TRENDYTECH

UPLIFT YOUR CAREER!

Assignment

Week5: Apache Hive Advance - Part1

Assignment

Problem Overview: Aim is to use the Big data tools and technologies for the analysis of the two given datasets based on covid-19 spread in India and the Testing associated with it and try to get some insights.The data is available between 1st-April to 10th June duration.

Detailed Problem Statement:

A. Need to create a data flow pipeline ,where data initially resides in RDBMS and needs to be brought to Hive so as to get a consolidated view of covid cases details as well as testing details,taken together in one table.

B. Optimizations during sqoop import/export , hive optimizations have to be considered.Also password encryption in sqoop to be used.

C.The data field is not in proper format and has to be taken care

D.Try to infer from the final consolidated table in Hive like whether there is any discrepency between Number of confirmed covid cases in the state Vs.Number of Positive Samples collected during testing.Which state shows least discrepency.

E.Run some more interesting queries from your end in hive to get more insights on the consolidated data .

Sample Query -For every state, find the total number of confirmed cases reported and also total number of positive samples tested,in the entire duration of 2months, starting with the state with the highest cases.

Dataset- Two csv files are provided -

1. **StatewiseTestingDetails.csv**
2. **Covid19_india.csv**

StatewiseTestingDetails.csv : This dataset contains statewise,daily count of Testing performed in India.
Columns : Seq,Date,State,TotalSamples,Negative,Positive

Covid19_india.csv: This dataset contains statewise,daily count of covid cases in India.
Columns : Sno,Date,State/UnionTerritory,Cured,Deaths,Confirmed

Note : If we scan the data we can see the data is cumulative for example TotalSamples shows cumulative total ,Confirmed field shows cumulative value.

Points to Note/Few Assumptions

1. We do not have data for every state in these two files.Also we do not have data for every date in these two files.Some States and dates might be missing in these datasets.

2. Joining of two datasets might not be of much signifiance in this case but this is to check the understanding of hive and whether in real time you can work on this type of dataflow involving hive and sqoop.

3. We assume that more data is added to both the tables on a frequent basis.So we need an incremental sqoop job

4. More queries on state and dates can come in future so , bucketing and partitioning to be used on these columns.

5.We assume that StatewiseTestingDetails.csv is the smaller dataset among the two.

Steps Overview:

Step 1. Data Preprocessing-
Data is assumed to be in Mysql initially. To mimic that-

- **Copy the given csv files from local to hdfs**
- **Creation of Mysql Tables-**
  Hint: Sno and Seq can be used as Primary Keys.Date can be varchar type
- **Sqoop export of data from HDFS to Mysql.**
  Hint :Use Staging table while exporting,use password encryption

- **Delete the data from hdfs post export**

Now data resides in Mysql.

Step 2- Sqoop Import-Bring the data from Mysql tables to HDFS
Hint: Create sqoop jobs for importing both tables
Use other necessary optimizations
Use Password Encryption

Step 3- Create Hive External Tables on top of data in HDFS.

Step 4- Create Optimized External tables in Hive:

Hint: File format -*use TextFile for time being* (though ORC is a better choice and can be used in future)
Use Dynamic Partitioning on State Column, Bucketing on Date Column
Note: *There might be frequent queries on State and Date, so we choose these columns for Partitioning and Bucketing, both tables can be Partitioned on State and Bucketed on Date column.*

Step 5: Load data to the optimized hive tables from normal hive tables.

*Hint: Use **INSERT OVERWRITE clause***
*Date has also to be formatted to proper hive format which is **yyyy-mm-dd** .Use **from_unixtime** , **unix_timestamp***

Step 6-Inner Join two tables in Hive and get a consolidated table.

Hint: Perform Map-side join of the two tables. Join columns can be 'date' and 'state' for better optimization.Here it is assumed that the State_Testing table is small enough to fit in memory.Use Hints

Step 7: Analysis - Do the required inference as mentioned in the problem statement, run queries and see the results.

For Example-
Ideally the number of samples tested positive and number of covid cases confirmed must be the same.*See which state/states have more*

consistent data collection like The number of positive samples (table1) match mostly with number of confirmed cases(table2), for which state.

.

For every state,find the total number of confirmed cases reported and also total number of positive samples tested,in the entire duration of 2months, starting with the state with the highest cases.

You can run additional queries as per your understanding to get more interesting insights from this consolidated data.

****