AADHAR

CHECKPOINT 1

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

LOADING DATA IN HDFS

hdfs dfs -mkdir aadhar

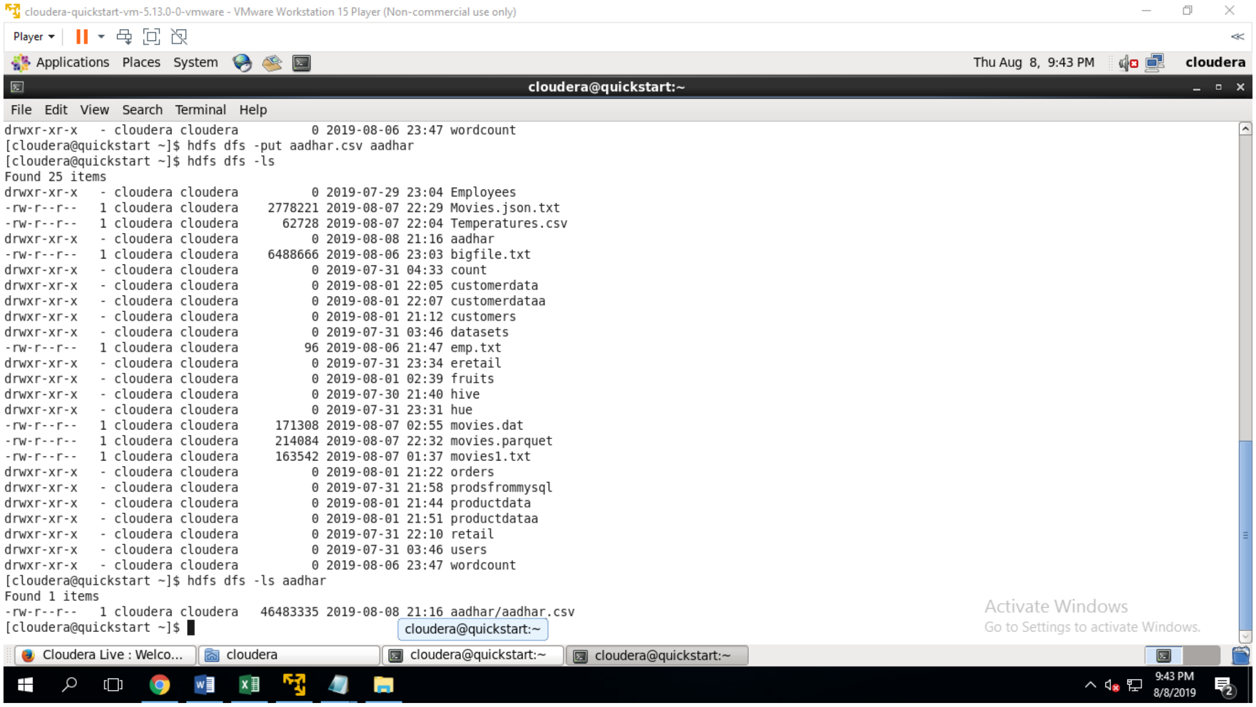
hdfs dfs -put aadhar.csv aadhar

hdfs dfs -ls

hdfs dfs -ls aadhar

Found 1 items

-rw-r--r--   1 cloudera cloudera   46483335 2019-08-08 21:16 aadhar/aadhar.csv



cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player
Applications Places System
Thu Aug 8, 9:43 PM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart:~
Activate Windows
Go to Settings to activate Windows.
Cloudera Live : Welco...
cloudera
cloudera@quickstart:~
cloudera@quickstart:~
9:43 PM
8/8/2019

```
drwxr-xr-x   - cloudera cloudera          0 2019-08-06 23:47 wordcount
[cloudera@quickstart ~]$ hdfs dfs -put aadhar.csv aadhar
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 25 items
drwxr-xr-x   - cloudera cloudera          0 2019-07-29 23:04 Employees
-rw-r--r--   1 cloudera cloudera    2778221 2019-08-07 22:29 Movies.json.txt
-rw-r--r--   1 cloudera cloudera      62728 2019-08-07 22:04 Temperatures.csv
drwxr-xr-x   - cloudera cloudera          0 2019-08-08 21:16 aadhar
-rw-r--r--   1 cloudera cloudera    6488666 2019-08-06 23:03 bigfile.txt
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 04:33 count
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 22:05 customerdata
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 22:07 customerdataa
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 21:12 customers
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 03:46 datasets
-rw-r--r--   1 cloudera cloudera         96 2019-08-06 21:47 emp.txt
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 23:34 eretail
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 02:39 fruits
drwxr-xr-x   - cloudera cloudera          0 2019-07-30 21:40 hive
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 23:31 hue
-rw-r--r--   1 cloudera cloudera     171308 2019-08-07 02:55 movies.dat
-rw-r--r--   1 cloudera cloudera     214084 2019-08-07 22:32 movies.parquet
-rw-r--r--   1 cloudera cloudera     163542 2019-08-07 01:37 movies1.txt
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 21:22 orders
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 21:58 prodsfrommysql
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 21:44 productdata
drwxr-xr-x   - cloudera cloudera          0 2019-08-01 21:51 productdataa
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 22:10 retail
drwxr-xr-x   - cloudera cloudera          0 2019-07-31 03:46 users
drwxr-xr-x   - cloudera cloudera          0 2019-08-06 23:47 wordcount
[cloudera@quickstart ~]$ hdfs dfs -ls aadhar
Found 1 items
-rw-r--r--   1 cloudera cloudera   46483335 2019-08-08 21:16 aadhar/aadhar.csv
[cloudera@quickstart ~]$ 
```

for loading data in Hive managed table

start Hive

hive> create database if not exists aadhar;

OK

Time taken: 1.435 seconds

hive> use aadhar;

OK

Time taken: 0.082 seconds

create table if not exists aadhar_in(registrar String,enrolment_Agency String,State String,district String,sub_District String,pin_code String,gender String,age int,aadhaarGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNumber int)

> row format delimited fields terminated by ','

> stored as textfile

> TBLPROPERTIES('serialization.null.format'='',

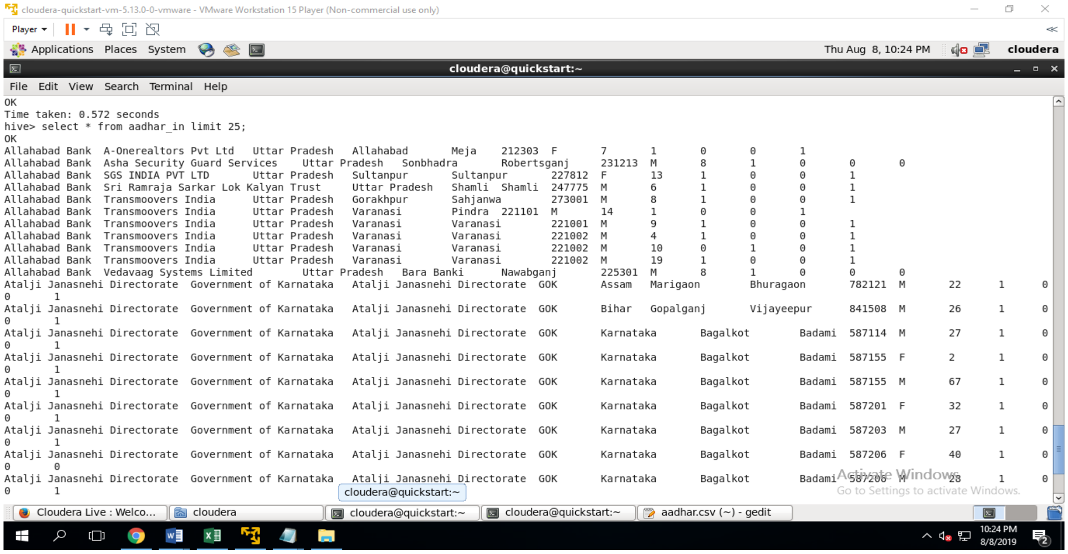
> 'skip.header.line.count'='1');

OK

Time taken: 0.071 seconds

load data inpath '/user/cloudera/aadhar/aadhar.csv' overwrite into table aadhar_in;

select * from aadhar_in limit 25;



```
OK
Time taken: 0.572 seconds
hive> select * from aadhar_in 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       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Bihar   Gopalganj       Vijayeepur      841508  M       26      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587114  M       27      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587155  F       2       1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587155  M       67      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587201  F       32      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587203  M       27      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587206  F       40      1       0
0       0
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587206  M       28      1       0
0       1
```

For creation of hive external table

hive> create external table if not exists aadhar_ex(registrar String,enrolment_Agency String,State String,district String,sub_District String,pin_code String,gender String,age int,aadhaarGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNumber int)

> row format delimited fields terminated by ','

> stored as textfile

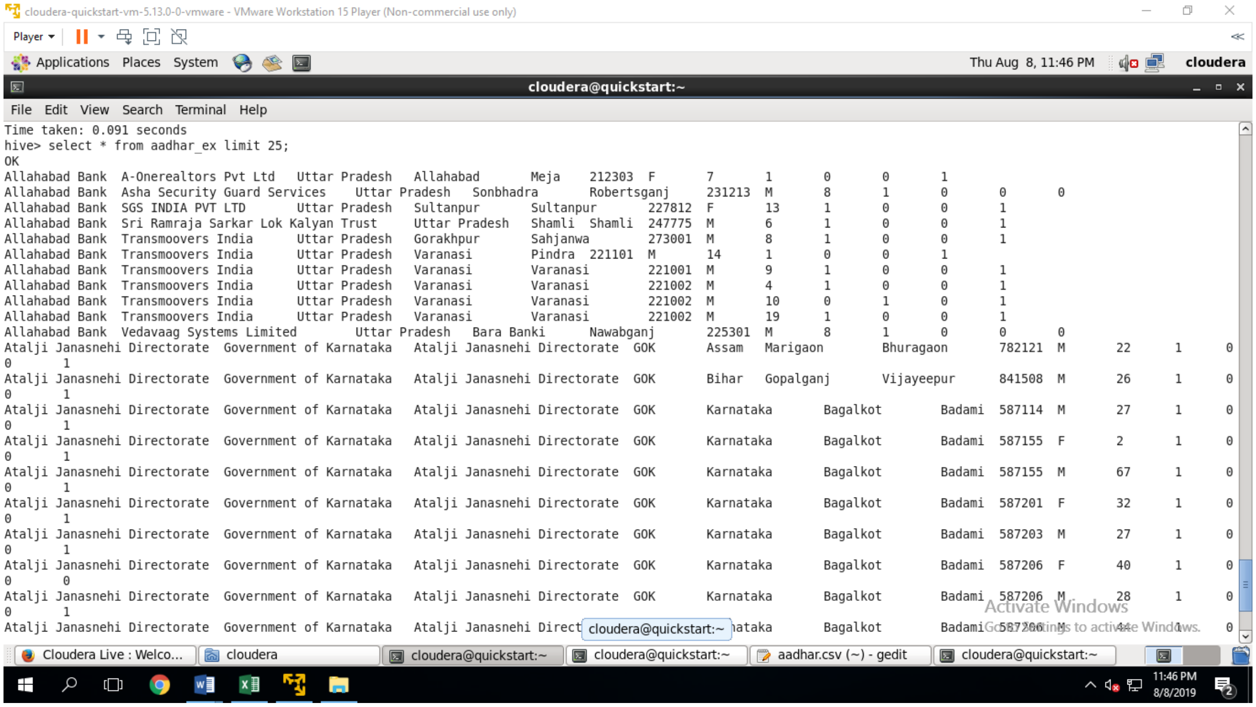
> location '/user/cloudera/aadhar/'

> TBLPROPERTIES('serialization.null.format'='',

> 'skip.header.line.count'='1');

OK

Time taken: 0.056 second



```
Time taken: 0.091 seconds
hive> select * from aadhar_ex 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       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Bihar   Gopalganj       Vijayeepur      841508  M       26      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587114  M       27      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587155  F       2       1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587155  M       67      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587201  F       32      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587203  M       27      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587206  F       40      1       0
0       0
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Directorate  GOK       Karnataka       Bagalkot        Badami  587206  M       28      1       0
0       1
Atalji Janasnehi Directorate  Government of Karnataka   Atalji Janasnehi Direct cloudera@quickstart:~ ataka   Bagalkot        Badami  587206  M       44      1       0
```

Creation of the table into DF

val aadharrdd=sc.textFile("/user/cloudera/aadhar/aadhar.csv")

val header=aadharrdd.first();

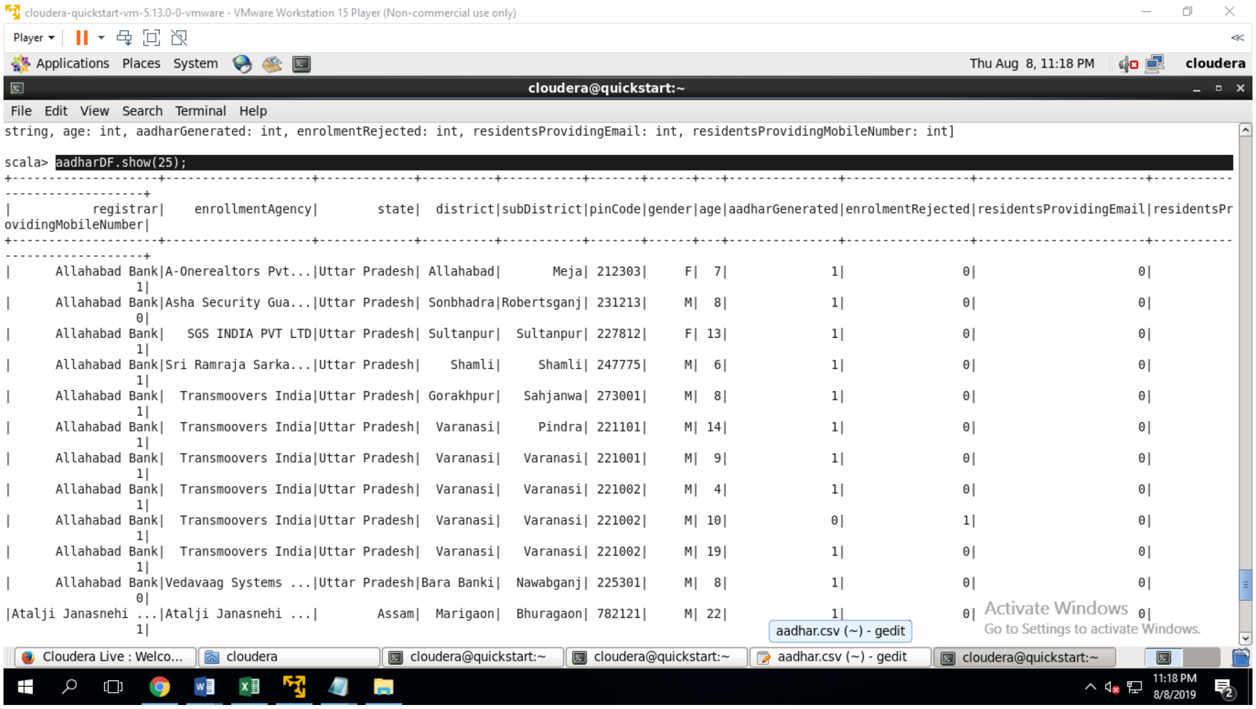
header: String = Registrar,Enrolment Agency,State,District,Sub District,Pin Code,Gender,Age,Aadhaar generated,Enrolment Rejected,Residents providing email,Residents providing mobile number

val firstaadharrdd=aadharrdd.filter(row=>row!=header);

val aadharDF=firstaadharrdd.map(_.split(",")).map{case Array(a,b,c,d,e,f,g,h,i,j,k,l) => (a,b,c,d,e,f,g,h.toInt,i.toInt,j.toInt,k.toInt,l.toInt)}.toDF("registrar","enrollmentAgency","state","district","subDistrict","pinCode","gender","age","aadharGenerated","enrolmentRejected","residentsProvidingEmail","residentsProvidingMobileNumber");

aadharDF: org.apache.spark.sql.DataFrame = [registrar: string, enrollmentAgency: string, state: string, district: string, subDistrict: string, pinCode: string, gender: string, age: int, aadharGenerated: int, enrolmentRejected: int, residentsProvidingEmail: int, residentsProvidingMobileNumber: int]

aadharDF.show(25);



```
string, age: int, aadharGenerated: int, enrolmentRejected: int, residentsProvidingEmail: int, residentsProvidingMobileNumber: int]

scala> aadharDF.show(25);
+--------------------+--------------------+-------------+----------+-----------+-------+------+---+---------------+-----------------+-----------------------+-----------
-------------------+
|           registrar|    enrollmentAgency|        state|  district|subDistrict|pinCode|gender|age|aadharGenerated|enrolmentRejected|residentsProvidingEmail|residentsPr
ovidingMobileNumber|
+--------------------+--------------------+-------------+----------+-----------+-------+------+---+---------------+-----------------+-----------------------+-----------
-------------------+
|      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|
```

CHECKPOINT 2

2. Describe schema

scala> aadharDF.printSchema

root

|-- registrar: string (nullable = true)

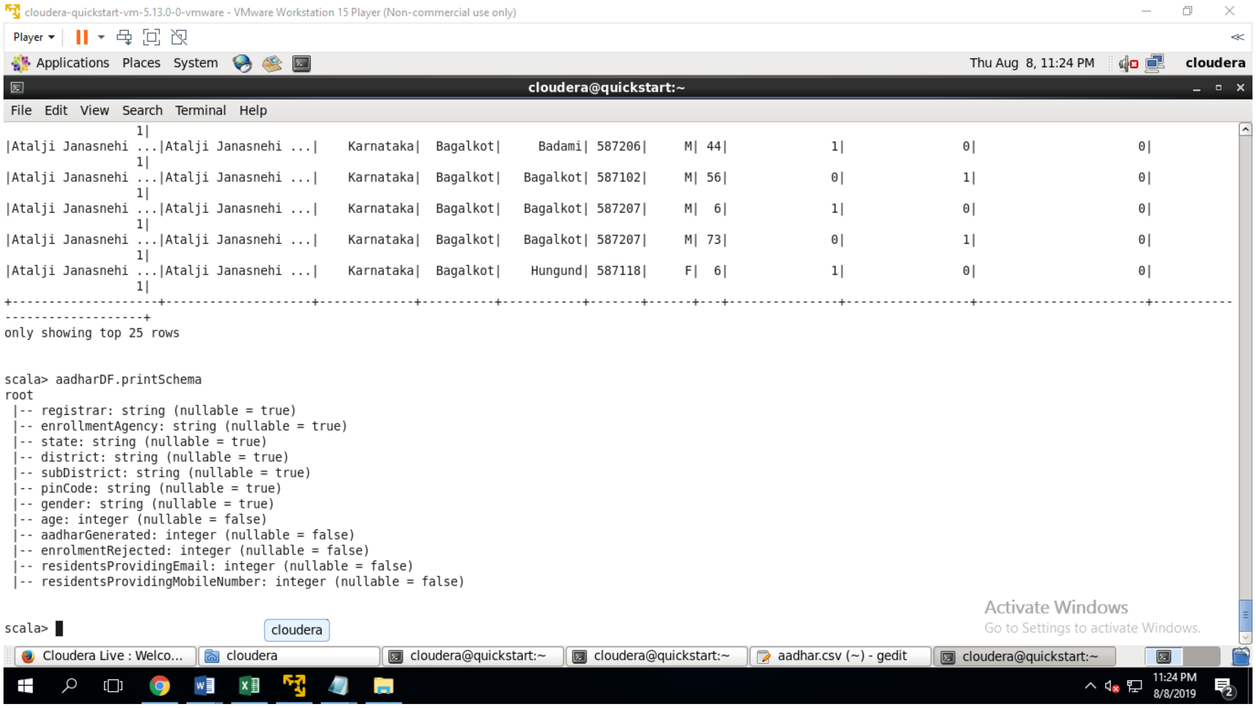
|-- enrollmentAgency: string (nullable = true)

|-- state: string (nullable = true)

|-- district: string (nullable = true)

|-- subDistrict: string (nullable = true)

```
 |-- pinCode: string (nullable = true)

 |-- gender: string (nullable = true)

 |-- age: integer (nullable = false)

 |-- aadharGenerated: integer (nullable = false)

 |-- enrolmentRejected: integer (nullable = false)

 |-- residentsProvidingEmail: integer (nullable = false)

 |-- residentsProvidingMobileNumber: integer (nullable = false)
```



cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player
Applications Places System
Thu Aug 8, 11:24 PM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

```
                   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> aadharDF.printSchema
root
 |-- registrar: string (nullable = true)
 |-- enrollmentAgency: string (nullable = true)
 |-- state: string (nullable = true)
 |-- district: string (nullable = true)
 |-- subDistrict: string (nullable = true)
 |-- pinCode: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- age: integer (nullable = false)
 |-- aadharGenerated: integer (nullable = false)
 |-- enrolmentRejected: integer (nullable = false)
 |-- residentsProvidingEmail: integer (nullable = false)
 |-- residentsProvidingMobileNumber: integer (nullable = false)


scala> █
```

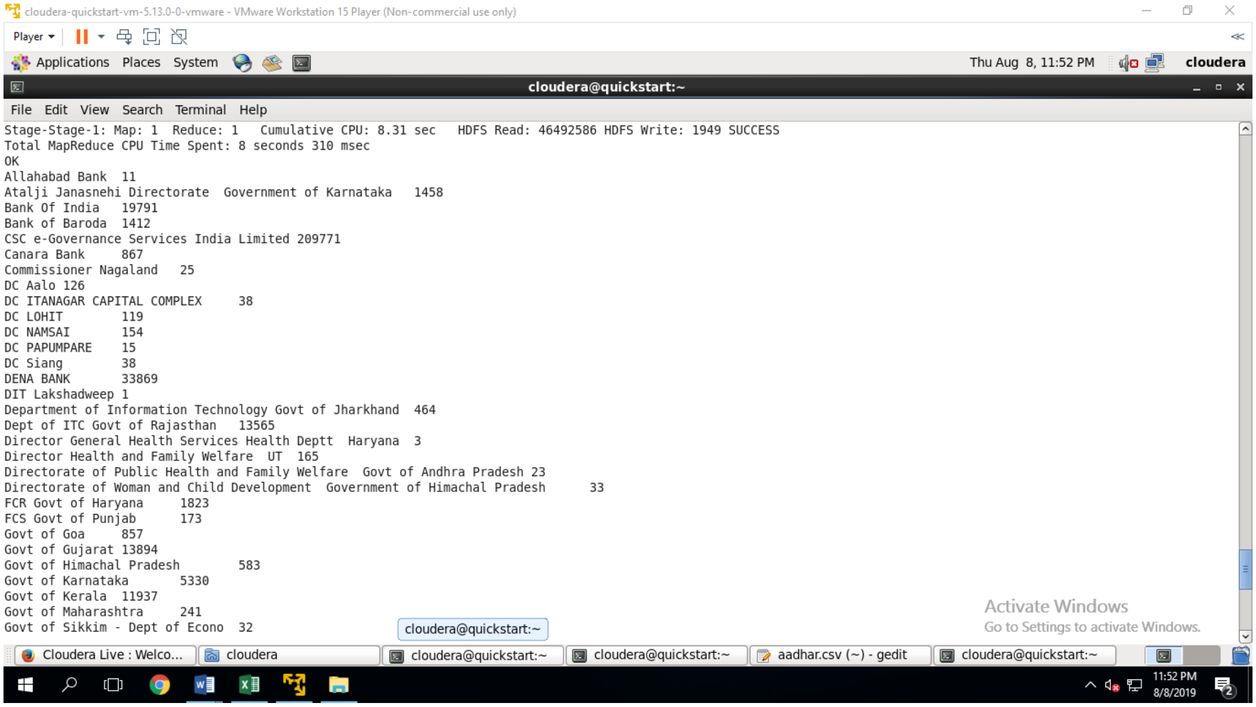
3. Find the count and names of registrars in the table

select registrar,count(*) from aadhar_ex groupby registrar;

```
OK

Allahabad Bank 11

Atalji Janasnehi Directorate  Government of Karnataka    1458

Bank Of India      19791

Bank of Baroda  1412

CSC e-Governance Services India Limited            209771

Canara Bank       867

Commissioner Nagaland             25
```

DC Aalo	126
DC ITANAGAR CAPITAL COMPLEX	38
DC LOHIT	119
DC NAMSAI	154
DC PAPUMPARE	15
DC Siang	38
DENA BANK	33869
DIT Lakshadweep	1
Department of Information Technology Govt of Jharkhand	464
Dept of ITC Govt of Rajasthan	13565
Director General Health Services Health Deptt Haryana	3
Director Health and Family Welfare UT	165
Directorate of Public Health and Family Welfare Govt of Andhra Pradesh	23
Directorate of Woman and Child Development Government of Himachal Pradesh	33
FCR Govt of Haryana	1823
FCS Govt of Punjab	173
Govt of Goa	857
Govt of Gujarat	13894
Govt of Himachal Pradesh	583
Govt of Karnataka	5330
Govt of Kerala	11937
Govt of Maharashtra	241
Govt of Sikkim - Dept of Econo	32
Govt of UT of Chandigarh	95
Govt. of Mizoram	3220
Govt. of Uttarkhand	44
IDBI Bank ltd	31
Information Technology & Communication Department	3958
Madhya Pradesh State Electronics Development Corporation Ltd.	17309
NSDL e-Governance Infrastructure Limited	54214
National Cooperative Consumers Federation Of India Limited	2590

```
Odisha Computer Application Center	1701

Punjab National Bank	1400

Punjab and Sind Bank	1543

RDD Govt of Tripura	606

Registrar General India BEL2	167

Registrar General India ECIL	757

Registrar General India Others	7

Registrar General of India ITI	55

Rural Development Department Bihar-1	640

Rural Development Dept  Govt. of Bihar	4145

Secretery IT J&K	110

State Bank of India	3422

Tamil Nadu eGovernance Agency	15468

U P Electronics Corporation Limited	293

U.P. Development Systems Corporation Ltd	4139

UIDAI-Registrar	19

UT Govt. Of Dadra & Nagar Haveli	46

UT Of Daman and Diu	50

UT of Puducherry	1

UTI Infrastructure Technology & Services Limited	2395

Union Bank	5536

Women and Child Development Govt. of Jharkhand	39

Time taken: 47.462 seconds, Fetched: 60 row(s)
```



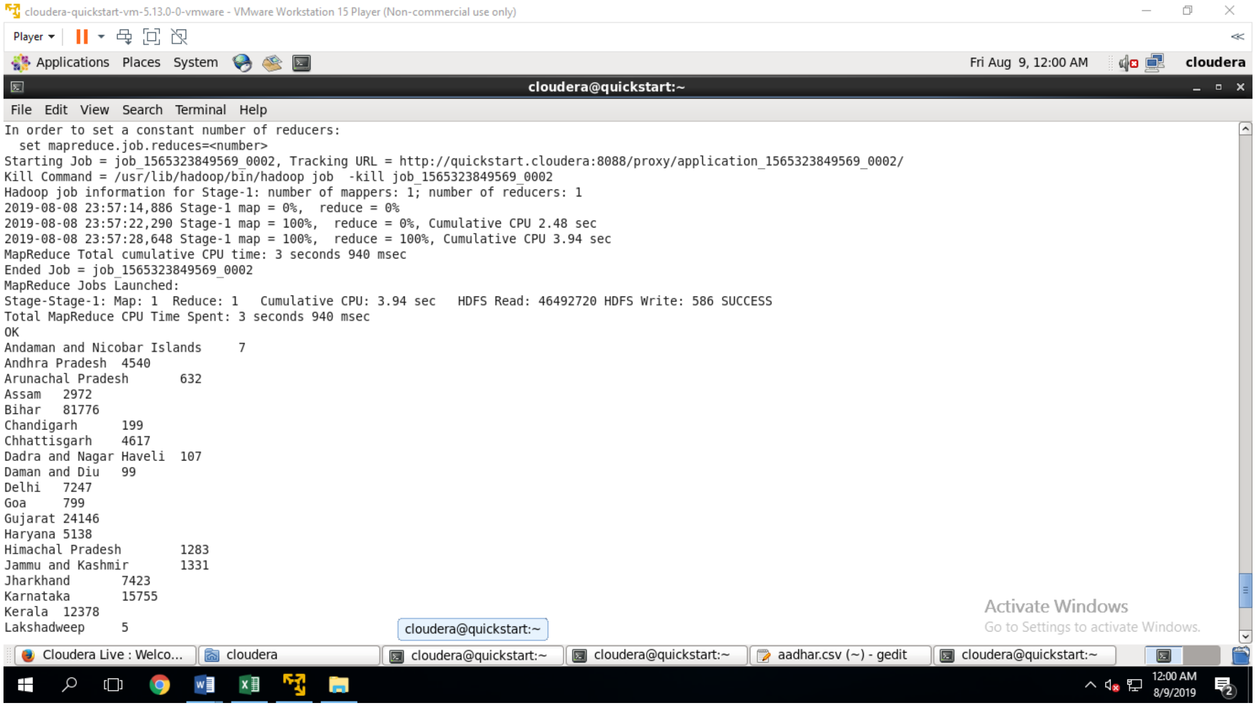
```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.31 sec   HDFS Read: 46492586 HDFS Write: 1949 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 310 msec
OK
Allahabad Bank  11
Atalji Janasnehi Directorate  Government of Karnataka   1458
Bank Of India   19791
Bank of Baroda  1412
CSC e-Governance Services India Limited 209771
Canara Bank     867
Commissioner Nagaland   25
DC Aalo 126
DC ITANAGAR CAPITAL COMPLEX     38
DC LOHIT        119
DC NAMSAI       154
DC PAPUMPARE    15
DC Siang        38
DENA BANK       33869
DIT Lakshadweep 1
Department of Information Technology Govt of Jharkhand  464
Dept of ITC Govt of Rajasthan   13565
Director General Health Services Health Deptt  Haryana  3
Director Health and Family Welfare  UT  165
Directorate of Public Health and Family Welfare  Govt of Andhra Pradesh 23
Directorate of Woman and Child Development  Government of Himachal Pradesh      33
FCR Govt of Haryana     1823
FCS Govt of Punjab      173
Govt of Goa     857
Govt of Gujarat 13894
Govt of Himachal Pradesh        583
Govt of Karnataka       5330
Govt of Kerala  11937
Govt of Maharashtra     241
Govt of Sikkim - Dept of Econo  32
```

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

```
select State,count(*) from aadhar_ex group by State;
```

```
OK
Andaman and Nicobar Islands	7
Andhra Pradesh4540
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
```

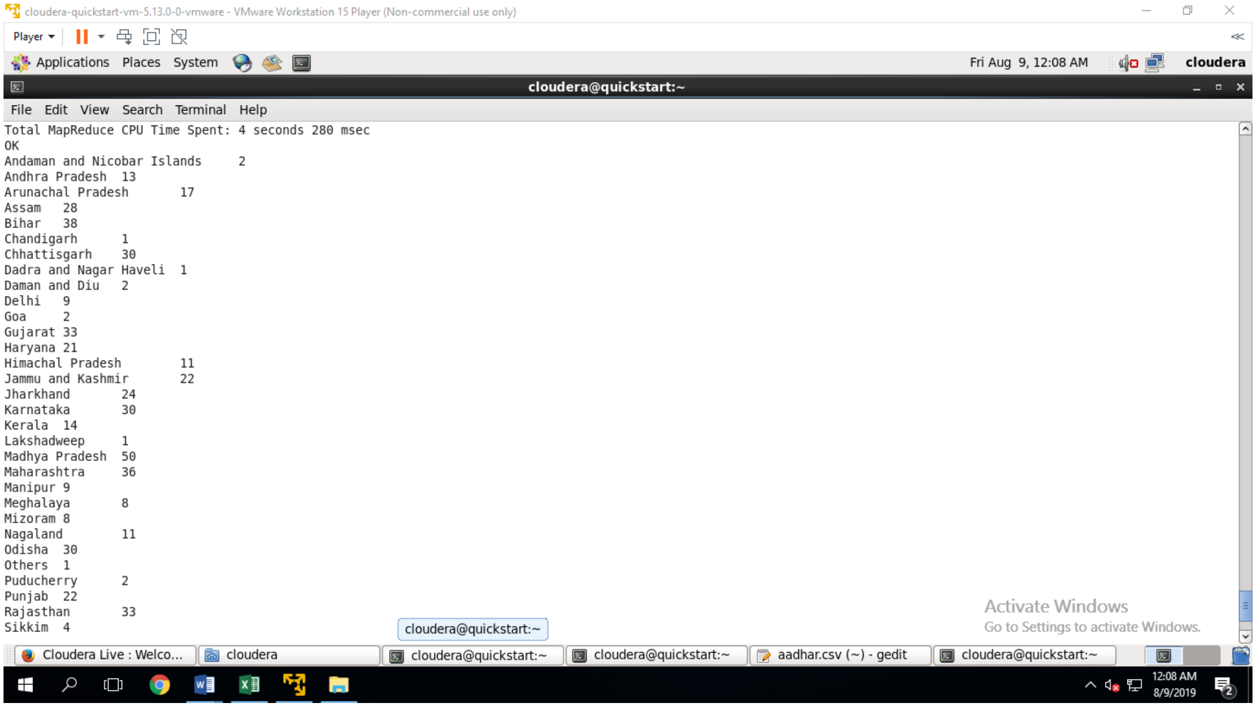
```
Himachal Pradesh        1283
Jammu and Kashmir       1331
Jharkhand       7423
Karnataka       15755
Kerala  12378
Lakshadweep     5
Madhya Pradesh          37360
Maharashtra     19783
Manipur         562
Meghalaya       259
Mizoram         3172
Nagaland        392
Odisha  11972
Others  12
Puducherry      85
Punjab  5888
Rajasthan       28659
Sikkim  48
Tamil Nadu      21196
Telangana       3768
Tripura 726
Uttar Pradesh   69476
Uttarakhand     6521
West Bengal     60485
Time taken: 23.244 seconds, Fetched: 37 row(s)
hive>
```



```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1565323849569_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-08 23:57:14,886 Stage-1 map = 0%,  reduce = 0%
2019-08-08 23:57:22,290 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.48 sec
2019-08-08 23:57:28,648 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.94 sec
MapReduce Total cumulative CPU time: 3 seconds 940 msec
Ended Job = job_1565323849569_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.94 sec   HDFS Read: 46492720 HDFS Write: 586 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 940 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
```

```
hive> select State,count(distinct(district)) from aadhar_ex group by State;

OK

Andaman and Nicobar Islands     2

Andhra Pradesh13

Arunachal Pradesh       17

Assam   28

Bihar   38

Chandigarh      1

Chhattisgarh    30

Dadra and Nagar Haveli 1

Daman and Diu 2

Delhi   9

Goa     2

Gujarat 33

Haryana         21

Himachal Pradesh        11

Jammu and Kashmir       22

Jharkhand       24
```

```
Karnataka	30
Kerala	14
Lakshadweep	1
Madhya Pradesh	50
Maharashtra	36
Manipur	9
Meghalaya	8
Mizoram	8
Nagaland	11
Odisha	30
Others	1
Puducherry	2
Punjab	22
Rajasthan	33
Sikkim	4
Tamil Nadu	32
Telangana	10
Tripura	8
Uttar Pradesh	75
Uttarakhand	13
West Bengal	19
Time taken: 22.608 seconds, Fetched: 37 row(s)
```



```
Total MapReduce CPU Time Spent: 4 seconds 280 msec
OK
Andaman and Nicobar Islands     2
Andhra Pradesh  13
Arunachal Pradesh       17
Assam   28
Bihar   38
Chandigarh      1
Chhattisgarh    30
Dadra and Nagar Haveli  1
Daman and Diu   2
Delhi   9
Goa     2
Gujarat 33
Haryana 21
Himachal Pradesh        11
Jammu and Kashmir       22
Jharkhand       24
Karnataka       30
Kerala  14
Lakshadweep     1
Madhya Pradesh  50
Maharashtra     36
Manipur 9
Meghalaya       8
Mizoram 8
Nagaland        11
Odisha  30
Others  1
Puducherry      2
Punjab  22
Rajasthan       33
Sikkim  4
```

select district,count(distinct(sub_district)) from aadhar_ex group by district;

Dhamtari	4
Dhanbad	9
Dhar	7
Dharmapuri	5
Dharwad	5
Dhemaji	5
Dhenkanal	10
Dholpur	5
Dhubri	8
Dhule	4
Dibrugarh	8
Dimapur	6
Dindigul	9
Dindori	2
Diu	1
Doda	4

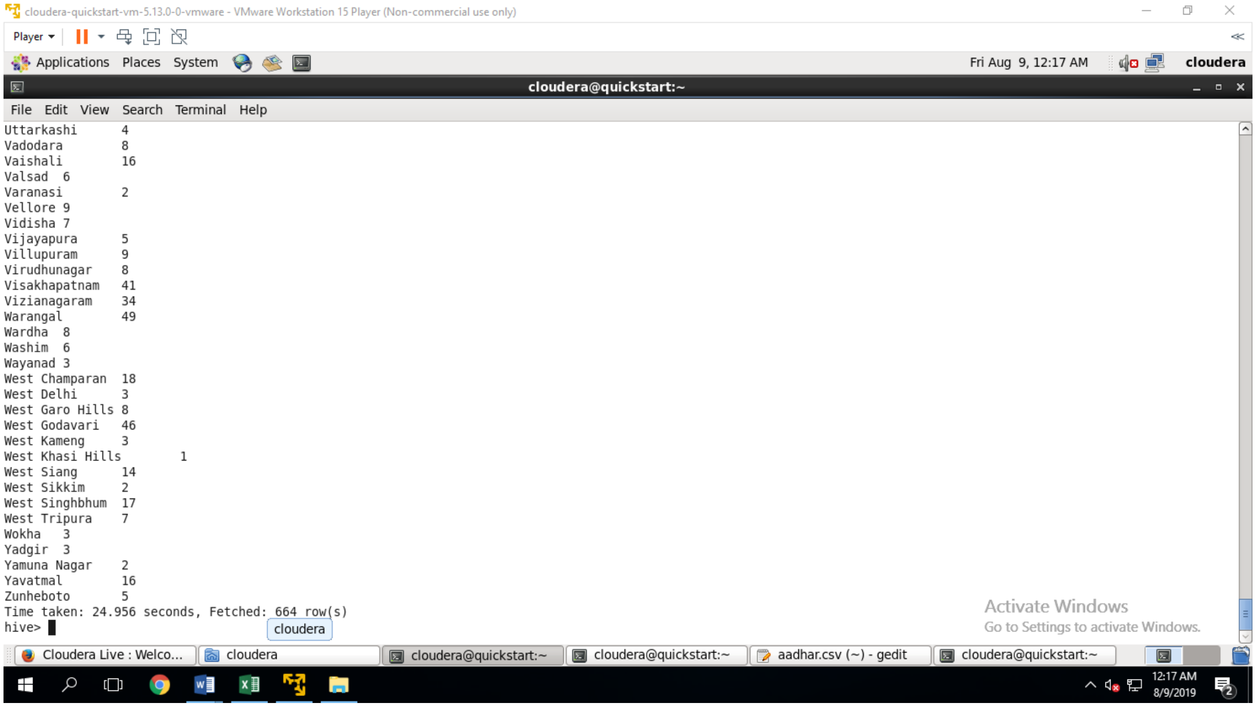
Dumka	10
Dungarpur	4
Durg	6
East Champaran	27
East Delhi	3
East Garo Hills	3
East Godavari	56
East Kameng	5
East Khasi Hills	5
East Siang	4
East Sikkim	4
East Singhbhum	12
Ernakulam	8
Erode	6
Etah	2
Etawah	5
Faizabad	5
Faridabad	2
Faridkot	2
Farrukhabad	2
Fatehabad	3
Fatehgarh Sahib	4
Fatehpur	3
Fazilka	3
Ferozepur	3
Firozabad	4
Gadag	5
Gadchiroli	10
Gajapati	8
Ganderbal	2
Gandhinagar	4

Ganganagar	9
Ganjam	22
Garhwa	14
Gariyaband	6
Gautam Buddha Nagar	4
Gaya	26
Ghaziabad	2
Ghazipur	4
Gir Somnath	6
Giridih	13
Goalpara	4
Godda	8
Golaghat	5
Gomati	6
Gonda	3
Gondiya	8
Gopalganj	14
Gorakhpur	6
Gumla	11
Guna	7
Guntur	53
Gurdaspur	3
Gurgaon	5
Gwalior	3
Hailakandi	4
Hamirpur	15
Hanumangarh	7
Hapur	3
Harda	3
Hardoi	5
Haridwar	3

Hassan	10
Hathras	4
Haveri	10
Hazaribagh	11
Hingoli	5
Hisar	4
Hojai	2
Hooghly	22
Hoshangabad	7
Hoshiarpur	4
Howrah	27
Hyderabad	17
Idukki	4
Imphal East	4
Imphal West	4
Indore	4
Jabalpur	4
Jagatsinghapur	7
Jaintia Hills	1
Jaipur	13
Jaisalmer	3
Jajapur	10
Jalandhar	5
Jalaun	4
Jalgaon	15
Jalna	8
Jalor	7
Jalpaiguri	13
Jammu	4
Jamnagar	6
Jamtara	4

Jamui	10
Janjgir-champa	10
Jashpur	8
Jaunpur	5
Jehanabad	8
Jhabua	5
Jhajjar	3
Jhalawar	7
Jhansi	4
Jharsuguda	10
Jhunjhunun	6
Jind	4
Jodhpur	7
Jorhat	5
Junagadh	11
K.v. Rangareddy	32
Kabeerdham	4
Kachchh	10
Kaimur (Bhabua)	11
Kaithal	2
Kalaburagi	9
Kalahandi	13
Kamrup	10
Kamrup Metro	4
Kancheepuram	18
Kandhamal	16
Kangra	23
Kanker	6
Kannauj	2
Kanniyakumari	4
Kannur	3

Kanpur Dehat	5
Kanpur Nagar	3



```
Uttarkashi      4
Vadodara        8
Vaishali        16
Valsad  6
Varanasi        2
Vellore 9
Vidisha 7
Vijayapura      5
Villupuram      9
Virudhunagar    8
Visakhapatnam   41
Vizianagaram    34
Warangal        49
Wardha  8
Washim  6
Wayanad 3
West Champaran  18
West Delhi      3
West Garo Hills 8
West Godavari   46
West Kameng     3
West Khasi Hills        1
West Siang      14
West Sikkim     2
West Singhbhum  17
West Tripura    7
Wokha   3
Yadgir  3
Yamuna Nagar    2
Yavatmal        16
Zunheboto       5
Time taken: 24.956 seconds, Fetched: 664 row(s)
hive> 
```

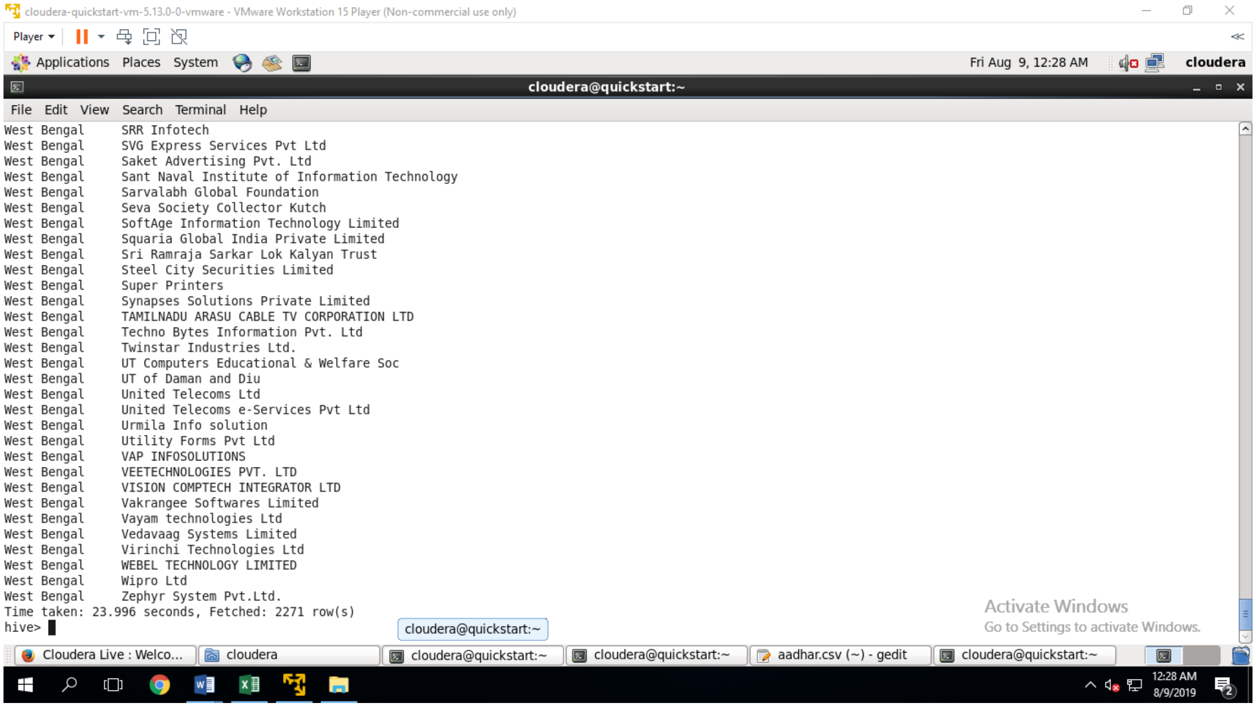
6. Find out the names of private agencies for each state.

select distinct(state),enrolment_Agency from aadhar_ex;

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: 23.996 seconds, Fetched: 2271 row(s)



```
West Bengal     SRR Infotech
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: 23.996 seconds, Fetched: 2271 row(s)
hive>
```

CHECKPOINT 3

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

hive> create table if not exists subaadhar as select state,sum(aadharGenerated) as aadharsum from aadhar_ex group by state;

select State,aadharsum from subaadhar order by aadharsum desc limit 3;

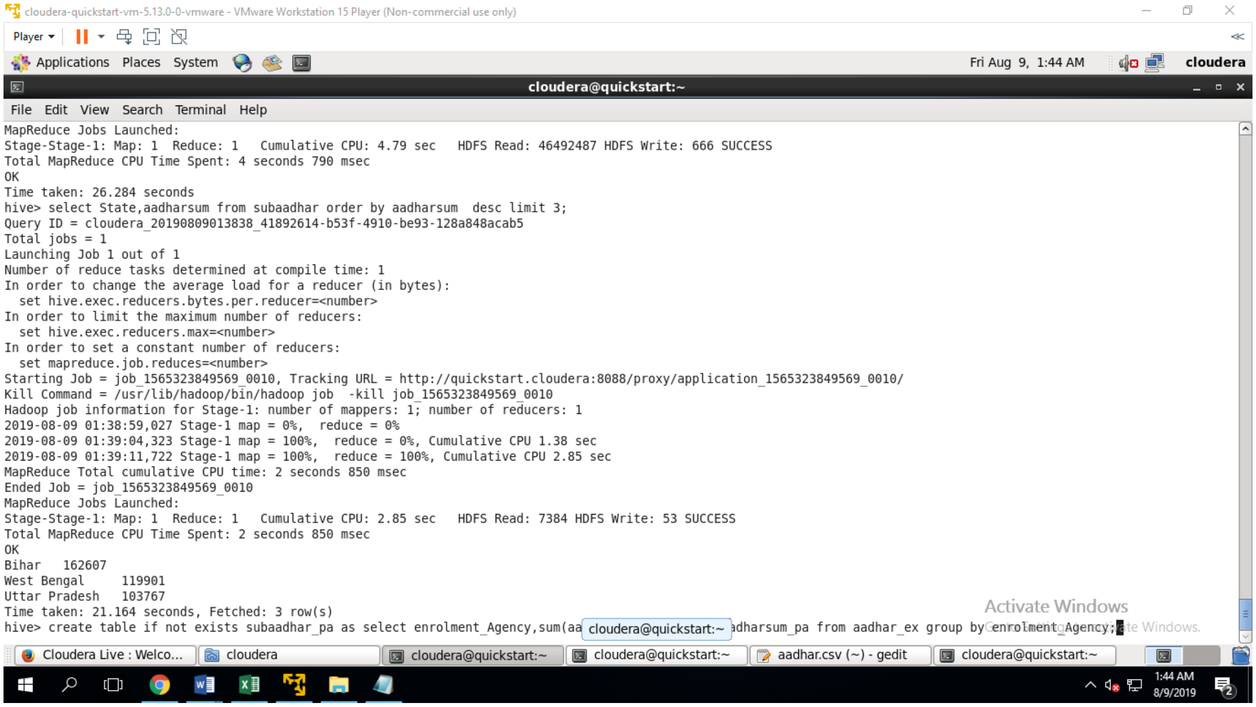
OK

Bihar	162607

West Bengal	119901

Uttar Pradesh	103767

Time taken: 21.164 seconds, Fetched: 3 row(s)



```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.79 sec   HDFS Read: 46492487 HDFS Write: 666 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 790 msec
OK
Time taken: 26.284 seconds
hive> select State,aadharsum from subaadhar order by aadharsum  desc limit 3;
Query ID = cloudera_20190809013838_41892614-b53f-4910-be93-128a848acab5
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_1565323849569_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 01:38:59,027 Stage-1 map = 0%,  reduce = 0%
2019-08-09 01:39:04,323 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.38 sec
2019-08-09 01:39:11,722 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.85 sec
MapReduce Total cumulative CPU time: 2 seconds 850 msec
Ended Job = job_1565323849569_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.85 sec   HDFS Read: 7384 HDFS Write: 53 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 850 msec
OK
Bihar   162607
West Bengal     119901
Uttar Pradesh   103767
Time taken: 21.164 seconds, Fetched: 3 row(s)
hive> create table if not exists subaadhar_pa as select enrolment_Agency,sum(aa                     adharsum_pa from aadhar_ex group by enrolment_Agency;
```

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

hive> create table if not exists subaadhar_pa as select enrolment_Agency,sum(aadhaargenerated) as aadharsum_pa from aadhar_ex group by enrolment_Agency;

select enrolment_Agency,aadharsum_pa from subaadhar_pa order by aadharsum_pa desc limit 3;

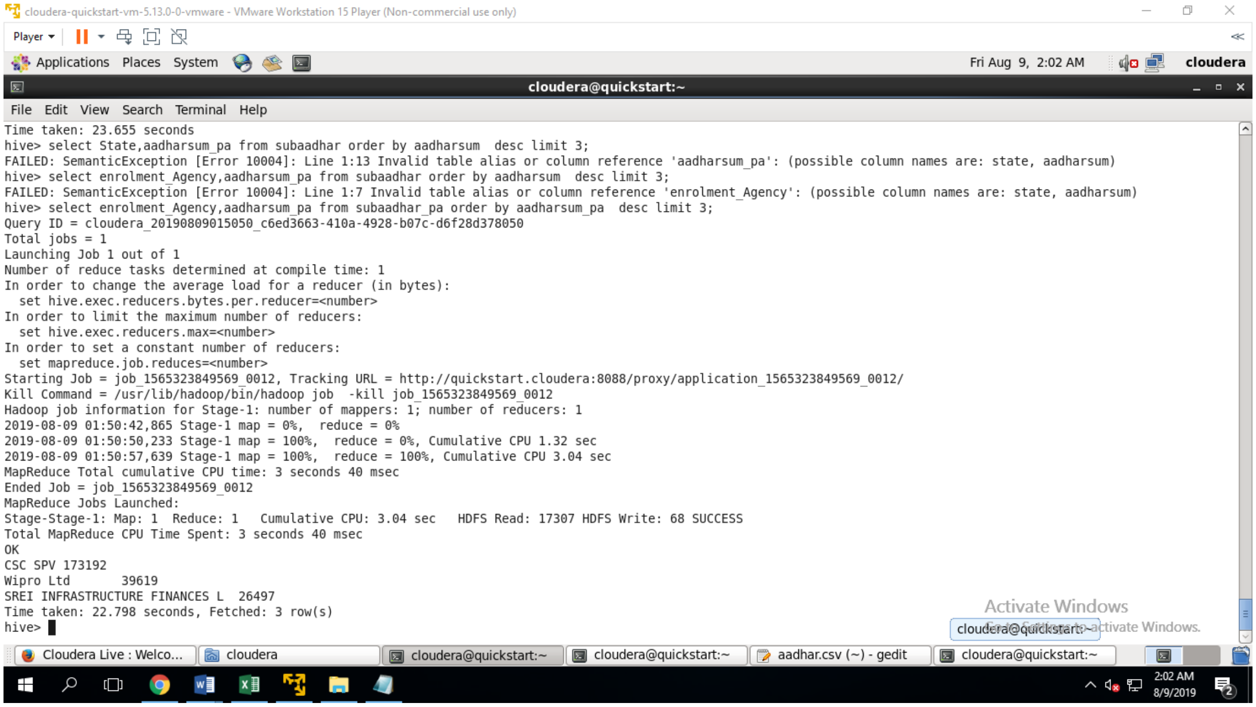
OK

CSC SPV 173192

Wipro Ltd 39619

SREI INFRASTRUCTURE FINANCES L 26497

Time taken: 22.798 seconds, Fetched: 3 row(s)



```
Time taken: 23.655 seconds
hive> select State,aadharsum_pa from subaadhar order by aadharsum  desc limit 3;
FAILED: SemanticException [Error 10004]: Line 1:13 Invalid table alias or column reference 'aadharsum_pa': (possible column names are: state, aadharsum)
hive> select enrolment_Agency,aadharsum_pa from subaadhar order by aadharsum  desc limit 3;
FAILED: SemanticException [Error 10004]: Line 1:7 Invalid table alias or column reference 'enrolment_Agency': (possible column names are: state, aadharsum)
hive> select enrolment_Agency,aadharsum_pa from subaadhar_pa order by aadharsum_pa  desc limit 3;
Query ID = cloudera_20190809015050_c6ed3663-410a-4928-b07c-d6f28d378050
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_1565323849569_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 01:50:42,865 Stage-1 map = 0%,  reduce = 0%
2019-08-09 01:50:50,233 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
2019-08-09 01:50:57,639 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.04 sec
MapReduce Total cumulative CPU time: 3 seconds 40 msec
Ended Job = job_1565323849569_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.04 sec   HDFS Read: 17307 HDFS Write: 68 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 40 msec
OK
CSC SPV 173192
Wipro Ltd       39619
SREI INFRASTRUCTURE FINANCES L  26497
Time taken: 22.798 seconds, Fetched: 3 row(s)
hive> 
```

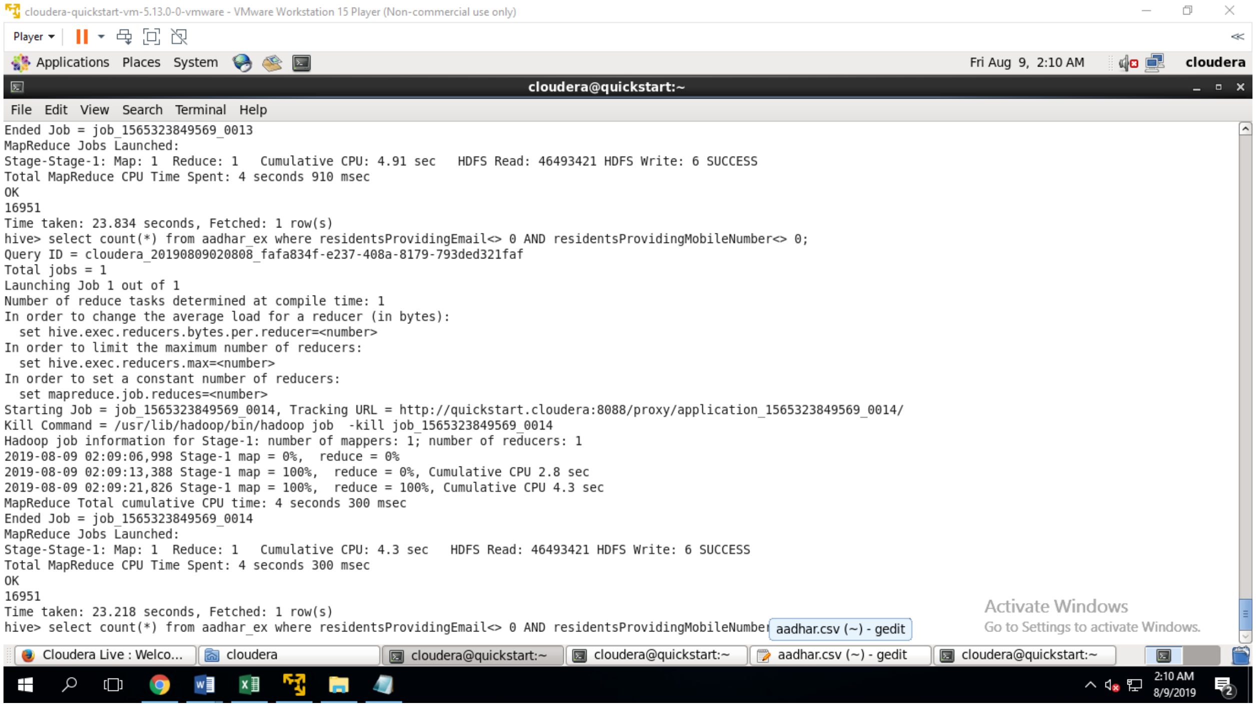
10. Find the number of residents providing email, mobile number? (Hint: consider non-zero values.)

hive> select count(*) from aadhar_ex where residentsProvidingEmail<> 0 AND residentsProvidingMobileNumber<> 0;

OK

16951

Time taken: 23.218 seconds, Fetched: 1 row(s)

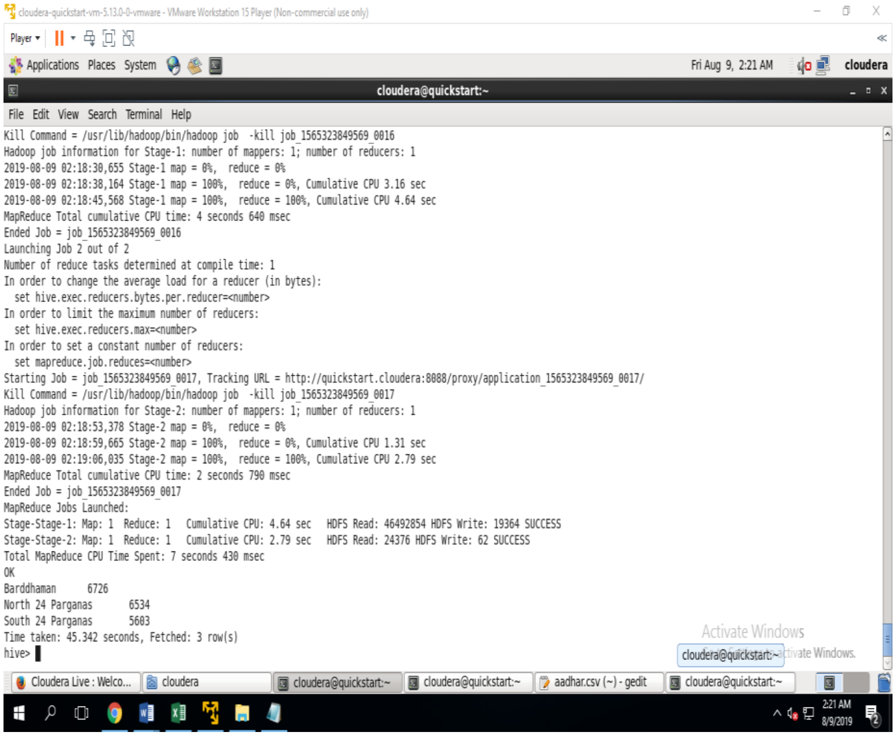


```
Ended Job = job_1565323849569_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.91 sec   HDFS Read: 46493421 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 910 msec
OK
16951
Time taken: 23.834 seconds, Fetched: 1 row(s)
hive> select count(*) from aadhar_ex where residentsProvidingEmail<> 0 AND residentsProvidingMobileNumber<> 0;
Query ID = cloudera_20190809020808_fafa834f-e237-408a-8179-793ded321faf
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_1565323849569_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 02:09:06,998 Stage-1 map = 0%,  reduce = 0%
2019-08-09 02:09:13,388 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.8 sec
2019-08-09 02:09:21,826 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.3 sec
MapReduce Total cumulative CPU time: 4 seconds 300 msec
Ended Job = job_1565323849569_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.3 sec   HDFS Read: 46493421 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 300 msec
OK
16951
Time taken: 23.218 seconds, Fetched: 1 row(s)
hive> select count(*) from aadhar_ex where residentsProvidingEmail<> 0 AND residentsProvidingMobileNumbe
```

11. Find top 3 districts where enrolment numbers are maximum?

SELECT district,count(*) as cnt from aadhar_ex where enrolmentRejected = 0 group by district order by cnt desc limit 3;

```
OK

Barddhaman      6726

North 24 Parganas       6534

South 24 Parganas       5603
```

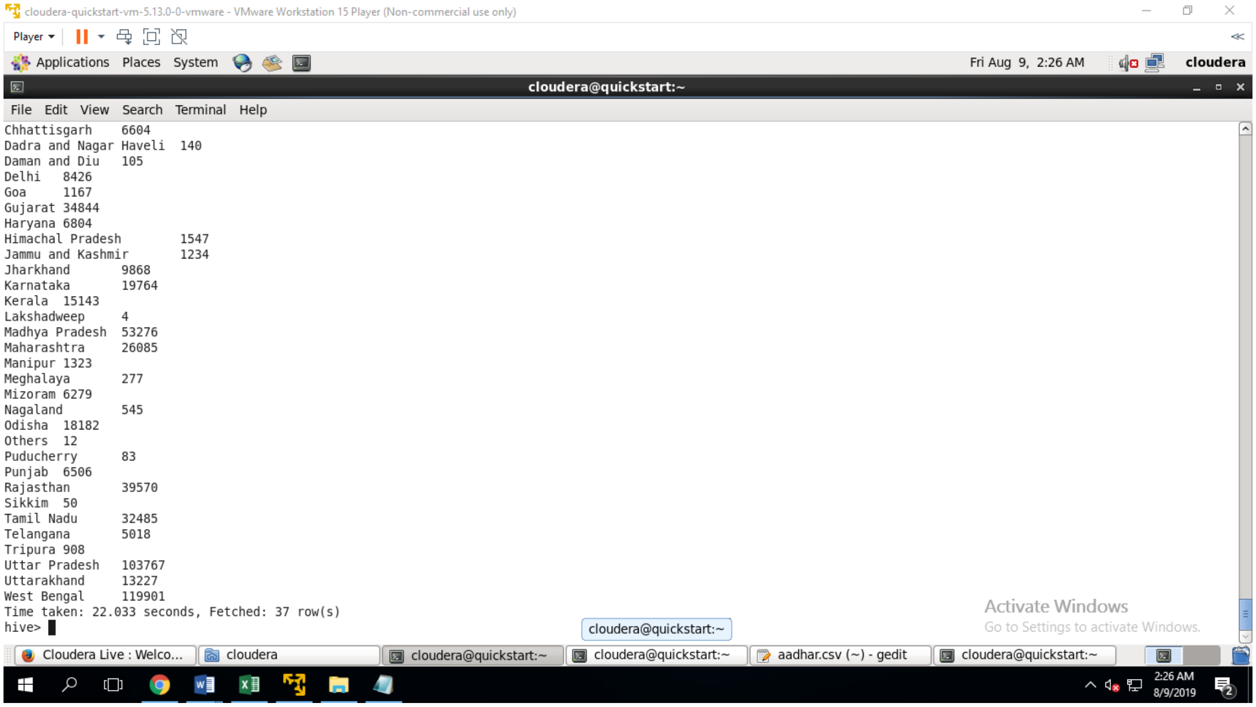


```
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 02:18:30,655 Stage-1 map = 0%,  reduce = 0%
2019-08-09 02:18:38,164 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.16 sec
2019-08-09 02:18:45,568 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.64 sec
MapReduce Total cumulative CPU time: 4 seconds 640 msec
Ended Job = job_1565323849569_0016
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_1565323849569_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0017
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-08-09 02:18:53,378 Stage-2 map = 0%,  reduce = 0%
2019-08-09 02:18:59,665 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.31 sec
2019-08-09 02:19:06,035 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.79 sec
MapReduce Total cumulative CPU time: 2 seconds 790 msec
Ended Job = job_1565323849569_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.64 sec   HDFS Read: 46492854 HDFS Write: 19364 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.79 sec   HDFS Read: 24376 HDFS Write: 62 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 430 msec
OK
Barddhaman      6726
North 24 Parganas       6534
South 24 Parganas       5603
Time taken: 45.342 seconds, Fetched: 3 row(s)
hive>
```

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

```
select state,sum(aadhaarGenerated)from aadhar_ex group by state;

OK

Andaman and Nicobar Islands     5

Andhra Pradesh5798

Arunachal Pradesh       913

Assam   3213

Bihar   162607

Chandigarh      259

Chhattisgarh    6604

Dadra and Nagar Haveli 140

Daman and Diu 105

Delhi   8426

Goa     1167

Gujarat 34844

Haryana         6804
```

```
Himachal Pradesh	1547
Jammu and Kashmir	1234
Jharkhand	9868
Karnataka	19764
Kerala	15143
Lakshadweep	4
Madhya Pradesh	53276
Maharashtra	26085
Manipur	1323
Meghalaya	277
Mizoram	6279
Nagaland	545
Odisha	18182
Others	12
Puducherry	83
Punjab	6506
Rajasthan	39570
Sikkim	50
Tamil Nadu	32485
Telangana	5018
Tripura	908
Uttar Pradesh	103767
Uttarakhand	13227
West Bengal	119901
Time taken: 22.033 seconds, Fetched: 37 row(s)
hive>
```



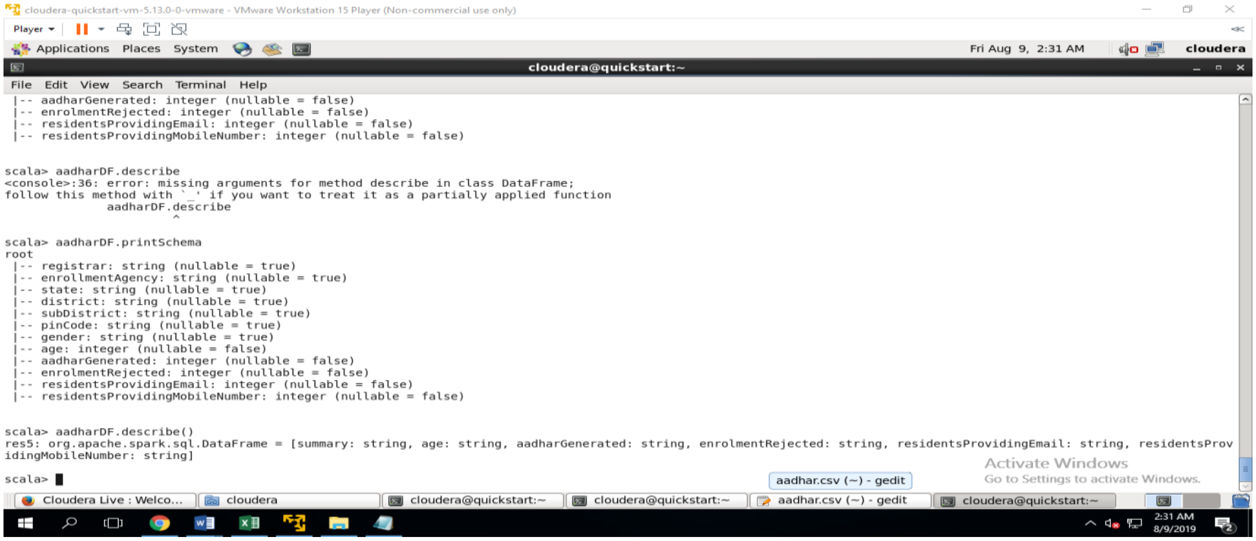
```
Chhattisgarh    6604
Dadra and Nagar Haveli  140
Daman and Diu   105
Delhi   8426
Goa     1167
Gujarat 34844
Haryana 6804
Himachal Pradesh        1547
Jammu and Kashmir       1234
Jharkhand       9868
Karnataka       19764
Kerala  15143
Lakshadweep     4
Madhya Pradesh  53276
Maharashtra     26085
Manipur 1323
Meghalaya       277
Mizoram 6279
Nagaland        545
Odisha  18182
Others  12
Puducherry      83
Punjab  6506
Rajasthan       39570
Sikkim  50
Tamil Nadu      32485
Telangana       5018
Tripura 908
Uttar Pradesh   103767
Uttarakhand     13227
West Bengal     119901
Time taken: 22.033 seconds, Fetched: 37 row(s)
hive>
```

CHECKPOINT 4

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

scala> aadharDF.describe()

res5: org.apache.spark.sql.DataFrame = [summary: string, age: string, aadharGenerated: string, enrolmentRejected: string, residentsProvidingEmail: string, residentsProvidingMobileNumber: string]



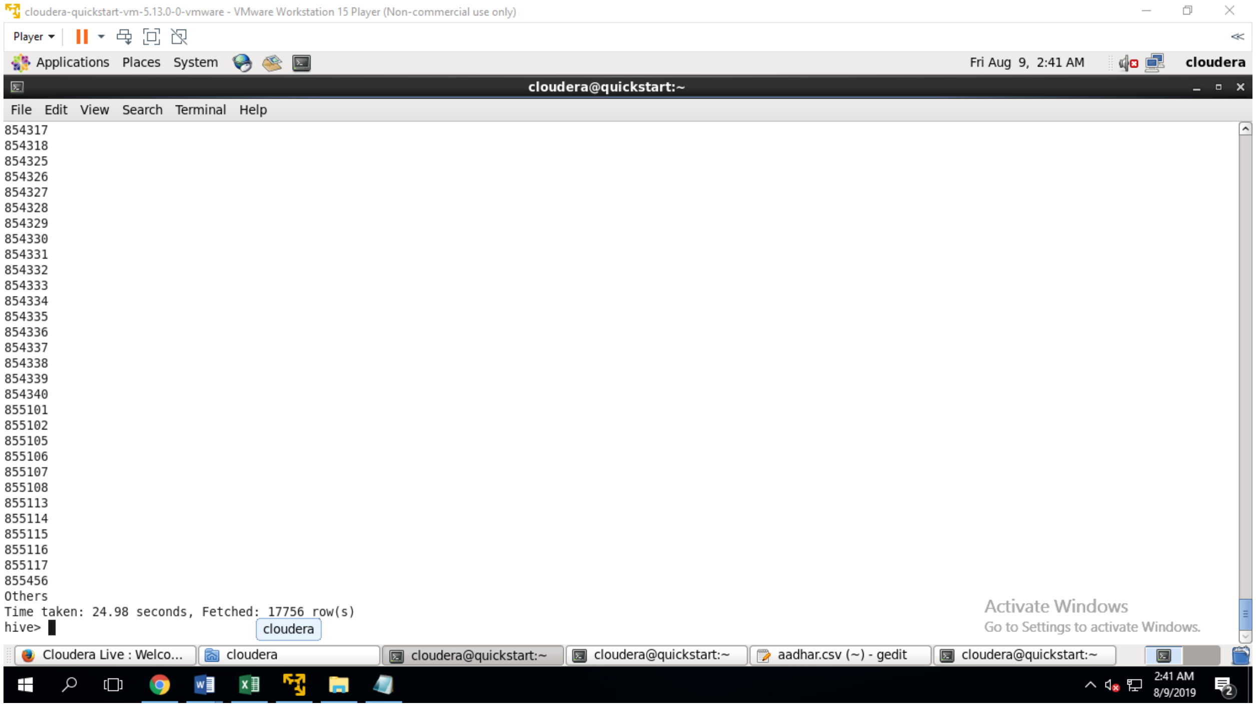
```
 |-- aadharGenerated: integer (nullable = false)
 |-- enrolmentRejected: integer (nullable = false)
 |-- residentsProvidingEmail: integer (nullable = false)
 |-- residentsProvidingMobileNumber: integer (nullable = false)


scala> aadharDF.describe
<console>:36: error: missing arguments for method describe in class DataFrame;
follow this method with `_' if you want to treat it as a partially applied function
              aadharDF.describe
                       ^

scala> aadharDF.printSchema
root
 |-- registrar: string (nullable = true)
 |-- enrollmentAgency: string (nullable = true)
 |-- state: string (nullable = true)
 |-- district: string (nullable = true)
 |-- subDistrict: string (nullable = true)
 |-- pinCode: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- age: integer (nullable = false)
 |-- aadharGenerated: integer (nullable = false)
 |-- enrolmentRejected: integer (nullable = false)
 |-- residentsProvidingEmail: integer (nullable = false)
 |-- residentsProvidingMobileNumber: integer (nullable = false)


scala> aadharDF.describe()
res5: org.apache.spark.sql.DataFrame = [summary: string, age: string, aadharGenerated: string, enrolmentRejected: string, residentsProvidingEmail: string, residentsProv
idingMobileNumber: string]

scala>
```

14. Write a command to see the correlation between “age” and “mobile_number”?

```
select corr(age,residentsProvidingMobileNumber)from aadhar_ex;

OK

-0.11754461896889339

Time taken: 24.314 seconds, Fetched: 1 row(s)
```

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

```
SELECT distinct(pin_code) from aadhar_ex;

854338

854339

854340

855101

855102

855105

855106

855107

855108

855113

855114

855115

855116

855117

855456

Others

Time taken: 24.98 seconds, Fetched: 17756 row(s)
```



```
854317
854318
854325
854326
854327
854328
854329
854330
854331
854332
854333
854334
854335
854336
854337
854338
854339
854340
855101
855102
855105
855106
855107
855108
855113
855114
855115
855116
855117
855456
Others
Time taken: 24.98 seconds, Fetched: 17756 row(s)
hive>
```

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

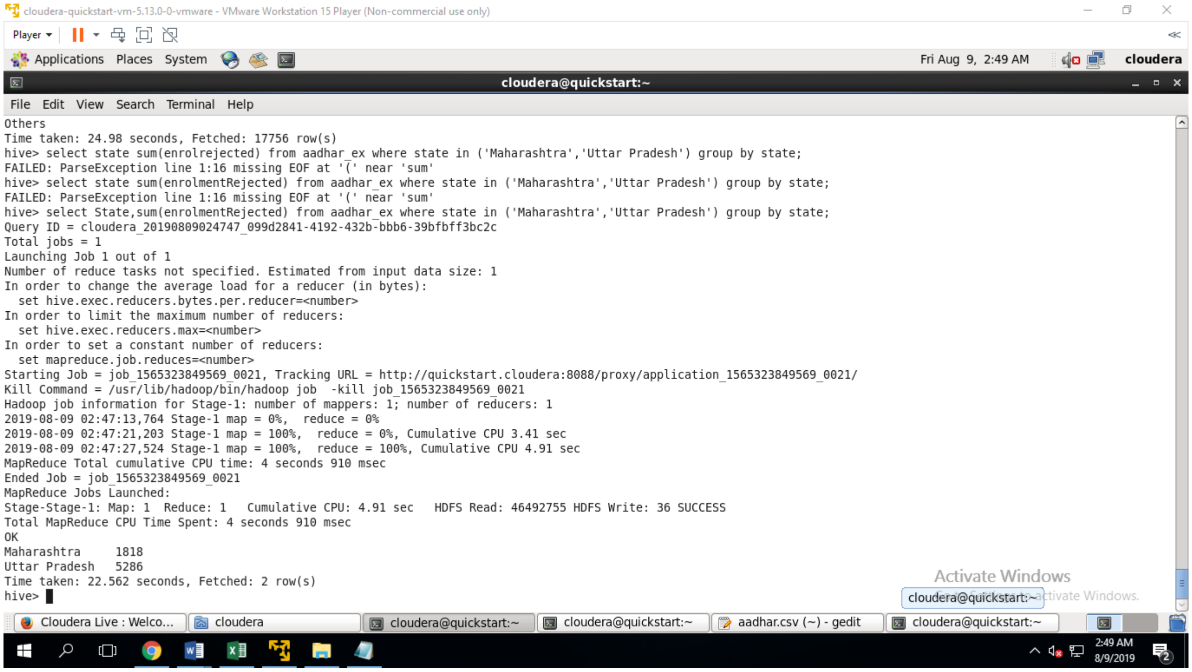
select State,sum(enrolmentRejected) from aadhar_ex where state in ('Maharashtra','Uttar Pradesh') group by state;

OK

Maharashtra 1818

Uttar Pradesh 5286

Time taken: 22.562 seconds, Fetched: 2 row(s)



```
Others
Time taken: 24.98 seconds, Fetched: 17756 row(s)
hive> select state sum(enrolrejected) from aadhar_ex where state in ('Maharashtra','Uttar Pradesh') group by state;
FAILED: ParseException line 1:16 missing EOF at '(' near 'sum'
hive> select state sum(enrolmentRejected) from aadhar_ex where state in ('Maharashtra','Uttar Pradesh') group by state;
FAILED: ParseException line 1:16 missing EOF at '(' near 'sum'
hive> select State,sum(enrolmentRejected) from aadhar_ex where state in ('Maharashtra','Uttar Pradesh') group by state;
Query ID = cloudera_20190809024747_099d2841-4192-432b-bbb6-39bfbff3bc2c
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_1565323849569_0021, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 02:47:13,764 Stage-1 map = 0%,  reduce = 0%
2019-08-09 02:47:21,203 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.41 sec
2019-08-09 02:47:27,524 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.91 sec
MapReduce Total cumulative CPU time: 4 seconds 910 msec
Ended Job = job_1565323849569_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.91 sec   HDFS Read: 46492755 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 910 msec
OK
Maharashtra     1818
Uttar Pradesh   5286
Time taken: 22.562 seconds, Fetched: 2 row(s)
hive> 
```

CHECKPOINT 5

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

hive> select state,(sum(aadhaarGenerated)*100)/(sum(aadhaarGenerated+enrolmentRejected))as male_members from aadhar_ex where gender='M' group by state order by male_members desc limit 3;

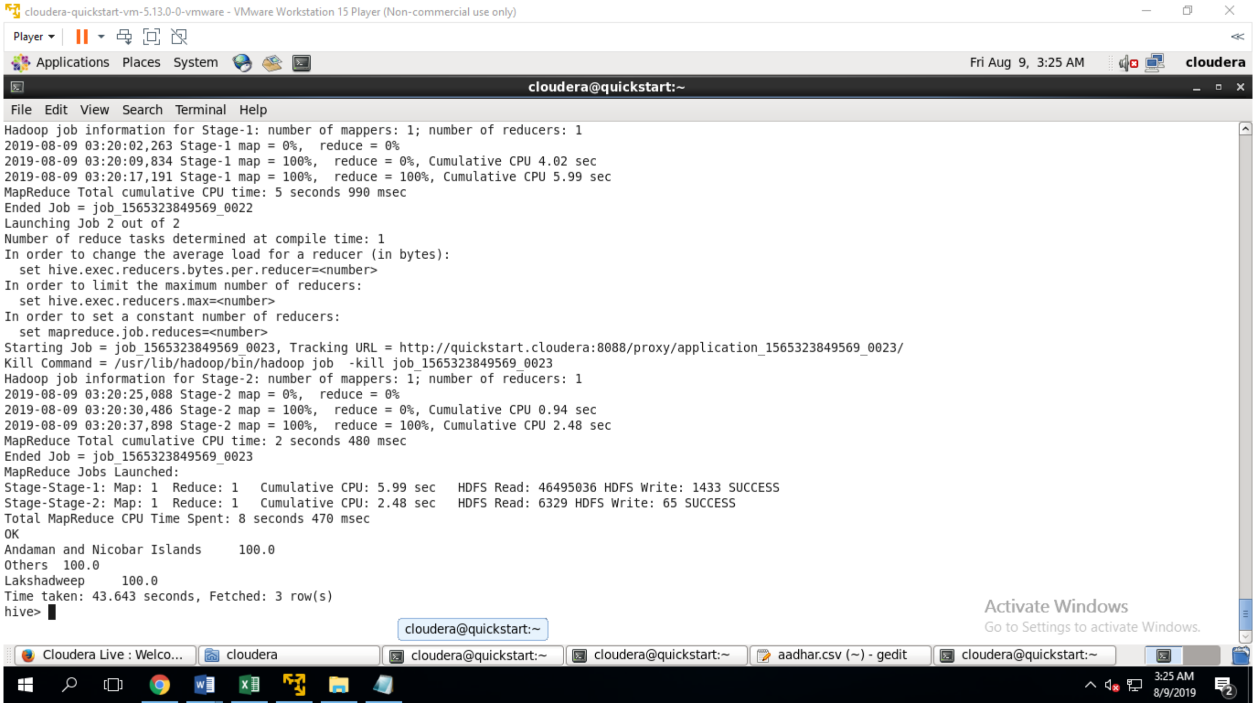
OK

Andaman and Nicobar Islands   100.0

Others   100.0

Lakshadweep   100.0

Time taken: 43.643 seconds, Fetched: 3 row(s)



```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 03:20:02,263 Stage-1 map = 0%,  reduce = 0%
2019-08-09 03:20:09,834 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.02 sec
2019-08-09 03:20:17,191 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.99 sec
MapReduce Total cumulative CPU time: 5 seconds 990 msec
Ended Job = job_1565323849569_0022
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_1565323849569_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0023
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-08-09 03:20:25,088 Stage-2 map = 0%,  reduce = 0%
2019-08-09 03:20:30,486 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.94 sec
2019-08-09 03:20:37,898 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.48 sec
MapReduce Total cumulative CPU time: 2 seconds 480 msec
Ended Job = job_1565323849569_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.99 sec   HDFS Read: 46495036 HDFS Write: 1433 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.48 sec   HDFS Read: 6329 HDFS Write: 65 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 470 msec
OK
Andaman and Nicobar Islands     100.0
Others  100.0
Lakshadweep     100.0
Time taken: 43.643 seconds, Fetched: 3 row(s)
hive> 
```

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

hive> select district,(sum(enrolmentRejected)*100/(sum(aadhaarGenerated+enrolmentRejected))) as female_rejections from aadhar_ex where gender = 'F' and state in('Andaman and Nicobar Islands','Others','Lakshadweep')group by district order by female_rejections desc limit 3;

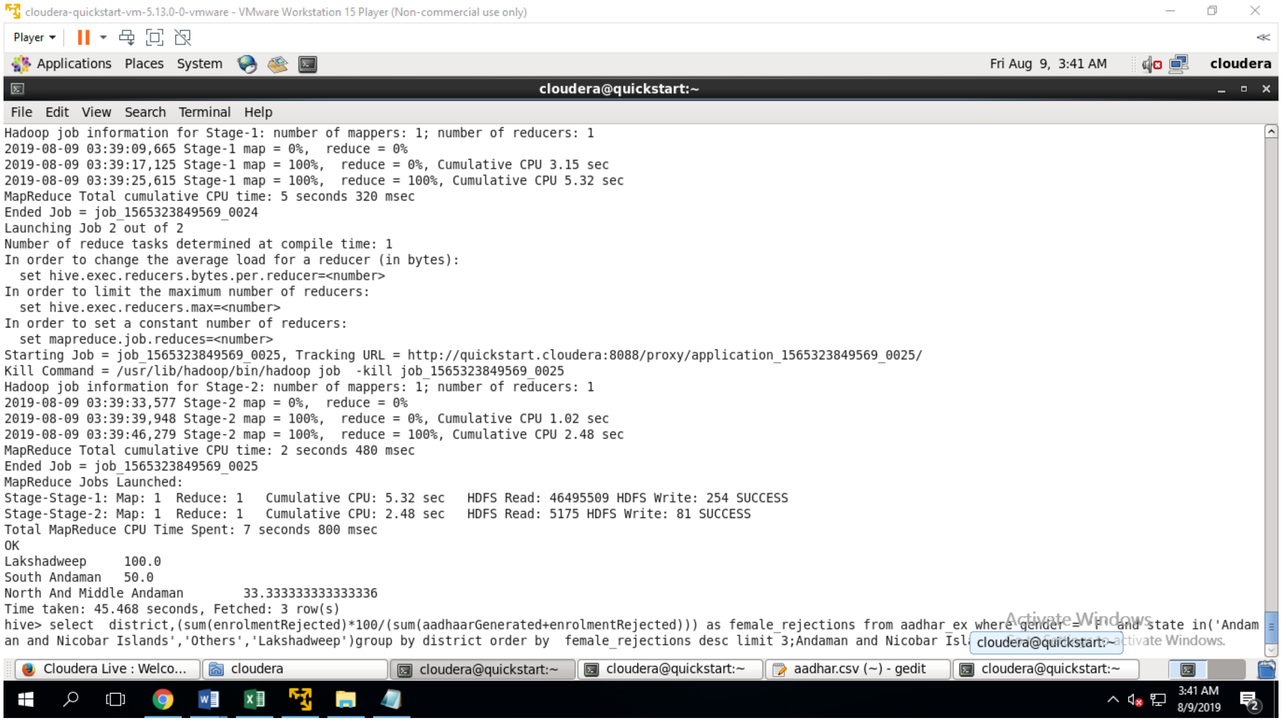
OK

Lakshadweep   100.0

South Andaman50.0

North And Middle Andaman   33.333333333333336

Time taken: 45.468 seconds, Fetched: 3 row(s)



```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 03:39:09,665 Stage-1 map = 0%,  reduce = 0%
2019-08-09 03:39:17,125 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.15 sec
2019-08-09 03:39:25,615 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.32 sec
MapReduce Total cumulative CPU time: 5 seconds 320 msec
Ended Job = job_1565323849569_0024
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_1565323849569_0025, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0025
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-08-09 03:39:33,577 Stage-2 map = 0%,  reduce = 0%
2019-08-09 03:39:39,948 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.02 sec
2019-08-09 03:39:46,279 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.48 sec
MapReduce Total cumulative CPU time: 2 seconds 480 msec
Ended Job = job_1565323849569_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.32 sec   HDFS Read: 46495509 HDFS Write: 254 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.48 sec   HDFS Read: 5175 HDFS Write: 81 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 800 msec
OK
Lakshadweep     100.0
South Andaman   50.0
North And Middle Andaman        33.333333333333336
Time taken: 45.468 seconds, Fetched: 3 row(s)
hive> select  district,(sum(enrolmentRejected)*100/(sum(aadhaarGenerated+enrolmentRejected))) as female_rejections from aadhar_ex where gender = 'F' and state in('Andam
an and Nicobar Islands','Others','Lakshadweep')group by district order by  female_rejections desc limit 3;Andaman and Nicobar Isl
```

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

select state,(sum(aadhaarGenerated)*100)/(sum(aadhaarGenerated+enrolmentRejected))as female_num from aadhar_ex where gender='F' group by state order by female_num desc limit 3;

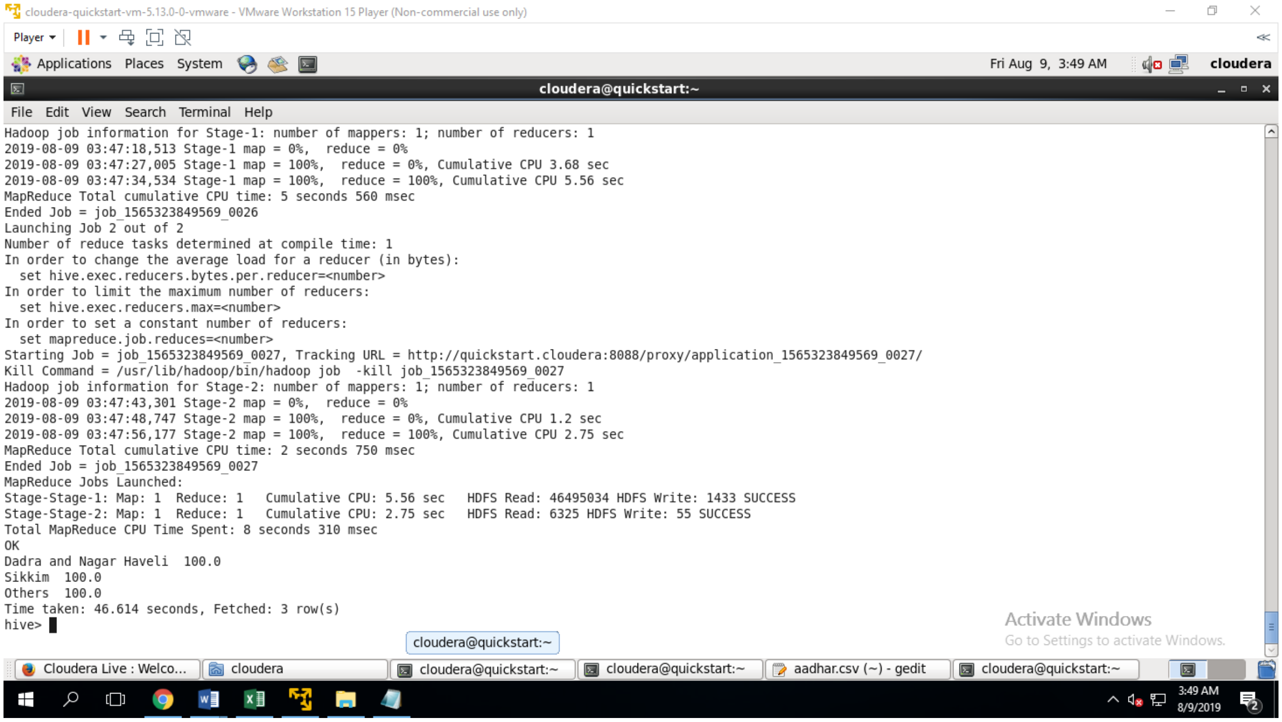
OK

Dadra and Nagar Haveli 100.0

Sikkim   100.0

Others   100.0

Time taken: 46.614 seconds, Fetched: 3 row(s)



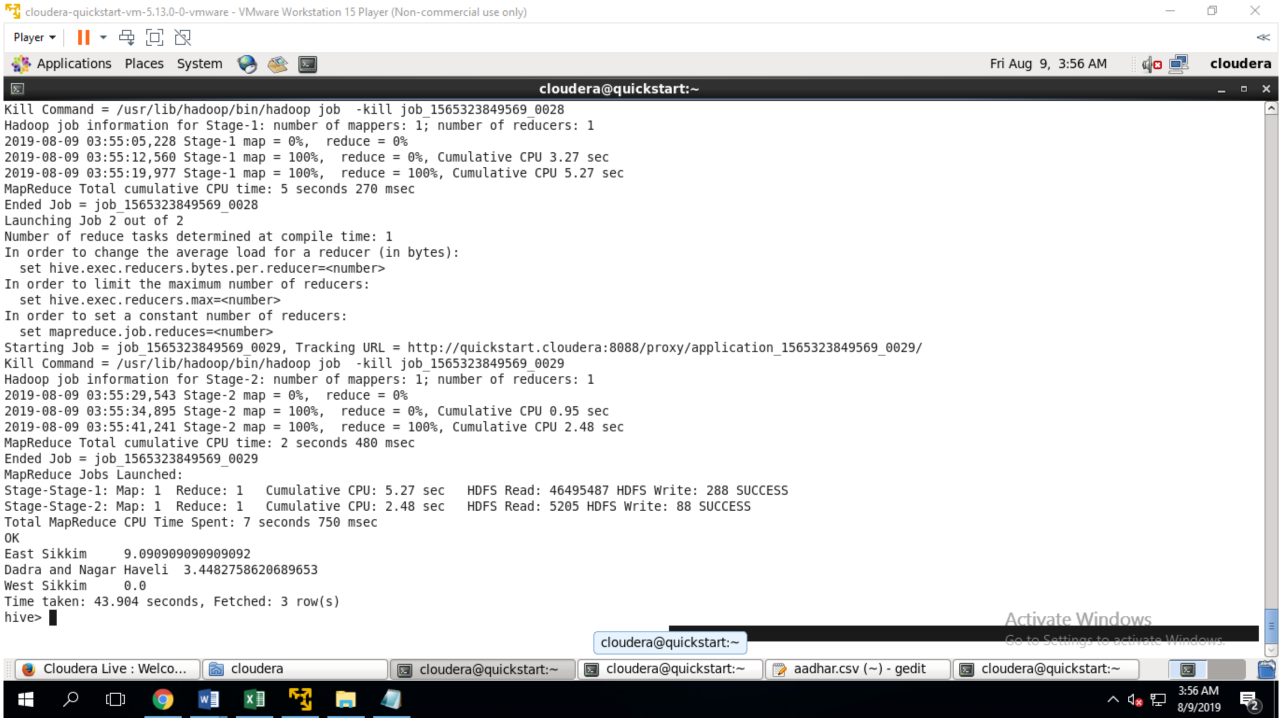
cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player
Applications Places System
Fri Aug 9, 3:49 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Activate Windows
Go to Settings to activate Windows.
Cloudera Live : Welco...
cloudera
cloudera@quickstart:~
aadhar.csv (~) - gedit
3:49 AM
8/9/2019

```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 03:47:18,513 Stage-1 map = 0%,  reduce = 0%
2019-08-09 03:47:27,005 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.68 sec
2019-08-09 03:47:34,534 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.56 sec
MapReduce Total cumulative CPU time: 5 seconds 560 msec
Ended Job = job_1565323849569_0026
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_1565323849569_0027, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0027/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0027
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-08-09 03:47:43,301 Stage-2 map = 0%,  reduce = 0%
2019-08-09 03:47:48,747 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.2 sec
2019-08-09 03:47:56,177 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.75 sec
MapReduce Total cumulative CPU time: 2 seconds 750 msec
Ended Job = job_1565323849569_0027
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.56 sec   HDFS Read: 46495034 HDFS Write: 1433 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.75 sec   HDFS Read: 6325 HDFS Write: 55 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 310 msec
OK
Dadra and Nagar Haveli  100.0
Sikkim  100.0
Others  100.0
Time taken: 46.614 seconds, Fetched: 3 row(s)
hive> 
```

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(enrolmentrejected)*100)/(sum(aadhaarGenerated+enrolmentRejected))as male_rejections from aadhar_ex where gender='M' and state in ('Dadra and Nagar Haveli','Sikkim','Others') group by district order by male_rejections desc limit 3;

```
OK
East Sikkim            9.090909090909092
Dadra and Nagar Haveli   3.4482758620689653
West Sikkim            0.0
      Time taken: 43.904 seconds, Fetched: 3 row(s)
```



```
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0028
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 03:55:05,228 Stage-1 map = 0%,  reduce = 0%
2019-08-09 03:55:12,560 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.27 sec
2019-08-09 03:55:19,977 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.27 sec
MapReduce Total cumulative CPU time: 5 seconds 270 msec
Ended Job = job_1565323849569_0028
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_1565323849569_0029, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0029/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0029
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-08-09 03:55:29,543 Stage-2 map = 0%,  reduce = 0%
2019-08-09 03:55:34,895 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.95 sec
2019-08-09 03:55:41,241 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.48 sec
MapReduce Total cumulative CPU time: 2 seconds 480 msec
Ended Job = job_1565323849569_0029
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.27 sec   HDFS Read: 46495487 HDFS Write: 288 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.48 sec   HDFS Read: 5205 HDFS Write: 88 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 750 msec
OK
East Sikkim     9.090909090909092
Dadra and Nagar Haveli  3.4482758620689653
West Sikkim     0.0
Time taken: 43.904 seconds, Fetched: 3 row(s)
hive> 
```

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

hive> create table if not exists aadhar_bucket(registrar String,enrolment_Agency String,State String,district String,sub_District String,pin_code String,gender String,age int,aadhaarGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNumber int)

> clustered by(age)into 10 buckets

> row format delimited fields terminated by ','

> stored as textfile

> TBLPROPERTIES('serialization.null.format'='',

> 'skip.header.line.count'='1');

OK

Time taken: 0.059 seconds

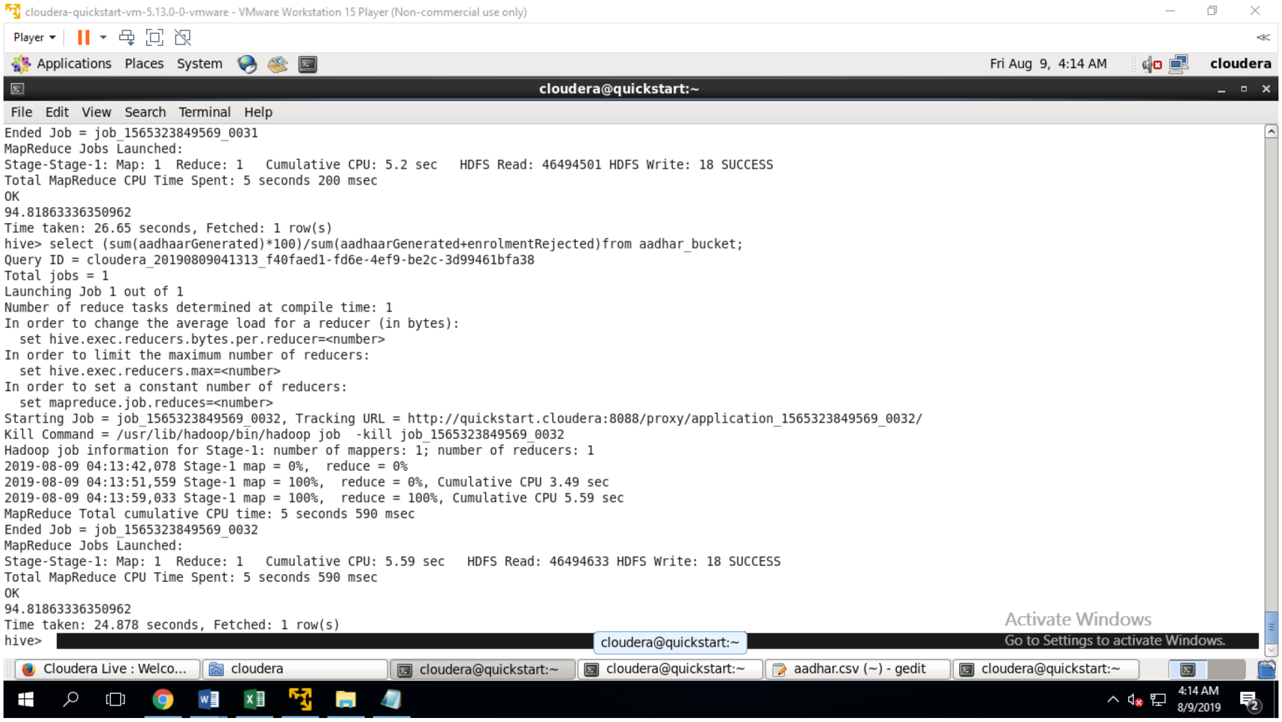
insert into aadhar_bucket select * from aadhar_ex;

select (sum(aadhaarGenerated)*100)/sum(aadhaarGenerated+enrolmentRejected)from aadhar_bucket;

OK

94.81863336350962

Time taken: 24.878 seconds, Fetched: 1 row(s)



cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player
Applications Places System
Fri Aug 9, 4:14 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Ended Job = job_1565323849569_0031
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.2 sec   HDFS Read: 46494501 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 200 msec
OK
94.81863336350962
Time taken: 26.65 seconds, Fetched: 1 row(s)
hive> select (sum(aadhaarGenerated)*100)/sum(aadhaarGenerated+enrolmentRejected)from aadhar_bucket;
Query ID = cloudera_20190809041313_f40faed1-fd6e-4ef9-be2c-3d99461bfa38
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_1565323849569_0032, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1565323849569_0032/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1565323849569_0032
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-08-09 04:13:42,078 Stage-1 map = 0%,  reduce = 0%
2019-08-09 04:13:51,559 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.49 sec
2019-08-09 04:13:59,033 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.59 sec
MapReduce Total cumulative CPU time: 5 seconds 590 msec
Ended Job = job_1565323849569_0032
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.59 sec   HDFS Read: 46494633 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 590 msec
OK
94.81863336350962
Time taken: 24.878 seconds, Fetched: 1 row(s)
hive>
```