



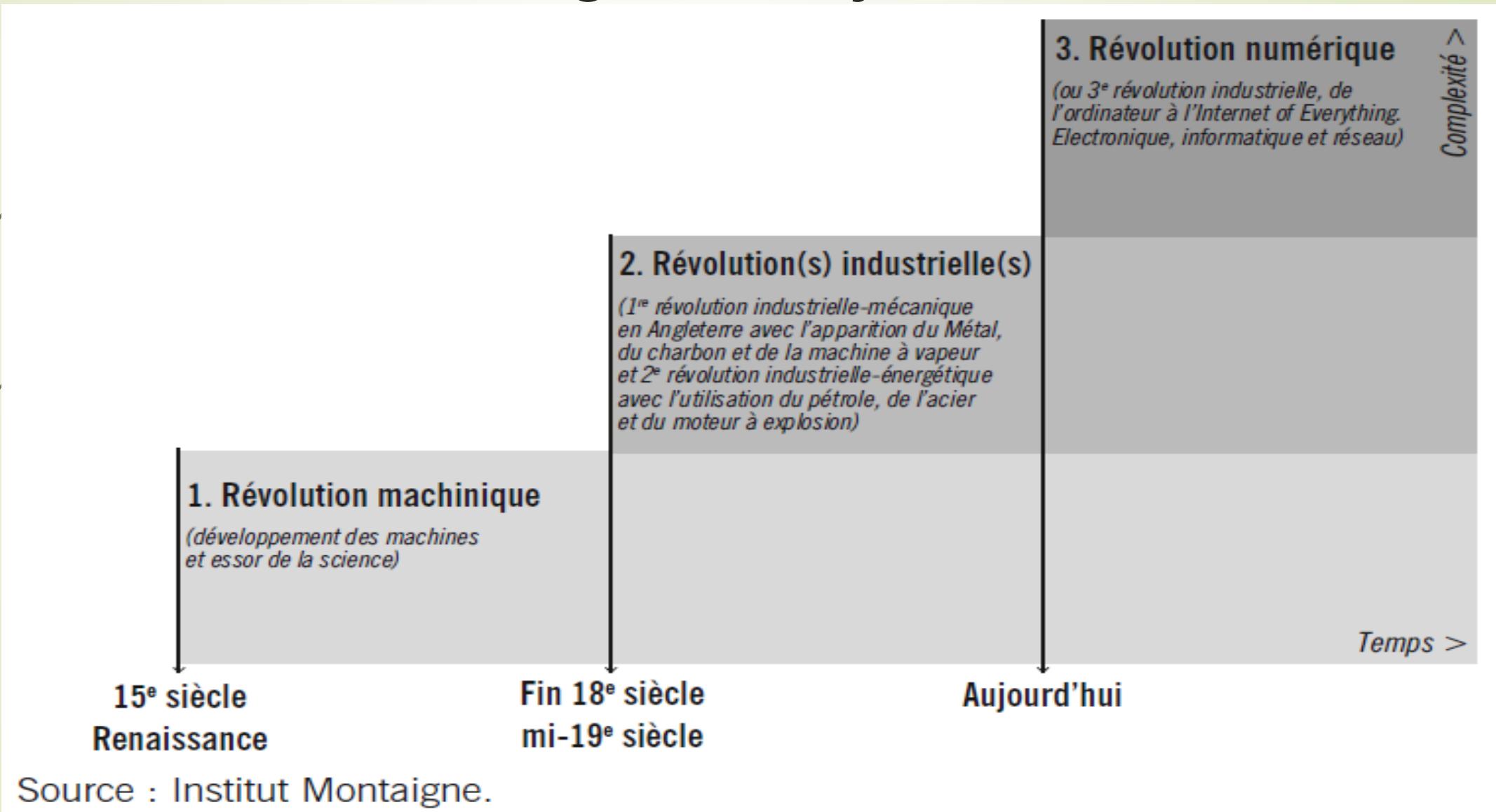
Dr. Ing. Ikbel DALY BRIKI

Révolution numérique

- La première grande révolution technique fut celle de la **machine** (Renaissance), dont la presse à imprimer typographique (Johannes Gutenberg) en 1450 reste un symbole.
- La **seconde** fut la révolution **mécanique** de l'ère industrielle.
- La **troisième** grande révolution technique de l'histoire moderne est la révolution **numérique**.
- Soutenue par les **objets connectés** et le **big data**, elle ouvre une nouvelle ère.



De Gutenberg aux objets connectés

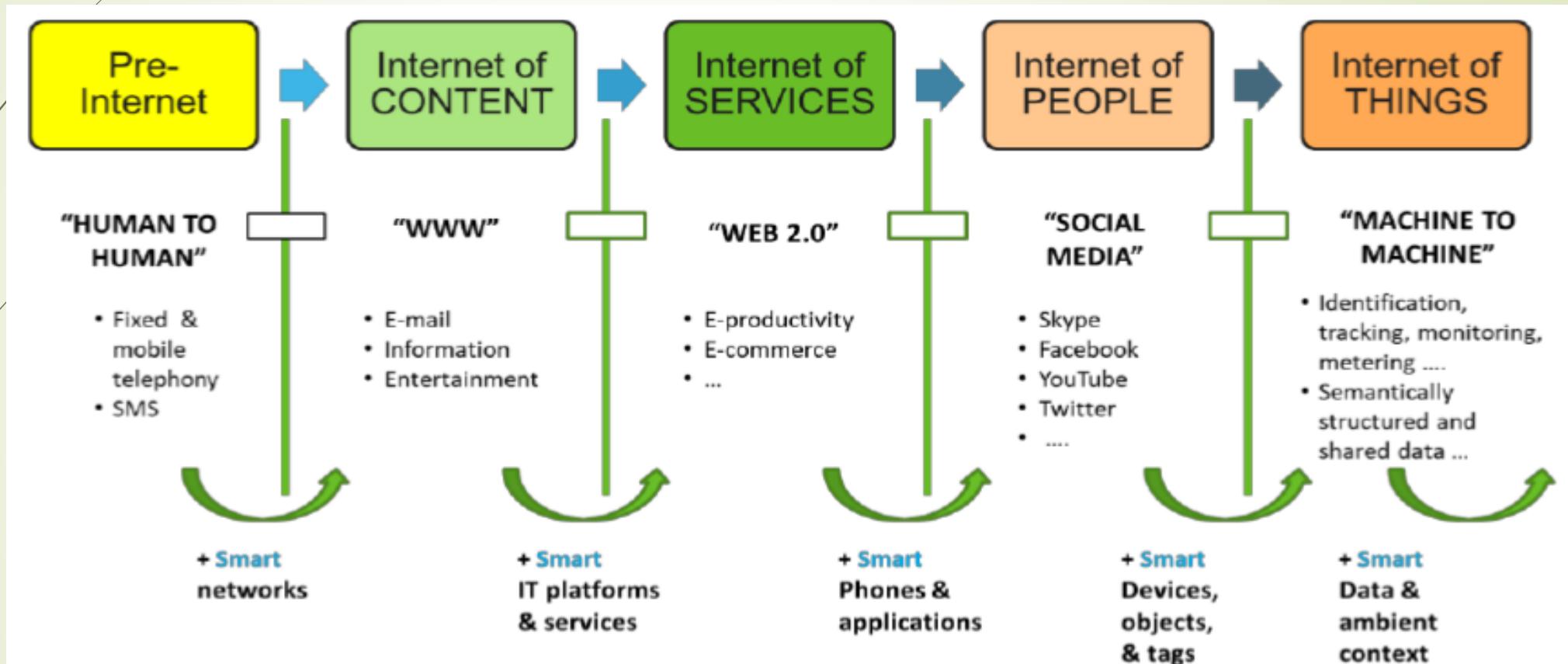


De Gutenberg aux objets connectés

Les trois grandes révolutions techniques de l'humanité, selon **Stéphane Vial**:

- **La première révolution** est celle de l'imprimerie, représentée par Johannes Gutenberg, qui a transformé la manière de diffuser l'information et de communiquer.
 - **La deuxième révolution** est celle des objets connectés, représentée par les objets connectés (IoT), qui permettent d'intégrer des dispositifs électroniques dans notre environnement quotidien et de collecter des données en temps réel.
 - **La troisième révolution** est celle des Big Data, qui concerne les jeux de données extrêmement volumineux qui ne peuvent pas être hébergés sur une seule machine
- ➔ contribution à la transformation de la manière dont nous communiquons, nous informons et nous connectons.

Evolution of the Internet of Things (IoT)



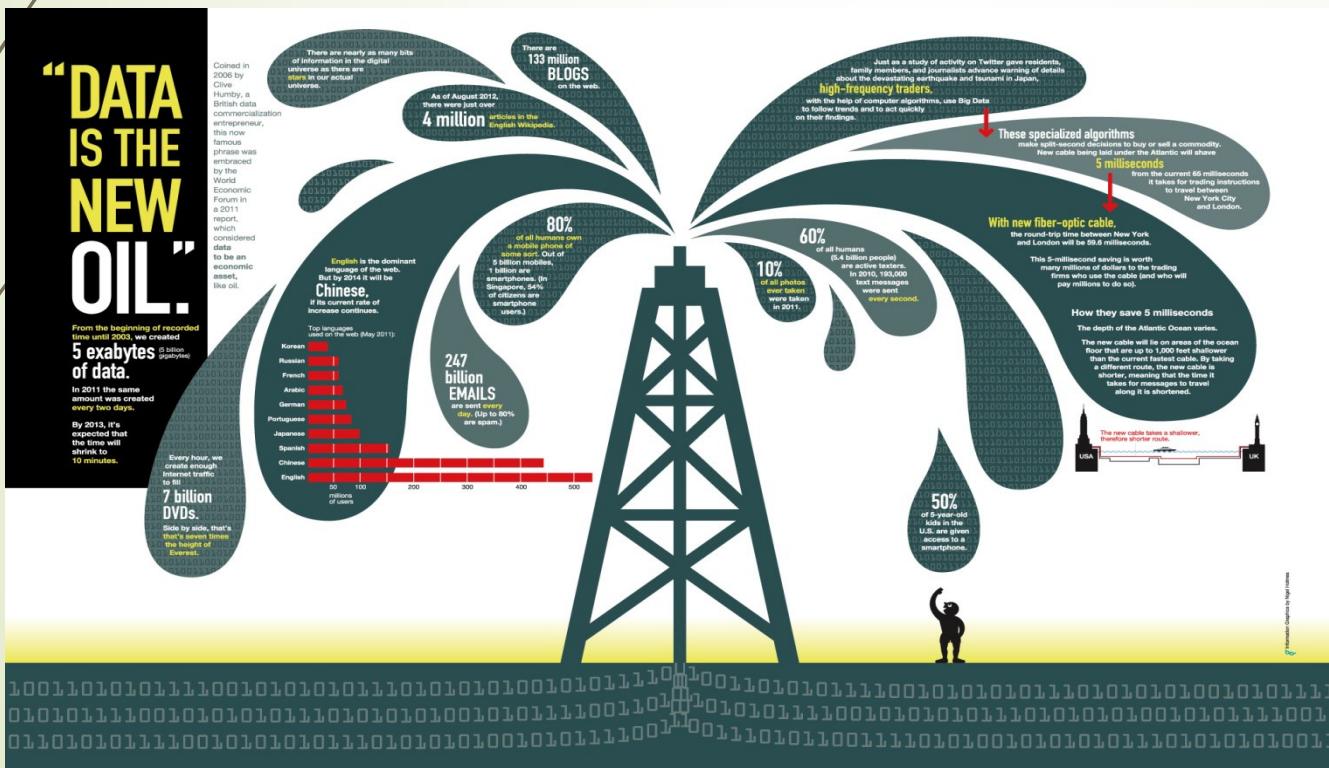
Source:

Internet of Things (IoT): The Next Cyber Security Target (Webinar)

Praveen Kumar Gandi, Head Information Security Services,

<https://www.slideserve.com/ClicTest/webinar-on-internet-of-things-iot-the-next-cyber-security-target>

Data is the new oil



« Les data sont le pétrole du XXI^e siècle »
 [Gilles Babinet (de l'ère numérique, un nouvel âge de l'humanité, janvier 2014)]

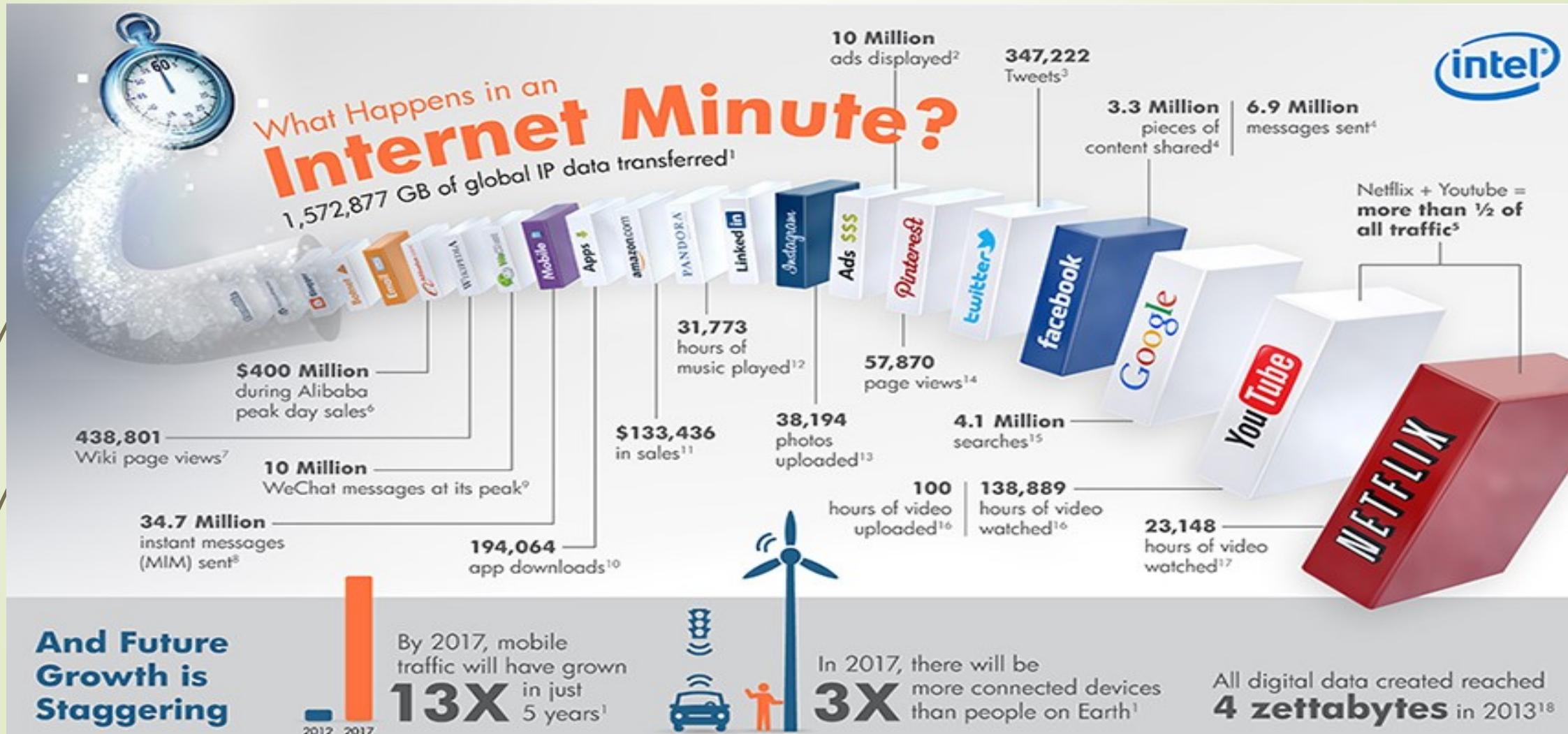
[Clive Huby, 2006 ; Infographie : Nigel Holmes]

Révolution numérique

- Les **objets connectés** occupent une place centrale:
 - outils au service des utilisateurs
 - collecteurs de données
- La **donnée** constitue la **matière première** de la révolution numérique.
- La **donnée** a été comparée au pétrole, ressource au cœur de la seconde révolution industrielle.
- Les **objets connectés** jouent pour le **Big data** le même rôle de catalyseur que la chimie ou l'automobile pour le pétrole.
- De même que le **pétrole brut** ne peut être utilisé comme combustible automobile, les **données brutes** ne sont pas pertinentes par elles-mêmes. Elles deviendraient en revanche créatrices de valeur une fois **analysées**.



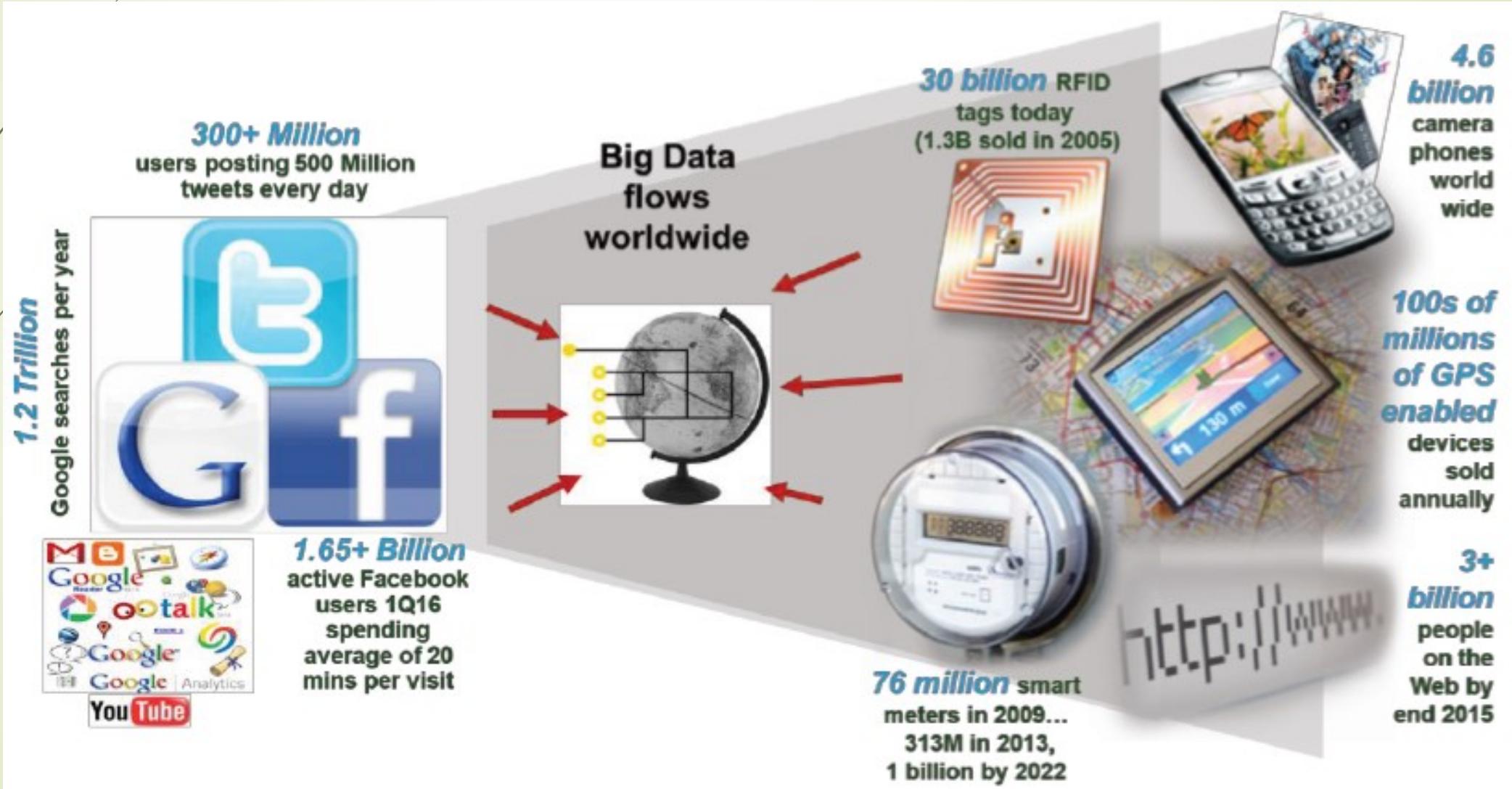
Sources des données massives



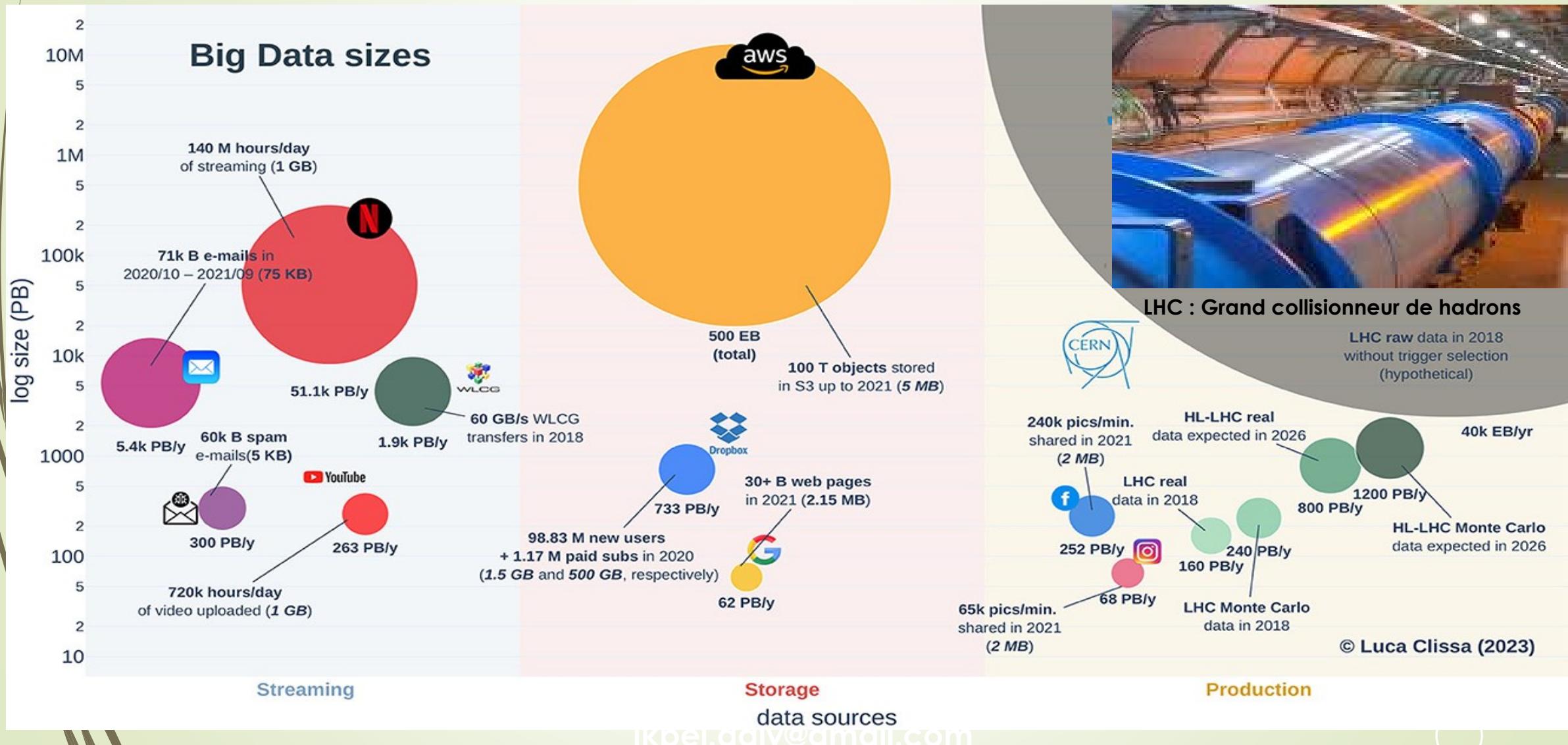
100 milliards d'adresses IP seront utilisées en **2025** (cent fois plus qu'en 2003).

1 zetta = un film qui dure **745 millions d'années.**

Growing interconnected & instrumented world



Sources des données massives



Sources des données massives

- ▶ Elles sont partout, c'est le déluge [The economist, 2010]



« DATA » (Donnée) vs « Information »

DATA : est une description élémentaire d'une réalité généralement que l'on observe ou que l'on mesure.

- ▶ Ex : Laptop, souris sans fil

Information : Ce que je peux DEDUIRE d'un ensemble de DATA (relation, context)

- ▶ Ex : un client achète un laptop et une souris sans fil

Connaissance : modèles qui résument le comportement des données.

- ▶ Ex : 80 % des clients qui achètent du laptop, achètent en plus une souris sans fil



« DATA » – Big Data



- ▶ **Métadonnées**
- ▶ 45° 50' 30" N 3° 15' 40" E
- ▶ 23/07/2018
- ▶ 12:30
- ▶ Canon



« DATA » en préfixe ou suffixe!

DATA en Préfixe :

- ▶ DATA base (1968 : Ted Codd et Modèle Relationnel), DBMS
- ▶ DATA bank
- ▶ DATA warehouse
- ▶ DATA mart
- ▶ DATA mining (OLAP, Corrélations, ..), Data Analytics, DATA Pumping (ETL)
- ▶ DATA Systems
- ▶ DATA mash up
- ▶ DATA SCIENCE

DATA en suffixe :

- ▶ Linked DATA, Web DATA (DBpedia, Web Sémantique)
- ▶ Meta DATA
- ▶ Open DATA
- ▶ Smart DATA
- ▶ BIG DATA

« DATA » - Structure

Donnée structurée (Structured data)

- ▶ La structure des données est bien définie
- ▶ **Exemple :** Base de données relationnelle, CSV, etc.

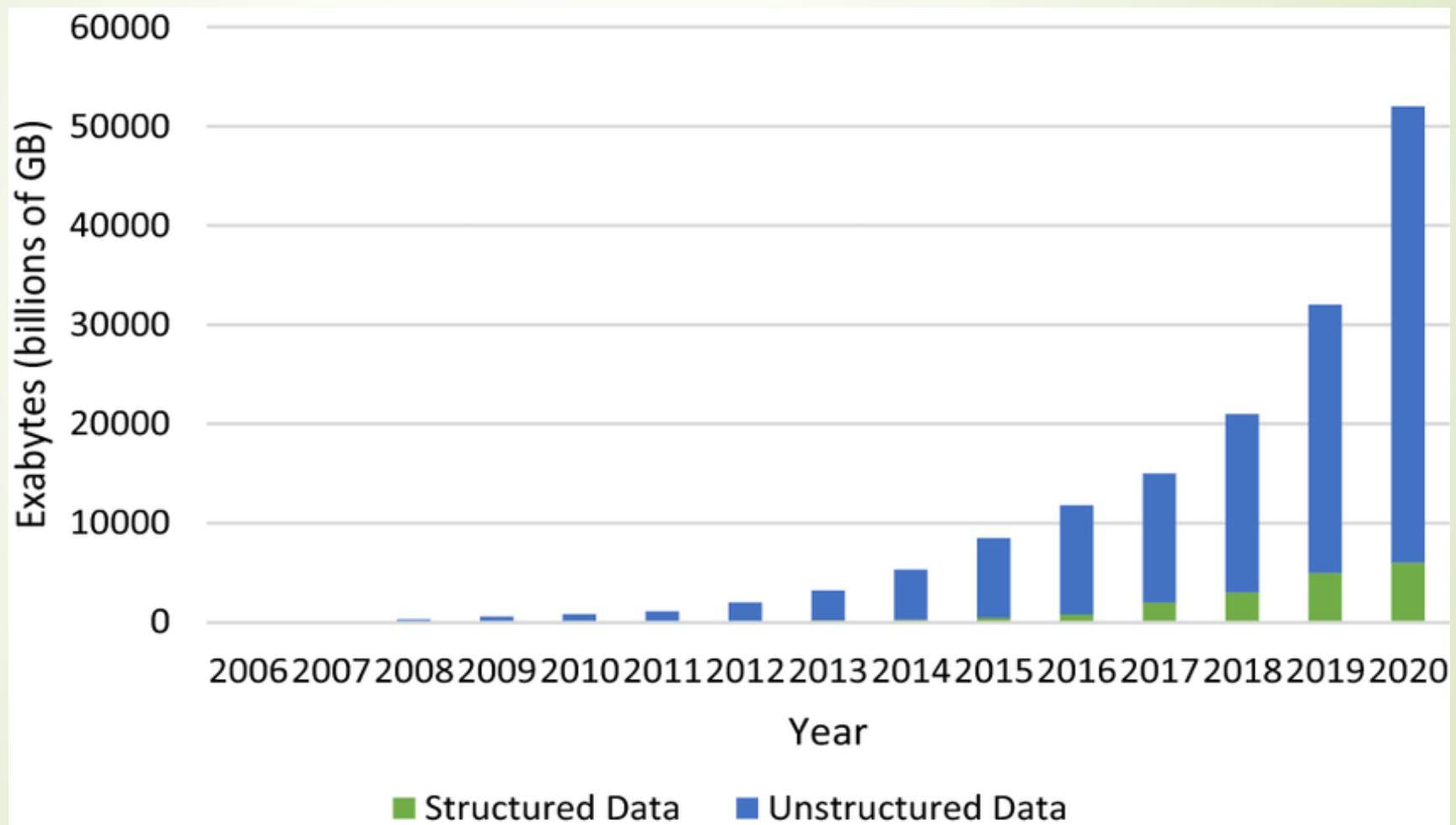
Donnée semi-structurée (Semi-structured Data)

- ▶ Contient des balises sémantiques, mais n'est pas conforme à la structure associée aux bases de données relationnelles typiques.
- ▶ même classe → attributs différents
- ▶ **Exemple :** XML, JSON, etc.

Donnée non structurée(Unstructured data)

- ▶ Les données ne suivent aucun schéma prédéfini
- ▶ **Exemple :** text, commentaires, données multimedia, etc.

Données non structurées : évolution



« DATA » - unité

International System of Units

Kilobyte	Kb	10^3
Megabyte	Mb	10^6
Gegabyte	Gb	10^9
Terabyte	Tb	10^{12}
Petabyte	Pb	10^{15}
Exabyte	Eb	10^{18}
Zettabyte	Zb	10^{21}
Yottabyte	Yb	10^{24}

Unité de mesure de base	Une page de texte	Un morceau de musique	Un film de 2 heures	6 millions de livres	Une pile de DVD de la hauteur de la tour Montparnasse	Toutes les informations produites en 2003	La totalité des données enregistrées en 2011	La NSA se dote en 2013 d'un datacenter de 300000 m ²
1	30 ko	5 Mo	1 Go	1 To	1 Po	5 Eo	1,8 Zo	1 Yo
Octet	kilo-octet ko	Mégo-octet Mo	Giga-octet Go	Téraoctet To	Pétaoctet Po	Exaoctet Eo	Zettaoctet Zo	Yottaoctet Yo
1000 0	1000 ko	1000 Mo	1000 Go	1000 To	1000 Po	1000 Eo	1000 Zo	1000 Zo
Unité	kilo	Méga	Giga	Téra	Péta	Exa	Zetta	Yotta

(Source : http://sciencesphysiques04.esy.es/documents/octets.htm#_Toc381598517)

Data at Rest

- ▶ Fait référence aux données stockées dans des systèmes de destination stables
- ▶ Les **données au repos** sont souvent définies comme des données qui ne sont pas utilisées
 - Données stockées dans une base de données en ligne
 - Données stockées sur disque
 - Données stockées extraits de bases de données en ligne ou hors ligne
 - Sauvegardes transférées sur disque
 - Les archives
- ▶ Les données sont un **instantané des informations** collectées et stockées, prêtes à être **analysées** pour la prise de décision



File Servers &
Network Shares



Document
Mgmt Systems



External
Storage



Databases



Endpoint,
laptops, PCs



Mobile
Devices



Cloud Storage

Data in motion

- ▶ **Flux de données** circulant sur tout type de réseau
- ▶ Données se déplaçant activement sur Internet
- ▶ Les **données en mouvement** sont le processus d'**analyse** des données à la volée **sans les stocker**

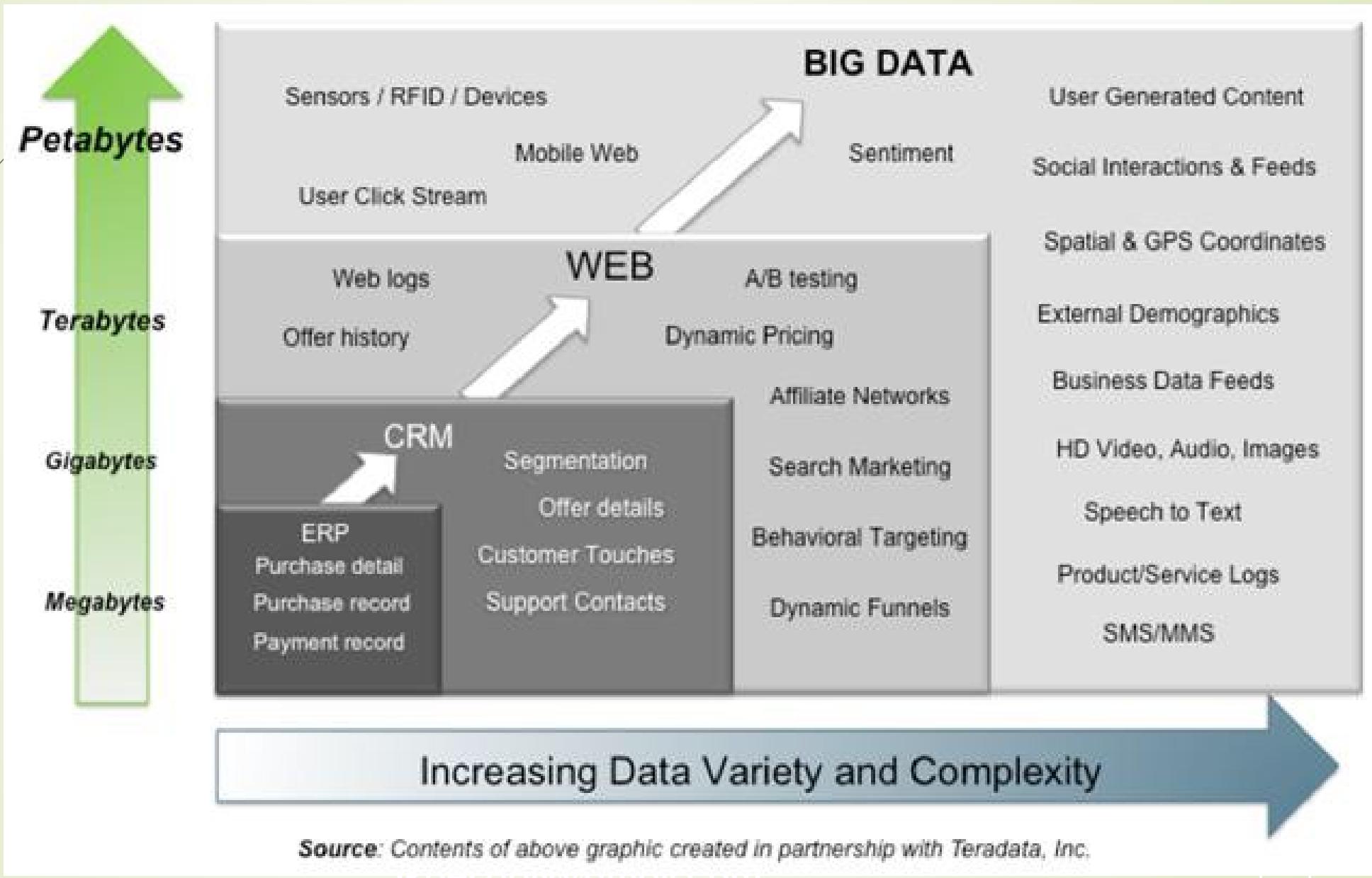


Les facettes des données

En science des données et Big Data, on rencontrera de nombreux types de données différents, et chacun d'entre eux nécessite des outils et des techniques différents. Les principales catégories de données sont les suivantes :

- ▶ Structured
- ▶ Unstructured
- ▶ Natural language
- ▶ Machine-generated
- ▶ Graph-based
- ▶ Audio, video, and image
- ▶ Streaming

Cielen, D., Meysman, A. D. B., & Ali, M. (2016). Introducing data science: Big data, machinelearning, and more, using Python tools . Shelter Island, NY: Manning Publications, pp. 4-8.



Big Data - Définition

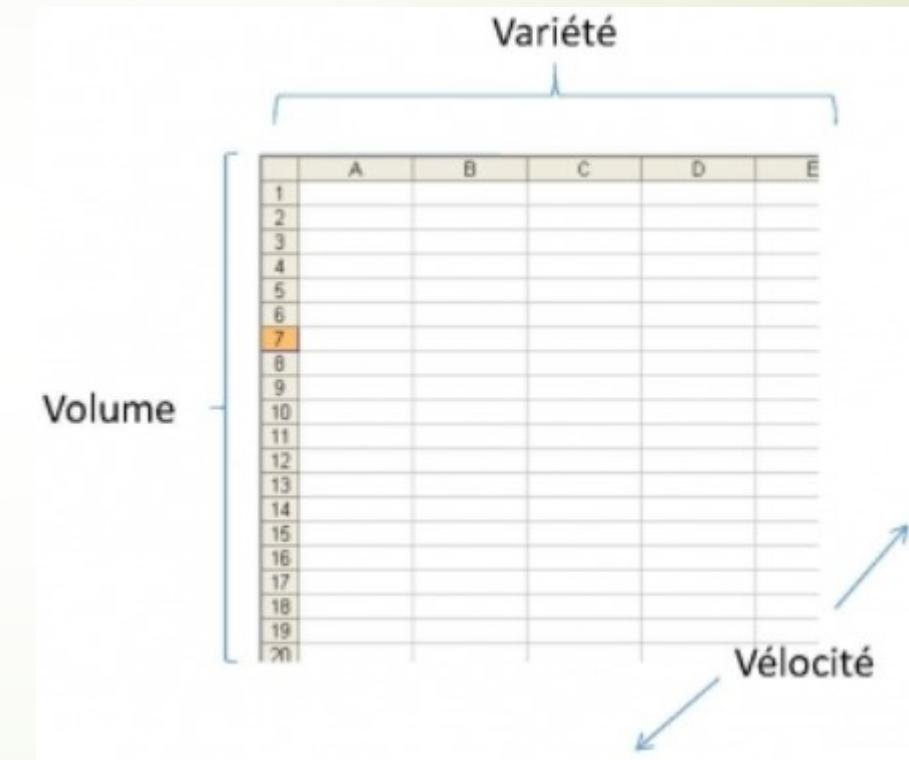
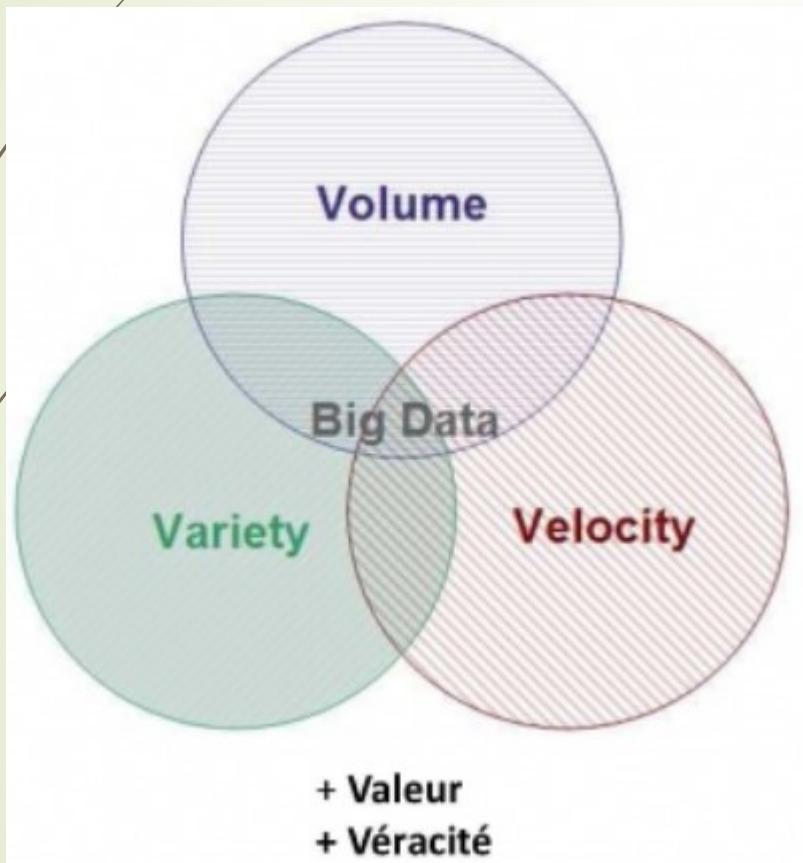
- **Grosses données**
- **Méga données**
- **Données massives**
- Ensembles de **données** qui deviennent tellement **volumineux** qu'ils en sont **difficiles** à travailler avec des **outils classiques** de gestion de base de données ou de gestion de l'information. On parle aussi de « **datamasse** ».
- « Le Big Data n'est pas qu'une technologie, mais bien une nouvelle structure d'information et de management. C'est donc une nouvelle façon d'interagir avec la réalité. » **[Gilles Babinet]**

Big Data - Définition

► “Le Big Data (ou mégadonnées) représente les collections de données caractérisées par un **volume**, une **vélocité** et une **variété** si grands que leur transformation en **valeur** utilisable requiert l'utilisation de technologies et de méthodes analytiques spécifiques.”

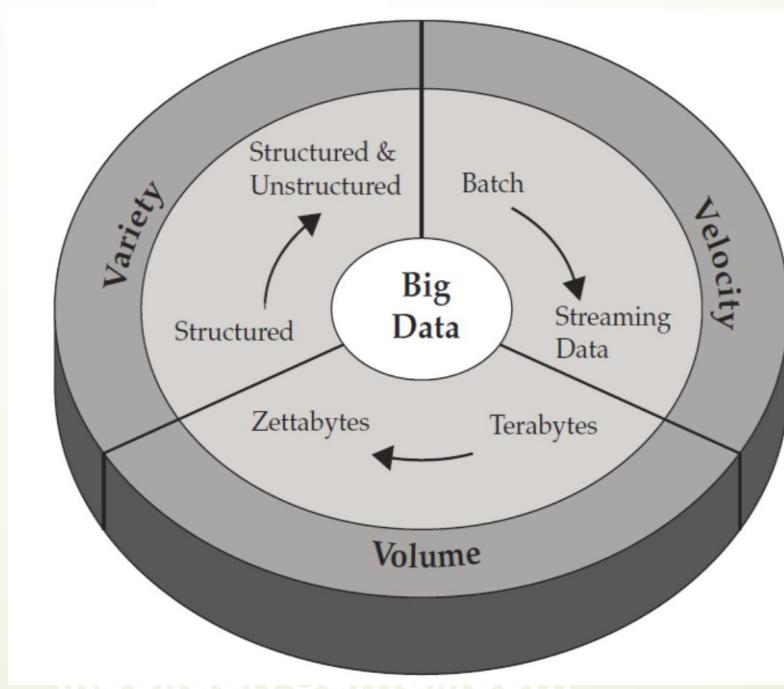


3Vs

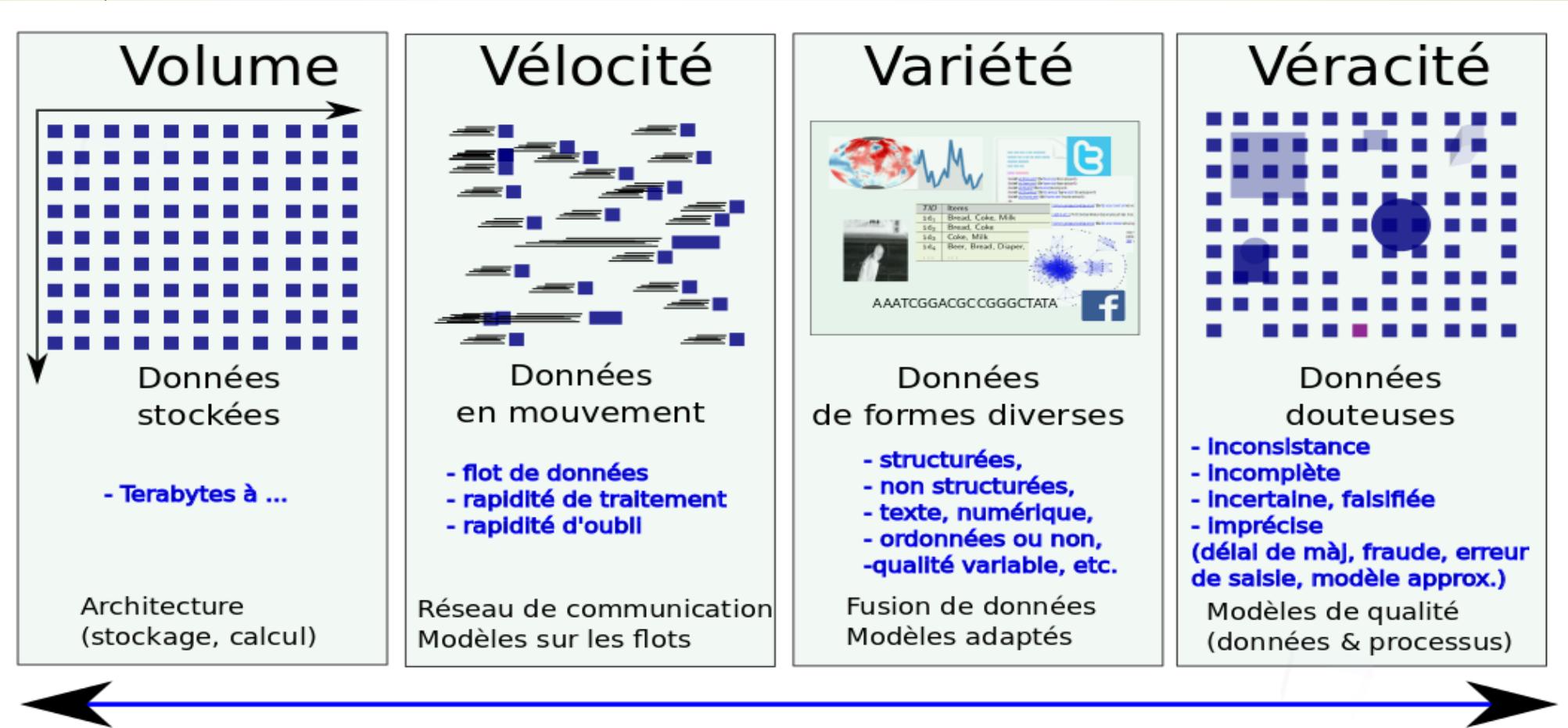


3Vs

- ▶ **Volume** - pas d'échantillonnage, on observe et mesure tout
- ▶ **Vélocité** - les données et les résultats sont souvent disponibles en temps réel
- ▶ **Variété** - puise dans les données textuelles, les photos, audio / vidéo et complète généralement les pièces manquantes en fusionnant plusieurs sources



4Vs



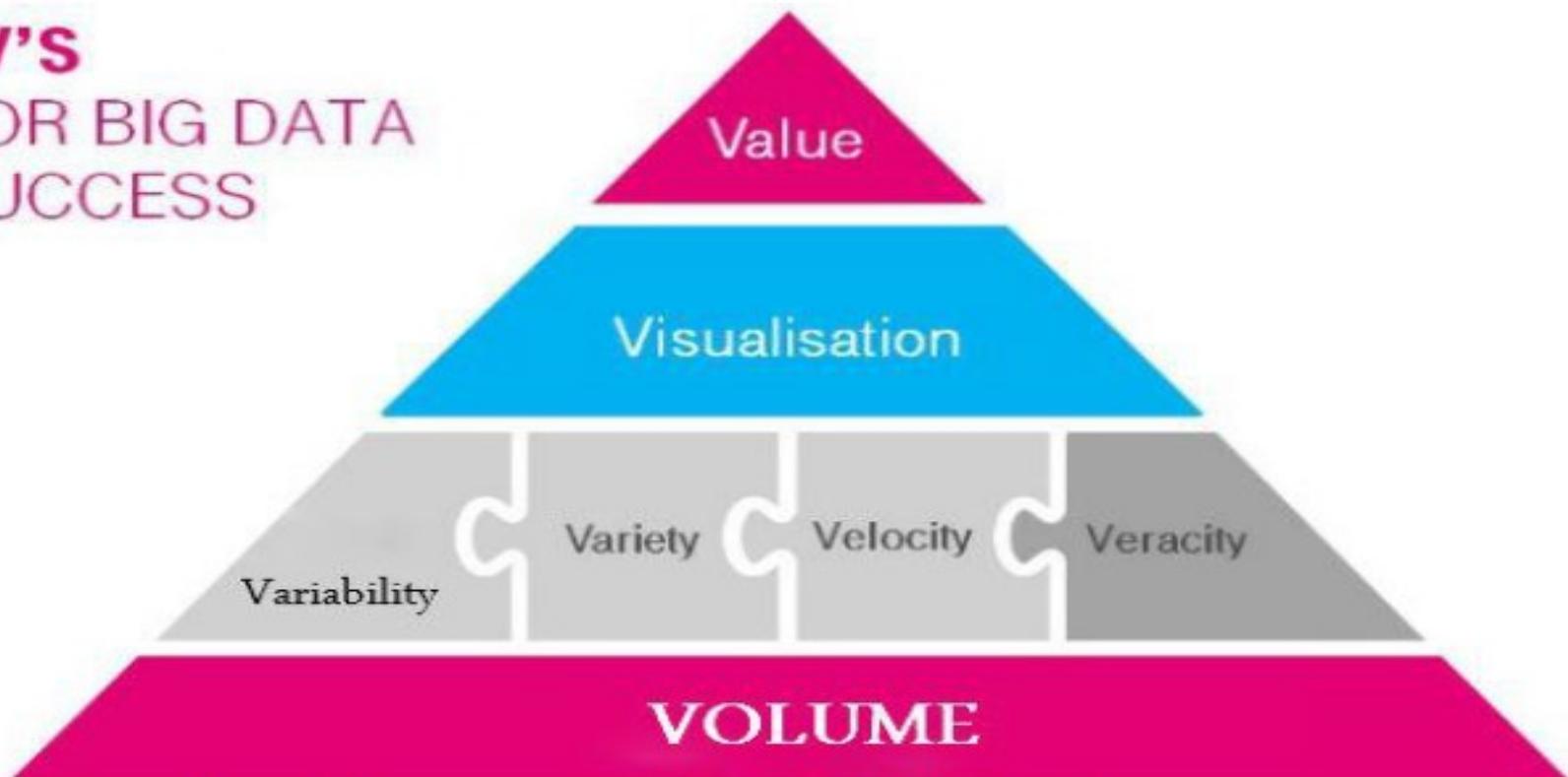
4Vs et +

- **Volume** : la quantité d'informations, trop volumineuse pour être acquise, stockée, traitée, analysée et diffusée par des outils standards.
- **Variété** : l'hétérogénéité des formats, de types, et de qualité des informations.
- **Vélocité** : l'aspect dynamique et/ou temporel des données, à leur délai d'actualisation et d'analyse.

« des données qui sont trop volumineuses ou ayant une arrivée trop rapide ou une variété trop grande pour permettre de les ranger directement dans des bases de données traditionnelles ou de les traiter par les algorithmes actuels » [J. Mothe].
- **Valeur** : la potentialité des données, en particulier en termes économiques (usage).
- **Véracité ou Validité** : la qualité des données et/ou aux problèmes éthiques liés à leur utilisation.

4Vs et +

**7V'S
FOR BIG DATA
SUCCESS**



<http://blogs.systweak.com/2017/03/big-data-vs-represents-characteristics-or-challenges-of-big-data>

4Vs et +

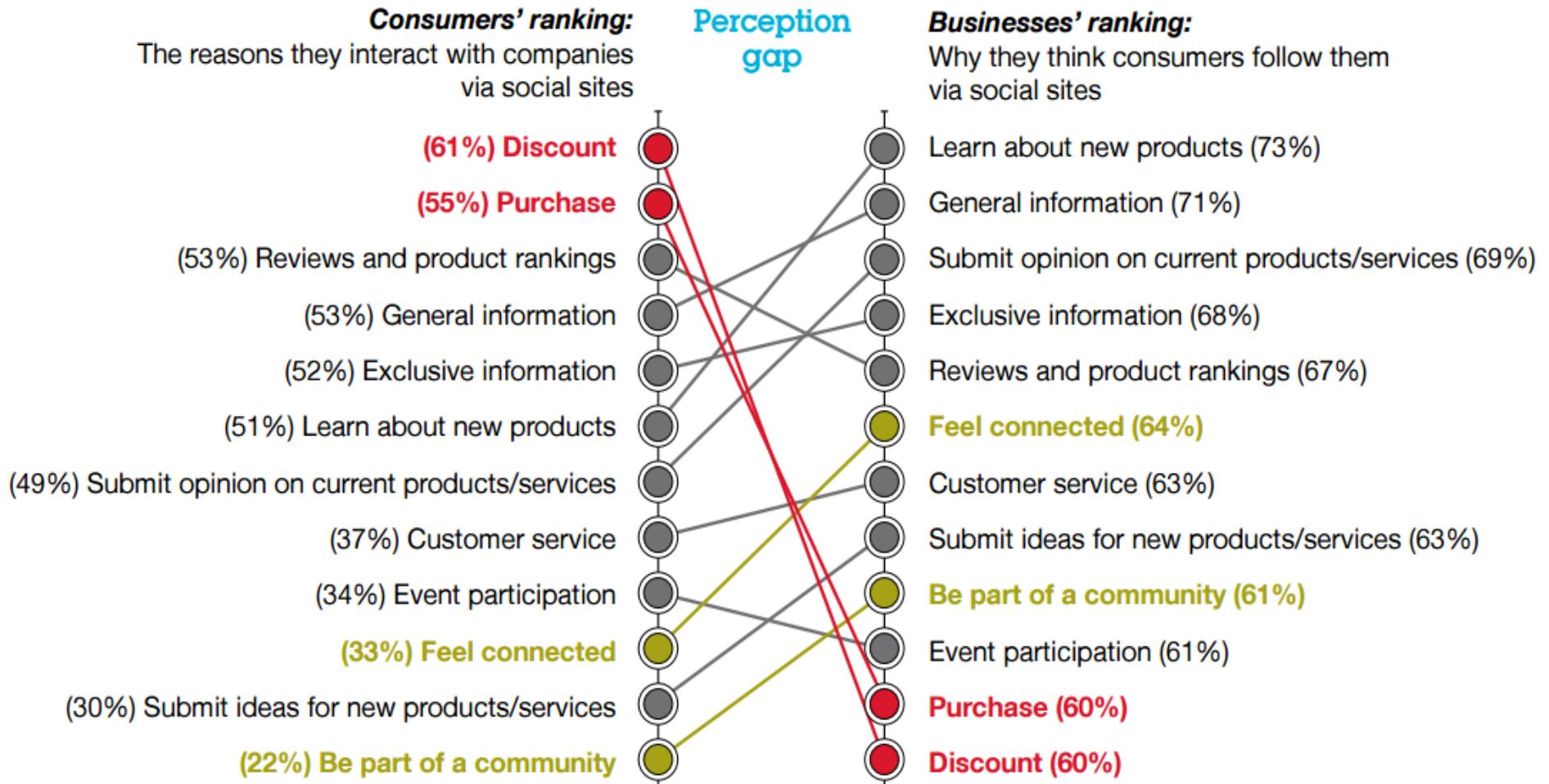
- Volume - how much data is there?
- Velocity - how quickly is the data being created, moved, or accessed?
- Variety - how many different types of sources are there?
- Veracity - can we trust the data?
- Validity - is the data accurate and correct?
- Viability - is the data relevant to the use case at hand?
- Volatility - how often does the data change?
- Vulnerability - can we keep the data secure?
- Visualization - how can the data be presented to the user?
- Value - can this data produce a meaningful return on investment?

<https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics>

Motivation – Data partout

- ▶ Internet www.internetworkstats.com
 - ▶ 3,5 milliard internautes
 - ▶ Équivalent à 46% de la population
- ▶ Facebook www.brandwatch.com
 - ▶ 1,7 milliard utilisateurs (6 nouveaux profiles par seconde)
 - ▶ Hive Data Warehouse stocke plus de 300 Po
 - ▶ Un taux journalier entrant d'environ 600 To
- ▶ Téléphonie Mobile www.statista.com
 - ▶ 4,61 milliard utilisateurs
 - ▶ >1,5 milliard de smartphones vendus
 - ▶ 7 Eo trafic de données mobiles

Motivation – Perception Gap



Note: Consumer: N=1056; Business: Learn N=333, General info N=336, Submit opinion N=334, Exclusive info N=333, Reviews/rankings N=333, Feel connected N=331, Customer service N=331, Submit ideas N=332, Community N=329, Event N=332, Purchase N=334, Discounts N=331.
Source: IBM Institute for Business Value analysis. CRM Study 2011.

Motivation – Enquête IBM

1 in 3 Les chefs d'entreprise prennent leurs décisions en se basant sur des informations auxquelles ils ne font pas confiance ou ne les ont pas.

1 in 2 Les chefs d'entreprise disent qu'ils n'ont pas accès aux informations dont ils ont besoin pour faire leur travail.

80% des données du monde sont non structurées

Motivation – Data partout

90%

of the world's
data was
created in the
last two years



80%

of the
world's data
today is
unstructured



20%

of available data
can be processed
by traditional
systems



1 in 2

business leaders
don't have access to
data they need

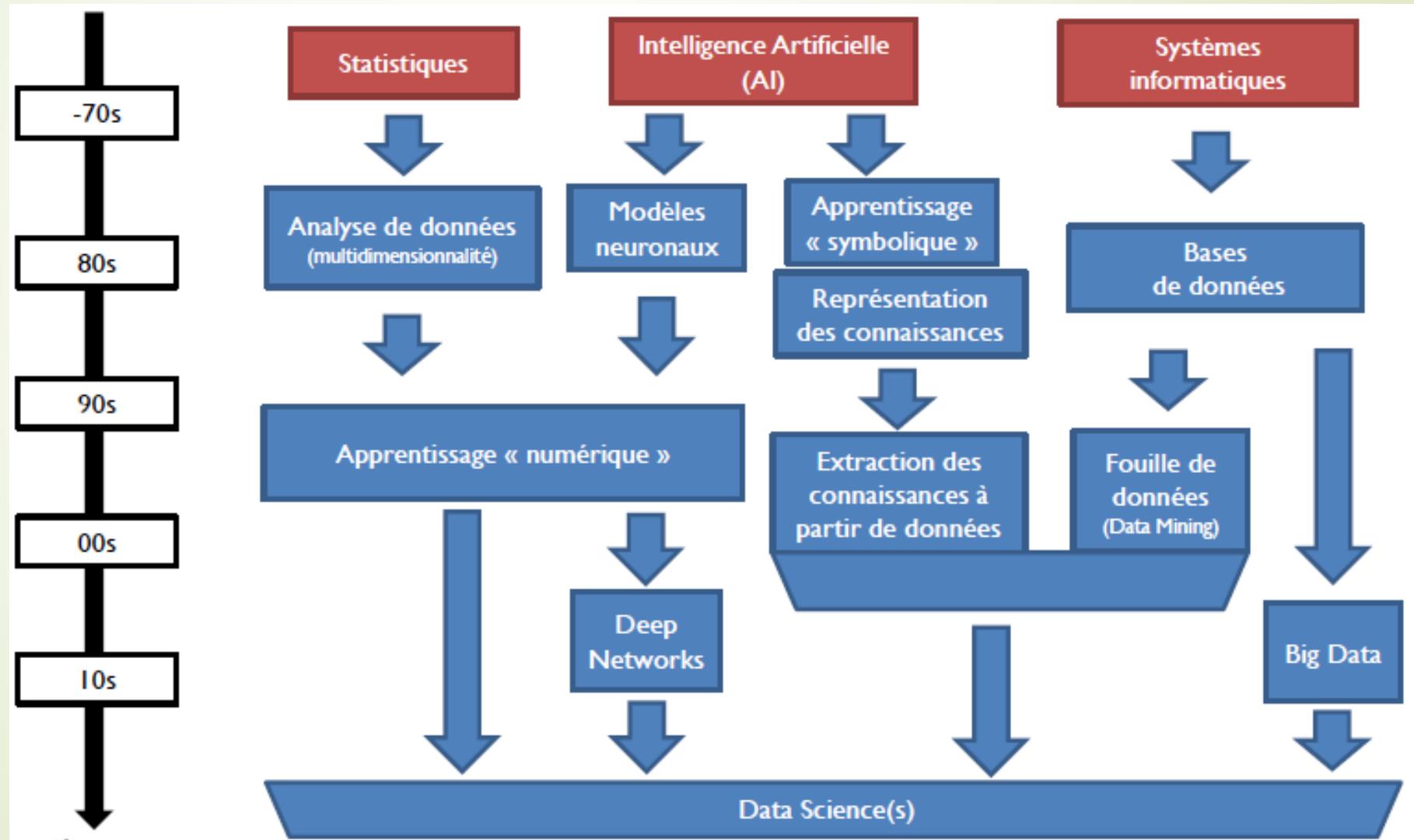
83%

of CIO's cited BI and
analytics as part of their
visionary plan

5.4X

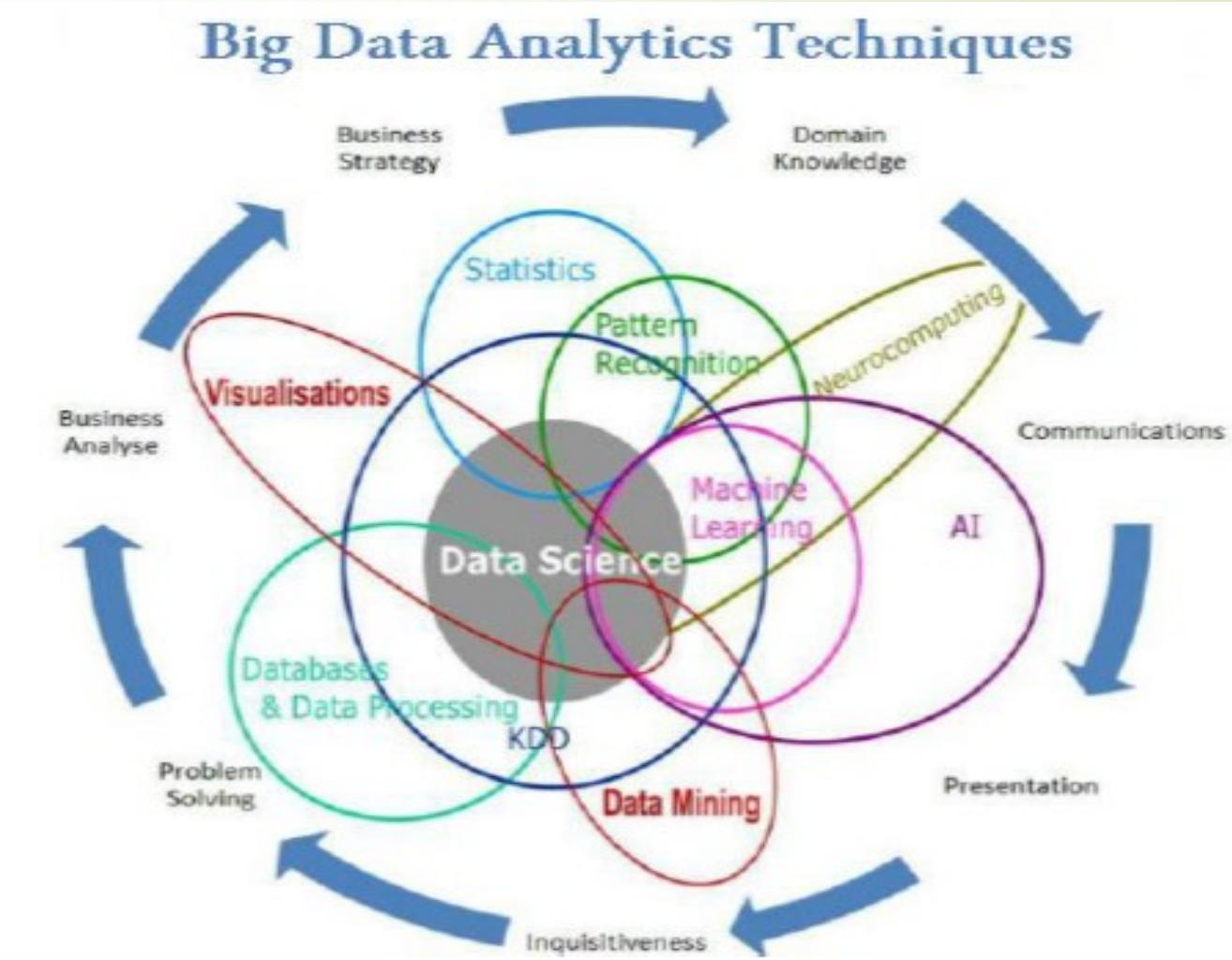
more likely that top
performers use
business analytics

Les origines scientifiques de la science des données



Big Data Analytics - Analyse des Données Massives

- ▶ Ensemble des processus et technologies permettant de stocker, traiter et analyser des données massives.
- ▶ → 26 Big Data Analytic techniques



Enjeux

- ▶ Sécurité:
- ▶ Hyper-surveillance
- ▶ Fuite de données
- ▶ Piratage



Enjeux

► Vie privée et données personnelles



► Anonymisation, conservation (archivage) et droit à l'oubli

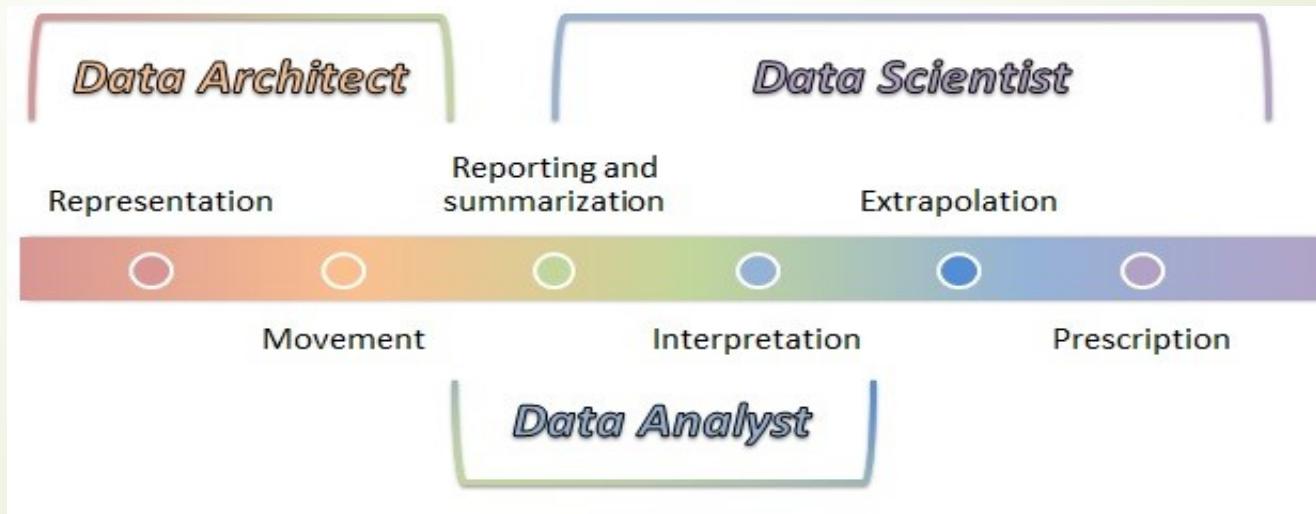


► Environnement

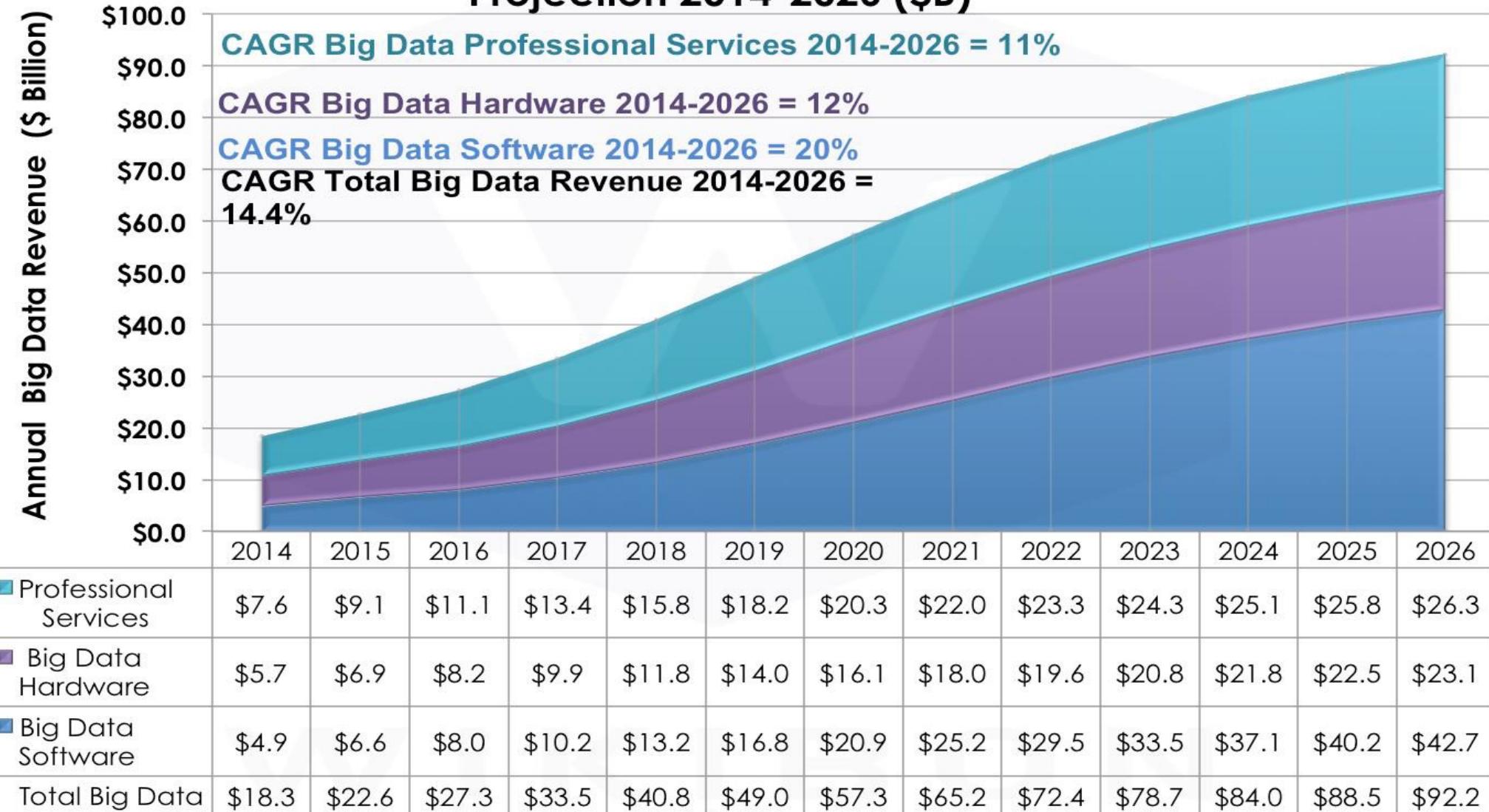


Nouveaux challenges, nouveaux besoins

- ▶ **Data Architect** : responsable des infrastructures
- ▶ **Data Scientist** : informaticien spécialisé dans l'analyse des données
- ▶ **Data Analyst** : responsable des opérations de bases de données et appui analytique à l'exploration de données

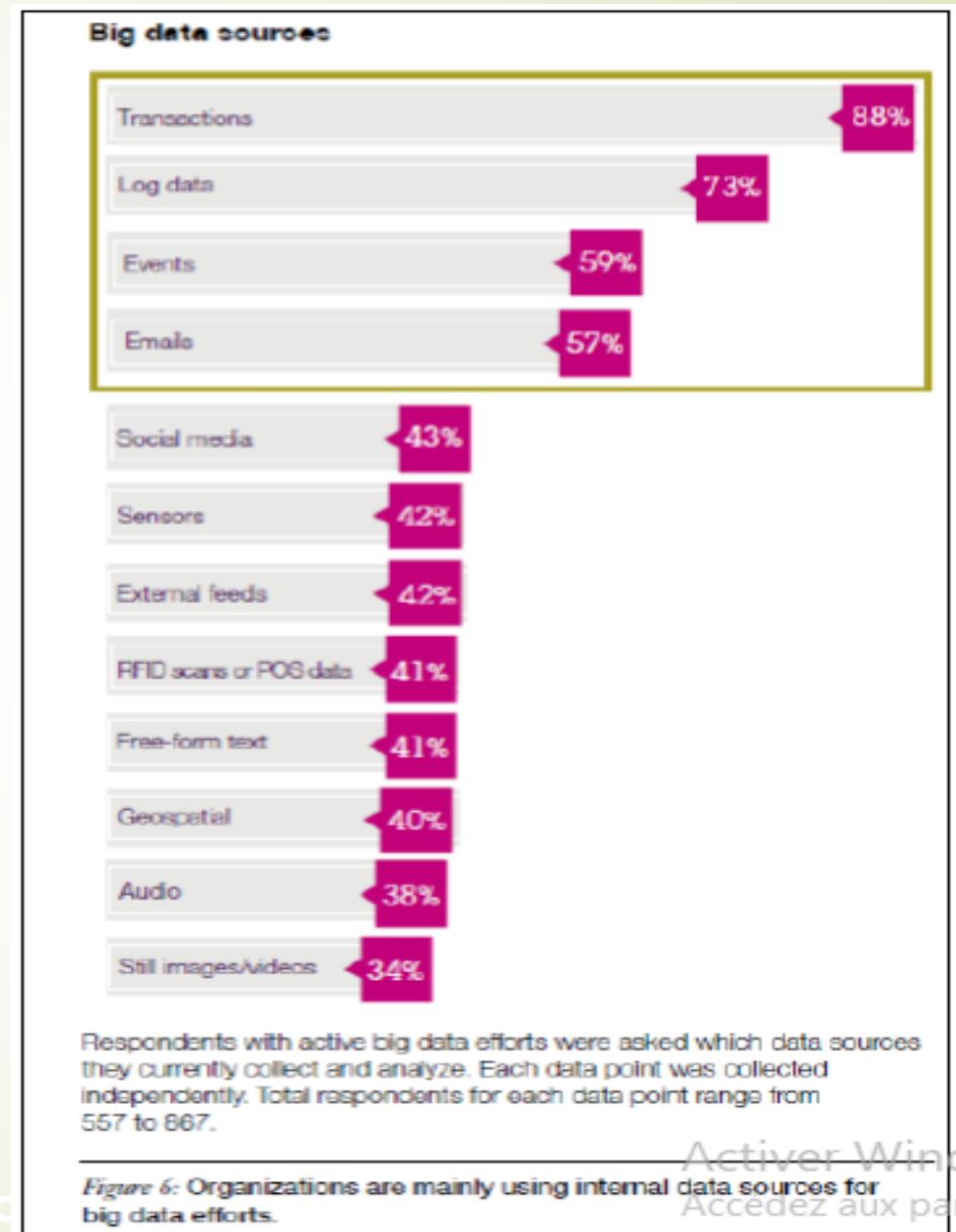
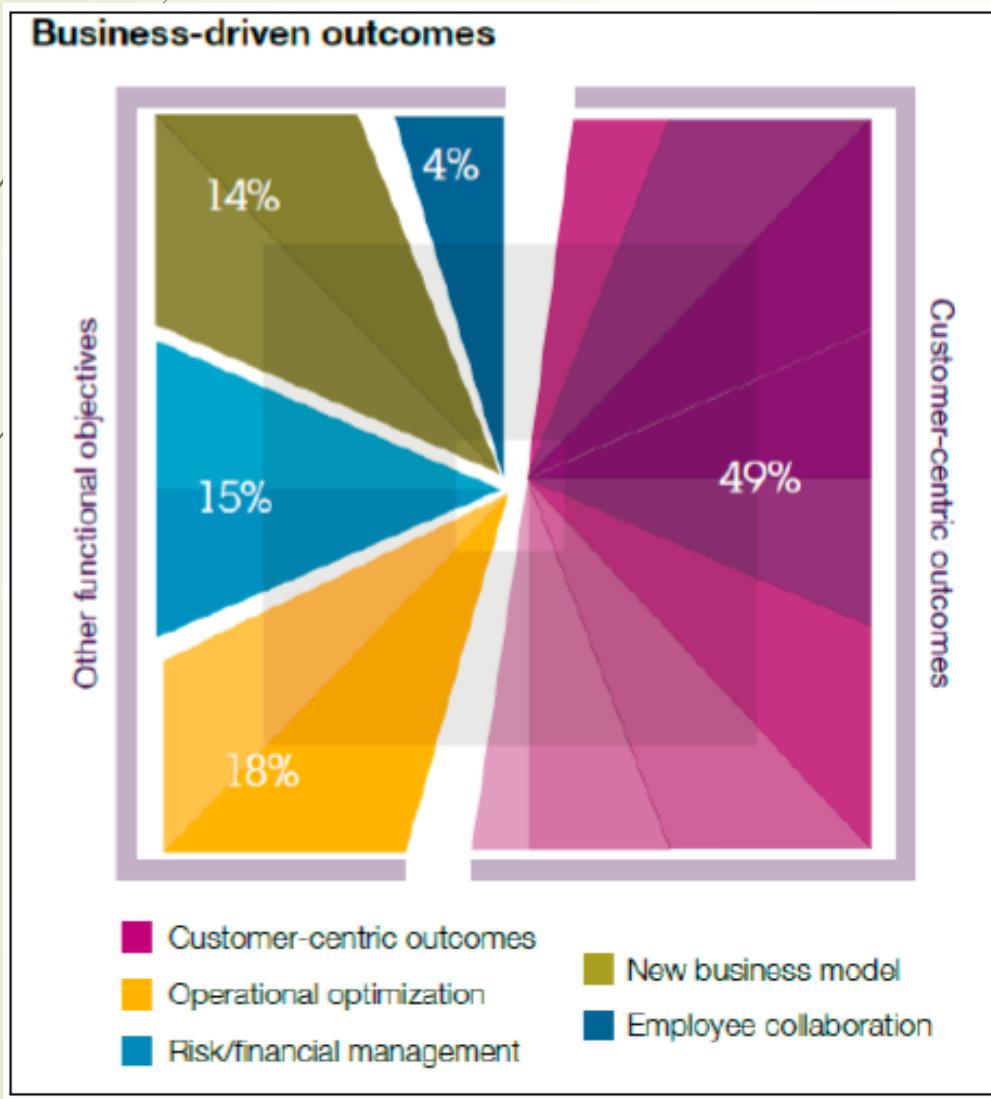


Wikibon Big Data Software, Hardware & Professional Services Projection 2014-2026 (\$B)



Source: © Wikibon Big Data Project, 2016

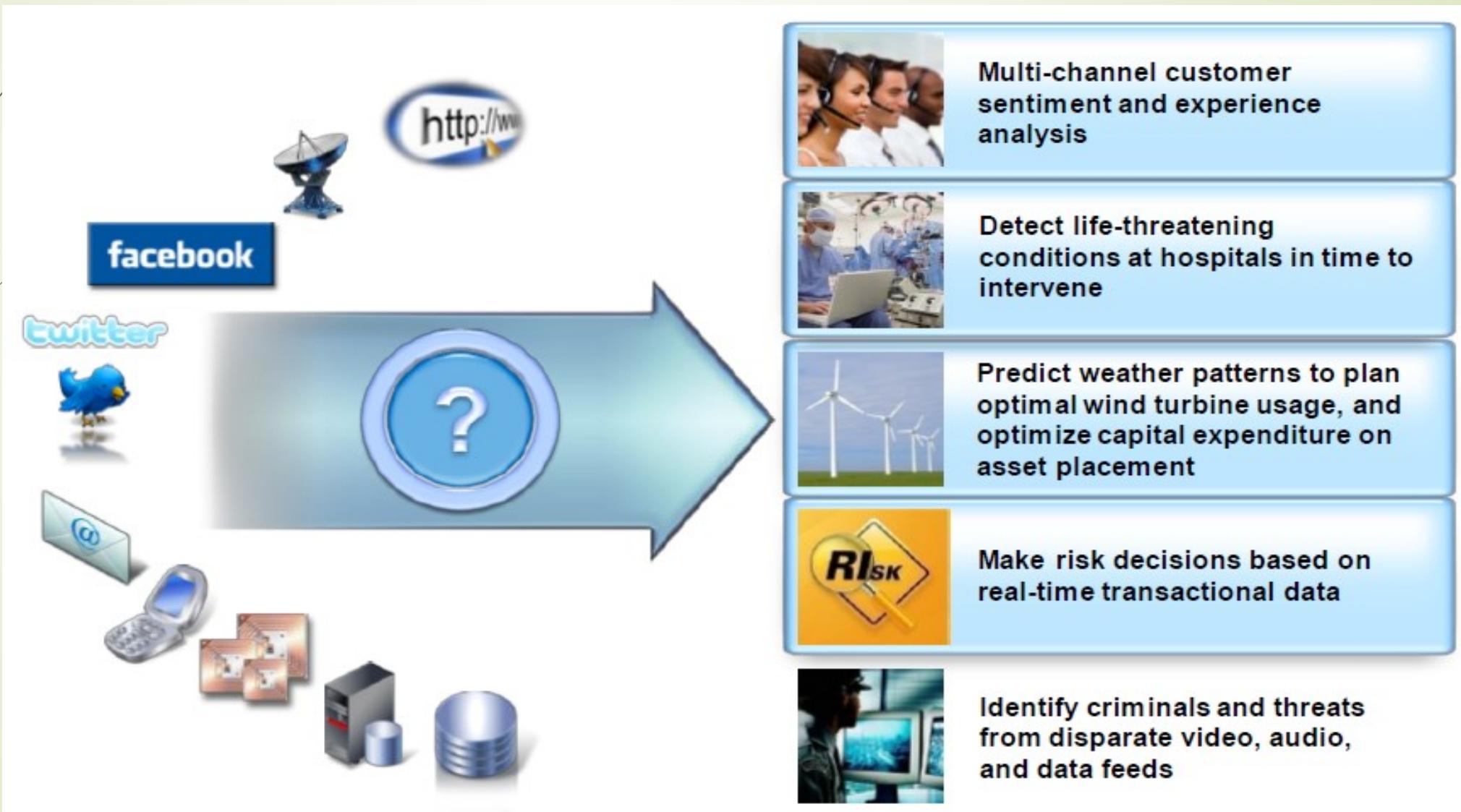
Outcomes & Sources



Cas d'utilisation de Big Data



Cas d'utilisation de Big Data



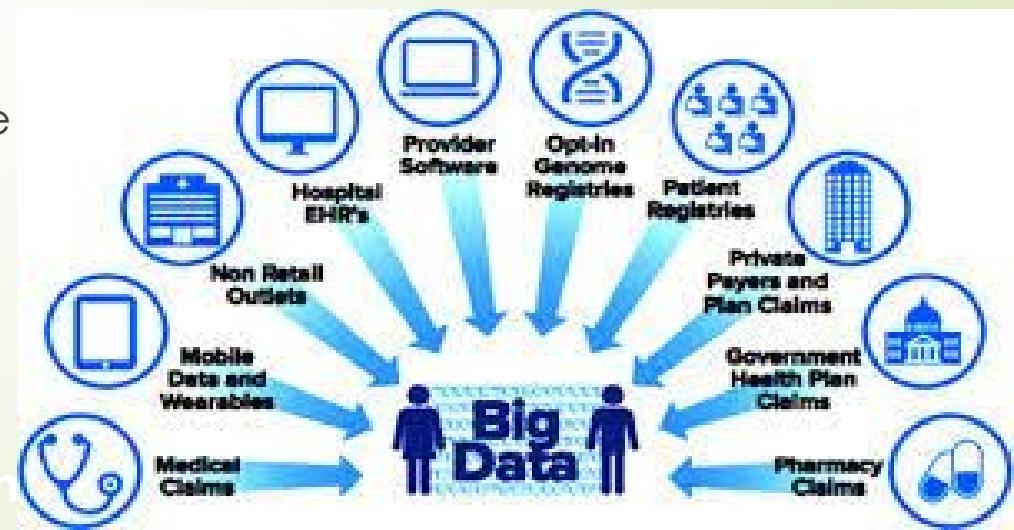
Santé et sciences de la vie

► Problème :

De grandes quantités d'informations en **temps réel** commencent à provenir des **dispositifs de surveillance** sans fil que les patients postopératoires et ceux atteints de maladies chroniques portent à la maison et dans leur vie quotidienne.

► Comment **l'analyse de données** volumineuses peut aider :

- Alerte précoce en cas d'épidémie
- Unité de soins intensifs et télésurveillance



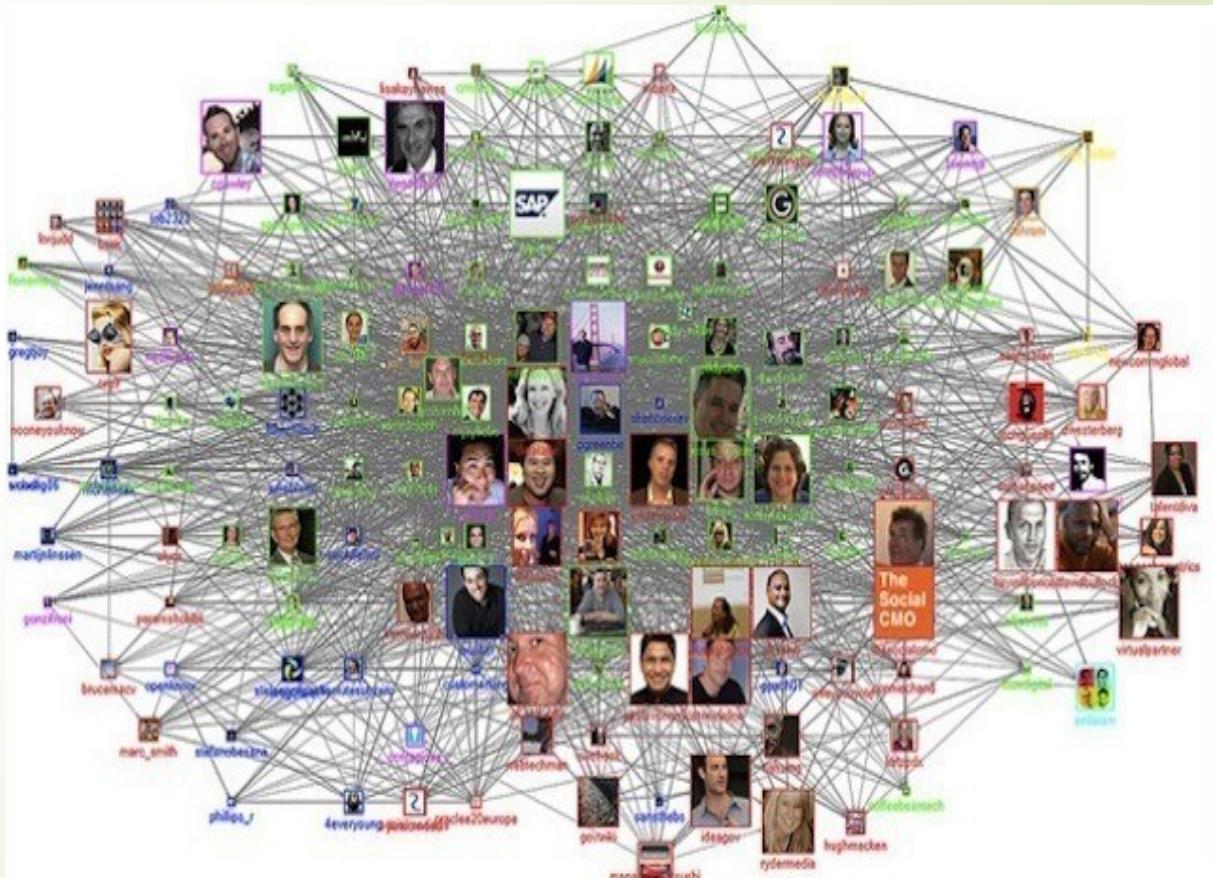
Finance

- ▶ **Problème** : Gérer les plusieurs pétaoctets de données qui augmentent de 40 à 100 % par an sous une pression croissante pour **prévenir les fraudes** et les plaintes auprès des régulateurs
- ▶ Comment **l'analyse de données volumineuses** peut aider :
 - ▶ Détection de fraude
 - ▶ Émission de crédit
 - ▶ Gestion des risques
 - ▶ Vision à 360° du client



Réseaux sociaux

- ▶ Analyse de connectivité
 - ▶ Analyse de communauté



Téléphonie mobile

- ▶ **Marketing: Profilage et prédition**
- ▶ **Besoin** : Comment anticiper à éventuel non-renouvellement d'un abonnement grâce à l'étude des sites concurrents visités par leurs clients?
- ▶ **Eléments** : Big data, échanges sur les réseaux sociaux, monitoring des conversations
- ▶ **Résultats** :
 - ✓ Plus de 81% des profils concernés par une potentielle résiliation détectés
 - ✓ 75% des clients conservés

Pharmacie

- ▶ **Santé: Analyse de données**
- ▶ **Besoin :** Comment réduire la durée d'un programme d'optimisation de molécules?
- ▶ **Eléments :** Big data, 400 composés pendant les 6 premiers mois du programme, 4 profils de composés présentant les meilleures propriétés
- ▶ **Résultat:**
 - ✓ temps de recherche divisé par 2

Smartgrid

- ▶ **Energie:** Optimisation et régulation de la distribution d'électricité
- ▶ **Besoin :** Comment lutter contre les fraudes?
- ▶ **Eléments :** Big data, sourcing des données avec des compteurs intelligents, échange avec les équipements de distribution hétérogènes
- ▶ **Résultats :**
 - ✓ 30 millions de compteurs déployés
 - ✓ Fraude réduite de 75%

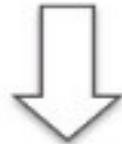
Approche traditionnelle vs Approche Big Data

Approche Traditionnelle

Analyse Structurée et Répétée

Responsables Métier

Déterminent quelles questions poser



Responsables IT

Structurent les données pour répondre à ces questions

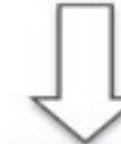


Approche Big Data

Analyse Itérative et Exploratoire

Responsables IT

Fournissent une plateforme pour permettre la découverte créative



Responsables Métier

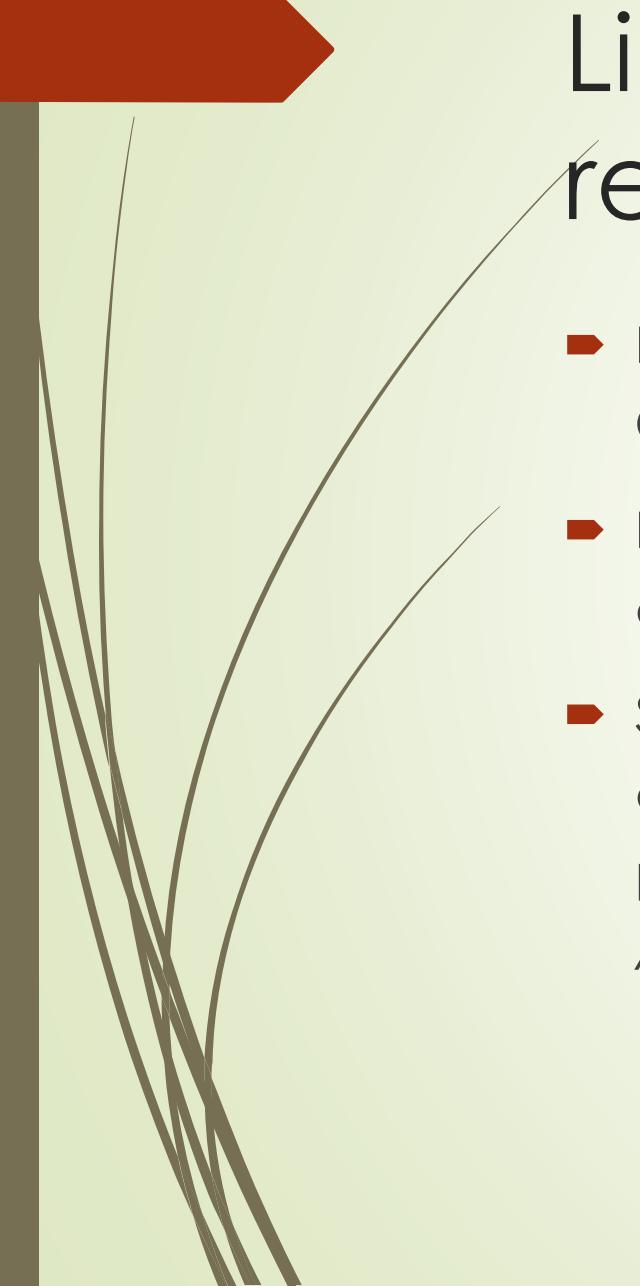
Explorent la plateforme pour déterminer quelles questions poser





Problématiques

- 
- ▶ **Volume → Stockage**
 - ▶ **Vélocité → Traitement**
 - ▶ **Variété → Collecte**



Limites des bases de données relationnelles et *Cloud Computing*

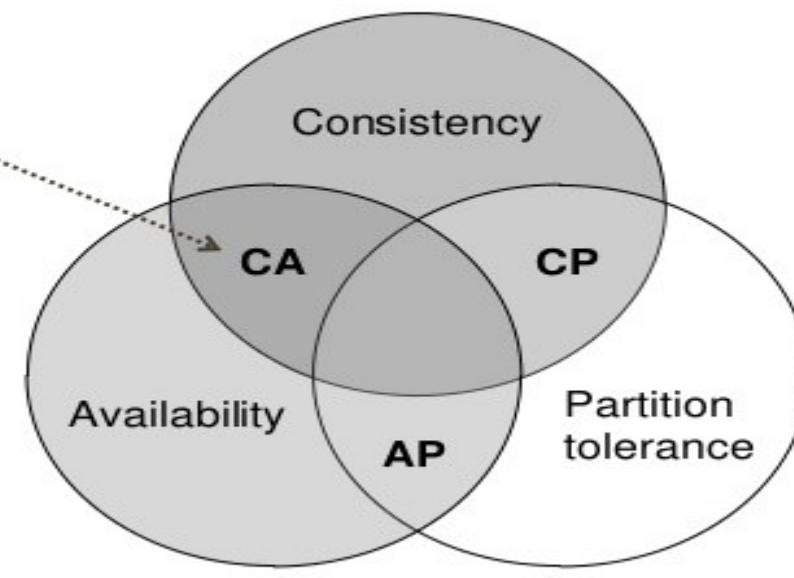
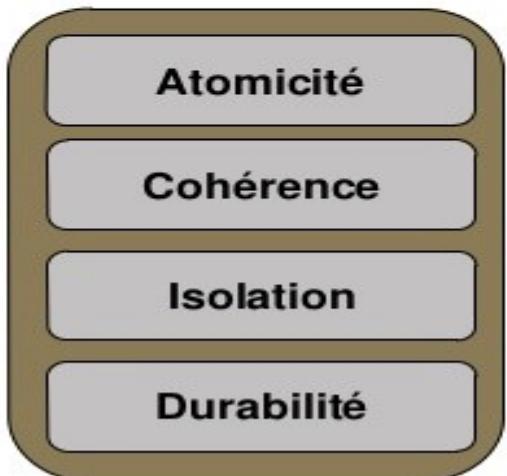
- ▶ **BD** : garantissent le maintien des propriétés **ACID (Atomicité, Cohérence, Isolation et Durabilité)**.
- ▶ **BD + gros volumes de données = distribution** des données sur différents disques permettant un **parallélisme** de l'exécution des requêtes
- ▶ Stockage **distribué** (partitionné) des données sur les clusters → Système distribué ne peut assurer à la fois la **cohérence**, la **disponibilité** et la possibilité d'être **partitionné** (D'après théorème CAP «Consistency, Availability, Partition tolerance » de Brewer).

Limites des bases de données relationnelles et Cloud Computing

Propriétés ACID vs Théorème CAP

BD relationnelle

4 / 4



2 / 3

Big Data

Limites des bases de données relationnelles et Cloud Computing

- ▶ Le **nuage** (cloud) est un ensemble de matériels, de raccordements réseau et de logiciels fournissant des services que des individus et des groupements peuvent exploiter à volonté depuis n'importe où.
 - ▶ Une mutualisation de ressources hétérogènes (partage de ressources).
 - ▶ Besoins de stockage
 - ▶ Variété et vélocité
 - ▶ Le cloud étant davantage un support de **stockage** qu'une solution de gestion de données.
- 
- 

éploiement de nouveaux serveurs

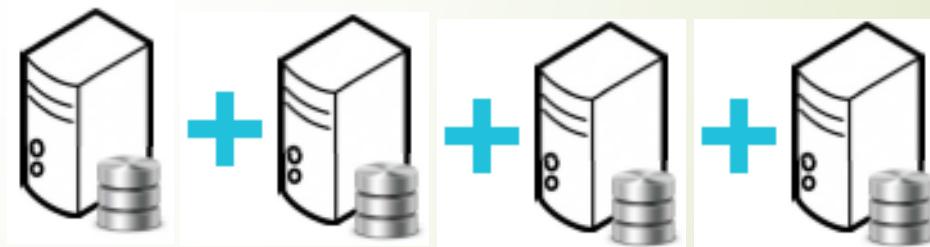
Stockage des données



Comment stocker ces données dont le volume ne cesse d'augmenter?



Partitionnement



Stockage des données

► **Théorème CAP** : Seules deux des trois propriétés suivantes peuvent être assurées :

- Consistency
- Availability
- Partitionning

- ✓ Tolérer une certaine perte de la consistance au profit du partitionnement et de la haute disponibilité
- ✓ Remplacer les propriétés ACID par les propriétés BASE
 - Basically Available
 - Soft State
 - Eventual Consistency



Stockage des données



→ Étant donné un système de stockage partitionné (sous forme de cluster), comment assurer :

- La répartition de charges
- La tolérance aux fautes
- La haute disponibilité

Stockage des données

Répartition de charges



- ✓ Données réparties sur l'ensemble des nœuds du cluster, selon une stratégie de partitionnement choisie (aléatoire, ordonnée..)
- ✓ Réalisation du traitement directement sur les nœuds de stockage

Stockage des données

► Tolérance aux fautes



- ✓ Duplication de toutes les données un nombre donné de fois
- ✓ Définition d'une stratégie de réPLICATION (simple, par topologie de réseau..)
- ✓ Principe de Rack Awareness

Stockage des données

Haute disponibilité



- ✓ Assurer une lecture et écriture instantanée des données
- ✓ Read and Write Anywhere
- ✓ Éviter les jointures et les transactions, tolérer les redondances
- ✓ Favoriser les traitements côté client pour décharger le système de stockage

Stockage des données

- Apparition d'autres types de systèmes de stockages:
 - Systèmes de fichiers distribués
 - Bases de données hautement distribuées (NoSQL)
 - Bases de données NewSQL

Traitement des données

► **Principes:**

- Déplacer le traitement vers les données
- Principe de **In-Memory Processing**
- Savoir être **Polyglotte**
 - Polyglot Programming: Plusieurs langages et paradigmes de programmation dans une seule application
 - Polyglot Persistence: Plusieurs technologies de stockage dans une seule application

Traitement des données

► Types de traitement:

- Batch Processing
- Stream Processing
- Micro-Batch Processing
- Real-time Processing

Hadoop



- Le traitement d'aussi grandes quantités de données impose des méthodes particulières :
 - Répartir les données sur plusieurs machines (de 5 à plusieurs millions d'ordinateurs):
 - système de fichiers spécial permettant de ne voir qu'un seul espace pouvant contenir des fichiers gigantesques et/ou très nombreux,
 - bases de données spécifiques,
 - algorithmes faciles à écrire,
 - exécutions faciles à paralléliser.

Hadoop



- High-Availability Distributed Object-Oriented Platform
- un **cadiciel** (framework) de référence **libre** et **opensource**
- Hadoop a été créé par **Doug Cutting** et fait partie des projets de la fondation logicielle **Apache** depuis **2009**.
- Plateforme distribuée pour le stockage et traitement de données massives

Composition:

- **HDFS** : un système de fichier distribué qui répartit les données sur de nombreuses machines,
- **YARN** : un mécanisme d'ordonnancement de programmes de type MapReduce (architecture pour un calcul parallèle de larges ensembles de données)

Hadoop

- Principe:
 - Division de données
 - Stockage des données sur un ensemble de machines (cluster)
 - Traitement des données localement (pas d'utilisation du serveur distribué pour copier les données)
- ➔ Avec la croissance des données, on a l'ajout des machines au cluster

Distributions de Hadoop

cloudera



IBM Open Platform



Amazon Elastic MapReduce

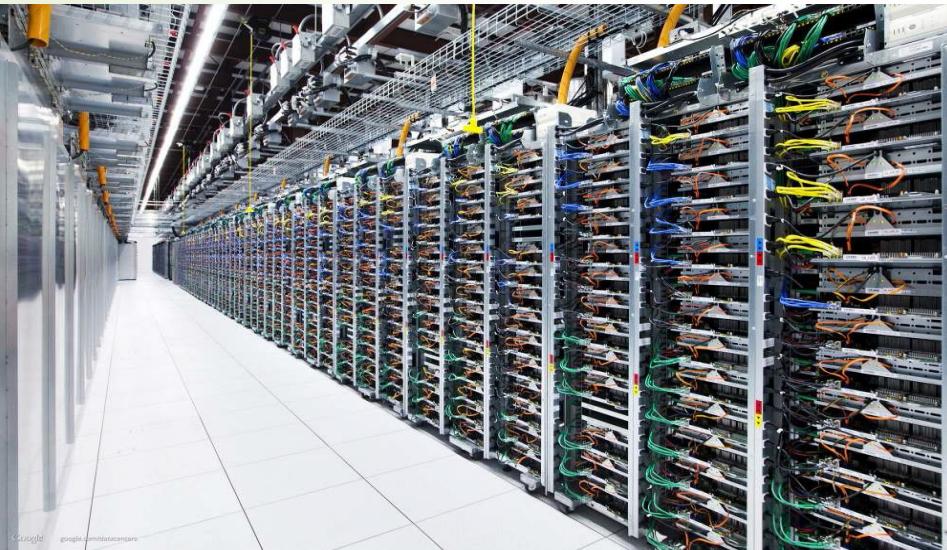
MAPR
DATA TECHNOLOGIES



Microsoft Azure's HDInsight

Data Center

- Un **centre de données** est un lieu (et un service) regroupant des équipements constituants du système d'information d'une ou plusieurs entreprise(s) (ordinateurs centraux, serveurs, équipements réseaux et de télécommunications, etc).
- Milliers d'ordinateurs connectés entre eux = un *cluster*



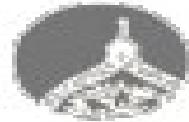
Serveur lame

- Un **serveur lame** (aussi appelé **serveur blade** ou **carte serveur** ; en anglais, **blade server**) est un serveur conçu pour un très faible encombrement. Alors qu'un **serveur en rack** n'est qu'un serveur traditionnel de taille un peu réduite, le serveur lame est beaucoup plus compact, car plusieurs composants sont enlevés.
- **Exemple:** 4 CPU multi-coeurs, 128 Go de RAM, 24 To de disques rapides, 5000€





Apache Hadoop Ecosystem



Ambari

Provisioning, Managing and Monitoring Hadoop Cluster

Scoop

Data Exchange



ZooKeeper
Coordination

Oozie

Workflow



Pig
Scripting



Mahout

Machine learning

R Connectors

Statistics



Hive
SQL Query

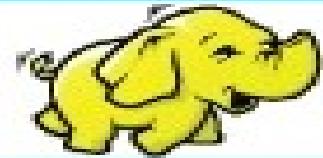
HBase

Columnar Store



Flume

Log collector



YARN Map Reduce v2
Distributed processing Framework

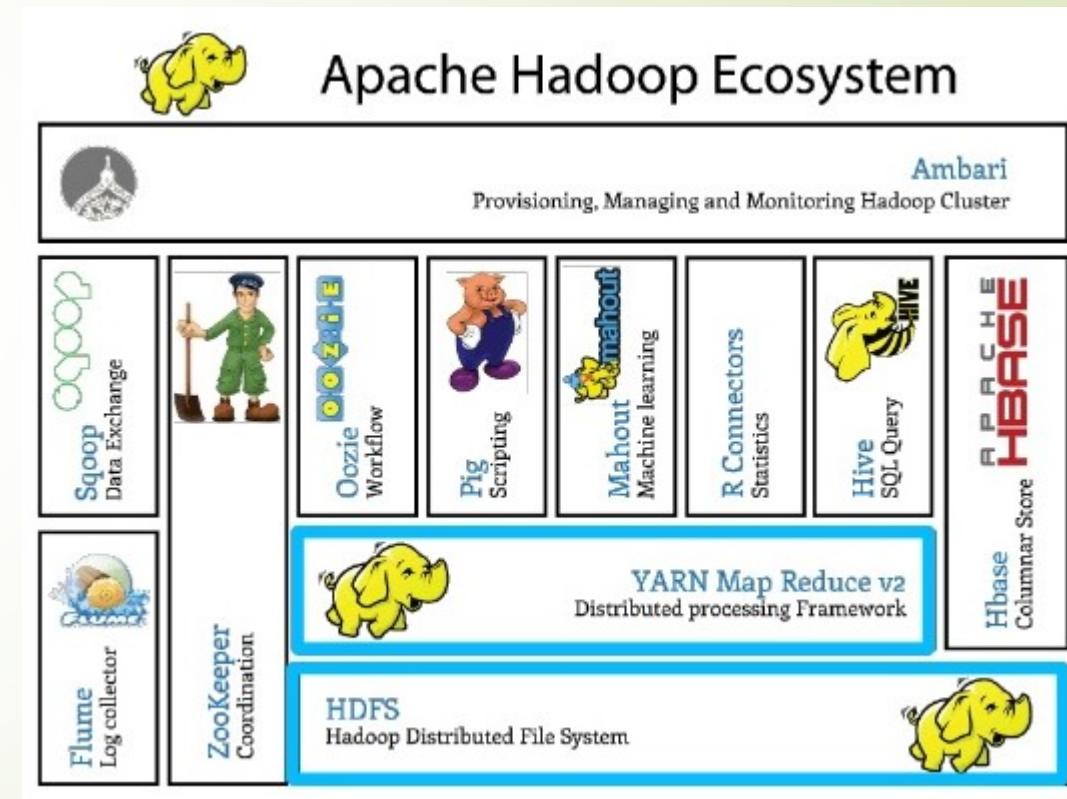
HDFS

Hadoop Distributed File System



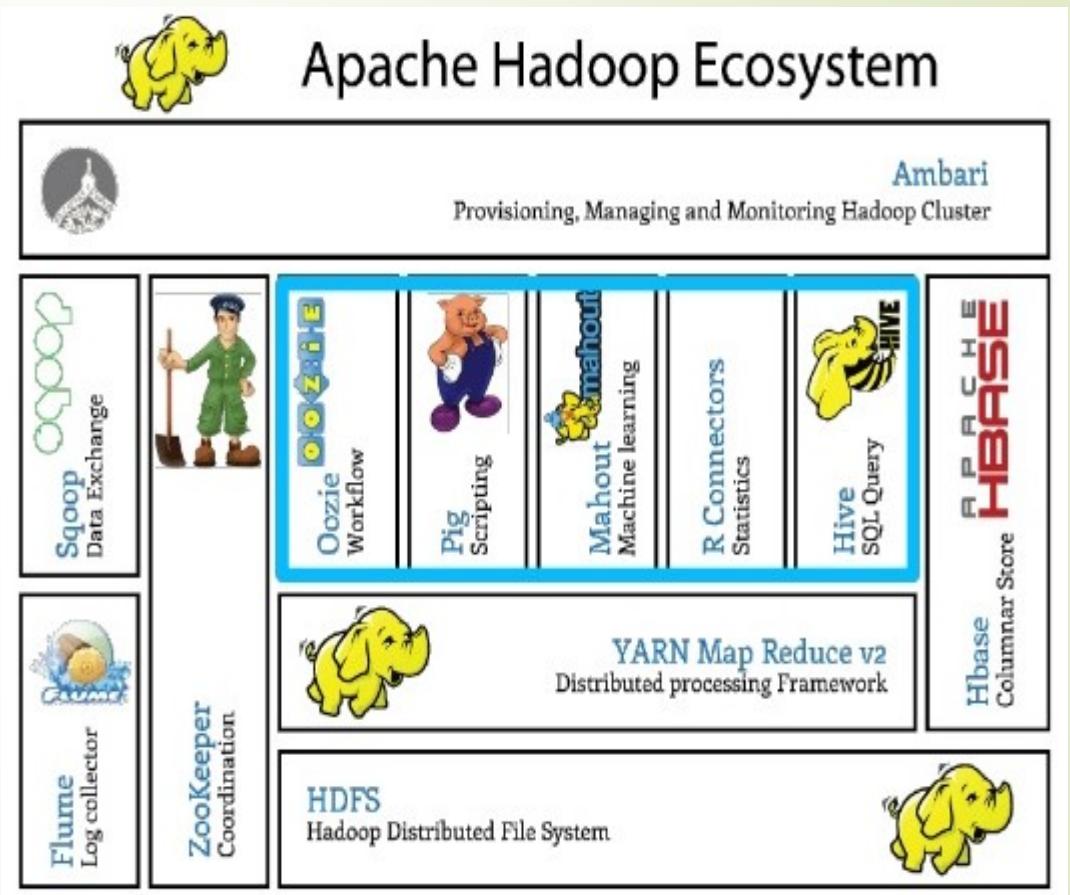
Ecosystème de Hadoop

- ▶ Autres outils pour :
 - Extraction et stockage des données (HDFS)
 - Simplification de traitement des données
 - Gestion et coordination des opérations
 - Surveillance du cluster (Monitoring)



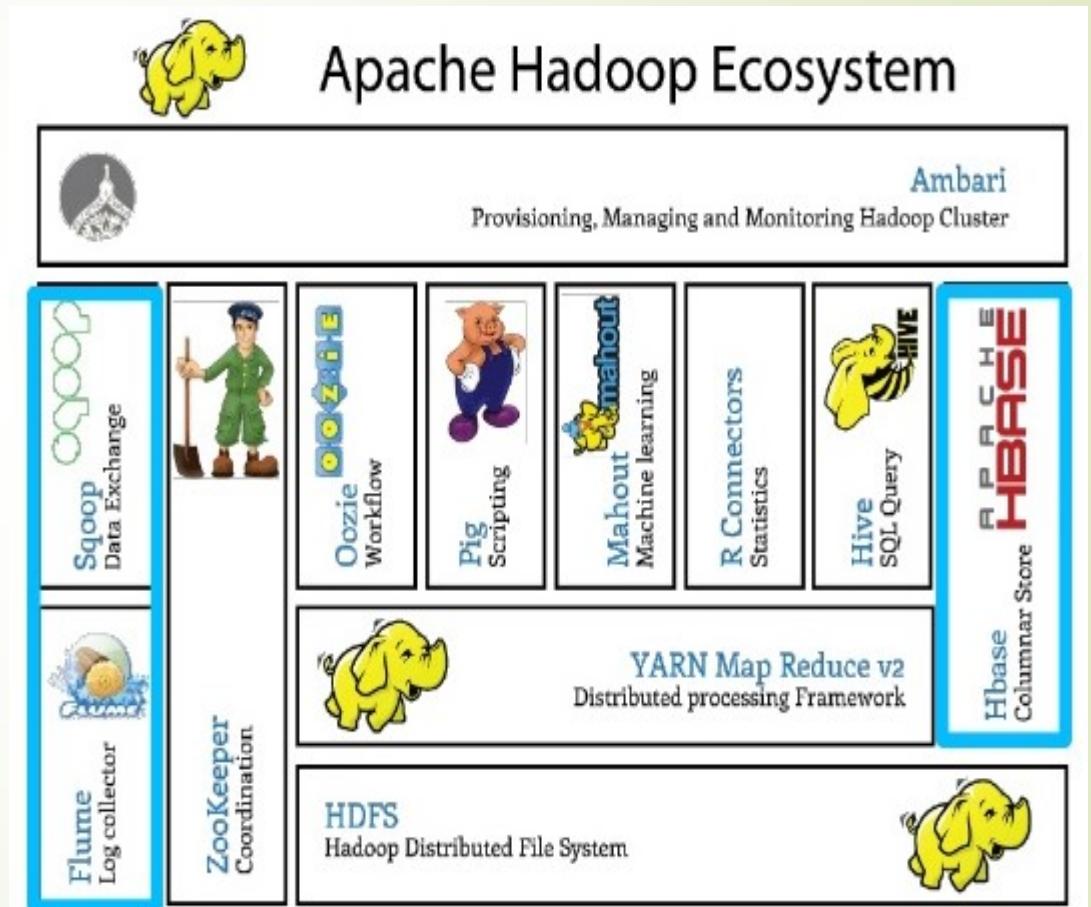
Ecosystème de Hadoop

- ▶ **Oozie**: l'ordonnancement des tâches de Map Reduce (jobs) par la définition des workflows
- ▶ **Pig**: langage de programmation de requêtes sur des fichiers HDFS (plus simple que Java) pour écrire des jobs MapReduce.
- ▶ **Mahout**: bibliothèque de Machine learning et mathématiques
- ▶ **R Connectors**: exécution des requêtes Map Reduce avec langage R
- ▶ **Hive**: base de données de Hadoop qui possède un langage d'interrogation, HiveQL, inspiré de SQL



Ecosystème de Hadoop

- ▶ **Hbase**: base de données NoSQL orientée colonnes
- ▶ **Impala**: requêtage de données à partir du HDFS (ou Hbase, Hive) par des requêtes Hive QL
- ▶ Outils pour la connexion HDFS et sources externes:
- ▶ **Sqoop**: manipulation des bases de données externes
- ▶ **Flume**: collecte de logs et stockage dans HDFS



Ecosystème de Hadoop

- Outils pour la gestion et l'administration de Hadoop:
 - ▶ **Ambari:** outil pour l'administration, la gestion et monitoring du cluster
 - ▶ **Zookeeper:** outil pour maintenir les informations de configuration, de nommage et de synchronisation distribuée

