Checkpoint 1

Question 1: Load the data into HDFS, Hive Managed table, Hive External table and Spark DataFrame.

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

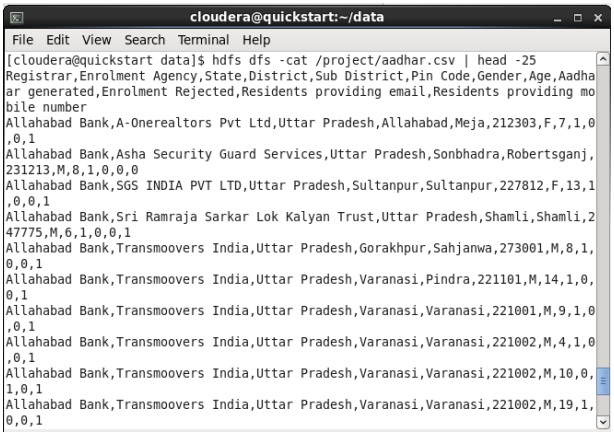
Solution

Load into hdfs

hdfs dfs -put aadhar.csv /project/

Printing

hdfs dfs -cat /project/aadhar.csv | head -25



Hive

Table

```
create table data (
    registrar String,
    private_agency String,
    state String,
    district String,
    sub_district String,
    pincode String,
```

gender String,

age int,

aadhar_genrated int,

rejected int,

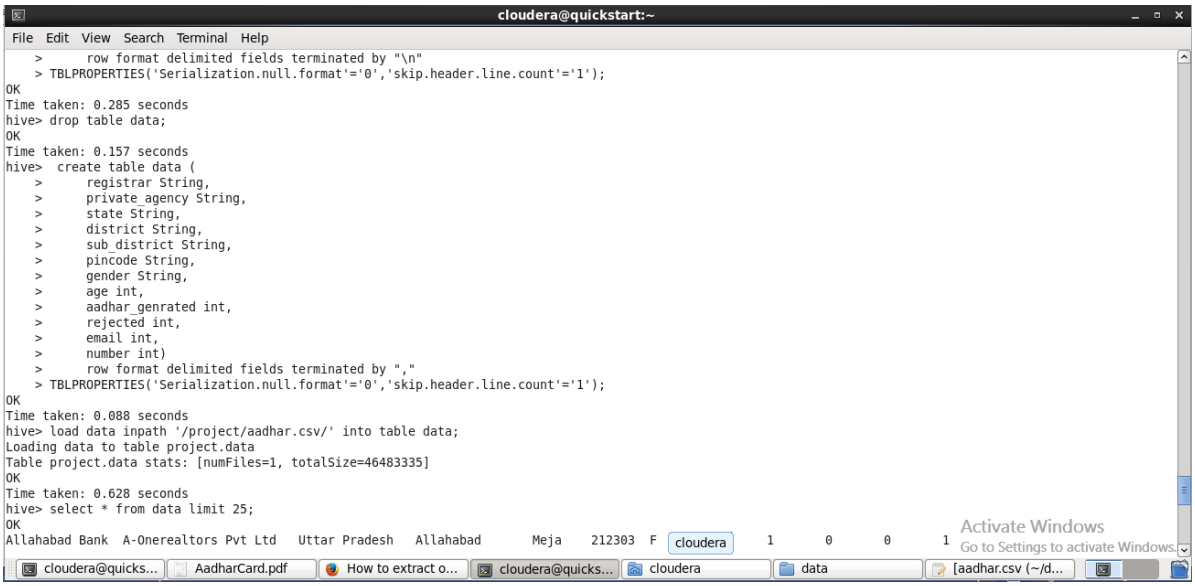
email int,

number int)

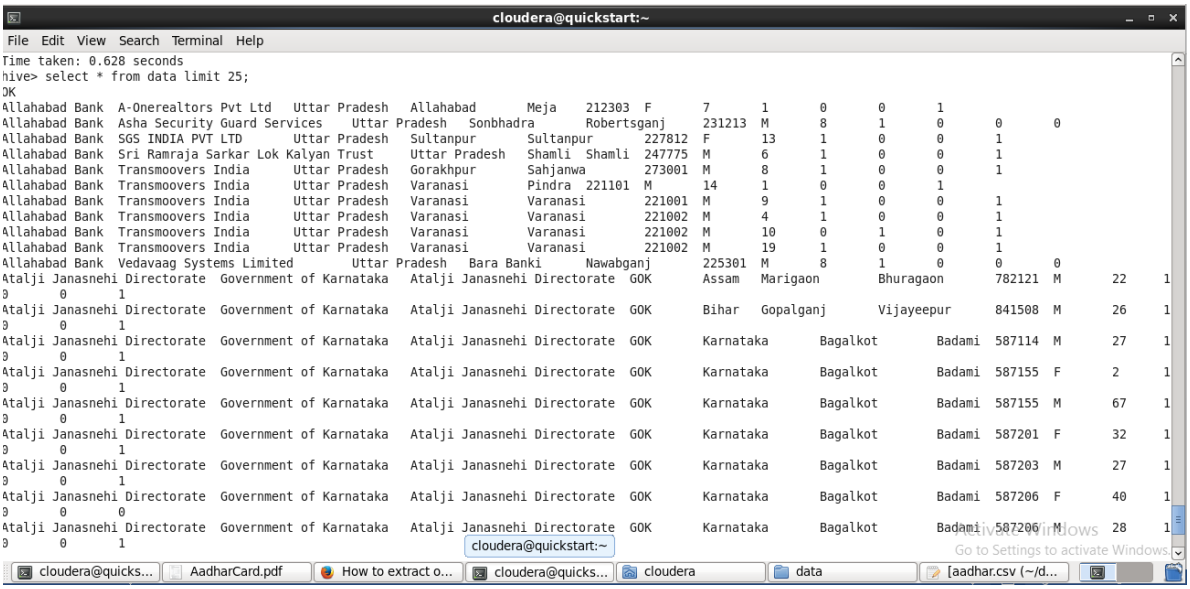
row format delimited fields terminated by ","

TBLPROPERTIES('Serialization.null.format'='0','skip.header.line.count'='1');

load data inpath '/project/aadhar.csv/' into table data;



hive> select * from data limit 25;



External table

Create external 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,

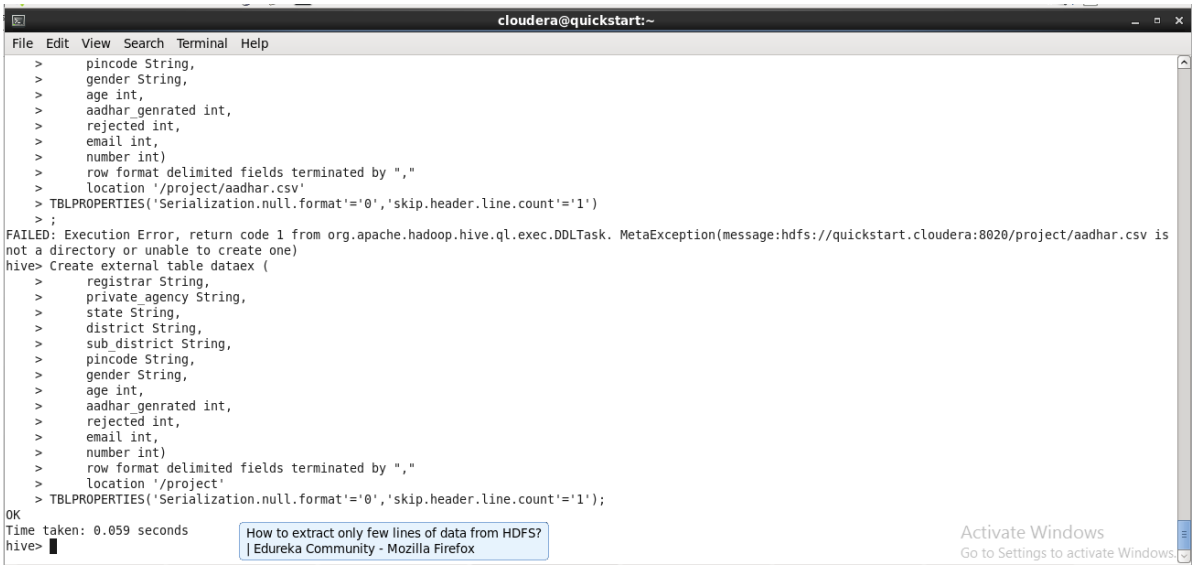
email int,

number int)

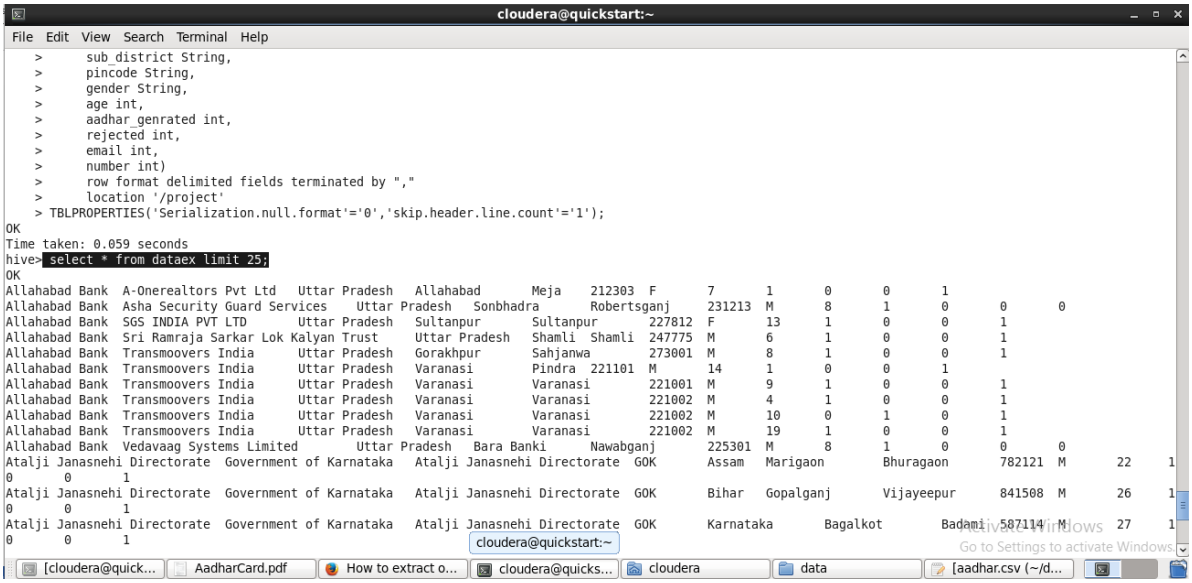
row format delimited fields terminated by ","

location '/project'

TBLPROPERTIES('Serialization.null.format'='0','skip.header.line.count'='1');



select * from dataex limit 25;



Dataframe

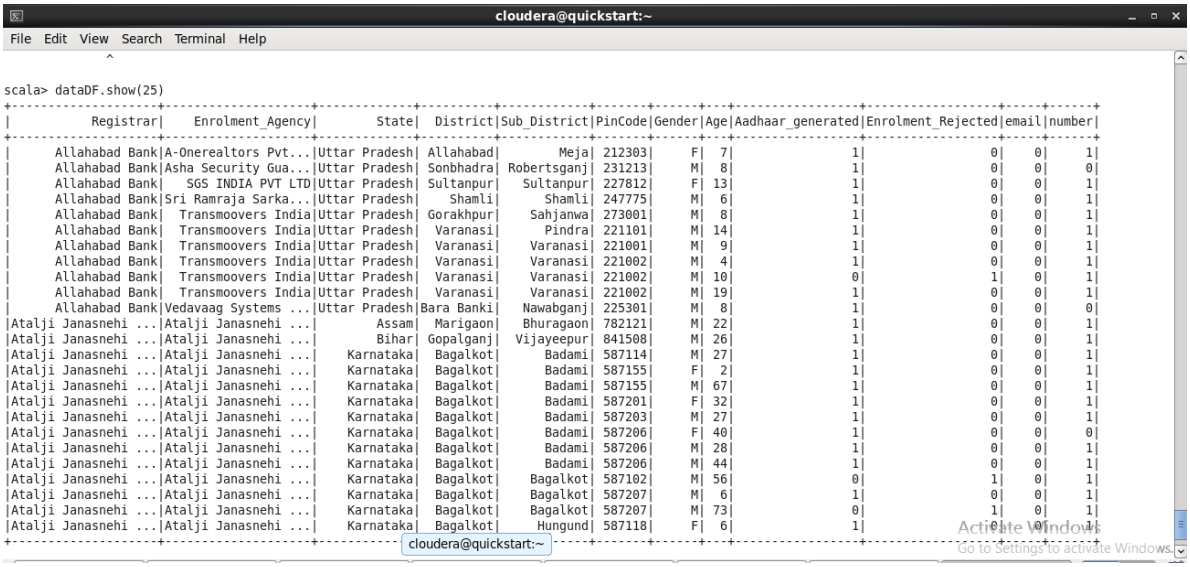
val datardd=sc.textFile("/project/aadhar.csv")

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","Pin
Code","Gender","Age","Aadhaar_generated","Enrolment_Rejected","email","number")



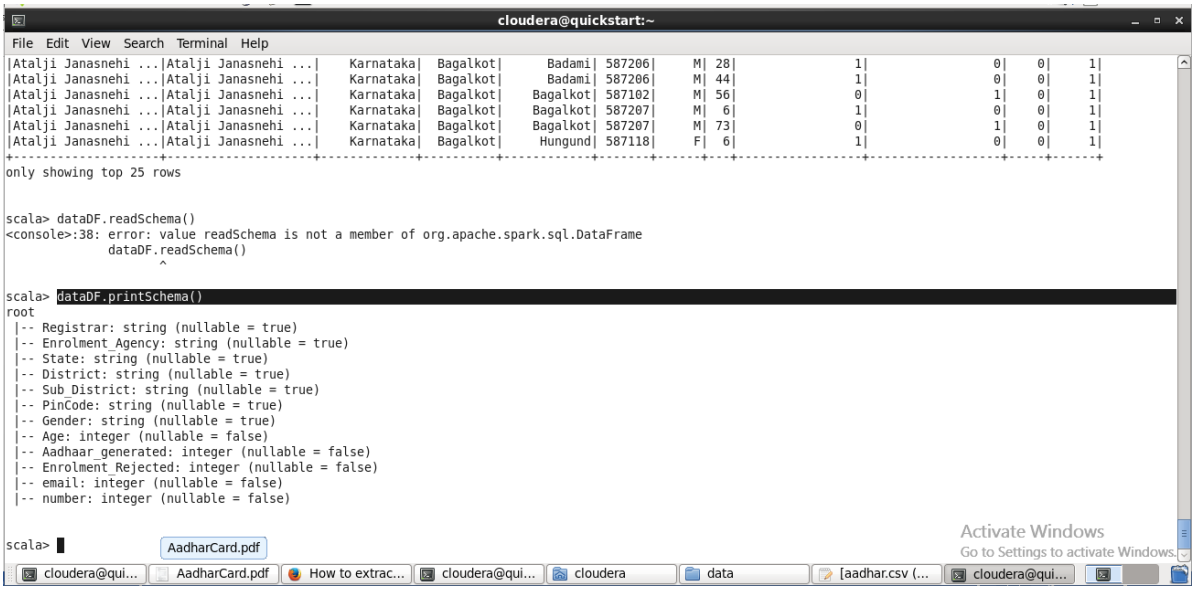
Registrar	Enrolment_Agency	State	District	Sub_District	PinCode	Gender	Age	Aadhaar_generated	Enrolment_Rejected	email	number
Allahabad Bank	A-Onerealtors Pvt...	Uttar Pradesh	Allahabad	Meja	212303	F	7	1	0	0	1
Allahabad Bank	Asha Security Gua...	Uttar Pradesh	Sonbhadra	Robertsganj	231213	M	8	1	0	0	0
Allahabad Bank	SGS INDIA PVT LTD	Uttar Pradesh	Sultanpur	Sultanpur	227812	F	13	1	0	0	1
Allahabad Bank	Sri Ramraja Sarka...	Uttar Pradesh	Shamli	Shamli	247775	M	6	1	0	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Gorakhpur	Sahjanwa	273001	M	8	1	0	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Pindra	221101	M	14	1	0	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221001	M	9	1	0	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	4	1	0	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	10	0	1	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	19	1	0	0	1
Allahabad Bank	Vedavaag Systems ...	Uttar Pradesh	Bara Banki	Nawabganj	225301	M	8	1	0	0	0
Atalji Janasnehi ...	Atalji Janasnehi ...	Assam	Marigaon	Bhuragaon	782121	M	22	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Bihar	Gopalganj	Vijayeepur	841508	M	26	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587114	M	27	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587155	F	2	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587155	M	67	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587201	F	32	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587203	M	27	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587206	F	40	1	0	0	0
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587206	M	28	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Badami	587206	M	44	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Bagalkot	587102	M	56	0	1	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Bagalkot	587207	M	6	1	0	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Bagalkot	587207	M	73	0	1	0	1
Atalji Janasnehi ...	Atalji Janasnehi ...	Karnataka	Bagalkot	Hungund	587118	F	6	1	0	0	1

Checkpoint 2

Q1:Describe the schema.

Solution:

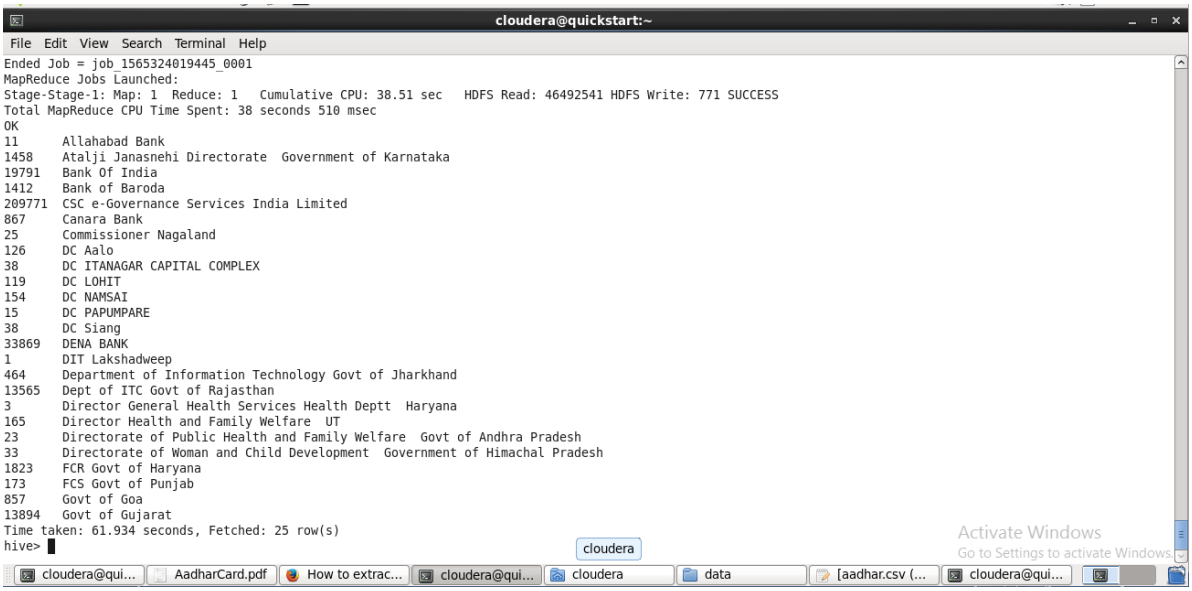
dataDF.printSchema()



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

Solution:

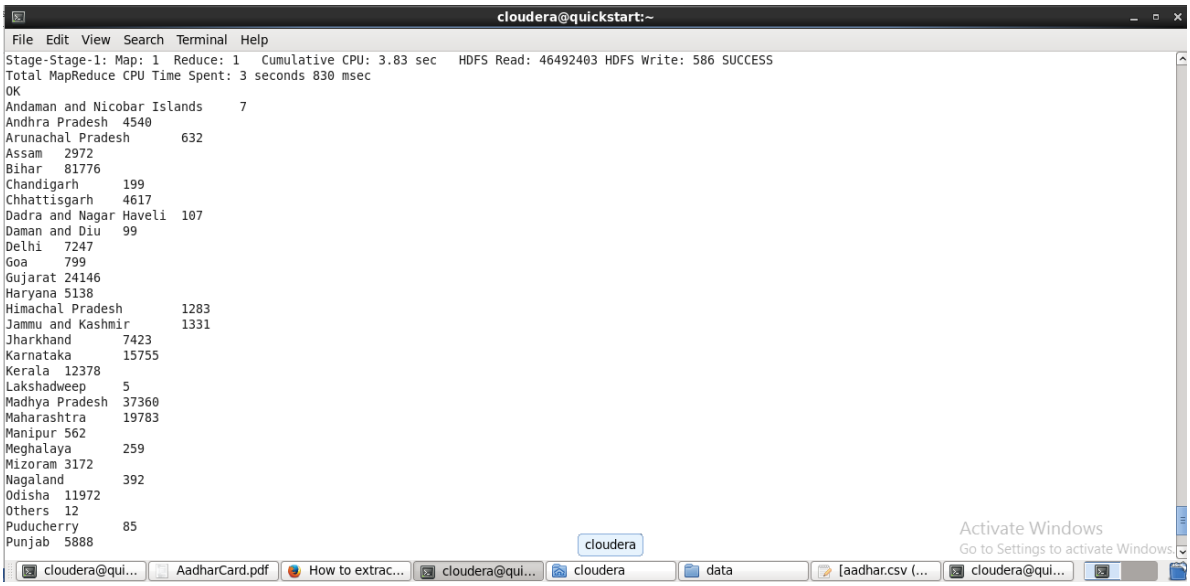
select count(registrar),registrar from data group by registrar limit 25;



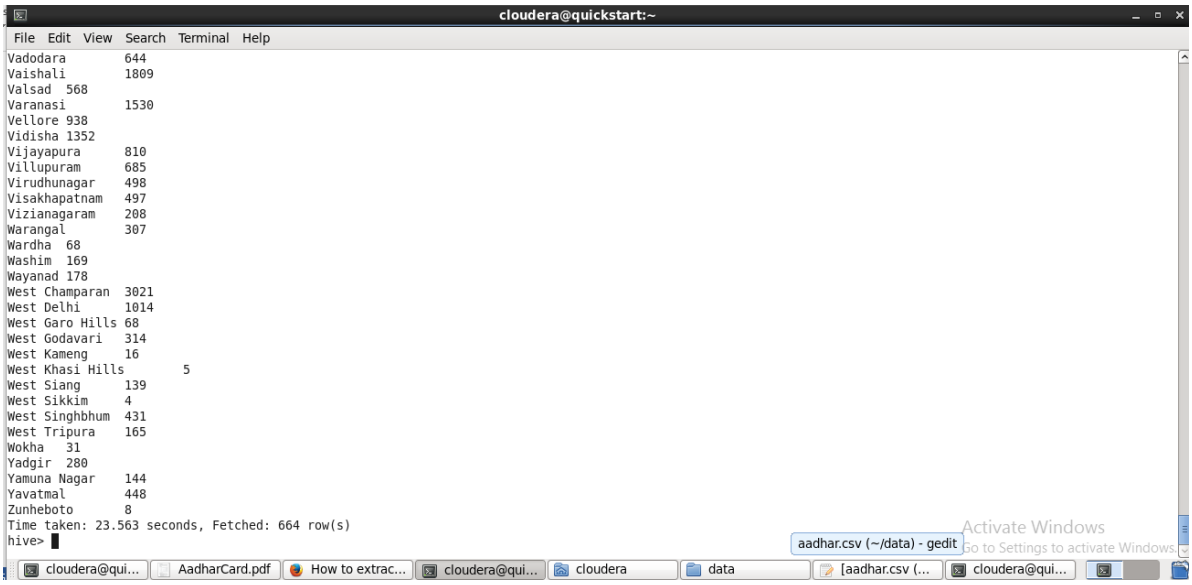
Q4: Find the number of states, districts in each state and sub-districts in each district.

Solution:

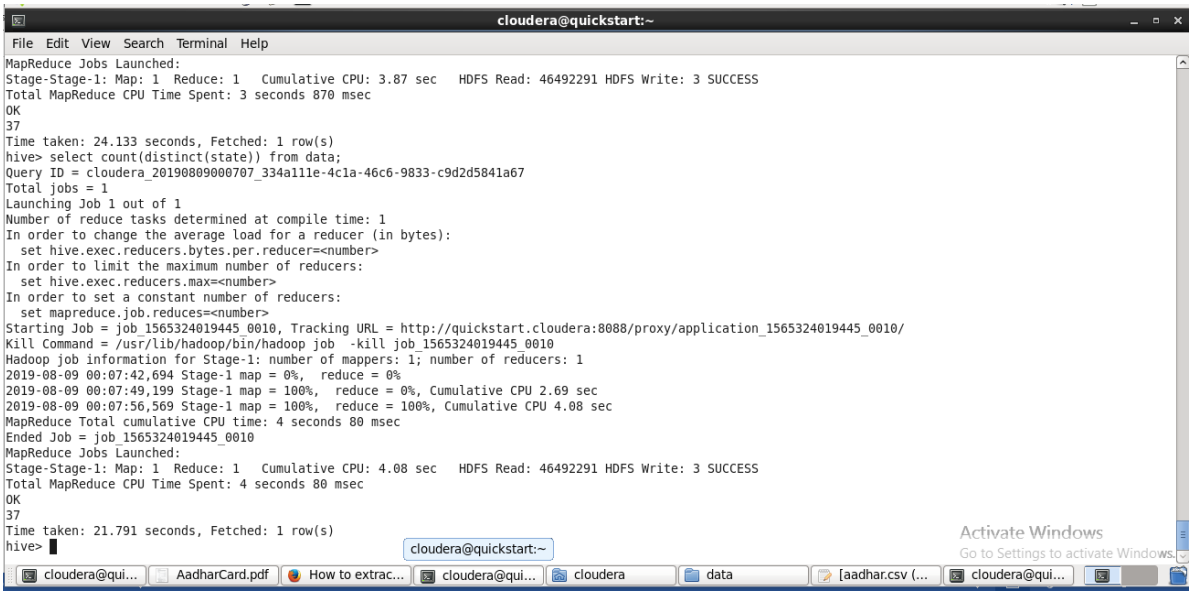
select state,count(district) from data group by state;



select district,count(sub_district) from data group by district;



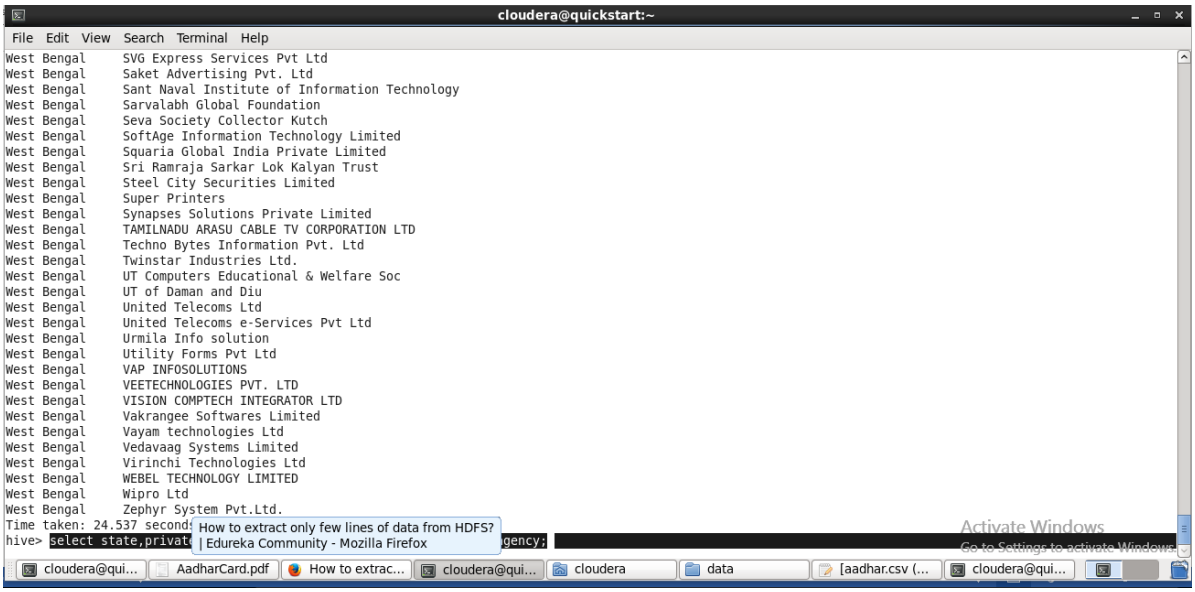
select distinct(state) from data;



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

Solution:

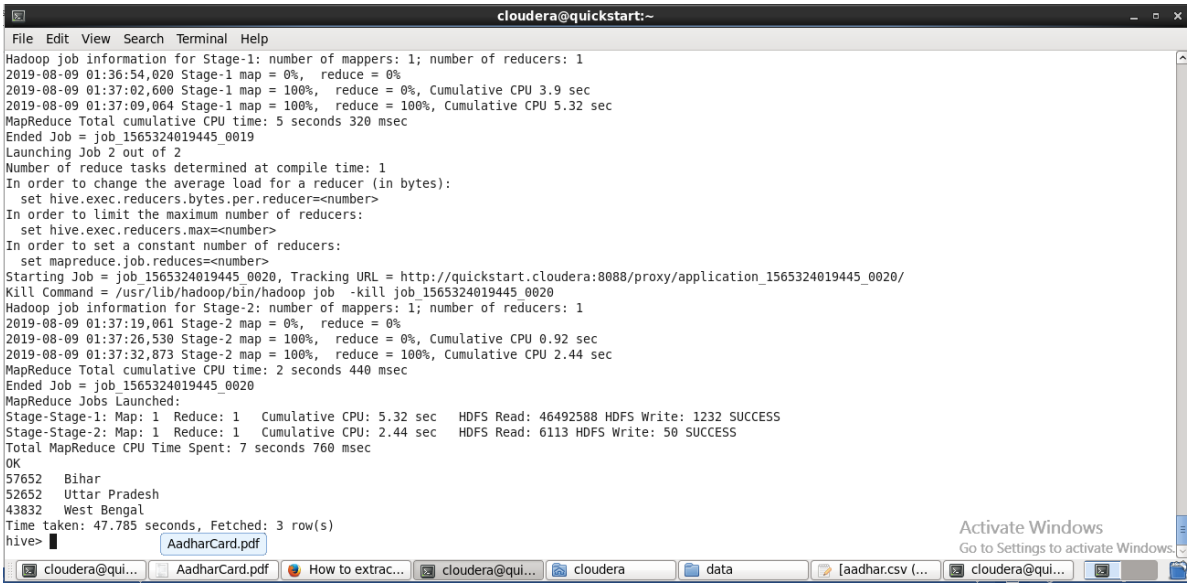
select state,private_agency from data group by state,private_agency;



Q8. Find top 3 states generating most number of Aadhaar cards?

Solution:

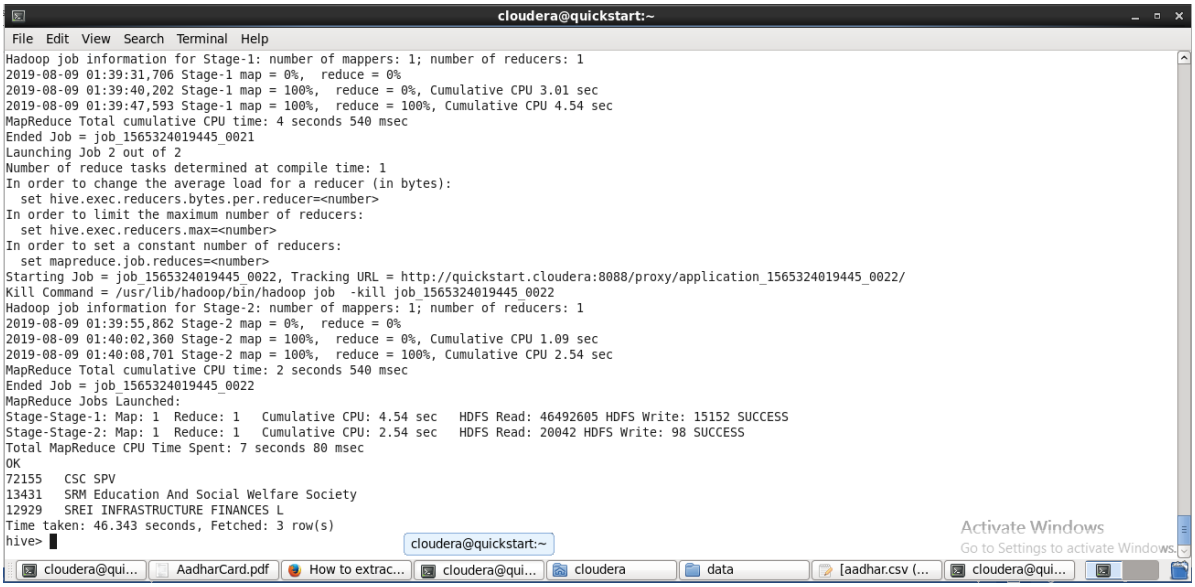
select count(*)as no,state from data where aadhar_genrated=1 group by state order by no desc limit(3);



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

Solution:

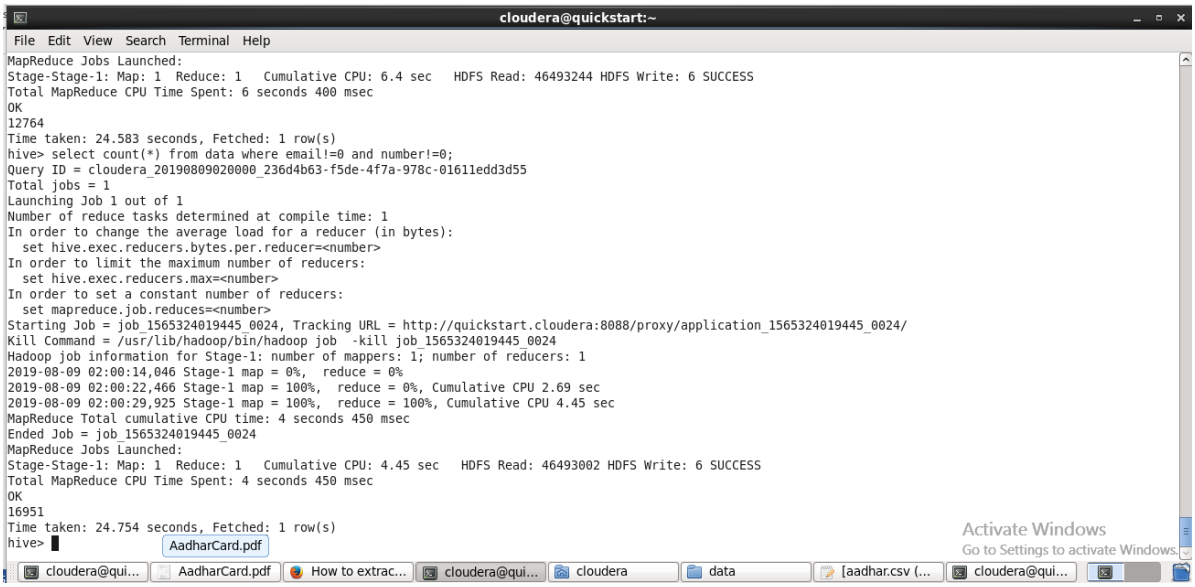
select count(*)as no,private_agency from data where aadhar_genrated=1 group by private_agency order by no desc limit 3;



Q10. Find the number of residents providing email, mobile number? (Hint: consider non-zero values.)

Solution:

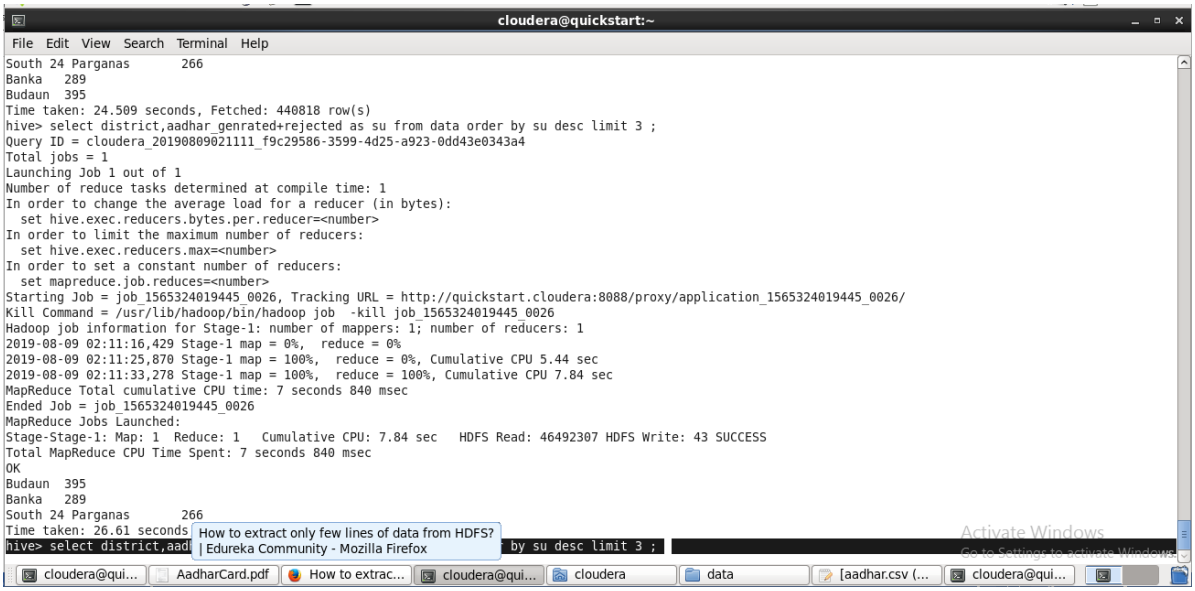
hive> select count(*) from data where email!=0 and number!=0;



Q11:Find top 3 districts where enrolment numbers are maximum?

Solution:

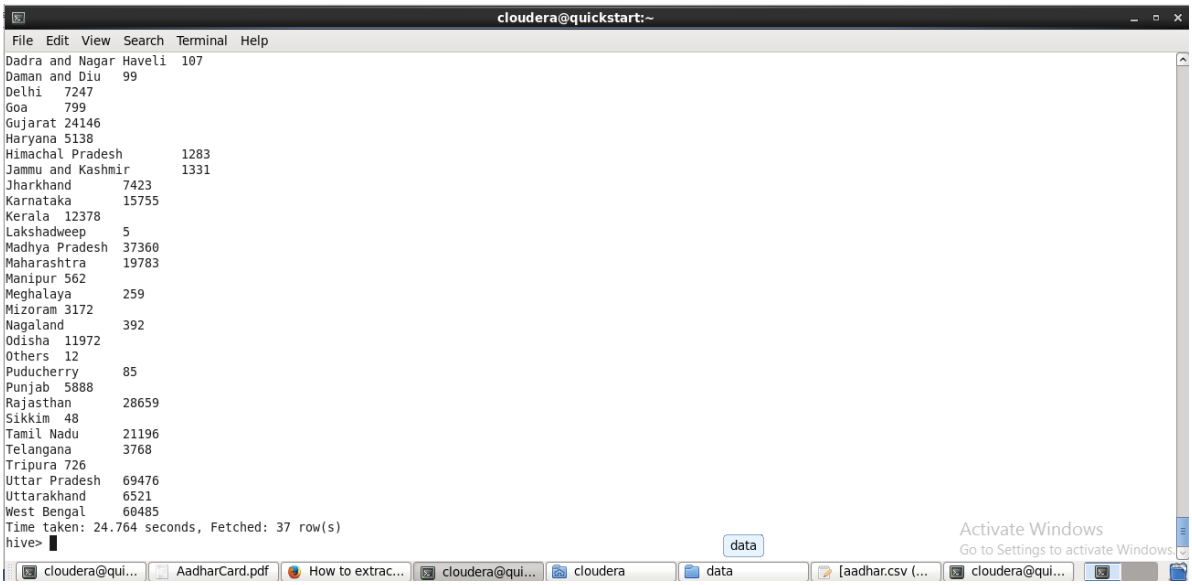
select district,aadhar_genrated+rejected as su from data order by su desc limit 3 ;



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

Solution:

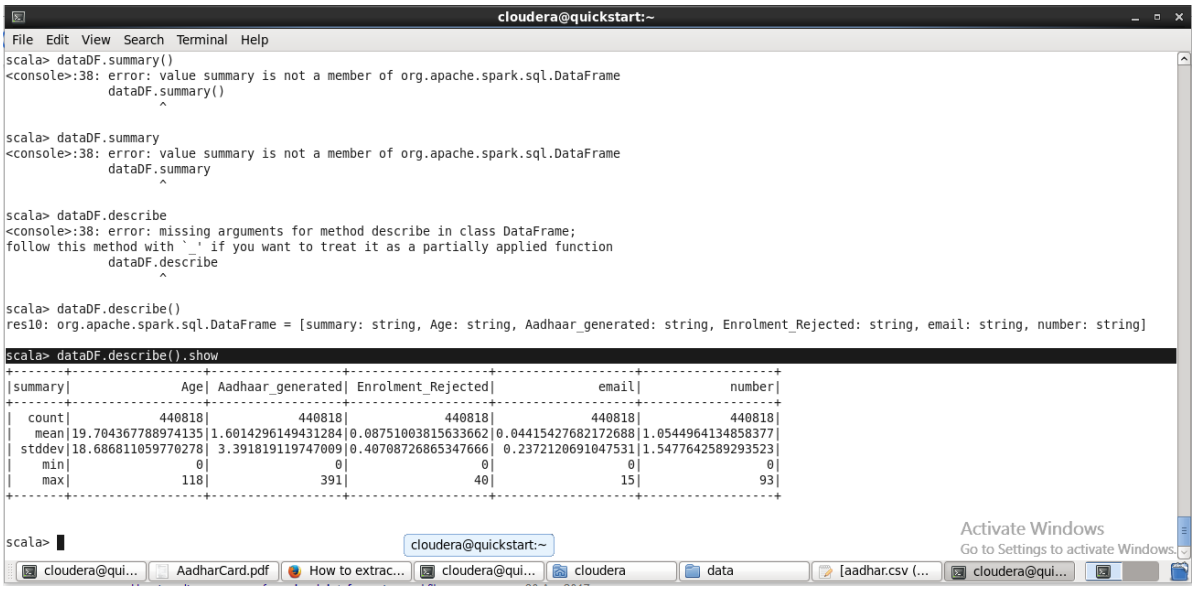
hive> select state,count(aadhar_genrated)as no from data group by state ;



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

Solution:

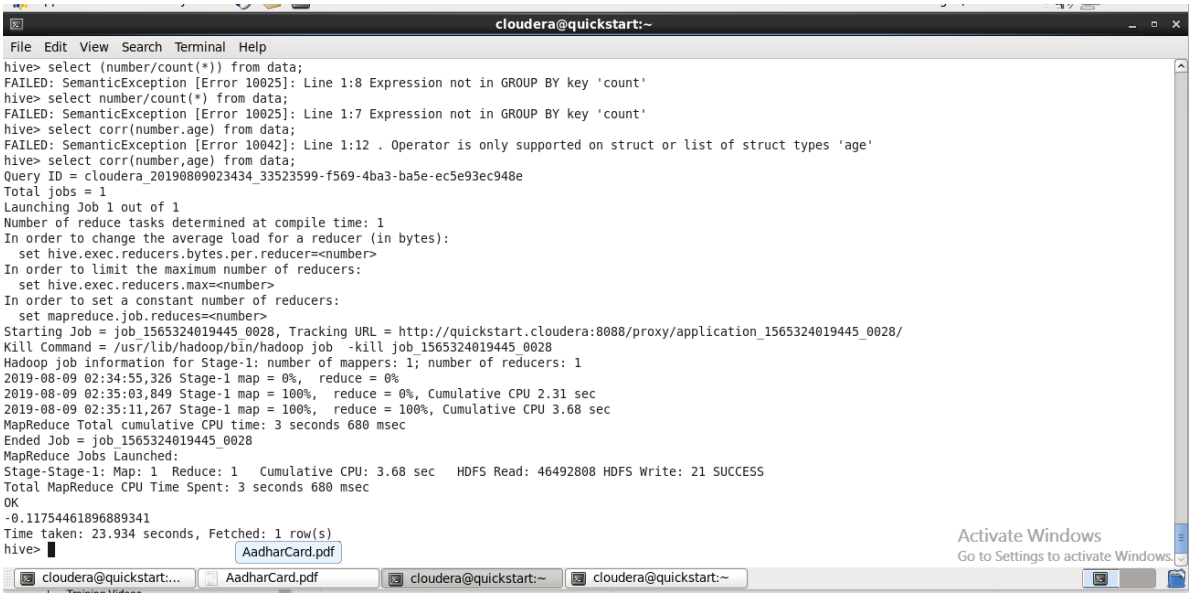
scala> dataDF.describe().show



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)

Solution:

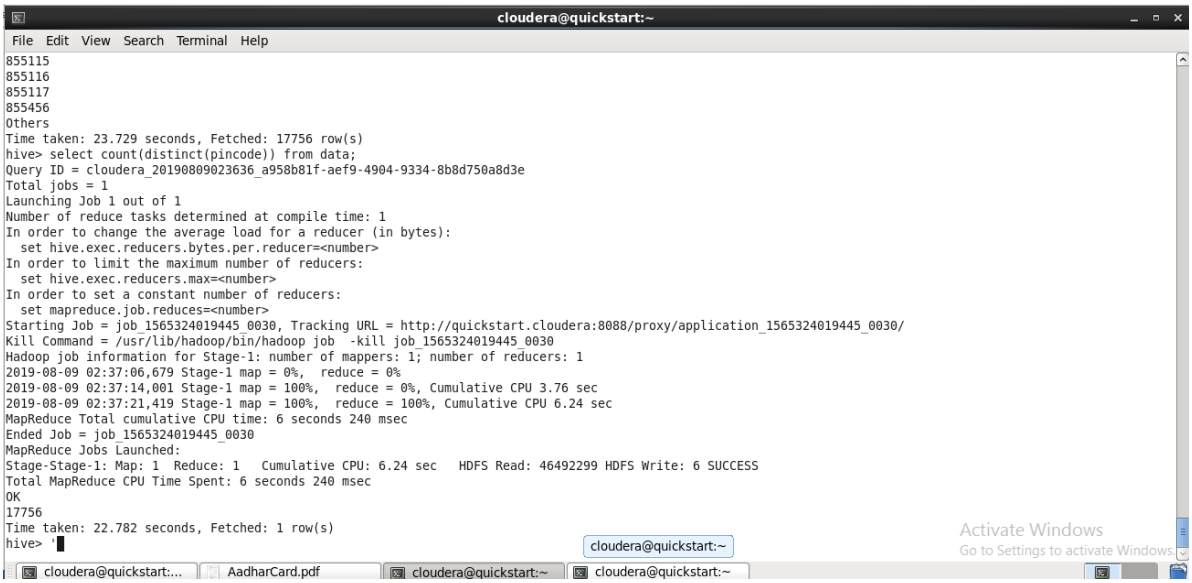
select corr(number,age) from data;



Q.15 Find the number of unique pincodes in the data?

Solution:

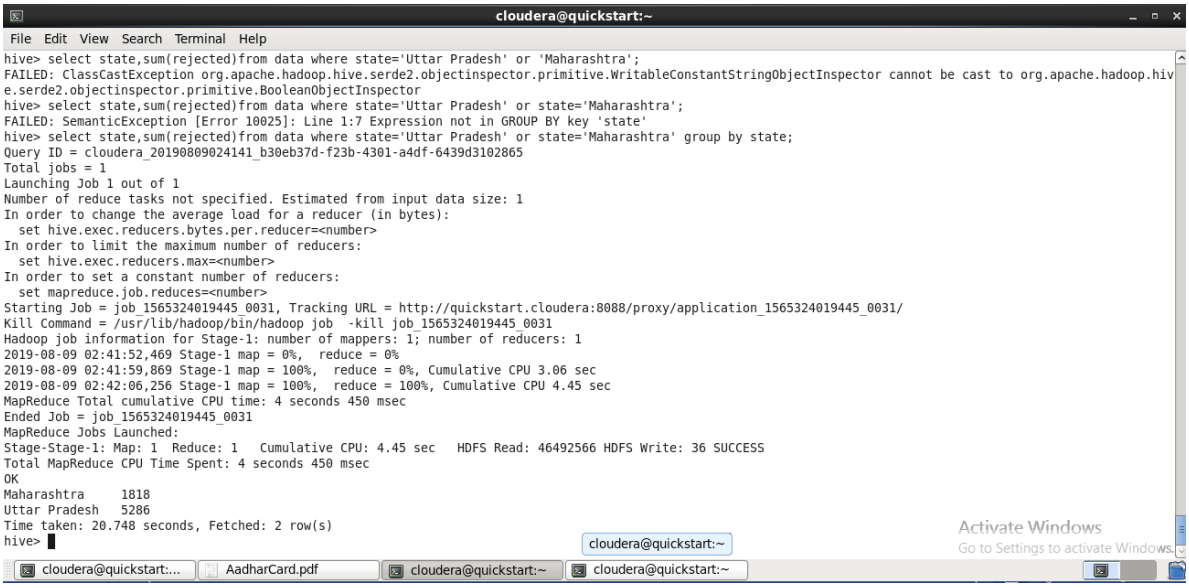
Select count(distinct(pincode)) from data;



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

Solution:

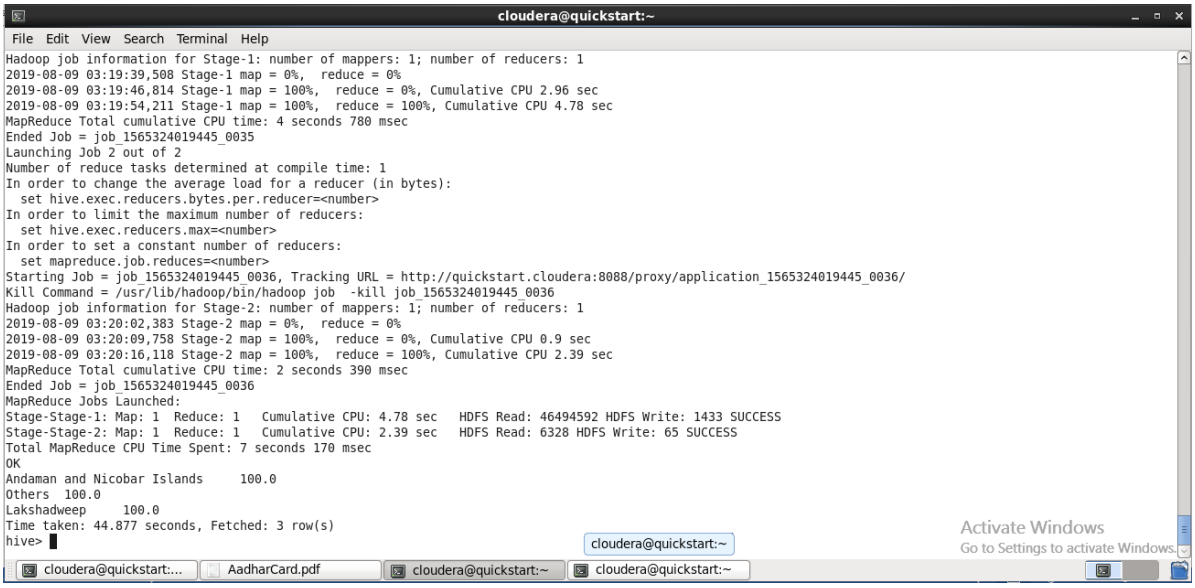
hive> select state,sum(rejected)from data where state='Uttar Pradesh' or state='Maharashtra' group by state;



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

Solution:

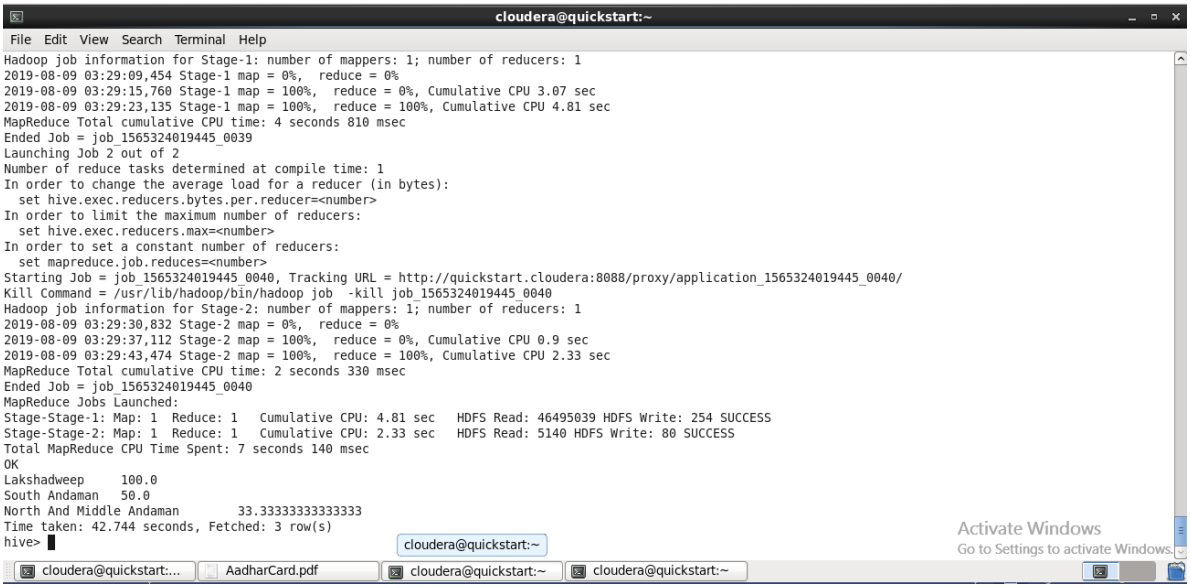
select state,(sum(aadhar_genrated)/sum(aadhar_genrated+rejected)*100) as percent from data where gender="M" group by state order by percent desc limit 3;



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.

Solution:

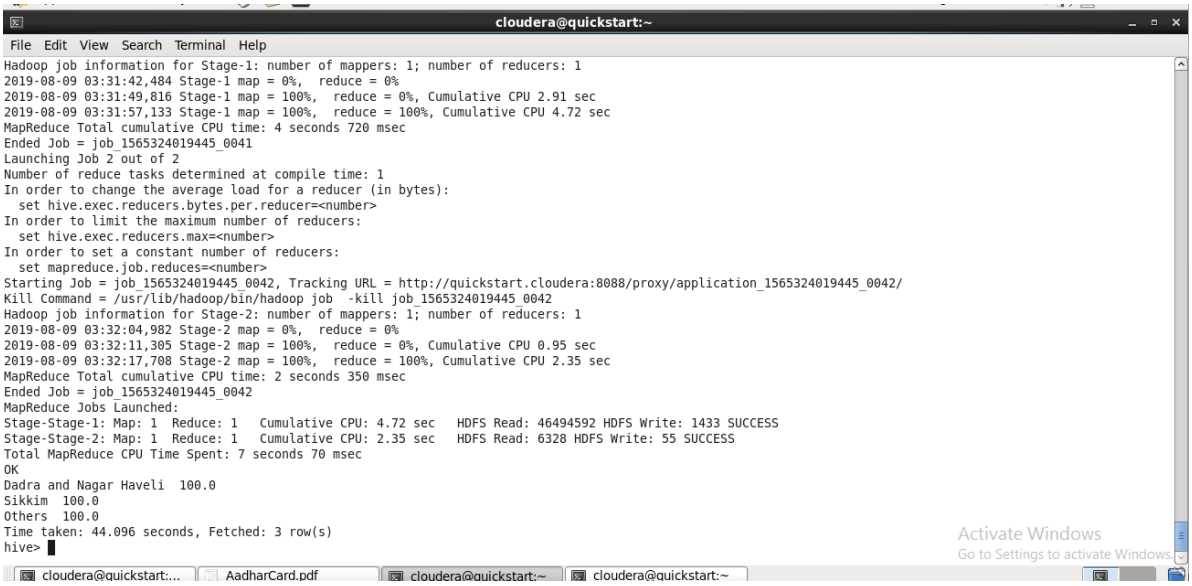
select district,(sum(rejected)/sum(aadhar_genrated+rejected)*100) as perc from data where gender="F" and state in("Andaman and Nicobar Islands","Others","Lakshadweep") group by district order by perc desc limit 3;



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

Solution:

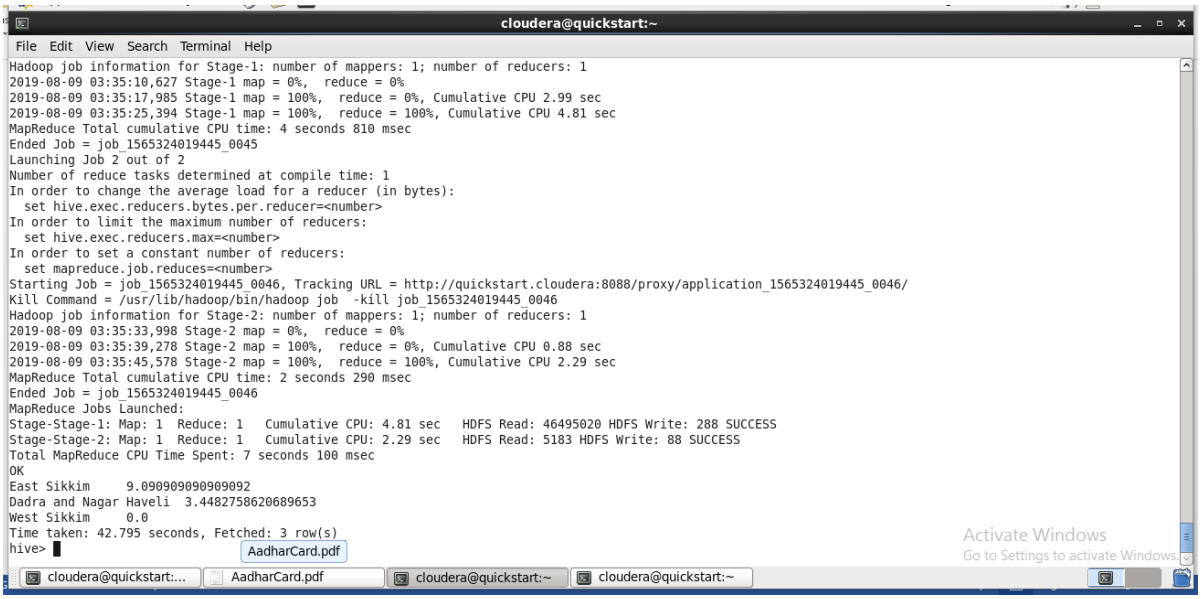
select state,(sum(aadhar_genrated)/sum(aadhar_genrated+rejected)*100) as percent from data where gender="F" group by state order by percent desc limit 3;



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.

Solution

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.

Solution:

```
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;