Introduction au Big Data

Hmida HMIDA

10 octobre 2022

Big Data

- Origine
  - 1997 : Cox et Ellsworth, "Application-controlled demand paging for out-of-core visualization" Proc. of the IEEE 8th conference on Visualization.
  - 2001 : Doug Laney, analyste du Meta Group puis Gartner, "3D Data Management: Controlling Data Volume, Velocity, and Variety."
  - 2004 : MapReduce
  - 2006 : Hadoop
  - 2010 : Spark
  - 2013 : Terme introduit dans Oxford English Dictionary

Big Data

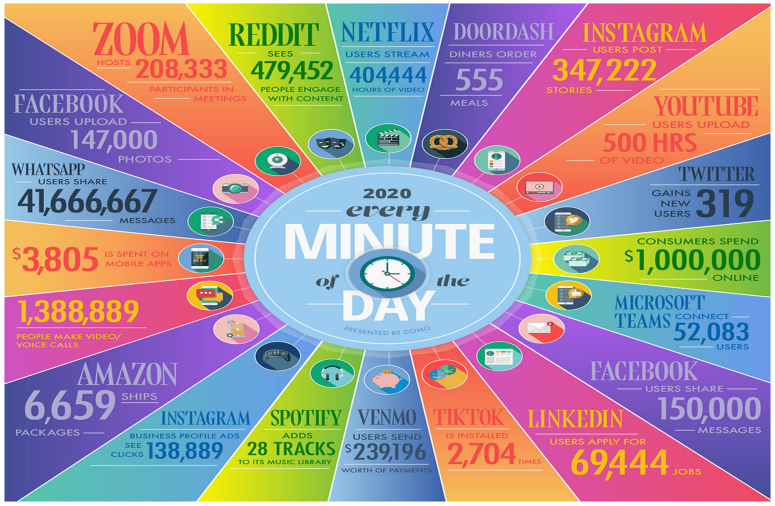
- Définition
  - Oxford English Dictionary : "sets of information that are too large or too complex to handle, analyse or use with standard methods"
  - Big Data :
    - ensemble de **technologies**, d'**architectures**, d'**outils** et de **procédures** permettant à une organisation de très rapidement **capter**, **traiter** et **analyser** de **larges quantités** et contenus **hétérogènes** et **changeants**, et d'en **extraire** les **informations pertinentes** à un **coût accessible**.
    - Génération massive de données complexes,
    - Stockage,
    - Extraction,
    - Analyse.

Big Data

- Chaque jour, nous générons 2,5 trillions d'octets de données
- 90% des données dans le monde ont été créées au cours des deux dernières années
- 90% des données générées sont non structurées
- Source:
  - Capteurs utilisés pour collecter les informations climatiques
  - Messages sur les médias sociaux
  - Images numériques et vidéos publiées en ligne
  - Enregistrements transactionnels d'achat en ligne
  - Signaux GPS de téléphones mobiles
  - ...
- Données appelées Big Data ou Données Massives

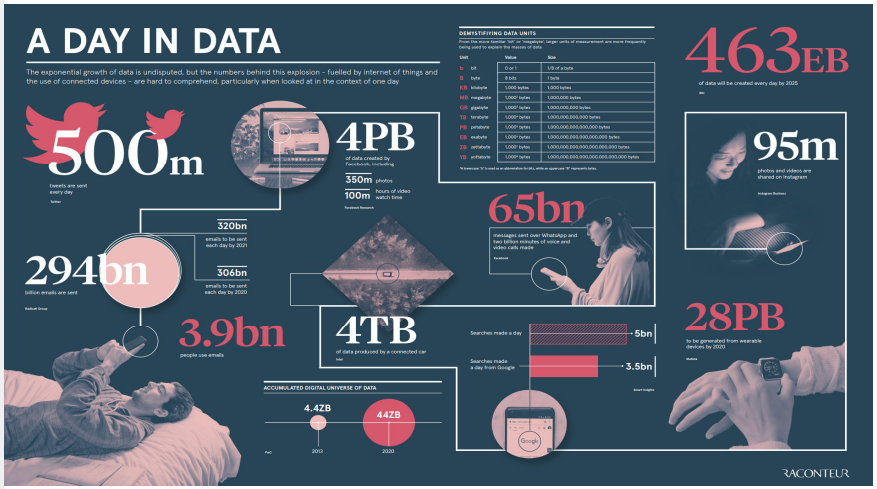
5

1 minute de données



6

Un jour de données



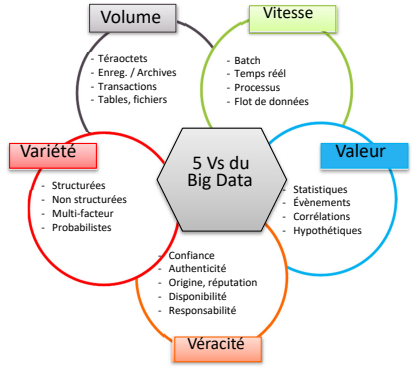
7

Types de « Big Data »

8

Dimensions du Big Data

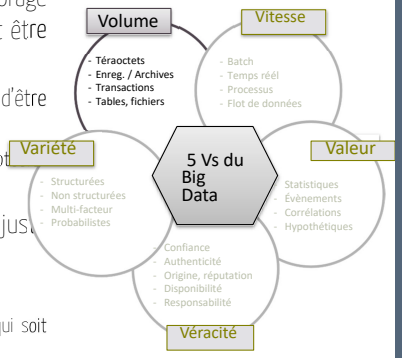
- Les Vs : 3 Vs, 4 Vs, …..
- Les 5 Vs
  - Volume (Volume)
  - Variété (Variety)
  - Vitesse (Velocity)
  - Véracité (Veracity)
  - Valeur (Value)



9

Volume

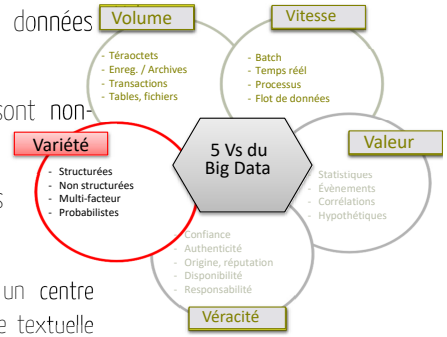
- Diminution du coût de stockage
- Les lieux de stockage fiables (comme des SAN: Storage Area Network) ou réseaux de stockage peuvent être très coûteux
  - Comment déterminer les données qui méritent d'être stockées
    - Transactions? Logs? Métier? Utilisateur? Capt Médicales? Sociales?
- ➔ Aucune donnée n'est inutile. Certaines n'ont jus pas encore servi.
  - Problèmes:
    - Comment stocker les données dans un endroit fiable, qui soit moins cher
    - Comment parcourir ces données et en extraire des informations facilement et rapidement?



10

Variété

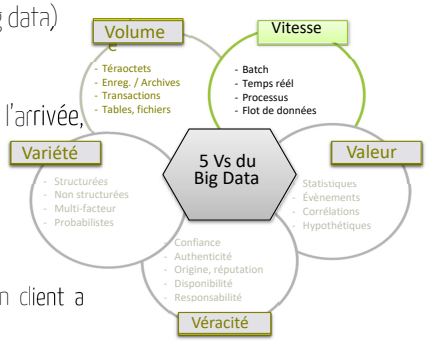
- Pour un stockage dans des bases de données ou dans des entrepôts de données, les données doivent respecter un format prédéfini.
- La plupart des données existantes sont non-structurées ou semi-structurées
- Données sous plusieurs formats et types
- On veut tout stocker:
  - Exemple: pour une discussion dans un centre d'appel, on peut la stocker sous forme textuelle pour son contenu, comme on peut stocker l'enregistrement en entier, pour interpréter le ton de voix du client



11

Vitesse

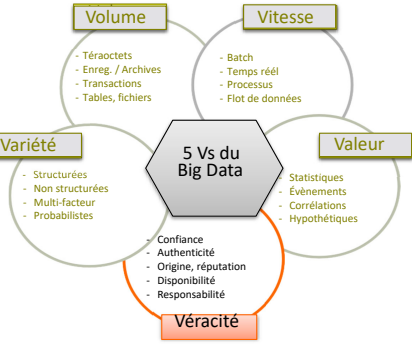
- Rapidité d'arrivée des données (Streaming data)
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des Teraoctets par jour
- Possibilité de traitement à la volée
- Exemple
  - Il ne suffit pas de savoir quel article un client a acheté ou réservé
  - Si si on sait que vous avez passé plus de 5mn à consulter un article dans une boutique d'achat en ligne, il est possible de vous envoyer un email dès que cet article est soldé.



12

Véracité

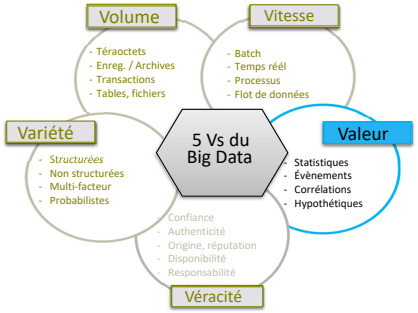
- Cela fait référence à la qualité et à la fiabilité des données.
  - Données bruitées, falsifiées, imprécises ...
  - Avec l'augmentation de la quantité, la qualité et précision se perden (abréviations, typos, déformations, source peu fiable...)
- Les solutions Big Data doivent remédier à cela en se référant au volume des données existantes
- Nécessité d'une (très) grande rigueur dans l'organisation de la collecte et le recoupement, croisement, enrichissement des données



13

Valeur

- Le V le plus important
- Il faut transformer toutes les données en valeurs exploitables :
  - les données sans valeur sont inutiles
- Atteindre des objectifs stratégiques de création de valeur pour les clients et pour l'entreprise dans tous les domaines d'activité



14

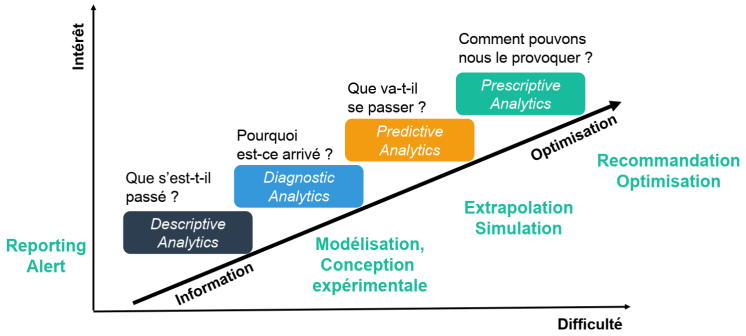
Défis

- Stocker ce volume de données dans les ressources disponibles ?
- Accéder aux données rapidement ?
- Combiner les différents formats ?
- Traiter ces données d'une manière performante, tolérante aux pannes et flexible ?
- Extraire les connaissances de façon interactive et rentable ?

15

Data Analytics

- Inspection, nettoyage, transformation et modélisation des données
  - Découverte d'informations pertinentes
  - Aide à la décision

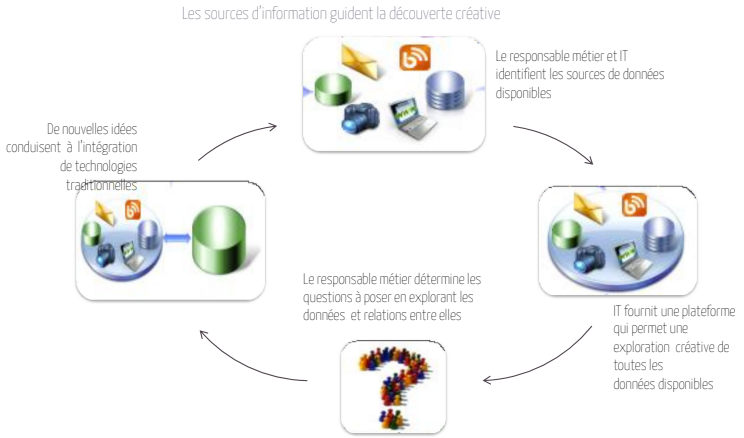


16

Approche traditionnelle

17

Approche Big Data



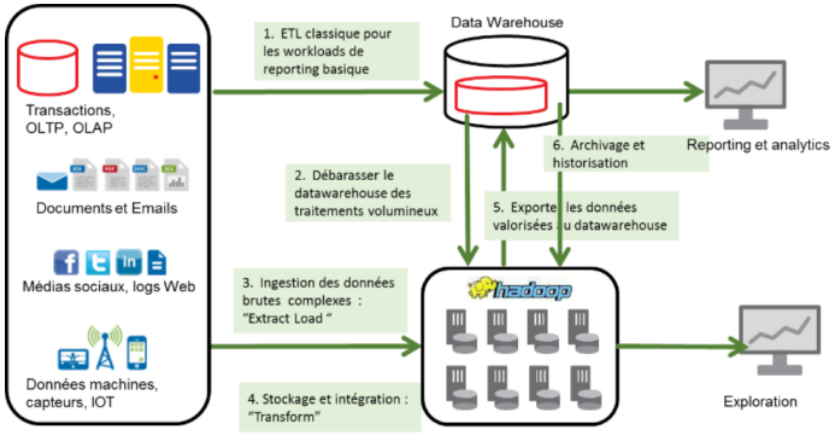
18

Big Data Analytics

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

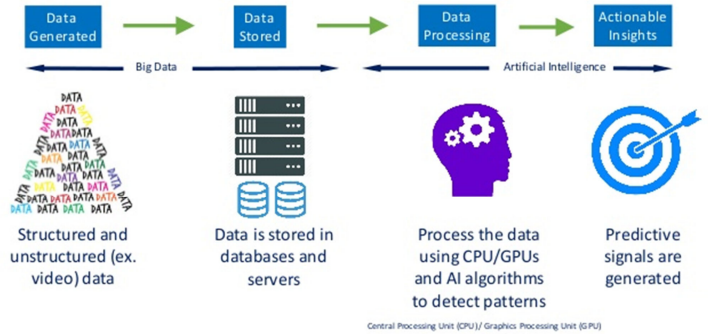
19

Intégration des approches



20

Big Data Analytics



21

Domaines d'application

22

Applications

	- Calculer le temps d'arrivée - Utilise les conditions du traffic pendant une fenêtre de temps
amazon	- Prédiction des achats en se basant sur l'historique des clients - Les clics, l'ajout au panier - Anticiper la livraison - Recommandation de produits
NETFLIX	- Recommandation de films
	- Analyser les données climatiques archivées - Prédire le temps

23

Applications (2)

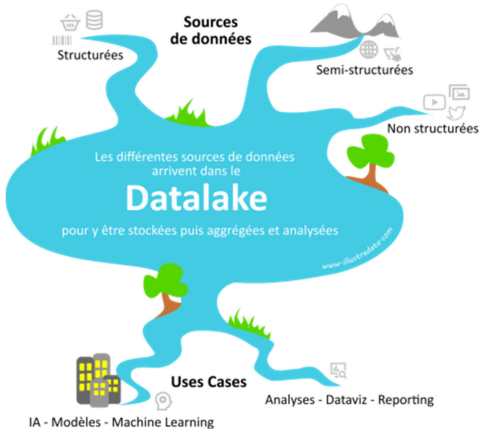
SAMSUNG	- Analyse des conversation sur les réseaux sociaux au moment du lancement du Iphone 5 en 2012 - Lancement du Galaxy S3 en 2013 (Best Smartphone award)
DEUTSCHER FUSSBALL-BUND	- Coupe du monde 2014 - Capteurs dans le protège-tibia - Analyse du comportement des joueurs - Simulation de situations de jeu
IBM	- Blue C.R.U.S.H. (Crime Reduction Utilizing Statistical History) - Il s'agit d'envoyer les policiers dans les « hot spots » - le nombre de meurtres et de cambriolages a diminué de 36% à Memphis. - Le vol de véhicules motorisés a chuté de 55% !

24

Data Science – Data Scientist

- Science de données
  - Englobe les activités, outils et méthodes qui permettent d'exploiter les données dans tous les domaines (science, médecine, marketing ...)
- 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 :
    - Data Architect : définir la plateforme technique et les solutions logicielles adaptées.
    - Data Analyst : prendront la suite en appliquant des algorithmes prédictifs

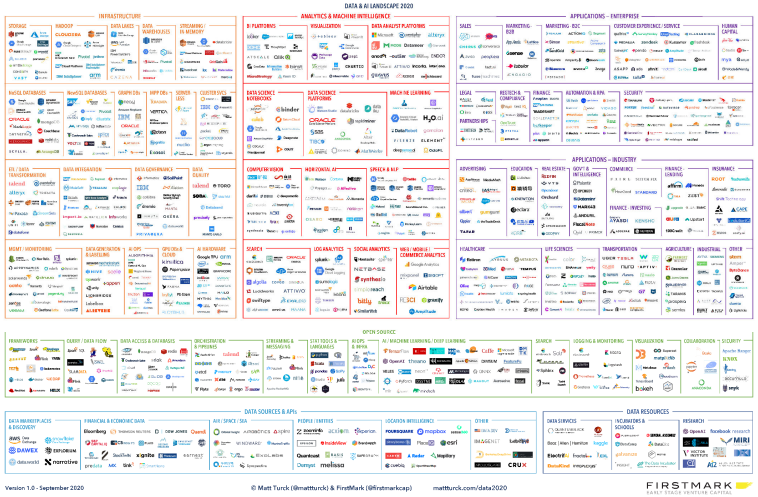
Datalake



Datalake

Caractéristique	Data Warehouse	Data Lake
Données	- Structurées (Relationnelles) - Système transactionnel, BD opérationnelle - Prétraitées	- Structurées, semi-structurées, non structurées - IoT, sites web, applications mobile, réseaux sociaux, … - Brutes (raw)
Schéma	- Conçu avant l'implémentation du DW - Schema-on-write	- Créé au moment de l'analyse - Schema-on-read
Coût/Performance	- Requêtes très rapides - Stockage coûteux	- Requêtes de plus en plus rapides - Stockage moins coûteux
Utilisateurs	Business Analysts	Data scientists, Data Developpers, Business Analysts
Analytics	Reporting, DataViz	Machine Learning, Analyse prédictive

Big Data and AI Landscape



Sources et Bibliographie

- Cours :
  - Cours de Big Data, Stéphane Vialle, Centrale Supélec
  - Cours de Big Data, Lilia Sfaxi, INSAT
  - Technologies pour le Big Data, Minyar Sassi Hidri, ENIT
- Livres :
  - Big Data Analytics with Hadoop 3, Sridhar Alla