

# Introduction au Big Data



Hmida HMIDA

20.11.22

2

## 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

3

## 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.

4

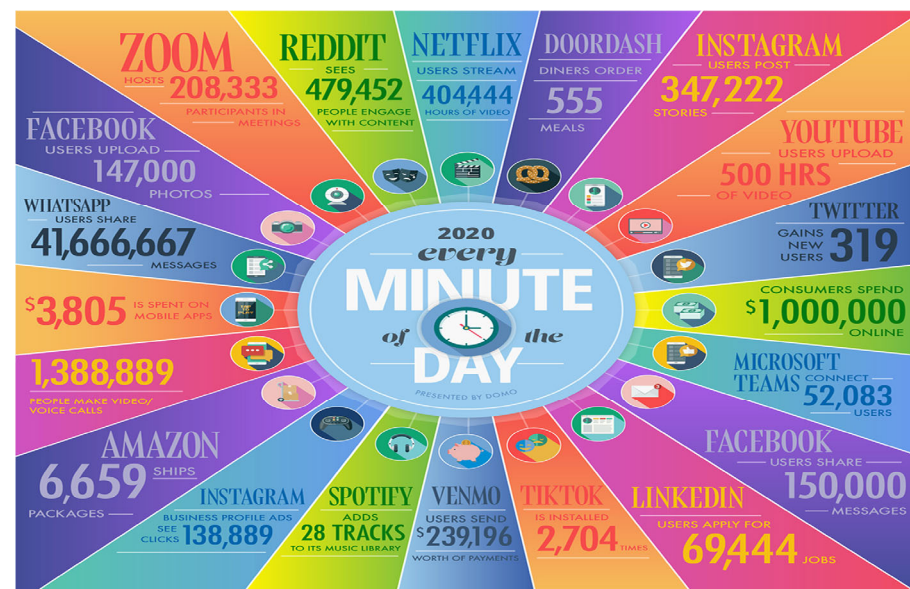
## Big Data



- > 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

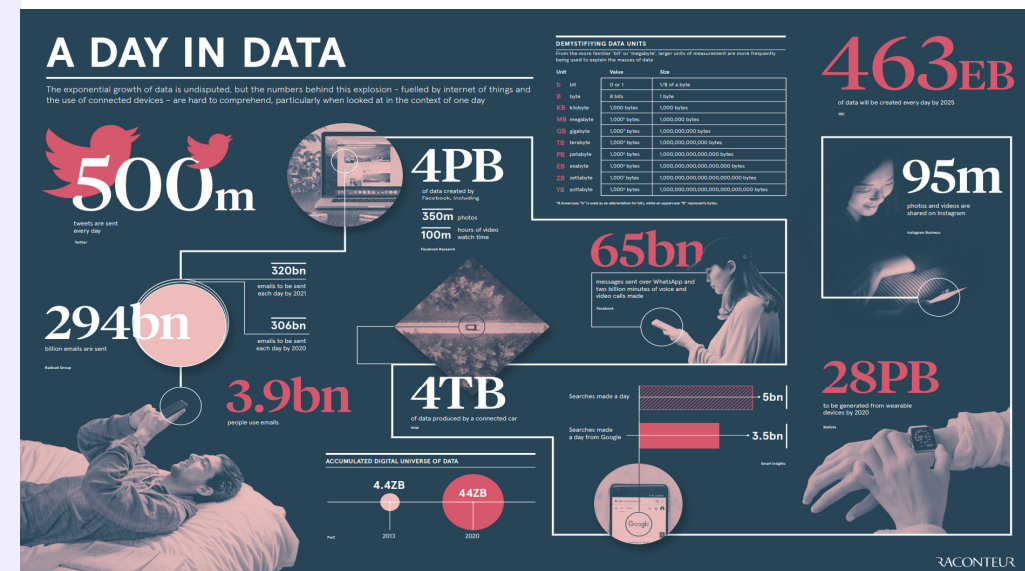
5

# 1 minute de données



6

# Un jour de données



7

# Types de « Big Data »



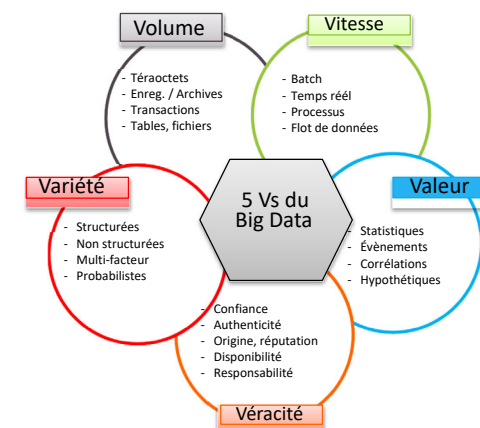
8

# Dimensions du Big Data

➤ Les Vs : 3 Vs, 4 Vs, ....

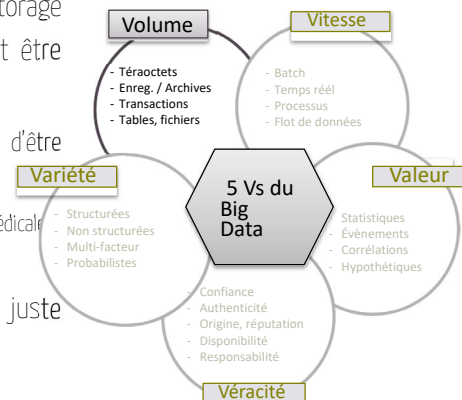
➤ Les 5 Vs

- ✓ Volume (Volume)
- ✓ Variété (Variety)
- ✓ Vitesse (Velocity)
- ✓ Véracité (Veracity)
- ✓ Valeur (Value)



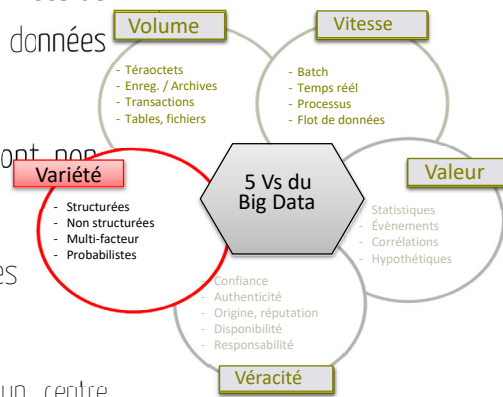
## Volume

- 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? Capteurs? Médical? Sociales?
- ➔ Aucune donnée n'est inutile. Certaines n'ont juste 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



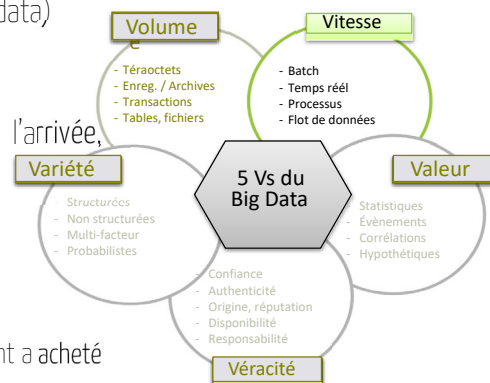
## Variété

- 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



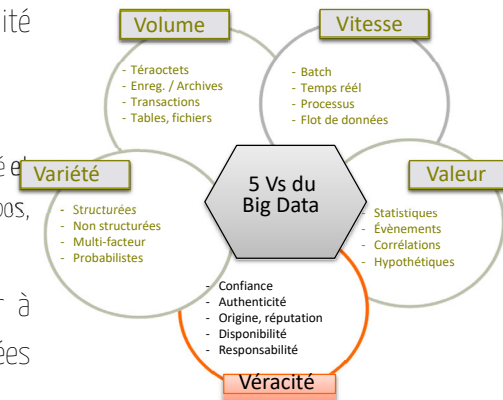
## Vitesse

- Rapidité d'arrivée des données (Streaming data)
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des Téraoctets 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.



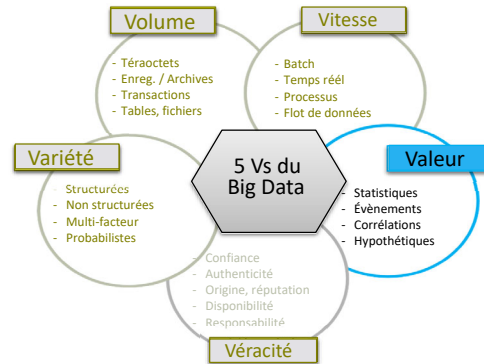
## Véracité

- 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 la précision se perdent (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



## 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

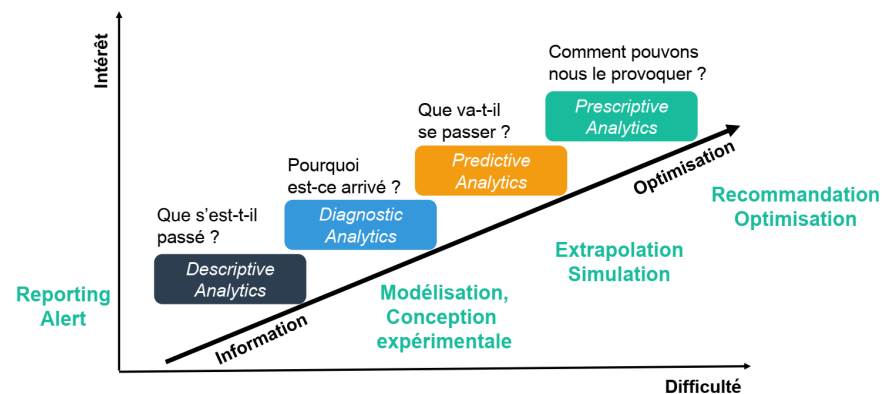
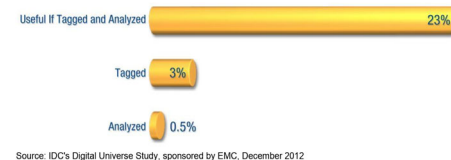


## Défis

- > 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 ?

## Data Analytics

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

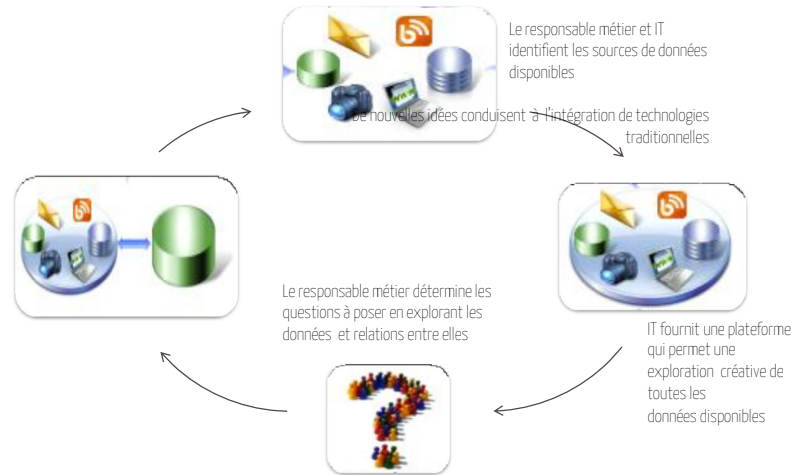


## Approche traditionnelle





Les sources d'information guident la découverte créative



## Traditional Analytics (BI)

## Big Data Analytics

### Focus on

- Descriptive analytics
- Diagnosis analytics

- **Predictive analytics**
- **Data Science**

### Data Sets

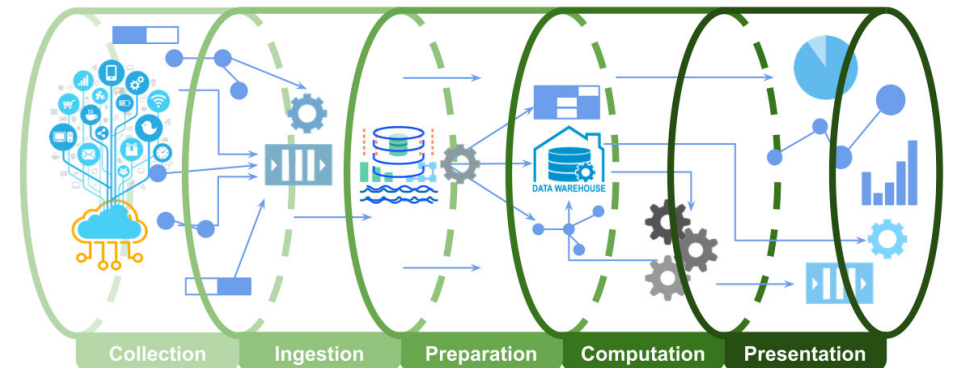
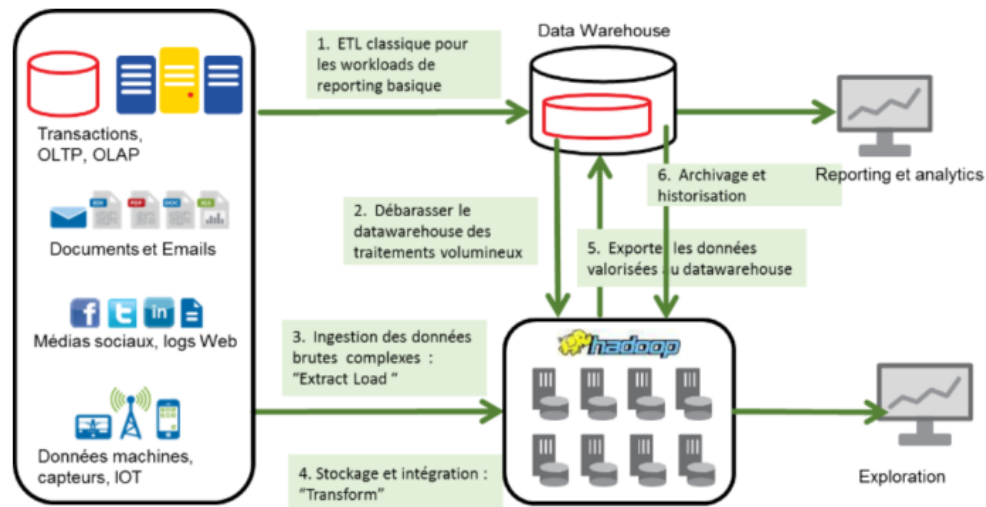
- Limited data sets
- Cleansed data
- Simple models

- 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





## Domaines d'application



<b>Banking</b> <ul style="list-style-type: none"> <li>Optimizing Offers and Cross-sell</li> <li>Customer Service and Call Center Efficiency</li> </ul>	<b>Insurance</b> <ul style="list-style-type: none"> <li>360° View of Domain or Subject</li> <li>Catastrophe Modeling</li> <li>Fraud &amp; Abuse</li> </ul>	<b>Telco</b> <ul style="list-style-type: none"> <li>Pro-active Call Center</li> <li>Network Analytics</li> <li>Location Based Services</li> </ul>	<b>Energy &amp; Utilities</b> <ul style="list-style-type: none"> <li>Smart Meter Analytics</li> <li>Distribution Load Forecasting/Scheduling</li> <li>Condition Based Maintenance</li> </ul>	<b>Media &amp; Entertainment</b> <ul style="list-style-type: none"> <li>Business process transformation</li> <li>Audience &amp; Marketing Optimization</li> </ul>
<b>Retail</b> <ul style="list-style-type: none"> <li>Actionable Customer Insight</li> <li>Merchandise Optimization</li> <li>Dynamic Pricing</li> </ul>	<b>Travel &amp; Transport</b> <ul style="list-style-type: none"> <li>Customer Analytics &amp; Loyalty Marketing</li> <li>Predictive Maintenance Analytics</li> </ul>	<b>Consumer Products</b> <ul style="list-style-type: none"> <li>Shelf Availability</li> <li>Promotional Spend Optimization</li> <li>Merchandising Compliance</li> </ul>	<b>Government</b> <ul style="list-style-type: none"> <li>Civilian Services</li> <li>Defense &amp; Intelligence</li> <li>Tax &amp; Treasury Services</li> </ul>	<b>Healthcare</b> <ul style="list-style-type: none"> <li>Measure &amp; Act on Population Health Outcomes</li> <li>Engage Consumers in their Healthcare</li> </ul>
<b>Automotive</b> <ul style="list-style-type: none"> <li>Advanced Condition Monitoring</li> <li>Data Warehouse Optimization</li> </ul>	<b>Chemical &amp; Petroleum</b> <ul style="list-style-type: none"> <li>Operational Surveillance, Analysis &amp; Optimization</li> <li>Data Warehouse Consolidation, Integration &amp; Augmentation</li> </ul>	<b>Aerospace &amp; Defense</b> <ul style="list-style-type: none"> <li>Uniform Information Access Platform</li> <li>Data Warehouse Optimization</li> </ul>	<b>Electronics</b> <ul style="list-style-type: none"> <li>Customer/ Channel Analytics</li> <li>Advanced Condition Monitoring</li> </ul>	<b>Life Sciences</b> <ul style="list-style-type: none"> <li>Increase visibility into drug safety and effectiveness</li> </ul>

## Applications



	<ul style="list-style-type: none"> <li>Calculer le temps d'arrivée</li> <li>Utilise les conditions du trafic pendant une fenêtre de temps</li> </ul>
	<ul style="list-style-type: none"> <li>Prédiction des achats en se basant sur l'historique des clients</li> <li>Les clics, l'ajout au panier</li> <li>Anticiper la livraison</li> <li>Recommandation de produits</li> </ul>
	<ul style="list-style-type: none"> <li>Recommandation de films</li> </ul>
	<ul style="list-style-type: none"> <li>Analyser les données climatiques archivées</li> <li>Prédire le temps</li> </ul>

## Applications (2)

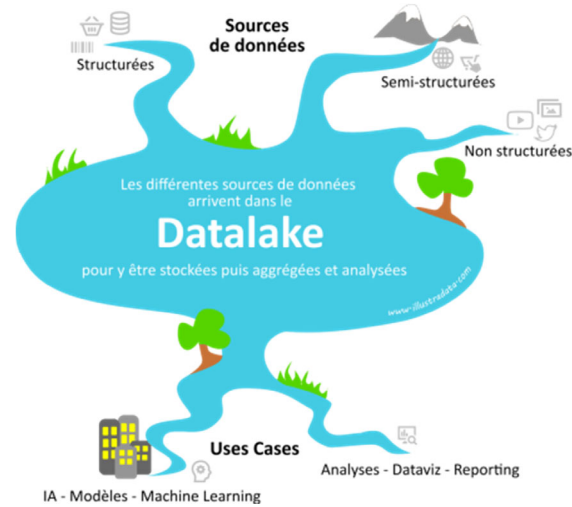


	<ul style="list-style-type: none"> <li>Analyse des conversations sur les réseaux sociaux au moment du lancement du iPhone 5 en 2012</li> <li>Lancement du Galaxy S3 en 2013 (Best Smartphone award)</li> </ul>
	<ul style="list-style-type: none"> <li>Coupe du monde 2014</li> <li>Capteurs dans le protège-tibia</li> <li>Analyse du comportement des joueurs</li> <li>Simulation de situations de jeu</li> </ul>
	<ul style="list-style-type: none"> <li>Blue C.R.U.S.H. (Crime Reduction Utilizing Statistical History)</li> <li>Il s'agit d'envoyer les policiers dans les « hot spots »</li> <li>le nombre de meurtres et de cambriolages a diminué de 36% à Memphis.</li> <li>Le vol de véhicules motorisés a chuté de 55% !</li> </ul>

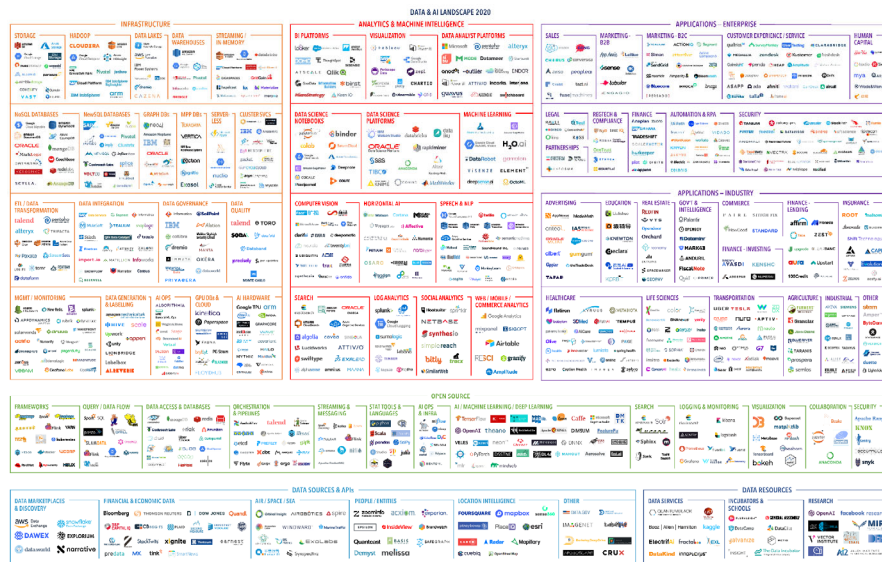
## Data Science – Data Scientist



- 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



Caractéristique	Data Warehouse	Data Lake
Données	<ul style="list-style-type: none"> <li>Structurées (Relationnelles)</li> <li>Système transactionnel, BD opérationnelle</li> <li>Prétraitées</li> </ul>	<ul style="list-style-type: none"> <li>Structurées, semi-structurées, non structurées</li> <li>IoT, sites web, applications mobile, réseaux sociaux, ...</li> <li>Brutes (raw)</li> </ul>
Schéma	<ul style="list-style-type: none"> <li>Conçu avant l'implémentation du DW</li> <li>Schema-on-write</li> </ul>	<ul style="list-style-type: none"> <li>Créé au moment de l'analyse</li> <li>Schema-on-read</li> </ul>
Coût/Performance	<ul style="list-style-type: none"> <li>Requêtes très rapides</li> <li>Stockage coûteux</li> </ul>	<ul style="list-style-type: none"> <li>Requêtes de plus en plus rapides</li> <li>Stockage moins coûteux</li> </ul>
Utilisateurs	Business Analysts	Data scientists, Data Developpers, Business Analysts
Analytics	Reporting, DataViz	Machine Learning, Analyse prédictive



### 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