25/25

Anay Shah
60004210073
TE Comps C12

Experiment 8

Aim:- To implement SparkQL in PySpark

Theory:- SparkQL is a Spark component that supports querying data either via SQL or via the Hive Query Language. It originated as the Apache Hive port to run on top of Spark (in place of MapReduce) and is now integrated with the Spark stack. Spark SQL introduces SchemaRDD, a new data abstraction that provides support for structured and semi-structured data. Features of SparkQL:-

i) Compatibility with Hive: Hive queries can be executed in SparkSQL as they are.
ii) Unified Data Access: Loading and querying data from various sources is possible.
iii) Standard Connectivity: SparkQL can connect to Java and Oracle using JDBC (Java Database Connectivity) and ODBC (Oracle Database Connectivity) APIs.
iv) Performance and Scalability: To make queries agile, alongside computing hundreds of nodes using the Spark engine, Spark SQL incorporates a code generator, a cost-based optimizer, and columnar storage. This provides complete mid-query fault tolerance

Advantages

i) Important advantage of SparkQL is that the loading and querying can be done for data from different sources. Hence, the data access is unified.

ii) It offers standard connectivity as Spark SQL can be connected through JDBC or ODBC.

iii) It can be used for faster processing of Hive tables.

Disadvantages

i) Creating or reading tables containing union fields is not possible with Spark QL.

ii) It does not ~~query~~ convery if there is any error in situations where the varchar is oversized.

iii) It does not support Hive transactions.

Conclusion:- Overall Spark SQL is a powerful tool for processing structured data in spark, offering the benefits of SQL queerying, DataFrame API & integration with spark's ecosystem, with considerations for performance optimization & data processing efficiency.