HOME WORK-5

SINDHU V

202001139

3rd yr CSE-C

1.What is a metastore in Hive?

Metastore is the **central repository of Apache Hive metadata**. It stores metadata for Hive tables (like their schema and location) and partitions in a relational database.

2. Where does the data of a Hive table get stored?

Hive stores its database and table metadata in a metastore, which is a database or file backed store that enables easy data abstraction and discovery.

3. Why Hive does not store metadata information in HDFS?

Hive stores metadata information in the metastore using RDBMS instead of HDFS. The reason for choosing RDBMS is to achieve low latency as HDFS read/write operations are time consuming processes.

4. What is the difference between local and remote metastore?

Local Metastore:- Here metastore service still runs in the same JVM as Hive but it connects to a database running in a separate process either on same machine or on a remote machine.

Remote Metastore:- Metastore runs in its own separate JVM not on hive service JVM.

5. What is the default database provided by Apache Hive for metastore?

Derby is the default database for the embedded metastore. Derby embeded JDBC driver class.

6. What is the difference between external table and managed table?

Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with the data.

7. Is it possible to change the default location of a managed table?

Yes, you can do it by using the clause – LOCATION '<hdfs_path>' we can change the default location of a managed table.

8. What is a partition in Hive?

The partitioning in Hive means dividing the table into some parts based on the values of a particular column like date, course, city or country.

9. Why do we perform partitioning in Hive?

Hive organizes tables into partitions. It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, and department. Using partition, it is easy to query a portion of the data.

10.What is dynamic partitioning and when is it used?

Dynamic programming is the stratergic approach to load data from the non-partitioned table where the single insert to partition table is called dynamic partition.

11.Suppose, you create a table that contains details of all the transactions done by the customers of year 2022: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ; Now, after inserting 50,000 tuples in this table, you want to know the total revenue generated for each month. But, Hive is taking too much time to process this query. How will you solve this problem and list the steps that you will be taking in order to do so?

CREATE TABLE partitioned_transaction (cust_id INT, amount FLOAT, country STRING) PARTITIONED BY (month STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

INSERT OVERWRITE TABLE partitioned_transaction PARTITION (month) SELECT cust_id, amount, country, month FROM transaction_details;