

Introduction au Big Data



Hmida HMIDA

10 octobre 2022

> Origine

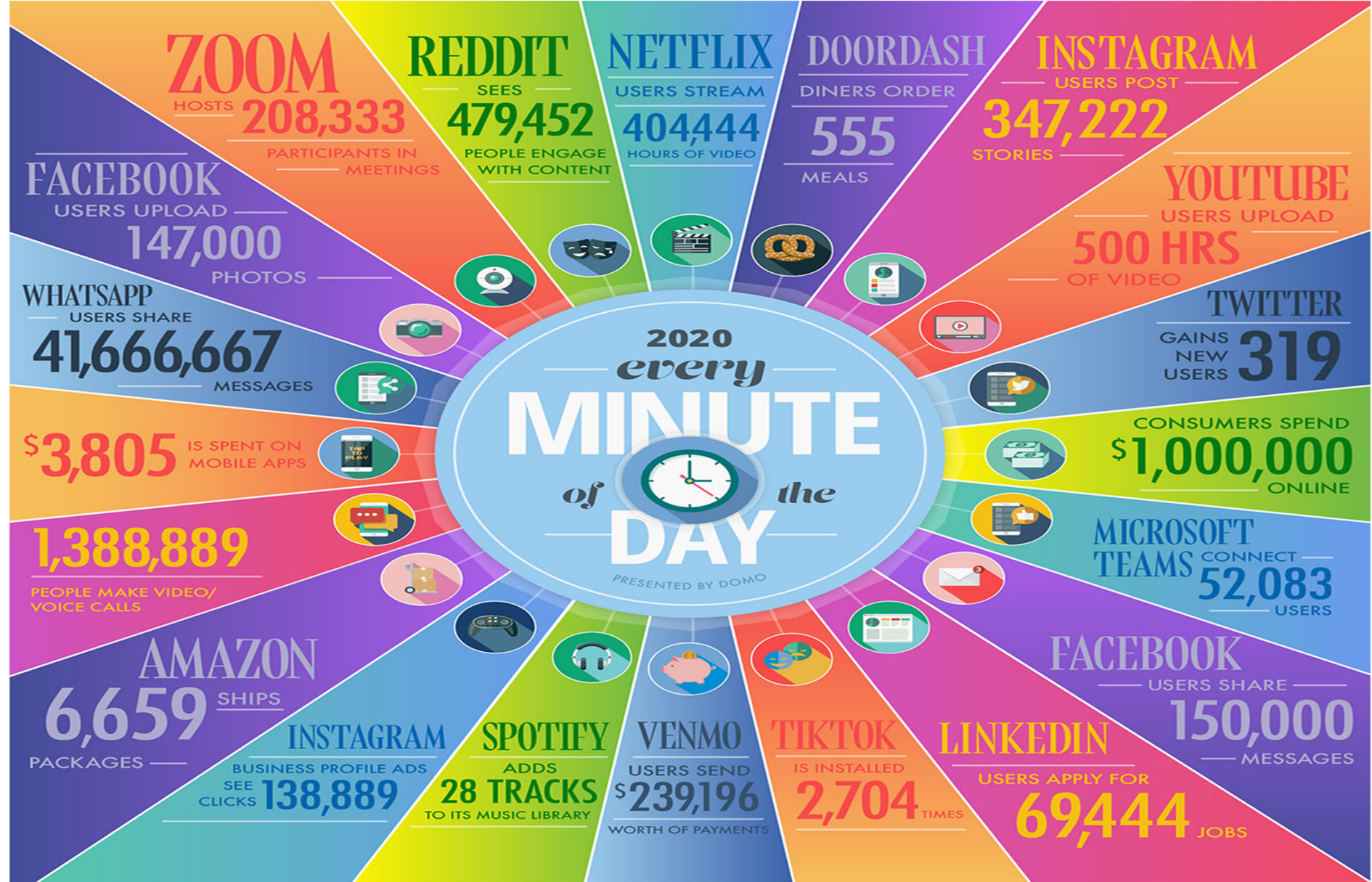
- ✓ 1997 : Cox et Ellsworth, “Application-controlled demand paging for out-of-core visualization” Proc. of the IEEE 8th conference on Visualization.
- ✓ 2001 : Doug Laney, analyste du Meta Group puis Gartner, “3D Data Management: Controlling Data Volume, Velocity, and Variety.”
- ✓ 2004 : MapReduce
-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

> 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.

- 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

1 minute de données



Un jour de données



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day
Twitter



4PB

of data created by Facebook, including

350m photos
100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

28PB

to be generated from wearable devices by 2020

Statista

Searches made a day **5bn**

Searches made a day from Google **3.5bn**

Smart Insights

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

PwC

2013

2020

Types de « Big Data »



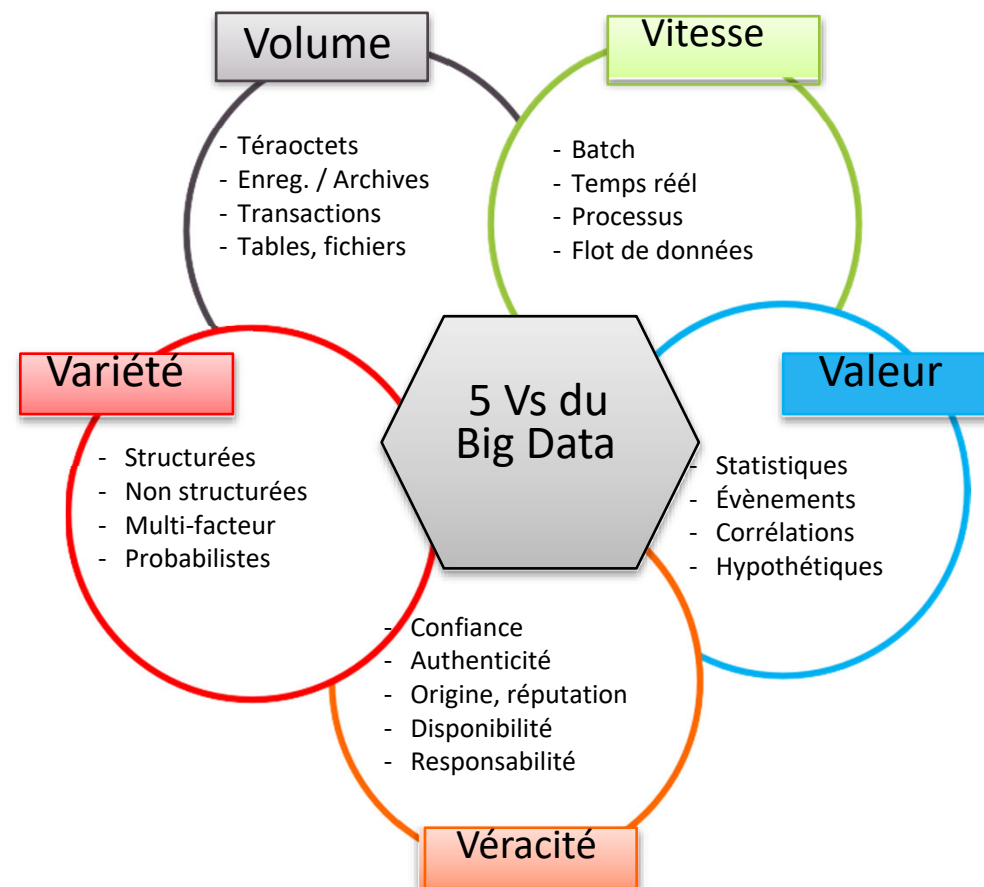
Dimensions du Big Data



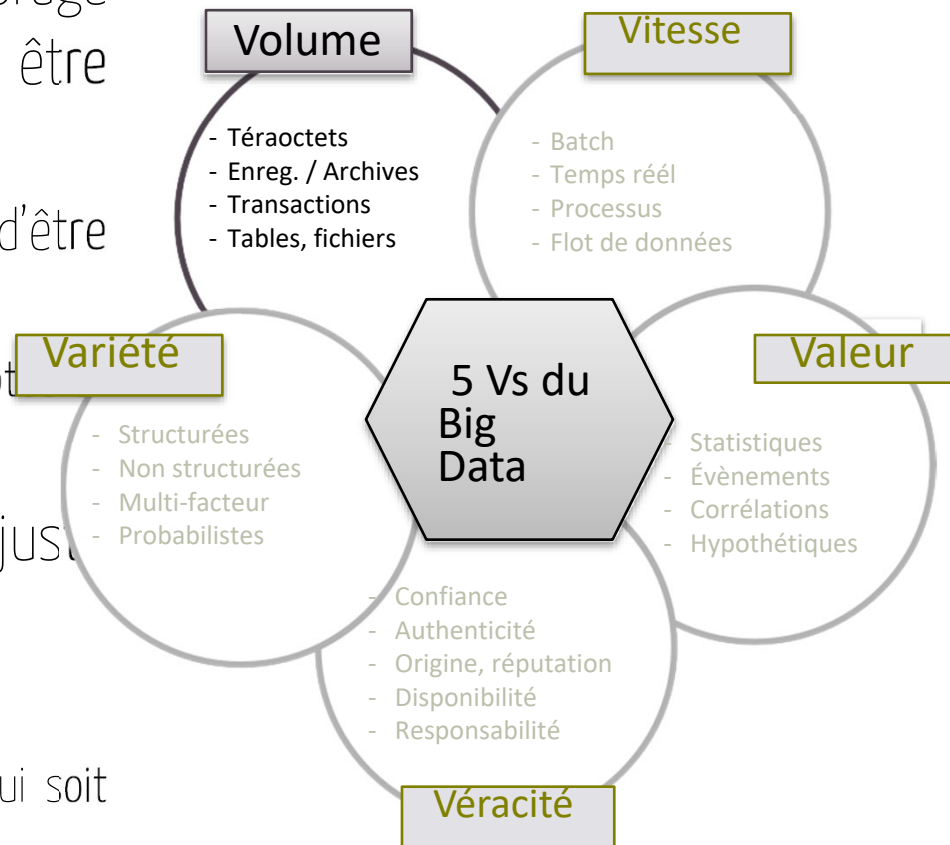
> Les Vs : 3 Vs, 4 Vs,

> Les 5 Vs

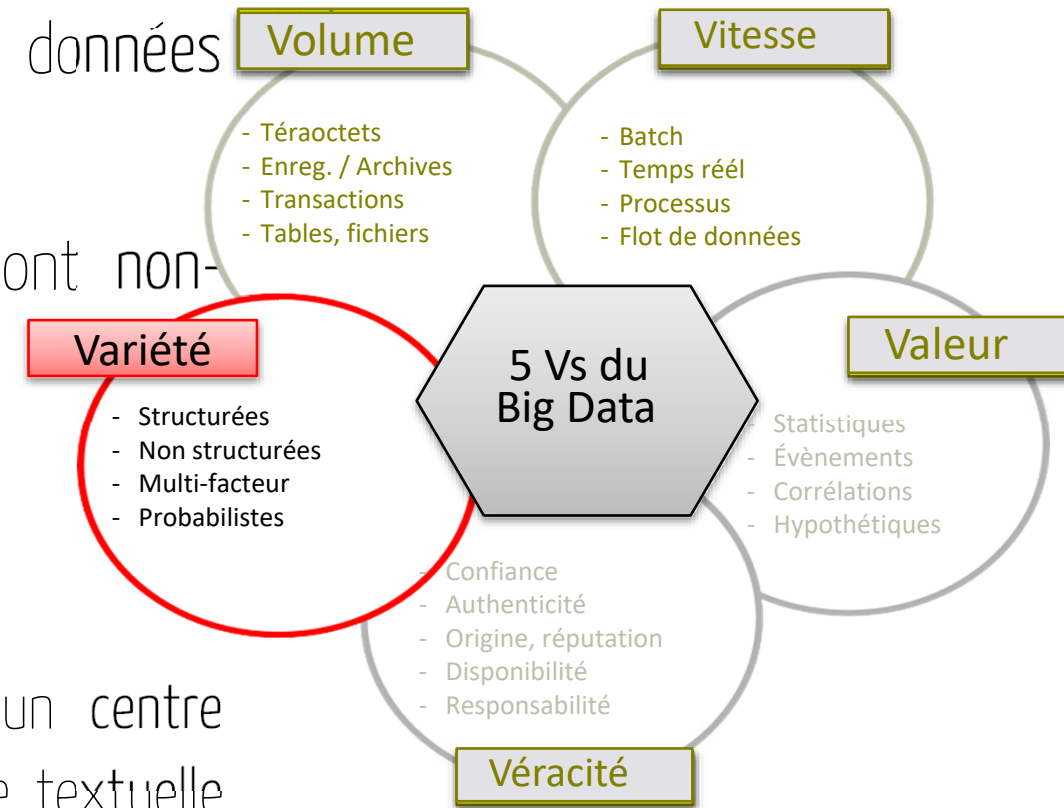
- ✓ Volume (Volume)
- ✓ Variété (Variety)
- ✓ Vitesse (Velocity)
- ✓ Véracité (Veracity)
- ✓ Valeur (Value)



- 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 just 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?



- 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



➤ Rapidité d'arrivée des données (Streaming data)

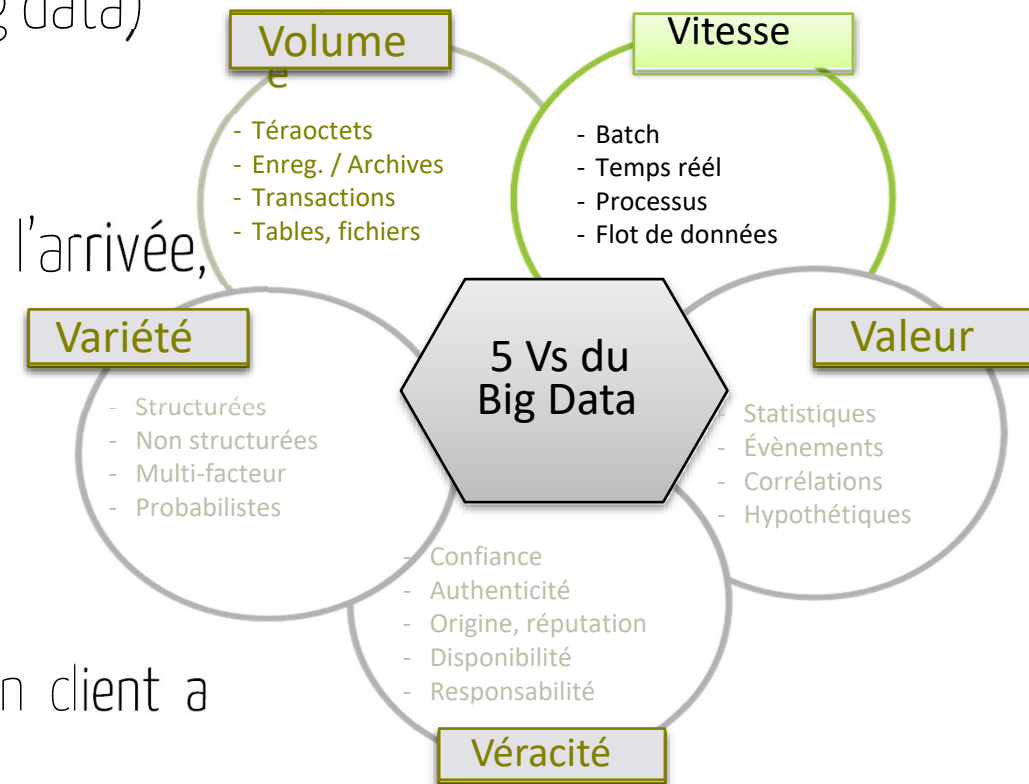
➤ Vitesse de traitement

➤ 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.

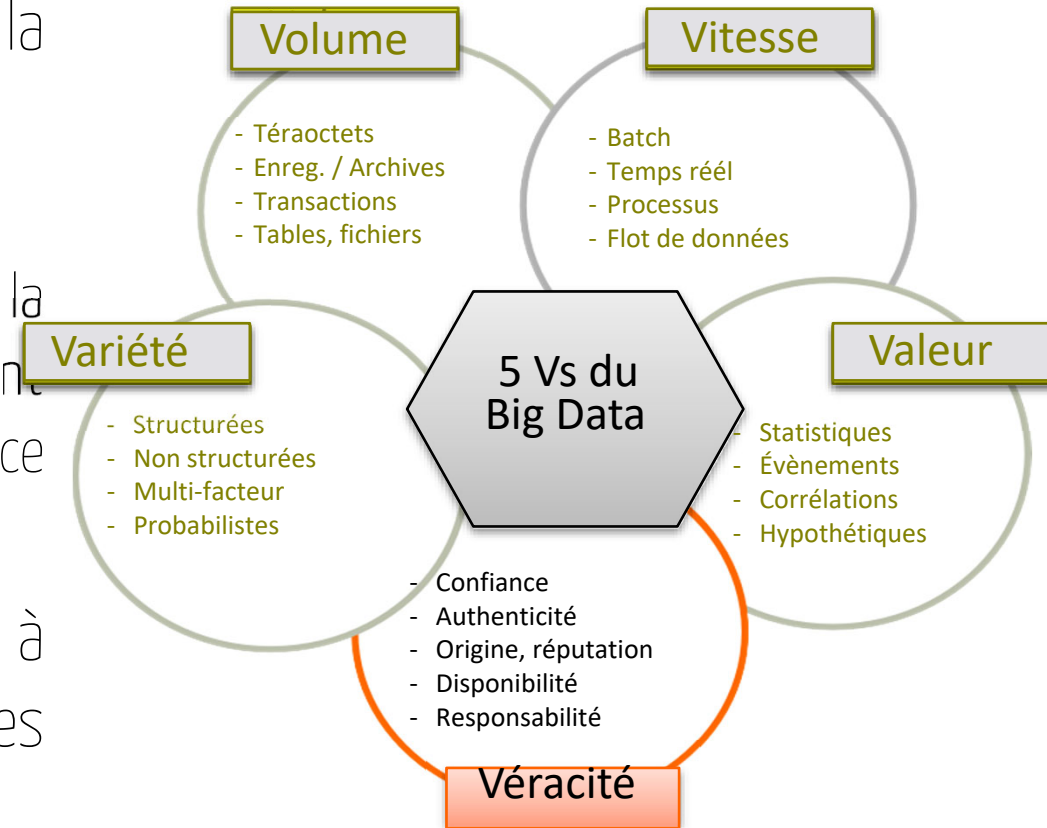


➤ Cela fait référence à la qualité et à la fiabilité des données.

- ✓ Données bruitées, falsifiées, imprécises ...
- ✓ Avec l'augmentation de la quantité, la qualité et précision se perdent (abréviations, typos, déformations, source peu fiable...)

➤ Les solutions Big Data doivent remédier à cela en se référant au volume des données existantes

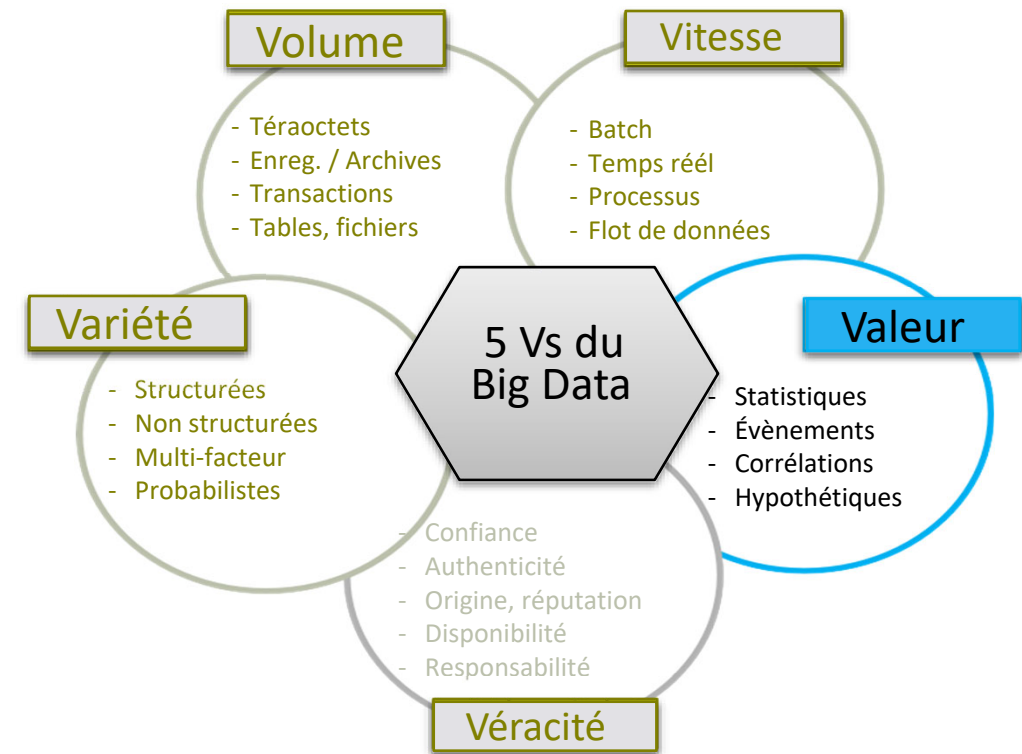
➤ Nécessité d'une (très) grande rigueur dans l'organisation de la collecte et le recoupement, croisement, enrichissement des données



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



- 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

Useful If Tagged and Analyzed

23%

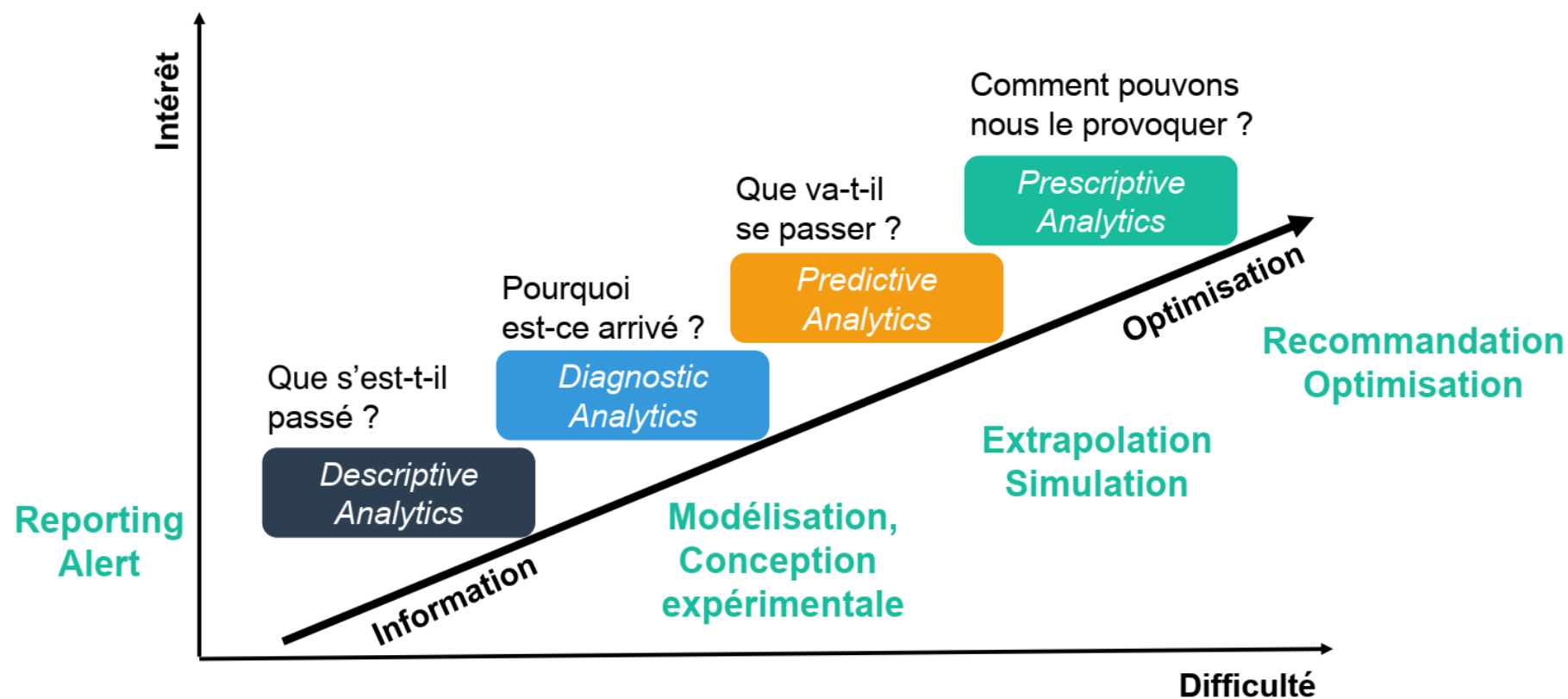
Tagged

3%

Analyzed

0.5%

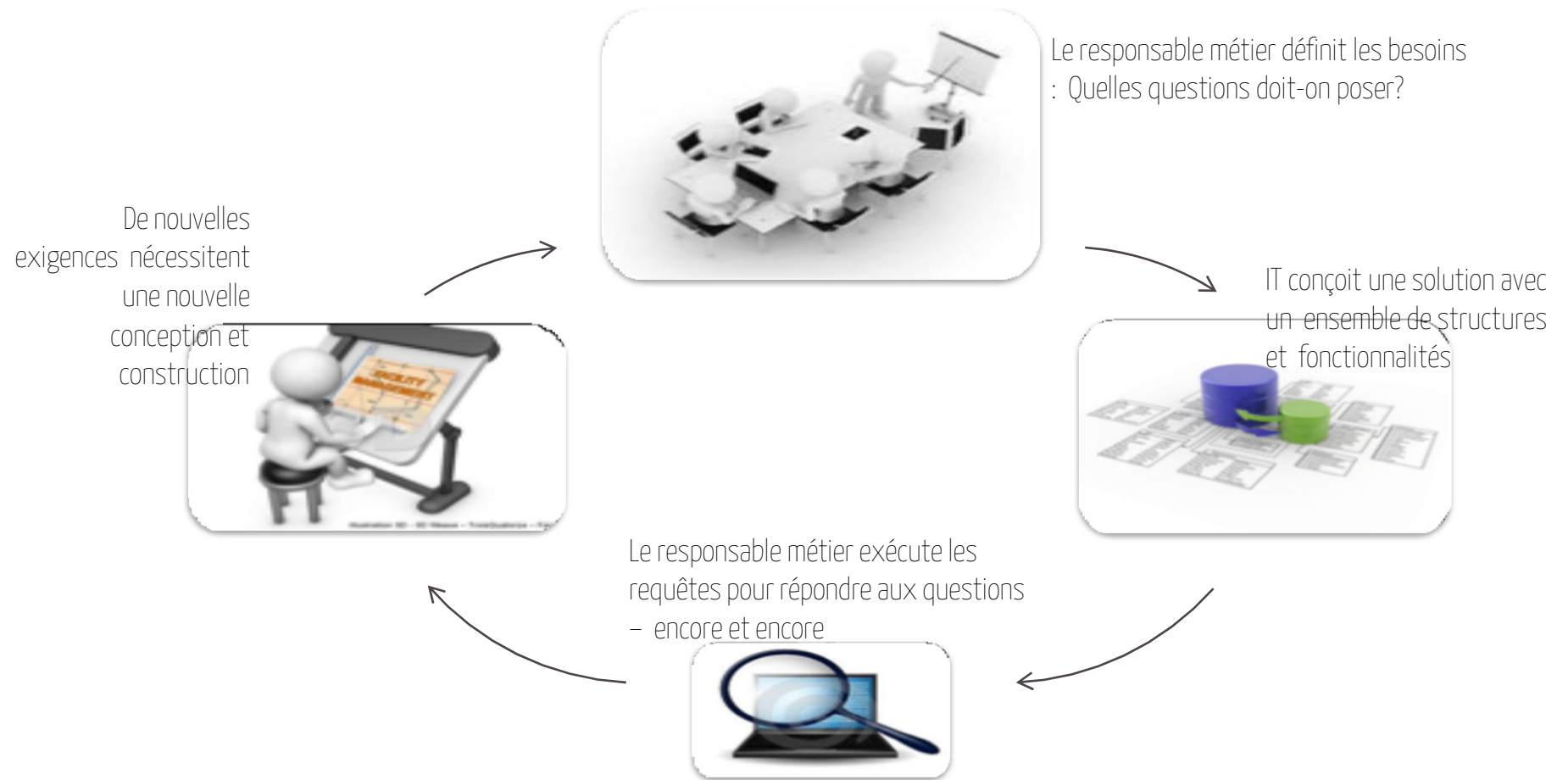
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012



Approche traditionnelle



Les besoins métier guident la conception de la solution

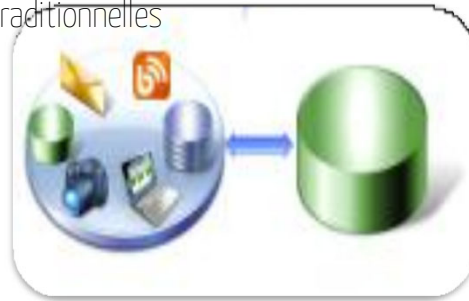


Approche Big Data



Les sources d'information guident la découverte créative

De nouvelles idées
conduisent à l'intégration
de technologies
traditionnelles



Le responsable métier et IT
identifient les sources de données
disponibles



IT fournit une plateforme
qui permet une
exploration créative de
toutes les
données disponibles

Le responsable métier détermine les
questions à poser en explorant les
données et relations entre elles

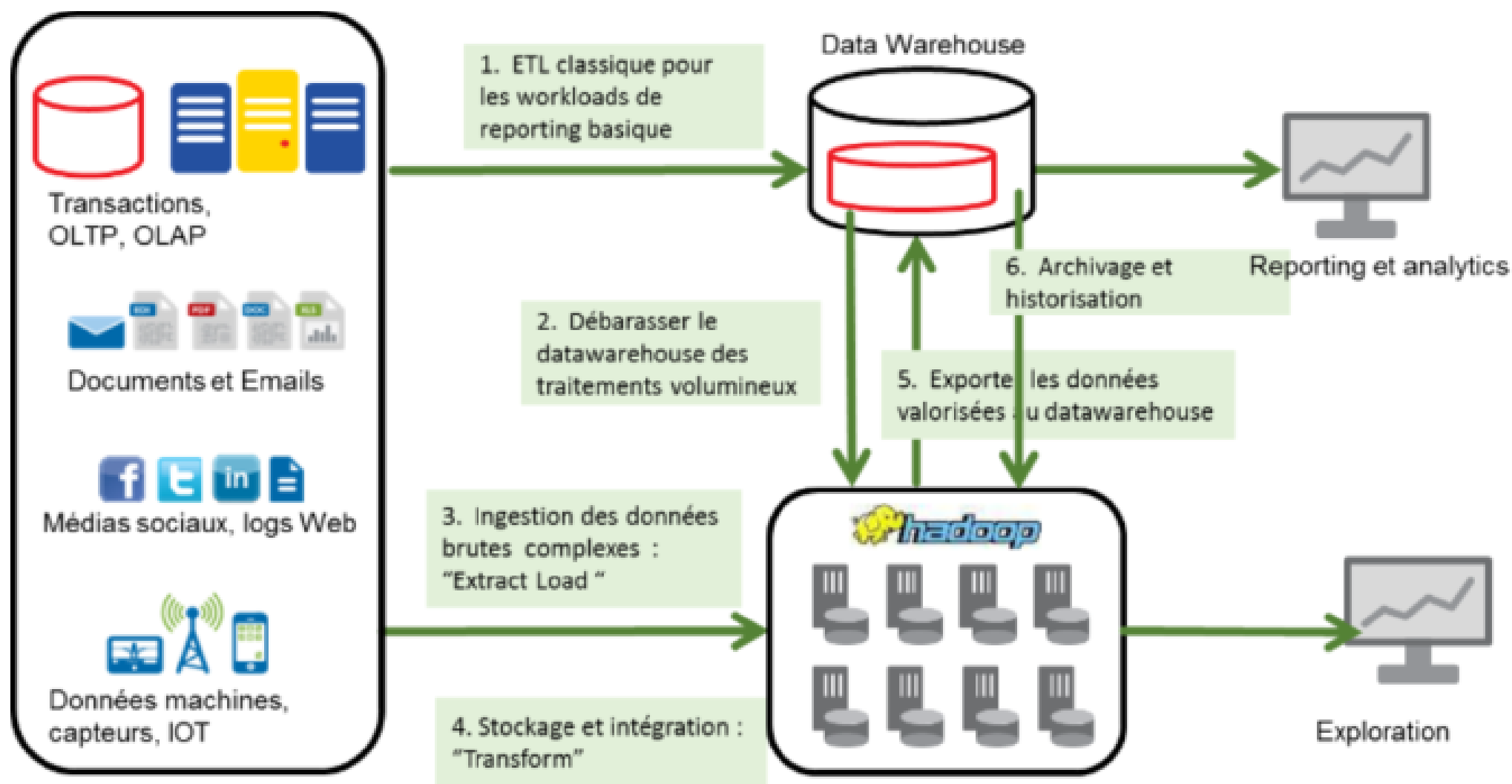


Big Data Analytics

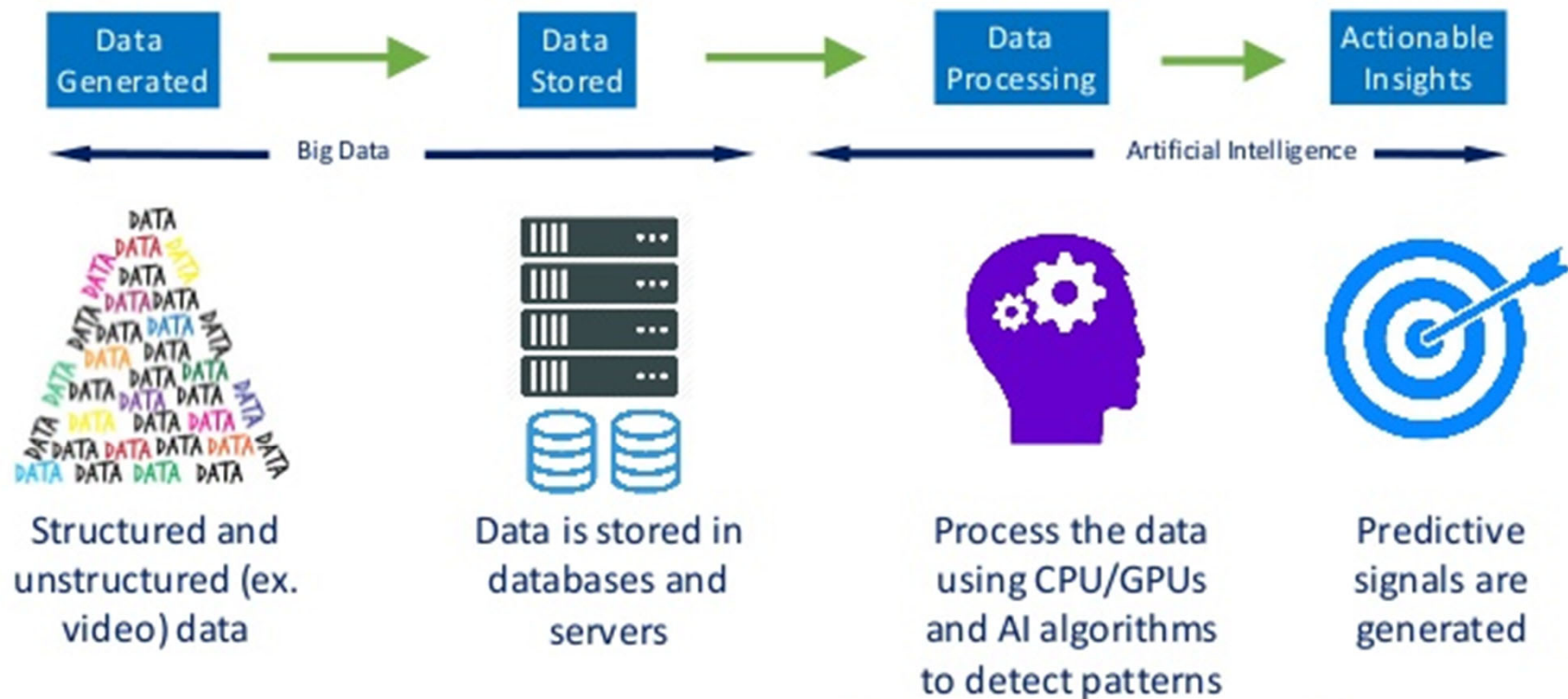


	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	<ul style="list-style-type: none">• Descriptive analytics• Diagnosis analytics		<ul style="list-style-type: none">• Predictive analytics• Data Science
Data Sets	<ul style="list-style-type: none">• Limited data sets• Cleansed data• Simple models		<ul style="list-style-type: none">• 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

Intégration des approches




The Process



Central Processing Unit (CPU)/ Graphics Processing Unit (GPU)

Domaines d'application



 **Banking**


- Optimizing Offers and Cross-sell
- Customer Service and Call Center Efficiency

 **Insurance**

- 360° View of Domain or Subject
- Catastrophe Modeling
- Fraud & Abuse

 **Telco**

- Pro-active Call Center
- Network Analytics
- Location Based Services

 **Energy & Utilities**

- Smart Meter Analytics
- Distribution Load Forecasting/Scheduling
- Condition Based Maintenance

 **Media & Entertainment**

- Business process transformation
- Audience & Marketing Optimization

 **Retail**

- Actionable Customer Insight
- Merchandise Optimization
- Dynamic Pricing

 **Travel & Transport**

- Customer Analytics & Loyalty Marketing
- Predictive Maintenance Analytics

 **Consumer Products**

- Shelf Availability
- Promotional Spend Optimization
- Merchandising Compliance

 **Government**

- Civilian Services
- Defense & Intelligence
- Tax & Treasury Services

 **Healthcare**

- Measure & Act on Population Health Outcomes
- Engage Consumers in their Healthcare

 **Automotive**

- Advanced Condition Monitoring
- Data Warehouse Optimization

 **Chemical & Petroleum**

- Operational Surveillance, Analysis & Optimization
- Data Warehouse Consolidation, Integration & Augmentation

 **Aerospace & Defense**

- Uniform Information Access Platform
- Data Warehouse Optimization

 **Electronics**

- Customer/ Channel Analytics
- Advanced Condition Monitoring

 **Life Sciences**

- Increase visibility into drug safety and effectiveness

Applications



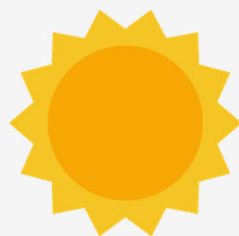
- Calculer le temps d'arrivée
- Utilise les conditions du trafic pendant une fenêtre de temps



- 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

Applications (2)



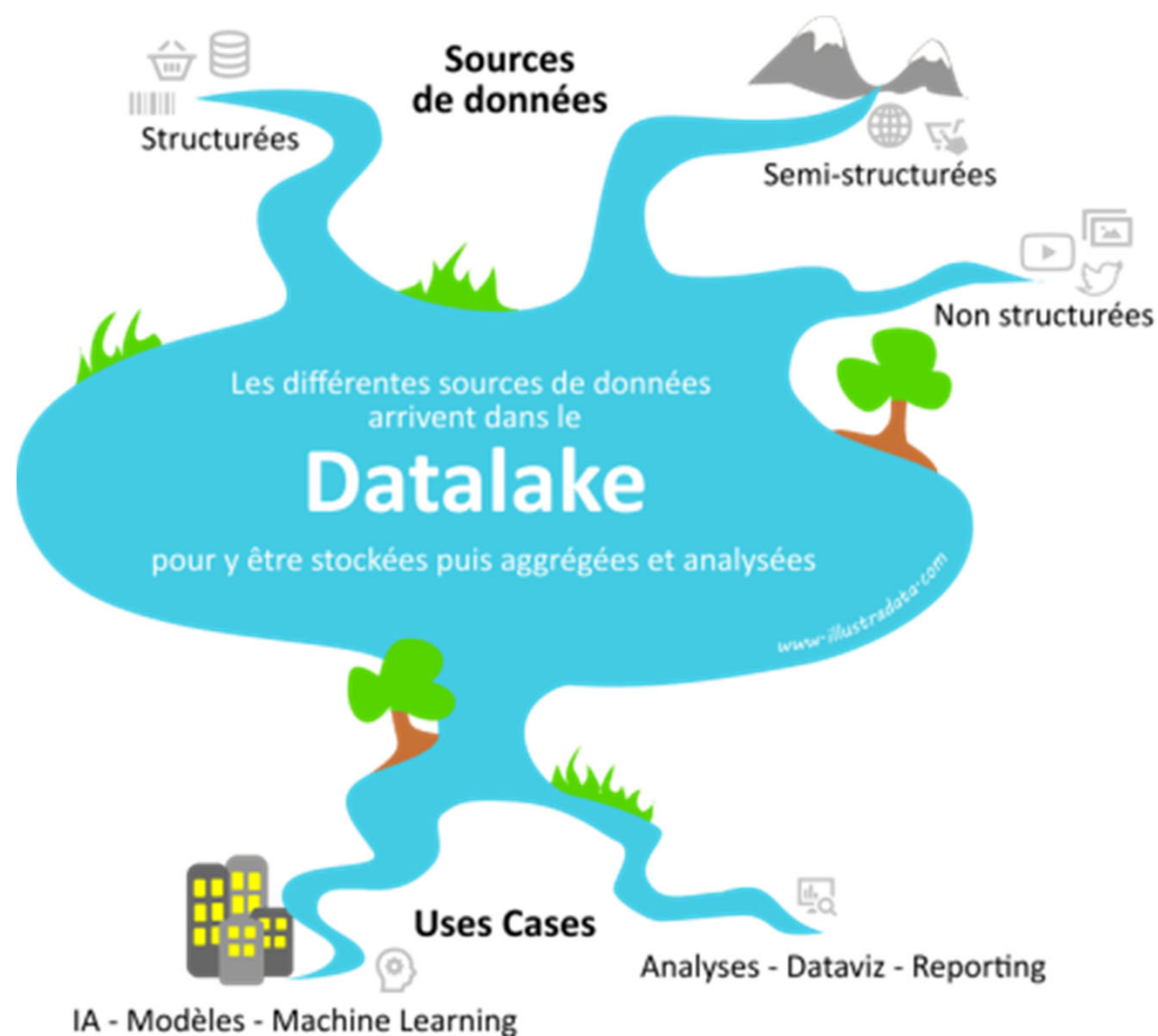
	<ul style="list-style-type: none">- Analyse des conversations sur les réseaux sociaux au moment du lancement du iPhone 5 en 2012- Lancement du Galaxy S3 en 2013 (Best Smartphone award)
	<ul style="list-style-type: none">- Coupe du monde 2014- Capteurs dans le protège-tibia- Analyse du comportement des joueurs- Simulation de situations de jeu
	<ul style="list-style-type: none">- 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% !

> Science de données

- ✓ Englobe les activités, outils et méthodes qui permettent d'exploiter les données dans tous les domaines (science, médecine, marketing ...)

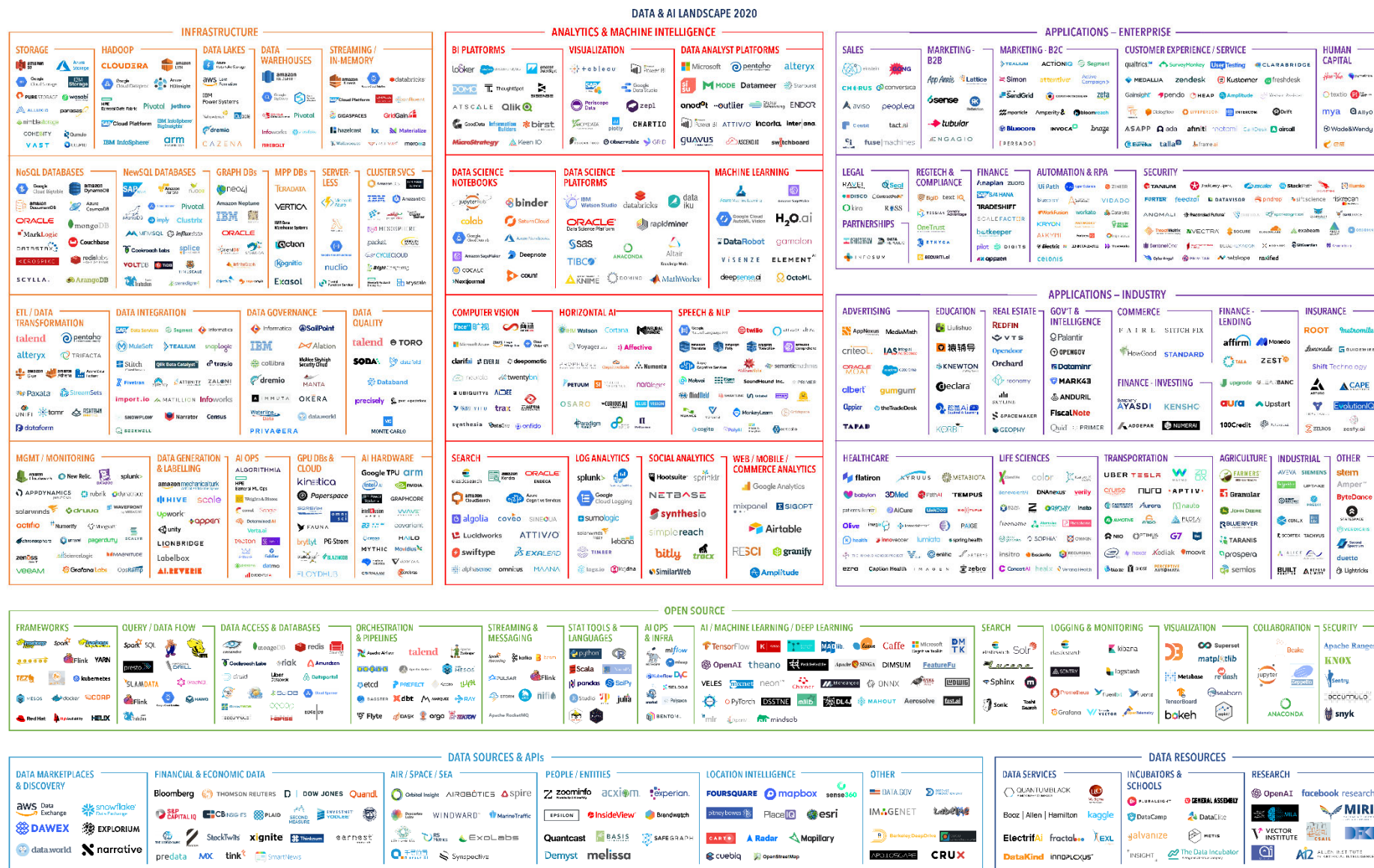
> Data Scientist :

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