TRENDYTECH

UPLIFT YOUR CAREER!

Assignment

Week3: Apache Sqoop - Moving Data into Hadoop

Assignment -Week3
total marks - 100

Qu 1) Suppose we have a **test_db** database in mysql. We have an input table **Customers** inside **test_db. (SQL Commands are given)**

Cust_Id	**Customer_Name**	**Purchase_Date**	**Item**	**City**	**Price**	**Cust_Type**
100	Rishi	2020-08-16	Mobile	Kanpur	10000	Regular
200	Venu	2019-05-04	Laptop	Bangalore	61000	Premium
300	Priya	2018-06-25	Mobile	Jaipur	20000	Premium
400	Rini	2019-01-30	Handbag	Pune	1000	Regular
700	Deepu	2019-12-12	Appliances	Mumbai	25000	Premium

The table has a Primary key on the Price column (which of course is not the right choice as prices may repeat when data grows).

Do the following: Share Snapshots of the command and Snapshot of the result in each case:

1) Before performing the sqoop import, using the sqoop command display the data present in mysql **Customers** table.The output of the command should not display on the console,rather should be redirected to log file named '**query.outpu**t'. Display the contents of the **query.output** file , share the Snapshot of the command and the output. - **(5 marks)**

2) Perform a single sqoop import inside the directory in hdfs named **sqoop_importdir**, considering all the following points: - **(20 marks)**

- Import all the columns except Cust_Type in hdfs.
- Include only the purchases made after **2019-01-01**
- The output data generated should have fields separated by | and rows separated by ; (semicolon)
- While importing, Nulls in the data , should be overridden with **'NA'**
- Redirect the log messages generated on screen to the files **log_out1** and **log_out2.**Display the contents of the **log_out2** file , when sqoop import is successful,share the snapshot of the number of records retrieved.
- Display the contents of the **sqoop_importdir**

- Now Again modify and run your sqoop import command ,so that cust_id column can be used to decide the input splits, as the Primary key column is not proper. **Also ensure that the output directory remains as sqoop_importdir, and the previously imported contents are automatically deleted and new contents are filled in the output directory.**
- Display the contents of the output directory now and the first 10 records from the mapper output files (hint: use head command)

- Now Suppose an outlier comes into the mysql table:

The new record inserted is :

Cust_Id	Customer_Name	Purchase_Date	Item	City	Price	Cust_Type
10000	Raman	2019/09/04	Misc	Cochin	9000	Regular

Mention the sqoop import command you will frame from your end to deal with such a situation to ensure even work distribution among mappers, using customized bounding val query.

Note: you got to know that cust_id 10000 is erroneous record and should not be taken care.

Qu 2) Suppose we have a database named **test_new_db** in mysql, We have three tables inside it:
City_Tbl (Consider this is the bigger table)
State_Tbl (Consider this is the smaller table)
Country_Tbl (Smaller Table)

City_Tbl: City_ID is the Primary Key Column

City_Name	City_ID
Bangalore	1000
Mumbai	1001
Chennai	1002
Kolkata	1003
Delhi	1004
Pune	1005
Nagpur	1006
Surat	1007
Kochi	1008

State_Tbl: No Primary Key Column

State_Name	Districts
Karnataka	30
TamilNadu	32
Goa	2
Kerala	14
Assam	33

Country_Tbl: No Primary Key Column

Name	**Country_Code**
Belgium	32
Brazil	55
France	33
Iran	98
India	91

A) Using a single sqoop import command,
Import all the tables present in test_new_db to hdfs excluding the Country_Tbl . You have to do it with a single sqoop command.

Also, City_Tbl should have 3 output files generated in hdfs. All the output files should be stored inside sqoop_all_tbl directory in hdfs, with sub-directories of each table name created inside the main directory. Share the snapshot of the command. (5 marks)

B) Show the contents of the output directory: (Share Snapshot)
(5 marks)

Qu 3) We have a **Categories** Table in **test_db** in Mysql. On this table both inserts and updates are performed from time to time.

Do the following:

A) Import the Categories table in hdfs but during the import ,do proper Null value handling:

- String Columns nulls should be replaced with ‘\\N’ (so that in file it should be read as \n and Non-string column nulls should be replaced with -1
- Use a warehouse directory
- We also want to see the query run by each mapper internally

Share the import command you will use,keeping in mind all of the above. Initially all records to be pulled in. **(10 marks)**

B) New Records are added to the table and also existing records are updated,(refer the mysql_commands text file for the insert and update commands), so import only those newly inserted/updated records from Categories table to hdfs. The delta records should get appended to existing directory.

Share the import command you will use this time, to get only delta records
(10 marks)

C) After this second import, how many records do you see in the hdfs folder now? Did you find any duplicate records, give details if any. **(5 marks)**

D) Create a new table in test_db named Categories_new. The command has been shared in mysql_commands text file.

This newly created table does not have a Primary key.

We want to do periodic imports and updates in this mysql table. But we do not want any duplicate records in the hdfs post import. Also we want to automate the process of import & want a good way to manage the password. Choose a different warehouse directory this time.

Note: The table creation command for Categories_New and fresh inserts and updates command has also been shared in mysql_commands file.

Share the commands you will use when:

- First time we need to pull all records in hdfs
- Second time to pull only the delta records,but without duplicates in hdfs

(25 marks)

E) How many records do you see this time in hdfs post second import? Do you see any duplicate records now? **(5 marks)**

F) Are any mapper files generated in hdfs this time after the second import? Explain. **(5 marks)**

G) Share the command you will use to see the last value of a Saved Sqoop Job. **(5 marks)**