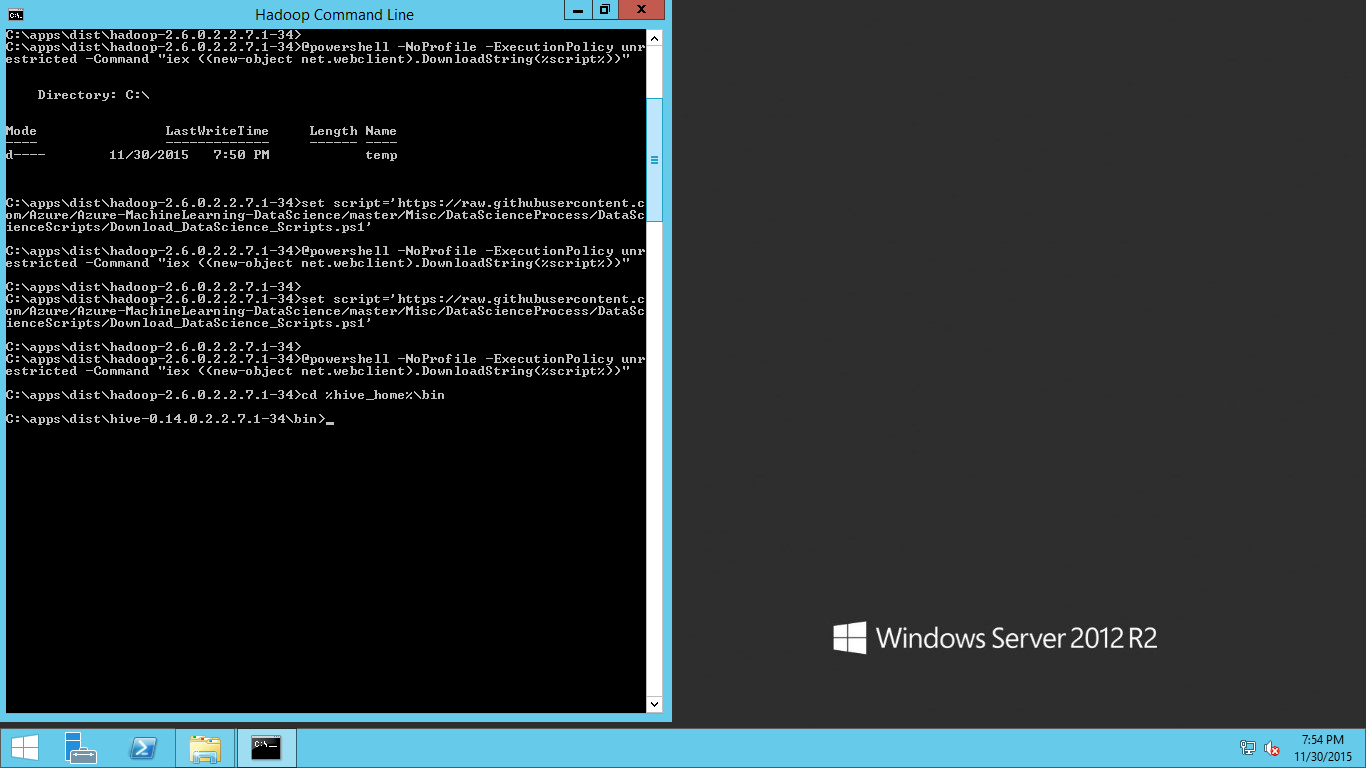
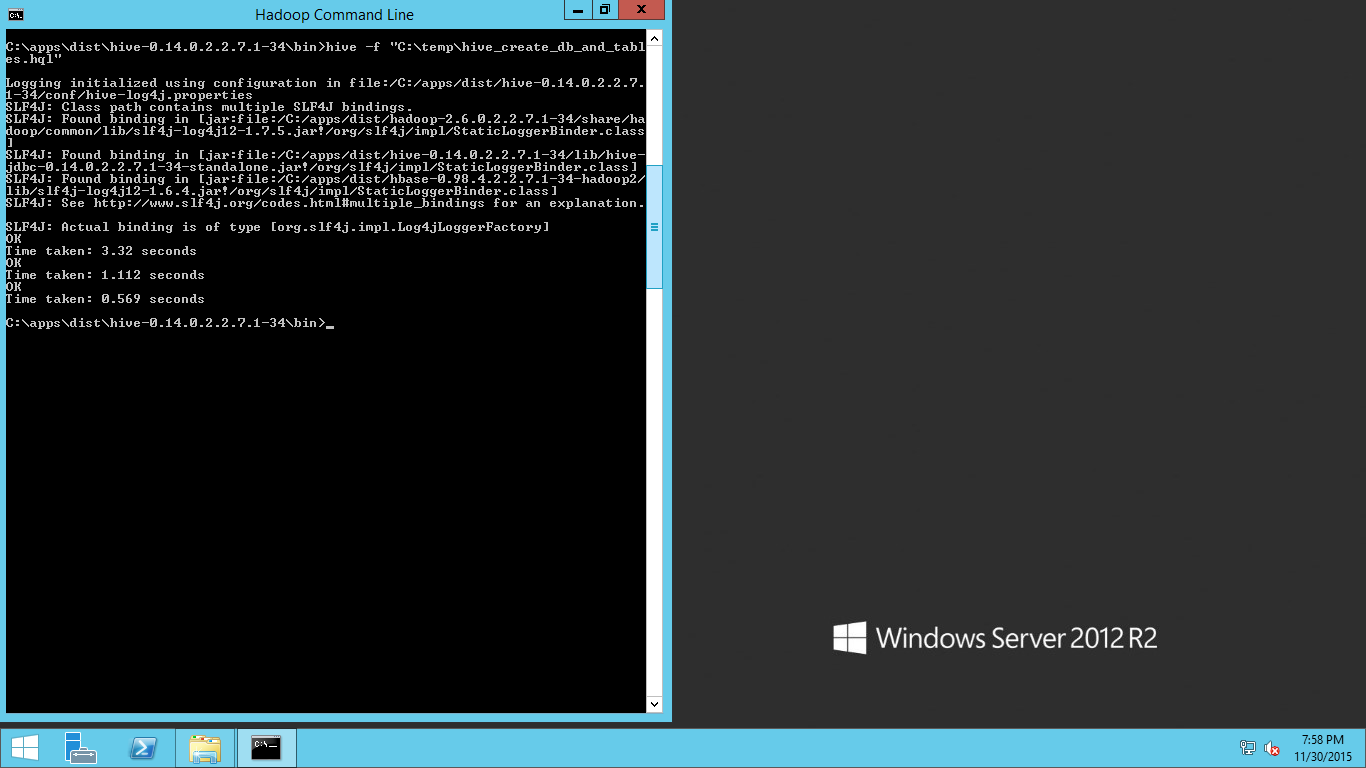
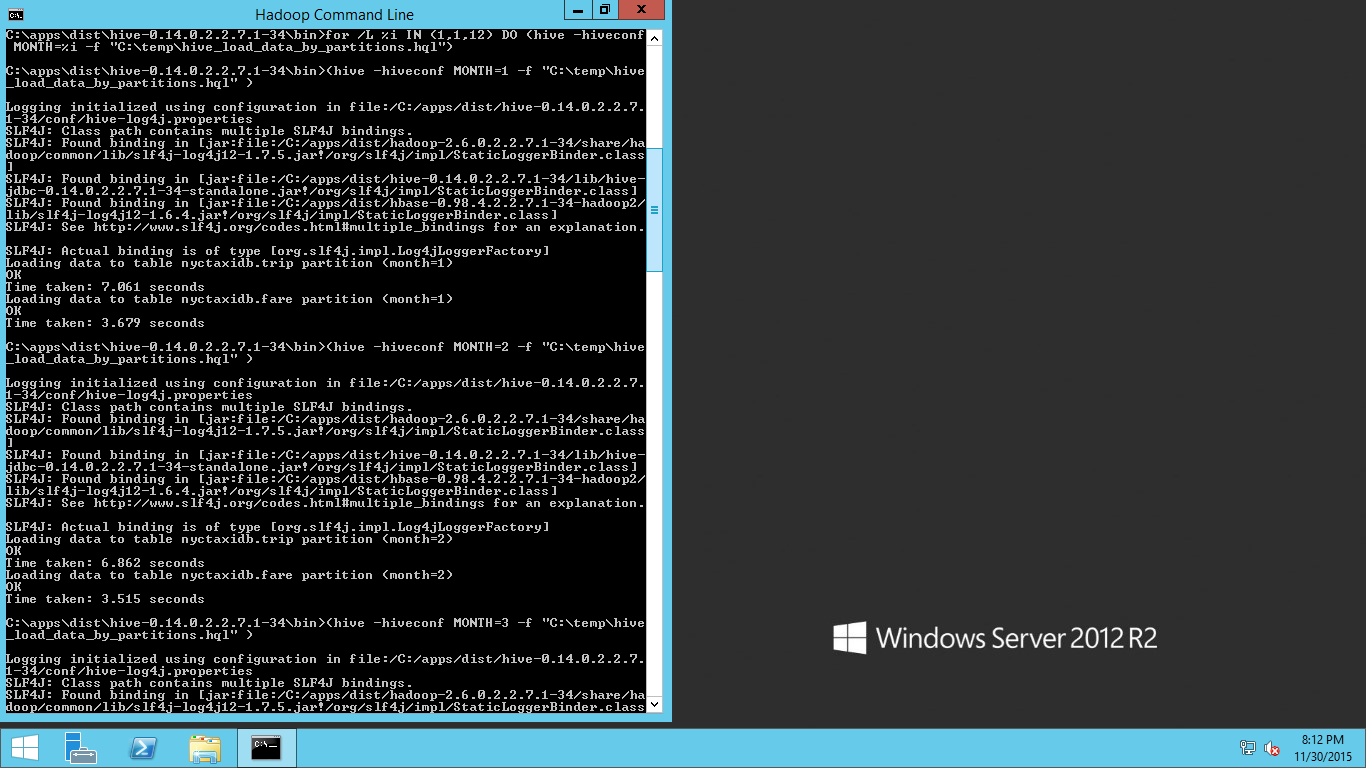
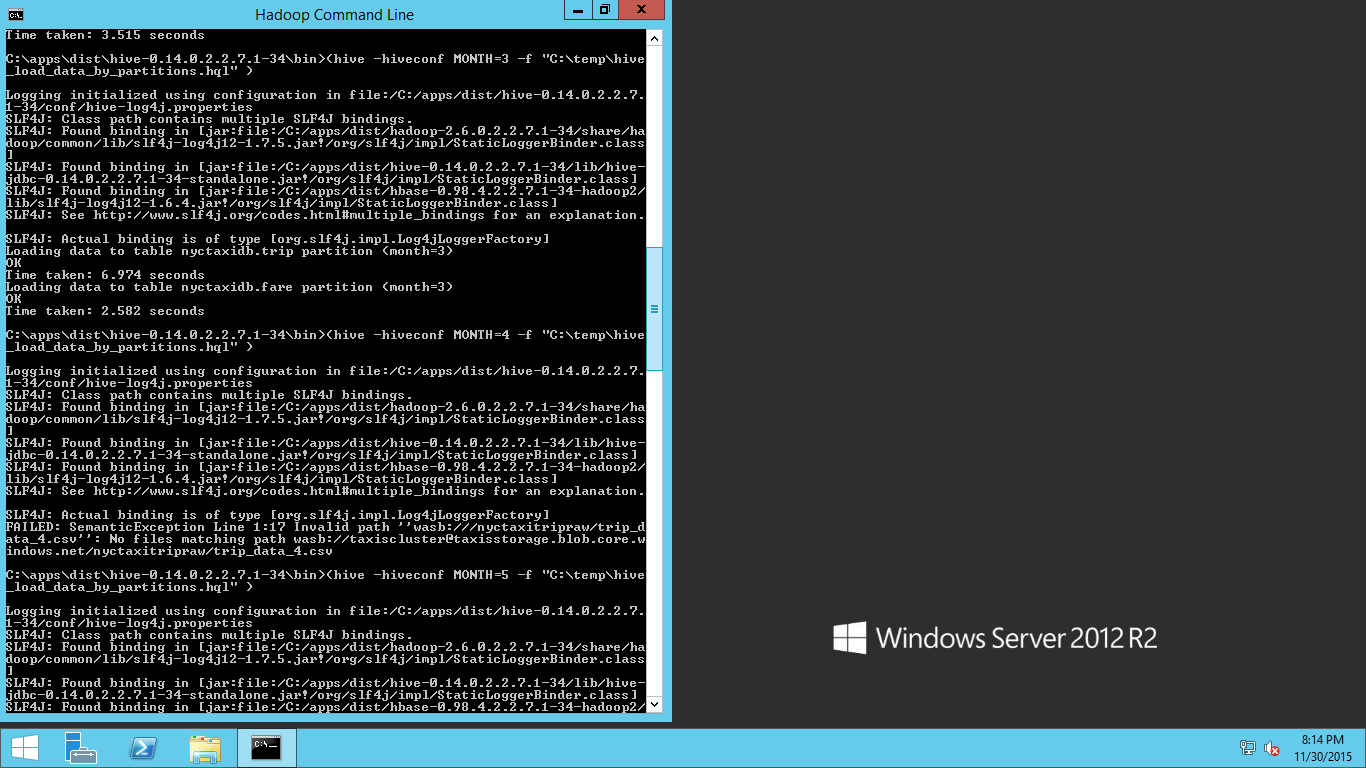


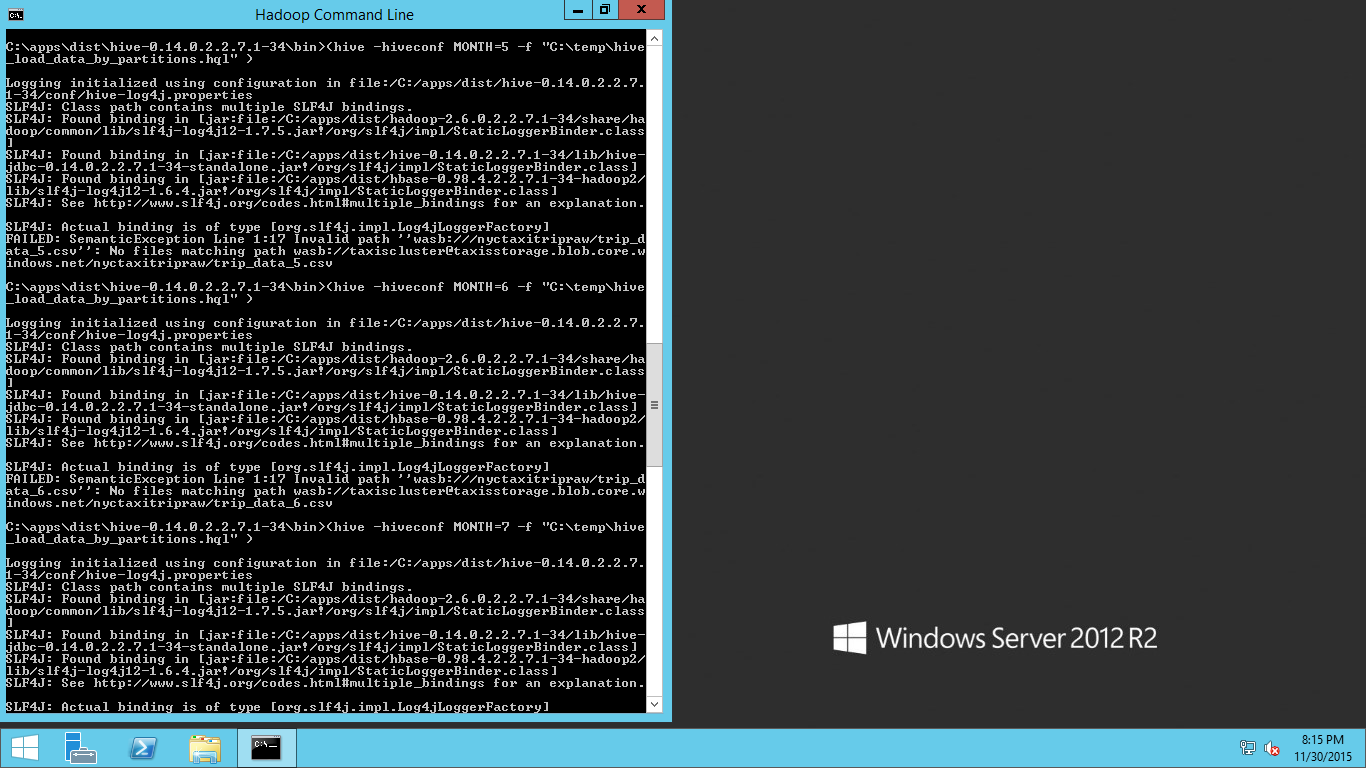
Remote desktop->connect->will download file and use it.

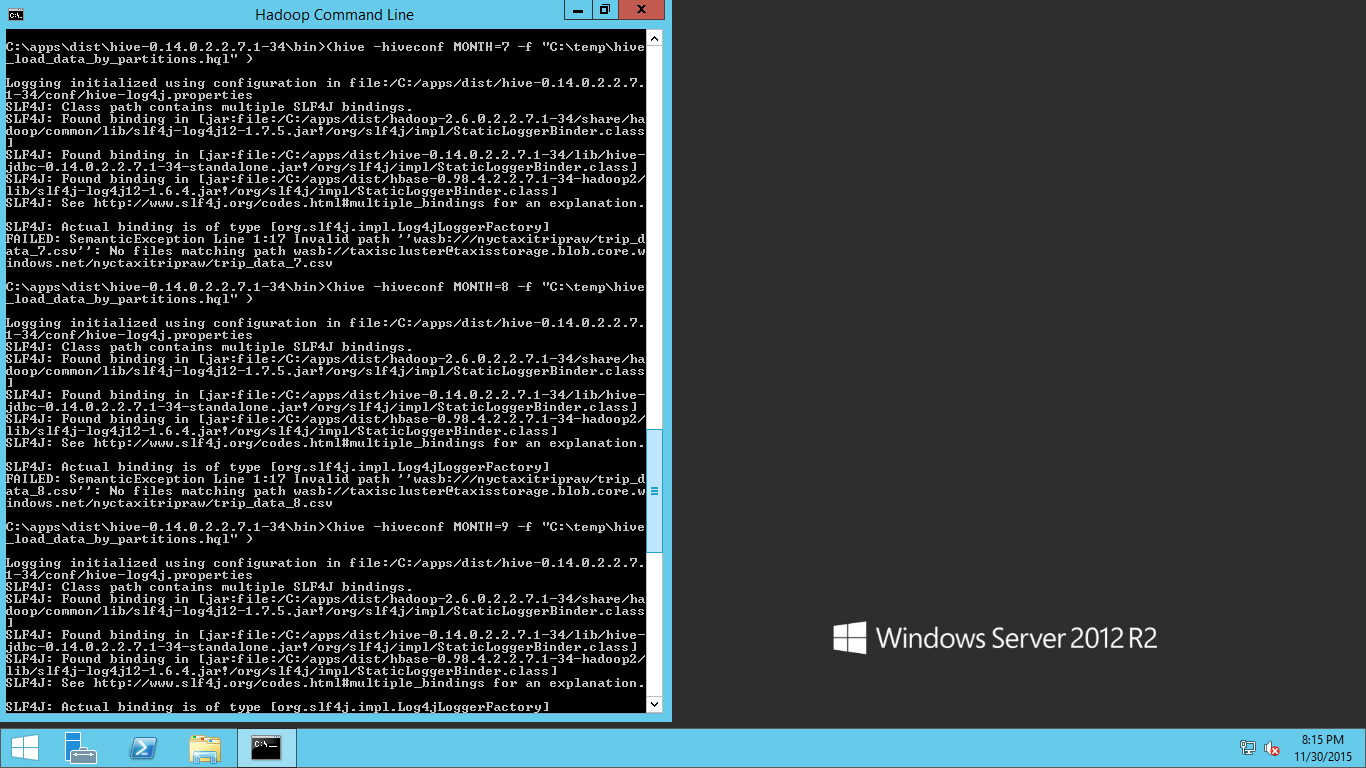


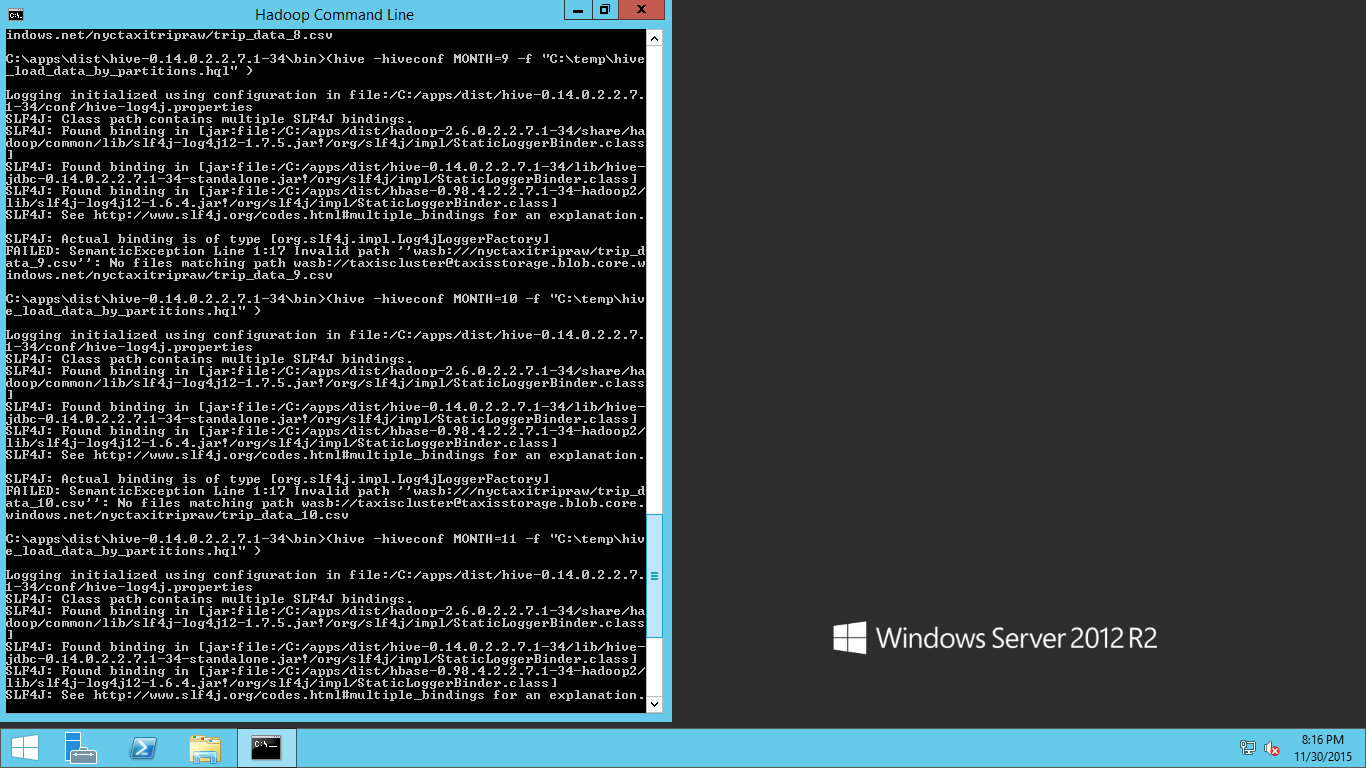


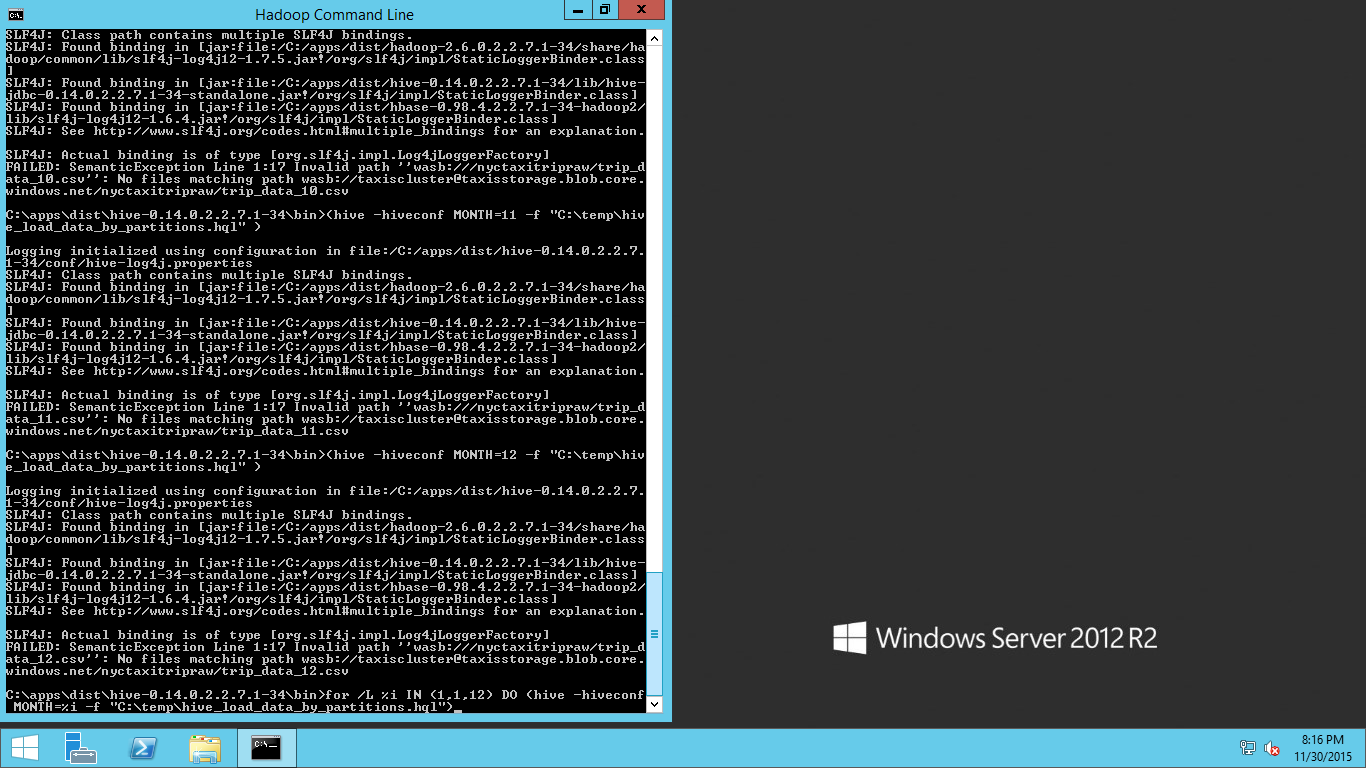


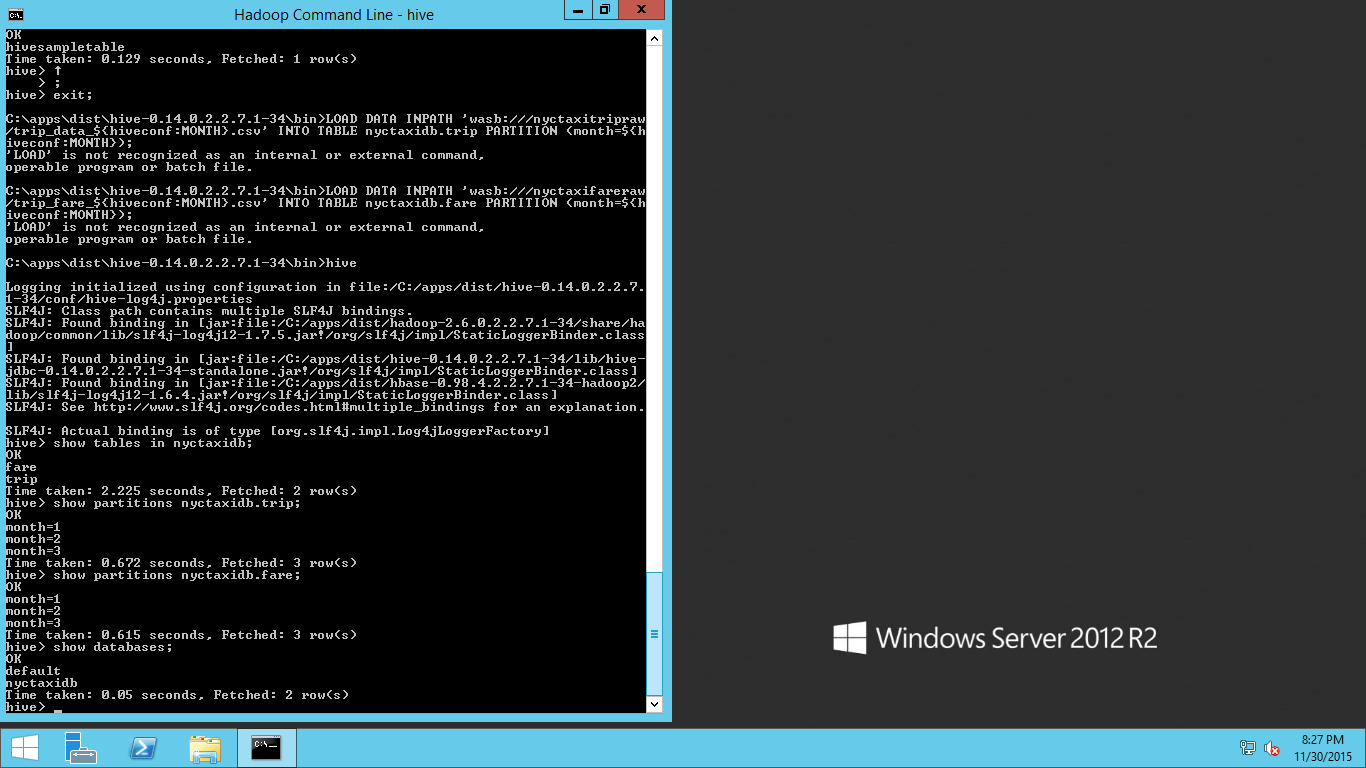




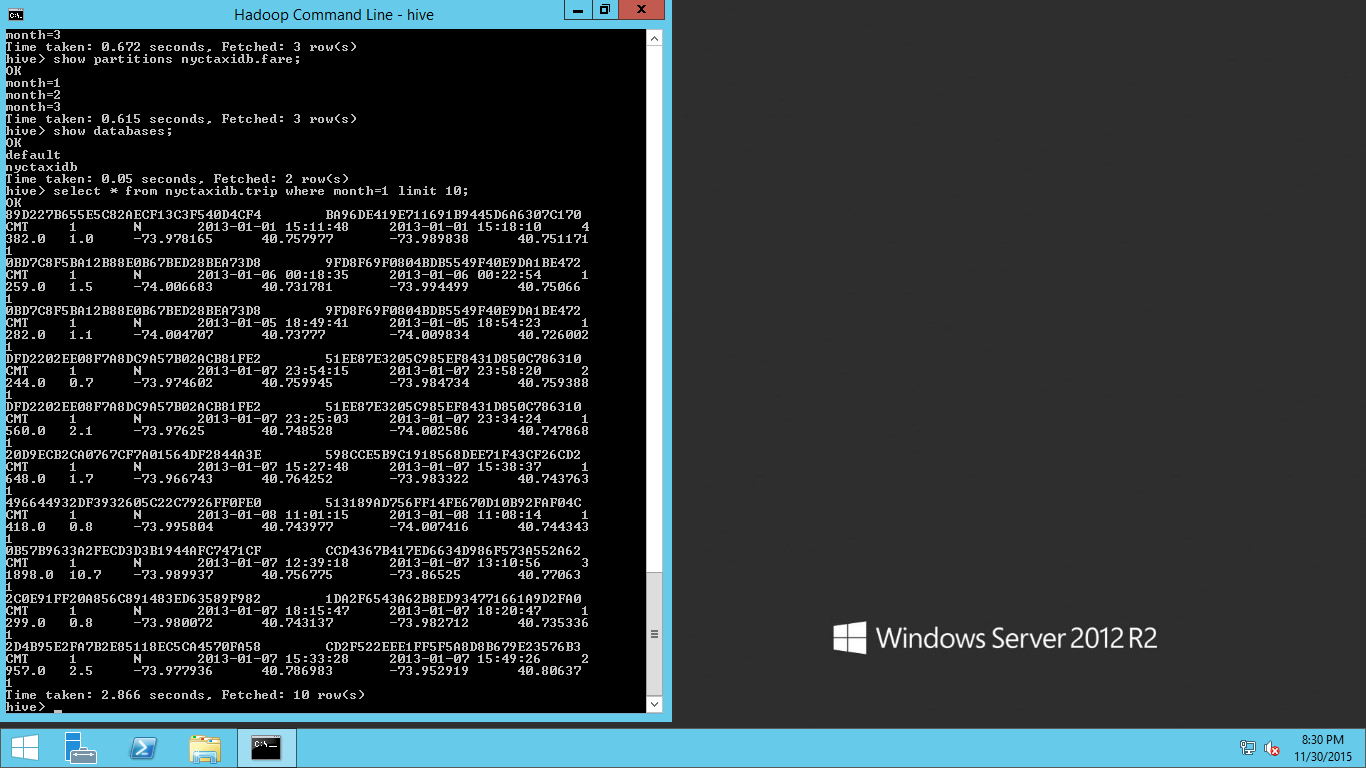




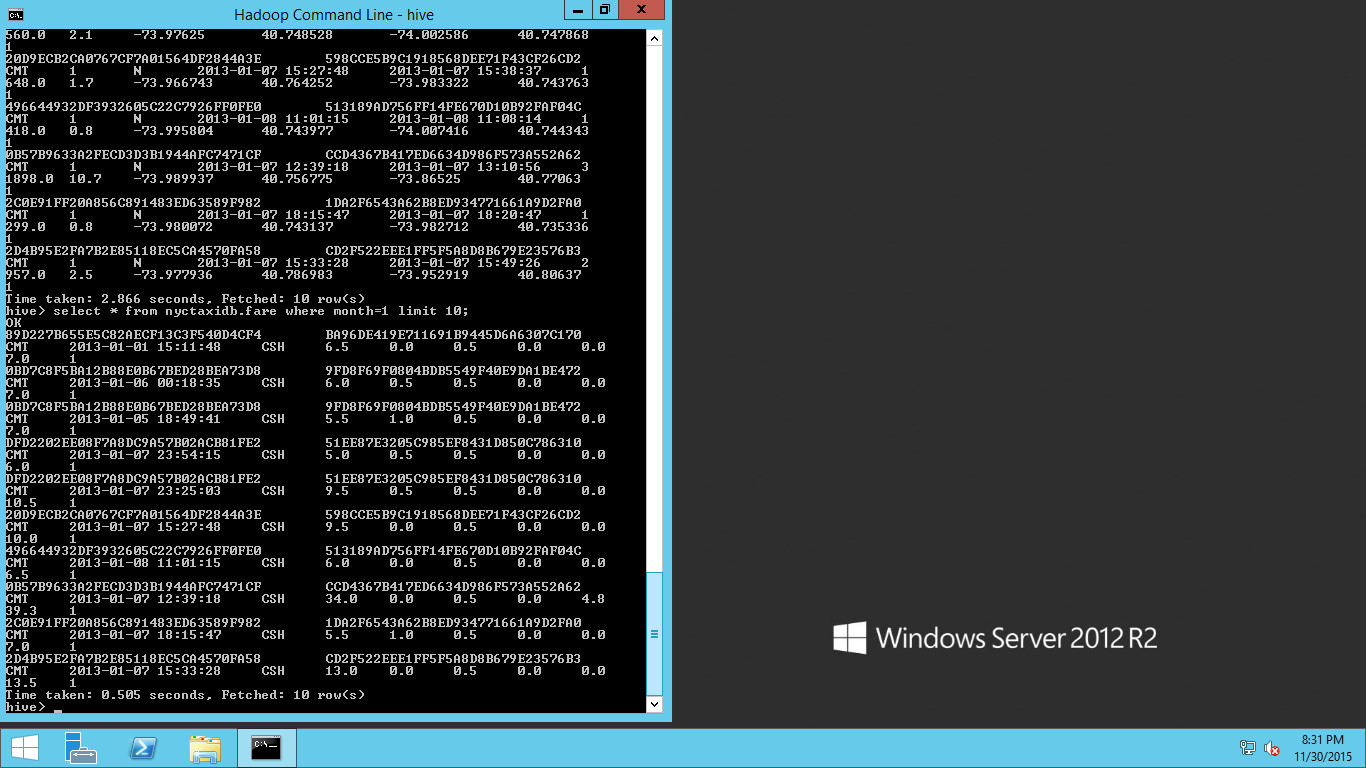




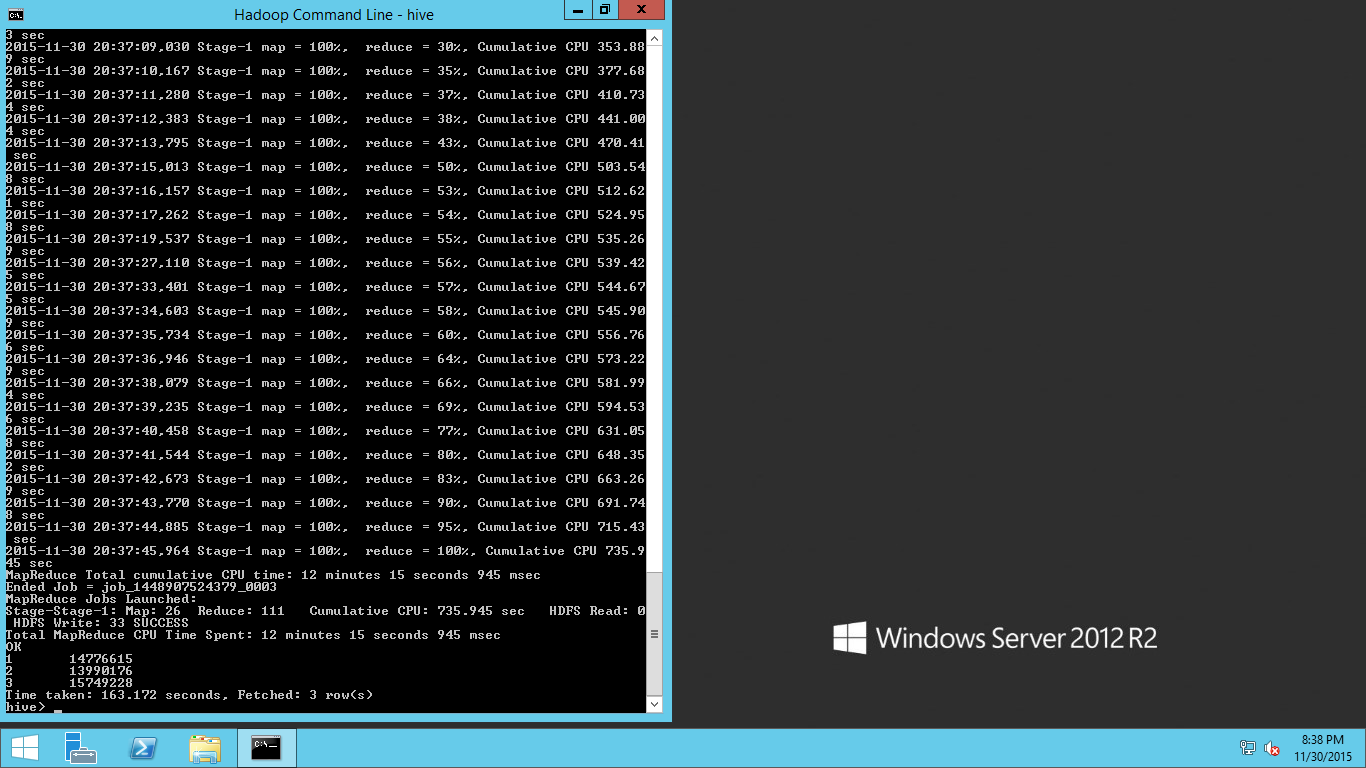
For trip 10 records month 1



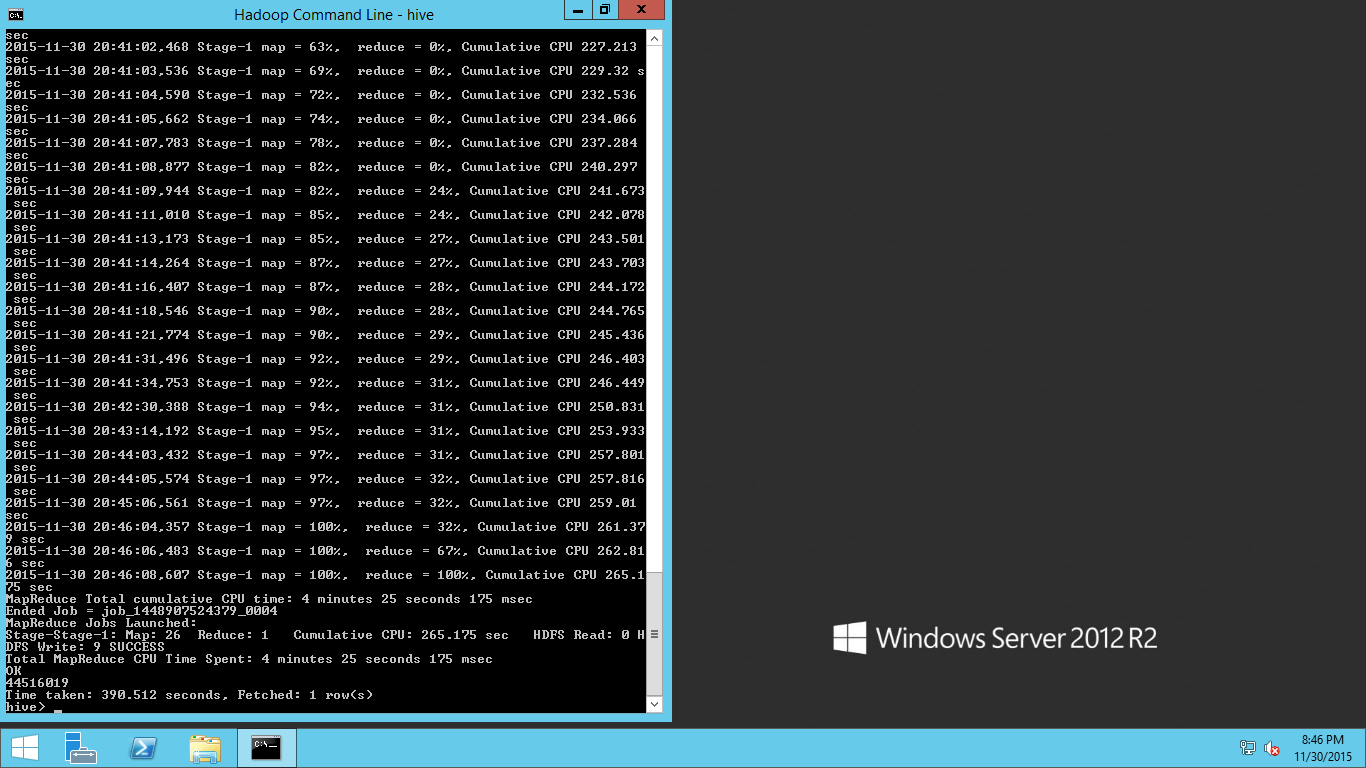
For month 1 fare table 10 records



**Exploration: View the number of records in each of the 12 partitions**

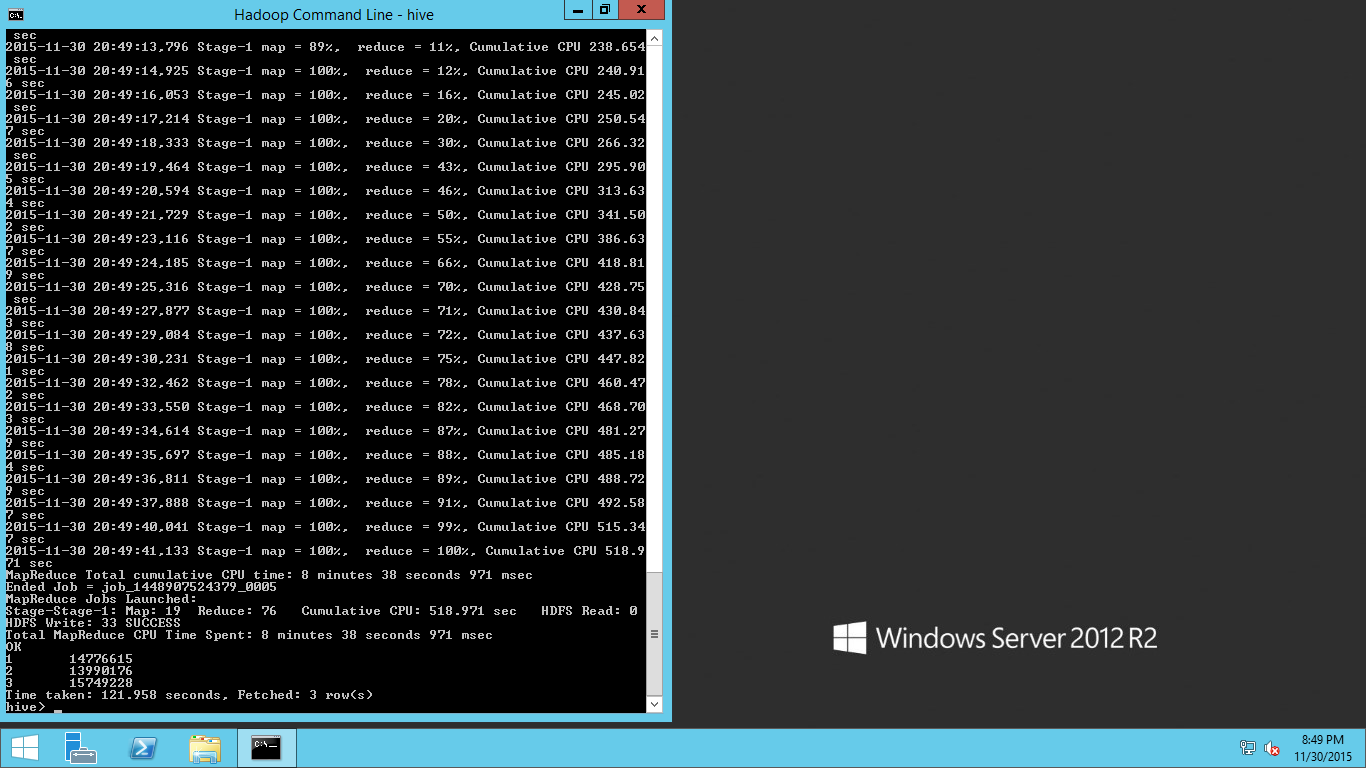


count the total number of records in our trip data set for 3 months



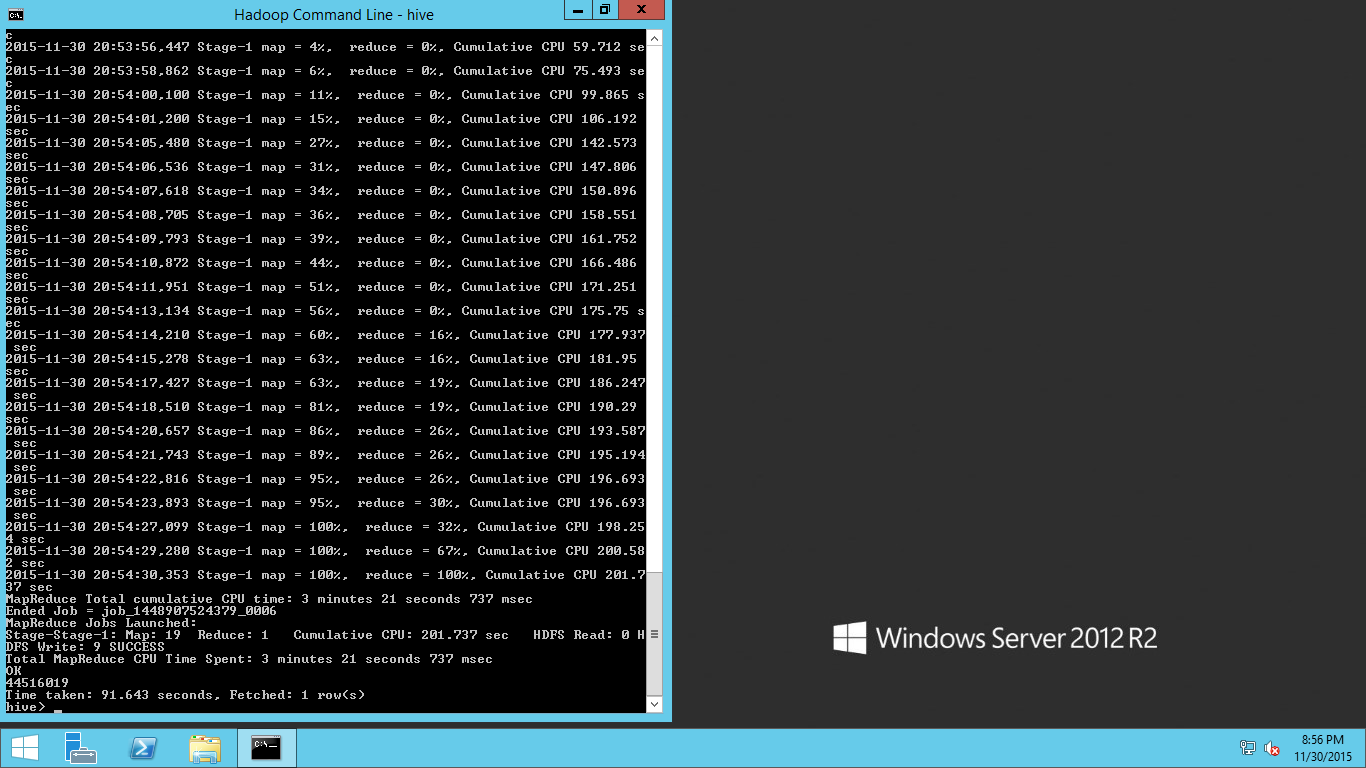
select month, count(\*) from nyctaxidb.fare group by month;

**shows month and fare for 3 months**

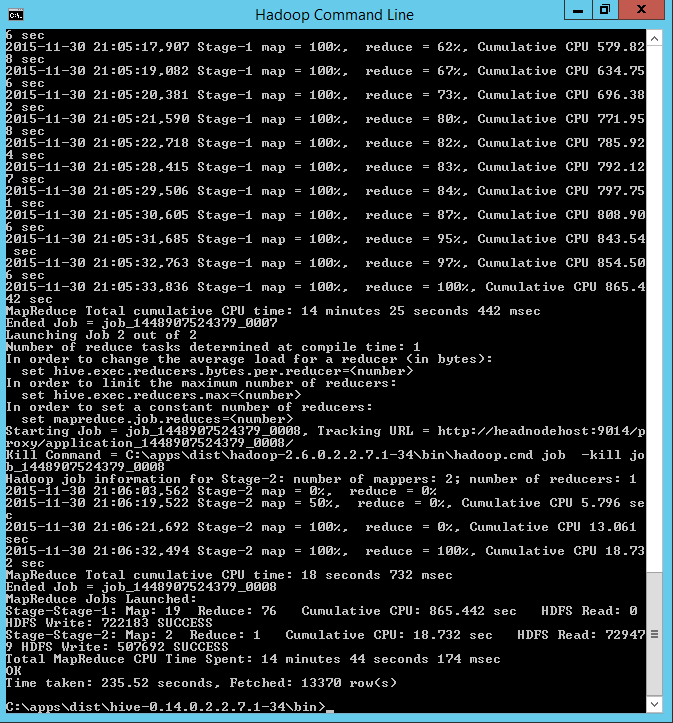


Note that the exact same number of trips per month is returned for both data sets. This provides the first validation that the data has been loaded correctly

Total count number of records



hive -f "C:\temp\sample\_hive\_trip\_count\_by\_medallion.hql" > C:\temp\queryoutput.tsv



SELECT medallion, COUNT(\*) as med\_count

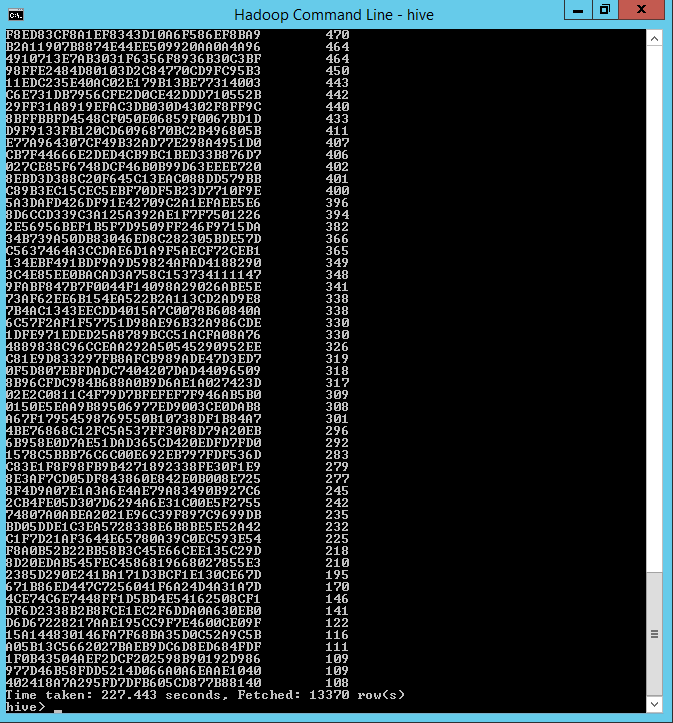
FROM nyctaxidb.fare

WHERE month<=3

GROUP BY medallion

HAVING med\_count > 100

ORDER BY med\_count desc;



SELECT medallion, COUNT(\*) as med\_count

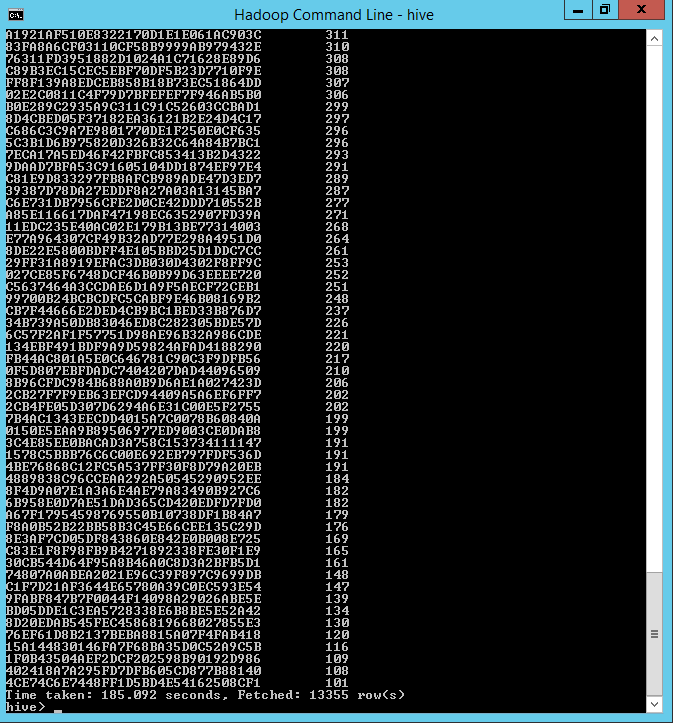
FROM nyctaxidb.fare

WHERE month<=2

GROUP BY medallion

HAVING med\_count > 100

ORDER BY med\_count desc;



SELECT medallion, hack\_license, COUNT(\*) as trip\_count

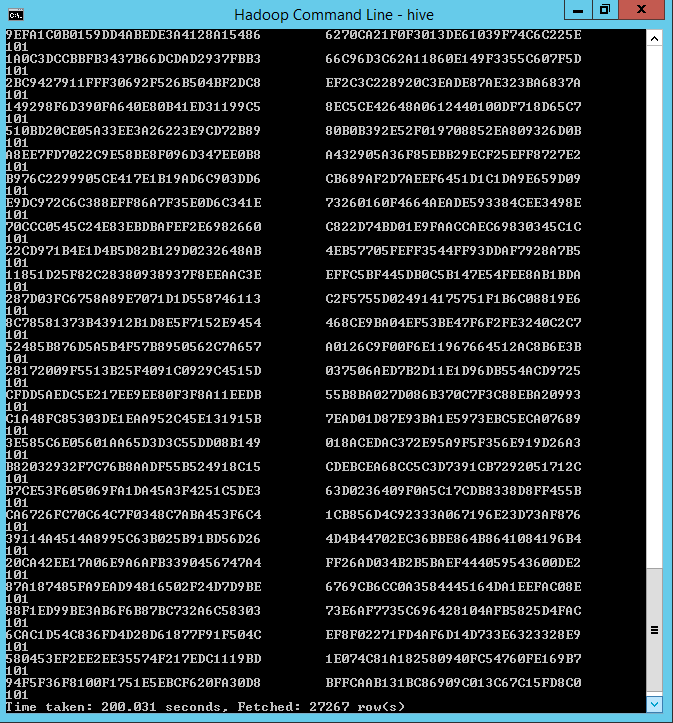
FROM nyctaxidb.fare

WHERE month=1

GROUP BY medallion, hack\_license

HAVING trip\_count > 100

ORDER BY trip\_count desc;



### Exploration: Class distributions in the multiclass setting

SELECT tip\_class, COUNT(\*) AS tipfreq

FROM

(

SELECT if(tip\_amount=0, 0,

if(tip\_amount>0 and tip\_amount<=5, 1,

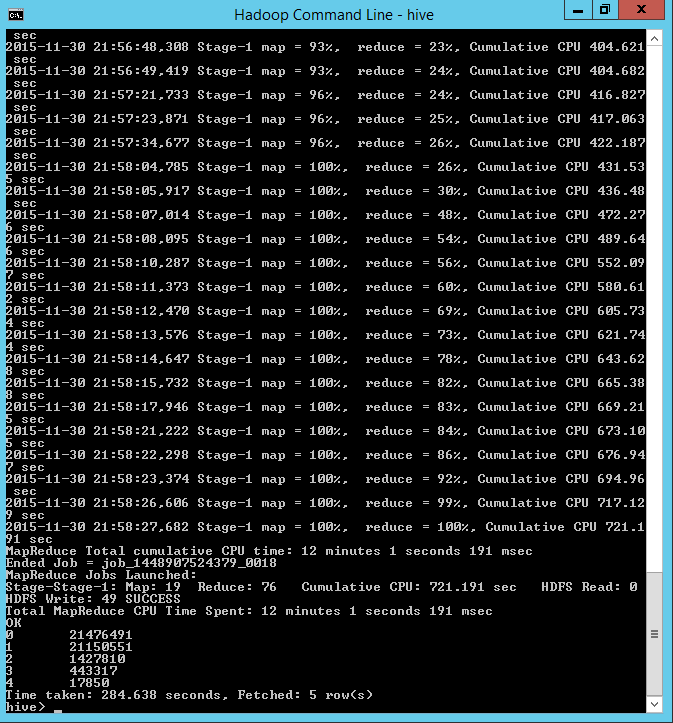
if(tip\_amount>5 and tip\_amount<=10, 2,

if(tip\_amount>10 and tip\_amount<=20, 3, 4)))) as tip\_class, tip\_amount

FROM nyctaxidb.fare

)tipclass

GROUP BY tip\_class;



SELECT tipped, COUNT(\*) AS tipfreq

FROM

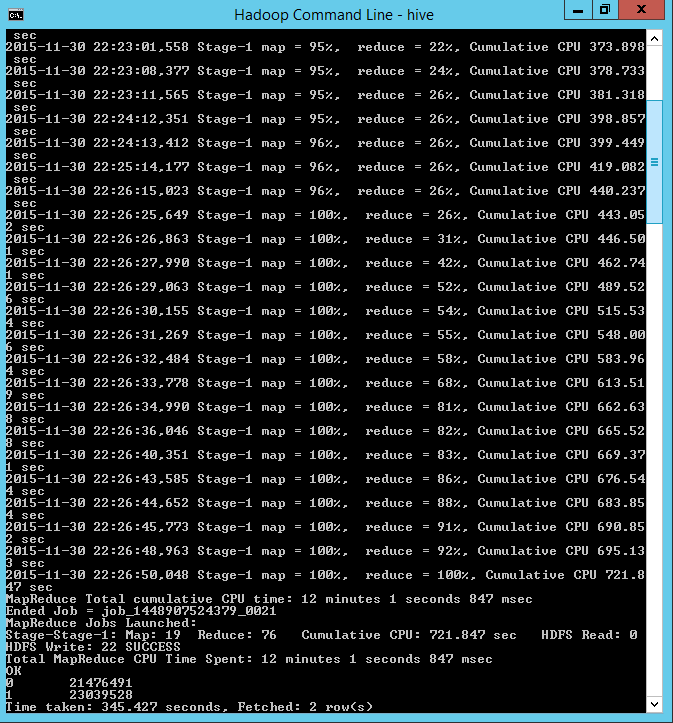
(

SELECT if(tip\_amount > 0, 1, 0) as tipped, tip\_amount

FROM nyctaxidb.fare

)tipclass

GROUP BY tipped;



set R=3959;

set pi=radians(180);

select pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude, trip\_distance, trip\_time\_in\_secs,

${hiveconf:R}\*2\*2\*atan((1-sqrt(1-pow(sin((dropoff\_latitude-pickup\_latitude)

\*${hiveconf:pi}/180/2),2)-cos(pickup\_latitude\*${hiveconf:pi}/180)

\*cos(dropoff\_latitude\*${hiveconf:pi}/180)\*pow(sin((dropoff\_longitude-pickup\_longitude)\*${hiveconf:pi}/180/2),2)))

/sqrt(pow(sin((dropoff\_latitude-pickup\_latitude)\*${hiveconf:pi}/180/2),2)

+cos(pickup\_latitude\*${hiveconf:pi}/180)\*cos(dropoff\_latitude\*${hiveconf:pi}/180)\*

pow(sin((dropoff\_longitude-pickup\_longitude)\*${hiveconf:pi}/180/2),2))) as direct\_distance

from nyctaxidb.trip

where month=1

and pickup\_longitude between -90 and -30

and pickup\_latitude between 30 and 90

and dropoff\_longitude between -90 and -30

and dropoff\_latitude between 30 and 90;

**output: queryoutputdir**

hdfs dfs -ls wasb:///queryoutput\_direct\_distance

