Étude de Cas Technique pour le Poste de Senior / Lead Data Analytics Platform

Introduction

Vous trouverez ci-dessous un cas technique que vous devrez réaliser en 3 heures. Ce cas est conçu pour évaluer vos compétences en SQL, votre capacité à organiser et structurer des données dans un datalake, ainsi qu'à effectuer des analyses avancées. Le cas inclut des défis liés aux fonctions de fenêtrage, à la gestion des doublons, aux auto-jointures, et aux procédures stockées, tout en intégrant des schémas RAW, SRC, FCT, DIM, et MART.

Cas : Analyse et Structuration des Données de Production d'Énergies Renouvelables

Objectif : Tester vos compétences en SQL pour l'analyse des données de production d'énergies renouvelables, la gestion des données de sites et de capteurs, et l'organisation des données dans un datalake en utilisant des schémas RAW, SRC, FCT, DIM, et MART.

Description : Vous recevez un ensemble de données de production d'énergie, de sites, et de capteurs sous format csv, comprenant les colonnes suivantes :

Données de Production :

- production_id
- site_id
- sensor_id
- production_date
- energy_generated (en kWh)
- energy_type (solaire, éolien, hydroélectrique, etc.)

Données de Sites :

- site_id
- site_name
- location
- installation_date
- total_capacity (en MW)

Données de Capteurs :

- sensor_id
- site_id
- sensor_type
- installation_date
- status (actif, inactif)

Les données contiennent des doublons et des anomalies que vous devrez traiter.

Vous êtes libre d'utiliser toute instance de base PostgreSQL qui serait à votre disposition pour vos tests.

Si besoin, vous trouverez en Annexe un guide pour lancer une base PostgreSQL en utilisant Docker. Vous y trouverez également les requêtes à exécuter pour importer la donnée brute à partir des fichiers CSV.

Tâches

Partie 1 : Chargement des Données dans le Schéma RAW

1. **Chargement des données brutes :**
    - Créer une table raw_production avec les colonnes suivantes :
        - production_id
        - site_id
        - sensor_id
        - production_date
        - energy_generated
        - energy_type
    - Créer une table raw_sites avec les colonnes suivantes :
        - site_id
        - site_name
        - location
        - installation_date
        - total_capacity
    - Créer une table raw_sensors avec les colonnes suivantes :
        - sensor_id
        - site_id
        - sensor_type
        - installation_date
        - status

Charger les données brutes fournies dans les tables correspondantes (voir Annexe).

Partie 2 : Nettoyage et Transformation des Données dans le Schéma SRC

2. **Nettoyage des données :**
    - Créer une table src_production avec les colonnes suivantes :
        - production_id
        - site_id
        - sensor_id
        - production_date

- energy_generated
- energy_type

Appliquer des transformations pour éliminer les lignes avec des valeurs manquantes ou incorrectes.
Identifier les productions en double (même site_id,sensor_id, production_date) et les supprimer en ne gardant que la plus récente (ne garder le production_id le plus grand).

- Créer une table src_sites avec les colonnes suivantes :
  - site_id
  - site_name
  - location
  - installation_date
  - total_capacity

  Appliquer des transformations pour nettoyer et normaliser les données des sites.
  Identifier et fusionner les enregistrements de sites en double basés sur site_name. Conserver le site_id le plus ancien et mettre à jour les autres colonnes avec les valeurs les plus récentes.

- Créer une table src_sensors avec les colonnes suivantes :
  - sensor_id
  - site_id
  - sensor_type
  - installation_date
  - status

  Appliquer des transformations pour nettoyer et normaliser les données des capteurs.

Partie 3 : Création des Tables Dimensionnelles dans le Schéma DIM

3. **Tables dimensionnelles :**
   - Créer une table dim_sites avec les colonnes suivantes :
     - site_id
     - site_name
     - location
     - installation_date
     - total_capacity
     - site_status : Statut du site basé sur la somme des statuts des capteurs.
       Si tous les capteurs du site sont actifs => ‘Operational’
       Si au moins un capteur du site est actif => ‘Partially Operational’
       Si aucun capteur du site n’est actif => ‘Non Operational’

- Créer une table dim_sensors avec les colonnes suivantes :
    - sensor_id
    - site_id
    - sensor_type
    - installation_date
    - status
    - sensor_age : Âge du capteur en jours basé sur la date d'installation.

Partie 4 : Création des Tables Factuelles dans le Schéma FCT

4. **Tables factuelles :**
    - Créer une table fct_production avec les colonnes suivantes :
        - production_id
        - site_id
        - sensor_id
        - production_date
        - energy_generated
        - energy_type
        - total_capacity : À partir de dim_sites
    - Ajouter les mesures calculées suivantes :
        - energy_efficiency : Efficacité énergétique calculée comme energy_generated / total_capacity
        - cumulative_energy : energie totale cumulée par sensor_id et energy_type

Partie 5 : Création des Tables de Mart dans le Schéma MART

5. **Tables MART :**
    - Créer une table mart_production_summary avec les colonnes suivantes :
        - month : Mois de la production
        - year : Année de la production
        - energy_type : Type d'énergie (solaire, éolien, hydroélectrique, etc.)
        - site_id
        - total_energy_generated : Montant total d'énergie produite pour le mois et le type d'énergie
        - average_daily_energy : Moyenne quotidienne d'énergie produite pour le mois et le type d'énergie
    - Créer une vue matérialisée mart_site_performance avec les colonnes suivantes :
        - site_id
        - total_energy_generated : Montant total d'énergie produite par le site
        - performance_ratio : Ratio de performance calculé comme total_energy_generated / total_capacity

Partie 6 : Analyse Avancée avec SQL

6. **Analyse des données de production :**

- Calculer le rendement moyen des capteurs par site pour les 10 dernières productions de chaque site. La table de sortie doit inclure :
    - site_id
    - production_date
    - avg_last_10_efficiency (total energy generated on last 10 records / total capacity)
- Écrire une requête SQL pour trouver toutes les paires de productions du même site où l'énergie générée de la première production est supérieure à 1000 kWh et l'énergie générée de la seconde production est supérieure à 2000 kWh, dans une période de 30 jours.
- Trouver tous les sites qui ont généré de l'énergie au moins une fois par mois pendant les six derniers mois.

Soumission

Veuillez soumettre vos requêtes SQL et votre documentation au sein d’un repository publique Gihub, GitLab ou BitBucket.

Assurez-vous que votre travail est bien structuré et facile à comprendre. Si vous avez des questions, n'hésitez pas à les poser.

Merci et bonne chance !

Annexe : Installation et Configuration de PostgreSQL en Local avec Docker

Installation de Docker

Si Docker n'est pas encore installé sur votre machine, vous pouvez le télécharger et l'installer à partir du site officiel : Docker (https://docs.docker.com/engine/install/).

Lancement d'un Conteneur PostgreSQL avec Docker

1. **Télécharger l'image PostgreSQL depuis Docker Hub :**

Ouvrez un terminal et exécutez la commande suivante pour télécharger l'image officielle de PostgreSQL :

```
docker pull postgres
```

2. **Lancer un conteneur PostgreSQL :**

Exécutez la commande suivante pour lancer un conteneur PostgreSQL :

```
docker run --name my_postgres -e POSTGRES_PASSWORD=mysecretpassword -d -p 5432:5432 postgres
```

Cette commande démarre un conteneur nommé my_postgres avec le mot de passe mysecretpassword pour l'utilisateur postgres par défaut, et expose le port 5432.

3. **Se connecter à PostgreSQL :**

Vous pouvez utiliser un client PostgreSQL comme psql ou un outil GUI comme pgAdmin pour vous connecter à la base de données PostgreSQL.

Avec psql :

```
docker exec -it my_postgres psql -U postgres
```

Avec pgAdmin :

- Téléchargez et installez pgAdmin depuis pgAdmin Download (https://www.pgadmin.org/download/).
- Ouvrez pgAdmin et ajoutez un nouveau serveur avec les paramètres suivants :
    - **Nom** : Local PostgreSQL
    - **Hôte** : localhost
    - **Port** : 5432
    - **Nom d'utilisateur** : postgres
    - **Mot de passe** : mysecretpassword

Création des Tables RAW

Une fois connecté à PostgreSQL, vous pouvez créer les tables RAW en exécutant les requêtes SQL suivantes :

1. **Table** raw_production **:**

```
CREATE TABLE raw_production (
    production_id SERIAL PRIMARY KEY,
    site_id INT NOT NULL,
    sensor_id INT NOT NULL,
    production_date DATE,
    energy_generated DECIMAL(10, 2),
    energy_type VARCHAR(50)
);
```

2. **Table** raw_sites **:**

```
CREATE TABLE raw_sites (
    site_id SERIAL PRIMARY KEY,
    site_name VARCHAR(100),
    location VARCHAR(100),
    installation_date DATE,
    total_capacity DECIMAL(10, 2)
);
```

3. **Table** raw_sensors **:**

```
CREATE TABLE raw_sensors (
    sensor_id SERIAL PRIMARY KEY,
    site_id INT NOT NULL,
    sensor_type VARCHAR(50),
    installation_date DATE,
    status VARCHAR(50)
);
```

Chargement des Données dans les Tables RAW

Pour charger les données des fichiers CSV dans les tables RAW, vous pouvez utiliser la commande COPY de PostgreSQL :

1. **Charger les données dans** raw_production **:**

```
COPY raw_production(production_id, site_id, sensor_id, production_date, energy_generated, energy_type)
FROM '/path/to/production_renewable.csv' DELIMITER ',' CSV HEADER;
```

2. **Charger les données dans** raw_sites **:**

```
COPY raw_sites(site_id, site_name, location, installation_date, total_capacity)
FROM '/path/to/sites_renewable.csv' DELIMITER ',' CSV HEADER;
```

3. **Charger les données dans** raw_sensors **:**

```
COPY raw_sensors(sensor_id, site_id, sensor_type, installation_date, status)
FROM '/path/to/sensors_renewable.csv' DELIMITER ',' CSV HEADER;
```

Assurez-vous de remplacer /path/to/ par le chemin correct vers vos fichiers CSV.

Résumé

Cette annexe vous guide à travers les étapes pour installer et configurer PostgreSQL en local en utilisant Docker, et fournit les requêtes SQL nécessaires pour créer les tables RAW et y charger les données. Assurez-vous de suivre chaque étape attentivement et de vérifier les connexions et les chemins de fichier pour garantir une configuration correcte.