

# **Introduction *BIG DATA***

**Licence Informatique – Technologies du Web**

*Laure Soulier*

**2016-2017**

# Plan du chapitre

---

- Historique des Bases de Données (BD)
- Emergence des *Big Data*
- Vocabulaire autour des *Big Data*
- Aperçu des approches et solutions *Big Data*

# Donnée, information, connaissance (1)

---

## □ **De la donnée à l'information**

- Une donnée est l'enregistrement d'une **observation**, **objet**, **fait** destiné à être interprété, traité par l'homme. La donnée est généralement **objective**

*Exemples :*

- *température = 35°*
- *âge = 2 mois*

- Une information est le **signifiant attaché à la donnée** ou à un ensemble de données par association. L'information est généralement **subjective, définie selon un contexte**

*Exemples*

- *(température=35°) : temps chaud*
- *(âge=2 mois) : nourrisson*

# Donnée, information, connaissance (2)

---

## □ **De l'information à la connaissance**

- Une connaissance est une information nouvelle, apprise par association d'informations de base, de règles, de raisonnement, d'expérience, d'expertise, etc. La donnée est généralement objective, peut être subjective.

*Exemple :*

*- temps chaud et enfant nourrisson alors risque de déshydratation*

# Bases de Données (BD) : c'est quoi ?

---

## ❑ **Des fichiers à la base de données**

- Un fichier est un **ensemble d'enregistrements physiques** qui représentent des données manipulées par plusieurs utilisateurs ayant une **vue unique** de ces données
- Une base de données est un **ensemble de données construit selon un schéma**, d'où peuvent être dérivés **différentes vues manipulées** par plusieurs utilisateurs

## ❑ **De la base de données à la banque de données**

- Une base de données est **ensemble structuré de données**, destinées à être exploitées par des applications cibles

*Exemple*

*Base de données personnel université (suivi de carrière, paie, ..)*

- Une banque de données comporte les **données de référence associées à un domaine donné**; elle est généralement structurée en un ensemble de bases de données.

# Bases de Données (BD) : c'est quoi ? (2)

---

## □ Base de Données (*BD*)?

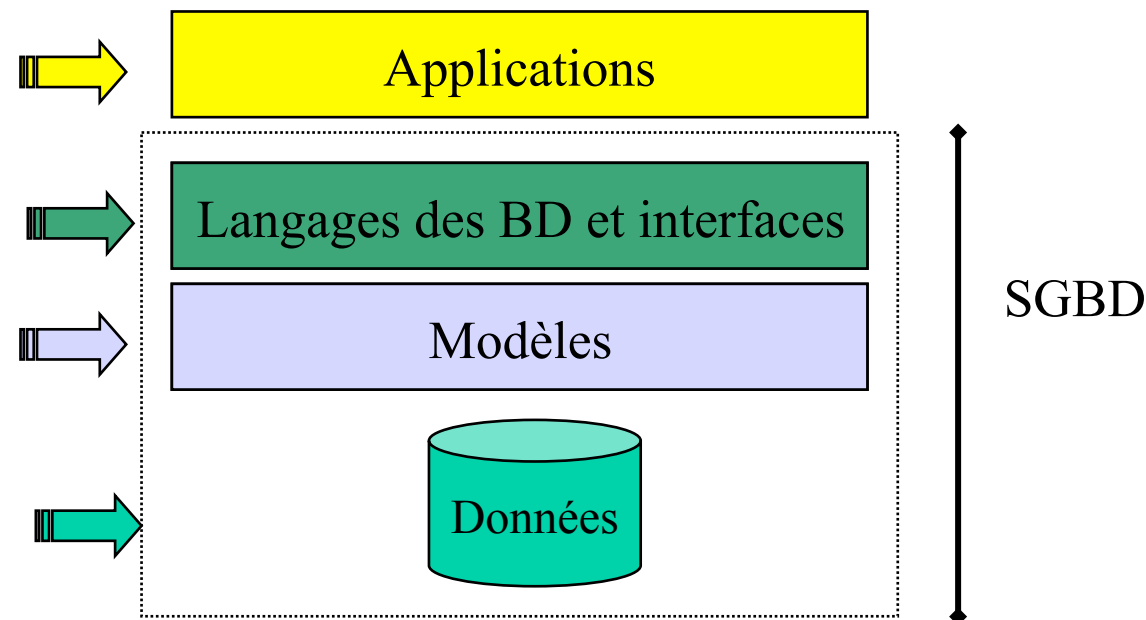
- Un **ensemble structuré** d'informations **agrégées ou élémentaires** accessibles par une communauté d'utilisateurs [Chrisment,08]
- Une collection de données qui intègre :
  - **une structure intégrée**,
  - **des liaisons sémantiques** entre données,
  - **des contraintes d'intégrité**,
  - **des vues** de différents utilisateurs.
- Une collection de données qui supporte des **opérations de manipulation et de recherche** de données :
  - cohérente,
  - sécurisée,
  - pérenne.

## BD et SGBD : Historique

### ☐ + 50 années de réflexion sur la gestion des données

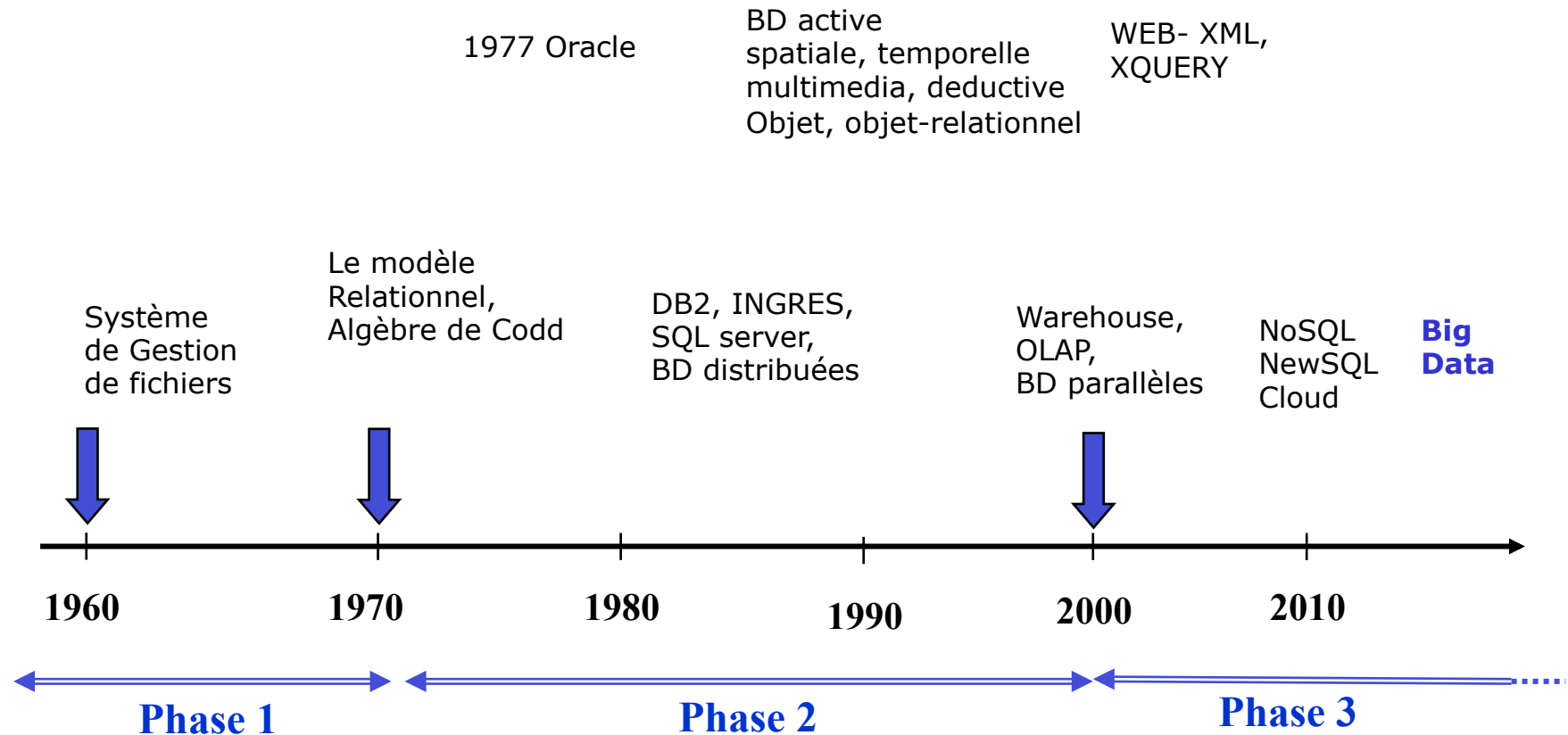
- Phase 1 : période préhistorique 1960 – 1969
- Phase 2 : période phare 1970 – 2000
- Phase 3 : période nouvelle 2000 – ...

*Evolution au niveau :*



*Eléments extraits de Tutoriel,  
M. Adiba, EDBT 2013*

# BD et SGBD : Historique



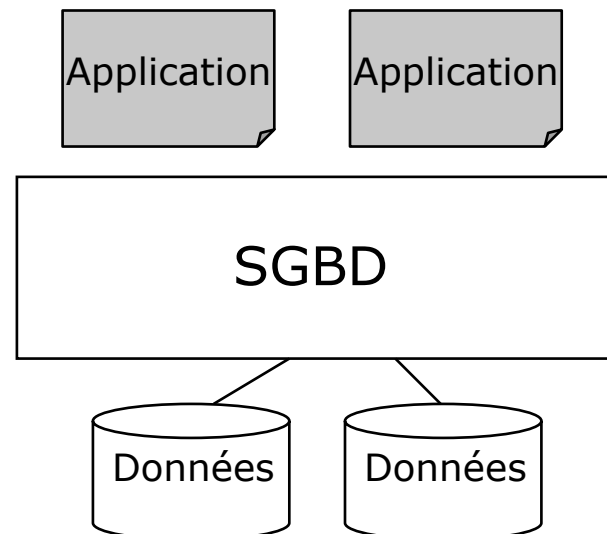


## BD et SGBD : quelques repères historiques (60-69)

---

### ❑ Introduction des principes de bases des BD/SGBD

- Plusieurs applications partagent des données
- Séparation des données et des traitements sur les données
- Données gérées par un serveur central
- Minimisation de la redondance et de l'inconsistance
- Amélioration du contrôle des données
- Accès par langage standardisé (COBOL)

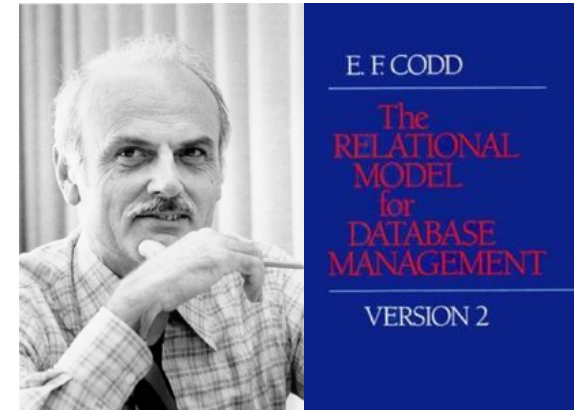


## BD et SGBD : quelques repères historiques (70-2000)

---

### ❑ Modèle relationnel

- Inventé par *Edgar Franck Codd* en 1970
- Fondements :
  - Algèbre relationnelle, logique de prédicat de 1<sup>er</sup> ordre
  - Indépendance des données, vue tabulaire
  - Langages : SQL, QUEL, QBE
  - Dépendances fonctionnelles, formes normales
- Prototypes SGBDR (1975)
- SGBDR commercialisés en 1980



## BD et SGBD : quelques repères historiques (70-2000)

---

### ❑ Notion de transaction et propriétés, *J. Gray* 1975

- Langage relationnel de requêtes
  - Langage déclaratif, non procédural
  - Indépendance des données/traitements
  - Introduction du principe d'optimisation des requêtes
- Propriétés ACID
  - Atomicité : principe de TOUT ou RIEN, une transaction est exécutée intégralement ou pas du tout
  - Cohérence : l'exécution de toute transaction assure le passage de la base d'un état cohérent vers un autre état cohérent
  - Isolation : une transaction est exécuté indépendamment des autres qui s'exécutent simultanément
  - Durabilité : les modifications opérées dans la base par une transaction sont pérennes
- Impact
  - Théorique : protocoles pour la gestion de la concurrence (exclusif, partagé, à deux phases,...)
  - Pratique : modules de gestion de la concurrence, contrôleurs de concurrence, algorithmes de reprise, gestion d'inter blocage

## BD et SGBD : quelques repères historiques (70-2000)

---

### ❑ Vers de nouveaux types de données...

- Temporel

- Extensions SQL : SQL2, TempSQL, TQUEL, TSQL2,...*R. Snodgrass, 1985, 1986*
- Introduction de propriétés temporelles, SQL2011  
(<http://www.sigmod.org/publications/sigmod-record/1209/pdfs/07.industry.kulkarni.pdf>)

- Spatial

- Extension de SQL à des objets spatiaux, *M. Egenhofer, 1994*
- Opérations et relations spatiales
- Requêtes interactives : localisation des régions par l'utilisateur

- Multimédia

- Introduction texte, image, audi, vidéo, *M. Adiba, 1996*
- Extensions SQL (temps, données continues,..)

## BD et SGBD : quelques repères historiques (70-2000)

---

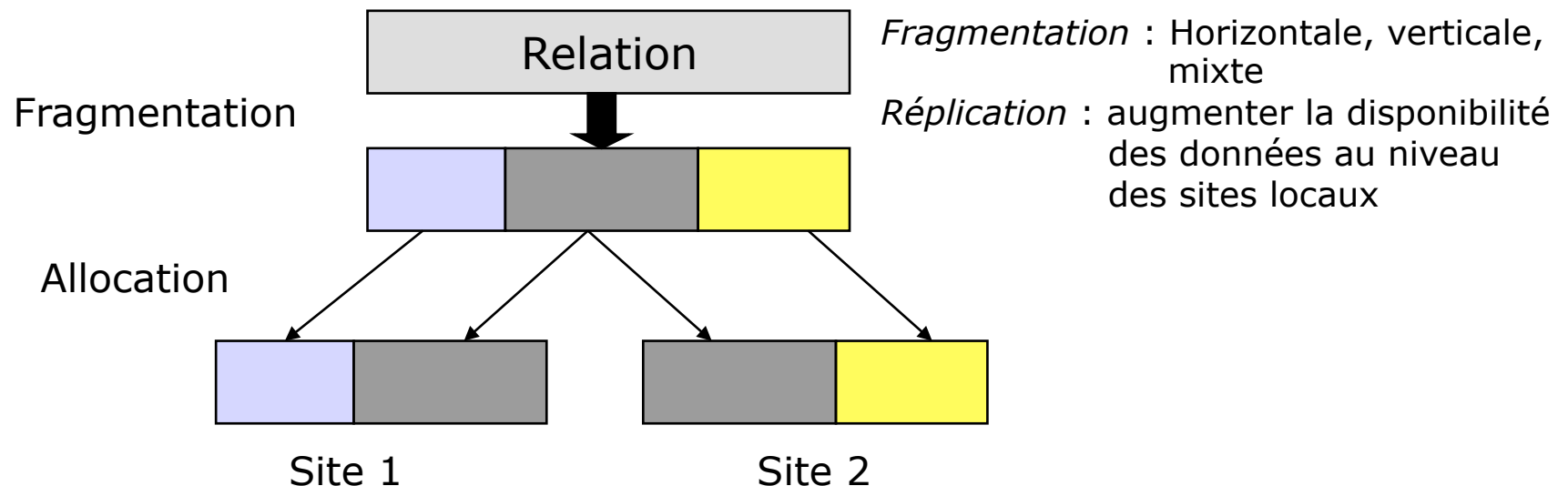
### ❑ Vers de nouveaux types de données...

- Document
  - Données + documents semi-structurés (XML)
  - GML (1971), SGML (1986), HTML (W3C, 1986), XML (W3C, 1998)
  - Introduction Xpath, Xquery, ..
- Objet, modèle NF2 (1985, 1986)
  - Non First Normal Form
  - Introduction des concepts classe, méthode, héritage
  - Extensions SQL : SQL2, OSQL
- Multimédia (1995)
  - Introduction texte, image, audi, vidéo, *M. Adiba*, 1996
  - Extensions SQL (temps, données continues,..)

## BD et SGBD : quelques repères historiques (70-2000)

### ❑ Données et systèmes distribués : milieu des années 70

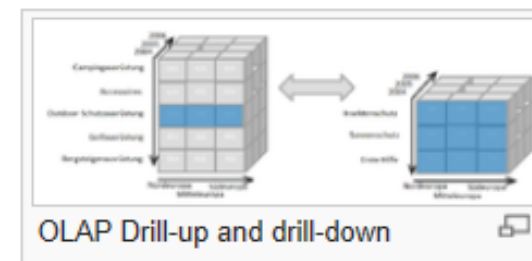
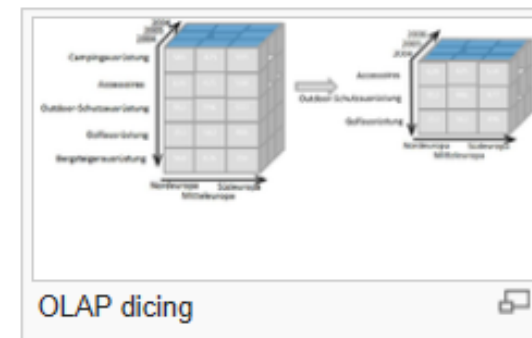
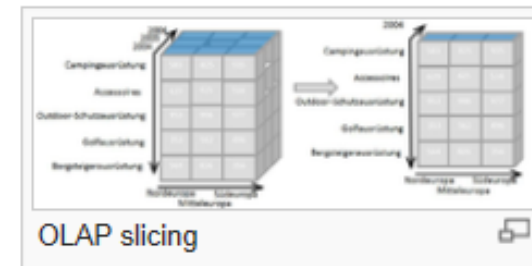
- Partition et replication (duplication) sur différents sites
  - Données
  - Shémas et catalogues de données
  - Système de contrôle (SGBD)
  - Infrastructure matérielle



## BD et SGBD : quelques repères historiques (70-2000)

### ❑ Data Warehouse & OLAP, 2000...

- OLAP : collection de données orientées « sujet », historisées, non volatiles consolidées dans une BD unique pour des besoins de gestion, prise de décision
- Schémas de données multidimensionnelles  
OLAP CUBE est une abstraction de l'opérateur relationnel de projection
- Requêtes décisionnelles plus complexes que pour une BD classique. Opérations de synthèse sur les données : rotation, *slicing*, *dicing*, forage vers le haut (*drill-up*), forage vers le bas (*drill-down*)







# Big Data, c'est quoi ? (1)

---

## □ Quelques définitions

- Définition 1 : « data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges » Oxford English Dictionary, « données de très grande taille, dont la manipulation et gestion présentent des enjeux du point de vue logistiques »
- Définition 2 : « an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications » Wikipédia, « englobe tout terme pour décrire toute collection de données tellement volumineuse et complexe qu'il devient difficile de la traiter en utilisant des outils classiques de traitement d'applications »
- Définition 3 : « datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze » McKinsey, 2011, « collections de données dont la taille dépasse la capacité de capture, stockage, gestion et analyse des systèmes de gestion de bases de données classiques »

## *Big Data*, c'est quoi ? (2)

❑ **Bien d'autres définitions encore...**

<http://datascience.berkeley.edu/what-is-big-data/>

## ❑ *Ce qu'on retient ...*

Volume des données,  
Complexité,  
Limites des outils classiques  
de gestion des données,  
Passage à l'échelle

□ □



Top recurring themes in our thought leaders' definitions (word cloud via Wordle [↗](#))

# *Big Data, pourquoi ? (1)*

---

## ❑ ***Explosion des volumes des données générées sur le web, web mobile...***

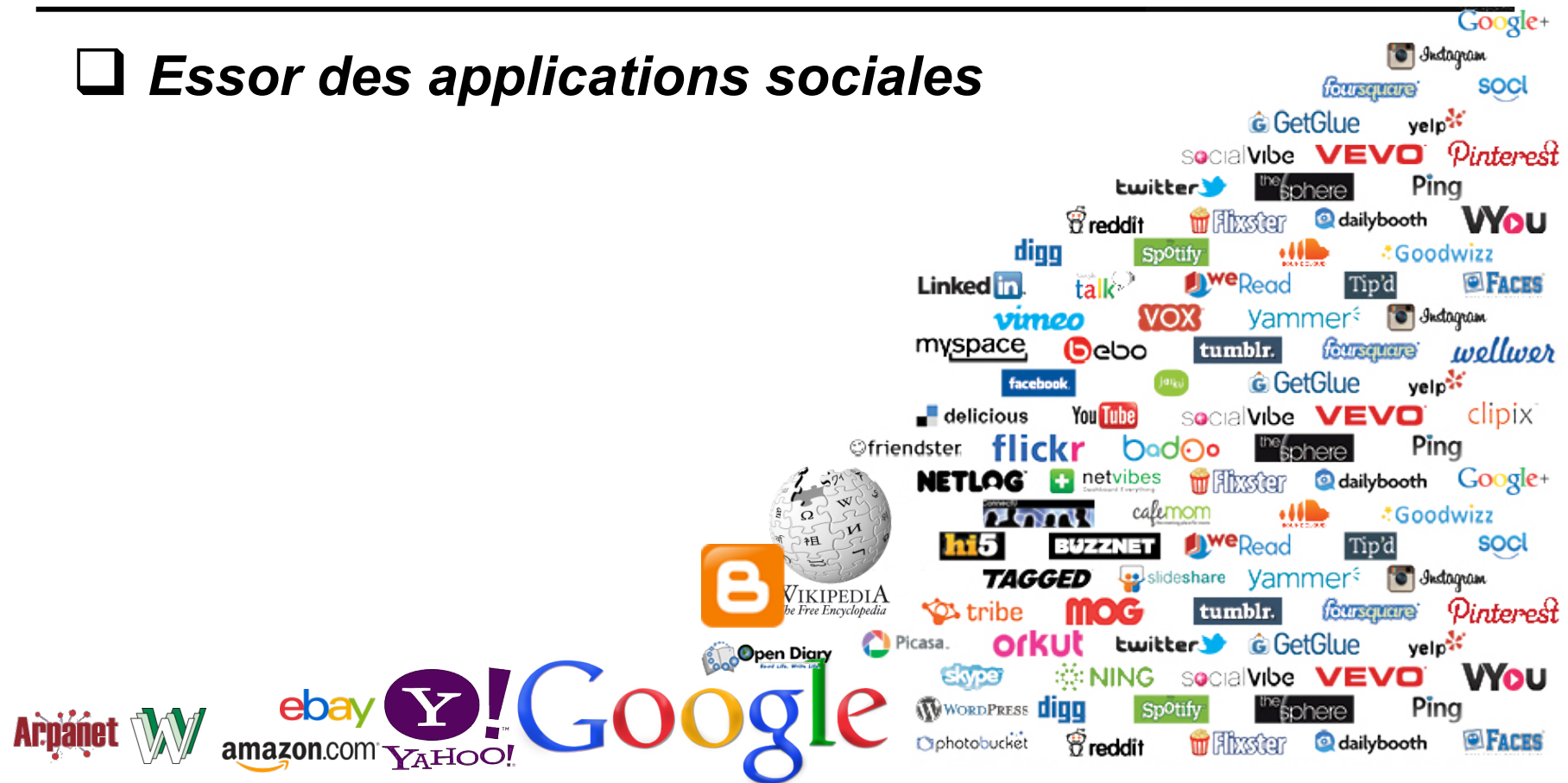
- Réseaux sociaux : Facebook, Twitter,...
- Moteurs de recherche : Google, Yahoo, Bing
- Internet des objets
- Sites commerciaux
- Appareils mobiles
- Capteurs
- Systèmes d'information des entreprises

## ❑+ ***Disponibilité, ouverture des données***

- *Open data* : données ouvertes au grand public
  - Gouvernement
  - Industries
  - Services : transports, météo, ...
  - ...

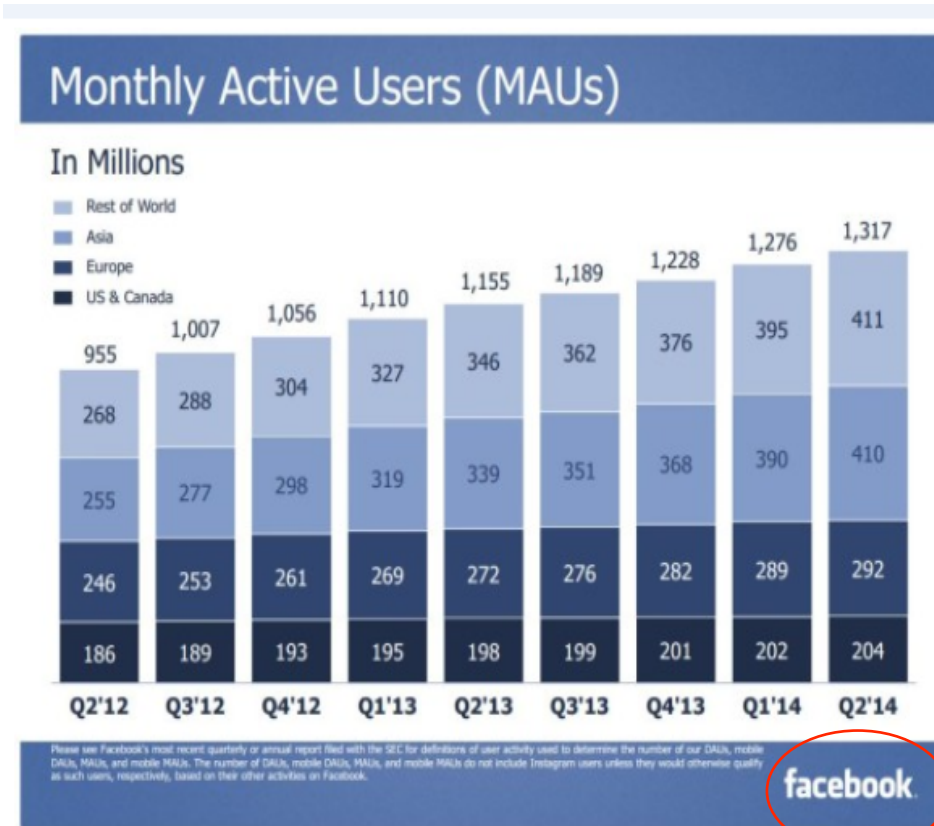
# Big Data, pourquoi ? (2)

## □ Essor des applications sociales

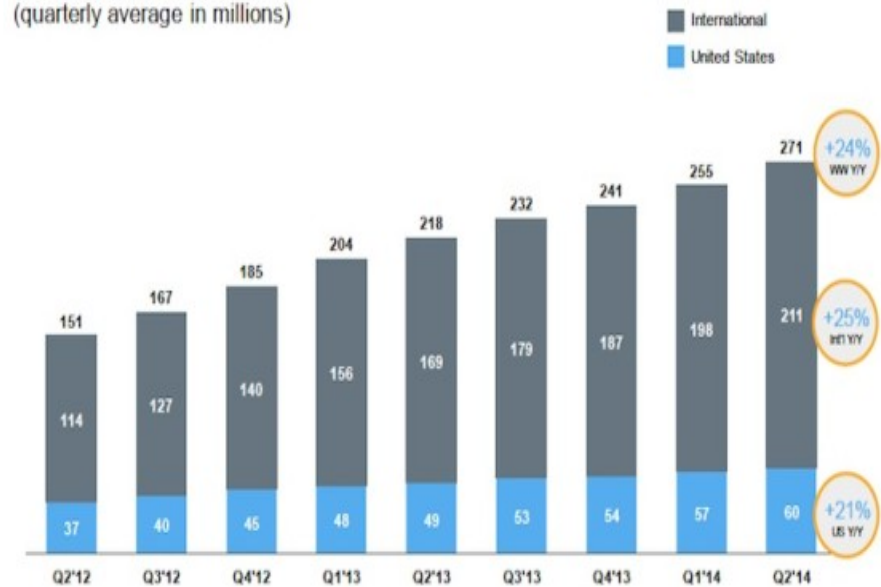
1972  
ARPANET1990  
WWW1994  
E-commerce1995  
Annuaire1998  
Recherche1999  
Blogs2001  
Wiki2003  
Réseaux sociaux

# Big Data, pourquoi ? (3)

## ❑ Chiffres à l'appui : utilisateurs des réseaux sociaux



Monthly active users  
(quarterly average in millions)



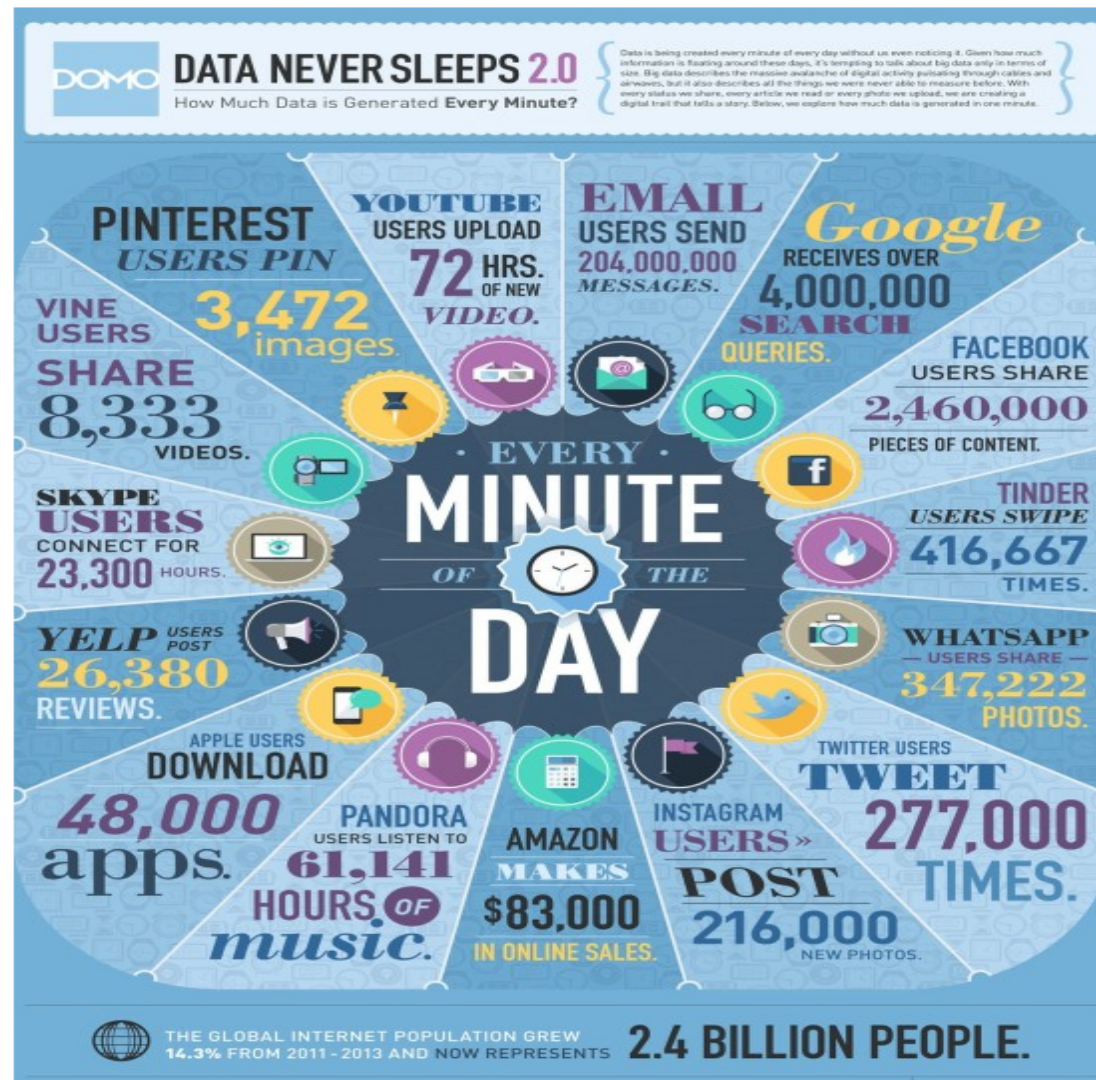
<http://www.blogdumoderateur.com/reseaux-sociaux/facebook/chiffres-facebook/>

<http://www.blogdumoderateur.com/reseaux-sociaux/twitter/chiffres-twitter/>



# Big Data, pourquoi ? (4)

## ❑ Chiffres à l'appui : volumes de données par minute sur le web



<http://www.blogdumoderateur.com/60-secondes-internet-2014/>

# *Big Data, pourquoi ? (5)*

---

## ❑ ***Explosion des volumes des données générées sur le web, web mobile...***

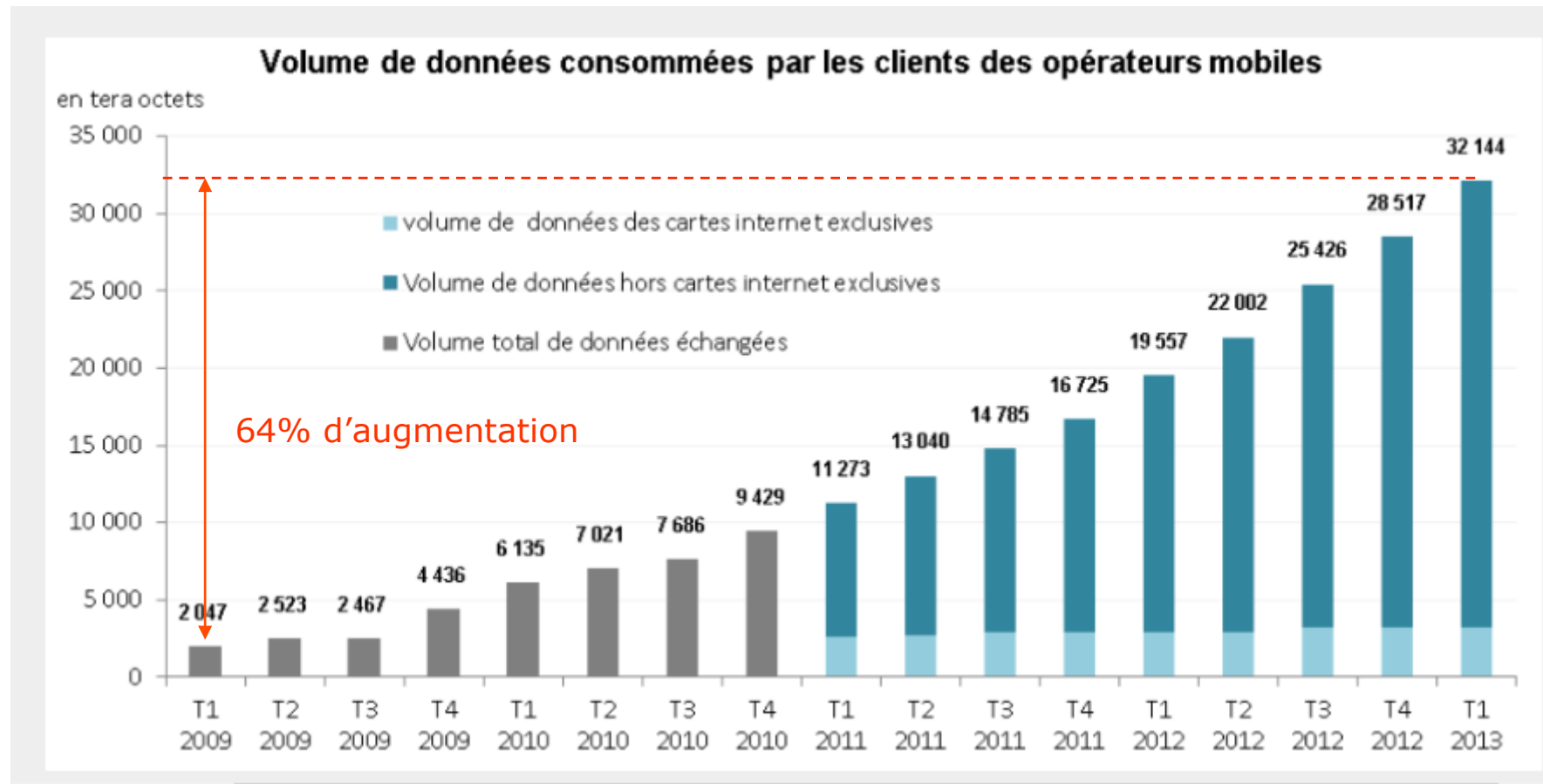
- Réseaux sociaux : Facebook, Twitter,...
- Moteurs de recherche : Google, Yahoo, Bing
- Internet des objets
- Sites commerciaux
- Appareils mobiles
- Capteurs
- Systèmes d'information des entreprises

## ❑+ ***Disponibilité, ouverture des données***

- *Open data* : données ouvertes au grand public
  - Gouvernement :
  - Industries
  - Services : transports, météo, ...
  - ...

# Big Data, pourquoi ? (6)

Croissance des volumes de données générées par les appareils mobiles en France



<http://thecallr.com/fr/blog/2013/07/29/une-augmentation-de-32-des-communications-telephoniques-mobiles-depuis-2012/>



# *Big Data*, pourquoi ? (5)

---

## ❑ **..+ Variété des données, peu de structure...**

- Image
- Vidéo,
- *Logs*,
- *Graphes*,
- *Son*,
- ..

## ❑ **..+ Dynamicité des données...**

- Flux de d'images (TV stream),...
- Flux de *tweets*
- Flux de données des capteurs
- ...

## ❑ **..+ Variété des sources**

- Mobiles
- Machine-Machine
- Machine-Homme
- Homme-Homme

# *Big Data, pourquoi ? (5)*

---

## ❑ **..+ Limites des SGBD**

- Capacités de stockage / traitement des SGBD
  - 1980 : Teradata database machine
  - 2010 : Oracle Exadata Database machine
- Nature/type des données
  - Structurée ou semi-structurées
- Vitesse de stockage
  - Temps de stockage ne suit pas le progrès en termes de vitesse des réseaux

## ❑ **...Passage à l'échelle des SGBD à quel coût ?**

# Big Data, pourquoi ? (6)

**Exercice** : Quel est le coût de stockage de 48 heures de vidéo extraites de Youtube dans une base ORACLE Exadata vs. système Big Data dédié

## Software License Costs for Reproducing YouTube Using Oracle

### Database Software

#### Per Machine:

CPUs per Exadata Server	16	
* Cores per CPU	8	
* Core Factor (Intel 7560)	0.5	
* Price of Database Enterprise Edition w/ options	\$107,000	
= <b>Cost per Exadata Machine (\$M)</b>	<b>\$6.8</b>	
* Number of Exadata Machines	26	
= <b>Cost of Database Licenses (\$M)</b>		<b>\$178</b>

### Middleware Software

#### Per Machine:

CPUs per Exalogic Server	30	
* Cores per CPU	12	
* Core Factor (Intel 7560)	0.5	
* Price of WebLogic Enterprise Edition	\$25,000	
= <b>Cost per Exalogic Machine (\$M)</b>	<b>\$4.5</b>	
* Number of Exalogic Machines	22	
= <b>Cost of Middleware Licenses (\$M)</b>		<b>\$99</b>

### Operating System Software

Total servers	70.0	
* Price of Linux Ultimate Edition	\$2,299	
= <b>Total Linux Support Costs</b>		<b>\$0</b>

### Exadata Storage Software

Total Exadata Cells	1,373	
* Disks per Cell	12	
= <b>Total Disks</b>	<b>16,476</b>	
* Price Exadata Software/Disk	\$10,000	
= <b>Cost of Exadata Storage Licenses (\$M)</b>		<b>\$165</b>

### Total License Cost (\$M)

**\$442**

### Maintenance Cost

Annual Maintenance Cost (\$M)

**\$97**

## Software Support Costs for Reproducing YouTube Using Open Source Software

### Data Management Software

Number of Data Management Server Racks	2	
* Servers per Rack	8	
* Annual Subscription per Server, Datameer Hadoop	\$12,000	
= <b>Cost of Database Support (\$M)</b>		<b>\$0.2</b>

### Middleware Software

Number of App Server Racks	54	
* Servers per Rack	8	
* Number of CPUs	2	
* Price of Jboss per CPU	\$1,406	
= <b>Cost of Middleware Support (\$M)</b>		<b>\$1.2</b>

### Operating System Software

Number of Racks	222	
* Servers per Rack	8	
* Price of Red Hat Enterprise Linux Support	\$6,498	
= <b>Cost of Database Support (\$M)</b>		<b>\$11.5</b>

### Total Support Cost (\$M)

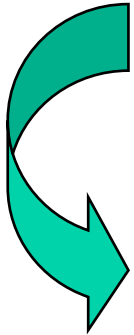
**\$12.9**

# *Big Data, à quoi ça sert ? (1)*

---

## ☐ *Explosion des domaines d'application utilisant les Big Data*

- Médical
- Marketing
- Politique
- Economie,
- ...



## *Pour ?*

- L'aide à la décision
- La prévision
- La découverte de nouvelles connaissances,...

# Big Data, à quoi ça sert ? (2)

---

## □ Quelques cas d'étude

- **Prédire les conflits mondiaux**

*L'outil GDELT, développé par l'université de Georgetown et accessible de manière open source, compile toutes les actualités (communiqués de presse, articles, discours...) parues depuis 1979. Il applique ensuite des techniques d'analyse sémantique et des algorithmes auto-apprenants pour faciliter la compréhension des événements récents et des principes de cause à effet pour arriver à prédire les conflits mondiaux –*

- **Gérer les catastrophes naturelles**

*En utilisant des outils de tracking, d'analyse sémantique et de visualisation en temps réel, l'Organisation Mondiale de la Migration a pu assister les forces locales en dégagant les urgences sanitaires, la localisation des ressources clés et en optimisant l'allocation des ressources sur le terrain lors du typhon qui a frappé les Philippines en 2013*

- **Faire de la veille sanitaire**

*Des scientifiques de l'université de Brigham Young essaient de simuler la localisation des mouches tsé-tsé dans le but d'aider à contrôler la propagation d'épidémies. De la même manière, la police de Chicago utilise le Big Data et la visualisation de données pour contrôler les populations de rats dans la ville.*

# Big Data, à quoi ça sert ? (2)

---

## □ **Autres cas d'étude**

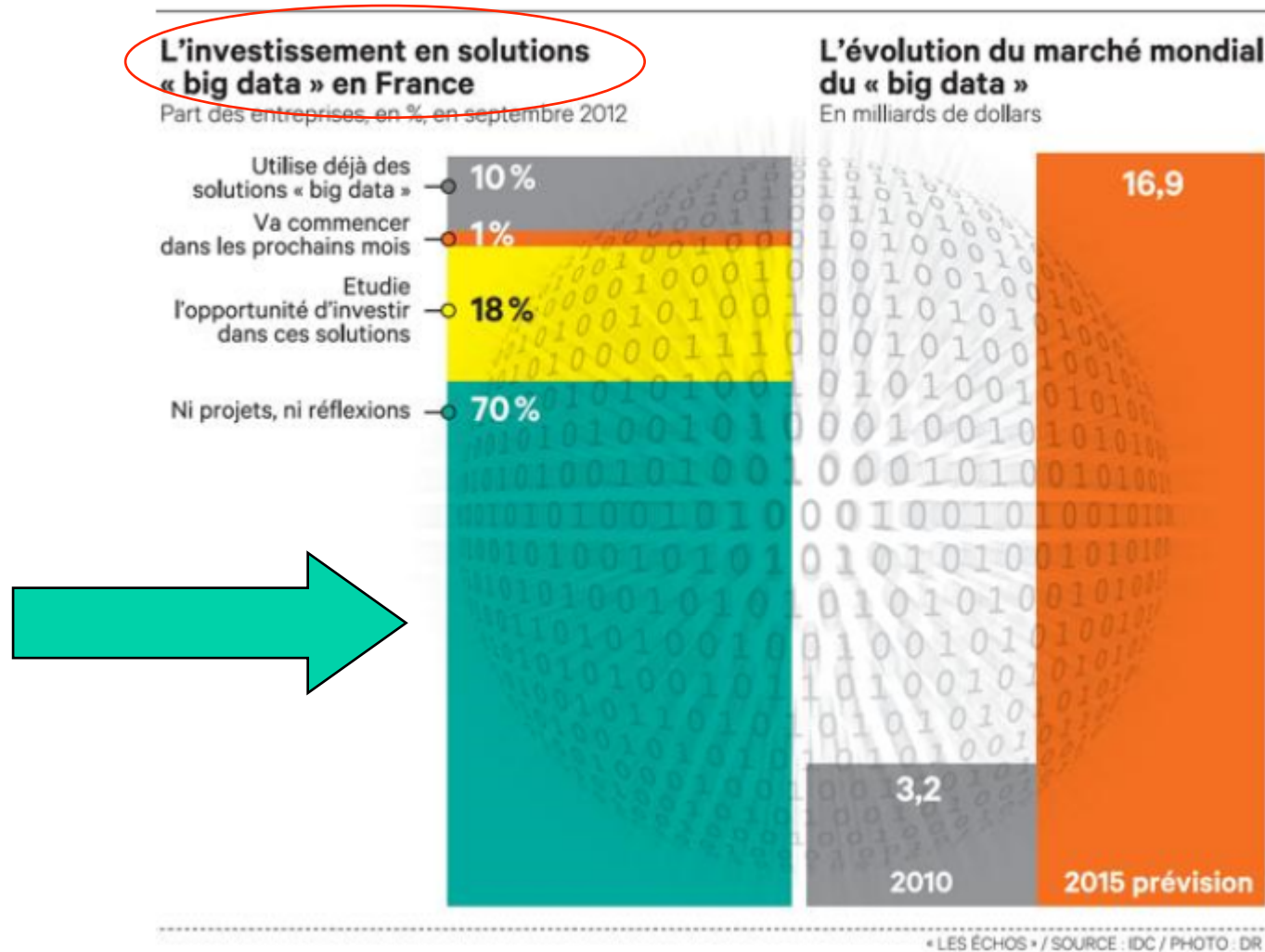
- **Cibler les clients sur le web**

*Dans le marketing web par exemple, le phénomène d'enchères en temps réel (Real-Time-Bidding – RTB), s'appuie sur de la data en mouvement pour proposer une publicité spécifique en fonction de l'utilisateur qui se connecte au site. L'entreprise Turn par exemple, classe l'utilisateur dans un segment lorsqu'il se connecte au site, en fonction de son historique de navigation et des informations issues de réseaux sociaux et lui affiche la publicité de l'annonceur ayant fait la meilleure enchère pour ce segment...en moins de 10 millisecondes - <http://www.data-business.fr/big-data-definition-enjeux-etudes-cas/#sthash.kRSvs3hq.dpuf>*

- **Bien d'autres...**

- *Secteur des Télécom.* : analyse de la qualité de service en temps réel
- *Secteur des banques* : prévention des fraudes et gestion du risque
- *Secteur des transports* : optimisation de trafics et des taux de remplissage
- *Secteur de l'éducation* : au travers des Massive Open Online Courses : pour comprendre les comportements des apprenants, et adapter les programmes
- ...

# Big Data, situation actuelle ? (1)



## *Big Data*, situation actuelle ? (2)

---

### ❑ En France... (source 01Business, 17/07/14)

« **Environ 10 % des entreprises françaises** en utiliseraient déjà (une solution Big Data) selon une étude de Steria de 2013, contre un tiers au niveau mondial.

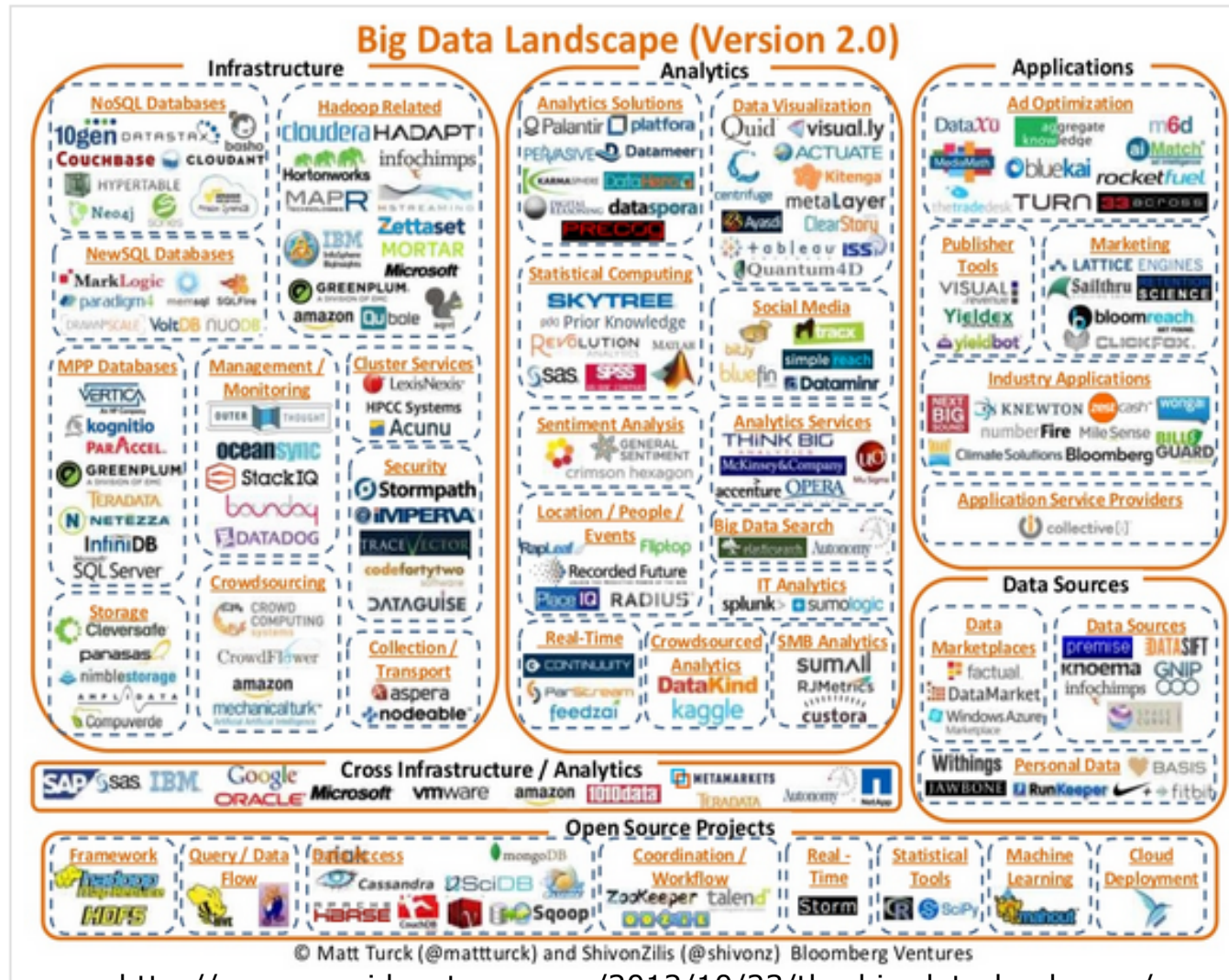
*« De nombreuses structures ont commencé à réaliser des POC (Proof of Concept), mais peu ont déroulé un projet de A à Z pour en tirer des enseignements et un retour sur investissement clair »*

*Gilbert Grenié*, associé de l'activité conseil au sein de PWC, partenaire de l'EBG pour le livre blanc Big Data

**Principales causes :** manque de compétences autour des big data : informatique pour les données massives, statistique, ...



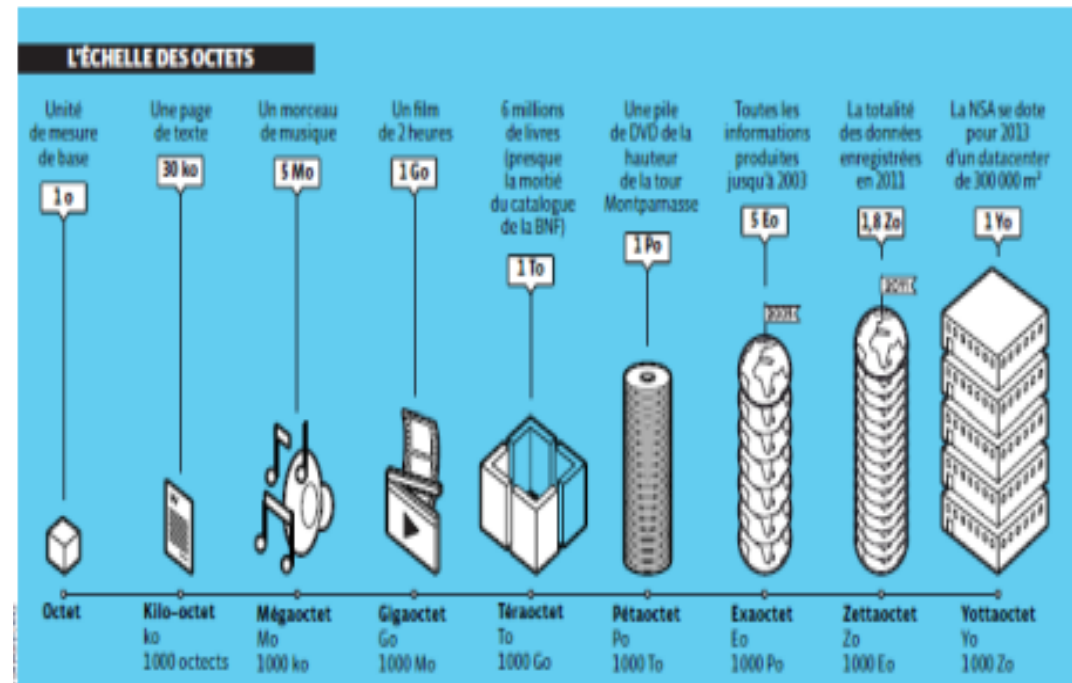
# Mots autour des Big Data



<http://www.ongridventures.com/2012/10/23/the-big-data-landscape/>

## Vocabulaire de base : unités de mesure de capacité de stockage

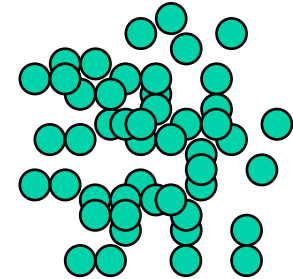
Unité de mesure	Eq. Octets
GigaOctets	$10^9$
TeraOctets	$10^{12}$
PetaOctets	$10^{15}$
Exaoctets	$10^{18}$
ZetaOctets	$10^{21}$



## Vocabulaire de base : Dimensions des Big Data ou les Big V

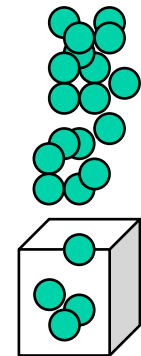
### □ **Volumétrie**

- Grande quantité de données
- Difficultés : stockage, recherche, partage, analyse, visualisation, ..



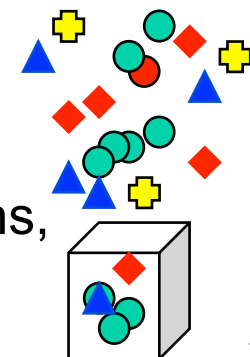
### □ **Vélocité**

- Flux continu de données : capteurs, appareils mobiles, réseaux sociaux...
- Difficultés : analyse et traitement des données à la volée, sans les avoir en intégralité (*one-pass processing*)



### □ **Variété**

- Différents formats : séquences, graphes, ..
- Difficulté d'intégration (jointure, association) par le sens, l'échelle, la qualité, ...



## Vocabulaire de base

Mot	Brève
MAP REDUCE	Principe de programmation qui consiste à distribuer et paralléliser le traitement sur plusieurs nœuds
HADOOP, HDFS (Hadoop Distributed File System)	Hadoop est une plate-forme informatique open-source de la fondation Apache, capable de gérer/traiter des big data sur une architecture distribuée. HDFS est le système de gestion de fichier de base qui supporte Hadoop
NOSQL	Technologie qui se différencie à la notion relationnelle des données, adaptée à des données peu structurées (nombre dynamique de colonnes, document, graphes,..
HBase, Cassandra, MongoDB, NE04J, Couche DB, Redis	SGBD qui supportent l'approche d'interrogation des données NOSQL
SAS, Talend, R, Python	Outils et ou environnements de programmation et analyse adaptés aux Big Data
Cloud computing	Ensemble de processus permettant d'offrir un espace de stockage sous forme de serveurs, accessibles à distance, sous forme de location. Utilile pour les entités (entreprises) qui ne souhaitent pas investir dans les infrastructures de stockage


# Quelles solutions pour le Big Data ?

---

## ❑ **Direction majeure**

Exploiter le parallélisme sur une architecture multi-processeurs

## ❑ **Comment ?**

- 
- Machines de bases de données
    - Pour les données massives, structurées, semi-structurées
    - Permet de pérenniser les solutions BD existantes => préservation des acquis, économie d'argent
    - Solutions propriétaires : ORACLE, MySQL, ..: amélioration des services à moindre coût
  - Environnement de programmation parallèle
    - MAP REDUCE , inventé par Google
    - Version logiciel libre (Open source) par Hadoop
    - Adapté aux données dynamiques, irrégulières, sans schéma qui sont inadaptées pour SQL, Xquery
  - Systèmes de Gestion de Bases de Données NoSQL
    - Pour les données non structurées : graphes, textes, ..

Avec possibilités de combinaisons de ces solutions