Spark SQL and DataFrames

Homework 5

SUJATHA.K

202001161

3rd yr CSE C

Q1. What is Spark SQL?

Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and dats

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

Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and dats

Q3. What is a Parquet file?

Apache Parquet is a free and open-source column-oriented data storage format in the Apache Hadoop ecosystem. It is similar to RCFile and ORC, the other columnar-storage file formats in Hadoop, and is compatible with most of the data processing frameworks around Hadoop

Q4. List the functions of Spark SQL.

Spark SQL Functions – Contents
String Functions.
Date & Time Functions.
Collection Functions.
Math Functions.
Aggregate Functions.
Window Functions

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

Hive, on one hand, is known for its efficient query processing by making use of SQL-like HQL(Hive Query Language) and is used for data stored in Hadoop Distributed File System whereas Spark SQL makes use of structured query language and makes sure all the read and write online operations are taken care of

Q6. Why is Spark SQL used?

Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data

Q7. Is Spark SQL faster than Hive?

Yes Spark SQL is faster than Hive but many students are confused and thinking if the spark is better than hive than why should people working on Hadoop and hive. Spark process data in-memory or distributed ram that makes processing speed faster but Hadoop hive store intermediate results into a disk and perform disk I/O operation that is why hive is slow.

but do not forget Hadoop is not meant for small data processing we made it for big data processing so if we are dealing with huge amounts of data (TB and PB's)distributed over a cluster then we will use Hadoop Hive.

Spark can only process data that can fit into distributed RAM