Ressource R5.C.07
Exploitation de la base de données
(Données massives)

IUT La Rochelle
BUT INFO3 Développeur IA

M. Sakkari

Objectifs

Ce cours a pour objectif de présenter les concepts fondamentaux de Big Data et comment ils ont changé les méthodes de gestion de données traditionnelles. Le cours présentera également divers autres aspects de Big Data comme la visualisation, afin d'offrir une vue concurrentielle de ce phénomène.

Chapitre I

Notions de base de Big data

I. Le Big Data c’est quoi ?

Big Data

- Données massives
- Mégadonnées

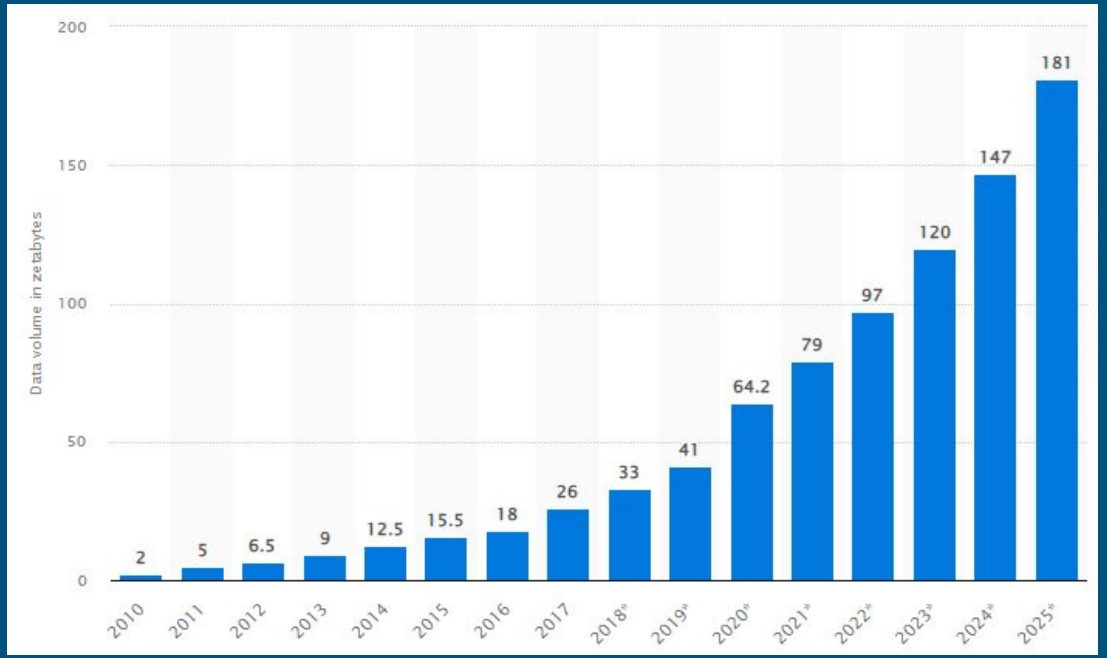
Question :

- Big Data = Le volume de données?

Réponse : **Faux**

I. Le Big Data c’est quoi ?

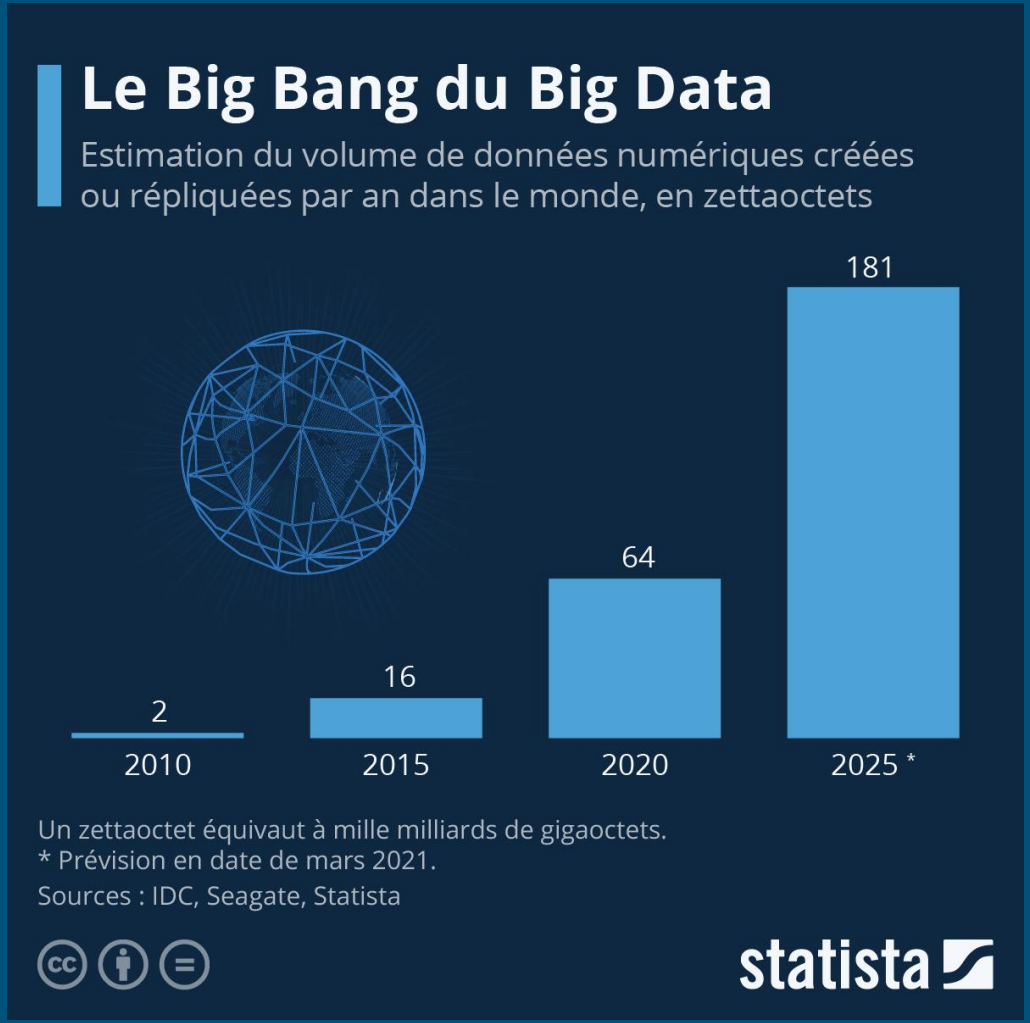
Selon ces dernières estimations, le volume de données numériques créées ou répliquées à l'échelle mondiale a été multiplié par plus de trente au cours de la dernière décennie, passant de 2 zettaoctets en 2010 à 79 zettaoctets l'année dernière (2021).



source : statista

I. Le Big Data c'est quoi ?

Comme le révèlent les prévisions, le **volume** de données générées dans le monde devrait dépasser 180 zettaoctets à l'horizon 2025, soit une **croissance** annuelle moyenne de près de 40 % sur cinq ans.



I. Le Big Data c'est quoi ?

Multiples de l'octet :
préfixes décimaux du SI et mésusages

Nom	Symbole	Valeur	Mésusage[a]
kilooctet	ko	10^3	2^{10}
mégaoctet	Mo	10^6	2^{20}
gigaoctet	Go	10^9	2^{30}
téraoctet	To	10^{12}	2^{40}
pétaoctet	Po	10^{15}	2^{50}
exaoctet	Eo	10^{18}	2^{60}
zettaoctet	Zo	10^{21}	2^{70}
yottaoctet	Yo	10^{24}	2^{80}

Multiples de l'octet :
préfixes binaires

Nom	Symbole	Valeur
kibioctet	Kio	2^{10}
mébioctet	Mio	2^{20}
gibioctet	Gio	2^{30}
tébioctet	Tio	2^{40}
pébioctet	Pio	2^{50}
exbioctet	Eio	2^{60}
zébioctet	Zio	2^{70}
yobioctet	Yio	2^{80}

I. Le Big Data c’est quoi ?

« Ce qui est étonnant, ce n’est pas que la production de données à stocker augmente, mais le rythme effréné de cette augmentation »

→ explique Jeff Fochtman, responsable marketing chez Seagate.

« Nous -mêmes sommes surpris. Et **la vague** de l’Internet des objets **ne fait que commencer** ».

→ Seagate est l'un des principaux fournisseurs d'octets dans le monde, avec en 40 ans d'histoire, plus de 3 zettaoctets à son actif.

I. Le Big Data c’est quoi ?

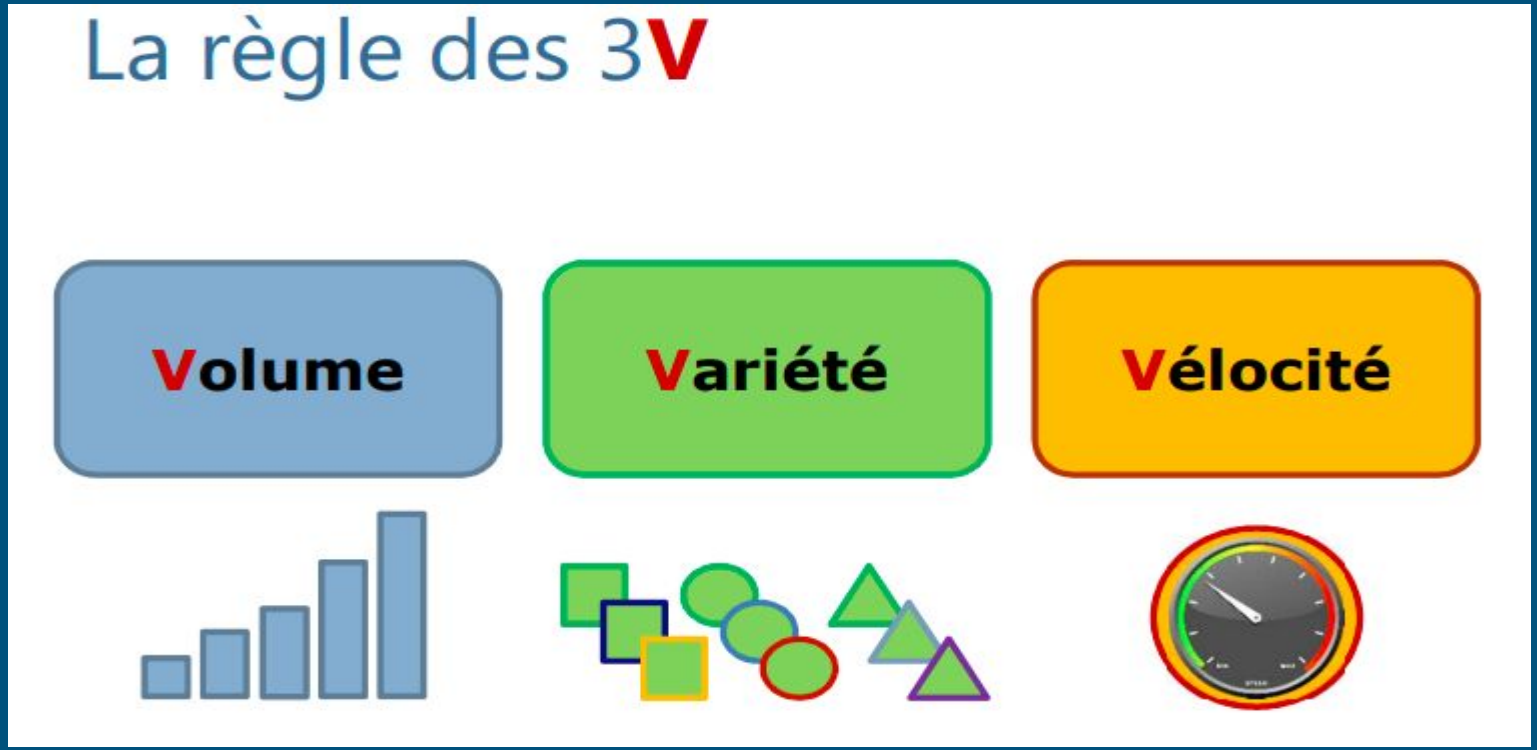
Ces données proviennent de partout :

- de capteurs utilisés pour collecter les informations climatiques,
- de messages sur les sites de médias sociaux,
- d’images numériques et de vidéos publiées en ligne,
- d’enregistrements transactionnels d’achats en ligne
- de signaux GPS de téléphones mobiles
- …

I. Le Big Data c’est quoi ?

- Le Big Data désigne **un très grand volume de données** souvent hétérogènes qui ont **plusieurs formes et formats** (texte, données de capteurs, son, vidéo, données sur le parcours, fichiers journaux, etc.), et comprenant des formats hétérogènes : **données structurées, non structurées et semi-structurées**.
- Le Big Data a une nature complexe qui nécessite des technologies puissantes et des algorithmes avancés pour son traitement et son stockage. Ainsi, il ne peut être traité en utilisant des outils tels que les SGBD traditionnels. La plupart des scientifiques et experts des données définissent le Big Data avec **le concept des 3V.**

II. Les 3V du Big Data



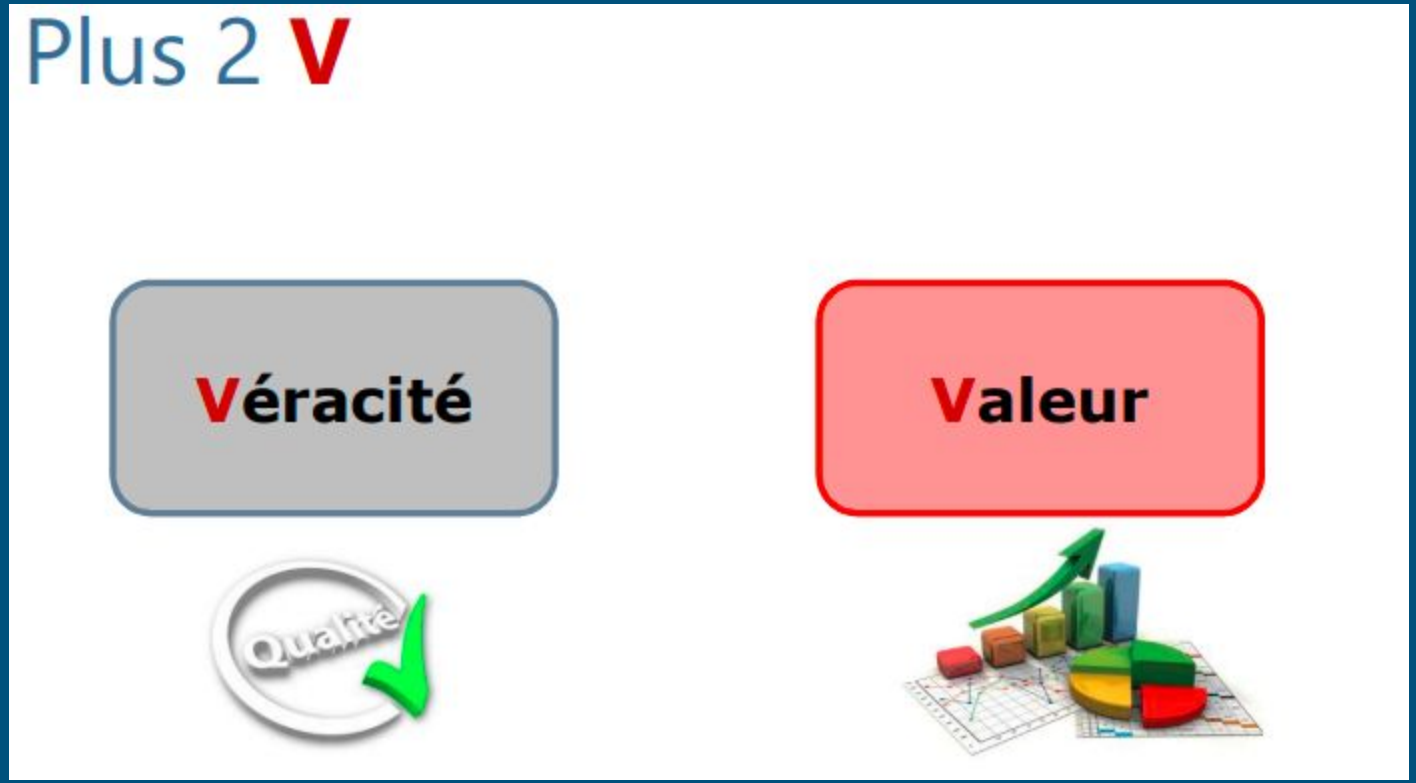
II. Les 3V du Big Data

- **Volume** : il représente l**a quantité de données générées**, stockées et exploitées.
- **Variété** : Les données volumineuses sont générées à partir de diverses sources distribuées dans **plusieurs formats** (vidéos, documents, commentaires, journaux,...). Les grands ensembles de données comprennent des données **structurées** et **non structurées,** publiques ou privées, locales ou distantes, partagées ou confidentielles, **complètes** ou **incomplètes**, etc

II. Les 3V du Big Data

- **Vélocité :** Les données sont **générées rapidement** et **doivent être traitées rapidement** pour extraire des informations utiles et des informations pertinentes. Par exemple, Wallmart (une chaîne internationale de détaillants à prix réduits) génère plus de 2,5 petabyte(PB) de données toutes les heures à partir des transactions de ses clients. YouTube est un autre bon exemple qui illustre la vitesse rapide du Big Data.

II. Les 3V du Big Data + 2V



II. Les 3V du Big Data + 2V

- **Véracité :** La véracité (ou validité) des données correspond à la fiabilité et l'exactitude des données, et la confiance que ces Big Data inspirent aux décideurs. Si les utilisateurs de ces données doutent de leur qualité ou de leur pertinence, il devient difficile d'y investir davantage.
- **Valeur** : Ce dernier V joue un rôle primordial dans les Big Data, la démarche Big Data n'a de sens que pour atteindre des objectifs stratégiques de création de valeur pour les clients et pour les entreprises dans tous les domaines.

III. Traitement à large échelle

- Maîtriser la donnée est un des grands **défis** sociétaux *(Un monde maîtrisant ses données de bout en bout, détiendrait un atout majeur pour relever les nombreux défis humains, environnementaux et économiques fixés par l'ONU pour l'horizon 2030)* avec l'objectif de disposer d'une donnée plus accessible, propre, intelligible, etc.. La donnée est considérée comme une matière première indispensable pour prendre de meilleures décisions à tous les niveaux de la société
- Le traitement des mégadonnées (à large échelle) est un ensemble de techniques ou de modèles de programmation permettant d'accéder à des données à grande échelle afin d'extraire des informations utiles pour soutenir et fournir des décisions.

III. Traitement à large échelle

- Les contraintes **5V** rendent plus difficile la gestion des méga-données
- Par ailleurs, les infrastructures informatiques ont fortement évolué avec l'essor du cloud computing (Une grappe de machines appelée cluster de calcul).
- Il est devenu possible de louer rapidement un cluster servant d'infrastructure pour gérer des méga-données. Ainsi des solutions logicielles parallèles et distribuées sont conçues pour offrir des solutions de gestion de données

III. Traitement à large échelle

La gestion de méga-données, au moyen d'une infrastructure distribuée de type cluster ou fédération de machines, soulève de nombreux défis:

1. Défi du passage à l'échelle

- Ce défi est tout d'abord posé par la contrainte de volume. Il s'agit de garantir que la solution de gestion de données continue de fonctionner lorsque la quantité devient très élevée.
- De plus, la contrainte de vélocité rend plus difficile le passage à l'échelle car l'adaptation doit être dynamique lorsque le volume des données fluctue.

III. Traitement à large échelle

→ Cela nécessite de concevoir une solution dite élastique, il s'agit de concevoir des algorithmes pour contrôler la distribution des données et des traitements.

- Distribuer les données sur plusieurs machines et adapter cette distribution en fonction du volume des données et du nombre de machines, et de la capacité de stockage des machines.
- Distribuer les traitements entre les machines. Pour faciliter le passage à l' échelle, les algorithmes doivent être décentralisés ce qui permet à chaque machine de traiter des données indépendamment. Cela nécessite également de coordonner les traitements des différentes machines pour être capable de traiter des requêtes plus complexes car accédant aux données de plusieurs machines.

III. Traitement à large échelle

2. Défi de la représentation des données

- Ce défi est posé par les contraintes de diversité et de complexité des données.
- Les données peuvent provenir de contextes très divers.
- Il n’y a pas de formalisme général pour « injecter » automatiquement les informations du contexte dans un système de gestion de données.
- Le défi est de définir ou compléter la représentation des données avec des informations spécifiques issues du contexte des données