NOTICE D'UTILISATION - TRAITEMENT DES DONNÉES MÉTÉOROLOGIQUES

PRÉSENTATION GÉNÉRALE

Ce script PySpark permet de traiter des données météorologiques stockées au format Parquet, en réalisant des transformations pour passer d'une couche de données brutes (bronze) à une couche de données traitées (silver).

PRÉREQUIS

- Apache Spark installé
- Python 3.x
- Java 8 (JDK 1.8)
- Connecteur MySQL (mysql-connector-j-9.3.0.jar)
- Configuration Hive fonctionnelle

STRUCTURE DES DOSSIERS

- bronze/ : Dossier contenant les données brutes au format Parquet
- silver/ : Dossier de destination pour les données traitées

DONNÉES TRAITÉES

Le script traite plusieurs types de données météorologiques :

- Vitesse du vent à différentes hauteurs (10m, 80m, 120m, 180m)
- Direction du vent à différentes hauteurs
- Température de l'air à différentes hauteurs
- Température du sol à différentes profondeurs
- Humidité du sol à différentes profondeurs

FONCTIONNALITÉS PRINCIPALES

1. **Importation des données** : Lecture des fichiers Parquet depuis le dossier bronze
2. **Conversion des types** : Transformation des colonnes vers des types appropriés
3. **Enrichissement des données** : Ajout de colonnes temporelles (année, mois, jour, heure, jour de la semaine)
4. **Calcul de moyennes** : Création de colonnes agrégées pour simplifier l'analyse
5. **Nettoyage des données** : Suppression des colonnes redondantes ou inutiles

6. **Stockage des résultats** : Écriture des données transformées en format Parquet et dans Hive

UTILISATION

1. Assurez-vous que les dossiers 'bronze' et 'silver' existent
2. Placez vos fichiers Parquet de données météorologiques dans le dossier 'bronze'
3. Exécutez le script avec la commande : `spark-submit traitement_meteo.py`

RÉSULTATS

Les données traitées sont :

- Stockées dans le dossier 'silver' au format Parquet, partitionnées par année
- Insérées dans une table Hive nommée 'silver.meteo', également partitionnée par année

REMARQUES

- Les valeurs manquantes ou NaN sont remplacées par 0 pour les colonnes numériques et par des chaînes vides pour les colonnes textuelles
- Les colonnes dont toutes les valeurs sont 0 ou vides sont automatiquement supprimées
- Les données brutes d'origine ne sont pas modifiées

MAINTENANCE ET ADAPTATION

Pour adapter ce script à d'autres types de données :

1. Modifiez les listes de colonnes (wind_speed, wind_direction, etc.)
2. Ajustez les fonctions de calcul de moyenne si nécessaire
3. Modifiez les paramètres de partitionnement selon vos besoins

CONTACT

Pour toute question ou problème concernant ce script, veuillez contacter l'équipe Data Engineering.

Date de dernière mise à jour : 03/05/2025