

Les notions de base de Méga Données (Big Data)

- L'expression « big data » est apparue en octobre 1997.
- Méga données est ajouté au dictionnaire français en 2014 pour la traduction de Big Data

Sources d'émergence de données

1 Evolution de la technologie



Sources d'émergence de données

1 Evolution de la technologie

2 Internet des objets



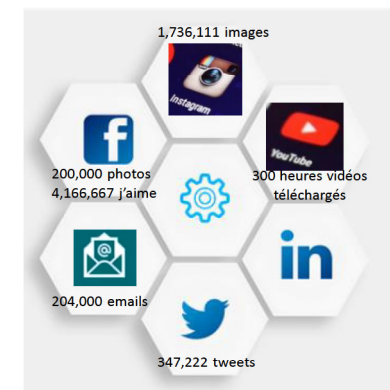
50 billions outils d'IOT en 2020

Sources d'émergence de données

1 Evolution de la technologie

2 Internet des objets

3 Les médias sociaux



Sources d'émergence de données

- 1 Evolution de la technologie
- 2 Internet des objets
- 3 Les médias sociaux
- 4 Autres sources



5

Définition des mégas données

C'est une collection des ensembles de données larges et complexes qu'il est devenu impossible de traiter utilisant les outils de systèmes de bases de données existants ou bien les applications traditionnelles de traitement de données



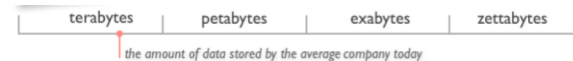
6

Caractéristiques des mégas données

Les mégas données sont caractérisées par leur 3 V en gestion de données: volume, variété et vélocité qui sont tous élevés. D'autres caractéristiques peuvent être considérées tels que la valeur et la véracité.

7

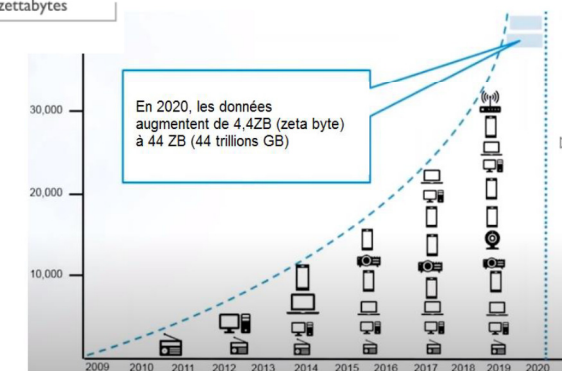
Caractéristiques des mégas données



[One petabyte = 1,024 terabytes]
One terabyte = 1,024 Giga Byte


- 1 **Volume:** le volume de données augmente exponentiellement au fil du temps

- Augmentation de données *44
*4 de 2009 à 2020



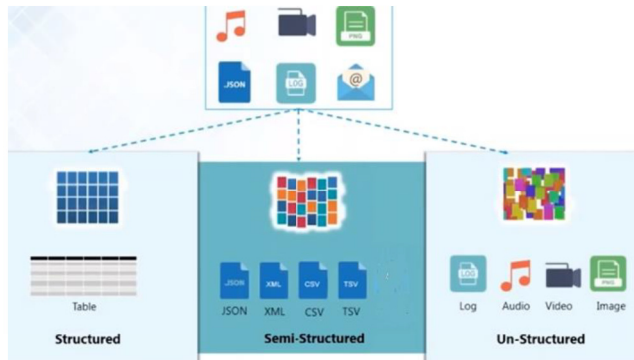
8

Caractéristiques des mégas données

- 2  **Variété:** différentes formes, bases de données traditionnelles, images, documents et dossiers complexes. Données structurées, semi-structurées et non structurées.


- Une seule application peut générer **plusieurs formats**

* **Données hétérogènes** => **problème d'intégration** de données complexes



9


Caractéristiques des mégas données

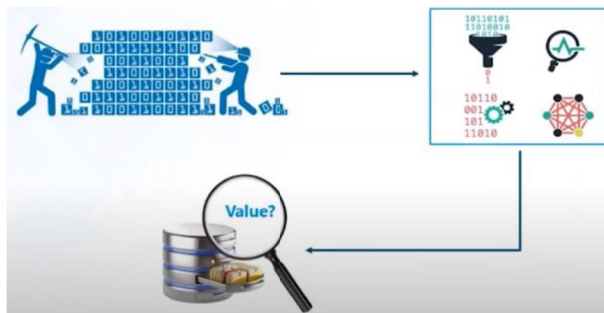
- 3  **Vélocité** (vitesse): La vitesse change constamment, à partir de données diffusées en provenance de plusieurs sources



10


Caractéristiques des mégas données

- 5  **valeur.** Les technologies de stockage et d'analyse des Big Data n'ont de sens que si elles apportent de la **valeur ajoutée**. Exploiter les données, c'est avant tout **répondre aux objectifs** d'utilisation des Big Data.



11

Caractéristiques des mégas données

- 4  **véracité. Qualité des données.** c'est l'un des enjeux majeurs de l'exploitation des Big Data. Il est nécessaire de multiplier les **précautions** pour minimiser les biais liés au **manque de fiabilité** du Big Data. Par exemple, les faux profils sur les réseaux sociaux, les fautes d'orthographe, les fraudes ...



Trouver les incertitudes et les incohérences dans les données

Reliability
Accuracy
Timeliness
Completeness
Consistency
Relevance

12

Types des mégas données

- des données structurées issues notamment de bases de données relationnelles (lignes et colonnes),
- des données semi-structurées (fichiers CSV, journaux, XML, JSON...),
- des données non structurées (emails, documents, PDF), des fichiers de type blob -binary large objects (images, audio, vidéo).

13

Chaîne de valeurs de mégas données



Génération

Enregistrement passif:

- Données structurées
- Opérations commerciales bancaires
- Enregistrements d'achat
- archives

génération active:

- Données semi-structurées ou structurées
- Contenu généré des utilisateurs, eg. réseaux sociaux

Production automatique

- Données de connaissance d'emplacement
- Données mobiles
- appareils compatibles Internet basés sur des capteurs

14

Chaîne de valeurs de mégas données



Acquisition

Collection

- Push based (poussé), vidéo surveillance
- Pull based (traction), eg, web crawler (robot d'exploration du web)

Transmission

- Transférer les données à un centre de données

Prétraitement

- Intégration
- Élimination de redondance
- Nettoyage

15

Chaîne de valeurs de mégas données



Storage (stockage)

Infrastructure

- Technologie de stockage (eg., HDD, SDD)
- Architecture réseau (eg., DAS, NAS, SAN)

Gestion de données

- Fichier système distribué (HDFS)
- Stockage No SQL

16

Chaîne de valeurs de mégas données



Analysis (analyse)

Objectives

- Analyse descriptive
- Architecture réseau (eg., DAS, NAS, SAN)

Méthodes

- Analyse statistique
- Clustering (groupement)
- Classification
- Modèles de programmation
 - Map reduce
 - Traitement de flux

17

Les défis de mégas données

Technologie et infrastructure

- Nouvelles architectures,
- modèles de programmation

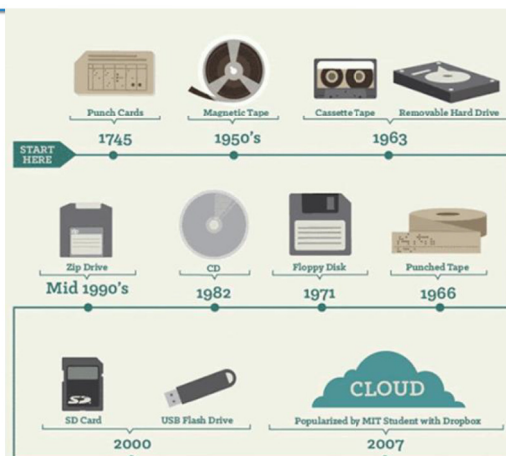
Gestion et analyse de données

- Nouvel accent sur les données
=> Data Science

18

Stockage des mégas données

Bref historique du stockage des données



19

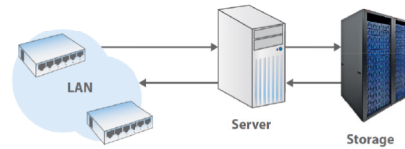
Stockage des mégas données

- Le stockage de données se fait sous un cluster (grappe-groupe) de machines
- Dans un cloud computing accessible via l'Internet dont le matériel de stockage est très varié.
 - o Stockage interne: média attaché à l'intérieur du serveur.
 - o Stockage externe: support connecté aux ports d'interface d'un serveur, à l'aide de canal en fibre Channel, USB, etc.

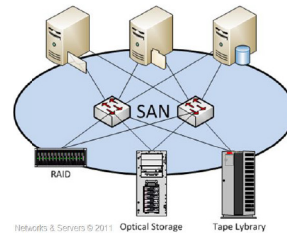
20

Stockage des mégas données

- **Stockage directement attaché** (Direct Attached Storage-DAS): c'est un stockage connecté à un serveur et accessible aux autres ordinateurs via l'accès préalable au serveur.



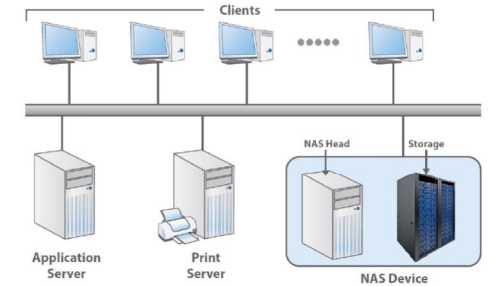
- **Réseau de stockage** (Storage Area Network - SAN) : SAN est un réseau haut débit de périphériques de stockage partagés. Les serveurs connectés à un SAN peuvent accéder à tous les périphériques de stockage connectés au SAN.



21

Stockage des mégas données

- **Stockage attaché à un réseau** (Network attached storage -NAS): peut être défini comme un périphérique de stockage de données au niveau des fichiers qui fournit un accès aux fichiers sur un réseau à des clients hétérogènes.



22