Partiel de data engineering

Mercredi 8 janvier 2025

Consignes

- Écrire les réponses dans le fichier reponses.nom.prenom.txt (remplacez nom et prenom par les vôtres).
- Toutes les questions valent le même nombre de points, donc répondez en priorité aux questions qui vous parle le plus.
- Envoyez le fichier à luc.marchand.pro@proton.me et maxencetallon@gmail.com.

Questions théoriques
(50 points)

Modèle relationnel

- Définir l’acronyme ACID.

- Expliquer OLAP et OLTP.

NoSQL

- Dans quel cadre le NoSQL est-il plus intéressant que le relationnel ?

- Qu’est-ce que le NoSQL impose à prendre en compte qu’un SGBD fait par défaut

Spark / Hadoop

- Pourquoi Hadoop a été une révolution dans le paysage de l'IT ?

- Expliquer avec vos mots l'approche Map Reduce.

- Expliquer la différence entre Spark et Hadoop.

DevOps / MLOps

- Pourquoi faire du devops ?

- Quelles sont les différentes étapes du cycle DevOps ?

- Comment le rôle du devops est complémentaire avec le rôle du data ing ?

DBT

- Expliquer avec vos mots comment vous utiliseriez DBT au sein d’un projet.

- Comment le mettriez-vous en place ?

Kafka

- A quoi sert Kafka ?

- De quoi est composé un message au sein de Kafka ?

- Sous quel format kafka traite-t-il les messages ?

Delta

- Lister quatre avantages à utiliser Delta Lake par rapport à un format parquet.
- Expliquer.

Sondage

- Quelle est votre perception du rôle de data engineer ?

Partie pratique
(50 points)

Questions Partie SQL

La base de données **factories** contient le résumé de production et de vente de plusieurs entreprises et de leurs usines respectives. Chaque entreprise possède plusieurs usines et chaque usine possède plusieurs lignes de production.

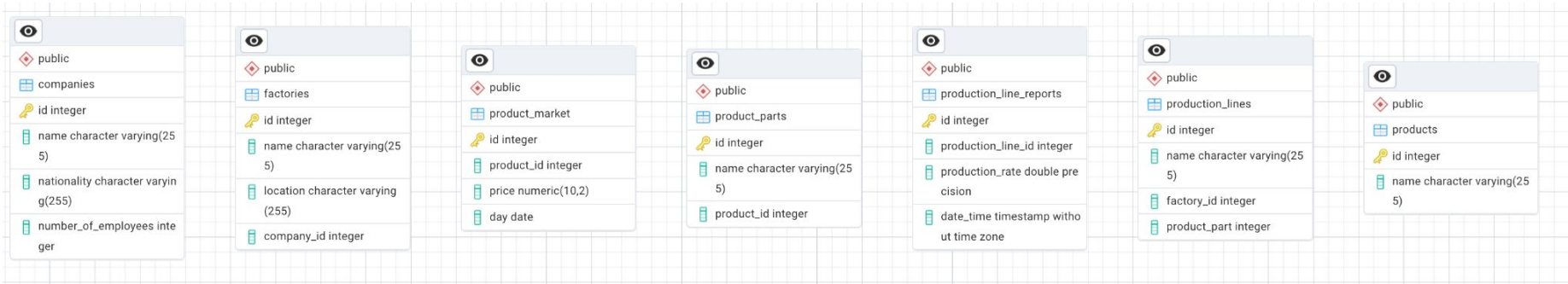
Chaque usine fabrique un seul produit et chaque ligne de production fabrique une partie de ce produit.

Il existe un rapport heure par heure de la vitesse de production des lignes de production (production_rate, produit/heure).

Il existe aussi une table qui contient les prix du marché de vente de chaque produit jour par jour (prix en euros par produit).

Astuce : Utiliser des sous-requêtes ou les CTE pour décomposer votre raisonnement.

Questions Partie SQL



Questions Partie SQL
Q 1

Calculez le nombre de produit fabriqué pour chaque compagnies durant l’année 2024 ?

Astuce : Utiliser des sous-requêtes ou les CTE pour rechercher le rendement le plus bas d’une ligne de production pour définir la vitesse de production d’un produit et arrondir.

Questions Partie SQL
Q 2

Quel produit est le plus rentable pour les usines françaises durant l’année 2024

Astuce : Utiliser des sous-requêtes ou les CTE pour calculer le chiffre d’affaire de chaque usine dans un premier temps.

Questions Partie SQL
Q 3

Quelle compagnie a le pire rendement sur les engrenages(gear) durant le mois février ?

Question Partie MongoDB

Connectez -vous sur NoSql Client sur le localhost:3000

Nom de collection : sensors

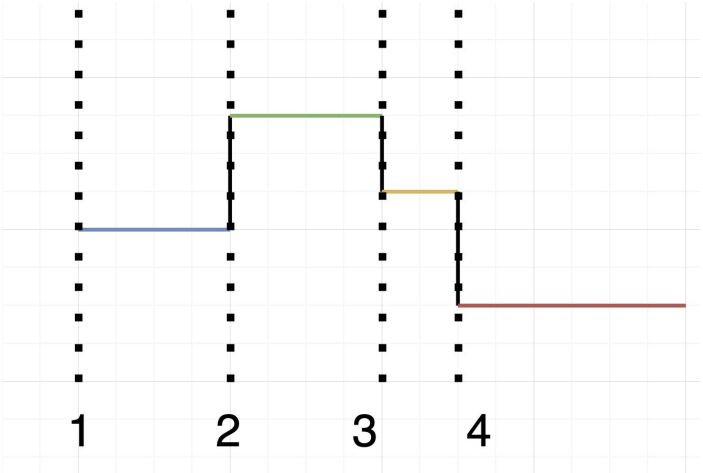
Q1 : Récupérer le nombre de capteurs installé avant le 31 octobres 2024 ? (utiliser les outils aggregate de mongo)

Q2 : Donner le nombre de capteur de qui ont eu leur date de calibration avant leurs date de maintenance.

Q3 : Donner la moyenne des capteurs de pression pour le mois de juillet ?

Questions Spark Sql

L'équipe de Data Engineers vous demande de migrer leur ancien code de nettoyage des données des capteurs vers l'outil Spark. Ils vous expliquent que les capteurs envoient des données chaque fois qu'une nouvelle donnée diffère de la précédente. Cela peut être conçu sous la forme d'un signal créneau.



sensor_ID	sensor_type	value	timestamp
2	pressure	170	2025-01-07T12:58:52.53'
2	pressure	175	2025-01-07T12:59:27'
2	pressure	173,25	2025-01-07T13:00:53'
2	pressure	165,5	2025-01-07T13:01:10'

Vous devez calculer la moyenne de chaque capteur par jour.
Astuce : Vous pouvez utiliser les fonctions de pyspark Window pour calculer les intervalles de temps entre chaque ligne.

Annexes Spark Sql

Script de connexion à la base de donnée en spark

```
from pyspark.sql import SparkSession

spark = SparkSession.builder
        .config("spark.jars", "/usr/local/spark/jars/postgresql-42.6.0.jar")
        .getOrCreate()

df = spark.read \
        .format("jdbc") \
        .option("url", "jdbc:postgresql://esme_postgresql:5432/sensors") \
        .option("driver", "org.postgresql.Driver") \
        .option("user", "postgres") \
        .option("password", "1234") \
        .option("dbtable", "sensor_readings") \
        .load()
```