Bucketed Table

1) Download the dataset in **STAGING\_AREA.**

**2)** Pull the contents in **LABS\_HOME**/demos

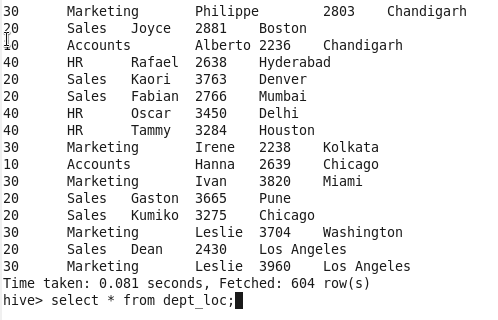
Data:- dept\_loc.txt



…



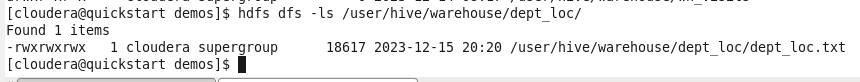
…



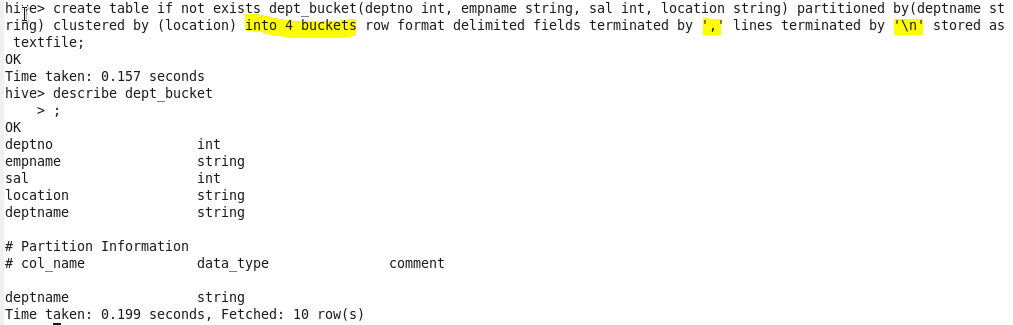
…

In cloudera:- /user/hive/warehouse/filename

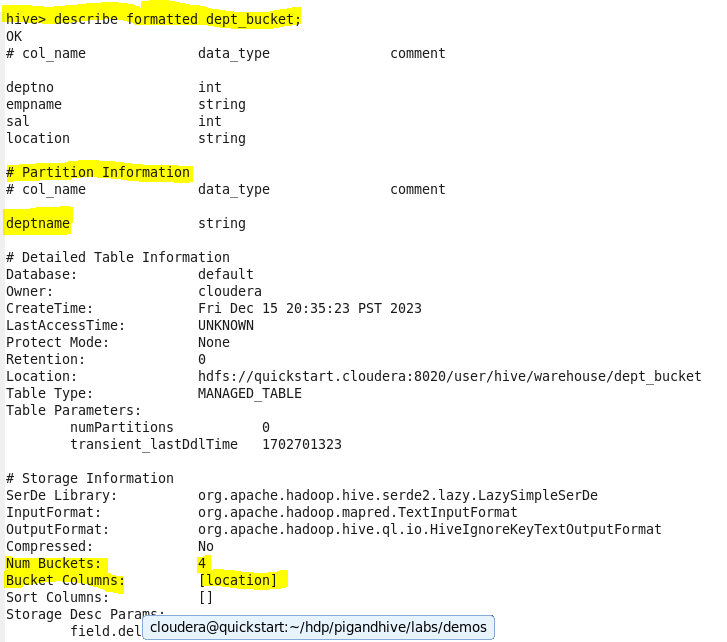
In Hive:- dfs –ls –r /user/hive/warehouse/filename



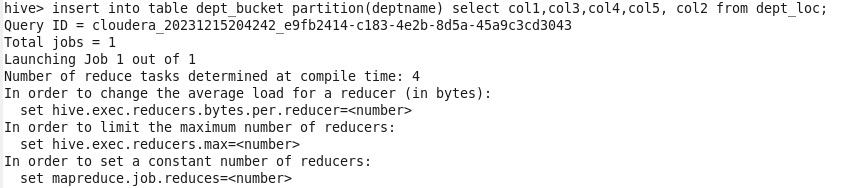
…



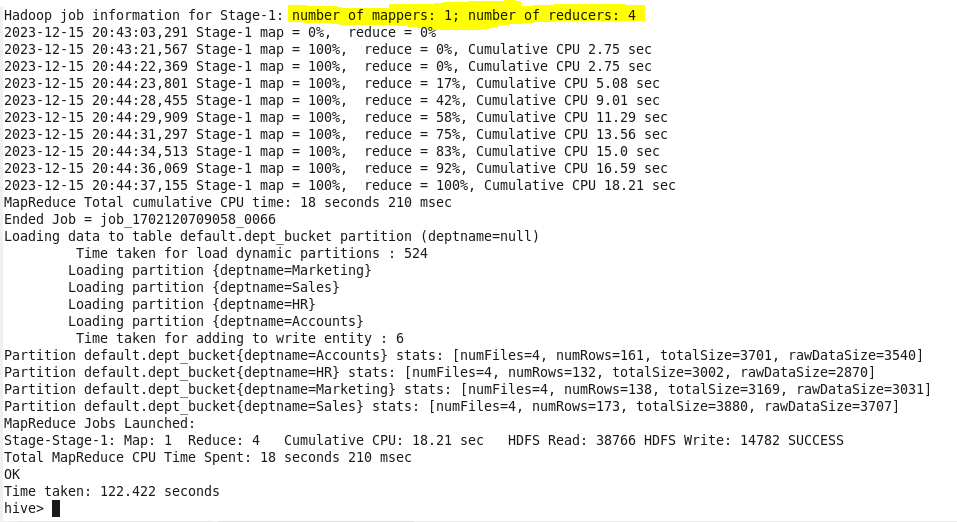
…



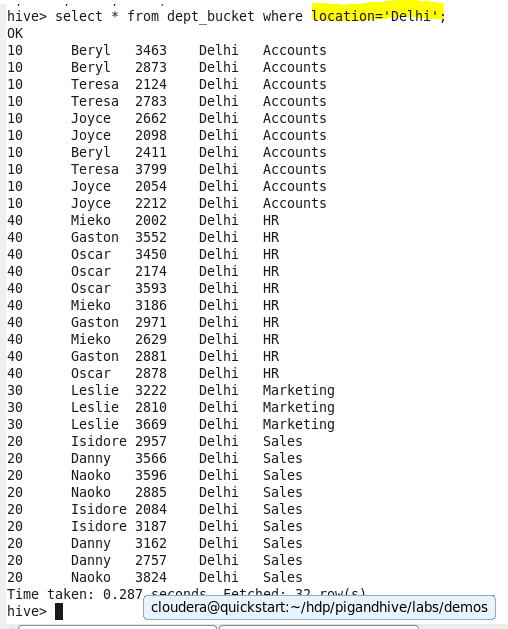
Load data alternative command :



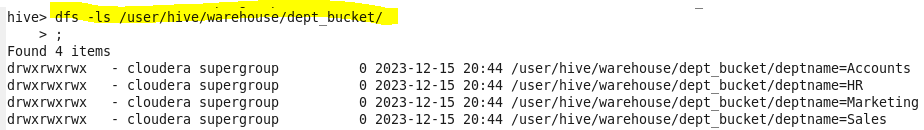
…



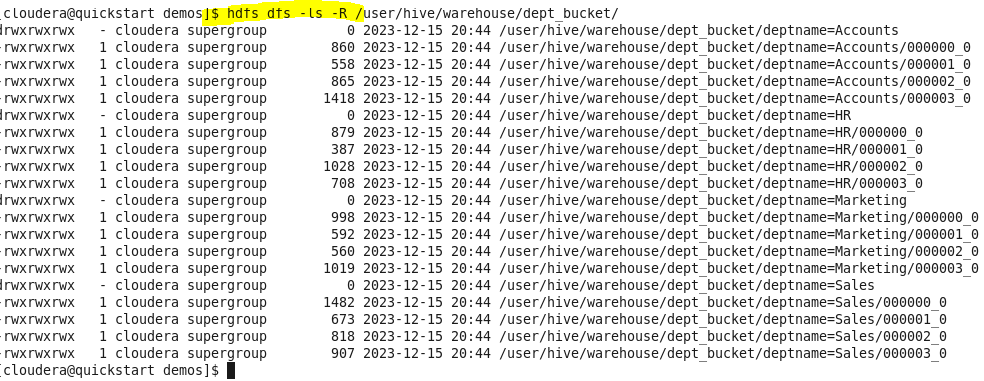
…



…



…



Cat buckets

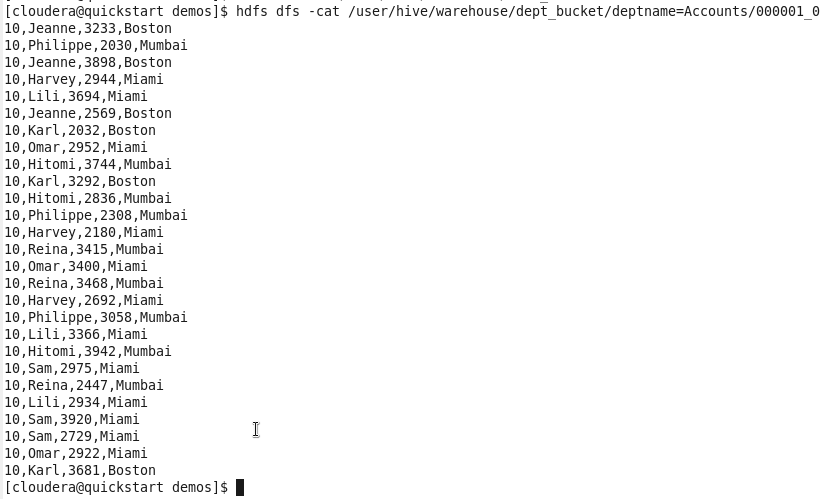
There are total 16 cities in location column, as deptname is partitioned and then total 4 buckets are created on location.

For deptname= Account

Buckets are as follows : 000000\_0, 000001\_0, 000002\_0, 000003\_0

in 1st bucket total 4 cities are there.





Automate :

LAB: ANALYZE BIG DATA WITH HIVE

Analyzing BigData with Hive

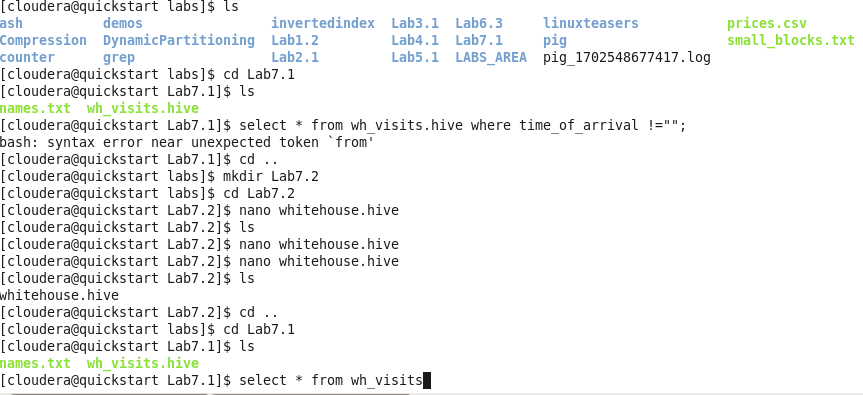
1) Refer the lab manual

2) Follow the instructions in lab manual

3) Translate the paths given in the lab manual to paths applicable for the cloudera sandbox.

**Note -** No data for this lab, the data needed is already there with you through previous labs.

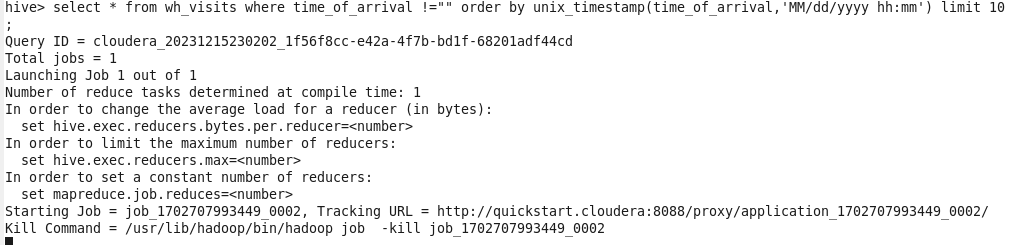
**Timeline =** 60 Minutes.



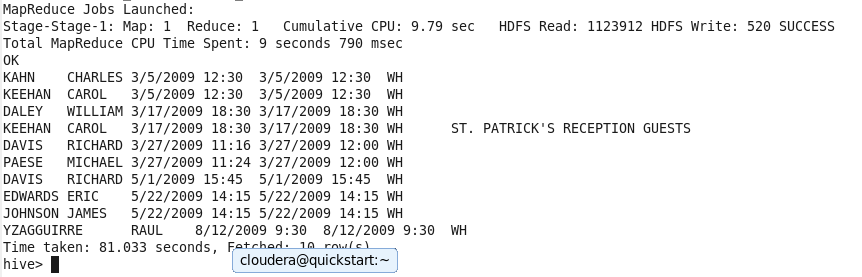
…



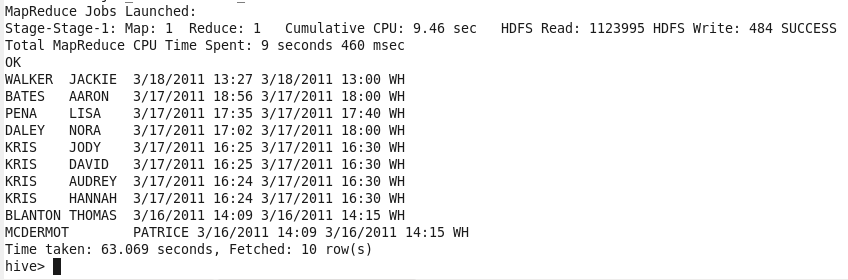
…



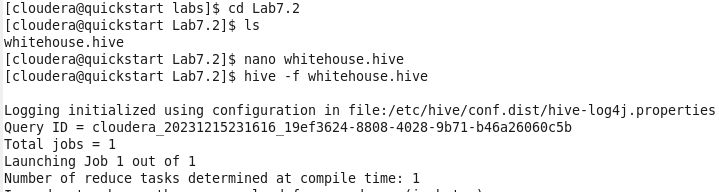
…



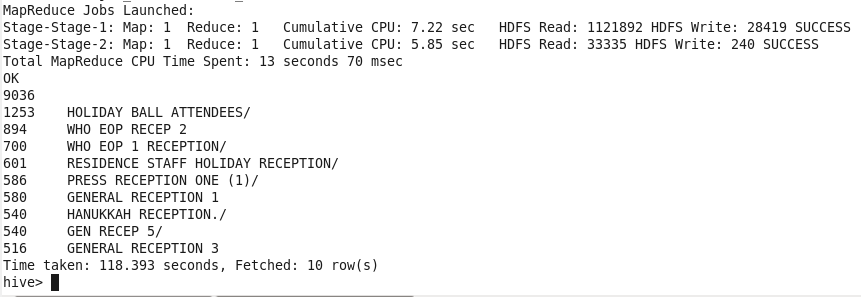
…



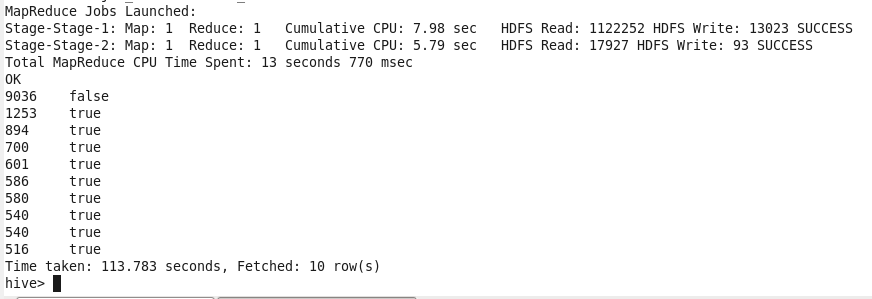
…



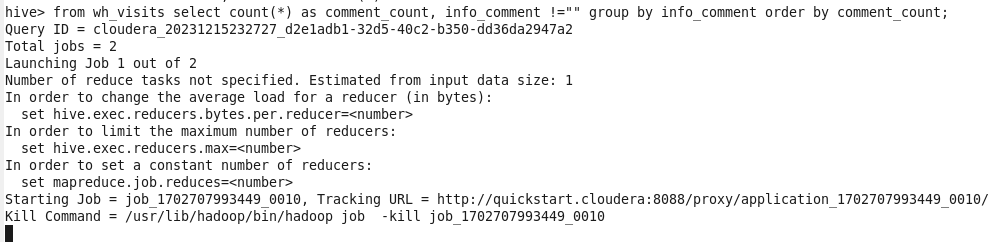
…



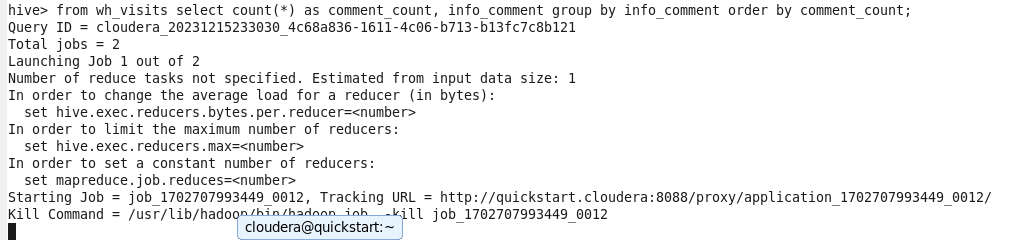
…



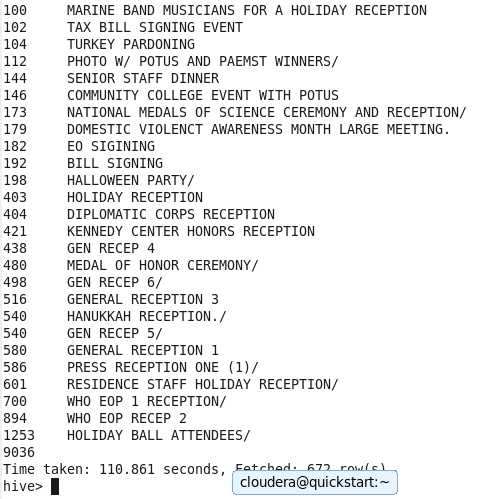
…



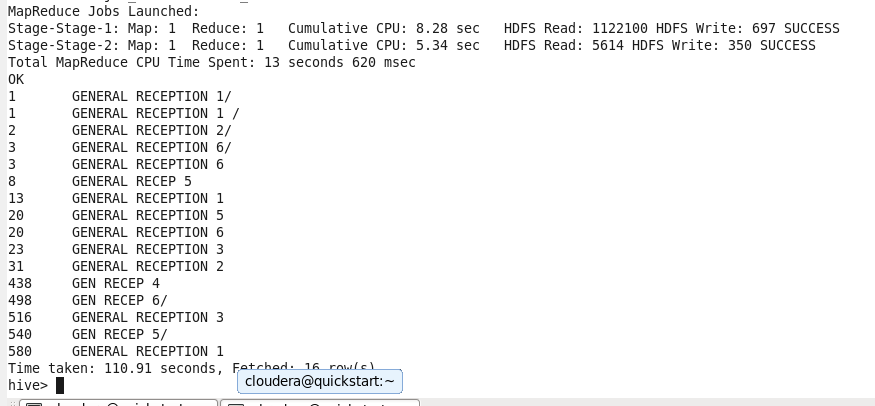
…

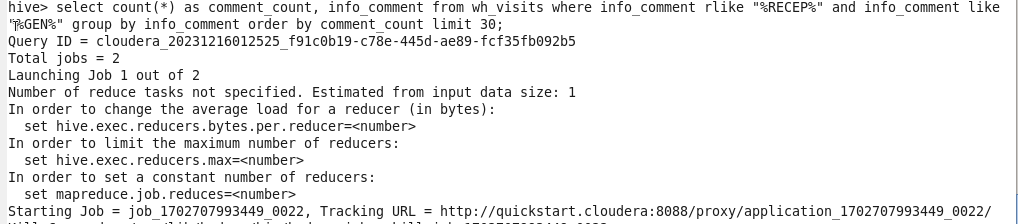


…

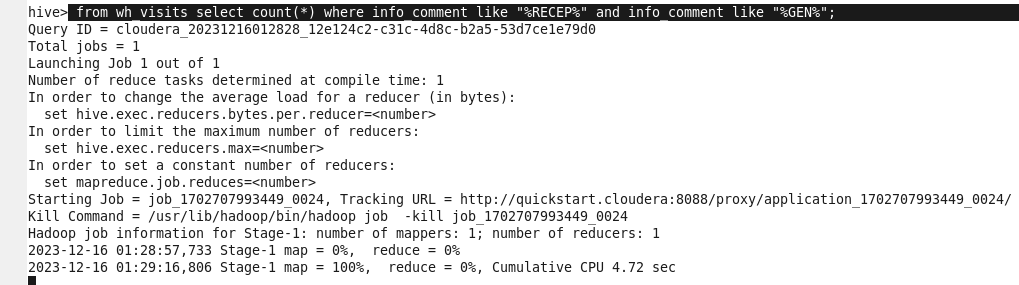


…

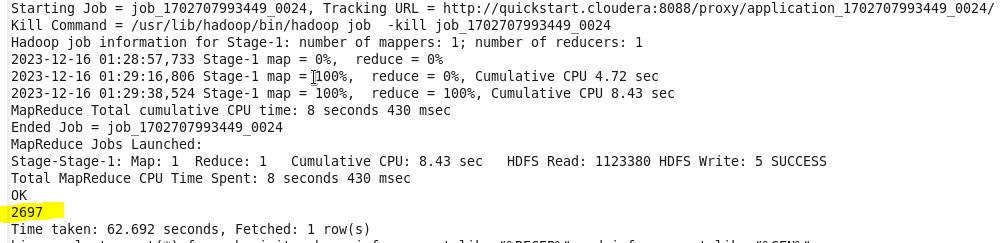




**…**

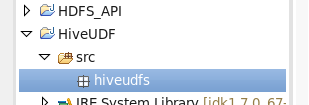


**…**

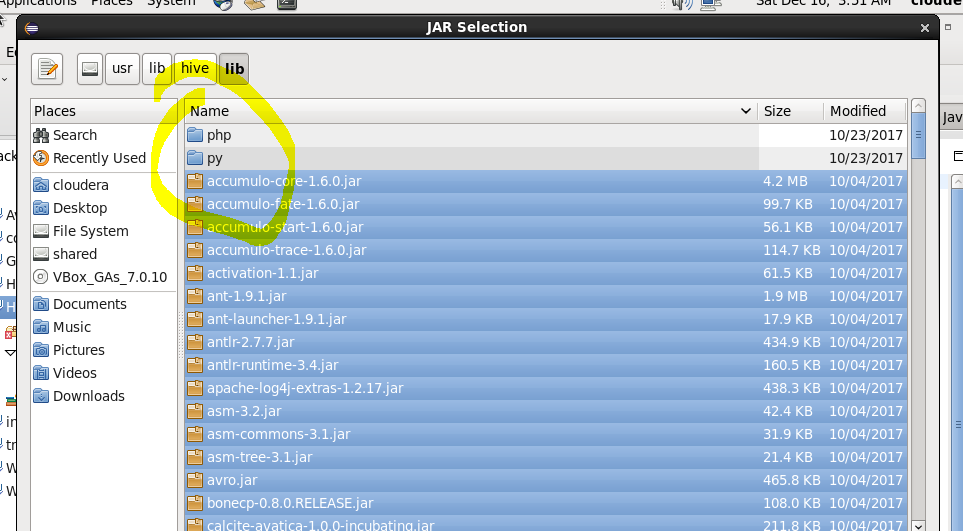


LAB: ANALYZE BIG DATA WITH HIVE



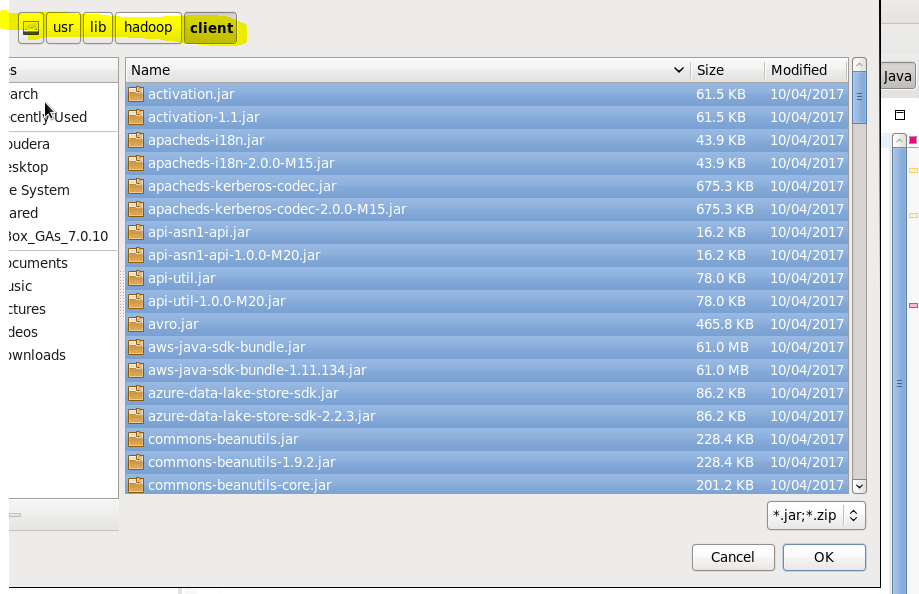


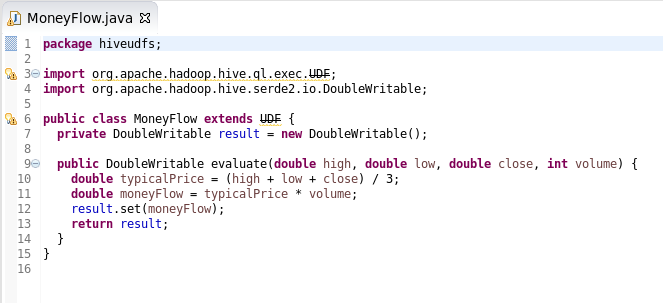
Path for packages



Don’t select folders as shown above..

Also import Hadoop jarfiles too

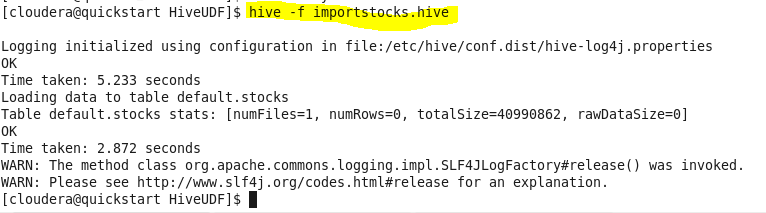




Create jar file and place into folder where rest files are present



Edit moneyflow.hive------- limit 100



… run below file too

