PLAN

Chapitre I : Les notions de base de Big Data

- Faits et Intérêts
- Sources
- Challenges
- Définition du concept de Big data
- Caractéristiques (5V)
- Généralités

Faits et Intérêts

90%

of the world's data was created in the last two years

80%

of the world's data today is unstructured

20%

of available data can be processed by traditional systems

1 in 2

business leaders don't have access to data they need

83%

of CIO's cited BI and analytics as part of their visionary plan

5.4X

more likely that top performers use business analytics

Les données ont une propriété intrinsèque → elles grandissent

Faits et Intérêts

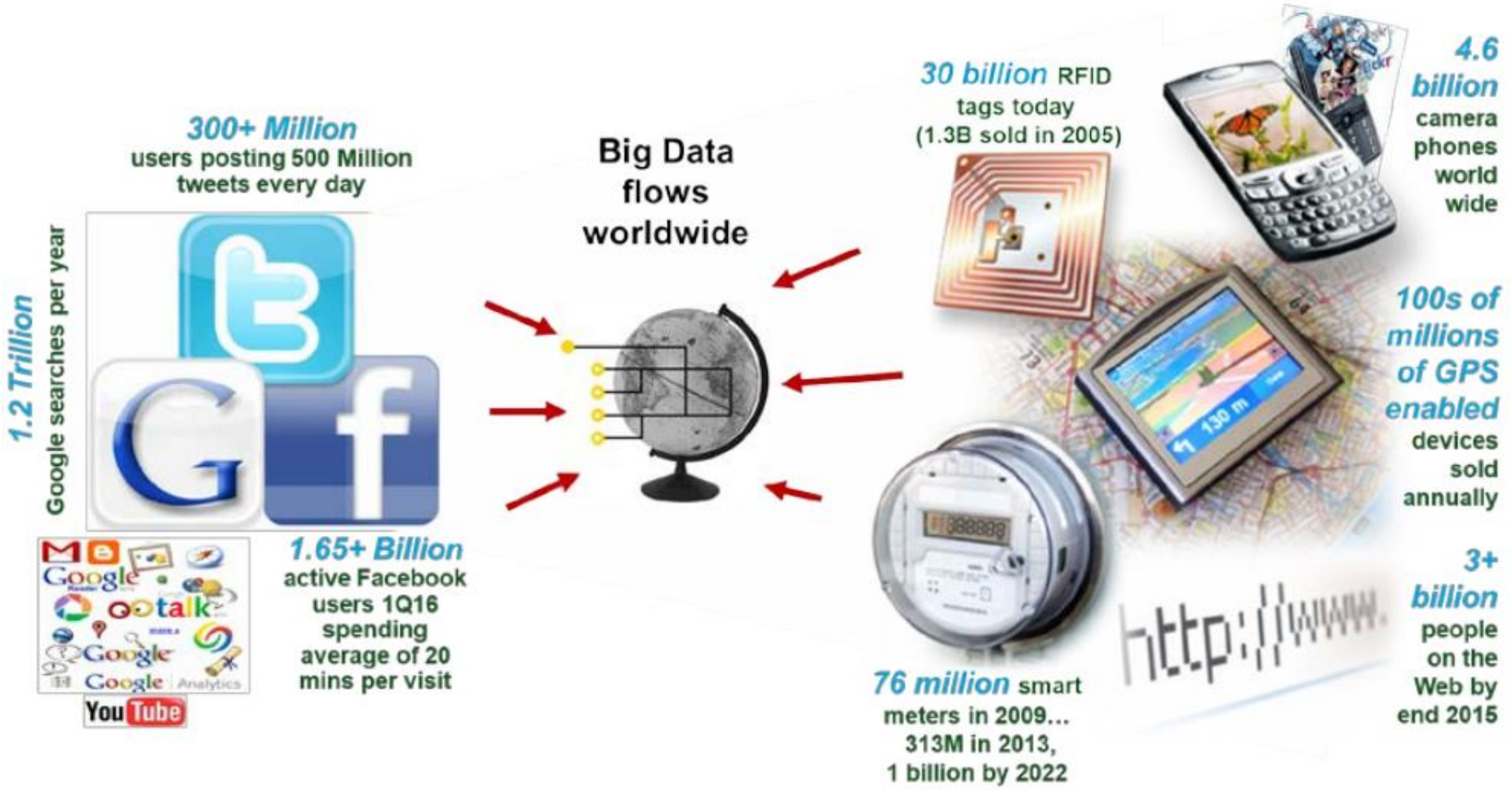
Chaque jour, nous générons 2,5 trillions d'octets de données.

- 90% des données au monde ont été créées au cours de la dernière décennie.
- 80% des données générées sont non structurées
- 20% uniquement des données disponibles peuvent être traitées par les systèmes traditionnels.
- Les dirigeants d'entreprises ne peuvent pas avoir accès aux données dont ils ont besoin.
- 83% des décideurs citent le BI et l'analyse de données dans leur business plan.

Sources

- Sources des données:
  - Capteurs utilisés pour collecter les informations climatiques
  - Messages sur les médias sociaux
  - Images numériques et vidéos publiés en ligne
  - Enregistrements transactionnels d'achat en ligne
  - Signaux GPS de téléphones mobiles
  - …
- Données appelées Big Data ou Données Massives

Sources


1.2 Trillion
Google searches per year
300+ Million
users posting 500 Million
tweets every day
1.65+ Billion
active Facebook
users 1Q16
spending
average of 20
mins per visit
Big Data
flows
worldwide
30 billion RFID
tags today
(1.3B sold in 2005)
4.6
billion
camera
phones
world
wide
100s of
millions
of GPS
enabled
devices
sold
annually
3+
billion
people
on the
Web by
end 2015
76 million smart
meters in 2009...
313M in 2013,
1 billion by 2022

Challenges

- Réunir un grand volume de données variées pour trouver de nouvelles idées
- Capturer des données créées rapidement
- Sauvegarder toutes ces données
- Traiter ces données et les utiliser

Définition du concept de Big data

- La notion de Big Data est un concept s'étant popularisé en 2012 pour traduire le fait que les entreprises sont confrontées à des volumes de données à traiter de plus en plus considérables et présentant un fort enjeux commercial et marketing.
- Ces Grosses Données en deviennent difficiles à travailler avec des outils classiques de gestion de base de données.
- Il s'agit d'un ensemble de technologies, d'architecture, 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.

Caractéristiques (5V)

- Extraction d'informations et décisions à partir des données, caractérisées par les 5 V:

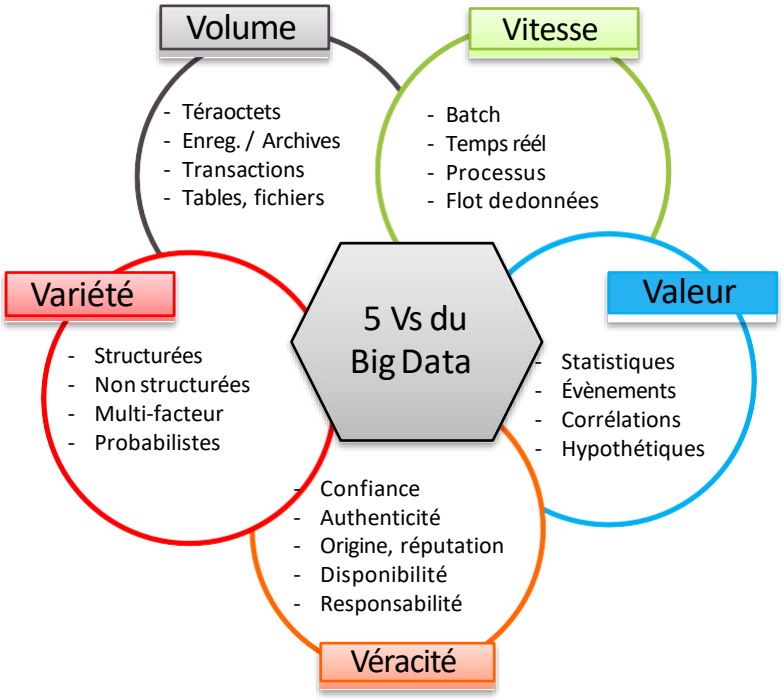
§ Volume (*Volume*)

§ Variété (*Variety*)

§ Vitesse (*Velocity*)

§ Véracité (*Veracity)*

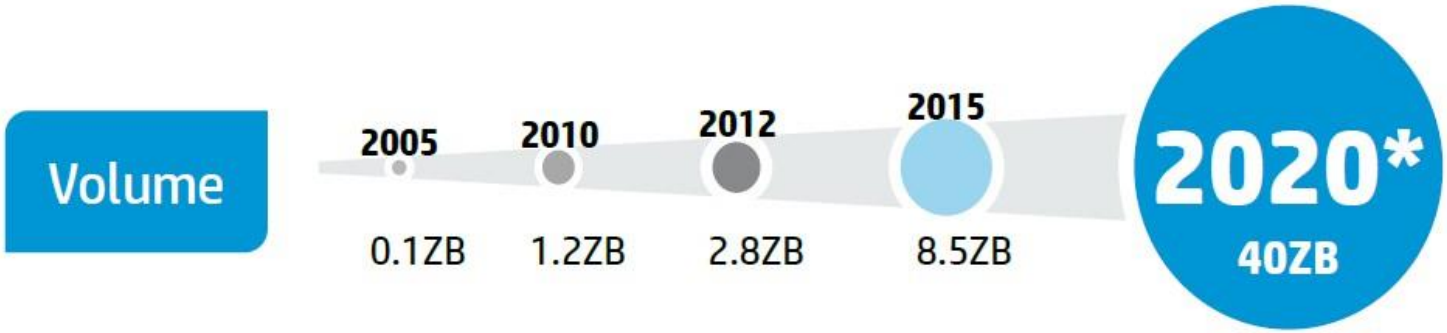
§ Valeur *(Value)*



Caractéristiques (5V)

Volume

- Croissance sans cesse des données à gérer de tout type, souvent en teraoctets voir en petaoctets.
- Chaque jour, 2.5 trillions d'octets de données sont générées.
- 90% des données créées dans le monde l'ont été au cours des 2 dernières années,
- Prévision d'une croissance de 800% des quantités de données à traiter d'ici à 5 ans.



Caractéristiques (5V)

Variété (Variety)

- Traitement des données sous forme structurée (bases de données structurée, feuilles de calcul, …) et non structurée (textes, sons, images, vidéos, données de capteurs, fichiers journaux, médias sociaux, signaux,…) qui doivent faire l'objet d'une analyse collective.

Caractéristiques (5V)

Vitesse (Velocity)

- Rapidité d'arrivée des données
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des teraoctets par jour.

  Sinon, risque de perte d'informations.
- Exemple

  Il ne suffit pas de savoir quel article a été acheté ou réservé par un client.

  Si on sait qu'un client a passé plus de 5mn à consulter un article dans une boutique d'achat en ligne, il est utile de lui envoyer un email dès que cet article est soldé.

Caractéristiques (5V)

Véracité (Veracity)

- Fait référence à la qualité de la fiabilité et la confiance des données.
- Données bruités, imprécises, prédictives, …
- Avec l'augmentation de la quantité, la qualité et la précision des données diminuent.

  Comment se trouver dans un déluge de hashtags ?

  Comment gérer les données partielles ou incomplètes ?

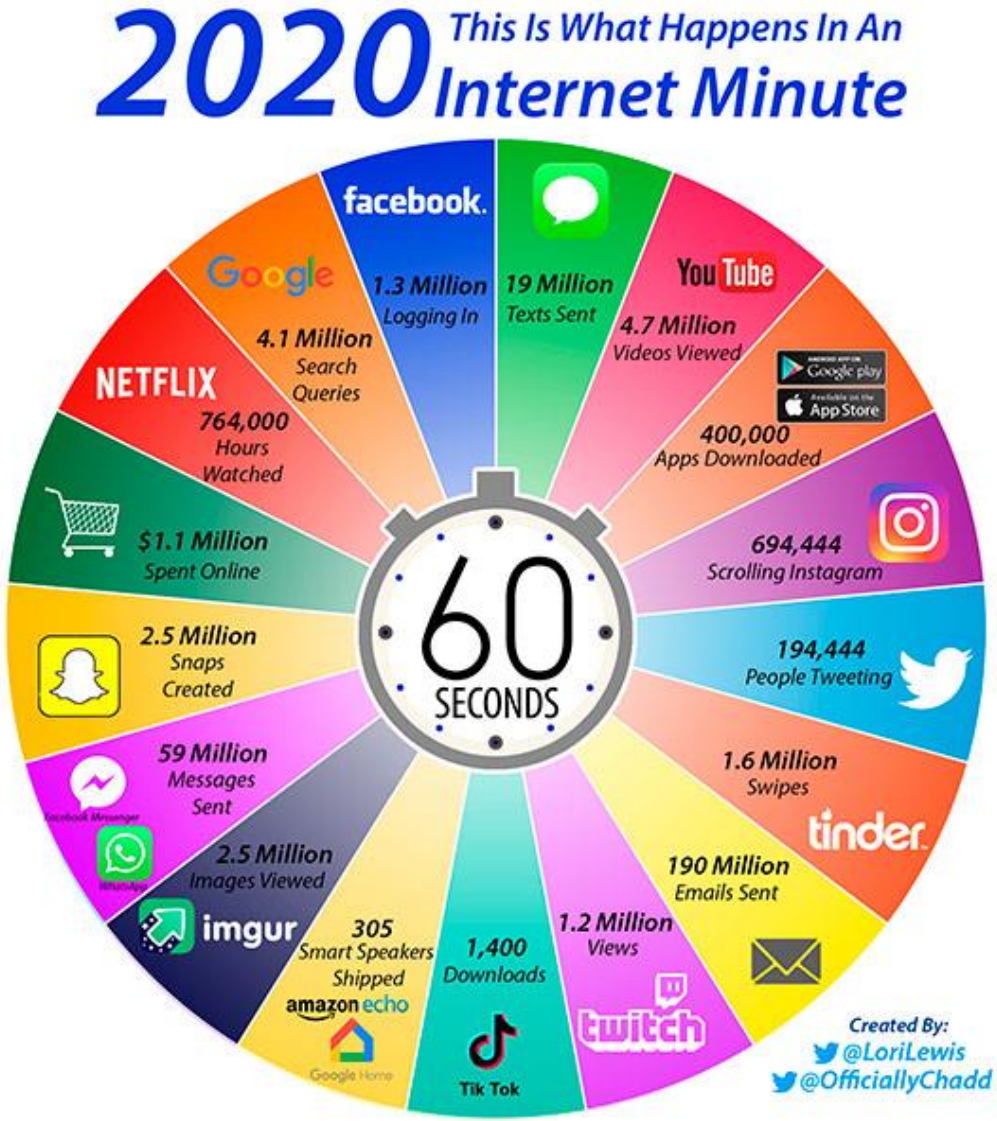
- Les solutions big data doivent remédier à cela
- Besoin d'une grande rigueur dans la collecte, l'enrichissement et le croisement des données

Caractéristiques (5V)

Valeur (Value)

- Le succès d’un projet Big Data n'a d'intérêt aux utilisateurs que s'il apporte de la valeur ajoutée et de nouvelles connaissances.
- Il faut transformer les données en valeurs exploitables
- Sans une réelle valeur, ce n’est qu’un gaspillage de ressources

Généralités



Généralités

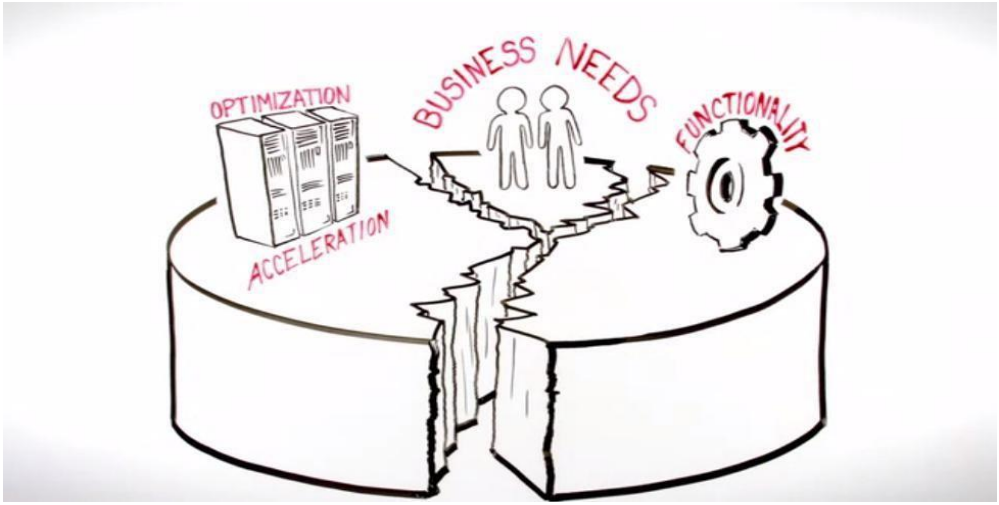
Quel est le problème posé par ces énormes quantités de données?

- Au paravent, quand les systèmes d'application de gestion de base de données ont été réalisés, ils ont été construits avec une échelle à l'esprit (limité). Même les organisations n'ont pas été préparées à l'échelle que nous produisons aujourd'hui.
- Comme les exigences de ces organisations ont augmenté au fil du temps, ils doivent repenser et réinvestir dans l'infrastructure. Actuellement, le coût des ressources impliquées dans l'extension de l'infrastructure, s'augmente avec un facteur exponentiel.
- De plus, il y aurait une limitation sur les différents facteurs tels que la taille de la machine, CPU, RAM, etc.
- Ces systèmes traditionnels ne seraient pas en mesure de soutenir l'échelle requise par la plupart des entreprises.

Généralités

ADAPTABILITE

- Dans ce nouveau contexte, les méthodes de traitement de ces données (capture, stockage, recherche, partage, analyse, visualisation) doivent être redéfinies car l'ensemble de ces données deviennent difficilement manipulables par les outils classiques.



Généralités

Comment le Big Data gère ces situations complexes?

- La distribution des données

  Système de Fichiers Distribués - DFS (Distributed File System)

- Le traitement en parallèle

  MapReduce de Google

- La tolérance aux pannes
- La réplication des données
- L'utilisation de matériel standard
- Flexibilité, évolutivité et scalabilité

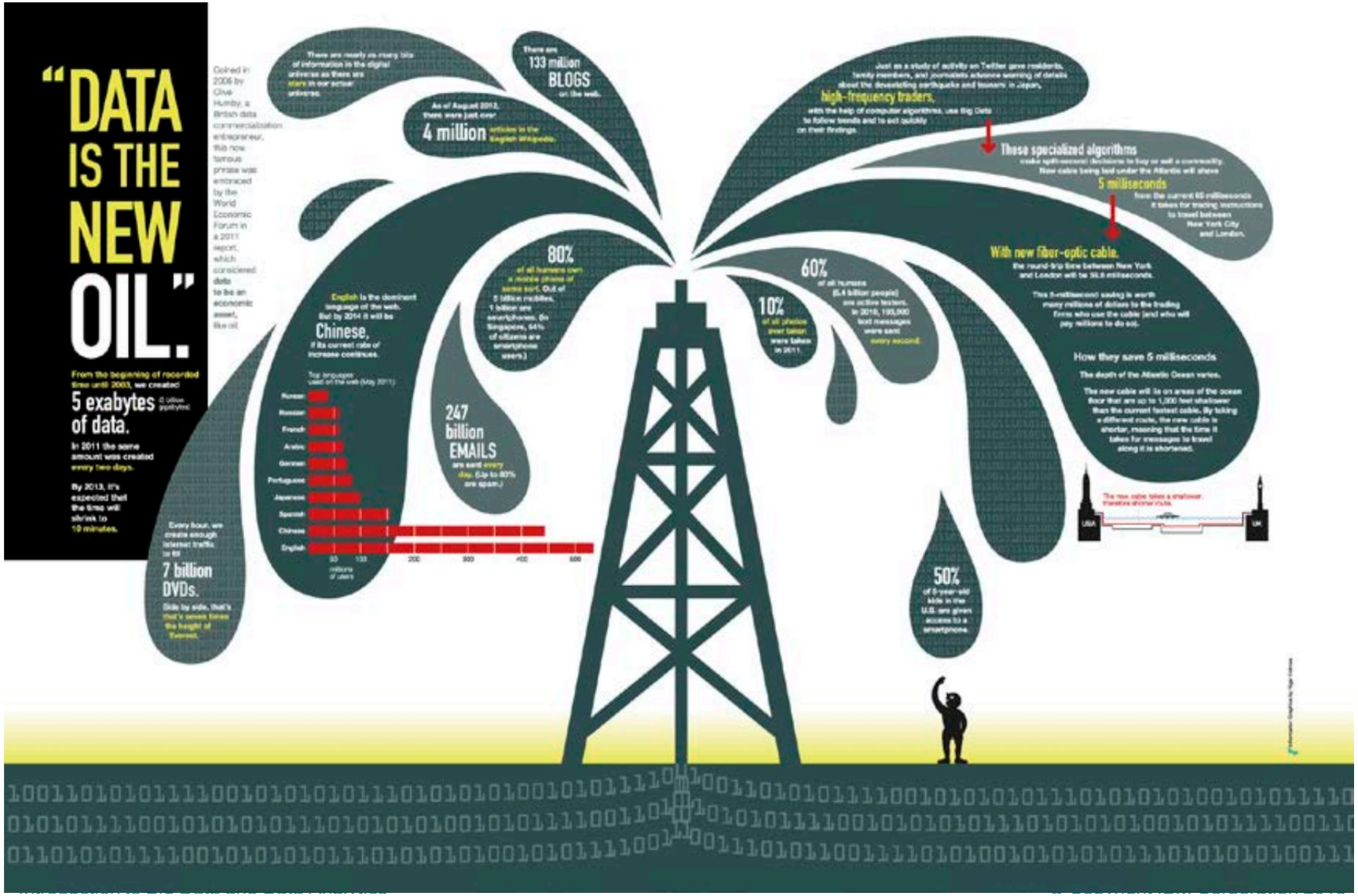
Généralités

Cas d'utilisation : Politique

- L'analyse de Big Data a joué un rôle important dans la campagne de réélection de Barack Obama, notamment pour analyser les opinions politiques de la population.
- Depuis l'année 2012, le Département de la défense américain investit annuellement sur les projets de Big Data plus de 250 millions de dollars.
- Le gouvernement américain possède six des dix plus puissants supercalculateurs de la planète.
- Les Etats-Unis d'Amérique possèdent le plus grand nombre des data centers au monde.
- En 2014, SIGMA conseil a utilisé le Big Data pour donner l'estimation du résultat de vote préliminaire en Tunisie.

“Data is the New Oil”

Coined in 2006 by Clive Huby, a British data commercialization entrepreneur.



Plateforme – Technologies - Outils

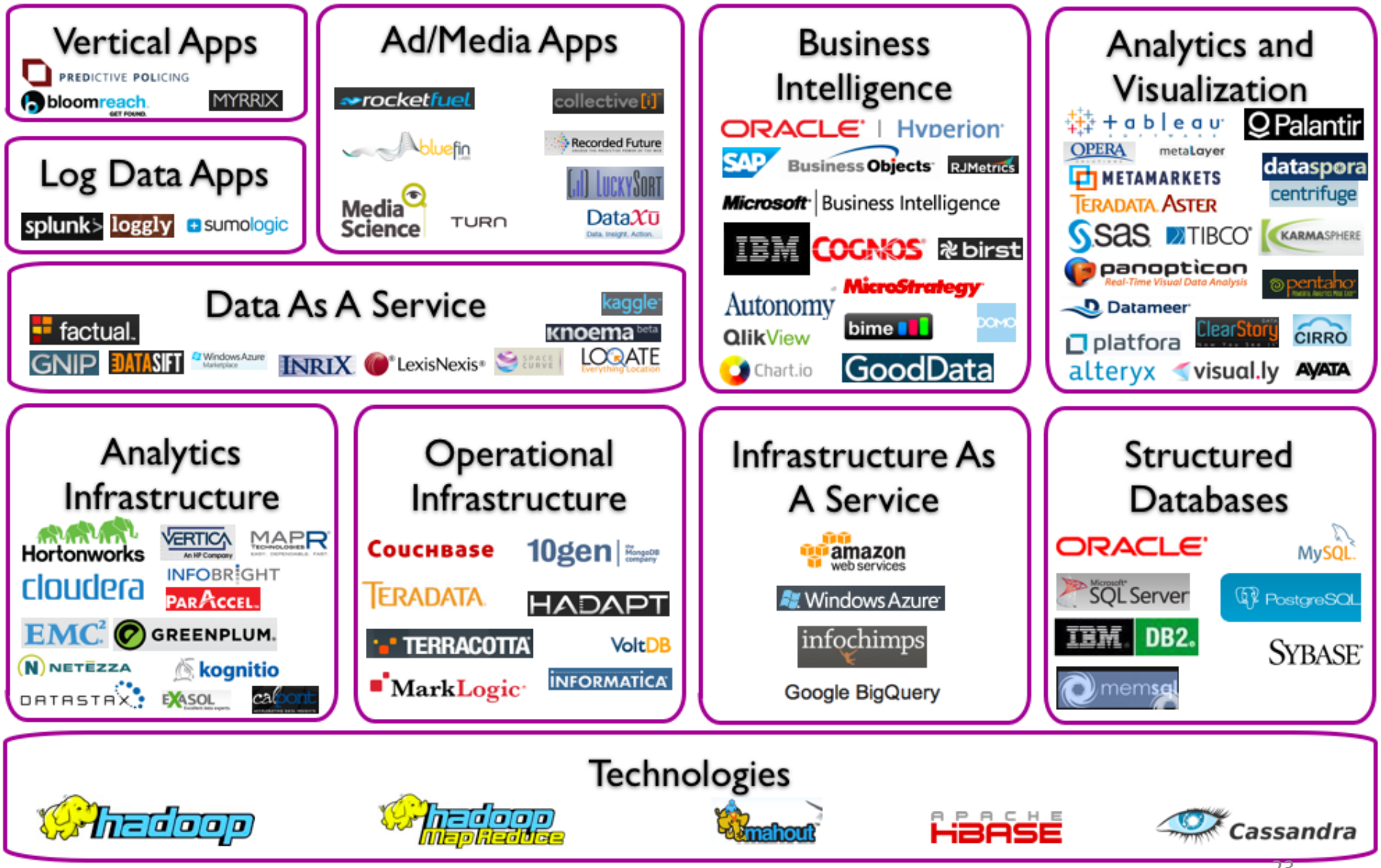
Société	Technologie développée	Type de technologie
Google	**Big Table**	Système de base de données distribuée propriétaire reposant sur GFS (Google File System). Technologie non Open Source, mais qui a inspiré Hbase qui est Open Source.
	MapReduce	Plate-forme de développement pour traitements distribués
Yahoo	**Hadoop**	Plateforme Java destinée aux applications distribuées et à la gestion intensive des données. Issue à l'origine de GFS et MapReduce.
	S4	Plateforme de développement dédiée aux applications de traitement continu de flux de données.
Facebook	**Cassandra**	Base de données de type NoSQL et distribuée.
	Hive	Logiciel d'analyse de données utilisant Hadoop.
Twitter	**Storm**	Plateforme de traitement de données massives.
	FockDB	Base de données distribuée de type graphe.
LinkedIn	**SenseiDB**	Base de données temps réel distribuée et semi-structurée.
	Voldemort	Base de données distribuée destinée aux très grosses volumétries.

Tableau : Quelques technologies Open Source du Big Data

Plateforme – Technologies - Outils

- Processing
  - Hadoop, Hive, Pig, mrjob, Caffeine
- NoSQL Databases
  - Hbase, MongoDB, Vertica, Cassandra, Neo4j, etc.
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Analytics
  - R, SAS, Python scikit-learn, Spark MLLib, Apache Mahout
- Search
  - Solr/Lucene, ElasticSearch

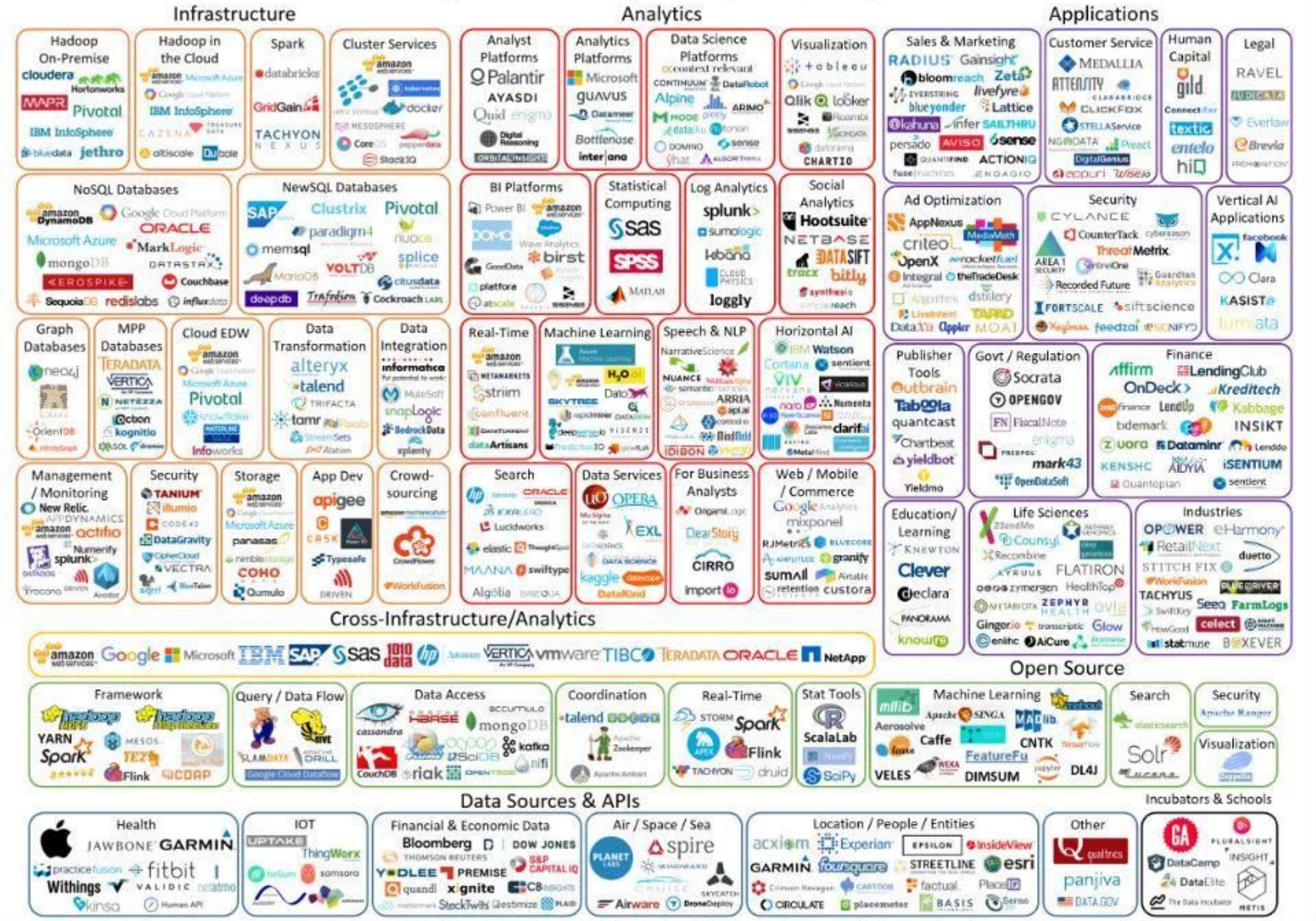
Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

Big Data Landscape 2020