Architecture et Technologies du Big Data

2.1 Introduction aux Architectures Big Data

Les architectures Big Data sont conçues pour gérer de grandes quantités de données en utilisant des systèmes distribués. Contrairement aux bases de données traditionnelles qui fonctionnent sur un seul serveur, les architectures Big Data se basent sur des clusters de serveurs pour le traitement parallèle des données.

2.2 Modèles de Traitement

- **Batch Processing** : Traitement des données par lots, idéal pour des analyses périodiques sur de grandes quantités de données.
- **Stream Processing** : Traitement des données en temps réel, utilisé pour analyser des données en continu, comme celles provenant des réseaux sociaux ou des capteurs IoT.

2.3 Technologies Clés

- **Hadoop** :
  - **HDFS (Hadoop Distributed File System)** : Un système de fichiers distribué qui permet de stocker des données sur plusieurs nœuds d'un cluster.
  - **MapReduce** : Un modèle de programmation qui permet de traiter de grandes quantités de données de manière parallèle et distribuée.
- **Apache Spark** :
  - Un moteur de traitement de données rapide et en mémoire, plus rapide que Hadoop MapReduce dans de nombreux cas. Il permet de traiter les données en batch ou en temps réel (streaming).
  - **RDD (Resilient Distributed Dataset)** : Une abstraction de données immuables qui permet de manipuler des données distribuées avec des opérations parallèles.

2.4 Comparaison entre Hadoop et Spark

- **Hadoop** : Bien adapté au traitement par lots et aux grandes quantités de données, mais plus lent en raison du stockage sur disque.
- **Spark** : Traite les données plus rapidement en raison du stockage en mémoire et est plus flexible, notamment pour les traitements en temps réel.