Ecole Normale Supérieure de l'Enseignement Technique

Département Mathématique et informatique

Examen de rattrapage du 2ème semestre 2024/2025

La note :…….	**Date : 03/06/2025**
Module : BIG DATA : Architectures orientées stockage	**Durée : 3h30**
Nom & Prénom :…….	**Filière :……**

Thème : *Analyse des Comportements Utilisateurs sur une Plateforme de Streaming Vidéo*
Technologies : HDFS · MapReduce · Hive · Superset · Cassandra
Environnement : Docker & Docker Compose

I. Objectif global

Implémenter un pipeline Big Data moderne permettant d'analyser de grands volumes de logs utilisateurs issus d'une plateforme de streaming, avec les objectifs suivants :

- Traitement et stockage distribué (HDFS, MapReduce)
- Analyse (Hive)
- Visualisation (Superset)
- Stockage NoSQL orienté colonnes (Cassandra)

II. Données fournies

- **logs_streaming.csv**

Ce fichier contient les **logs de navigation** générés automatiquement lors de chaque session de visionnage par les utilisateurs. Chaque ligne correspond à une **interaction réelle d'un utilisateur avec une vidéo**, incluant :

- user_id : identifiant unique de l'utilisateur

- video_id : identifiant de la vidéo regardée
- category : catégorie de la vidéo (film, série, documentaire...)
- duration : durée du visionnage (en secondes)
- timestamp : date et heure de la session
- device_type : appareil utilisé (mobile, TV, PC...)
- location : ville ou pays de l'utilisateur

- **video_metadata.csv**

Ce fichier contient les **métadonnées descriptives des vidéos disponibles sur la plateforme**. Il ne contient **aucune information sur l'activité des utilisateurs**, mais décrit les vidéos elles-mêmes. Chaque ligne correspond à **une vidéo unique** avec :

- video_id : identifiant de la vidéo (permet de faire le lien avec logs_streaming.csv)
- title : titre de la vidéo
- release_year : année de sortie
- language : langue principale de la vidéo
- duration : durée totale de la vidéo (en minutes)

Lien entre les deux fichiers : video_id est la clé de jointure. Le fichier logs_streaming.csv représente **les usages réels**, tandis que video_metadata.csv représente **l'offre de contenu**.

NB : Les fichiers logs_streaming.csv et video_metadata.csv sont fournis en pièces jointes. Veuillez les utiliser pour importer les données dans HDFS et créer les tables nécessaires dans Hive.

III. Tâches à réaliser

Partie 1 – HDFS

1. Créez les dossiers HDFS suivants :
   `/streaming/users/` et `/streaming/metadata/`.
2. Placez le fichier `logs_streaming.csv` dans `/streaming/users/`.

3. Utilisez le fichier `video_metadata.csv` fourni en pièce jointe (contenant les colonnes : `video_id, title, release_year, language, duration`) et placez-le dans le répertoire `/streaming/metadata/` sur HDFS.

Partie 2 – MapReduce

1. Implémentez un job MapReduce pour **compter le nombre de types d'appareils utilisés par chaque utilisateur**.
2. Implémentez un second job pour **calculer la durée moyenne de visionnage par jour de la semaine**, à partir du champ `timestamp`.
3. Sauvegardez les résultats dans `/streaming/output/device_count` et `/streaming/output/daily_avg`.

Partie 3 – Hive

1. Créez deux tables externes Hive :
    - `logs_streaming` à partir de `/streaming/users/logs_streaming.csv`
    - `video_metadata` à partir de `/streaming/metadata/video_metadata.csv`
2. Écrivez les requêtes HiveQL suivantes :
    - a. Afficher les 5 catégories ayant le plus long temps total de visionnage.
    - b. Afficher les 10 vidéos les plus visionnées en 2023 (avec jointure sur `video_metadata`).
    - c. Calculer la répartition des utilisateurs par type d'appareil en pourcentage.

Partie 4 – Superset

1. Connectez Superset à la base Hive via SQLLab.
2. Créez un tableau de bord avec au moins 3 visualisations :
    - a. Histogramme : durée moyenne de visionnage par catégorie
    - b. Camembert ou carte : répartition géographique des utilisateurs
    - c. Série temporelle : évolution hebdomadaire du nombre de vidéos vues

Partie 5 – Cassandra

1. Déployez Cassandra via Docker Compose.
2. Créez le `keyspace streaming` et la table `user_comments` :
3. Insérez au moins 5 commentaires fictifs.
4. Écrivez les requêtes suivantes :
    - a. Lister les commentaires des vidéos ayant reçu plus de 2 sentiments négatifs.
    - b. Afficher le nombre de commentaires positifs par catégorie.

IV. Livrable demandé

Un **rapport PDF** incluant :

- Explication de l'environnement Docker Compose utilisé (fichier `docker-compose.yml`)
- Étapes réalisées avec captures d'écran
- Code des jobs MapReduce
- Requêtes HiveQL et Cassandra
- Visualisations Superset
- Résultats observés
- Problèmes rencontrés et solutions