4] Importing data using sqoop

1) Download the attached files in **STAGING\_AREA**

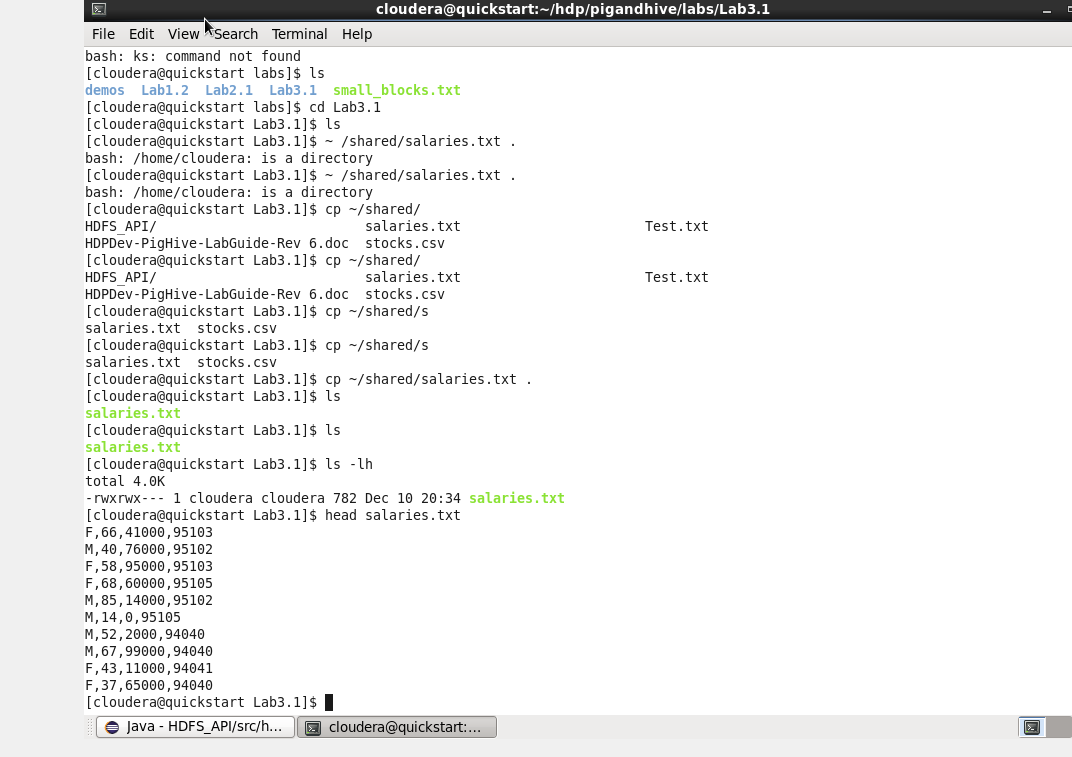
2) Create a folder Lab3.1 in **LABS\_HOME**

3) Copy the files in step 1 to step 2 location

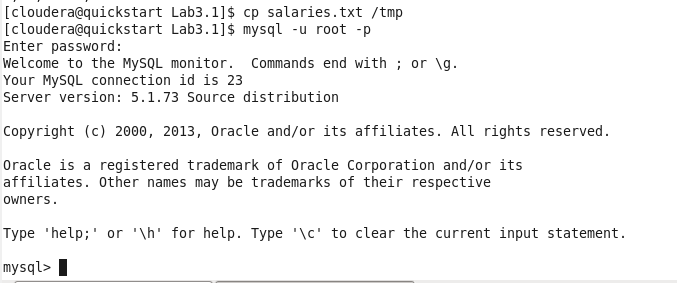
4) Wait for the instructions from the trainer.

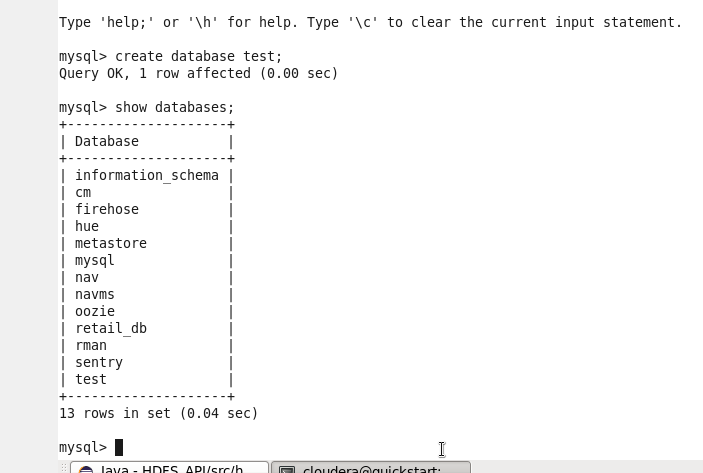
**Dataset -**

* data/salaries.txt

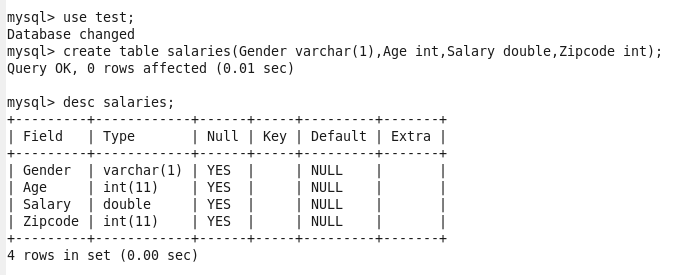


Cccc



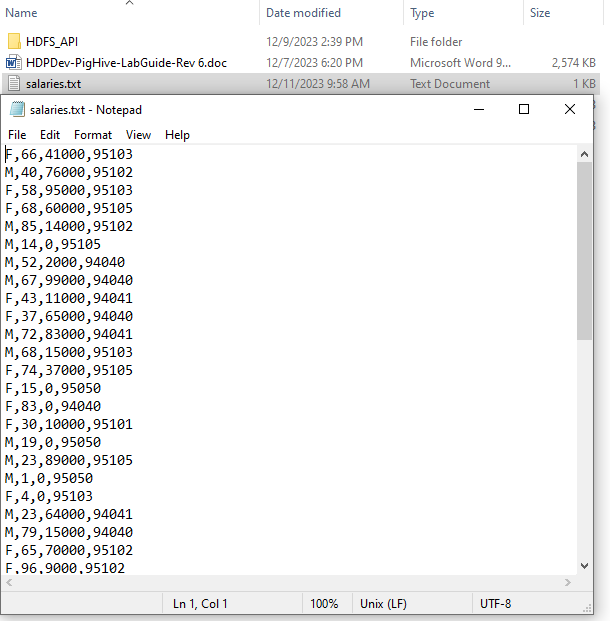


Cccc



Cccc

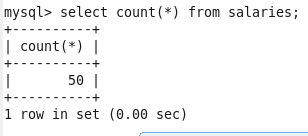
We have salaries.txt loaded in hdfs, so by giving below command we are loading data from salaries.txt to our table



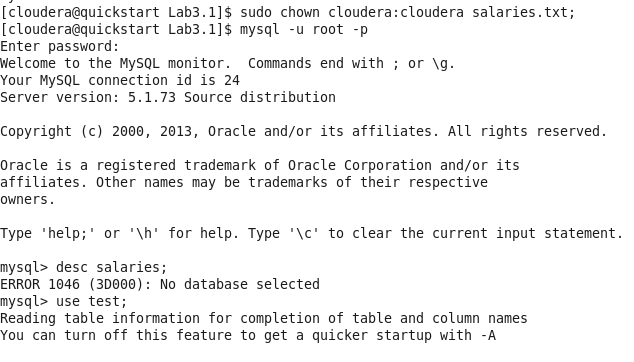
Cccc



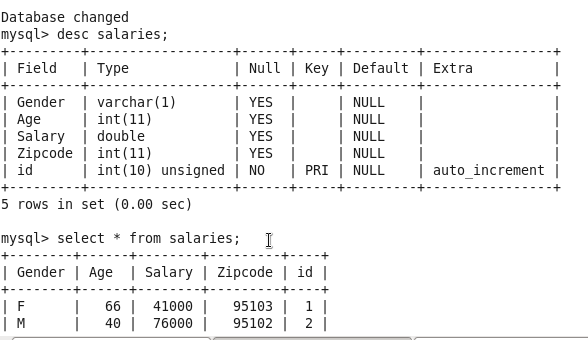
Cccc



Cccc



Cccc

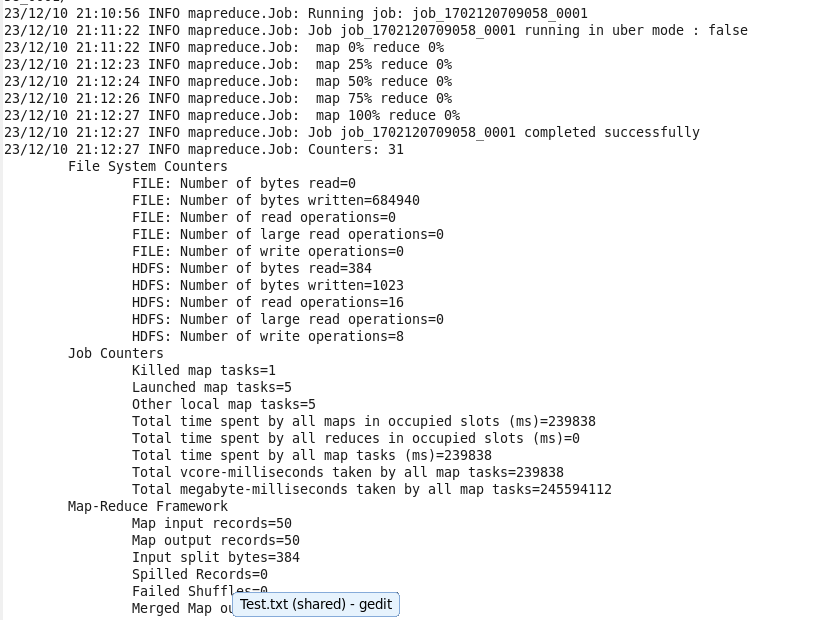


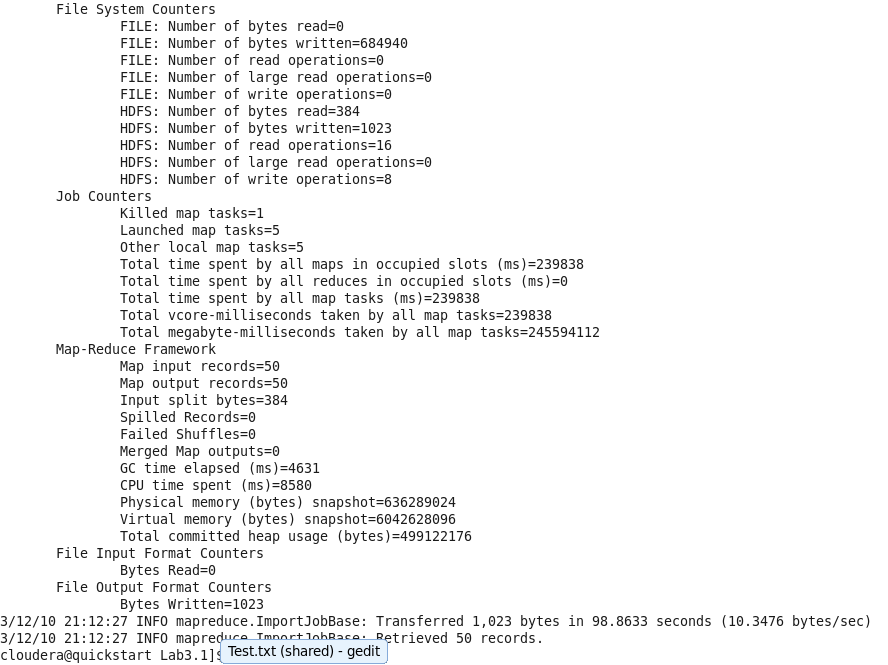
Cccc

Sqoop commands

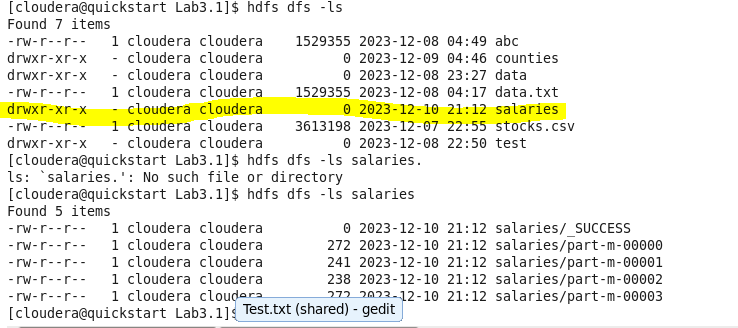
1) Import the Table into HDFS

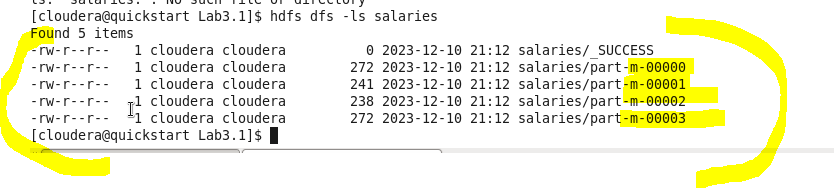
sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/test --driver com.mysql.jdbc.Driver --username root -password cloudera --table salaries





Cccc

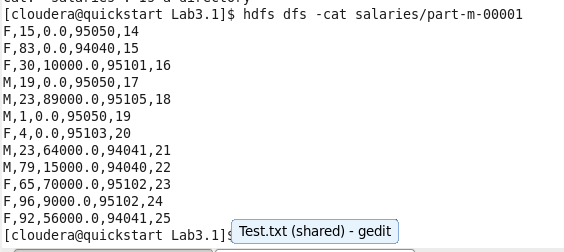




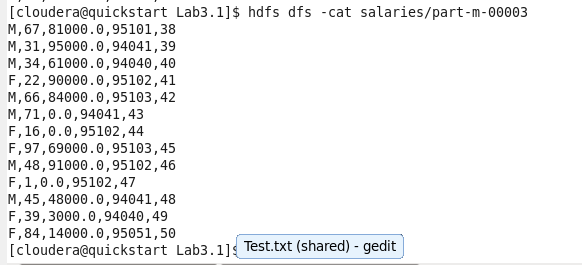
Here mappers are generated as shown above- 4 mappers are created and the records are distributed accordingly.

Look at the sizes, you will see almost equal distribution of record

1st mapper- 272 byte of data, 2nd mapper- 241, etc

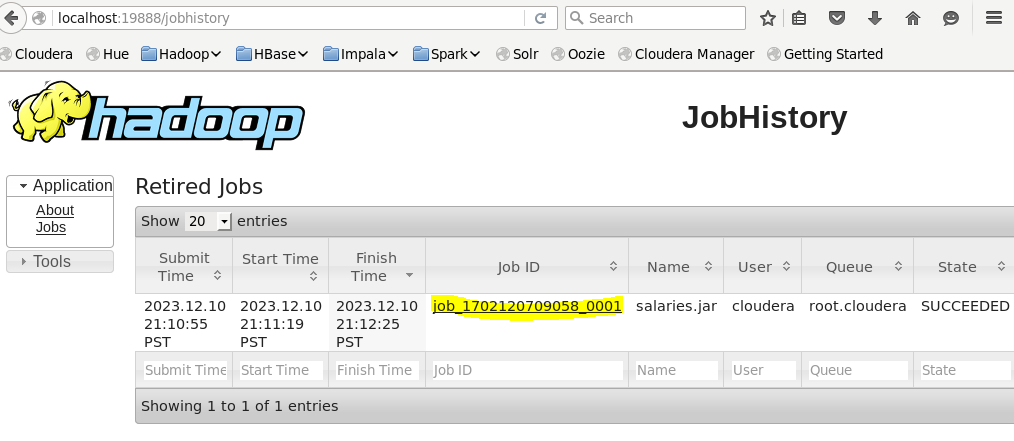


The 12 entries from 2nd mapper as we can see (i.e. 00001)



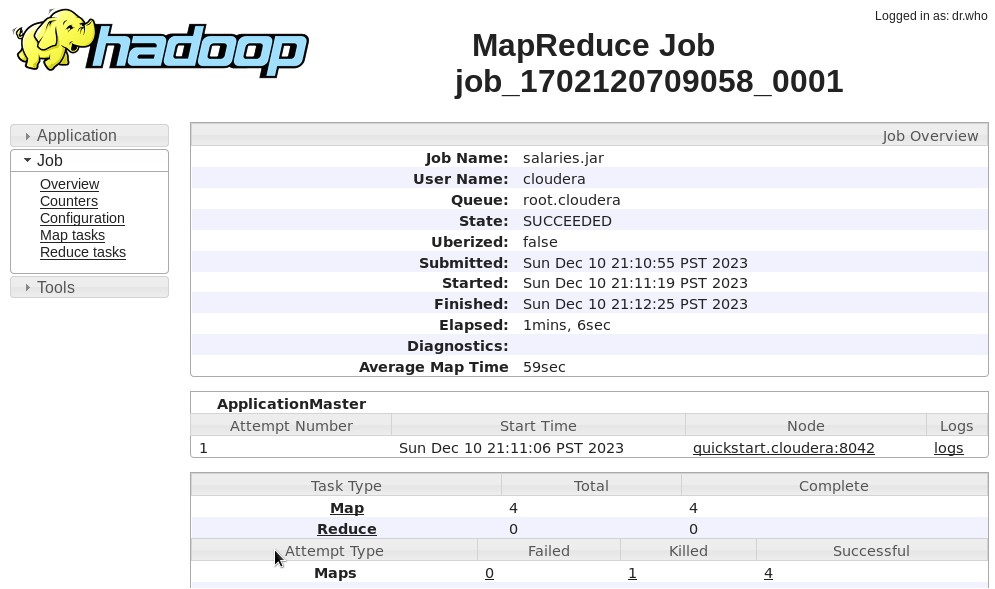
-m-00003

We cannot get the schemas means table (column name, etc) here because it is Metadata, Can get by using HIVE.

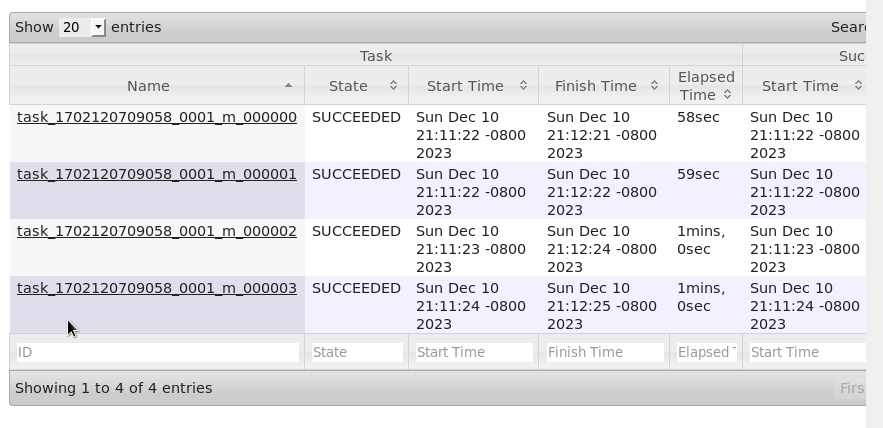


Job ID : same for both.

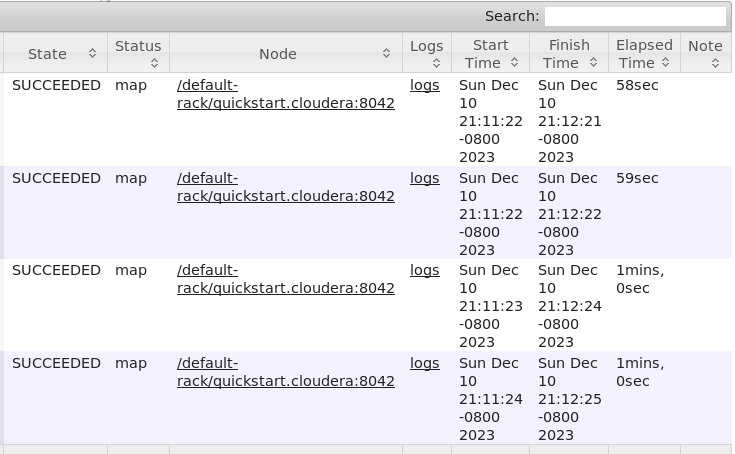




Click on Map:



Cccc

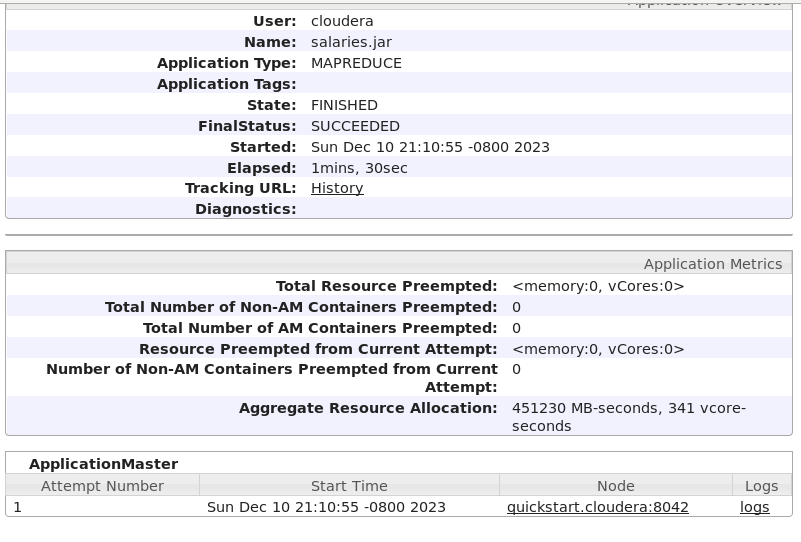


Click on Logs:

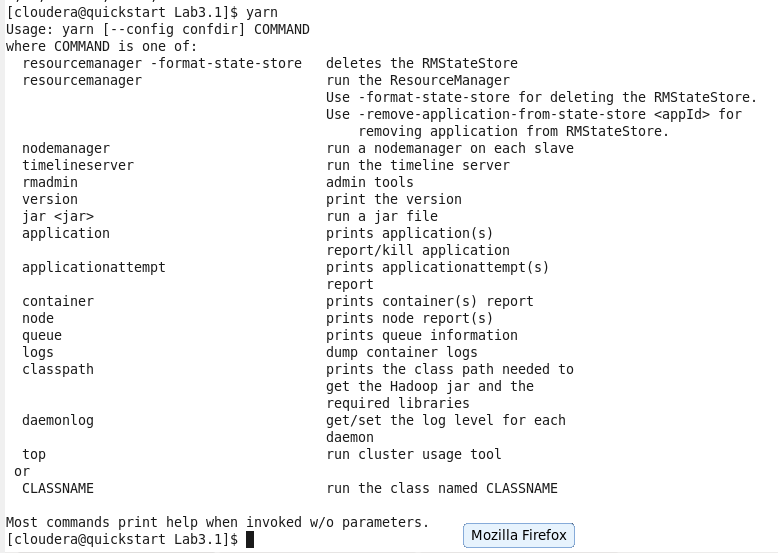
Shared data across 4 map process accessing same data



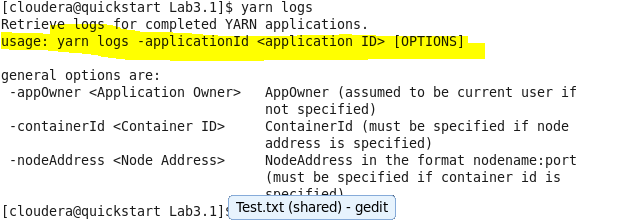
Cccc



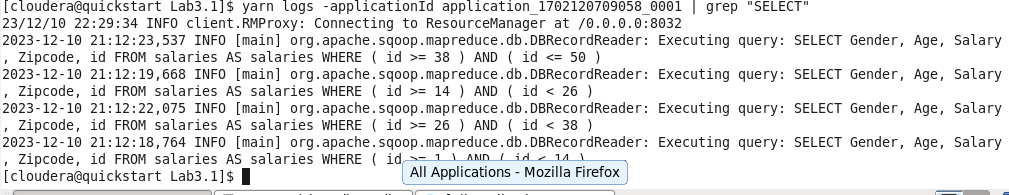
Yarn



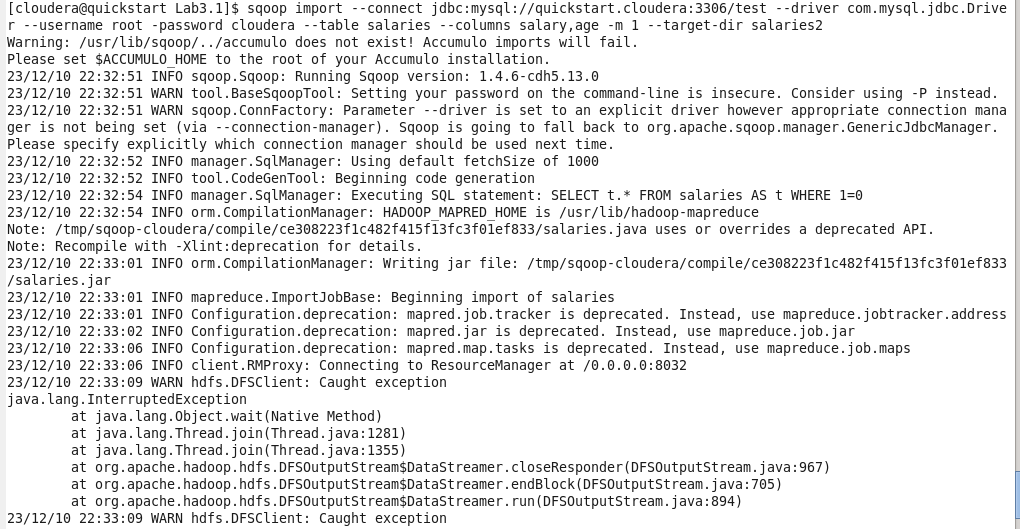
Logs : container



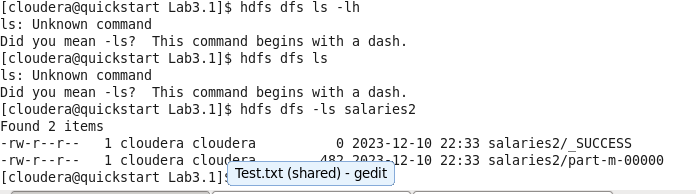
Cccc



2) sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/test --driver com.mysql.jdbc.Driver --username root -password cloudera --table salaries --columns salary,age -m 1 --target-dir salaries2



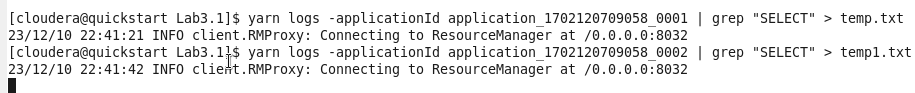
Cccc



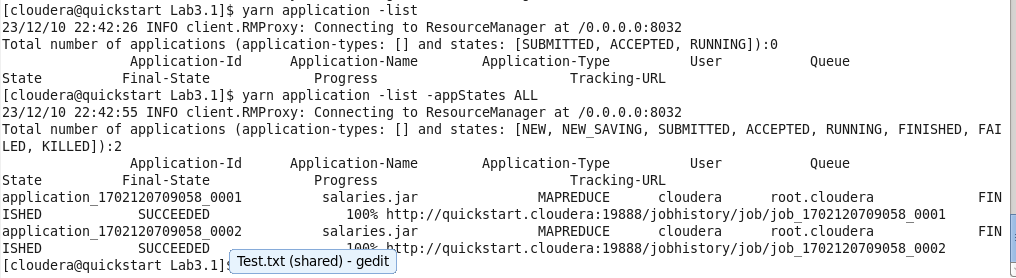
Ap id :02



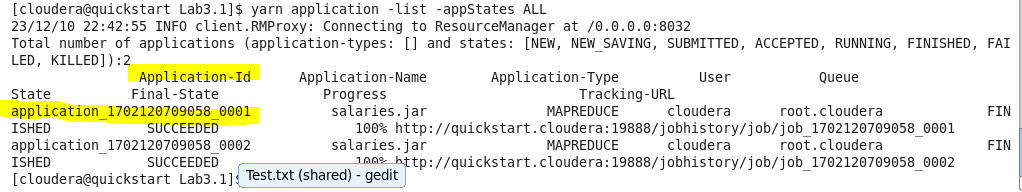
Cccc



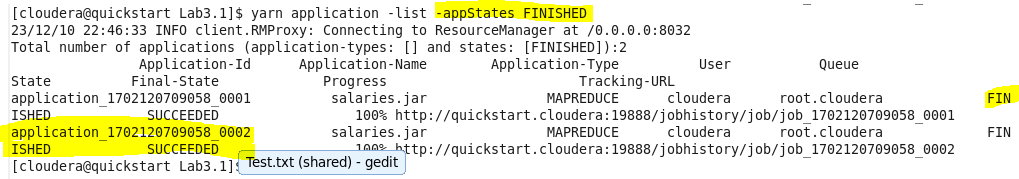
Cccc



To see the ApplicationID

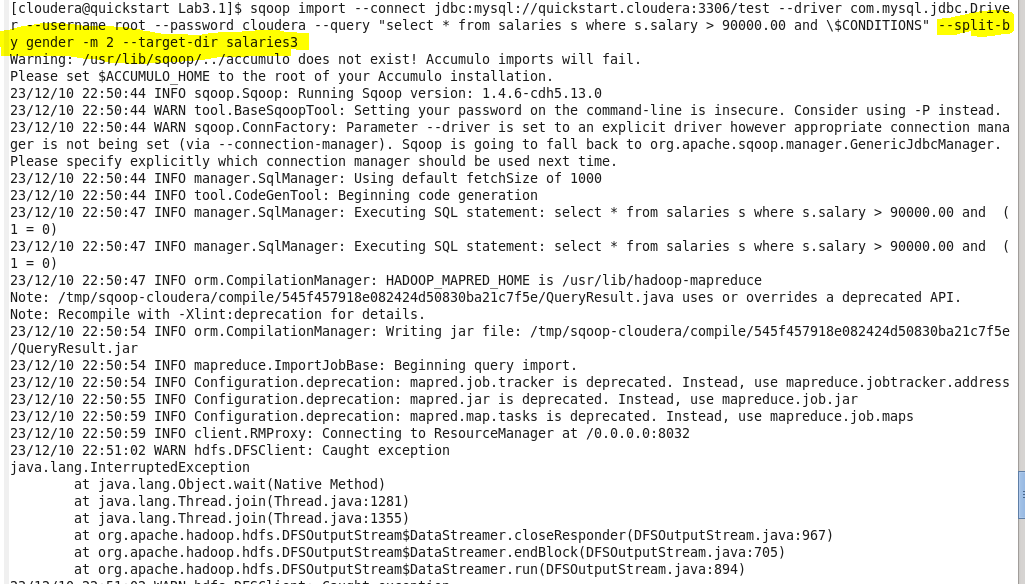


Cccc

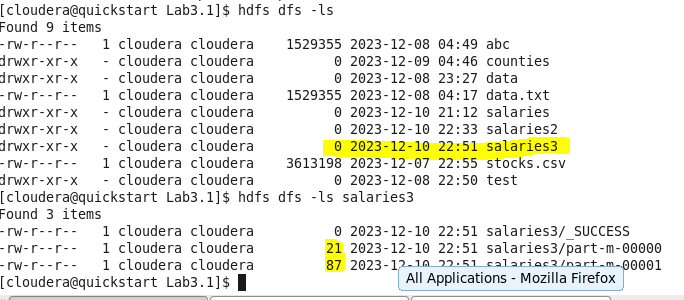


3) Importing from a Query

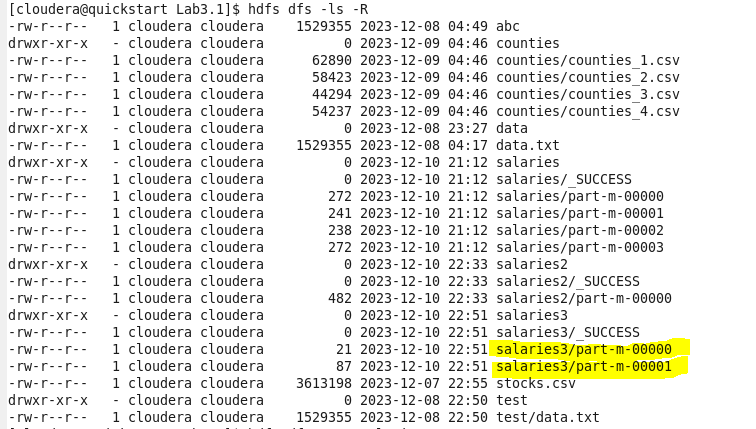
sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/test --driver com.mysql.jdbc.Driver --username root --password cloudera --query "select \* from salaries s where s.salary > 90000.00 and \$CONDITIONS" --split-by gender -m 2 --target-dir salaries3



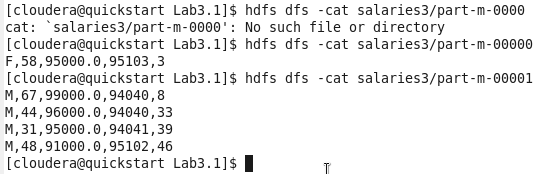
Cccc



Cccc

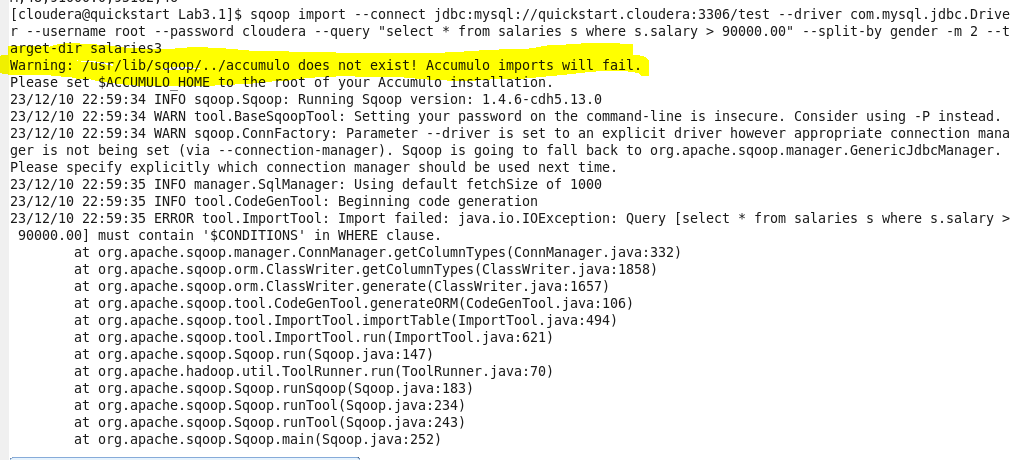


Cccc



Cccc

If we remove and \$CONDITIONS so our query will fail similarly –splitby condition is also necessary.



Here we removed split by condition, so it will fail again

