211720104151 BIG DATA

SURYA K HOME WORK 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 Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine.

1. Start the Spark shell. dse spark.
2. Use the sql method to pass in the query, storing the result in a variable.
   val results = spark.sql("SELECT * from my_keyspace_name.my_table")
3. Use the returned data.

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?

- String Functions.
- Date & Time Functions.
- Collection Functions.
- Math Functions.
- Aggregate Functions.
- Window Functions.

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

Hive is a distributed data warehouse platform which can store the data in form of tables like relational databases whereas Spark is an analytical platform which is used to perform complex data analytics on big data.

SparkSQL is a special component on the sparkCore engine that support SQL and HiveQueryLanguage without changing any syntax.

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?

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

Hive is the best option for performing data analytics on large volumes of data using SQLs. Spark, on the other hand, is the best option for running big data analytics. **It provides a faster, more modern alternative to MapReduce**.