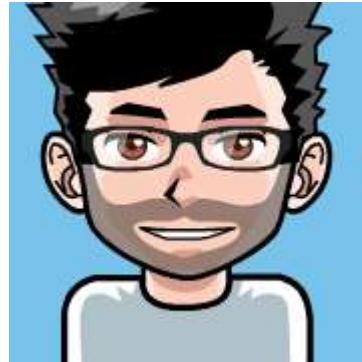


# PANORAMA DU BIG DATA ENJEUX & FONDAMENTAUX

MICHAEL TACHE





## MICHAEL TACHE

[michael.tache@formation.bigdata.com](mailto:michael.tache@formation.bigdata.com)

- Architecte Data, Responsable du service Big Data (Informatique Caisse des Dépôts et des Consignations)
- Responsable du projet de refonte de l'entrepôt de données B2B sur une architecture Big Data à EDF
- Directeur de projets Big Data, BI chez Sopra
  - Clients: Carrefour, Snecma, Safran, EDF, RTE, Samsung, La Poste
- Formateur ([www.formation-bigdata.com](http://www.formation-bigdata.com))
- Ingénieur en système d'information à EFREI
- Futsal ....

# SOMMAIRE

- ✓ **Intro Big Data**
- ✓ **Le Datalake**
- ✓ **La gouvernance**
- ✓ **La GDPR (General Data Protection Regulation)**
- ✓ **Les cas d'usage**
- ✓ **Les Technologies**
  - ✓ Hadoop
  - ✓ Les bases NoSQL

# OBJECTIFS

Présenter les concepts et les enjeux du Big Data avec un focus sur les technologies open source.

- Découvrir les cas d'usages et les technologies Big Data
- Comprendre le principe du Datalake
- Comprendre l'importance de la gouvernance des données et des processus qualité
- Découvrir l'écosystème Hadoop
- Découvrir les principes du NoSQL et les différentes bases associées

# COURS ET ÉVALUATION

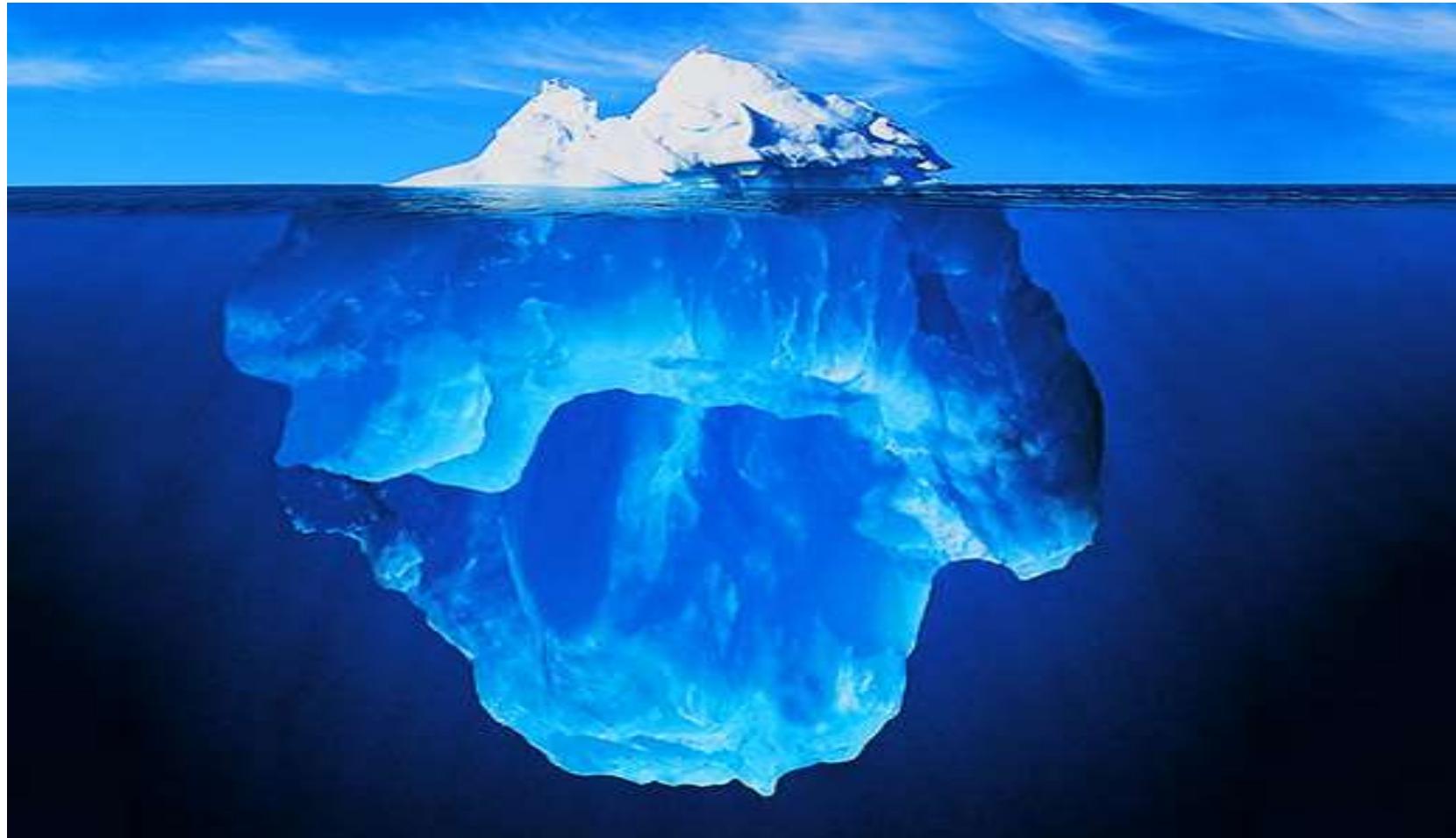
## Cours:

- 2 Séances
- Les dates prévues:
  - Vendredi 12/10 de 09h à 18h
  - Samedi 13/10 de 09h à 18h

## Evaluation:

- QCM ou questions ouvertes

# **ENCORE BEAUCOUP DE DONNÉES NON EXPLOITÉES**



# LA DATA EST LE NOUVEAU PÉTROLE



# **LES OUTILS NE SONT PLUS ADAPTÉS**



# Google



DATA

# DE LA DONNÉE AU BIG DATA

**Octet** : Grain de riz



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet : Bol de riz



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet : Bol de riz

Mégoctet : 8 sacs de riz

**Gigaoctet : 3 semi-remorques**



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

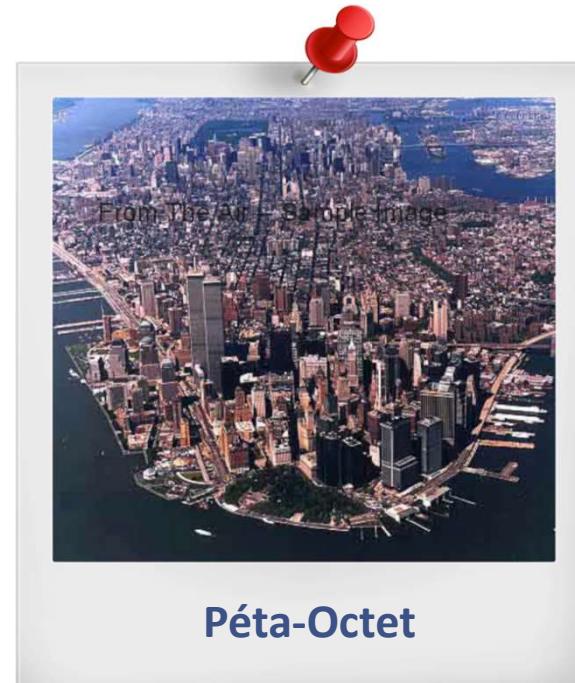
Kilo-octet : Bol de riz

Mégoctet : 8 sacs de riz

Gigaoctet : 3 semi-remorques

Téraoctet : 2 porte-containers

**Pétaoctet : Manhattan couverte de riz**



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet : Bol de riz

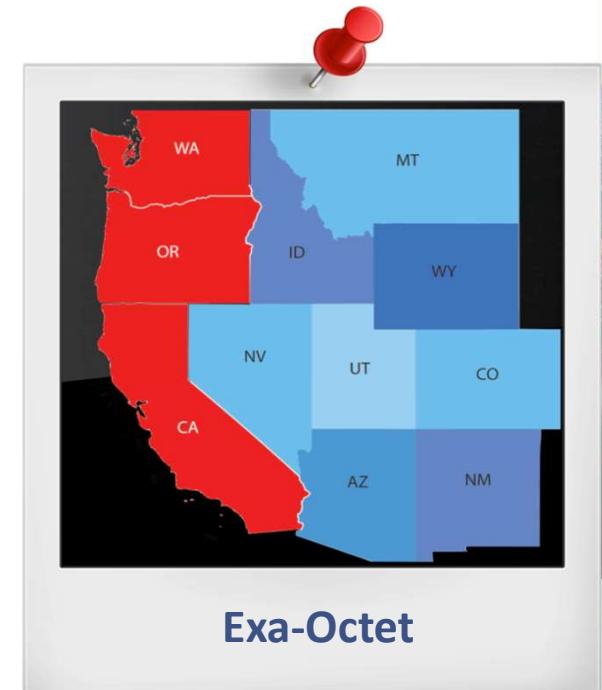
Mégaoctet : 8 sacs de riz

Gigaoctet : 3 semi-remorques

Téraoctet : 2 porte-containers

Pétaoctet : Manhattan couverte de riz

**Exaoctet : Côte ouest des USA couverte de riz**



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet : Bol de riz

Mégaoctet : 8 sacs de riz

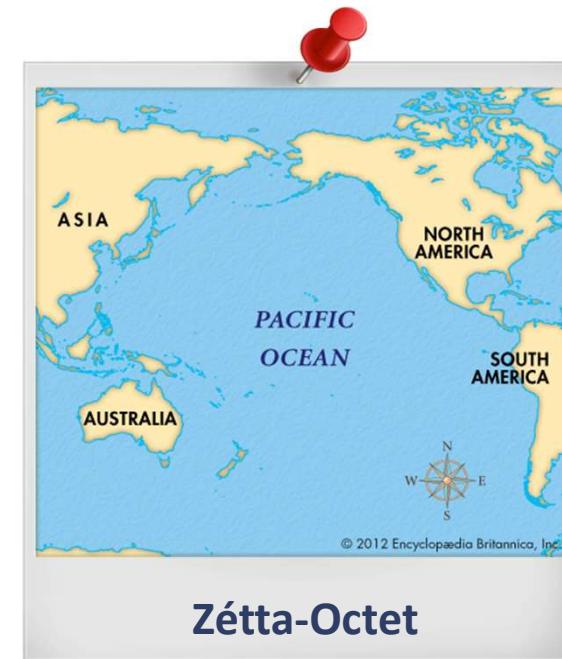
Gigaoctet : 3 semi-remorques

Téraoctet : 2 porte-containers

Pétaoctet : Manhattan couverte de riz

Exaoctet : Côte ouest des USA couverte de riz

**Zétta-octet : couvrir l'océan pacifique de riz**



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet : Bol de riz

Mégaoctet : 8 sacs de riz

Gigaoctet : 3 semi-remorques

Téraoctet : 2 porte-containers

Pétaoctet : Manhattan couverte de riz

Exaoctet : Côte ouest des USA couverte de riz

Zéttaoctet : couvrir l'océan pacifique de riz

**Yottaoctet : Remplir le volume de la terre de riz**



# DE LA DONNÉE AU BIG DATA

**Octet** : Grain de riz



**Kilo-octet** : Bol de riz



**Mégaoctet** : 8 sacs de riz

**Gigaoctet** : 3 semi-remorques



**Téraoctet** : 2 porte-containers

**Pétaoctet** : Manhattan couverte de riz



**Exaoctet** : Côte ouest des USA couverte de riz

**Zettaoctet** : Remplir l'océan pacifique de riz

---

**Yottaoctet** : Remplir le volume de la terre de riz



# DE LA DONNÉE AU BIG DATA

Octet : Grain de riz

Kilo-octet (3) : Bol de riz

Mégaoctet (6) : 8 sacs de riz

Gigaoctet (9) : 3 semi-remorques

Téraoctet (12) : 2 porte-containers

Pétaoctet (15) : Manhattan couverte de riz

Exaoctet (18) : Côte ouest des USA couverte de riz

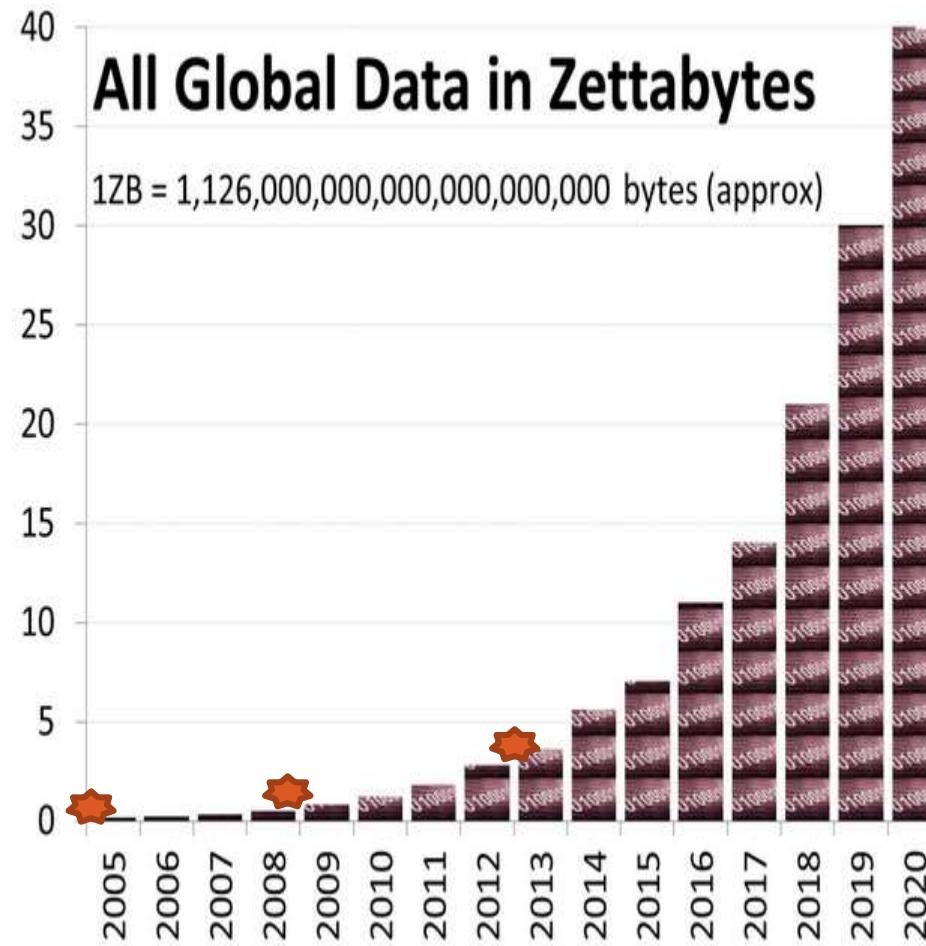
Zettaoctet (21) : Remplir l'océan pacifique de riz

Yottaoctet (24) : Remplir le volume de la terre de riz

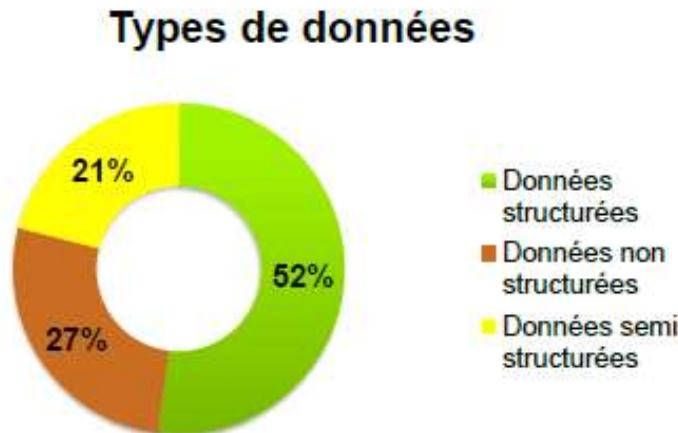


*Tous les 2 jours, nous  
générons 5 Exaoctets*

# DE LA DONNÉE AU BIG DATA



# DE LA DONNÉE AU BIG DATA



Les données non structurées vont croître **5 fois** plus vite que les données structurées.

(Source: Tata Consultancy Services)

La complexité des données augmente

# BIG DATA : POURQUOI MAINTENANT ?

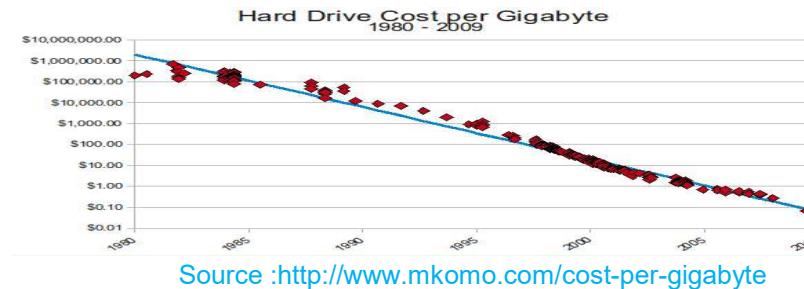
## 1. La puissance des micro-processeurs



L'Iphone est aussi puissant que les calculateurs qui existaient dans les années 70 ou 80.

## 2. Les coûts de stockage

Pour 1Go : de 1 000 000 \$ en 1980 à moins de 1\$ aujourd'hui



## 3. La vitesse des réseaux

De 9 600 bites/s à 10 Gb/s

**QUE LA FORCE  
DES DONNÉES  
SOIT AVEC TOI**

**GRÂCE À LA  
PLATEFORME  
BIG DATA**

- Les concepts du Data

# THE DATA AWAKENS

# THE DATA AWAKENS



*C'est une époque où la guerre des données fait rage*



Le Big Data permet d'explorer des données massives.

Le challenge est de réussir à transformer ces données massives en information qui guide les décisions et permet de créer de la valeur pour les entreprises.

Les clés de succès sont:

1. Une organisation et un processus adapté
2. Une connaissance même rudimentaire des techniques et méthodologies

Définition du « Big Data »

Big Data = Les 3V

# Volumétrie

Avant le Big Data



Big Data



# LES 3 DIMENSIONS DU BIG DATA

## Usage des technologies Big Data

Aujourd'hui, les **limites** du SI structurent le besoin



Une **croissance exponentielle** des données : tous les 2 ans les données doublent de volume, le TeraByte\* devient l'unité standard



# Variété

Avant le Big Data



Big Data



# LES 3 DIMENSIONS DU BIG DATA

## Usage des technologies Big Data

Aujourd'hui, les **limites** du SI structurent le besoin



Une **croissance exponentielle des données** : tous les 2 ans les données doublent de volume, le TeraByte\* devient l'unité standard

Une **multiplication des sources** (mobiles, puces RFID, réseaux sociaux, open data, etc.) et la prépondérance des **formats plus ou moins structurés**

# Vélocité

Avant le Big Data



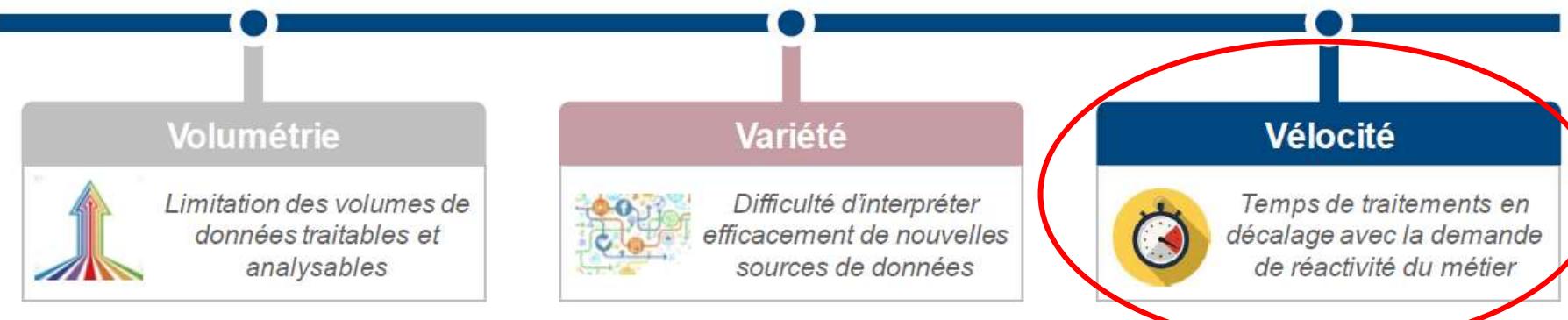
Big Data



# LES 3 DIMENSIONS DU BIG DATA

## Usage des technologies Big Data

Aujourd'hui, les **limites** du SI structurent le besoin

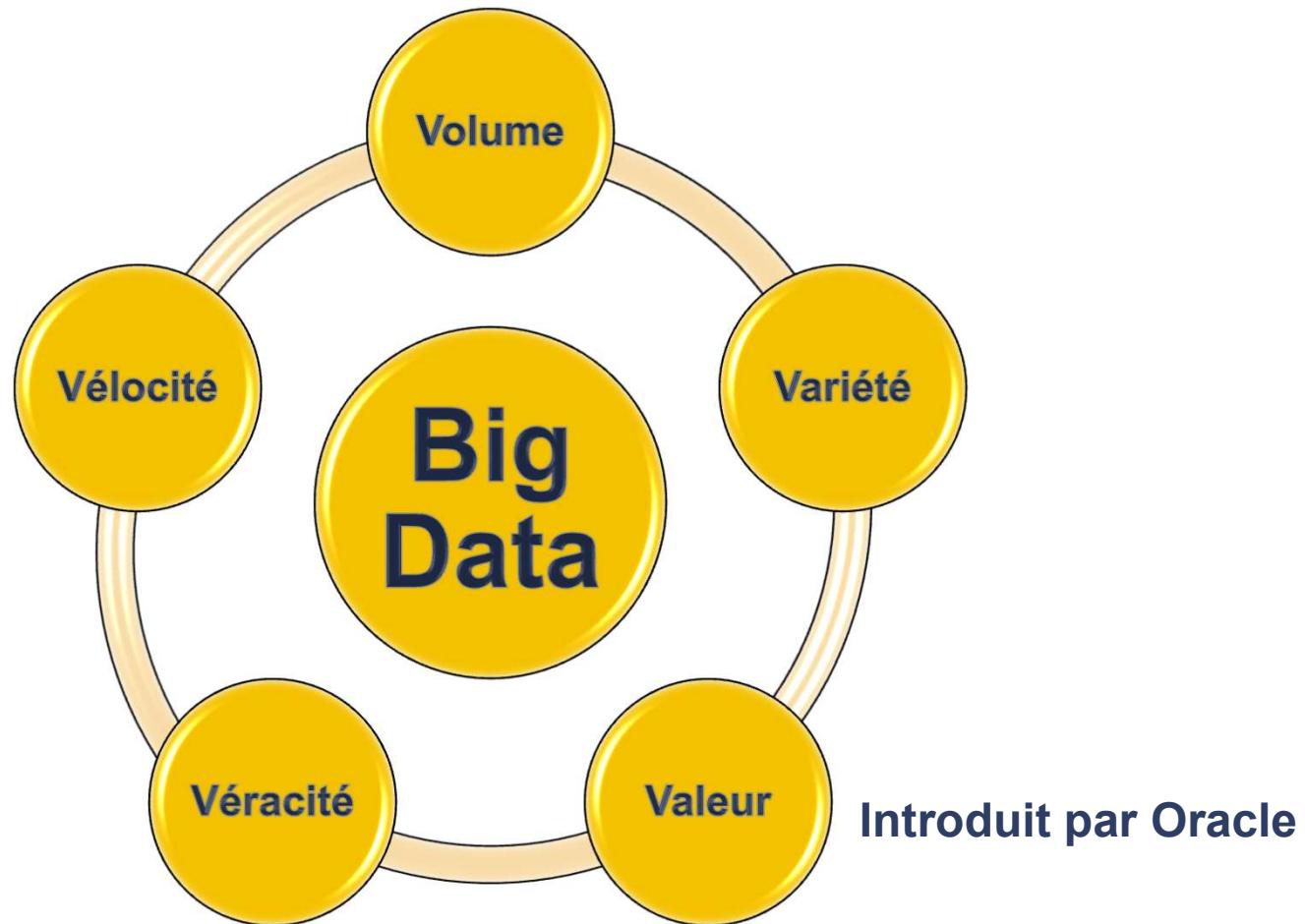


Une **croissance exponentielle des données** : tous les 2 ans les données doublent de volume, le TeraByte\* devient l'unité standard

Une **multiplication des sources** (mobiles, puces RFID, réseaux sociaux, open data, etc.) et la prépondérance des **formats plus ou moins structurés**

Une **fréquence élevée de diffusion, partage et traitement**

# **BIG DATA : LES 5 V**



# **BIG DATA : DÉFINITION**

Le Big Data désigne la problématique d'avoir un ensemble de données à traiter tellement volumineux qu'il devient très difficile, voire impossible, de le faire avec les outils existants.

wikipedia

# **BIG DATA : DÉFINITION**

Le Big Data est l'ambition de tirer un avantage économique de l'analyse quantitative des données internes et externes de l'entreprise.

*Cap Gemini*

# ENJEUX & OPPORTUNITÉS

# THE FORMULA 1 DATA JOURNEY

## Compare your data journey with Formula 1 telemetry

How will we make our data count? This simple question brings forth a huge number of considerations for an effective telemetry data journey.

In Formula 1 these include a wide variety of parameters:

- Intense pressure for real time analysis and response
  - Huge data rates
  - Available frequencies and acceptable latency  
  - System reliability in hostile environments
  - Regulatory constraints

COLLECT

How is the telemetry data acquired?

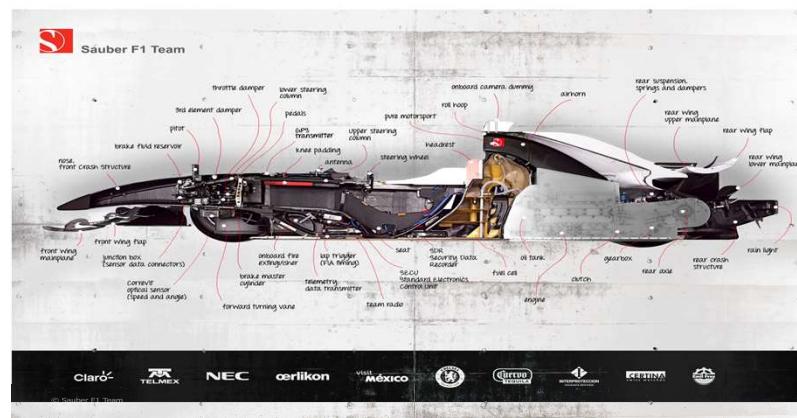
Each car has 120+ sensors. These measure everything from oil, water, exhaust and tyre temperatures to speed, engine revs per minute (RPM), clutch fluid pressure, G-force and even the driver's heartbeat.

The number of sensors isn't exact because from track to track and from practice to race they are added or removed depending on need.

Data is collected in the Electronic Control Unit which is 'the brain' of the car, it also acts as the primary logger recording more than 500 parameters. The data is sent via the telemetry antenna back to the pits in real time.

The cars also have a high performance logger with 1Gb memory taking up to 200 channels at a maximum sample rate of 1KHz per channel.

source:<https://www.metasphere.co.uk/telemetry-data-journey-f1/>



**Le « Big Data » est-il seulement une question de technologie, de collecte et traitement de la donnée ???**

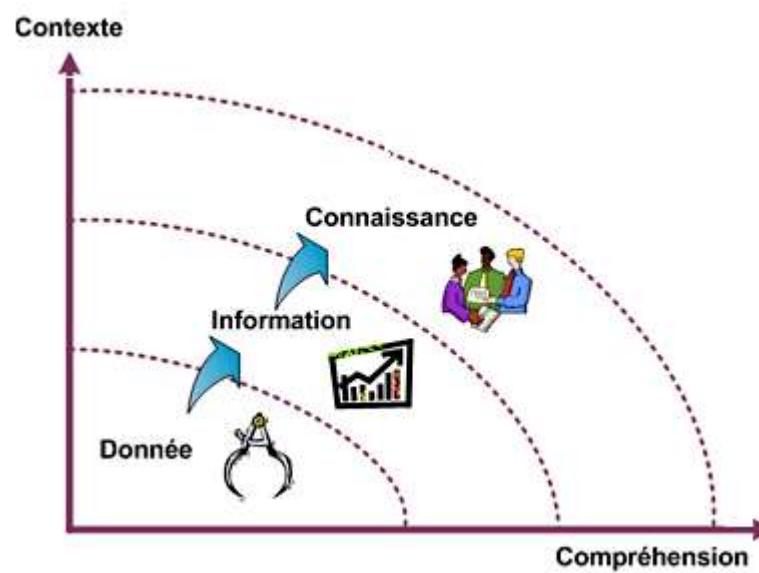
# LES OBJECTIFS PRINCIPAUX DES PROJETS BIG DATA

Gains	Entreprises ayant constaté un gain
Meilleure prise de décision	69%
Amélioration des processus opérationnels	54%
Amélioration de la connaissance client	52%
Réduction des coûts	47%

Source : <http://barcresearch.com/research/bigdatausecases2015/>

# DE LA DONNÉE, À L'INFORMATION, À LA CONNAISSANCE

- Une **donnée** est le résultat direct d'une mesure
- Une **information** est une donnée à laquelle un sens et une interprétation ont été donnés
- La **connaissance** est le résultat d'une réflexion sur les informations analysées



# PLUSIEURS TYPES DE DONNÉES

Données qualitatives

Données quantitatives

Données catégorielles

Données discrètes

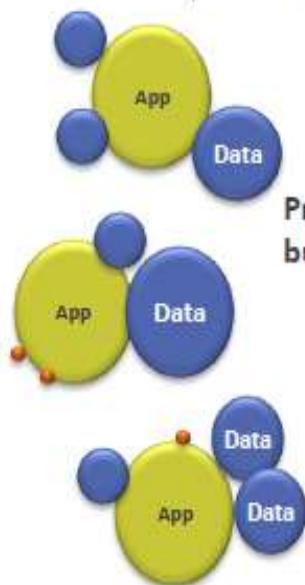
Données continues



# ...ET SUR LE MODE DE GESTION DE LA DONNÉE

## Ce que nous faisions avant

Copier les données vers les applications

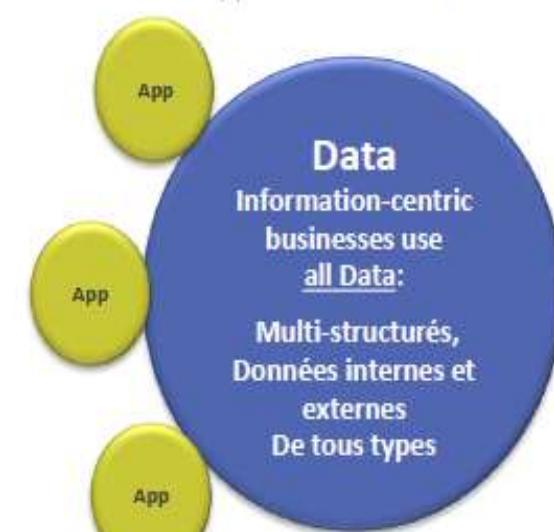


### Process-centric businesses use:

- Données structurées principalement
- Données internes seulement
- Données «importantes» uniquement
- Plusieurs copies de données

## Ce que nous faisons maintenant

Porter les applications aux données



# **BIG DATA & MARKETING**

En plein Révolution Numérique « Digitalisation »

Phénomène d'Uberisation



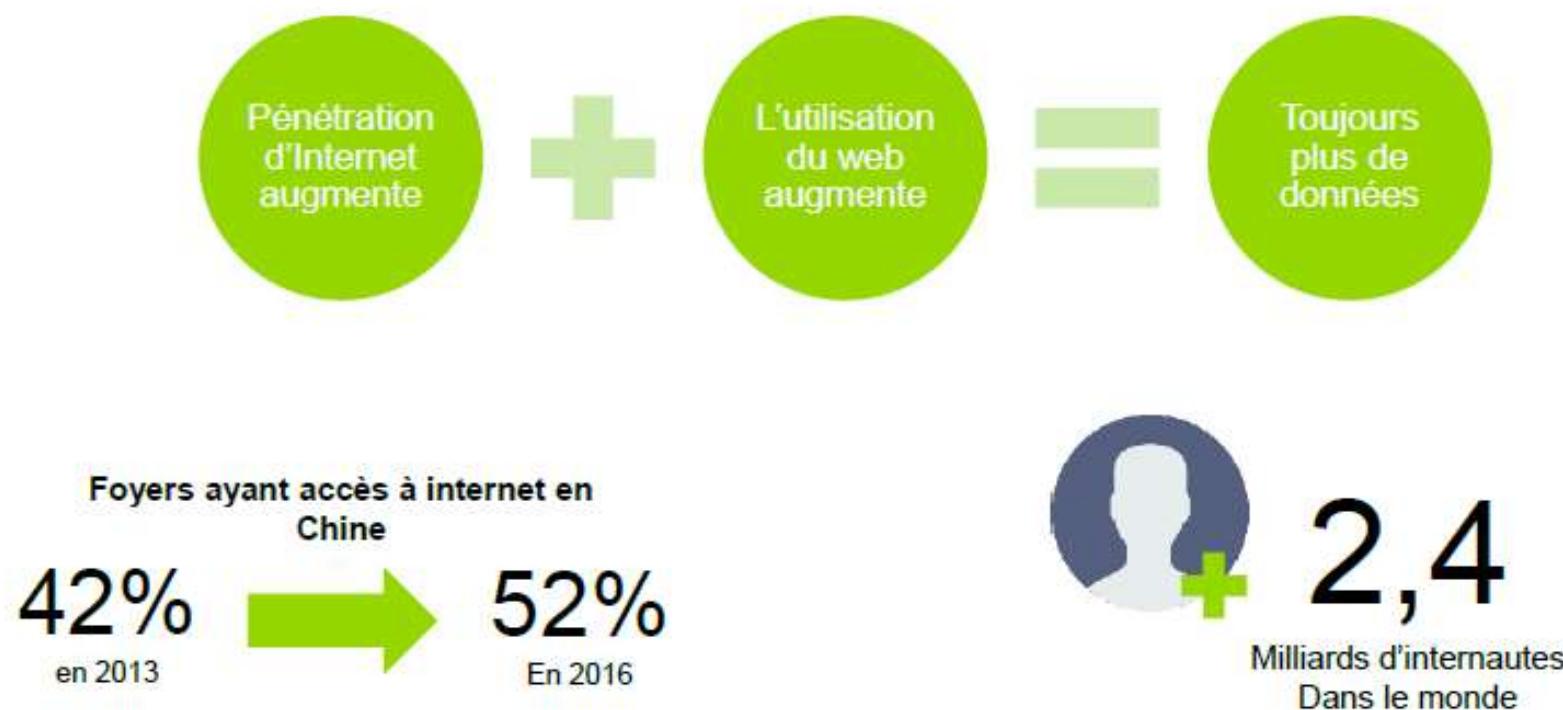
# LA REVOLUTION NUMERIQUE

Les nouvelles technologies ont une incidence forte sur nos vies quotidiennes. Nos réflexes d'achat en ligne ou d'échange de biens, notre consommation de l'information, transforment nos vies.

Les changements induits par internet sur l'information, la politique ou la participation citoyennes ne laissent personne indifférent : des régimes les plus autoritaires, soucieux d'encadrer et de verrouiller la pratique des réseaux sociaux à ceux qui feront le choix de les détourner pour mieux porter la propagande d'État, les usages sont aussi pervers qu'utopistes et nous rappellent à la nécessité de nous garder de tout angélisme face à un changement qui, comme dans toute révolution porte ses espoirs et sa part d'ombre.

*Rémy Rieffer*

# AUGMENTATION DES ACCES INTERNET



# BIG DATA & MARKETING

Relation client Multicanal



# BIG DATA & MARKETING

“ Il n'y a qu'un patron : Le client, il peut licencier tout le personnel, depuis le directeur jusqu'à l'employé, tout simplement en allant dépenser son argent ailleurs ” - *Sam Walton*

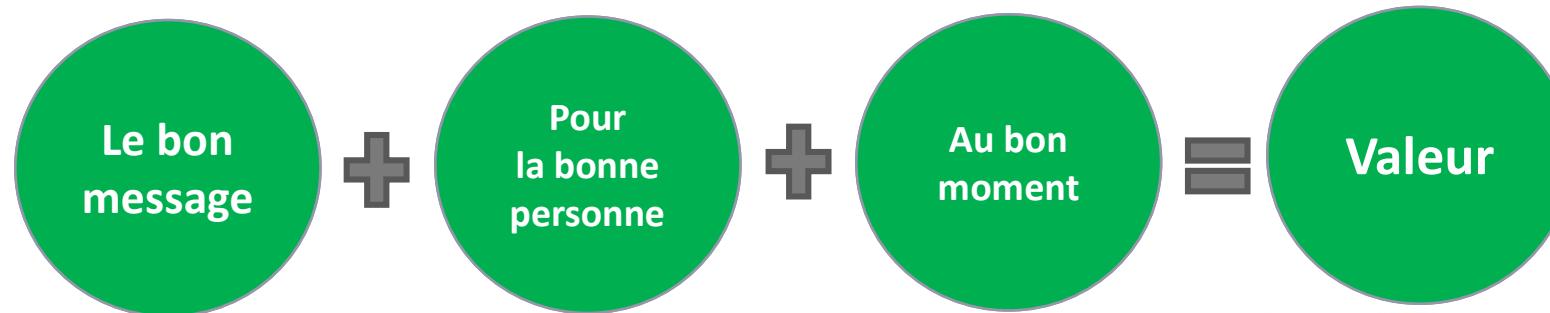
# BIG DATA & MARKETING

**Importance de la mise en place d'outils Big Data  
autour des données**

*Degré d'importance des données client en fonction de leur origine*

infographie IDC réalisée par  
l'Observatoire Big Data en  
France :

# BIG DATA & MARKETING



# BIG DATA & MARKETING

- Le Big data explore déjà de nouveaux usages dans plusieurs spécialités comme :

- Le marketing relationnel avec la segmentation et le ciblage plus fins des clients (les plus rentables, les plus à risque...)
- Le web analytics avec l'optimisation des parcours en ligne.
- Le marketing prédictif avec l'anticipation des besoins et des évolutions.

# BIG DATA & MARKETING

- Le Big Data va contribuer à transformer le marketing en améliorant ses capacités d'exécution :

- Accroître la réputation de la marque.
- Implémentation de promotions ou d'offres ciblées
- Anticiper les comportements plutôt que réagir aux situations.
- Comprendre chaque client dans son unicité.
- Développer les produits de demain.
- Explorer toutes les informations clients disponibles.

# BIG DATA & MARKETING

## Revenu publicitaire sur son site

- Optimiser la sélection des annonces en fonction des profils
- Identifier et bannir les robots



# BIG DATA & MARKETING

- « *Un client est à respecter sous peine de sanction immédiate. Submergé de messages plus ou moins bien ciblés le client est devenu exigeant. Il vous a donné la permission d'être en contact avec lui respectez la ! Demandes lui par quel canal et combien de fois vous pouvez lui envoyer un message, et surtout ciblez les messages... » Seth Godin*

# **SOURCES DE DONNÉES**

# LES SOURCES DE DONNÉES

Les données internes :

- ERP
- CRM
- Billing
- ...

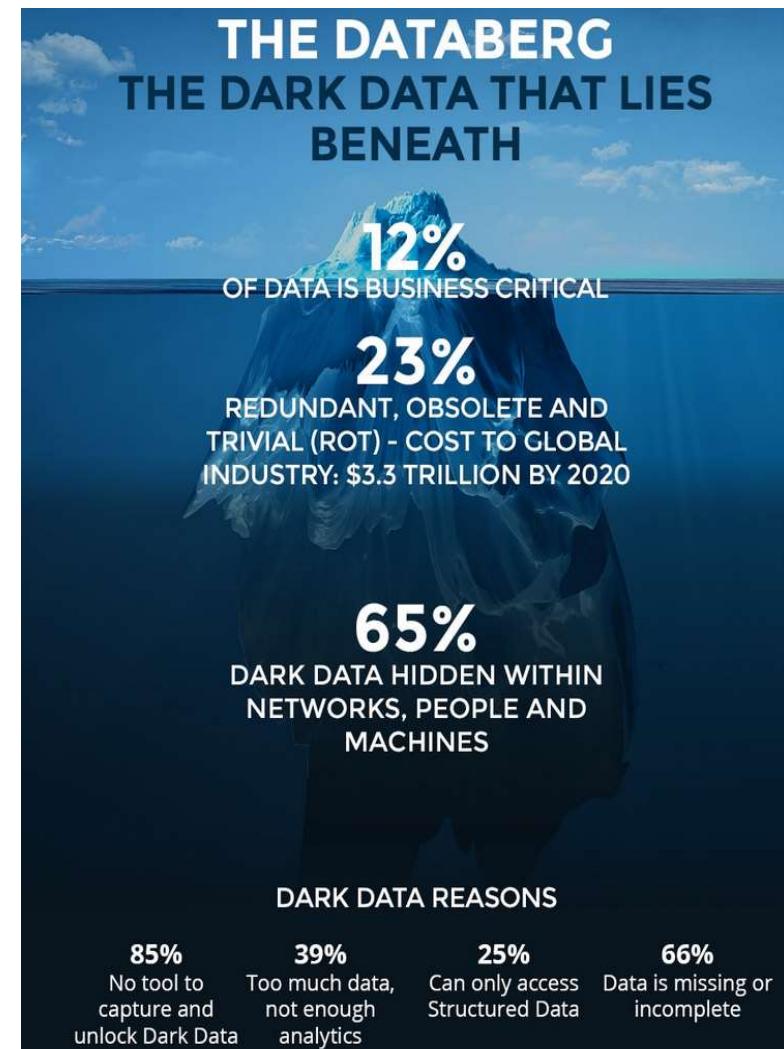


# LES SOURCES DE DONNÉES

Dark Data : Les données internes non exploitées

- Mail
- Rapport
- Log,...

« données potentiellement utiles qui pourraient être obtenues à partir de processus métier, mais ne sont actuellement pas mises à profit »



# LES SOURCES DE DONNÉES

- Web / Réseaux sociaux / Mobile :

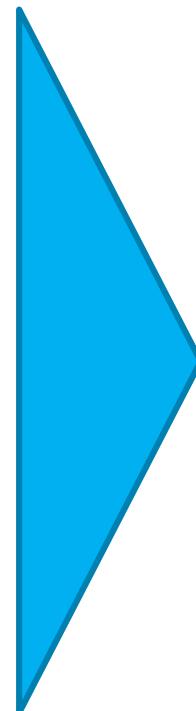
- Facebook
- Tweets
- Géolocalisation
- ClickStream



# IOT



Les objets connectés deviennent les plus gros producteurs de données



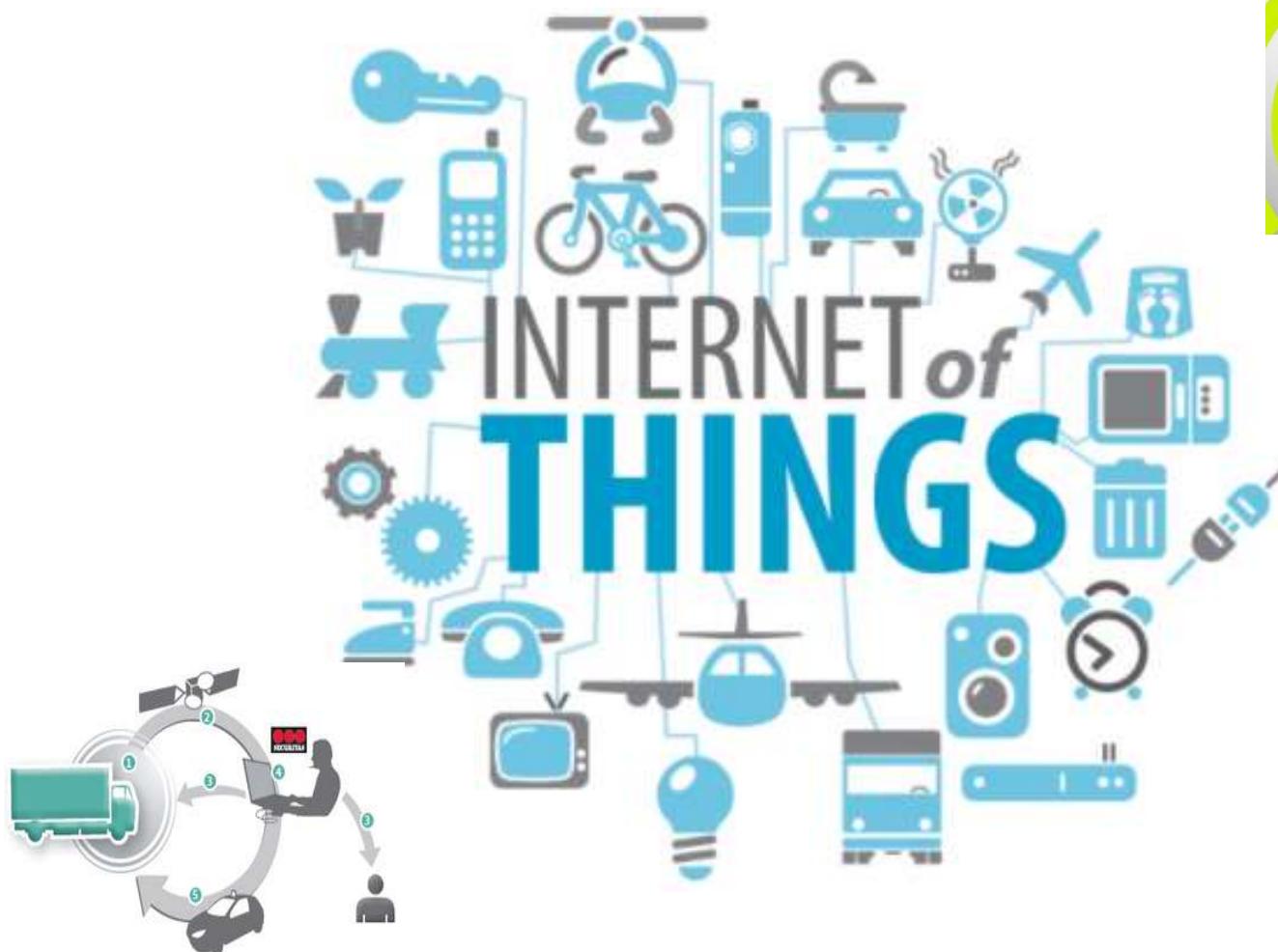
8,4  
Mds d'IOT

+ 31%

2016 vs 2018

*Source: Gartner*

# IOT



# LES SOURCES DE DONNÉES

- Les données externes :

- INSEE ( Données juridiques )
- Météo
- Entreprises privées



# **LES SOURCES DE DONNÉES : OPEN DATA**

**Publication en 2009 de W3C sur l'ouverture des données**

- Participation
- Collaboration
- L'amélioration
- L'innovation

**La puissance publique cherche de la croissance.**

- Les données sont un relai de croissance immense

**Ouverture des données publiques pour une réutilisation et une  
création de valeur grâce à l'industrie numérique**

# **LES SOURCES DE DONNÉES : OPEN DATA**

- **Données collectées par des organismes publics**
- **Données non-nominatives**
- **Données ne relevant pas de la vie privée ni de la sécurité**
- **Données sont libres de droit**

# **LES SOURCES DE DONNÉES PUBLIQUES FRANÇAISES: OPEN DATA**

- INSEE
  - EUROSTAT
  - Banque de France
  - Cabinet du Premier Ministre
  - Ministères
  - ARCEP (télécommunications)
  - BNF
  - Villes et conseils régionaux
  - IGN
  - La Poste
  - SNCF
  - Comité National Routier
  - Caisse d'Allocations Familiales
  - Office National des Forêts
  - AIRPARIF
  - Réseau National de Surveillance Aérobiologique (RNSA)
  - Institut National d'études démographiques
  - ....

Les principaux portails de données  
[data.gouv.fr](http://data.gouv.fr), [api.gouv.fr](http://api.gouv.fr), [data-publica.com](http://data-publica.com)



# **LES SOURCES DE DONNÉES : OPEN DATA**

<https://youtu.be/sXICdsQZcDA>

# LES SOURCES DE DONNÉES MARKET PLACE



Infochimps Data Marketplace (acquired by CSC)



Factual: Data on over 600K consumer packaged goods, ingredients and nutrition information



FIND AND UNDERSTAND DATA

Microsoft Azure  
Marketplace

DataMarket: Provides Open Data sets by country, industry and provider

Microsoft Azure Marketplace



DataSift: Data aggregated from various social media sources like Twitter, Facebook,  
WordPress blogs...



Twitter: Streaming API, Search and Firehose (service offered by DataSift and GNIP)



GNIP: Complete Twitter data



Driving Intelligence **Inrix: Data for automotive and transport traffic**

# LES SOURCES DE DONNÉES : OPEN DATA



Y A D Frite!

## 2 TYPES DE PROJETS BIG DATA

- **Projets de Data Science** = Datamining 2.0, Big Data Analytics
  - Projets métiers, avec des innovations / explorations fortes
  - Approche « Test & Learn », plutôt que des grands cadrages
  - Le sourcing de données est au cœur de l'expérience
- **Projets SI capacitaires** : Augmentation de la capacité en réduisant les coûts des systèmes d'informations
  - Projets DSI : transformation des systèmes décisionnels et de traitements de données
  - Intégration du « décisionnel » aux systèmes opérationnels, en collecte et restitution (indicateurs, rapports, scores, prédictions)
  - Permettre la réalisation de projets de Data Science au sein du SI

# COÛT D'UN PROJET BIG DATA

Appliance Datawarehouse  
traditionnelle

60 K€ / To

versus

Plateforme Big Data  
Hadoop Open Source

10 K€ / To  
(à périmètre comparable)

*Source Capgemini*

# CAS D'ETUDE

**Scénario :** Vous êtes fournisseur de gaz. Vous remarquez qu'un nombre important de facture ne sont pas recouvrées par les clients B2B lorsqu'un client quitte son local

En plus de perdre de l'argent car le client est parti sans payer sa dernière facture, vous perdez aussi de l'argent car le nouveau professionnel dans le local utilise votre gaz sans avoir contracté avec vous. Pour peu qu'il décide de contracter avec un autre fournisseur, le manque à gagner est encore plus important.

Votre objectif est de limiter les impayés et de contractualiser au plus tôt avec les nouveaux propriétaires ou locataires:

**Action :**

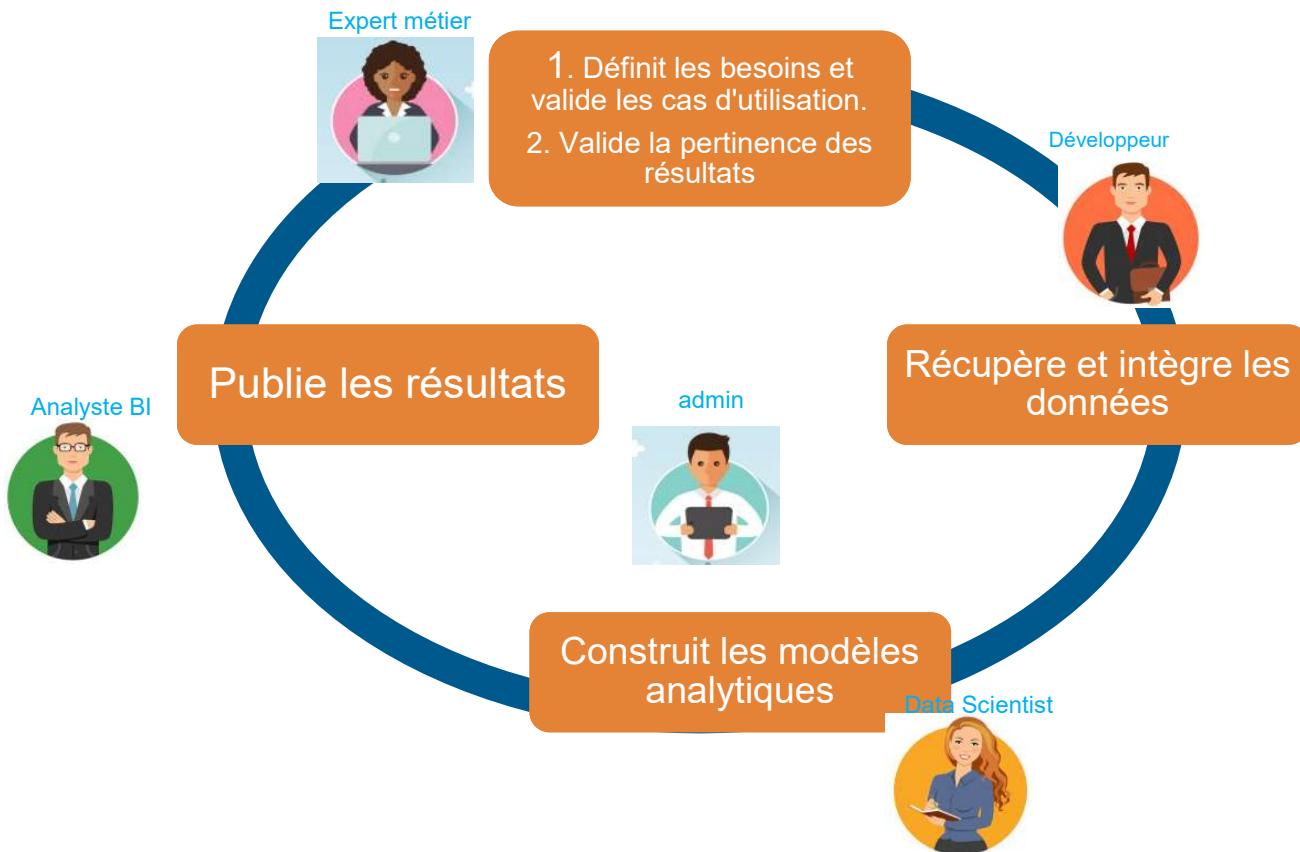
- Quelles sont les données internes et externes dont vous avez besoin ?
- Expliquer comment utiliser les différentes données pour répondre à votre besoin.

# CAS D'ETUDE

**Scénario :** Je suis un des leaders dans le télécoms et je souhaiterais monétiser mes données

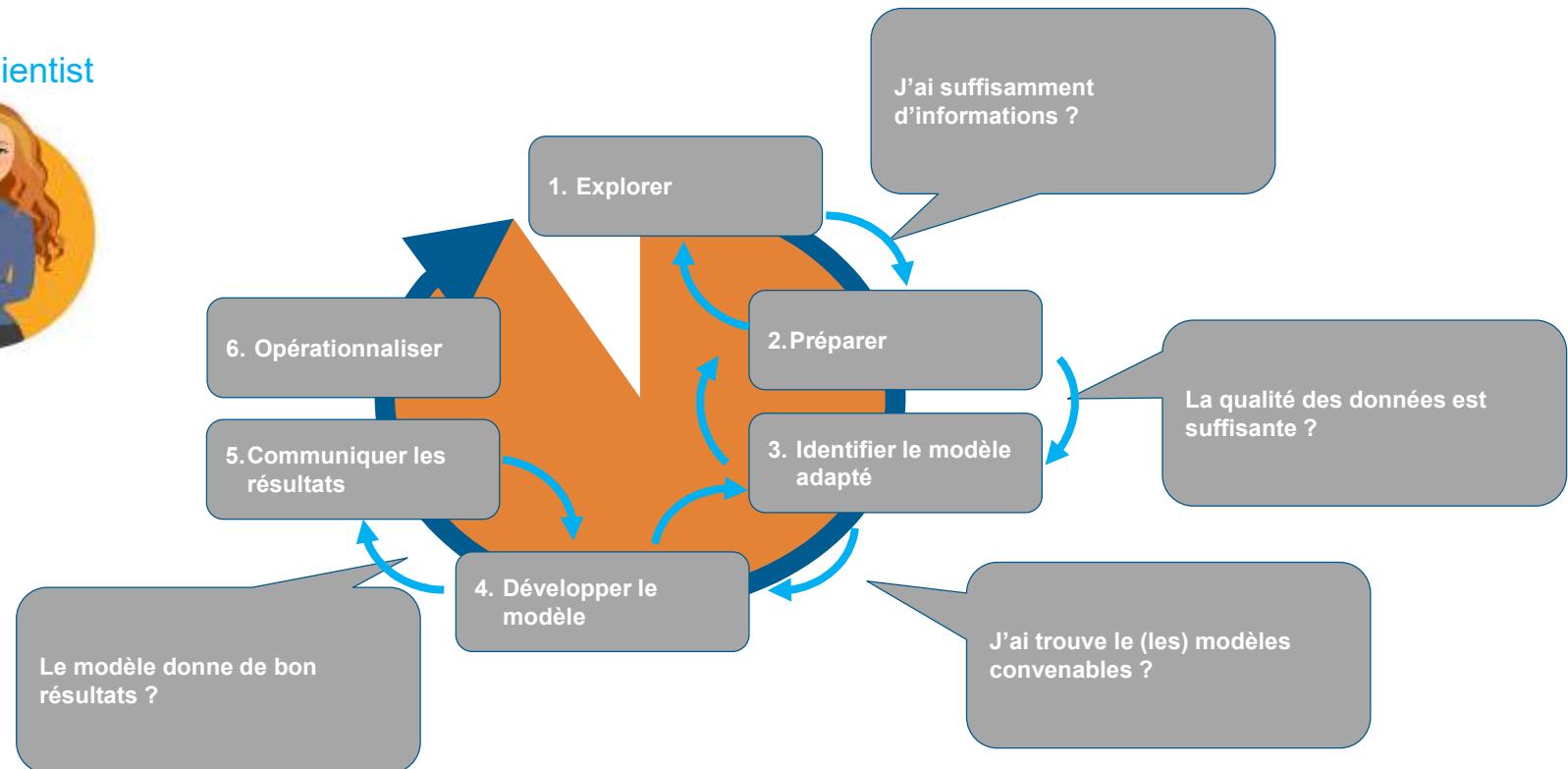
Pour quels cas d'usage et donc pour quelles industries, mes données pourraient avoir de la valeur ?:

# PROJET DATASCIENCE

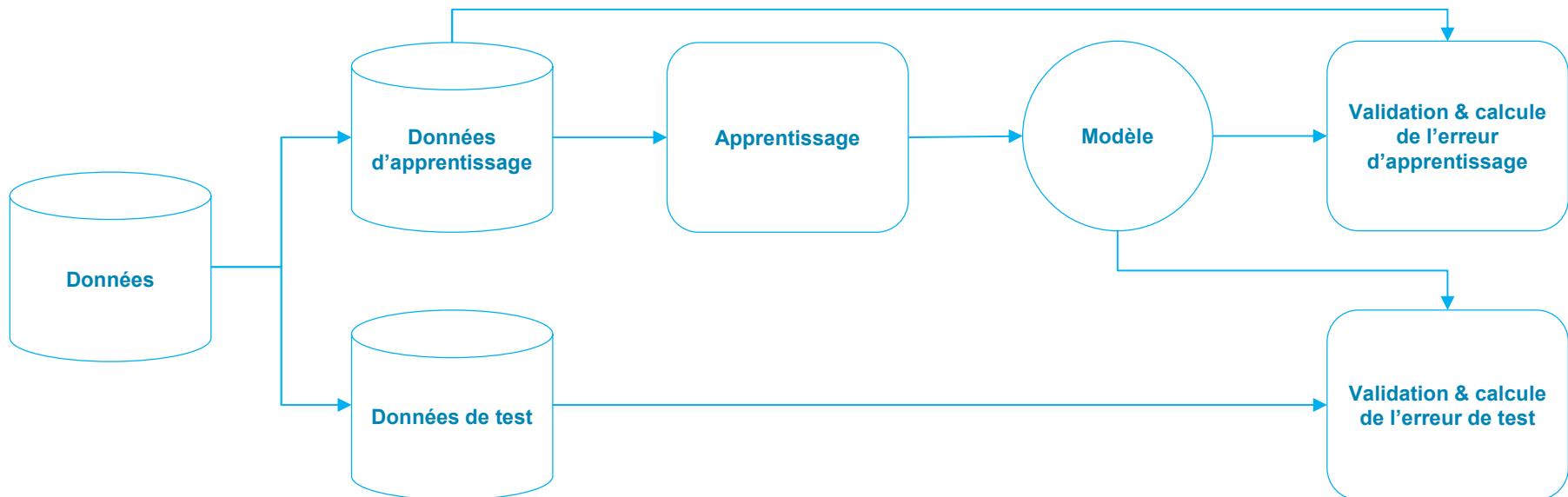


# PROCESSUS DE LA MODÉLISATION ANALYTIQUE

Data Scientist

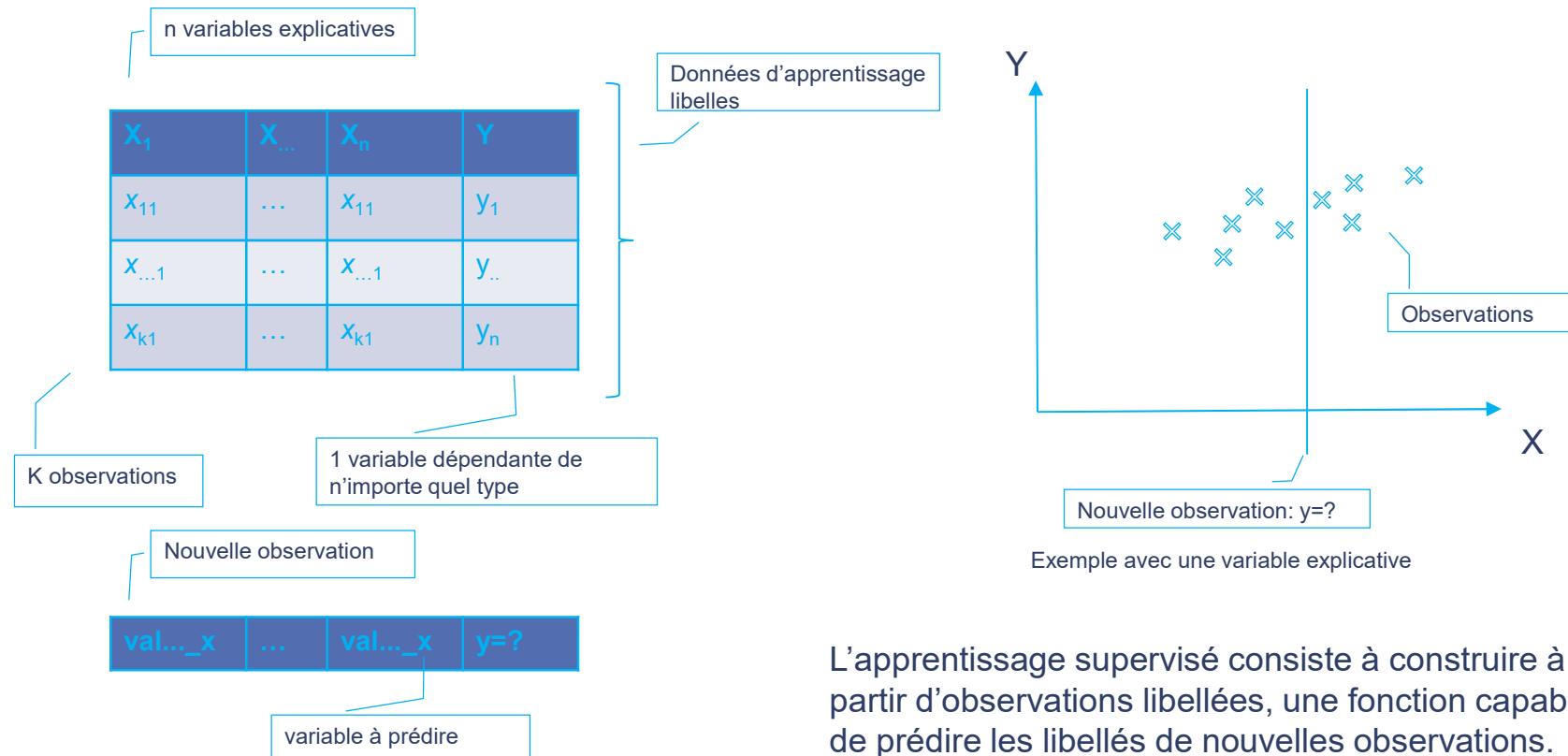


# PROCESSUS ET VALIDATION D'UN MODÈLE ANALYTIQUE(APPRENTISSAGE SUPERVISÉ)



Un modèle doit être paramétré de manière à réduire l'erreur sur les données de test et savoir gérer de nouvelles données également.

# APPRENTISSAGE SUPERVISÉ



# APPRENTISSAGE SUPERVISÉ



Iris Setosa



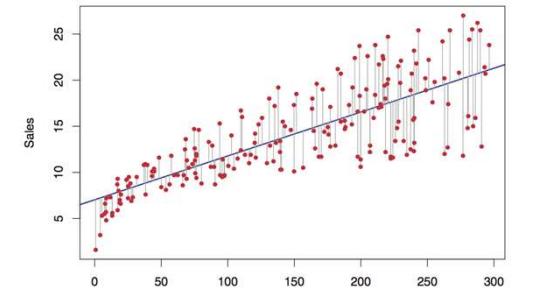
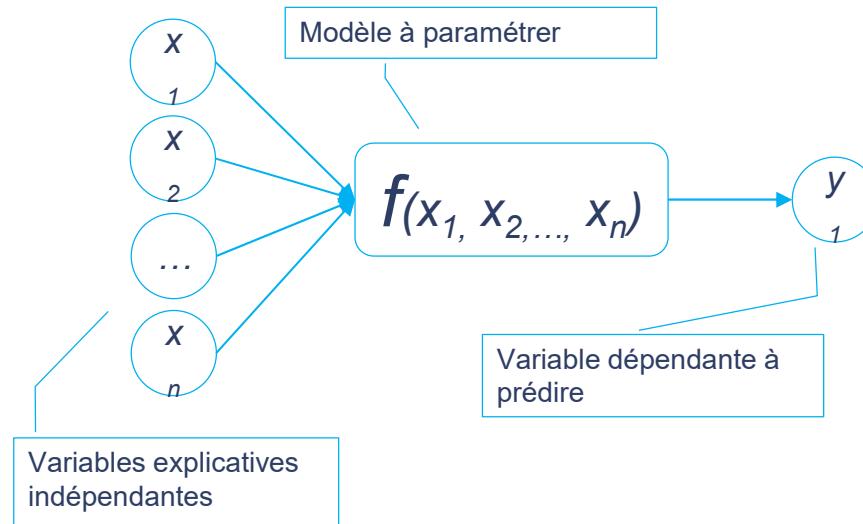
Iris Versicolor



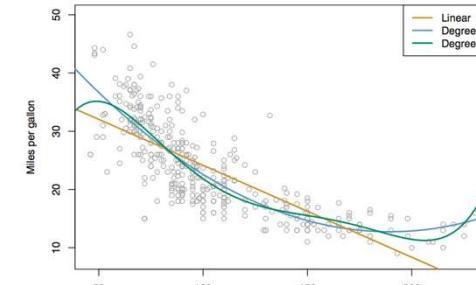
Iris Virginica

longueur des sépales	largeur des sépales	longueur des pétales	largeur des pétales	Espèce
5.1	3.5	1.4	0.2	I. setosa
4.3	3.0	1.1	0.1	I. setosa
5.5	2.3	4.0	1.3	I. versicolor
4.9	2.4	3.3	1.0	I. versicolor
4.9	2.5	4.5	1.7	I. virginica
5.8	2.8	5.1	2.4	I. virginica
7.2	3.2	6.0	1.8	?

# APPRENTISSAGE SUPERVISÉ : RÉGRESSION



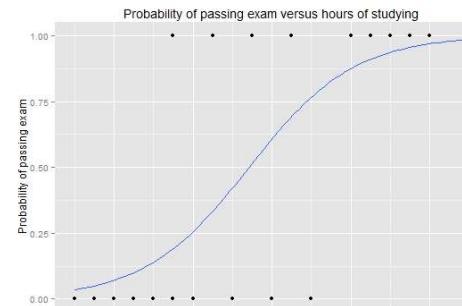
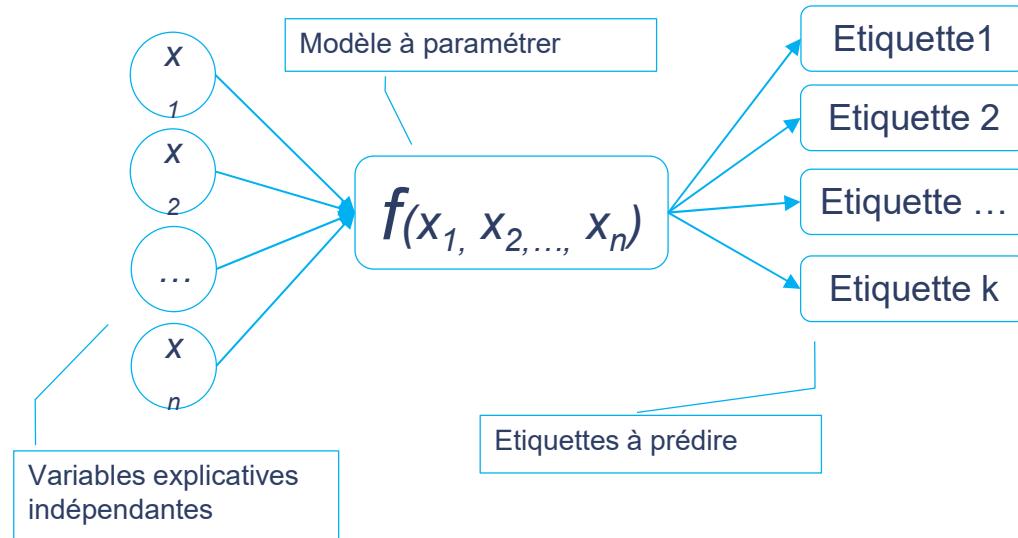
Régression linéaire  $y = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$



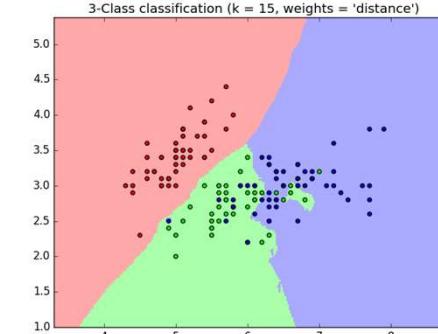
Régression polynomiale  $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \dots + \theta_n x_n^n$

La régression consiste à prédire une variable continue à partir de variables explicatives.  
Les paramètres  $\theta$  de la fonction  $f$  sont identifiés par apprentissage de l'historique des observations.

# APPRENTISSAGE SUPERVISÉ :CLASSIFICATION



Classification logistique (binaire) – source wikipedia



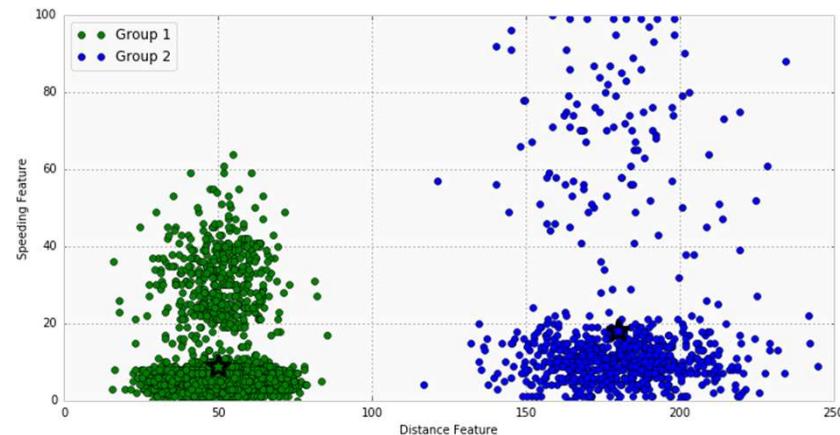
Classification Nearest Neighbor (Multiclasse) source scikit learn

La classification consiste à produire une étiquette à partir d'une observation.

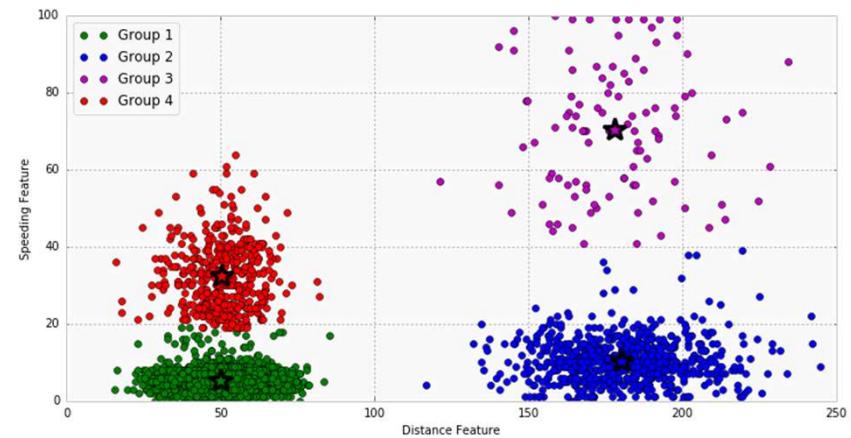
Les paramètres de la fonction  $f$  sont identifiés par apprentissage de l'historique des observations.

Sources des images: The elements of statistical learning

# APPRENTISSAGE NON SUPERVISÉ



K-Means: k=2



K-Means: k=4

La segmentation consiste à identifier des critères communs entre les observations pour ensuite les organiser par groupes.

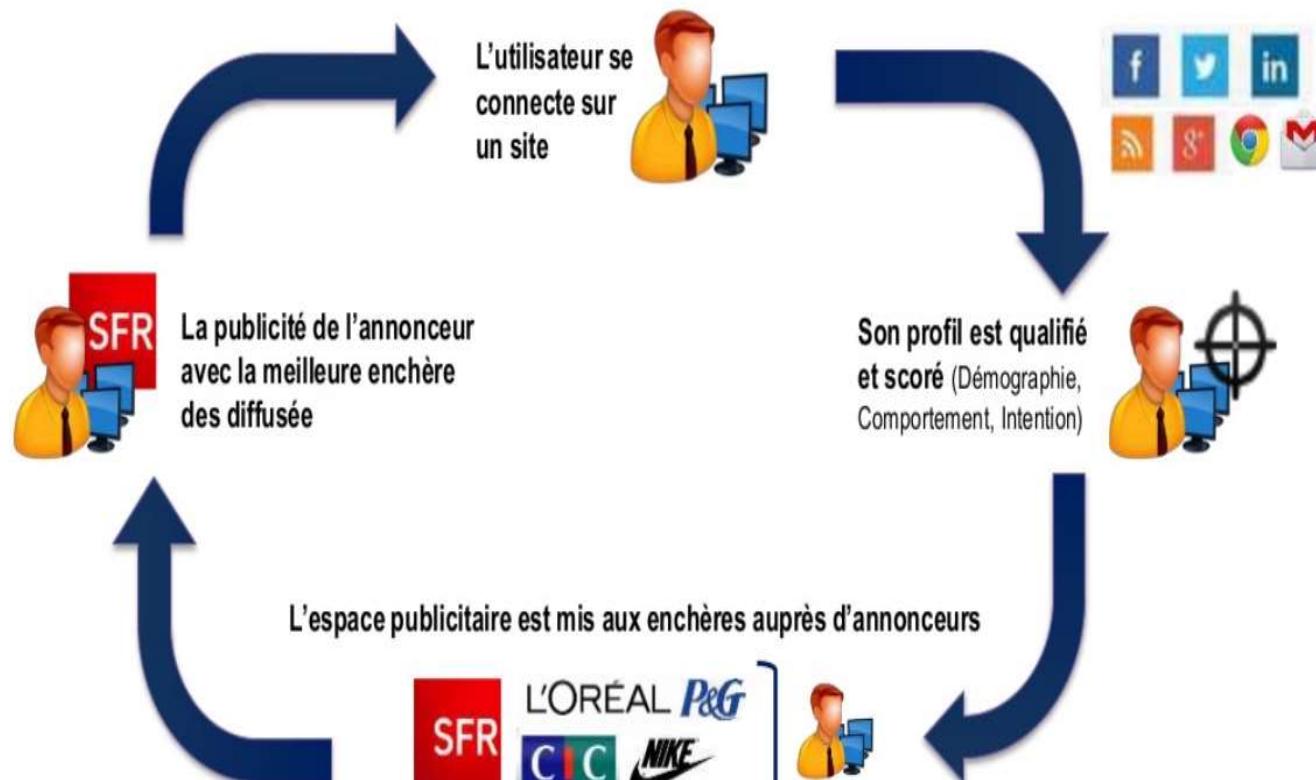
Source: <https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data-science-tutorials>

# LES USAGES



# USE CASE – REAL TIME BIDDING

RTB : Enchères en temps réel pour la diffusion de publicités display (exemple : publicités youtube)



# USE CASE – CINÉMA

Prédire l'impact de la sortie d'un nouveau film MAN 3

Réponse aux questions :

- ✓ Est-ce que le film est attendu ?.
- ✓ Quels sont les premiers ressentis à la sortie du film?



# USE CASE – MARKETING

- Comment le département marketing du supermarché a développé son outil de prédiction de grossesse :



- Analysant les données démographiques et historiques d'achat des millions de clients
- Identifier une liste de 25 produits que les femmes enceintes sont plus susceptibles d'acheter
- Score de prévisibilité de grossesse



80

# BIG DATA : COMMENT DÉCONGESTIONNER LE TRAFIC URBAIN ?

Le «Big Data» et les technologies numériques révolutionnent l'organisation de la ville.

GPS

App  
Géolocalisation

Gestion du  
trafic



Stationn-  
ement

Transport  
public

Smart City

# LA BIG DATA AU SERVICE DE BARACK OBAMA

## Le pouvoir du "Big data" : Obama premier Président élu grâce à sa maîtrise de traitement de données ?

Le "Big data", la collecte et le traitement automatisé de gigantesques quantités de données, a aidé Barack Obama à remporter un second mandat. Une méthode déjà utilisée depuis longtemps dans le monde de l'économie et de la finance.



Je m'abonne  
à partir de 4,90€

Ajouter au classeur

Suivre ce contributeur

Lecture zen

A+ A-

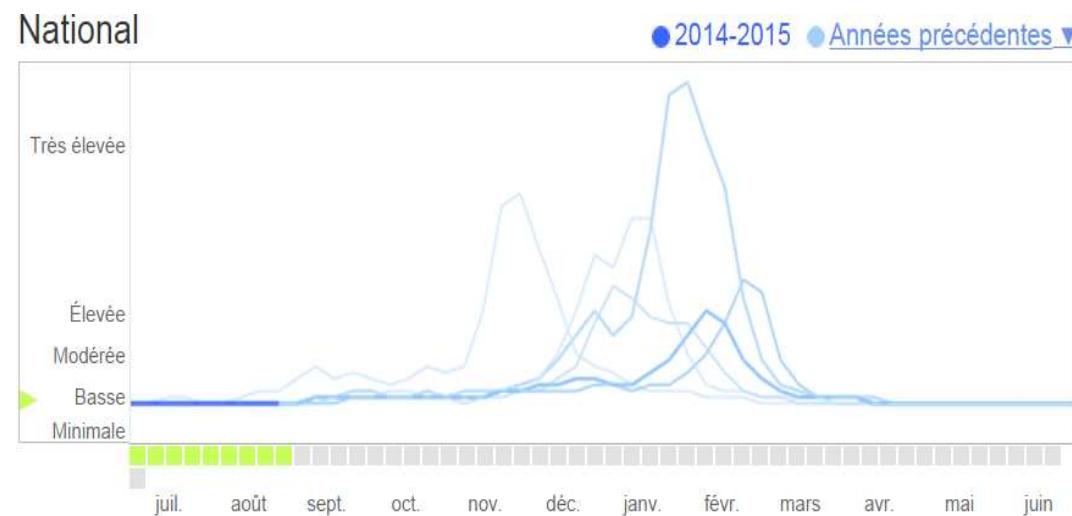


# USE CASE – SANTÉ

**Pouvons-nous aujourd'hui prévoir une épidémie de grippe?**

Google Flu Trends

1. Analyse des données des recherches effectués sur Google



# USE CASE - BANQUE

## Challenges métiers

- Amélioration de la détection de fraude sur les transactions bancaires en temps réel.
- Moteur de règles archaïque.
- Trop de "fausses alarmes"
- Besoin d'une expertise externe solide et long terme sur le deep learning et l'architecture big data

## Solution

- Construire un modèle prédictif et une base pour le scoring temps réel..
- Déployer une solution analytique de type deep learning pour améliorer les performances
- Think Big Analytics est devenu un partenaire de confiance et une locomotive pour l'implantation de future projets IA au sein de l'entreprise.

## Avantages

- Economie de plusieurs millions de \$ chaque mois.
- Fausses alarmes réduites de 50% et détection améliorées de 60%.
- Mise en production du plateforme analytique temps réel.
- Détection de fraudes accélérée

## Outils et technologies



TERADATA.

# USE CASE - TELECOM

Orange Business Services

Améliorer la proactivité vers les clients



- 220 pays couverts en réseau, 29 pays couverts en mobiles
- 156 000 collaborateurs
- 263 millions de clients dans le monde



## Enjeux

- Anticiper la détection des anomalies de 350 000 équipements réseau (routeurs, switchs...)
- Améliorer la proactivité vers les clients
- Optimiser les plannings d'équipes d'interventions

## Solution

- Prestations de Service Data :
- Développement d'algorithmes de machine learning pour prédiction en « temps réel »
- Équipe multidisciplinaire (Chef de Projet métier, Architecte big data, Data Scientist, Data Engineer, Chef de projet Data)
- PoC -> étude architecture -> pré-Prod -> Prod
- Technologies : Hadoop et Spark

## Bénéfices

- > 80% des anomalies « sérieuses » détectées
- 50% réduction délai de traitement
- Industrialisation en cours

# USE CASES

## Industrie



- Produit comme un service
- Qualité, innovation R&D
- Maintenance préventive

## Assurance



- Fraudes et risques
- Connaissance client, Vision 360°, scoring temps-réel
- Recommandation client
- Tarification personnalisée

## Secteur public



- Services informationnels
- Fraudes, abus
- Sécurité publique

## Distribution



- Offres temps réel et service personnalisés
- Optimisation de l'expérience magasin
- Pricing dynamique

## Santé



- Gestion des effets indésirables
- Traitements personnalisés.
- Amélioration des diagnostics

## Telecom



- Parcours clients multi-canaux
- Partage de données de géo localisation
- Fraudes et analyse du comportement client

## Banques



- Parcours clients multi-canaux
- Fraude, anti blanchiment
- Partage des données consommateurs pour personnalisation

## Transports, loisirs



- Planification et gestion des evts liés à la logistique
- Service client temps réel
- Economie d'énergie

## Produits gde conso.



- Analyse de sentiments et retour produits
- Relation personnalisée avec le consommateur

# QUELQUES EXEMPLES PRATIQUES DE MACHINE LEARNING

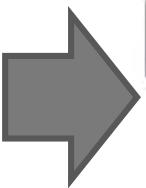
Notes  
Personal  
Travel  
Work  
Moins ▲  
Tous les chats  
Tous les messages  
**Spam (142)**  
Corbeille

<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Lesproteines.com	La grande nouveauté de la rentrée 2016 - pré commande ouverte - livraison à partir du 1... - Visualis
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Fenêtres sur mesures	Vos 3 devis offerts pour des fenêtres qui isolent mieux - Veuillez ajouter info@leboninvestissement4.
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Marie de l'équipe OuiCar	Un OuiCar encore plus simple ! - Application, déménagements... nous allons vous changer la vie ! OUI
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	finanzen.fr	Investissez dans l'immobilier et effacez vos impôts pendant 12 ans - ✓ Achetez un bien neuf dans ur
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Casmio	20 Spins just for registering - Start with 20 Spins If you don't wish to receive further emails you can uns
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Lesproteines.com	Jusqu'à lundi: exceptionnel Creapure offerte avec chaque Instant Whey Reflex (2,2 et 4,... - Visualis
<input type="checkbox"/>	<input type="star"/>	<input type="checkbox"/>	Voyage Privé	Junior Suite de luxe en Jamaïque, city-break dans quartier chic à Rome, escapade de cha... - Elles

SPAM

# QUELQUES EXEMPLES PRATIQUES DE MACHINE LEARNING

Les clients ayant acheté cet article ont également acheté



HDMI adaptateur -  
TUROBOT Chipset HDMI  
mâle vers VGA femelle  
Cordon vidéo Câble...

★★★★★ 75

EUR 6,99 ✓Premium

Patuoxun 1080p  
Thunderbolt Mini  
Displayport vers adaptateur  
VGA haute qualité pour...

★★★★★ 101

EUR 7,59 ✓Premium

Patuoxun Mini Display Port  
Thunderbolt vers DVI/  
VGA/ HDMI/ TV/ AV/  
HDTV - Adaptateur Câble...

★★★★★ 28

EUR 12,55 ✓Premium

Film VERRE TREMPE 9H  
pour Wiko Ridge 4G Ultra  
Transparent Ultra Résistant  
INRAYABLE INVISIBLE

★★★★★ 120

EUR 6,98

Intel Haswell Processeur  
Pentium G3220 3.00 GHz  
3Mo Cache Socket 1150  
Boîte (BX80646G3220)

★★★★★ 30

EUR 64,38

# SES APPLICATIONS

Maintenance prédictive

Fraude

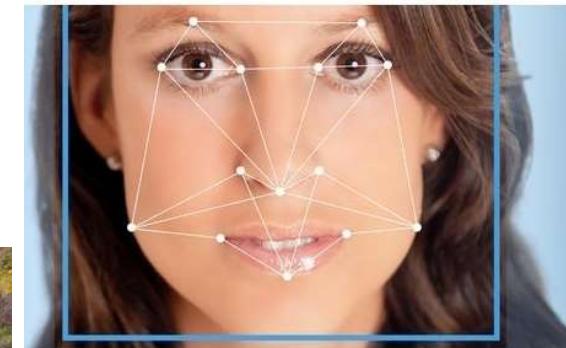
Reconnaissance faciale

Voiture autonome

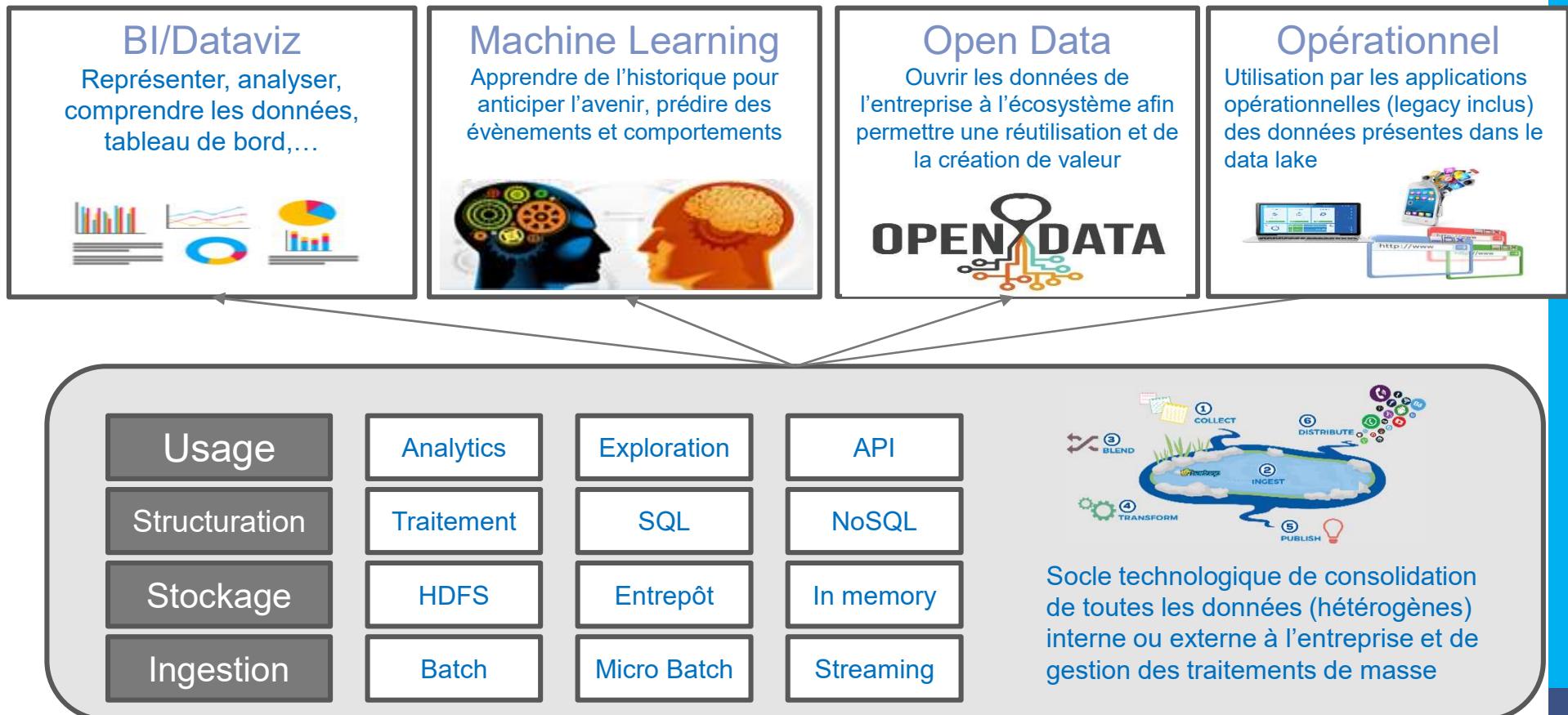
Diagnostic médical

Création de commerce

...

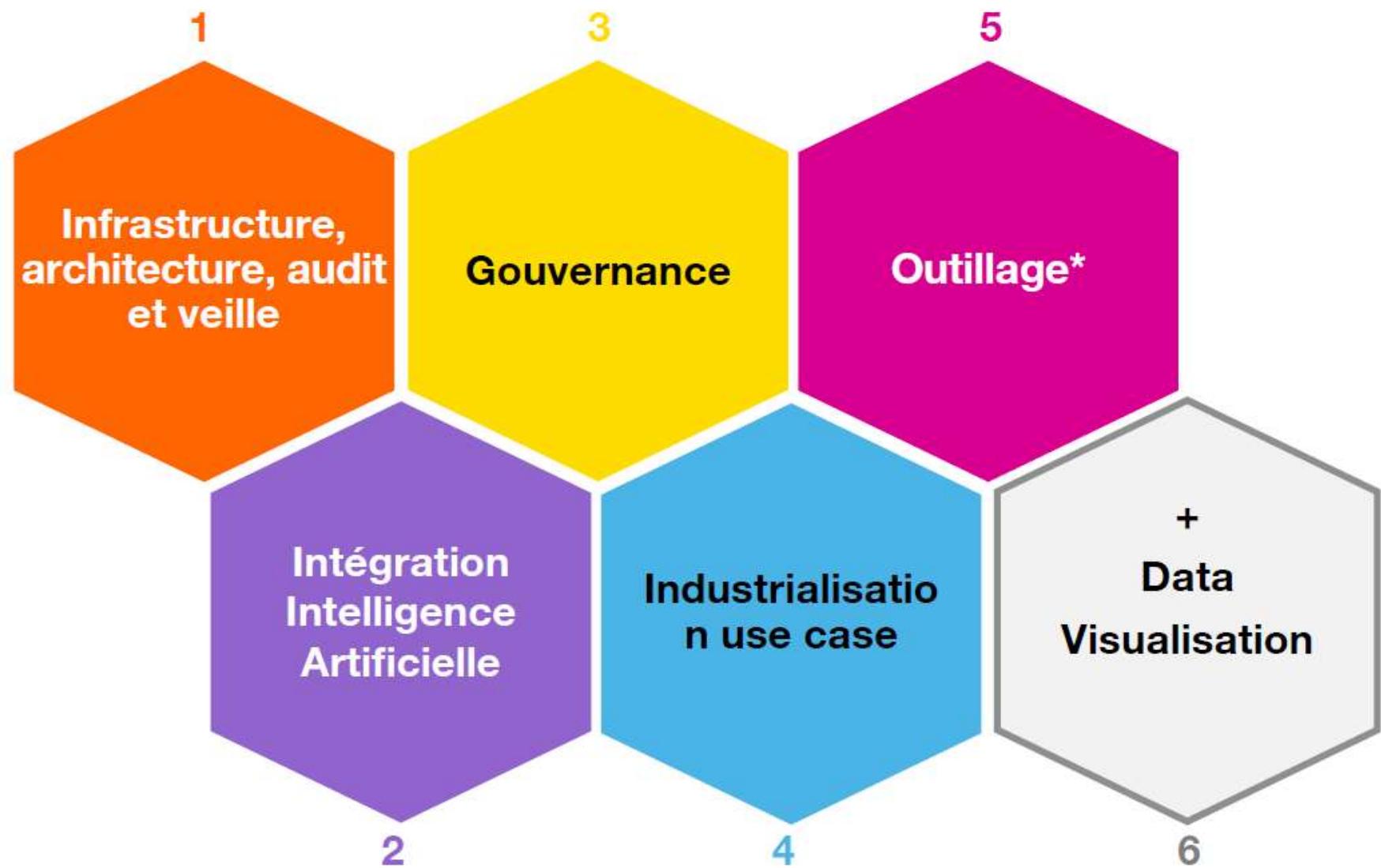


# LES ENJEUX ET LES USAGES DU BIG DATA



Volume Variété Vélocité

# PROJET BIG DATA : LES THÈMES A TRAITER



# REX SEMINAIRE

Comitologie stratégique  
N → N-1 → N-2

DataScience  
Où et comment s'insère  
les métiers ?

Liste des cas d'usages  
CDC

Organiser la filière SI de  
la gouvernance Data

Promouvoir une culture  
Data

Gouvernance centralisée  
de la donnée

Désiloter les données  
pour les utilisateurs

Gain à travailler  
collectivement

Trouver un projet  
donnant l'opportunité de  
désiloter

Pédagogie,  
communication et  
parcours de formation

Besoin d'un pipeline  
client simple

Hybridation de  
l'architecture

Comment désiloter la  
donnée?

Data Scientist et IT  
Comment se mélagent  
ils ?

# REX SEMINAIRE

Quid de la gestion de la dette technologique

Améliorer la performance de la BI en utilisant le BigData

Le postulat «Tout le décisionnel dans la BIODATA » est il réaliste et viable ??  
→ NON RECOMMANDÉ

Processus de construction des objets métiers  
• Quelle organisation déployer ?

Organisation à mettre à plat sur les sujets Data

Définir l'architecture logique et fonctionnelle de la CDC

Cycle d'évolution des objets métiers  
• Qui fait quoi dans l'organisation ?

Monitoring

Quelle vision client avons-nous de la CDC?  
Une vision client Cross Direction ?

Comment construire des principes d'architecture ?

Google Analytics

# REX SEMINAIRE

## Gouvernance

- Quelles métadonnées ajouter à nos données dans le dictionnaire de données ?

Quid de la gestion des métadonnées au niveau des flux

Maintenance et enrichissement de l'écosystème Big Data dans le SI :

- Quelle charge induite?
- Quelles compétences?
- Quelle validations?

Création d'environnement automatique de développement de prototypage

## Industrialiser les POC

Réduire le délai entre l'expérimentation et l'industrialisation

Construire un programme métier Data

## Projet POC

Remplacer Sylab (Lab/Lat actuelle) par le détecteur de fraude dans le Big Data

Avoir une plateforme BigData DEVOPS compliant

## Projet POC

Réintégrer les données générées par les DataScientist dans les applis sources (ex : score risque client réintégré dans Caligéo)

Cycle de la gouvernance :  
Quid de la qualité, de la transformation, de la source, de la durée du cycle?

**Industrialisation Data Science**  
**Test de validation du modèle dans le temps**  
• Comment gérer la maintenance?

**Prérequis pour l'industrialisation des modèles du Build au Run**

I.A gouvernant tous les IOT ?

Quelle plateforme pour la CDC ?

Gouvernance du code  
• Comment faire ?

Données brutes VS données dites « métiers »  
• Quelles différence ?  
• 2 gouvernances ?  
• 2 types de stockage ?  
2 cycles de vie ?

Gestion des référentiels dans le Datalake

Enrichissement des référentiels via les caches présents dans le Datahub Datalake

Comment synchroniser les système référentiel source?

# **LANCEMENT ET GESTION D'UN PROJET BIG DATA**

# QU'EST-CE QU'UN PROJET « BIG DATA » ?

- Les projets peuvent être très variés, du fait des différents domaines couverts par la notion même de big data. Il est intéressant en premier lieu de s'interroger si l'entreprise, par certaines caractéristiques, est concernée par une problématique de type big data.
- Comme pour tout projet, il s'agit donc de répondre à un besoin défini dans des délais fixés et dans les limites d'un budget alloué. Opérationnellement un projet big data nécessite une structure, une organisation et les ressources adéquates.

# L'ENTREPRISE EST-ELLE CONFRONTÉE AU BIG DATA ?

On peut estimer qu'une entreprise est confrontée au big data si elle se trouve dans au moins une des situations suivantes :

- Les bases de données sont saturées.
- Le système d'information de l'entreprise est, ou devra prochainement être, connecté à des systèmes externes.
- Les données ne sont pas toutes structurées ;
- Certaines informations sont produites en temps réel.
- Les besoins de puissance de calcul peuvent varier de façon importante.
- L'analyse des données est peu développée.
- La notion d'analyse prédictive est absente de l'entreprise.
- Un projet visant à traiter un des points précédents est typiquement un projet big data

# PROJET PLURIDISCIPLINAIRE

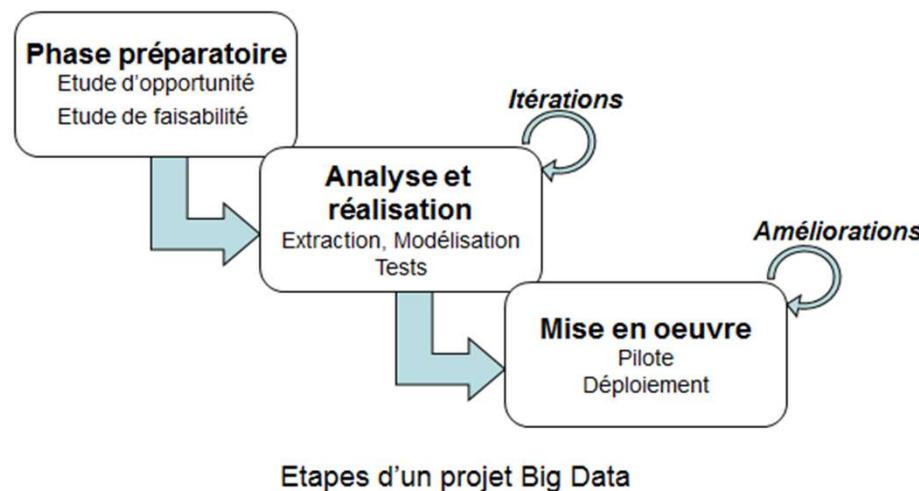
L'organisation d'un projet big data dépend de l'objectif, du contexte et du périmètre à couvrir. Dans tous les cas le projet doit prévoir une structure avec des rôles bien définis. Comme pour la plupart des projets, on aura généralement :

- Une maîtrise d'ouvrage porteuse du besoin et représentant les bénéficiaires, c'est-à-dire les utilisateurs finaux du projet.
- Une maîtrise d'œuvre en charge d'exécuter les travaux. Celle-ci peut être interne ou externe à l'entreprise.
- Au sein de la maîtrise d'œuvre, on trouvera dans pratiquement tous les cas de figure.
  - Un chef de projet ou un product owner.
  - Une équipe fonctionnelle pour formaliser les besoins métier, les mettre en œuvre, tester, former, accompagner.
  - Une équipe technique pour concevoir et mettre en œuvre la technologie requise.

# L'APPROCHE PROGRESSIVE ET ITÉRATIVE

Dans ce type de projets, le modèle classique du cycle en V est déconseillé. Il faut au contraire construire la solution progressivement, en prévoyant quelques itérations comprenant des interactions avec les futurs utilisateurs.

Les méthodes incrémentales sont donc adaptées à ce type de projet. Il s'agit de diviser le projet en incréments.



# LES FACTEURS CLÉS DE SUCCÈS

La première chose importante à réaliser est de s'assurer du bien-fondé du projet lui-même. L'étude d'opportunité menée en phase d'avant-projet doit clairement mettre en évidence les apports du projet pour l'entreprise. Il faut ainsi que le projet :

- Apporte de la valeur au métier de l'entreprise.
- Soit générateur d'un avantage concurrentiel.
- Soit porté par une direction métier et complètement approuvé par la direction générale.
- Ne soit pas le résultat d'un effet de mode.
- La disponibilité de ressources adéquates.

# PALLIER LES RISQUES

Le risque lié à une technologie pas encore « sèche » ne doit pas être sous-estimé. Pour limiter ce type de risque, il convient de se poser les questions fondamentales telles que :

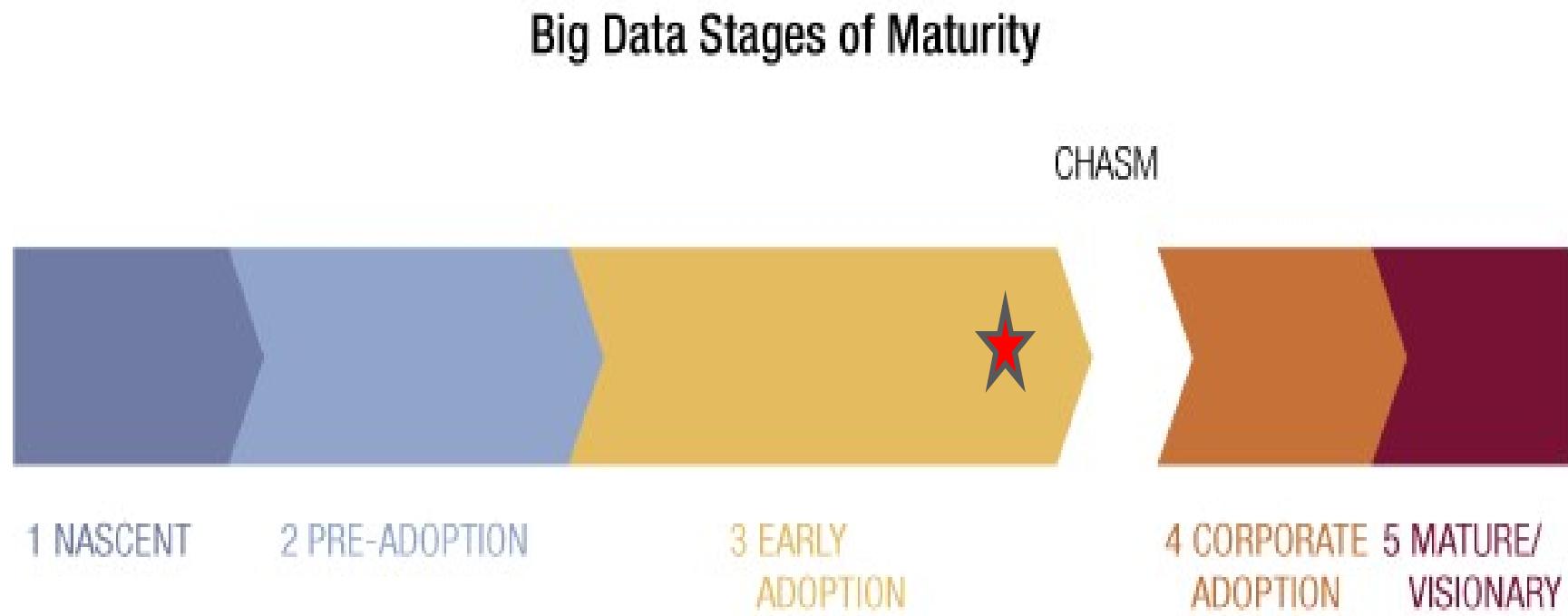
- L'outil envisagé est-il réellement adapté à ce que je veux faire ?
- S'agit-il d'une technologie fiable sur laquelle je peux miser pour les cinq à dix ans à venir ?
- Le fournisseur est-il pérenne ?
- L'outil s'intègre-t-il facilement dans mon paysage actuel ?
- La technologie envisagée respecte-t-elle des standards de normalisation ?

# CONCLUSION

Un projet big data doit s'accompagner d'une réflexion globale sur la gouvernance des données.

Plus que jamais, la prise de conscience des dirigeants de la valeur des données pour l'entreprise doit être présente. Ensuite la volonté de réussir, le pragmatisme et la mise en place des équipes adéquates permettent à coup sûr d'assurer la réussite de ces projets.

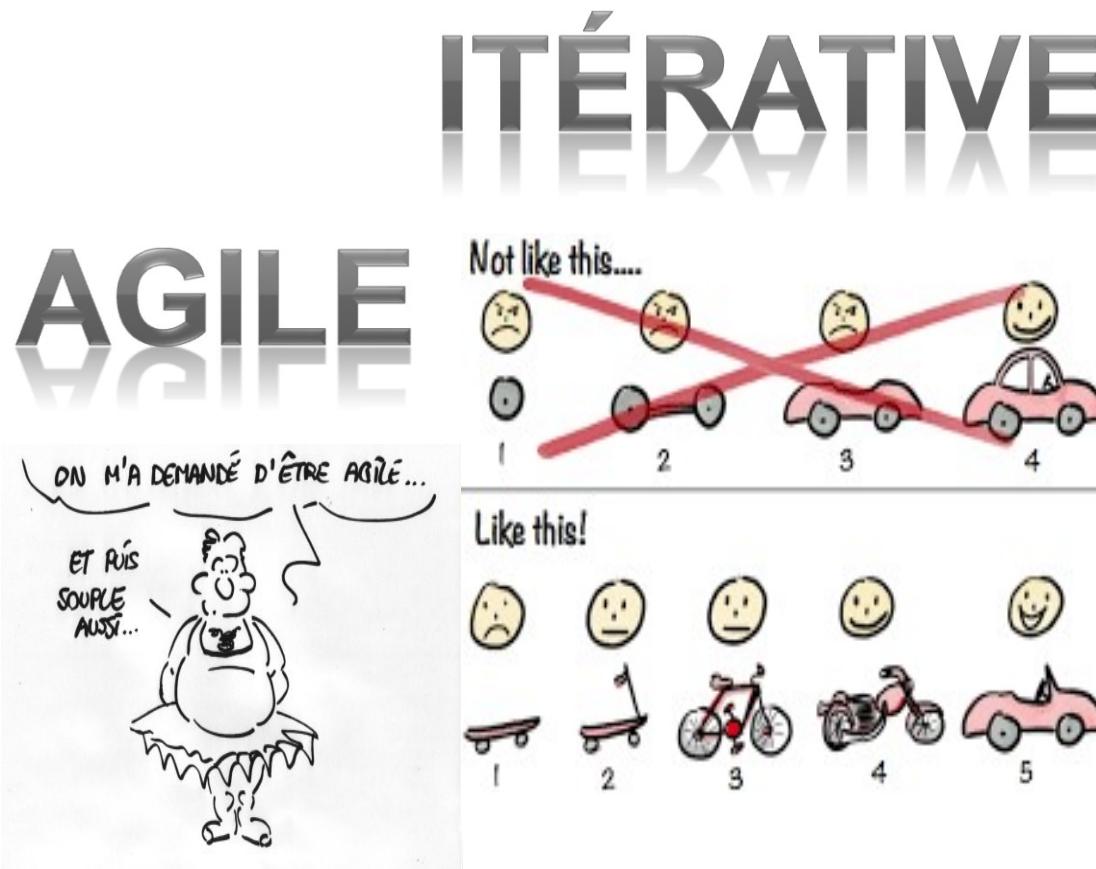
# MODÈLE DE MATURITÉ DU BIG DATA



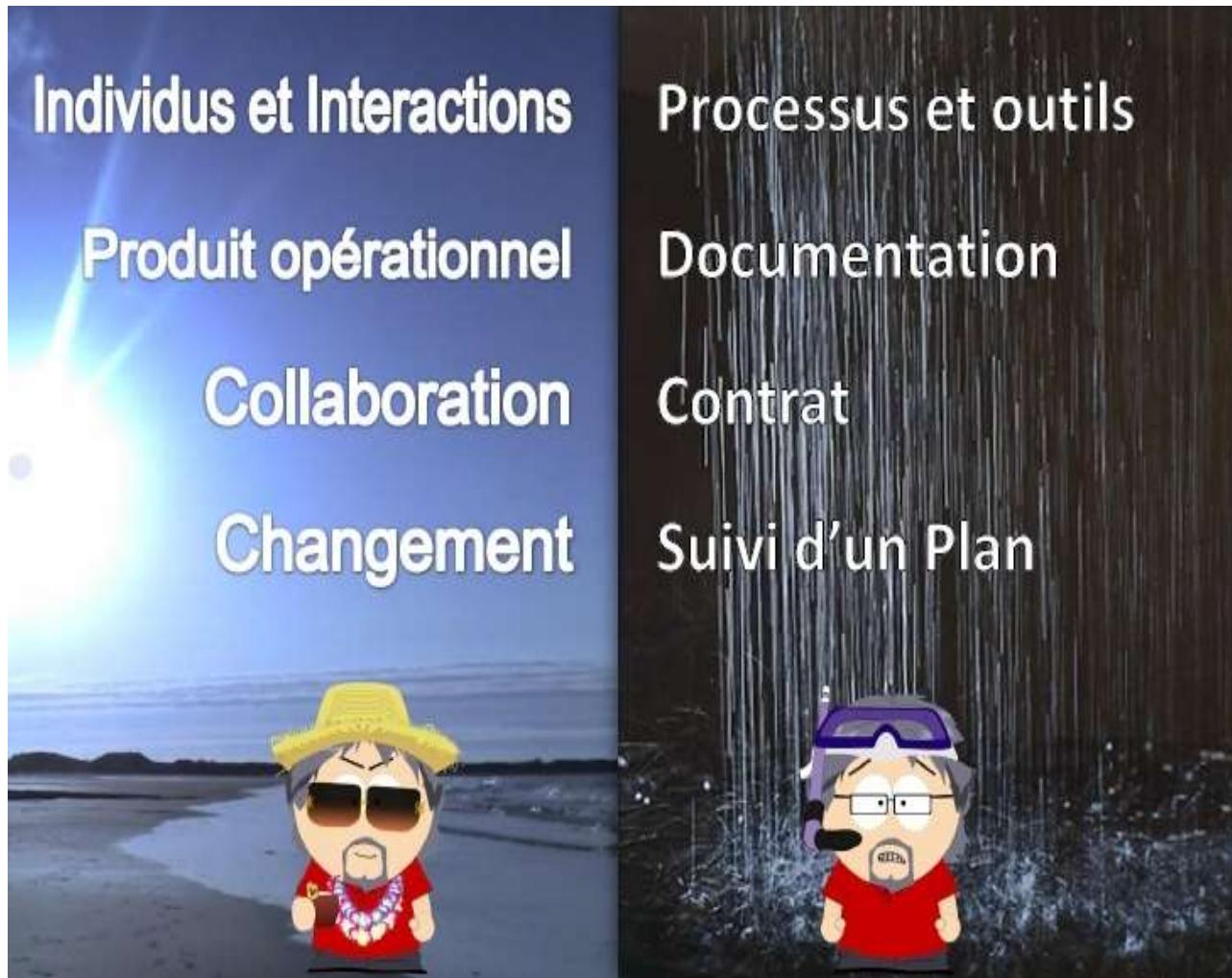
*La majorité des sociétés françaises*

Google : tdwi Big Data Maturity Model

# DÉMARCHE ET MÉTHODE

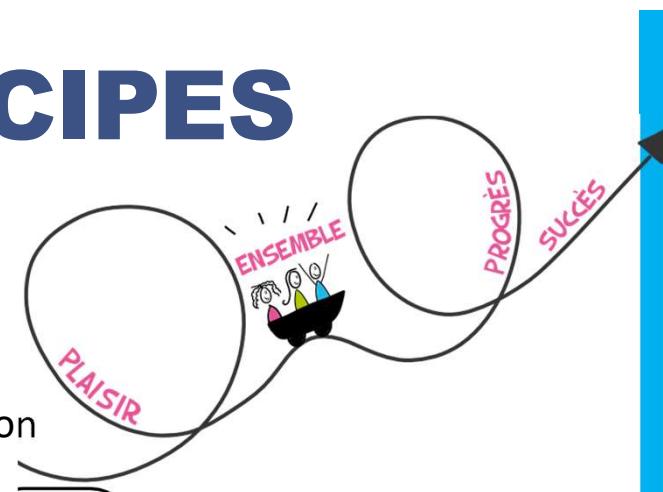


# MANIFESTE AGILE: LES 4 VALEURS

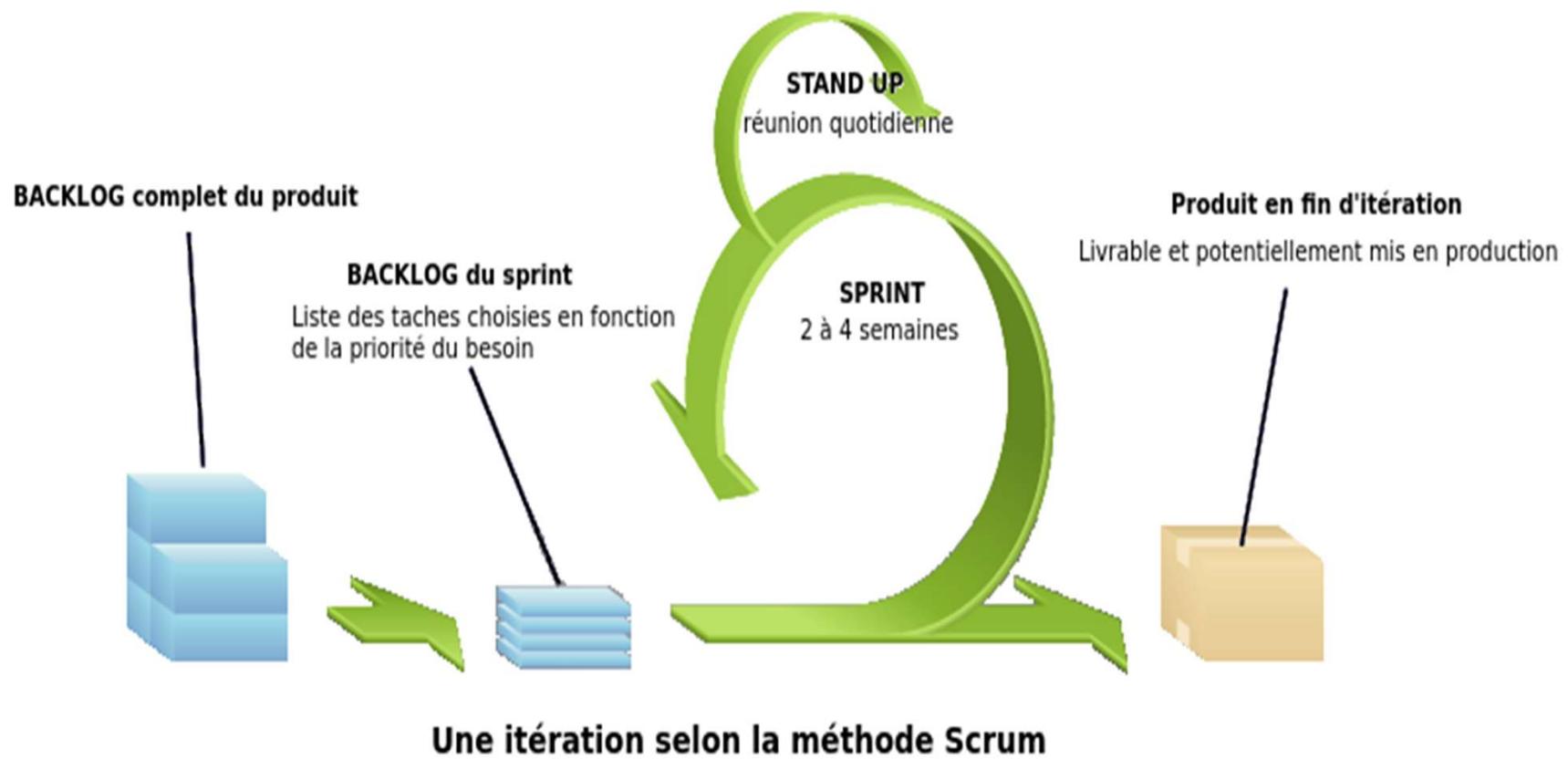


# AGILE: LES 12 PRINCIPES

1. Prioriser la satisfaction du client
2. Accepter les changements
3. Livrer en permanence des versions opérationnelles de l'application
4. Assurer, le plus souvent possible, une coopération entre l'équipe du projet et le client
5. Construire des projets avec des intervenants motivés
6. Favoriser le dialogue direct
7. Mesurer l'avancement du projet en fonction de l'opérationnalité du produit
8. Adopter un rythme constant et soutenable pour tout les intervenants du projet
9. Contrôler continuellement l'excellence de la conception et la bonne qualité technique
10. Privilégier la simplicité en évitant le travail inutile
11. Auto-organiser et responsabiliser les équipes
12. Améliorer régulièrement l'efficacité de l'équipe en ajustant son comportement



# MÉTHODE AGILE



# BI vs Big Data

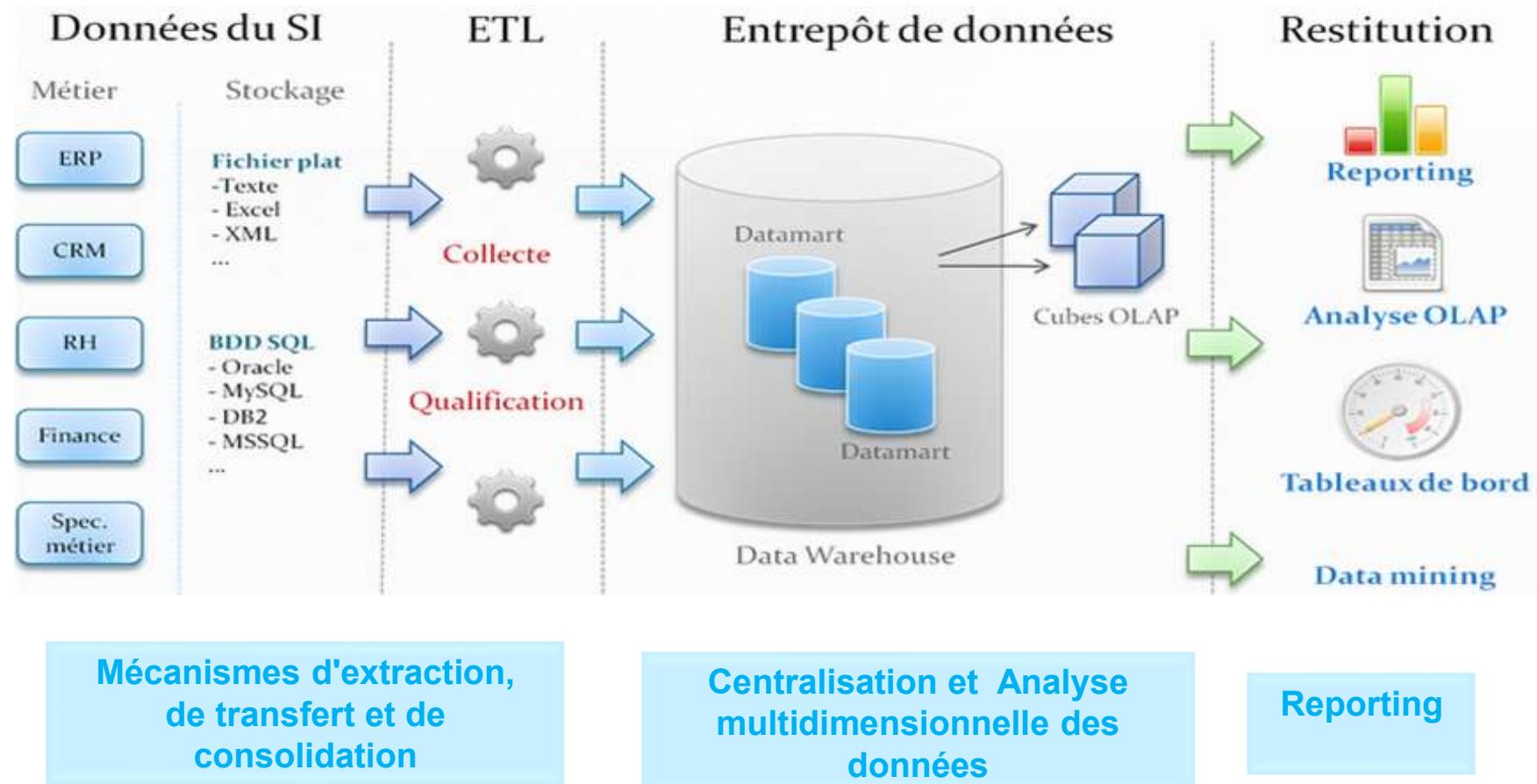
# DIFFÉRENCES TRANSACTIONNEL /DÉCISIONNEL

	Transactionnel	Décisionnel
Données	<ul style="list-style-type: none"><li>- sur une période courte</li><li>- détaillées</li><li>- personnelles</li><li>- mises à jour</li><li>- temps de validité, de transaction</li></ul>	<ul style="list-style-type: none"><li>- historisées</li><li>- agrégées</li><li>- peuvent être anonymes</li><li>- recalculées</li><li>- temps de validité, de transaction, d'extraction</li></ul>
Traitements	<ul style="list-style-type: none"><li>- requêtes simples</li><li>- répétitives</li><li>- très sensible aux performances</li></ul>	<ul style="list-style-type: none"><li>- requêtes complexes</li><li>- variées</li><li>- échelle de performance différente</li></ul>

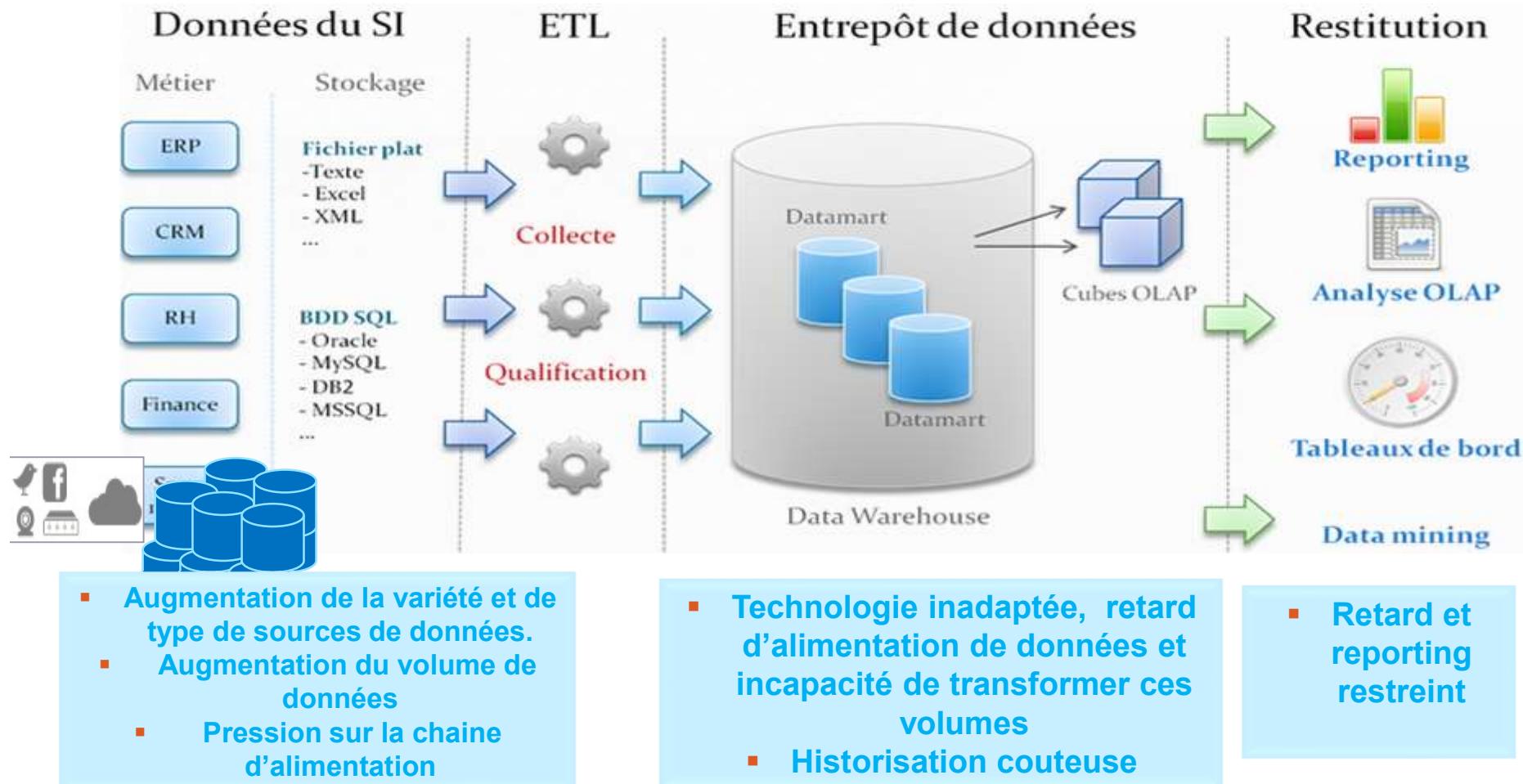
# DIFFÉRENCES TRANSACTIONNEL /DÉCISIONNEL

	Transactionnel	Décisionnel
Conception	- orientée fonction - relativement statique	- orientée sujet - évolutif
Utilisateurs	- agents opérationnels	- manager
Nbres : . utilisateurs	- milliers	- centaines
. tuples accédés	- dizaines/centaines	- millions
. base de données	- centaines de MB/GB	- centaines de GB/TB

# ARCHITECTURE DÉCISIONNELLE

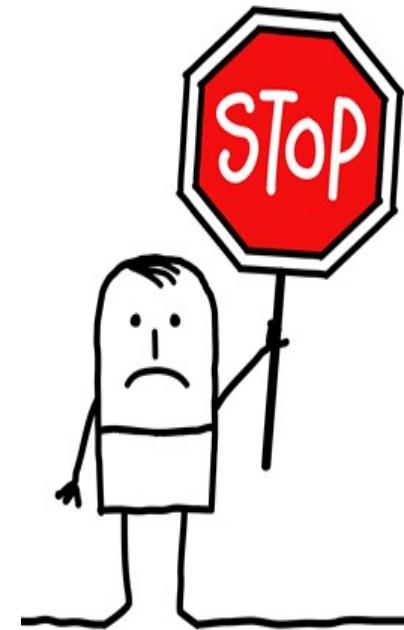


# ARCHITECTURE DECISIONNELLE



## 4 - LES LIMITES DE LA BI

- L'analyse des données en temps réel
- intégrations des données non structurées
- Ces bases de données n'adaptent pas leur capacité de stockage en fonction des montées en charge



# LES ENJEUX DATA DES ENTREPRISE

La majorité des entreprises prennent **le virage de la data**, elles ont comme projection de devenir des « **Data Driven company** ». Et cette prise de conscience se traduit par la nécessité de revoir les process et **les architectures orientées autour des données**.

La maturité d'une entreprise se mesure par sa capacité à gouverner ses processus au travers des données.

Les sociétés Data Driven visent un niveau supérieur. Leur Business Model se veut évolutif, prédictif et prospectif.

Comment ? en intégrant l'IA et dans les processus, en exposant et monétisant la donnée.

# LE DATALAKE

“If you think of a datamart as a store of bottled water – cleaned and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

# LE DATALAKE

L'analyste à besoin de données brutes, variées, provenant de sources différentes.

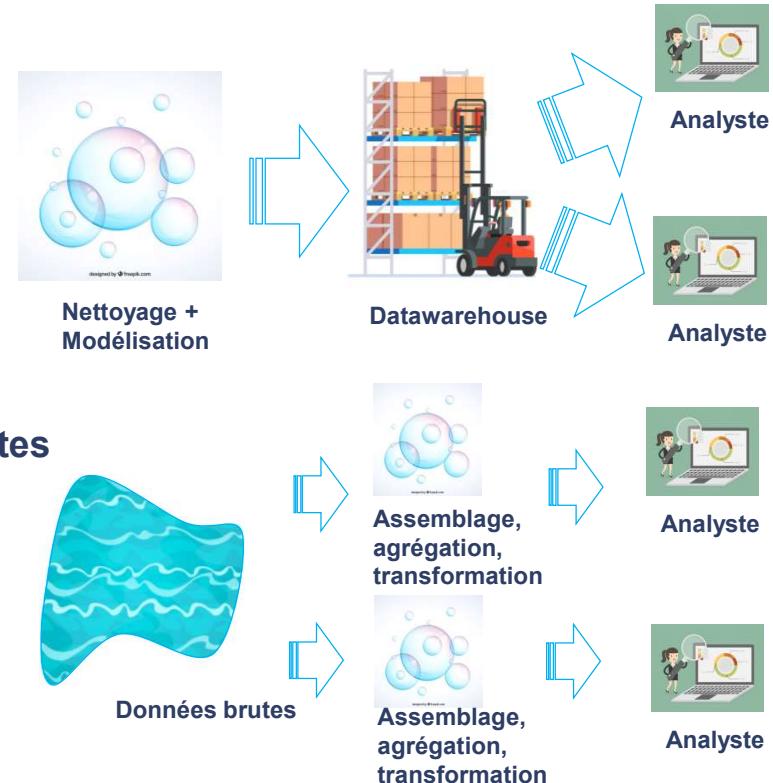
Ces données peuvent devenir rapidement volumineuses et difficiles à explorer.

Le Data Lake est un concept qui répond à ces besoins.

Le Data Lake permet de rassembler les données de différentes sources comme pour un datawarehouse.

➤ Dans un datawarehouse les données sont nettoyées, organisées, agrégées. Les analyses sont effectuées directement sur les données prétraitées.

➤ Dans Data Lake les données brutes sont stockées tel quel. C'est la responsabilité du consommateur de la donnée de lui appliquer un modèle et de lui donner un sens.



# Datalake et GBDR



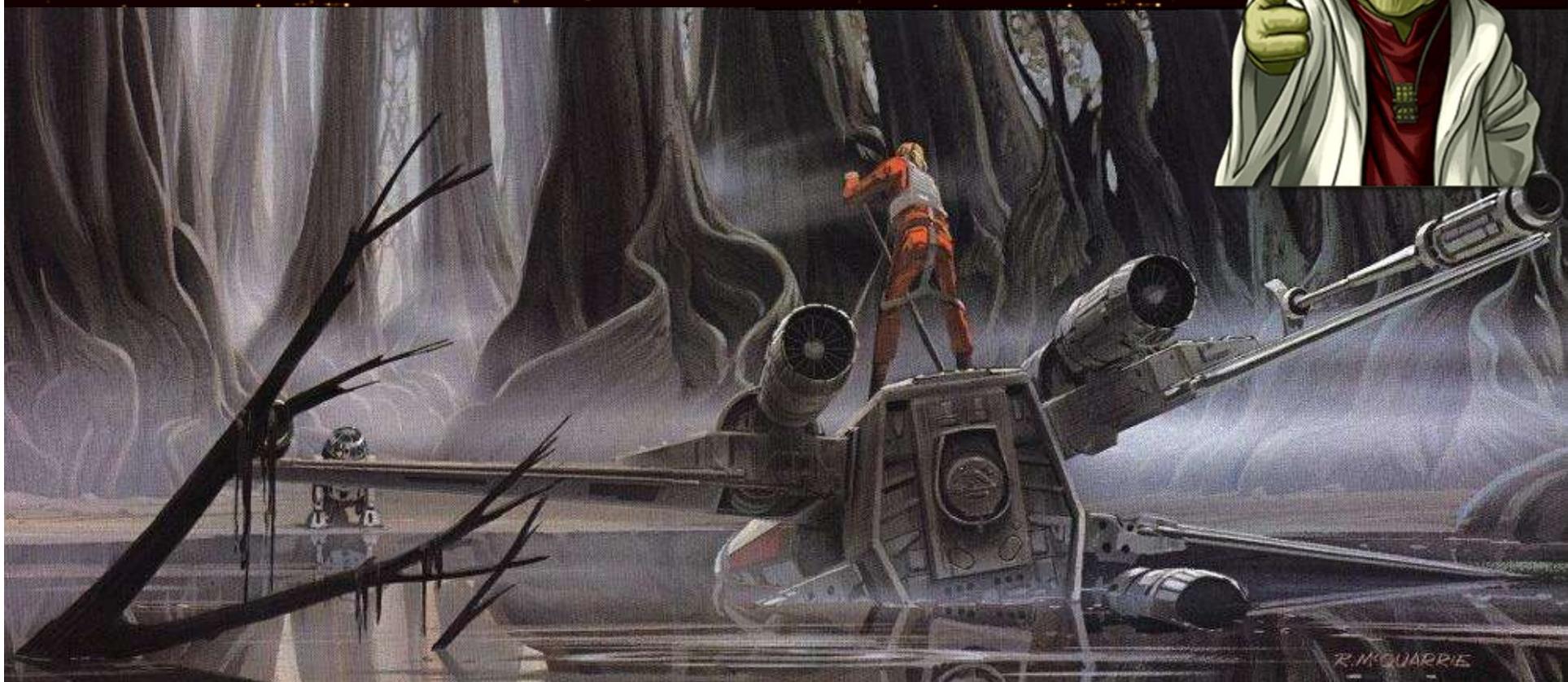
# Datalake et décisionnel



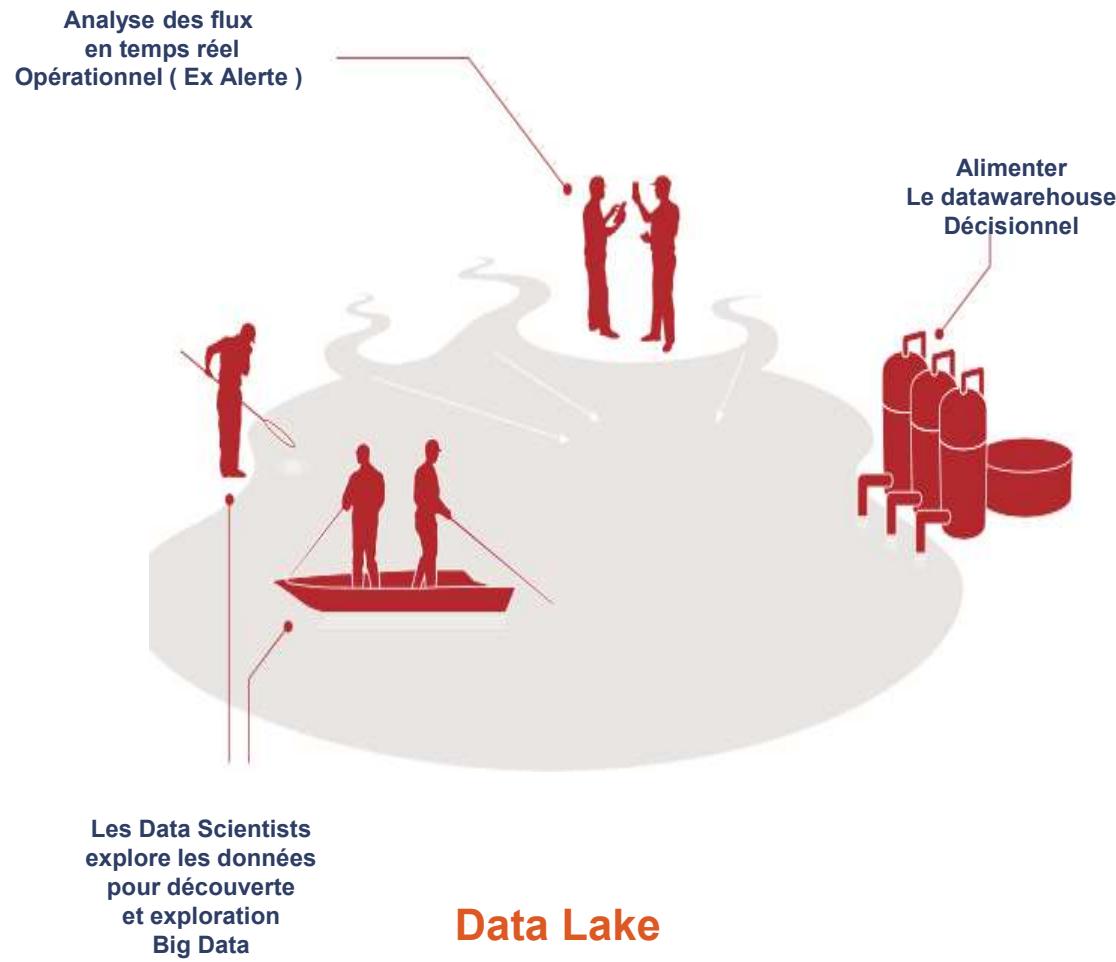
Rigoureux tu devras être  
pour que ton datalake...



en marécage ne se transforme...

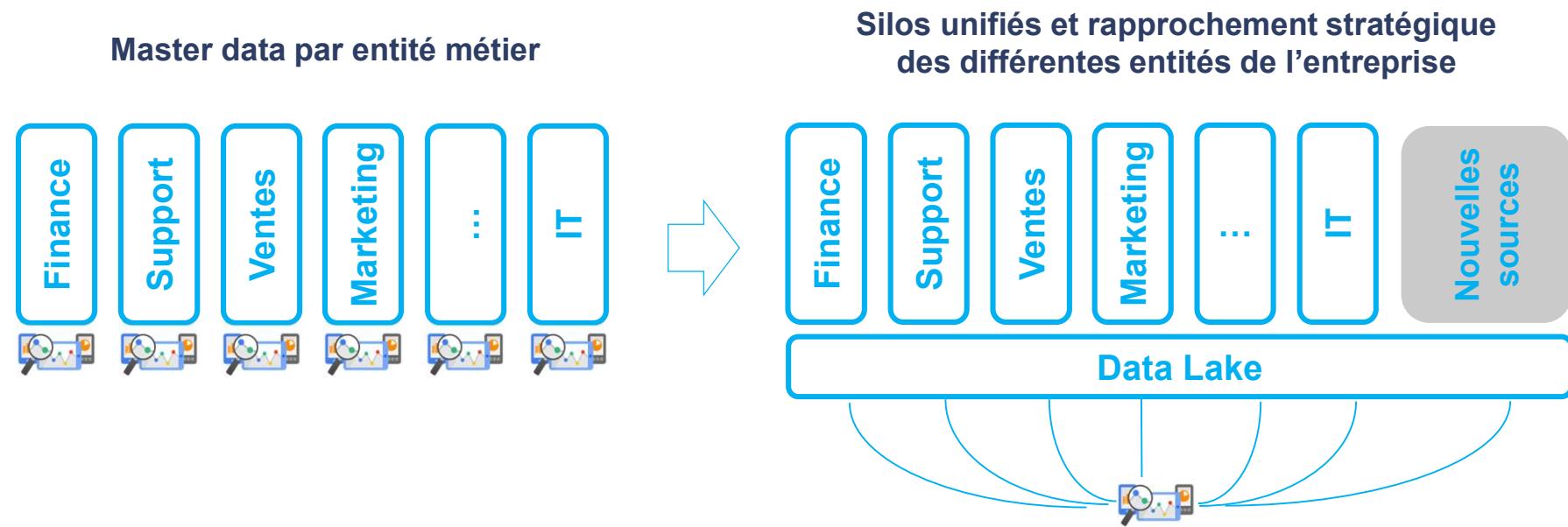


# LE DATALAKE



BI

# LE DATALAKE



# LE DATALAKE

... mais pourquoi tout stocker sans savoir pourquoi ?

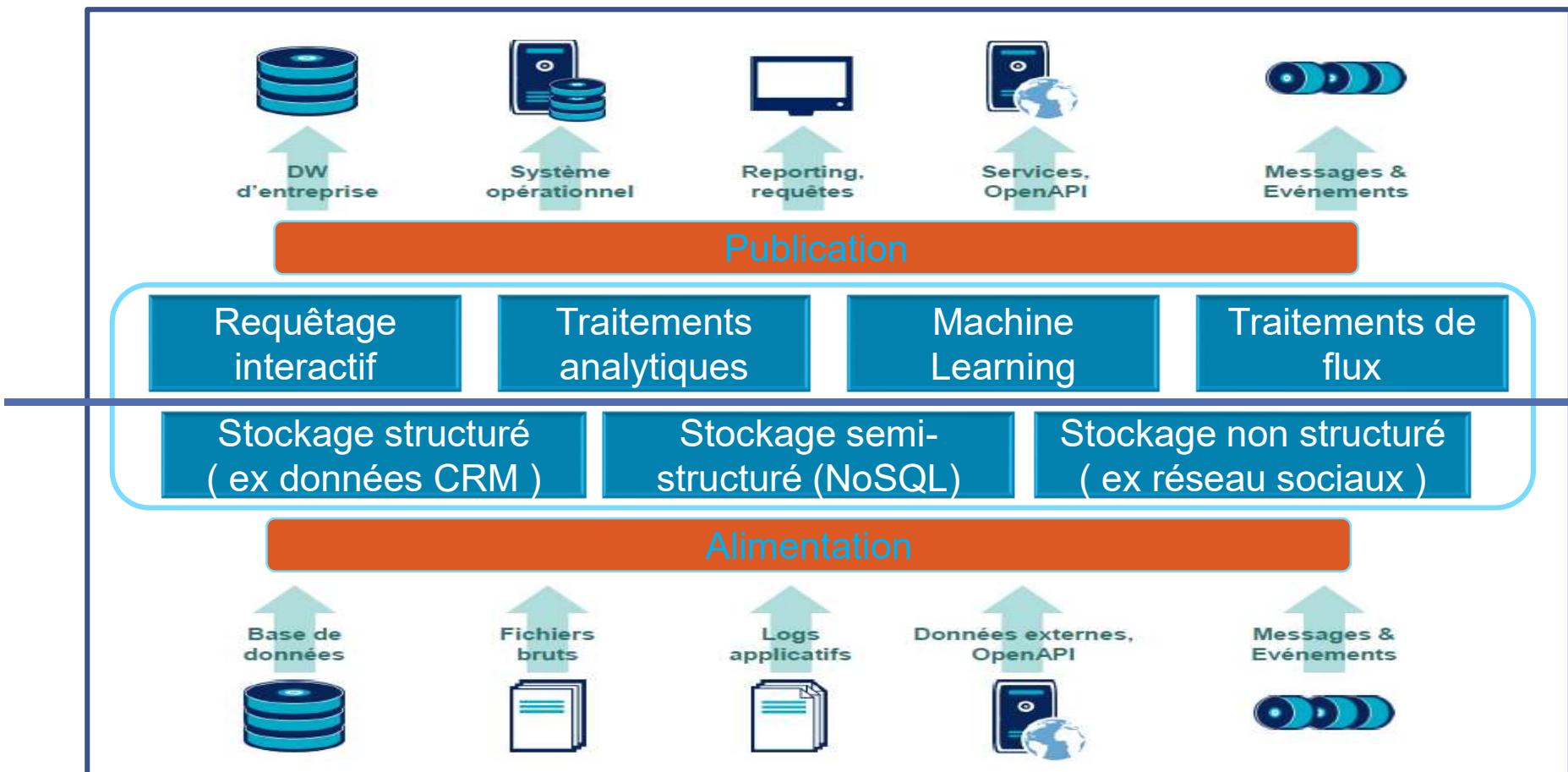
#1 : ça peut servir un jour. En plus, ils le font tous dans la Silicon Valley.



#2 : et puis, vu ce que ça coûte maintenant, ce serait dommage de s'en priver.



# ARCHITECTURE DATA LAKE



# 4 – L'ÉTAGE DE DONNÉES CHEZ UN ÉNERGÉTIEN

Etage  
3+

Exposition des données. Etages des vues métiers, indicateurs.



Transformation spécifiques aux cas d'usages,  
règles de gestions, filtres, périmètres.

Etage  
2

Objets Métiers mutualisés.



Réconciliation et mutualisation des sources.  
Modélisation proche des concepts métiers.

Etage  
1

Flux structurés, mise à plat. Etage de données flux.

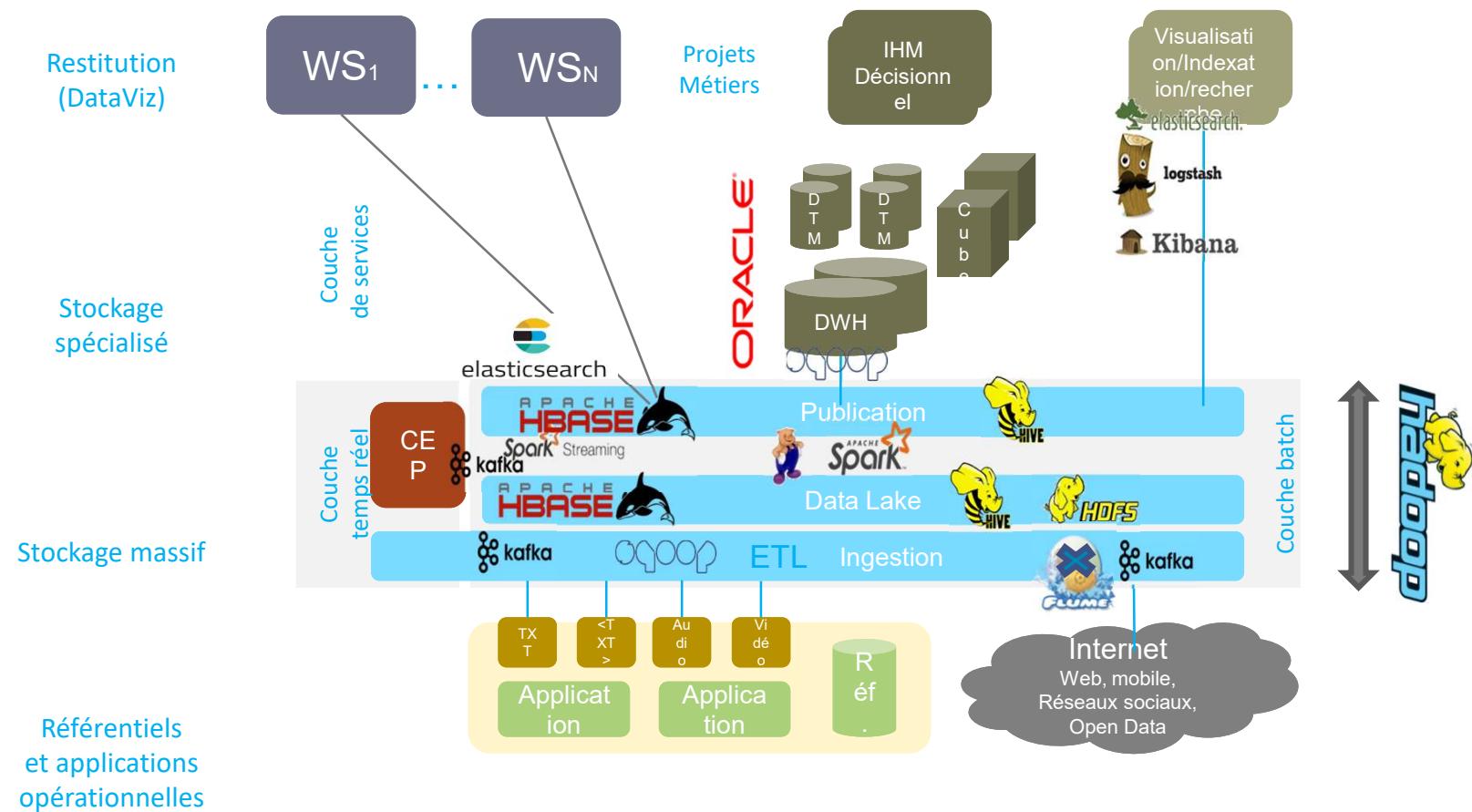


Structuration et typage des données.  
Modélisation au plus proche du flux

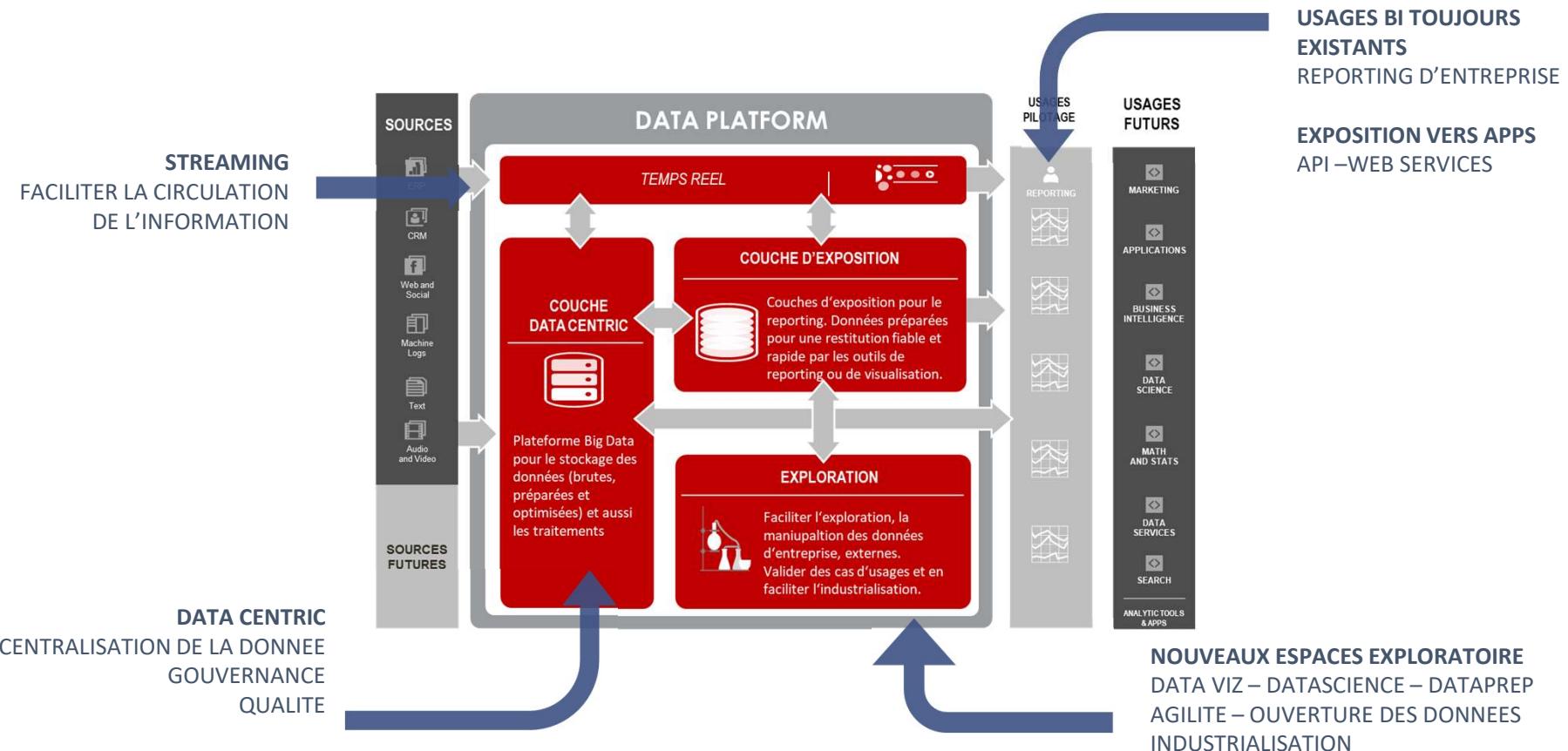
Etage  
0

Réception de flux bruts provenant des applications ou des pivots.  
Etage brut archive

## 4 - DATA LAKE



# Evolution de la BI vers des Data Platform



# SYSTÈME RELATIONNEL VS HADOOP

Relational	Hadoop
Required on write	schema
Reads are fast	speed
Standards and structured	governance
Limited, no data processing	processing
Structured	data types
Interactive OLAP Analytics Complex ACID Transactions Operational Data Store	best fit use
	Data Discovery Processing unstructured data Massive Storage/Processing

*Relational Database vs. Hadoop*

# **LES RÔLES ET MÉTIERS**

Le Big Data ne supprime pas des rôles au sein de l'entreprise  
mais il en crée d'avantage

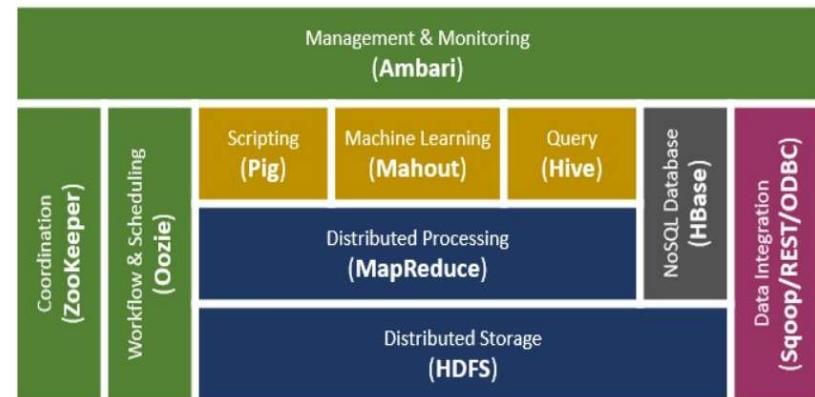
# LE CHEF DE PROJET BIG DATA

- Son rôle n'a pas fondamentalement changé avec l'arrivée du Big Data.
- Il manage une équipe et s'assure que les objectifs du projet sont atteints en respectant les contraintes de budget, de délai, de qualité,...
- Il s'assure que les compétences nécessaires à la réalisation du projet sont dans son équipe.
- Cadrage de projet en adéquation avec le enjeux métiers et IT.
- Garant des chiffages d'où l'importance d'avoir des compétences techniques.



# DATA ENGINEER: DÉVELOPPEUR HADOOP & SPARK

- Développe les cas d'usages sous Hadoop & Spark et depuis peu sur les ELT type Talend
- Valide la conception technique avec l'architecte
- Valide la phase de recette avec l'équipe MOA
- Maintient et fait évoluer les chaines existantes
- Capacité d'apprentissage, Curieux
- Compétences techniques :
  - Scala, python
  - Spark
  - Kafka
  - Java
  - SQL, NoSQL
  - Hadoop
  - Architectures distribuées
  - ETL/ELT
  - Intégration continue



# BUSINESS ANALYSTE : L'EXPLORATEUR DES DONNÉES

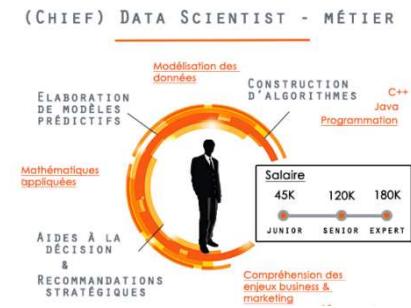
- Analyser les besoins des métiers afin de formaliser le plus fidèlement possible les cahiers des charges ou spécifications.
- Comprendre les interactions entre les différents objets de l'entreprise (au sens modélisation du terme)
- Modéliser les chaînes de collecte, de stockage, de traitement et de restitution des données
- Identifier et développer les référentiels et outils de Master Data Management (MDM) essentiels à la bonne manipulation des données, tout en assurant leur intégrité et leur qualité
- Conception NoSQL
- Compétences :

- Communication
- Capacité d'analyse dans le détail
- Curieux
- SQL, R
- Outils de Dataviz type Tableau, Qlik,...



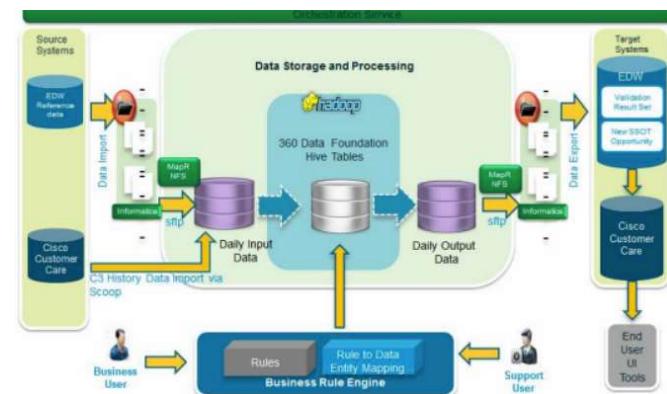
# DATA SCIENTIST : LE MOUTON À CINQ PATTES

- Conçoit des modèles servant à faire parler les données et d'en dégager des indicateurs viables intelligibles pour les Direction Générale, marketing, financière, Ressource humaine ...
- Il faut que le Data scientist comprenne la problématique marketing, marché, commerciale, fidélisation client, RH, etc.
- Expert de la modélisation statistique et algorithmique pour répondre aux problématiques demandés selon les cas d'usage.
- Analyser les données et restituer les résultats, de façon à ce qu'ils répondent à une stratégie (le plus souvent commerciale).
- Ce qui est important c'est la capacité à coder et à travailler avec les outils statistiques et d'analyse prédictive
- Compétences techniques :
  - SQL ,R, Python, Spark est un plus
  - Machine learning
  - Triple compétence : statistique , développement informatique et métier (marketing, finance,...)



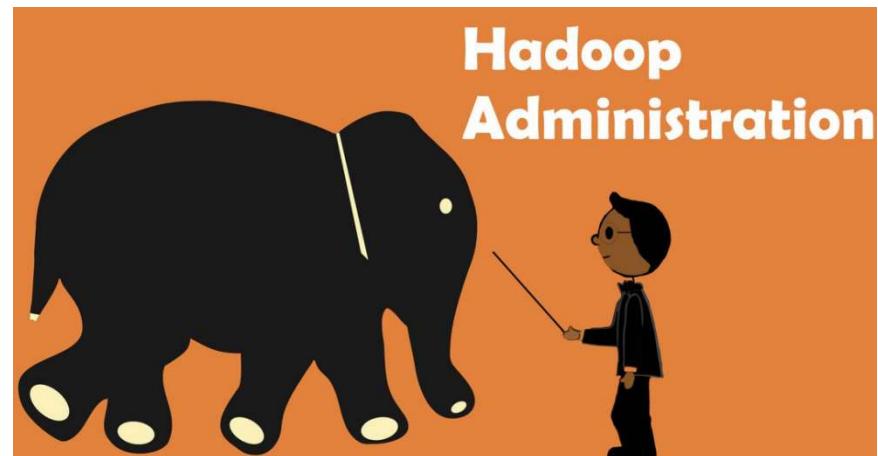
# ARCHITECTE BIG DATA :

- L'architecte s'appuie sur différents patterns et s'inscrit dans une démarche collaborative.
- son rôle est de concevoir des architectures en réponses à un besoin.
- Cette architecture s'appuie sur une infrastructure. Bien que l'architecture ne s'occupe pas directement de l'infrastructure, il a une bonne connaissance de celle-ci.
- Un appui indispensable pour les Data Engineers
- Définition des patterns de développements
- Garant de la conception technique et des montées de charges
- L'architecte Big Data a une vision transverse sur l'ensemble du SI afin d'assurer une cohérence dans les choix



# ADMINISTRATEUR HADOOP

- Administrer les clusters Hadoop, ElasticSearch..
- Ajout des machines de stockages
- Montée de version de cluster
- Gère les accès au cluster
- En relation avec le support de la distribution
- Garant de l'architecture matérielle
- Compétences:
  - Linux
  - Réseaux



# DATA OWNER

1 | 3 | 4 | 5



## Responsabilités

- Régler les problèmes de qualité de données à la source comme les données manquantes, dupliquées ou corrompues.

## Collaboration:

- Analyste
- Data scientist
- Data steward

En Big Data le problème de la qualité de la données est amplifié. Les systèmes prédictifs sont très sensibles au bruit et une mauvaise qualité de données entraîne de l'incohérence dans les résultats.

# DATA STEWARD

1 | 3 | 4 | 5



## Responsabilités

- Définir les métadonnées
- Définir les règles de qualité de la donnée
- Définir les règles et conventions de nommage
- Définir les règles d'accès et de sécurité
- Définir les besoins en monitoring
- Définir les règles de rétention et d'archivage

## Collaboration:

- Métier
- Data quality manager
- Business

C'est un rôle qui est à la fois IT et métier. Il doit y avoir un data steward par domaine métier. Les data steward doivent communiquer pour convenir d'une stratégie de gouvernance commune.

# CHIEF DATA OFFICER

1 | 3 | 4 | 5



## Responsabilités

- Effectuer un inventaire des données de l'entreprise
- Documenter l'usage des données
- Identifier les sources de données à acquérir
- Evaluer la valeur économique de la données
- Identifier des cas d'utilisation pour exploiter la donnée
- Développer des stratégies pour exploiter l'IOT
- Appliquer la gouvernance de la donnée

## Collaboration:

- Tous les acteurs

LE CDO est un membre du comité de direction qui traite les sujets liés à la donnée. Il est responsable de la bonne utilisation de la donnée à des fins d'optimisation des processus métier de la création de valeur.

# DATA PROTECTION OFFICER

1 | 3 | 4 | 5



## Responsabilités

- Garantir la conformité par rapport à la GDPR
- Assurer la conformité légale avec les aspects juridiques
- Préserver l'image de l'entreprise et la confiance vis à vis des parties prenantes
- Coopérer avec la CNIL sur les aspects juridiques et réglementaires

## Collaboration:

- Métier
- CDO
- DSI

La GDPR rend ce rôle obligatoire pour les entreprises qui traitent des données personnelles et sensibles. Ce rôle peut être tenu par un consultant externe.

# LA GOUVERNANCE & LA GDPR

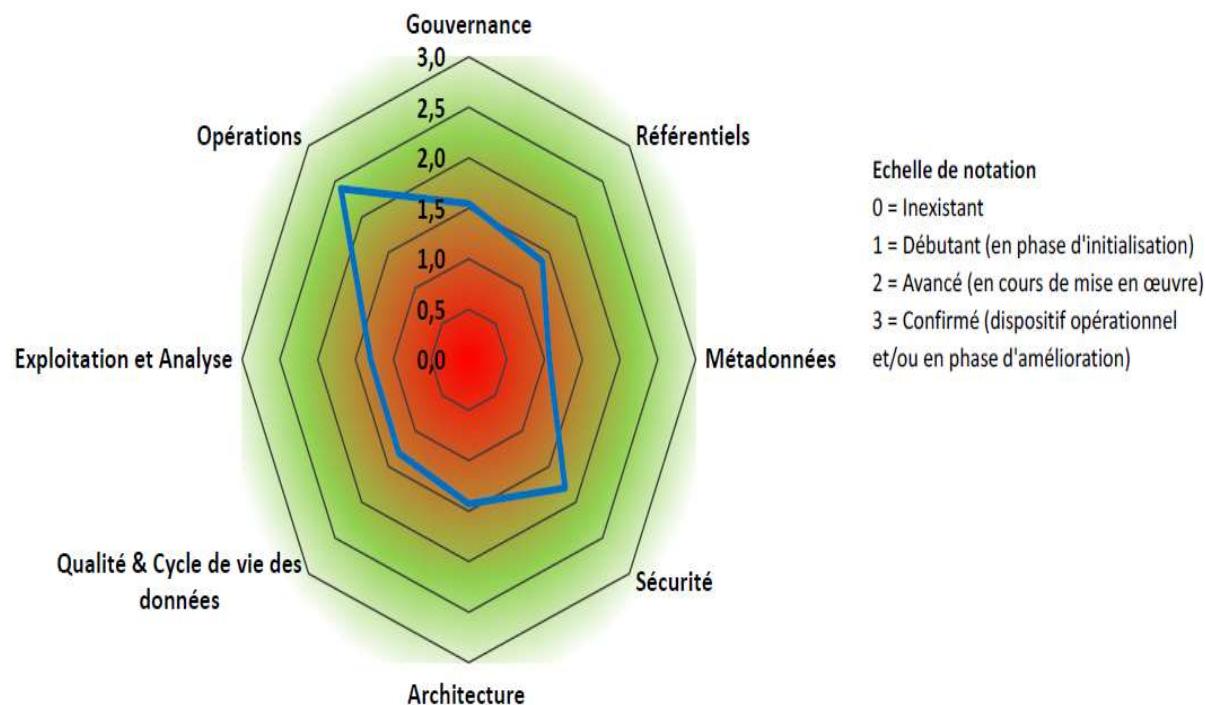


25th of May  
**2018**

# LA GOUVERNANCE

Ne pas traiter un projet Big Data seulement sous l'angle technique.

Maturité de la gestion des données  
dans les entreprises du groupe de travail CIGREF



# LA GOUVERNANCE DE LA DONNÉE

**La donnée est devenue, avec le capital, les ressources humaines, les clients, les processus et le système d'information, l'un des principaux actifs de l'entreprise, quel que soit son secteur d'activité.**

**L'information ne constituera une nouvelle richesse pour l'entreprise que si on sait l'exploiter et la valoriser**

**C'est la gouvernance qui décrit les responsabilités, fixe les règles et contrôle leur application.**

# LA GOUVERNANCE

**La gouvernance des données est l'ensemble des moyens mis en place par une entreprise pour gérer le patrimoine de données:**

- **Monitoring**
- **Sécurité**
- **Politiques, règles et procédures**
- **Conformité avec les lois et réglementations**

# LA GOUVERNANCE

**La gouvernance des données assure une communication fluide entre l'IT et le métier. Elle rétablit la confiance entre tous les acteurs.**

## ➤ L'IT:

**Possède une vue globale du patrimoine des données et peut ainsi gérer la mutualisation des technologies et processus à travers les entités.**

**Fournit des services autour de la donnée de manière agile**

**Contrôle la bonne application des règles et standards**

## ➤ Le métier:

**Préserve la qualité et la cohérence des données**

**Permet aux différents acteurs de partager un vocabulaire commun**

**Réduit le risque pour le business à travers une bonne séparation des rôle et responsabilités.**



# GOUVERNANCE DE LA DONNÉE

## Instance de pilotage :

- définition des règles, des priorités
- stratégie de référentiels, des données à collecter
- suivi d'indicateurs
- Politique de sécurité et de gestion des données personnelles,...

## Déterminer des propriétaires et des responsables de la donnée :

- Chief Data Officer, Data Owner , Data Stewart

## Déterminer des processus d'évolution et de consommation des données

- Contrat entre producteur/consommateur
- *Permet d'identifier les cas d'usage impactés par une évolution ou une anomalie*

# GOUVERNANCE DE LA DONNÉE

## Il faut pouvoir auditer et qualifier les données

- Disposer d'un dictionnaire de données ➔ partager un langage commun
- Identifier les règles de contrôle et de qualité de données à mettre en place

## Définir le cycle de vie de la donnée en amont

- Accessibilité,
- Fréquence d'archivage et de suppression,
- Criticité sur les temps de restaurations

Définir des règles de noms suivant les technologies de stockage ou de traitement

# ENJEUX DE LA GOUVERNANCE

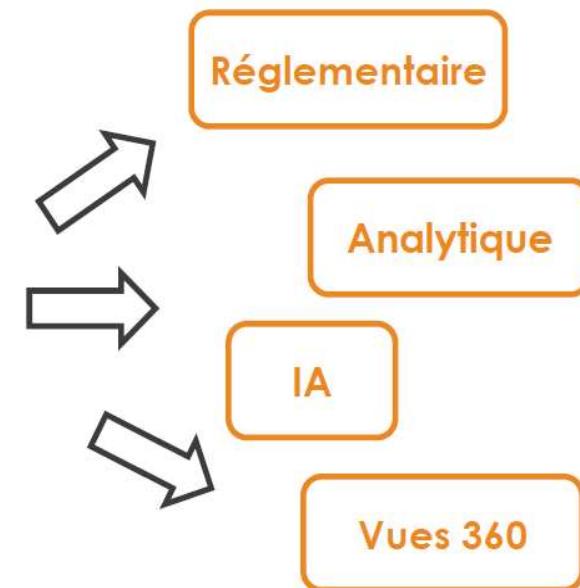
Rechercher et explorer mes données à partir de critères métiers

Comprendre la signification de mes données

Offrir une vision de bout en bout d'une donnée produite à partir des données sources

Mesurer la qualité d'une donnée et son usage

Faciliter l'extraction de pistes d'audits



TERADATA.

# **SANS GOUVERNANCE DE LA DONNÉE**

**Doublons d'intégration dans le Data Lake**

**Les mêmes données sont stockées sous différents libellés**

**Le même indicateur est calculé plusieurs fois avec des règles de gestion différentes**

**L'évolution d'un champ n'est pas répercuté sur les doublons.**

**Difficile d'identifier les impacts techniques d'une évolution**

**Difficile d'identifier les cas d'usages impactés par une évolution →  
Augmentation des délais de validation**

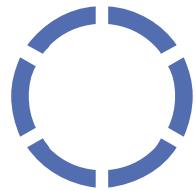
# Focus : Gouvernance et Qualité

## INTÉGRATION



INTEGRATION DE DONNEES  
DATA SERVICES

## QUALITÉ DES DONNÉES ET ENRICHISSEMENT



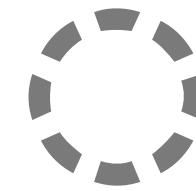
PROFILING DES DONNEES  
GOUVERNANCE  
PARSING/STANDARDISATION  
NORMALISATION  
VALIDATION POSTALE  
MATCHING / CONSOLIDATION  
ENRICHISSEMENT

## GÉO SPATIAL



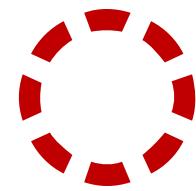
GEOCODAGE  
AIDE SAISIE ADRESSE  
CALCUL ITINERAIRE  
WEBSERVICES  
COUPABLE CARTO / BI

## MASTER DATA MANAGEMENT



CONNAISSEMENT  
CARTOGRAPHIER  
ANALYSER  
CENTRALISER  
ADMINISTRER / PILOTER  
DIFFUSER

## BUSINESS SERVICES



CONNECTEURS SAP, SIEBEL,  
DYNAMICS, SALES FORCE ...

**EVALUER → MESURER → INTEGRER → AMELIORER → CORRIGER → DIFFUSER → EVALUER**

# GDPR EN 4 CHIFFRES



Le règlement 679/2016 appelé GDPR s'applique à toutes les entreprises qui détiennent ou traitent des données personnelles de citoyens européens

GDPR est traduit en Français en Règlement Général sur la Protection des Données (RGPD)

Le périmètre des données personnelles concernées est large



composent le  
règlement



signataires  
pour un champ  
d'application  
mondial



pour se conformer  
Entrée en application  
en Mai 2018



ou 20 M€ d'amende  
possible

# GDPR : ZOOM SUR 8 DROITS POUR LES INDIVIDUS

Article 7  
Consentement explicite

Articles 11 à 14  
Information de la personne

Articles 15  
Droit d'accès

Article 16  
Droit de rectification

Article 17  
Droit à l'oubli

Article 20  
Droit à la portabilité

Article 21  
Droit d'opposition

Article 22  
Décisions automatisées (profilage)

Source Buisness & Decision

# GDPR : ZOOM SUR 8 OBLIGATIONS POUR LES ORGANISATIONS

Article 5  
Responsabilité et  
« Accountability »

Article 25  
Privacy  
by Design  
by Default

Article 30  
Registre des traitements

Article 32  
Sécurisation des  
traitements

Article 33 et 34  
Fuites de données

Article 35  
Data Privacy Impact  
Assessment (DPIA)

Article 37 à 39  
Data Protection Officer  
(DPO)

Articles 44 à 49  
Transferts de données

Source Buisiness & Decision

# DONNÉES PERSONNELLES & SENSIBLES

**Données Personnelles : « toute donnée permettant d'identifier directement ou indirectement une personne physique »**

**Une donnée anonymisée ou pseudonymisées n'est pas une donnée personnelle et n'est pas soumis à la réglementation.**

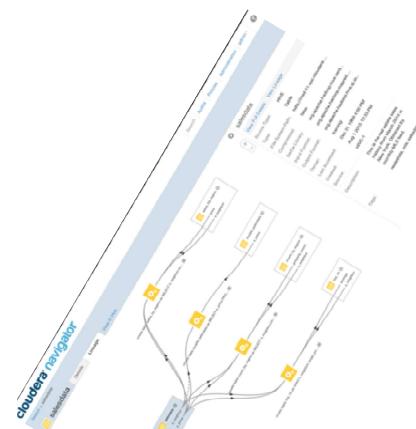
**Données sensibles : les données à caractère personnel qui font apparaître, directement ou indirectement:**

- les origines raciales ou ethniques
- les opinions politiques
- Les opinions philosophiques
- Les opinions religieuses
- l'appartenance syndicale des personnes
- Les données qui sont relatives à leur santé
- Les données qui sont relatives à leur vie sexuelle

**Ces données sensibles ne peuvent être collectées ou sauvegardées sans l'accord directe du régulateur (déclaration à la CNIL)**

# QUELLES SOLUTIONS POUR GDPR ?

■ La mise en conformité à GDPR va nécessiter le déploiement de multiples briques technologiques



## Sécurité

- Cryptage et sécurité des données
- Protection contre les intrusions et le vol de données
- Identity and Access Management
- Gestion des accès et droits sur les données



## Data Management

- Centralisation et gestion de la qualité des données
- Gestion des données de référence
- Traçabilité des données (Metadata et Data Lineage)
- Anonymisation / Pseudonymisation et génération de jeux de données



## Conduite du changement

- Revue des procédures clients
- Intégration du privacy-by-design dans les méthodos projet
- Changement de comportement face aux questions de sécurité



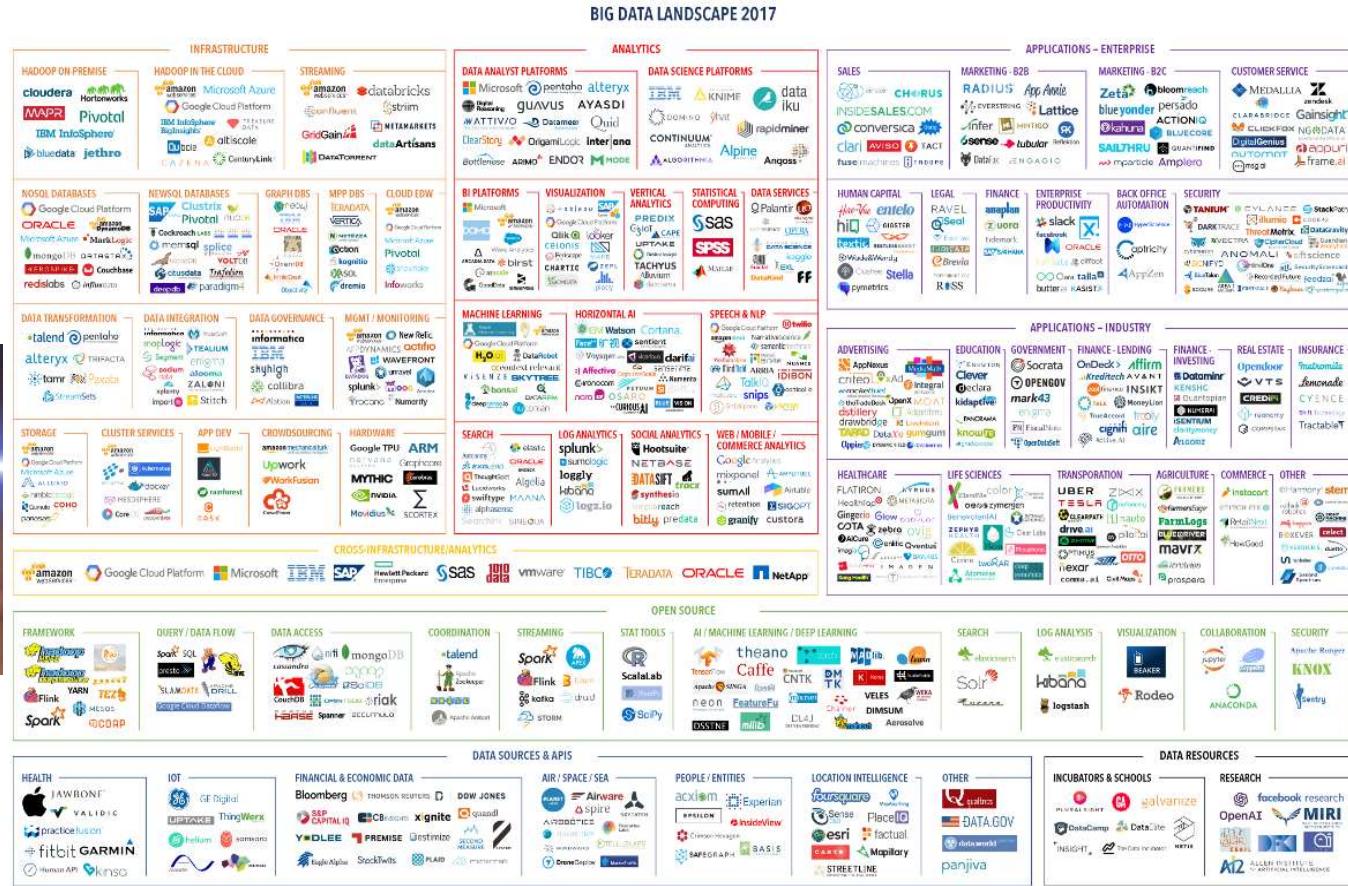
# LES TECHNOLOGIES



# L'ECOSYSTEME BIG DATA



*Ne nous laissons pas impressionner  
par le coté obscur !!*



Last updated 4/5/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)

[mattturck.com/bigdata2017](http://mattturck.com/bigdata2017)

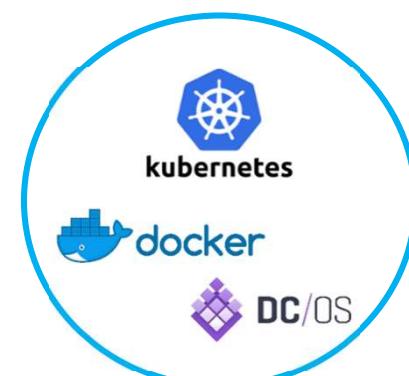
## LES NOUVELLES TENDANCES



Streaming/Temps réel



DevOps



Containers/Orchestration



Cloud

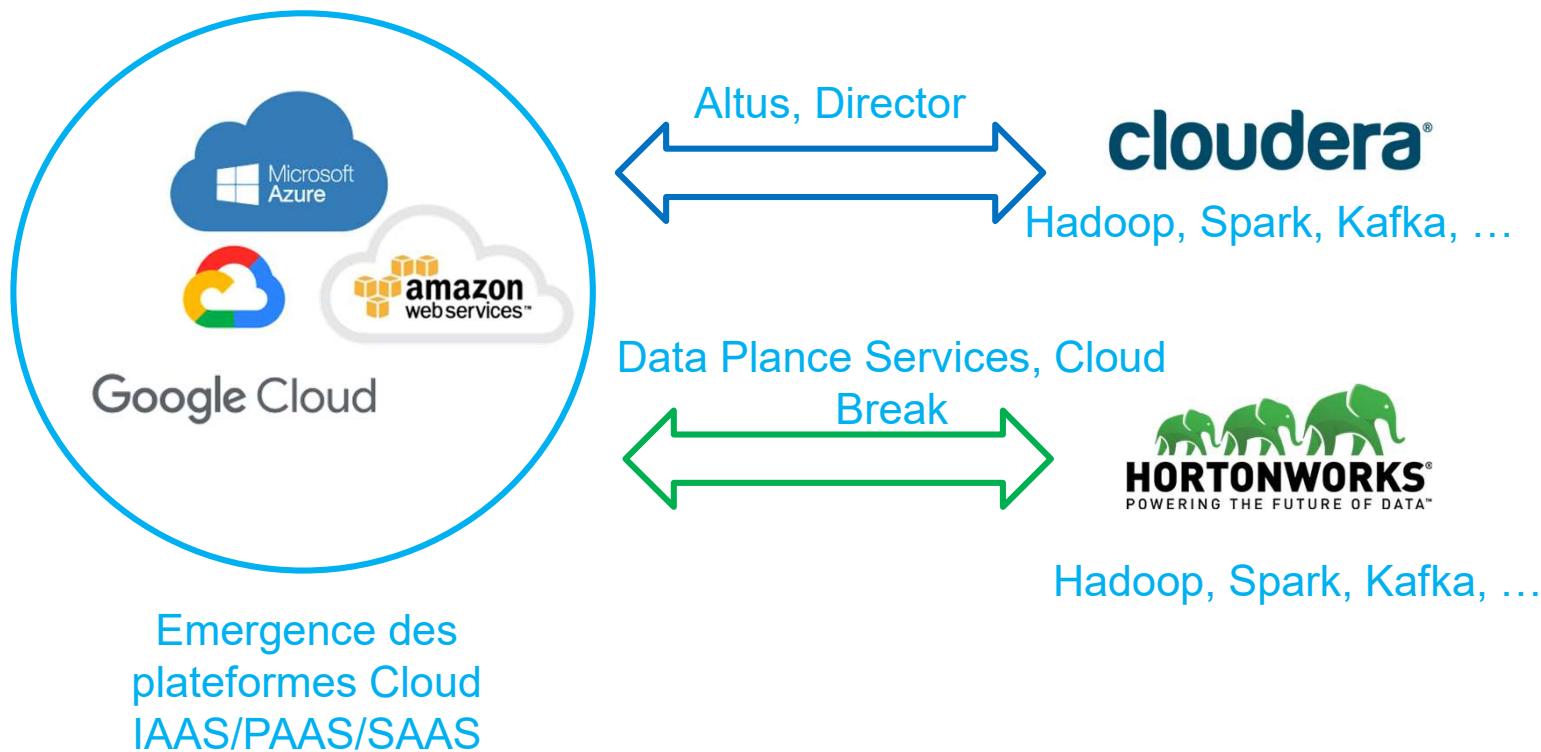


Plateformes Hybrides

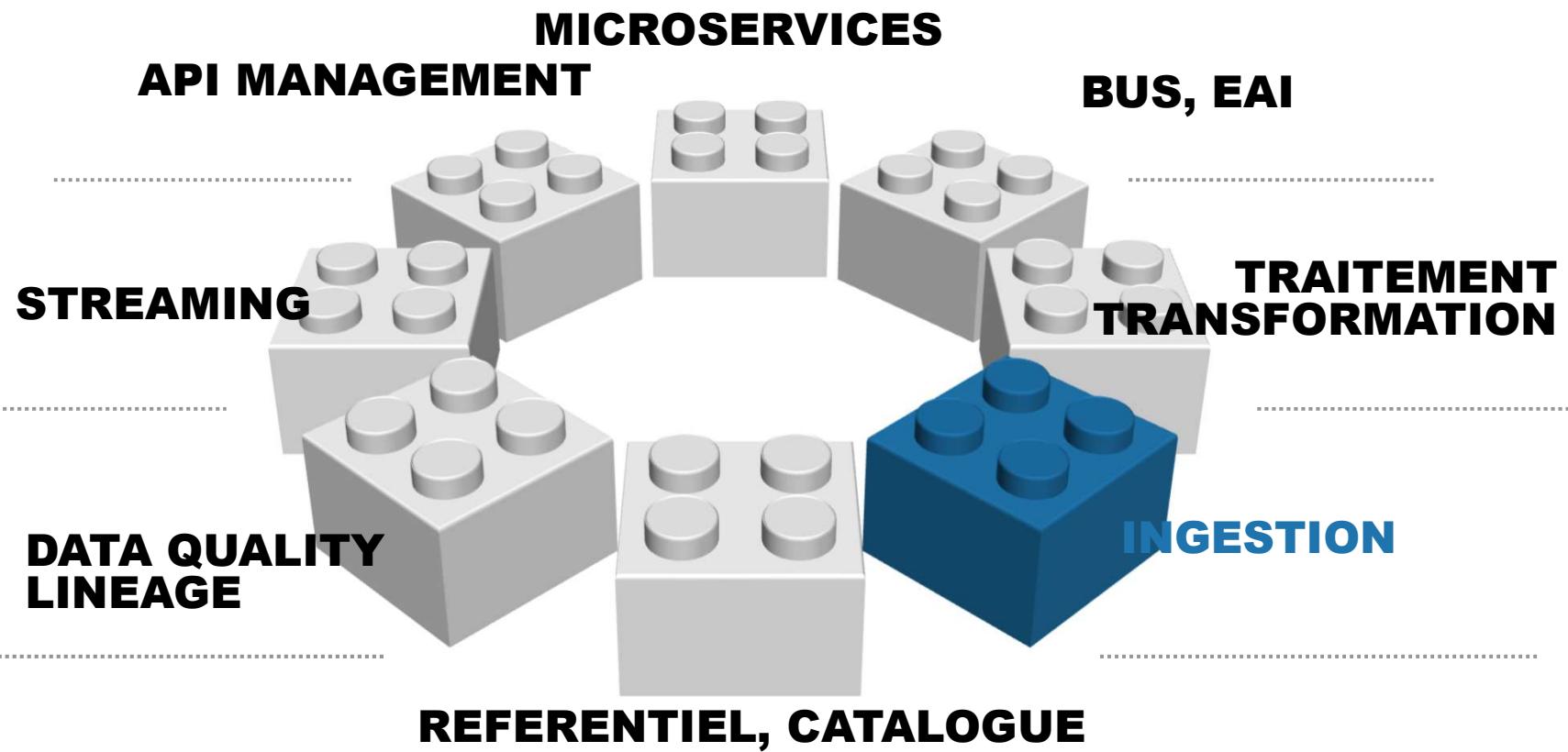


DataLab, IA, ...

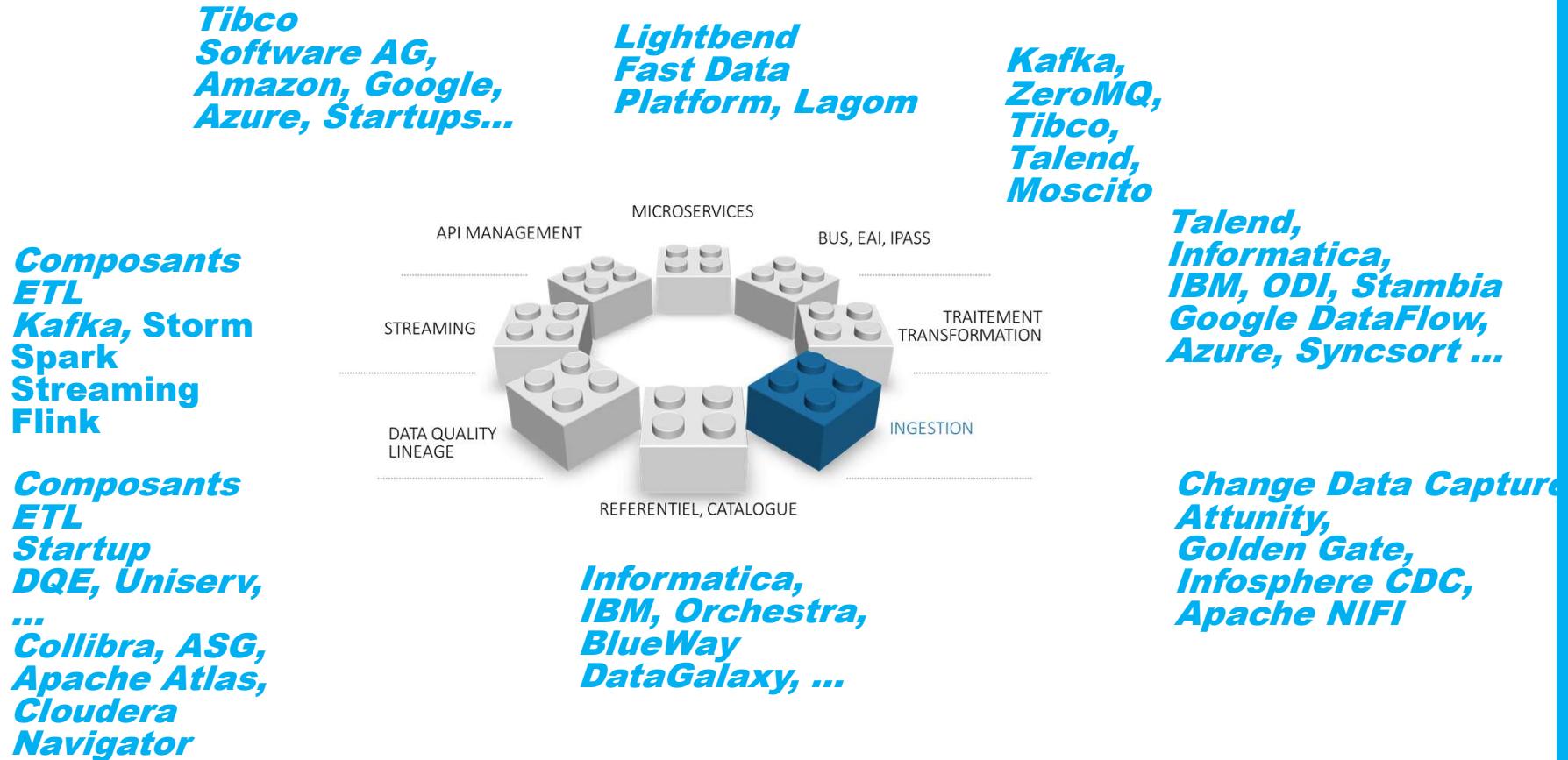
## PLATEFORMES HYBRIDES



# Fondements basés sur le traitement de la donnée



# Focus rapide sur les solutions



# FOCUS SUR HADOOP: DISTRIBUTIONS

Business Intelligence

ORACLE®

TERADATA.



Microsoft

Big Data

cloudera

MAPR®



# FOCUS SUR HADOOP: DISTRIBUTIONS

## ➤ Cloudera

Le vétéran ce qu'il lui donne une légitimité et un nombre de clients supérieur à ces concurrents. Un autre avantage est de disposer dans ses rangs de Doug Cutting le créateur d'Hadoop. Cloudera est très prompt à sortir les dernières versions d'Hadoop.

## ➤ Hortonworks

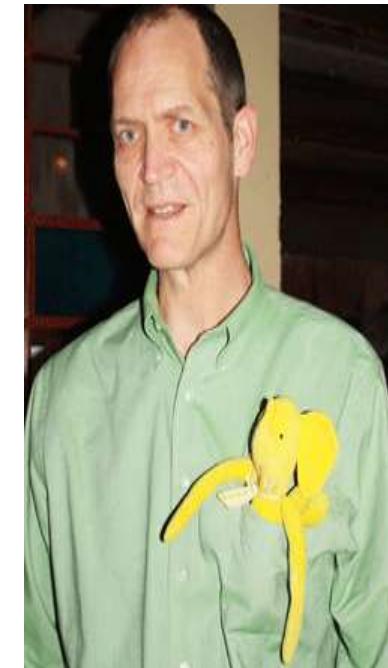
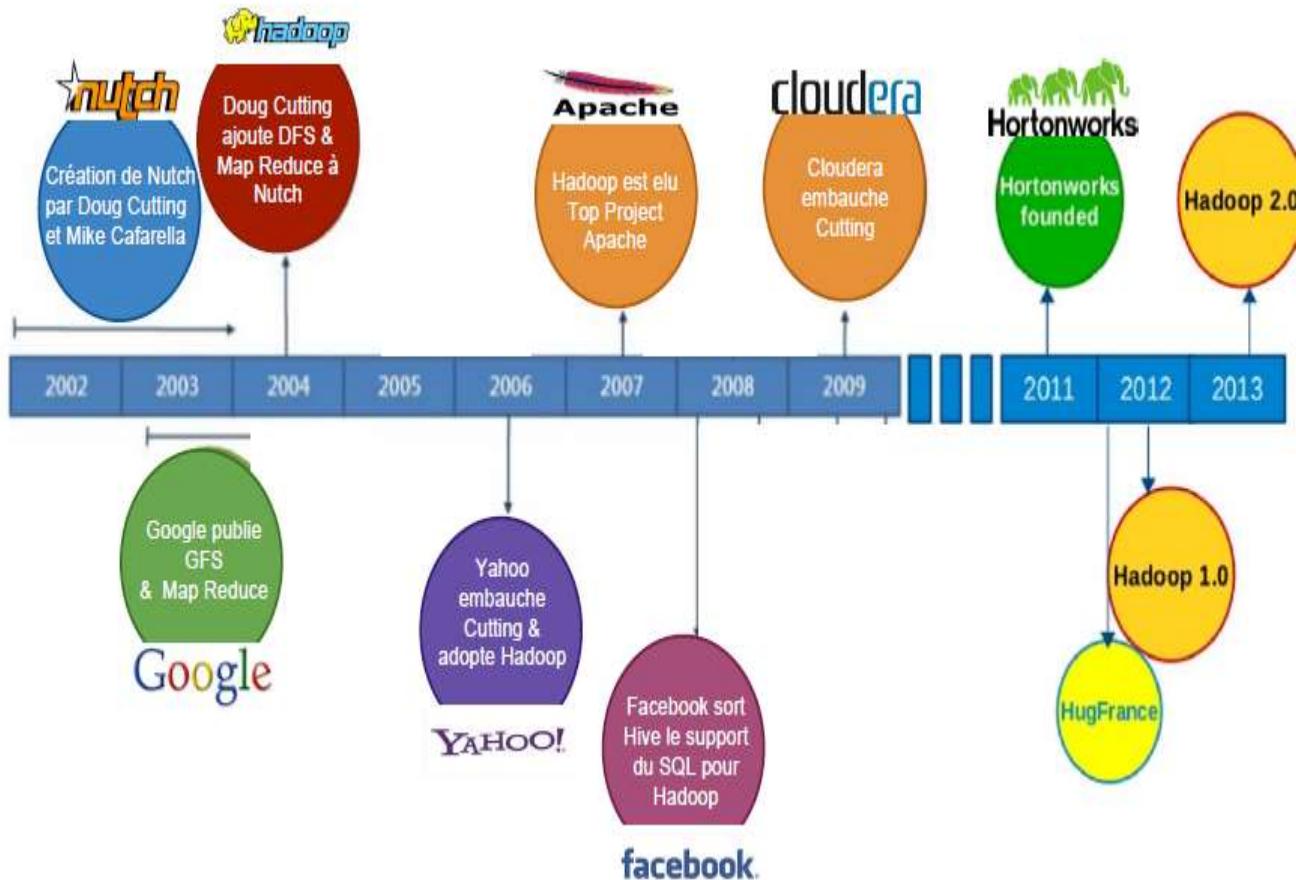
C'est la seule plateforme 100% Apache Hadoop. La stratégie assumée d'Hortonworks est de se baser sur les versions stables et testées d'Apache Hadoop. Malgré sa relative jeunesse, Hortonworks est la seule distribution en bourse et a signé des partenariats importants. IBM a décidé d'arrêter le développement et la vente de BigInsights et deviennent partenaire Hortonworks.

## ➤ MapR

La plus éloignée d'Apache Hadoop car elle intègre des versions propriétaires de MapReduce et HDFS.

C'est une solution avec plusieurs composants spécifiques.

# FOCUS SUR HADOOP : HISTORIQUE



# FOCUS SUR HADOOP : POURQUOI ?

## ➤ Problématique

Lire un fichier de 3TB sur un disque prend presque 4 heures

- Nous ne pouvons traiter les données avant de l'avoir lu la totalité du fichier
- Nous sommes limités par la vitesse du disque

## ➤ Solution : utiliser plusieurs disques en parallèle

- La vitesse de transfert d'un disque est à peu près de 210 MB/s
  - 4 heures pour 3 TB
- Si nous mettons 1000 disques en parallèle:
  - Moins de 15 secondes pour lire 3TB

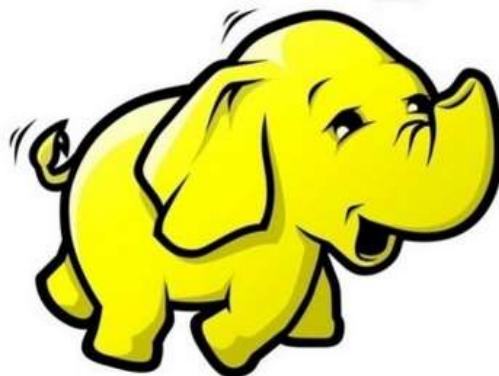
## ➤ Tolérance à la panne → c'est le logiciel qui traite la panne hardware

- 1 serveur risque de tomber en panne tous les 5 ans (1825 jours)
- Un cluster de 2000 nœuds doit gérer une panne par jour

# FOCUS SUR HADOOP : POURQUOI ?

Apache Hadoop est un framework qui va permettre le traitement de données massives et distribué sur un cluster allant d'une à plusieurs milliers de machines.

- C'est une solution logicielle



# FOCUS SUR HADOOP : POURQUOI ?

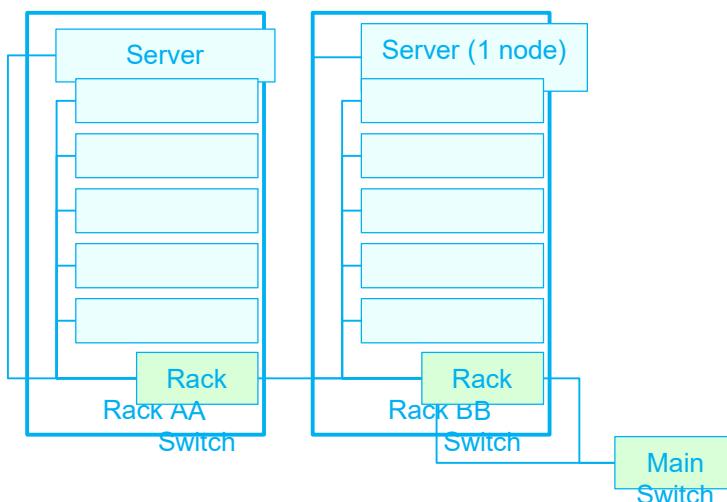
- Le projet Hadoop repose sur 3 composants majeurs:
  - Stockage des données : **HDFS** (*Hadoop Distributed File System*)
  - Traitement des données : **MapReduce**
  - Gestion des ressources : **YARN**
- Principe :
  - Diviser les données
  - Les sauvegarder sur une collection de machines, appelées cluster
  - Traiter les données directement là où elles sont stockées, plutôt que de les copier à partir d'un serveur distribué
- Il est possible d'ajouter des machines à votre cluster, au fur et à mesure que les données augmentent

# Topologie: Cluster

Cluster : Ensemble de serveurs sur un réseau

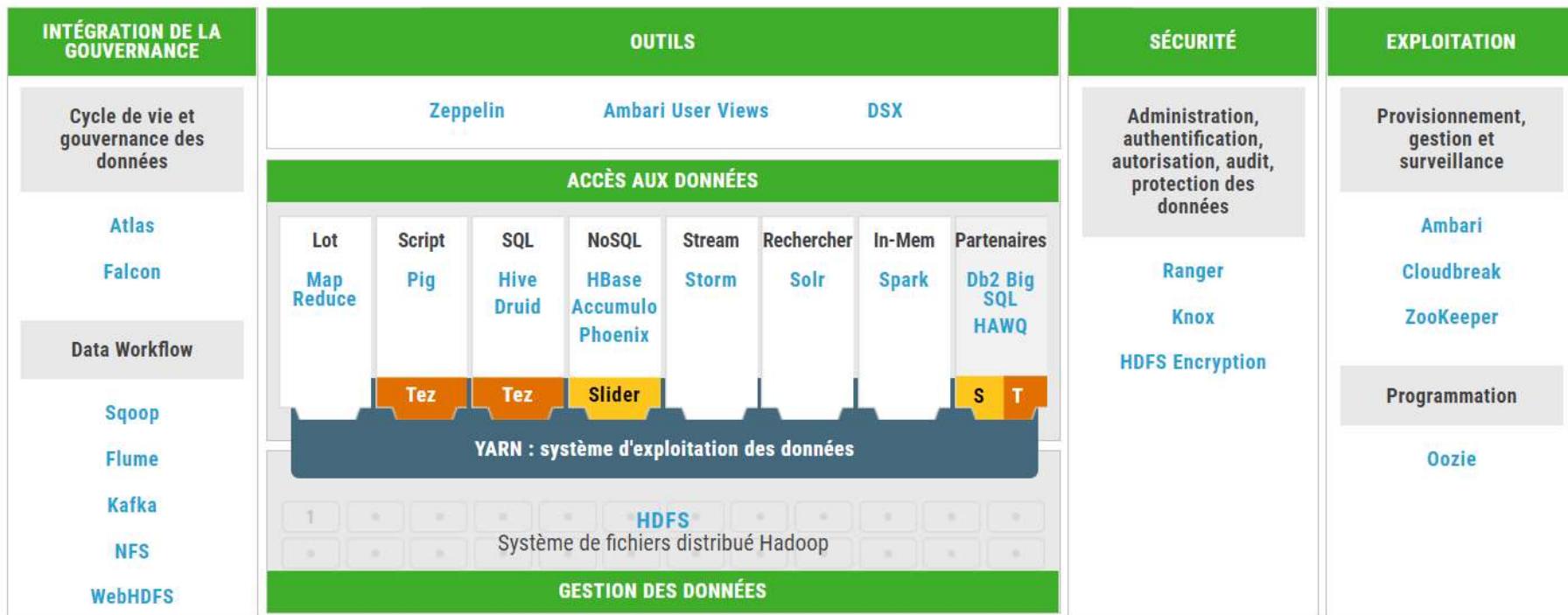
## Topology classique d'un cluster Hadoop:

- Plusieurs serveurs qui tournent sur plusieurs rack
- Chaque rack contient entre 24 et 40 serveurs
- Connectés à 1GB



Yahoo : plus de 40 000 noeuds  
167

# FOCUS SUR HADOOP : HDP



# FOCUS SUR HADOOP : HDFS

## □ Hadoop Distributed File System

### □ Définition :

- Gestionnaire d'espace de stockage
  - Stocker et mettre à disposition des données
  - Organiser avec des répertoires et des fichiers (arborescence)

### □ Exemples :

- NTFS              (Windows)
- FAT32            (Clé USB)
- HFS              (Mac)
- ext3/ext4        (Linux)

# FOCUS SUR HADOOP : HDFS

## □ Particularités :

- Plusieurs supports de stockage
- On ne sait pas où sont stockés les fichiers
- Scalabilité horizontale
- Plusieurs répliques en même temps
- Très haute disponibilité
- Tolérance aux pannes

## □ Ratio de réPLICATION :

- Espace disque / Tolérance aux pannes

## □ Exemples de systèmes de fichiers distribués:

- GFS
- OrangeFS
- HDFS

# FOCUS SUR HADOOP : HDFS

## □ Particularités :

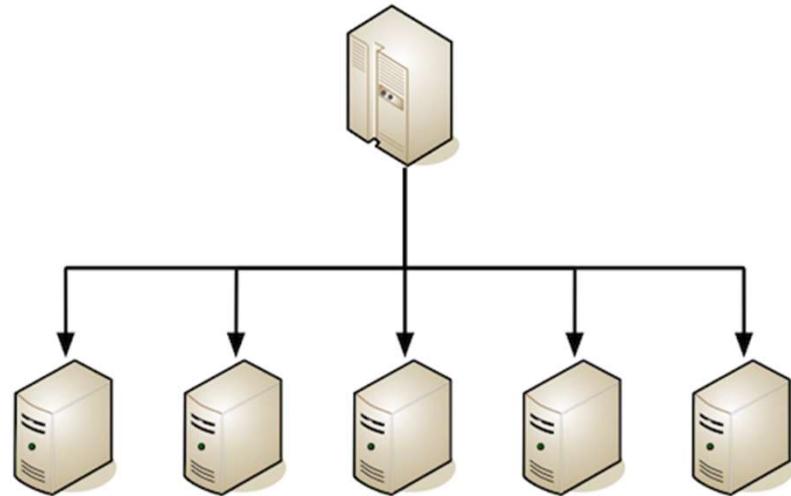
- Découpage du fichier en blocs
- Blocs de taille fixe ( taille à définir)
- Espace de stockage découpé par bloc
- Si un fichier (1MB) est plus petit que le bloc :
  - Hadoop s'appuie sur le syst de gestion de gest de fichier natif, par exemple ext3, l'espace perdu sera approximativement de la taille d'un bloc dans le système natif, 4 ou 8 Ko dans le cas ext3.

## □ Résumé :

- Système de fichiers
- Par blocs
- Distribué
- write once, read multiple

# FOCUS SUR HADOOP : HDFS

- Economique
- Tolérant aux pannes
- Evolutif scalabilité horizontale



# FOCUS SUR HADOOP : HDFS

□ Un cluster HDFS est basé sur une architecture Maître/Eclave

□ Un cluster HDFS est composé de:

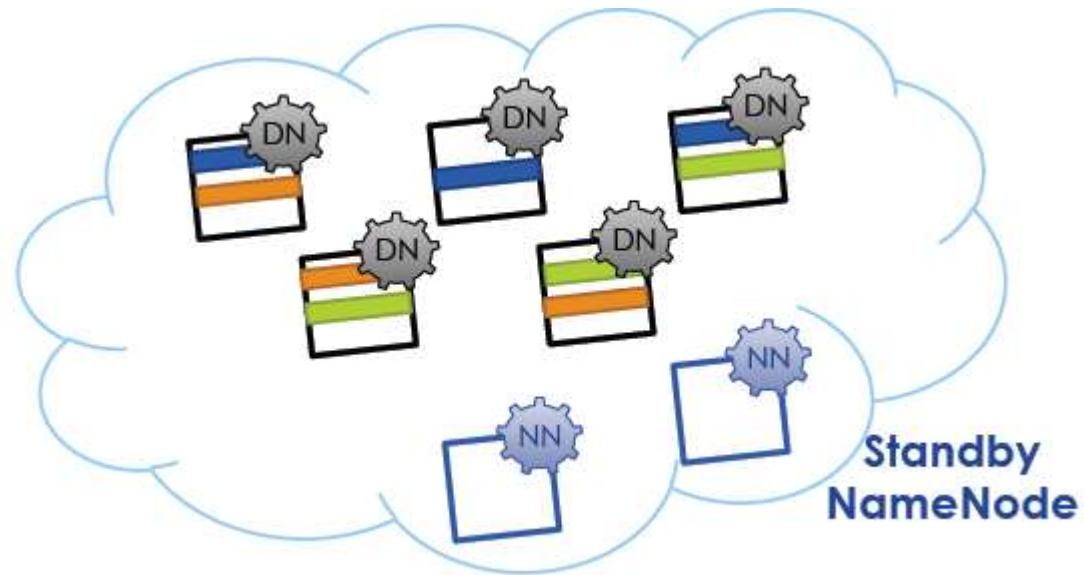
□ Un NameNode

- Gère l'espace de noms dans le système de fichier (NameSpace).
- Il a une connaissance des DataNodes.
- Gère l'accès aux données par les programmes clients.
- Fonctionne sur la mémoire vive pour plus de réactivité.
- Stock sur le disque le NameSpace dans un fsimage (principalement) et un editlog (pour les modifications).

□ Plusieurs DataNodes

- Espace de stockage des données.

# FOCUS SUR HADOOP : HDFS



# HDFS : LECTURE

- ❑ Hadoop est plus efficace pour traiter de gros fichiers plutôt qu'un grand nombre de petit fichier
- ❑ Chaque info relative à un fichier (nom du fichier, permission, localisation des blocs,...) consomme 200 octets de RAM
- ❑ Un fichier de 1 Go avec une taille de bloc de 128Mo consommera :
  - 200 octets pour le nom du fichier
  - $8 * 3 * 200 = 4\,800$  octets de RAM
- ❑ 1000 fichiers de 1 Mo avec une taille de bloc de 128Mo consommeront :
  - 200 octets pour le nom du fichier
  - $1000 * 3 * 200 = \text{600 000 octets}$  de RAM

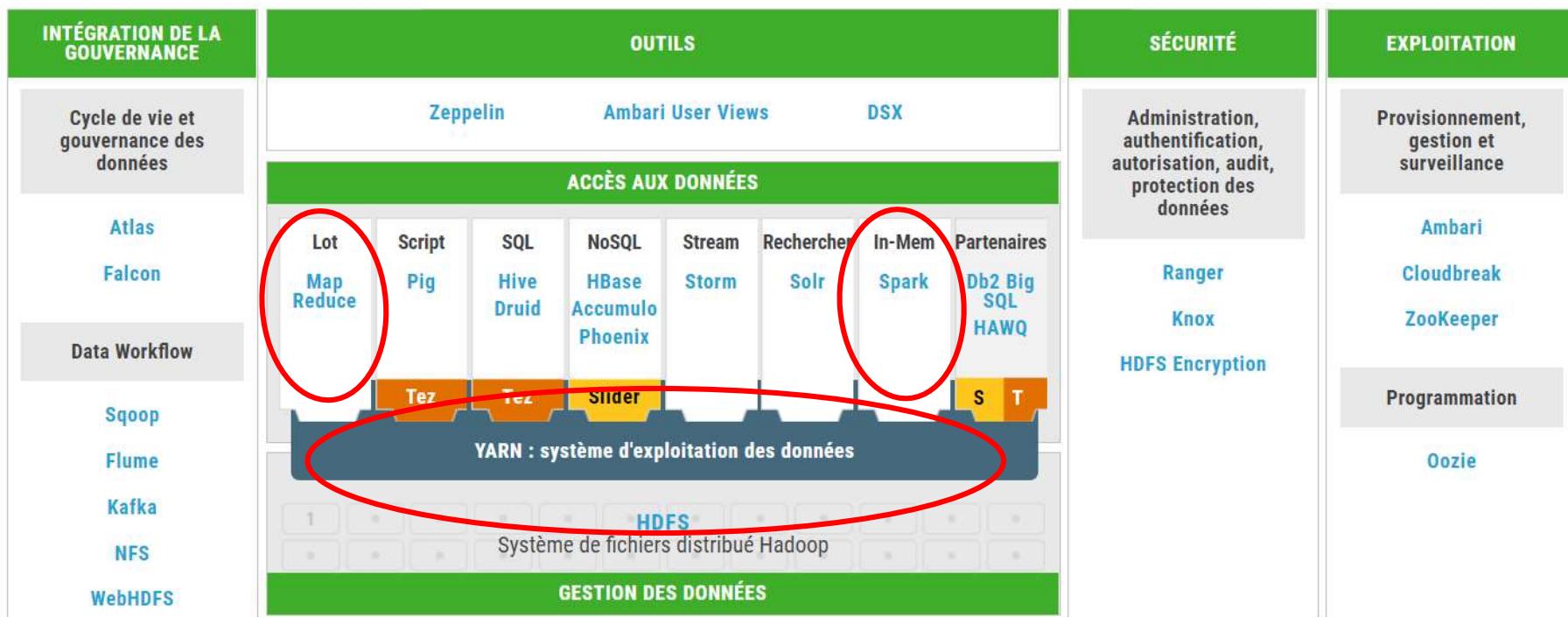
# HDFS : LES COMMANDES

- Il y a environ 30 commandes pour manipuler HDFS

- Lancer les commandes HDFS :  
**hdfs dfs**

**-cat** : affiche le contenu d'un fichier  
**-text** : comme cat mais sait afficher des données compressées  
**-chgrp, -chmod, -chown** : modification des permissions  
**-put, -get, -copyFromLocal, -copyToLocal** : import / export entre le système de fichier local et HDFS  
**-ls, -lsr** : liste les fichiers / répertoires  
**-mv, -moveFromLocal, -moveToLocal** : déplace les fichiers  
**-stat** : informations statistiques sur les ressources (taille des blocs, nombre de blocs, type de fichiers, etc.)

# FOCUS SUR HADOOP : HDP



# LE CALCUL DISTRIBUÉ

Désigne l'exécution d'un traitement informatique sur une multitude de machines différentes (un *cluster de machines*) de manière transparente.

## Problématiques:

- Accès et partage des ressources pour toutes les machines.
- Extensibilité: on doit pouvoir ajouter de nouvelles machines pour le calcul si nécessaire.
- Hétérogénéité: les machines doivent pouvoir avoir différentes architectures
- Tolérance aux pannes: une machine en panne faisant partie du cluster ne doit pas produire d'erreur pour le calcul dans son ensemble.
- Transparence: le *cluster dans son ensemble doit être utilisable comme une seule et même machine « traditionnelle »*.

# LE CALCUL DISTRIBUÉ

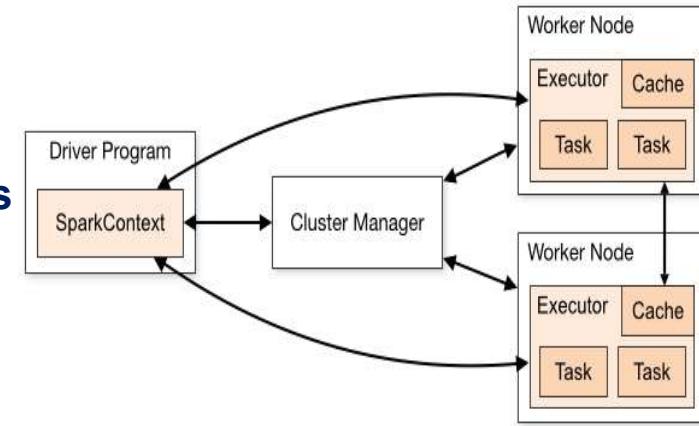
Pour manipuler de grands volumes de données, des architectures classiques peuvent répondre au besoin.

Les serveurs dans de telles architectures sont constitués d'un nœud maître et de plusieurs calculateurs ou dispositifs de stockage appelés workers ou esclaves. L'ensemble de ces machines est appelé « Cluster ».

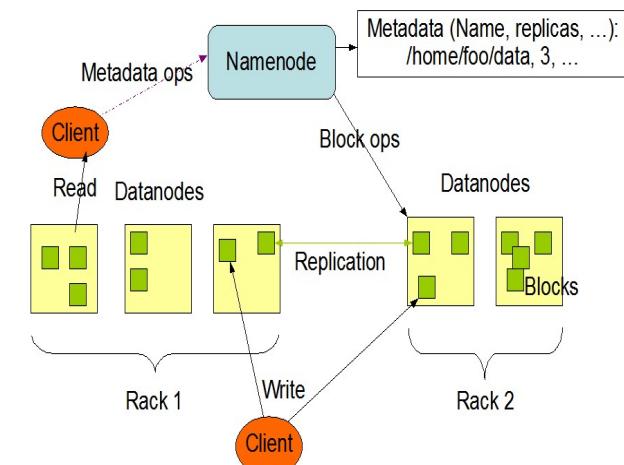
Quand en plus du volume des données, le temps de réponse devient aussi un critère important, des architectures basées sur le stockage et le calcul distribué deviennent nécessaires.

Les architectures distribuées permettent d'accélérer les temps de traitement: agrégation, assemblage, calcul, transformation, exploration.

Ces systèmes permettent aussi de stocker et traiter des données non structurées et facilitent l'exploration.

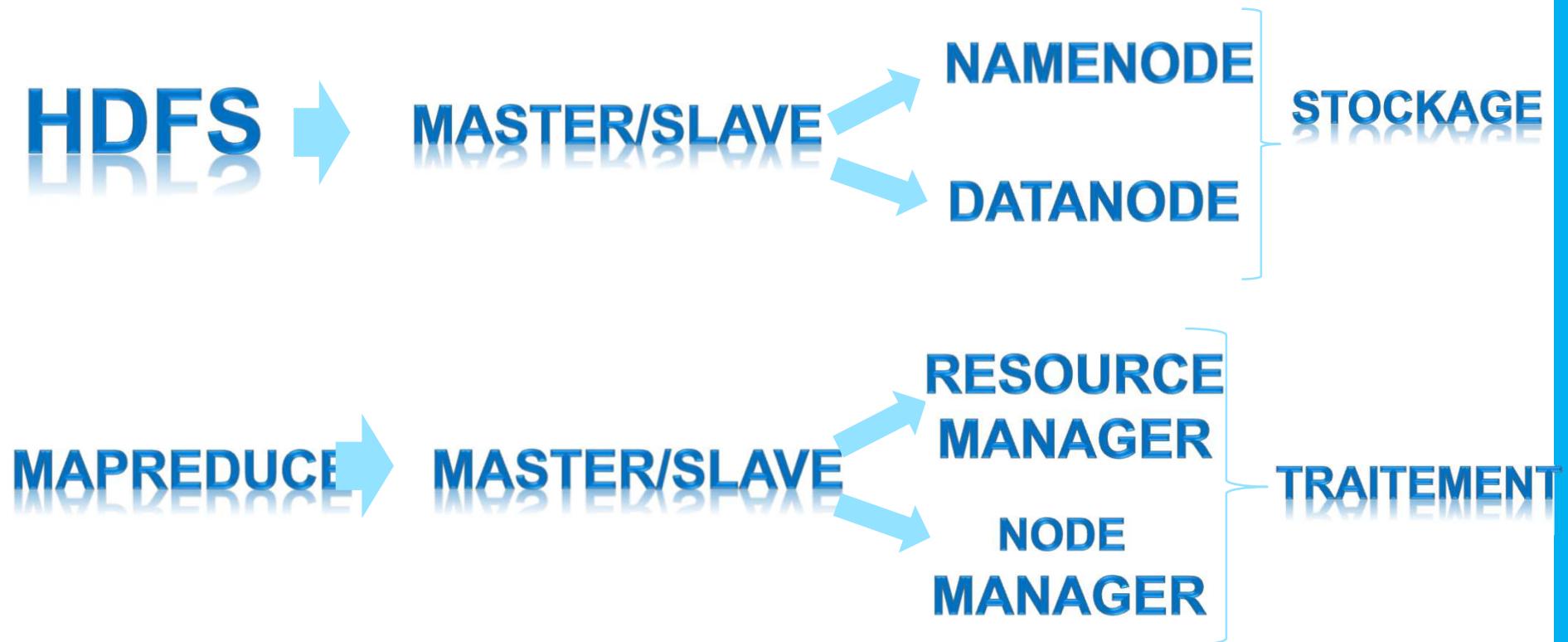


HDFS Architecture



Sources: <https://spark.apache.org/docs/latest/cluster-overview.html>  
[https://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)

# MAP REDUCE

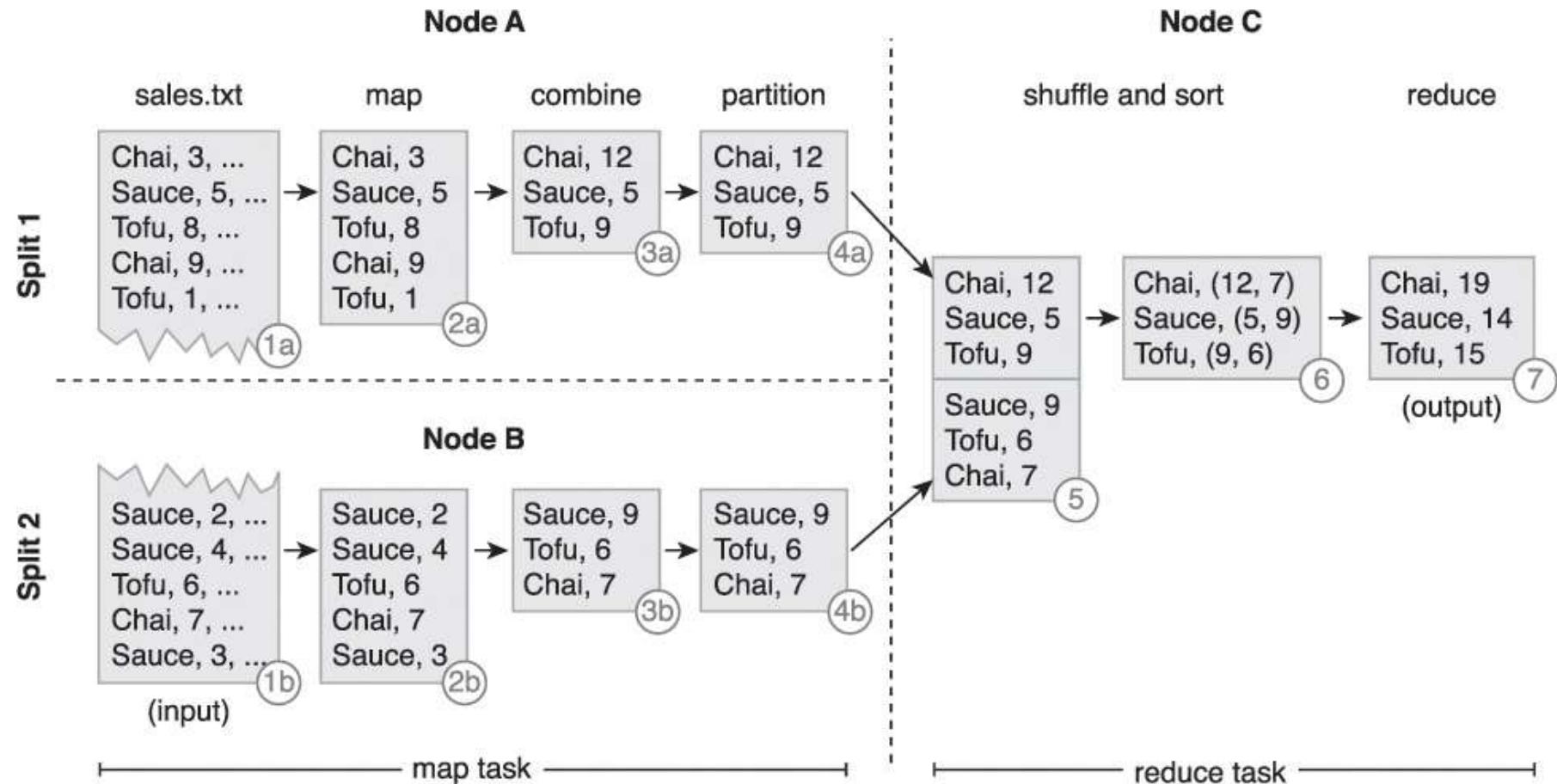


# MAP REDUCE

**Paradigme de calcul distribué.** Au lieu de parcourir les données séquentiellement, il faut pouvoir les découper pour exécuter une partie du traitement sur une machine du cluster.

C'est un paradigme qui existait depuis longtemps mais c'est le *whitepaper* issu du département de recherche de Google publié en 2004 (« MapReduce: Simplified Data Processing on Large Clusters ») qui a permis de normer et d'étendre ce modèle d'architecture .

# MAP REDUCE



Source: Big Data Fundamentals: Concepts, Drivers & Techniques  
Auteur: Paul Buhler; Wajid Khattak; Thomas Erl  
Publié par Prentice Hall, 2016

# LE CALCUL DISTRIBUÉ

On distingue deux types de traitements distribués:

- ❑ Mode batch: Stockage et ensuite traitement de la donnée
- ❑ Mode Streaming: Aucun stockage, traitement de la donnée en une seule passe.

Exemple de traitement en mode batch: Deezer calcul consolide les préférences utilisateurs pendant la nuit.

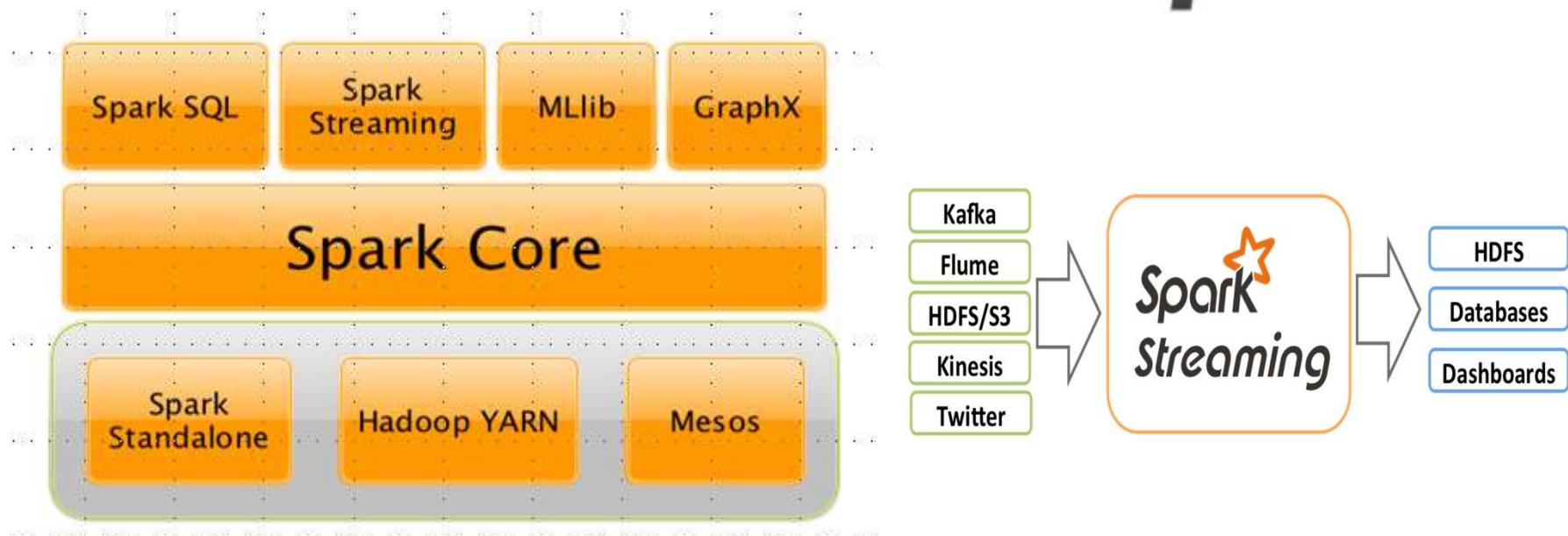
Exemple de traitement en mode streaming: Analyse des logs provenants de la plateforme de production pour détecter des intrusions.

# SPARK

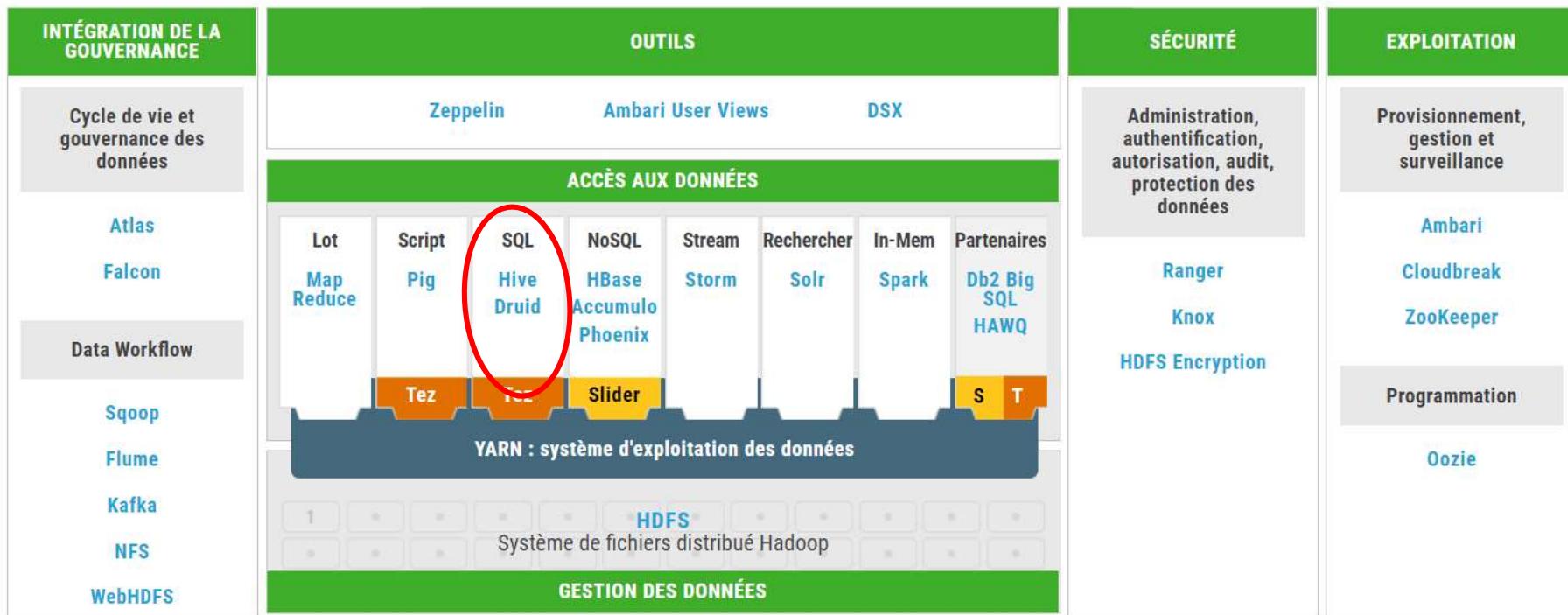
Framework de calcul distribué (fonctionne en mémoire)

Api Scala, java, python, R

Beaucoup plus rapide que Map Reduce



# FOCUS SUR HADOOP : HDP



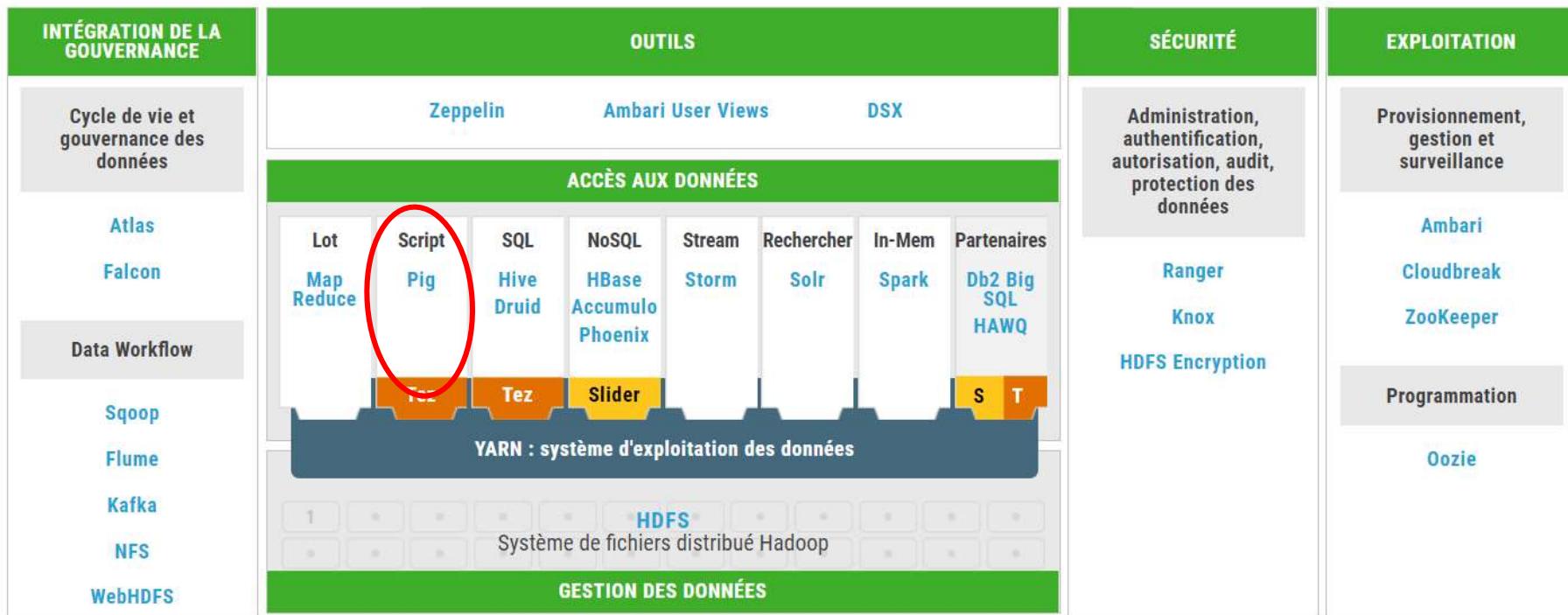
# HIVE

- Entrepôt de données pour Hadoop Langage semblable à SQL, appelé HiveQL ( HQL )

**N'est pas conçu pour :**

- Le traitement des transactions en ligne
- Des requêtes en temps réel

# FOCUS SUR HADOOP : HDP

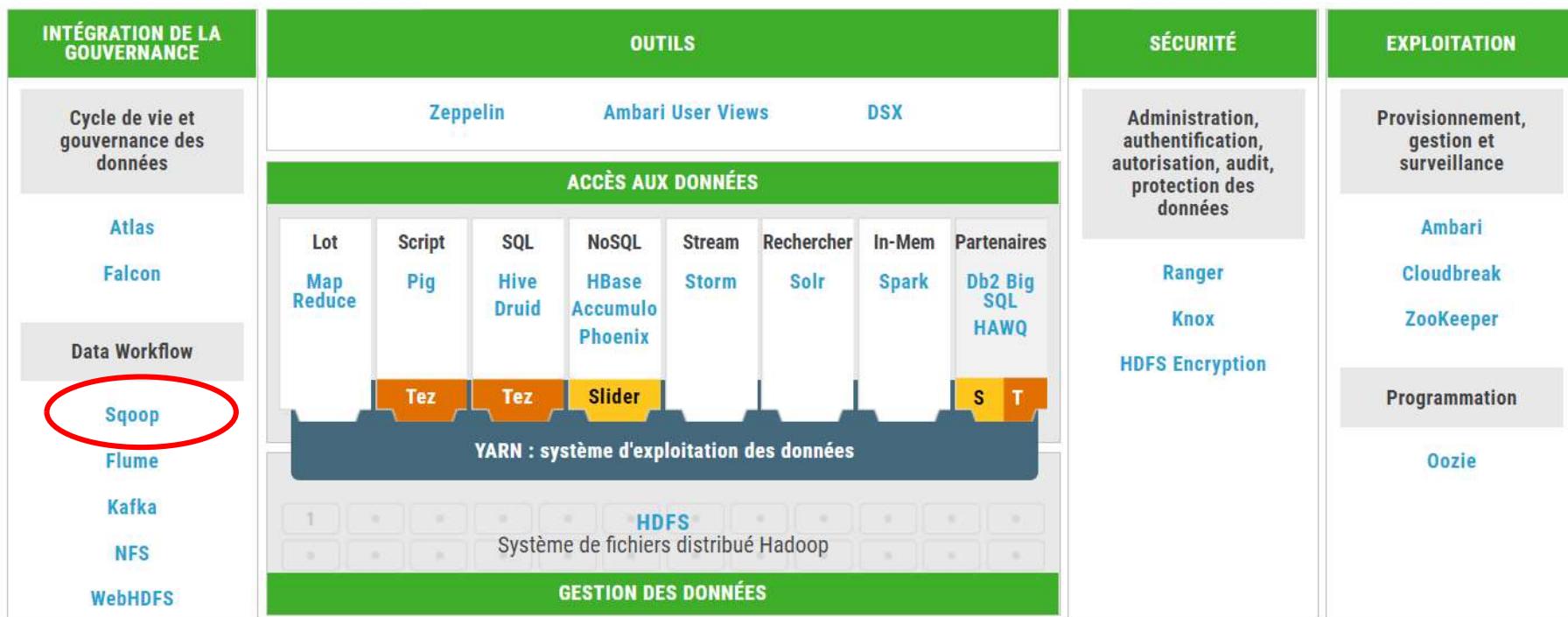


# PIG

## Créé chez Yahoo!

- Une plate-forme très simple pour traiter les Big Data
- **PigLatin** : langage dont le traitement est en flux, simple, très efficace
- **Pig Engine** : parse, optimise et exécute automatiquement les scripts PigLatin comme une série de jobs MapReduce au sein d'un cluster Hadoop

# FOCUS SUR HADOOP : HDP



# SQOOP

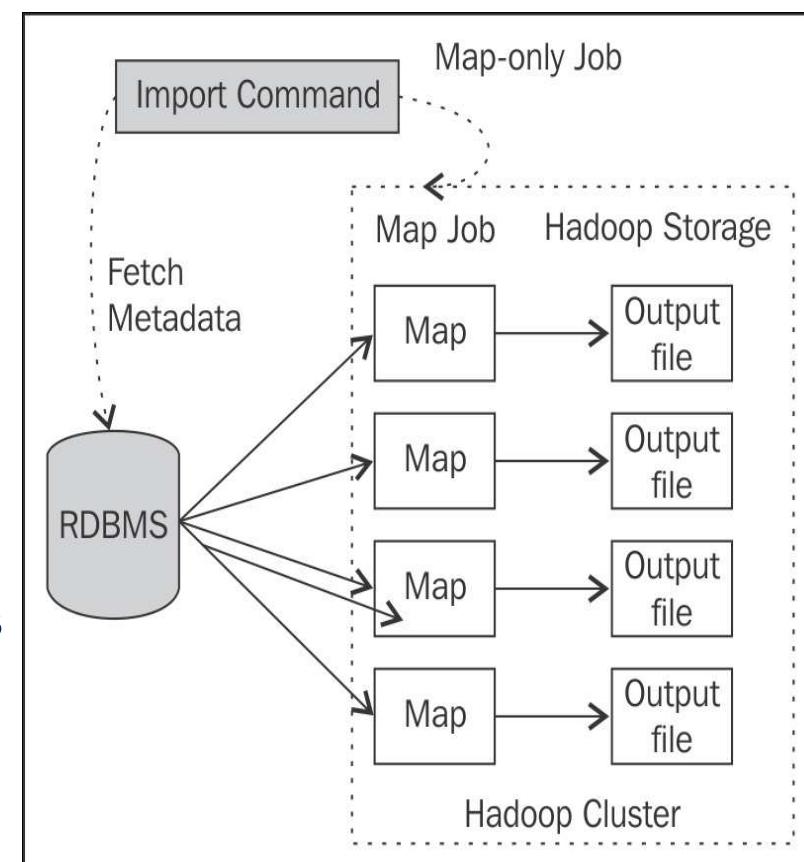


Sqoop permet d'importer des données dans le Data Lake depuis de données relationnelles et vice-versa.

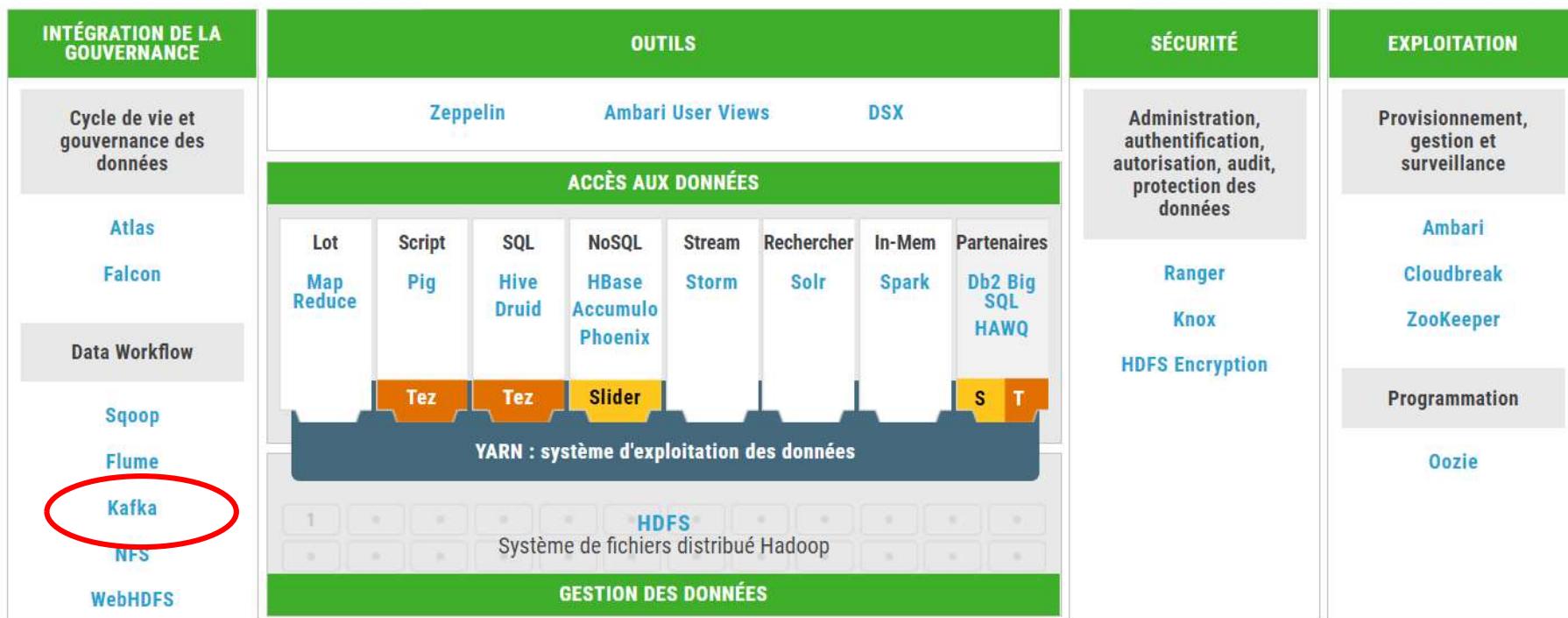
Dans le Data Lake Sqoop injecte les données sous formes de fichiers CSV ou directement dans Hive ou Hbase.

Fonctionnalités majeures:

- Importer des tables entières
- Importer un sous ensemble des données
- Importer de manière incrémentale

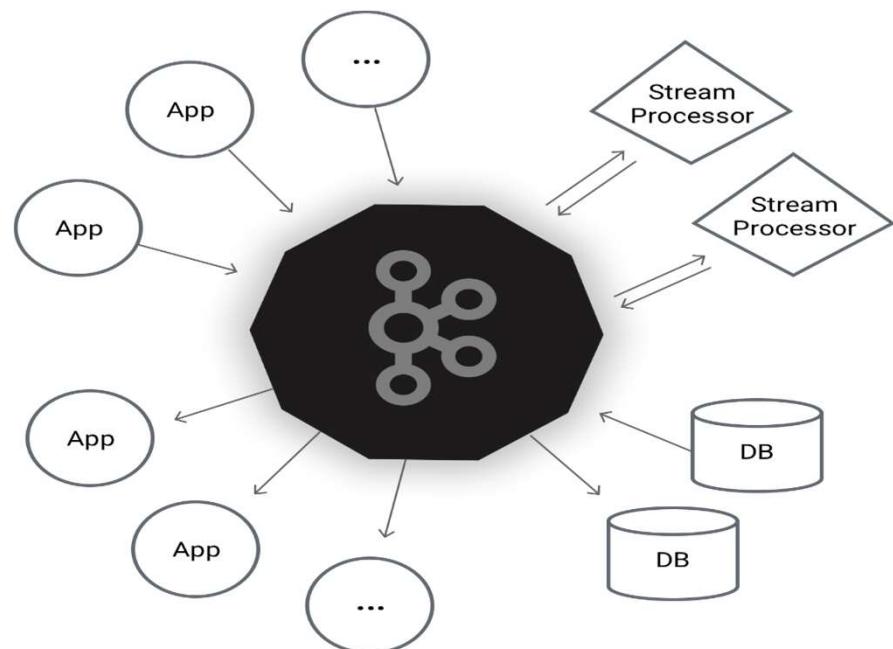
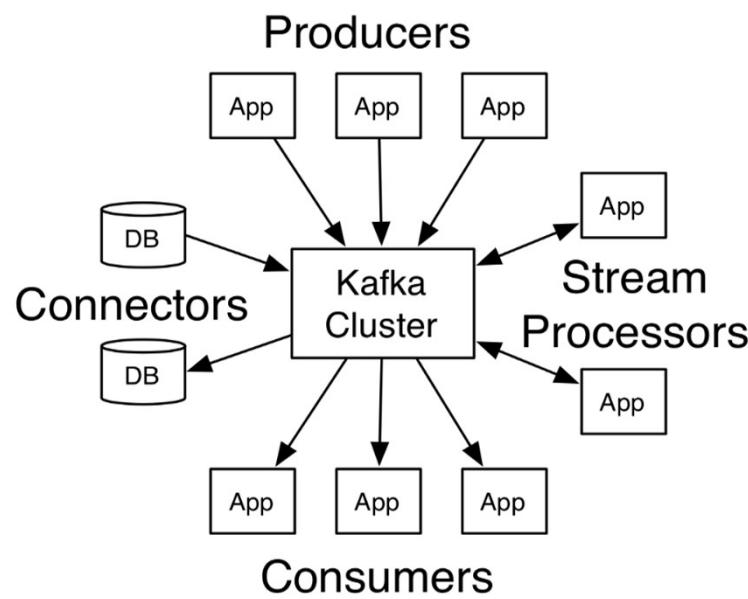


# FOCUS SUR HADOOP : HDP



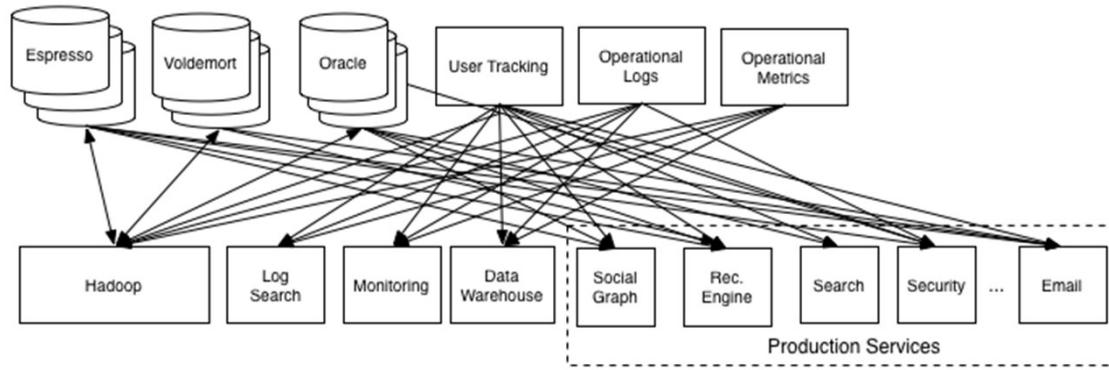
# KAFKA

- Apache Kafka est une plateforme de streaming distribué. C.-à-d. :
  - On peut publier et souscrire à un Streaming d'enregistrement
  - Permet de stocker nos enregistrements (tolérance aux pannes, durée de rétention)
  - Permet de traiter des flux d'enregistrement au fur et à mesure qu'ils se produisent.

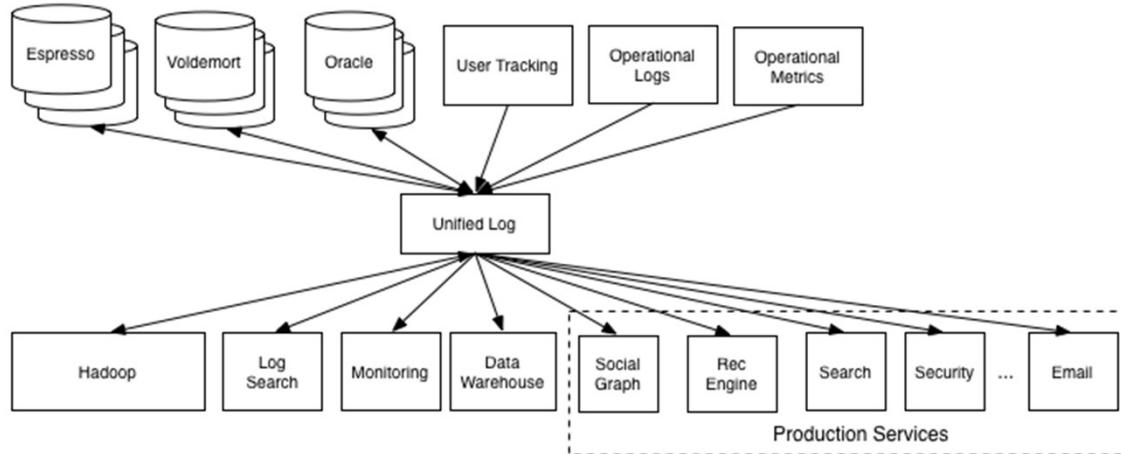


# POURQUOI KAFKA?

On est parti de :



À :



# **MOTEUR D'INDEXATION : SOLR, ELASTICSEARCH**

**Moteurs de recherche full text**

**Permet de faire de la recherche multicritère**

**Les 2 produits sont basés sur les Indexes Lucène**

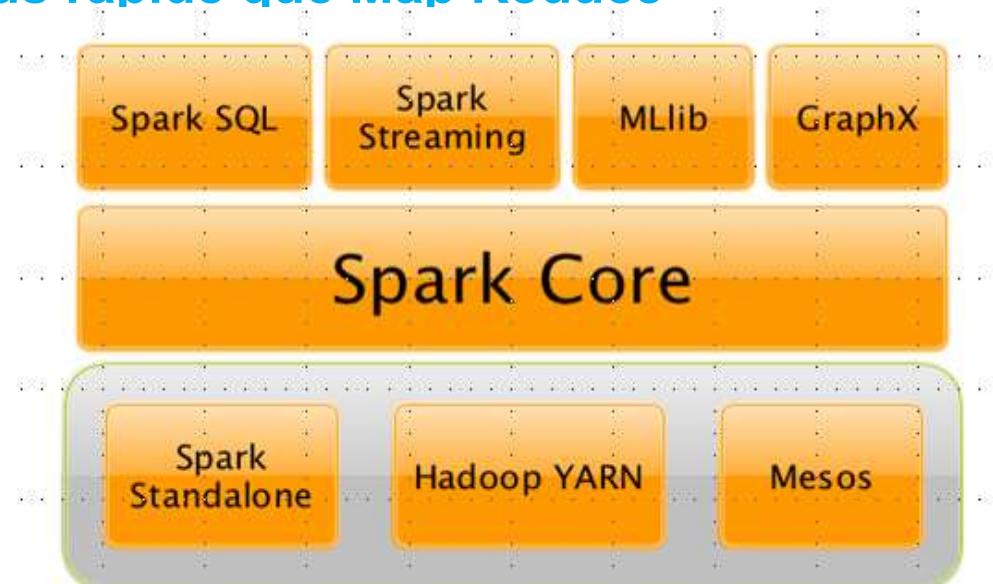


# SPARK

Framework de calcul distribué (fonctionne en mémoire)

Api Scala, java, python

Beaucoup plus rapide que Map Reduce



# NIFI

Outil de médiation de données permettant de mettre en place très rapidement des échanges entre systèmes internes ou externes. Il supporte quasiment tous les protocoles et dispose de plus de 140 connecteurs, à la fois en lecture (Get) et écriture (Put), en particulier pour HDFS, Hbase, ElasticSerach, Kafka, JDBC, etc.

Ce n'est pas un ETL. Il n'a pas vocation à transformer et enrichir les données, mais il peut changer le format des données (JSON, XML, CSV, AVRO, ...) et les aggréger.



# **NIFI CE N'EST PAS ....**

## **Un moteur de traitement distribué**

- Permettant de faire du CEP (Complex Event processing)
- Permettant de faire des jointures ou des opérations comme Spark/Storm/Flink

**Nifi n'a pas de dépendance avec d'autres outils big data comme Hadoop ou Zookeeper. Nifi n'a besoin que de java pour fonctionner**

## **ETL type Informatica/Pentaho/Talend/SSIS**

**Un espace où l'on persiste la donnée sur le long terme. Garde la donnée temporairement pour retraitier la données en cas d'échec par exemple.**

# **NIFI PERMET.....**

**Garantit la livraison des données**

**Buffering, gestion du back pressure**

**Gestion des priorités dans les queues**

**Custom extension ( on peut écrire de nouveau processor,...)**

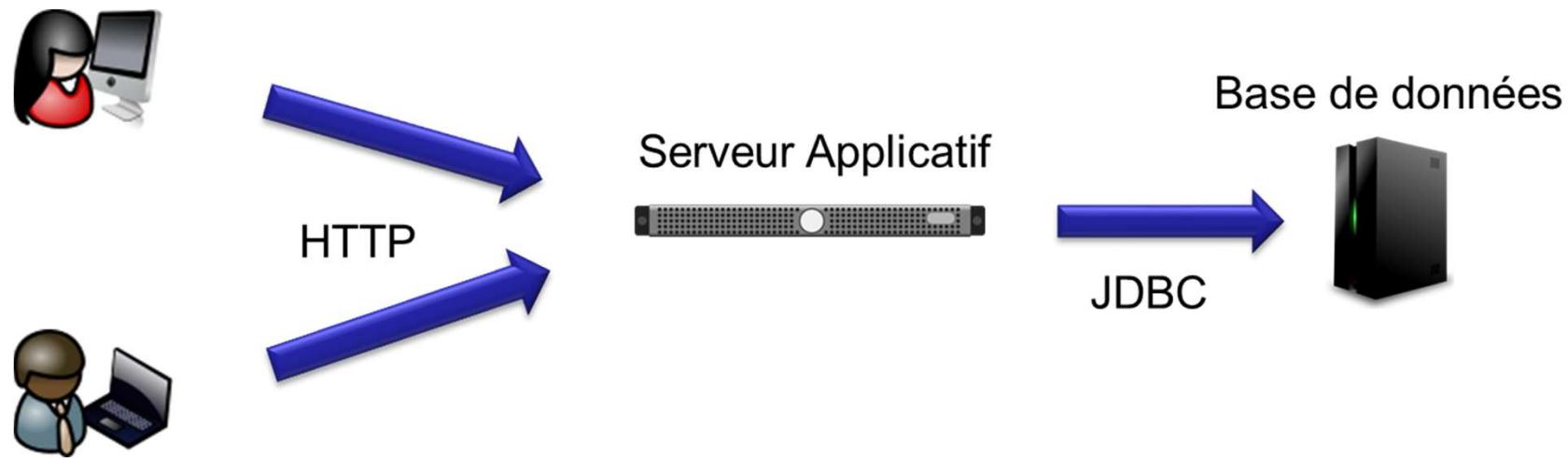
**Interface graphique (drag and drop, statistiques, ....)**

**Suivi de la provenance des données**

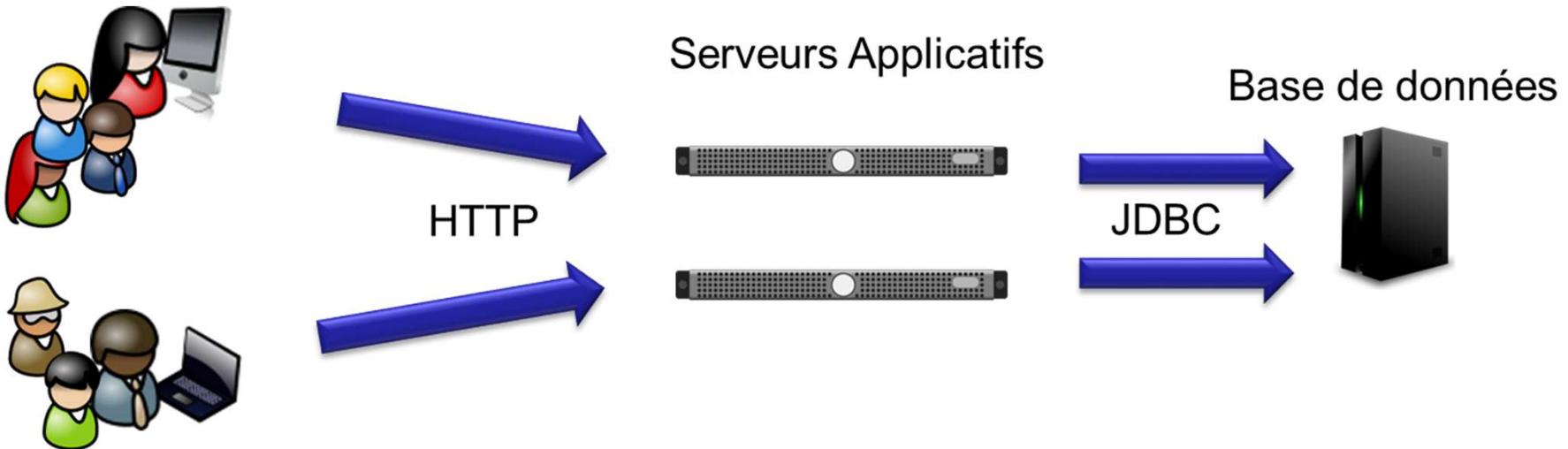
# BASES DE DONNÉES NOSQL?



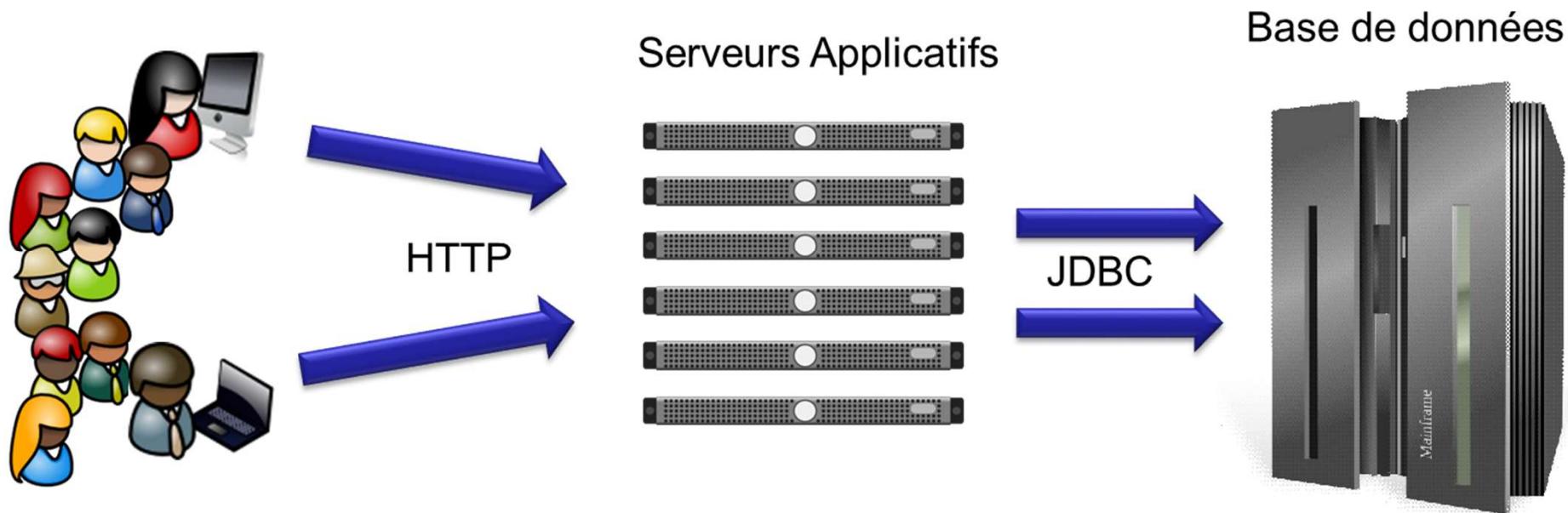
# BASES DE DONNÉES NOSQL?



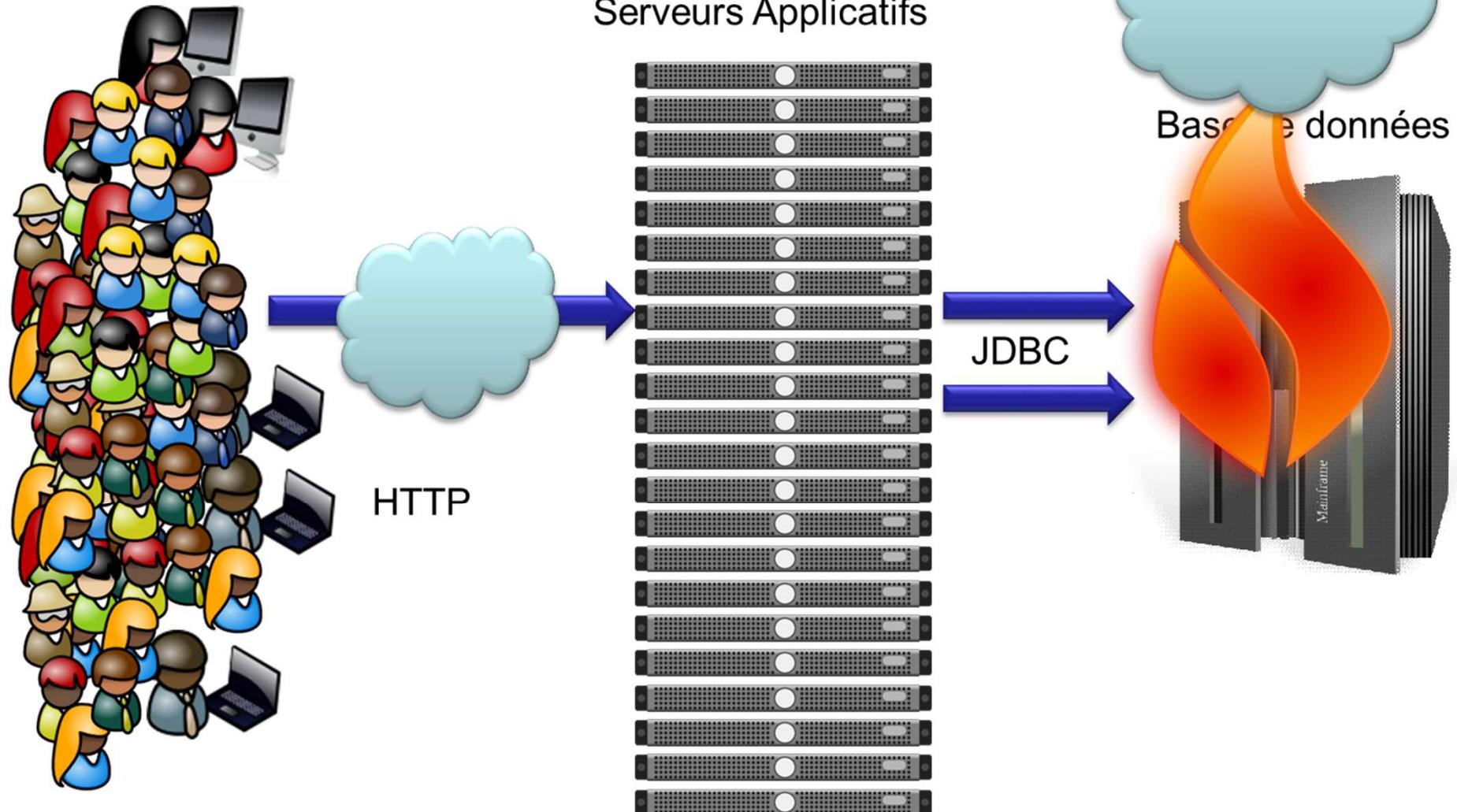
# BASES DE DONNÉES NOSQL?



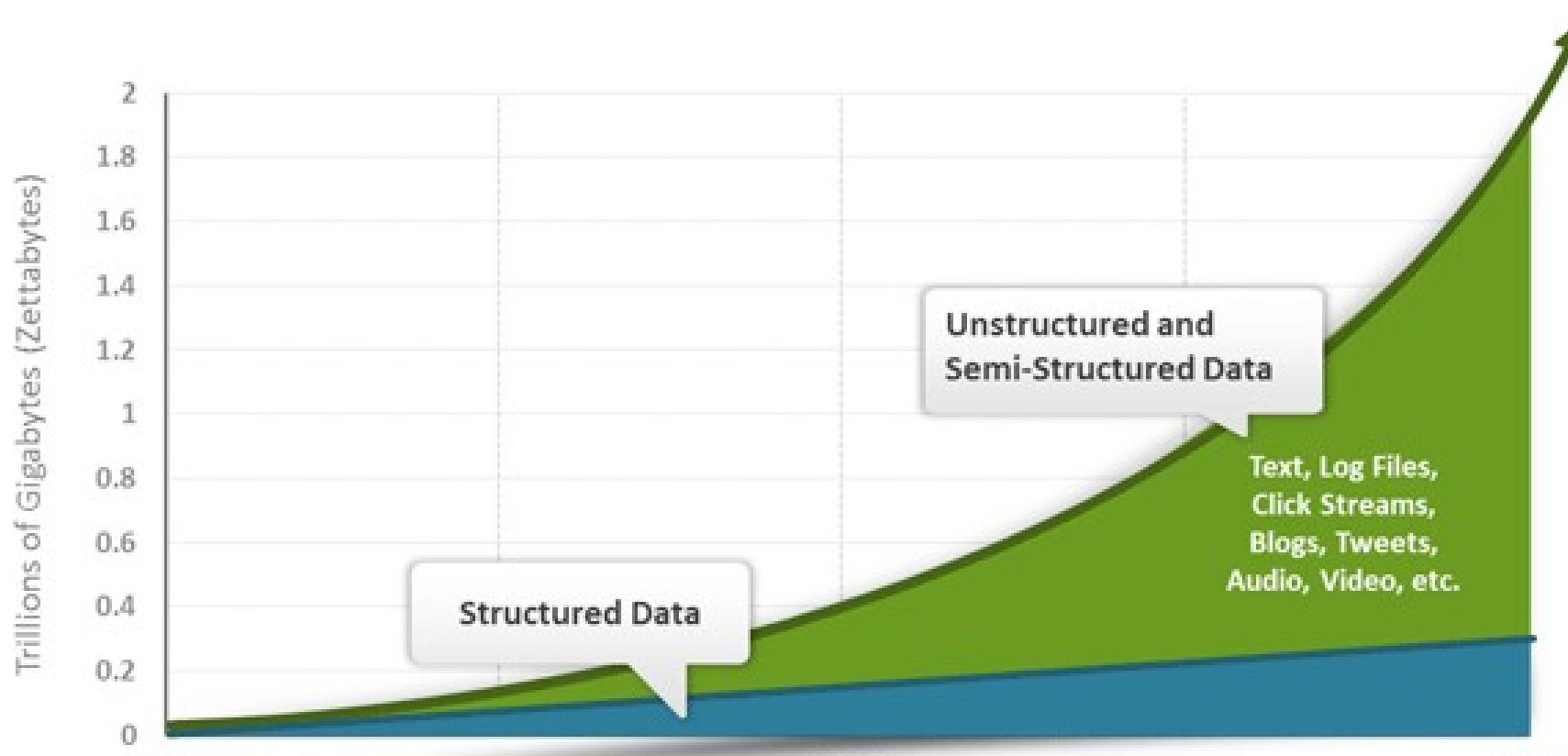
# BASES DE DONNÉES NOSQL?



# BASES DE DONNÉES NOSQL?



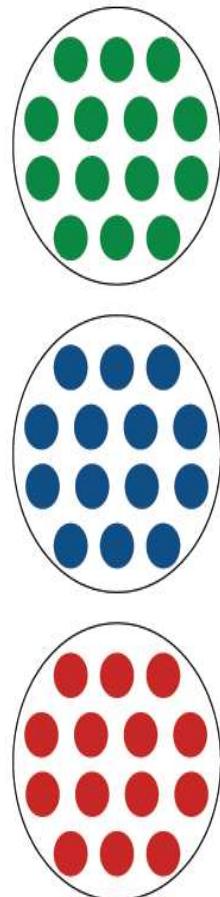
# BASES DE DONNÉES NOSQL?



Source: IDC 2011 Digital Universe Study (<http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>)

# LE PRINCIPE DES RGBD

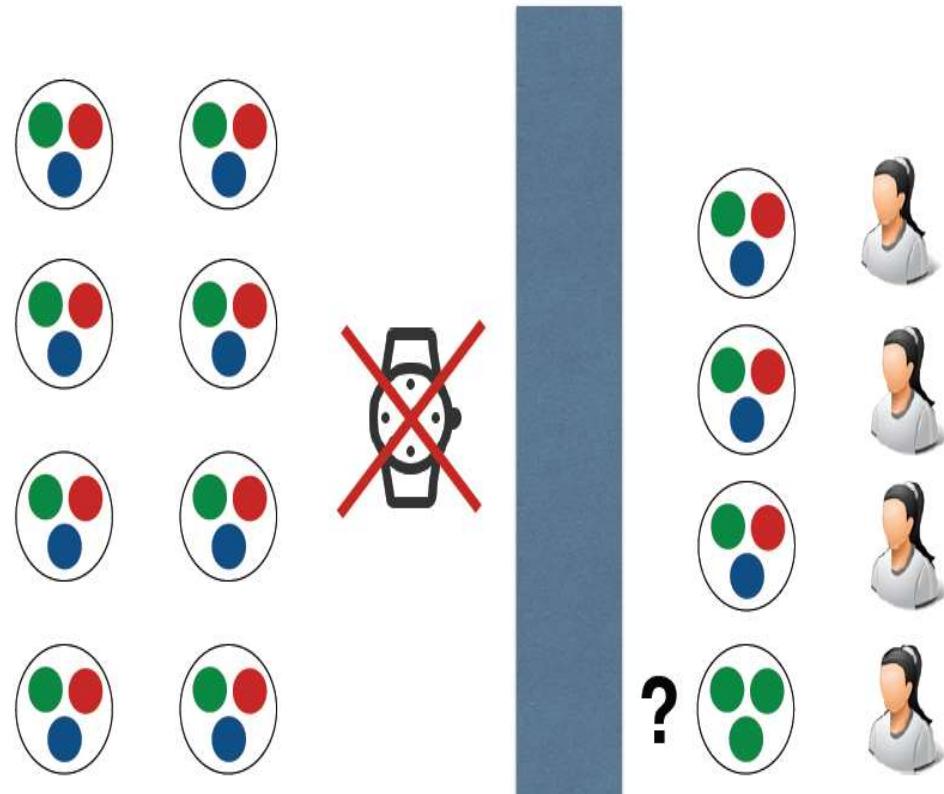
La chocolaterie



I-4. Penser usage

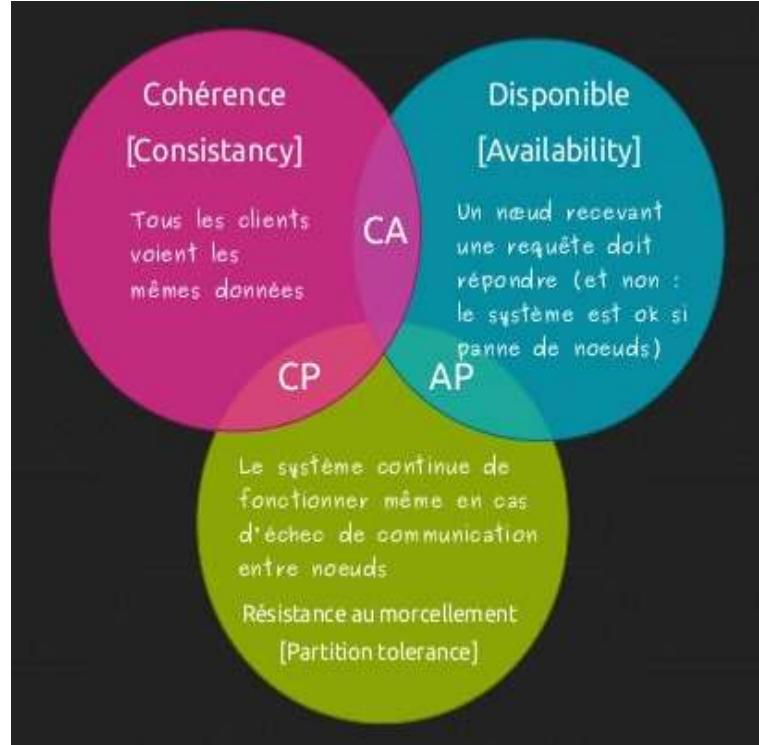
# LE PRINCIPE DES BASES NOSQL

La chocolaterie



I-4. Penser usage

# THÉORÈME DE CAP



Théorème CAP : « Il est impossible de satisfaire les trois propriétés CAP en même temps »

# LE CHOIX D'AMAZON

Lors qu'un client  
clique  
sur le bouton  
« acheter »  
Faut-il ?



# PERFORMANCES PRÉdictibles

- **La performance des opérations doit être prédictible**
- **Amazon:**
  - ✓ Perte de 1 % de chiffre d'affaire si le temps d'affichage des pages augmente de 0,1 s
  - ✓ Plan qualité interne : Temps de réponse doit être < 300 ms pour 99,9 % des requêtes pour un pic de 500 requêtes par secondes
- **Google pénalise les sites dont les pages s'affichent en plus de 1,5 s**

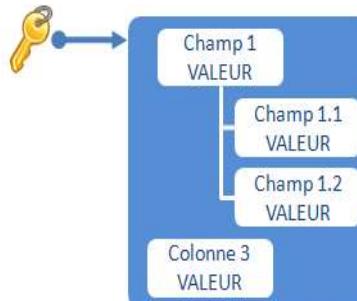
# LES TYPES DE BASES NOSQL

## Clé-valeur : ultra-rapide, ultra simplifié

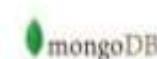


Représentation simple  
Permet d'indexer des informations diverses via une clé.  
Adaptée à la gestion des caches ou accès rapide à l'info.

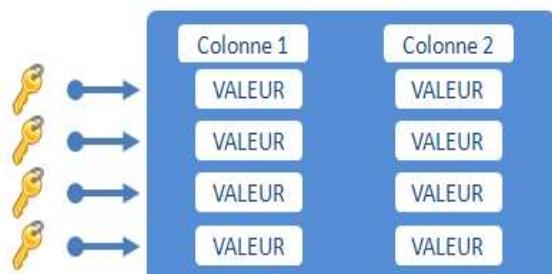
## Document : j'enregistre ce que je veux, comme je veux



Modélisation souple permettant de stocker des documents au format JSON dans des collections.  
Stockage de masse.  
Outils : MongoDB



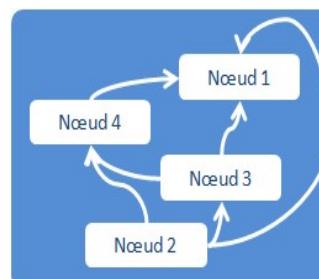
## Colonne : performance et consistance



Extension du système clé-valeur.  
Permet de stocker un grand nombre d'informations sur une même ligne, permettant de stocker des informations du type one-to many.  
Outils : HBASE, Cassandra



## Graphe : Analyse des relations



Modélisation optimisée pour les problèmes de graphes.  
Stockage provisoire pour traiter des données de type réseaux (sociaux, électrique, ...)  
Outils : Neo4j, Titan

# BASES CLÉS / VALEURS

## A utiliser pour ...

- ✓ De l'information très volatile
- ✓ Session utilisateur, données d'un panier d'achat
- ✓ De l'information très peu volatile et accédée très fréquemment
- ✓ Descriptions produits, paramétrage applicatif

## A éviter pour ...

- ✓ Des données possédant des relations
- ✓ Relations entre agrégats entre les données de différents ensemble de clés
- ✓ Des opérations impliquant de multiples clés
- ✓ Des besoins de requêtage par les données

# BASES ORIENTÉES DOCUMENTS

## A utiliser pour ...

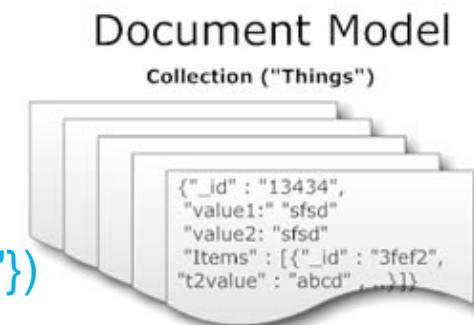
- ✓ Données avec partie structurée et partie non structurée
- ✓ Données de publication variables
- ✓ CMS, Blogging avec commentaires, contenu dynamique, etc ...

## A éviter pour ...

- ✓ Opérations nécessitant consistance sur plusieurs agrégats
- ✓ Des structures d'agrégat très changeante avec des besoins de requêtage forts
- ✓ Inconvénient du schemaless

## Syntaxe d'insertion:

```
db.vehicule.insertOne({marque : "Peugeot",annee :"2005"})
```



# BASES ORIENTÉES COLONNES

## A utiliser pour ...

- ✓ Données avec partie structurée et partie non structurée
- ✓ Evénements des applicatifs (*sharding* possible par application)
- ✓ Données de publication variables
- ✓ Compteurs et analytiques
- ✓ Données avec TTL

## • A éviter pour ...

③ HBase version

HBase version						
Id	Name	Ebay	Google	Facebook	(other columns)	MadBillFans.com
1	Dick	507,018	690,414	723,649	.....	675,230
2	Jane	716,426	643,261	.....	.....	856,767

- ✓ Des besoins de requêtage complexes
- ✓ Des besoins de calcul d'agrégation

## Syntaxe d'insertion:

```
put 'table', 'clé', 'index ligne', 'famille:colonne', 'valeur'
```

Exemple:

```
put 'client', '1', 'profil:prénom', 'Jean'
```

```
put 'client', '1', 'profil:nom', 'Dupont'
```

```
put 'client', '1', 'profil:dossier', '1254587'
```

# BASES ORIENTÉES GRAPHS

## A utiliser pour ...

- ✓ Les moteurs de recommandations :  
« Les autres clients ayant acheté ce produit ont aussi acheté ... »
- ✓ Les données naturellement connectées  
Réseaux sociaux
- ✓ Les services basés sur la localisation ou le calcul d'itinéraires

## A éviter pour ...

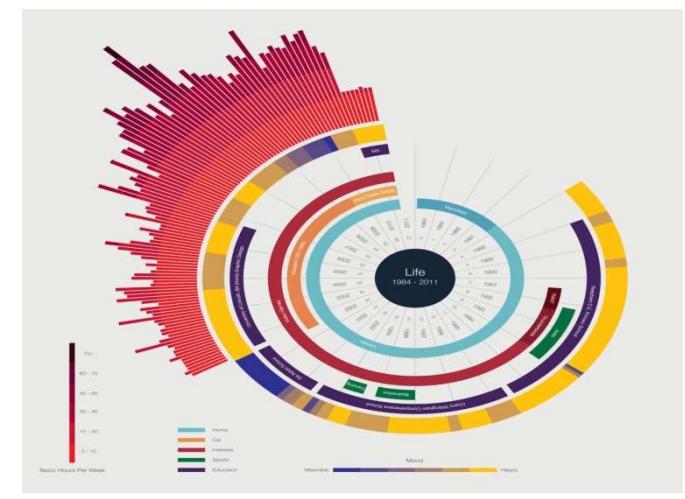
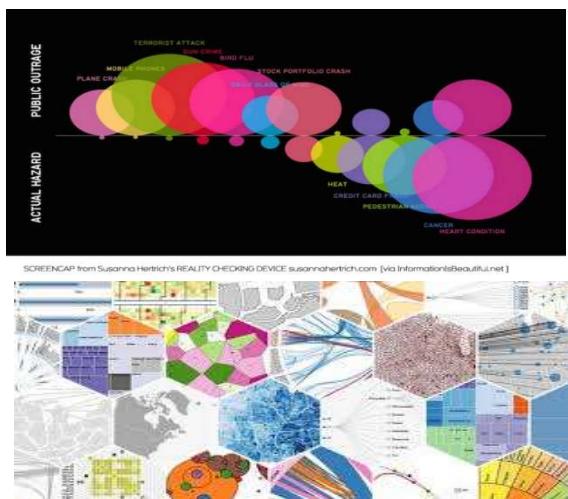
- ✓ Des structures d'agrégat très changeante avec des besoins de requêtage forts
- ✓ Inconvénient du schemaless

# LA « DATAVIZ »

Le terme “Dataviz” désigne la Data Visualization, c'est-à-dire l'ensemble des techniques qui mettent en œuvre la **visualisation de données**

Concrètement, il s'agit de transformer un ensemble de **données** brutes et souvent complexes en une ou plusieurs représentation(s) visuelle(s)

Ces dernières permettant de comprendre en un clin d'œil ce que les données signifient



# HISTOIRE DE LA DATAVIZ



**William Playfair (1759 – 1823)**

Ingénieur et économiste écossais, est l'un des pionniers de la représentation graphique de données statistiques

“En fait de calculs et de proportions,  
le plus sûr moyen de frapper l'esprit est de parler aux yeux.”

William Playfair - 1780

**En 1786, l'inventeur de trois type de représentations graphiques**

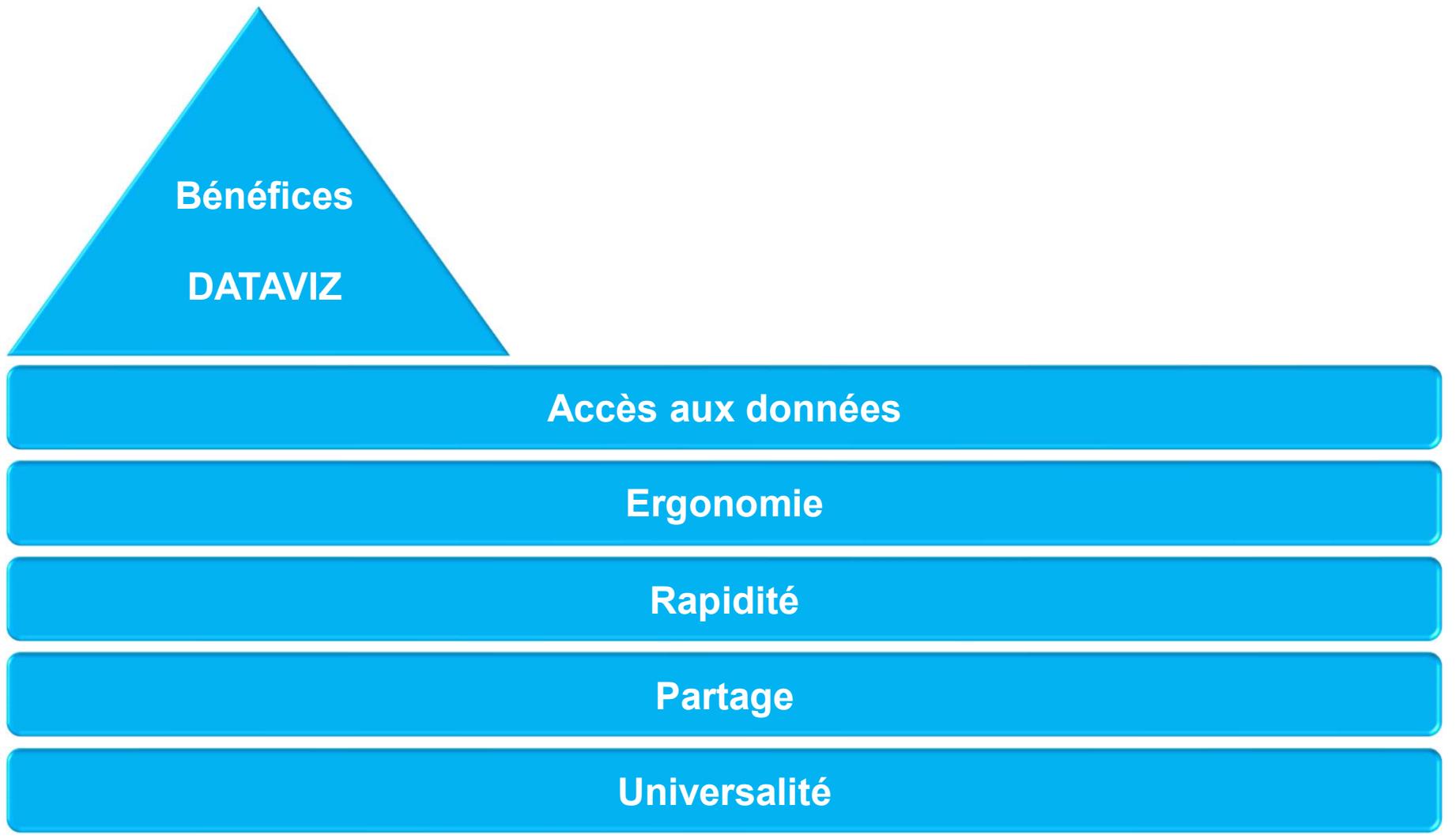
# **DATAVIZ**

## **À QUOI ÇA SERT ?**



# **DATAVIZ**

## **LES BÉNÉFICES**

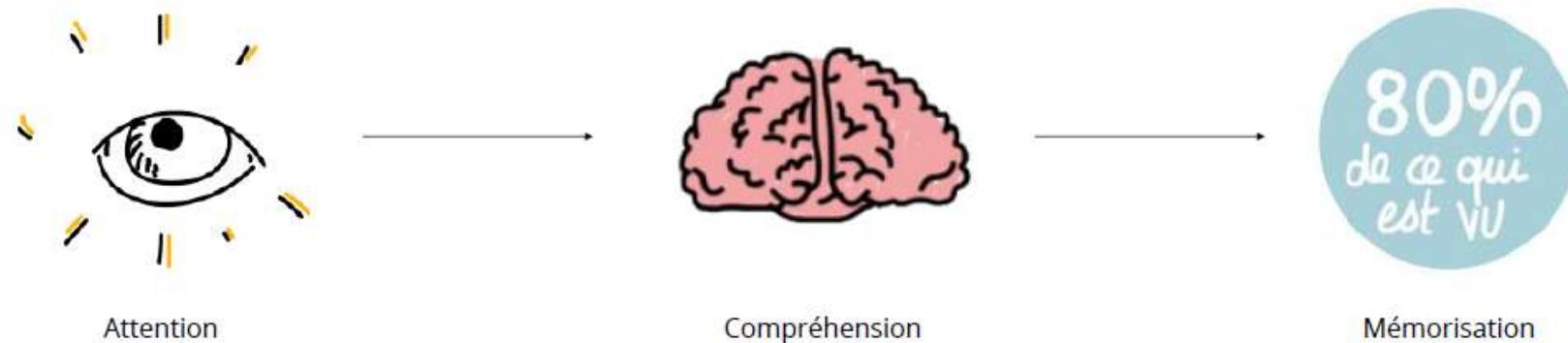


# DATAVIZ PRINCIPES THÉORIQUES



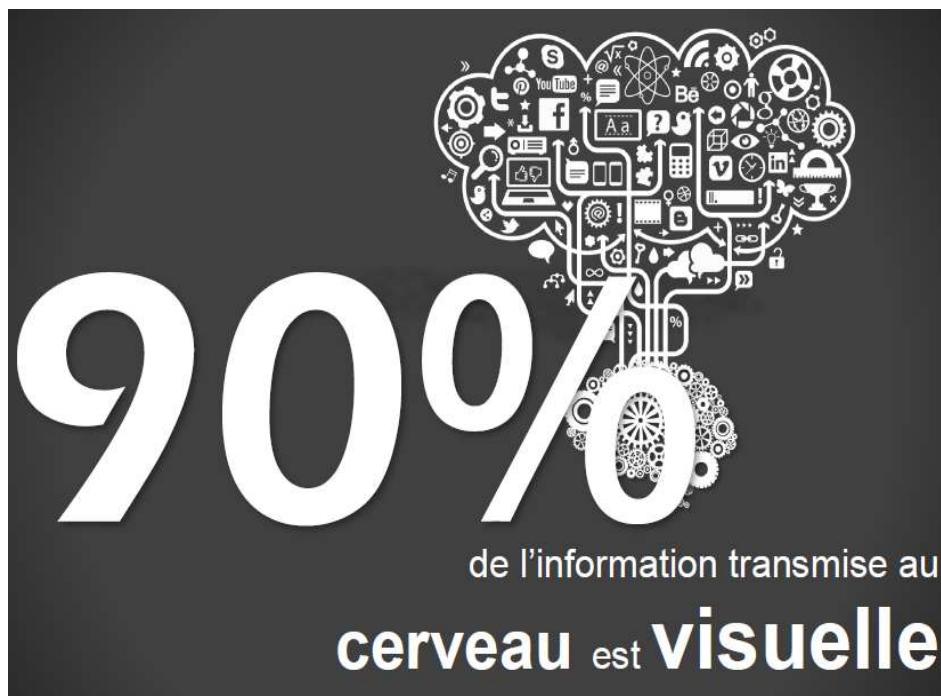
“Une bonne image est celle qui verse **directement l'information** dans nos **yeux** presque sans avoir besoin de **l'analyser**”

David McCandless - 2010



# SACHEZ QUE...

- **90% de l'information arrivant au cerveau est visuelle et les images y sont traitées 60 000 fois plus vite que le texte**
- **50% de notre cerveau est impliqué dans un processus d'analyse visuelle**



A large white graphic of the number "50%" is overlaid on a dark background. Below the percentage, the text "de notre cerveau" is written in white. Below that is a stylized brain composed of gears. At the bottom, the text "est impliqué dans un processus d'analyse visuelle" is written in white.

50%

de notre cerveau

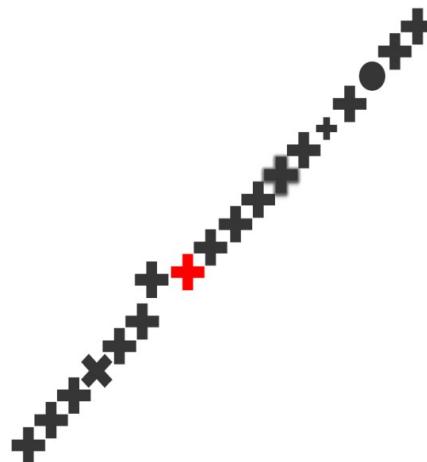
est impliqué dans un processus d'analyse  
visuelle

# TOUJOURS CONCERNANT NOTRE CERVEAU



Le cerveau humain est visuellement connecté

Le cerveau humain est conçu pour « détecter & reconnaître des modèles »



# « UNE IMAGE VAUT MIEUX QUE 1000 MOTS »

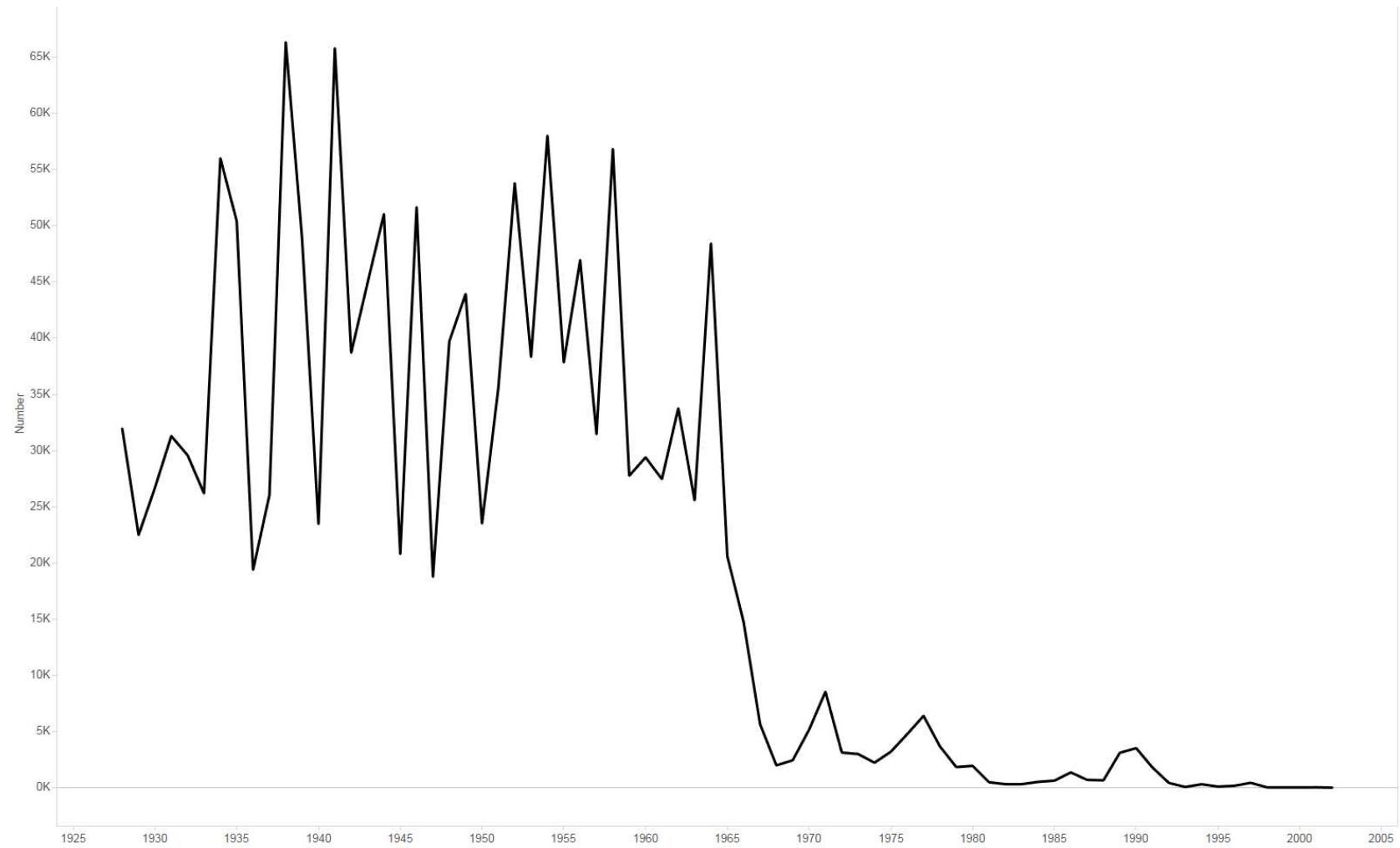
Une image  
vaut mieux que  
**1 000** mots

Confucius (551 av J - 479 av J)

# 1000 MOTS.....

	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	DC	Delaware	Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa	Kansas
1928		205	673	284	184	398	234	43	175	367		11	331	680	86	236
1929		57	199	315	237	620	39	62	89	97		88	2,222	616	171	807
1930		212	196	2,524	993	118	79	28	530	335		99	923	210	776	863
1931		222	74	1,796	669	870	373	265	289	186		29	2,317	1,341	113	131
1932		12	6	717	183	301	26	3	21	95		5	1,428	213	9	549
1933		135	475	1,388	19	328	48	18	97	524		94	953	316	108	407
1934		76	765	1,570	2,112	316	711	269	1,187	2,122		179	2,700	1,525	917	836
1935		63	241	1,714	1,016	1,779	92	125	102	26		82	3,231	628	1,889	1,785
1936		212	15	2,841	44	249	187	203	29	8		133	57	165	14	41
1937		294	64	310	96	799	146	184	55			102	1,299	935	15	53
1938		82	501	1,017	620	207	183	60	1,313	1,576		878	6,933	1,322	392	770
1939		110	236	4,513	536	1,505	357	39	232	232		898	1,947	443	290	342
1940		187	304	2,391	281	525	181	14	190	283		186	737	32	416	639
1941		203	479	607	641	631	370	431	1,337	1,207		44	4,497	1,156	340	1,169
1942		365	365	6,930	336	656	134	50	363	450		156	768	481	402	720
1943		119	239	4,067	879	506	132	331	175	352		402	1,942	688	393	797
1944		557	437	4,371	511	622	238	65	416	777		116	1,331	374	386	863
1945		160	159	2,421	183	284	77	43	217	191		136	687	187	151	419
1946		266	222	3,976	1,684	636	427	66	311	459		227	1,939	1,045	500	1,077
1947		134	383	755	141	1,072	44	4	144	264		32	770	245	1,248	474
1948		556	259	3,854	792	245	194	82	363	147		143	2,649	1,335	907	92
1949		242	1,114	2,230	630	1,570	161	68	326	989		238	408	283	172	1,735
1950		181	246	777	367	578	110	115	225	341	7	209	1,530	610	1,083	179
1951		855	812	3,827	959	387	90	54	186	641	784	190	880	433	326	941
1952		470	344	5,584	473	2,228	405	50	280	587		81	2,195	547	305	661
1953		494	1,536	3,666	830	557	63	46	98	314		121	1,095	657	978	1,336
1954	180	257	235	3,409	193	603	314	170	968	554	37	446	2,170	2,319	986	150
1955	106	721	260	4,194	460	2,189	37	10	115	229	473	44	728	364	670	383
1956	213	531	702	1,916	902	248	130	75	444	362	406	209	2,707	1,390	1,009	905
1957	106	531	193	3,403	189	727	62	19	602	392	456	240	509	501	721	3
1958	106	977	546	2,149	843	997	102	38	1,721	646	115	259	1,845	1,550	2,909	34
1959	178	696	158	2,708	490	863	34	49	362	110	382	126	438	277	1,053	
1960	201	326	188	1,379	580	1,037	86	357	265	29	621	240	1,439	635	252	
1961	232	510	233	2,276	253	476	43	158	729	53	14	166	700	361	641	
1962	148	442	383	1,754	679	1,344	67	224	262	64	180	124	1,074	555	1,076	
1963	197	782	341	1,283	699	432	41	157	468	29	345	339	919	435	824	
1964	165	451	226	2,340	282	351	37	34	707	45	55	125	1,267	2,272	5,251	
1965	16	84	430	1,263	527	844	17	69	341	71	290	165	411	156	832	
1966	89	362	207	997	116	74	50	32	385	57	9	194	1,240	523	385	
1967	18	93	512	285	113	9	4	6	125	9	16	59	104	46	104	30
1968	7	20	1	153	37	59	2	3	34	2	8	4	128	80	36	8

# UNE IMAGE



## **EXPÉRIMENTATION 1/2**

Comptez le nombre de 3 ?

2164597546321546497752316164549334  
2113464973461549764316454797413461  
5349742431975724169436794513649121  
5487887739898831515125394545211222

## EXPÉRIMENTATION 1/2

Et maintenant ?

2164597546321546497752316164549334  
2113464973461549764316454797413461  
5349742431975724169436794513649121  
5487887739898831515125394545211222

# EXPÉRIMENTATION 1/2

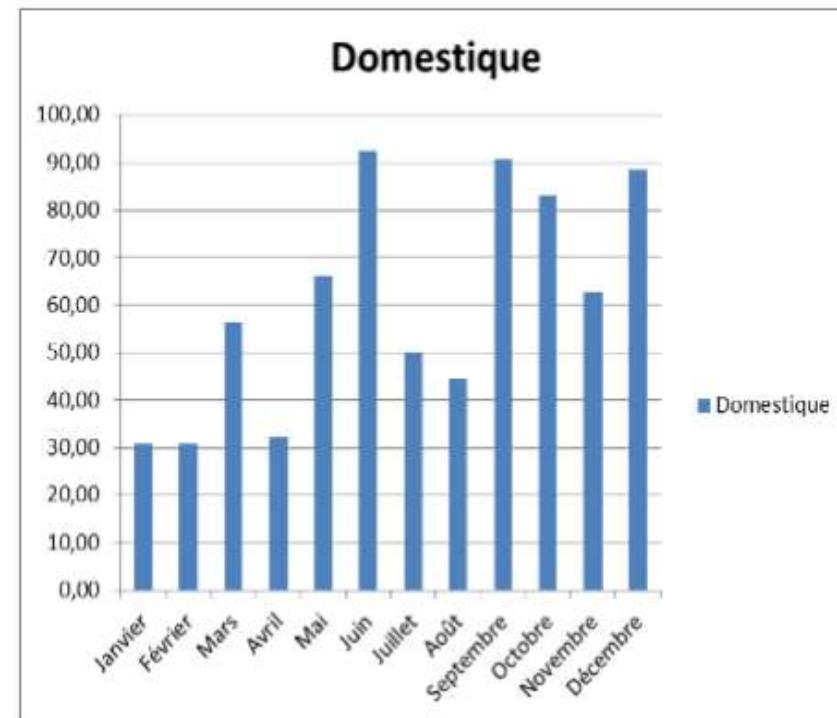
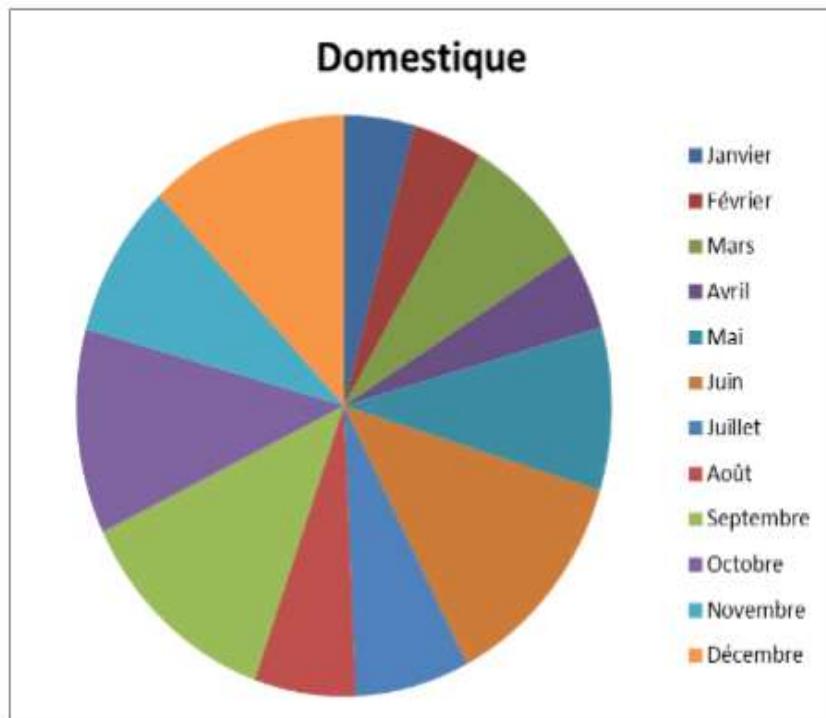
## Tableau VS Carte

	2010	2011	2012	2013	2014
U.K.	30	35	40	25	30
Belgium	10	15	20	20	15
France	35	40	20	20	25
Italy	10	10	15	10	20
Norway	20	25	25	35	45
Spain	5	15	10	15	20
Sweden	20	30	30	45	40
Germany	40	50	40	35	40
Finland	35	40	40	35	45
Danemark	5	5	15	20	20



# QUESTION (TRÈS TRÈS SIMPLE ?)

Quelle est la meilleure représentation?



# **DES PERCEPTIONS IMPORTANTES**

**Les couleurs**

**....mais aussi le contraste, la luminosité, les codes**

**Les formes**

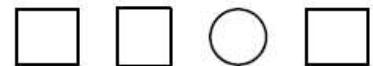
**...mais aussi géométrie, taille, position, profondeur**

**Une troisième perception à ne pas négliger: le contexte, la culture, l'environnement**

# DES PERCEPTIONS IMPORTANTES

Les leviers d'action:

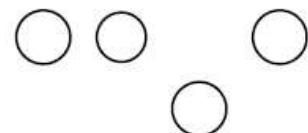
Formes



Couleurs



Position



Taille

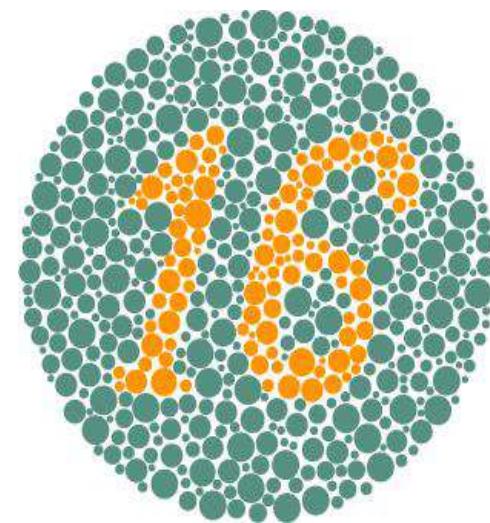
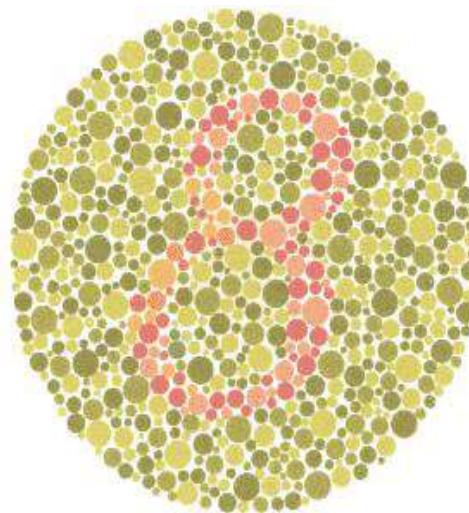
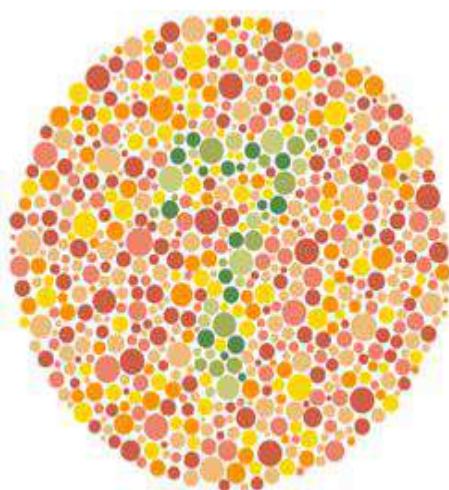


Direction



# PERCEPTION DES COULEURS

Avons-nous la même lecture d'une information ?

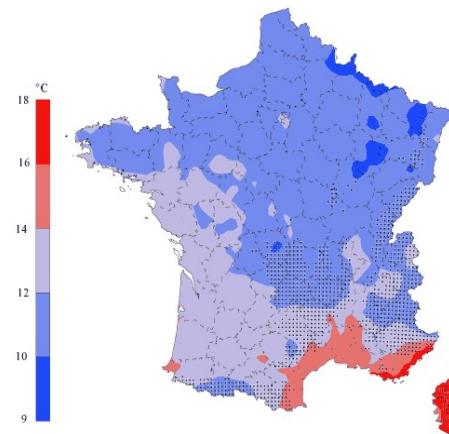


- il y a plus 2 670 000 de daltoniens en France !

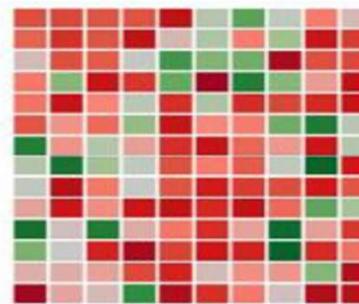
# PERCEPTION DES COULEURS

Ne suis-je pas inconsciemment influencé par ce qu'on m'a toujours appris ?

- Dans votre vie au quotidien



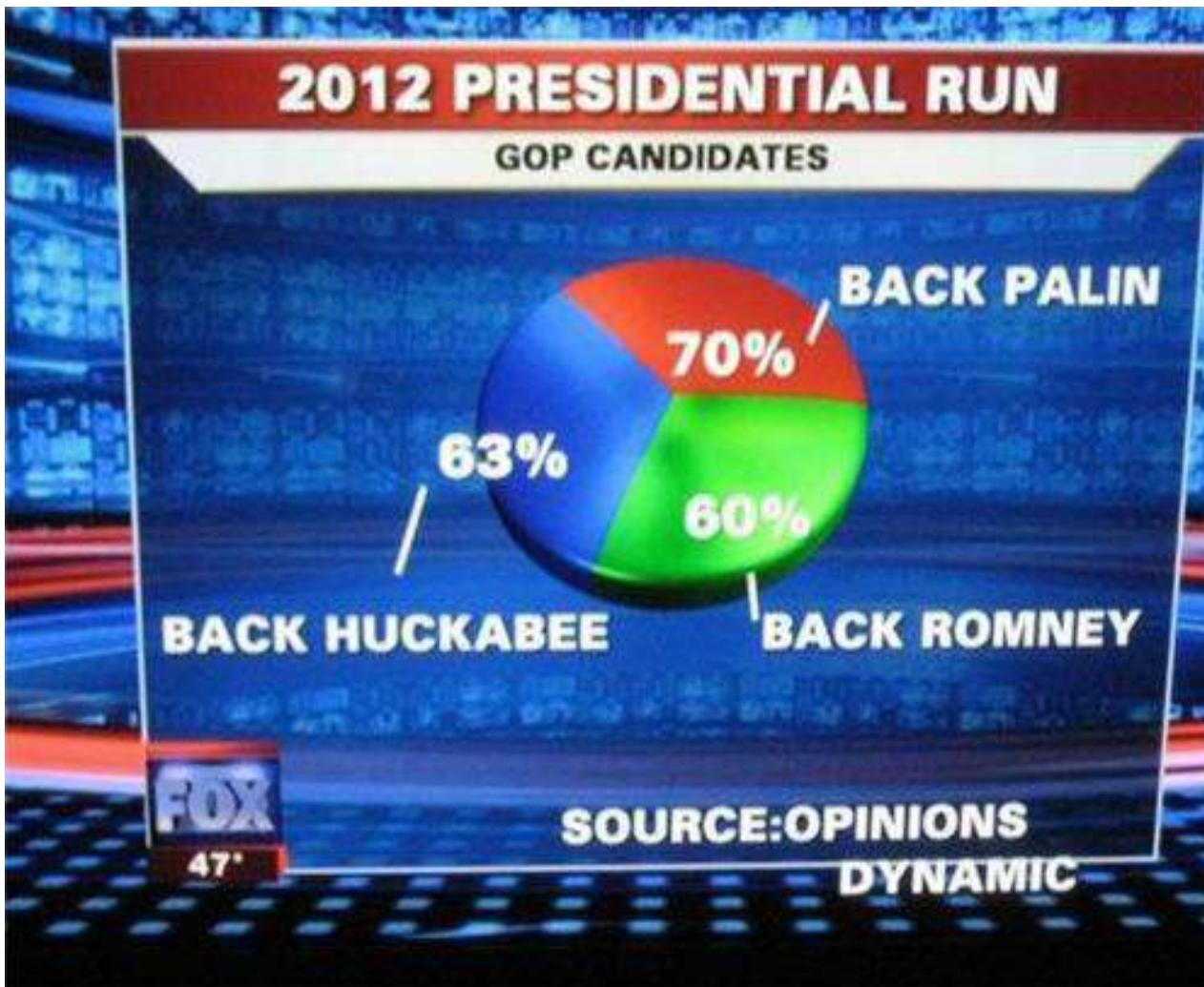
- L'influence du travail



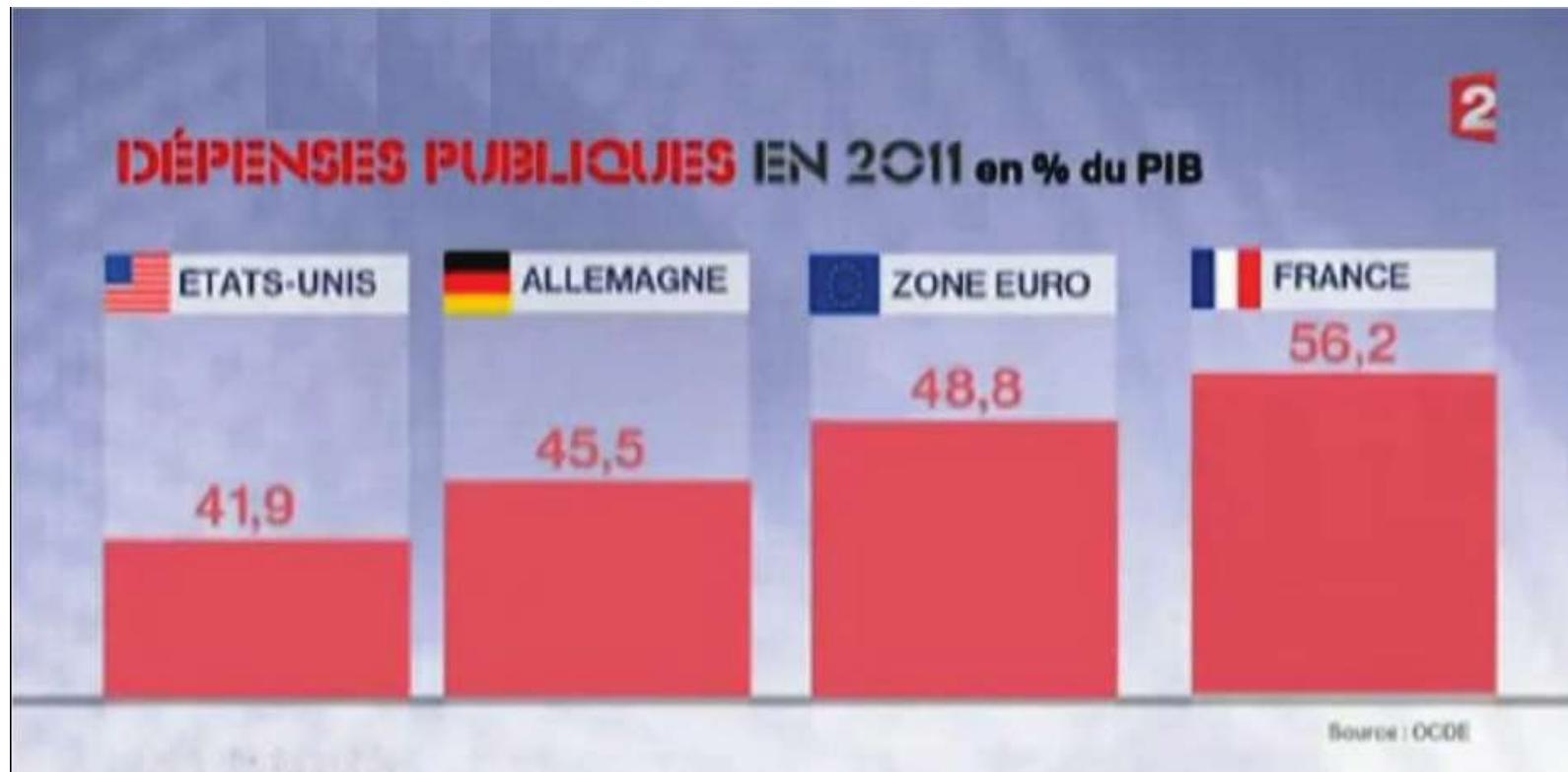
# FAILURES ^^



**WTF !**



# EN FRANCE AUSSI...



# LE CLOUD

236

# LE CLOUD

National Institute of Standards and Technology (NIST):

Le *cloud computing* est l'accès via un réseau de télécommunication, à la demande et en libre-service, à des ressources informatiques partagées configurables. Il s'agit donc *d'une délocalisation de l'infrastructure informatique*.

# **LE CLOUD**

**Le calcul comme un service**

**Il est moins couteux de louer de la capacité de stockage et de calcul que de monter un centre de calcul ?**

**Croissance ou décroissance des ressources à la demande**

**On paye les ressources utilisées**

**La gestion de l'infrastructure est déléguée**

# BIG DATA DANS LE CLOUD

## ■ Catégorie IaaS : Infrastructure as a Service

- Matériel fourni sous forme de VM sur lesquelles ont installé son image disque
- Amazone EC2, GoGRID, Orange

## ■ Catégorie « Platform-as-a-Service »

- l'entreprise cliente maintient les applications proprement dite
- le fournisseur cloud maintient la plate-forme d'exécution de ces applications (carte mère, réseau, stockage,...)
- Windows Azure, Amazone S3, Google Apps

## ■ Catégorie « Software-as-a-Service »

- Les logiciels sont installés sur des serveurs distants (CRM, visioconf, logiciels collaboratifs,...)

## ■ Catégorie « Hadoop-as-a-Service » & « Analytics-as-a-Service »

# CHOIX DU BIG DATA DANS LE CLOUD

1

Certaines DSI ont des contraintes d'hébergement (**place disponible, coût, approvisionnement, compétences**) qui ne leur permettent pas de mettre en place **rapidement** un lot de machines adaptées à Hadoop (commodity hardware, disques JBOD) pour construire un cluster de petite ou moyenne taille.

2

Dans d'autre cas, il s'agit de projets d'innovation pour lesquels **on ne sait pas déterminer** précisément à l'avance **la taille du cluster** dont on aura besoin, ni quel sera son taux d'utilisation.

# Big Data dans le Cloud



# **RETOURS D'EXPÉRIENCE**

**242**

# CARREFOUR ET SON DATALAKE

## Constat

- SI constitué d'une succession de couches applicatives et logiciels avec d'innombrables transferts de données par batch
- Impossible d'avoir une vision temps réel des stocks
  - Batch : caisses => Back office => Supply chain

## Problèmes

- Fiabilité des informations pour le e-commerce
- Cohérence des données au niveau de chacune des applications mais du fait des échanges par batch, incohérente au niveau du SI dans son ensemble

# CARREFOUR ET SON DATALAKE

## Objectif

- Décloisonner, désiloter l'ensemble du SI pour disposer d'une information fiable en temps réel
- Mettre en place un cluster où vont se déverser potentiellement toutes les données générées par les applications du groupe.
- Se servir des technologies Big Data pour améliorer les processus opérationnels

## Les 3 axes du projets

- Le socle technique
- Les cas d'usage
- La gouvernance des données

# CARREFOUR ET SON DATALAKE

## Le socle

- Association de 25 briques open sources:
  - Hadoop distribution Cloudera (HDFS, Hive, ...)
  - Base NoSQL Cassandra
  - ElasticSearch, Kafka, R
- Fortes contraintes sur le Datalake => Toutes les données de l'entreprise vont être amenées à y être intégrées (tickets de caisse, stocks, commandes,..., logs,...)
- Il y a eu une réflexion globale sur le stockage des données
- Mise en place d'une architecture Lambda afin de garantir le niveau de performance attendu

# CARREFOUR ET SON DATALAKE

## Les cas d'usage

- Dissocier le socle des usages est important
- « On a souvent tendance à aborder le Big Data via les usages. C'est l'élément important, mais si l'on veut construire quelque chose qui est industrialisable, il faut absolument travailler sur cette notion de socle. C'est la partie cachée de l'iceberg mais elle doit être capable, au fil du temps, de résister à tous les cas d'usage que l'on veut bâtir dans les années à venir. » Jean-Christophe Brun (directeur du centre de solution BI et Big Data de Carrefour)
- Identifier les usages transactionnels et décisionnels

# CARREFOUR ET SON DATALAKE

## La gouvernance de la donnée

- Mis en place d'une organisation devant assurer la maîtrise des données devant alimenter le Data Lake
- Nomination de Data Owners
- Mise en place d'outils de traçabilité pour garantir la qualité de la donnée dans la durée.

# CARREFOUR ET SON DATALAKE

## Migration progressive

- Dès qu'une application est interfacée avec le Data Lake, ses données deviennent accessibles à l'ensemble des nouvelles applications sous forme de Web Services
- Data Lake = BDD de l'application
- Exemples:
  - Nouvelle application SAV : Développement de l'IHM et des processus applicatif; Les données sont stockées dans le DataLake

# CARREFOUR ET SON DATALAKE

## Impact organisationnel

- Évolution des méthodes de travaille et des interactions
  - Métiers ↔ équipe projet ↔ exploitant
- **Agilité**, adaptation nécessaire au changement (évolution régulière des logiciels Big Data,...)
- Les changements doivent s'opérer de la manière la plus fluide jusqu'à la mise en production ce qui implique d'aller vers une approche type **DevOps**