

*Ejercicio de Laboratorio 4: Databricks*

Instituto Politécnico Nacional.

Escuela Superior de Cómputo.

Licenciatura en ciencia de datos.

Bases de Datos Avanzadas

Emiliano López Méndez.

**Introduccion**

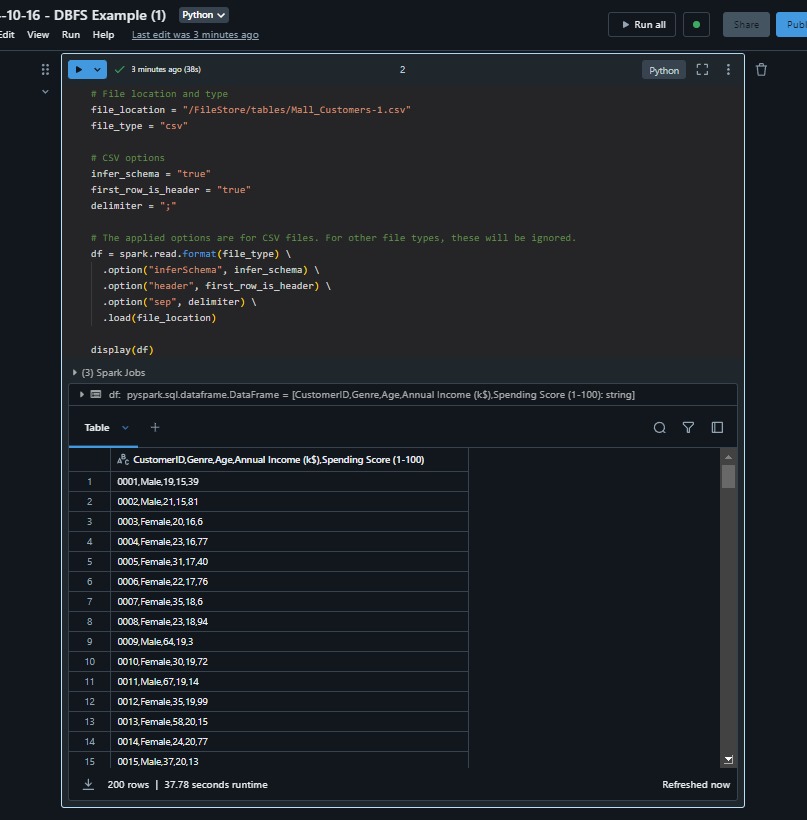
En este laboratorio se explorarán las capacidades de Databricks, una plataforma basada en la nube diseñada para procesar grandes volúmenes de datos utilizando Apache Spark. La actividad consiste en cargar un conjunto de datos, ejecutar consultas tanto en PySpark como en SQL, y almacenar los resultados en formato Parquet. A través de este ejercicio, se pretende reforzar las habilidades en el uso de herramientas de análisis avanzado de datos y comprender cómo utilizar vistas temporales y tablas en un entorno distribuido. Además, se practicará el manejo eficiente de datos estructurados, esencial para proyectos de Big Data.

**Desarrollo de la Actividad:**

El desarrollo de este laboratorio se estructura en varias etapas para garantizar un aprendizaje progresivo en el uso de Databricks y Apache Spark. En la primera parte, se carga el dataset Mall\_customers.csv en el entorno de Databricks, permitiendo explorar su integración con un espacio de almacenamiento en la nube. Posteriormente, se crea una vista temporal con PySpark, lo que habilita la ejecución de consultas mediante la función spark.sql. En la siguiente fase, se aprovecha el modo SQL del notebook para realizar consultas adicionales y generar tablas temporales. Finalmente, se almacena el resultado de las consultas en archivos Parquet, un formato eficiente para datos analíticos, consolidando así la comprensión del manejo de datos en un entorno distribuido.

**Parte 1 Carga de archivo**

En esta sección se procede a cargar el dataset Mall\_customers.csv en el entorno de Databricks. La carga del archivo es fundamental para que los datos estén disponibles en el espacio de almacenamiento de la plataforma y puedan ser utilizados en consultas posteriores. A través de este paso, se garantiza que el archivo se integre correctamente y sea accesible mediante PySpark, facilitando así el análisis de la información en un entorno de Big Data.



Texto

Descripción generada automáticamente

**Parte 2 Consultas con spark.sql**

En esta etapa, se realizan consultas sobre los datos utilizando PySpark mediante la función spark.sql. Esta función permite ejecutar instrucciones SQL sobre las vistas temporales generadas, proporcionando un entorno flexible y eficiente para el análisis de datos. La ejecución de estas consultas permite obtener conocimientos específicos del dataset, como estadísticas descriptivas y segmentación de los clientes. De este modo, se pone en práctica la capacidad de trabajar con datos estructurados mediante Spark en Databricks.

Texto

Descripción generada automáticamente

Captura de pantalla de un celular

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

**Parte 3 Utilizando el modo sql realizar 4 consultas y tabla temporal**

En esta parte del laboratorio se utiliza el modo SQL del notebook para realizar consultas adicionales sobre los datos cargados. Se aprovecha la familiaridad con las consultas SQL para explorar la información disponible en el dataset. Además, se procede a la creación de una tabla temporal mediante la inserción de los datos desde la vista temporal generada previamente. Esta tabla temporal facilita la organización y reutilización de datos en consultas posteriores, mejorando la eficiencia del análisis y la manipulación de la información dentro del entorno de Databricks.

Captura de pantalla de un celular

Descripción generada automáticamente

Captura de pantalla de un celular

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

**Parte 4 Almacenar resultados en archivo parquet**

En esta última fase, se explora cómo almacenar los resultados de las consultas en archivos Parquet. Este formato es ampliamente utilizado en Big Data debido a su capacidad para manejar grandes volúmenes de datos de manera eficiente, con compresión y particionado incorporados. Al finalizar esta parte, los resultados generados estarán disponibles en Parquet, lo que facilita su reutilización y optimiza su almacenamiento para futuras consultas y análisis.

Captura de pantalla de computadora

Descripción generada automáticamente

**Conclusión**

A lo largo de este laboratorio se exploraron las capacidades de Databricks para cargar, consultar y almacenar datos de manera eficiente utilizando PySpark y SQL. El trabajo permitió reforzar conocimientos sobre vistas temporales, consultas relacionales y el uso de tablas temporales, además de familiarizarse con el formato Parquet para el almacenamiento de resultados. Estas habilidades son fundamentales en el análisis de grandes volúmenes de datos y preparan al estudiante para enfrentar retos en entornos de Big Data, optimizando tanto el procesamiento como la gestión de la información.