# Big data

## C’est quoi le big data ?

### Big data : constat

• Explosion des capacités de stockage et de calculs des ordinateurs

• Augmentation exponentielle du nombre de données générées et stockées

• Augmentation exponentielle du nombre de données

• Développement des technologies de virtualisation et du cloud computing

### Big data : loi de Moore

• La loi de Moore stipule que les ordinateurs deviennent au fil du temps plus petits, plus rapides, moins chers, à mesure que les transistors sur circuits intégrés deviennent plus efficaces.

• Aujourd’hui, on est arrivé à une limite physique. On n’arrive plus à réduire les transistors.

### Big data : volume de données

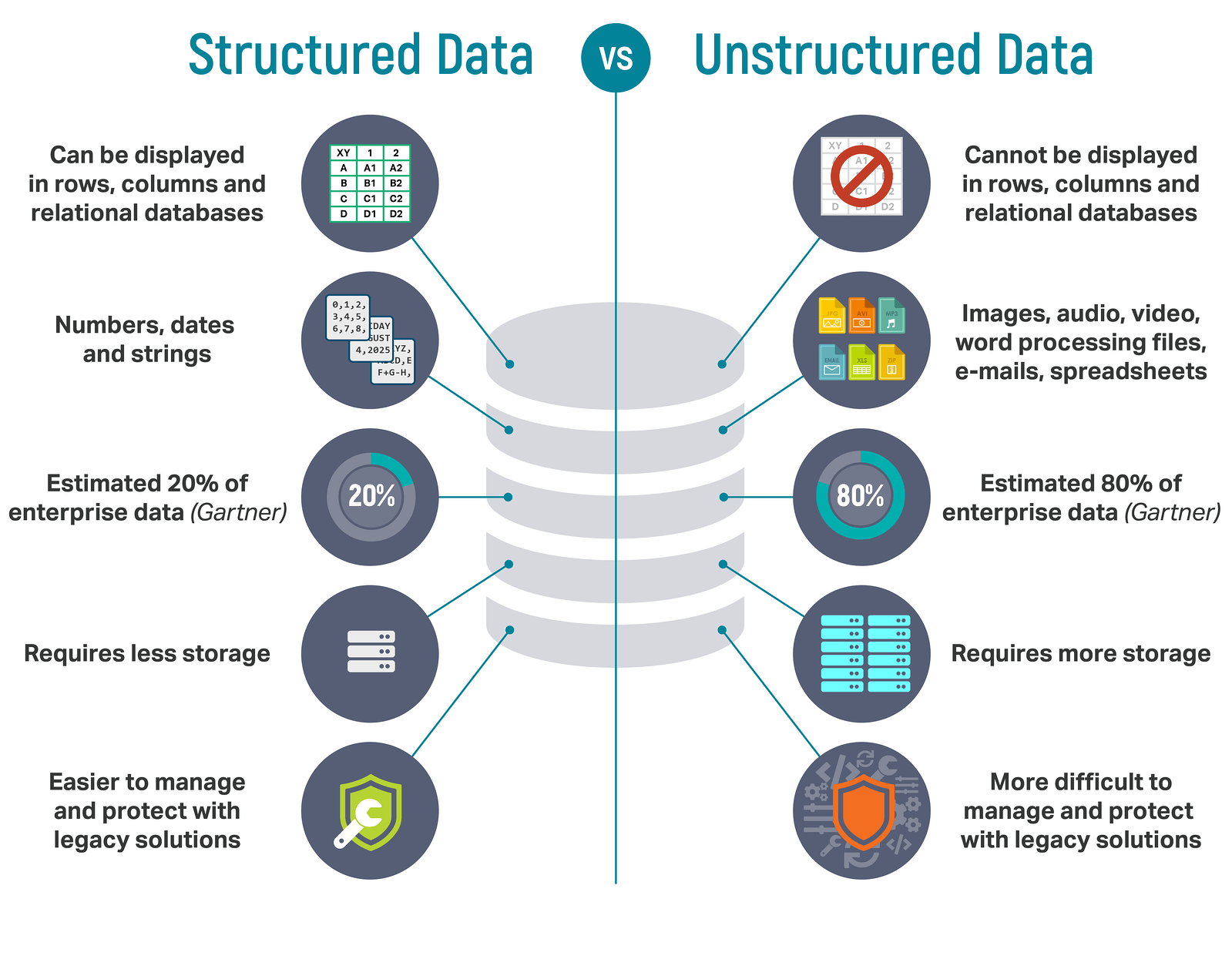
• Génération sans cesse croissante de données

• En 2010, le monde entier avait généré 2 zettaoctets, en 2015 on a multiplié par 8 ce chiffre et en 2020, on est à 181 zettaoctets.

• Rappel : 1 Zo = 1000 000 000 000 Go = 1021 octets

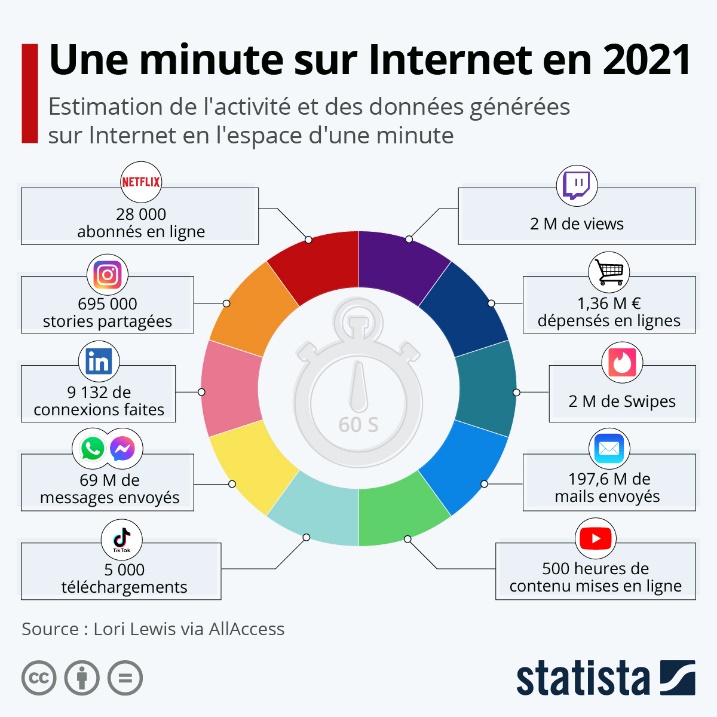
• Jusqu’à 2010, 80% de la data générée dans le monde était à usage humain. En 2020, on est à 18% et ça continue de descendre. Les données qui ne sont pas à usage humain sont pour les machines qui communiquent avec les autres machines.

### Données structurées vs non structurées



• Utilisation de données semi ou non-structurées de plus en plus importante.

### Big data : données connectées



### Limitation des systèmes classiques

• Données trop volumineuses pour être stockées sur un seul serveur

• Données trop variées pour être stockées facilement

• Données changeantes trop rapidement pour être stockées/traitées facilement

• Données trop complexes pour être traitées facilement

### Le big data

• **Big data :** ensemble de technologies et de méthodes permettant de stocker, traiter et analyser des données massives, variées et changeantes.

Objectif : permettre un traitement efficace de données massives, variées et changeantes **à un coup financier, humain et temporel raisonnable**.

### Big data : historique

• Concept ancien issu des années 70 avec les premiers datacenters

• Le terme serait apparu en 1997.

### Big data : tout est information

• Principe que l’homme et la totalité du monde qui l’entoure peuvent être représentés comme « des ensembles informationnels, dont la seule différence avec la machine est leur niveau de complexité. »

• « La vie deviendrait alors une suite de 0 et de 1, programmable et prédictible.

### Big data : mapreduce et NoSQL

• Concrétisation dans les années 2005 par Google qui déploie un algorithme sur des opérations massives : le MapReduce qui deviendra le projet Hadoop.

• Dans les années 2009, le déploiement de stockage open-source, distribués et non-relationnels : le NoSQL

### Big data : concepts

• Le big data repose sur le concept de **parallélisme** et de **distribution** des algorithmes et de traitement et du stockage des données.

• Diviser pour mieux régner, c’est-à-dire répartir le stockage ou le traitement des données massives sur plusieurs machines.

### Big data : domaines d’application

• L’utilisation du Big Data est très large et touche de nombreux domaines :

— **Commerce :** analyse des données clients, prévision, recommandation, etc

— **Finance :** analyse des données financières, prévision, etc.

— **Santé :** analyse des données médicales, prévision, etc.

— **Energie :** analyse des données énergétiques, prévision, etc.

— **Transport :** analyse des données de transport, prévision, etc.

— **Agriculture :** analyse des données agricoles, prévision, etc.

### Exemple : en santé

• Recherche sur l’effet d’un médicament dans une maladie

• Collecter des données à partir de milliers/millions de patients

• Sources multiples et variées (dossiers médicaux électroniques, des bases de données d’assurance maladie, des registres et des enquêtes de santé publique).

• Analyses des données

• Identifier tendances, corrélations avec les modèles, efficacité du médicament, facteurs de risque, traitements associés, et les impacts sur la qualité de vie

• Aide à la décision, adaptabilité des traitements, prévision des effets secondaires ou de l’évolution de la maladie

### Big data : domaine technique

• Croisement de nombreuses spécialités techniques :

— Informatique transactionnelle (principe ACID, etc.)

— Informatique décisionnelle (BI, prise de décision, etc.)

— Informatique en tems réel (temps de réponse critique)

— Stockage et tri de données (besoin volume, rapidité, etc.)

— Traitement et analyse de données (catégorisations, synthèse, prédictions, représentations, etc.)

## Caractéristiques du big data

### Les 3 V

• Le Big Data repose sur le principe des 3 V :

— Volume

— Vitesse

— Variété

### Volume

• Volume : masse croissante de données

• Nécessite des besoins spécifiques pour :

— stocker les données

— transporter et échanger les données

— traiter et analyser les données

• Volume : producteurs

— Entreprises commerciales (Google, Amazon, Facebook, Twitter, etc.

— Analyses de données de santé (les données médicales)

— Analyses météorologiques/spatiales

### Vitesse

• Vitesse : rapidité avec laquelle les données sont générées, traitées ou modifiées

— Répond aux besoins des processus chronosensibles (bourses, stream, etc.)

— Risque pour l’Homme de perdre le contrôle sur les données (traitement à la nanoseconde)

### Variété

• Variété : diversité des types de données

— Données structurées (tableaux, fichiers, etc.) : 20%

— Données semi ou non structurées (XML, JSON, textes, images, etc.) : 80%

Variété : données structurées : données qui a une structure prédéfinie (tableau, fichier, etc.)

— Ce sont toutes les données observables avec un schéma de typage bien définis et fixe

— Généralement, elles sont observables en tableau ou base de données relationnelles traditionnelles (adresses, ventes d’articles, etc.)

Variété : données non-structurées : données qui sont qualitatives et non quantitatives

— toutes les données qui ne sont pas observables avec un schéma de typage bien définis et fixe

— sont une conglomération de nombreuses données de différents types qui sont stockées dans leurs formats en mode natif (vidéo, audio, etc.)

Variété : données semi-structurées : données intermédiaires entre les données structurées et non-structurées

— données qui ont une structure mais qui ne sont pas fixes ou qui n’ont pas de structure prédéfinie mais possède des métadonnées typées (xml, JSON, etc.)

### Big data : extension des 3V

Aujourd’hui, les principes des 3V se sont largement étendus. Par extension, on y trouve d’autres grands V comme : volatilité, valeur, vulnérabilité, véracité, validité, visibilité

### Volatilité

Notion de durée de vie d’une donnée

* L’estimation de la durée de vie d’une donnée
* De sa prise en charge et son traitement
* De son obsolescence sont à prendre en compte

### Valeur

• Notion de profit à tirer d’une donnée

* Extraire une valeur d’une masse d’information est une problématique majeure
* Notions de regroupement, filtres, classification, hiérarchisation, etc.

### Vulnérabilité

• Notion de sécurité des données

— Un gros volume de données est une cible de choix pour les pirates informatiques

— Induit la nécessité d’une structure sécurisée adaptée

### Véracité

• Notion de fiabilité et confiance des données

— Nécessite des outils de vérification et de validation des données (recoupement, enrichissement, etc.)

### Validité

### Visibilité

### A

## A

A

• A

—

## A

A

•

—

7V du big data