Hive_Scenerio_Based_Quetions

26 October 2022   15:13

1.Will the reducer work or not if you use "Limit 1" in any HiveQL query?
=> No.

2.Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same time?
=> It will not connecting with other clients because default hive database 'Derby' that allows to connect with only single client at time.

3.Suppose, I create a table that contains details of all the transactions done by the customers: CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ; Now, after inserting 50,000 records in this table, I want to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?
=> We can do partitioning on "Month" column ,so when we run query for particular month then it will only go to the respective directory only inplace of traversing the whole dataset, So the latency will be decrease from previous.

4.How can you add a new partition for the month December in the above partitioned table?
=> Yes with the help of 'Alter' Command we can add new partition.
ALTER TABLE table_name ADD PARTITION (month='Dec') LOCATION '/table_name';

5.I am inserting data into a table based on partitions dynamically. But, I received an error – FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column. How will you remove this error?
=> We can use this commands to remove this error and set the partitioning in non strict mode so that it will allow to create dynamic partitioning without creating one static partitioning .
 set hive.exec.dynamic.partition=true;
 Set hive.exec.dynamic.partition.mode=nonstrict

6.Suppose, I have a lot of small CSV files present in the input directory in HDFS and I want to create a single Hive table corresponding to these files. The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files. So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?
=>Using SequenceFile format and grouping these small files together to form a single sequence file can solve this problem

7.Suppose, I have a CSV file – 'sample.csv' present in '/temp' directory with the following entries: id first_name last_name email gender ip_address How will you consume this CSV file into the Hive warehouse using built-in SerDe?

=>

8.Is it possible to add 100 nodes when we already have 100 nodes in Hive? If yes, how?
=>Yes
Step 1: Take a new system; create a new username and password
Step 2: Install SSH and with the master node setup SSH connections
Step 3: Add ssh public_rsa id key to the authorized keys file
Step 4: Add the new DataNode hostname, IP address, and other details in /etc/hosts slaves file:

9.LOAD DATA LOCAL INPATH 'Home/country/state/' OVERWRITE INTO TABLE address;The following statement failed to execute. What can be the cause?

=>