211720118030

Spark SQL and Data Frames

Homework 6

Q1. What is Spark SQL?

Spark SQL is a Spark module for structured data processing. It integrates relational processing with Spark's functional programming. It provides support for various data sources and helps to construct SQL queries with code transformations which resulting in a very powerful tool.

Q2. Is there a module to implement SQL in Spark? How it works?

> Using SparkSession we can access Spark functionality by importing the class and create an instance in the code. To issue any SQL query, use the sql() method on the SparkSession instance, spark , such as spark. sql("SELECT * FROM myTableName") .

> Spark SQL provides Data frame APIs which perform relational operations on both external data sources and Spark's built-in distributed collections. Spark runs on both Windows and UNIX-like systems. It can run locally by installing java on our system path, or the JAVA_HOME environment variable pointing to a Java installation.

Q3. What is a Parquet file?

Parquet is an open source file format built to handle flat columnar storage data formats. Parquet operates well with complex data in large volumes. It is known for its both performable data compression and its ability to handle a wide variety of encoding types.

Q4. List the functions of Spark SQL.

>String Functions

>Math Functions

>Aggregate Functions

>Collection Functions

>Date & Time Functions

>Window Functions.

Q5. How is Spark SQL different from HQL and SQL?

Hive Query Language (HQL) and is used for data stored in Hadoop Distributed File System (HDFS) whereas Spark SQL makes use of structured query language and makes sure all the read and write online operations.

Q6. Why is Spark SQL used?

Spark SQL is used because of in-memory computing which results better performance. It includes a cost based optimizer, columnar storage and code generation for faster execution of queries. Spark SQL has various performance tuning options like memory settings, codegen, batch sizes and compression codes.

Q7. Is Spark SQL faster than Hive?

Spark SQL is faster than Hive when it comes to processing speed because the operations in Hive are slower than Spark SQL in terms of memory and disk processing as Hive runs on top of Hadoop.