Université Tunis El Manar
Faculté des Sciences Economiques et de Gestion de Tunis
Responsable du cours : Besma Ben Amara

1MP ISIDS / Traitement du Big Data
Projet Spark_Avril 2025

Projet Spark Version 3: Analyse de la consommation mondiale d'eau

Présentation du Sujet et du Dataset

L'objectif de ce mini-projet est d'analyser la consommation d'eau à travers le monde en utilisant PySpark. Vous allez manipuler des RDD, utiliser l'API Spark SQL et générer des visualisations pour interpréter les tendances de consommation d'eau.

Le dataset fourni contient des informations sur la consommation d'eau par pays et par année, incluant la consommation totale, l'utilisation agricole, industrielle et domestique, ainsi que l'impact des précipitations et l'épuisement des eaux souterraines.

Le travail demandé est le prétraitement des données, analyse avec Spark SQL, et visualisation des résultats. Ecrire un script PySpark qui de manipuler des RDD et d'exécuter des requêtes SQL sur le dataset de la consommation mondiale d'eau. Vous pouvez l'exécuter dans un environnement Spark pour analyser les tendances d'utilisation de l'eau à travers différents pays.

Objectifs

- Manipuler des RDD en PySpark : transformation et filtrage des données.
- Utiliser Spark SQL pour analyser les tendances de consommation d'eau.
- Générer des visualisations pour interpréter les résultats.

Travail demandé :

Partie 1 : Manipulation des RDD

1. Charger le fichier CSV en RDD et supprimer l'en-tête.
2. Transformer les données pour les rendre exploitables (split, conversion des types).
3. Appliquer les transformations suivantes :
    - map : Transformer chaque ligne en une structure exploitable.
    - filter : Supprimer les valeurs nulles ou aberrantes.
    - reduce : Calculer la consommation totale d'eau par pays.
    - sortByKey : Trier les pays par ordre alphabétique.

Partie 2 : Utilisation de Spark SQL

4. Convertir l'ensemble de données en DataFrame Spark et créer une vue temporaire.
5. Exécuter les requêtes suivantes :
    - Identifier les pays avec la consommation d'eau la plus stable au fil des années.
    - Étudier les tendances de consommation d'eau dans les régions arides.
    - Analyser les pics de consommation d'eau et proposer des explications.
    - Comparer la consommation d'eau entre pays développés et en développement.
    - Déterminer si les politiques de conservation de l'eau ont eu un impact significatif sur la consommation totale.

Partie 3 : Visualisation des Résultats

6. Convertir les résultats des requêtes SQL en DataFrame Pandas.
7. Générer les visualisations suivantes avec Matplotlib :
    - Un graphique en barres montrant la variation de la consommation d'eau entre pays développés et en développement.
    - Un graphique en nuage de points analysant la stabilité de la consommation d'eau par pays.
    - Une visualisation temporelle montrant les pics de consommation d'eau par région et par année.

Partie 4 : Analyse des résultats

8. Répondre aux questions d'analyse suivantes :
    - Quels pays montrent la plus grande stabilité dans leur consommation d'eau ?
    - Quelles tendances spécifiques observe-t-on dans les régions arides ?
    - Quels événements pourraient expliquer les pics de consommation d'eau ?
    - La consommation d'eau diffère-t-elle significativement entre pays développés et en développement ?
    - Les politiques de conservation ont-elles un impact mesurable sur la consommation ?

Livrables

- Un script PySpark contenant les transformations, requêtes SQL et visualisations.
- Un rapport expliquant les résultats et répondant aux questions d'analyse.
- Deadline pour la remise du travail **: 21 avril 2025 à 11h10 Salle E**

Input_Dataset: Google Classroom: cleaned_global_water_consumption.csv