How to Handle Incremental Load in Hive?

Suppose we have a input directory or a staging directory , in project we will be getting daily data in the form of files .

Day 1 we got Day1.csv

Day 2 we got Day2.csv

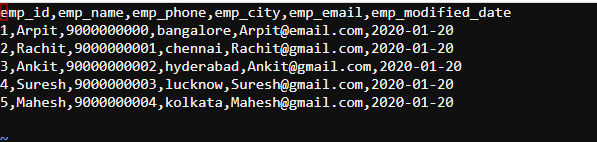
May be this Day 2 file can have new data or may be having some data which have changes in the day 1 .

Now the requirement is we need to read all the data on daily basis and we have to analyses the incremental load .

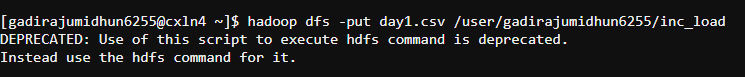
Incremental Load in the sense what new data we have , and what changes in the previous data we have.

We will analyze this incremental or delta data in hive and finally we will create a view which shows the final incremented data.

Suppose Day 1 we got the file Day1.csv



We are moving this to hdfs as shown below …

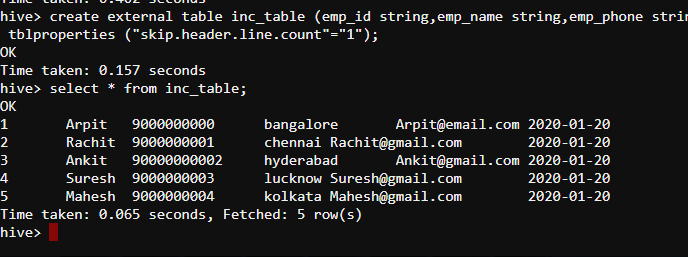


Now we will creating the Hive table on top of this .

hive>

create external table inc\_table (emp\_id string,emp\_name string,emp\_phone string,emp\_city string,emp\_email string,modified\_date date)

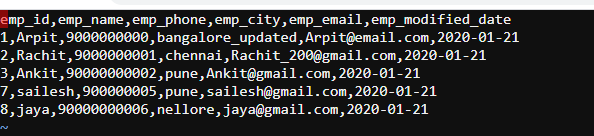
row format delimited fields terminated by ',' location '/user/gadirajumidhun6255/INC\_load' tblproperties ("skip.header.line.count"="1")



Now day1.csv data got loaded in inc\_table.

All the data has come on Jan 20,2020.

Now lets create another dataset day2.csv as shown below .



We have updated some of the values for first 3 employees and introduced 2 new employees.

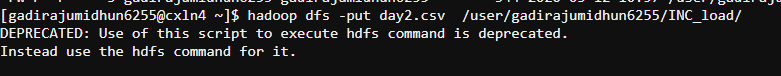
For first employee 🡪 Bangalore is updated to Bangalore\_updated

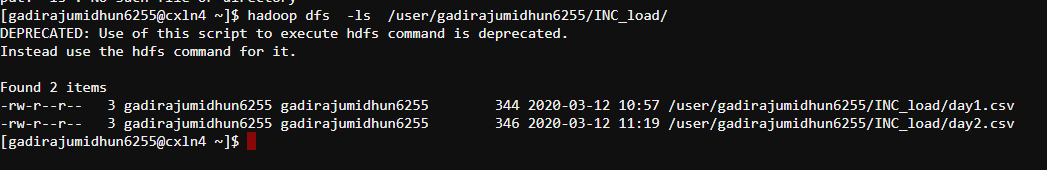
For second employee 🡪 Rachit email is updated to [Rachit\_2000@gmail.com](mailto:Rachit_2000@gmail.com)

For 3rd employee city is updated from Hyderabad to Pune.

7th and 8th employee id is introduced .

Now lets move to INC\_load folder in HDFS .

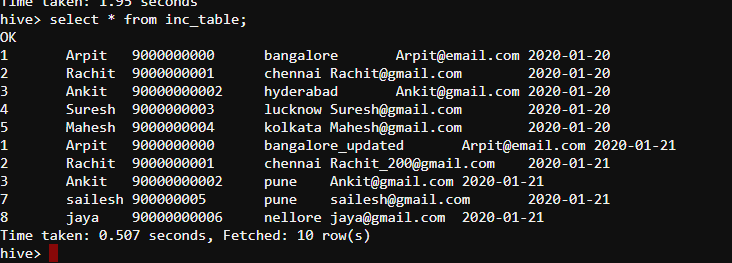




Now already a external table is created on this folder .

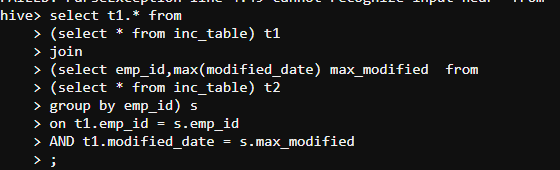
use hive\_incremented\_load;

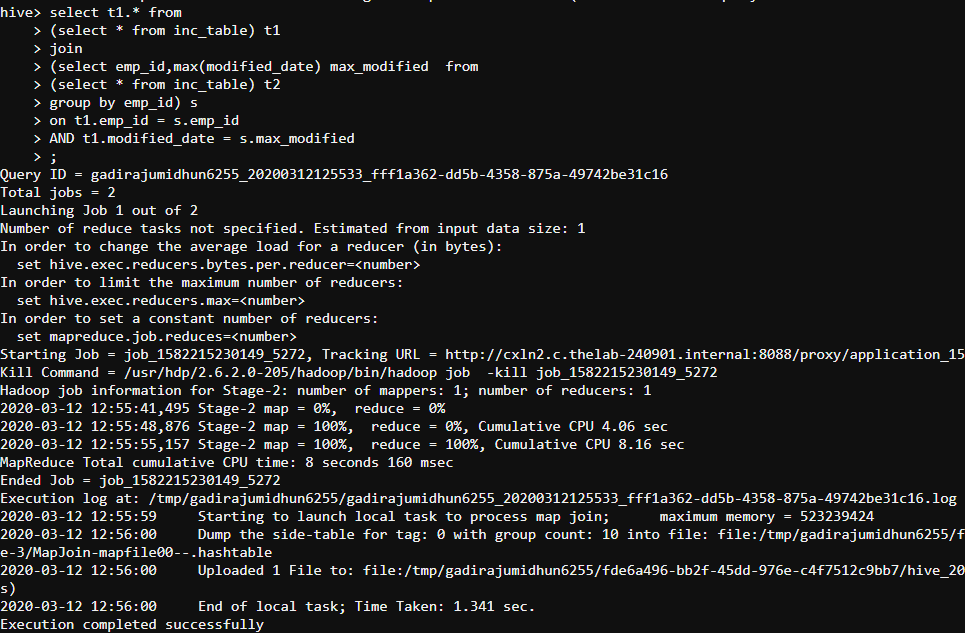
Now if we give select \* , we will get the all the day2 data along with day 1 data since we have created hive external table on top of HDFS folder.



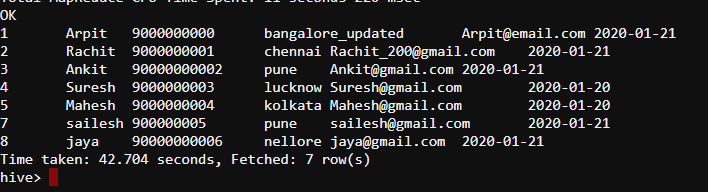
But we need only the changes to be reflected .Means only incremental data to be loaded.

We will write a query , which will only give us the original records with updated values .

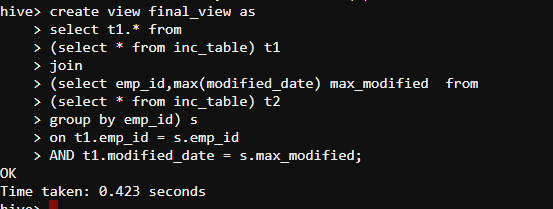




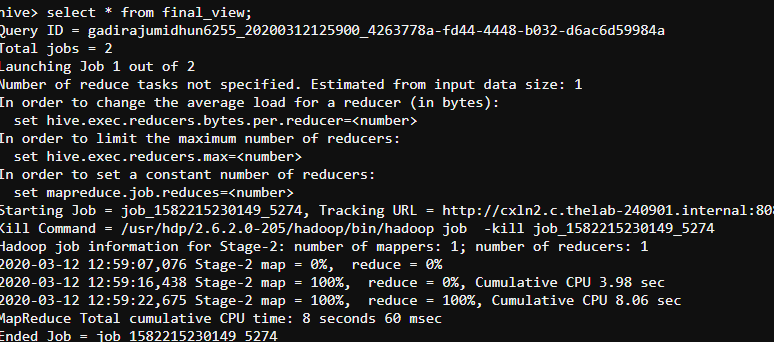
Finally the below is the required output …

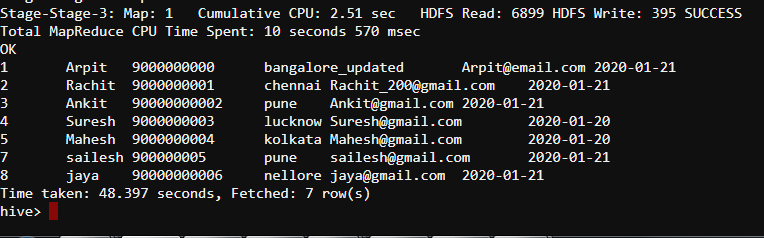


Now we are creating a view on above query …

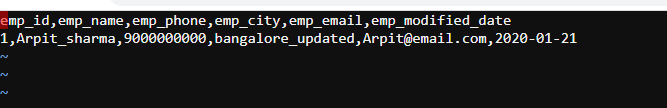


Whenever we are selecting the final\_view, we will get the incremental load output as shown below …





Now we are creating the Day 3 file …



Move to HDFS under INC\_load .

Now select the view , we will get the updated value …