Programa: Ciencia de Datos

Curso: Big Data

Proyecto Final

Machine Learning para Big Data y PostgreSQL

Alonso Nuñez Sanchez

Esta parte del trabajo final del curso agrupa los conocimientos adquiridos para resolver un problema real con datos reales.

Muestra la aplicación de Machine Learning usando librerias de Spark y MLLib.

Usa distintos algoritmos de Clasificación para tratar un conjunto de datos de más de 22.000 registros. Cada registro es un día de operación de una tienda.

Estos datos corresponden a las ventas diarias de un negocio durante un año. La clase para clasificar corresponde un día de altas ventas (1) o bajas (0), con la intensión de pronosticar el tipo de venta (alta o baja) de acuerdo al día de la semana, la cantidad de cajas registradoras y horas de operación.

Los datos son leídos de una Base de Datos de postgreSQL y posteriormente manejados en memoria en estructuras data frame de Spark

A continuación, cómo se creó la Base de Datos en postgresql, que posteriormente se usará en Spark

Prerequisito: tener instalado y configurado PostgreSQL esto se documentó durante el curso en otra asignación

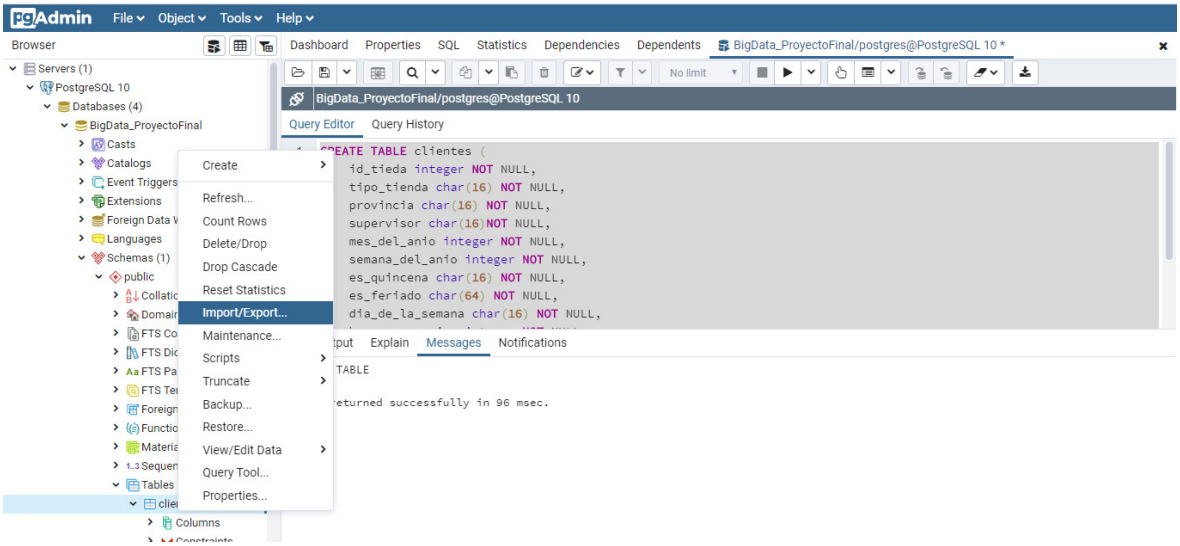
1. **Crear la Base de Datos**

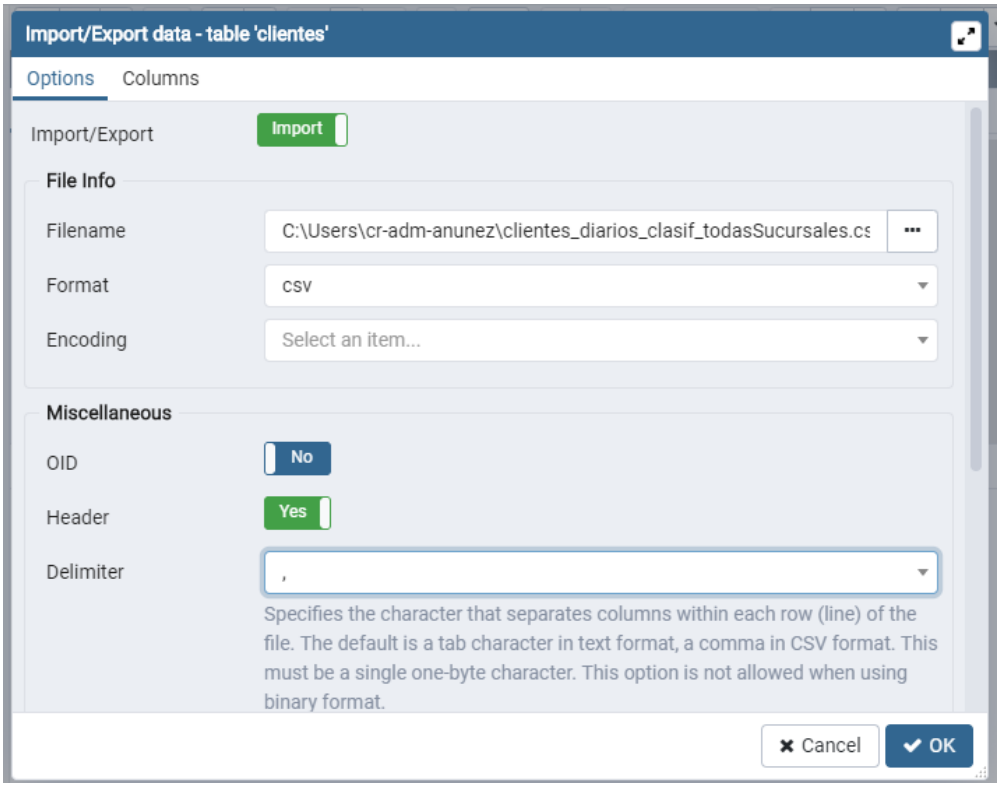
   CREATE DATABASE BigData_ProyectoFinal

2. Crear la tabla

```
CREATE TABLE clientes (
        id_tieda integer NOT NULL,
        tipo_tienda char(16) NOT NULL,
        provincia char(16) NOT NULL,
        supervisor char(16)NOT NULL,
        mes_del_anio integer NOT NULL,
        semana_del_anio integer NOT NULL,
        es_quincena char(16) NOT NULL,
        es_feriado char(64) NOT NULL,
        dia_de_la_semana char(16) NOT NULL,
        horas_operacion integer NOT NULL,
        pos_en_uso integer NOT NULL,
        clientes_totales integer NOT NULL,
        clientes_vip integer NOT NULL,
        label integer NOT NULL
);
```

3. Importar datos





De esta forma, queda creada la Base de Datos, la tabla, y el contenido que será usado posteriormente en Spark

4.
5. **Leer datos desde Spark, usando el conector JDBC**

Desde la notebook de Jupyter se ejecuta el siguiente código, usando el driver correspondiente, credenciales y datos de conexión

```
spark = SparkSession \
    .builder \
    .appName("Basic JDBC pipeline") \
    .config("spark.driver.extraClassPath", "C:\Tarea3\postgresql-42.2.9.jar") \
    .config("spark.executor.extraClassPath", "C:\Tarea3\postgresql-42.2.9.jar") \
    .getOrCreate()

# Lee los datos de la tabla "clientes" en la base de datos "BigData_ProyectoFinal"
# usando el conector JDBC de Postgresql y crea una estructura tipo dataframe de Spark
df = spark \
    .read \
    .format("jdbc") \
    .option("url", "jdbc:postgresql://localhost:5432/BigData_ProyectoFinal") \
    .option("user", "postgres") \
    .option("password", "zeroone") \
    .option("dbtable", "clientes") \
    .load()
```