Livrables - Partie 1 : Conception du Data Lake

1. Document de conception détaillé de l'architecture du Data Lake

L'architecture du Data Lake suit le modèle en trois couches Bronze, Silver, Gold dans un environnement local. Les données sont ingérées, transformées, et agrégées à travers ces trois couches pour une analyse efficace et stratégique.

1. **Bronze Layer (Raw Zone) :**
    - **Description :**
      Cette couche contient les données brutes collectées directement à partir des différentes sources sans aucune transformation ou nettoyage.
    - **Technologies utilisées :**
        - **Système local de stockage pour les données brutes. Cela pourrait inclure des fichiers locaux dans des formats comme CSV, JSON, et des fichiers texte.**
        - **Kafka pour les flux de données en temps réel (par exemple, les publicités en ligne).**
    - **Cas d'utilisation :**
        - **Stockage des données provenant de sources diverses : fichiers CSV (transactions), logs (texte), JSON (médias sociaux), et flux Kafka (publicités).**
2. **Silver Layer (Processed Zone) :**
    - **Description :**
      Cette couche contient les données nettoyées et transformées. Les données sont préparées pour une analyse plus approfondie : gestion des valeurs manquantes,

ajustement des types de données et filtrage des données incorrectes.

- **Technologies utilisées :**
  - **Apache Spark pour le traitement et la transformation des données.**
  - **Système local de stockage pour les données traitées et transformées.**
- **Cas d'utilisation :**
  - **Transformation des données brutes dans la couche Bronze pour les rendre prêtes à l'analyse.**

3. **Gold Layer (Aggregated Data) :**
   - **Description :**
     Cette couche contient les données agrégées et prêtes pour l'analyse. Elle inclut les résultats de calculs comme les métriques ou les agrégations effectuées à partir des données de la couche Silver.
   - **Technologies utilisées :**
     - **Apache Spark pour effectuer les calculs des métriques et agrégations.**
     - **Système local de stockage pour les résultats finaux et prêts pour l'analyse.**
   - **Cas d'utilisation :**
     - **Calcul de métriques stratégiques (ex. : ventes totales, nombre de clics publicitaires) sur les données nettoyées et transformées.**

---

2. Documentation technique sur les choix de technologies et de solutions

La conception du Data Lake repose sur des technologies open-source et adaptées à un environnement local :

1. **Stockage local des données (pas de Cloud)**
    - **Usage :**
      Stockage des données à chaque étape (Bronze, Silver, Gold) sur des systèmes de fichiers locaux (par exemple, des fichiers sur disque dur ou SSD).
    - **Pourquoi stockage local ?**
        - **Nous avons choisi un stockage local, car l'infrastructure Cloud n'est pas utilisée dans ce projet. Les données sont stockées directement dans le système de fichiers local, ce qui est suffisant pour un projet de taille gérable.**
2. **Apache Spark**
    - **Usage :**
      Traitement des données en batch et en streaming, notamment pour le nettoyage, la transformation, et le calcul des métriques.
    - **Pourquoi Apache Spark ? :**
        - **Spark est rapide et distribué, parfaitement adapté pour traiter de grandes quantités de données en local.**
        - **Il permet un traitement en parallèle, essentiel pour le nettoyage et l'agrégation des données de manière efficace.**
3. **Apache Kafka & Zookeeper**

- **Usage :**
  Kafka gère les flux de données en temps réel pour des cas comme les publicités en ligne.
- **Pourquoi Kafka ? :**
  - **Kafka est idéal pour gérer des flux en temps réel avec une faible latence et une grande scalabilité, même en environnement local.**
  - **Zookeeper est utilisé pour coordonner les brokers Kafka et assurer la gestion du cluster Kafka.**

---

3. Flux de Données

Le flux de données suit un chemin bien défini, avec une séparation claire entre les trois couches de l'architecture médaillon :

1. **Sources de données :**
   - **CSV (Transactions clients), JSON (Médias sociaux), Texte (Logs des serveurs), Kafka (Publicités).**
2. **Raw Zone (Bronze) :**
   - **Les données brutes sont ingérées directement dans cette zone de stockage local sans transformation.**
3. **Traitement ETL (Apache Spark) :**
   - **Les données de la Raw Zone sont traitées pour nettoyer, enrichir et transformer les informations dans la couche Silver.**
4. **Processed Zone (Silver) :**
   - **Les données transformées sont stockées ici pour être analysées plus avant.**

5. **Kafka Producer :**
    - **Envoi des données publicitaires en temps réel vers Kafka, en continu.**
6. **Kafka Consumer :**
    - **Consommation et traitement des données en temps réel via Apache Spark Streaming, avec analyse instantanée.**