# Checkpoints

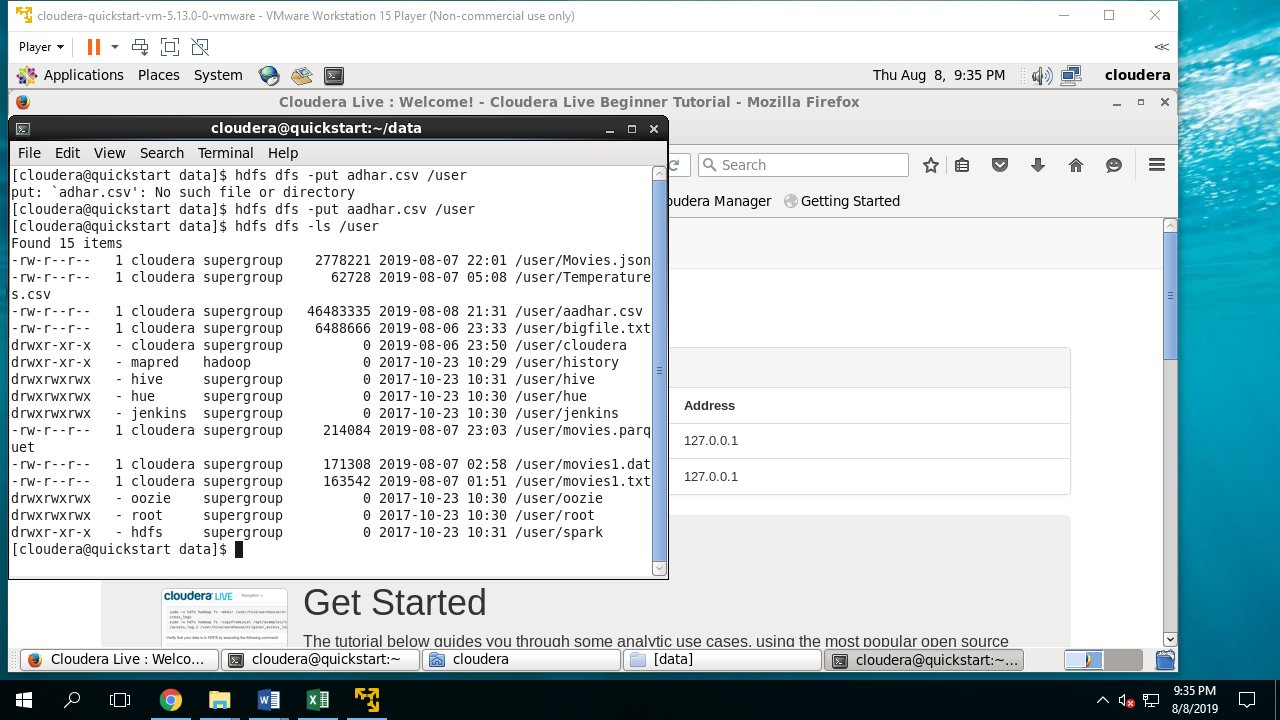
## Checkpoint 1

Load the data into HDFS, Hive Managed table, Hive External table and Spark DataFrame.

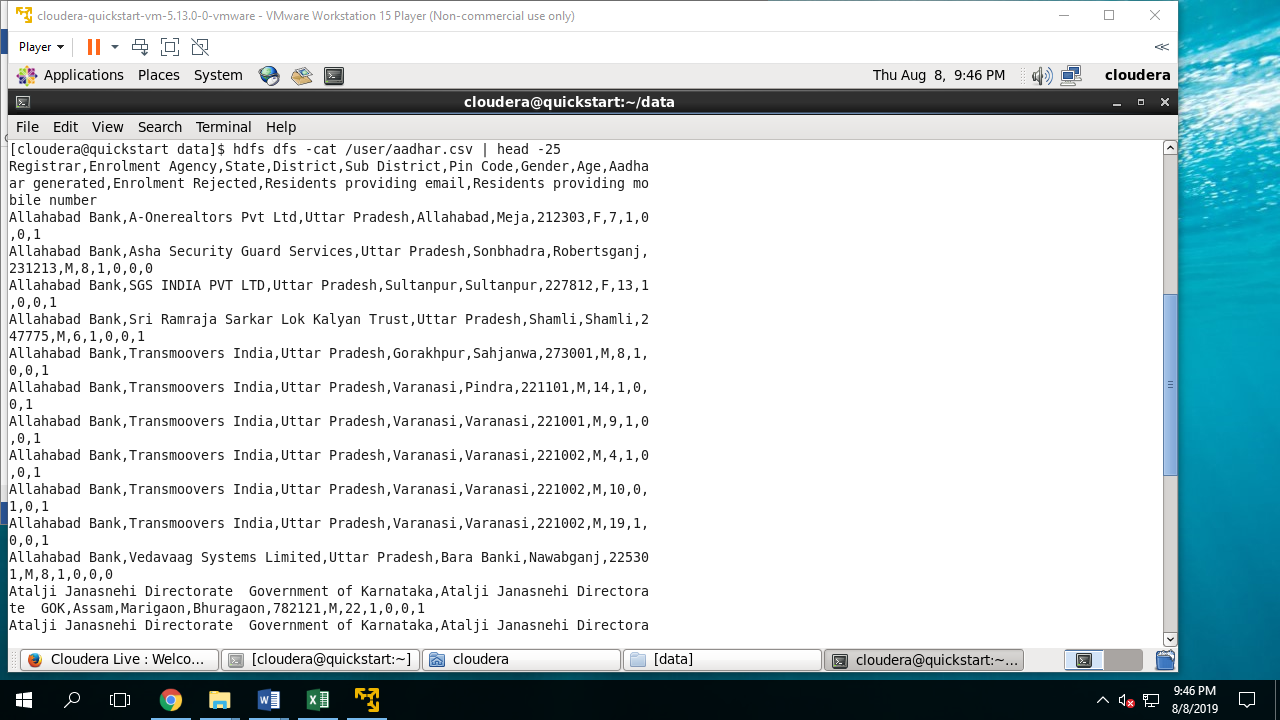
Commit the screenshot of the view/result of the top 25 rows from each individual store (HDFS, Hive – Managed/External and Spark DataFrame

Load the data into HDFS

Hdfs dfs –put aadhar.csv /user



1. Commit the screenshot of the view/result of the top 25 rows from each individual store (HDFS, Hive – Managed/External and Spark DataFrame).



**HIVE MANGED TABLE**

create table adhar (

> registrar String,

> private\_agency String,

> state String,

> district String,

> sub\_district String,

> pincode String,

> gender String,

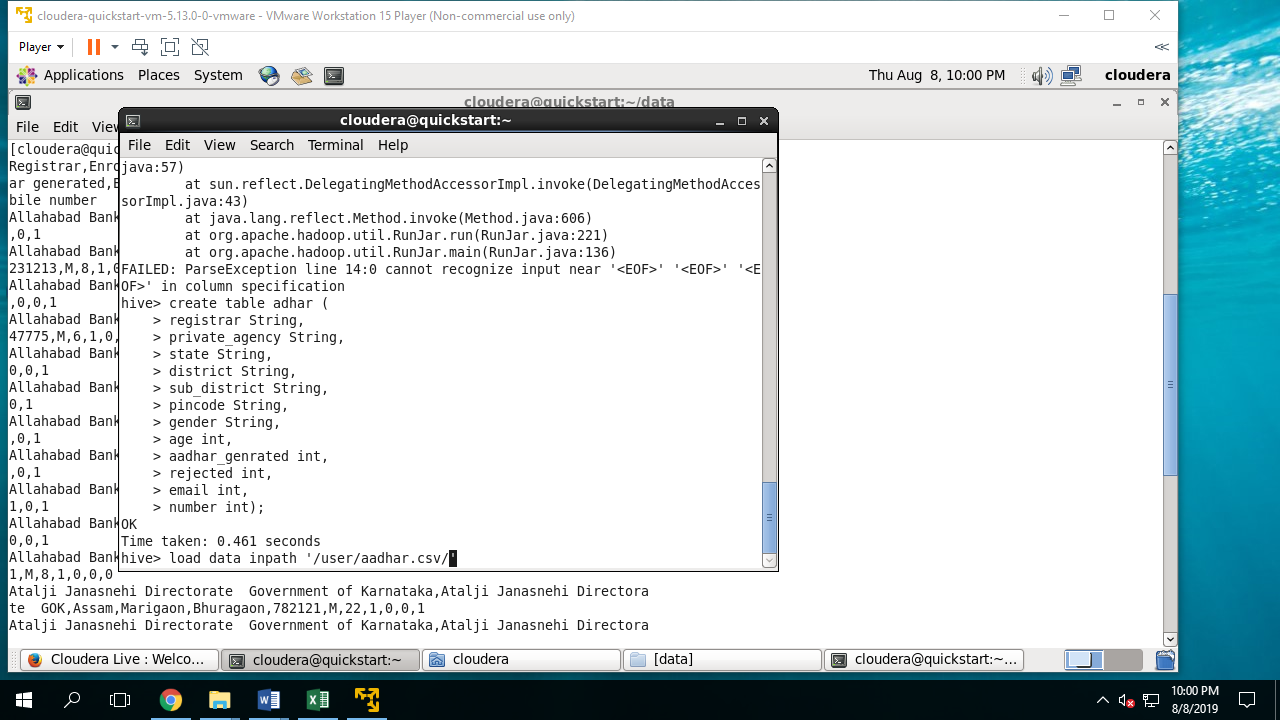
> age int,

> aadhar\_genrated int,

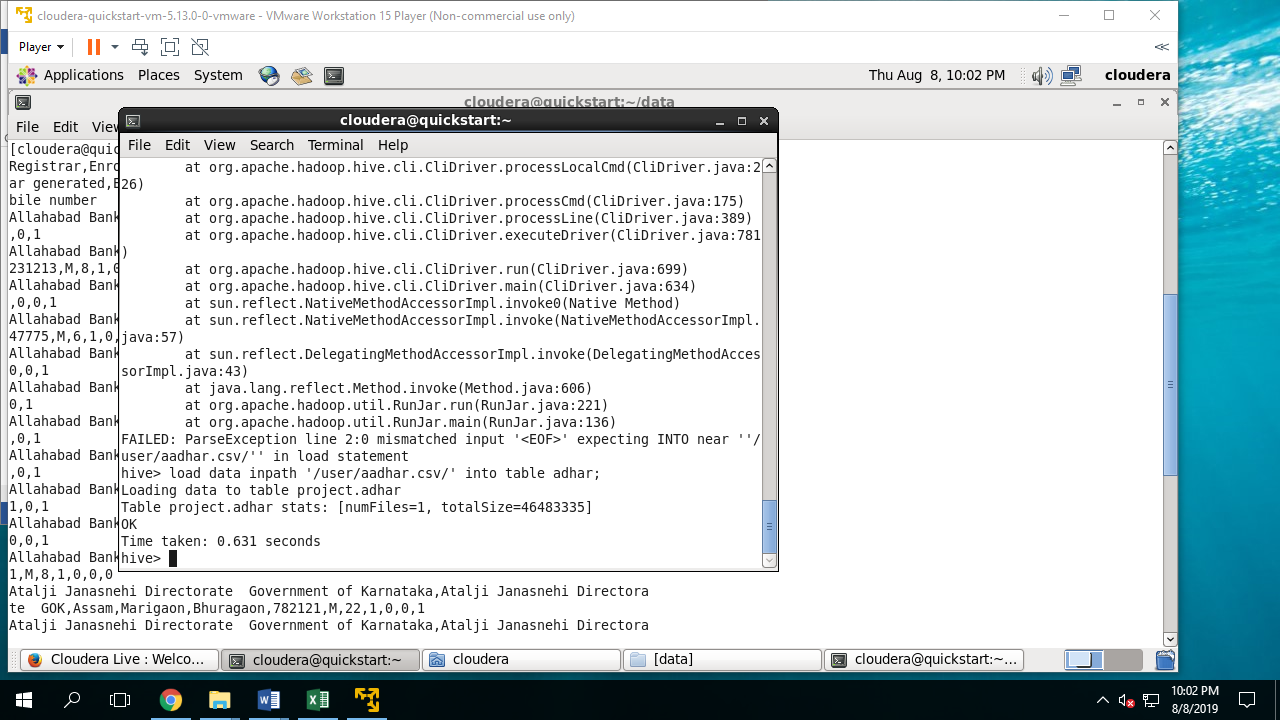
> rejected int,

> email int,

> number int);



load data inpath '/user/aadhar.csv/' into table adhar;



create table adhar (

registrar String,

private\_agency String,

state String,

district String,

sub\_district String,

pincode String,

gender String,

age int,

aadhar\_genrated int,

rejected int,

email int,

number int)

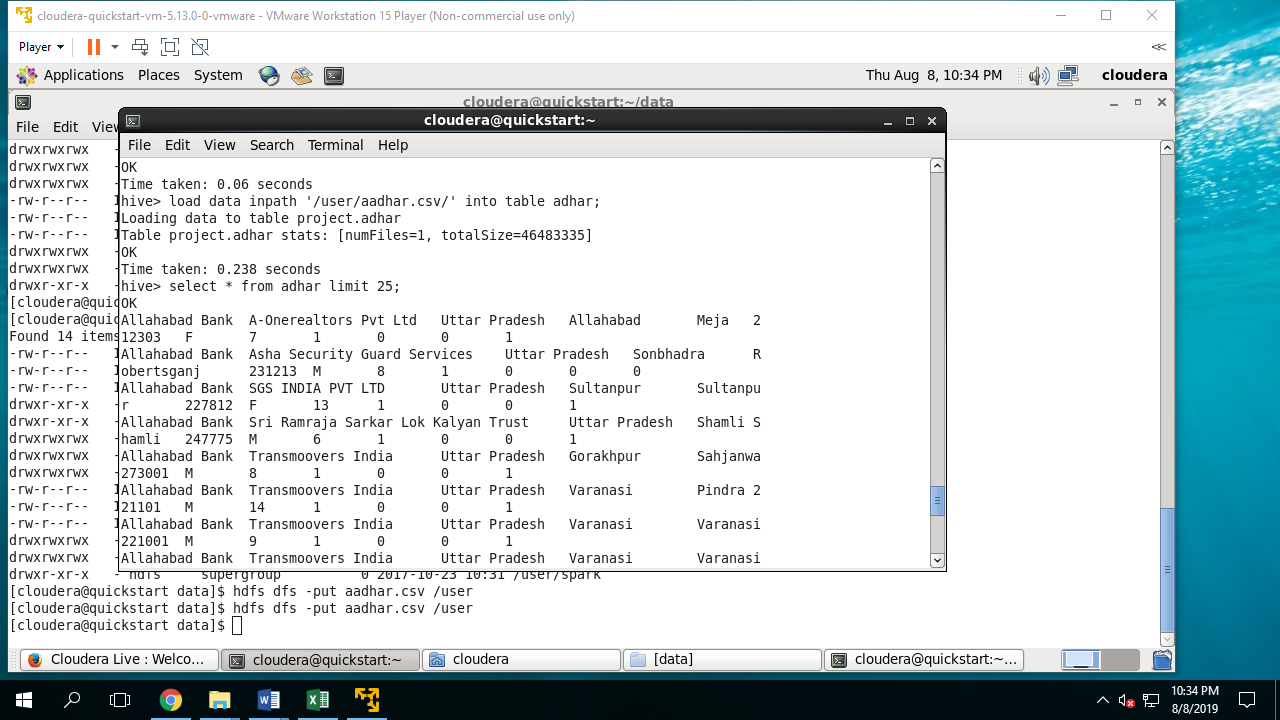
row format delimited fields terminated by ","

TBLPROPERTIES('skip.header.line.count'='1');

**LOAD DATA INTO TABLE**

load data inpath '/user/aadhar.csv/' into table adhar;

select \* from adhar limit 25;



create table dataex (

registrar String,

private\_agency String,

state String,

district String,

sub\_district String,

pincode String,

gender String,

age int,

aadhar\_genrated int,

rejected int,

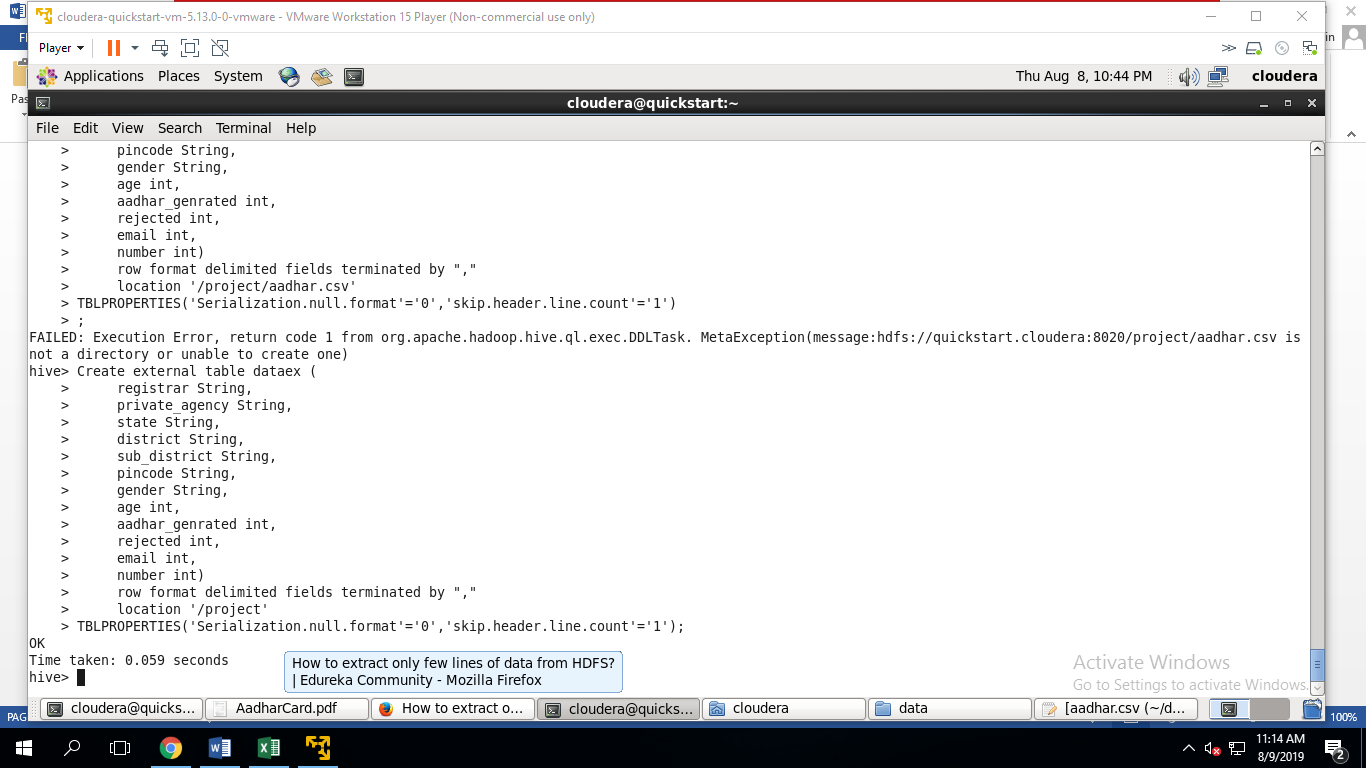
email int,

number int)

row format delimited fields terminated by ","

location ‘project’

TBLPROPERTIES('skip.header.line.count'='1');



select \* from dataex limit 25;



**DATA FRAME**

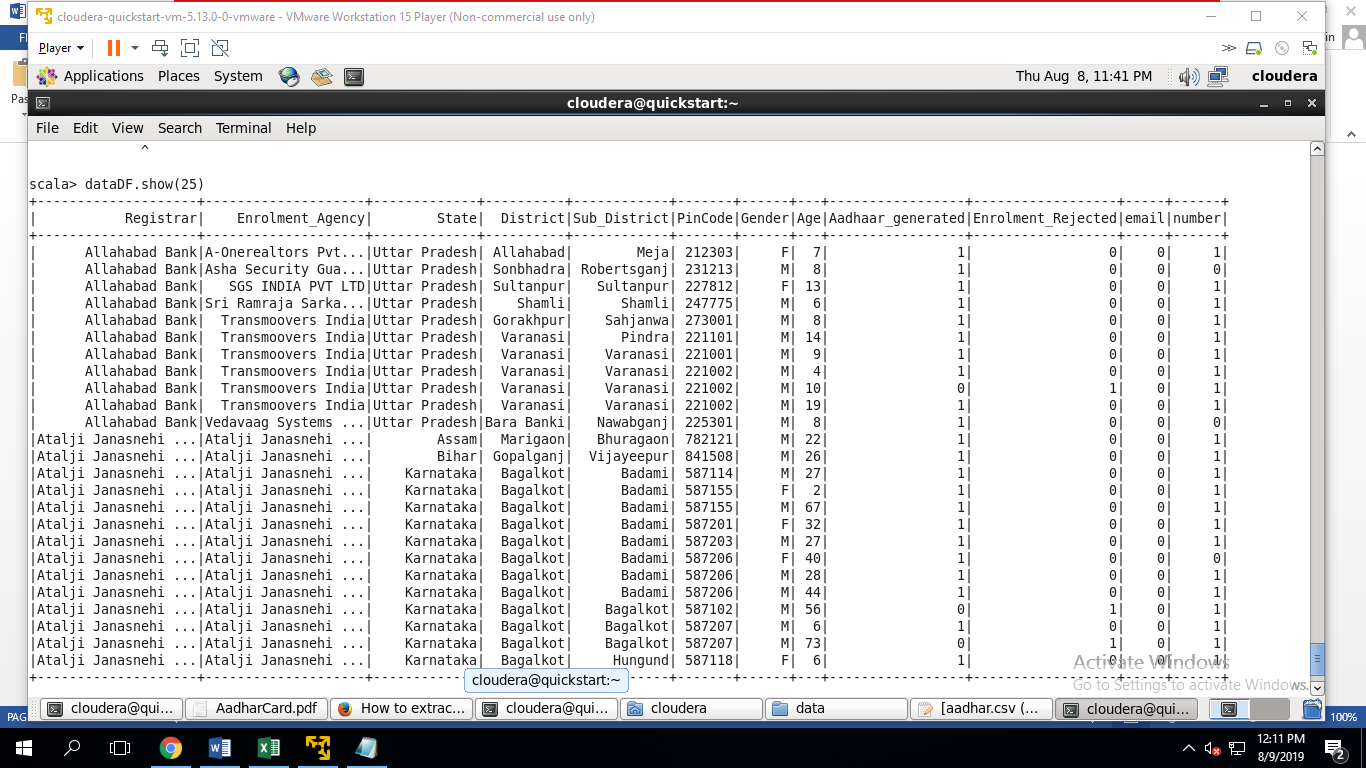
val datardd=sc.textFile("/project/aadhar.csv")

val h=datardd.first()

var rdddata=datardd.filter(r =>r!=h)

var rddfin=rdddata.map(x=>(x.split(",")(0),x.split(",")(1), x.split(",")(2),x.split(",")(3),x.split(",")(4),x.split(",")(5),x.split(",")(6),x.split(",")(7).toInt,x.split(",")(8).toInt,x.split(",")(9).toInt,x.split(",")(10).toInt,x.split(",")(11).toInt))

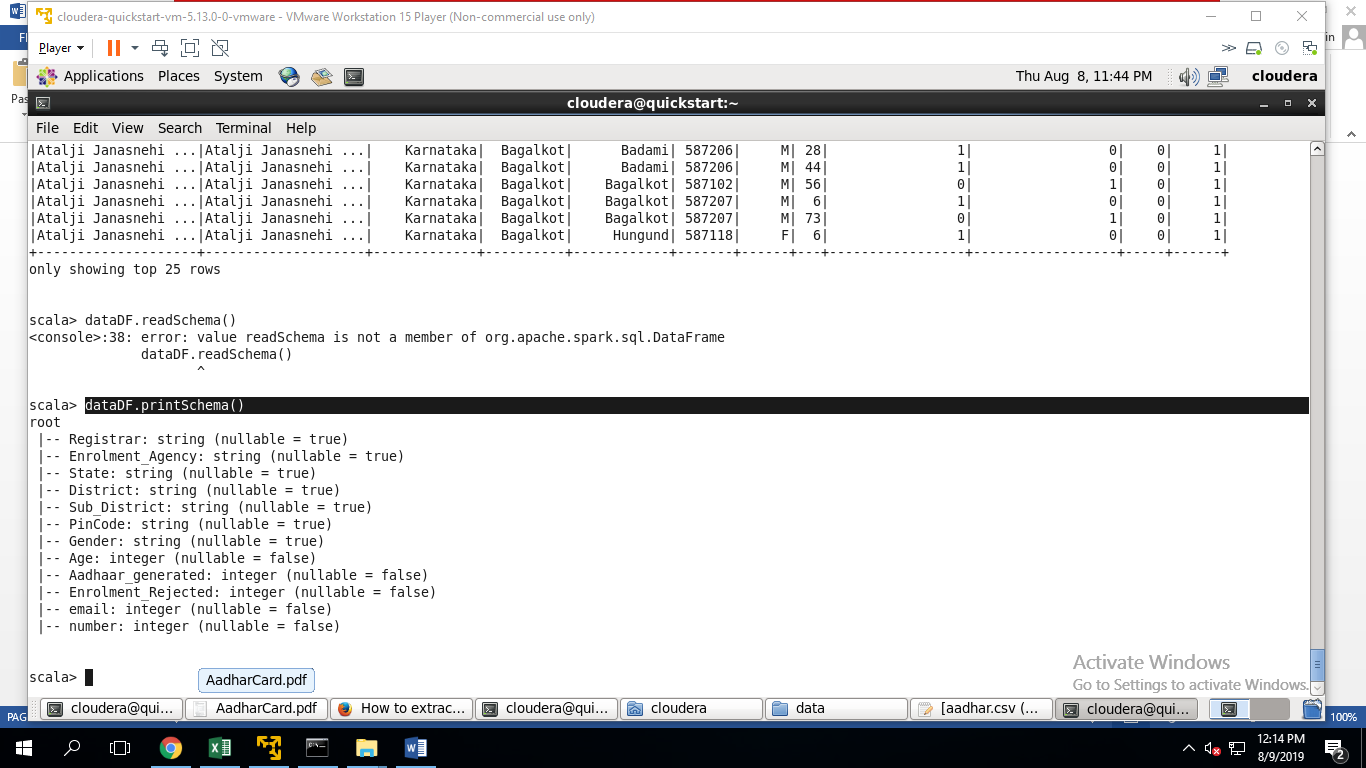
val dataDF=rddfin.toDF("Registrar","Enrolment\_Agency","State","District","Sub\_District","PinCode","Gender","Age","Aadhaar\_generated","Enrolment\_Rejected","email","number")



## Checkpoint 2

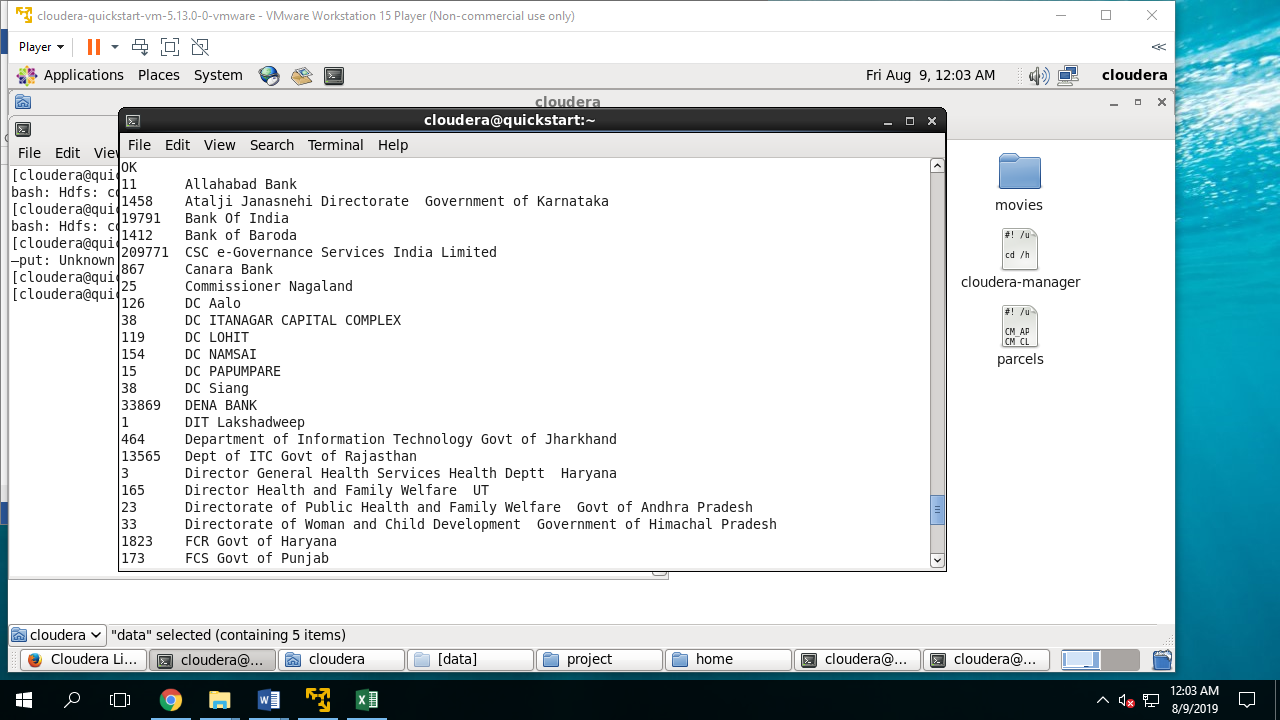
**Q-2** Describe the schema.

dataDF.printSchema()



**Q3 Find the count and names of registrars in the table**.

**select count(registrar),registrar from adhar group by registrar;**

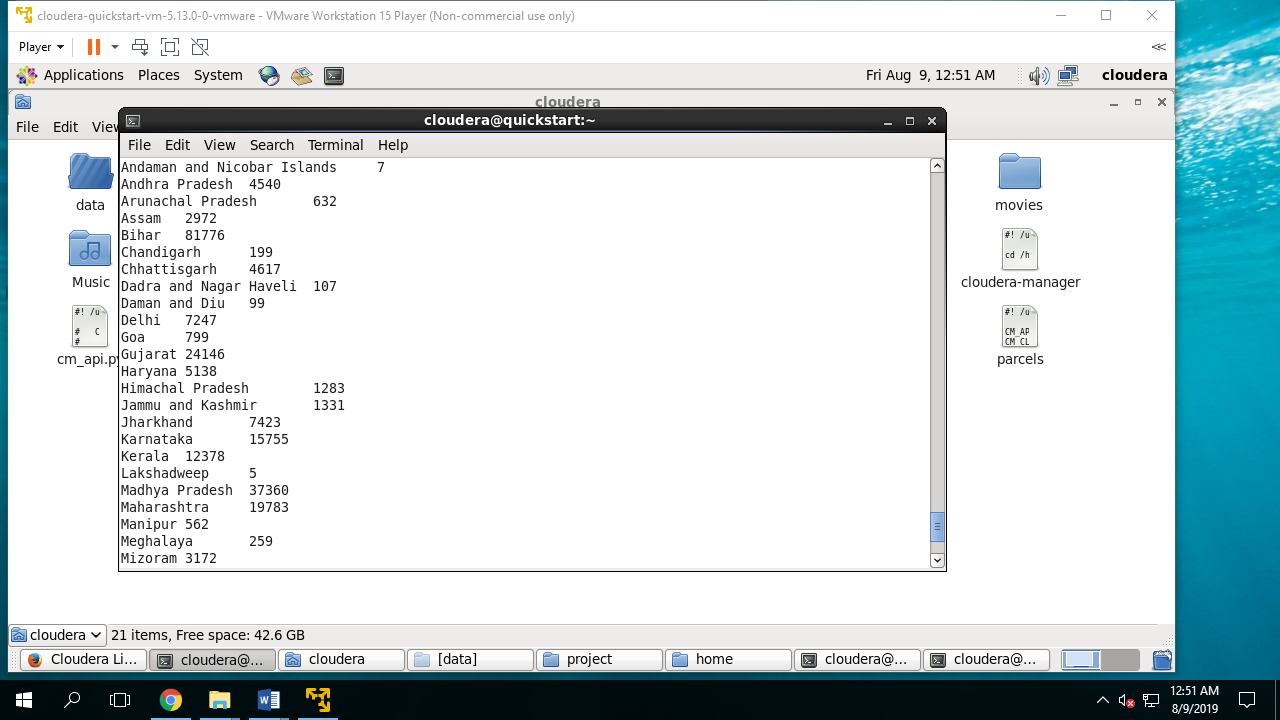


**Q-4 Find the number of states, districts in each state and sub-districts in each district.**

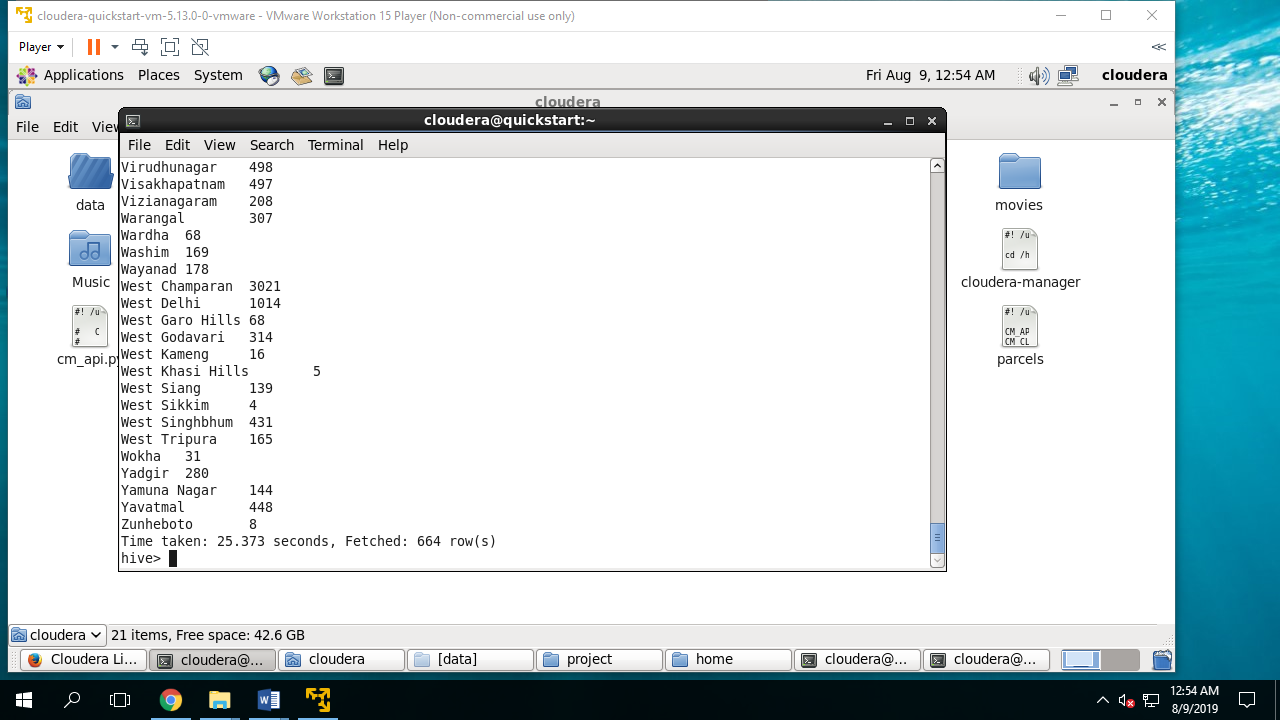
1. **select count(distinct(state)) from adhar;**



**(B) select state,count(district) from adhar group by state;**

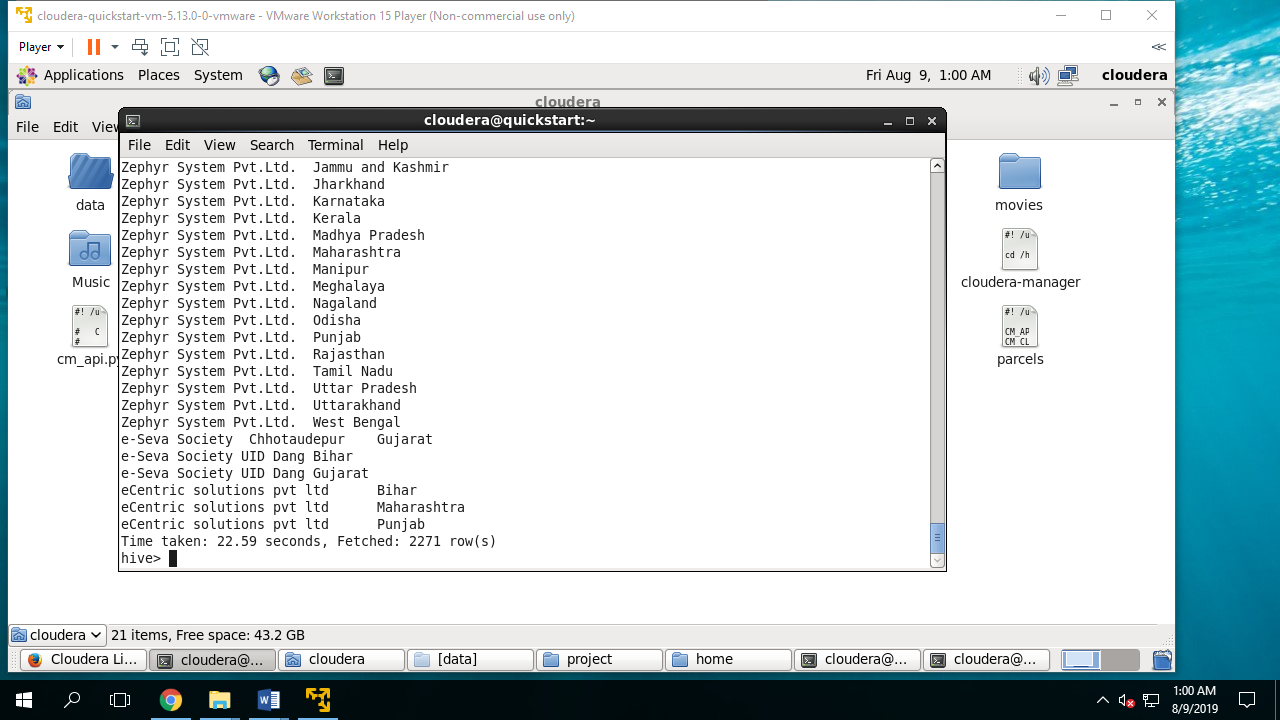


**(C) SELECT DISTRICT,COUNT(SUB\_DISTRICT) FROM ADHAR GROUP BY DISTRICT**



**Q-6 Find out the names of private agencies for each state.**

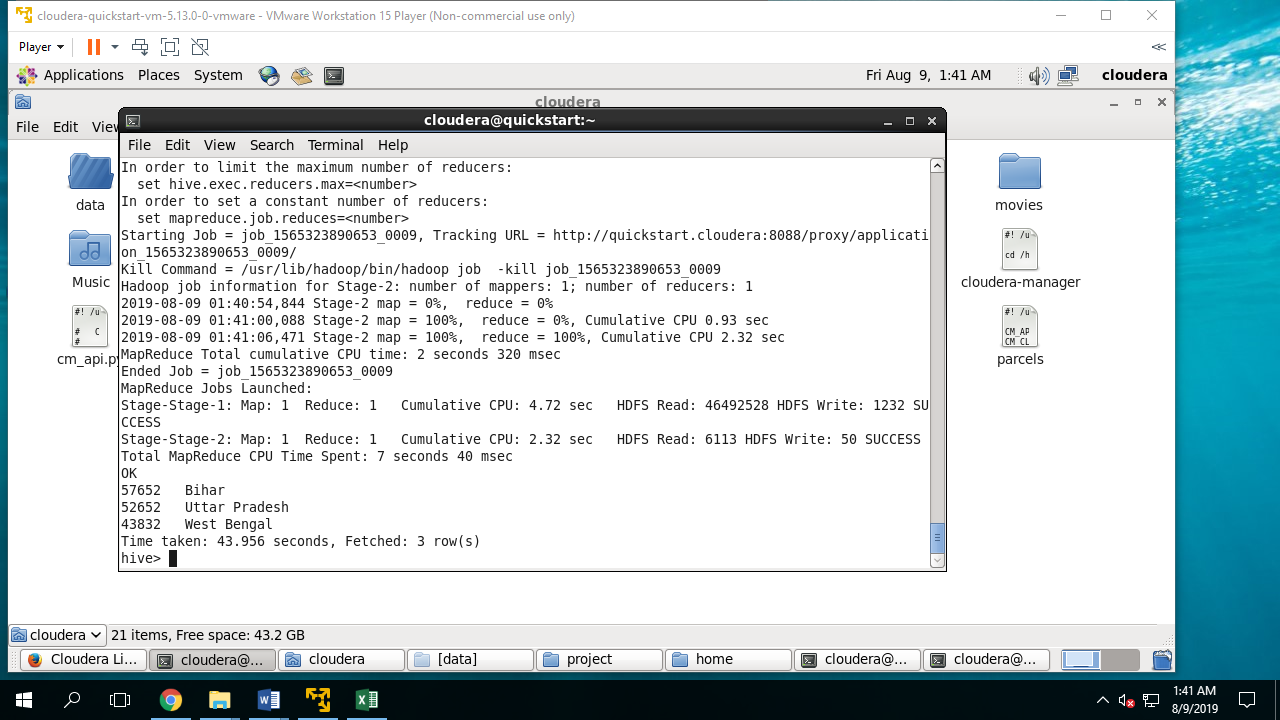
**SELECT PRIVATE\_AGENCY,STATE FROM ADHAR GROUP BY PRIVATE\_AGENCY,STATE;**



## Checkpoint 3

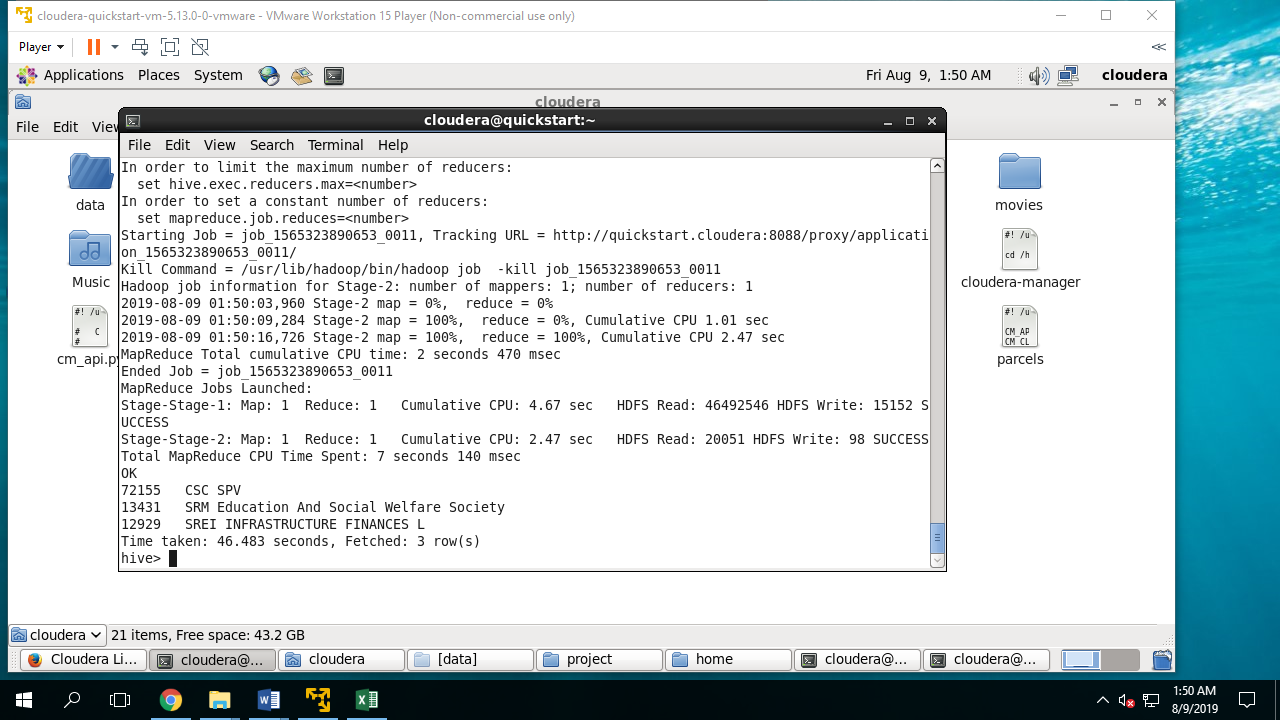
**Q-8 Find top 3 states generating most number of Aadhaar cards?**

**Select count(\*) as no,state from adhar where adhar\_genrated=1 group by state order by no desc limit 3;**



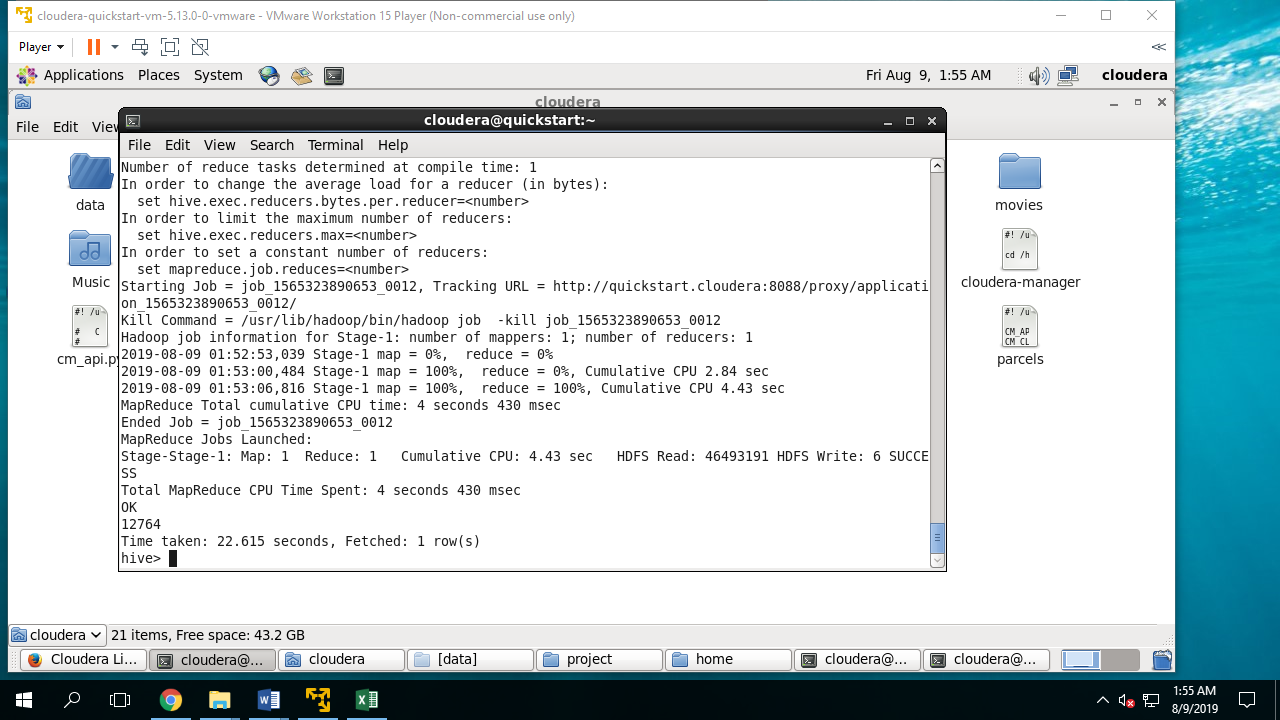
**Q-9 Find top 3 private agencies generating the most number of Aadhar cards?**

**Select count(\*) as no,private\_agency from adhar where aadhar\_genrated=1 group by private\_agency order by no desc limit 3;**



**Q-10 Find the number of residents providing email, mobile number? (Hint: consider non-zero values.)**

**Select count(\*) from adhar where email=1 and number=1;**



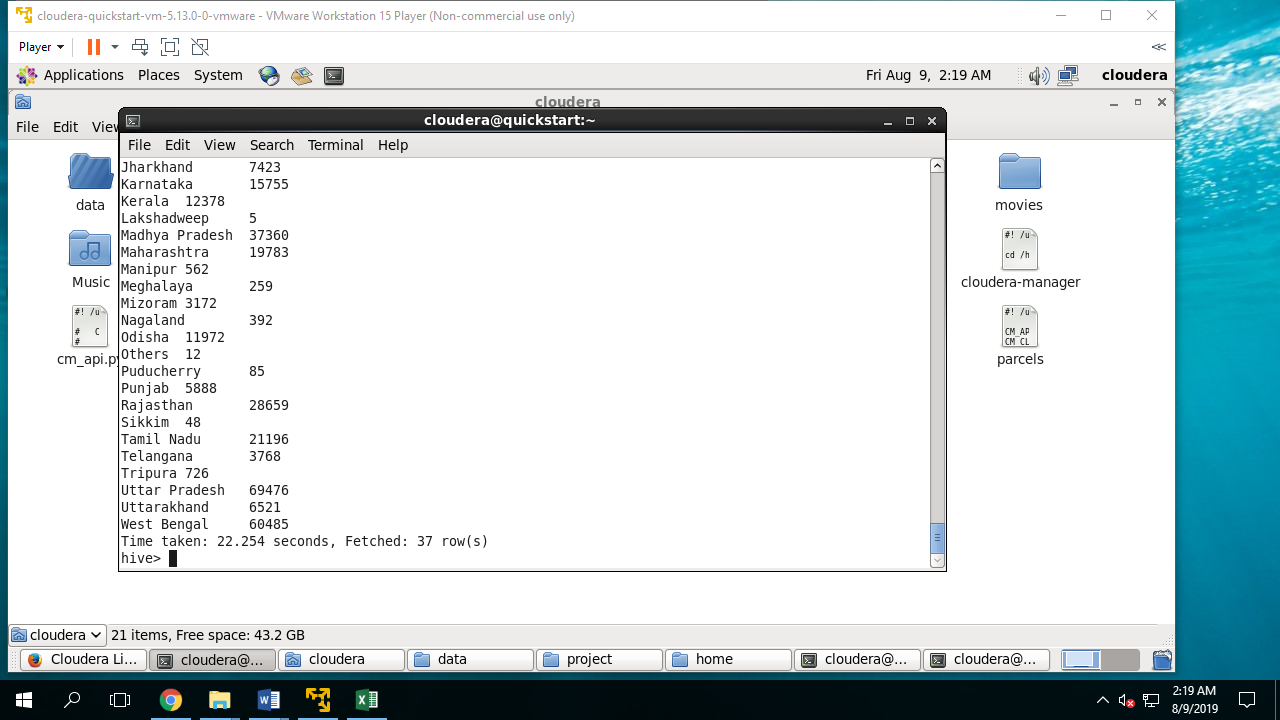
**Q-11 Find top 3 districts where enrolment numbers are maximum?**

**Select district,aadhar\_genrated+rejected as su from adhar order by su desc limit 3;**



**Q-12 Find the no. of Aadhaar cards generated in each state?**

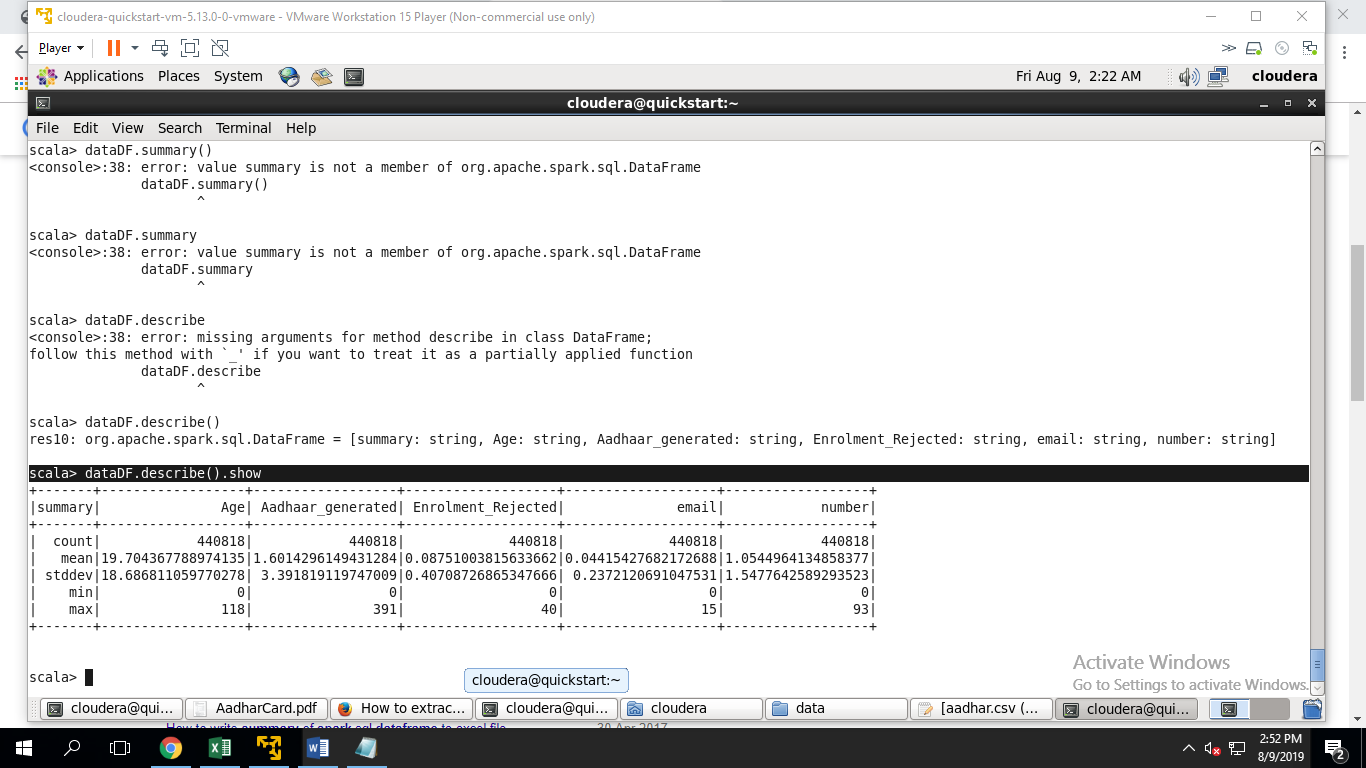
**Select state,count(aadhar\_genrated) from adhar as no group by state;**



**CHECK POINT 4**

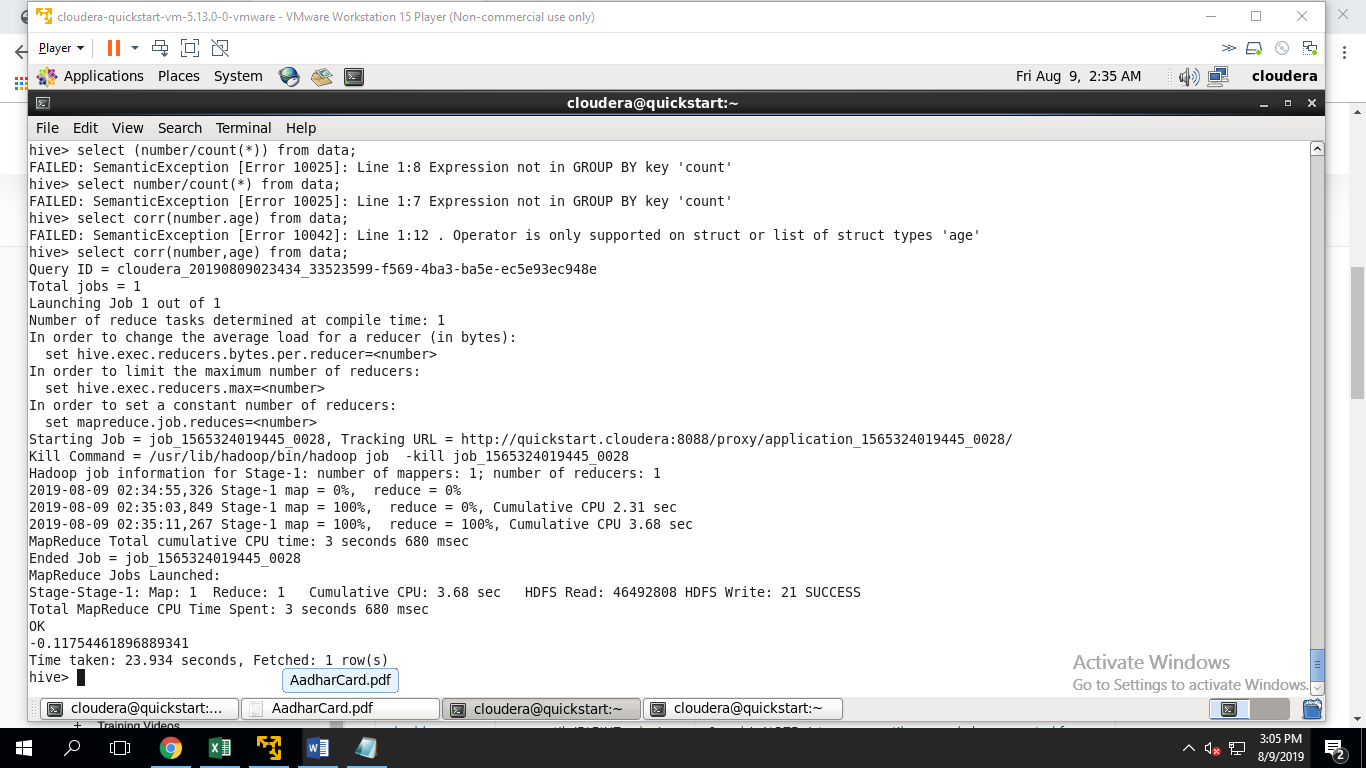
Q.13 Create a data frame using the file and provide its summary

scala> dataDF.describe().show



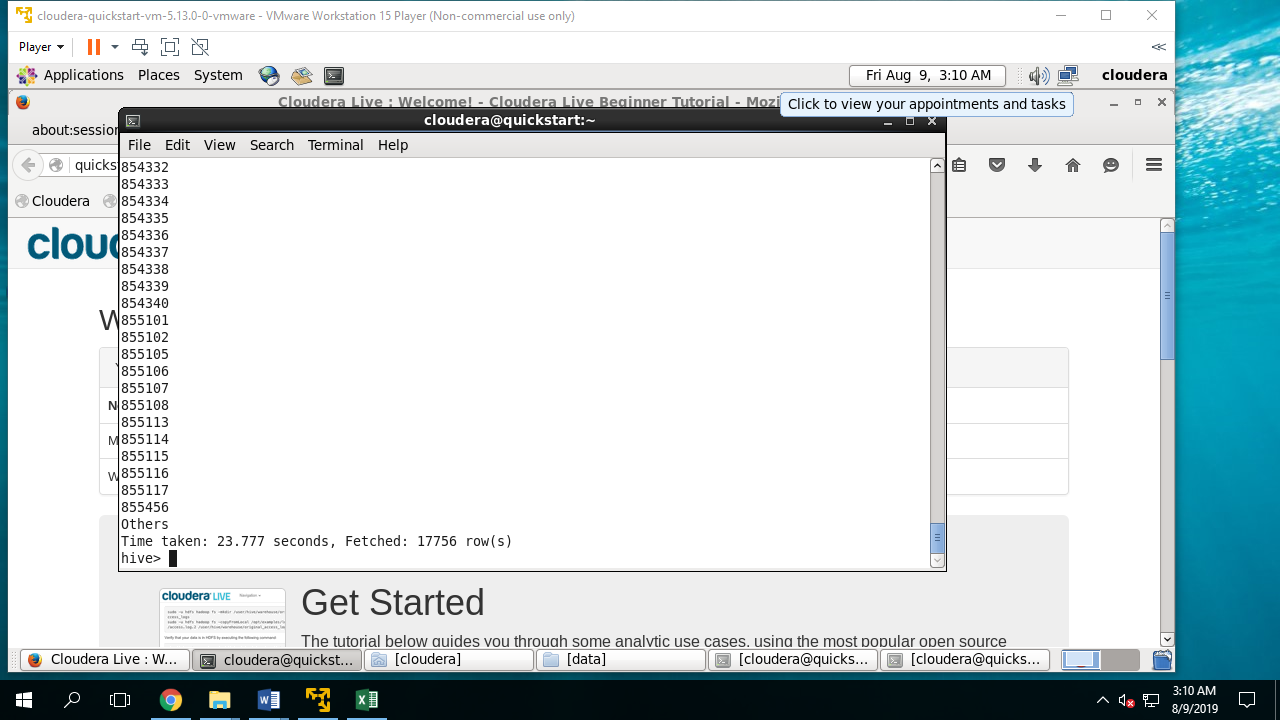
Q.14 **Write a command to see the correlation between “age” and “mobile\_number”? (Hint: Consider the percentage of people who have provided the mobile number out of the total applicants)**

select corr(number,age) from data;



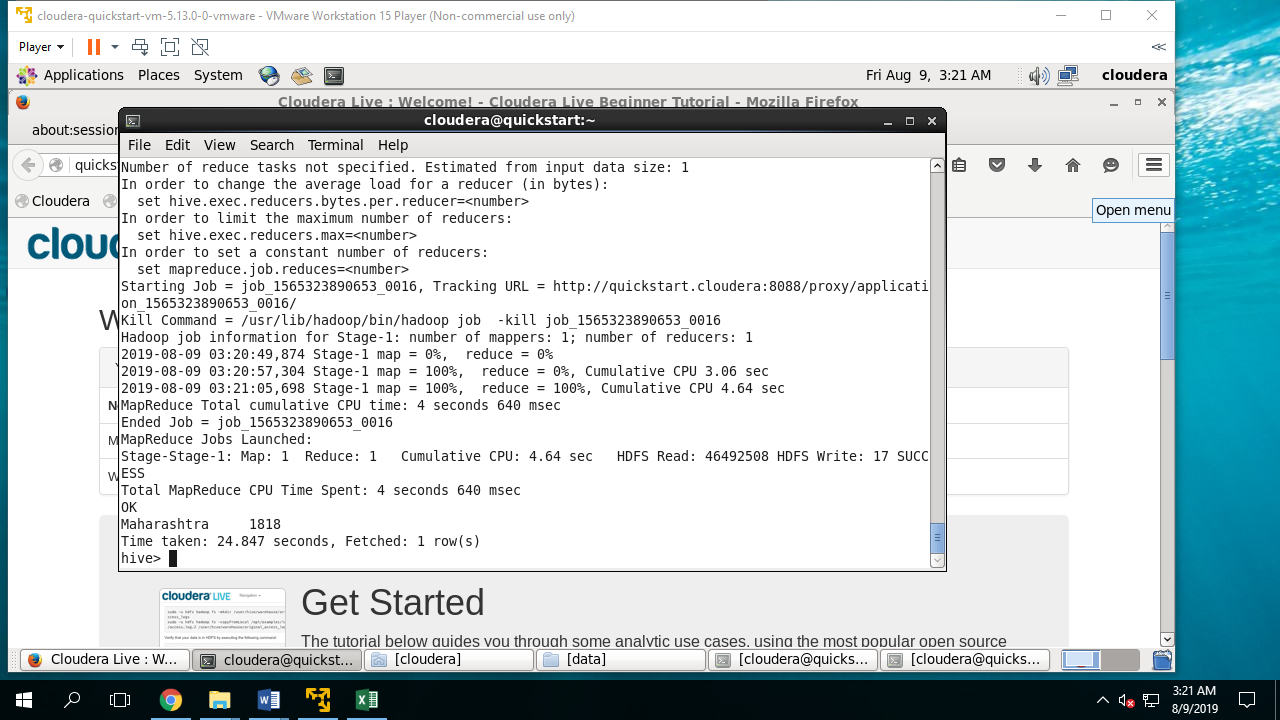
**Q-15 Find the number of unique pincodes in the data?**

**Select distinct(pincode) from adhar;**



**Q-16 Find the number of Aadhaar registrations rejected in Uttar Pradesh and Maharashtra?**

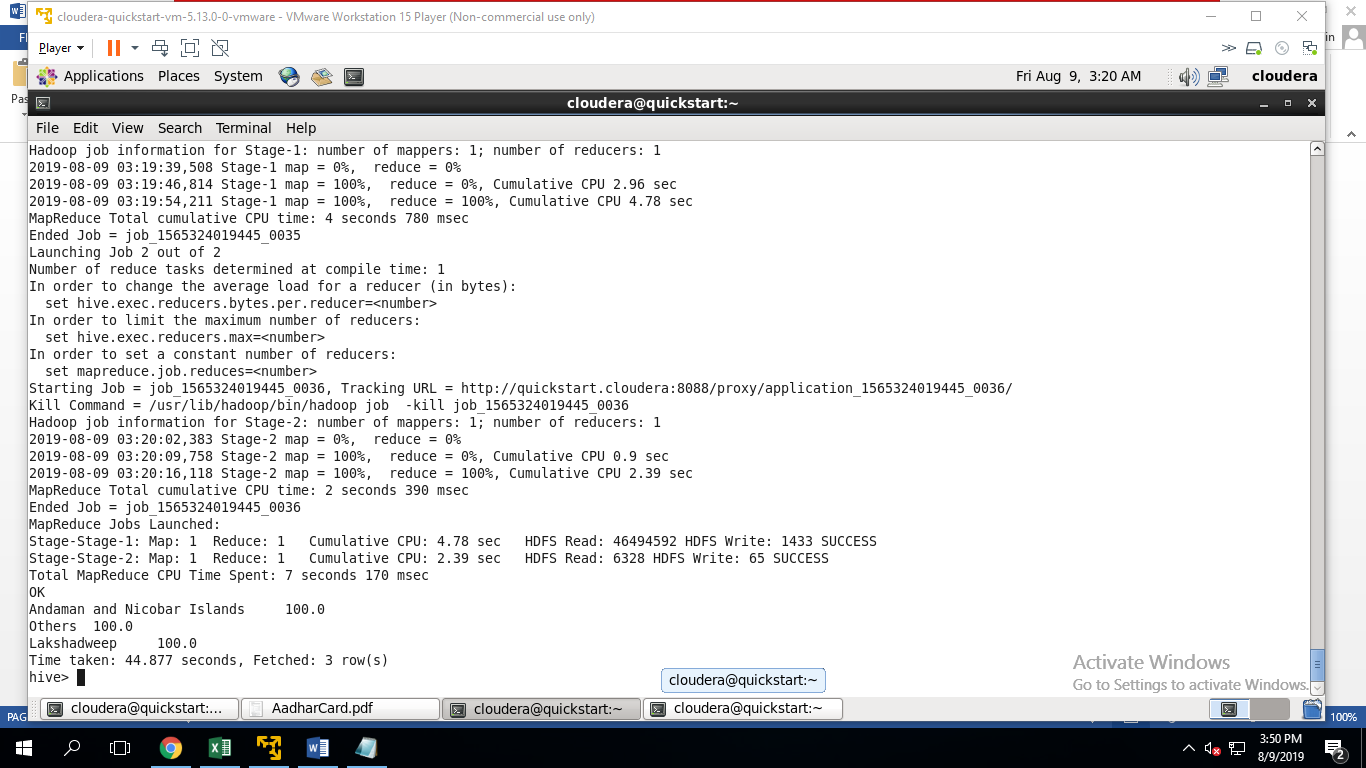
**Select state,sum(rejected) from adhar where state ="uttar pradesh" or state ="Maharashtra" group by state;**



## Checkpoint 5

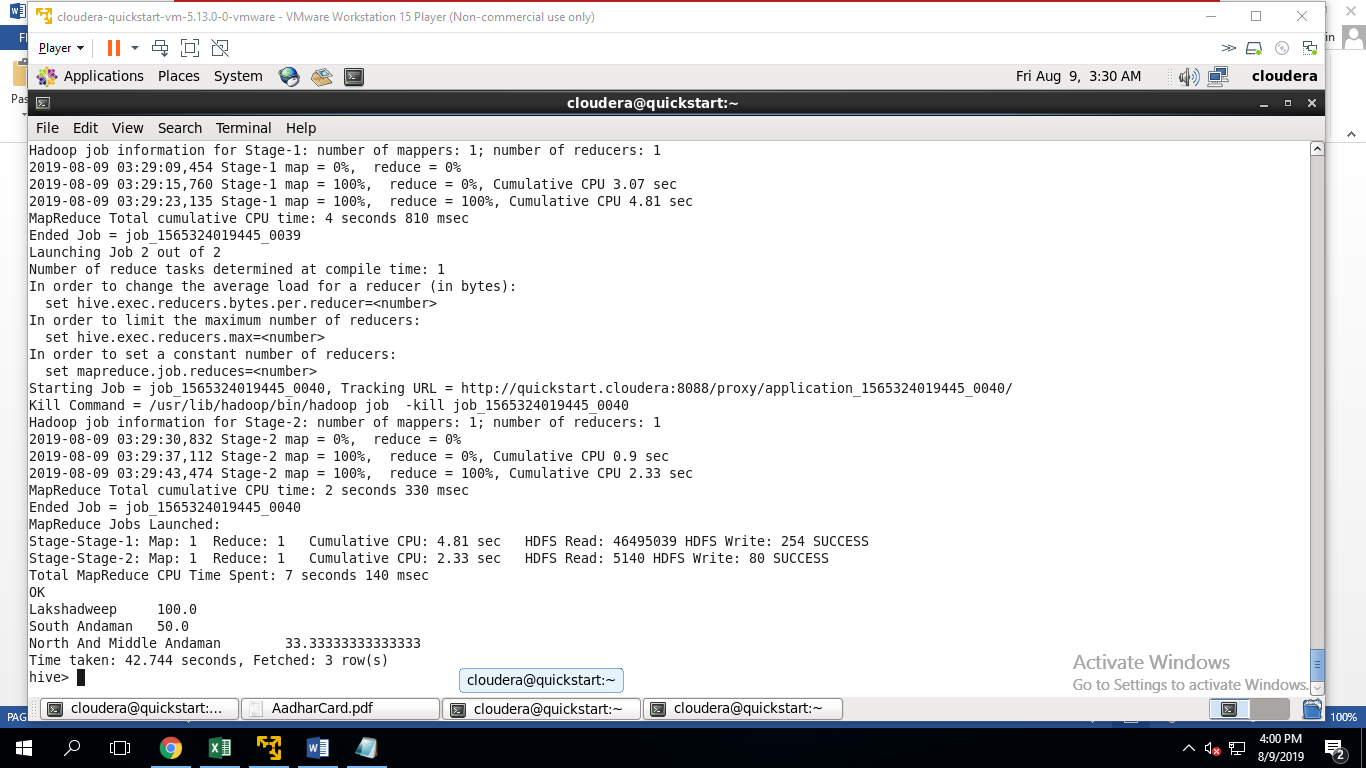
**Q-17 The top 3 states where the percentage of Aadhaar cards being generated for males is the highest**

select state,(sum(aadhar\_genrated)/sum(aadhar\_genrated+rejected)\*100) as percent from data where gender="M" group by state order by percent desc limit 3;



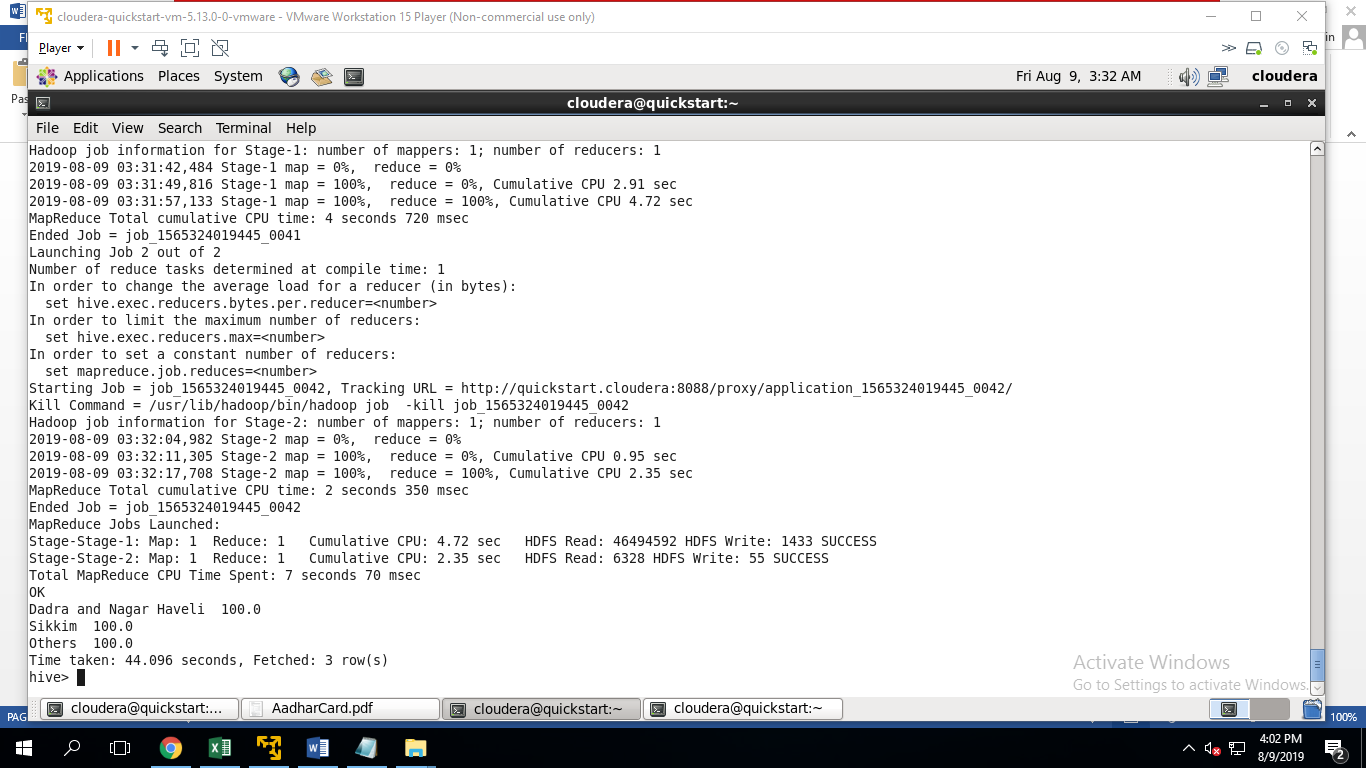
Q18. **In each of these 3 states, identify the top 3 districts where the percentage of Aadhaar cards being rejected for females is the highest.**

select district,(sum(rejected)/sum(aadhar\_genrated+rejected)\*100) as perc from data where gender="F" and state in("Andaman and Nicobar Islands","Others","Lakshadweep") group by district order by perc desc limit 3;



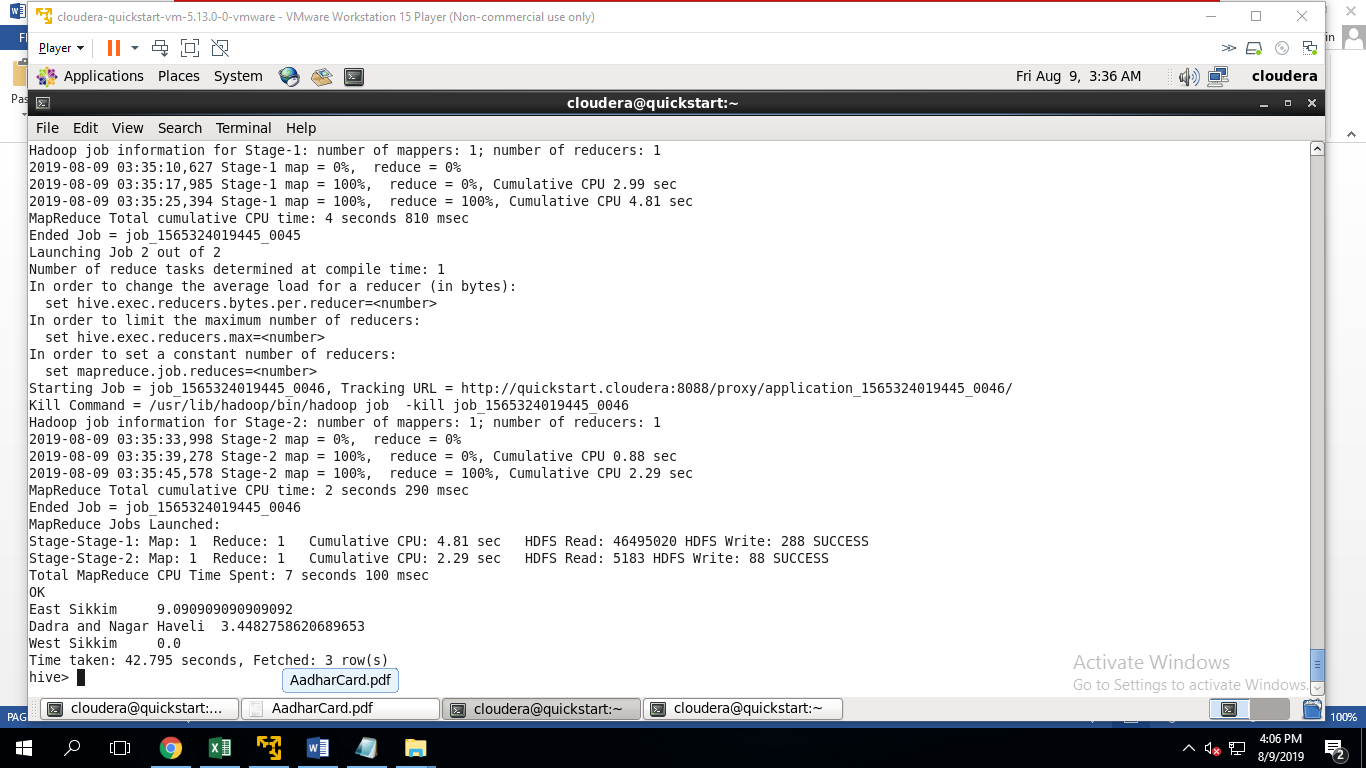
Q.19 **The top 3 states where the percentage of Aadhaar cards being generated for females is the highest**.

select state,(sum(aadhar\_genrated)/sum(aadhar\_genrated+rejected)\*100) as percent from data where gender="F" group by state order by percent desc limit 3;



Q.**20 In each of these 3 states, identify the top 3 districts where the percentage of Aadhaar cards being rejected for males is the highest.**

select district,(sum(rejected)/sum(aadhar\_genrated+rejected)\*100) as perc from data where gender="M" and state in("Dadra and Nagar Haveli","Others","Sikkim") group by district order by perc desc limit 3;



**Q.21 The summary of the acceptance percentage of all the Aadhaar cards applications by bucketing the age group into 10 buckets.**

set hive.exec.dynamic.partition.mode=nonstrict

create table user\_buck (

registrar String,

private\_agency String,

state String,

district String,

sub\_district String,

pincode String,

gender String,

age int,

aadhar\_genrated int,

rejected int,

email int,

number int)

clustered by(age) into 10 buckets

row format delimited fields terminated by ',';

insert into user\_buck select registrar String,

private\_agency String,

state String,

district String,

sub\_district String,

pincode String,

gender String,

age int,

aadhar\_genrated int,

rejected int,

email int,

number int from data;

select (sum(aadhar\_genrated)/sum(aadhar\_genrated+rejected)\*100) as Acceptance\_percentage from user\_buck limit 5; 