

Master 1

FRAMEWORK LOGICIEL POUR LE BIG DATA

COURS 1

R.JAZIRI

PLAN DU COURS

- Big data, les fondamentaux
- Schémas organisationnels et gouvernance des données (ETL, Base de données, ...),
- Introduction aux fondements technologiques (Hadoop, cloud, H2O)
- Stratégies de valorisation des données en entreprise (typologie générale, focus sectoriels, études de cas concrètes...)

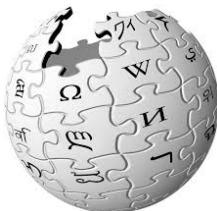
BIG DATA, LES FONDAMENTAUX

- De l'informatique décisionnelle au Big data
- Un nouveau contexte des données
 - Internet des objets
 - Open data
- Le big data
- Des exemples
- De nouveaux métiers

INFORMATIQUE DECISIONNELLE

Définition

WIKIPEDIA



L'**informatique décisionnelle** est l'informatique à l'usage des décideurs et des dirigeants d'entreprises.

Elle désigne les moyens, les outils et les méthodes qui permettent de **collecter, consolider, modéliser et restituer** les données, d'une entreprise en vue d'offrir une **aide à la décision** et de permettre à un décideur d'avoir une vue d'ensemble de l'activité traitée.

L'informatique décisionnelle peut se décliner selon trois axes :

- ▶ **Business Intelligence** = Visualisation et exploration des données
- ▶ **Business Analytics** = Statistique & Datamining
- ▶ **Big Data & Data Science** = Evolution technologique et méthodologique

L'EVOLUTION DE L'INFORMATIQUE DECISIONNELLE...

L'opportunité du BIG DATA et l'évolution vers la DATA SCIENCE



La poussée du Data déluge...

SIO

Data WEB
d'entreprise

Data WEB externes :
Réseaux sociaux, objets
connectés, open data...



Les opportunités
technologiques...

ARCHITECTURES
DISTRIBUEES



IN MEMORY, GRID,
IN DATA BASE, SSD...

ption Décisionnel

Big Data

Modernisation
des technologies,
des outils,
des méthodes...



- ▶ Data Déluge & Web Crawling
- ▶ Architectures distribuées
- ▶ In memory & In database
- ▶ Dataviz
- ▶ Déstructuration & faible densité
- ▶ Text mining & variété
- ▶ Open source
- ▶ Machine learning...

Data Science & Big Data Analytics

COMBIEN D'INFORMATION ?

- **On pourra bientôt tout indexer et enregistrer**
- **La majeure partie de cette information sera inconnue des humains**
- → **Synthétiser les données, repérer des tendances détecter des anomalies sont donc des techniques essentielles**



D'après Jim Gray et Gordon Bell
Microsoft Research

Data déluge

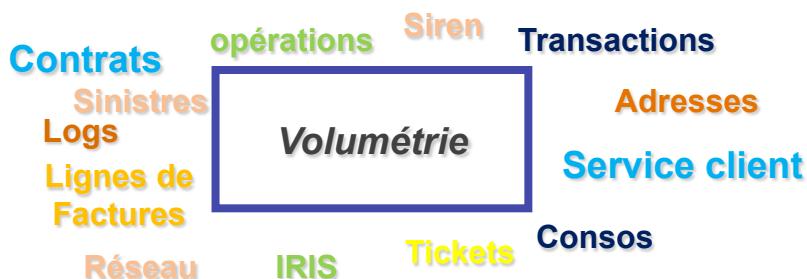
80% des data ont été produites
lors des 2 dernières années...

Elles adressent un nombre croissant
de domaines totalement nouveaux
dont la culture d'usage
ne dispose pas de recul

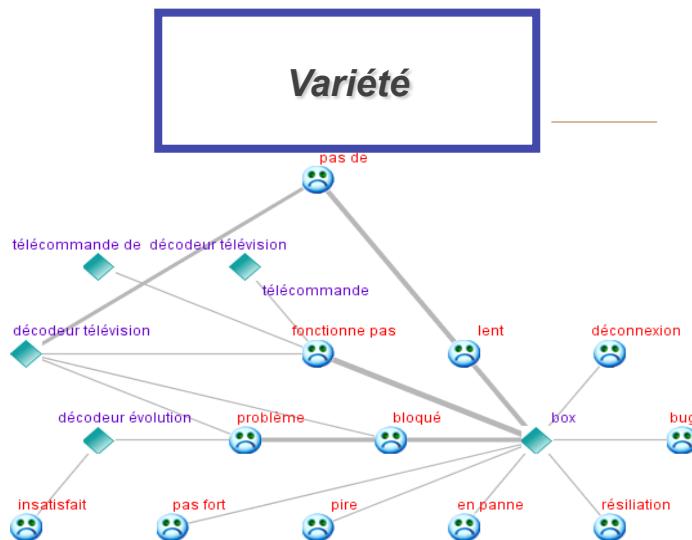
Data déluge

Le site de Planetscope <http://www.planetscope.com> estime à 3 millions le nombre d'e-mails envoyés dans le monde chaque seconde, soit plus de 200 milliards par jour en comptant les spams qui représentent 90% des flux, pour une population d'internautes de 2,5 milliards d'individus.

On prévoit 10 Zo (zettaoctets), soit 10400 milliards de gigaoctets de données déversés tous les mois sur Internet.



Exemples de domaines de données



Carte de sentiments réalisés sur des données twitter relatives à un contact auprès du SAV d'un opérateur télécom

SAM - Windows Internet Explorer
http://firweb01.dvn10/SEM/WebApp/FormuleProche.aspx?FrmCd=1003481%20&FmRefCd=1067594%20&Ld=714123%2014A+41340/690039
SAM

Analog Microbio

Demande : Recherche de formules proches pour la formule 1003481 A - Evaluation du 10/02/2011 (V2.2011-6)

Système d'Apprentissage Microbiologie

Formules proches	Distance	Date validation	Provenance	Stade test	pH réel	Résultats Observés	Prédictions SAM
						A B C D	A B C D
<input checked="" type="checkbox"/> 1003481_A	0 (100)	10/02/2011	ND	ND		■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 1003481	0.18 (100)	04/09/2007	LABO	T2-CONF	5.4	■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 1003463_A	4.42 (69)	21/06/2007	LABO	T2-CONF	5.6	■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 1003470	8.14 (57)	17/07/2007	LABO	T2-CONF	5.6	■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 1003471	8.14 (57)	17/07/2007	LABO	T2-CONF	5.6	■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 1064256	43.63 (38)	20/08/2008	LABO			■ ■ ■ ■	■ ■ ■ ■
<input type="checkbox"/> 714123_14A (ref.)	47.51 (38)	20/08/2008				■ ■ ■ ■	■ ■ ■ ■

Vitesse (performance + réactivité business)

Légende : Proximité de Confiance : 1 = Peu Proche 5 = Très Proche

Comparaison Fermer

Application de test et de recherche de formulation microbiologique utilisé par des centaines d'utilisateur délivré via un web service interrogeant un serveur de calcul statistique

Les 3 V et+ du Big Data

Vérité

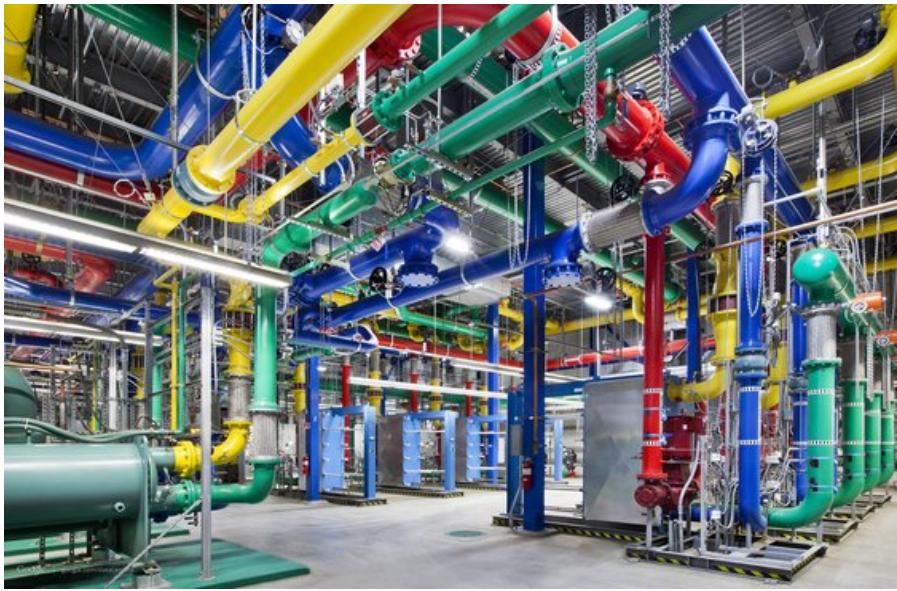
Visualisation

Valorisation

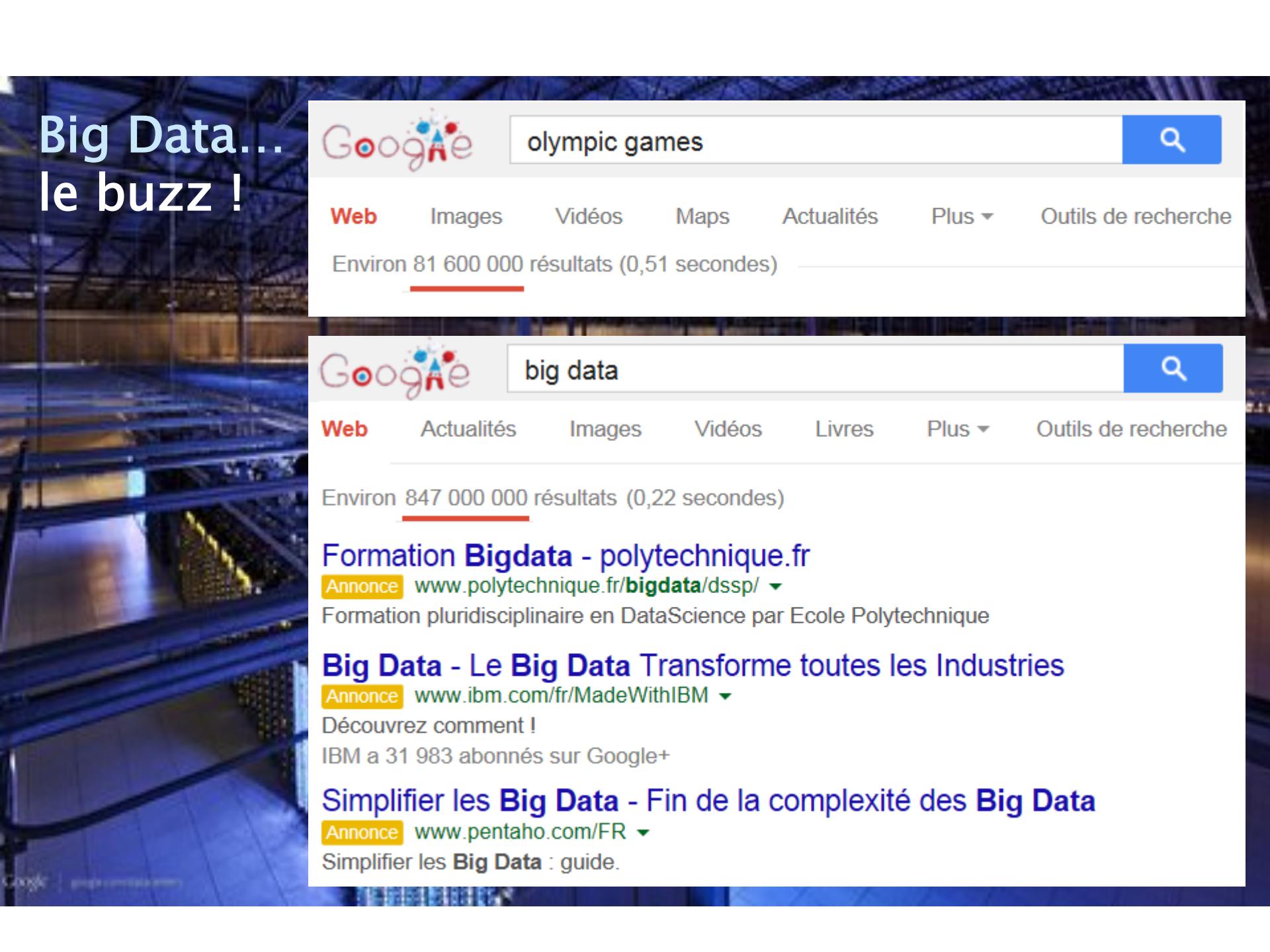
Qualité et Gouvernance des données :
DATA PROFILING
DATA CLEANING et RULING
DATA MONITORING

Intelligence des données :
Injecter dans les processus BI et analytics, des critères robustes, explicites et pertinents

Les pionniers du Big Data



Big Data... le buzz !

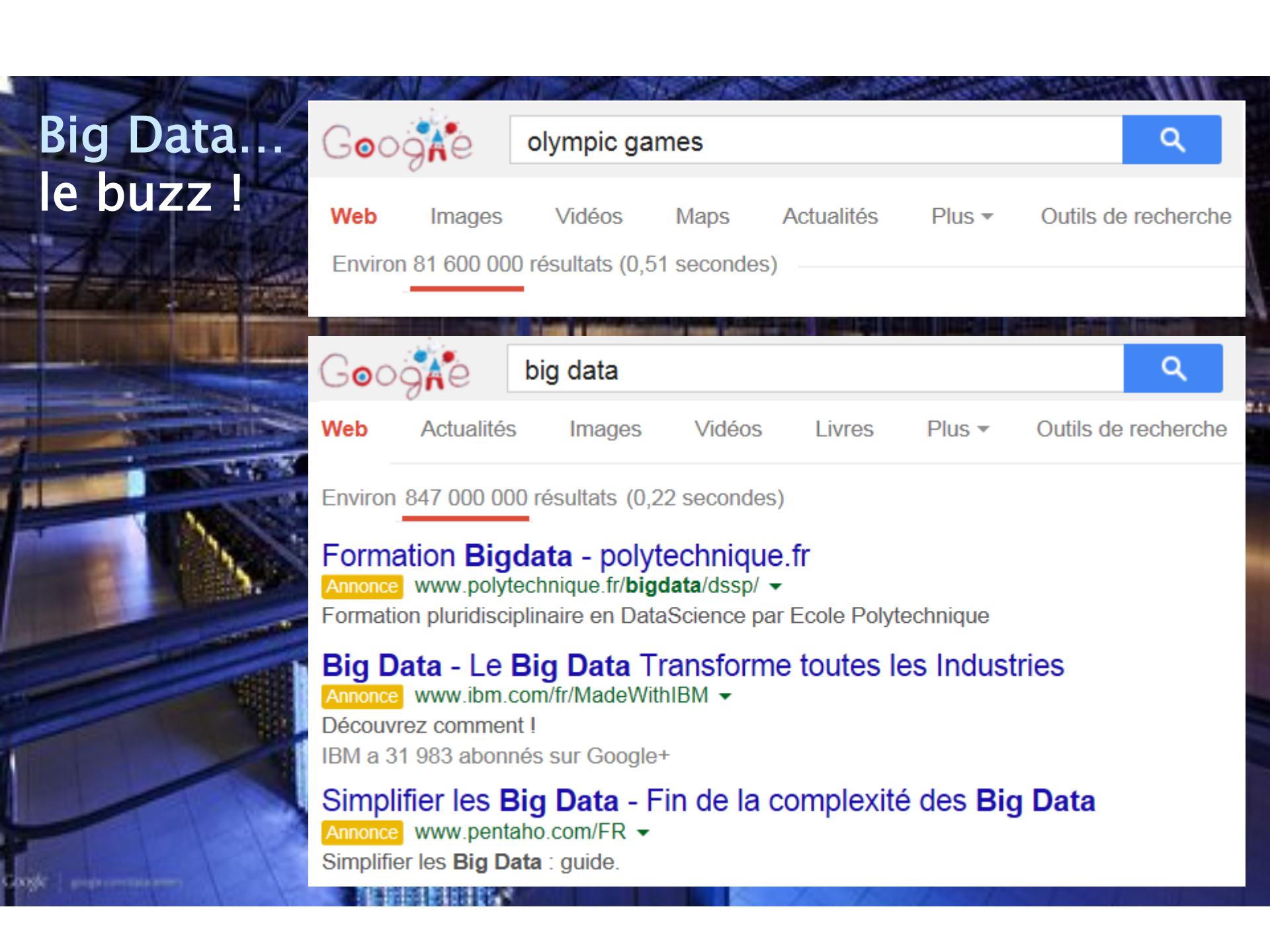


Google

olympic games

Web Images Vidéos Maps Actualités Plus Outils de recherche

Environ 81 600 000 résultats (0,51 secondes)



Google

big data

Web Actualités Images Vidéos Livres Plus Outils de recherche

Environ 847 000 000 résultats (0,22 secondes)

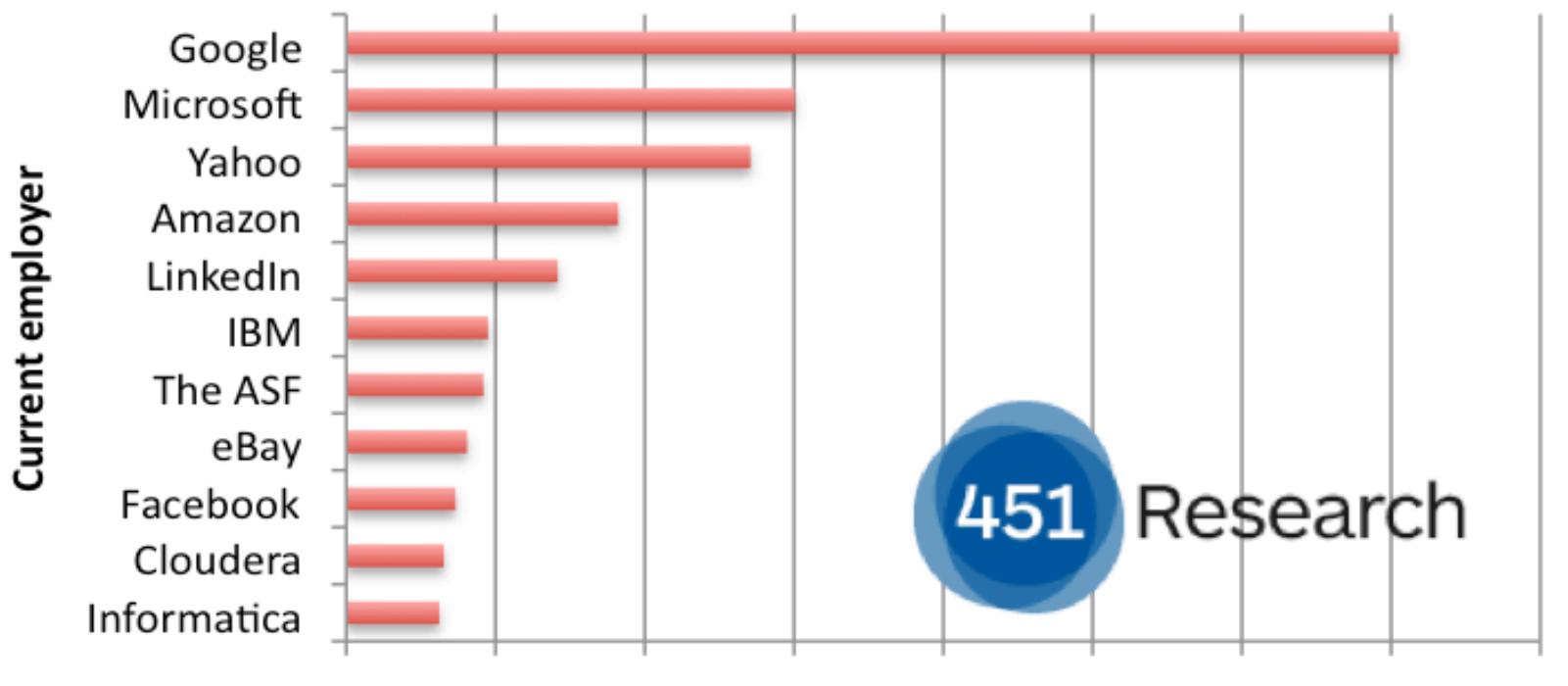
Formation Bigdata - polytechnique.fr
Annonce www.polytechnique.fr/bigdata/dssp/ ▾
Formation pluridisciplinaire en DataScience par Ecole Polytechnique

Big Data - Le Big Data Transforme toutes les Industries
Annonce www.ibm.com/fr/MadeWithIBM ▾
Découvrez comment !
IBM a 31 983 abonnés sur Google+

Simplifier les Big Data - Fin de la complexité des Big Data
Annonce www.pentaho.com/FR ▾
Simplifier les Big Data : guide.

Les pionniers du Big Data... en 1^{ère} ligne pour le buzz...

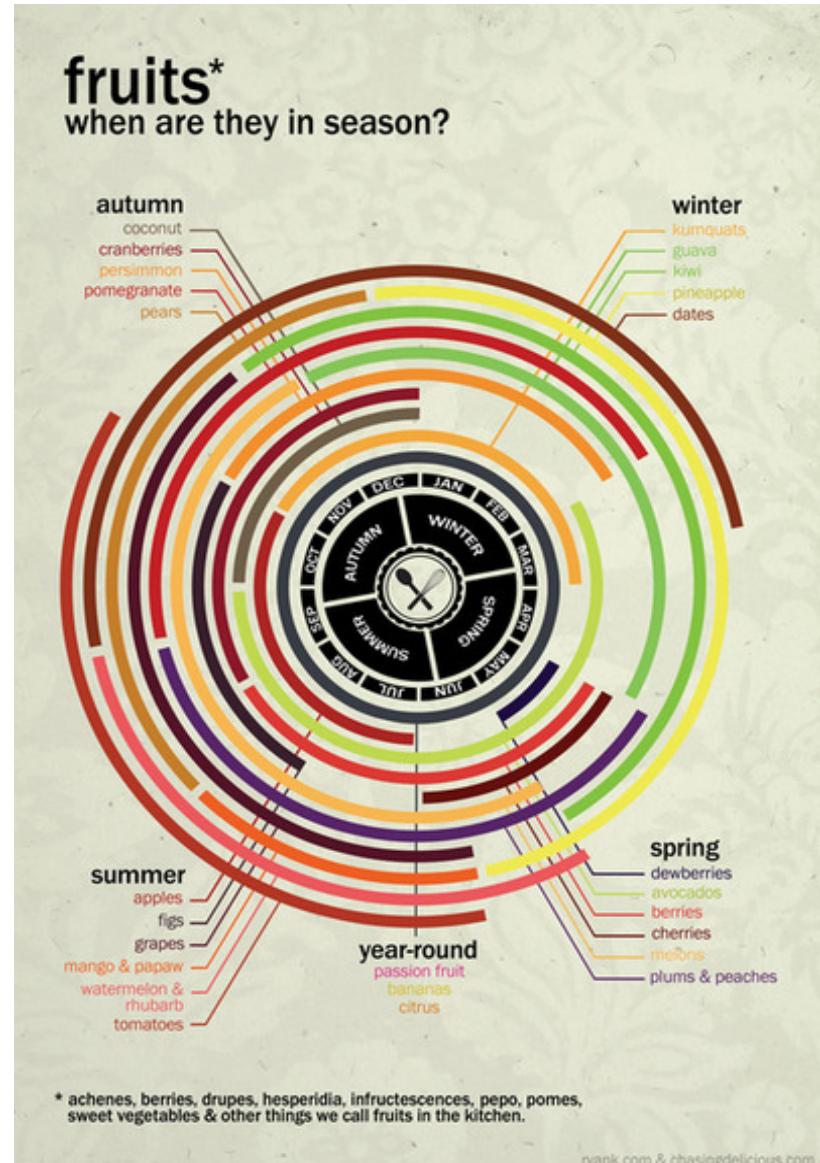
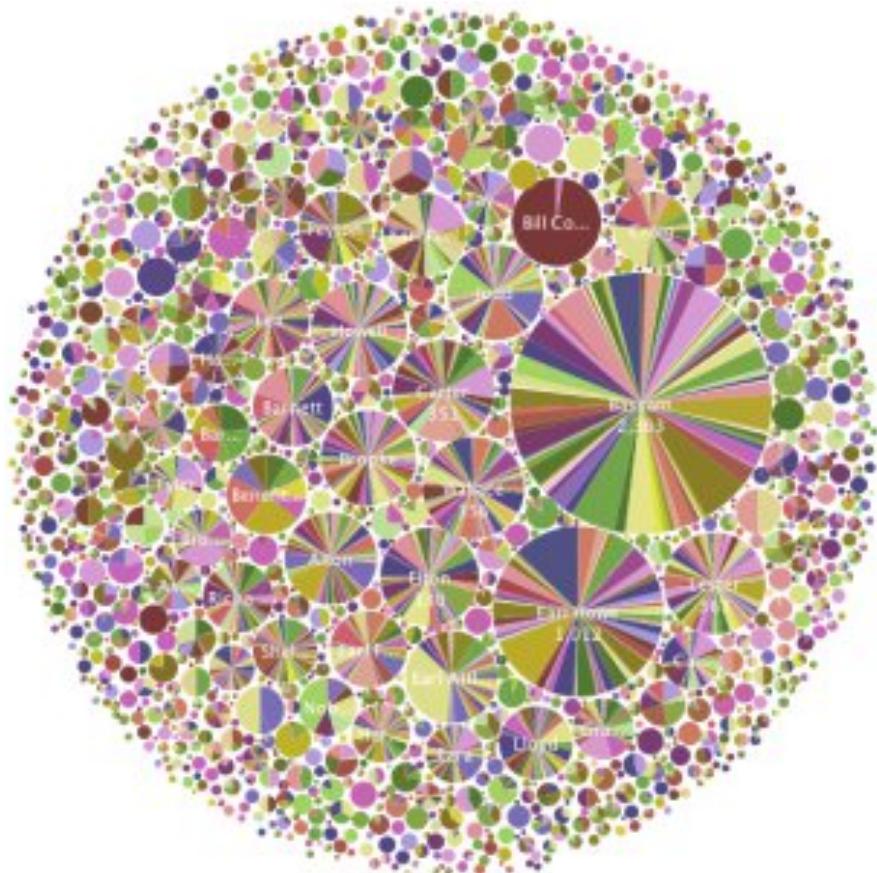
Biggest employers of MapReduce skills



20% détenus
par le top 10
< 1 million de sal.

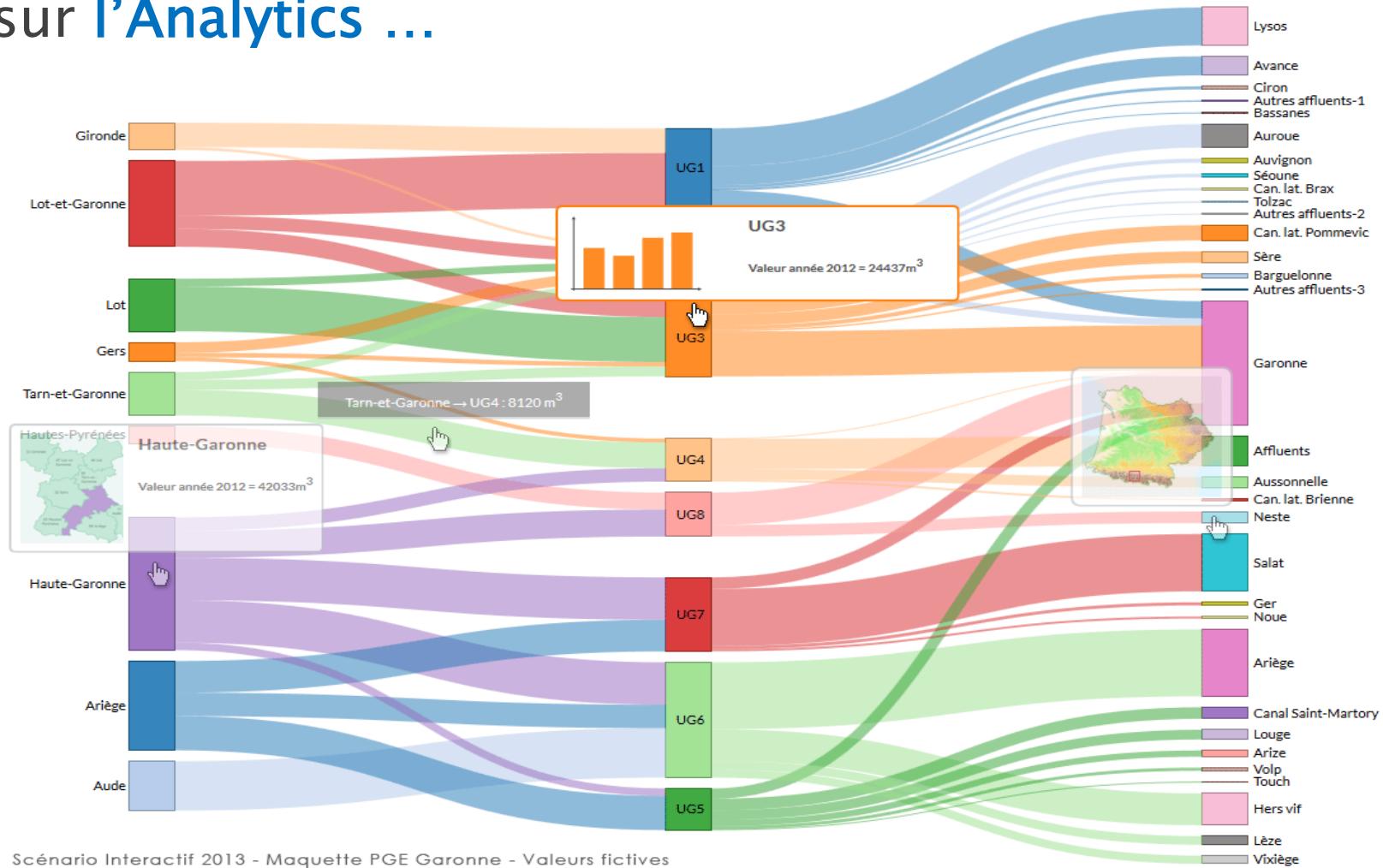
% of total LinkedIn profiles mentioning MapReduce

Le Big Data fait main basse sur la DataViz et sur l'Analytics



Le Big Data fait main basse sur la DataViz et sur l'Analytics ...

Phénomène de mode... ou révolution culturelle ?



Pas assez **BIG**...

Quelle taille faut-il pour être **Big** ?

Wikipédia :

« difficulté à travailler les données avec les moyens classiques »

Donc le caractère **Big** est relatif au contexte et à l'usage.

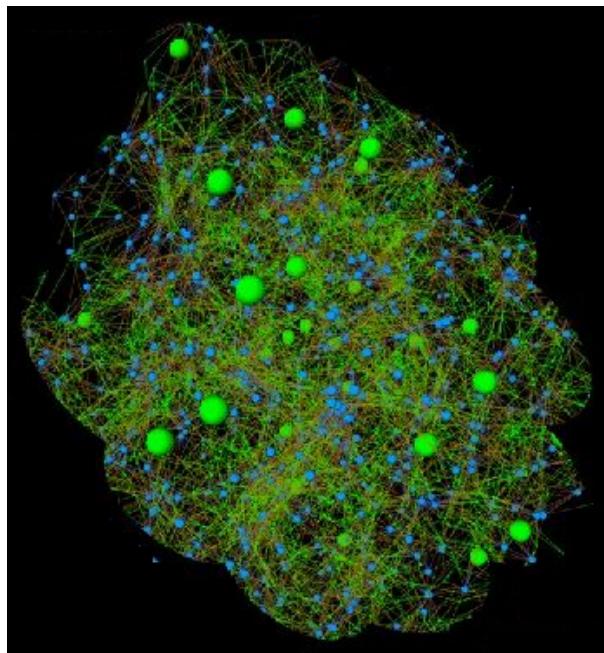
Dès que le besoin se fait sentir de dépasser les « moyens classiques »... le projet peut devenir **Big**

Le principe est déjà intégré dans le concept de
« Not So Big Data »

La **BIG** complexité

La VOLUMETRIE est un indicateur bien trop simpliste pour caractériser le **Big...**

De très gros volumes peuvent ne pas poser de problème à l'usage (**ex** : accès directs à des données indexées...).



Des volumes non nécessairement gigantesques peuvent engendrer une complexité algorithmique colossale...

Exemple : Fichier de 20 millions de clients à enrichir par des données de 10 000 panélistes, **par calcul de la moyenne des 100 panélistes les plus proche de chaque client...**

Pour affronter cet univers gigantesque,
complexe et
encore inconnu...

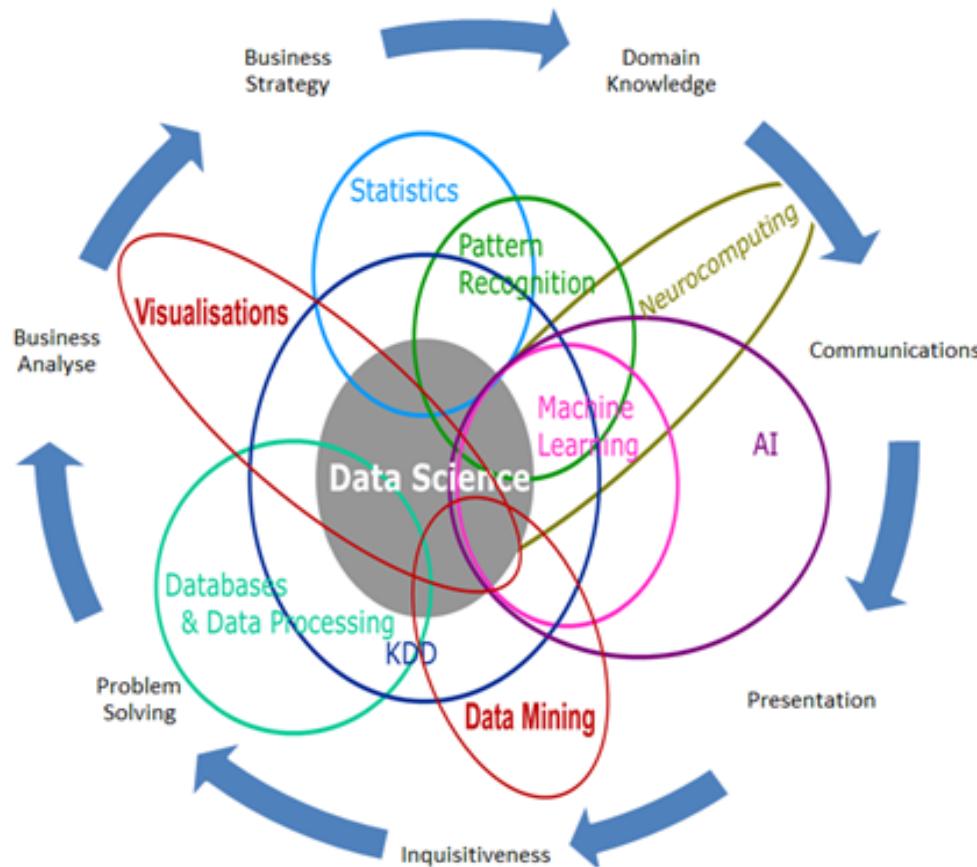
... il fallait de nouveaux « héros »
et une nouvelle discipline :

Les **data scientists**
Et la **data science**

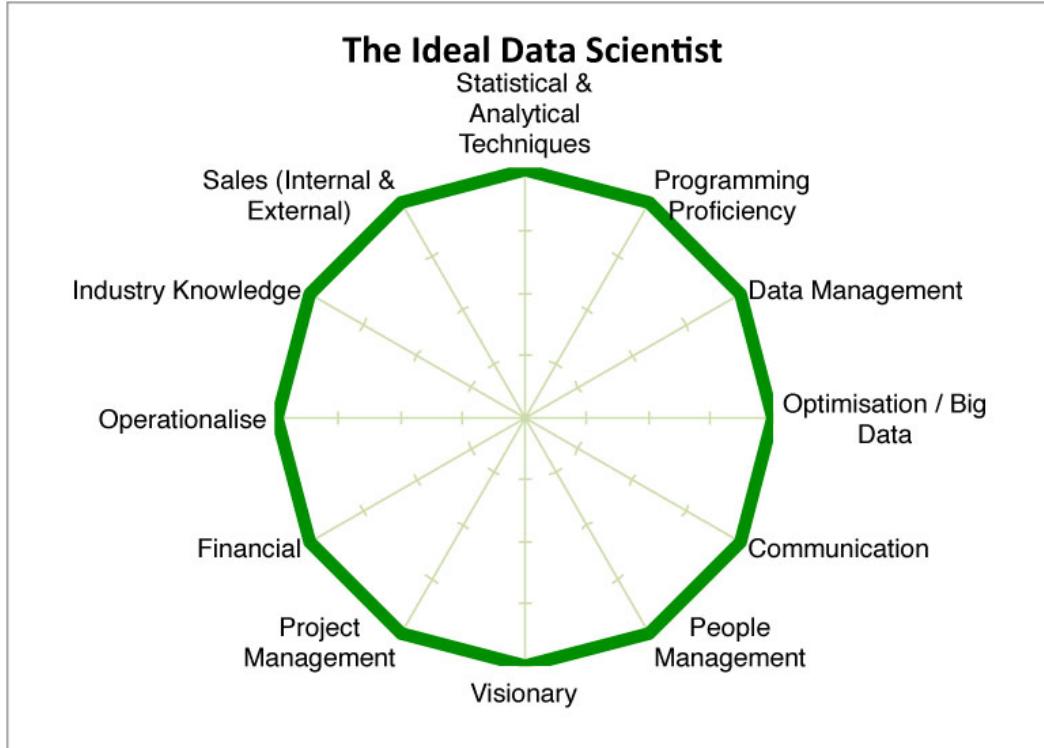


... La **data science**

Data Science Is Multidisciplinary



... Les data scientists



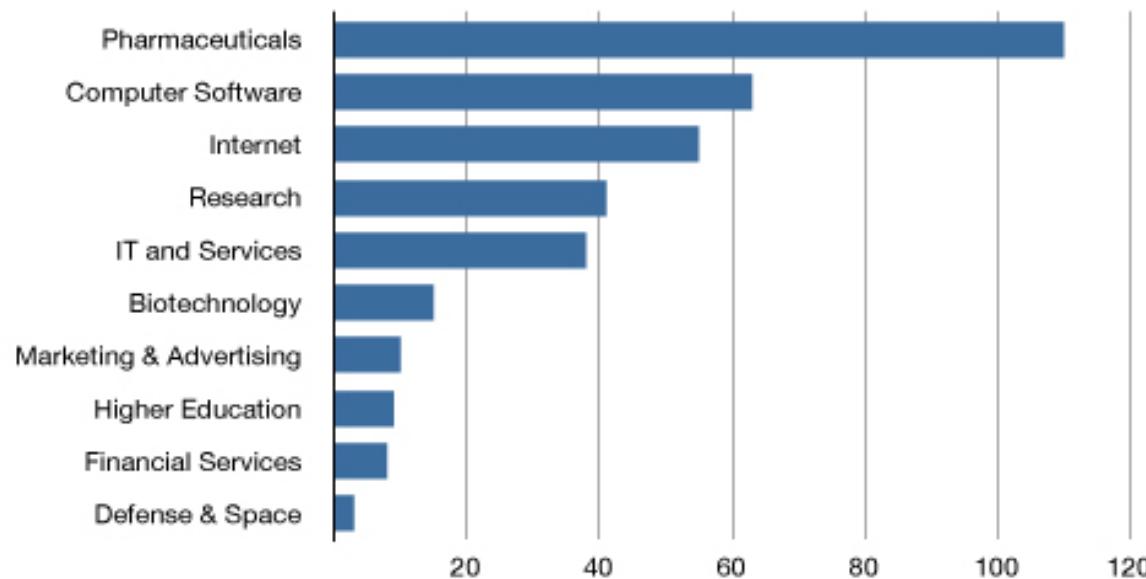
Profile of a Data Scientist



... savent tout faire.

... Les data scientists

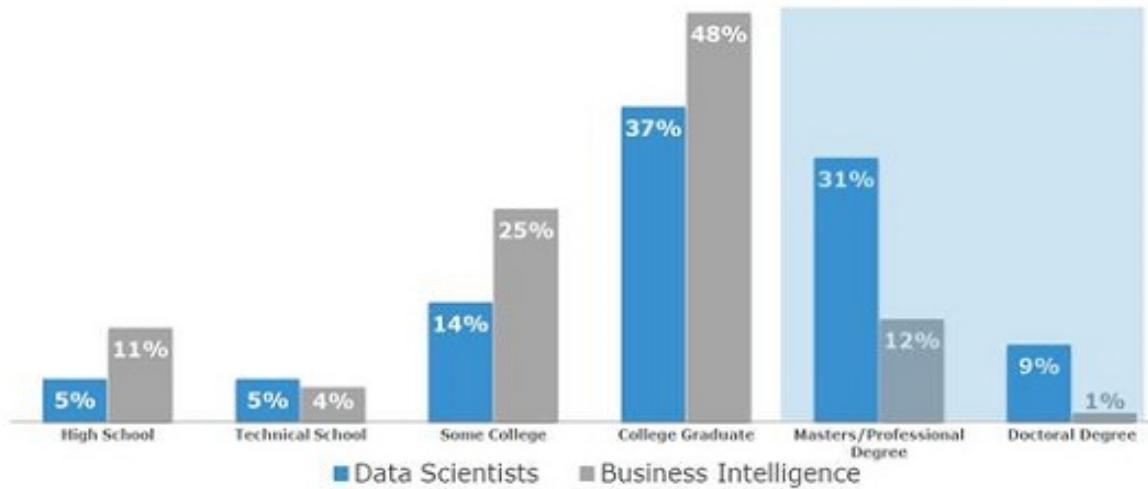
Distribution by Industry of “Data Scientists” in the U.S.



... ne font pas que du business

... Les data scientists

Data science requires greater education



40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

EMC²

... ont fait des études.



Exploiter la mutation progressive des technologies et méthodes...

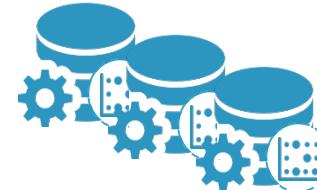
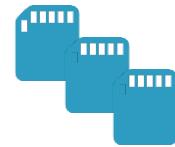
... pour mieux répondre aux questionnements métiers

ETL +
Restitutions BI
&
Analytiques

In Database
In Memory

Distribué

Dataviz in memory & in hadoop
Analytics in memory & in hadoop
Solutions Intégrées et dynamisme de l'open source
Accès au cloud (SAAS...)



BI &
Analytics

Évolution...

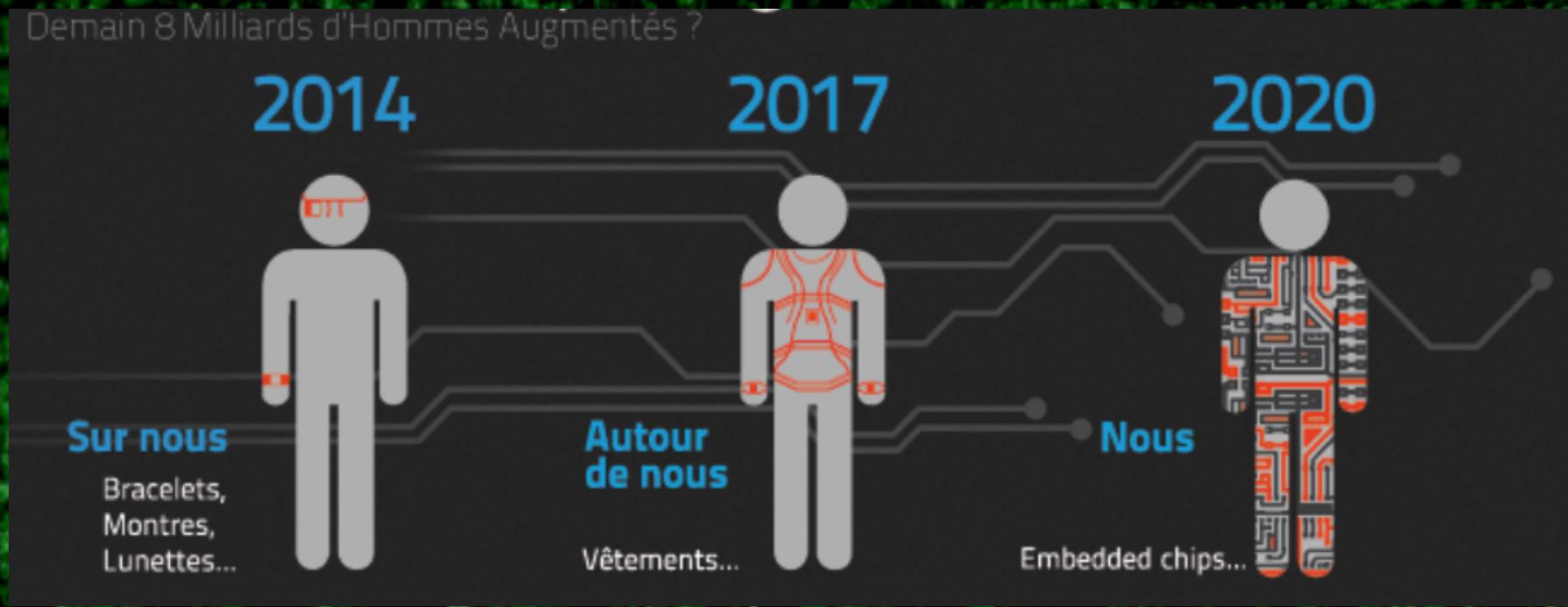
Évolution...

... Les cas d'usages

- Déetecter des comportements frauduleux lors de transactions financières en ligne, (**Carte bleu**)
 - Estimer un taux de transformation sur un site marchand en fonction du nombre de clics sur certaines pages, (**Twenga**)
 - Prédire les risque de non solvabilité d'un client en fonction de des ressources et de son profils sociprofessionnel, (**Secteur Bancaire**)
 - Anticiper les intensions de résiliation d'un service en fonction des activités d'un souscripteur, (**Orange**)
 - Découvrir les préférences d'un client que l'on souhaite retenir pour lui suggérer des produits et des services adaptés à ses gouts et ses besoins, ...
- Ces préoccupations ne sont pas nouvelles, c'est le quotidien des RSSI, actuaires et marqueteurs

Les industries fournissent des machines équipées de nombreux capteurs et échangent entre elles des informations via le « CLOUD ».

Les vêtements connectés, attendus à courte échéance, seront aussi des producteurs d'informations. Certaines machines à café, sont déjà équipées de moteurs d'analyses prédictives.

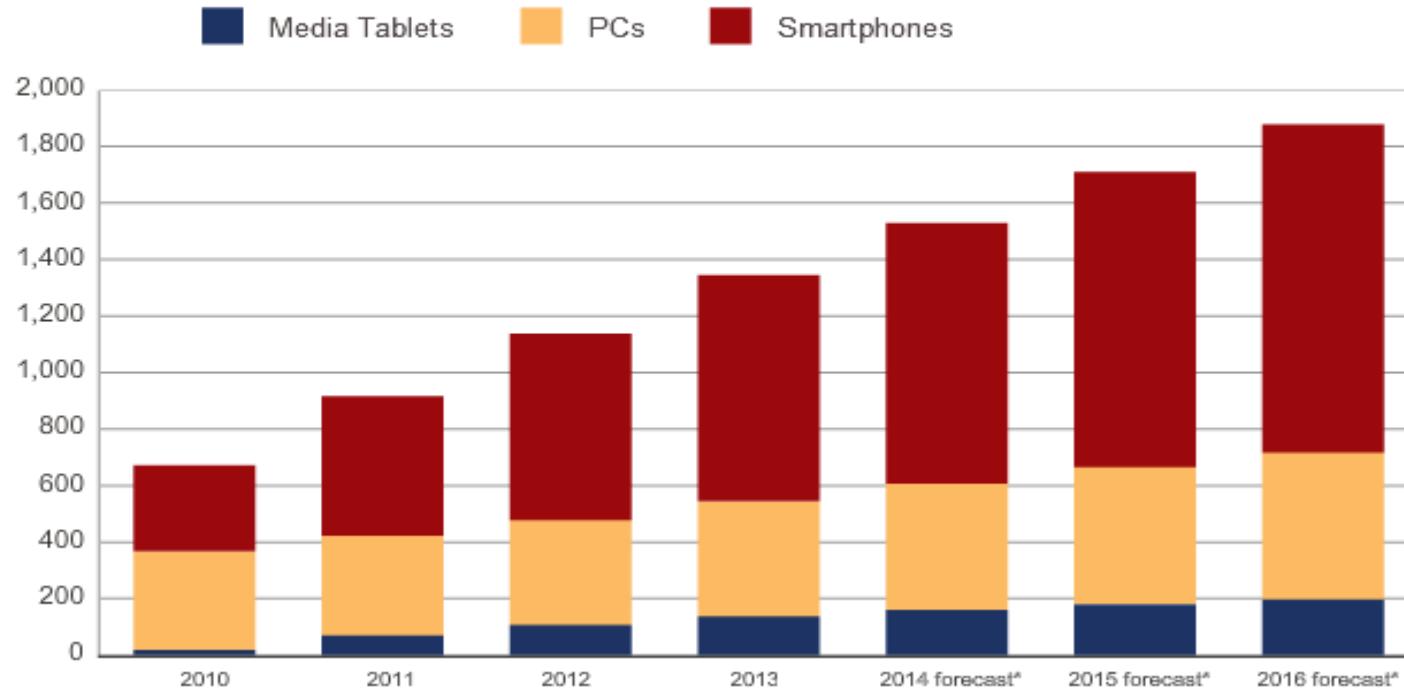


Motivations du Big Data

Evolution du nombre
de machines connectées...

... 1 à 2 milliards de
machines dans le monde

**Worldwide Smart Connected Device Shipments,
2010-2016 (Unit Millions)**



Les villes intelligentes comptent les voitures,
adaptent les feux tricolores,
gèrent l'éclairage urbain et des caméras de surveillance,
Supervisent la bonne marche et le trafic de leurs réseaux...

... elles sont donc des productrices de données massives.



L'omniprésence des données : Linky

Le nouveau compteur intelligent Linky

- Un projet Smart Grid : Linky est l'objet connecté qui sera déployé dans 35 millions de foyers français d'ici 2020.
- Un suivi de la consommation toutes les 10 minutes : *Big Data par excellence*
- C'est aussi pour prévoir le pics de consommation, éviter le lancement de centrales électriques, détecter les fraudes



L'omniprésence des données : Linky

- Linky peut-il aussi savoir quel appareil ménager j'utilise ? combien de téléviseurs je possède ? A quelle heure je prends ma douche ? mes habitudes et donc mon profil de consommateur : **Une publicité adaptée**



L'omniprésence des données : Freebox Révolution

La Freebox Révolution

- Peu de données à la signature d'un contrat



- Et après ...

- Plein de nouvelles données : à quel heure vous rentrez chez vous, à quel heure vous couchez, vos préférences télé, etc.
- Une bonne analyse de vos habitudes conduira à vous recommander la bonne chaîne au bon moment (le client est satisfait mais le FAI aussi)



L'omniprésence des données : Applis mobile

The screenshot shows the Dismoiou website homepage. At the top, there's a navigation bar with the logo "dismoiou" (a speech bubble icon), "Paris", "Autres villes : Lyon · Marseille · Toulouse · Bordeaux · Nice · Nantes · Plus »", and links for "App mobile", "Inscription", and "Connexion". Below the navigation, there's a large image of a street scene with people at outdoor tables. Overlaid on this image is a white rectangular modal box with the title "Inscrис-toi pour continuer" (Sign up to continue) in bold black font. Below the title, it says "Inscrис-toi en un click et découvre les bonnes adresses de tes amis" (Sign up in one click and discover the good addresses of your friends). There are two buttons: a blue "Connexion Facebook" button with a white "f" icon, and a smaller "Ou inscrис-toi par email" (Or sign up via email) link. To the right of the main image, there's a smaller, semi-transparent overlay box with the text "partage tes" and a "Rechercher" (Search) button.



Dismoiou : découvre de nouvelles adresses et partage tes expériences !

Avec Dismoiou, redécouvre ta ville, ne passe plus à côté d'une bonne adresse, cumule des Points Maximiles et échange-les contre des cadeaux.

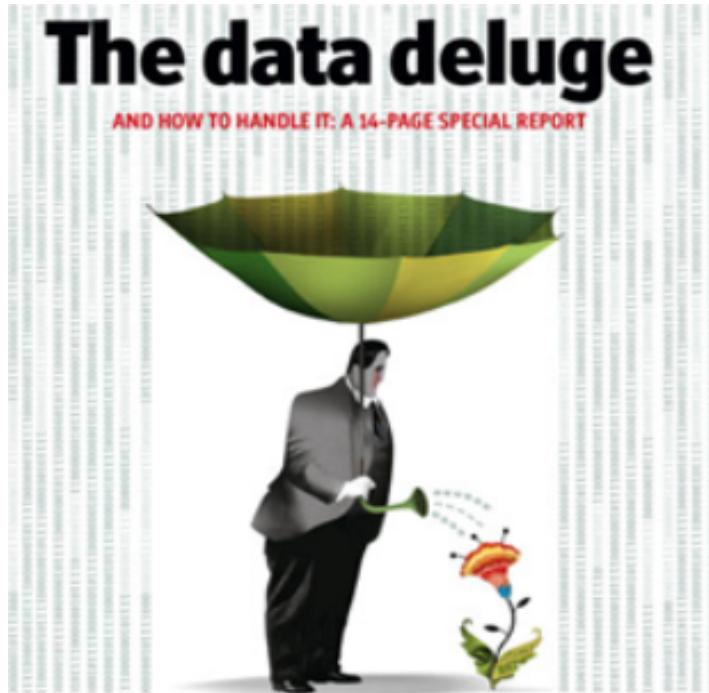
[Connexion Facebook](#)

Télécharge l'application



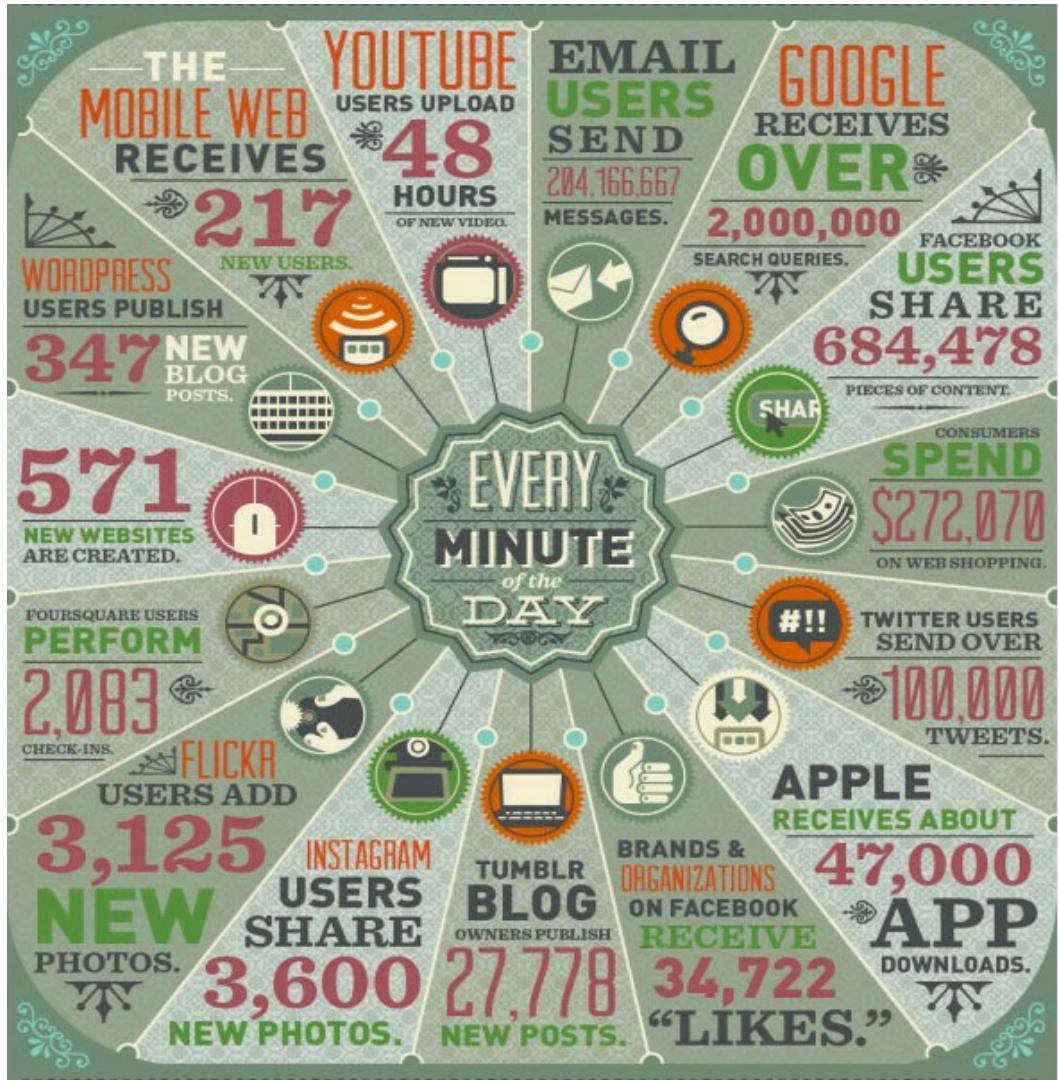
[En savoir plus sur les applications mobile](#)

Motivations du Big Data : data deluge & digital economy



1 à 2 Zeta Octets par an
40 Zeta Octets en 2020

$$1 \text{ ZO} = 10^{21} \text{ Octets}$$
$$= 1 \text{ milliard} * 1 \text{ TO}$$



DONNÉES ET INFORMATION

- Les entreprises produisent des quantités énormes de données
 - Sources: affaires, science, géographie, environnement...
- Les données sont des ressources de valeur
- Besoin de techniques pour analyser les données et extraire des informations / connaissances automatiquement
 - Données = faits
 - Information = modèle / motif (pattern) dans les données

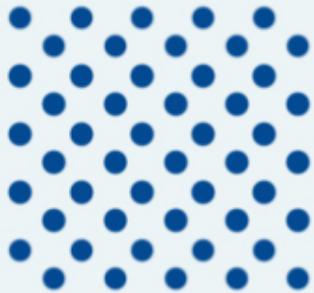
VOCABULAIRE

- **Population** : ensemble des éléments considérés dans une étude particulière,
- **Echantillon** : sous ensemble de la population,
- **Variables** : ensemble de caractéristiques d'une population,
 - **Quantitatives** : nombres sur lesquels les opérations usuelles (somme, moyenne, ...) ont un sens.
 - **Variable continue**: peut prendre n'importe quelle valeur réelle appartenant à un intervalle donné.
 - **Variable discrète** : peut prendre que l'une des valeurs définie dans une liste restreint de valeurs possibles.
 - **Qualitatives**: catégorielles: appartenance à une catégorie de données.
 - **Variable nominale** : (exemple: sexe)
 - **Variable ordinaire** : valeurs ordonnées (exemple : très résistant, assez résistant,)

DE QUOI S'AGIT-IL ?

Dimensions du Big Data

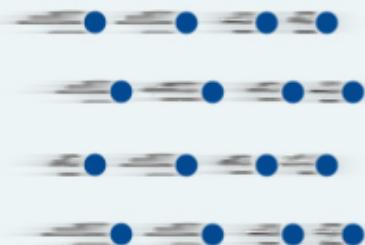
Volume



Données brutes
d'origine

Terabytes à exabytes
de données
disponibles

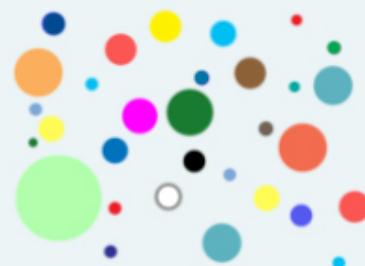
Vitesse



Données
dynamiques

Analyse en temps réel,
décision en une
fraction de seconde

Variété



Données
hétérogènes

Format structuré,
non structuré, texte,
multimédia

Véracité



Données
incertaines

Cohérence, fiabilité,
qualité et prédictibilité
des données

PROJET BIG DATA

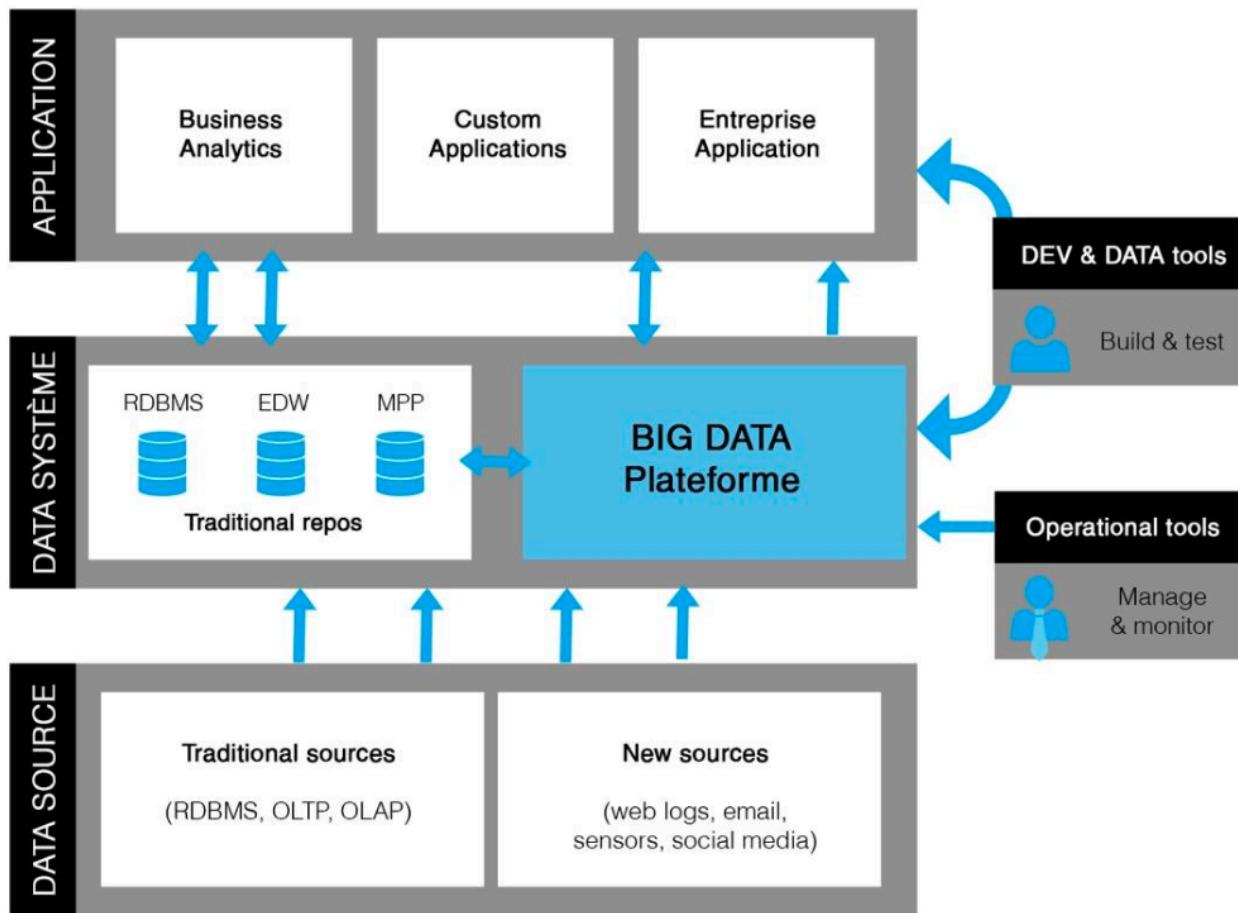
- ▶ Vue 360 Clients
- ▶ Analyse comportementale multicanal digital
- ▶ Système d'archive de données
- ▶ Datawarehouse offload et remplacement ETL
- ▶ Infrastructure Log Analysis
- ▶ Analyse de risques / fraudes
- ▶ Maintenance industrielle (IoT)
- ▶ Nouveaux business modèle : le tout connecté (équipement, voitures, ..)
- ▶ Segmentation clients en temps réels
- ▶ KPI temps réels / Analyse cross systèmes

CONCRETEMENT CA DONNE QUOI ?

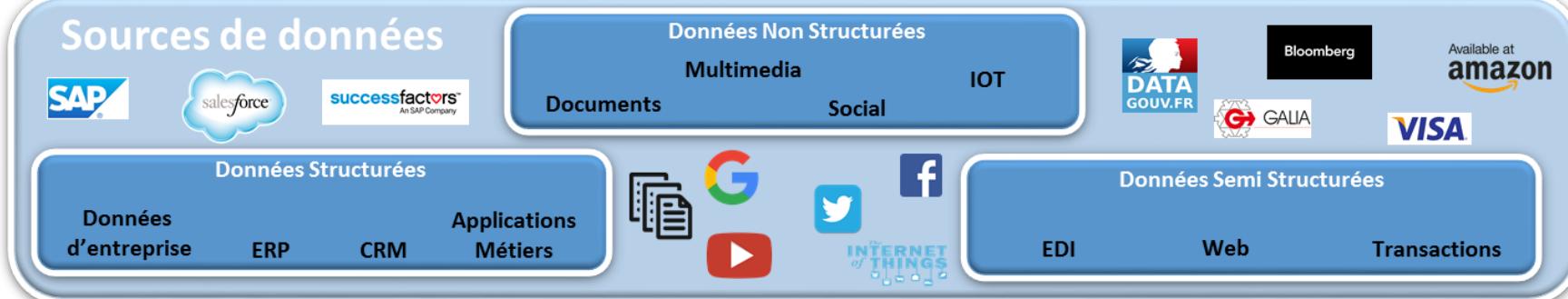
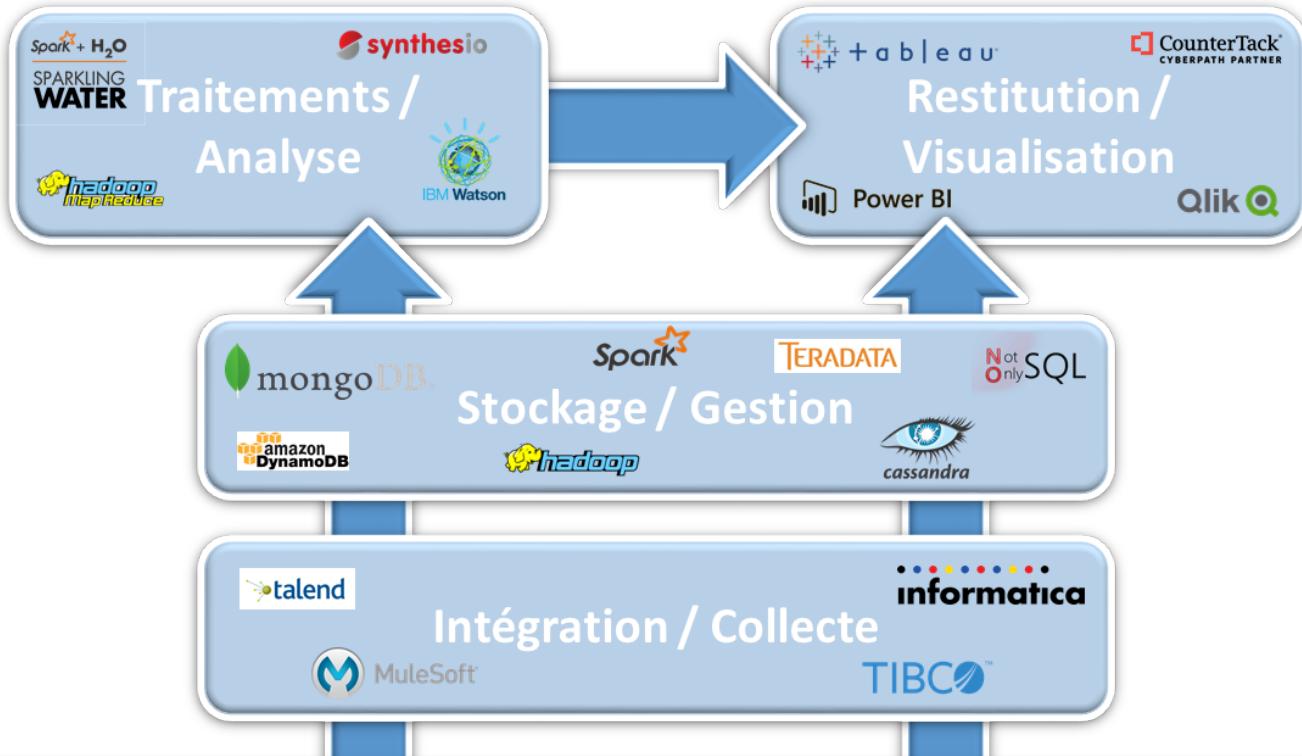
EXEMPLES PRATIQUES

- ▶ E-reputation
- ▶ Projet Velib
- ▶ Projet pipeline TOTAL Petrochimie
- ▶ UBER
- ▶ Google Analytics
- ▶ Facebook Ads

LES TECHNOLOGIES CONCERNÉES



LES OUTILS



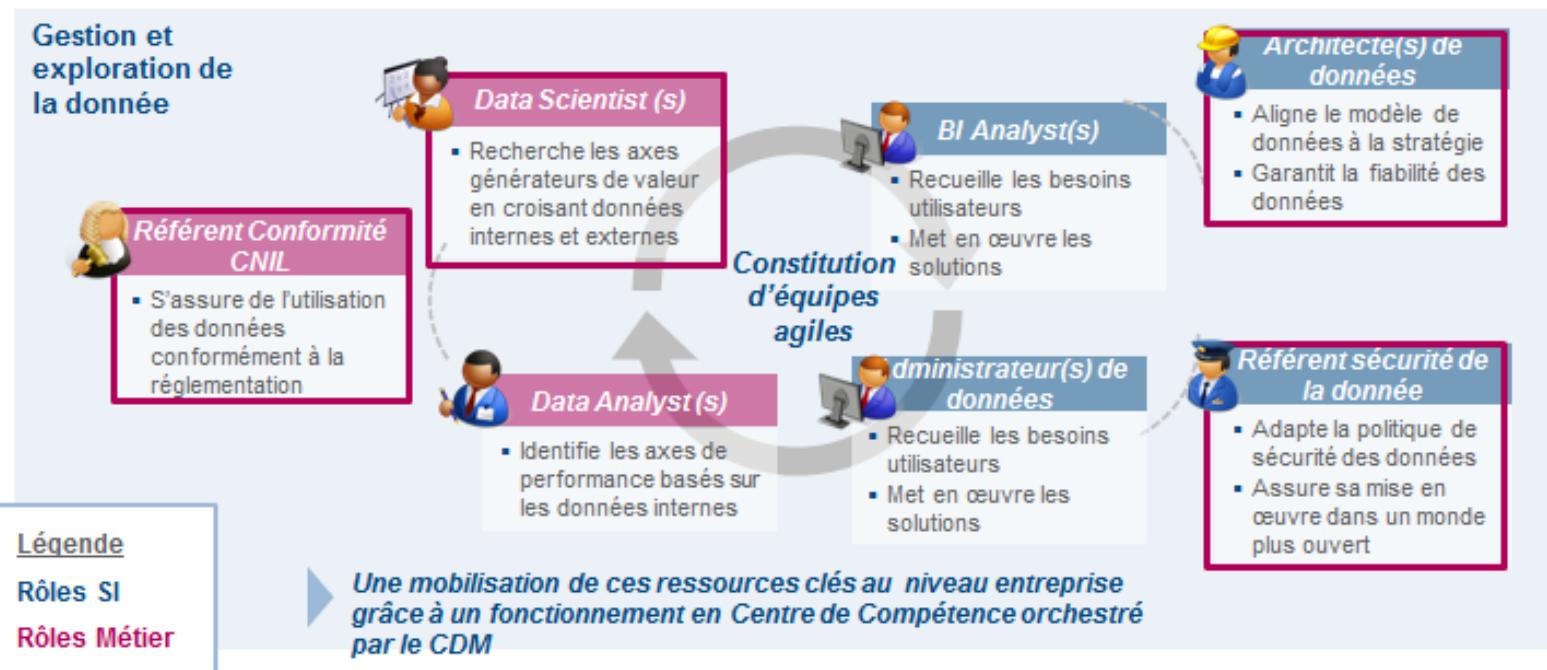
DÉMYSTIFICATION DU BIG DATA

Le BigData est un ensemble d'algorithmes et d'outils non classiques qui rendent possibles le stockage, l'analyse et la visualisation de quantités très importantes (tera/peta/exaoctets) de données non structurées ou mixtes au moyen de centaines ou de milliers de machines travaillant en parallèle.



JE NE VEUX PAS TRAVAILLER ... ! PAM PAM PARAM... :-D

LES DIFFÉRENTS MÉTIERS DU BIG DATA



DES ARÔMES ET INGRÉDIENTS QUI SE DÉCLINENT DANS TOUS LES
SECTEURS D'ACTIVITÉ

POUR QUI, POUR QUOI ?

Industrie



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

Distribution



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

Banques



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

Assurance



- Fraudes et risques
- Recommandation client
- Tarification à l'usage, personnalisation

Santé



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

Transports, loisirs



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

Secteur public



- Services informationnels
- Fraudes, abus
- Sécurité publique
- Personnalisation de la relation citoyen

Telecom



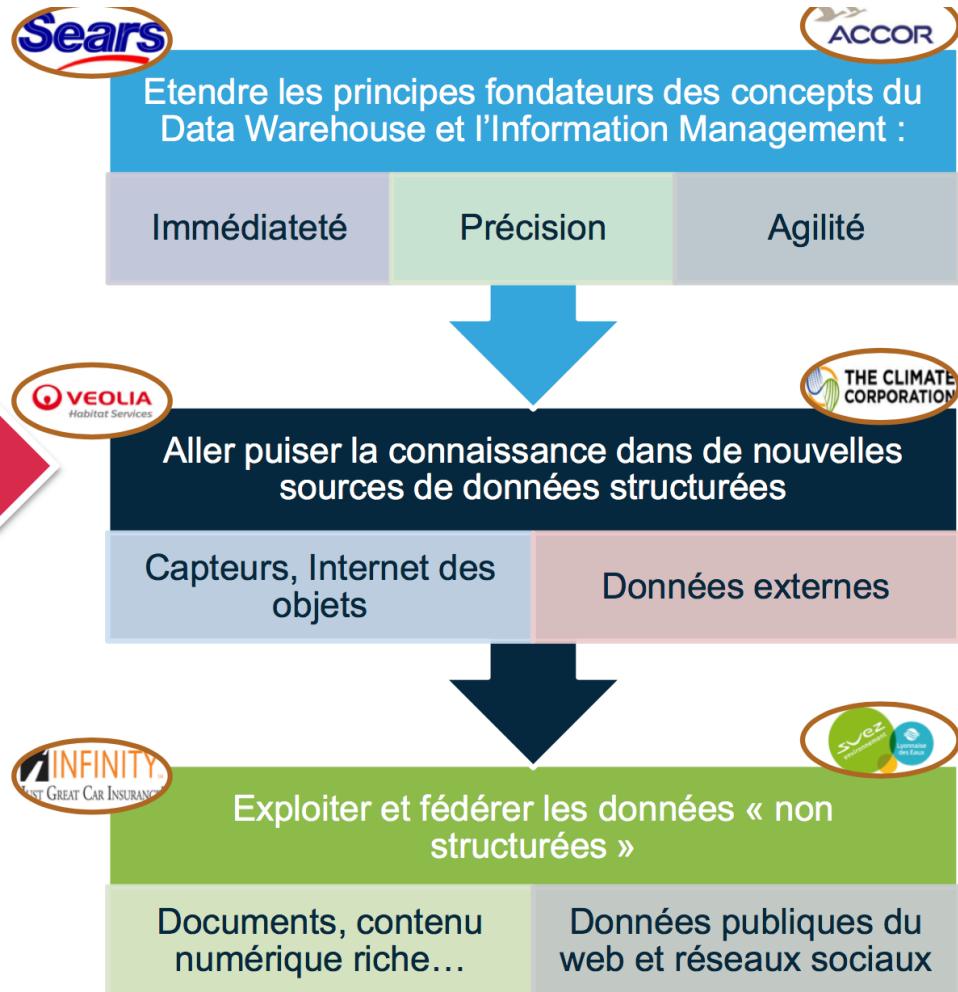
- Parcours clients multi-canaux
- Partage de données de géo localisation
- Fraudes et analyse du comportement client

Produits gde conso.



- Analyse de sentiments et retour produits
- Relation personnalisée avec le consommateur
- Produit comme un service

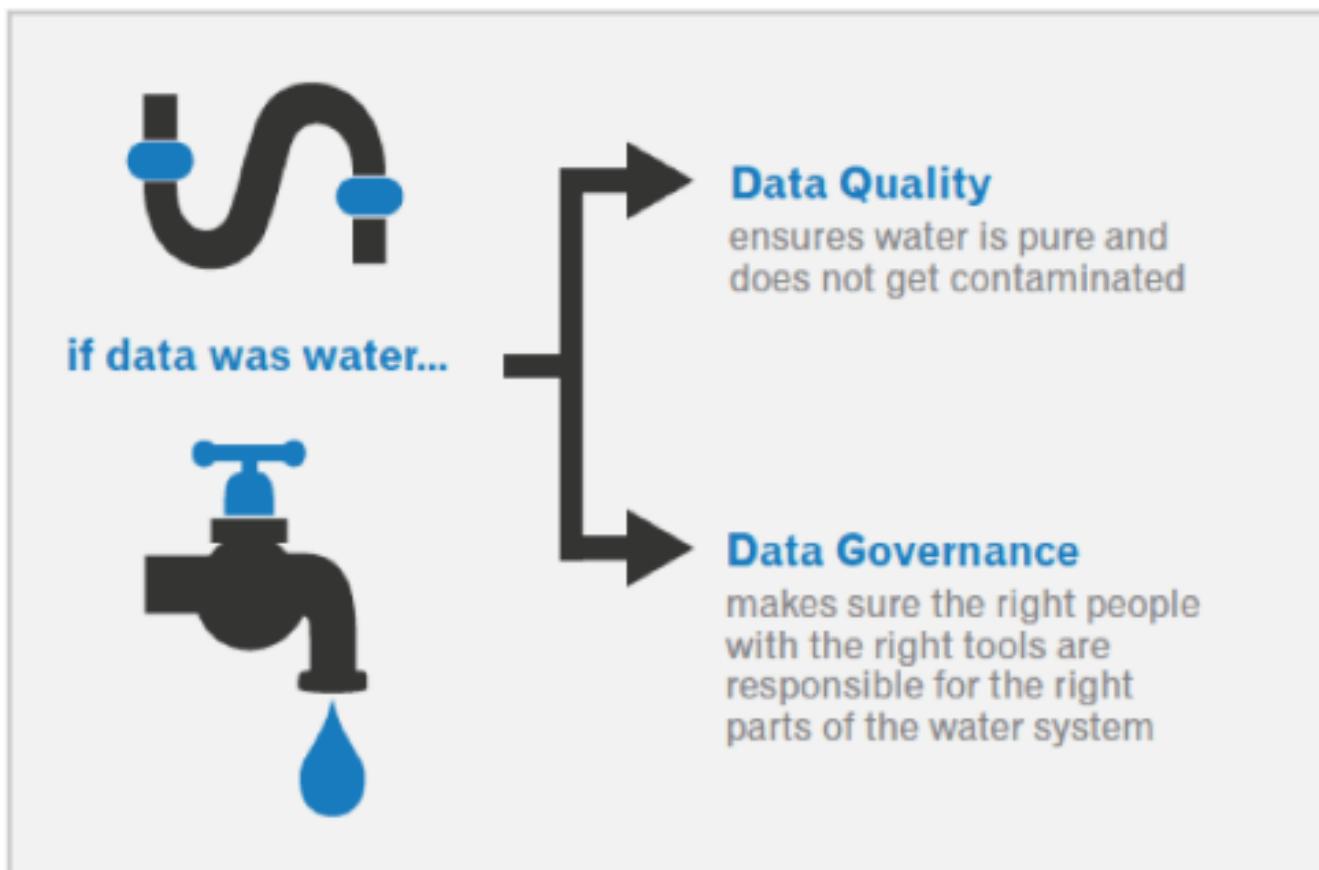
BI OU BIG DATA ?



LA QUALITÉ DES DONNÉES

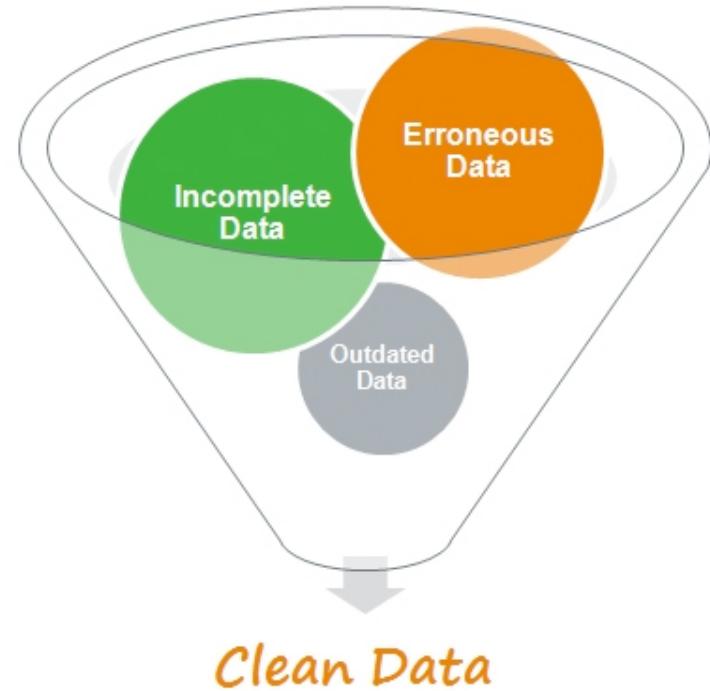
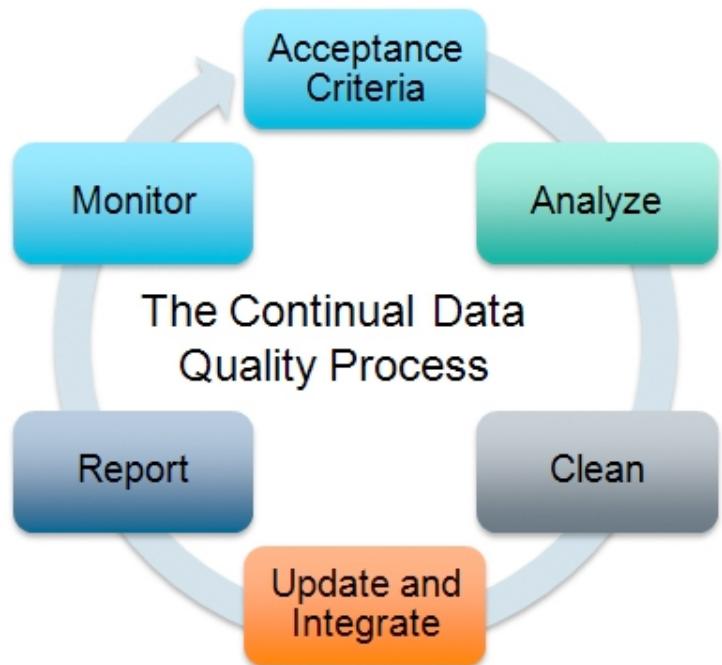


LA QUALITÉ DES DONNÉES

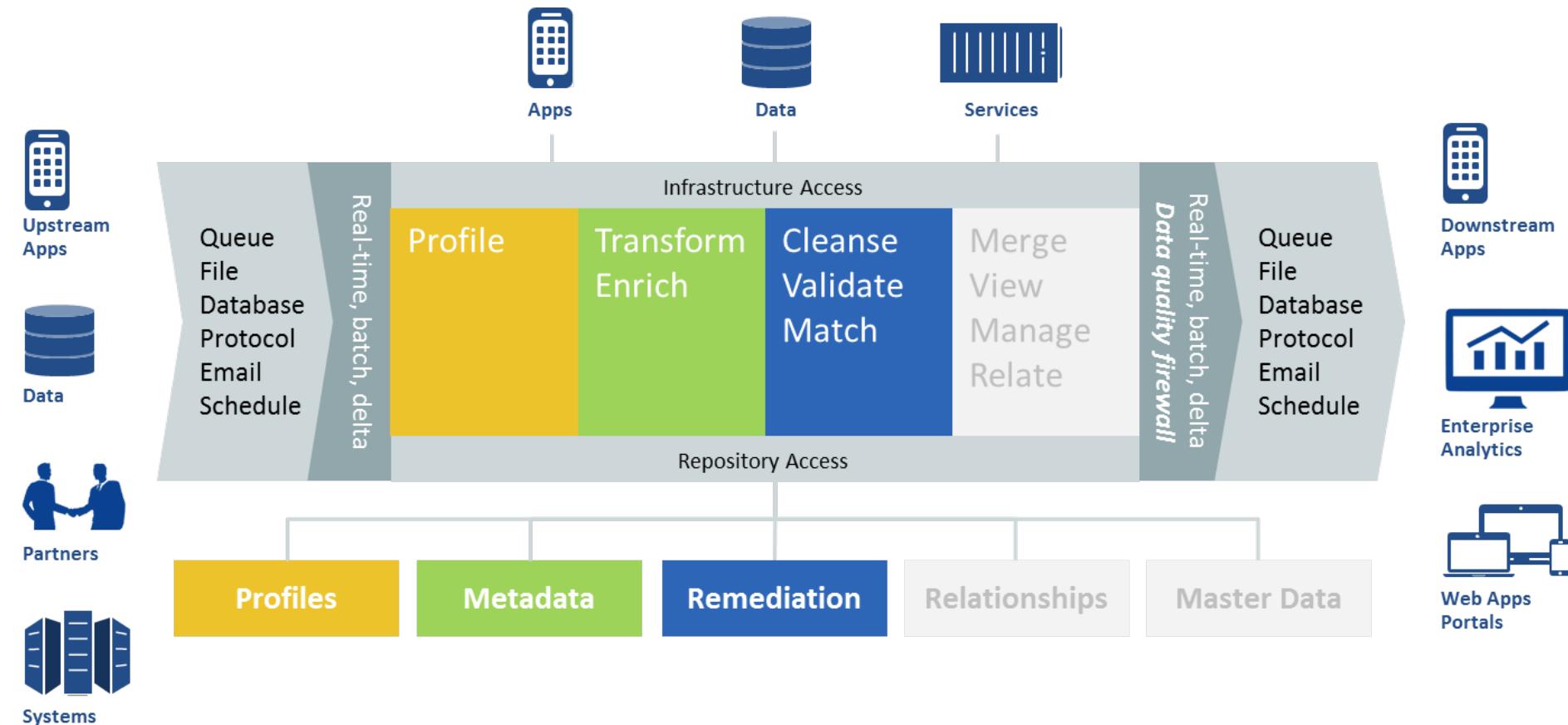


ET LE PROCESS ..

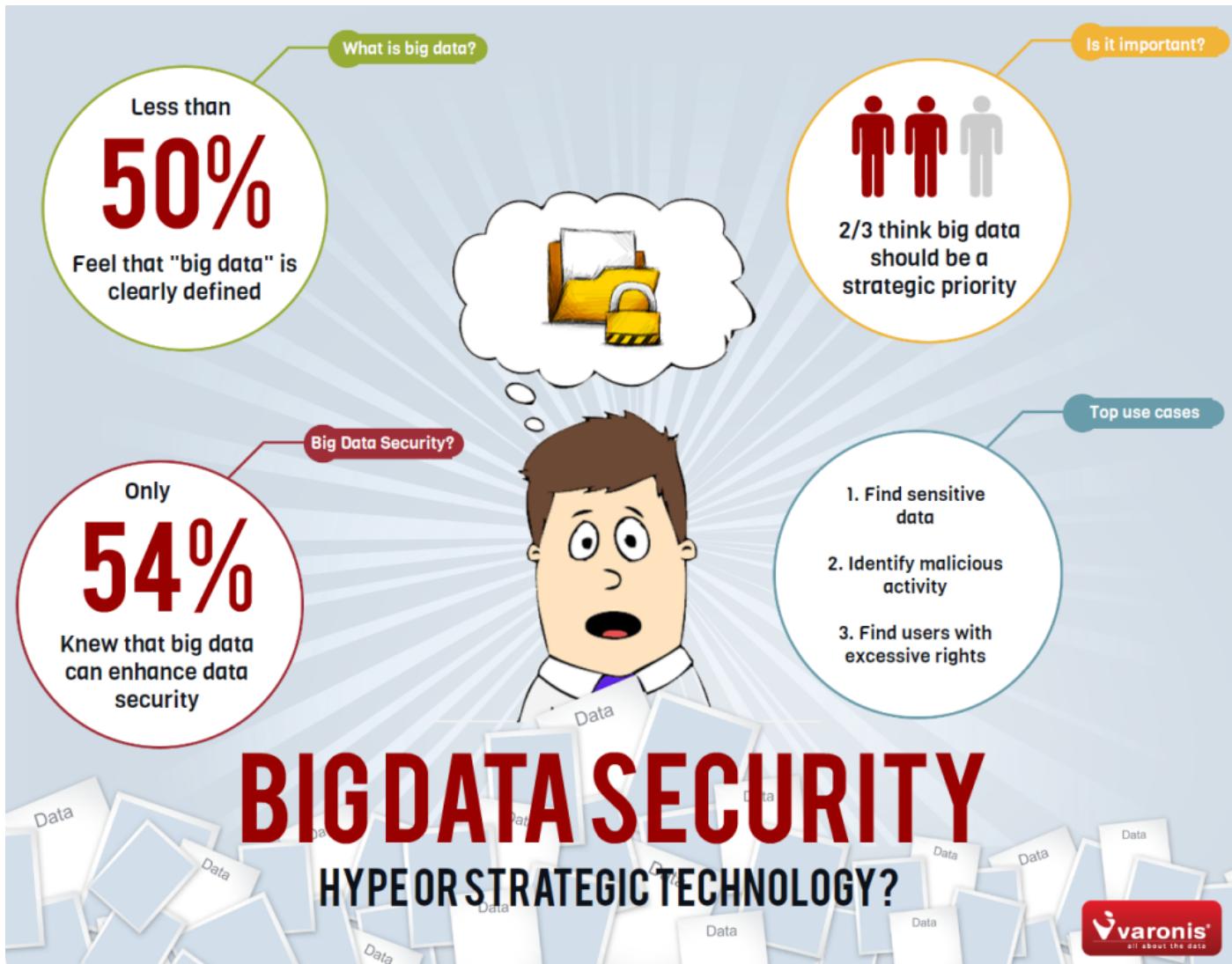
LA QUALITÉ DES DONNÉES



LE TRAITEMENT DES DONNÉES

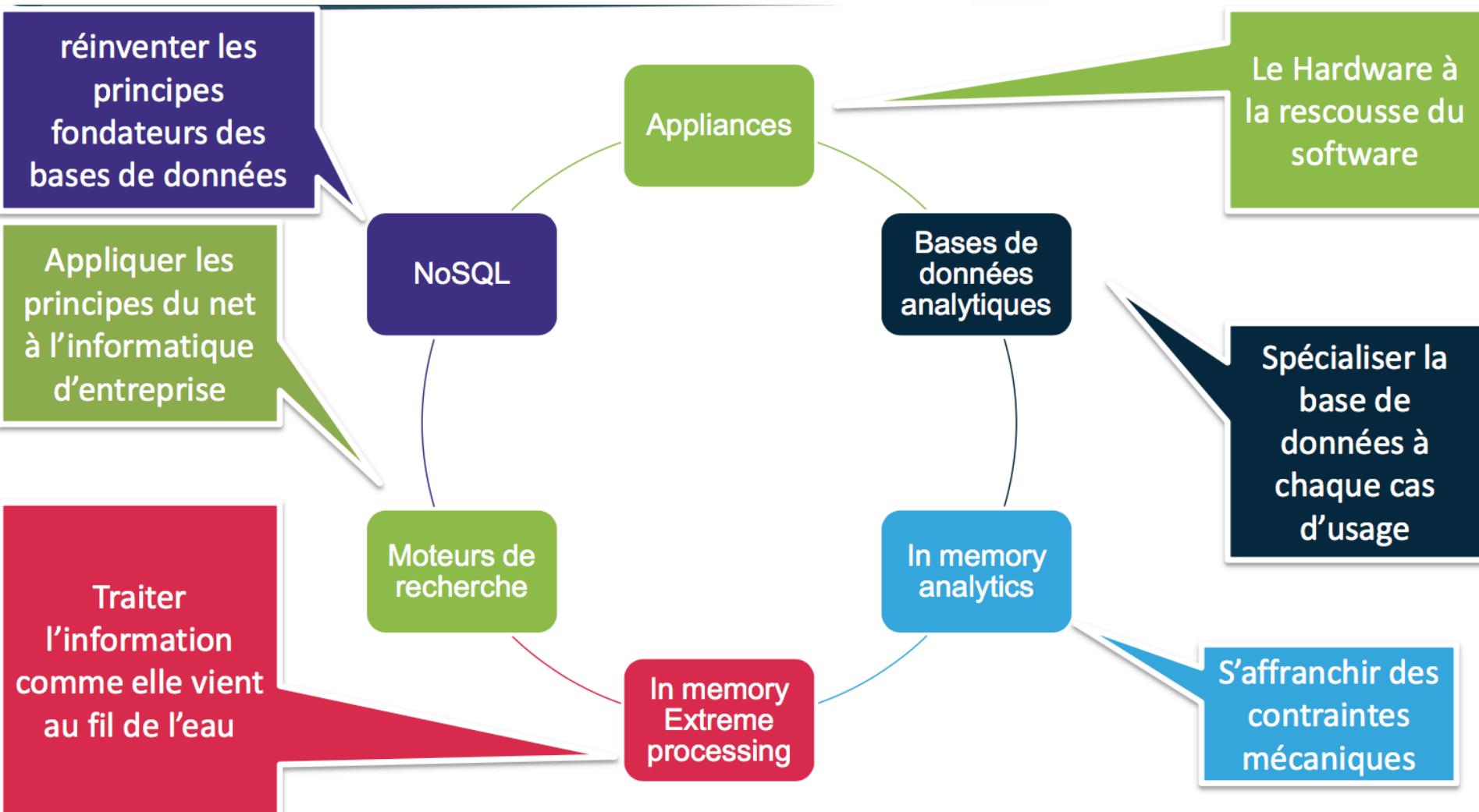


LA PROTECTION DES DONNÉES



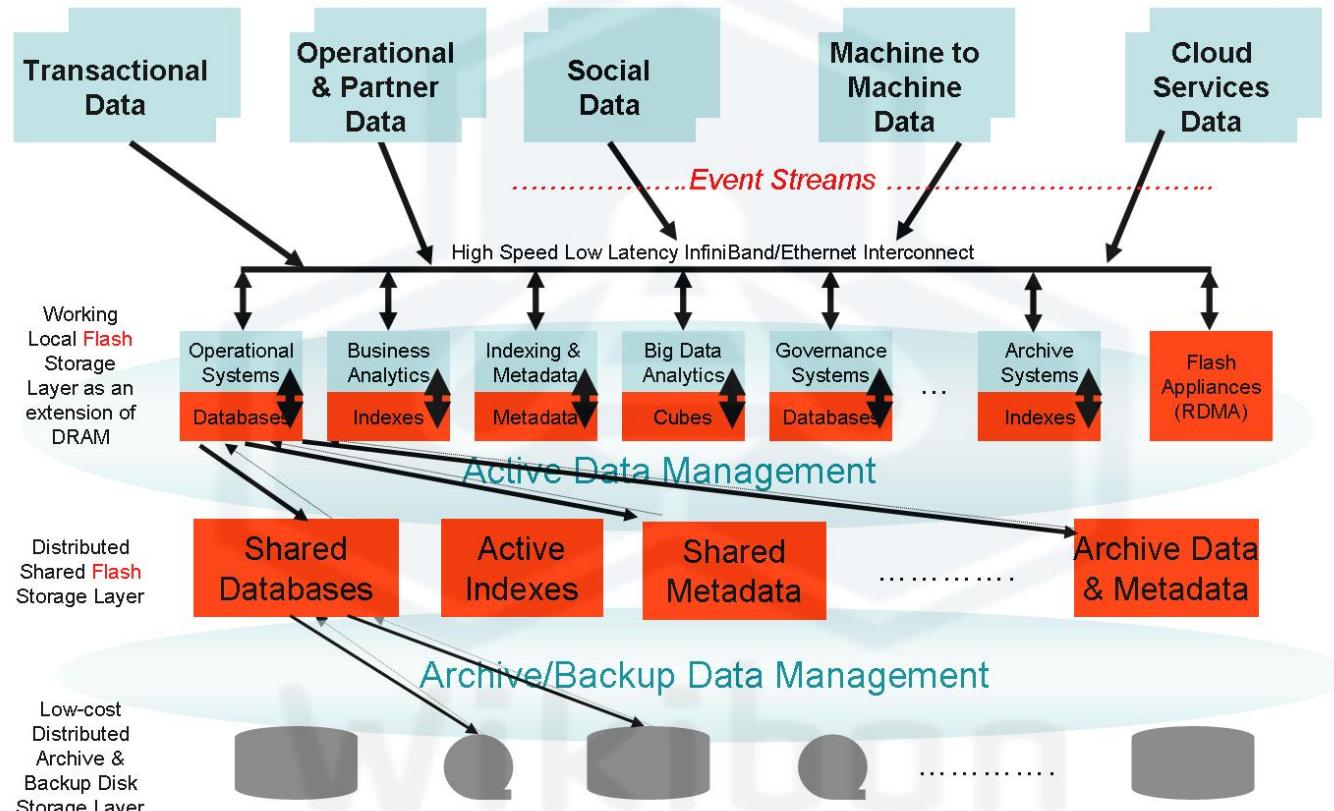
ZOOM SUR LES TECHNOLOGIES DE GESTION DES DONNÉES (« STOCKAGE »):
LA TAILLE UNIQUE N'EXISTE PAS, LA SOLUTION DÉPEND DES USAGES

L'IMAGE DE LA DONNÉE



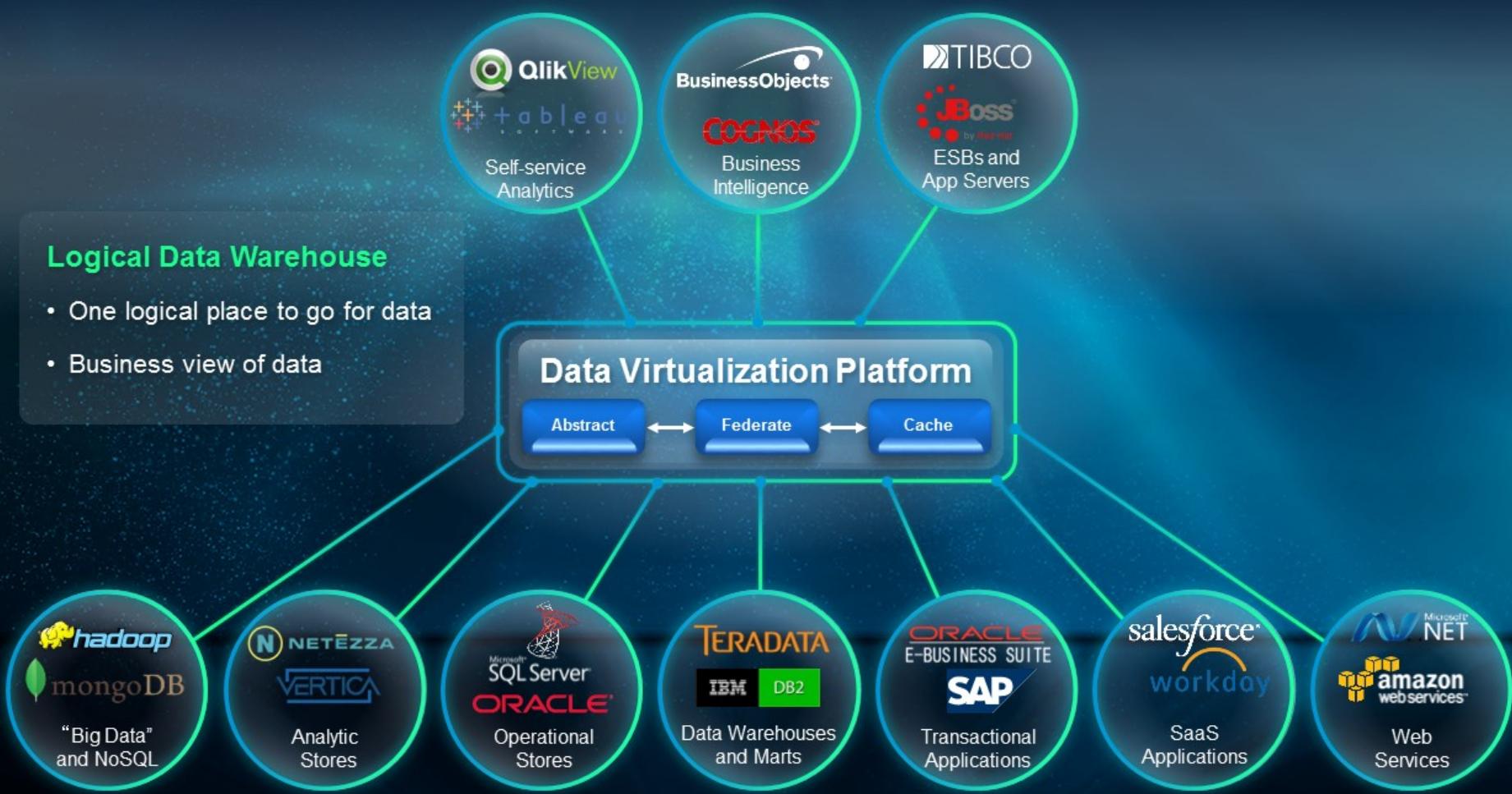
LES OUTILS DE STOCKAGE

Real-time Big Data Processing



Parallel Processing of Transactional, Analytic, Operational & Archive Systems

LES SOLUTIONS DE VISUALISATION



ANALYSE DES CARACTÉRISTIQUES DES PROJETS EN BIG DATA

- *Globalement, 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, et ne permettent plus d'obtenir des résultats dans des délais raisonnables ;
 - Le système d'information de l'entreprise est, ou devra prochainement être, connecté à des systèmes externes comportant une volumétrie de données élevée ;
 - Les données ne sont pas toutes structurées ;

ANALYSE DES CARACTÉRISTIQUES DES PROJETS EN BIG DATA

- *Ou encore :*
 - Certaines informations sont produites en temps réel par des capteurs, ou des événements extérieurs à l'entreprise (générés par des clients par exemple) ;
 - Les besoins de puissance de calcul peuvent varier de façon importante ;
 - L'analyse des données est peu développée ou nécessiterait de l'être davantage ;
 - La notion d'analyse prédictive est absente de l'entreprise et constituerait pourtant un atout majeur face à la concurrence.

LES PROJETS EN BIG DATA

MYTHES ET RÉALITÉS

- Le Big data sort des schémas classiques, ce qui crée des freins réels liés à l'inadéquation de l'organisation ou des compétences de l'entreprise.
- Il y a des mythes qu'il convient de faire disparaître.

LES PROJETS EN BIG DATA

MYTHES

"Les projets Big Data sont avant tout techniques et visent à gérer d'importantes volumétries de données".

LES PROJETS EN BIG DATA

RÉALITÉS

- Il y a des projets dont l'objectif est de résoudre des problèmes de performance lorsque les technologies classiques atteignent leurs limites.
- Les enjeux sont alors en effet surtout portés par les DSI
- les DSI privilégient plutôt des solutions où la technologie masque la complexité des volumes (par exemple *Exadata, Teradata*).
- Sa mise en place sera en pratique souvent moins disruptive en termes de pratiques de développements.

LES PROJETS EN BIG DATA RÉALITÉS

Les projets Big Data avec les enjeux et les ROI (Return on investment) les plus importants sont ceux où l'innovation porte fortement sur les équipes métier

LES PROJETS EN BIG DATA

RÉALITÉS: EXEMPLE

Projet :

Analyse massive du comportement des clients dans des entreprises BtoB (business to business) pour assister les commerciaux avec des recommandations de produits : ce type de projet touche directement les équipes de vente.

Apport Business :

Augmentation du chiffre d'affaires par client grâce à la vente de produits supplémentaires (*les gains peuvent se compter en dizaines de millions d'euros bien au-delà de l'investissement IT*)

Les projets Big data générant les opportunités business les plus importantes doivent être portés par les directions métier, voire les directions générales. Ce serait une erreur que de penser que seules les DSI sont concernées.

LES PROJETS EN BIG DATA

MYTHES

"Il faut rationaliser les référentiels et nettoyer les données avant de les analyser".

LES PROJETS EN BIG DATA

- L'approche classique génère :
 - Un fort besoin de référentialisation des données pour pouvoir rapprocher les différentes sources.
 - Des projets d'amélioration de la qualité de données en vue de la nouvelle exploitation des données.

LES PROJETS EN BIG DATA

APPROCHE PRÉDICTIVE

- La probabilité que le client achète le produit qui lui serait proposé, il n'est pas nécessaire que les données soient fiables à 100 %.
- c'est le propre des algorithmes d'analyse proche de l'intelligence artificielle (ou de **Machine Learning**) de faire ce type de rapprochement et de distinguer les données pertinentes de celles qui ne sont pas porteuses d'information.
- La rationalisation des référentiels et le nettoyage des données sont rarement un prérequis incontournable à la mise en œuvre d'un projet Big data, typiquement en cas d'analyse prédictive.

LES PROJETS EN BIG DATA

RÉALITÉ

Les projets Big data ont bien souvent un fort degré d'incertitude.

Première incertitude : pouvons-nous extraire depuis nos bases de données des informations créatrices de valeur pour l'entreprise ?

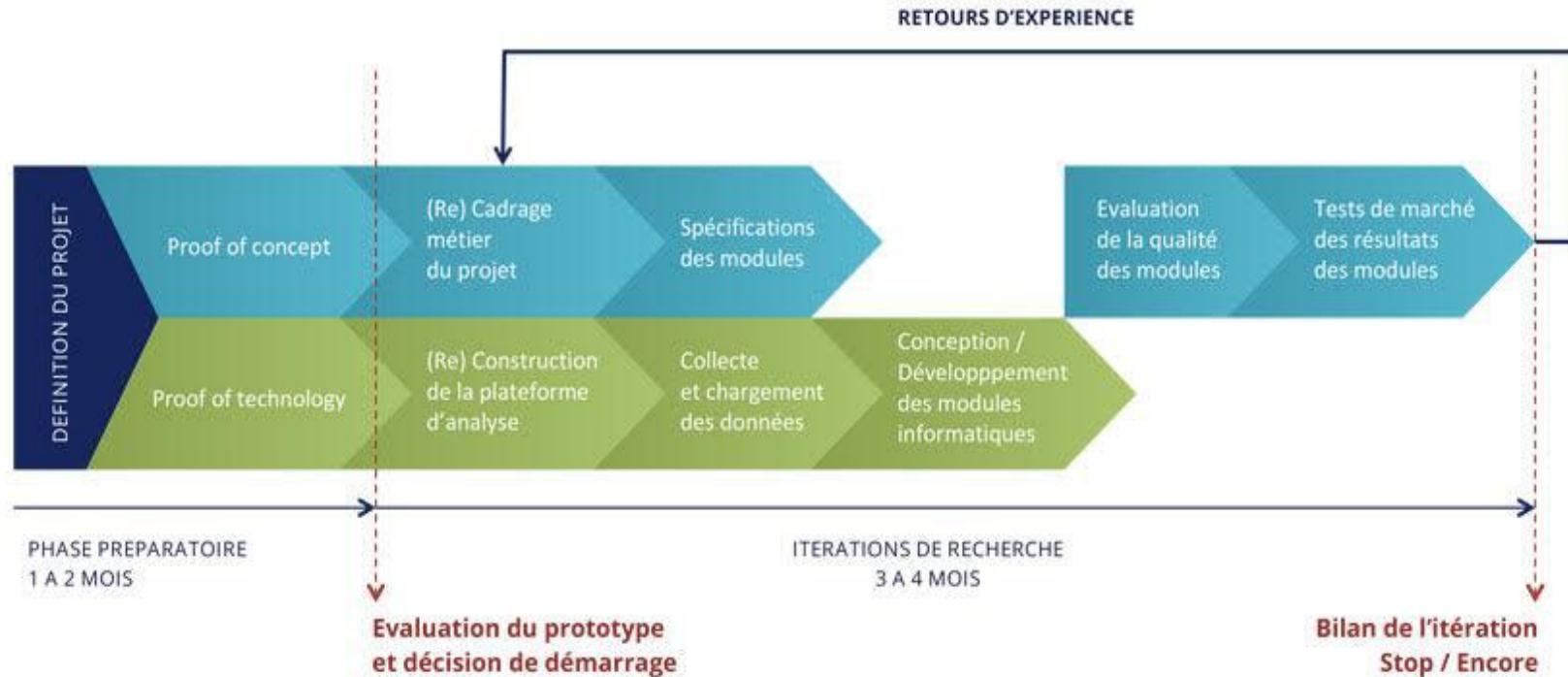
Seconde incertitude : y a-t-il des solutions techniques que nous maîtrisons suffisamment pour exploiter ces données ?

LES PROJETS EN BIG DATA RÉALITÉ

- Projet :
 - Analyse massive du comportement des clients dans des entreprises BtoB (business to business) pour assister les commerciaux avec des recommandations de produits : ce type de projet touche directement les équipes de vente.
- Apport Business :
 - Augmentation du chiffre d'affaires par client grâce à la vente de produits supplémentaires (*les gains peuvent se compter en dizaines de millions d'euros bien au-delà de l'investissement IT*)
 - Les projets Big data générant les opportunités business les plus importantes doivent être portés par les directions métier, voire les directions générales. Ce serait une erreur que de penser que seules les DSI sont concernées.

DÉMARCHE D'ORGANISATION PROJETS BIG DATA

SCHÉMA GÉNÉRAL DU PROJET BIG DATA



DÉMARCHE D'ORGANISATION PROJETS BIG DATA

- 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.

DÉMARCHE D'ORGANISATION PROJETS BIG DATA

- 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.

DÉMARCHE D'ORGANISATION

PROJETS BIG DATA

- Au sein de la maîtrise d'œuvre, on trouvera dans pratiquement tous les cas de figure :
 - un chef de projet ou un directeur de projet ;
 - 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.

DÉMARCHE D'ORGANISATION PROJETS BIG DATA

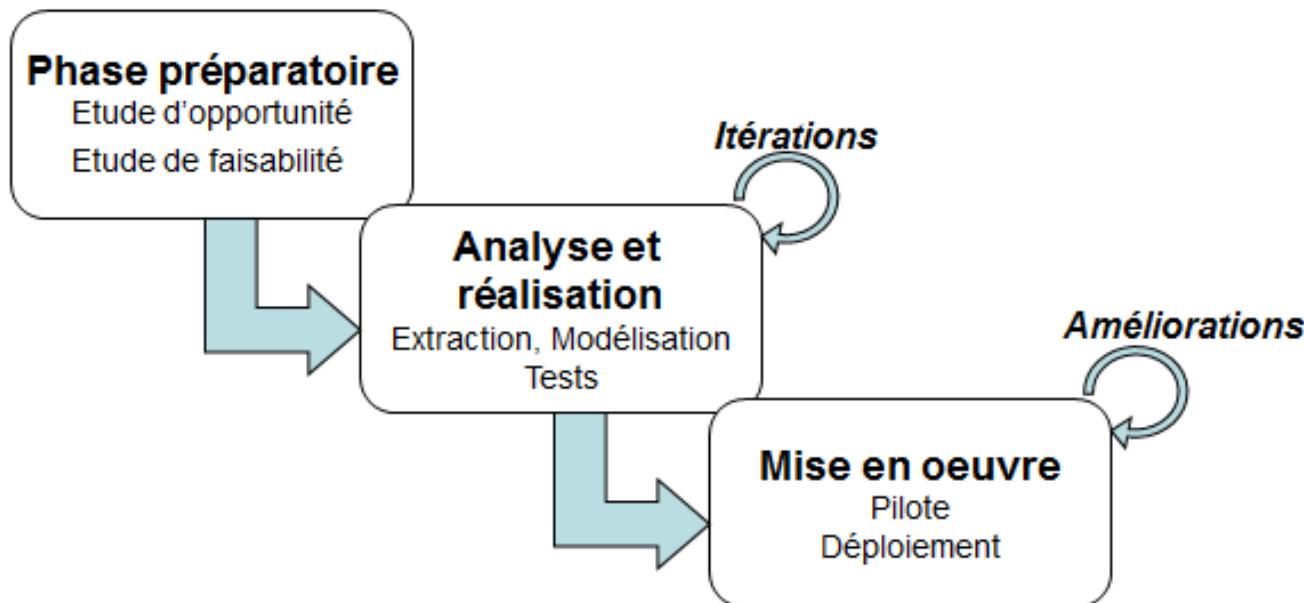
- Dans ce type de projets, le modèle classique du cycle en V est déconseillé.
- Il faut construire la solution progressivement, en prévoyant quelques itérations comprenant des interactions avec les futurs utilisateurs.

DÉMARCHE D'ORGANISATION

PROJETS BIG DATA

- Les méthodes incrémentales sont donc adaptées à ce type de projet.
- Diviser le projet en parties fonctionnelles cohérentes, chaque incrément pouvant être testé séparément et faisant l'objet de plusieurs itérations.
- L'objectif est d'impliquer les utilisateurs dans le développement, la fourniture des exigences et l'évaluation des itérations.

DÉMARCHE D'ORGANISATION PROJETS BIG DATA



Etapes d'un projet Big Data

CADRAGE MÉTIERS

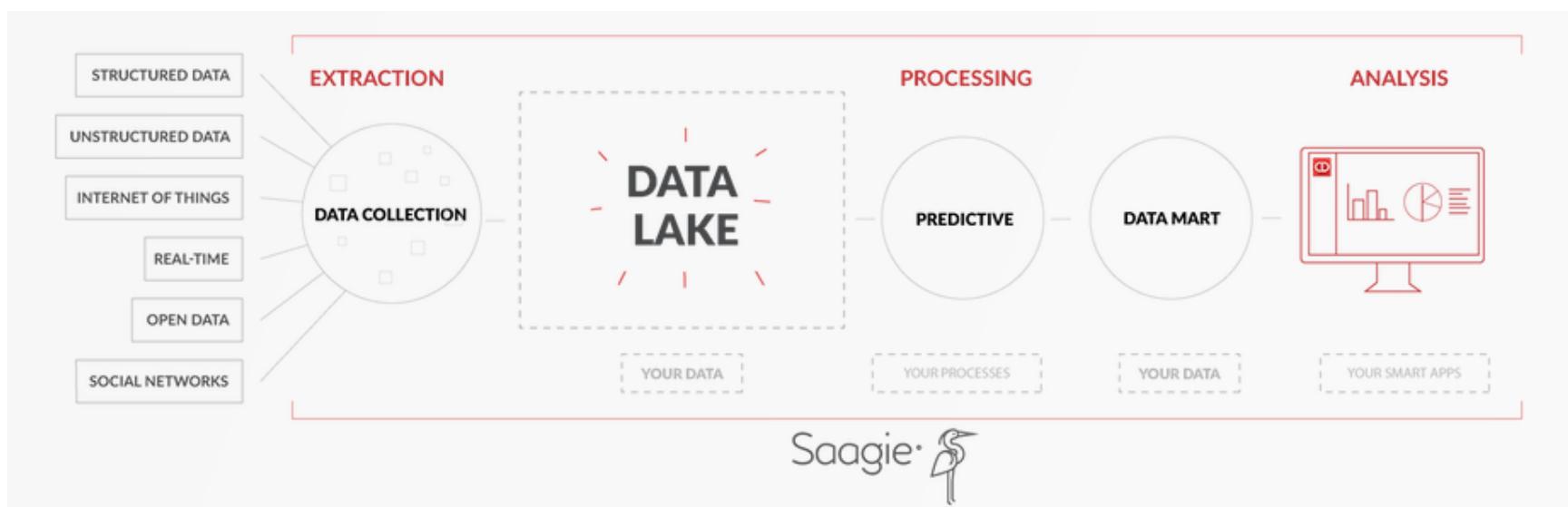
- La première chose importante à réaliser est de s'assurer du bien-fondé du projet lui-même.
- L'étude d'opportunité menée en phase d'avant-projet doit clairement mettre en évidence les apports du projet pour l'entreprise.

CADRAGE MÉTIERS

LES FACTEURS CLÉS DE SUCCÈS

- Il faut 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.

IDENTIFICATION DE NOUVELLES SOURCES DE DONNÉES





Marketing

- Définition des nouvelles offres et services
- Prédiction des appétences en termes d'achat
- Etc.

Ventes

- Profilage des comportements clients
- Prévision des risques d'attrition
- Détection des potentiels de ventes supplémentaires
- Etc.

Relation client

- Personnalisation de la relation
- Fidélisation des clients
- Etc.

RH

- Analyse prédictive du Turnover
- Identification des hauts potentiels
- Détection des talents en école
- Etc.

Finance / Contrôle de gestion

- Analyse de la performance des activités
- Etc.

IT

- Analyse de logs
- Traitement des incidents IT
- Etc.

DÉFINITION DES BESOINS TECHNOLOGIQUES

- Aussi 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 ?

PHASES DES PROJETS BIG DATA

