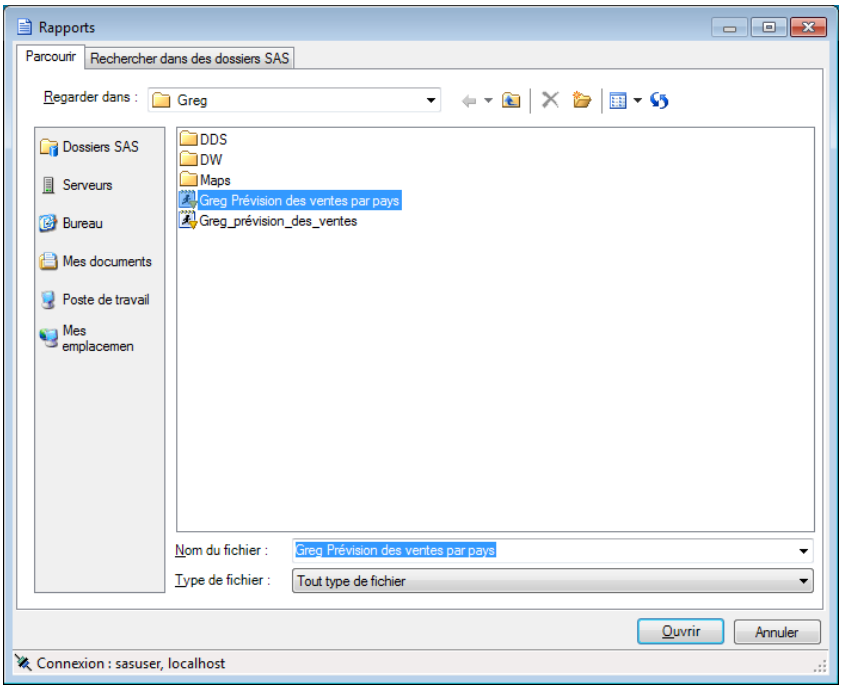
S'il n'y a pas de connexion, ajoutez-en une :

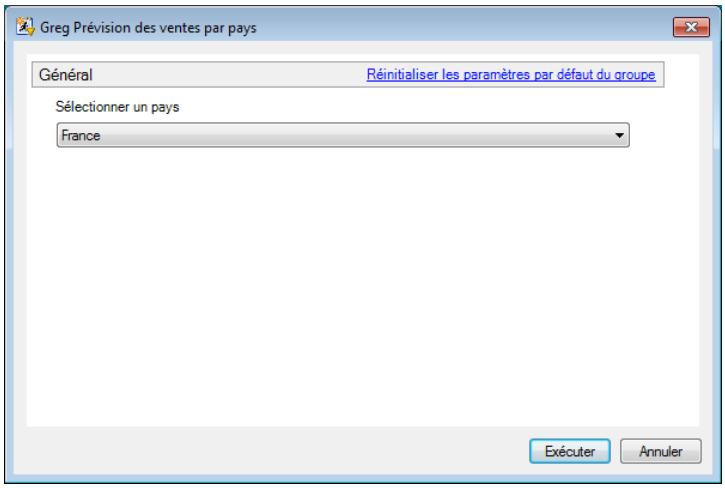


Il faut rentrer un nom de connexion, le nom du serveur, l'utilisateur et son mot de passe. Enregister.

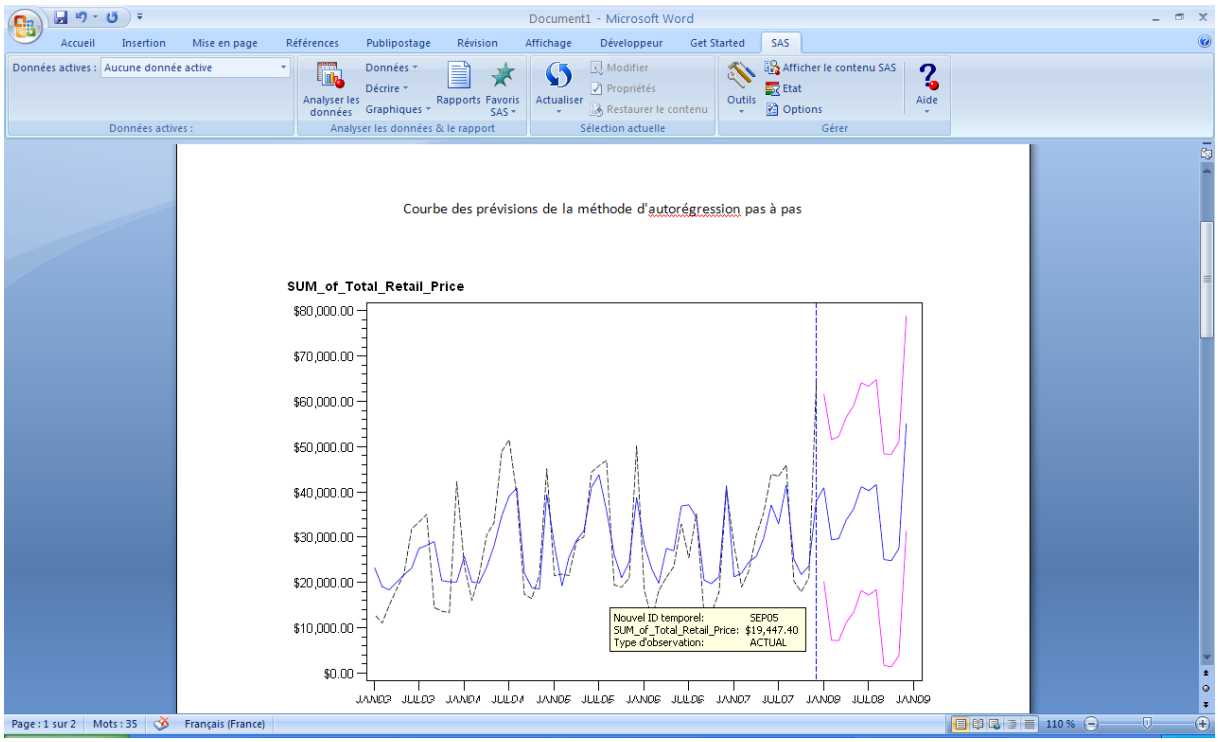


Dans l'onglet SAS → Rapport



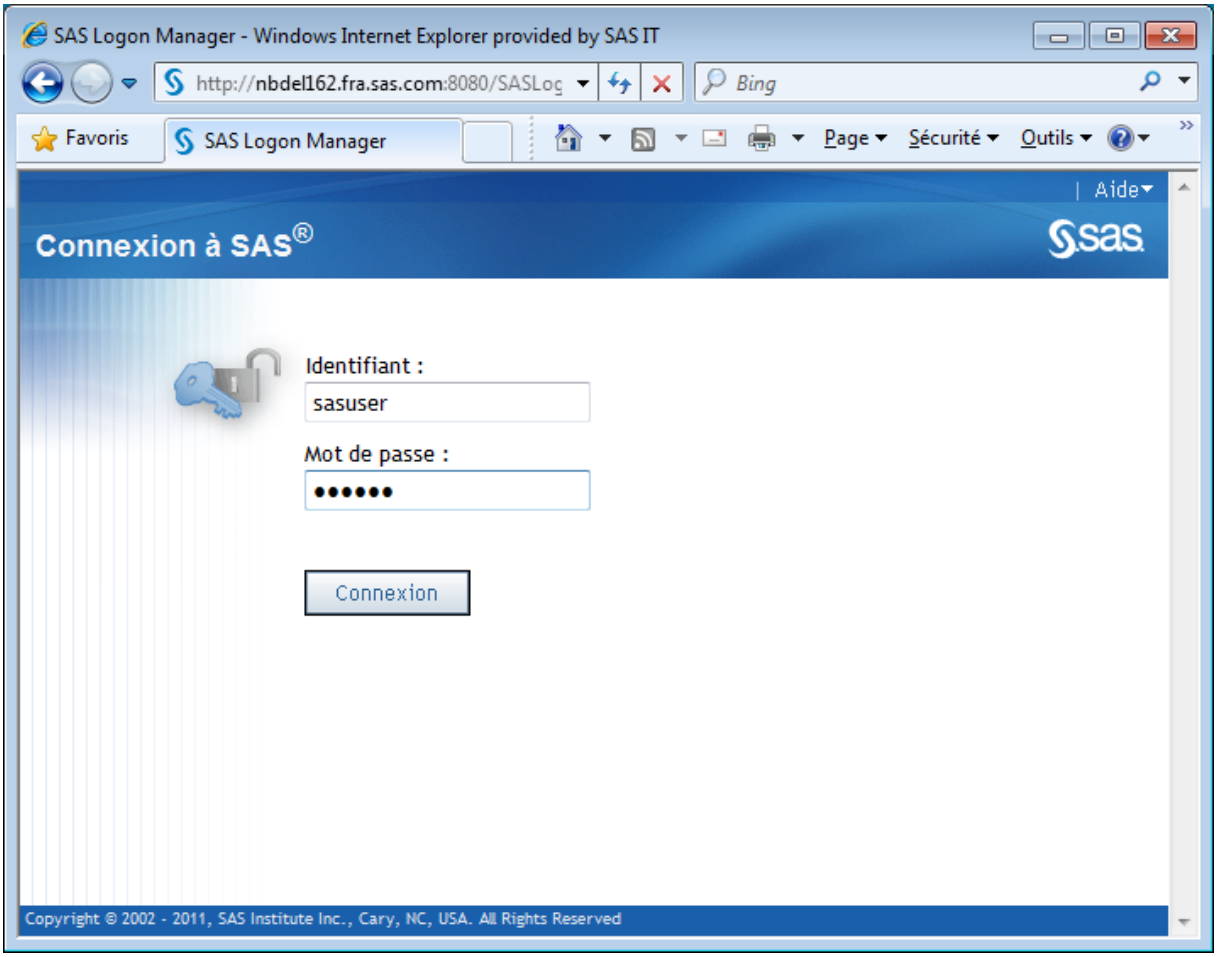


Sélection un pays



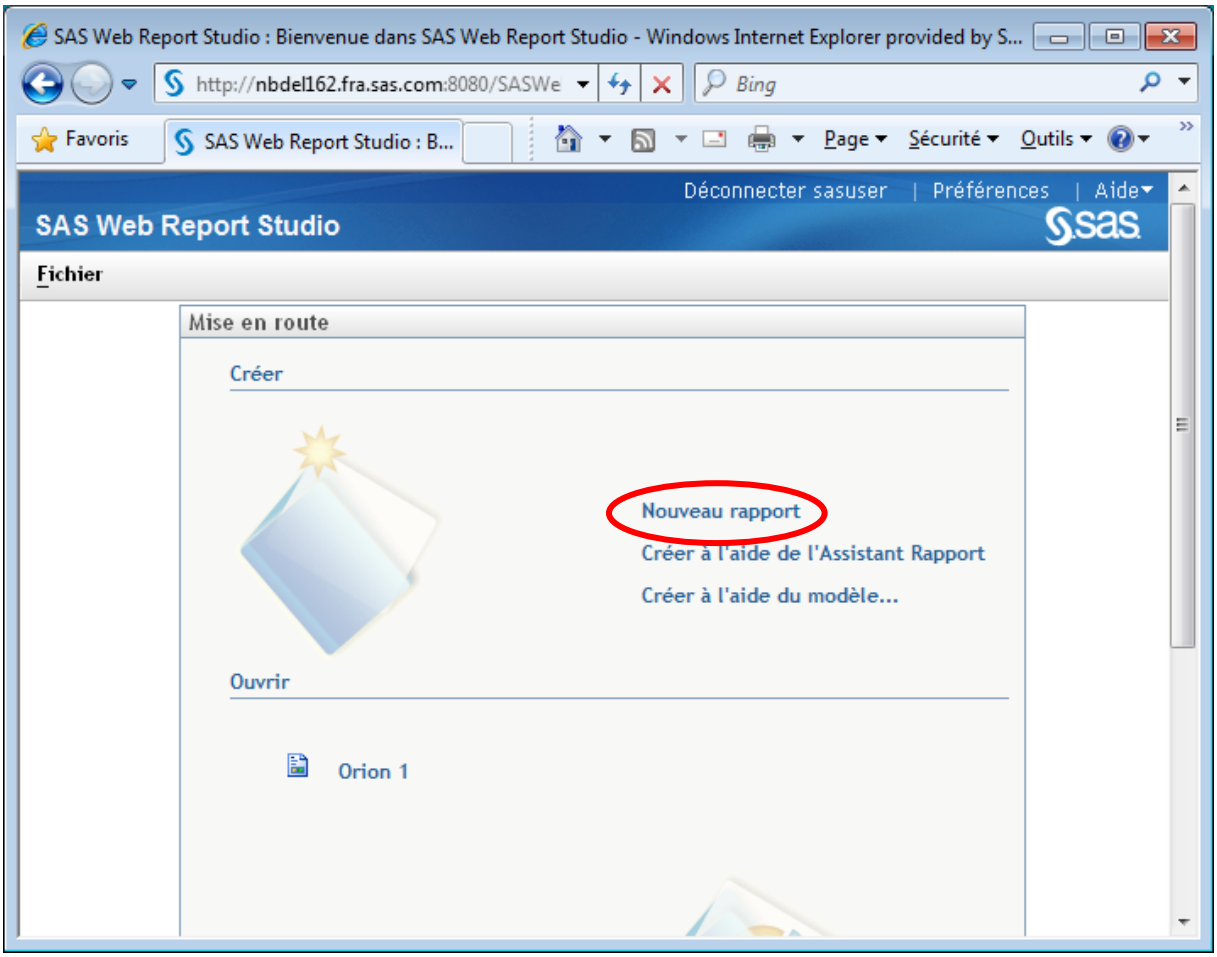
Et voilà.

Utilisation de la procédure stockée dans un rapport Web Report Studio

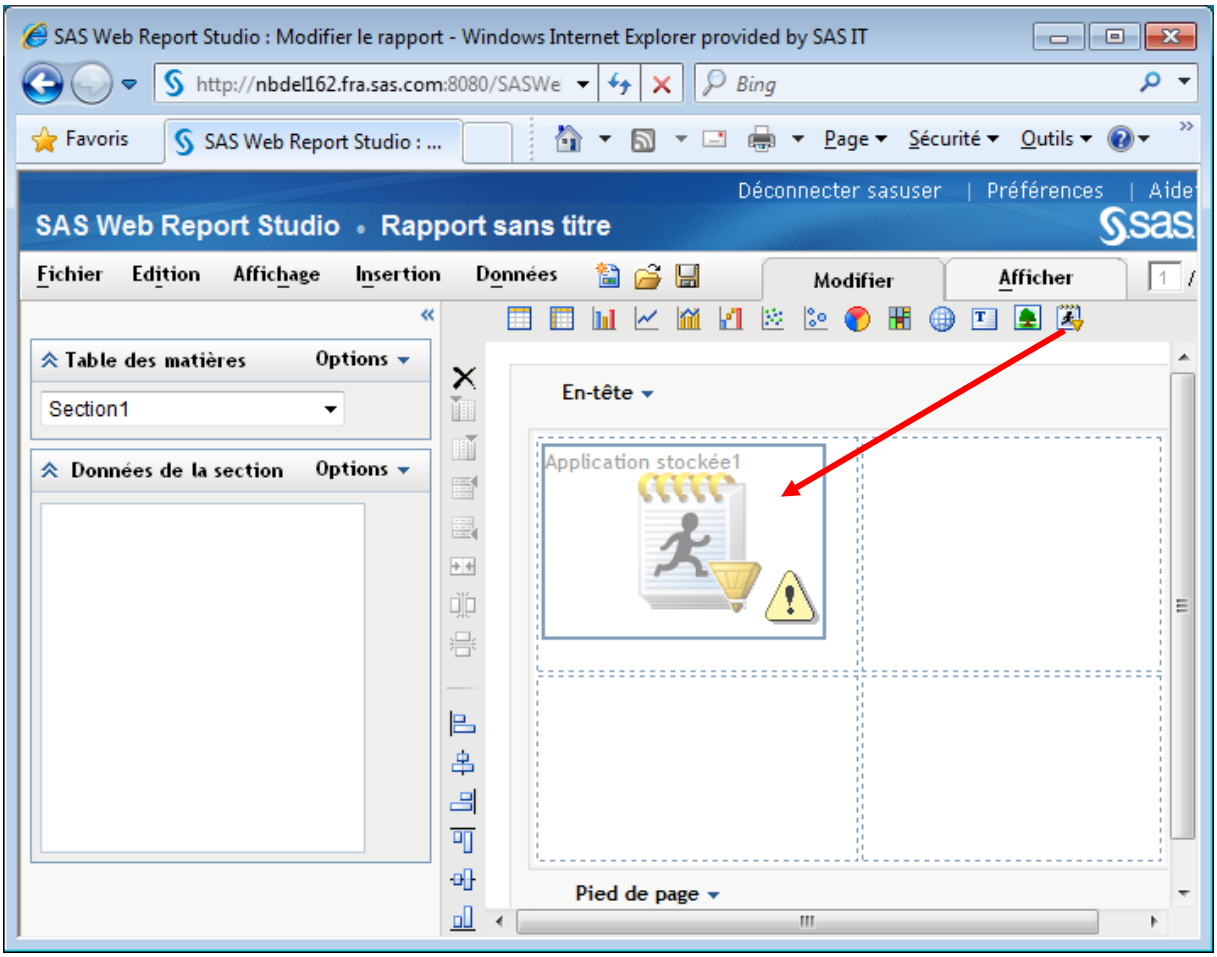


Lancer SAS Web Report Studio : http://nomduserveur:8080/SASWebReportStudio

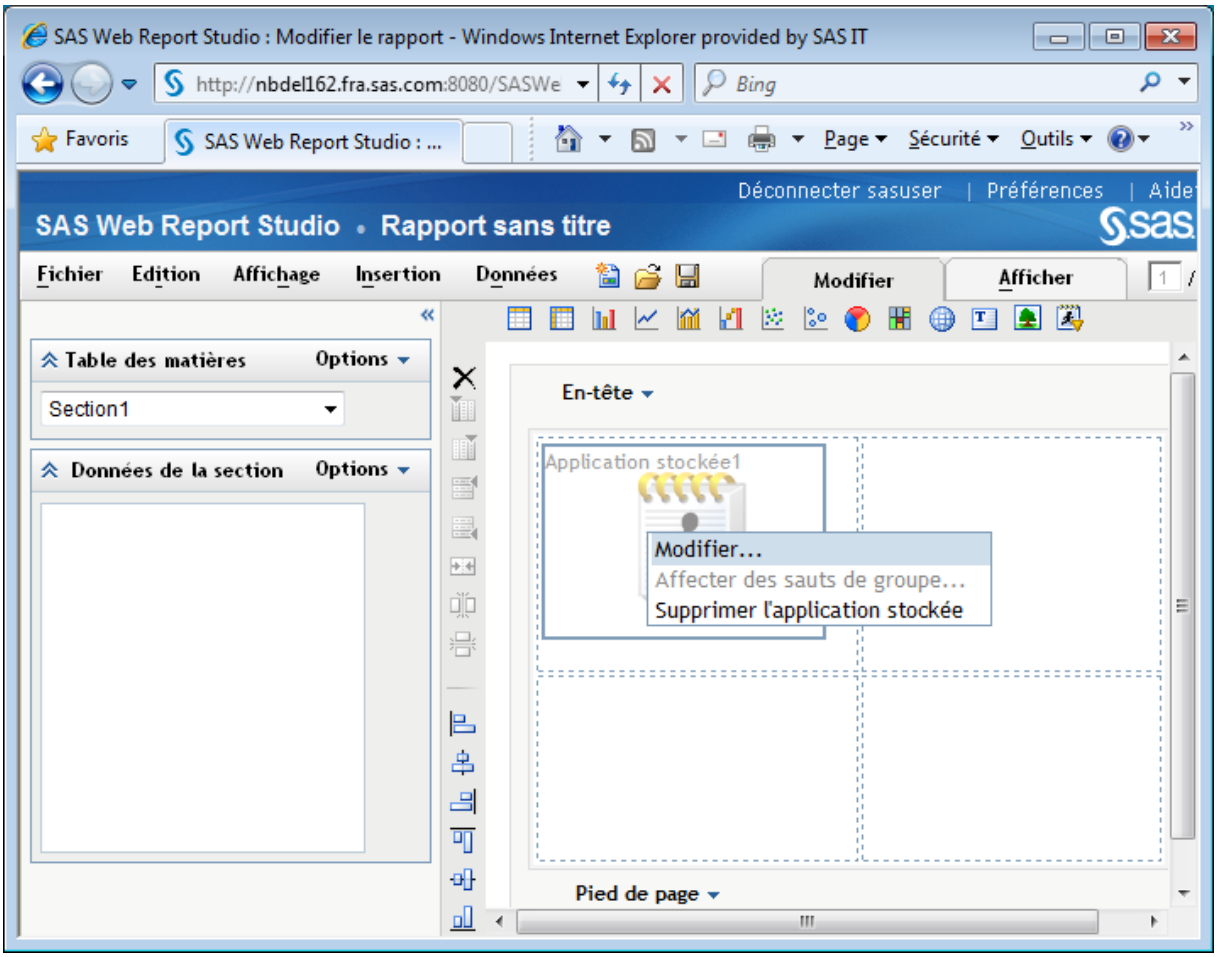
Entrer l'identifiant et le mot de passe.



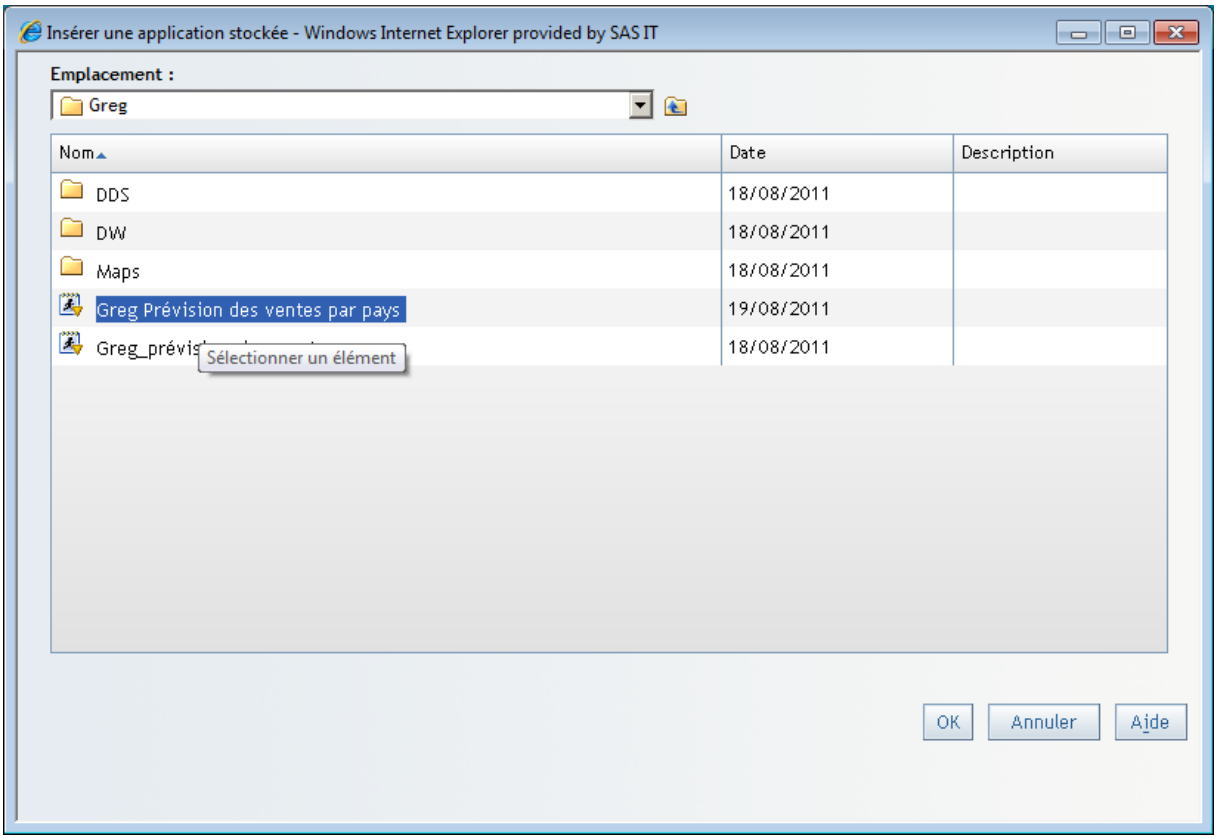
Créer un nouveau rapport :



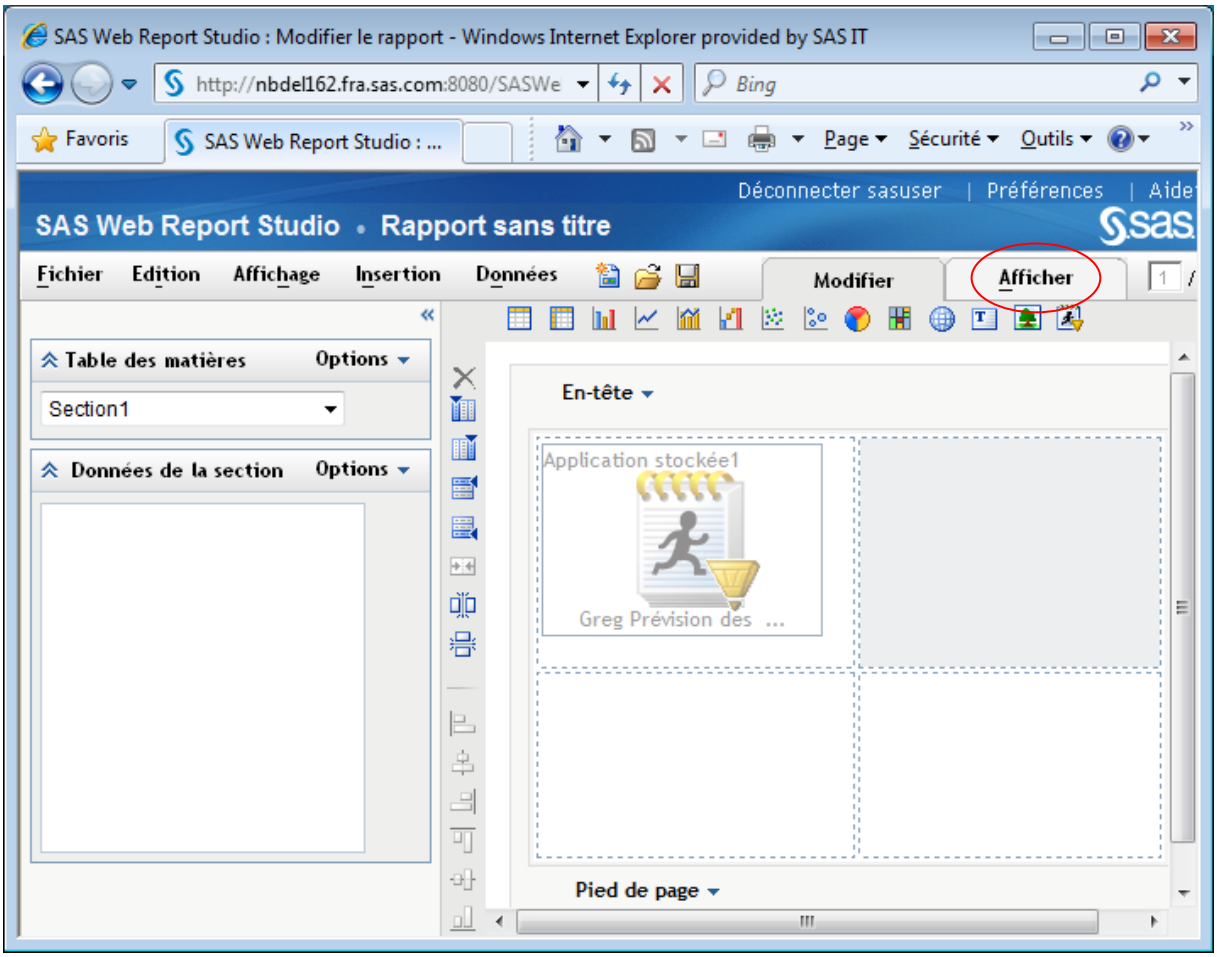
Ajouter la procédure stockée.



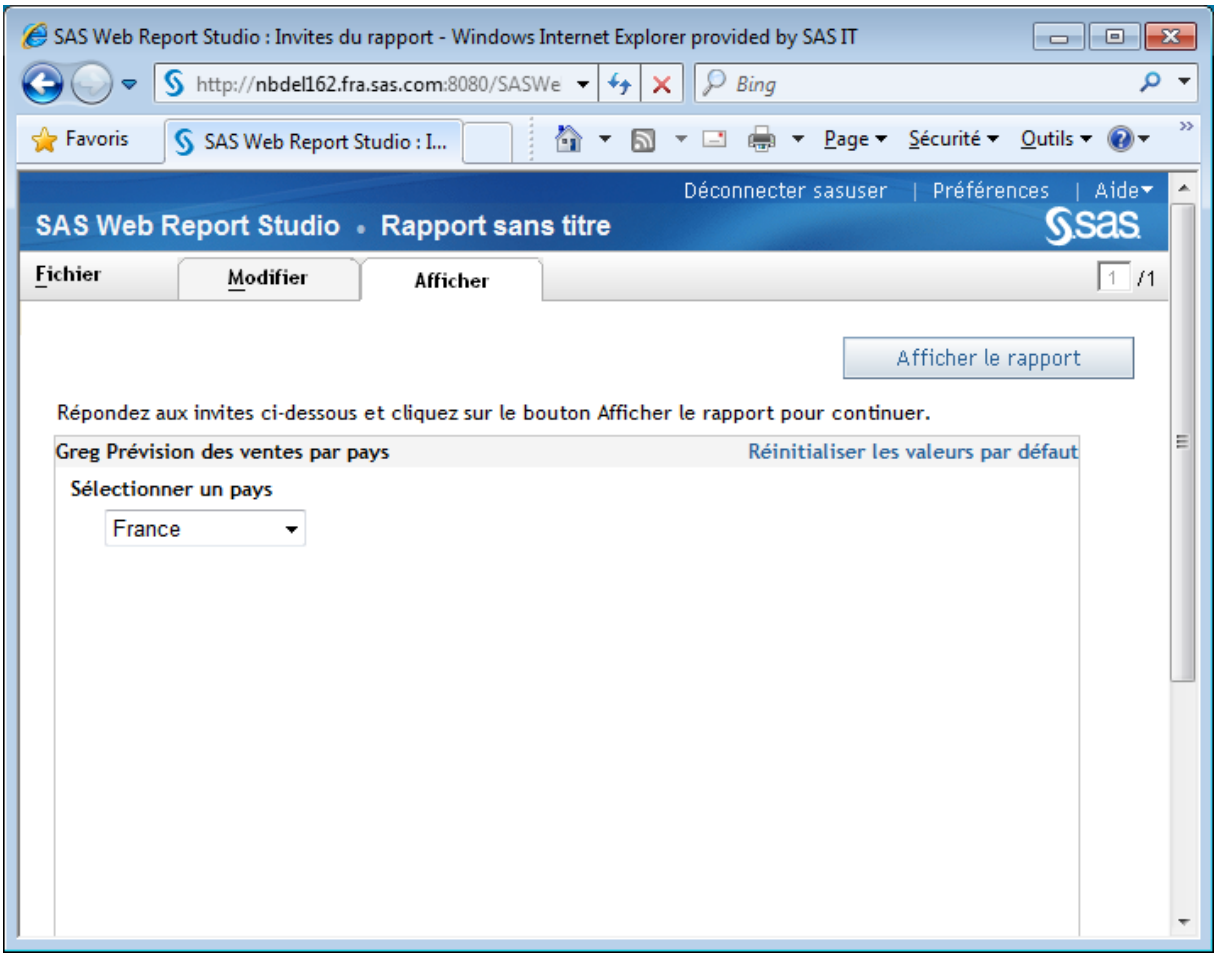
Clic-droit → Modifier



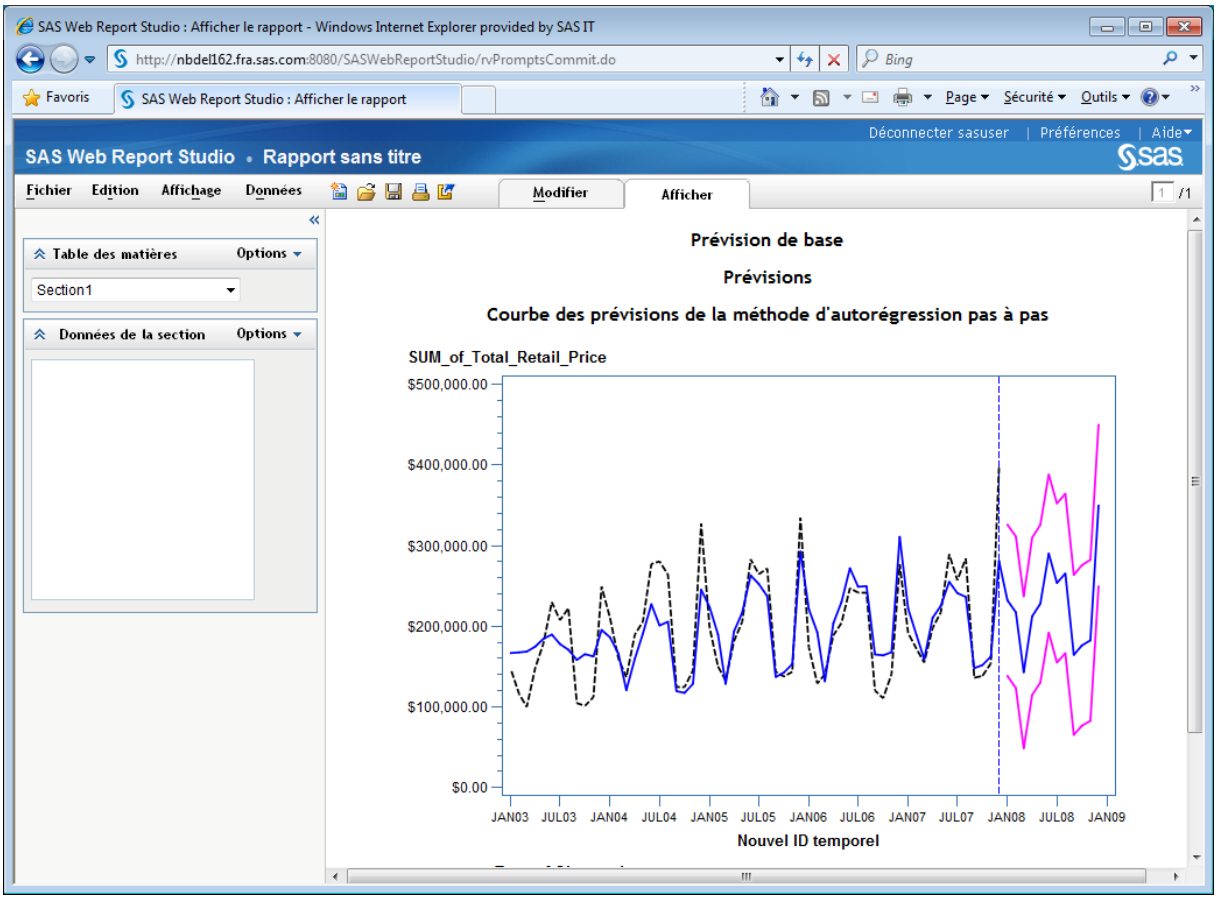
Sélectionner votre procédure stockée



Afficher le rapport

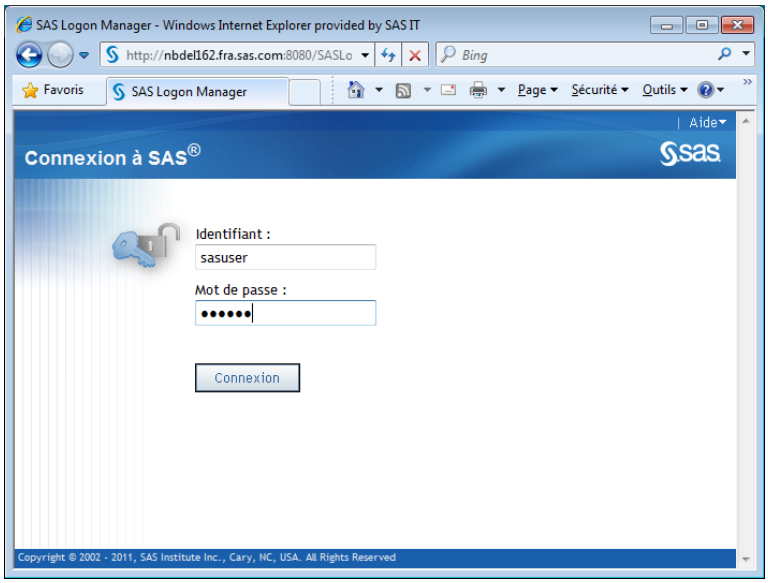


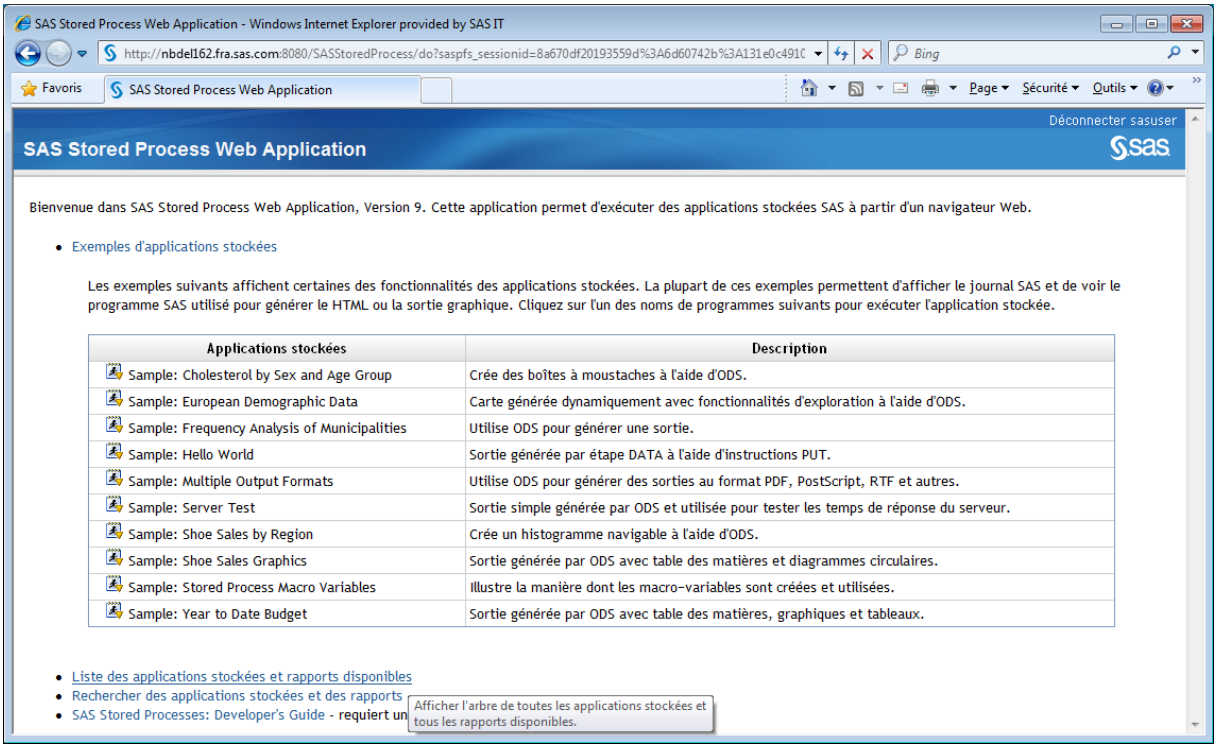
Sélectionner un pays
Afficher le rapport


SAS Web Report Studio : Afficher le rapport - Windows Internet Explorer provided by SAS IT
http://nbdel162.fra.sas.com:8080/SASWebReportStudio/rvPromptsCommit.do
Bing
Favoris
SAS Web Report Studio : Afficher le rapport
Page
Sécurité
Outils
Déconnecter sasuser
Préférences
Aide
SAS Web Report Studio • Rapport sans titre
Fichier
Edition
Affichage
Données
Modifier
Afficher
1 /1
Table des matières
Options
Section1
Données de la section
Options
Prévision de base
Prévisions
Courbe des prévisions de la méthode d'autorégression pas à pas
SUM_of_Total_Retail_Price
$500,000.00
$400,000.00
$300,000.00
$200,000.00
$100,000.00
$0.00
JAN03
JUL03
JAN04
JUL04
JAN05
JUL05
JAN06
JUL06
JAN07
JUL07
JAN08
JUL08
JAN09
Nouvel ID temporel

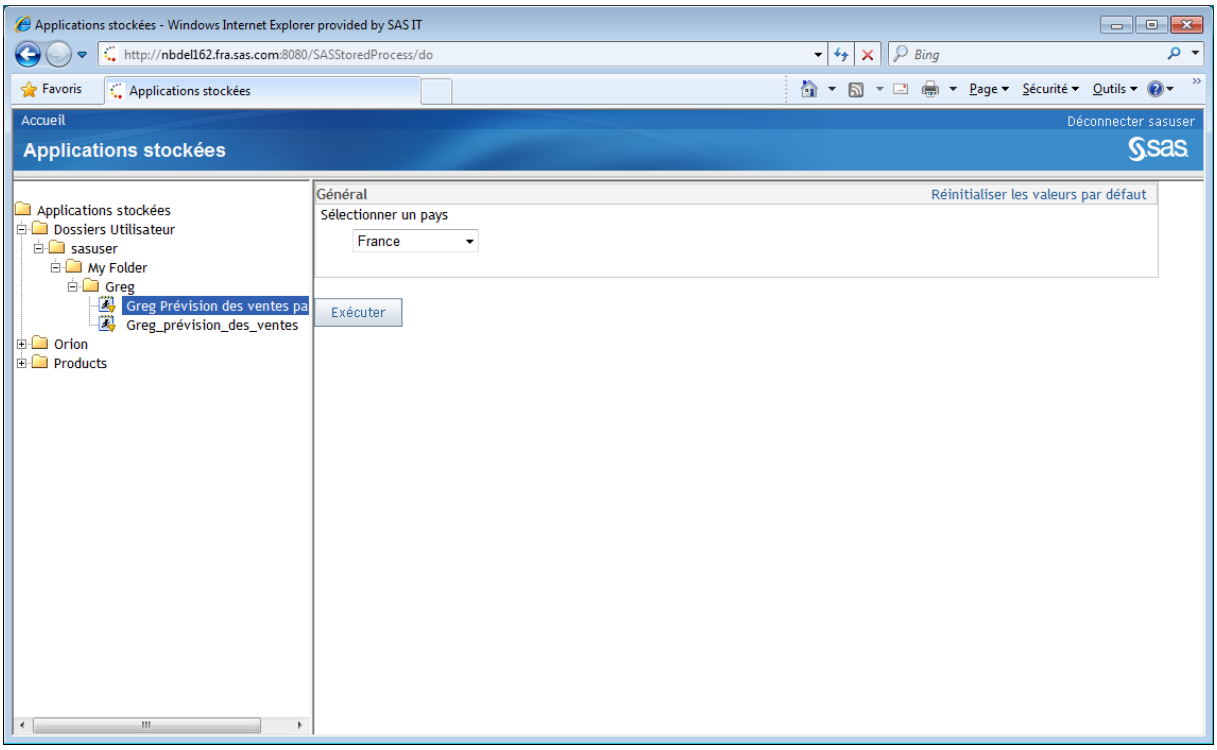
Utilisation de la procédure stockée depuis SAS Web Stored Process

Entrer l'adresse suivant dans un navigateur Web.
http://nomduserveur:8080/SASStoredProcess

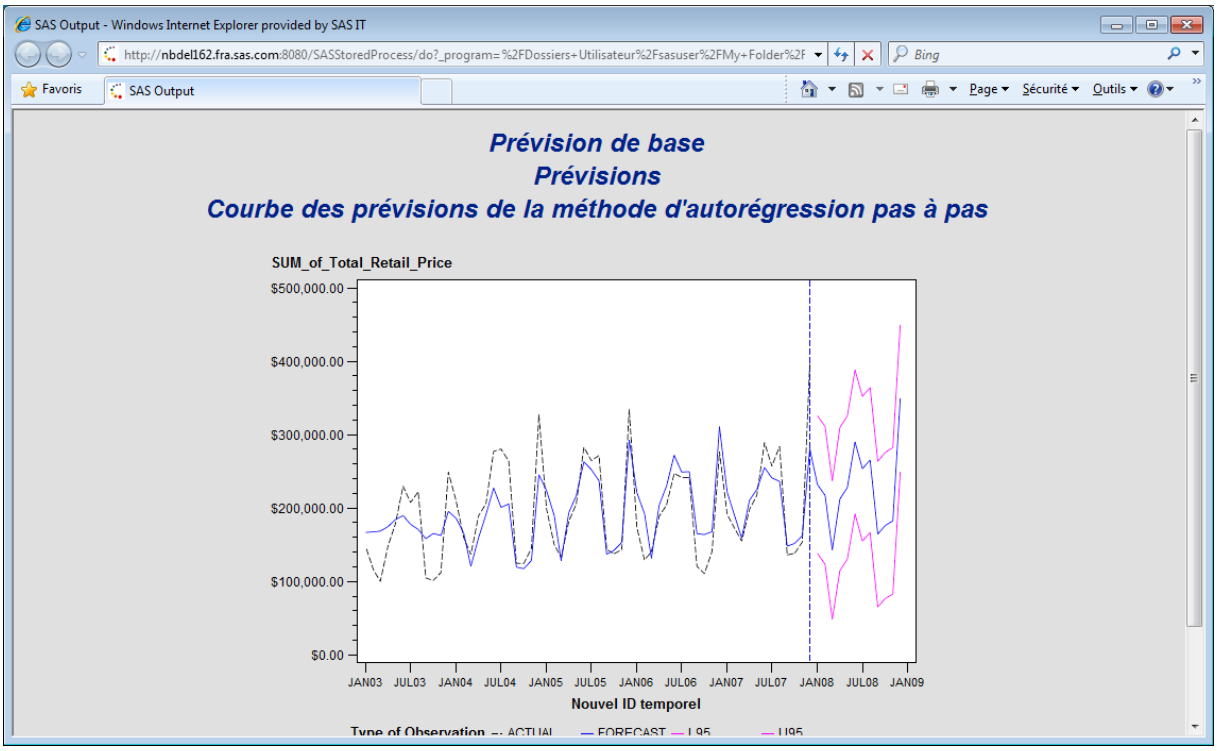




Sélectionner « List Available Stored Processes ».

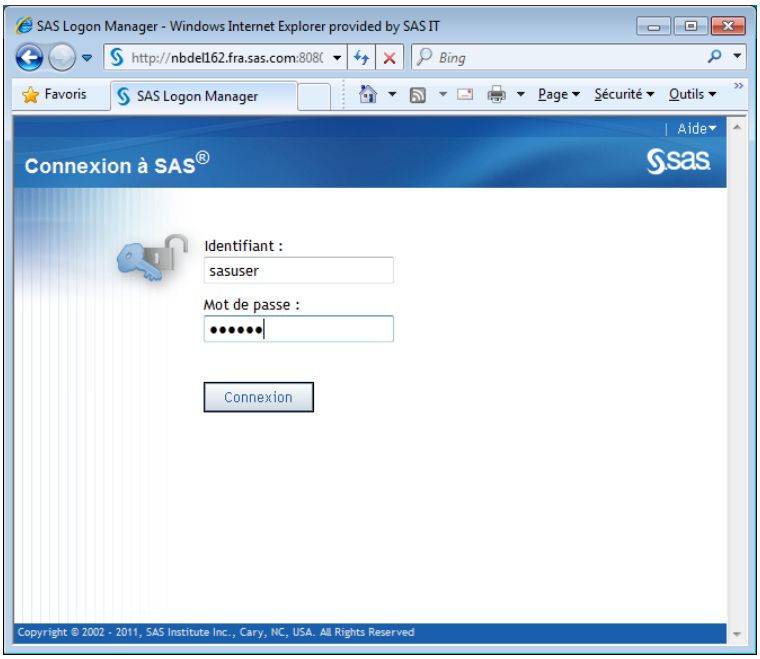


Sélectionner votre procédure stockée
Sélectionner le pays,

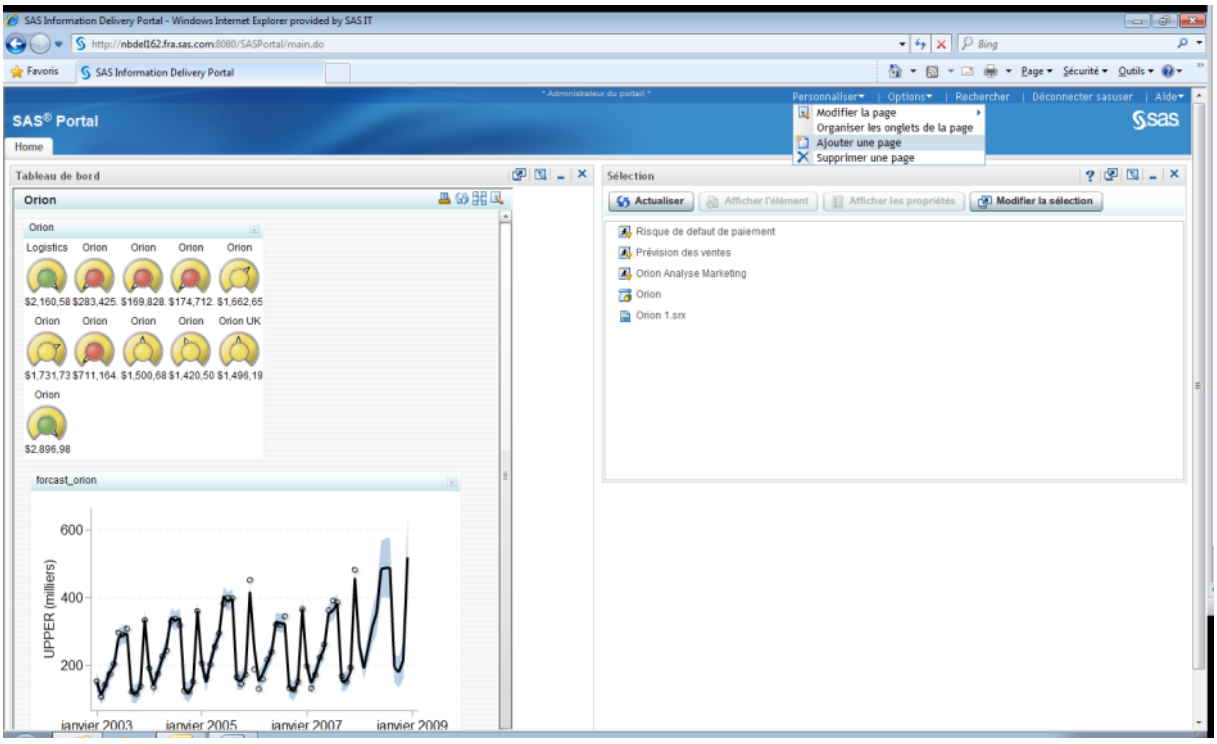


Utilisation de la procédure stockée depuis le Portail

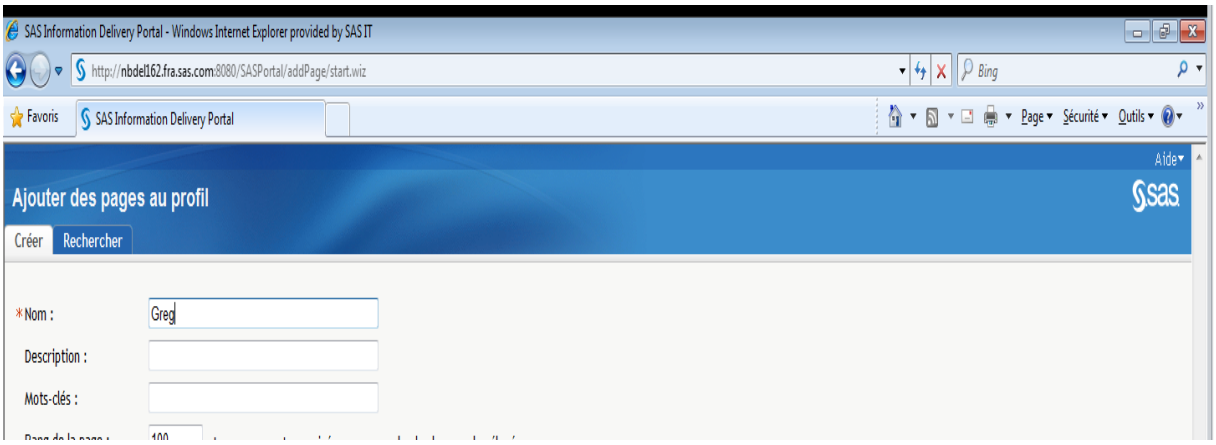
Entrer l'adresse suivante dans un navigateur Web.
http://nomduserveur:8080/SASPortal



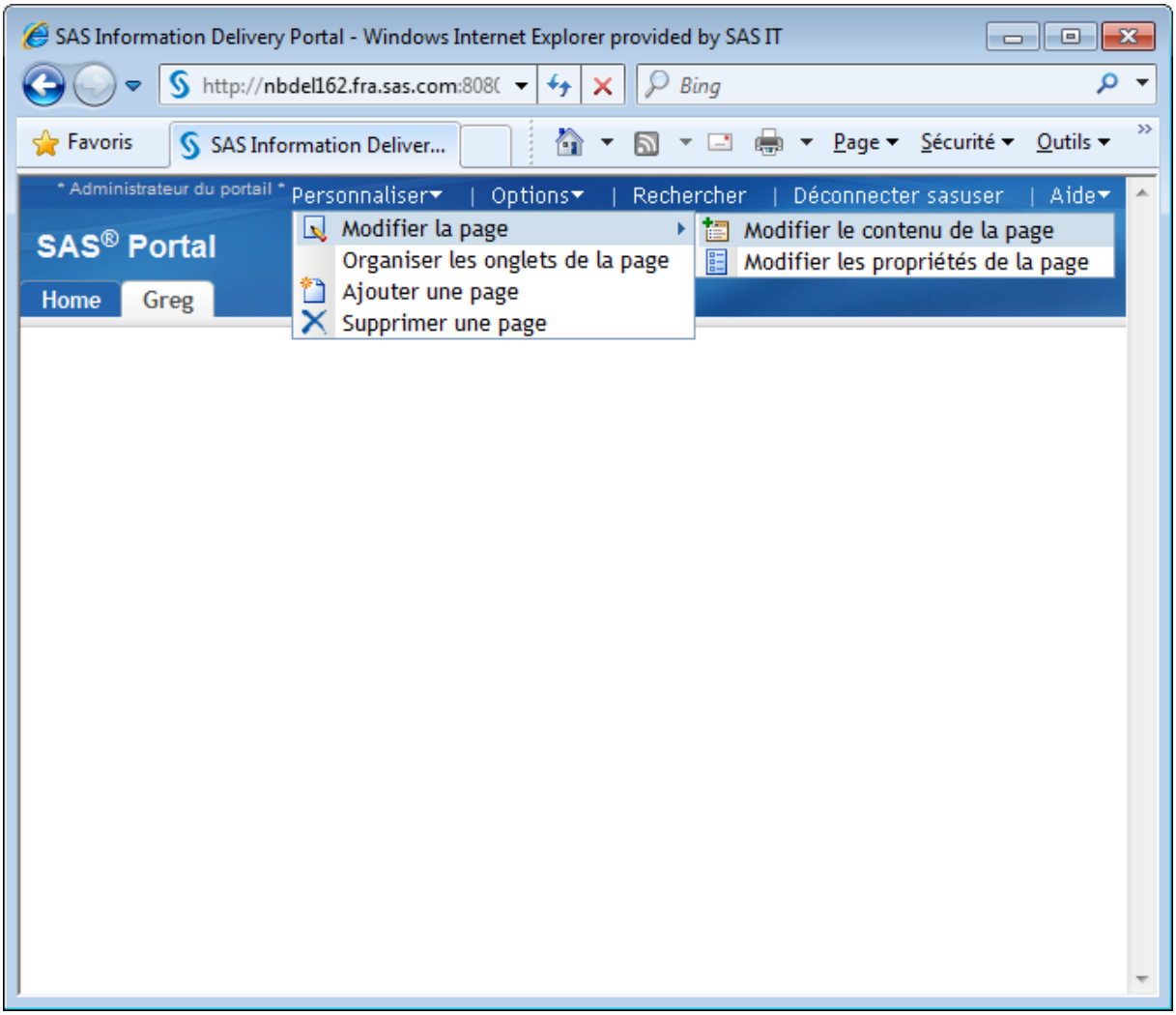
Entrer votre identifiant et votre mot de passe.



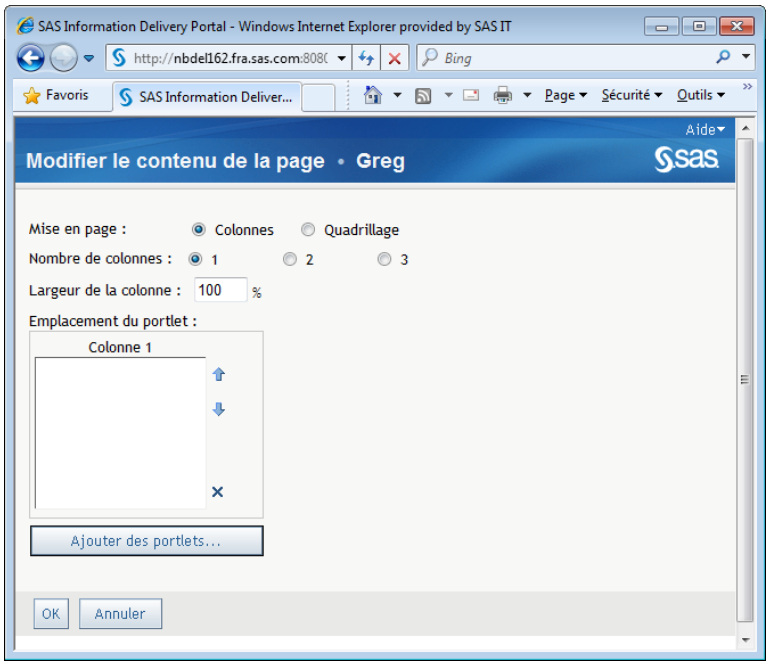
Si c'est la première fois que vous vous connectez au portail, il n'y a rien, il faut donc aller dans Personnaliser → ajouter une page.



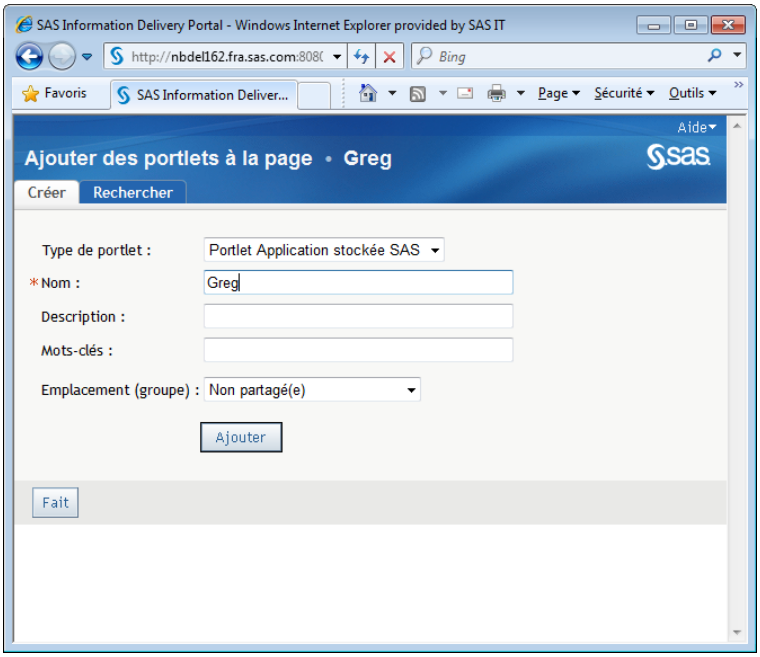
Donnez un nom à votre page
Cliquez sur Ajouter
Puis sur Fait
Vous obtenez une page blanche à votre nom.



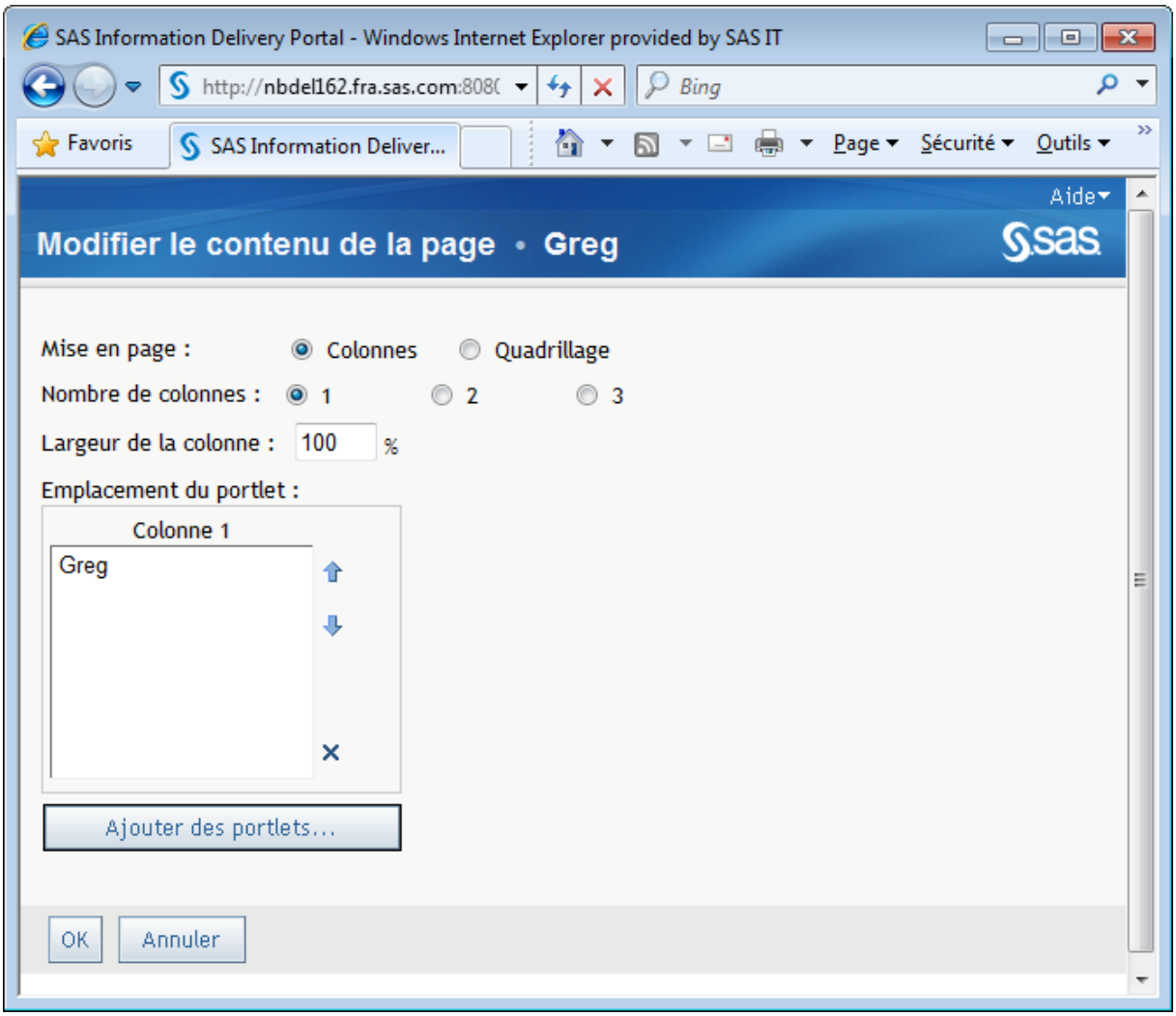
Dans les Personnaliser → Modifier la page → Modifier le contenu de la page



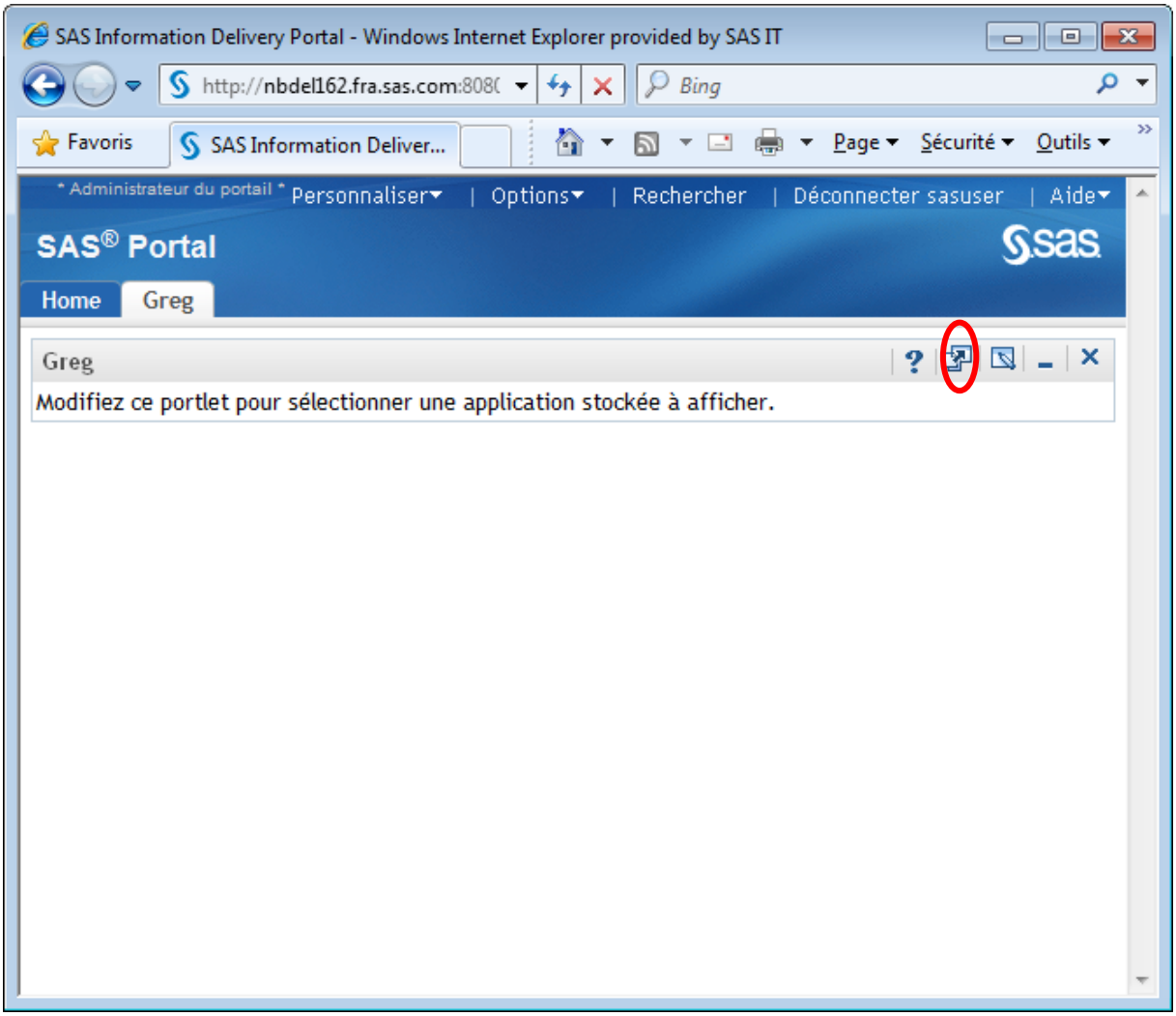
Ajouter un portlets



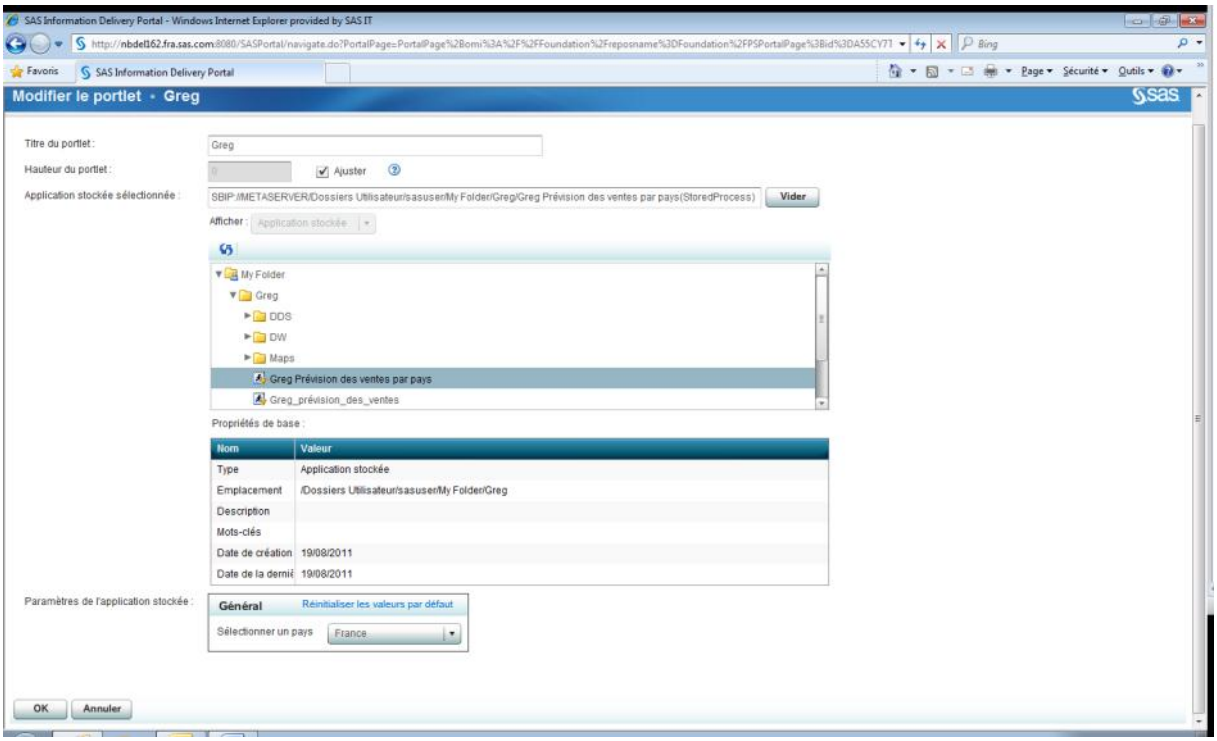
Sélectionner le portlet d’application stockée
Lui donner un nom
Ajouter
Fait


SAS Information Delivery Portal - Windows Internet Explorer provided by SAS IT
http://nbdel162.fra.sas.com:808(
Bing
Favoris
SAS Information Deliver...
Page
Sécurité
Outils
Aide
Modifier le contenu de la page • Greg
sas
Mise en page :
Colonnes
Quadrillage
Nombre de colonnes :
1
2
3
Largeur de la colonne :
100
%
Emplacement du portlet :
Colonne 1
Greg
Ajouter des portlets...
OK
Annuler

OK



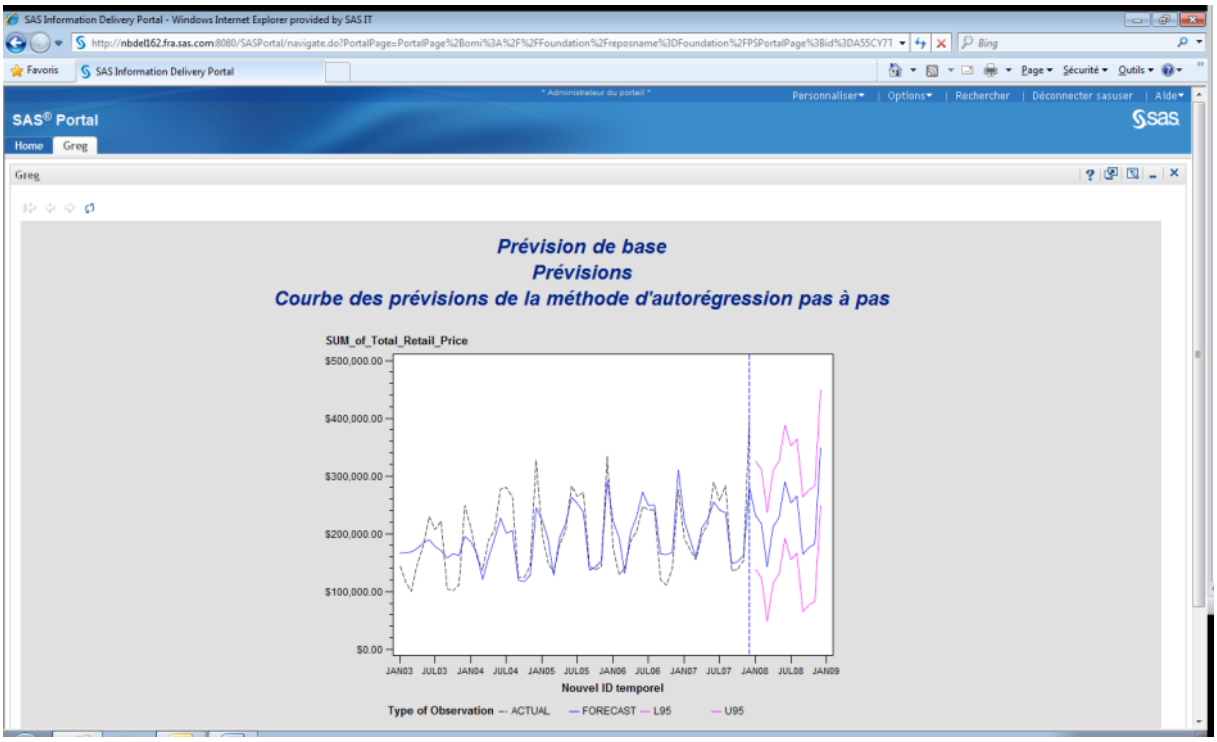
Modifier le contenu



Sélectionner votre dossier,

Sélectionner votre procédure

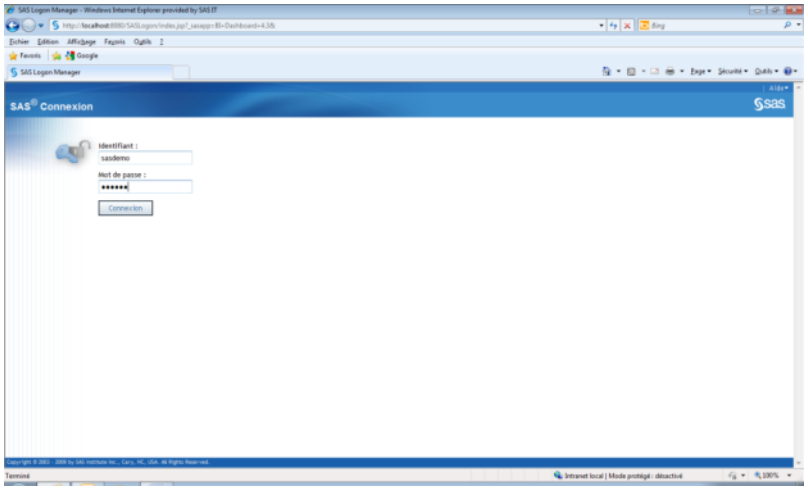
Cliquer sur OK, en bas à gauche

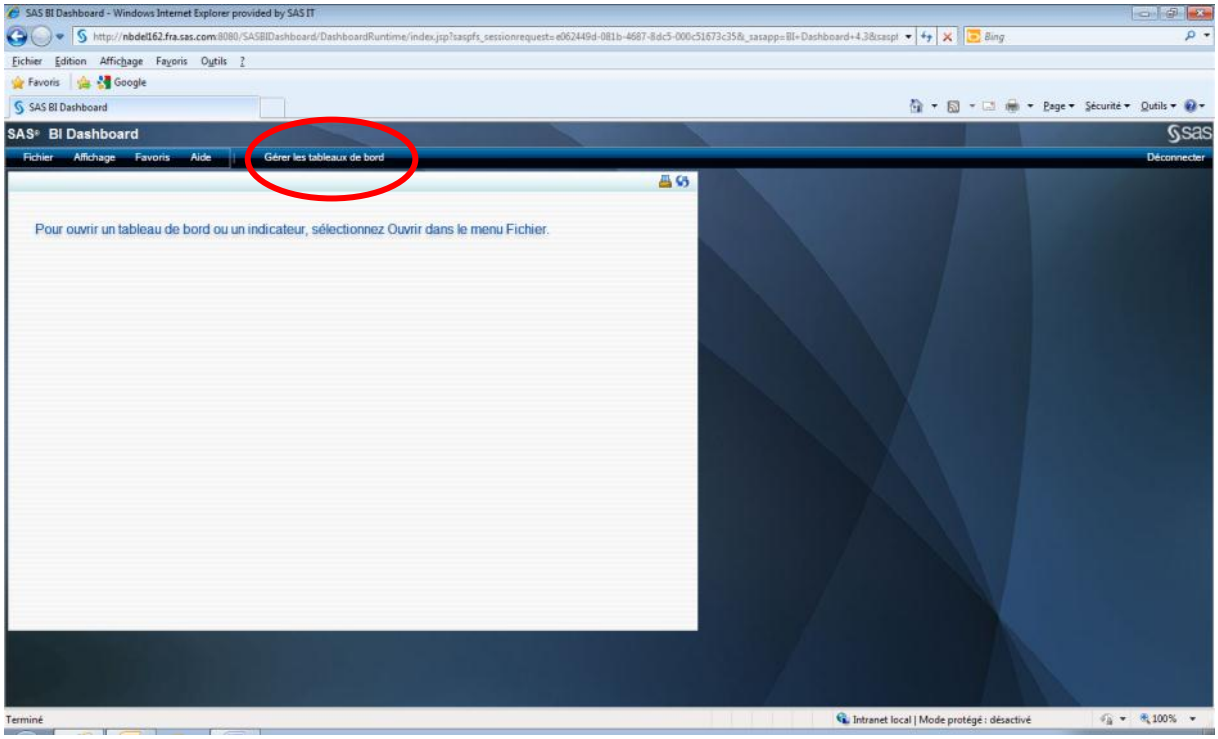


Créer un tableau de Bord

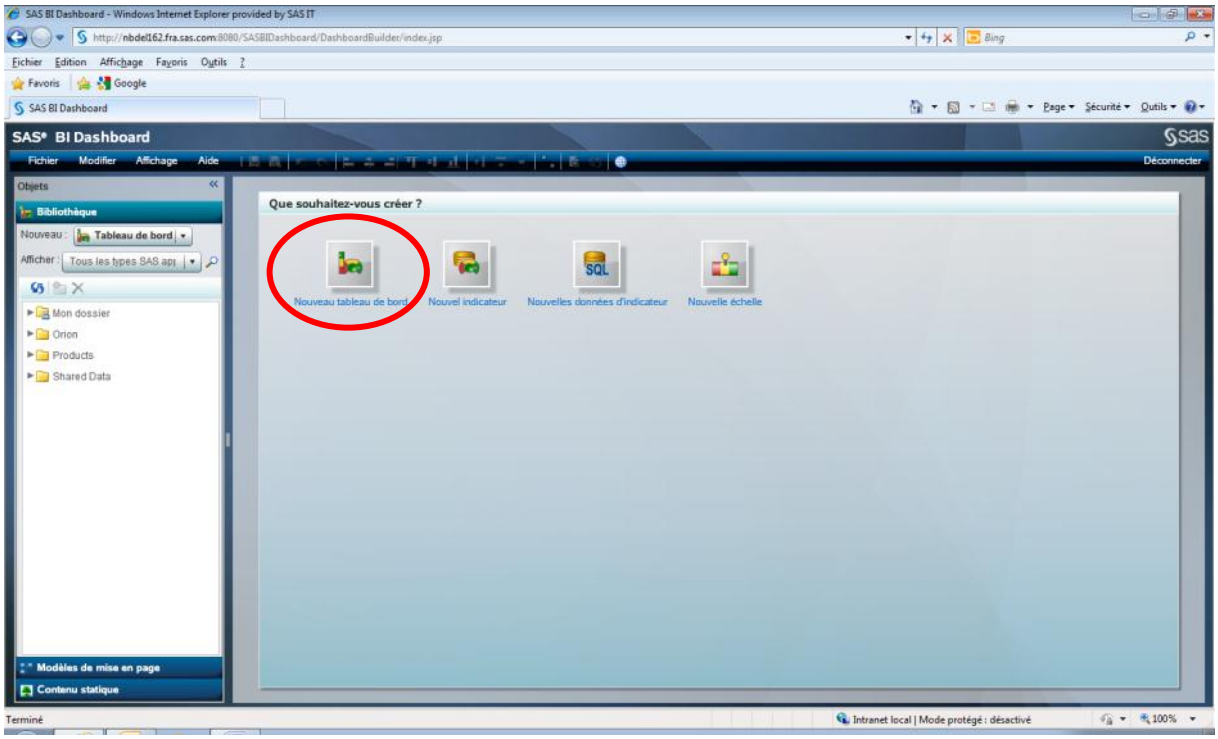
Avec BI Dashboard 4.3

Connectez-vous à la page suivante avec votre identifiant et votre mot de passe :
http://nom_du_serveur:8080/SASLogon/index.jsp?_sasapp=BI+Dashboard+4.3&

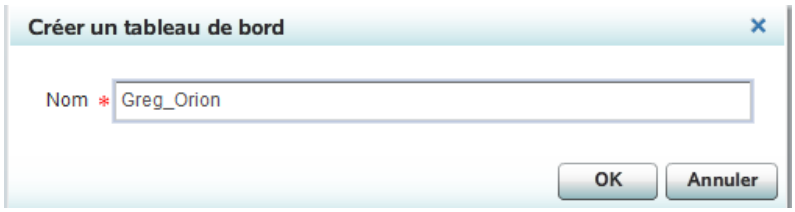




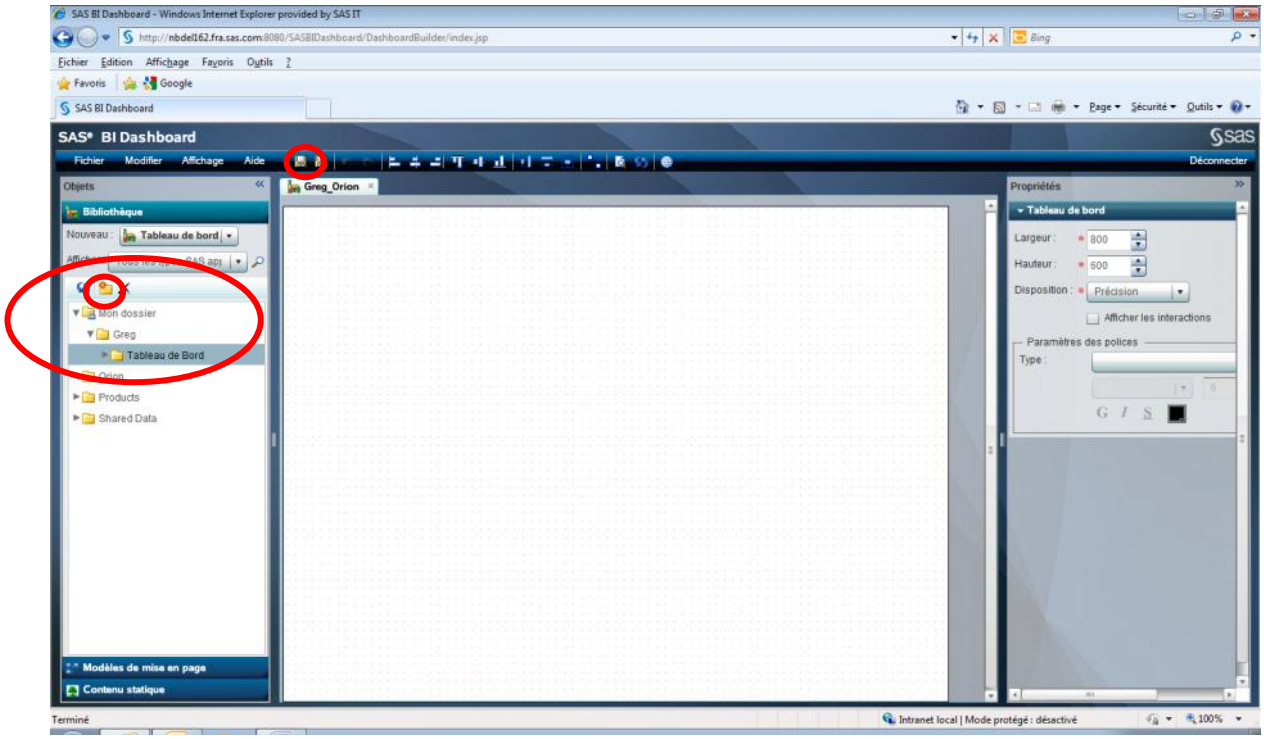
Cliquez sur Gérer les tableaux de bord.



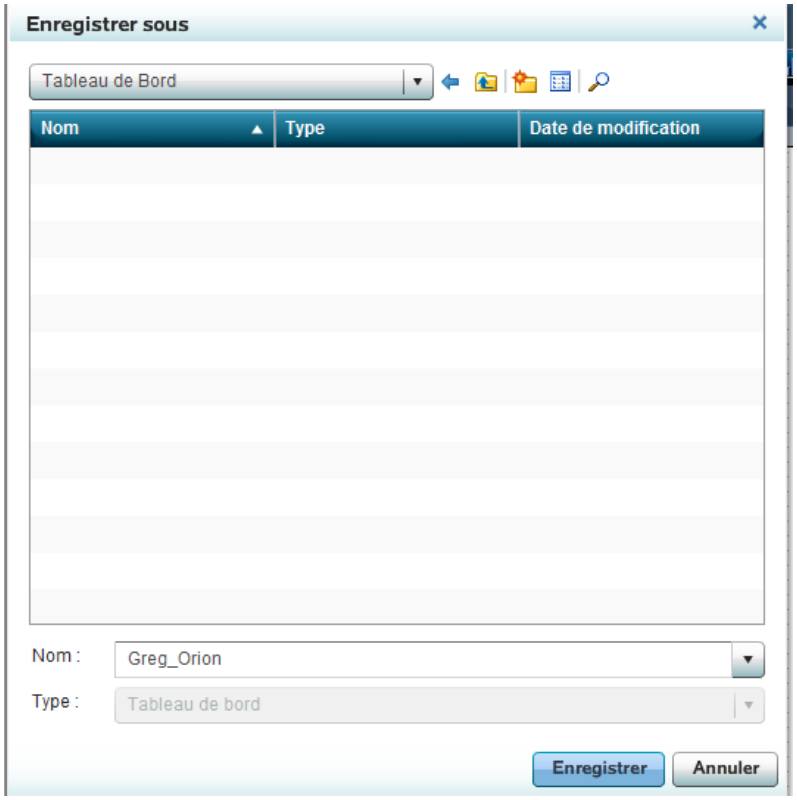
On peut commencer par créer un nouveau tableau de Bord



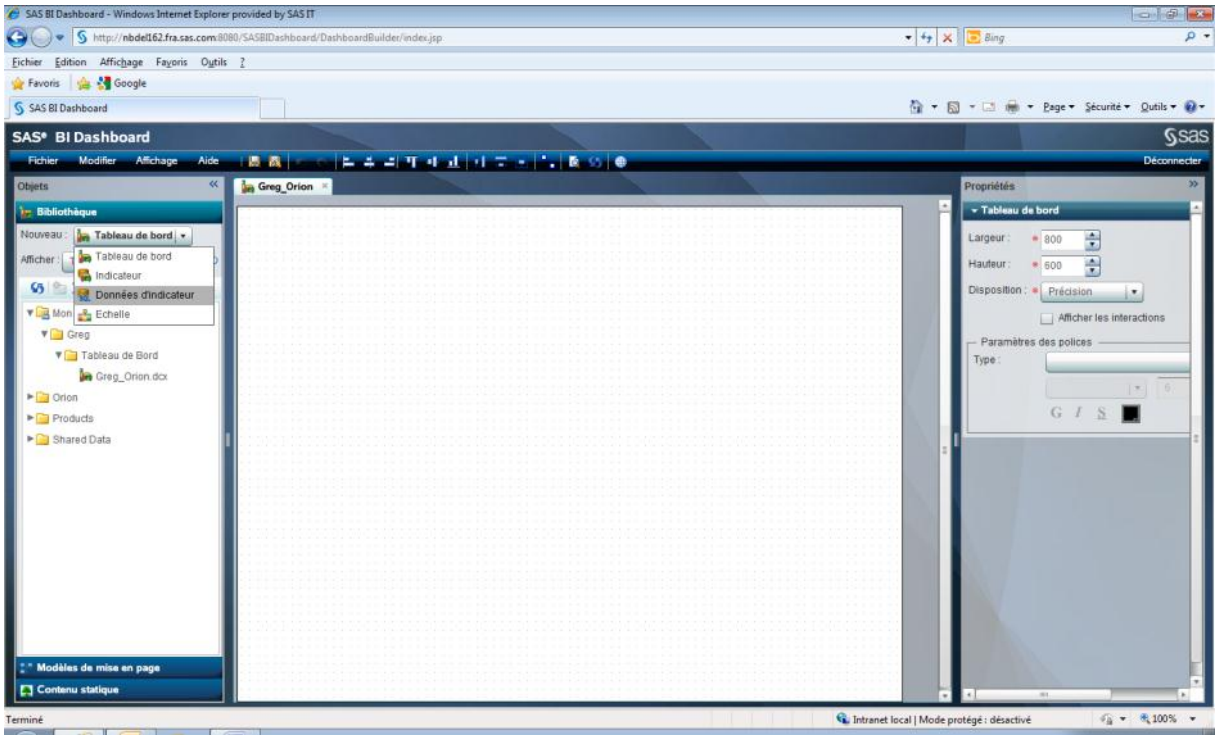
Donnez-lui un nom
OK



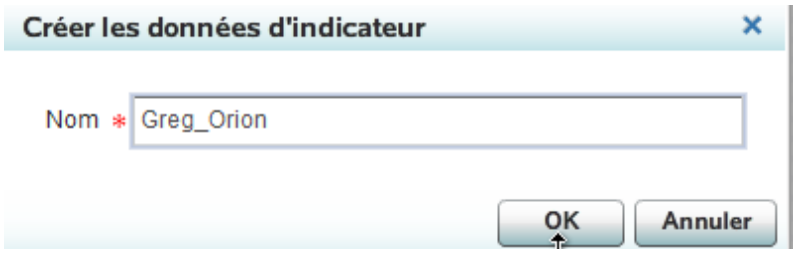
Vous pouvez vous créer un nouveau dossier pour enregistrer les éléments de votre tableau de bord.



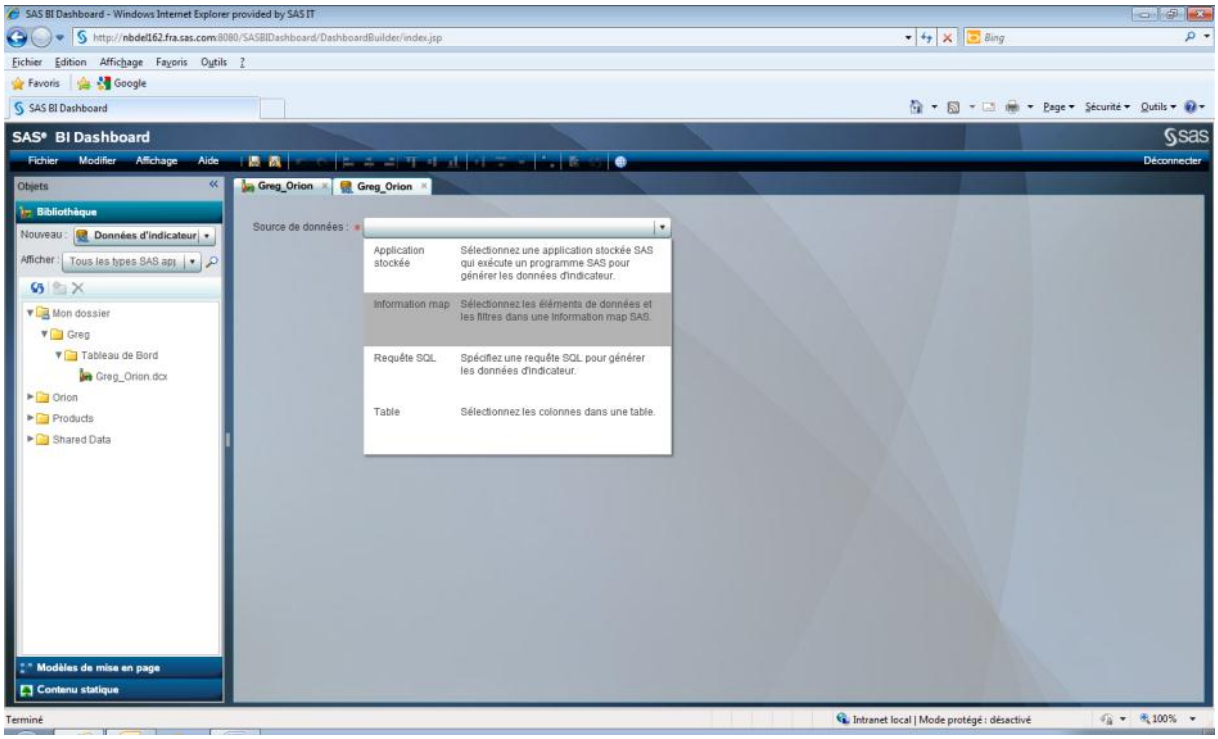
Enregistrer votre tableau de bord
Enregistrer



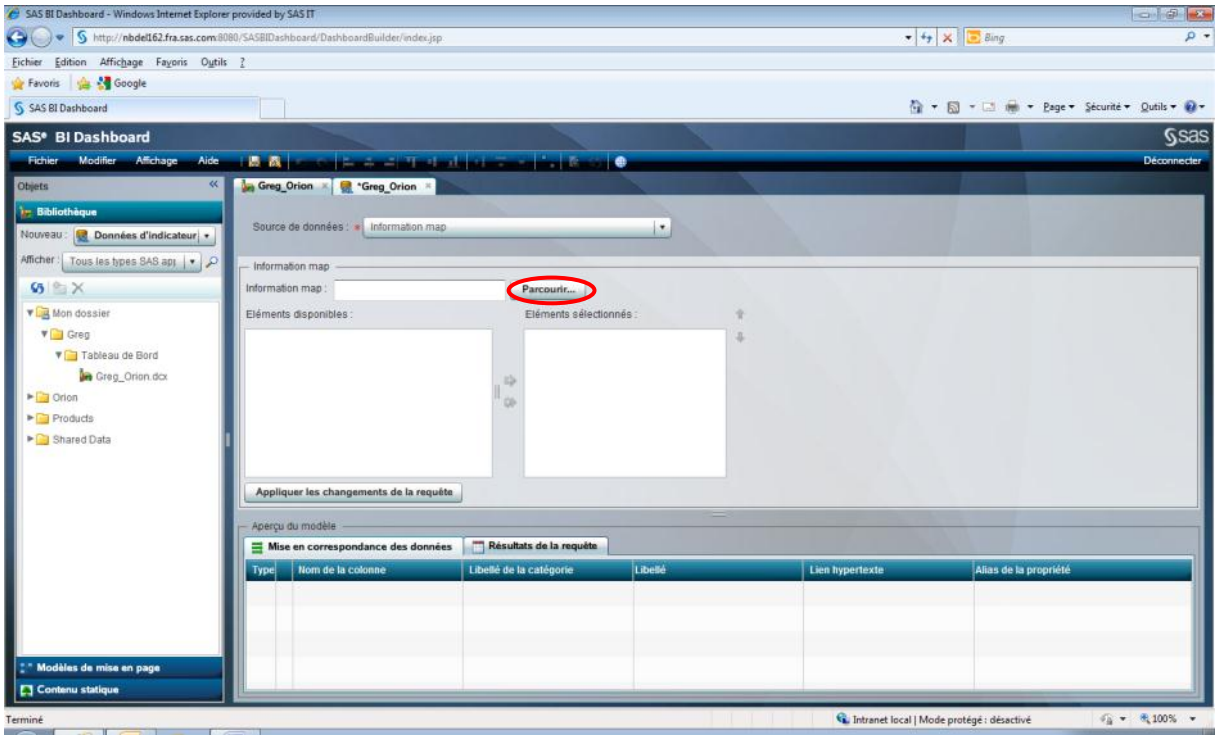
Définissez de nouvelles données pour votre tableau de bord



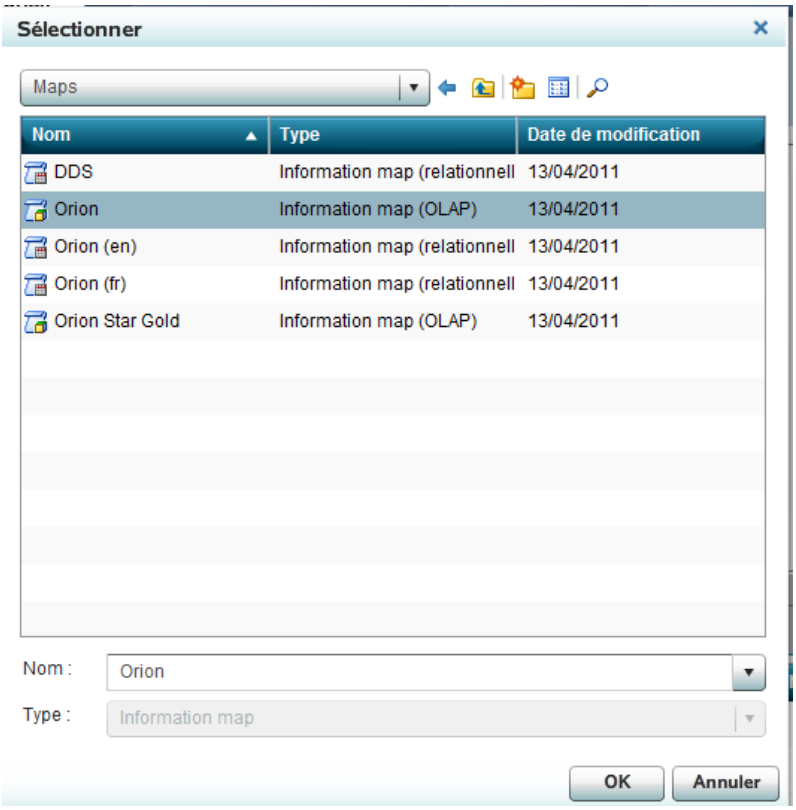
Enregistrer vos données
OK



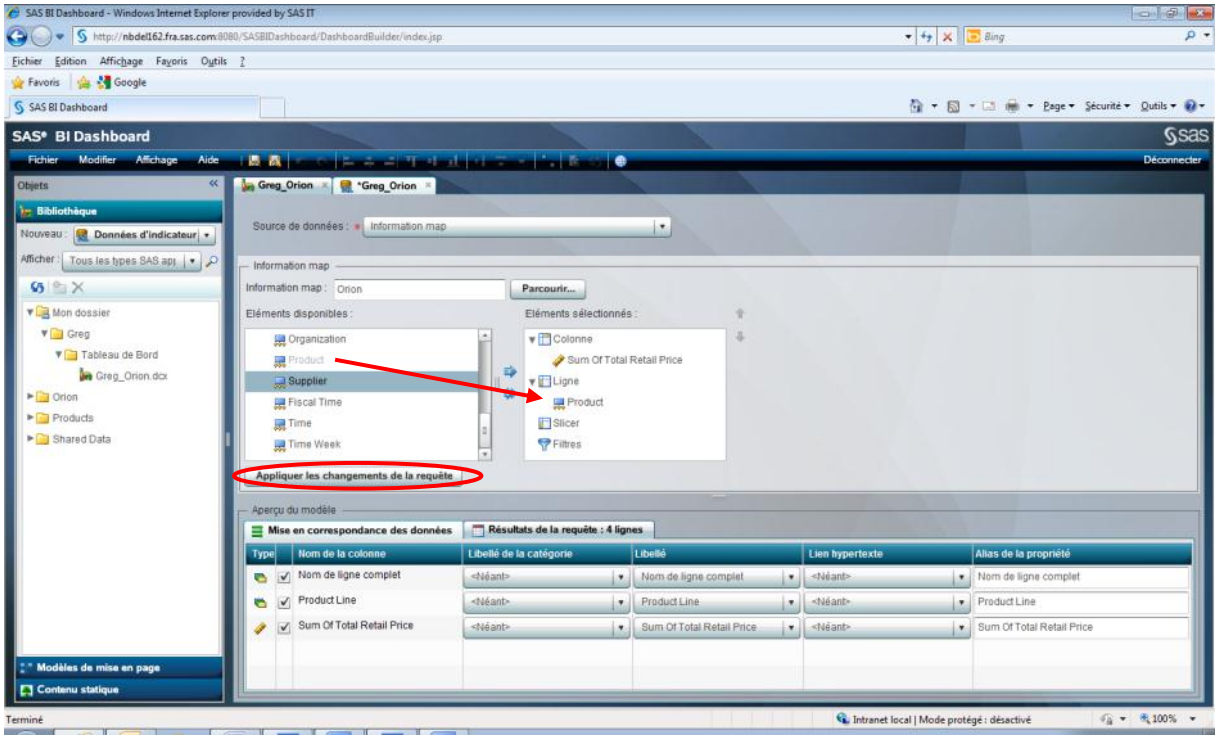
Sélectionner une source de données, par exemple une Information Map.



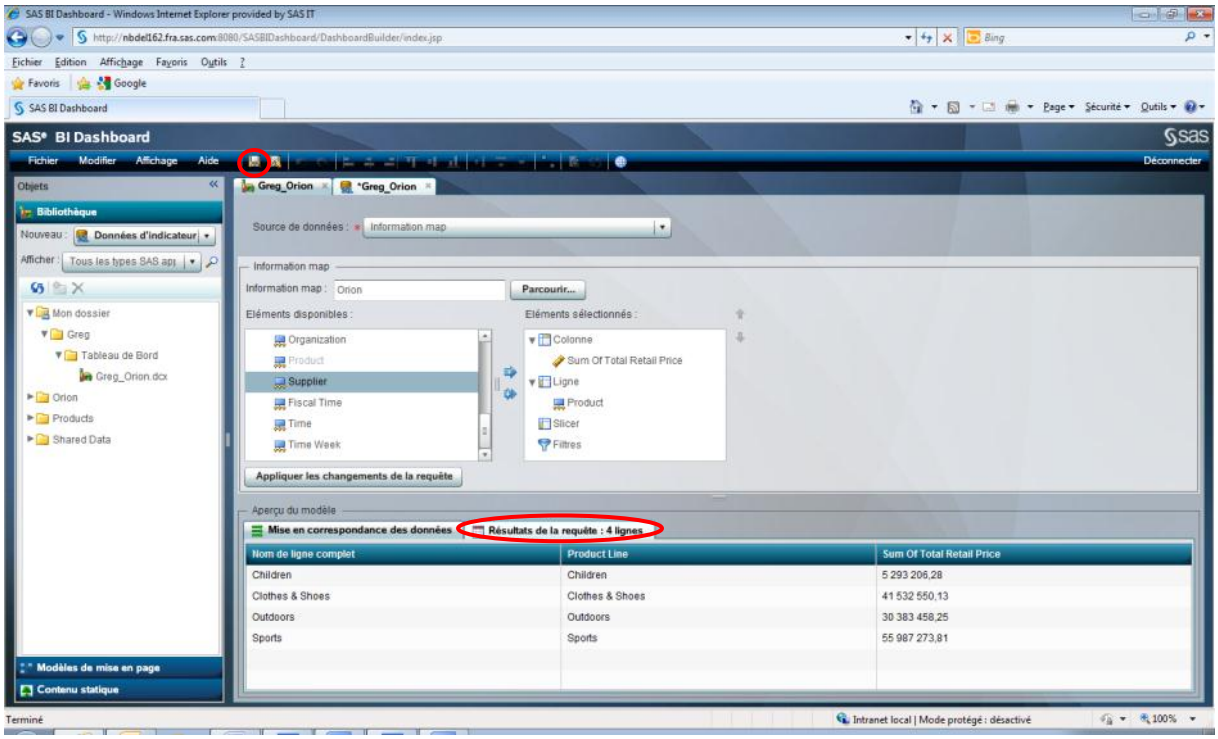
Allez chercher votre information Map



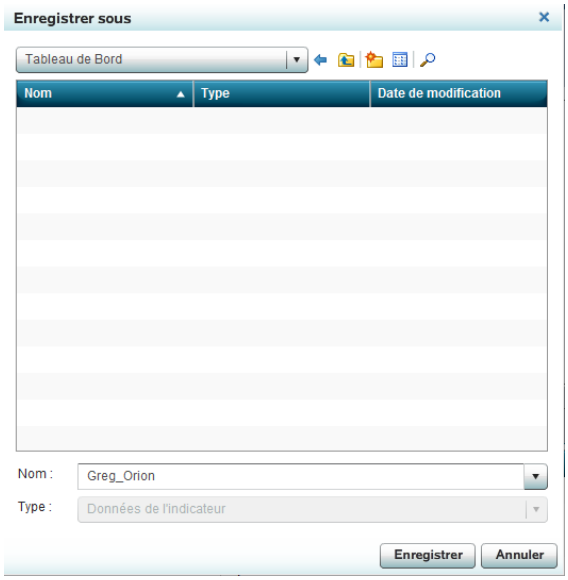
Sélectionner votre information Map (votre cube par exemple).



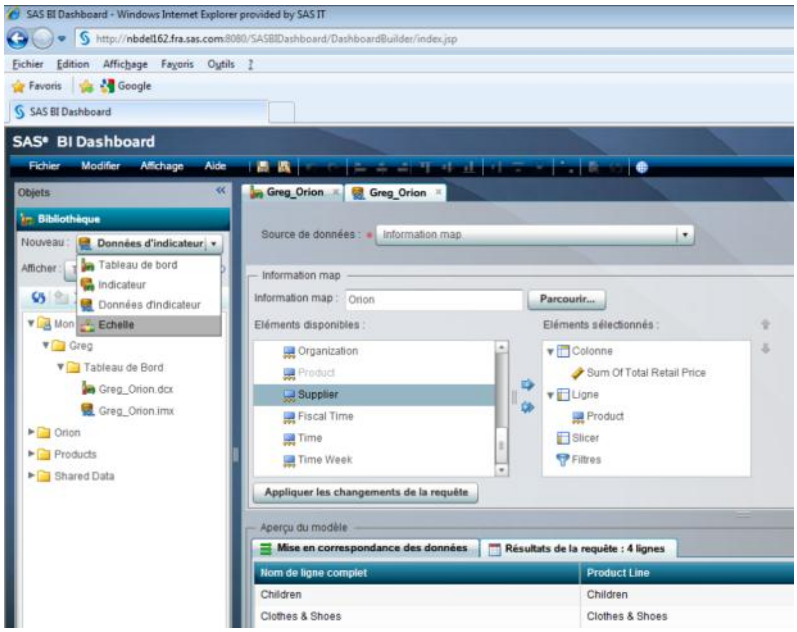
Sélectionner par exemple le chiffre d'affaire en colonne et la hiérarchie produit en ligne.
Appliquer les changements de la requête



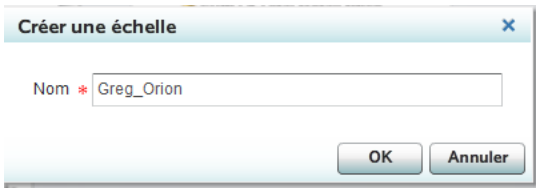
Vous pouvez regarder les résultats de la requête.
Enregistrez votre sélection de données.



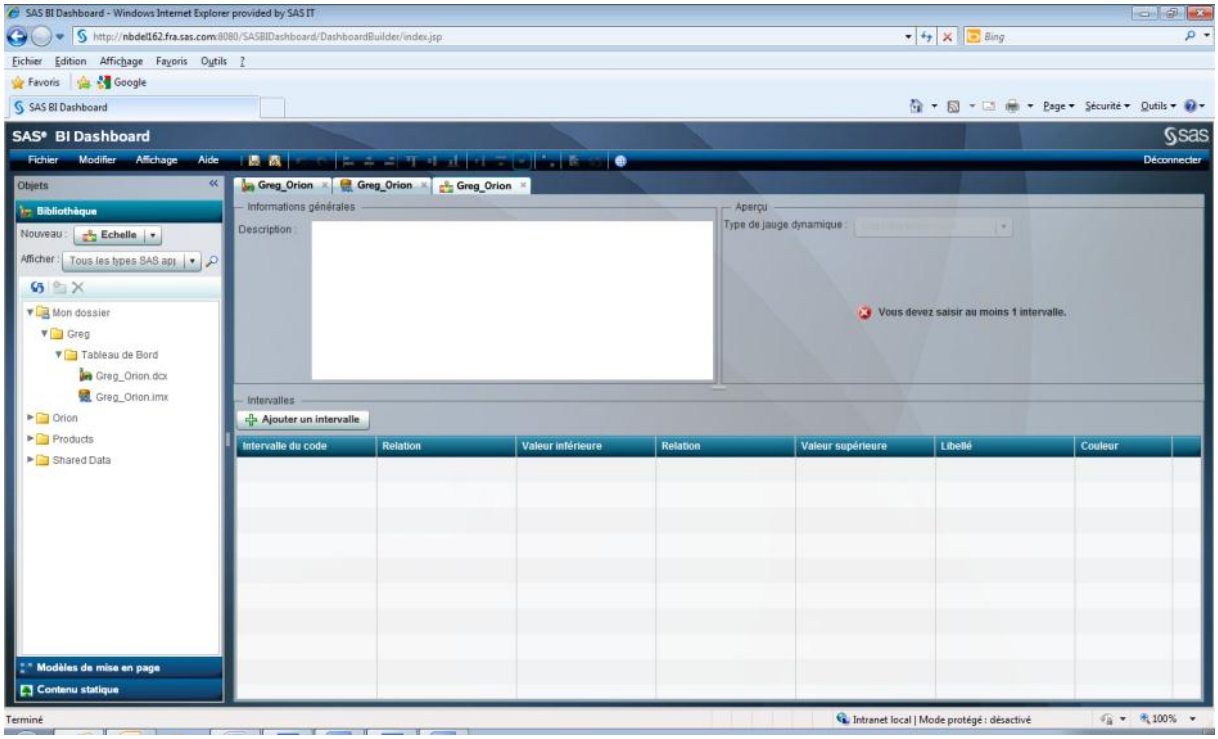
Donner un nom à votre sélection de données.
Enregistrer.



Créez une nouvelle échelle

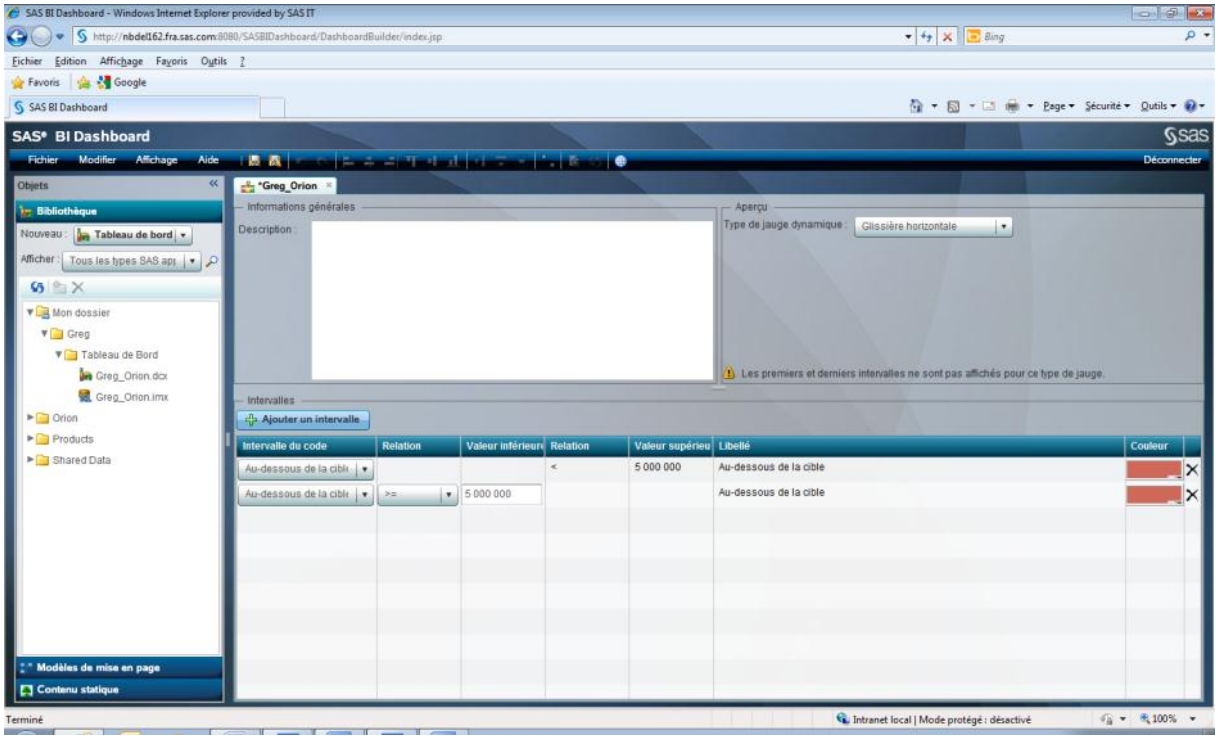


Lui donner un nom
OK

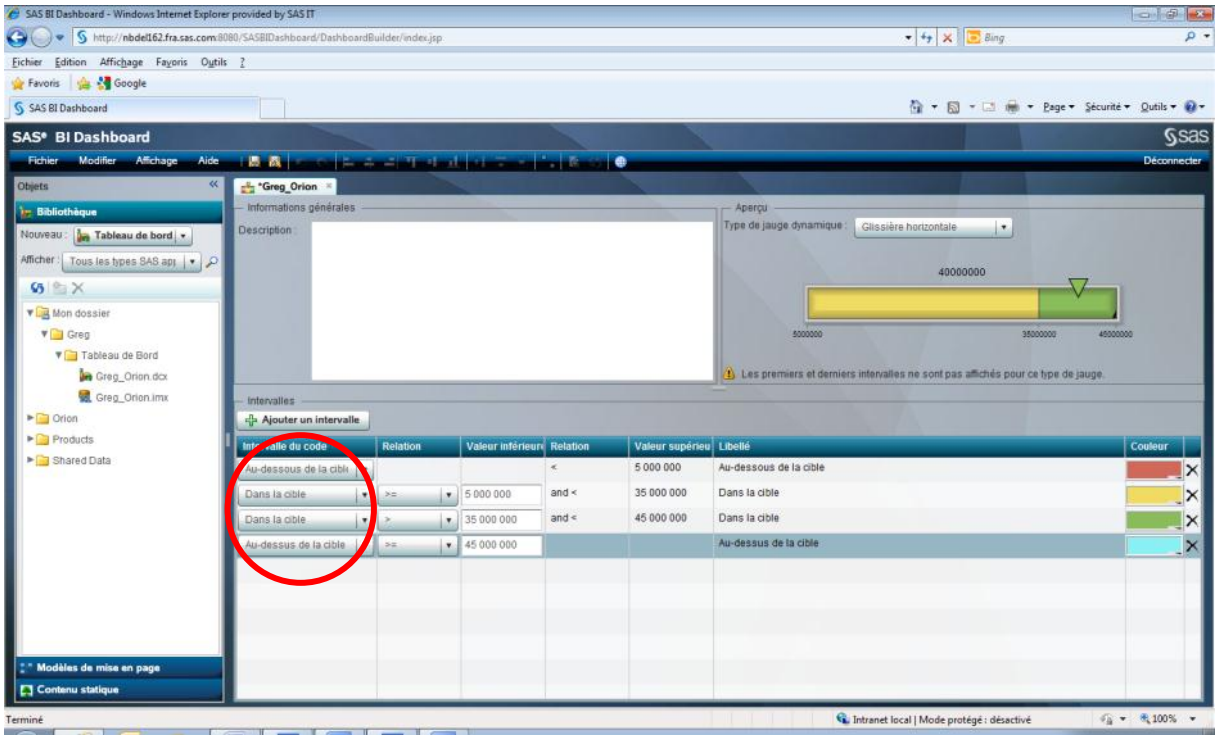


Ajouter un intervalle

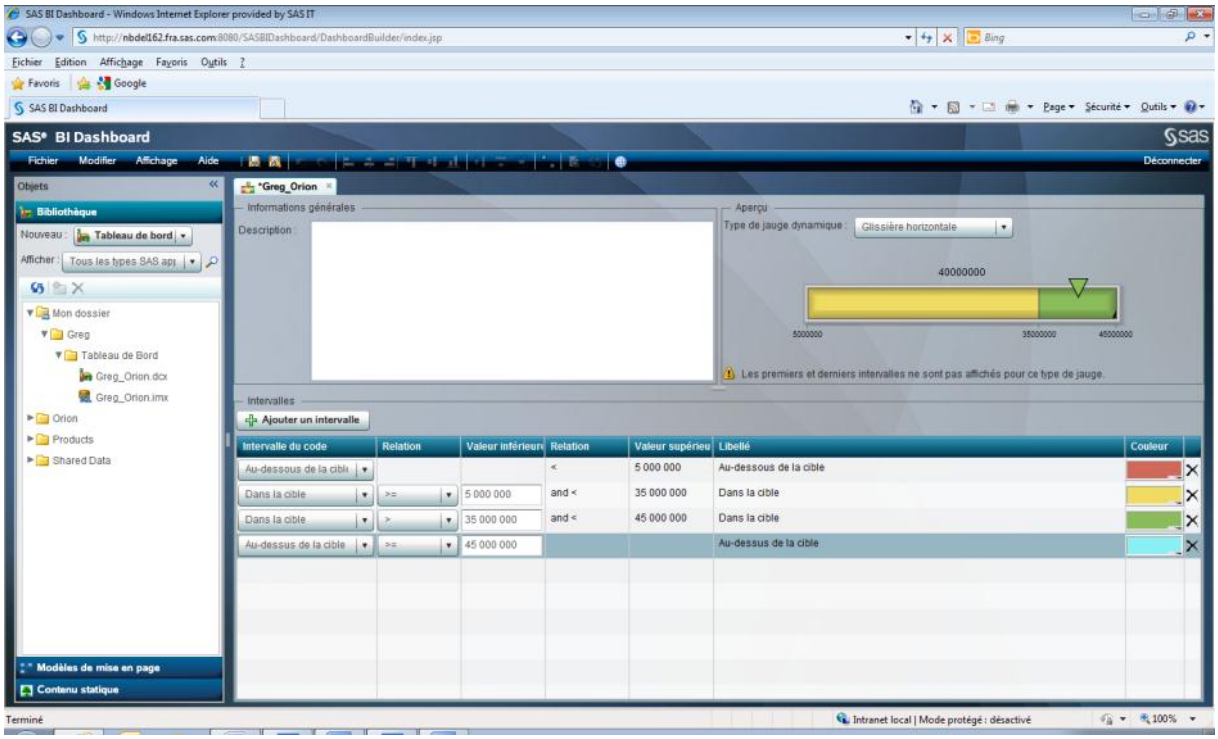
5 000 000 par exemple
OK



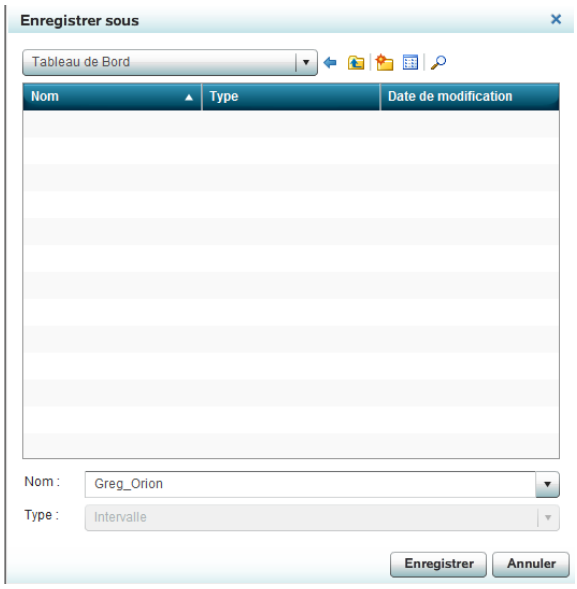
Ajouter d'autres intervalles : par exemple 35 000 000 et 45 000 000.



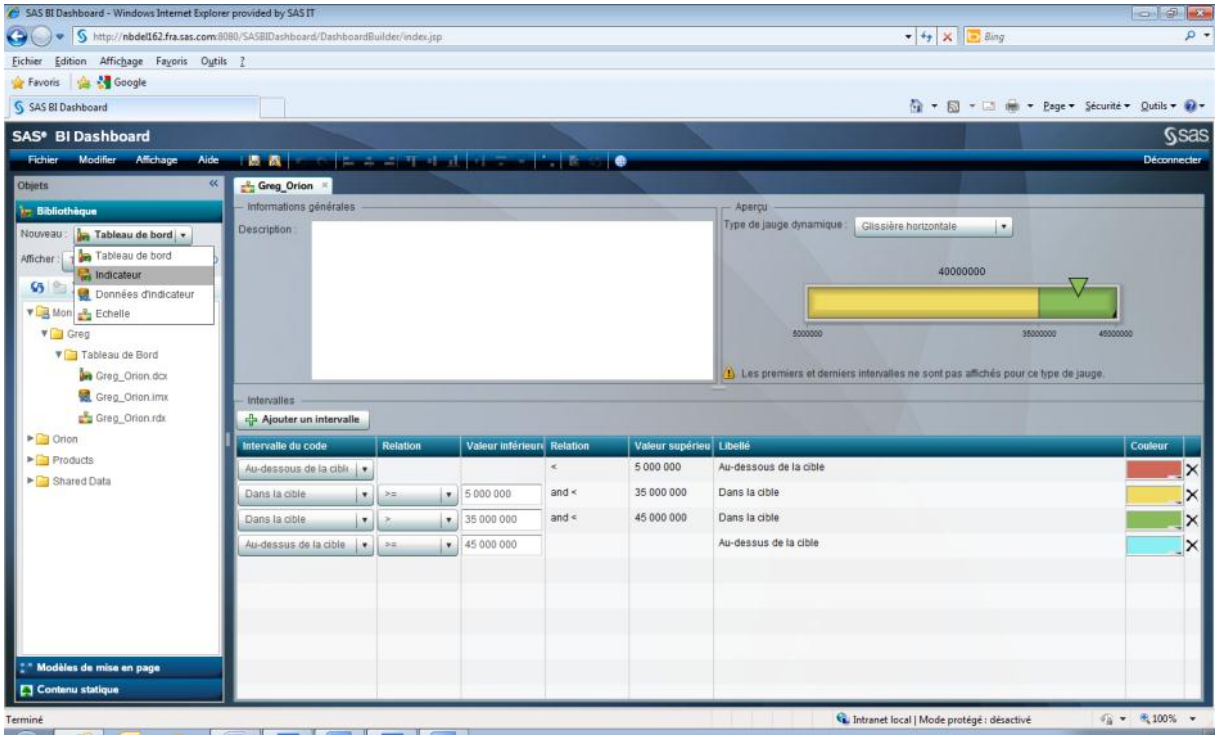
Modifier les intervalles du code



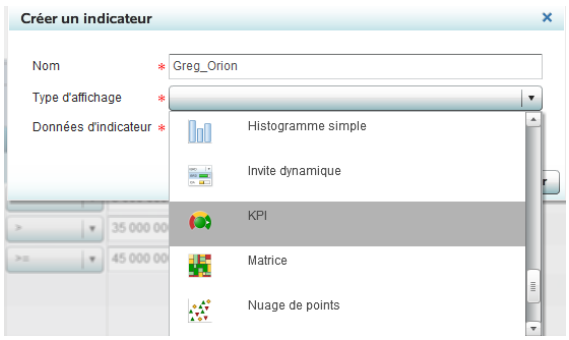
Modifiez les couleurs si besoin
Enregistrez votre échelle

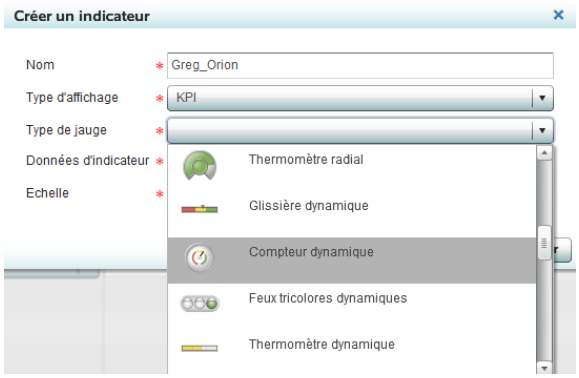


Enregistrer

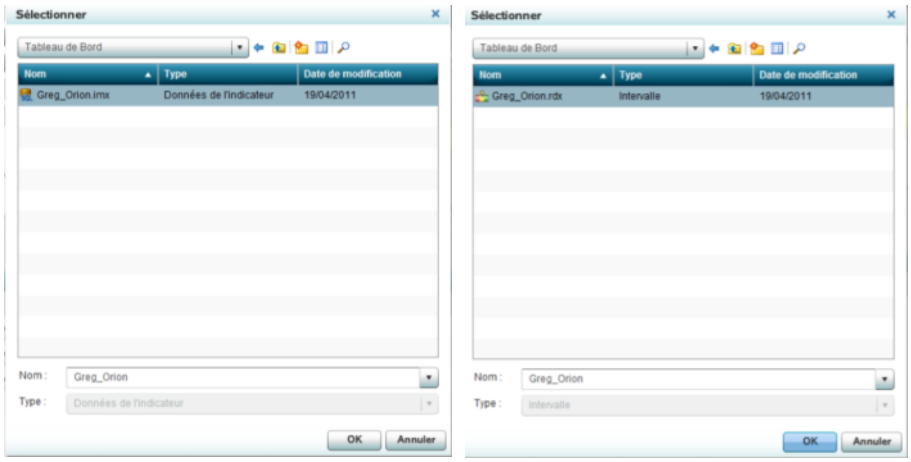


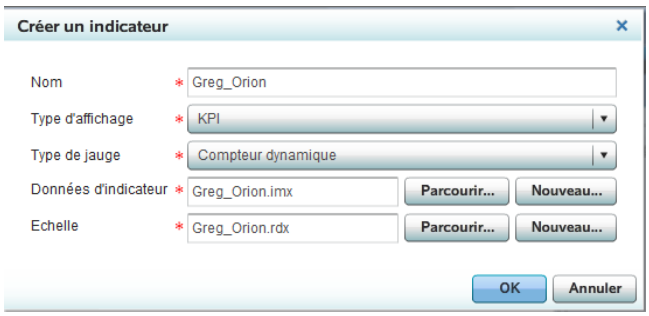
Créer un nouvel indicateur



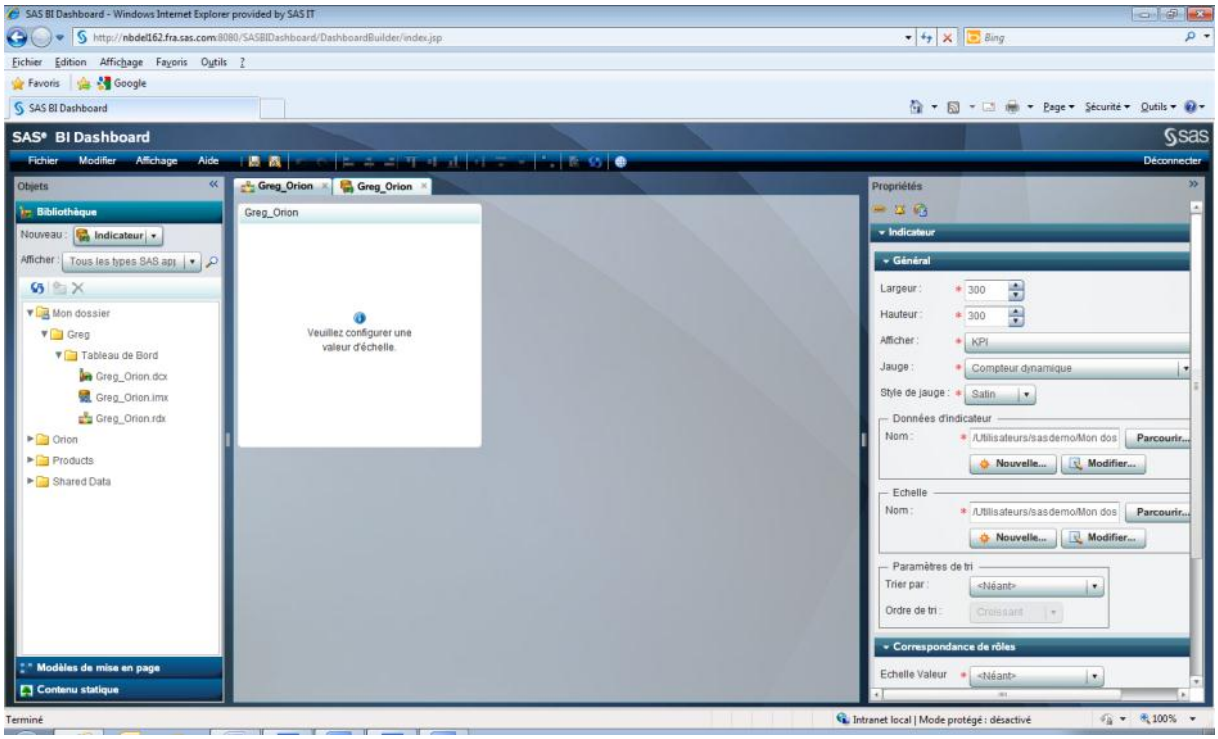


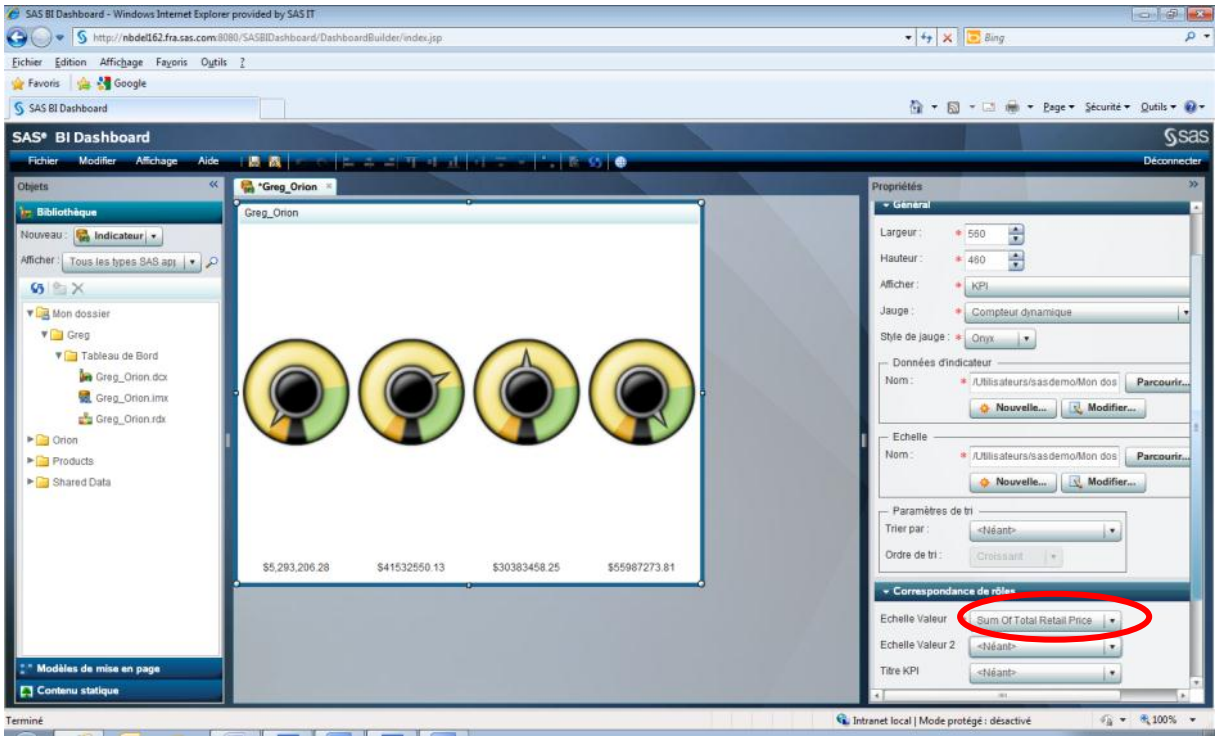
Lui donner un nom,
Sélectionner le type d'affichage
Pour un KPI, sélectionner le type de jauge
Sélectionner vos données et votre échelle



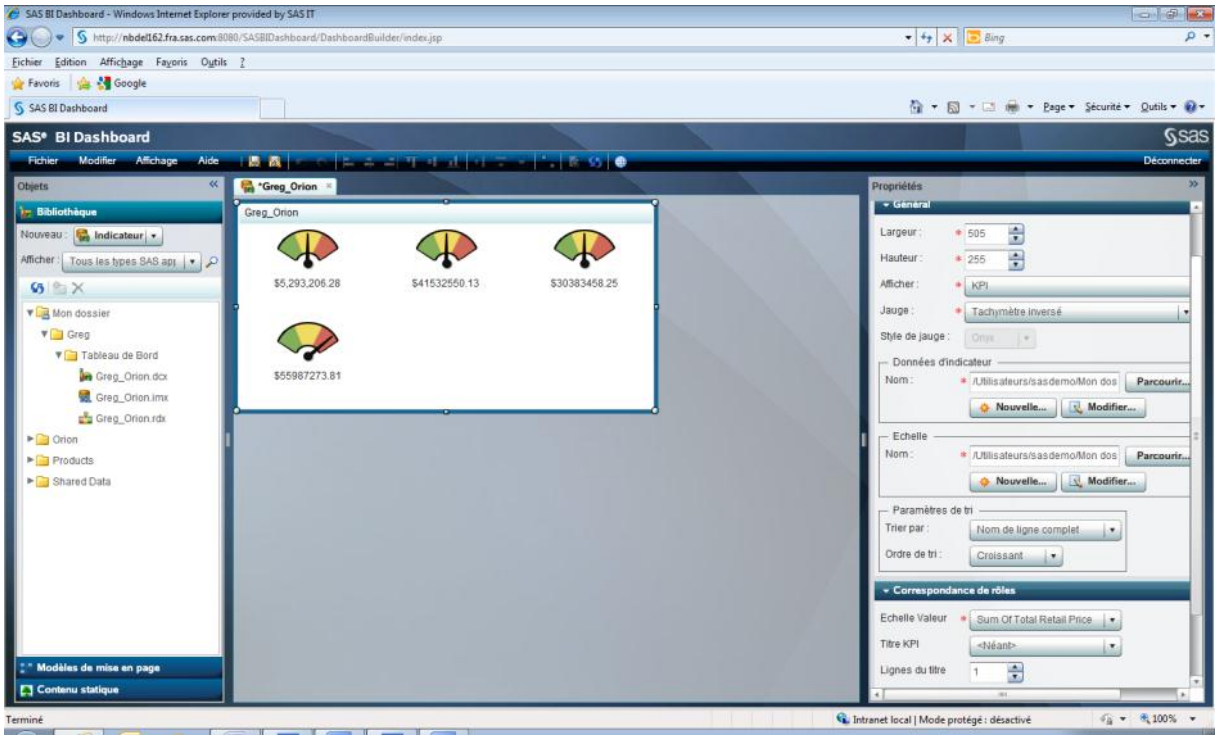


OK

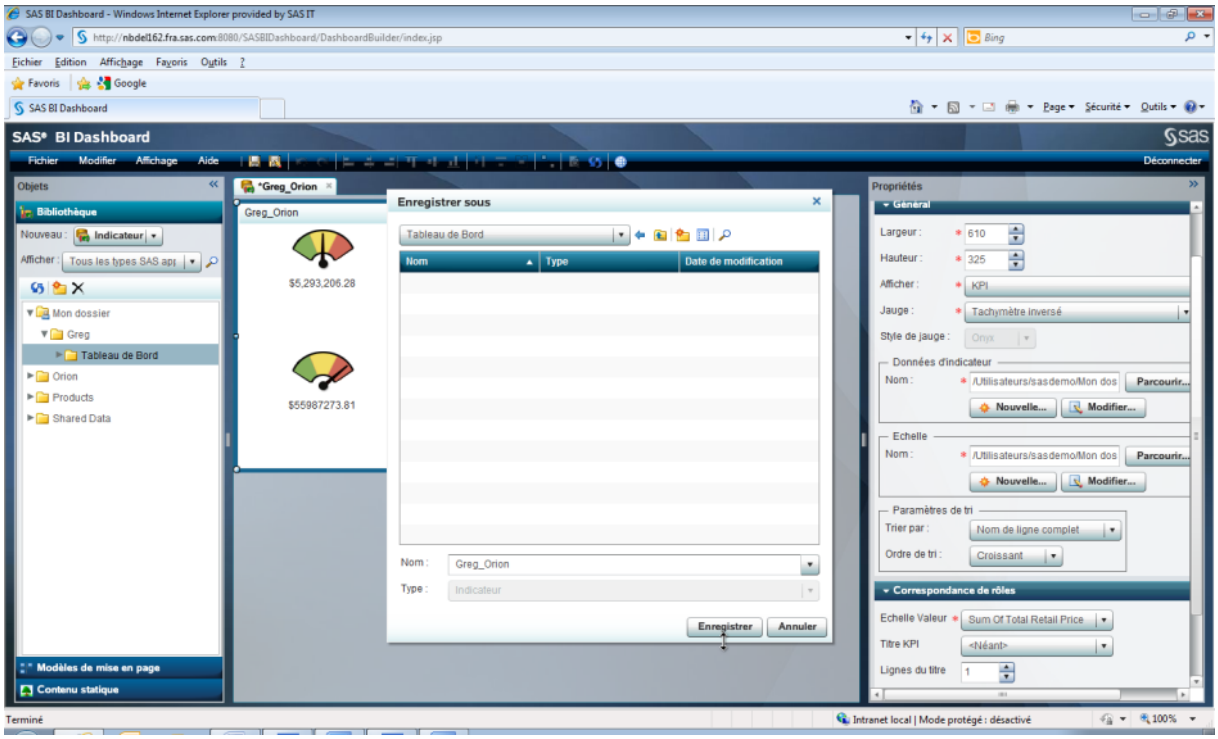


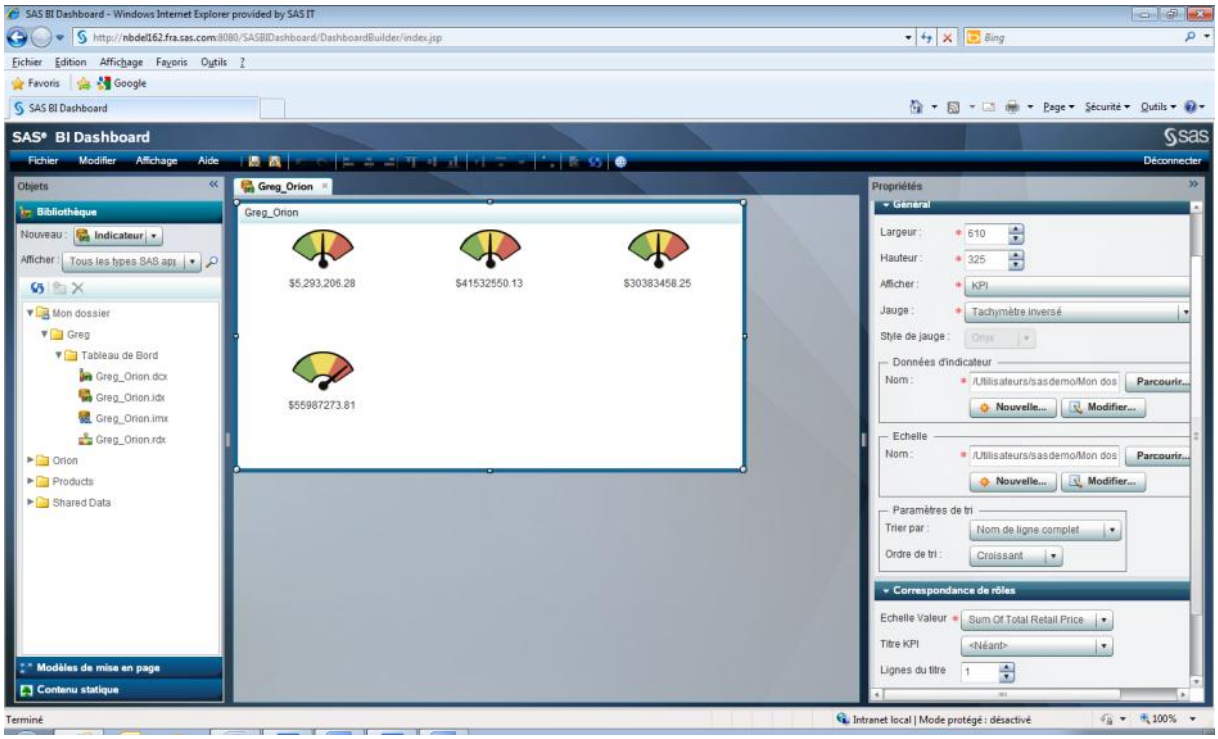


Sélectionner votre échelle de valeur sur la somme du chiffre d'affaire.

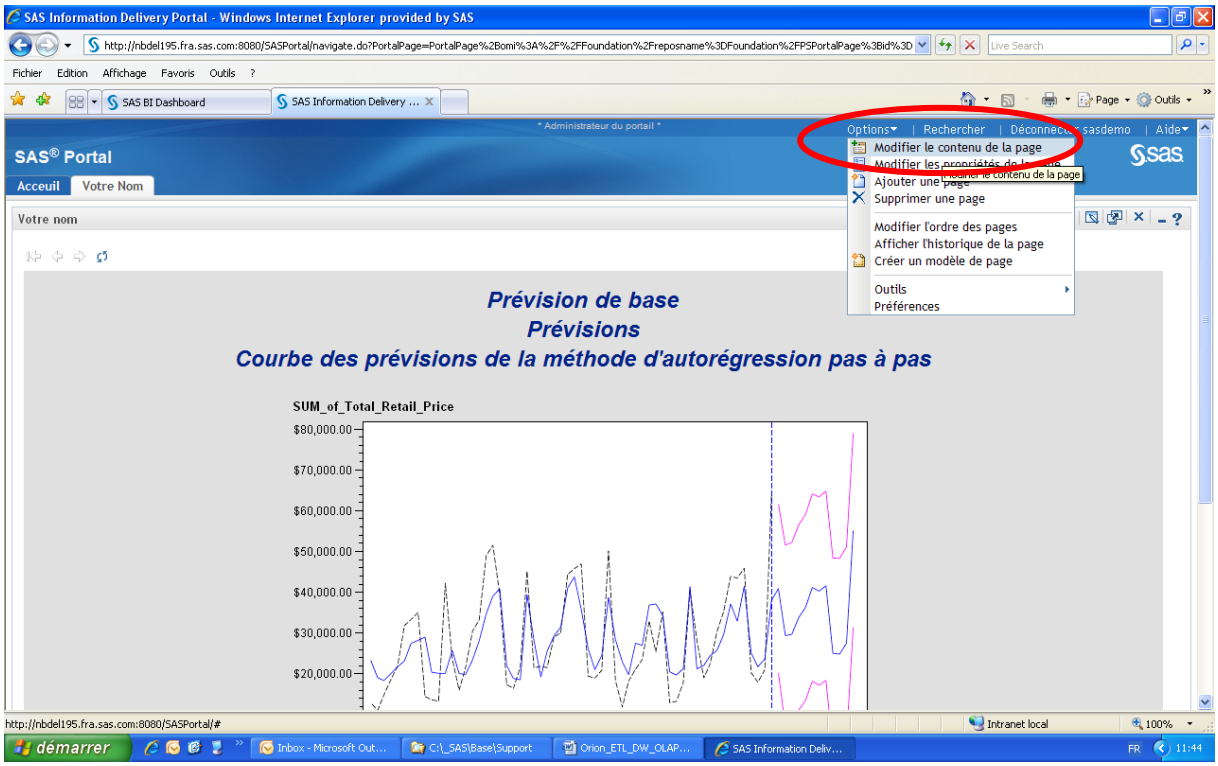


Enregistrer votre tableau de bord.



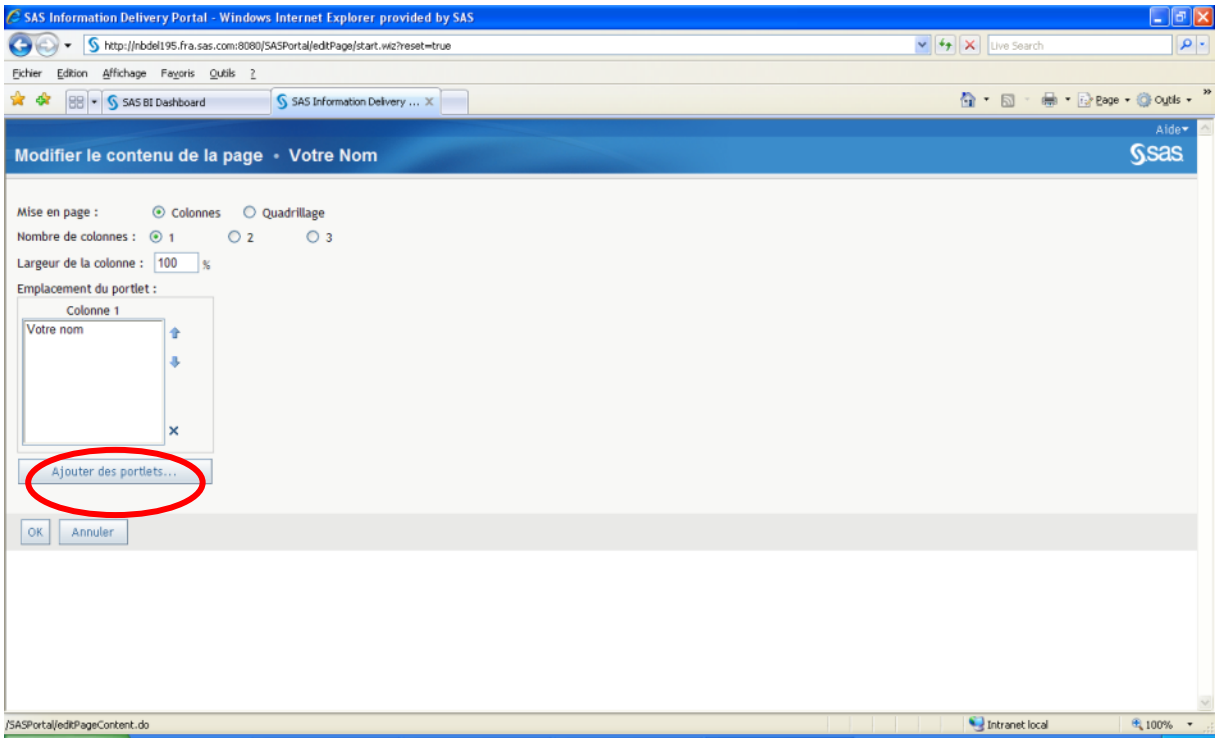

SAS BI Dashboard - Windows Internet Explorer provided by SAS IT
SAS® BI Dashboard
Fichier Modifier Affichage Aide
Déconnecter
Objets
Bibliothèque
Nouveau : Indicateur
Afficher : Tous les types SAS apr
Mon dossier
Greg
Tableau de Bord
Greg_Orion.dcx
Greg_Orion.idx
Greg_Orion.imx
Greg_Orion.rdx
Orion
Products
Shared Data
Modèles de mise en page
Contenu statique
Greg_Orion
$5,293,206.28
$41532550.13
$30383458.25
$55987273.81
Propriétés
Général
Largeur : 610
Hauteur : 325
Afficher : KPI
Jauge : Tachymètre inversé
Style de jauge : Onyx
Données d'indicateur
Nom : /Utilisateurs/sasdemo/Mon dos
Parcourir...
Nouvelle...
Modifier...
Echelle
Nom : /Utilisateurs/sasdemo/Mon dos
Parcourir...
Nouvelle...
Modifier...
Paramètres de tri
Trier par : Nom de ligne complet
Ordre de tri : Croissant
Correspondance de rôles
Echelle Valeur : Sum Of Total Retail Price
Titre KPI : <Néant>
Lignes du titre : 1
Terminé
Intranet local | Mode protégé : désactivé
100%

Afficher le tableau de bord dans le portail

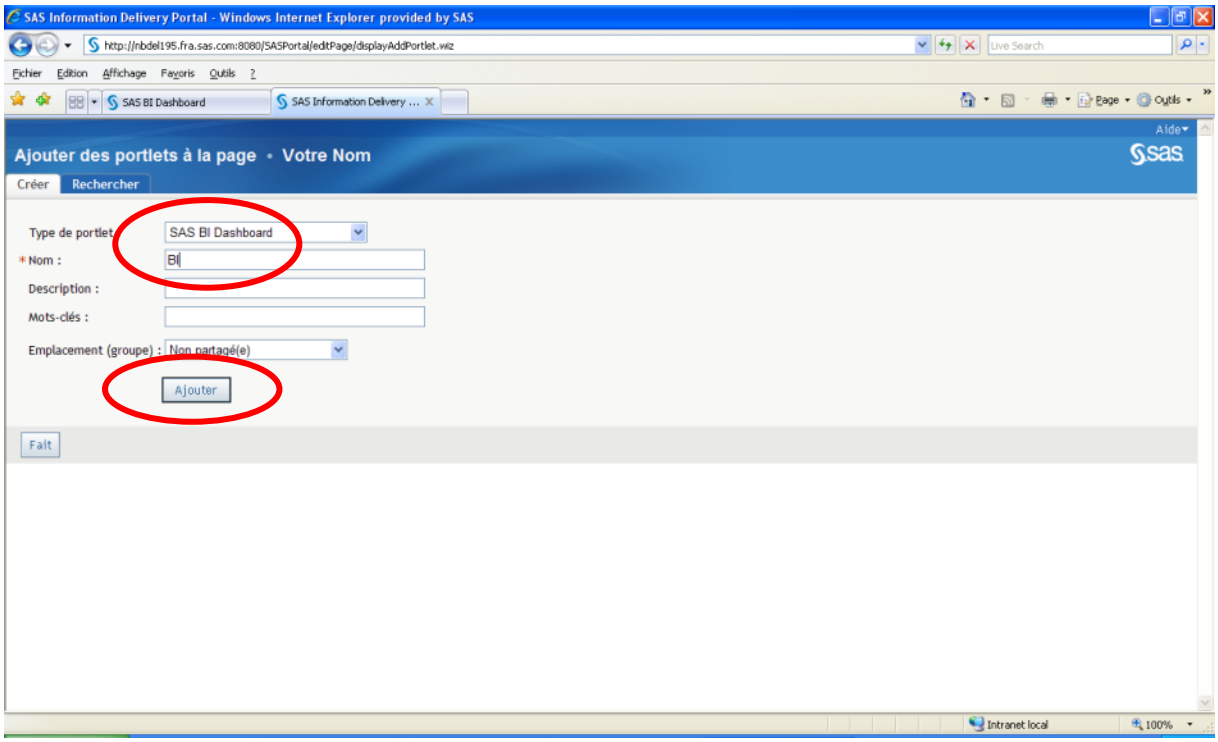


Ouvrir le Portail : http://nomdelamachine:8080/SASPortal et entrer le nom et le mot de passe.

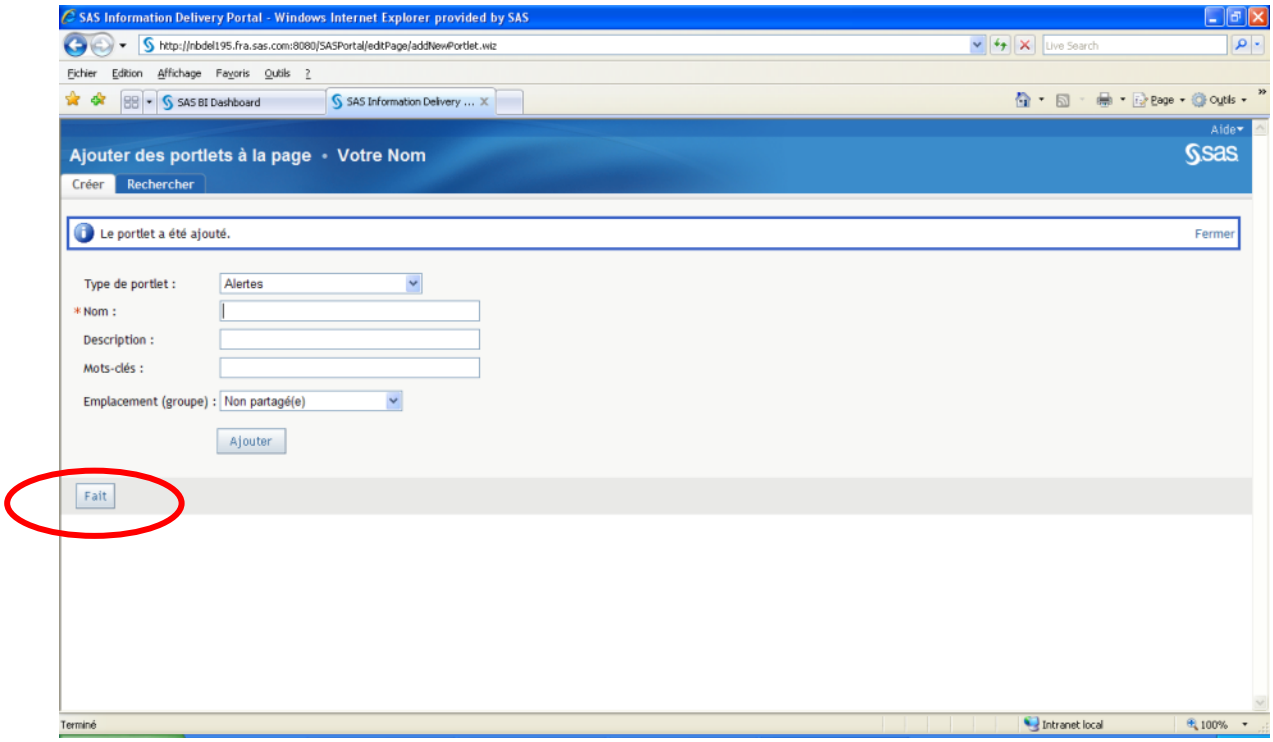
Modifier le contenu de la page



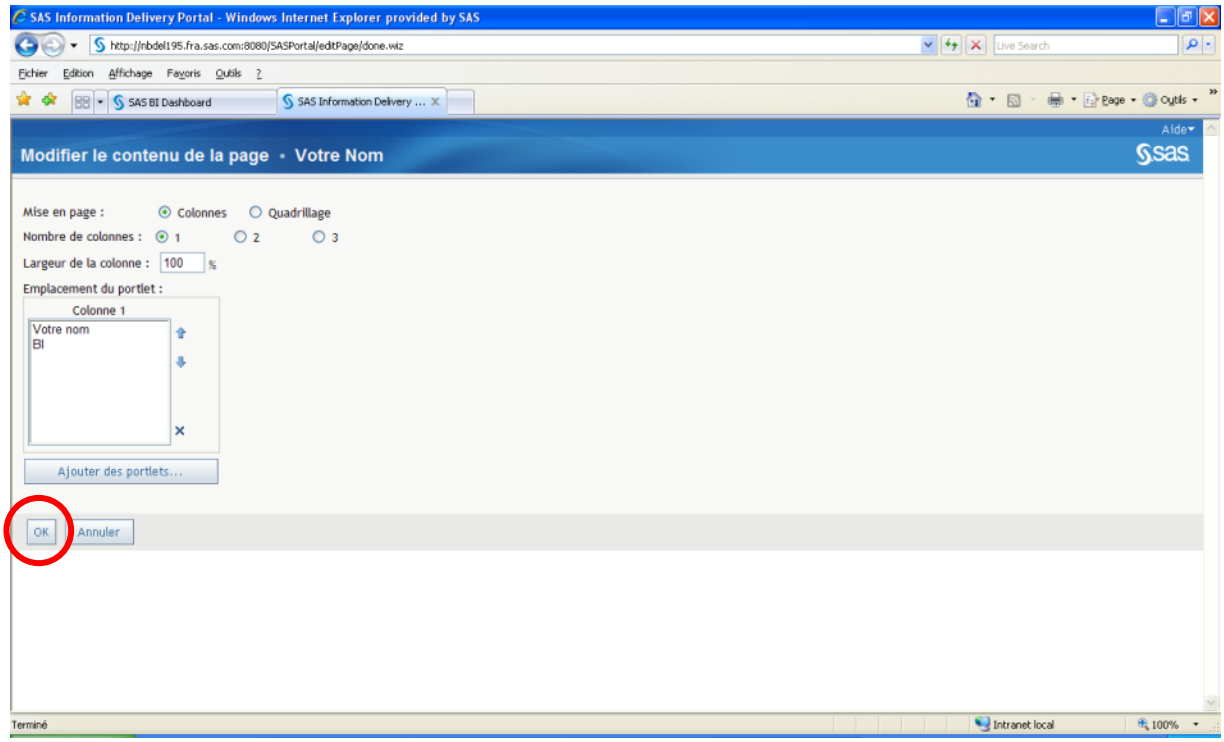
Ajouter un portlet



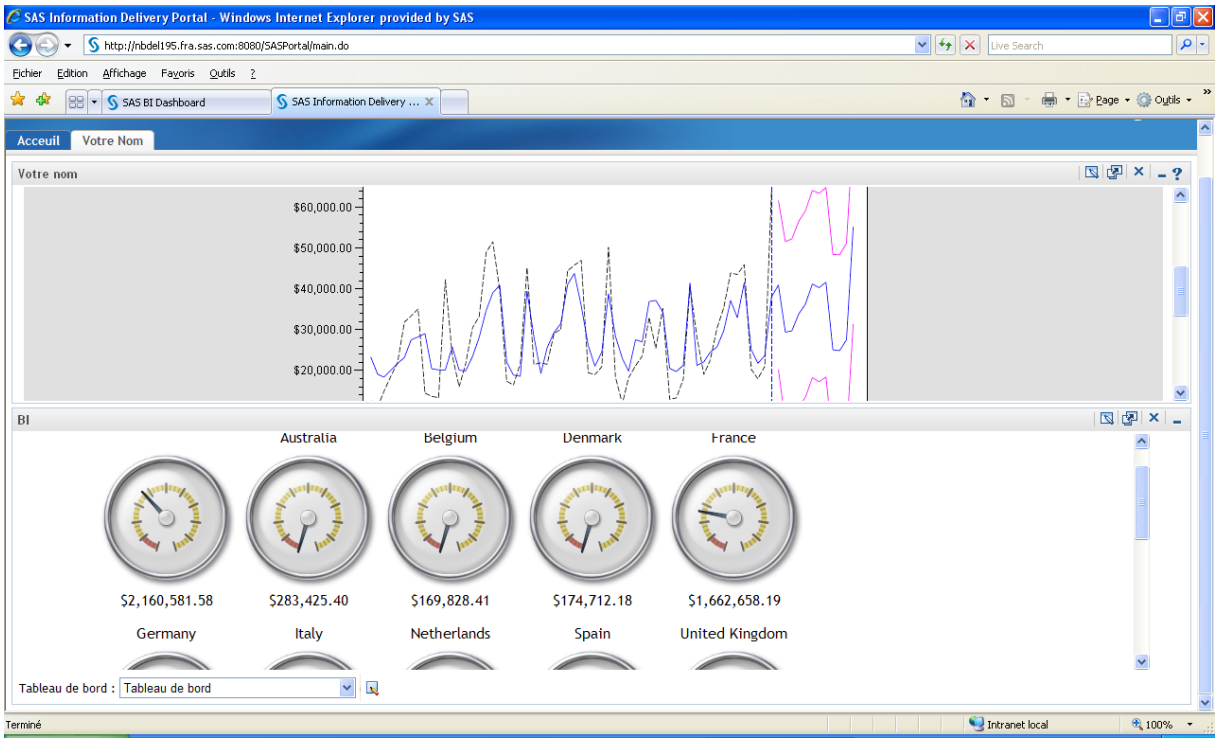
Ajouter un SAS BI Dashboard portlet et lui donner un nom.

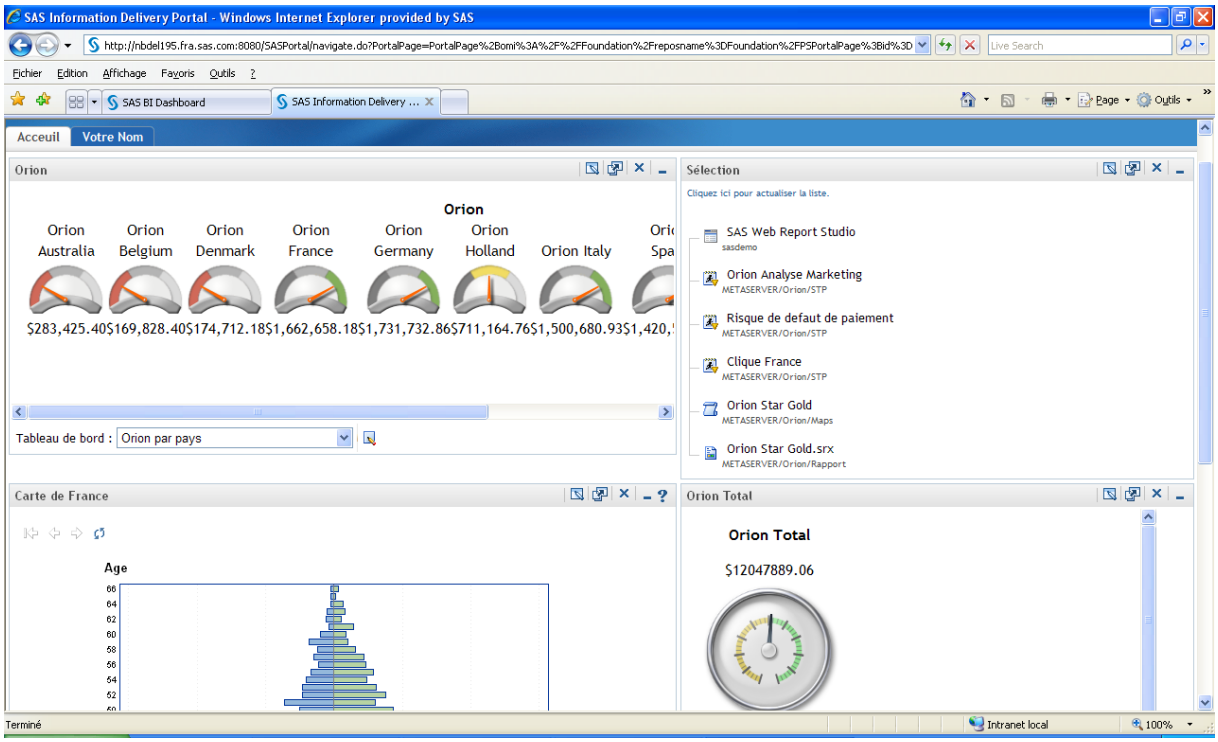

SAS Information Delivery Portal - Windows Internet Explorer provided by SAS
http://nbdel195.fra.sas.com:8080/SASPortal/editPage/addNewPortlet.wiz
Ajouter des portlets à la page • Votre Nom
Créer
Rechercher
Le portlet a été ajouté.
Fermer
Type de portlet :
Alertes
* Nom :
Description :
Mots-clés :
Emplacement (groupe) :
Non partagé(e)
Ajouter
Fait

Fait


SAS Information Delivery Portal - Windows Internet Explorer provided by SAS
http://nbdel195.fra.sas.com:8080/SASPortal/editPage/done.wiz
Modifier le contenu de la page • Votre Nom
Mise en page :
Colonnes
Quadrillage
Nombre de colonnes :
1
2
3
Largeur de la colonne :
100
%
Emplacement du portlet :
Colonne 1
Votre nom
BI
Ajouter des portlets...
OK
Annuler

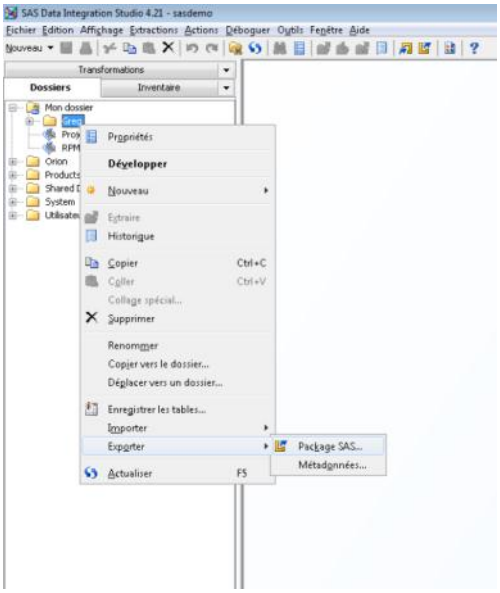
OK

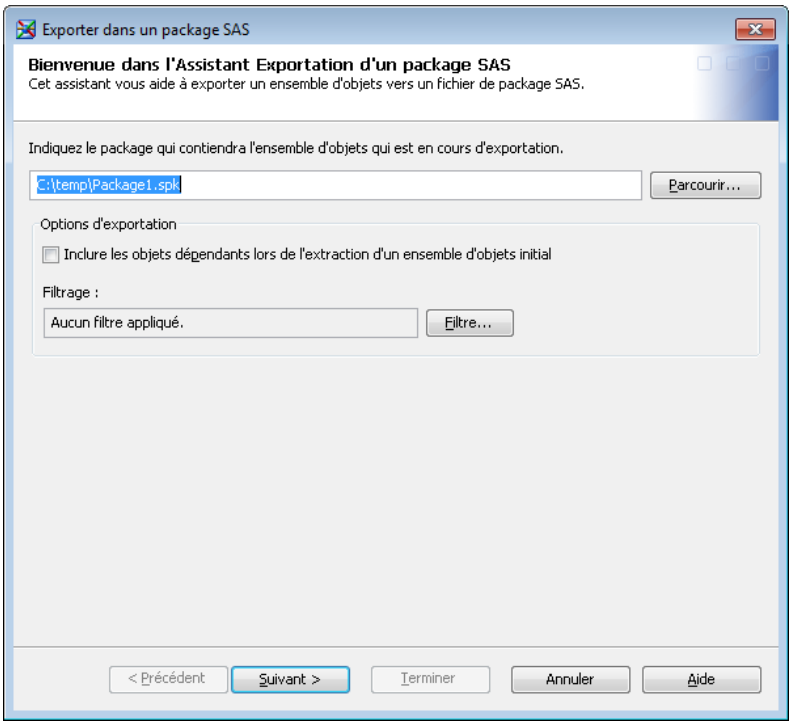

SAS Information Delivery Portal - Windows Internet Explorer provided by SAS
http://nbdel195.fra.sas.com:8080/SASPortal/main.do
Live Search
Fichier Edition Affichage Favoris Outils ?
SAS BI Dashboard
SAS Information Delivery ...
Page Outils
Acceuil
Votre Nom
Votre nom
$60,000.00
$50,000.00
$40,000.00
$30,000.00
$20,000.00
BI
Australia
Belgium
Denmark
France
$2,160,581.58
$283,425.40
$169,828.41
$174,712.18
$1,662,658.19
Germany
Italy
Netherlands
Spain
United Kingdom
Tableau de bord :
Tableau de bord
Terminé
Intranet local
100%


SAS Information Delivery Portal - Windows Internet Explorer provided by SAS
Acceuil
Votre Nom
Orion
Orion
Orion Australia
Orion Belgium
Orion Denmark
Orion France
Orion Germany
Orion Holland
Orion Italy
$283,425.40
$169,828.40
$174,712.18
$1,662,658.18
$1,731,732.86
$711,164.76
$1,500,680.93
Tableau de bord : Orion par pays
Sélection
Cliquez ici pour actualiser la liste.
SAS Web Report Studio
sasdemo
Orion Analyse Marketing
METASERVER/Orion/STP
Risque de defaut de paiement
METASERVER/Orion/STP
Clique France
METASERVER/Orion/STP
Orion Star Gold
METASERVER/Orion/Maps
Orion Star Gold.srx
METASERVER/Orion/Rapport
Carte de France
Age
Orion Total
Orion Total
$12047889.06
Terminé
Intranet local
100%

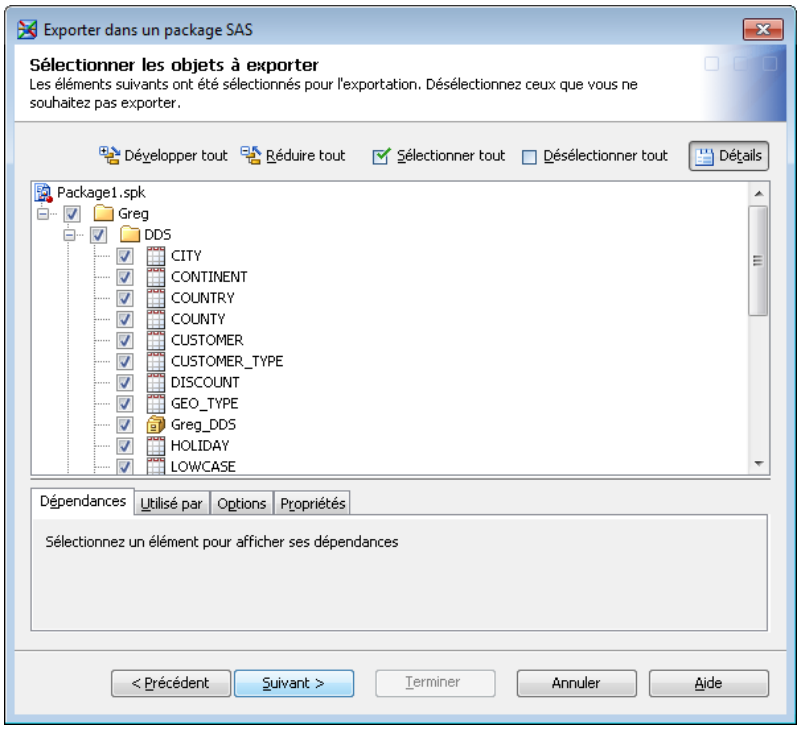
Export des métadonnées de votre travail

Pour exporter les métadonnées de ce que vous avez fait, depuis SAS Data Integration Studio, clic-droit sur votre dossier → Exporter





Exportez vos métadonnées en spécifiant un fichier dans un dossier.
Suivant



Suivant
Suivant
Terminer

Guide de démarrage avec SAS Enterprise Miner

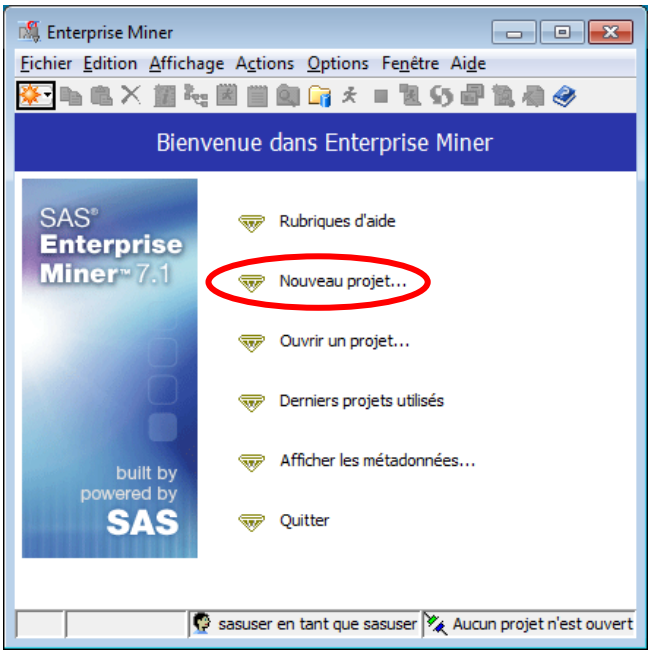
Une banque américaine souhaite créer un modèle de crédit *scoring*. Pour cela, sur un historique de 5960 clients pour lesquels le dossier a été clôturé, nous allons chercher à créer le meilleur modèle nous permettant de prédire le remboursement ou non d'un prêt, en fonction de plusieurs variables.
Sur l'ensemble de ses clients, 80% ont remboursé leur crédit sans incident (retard dans le paiement d'une traite, décès, etc.) et 20% ont eu un incident.

Démarrer SAS Enterprise Miner

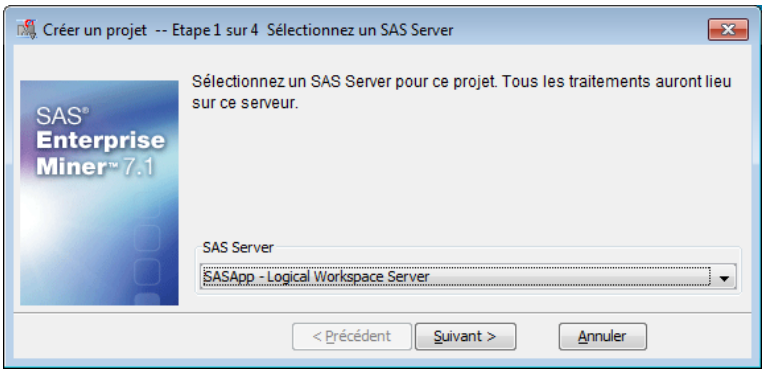
La table HMEQ de cet exemple se trouve dans la bibliothèque SAMPSIO qui se crée automatiquement au démarrage de SAS Enterprise Miner. (Bibliothèque d'exemple SAS)

Lancez SAS
Depuis Démarrer → Programmes → SAS → Analytics → SAS Enterprise Miner Client 7.1

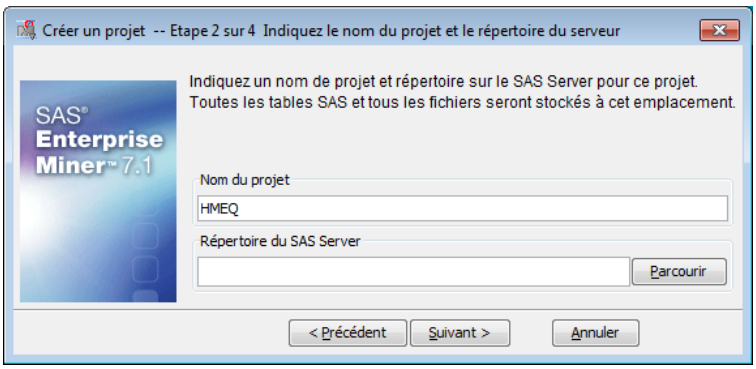
Entrer l'**Identifiant** et son **Mot de passe**
Cliquer sur **Connexion**
Dans le cas d'une installation en client-serveur, décocher **« station de travail personnelle »** et entrer le nom du serveur.



Créer un **nouveau projet**

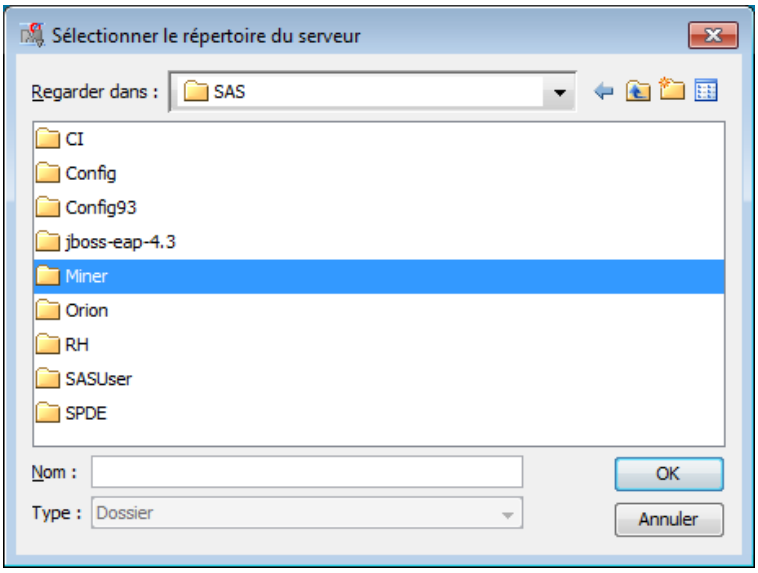


Suivant

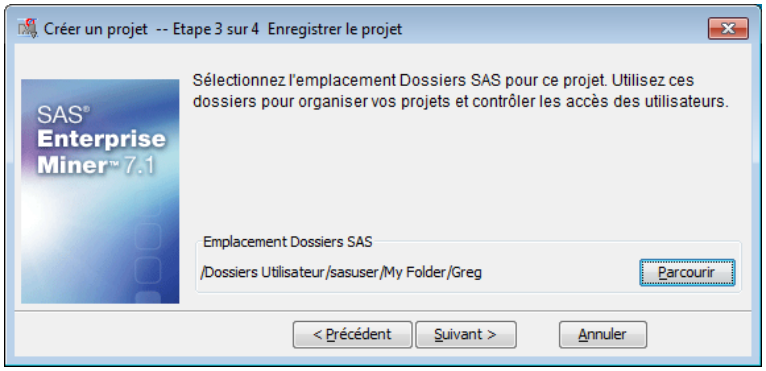


Lui donner un nom : **Votre_nom_HMEQ**

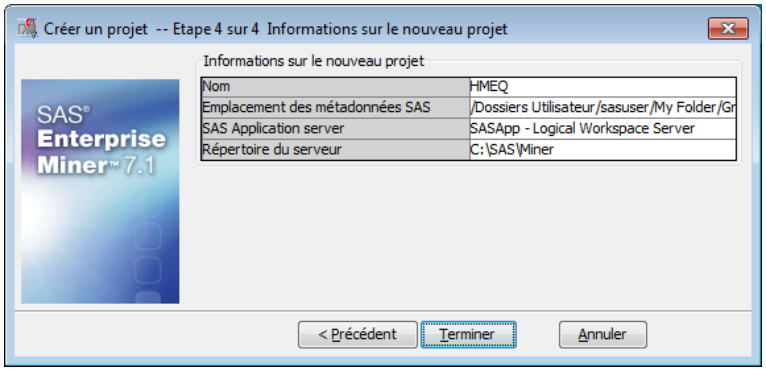
Si vous en avez la possibilité, cliquer sur **Parcourir** pour définir le dossier où vous stockerez votre projet.



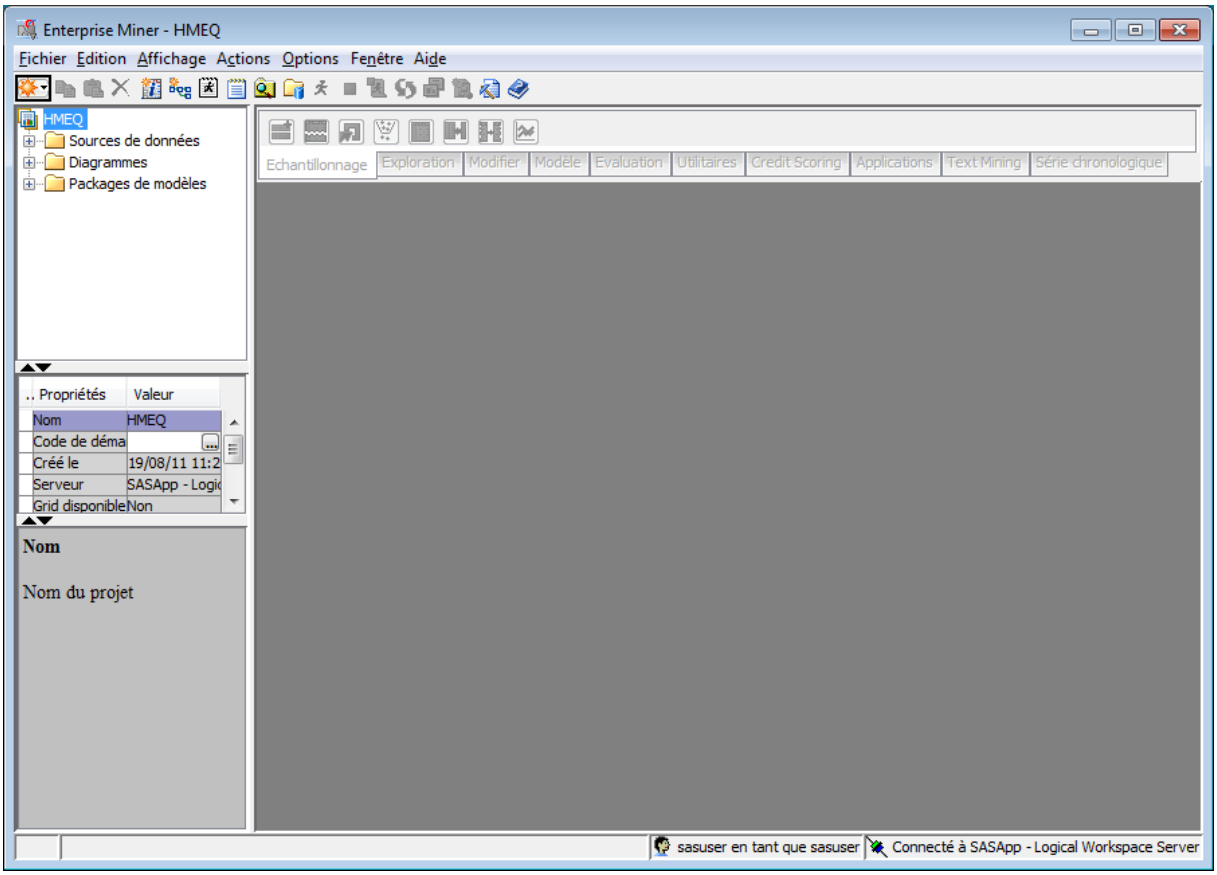
Sélectionner le dossier où vous souhaitez enregistrer votre projet.
OK
Suivant



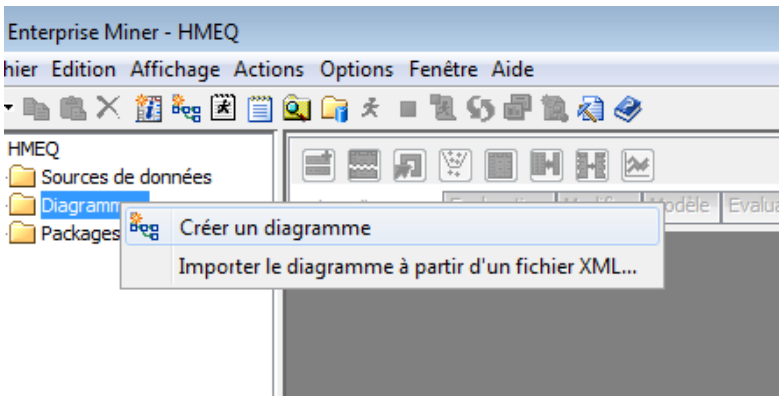
Suivant (votre projet sera donc dans votre dossier.)



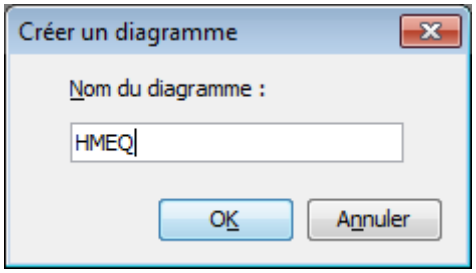
Terminer



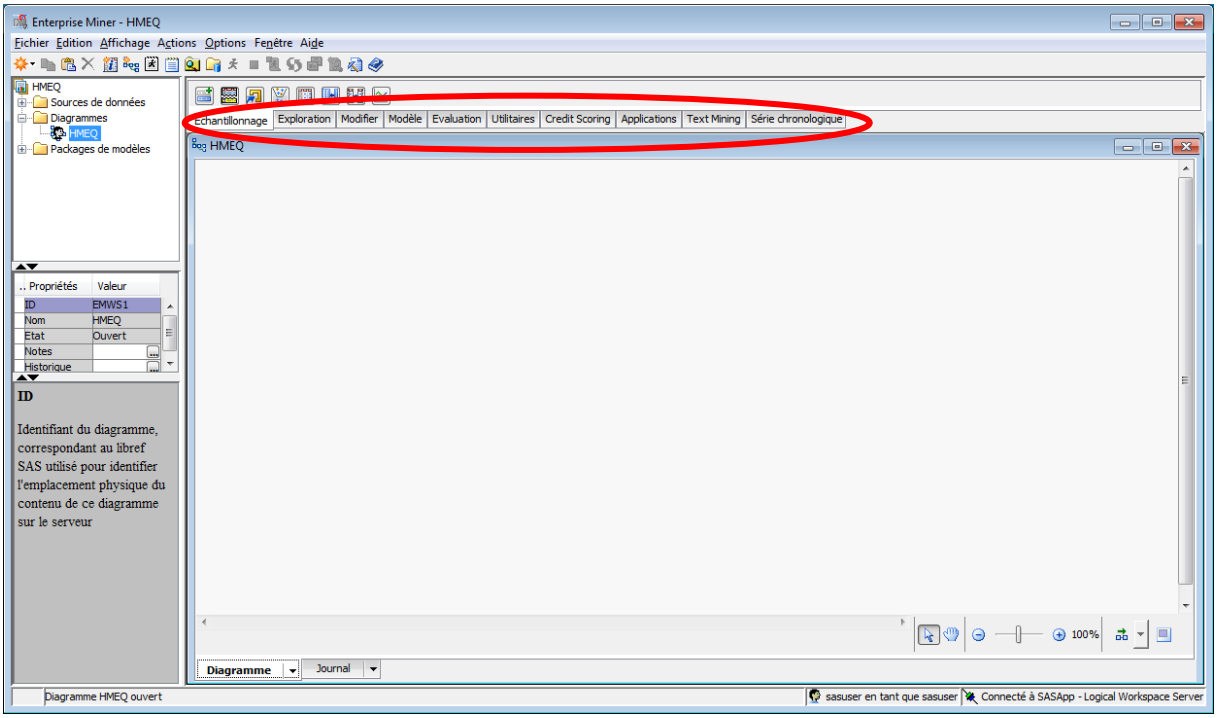
La fenêtre ci-dessus s'ouvre.



Pour créer un nouveau diagramme, un clic-droit sur **Diagrammes → Créer un diagramme**

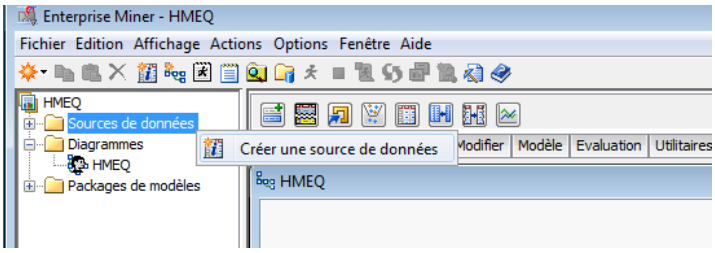


Lui donner un nom : HMEQ
OK

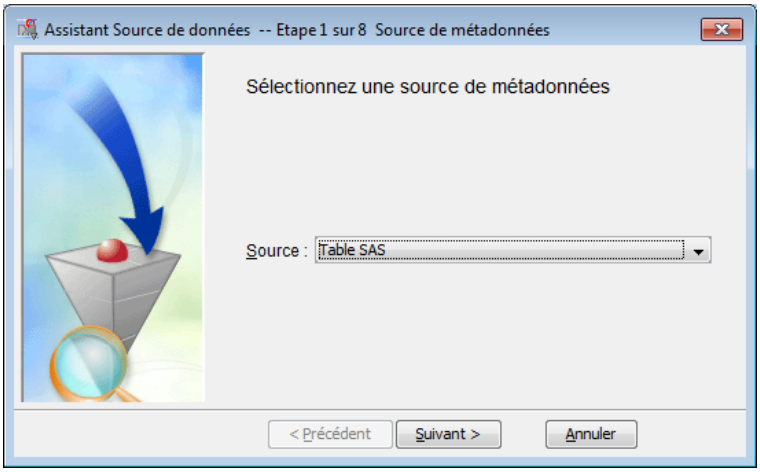


Les outils d'Enterprise Miner sont ordonnés suivant la méthodologie SEMMA : Sample – Explore – Modify – Model – Assess ; soit en français : Echantillonnage – Exploration – Modification – Modélisation – Evaluation.

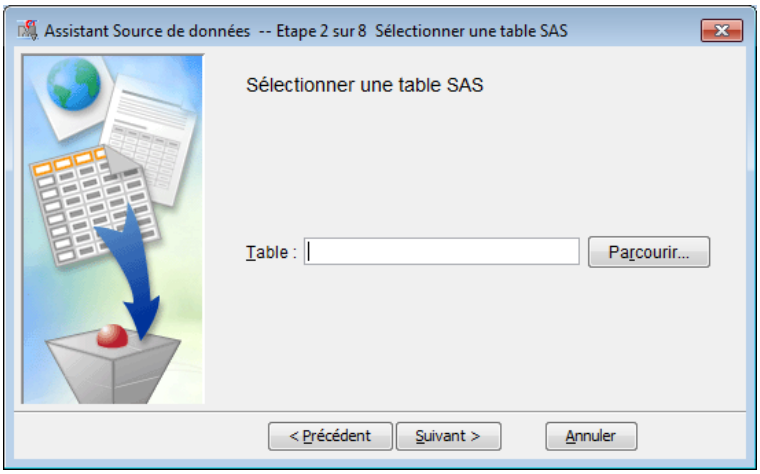
Définition de la table à utiliser



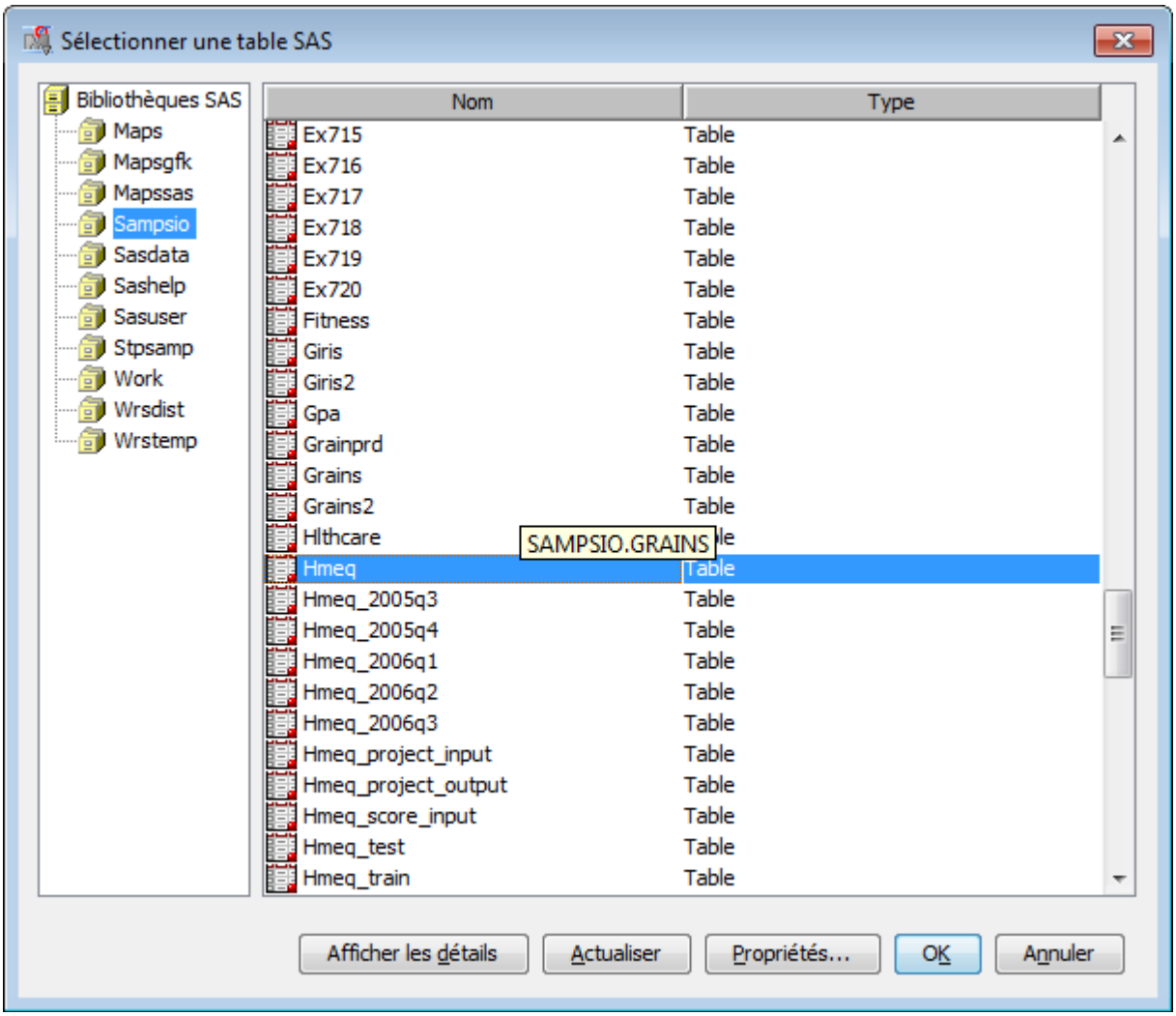
Avant tout projet de *Data Mining* avec Enterprise Miner, il faut commencer par définir la table que l'on utilisera et le rôle de chaque variable.
Clic-droit sur **Sources de données → Créer une source de données**



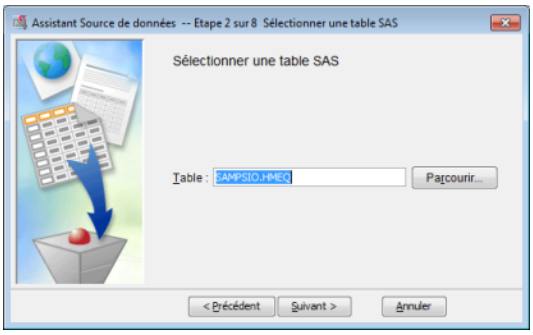
Suivant

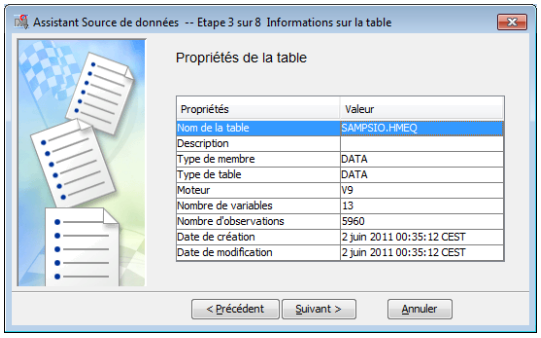


Parcourir

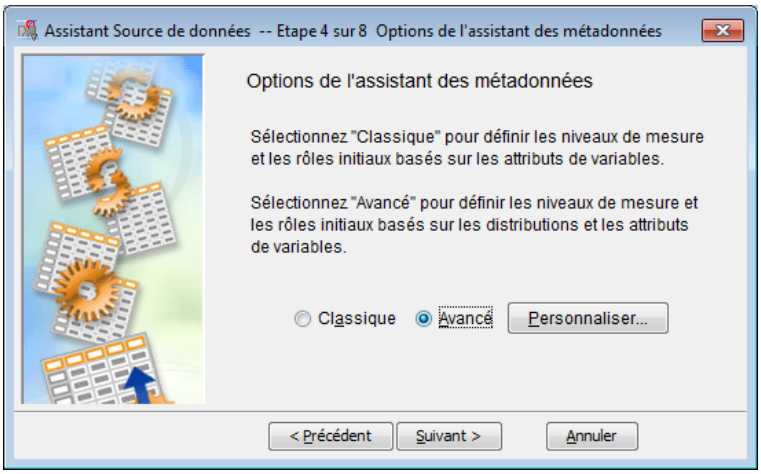


Dans la bibliothèque **Sampsio**, sélectionner la table **HMEQ**
OK



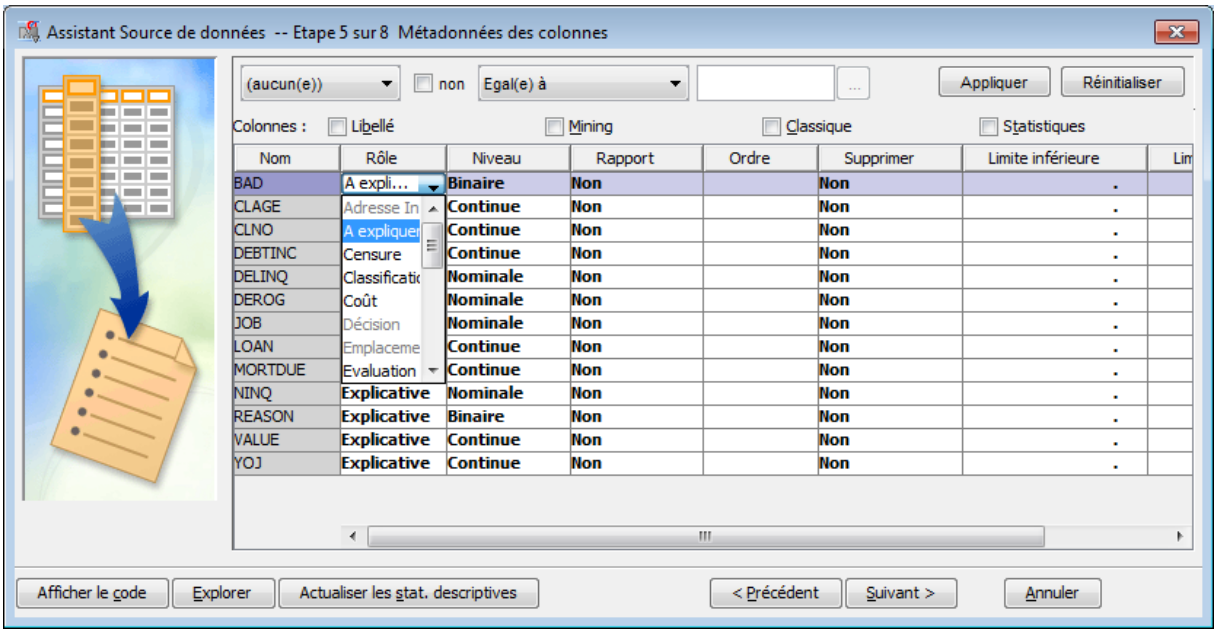


Suivant, suivant



Sélectionner l'option **Avancé**

Suivant



Affecter le rôle **A expliquer** à la variable **Bad**

Les Variables sont :

- Bad : Est-ce que la personne a remboursé son crédit sans incident (0) ou avec (1)
- Loan : Montant du prêt demandé
- Mortdue : Montant de l'hypothèque
- Value : Valeur de la propriété
- Reason : Motif du prêt (DebtCon = consolidation financière ; Homelmp = prêt immobilier)
- Job : Travail du demandeur (Mgr, Office, Other, ProfExe, Sales, Self, non renseigné)
- Yoj : Nombre d'années dans le présent travail
- Derog : Nombre de dérogations
- Delinq : Nombre de litiges
- Clage : Age de la plus ancienne affaire en mois
- Ninq : Nombre de demandes récentes de crédit
- Clno : Nombre d'articles du client dans la banque
- Debtinc : Ratio dette sur revenu

Nous avons donc 12 variables explicatives (variables d'entrée) et une variable à expliquer, variable cible.
Par défaut, toutes les variables ont pour rôle 'explicative', il faut donc spécifier que le rôle de la variable 'Bad' est 'à expliquer'.

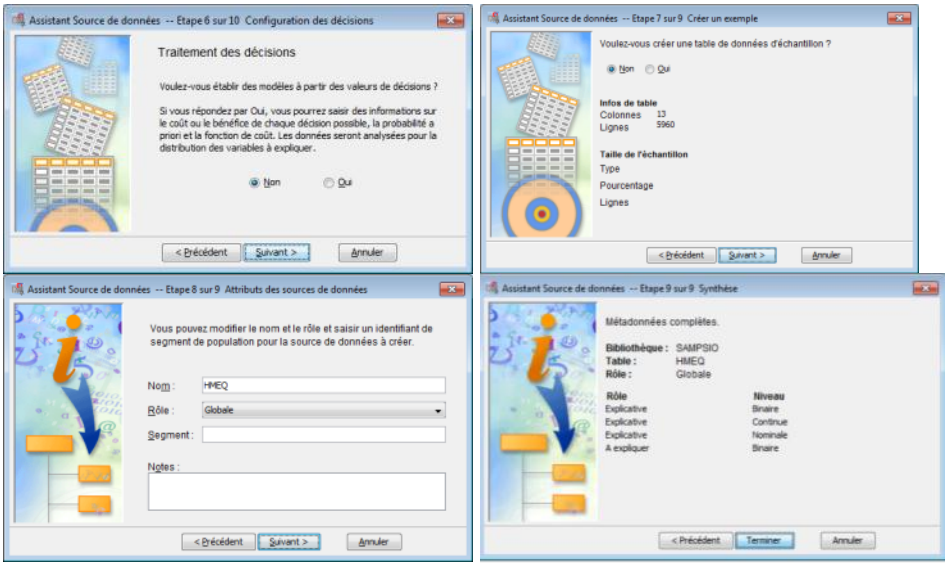
Le **Niveau** est très important :
Il peut être de Quatre types :

1. **Binaire** : deux valeurs, et uniquement deux.
    - Exemple : situation marital : Marié ou non. Si uniquement deux possibilités, la variable est binaire alors que le statut marital (Marié, célibataire, veuf, etc.) n'est pas une variable binaire.
2. **Nominale** : plus de deux valeurs distinctes, sans ordre
    - Exemple : la variable travail : Mgr, Office, Other, ProfExe, Sales, Self
3. **Ordinale** : plus de deux valeurs distinctes et ordonnées (par un nombre entier)
    - Exemple 1 : Type de carte de crédit :
        - 0 pour pas de carte
        - 1 pour carte de retrait
        - 2 pour carte de credit
        - 3 pour carte Gold
        - 4 pour carte Platinum
    - Exemple 2 : groupe d'âge :
        - 1 pour les 0 – 6 ans
        - 2 pour les 7 – 11 ans
        - 3 pour les 12 – 25 ans
        - 4 pour les 26 – 59 ans
        - 5 pour les plus de 60 ans

    Ce n'est pas parce que vous êtes dans la classe 4 que vous êtes deux fois plus vieux que ceux de la seconde classe. Vous êtes juste dans une classe supérieure.
4. **Continue** : Nombre entier ou réel
    - Exemple : Montant du prêt

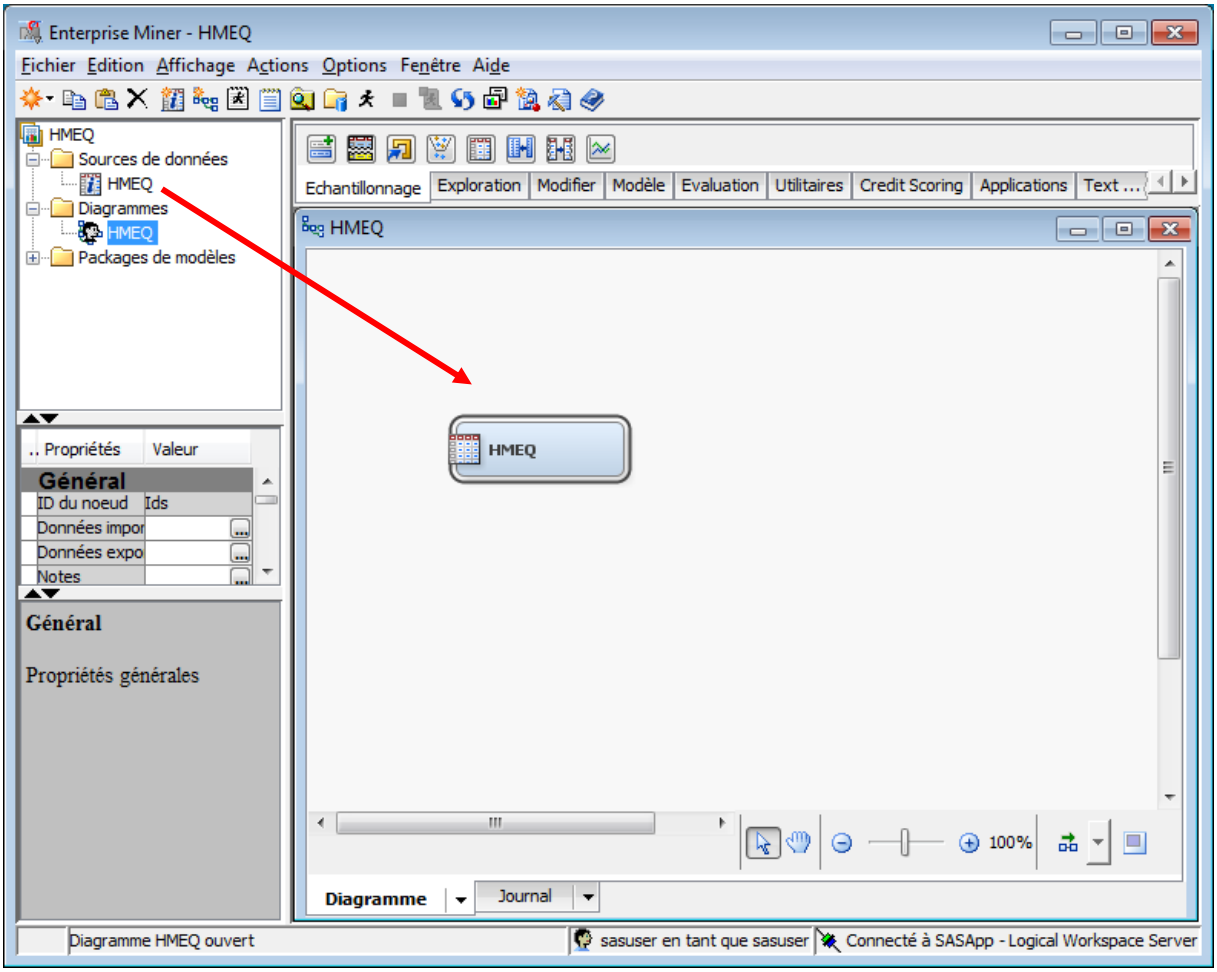
Les niveaux binaire, nominale et ordinale sont des niveaux de variables qualitatives alors que le niveau continu est pour des variables quantitatives. Pour les variables quantitatives, on peut utiliser les opérateurs mathématiques comme l'addition, la multiplication, la soustraction ou la division, alors que ce n'est pas possible avec les variables qualitatives, dites de classe.

Lorsque vous avez les mêmes informations que ci-dessus, vous avez sélectionné la table que vous voulez analyser (HMEQ) et vérifié le rôle et le niveau des variables.

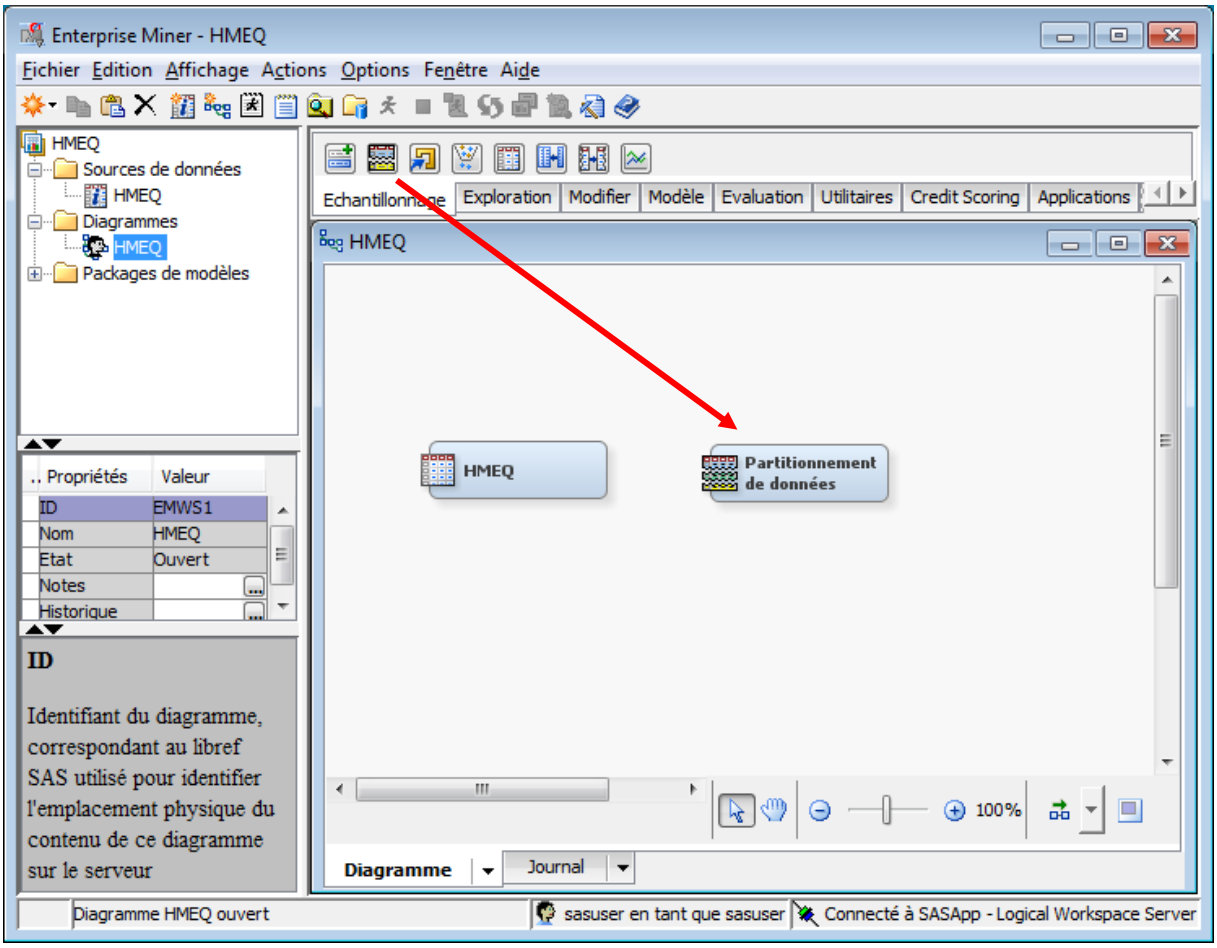


Suivant, Suivant, Suivant, Terminer Vous avez dans le dossier des sources de données, la table HMEQ, où vous avez défini le rôle de chaque variable.

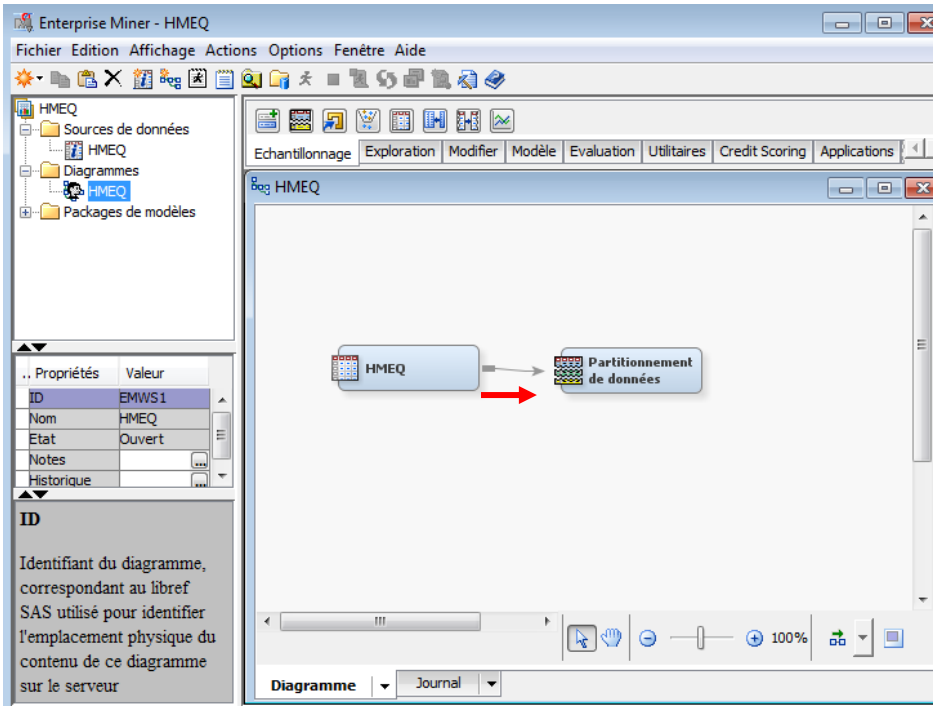
Création du diagramme



Sélectionnez la table HMEQ que vous venez de définir, des sources de données, glissez et lâchez la dans l'espace de travail.



Dans l'onglet **Echantillonnage**, sélectionner l'outil de **Partitionnement des données**, glissez et lâchez le dans l'espace de travail.



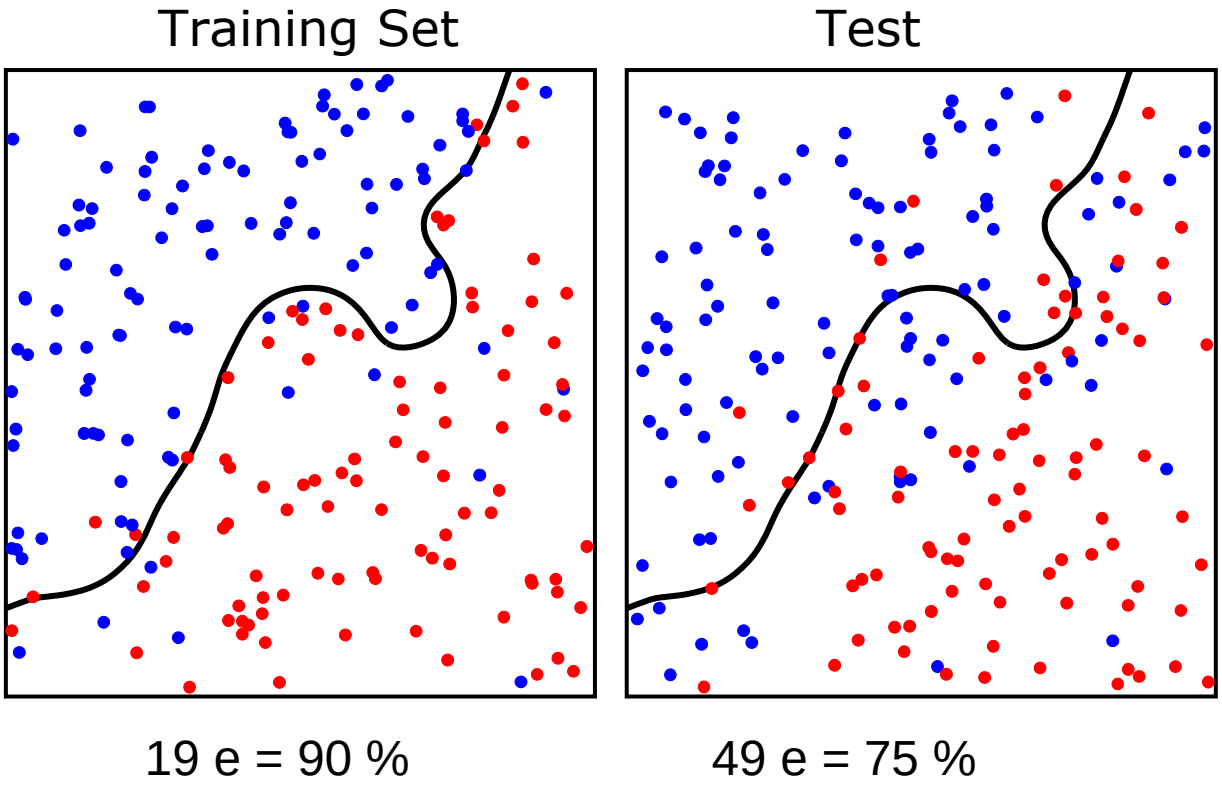
Pour tracer la flèche du flux de processus de Data Mining, cliquez sur le point à droite de l'objet HMEQ sur plan de travail, un petit crayon apparait, glissez jusqu'à l'icône 'Data Partition' et lâchez.

L'outil Data Partition permet de partitionner notre table d'entrée en deux ou trois échantillons. Dans notre cas, nous allons partitionner la table HMEQ en deux sous-ensembles, 60% pour apprendre, et le reste, soit 40%, pour valider que notre modèle est robuste.

En effet, nous cherchons à construire un bon modèle robuste, c'est-à-dire un modèle qui ait bien appris, qui différencie bien les bons des mauvais clients, mais qui soit capable aussi de définir la probabilité d'un client de rembourser son crédit, sur des données différentes de celles sur lesquelles il a appris.

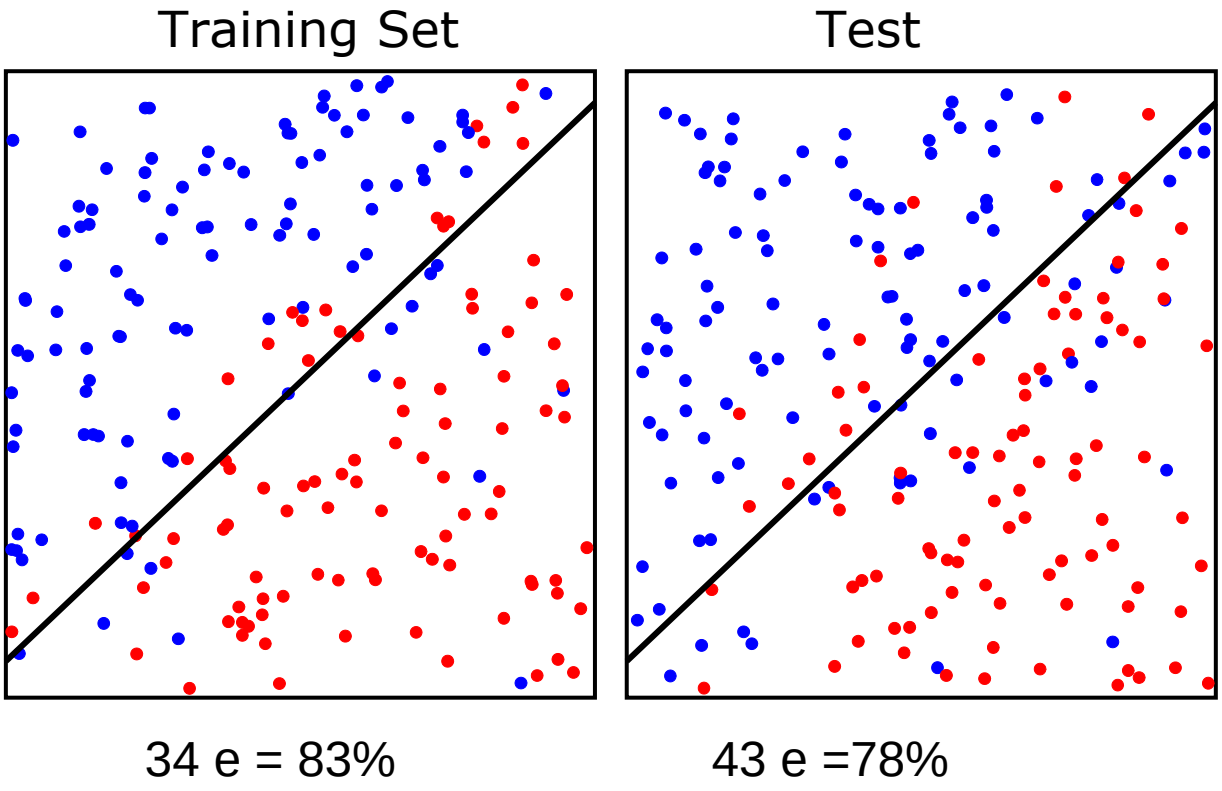
Prenons un exemple : Nous essayons ici de départager les rouges des bleus.

Voici ci-dessous un modèle en sur apprentissage, qui a trop bien appris sur les données d'apprentissages (Training Set) seulement 19 erreurs (bleu classé chez les rouge et vice-versa) , mais qui, sur des données légèrement différentes de celles sur lesquelles il a appris, il fait 49 erreurs, soit seulement 75% de bien classé.

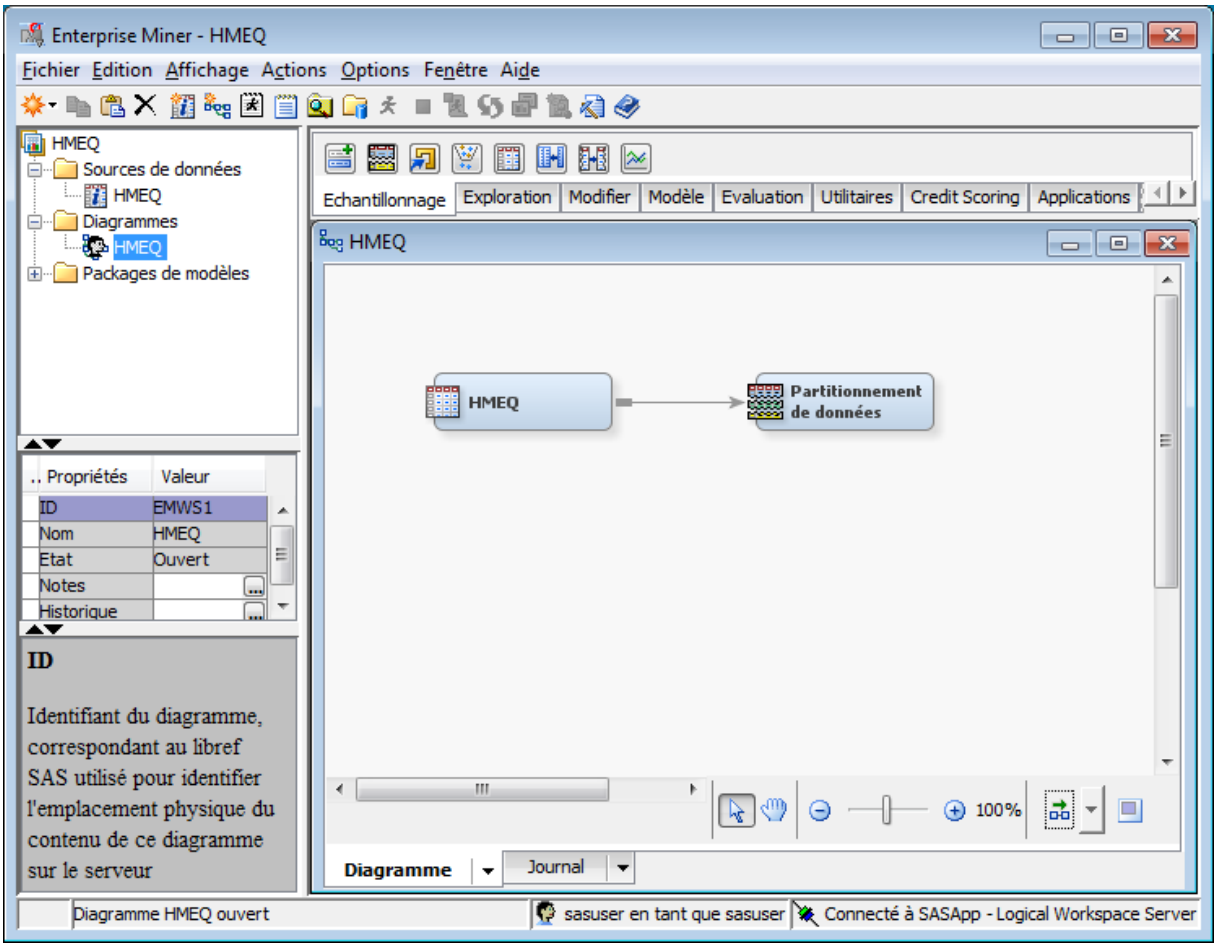


Le modèle suivant a moins bien appris, si on l'applique sur les données sur lesquelles il a appris, il fait 34 erreurs, par contre sur la table de test (Test Set : Base de données différente de celle sur laquelle le modèle a appris) le modèle ne fait que 43 erreurs, soit un gain de trois pourcent par rapport au précédent.

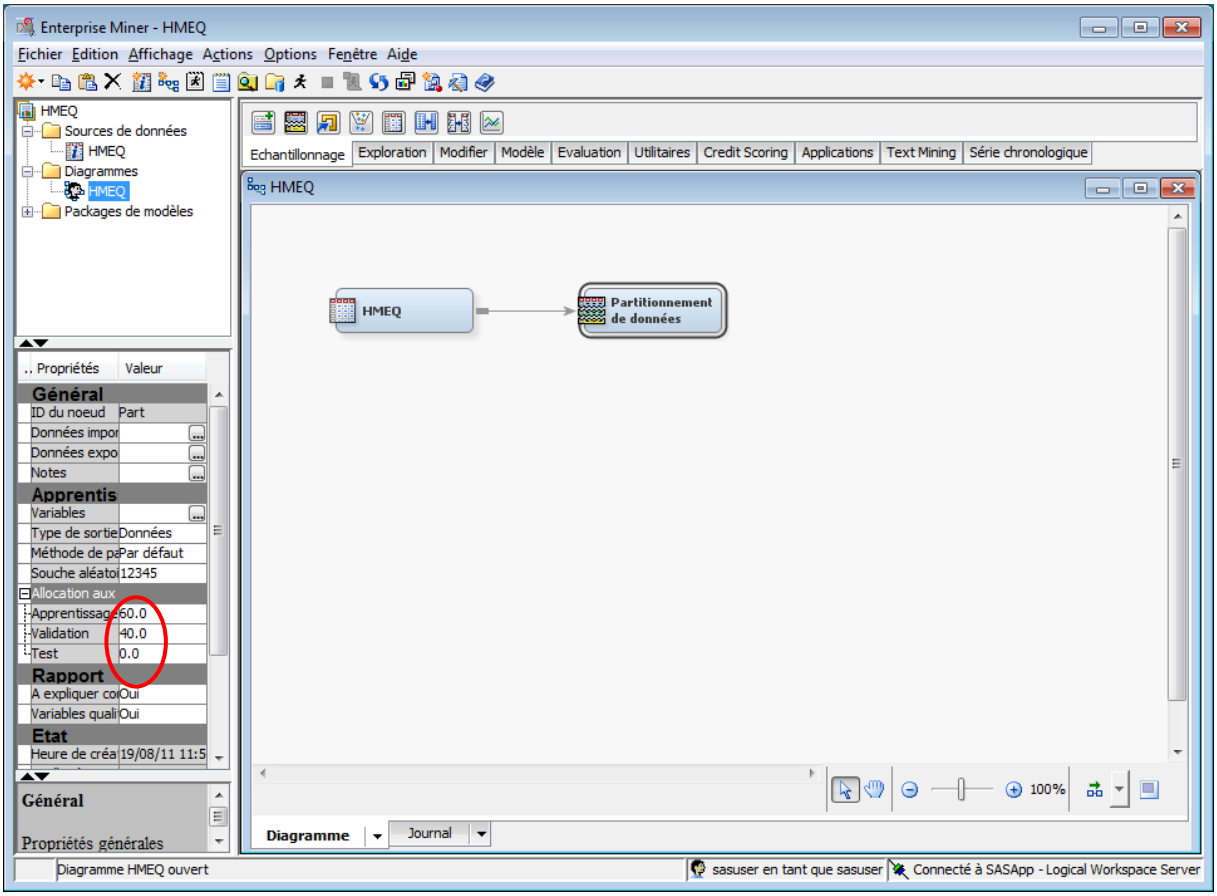
Le but du *Data Mining* n'est pas de construire un modèle parfait, mais un bon modèle robuste, c'est-à-dire un modèle qui donne de relativement bons résultats sur des données différentes de celles sur lesquelles il a appris.



L'outil 'Data Partition', permet donc de séparer notre base en deux échantillons distincts, 60% pour apprendre, et 40% pour valider, ce qui nous permettra d'affirmer ou non la robustesse' de notre modèle.

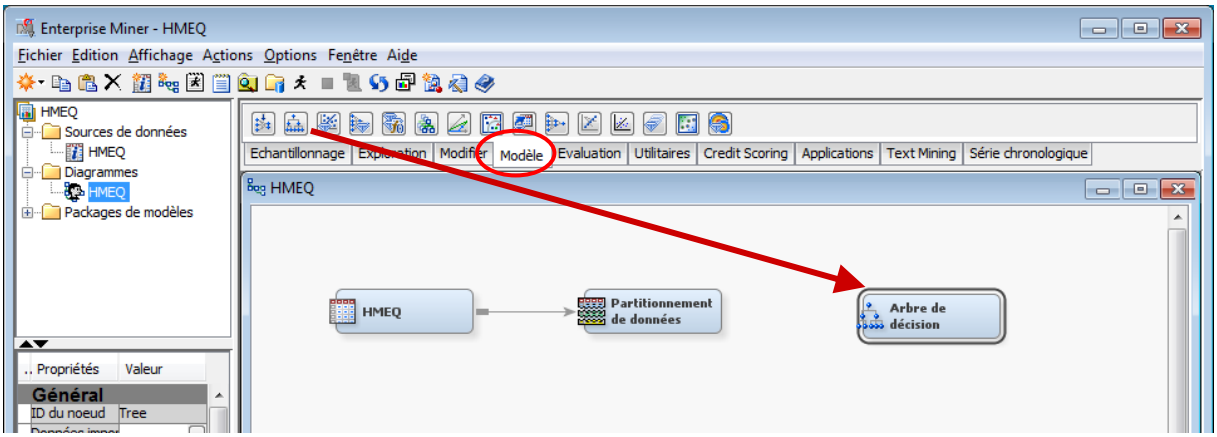


Une fois la flèche tracée, cliquez sur ‘Data Partition’ pour le sélectionner.
Dans notre exemple, nous n’utiliserons pas de table de Test.
Nous allons donc spécifier que l’on souhaite partitionner la table de départ en deux sous-ensembles : 60% pour la table d’apprentissage et 40% pour la table de validation.
Objectif : nous aurons donc une table de données différente de celle sur laquelle le modèle aura appris, ce qui nous permettra de valider la robustesse de notre modèle.



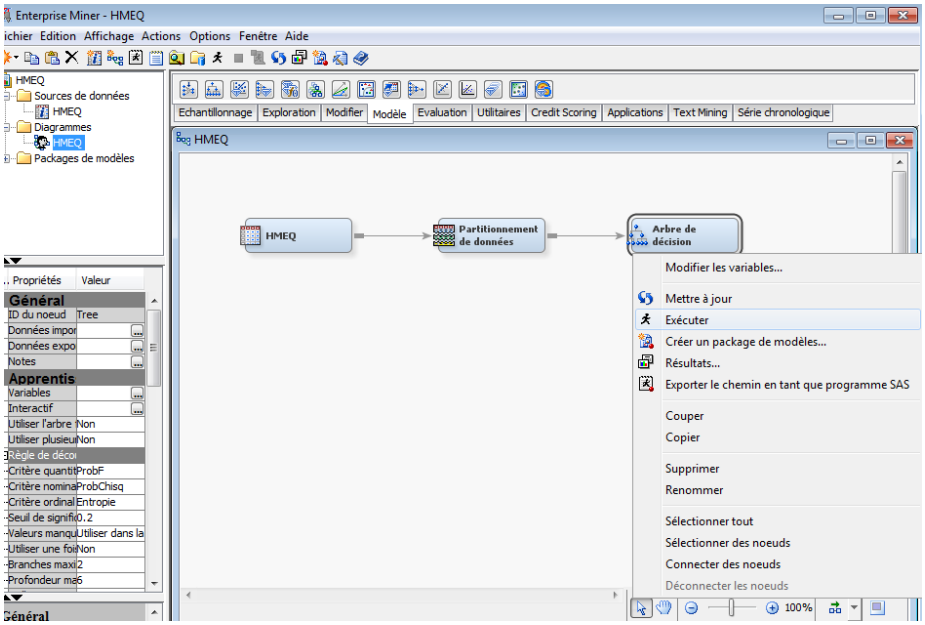
Entrez les chiffres 60 pour la table d'apprentissage, 40 pour la table de validation et 0 pour celle de Test dans la partie des propriétés, à gauche du processus.

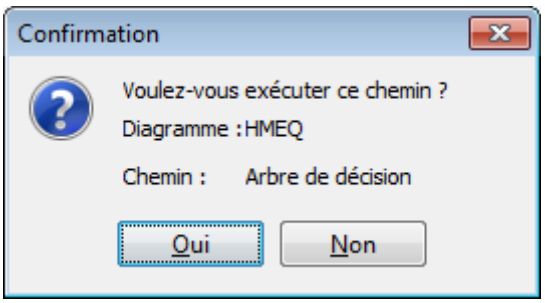
Arbre de décision



Suite au partitionnement des données, nous pouvons mettre en compétition plusieurs modèles.
Commençons par un arbre de décision.
Dans l'onglet Modélisation, sélectionnez l'arbre de décision, glissez et lâchez le dans le diagramme.
Une fois l'arbre posé sur le plan de travail, reliez le Partitionnement des données à l'arbre.

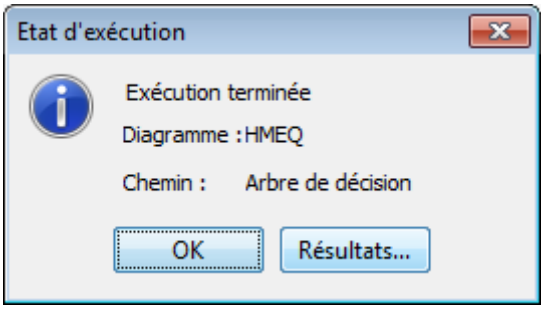
Pour l'exécuter, cliquez droit sur l'icône l'arbre, cliquez gauche sur Exécuter.



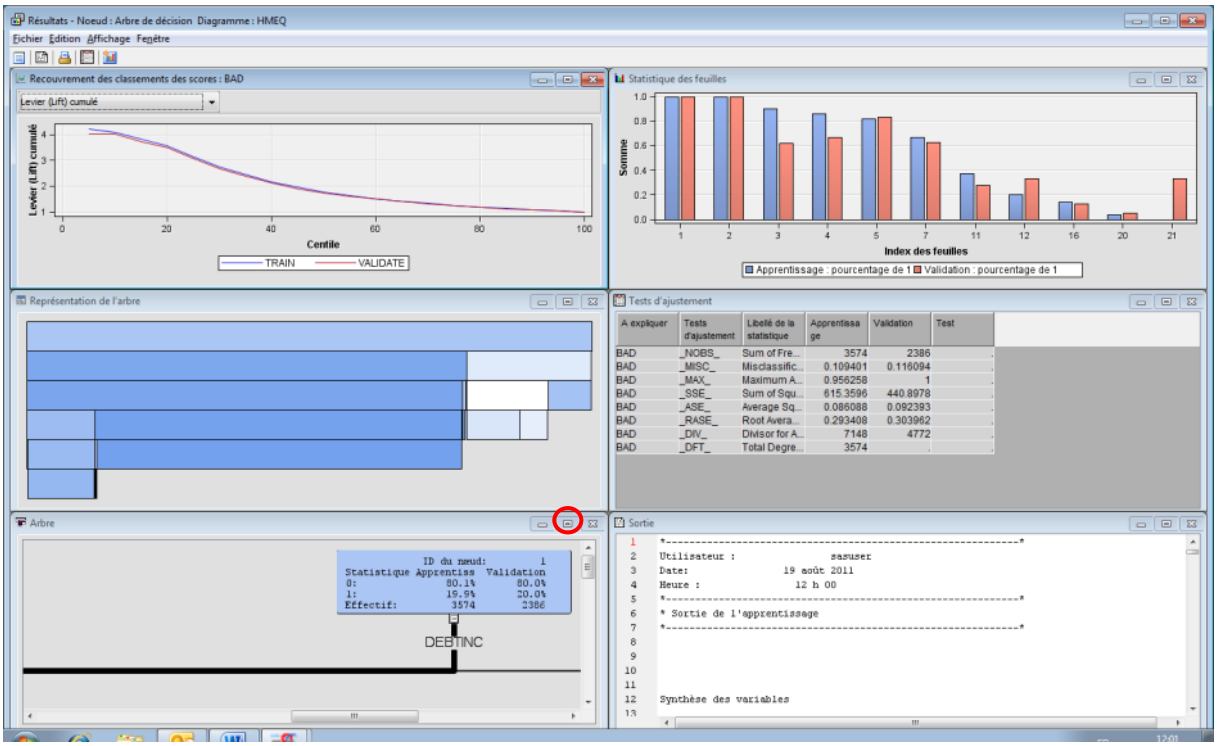


Oui

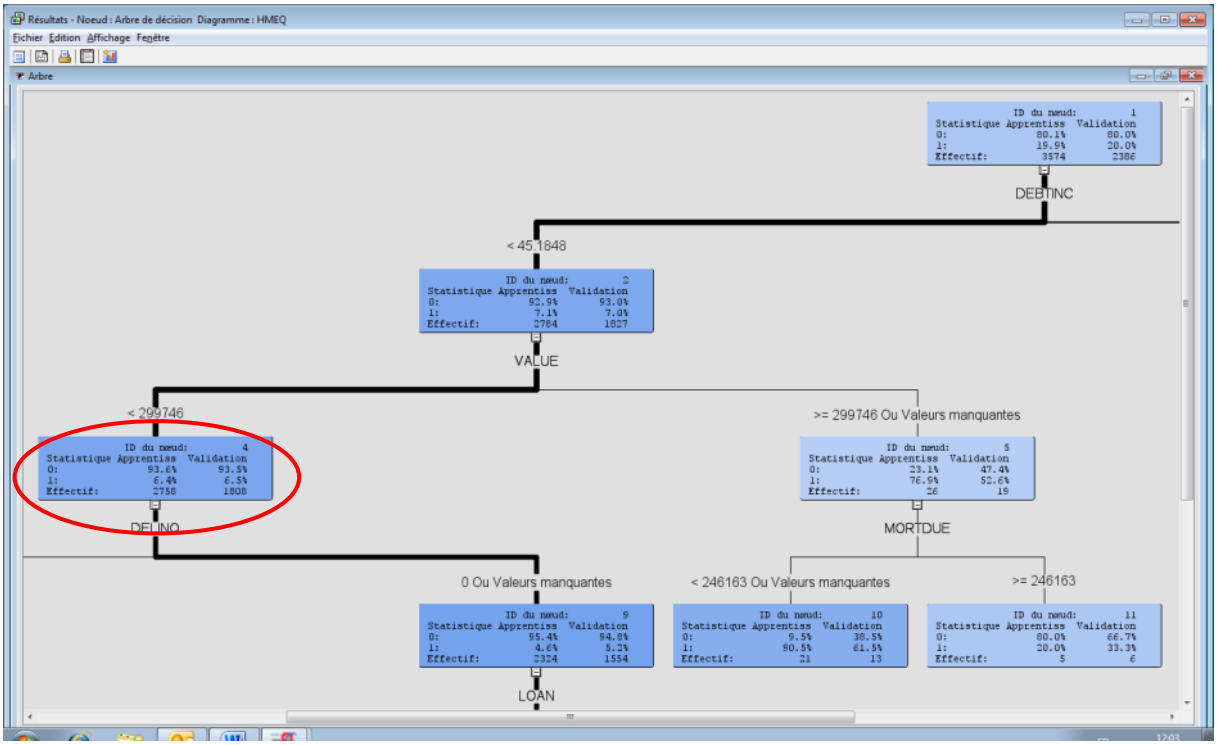
Lorsque l'exécution est terminée



Cliquez sur **Résultats** pour regarder les résultats.
Si vous avez cliquez sur OK, clic-droit sur l'arbre → résultat.



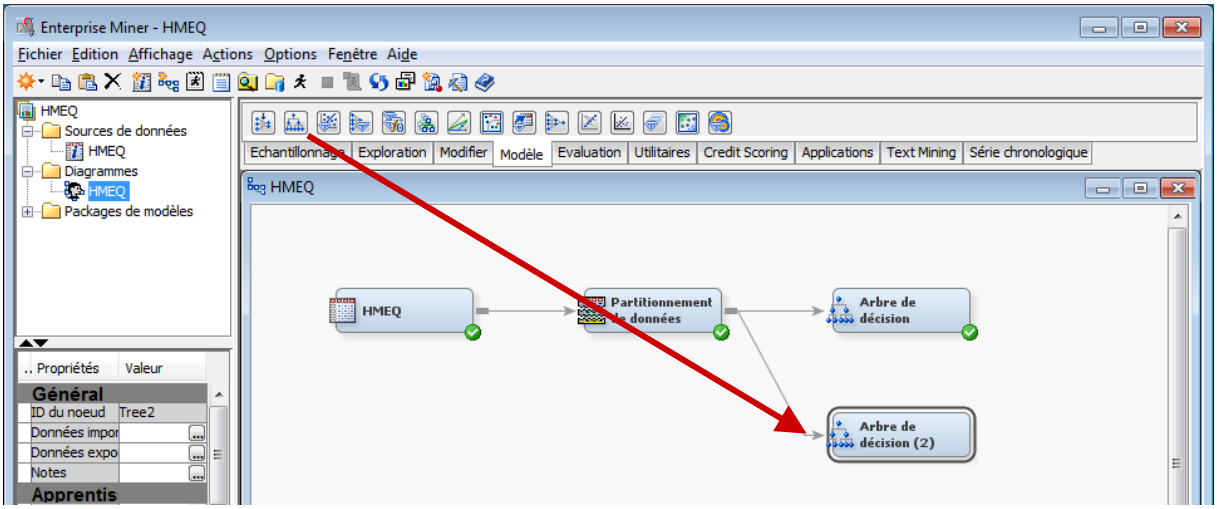
Dans les résultats de l'arbre décision, agrandir la fenêtre de l'arbre (en bas à gauche dans notre cas) pour afficher l'arbre de décision.



Pour chaque nœud, la colonne du milieu se réfère à la table d'apprentissage et celle de droite à la table de validation. Les deux premières lignes indiquent le pourcentage de bons (0) et celui de mauvais (1) payeurs. La troisième et dernière ligne indique le nombre total d'individus dans le nœud. Le nœud tout en haut est appelé la racine de l'arbre. Il divise en deux branches et ainsi de suite jusqu'aux feuilles de l'arbre qui sont les terminaisons les plus en bas de l'arbre.

Interprétation du nœud en bas à gauche (encerclé en rouge): parmi les personnes ayant un ratio dette sur revenu (Debtinc) inférieur à 45.18 et une valeur de la propriété (value) inférieur à 299 746$; 6.4 % de ces client sont des mauvais payeurs (1).

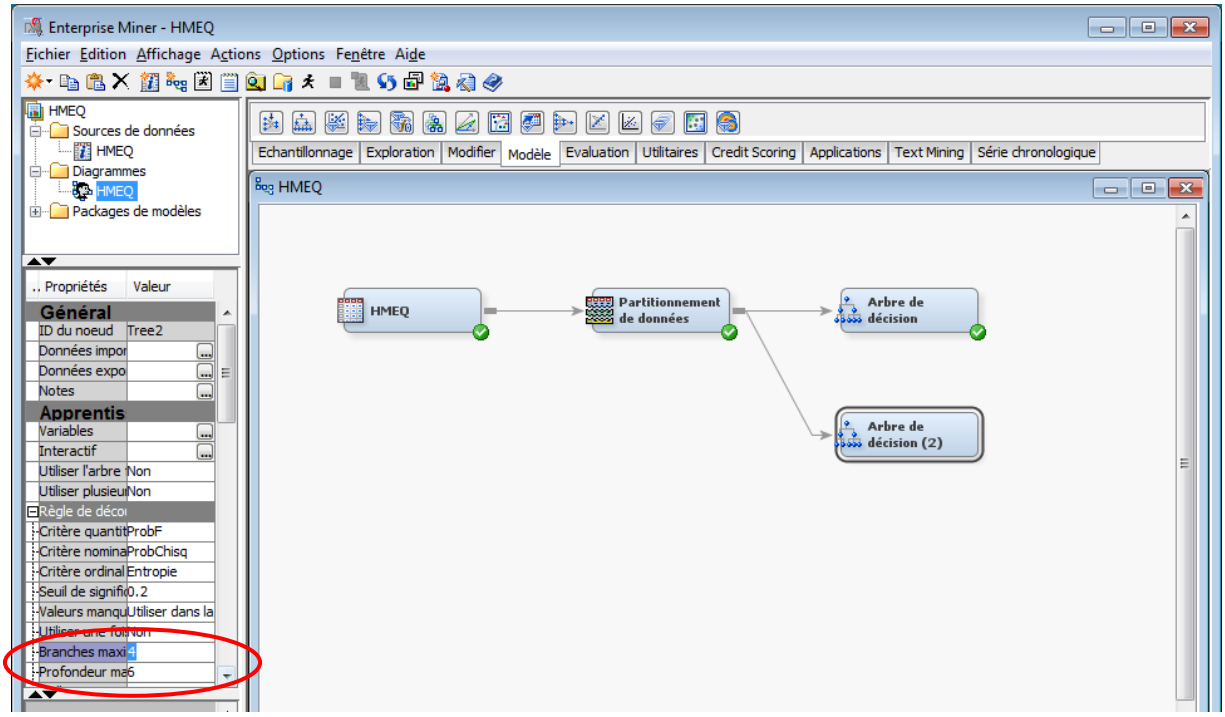
Pour fermer les résultats, cliquez sur la croix en haut à droite, ce qui vous ramène au plan de travail.



Nous allons maintenant créer un deuxième arbre de décision avec des paramètres différents du premier, à savoir un arbre à quatre branches.

Pour cela, ajouter à votre diagramme un arbre de décision, puis tracer une flèche de l'outil de partitionnement des données à l'arbre.

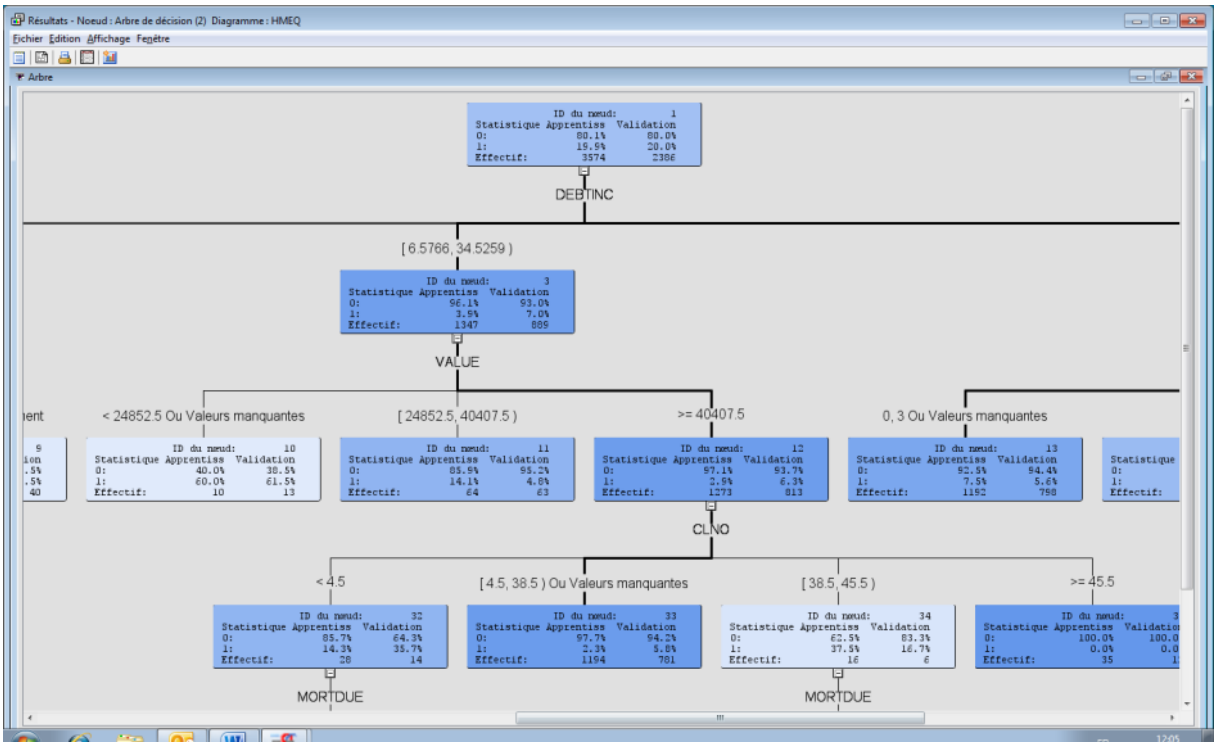
Dans les propriétés de ce nouvel arbre ; sélection le second arbre en cliquant dessus :



Sur la ligne Branches Maximum, entrer le chiffre 4.

Clic-droit sur l'arbre modifié →Exécuter.

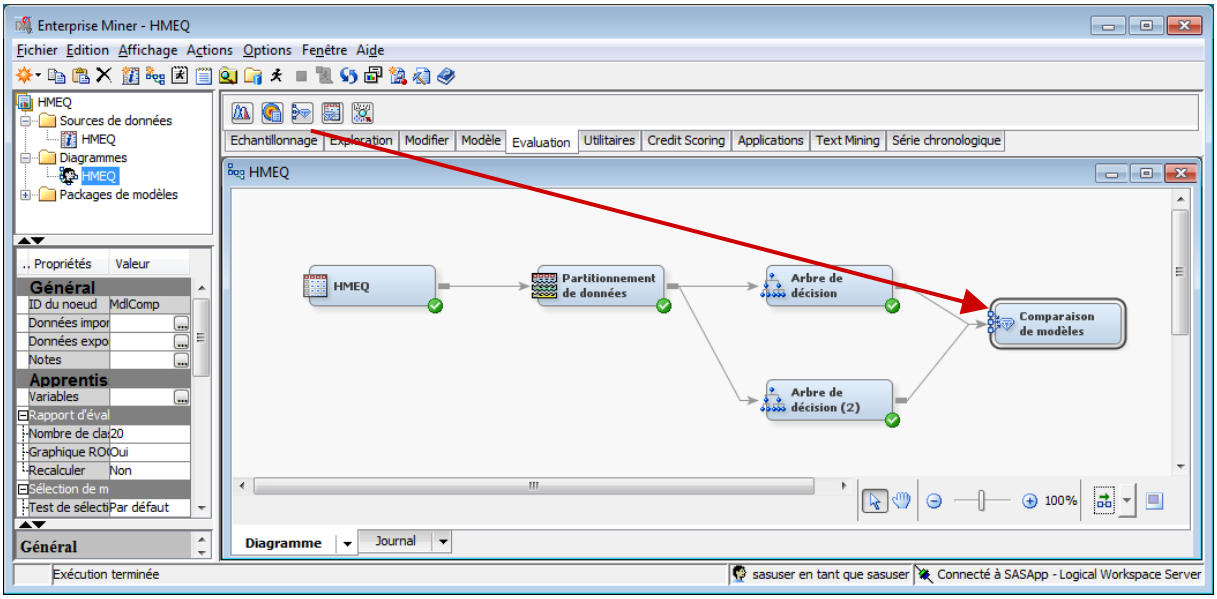
Regarder les résultats.



L’arbre créé à deux, trois ou quatre branches à chaque niveau.

Comparaison de modèles

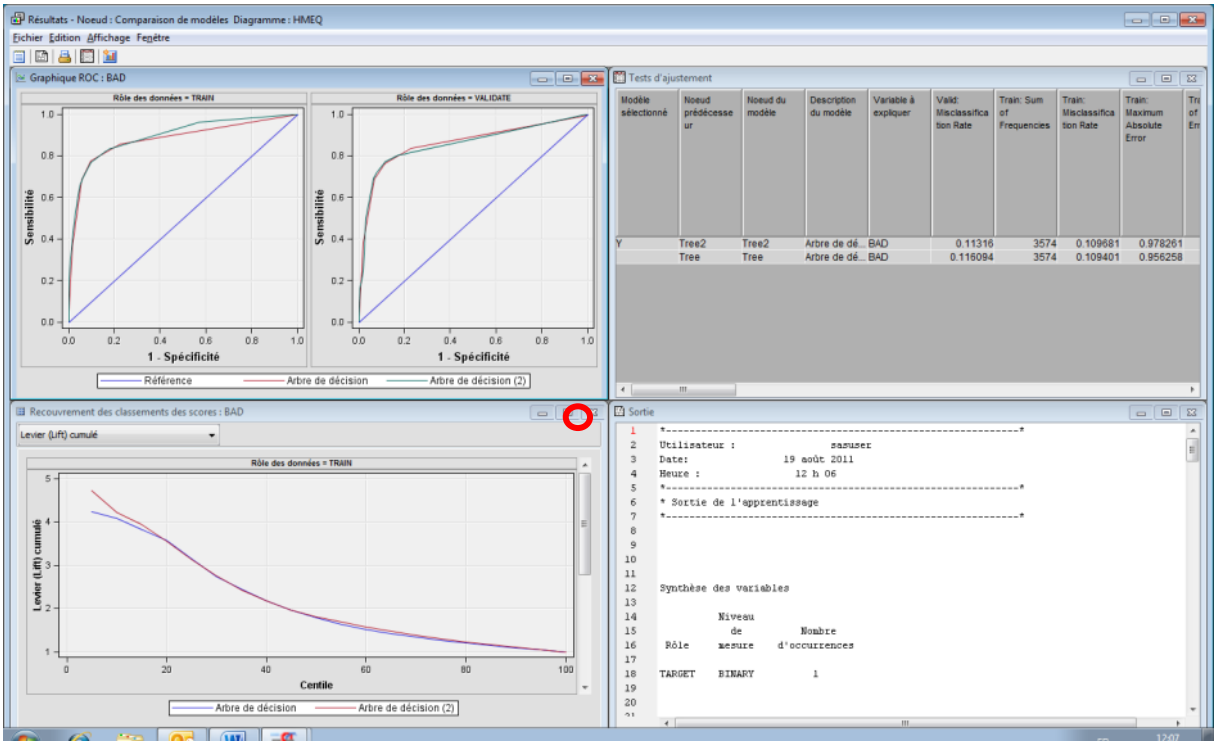
Pour pouvoir comparer les arbres de décision construits précédemment, nous allons utiliser l'outil de comparaison de modèle.



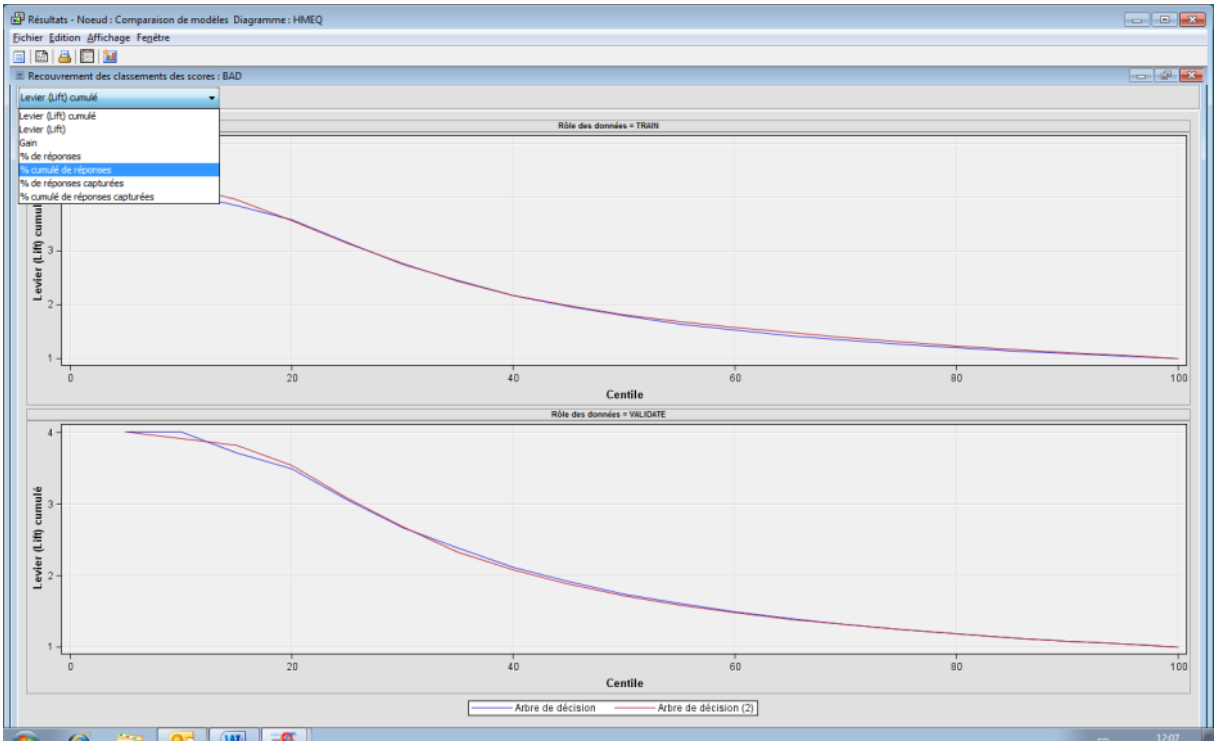
Sélectionner l'outil de comparaison de modèles dans l'onglet évaluation.
Après avoir ajouté la comparaison de modèles au flux de processus, reliez les deux arbres à cet outil.

Vous pouvez mettre en compétition plusieurs modèles, comme différents arbres de décision, des régressions (dans ce cas d'une variable cible binaire, elles sont logistiques), des réseaux neuronaux, etc.

Clic-droit sur cet outil → exécuter le processus.
Ouvrir les résultats.



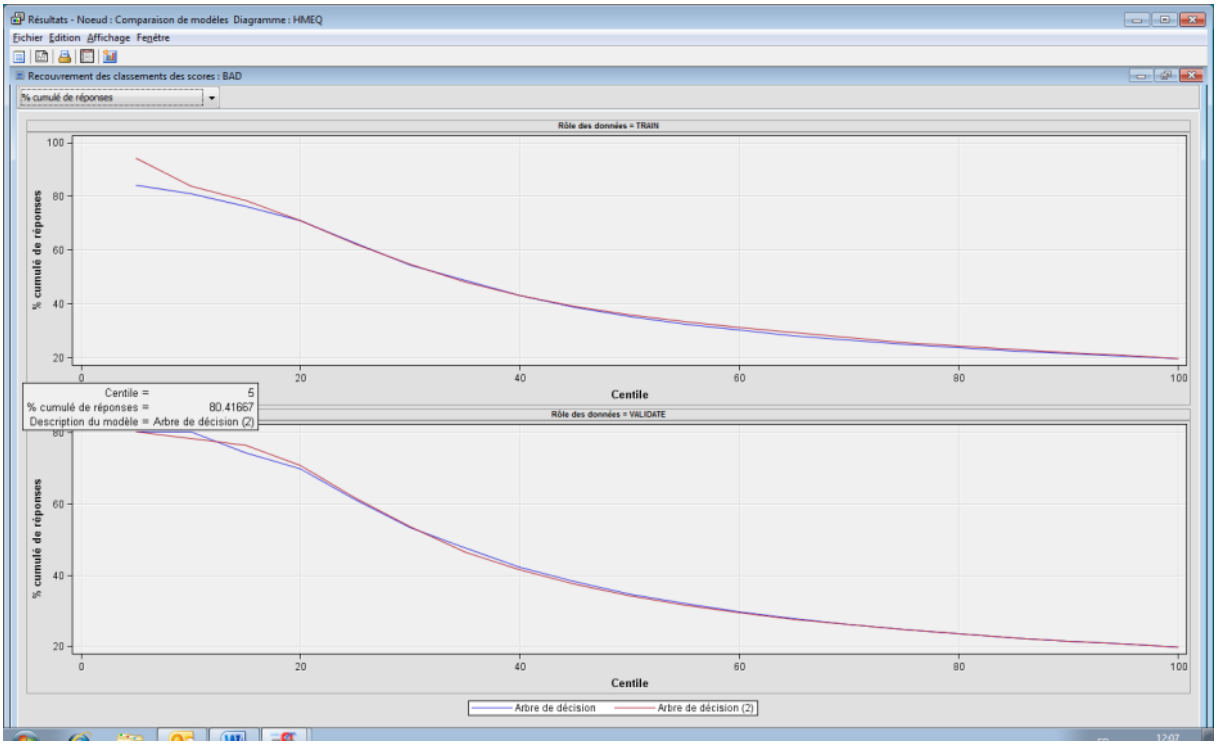
Agrandir la fenêtre de recouvrement des classements des scores



Sélectionnez la courbe du pourcentage de réponses cumulées, qui montre que les deux modèles d'arbre de décision ont des performances similaires.

L'axe horizontal présente les déciles de population ordonnés suivant les probabilités décroissantes d'être un client qui ne rembourse pas. A l'origine (en zéro), se trouve le client ayant la plus forte probabilité données par chaque modèle, de ne pas rembourser son crédit.

L’axe vertical représente le pourcentage cumulé par décile de client qui effectivement ne rembourse pas leur crédit.
Cette courbe est tracée sur les données de la table de validation.



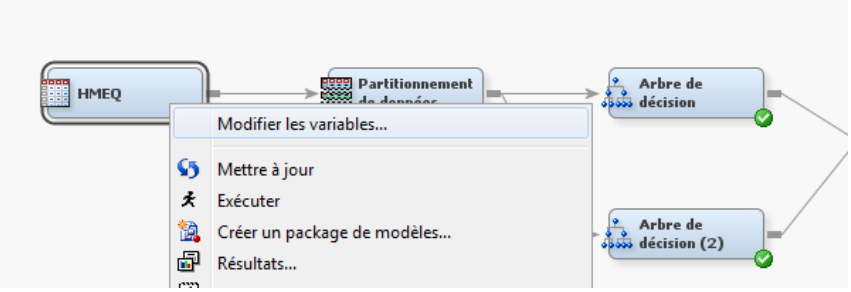
Interprétation de la courbe rouge de l’arbre de décision (2), à quatre branches, sur la table de validation au point 5 – 80.4 (en approchant la souris de ce point, un encadré apparaît.

Si l’on sélectionne 5% de la population de la table de validation, ayant la probabilité la plus forte, d’après l’arbre de décision (2), parmi ces 5%, 80.41% sont effectivement des mauvais payeurs.

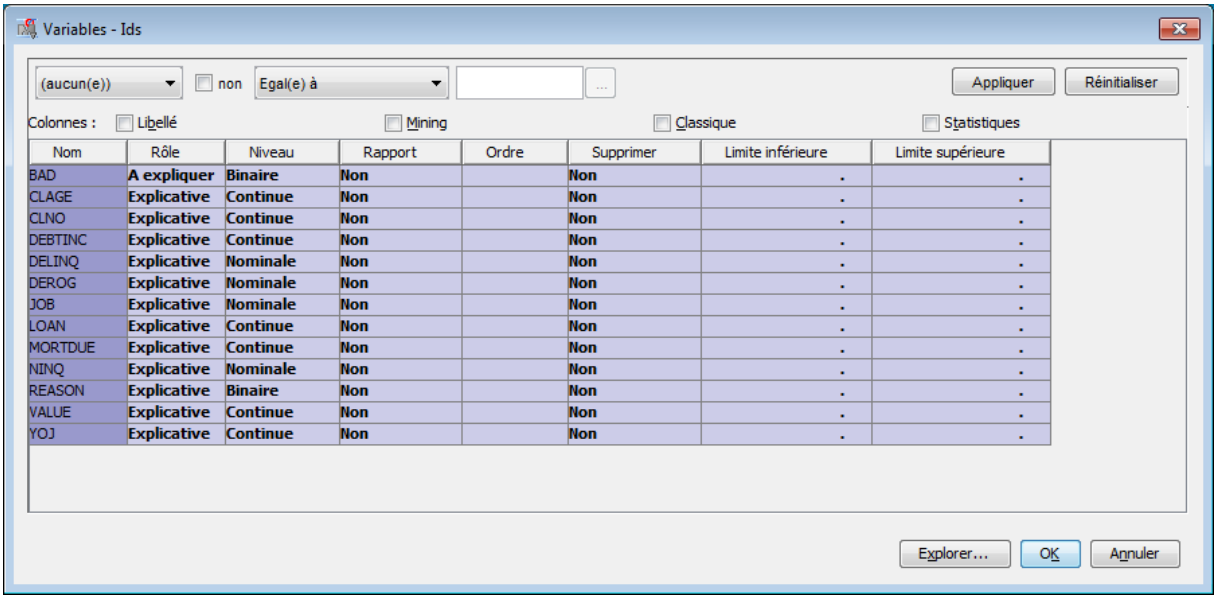
Fermer les résultats de l’outil de comparaison de modèles et revenir sur le plan de travail.

Remplacement des valeurs manquantes

L'arbre de décision est un modèle qui supporte les valeurs manquantes alors que la régression et les réseaux neuronaux ne les supportent pas.
Il est donc important de placer l'outil de remplacement des valeurs manquantes avant la régression et le réseau de neurone s'il y a des valeurs manquantes.



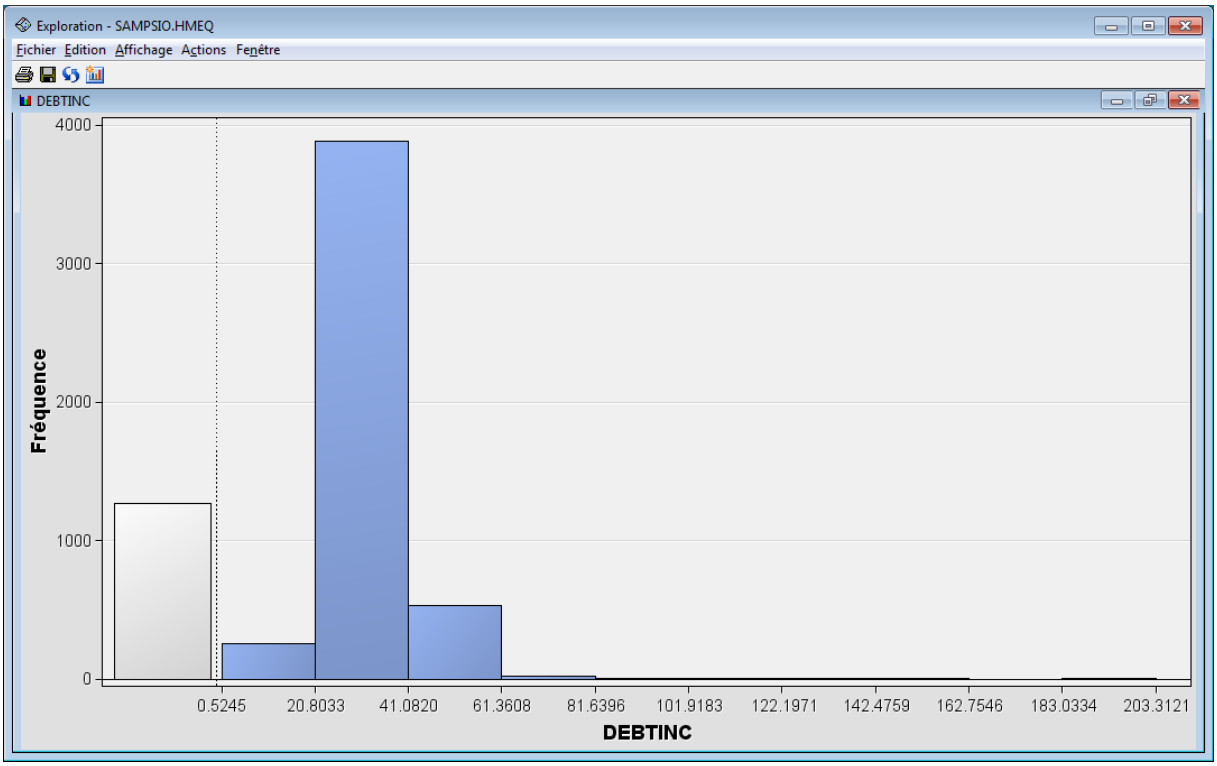
Revenons à notre processus et jetons un coup d'œil à nos variables. Clic-droit sur l'a table HMEQ de départ → Modifier les variables…



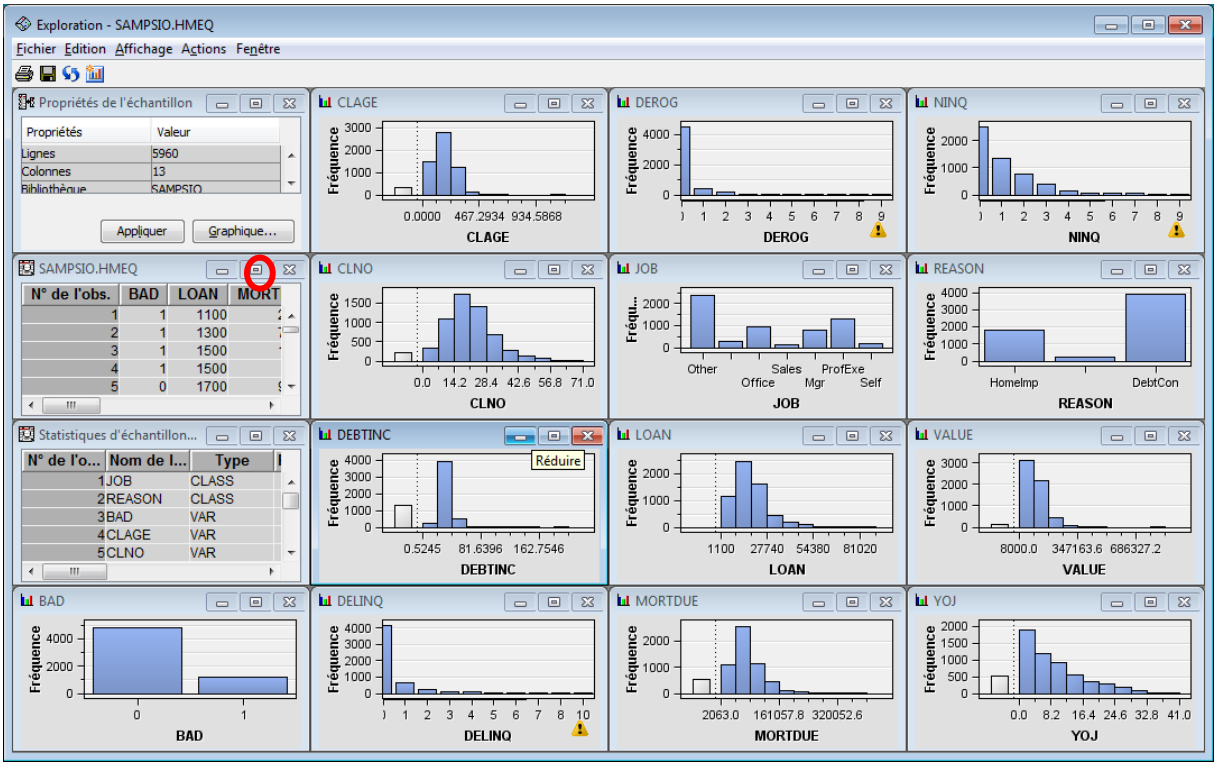
Nom	Rôle	Niveau	Rapport	Ordre	Supprimer	Limite inférieure	Limite supérieure
BAD	A expliquer	Binaire	Non		Non	.	.
CLAGE	Explicative	Continue	Non		Non	.	.
CLNO	Explicative	Continue	Non		Non	.	.
DEBTINC	Explicative	Continue	Non		Non	.	.
DELINQ	Explicative	Nominale	Non		Non	.	.
DEROG	Explicative	Nominale	Non		Non	.	.
JOB	Explicative	Nominale	Non		Non	.	.
LOAN	Explicative	Continue	Non		Non	.	.
MORTDUE	Explicative	Continue	Non		Non	.	.
NINQ	Explicative	Nominale	Non		Non	.	.
REASON	Explicative	Binaire	Non		Non	.	.
VALUE	Explicative	Continue	Non		Non	.	.
YOJ	Explicative	Continue	Non		Non	.	.

Sélectionner toutes les variables
Explorer

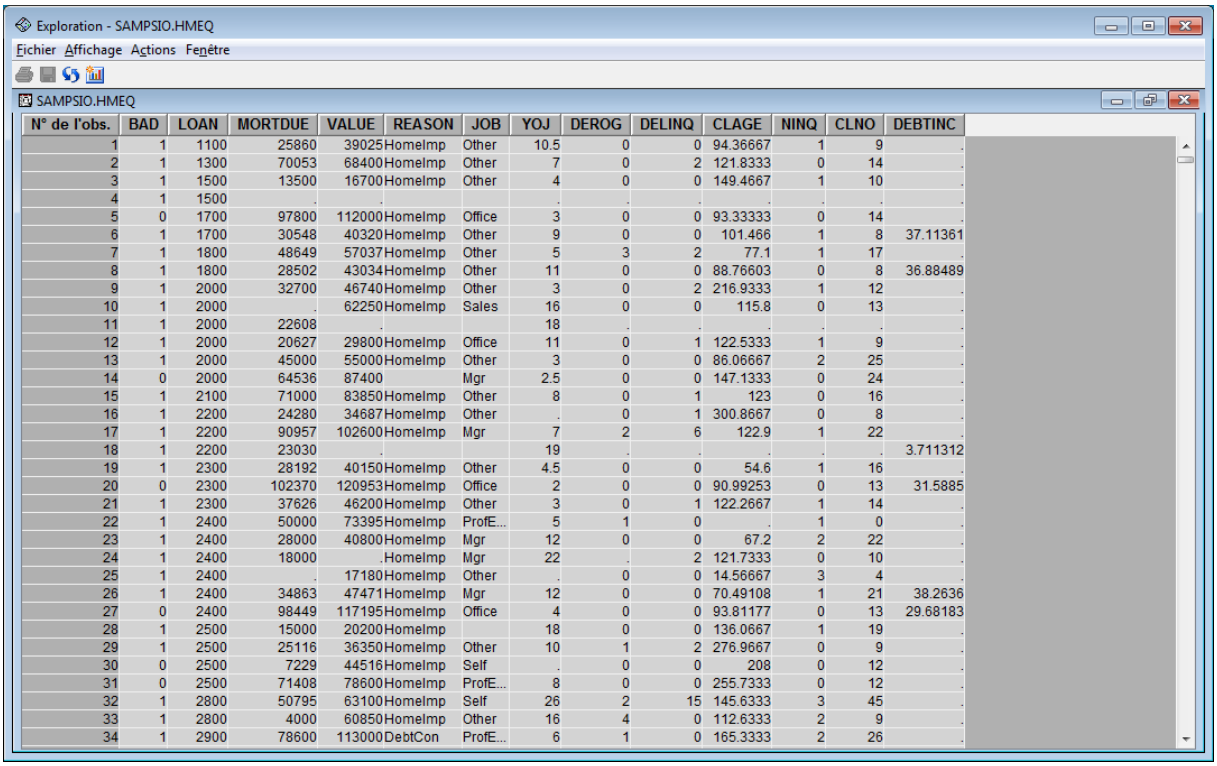
Prenons le diagramme de la variable Debtinc,



On remarque qu'il y a beaucoup de personnes (1267), pour lesquelles la variable du taux d'endettement n'est pas renseignée.

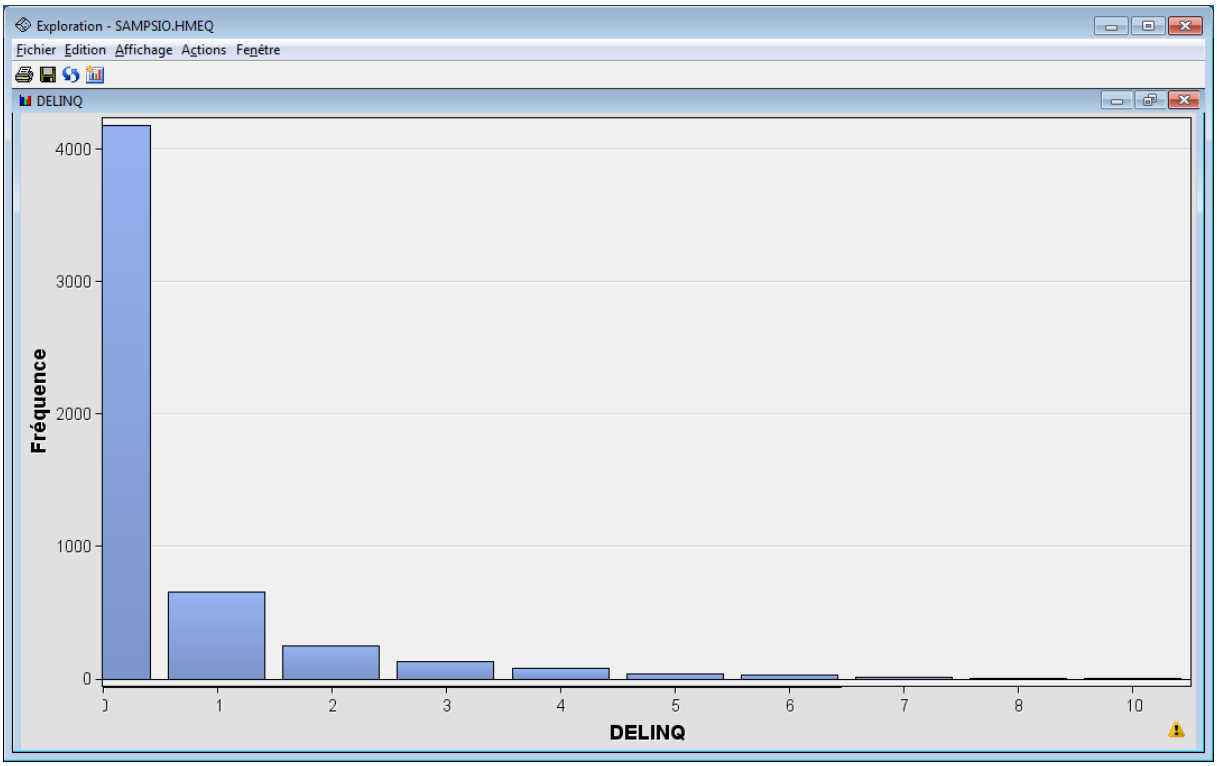


Réduire la fenêtre et agrandir la fenêtre Sampsio.HMEQ

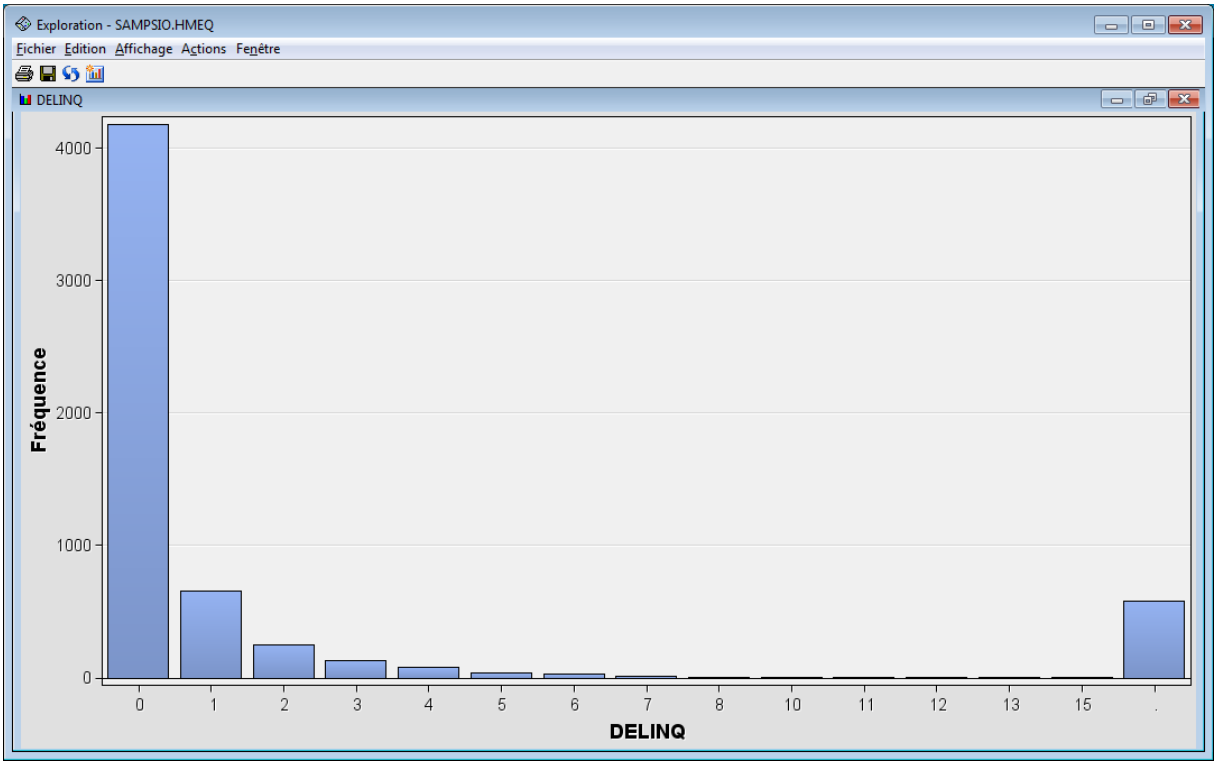


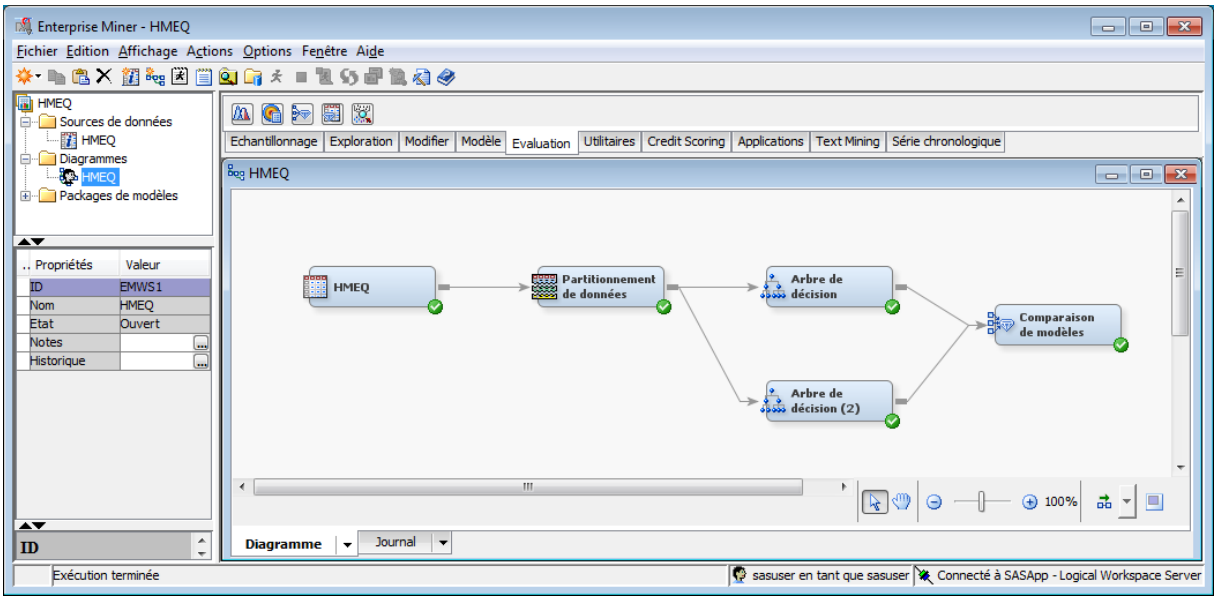
N° de l'obs.	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
1	1	1100	25860	39025	HomeImp	Other	10.5	0	0	94.36667	1	9	.
2	1	1300	70053	68400	HomeImp	Other	7	0	2	121.8333	0	14	.
3	1	1500	13500	16700	HomeImp	Other	4	0	0	149.4667	1	10	.
4	1	1500	.	.	.	.	.	.	.	.	.	.	.
5	0	1700	97800	112000	HomeImp	Office	3	0	0	93.33333	0	14	.
6	1	1700	30548	40320	HomeImp	Other	9	0	0	101.466	1	8	37.11361
7	1	1800	48649	57037	HomeImp	Other	5	3	2	77.1	1	17	.
8	1	1800	28502	43034	HomeImp	Other	11	0	0	88.76603	0	8	36.88489
9	1	2000	32700	46740	HomeImp	Other	3	0	2	216.9333	1	12	.
10	1	2000	.	62250	HomeImp	Sales	16	0	0	115.8	0	13	.
11	1	2000	22608	.	.	.	18	.	.	.	.	.	.
12	1	2000	20627	29800	HomeImp	Office	11	0	1	122.5333	1	9	.
13	1	2000	45000	55000	HomeImp	Other	3	0	0	86.06667	2	25	.
14	0	2000	64536	87400	.	Mgr	2.5	0	0	147.1333	0	24	.
15	1	2100	71000	83850	HomeImp	Other	8	0	1	123	0	16	.
16	1	2200	24280	34687	HomeImp	Other	.	0	1	300.8667	0	8	.
17	1	2200	90957	102600	HomeImp	Mgr	7	2	6	122.9	1	22	.
18	1	2200	23030	.	.	.	19	.	.	.	.	.	3.711312
19	1	2300	28192	40150	HomeImp	Other	4.5	0	0	54.6	1	16	.
20	0	2300	102370	120953	HomeImp	Office	2	0	0	90.99253	0	13	31.5885
21	1	2300	37626	46200	HomeImp	Other	3	0	1	122.2667	1	14	.
22	1	2400	50000	73395	HomeImp	ProfE...	5	1	0	.	1	0	.
23	1	2400	28000	40800	HomeImp	Mgr	12	0	0	67.2	2	22	.
24	1	2400	18000	.	HomeImp	Mgr	22	.	2	121.7333	0	10	.
25	1	2400	.	17180	HomeImp	Other	.	0	0	14.56667	3	4	.
26	1	2400	34863	47471	HomeImp	Mgr	12	0	0	70.49108	1	21	38.2636
27	0	2400	98449	117195	HomeImp	Office	4	0	0	93.81177	0	13	29.68183
28	1	2500	15000	20200	HomeImp	.	18	0	0	136.0667	1	19	.
29	1	2500	25116	36350	HomeImp	Other	10	1	2	276.9667	0	9	.
30	0	2500	7229	44516	HomeImp	Self	.	0	0	208	0	12	.
31	0	2500	71408	78600	HomeImp	ProfE...	8	0	0	255.7333	0	12	.
32	1	2800	50795	63100	HomeImp	Self	26	2	15	145.6333	3	45	.
33	1	2800	4000	60850	HomeImp	Other	16	4	0	112.6333	2	9	.
34	1	2900	78600	113000	DebtCon	ProfE...	6	1	0	165.3333	2	26	.

On remarque que notre table est un vrai Emmental, il y a beaucoup de valeurs manquantes, non renseignées, des trous.

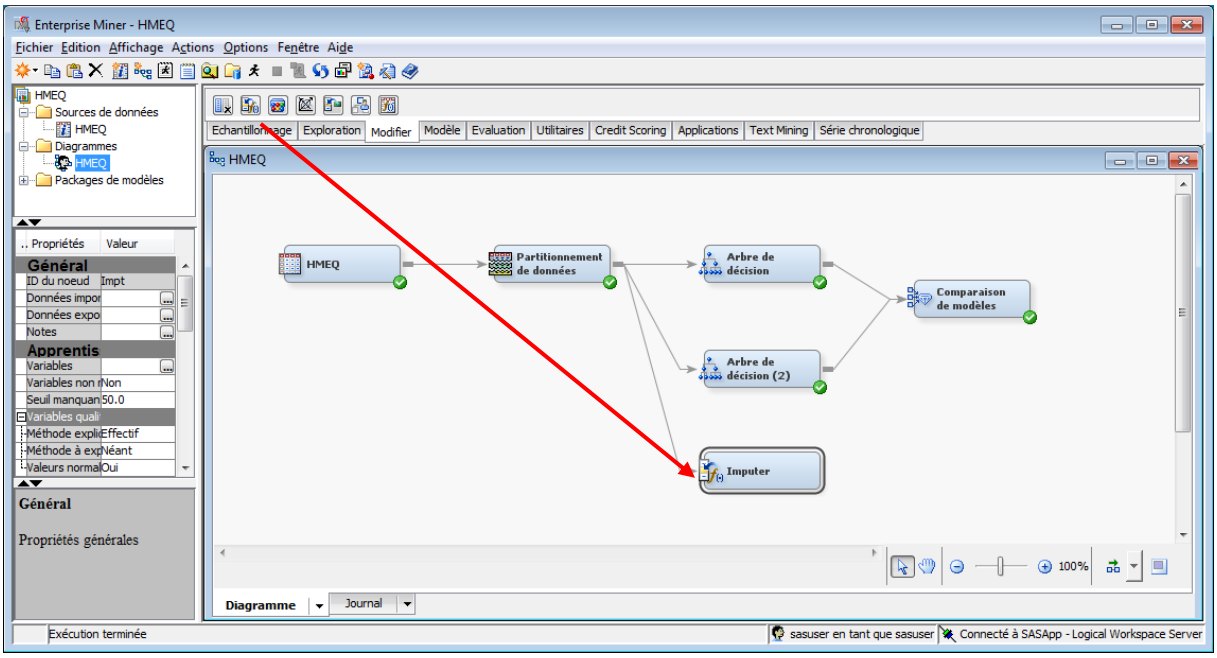


Si on regarde la variable Delinc (nombre d'articles litigieux), il suffit de cliquer sur le petit triangle « Attention » pour afficher toutes les modalités.

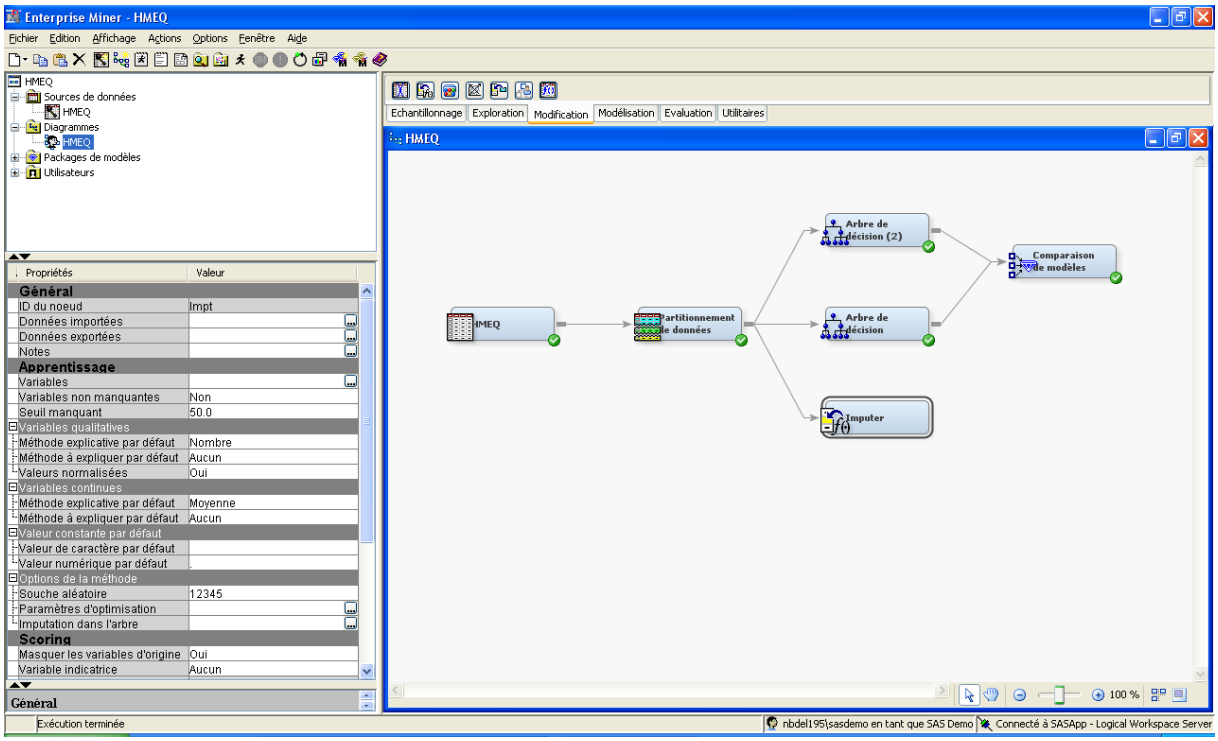




Revenons au processus,
Dans l'onglet modification, prendre l'outil d'imputation et le placer après l'outil de partitionnement des données comme ci-dessous.

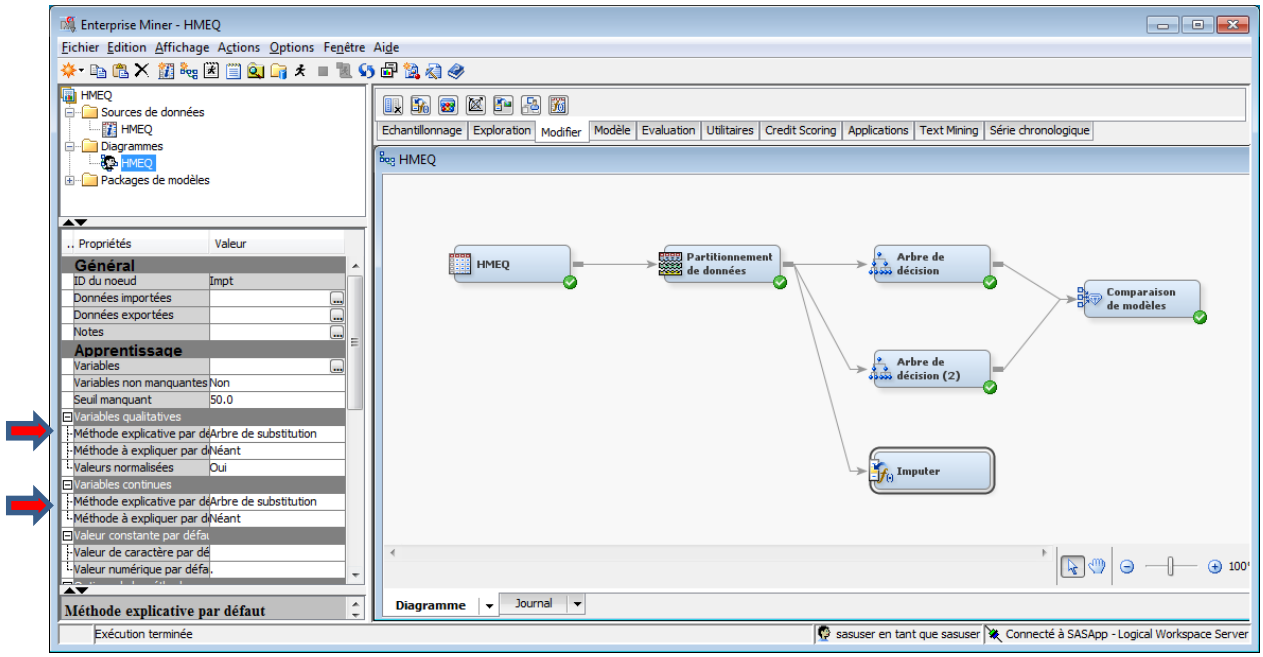


Sélectionner **Imputer** :



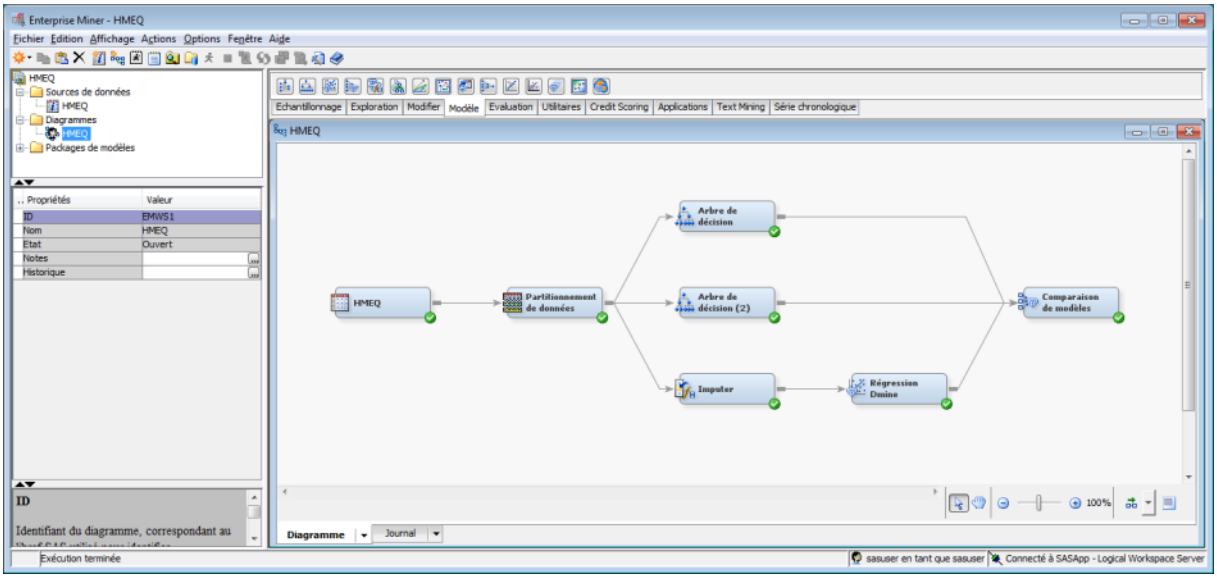
Pour les variables quantitatives, la méthode par défaut est moyenne c'est-à-dire que s'il y a une valeur non renseignée, elle est remplacée par la moyenne de toutes celles renseignées. On peut aussi sélectionner la médiane.
La méthode basée sur la distribution remplace toutes les valeurs manquantes de telle sorte qu'elles suivent la distribution de base.
La méthode de l'arbre de substitution est un peu plus sophistiquée. L'idée est de construire un arbre pour la variable où l'on souhaite remplacer les valeurs manquantes sur toutes les lignes où l'on connaît l'information, et de l'appliquer sur toutes celles où l'on souhaite remplacer la valeur manquante. Étant donné que l'arbre de décision est un modèle supportant les valeurs manquantes, cela permet de remplacer les valeurs manquantes d'une façon plus juste, à savoir, en utilisant les autres variables d'entrée renseignées.
On peut aussi utiliser une valeur par défaut. Pour la valeur par défaut, dans l'onglet valeur constante par défaut, puis dans variable numérique par défaut, définir cette valeur. Dans le cas où il y a trop de valeurs manquantes, l'outil de filtrage permet de supprimer les variables ayant trop de valeurs manquantes par rapport à un seuil.

Dans les variables qualitatives, on ne peut pas calculer la moyenne, la méthode par défaut est de remplacer les valeurs manquantes par la valeur la plus fréquemment rencontrée (le mode). On peut aussi utiliser les méthodes de remplacement par arbre de décision, ou bien définir une valeur constante.
Fermer l'outil de remplacement des valeurs manquantes.



Dans les propriétés, dans la partie des variables qualitatives et dans celle des variables continues, sélectionner la méthode explicative par défaut de l'arbre de substitution.

Régression



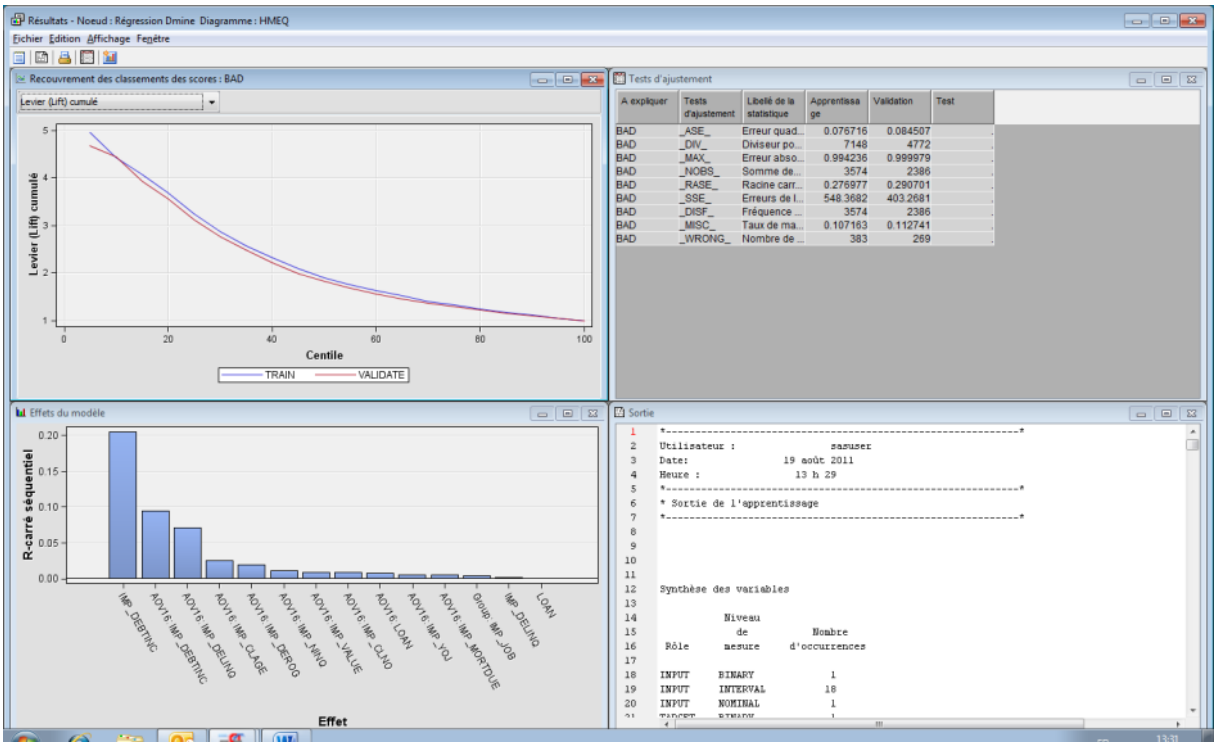
Ajouter la régression.

La régression est un modèle mathématique très utilisé.

La régression linéaire simple est une équation du type Y = a * X + b ou a et b sont des valeurs à déterminer.
La généralisation à p variables s'appelle la régression linéaire multiple.
Y = a0 + a1•X1 + a2•X2 + … + ap•Xp

Dans notre cas, la variable à expliquer est binaire, nous utiliserons donc une régression logistique généralisée.
La probabilité que la variable à expliquer soit égale à 1, sera fonction des autres variables.

Exécuter la régression.



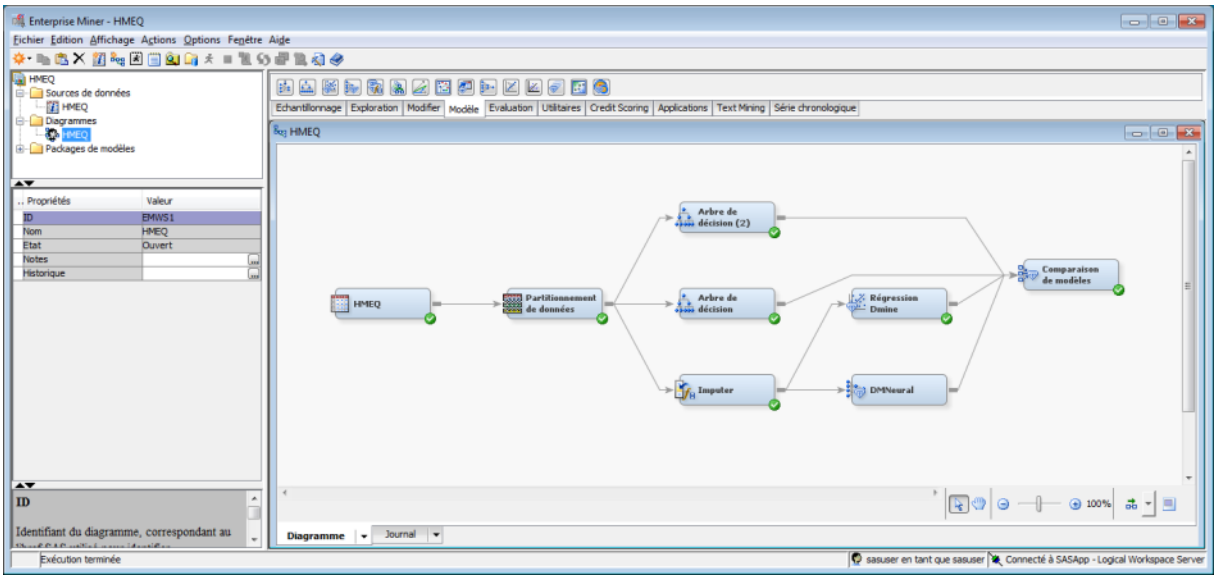
Dans les résultats de la régression, dans le graphique d'effets, on a les paramètres de la régression triés de manière décroissante selon leur effet.

L' « effect» nous donne l'importance de la variable. Il faut le lire en valeur absolue.

Nous avons ici les paramètres de notre régression logistique.
La odds ratio est égale à « Intercept », plus la somme des variables d'entrée, pondérées par les coefficients de la régression (Parameter Estimate)

$\log(\frac{\hat{p}}{1-\hat{p}})$= Intercept +Σ Parameter * Input

Réseau de neurones



Ajouter le réseau de neurones.

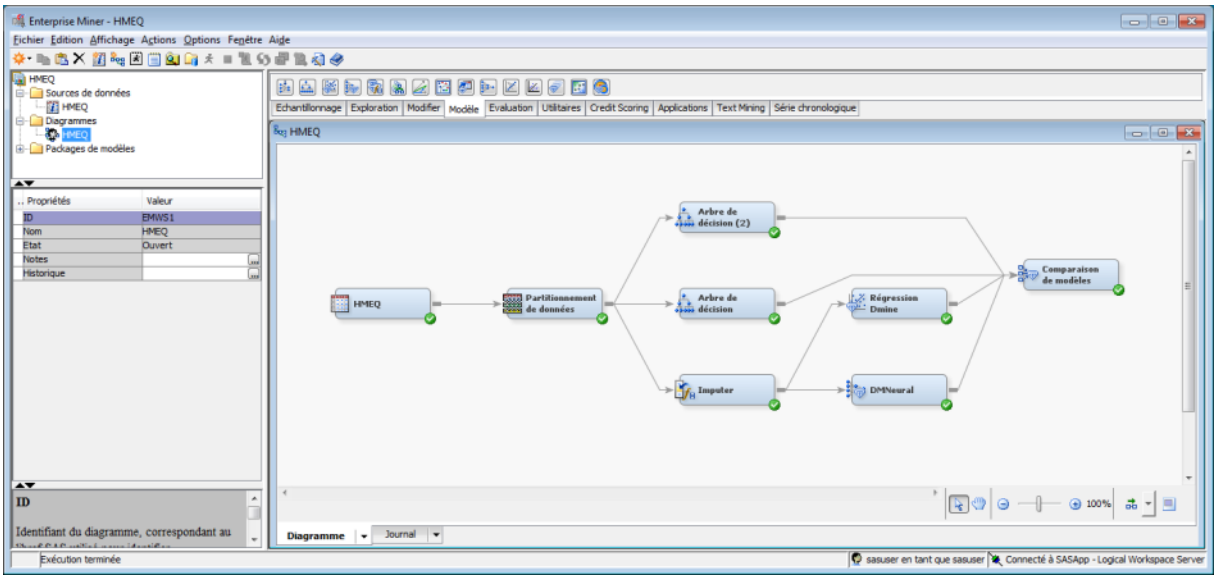
Un réseau de neurone est quasiment une boite noire. Par contre, ils sont très utiles pour modéliser des comportements atypiques.

Résumé sur les modèles

Modèle			
Interprétation	+++ Facile à interpréter. Très apprécié en marketing. On voit vite les variables importantes	+ Interprétation : on connaît l'importance des variables, on a une artillerie statistique importante pour l'utiliser.	--- « Boite noire » On ne peut presque pas les interpréter.
Caractéristiques principales	Trop carré : Par exemple les moins de 35.2 ans sont risqués et les plus de 35.2 ans ne sont pas risqués.	La régression permet : plus le client est âgé, moins il est risqué	compliqué
Utilisation majeur	Très utile pour définir simplement de grand groupe, très lisible	Très utile pour modéliser des « choses » normales.	Très utile pour modéliser des « choses » atypiques.
Gestion des valeurs manquantes	L'arbre gère les valeurs manquantes comme une modalité	Les individus avec des valeurs manquantes ne sont pas pris en compte	Les individus avec des valeurs manquantes ne sont pas pris en compte

Globalement, On ne peut pas dire qu'un modèle est meilleur que les autres. La méthode empirique est de les tester et de sélectionner le meilleur avec « assessement » pour le cas sur lequel on est.

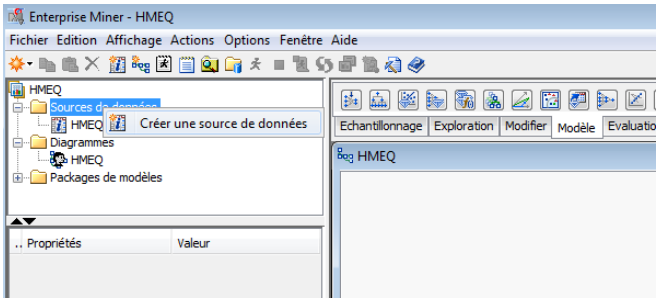
Scoring



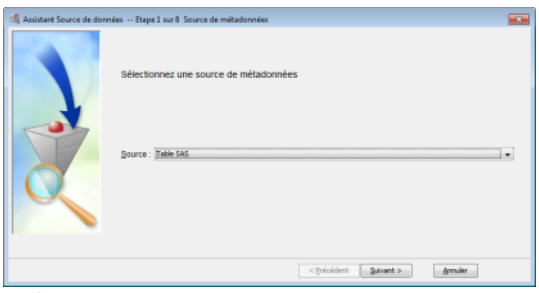
Application du meilleur modèle sur une table à scorer.

En partant du diagramme suivant, si l'on souhaite appliquer notre score sur une autre base de données, une base de données où l'on ne connaît pas la réponse, voici les étapes à suivre.
Pour plus de simplicité, je vous propose d'utiliser comme table à scorer, la table HMEQ de la bibliothèque SAMPSIO. Cette table contient la colonne 'target' : BAD, mais cette colonne ne sera pas utilisée.
Pour scorer une table, il faut absolument que toutes les variables qui avaient pour rôle 'input' dans la phase de construction du modèle, se retrouvent dans cette table, avec mêmes noms de colonne.

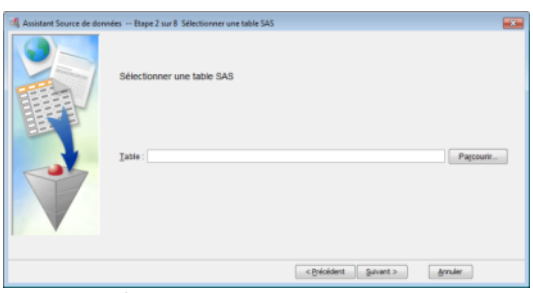
Dans notre cas, nous souhaitons appliquer notre modèle sur la base SAMPSIO.HMEQ. Dans les sources de données, il faut créer une table qui aura pour rôle d'être à scorer.



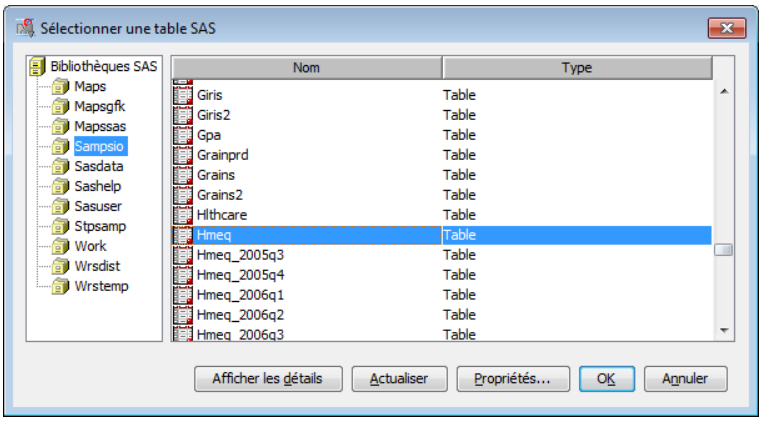
Clic-droit sur sources de données → créer une source de données



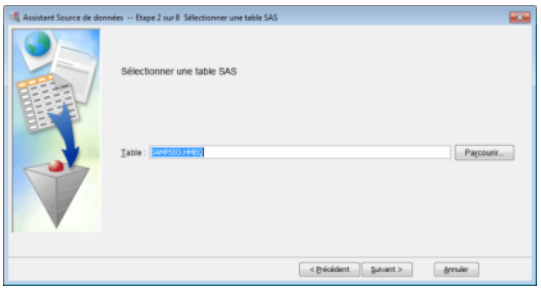
Suivant



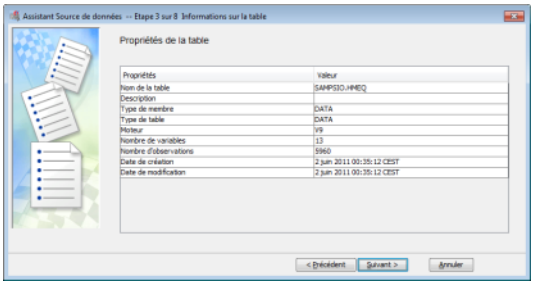
Parcourir



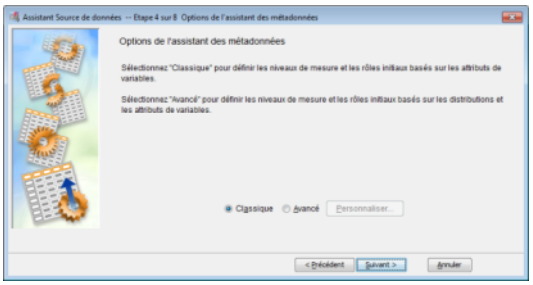
Sélectionner la table HMEQ de la bibliothèque Sampsio
OK



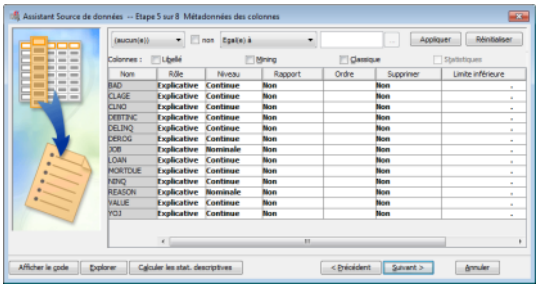
Suivant



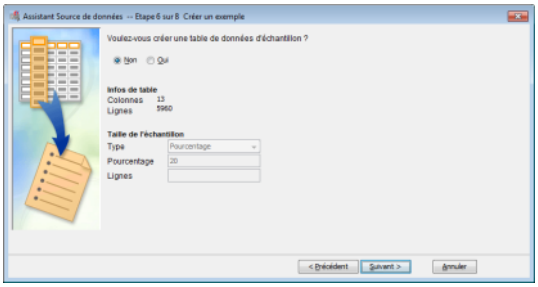
Suivant



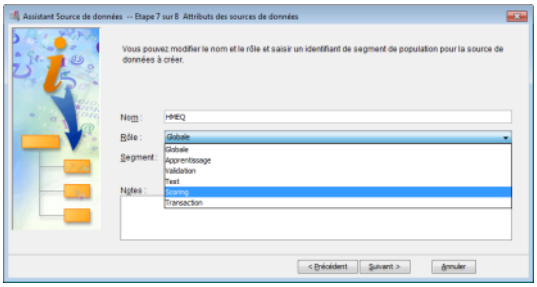
Suivant



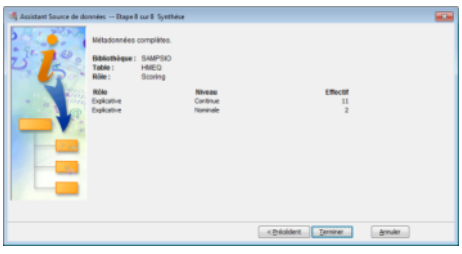
Suivant



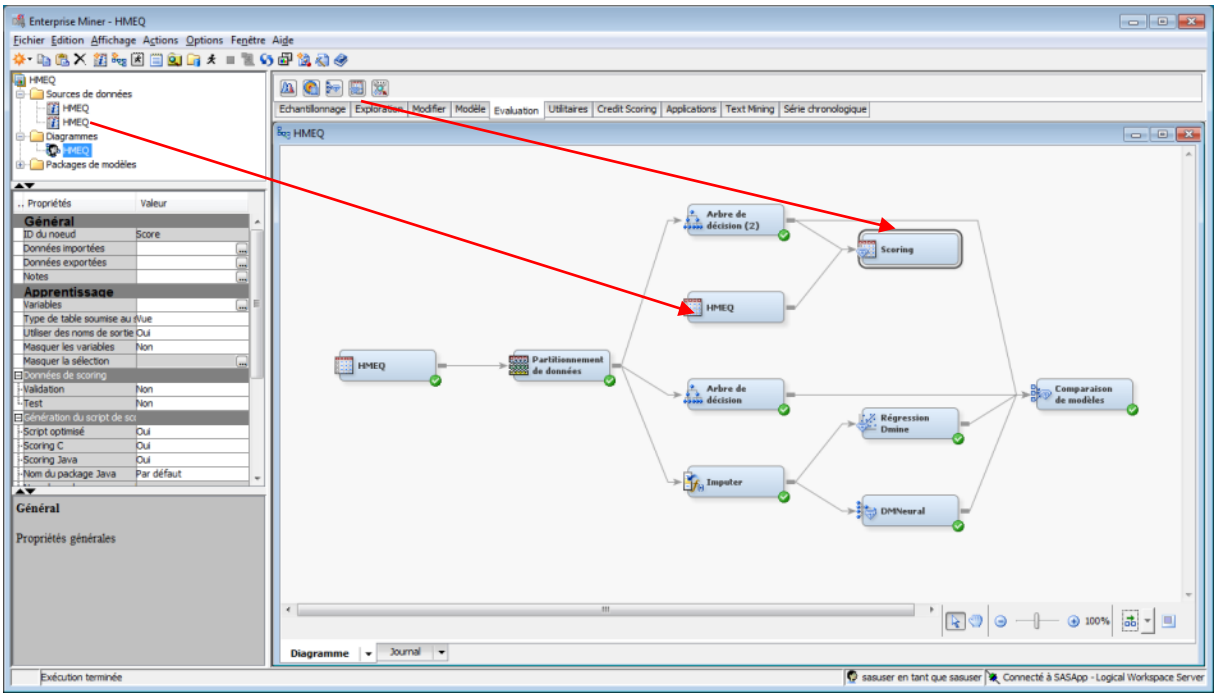
Suivant



Sélectionner le rôle Scoring

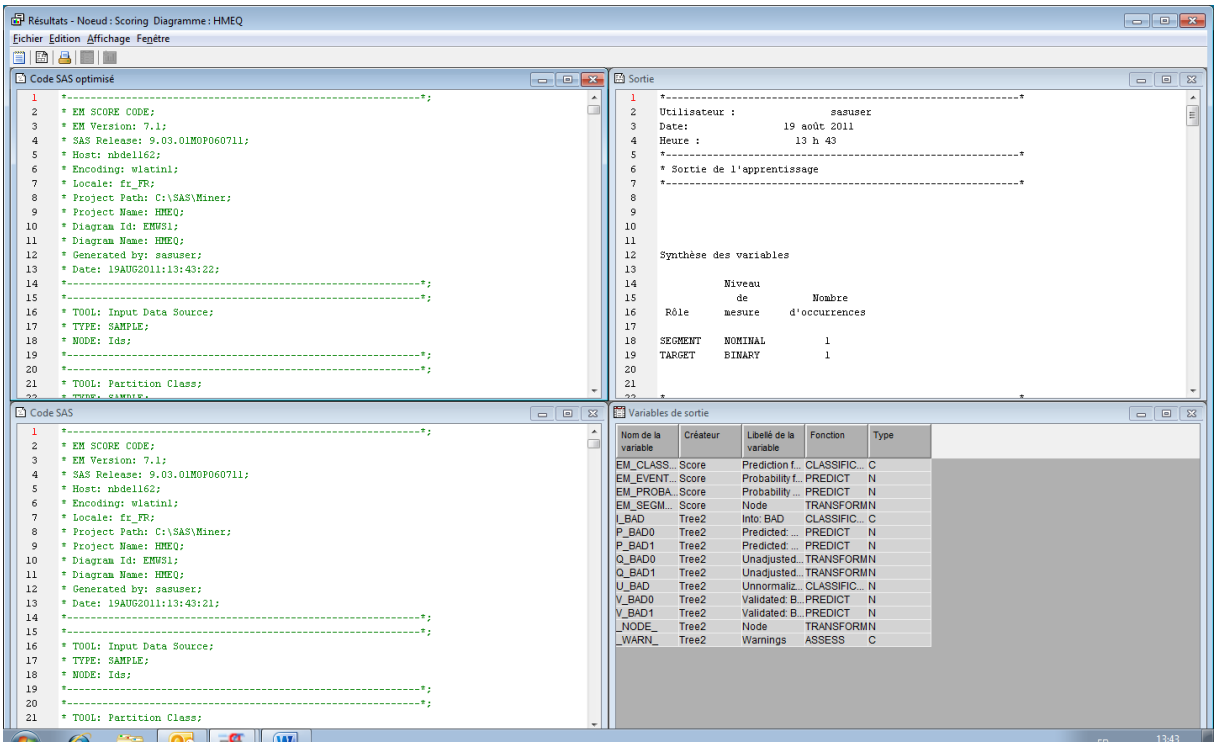


Terminer

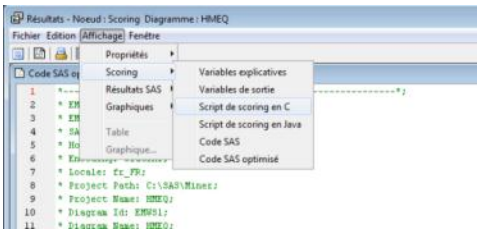


Ajouter l'outil de scoring de l'onglet Evaluation et la nouvelle table HMEQ des sources de données. Tracer les flèches comme si dessus.
En entrée de l'outil de scoring, on a un lien depuis notre meilleur modèle et un depuis notre table à scorer.

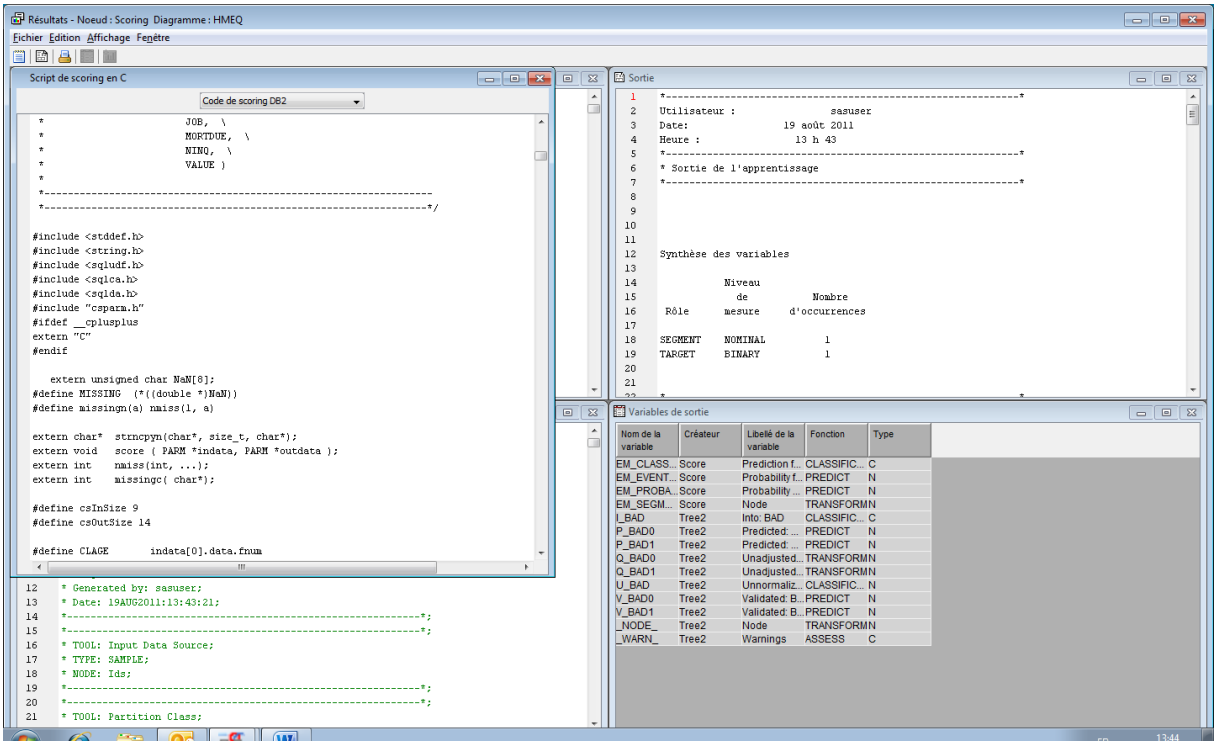
Exécuter le score



Dans les résultats du score, on obtient le code SAS



Si l'on souhaite le code en C ou en Java, dans le menu Affichage → Scoring → Script de scoring en C ou en Java

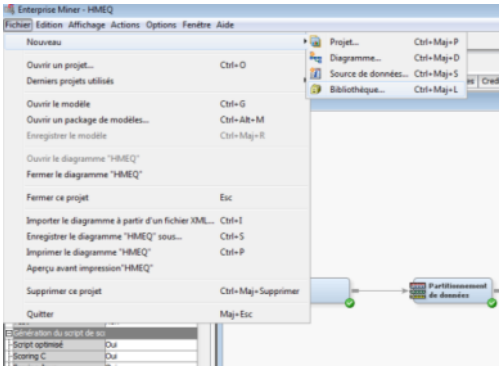


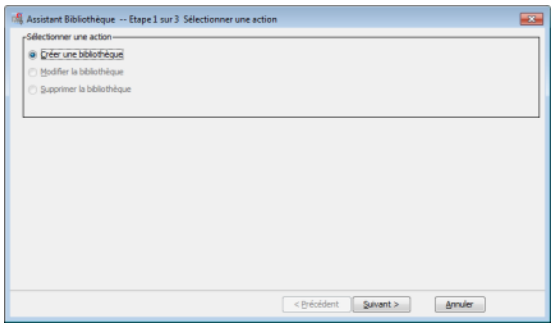
Sélectionner le Script de scoring

Création d'une bibliothèque dans Enterprise Miner :

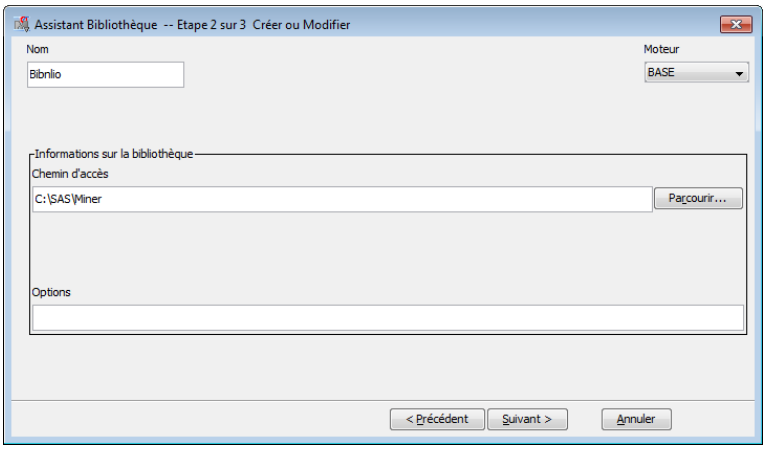
Une bibliothèque SAS est notamment un raccourci vers un dossier où se trouvent des données, des fichiers, ou vers une base de données.

Pour créer une bibliothèque pointant vers un répertoire Windows où se trouvent des fichiers de données SAS, aller dans **Fichier → Nouveau → Bibliothèque**



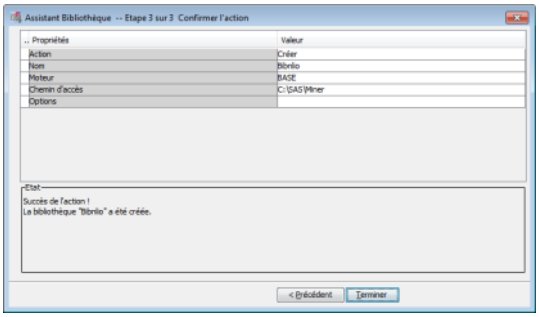


Suivant



Donner un nom de maximum 8 caractères, composé uniquement de lettres de l'alphabet, de chiffres ou du souligné ; pas de caractères spéciaux et ne commençant pas par un chiffre.

Entrer le chemin vers le dossier où se trouve les tables SAS, à l'aide du bouton **Parcourir** si besoin.



Terminer

TP Manipulation de données avec SAS Enterprise Guide et modélisation prédictive avec SAS Enterprise Miner

Introduction

L'objectif général de cette partie est de présenter l'apport de la modélisation pour prédire l'avenir, grâce au Data Mining. Généralement, un TP de Data Mining est fait avec une table toute prête. Ici, vous vous rendrez compte qu'une grande partie du travail, est de construire la base d'analyse.

La question est :

« Comment augmenter les ventes dans la prochaine période creuse ? »

L'idée est :

Modéliser le profil du client qui achète pendant une période creuse, l'appliquer sur la base de données de l'ensemble des clients et lancer une campagne marketing sur les plus forts potentiels.
Cette modélisation sera construite sur et pour les clients ayant la carte de fidélité Orion Star Gold, clients pouvant être relancés par courrier.

Public :

1) Pour faire une bonne modélisation, il faut des compétences en :
- Informatiques : créer une bonne base de données de Data Mining est un projet beaucoup plus complexe qu'il n'y parait. La modélisation grâce au Data Mining nécessite de très grosses bases de données qu'il faut alimenter par des processus souvent très complexes, stocker et faire vivre. Il est fréquent avec une table de plusieurs millions de transaction et quelques colonnes (3 ~10), de créer une table ayant une seule ligne par client et plusieurs centaines de variables (100 ~ 3000 colonnes).
- Statistiques : la modélisation Data Mining utilise massivement des méthodes statistiques et même si la loi des grands nombres facilite largement le processus, elle nécessite des compétences solides en analyse de données.
- Métier : Sans la connaissance métier, il est difficile de construire une bonne base de données permettant de faire des modèles performants. Il est primordial d'avoir de l'intuition pour faire de bonne modélisation. C'est-à-dire qu'il est nécessaire d'avoir une vraie culture fonctionnelle pour améliorer la construction des bases de données et pour l'interprétation pratique des résultats. Dans notre cas, la société fictive Orion Star, nous nous concentrerons sur le marketing. Qui mieux qu'une personne du marketing pour rappeler que pour modéliser le comportement d'un client, les variables pertinentes sont souvent récence[1], fréquence, montant, âge etc.

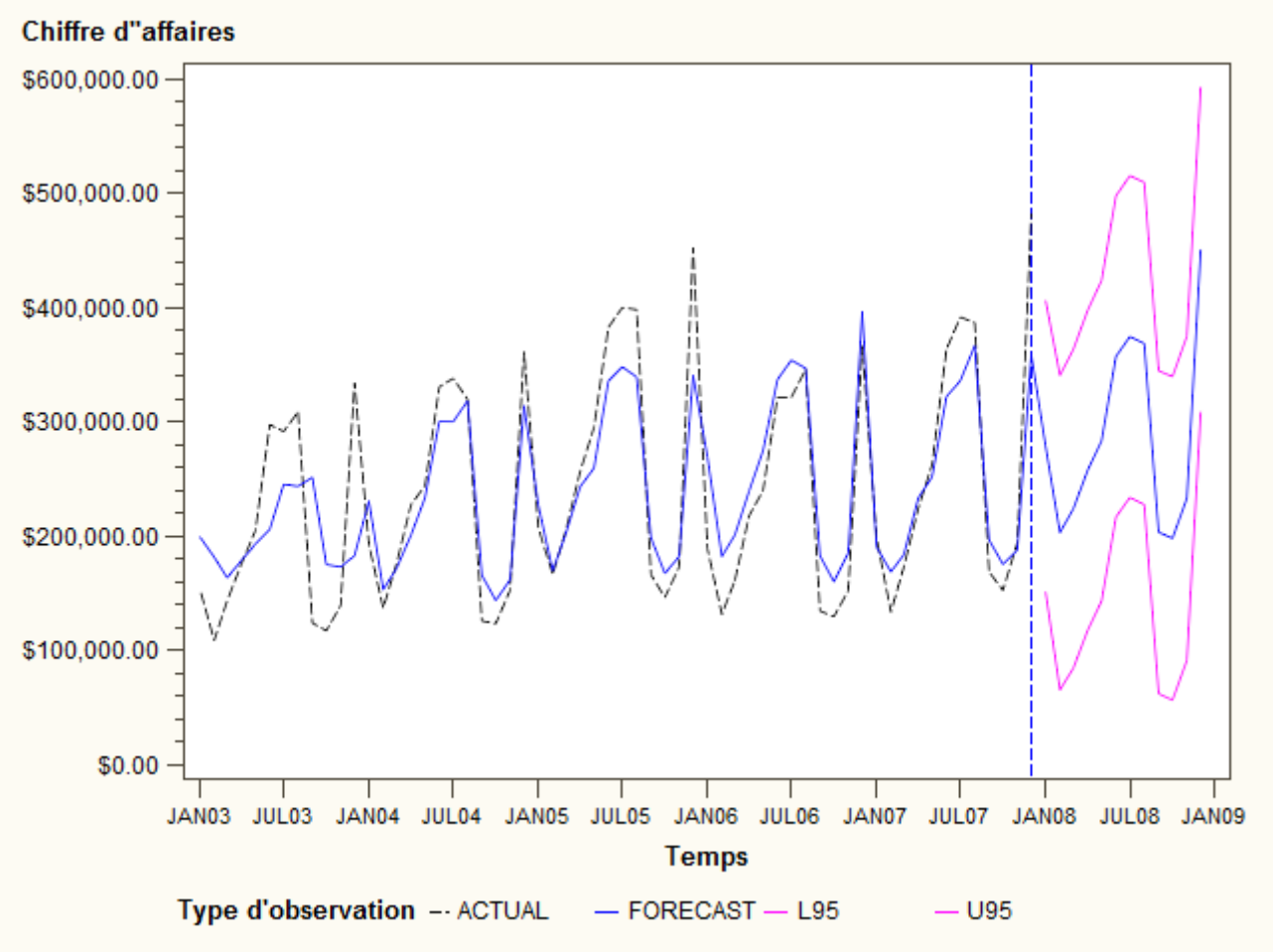
2) Étant donné qu'il est rare qu'une seule personne cumule toutes ces compétences, il est fréquent de devoir créer des synergies entre des individus aux univers différents, afin qu'ils travaillent bien ensemble. Ce chapitre s'adresse donc aux trois publics de telle sorte que chacun puisse acquérir un vocabulaire lui permettant une meilleure communication avec les deux autres parties prenantes.

Présentation du processus :

Dans le cadre du cas de la société Orion Star, nous sommes aujourd'hui le 1 janvier 2008 et l'on souhaiterait modéliser le comportement du client qui achète pendant la prochaine période creuse, c'est-à-dire entre février et mars 2008.

[1] Récence : temps écoulé depuis le dernier achat.

Globalement, l'activité de la société Orion Star est particulièrement saisonnière avec des sommets l'été et au mois de décembre. Les deux périodes creuses sont février – mars et septembre – octobre – novembre.



La courbe de prévisions des ventes ci-dessus, nous montre globalement une activité cyclique stable.

Remarquons un pic pour l'année 2006 et une petite chute en 2007 mais il n'est pas, après vérification, nécessaire d'introduire un coefficient de proportionnalité pour garder une cohérence chronologique sur l'année à venir, c'est-à-dire que sur les 5 périodes d'historique, rien ne laisse prévoir une hausse ou une baisse significative des ventes.

Positionnons-nous un an auparavant, soit au premier janvier 2007, date à laquelle nous connaissons quatre ans d'historique et les ventes sur les mois à venir de février et avril 2007. Nous pouvons alors modéliser le comportement d'un client qui achète dans la prochaine période creuse, le connaissant avec un historique de quatre ans.

Projet :

Ayant défini le projet, et connaissant l'ensemble des données disponibles, le Data Mart Orion, nous allons maintenant :

1. Créer la base de données d'apprentissage rassemblant un maximum d'information au 1er janvier 2007, par client :
    - Cible : est-ce que le client a acheté entre février et mars 2007
    - Variables décrivant le client :
        i. Sexe
        ii. Age au premier janvier 2007
        iii. Pays
        iv. Type de client
        v. Etc.
    - Variables décrivant le comportement d'achat du client du 1er janvier 2003 au 31 décembre 2006 :
        i. Quantité de produits achetés en moyenne, par commande, par période de temps, globalement, etc.
        ii. Montant des commandes, par période de temps, globalement, etc.
        iii. Propension à acheter les produits remisés,
        iv. Type de produits achetés,
        v. Période d'achat
        vi. Fréquence des achats
        vii. Récence des achats
        viii. Canal de commande
        ix. Etc.

2. Création de la base de données à scorer, par clients, au 1er janvier 2008. Cette base de données est quasiment identique à la précédente à la différence qu'il y a
    - une colonne de moins, la variable cible
    - L'âge est calculé au 1er janvier 2008
    - Et les variables décrivant le comportement d'achat du client, prennent en compte les quatre ans, du 1er janvier 2004 au 31 décembre 2007.
3. Création de modèle de Data Mining prédictif sur les variables les plus pertinentes
4. Validation, choix du modèle le plus pertinent et calcul du retour sur investissement à priori.
5. Application de ce modèle sur la base de données des clients

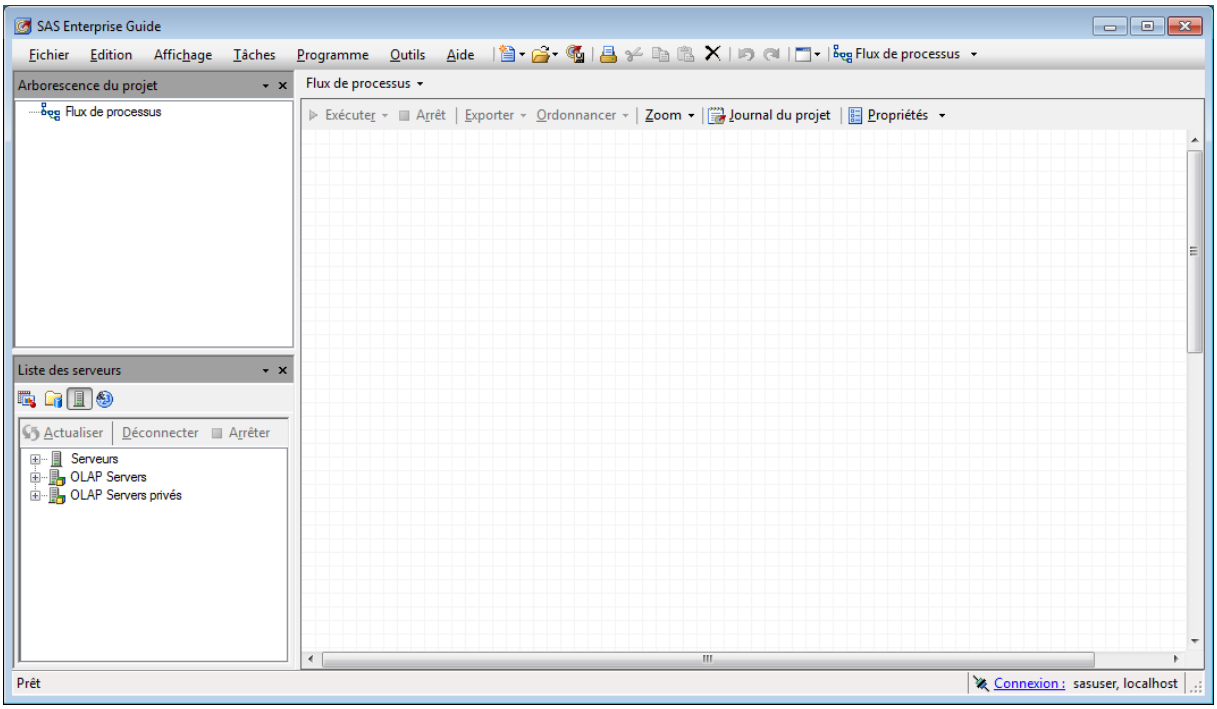
Le département « marketing opérationnel » recevra donc une base de données des clients à relancer.

Objectif du cas et RoI

Admettons que l'envoie d'une lettre coûte 2€ et que la marge nette soit de 6€ si la personne achète. Vous pouvez remarquer que selon cette hypothèse, un taux de retour de 17% coûte et un taux de retour de 33% est le seuil critique.
L'objectif de ce cas est de suivre une première fois le tutoriel afin de mieux comprendre le problème, mais surtout de tenter d'améliorer la performance du modèle. Si vous ne faites que suivre ce tutoriel, vous obtiendrais un taux de retour sur le premier décile, à peine supérieur au seuil critique.
Il faut donc trouver d'une part des variables à rajouter en plus à la table, en espérant en trouver qui soient discriminante, et d'autre part, chercher de nouveau modèle de Data Mining. Si vous cherche correctement, vous pouvez obtenir un modèle dont le taux de retour est supérieur à 45%.

Création de la base d'apprentissage

Ouvrir SAS Enterprise Guide
Depuis Démarrer → Programmes → SAS → Enterprise Guide 4.3

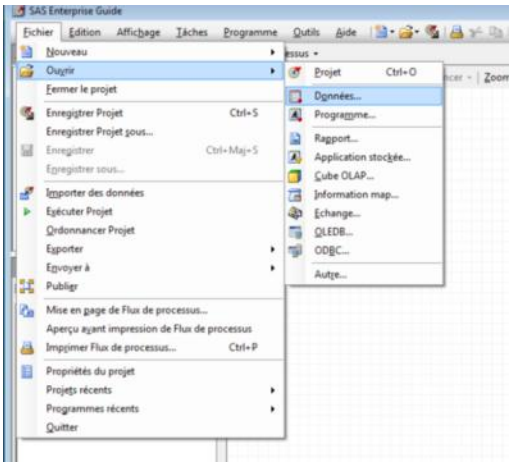


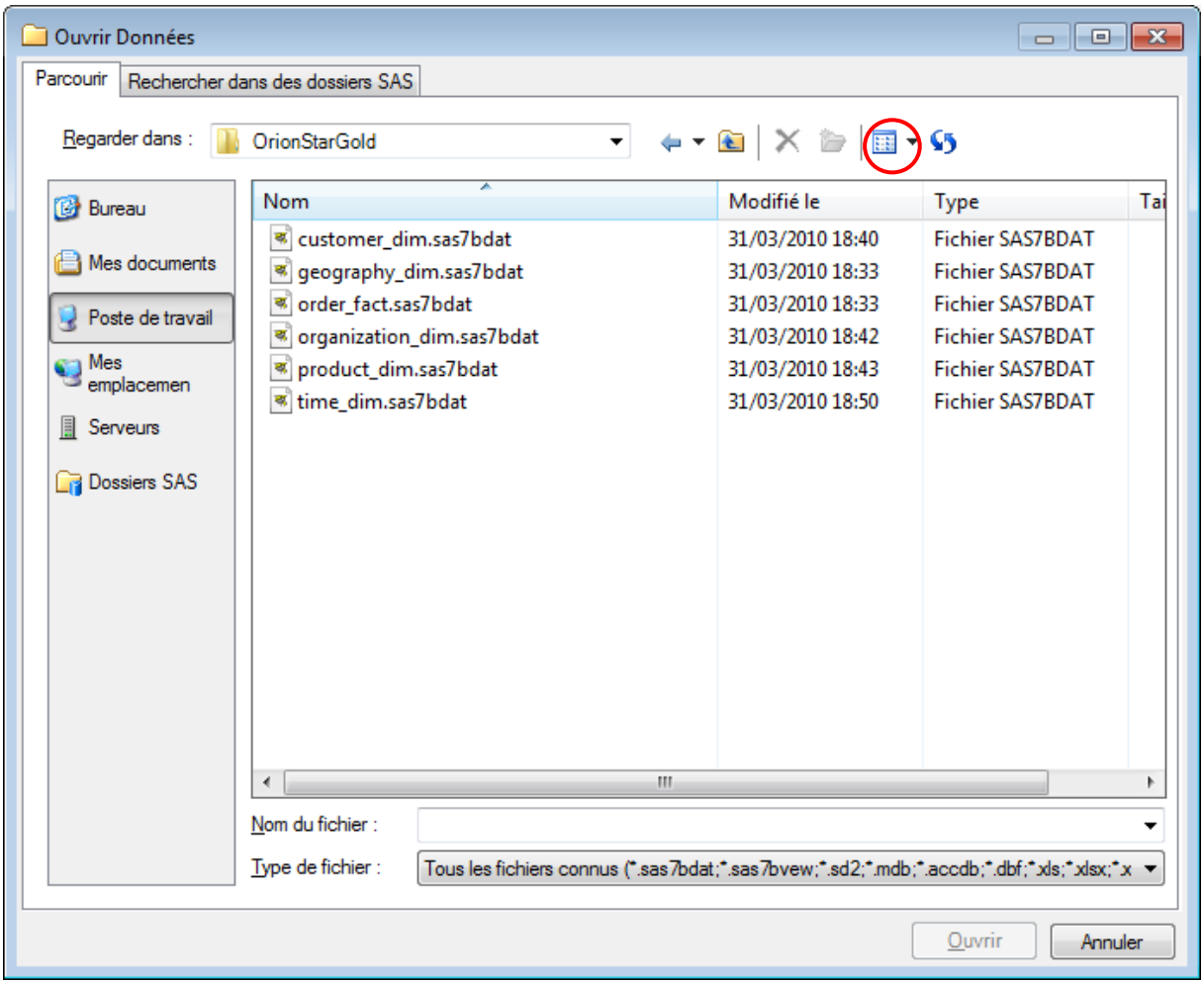
Ouvrir les tables :

- customer_dim.sas7bdat
- order_fact.sas7bdat
- product_dim.sas7bdat
- time_dim.sas7bdat

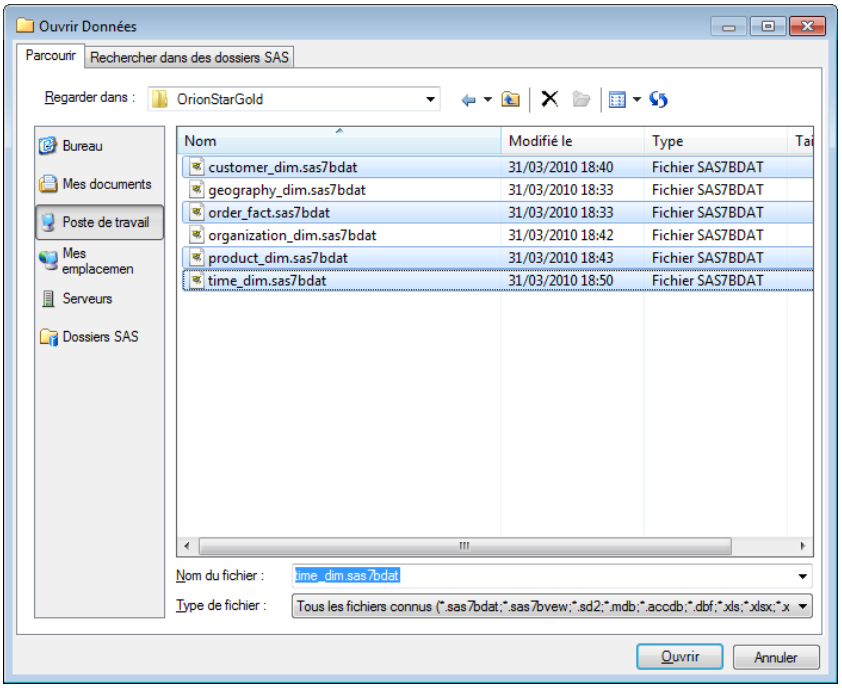
Créer la table de synthèse, jointure de ces quatre tables ci-dessus, avec un filtre sur la date (<= ‘31DEC2006’d) et sélectionner les bonnes colonnes.

Ouvrir des données :

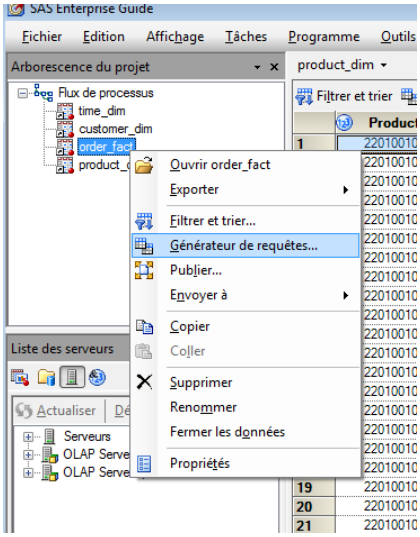




Si le nom des tables s'affiche mal, cliquer sur le bouton en haut à droite.
Depuis votre poste de travail → Dans le dossier OrionStarGold:



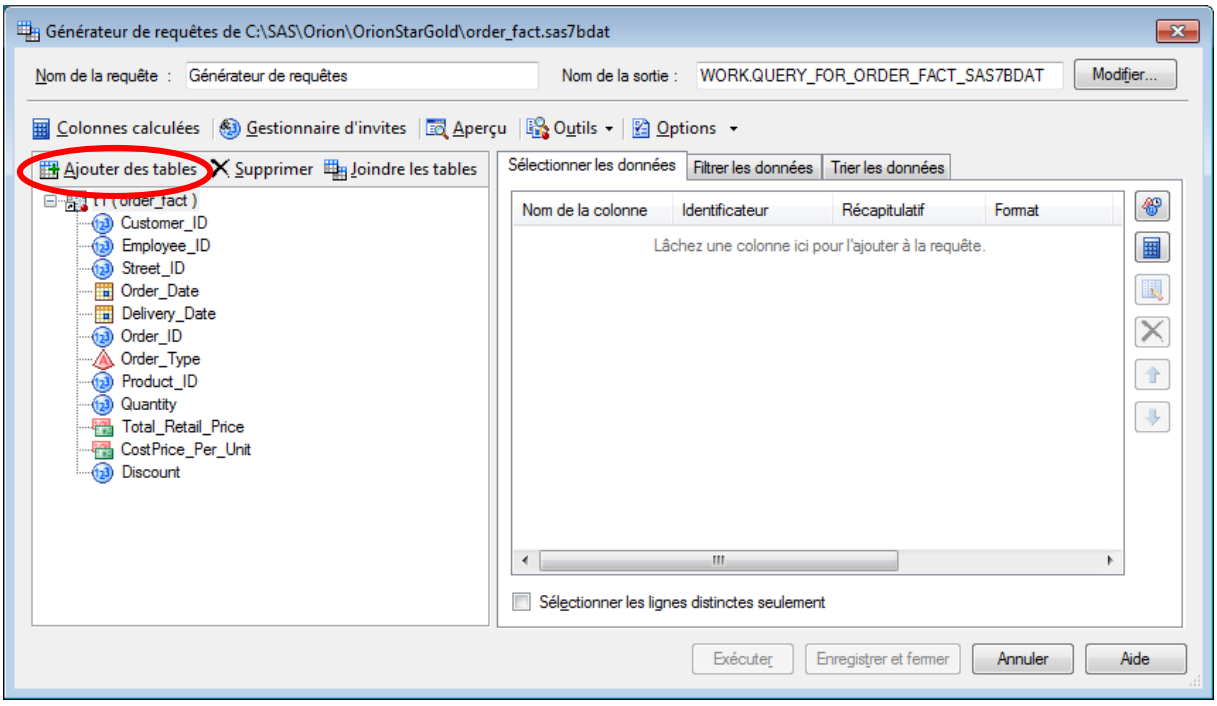
Sélectionner les tables Customer_Dim, Order_Fact, Product_Dim et Time_Dim



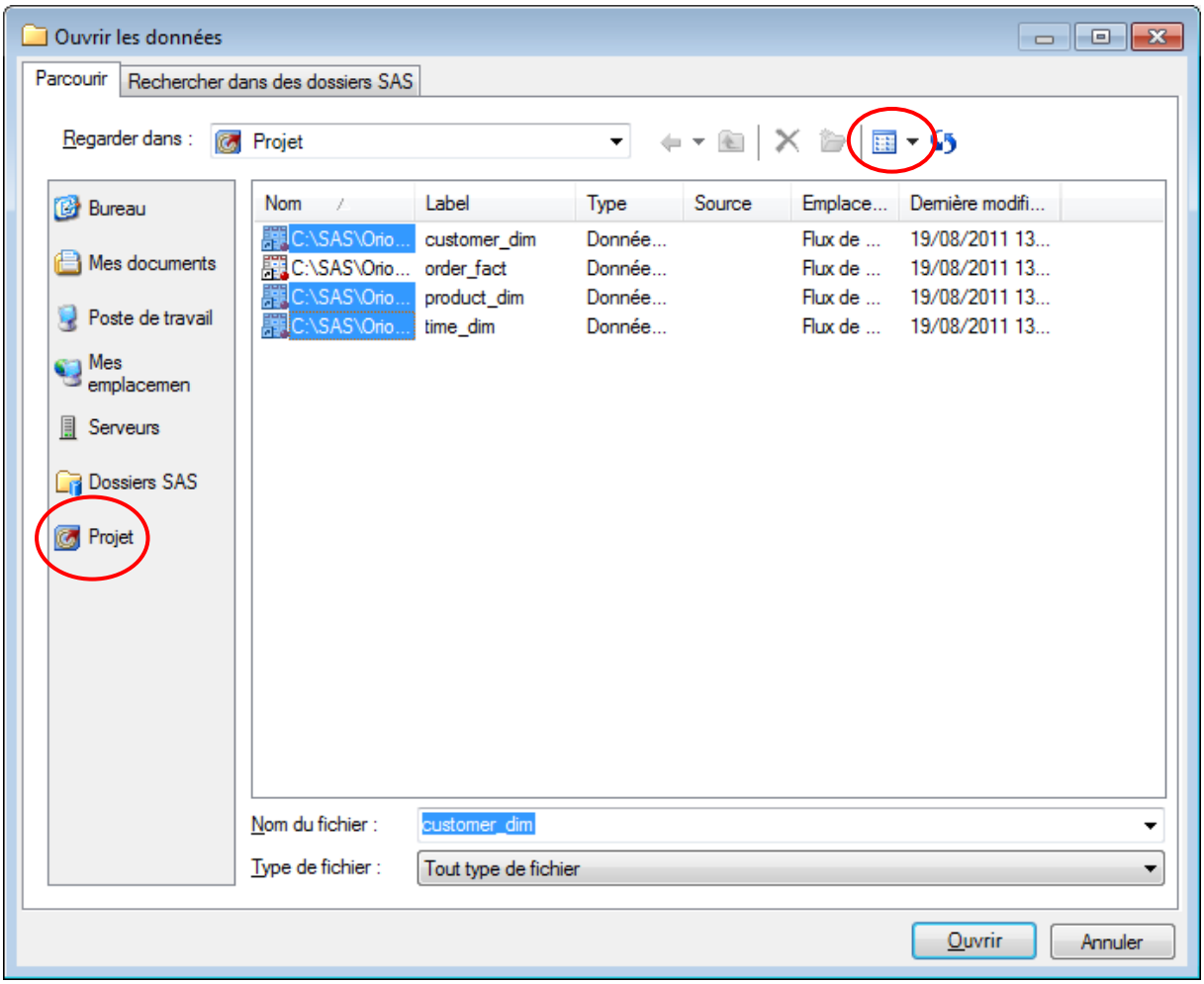
Créer une requête :

Depuis la fenêtre Flux de processus

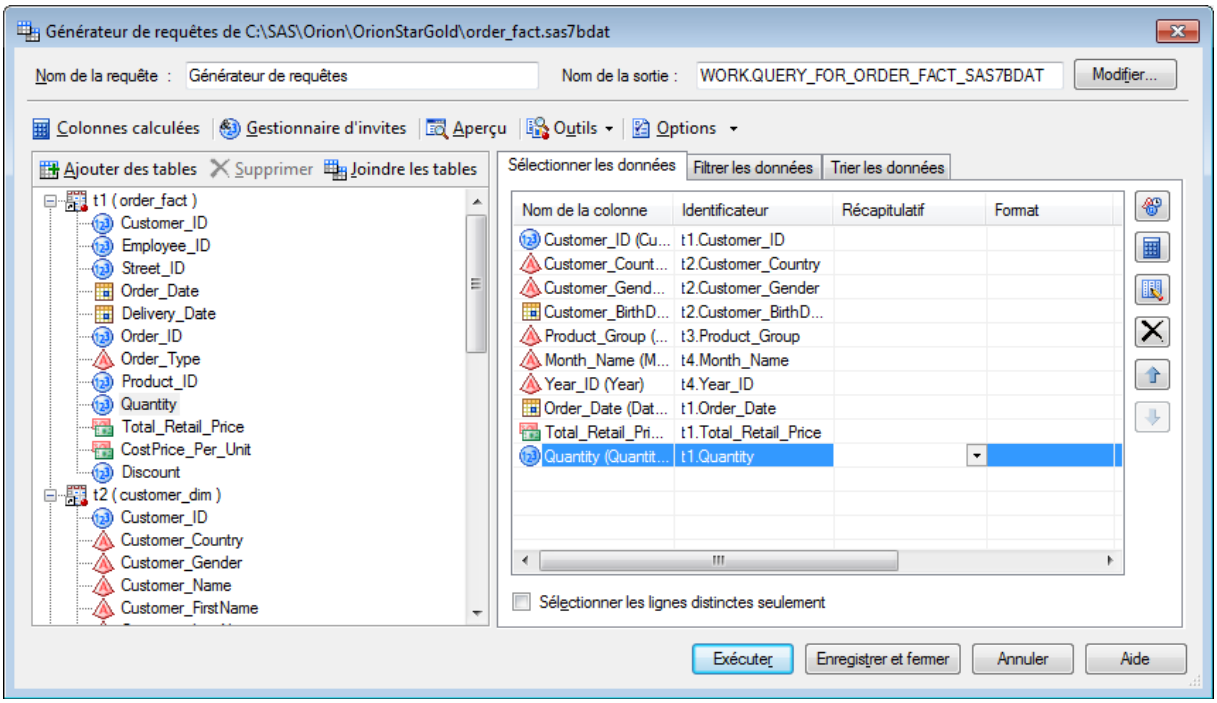
Clique droit sur une table, par exemple sur la table Order_Fact → Générateur de requêtes



Cliquer sur « Ajouter des Tables »

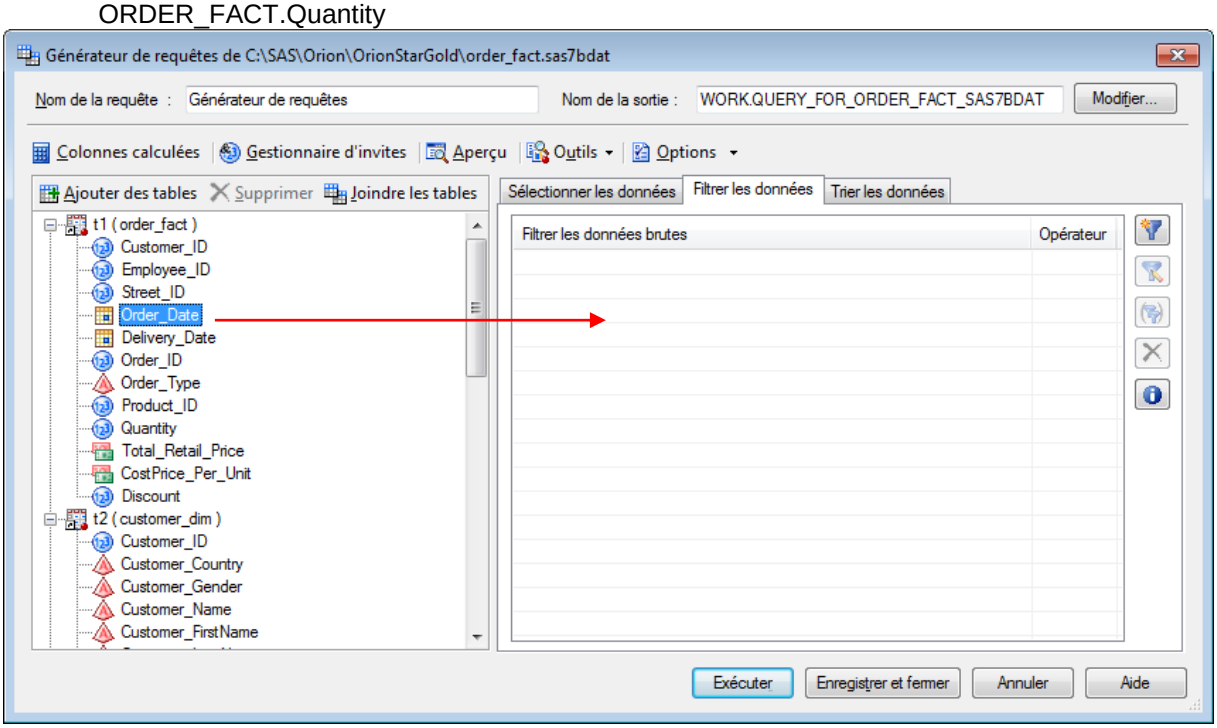


Sélectionner depuis le « Projet »
Si le nom des tables s'affiche mal, cliquer sur le bouton en haut à droite.
Sélectionner les trois autres tables. Si vous avez créé la requête depuis la table Order_Fact, il faut donc sélectionner les tables Customer_Dim, Protuct_Dim et Time_Dim.

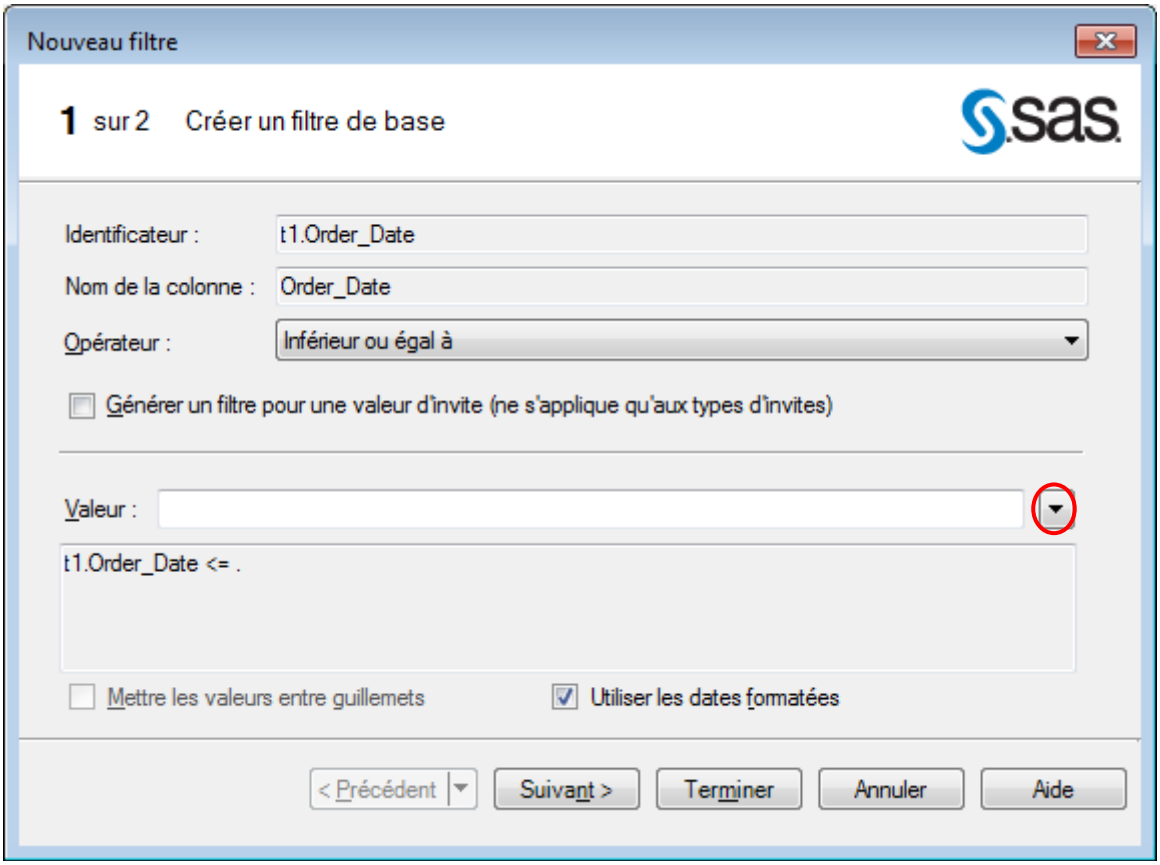


Sélectionner les colonnes comme ci-dessus :

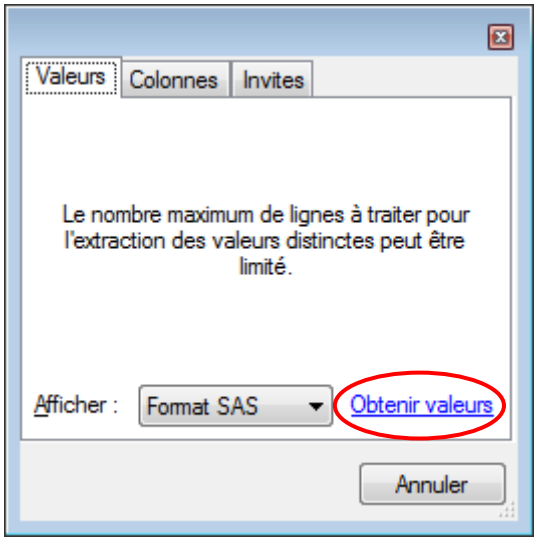
ORDER_FACT.Customer_ID
CUSTOMER_DIM.Customer_Country
CUSTOMER_DIM.Customer_Gender
CUSTOMER_DIM.Customer_BirthDate
PRODUCT_DIM.Product_Group
TIME_DIM.Month_Name
TIME_DIM.Year_ID
ORDER_FACT.Order_Date
ORDER_FACT.Total_Retail_Price
ORDER_FACT.Quantity



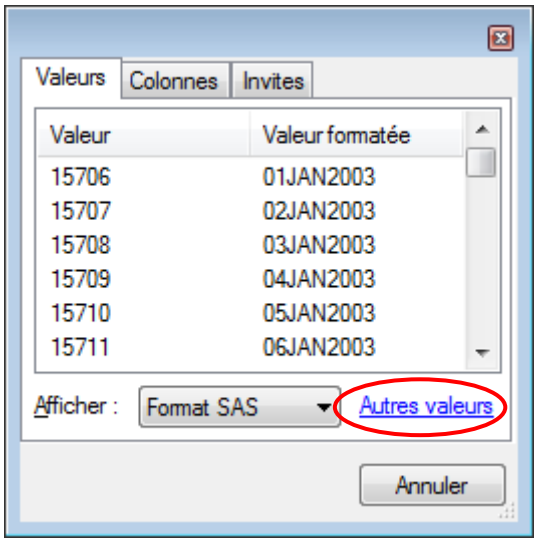
Dans l'onglet « Filtrer les données »
Glisser la colonne Order_date dans la zone de filtre.



Sélectionner l'opérateur « inférieur ou égale à »
Cliquer sur le bouton à droite de valeur

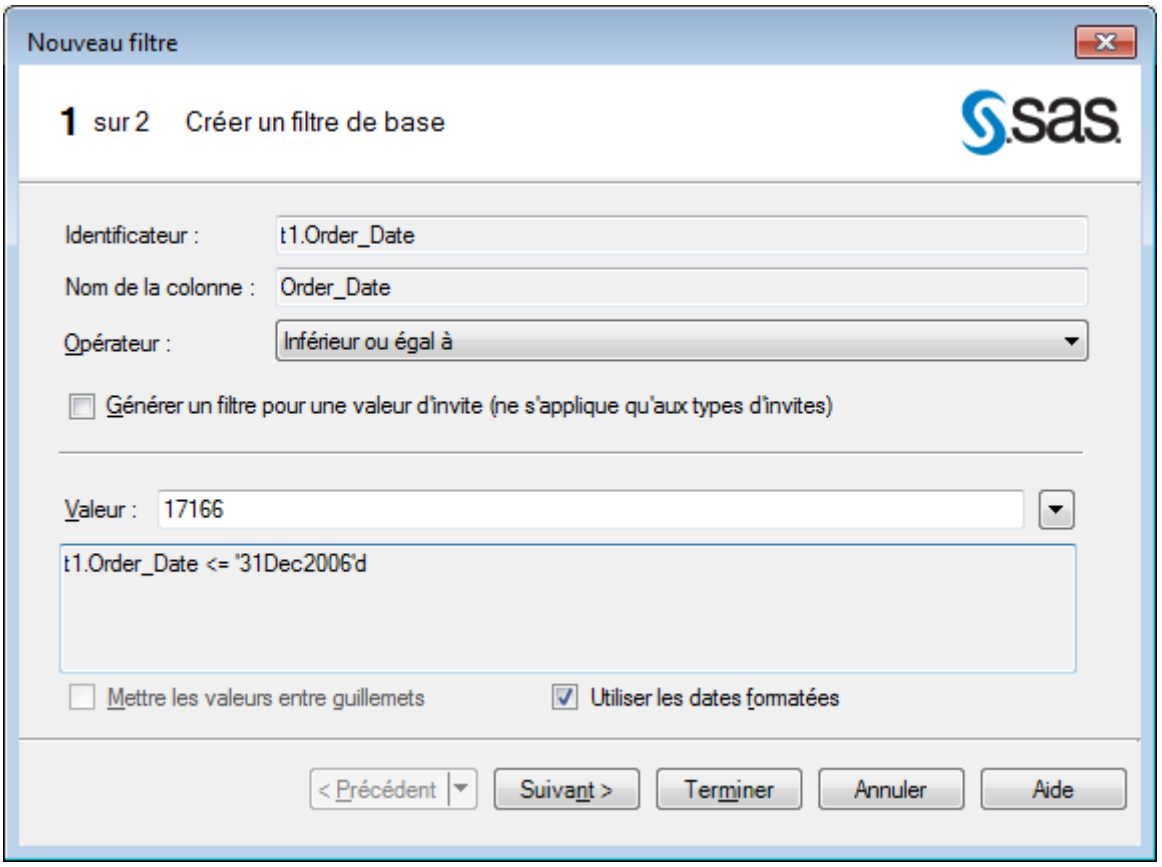


Cliquer sur « Obtenir valeur »



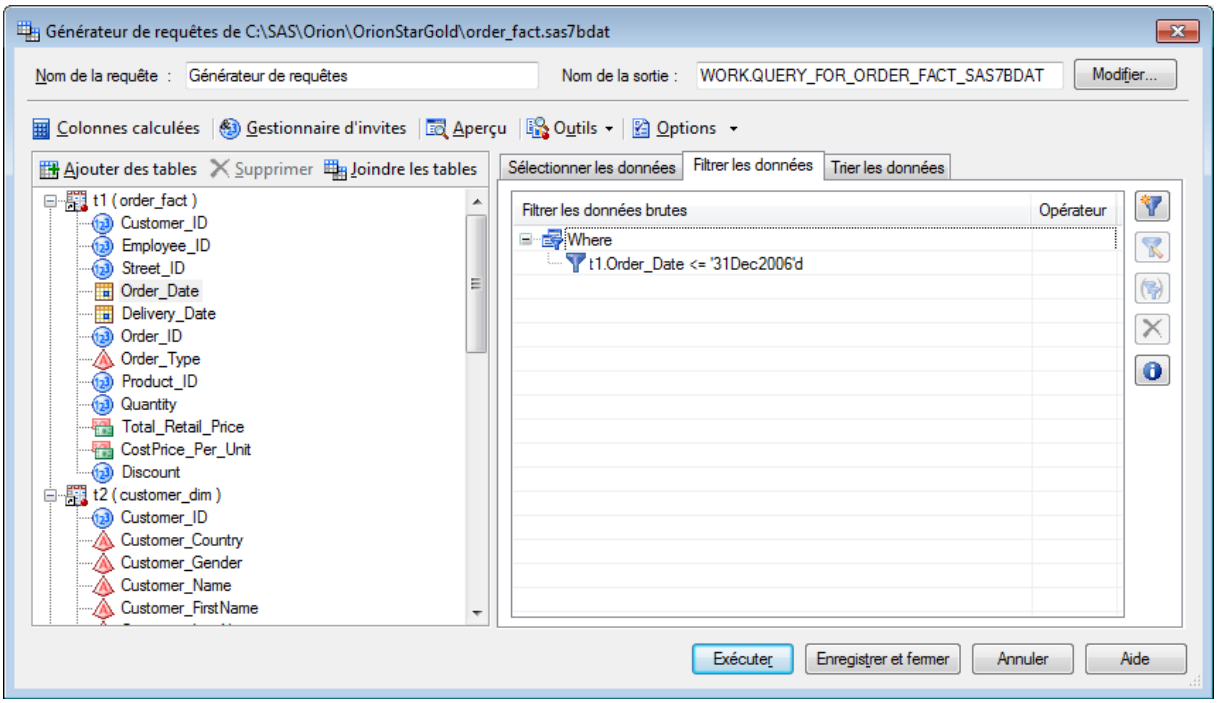
Cliquer plusieurs fois sur Autres Valeurs

Sélectionner le 31 décembre 2006

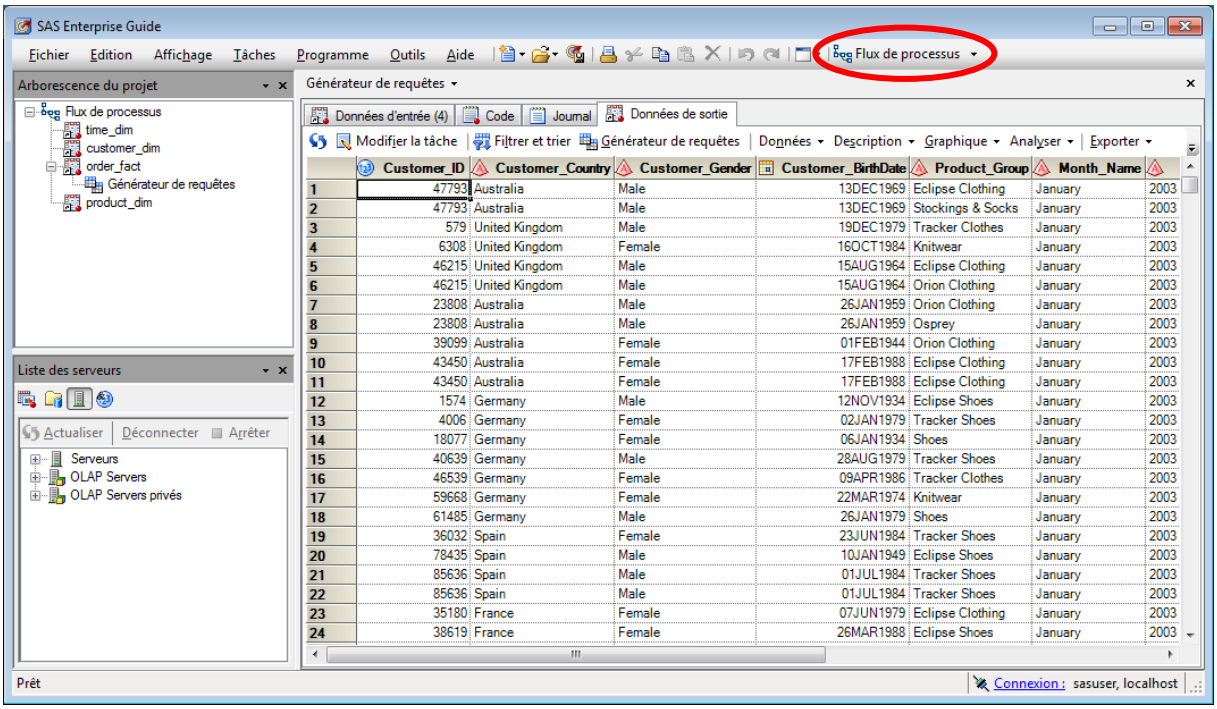


La valeur '31Dec2006'd est stocké dans SAS en 17 166, ce qui correspond au nombre de jours entre le premier janvier 1960 et le 31 décembre 2006.

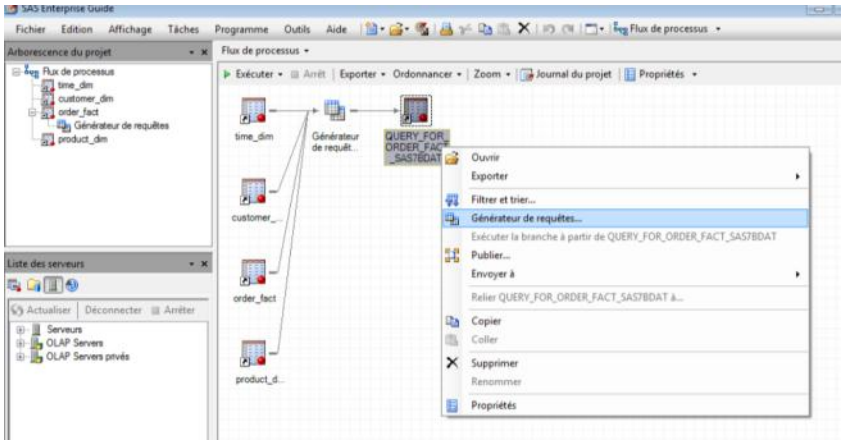
Terminer



Exécuter la requête.

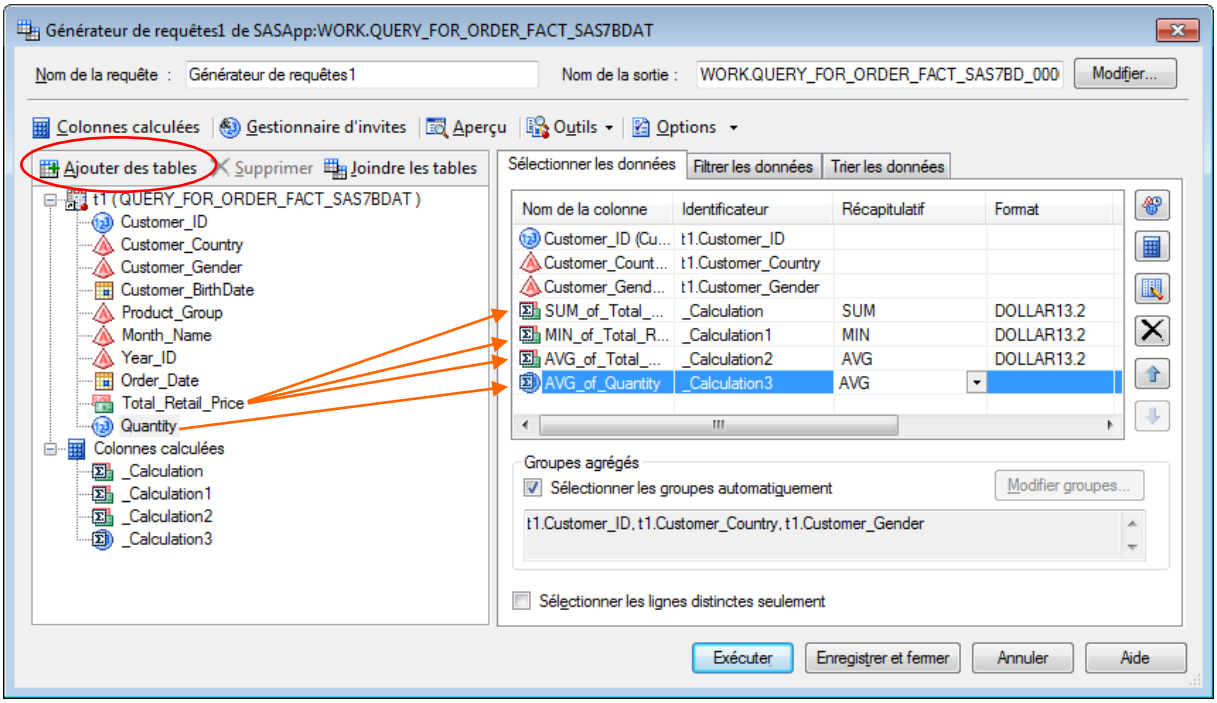


Revenir dans la fenêtre Flux de processus.



Création de la première table intermédiaire

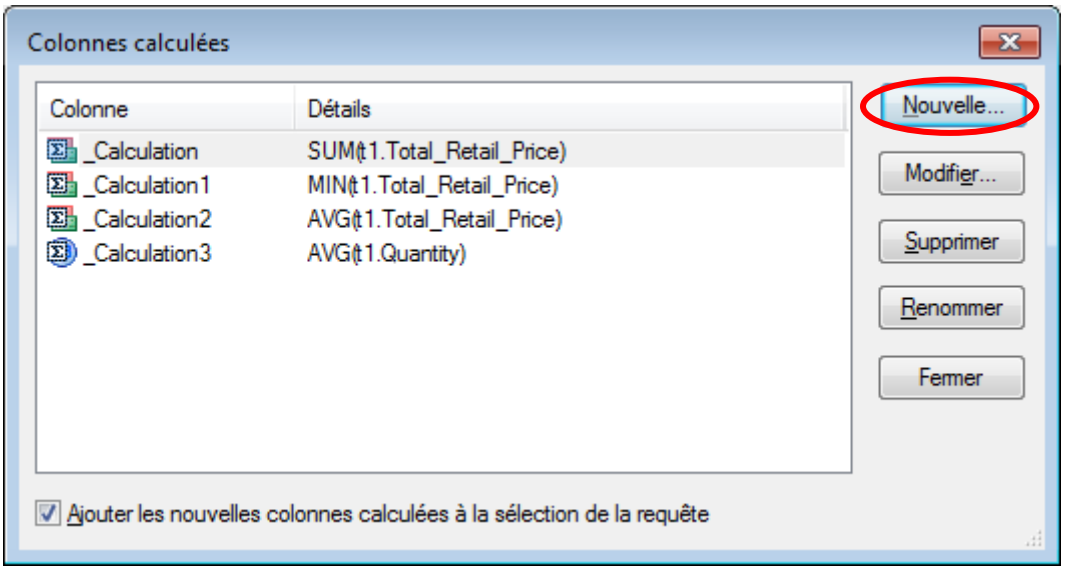
Clique droit sur la table base → Générateur de requêtes



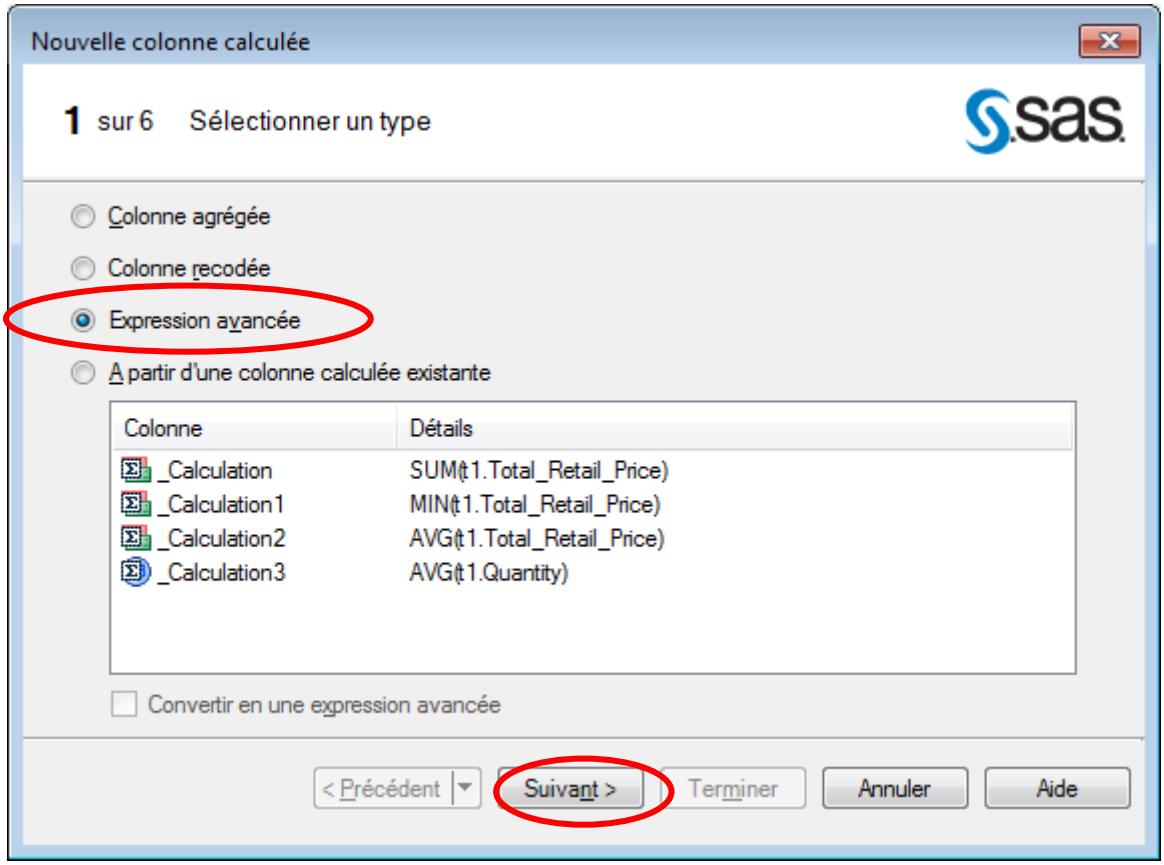
Sélectionner les colonnes :

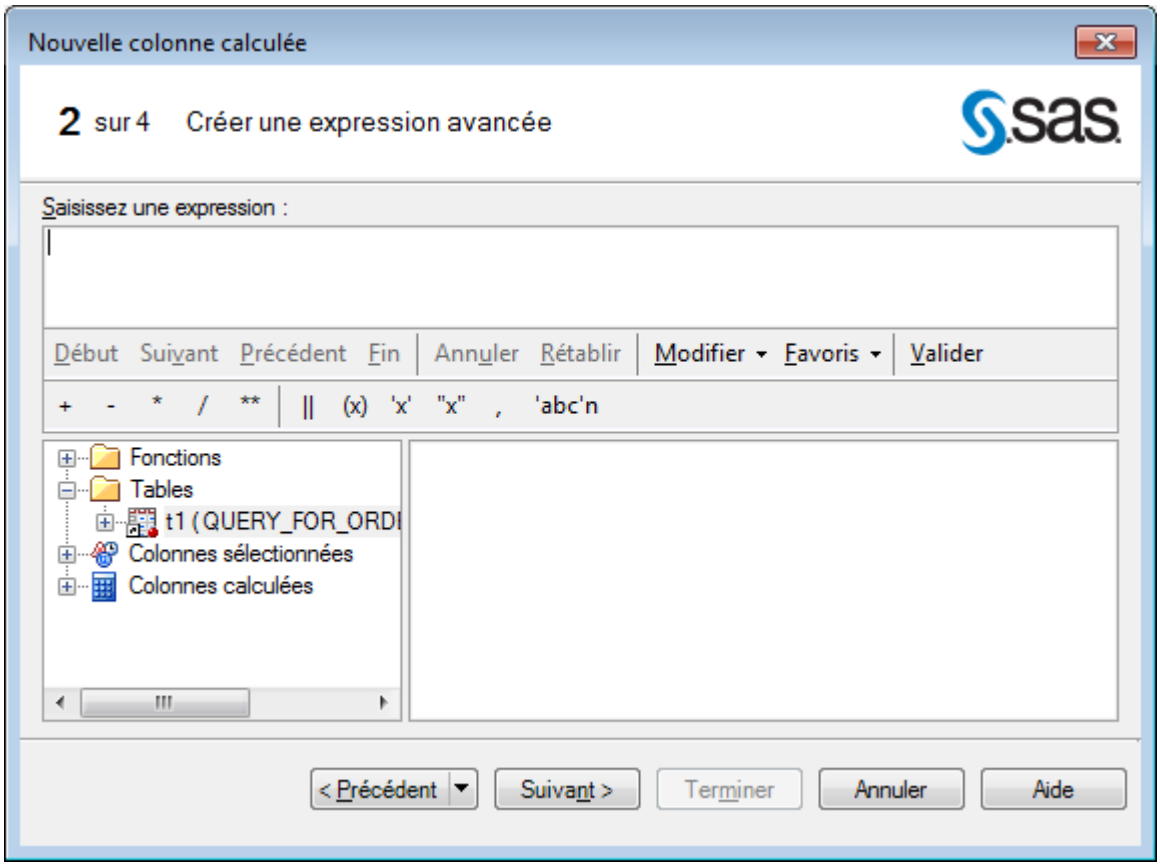
Custommer_ID
Customer_Country
Customer_Gender
Total_retail_price → SUM (il faut sélectionner sum dans récapitulatif, le nom de la colonne devient automatiquement SUM_of_Total_Retail_Price)
Total_retail_price → AVG
Total_retail_price → MIN
Quantity → AVG

Cliquer sur le bouton « colonnes calculées »



Créer une nouvelle colonnes calculées

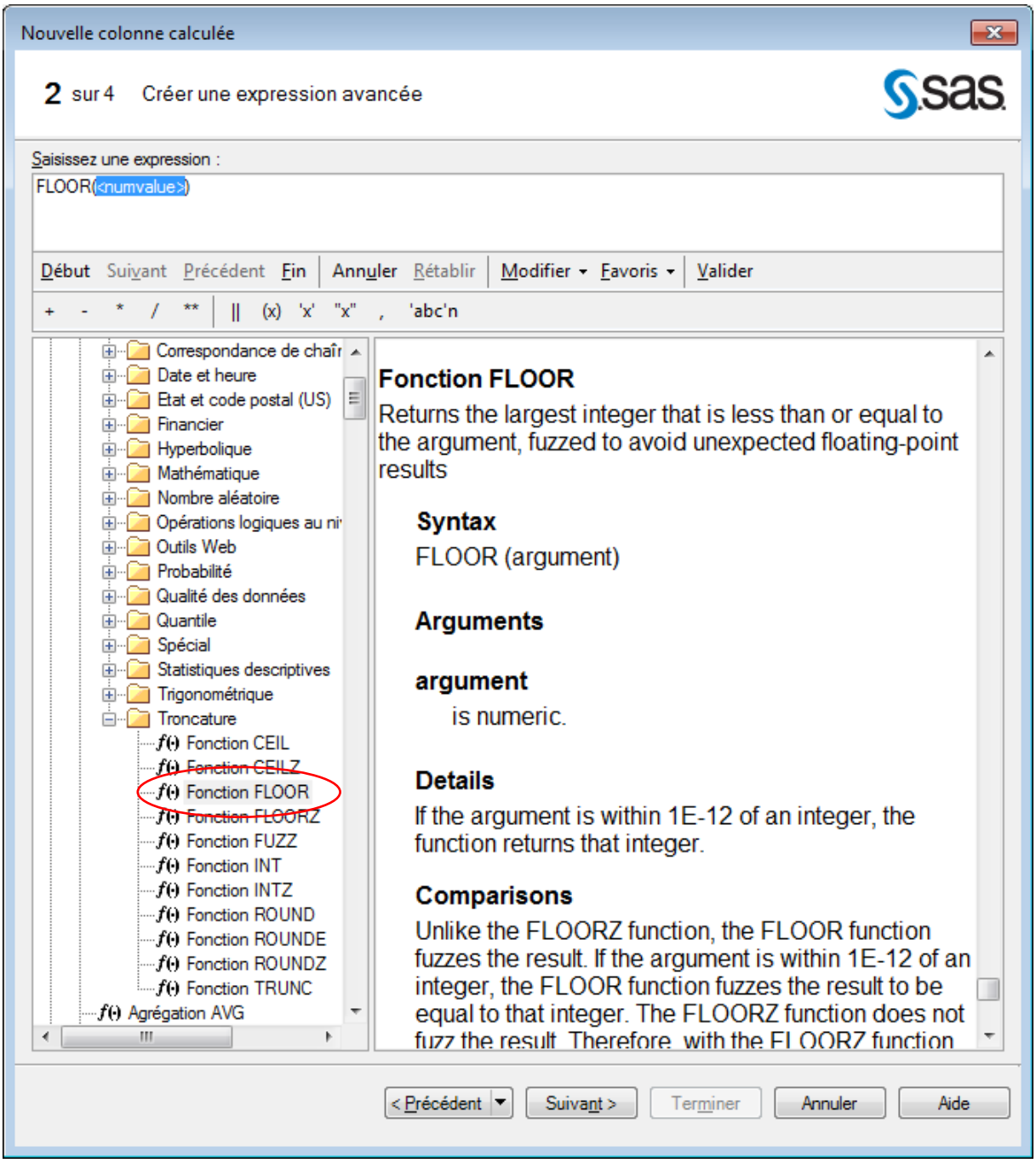




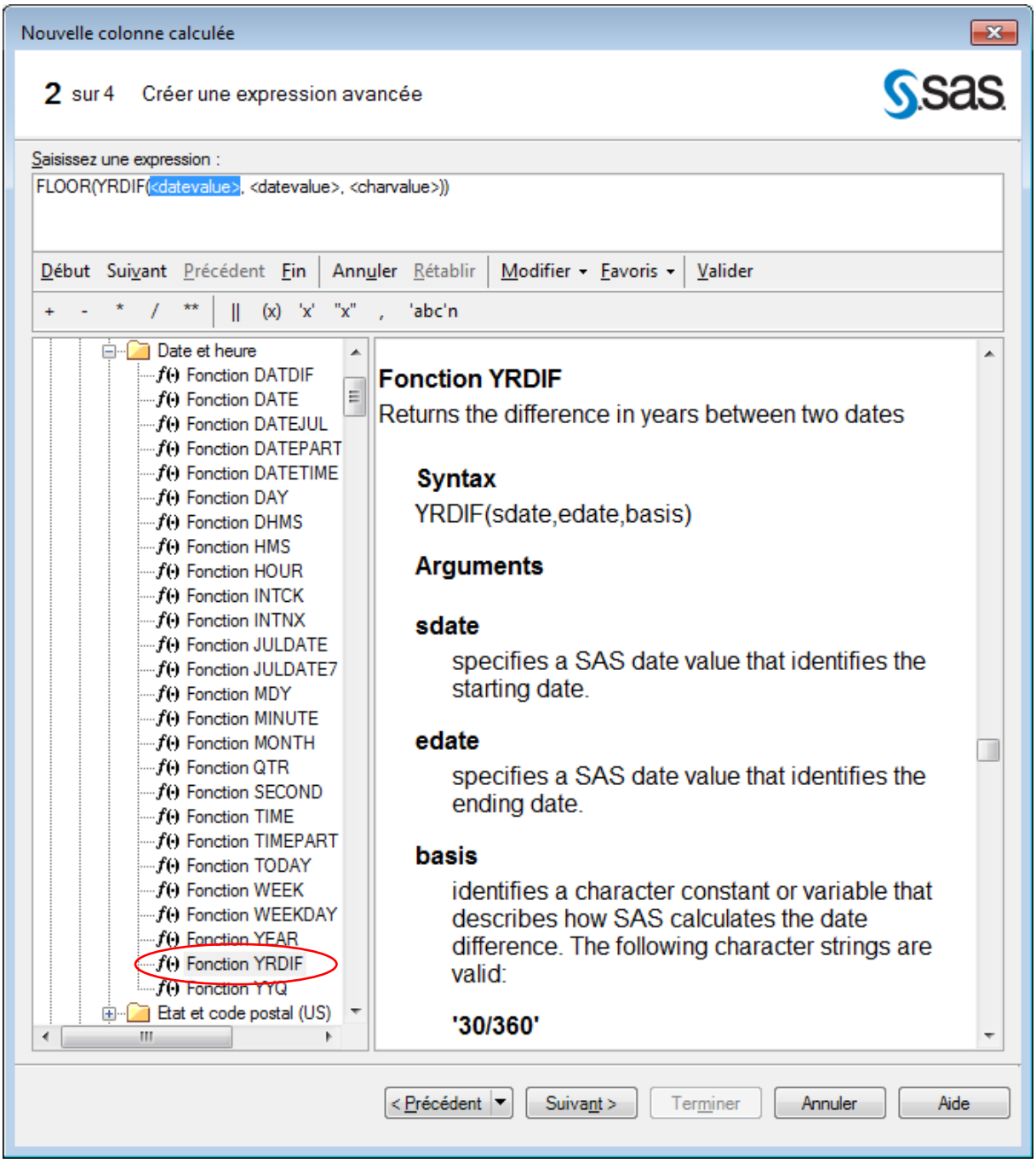
Calcul de l'âge du client au premier janvier 2007
L'expression de l'âge est une troncature de l'argument (floor) d'une différence de date (dans date et heure) sélectionnez YRDIF.
Dans source de données, sélectionnez pour le premier argument la colonne Birth_Date, pour le deuxième argument, 17166, et tapez 'actual' pour le dernier.

L'expression est donc :
FLOOR (YRDIF (t1.Customer_BirthDate, 17166,'actual'))

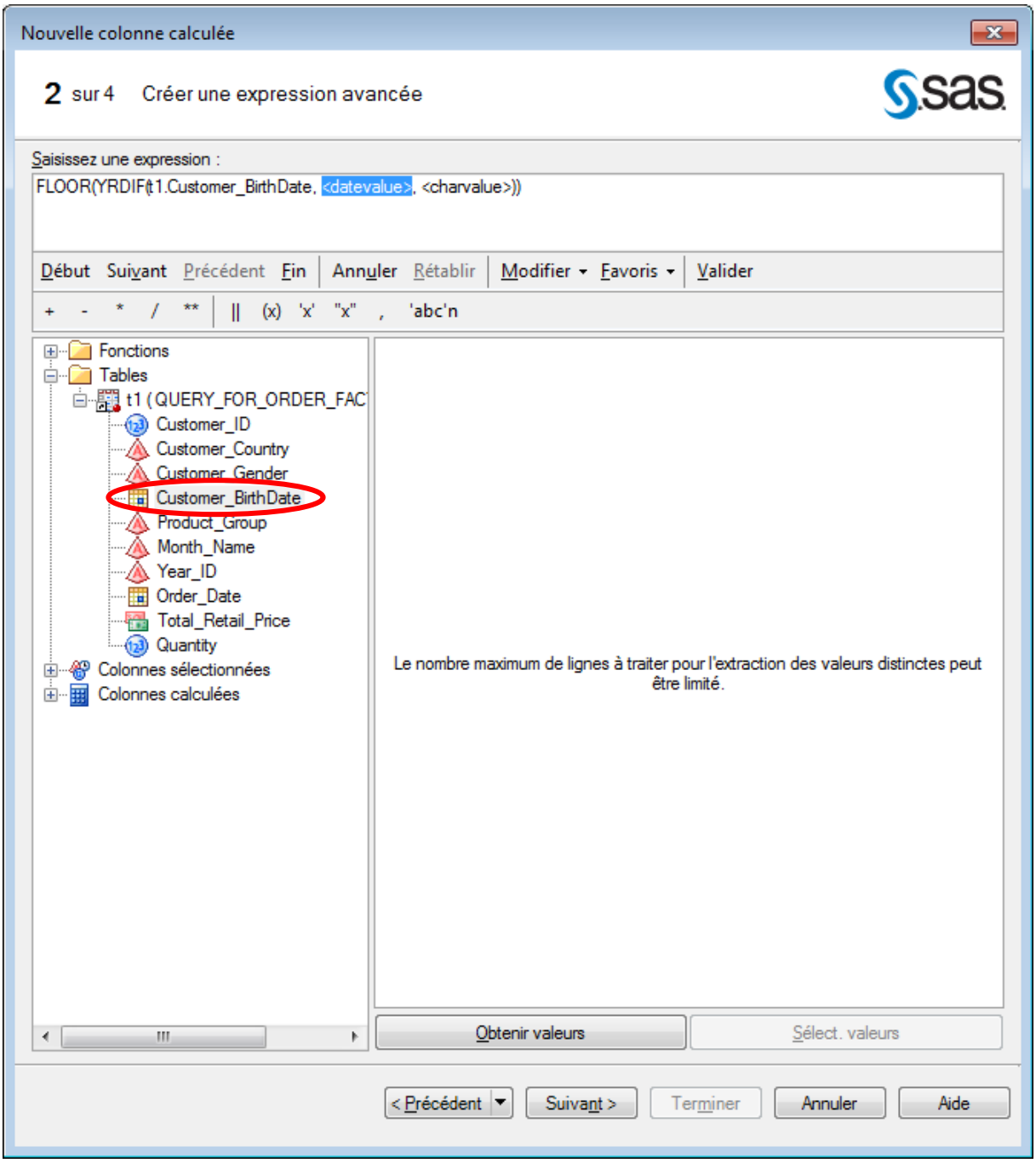
Soit, en copie d'écran :



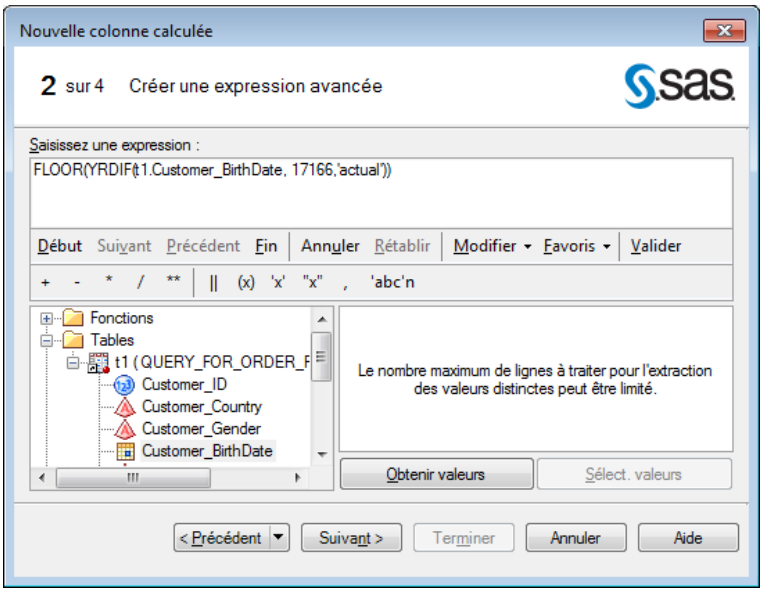
Dans les fonctions, dans Catégories, dans troncature, sélectionner la fonction FLOOR (ou bien INT, ou FLOORZ)



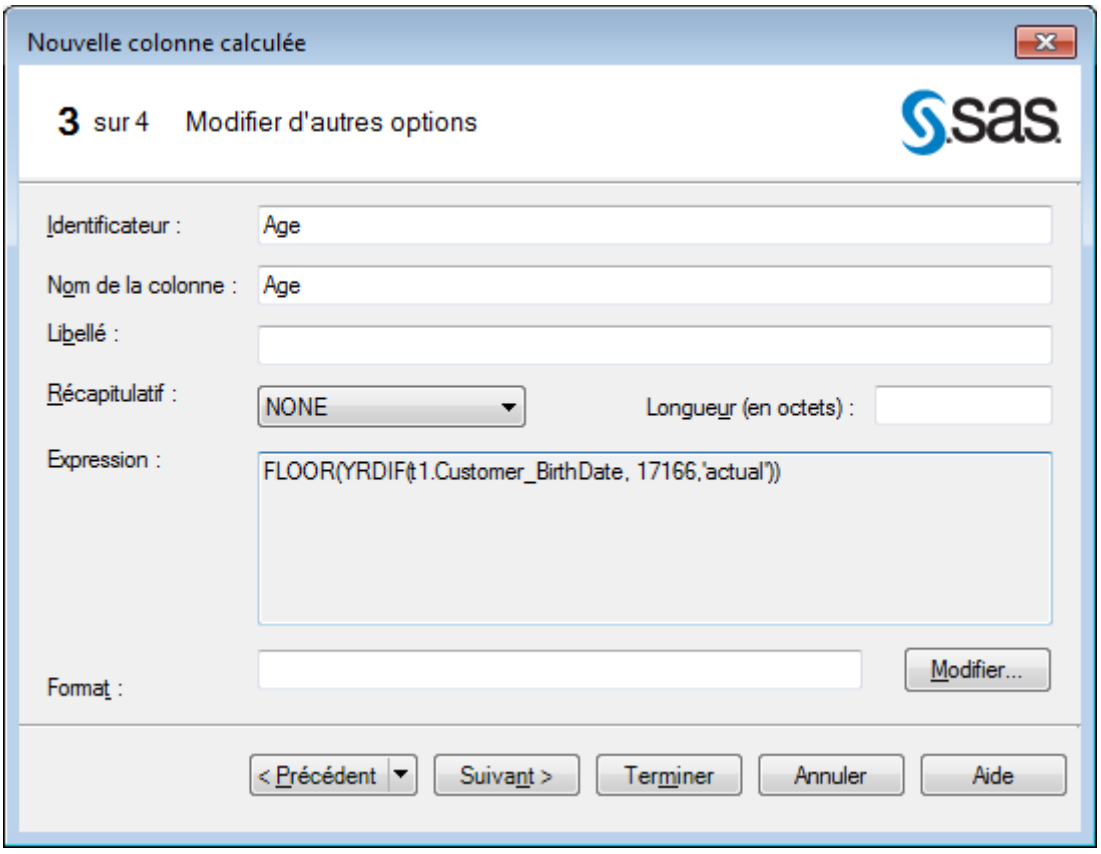
A la place <nomvalue>, sélectionner la fonction YRDIF (dans les fonctions, dans Catégories, dans date et heure)



A la place de la première date, insérer la date de naissance
A la place de la seconde date, insérer 17166
A la place du calendrier (dernière place), insérer ‘actual’ (actual entre simple cote – cote de la touche 4)

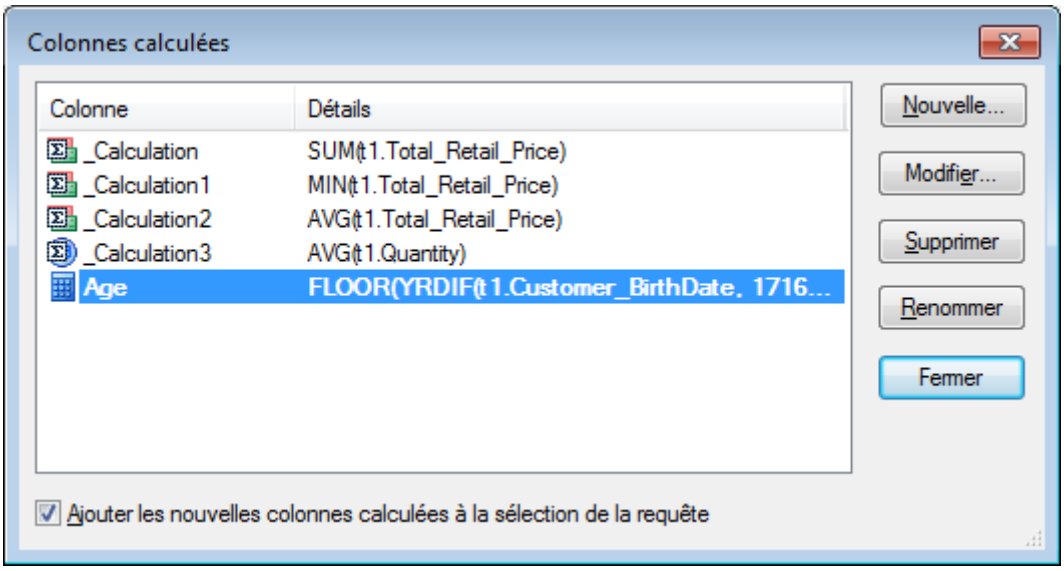


Suivant

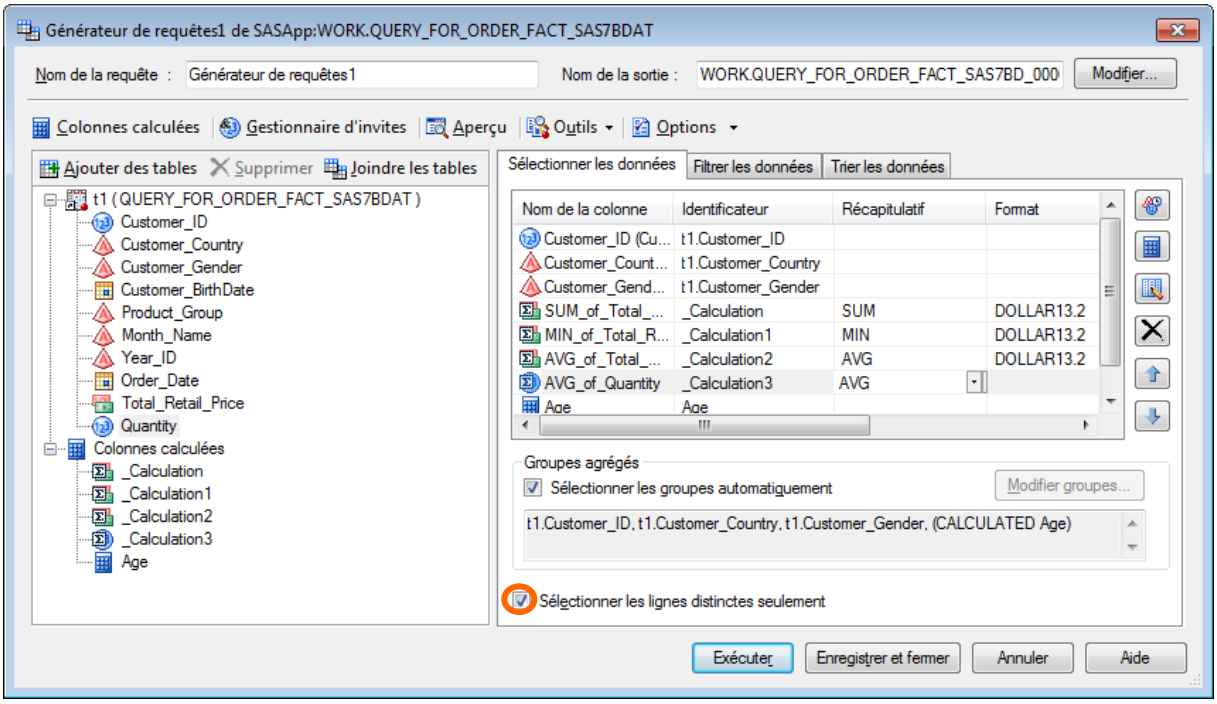


Renommer la colonne Age

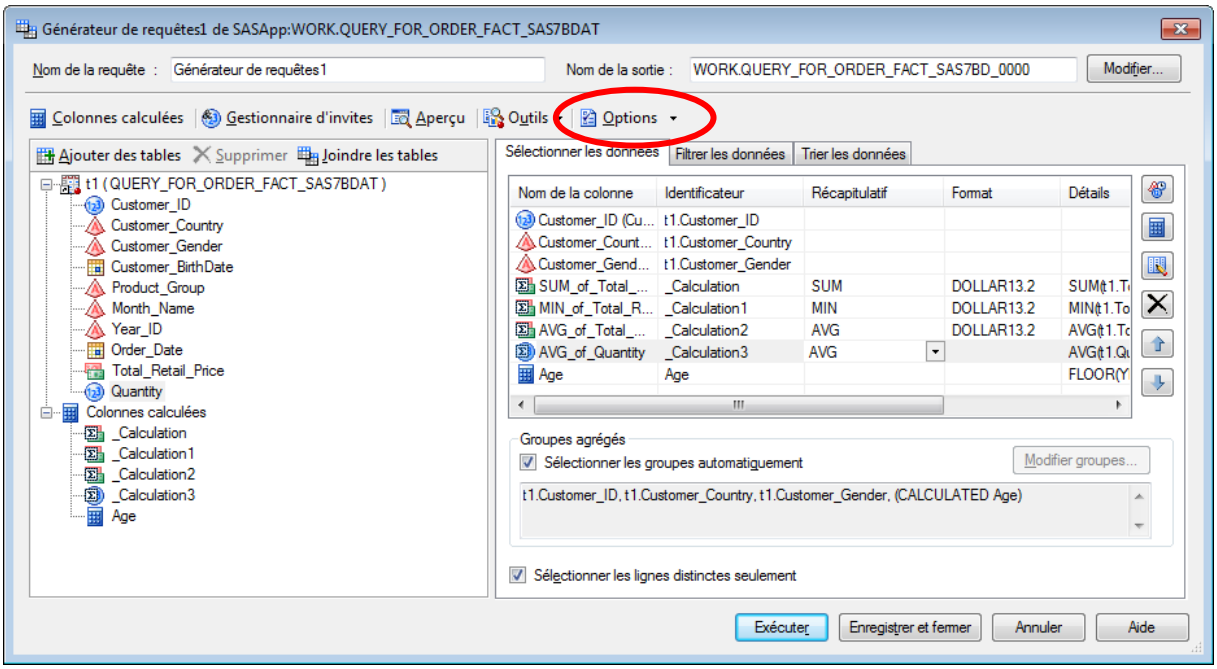
Terminer



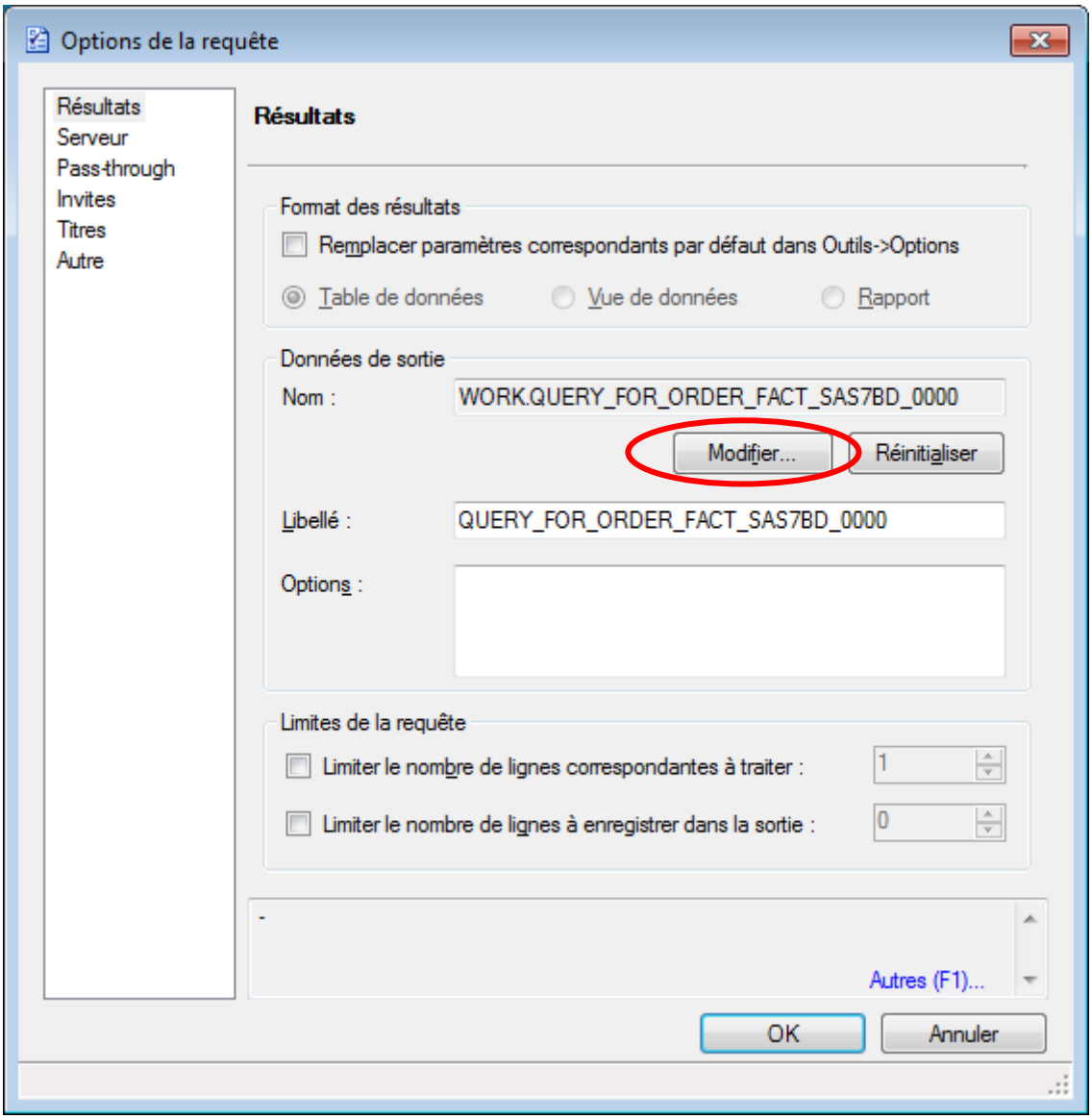
Fermer



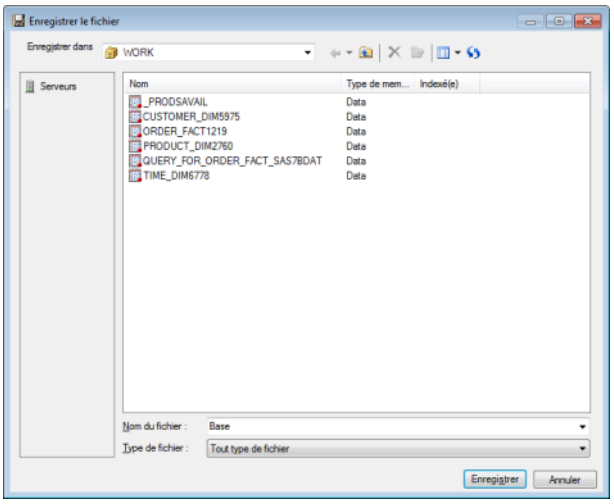
Sélectionner les lignes distinctes seulement,

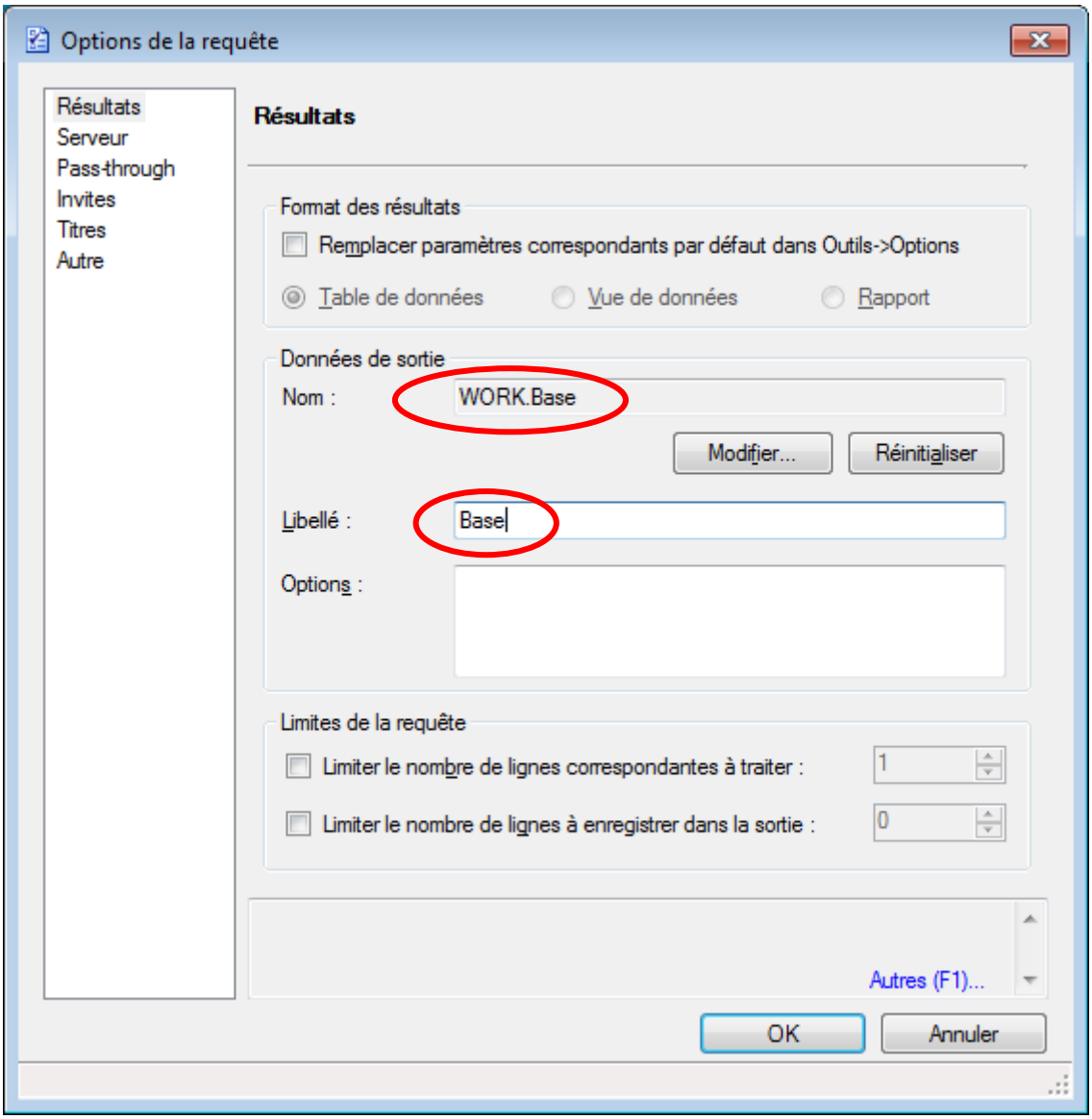


Dans les options



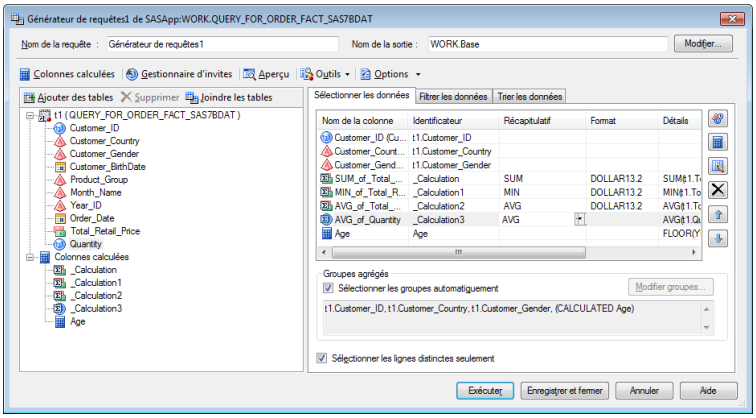
Modifier le nom de la table

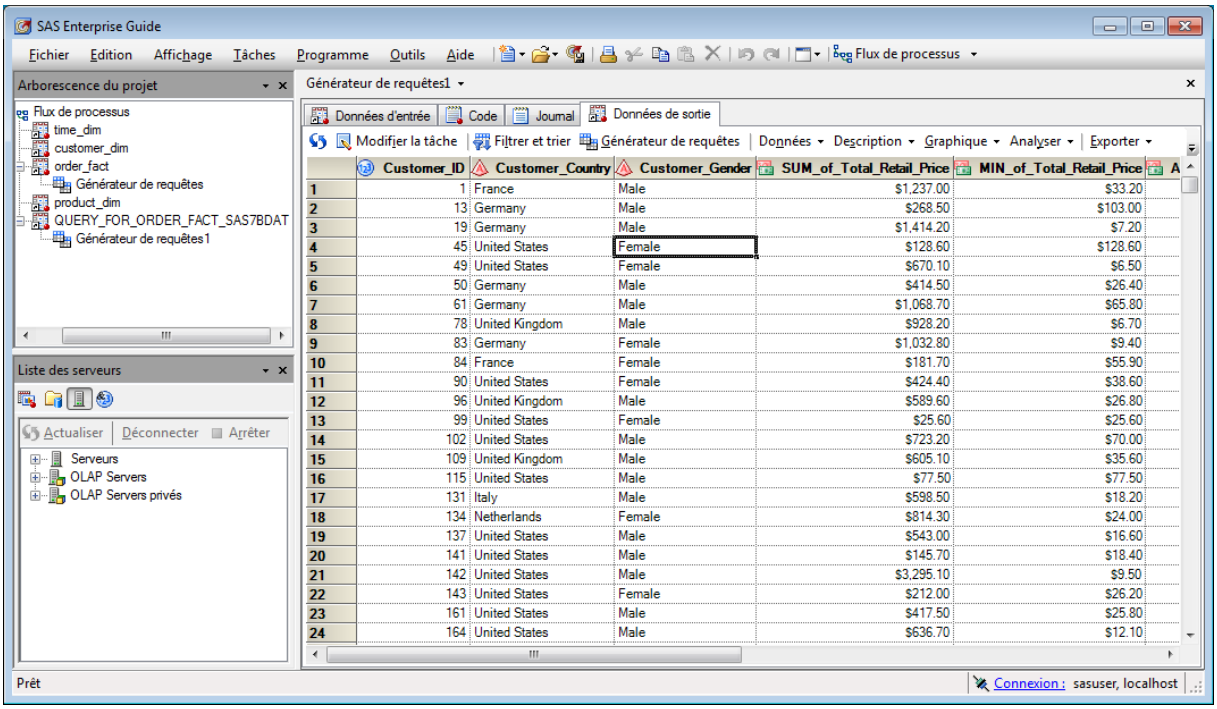




Par exemple « base », ce qui donne « work.base » si elle est dans la bibliothèque de travail.

Exécuter la requête.

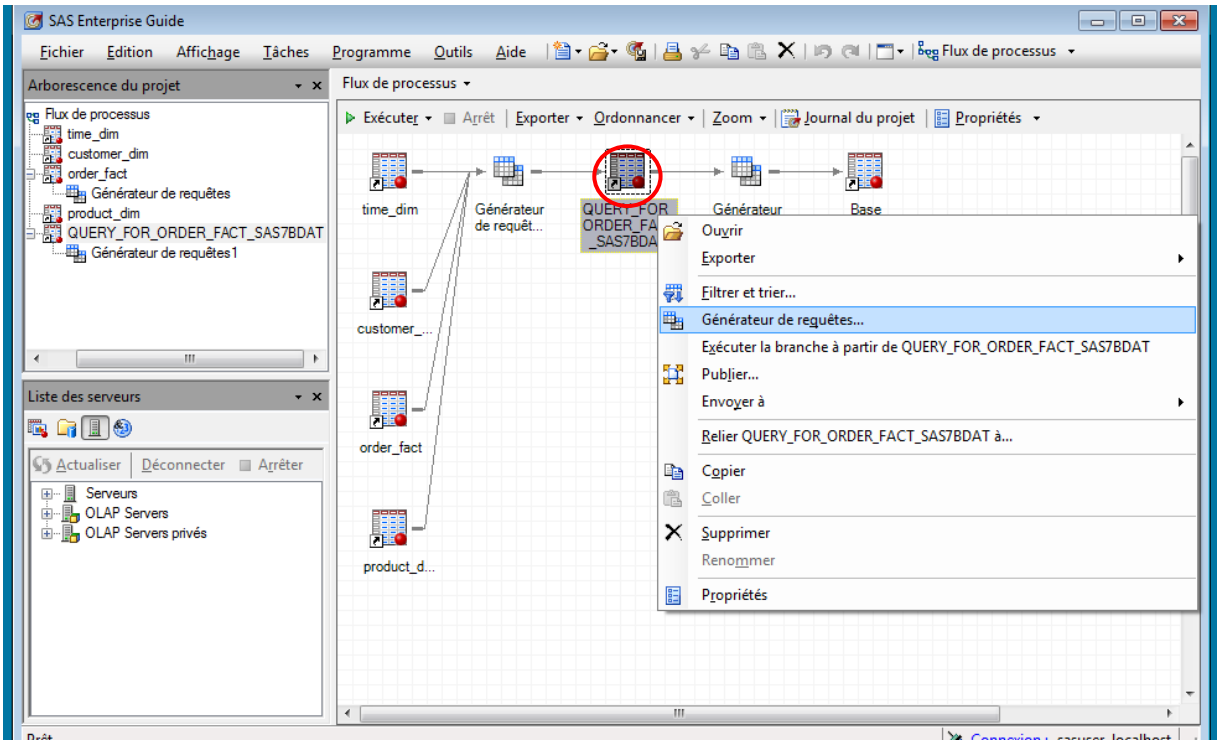




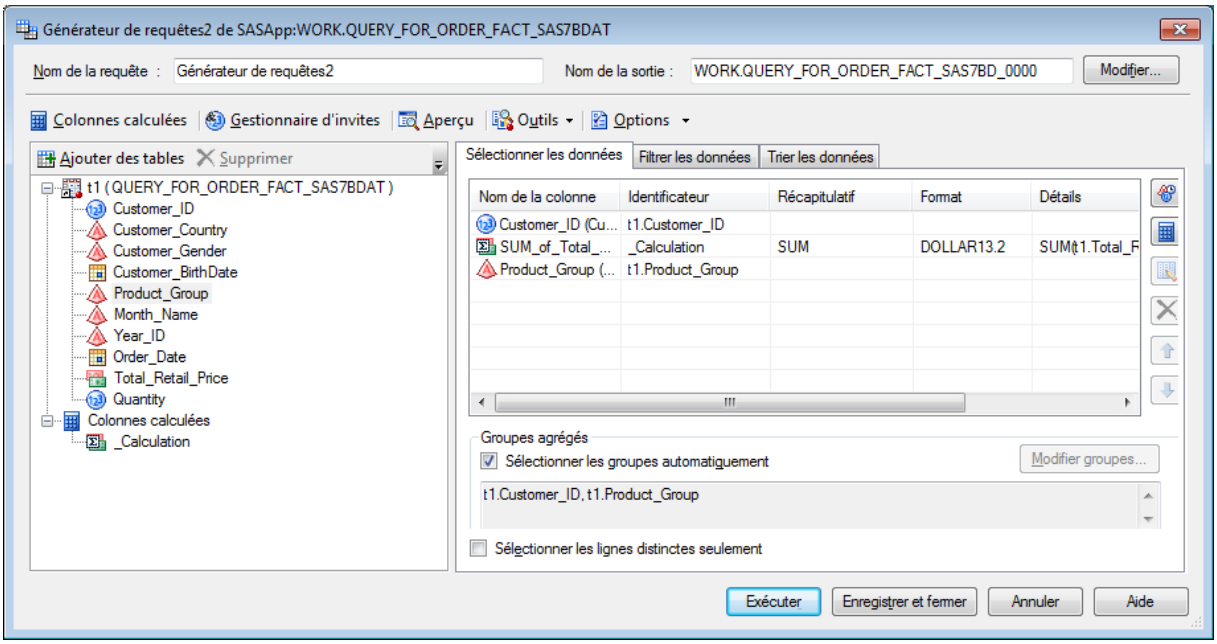
	Customer_ID	Customer_Country	Customer_Gender	SUM_of_Total_Retail_Price	MIN_of_Total_Retail_Price	A
1	1	France	Male	$1,237.00	$33.20	
2	13	Germany	Male	$268.50	$103.00	
3	19	Germany	Male	$1,414.20	$7.20	
4	45	United States	Female	$128.60	$128.60	
5	49	United States	Female	$670.10	$6.50	
6	50	Germany	Male	$414.50	$26.40	
7	61	Germany	Male	$1,068.70	$65.80	
8	78	United Kingdom	Male	$928.20	$6.70	
9	83	Germany	Female	$1,032.80	$9.40	
10	84	France	Female	$181.70	$55.90	
11	90	United States	Female	$424.40	$38.60	
12	96	United Kingdom	Male	$589.60	$26.80	
13	99	United States	Female	$25.60	$25.60	
14	102	United States	Male	$723.20	$70.00	
15	109	United Kingdom	Male	$605.10	$35.60	
16	115	United States	Male	$77.50	$77.50	
17	131	Italy	Male	$598.50	$18.20	
18	134	Netherlands	Female	$814.30	$24.00	
19	137	United States	Male	$543.00	$16.60	
20	141	United States	Male	$145.70	$18.40	
21	142	United States	Male	$3,295.10	$9.50	
22	143	United States	Female	$212.00	$26.20	
23	161	United States	Male	$417.50	$25.80	
24	164	United States	Male	$636.70	$12.10	

Création de la table du chiffre d'affaires par groupe de produit et par client

Pour créer cette table, depuis la table de base, nous allons créer une table de la somme du chiffre d'affaires par client et par groupe de produit, trier par client et par groupe de produit. Puis nous allons transposer les produits afin d'avoir une table avec une ligne par client et le chiffre d'affaires par groupe de produit.



Depuis la table base, créer une requête.

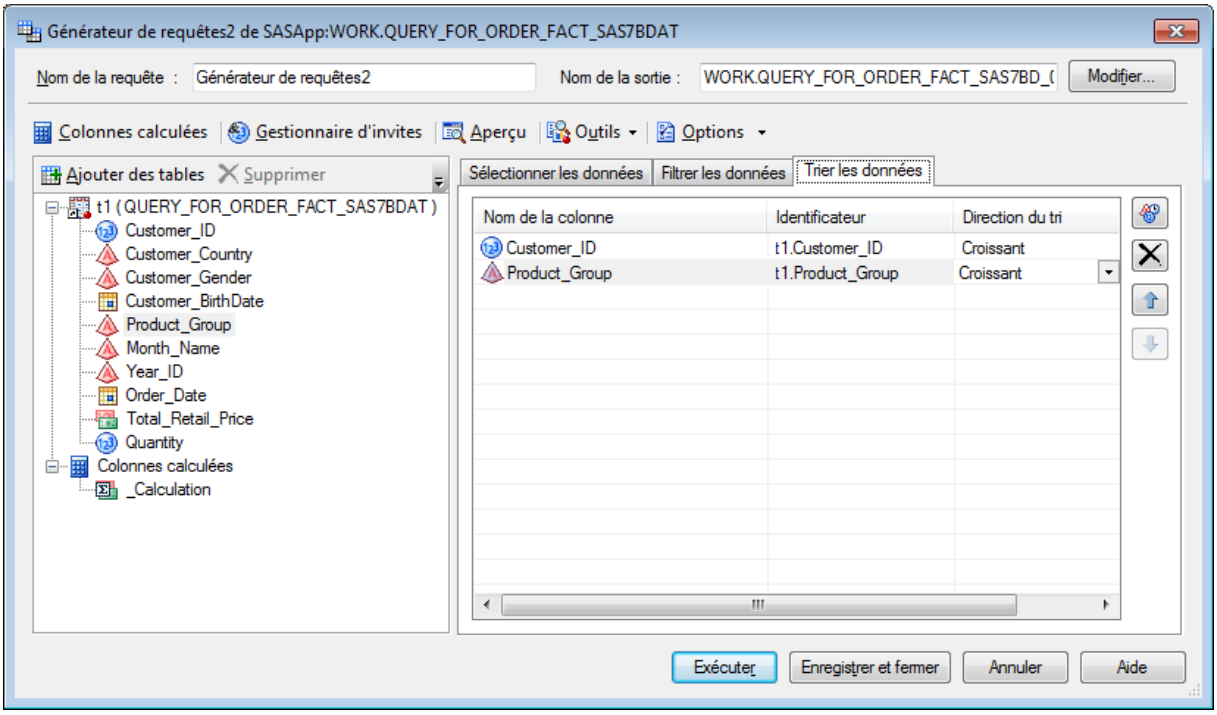


Sélectionner les colonnes :

Customer_ID

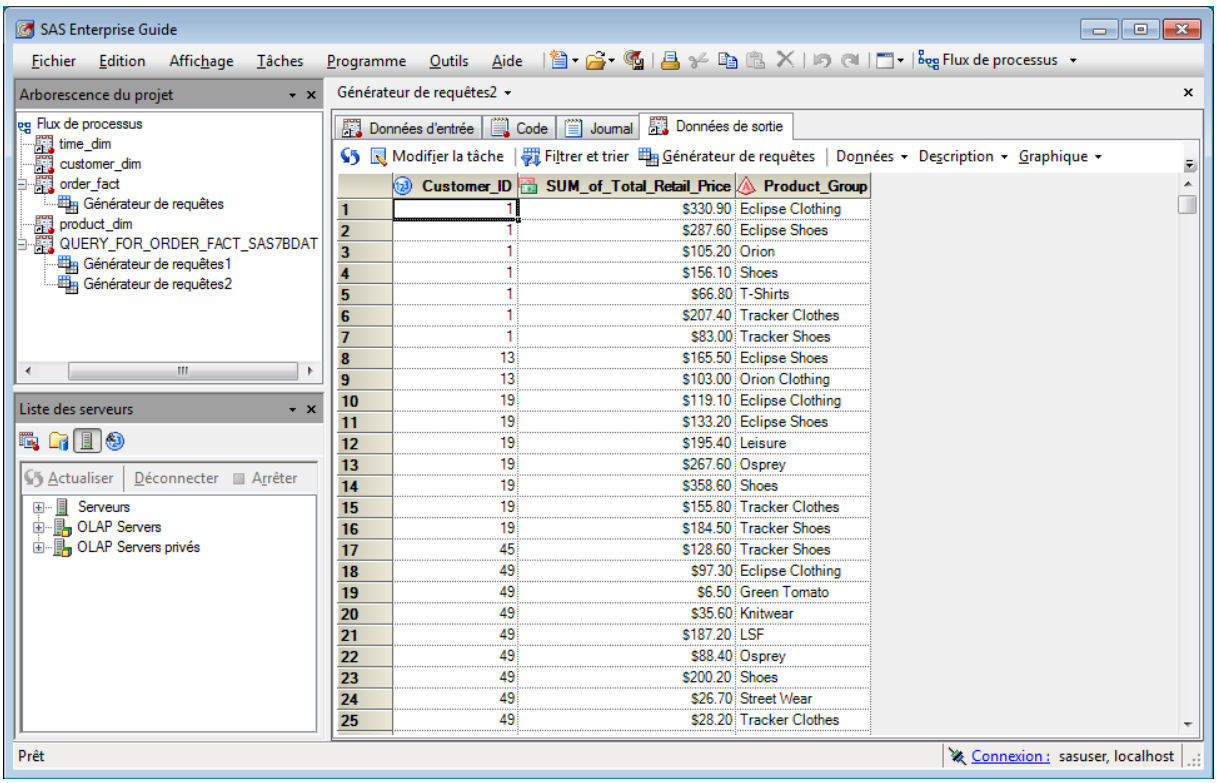
Product_Group
Total-Retail_Price

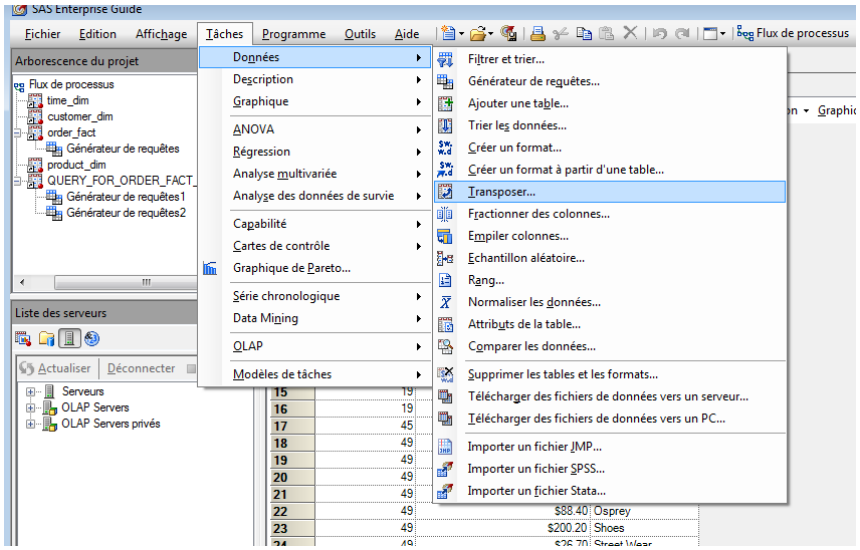
Sélectionner la somme du chiffre d'affaires



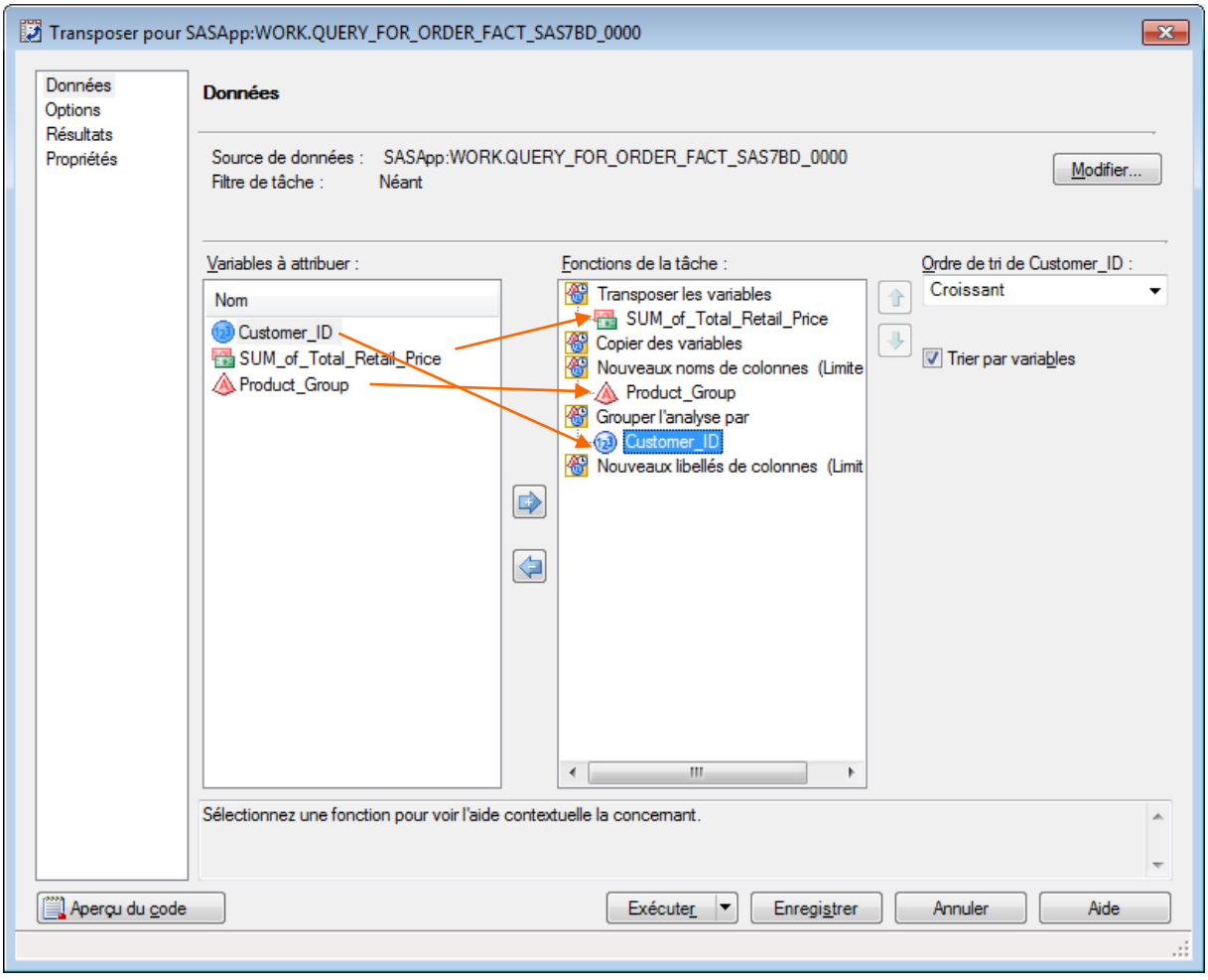
Dans l'onglet « Trier les données »
Sélectionner Customer_ID puis Product_Group.

Executer

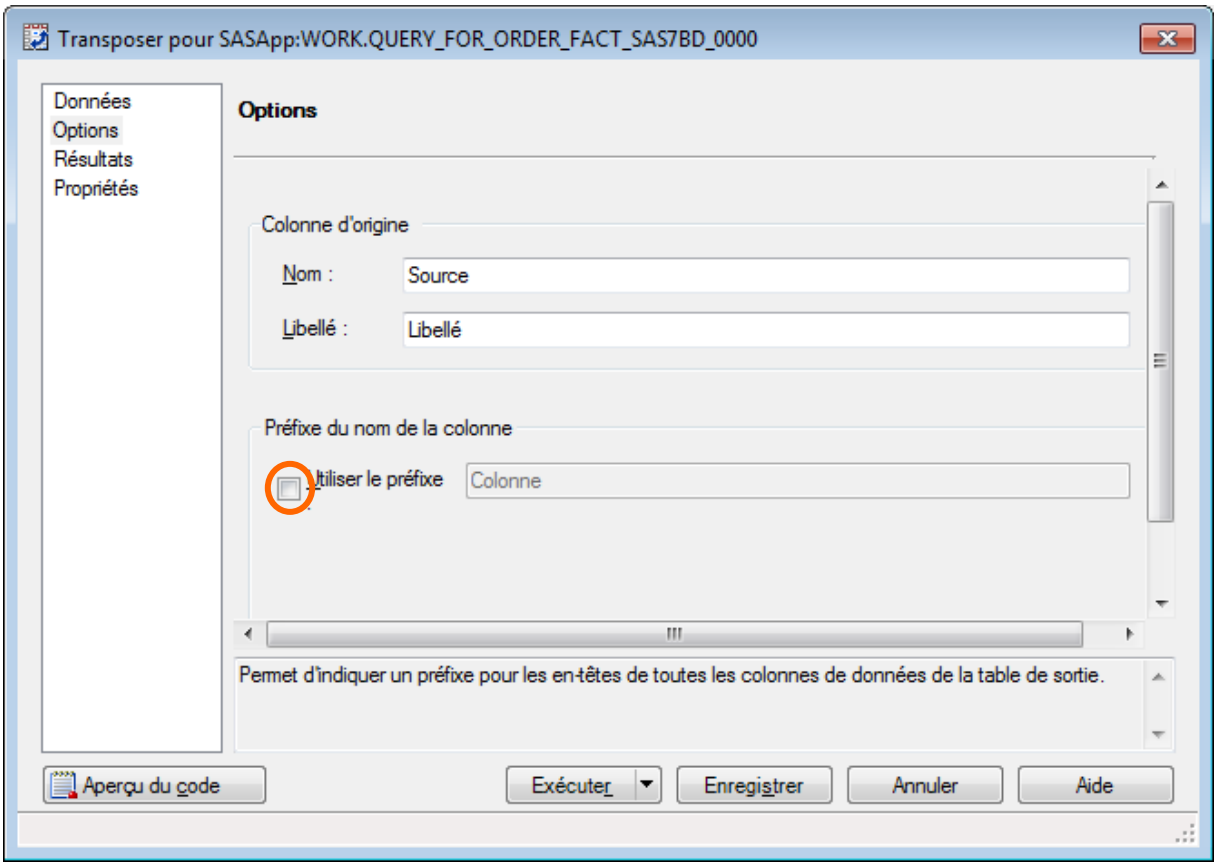




Dans le menu des tâches → Données → utiliser la fonction « transposer »

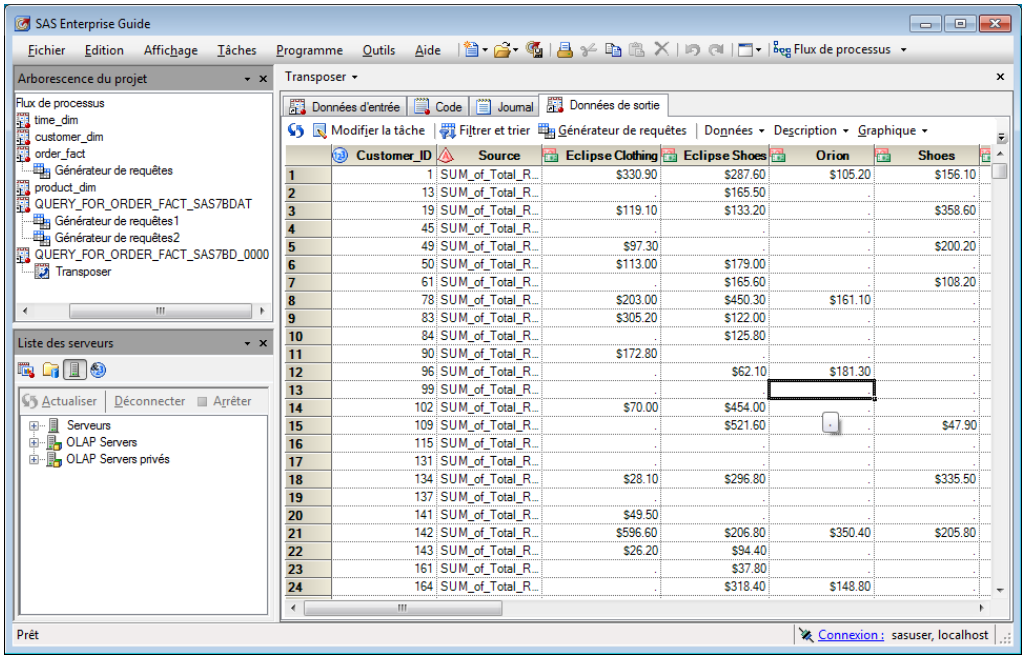


Affecter la somme du chiffre d'affaires à la fonction « transposer les variables ».
Affecter le groupe de produit à la fonction « nouveaux noms de colonne »
Et Affecter le numéro du client à « grouper l'analyse par »



Désélectionner l'option « utiliser le préfixe ».

Exécuter

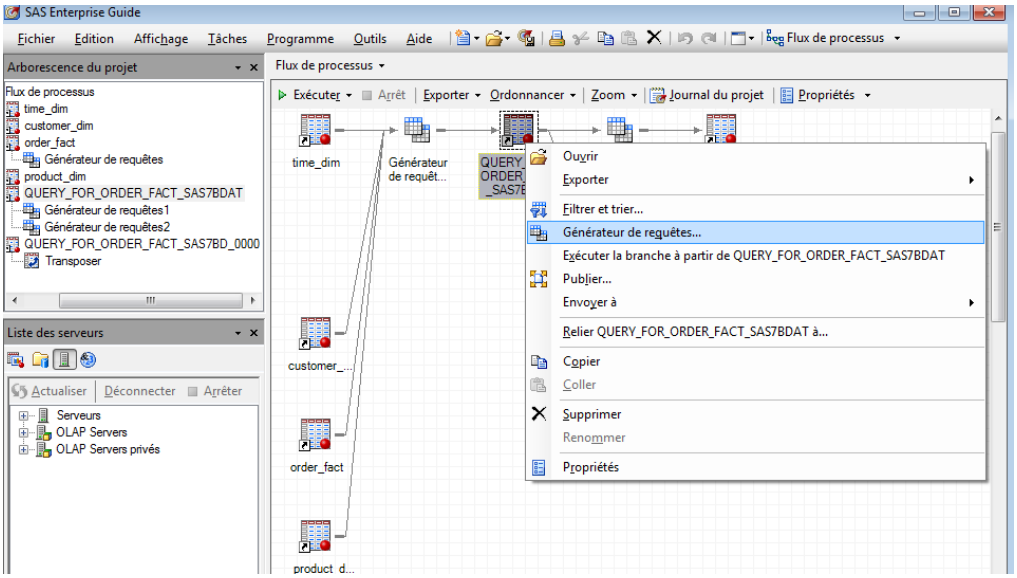


	Customer_ID	Source	Eclipse Clothing	Eclipse Shoes	Orion	Shoes
1	1	SUM_of_Total_R...	$330.90	$287.60	$105.20	$156.10
2	13	SUM_of_Total_R...	.	$165.50	.	.
3	19	SUM_of_Total_R...	$119.10	$133.20	.	$358.60
4	45	SUM_of_Total_R...	.	.	.	.
5	49	SUM_of_Total_R...	$97.30	.	.	$200.20
6	50	SUM_of_Total_R...	$113.00	$179.00	.	.
7	61	SUM_of_Total_R...	.	$165.60	.	$108.20
8	78	SUM_of_Total_R...	$203.00	$450.30	$161.10	.
9	83	SUM_of_Total_R...	$305.20	$122.00	.	.
10	84	SUM_of_Total_R...	.	$125.80	.	.
11	90	SUM_of_Total_R...	$172.80	.	.	.
12	96	SUM_of_Total_R...	.	$62.10	$181.30	.
13	99	SUM_of_Total_R...	.	.	.	.
14	102	SUM_of_Total_R...	$70.00	$454.00	.	.
15	109	SUM_of_Total_R...	.	$521.60	.	$47.90
16	115	SUM_of_Total_R...	.	.	.	.
17	131	SUM_of_Total_R...	.	.	.	.
18	134	SUM_of_Total_R...	$28.10	$296.80	.	$335.50
19	137	SUM_of_Total_R...	.	.	.	.
20	141	SUM_of_Total_R...	$49.50	.	.	.
21	142	SUM_of_Total_R...	$596.60	$206.80	$350.40	$205.80
22	143	SUM_of_Total_R...	$26.20	$94.40	.	.
23	161	SUM_of_Total_R...	.	$37.80	.	.
24	164	SUM_of_Total_R...	.	$318.40	$148.80	.

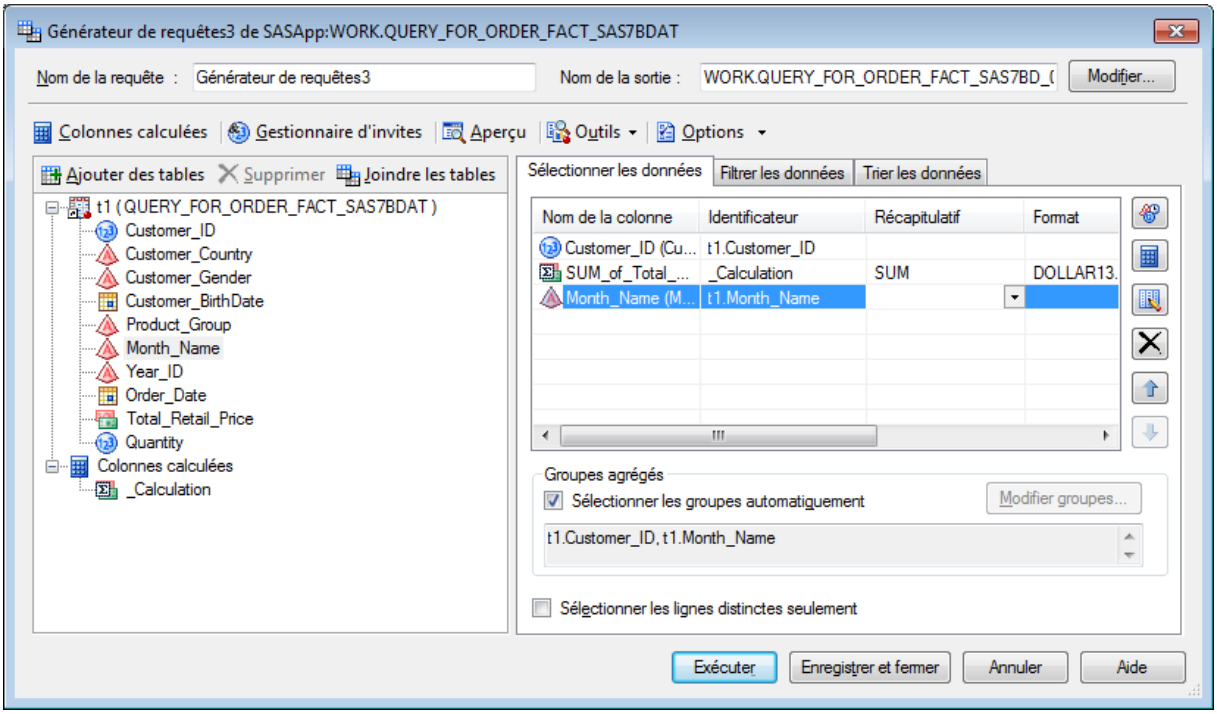
Nous avons une table avec le chiffre d'affaires par numéro de client en ligne et par groupe de produit en colonne.

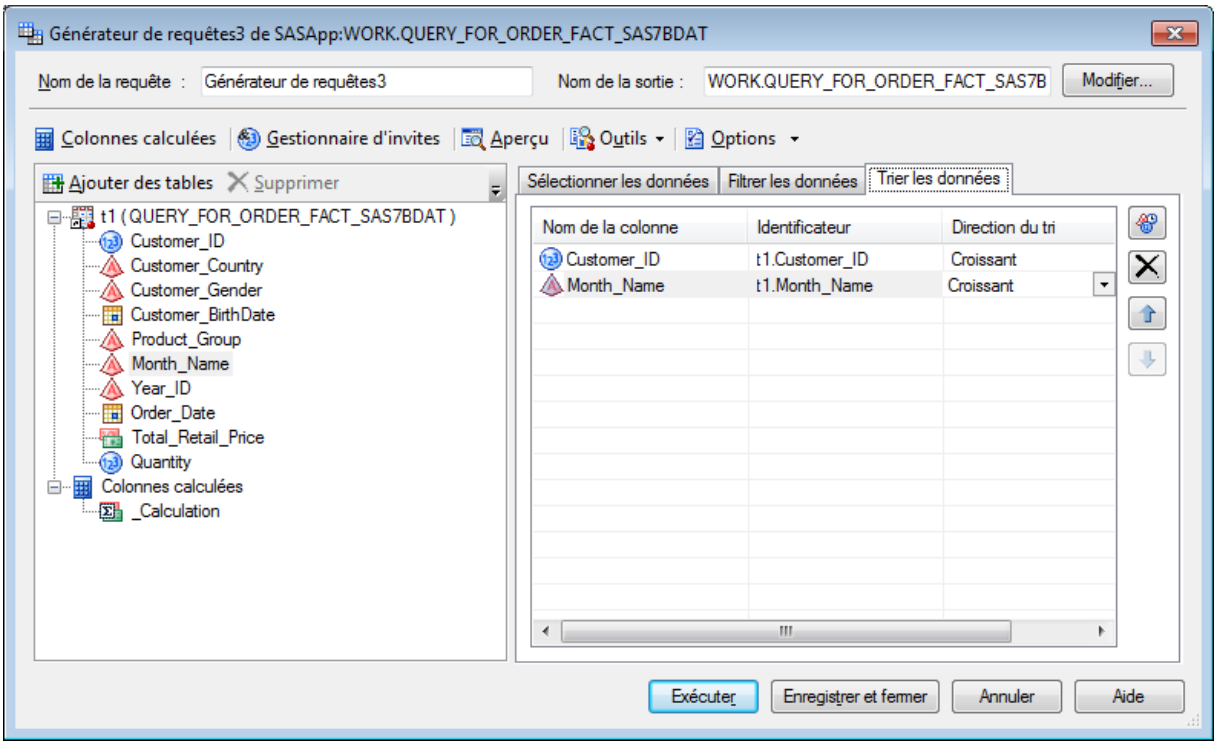
Création de la table du chiffre d'affaires par mois et par client

Pour créer cette table, nous allons reprendre le même processus que précédemment ; c'est-à-dire, créer une table de la somme du chiffre d'affaires par numéro de client et par mois et la transposer.

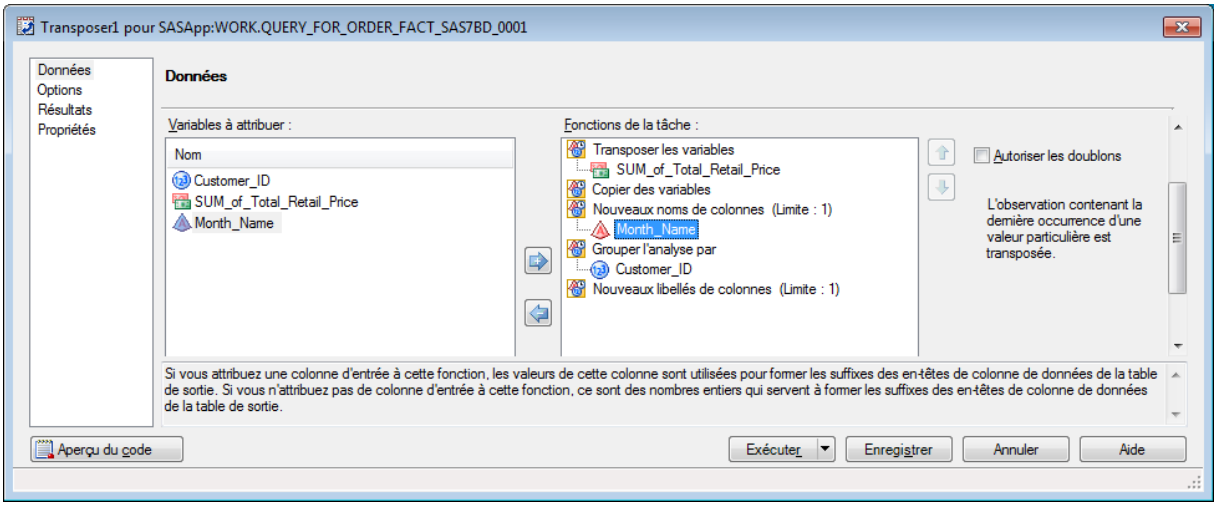


Depuis la table de « base »
Créer une requête, de la somme du chiffre d'affaires par mois et par numéro de clients

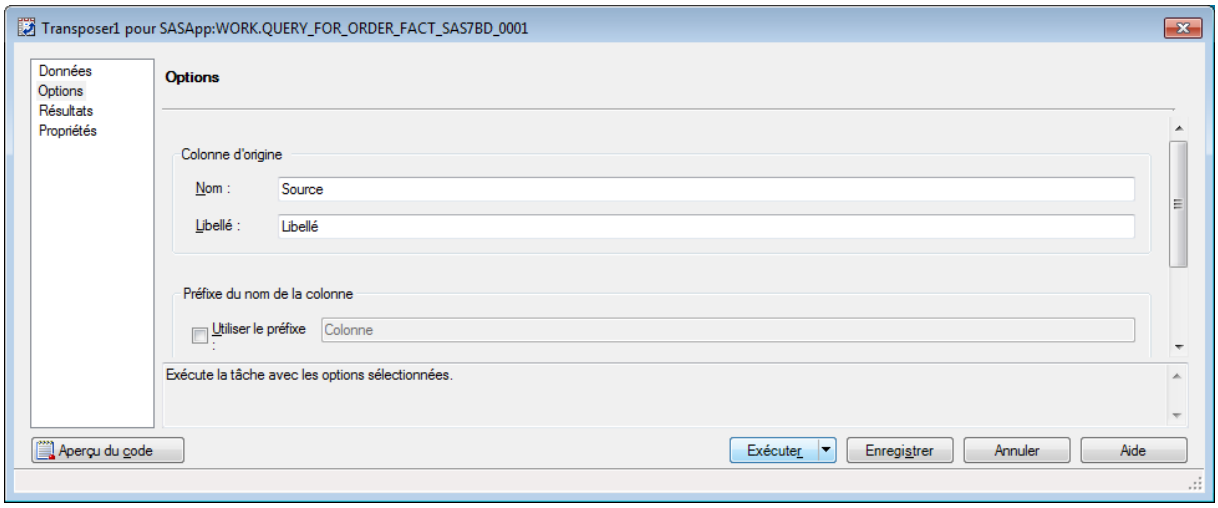


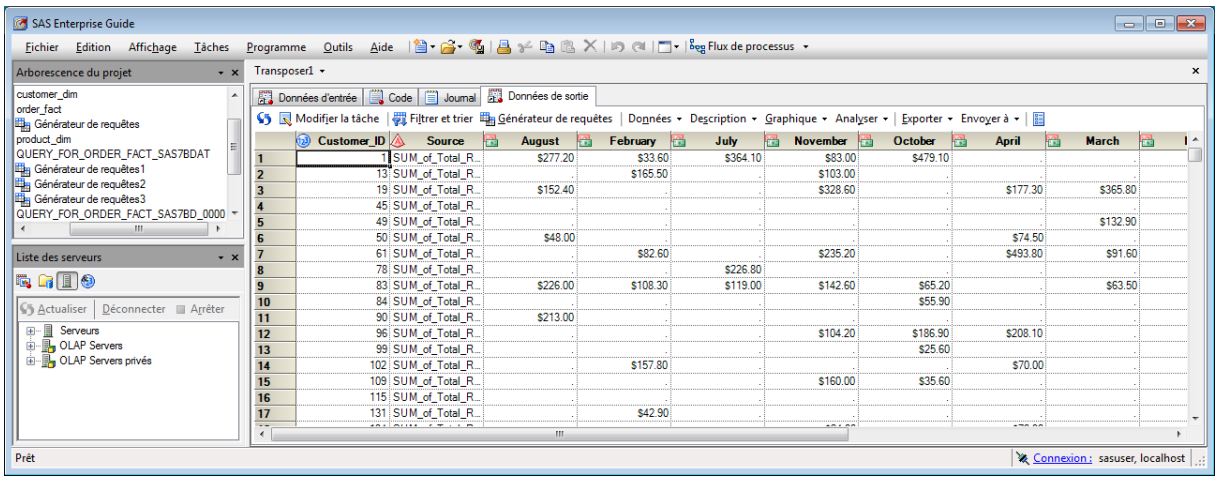


Trier par numéro de clients et par mois.
Exécuter



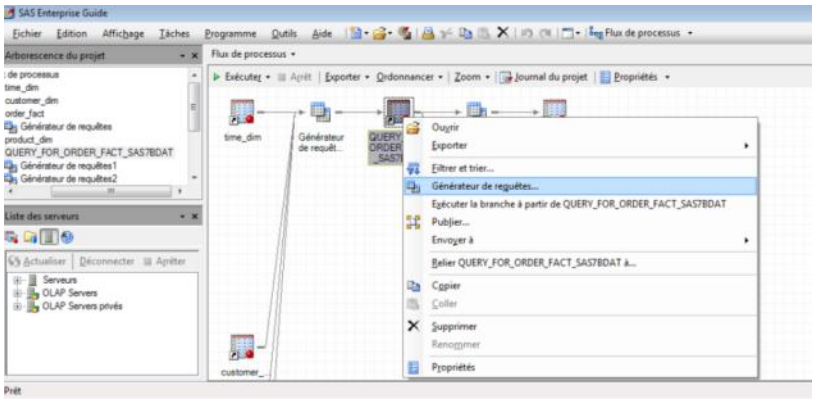
Transposer la table


Transposer1 pour SASApp:WORK.QUERY_FOR_ORDER_FACT_SAS7BD_0001
Données
Options
Résultats
Propriétés
Options
Colonne d'origine
Nom :
Source
Libellé :
Libellé
Préfixe du nom de la colonne
Utiliser le préfixe
Colonne
Exécute la tâche avec les options sélectionnées.
Aperçu du code
Exécuter
Enregistrer
Annuler
Aide

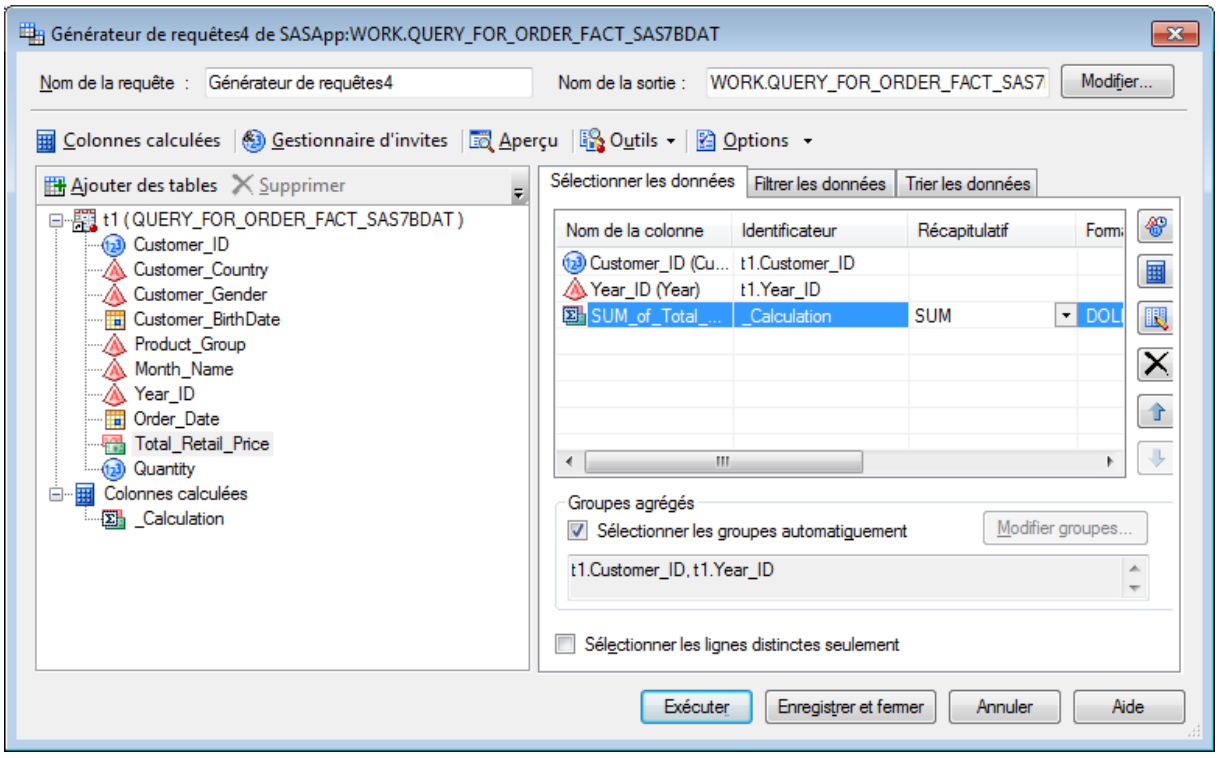

SAS Enterprise Guide
Fichier
Edition
Affichage
Tâches
Programme
Outils
Aide
Flux de processus
Arborescence du projet
customer_dim
order_fact
Générateur de requêtes
product_dim
QUERY_FOR_ORDER_FACT_SAS7BDAT
Générateur de requêtes1
Générateur de requêtes2
Générateur de requêtes3
QUERY_FOR_ORDER_FACT_SAS7BD_0000
Liste des serveurs
Actualiser
Déconnecter
Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Transposer1
Données d'entrée
Code
Journal
Données de sortie
Modifier la tâche
Filtrer et trier
Générateur de requêtes
Données
Description
Graphique
Analyser
Exporter
Envoyer à
Customer_ID
Source
August
February
July
November
October
April
March
Prêt
Connexion : sasuser, localhost

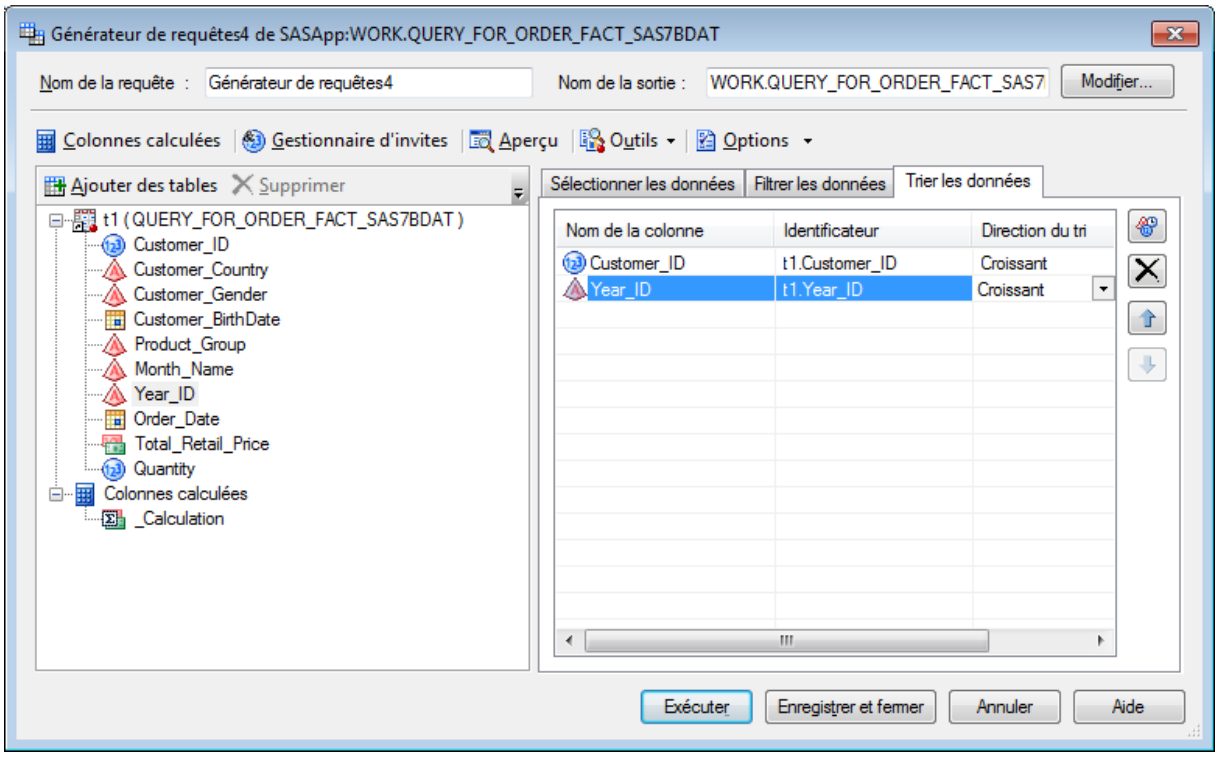
Création de la table du chiffre d'affaires par année et par client

Pour créer cette table, nous allons reprendre le même processus que précédemment ; c'est-à-dire, créer une table de la somme du chiffre d'affaires par numéro de client et par année et la transposer.

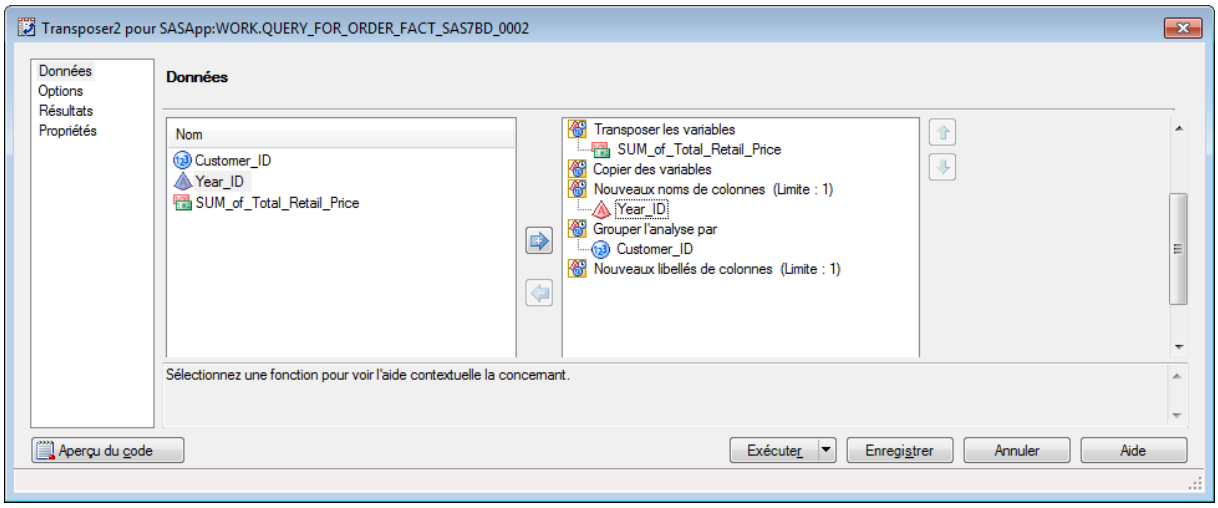


Depuis la table de « base »
Créer une requête, de la somme du chiffre d'affaires par année et par numéro de clients

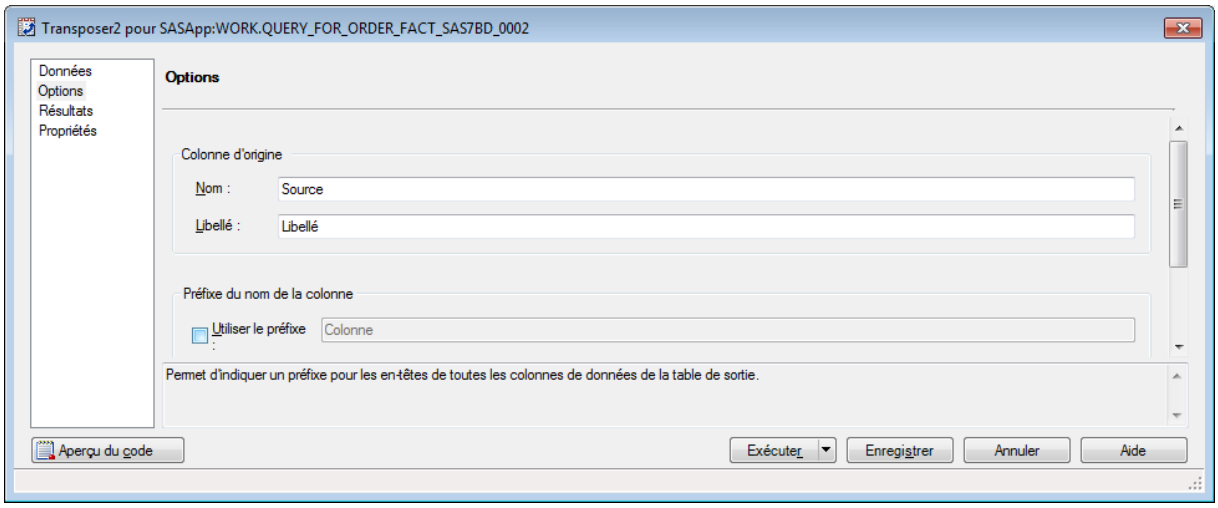


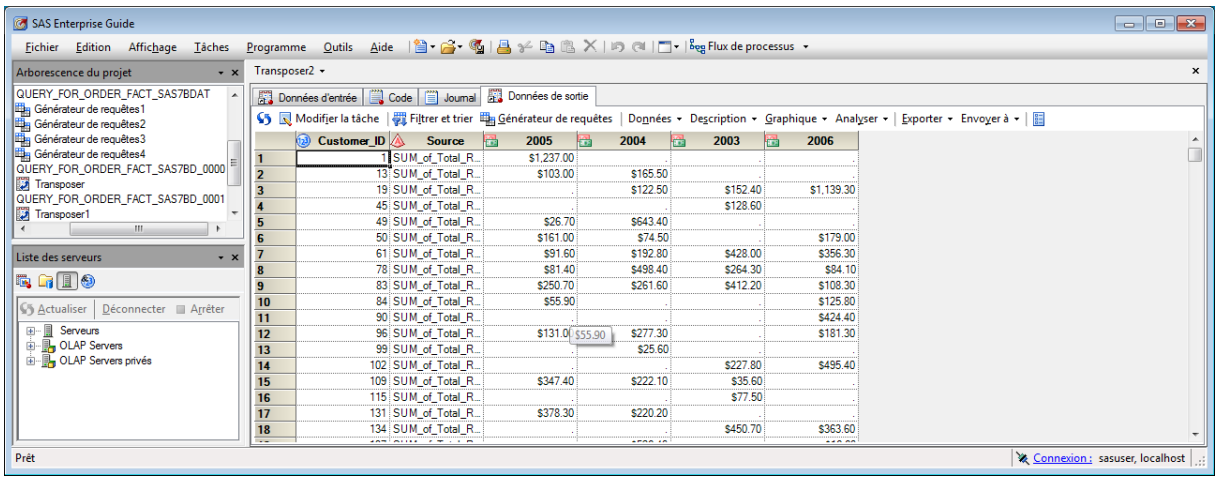


Trier par numéro de client et par année.
Exécuter



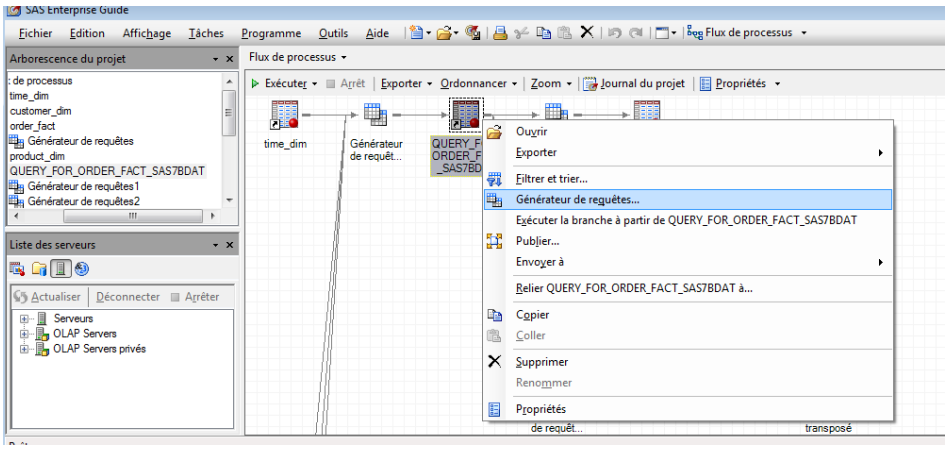
Transposer la table

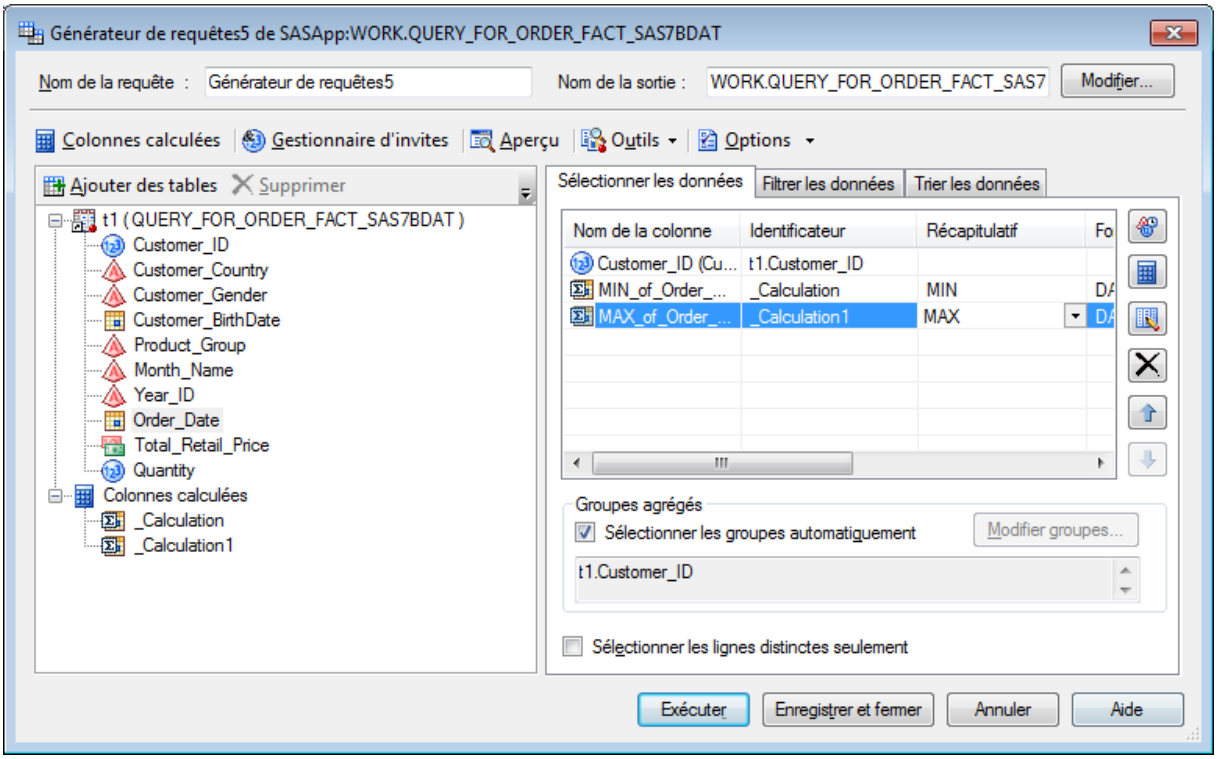

Transposer2 pour SASApp:WORK.QUERY_FOR_ORDER_FACT_SAS7BD_0002
Données
Options
Résultats
Propriétés
Options
Colonne d'origine
Nom :
Source
Libellé :
Libellé
Préfixe du nom de la colonne
Utiliser le préfixe :
Colonne
Permet d'indiquer un préfixe pour les en-têtes de toutes les colonnes de données de la table de sortie.
Aperçu du code
Exécuter
Enregistrer
Annuler
Aide


SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
Transposer2
Données d'entrée Code Journal Données de sortie
Modifier la tâche Filtrer et trier Générateur de requêtes Données Description Graphique Analyser Exporter Envoyer à
Customer_ID Source 2005 2004 2003 2006
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Prêt
Connexion : sasuser, localhost

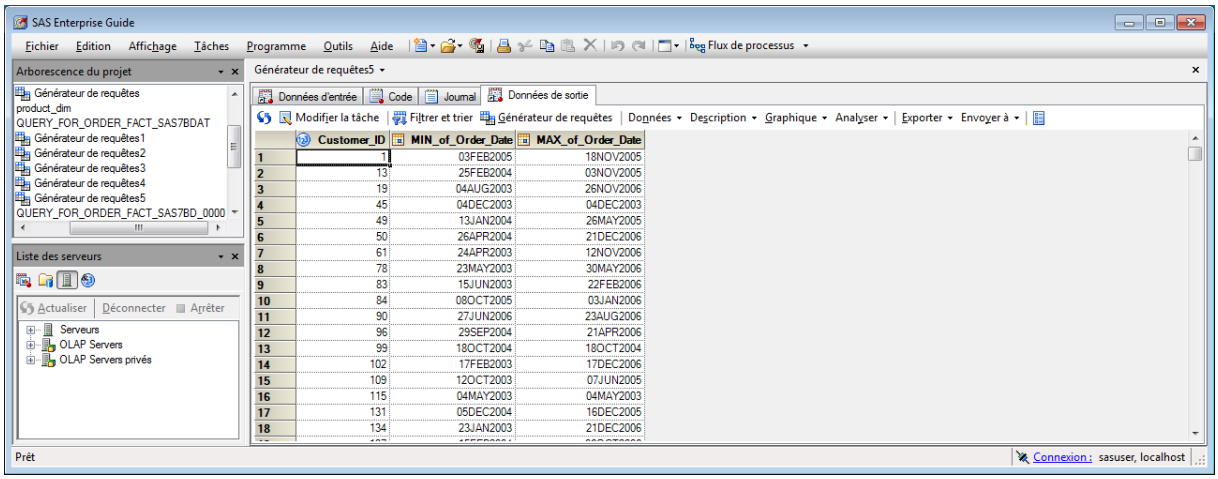
Création des colonnes de récence et d'ancienneté

Créer une requête sur la table de base.

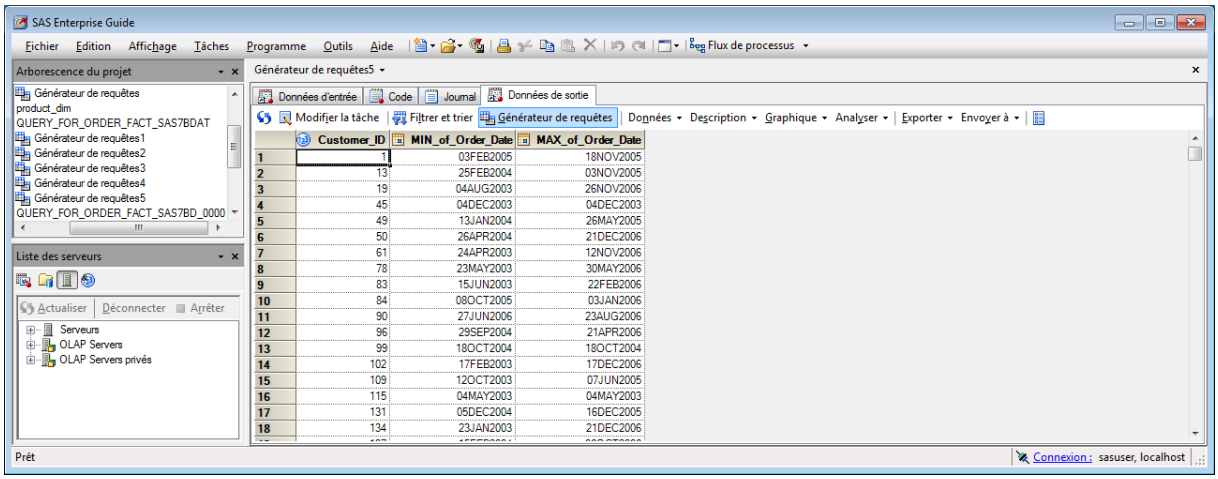


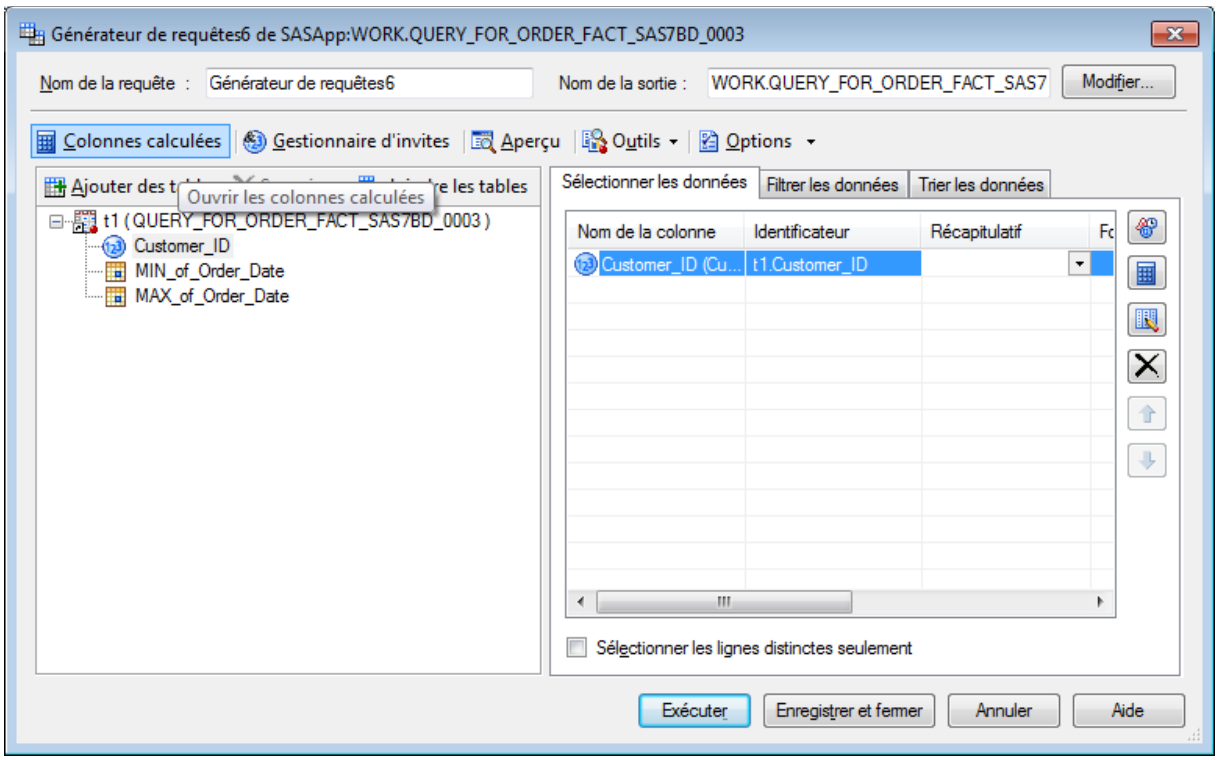


Sélectionner le minimum et le maximum de la date de commande par client.
Exécuter.

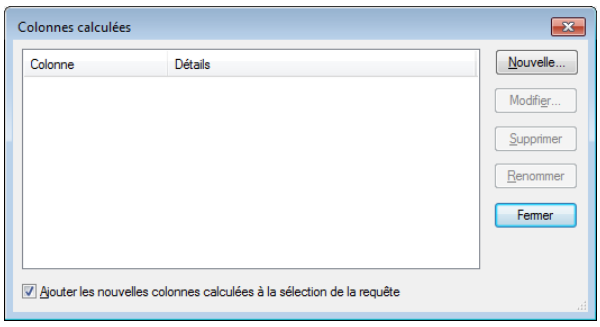


Créer une nouvelle requête sur cette table,

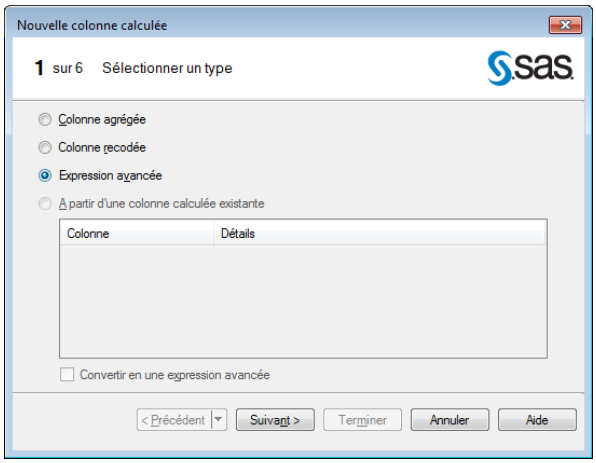




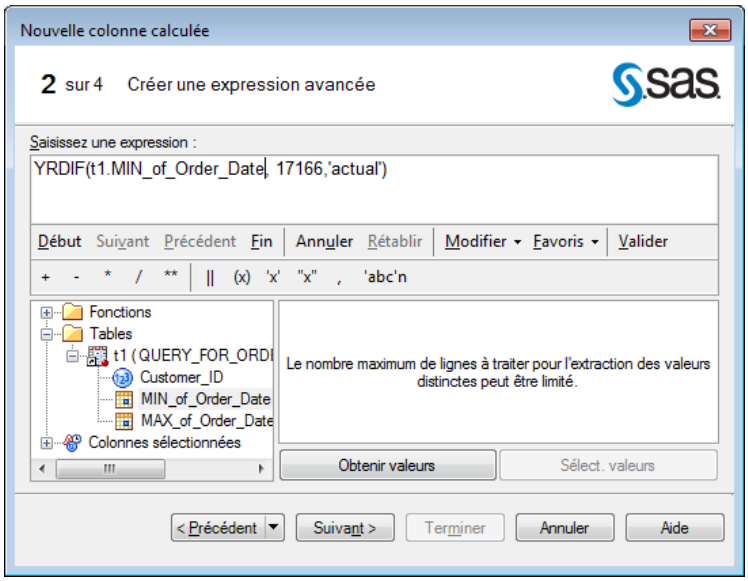
Sélectionner le numéro de client
Ajouter deux colonnes, la recense et l'ancienneté.
L'ancienneté est égale à la différence entre la plus ancienne commande et la date d'aujourd'hui, le premier janvier 2001, soit l'expression : YRDIF(t1.MIN_of_Order_Date, 17166,'actual')
La recense est égale à la différence entre la plus récente commande et la date d'aujourd'hui, le premier janvier 2001, soit l'expression : YRDIF(t1.MAX_of_Order_Date, 17166,'actual')



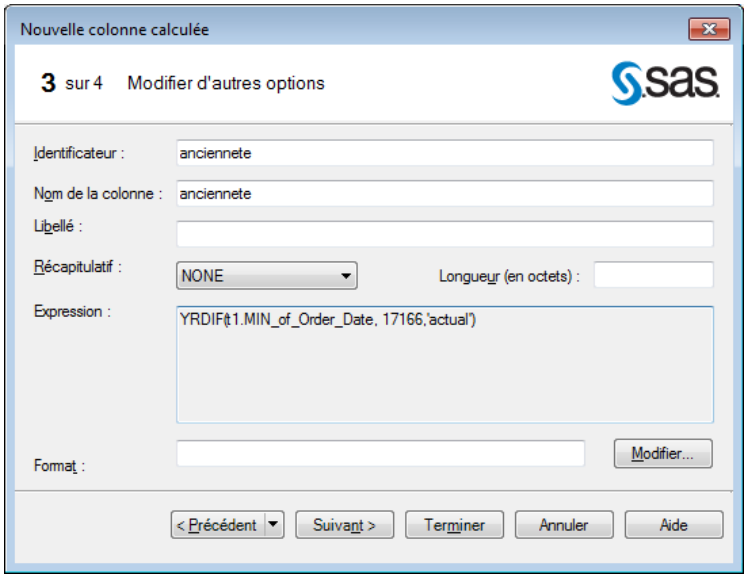
Nouvelle



Expression avancée, suivant

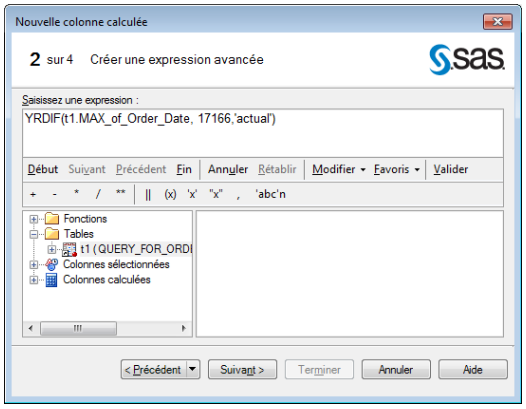


YRDIF(t1.MIN_of_Order_Date, 17166,'actual')
Suivant

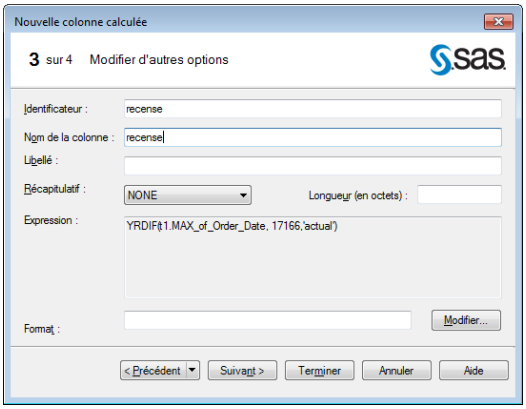


Ancienneté
Terminer

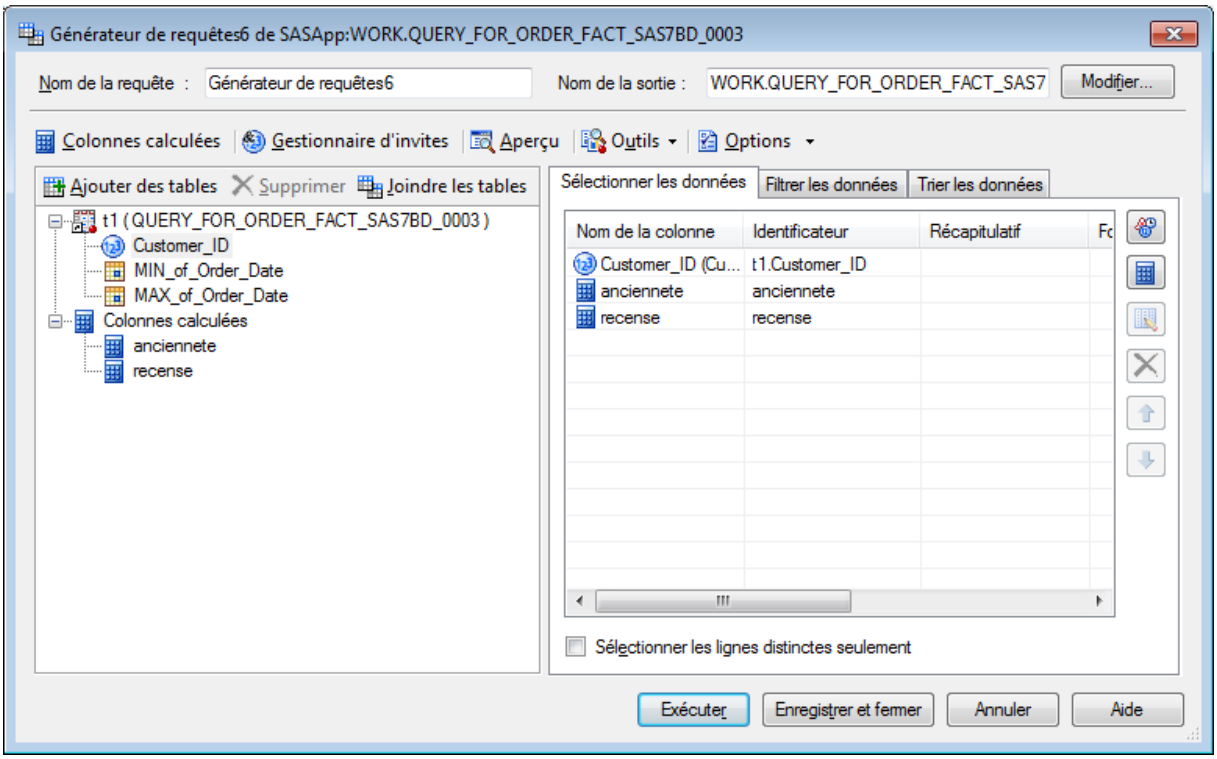
Nouvelle, Expression avancée, suivant



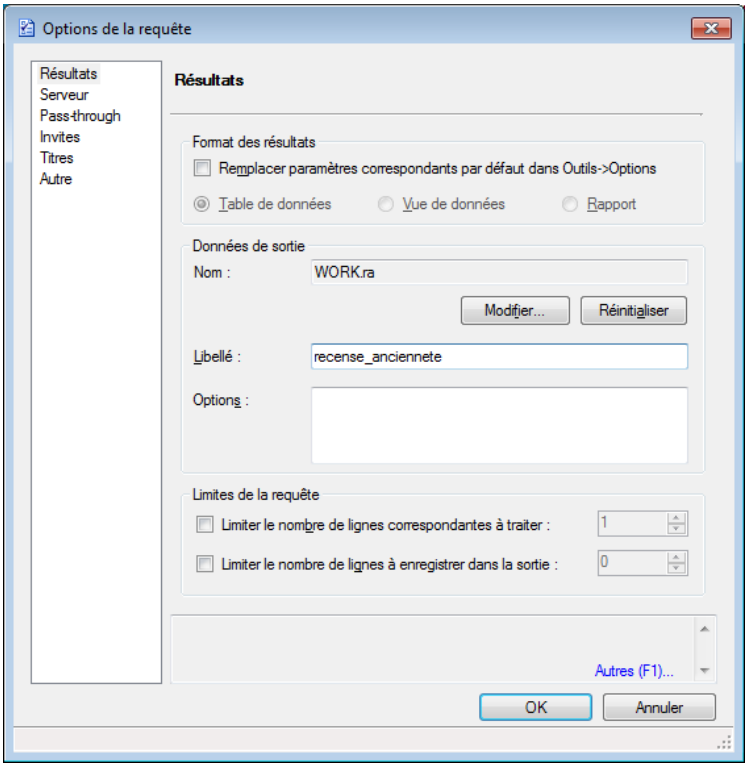
Suivant,



Terminer, Fermer

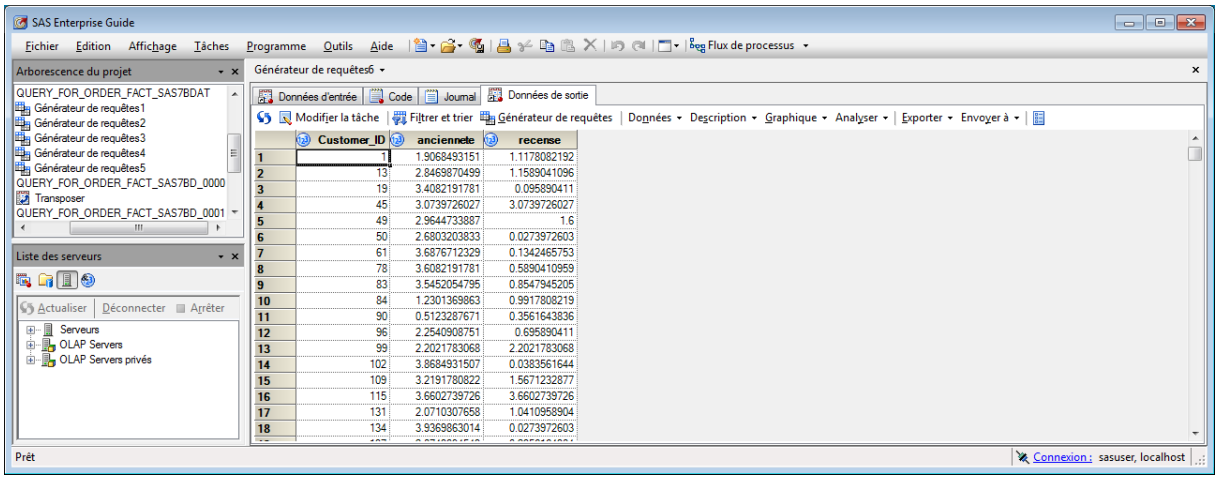


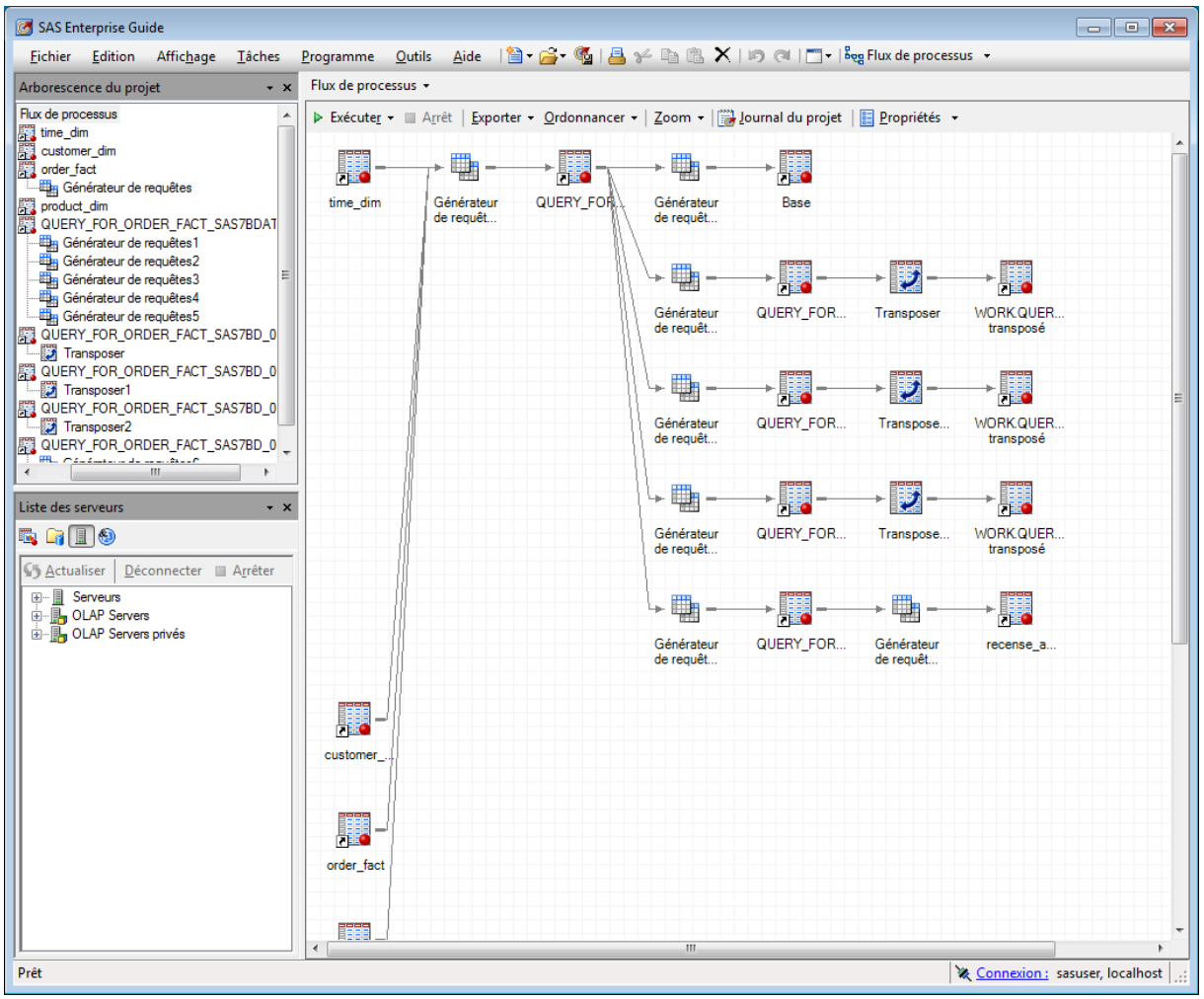
Dans les options de la table,



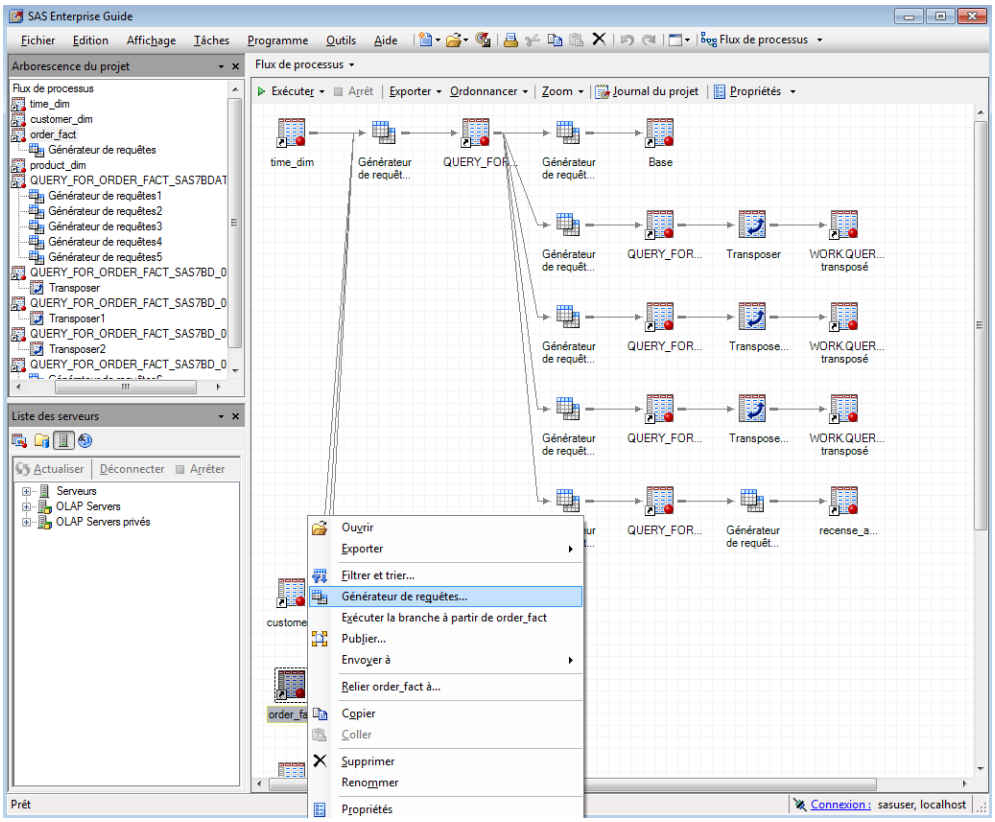
Renommer la table RA par exemple
OK

Exécuter

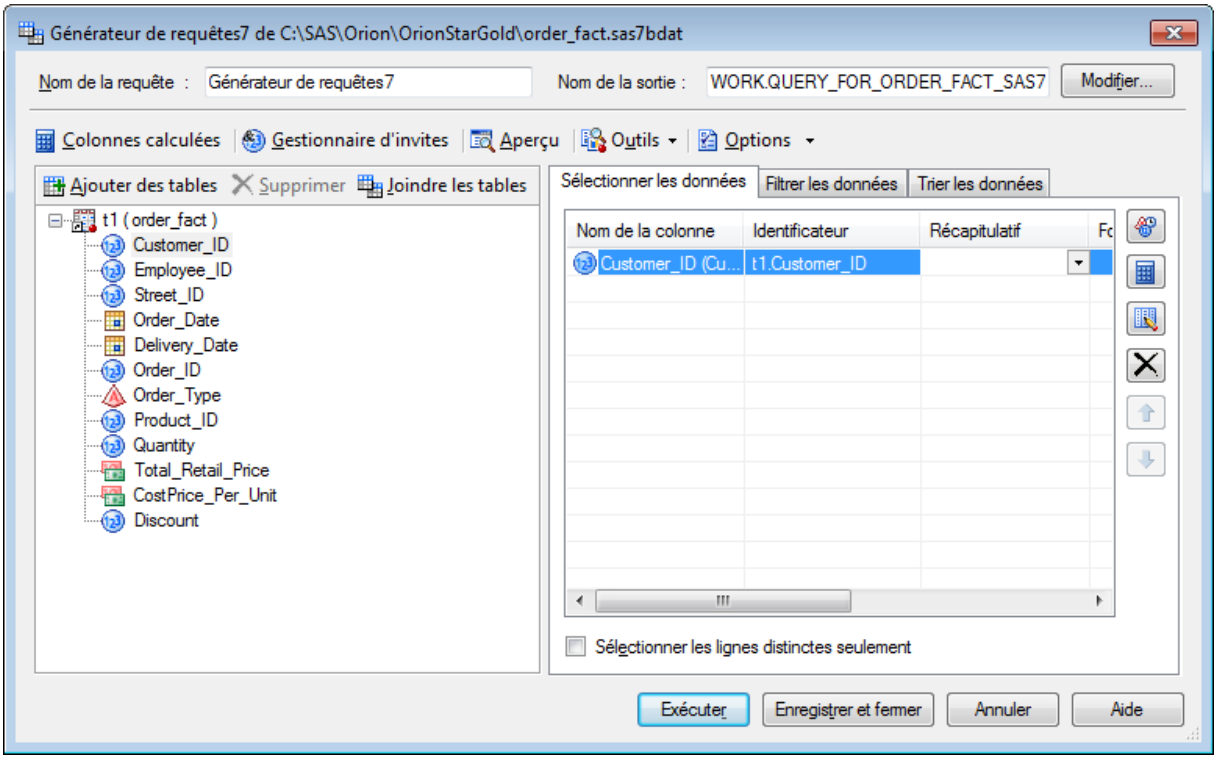

SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
QUERY_FOR_ORDER_FACT_SAS7BDAT
Générateur de requêtes1
Générateur de requêtes2
Générateur de requêtes3
Générateur de requêtes4
Générateur de requêtes5
QUERY_FOR_ORDER_FACT_SAS7BD_0000
Transposer
QUERY_FOR_ORDER_FACT_SAS7BD_0001
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Générateur de requêtes6
Données d'entrée Code Journal Données de sortie
Modifier la tâche Filtrer et trier Générateur de requêtes Données Description Graphique Analyser Exporter Envoyer à
Customer_ID ancienneté recense
1 1 1.9068493151 1.1178082192
2 13 2.8469870499 1.1589041096
3 19 3.4082191781 0.095890411
4 45 3.0739726027 3.0739726027
5 49 2.9644733887 1.6
6 50 2.6803203833 0.0273972603
7 61 3.6876712329 0.1342465753
8 78 3.6082191781 0.5890410959
9 83 3.5452054795 0.8547945205
10 84 1.2301369863 0.9917808219
11 90 0.5123287671 0.3561643836
12 96 2.2540908751 0.695890411
13 99 2.2021783068 2.2021783068
14 102 3.8684931507 0.0383561644
15 109 3.2191780822 1.5671232877
16 115 3.6602739726 3.6602739726
17 131 2.0710307658 1.0410958904
18 134 3.9369863014 0.0273972603
Prêt
Connexion : sasuser, localhost


SAS Enterprise Guide
Fichier Edition Affichage Tâches Programme Outils Aide
Flux de processus
Arborescence du projet
Flux de processus
time_dim
customer_dim
order_fact
Générateur de requêtes
product_dim
QUERY_FOR_ORDER_FACT_SAS7BDAT
Générateur de requêtes1
Générateur de requêtes2
Générateur de requêtes3
Générateur de requêtes4
Générateur de requêtes5
QUERY_FOR_ORDER_FACT_SAS7BD_0
Transposer
QUERY_FOR_ORDER_FACT_SAS7BD_0
Transposer1
QUERY_FOR_ORDER_FACT_SAS7BD_0
Transposer2
QUERY_FOR_ORDER_FACT_SAS7BD_0
Liste des serveurs
Actualiser Déconnecter Arrêter
Serveurs
OLAP Servers
OLAP Servers privés
Flux de processus
Exécuter Arrêt Exporter Ordonnancer Zoom Journal du projet Propriétés
time_dim
Générateur de requêt...
QUERY_FOR...
Générateur de requêt...
Base
Générateur de requêt...
QUERY_FOR...
Transposer
WORK.QUER... transposé
Générateur de requêt...
QUERY_FOR...
Transpose...
WORK.QUER... transposé
Générateur de requêt...
QUERY_FOR...
Transpose...
WORK.QUER... transposé
Générateur de requêt...
QUERY_FOR...
Générateur de requêt...
recense_a...
customer_...
order_fact
Prêt
Connexion : sasuser, localhost

Création de la colonne Target

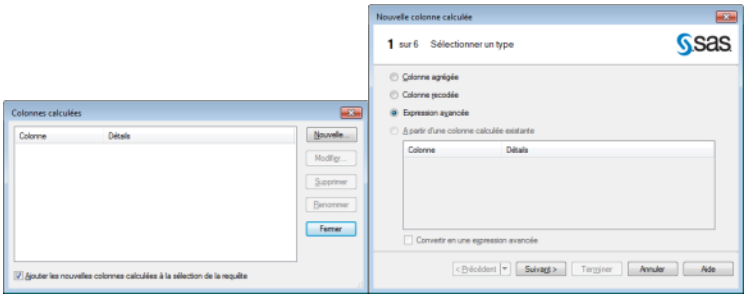


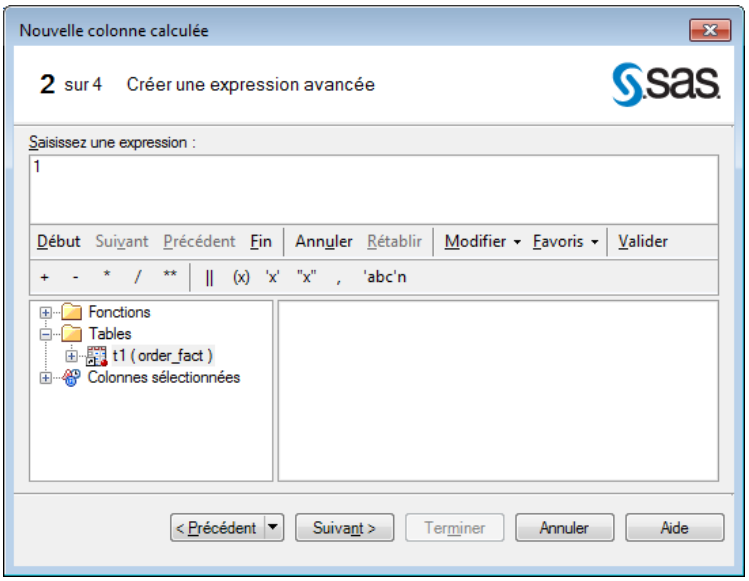
Créer une requête sur la table order_fact



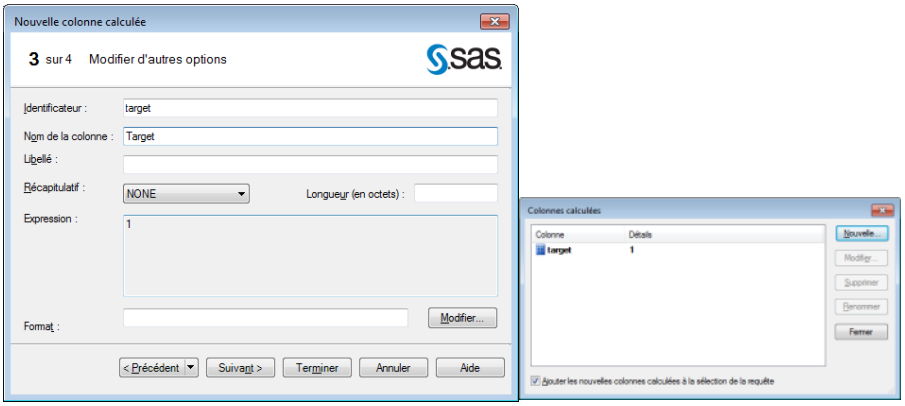
Sélectionner la colonne customer_id

Créer une colonne calculée dont l'expression est « 1 »

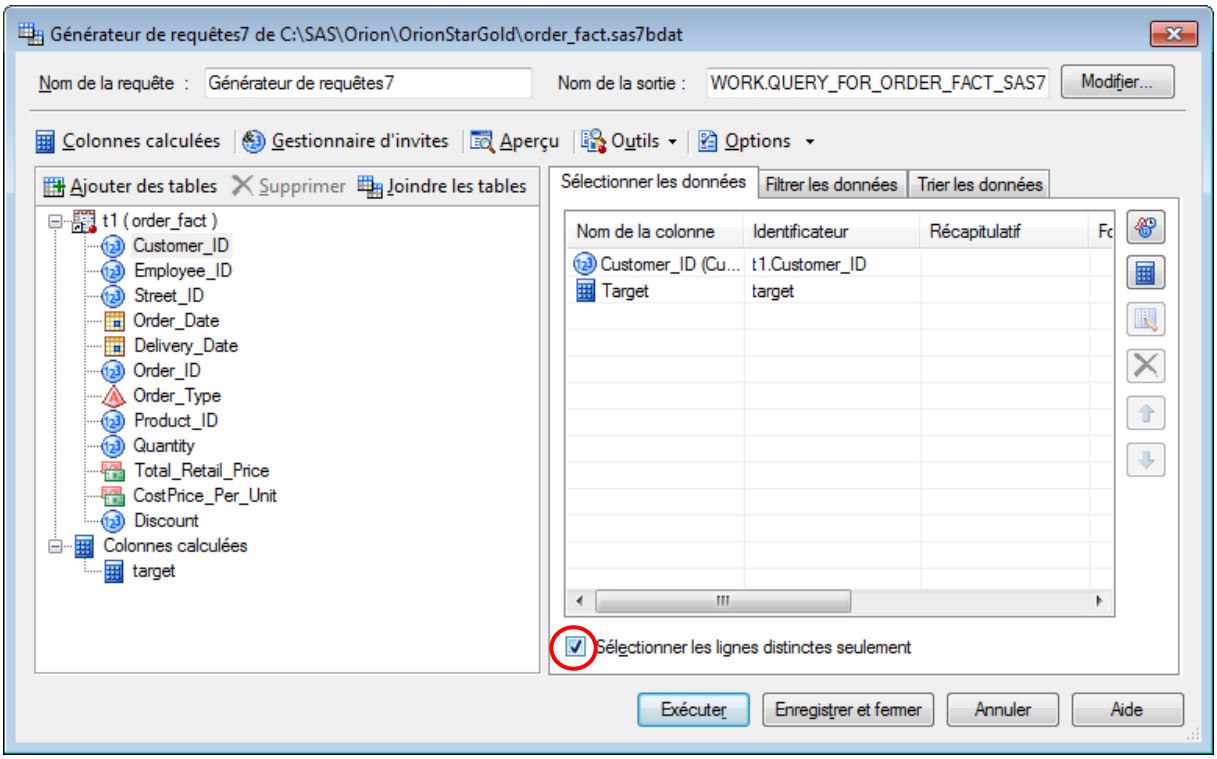




Taper « 1 »
Suivant

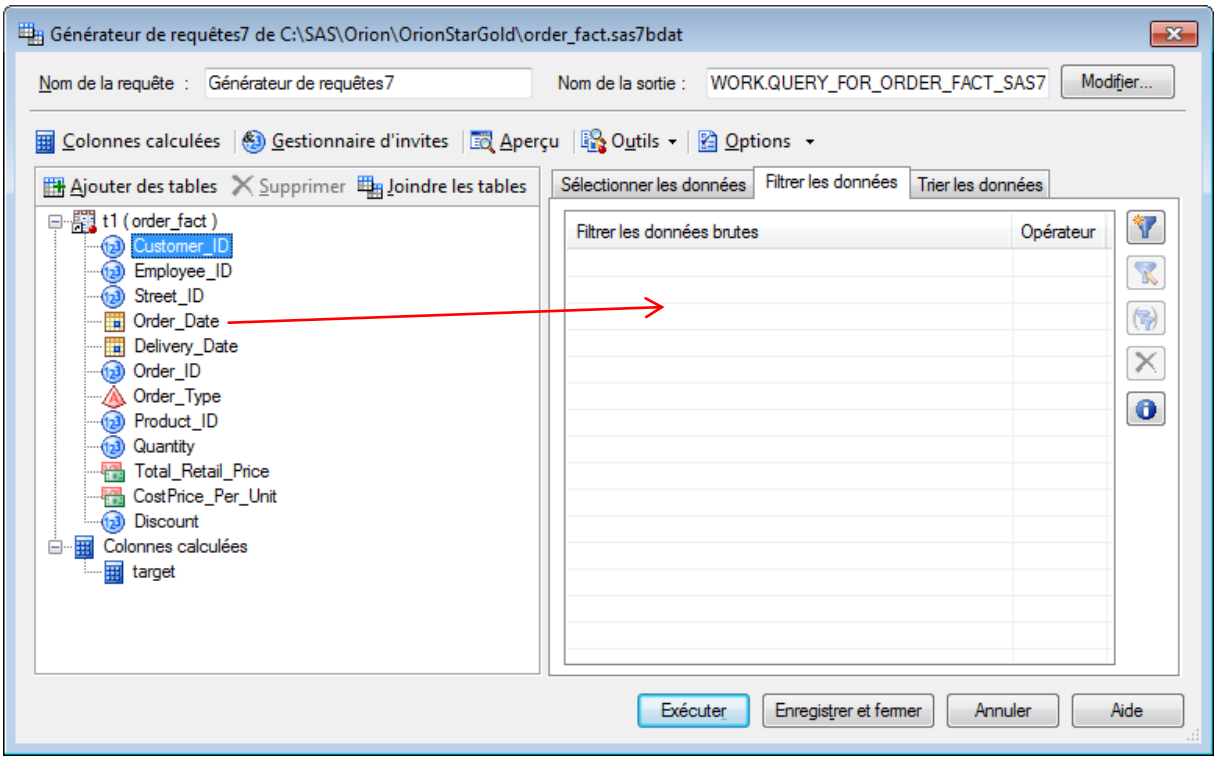


Renommer la colonne « Target »
Suivant, Fermer



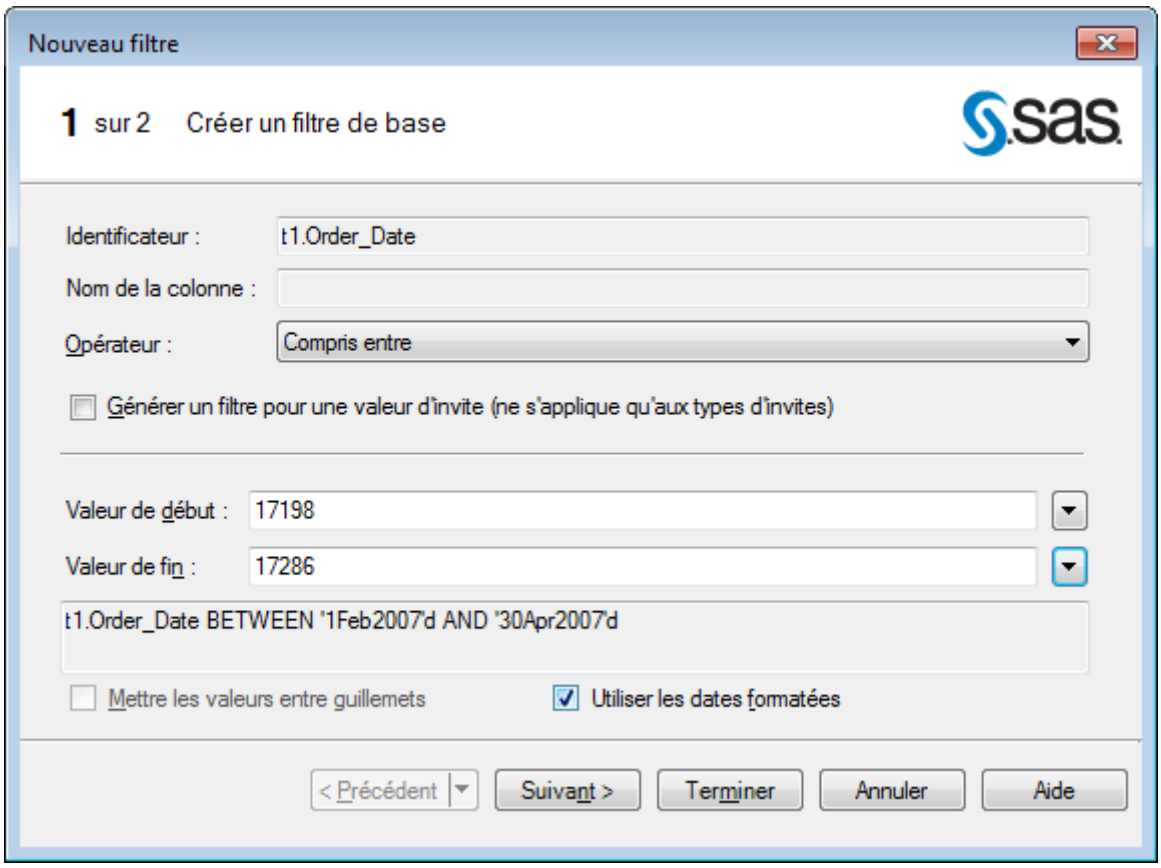
Sélectionner les valeurs distinctes seulement,

Dans « filtrer les données », sélectionner la colonne order_date

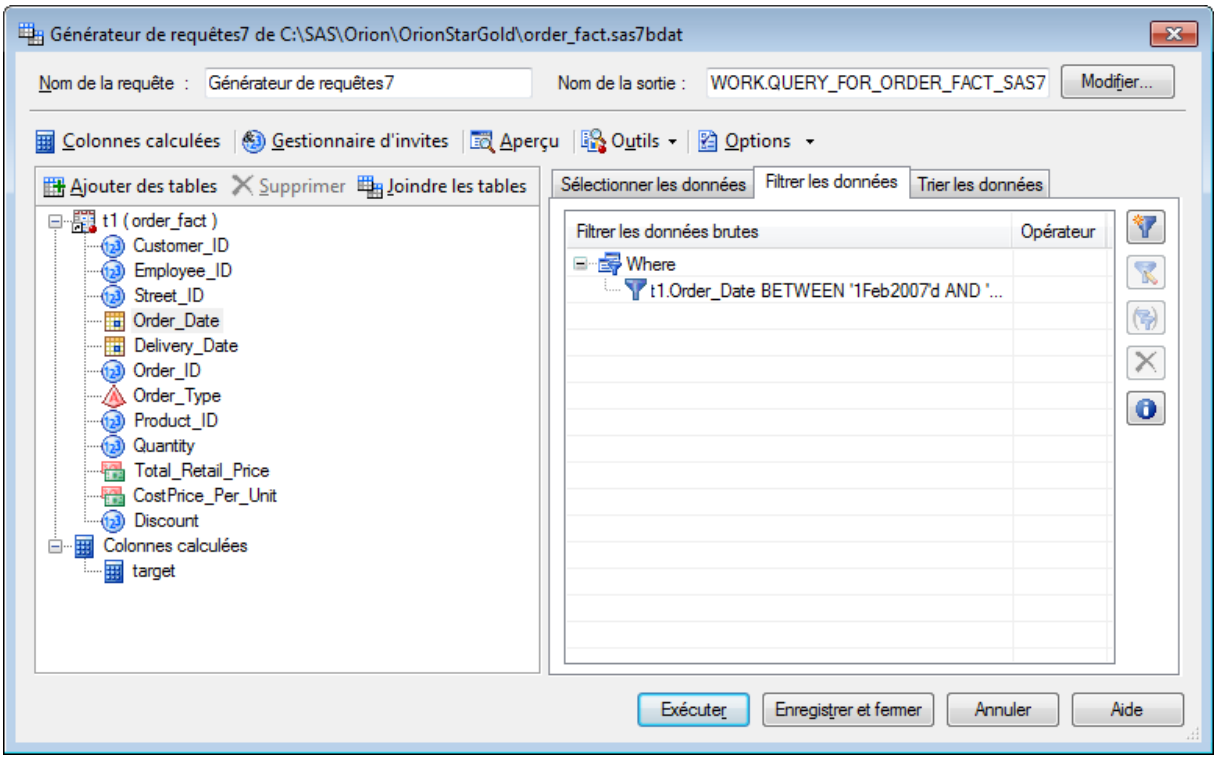


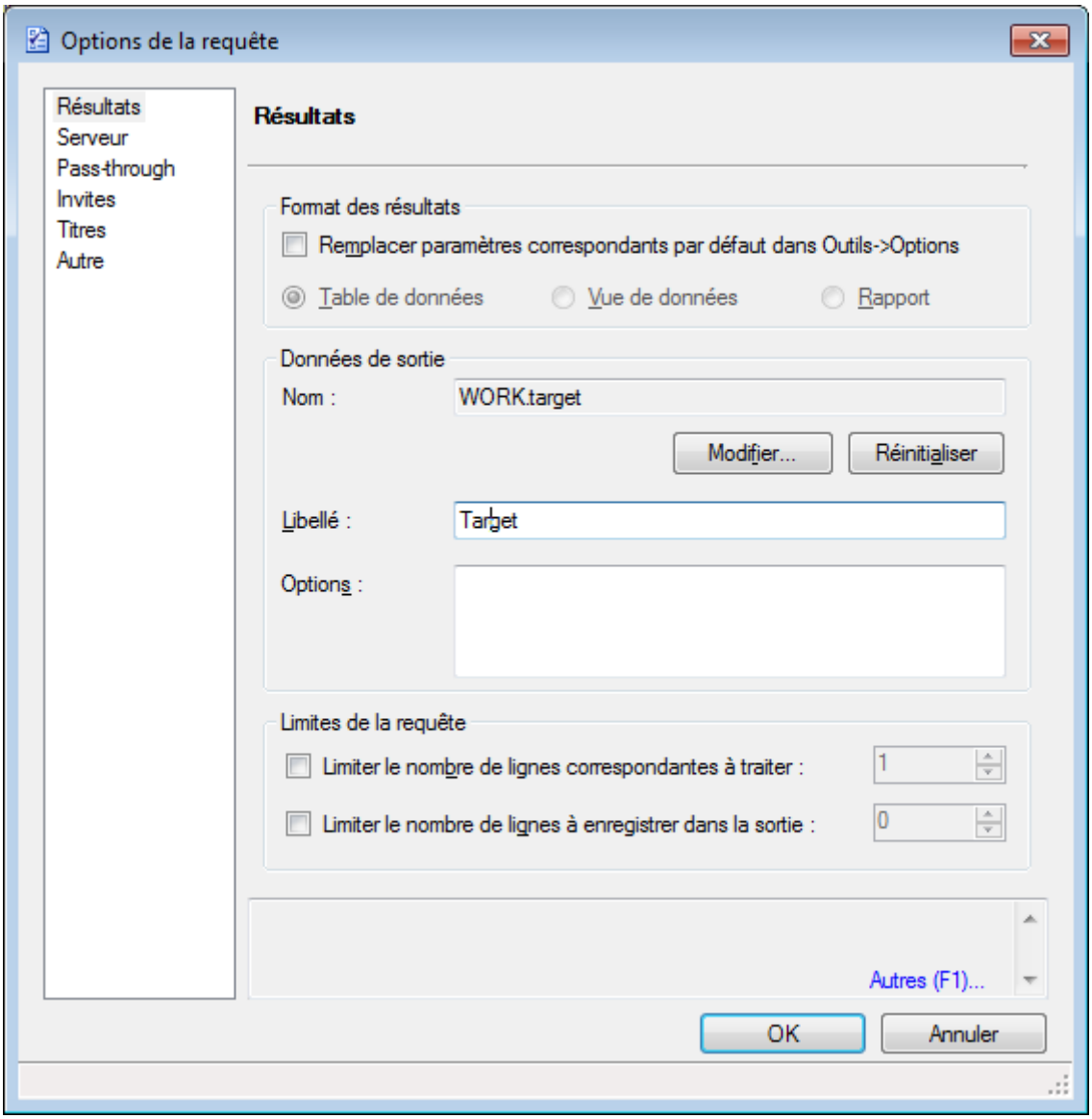
Exécuter

Créer un filtre de telle sorte que la date soit comprise entre le ‘01Feb2007’d et le ‘30Apr2007’d

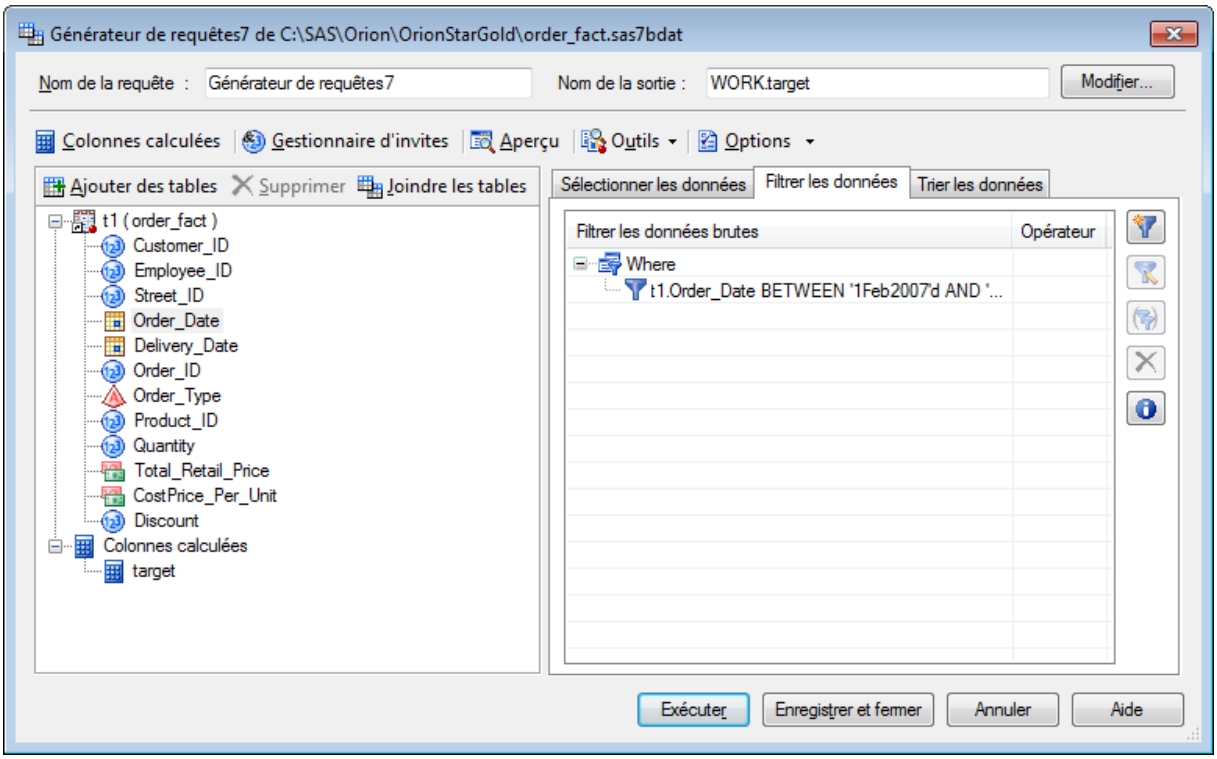


Terminer

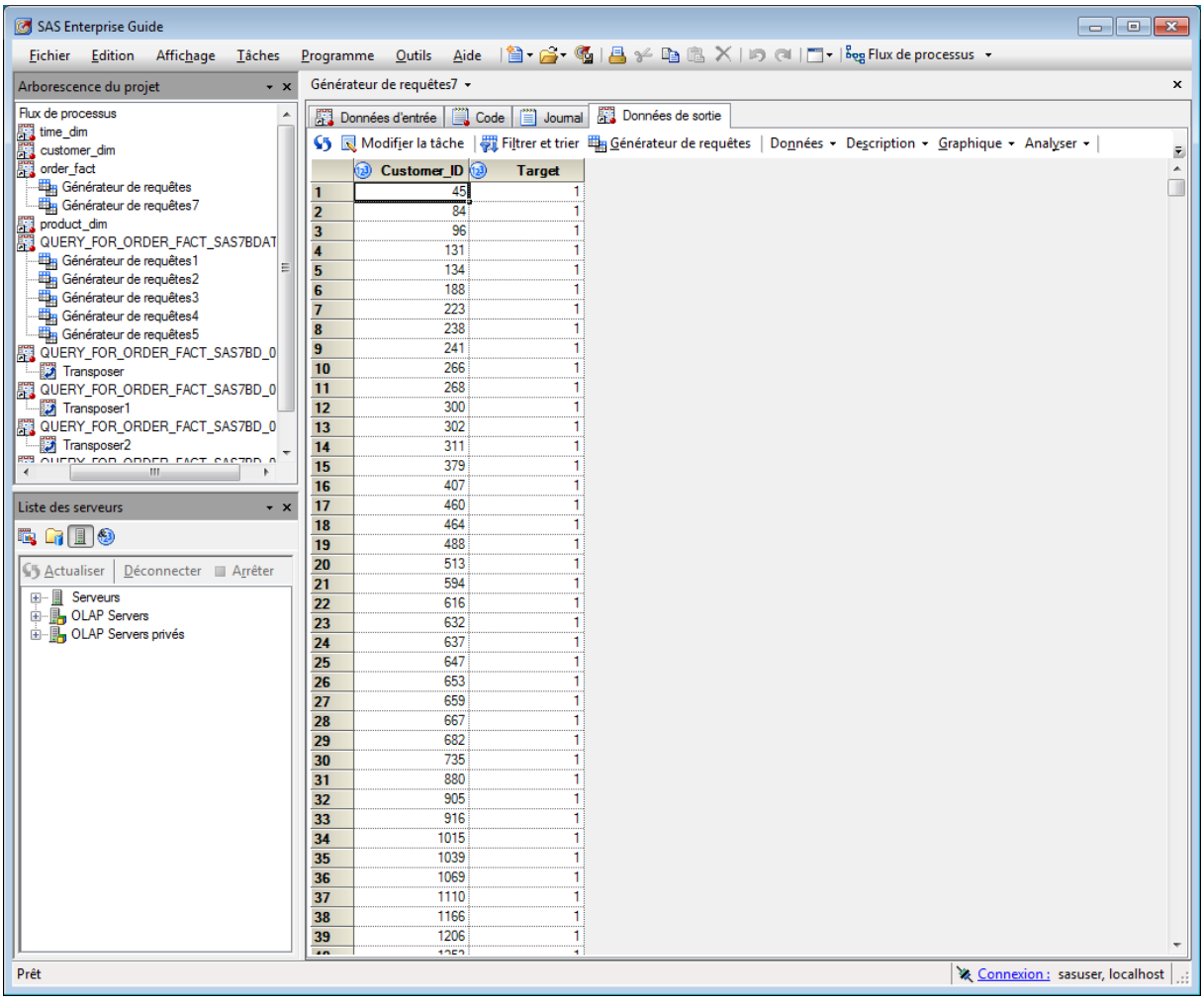

Générateur de requêtes7 de C:\SAS\Orion\OrionStarGold\order_fact.sas7bdat
Nom de la requête :
Générateur de requêtes7
Nom de la sortie :
WORK.QUERY_FOR_ORDER_FACT_SAS7
Modifier...
Colonnes calculées
Gestionnaire d'invites
Aperçu
Outils
Options
Ajouter des tables
Supprimer
Joindre les tables
t1 (order_fact)
Customer_ID
Employee_ID
Street_ID
Order_Date
Delivery_Date
Order_ID
Order_Type
Product_ID
Quantity
Total_Retail_Price
CostPrice_Per_Unit
Discount
Colonnes calculées
target
Sélectionner les données
Filtrer les données
Trier les données
Filtrer les données brutes
Opérateur
Where
t1.Order_Date BETWEEN '1Feb2007'd AND '...
Exécuter
Enregistrer et fermer
Annuler
Aide



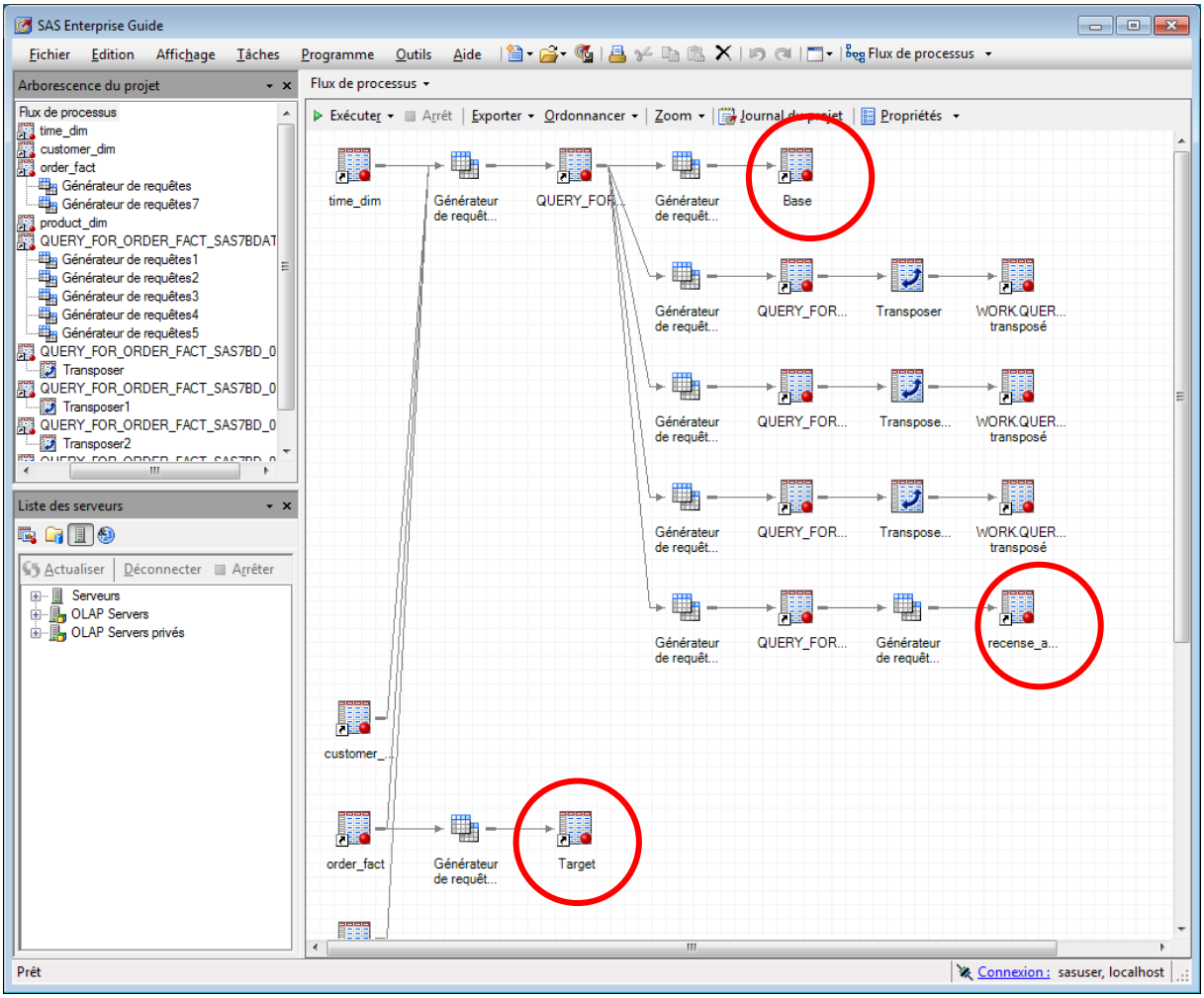
Dans les options de la requête, renommer la table en target par exemple



Exécuter la requête

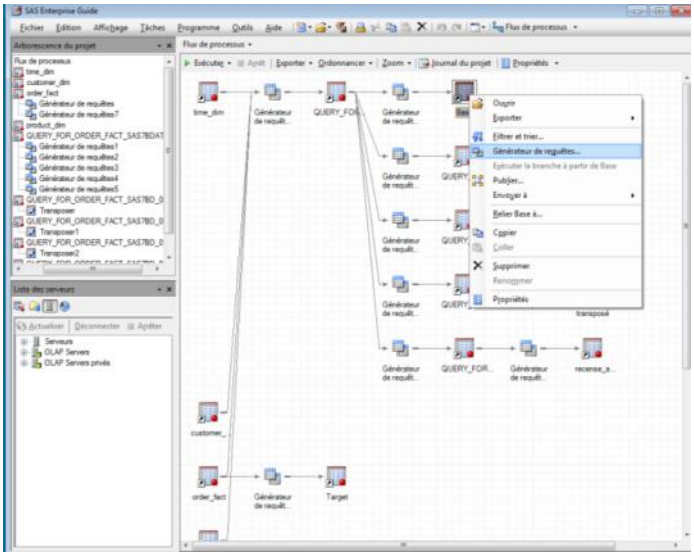


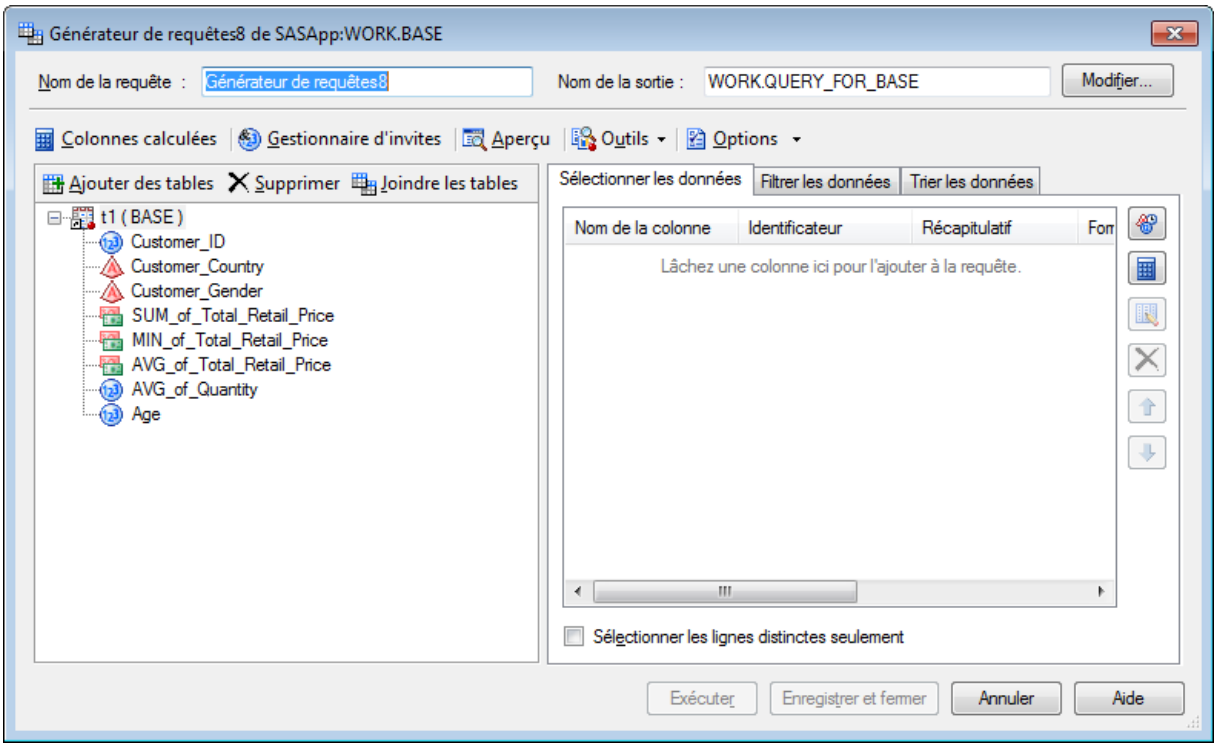
Création de la table d'apprentissage



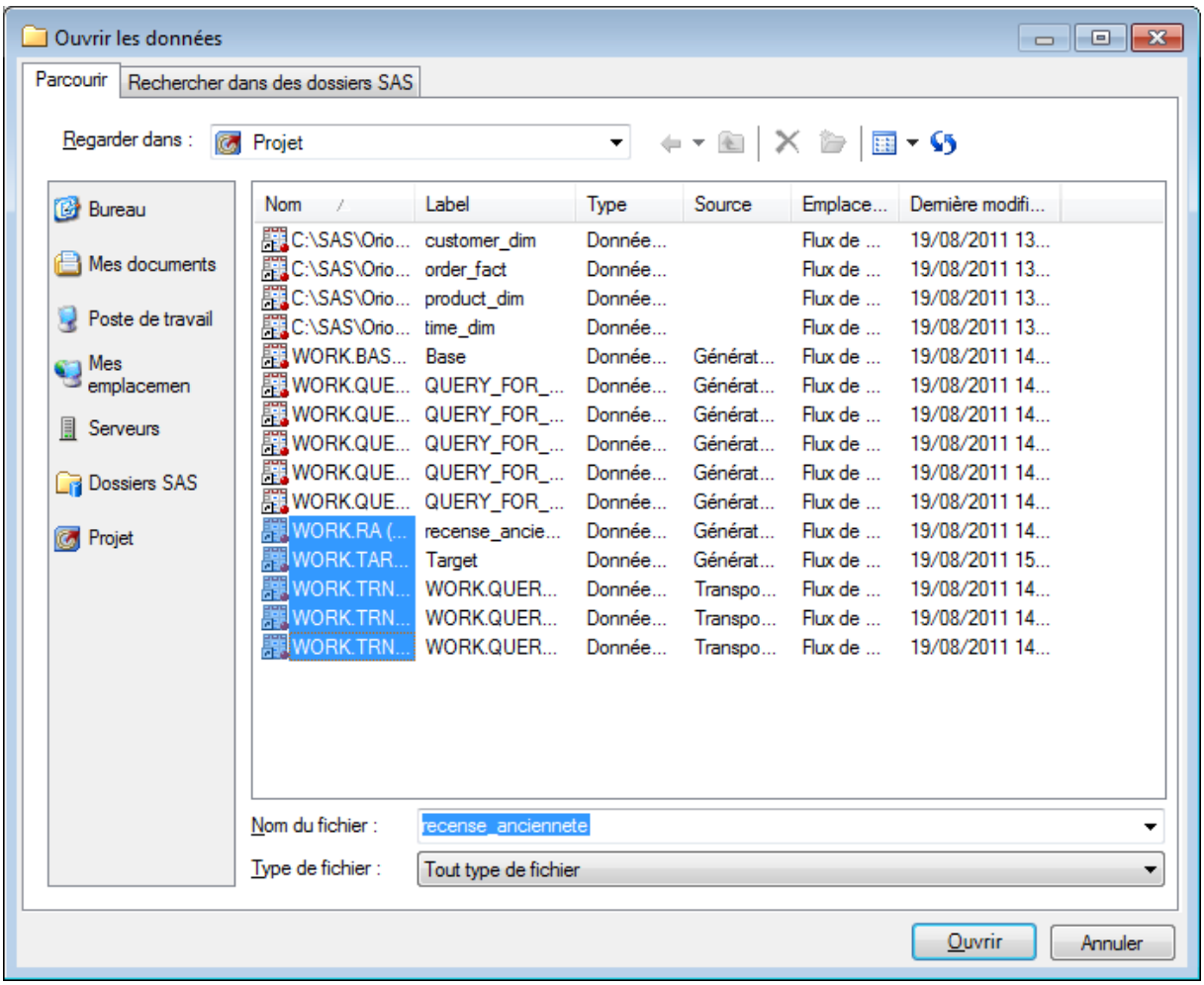
Lire et noter le nom de toutes les tables intermédiaires (fin de ligne de processus).

Depuis la fenêtre du flux de processus, créer une requête sur la première table intermédiaire : la table Base ici.



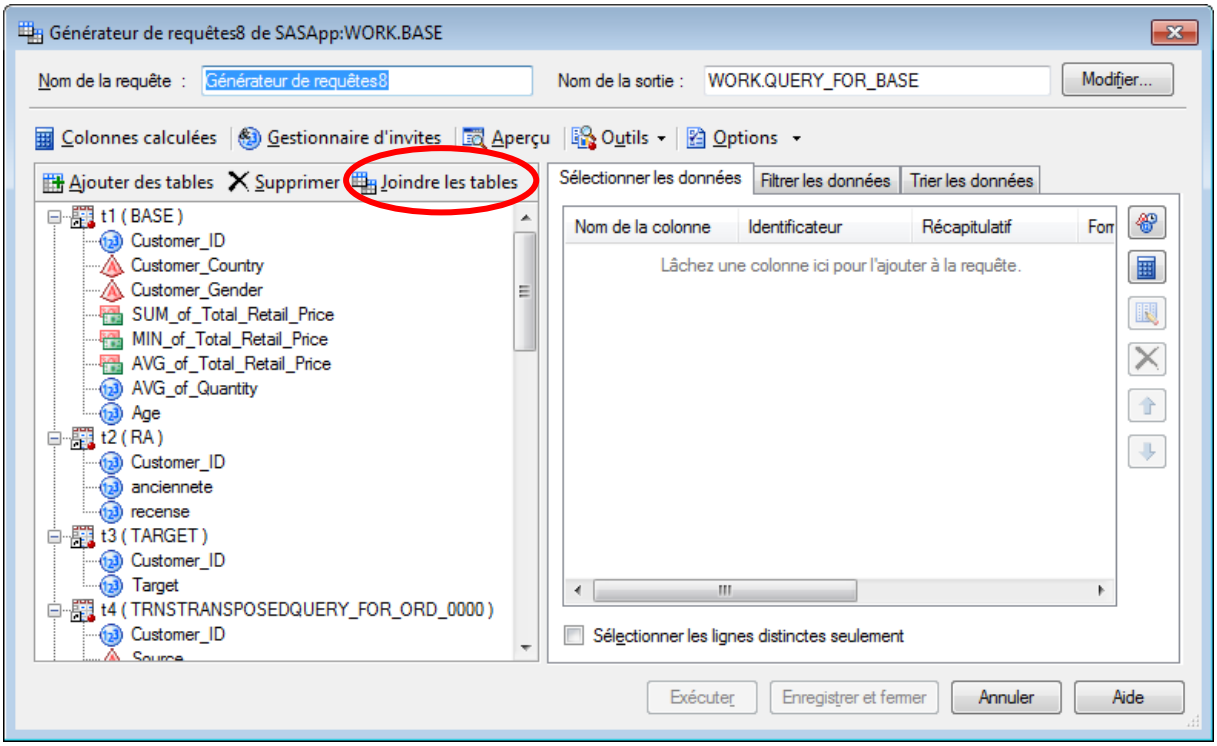


Ajouter les tables



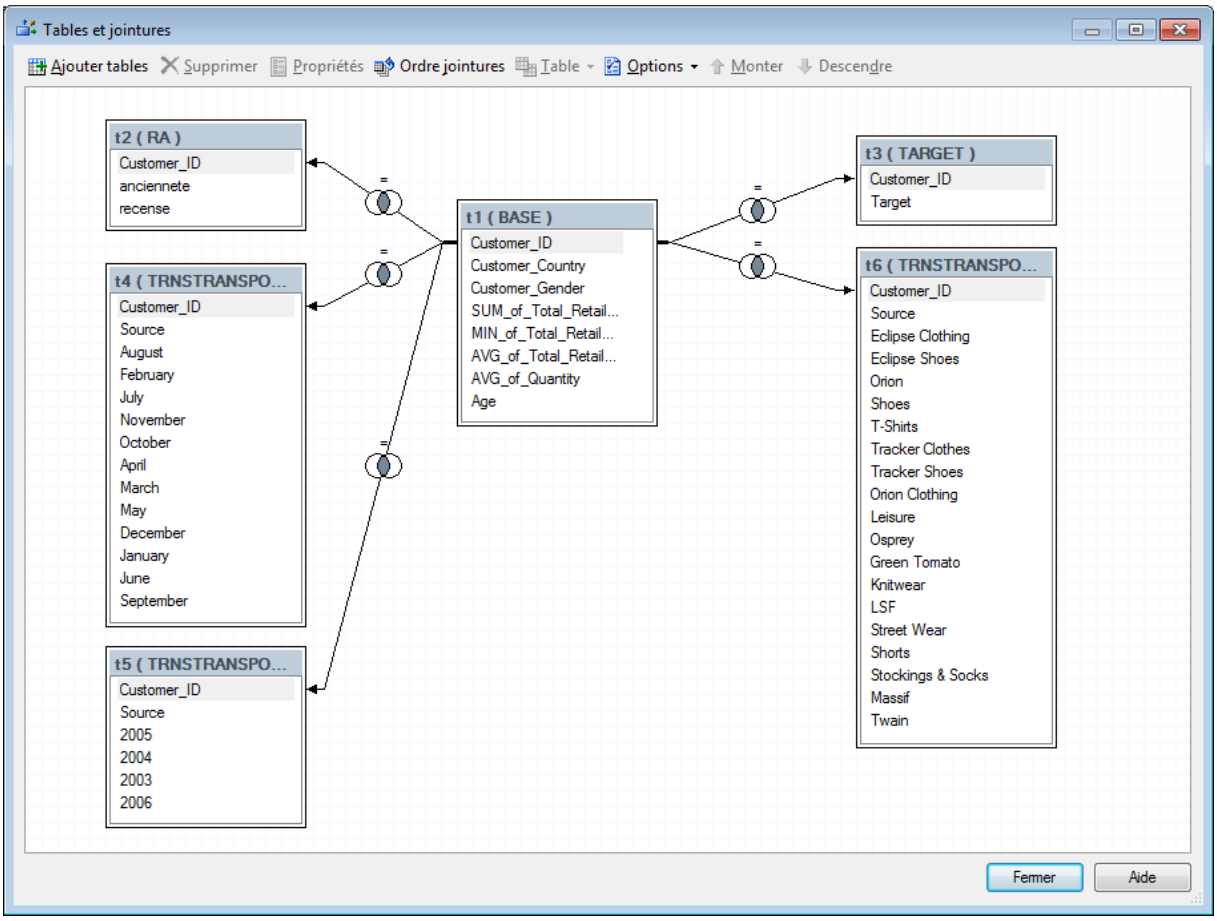
Depuis le projet

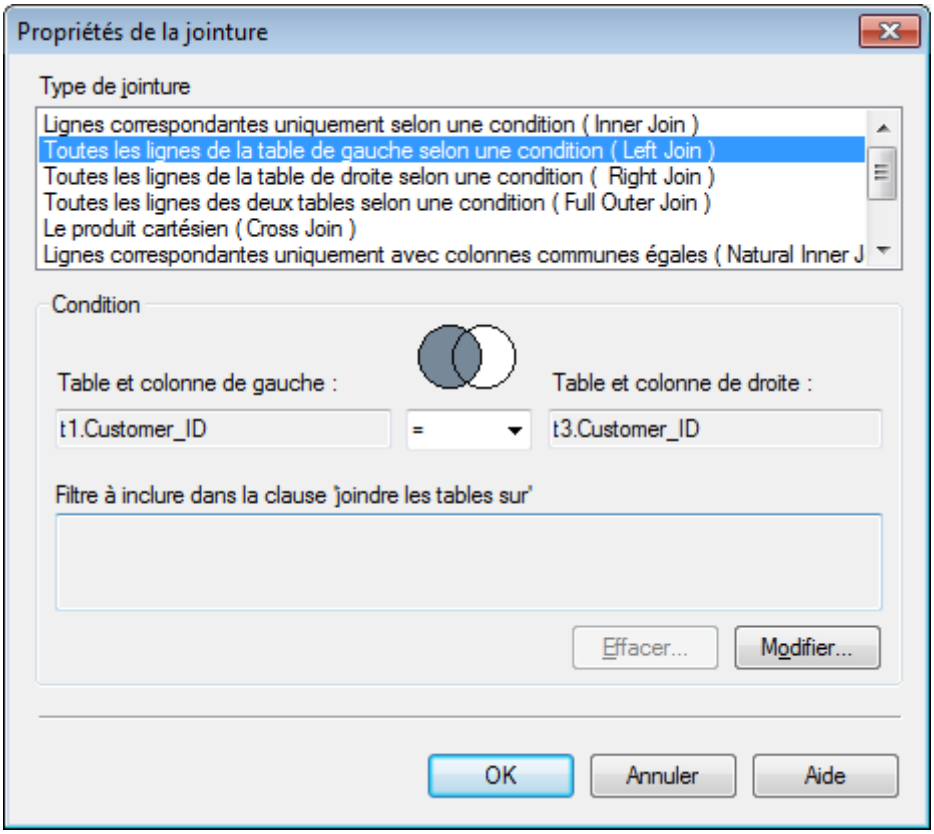
Sélectionner les tables intermédiaires : Les 3 tables transposées, la table de l'ancienneté et de la recense, ainsi que celle de la Target.



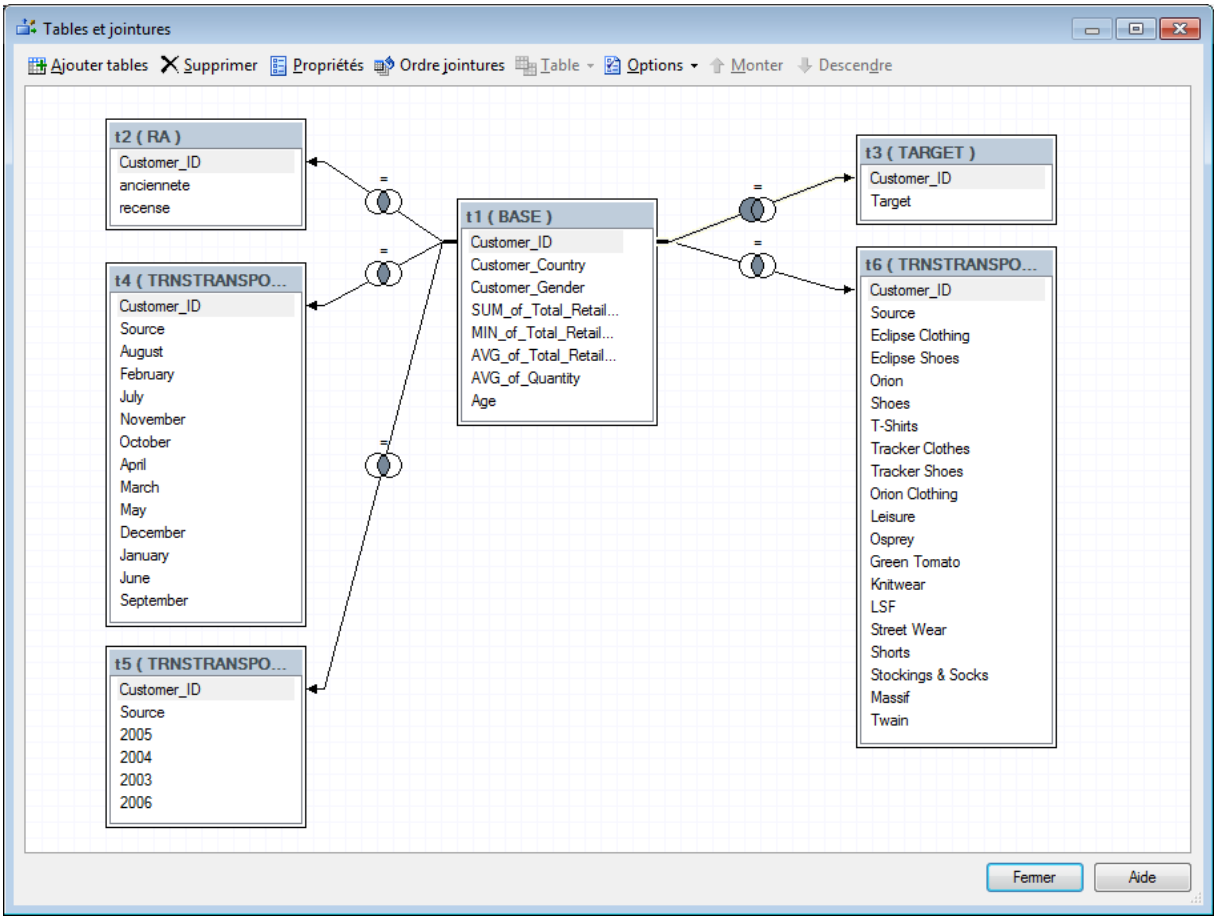
Dans l'onglet jointure (Si vous n'avez pas l'onglet jointure, agrandir la fenêtre.)

Modifier la jointure entre la table intermédiaire où l'on a calculé la cible (target) et celle avec le pays, le sexe, etc.


Tables et jointures
Ajouter tables
Supprimer
Propriétés
Ordre jointures
Table
Options
Monter
Descendre
t2 (RA)
Customer_ID
ancienneté
recense
t4 (TRNSTRANSPO...
Customer_ID
Source
August
February
July
November
October
April
March
May
December
January
June
September
t5 (TRNSTRANSPO...
Customer_ID
Source
2005
2004
2003
2006
t1 (BASE)
Customer_ID
Customer_Country
Customer_Gender
SUM_of_Total_Retail...
MIN_of_Total_Retail...
AVG_of_Total_Retail...
AVG_of_Quantity
Age
t3 (TARGET)
Customer_ID
Target
t6 (TRNSTRANSPO...
Customer_ID
Source
Eclipse Clothing
Eclipse Shoes
Orion
Shoes
T-Shirts
Tracker Clothes
Tracker Shoes
Orion Clothing
Leisure
Osprey
Green Tomato
Knitwear
LSF
Street Wear
Shorts
Stockings & Socks
Massif
Twain
Fermer
Aide


Propriétés de la jointure
Type de jointure
Lignes correspondantes uniquement selon une condition (Inner Join)
Toutes les lignes de la table de gauche selon une condition (Left Join)
Toutes les lignes de la table de droite selon une condition (Right Join)
Toutes les lignes des deux tables selon une condition (Full Outer Join)
Le produit cartésien (Cross Join)
Lignes correspondantes uniquement avec colonnes communes égales (Natural Inner J
Condition
Table et colonne de gauche :
Table et colonne de droite :
t1.Customer_ID
=
t3.Customer_ID
Filtre à inclure dans la clause 'joindre les tables sur'
Effacer...
Modifier...
OK
Annuler
Aide

Sélectionner toutes les lignes de la table de gauche



Fermer

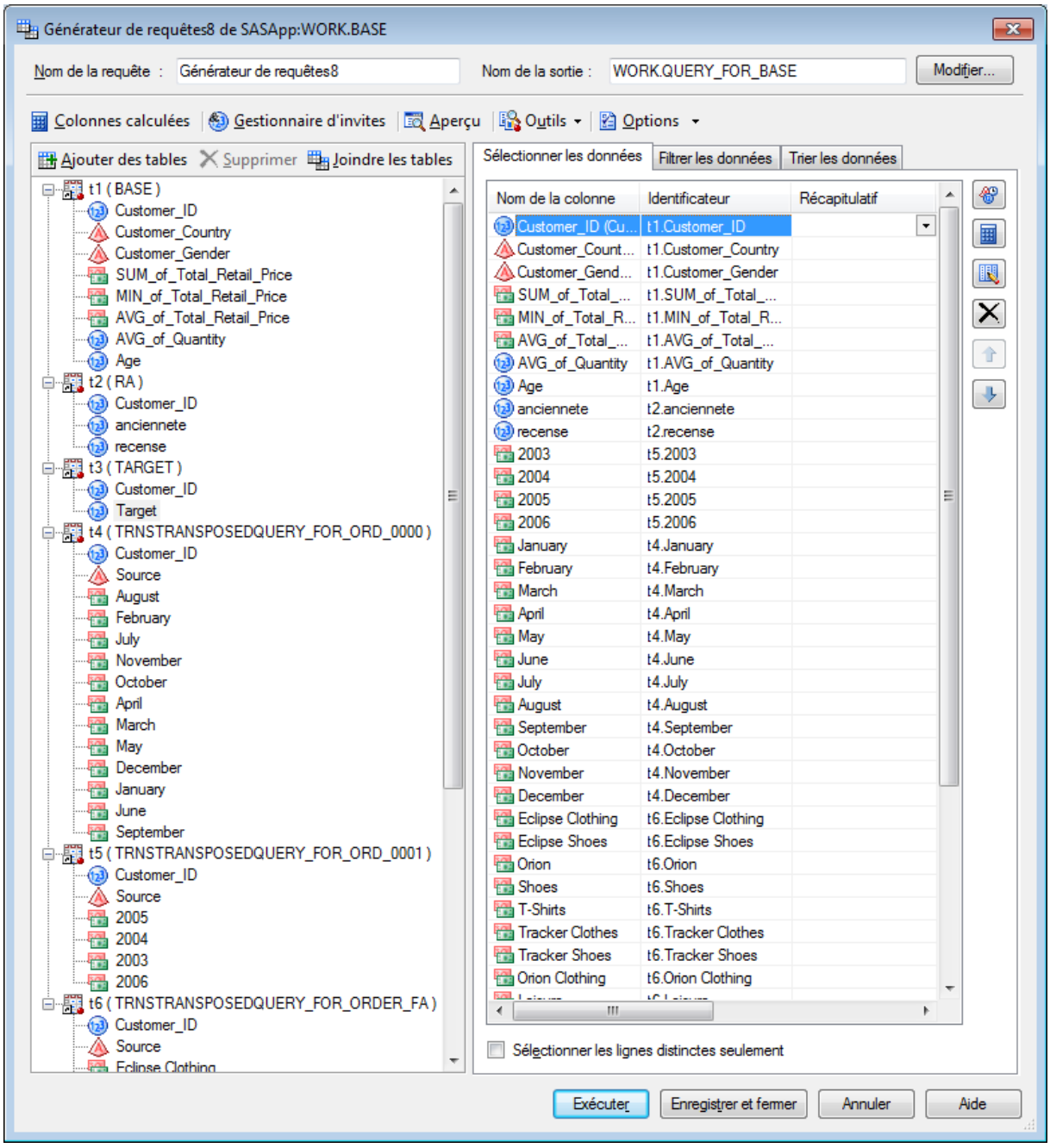
Sélectionner les colonnes

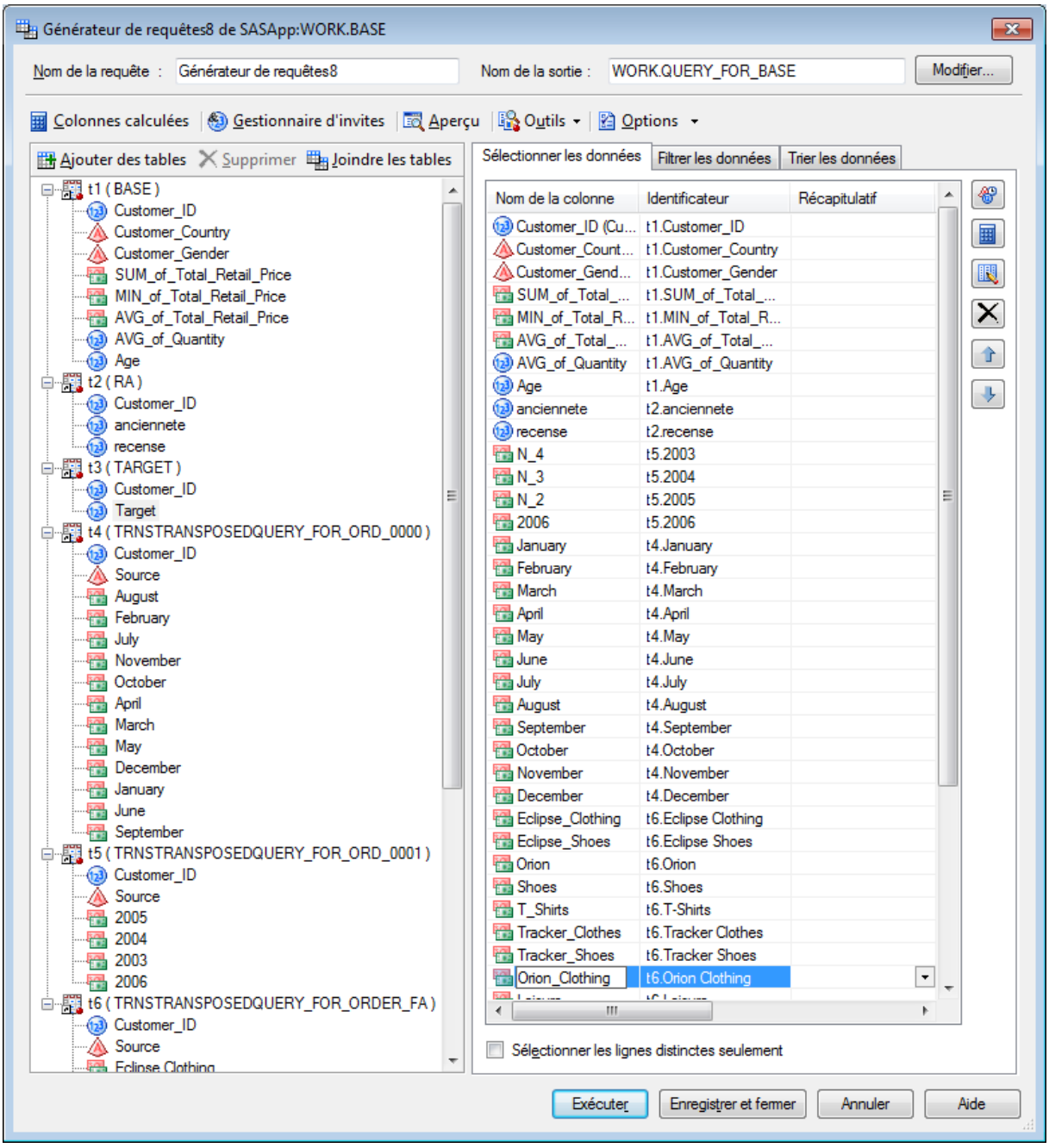
Toutes les colonnes de la table intermédiaire de base

La colonne de la récence et celle de l'ancienneté

Pour toutes les tables « transposées » sélectionner toutes les colonnes, sauf les deux premières,

La colonne Target


Générateur de requêtes8 de SASApp:WORK.BASE
Nom de la requête : Générateur de requêtes8
Nom de la sortie : WORK.QUERY_FOR_BASE
Modifier...
Colonnes calculées
Gestionnaire d'invites
Aperçu
Outils
Options
Ajouter des tables
Supprimer
Joindre les tables
t1 (BASE)
Customer_ID
Customer_Country
Customer_Gender
SUM_of_Total_Retail_Price
MIN_of_Total_Retail_Price
AVG_of_Total_Retail_Price
AVG_of_Quantity
Age
t2 (RA)
Customer_ID
anciennete
recense
t3 (TARGET)
Customer_ID
Target
t4 (TRNSTRANSPOSEDQUERY_FOR_ORD_0000)
Customer_ID
Source
August
February
July
November
October
April
March
May
December
January
June
September
t5 (TRNSTRANSPOSEDQUERY_FOR_ORD_0001)
Customer_ID
Source
2005
2004
2003
2006
t6 (TRNSTRANSPOSEDQUERY_FOR_ORDER_FA)
Customer_ID
Source
Sélectionner les données
Filtrer les données
Trier les données
Nom de la colonne
Identificateur
Récapitulatif
Customer_ID (Cu...
t1.Customer_ID
Customer_Count...
t1.Customer_Country
Customer_Gend...
t1.Customer_Gender
SUM_of_Total_...
t1.SUM_of_Total_...
MIN_of_Total_R...
t1.MIN_of_Total_R...
AVG_of_Total_...
t1.AVG_of_Total_...
AVG_of_Quantity
t1.AVG_of_Quantity
Age
t1.Age
anciennete
t2.anciennete
recense
t2.recense
2003
t5.2003
2004
t5.2004
2005
t5.2005
2006
t5.2006
January
t4.January
February
t4.February
March
t4.March
April
t4.April
May
t4.May
June
t4.June
July
t4.July
August
t4.August
September
t4.September
October
t4.October
November
t4.November
December
t4.December
Eclipse Clothing
t6.Eclipse Clothing
Eclipse Shoes
t6.Eclipse Shoes
Orion
t6.Orion
Shoes
t6.Shoes
T-Shirts
t6.T-Shirts
Tracker Clothes
t6.Tracker Clothes
Tracker Shoes
t6.Tracker Shoes
Orion Clothing
t6.Orion Clothing
Sélectionner les lignes distinctes seulement
Exécuter
Enregistrer et fermer
Annuler
Aide

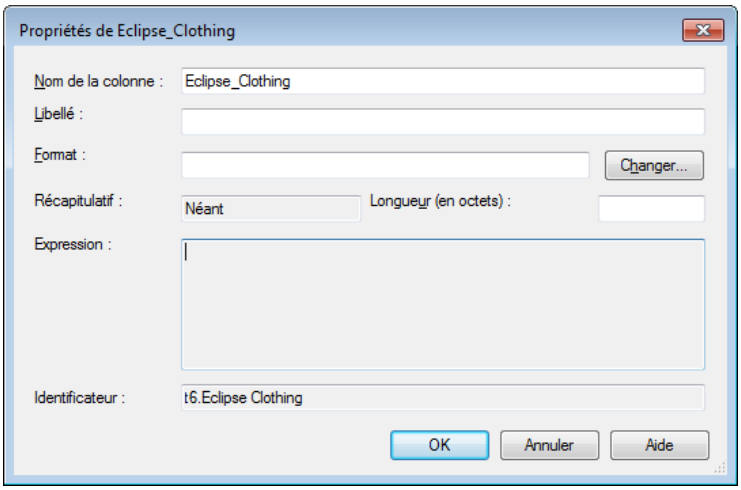


Renommer toutes les colonnes de telle sorte que leur nom soit un nom qui ne commence pas par un chiffre, qu'il soit sans espace et sans caractère spécial. Le nom de la colonne ne peut contenir de telles choses.

Remarque : Il n'y a alors plus de simple quotte sur le nom.

Renommer

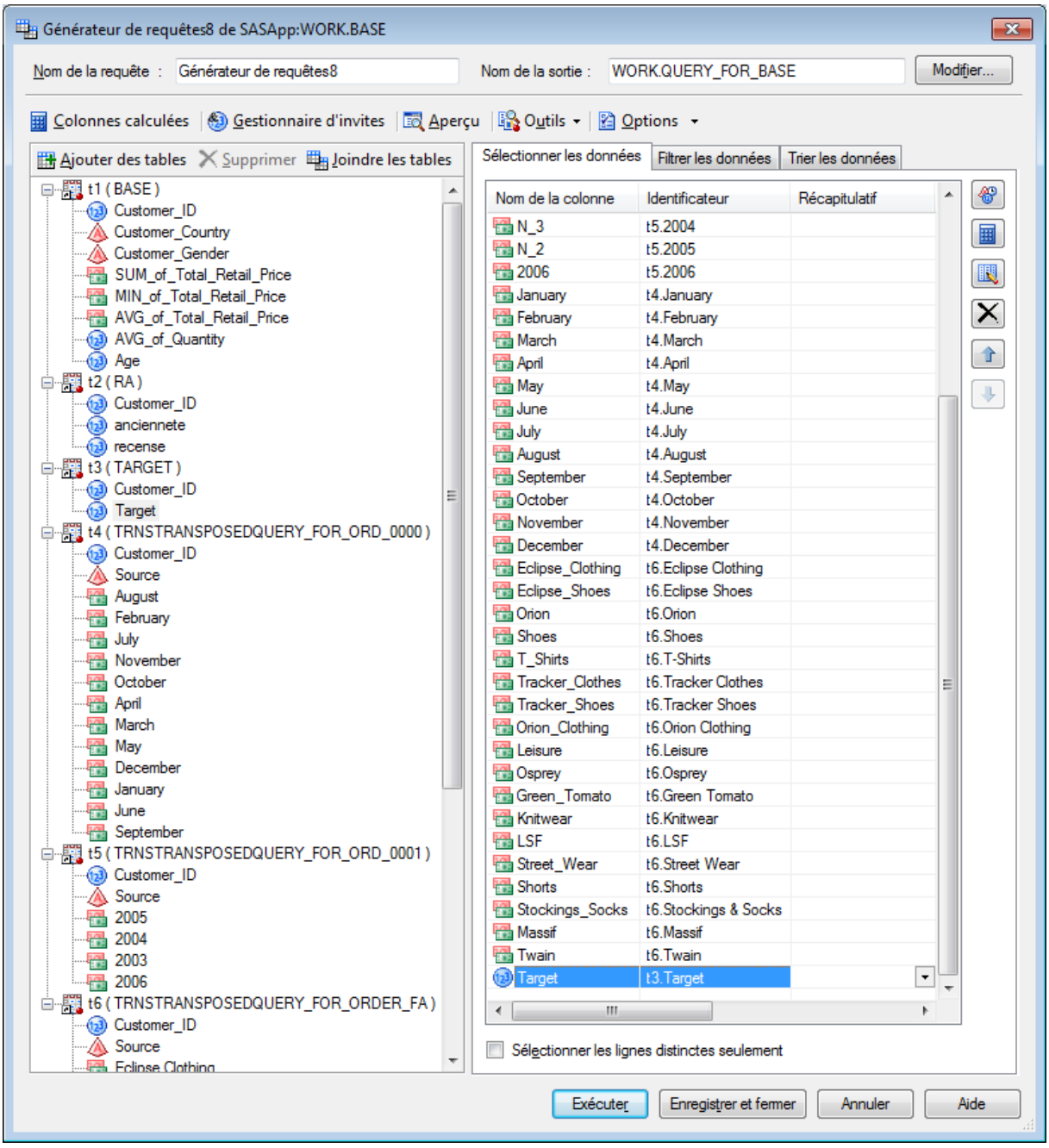
2006 en N_1 pour année n moins un
2005 en N_2
2004 en N_3
2003 en N_4



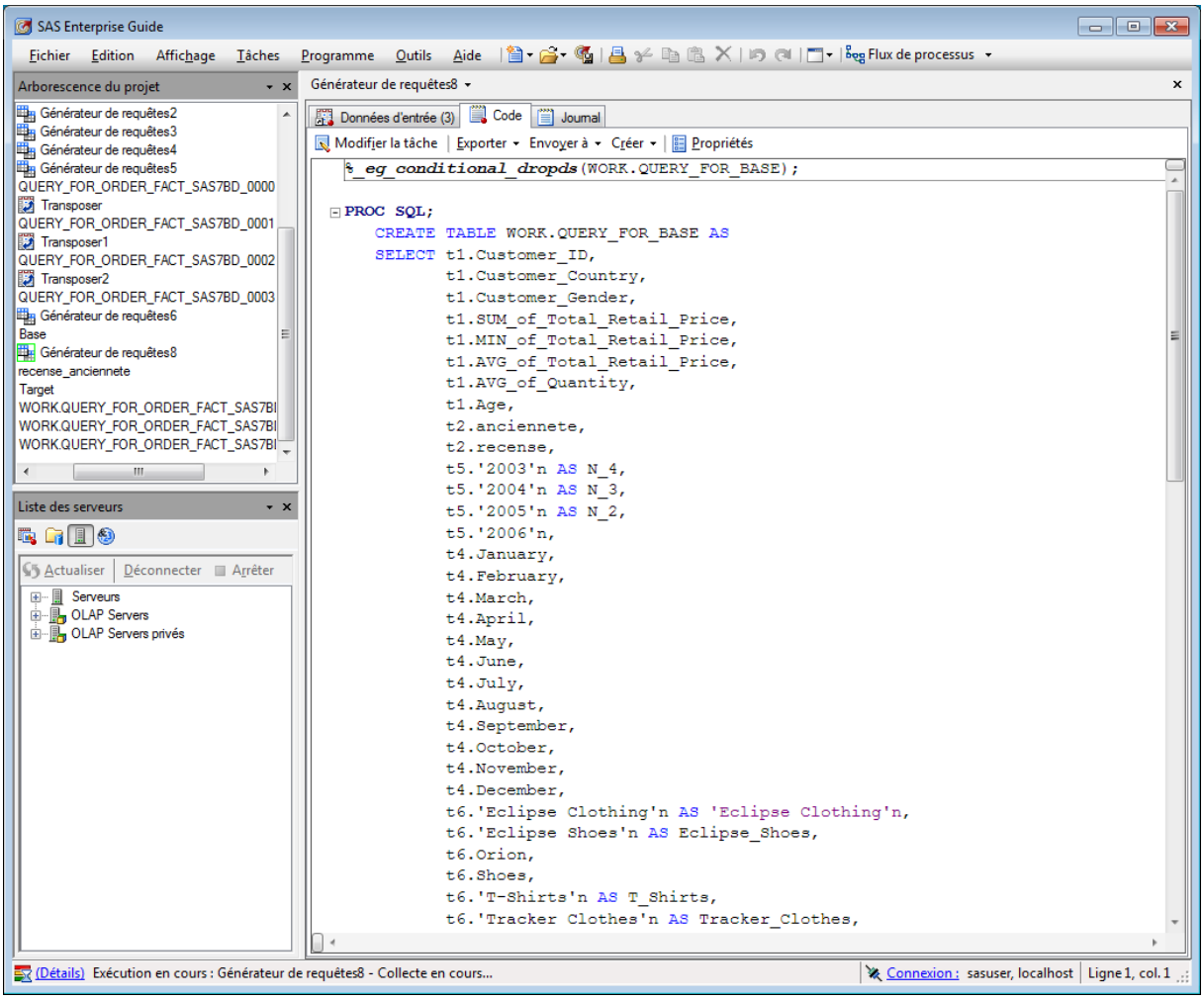
Il faut que le nom de la colonne soit sans espace ni carractère spéciaux.

Alias = Customer_ID
Alias = Eclipse_Clothing
Alias = Eclipse_Shoes
Alias = T_Shirts
Alias = Tracker_Shoes
Alias = Green_Tomato
Alias = Orion_Clothing
Alias = Street_Wear
Alias = Stockings_Socks

Ancienneté sans accent = anciennete


Générateur de requêtes8 de SASApp:WORK.BASE
Nom de la requête : Générateur de requêtes8
Nom de la sortie : WORK.QUERY_FOR_BASE
Modifier...
Colonnes calculées
Gestionnaire d'invites
Aperçu
Outils
Options
Ajouter des tables
Supprimer
Joindre les tables
t1 (BASE)
Customer_ID
Customer_Country
Customer_Gender
SUM_of_Total_Retail_Price
MIN_of_Total_Retail_Price
AVG_of_Total_Retail_Price
AVG_of_Quantity
Age
t2 (RA)
Customer_ID
anciennete
recense
t3 (TARGET)
Customer_ID
Target
t4 (TRNSTRANSPOSEDQUERY_FOR_ORD_0000)
Customer_ID
Source
August
February
July
November
October
April
March
May
December
January
June
September
t5 (TRNSTRANSPOSEDQUERY_FOR_ORD_0001)
Customer_ID
Source
2005
2004
2003
2006
t6 (TRNSTRANSPOSEDQUERY_FOR_ORDER_FA)
Customer_ID
Source
Eclipse Clothing
Sélectionner les données
Filtrer les données
Trier les données
Nom de la colonne
Identificateur
Récapitulatif
N_3 t5.2004
N_2 t5.2005
2006 t5.2006
January t4.January
February t4.February
March t4.March
April t4.April
May t4.May
June t4.June
July t4.July
August t4.August
September t4.September
October t4.October
November t4.November
December t4.December
Eclipse_Clothing t6.Eclipse Clothing
Eclipse_Shoes t6.Eclipse Shoes
Orion t6.Orion
Shoes t6.Shoes
T_Shirts t6.T-Shirts
Tracker_Clothes t6.Tracker Clothes
Tracker_Shoes t6.Tracker Shoes
Orion_Clothing t6.Orion Clothing
Leisure t6.Leisure
Osprey t6.Osprey
Green_Tomato t6.Green Tomato
Knitwear t6.Knitwear
LSF t6.LSF
Street_Wear t6.Street Wear
Shorts t6.Shorts
Stockings_Socks t6.Stockings & Socks
Massif t6.Massif
Twain t6.Twain
Target t3.Target
Sélectionner les lignes distinctes seulement
Exécuter
Enregistrer et fermer
Annuler
Aide

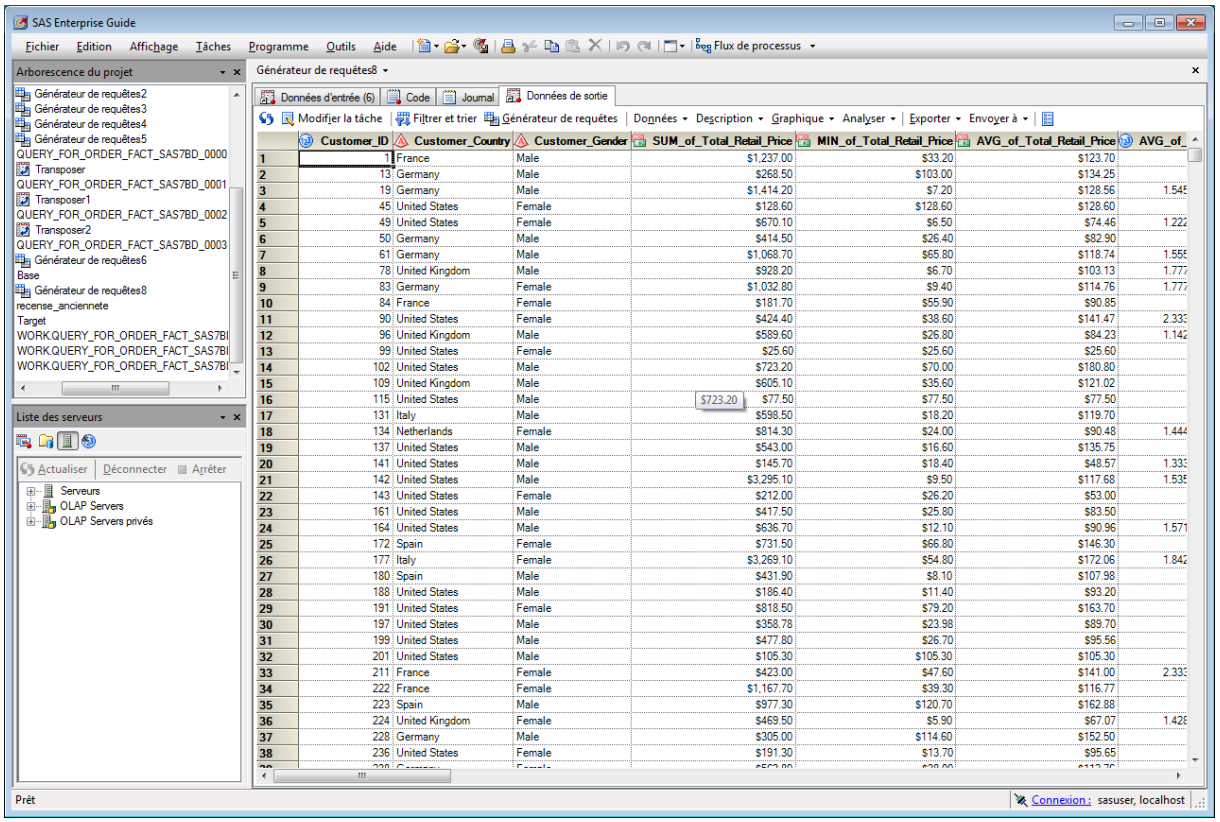
Exécuter

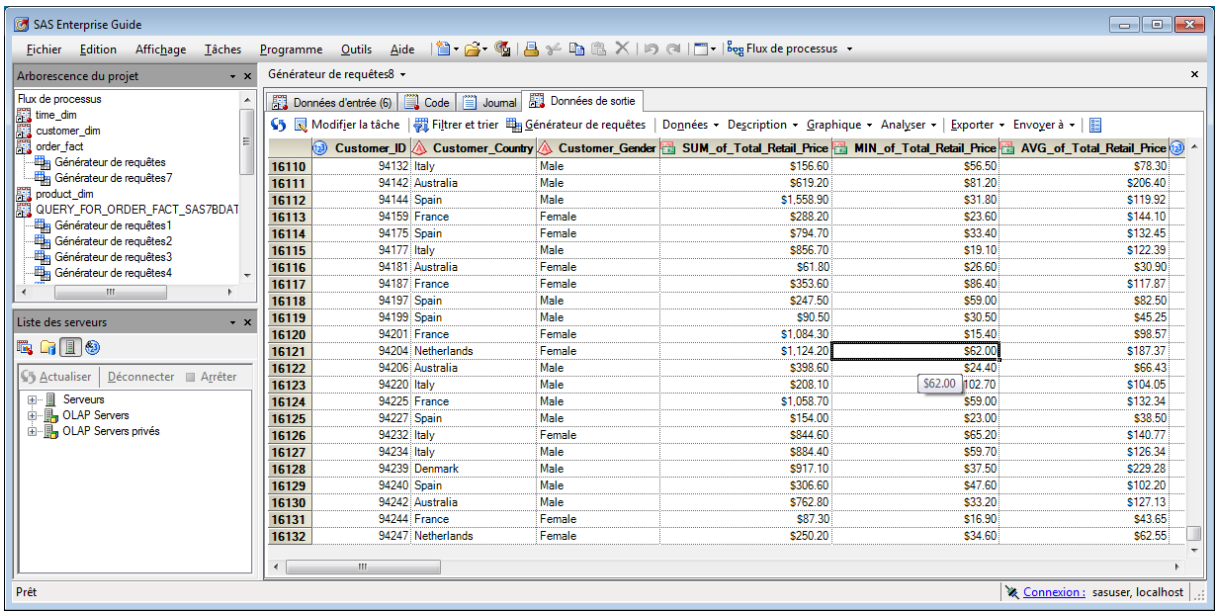


Pour vérifier que tous les noms sont bien valides, vous pouvez regarder le code. Si une ligne fini par un « 'n » en bordeaux, comme

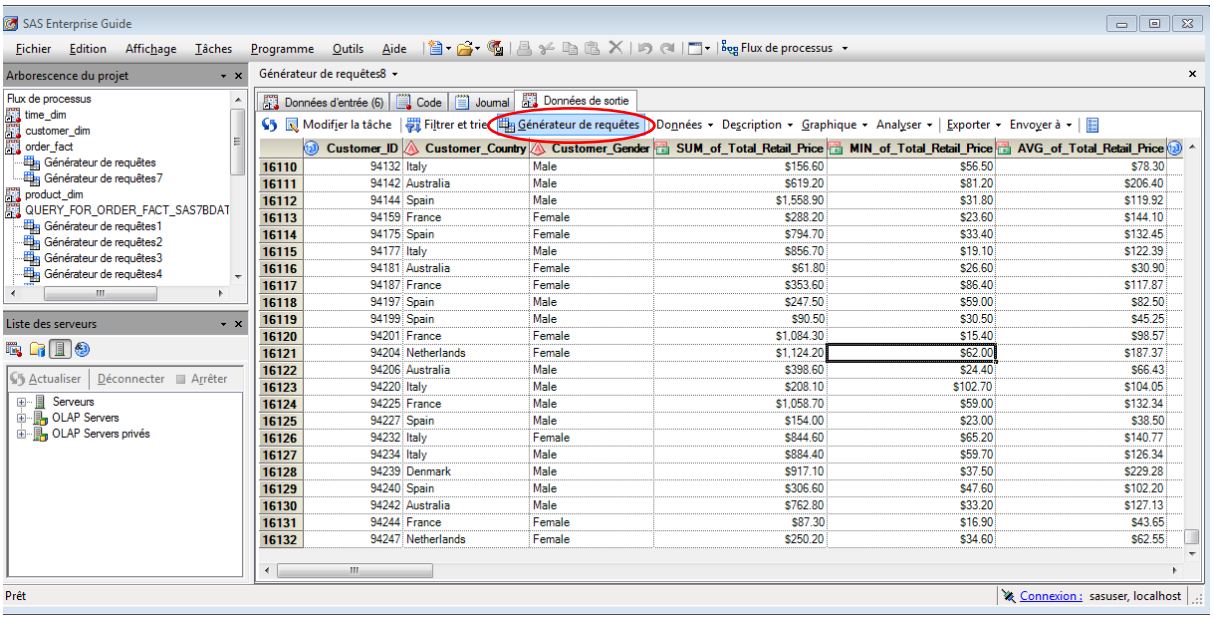
```
t6.'Eclipse Clothing'n AS 'Eclipse Clothing'n,
```

c'est qui y a un problème, il faut modifier la tâche.

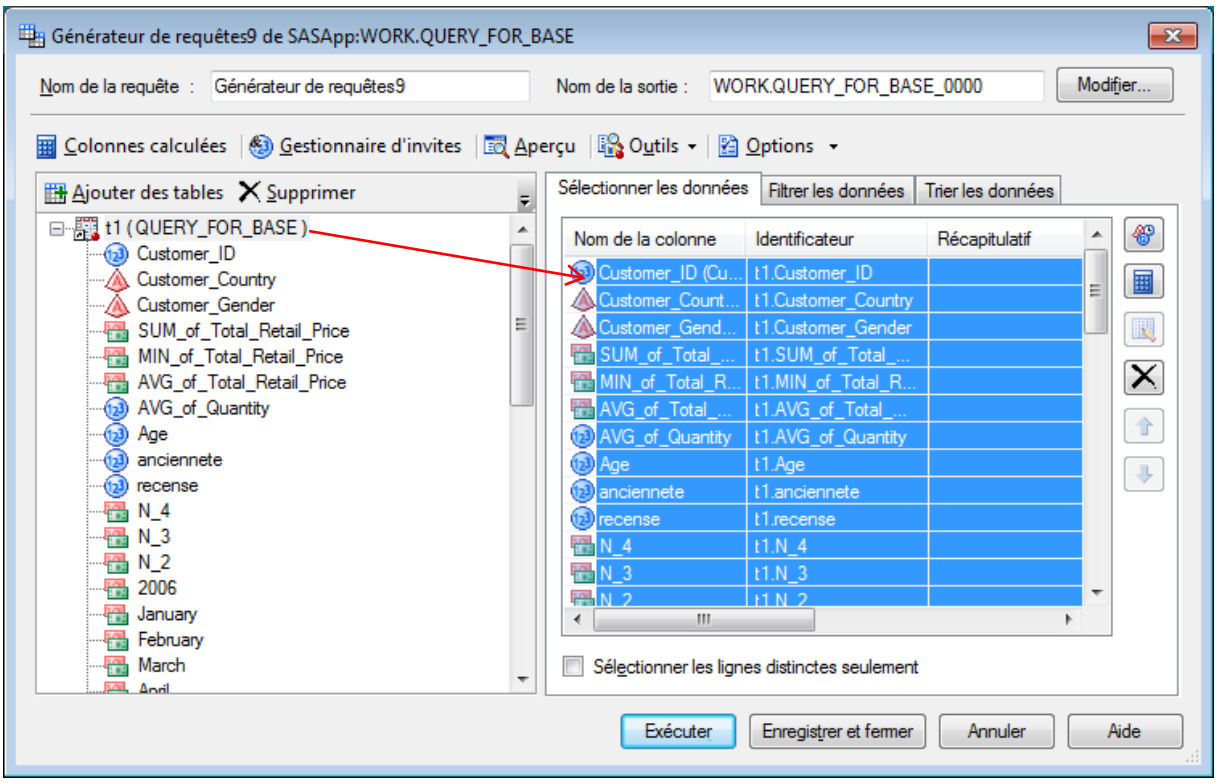




Vous devez avoir 16 132 lignes.

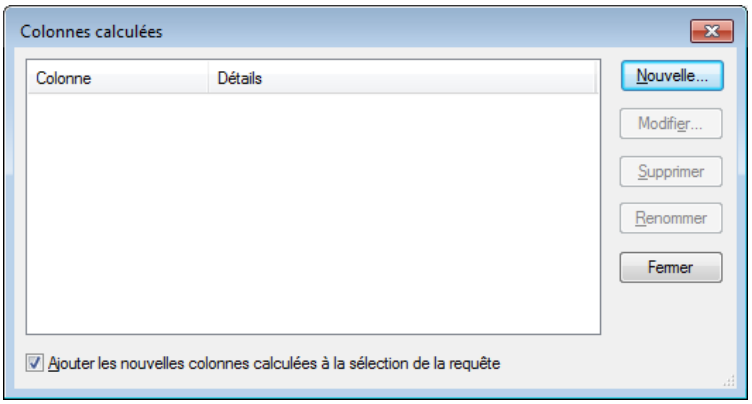


Générer une requête pour recoder la variable à expliquer en 1 ou 0.

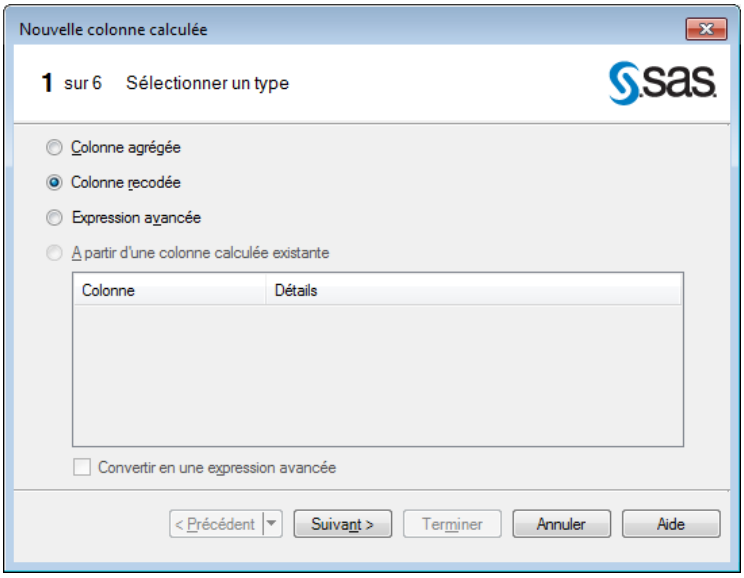


Ajouté toutes les lignes

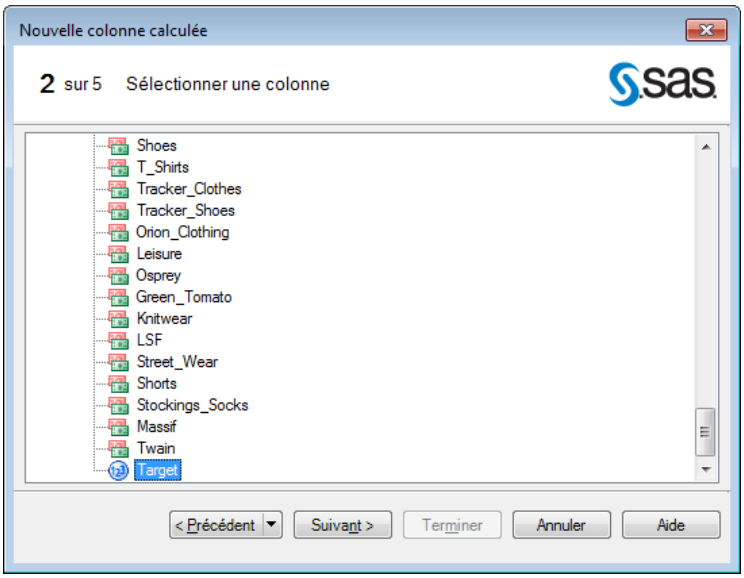
Dans colonnes calculées



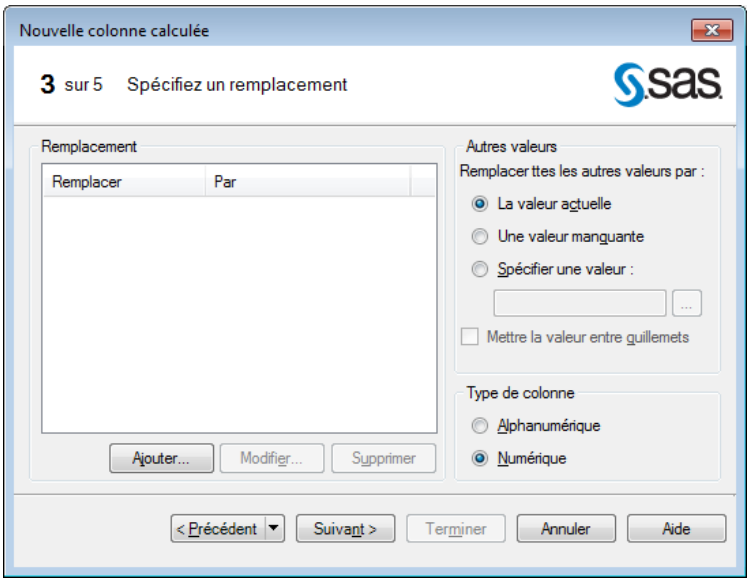
Nouvelle



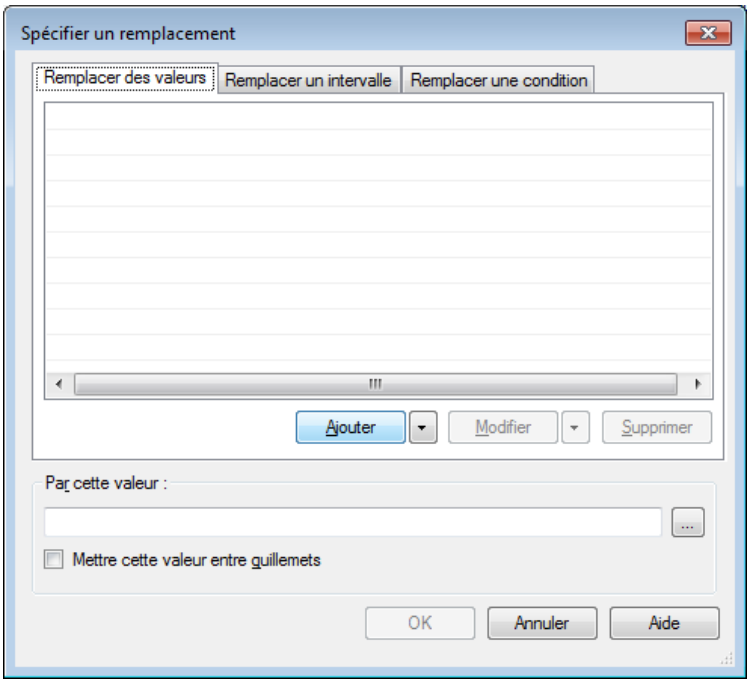
Colonne recodée, suivant



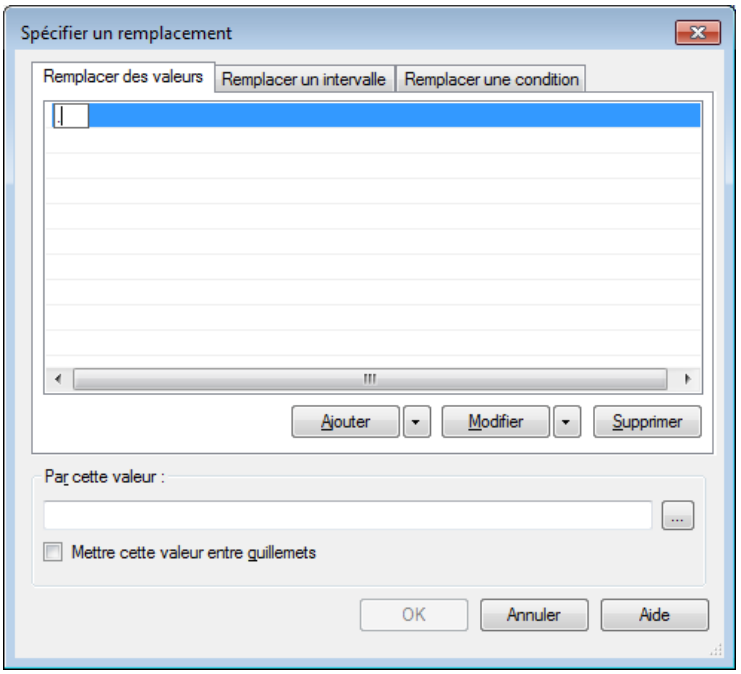
sélectionnez ‘Target ‘, suivant



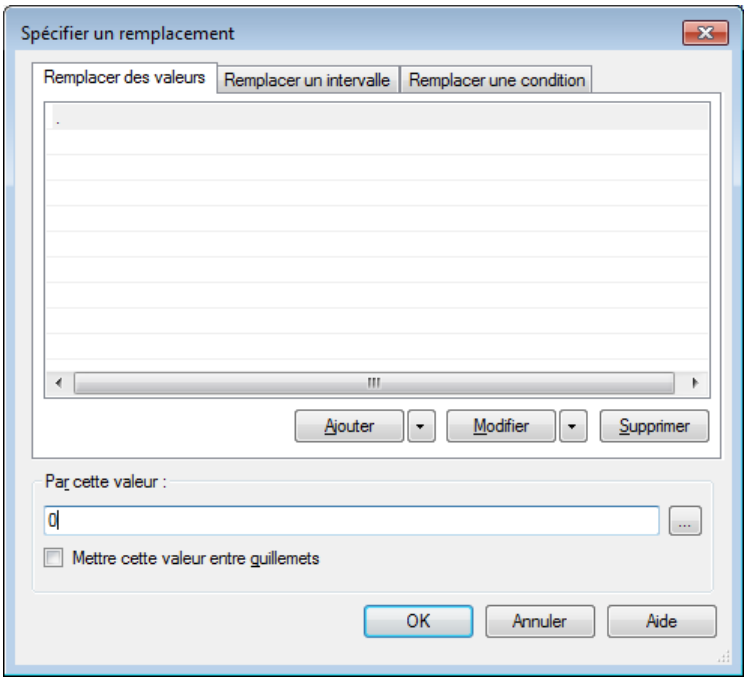
Ajouter



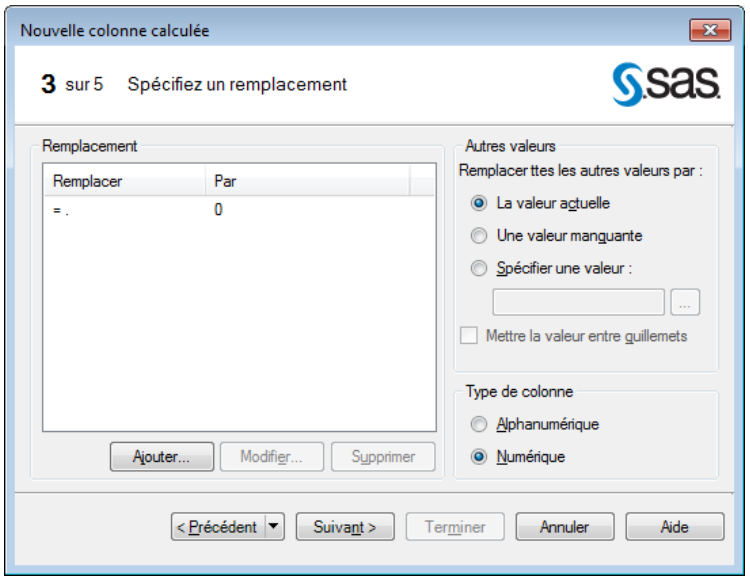
Ajouter



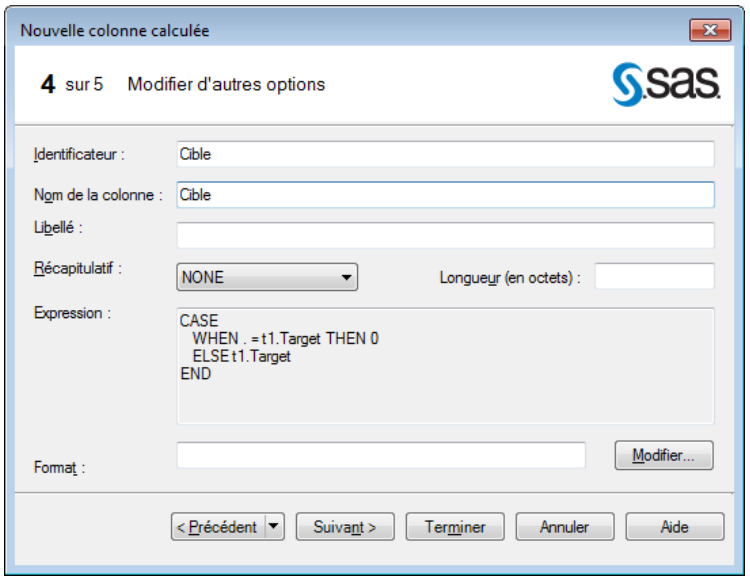
point : « . »



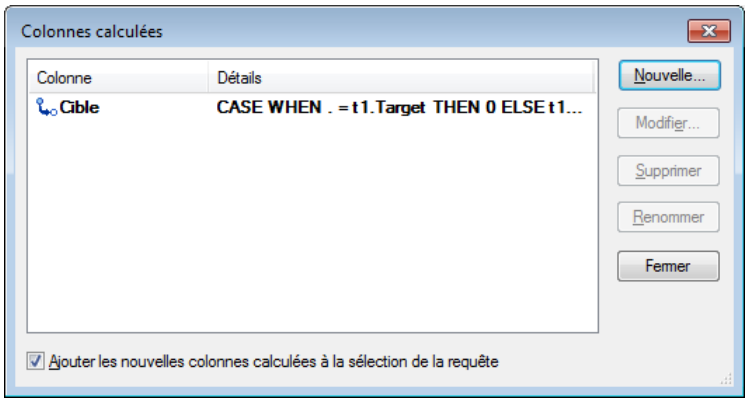
zéro : « 0 », OK



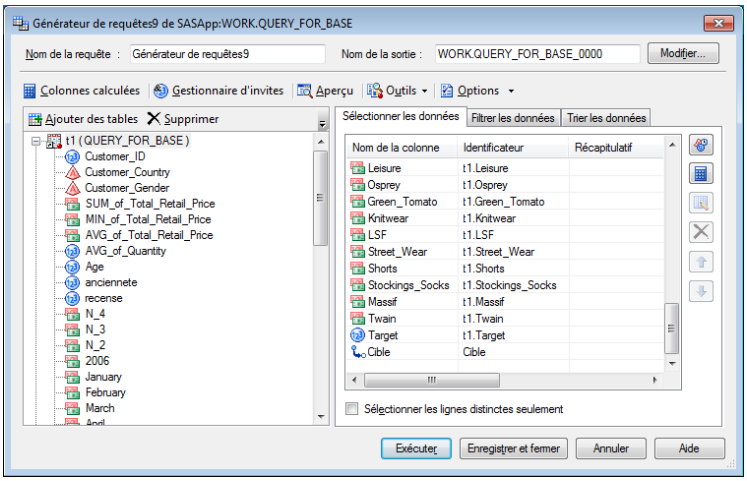
suivant



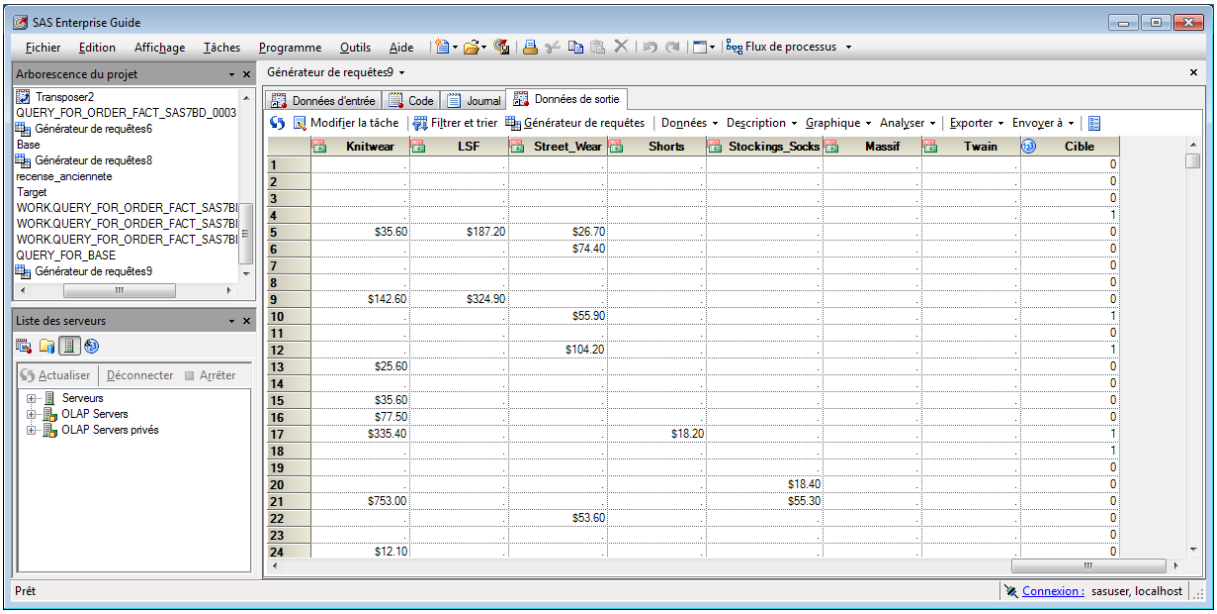
Renommer la colonne, Terminer



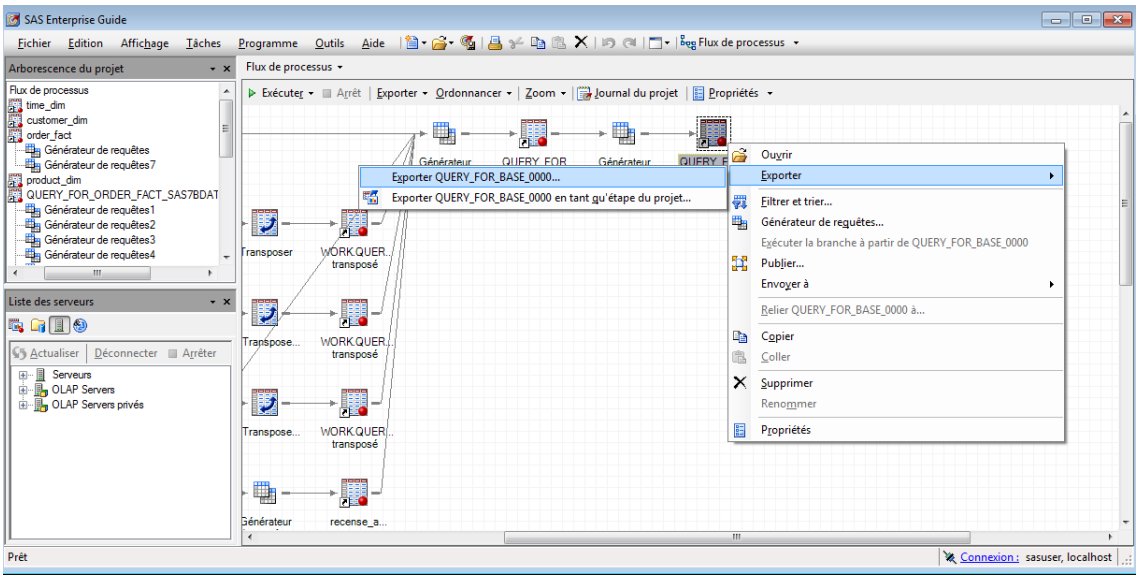
Fermer

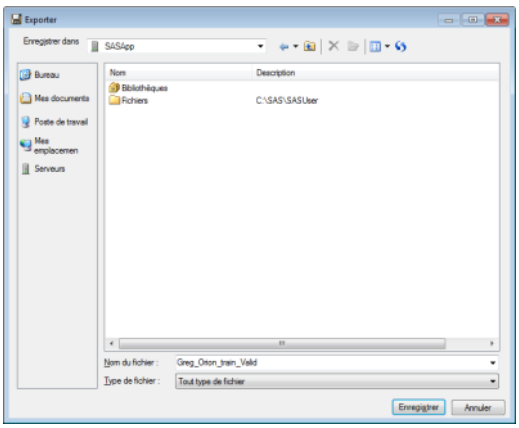


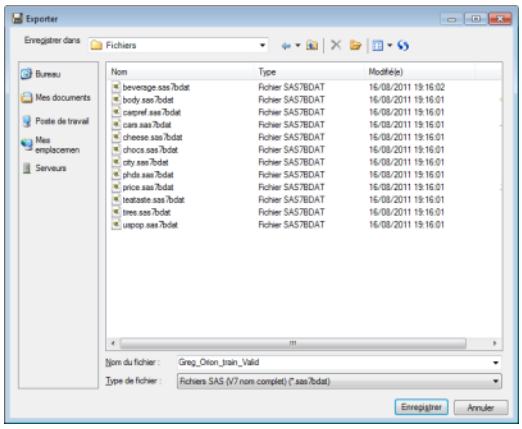
Supprimer la colonne Target,
Exécuter



Exporter la table sur l'ordinateur local

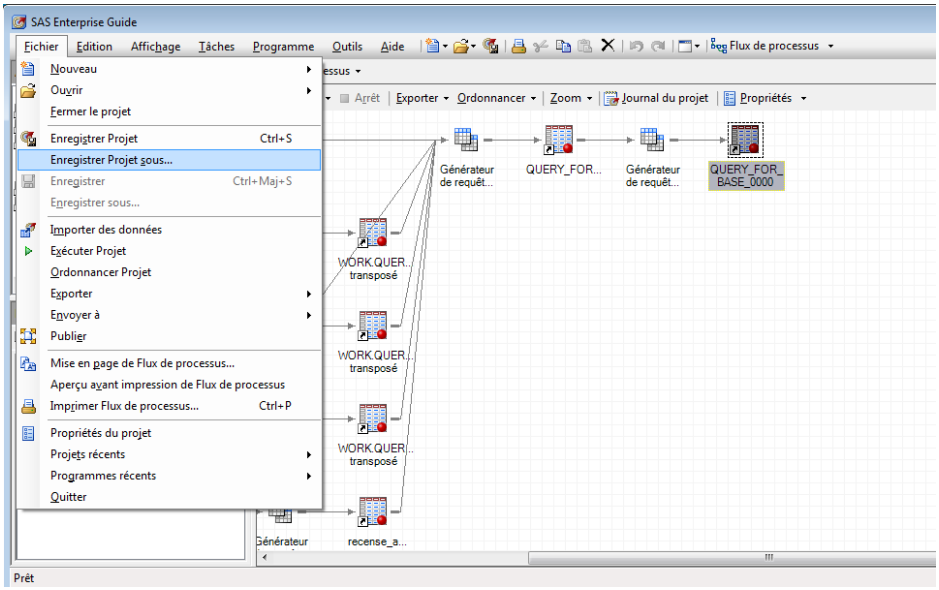


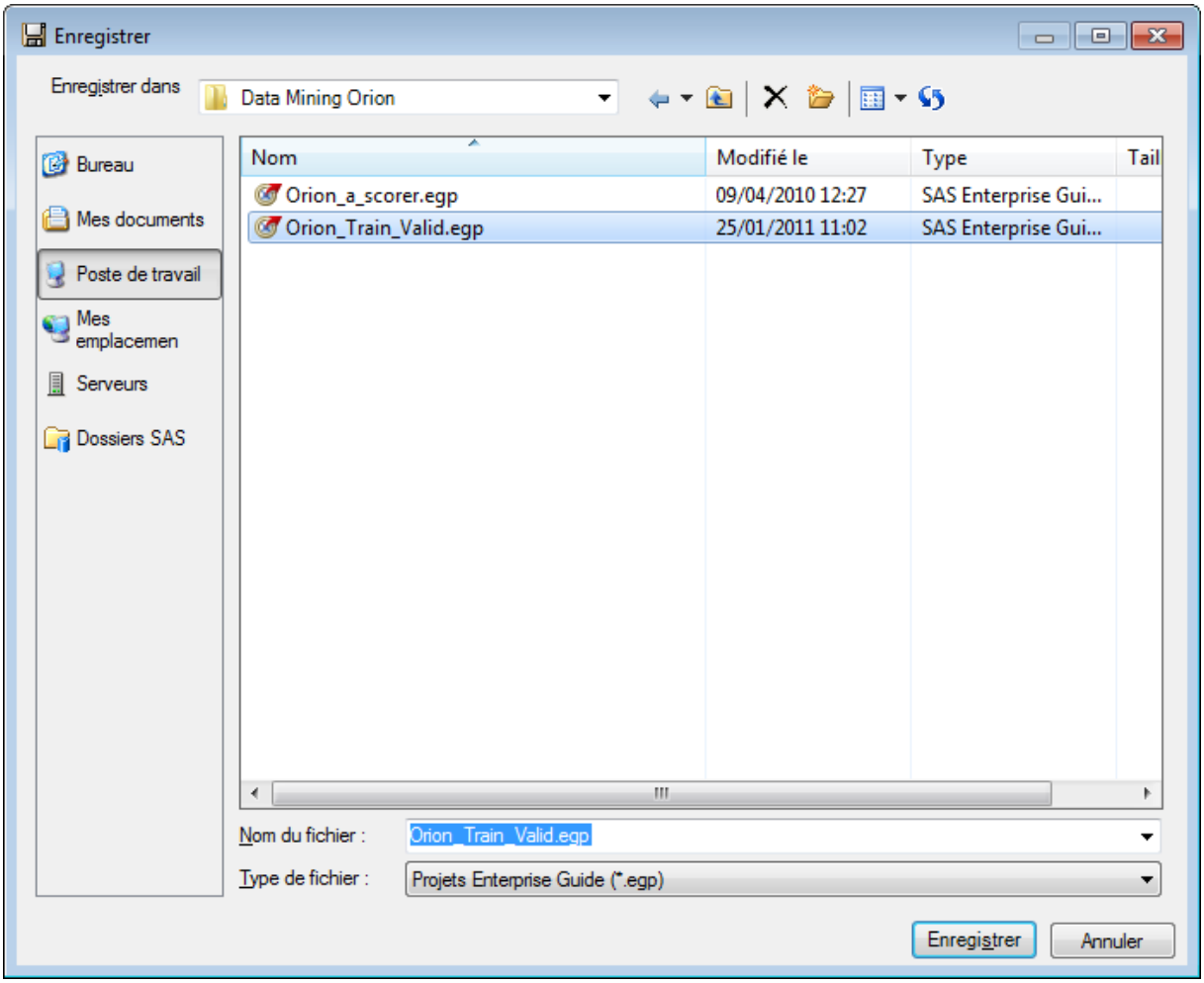




Sur le serveur SASApp, dans Fichiers
La nommer Votre_nom_Orion_Train_Valid
Enregistrer

Enregistrer le projet



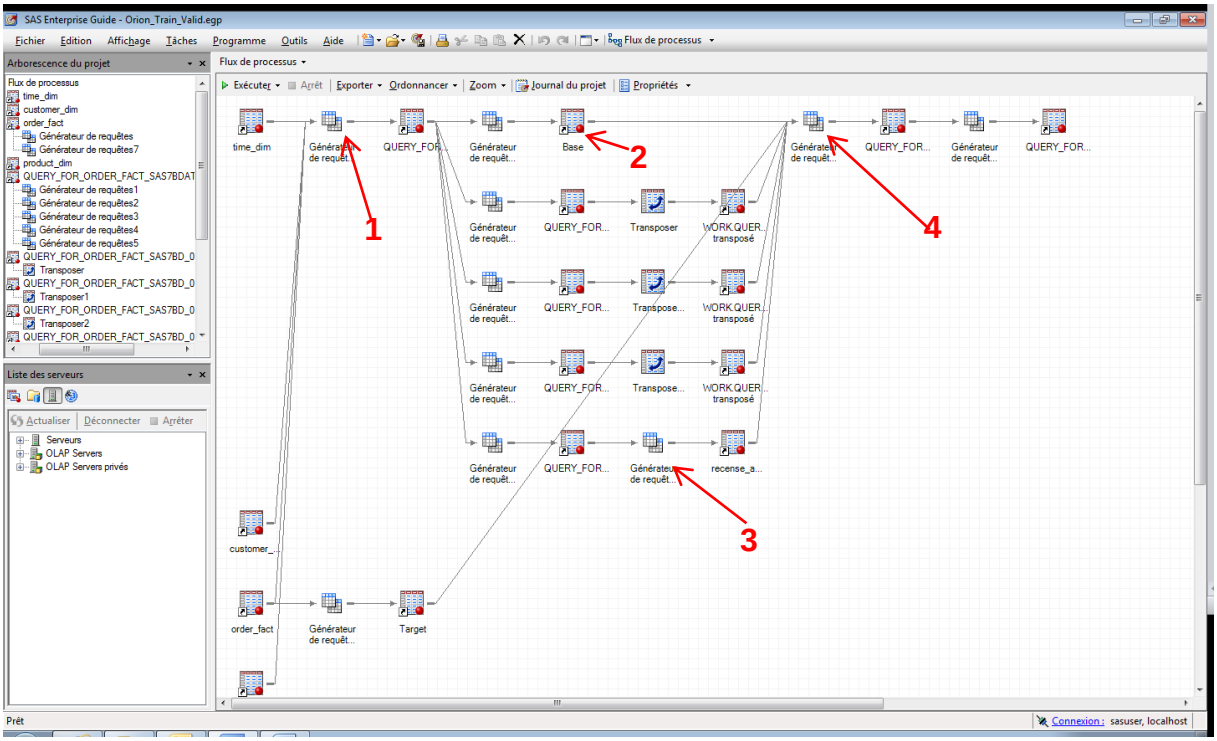


Sur votre bureau, par exemple
Enregistrer

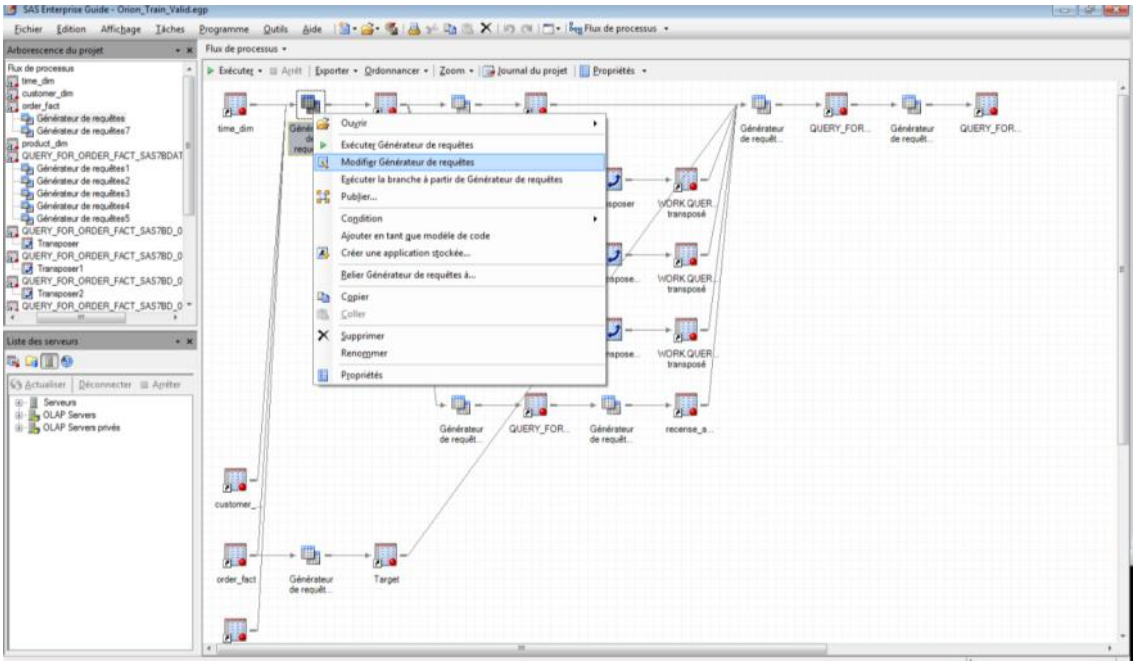
Création de la table à scorer

Nous avons créé la table d'apprentissage et de validation. Nous allons maintenant créer la table à scorer. C'est presque la même, sauf qu'il faut modifier :

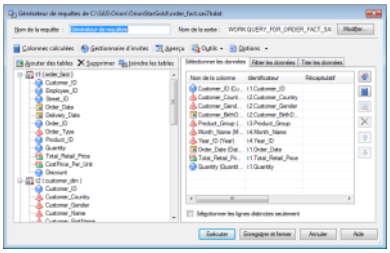
1. le filtre sur la première requête n'est plus « inférieur au 1er Janvier 2007 », mais « supérieur ou égale au 1er janvier 2004 »
2. les clients ont pris un an, leur âge n'est plus au 1er Janvier 2007 mais au 1er janvier 2008.
3. De même pour la recense et l'ancienneté, elles ne sont plus au 1er Janvier 2007 mais au 1er janvier 2008.
4. Dans la dernière requête, il faut supprimer la jointure avec la « Target ». Les années relative change, 2003 n'est plus prise en compte, 2004 (N_3) devient N_4, 2005 (N_2) devient N_3, 2006 (N_1) devient N_2, et 2007 est N_1.

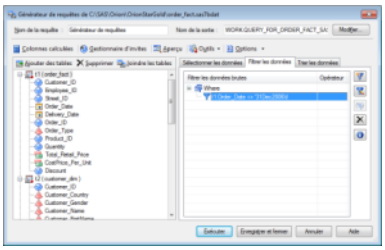


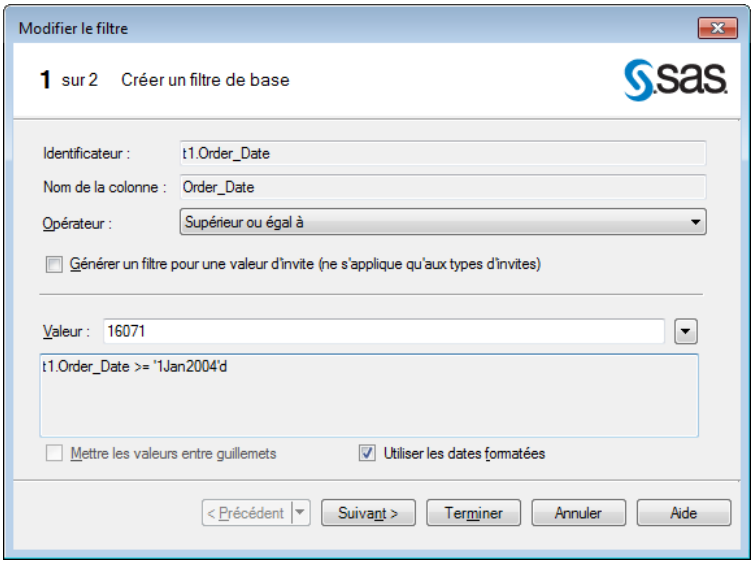
1) Ouvrir la première requête sur order_fact et modifier le filtre



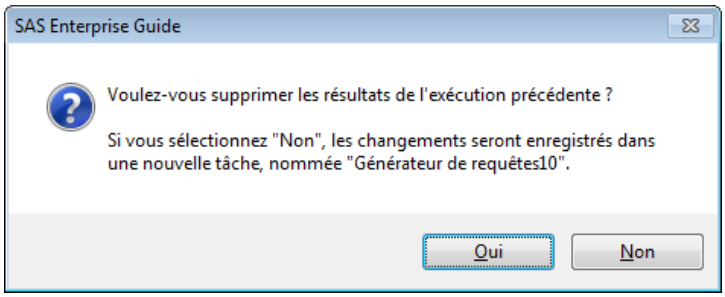
Modifier le générateur de la première requête





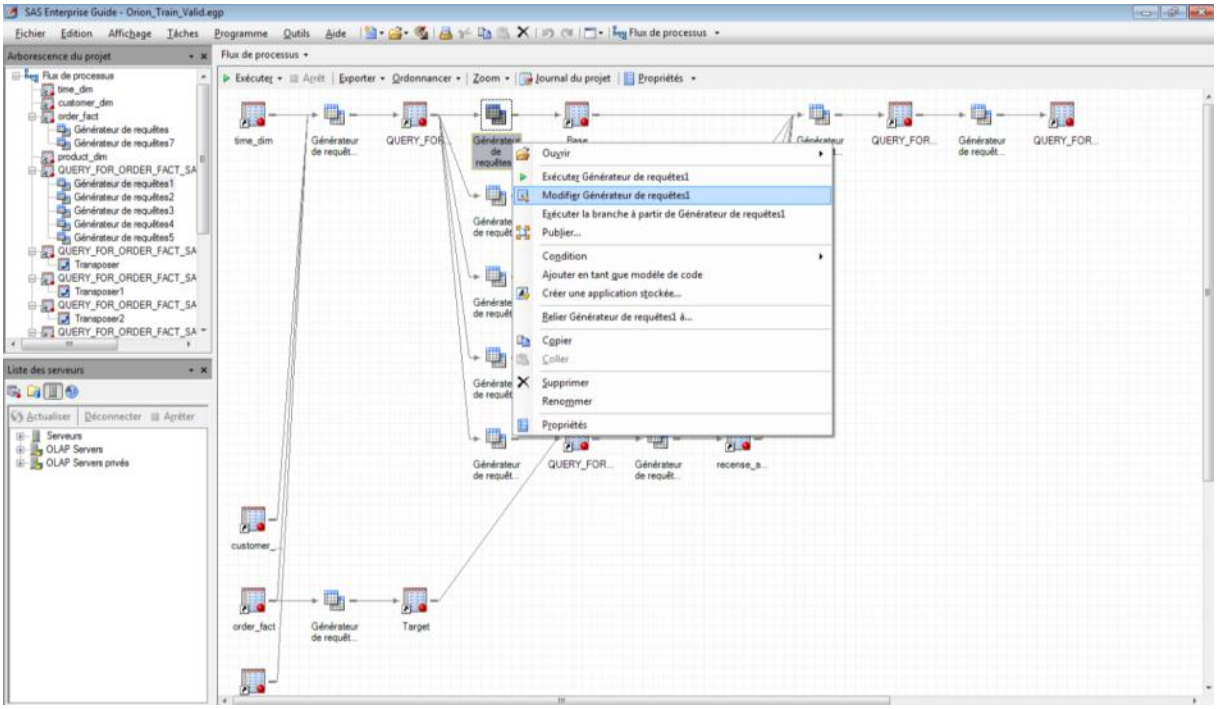


Dans l’onglet du filtre
Filtrer la table sur une date supérieure au ‘01Jan2004’d
Suivant
Enregistrer et fermer la requête

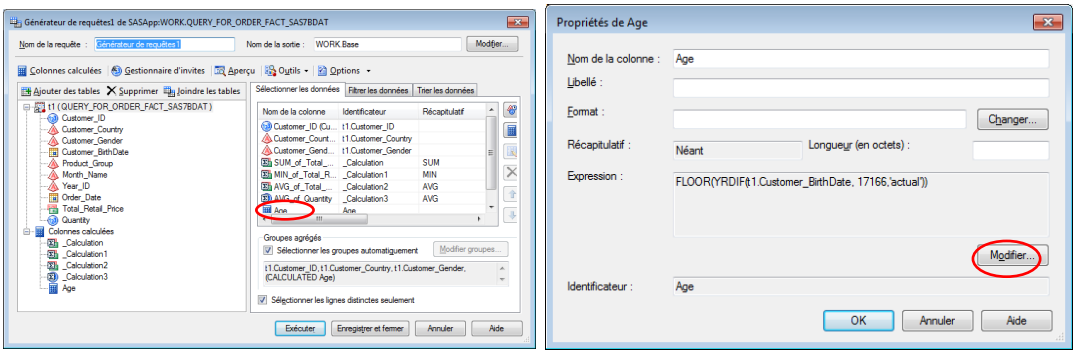

SAS Enterprise Guide
Voulez-vous supprimer les résultats de l'exécution précédente ?
Si vous sélectionnez "Non", les changements seront enregistrés dans une nouvelle tâche, nommée "Générateur de requêtes10".
Oui
Non

Oui

2) Ouvrir la requête sur la première table intermédiaire

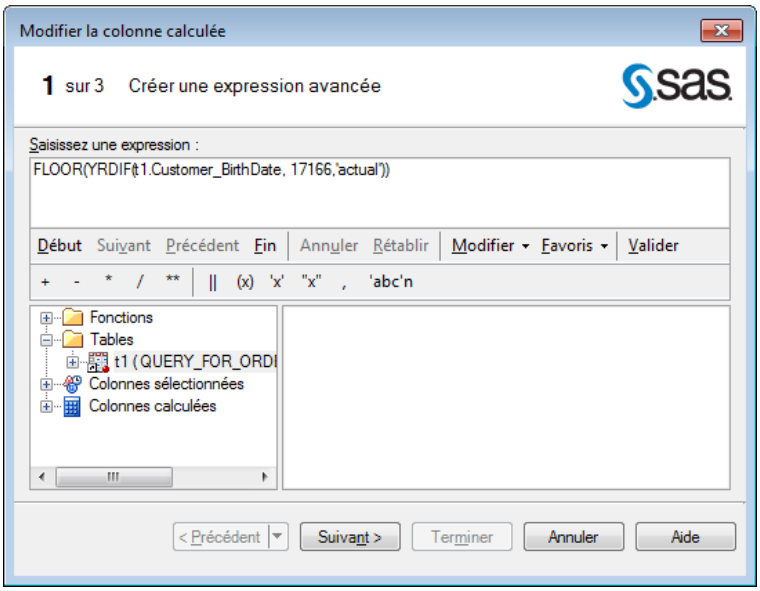


Modifier la tâche

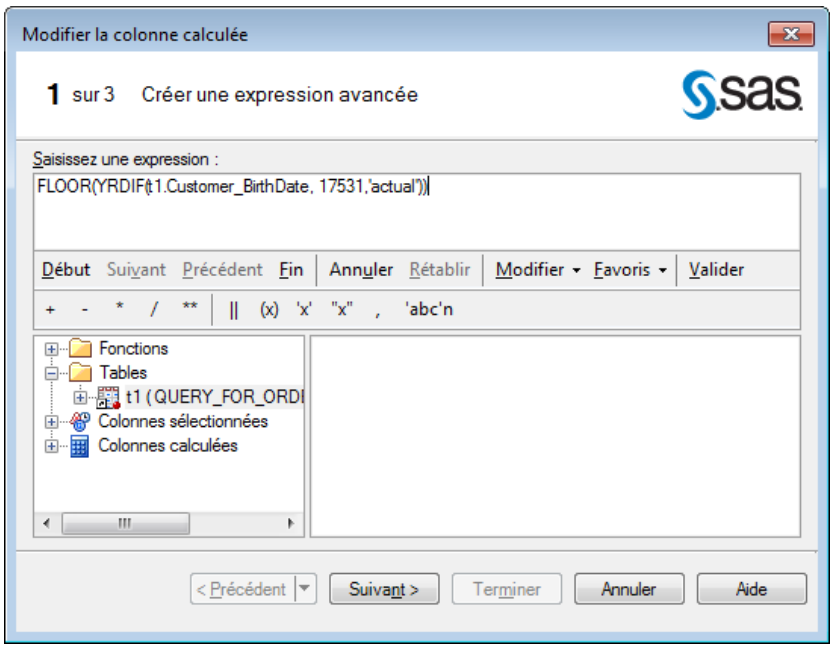


Sélectionner l'âge

Modifier l'expression

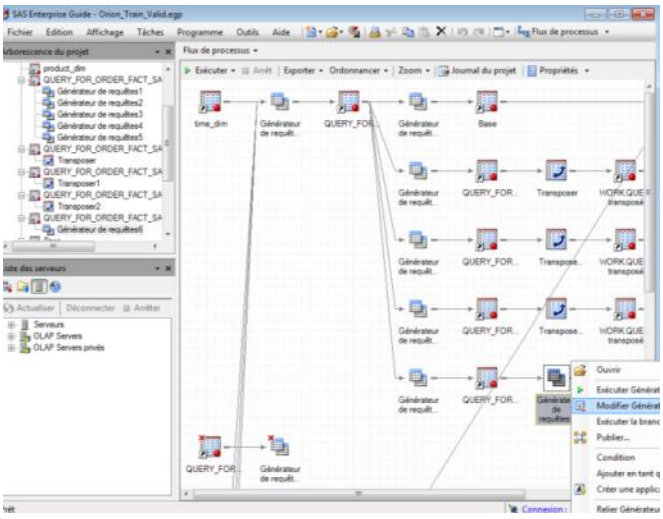


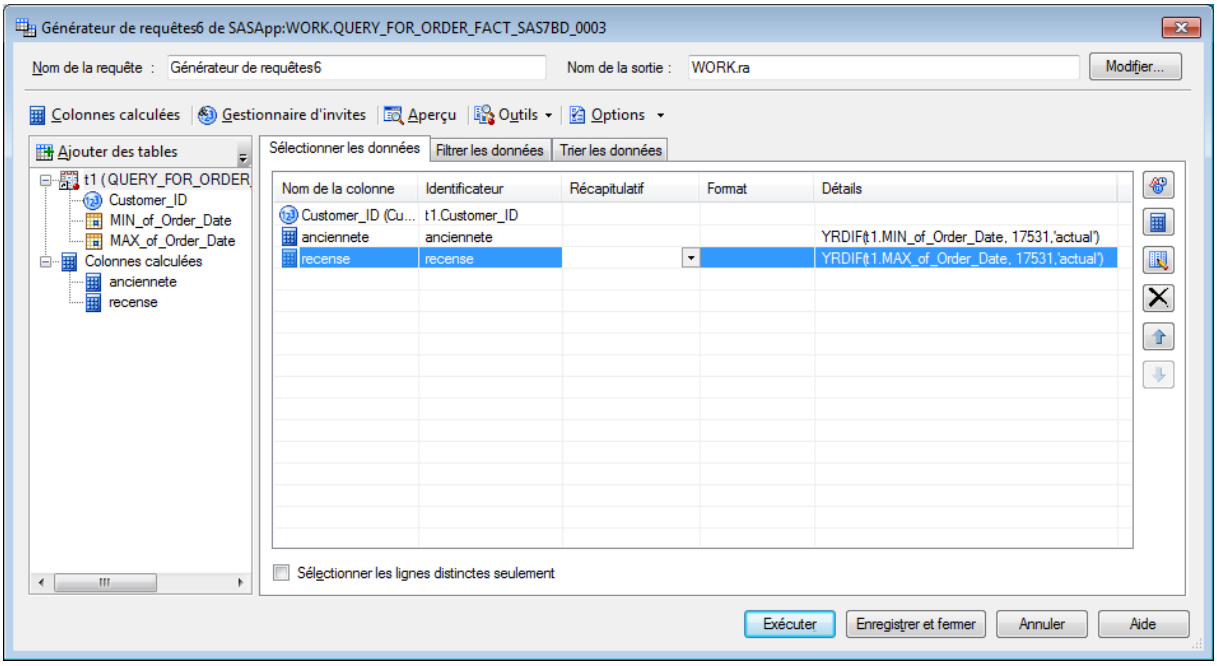
Modifier l'âge de telle sorte qu'il soit par rapport au premier janvier 2008 (17531 = 17166 + 365).



Suivant, Terminer, OK
Enregistré et fermer la requête
Oui pour supprimer les résultats de l'exécution précédente.

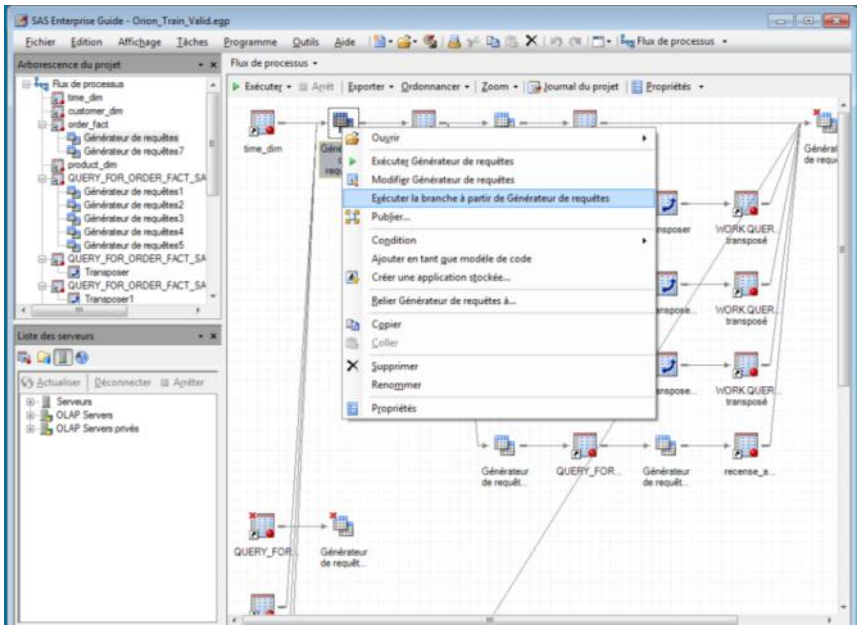
3) Modifier l'ancienneté et la recensent de tel sorte qu'elles soient calculer par rapport au 1er janvier 2008 (17531).

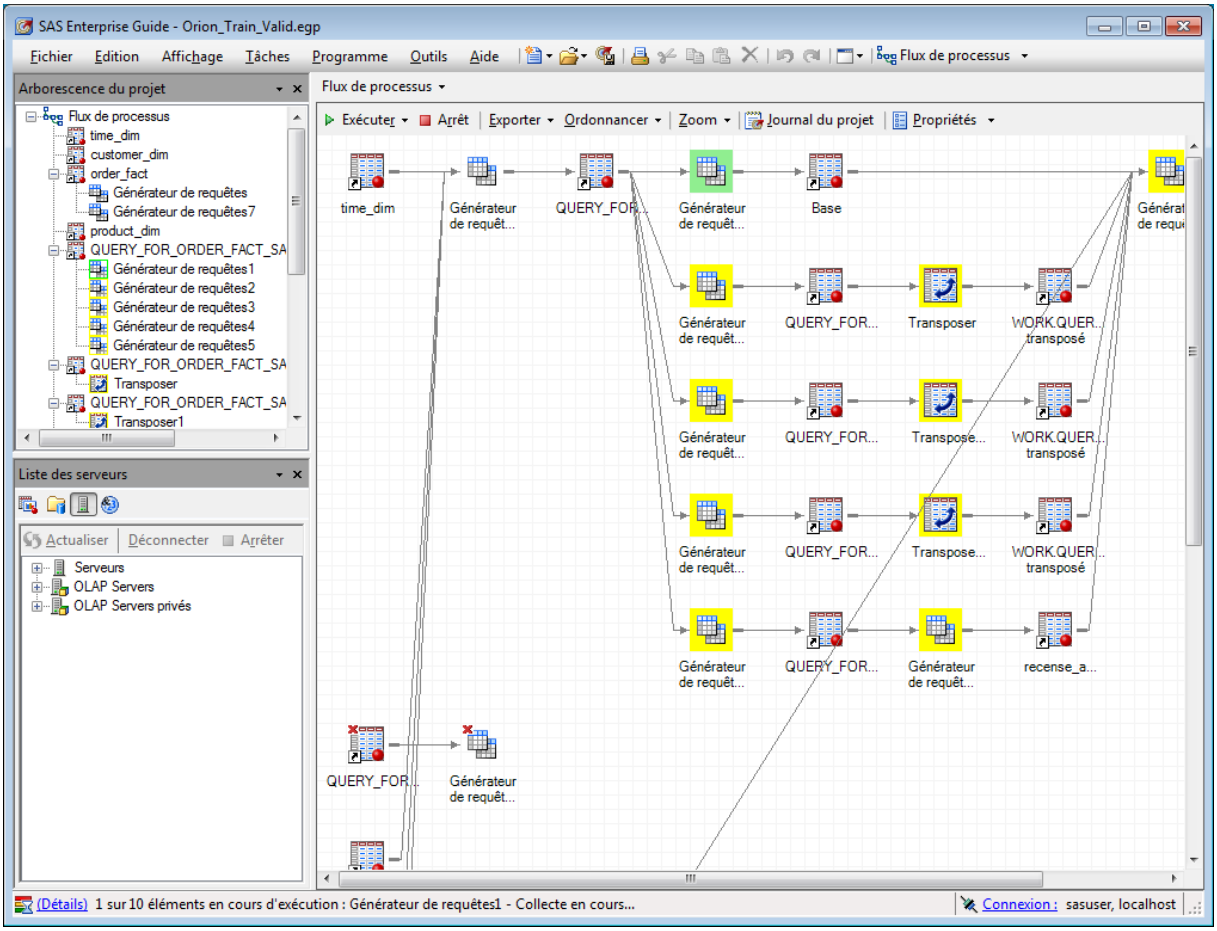




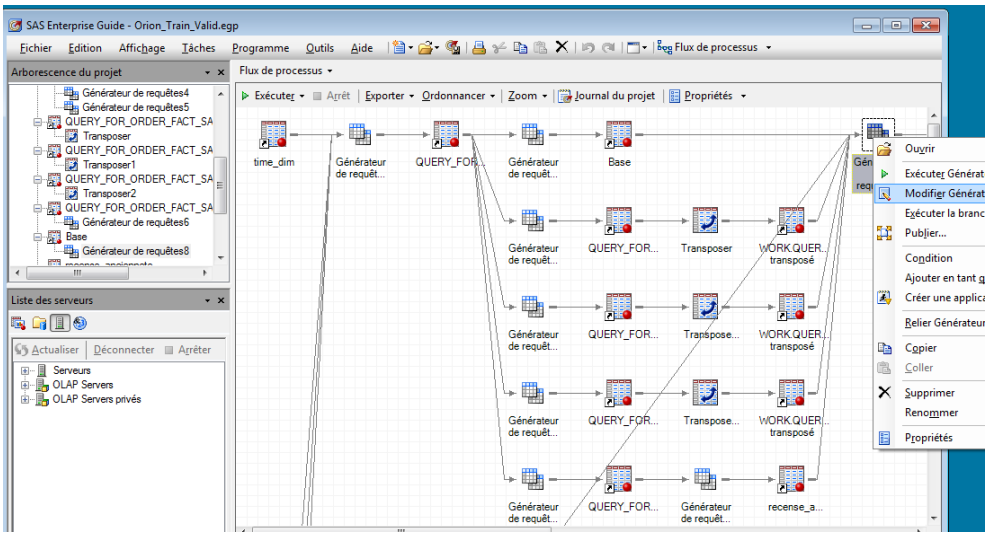
Enregistré et fermer la requête
Oui pour supprimer les résultats de l'exécution précédente.

4) Avant de modifier la dernière requête, exécuter la branche à partir de la première requête.



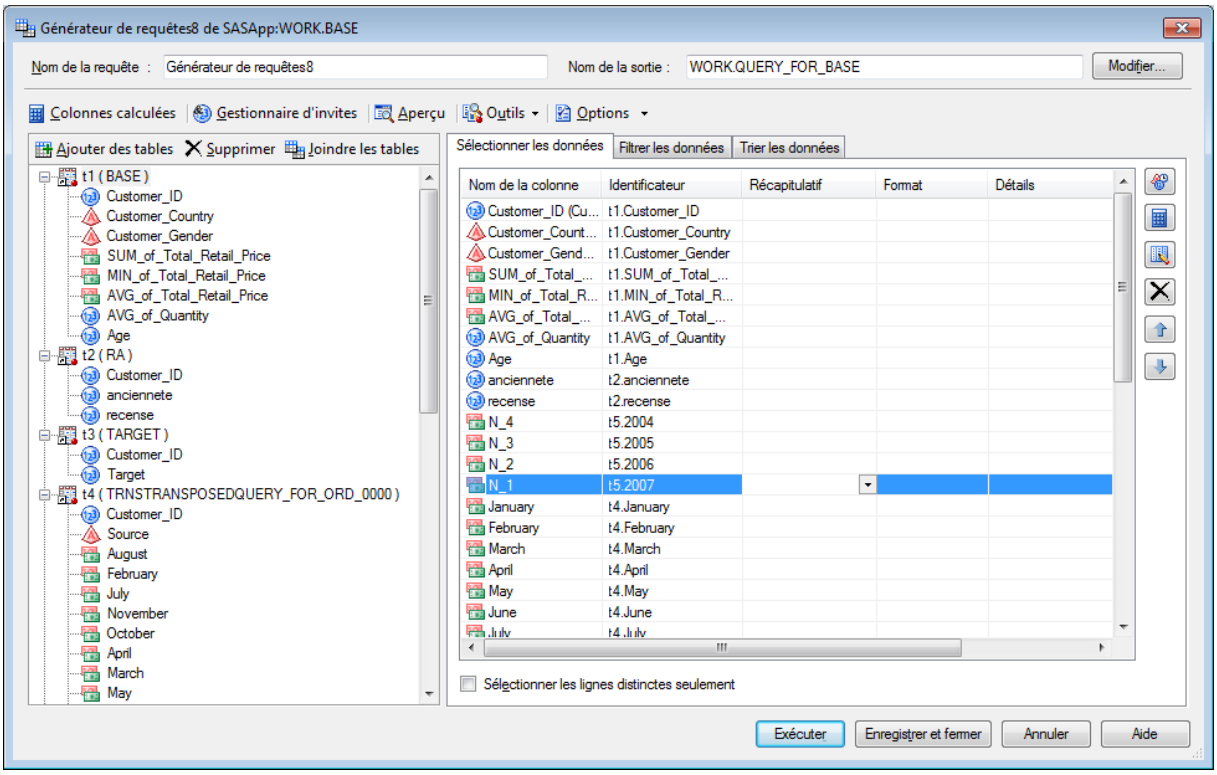


Il y a des erreurs !



Modifier la tâche

Supprimer les colonnes Target et 2003 (N_4)



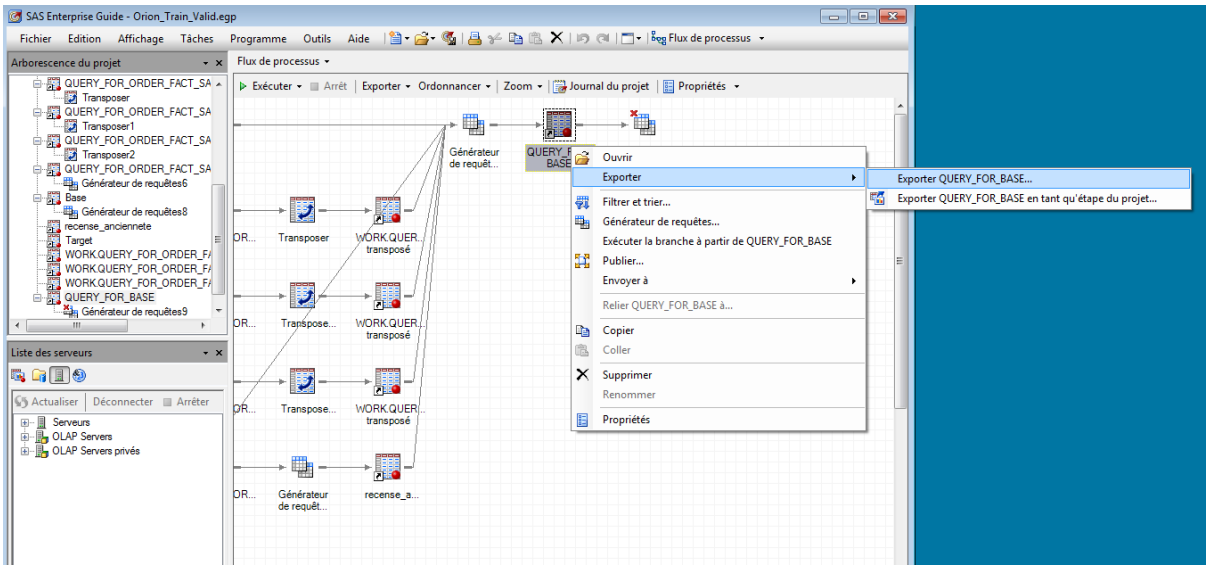
Renommer dans l'ordre

N_3 en N_4,
N_2 en N_3,
N_1 en N_2,

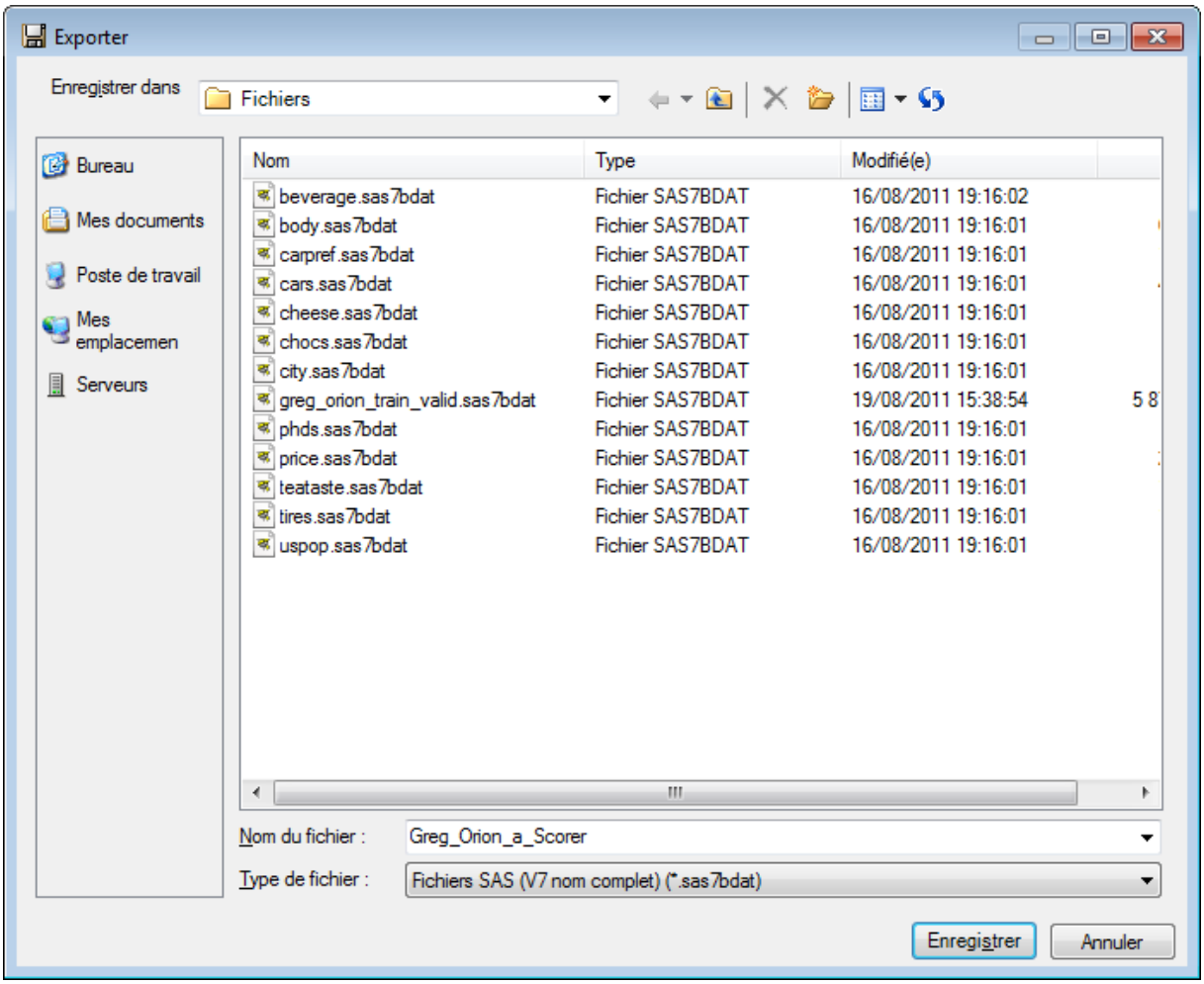
Ajouter la colonne 2007 et renommer là en N_1

Exécuter la requête

Oui pour remplacer la table précédente

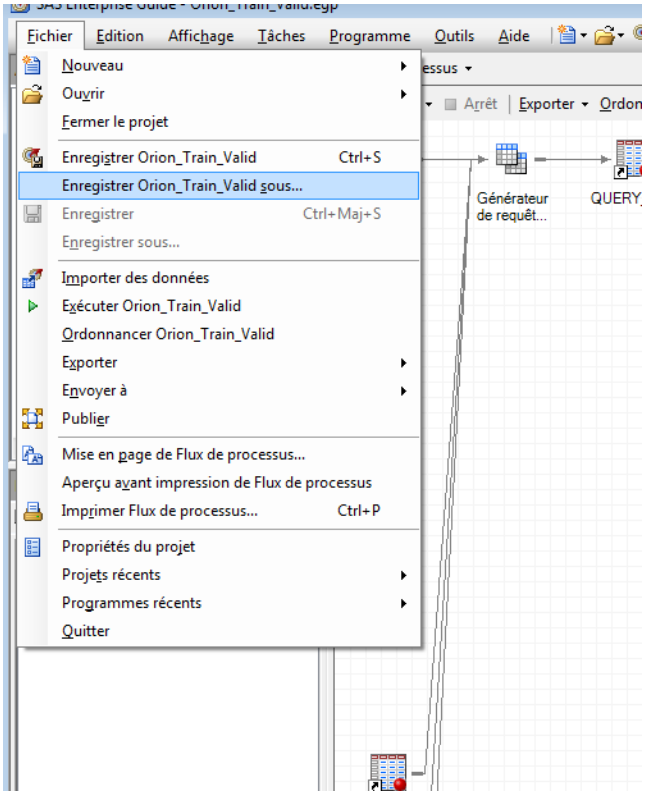


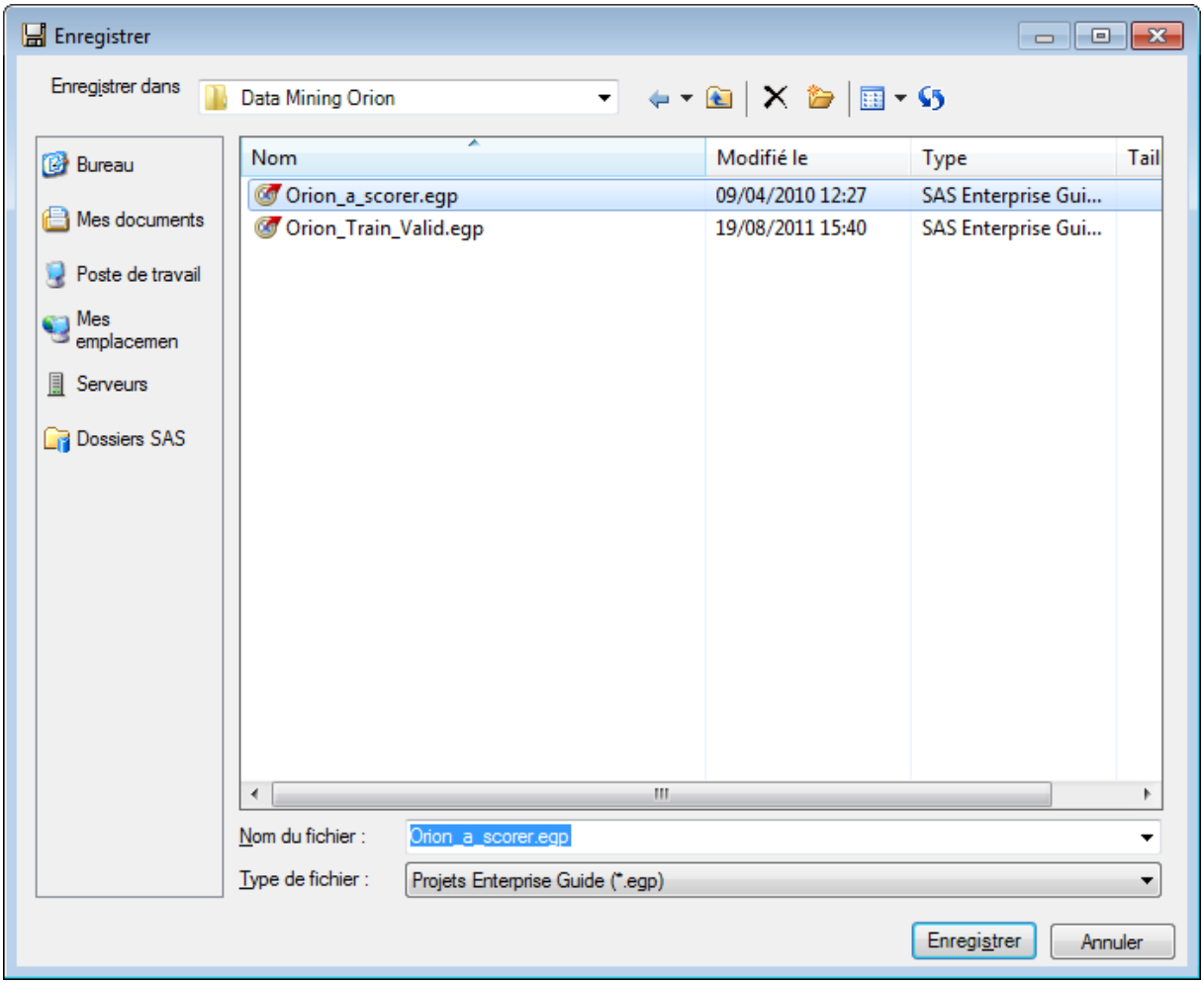
Exporter la table en tant que table votre_nom_Orion_a_scorer
Il n'y a pas de colonne cible à recoder.



Enregistrer

Enregistrer le projet sous Orion_a_scorer



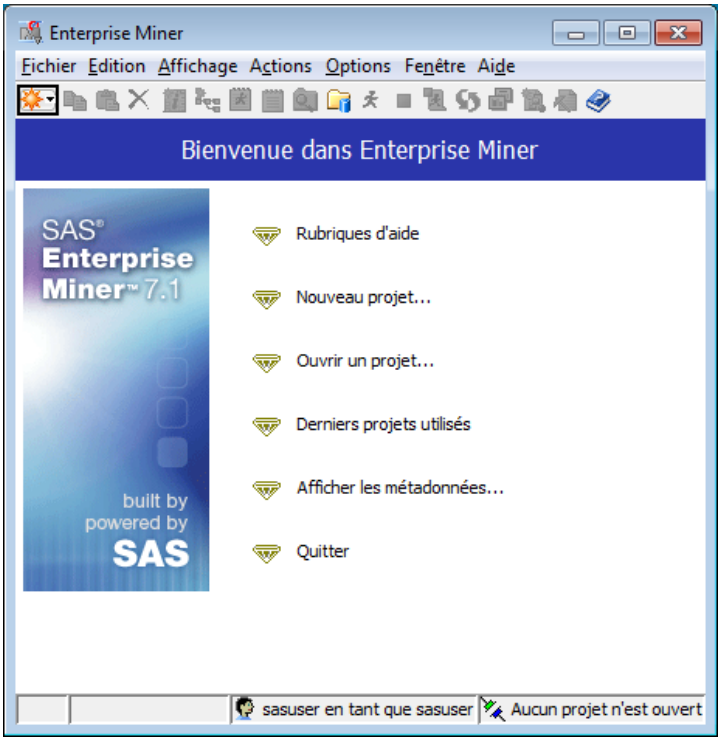


Vous devez avoir dans votre répertoire un projet pour créer la table orion_train_and_valid et un projet pour créer la table Orion_a_scorer.

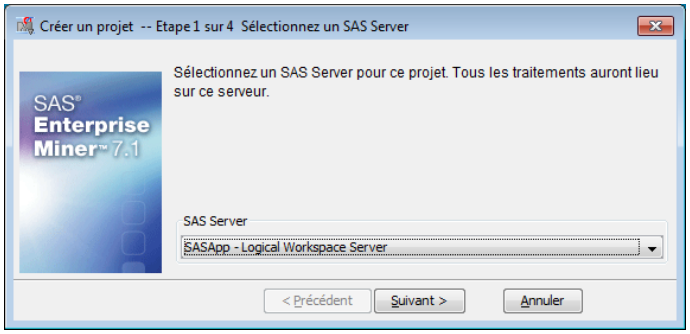
Création d'un modèle de Data Mining

Depuis Démarrer → Programmes → SAS → SAS Enterprise Miner Client

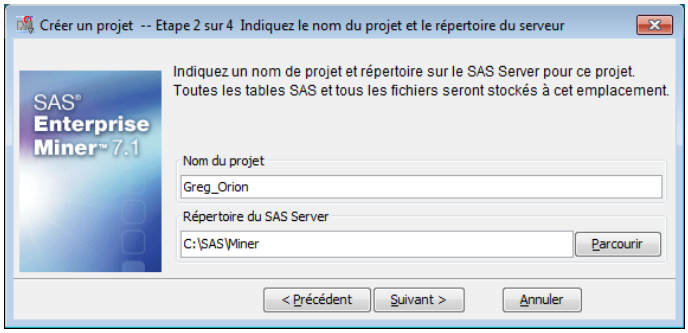
Entrer l'**Identifiant** et son **Mot de passe**
Cliquer sur **Connexion**



Créer un **nouveau projet**

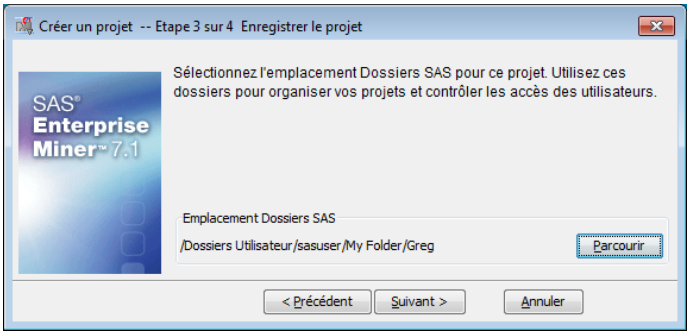


Suivant

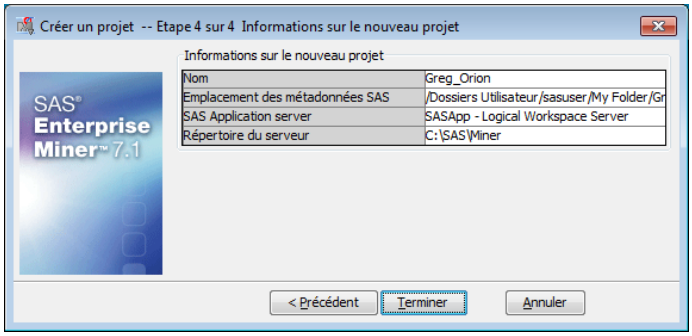


Lui donner un nom : Orion
Si c'est possible, Cliquer sur Parcourir
Sélectionner le dossier où vous souhaitez enregistrer votre projet.

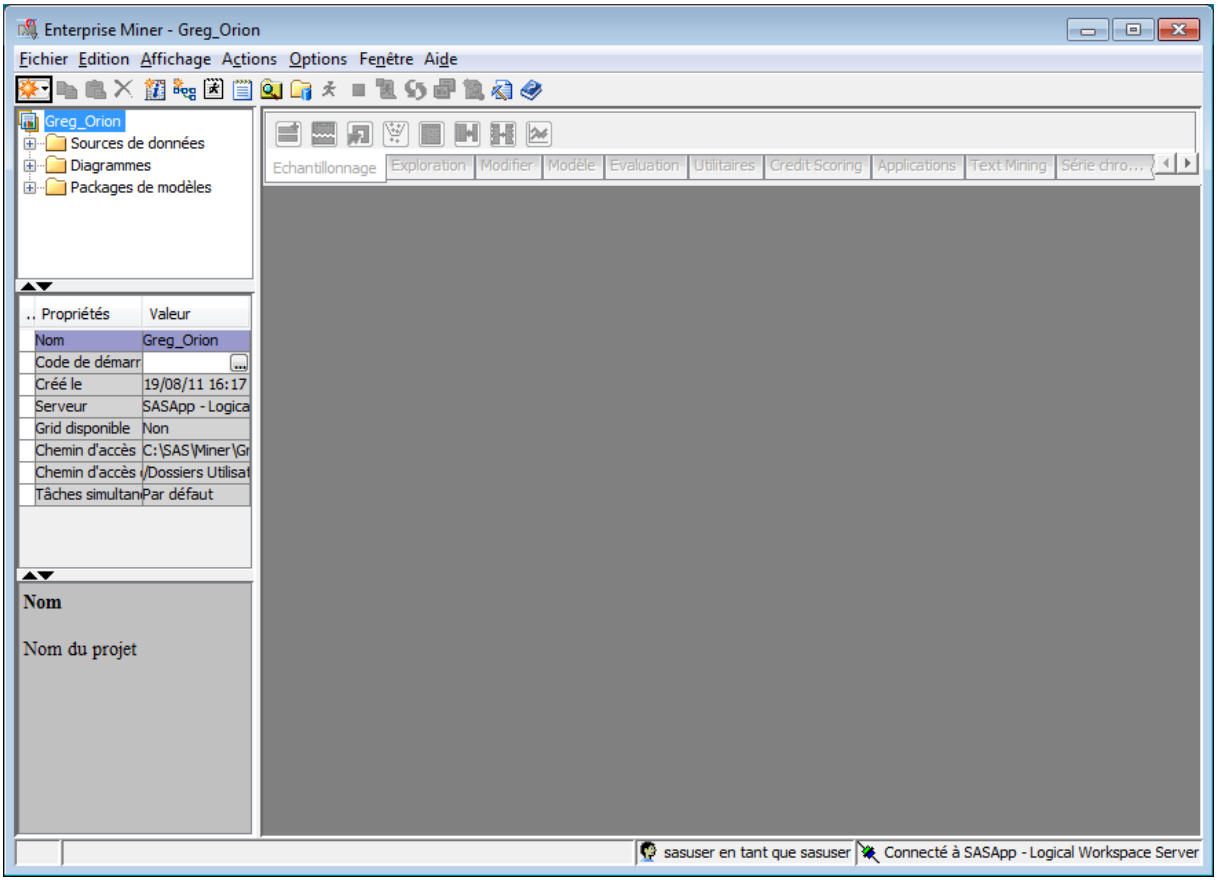
Suivant



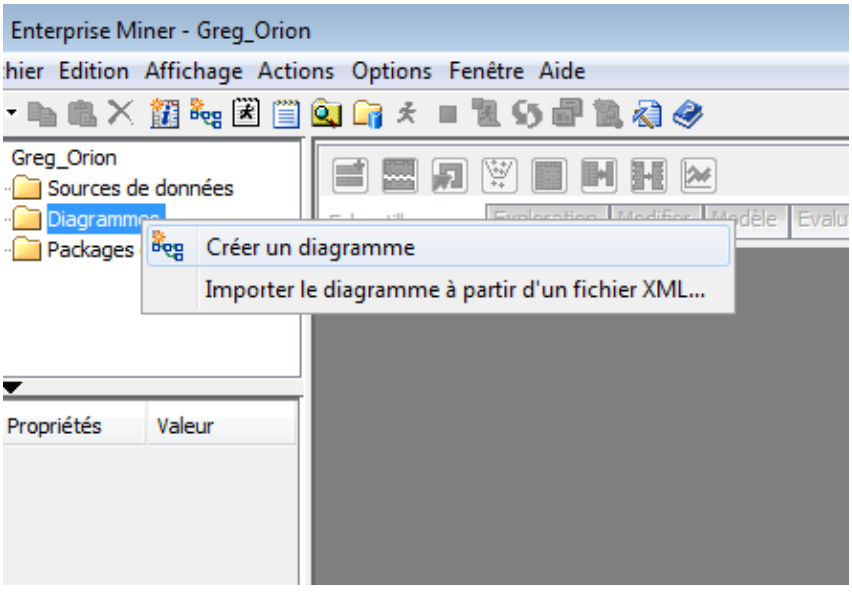
Enregistrer votre projet dans votre dossier
Suivant



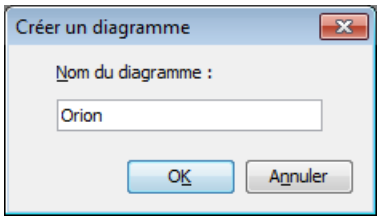
Terminer



La fenêtre ci-dessus s'ouvre.



Pour créer un nouveau diagramme, un clic droit sur **Diagrammes → Créer un diagramme**

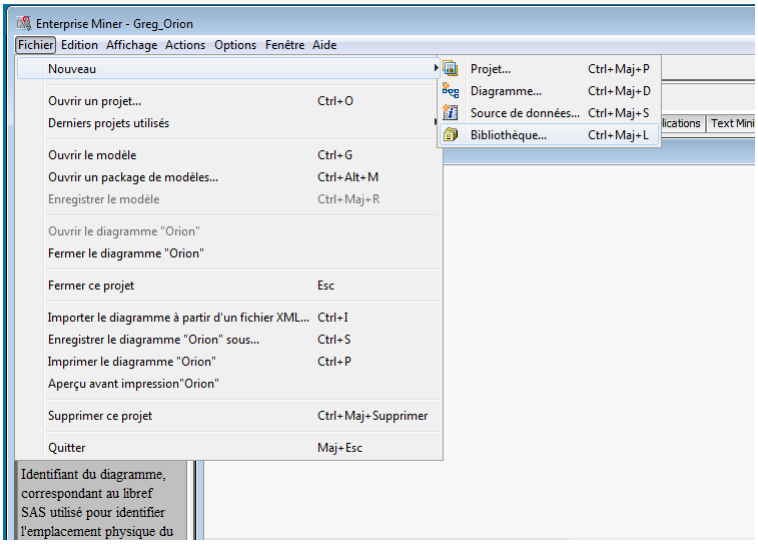


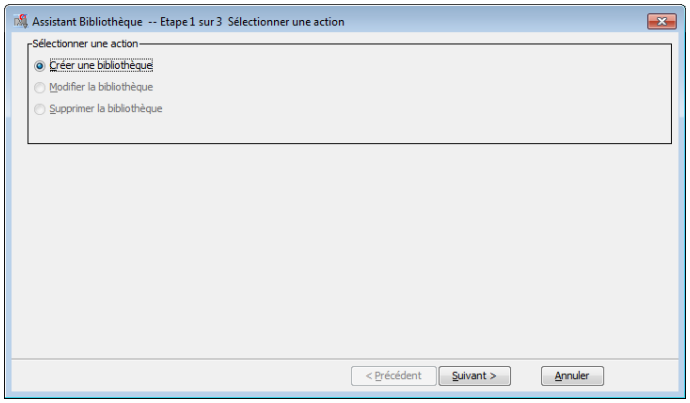
Lui donner un nom : Orion
OK

Création d'une bibliothèque :

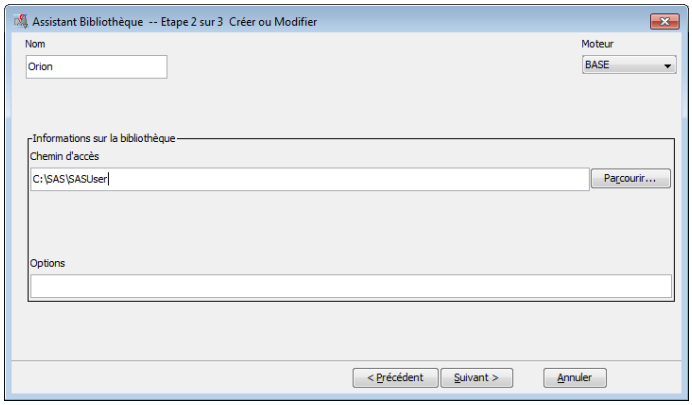
Une bibliothèque SAS est notamment un raccourci vers un dossier où se trouvent des données, des fichiers, ou vers une base de données.

Pour créer une bibliothèque pointant vers un répertoire Windows où se trouvent des fichiers de données SAS, aller dans **Fichier → Nouveau → Bibliothèque**



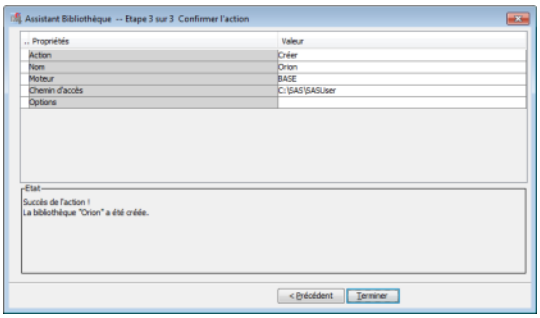


Suivant



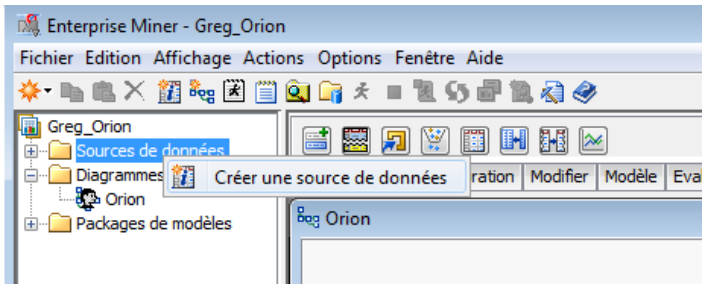
Donner un nom de maximum 8 caractères, composé uniquement de lettres de l'alphabet, de chiffres ou du souligné ; pas de caractères spéciaux et ne commençant pas par un chiffre.

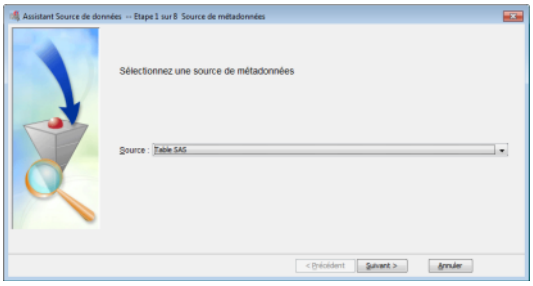
Entré le chemin vers le dossier où se trouve les tables SAS Orion_train_and_valid et Orion_a_scorer, à l'aide du bouton **Parcourir** si besoin.

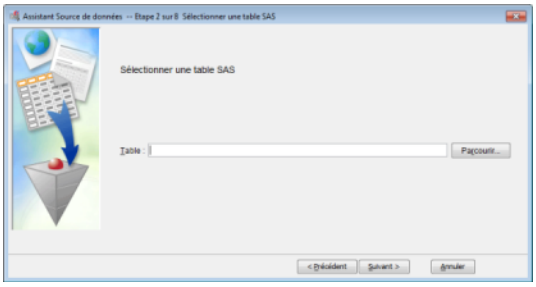


Terminer

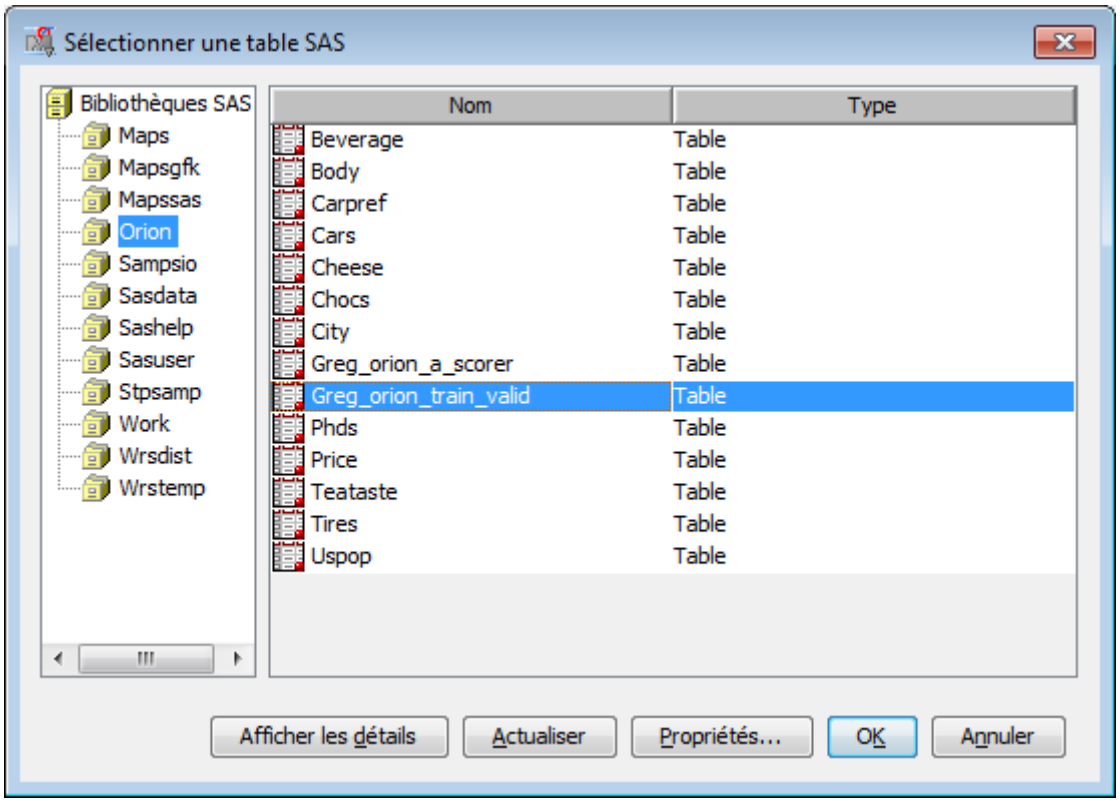
Définition de la table à utiliser





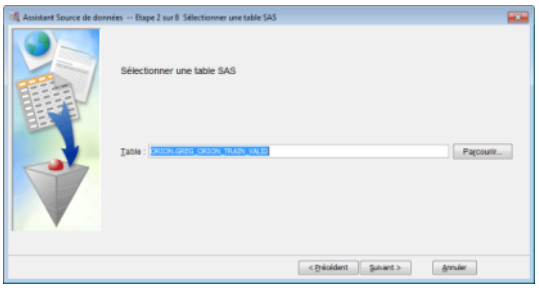


Suivant Parcourir

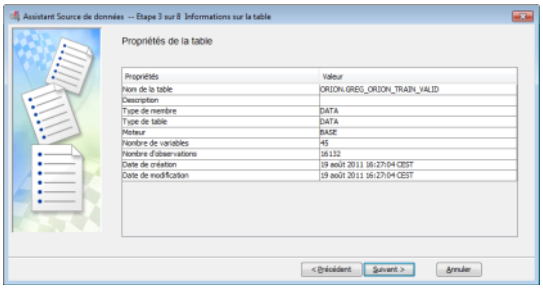


Sélectionner la bibliothèque Orion, cliquer sur actualiser pour voir apparaitre les données.

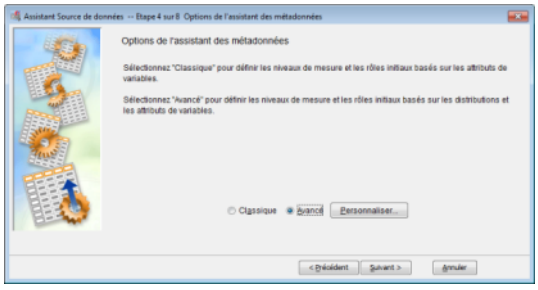
Sélectionner la table Orion_Train_And_Valid
OK



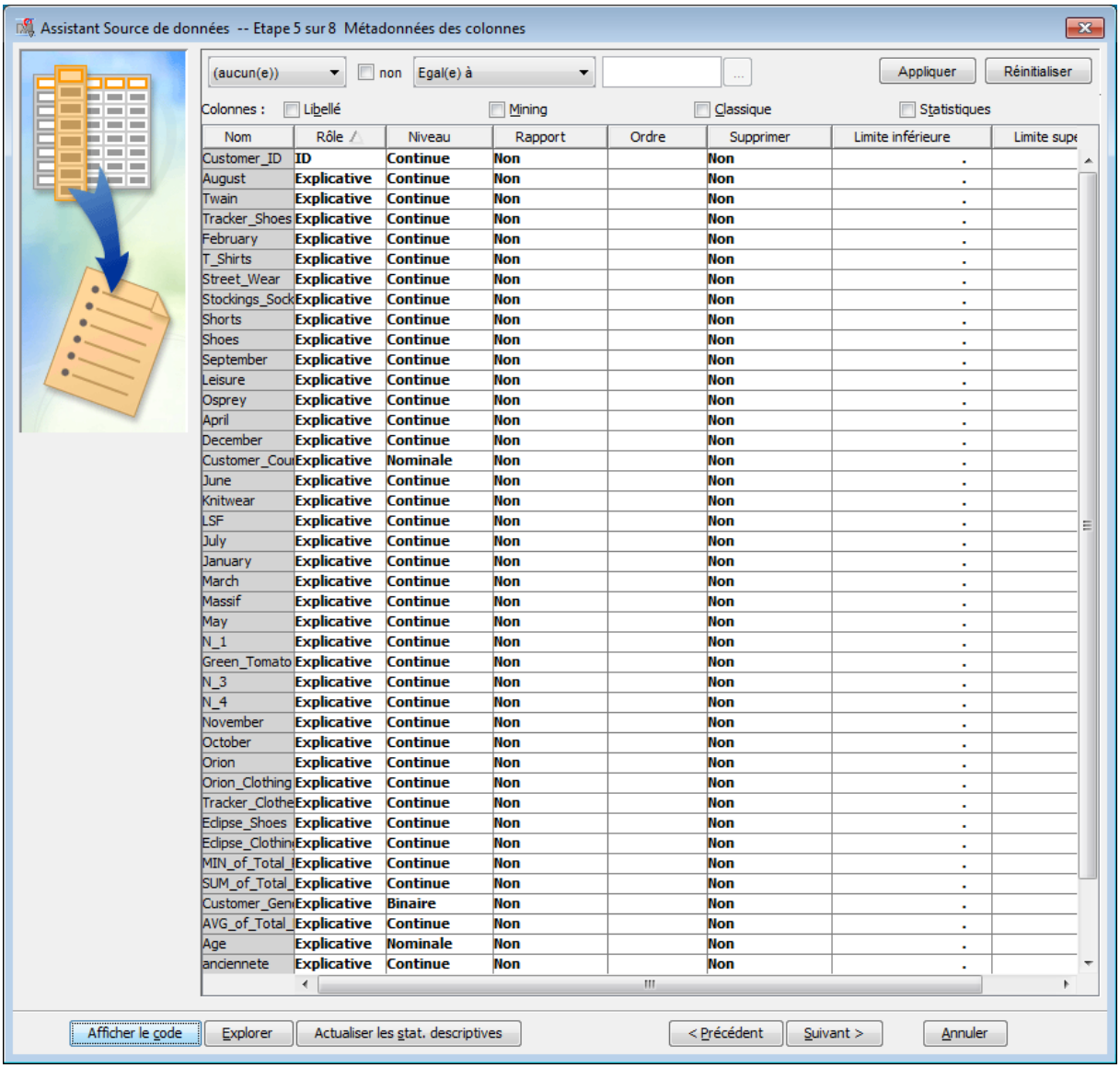
Suivant

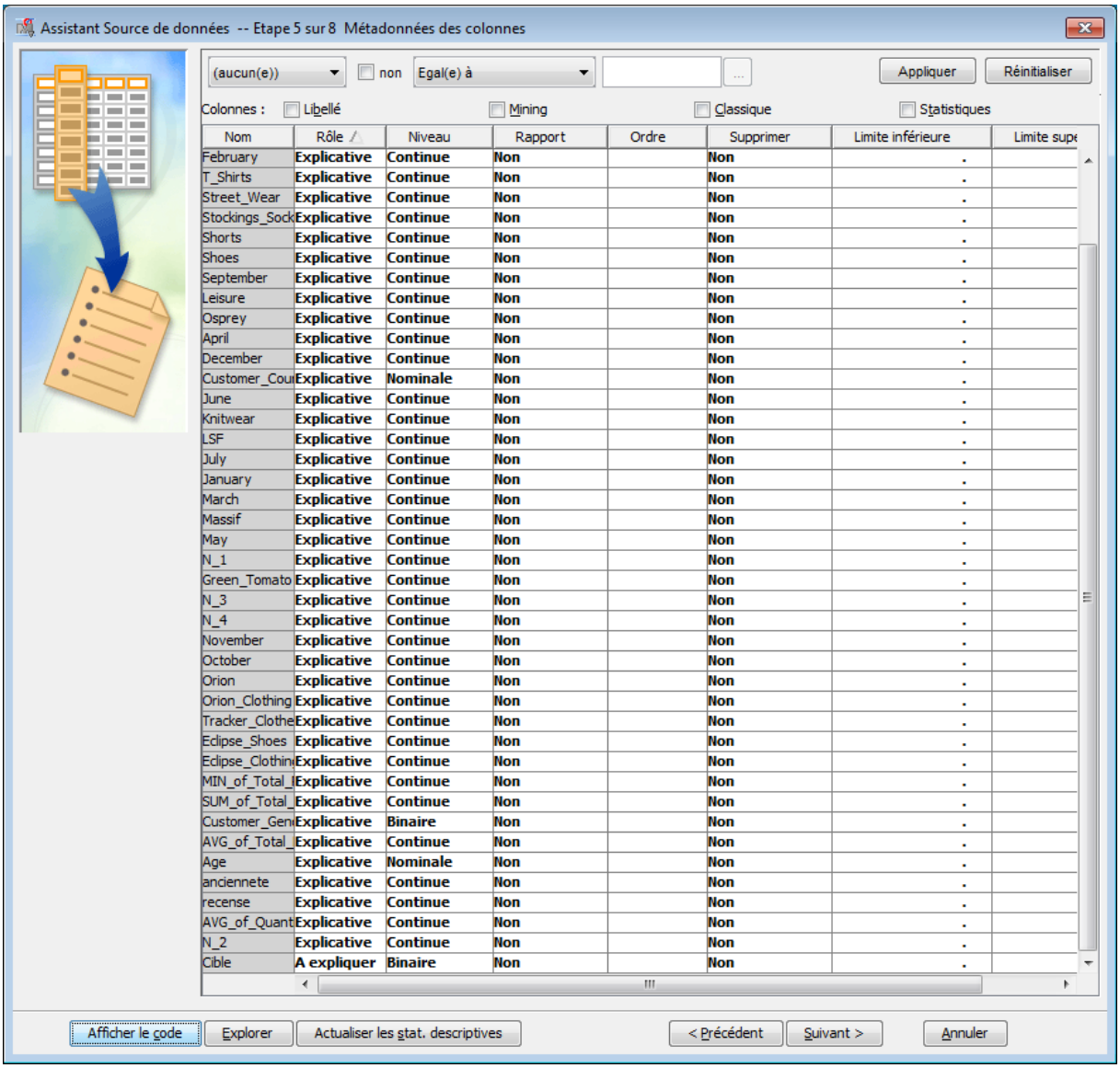


Suivant



Avancé, Suivant


Assistant Source de données -- Etape 5 sur 8 Métadonnées des colonnes
(aucun(e))
non
Egal(e) à
Appliquer
Réinitialiser
Colonnes :
Libellé
Mining
Classique
Statistiques
Nom
Rôle
Niveau
Rapport
Ordre
Supprimer
Limite inférieure
Limite supé
Customer_ID ID Continue Non Non .
August Explicative Continue Non Non .
Twain Explicative Continue Non Non .
Tracker_Shoes Explicative Continue Non Non .
February Explicative Continue Non Non .
T_Shirts Explicative Continue Non Non .
Street_Wear Explicative Continue Non Non .
Stockings_Sock Explicative Continue Non Non .
Shorts Explicative Continue Non Non .
Shoes Explicative Continue Non Non .
September Explicative Continue Non Non .
Leisure Explicative Continue Non Non .
Osprey Explicative Continue Non Non .
April Explicative Continue Non Non .
December Explicative Continue Non Non .
Customer_Cou Explicative Nominale Non Non .
June Explicative Continue Non Non .
Knitwear Explicative Continue Non Non .
LSF Explicative Continue Non Non .
July Explicative Continue Non Non .
January Explicative Continue Non Non .
March Explicative Continue Non Non .
Massif Explicative Continue Non Non .
May Explicative Continue Non Non .
N_1 Explicative Continue Non Non .
Green_Tomato Explicative Continue Non Non .
N_3 Explicative Continue Non Non .
N_4 Explicative Continue Non Non .
November Explicative Continue Non Non .
October Explicative Continue Non Non .
Orion Explicative Continue Non Non .
Orion_Clothing Explicative Continue Non Non .
Tracker_Clothe Explicative Continue Non Non .
Eclipse_Shoes Explicative Continue Non Non .
Eclipse_Clothin Explicative Continue Non Non .
MIN_of_Total_I Explicative Continue Non Non .
SUM_of_Total_ Explicative Continue Non Non .
Customer_Gen Explicative Binaire Non Non .
AVG_of_Total_ Explicative Continue Non Non .
Age Explicative Nominale Non Non .
ancienne te Explicative Continue Non Non .
Afficher le code
Explorer
Actualiser les stat. descriptives
< Précédent
Suivant >
Annuler

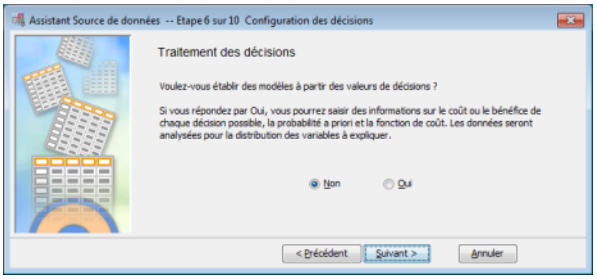


Sélectionner toutes les variables rejetées (sauf la cible) et affecter le rôle Explicative.

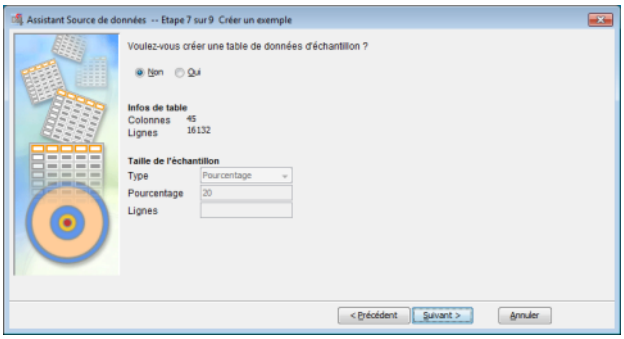
La variable Cible est une variable à expliquer.

On obtient donc des variables explicatives, un identifiant : Customer_ID ; et une variable à expliquer : Cible.

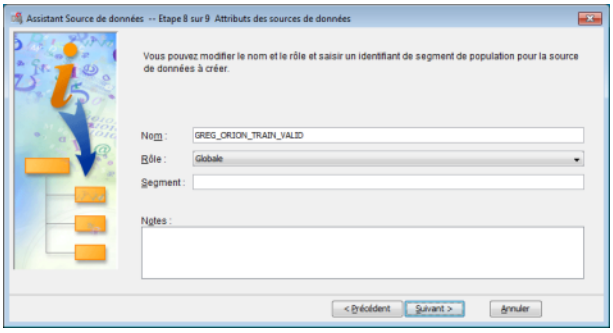
Suivant



Suivant

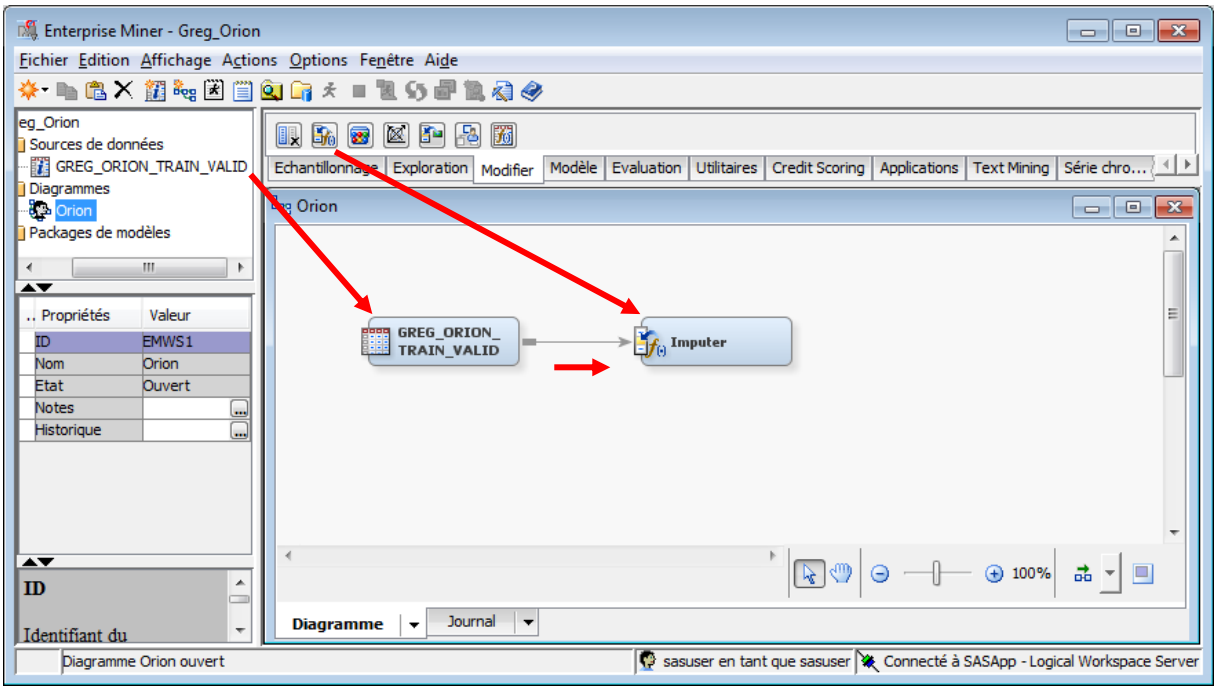


Suivant

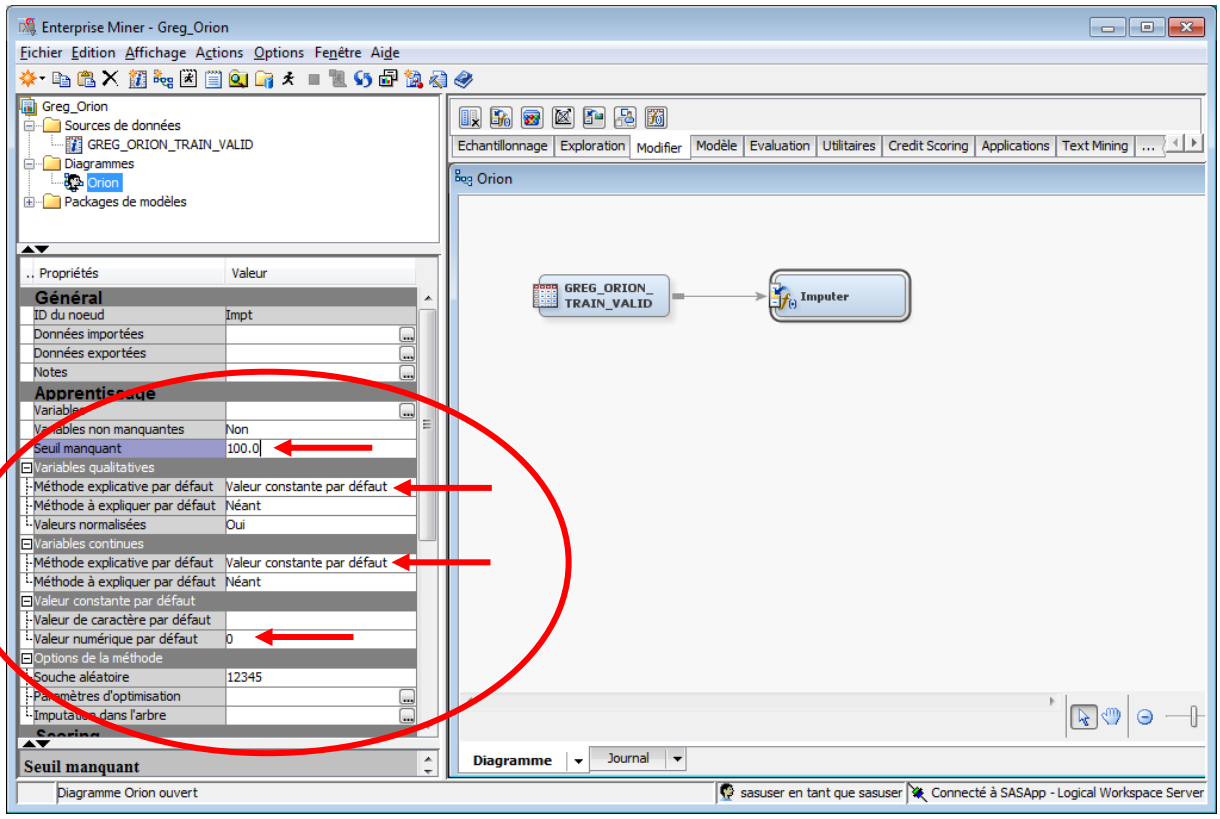


Suivant,
Terminer

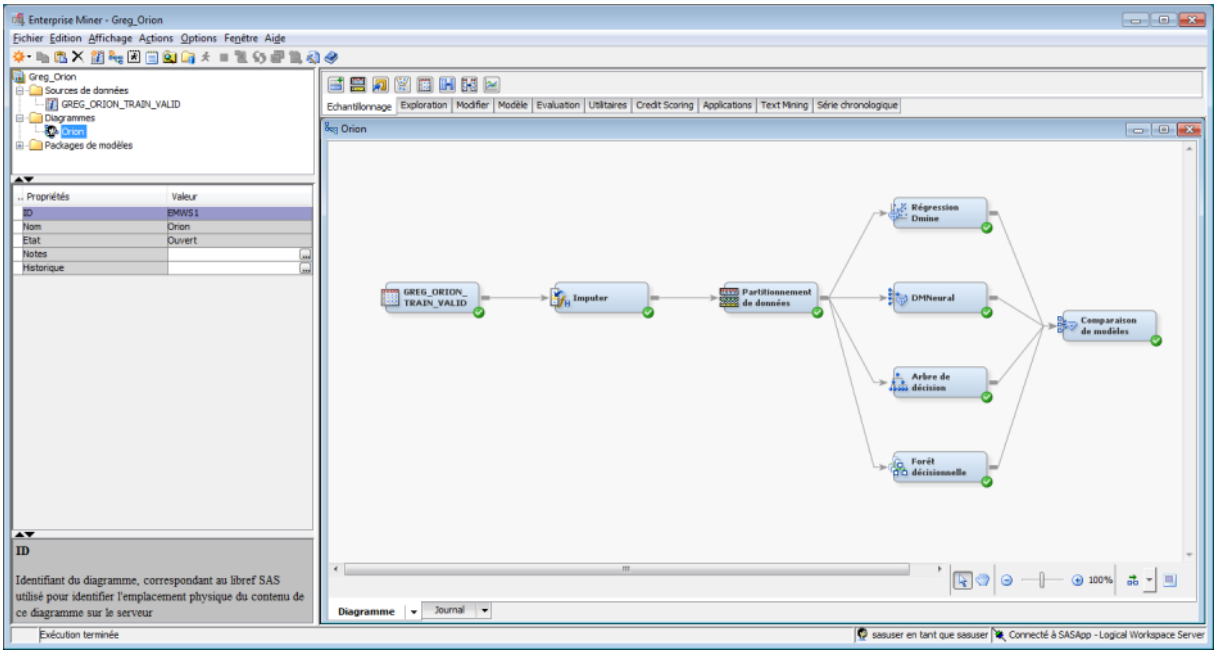
Dans le diagramme Orion, ajouter la table Orion_train_and_valid, ajouter l'outil d'imputation (depuis l'onglet Modification) et lier les deux outils.



Sélectionner l'outil d'imputation,

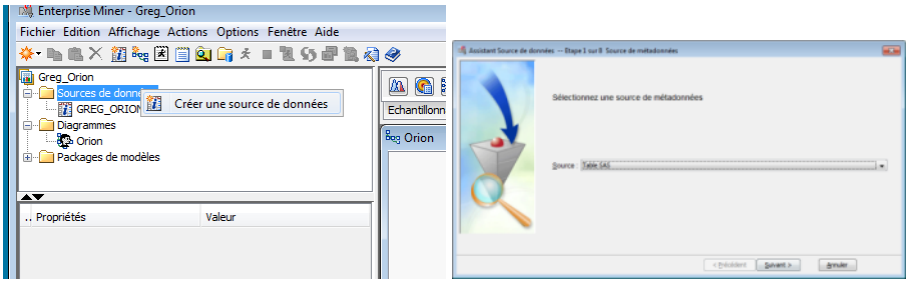


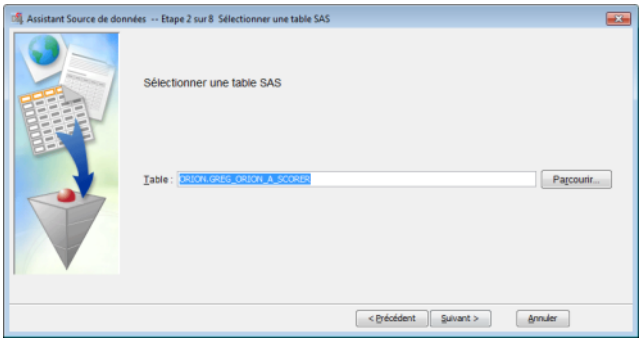
Entrer le seuil manquant à 100%
Entrer la valeur constante par défaut comme méthode explicative par défaut des variables qualitatives et continues.
Saisir la valeur 0 comme valeur numérique constante par défaut.

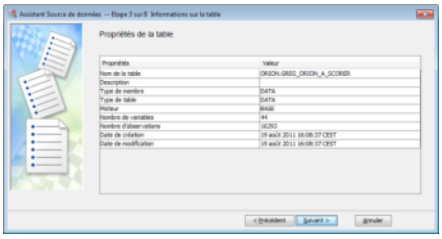


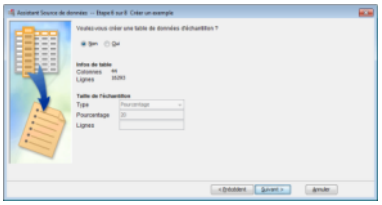
Construire un processus comme celui-ci-dessus.

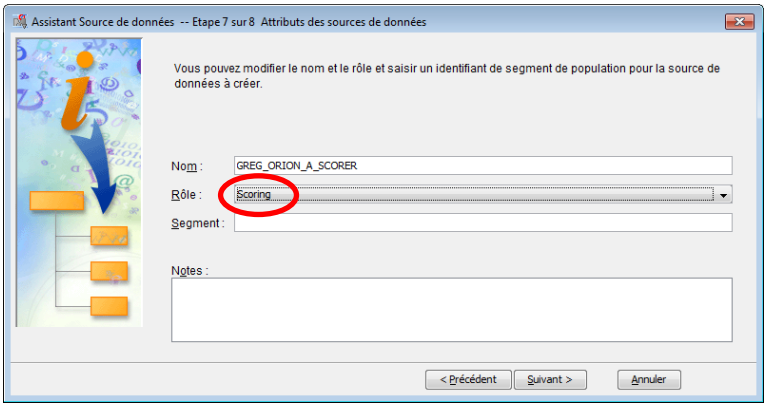
Pour le partitionnement des données, 60% pour apprendre, 40% pour valider.

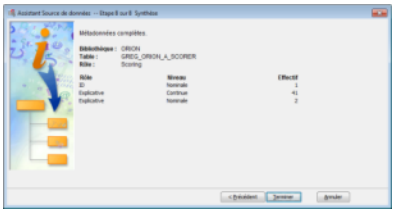




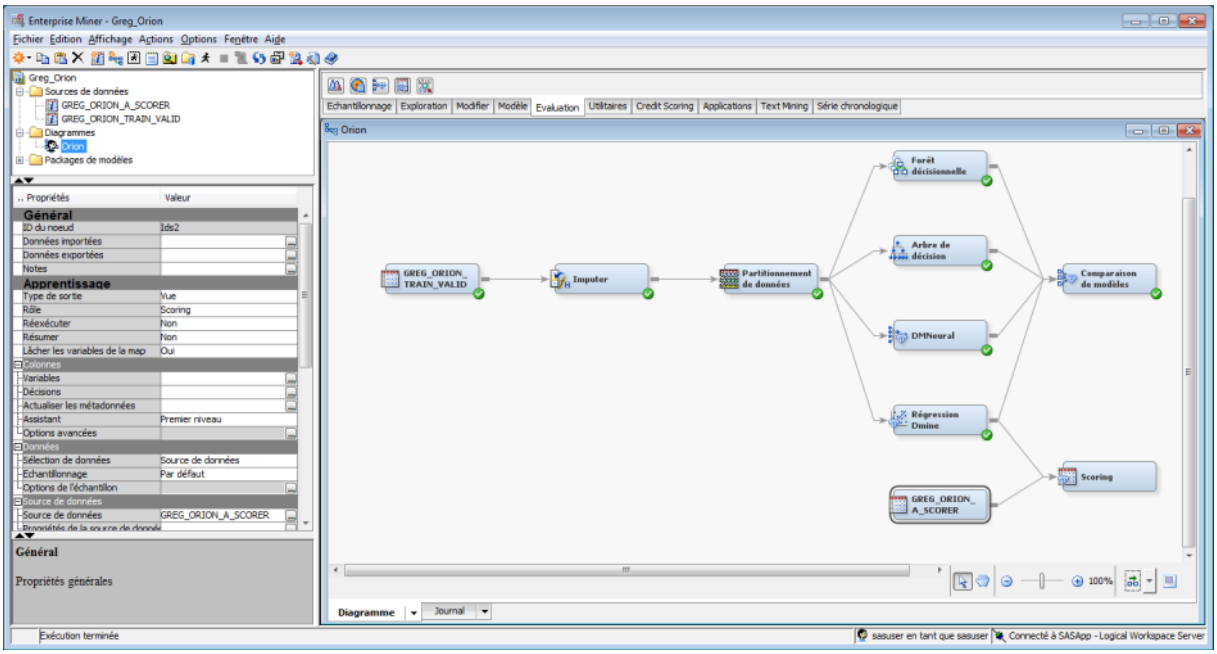








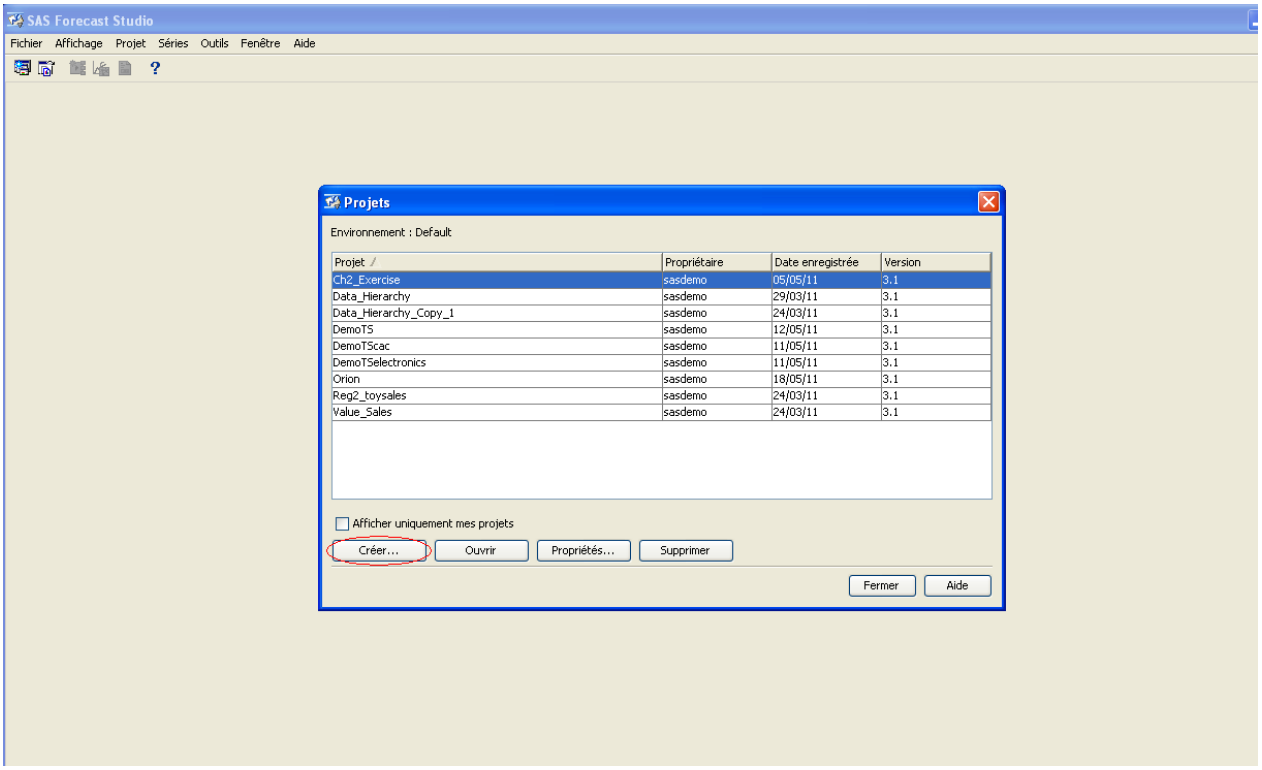
Appliquer votre meilleur modèle à votre table à scorer.


Enterprise Miner - Greg_Orion
Fichier Edition Affichage Actions Options Fenêtre Aide
Greg_Orion
Sources de données
GREG_ORION_A_SCORER
GREG_ORION_TRAIN_VALID
Diagrammes
Orion
Packages de modèles
Propriétés
Valeur
Général
ID du noeud
Ids2
Données importées
Données exportées
Notes
Apprentissage
Type de sortie
Vue
Rôle
Scoring
Réexécuter
Non
Résumer
Non
Lâcher les variables de la map
Oui
Colonnes
Variables
Décisions
Actualiser les métadonnées
Assistant
Premier niveau
Options avancées
Données
Sélection de données
Source de données
Échantillonnage
Par défaut
Options de l'échantillon
Source de données
Source de données
GREG_ORION_A_SCORER
Général
Propriétés générales
Echantillonnage
Exploration
Modifier
Modèle
Evaluation
Utilitaires
Credit Scoring
Applications
Text Mining
Série chronologique
Orion
GREG_ORION_ TRAIN_VALID
Imputer
Partitionnement de données
Forêt décisionnelle
Arbre de décision
DMNeural
Régression Dmine
Comparaison de modèles
Scoring
GREG_ORION_ A_SCORER
100%
Diagramme
Journal
Exécution terminée
sasuser en tant que sasuser
Connecté à SASApp - Logical Workspace Server

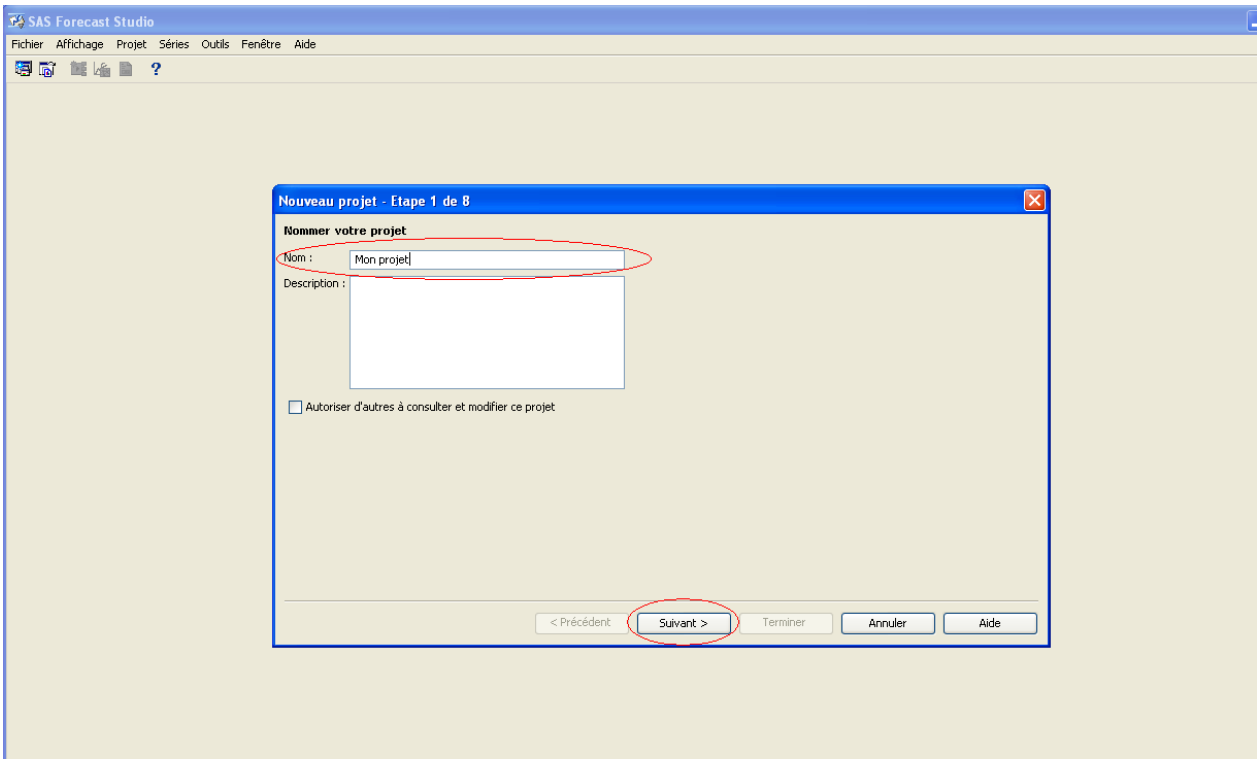
Guide pratique d'utilisation de SAS Forecast Studio 3.1

Ce chapitre a été réalisé par Jérémy Noël.
Réalisation de prévisions en fonction d'une variable qualitative de classification.

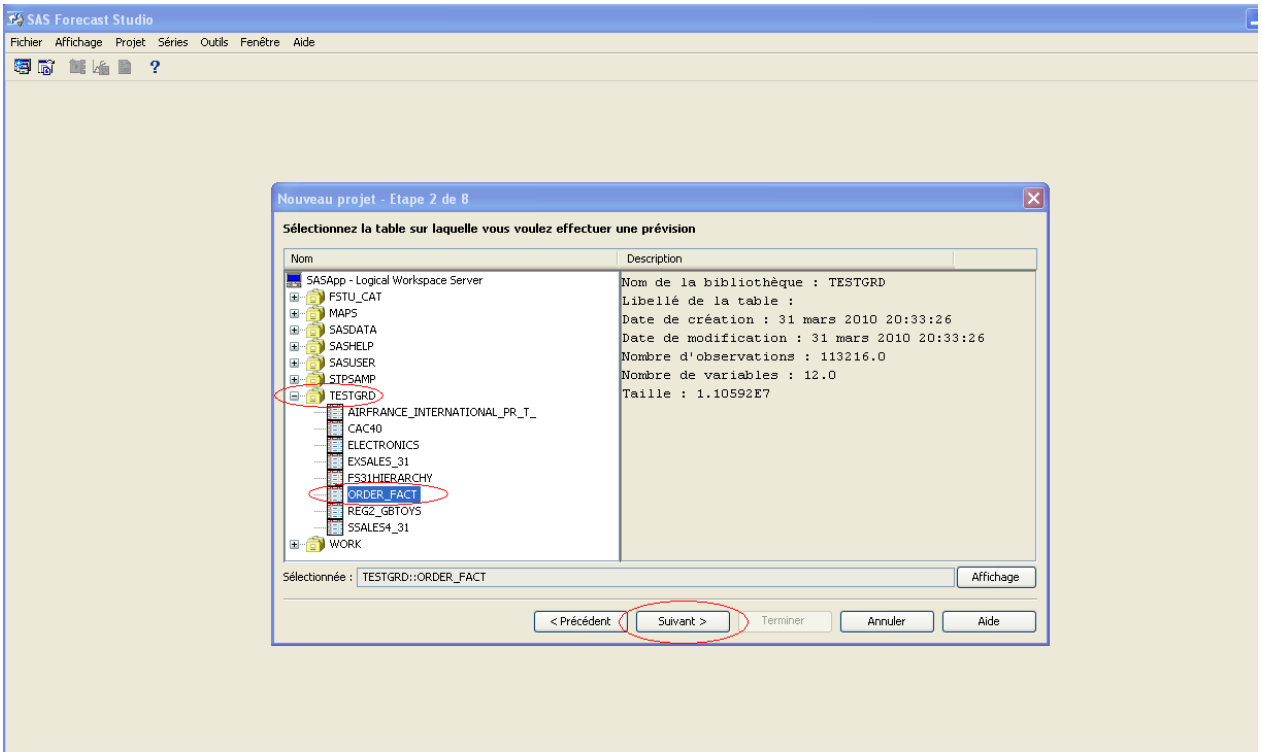
Les données utilisées pour illustrées ce guide pratique, sont les données de l'étude de cas « Orion Star », un exemple fictif, propriété de l'éditeur de logiciel SAS, Copyright© 2003 SAS Institute Inc., Cary, NC, USA. La société vend des articles de sport et d'extérieur par différents canaux : Magasins, Catalogue et Internet.
Après avoir ouvert SAS Forecast Studio, vous arrivez sur la page d'accueil avec la liste de tous les projets existants, et quelques boutons d'actions. **Cliquez** sur « Créer… ».



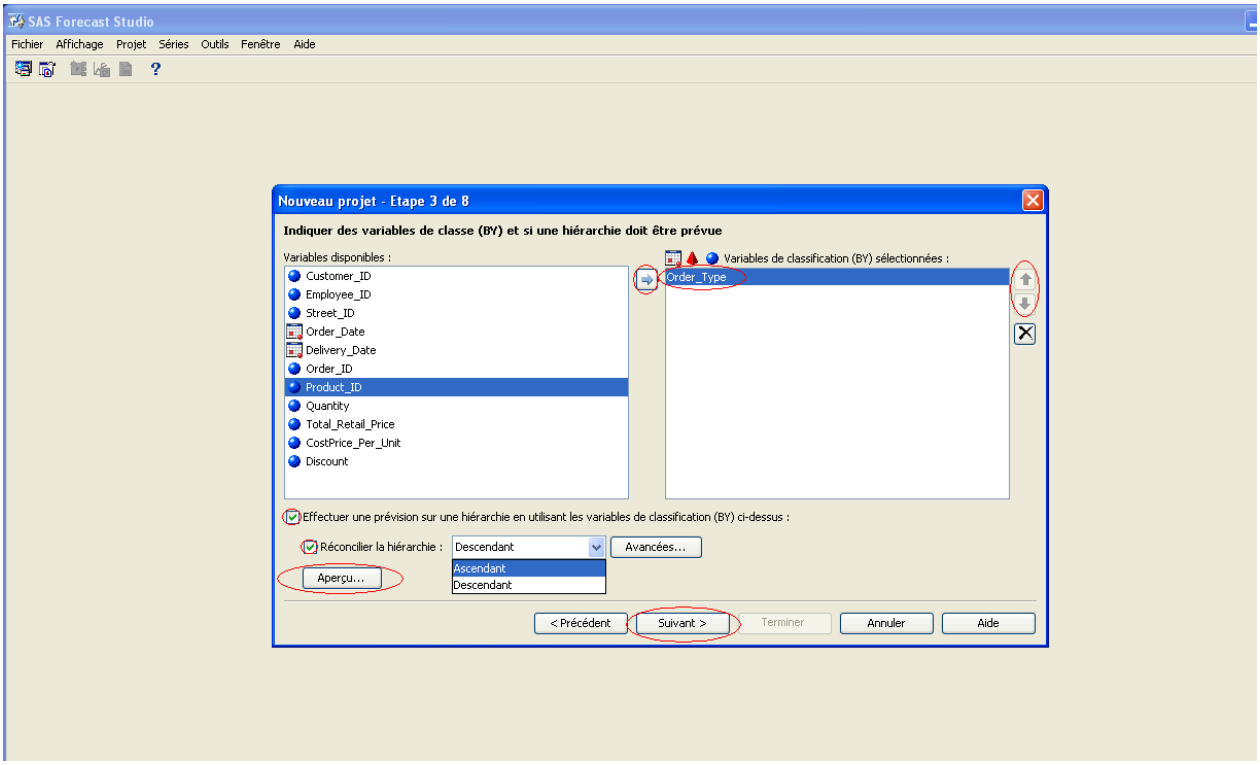
Tapez le nom de vous voulez donner au projet, par exemple « Mon_projet », puis **cliquez** sur « Suivant >».


SAS Forecast Studio
Fichier Affichage Projet Séries Outils Fenêtre Aide
Nouveau projet - Etape 1 de 8
Nommer votre projet
Nom :
Mon projet
Description :
Autoriser d'autres à consulter et modifier ce projet
< Précédent
Suivant >
Terminer
Annuler
Aide

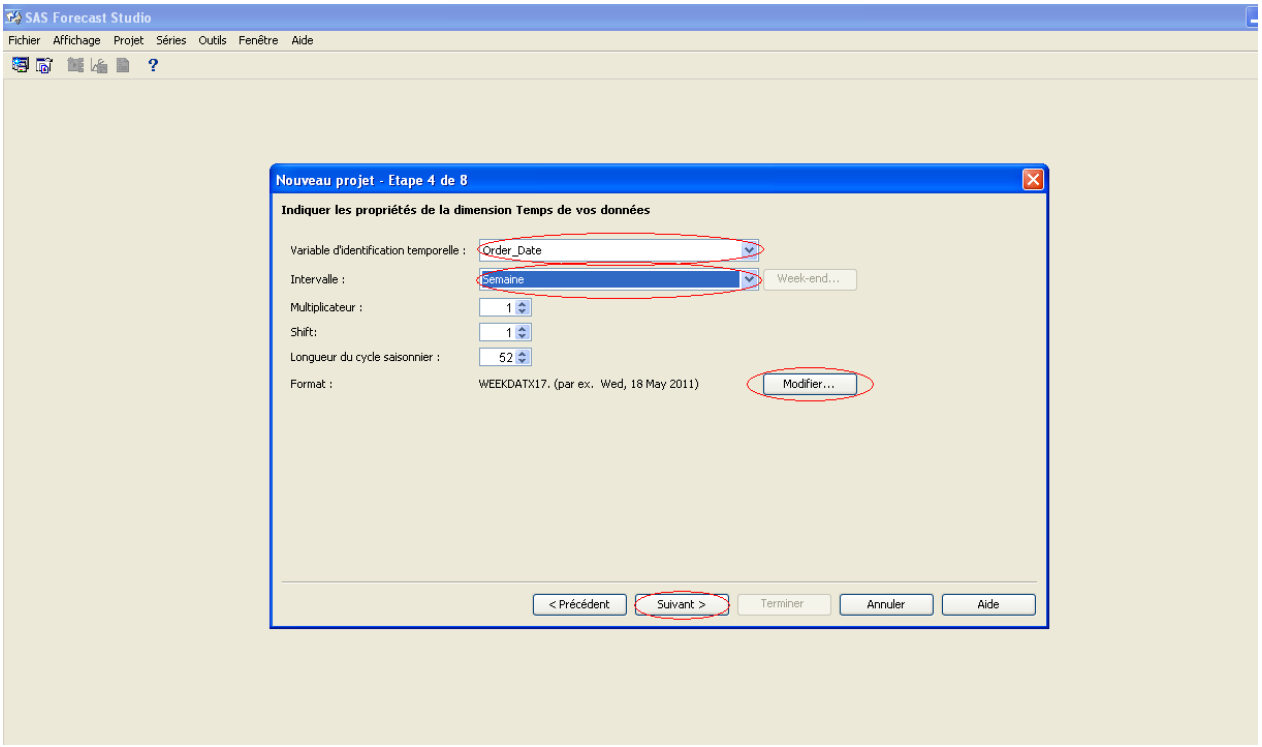
Sélectionnez la bibliothèque où la table se trouve, puis la table elle-même. Dans notre cas la bibliothèque Orion Star Gold, puis la table ORDER_FACT. On peut remarquer que la partie droite de la fenêtre contient une brève description de la table sélectionnée. **Cliquez** sur « Suivant >».



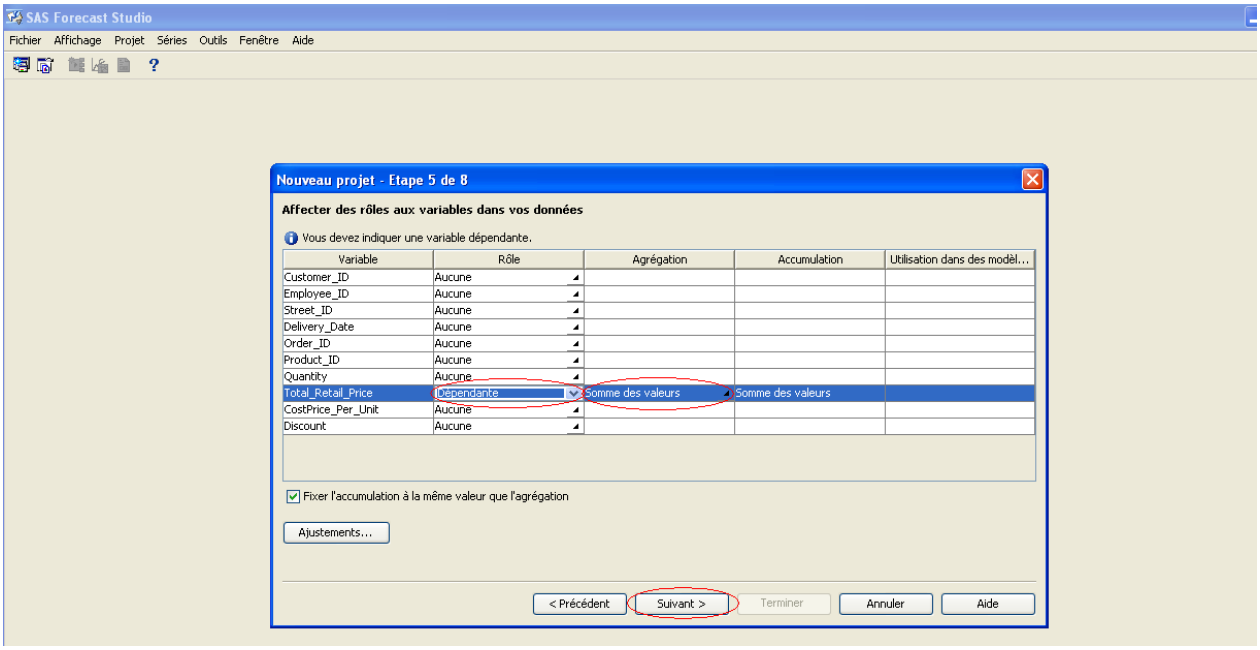
Choisissez la variable de classification, c'est-à-dire celle qui va permettre au logiciel de créer des séries en fonction des valeurs de celle-ci sur lesquelles il effectuera par la suite les différentes prévisions, puis **cliquez** sur la flèche horizontale bleue au milieu de la fenêtre. Cette dernière action va permettre de sélectionner définitivement la variable. Il est possible de sélectionner plusieurs variables de classification, les flèches de droite permettent l'ordonnancement de celles-ci, le bouton « Aperçu… » vous présente un schéma de l'organisation de vos données. Deux choses importantes sur cette étape primordial de ce type de prévisions, **vérifiez** que les deux cases en bas à droite soient bien cochées et **choisissez** le type de réconciliation qui s'adapte la mieux à vos données, dans notre cas « Ascendant » nous semble le plus approprié. **Cliquez** sur « Suivant > ».



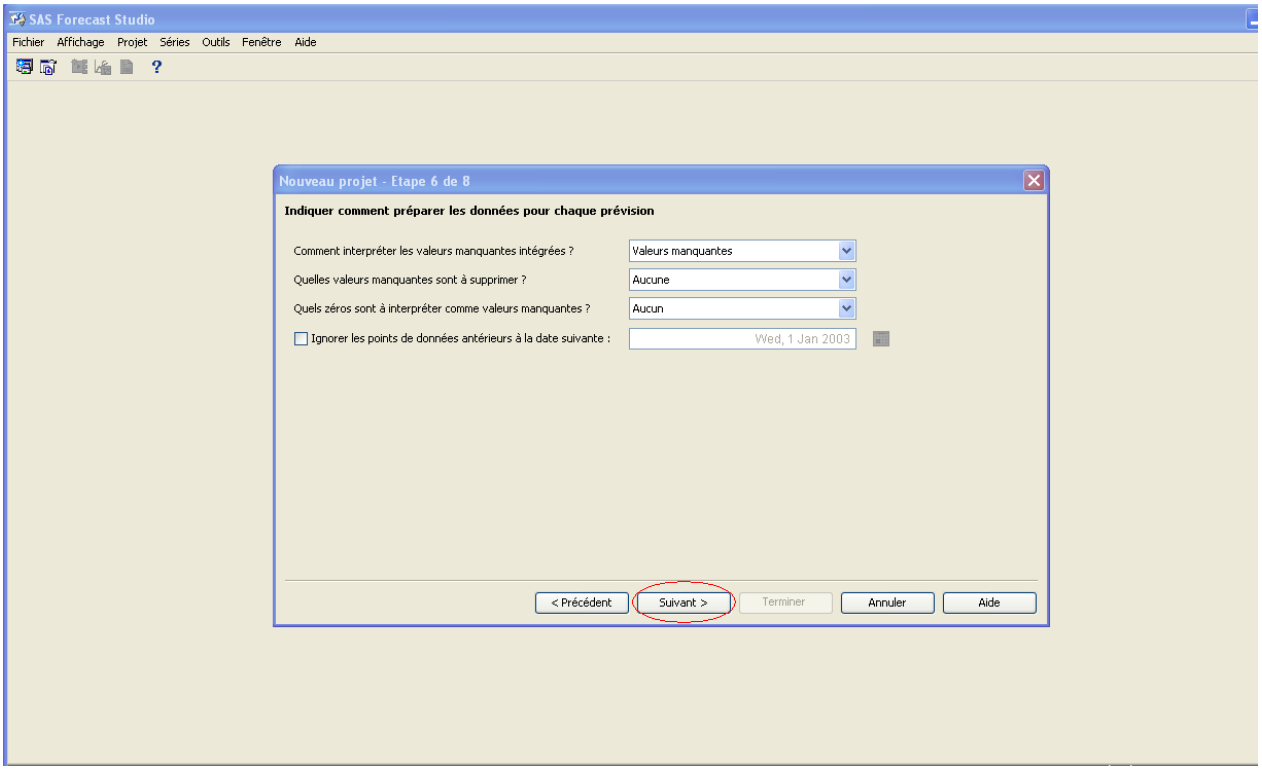
Sélectionnez la variable d'identification temporelle que vous voulez, Order_Date dans notre exemple, le logiciel déterminera s'il y arrive l'intervalle de temps dans lequel vos données sont présentées. Cependant, vous **pouvez choisir** un intervalle de temps supérieur si cela vous semble plus adéquat, avec l'application d'une option que nous verrons dans l'étape suivante. Dans notre cas, nous changeons l'intervalle de jour en semaine, qui nous apparaît être le plus adapté. Ainsi, cette option vous permet de changer de temporalité, c'est-à-dire d'intervalle de temps, et cela de manière simple, facile et efficace. Vous **pouvez aussi changer** le format de la variable si vous le souhaitez, chose que nous ne ferons pas ici, en cliquant sur le bouton « Modifier… ». **Cliquez** sur « Suivant > ».



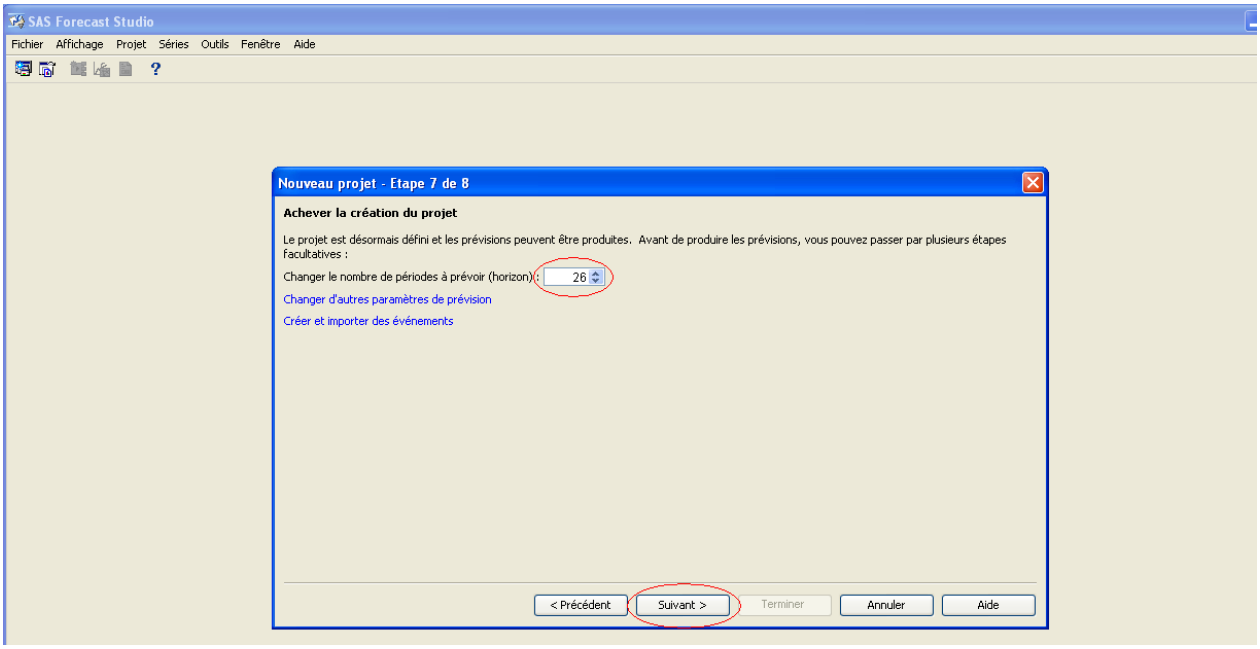
Changez le « Rôle » en « Dépendante » pour la variable que vous voulez définir comme telle, dans notre cas la variable Total_Retail_Price obtient le rôle de « Dépendante ». **Changez** le rôle des autres variables en « Indépendante » si vous voulez qu'elles deviennent des variables explicatives du modèle, nous ne changerons rien concernant les autres variables. L'option importante lorsque les données initiales ne présentent pas une valeur unique par données temporelles (plusieurs données par jour pour des données journalières ou des données pour chaque jour alors que l'on a choisi un intervalle hebdomadaire ce qui est notre choix ici) est l'agrégation des données, la somme de ces valeurs est en général, et en particulier dans notre situation, choisie Pour ce faire, **sélectionnez** « Somme des valeurs » dans la colonne « Agrégation », puis **cliquez** sur « Suivant > ».



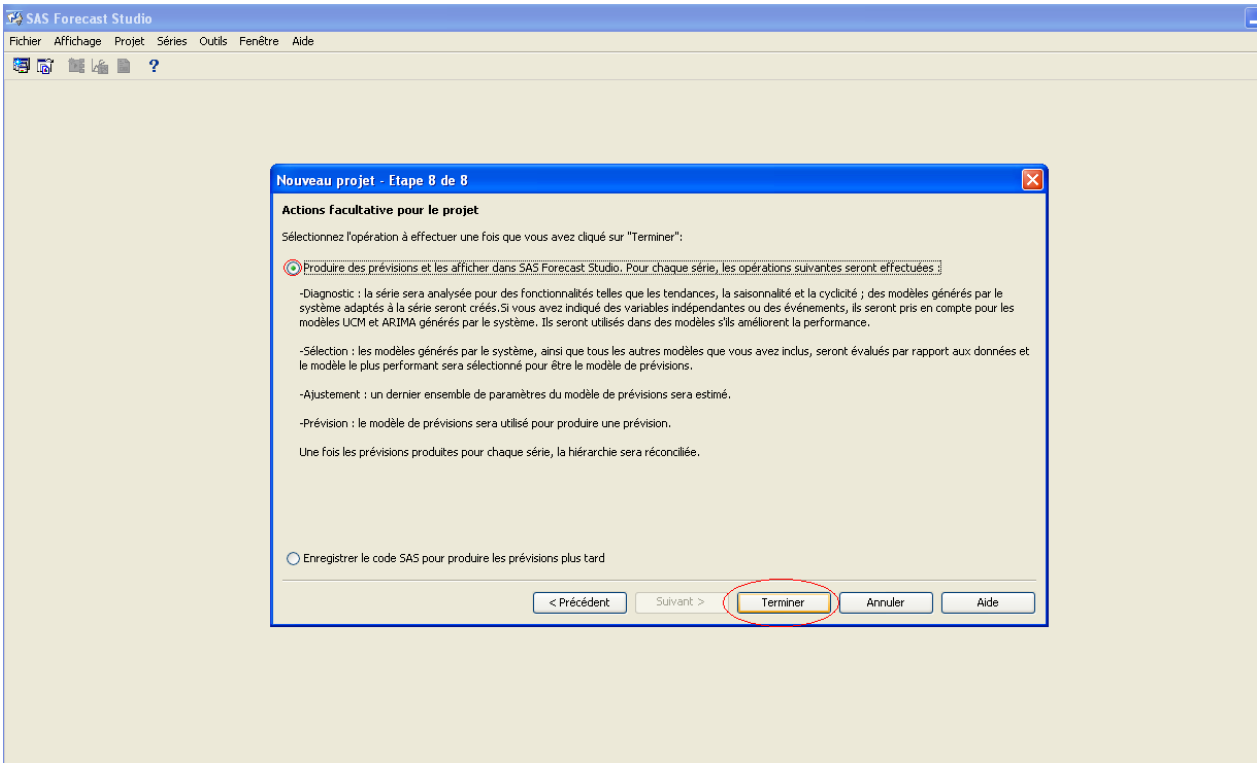
Définissez le traitement que vous voulez voir appliquer aux valeurs manquantes et à leur interprétation. Dans notre cas, les valeurs par défaut sont convenables, donc **cliquez** sur « Suivant > ».



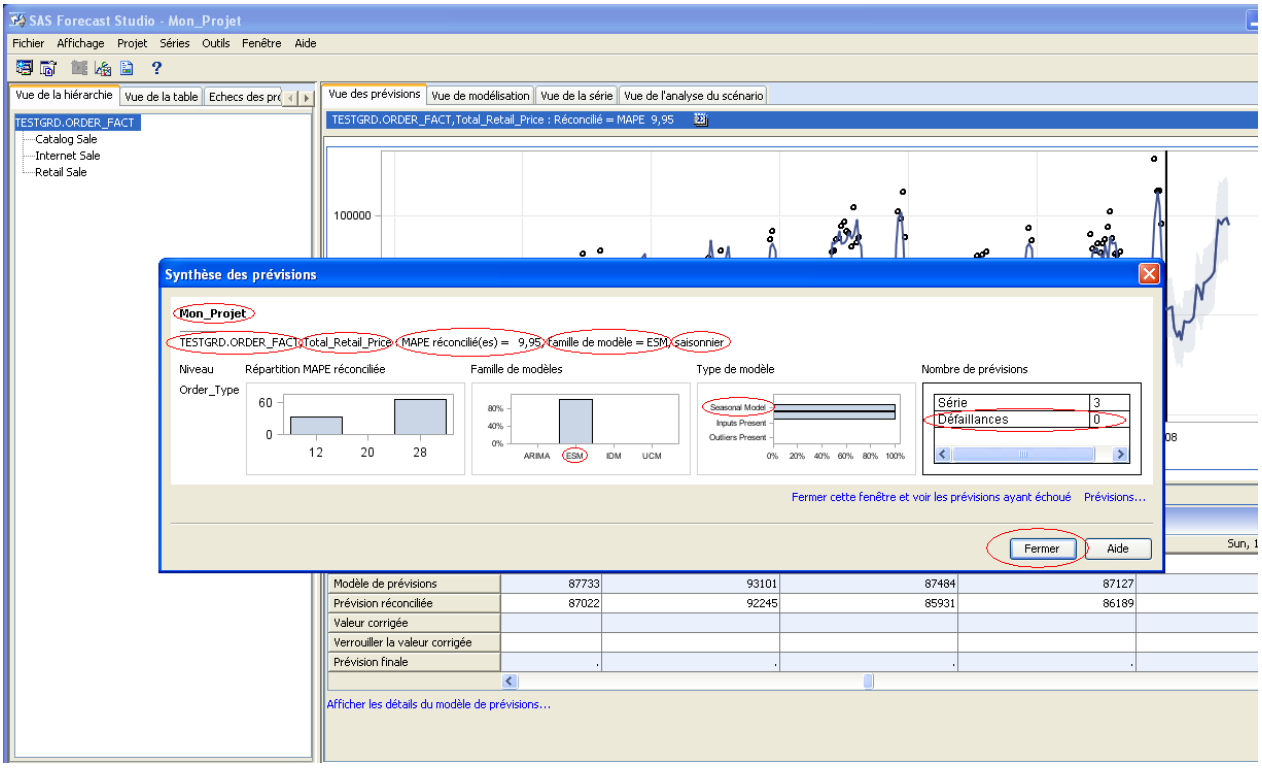
Sélectionnez le nombre de périodes, l'horizon, que le logiciel devra prévoir. **Sélectionnez** par exemple 26 qui correspond à une demie année. **Cliquez** sur « Suivant > ».



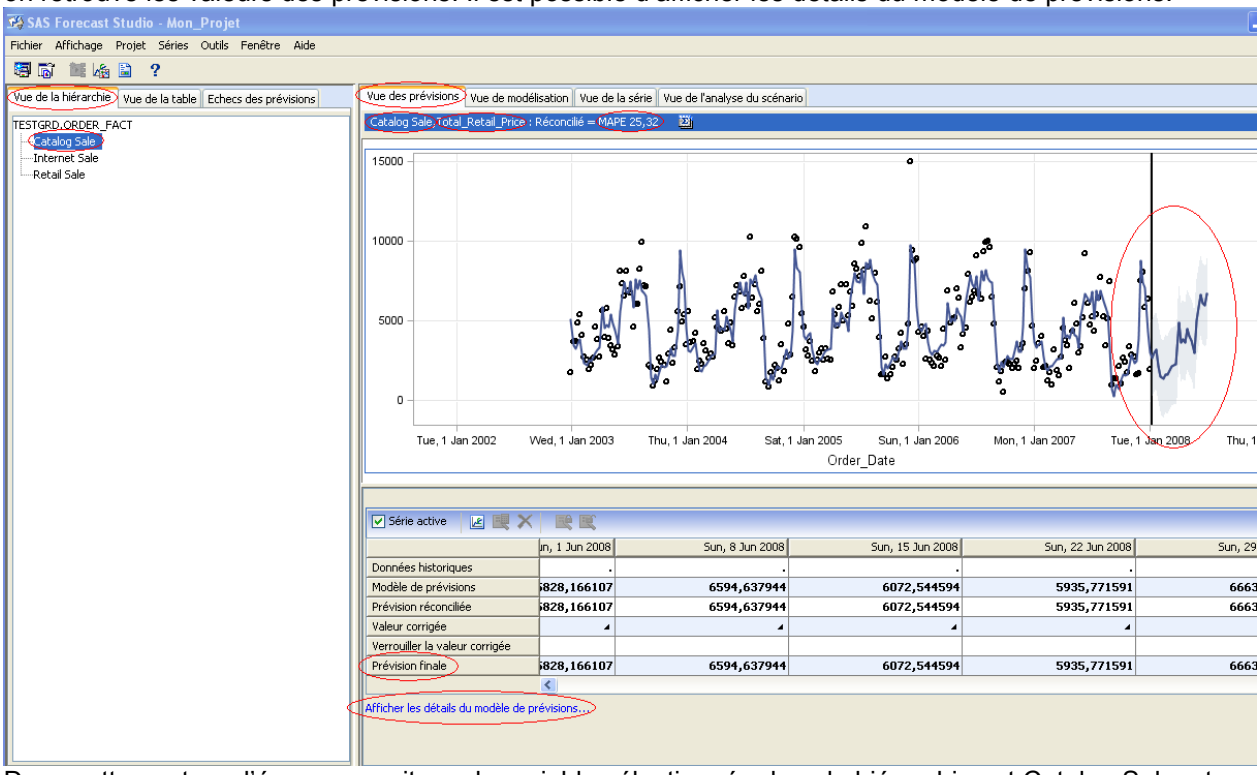
L’étape finale vous permet de sauver le code SAS que Forecast Studio aurait généré pour produire les prévisions avec les paramètres choisis et ainsi pouvoir exécuter le projet ultérieurement. **Gardez** la sélection initiale, qui va nous permettre de produire les prévisions immédiatement et les afficher dans Forecast Studio. **Cliquez** sur « Terminer ».



Le fait de cliquer lance les prévisions. Une fenêtre de « Synthèse des prévisions » apparaît. On retrouve le nom du projet en haut à gauche, la table utilisée, la variable dépendante, la valeur de la statistique MAPE réconciliée, la famille et le type de modèle. Ces trois derniers éléments sont détaillés sous forme graphique en dessous. Dans notre cas, on remarque que la MAPE d'une des trois séries (une série pour chaque valeur de la variable Order_Type) tourne autour de 12 et les deux autres aux alentours de 28. On constate que les séries modélisées sont des ESM (Exponential Smoothing Model) saisonniers. Le dernier tableau indique qu'il n'y a eu aucune défaillance lors des prévisions. **Fermez** cette fenêtre.

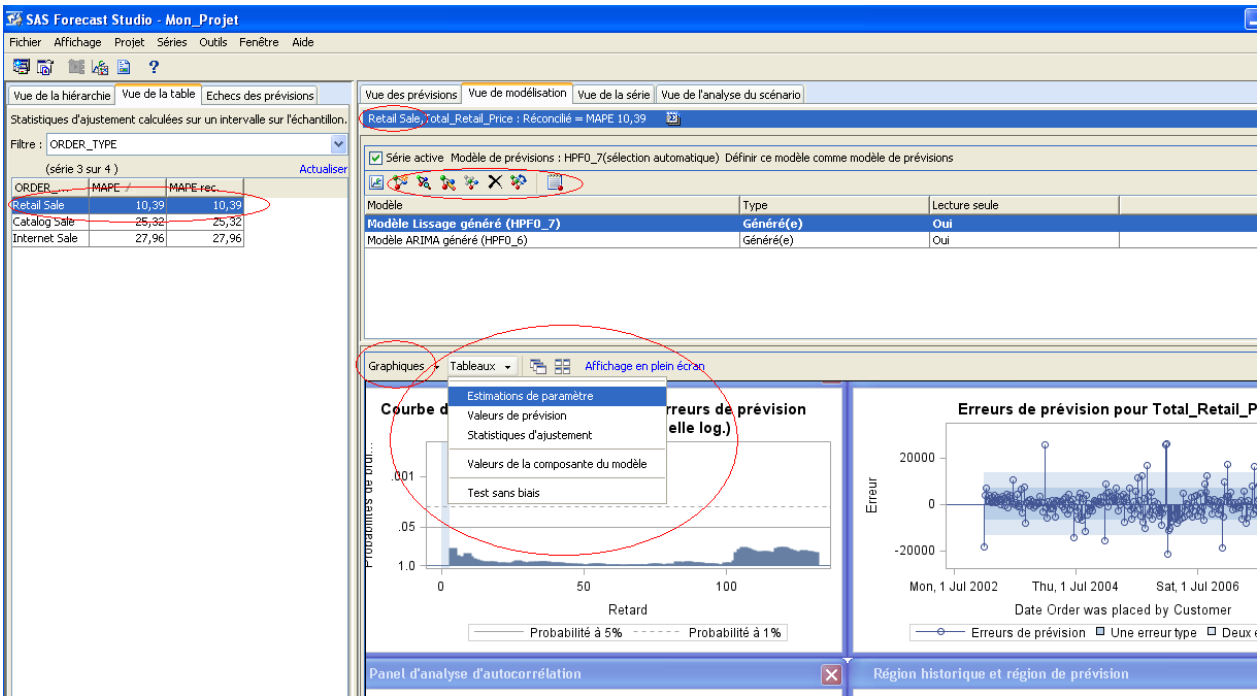


La page qui s'affiche est décomposée en plusieurs parties. A gauche, nous retrouvons les onglets permettant de choisir la série à afficher. Le premier onglet offre une « Vue de la hiérarchie », tandis que le deuxième offre une « Vue de la table » permettant le filtrage des séries et quelques statistiques d'ajustements et enfin, le dernier concerne les « Echecs des prévisions » et est donc inutile ici. La partie de droite, qui est la partie principale est aussi décomposée en quatre onglets. Le premier offre une « Vue des prévisions », avec en haut la variable ou la table contenant la série à modéliser, la variable sur laquelle est faite les prévisions et la valeur de la MAPE. En dessous se trouve un graphique représentant les prévisions du modèle et l'intervalle de confiance associé. En bas on retrouve les valeurs des prévisions. Il est possible d'afficher les détails du modèle de prévisions.



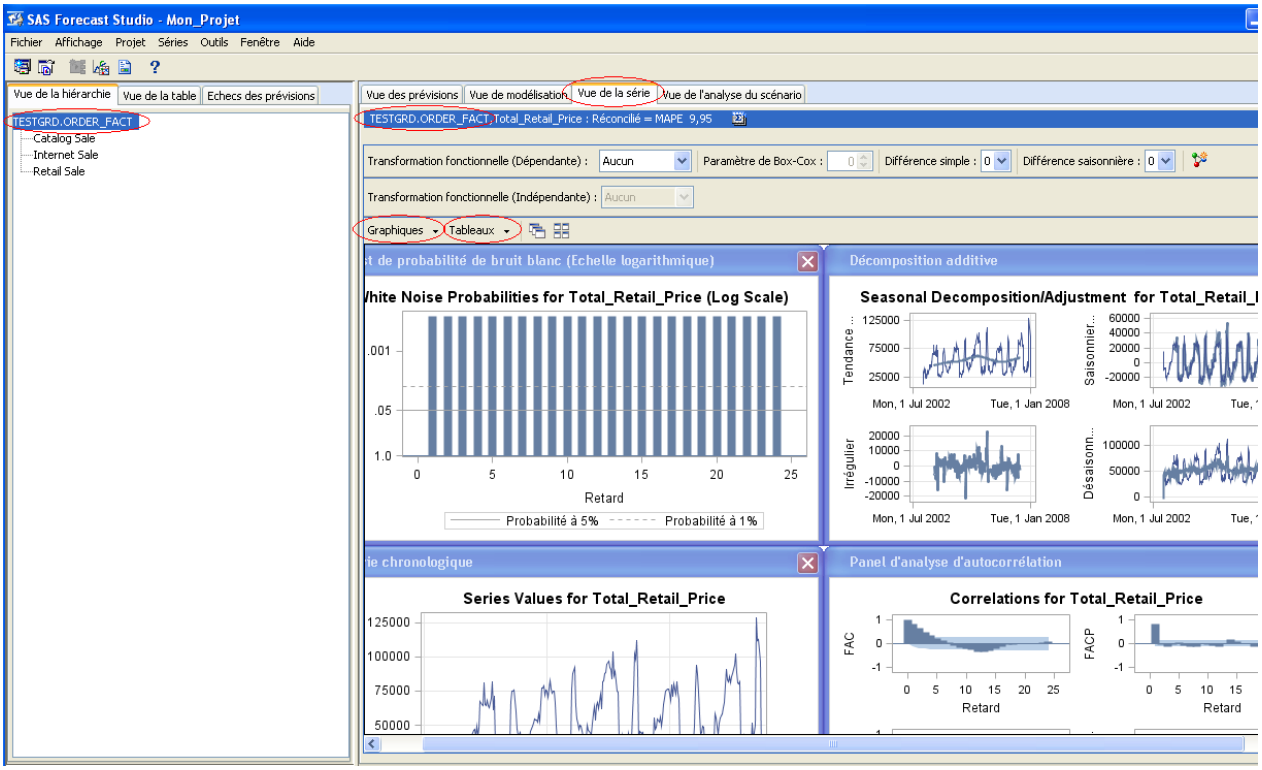
Dans cette capture d'écran, on voit que la variable sélectionnée dans la hiérarchie est Catalog Sale, et donc dans la partie de droite on constate que cette série est modélisée.

Dans la « Vue de modélisation », on retrouve les mêmes informations tout en haut. La partie dessous offre divers possibilités d'actions, comme la création ou modification de modèle ou la comparaison de modèles, et l'affichage de quelques informations générales sur les modèles générés. La partie du bas affiche les graphiques et tableaux que vous souhaitez voir.



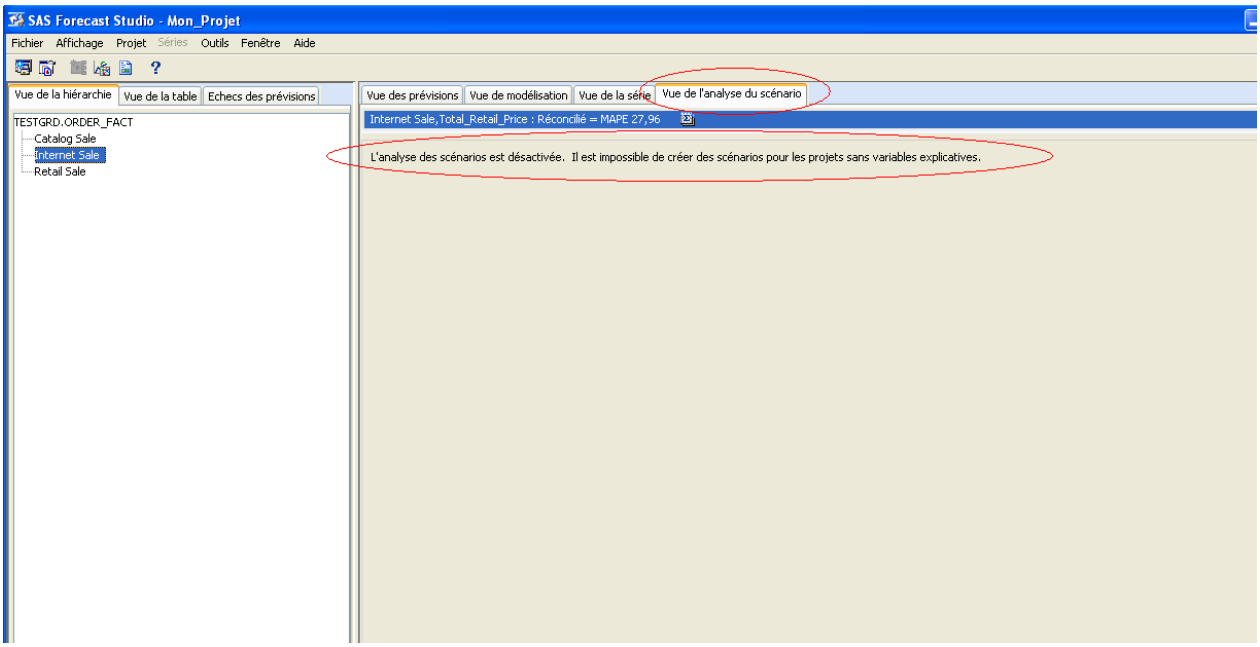
On constate que la série sélectionnée est celle qui est en haut du tableau dans la « Vue de la table ». Les graphiques et tableaux font référence au modèle sélectionné dans la partie au-dessus.

Le troisième onglet, « Vue de la série », offre quelques graphiques concernant la série.

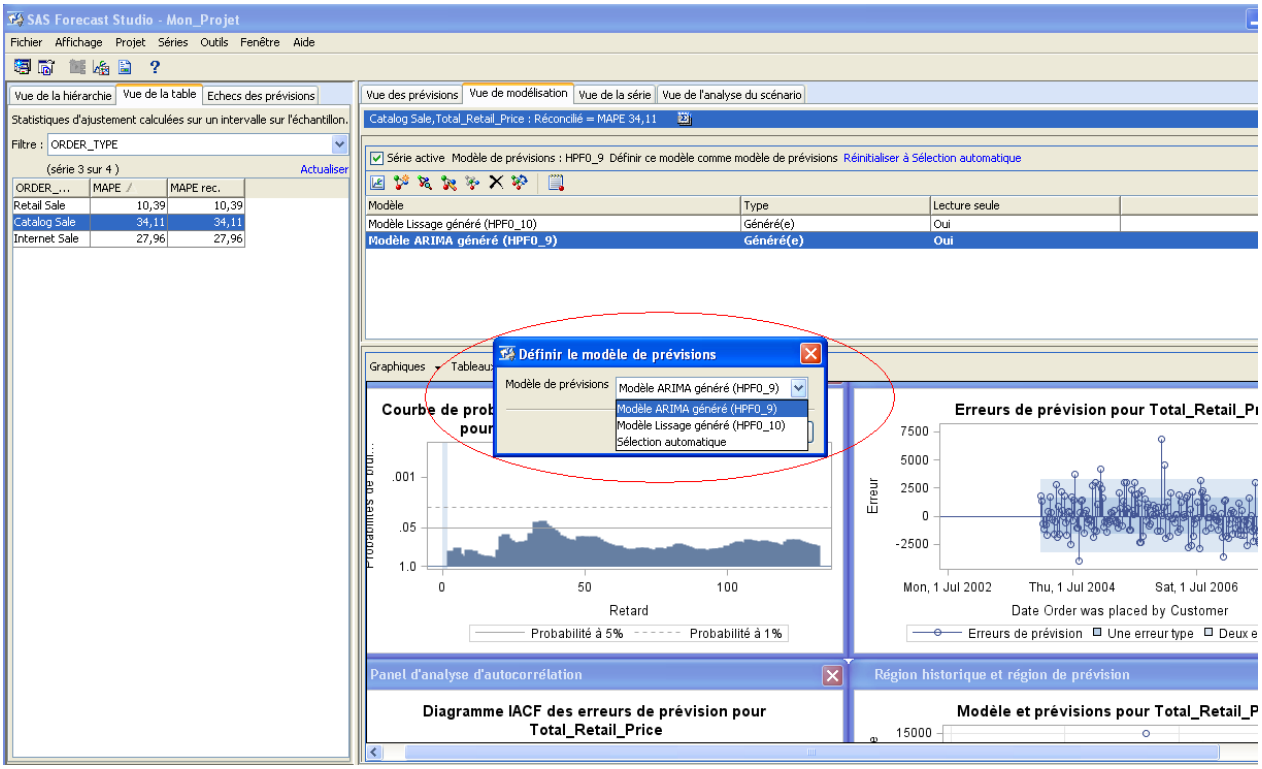


Dans la capture d'écran ci-dessus, on a sélectionné la table générale sans la distinction en fonction des canaux de vente. Les valeurs de cette série est la réconciliation des trois séries par canal. On voit également que les graphiques affichés, par défaut, sont assez complets mais que d'autres graphiques sont disponibles dans « Graphiques » et « Tableaux ».

Le dernier onglet, « Vue de l'analyse du scénario », n'est pas disponible dans notre cas puisque les scénarii analysent le comportement de notre variable dépendante en fonction des événements passés ou futurs de variables explicatives du modèle, qui sont absentes de notre modèle. Ce type d'analyse est intéressant et donne des résultats époustouflants que vous découvrirez facilement avec le support de cours « Forecasting Using SAS® Forecast Server Software ».



Il est possible de changer de modèle sélectionné pour le modèle de son choix. En effet, vous constaterez que Forecast Studio sélectionne de manière automatique le modèle ayant la MAPE la plus petite, malheureusement celui-ci peut ne pas vous convenir (il peut ne pas « passer » les tests de bruits blancs, comme celui de Ljung-Box). Ainsi, si vous voulez sélectionner l'autre modèle, le logiciel n'estime que deux modèles par défaut (un lissage et un ARIMA), **allez** dans le menu « Séries », puis **cliquez** sur « Définir le modèle de prévisions… », il ne vous reste plus qu'à sélectionner le modèle que vous désirez.



Cependant, cette modification ne sera effective, comme le logiciel vous le précisera, tant que le modèle ne sera pas diagnostiqué ou ré estimé. Seconde chose importante à noter un bouton « Réconcilier » apparaitra en haut à droite et le logiciel insistera sur sa nécessité, donc **pensez à cliquer** sur ce bouton (surtout en cas de réconciliation ascendante et que vous modifiez une série de base).

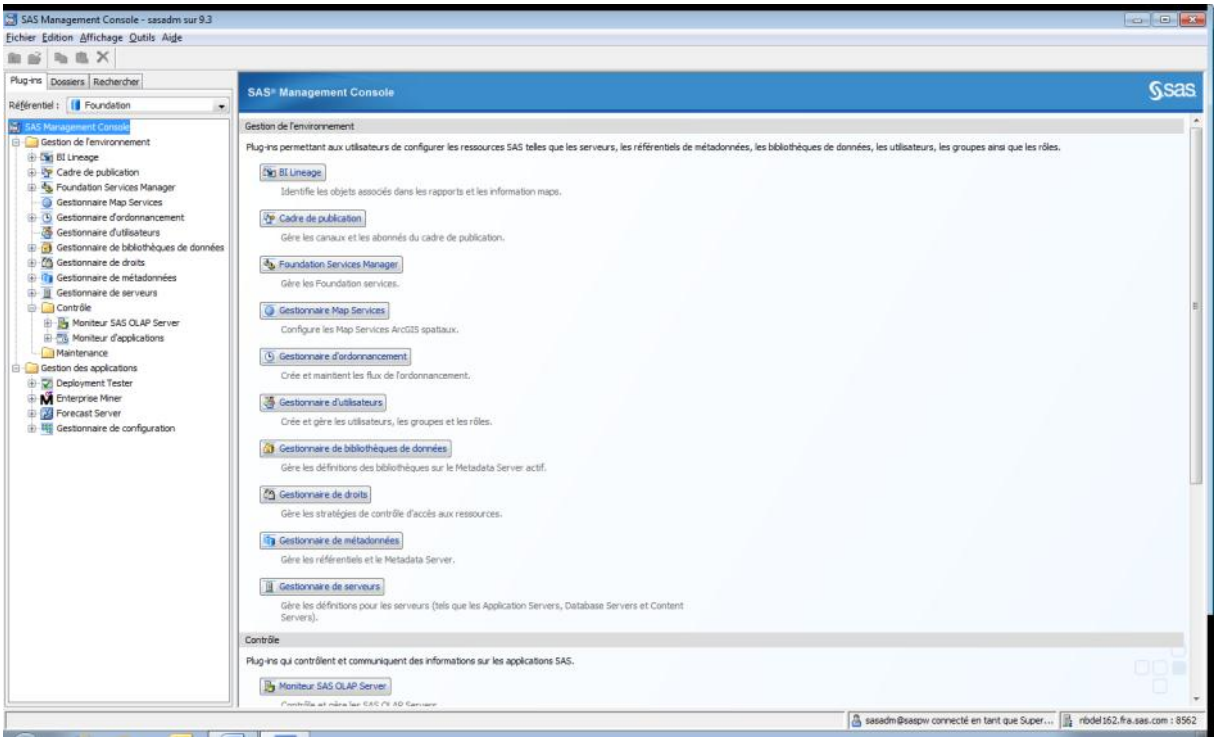
Administration

SAS *Management Console* (SMC)
La SAS *Management Console* est l'outil central de création et de gestion :

- des référentiels de métadonnées
- des serveurs
- des utilisateurs et de leurs droits
- des sécurités (authentifications, autorisations)

Mais elle permet également la gestion :

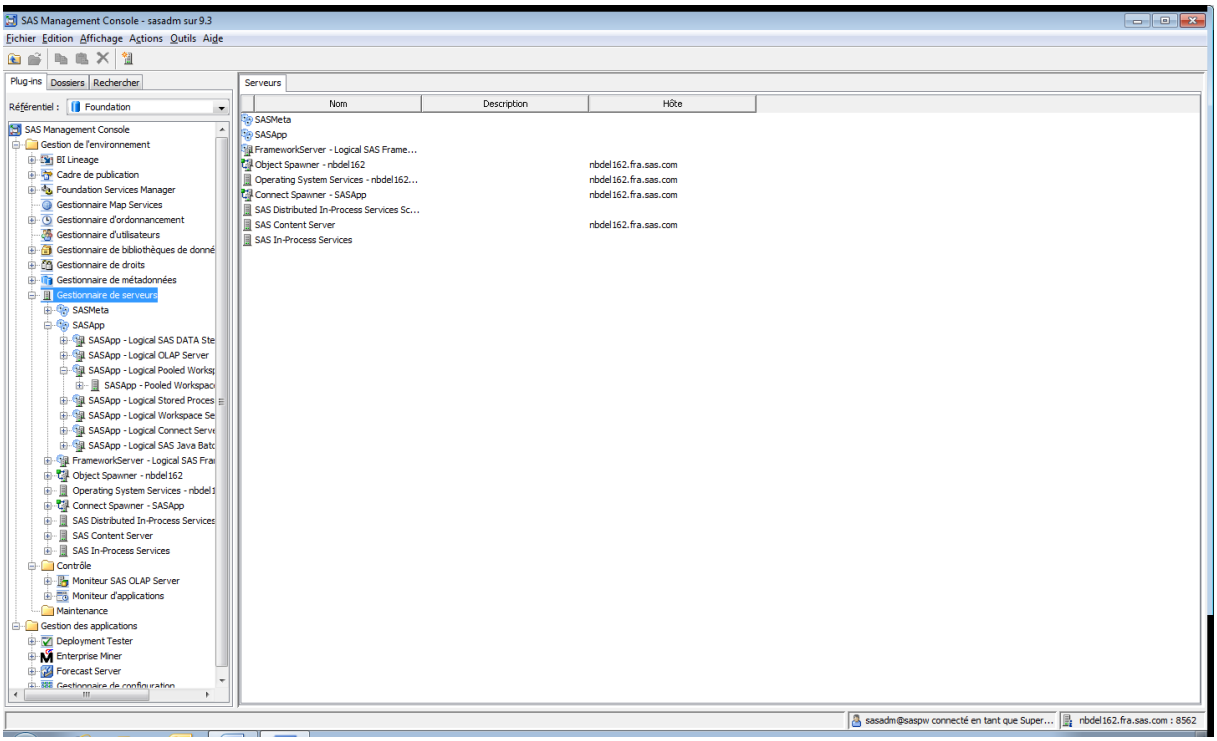
- des bibliothèques de données
- de la planification des traitements
- des licences SAS
- Etc.…
-



Présentation des *Application Servers*
Les « *Applications Servers* » représentent des points d'accès banalisés à une session SAS depuis un client externe.
Ils possèdent des caractéristiques propres qui en définissent la nature et le mode d'accès.

Les serveurs

Il existe différents types d'« *Application Servers* »
Tous ces serveurs sont gérés au niveau de la SMC.



Voici une présentation des principaux serveurs :

Le Metadata Server (ou serveur de métadonnées)

Toutes les informations relatives à notre environnement sont stockées dans des référentiels sous la forme de métadonnées.
Le serveur de métadonnées nous permet d'accéder à ces informations et de les exploiter dans toutes nos applications.
Il est par défaut sur le port 8561.

Le Workspace Server (ou serveur d'espace de travail)

C'est le serveur SAS chargé d'exécuter le code SAS (en dehors des processus stockés).
Par exemple, lors de la définition d'une structure multi-dimensionnelle, c'est lui qui est chargé de sa création.
C'est également lui qui accède aux bibliothèques et aux données.

Le Stored Process Server (ou serveur de procédures stockées)

C'est le serveur SAS chargé d'exécuter le code SAS relatif aux processus stockés.

L'OLAP Server (ou serveur OLAP)

C'est le serveur SAS qui permet d'accéder aux structures multidimensionnelles.
Il utilise le langage MDX.

Pour le serveur d'espace de travail et celui des procédures stockées, sous un serveur logique, peuvent être définis plusieurs serveurs physiques, ce qui permet la répartition de charge ou le *pooling.*

Gestion des référentiels

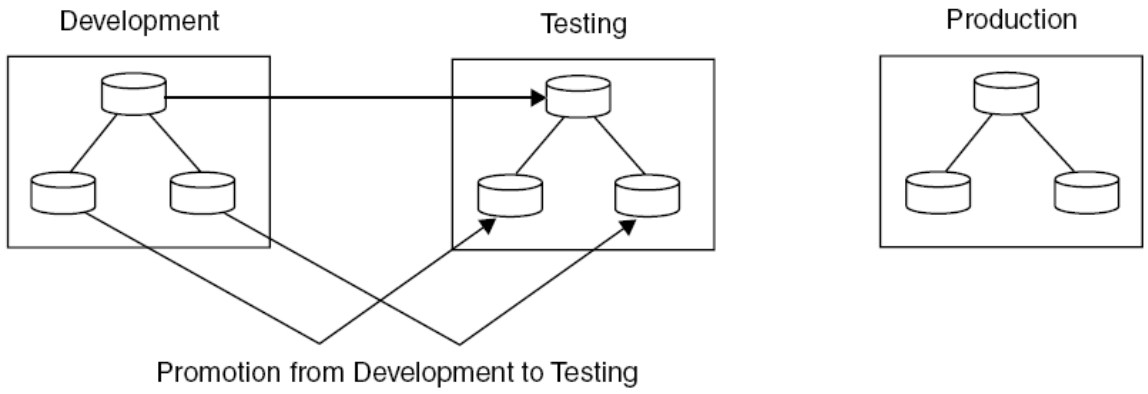
Toutes les informations sur l'environnement sont stockées de manière centralisée dans des référentiels de métadonnées
Intérêts de cette centralisation :

- point unique de stockage des informations
- environnement sécurisé
- organisation par métiers
- réutilisation des informations définies
- partage des informations
- contrôle des informations (check-in / check-out …)
- ouverture vers d'autres applications (CWM)

Ces référentiels, qui peuvent dépendre les uns des autres, permettent d'organiser les informations par métier, par projet …
Les référentiels peuvent également être associés à des processus de réplication ou de promotion

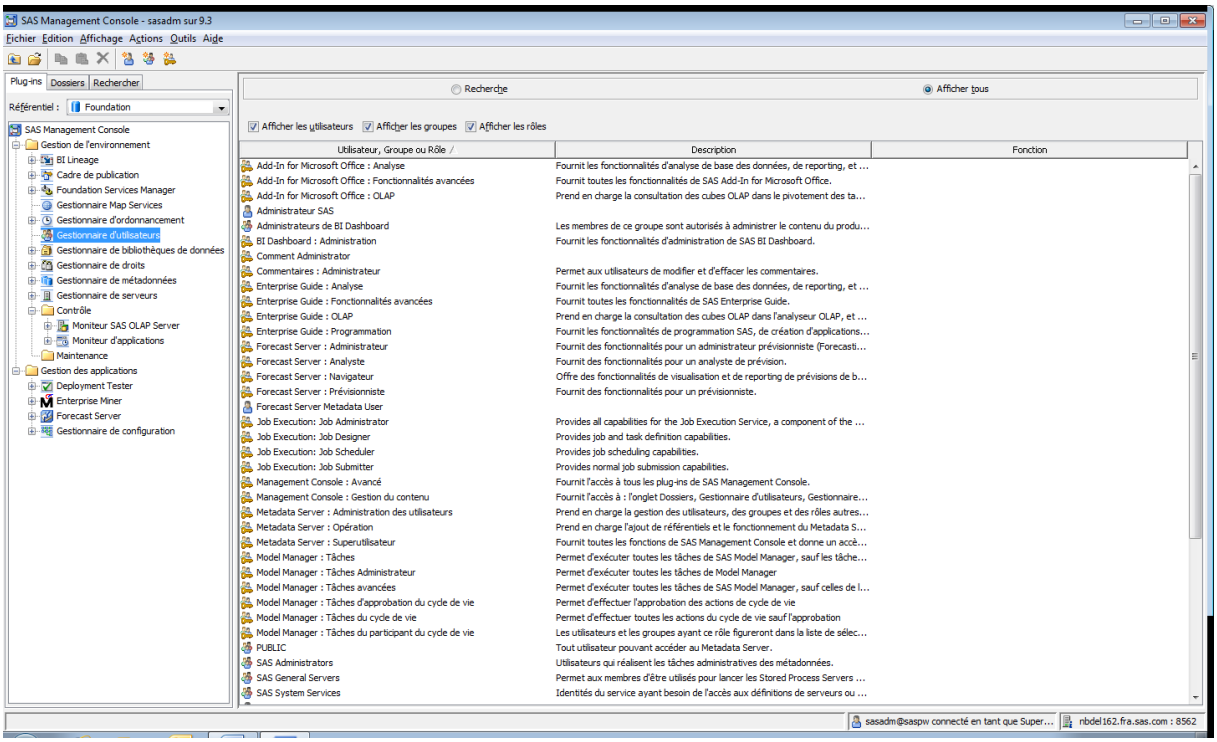
Promotion & Réplication

Passer un référentiel de développement à un de test ou de production



Gestion des utilisateurs

La SASMC permet de créer et de gérer les utilisateurs et les groupes d'utilisateurs.
On fait ici la distinction entre les utilisateurs logiques et les utilisateurs physiques.



Les utilisateurs et groupe de base :

Les différents utilisateurs de base :

- **SAS Administrator : sasadm**
  - Sert uniquement à administrer les métadonnées
  - Ne doit pas être utilisé depuis un client SAS autre que SMC
- **SAS Demo User**:
  - Utilisateur utilisé pour tester les différentes éléments

<u>Les différents groupes par défaut</u>

- **PUBLIC** :
  - Tous les utilisateurs identifiés possèdent le profil Public
- **SASUSER** :
  - Tous les utilisateurs disposant d'un profil dans les Metadonnées possèdent le profil SASUSER

- **<u>L'authentification primaire</u>**

L'authentification primaire est l'étape de validation de l'identité d'un utilisateur. Cette étape est assurée par le serveur de métadonnées et consiste uniquement à valider que le login et le mot de passe de l'utilisateur sont reconnus.
Pour cette étape, le serveur de métadonnées se repose sur les accréditations d'un tiers qui peut être :

- Le système d'exploitation
- Un annuaire d'entreprise (LDAP ou Active Directory)

A l'issue de l'authentification primaire, le Metadata Server va comparer le login utilisé pour la connexion aux différents logins du référentiel.

- Si le login est référencé, l'utilisateur est identifié et dispose des droits d'accès défini par son profil.

- Si le login n'est pas référencé, l'utilisateur n'est pas identifié et dispose des droits d'accès du groupe PUBLIC.
-

- **L'authentification complémentaire**

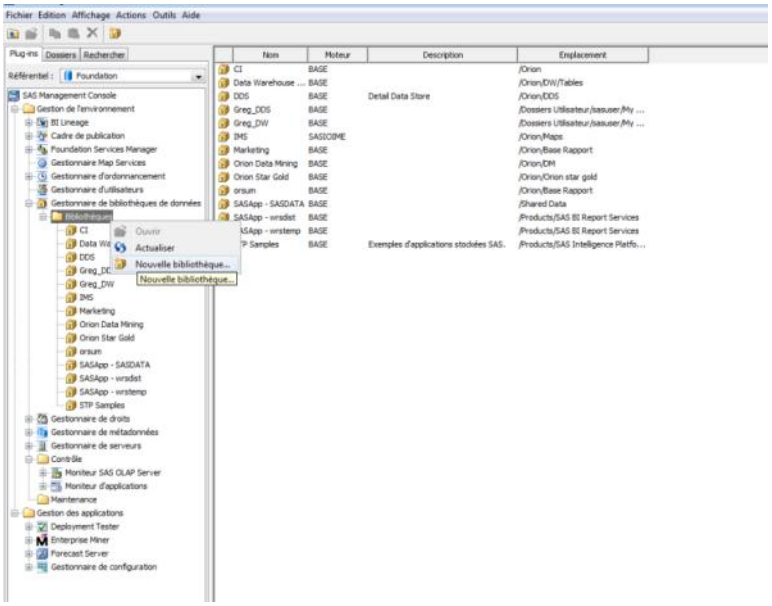
Une fois authentifiés sur le Metadata Server, les utilisateurs vont pouvoir accéder aux autres serveurs SAS. Pour cela, une deuxième étape d'authentification est nécessaire.
Par exemple, pour lancer une session Workspace Server, il faut indiquer à l'Object Spawner quel login système utiliser.

Pour déterminer le compte à utiliser, le serveur de métadonnées :

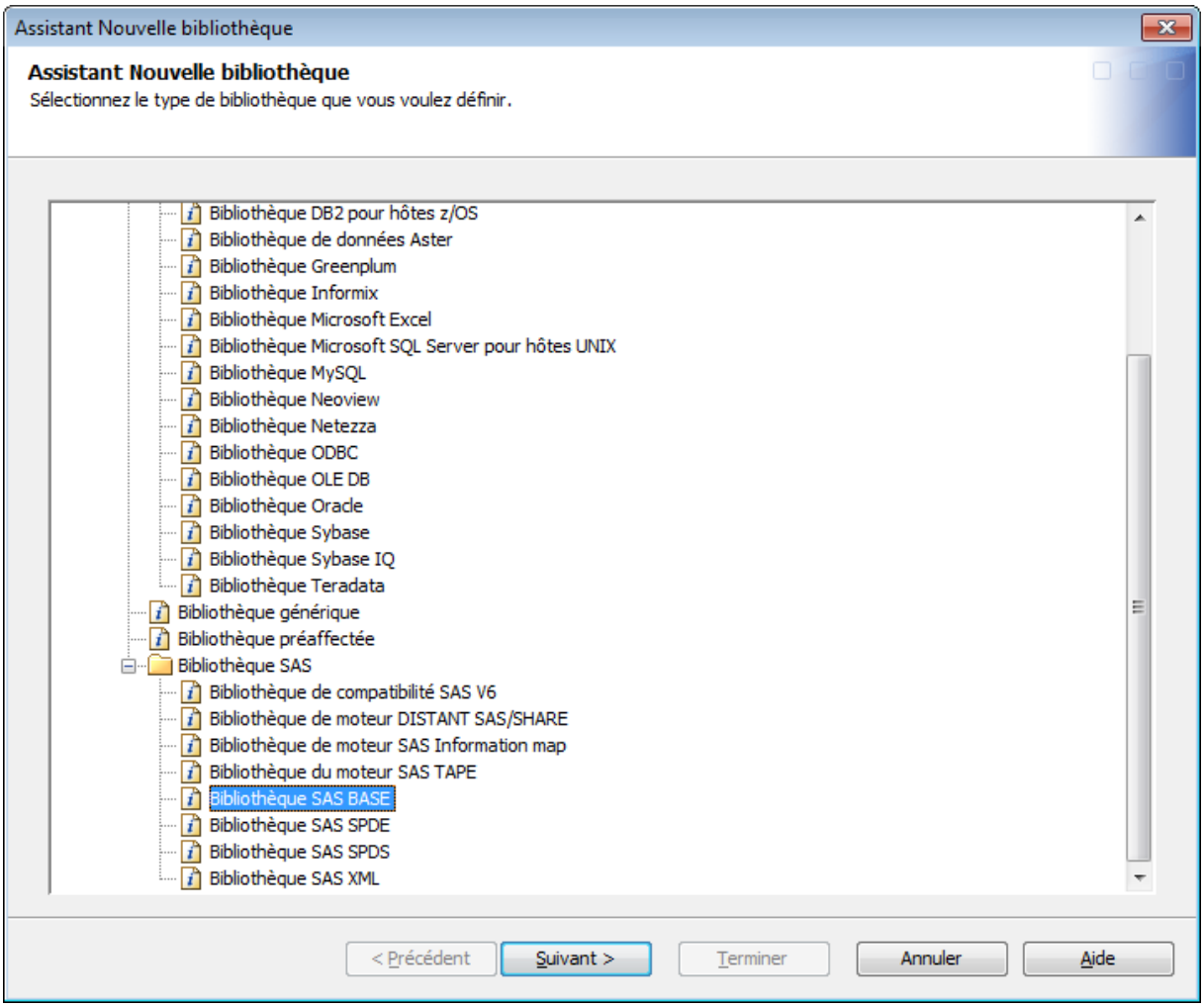
- Recherche le domaine d'authentification du serveur nécessitant l'authentification complémentaire
- Valide que l'utilisateur dispose bien d'un login pour accéder à ce domaine d'authentification
- Utilise le login et le mot de passe attribuée au profil de l'utilisateur pour ce domaine d'authentification

Création d'une bibliothèque :

Bibliothèque SAS Base



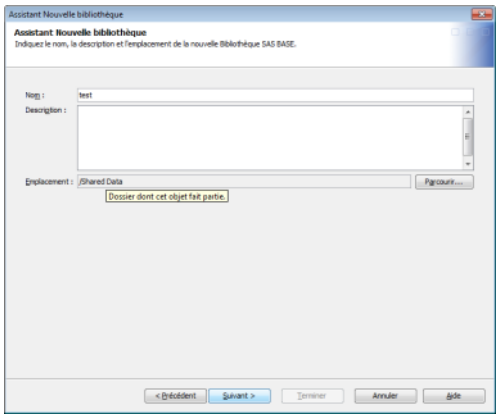
Dans la SAS Management Console (SMC) Dans le Gestionnaire de Bibliothèque (en anglais Library) → Bibliothèque SAS. Clic-droit de la souris, nouvelle bibliothèque.



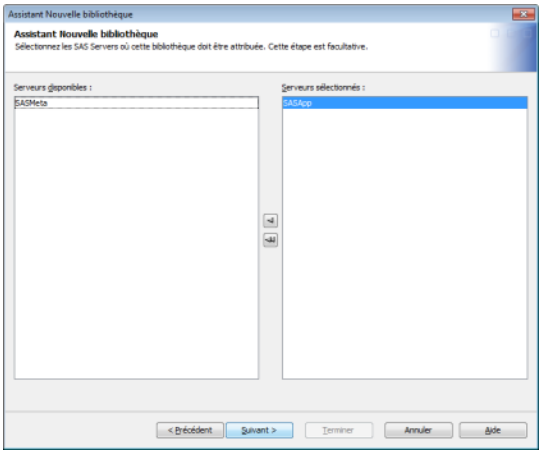
Parmi les bibliothèques SAS, sélectionner celle de type SAS Base. Cliquer sur suivant.

Remarque sur les différents types de bibliothèque SAS :

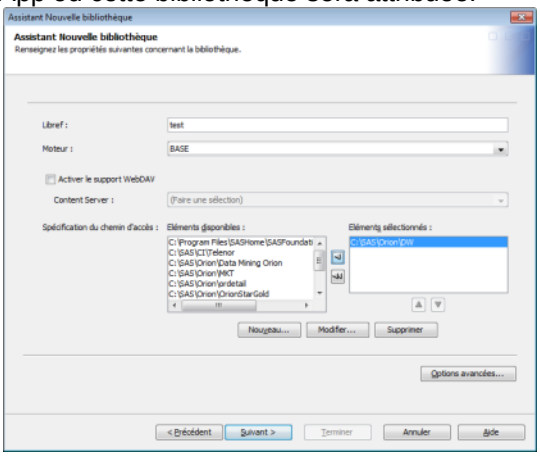
- Bibliothèque de compatibilité SAS V6, permet d'utiliser des tables SAS d'extension .sd2
- Bibliothèque de moteur de données à performance évolutive. SPDE est un module intégré maintenant au socle de base de SAS, Scalable performance Data Engine. Les bibliothèques SPDE permettent l'enregistrement des tables est des indexes associés sur de multiples partitions, afin de paralléliser correctement les différentes tâches.
    - Partitionnement hash et valeur : données, index
    - Index décisionnels : B-tree et BitMap
    - Parallélisation implicite : lecture et écriture
    - Archivage automatique des historiques
    - Backup/restore incrémental
    - Administration centralisée
- Bibliothèque de moteur SAS Scalable performance Data Server. SPDS est un module complémentaire de SAS/BASE. Les bibliothèques SPDS permettent l'enregistrement des table et des indexes associés sur de multiples partitions, afin de paralléliser au maximum les différentes tâches. SPDS devient nécessaire pour les très grosses volumétries.



Donner un nom explicite à la bibliothèque SAS.
Cliquer sur suivant.



Sélectionner le serveur SASApp où cette bibliothèque sera attribuée.



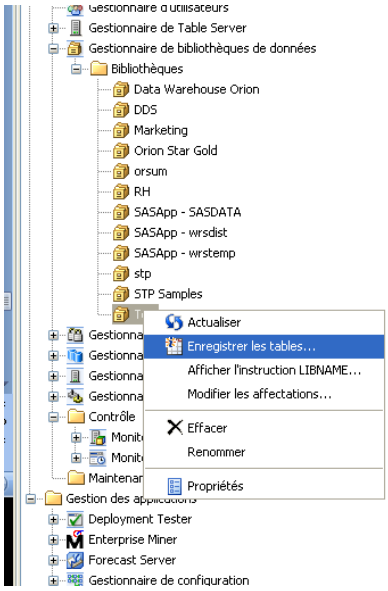
Donner un libref à la bibliothèque. Un libref ici est un nom sans espace, sans caractère spéciaux, de 8 caractères maximum.
Cliquer sur nouveau pour créer un nouveau chemin d'accès,

Entrer le chemin du répertoire auquel vous souhaiter affecter la librairie.

Lorsque le libref et le chemin sont définis, cliquer sur suivant.

Vérifier les informations, cliquer sur terminer.

Si vous sélectionnez la bibliothèque, vous ne verrez aucune métadonnée des tables qui se trouvent éventuellement dans le répertoire défini.
Créer une bibliothèque, c'est uniquement définir au niveau des métadonnées de SAS ces informations.
Pour importer les métadonnées des tables, vous pouvez le faire dans SAS® Data Integration Studio, ou directement dans la console d'administration :



Clic-droit, « enregistrer les tables » ou plus exactement, importer les métadonnées des tables.

Quelques modules SAS:

Une présentation commerciale succincte se trouve à la page :
http://www.sas.com/products/index.html
Si vous souhaitez des informations techniques, le site du support est :
http://support.sas.com
et plus spécialement la documentation en ligne qui est très riche :
http://support.sas.com/documentation/onlinedoc/

SAS/BASE : le module SAS/BASE est le socle de base nécessaire à l'utilisation de tout autre produit SAS. C'est tout d'abord un langage de quatrième génération (L4G) particulièrement performant pour la manipulation de données, la lecture et l'écriture dans des tables, SAS ou autre, l'édition d'état standard ou personnalisé et pour quelques analyses statistiques simples.

SAS/GRAPH : Le module SAS/GRAPH permet de générer et d'interagir avec tout type de graphique : histogramme, camembert, courbes multiples, cartographie, représentation 3D, image, vidéo, etc.

SAS/ACCESS : les modules SAS/ACCESS permettent de s'intégrer à différents systèmes de gestion de base de données (SGBD) du marché et à certains progiciels de gestion intégré (PGI ou ERP en anglais : Enterprise Resource Planning). Voici une liste d'exemples non exhaustive SAS/ACCESS Interface to SAP BW, to Teradata, to DB2, to MySQL, to ODBC,to OLE DB, to ORACLE, to PC Files, to PeopleSoft, Etc.

SAS/STAT: Le module SAS/STAT regroupe l'essentiel des fonctions d'analyses statistiques proposées par SAS.

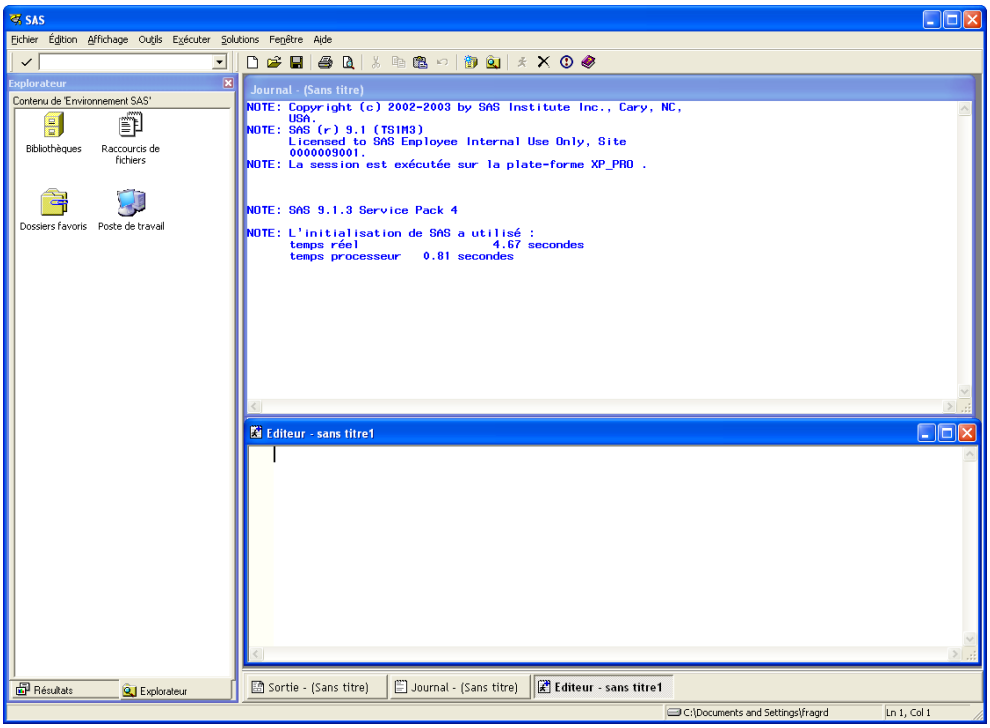
SAS/ETS (Econometric and Time Series): Le module SAS/ETS regroupe l'essentiel des fonctions d'économétrie, de séries chronologiques. (ARIMA, VARMAX, TIMESERIES…).

SAS/OR (Operational Research) : Le module SAS/OR regroupe des fonctions avancées de recherche opérationnelle comme la programmation linéaire, la programmation non-linéaire ou bien l'optimisation de graphe. L'un des problèmes le plus fréquent d'optimisation est de minimiser ou de maximiser une équation sous contraintes. Par exemple, maximiser la marge, minimiser les coûts ou minimiser le temps sous contraintes.
SAS/OR comprend aussi les algorithmes génétiques qui reproduisent le processus biologique de sélection et d'évolution et permettent d'identifier de « bonnes » solutions aux problèmes combinatoires dans des espaces larges, complexes ou peu structurés (objectif ou domaine non-continus).

SAS/IML (Interactive Matrix Language) : Le module SAS/IML est un langage de programmation matriciel.

SAS/CONNECT permet de connecter entre eux des ordinateurs sur lesquels est installé SAS. Vous partagerez au moyen de cette application vos tables de données, des programmes écrits sur un ordinateur pourront être exécutés sur un second.

Tous ces modules nécessitent de programmer en SAS. Ci-dessous l'interface SAS/BASE Windows, pour les programmeurs SAS.

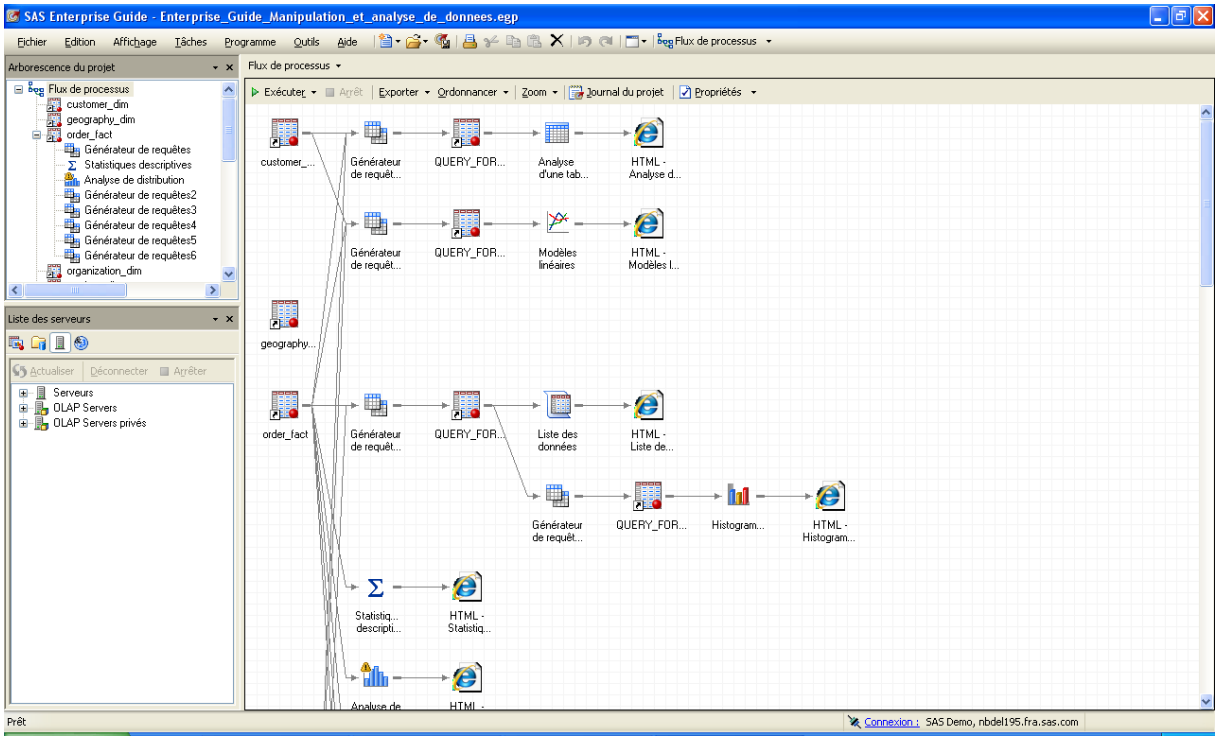

SAS
Fichier Édition Affichage Outils Exécuter Solutions Fenêtre Aide
Explorateur
Contenu de 'Environnement SAS'
Bibliothèques
Raccourcis de fichiers
Dossiers favoris
Poste de travail
Résultats
Explorateur
Journal - (Sans titre)
NOTE: Copyright (c) 2002-2003 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) 9.1 (TS1M3)
Licensed to SAS Employee Internal Use Only, Site 0000009001.
NOTE: La session est exécutée sur la plate-forme XP_PRO .
NOTE: SAS 9.1.3 Service Pack 4
NOTE: L'initialisation de SAS a utilisé :
temps réel 4.67 secondes
temps processeur 0.81 secondes
Editeur - sans titre1
Sortie - (Sans titre)
Journal - (Sans titre)
Editeur - sans titre1
C:\Documents and Settings\fragrd
Ln 1, Col 1

Quelques packages :

Les packages ci-dessous sont présentés à titre d'information, la liste des modules intégrés n'est pas toujours présentée et peut être soumise à des changements.

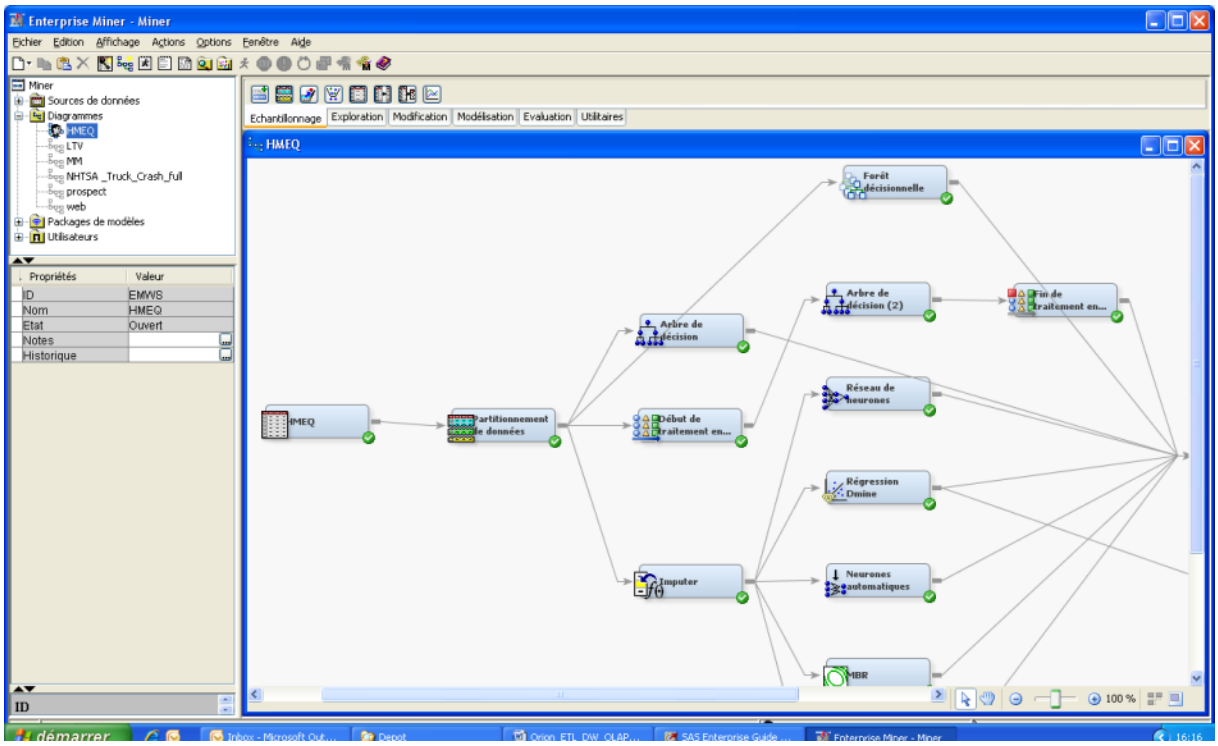
SAS Enterprise Guide

SAS Enterprise Guide est une interface permettant de manipuler des données, de faire des analyses statistiques et des rapports. Cette application développée en « .NET » est donc une application uniquement Windows. Enterprise Guide génère du code SAS, qui est exécuté par un moteur SAS base. SAS Enterprise Guide ne peut être utilisé sans une connexion à un socle SAS/BASE. SAS Enterprise Guide génère du code SAS qui peut être exécuté soit par la plate-forme SAS, soit directement par SAS/BASE.



SAS Enterprise Miner

SAS Enterprise Miner est le module de Data Mining de SAS. C'est une « usine » à modèle permettant l'industrialisation de l'exploitation intensive de gigantesques bases de données.



Dans la barre des tâches (en haut) se trouve les outils de SAS Enterprise Miner ordonnés selon la méthodologie SEMMA : (Sample – Explore – Modify – Model - Assess)
On peut noter :

Sample : préparer les échantillons de Data Mining

Input Data Source : On l'utilise pour définir la table sur laquelle on va travailler, le rôle des variables et leurs mesures.

Data Partition : On l'utilise pour partitionner la table d'entrée en deux ou trois sous-ensembles, dont une partie des données pour construire le modèle, et une pour le valider.

Explore : avant de faire des modèles, explorer les données et les analyser

Multiplot est un outil rapide pour visualiser les distributions, et donc détecter les valeurs manquantes, les valeurs aberrantes, etc.

Insight : outil d'analyse descriptive des données, analyses interactives simples et multivariées.

Modify : suite à l'exploration des données, modifier les données.

Replacement permet le remplacement des valeurs manquantes par différentes méthodes.

Transforme Variable permet de transformer les variables

Model : création de plusieurs modèles de différentes familles

Tree : pour la création d'arbres de décision

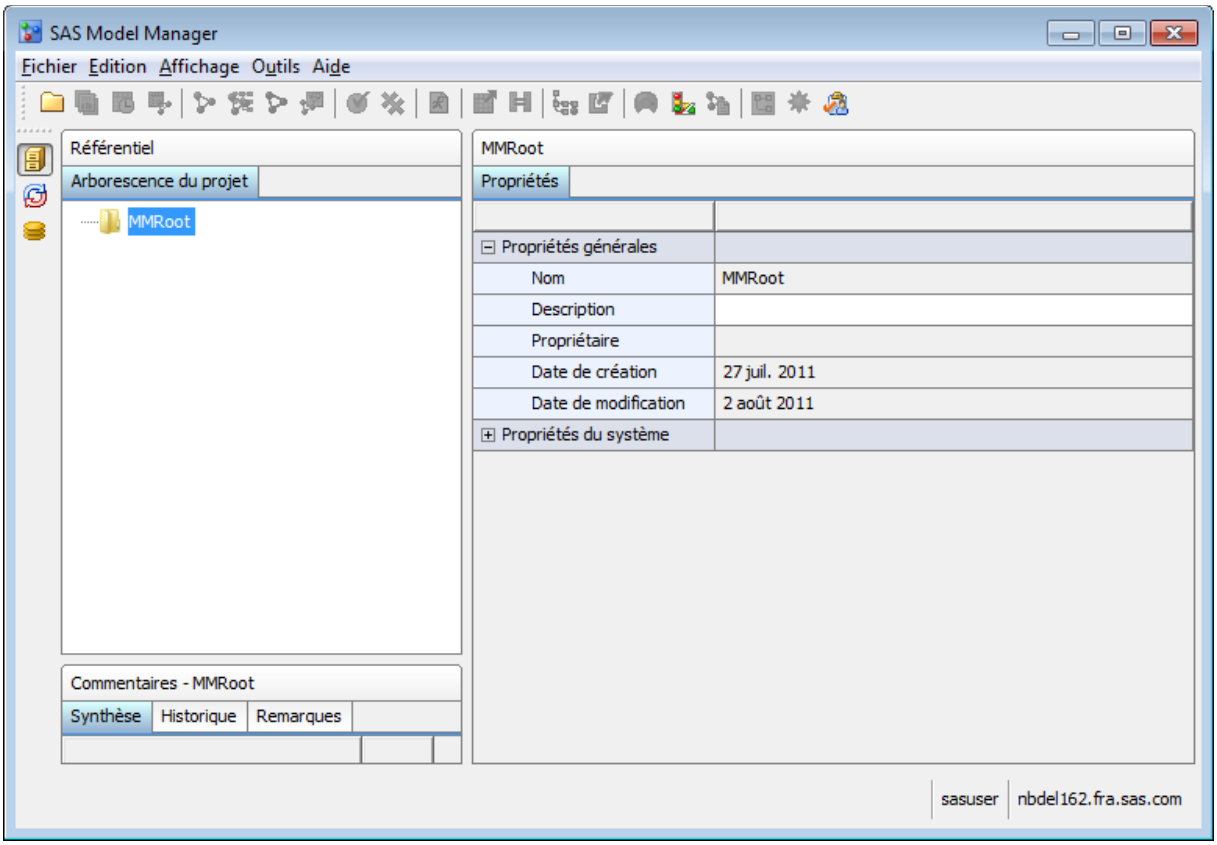
Regression : pour la création de régressions

Neural Network : pour la création de réseau de neurones

Assess : comparaison des modèles pour sélectionner le plus pertinent.

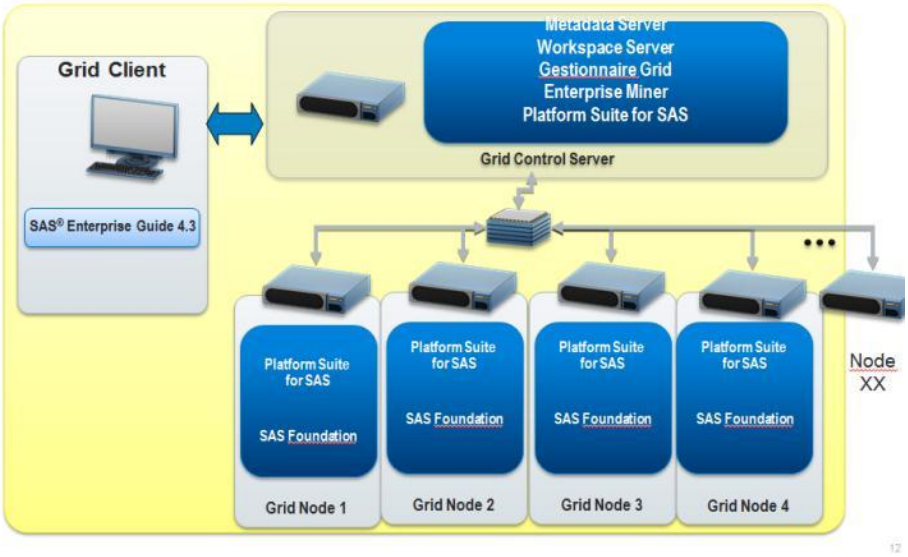
Assessment : permet de comparer les modèles.

SAS Model Manager

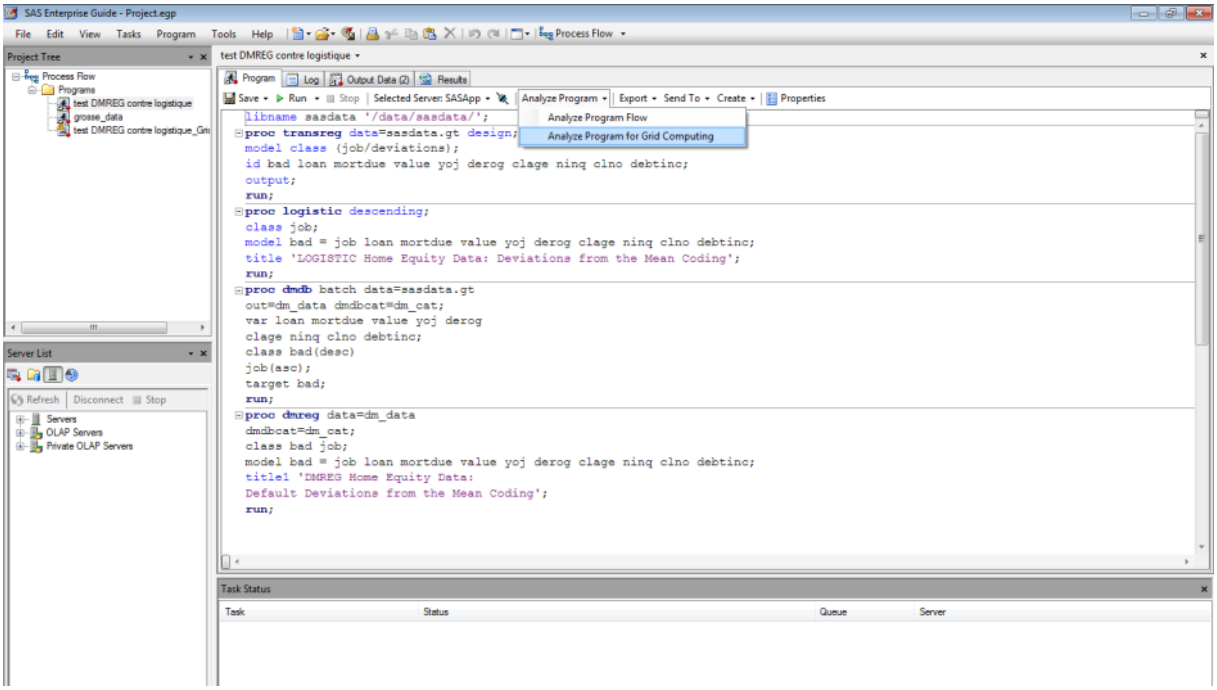


SAS Model Manager permet de suivre les modèles de Data Mining tout au long de leur vie, comparer leur performance, leur évolution au cours du temps.
Par exemple, Enterprise Miner permet de générer des modèles et Model Manageur permet suivre la performance de ces modèles lorsqu'ils sont appliqués sur les vrais cas. On peut ainsi comparer la performance estimée d'un modèle a priori, à sa performance lorsqu'il est appliqué dans la « vraie vie ». De plus, on peut suivre la performance d'un modèle au cours du temps et détecter ainsi plus rapidement lorsqu'il commence à moins bien fonctionner. Le Data Mining devra alors chercher à l'améliorer de nouveau.

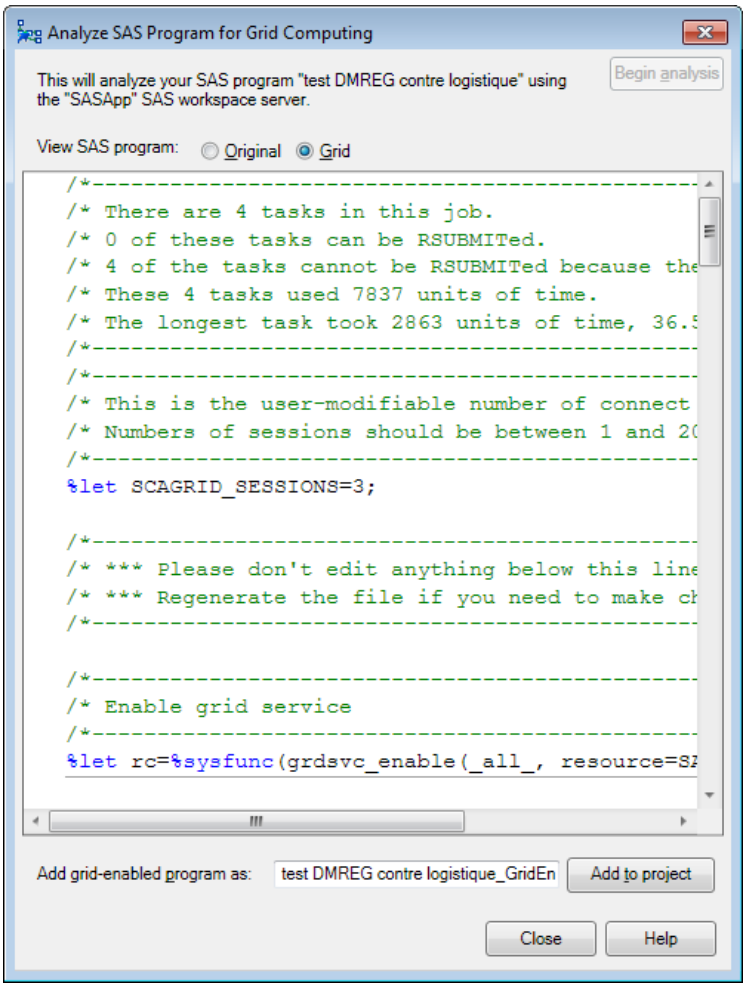
SAS Grid Computing



L'architecture en grille de calcul est l'une des technologies importantes qui permet de manipuler et d'analyser des quantités astronomiques d'information et parallélisant les calculs.

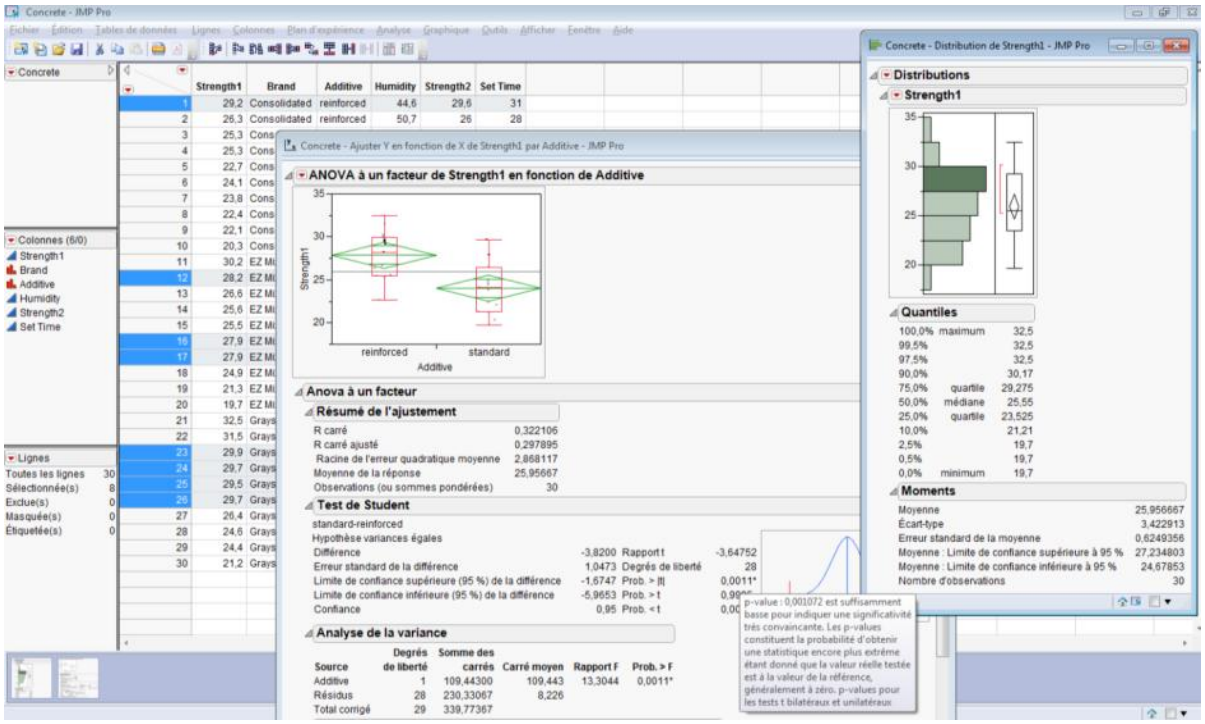


Par exemple, avec Enterprise Guide, on prend un code SAS et un bouton permet de faire l'analyse de ce code pour le rendre parallélisable sur une grille


Analyze SAS Program for Grid Computing
This will analyze your SAS program "test DMREG contre logistique" using the "SASApp" SAS workspace server.
Begin analysis
View SAS program:
Original
Grid
/* There are 4 tasks in this job.
/* 0 of these tasks can be RSUBMITed.
/* 4 of the tasks cannot be RSUBMITed because the
/* These 4 tasks used 7837 units of time.
/* The longest task took 2863 units of time, 36.
/* This is the user-modifiable number of connect
/* Numbers of sessions should be between 1 and 2
%let SCAGRID_SESSIONS=3;
/* *** Please don't edit anything below this line
/* *** Regenerate the file if you need to make ch
/* Enable grid service
%let rc=%sysfunc(grdsvc_enable(_all_, resource=SA
Add grid-enabled program as:
test DMREG contre logistique_GridEn
Add to project
Close
Help

JMP

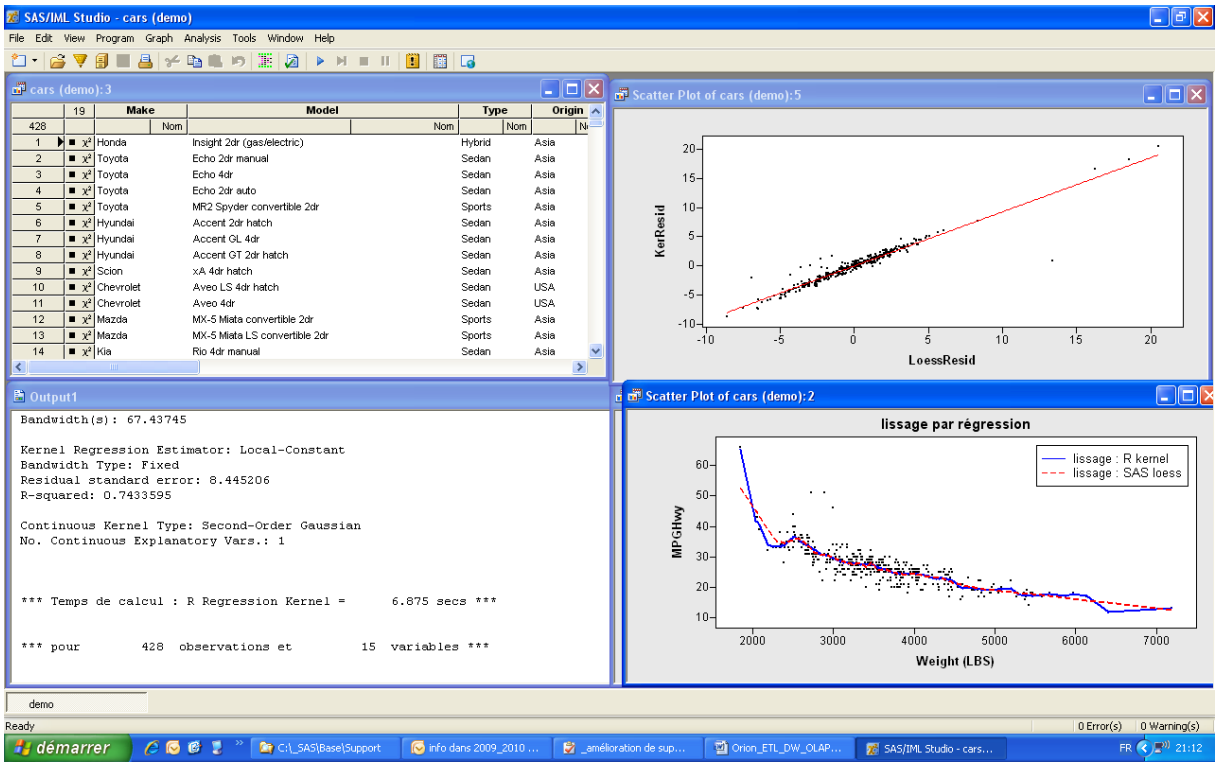
JMP est un outil d'analyses et d'explorations interactives des données pour ordinateur personnel Microsoft ou Mac.
Revenons sur un peu sur l'histoire de SAS. Il était une fois, à la fin des années soixante, quatre chercheurs à l'université de Caroline du Nord qui avaient développé un programme informatique de manipulation et d'analyse statistique. En 71, ils ont vendu leur première licence à une compagnie privée et devant leur succès, ont dû en 76 créer la société SAS pour commercialiser leur programme. Après s'être redistribuer leurs parts, SAS appartient maintenant à Jim Goodnight et John Sall.
A la fin des années 80, John Sall a commencé de développer un autre programme pour Macintosh : le John Macintosh Programme : JMP. JMP est donc un logiciel développé depuis plus de vingt ans par une équipe de SAS.



Vous pouvez télécharger une version d'évaluation depuis le site http://www.jmp.com.

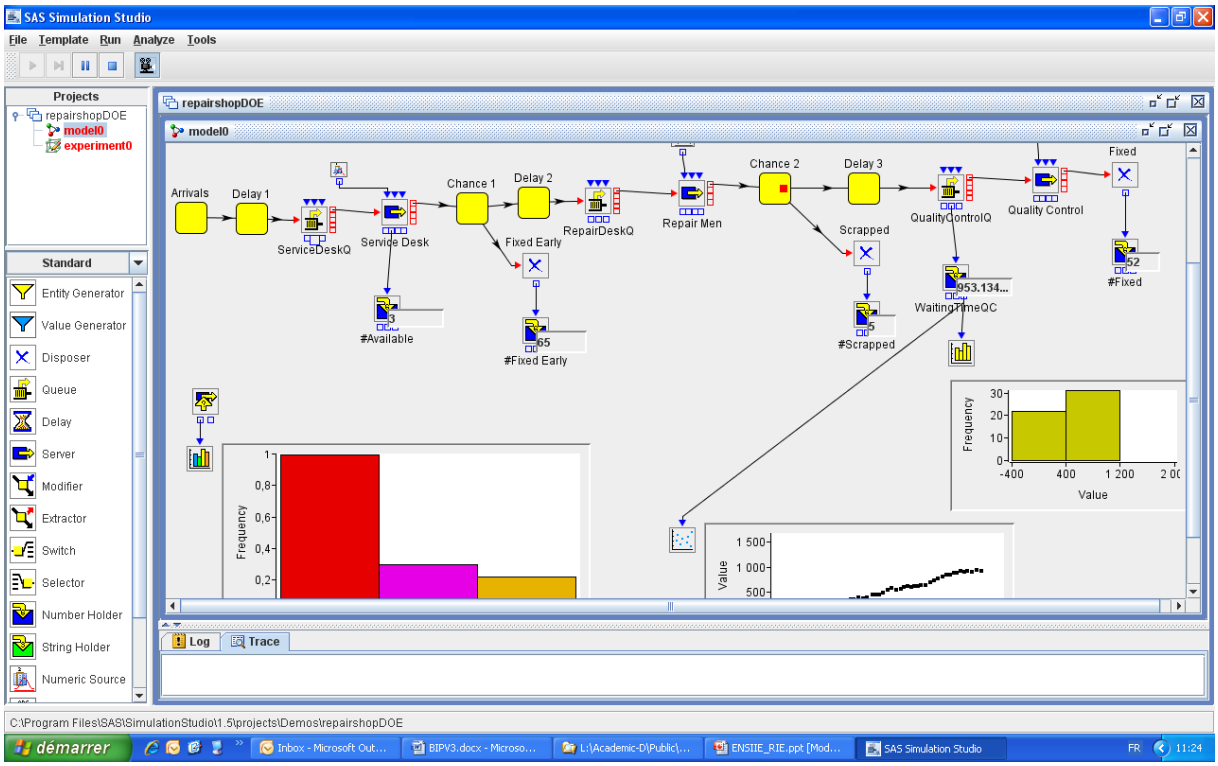
SAS IML Studio

Interface sortie avec SAS 9.2, permet notamment l'exploration interactive des données. SAS IML Studio est l'évolution de SAS/Insight et d'IML workshop.
Le langage IML (proc IML) est un langage de programmation matriciel.
Le langage IML + que l'on peut coder dans IML Studio, est un langage objet ce qui permet d'encapsuler du code IML, du code SAS, ou du code C, Java ou R (logiciel libre de statistiques).



SAS Simulation Studio

SAS Simulation Studio est une interface pour la modélisation et l'analyse de files d'attente, l'évaluation de scénarios alternatifs, analyse des performances.



AppDev Studio

SAS AppDev Studio est l'outil de développement basé sur Eclipse permettant notamment de :

- Créer des interfaces clientes
- Créer des servlets, des applets, pour construire des applications clientes sur les technologies Web
- Créer des portlets pour compléter le Portail
- Créer des « plug-in » à ajouter aux clients Java
- Développer des applications analytiques et de pilotage en client léger, Java ou Windows
- Exploiter au mieux la puissance du serveur SAS dans des applications de pilotage sur mesure
- Un environnement de développement complet et autonome, pour concevoir tous types d'applications

SAS® Data Integration Server

SAS® Data Integration Studio est une interface cliente Java de la plate-forme décisionnelle SAS. Elle fait partie des packages « SAS Data Integration Server » et « SAS Enterprise Data Integration Server ».

SAS Data Integration Server comprend notamment SAS/Base, SAS/CONNECT, ETL Studio et LSF.

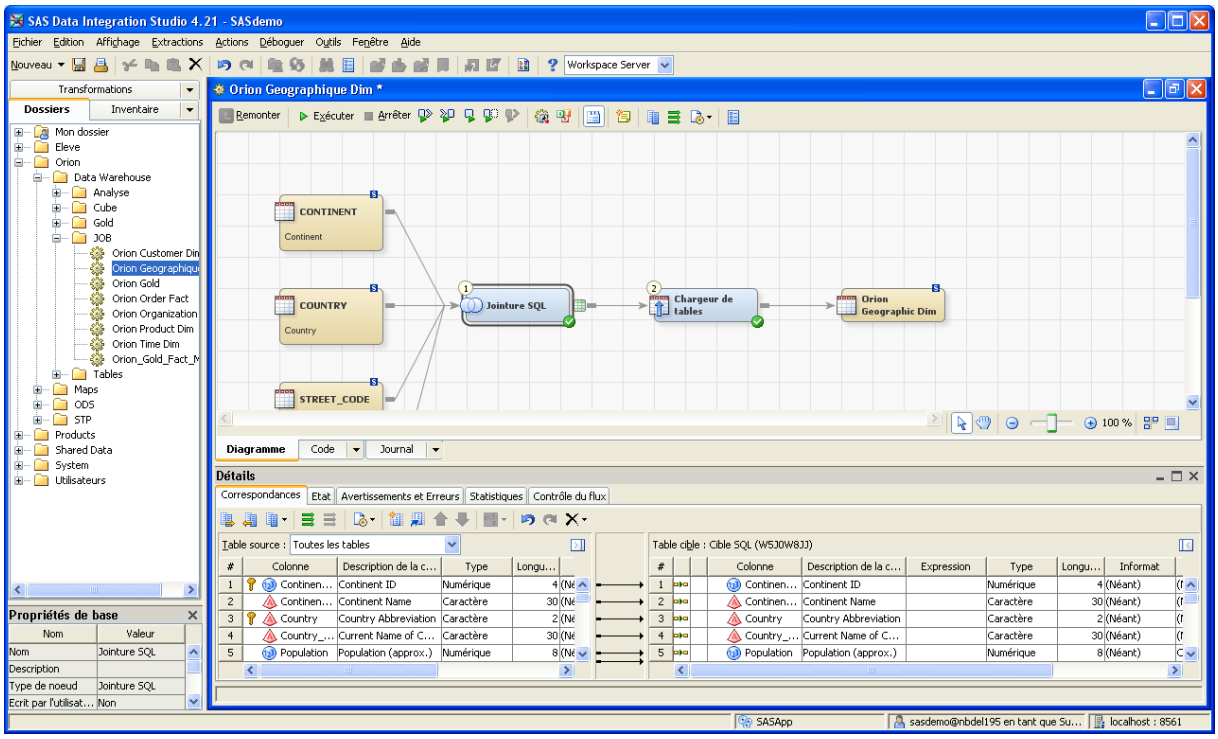
SAS Enterprise Data Integration Server comprend notamment SAS/Base, SAS/CONNECT, SAS/SHARE, 2 SAS/Access au choix (), ETL Studio, SAS Data Quality Server et LSF.

LSF est un Ordonnanceur permettant comme son nom l'indique d'ordonnancer des flux de processus.

ETL Studio permet de créer des flux de processus, qui une fois créés, vont être déployés pour l'ordonnancement.

SAS Data Quality Server est intégré à la plate-forme SAS suite au rachat de la société Data Flux en 2000. La qualité des données, c'est notamment :

- Audit de données
- Définition et suivi des indicateurs de qualité
- Définition des standards
- Correction des données
- Intégration : dé-doublonnage, consolidation, augmentation



SAS® Intelligence Storage Server

SAS 9.1 Base, SAS Integration Technologies, SAS Scalable Performance Data Server, SAS OLAP Server, SAS OLAP Cube Studio, SAS OLAP Server.
Stockage : SAS Scalable Performance Data Engine (SPDE – moteur de donnée à performance évolutive) est un SGBD interne à SAS et Scalable Performance Data Server SDPS et un SGBD SAS indépendant. Pour le stockage des informations dans le Data Warehouse, SAS propose de multiples technologies en fonction des problématiques. Toutes ces technologies sont des solutions décisionnelles et ne sont pas du tout destinées au stockage opérationnel. Les « rollback segment » et la gestion des contraintes d'intégrité ne sont pas des problèmes décisionnels. Tout comme ces bases définies et dédiées pour le décisionnel et qui sont inadaptées à l'opérationnel, les bases de données opérationnelles, sont généralement très mal adaptées au décisionnel, leurs performances mauvaises car elles gèrent des problèmes opérationnels, ce qui les surchargent. Pour les petits Data Warehouse, cela passe, mais pas pour les complexes. Parmi les technologies proposées par l'éditeur SAS, notons :

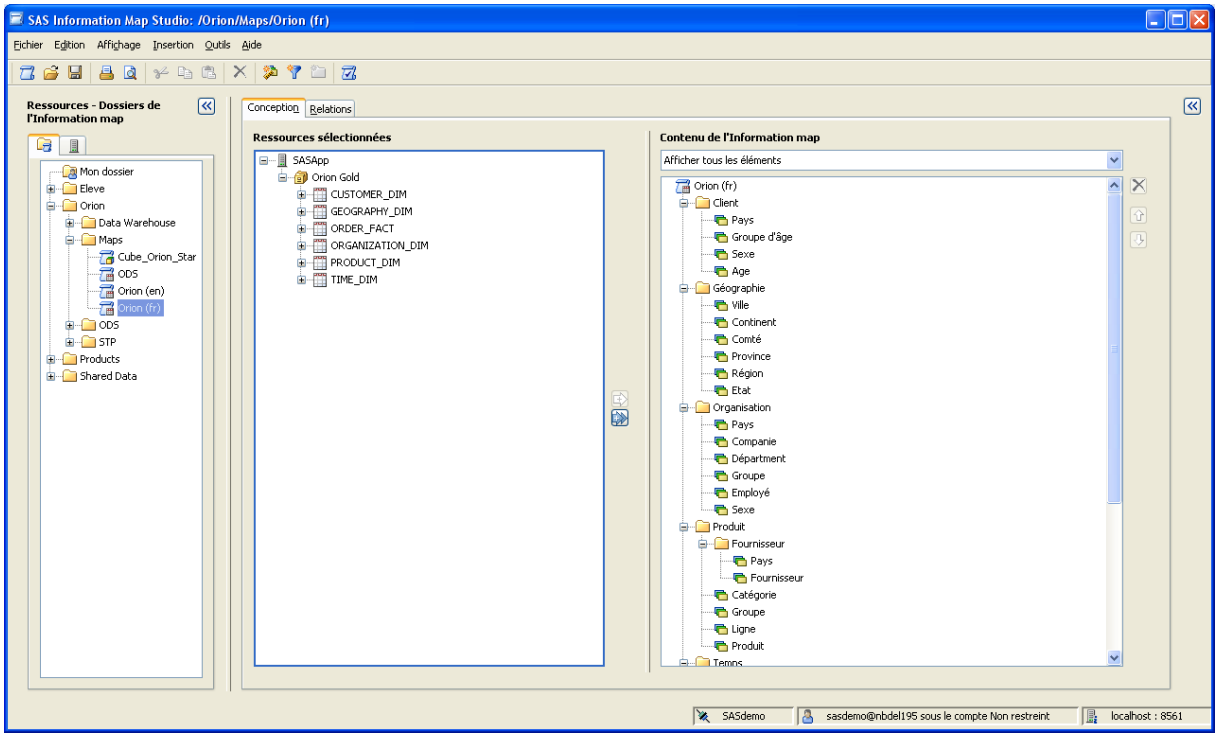
- les tables SAS Base sont destinées aux faibles volumétries
- les bases SPDE aux fortes volumétries
- les bases SPDS aux très fortes volumétries
- SAS OLAP server permet la création et la gestion des tables multidimensionnelles. Les hyper cubes sont en fait des tables SPDS et peuvent donc être parallélisées par le moteur SPDE ou SPDS.
- le module SAS/SHARE permet le partage en lecture et en écriture des tables
- les vues permettent de ne pas trop alourdir le Data Warehouse
- Sur la base d'Information Map, on peut créer des Data Mart virtuels (voir la partie stockage).

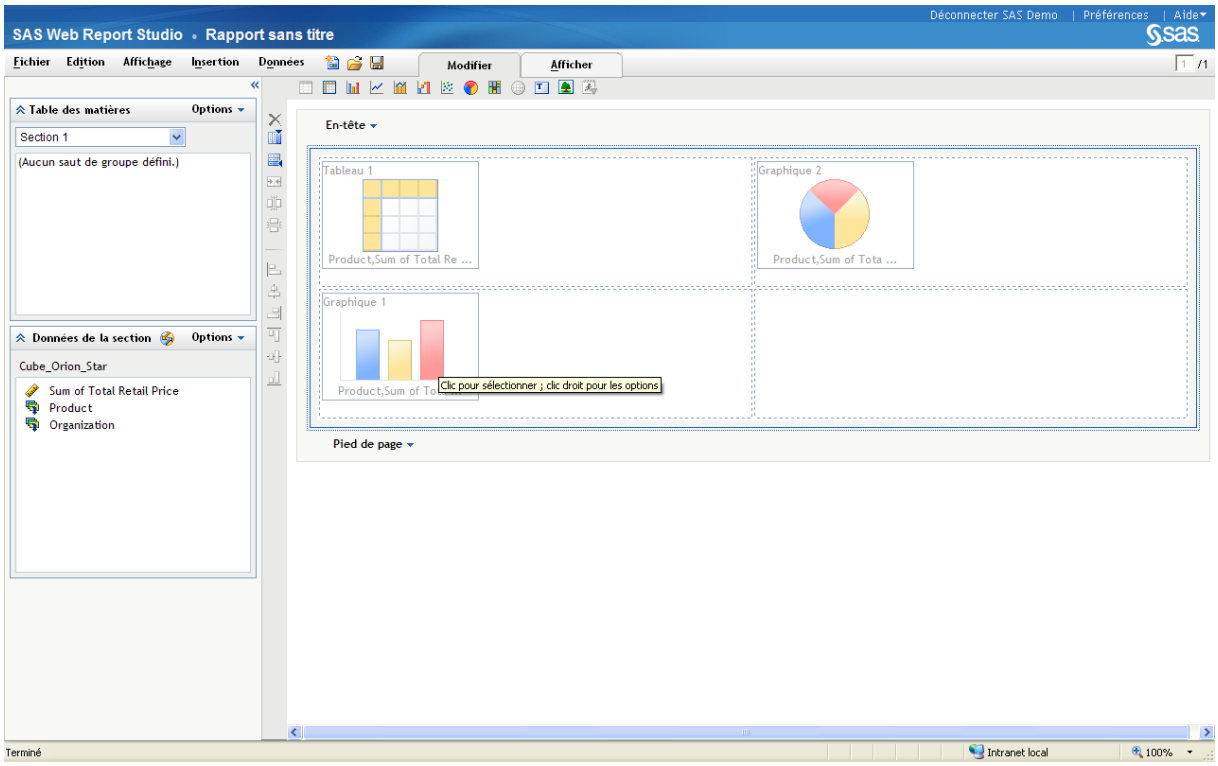
SAS® Business Intelligence Server

SAS 9.1 Base
SAS/Graph
SAS Integration Technologies
Enterprise Guide
SAS Office Integration Components
SAS Web Report Studio
SAS Information Map Studio

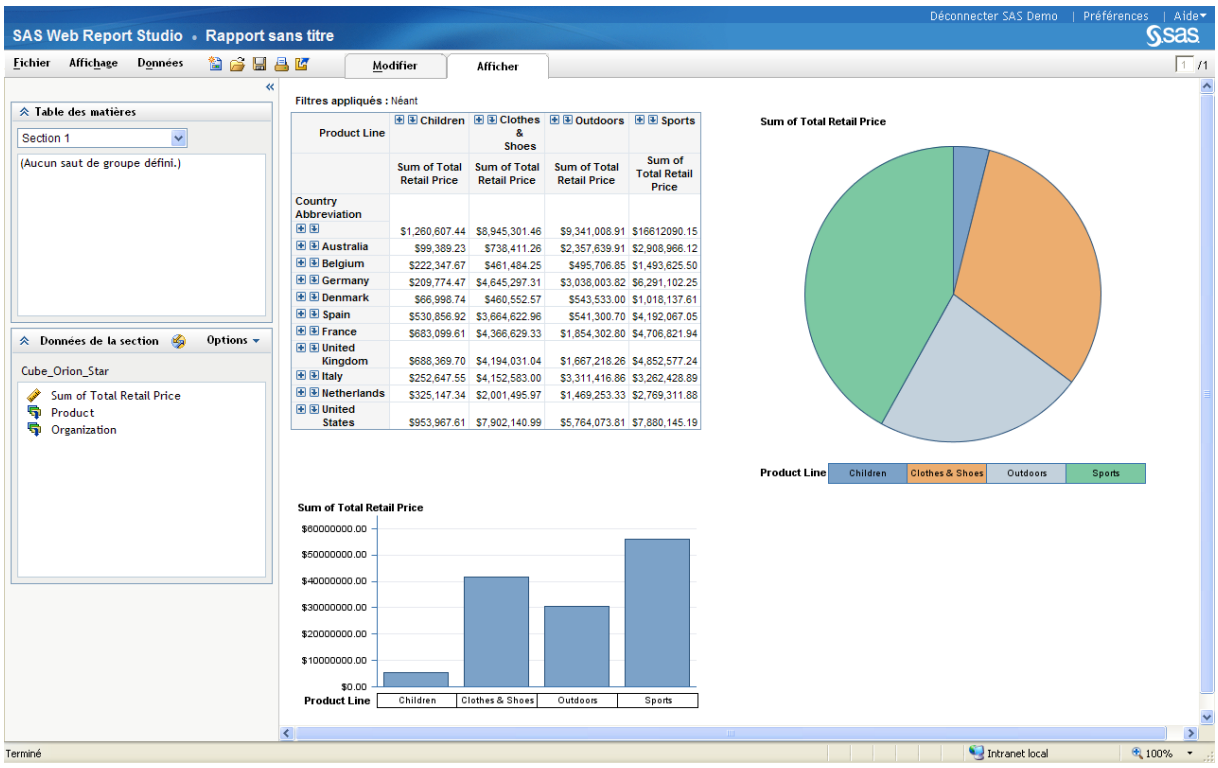
Reporting de masse : l'interface cliente Java SAS Information Map Studio, permet de créer des « information Map » que l'on peut comparer par bien des points aux « univers » de l'éditeur Business Object, créé avec l'interface BO – Designer. L'interface client léger de Reporting de masse SAS Web Report Studio permet à des utilisateurs non informaticiens de créer leurs rapports simplement. SAS Web Report Studio peut être comparé en plusieurs point avec l'outil BO.

SAS Information Map Studio permet de créer des Informations Map. Ci-dessus, dans la partie gauche, se trouve le référentiel de métadonnées SAS, au milieu, les données physiques, qui comme ci-dessous, peuvent avoir des relations complexes, ce qui est incompréhensible pour les non informaticiens, et dans la partie droite de l'interface ci-dessus se trouve l'Information Map, élément formaté pour un groupe d'utilisateur non informaticien. Cette Information Map permet de rendre disponible l'information de manière compréhensible pour chaque utilisateur. Voir le chapitre sur les Map.

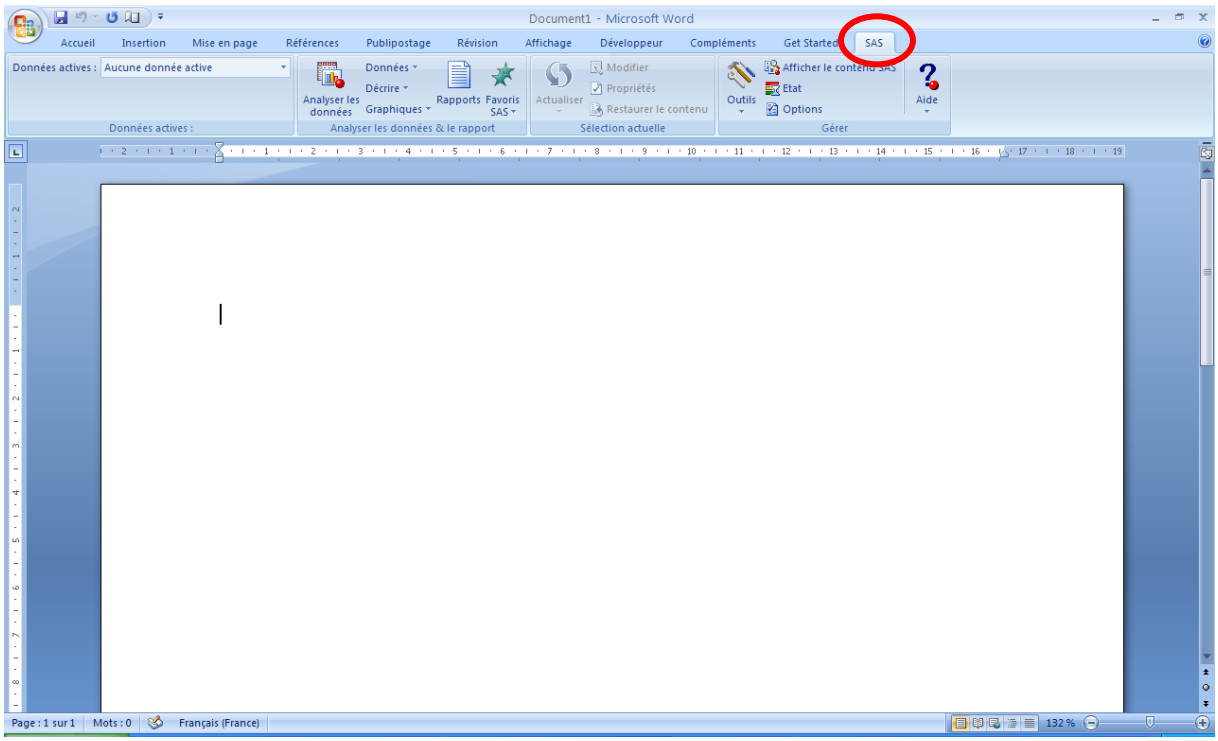




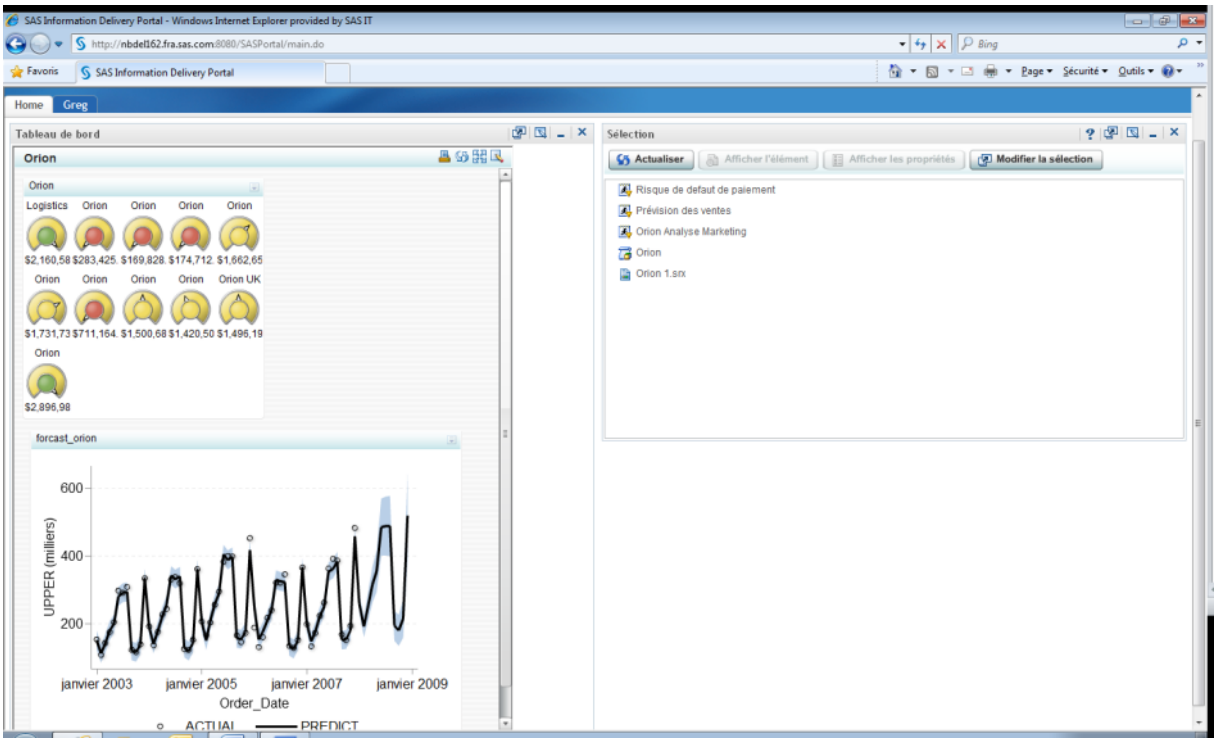
SAS Web Report Studio est une interface Web qui permet de créer des rapports. Ci-dessus, exemple de la fenêtre générique de création de rapport. Ci-après, exemple du rapport simple créé.



SAS Office Integration Components offre la possibilité d'ouvrir des tables SAS, d'exécuter des programmes SAS depuis la suite office.


Document1 - Microsoft Word
Accueil
Insertion
Mise en page
Références
Publipostage
Révision
Affichage
Développeur
Compléments
Get Started
SAS
Données actives :
Aucune donnée active
Analyser les données
Données
Décrire
Graphiques
Rapports
Favoris SAS
Actualiser
Modifier
Propriétés
Restaurer le contenu
Outils
Afficher le contenu SAS
Etat
Options
Aide
Données actives :
Analyser les données & le rapport
Sélection actuelle
Gérer
Page : 1 sur 1
Mots : 0
Français (France)
132 %

SAS Information Delivery Portal :



Le portail décisionnel SAS est le point d'entrée vers un ensemble d'applications Web décisionnelle.

Certification

Ces certifications ont une validité internationale. Trop d'étudiants mettent « SAS » sur leur CV, mais n'ont qu'une connaissance limitée de SAS. Ces certifications garantissent un véritable niveau. Il faut vraiment les préparer avant de tenter l'examen. Plus d'information à la page :
http://support.sas.com/certify/creds/index.html

Voici les certifications actuelles :

1. Certification Programmation de base
2. Certification Programmation avancée
3. Certification Plate-forme SAS® 9.2
4. Certification Predictive Modeling using SAS Enterprise Miner
5. Certification Business Intelligence SAS® 9.2
6. Certification Data Integration SAS® 9.2

Certification Programmation de base

La majeure partie des postes de consultant sur la plateforme décisionnelle SAS, proposés aux étudiants, nécessite de programmer en SAS. Le langage SAS est le socle de la plateforme SAS.
Attention : il faut connaitre parfaitement le langage SAS, c'est-à-dire qu'en regardant un programme, on sait sans l'exécuter exactement la sortie qu'il va produire ou pourquoi il ne fonctionne pas. Cela nécessite un travail conséquent.
Le livre SAS 9.2 de Sébastien Ringuedé, regroupe notamment tous les éléments pour préparer cette certification SAS.
La certification programmation de base est composées de 70 questions QCM (1 « chance » sur 4) en deux heures– Il faut 65% de réponses justes.
Après avoir passée la certification de programmation de base, vous pouvez passer celle de niveau avancé : Certification Programmation avancée

Certification Business Intelligence SAS®

Les éléments nécessaires mais non suffisant sans travail et expérience sur le sujet sont présentés lors des formations SAS® Business Intelligence - Les outils de restitution : niveau 1 (V92SBA1) et SAS® Business Intelligence - Les outils de restitution : niveau 2 (V92SBA2) (ces deux formations sont condensées dans le cours « Creating Business Intelligence for Your Organization: Fast Track » : SBAFT).
La certification est composées de 64 questions QCM en deux heures– Il faut 67% de réponses justes.

Voici des idées de questions :
Adobe flash plug in est-il nécessaire pour le BI Dashboard ? (Attention à la version du Dashboard)
Est-ce que le %stpbegin génère des macros de réserves et prépare les prompts de paramètre ?
Est-ce que l'on peut sauvegarder des rapports SAS Web Report Studio en html ou en pdf ?
Que sont que les BI Dashboard alert ? Personal Alert ? Dashboard Alert ? Data Alert ? Indicator Alert ?
Si on ajoute un filtre comme un pré-filtre général, est –ce que cela oblige tout le monde à l'utiliser ?
Est-ce que dans la configuration par défaut de l'add in à office demande de s'identifier ? Celle de la SAS Management Console ? Celle de SAS Information Map Studio ? Celle de SAS OLAP Cube Studio ?
Quelle est la différence entre SAS.IdentityGroups et SAS.IdentityGroupName ?
Un utilisateur exécute l'ordonnancement d'un rapport dans SAS Web Report Studio, un autre utilisateur, une heure plus tard, distribue ce rapport. Est-ce que l'action de l'un impact celle de l'autre?

Certification Data Integration SAS®

Les connaissances nécessaires mais non suffisant sans travail et expérience sur le sujet à la Certification Data Integration sont dispensées lors de la formation SAS® Data Integration Studio : Niveau 1 (V92DI1)
La certification est composées de 78 questions QCM en deux heures– Il faut 70% de réponses justes.

Certification Predictive Modeling

Les connaissances nécessaires mais non suffisant sans travail et expérience sur le sujet à la Certification Data Mining sont dispensées lors de la formation SAS® Enterprise Miner™ : applications des techniques de Data Mining (AAEM_6).

Glossaire

ABC/ABM (Activ*ity Based Costing/Activity Based Management)* : La méthodologie et les outils ABC/ABM, permettent de dégager la contribution des activités au résultat selon les axes d'analyse souhaités. SAS® ABM est un module d'analyse des performances de l'entreprise par activité, pour le contrôle de gestion. SAS® ABM est aussi une composante importante de SAS® for Sustainability Management.

ACP : Analyse en Composantes Principalees. Analyse statistique multivariée très utilisée par les services marketing notamment.

Agrégat : Valeur synthétique obtenue par la combinaison ou la somme d'informations de détail. Exemple : la somme du chiffre d'affaire pour une société, par année.

AI : *Analytic Intelligence*, maillon de l'*Intelligence Value Chaine*, qui va du contrôle qualité à l'analyse des séries chronologiques en passant par la recherche opérationnelle, les statistiques avancées et le *data mining* ; permet de transformer la donnée en connaissance.

AI (2) : Artificial intelligence : Intelligence artificiel. Tout outil de Data Mining se respectant, comportent des algorithmes d'intelligence artificiel.

AID : Nom d'un algorithme d'arbre de décision (voir Arbre de Décision)

Analyse prédictive : Utilisation des méthodes mathématiques, statistiques, économétriques ou de datamining pour concevoir des modèles et obtenir des prévisions. La plupart des algorithmes prédictifs fonctionnent en apprentissage supervisé.

ANOVA : *ANalyse On Variance, littéralement :* Analyse de la variance. Ensemble d'analyses statistiques basées sur les analyses de la variance comme le test en T (T test ou test de Student), les analyses de variance à une dimension, les analyses de variance non-paramétriques à une dimension, les modèles linéaires ou les modèles mixtes.

Applet : Application Java non exécutable, pouvant être interprétée par une machine virtuelle java. Lors de la navigation dans une page Web, une Applet peut être chargée sur l'ordinateur de l'utilisateur, et sera interprétée (par abus de langage, exécutée) par le JRE du navigateur Web.

Application analytique : Solutions logicielles clés en main bâties pour répondre au besoin d'aide à la décision d'un département donné de l'entreprise ou d'un secteur économique.

Apprentissage supervisé : Algorithme de Data Mining capable d'apprendre sur des données où l'on connaît la réponse afin d'être appliqué sur d'autre données où l'on ne connaît pas la réponse, pour la déduire.

Attrition (ou *churn en anglais*) : Le potentiel d'attrition d'un client, est sa capacité à partir à la concurrence. Dans beaucoup de société, des modèles de *Data Mining* permettent de calculer pour chaque client, sa probabilité de départ.

BASE/SAS : BASE/SAS est le socle de SAS, premier module développé à l'origine, à la fin des années 60 par PhD Jim Goodnight. Toute licence SAS comprend ce module. BASE/SAS comprend le langage de programmation de quatrième génération SAS.

Base de données : regroupement organisé de toutes les données recueillies, calculées ou extrapolées sur un client ou un prospect. Une information appropriée du client ou du prospect sur les données le concernant doit être réalisée conformément à la Loi sur l'Informatique et les Libertés (pour laquelle l'organe de référence est la CNIL). La gestion des bases de données peut obéir à différentes logiques selon les besoins des utilisateurs.

BI : *Business Intelligence* historiquement traduit en français par informatique décisionnelle.

Check-Out/Check-In : Dans un projet ETL, l'administrateur de la plate-forme décisionnelle créé un référentiel de métadonnées dit « projet » pour chaque développeur ETL. Ceux-ci peuvent alors extraire (Check-out) du référentiel principal une partie du projet, dans leur référentiel dédié. Les autres utilisateurs ne peuvent pas modifier cet élément. Une fois les modifications faites, l'utilisateur réintègre, (Check-in) le processus dans le référentiel principal, avec la version N+1. Le Check-Out/Check-In permet donc à plusieurs développeurs ETL de travailler sans conflit sur un même projet et la gestion des versions.

CHAID : Algorithme d'arbre de décision (voir Arbre de Décision)

CPU : Central Processing Unit : Unité centrale d'un ordinateur exécutant les instructions. On mesure souvent la puissance des machines notamment en fonction du nombre de CPU, cœur de processeur. Une machine de 4 processeurs dual-core (double cœur) est souvent considérée comme une machine de 8 CPU.

CRM : le Customer Relationship Management se traduit en français par GRC : Gestion de la relation Client.

CRM ANALYTIQUE : Consiste à analyser les données présentes dans le référentiel clients/prospects pour en dégager les informations caractéristiques de la population cible de l'entreprise.

CRM OPERATIONNEL : Ensemble de fonctions permettant d'assurer la gestion quotidienne de la relation avec le client à travers l'ensemble des points de contact.

CROSS SELLING / VENTES CROISÉES : Technique marketing de vente directe qui consiste à offrir un ou plusieurs produits complémentaires s'ajoutant au produit que l'acheteur a l'intention d'acquérir.

CWM : Common Warehouse Model, selon le site web www.cwm.org, CWM est une norme pour les outils décisionnels permettant de transférer des fichiers de métadonnées entre différentes applications décisionnelles de différents éditeurs de logiciel. Par exemple, intégrer une l'application de Reporting Business Object® à la plate-forme SAS®.

DDS : Detail Data Store.

DSS : Decision Support System : autre traduction d'informatique décisionnelle

DM : Data Mining : traduit en français par fouille de donnée,

Data Miner : Personne faisant du Data Mining

DM (2) : *Data Mart*

Data Mart : Entrepôt de donnée à but décisionnel spécialisé pour une activité définie, orientée métier. Le *Data Mart* est un sous-ensemble du référentiel *Data Warehouse*. On peut avoir un Data Warehouse monde, un Data Mart pour le Japon, un autre pour l'Europe, etc. On peut avoir un Data Warehouse pour toute l'entreprise, un Data Mart pour les ressources humaines, un autre pour le marketing, etc.

DSI : Directeur des Systèmes d'Information

DW : *Data Warehouse* : Entrepôt de donnée à but décisionnel global à l'entreprise.

EIS : *Executive Information System*, Ce terme du début des années 90 équivaut à systèmes de pilotage destinés aux décideurs. Avec les EIS est apparue la navigation multidimensionnelle dans les données.

ERP : Enterprise Ressource Planning, en français, PGI : Progiciel de Gestion Intégré. Le leader sur du marché est SAP, suivie de People Soft, racheté en 2005 par Oracle.

ETL : *Extraction, Transformation, Load*, en français, extraire transformer et charger. On trouve sur le marché des ETL, des outils indépendants comme Informatica, ou intégrés à une plate-forme décisionnelle comme SAS Enterprise Data Integration Server.

Fidélisation : Ensemble des techniques de marketing direct visant à établir un dialogue continu avec ses clients pour fidéliser ceux-ci au produit, au service, à la marque.

Fonctionnel : caractéristiques liées à l'usage fait de l'information et différenciant la donnée brute de l'information finale.

FM : SAS® *Financial Management* est une solution SAS pour les directions financière.

Fonctionnel : Utilisateur fonctionnel = utilisateur métier ; voir Métier

Gestion de la performance

Gestion des campagnes marketing : Définit l'ensemble des opérations de marketing direct prévues pour soutenir le lancement et la commercialisation d'une offre produit / service.

Giga : voir Téra

HCM : *Human Capital Management* : Gestion des Ressources Humaine. On rencontre souvent ce sigle pour désigner le *Data Mart* RH. SAS Propose la solution SAS *Human Capital Management*.

HOLAP : Hybride OLAP (Voir OLAP)

IA : Intelligence Analytique

IA (2) : Intelligence Artificielle

ID : Informatique Décisionnelle

IML : Le logiciel SAS/IML® est un langage matriciel permettant de développer des méthodes d'analyse spécialisées de modélisation ou de statistiques. Un grand nombre de fonctions de manipulation de matrices (numériques ou caractères) ainsi que des fonctions de calcul matriciel y sont implémentées.

IS : Intelligent Storage : Stockage des données décisionnelles

IVC : Intelligence Value Chain : chaîne du décisionnel qui comprend les phases de Plan, ETL, Intelligent Storage, Business Intelligence & Analytique Intelligence. Le but de L'IVC est de transformer la donnée en intelligence.

JDK : Java Development Kit

JRE : Java Runtime Environnement. Machine virtuelle Java interprétant un programme Java précompilé.

JSP Java Server Page

KDD : Knowledge Data Discovery

Ki^2 : Test statistique

KPI : Key Performance Indicator

L4G : Langage de $4^{ème}$ génération, langage de programmation informatique. SAS est un L4G.

LDAP : LDAP (*Lightweight Directory Access Protocol*, traduisez protocole d'accès aux annuaires léger et prononcez "èl-dap") est un protocole standard permettant de gérer des annuaires, c'est-à-dire d'accéder à des bases d'informations sur les utilisateurs d'un réseau par l'intermédiaire de protocoles TCP/IP.

MDDB : *Multi-Dimentionnal Data Base* : structures de données multidimensionnelles qui permettent le stockage de données consolidées réparties sur plusieurs axes d'agrégation (par exemple, l'axe temporel peut être divisé par années, trimestres, mois; la géographie par pays, région, ville; les produits par gamme, segment...).

Métadonnées : littéralement : données sur la donnée. Il s'agit de l'ensemble des informations qui permettent de qualifier une donnée, par sa provenance, son format, sa date de mise à jour, les droits d'administration, de modification, de lecture, etc. qui lui sont attachés, etc. La métadonnée est une étiquette attachée à la donnée qui doit permettre sa traçabilité depuis son origine, afin notamment d'en garantir la qualité.
Un serveur de métadonnées gère les données sur les données, les utilisateurs, les serveurs, les processus, etc. Il permet donc d'administrer une plate-forme.

Métier : les métiers génériques de l'entreprise sont la production, les ressources humaines, le marketing, les forces de ventes. Les utilisateurs métiers d'un système décisionnel sont donc des utilisateurs généralement non informaticiens, ayant besoin d'accéder de manière simple à la connaissance.

MOA : Maîtrise d'ouvrage

MOE : Maîtrise d'œuvre

Multi-threading : Technologie logicielle permettant de découper une tâche en sous-tâches logiques à l'intérieur d'un même processus de traitement et de le répartir sur plusieurs processeurs simultanément.

SAS NLS : National Language Support n'est pas un module en tant que tel mais c'est une application qui permet à SAS de fonctionner au mieux sur des systèmes non américains. Le fichier de configuration par défaut C:\Program Files\SAS\SASFoundation\9.2\ SASV9.CFG pointe vers le véritable fichier de configuration, par exemple : C:\Program Files\SAS\SASFoundation\9.2\nls\fr\SASV9.CFG

ODS : Output Delivery System. Partie du module SAS/BASE permettant de définir vers différentes sorties (Word, PDF, HTML, Latex, etc.) le format de sortie de procédures SAS.

ODS : Operational Data Store ou Detail Data Store, ou Data Store par simplification. Espace d'entrée de l'entrepôt de données où est stockée une copie souvent quasi conforme des données des systèmes de production.

OLAP : Désigne une catégorie d'outils d'exploration de données permettant de visualiser des indicateurs agrégés et détaillés selon plusieurs axes : temps, géographie, produits par exemple. S'ils s'appuient sur un cube OLAP, on parle de MOLAP, par opposition aux outils ROLAP sur SGBD/R.

OR : *Operational research* : recherche opérationnelle : Le logiciel SAS/OR offre des outils de résolution de problèmes de recherche opérationnelle, de gestion de projet et de programmation non-linéaire.

Parallélisation : Technologie logicielle permettant de répartir plusieurs processus de traitement en parallèle sur plusieurs processeurs ou plusieurs machines organisées en grappe.

Plan : Première phase de l'*Intelligence Value Chain*, phase de compréhension de la problématique de l'entreprise et du besoin des utilisateurs, d'analyse des ressources disponible, pour définir le périmètre de faisabilité du projet décisionnel.

Plate-forme décisionnelle : Système décisionnel complet, intégré, modulaire et ouvert, administrable depuis une console unique et permettant la traçabilité de tous les traitements depuis l'extraction des données jusqu'aux applications analytiques grâce à un système de méta –données transversal et un langage de traitement et d'analyse des données unique. La plate-forme décisionnelle permet de mutualiser les compétences de l'entreprise quel que soit le type de projet décisionnel et de rationaliser l'architecture.

Portail décisionnel : Application web qui offre un point d'accès unique, sécurisé et personnalisable aux informations et aux applications décisionnelles pour tous les utilisateurs de l'entreprise. Il peut s'agir d'un sous -ensemble d'un portail d'entreprise.

PMML : (Predictive Modeling Markup Language) le PMML est un standard basé sur XML d'échange de résultats de Data Mining.

Portail décisionnel : Application web qui offre un point d'accès unique, sécurisé et personnalisable aux informations et aux applications décisionnelles pour tous les utilisateurs de l'entreprise. Il peut s'agir d'un sous-ensemble d'un portail d'entreprise.

QC : Le logiciel SAS/QC® propose l'ensemble des outils statistiques nécessaires au contrôle de la qualité.

Query and Reporting : littéralement : requête et création de rapport. Outil ergonomique permettant à un non spécialiste d'interroger facilement une base de données puis de mettre en forme un rapport.

ROI : Return On Investment, littéralement, retour sur investissement. Calcul du retour sur investissement, de plus en plus fondé sur le rapport entre le TCO (voir plus bas) et l'ensemble des gains mesurables.

SAS : A l'origine, « Statistical Analysis System ». Se prononce "sas" et non S.A.S. En fonction du contexte, SAS désigne le produit (SAS/BASE), la plate-forme, le langage (un programme en code SAS) ou bien la société (en France, la raison sociale est « SAS Institute s.a.s. ») qui les commercialise.

Scoring : Un score est une note. Les modèles prédictifs de *Data Mining* permettent de noter un individu sur sa propension à, par exemple, être un bon client. Scorer une base de client, c'est utiliser un modèle mathématique ayant appris sur un historique, pour affecter à chaque client sa note, souvent sa probabilité, à acheter ce produit, ou bien à partir, etc. Faire du scoring, c'est donc apprendre sur un historique, pour simuler le futur.

Segmentation : Caractérise les méthodes grâce auxquelles un fichier pourra être segmenté, c'est-à-dire défini plus finement dans le détail des cibles qui le composent.

SEM : SAS *Enterprise Miner* est la solution SAS pour le *Data Mining*. Il a été conçu pour être utilisé aussi bien par des utilisateurs métiers (chefs de produit, de campagne…) que par des professionnels du *Data Mining*. SAS *Enterprise Miner* intègre l'ensemble des étapes d'un processus de *Data Mining* :
- accès et préparation des données
- visualisation, exploration, manipulation des données
- modélisation (arbres de décision, réseaux de neurones, modèles de régressions, classifications, associations…)
- comparaison et validation des modèles
- industrialisation des résultats

SAS Enterprise Miner offre une ergonomie d'utilisation semblable à celle des outils bureautiques.
Les processus créés peuvent être sauvegardés et distribués de façon simple à tout type d'utilisateur (averti ou novice) dans l'entreprise.

SGBD : Système de Gestion de Base de Données

SIAD : Système Interactif d'Aide à la Décision ou Système d'Information pour l'Aide à la Décision : Environnement permettant de stocker, de structurer et de restituer l'information décisionnelle.

SID : Système d'Information Décisionnelle

SID : Le SAS Installation Data, anciennement appelé setinit, est la clé d'accès de SAS. Lorsqu'il arrive à expiration, SAS ne fonctionne plus.

SSII, ou SS2I : une Société de Services en Ingénierie Informatique est une société de services spécialisée en informatique.

Suite décisionnelle : Système décisionnel constitué par rachats ou développements successifs. La suite décisionnelle demande généralement une administration et des développements multiples et hétérogènes limitant la mutualisation des compétences et la simplicité d'administration.

SIM : Système d'Information Marketing : Stock de données collectées et/ou analysées sur des clients ou des prospects. Aux données internes, dont les informations ont été recueillies par l'entreprise elle-même, peuvent être croisées des données externes louées ou achetées à des entreprises spécialisées.

TAID : Algorithme d'arbre de décision

TCO (Total Cost of Ownership) : Calcul du coût de possession qui, au-delà du coût d'acquisition, prend en compte les dépenses liées à la mise en œuvre, à la formation, à l'exploitation ou à la maintenance.

Téra :

Un kilooctet = 1000 octet = 10^3 octet
Un mégaoctet = 1000 kilooctet = 10^6 octet
Un gigaoctet = 1000 kilooctet = 10^9 octet
Un téraoctet = 1000 gigaoctet = 10^{12} octet
Un pétaoctet = 1000 téraoctet = 10^{15} octet

TCP/IP : Terminal Control Protocol / Internet Protocol. Protocoles de niveaux 4 et 3 respectivement utilisé sur internet.

UP SELLING : Terme marketing désignant la méthode de vente qui consiste à offrir un produit de gamme supérieure, le client est poussé à dépasser le prix qu'il s'était fixé pour l'achat d'un produit.

WA : *WareHouse Administrator* (nom de l'ETL de la version 6 à 8 de SAS)

Wizard : Si la traduction littérale est 'magicien', le *Wizard* est en informatique un petit guide en quelques fenêtres qui n'a rien de magique. Vous trouverez notamment un *Wizard* pour vous accompagner dans l'installation d'un produit, il vous demande souvent l'emplacement où que vous souhaitez installer le produit (C:\Program Files\...) etc…

WIK : Web Infrastructure Kit, est un ensemble de composants permettant la connexion entre la partie Mid tier et la partie Server Tier.

WRS : SAS® Web Report Studio est l'interface Web légère de *Reporting* de masse. Il n'est pas nécessaire d'installer quelque chose sur le poste client, l'utilisateur accède à cette application depuis Internet Explorer.

XML (Extensible Markup Language ou langage de balisage extensible) est un standard du World Wide Web Consortium qui sert de base pour créer des langages balisés spécialisés; c'est un « méta langage ».

YTD : Year To Date : Beaucoup de mesure, de KPI présentent les chiffres en YTD, chiffre du jour par rapport à ceux de l'année dernière à la même date.

Bibliographie fondamentale, à lire !

1. Le Data Warehouse : Guide de conduite de projet, Ralph Kimball, Editions Eyrolles (Il est fortement recommandé de le lire)
2. Entrepôts de données. Guide pratique de modélisation dimensionnelle, Ralph Kimball, 2ème édition Eyrolles
3. Data Mining et Statistiques décisionnelles, Stéphane TUFFERY, Editions Technip (Pour ceux qui s'intéresse au Data Mining, il est fortement recommandé de le lire)
4. SAS 9.2 - Introduction SAS : Sébastien Ringuedé, Edition Pearson (Très bon livre en français pour commencer à programmer en SAS, nécessaire et suffisant pour préparer la certification SAS)

Site Web

www.sas.com (site commercial de SAS)
support.sas.com (site des utilisateurs de SAS)
www.bettermanagement.com (il est fortement conseillé de s'abonner à plusieurs lettres d'information. C'est un site SAS)
http://decisio.info/ (il est fortement conseillé de s'abonner la lettre d'information. C'est un site SAS)
http://www.sas.com/apps/whitepapers/whitepaper.jsp (Il faut s'enregistrer. Cela n'engage en rien. Il n'y aura pas de relance commerciale. Il est fortement conseillé de télécharger et de lire plusieurs livres blancs)
http://support.sas.com/learn/statlibrary/ (analyse de données avec SAS Enterprise Guide)
http://www.stat.ucl.ac.be/cours/stat2020/documents/manuels_logiciels/SASV9-Preudhomme.pdf (un