Projet de Groupe Big Data

Conception d’un pipeline de données avec architecture Lambda

Objectif

Étudier et concevoir un pipeline de données complet en utilisant l’architecture Lambda, intégrant ingestion en temps réel, traitement batch et visualisation des données. Le projet doit inclure la mise en place d’un environnement isolé pour tous les services et fournir une documentation détaillée.

Description du projet

1. **Architecture du pipeline**
    - Choisir un modèle Lambda adapté.
    - Identifier les couches : **batch** et **streaming**, ainsi que les **sources de données** et **destinations**.
    - Vous pouvez sélectionner librement un domaine (IoT, médias sociaux, logs Web, météo, etc.) et utiliser une API publique de ce domaine pour simuler la collecte de données.
    - Une liste d’APIs publiques est disponible ici : Public APIs.
2. **Frontends**
    - **Frontend 1 : Dashboard temps réel**
        - Visualiser les métriques du pipeline en temps réel.
        - Possibilité d’utiliser un outil BI (Grafana, Superset, Kibana).
        - Offrir des boutons de contrôle pour interagir avec le pipeline.
    - **Frontend 2 : Panneau de filtrage de données**
        - Permettre de sélectionner et d’exécuter plusieurs requêtes sur vos différents endpoints.
        - Stocker les résultats dans HDFS local.
        - Alerter l’utilisateur lorsque les données sont prêtes.

    - Permettre d’exécuter des requêtes supplémentaires (groupement, filtrage, agrégation) sur les données HDFS.
    - Assurer que tous les résultats soient triables et que les données soient correctement standardisées.
3. **Isolation des services**
    - Utiliser **Docker / Docker Compose** pour isoler tous les services : Kafka, Zookeeper, Spark, Airflow, base de données, etc.
4. **Backend (optionnel)**
    - Peut gérer l’ingestion et la publication vers Kafka ou déclencher des DAGs Airflow.
5. **Documentation et rapport**
    - Fournir un document détaillant :
        - L’architecture mise en place.
        - Le code développé (DAGs Airflow, scripts Python/Spark).
        - L’enchaînement des tâches et l’implémentation du modèle Lambda.

Étapes à suivre

1. **Mise en place de l’environnement**
    - Installation des services nécessaires.
    - Configuration et validation de la connectivité entre Kafka, Spark, Airflow, backend, BI et frontends avec quelques tests basiques
2. **Ingestion et couche speed**
    - Collecte des données via API.
    - Publication des événements dans Kafka.
    - Visualisation temps réel dans le 1er dashboard.
3. **Couche batch**
    - Traitement batch avec Spark.
    - Stockage des résultats dans HDFS.
4. **Définition des DAGs Airflow**

- Orchestration des tâches batch et streaming.

5. **Front-end de filtrage de données**
    - Exécution et suivi des requêtes.
    - Gestion des résultats, tri et agrégation.
6. **Documentation finale**
    - Décrire le pipeline complet, les choix techniques et les scripts développés.
