Fundamentos de

Big Data

Sesión 4

¿Qué es Spark SQL?

Módulo de Apache Spark diseñado para trabajar con datos estructurados.

Características:

1. **Compatibilidad con SQL estándar**
2. **Uso de DataFrames y Datasets**
3. **Optimización con Catalyst**
4. **Compatibilidad con múltiples fuentes de datos**
5. **Interoperabilidad con RDDs**

DataFrame

Colección distribuida de datos

Ejecución diferida (lazy evaluation)

Creación a partir de listas, diccionarios, CSV, JSON, Parquet o RDD

Formatos de Archivo: JSON y Parquet

JSON

- Formato ligero y basado en texto.
- Auto-descriptivo y legible.
- Compatible con la mayoría de los lenguajes de programación.

Ejemplo de lectura y escritura en JSON:

```
df_json = spark.read.json("datos.json")
df_json.write.json("salida.json")
```

Parquet

- Formato de almacenamiento columnar.
- Optimizado para consultas rápidas y compresión eficiente.
- Recomendado para grandes volúmenes de datos.

Ejemplo de lectura y escritura en Parquet:

```
df_parquet = spark.read.parquet("datos.parquet")
df_parquet.write.parquet("salida.parquet")
```

Lectura y Escritura Distintas Fuentes

1 **Lectura desde una base de datos MySQL**

```
df_mysql = spark.read.format("jdbc") \
    .option("url", "jdbc:mysql://localhost:3306/mi_base") \
    .option("dbtable", "usuarios") \
    .option("user", "root") \
    .option("password", "1234") \
    .load()
df_mysql.show()
```

2 **Escritura en una base de datos PostgreSQL**

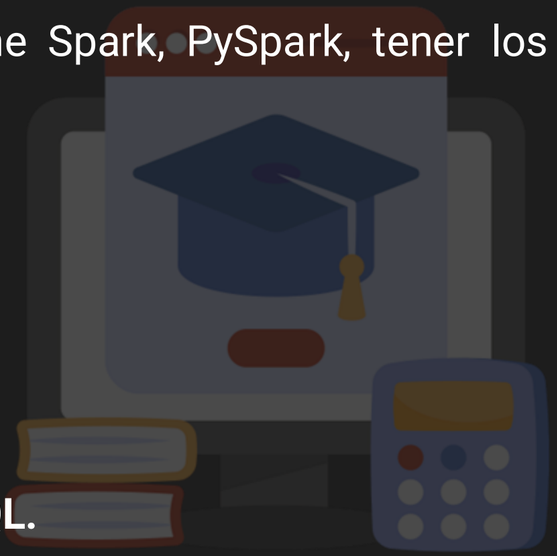
```
df.write.format("jdbc") \
    .option("url", "jdbc:postgresql://localhost:5432/mi_base") \
    .option("dbtable", "usuarios") \
    .option("user", "postgres") \
    .option("password", "1234") \
    .save()
```

Actividad Práctica Guiada

Objetivo: Implementar Spark SQL para procesar datos estructurados, usando DataFrames, consultas SQL, UDFs y Lectura/Escritura en distintos formatos.

Requisitos:

1. **Requisitos previos:** Instalación de Apache Spark, PySpark, tener los datos en formato JSON o CSV (revisar guía de estudio).
2. **Configurar una sesión de Spark.**
3. **Cargar los datos en un DataFrame.**
4. **Explorar el esquema del DataFrame.**
5. **Registrar el DataFrame como una tabla SQL.**
6. **Crear una Función Definida por el Usuario (UDF).**
7. **Guardar los datos en formato Parquet.**
8. **Leer datos desde una base de datos.**



El detalle de la actividad se encuentra en la guía de estudio de la sesión.

Preguntas

Sección de preguntas

Fundamentos de

Big Data

Continúe con las actividades