

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).

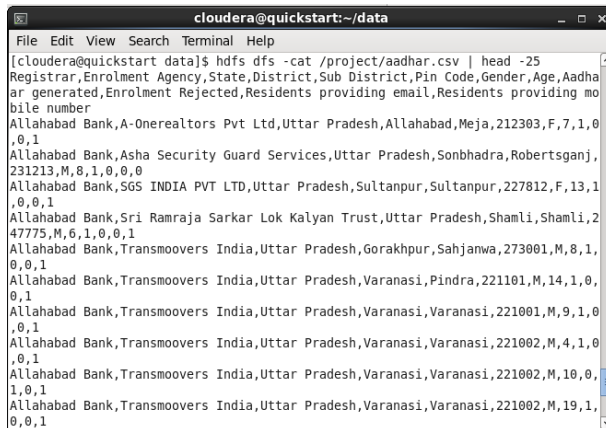
Solution

Load into hdfs

```
hdfs dfs -put aadhar.csv /project/
```

Printing

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



```
cloudera@quickstart:~/data
File Edit View Search Terminal Help
[cloudera@quickstart data]$ hdfs dfs -cat /project/aadhar.csv | head -25
Registrar,Enrolment Agency,State,District,Sub District,Pin Code,Gender,Age,Aadhaar generated,Enrolment Rejected,Residents providing email,Residents providing mobile number
Allahabad Bank,A-Onerealtors Pvt Ltd,Uttar Pradesh,Allahabad,Meja,212303,F,7,1,0,0,1
Allahabad Bank,Asha Security Guard Services,Uttar Pradesh,Sonbhadra,Robertsganj,231213,M,8,1,0,0,0
Allahabad Bank,$GS INDIA PVT LTD,Uttar Pradesh,Sultanpur,Sultanpur,227812,F,13,1,0,0,1
Allahabad Bank,Sri Ramaja Sarkar Lok Kalyan Trust,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
```

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,

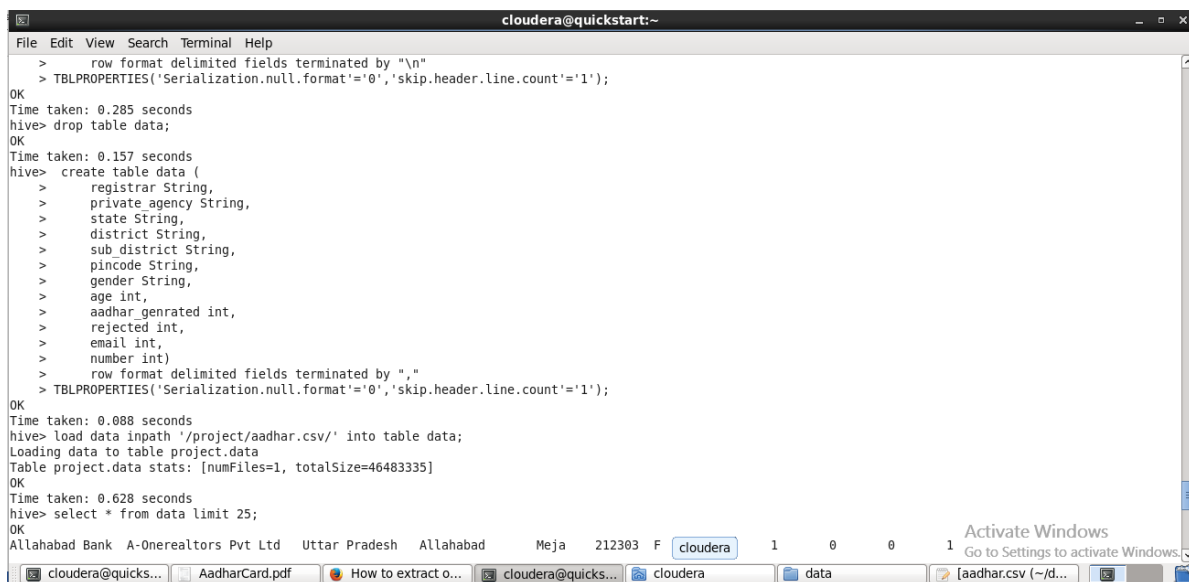
email int,

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;



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
> row format delimited fields terminated by "\n"  
> TBLPROPERTIES('Serialization.null.format'='0','skip.header.line.count'='1');  
OK  
Time taken: 0.285 seconds  
hive> drop table data;  
OK  
Time taken: 0.157 seconds  
hive> 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');  
OK  
Time taken: 0.088 seconds  
hive> load data inpath '/project/aadhar.csv/' into table data;  
Loading data to table project.data  
Table project.data stats: [numFiles=1, totalSize=46483335]  
OK  
Time taken: 0.628 seconds  
hive> select * from data limit 25;  
OK  
Allahabad Bank A-Onerealtors Pvt Ltd Uttar Pradesh Allahabad Meja 212303 F cloudera 1 0 0 1  
Activate Windows  
Go to Settings to activate Windows.  
cloudera@quicks... AadharCard.pdf How to extract o... cloudera@quicks... cloudera data [aadhar.csv (~d...
```

```
hive> select * from data limit 25;
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Time taken: 0.628 seconds  
hive> select * from data limit 25;  
OK  
Allahabad Bank A-Onerealtors Pvt Ltd Uttar Pradesh Allahabad Meja 212303 F 7 1 0 0 1  
Allahabad Bank Asha Security Guard Services 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 Sarkar Lok Kalyan Trust 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 Limited Uttar Pradesh Bara Banki Nawabganj 225301 M 8 1 0 0 0  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Assam Marigaon Bhuragaon 782121 M 22 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Bihar Gopalganj Vijayeeepur 841508 M 26 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587114 M 27 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587155 F 2 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587155 M 67 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587201 F 32 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587203 M 27 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587206 F 40 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587206 M 28 1
```

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');

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
> pincode String,  
> gender String,  
> age int,  
> aadhar_generated int,  
> rejected int,  
> email int,  
> number int)  
> row format delimited fields terminated by ","  
> location '/project/aadhar.csv'  
> TBLPROPERTIES('Serialization.null.format'='0','skip.header.line.count'='1')  
>  
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:hdfs://quickstart.cloudera:8020/project/aadhar.csv is  
not a directory or unable to create one)  
hive> Create external table dataex (  
> registrar String,  
> private_agency String,  
> state String,  
> district String,  
> sub_district String,  
> pincode String,  
> gender String,  
> age int,  
> aadhar_generated 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');  
OK  
Time taken: 0.059 seconds  
hive> |
```

select * from dataex limit 25;

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
> sub_district String,  
> pincode String,  
> gender String,  
> age int,  
> aadhar_generated 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');  
OK  
Time taken: 0.059 seconds  
hive> select * from dataex limit 25;  
OK  
Allahabad Bank A-Onerealtors Pvt Ltd Uttar Pradesh Allahabad Meja 212303 F 7 1 0 0 1 0 0  
Allahabad Bank Asha Security Guard Services Uttar Pradesh Sonbhadra Robertsganj 231213 M 8 1 0 0 0 0  
Allahabad Bank SGS INDIA PVT LTD Uttar Pradesh Sultanpur Sultanpur 227812 F 13 1 0 0 1 0  
Allahabad Bank Sri Ramraja Sarkar Lok Kalyan Trust Uttar Pradesh Sahjanwa Sahjanwa 247775 M 6 1 0 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Gorakhpur Pindra 273001 M 8 1 0 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221101 M 14 1 0 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221001 M 9 1 0 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221002 M 4 1 0 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221002 M 10 0 1 0 1 0  
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221002 M 19 1 0 0 1 0  
Allahabad Bank Vedavaag Systems Limited Uttar Pradesh Bara Banki Nawabganj 225301 M 8 1 0 0 0 0  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Assam Marigaon Bhuragaon 782121 M 22 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Bihar Gopalganj Vijayeeupur 841508 M 26 1  
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Badami 587114 M 27 1  
cloudera@quickstart:~
```

Dataframe

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

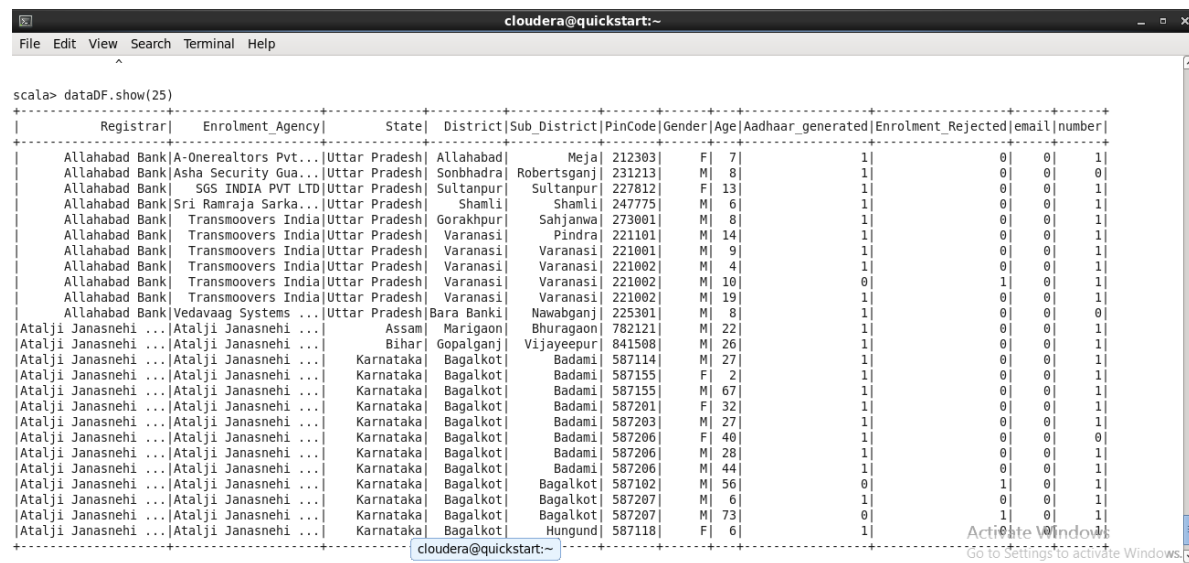
```
val h=datardd.first()
```

```
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")
```



```
scala> dataDF.show(25)
```

Registrar	Enrolment_Agency	State	District	Sub_District	PinCode	Gender	Age	Aadhaar_generated	Enrolment_Rejected	email	number
Allahabad Bank	A-One Realtors 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	Vijayepur	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:

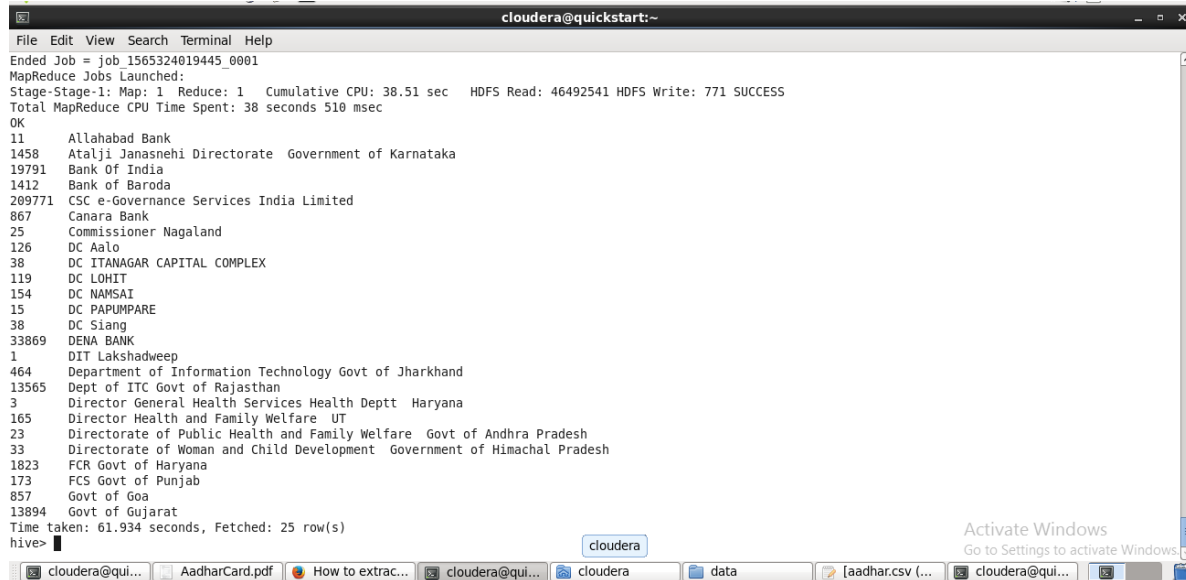
```
dataDF.printSchema()
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[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|  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
only showing top 25 rows  
  
scala> dataDF.readSchema()  
<console>:38: error: value readSchema is not a member of org.apache.spark.sql.DataFrame  
dataDF.readSchema()  
      ^  
  
scala> dataDF.printSchema()  
root  
|-- Registrar: string (nullable = true)  
|-- Enrolment Agency: string (nullable = true)  
|-- State: string (nullable = true)  
|-- District: string (nullable = true)  
|-- Sub District: string (nullable = true)  
|-- PinCode: string (nullable = true)  
|-- Gender: string (nullable = true)  
|-- Age: integer (nullable = false)  
|-- Aadhaar generated: integer (nullable = false)  
|-- Enrolment Rejected: integer (nullable = false)  
|-- email: integer (nullable = false)  
|-- number: integer (nullable = false)  
  
scala>   
AadharCard.pdf  
cloudera@qui... AadharCard.pdf How to extrac... cloudera@qui... cloudera data [aadhar.csv (... cloudera@qui...  
Activate Windows  
Go to Settings to activate Windows.
```

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

Solution:

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

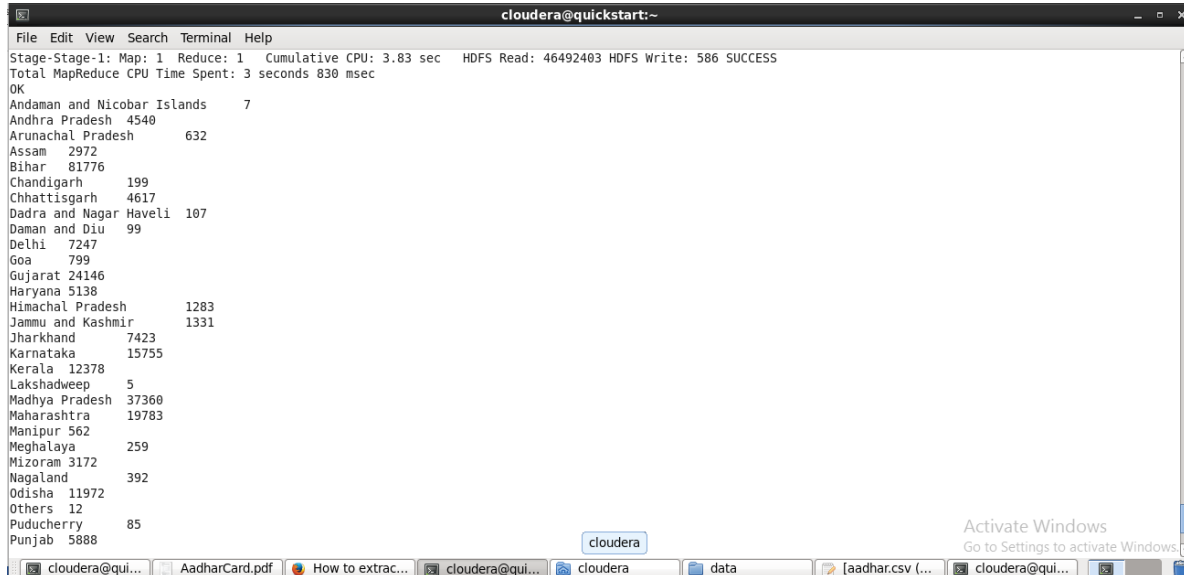


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Ended Job = job_1565324019445_0001  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 38.51 sec HDFS Read: 46492541 HDFS Write: 771 SUCCESS  
Total MapReduce CPU Time Spent: 38 seconds 510 msec  
OK  
11 Allahabad Bank  
1458 Atalji Janasnehi Directorate Government of Karnataka  
19791 Bank Of India  
1412 Bank of Baroda  
209771 CSC e-Governance Services India Limited  
867 Canara Bank  
25 Commissioner Nagaland  
126 DC Aalo  
38 DC ITANAGAR CAPITAL COMPLEX  
119 DC LOHIT  
154 DC NAMSAI  
15 DC PAPUMPARE  
38 DC Siang  
33869 DENA BANK  
1 DIT Lakshadweep  
464 Department of Information Technology Govt of Jharkhand  
13565 Dept of ITC Govt of Rajasthan  
3 Director General Health Services Health Deptt Haryana  
165 Director Health and Family Welfare UT  
23 Directorate of Public Health and Family Welfare Govt of Andhra Pradesh  
33 Directorate of Woman and Child Development Government of Himachal Pradesh  
1823 FCR Govt of Haryana  
173 FCS Govt of Punjab  
857 Govt of Goa  
13894 Govt of Gujarat  
Time taken: 61.934 seconds, Fetched: 25 row(s)  
hive>
```

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

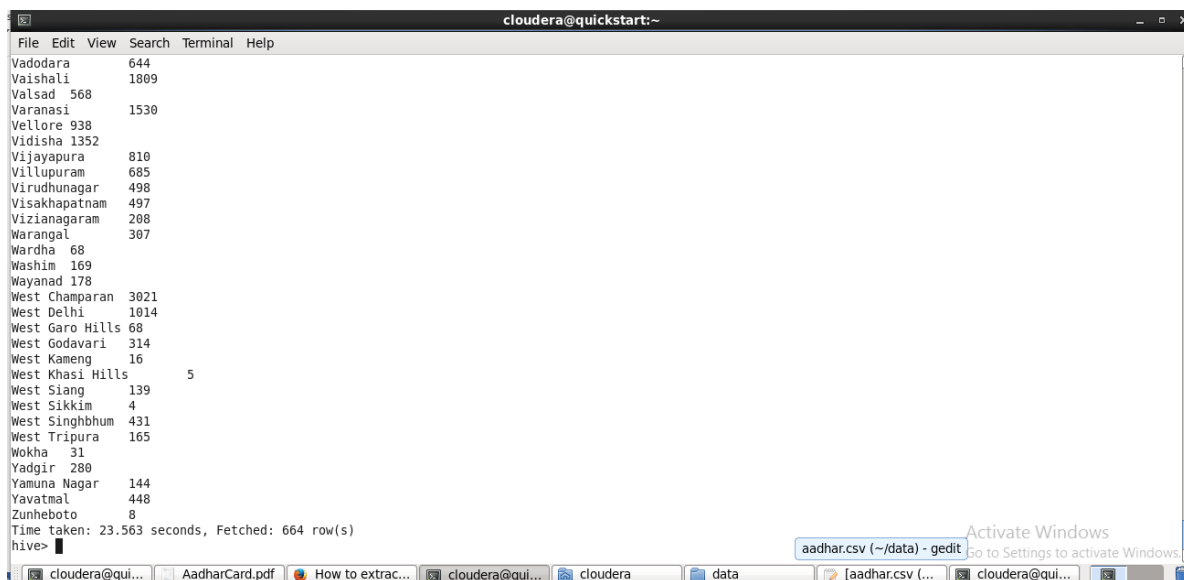
Solution:

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



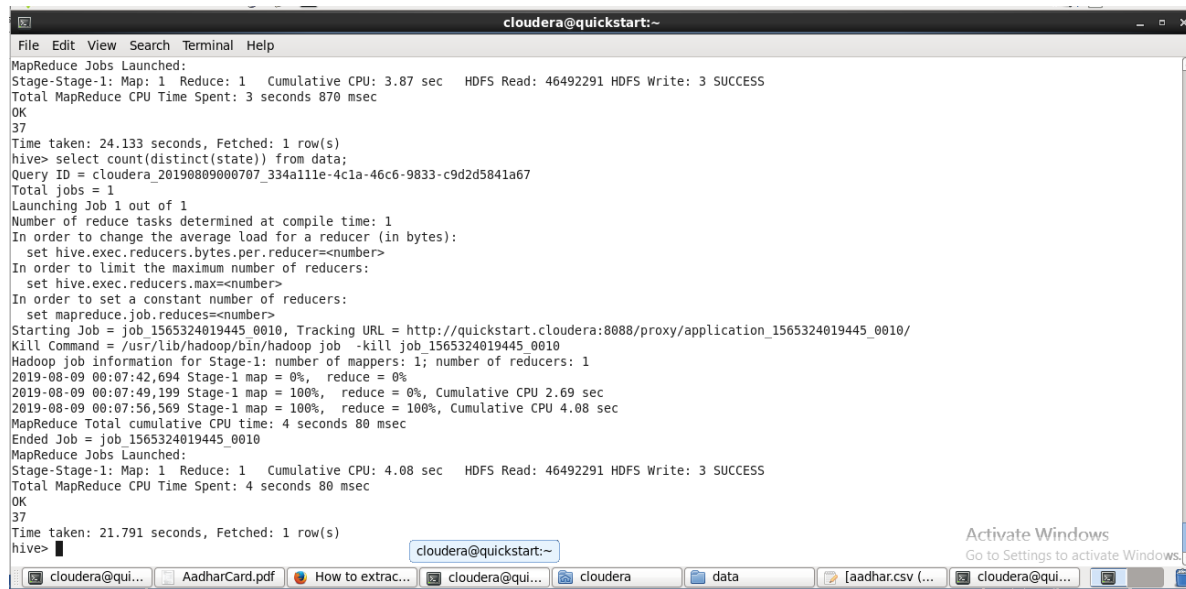
```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.83 sec HDFS Read: 46492403 HDFS Write: 586 SUCCESS  
Total MapReduce CPU Time Spent: 3 seconds 830 msec  
OK  
Andaman and Nicobar Islands 7  
Andhra Pradesh 4540  
Arunachal Pradesh 632  
Assam 2972  
Bihar 81776  
Chandigarh 199  
Chhattisgarh 4617  
Dadra and Nagar Haveli 107  
Daman and Diu 99  
Delhi 7247  
Goa 799  
Gujarat 24146  
Haryana 5138  
Himachal Pradesh 1283  
Jammu and Kashmir 1331  
Jharkhand 7423  
Karnataka 15755  
Kerala 12378  
Lakshadweep 5  
Madhya Pradesh 37360  
Maharashtra 19783  
Manipur 562  
Meghalaya 259  
Mizoram 3172  
Nagaland 392  
Odisha 11972  
Others 12  
Puducherry 85  
Punjab 5888
```

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



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Vadodara 644  
Vaishali 1809  
Valsad 568  
Varanasi 1530  
Vellore 938  
Vidisha 1352  
Vijayapura 810  
Villupuram 685  
Virudhunagar 498  
Visakhapatnam 497  
Vizianagaram 208  
Warangal 307  
Wardha 68  
Washim 169  
Wayanad 178  
West Champaran 3021  
West Delhi 1014  
West Garo Hills 68  
West Godavari 314  
West Kameng 16  
West Khasi Hills 5  
West Siang 139  
West Sikkim 4  
West Singhbhum 431  
West Tripura 165  
Wokha 31  
Yadgir 280  
Yamuna Nagar 144  
Yavatmal 448  
Zunheboto 8  
Time taken: 23.563 seconds, Fetched: 664 row(s)  
hive>
```


select distinct(state) from data;



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.87 sec HDFS Read: 46492291 HDFS Write: 3 SUCCESS  
Total MapReduce CPU Time Spent: 3 seconds 870 msec  
OK  
37  
Time taken: 24.133 seconds, Fetched: 1 row(s)  
hive> select count(distinct(state)) from data;  
Query ID = cloudera_20190809000707_334a111e-4c1a-46c6-9833-c9d2d5841a67  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0010/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0010  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 00:07:42,694 Stage-1 map = 0%, reduce = 0%  
2019-08-09 00:07:49,199 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.69 sec  
2019-08-09 00:07:56,569 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.08 sec  
MapReduce Total cumulative CPU time: 4 seconds 80 msec  
Ended Job = job_1565324019445_0010  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.08 sec HDFS Read: 46492291 HDFS Write: 3 SUCCESS  
Total MapReduce CPU Time Spent: 4 seconds 80 msec  
OK  
37  
Time taken: 21.791 seconds, Fetched: 1 row(s)  
hive>
```

cloudera@quickstart:~

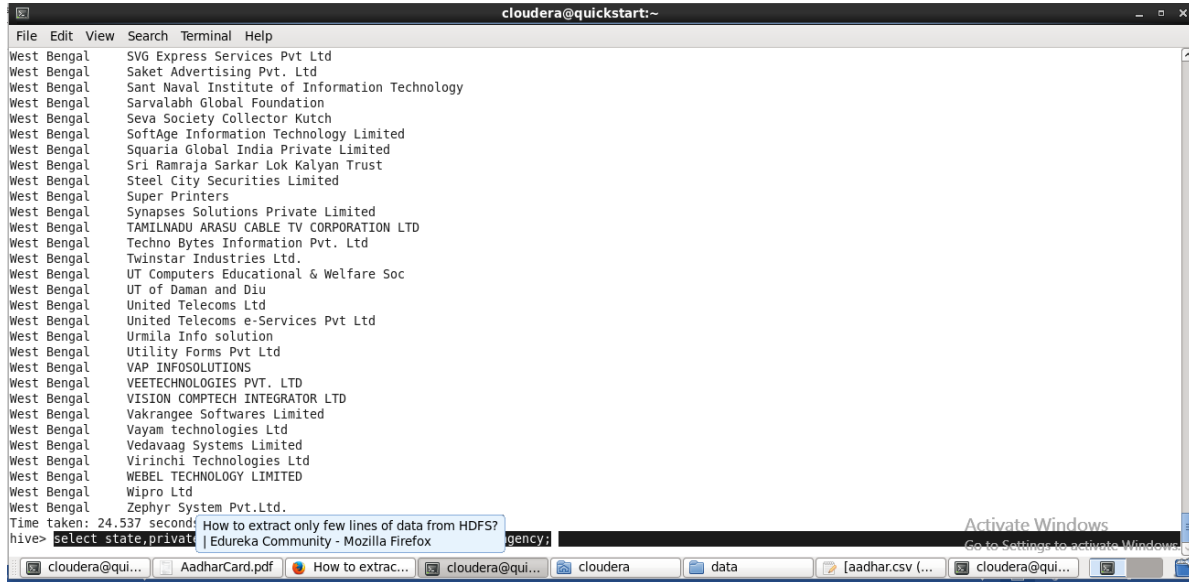
cloudera@qui... AadharCard.pdf How to extrac... cloudera@qui... cloudera data [aadhar.csv (... cloudera@qui...

Activate Windows
Go to Settings to activate Windows.

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

Solution:

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

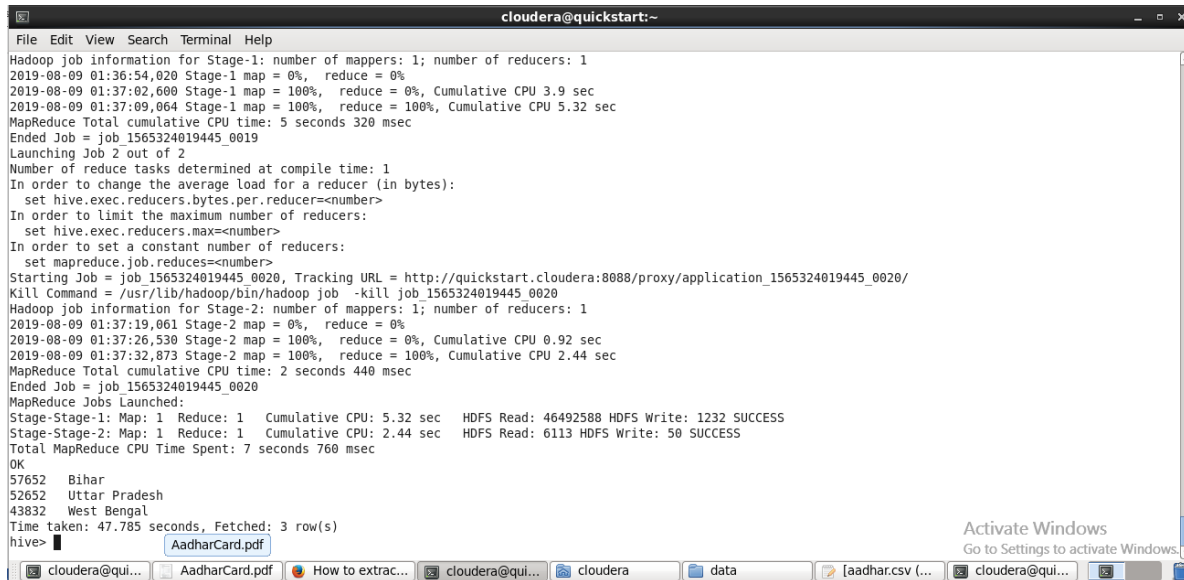


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
West Bengal SVG Express Services Pvt Ltd  
West Bengal Saket Advertising Pvt. Ltd  
West Bengal Sant Naval Institute of Information Technology  
West Bengal Sarvalabh Global Foundation  
West Bengal Seva Society Collector Kutch  
West Bengal SoftAge Information Technology Limited  
West Bengal Squaria Global India Private Limited  
West Bengal Sri Ramraja Sarkar Lok Kalyan Trust  
West Bengal Steel City Securities Limited  
West Bengal Super Printers  
West Bengal Synapses Solutions Private Limited  
West Bengal TAMILNADU ARASU CABLE TV CORPORATION LTD  
West Bengal Techno Bytes Information Pvt. Ltd  
West Bengal Twinstar Industries Ltd.  
West Bengal UT Computers Educational & Welfare Soc  
West Bengal UT of Daman and Diu  
West Bengal United Telecoms Ltd  
West Bengal United Telecoms e-Services Pvt Ltd  
West Bengal Urmila Info solution  
West Bengal Utility Forms Pvt Ltd  
West Bengal VAP INFOSOLUTIONS  
West Bengal VEETECHNOLOGIES PVT. LTD  
West Bengal VISION COMPTECH INTEGRATOR LTD  
West Bengal Vakrangee Softwares Limited  
West Bengal Vayam technologies Ltd  
West Bengal Vedavaag Systems Limited  
West Bengal Virinchi Technologies Ltd  
West Bengal WEBEL TECHNOLOGY LIMITED  
West Bengal Wipro Ltd  
West Bengal Zephyr System Pvt.Ltd.  
Time taken: 24.537 seconds; How to extract only few lines of data from HDFS?  
hive> select state,private_agency; 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);
```



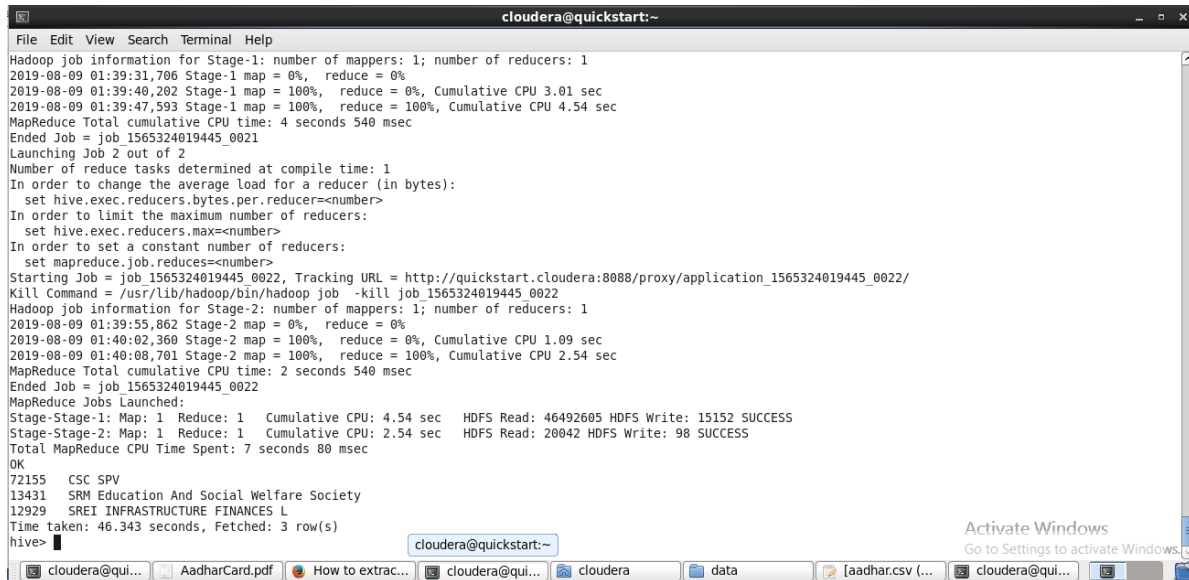
```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 01:36:54,020 Stage-1 map = 0%, reduce = 0%  
2019-08-09 01:37:02,600 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.9 sec  
2019-08-09 01:37:09,064 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.32 sec  
MapReduce Total cumulative CPU time: 5 seconds 320 msec  
Ended Job = job_1565324019445_0019  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0020/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0020  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 01:37:19,061 Stage-2 map = 0%, reduce = 0%  
2019-08-09 01:37:26,530 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.92 sec  
2019-08-09 01:37:32,873 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.44 sec  
MapReduce Total cumulative CPU time: 2 seconds 440 msec  
Ended Job = job_1565324019445_0020  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.32 sec HDFS Read: 46492588 HDFS Write: 1232 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.44 sec HDFS Read: 6113 HDFS Write: 50 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 760 msec  
OK  
57652 Bihar  
52652 Uttar Pradesh  
43832 West Bengal  
Time taken: 47.785 seconds, Fetched: 3 row(s)  
hive> █
```

Windows taskbar at the bottom shows: cloudera@qui..., AadharCard.pdf, How to extrac..., cloudera@qui..., cloudera, data, [aadhar.csv (...), cloudera@qui..., and an 'Activate Windows' watermark.

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;
```

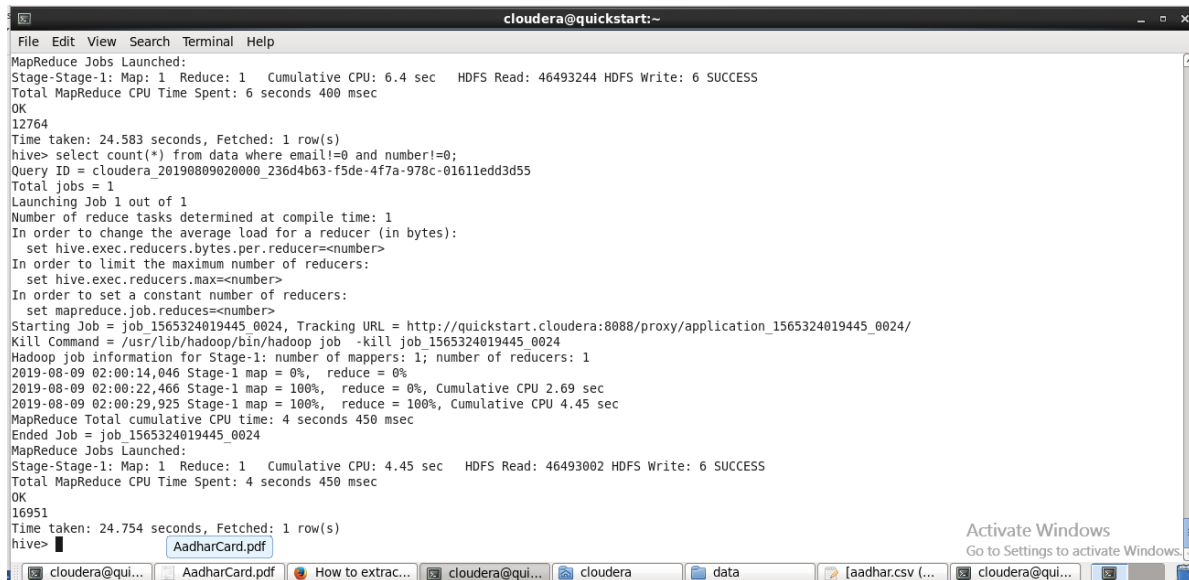


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 01:39:31,706 Stage-1 map = 0%, reduce = 0%  
2019-08-09 01:39:40,202 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.01 sec  
2019-08-09 01:39:47,593 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.54 sec  
MapReduce Total cumulative CPU time: 4 seconds 540 msec  
Ended Job = job_1565324019445_0021  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0022, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0022/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0022  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 01:39:55,862 Stage-2 map = 0%, reduce = 0%  
2019-08-09 01:40:02,360 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.09 sec  
2019-08-09 01:40:08,701 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.54 sec  
MapReduce Total cumulative CPU time: 2 seconds 540 msec  
Ended Job = job_1565324019445_0022  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.54 sec HDFS Read: 46492605 HDFS Write: 15152 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.54 sec HDFS Read: 20042 HDFS Write: 98 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 80 msec  
OK  
72155 CSC SPV  
13431 SRM Education And Social Welfare Society  
12929 SREI INFRASTRUCTURE FINANCES L  
Time taken: 46.343 seconds, Fetched: 3 row(s)  
hive>
```

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;
```



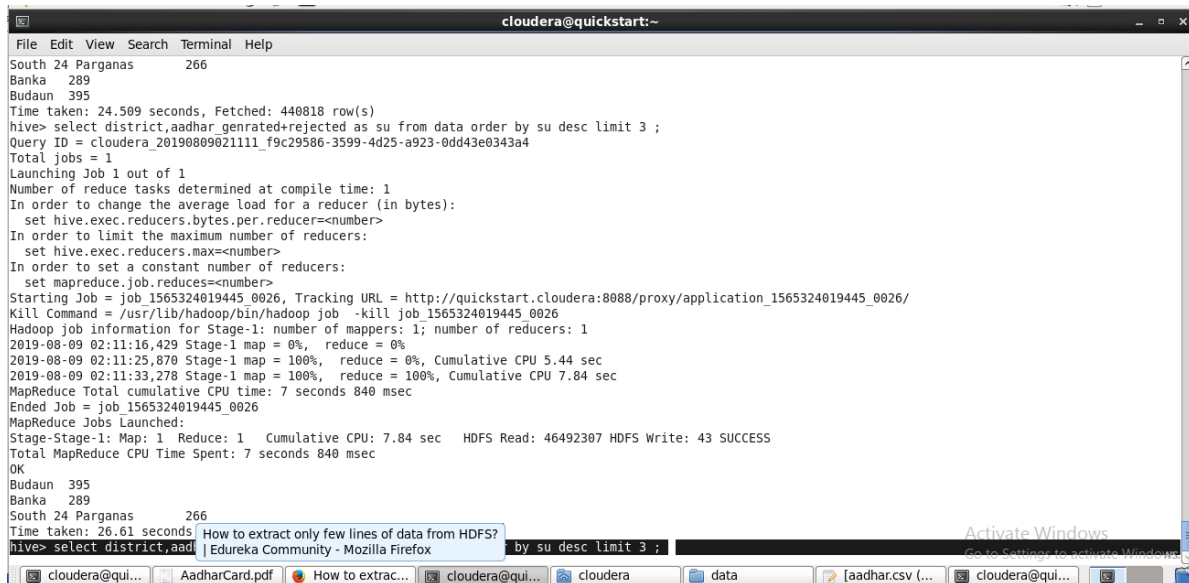
The screenshot shows a terminal window titled 'cloudera@quickstart:~'. It displays the output of a Hive query: 'select count(*) from data where email!=0 and number!=0;'. The query ID is 'cloudera_20190809020000_236d4b63-f5de-4f7a-978c-01611edd3d55'. The output shows 'Total jobs = 1' and 'Launching Job 1 out of 1'. The job information for Stage-1 shows 1 map and 1 reduce task. The cumulative CPU time is 6.4 seconds. The query result shows 1 row(s) fetched. The terminal also shows the execution of a second Hive query: 'select count(*) from data where email!=0 and number!=0;'. The query ID is 'cloudera_20190809020000_236d4b63-f5de-4f7a-978c-01611edd3d55'. The output shows 'Total jobs = 1' and 'Launching Job 1 out of 1'. The job information for Stage-1 shows 1 map and 1 reduce task. The cumulative CPU time is 4.45 seconds. The query result shows 1 row(s) fetched. The terminal also shows the execution of a third Hive query: 'select count(*) from data where email!=0 and number!=0;'. The query ID is 'cloudera_20190809020000_236d4b63-f5de-4f7a-978c-01611edd3d55'. The output shows 'Total jobs = 1' and 'Launching Job 1 out of 1'. The job information for Stage-1 shows 1 map and 1 reduce task. The cumulative CPU time is 4.45 seconds. The query result shows 1 row(s) fetched.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.4 sec HDFS Read: 46493244 HDFS Write: 6 SUCCESS  
Total MapReduce CPU Time Spent: 6 seconds 400 msec  
OK  
12764  
Time taken: 24.583 seconds, Fetched: 1 row(s)  
hive> select count(*) from data where email!=0 and number!=0;  
Query ID = cloudera_20190809020000_236d4b63-f5de-4f7a-978c-01611edd3d55  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0024, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0024/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0024  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 02:00:14,046 Stage-1 map = 0%, reduce = 0%  
2019-08-09 02:00:22,466 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.69 sec  
2019-08-09 02:00:29,925 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.45 sec  
MapReduce Total cumulative CPU time: 4 seconds 450 msec  
Ended Job = job_1565324019445_0024  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.45 sec HDFS Read: 46493002 HDFS Write: 6 SUCCESS  
Total MapReduce CPU Time Spent: 4 seconds 450 msec  
OK  
16951  
Time taken: 24.754 seconds, Fetched: 1 row(s)  
hive>   
AadharCard.pdf
```

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 ;
```



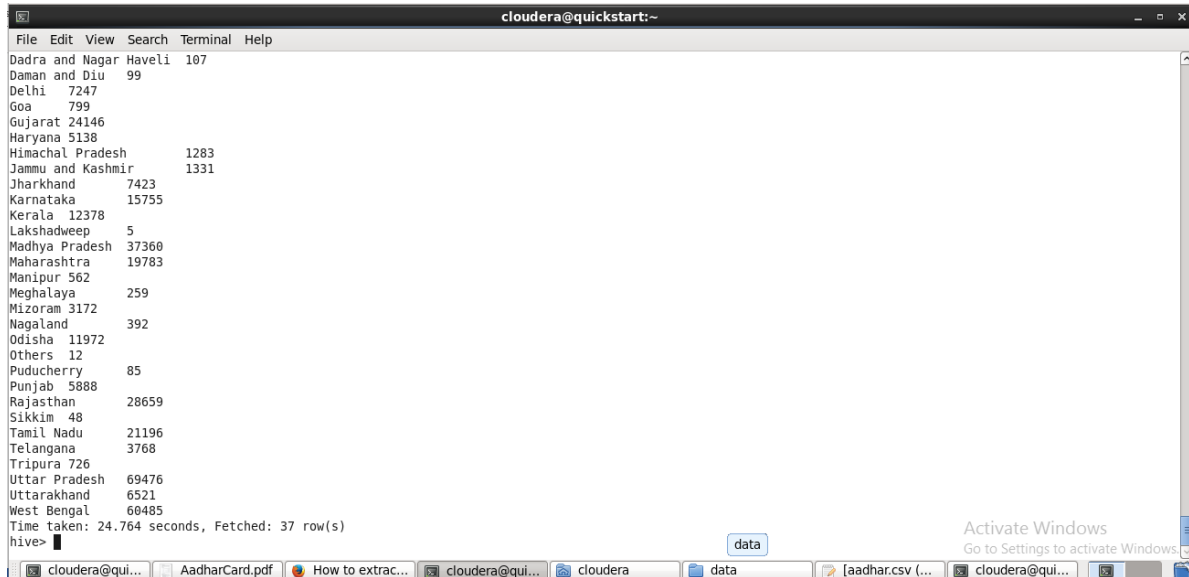
The screenshot shows a terminal window titled 'cloudera@quickstart:~'. The terminal displays the output of a Hive query and the logs of a MapReduce job. The query results show the top 3 districts by enrolment number: South 24 Parganas (266), Banka (289), and Budaun (395). The job logs show the execution of the query, including the number of mappers and reducers, and the cumulative CPU time. The terminal also shows the command to kill the job and the command to extract data from HDFS.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
South 24 Parganas      266  
Banka      289  
Budaun     395  
Time taken: 24.509 seconds, Fetched: 440818 row(s)  
hive> select district,aadhar_genrated+rejected as su from data order by su desc limit 3 ;  
Query ID = cloudera_20190809021111_f9c29586-3599-4d25-a923-0dd43e0343a4  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0026, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0026/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0026  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 02:11:16,429 Stage-1 map = 0%, reduce = 0%  
2019-08-09 02:11:25,870 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.44 sec  
2019-08-09 02:11:33,278 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.84 sec  
MapReduce Total cumulative CPU time: 7 seconds 840 msec  
Ended Job = job_1565324019445_0026  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.84 sec HDFS Read: 46492307 HDFS Write: 43 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 840 msec  
OK  
Budaun      395  
Banka      289  
South 24 Parganas      266  
Time taken: 26.61 seconds | How to extract only few lines of data from HDFS?  
hive> select district,aad | Edureka Community - Mozilla Firefox by su desc limit 3 ;  
Activate Windows  
Go to Settings to activate Windows.
```

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 ;
```

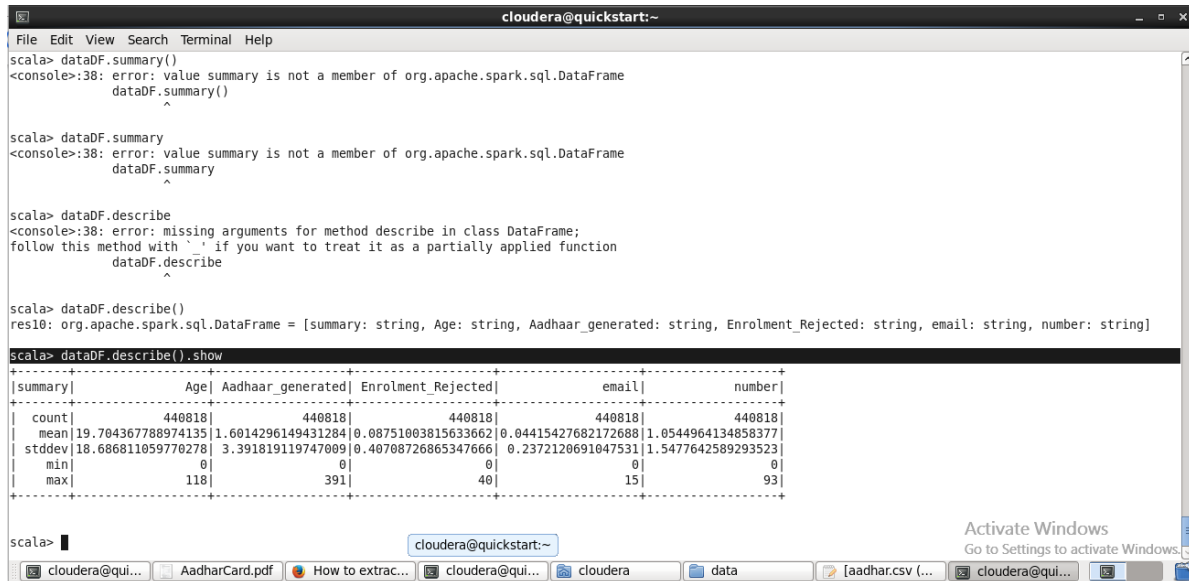


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Dadra and Nagar Haveli 107  
Daman and Diu 99  
Delhi 7247  
Goa 799  
Gujarat 24146  
Haryana 5138  
Himachal Pradesh 1283  
Jammu and Kashmir 1331  
Jharkhand 7423  
Karnataka 15755  
Kerala 12378  
Lakshadweep 5  
Madhya Pradesh 37360  
Maharashtra 19783  
Manipur 562  
Meghalaya 259  
Mizoram 3172  
Nagaland 392  
Odisha 11972  
Others 12  
Puducherry 85  
Punjab 5888  
Rajasthan 28659  
Sikkim 48  
Tamil Nadu 21196  
Telangana 3768  
Tripura 726  
Uttar Pradesh 69476  
Uttarakhand 6521  
West Bengal 60485  
Time taken: 24.764 seconds, Fetched: 37 row(s)  
hive>
```

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

Solution:

```
scala> dataDF.describe().show
```

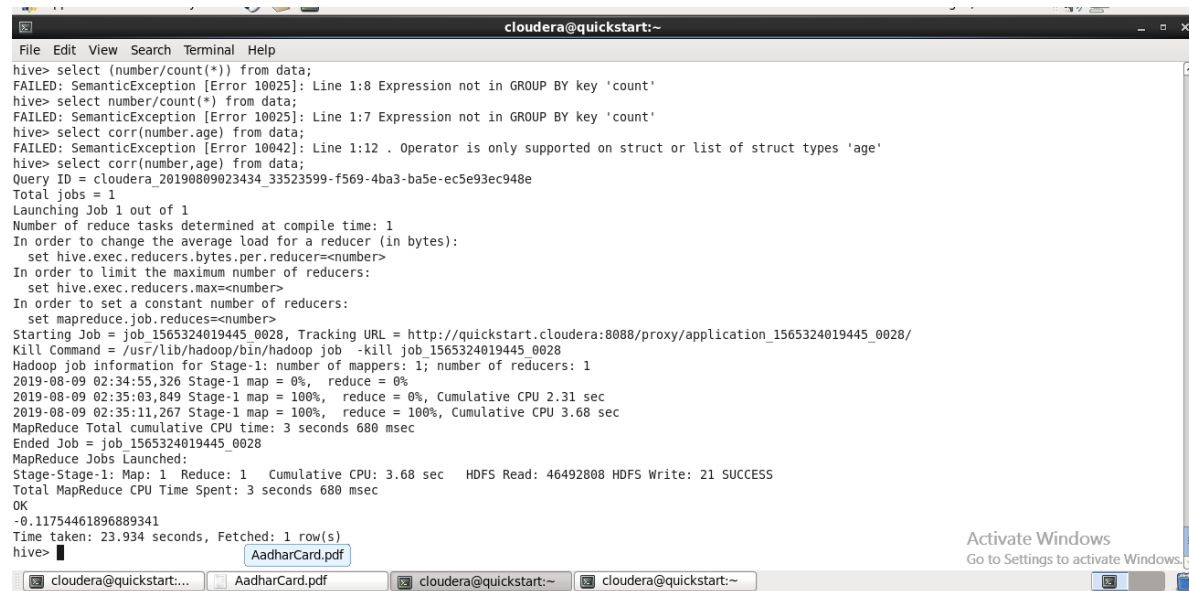


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
scala> dataDF.summary()  
<console>:38: error: value summary is not a member of org.apache.spark.sql.DataFrame  
  dataDF.summary()  
      ^  
  
scala> dataDF.summary  
<console>:38: error: value summary is not a member of org.apache.spark.sql.DataFrame  
  dataDF.summary  
      ^  
  
scala> dataDF.describe  
<console>:38: error: missing arguments for method describe in class DataFrame;  
follow this method with ` ` if you want to treat it as a partially applied function  
  dataDF.describe  
      ^  
  
scala> dataDF.describe()  
res10: org.apache.spark.sql.DataFrame = [summary: string, Age: string, Aadhaar_generated: string, Enrolment_Rejected: string, email: string, number: string]  
  
scala> dataDF.describe().show  
+-----+-----+-----+-----+-----+-----+  
|summary|      Age| Aadhaar_generated| Enrolment_Rejected|      email|      number|  
+-----+-----+-----+-----+-----+-----+  
| count|    440818|         440818|         440818|    440818|    440818|  
| mean|19.704367788974135|1.6014296149431284|0.08751003815633662|0.04415427682172688|1.0544964134858377|  
| stddev|18.686811059770278| 3.391819119747009|0.40708726865347666| 0.2372120691047531|1.5477642589293523|  
| min|      0|      0|      0|      0|      0|  
| max|    118|    391|    40|    15|    93|  
+-----+-----+-----+-----+-----+-----+  
  
scala>   
cloudera@quickstart:~  
cloudera@qui... AadharCard.pdf How to extrac... cloudera@qui... cloudera data [aadhar.csv (... cloudera@qui...
```


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;



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
hive> select (number/count(*)) from data;  
FAILED: SemanticException [Error 10025]: Line 1:8 Expression not in GROUP BY key 'count'  
hive> select number/count(*) from data;  
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'count'  
hive> select corr(number,age) from data;  
FAILED: SemanticException [Error 10042]: Line 1:12 . Operator is only supported on struct or list of struct types 'age'  
hive> select corr(number,age) from data;  
Query ID = cloudera_20190809023434_33523599-f569-4ba3-ba5e-ec5e93ec948e  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0028, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0028/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0028  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 02:34:55,326 Stage-1 map = 0%, reduce = 0%  
2019-08-09 02:35:03,849 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.31 sec  
2019-08-09 02:35:11,267 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.68 sec  
MapReduce Total cumulative CPU time: 3 seconds 680 msec  
Ended Job = job_1565324019445_0028  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.68 sec HDFS Read: 46492808 HDFS Write: 21 SUCCESS  
Total MapReduce CPU Time Spent: 3 seconds 680 msec  
OK  
-0.11754461896889341  
Time taken: 23.934 seconds, Fetched: 1 row(s)  
hive>
```

```

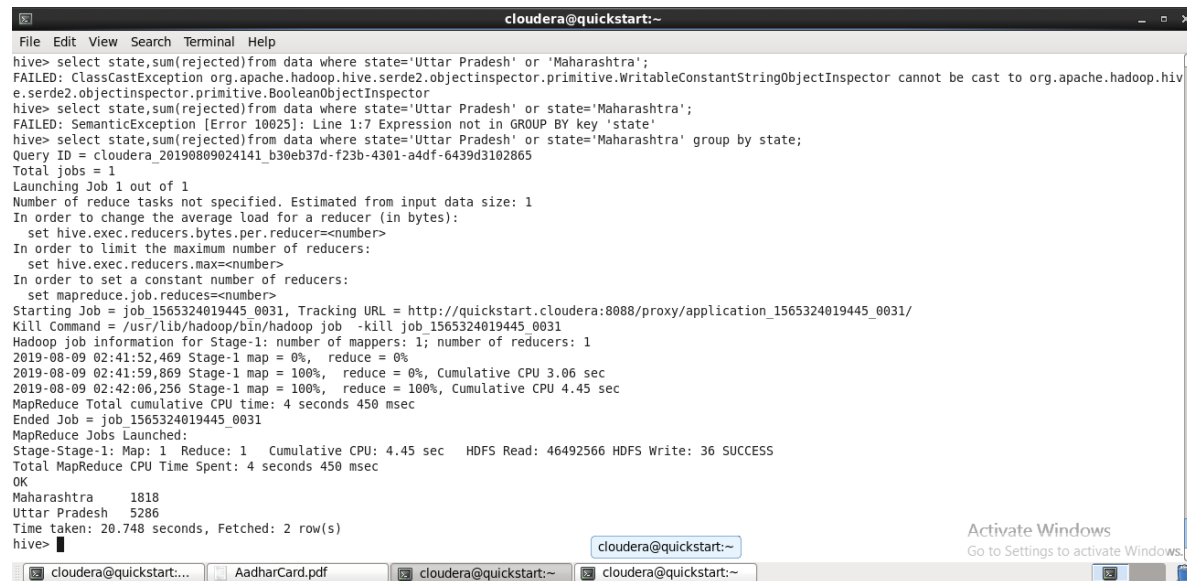
cloudera@quickstart:~
File Edit View Search Terminal Help
855115
855116
855117
855456
Others
Time taken: 23.729 seconds, Fetched: 17756 row(s)
hive> select count(distinct(pincode)) from data;
Query ID = cloudera_20190809023636_a958b81f-aef9-4904-9334-8b8d750a8d3e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
    set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
    set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
    set mapreduce.job.reducers=<number>
Starting Job = job_1565324019445_0030, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0030/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0030
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 02:37:06,679 Stage-1 map = 0%, reduce = 0%
2019-08-09 02:37:14,001 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.76 sec
2019-08-09 02:37:21,419 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.24 sec
MapReduce Total cumulative CPU time: 6 seconds 240 msec
Ended Job = job_1565324019445_0030
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.24 sec HDFS Read: 46492299 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 240 msec
OK
17756
Time taken: 22.782 seconds, Fetched: 1 row(s)
hive> █
cloudera@quickstart:~
cloudera@quickstart:... AadharCard.pdf cloudera@quickstart:~ cloudera@quickstart:~

```

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;
```

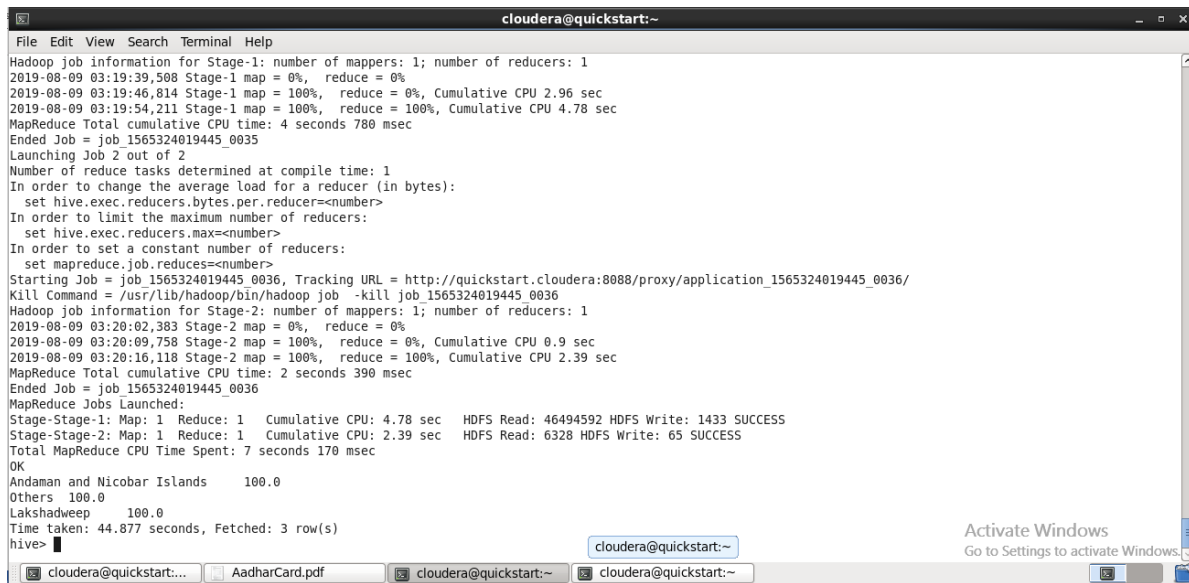


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
hive> select state,sum(rejected)from data where state='Uttar Pradesh' or 'Maharashtra';  
FAILED: ClassCastException org.apache.hadoop.hive.serde2.objectinspector.primitive.WritableConstantStringObjectInspector cannot be cast to org.apache.hadoop.hiv  
e.serde2.objectinspector.primitive.BooleanObjectInspector  
hive> select state,sum(rejected)from data where state='Uttar Pradesh' or state='Maharashtra';  
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'state'  
hive> select state,sum(rejected)from data where state='Uttar Pradesh' or state='Maharashtra' group by state;  
Query ID = cloudera_20190809024141_b30eb37d-f23b-4301-a4df-6439d3102865  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks not specified. Estimated from input data size: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0031, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0031/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0031  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 02:41:52,469 Stage-1 map = 0%, reduce = 0%  
2019-08-09 02:41:59,869 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.06 sec  
2019-08-09 02:42:06,256 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.45 sec  
MapReduce Total cumulative CPU time: 4 seconds 450 msec  
Ended Job = job_1565324019445_0031  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.45 sec HDFS Read: 46492566 HDFS Write: 36 SUCCESS  
Total MapReduce CPU Time Spent: 4 seconds 450 msec  
OK  
Maharashtra 1818  
Uttar Pradesh 5286  
Time taken: 20.748 seconds, Fetched: 2 row(s)  
hive>
```

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

Solution:

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

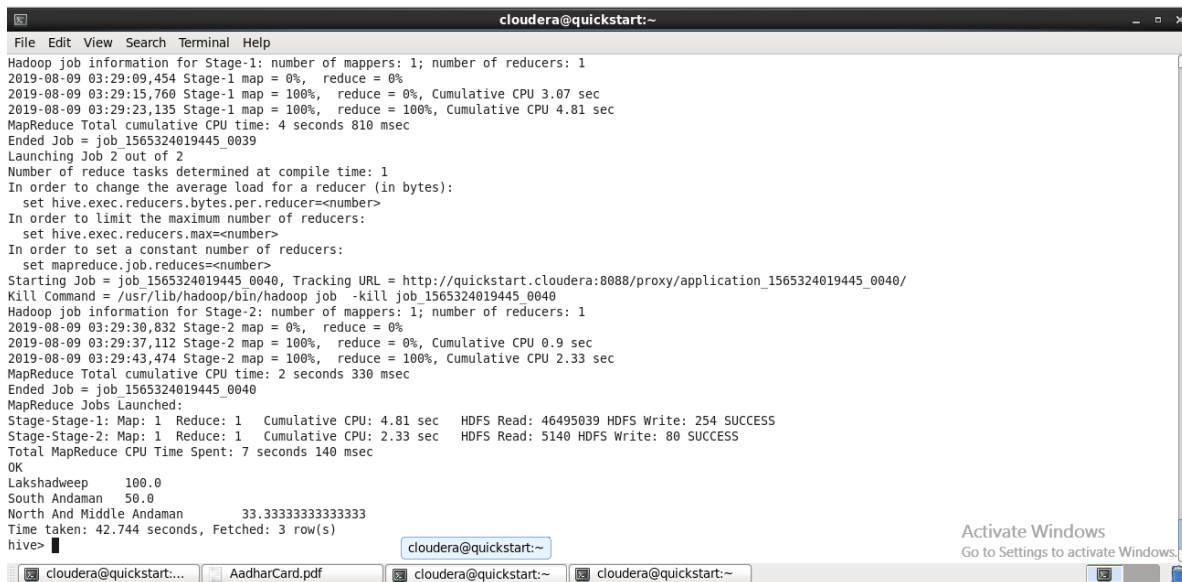


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 03:19:39,508 Stage-1 map = 0%, reduce = 0%  
2019-08-09 03:19:46,814 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.96 sec  
2019-08-09 03:19:54,211 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.78 sec  
MapReduce Total cumulative CPU time: 4 seconds 780 msec  
Ended Job = job_1565324019445_0035  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0036, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0036/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0036  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 03:20:02,383 Stage-2 map = 0%, reduce = 0%  
2019-08-09 03:20:09,758 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.9 sec  
2019-08-09 03:20:16,118 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.39 sec  
MapReduce Total cumulative CPU time: 2 seconds 390 msec  
Ended Job = job_1565324019445_0036  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.78 sec HDFS Read: 46494592 HDFS Write: 1433 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.39 sec HDFS Read: 6328 HDFS Write: 65 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 170 msec  
OK  
Andaman and Nicobar Islands    100.0  
Others    100.0  
Lakshadweep    100.0  
Time taken: 44.877 seconds, Fetched: 3 row(s)  
hive>
```

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;`

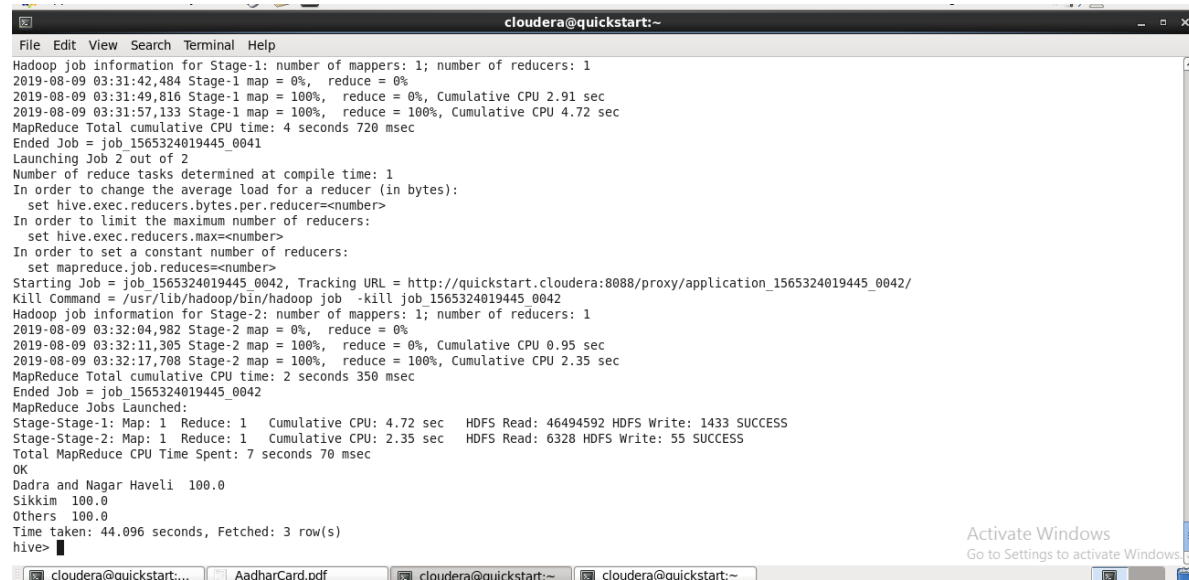


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 03:29:09,454 Stage-1 map = 0%, reduce = 0%  
2019-08-09 03:29:15,760 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.07 sec  
2019-08-09 03:29:23,135 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.81 sec  
MapReduce Total cumulative CPU time: 4 seconds 810 msec  
Ended Job = job_1565324019445_0039  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0040, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0040/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0040  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 03:29:30,832 Stage-2 map = 0%, reduce = 0%  
2019-08-09 03:29:37,112 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.9 sec  
2019-08-09 03:29:43,474 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.33 sec  
MapReduce Total cumulative CPU time: 2 seconds 330 msec  
Ended Job = job_1565324019445_0040  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.81 sec HDFS Read: 46495039 HDFS Write: 254 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.33 sec HDFS Read: 5140 HDFS Write: 80 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 140 msec  
OK  
Lakshadweep 100.0  
South Andaman 50.0  
North And Middle Andaman 33.33333333333333  
Time taken: 42.744 seconds, Fetched: 3 row(s)  
hive>
```

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

Solution:

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

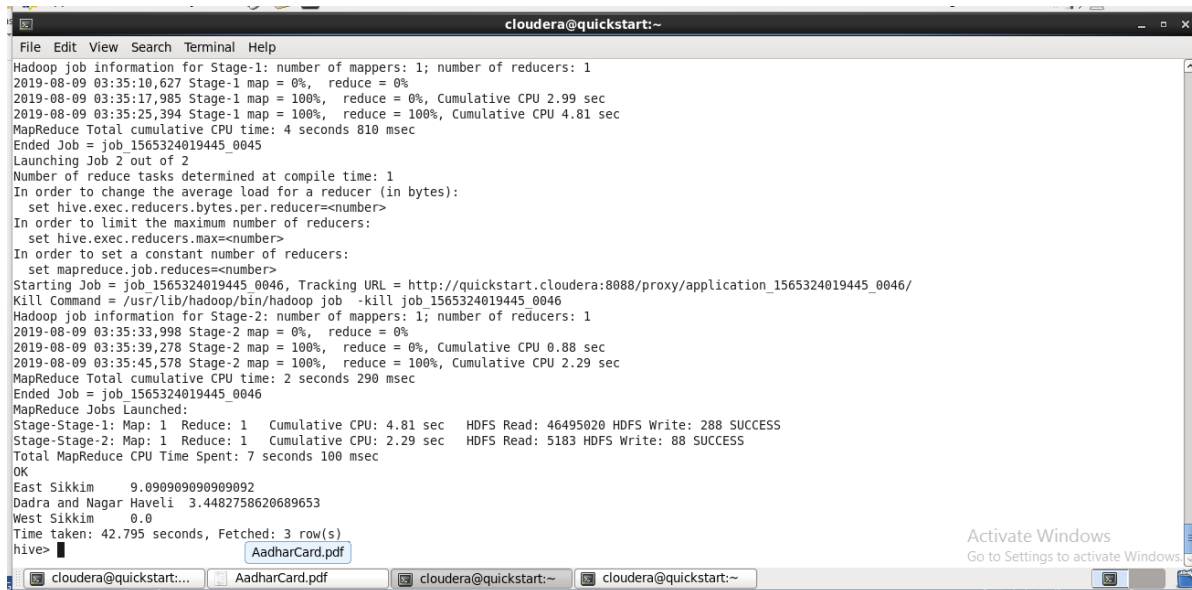


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 03:31:42,484 Stage-1 map = 0%, reduce = 0%  
2019-08-09 03:31:49,816 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.91 sec  
2019-08-09 03:31:57,133 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.72 sec  
MapReduce Total cumulative CPU time: 4 seconds 720 msec  
Ended Job = job_1565324019445_0041  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0042, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0042/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0042  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 03:32:04,982 Stage-2 map = 0%, reduce = 0%  
2019-08-09 03:32:11,305 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.95 sec  
2019-08-09 03:32:17,708 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.35 sec  
MapReduce Total cumulative CPU time: 2 seconds 350 msec  
Ended Job = job_1565324019445_0042  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.72 sec HDFS Read: 46494592 HDFS Write: 1433 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.35 sec HDFS Read: 6328 HDFS Write: 55 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 70 msec  
OK  
Dadra and Nagar Haveli 100.0  
Sikkim 100.0  
Others 100.0  
Time taken: 44.096 seconds, Fetched: 3 row(s)  
hive>
```

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;



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 03:35:10,627 Stage-1 map = 0%, reduce = 0%  
2019-08-09 03:35:17,985 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.99 sec  
2019-08-09 03:35:25,394 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.81 sec  
MapReduce Total cumulative CPU time: 4 seconds 810 msec  
Ended Job = job_1565324019445_0045  
Launching Job 2 out of 2  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0046, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0046/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0046  
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1  
2019-08-09 03:35:33,998 Stage-2 map = 0%, reduce = 0%  
2019-08-09 03:35:39,278 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.88 sec  
2019-08-09 03:35:45,578 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.29 sec  
MapReduce Total cumulative CPU time: 2 seconds 290 msec  
Ended Job = job_1565324019445_0046  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.81 sec HDFS Read: 46495020 HDFS Write: 288 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.29 sec HDFS Read: 5183 HDFS Write: 88 SUCCESS  
Total MapReduce CPU Time Spent: 7 seconds 100 msec  
OK  
East Sikkim 9.090909090909092  
Dadra and Nagar Haveli 3.4482758620689653  
West Sikkim 0.0  
Time taken: 42.795 seconds, Fetched: 3 row(s)  
hive>
```

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,

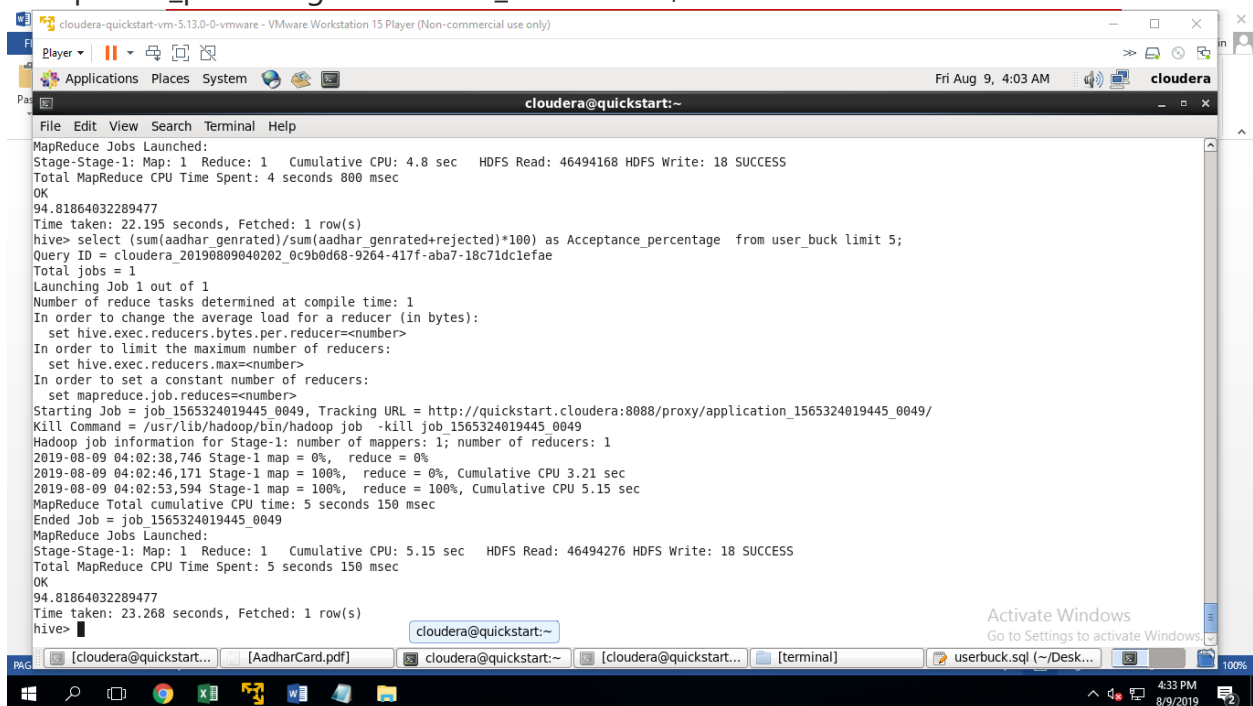
aadhar_genrated int,

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;



The screenshot shows a terminal window titled "cloudera@quickstart:~" within a VMware Workstation 15 Player. The terminal displays the output of several Hadoop and Hive commands. It shows the successful execution of a MapReduce job, followed by a Hive query that calculates the acceptance percentage for user buckets. The output of the Hive query is a single row with the value 94.81864032289477. The terminal also shows the execution of a Hive command to set the number of reducers to 1 and the execution of a Hive command to set the number of reducers to 1. The terminal window is part of a desktop environment with a taskbar at the bottom showing various applications and the system clock.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.8 sec HDFS Read: 46494168 HDFS Write: 18 SUCCESS  
Total MapReduce CPU Time Spent: 4 seconds 800 msec  
OK  
94.81864032289477  
Time taken: 22.195 seconds, Fetched: 1 row(s)  
hive> select (sum(aadhar_genrated)/sum(aadhar_genrated+rejected)*100) as Acceptance_percentage from user_buck limit 5;  
Query ID = cloudera_20190809040202_0c9b0d68-9264-417f-aba7-18c71dc1efae  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):  
  set hive.exec.reducers.bytes.per.reducer=<number>  
In order to limit the maximum number of reducers:  
  set hive.exec.reducers.max=<number>  
In order to set a constant number of reducers:  
  set mapreduce.job.reduces=<number>  
Starting Job = job_1565324019445_0049, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565324019445_0049/  
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1565324019445_0049  
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1  
2019-08-09 04:02:38,746 Stage-1 map = 0%, reduce = 0%  
2019-08-09 04:02:46,171 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.21 sec  
2019-08-09 04:02:53,594 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.15 sec  
MapReduce Total cumulative CPU time: 5 seconds 150 msec  
Ended Job = job_1565324019445_0049  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.15 sec HDFS Read: 46494276 HDFS Write: 18 SUCCESS  
Total MapReduce CPU Time Spent: 5 seconds 150 msec  
OK  
94.81864032289477  
Time taken: 23.268 seconds, Fetched: 1 row(s)  
hive>
```