BIG DATA ANALYTICS

Spark SQL and DataFrames

Homework 6

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 data.

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

Spark SQL is a module in Spark which integrates relational processing with Spark's functional programming API. It supports querying data either via SQL or via the Hive Query Language.

Spark SQL blurs the line between RDD and relational table. It offers much tighter integration between relational and procedural processing, through declarative DataFrame APIs which integrates with Spark code.

It also provides higher optimization. DataFrame API and Datasets API are the ways to interact with Spark SQL.

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 performant data compression and its ability to handle a wide variety of encoding types.

Q4. List the functions of Spark SQL.

The following are the various functions offered by Spark :

- 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?

- It runs SQL queries over imported data and existing RDDs
- Easily write RDDs out to Hive tables or Parquet files
- Import relational data from Parquet files and Hive tables
- It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.
- It also provides powerful integration with the rest of the Spark ecosystem (e.g., integrating SQL query

Q7. Is Spark SQL faster than Hive?

Spark SQL is faster than Hive. For example, if it takes 5 minutes to execute a query in Hive then in Spark SQL it will take less than half a minute to execute the same query.

The operations in Hive are slower than Apache Spark in terms of memory and disk processing as Hive runs on top of Hadoop.