Chap I: Le Big-Data

Tables des matières

1. Le Big-Data
2. Les outils de Stockages de données
3. L'avènement du Cloud
4. Les notions de Streaming et de Batch
5. Les différentes Data-Plateformes
6. L'environnement Hadoop

Tour de table

About Me

Data-Engineer, 3 ans d’expériences avec les techno Big-Data et Cloud.

Introduction

“ There was 5 exabytes of information created between the dawn of civilization through 2003, but that much informations now created every 2 days, and the pace is increasing."

Eric Schmidt, Former PDG Google, 2010

Focus sur la quantité de données produites par les réseaux sociaux à chaque minutes de la journée

527,760 photos

120 nouveaux utilisateurs

4,146,600 vidéos regardées

456,000 tweets

46,740 nouvelles photos

Définition

Le Big Data (ou mégadonnées) représente les collections de données caractérisées par un **volume**, une **vélocité** et une **variété** si grands que leur transformation en valeur utilisable requiert l'utilisation de **technologies et de méthodes analytiques spécifiques**.

Caractéristiques du big-data

Volume	Vélocité	Variété	Véracité	Variabilité	Valeur
Quantité importante de données issues de multiples sources.	La vitesse à laquelle la donnée est créée et consommée.	Différentes types de données (Structurées, non-structurées, semi-structurées)	A quelle mesure la donnée est fiable	Comportement de la donnée au cours du temps (anomalies,).	La valeur associée à l'information extraite de la donnée

Pourquoi le Big-Data

Pourquoi le Big-Data

- Augmentation exponentielle de la quantité de données (non structurées notamment: chat, blog, web, musique, photo, vidéo, etc)
- L 'augmentation de la capacité de stockage à un coût relativement bas
- Nouvelles techniques d'analyse/d’exploitation de la donnée (Visualisation, Data-Science, etc)
- De nouvelles technologies plus adaptées (Hadoop)
- L'avènement du CLoud

Les Outils de stockages

Les différentes types de bases de données

- Les bases de données relationnelles (Sql)
- Les bases de données Nosql
- Le Stockage Objet

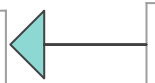
Les bases de données relationnelles

Définition

Une base de données relationnelle est un ensemble d'éléments de données dotés de relations prédéfinies entre eux. Ces éléments sont organisés en un jeu de tables composées de colonnes et de lignes.

students

id_stud	name	dept
1	Abdou	2
2	Jeanne	1
3	Ibou	3



id_dept	name
1	Maths
2	Infos
3	Bio

départements

Quelques notions clés

Attribut: Un attribut est un identificateur (un nom) décrivant une information stockée dans une base.
Schéma de relation: Un schéma de relation précise le nom de la relation ainsi que la liste des attributs avec leurs domaines.
Clé primaire: La clé primaire d'une relation est un attribut unique qui identifie de manière unique un élément d'une table.
Clé étrangère: Une clé étrangère dans une relation est formée d'un ou plusieurs attributs qui constituent une clé primaire dans une autre relation.
Transaction: Une transaction de base de données désigne une ou plusieurs instructions SQL exécutées en tant que séquence d'opérations (requête).

id_stud	name	dept
1	Abdou	2
2	Jeanne	1
3	Ibou	3

Propriété ACID

- **Atomicité:** La transaction est exécutée dans son intégralité ou invalidée dans son intégralité.
- **Cohérence:** requiert que les données écrites dans la base de données dans le cadre d'une transaction soient conformes à toutes les règles et restrictions définies.
- **Isolation:** L'indépendance des différentes transactions en cas de concurrence.
- **Durabilité:** requiert que toutes les modifications réalisées sur la base de données soient permanentes une fois la transaction exécutée avec succès

Avantages et Inconvénients

- **Les ++**
  - Un modèle de données simple
  - Préservation de l’intégrité des données
  - Un système de requêtage assez mature et facile à prendre en main

- **Les - -**
  - Un schéma parfois très rigide moins pratique pour des données qui varient constamment
  - Scalabilité horizontale parfois difficile à mettre en place
  - Souvent moins performant que le modèle non-relationnel

Cas d’usage

- Données structurées avec de fortes contraintes sur le modèle de données
- Pour des transactions complexes et/ou fréquentes sur les données
- Quand on a des relations entre les différentes entités présentes.

Exemples de systèmes de bases de données relationnelles

Le modèle non relationnel, le NoSql

Le NoSQL est un ensemble de technologies de base de données qui **repose sur un modèle différent de celui des BDD relationnelles**. Il se passe de l'exigence d’avoir un schéma prédéfini offrant ainsi plus de flexibilité avec la possibilité de stocker des données non-structurées, semi-structurées et structurées.

Il existe 4 grandes familles de base de données NoSql: Les bases de données orientées **clé-valeur,** les bases de données **document**, les bases de données orientées **colonne** et les bases de données type **graphe**.

Le modèle clé-valeur

Clé	Valeur
1	https://adresseweb.com
2	356
3	**mail**: monmail@gmail.com **date**: 25/10/2020 13:42:12

Exemples:

-  DynamoDB
  amazon web services
- redis

Le modèle document (basé sur le json et le xml)

Document (id: 5baf47)

```
{
 "nom": "Liquide vaisselle",
 "images": [
   "https://...",
   "https://..."
 ],
 "specs": {
   "parfum": "Orange"
 }
}
```

Document (id: ea53aa)

```
{
 "nom": "Shampooing",
 "images": [
   "https://...",
   "https://..."
 ],
 "specs": {
   "parfum": "Vanille"
 }
}
```

Document (id: d710bb)

```
{
 "nom": "Fromage blanc",
 "images": [
   "https://...",
   "https://..."
 ],
 "specs": {
   "mat_grasses": "0%"
 }
}
```

Exemples:

- 

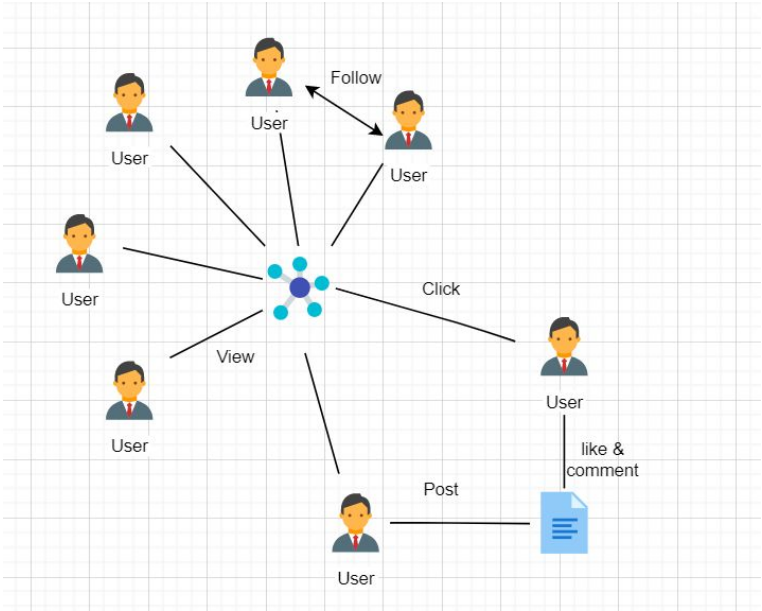
- 

Le modèle orienté colonne

Exemples:

- APACHE HBASE
- cassandra

Le modèle Graphe



Exemple:

- neo4j

Le Stockage Objet (un modèle souvent basé sur le cloud)

Le **stockage Objet** est un format de stockage dans lequel les données sont stockées sous forme d'unités discrètes appelées objets. Chaque unité a un identifiant unique ou une clé qui permet de la retrouver où qu'elle soit stockée dans un système distribué.

Les avantages

- **Évolutivité :** une architecture simple

- **Données à la demande :** avec le stockage d'objets, il est plus facile de payer seulement pour la capacité de stockage effectivement utilisée.

- **Supporte multiples Api :** vous pouvez accéder et gérer les données dans des systèmes de stockage d'objets par différentes manières

- **Meilleure intégrité des données :** grâce au codage à effacement, les systèmes de stockage d'objets peuvent protéger l'intégrité des données en reconstruisant des morceaux de vos données et en effectuant des contrôles d'intégrité pour prévenir la corruption.
- **Disponibilité des données:** Profitant du cloud, ces systèmes offrent une haute disponibilité de la donnée.
- **Scalabilité** (jusqu'à plusieurs Teras de données)

cas d’usage

- En tant que **Datalake**
- Pour les gros volumes de données non-structurées (Photos, vidéos, fichiers etc)
- Machine-Learning, Data-Science
- Backup and archivage

Exemples

Azure Data Lake Storage Gen2

L’avènement du cloud