PRAGMA SAS
Chapter de data, data rangers
Prueba de ingeniería de datos

Planteamiento del reto:

En este reto de ingeniería de datos, queremos que nos demuestres un poco de tus habilidades en el manejo de pipelines de datos, y cómo integras la captura, procesamiento y consulta de información. Para este caso particular queremos que crees un pipeline de datos, que permita obtener información y/o estadísticas de los datos que vas cargando en **micro Batches**.

En Big Data es común encontrar que las cargas en Batch no son suficientes ya que la información se necesita en real time (o near real time), y/o las capacidades de cómputo de tu organización no tienen el suficiente poder para hacer cargas masivas de toda la data. Es entonces cuando es necesario implementar la conocida estrategia "divide y vencerás", y hacemos uso de la ingesta por micro Batches, es decir, tomar una parte de la data, procesarla, disponerla, y repetir el proceso con una nueva fracción de la data faltante. Esto es útil porque no es necesario hacer uso exhaustivo del poder computacional, ya que no toda la data se carga en memoria al mismo tiempo.

Especificación de la data:

Tenemos un conjunto de seis (6) archivos en formatos .CSV llamados "2012-1.csv", "2012-2.csv", ..., "2012-5.csv" y "validation.csv". Todos ellos contienen únicamente tres (3) campos: timestamp, price, user_id. Puedes asumir que los nombres de los archivos .CSV son únicos y están ordenados por hora, es decir, "2012-1.csv" contiene eventos/transacciones que ocurrieron antes de los eventos en "2012-2.csv".

Enlace a los archivos:
https://drive.google.com/file/d/1ejZpGTvZa81ZGD7IRWjObFeVuYbsSvuB/view?usp=sharing

Requerimientos:

1. Descarga la carpeta comprimida que contiene los datos y déjalos en una carpeta exclusiva para este reto que vas a realizar.
2. Construye un pipeline que sea capaz de:
   a. Cargar todos los archivos .CSV excepto el llamado "validation.csv" (El pipeline no debe contener todo el conjunto de datos, es decir, los 5 archivos .CSV al mismo tiempo en memoria en cualquier momento).
   b. Almacena los datos de los archivos .CSV, en una base de datos de tu elección (ejemplo: PostgreSQL, MySQL, etc). El diseño de esta base de datos dependerá de ti, crea la tabla o tablas que creas necesarias con el esquema que creas es adecuado, pero ten presente que todos los .CSV

deben ir en la misma base de datos.

c. A medida que los datos son cargados, realiza un seguimiento de:

• Recuento (Count) de filas cargadas a la base de datos.
• Valor medio, mínimo y máximo del campo “price”.

Se espera que en la ejecución de este pipeline al menos después de que se cargue cada .CSV (pero idealmente después de la inserción de cada fila), estas estadísticas se actualicen para reflejar los nuevos datos. Las actualizaciones del pipeline **NO deben tocar los datos ya cargados**, es decir, hacer “SELECT avg (price) …” para cada actualización no es una buena solución al problema.

3. Comprobación de resultados:

- Imprime el valor actual de las estadísticas en ejecución.
- Realiza una consulta en la base de datos del: recuento total de filas, valor promedio, valor mínimo y valor máximo para el campo “price”.
- Ejecuta el archivo “validation.csv” a través de todo el pipeline y muestra el valor de las estadísticas en ejecución.
- Realice una nueva consulta en la base de datos después de cargar “validation.csv”, para observar cómo cambiaron los valores del: recuento total de filas, valor promedio, valor mínimo y valor máximo para el campo “price”.

Algunas reglas y consideraciones del reto:

- Puedes utilizar cualquier Framework o librería que desees.
- Puedes utilizar cualquier base de datos que desees, lo importante es que muestres cómo te conectas a ella, cómo poblas la(s)
- tabla(s) y cómo realizas las consultas.
- Puedes hacer uso de alguna interfaz gráfica para
- administrar/manipular tu base de datos (ejemplo PgAdmin), o
- puedes hacer uso de línea de comandos.
- Puedes usar cualquier código existente que tengas a disposición.
- No hay una forma definida de resolver esta tarea. Queremos ver la forma en la que piensas para resolver un problema así.
- Las estadísticas se pueden almacenar de la forma que desees: en base de datos, en memoria, en un archivo.
- No te preocupes por el rendimiento, el objetivo es una solución
- funcional.
- Si no logras terminar, no te preocupes, queremos saber hasta
- dónde puedes llegar

Entregables:

Para este reto, te solicitamos por favor que nos hagas llegar en una carpeta .ZIP, o un enlace a algún repositorio en nube (ejemplo Drive), los siguientes elementos:

- Notebook o script en donde tengas el pipeline escrito y documentado el paso a paso.