

## 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
```

```
cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player | [Icons] | Thu Aug 8, 9:36 PM | cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
-rw-r--r-- 1 cloudera cloudera 62728 2019-08-07 05:10 Temperatures.csv
drwxr-xr-x - cloudera cloudera 0 2019-08-01 00:06 _sqoop
-rw-r--r-- 1 cloudera cloudera 46483335 2019-08-08 21:14 aadhar.csv
-rw-r--r-- 1 cloudera cloudera 6488666 2019-08-06 23:03 bigfile.txt
drwxr-xr-x - cloudera cloudera 0 2019-08-06 23:53 count
drwxr-xr-x - cloudera cloudera 0 2019-08-01 22:05 customers
-rw-r--r-- 1 cloudera cloudera 467 2019-08-06 22:02 data.txt
drwxr-xr-x - cloudera cloudera 0 2019-07-29 02:28 datasets
drwxr-xr-x - cloudera cloudera 0 2019-07-30 00:45 deleveries
drwxr-xr-x - cloudera cloudera 0 2019-07-31 23:40 empl
-rw-r--r-- 1 cloudera cloudera 95 2019-08-06 22:13 employees.txt
-rw-r--r-- 1 cloudera cloudera 8754 2019-08-08 03:26 fbfriends.csv
drwxr-xr-x - cloudera cloudera 0 2019-07-30 00:06 ipl
drwxr-xr-x - cloudera cloudera 0 2019-08-07 04:07 movOutput
-rw-r--r-- 1 cloudera cloudera 171308 2019-08-07 02:43 movies.dat
-rw-r--r-- 1 cloudera cloudera 214084 2019-08-07 23:04 movies.parquet
-rw-r--r-- 1 cloudera cloudera 163542 2019-08-07 01:38 movies1.txt
drwxr-xr-x - cloudera cloudera 0 2019-07-31 23:00 ooziedb
drwxr-xr-x - cloudera cloudera 0 2019-07-31 23:32 ooziedb1
drwxr-xr-x - cloudera cloudera 0 2019-08-01 21:26 orders
drwxr-xr-x - cloudera cloudera 0 2019-08-07 00:01 output
-rw-r--r-- 1 cloudera cloudera 45 2019-08-06 22:13 persons.txt
drwxr-xr-x - cloudera cloudera 0 2019-07-31 21:46 prodsfrommysql
drwxr-xr-x - cloudera cloudera 0 2019-08-01 21:42 productdata
drwxr-xr-x - cloudera cloudera 0 2019-08-01 21:10 products
-rw-r--r-- 1 cloudera cloudera 24594131 2019-08-07 03:03 ratings.dat
drwxr-xr-x - cloudera cloudera 0 2019-07-31 22:03 retail
drwxr-xr-x - cloudera cloudera 0 2019-07-31 22:56 retailcust
drwxr-xr-x - cloudera cloudera 0 2019-07-31 21:57 retaildata
drwxr-xr-x - cloudera cloudera 0 2019-07-31 22:08 retaildept
-rw-r--r-- 1 cloudera cloudera 37 2019-08-06 22:13 students.txt
drwxr-xr-x - cloudera cloudera 0 2019-07-30 21:21 users
cloudera@quickstart ~]$
```

## //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;

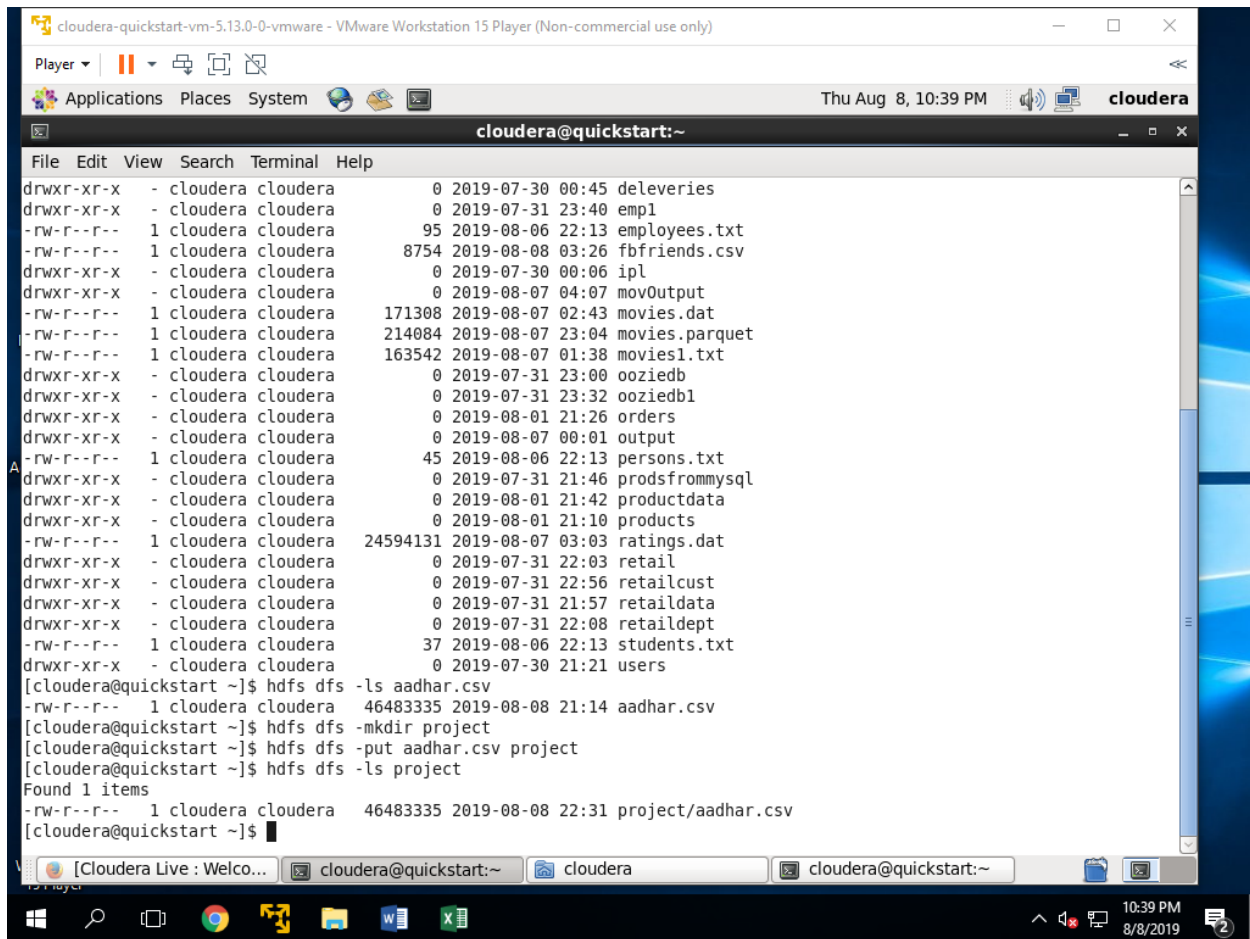
The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)". The terminal displays the following commands and output:

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

OK

Allahabad Bank	A-Onerealtors Pvt Ltd	Uttar Pradesh	Allahabad	Meja	212303	F	7	1	0
Allahabad Bank	Asha Security Guard Services	Uttar Pradesh	Sonbhadra	Robertsganj	231213	M	8	0	0
Allahabad Bank	SGS INDIA PVT LTD	Uttar Pradesh	Sultanpur	Sultanpur	227812	F	13	1	1
Allahabad Bank	Sri Ramaja Sarkar Lok Kalyan Trust	Uttar Pradesh	Shamli	Shamli	247775	M	6	1	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Gorakhpur	Sahjanwa	273001	M	8	1	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Pindra	221101	M	14	1	0
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221001	M	9	1	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	4	1	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	10	0	1
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	19	1	1
Allahabad Bank	Vedavaag Systems Limited	Uttar Pradesh	Bara Banki	Nawabganj	225301	M	8	0	0
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Assam	Marigaon	782121	M	22	1
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Bihar	Gopalganj	841508	M	26	1
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Karnataka	Bagalk	587114	M	27	1
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Karnataka	Bagalk	587155	F	2	1
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Karnataka	Bagalk	587155	M	67	1
Atalji Janasnehi Directorate	Government of Karnataka	Atalji Janasnehi Directorate	GOK	Karnataka	Bagalk				1

```
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');

```

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player | Applications | Places | System | Thu Aug 8, 10:43 PM | cloudera

cloudera@quickstart:~
hive> select * from aadharInfoExternal 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
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
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Assam Marigaon B
huragaon 782121 M 22 1 0 Atalji Janasnehi Directorate GOK Bihar Gopalganj V
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ijayeeppur 841508 M 26 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587114 M 27 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587155 F 2 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587155 M 67 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
0 1

```

```

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)
Player | Applications | Places | System | Thu Aug 8, 10:44 PM | cloudera

cloudera@quickstart:~
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
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Assam Marigaon B
huragaon 782121 M 22 1 0 Atalji Janasnehi Directorate GOK Bihar Gopalganj V
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ijayeeppur 841508 M 26 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587114 M 27 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587155 F 2 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587155 M 67 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587201 F 32 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587203 M 27 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587206 F 40 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587206 M 28 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Badami 587206 M 44 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Bagalkot 587102 M 56 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Bagalkot 587207 M 6 1 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Bagalkot 587207 M 73 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK
ot Hungund 587118 F 6 1 0 Atalji Janasnehi Directorate GOK Karnataka Bagalk
0 1
Time taken: 0.054 seconds, Fetched: 25 row(s)
hive>

```

//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")
```

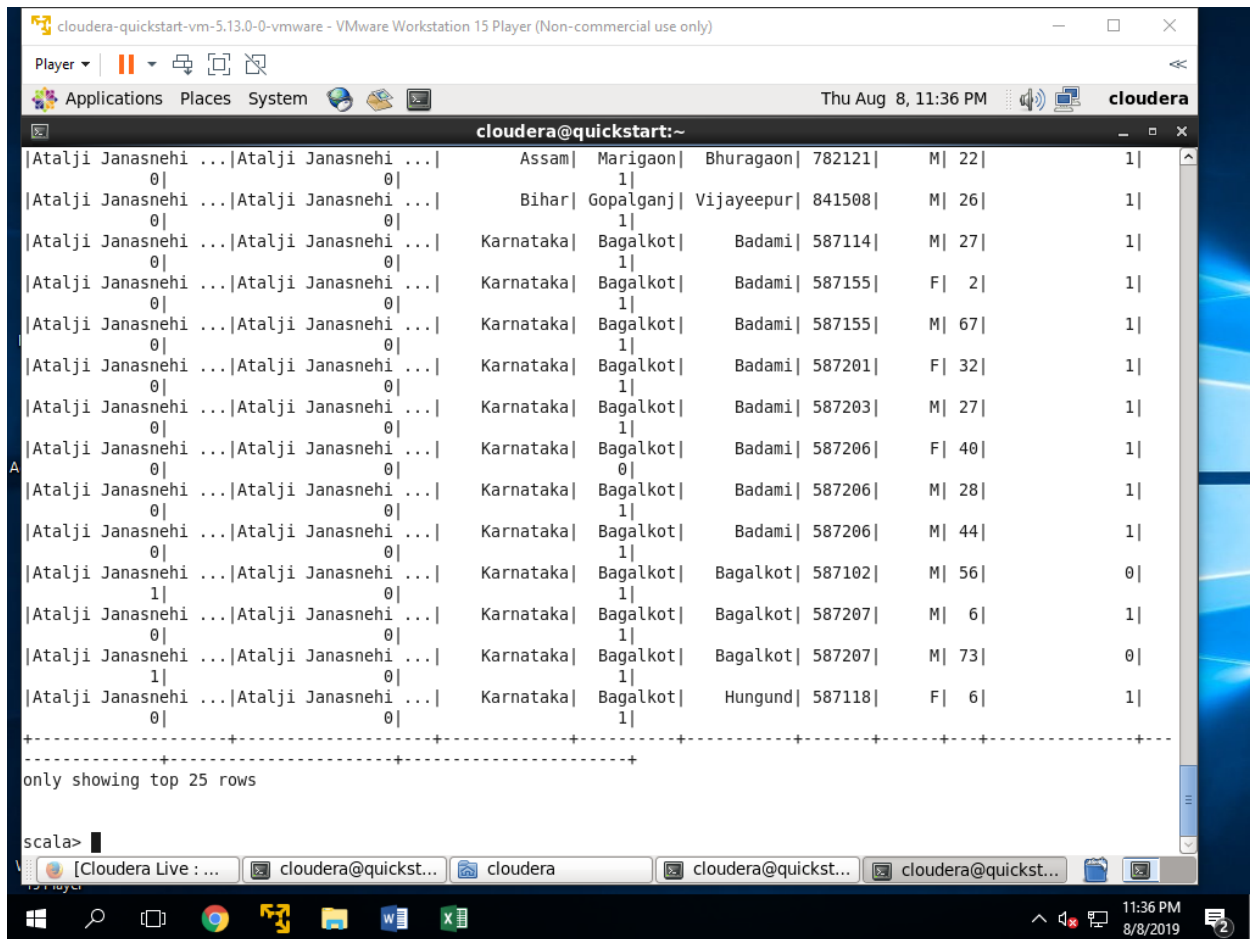
```
aadharDF.show(25)
```

The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)". The terminal displays the following commands and output:

```
scala> val aadharDF=splitData.toDF("Registrar","EnrolmentAgency","State","District","SubDistrict","Pincode","Gender","Age","AadharGenerated","EnrolmentRejected","ResidentsProvidingEmail","ResidentsProvidingMobNo")
aadharDF: org.apache.spark.sql.DataFrame = [Registrar: string, EnrolmentAgency: string, State: string, District: string, SubDistrict: string, Pincode: string, Gender: string, Age: int, AadharGenerated: int, EnrolmentRejected: int, ResidentsProvidingEmail: int, ResidentsProvidingMobNo: int]

scala> aadharDF.show(25)
```

Registrar	EnrolmentAgency	State	District	SubDistrict	Pincode	Gender	Age	AadharGenerated	EnrolmentRejected	ResidentsProvidingEmail	ResidentsProvidingMobNo
Allahabad Bank	A-Onerealtors Pvt...	Uttar Pradesh	Allahabad	Meja	212303	F	7	1			
Allahabad Bank	Asha Security Gua...	Uttar Pradesh	Sonbhadra	Robertsganj	231213	M	8	1			
Allahabad Bank	SGS INDIA PVT LTD	Uttar Pradesh	Sultanpur	Sultanpur	227812	F	13	1			
Allahabad Bank	Sri Ramraja Sarka...	Uttar Pradesh	Shamli	Shamli	247775	M	6	1			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Gorakhpur	Sahjanwa	273001	M	8	1			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Pindra	221101	M	14	1			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221001	M	9	1			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	4	1			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	10	0			
Allahabad Bank	Transmoovers India	Uttar Pradesh	Varanasi	Varanasi	221002	M	19	1			



## 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	
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: 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 residentsprovidingmobileneno !=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**

Bardhaman 6726

North 24 Parganas 6534

South 24 Parganas 5603

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

`select state,sum(aadhaarGenerated)as no from aadhaarInfo group by state;`

### OUTPUT

Andaman and Nicobar Islands	5
Andhra Pradesh	5798
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,residentsprovidingmobilenos) 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