Tarea #2 - Big Data

Objetivo

Introducir a los estudiantes al procesamiento de tipos de datos complejos, agregaciones y métricas con Apache Spark.

Resultados esperados

Para esta asignación se espera que los estudiantes concluyan dos entregables relacionados:

- Un programa principal que recibirá como entrada un patrón de archivos tipo YAML que contienen las transacciones diarias de diferentes cajas en supermercados y generará múltiples archivos de salida, descritos posteriormente.
- Una serie de pruebas unitarias que permitan corroborar la correctitud de las diferentes funciones internas al programa.

Entrega: Archivo comprimido con código y PDF en TEC Digital a más tardar el 27 de Junio de 2024 a las 11:00 PM

Datos de entrada

Cada archivo de entrada YAML describe las compras realizadas en una caja de un supermercado. La expectativa es que cada corrida del programa cargará un número posiblemente grande de estos archivos.

El formato de cada archivo es:

- Una atributo llamado **numero_caja** que sirve como identificador
- Un grupo de **compras** que contiene la lista de cada una de las compras hechas por clientes.
- Cada **compra** es una colección de productos
- Cada producto tiene tres atributos: **nombre**, **cantidad**, **precio_unitario**

El siguiente es un ejemplo del contenido que tendría un archivo:

```
---
- numero_caja: 42
- compras:
    - compra:
        - producto:
            - nombre: manzana
              cantidad: 6
              precio_unitario: 200
        - producto:
            - nombre: brocoli
              cantidad: 1
              precio_unitario: 450
    - compra:
        - producto:
            - nombre: aguacate
              cantidad: 1
              precio_unitario: 9000
```

Para la ejecución del programa principal, los estudiantes deberán proveer 5 archivos de prueba con al menos 10 compras diferentes cada uno; cada compra con un número variable de productos.

Los archivos deben contener YAML válido.

Programa principal (25 puntos)

Se espera que los estudiantes entreguen un manual en PDF con las instrucciones para ejecutar el programa principal. Idealmente esto debería realizarse con una simple llamada a "spark-submit programaestudiante.py caja*.yaml"

Cualquier detalle necesario para la ejecución debe agregarse en este documento. La imposibilidad de ejecución del programa impedirá la obtención de los puntos.

El producto de la ejecución del programa será una serie de archivos de texto:

- total_productos.csv: contiene 2 columnas que representan el nombre de cada producto y la cantidad total de ese producto vendida en todas las cajas.
- total_cajas.csv: contiene 2 columnas que representan el identificador de cada caja y el **total vendido** por esa caja
- metricas.csv: contiene 2 columnas que representan el tipo de métrica y su valor. En particular deberá generarse las siguientes métricas

- caja_con_mas_ventas: identificador de la caja con más ventas (ventas en dinero)
- caja_con_menos_ventas: identificador de la caja con menos ventas (ventas en dinero)
- percentil_25_por_caja: si se ordenan todas las cajas de menor cantidad de ventas a mayor, **cuál valor monetario** representa el percentil 25
- percentil_50_por_caja
- percentil_75_por_caja
- producto_mas_vendido_por_unidad: nombre del producto que tuvo más unidades vendidas
- producto_de_mayor_ingreso: nombre del producto que generó más cantidad de dinero (e.g. cantidad * precio)

Pruebas esperadas

Para realizar las pruebas unitarias se espera que los estudiantes piensen en las diferentes partes necesarias para conseguir el objetivo final. **Éstas deberán arrancar de datos que se encuentren en memoria (no archivos)**. En este caso, pueden arrancar de dataframes en los que cada fila es el string del YAML para una caja.

Los estudiantes deberán diseñar sus propias pruebas unitarias. Para efectos de evaluación se espera que haya suficientes pruebas para probar las diferentes áreas funcionales. Se espera que cada área funcional tenga su propia función de entrada:

- **Total de productos: 20 puntos**
- **Total de cajas: 20 puntos**
- **Métricas: 5 puntos cada una**

Las pruebas deben cubrir casos excepcionales. En la revisión se reserva el derecho de agregar pruebas unitarias adicionales en cada apartado para asegurar el correcto funcionamiento.

Se recuerda a los estudiantes que la nota será completamente derivada de las pruebas unitarias. Deberá ser posible ejecutar las pruebas simplemente al correr el comando pytest en la carpeta que se entrega con el código.

Agregados en Base de Datos (15 Puntos extra)

En un ambiente de producción se esperaría que los resultados no sean almacenados en archivos CSV, sino en algún servicio sencillo de consultar. Como punto opcional de la

asignación se plantea a los estudiantes reutilizar el código hecho para generar el archivo **CSV de métricas únicamente** y expandirlo. Los requerimientos son los siguientes:

- Permitir que cada compra tenga un atributo opcional **fecha**, con formato "YYYY/MM/DD" (año, mes y día). Dado que es opcional, los estudiantes deben asegurarse que los archivos de prueba pedidos previamente puedan servir como entrada al programa principal. Esto es, no será válido que la información para los puntos extra tenga que ser provista para revisar la tarea base.
- Debe existir código que produzca un dataframe con tres columnas: las dos columnas originales de métricas (nombre de métrica y valor) además de una tercera que representará la fecha, extraída del atributo YAML opcional, mencionado en el punto anterior. Éste debe ser probado con pytest.
- Debe proveerse un script de SQL para crear una tabla en postgresql donde se insertarán los datos.
- Crear un programa principal, **aparte** del solicitado para la tarea primaria, que reciba por parámetros de consola el nombre del archivo YAML y los parámetros de conexión necesarios para insertar en la base de datos (dirección de la máquina o host, usuario, password y nombre de la base de datos). Se pide que se parametrice ya que, para la revisión, se ejecutará el script de SQL para crear la tabla en la base de datos y, posteriormente, se dará al programa principal la información de la base de datos donde se creó la tabla.
- Agregar una sección adicional en el manual explicando el uso del programa y cambios realizados.

Condiciones para puntos extra:

- Debe ser posible cumplir con los requerimientos básicos, aparte de los puntos extra. Por ejemplo, los estudiantes deben proveer los archivos YAML sin atributos de fechas y toda la infraestructura debe funcionar igual. Lo anterior a manera de ejemplo. En general, no se podrá justificar errores en la tarea primaria debido a código para puntos extra
- Los estudiantes deben proveer archivos YAML, adicionales, que sí incluyan fechas.
- Se debe crear un archivo "main" aparte.
- Los estudiantes deben documentar en detalle en el archivo PDF qué diferencias de implementación hubo, además de documentar cómo ejecutar el programa primario adicional.
- Deben proveer pruebas unitarias que revisen los dataframes de 3 columnas mencionadas para poder optar por los puntos extra.