

Chp1 – Introduction au big data



Plan module

- **Introduction**
- Écosystème Hadoop
- HDFS
- MapReduce
- Langages de requête Hadoop : Pig, Hive
- SGBDNR
 - Différences entre une BDNR et une BD relationnelle
 - Typologies des BD non relationnelles
- Etude d'un SGBDNR : HBase

Plan

- Introduction
- Les caractéristiques du big data
- Le processus Big data
- Les data scientists
- Domaines d'application du Big data
- Challenges

Introduction

- Vieux paradigme
 - Déploiement de technologie pour améliorer la productivité
 - Des données sont créées
- Nouveau paradigme
 - Les données sont les matières premières du monde des affaires
 - Les valeurs des données et l'analyse de celle-ci ne sont plus remise en question.

Introduction

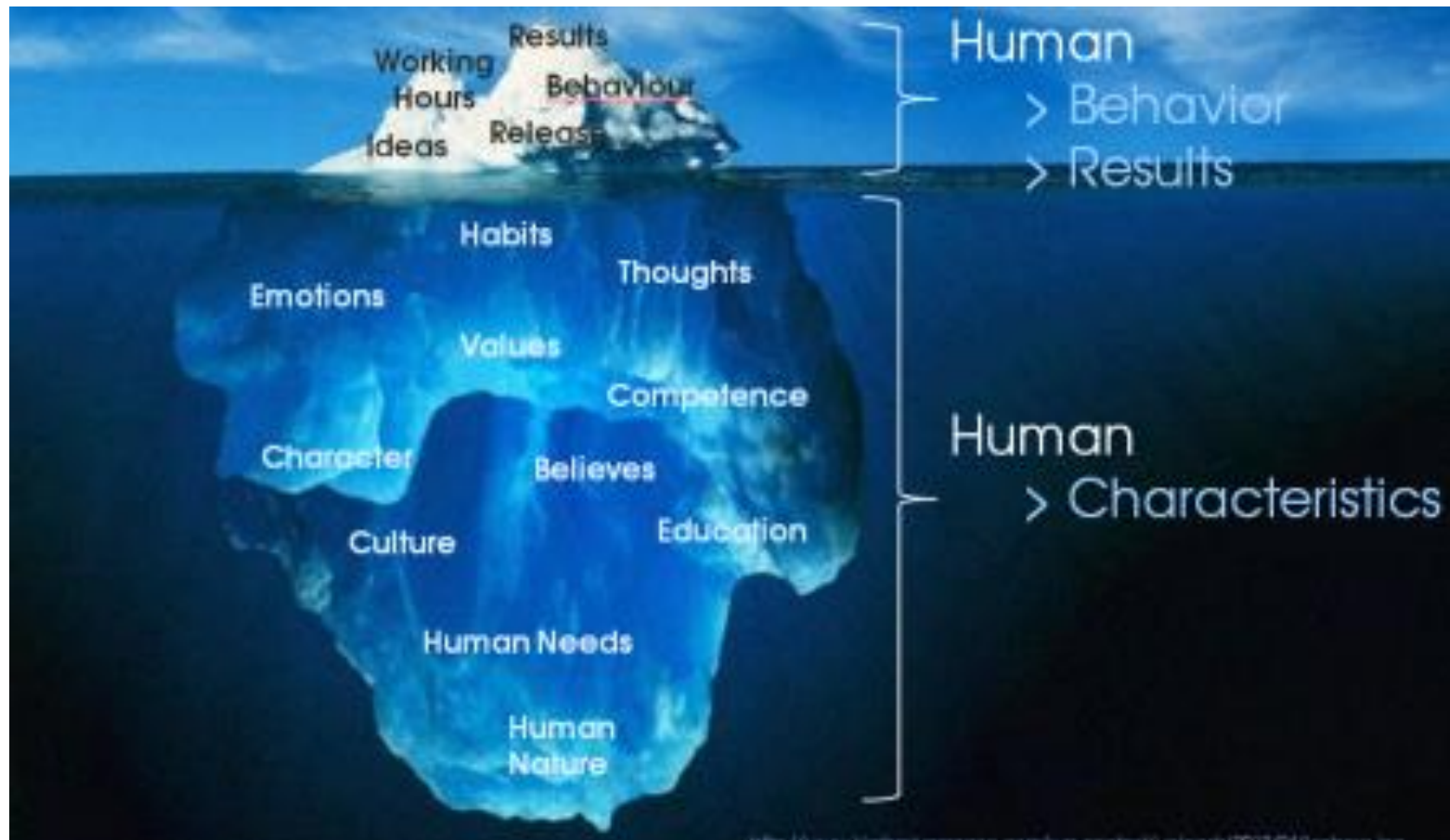
- **Les données sont de plus en plus précieuses.**

Les entreprises cherchent à libérer toute la valeur potentielle de leurs données afin d'en tirer des avantages concurrentiels.

« que pouvons-nous faire avec ces données ? »

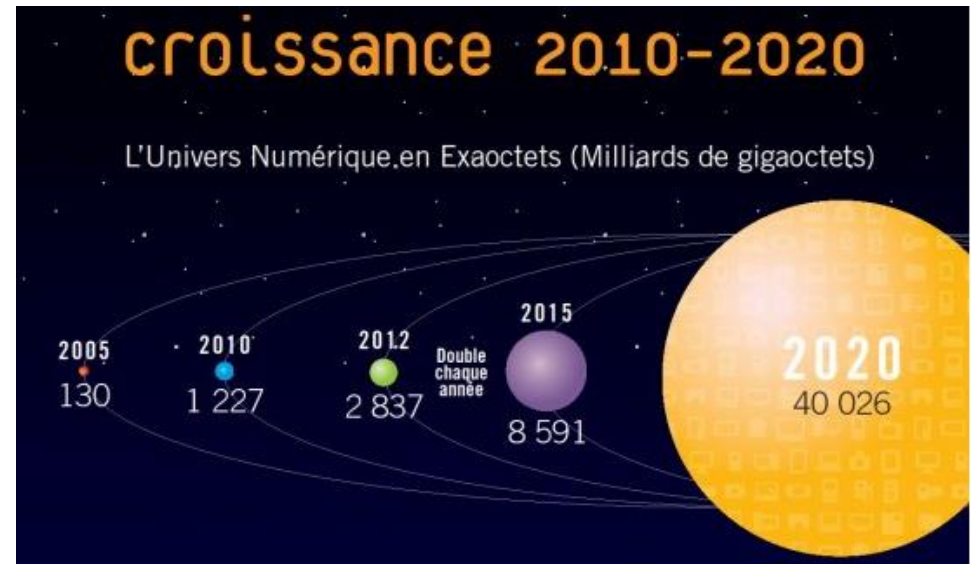
- **Le traitement en temps réel contribue à l'analyse prédictive.** L'analyse prédictive permet aux entreprises d'avoir une idée beaucoup plus claire de l'avenir et peut ouvrir d'excellentes opportunités de génération de valeur à partir des données.

Les sources du big data



Expansion volume de données

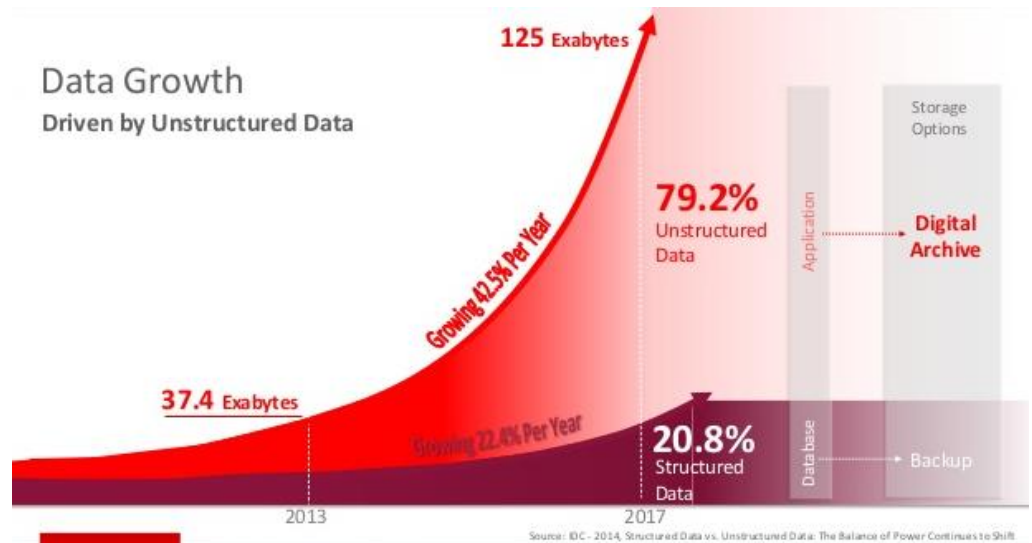
- Plus de données créées dans les trois dernières années que dans les 100 années qui les ont précédé
- Le total des données existantes ont quadruplé dans les trois dernières années
- 90% des données dans le monde ont été créées au cours des deux dernières années.



Expansion volume de données

- Source:

- Capteurs
- Messages sur les réseaux sociaux
- Images numériques et vidéos publiées en ligne
- Enregistrements transactionnels d'achat en ligne
- Signaux GPS de téléphones mobiles
- Cliques web
- Requêtes serveur



Sources des données : RFID tags

- Agriculture
- Sport
- Informations climatiques
- Traffic routier <http://www.511.org/>

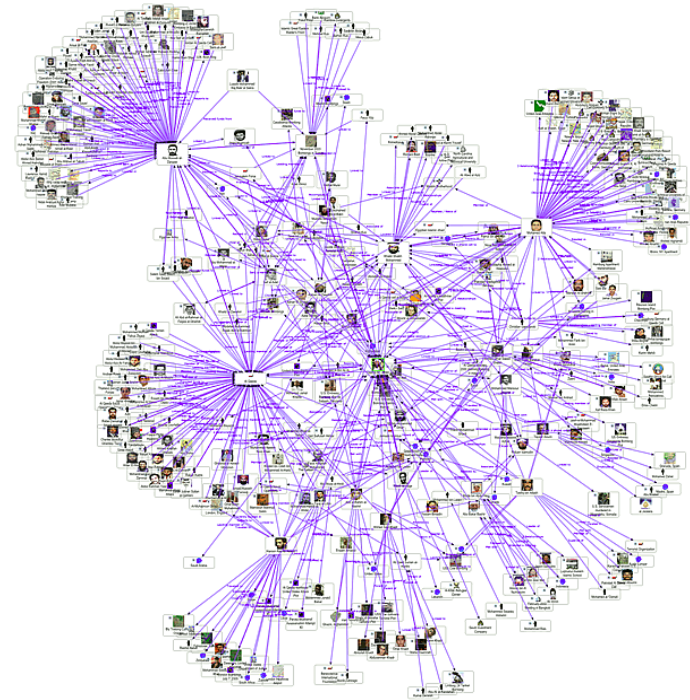


Source de données : Réseaux sociaux

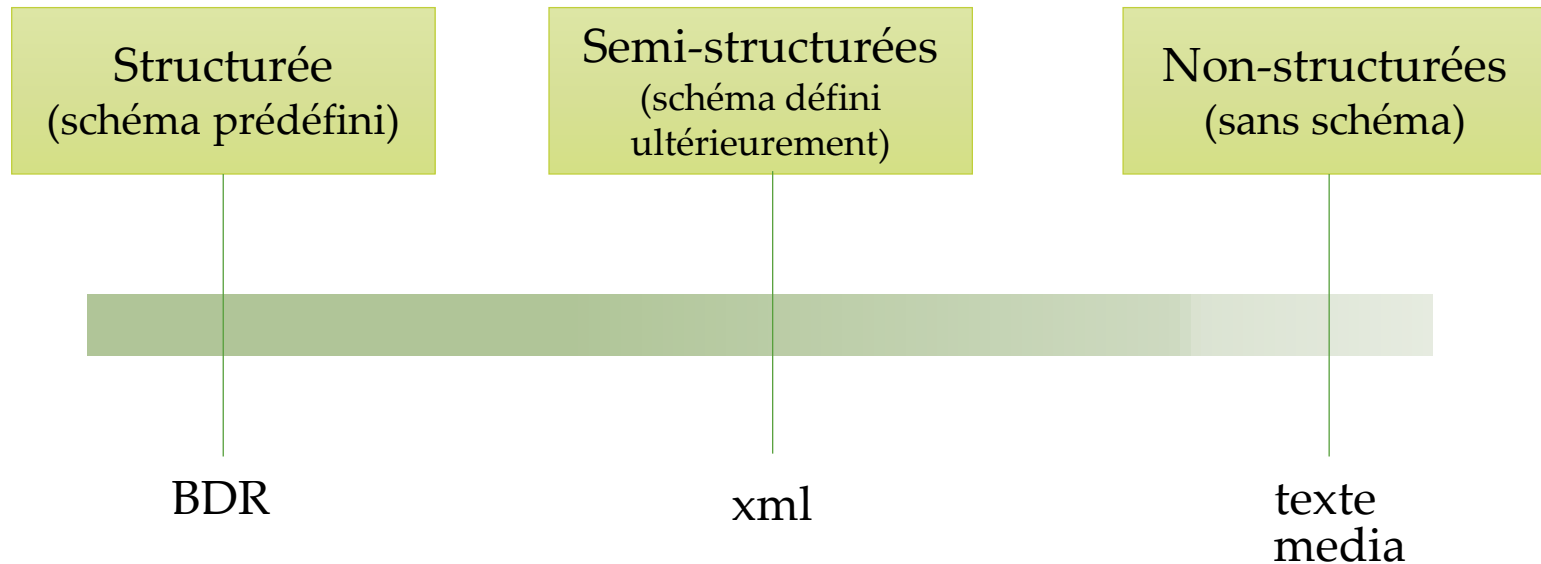


Les réseaux de données

- La plupart des données sont structurées sous format graph
 - Réseaux sociaux
 - Réseaux de télécommunication
 - Réseaux informatiques
 - Réseaux du trafic
 - ...



Modèles de données



Données structurées : Relation

- Modèle relationnel de données
- Une relation est une table avec des lignes et des colonnes
- Chaque relation a un schéma définissant les types de ses colonnes
- Le schéma prédéfini est statique

Données semi-structurées : Fichier log

```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/
1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/
1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-
logo.gif HTTP/1.0" 200 1713
```

Données semi-structurées : Documents PDF

```
HEADER      APOPTOSIS                                23-DEC-12    3J2T
TITLE       AN IMPROVED MODEL OF THE HUMAN APOPTOSOME
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: APOPTOTIC PROTEASE-ACTIVATING FACTOR 1;
COMPND      3 CHAIN: A, B, C, D, E, F, G;
COMPND      4 SYNONYM: APAF-1;
COMPND      5 ENGINEERED: YES;
COMPND      6 MOL_ID: 2;
COMPND      7 MOLECULE: CYTOCHROME C;
COMPND      8 CHAIN: H, I, J, K, L, M, N
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE      3 ORGANISM_COMMON: HUMAN;
SOURCE      4 ORGANISM_TAXID: 9606;
SOURCE      5 GENE: APAF-1, APAF1, KIAA0413;
SOURCE      6 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE      7 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
KEYWDS      APOPTOSIS PROTEASE ACTIVATING FACTOR-1, APAF-1, CYTOCHROME C,
KEYWDS      2 APOPTOSIS
```

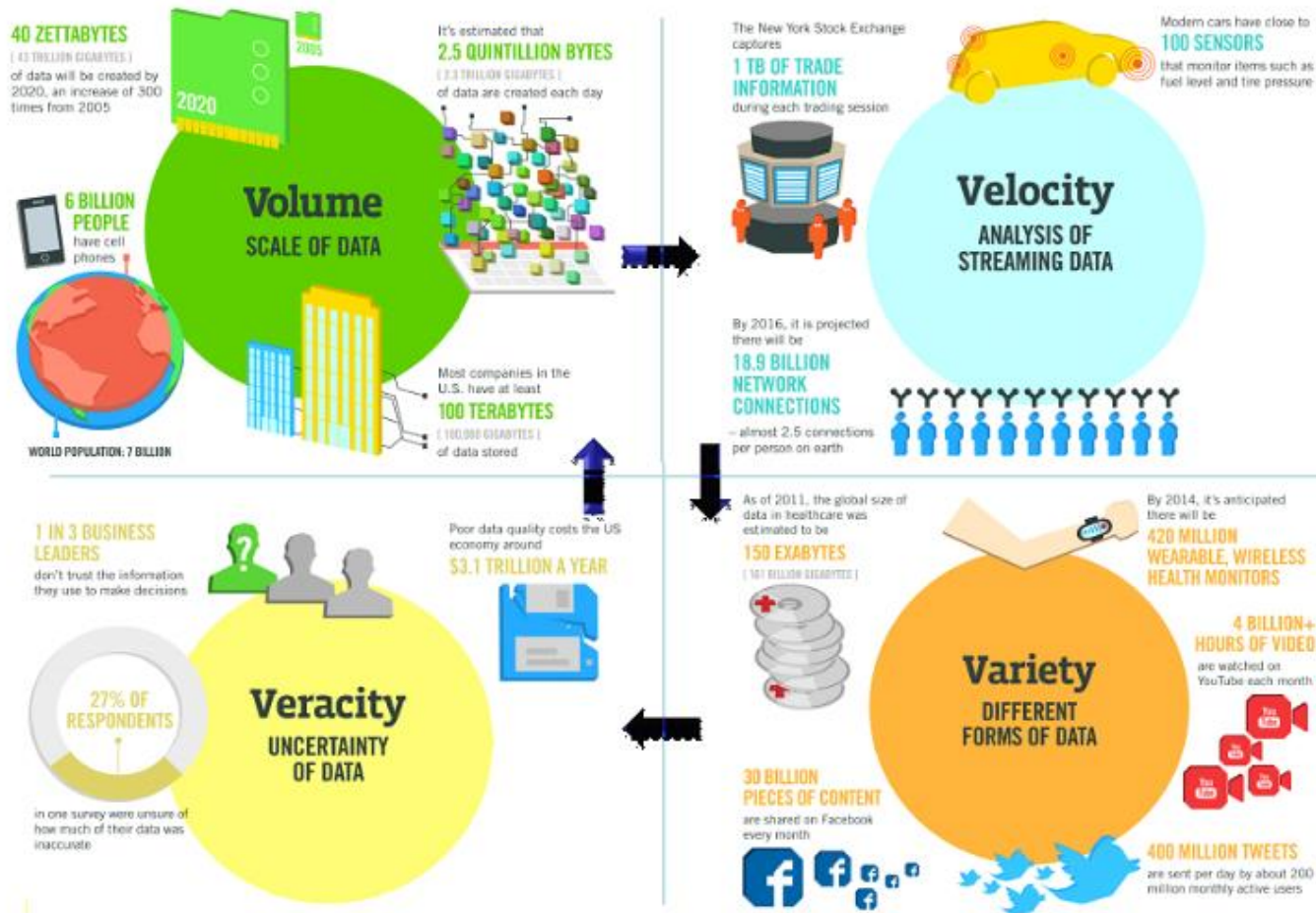
Données non-structurées

- Une seule colonne de type binaire ou chaîne de caractère
- Exemples:
 - Post Facebook
 - image Instagram
 - vidéo
 - Blog
 - Article journal
 - ...

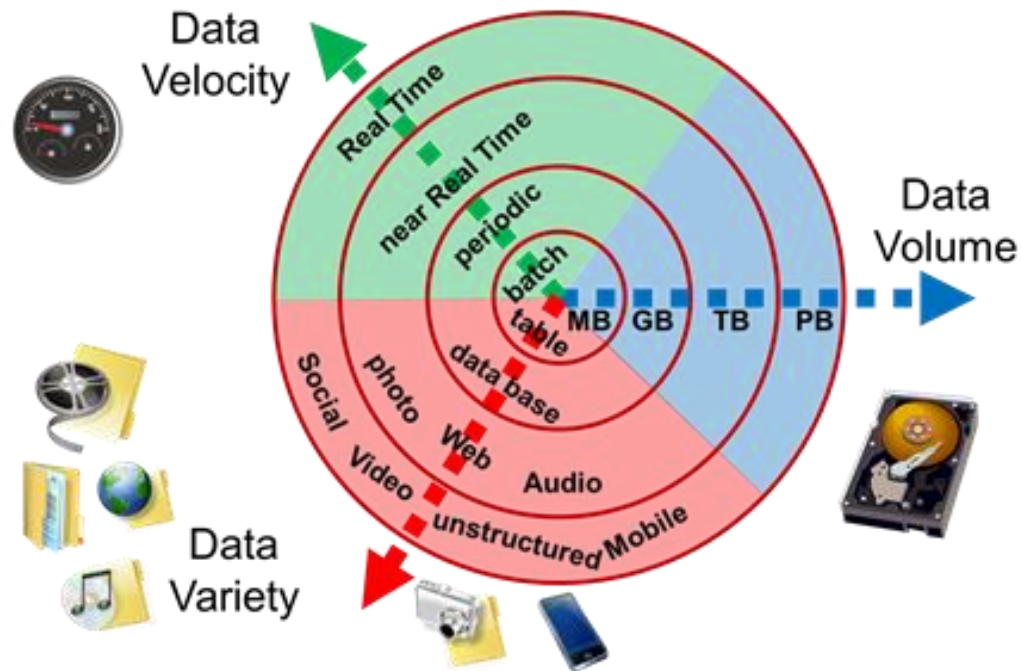
Définition

- Le Big Data est devenu un centre d'intérêt majeur pour le monde IT.
- En général, ce terme renvoie à des types de données relativement nouveaux (vidéo, images, son, etc.) qui génèrent des fichiers volumineux.
- Il désigne aussi de grands ensembles de petits volumes de données (commentaires sur les sites Web des réseaux sociaux, photos du fonds marin, images des caméras de surveillance du trafic) qui prennent leur sens lorsqu'ils sont combinés.
- Le plus souvent, ces Big Data connaissent une croissance rapide et certains ensembles de données modestes seront amenés à se développer pour devenir des Big Data.

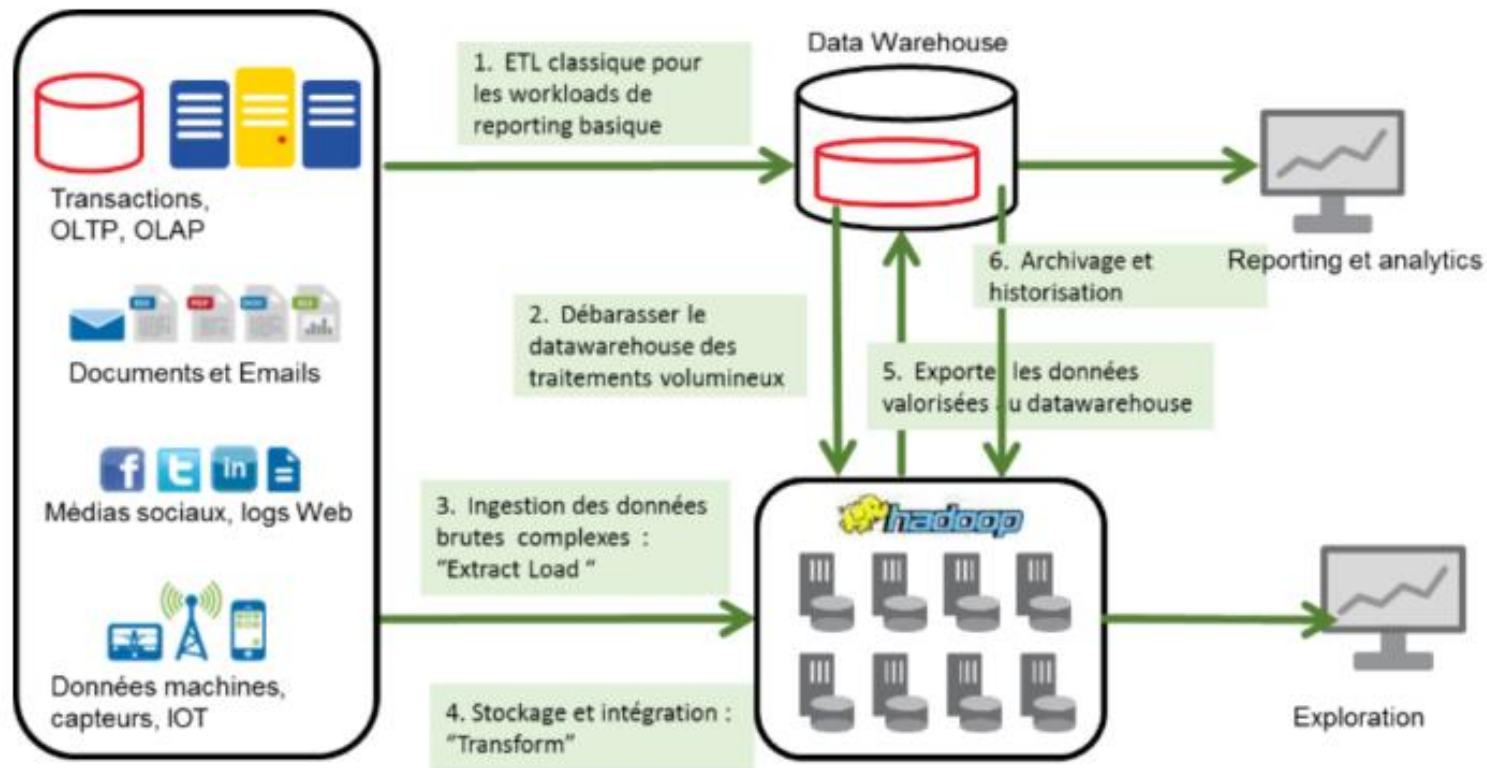
Caractéristiques



Caractéristiques



Processus Big data



La science de données

- Science de données : extraction intelligentes et efficace des connaissances à partir des big data
- *La* Science de données englobe les activités, outils et méthodes qui permettent d'exploiter les données dans tous les domaines (science, médecine, marketing ...)

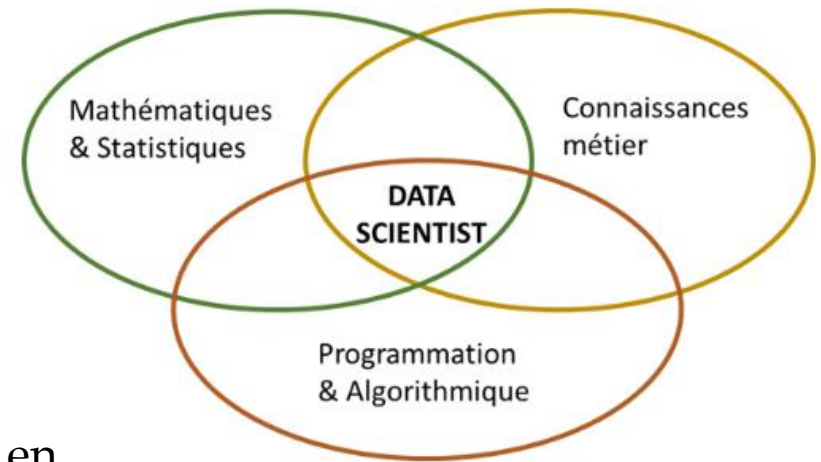
Data Scientist :

- un nouveau métier caractéristique du Big Data,
- On le retrouve en tête du classement des « jobs les plus sexy du 21^e siècle », publié par la [Harvard Business Review](#).

Un nouveau métier : le Data Scientist

- On associe trois compétences fortes chez un data scientist :
 - les méthodes mathématiques et statistiques,
 - la programmation
 - la compréhension des enjeux métier.

- On distingue deux catégories de *data scientists* :
 - Les **data architects** : définir la plateforme technique et les solutions logicielles adaptées.
 - Les **data analysts** : prendront la suite en appliquant des algorithmes prédictifs



Domaines d'application

Figure 7: Big Data Analytics Market Size by Business Category



Source: Heavy Reading

Big Data & Marketing prédictif

- Marketing : Prédiction basée sur l'intuition et l'irrationnel
 - sélection arbitraire de quelques facteurs, qui doivent, permettre de créer des segments ou des scoring pertinents : l'âge, le sexe...
- Marketing prédictif : des prévisions basées sur des données et des probabilités.
 - traitement en temps réel d'un grand volume de données : connaissance et définition des besoins et des attentes des clients



CE QUE LES ENTREPRISES Y GAGNENT

AMÉLIORATION DE L'EXPÉRIENCE
CONSOMMATEUR



MEILLEURE CAPACITÉ
OPÉRATIONNELLE

MONÉTISATION DES
DONNÉES

CE QUE LES CONSOMMATEURS APPRÉCIENT

OFFRES TRANSACTIONNELLES

(cadeaux, bons de réduction..)



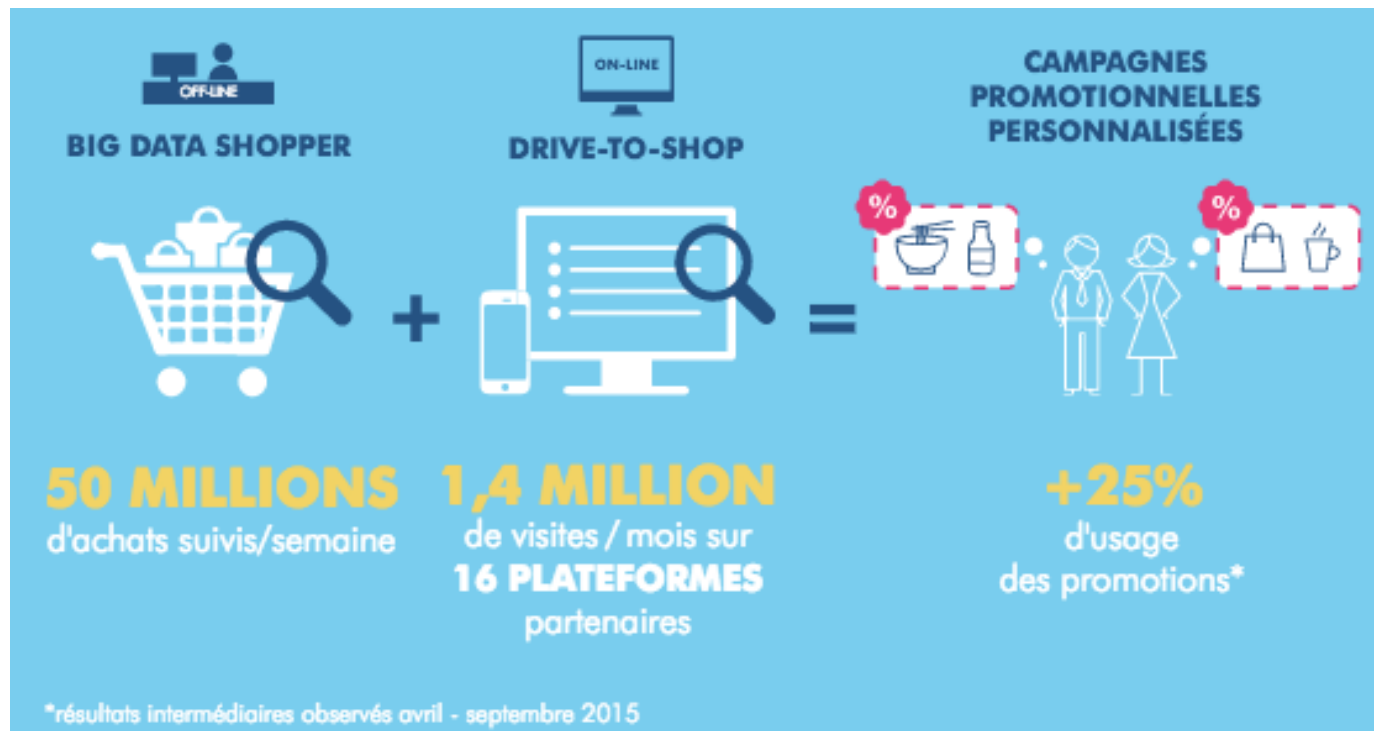
OFFRE PERSONNALISÉE

(sélection de produits sur mesure,
Choix moyen de comm)

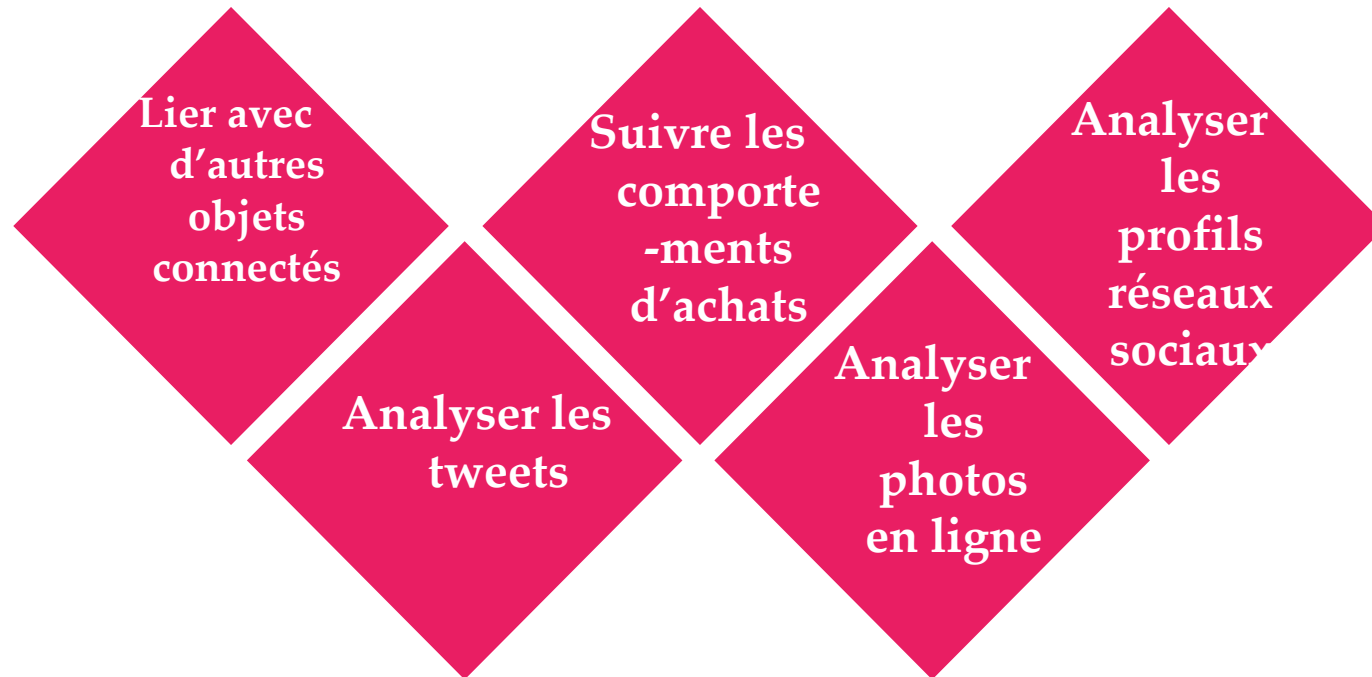
SERVICES INNOVANTS

(applications smart phones...)

Big Data & Marketing prédictif



Comment ?



Exemple : shoes bar

- Avant la visite
 - Interpeller la clientèle Cibler et recruter de nouveaux clients
 - Inciter à la visite avec des actualités et des incentives (couponing, jeux)
- Pendant le shopping
 - Guider le visiteur dans la boutique
 - Offrir des informations supplémentaires
 - Permettre l'achat ou la commande sans contact
 - Proposer des expériences d'achat interactives et inédites via des écrans digitaux
- Après la visite
 - Prolonger l'expérience d'achat
 - Retargeter en vue d'une visite future
 - Proposer des contenus additionnels pour favoriser la préférence de marque

MEILLEURE CONNAISSANCE CLIENT

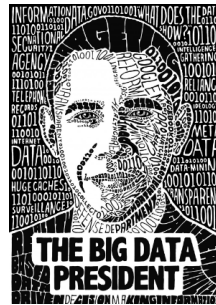
GEOLOCALISATION

PARCOURS CLIENT

HABITUDES D'ACHAT PRÉFÉRENCES

Objectif: Proposer le bon produit, au bon moment, au bon endroit et au bon client

Big Data & politique : Election présidentielle



Big Data

92 %

Octobre 2012

Sondage

45 %

Big Data & politique : Autres exemples

- Élection présidentielle française de 2012 : François Hollande
- La Tunisie : élection présidentielle 2014 <http://tn.webradar.me/>
- ...

Big data et administration publique

- Dans l'administration publique, des quantités extraordinaires de données sont accumulées au cours de l'exécution des services publics :
 - La gestion des prestations d'aide sociale et de la santé publique,
 - La délivrance des passeports et permis de conduire.
 - La gestion des taxes et recettes ...

Big data et sport

- le Big Data: arme secrète de l'Allemagne au Mondial de football

<http://www.01net.com/editorial/623742/le-big-data-le-douzieme-homme-de-lequipe-allemande-de-foot/>



Big data et sport






- Les paris sportifs : <https://www.numberfire.com/>



Big Data et crime

- Blue C.R.U.S.H. (Crime Reduction Utilizing Statistical History) est un logiciel qui prélève et rassemble avec l'aide de caméras et des forces de police un maximum de données sur les délits qui surviennent dans un territoire.
- Il s'agit d'envoyer les policiers dans les « hot spots »; là où la probabilité qu'un crime survienne est la plus élevée, et ainsi arrêter un délit avant qu'il ne se produise.
- Depuis son lancement il y a 7 ans,
 - le nombre de meurtres et de cambriolages a diminué de 36% à [Memphis](#).
 - Le vol de véhicules motorisés a chuté de 55% !

Détection des fraudes

Vertical	Type of Fraud	Pattern of Fraud
 Financial Services	Account takeover	Many transactions between \$9–\$10K
 Healthcare	Physician billing	Physician billing for drugs outside their expertise area
 E-tailing	Account takeover	Many accounts accessed from one IP
 Telecom	Roaming abuse	Excessive roaming on partner network by unlimited use customers
 Online Education	Student loan fraud	Student IP in “high-risk” country and student absent from classes and assignments

Autres exemples :

- Department de la santé et services de l'humanité.
- Institut national de santé: améliorer l'utilisation de l'imagerie dans les recherche sur le cancer
- Département d'énergie : permettre d'obtenir des observations précises des phénomènes atmosphériques.

Challenges : Entreprise

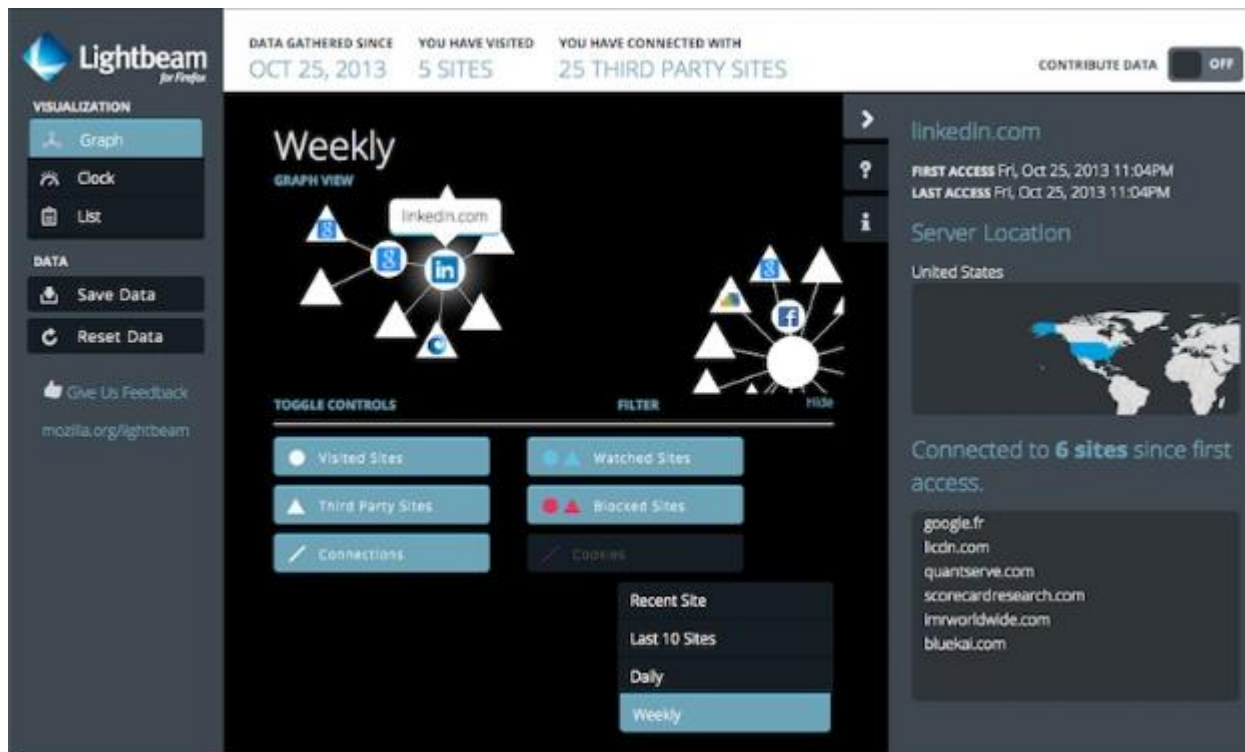
- La croissance des données entraîne en particulier une hausse des coûts du matériel, du logiciel, de la maintenance associée, de l'administration et des services.
- Le Big Data exige un nouvel ensemble de compétences au sein de l'entreprise.
- Les projets d'analyse Big Data nécessitent des équipes multidisciplinaires, et une collaboration active doit être engagée entre le service informatique et les **data scientists**.

Challenges : Sécurité



Challenges : Sécurité

- Il y a des sites qui nous suivent discrètement lorsqu'on navigue sur le web.



Challenges : Sécurité

- L'Open Data Santé : rendre publiques les données de la sécurité sociale
- Détection localisation et recherche web
- Détection transaction bancaire

➡ Respect de la liberté civile.