

CHECKPOINT 1

//Creating a Directory in hdfs

```
hdfs dfs -mkdir Aadhar;
```

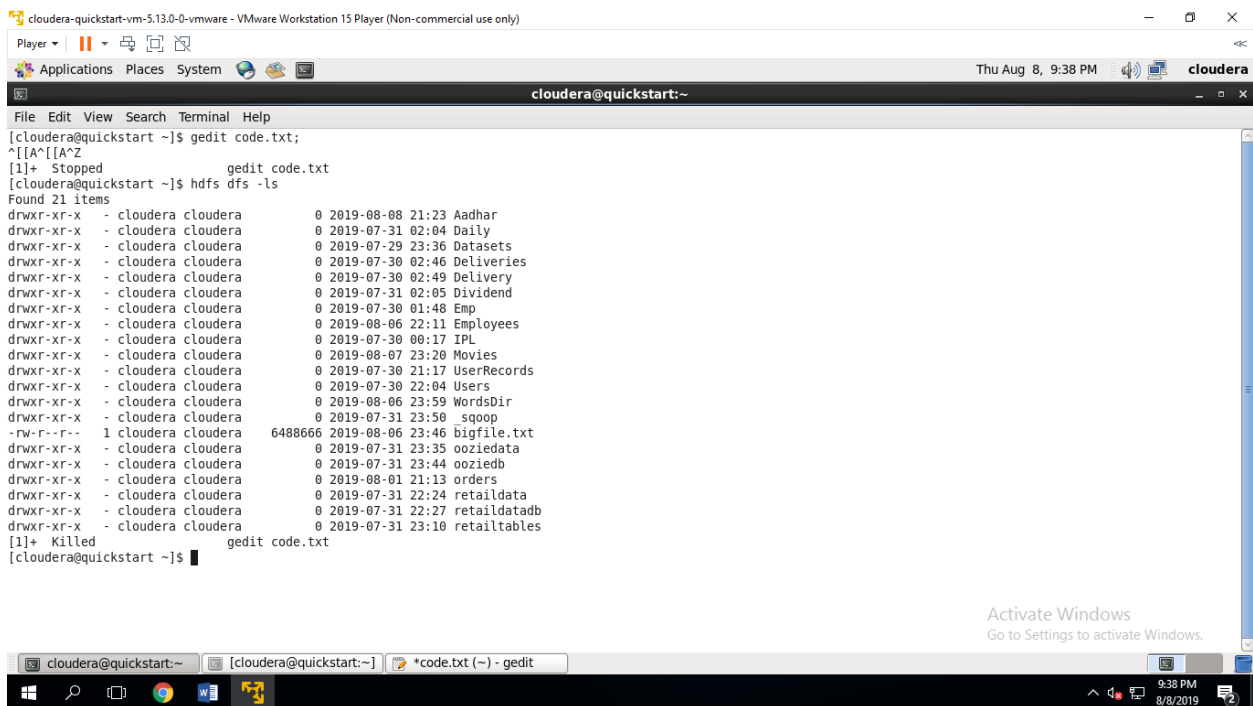
//Loading the dataset into the hdfs directory

```
hdfs dfs -put aadhar.csv Aadhar;
```

//Making sure that the directory is created

```
hdfs dfs -ls;
```

SCREENSHOT:-



The screenshot shows a terminal window titled 'cloudera@quickstart:~' within a VMware Workstation 15 Player. The terminal displays the following commands and output:

```
[cloudera@quickstart ~]$ gedit code.txt;
^[[A^[[A^Z
[1]+  Stopped                  gedit code.txt
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 21 items
drwxr-xr-x - cloudera cloudera      0 2019-08-08 21:23 Aadhar
drwxr-xr-x - cloudera cloudera      0 2019-07-31 02:04 Daily
drwxr-xr-x - cloudera cloudera      0 2019-07-29 23:36 Datasets
drwxr-xr-x - cloudera cloudera      0 2019-07-30 02:46 Deliveries
drwxr-xr-x - cloudera cloudera      0 2019-07-30 02:49 Delivery
drwxr-xr-x - cloudera cloudera      0 2019-07-31 02:05 Dividend
drwxr-xr-x - cloudera cloudera      0 2019-07-30 01:48 Emp
drwxr-xr-x - cloudera cloudera      0 2019-08-06 22:11 Employees
drwxr-xr-x - cloudera cloudera      0 2019-07-30 00:17 IPL
drwxr-xr-x - cloudera cloudera      0 2019-08-07 23:20 Movies
drwxr-xr-x - cloudera cloudera      0 2019-07-30 21:17 UserRecords
drwxr-xr-x - cloudera cloudera      0 2019-07-30 22:04 Users
drwxr-xr-x - cloudera cloudera      0 2019-08-06 23:59 WordsDir
drwxr-xr-x - cloudera cloudera      0 2019-07-31 23:50 _sqoop
-rw-r--r-- 1 cloudera cloudera 6488666 2019-08-06 23:46 bigfile.txt
drwxr-xr-x - cloudera cloudera      0 2019-07-31 23:35 ooziedata
drwxr-xr-x - cloudera cloudera      0 2019-07-31 23:44 ooziedb
drwxr-xr-x - cloudera cloudera      0 2019-08-01 21:13 orders
drwxr-xr-x - cloudera cloudera      0 2019-07-31 22:24 retaildata
drwxr-xr-x - cloudera cloudera      0 2019-07-31 22:27 retaildatadb
drwxr-xr-x - cloudera cloudera      0 2019-07-31 23:10 retailtables
[1]+  Killed                  gedit code.txt
[cloudera@quickstart ~]$
```

The terminal window is part of a VMware Workstation 15 Player. The top bar shows 'cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)'. The bottom bar shows the system tray with the date 'Thu Aug 8, 9:38 PM' and the username 'cloudera'. The taskbar at the bottom shows the 'cloudera@quickstart:~' window, a terminal window, and a file explorer window titled '*code.txt (~) - gedit'. The system tray also shows the time '9:38 PM' and the date '8/8/2019'.

//Inside Hive terminal show all the databases present

```
show databases;
```

//Create a database Aadhar if not exists in hive

```
Create database if not exists Aadhar;
```

```
//Use the database
```

```
Use Aadhar;
```

```
//1. Creating an external table for the data in the location where the dataset is present
```

```
create external table aadharData(registrar string, enrollmentAgency string,state string,district
string,subDistrict string,pincode string,gender string,age int,aadharGenerated int,enrollmentRejected
int,residentsProvidingEmail int,residentsProvidingMobileNumber int)row format delimited fields
terminated by ',' location '/user/cloudera/Aadhar'
TBLPROPERTIES('serialization.null.format','skip.header.line.count'='1');
```

```
//1. Creating a managed table in the database
```

```
create table aadharDataManaged(registrar string, enrollmentAgency string,state string,district
string,subDistrict string,pincode string,gender string,age int,aadharGenerated int,enrollmentRejected
int,residentsProvidingEmail int,residentsProvidingMobileNumber int)row format delimited fields
terminated by ',' TBLPROPERTIES('serialization.null.format','skip.header.line.count'='1');
```

```
//Loading data into the managed table
```

```
load data inpath '/user/cloudera/Aadhar/aadhar.csv' into table aadharDataManaged;
```

```
//List of tables created in the database
```

```
Show tables;
```

SCREENSHOT:-

```
hive> show tables;
OK
aadhardata
aadhardatamanaged
Time taken: 0.012 seconds, Fetched: 2 row(s)
hive> █
```

```
//Top 25 rows of the table aadharDataManaged
```

```
Select * from aadharDataManaged limit 25;
```

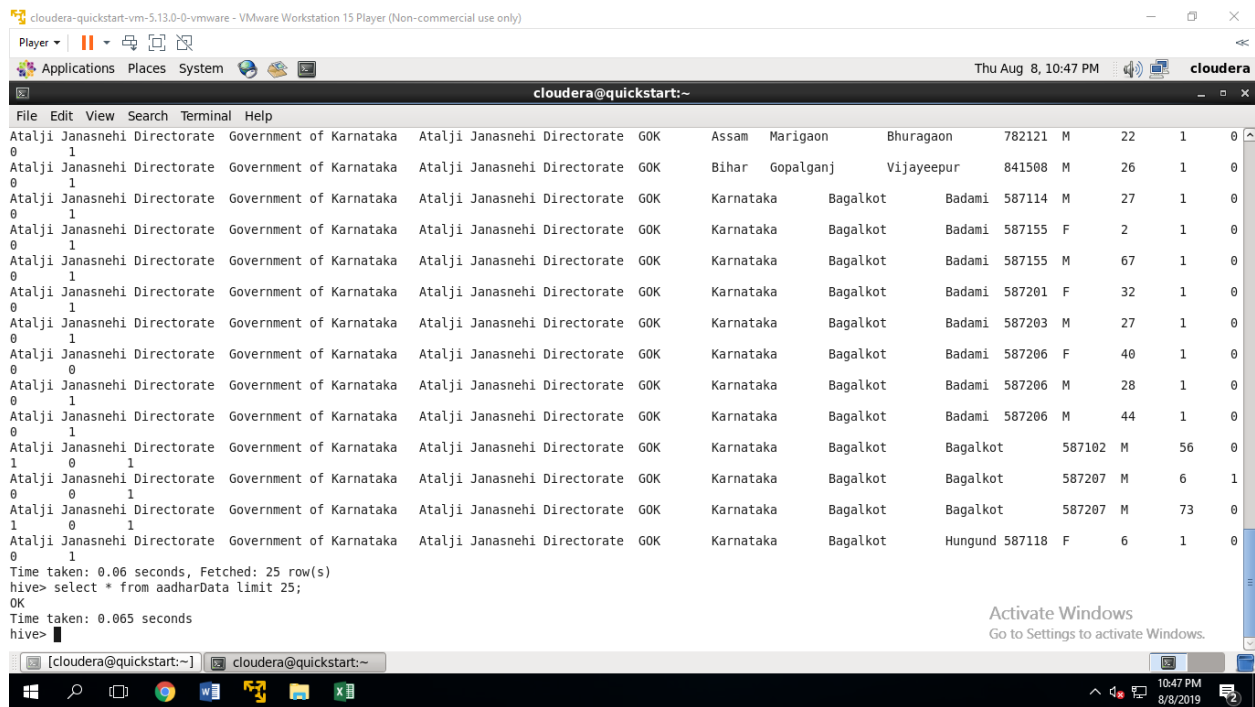
```

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:~
File Edit View Search Terminal Help
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221002 221002 M 10 0 1 0 1
Allahabad Bank Transmoovers India Uttar Pradesh Varanasi Varanasi 221002 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 Bhuragaon 782121 M 22 1 0
0 1
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Bihar Gopalganj Vijayepur 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 1
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 Directorate GOK Karnataka Bagalkot Badami 587206 M 44 1 0
0 1
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Bagalkot 587102 M 56 0
1 0 1
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Bagalkot 587207 M 6 1
0 0 1
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Bagalkot 587207 M 73 0
1 0 1
Atalji Janasnehi Directorate Government of Karnataka Atalji Janasnehi Directorate GOK Karnataka Bagalkot Hungund 587118 F 6 1 0
0 1
Time taken: 0.06 seconds, Fetched: 25 row(s)
hive>

```

```
Select * from aadharData limit 25;
```

SCREENSHOT:-



```
//1. Create Spark DataFrame
```

```
//Loading csv from hdfs as RDD
```

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

```
//Get headers from first row
```

```
val header=aadharRDD.first();
```

```
//Construct Final RDD without headers
```

```
val aadharFinalRDD=aadharRDD.filter(row=>row!=header);
```

```
//Create a dataframe
```

```
val aadharDF = aadharFinalRDD.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","distri
ct","subDistrict","pinCode","gender","age","aadharGenerated","enrolmentRejected","residentsPr
ovidingEmail","residentsProvidingMobileNumber");
```

```
//1. Show top 25 rows of the DataDrame
```

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

SCREENSHOT:-

The screenshot shows a Cloudera Quickstart VM window with a terminal running a Scala command. The command is `scala>` . The output is a table with 10 columns: `Atalji Janasnehi ...`, `Atalji Janasnehi ...`, `Bihar`, `Gopalganj`, `Vijayeeepur`, `841508`, `M`, `26`, `1`, `0`, `0`. The table contains 25 rows of data. The first row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Bihar | Gopalganj | Vijayeeepur | 841508 | M | 26 | 1 | 0 | 0`. The second row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587114 | M | 27 | 1 | 0 | 0`. The third row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587155 | F | 2 | 1 | 0 | 0`. The fourth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587155 | M | 67 | 1 | 0 | 0`. The fifth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587201 | F | 32 | 1 | 0 | 0`. The sixth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587203 | M | 27 | 1 | 0 | 0`. The seventh row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | F | 40 | 1 | 0 | 0`. The eighth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | M | 28 | 1 | 0 | 0`. The ninth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | M | 44 | 1 | 0 | 0`. The tenth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587102 | M | 56 | 0 | 1 | 0`. The eleventh row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587207 | M | 6 | 1 | 0 | 0`. The twelfth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587207 | M | 73 | 0 | 1 | 0`. The thirteenth row is `Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Hungund | 587118 | F | 6 | 1 | 0 | 0`. The output is truncated with `only showing top 25 rows`. The terminal window has a title bar `cloudera@quickstart:~` and a menu bar `File Edit View Search Terminal Help`. The taskbar at the bottom shows the Cloudera logo and the date `11:54 PM 8/8/2019`.

```
cloudera@quickstart:~$ scala> 
```

| Atalji Janasnehi ... | Atalji Janasnehi ... | Bihar | Gopalganj | Vijayeeepur | 841508 | M | 26 | 1 | 0 | 0 |
|----------------------|----------------------|-----------|-----------|-------------|--------|---|----|---|---|---|
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587114 | M | 27 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587155 | F | 2 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587155 | M | 67 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587201 | F | 32 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587203 | M | 27 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | F | 40 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | M | 28 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Badami | 587206 | M | 44 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587102 | M | 56 | 0 | 1 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587207 | M | 6 | 1 | 0 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Bagalkot | 587207 | M | 73 | 0 | 1 | 0 |
| Atalji Janasnehi ... | Atalji Janasnehi ... | Karnataka | Bagalkot | Hungund | 587118 | F | 6 | 1 | 0 | 0 |

only showing top 25 rows

scala>

cloudera

cloudera@quickstart:~

cloudera@quickstart:~

cloudera

cloudera@quickstart:~

11:54 PM 8/8/2019

CHECKPOINT 2

//2. Describe the external table aadharData

describe formatted aadharData;

OUTPUT:-

| # | col_name | data_type | comment |
|---|----------|-----------|---------|
|---|----------|-----------|---------|

| | | | |
|---|-----------|--------|--|
| 1 | registrar | string | |
|---|-----------|--------|--|

| | | | |
|---|------------------|--------|--|
| 2 | enrollmentagency | string | |
|---|------------------|--------|--|

| | | | |
|---|-------|--------|--|
| 3 | state | string | |
|---|-------|--------|--|

| | | | |
|---|----------|--------|--|
| 4 | district | string | |
|---|----------|--------|--|

| | | | |
|---|-------------|--------|--|
| 5 | subdistrict | string | |
|---|-------------|--------|--|

| | |
|--------------------------------|--------|
| pincode | string |
| gender | string |
| age | int |
| aadhargenerated | int |
| enrollmentrejected | int |
| residentsprovidingemail | int |
| residentsprovidingmobilenumber | int |

Detailed Table Information

| | |
|-----------------|--|
| Database: | aadhar |
| Owner: | cloudera |
| CreateTime: | Thu Aug 08 22:04:16 PDT 2019 |
| LastAccessTime: | UNKNOWN |
| Protect Mode: | None |
| Retention: | 0 |
| Location: | hdfs://quickstart.cloudera:8020/user/cloudera/Aadhar |
| Table Type: | EXTERNAL_TABLE |

Table Parameters:

| | |
|---------------------------|------------|
| COLUMN_STATS_ACCURATE | false |
| EXTERNAL | TRUE |
| numFiles | 1 |
| numRows | -1 |
| rawDataSize | -1 |
| serialization.null.format | |
| skip.header.line.count | 1 |
| totalSize | 46483335 |
| transient_lastDdlTime | 1565327056 |

Storage Information

SerDe Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat: org.apache.hadoop.mapred.TextInputFormat
OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed: No
Num Buckets: -1
Bucket Columns: []
Sort Columns: []
Storage Desc Params:
field.delim ,
serialization.format ,

//2. Describe the managed table aadharDataManaged
describe formatted aadharDataManaged;

OUTPUT:-

| # | col_name | data_type | comment |
|---|-------------------------|-----------|---------|
| | registrar | string | |
| | enrollmentagency | string | |
| | state | string | |
| | district | string | |
| | subdistrict | string | |
| | pincode | string | |
| | gender | string | |
| | age | int | |
| | aadhargenerated | int | |
| | enrollmentrejected | int | |
| | residentsprovidingemail | int | |

residentsprovidingmobilenumber int

Detailed Table Information

Database: aadhar

Owner: cloudera

CreateTime: Thu Aug 08 22:14:57 PDT 2019

LastAccessTime: UNKNOWN

Protect Mode: None

Retention: 0

Location: hdfs://quickstart.cloudera:8020/user/hive/warehouse/aadhar.db/aadhardatamanaged

Table Type: MANAGED_TABLE

Table Parameters:

 COLUMN_STATS_ACCURATE true

 numFiles 1

 serialization.null.format

 skip.header.line.count 1

 totalSize 46483335

 transient_lastDdlTime 1565327780

Storage Information

SerDe Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

InputFormat: org.apache.hadoop.mapred.TextInputFormat

OutputFormat: org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat

Compressed: No

Num Buckets: -1

Bucket Columns: []

Sort Columns: []

Storage Desc Params:

 field.delim ,

serialization.format ,

//2. Describe the Data Frame Schema

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)
```

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

select registrar,count(*) from aadharData group by registrar;

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 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 |

//4. count the number of states

```
select count(distinct(state)) as No_of_States from aadharData;
```

OUTPUT:-

37

//4. count number of district in each state

```
select state,count(distinct(district)) from aadharData 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 | |

//4. count number of sub-district in each district

select district,count(distinct(subDistrict)) from aadharData group by district limit 25;

OUTPUT:-

| | |
|----------------|----|
| Adilabad | 41 |
| Agra | 6 |
| Ahmadnagar | 14 |
| Ahmedabad | 9 |
| Aizawl | 5 |
| Ajmer | 8 |
| Akola | 7 |
| Alappuzha | 6 |
| Aligarh | 5 |
| Alirajpur | 2 |
| Allahabad | 8 |
| Almora | 7 |
| Alwar | 12 |
| Ambala | 3 |
| Ambedkar Nagar | 4 |
| Amethi | 4 |
| Amravati | 14 |
| Amreli | 11 |
| Amritsar | 4 |
| Amroha | 3 |
| Anand | 8 |
| Ananthapuramu | 61 |
| Anantnag | 4 |
| Angul | 19 |
| Anjaw | 2 |

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

select distinct(state),enrollmentAgency from aadharData;

OUTPUT:-

| | |
|---------------|---|
| Uttar Pradesh | Yashi Informatics LLP |
| Uttar Pradesh | Yuvaan Infotech |
| Uttar Pradesh | Zephyr System Pvt.Ltd. |
| Uttarakhand | A I Soc for Electronics and Comp Tech |
| Uttarakhand | A-Onerealtors Pvt Ltd |
| Uttarakhand | AKSH OPTIFIBRE LIMITED |
| Uttarakhand | AVVAS INFOTECH PVT LTD |
| Uttarakhand | Abha Systems And Consultancy |
| Uttarakhand | Abhipra Capital Ltd |
| Uttarakhand | Akshaya |
| Uttarakhand | Alankit Limited |
| Uttarakhand | Amar Constructions |
| Uttarakhand | BASIX |
| Uttarakhand | Binary Systems |
| Uttarakhand | CALANCE SOFTWARE PRIVATE LTD |
| Uttarakhand | CHIPS |
| Uttarakhand | CMS Computers Ltd |
| Uttarakhand | COMTECHINFO SOLUTIONS PVT.LTD |
| Uttarakhand | CSC SPV |
| Uttarakhand | CSC e-Governance Services India Limited |
| Uttarakhand | Care Educational & Welfare Society |
| Uttarakhand | Conatus Infocom Pvt. Ltd |
| Uttarakhand | DATASOFT COMPUTER SERVICES(P) |
| Uttarakhand | Department of IT Govt. of HP |
| Uttarakhand | Digitcom Systems Pvt. Ltd. |
| Uttarakhand | District E-Seva Society Gandhinagar |

| | |
|-------------|--|
| Uttarakhand | District E-Seva Society Navsari |
| Uttarakhand | District IT Society Gurgaon |
| Uttarakhand | District IT Society Hisar |
| Uttarakhand | District IT Society Jhajjar |
| Uttarakhand | District IT Society Karnal |
| Uttarakhand | District IT Society Rewari |
| Uttarakhand | District IT Society Yamuna Nagar |
| Uttarakhand | District Magistrate & Collector West Tripura District |
| Uttarakhand | EDCS GOK |
| Uttarakhand | Electronics Corporation of Tamil Nadu Limited |
| Uttarakhand | FINANCIAL INFORMATION NETWORK |
| Uttarakhand | Home Life Buildcon Pvt Ltd |
| Uttarakhand | IAP COMPANY Pvt. Ltd |
| Uttarakhand | Indotech Engineering Products |
| Uttarakhand | KRISHNAURAM SHIKSHA EVAM JAN KALYAN SAMITI |
| Uttarakhand | Karvy Data Management Services |
| Uttarakhand | Late Smt. Nirmala Singh Seva Samiti |
| Uttarakhand | M/s Gold Square Builders & Promoters Pvt. Ltd. |
| Uttarakhand | M/s. Goa Electronics Ltd |
| Uttarakhand | MEGHA VINCOM PVT LTD |
| Uttarakhand | MPOnline Limited |
| Uttarakhand | Mahaonline Limited |
| Uttarakhand | Make India Smart Private Limited |
| Uttarakhand | Matrix Processing House |
| Uttarakhand | N.K. Sharma Enterprises Ltd. |
| Uttarakhand | NPS Technologies Pvt. Ltd |
| Uttarakhand | National Cooperative Consumers Federation of India Limited |
| Uttarakhand | Nekton IT India Pvt Ltd. |
| Uttarakhand | Offshoot Agency Pvt. Ltd. |

| | |
|-------------|--|
| Uttarakhand | Ojus G Enterprises |
| Uttarakhand | Ojus Healthcare Private Limited |
| Uttarakhand | Omnitech Infosolutions Ltd |
| Uttarakhand | Osiris Infotech Pvt. Ltd. |
| Uttarakhand | Prakash Computer Services |
| Uttarakhand | Promind Solutions P Limited |
| Uttarakhand | RBS multisolutions private limited |
| Uttarakhand | RELIGARE SECURITIES LTD |
| Uttarakhand | Radiant Haroti Industries India Ltd |
| Uttarakhand | Raj Construction Co. |
| Uttarakhand | Rajcomp Info Services Ltd |
| Uttarakhand | SGS INDIA PVT LTD |
| Uttarakhand | SREI INFRASTRUCTURE FINANCES L |
| Uttarakhand | SRM Education And Social Welfare Society |
| Uttarakhand | SRR Infotech |
| Uttarakhand | SVG Express Services Pvt Ltd |
| Uttarakhand | Silver Touch Technologies Ltd |
| Uttarakhand | SoftAge Information Technology Limited |
| Uttarakhand | Sri Ramraja Sarkar Lok Kalyan Trust |
| Uttarakhand | State Health Society |
| Uttarakhand | Steel City Securities Limited |
| Uttarakhand | Synapses Solutions Private Limited |
| Uttarakhand | Twinstar Industries Ltd. |
| Uttarakhand | UIDAI-EA |
| Uttarakhand | Utility Forms Pvt Ltd |
| Uttarakhand | Vakrangee Softwares Limited |
| Uttarakhand | Virinchi Technologies Ltd |
| Uttarakhand | Wipro Ltd |
| Uttarakhand | Yash Ornaments Pvt. Ltd |

Uttarakhand Zephyr System Pvt.Ltd.
West Bengal A I Soc for Electronics and Comp Tech
West Bengal A-Onerealtors Pvt Ltd

CHECKPOINT 3

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

```
select state,sum(aadharGenerated) as cnt from aadharData group by state order by cnt desc limit 3;
```

OUTPUT:-

Bihar 162607
West Bengal 119901
Uttar Pradesh 103767

//9. Find top 3 private agencies generating the most number of Aadhaar cards

```
select enrollmentAgency,sum(aadharGenerated) as cnt from aadharData group by enrollmentAgency  
order by cnt desc limit 3;
```

OUTPUT:-

CSC SPV 173192
Wipro Ltd 39619
SREI INFRASTRUCTURE FINANCES L 26497

//10. Find the number of residents providing email, mobile number

```
select count(*) from aadharData where residentsProvidingEmail<>0 and  
residentsProvidingmobilenumber<>0;
```

OUTPUT:-

16951

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


```
select district,count(*) as cnt from aadharData where enrollmentRejected=0 group by district order by cnt 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(aadharGenerated) from aadharData 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 |

CHECKPOINT 4

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

//Create Spark DataFrame

//Loading csv from hdfs as RDD

```

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

//Get headers from first row

val header=aadharRDD.first();

//Construct Final RDD without headers

val aadharFinalRDD=aadharRDD.filter(row=>row!=header);

//Create a dataframe

val aadharDF = aadharFinalRDD.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.show(25)

```

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

```
select corr(age,residentsprovidingmobilenumber) from aadharData;
```

OUTPUT:-

-0.11754461896889339

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

```
select distinct(pincode) from aadharData;
```

OUTPUT:-

854337

854338

854339

854340

855101

855102

855105

855106

855107

855108

855113

855114

855115

855116

855117

855456

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

```
select state,sum(enrollmentRejected) from aadharData where state in('Uttar Pradesh','Maharashtra') group by state;
```

Maharashtra 1818

Uttar Pradesh 5286

CHECKPOINT 5

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

```
select state,sum(aadharGenerated)*100/(sum(aadharGenerated+enrollmentRejected)) as cnt from aadharData where gender='M' group by state order by cnt desc limit 3;
```

OUTPUT:-

Andaman and Nicobar Islands 100.0

Others 100.0

Lakshadweep 100.0

//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(enrollmentRejected)*100/(sum(aadharGenerated+enrollmentRejected)) as cnt from aadharData 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 Andaman 50.0

North And Middle Andaman 33.333333333333336

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

```
select state,sum(aadharGenerated)*100/(sum(aadharGenerated+enrollmentRejected)) as cnt
from aadharData where gender='F' group by state order by cnt desc limit 3;
```

OUTPUT:-

Dadra and Nagar Haveli 100.0

Sikkim 100.0

Others 100.0

//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(enrollmentRejected)*100/(sum(aadharGenerated+enrollmentRejected)) as cnt
from aadharData 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

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

```
select sum(aadharGenerated)*100/(sum(aadharGenerated+enrollmentRejected)) from  
aadharBucket;
```

OUTPUT:-

94.81864032289477