

Énoncé de Projet : Big Data Processing

Analyse des Tendances de Consommation Énergétique avec Machine Learning

Contexte

Avec l'augmentation des préoccupations environnementales et la nécessité de gérer efficacement les ressources énergétiques, il est essentiel d'analyser les tendances de consommation énergétique à grande échelle.

Ce projet vise à traiter et analyser des données énergétiques massives pour extraire des informations précieuses qui peuvent aider les entreprises et les gouvernements à prendre des décisions éclairées.


Objectif

Développer une application de traitement de Big Data qui collecte, traite et analyse les données de consommation énergétique provenant de diverses sources, telles que des capteurs IoT, des bases de données publiques et des fichiers CSV.

L'application utilisera Hadoop pour le stockage et Spark pour le traitement et l'analyse des données, en intégrant Spark MLlib pour appliquer des modèles de machine learning.

Étapes du Projet

1. Collecte des Données :

- utiliser le dataset **Climate Change: Earth Surface Temperature Data** à télécharger depuis:  **Climate Change: Earth Surface Temperature Data**
<https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data>

2. Stockage des Données :

- Utiliser Hadoop pour stocker les données collectées.

3. Traitement des Données avec Spark :

- Utiliser Apache Spark pour lire les données stockées dans HDFS.
- Effectuer des transformations et des actions sur les données, comme le nettoyage des données, le filtrage des enregistrements, et l'agrégation des données par période (jour, semaine, mois).

4. Analyse avec Spark MLlib :

- Préparation des Données : Convertir les données en format adapté pour le machine learning (par exemple, en utilisant des DataFrames et des VectorAssembler).
- Modélisation :

- Utiliser des algorithmes de régression pour prédire la consommation énergétique future en fonction de variables historiques (ex. : régression linéaire).
- Appliquer des techniques de classification pour identifier les types de consommation (ex. : classification des utilisateurs en fonction de leurs habitudes de consommation).
- Évaluation des Modèles : Utiliser des métriques telles que RMSE (Root Mean Squared Error) pour les modèles de régression et la précision pour les modèles de classification.

5. Visualisation :

- Utiliser des bibliothèques Python telles que Matplotlib ou Seaborn pour visualiser les résultats de l'analyse et des prédictions (par exemple, des graphiques de tendance, des courbes de précision, etc.).
- Visualiser la performance des modèles avec des courbes ROC ou des matrices de confusion.

6. Rapports et Recommandations :

- Générer des rapports sur les résultats de l'analyse, soulignant les tendances clés et fournissant des recommandations sur l'optimisation de la consommation énergétique.
- Présenter les résultats à l'aide d'un tableau de bord interactif (par exemple, avec Dash ou Streamlit) pour permettre aux utilisateurs d'explorer les données. **(facultatif)**

Technologies et Outils

- Langage : Python
- Stockage : Hadoop HDFS
- Traitement : Apache Spark (PySpark)
- Machine Learning : Spark MLlib
- Visualisation : Matplotlib, Seaborn, Dash ou Streamlit

Livrables

- Code source de l'application de traitement de données.
- Rapport d'analyse avec visualisations et recommandations.