HUPI

valoriser analytiquement votre différence

FORMATION

HUPI-INTERACTIF & HUPI-LINK & HUPI-ANALYTICS

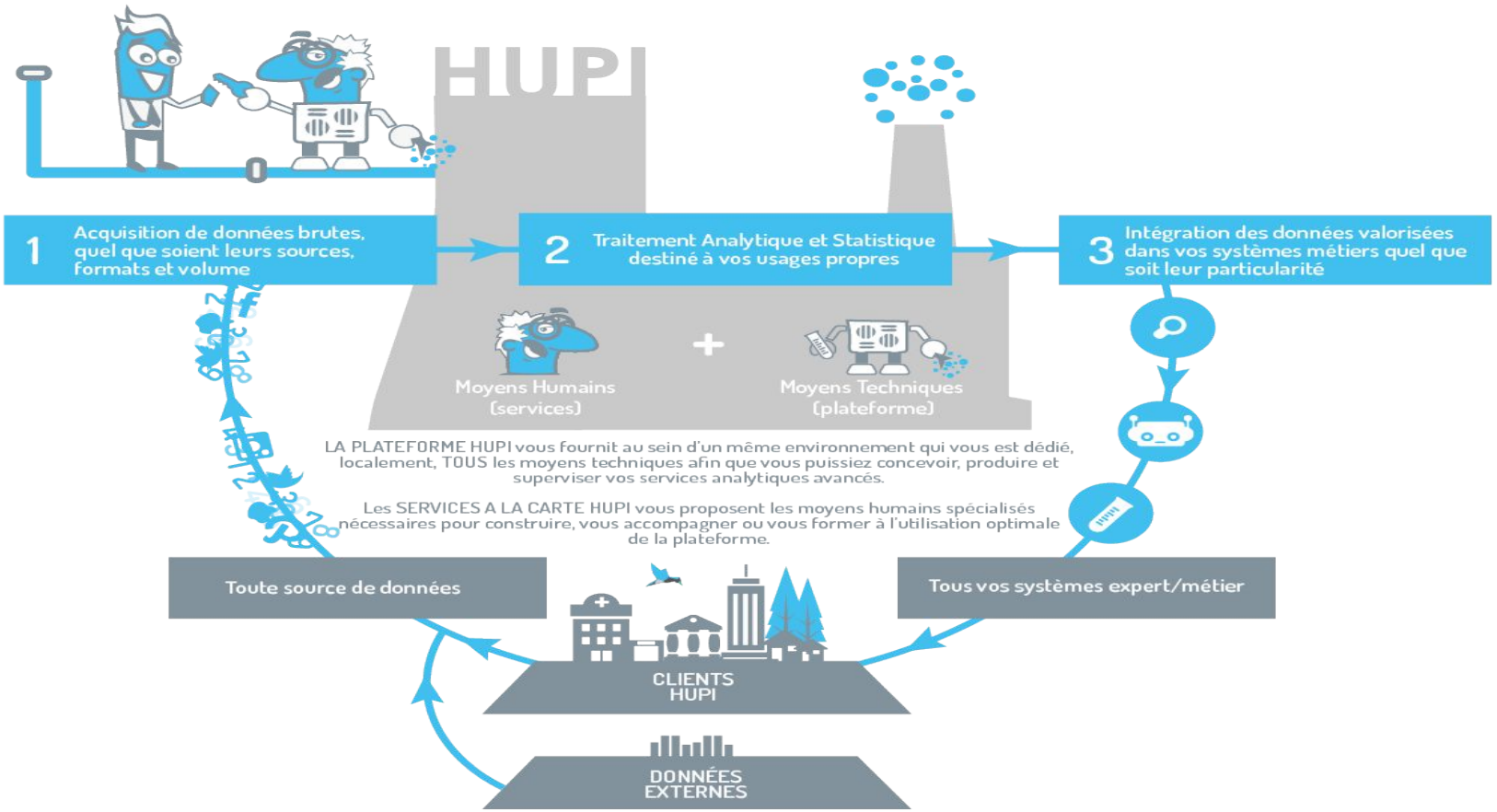
2018

HUPI,
C'EST LA CAPACITÉ D'ACTIONNER DES RECOMMANDATIONS AUTOMATISÉES DANS VOS SYSTÈMES EXPERTS

HUPI

VALORISER ANALYTIQUEMENT VOTRE DIFFÉRENCE

NOTRE SOLUTION (des moyens techniques animés (si besoin) par des moyens humains)



NOTRE MISSION, NOTRE PROMESSE

L'ANALYTIQUE POUR TOUS, UNE COMPÉTENCE A ANCRER ICI

NOTRE MISSION

HUPI s'engage à rendre la compétence de traitement et de valorisation analytique des flux de données accessible à tous

Valorisez analytiquement votre différence, quelle que soit votre taille et votre activité….

...Et devenez acteur de la transition analytique de votre territoire..

...En expérimentant et produisant en continu des services innovants issus de l'analyse scientifique de vos données….

…Et en s'appropriant les compétences essentielles du Traitement et de la Valorisation Analytiques des Flux de Données, rendues accessibles par l'utilisation de l'Usine HUPI

OBJECTIFS

Cette formation a pour objectif de donner à ses participants:

- une vision générale de l'industrie des plateformes de traitement analytique, de l'architecture type, des technologies utilisées, des problématiques d'intégration, ainsi que des principaux usages métiers et services rendus
- une expérience pratique et un premier niveau d'expertise opérationnelle dans l'utilisation de ces technologies au sein d'une plateforme analytique: instanciation d'un environnement de traitement analytique, intégration de flux de données, développement et exécution de modèles analytiques *(exemple Machine Learning)*, restitution de résultats et la création d'une recommandation.

Basé sur des cas pratiques, à l'issu de cette formation, vous serez capables de comprendre et mettre en place une chaîne complète de valorisation de la donnée.

PUBLIC

Public ciblé

- Ingénieurs, data scientists, data analyste, développeurs, consultants en système décisionnels

Vos intérêts

- L'analyse et la fouille des données
- La recherche de nouveaux services liés à la valorisation des données
- Les processus "industriels" de traitement des données
- Les architectures et les technologies Big Data

Prérequis

- Notions de développement informatique
- Notions de mathématiques statistiques
- Notions de SQL

DÉROULÉ DE LA FORMATION

- Concepts Généraux de Votre Usine de Traitement et de Valorisation des Données
- Architecture fonctionnelle et technique: Présentation des différentes technologies et de leur intégration
- Connecteurs aux sources de données
- Stockage (NoSQL, HDFS)
- Analytique (langage R, Hadoop, HIVE /Spark).
- Calcul distribué (MapReduce)
- Visualisation des résultats de vos traitements analytiques
- API dynamique en temps réel (NRT) de restitution au sein de systèmes experts

DÉROULÉ DE LA FORMATION

- “Entrez” dans votre usine
- Présentation des fonctionnalités principales
- Support HUPI
- Administration de votre Usine
- Partager vos représentations graphiques (DataVisualisation)
- Accès aux données de vos systèmes experts (RestFUL)

Travaux Pratiques

Au cours de cette formation, vous allez “Développer une solution complète de Recommandation eCommerce”

Plan :

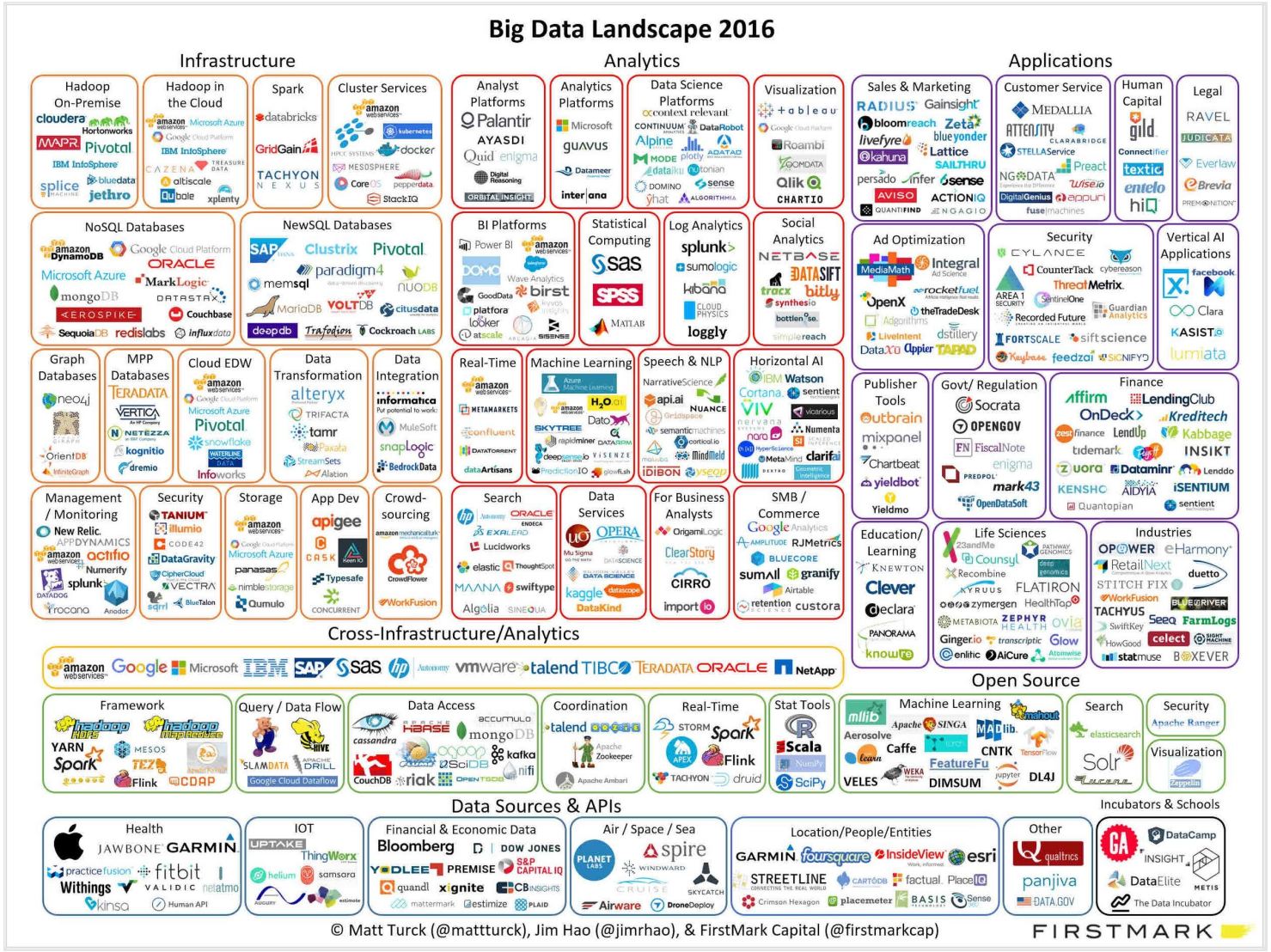
- Créer un “mini” site Internet
- Capter les comportements de vos utilisateurs
- Mesurer l’activité de votre site
- Créer votre premier modèle(s) de recommandations
- Déployer votre modèle
- Mesurer les résultats

Mais avant ....

Que savez-vous des plateformes “Big Data” / Intelligence Artificielle ?

https://docs.google.com/spreadsheets/d/1xfVUUFIUoyBfvhFpcQSxNZJlHFn9TI0XXpPDESkN6as/edit?usp=sharing

Landscape 2017 Big Data Platform



DÉROULÉ DE LA FORMATION

- **Jour 1:**
  - Matinée
    - Présentation HUPI
    - Les produits HUPI-INTERACTIF et HUPI-LINK
  - Après-midi
    - Mise en œuvre de Dashboards
- **Jour 2:**
  - Matinée
    - Dashboards experts
    - Réflexion sur cas pratiques
  - Après-midi
    - Exercices pratiques
    - Présentation HUPI-ANALYTICS

PRÉSENTATION HUPI

- Jour 1: Matinée
  - Concepts
  - Votre Usine
  - L'architecture
  - Support
  - La Datavisualization
  - Cas d'usage
  - Entrez dans votre usine
  - Fonctionnalités
  - Administration
  - Créer son premier Thème

- Jour 1: Après-midi
  - Partager vos Widgets
  - Les Dashboards
  - Les Widgets
  - Filtres
  - Customisation
  - Accès aux données systèmes experts

LÉGENDES

Une information

Trucs et Astuces

Opération devant être réalisée avec précaution

Exercices

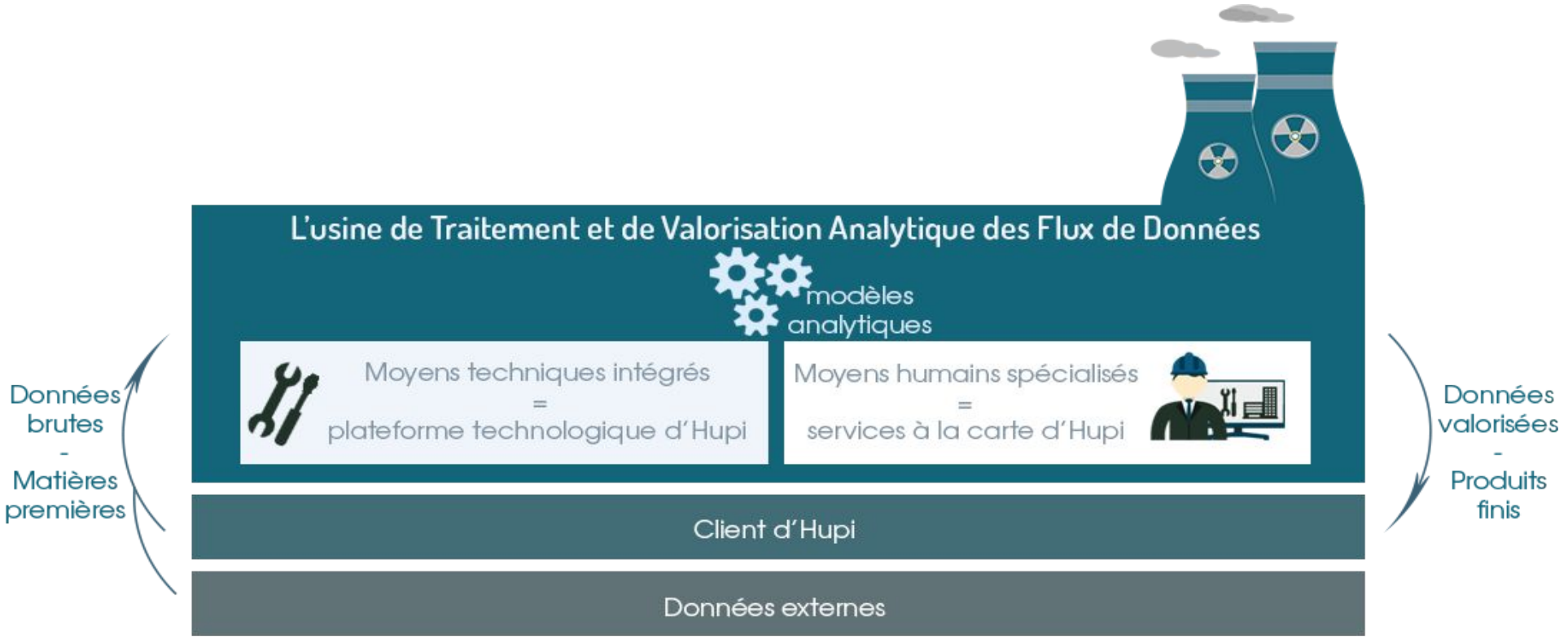
NIVEAU 1 — Niveau accessible à tous les utilisateurs (aucun prérequis technique)

NIVEAU 2 — Niveau accessible aux utilisateurs ayant des connaissances des données (peu de prérequis technique)

NIVEAU 3 — Niveau accessible à des utilisateurs orienté « développement » (prérequis technique)

NOTRE SOLUTION

DES MOYENS TECHNIQUES ANIMÉS (SI BESOIN) PAR DES MOYENS HUMAINS



Chaque Usine permet ainsi la conception, la mise en œuvre, l'expérimentation, la production en continu, et l'intégration dans vos systèmes métiers de vos services analytiques avancés

HUPI

CARACTÉRISTIQUES ET BÉNÉFICES

DES BÉNÉFICES IMMÉDIATS

PRINCIPAUX BÉNÉFICES

#1. Simplifiez, accélérez et dé-risquez

#2. Libérez vous du casse tête lié à la réalisation et résolution de tâches secondaires

#3. Devenez expert et autonome, pour les fonctions ou processus de votre choix

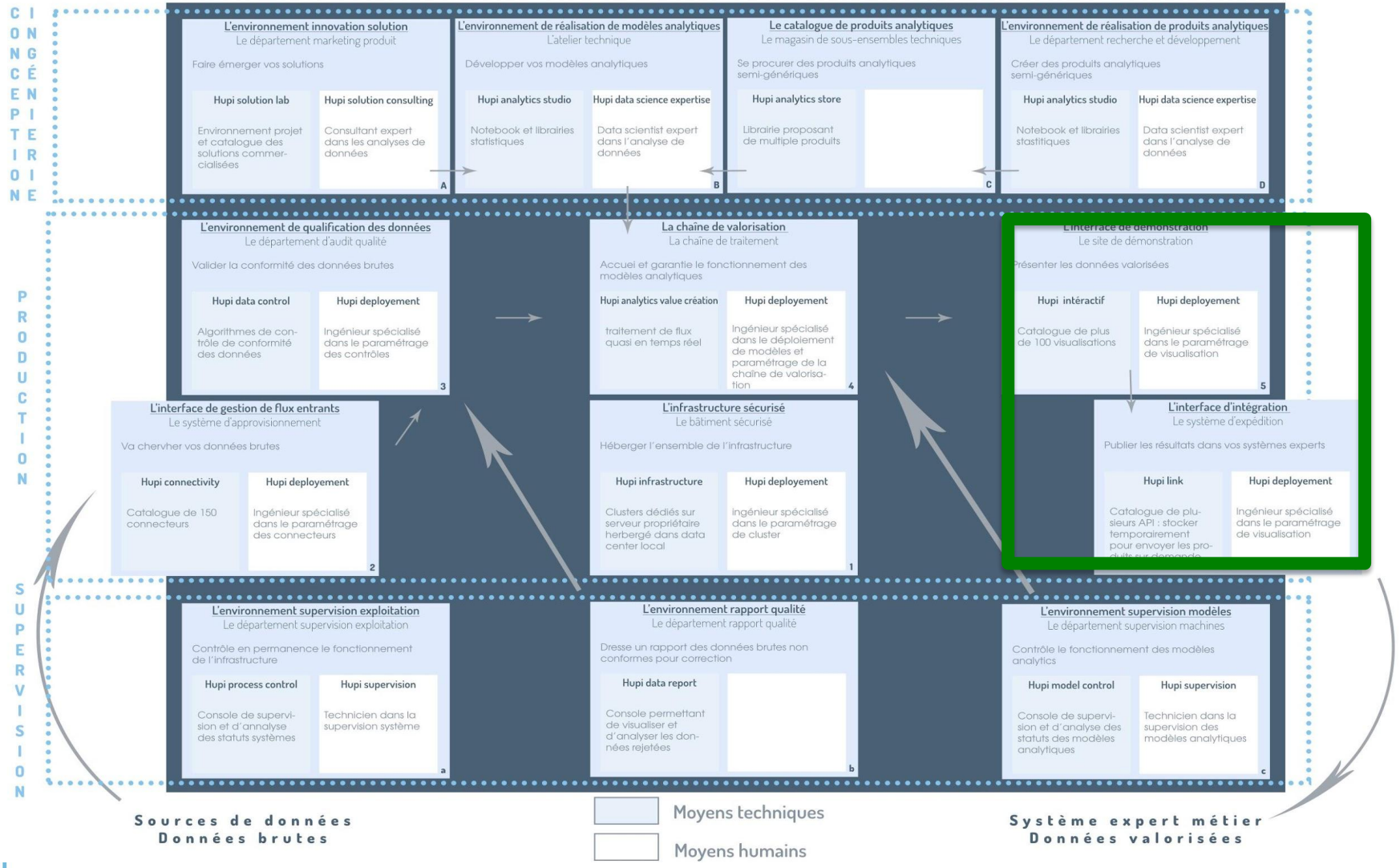
CARACTÉRISTIQUES DE LA PLATEFORME

- ✓ Unique, “tout en un”, intégrée, clés en main
- ✓ Intégrant les dernières technologies d’analyse et de traitement des données
- ✓ Dédiée, ultra performante, scalable et dimensionnée en fonction des besoins
- ✓ Accessible, intuitive et 100% web
- ✓ Ouverte et interfaçable avec tout système métier
- ✓ Locale et sécurisée

SERVICES

- ✓ Capable de servir autant les non-initiés que les experts
- ✓ Services à la carte, personnalisés et de proximité
- ✓ A vocation métier, technologique ou scientifique

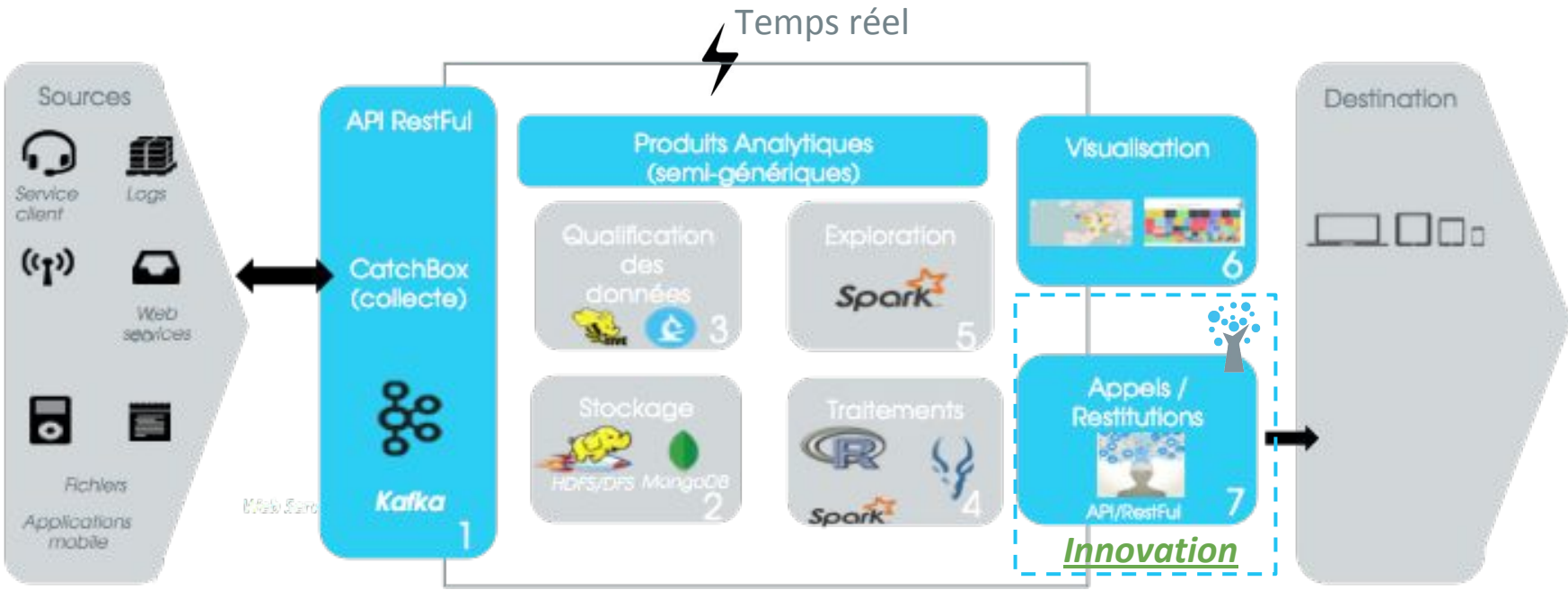
VOTRE USINE


CONCEPTION INGÉNIERIE
PRODUCTION
SUPERVISION
L'environnement innovation solution
Le département marketing produit
Faire émerger vos solutions
Hupi solution lab
Environnement projet et catalogue des solutions commercialisées
Hupi solution consulting
Consultant expert dans les analyses de données
A
L'environnement de réalisation de modèles analytiques
L'atelier technique
Développer vos modèles analytiques
Hupi analytics studio
Notebook et librairies statistiques
Hupi data science expertise
Data scientist expert dans l'analyse de données
B
Le catalogue de produits analytiques
Le magasin de sous-ensembles techniques
Se procurer des produits analytiques semi-génériques
Hupi analytics store
Librairie proposant de multiple produits
C
L'environnement de réalisation de produits analytiques
Le département recherche et développement
Créer des produits analytiques semi-génériques
Hupi analytics studio
Notebook et librairies stastitiques
Hupi data science expertise
Data scientist expert dans l'analyse de données
D
L'environnement de qualification des données
Le département d'audit qualité
Valider la conformité des données brutes
Hupi data control
Algorithmes de contrôle de conformité des données
Hupi deployement
Ingénieur spécialisé dans le paramétrage des contrôles
3
La chaîne de valorisation
La chaîne de traitement
Accueil et garantie le fonctionnement des modèles analytiques
Hupi analytics value création
traitement de flux quasi en temps réel
Hupi deployement
Ingénieur spécialisé dans le déploiement de modèles et paramétrage de la chaîne de valorisation
4
L'interface de démonstration
Le site de démonstration
Présenter les données valorisées
Hupi intéractif
Catalogue de plus de 100 visualisations
Hupi deployement
Ingénieur spécialisé dans le paramétrage de visualisation
5
L'interface de gestion de flux entrants
Le système d'approvisionnement
Va chervher vos données brutes
Hupi connectivity
Catalogue de 150 connecteurs
Hupi deployement
Ingénieur spécialisé dans le paramétrage des connecteurs
2
L'infrastructure sécurisé
Le bâtiment sécurisé
Héberger l'ensemble de l'infrastructure
Hupi infrastructure
Clusters dédiés sur serveur propriétaire herbergé dans data center local
Hupi deployement
ingénieur spécialisé dans le paramétrage de cluster
1
L'interface d'intégration
Le système d'expédition
Publier les résultats dans vos systèmes experts
Hupi link
Catalogue de plusieurs API : stocker temporairement pour envoyer les produits sur demande
Hupi deployement
Ingénieur spécialisé dans le paramétrage de visualisation
L'environnement supervision exploitation
Le département supervision exploitation
Contrôle en permanence le fonctionnement de l'infrastructure
Hupi process control
Console de supervision et d'annalyse des statuts systèmes
Hupi supervision
Technicien dans la supervision système
a
L'environnement rapport qualité
Le département rapport qualité
Dresse un rapport des données brutes non conformes pour correction
Hupi data report
Console permettant de visualiser et d'analyser les données rejetées
b
L'environnement supervision modèles
Le département supervision machines
Contrôle le fonctionnement des modèles analytics
Hupi model control
Console de supervision et d'analyse des statuts des modèles analytiques
Hupi supervision
Technicien dans la supervision des modèles analytiques
c
Sources de données
Données brutes
Moyens techniques
Moyens humains
Système expert métier
Données valorisées

ARCHITECTURE : UNE PLATEFORME TOUT INTEGRÉE

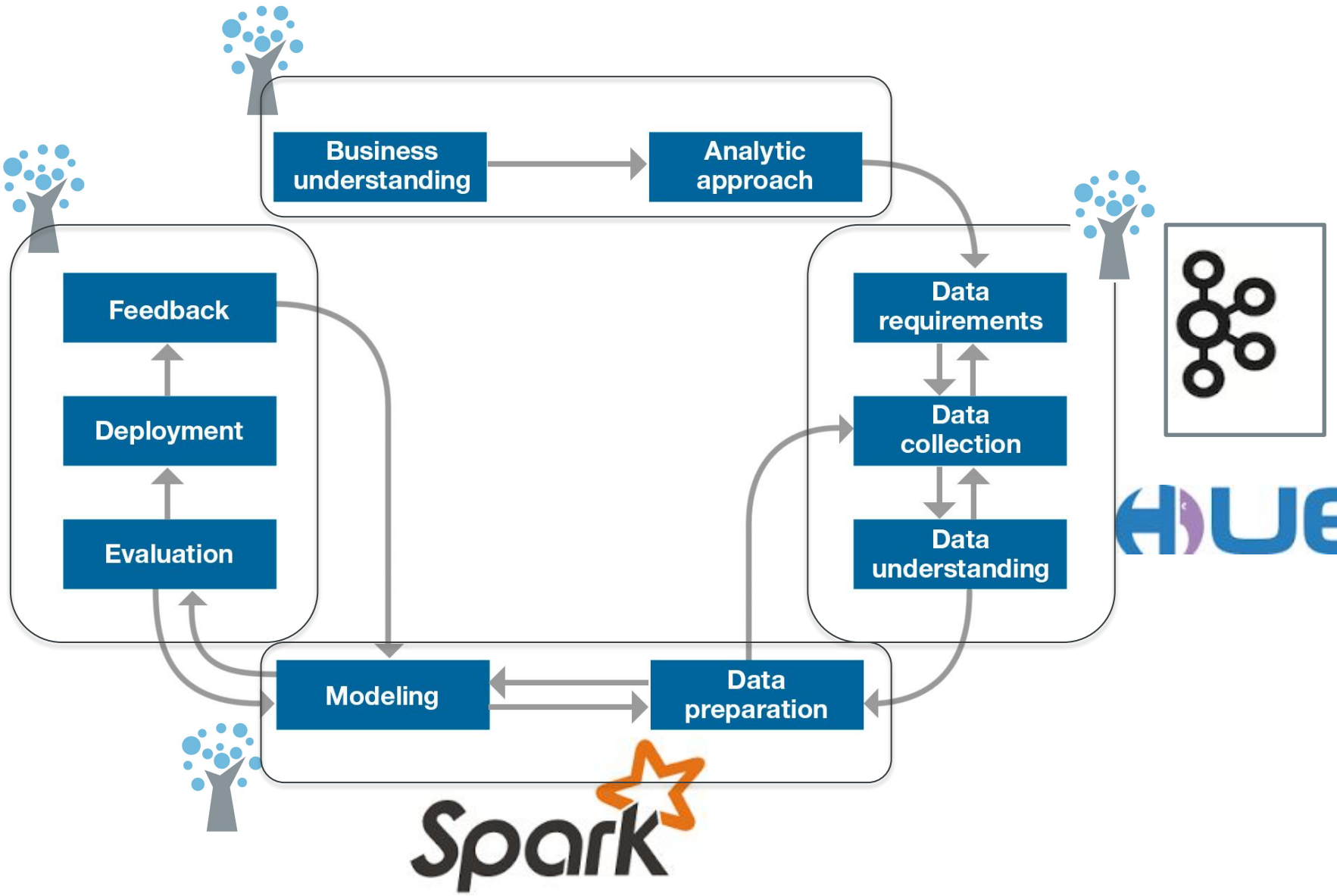
Plateforme hébergée sur nos serveurs dans un Datacenter local

Environnement dédié pour chaque client



Collecte : connecteur pour la collecte massive, sécurisée et standardisée de tout type de flux de données (dont API pour réception)	**Stockage :** massif des données en haute performance clusterisée	**Qualification :** analyse de données massives et hétérogènes pour leur préparation	**Traitement :** développement de calcul statistique distribué en temps réel utilisé pour la Création des Modèles Analytiques de Traitement des données	**Visualisation :** dashboards et bibliothèque graphique pour la Visualisation des Résultats des Traitements Analytiques	**Appels / Restitution :** interrogation des modèles analytiques pour la réintégration en temps réel des Résultats au sein des systèmes d'information.

LES TECHNOLOGIES: CAPTATION DE LA DONNÉE



SUPPORT

Pour toutes vos questions :

- Comment créer un thème ?
- Je veux ajouter des données ?
- Je souhaite une nouvelle fonctionnalité
- J'ai une question sur un fonctionnement ?

Plusieurs adresses :

https://support.hupi.io (accessible depuis votre portail)

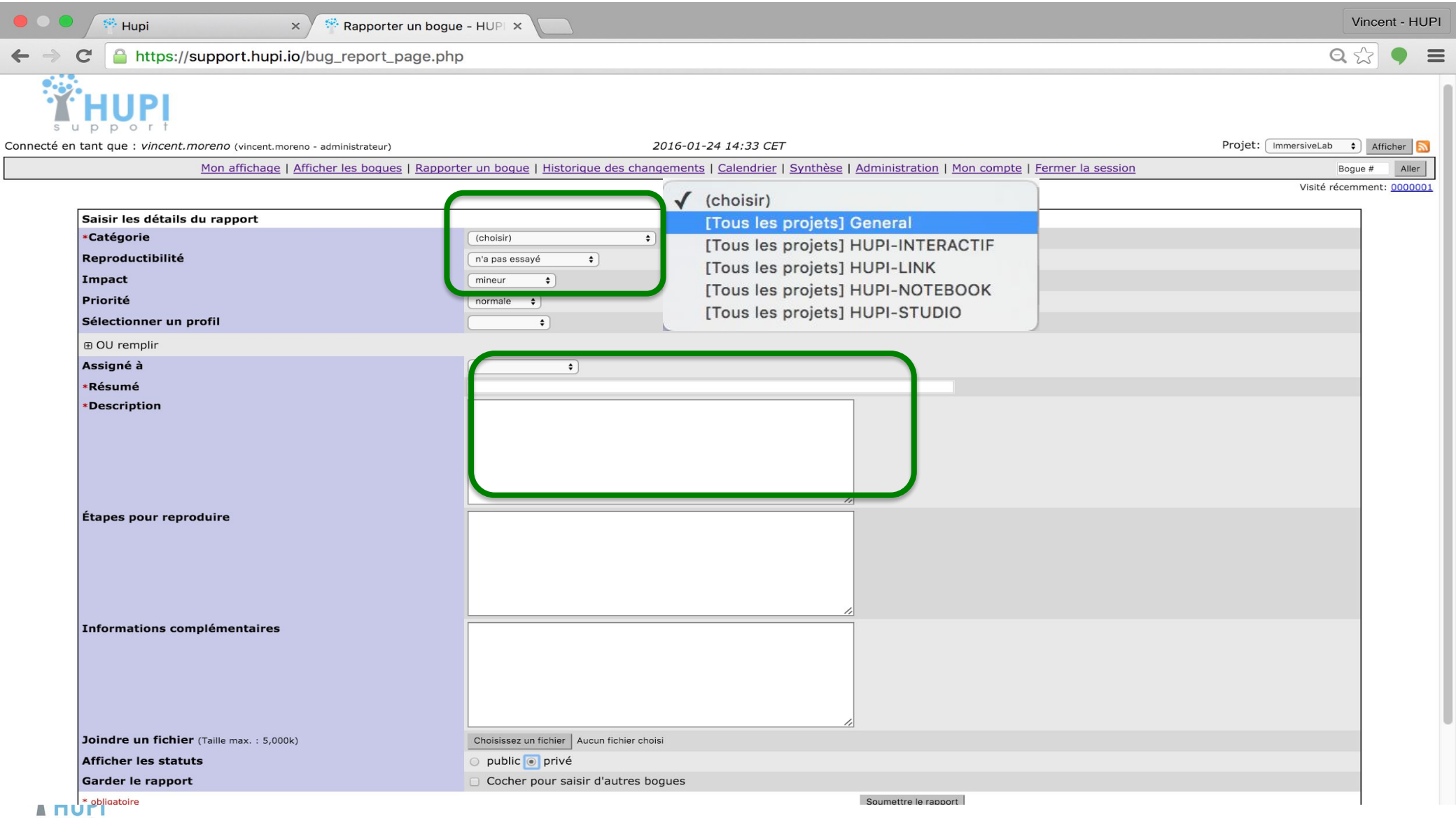
http://www.hupi.fr

https://github.com/hupi-analytics/ Nos projets open-source

Notre ambition, vous apportez les réponses les plus adaptées

NIVEAU 1

SUPPORT



VOCABULAIRE

NIVEAU 1

- **ACCOUNTS**

Correspond au nom d'une usine, permet d'accéder à toutes les données

NIVEAU 1

- **THEMES**

Contient un ensemble de dashboards

NIVEAU 1

- **DASHBOARDS**

Regroupement logique de Widgets

NIVEAU 2

- **WIDGETS**

Représentation graphique des données

NIVEAU 1

- **IFRAME**

Elément permettant de partager publiquement des widgets

NIVEAU 3

- **HDR-EndPoints**

Regroupement logique de données, interface de restitution des données (flux de données, prédiction)

NIVEAU 3

- **HDR-Query Engines**

Moteur pour accéder à des données internes ou externes dans différentes sources.

HUPI

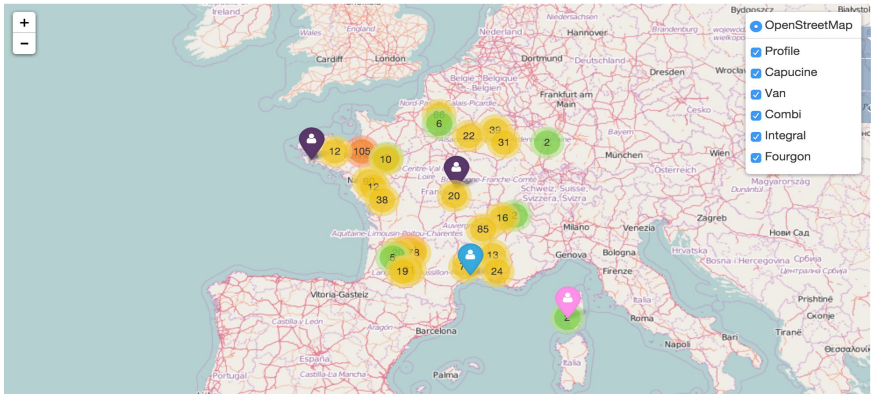
LA DATAVISUALIZATION

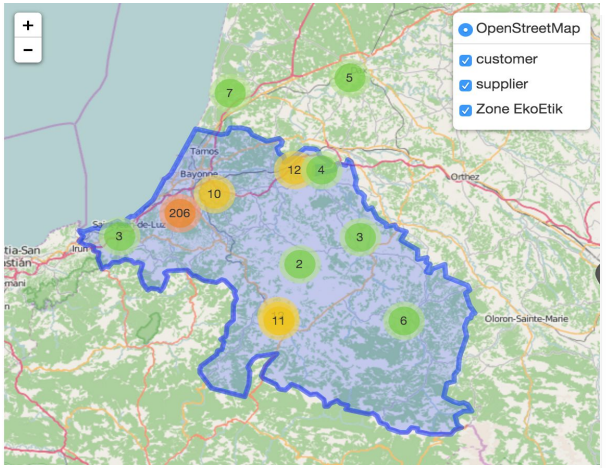
90% de l'information transmise au cerveau est visuelle et les images sont traitées par le cerveau 60 000 fois plus vite que le texte *(sources: 3M Corporation & Zabisco)*

77% déclarent qu'elle améliore la prise de décision grâce à de meilleures analyses

41% considèrent qu'elle améliore la collaboration et le partage de l'information

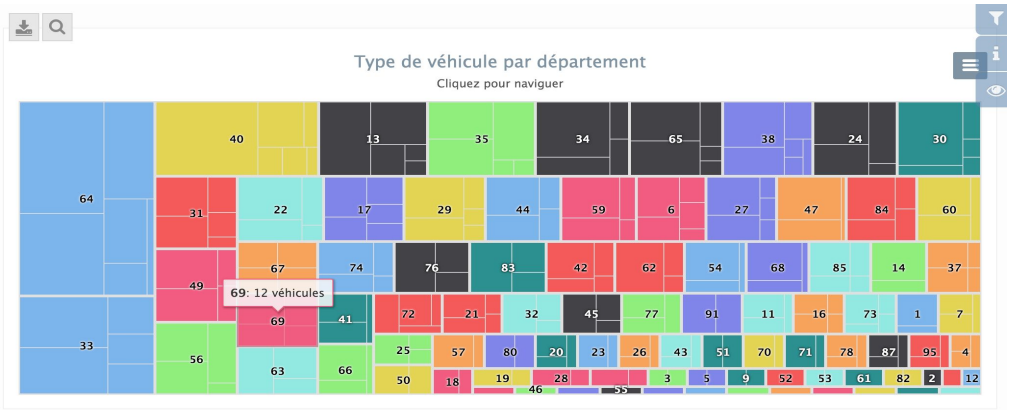
LA DATAVISUALIZATION

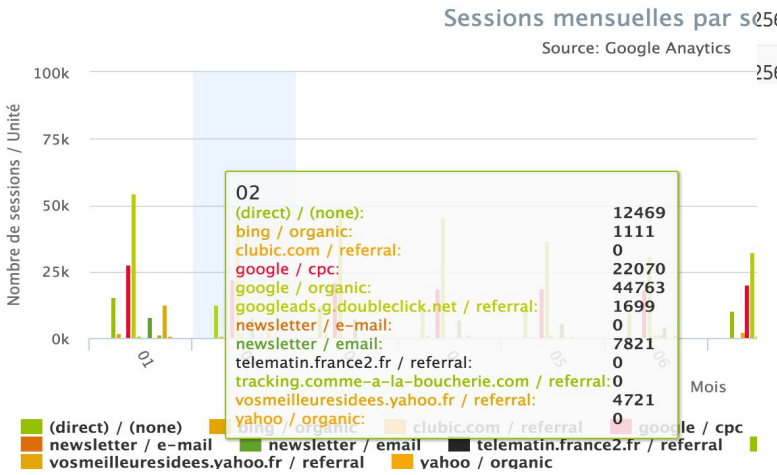




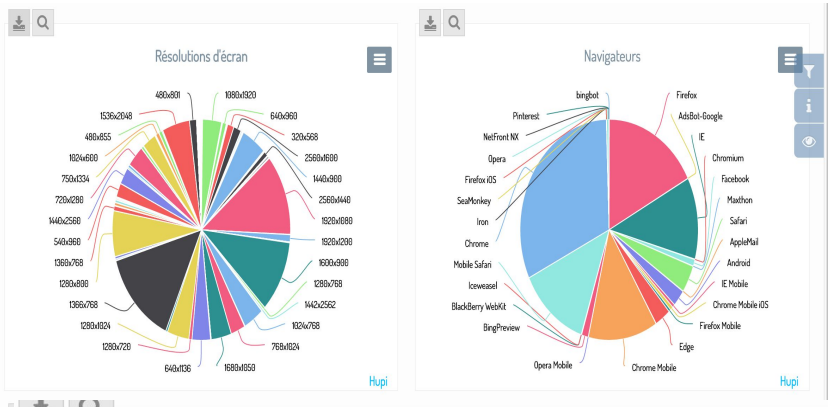
Prédiction sur les pertes de nouveau par rapport aux producteurs des deux premiers mois d'achat

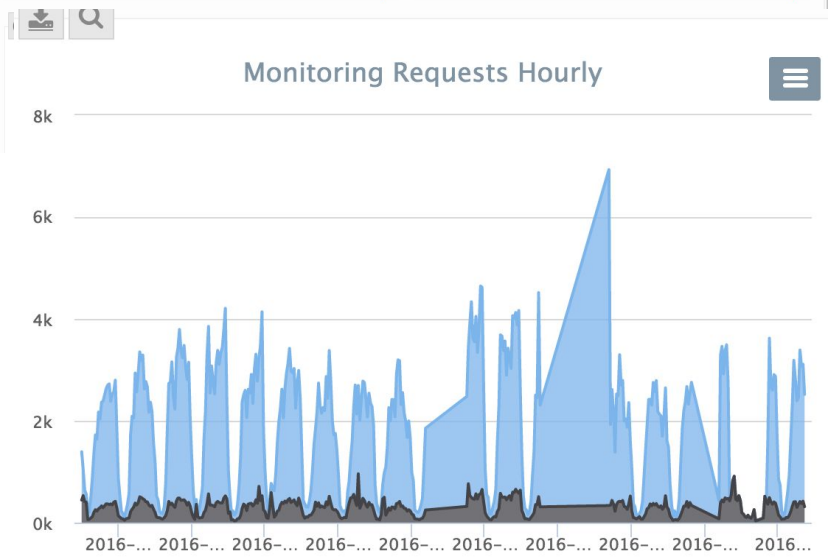
Entity_name	Valeur_predic
Tania	0.0256410256
Stéphanie	0.0256410256
Sophie	0.0256410256
Sandrine	0.0256410256
Sandra	0.0256410256
Roberte	0.0256410256
Quentir	0.0256410256
Muriel	0.0256410256

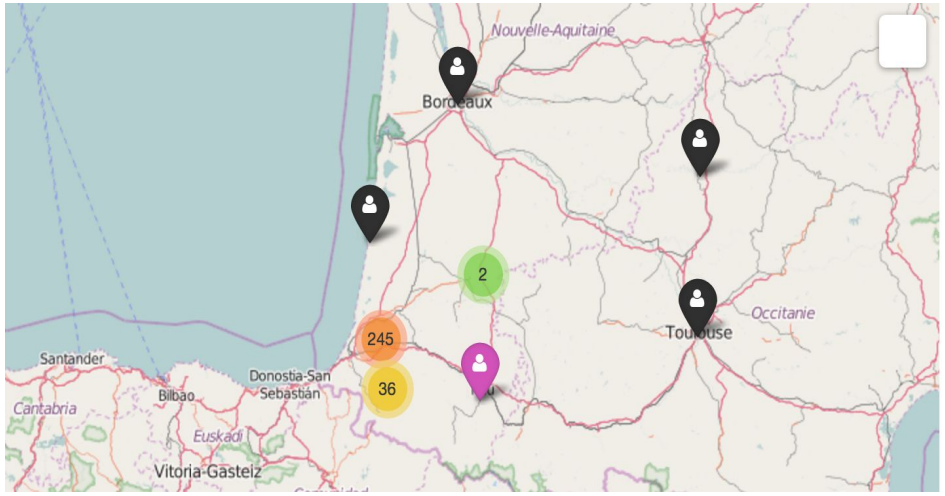


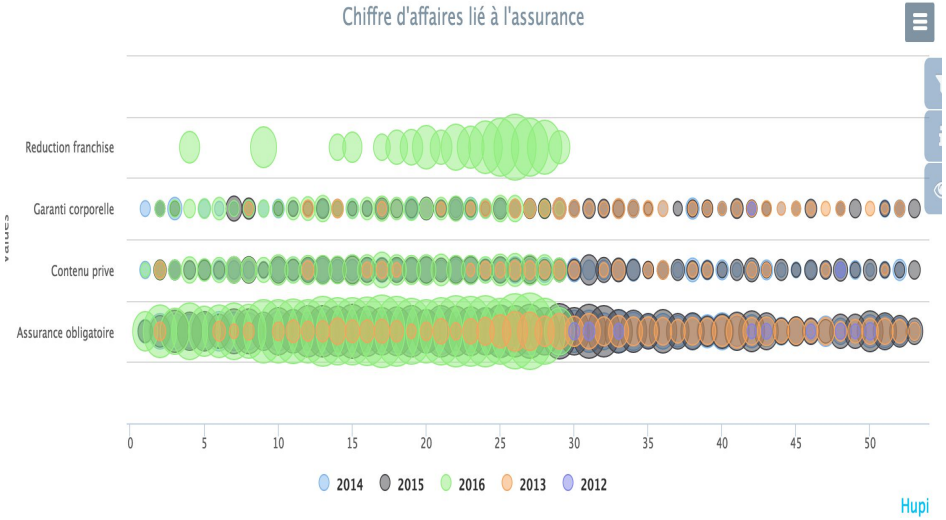


Exemples de Dashboard

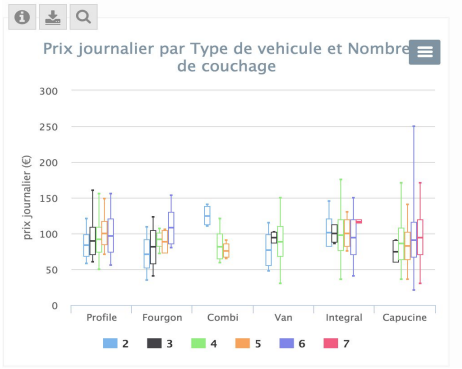


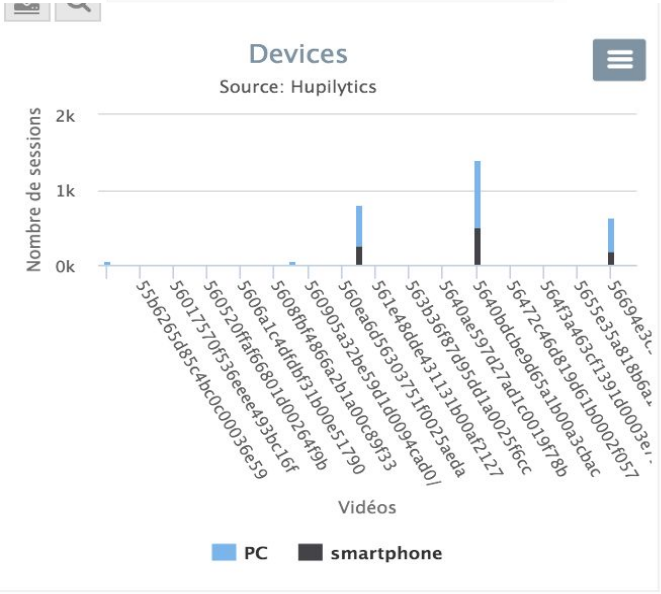


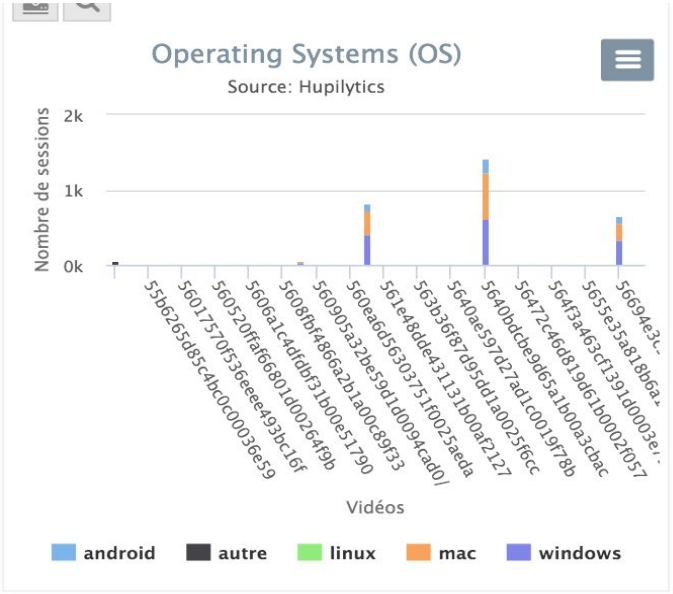


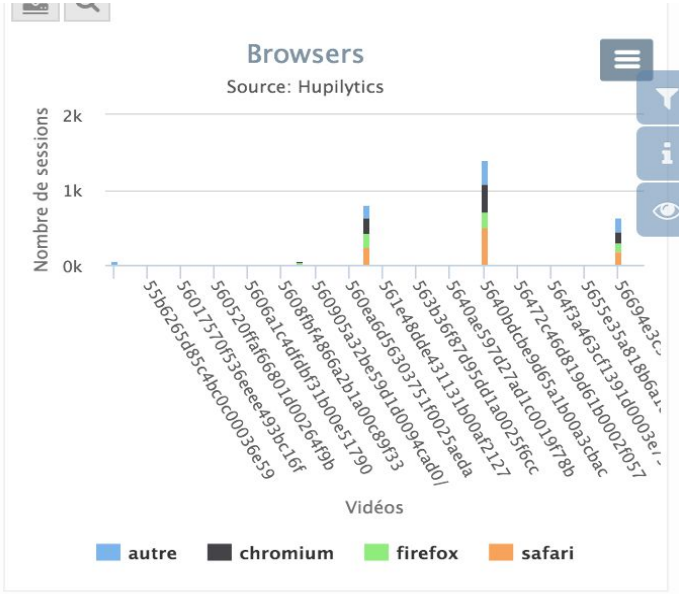


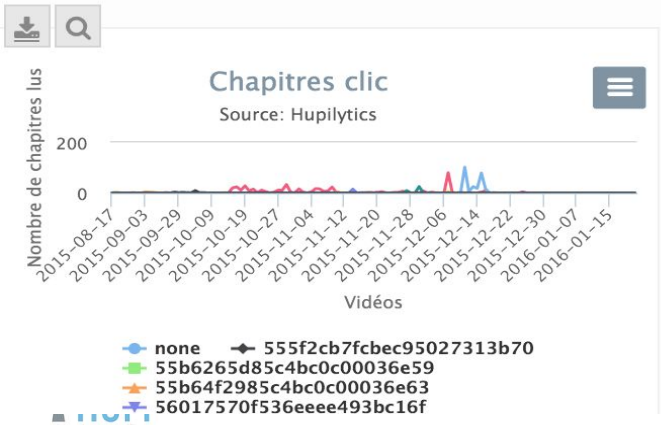
LA DATAVISUALIZATION

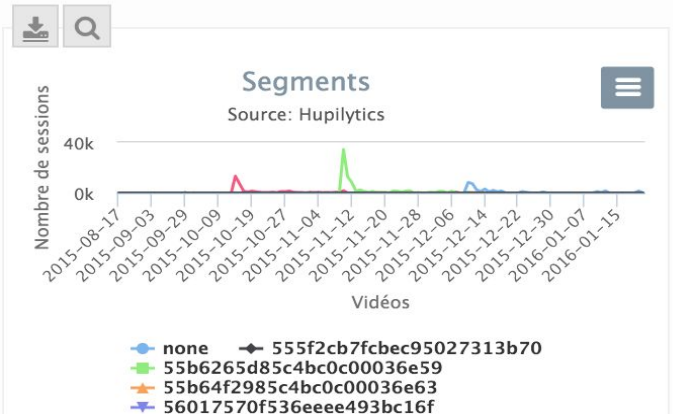


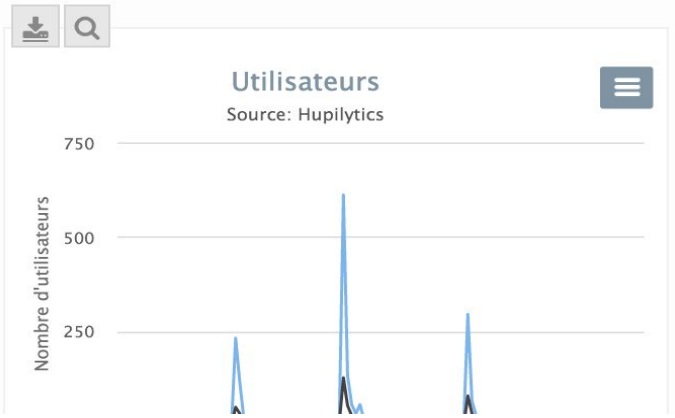












CAS D'USAGE

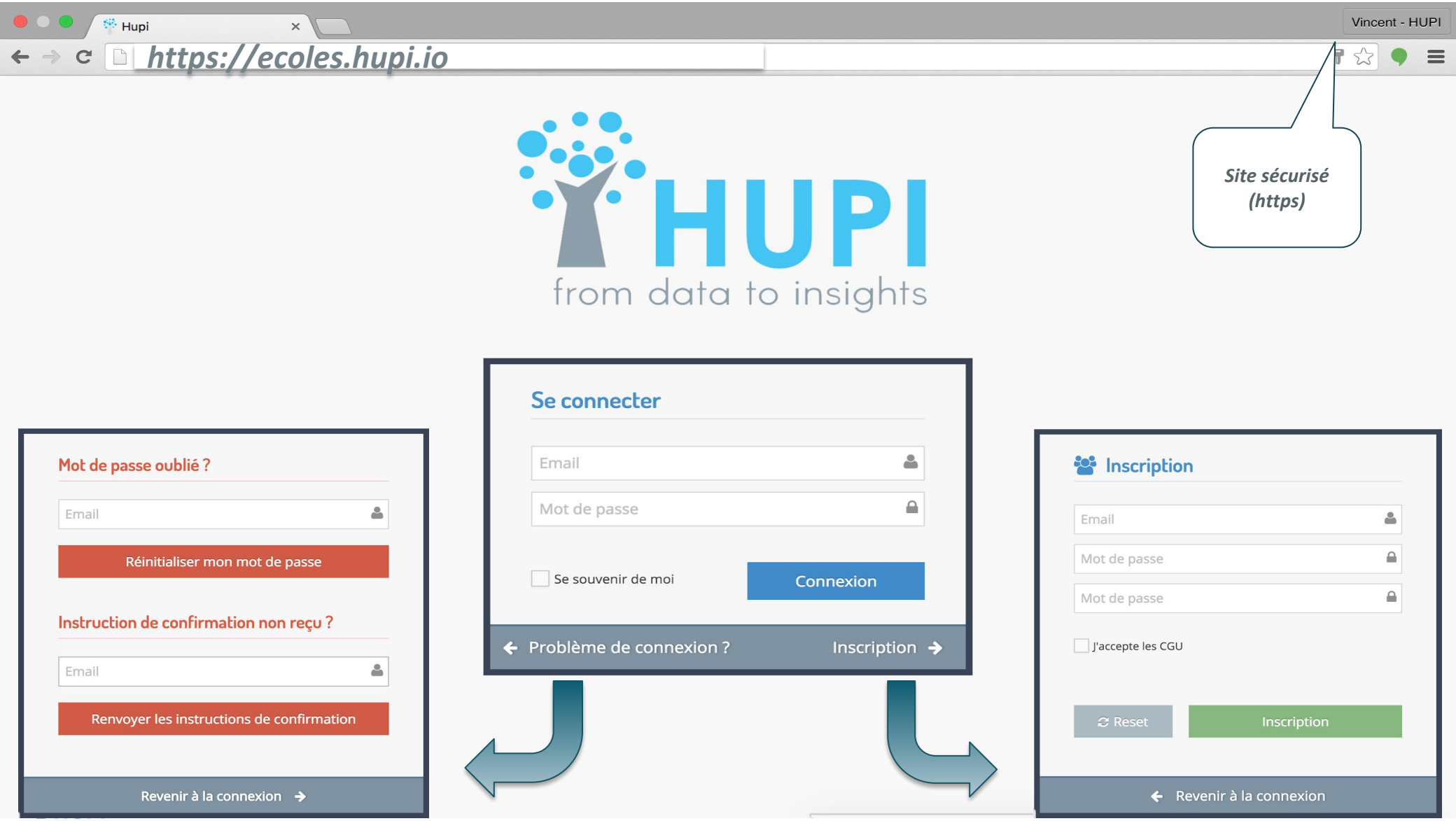
- Différencier les Dashboards de synthèse des Dashboards réservés à l'analyse de données

- Associations
- Classifications
- Clustering
- A/B Testing

- Réseaux neuronnaux
- Prévisions temporelle
- Géolocalisation

Conseils pour le choix de Datavisualization fonction des problèmes - http://www.datavizcatalogue.com/index.html

ENTREZ DANS VOTRE USINE

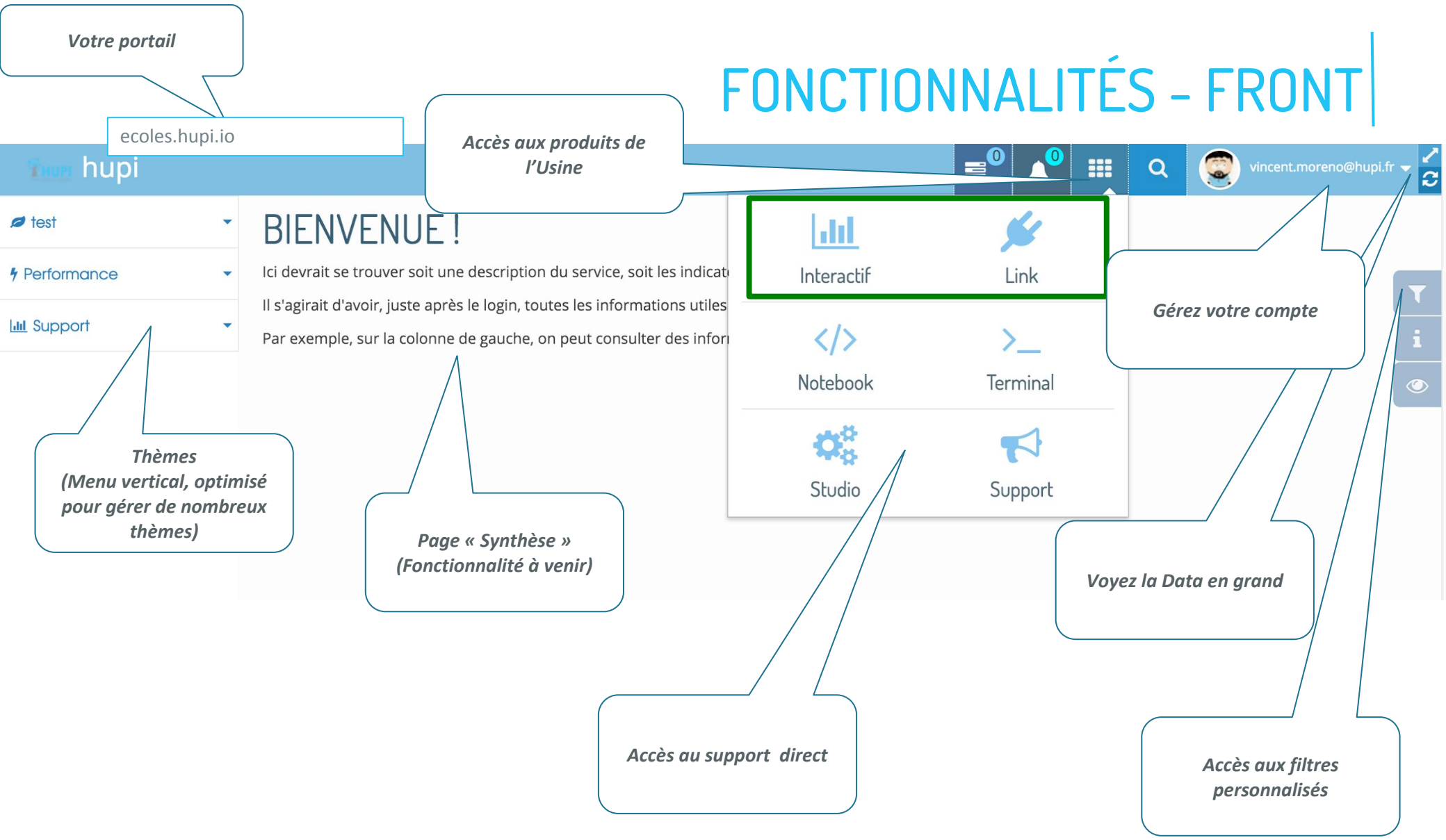


ENTREZ DANS VOTRE USINE

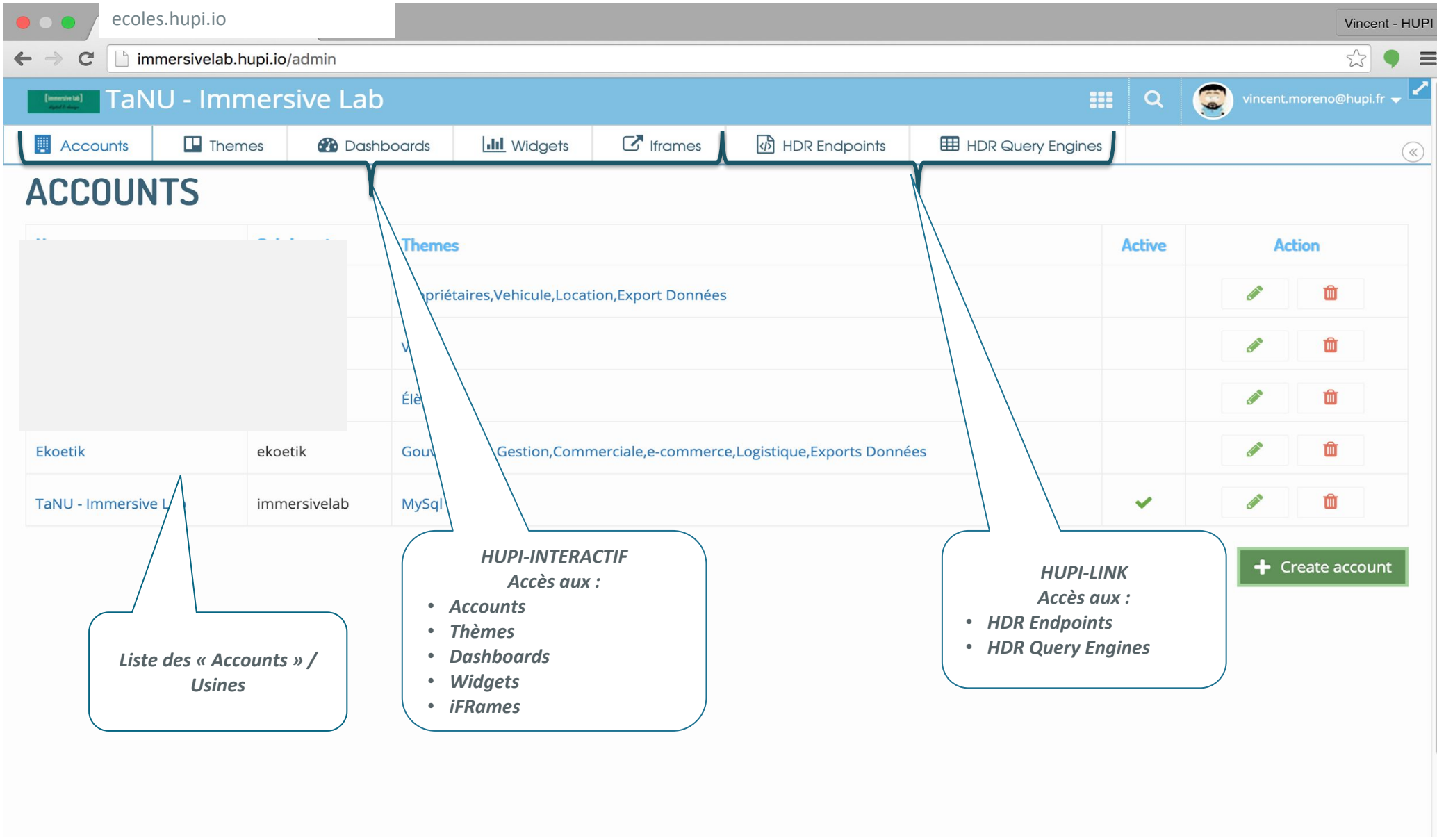
Vos URLS:

URL	Commentaires
https://ecoles.hupi.io	Votre portail, point d'entrée de toutes vos applications
https://ecoles-studio.hupi.io	Accès direct au studio
https://ecoles-terminal.hupi.io	Accès direct au Terminal (R, pyspark, python, shell..)
https://ecoles-notebook.hupi.io	Accès direct au Notebook (Spark/Scala)
https://api.dataretriever.hupi.io/private/ecoles	Votre API dynamique HUPI-LINK
ecoles.kafka.pro.hupi.loc	Vos messages en temps réel pour traitement
ecoles.mongo.pro.hupi.loc	Votre base de données ecoles
http://api.catchbox.hupi.io/v1/ecoles/<dynamic endpoint>	Votre API dynamique pour l'entrée de vos flux de données

FONCTIONNALITÉS - FRONT

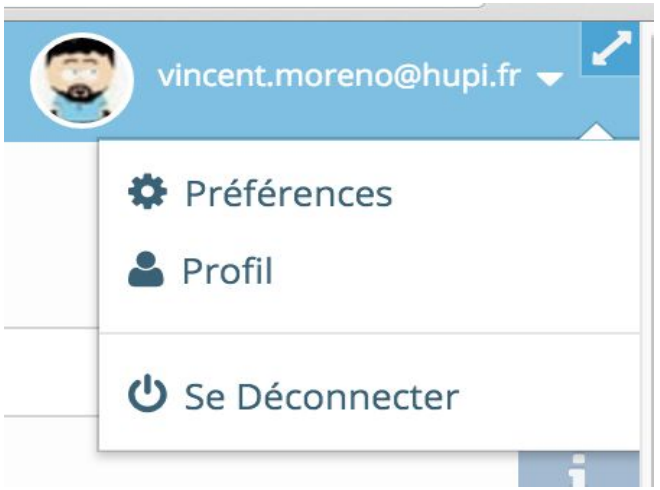


FONCTIONNALITÉS - ADMINISTRATION



FONCTIONNALITÉS - UTILISATEUR

- Gérer votre profil
- Gérer vos préférences
  - Décider quelle période vous intéresse le plus
  - Cette période sera appliquée par défaut dans tous les Dashboards



NIVEAU 1

ADMINISTRATION

- Gestion des Usines « Accounts »
  - Utilisateurs (profil viewer)
    - Accès à tous les Thèmes en mode lecture seule
    - Accès à toutes les options (export, zoom, filtres, profil,etc…)
    - Aucun accès aux applications
  - Administrateurs (profil administrateur)
    - Accès aux applications HUPI-INTERACTIF et HUPI-LINK en tant qu'administrateur
    - Autorise un utilisateur à devenir « Administrateur » et « Expert »
  - Experts (profil « Datascientist »)
    - Accès à toutes les solutions pour développer des recommendations en temps réel

Un expert peut travailler sur des usines différentes avec un seul compte. Formez une communauté

ADMINISTRATION - EXERCICES

- Inscrire un nouvel utilisateur
- Autoriser l'accès au nouvel utilisateur
  - Que se passe t'il ?
- Ajouter l'utilisateur comme Administrateur
  - Que se passe t'il ?
- Ajouter l'utilisateur comme Expert
  - Que se passe t'il ?
- Retirer l'utilisateur de votre Usine
- Supprimer le compte de l'utilisateur

NIVEAU 1

CRÉER VOTRE PREMIER THÈME

Objectifs:

- Comprendre les mécanismes et les interactions entre les Thèmes, les Dashboards et les Widgets
- Manipuler les différents objets
- Assimiler les étapes jusqu'à la publication

CRÉER VOTRE PREMIER THÈME

1. Créez votre Thème
2. Personnalisez votre Thème (icône, nom)
   - Pourquoi je ne le vois pas ?
3. Identifiez les Dashboards à votre disposition
4. Ajouter et organisez les Dashboards à votre Thème
   - Que deviens le Thème « MySQL » ?

Je supprime mon Thème car il n'est plus intéressant. Tous les Dashboards associés sont supprimés ! (mais pas les widgets).

Mon thème pourrait m'être demandé lors d'une prochaine présentation, je le dépublie mais le conserve au besoin. Je peux surprendre mes interlocuteurs par ma réactivité ☺

PAUSE ?

- Au retour de la pause, peut-être aurez-vous des questions ?
- Nous prendrons un instant ☺

NIVEAU 1

PARTAGEZ VOS WIDGETS

Objectifs:

- Apprendre à partager de manière publique un widget
- Identifier comment l'intégrer dans tous les types de sites Internet ou dans des applications Web
- Identifier les filtres possibles
- Jouer avec les intéractions

Exercice:

- Partager le widget

LES OS

Ordonner et placer vos widgets

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

TANU – IMMERSIVE LAB

Delete

Back to index

Name : Tables et Champs

Icon : fa fa-database

Theme : MySql

Skeleton : 50_50

Description :

Choisissez votre icône parmi plus de 150 icones

Widgets :

Gauche_5050

Liste des tables - csv

+add widget

Droite_5050

Liste des champs des tables - csv

+add widget

Filters :

Name	Slug	Type	
Table	table_name	text	

+add filter

Choisissez le template de votre page

75_25
25_75
✓ 50_50
50_50_footer
header_50_50
33_33_33

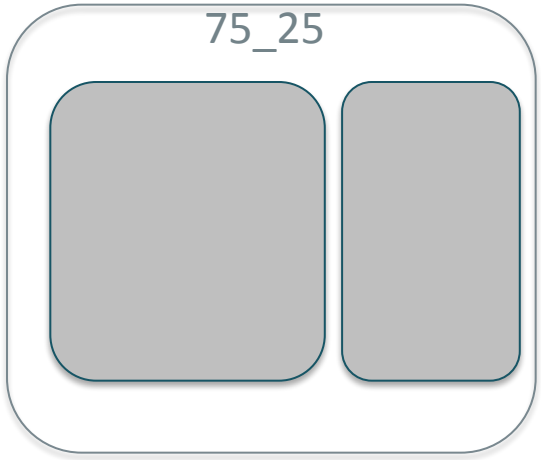
Ajouter des filtres à vos Dashboards, 3 types:

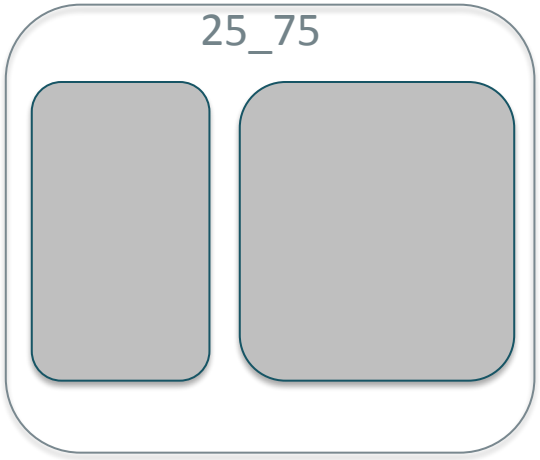
- *Text*
- *Number*
- *Date*

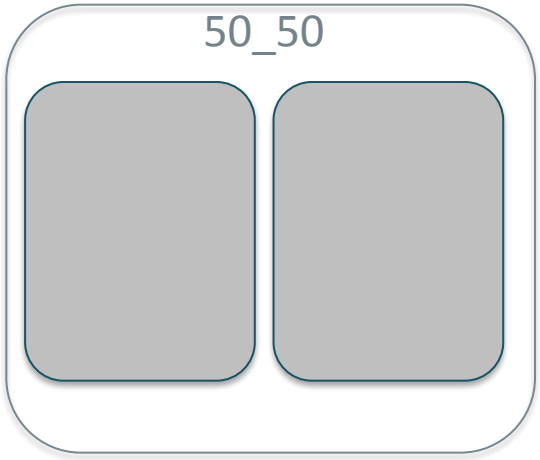
Les filtres sont appliqués à tous les Widgets qui l'acceptent.
Le « slug » des filtres sera utilisé dans les widgets pour faire le lien.

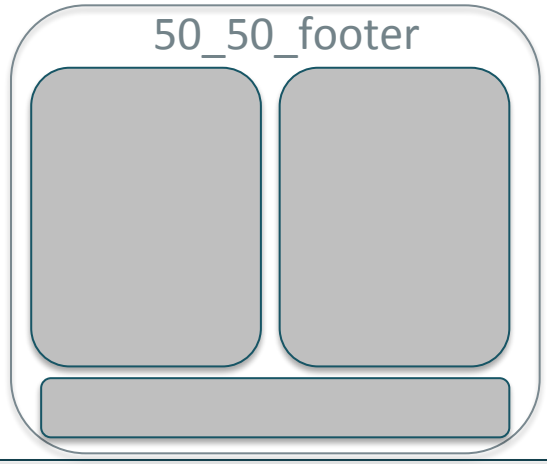
75_25
25_75
✓ 50_50
50_50_footer
header_50_50
33_33_33

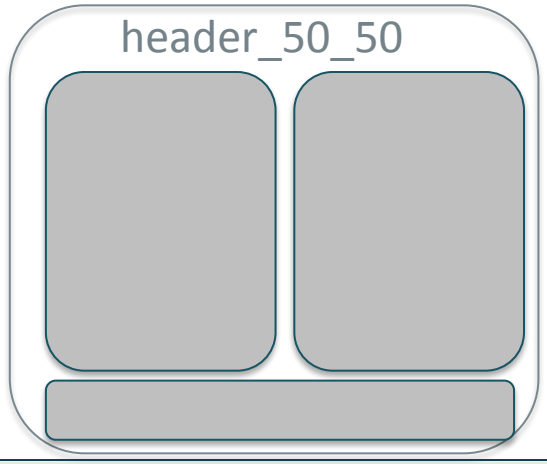
DASHBOARDS SKELETON

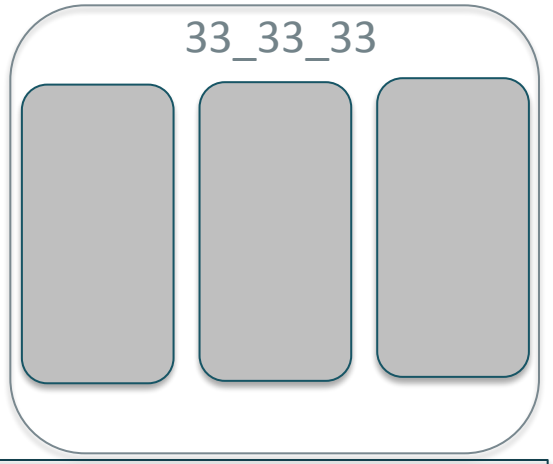












Je modifie un skeleton, je dois remettre les Widgets

NIVEAU 2

LES WIDGETS

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

TANU - IMMERSIVE LAB

Back to index

Name : Répartition des questions

Description :

URI : immersivelab / questions_categories

Render Type : column_stacked_normal

Cache Expiration :

Widget Options :

```
{
  "title" : {"text": "Questions
mises à jour",
          "x": -20 },
  "colors": ["#97C000",
"#E3A606", "#F7A708",
"#F20529", "#BBCB07",
"#BDCD06", "#DE6E07",
```

Afficher une description détaillée des informations
Insérer des URL pour dynamiser

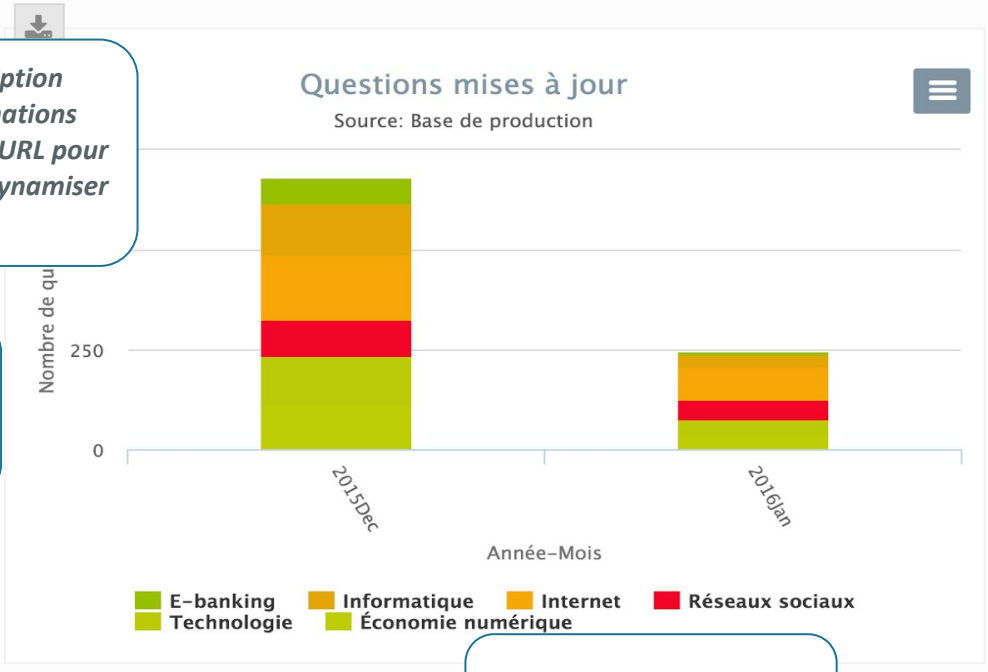
Nom de la Datavisualization

Temps en seconde pour rafraichir vos Widgets

Imaginez et personnalisez vos widgets

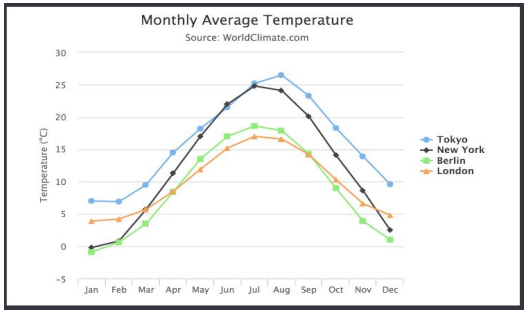
widget preview

Testez vos widgets

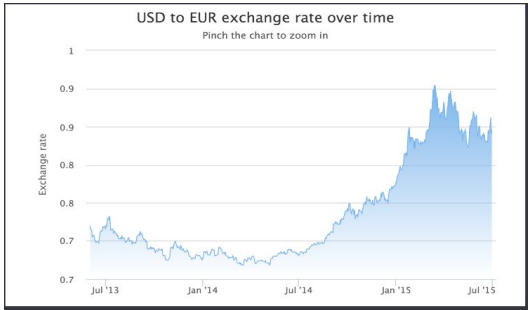


HUPI

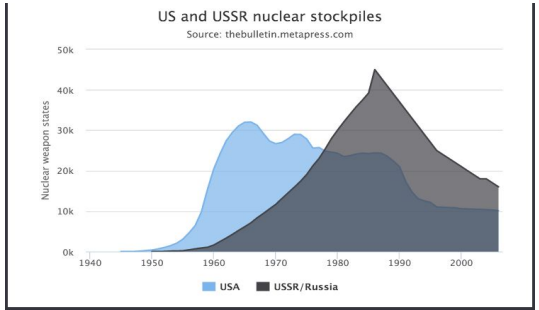
WIDGETS
RENDER_TYPE



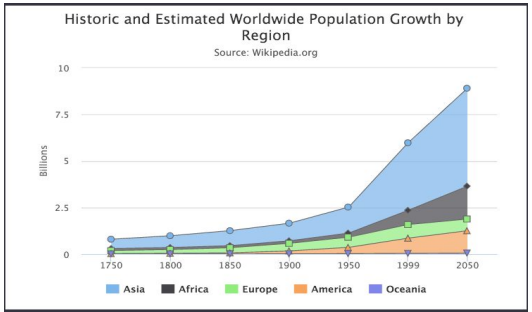
Basic line



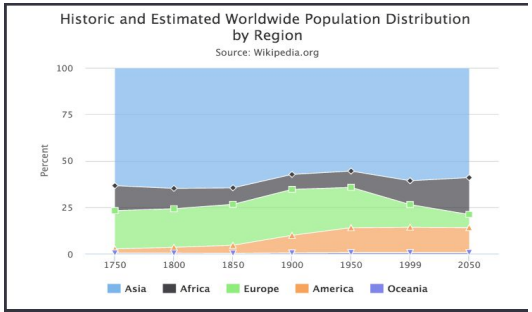
Time series, zoomable



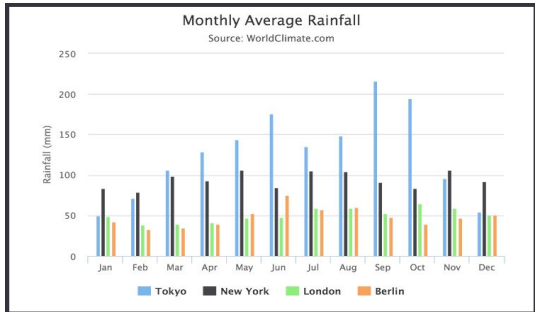
Basic area



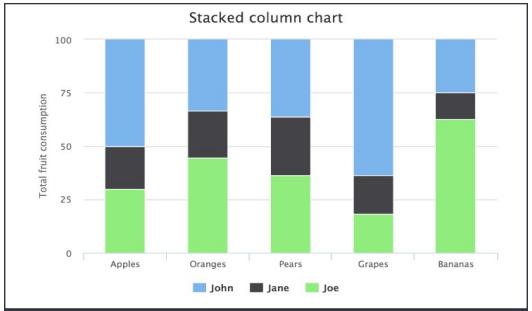
Stacked area



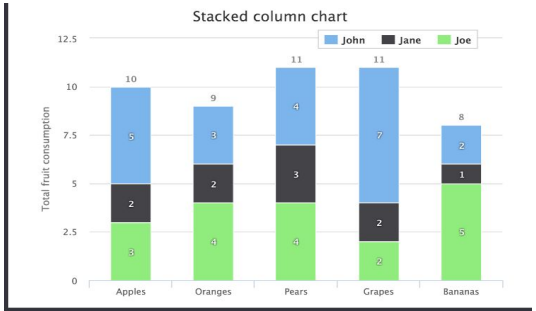
Percentage area



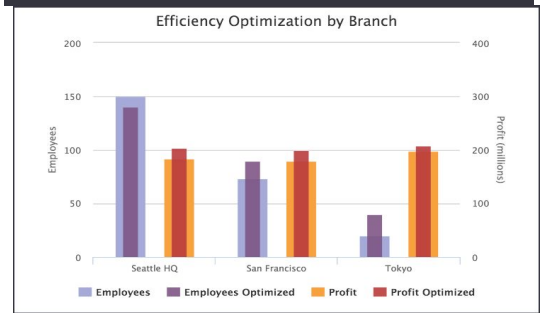
Basic column



Stacked percentage column

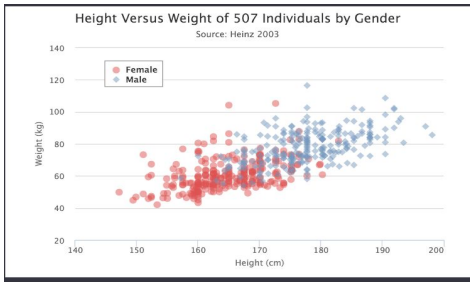


Stacked column



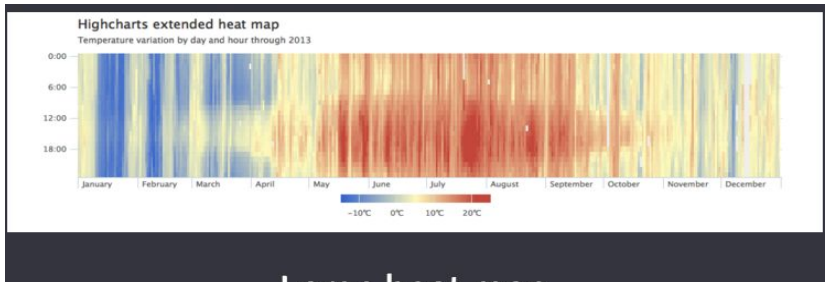
Fixed placement columns

WIDGETS
RENDER_TYPE

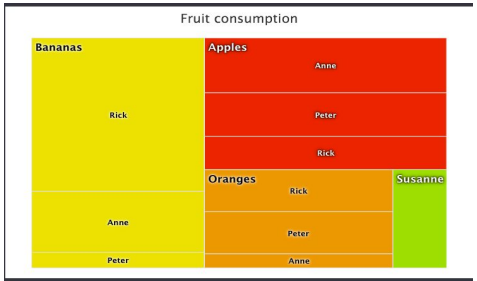


Scatter plot

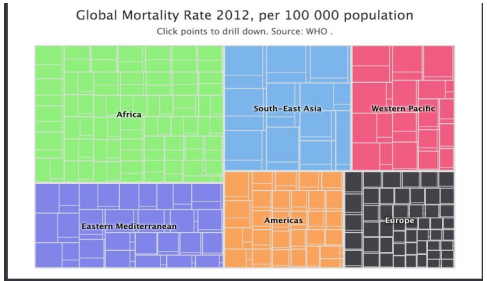
Heat map



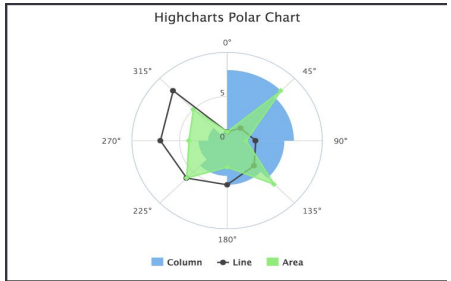
Large heat map



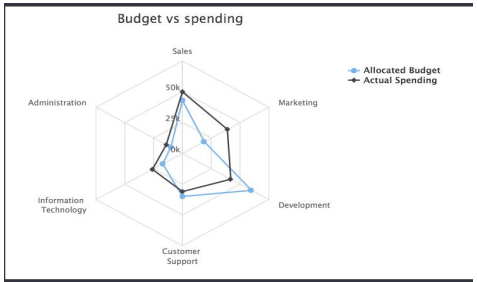
Tree map with levels



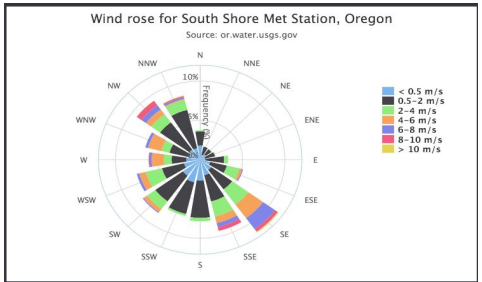
Large tree map



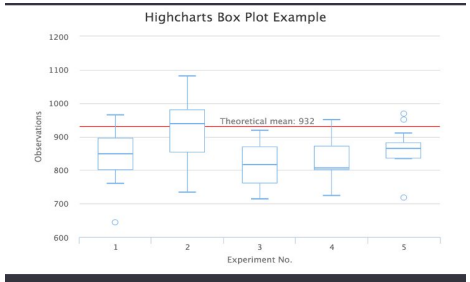
Polar chart



Spiderweb

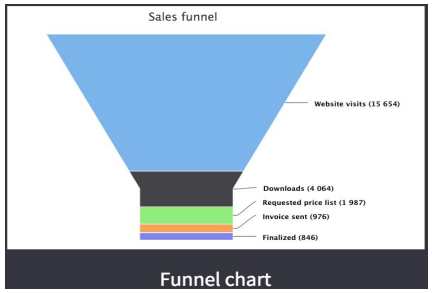


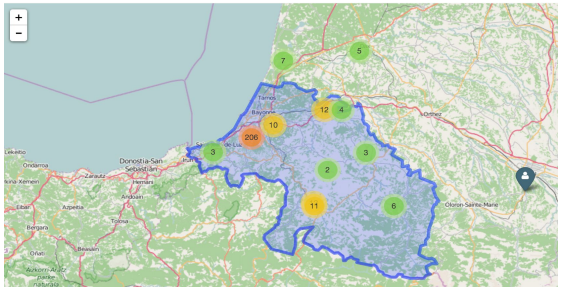
Wind rose



Box plot

WIDGETS
RENDER_TYPE





HUPI intègre au fil de l'eau des Datavisualizations de HighCharts et 3D.js

NIVEAU 2

WIDGETS

Objectifs:

- Construire un widget
- Intégrer les données du HUPI-LINK
- Afficher les résultats dans différents « render-type »

« endpoint » : Nom logique pour accéder aux données

HUPI

NIVEAU 2

WIDGETS

Exercice:

- Créer un nouveau widget basé sur un « endpoint »
- Tester le widget
- Affecter le widget à un Dashboard existant ou un nouveau
- Visualiser le widget
- Personnalisez votre Dashboard

NIVEAU 3

WIDGETS CUSTOMIZATION

```
{
  "title" : {"text": "Questions mises à jour",
          "x": -20 },
  "colors": ["#97C000", "#E3A606", "#F7A708", "#F20529", "#BBCB07", "#BDCD06", "#DE6E07", "#63A12E",
"#242524"] ,
  "legend" : { "enabled": true },
  "yAxis": {  "min": 0,
              "title": {"text": " Nombre de questions"}
            },
  "xAxis": {  "min": 0,
              "title": {"text": "Année-Mois"}
            },
"subtitle": {
          "text": "Source: Base de production",
          "x": -20
      },
"tooltip": {
          "valueSuffix" : " ",
          "headerFormat": "<span style=font-size:15px>{point.key}</span><table>",
          "pointFormat": "<tr><td style=color:{series.color};padding:0>{series.name}: </td> <td
style=padding:0><b>{point.y:.0f} TaNU</b></td></tr>",
          "footerFormat": "</table>",
          "shared": true,
          "useHTML": true
      },
"plotOptions": {
          "column": {
              "pointPadding": 0.2,
              "borderWidth": 0
          }
      },
 "chart": {

          "zoomType": "x",
          "panning": true,
          "panKey": "shift"
      }
}
```

Je crée un widget « caché » et je conserve ces options pour les copie/coller, je gagne du temps mais aussi j'assure une cohérence graphique à mes Dashboards

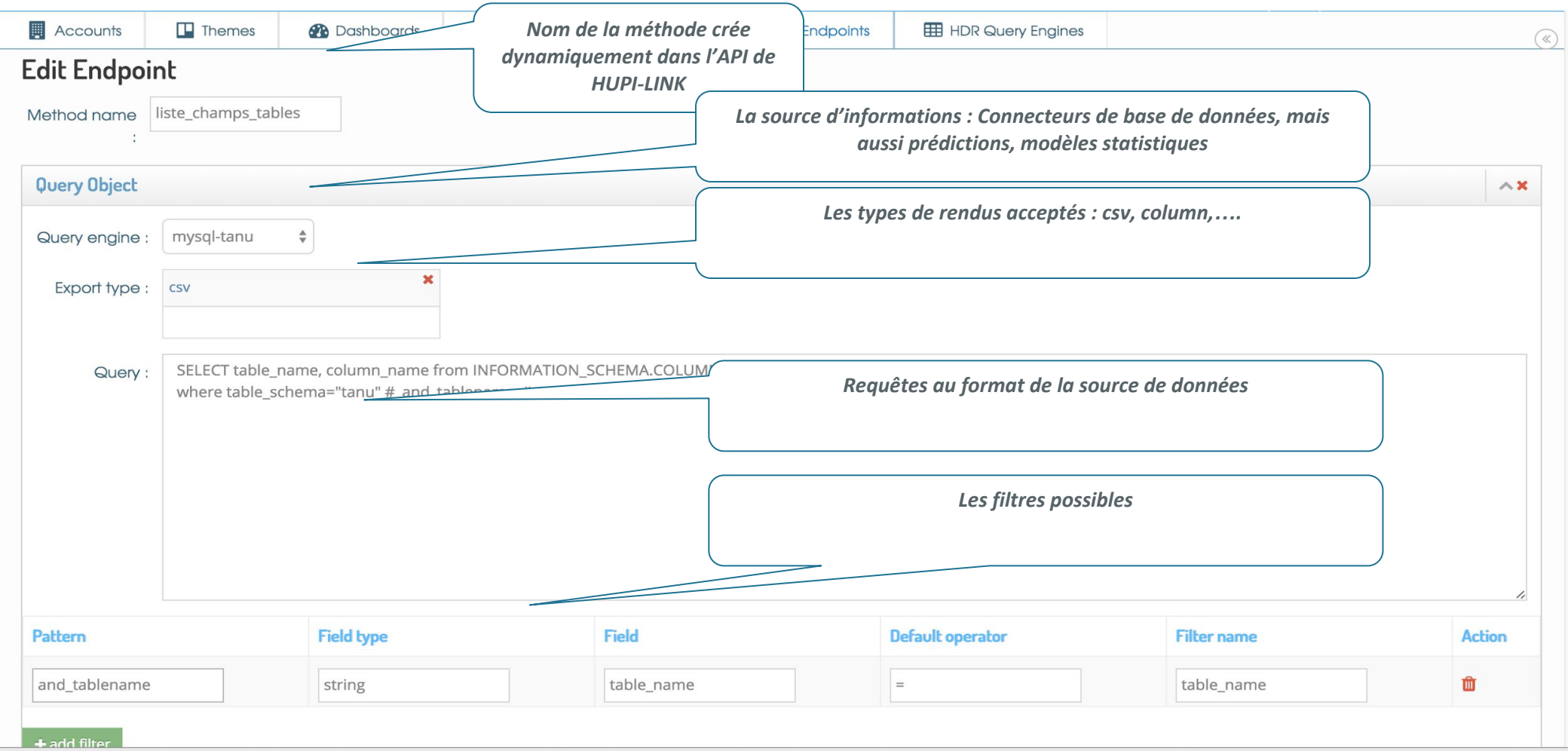
Retrouver toutes les options possibles : http://api.highcharts.com/highcharts

HUPI

NIVEAU 3

HUPI-LINK

ACCÈS AUX DONNÉES – PARTAGE SYSTÈMES EXPERTS



HUPI

- Exemples de requêtes Mongo

```
{
    "collection": "number_visits_perNumberVisitors",
    "query": [{
      "operator": "aggregate",
      "pipeline": [#_match_f1_#{
                        "$group": {
                                "_id":"$number_visits",
                                "value": {
                                        "$sum": "$number_visitors"
                                }
                        }
                },{
      "$project": {
        "category": "$_id",
        "value": "$value"
      }
    }, {
      "$sort": {
        "category":1
      }
    }]
  }]
}
```

Renommer les variables pour leur prise en compte dans les widgets

Quels sont les règles et formats à suivre

https://github.com/hupi-analytics/data-retriever

Pattern	Field type	Field	Default operator	Filter name
match_f1	int	event_timestamp	$gte	start_date
match_f1	int	event_timestamp	$lte	end_date

Les filtres Date par défaut :

- start_date
- end_date

#_where_[f1]_#
#_and_[f2]_#
#_match_[f4]_#

NIVEAU 3

HUPI-LINK : ACCÈS SYSTÈMES EXTERNES

- Comment tester si le service est opérationnel ?

```
curl -H "Content-Type: application/json" -H "Accept-Version: v1" -H "X-API-Token: my-private-token1"
-X POST -d '{ "client": "ecoles", "render_type": "csv", "filters": { "question_id": "2485" } }'
http://api.dataretriever.hupi.io/private/ecoles/???????
```

https://github.com/hupi-analytics/data-retriever

Utiliser l'outil POSTMAN, très pratique pour analyser vos API et vérifier les résultats

HUPI

HUPI

valoriser analytiquement votre différence

FORMATION

HUPI-INTERACTIF & HUPI-LINK & HUPI-ANALYTICS

2ème Journée

JOURNÉE 2

- Jour 2: Matinée
  - Rappels
  - Dashboards experts
  - API expert
  - Cas Exercices pratiques
- Jour 2: Après-midi
  - Présentation HUPI-STUDIO
  - Présentation HUPI-NOTEBOOK

RAPPELS

Vos retours:

- Les Plus ?
- Quels sujets souhaitez-vous que nous repassions plus en détails ?
- Quels sont les nouveaux services que vous imaginez ?

NIVEAU 1

RAPPELS

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

- ACCOUNTS:
  - Votre usine, en tant qu'administrateur, vous pouvez administrer plusieurs usines
  - Vous habilitez tous les utilisateurs qui se sont déjà enregistrés à être « Administrateur » ou « Viewer »
  - Vous définissez le nom de votre usine
  - Vous choisissez les thèmes et les ordonnez

RAPPELS

- THEMES:
  - Vous définissez de nouveaux thèmes
  - Vous leur associez un icône reconnaissable
  - Vous ordonnez les Dashboards
  - Vous partagez vos Thèmes avec d'autres usines

NIVEAU 1

RAPPELS

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

- Dashboards:
  - Vous créer des dashboards et leur associez un icône
  - Vous choisissez les « skeleton »
  - Vous ajoutez une description, visible pour les utilisateurs
  - Vous ajoutez tous les filtres que vous souhaitez
  - Vous ordonnez vos widgets

- WIDGETS:
  - Vous créer des widgets basés sur des « endpoints » [votre API privé]
  - Vous pouvez ajouter des descriptions « i » affiché sur chaque widget. *Pensez à mettre des liens* et décrire l'information
  - Choisissez votre « render type » ou encore le type de gra  que vous souhaitez
  - Passez des paramètres aux API
  - Personnalisez vos widgets
  - Donnez la possibilité de télécharger les données

Ne pas oublier l'export type « csv »

- IFRAMES:
  - Choisissez quels widgets vous souhaitez partager avec vos collègues, intégrez les dans vos sites pour les rendre dynamiques, envoyez les par emails pour échanger
  - Pensez à utiliser les filtres associés (cliquer sur le bouton de partage pour voir comment utiliser les filtres)

NIVEAU 1

RAPPELS

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

- HDR Endpoints:
  - Définissez un nouvelle API (un nouveau service dynamique d'accès à des données)
  - Choisissez le type d'export de vos données
  - Voir la documentation pour le type des exports

  - L'export type « csv » est le plus simple pour tester vos données
  - Intégrez vos filtres personnalisés (venant des Dashboards ou de vos appels API)

https://github.com/hupi-analytics/data-retriever

HUPI

NIVEAU 1

RAPPELS

Accounts | Themes | Dashboards | Widgets | Iframes | HDR Endpoints | HDR Query Engines

- HDR Query Engines:
  - Ajoutez vos propres accès à vos données sur vos bases externes à votre usine
  - Par défaut, les « Engines » internes à votre usine ne sont pas présentés:
    - MongoDB
    - Impala
    - Prediction
    - Search

Nouveaux services pour les intégrer dans vos systèmes experts

Nomenclatures:

- Bien nommer les widgets pour les retrouver facilement
  - Par catégories numérotées, exemple :
    - 1000 – Economie
      - 1100 – Chiffres d'affaires
    - 2000 – Produits
      - Etc.…
  - Par nom logique combinés, exemple:
    - Economie_chiffres_d_affaires
  - Par rédacteurs
    - Arnaud_economie_chiffre_d_affaires

NIVEAU 2

DASHBOARDS EXPERTS

- Optimisez vos temps de réponse:
  - Utiliser la fonction « cache_expiration »
- Comment intégrez de nouvelles sources de données
  - Impala
  - Mysql
  - Postgres
  - Search
  - prediction

Header Params
Content-Type: "application/json"
Accept-Version: "v1"
X-API-Token: "your_token »

POST http://api.dataretriever.hupi.io/private/(:module_name)/(:method_name)

module_name: (String)
method_name: (String)
body:

```
{
  "client": client_name,
  "render_type": render_type,
  "filters": {
    filter_name1: { "operator": operator, "value": value},
    filter_name2: value
  }
}
```

PAUSE ?

- Au retour de la pause, peut-être aurez-vous des questions ?
- Nous prendrons un instant ☺

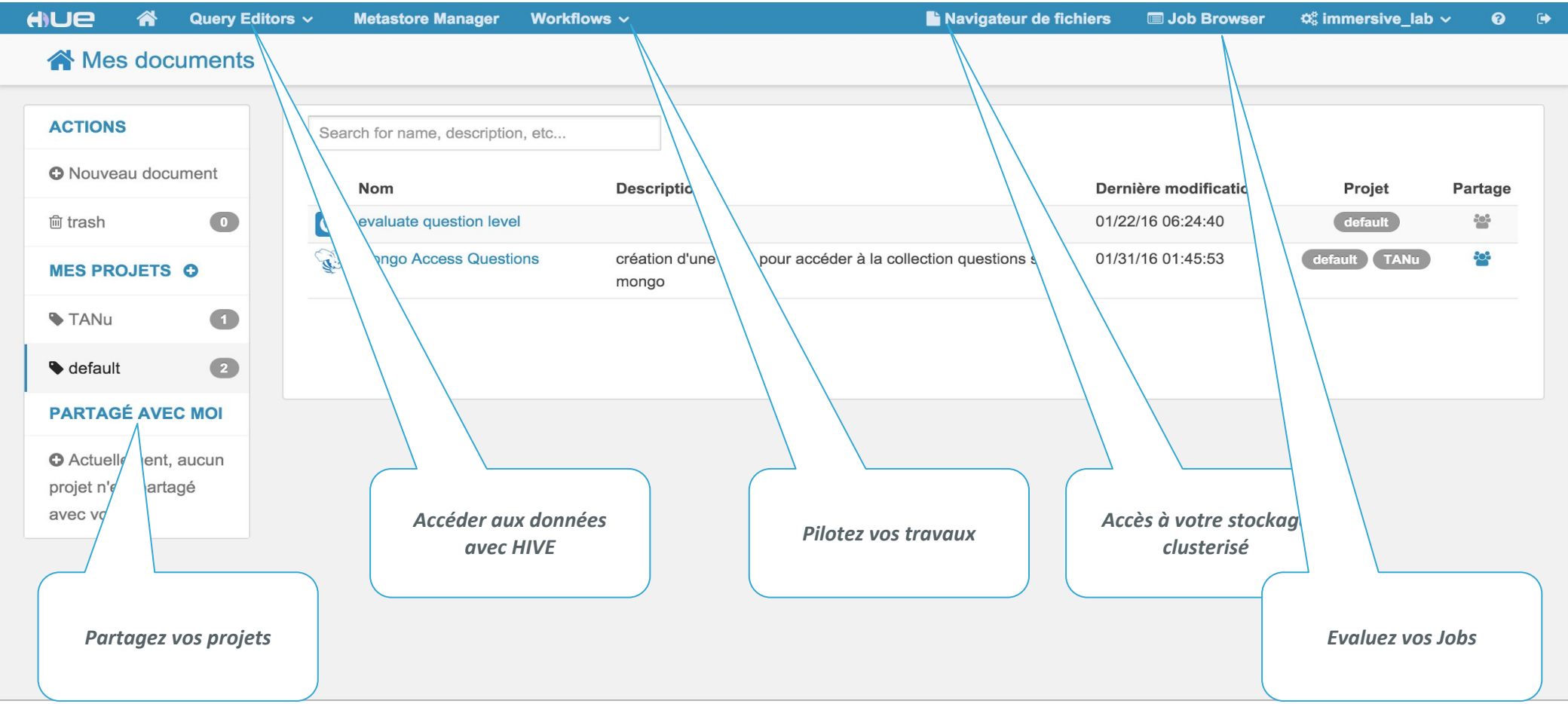
NIVEAU 2

HUPI - STUDIO

- Le HUPI-STUDIO est un environnement dédié qui vous permet :
  - D'importer vos données structurées, semi-structurées, non-structurées – HDFS
  - Exploiter en SQL vos données
  - Ordonnancez vos batchs
  - Partagez vos projets

NIVEAU 2

HUPI - STUDIO

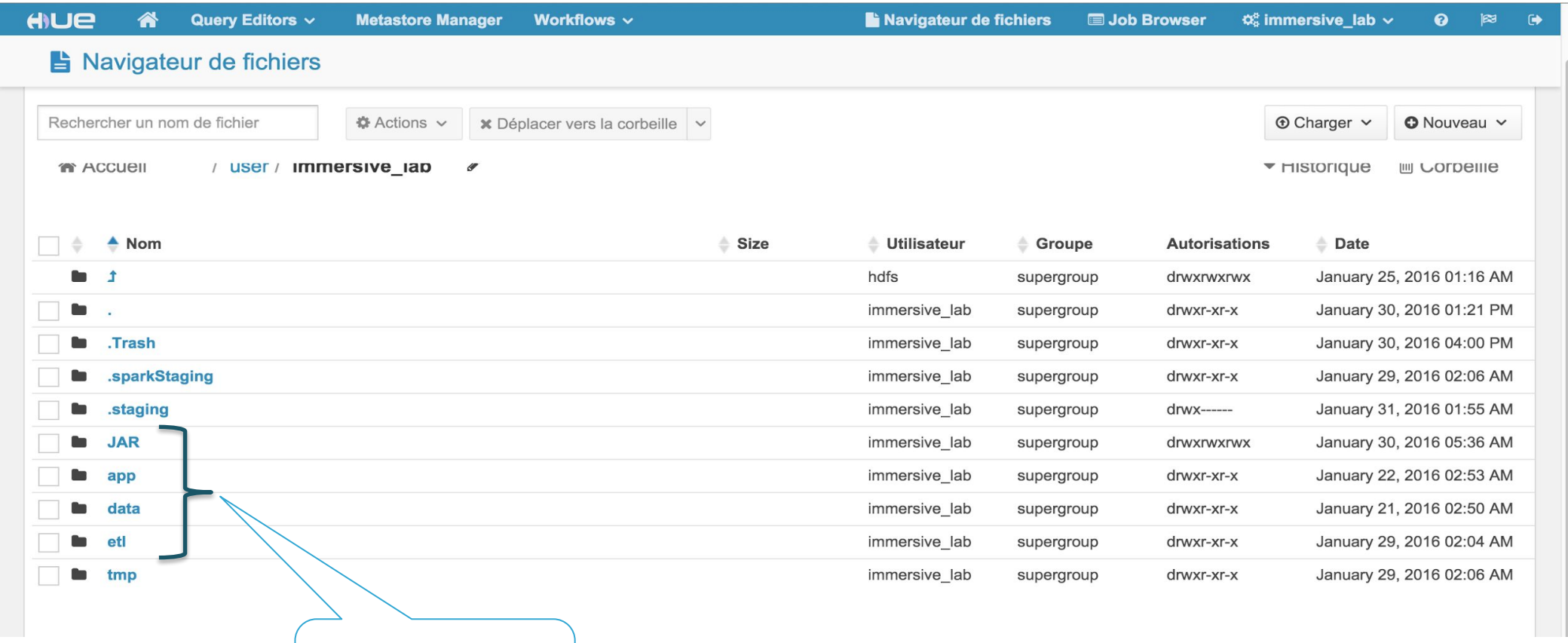


HUPI

NIVEAU 2

HUPI – STUDIO

NAVIGATEUR DE FICHIERS



NIVEAU 2

HUPI – STUDIO

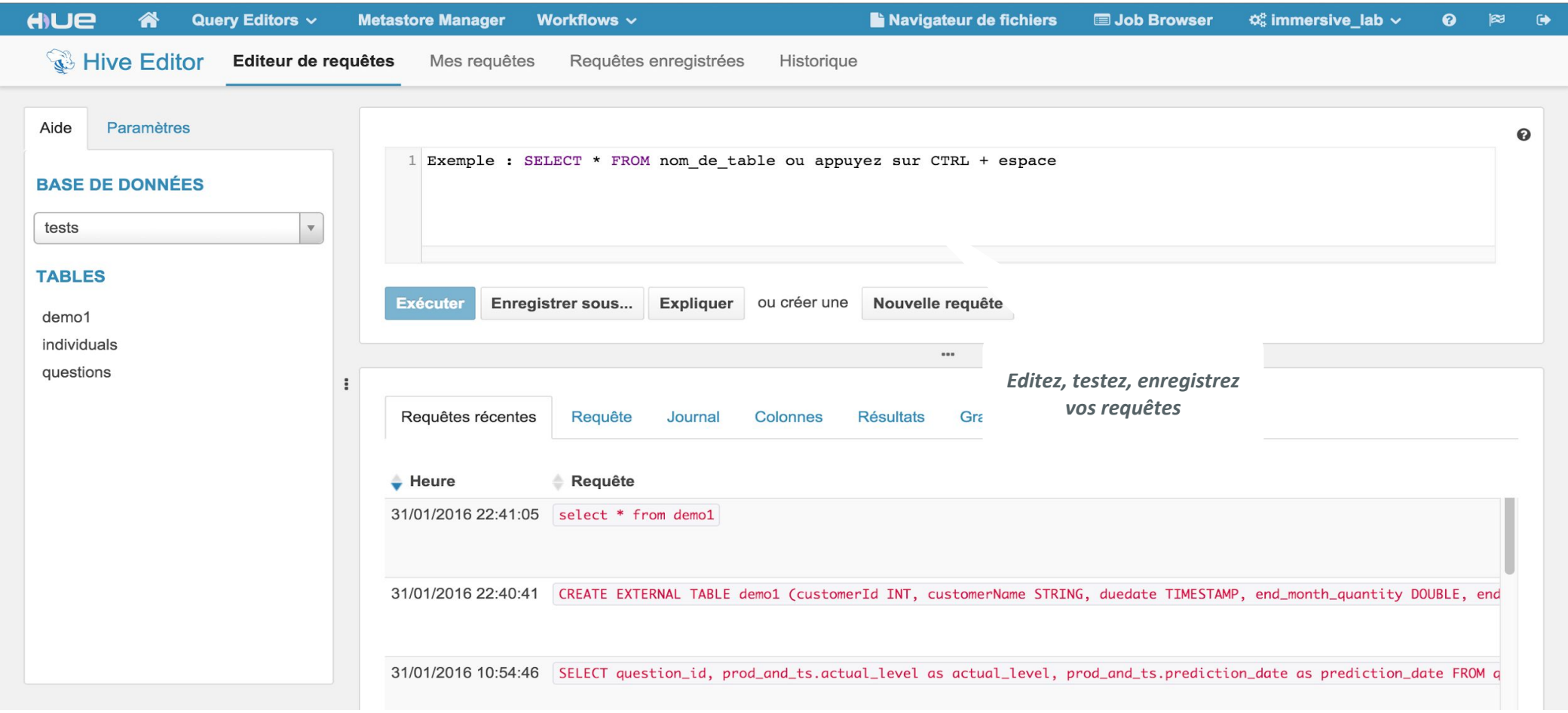
NAVIGATEUR DE FICHIERS

Exercices:

- Ajouter des fichiers de données
- Modifier les autorisations
- Afficher les informations contenues

HUPI – STUDIO

QUERY EDITORS



- Les moteurs d'accès aux données sont bien déconnectés du stockage

Langage: HQL (proche du SQL)

```
CREATE EXTERNAL TABLE demo1 (customerId INT,
customerName STRING, duedate TIMESTAMP,
end_month_quantity DOUBLE, end_month_price
DOUBLE, start_month_quantity DOUBLE,
start_month_price DOUBLE)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/user/ecoles/Formation/Data';
```

Vous pouvez importer tout type de fichiers : CSV/ TSV / JSON / XML VIDEO /

- Les moteurs d'accès aux données sont bien déconnectés du stockage

Langage: HQL (proche du SQL)


```
CREATE TABLE click_data
(
  pid INT,
  vid STRING,
  score INT,
  uid INT,
  v_index INT
)
STORED BY 'com.mongodb.hadoop.hive.MongoStorageHandler'
WITH SERDEPROPERTIES('mongo.columns.mapping'='{"id":"_id"}')
TBLPROPERTIES('mongo.uri'='mongodb://ecoles.node1.pro.hupi.loc:2
7017/test.click_data');
```

Vous pouvez importer tout type de fichiers : CSV/ TSV / JSON / XML VIDEO /

Exercices:

- Ajouter un fichier au format « TSV » ou « CSV »
- Créer une table dans la table TEST avec HIVE
- Visualisez les informations
- Créer une nouvelle API (Endpoint)
- Ajouter un Dashboard

Job Browser

Nom d'utilisateur immersive_lab Texte Rechercher un texte

Réussi | En cours d'exécution | Echoué | Détruit

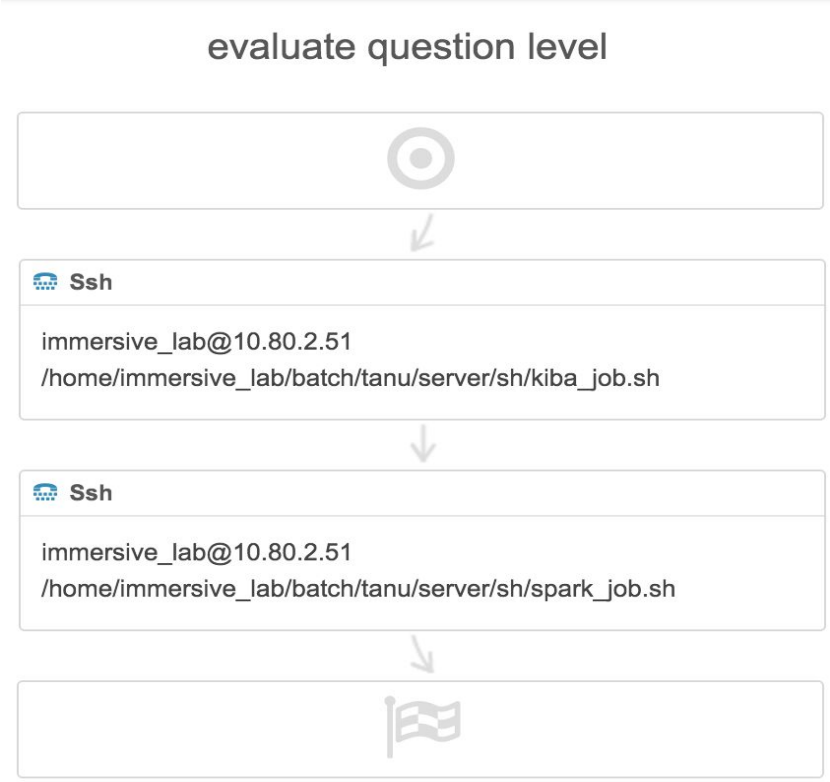
Journaux	ID	Nom	Application Type	Statut	Utilisateur	Maps	Reduces	File d'attente	Priorité	Durée	Envoyé
	1453413556154_0029	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	20s	01/31/16 01:54:47
	1453413556154_0028	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	18s	01/30/16 14:17:45
	1453413556154_0027	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	18s	01/30/16 14:02:43
	1453413556154_0026	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	18s	01/30/16 14:01:46
	1453413556154_0025	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	19s	01/30/16 14:00:50
	1453413556154_0024	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	17s	01/30/16 13:59:26
	1453413556154_0023	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	17s	01/30/16 13:57:16
	1453413556154_0022	SELECT question_id, prod...prod_and_ts(Stage-1)	MAPREDUCE	SUCCEEDED	immersive_lab	100%	100%	root.immersive_lab	N/A	18s	01/30/16 13:52:43

NIVEAU 2

HUPI – STUDIO

ORDONNANCEUR - OOZIE

Lancez manuellement ou ordonnancez vos traitements



Créer vos scripts shell et lancer les avec votre ordonnanceur
Créer vos programmes spark (pyspark/scala) et ordonnancez les
Activer des SLA (Service Level Agreement)

Exercices:

- Visualisez le traitement
- Lancer manuellement le traitement
- Planifier le traitement
- Créer un nouveau workflow

HUPI - NOTEBOOK

- Environnement interactif dédié au développement de calculs statistiques distribués
- Partagez vos projets
- Tester des modèles statistiques

NIVEAU 3

HUPI - NOTEBOOK

hupi **SPARK NOTEBOOK**

Files | Running | Clusters

To import a notebook, drag the file to the listing below or **click here.**

New

Process tournant dans votre usine

- ..
- adam
- anomalyDetection
- cassandra
- core
- graphx
- machine-learning
- misc
- mllib
- sql
- streaming
- tachyon
- viz

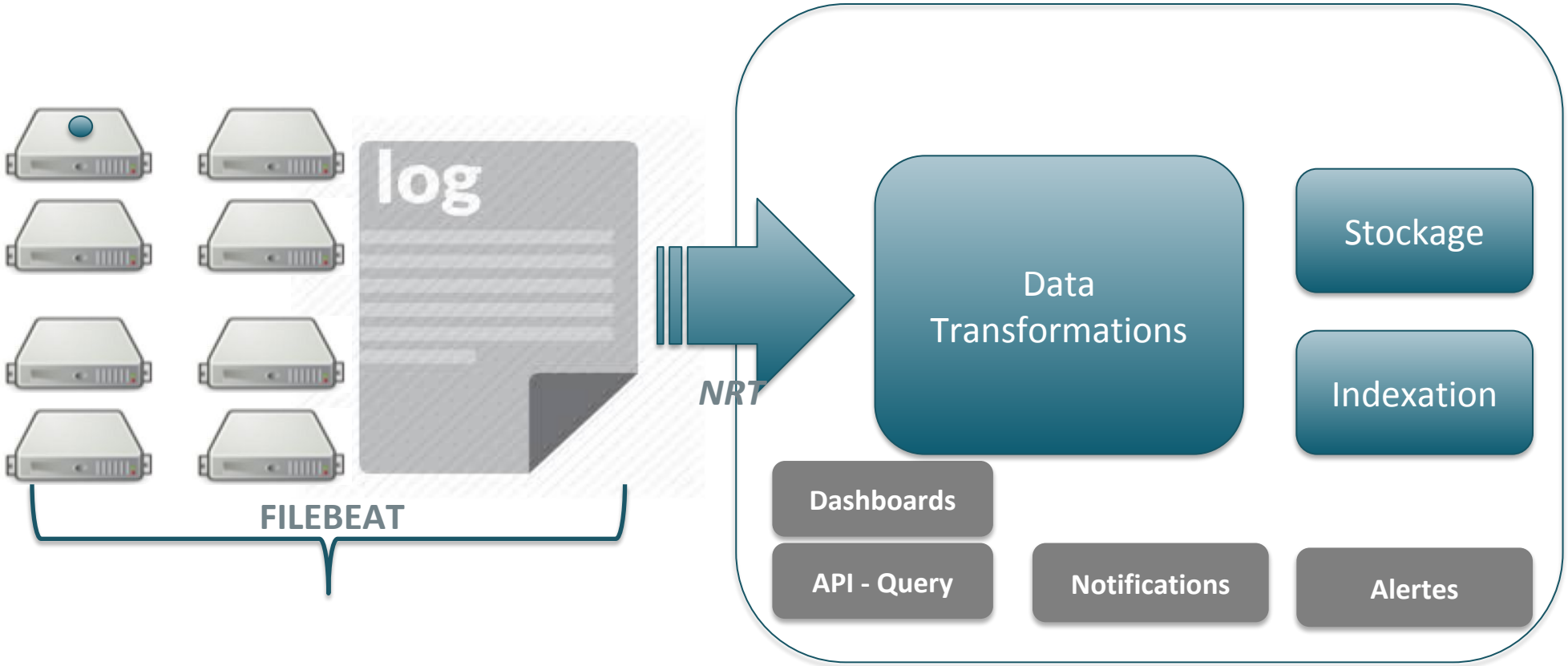
SCÉNARIOS DE FLUX DE DONNÉES

- HUPILYTICS
  - Evaluez la performance de votre site ou de vos applications web
- Logs
  - Applicatifs
  - Systèmes

https://github.com/hupi-analytics/hupilytics

HUPI vous accompagne dans la connexion de vos logs et la mise en place des flux

SCÉNARIOS DE FLUX DE DONNÉES



DÉMONSTRATION DE FLUX

- Affichage des flux entrants
- Affichage du stockage
- Visualisation des traitements

HUPI

valoriser analytiquement votre différence

FORMATION

HUPI-ANALYTICS

3ème Journée

AGENDA

- Introduction au Machine Learning
- Basics
- Classification
- Clustering
- Regression
- Use-Cases
- Exercices (SPARK MLIB)

A QUOI ÇA SERT

Les principaux scénarios fonctionnels qui peuvent être adressés avec du Machine Learning

Analyse de désabonnements (churn)	Ciblage publicitaire	Détection d'images et classification
Monitoring d'équipements	Recommandations	Prévisions
Spam filtering	Détection de fraudes	Détection d'anomalies

QU'EST CE QUE LE MACHINE LEARNING

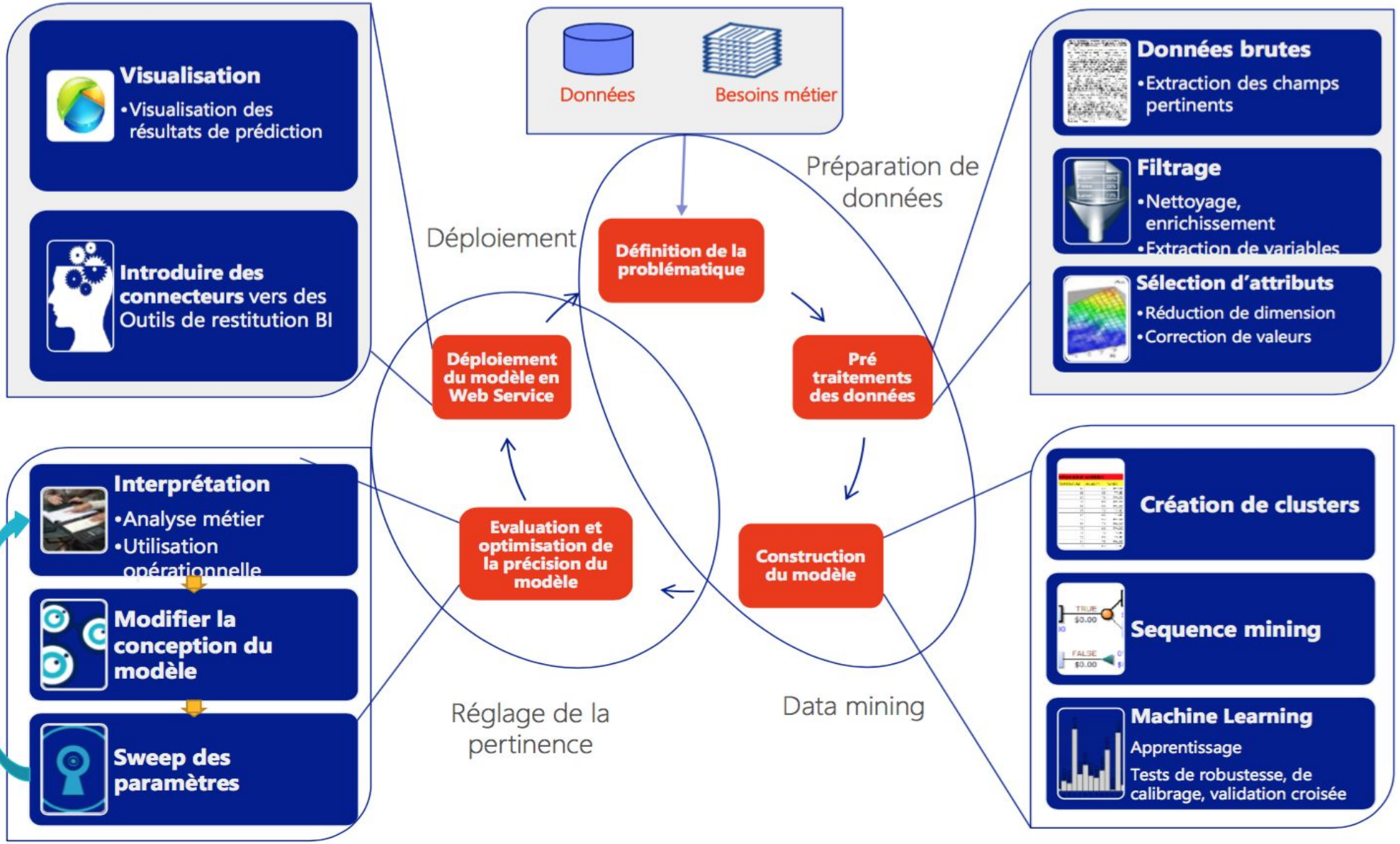
- Approximation d'un pattern à partir d'exemples
- Algorithme qui s'adpate
- Prédiction de l'avenir (proche) en fonction du passé

Le Machine Learning à pour but de prédire et le Data Mining a pour but de comprendre

CRITERES POUR UNE BONNE PROBLEMATIQUE DE ML

- Disposer d'exemples de résultats attendus (label)
  - Désabonnement client : enregistrement de clients actuels (loyaux) + clients ayant quittés (churn)
- De variables pertinentes
  - Informations clients : age, sexe, code postal, historique des achats
- Et accepter un certain niveau d'incertitude, de tolérance à l'erreur
  - Réussir à identifier certains clients mais pas tous

PROCESSUS D'UNE ITERATION



Avançons…..

Que voulons-nous dire par

APPLE

LEARNING (TRAINING)

Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc…

Features:
1. Sky Blue
2. **Logo**
3. Shape
etc…

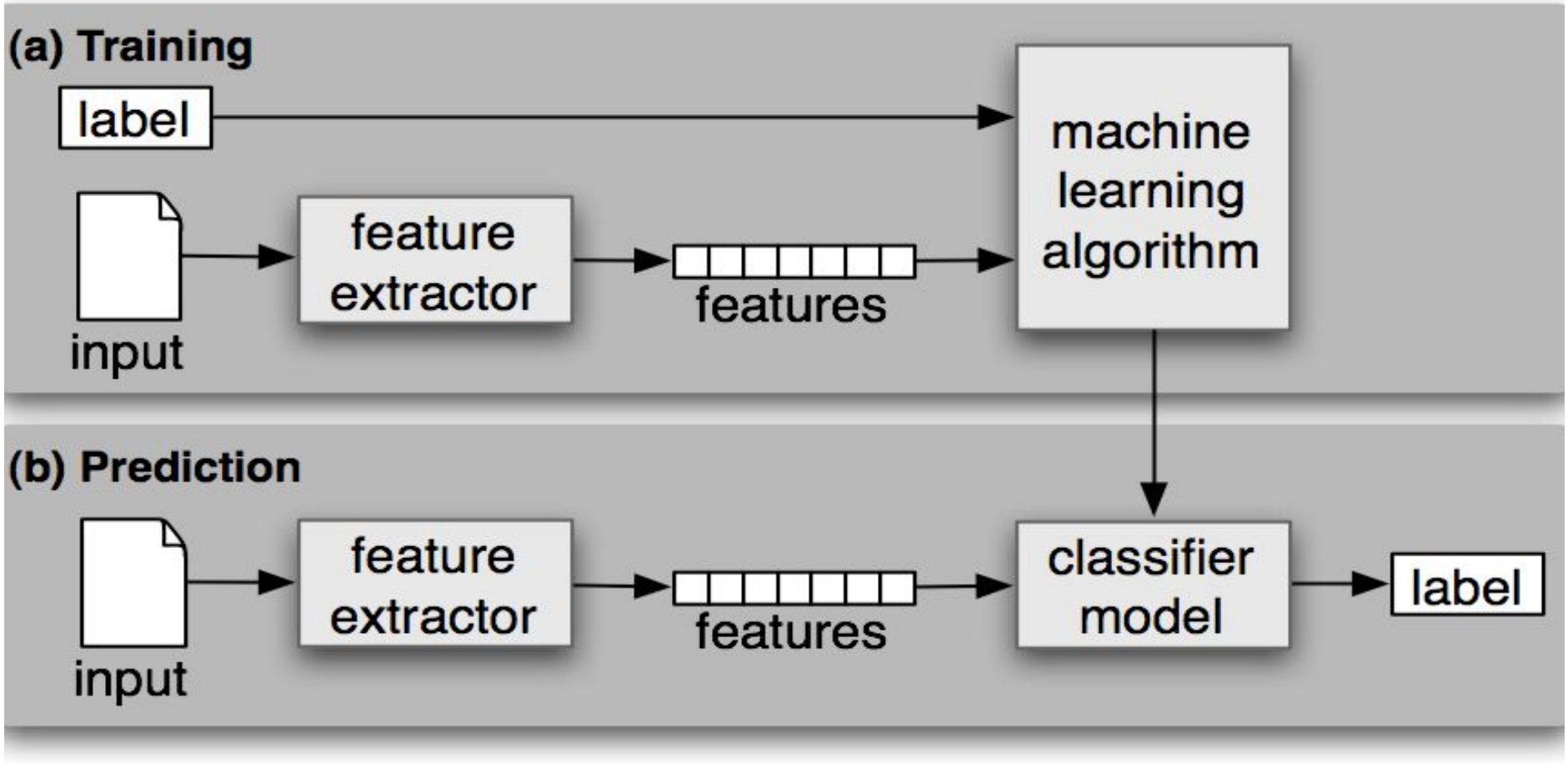
Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc…

TERMINOLOGIE

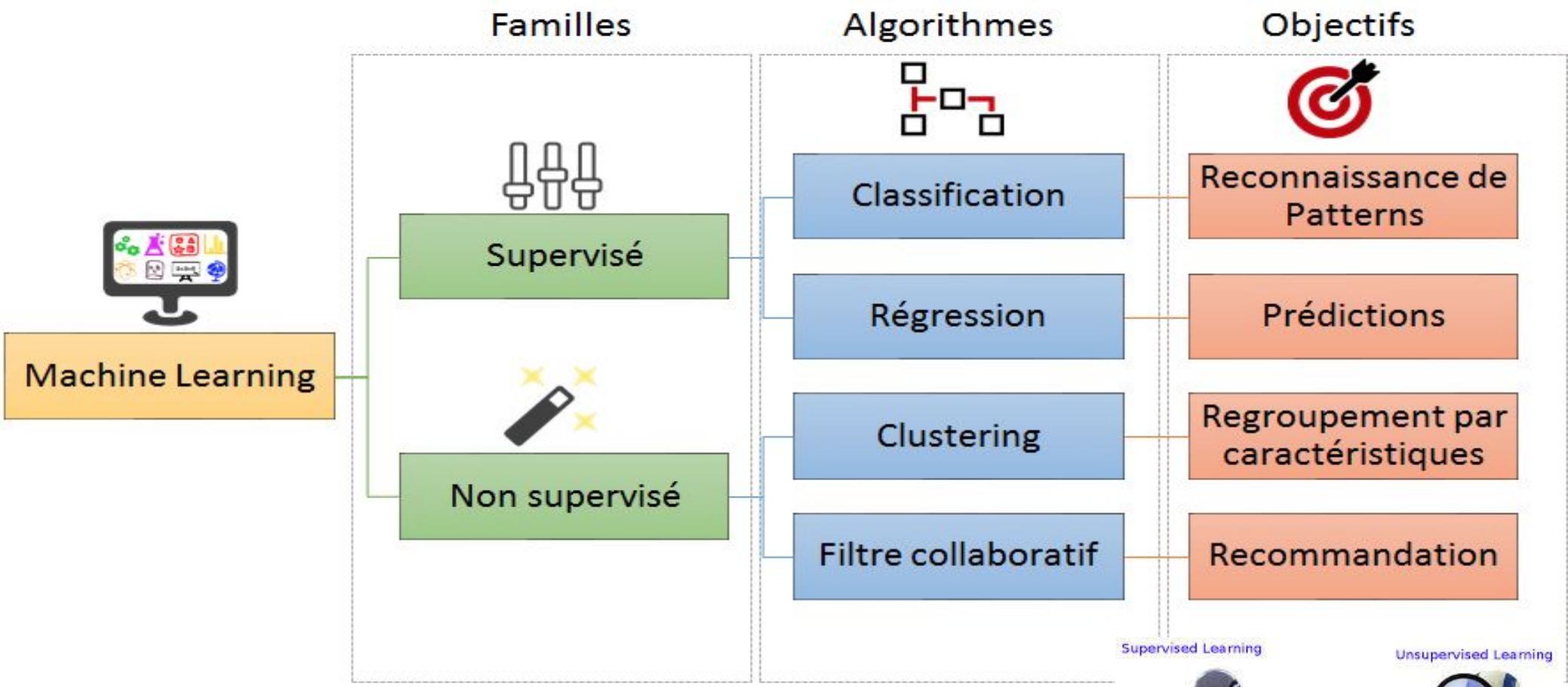
- Features
  - Le nombre d'informations ou traits distincts qui peuvent être utilisés pour décrire chaque élément d'une manière quantitative

  .
- Samples
  - Un échantillon est un élément à traiter (par exemple classer). Il peut être un document , une image, un son , une vidéo , une ligne dans le fichier de base de données ou CSV , ou ce que vous pouvez décrire avec un ensemble fixe de caractères quantitatifs .
- Feature vector
  - est un vecteur à n dimensions de caractéristiques numériques qui représentent un objet
- Feature extraction
  - Préparation d'un "feature vector"
  - Transformer la donnée avec ses dimensions dans un espace avec moins de dimensions
- Training/Evolution set
  - Jeux de données pour découvrir potentiellement des relations prédictives

WORKFLOW


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

FAMILLES

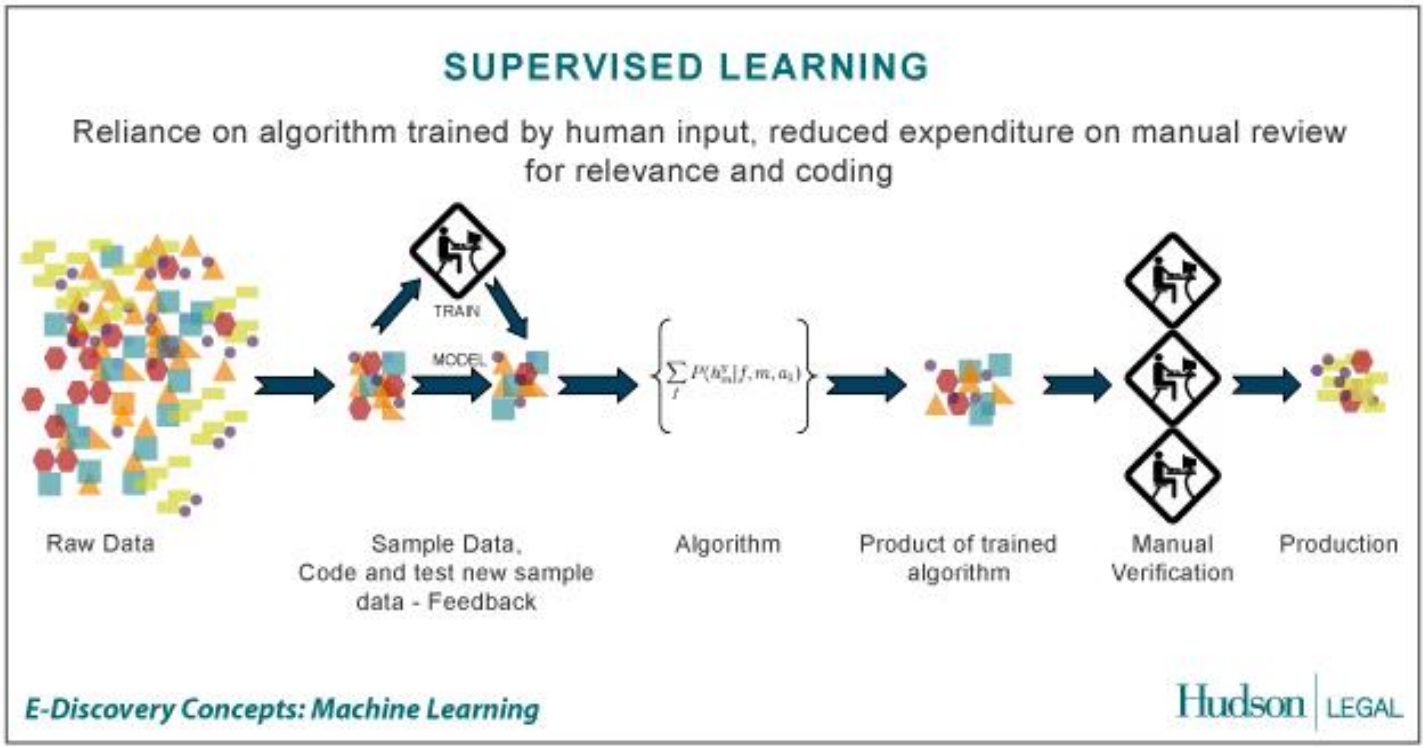

Familles
Algorithmes
Objectifs
Machine Learning
Supervisé
Non supervisé
Classification
Régression
Clustering
Filtre collaboratif
Reconnaissance de Patterns
Prédictions
Regroupement par caractéristiques
Recommandation


Supervised Learning


Unsupervised Learning
dataaspirant.wordpress.com

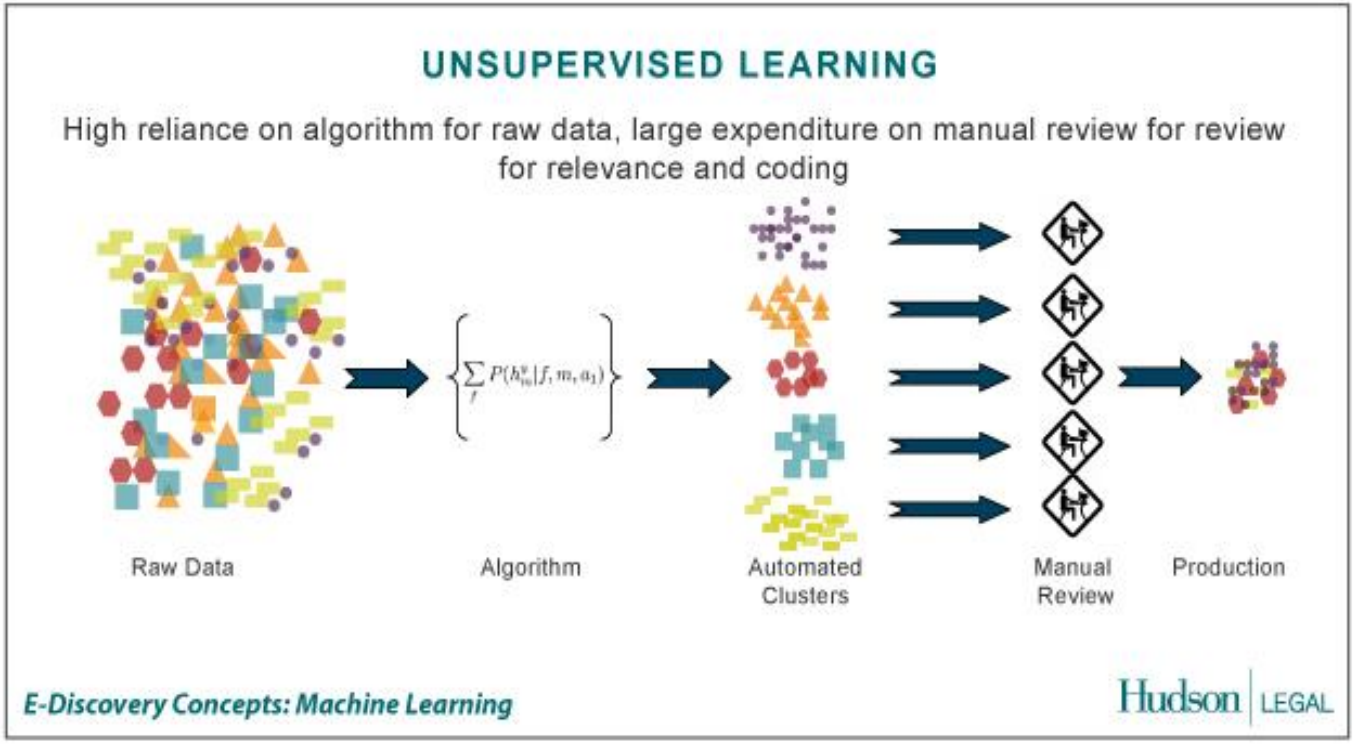
SUPERVISED LEARNING

- Les classes du jeux de données “Training” sont connues



UNSUPERVISED LEARNING

- Les classes des jeux de données ne sont pas connues



Credit: http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth

USE-CASES

- Spam Email Détection
- Traductions (Language traduction)
- Recherche d'Images (Similarité)
- Clustering (KMeans) : Recommendations
- Classification : Google News
- Rating a Review/Comment: Yelp
- Fraud detection : Credit card Providers
- Decision Making : e.g. Bank/Insurance sector
- Sentiment Analysis
- Speech Understanding – iPhone with Siri
- Face Detection – Facebook's Photo tagging

CLASSIFICATION IN ACTION

ISN'T IT EASY?

IT'S NOT (SNAPSHOT OF SPAM FOLDER)

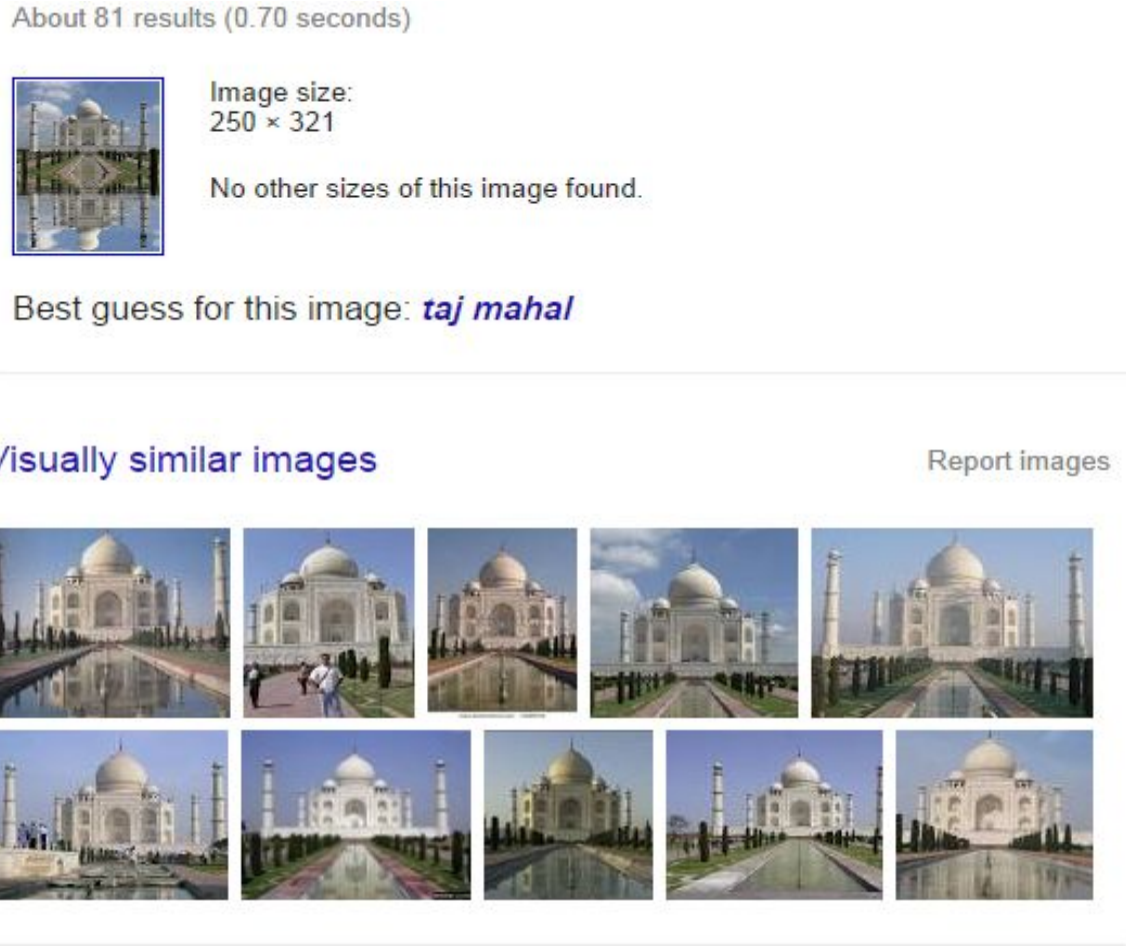
Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab -- If you do not want to receive any more newsletters, please click here	9:40 pm
iEntry	**Welcome iEntry Member** - Ultimate Guide To Assessing	9:23 pm
New-Zealand-Jobs.067L	**Come to New Zealand to find a great job and settle here (Search for all Jobs from diffe...** - Search for all Jobs from different kinds of industries Find a Job in Enchanting	8:18 pm
CarSizzler	**Assured Free Luxurious Ride worth Rs.300 with Uber Cabs** - Home Home Buy New Car Buy New Car Sell Car Sell Car Tech Tics Tip & Tale Facebook 41727 others	6:05 pm
Supermarket Promotion	**Enjoy Rs.1700 voucher valid at any supermarket!** - If you are unable to view this mailer Click here HOW TO CONTACT US? BY EMAIL: support@savethedeals.in	4:51 pm
Entireweb Newsletter	**Hire an SEO the Right Way – 6 Tips You Must Remember for Life** - Unsubscribe me View web version Become a fan on Facebook Follow us on Twitter September 5th, 2	1:24 pm
Max Bupa	**A policy that understands your family's medical need** - open in fresh tab -- If you do not want to receive any more newsletters, pleas	11:08 am
Scoop.it	**Your Scoop.it Daily Summary - How to Maximize Your LinkedIn Publishing Exposure | SME a...** - Scoop.it Facebook Twitter G+ H re the top	9:30 am
standard charterer Bank	**Instant approval on your Credit Card`**	7:27 am
CAR TRADE	**Sell your car at no cost at all** - If you are having trouble viewing this email,view web version | View this message in your mobile |	Sep 4
Uday	**VPS Web Hosting Services Provider** - Dear Sir, I am Uday Sharma, Business development executive. We are providing quality VPS hosting for	Sep 4
Mark Regan, SPN	**How to Find Your Most Valuable Keywords [Free Guide]** - This is a SiteProNews/ExactSeek Webmaster Exclusive Mailing! To drop your subscription, use the link	Sep 4
HDFC Bank	**LOAN upto Rs 25 lac - Disbursal in 2 days** - open in fresh tab You have received this mailer from Shop@Best on behalf of HDFC Bank because you	Sep 4
CAR TRADE	**Sell your car at no cost at all** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
ICICI Bank	**Home Loan Interest Rate starting from 10.15%*. Get Instant Approval!** - open in fresh tab -- If you do not want to receive any more newsletters, please Click Here	Sep 4
calculateyourwealth	**It's good when your bank helps you manage your wealth and fulfill your ambitions** - Calculate Now Dreams you wish to realize in your lifetime require enough wealth. (	Sep 4
Angel Broking	**Get Low Brokerage - Free Demat & Trading Account** - open in fresh tab -- If you do not want to receive any further newsletters, please click here	Sep 4
Bankbazaar	**7 Minute Instant Online Approval for your PESONAL LOAN** - Now get instant online Personal Loan approval in 7 minutes by BankBazaar.com from leading Banks in	Sep 4
Jayde	**Welcome To The Jayde Newsletter!** - Welcome To WebProNews Welcome To The Jayde Newsletter! Before we begin, make sure to add	Sep 4
ineedhits noreply	**[ineedhits] Your ineedhits Account and Password** - ACCOUNT CREATION Account ID : A1588368 Dear Rah, Welcome to ineedhits. Yo	Sep 4
Rekha	**Mobility Apps for Your Business** - While we look at the span of last 20 years, we could broadly look at two distinct eras, - Life in	Sep 4
SlideShare Newsletter	**Top Tips From the World Champions of PowerPoint** - View online version Remember to display images Meet the PowerPoint World Champs Top Tips From the	Sep 4
Dilshad Pathan	**Feeling Hesitate to Discuss personal Health Queries** - My Life Care Follow Us on facebook twitter linkedin Google+ Feeling Hesitate to Discuss personal	Sep 4
Vaishu	**TAKE YOUR PICK. Register in SimplyMarry** - TAKE YOUR PICK. Register in SimplyMarry -- Regards Vaishu	Sep 4

Not a Spam

Not a Spam

SIMILARITÉ/DUPLICATION IMAGES



Credit: https://www.google.co.in/

Rappel:

Features ?

(Feature Extraction)

peut-être:

- Width
- Height
- Contrast
- Brightness
- Position
- Hue
- Colors

Check this :

LIRE (Lucene Image REtrieval) library -

https://code.google.com/p/lire/

MACHINE LEARNING TECHNIQUES

TECHNIQUES

- prédire les classes à partir d'observations

- regrouper les observations en des groupes "reconnaissables" pour prédire les valeurs depuis les observations

CLASSIFICATION

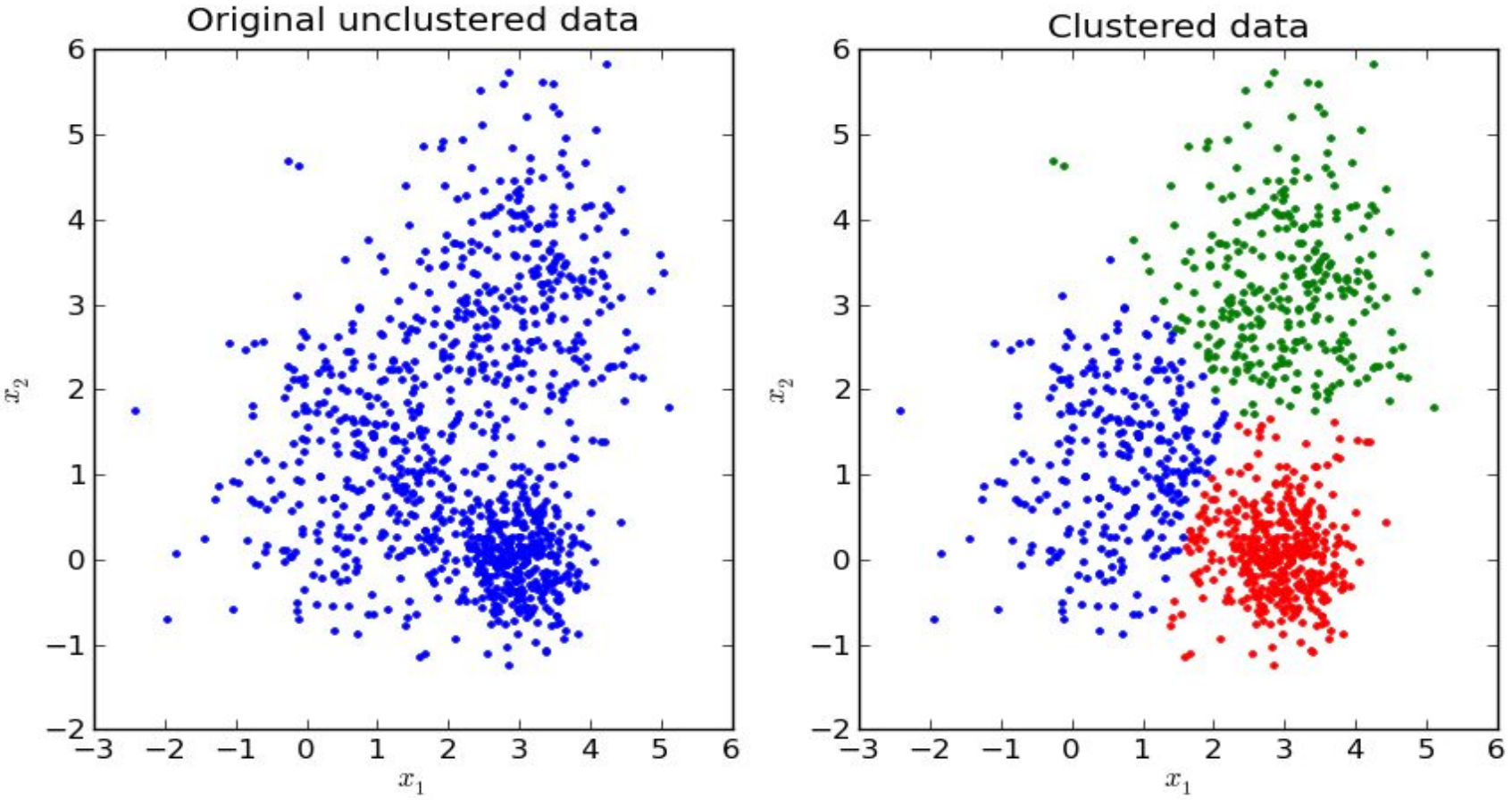
- Classifier un document en une catégorie prédéfinie
- Les documents peuvent être du texte, des images
- Naive Bayes Classifier est le plus populaire
- Etapes:
  - Step1 : “Train” le programme (Construire un Modèle) avec un jeux de données avec une catégorie par exemple sports, cricket, news,
  - Le “Classifier” calcule les probabilités pour chaque mot, la probabilité qu’un document appartienne à l’une des catégories
  - Step2 : Tester avec un autre jeu de données le modèle

- http://en.wikipedia.org/wiki/Naive_Bayes_classifier

CLUSTERING

- **clustering** action de regrouper un ensemble d'objets de telle manière que les objets qui sont dans un même groupe(appelé **cluster**) sont similaires entre eux
- Les objets ne sont pas prédéfinis
- Par exemple les mots clés suivants
  - "man's shoe"
  - "women's shoe"
  - "women's t-shirt"
  - "man's t-shirt"
  - Peuvent être clusterisé en deux catégories "shoe" and "t-shirt" ou"man" and "women"
- **K-means clustering** et **Hierarchical clustering** sont les plus connus

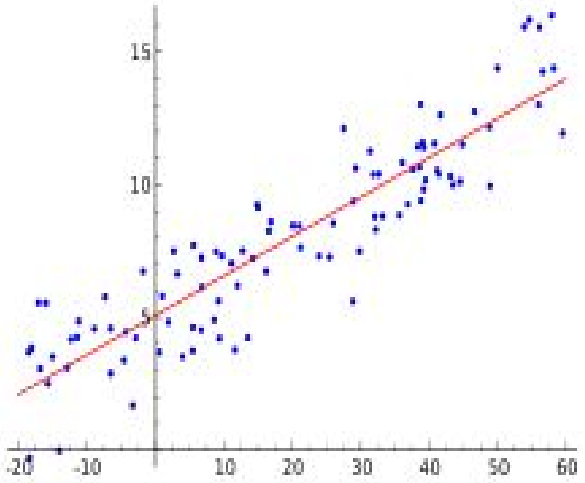
K-MEANS CLUSTERING



- partition ***n*** observations dans ***k*** clusters dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche
- http://en.wikipedia.org/wiki/K-means_clustering

REGRESSION

- Mesure d'une relation entre la valeur moyenne d'une variable et les valeurs correspondantes des autres variables
- **regression analysis** processus de statistique pour estimer les relations entre différentes variables
- Regression veut dire **prédire** la valeur de sortie à partir d'un "training data"
- Logistic regression (binary regression) est la plus populaire
- http://en.wikipedia.org/wiki/Logistic_regression



CLASSIFICATION VS REGRESSION

- Classification permet de grouper les sorties en classes
- classification pour prédire le type de tumeur, par exemple dangereuse ou non avec un “traing data”

- Regression permet de prédire une valeur en sortie à partir d’un “training data”.
- regression pour **prédire** le prix d’une maison à partir de jeux de donénes
- Si on cherche un nombre alors c’est un problème de régression

HADOOP CONCEPTS

HADOOP : Big Data Technology

Impact of Hadoop on Big Data and Global Business Performance

What is Big Data - Intelligence Artificial

Big data is massive and messy, and it’s coming at you uncontrolled. Data are gathered to be analyzed to discover patterns and correlations that could not be initially apparent, but might be useful in making business decisions in an organization. These data are often personal data, which are useful from a marketing viewpoint to understand the desires and demands of potential customers and in analyzing and predicting their buying tendencies.

HADOOP : Big Data Technology

Impact of Hadoop on Big Data and Global Business Performance

Breaking through Silos

Silos are a result of hierarchies of the organization, which require organizing people into economically effective groups. Data silos become a barrier that impedes decision-making and organizational performance. Enterprises are facing many challenges to glean insight with Big Data Analytics that trapped in the data silos exist across business operations. Through the effective handling of big data can stymie data silos and the enterprise can leverage available data into emerging customer trends or market shifts for insights and productivity.

Excercice

How you will do it ?

Give me different solutions to count the occurence of each words that appears in this document:

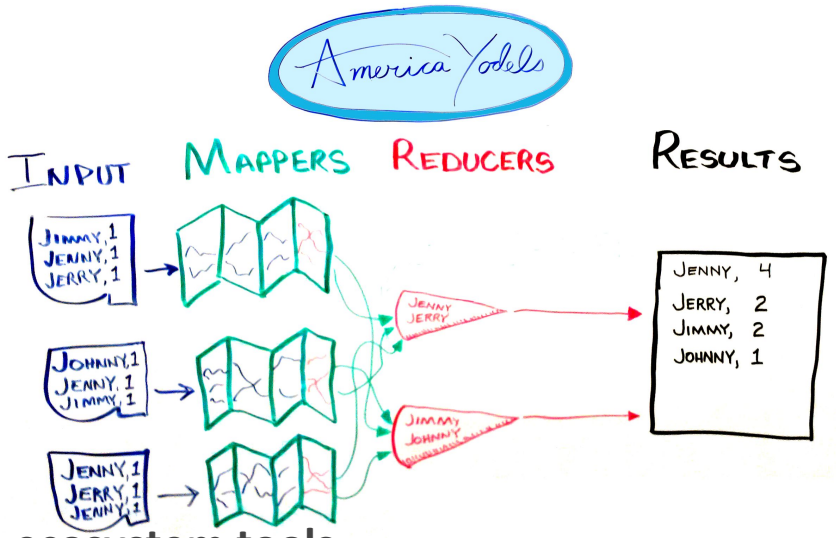
To be, or not to be, that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep
No more; and by a sleep, to say we end
the heart-ache, and the thousand natural shocks
that Flesh is heir to? 'Tis a consummation
devoutly to be wished. To die, to sleep,
To sleep, perchance to Dream; aye, there's the rub,
for in that sleep of death, what dreams may come,
when we have shuffled off this mortal coil,
must give us pause. There's the respect
that makes Calamity of so long life:
For who would bear the Whips and Scorns of time,
the Oppressor's wrong, the *proud* man's Contumely, [F: *poor*]
the pangs of *despised* Love, the Law's delay, [F: *disprized*]
the insolence of Office, and the spurns
that patient merit of the unworthy takes,
when he himself might his Quietus make
with a bare Bodkin? Who would Fardels bear, [F: *these Fardels*]
.

to grunt and sweat under a weary life,
but that the dread of something after death,
the undiscovered country, from whose bourn
no traveller returns, puzzles the will,
and makes us rather bear those ills we have,
than fly to others that we know not of.
Thus conscience does make cowards of us all,
and thus the native hue of Resolution
Is sicklied o'er, with the pale cast of Thought,
And enterprises of great *pitch* and moment, [F: *pith*]
with this regard their Currents turn *awry*, [F: *away*]
And lose the name of Action. Soft you now,
The fair Ophelia? Nymph, in thy Orisons
Be all my sins remember'd

AND: 12
……

cf https://en.wikipedia.org/wiki/To_be,_or_not_to_be

HADOOP : Big Data Technology



In summary: Hadoop = Map Reduce + HDFS + ecosystem tools.

Map Reduce aims to execute batch processing on a distributed system.

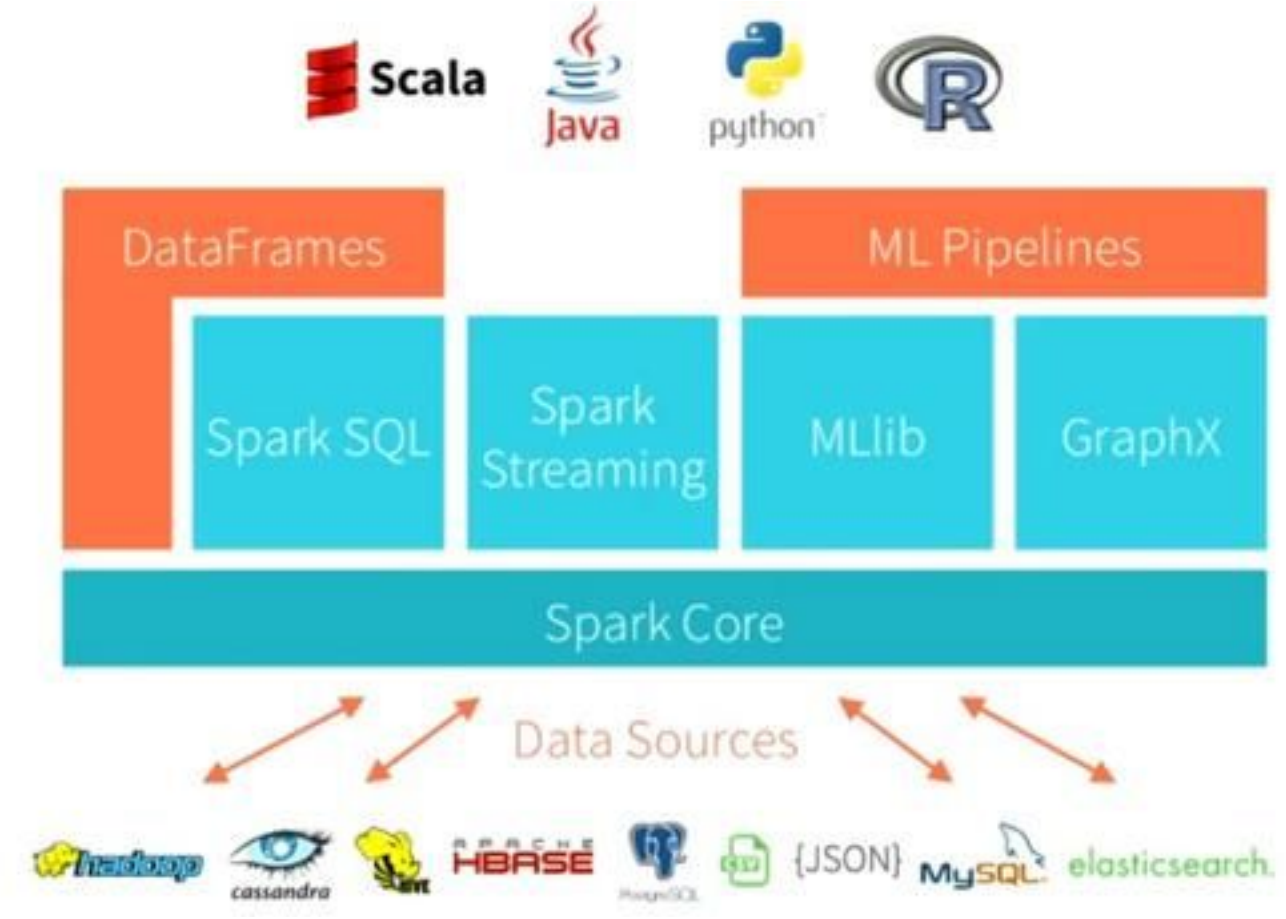
The idea is that we manage data that is a priori unstructured, so we have a step of extracting and processing the relevant data.

What is called **Map**.

It is to read information in a file, to extract the part which concerns us, and to present it in the form of key value. Finally, we obtain a looooooongue list, distributed, of key - value pairs.

A step of processing the relevant data, which bears the sweet name of Reduce. As much as the Map brings us back lists of key - value pairs, as much the reduce will treat, and aggregate, these values.

Les Technologies
Les grands principes du calcul distribué



CALCUL DISTRIBUÉ : SPARK

IMAGINEZ SI VOUS POUVIEZ EXÉCUTER SUR DE LARGES VOLUMES DE DONNÉES, DES ITÉRATIONS EN CONTINU

Les grands principes du calcul distribué:

- De nombreux traitements statistiques sont très souvent réalisés sur une machine ➔ la distribution permet de découper ces traitements sur plusieurs noeuds (machines) et donc de pouvoir découpler les problématiques liées à la “taille” d’une machine aux temps de traitements

- SPARK est un framework de type “LAZY”, les opérations de calcul ne sont effectuées que lors d’un appel de restitution

Les bénéfices

- Les avantages sont la capacité à pouvoir traiter de larges volumes de données en parallèle
- Une réduction du cout des calculs (moins cher d’avoir plusieurs serveurs qu’un “gros” serveur

- SPARK est jusquà 100x plus rapide que le MapReduce de HADOOP

Les Technologies : Modélisation

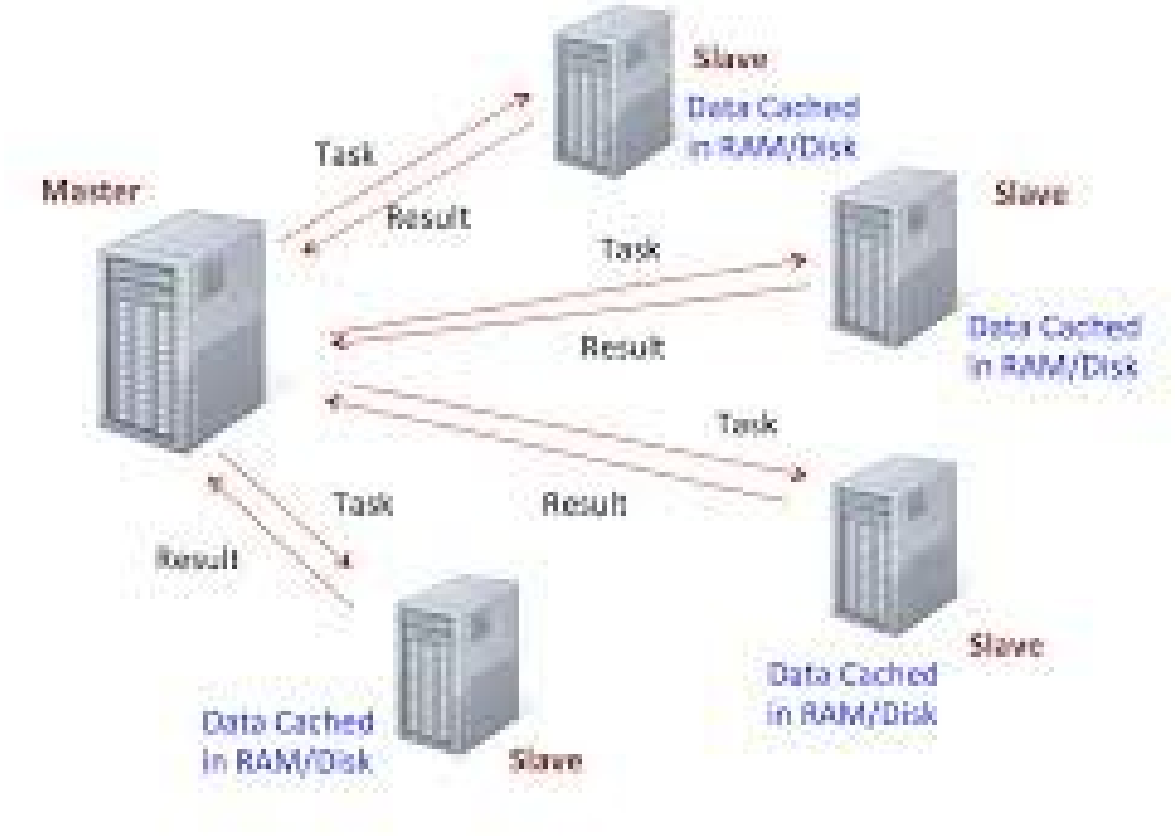
Apache Spark

Scaling > 10 000 computers

- Stockage et traitement de **exabytes** de données
  - 1 To (TeraByte) = 1 000 Go (HUPI aujourd'hui, 90 To)
  - 1 Po (PetaByte) = 1 000 To
  - 1 Eo (ExaByte) = 1 000 Po
- Tolérance aux pannes lors des échecs (relance automatique)
- 2 tiers des serveurs peuvent "crasher" et pourtant la requête sera effectuée
- Indépendance du matériel et des logiciels

Les Technologies
Les grands principes du calcul distribué

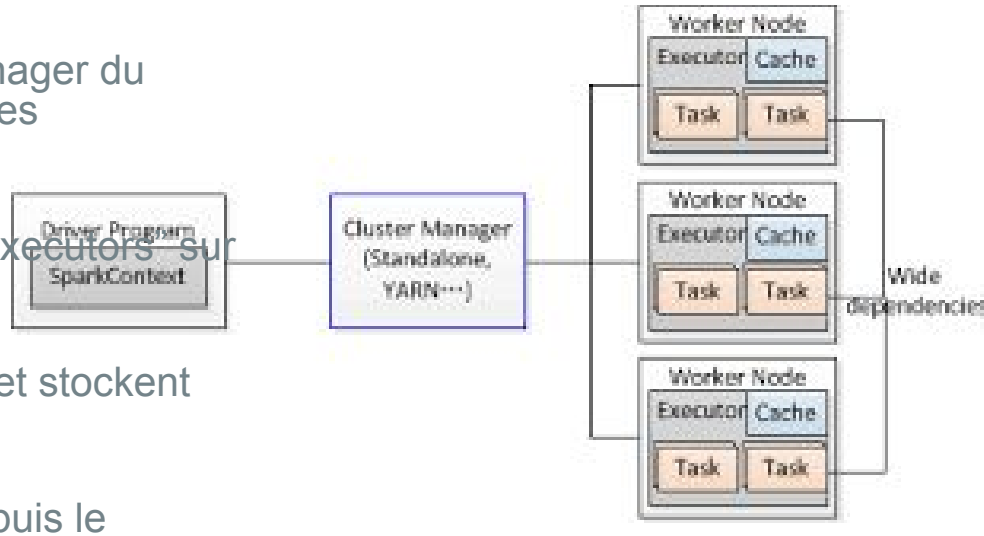
- Les clusters



ARCHITECTURE SPARK

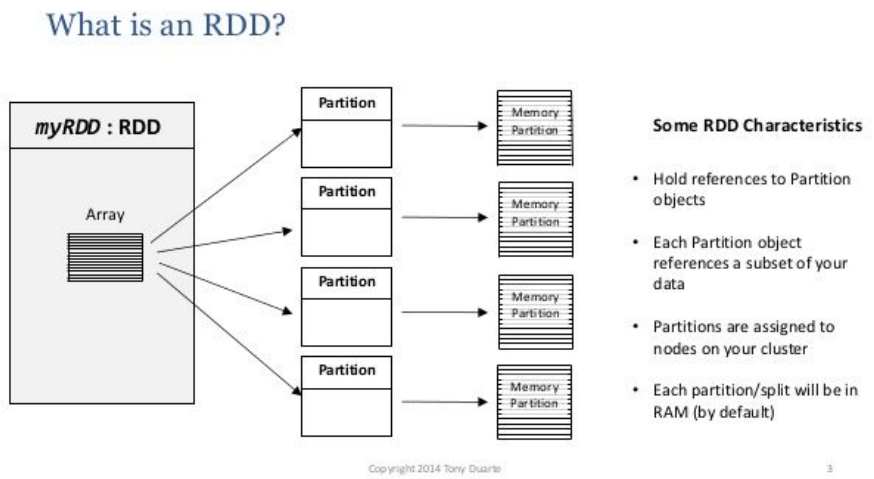
Architecture:

- Les applications Spark s'éxécutent en processus indépendants sur les clusters, coordonnés par l'objet "SparkContext" dans le programme "main", appelé le "Driver Program"
- Le SparkContext se connecte aux Resources Manager du cluster (gestion des ressources pour l'ensemble des applications)
- Une fois connecté au cluster, Spark obtient des "executors" sur chaque noeud du cluster
- Les "executors" sont les processus qui exécutent et stockent les résultats
- Ensuite Spark envoie le code aux "executors" (depuis le SparkContext)
- Finalement, le SparkContext envoie les tâches devant être exécutées



SPARK
RESILIENT DISTRIBUTED DATASET

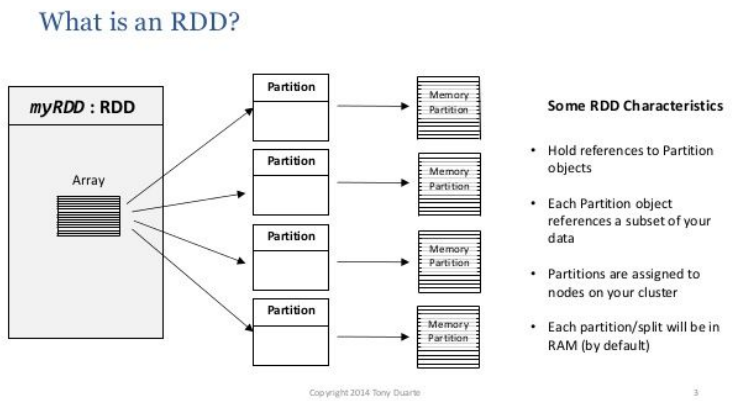
- Resilient Distributed Datasets (RDDs) permettent de réutiliser efficacement les données dans une large famille d'applications.
- Les RDDs sont tolérants à la panne et proposent des structures de données parallèles qui laissent les utilisateurs :
  - persister explicitement les données intermédiaires en mémoire,
  - controller leur partitionnement afin d'optimiser l'emplacement des données,
  - manipuler les données en utilisant un ensemble important d'opérateurs.



RESILIENT DISTRIBUTED DATASET

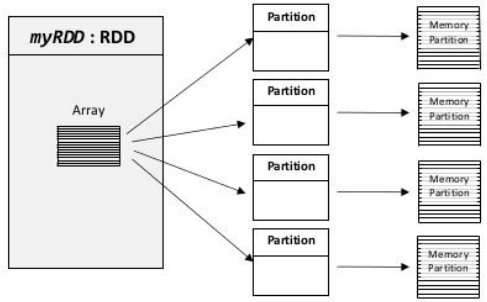
- Un RDD est une collection partitionnée d'enregistrements en lecture seule qui ne peut être créée que par des opérations déterministes :
  - soit à partir de données présentes dans un stockage stable,
  - soit à partir d'autres RDDs.
- Ces opérations sont appelées transformations pour les différentier des autres opérations. il y a : map, filter et join.

- Les RDDs n'ont pas besoin d'être matérialisés tout du long puisqu'un RDD dispose de suffisamment d'informations sur la façon dont il a été produit à partir d'un autre ensemble de données pour pouvoir être recalculé. Ainsi, un programme ne peut faire référence à un RDD s'il n'est pas capable de le reconstruire suite à une panne.



SPARK
RESILIENT DISTRIBUTED DATASET

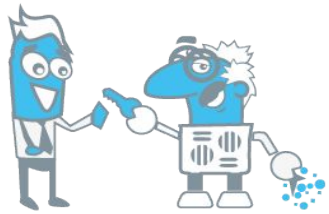
Transformations	$map(f : T \Rightarrow U)$	:	RDD[T] ⇒ RDD[U]
	$filter(f : T \Rightarrow Bool)$	:	RDD[T] ⇒ RDD[T]
	$flatMap(f : T \Rightarrow Seq[U])$	:	RDD[T] ⇒ RDD[U]
	$sample(fraction : Float)$	:	RDD[T] ⇒ RDD[T] (Deterministic sampling)
	$groupByKey()$	:	RDD[(K, V)] ⇒ RDD[(K, Seq[V])]
	$reduceByKey(f : (V, V) \Rightarrow V)$	:	RDD[(K, V)] ⇒ RDD[(K, V)]
	$union()$	:	(RDD[T], RDD[T]) ⇒ RDD[T]
	$join()$	:	(RDD[(K, V)], RDD[(K, W)]) ⇒ RDD[(K, (V, W))]
	$cogroup()$	:	(RDD[(K, V)], RDD[(K, W)]) ⇒ RDD[(K, (Seq[V], Seq[W]))]
	$crossProduct()$	:	(RDD[T], RDD[U]) ⇒ RDD[(T, U)]
	$mapValues(f : V \Rightarrow W)$	:	RDD[(K, V)] ⇒ RDD[(K, W)] (Preserves partitioning)
	$sort(c : Comparator[K])$	:	RDD[(K, V)] ⇒ RDD[(K, V)]
	$partitionBy(p : Partitioner[K])$	:	RDD[(K, V)] ⇒ RDD[(K, V)]
Actions	$count()$	:	RDD[T] ⇒ Long
	$collect()$	:	RDD[T] ⇒ Seq[T]
	$reduce(f : (T, T) \Rightarrow T)$	:	RDD[T] ⇒ T
	$lookup(k : K)$	:	RDD[(K, V)] ⇒ Seq[V] (On hash/range partitioned RDDs)
	$save(path : String)$	:	Outputs RDD to a storage system, *e.g.*, HDFS



Some RDD Characteristics

- Hold references to Partition objects
- Each Partition object references a subset of your data
- Partitions are assigned to nodes on your cluster
- Each partition/split will be in RAM (by default)

ARCHITECTURE OF A BIG DATA PLATFORM



Visualisation




6

Produits Analytiques
(semi-génériques)

Sources

Service client


Logs


Objets connectés *Web services*


Applications mobile *Fichiers*

API RestFul

CatchBox
(collecte)

Kafka

1

Traitements

4

Exploration

5

Stockage

2

Qualification
des données

3

Appels /
Restitutions

API/RestFul

7

8

Destination

Exercice : Scandi-Vie

Un site eCommerce Scandivie (https://www.scandi-vie.com)

- Tracking des évènements (cf. github hupi / hupilytics)
- Sauvegarde des évènements en HDFS
  - pages cliquées
  - pages produits visitées

Vos premiers programmes

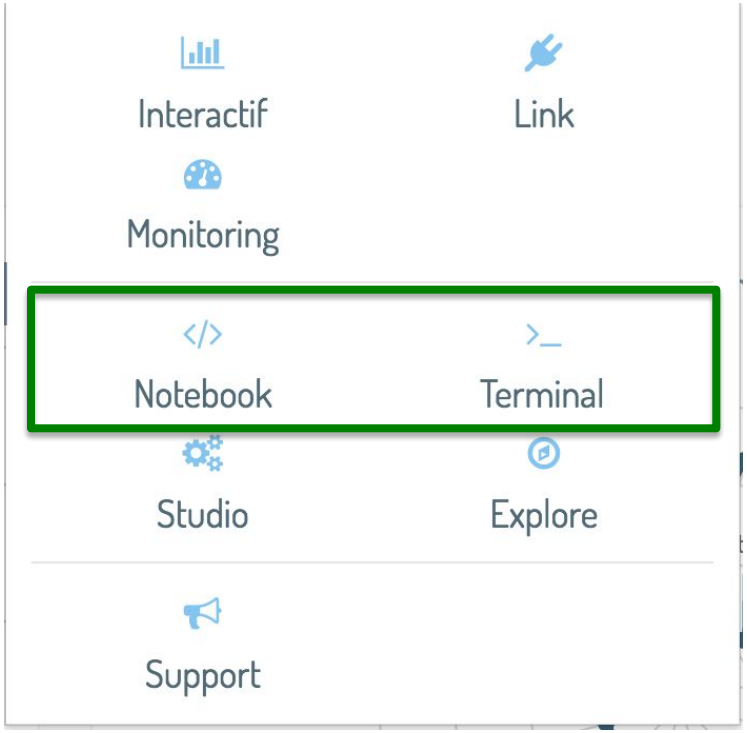
- Calcul d’indicateurs simples et les afficher
- Développer un modèle de similarité de produits

OUTILS NOTEBOOKS

Objectifs:

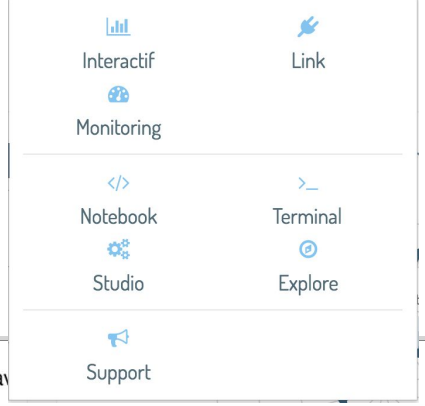
Les outils de type Notebook permettent de tester rapidement et de valider les modèles analytiques

- “Notebook” permet l’accès au développement en *Spark – Scala*.
- “Terminal” permet l’accès au développement en Spark – Python



HUPI recommande de développer en Scala:

- Versions plus fréquentes
- Meilleures performances



TERMINAL - SPARK

JUPYTER

jupyter nb10-sql-dataframes (unsav

File Edit View Insert Cell Kernel Help Python 2

Markdown CellToolbar

Spark SQL and Data Frames

Introduction to Spark with Python, by Jose A. Dianes

This notebook will introduce Spark capabilities to deal with data in a structured way. Basically, everything turns around the concept of *Data Frame* and using *SQL language* to query them. We will see how the data frame abstraction, very popular in other data analytics ecosystems (e.g. R and Python/Pandas), it is very powerfull when performing exploratory data analysis. In fact, it is very easy to express data queries when used together with the SQL language. Moreover, Spark distributes this column-based data structure transparently, in order to make the querying process as efficient as possible.

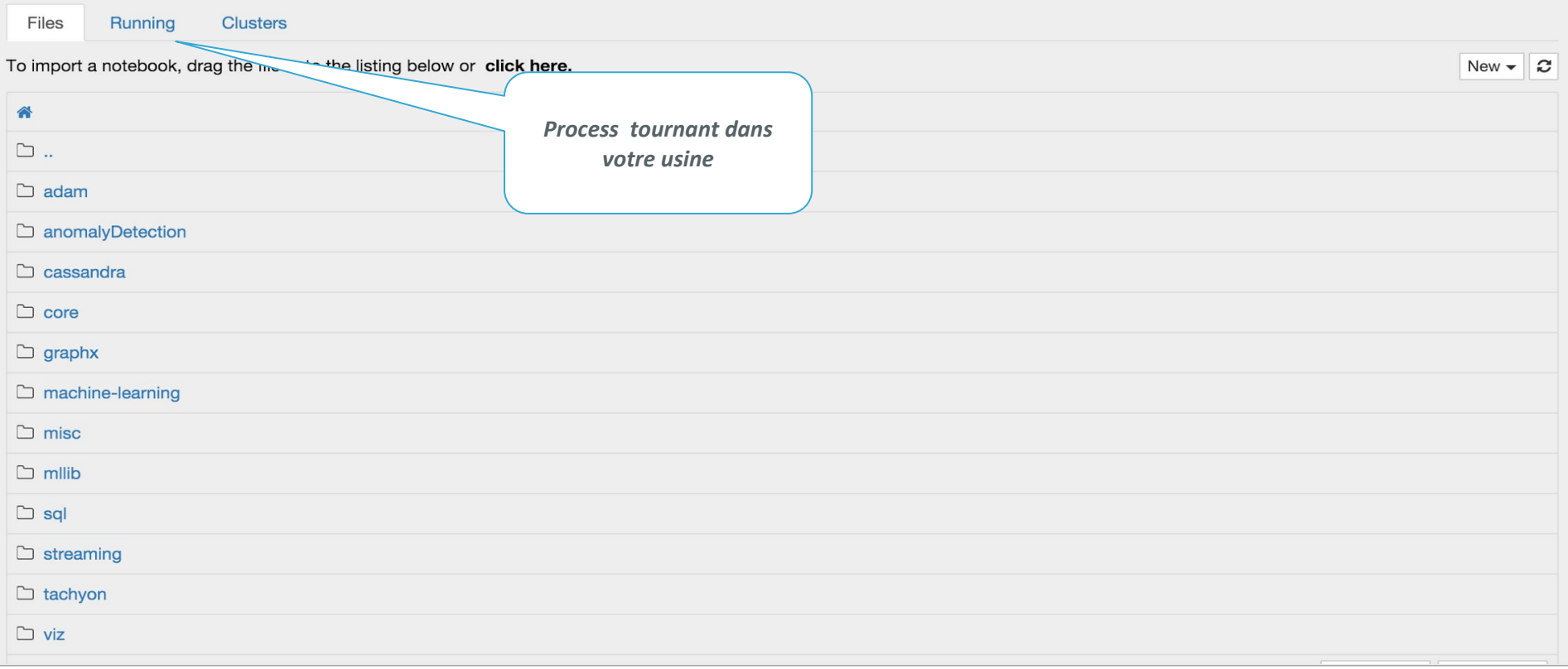
Getting the data and creating the RDD

As we did in previous notebooks, we will use the reduced dataset (10 percent) provided for the KDD Cup 1999, containing nearly half million nework interactions. The file is provided as a Gzip file that we will download locally.

In [1]:
```
import urllib
f = urllib.urlretrieve ("http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz", "kddcup.data_10_percent.gz")
```

In [2]:
```
data_file = "./kddcup.data_10_percent.gz"
raw_data = sc.textFile(data_file).cache()
```

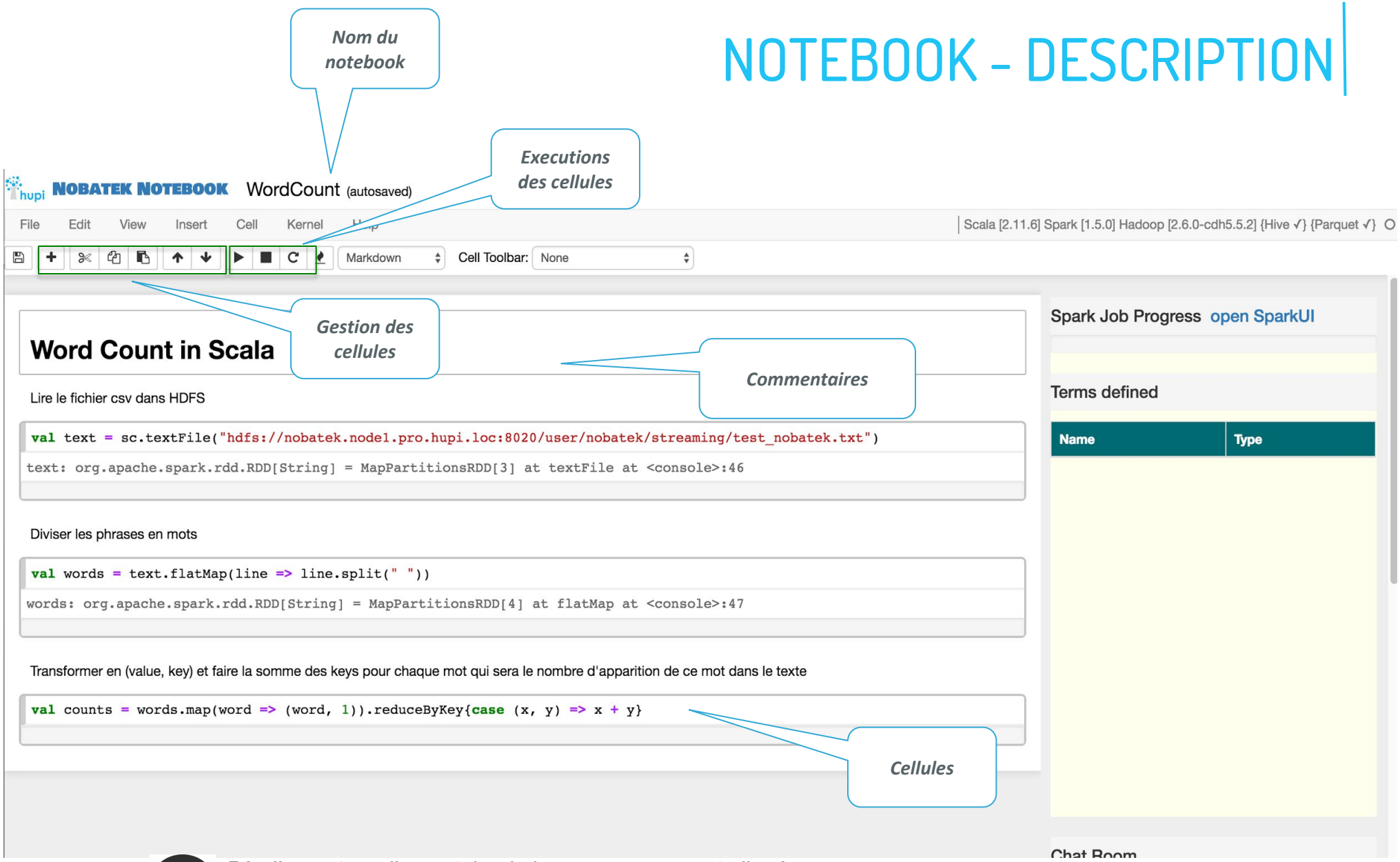
hupi SPARK NOTEBOOK



- Gestion des process actifs

Les processus sont limités, vérifier que les programmes ne sont pas trop nombreux

NOTEBOOK - DESCRIPTION



Dès l'ouverture d'un notebook, les ressources sont allouées

EXERCICES

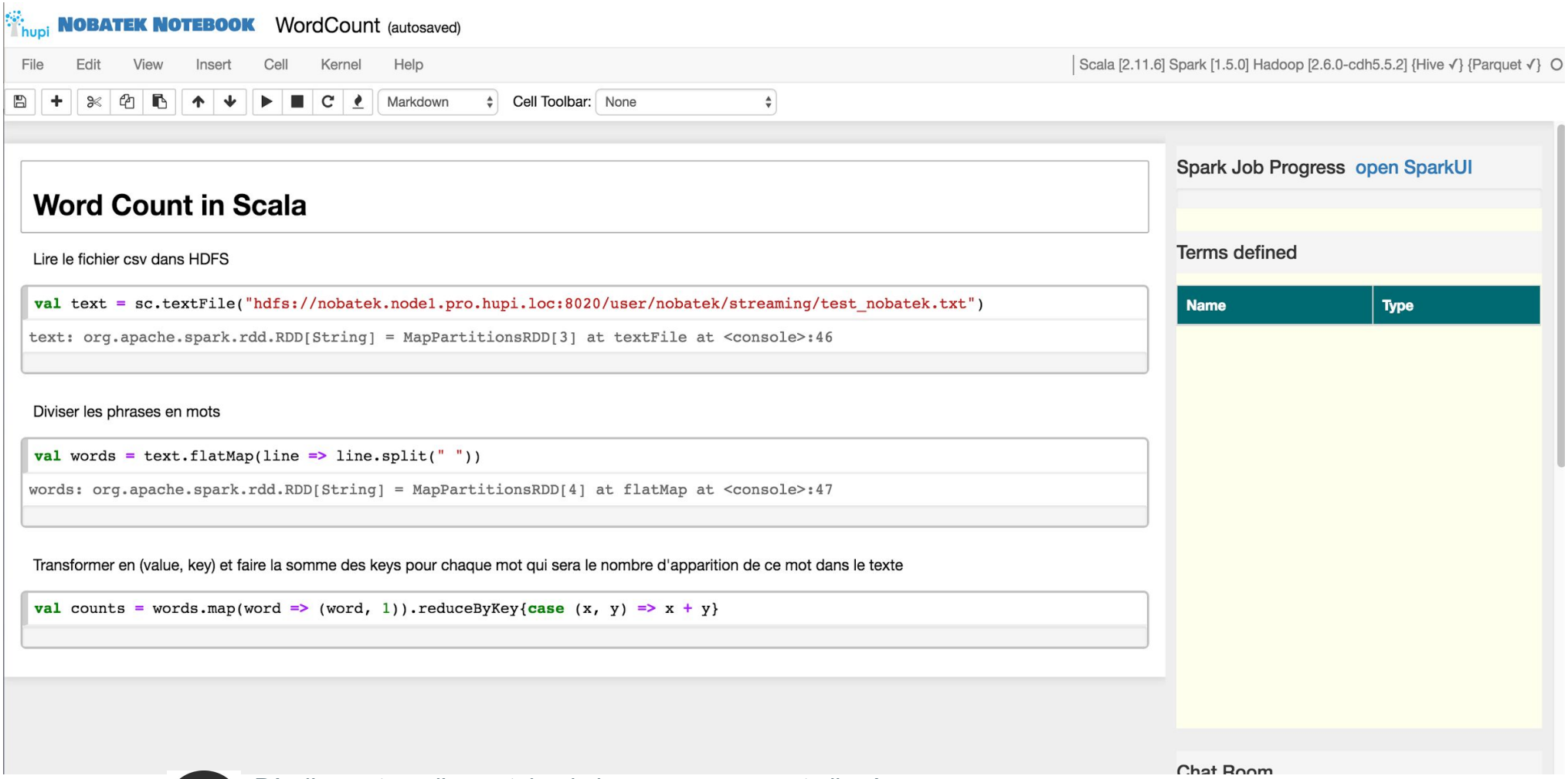
Exercices pour découvrir les grandes fonctionnaliatés

- WordCount
- Manipulation des RDD
- Manipulation des DATAFRAMES

Exercices pour découvrir la modélisation analytique:

- Découverte d'élements qui apparaissent souvent ensemble : Frequent Pattern mining FP-Growth
- Prédiction des consommations d'énergie : Régression linéaire
- Aide à la décision : Darwin-ecoles

NOTEBOOK - WORDCOUNT



Dès l'ouverture d'un notebook, les ressources sont allouées

NOTEBOOK – RDD

hupi **NOBATEK NOTEBOOK** RDD Manipulation (autosaved)

File Edit View Insert Cell Kernel Help Scala [2.11.6]

Markdown Cell Toolbar: None

Création un RDD

```
val liste = List("pandas", "i like pandas")
val lines = sc.parallelize(liste)
// ou val lines = sc.parallelize(List("pandas", "i like pandas"))
```

```
liste: List[String] = List(pandas, i like pandas)
lines: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[1] at parallelize at <console>:48
```

Operations sur RDD

On lit le fichier "test_nobatek.txt"

```
val text = sc.textFile("hdfs://nobatek.node1.pro.hupi.loc:8020/user/nobatek/streaming/test_nobatek.txt")
```

```
text: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at textFile at <console>:46
```

count()

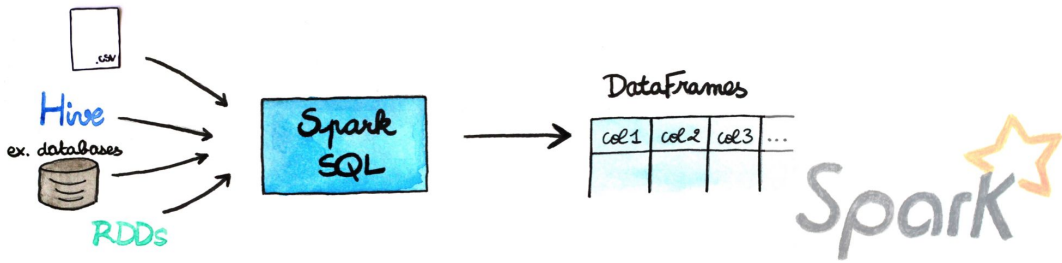
Pour compter nombre de lignes de RDD

```
text.count()
```

SPARK
DATAFRAMES

Objectif:

- Pouvoir manipuler des données sous forme d'un tableau avec des colonnes nommées
- Les DATAFRAMES peuvent être construites à partir de
  - sources de fichiers structurés
  - De RDD
  - Hive
  - De base de données externes




```python
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)

df = sqlContext.read.json("examples/src/main/resources/people.json")

# Displays the content of the DataFrame to stdout
df.show()
```

SPARK DATAFRAMES

hupi **NOBATEK NOTEBOOK** DataFrame (autosaved)

File Edit View Insert Cell Kernel Help | Scala [2.11.6]

Markdown Cell Toolbar: None

1/ Lire un JSON file de HDFS et créer un dataframe temporaire

Importer des packages nécessaires

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._
import org.apache.spark.sql._
```

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._
import org.apache.spark.sql._
```

Créer un SQLContext

```
val sqlContext = new SQLContext(sc)
```

```
sqlContext: org.apache.spark.sql.SQLContext = org.apache.spark.sql.SQLContext@3567c360
```

DÉCOUVERTE DE PATTERNS

hupi NOBATEK NOTEBOOK FP-Growth (autosaved)

File Edit View Insert Cell Kernel Help Scala [2.11.6]

Markdown Cell Toolbar: None

Règles d'association ¶

But :

- Mettre en évidence les éléments qui apparaissent souvent ensemble
- On espère obtenir un résultat comme : Si batiment + °C alors KWH_ELEC + KWH_CVC ou bien si BatA-LOT1 => KWH_ELEC + °C par exemple

Algorithme choisi : FP-Growth utilise une structure d'arbre (FP-tree) pour stocker une forme compressée d'une base de données. FP-growth adopte une stratégie de découpage pour décomposer les taches d'exploration de données et les bases de données. Il utilise une méthode « pattern fragment growth » pour éviter le coûteux processus de génération et de test des candidats, utilisé par Apriori.

Ici, on ne prend en compte que les lots des batiments, niveau de consommation d'électricité, celui de chauffrage et la température. On ne prend pas l'eau car c'est mesuré par batiment (RDC de BatA + étages de BatA + 2eme étage de BatC). Comme unité ici est lot et que il y a une grande différence entre le nombre des lots pour chaque batiment (A : Lot01->Lot19 + Commun A (BatA-N1 ou BatA-N2) / C : Lot20 + Lot21 + Commun C (dans l'input c'est BatC-NULL)), du coup, ce sera pas évident de prendre en compte cette variable..

Création de DataFrame

Lecture des fichiers JSON

PRÉDICTIONS

hupi **NOBATEK NOTEBOOK** Linear Regression With SGD (autosaved)

File Edit View Insert Cell Kernel Help Scala [2.11.6]

Markdown Cell Toolbar: None

```
plit at <console>:87
testData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = MapPartitionsRDD[125] at randomSplit
at <console>:87
model: org.apache.spark.mllib.regression.LinearRegressionModel = org.apache.spark.mllib.regression.LinearRe...
```

```
val MSE = valuesAndPreds.map{case(v, p) => math.pow((v - p), 2)}.mean()
println("training Mean Squared Error = " + MSE)
```

```
training Mean Squared Error = 1.2442576813043335E8
MSE: Double = 1.2442576813043335E8
```

```
// Save and load model
model.save(sc, "myModelPath")
val sameModel = LinearRegressionModel.load(sc, "myModelPath")
```

```
// Pour tester
val features = Vectors.dense(List(250d,20d).toArray)
val pred = model.predict(features)
```

```
features: org.apache.spark.mllib.linalg.Vector = [250.0,20.0]
pred: Double = 81.38141621174667
```

hupi **NOBATEK NOTEBOOK** Darwin_Nobatek Last Checkpoint: 01/02/2015 (autosaved)

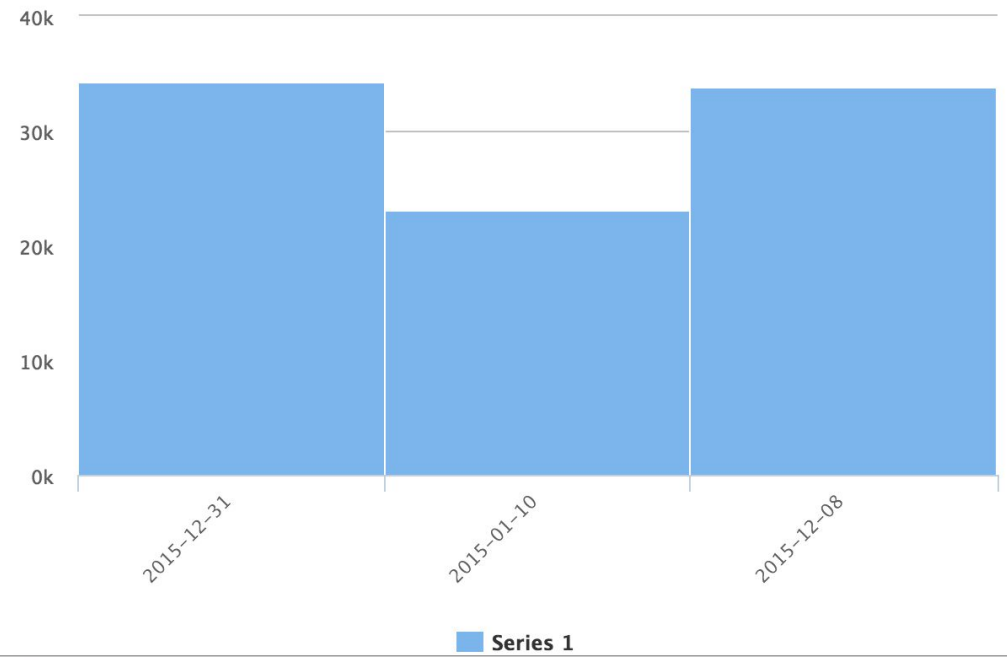
File Edit View Insert Cell Kernel Help | Scala [2.11.6]

Markdown Cell Toolbar: None

Par jour

```
val j1 = meanKWH_per_day.rdd.map{x:Row => (x.getAs[String](0),x.getAs[Double](1))}
val k1 = j1.collect()
PlotH(com.quantifind.charts.highcharts.Histogram.histogram(Seq(k1(0), k1(1), k1(2))))
```

```
j1: org.apache.spark.rdd.RDD[(String, Double)] = MapPartitionsRDD[582] at map at <console>:84
k1: Array[(String, Double)] = Array((2015-12-31,34321.72), (2015-01-10,23055.08), (2015-12-08,33811.82), (2014-08-26,1882
9.26), (2014-12-31,22673.15), (2015-09-15,30680.38))
res72: notebook.front.third.wisp.PlotH = <PlotH widget>
```



COMPTEZ EN TEMPS RÉEL LE NOMBRE DE VISITEURS D'UN SITE PAR INTERVALLE DE TEMPS

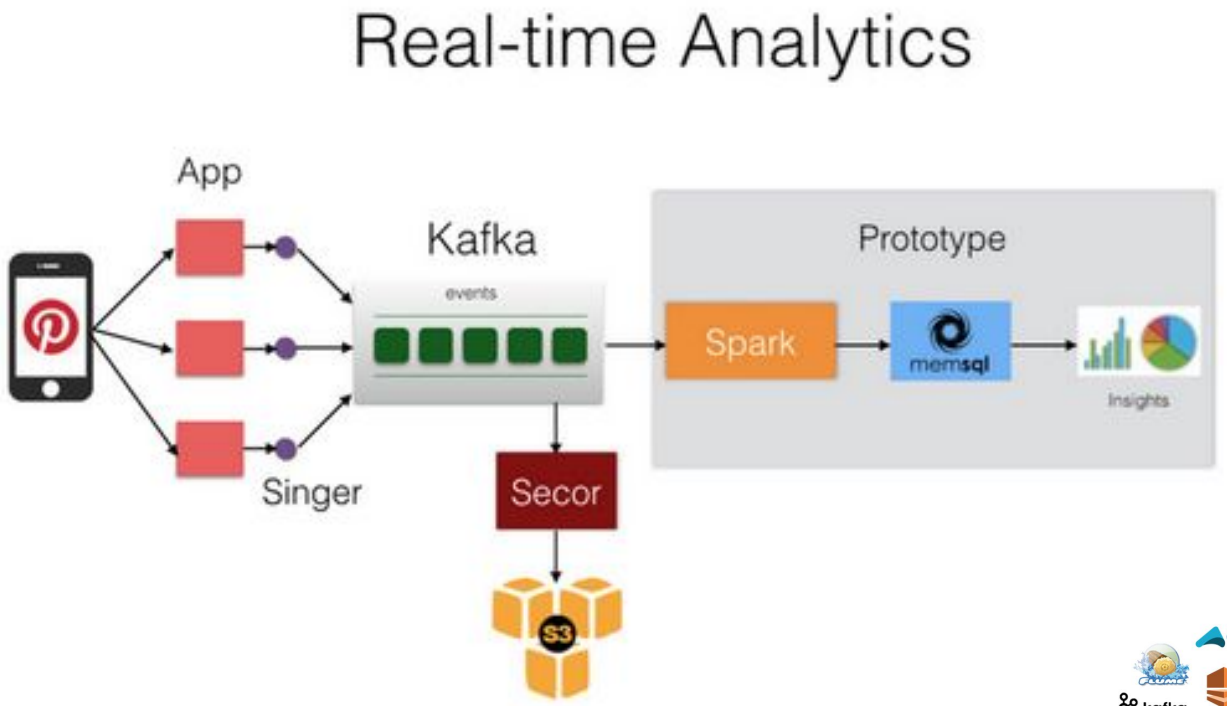
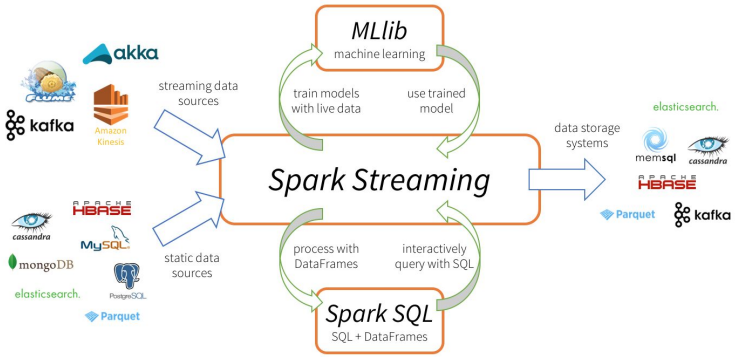


Figure 1: All elements of the real-time analytics platform



SPARK EXERCICES STREAMING

- En utilisant l'interface CATCHBOX, nous allons insérer des données en streaming et les exploiter
- Comment fonctionne l'API Catchbox:
  - API RESTFul
  - https://github.com/hupi-analytics voir quels sont les configurations pour le message des messages
- Le flux de données est stocké dans un "Topic"
- Ecoutons le flux du "topic"

INTÉGREZ VOS MODELES

- HUPI propose pour mettre à disposition les modèles en temps 2 services
  - Openscoring
  - http

OpenScoring:

Dans le module de prédiction, intégrez vos modèles au format PMML (supporté par une grande communauté).