BIG DATA PROJECT

CHECKPOINT 1

```
1)CREATE TABLE IF NOT EXISTS AADHAR_DATA(REGISTRAR STRING,PRIVATE_AGENCY STRING,STATE STRING,DISTRICT STRING,SUBDISTRICT STRING,PINCODE STRING,GENDER STRING,AGE INT,AADHAR_GENERATED INT,REJECTED INT,PROVIDE_EMAIL INT,PROVIDE_MOBILE INT)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
LOAD DATA LOCAL INPATH '/home/cloudera/aadhar.csv' INTO TABLE AADHAR_DATA;
```

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/project2' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE SELECT * FROM AADHAR_DATA LIMIT 25;
```

```
2)CREATE EXTERNAL TABLE IF NOT EXISTS AADHAR_DATA_EXTERNAL(REGISTRAR STRING,PRIVATE_AGENCY STRING,STATE STRING,DISTRICT STRING,SUBDISTRICT STRING,PINCODE STRING,GENDER STRING,AGE INT,AADHAR_GENERATED INT,REJECTED INT,PROVIDE_EMAIL INT,PROVIDE_MOBILE INT)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
LOAD DATA LOCAL INPATH '/home/cloudera/aadhar.csv' INTO TABLE AADHAR_DATA_EXTERNAL;
```

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/project2' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE SELECT * FROM AADHAR_DATA_EXTERNAL LIMIT 25;SCALA
```

3)

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

val first_header=aadharrdd.first()

val final_details=aadharrdd.filter(w=>w!=first_header)

val aadhardet=final_details.map(w=>(w.split(",")(0),w.split(",")(1),w.split(",")(2),w.split(",")(3),w.split(",")(4),w.split(",")(5),w.split(",")(6),w.split(",")(7).toInt,w.split(",")(8).toInt,w.split(",")(9).toInt,w.split(",")(10).toInt,w.split(",")(11).toInt));

valaadharframe=aadhardet.toDF("registerar","private_agency","state","district","sub_district","pincode","gender","age","aadhar_generated","rejected","noemails","nomobile");
```

aadharframe.show(25)

CHECKPOINT2

2) aadharframe.schema

3)

caseclassaadharnew(registerar:String,private_agency:String,state:String,district:String,sub_district:String,pincode:String,gender:String,age:Int,aadhar_generated:Int,rejected:Int,noemails:Int,nomobile:Int);

aadharframe.registerTempTable("aadhar")

sqlContext.sql("select count(distinct(registerar)) from aadhar").show()

4) sqlContext.sql("select state,count(district) from aadhar group by state").show()

sqlContext.sql("select district,count(sub_district) from aadhar group by district").show()

5) sqlContext.sql("select state,count(gender=='M') as Male,count(gender=='F') as Female from aadhar group by state").show()

6)sqlContext.sql("Select state,private_agency from aadhar group by state,private_agency").show

CHECKPOINT 3

8) sqlContext.sql("select state,sum(aadhar_generated) as Number_of_aadhar from aadhar group by state sort by Number_of_aadhar desc Limit 3").show

9) sqlContext.sql("select private_agency,sum(aadhar_generated) as Number_of_aadhar from aadhar group by private_agency sort by Number_of_aadhar desc Limit 3").show

10) sqlContext.sql("select sum(noemails) as emails,sum(nomobile) as mobile from aadhar ").show

11) sqlContext.sql("select sum(aadhar_generated) as number_of_enrolment ,district from aadhar group by district sort by number_of_enrolment desc limit 3 ").show

12) sqlContext.sql("select state,sum(aadhar_generated) as Number_of_aadhar from aadhar group by state").show

CHECKPOINT 4

13) aadharframe.printSchema

14) aadharframe.select(corr('age,'nomobile)).show()

15) sqlContext.sql("select count(distinct(aadhar_generated)) as Number_of_pincodes from aadhar ").show

```
16) sqlContext.sql("select sum(rejected) from aadhar where state like 'Uttar Pradesh' or state like
'Maharashtra' ").show
```

CHECKPOINT 5

17

```
sqlContext.sql(" Selectstate,round((sum(aadhar_generated)/sum(aadhar_generated+rejected))*100,2)
Percentage_of_aadhar from aadhar where gender like 'M' group by state order by
Percentage_of_aadhar desc limit 3" ).show
```

```
18. sqlContext.sql("Select state,district,round((sum(rejected)/sum(aadhar_generated+rejected))*100,2)
Percentage_of_rejected from aadhar where gender like 'F' and state like 'Andaman and Nicobar Islands'
or state like 'Lakshadweep' or state like 'Others' group by state,district order by Percentage_of_rejected
desc" ).show
```

19

```
sqlContext.sql(" Select state,round((sum(aadhar_generated)/sum(aadhar_generated+rejected))*100,2)
Percentage_of_aadhar from aadhar where gender like 'F' group by state order by Percentage_of_aadhar
desc limit 3" ).show
```

```
20. sqlContext.sql("Select state,district,round((sum(rejected)/sum(aadhar_generated+rejected))*100,2)
Percentage_of_rejected from aadhar where gender like 'M' and state like 'Sikkim' or state like 'Others'
or state like 'Dadra and Nagar Haveli' group by state,district order by Percentage_of_rejected desc limit
3" ).show
```

```
21. create table aadhar_bucket(registrar string,private_agency string,state string,district
string,sub_district string,pincode string,gender string, age int,aadhar_generated int,rejected
int,noemails int,nomobiles 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');

Insert into aadhar_bucket select * from aadhar_data;

select round((sum(aadhar_generated)/sum(aadhar_generated+rejected))*100,2) from aadhar_bucket;
```