Problem Statement

You are working for the data analysis team and wish to analyse the data in hand for various demographic parameters. The analysis at hand involves basic data preparation, processing and understanding. Further, you also wish to forecast the effects of certain information on the overall Aadhaar number generation. The metadata/dictionary is provided below:

Metadata/Data Dictionary

Sr.No.	Name of the field	Description
1	date	This is the registration date.
2	registrar	This is the name of the registrar office, generally, this is a government approval body governing the process.
3	private_agency	This is the name of the private agency working for registration of Aadhaar cards in a particular area/region.
4	state	This is the name of the state/union territory.
5	district	This is the name of the district.
6	sub_district	This is the name of the sub-districts/major cities in a particular district.
7	pincode	This is the postal code of an area.
8	gender	This is the gender of the group*.
9	age	This is the age of the group*.
10	aadhaar_generated	This is the total number of Aadhaar cards generated on a particular day
11	rejected	This is the total number of enrolments rejected on a particular day
12	mobile_number	This is the count of residents who have provided the mobile number at the time of enrolment
13	email_id	This is the count of residents who have provided the email id at the time of enrolment

(*: explained in the example below).

Note: The dataset does not contain the headers. You should use the header names in the order as mentioned above.

You can understand the data dictionary better by the following example: A row with data - 20150420, Allahabad Bank, A-Onerealtors Pvt Ltd, Uttar Pradesh, Ambedkar Nagar, Akbarpur, 224155, F, 15, 5, 0, 0, 4 indicates that

- On 20 Apr 2014 (date), for A-Onerealtors Pvt Ltd (private_agency) registered with Allahabad Bank (registrar) at PIN code 224155, Akbarpur (sub_district), Ambedkar Nagar (district), Uttar Pradesh (state)
- Among the group of women aged 15
- There were 5 Aadhar numbers generated and 0 were rejected
- Out of the 5 that applied, none had an email ID and 4 had mobile numbers

Checkpoints

Checkpoint 1

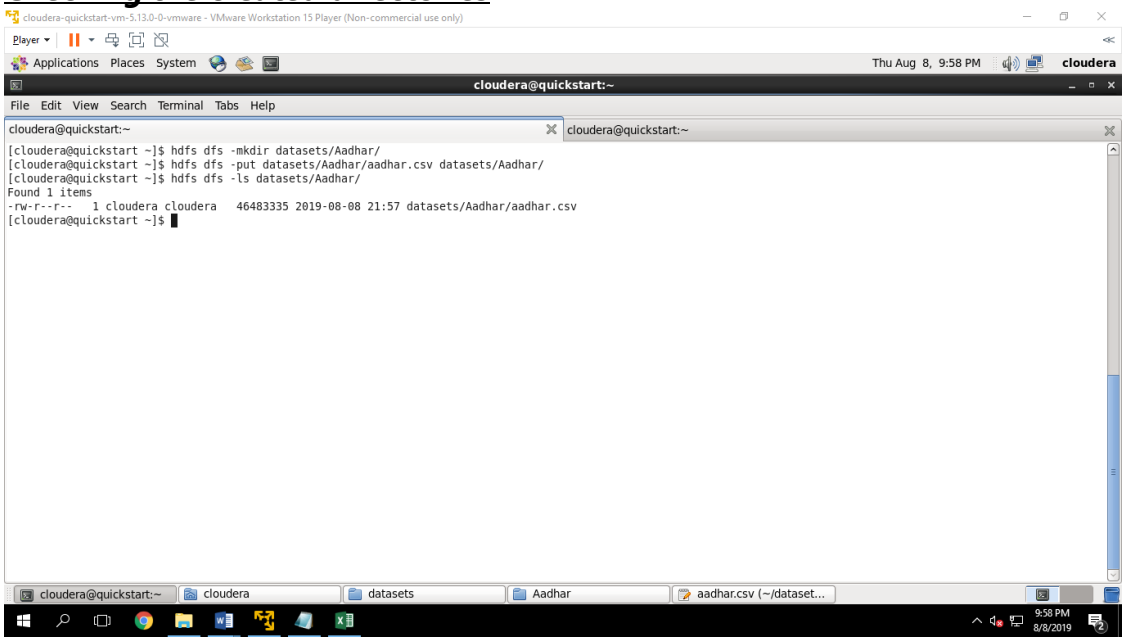
Load the data into HDFS, Hive Managed table, Hive External table and Spark DataFrame.

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

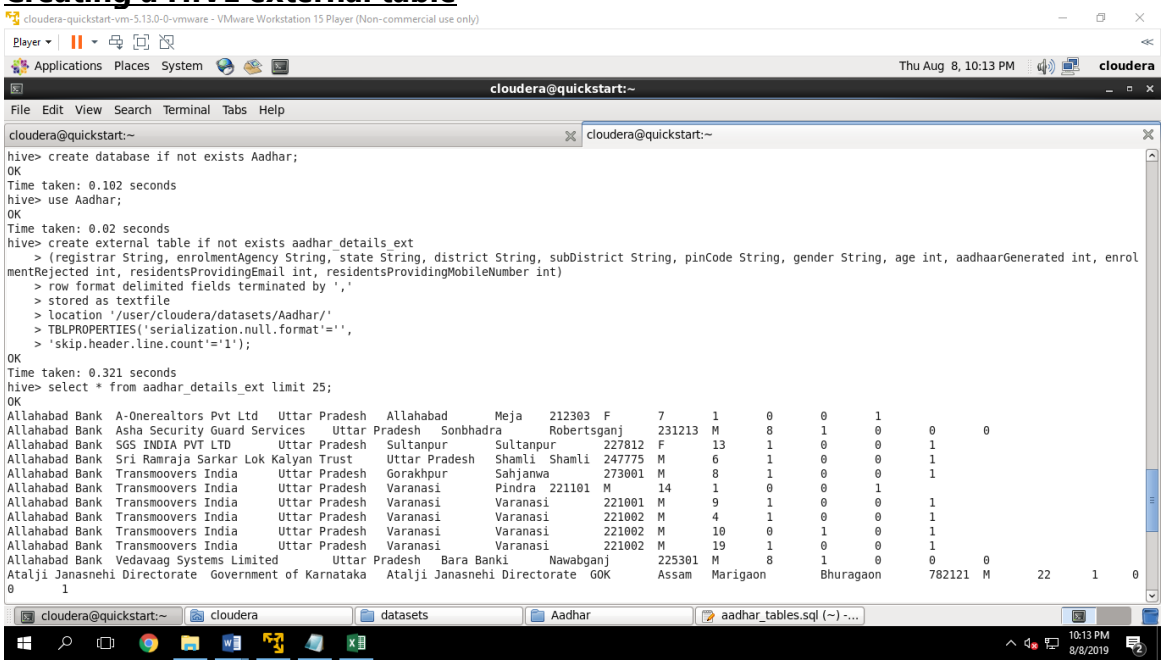
Loading the data ino HDFS

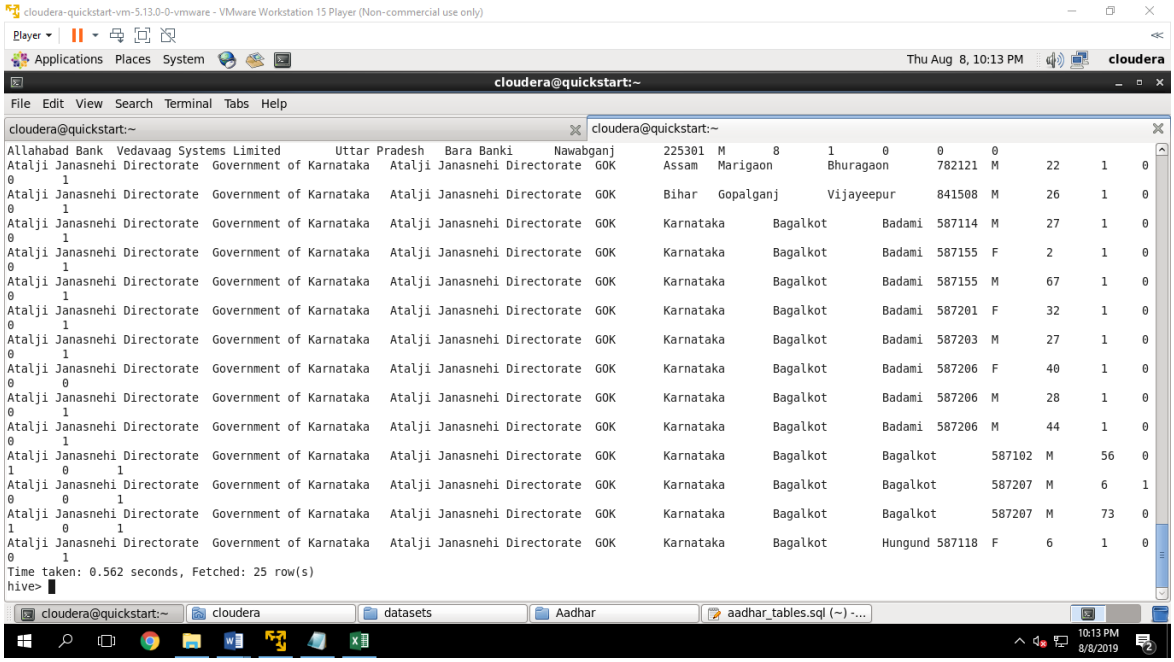
$ hdfs dfs -put datasets/Aadhar/aadhar.csv datasets/Aadhar

Checking the created directories

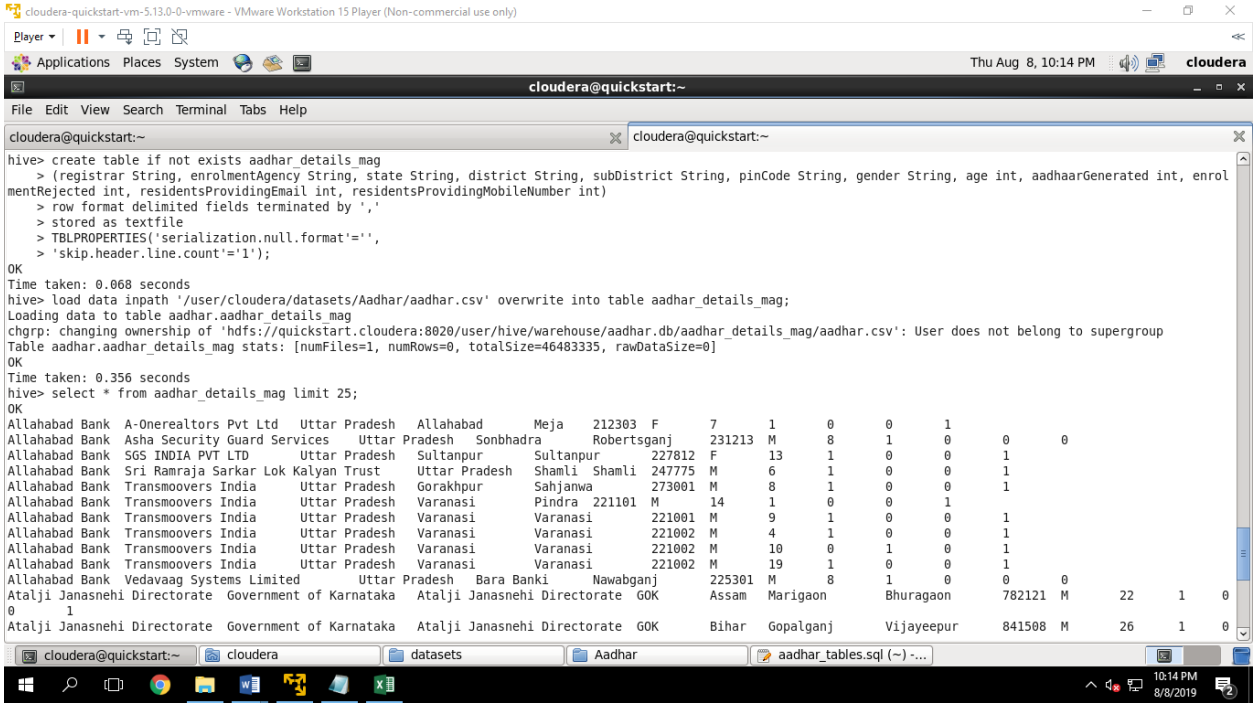


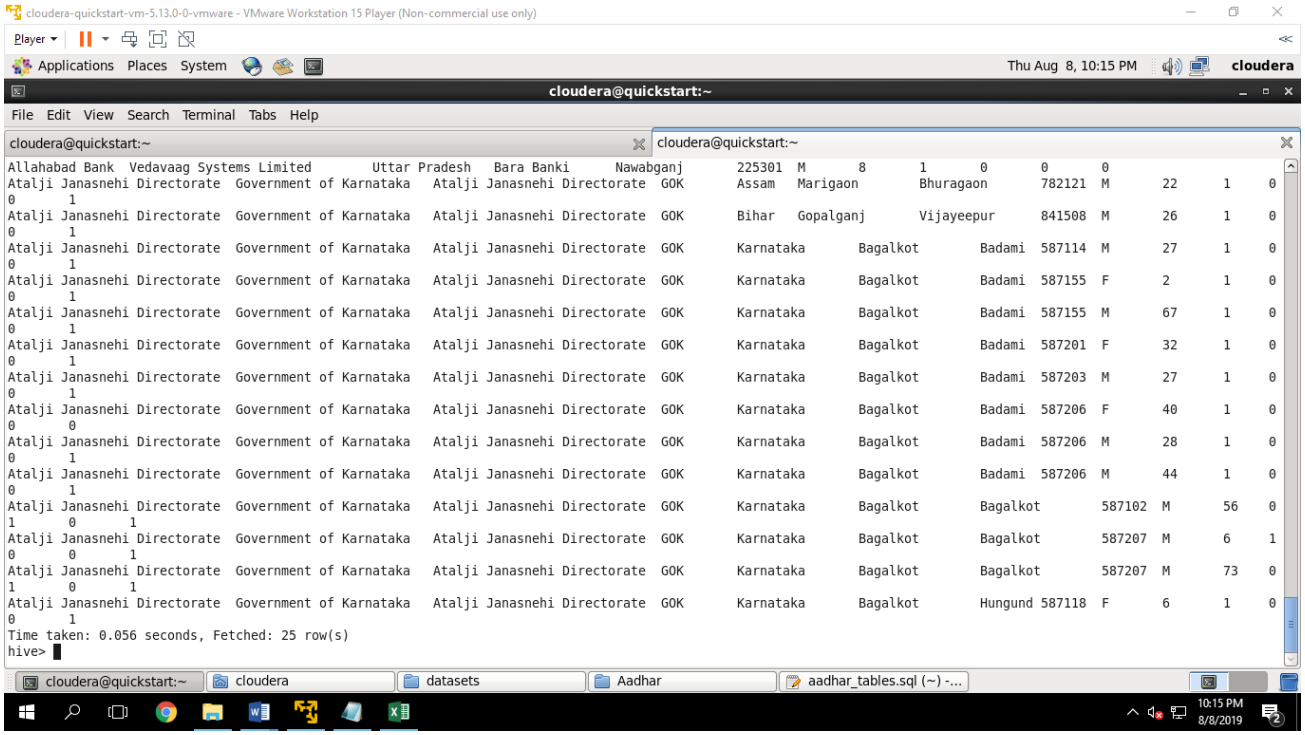
Creating a HIVE external table





Creating HIVE internal table





Creating Spark Dataframe

1. Loading CSV from HDFS as RDD

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

2. Get headers from the first row

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

3. Construct Final RDD without headers

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

4. Create 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","re
sidentsProvidingEmail","residentsProvidingMobileNumber");
```

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)

Applications Places System Thu Aug 8, 11:24 PM cloudera

cloudera@quickstart:~

File Edit View Search Terminal Tabs Help

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

cloudera

cloudera@quickstart:~ | cloudera | datasets | Aadhar | aadhar_tables.sql (~) -...

11:24 PM 8/8/2019

cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation 15 Player (Non-commercial use only)

Applications Places System Thu Aug 8, 11:25 PM cloudera

cloudera@quickstart:~

File Edit View Search Terminal Tabs Help

```
|Atalji Janasnehi ...|Atalji Janasnehi ...|        Assam|  Marigaon|  Bhuragaon| 782121|     M| 22|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|        Bihar| Gopalganj| Vijayeepur| 841508|     M| 26|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587114|     M| 27|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587155|     F|  2|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587155|     M| 67|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587201|     F| 32|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587203|     M| 27|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587206|     F| 40|              1|                0|                      0|
                   0|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587206|     M| 28|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|     Badami| 587206|     M| 44|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|   Bagalkot| 587102|     M| 56|              0|                1|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|   Bagalkot| 587207|     M|  6|              1|                0|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|   Bagalkot| 587207|     M| 73|              0|                1|                      0|
                   1|
|Atalji Janasnehi ...|Atalji Janasnehi ...|    Karnataka|  Bagalkot|    Hungund| 587118|     F|  6|              1|                0|                      0|
                   1|
+--------------------+--------------------+-------------+----------+-----------+-------+------+---+---------------+-----------------+-----------------------+-----------
--------------------+
only showing top 25 rows
```

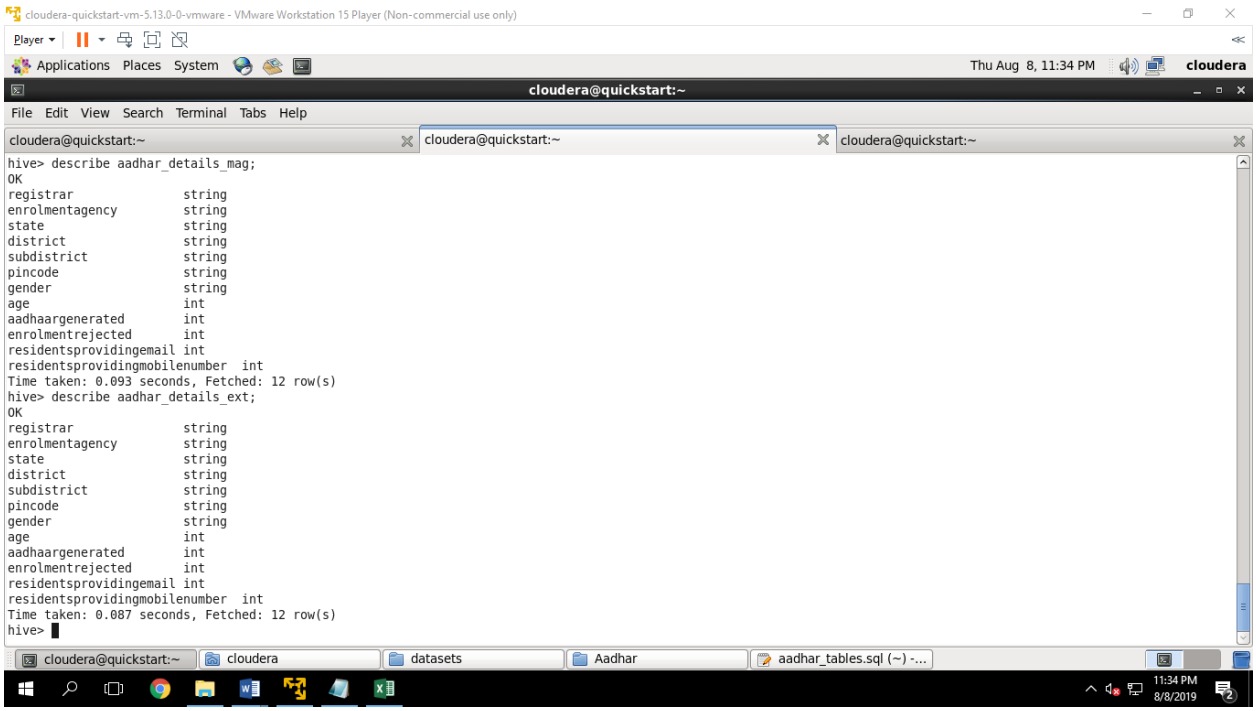
cloudera@quickstart:~ | cloudera | datasets | Aadhar | aadhar_tables.sql (~) -...

11:25 PM 8/8/2019

Checkpoint 2

2. Describe the schema.

Description of managed and external HIVE tables:



Description of Spark Dataframe:

```
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)
```

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

hive> select registrar as Registrars, count(*) as Number from aadhar_details_mag 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
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 Pradesh583
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 Department3958
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 Agency15468
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

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

hive> select count(distinct(state)) from aadhar_details_mag;

OUTPUT:
37

hive> select state, count(distinct(district)) from aadhar_details_mag group by state;

OUTPUT:
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
Tamil Nadu 32
Telangana 10
Tripura 8
Uttar Pradesh 75
Uttarakhand 13
West Bengal 19
```

```
hive> select district, count(distinct(subDistrict)) from aadhar_details_mag group by state;
```

OUTPUT:

```
....
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: 27.241 seconds, Fetched: 664 row(s)
```

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

hive> select distinct(state), enrolmentagency from aadhar_details_mag;

OUTPUT:

....

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: 35.084 seconds, Fetched: 2271 row(s)

Checkpoint 3

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

hive> create table if not exists aadhar_by_state as select state, sum(aadhaarGenerated) as totalAadhars from aadhar_details_mag group by state;

hive> select * from aadhar_by_state order by totalaadhars desc limit 3;

OUTPUT:

Bihar	162607
West Bengal	119901
Uttar Pradesh	103767

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

hive> create table if not exists aadhar_by_agency as select enrolmentAgency, sum(aadhaarGenerated) as totalAadhars from aadhar_details_mag group by state;

hive> select * from aadhar_by_agency order by totalaadhars desc limit 3;

OUTPUT:

CSC SPV 173192

Wipro Ltd 39619

SREI INFRASTUCTURE FINANCES L 26497

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

```
hive> select count(*) from aadhar_details_mag where residentsProvidingEmail <> 0 and residentsProvidingMobileNumber <> 0;
```

OUTPUT:

16951

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

```
hive> select district, count(*) as cnt from aadhar_details_mag where enrolmentRejected = 0 group by district order by cnt desc limit 3;
```

OUTPUT:

Barddhaman 6726

North 24 Parganas 6534

South 24 Parganas 5603

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

```
hive> select state, sum(aadhaarGenerated) from aadhar_details_mag 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: 20.834 seconds, Fetched: 37 row(s)
```

Checkpoint 4

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

```
val aadharRDD = sc.textFile("/user/cloudera/datasets/Aadhar/aadhar.csv");
val header = aadharRDD.first()
val aadharFinalRDD = aadharRDD.filter(row => row!=header);
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.describe();
```

12. **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)**

```
hive> select corr(age, residentsProvidingMobileNumber) from aadhar_details_mag;
```

OUTPUT:
-0.11754461896889339

13. **Find the number of unique pincodes in the data?**

```
hive> select distinct(pinCode) from aadhar_details_mag;
```

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

```
hive> select state, sum(enrolmentRejected) from aadhar_details_mag where state in
('Maharashtra', 'Uttar Pradesh') group by state;
```

OUTPUT:

Maharashtra	1818
Uttar Pradesh	5286

Checkpoint 5

On the given dataset, perform EDA and find:

15. **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_numbers from
aadhar_details_mag where gender = 'M' group by state order by male_numbers desc limit
3;
```

OUTPUT:
Andaman and Nicobar Islands 100.0
Others 100.0
Lakshadweep 100.0

16. 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_details_mag where gender = 'F' and state in ('Andaman
and Nicobar Islands', 'Others', 'Lakshadweep') group by district order by female_rejections
desc limit 3;
```

OUTPUT:
Lakshadweep 100.0
South Andaman 50.0
North And Middle Andaman 33.333333333333336

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

```
hive> select state, (sum(aadhaarGenerated) *
100)/(sum(aadhaarGenerated+enrolmentRejected)) as female_numbers from
aadhar_details_mag where gender = 'F' group by state order by female_numbers desc limit
3;
```

OUTPUT:
Dadra and Nagar Haveli 100.0
Sikkim 100.0
Others 100.0

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

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

OUTPUT:
East Sikkim 9.090909090909092
Dadra and Nagar Haveli 3.4482758620689653
West Sikkim 0.0

19. 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 age_bucket
(registrar String, enrolmentAgency String, state String, district String, subDistrict String,
pinCode 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 '\t'
stored as textfile
TBLPROPERTIES('serialization.null.format'='', 'skip.header.line.count'='1');

set hive.enforce.bucket = true;

insert into age_bucket select * from aadhar_details_mag;

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

OUTPUT:
94.81863336350962