Checkpoint 1

Q1: Load the data into HDFS, Hive Managed table, Hive External table and Spark DataFrame.

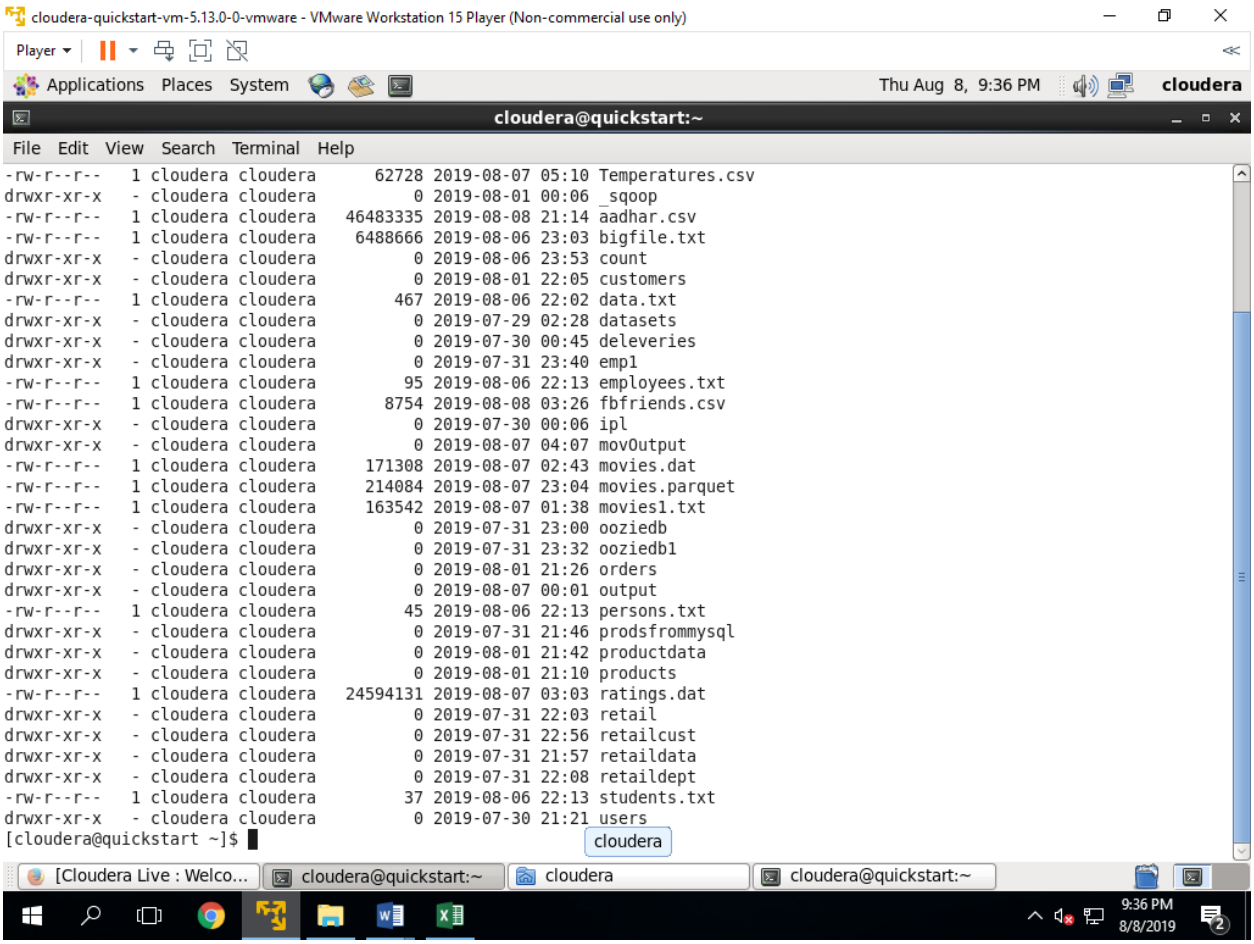
1. **Commit the screenshot of the view/result of the top 25 rows from each individual store (HDFS, Hive – Managed/External and Spark DataFrame).**

1: //loading the aadhar.csv file in hdfs by first placing it in the home folder and then putting it in HDFS using

hdfs dfs -put /home/cloudera/aadhar.csv

//to make sure that the file is there in hdfs

hdfs dfs –ls



//creating and using the database in hive

create database if not exists aadhar;

use aadhar;

//creating the table and loading the data from the hdfs

create table if not exists aadharInfo(registrar string,enrolment_agency string,state string,district string,subDistrict string,pincode string,gender string,age int,aadharGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNo int)

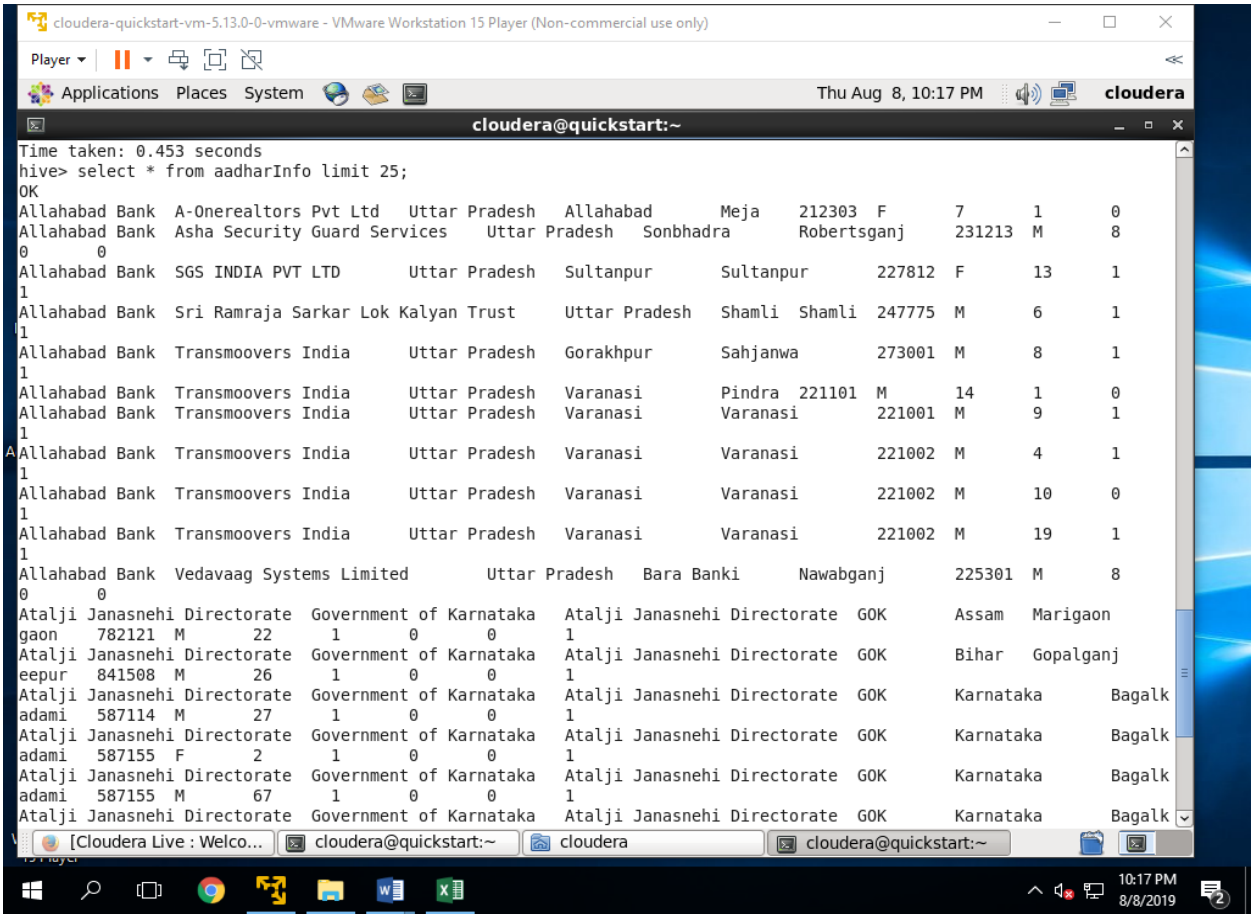
>row format delimited fields terminated by ','

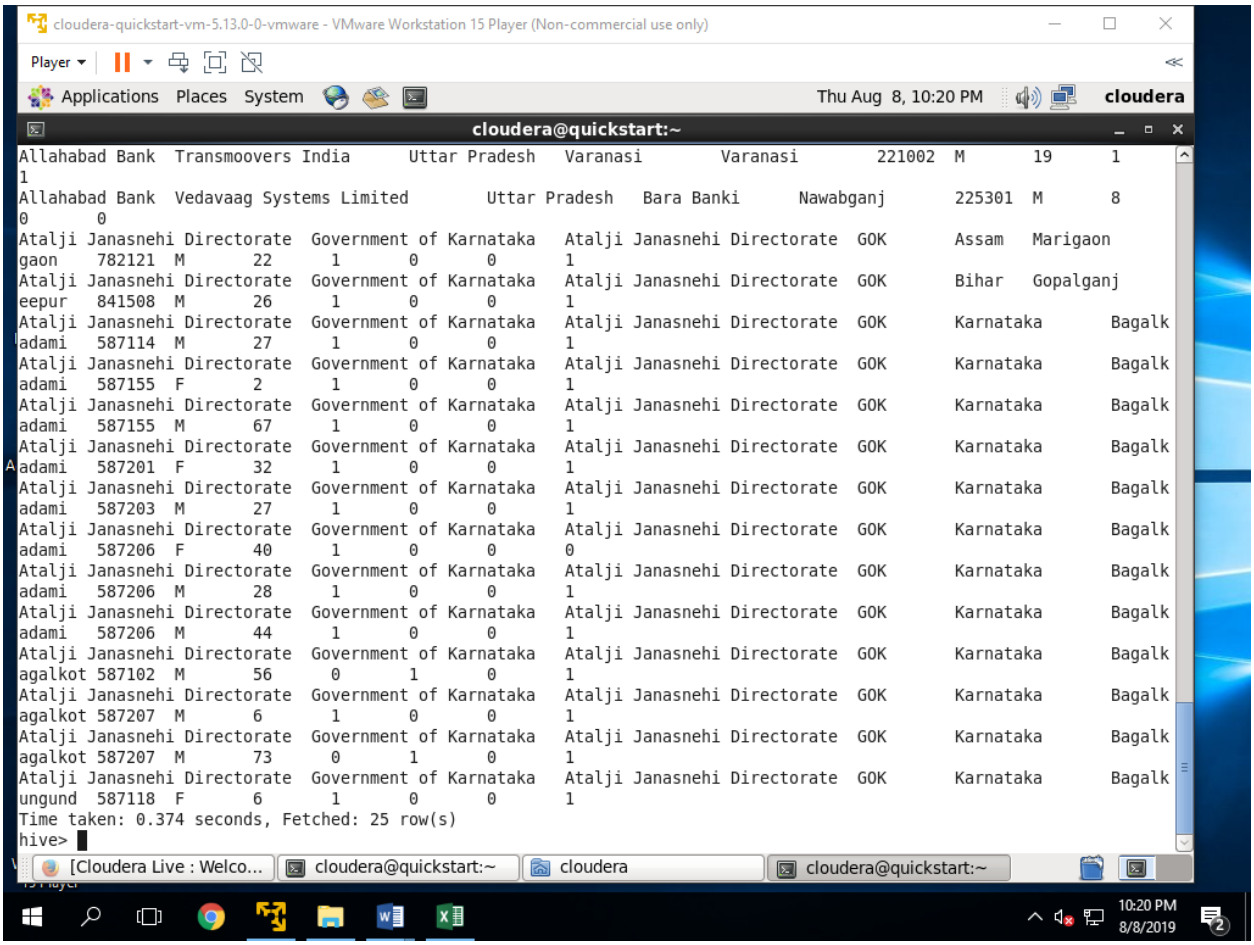
>TBLPROPERTIES('skip.header.line.count'='1');

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

//result of top 25 rows stored in hive managed table

select * from aadharInfo limit 25;



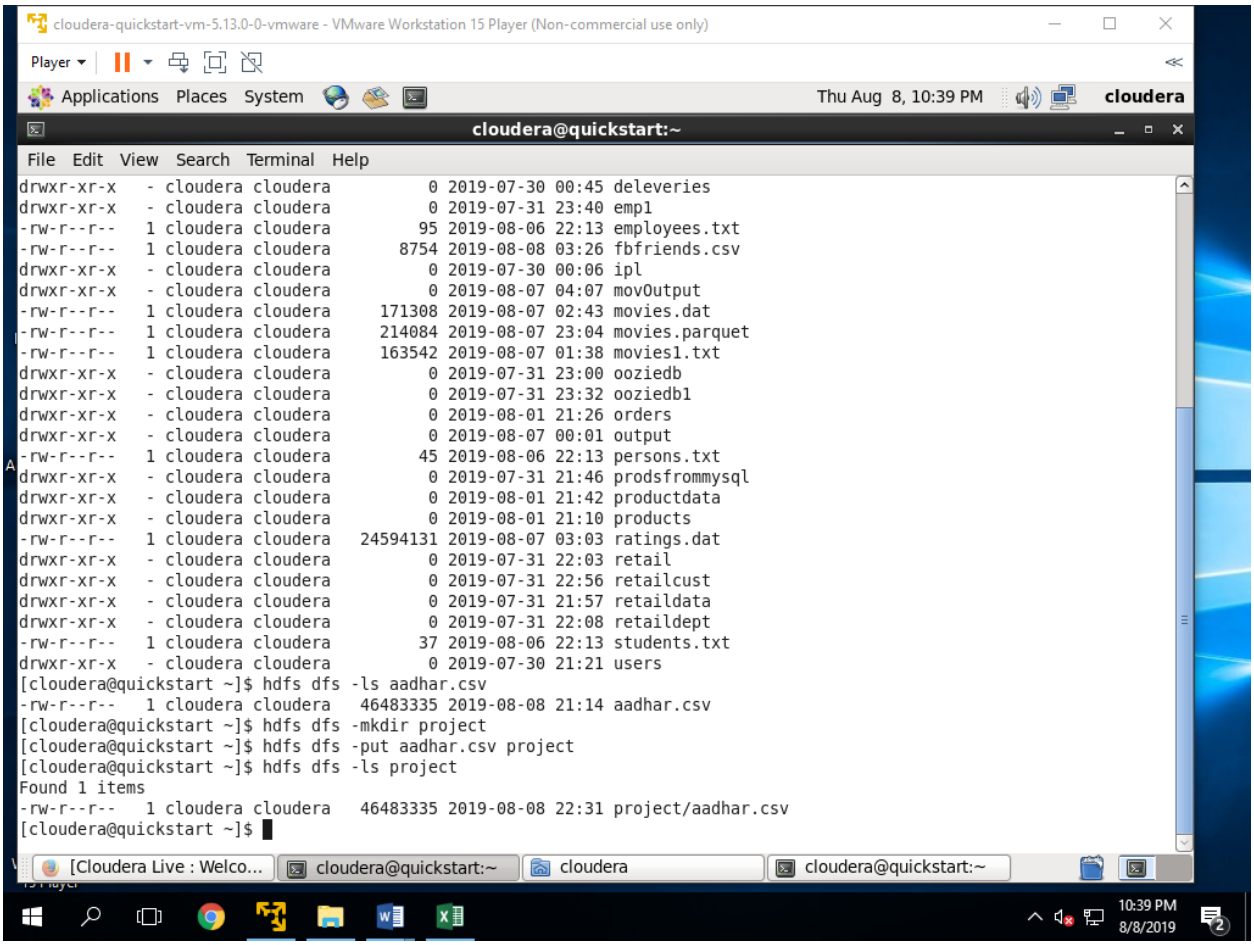


//making a project directory in hdfs and putting the aadhar.csv file in that directory

hdfs dfs -mkdir project

hdfs dfs -put aadhar.csv project

hdfs dfs -ls project



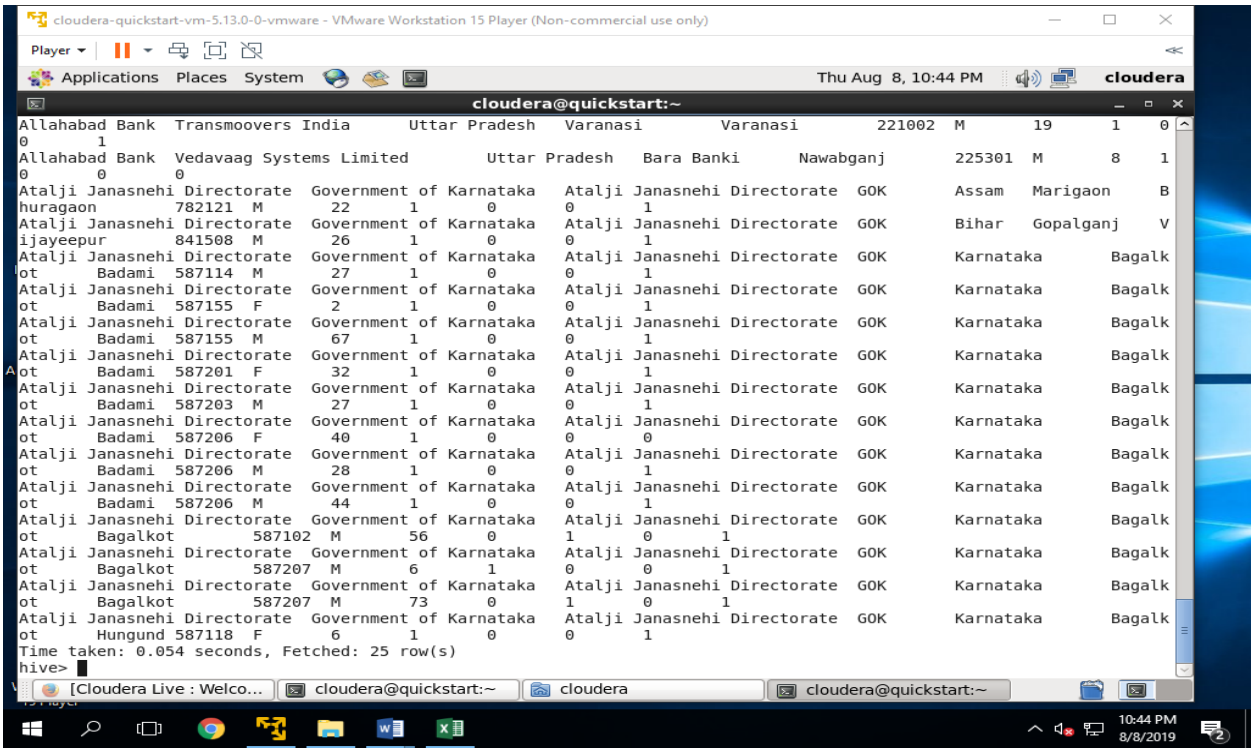
//creating the external table in hive and loading data from the project directory

create external table if not exists aadharInfoExternal(registrar string,enrolment_agency string,state string,district string,subDistrict string,pincode string,gender string,age int,aadharGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNo int)

> row format delimited fields terminated by ','

> location '/user/cloudera/project'

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



//creating dataframe in spark

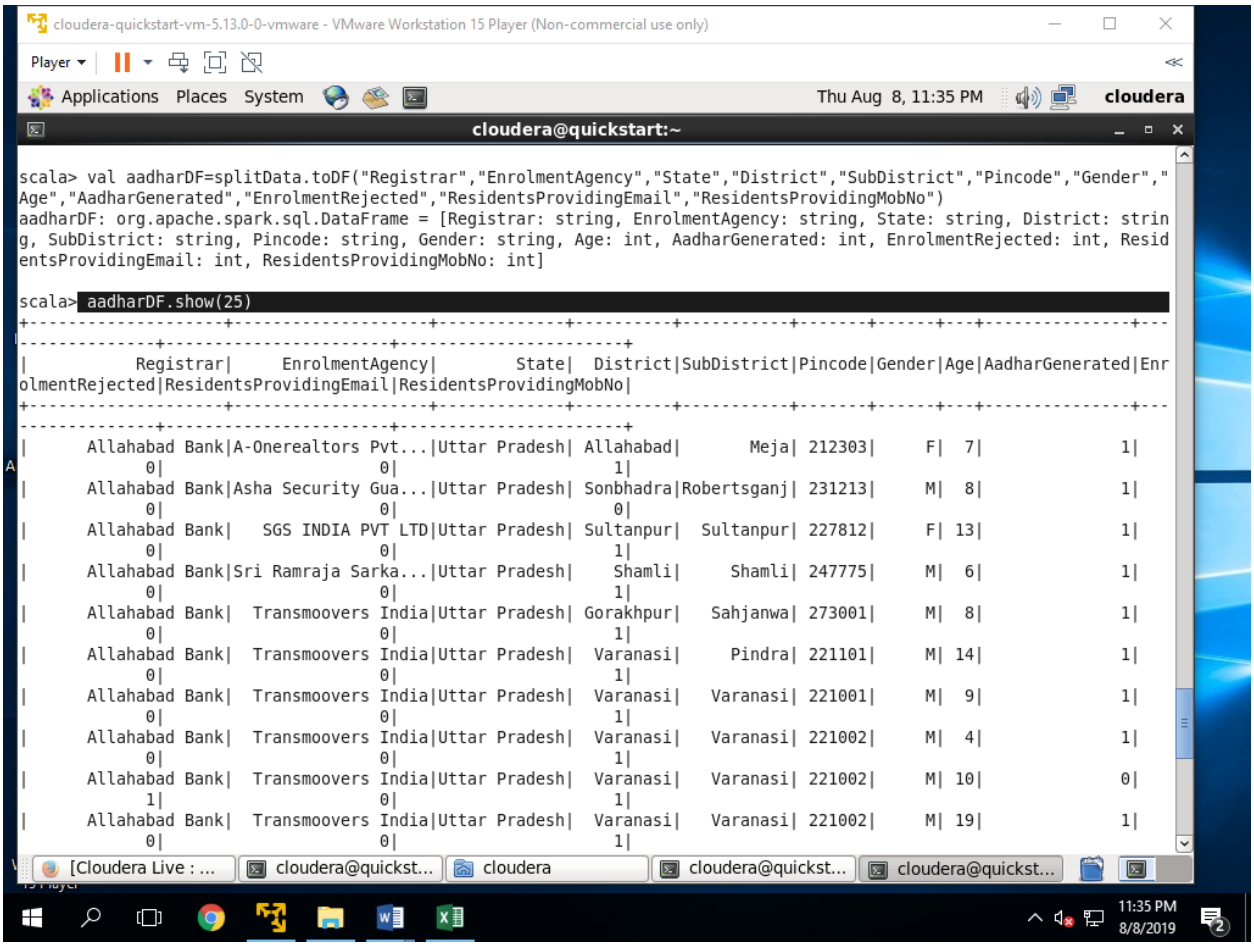
val aadharRDD=sc.textFile("/user/cloudera/project/aadhar.csv")

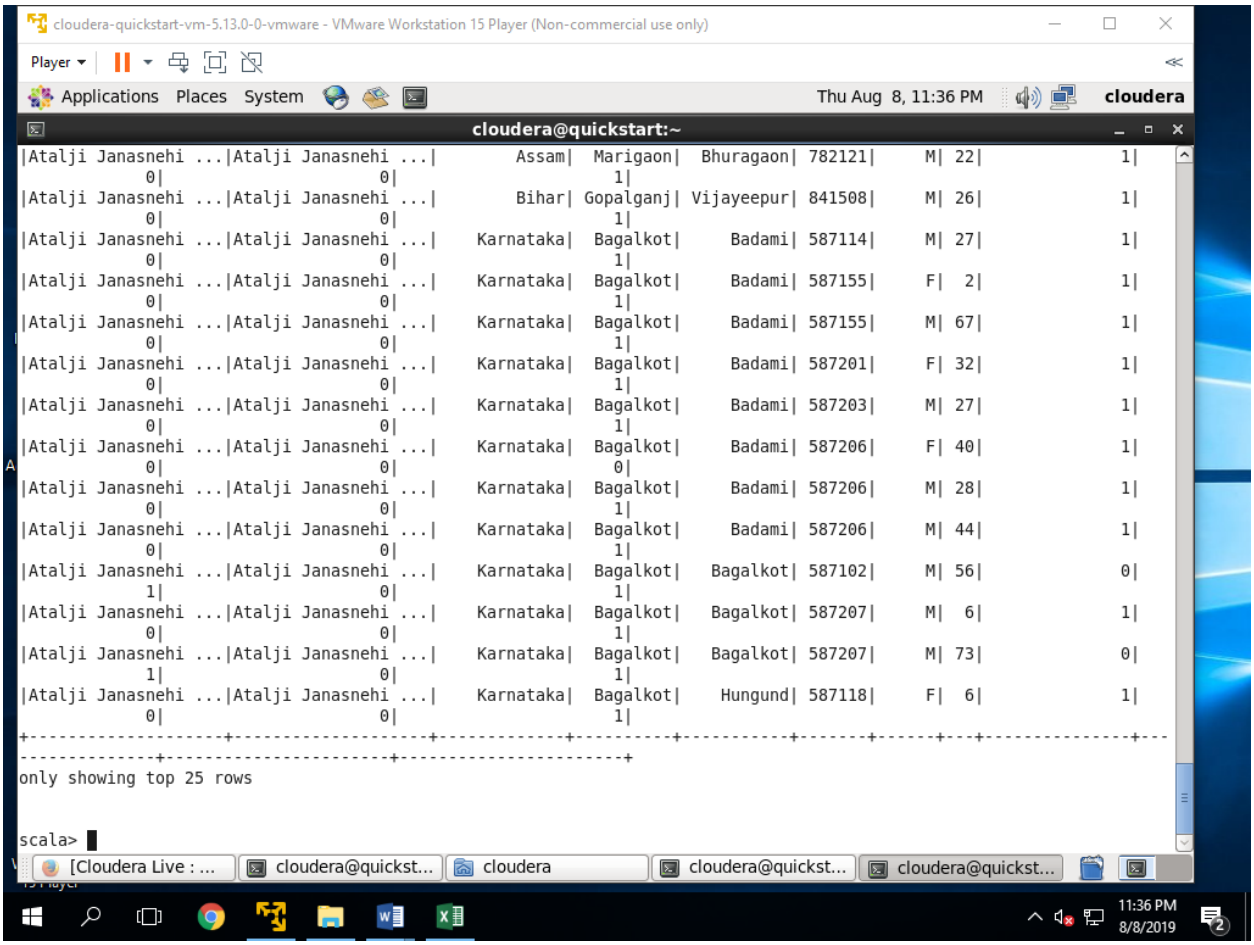
//removing the header

```
val remHeader=aadharRDD.first()

 val final_data=aadharRDD.filter(row=>row!=(remHeader))

val splitData=final_data.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))
```

//creating dataframe

```
val aadharDF=splitData.toDF("Registrar","EnrolmentAgency","State","District","SubDistrict","Pincode","Gender","Age","AadharGenerated","EnrolmentRejected","ResidentsProvidingEmail","ResidentsProvidingMobNo")

aadharDF.show(25)
```





Checkpoint 2

Q2: **Describe the schema.**

aadharDF.printSchema

OUTPUT

root
 |-- Registrar: string (nullable = true)
 |-- EnrolmentAgency: 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)
|-- ResidentsProvidingMobNo: integer (nullable = false)

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

select registrar,count(registrar) from aadharInfo group by registrar limit 10;

OUTPUT

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 Aalo126
DC ITANAGAR CAPITAL COMPLEX 38
DC LOHIT 119
Time taken: 34.019 seconds, Fetched: 10 row(s)

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

a) select count(distinct(state)) from aadharInfo;

37
Time taken: 21.057 seconds, Fetched: 1 row(s)

b) select state,count(distinct(district)) from aadharInfo group by state;

OUTPUT

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: 20.829 seconds, Fetched: 37 row(s)

c) select district,count(distinct(subDistrict)) from aadharInfo group by district;

OUTPUT

Uttara Kannada 13
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
Wardha8
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: 23.728 seconds, Fetched: 664 row(s)

Q6: **Find out the names of private agencies for each state.**

select distinct(state),enrolment_agency from aadharInfo;

OUTPUT

West Bengal SRM Education And Social Welfare Society
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: 28.2 seconds, Fetched: 2271 row(s)

Checkpoint 3

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

```
select state,sum(aadharGenerated)as no from aadharInfo group by state order by (no) desc limit
3;
```

OUTPUT

Bihar 162607

West Bengal 119901
Uttar Pradesh 103767
Time taken: 44.134 seconds, Fetched: 3 row(s)

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

```
select enrolment_agency,count(aadharGenerated)as no from aadharInfo group by
   enrolment_agency order by (no) desc limit 3;
```

OUTPUT

CSC SPV 100357
SRM Education And Social Welfare Society 18101
SREI INFRASTRUCTURE FINANCES L 16972
Time taken: 41.707 seconds, Fetched: 3 row(s)

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

```
 select count(residentsprovidingemail)as NoOfEmailProvider from aadharInfo where
     (residentsprovidingemail !=0 AND residentsprovidingmobileno !=0);
```

OUTPUT

16951
Time taken: 26.638 seconds, Fetched: 1 row(s)

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

```
select district,count(enrolmentRejected)as noofEnrolments from aadharInfo where
     enrolmentRejected =0 group by district order by (noofenrolments) desc limit 3;
```

OUTPUT

Barddhaman 6726
North 24 Parganas 6534
South 24 Parganas 5603

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

select state,sum(aadharGenerated)as no from aadharInfo group by state;

OUTPUT

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: 23.881 seconds, Fetched: 37 row(s)

Checkpoint 4

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

//creating dataframe in spark

```
val aadharRDD=sc.textFile("/user/cloudera/project/aadhar.csv")
```

//removing the header

```
val remHeader=aadharRDD.first()

 val final_data=aadharRDD.filter(row=>row!=(remHeader))

val
splitData=final_data.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))
```

//creating dataframe

```
val
aadharDF=splitData.toDF("Registrar","EnrolmentAgency","State","District","SubDistrict","Pincode","Gender","Age","AadharGenerated","EnrolmentRejected","ResidentsProvidingEmail","ResidentsProvidingMobNo")

aadharDF.show(25)
```

//providing the summary

```
aadharDF.describe()
res16: org.apache.spark.sql.DataFrame = [summary: string, Age: string, AadharGenerated: string, EnrolmentRejected: string, ResidentsProvidingEmail: string, ResidentsProvidingMobNo: string]
```

14: Write a command to see the correlation between "age" and "mobile_number"? (Hint: Consider the percentage of people who have provided the mobile number out of the total applicants)

select corr(age,residentsprovidingmobileno) from aadharInfo;

OUTPUT

-0.11754461896889339
Time taken: 23.431 seconds, Fetched: 1 row(s)

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

select count(distinct(pincode)) from aadharInfo;

OUTPUT

17756
Time taken: 21.717 seconds, Fetched: 1 row(s)

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

select state,sum(enrolmentRejected)as noofEnrolmentsRejected from aadharInfo where (state in ("Uttar Pradesh","Maharashtra") AND enrolmentRejected!=0) group by state;

OUTPUT

Maharashtra 1818
Uttar Pradesh 5286
Time taken: 22.982 seconds, Fetched: 2 row(s)

Checkpoint 5

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

select state,((sum(aadharGenerated)/(sum(aadharGenerated+enrolmentRejected)))*100)as noOfAadharGenerated from aadharInfo where gender="M" group by state order by (noOfAadharGenerated) desc limit 3;

OUTPUT

Andaman and Nicobar Islands 100.0
Others 100.0
Lakshadweep 100.0

Time taken: 41.915 seconds, Fetched: 3 row(s)

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

select district,((sum(enrolmentRejected)/(sum(aadharGenerated+enrolmentRejected)))*100)as cnt from aadharInfo where (gender="F" AND state in("Andaman and Nicobar Islands","Others","Lakshadweep")) group by district order by (cnt) desc limit 3;

OUTPUT

Lakshadweep 100.0
South Andaman50.0
North And Middle Andaman 33.33333333333333
Time taken: 42.232 seconds, Fetched: 3 row(s)

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

select state,((sum(aadharGenerated)/(sum(aadharGenerated+enrolmentRejected)))*100)as noOfAadharGenerated from aadharInfo where gender="F" group by state order by (noOfAadharGenerated) desc limit 3;

OUTPUT

Dadra and Nagar Haveli 100.0
Sikkim 100.0
Others 100.0
Time taken: 46.093 seconds, Fetched: 3 row(s)

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)/(sum(aadharGenerated+enrolmentRejected)))*100)as cnt from aadharInfo where (gender="M" AND state in("Dadra and Nagar Haveli","Sikkim","Others")) group by district order by (cnt) desc limit 3;

OUTPUT

East Sikkim 9.090909090909092
Dadra and Nagar Haveli 3.4482758620689653
West Sikkim 0.0
Time taken: 46.008 seconds, Fetched: 3 row(s)

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

//creating a bucketed table

```
create table if not exists aadharInfo_bucket(registrar string,enrolment_agency string,state string,district string,subDistrict string,pincode string,gender string,age int,aadharGenerated int,enrolmentRejected int,residentsProvidingEmail int,residentsProvidingMobileNo int)
   > clustered by (age) into 10 buckets
   > row format delimited fields terminated by ',';
```

//inserting data from staging table

```
insert into aadharInfo_bucket select * from aadharInfo;

select ((sum(aadharGenerated)/(sum(aadharGenerated+enrolmentRejected)))*100)as acceptancePercentage from aadharInfo_bucket;
```

OUTPUT

94.81864032289477