Conception, Développement, et Exploitation d'un Data Lake pour une Entreprise Digitale

Scénario :

Une entreprise de commerce en ligne souhaite exploiter au maximum ses données en centralisant toutes ses sources de données dans un **Data Lake**. Elle reçoit des données provenant de diverses sources :

- Transactions clients (base de données relationnelle). SQLITE
- Logs des serveurs web (données non structurées). Fichier texte
- Données des médias sociaux (semi-structurées, format JSON).
- Flux de données en temps réel des campagnes de publicité en ligne. Script or Kafka

Les étudiants doivent concevoir un Data Lake qui permet de collecter, transformer, analyser et gouverner ces données pour aider l'entreprise à prendre de meilleures décisions commerciales.

Partie 1 : Conception du Data Lake

Analyse et Conception

- **Description :**
  Faire l'analyse des besoins en données de l'entreprise et faites-en la conception logique et physique d'un data lake.

- **Livrables :**
  - Un document de conception détaillé décrivant l'architecture du Data Lake.
  - Un diagramme des flux de données (comment les données sont collectées, stockées et traitées).
  - Choix des technologies pour le stockage (HDFS, Amazon S3 ou file system si le travail se fait sur une machine locale). blabla

Création de l'infrastructure

- **Description :**
  - Implémentation de l'infrastructure du Data Lake sur une une machine local ou sur une plateforme Cloud (AWS, Azure, ou Google Cloud).
  - Utilisation de clusters Hadoop ou de solutions similaires pour le stockage.
  - Utilisation de Spark/Kafka/Kafka stream pour le traitement distribué ou une solution similaire.
- **Livrables :**
  - Scripts pour déployer l'infrastructure. shell fill db & query jsons etc
  - Documentation technique sur les choix de technologies et de solutions Cloud.

Partie 2 : Ingestion et Transformation des Données

Ingestion de Données Brutes

- **Description :**
  - Mise en place de l'ingestion des données provenant de différentes sources : bases de données SQL, logs, flux de données en temps réel (via Apache Kafka, AWS Kinesis).
  - Mise en place d'un pipeline d'ingestion pour des données structurées, semi-structurées et non structurées.
  - Pour chaque type d données, générez un exemple de jet de données pour illustrer l'exemple
- **Livrables :**
  - Pipeline d'ingestion en temps réel ou batch pour les différentes sources de données.
  - Exemple de datasets ingérés dans le Data Lake.

Pipelines de Transformation des Données

- **Description :**
  - Conception et développement de pipelines ETL (ou ELT) pour transformer et enrichir les données dans le Data Lake.
  - Utilisation d'Apache Spark ou une technologie similaire pour effectuer ces transformations.
- **Livrables :**
  - Code Spark pour nettoyer et transformer les données, ou avec une autre technologies.
  - Documentation sur les différentes transformations appliquées aux données.

Spark submit shell

Partie 3 : Analyse et Exploitation des Données

- **Description :**
  - Implémentation d'une API pour exposer les données du data lake.
  - L'API devra exécuter des repûtes pour extraire des données
  - Création de rapports et de tableaux de bord sur les données collectées.
  - Bonus: Utilisation d'outils BI (comme Tableau ou Power BI) pour visualiser les résultats.
- **Livrables :**
  - Code API
  - Documentation des endpoints implémentés et les résultats qu'ils renvoient

Expose raw zone

Partie 4 : Sécurité, Gouvernance et Qualité des Données

- **Description :**
  - Mise en place de la gouvernance des données : catalogage, suivi des métadonnées, versioning des datasets.
  - Sécurisation des accès aux données via des solutions comme AWS IAM, Azure Active Directory, ou Kerberos pour Hadoop.
- **Livrables :**
  - Document décrivant la Politique de sécurité des données, les politiques d'accès

Blabla

Gestion de la Qualité des Données

- **Description :**
  - Mise en place de solutions pour assurer la qualité des données ingérées et transformées : détection d'anomalies, gestion des valeurs manquantes.
  - Surveillance des pipelines de données pour garantir leur fiabilité.
- **Livrables :**
  - Scripts ou outils pour la vérification de la qualité des données.
  - Rapport de qualité des données avec des indicateurs clés (anomalies détectées, données manquantes, etc.).