

LEBIGDATA

BLENDING



Les principes du Big Data

Les origines du Big Data

Les origines du Big Data



Les origines du Big Data



1ère explosion des données

1448 poursuit ses recherches et, deux ans plus tard, persuade le riche banquier **Johann Fust** de l'aider à financer son projet.

1454 naissance de l'imprimerie à caractères mobiles

1468 : le 3 février, Gutenberg meurt, et lègue son invention à l'humanité.

1964 : L'IBM 360

Le PDP-8 de DEC (Digital Equipment Corporation)

1965 : Le premier système expert (Dendral) par Edward Feigenbaum



l'ordinateur IBM 360

La démocratisation de l'informatique dans les entreprises les administrations et les milieux scientifiques permet de traiter et de stocker de plus en plus de données dans de multiples domaines

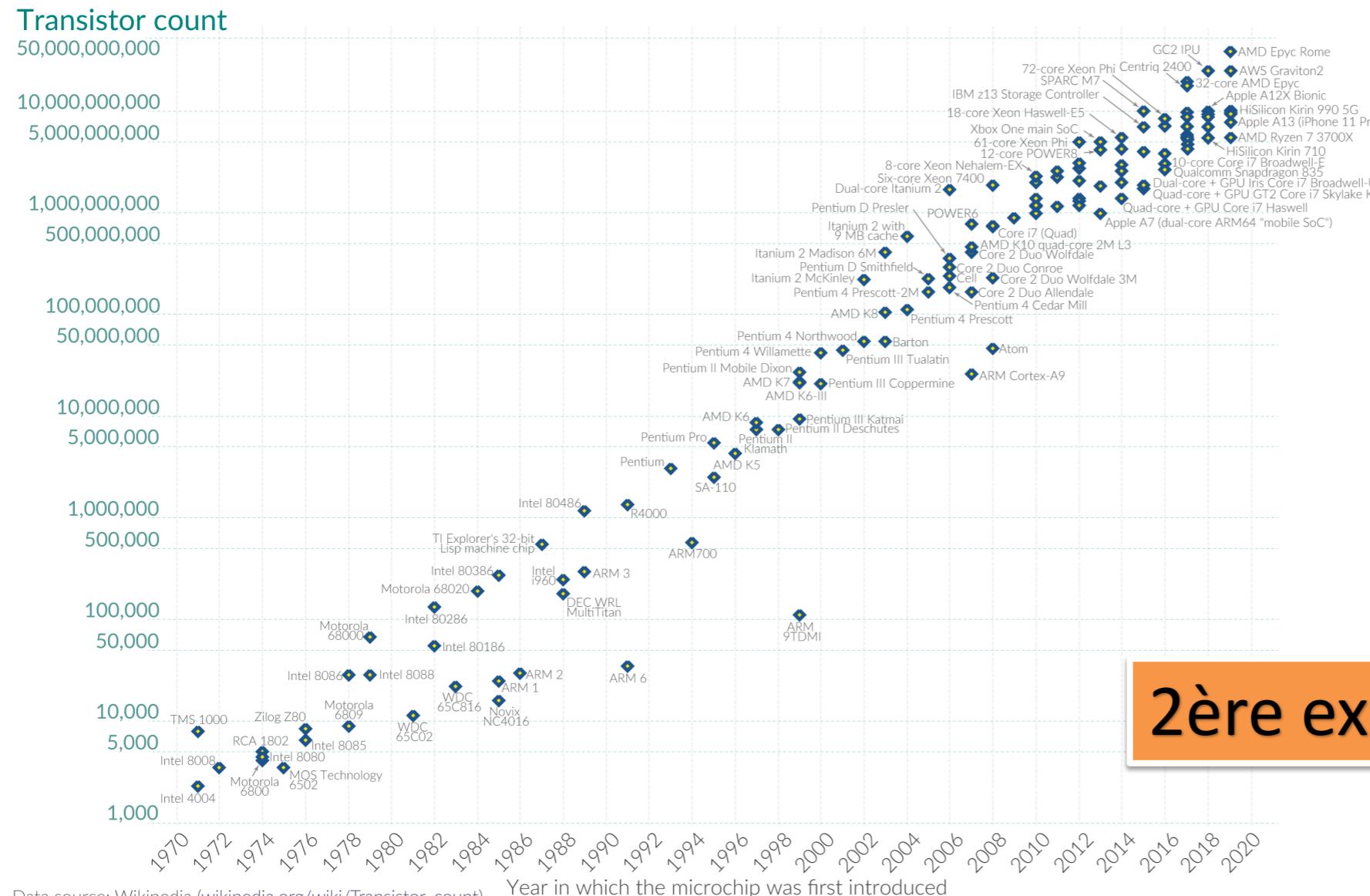
Dendral est système expert ou programme interactif créé en 1965, par les informaticiens Edward Feigenbaum, Bruce Buchanan, le médecin Joshua Lederberg et le chimiste Carl Djerassi

Les origines du Big Data

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World
in Data



La loi de Moore fixe un cycle de dix-huit mois pour les doublements de nombre de transistors, rendant les ordinateurs rapidement obsolètes

2ère explosion des données

Data source: Wikipedia ([wikipedia.org/wiki/Transistor](https://en.wikipedia.org/w/index.php?title=Transistor&oldid=10000000))

OurWorldinData.org – Research and data to make progress against the world's largest problems

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser

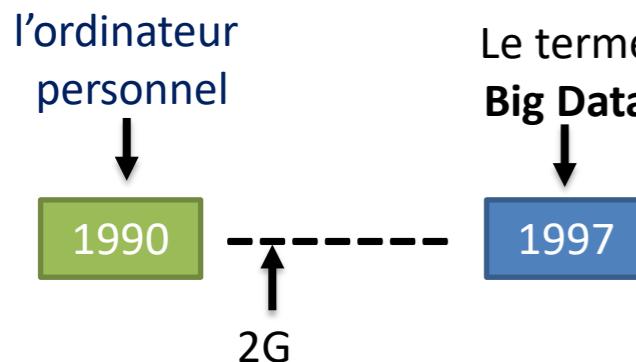
Les origines du Big Data

En juillet 1992, France Télécom lance le premier réseau de téléphonie mobile 2G à la norme GSM : itineris.

Cette décennie a aussi été marquée bien sûr par l'ouverture de l'Internet au commerce, fin 1992, puis par l'expansion de la Toile. La convergence de l'informatique, de l'Internet, et des télécommunications a été décrite par une nouvelle expression les « **technologies de l'information et de la communication** »



Le terme « **Big Data** » est apparu pour la première fois en 1997 dans un article publié par deux **chercheurs de la Nasa** : **Michael Cox, et David Ellsworth**. A cette époque, les deux chercheurs mettent déjà en lumière l'augmentation du volume de données produit par la population et la difficulté des systèmes à pouvoir traiter ce volume grandissant. Les spécialistes annoncent aujourd'hui une augmentation de 4 300 % de la génération annuelle des données d'ici à 2020. (Source : CSC)



Managing Big Data for Scientific Visualization

Michael Cox[†]
MRJ/NASA Ames Research Center
Microcomputer Research Lab, Intel Corporation

David Ellsworth[‡]
MRJ/NASA Ames Research Center

1-May-97

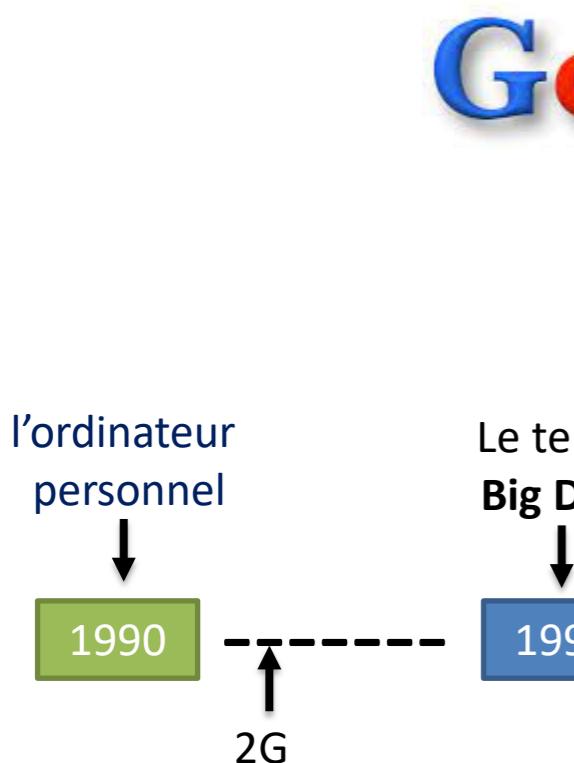
Introduction

Many areas of endeavor have problems with big data. Some classical business applications have faced big data for some time (e.g. airline reservation systems), and newer business applications to exploit big data are under construction (e.g. data warehouses, federations of databases). While engineering and scientific visualization have also faced the problem for some time, solutions are less well developed, and common techniques are less well understood. In this section we offer some structure to understand what has been done to manage big data for engineering and scientific visualization, and to understand and go forward in areas that may prove fruitful. With this structure as backdrop, we discuss the work that has been done in management of big data, as well as our own work on demand-paged segments for fluid flow visualization.

Our primary goal is to enable the scientist or engineer to extract information from his or her data. Many authors begin with the assumption that interactivity is the most important goal of visualization systems (cf. [1]). While we agree that interactivity is important, it is not always possible on big data sets. We encounter practicing scientists and engineers whose single most important goal is to understand their data; they are willing to live with off-line algorithms that give them information that interactive visualization

4 septembre 1998, Menlo Park, Californie, États-Unis

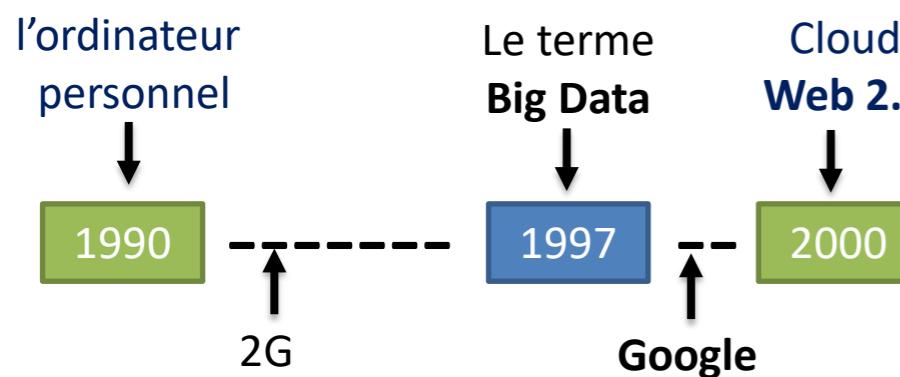
Les fondements de l'histoire de l'entreprise Google commencent par la rencontre de deux étudiants de l'université Stanford en 1995. Sergey Brin alors âgé de vingt-trois ans et Larry Page de vingt-quatre ans sont « pratiquement en désaccord sur tout »³⁸. Cela ne les empêche pourtant pas, en janvier 1996, de commencer à travailler sur un nouveau moteur de recherche



Cloud Web 2.0

Du point de vue des techniques de développement web, le terme a été également beaucoup utilisé dans la seconde moitié des années 2000 pour désigner la généralisation de l'utilisation des technologies dites AJAX qui permettent de modifier l'apparence d'une page web en fonction des instructions données par le serveur sans avoir à la recharger, ce qui donne à un site web des possibilités comme interagir avec l'utilisateur lors du remplissage d'un formulaire.

3ère explosion des données



AJAX Contact Form

Name:

Sujet: Choisissez

Email:

Message:

Regarding:

Message de votre commande:

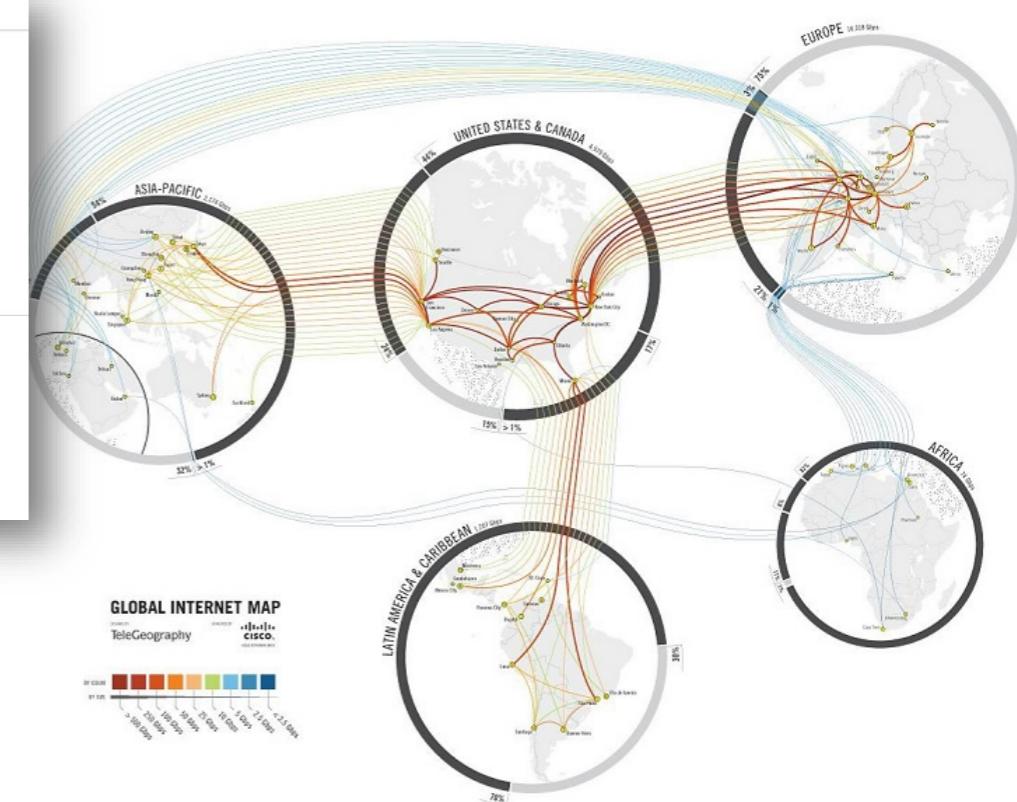
Annexe: Choisissez un fichier Aucun fichier choisi

Envoyer

ERREUR pour le propriétaire du site : Clé de site non valide

reCAPTCHA

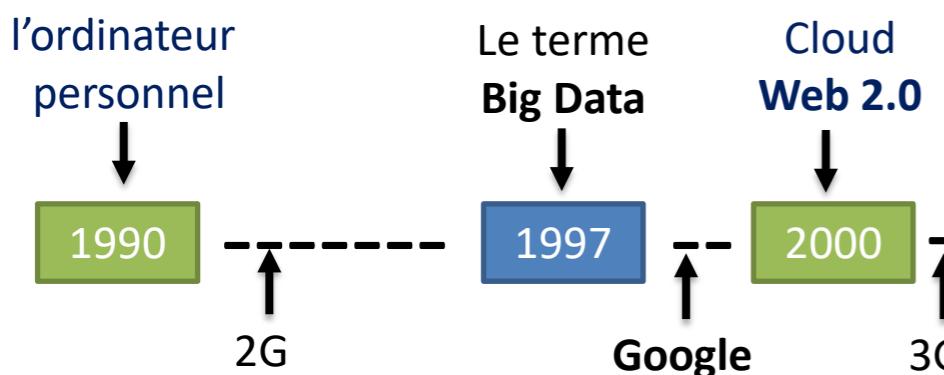
Confidentialité - Conditions



Les origines du Big Data

3G :

communications plus rapides notamment pour la voix, la télécopie, l'Internet de n'importe quel endroit et à tout moment. L'UIT IMT-2000 est la norme internationale de la 3G a ouvert la voie à de nouvelles applications et services comme par exemple le divertissement multimédia, la localisation des services, ...



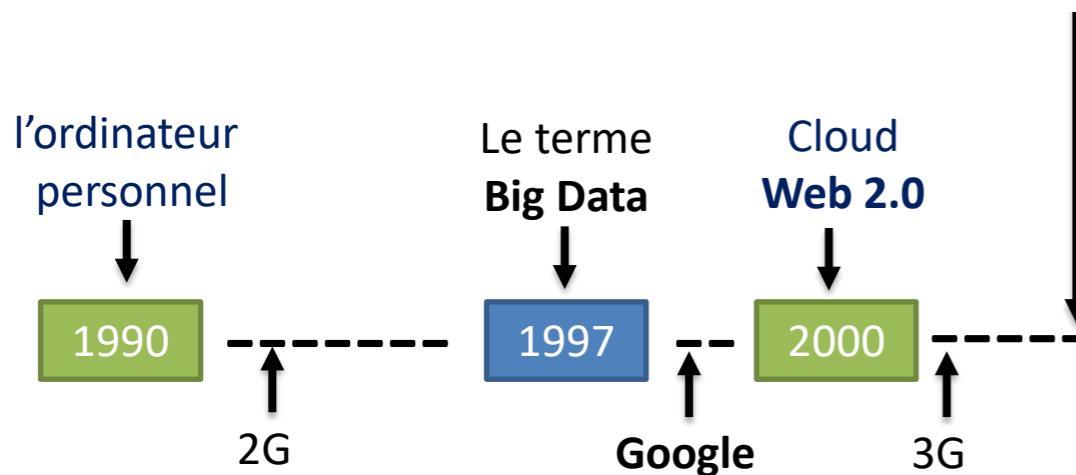
février 2004, Cambridge, Massachusetts, États-Unis



Le réseau social est né : en 24 heures, 1200 étudiants s'inscrivent sur le service. En un mois, 50% des étudiants d'Harvard deviennent membres du site.

En mars 2004, le réseau social intègre trois nouvelles universités : Stanford, Columbia et Yale. En juin, 150 000 étudiants sont inscrits au sein de 30 campus et les premiers investisseurs viennent soutenir le réseau :

facebook

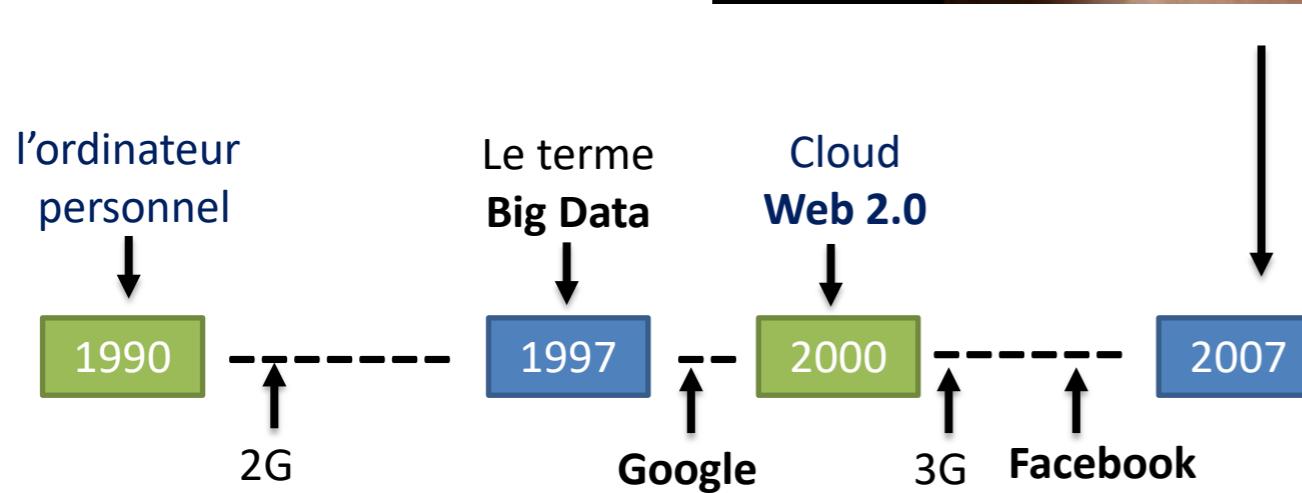


Les origines du Big Data

Janvier 2007 que Steve Jobs présente le tour premier Iphone

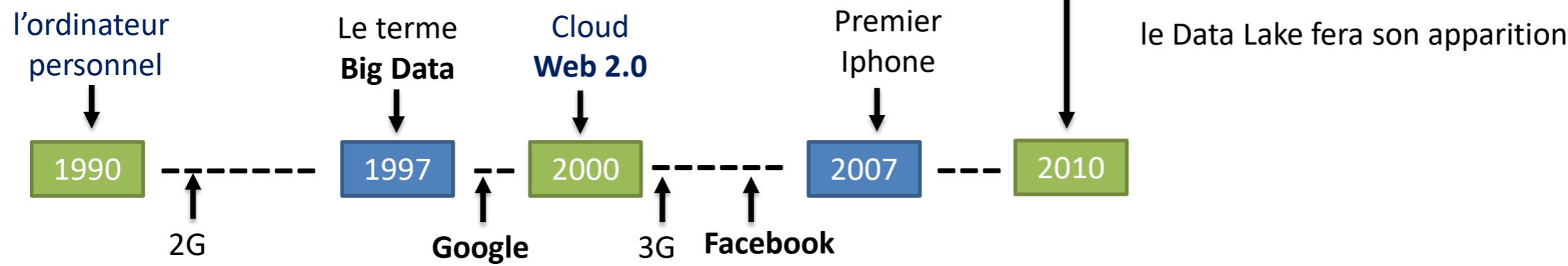


Le volume des données stockées est en pleine expansion

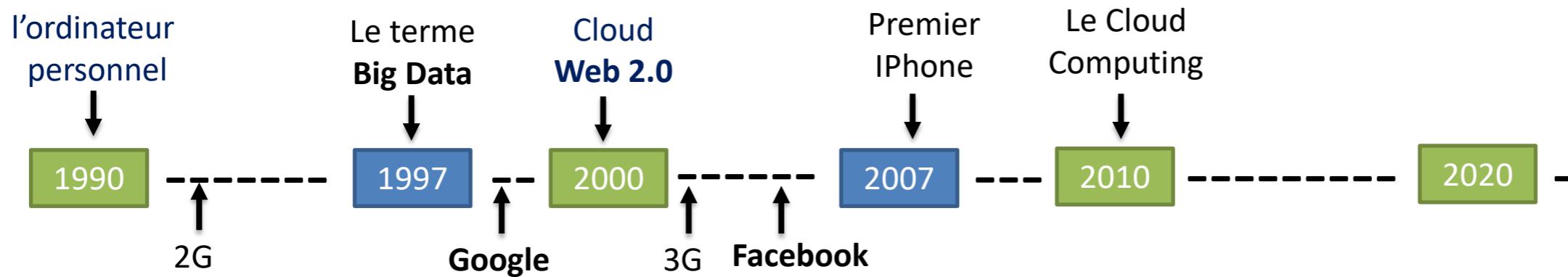


Les origines du Big Data

Le Cloud Computing (le Nuage informatique).



Les origines du Big Data



Le nombre d'utilisateurs connectés à internet depuis un smartphone dépasse celui des utilisateurs connectés via un PC

Les ventes de tablettes dépassent celles des PC

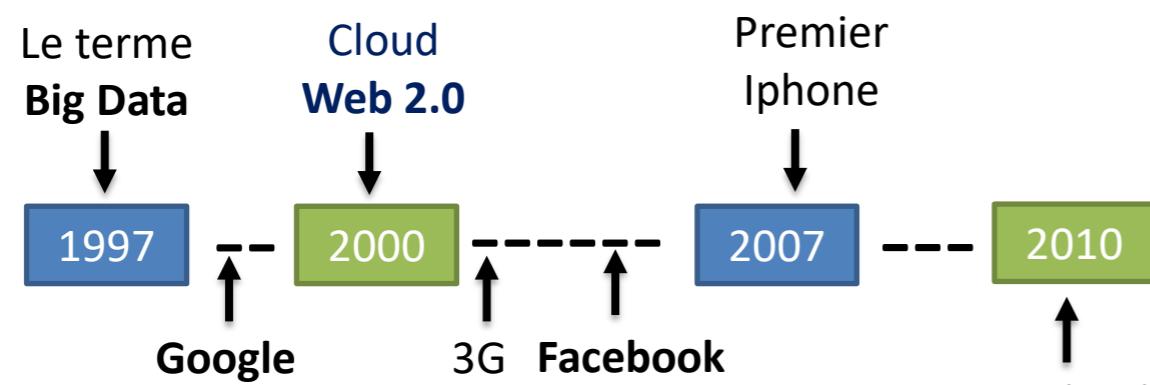
- Émergence du marché des objets connectés1
- Le nombre de sites web dans le monde dépasse le milliard
- Le temps d'utilisation des applications sur mobiles a dépassé l'utilisation d'internet sur ordinateur.

- Les recherches Google sur mobile ont dépassé celles faites à partir d'un ordinateur classique

Forte croissance de l'utilisation du (Cloud computing) a l'occasion de la crise sanitaire

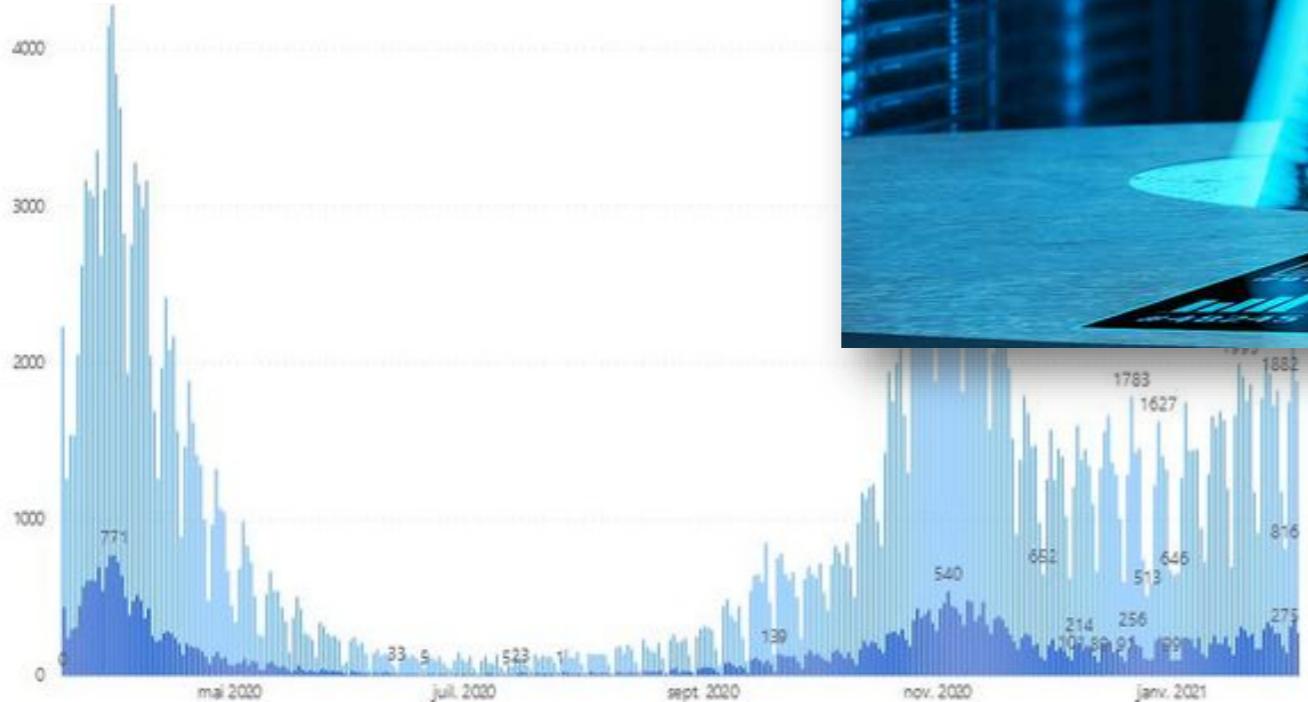
2020

- Entrée en vigueur du règlement général sur la protection des données (RGPD).
- développement de l'internet des objets : 5,5 millions de nouveaux objets se connectent au réseau chaque jour



Les origines du Big Data

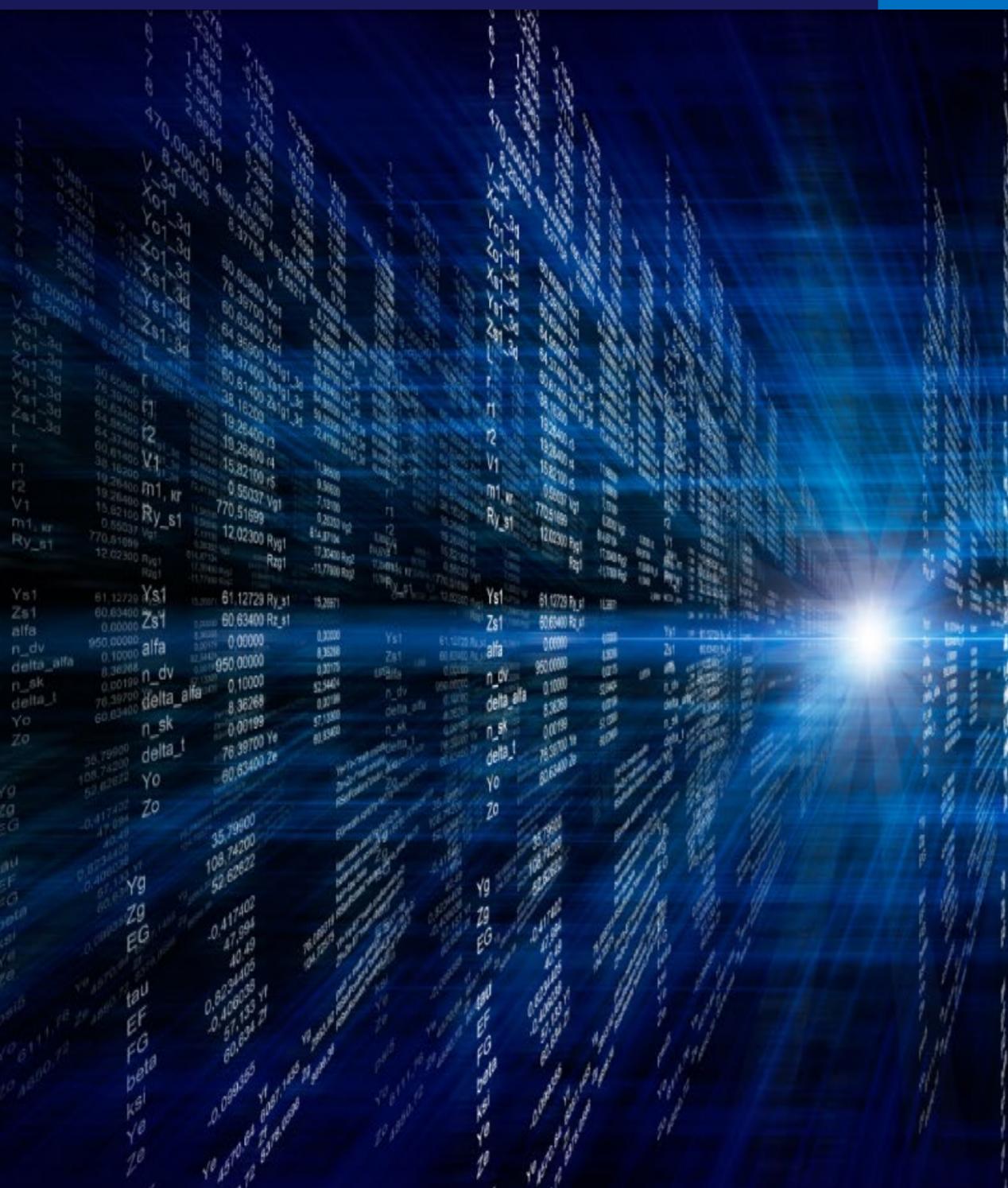
. Dans ce contexte de la pandémie



Le Big Data qu'est-ce que c'est ?

- **Définition**
- **Les 3v +1**
- **Les points faibles**
- **Les risques**
- **A qui bénéficie le big data**

Définition



Le Big Data qu'est-ce que c'est ? - Définition

Le Big Data, littéralement « **grosses données** », « **méga données** », ou encore « **données massives** », désigne des ensembles de données devenus si volumineux qu'ils arrivent à dépasser l'intuition, les capacités humaines d'analyse et remettent en cause l'utilisation des outils informatiques classiques.

Ces données peuvent être de nature:

- Personnelle
- Professionnelle
- Institutionnelle

Provenir de différentes sources d'information circulant par le biais des différents réseaux numériques :

- Texte
- Vidéo
- Audio
- base de données
- etc...)

Le Big Data qu'est-ce que c'est ? - Définition

Aujourd'hui, la puissance du Big Data est telle que le volume des données a dépassé nos capacités d'analyse traditionnelles. Produites à chaque instant n'importe où dans le monde, ces « mégadonnées » ou « données massives » prennent de plus en plus d'ampleur, de façon structurée ou non, en temps réel ou non.



Pour être qualifiées de "**données du Big Data**", ces données doivent répondre au critère **des trois V**

Volume Vélocité Variété

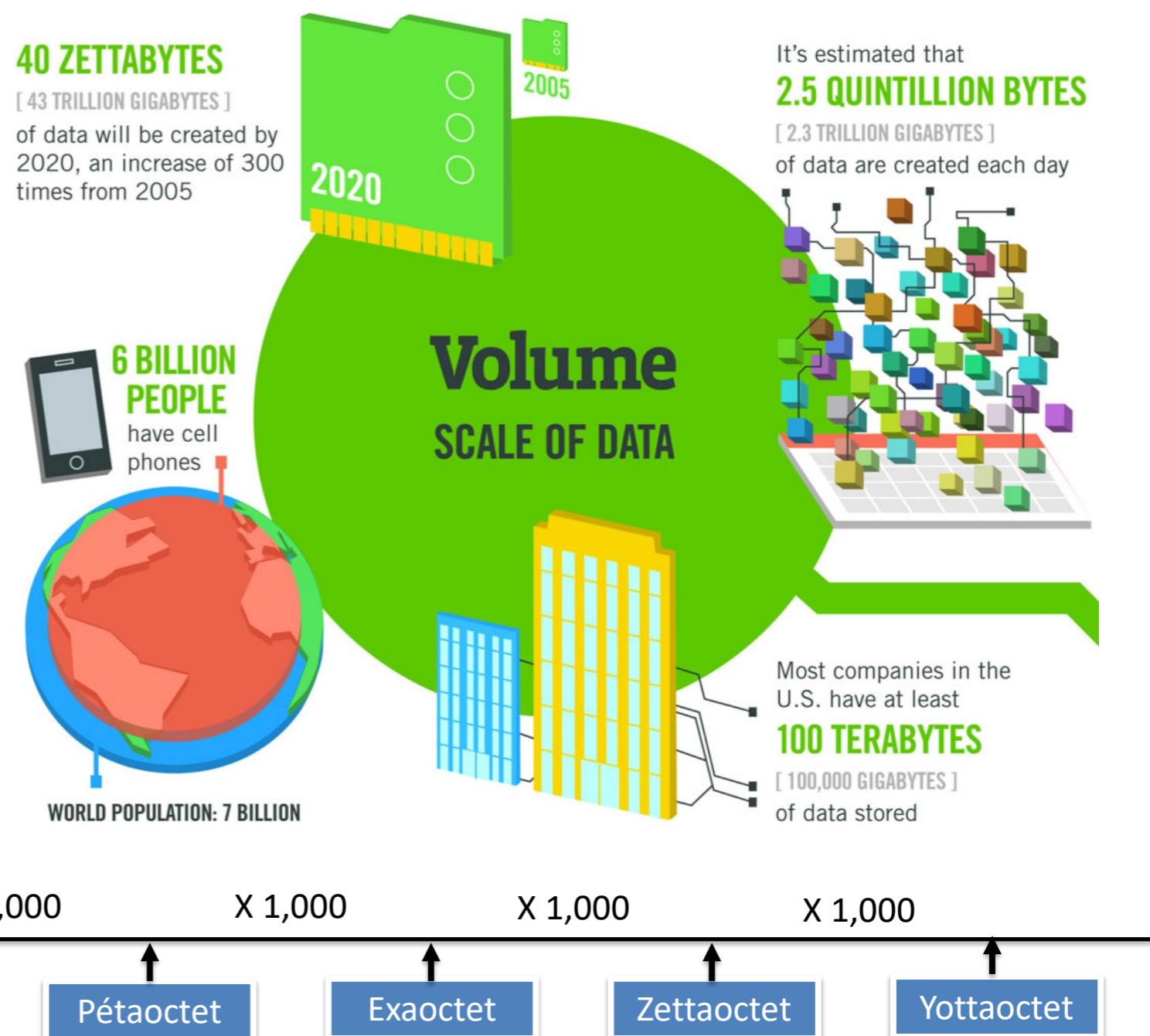
3V

Critère des trois V :

V - pour **Volume**

(plus ou moins massif)

- WEB
- Smartphone
- Capteur
- Données Publiques



Critère des trois V :

V - pour Volume

(plus ou moins massif)

V - pour Velocity (rapidité)

(de la production, de la collecte et d'analyse)

Chaque activité réalisée sur internet est désormais traquée avec précision

Être capable de traiter et analyser ce flux continu de données

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

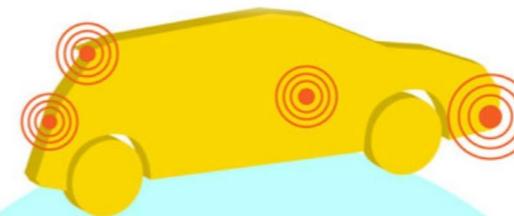
during each trading session



By 2016, it is projected there will be

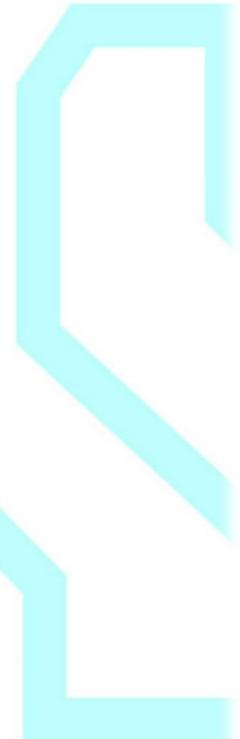
18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity
ANALYSIS OF STREAMING DATA



Critère des trois V :

V - pour Volume

(plus ou moins massif)

V - pour Velocity (rapidité)

(de la production, de la collecte et d'analyse)

V - pour Variété

(nature et niveau de structuration)

Parfaitement Structurées	Semi Structurées	Non Structurées
Tables de Base de données	<ul style="list-style-type: none"> • Fichier JSON • Fichier XML • Fichier CSV • Fichier HTML 	<ul style="list-style-type: none"> • Vidéo • Audio • Binaires • Images

As of 2011, the global size of data in healthcare was estimated to be

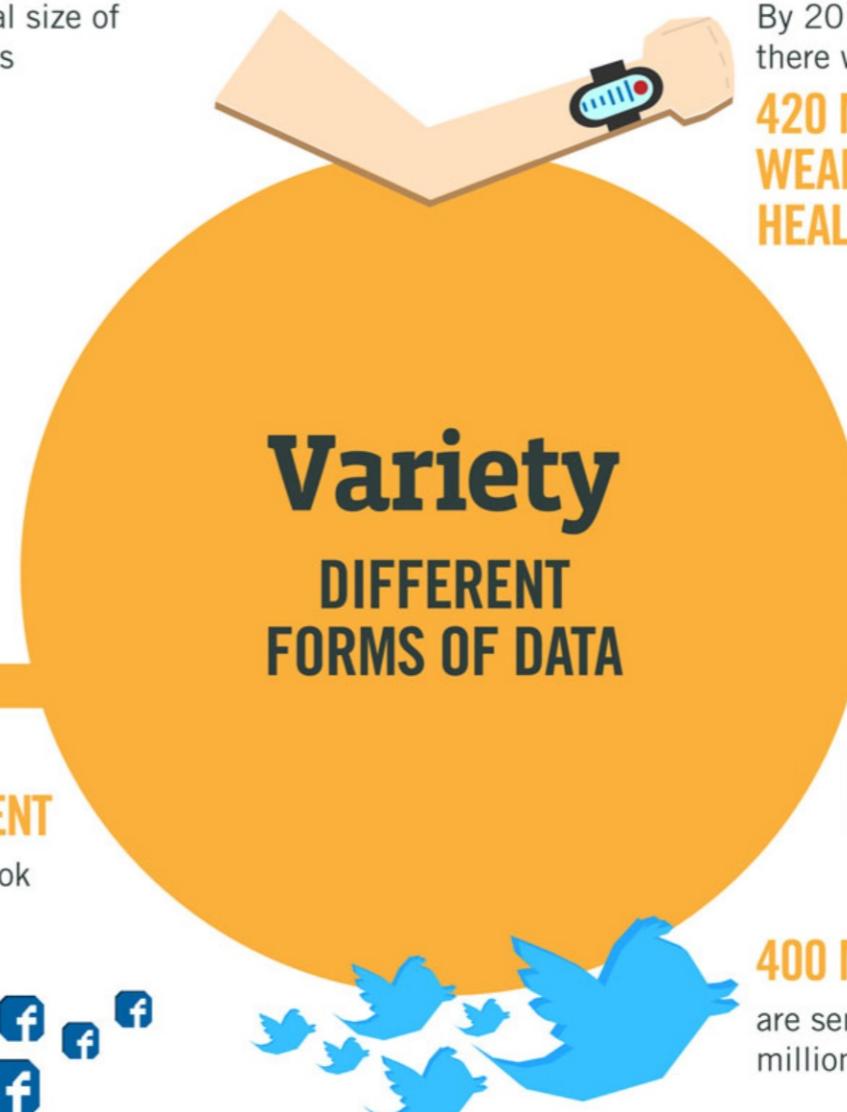
150 EXABYTES

[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT

are shared on Facebook every month



By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Critère des trois V :

V - pour **Volume**

(plus ou moins massif)

V - pour **Velocity** (rapidité)

(de la production, de la collecte et d'analyse)

V - pour **Variété**

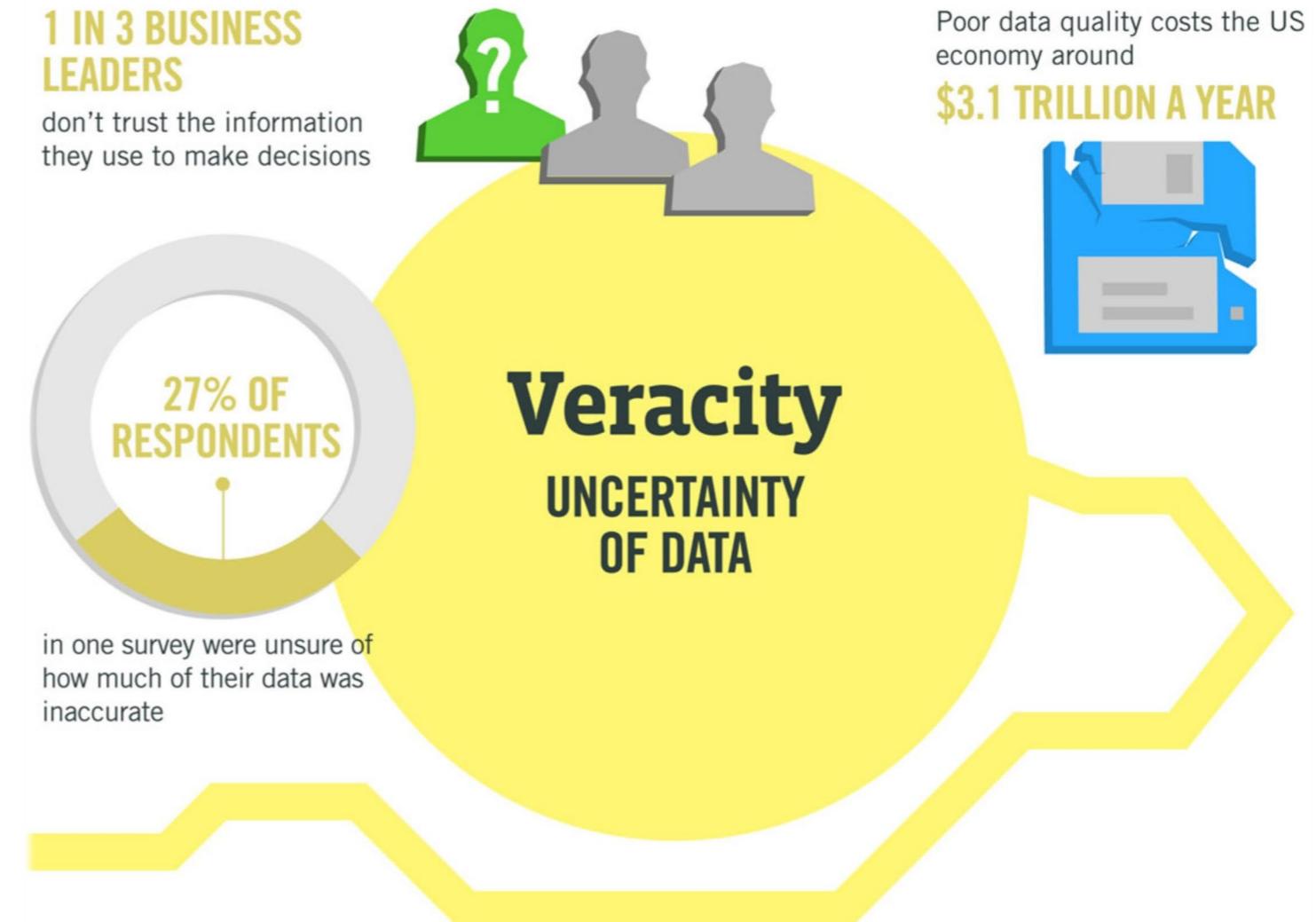
(nature et niveau de structuration)

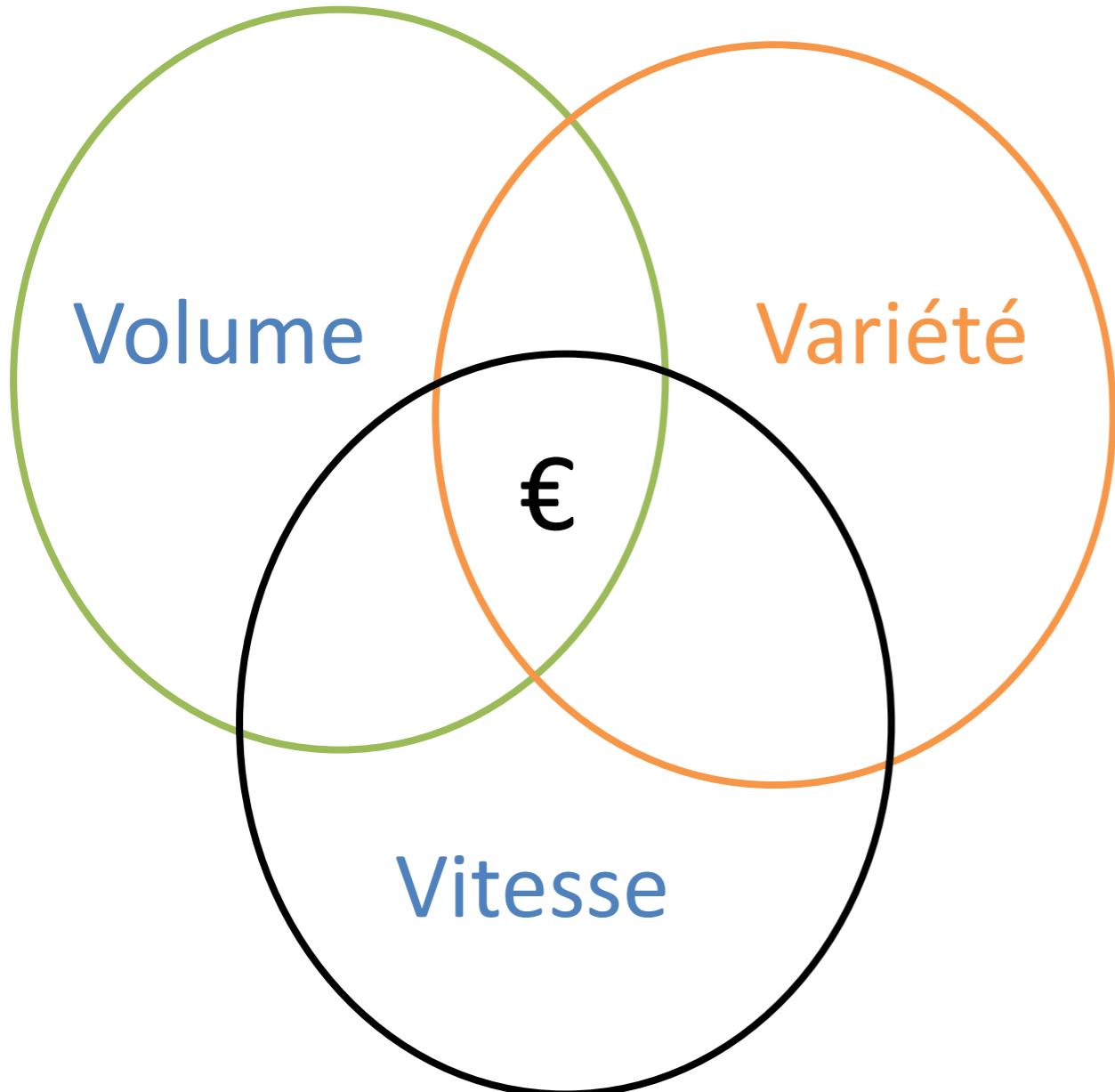
+

V - pour **Véracité**

(l'exactitude des données)

Pour comprendre le Big Data,
IBM synthétise ce phénomène par
quatre spécificités majeures : les 4 V. **Le Volume, la Variété, la Vélocité et la Véracité.**

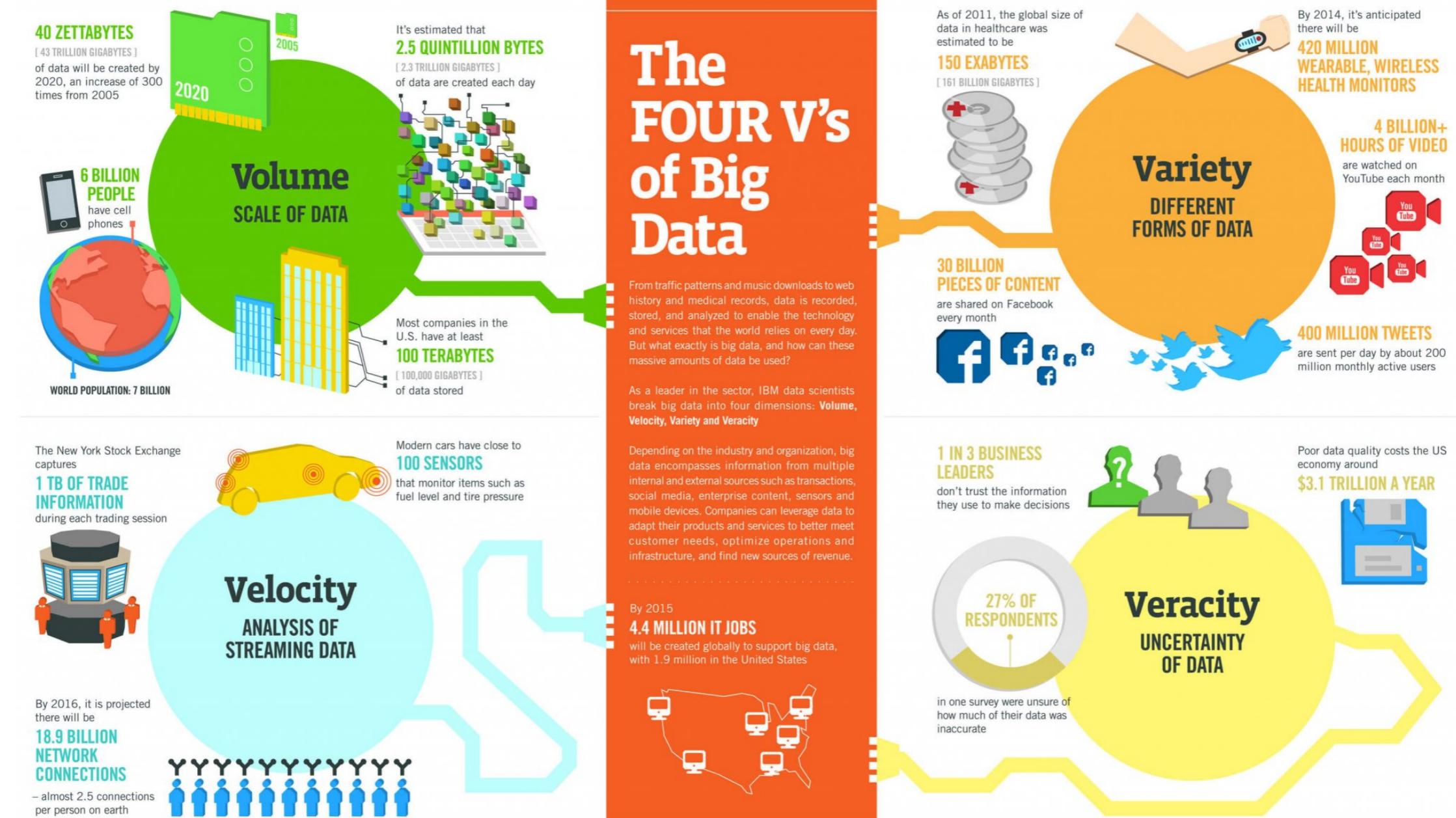




La maîtrise des 3v génère de la
Valeur



- Valeur stratégique
- Valeur opérationnelle
- Nouvelles opportunités



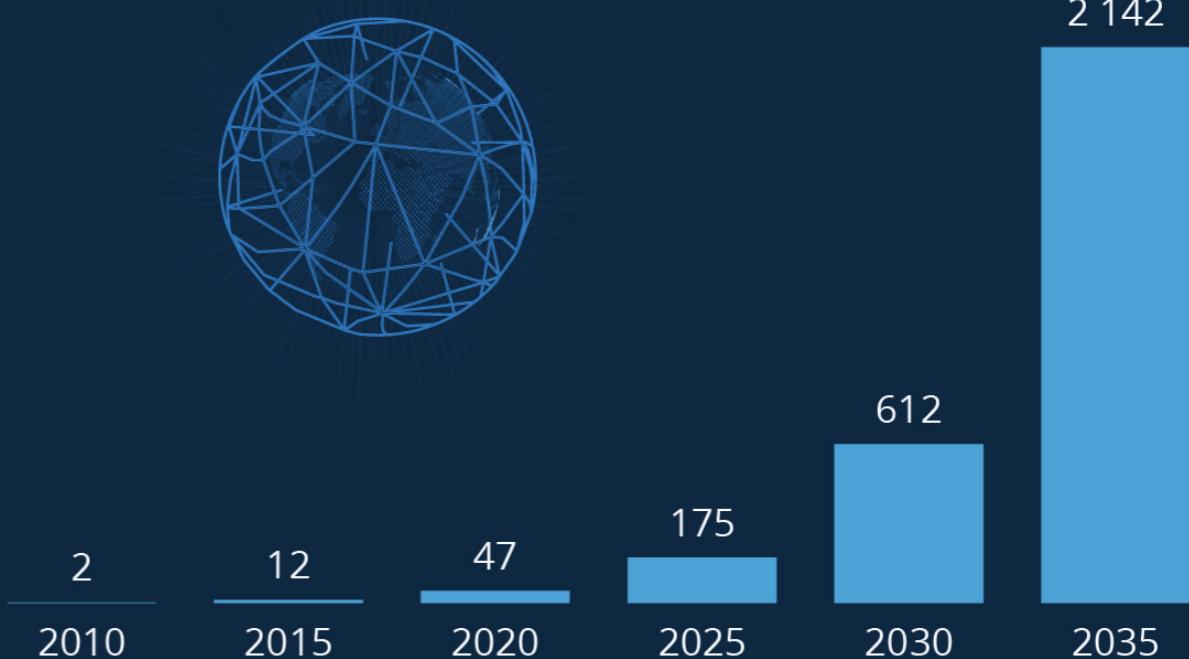
Le Big Data est un terme utilisé pour décrire une collection de données qui croît de façon exponentielle avec le temps.

En fait, ces données sont si volumineuses et complexes qu'aucun des outils traditionnels de gestion des données n'est capable de les stocker ou de les traiter efficacement



Le big bang du big data

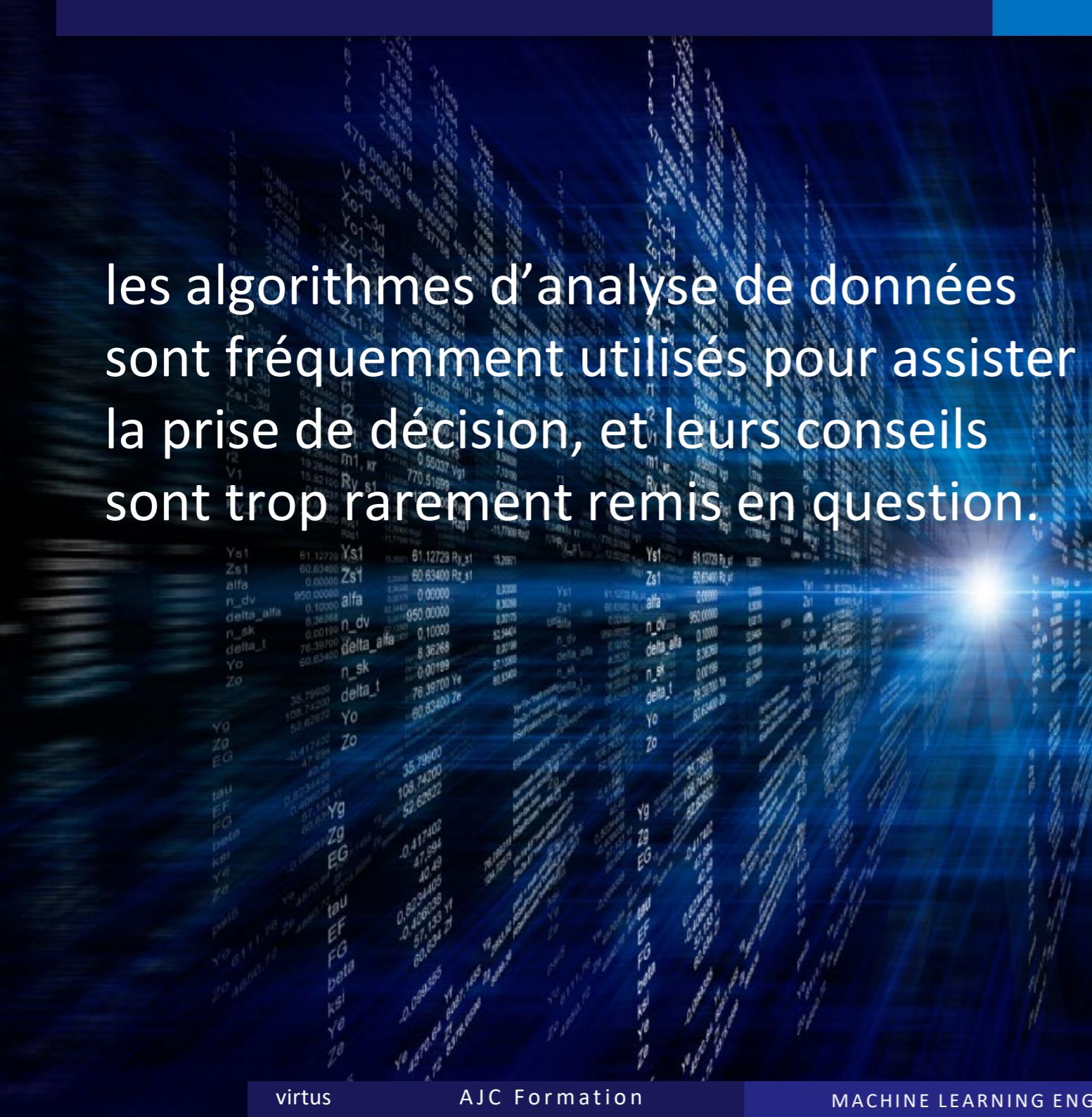
Volume annuel de données numériques créées à l'échelle mondiale depuis 2010, en zettaoctets *



* Prévisions de 2020 à 2035. Un zettaoctet équivaut à mille milliards de gigaoctets.
Source : Statista Digital Economy Compass 2019



Les points faibles



les algorithmes d'analyse de données
sont fréquemment utilisés pour assister
la prise de décision, et leurs conseils
sont trop rarement remis en question.

Les risques Du Big Data

Le Big Data qu'est-ce que c'est ? – Les risques Du Big Data

- **la vie privée** (le paradoxe des consommateurs)

Chaque minute Google enregistre 3,8 millions de requêtes différentes sur son moteur de recherche, et Facebook autant de « likes ».



Adresses Ip



Géo localisation

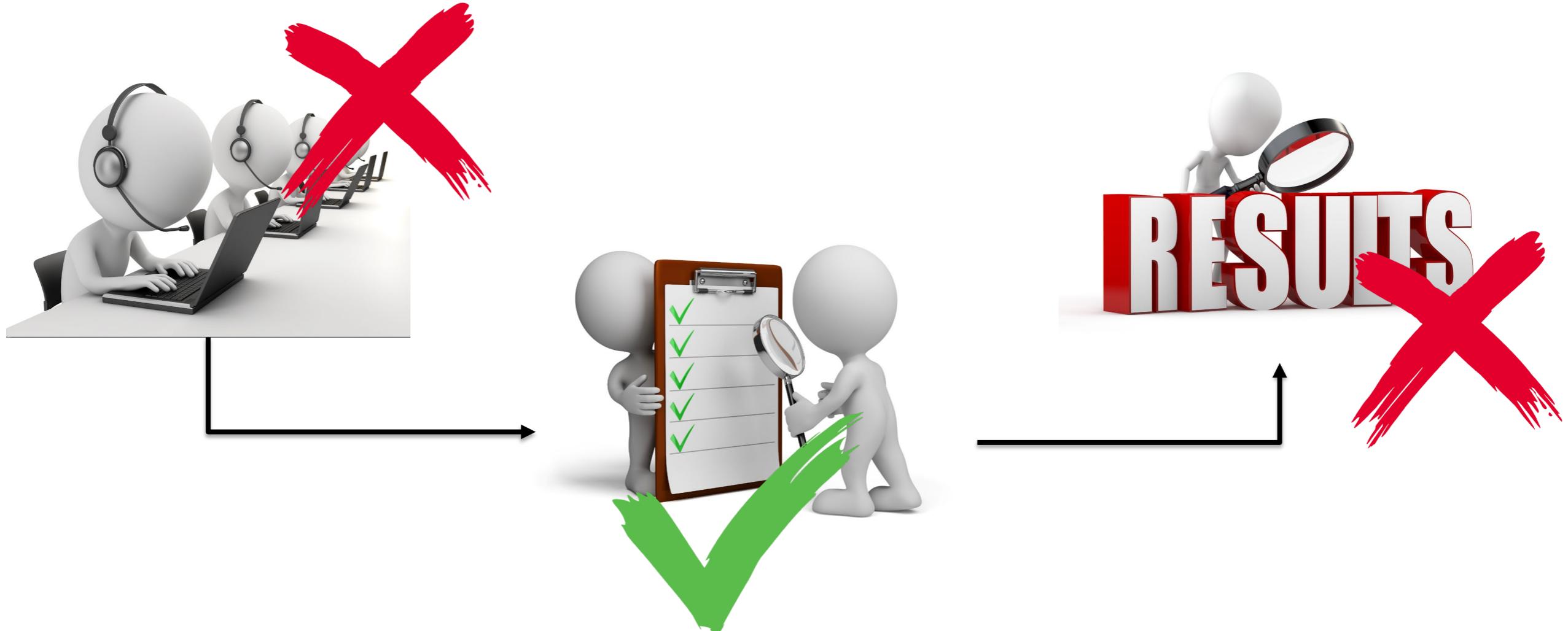


Informations



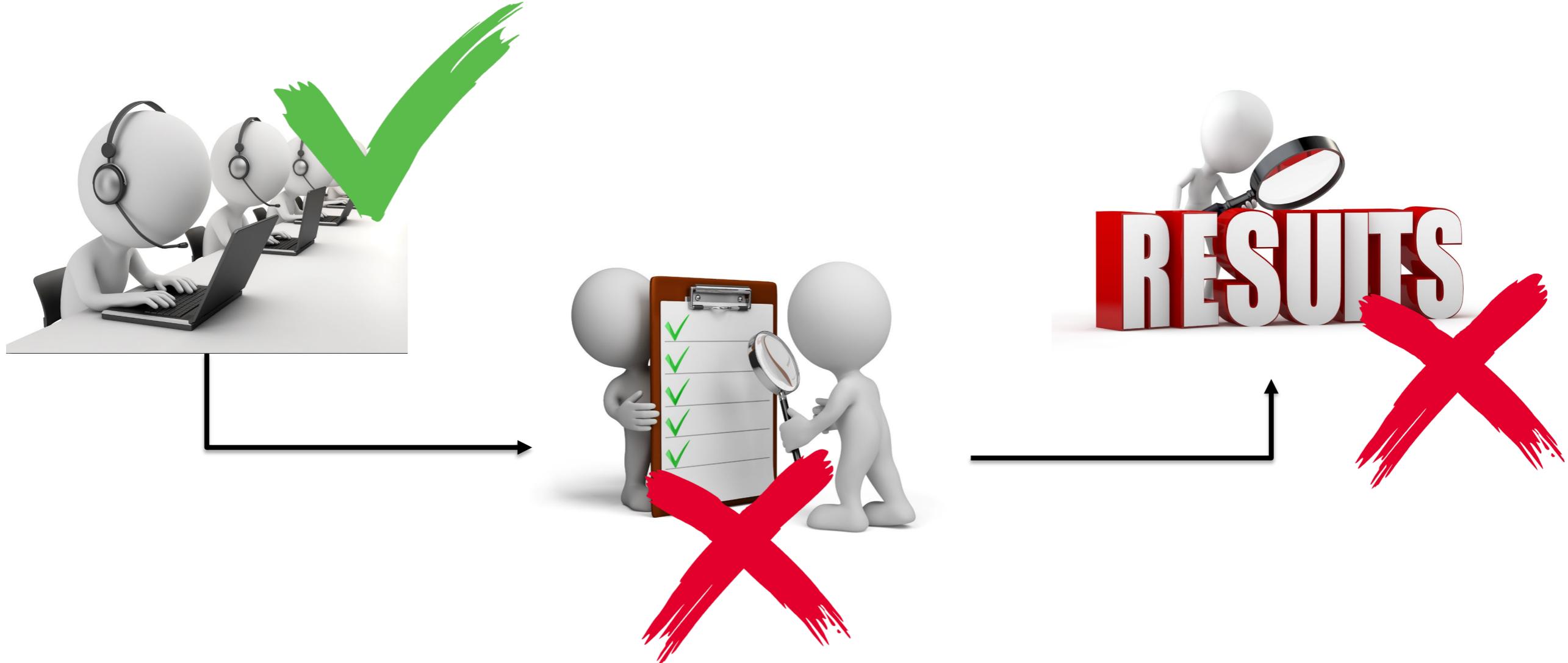
Le Big Data qu'est-ce que c'est ? – Les risques Du Big Data

- **la vie privée** (le paradoxe des consommateurs)
- **Analytique** (Récolter les données d'une manière erronée)



Le Big Data qu'est-ce que c'est ? – Les risques Du Big Data

- **la vie privée** (le paradoxe des consommateurs)
- **Analytique** (Récolter les données d'une manière erronée)
- **Stratégiques** (un problème d'analyse)



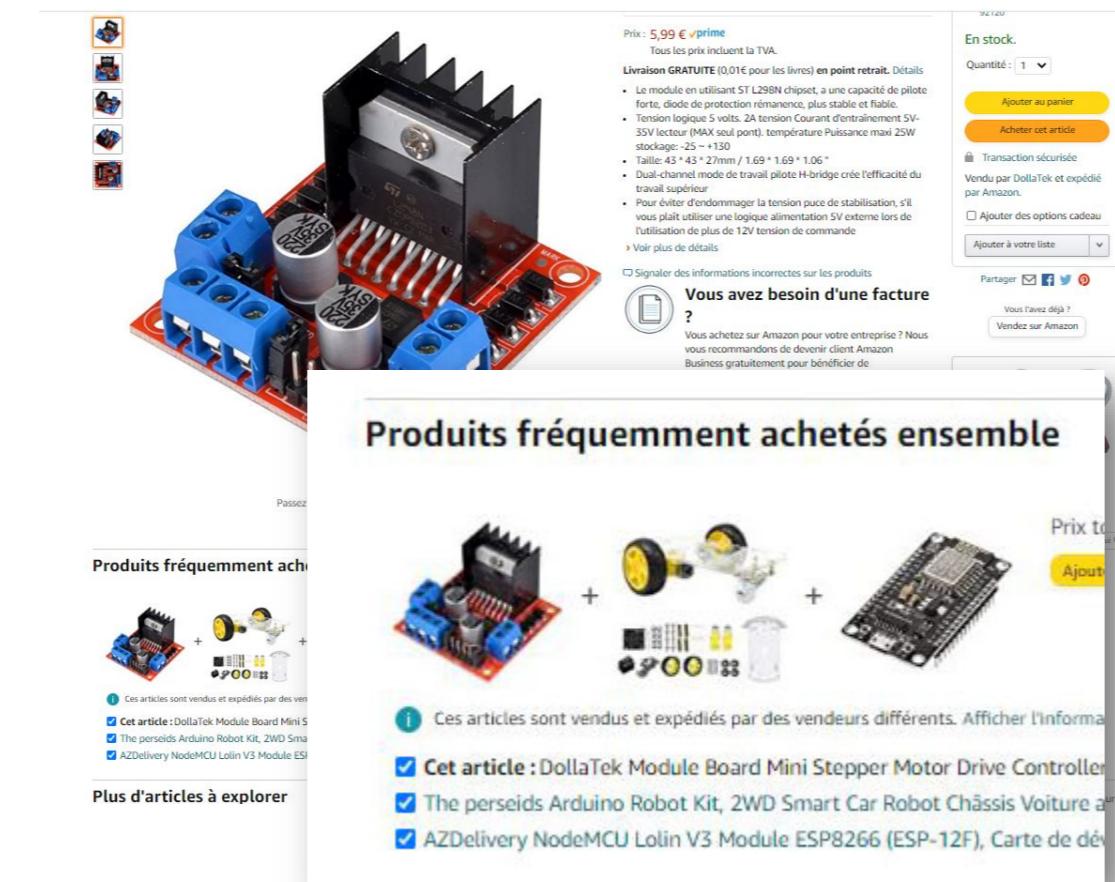
A qui bénéficie le big data



Le Big Data qu'est-ce que c'est ? – A qui bénéficie le big data

Le Big Data est appliqué dans **tous les domaines liés au Web**.

Un exemple d'outil de Big Data dans le domaine de l'e-commerce est la fameuse phrase « ceux qui ont acheté le produit X ont aussi acheté... ». Ces recommandations naissent à partir de l'évaluation de millions de données d'achats d'autres clients.



Le Big Data qu'est-ce que c'est ? – A qui bénéficie le big data

Voici quelques domaines qui tirent profit du Big Data :

La recherche médicale : grâce à l'évaluation des données massives, les médecins peuvent trouver de meilleures solutions de thérapie et de traitement pour leurs patients.

L'industrie : grâce à l'utilisation des données des machines de la chaîne de production par exemple, les entreprises peuvent augmenter l'efficacité de leur production et travailler de manière plus durable.

L'économie : il permet aux entreprises de mieux connaître leurs clients et de leur proposer des offres toujours plus adaptées à leurs besoins.

L'énergie : les données sur la consommation d'énergie permettent à long terme d'adapter l'offre aux besoins des utilisateurs dans le but de rendre l'approvisionnement énergétique plus responsable et durable.

Le marketing : le Big Data est utilisé dans le marketing pour mieux cibler les clients. L'objectif est, entre autres, d'améliorer les relations avec les consommateurs et d'augmenter le taux de conversion.

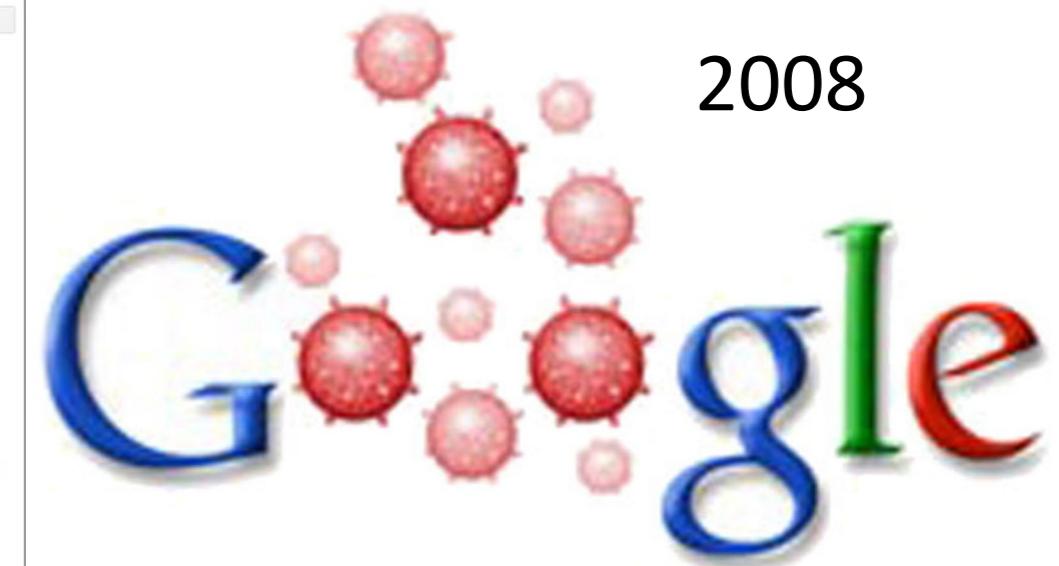
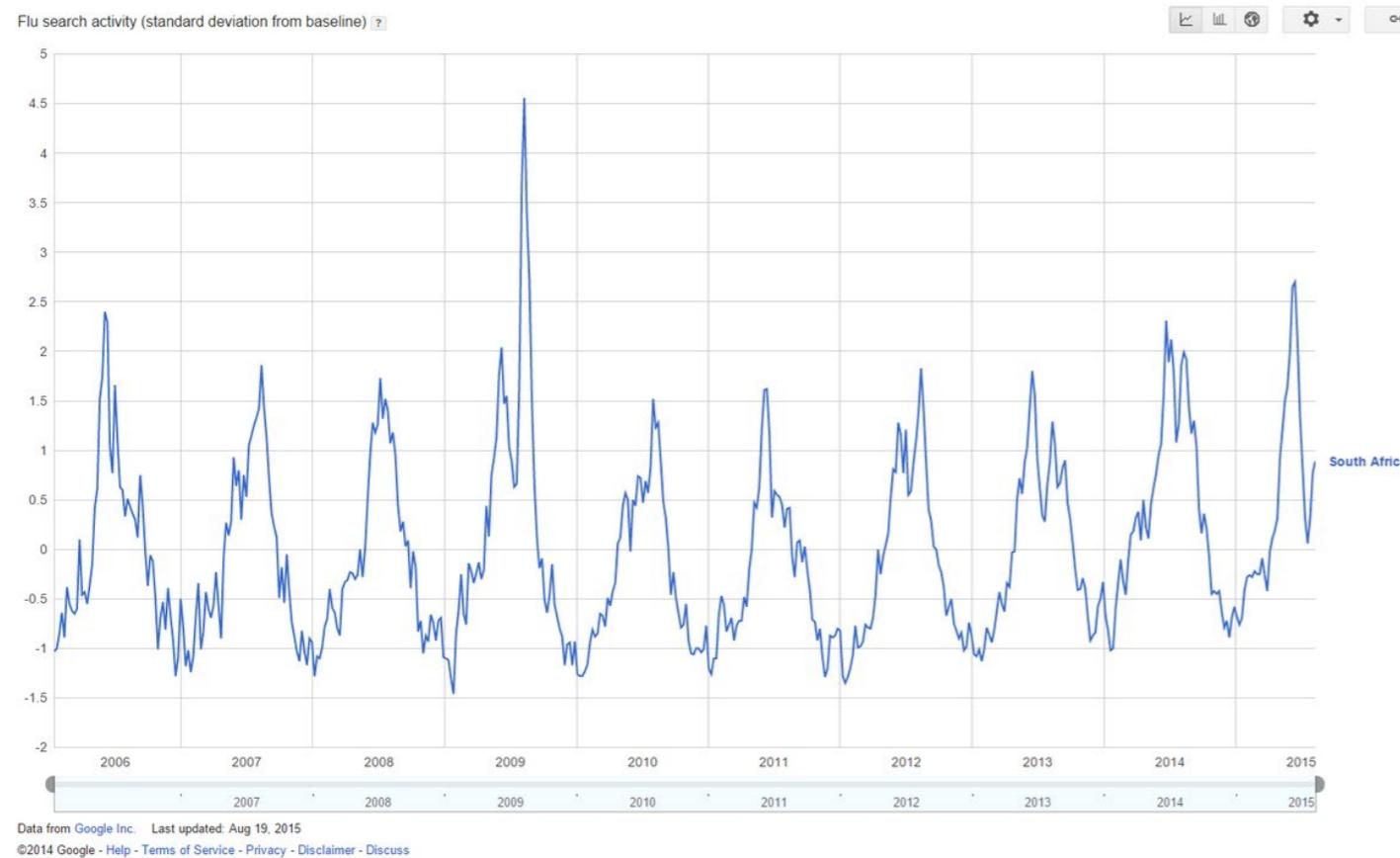
Le secteur bancaire : le Big Data permet à une banque de proposer des services adaptés au profil de ses clients ou de mieux anticiper ses risques de défaut ou de liquidité.

De nouveaux métiers sont apparus pour gérer toutes ces données, mais aussi pour les exploiter.

On a en effet découvert qu'on pouvait apprendre beaucoup de leur étude.

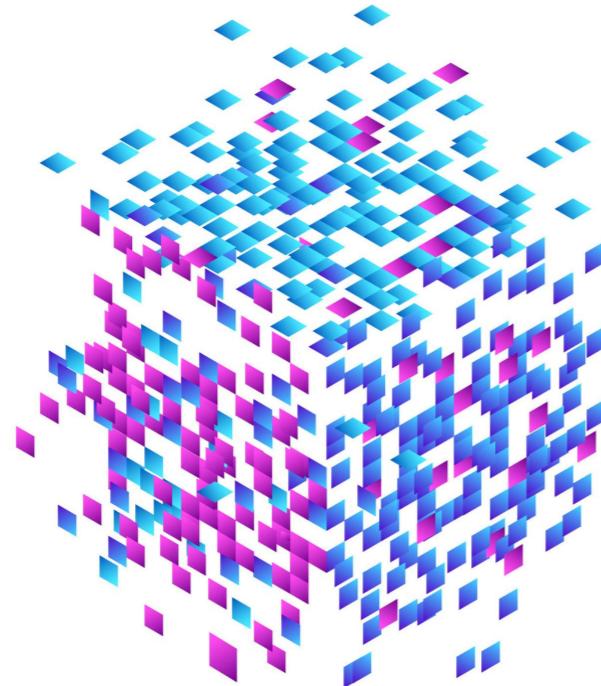


Le Big Data qu'est-ce que c'est ? – A qui bénéficie le big data

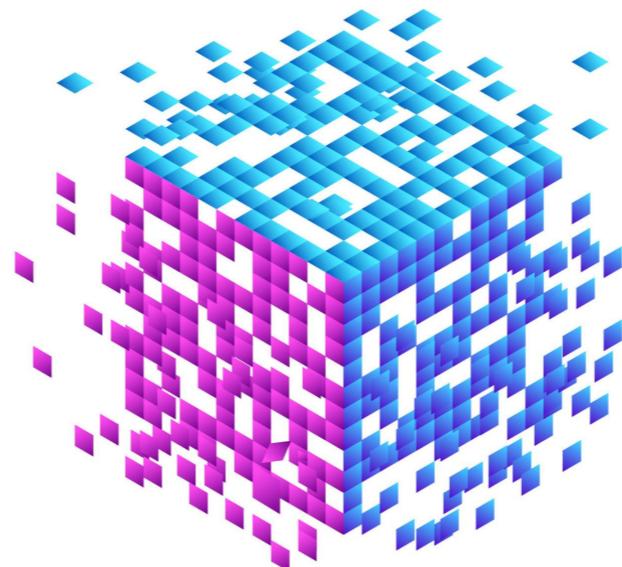


Son système *Google Flu Trends* serait plus rapide que les services d'alerte habituels

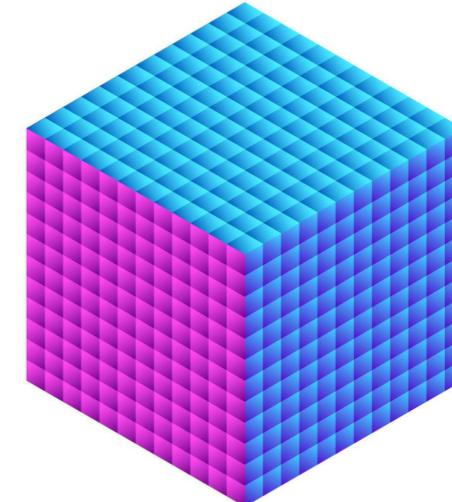
BIG DATA



ANALYTICS



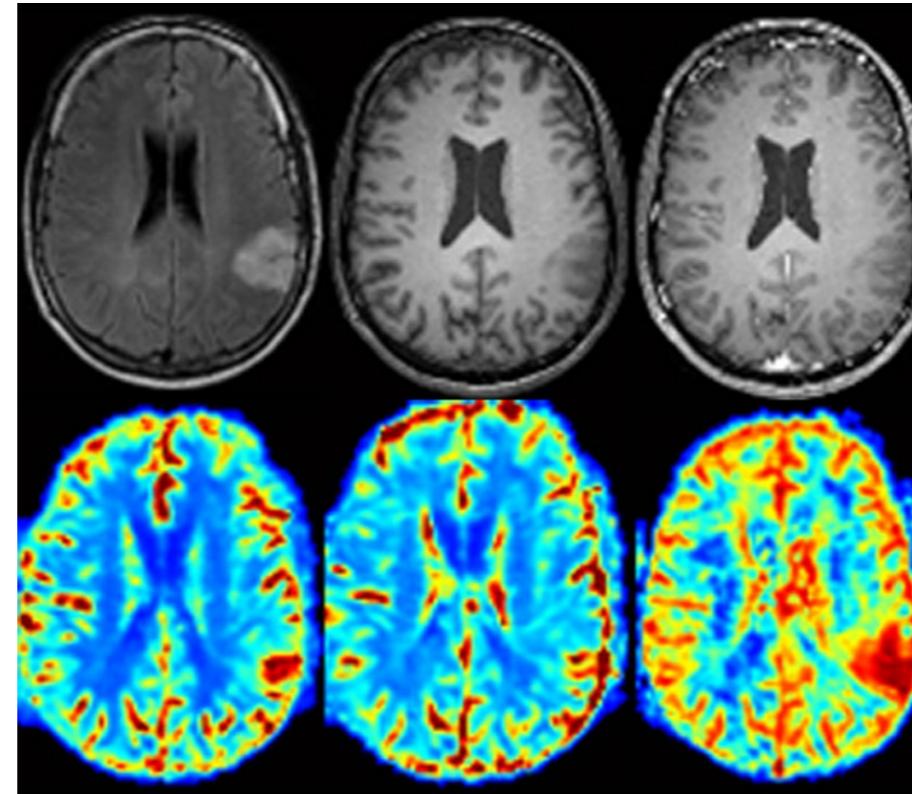
DECISIONS



Le Big Data qu'est-ce que c'est ? – A qui bénéficie le big data

Des algorithmes, qui apprennent des situations précédentes pour prévoir ce qui pourrait se passer, existent dans différents domaines : identification de personnes sur des photos, sur Facebook par exemple⁶, ou détection de cellules malades sur des images d'IRM.

Un programme informatique capable de détecter et d'identifier automatiquement des lésions cérébrales



<https://presse.inserm.fr/un-programme-informatique-capable-de-detecter-et-didentifier-automatiquement-des-lesions-cerebrales/31634/>

Le Big Data qu'est-ce que c'est ? – A qui bénéficie le big data

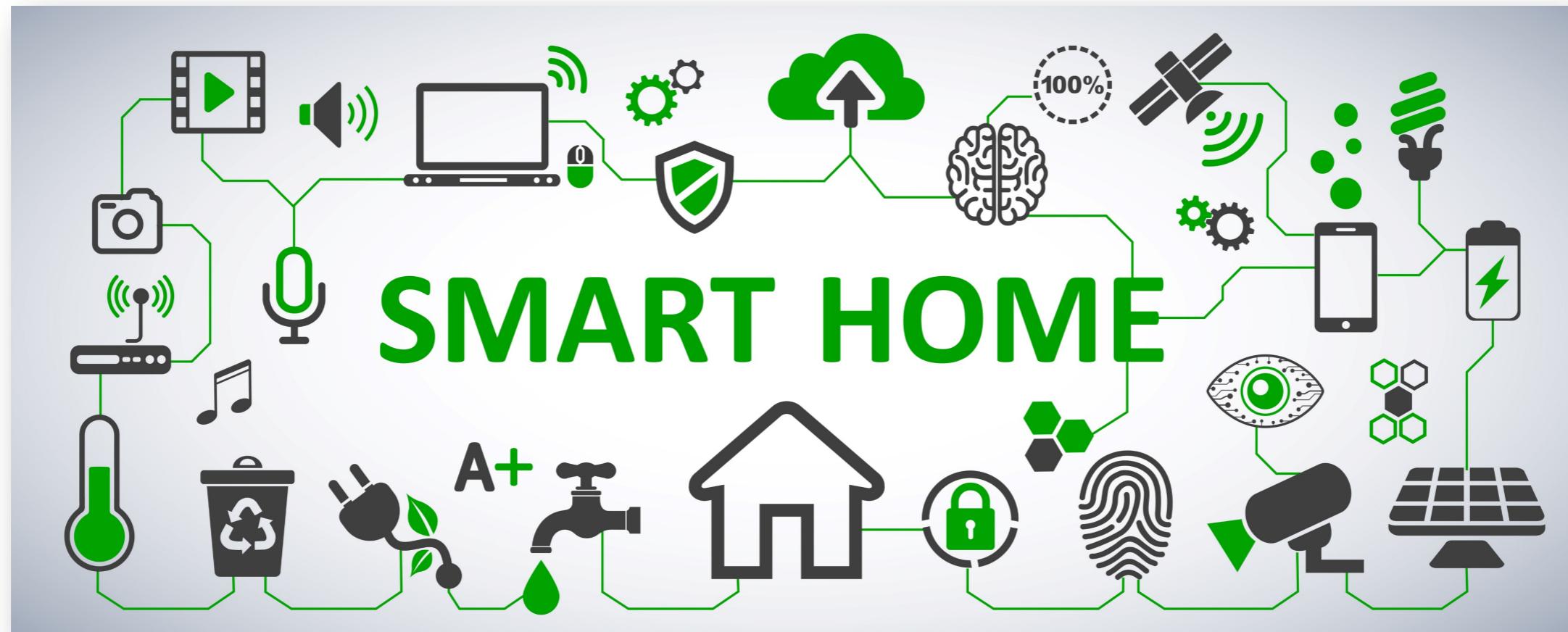
Déjà certains algorithmes apprennent et décident à partir de situations existantes qui peuvent être biaisées. Ainsi, les algorithmes de Google proposent-ils des offres d'emploi. Mais une étude a montré que les postes les plus prestigieux étaient proposés aux hommes plutôt qu'aux femmes.

Les formules mathématiques font aussi des erreurs – **on parle de faux positifs**

<https://www.theguardian.com/uk-news/2014/dec/02/youre-the-bomb-are-you-at-risk-from-anti-terrorism-algorithms-automated-tracking-innocent-people>



Une accélération grâce aux objets connectés



Les évolutions technologiques derrière le Big Data

- **Comment s'y retrouver**
- **Le traitement**



technologiques derrière le Big Data

Comment s'y retrouver

Les évolutions technologiques derrière le Big Data – la rencontre avec hadoop

1997



Doug Cutting

Projet
Lucene



2000



Mike Cafarella

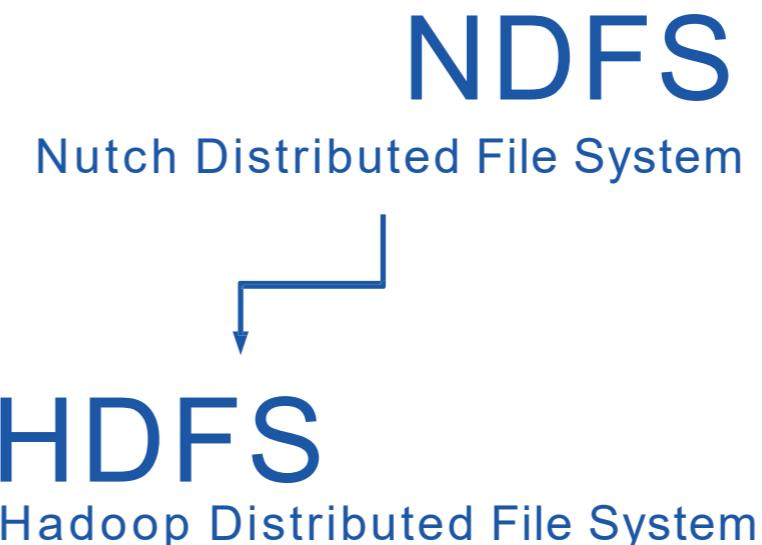
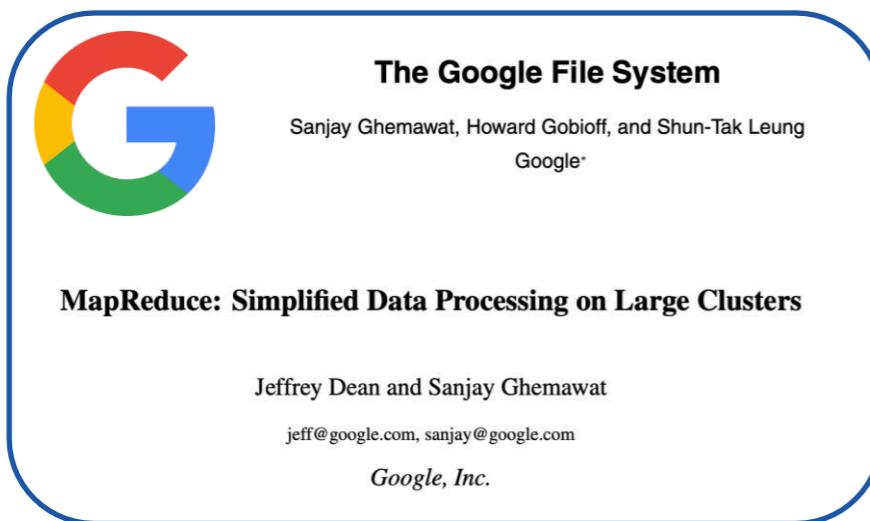
Projet
Nutch



Indexer les pages Web

Manque de performance de Nutch

- Supporter les pannes de serveurs sans incidence sur la recherche et l'indexation
- Distribuer les traitements dynamiquement sur d'autres serveurs
- Supporter la perte des données d'un disque dur



2006



yahoo!
CLOUDERA



Apache
STORM



APACHE
kafka™

Le big bang du big data tant matériel que logiciel la rupture

- Nécessite de nouvelles architectures informatiques

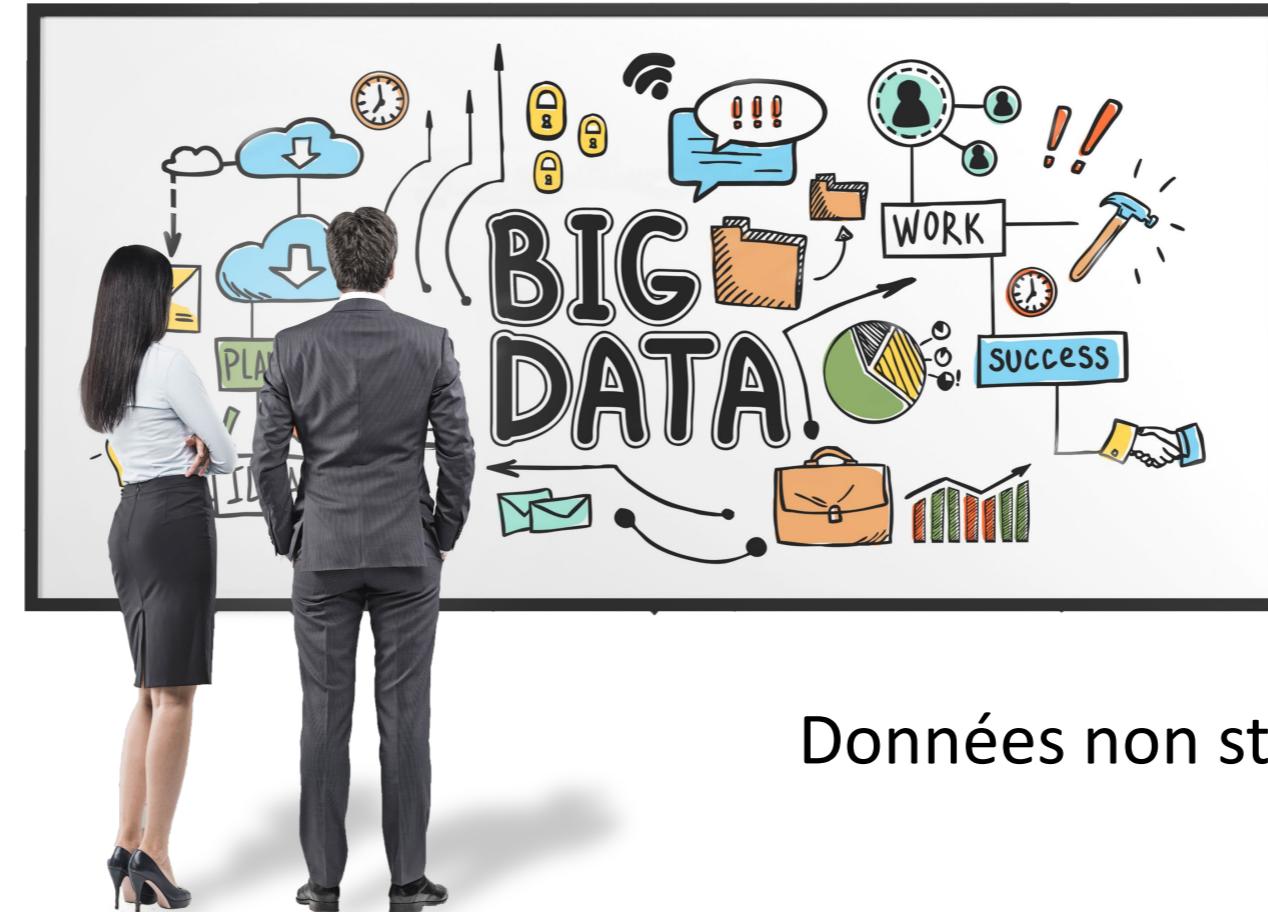
la rupture technologique a été lancée par Google et suivi d'Amazon Facebook Twitter et Netflix leur solution répondre à **4 besoins spécifiques :**

- le besoin de paralléliser le traitement sur un stockage distribué
- le besoin de permettre une montée en charge rapide
- le besoin de gérer des données non structurées et variées
- le besoin d'analyser des données en flux continu

face à ces 4 besoins les technologies principales mises en œuvre sont

Parallélisme = **MapReduce**

Montée en charge =
Cloud Computing



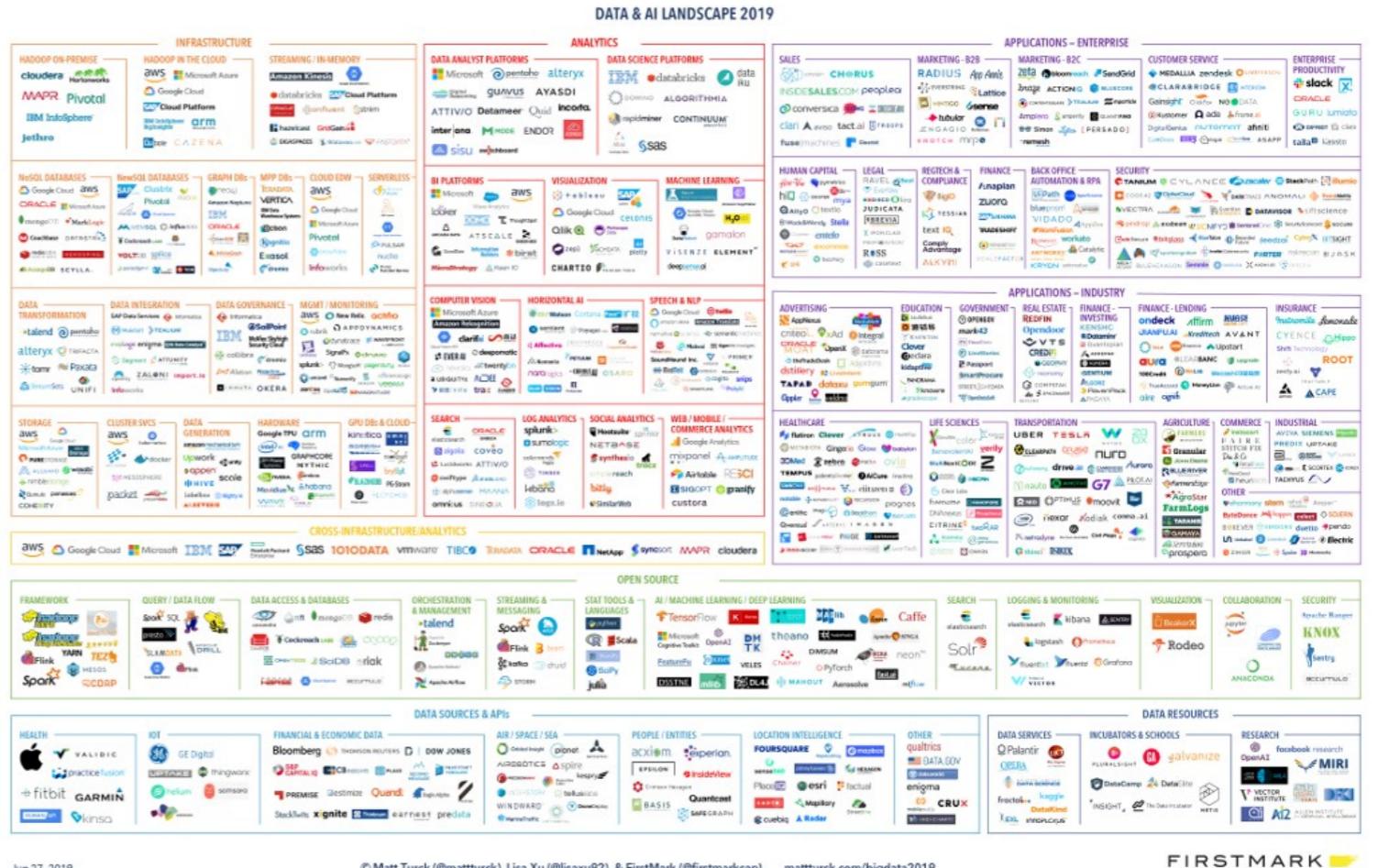
Flux continu =
outils comme **Kafka**

Données non structurées = **NoSQL**

Les évolutions technologiques derrière le Big Data – comment s'y retrouver

le Big data n'est pas une activité unique mais se découpe en une multitude d'activités différentes

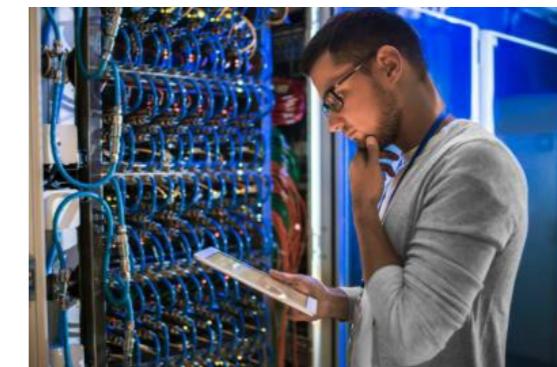
- Gestion de l'infrastructure
 - Les Outils d'analytiques
 - Bases de données NoSQL
 - Outils de visualisation
 - Outils d'intégration de données
 - Langages de programmation dédiés
 - Algorithmes spécifiques pour la parallélisation des traitements
 - Intelligence artificielle avec le Machine Learning et le Deep Learning



Les évolutions technologiques derrière le Big Data – comment s'y retrouver

Le Big Data s'adresse à des profils totalement différents

- Chief Data Officer, en charge de la **gouvernance** des données pour l'entreprise
- Data Engineer, spécialisé dans la **gestion** des données
- Data Scientist, spécialiste de l'**analyse** des données massives
- Architecte Big Data, **concepteur** des solutions
- Développeur Big Data, maîtrisant les **langages** de développeur
- Administrateur Big Data, en charge de l'**opérabilité** de la plate-forme de Big Data



Les évolutions technologiques derrière le Big Data – comment s'y retrouver

Hadoop est le socle technique du Big data

Le Framework Open Source Java de la fondation Apache. Conçu pour le développement de systèmes de fichiers distribués permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données

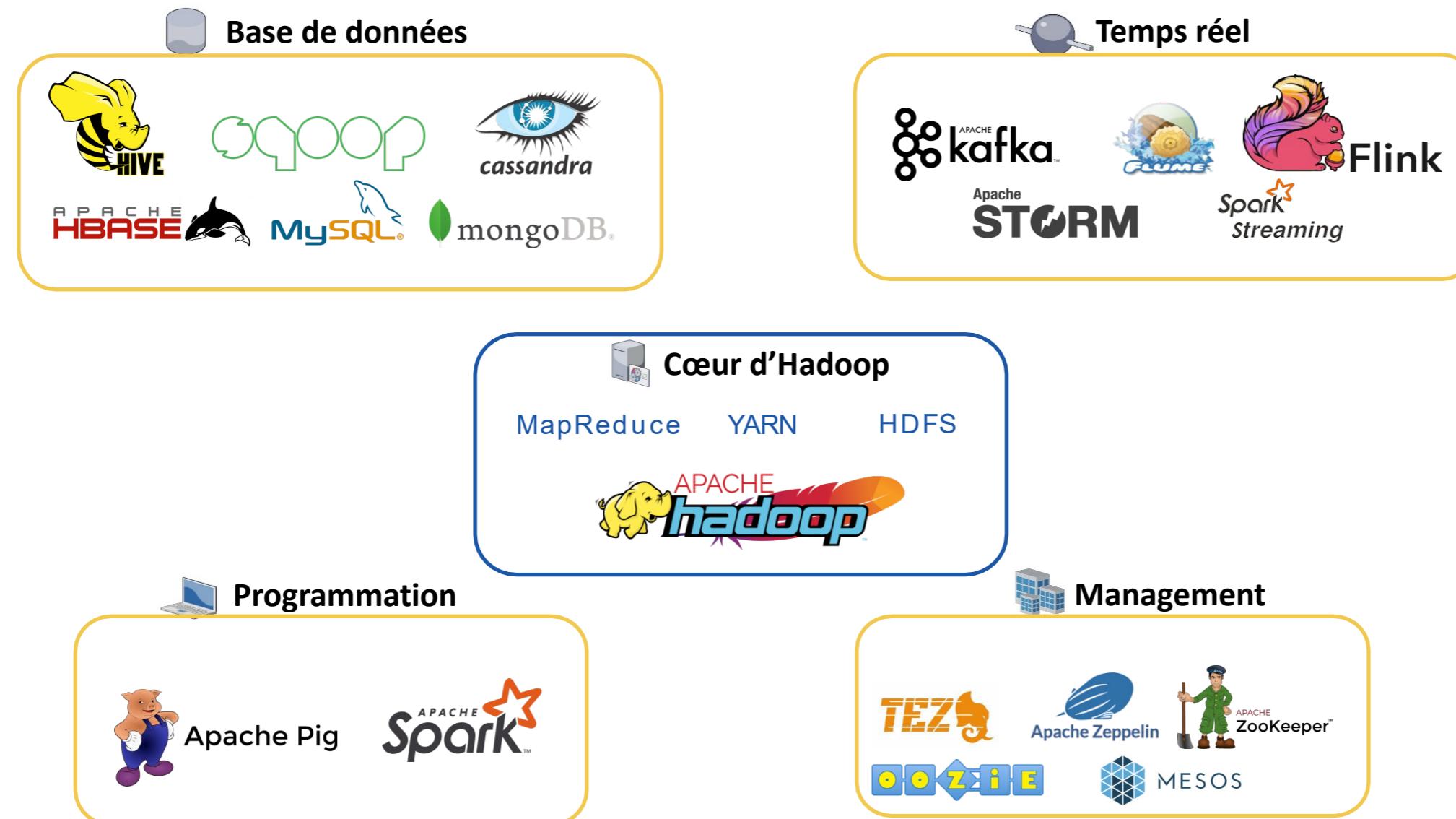
Architecte

Développeur

Administrateur



Le traitement





Cœur d'Hadoop

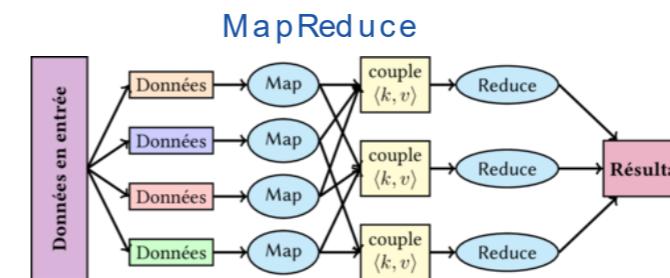
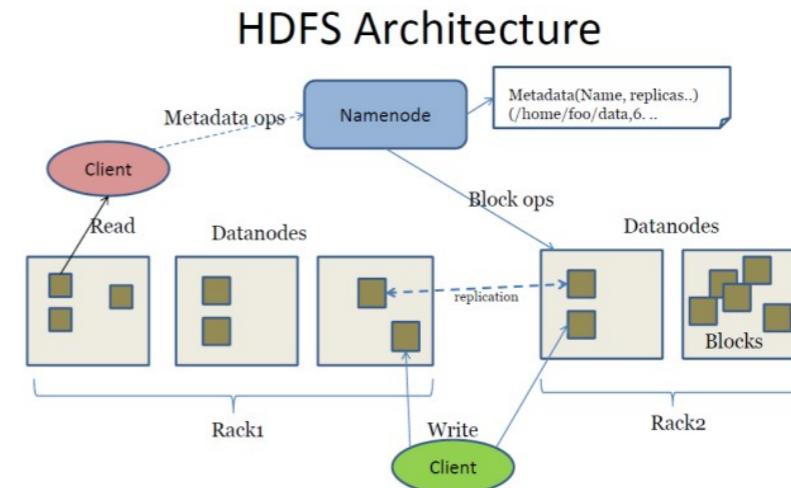
MapReduce YARN HDFS



HDFS : système de fichiers d'Hadoop

MapReduce : parallélise le traitement des données

YARN : modèle de traitement pour HDFS



Les évolutions technologiques derrière le Big Data – Le traitement

- HDFS : le système de fichier distribué d'Apache Hadoop



HDFS – Hadoop Distributed File System



Définition: bloc

Un *bloc* est une zone mémoire contigüe de taille fixe stockée sur disque, lue ou écrite solidairement. *Le bloc est l'unité d'entrée/sortie entre la mémoire secondaire et la mémoire principale.*

Les évolutions technologiques derrière le Big Data – Le traitement

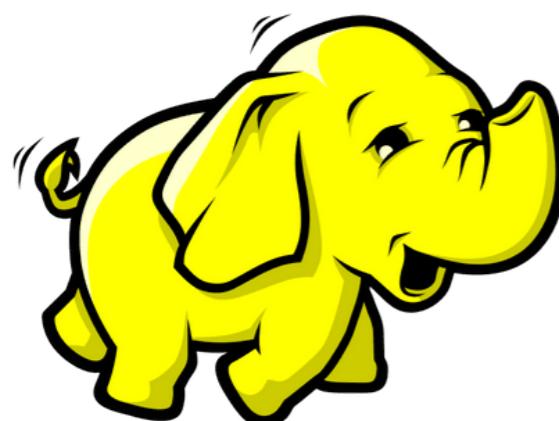


HDFS – Hadoop Distributed File System



1 = 65 M -512 M

1 = 8k



Un bloc = 1

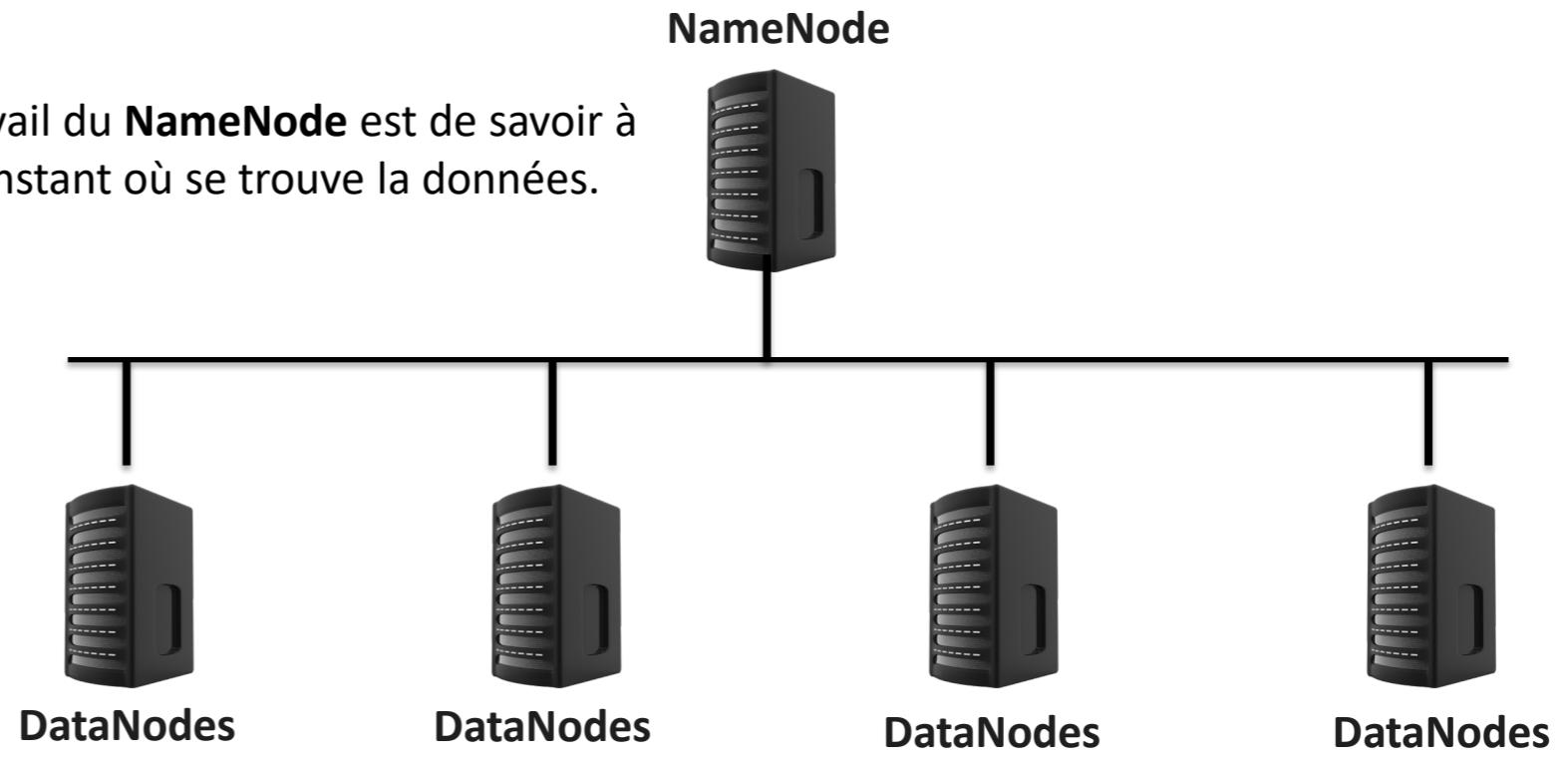


HDFS – Hadoop Distributed File System

Le système de fichiers HDFS comporte deux principaux composants qui sont le **NameNode** et les **DataNodes**.

les **DataNodes** qui se chargent du stockage des données et nécessitent donc plus de ressources de stockage

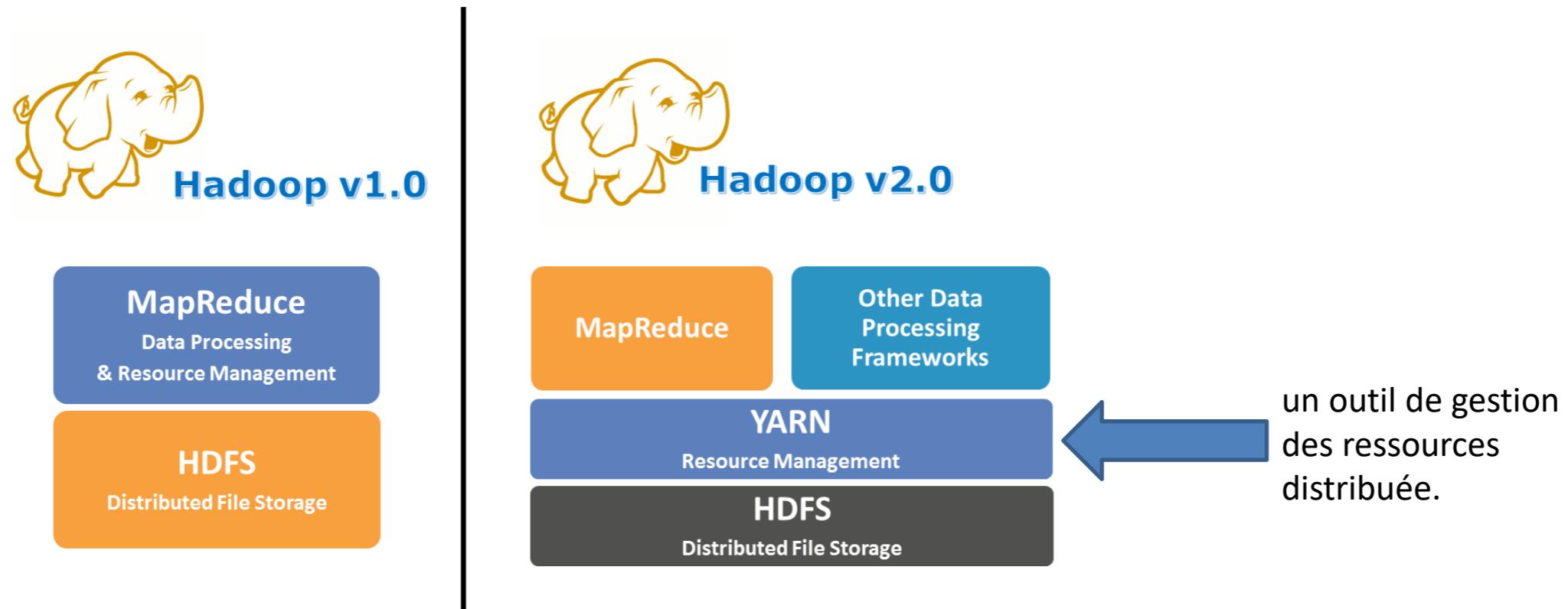
le travail du **NameNode** est de savoir à tout instant où se trouve la données.



Les évolutions technologiques derrière le Big Data – Le traitement

YARN – Yet Another Resource Negotiator

YARN est un gestionnaire de ressources au sein d'un cluster Hadoop. Il a pour but de planifier et allouer les ressources au sein des clusters Hadoop. Un peu comme HDFS qui se charge de la gestion du stockage, YARN se charge de la puissance de calcul.

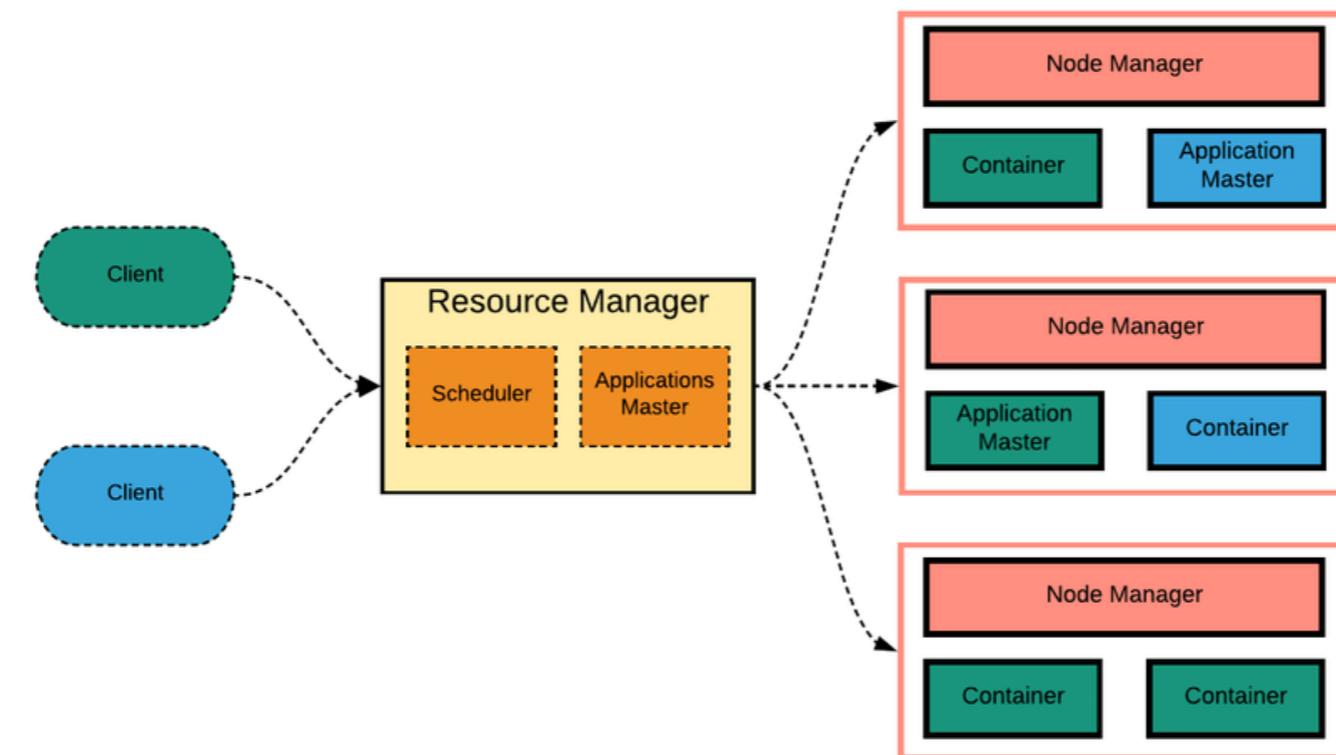


La description de Hadoop comme possédant 2 couches (MapReduce et HDFS) est correcte pour la version 1 de Hadoop. Depuis la version 2, Hadoop a adopté une troisième couche : YARN ("Yet Another Resource Negotiator"), un outil de gestion des ressources distribuée.

Les évolutions technologiques derrière le Big Data – Le traitement

YARN est l'ensemble de trois principaux composants :

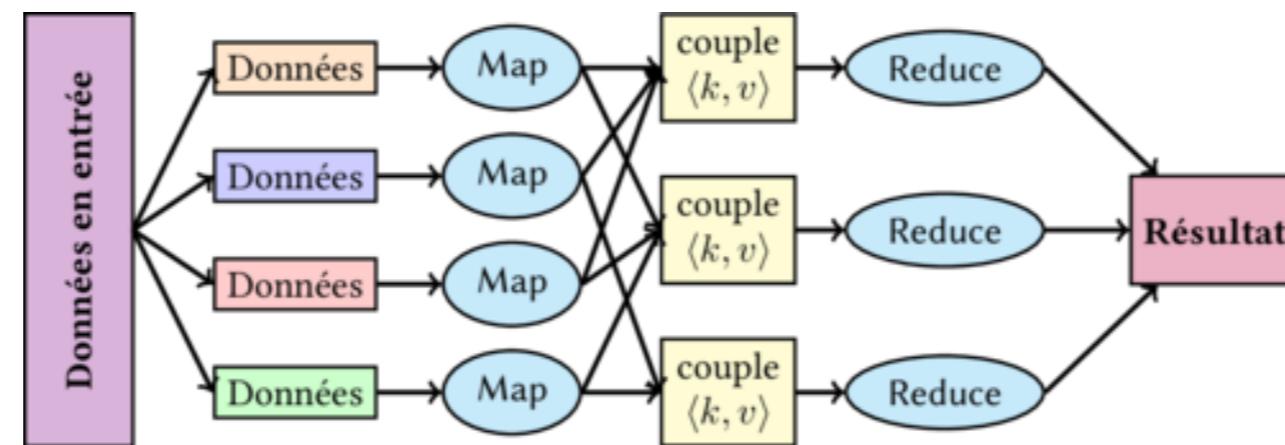
- Le “resource manager” qui est chargé de recevoir les demandes de calculs et de les dispatcher aux **nodes managers**
- Les “**nodes managers**” qui sont installés sur chaque **datanode** se chargeront d’effectuer des calculs au sein de ce **datanode**.
- L’**application manager** qui fonctionne comme une interface de négociation de ressource **entre le resource manager et des nodes managers**. C'est grâce à lui que le **resource manager** arrive à connaître la puissance de calcul disponible dans chaque conteneur (l'environnement dans lequel les processus de chaque **datanode** sont exécutés).





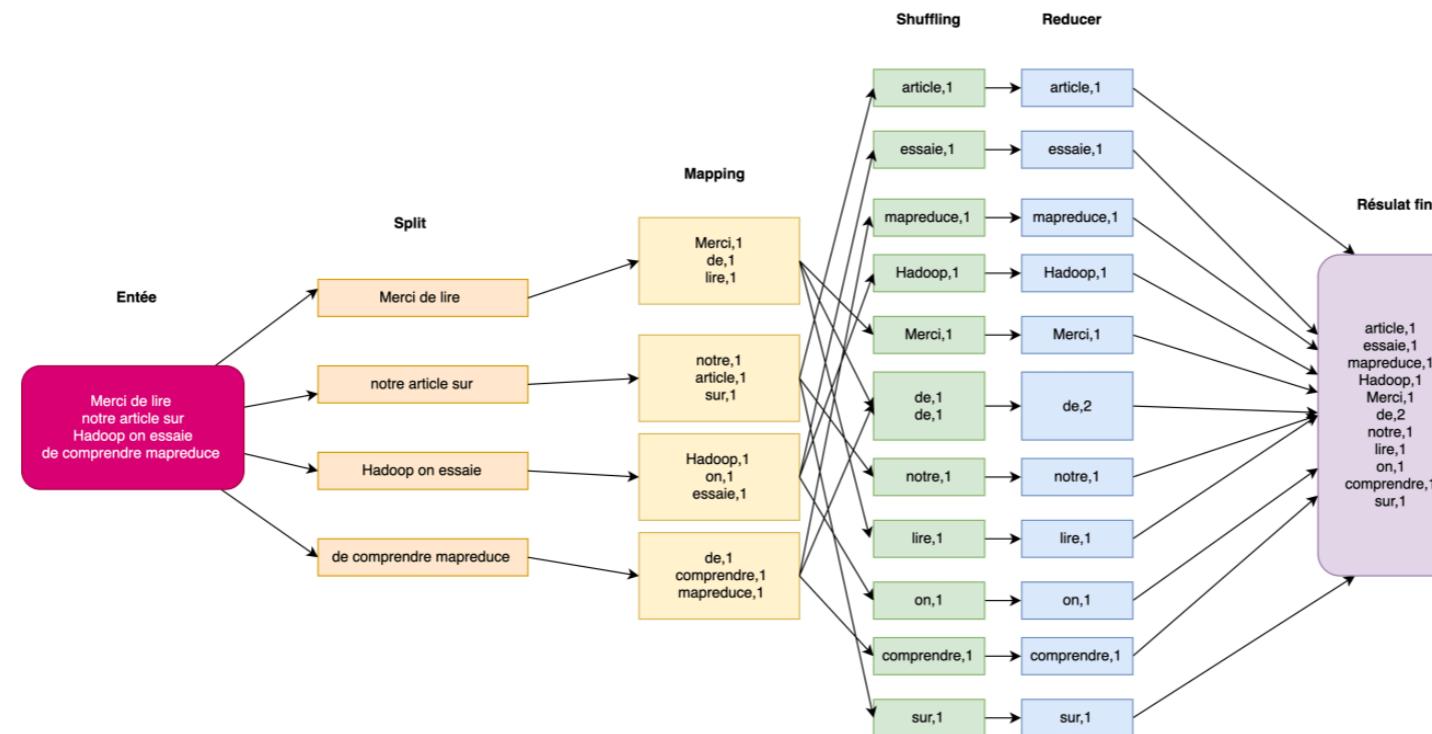
C'est un modèle de programmation présent au sein d'Hadoop qui permet d'effectuer des calculs en parallèle. L'idée derrière MapReduce est de pouvoir diviser tous les traitements en deux parties (**Map** et **Reduce**)

"pour faire simple". Les opérations de Map consistent à filtrer, trier et regrouper les données sous forme de clé-valeur (tuple). Ensuite, les opérations de Reduce auront pour but de combiner ces tuples afin d'obtenir le résultat souhaité.

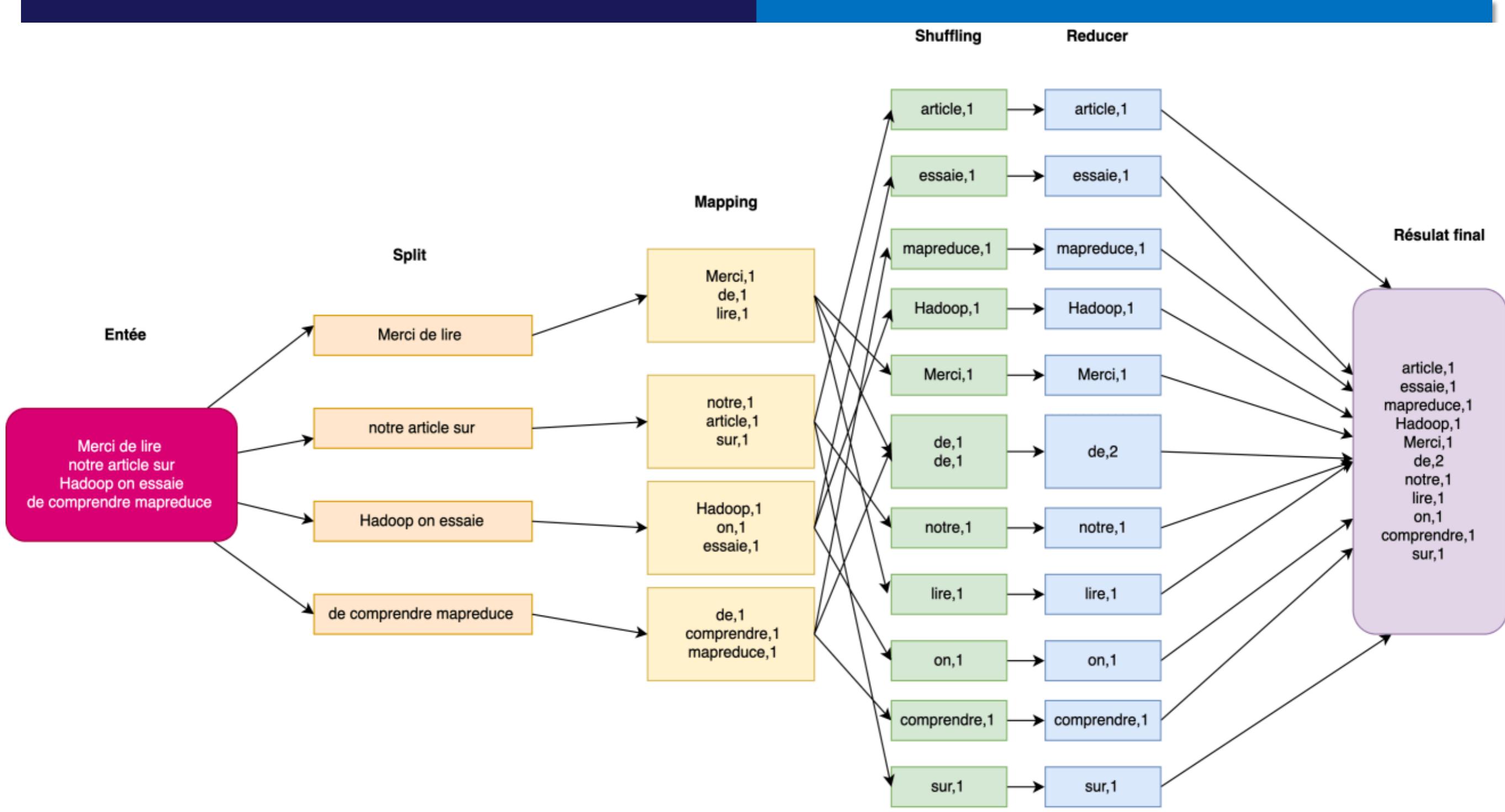


Un exemple concret :

Imaginons qu'on aimerait compter l'occurrence de chaque mot dans un texte de façon distribuée. L'objectif sera de découper le texte en petit morceau et compter l'occurrence des mots de chaque morceau du texte et en faire la somme. C'est le rôle de MapReduce.



Ledatascientist





Base de données



Hive : base de données de type DataWarehouse (entrepôt de données)

HBase : base de données NoSQL avec stockage en colonnes

Sqoop : logiciel d'import/export de données relationnelles

MySQL : base de données relationnelles

Cassandra : base de données NoSQL

MongoDB : base de données avec stockage en documents



Programmation



Apache Pig



Pig : langage de programmation (Pig Latin) de haut niveau masquant la complexité MapReduce

Spark : plate-forme proposant des APIs de manipulation de données d'un cluster HDFS

```
input_lines = LOAD '/tmp/my-copy-of-all-pages-on-internet' AS  
(line:chararray);  
  
-- Extract words from each line and put them into a pig bag  
-- datatype, then flatten the bag to get one word on each row  
words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS  
word;  
  
-- filter out any words that are just white spaces  
filtered_words = FILTER words BY word MATCHES '\\w+';  
  
-- create a group for each word  
word_groups = GROUP filtered_words BY word;  
  
-- count the entries in each group  
word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS  
count, group AS word;  
  
-- order the records by count  
ordered_word_count = ORDER word_count BY count DESC;  
STORE ordered_word_count INTO '/tmp/number-of-words-on-internet';
```



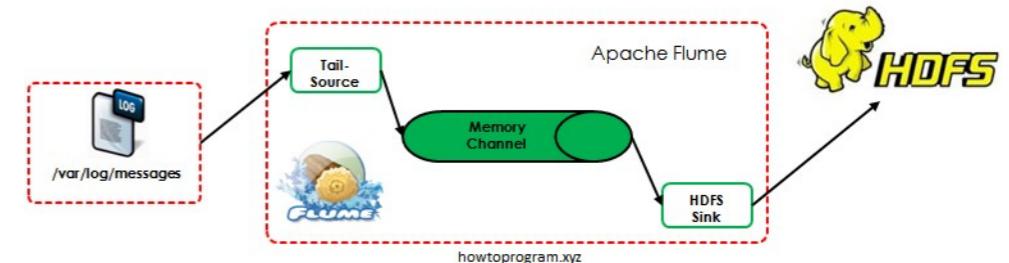
Kafka : récupérer des données provenant d'objets connectés

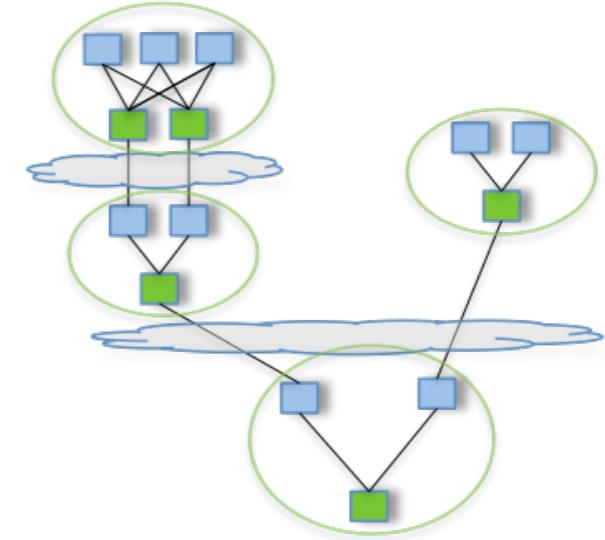
Flume : récupérer des données de fichiers log

Flink : traiter des flux de données

Storm : moteur de déploiement de calcul sur un flux de données

SparkStreaming : gestion de flux de données en temps réel





Tez : permet une représentation graphique de tâches (DAG)

Zeppelin : offre une mise en forme visuelle de données traitées par Spark

ZooKeeper : propose une service de configuration et de coordination des traitements distribués sur Hadoop

Oozie : sert à ordonner les traitements Hadoop

Mesos : optimise l'utilisation des ressources Hadoop