211720104133 **BIG DATA**

SATHYANATHAN H **HOMEWORK 6**

#Question 1. 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.

#Question2. IS ther 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 data.

#Question3. What is Parquet file?

Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk.

Parquet is **a columnar format that is supported by many other data processing systems**. Spark SQL provides support for both reading and writing Parquet files that automatically preserves the schema of the original data.

#Question4. List the functions of Spark SQL.

Spark SQL Functions

- String Functions.
- Date & Time Functions.
- Collection Functions.

- Math Functions.
- Aggregate Functions.
- Window Functions.

#Question5. How Is Spark SQL different from HQL and SQL?

SparkSQL is a special component on the sparkCore engine that support SQL and HiveQueryLanguage without changing any syntax. It's possible to join SQL table and HQL table.

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**.

#Question6. Why is Spark SQL used?

Spark SQL is used as the loading and querying can be done for data from different sources. Hence, the data access is unified. It offers standard connectivity as Spark SQL can be connected through JDBC or ODBC. It can be used for faster processing of Hive tables.Spark provides a faster and more general data processing platform. Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop**.**

#Question 7. Is Spark SQL faster than Hive?

Yes. Spark SQL query execution is in-memory while Hive SQL query gets transformed in to map reduce jobs which persists intermediate results to disk. So Spark SQL is usually 100 times faster if not 1000s. Hive provides a way to write SQL queries over tons of data that does not fit in memory like 100+ TB of data over 50+ nodes. Spark SQL is usually for data sets that can be more or less fits all in memory, definitely < 100 TB beyond which Spark will run out of memory and will have no advantage rather problems to give results.pdf