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 1: 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 :
    - Obtenir les 5 pays avec la plus grande consommation totale d'eau.
    - Trouver le pays ayant la plus grande consommation d'eau par habitant.
    - Identifier l'année avec la consommation maximale et minimale d'eau.
    - Calculer la consommation moyenne d'eau par habitant pour chaque pays.
    - Identifier l'impact des précipitations sur la consommation d'eau.

Partie 3 : Visualisation des Résultats

6. Convertir les résultats des requêtes SQL en DataFrame Pandas.
7. Générer les visualisations suivantes avec Matplotlib :
    - Un histogramme des 5 pays ayant la plus grande consommation totale d'eau.
    - Un graphique en barres montrant la consommation moyenne d'eau par habitant.
    - Une courbe illustrant l'évolution de la consommation maximale et minimale d'eau par année.

Partie 4 : Analyse des résultats

8. Répondre aux questions d'analyse suivantes :
    - Quels sont les 5 pays ayant la plus grande consommation totale d'eau ?
    - Quel pays a la plus grande consommation d'eau par habitant ?
    - Comment évolue la consommation maximale d'eau au fil des années ?
    - Quelle année a enregistré la plus faible consommation totale d'eau ?
    - Peut-on observer une tendance générale de l'évolution de la consommation d'eau dans le temps ?
    - Quels facteurs pourraient expliquer les tendances observées dans les graphiques ?

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