APACHE SPARK SQL

Profesores:

Miguel Angel Sánchez Hernández
Omar Mendoza González

Alumna:

Belem Anahi Mendieta Hernández

Contenido

- ¿Qué es Apache Spark?
- Fuentes de datos compatibles
- Componentes esenciales de Spark SQL
- ¿Qué es Spark SQL?
- Arquitectura de Spark SQL
- Ventajas de Spark SQL

¿Qué es Apache Spark?

Es un motor de procesamiento distribuido en memoria, diseñado para ser más rápido y flexible, capaz de procesar grandes volúmenes de datos de manera eficiente.

Características clave:

- Más rápido que MapReduce.
- Procesamiento en memoria.
- Compatible con YARN y HDFS.

Fuentes de datos compatibles

Spark soporta múltiples formatos:

CSV

JSON

Parquet

ORC

Avro

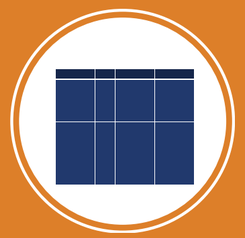
JDBC/SQL

Hive tables

Componentes esenciales de Spark SQL

SparkSession

Punto de entrada a Spark SQL.



DataFrames

Tablas distribuidas con columnas y tipos.

DataSets

Versión tipada de DataFrames

RDD

Colección de objetos dividida en particiones y procesada en paralelo.

SQL API

Ejecuta consultas directamente

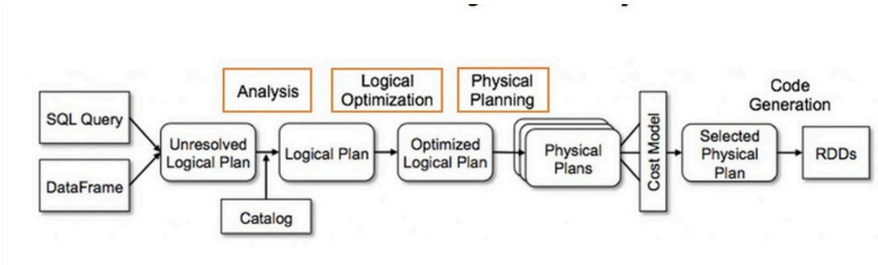
¿Qué es Spark SQL?

Es el módulo de Apache Spark para trabajar con datos estructurados utilizando:

- SQL estándar
- DataFrames
- Datasets.

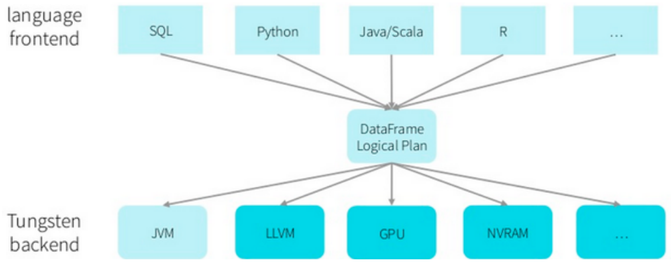
Permite ejecutar consultas distribuidas de forma optimizada y muy eficiente.

Arquitectura de Spark SQL



Catalyst Optimizer

- Optimiza automáticamente las consultas.
- Genera planes lógicos y físicos eficientes.



Tungsten Engine

- Motor físico optimizado.
- Manejo de memoria de bajo nivel para máxima velocidad.

Ventajas de Spark SQL

- Muy rápido (optimización automática).
- Fácil para usuarios de SQL.
- Escala desde una computadora hasta un clúster de cientos de nodos.
- Se integra con MLlib, Streaming y GraphX.