Web: Inceptez.com Mail: info@inceptez.com Call: 7871299810, 7871299817

HDFS Sqoop Hive Shell Script UseCase for Self Learning

The Scope of this is use case is to improve the self learning capabilities (searching online, our documents, project documents etc.,), with automation of data pipeline from DB till Hive. The below script should be capable enough to run for the first time, daily incremental run and rerun incase of any change happen in the source DB.

1. Create and load the below data into mysql table orders table

create table custdb.orders(id integer,product varchar(100),qty integer,amount float,updts timestamp);

insert into custdb.orders values(1,'laptop',2,80000,'2019-09-20 00:00:01'),(2,'mobile',1,10000,'2019-09-20 00:00:02'),(3,'tablet',1,20000,'2019-09-20 00:00:03'),

(4,'laptop',1,50000,'2019-09-20 00:00:03');

2. First try to execute all the steps directly without writing any shell script, then try to convert writing a Linux shell script which should take arguments such as updts from when the data has to be imported and type of data to store with (orc or parquet) in the final hive table.

Example:

bash loaddata.sh 2019-09-20\ 00:00:00 parquet

Note:

A. Provide proper logging, check for the execution state ie succes or failure of the sqoop import and hive data loads and terminate the job as needed.

B. Clue: Run Hive using (hive -e) option

3. Import the above data into HDFS (overwrite the data if already exists) using incremental option with --check-column of updts and --last-value $1 to

import inserted or updated data;

4. Load the above incrementally imported data into hive table (decide on the type of table managed/external).

5. Based on the second argument $2 (orc or parquet), create a hive partitioned table (decide on the type of table managed/external, decide on the name of

the table based on second argument) eg: if orc then create the table as orders_partition_orc else if parquet then create the table as orders_partition_orc

and load the with partition column as upddt which is the date portion of updts.

6. Iterate from step 1 (insert with 2019-09-21 00:00:01 data) till step 4 for loading a different date partition in the final hive table.