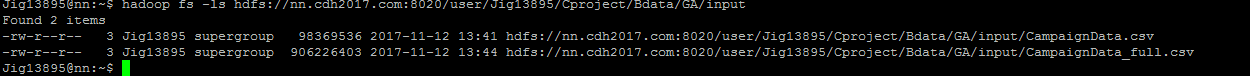
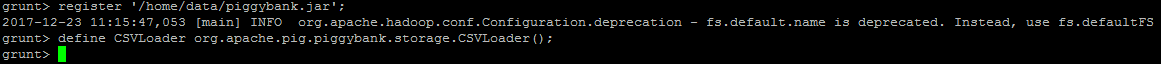
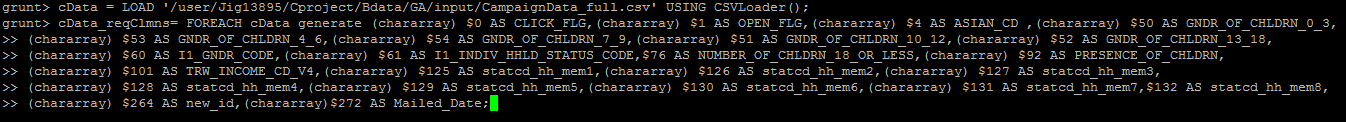
--move the input file into Hdfs path and check for the input file

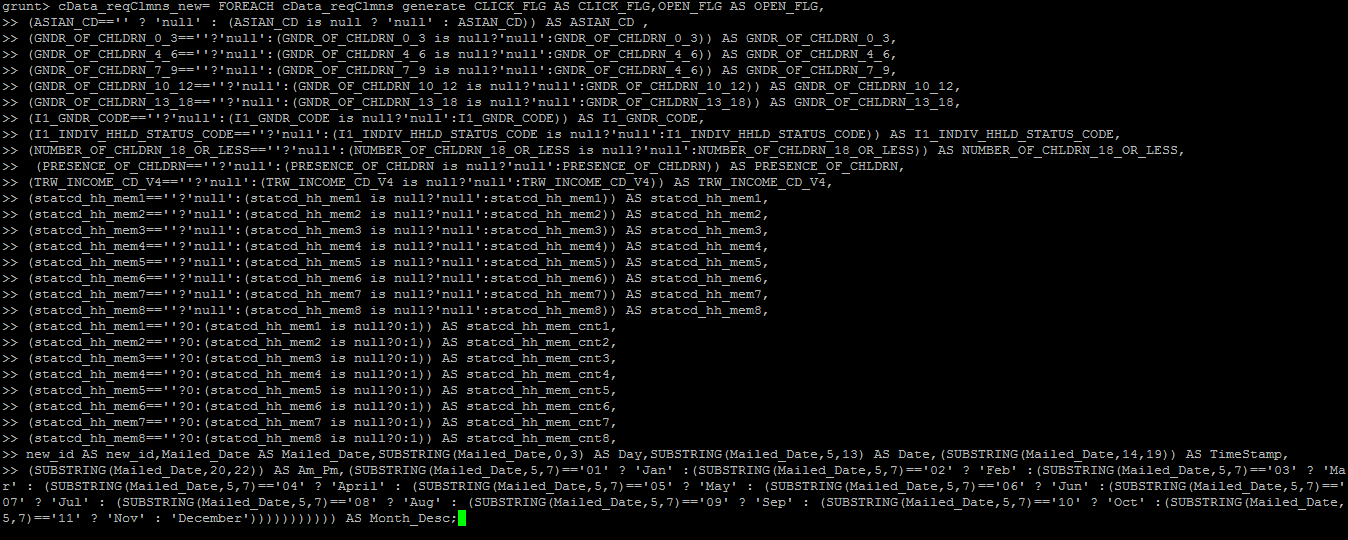


--register and define the jar to load the csv file

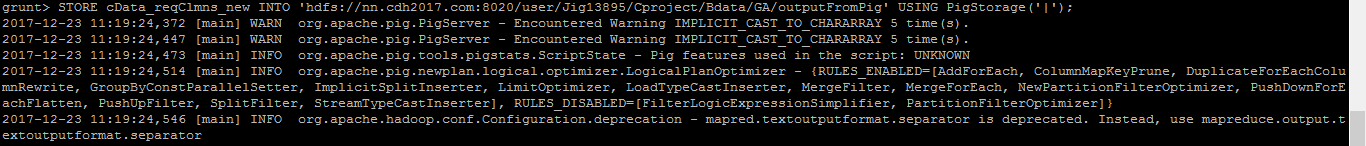


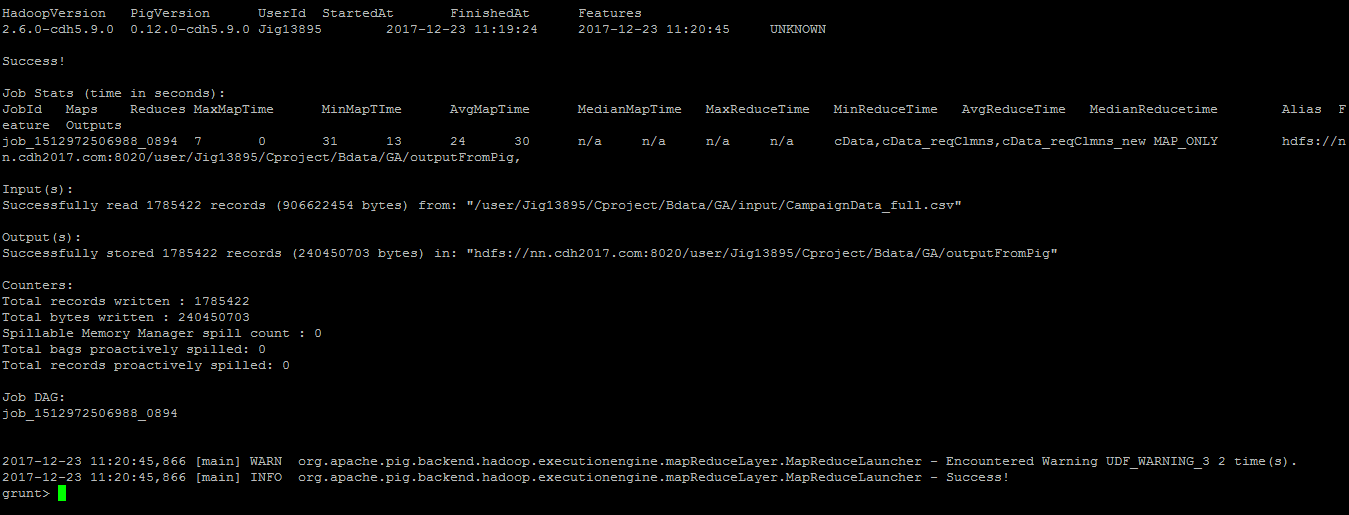
--do the data cleaning and transformation in pig



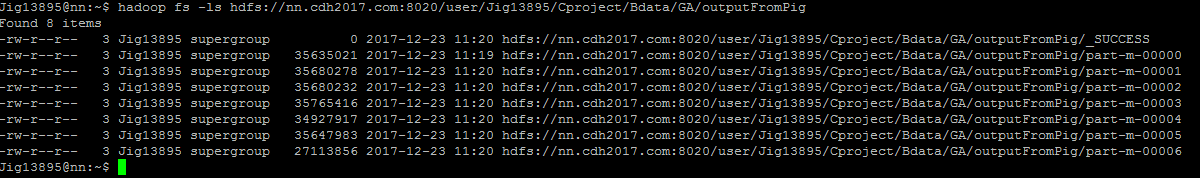


Store the output from pig in hdfs

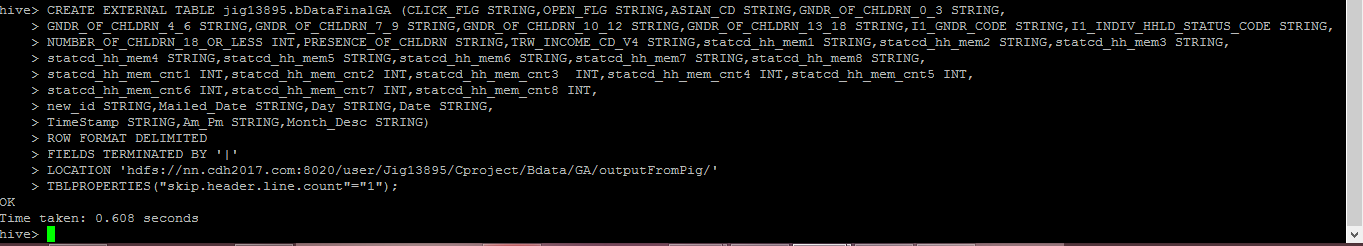




--check for the output file

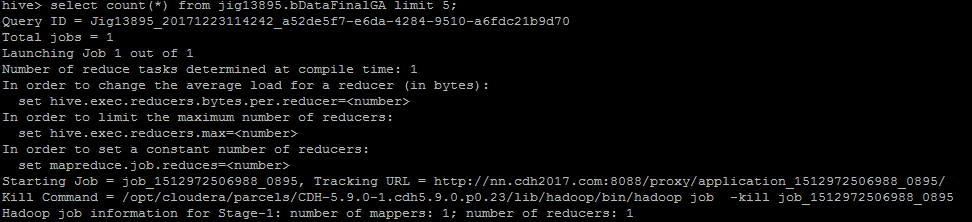


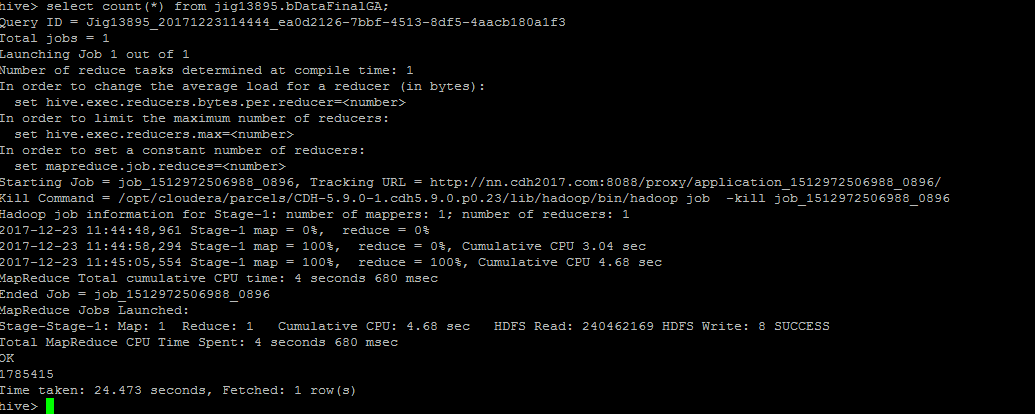
--create table in hive;



--check the data in the table

select count(\*) from jig13895.bDataFinalGA;





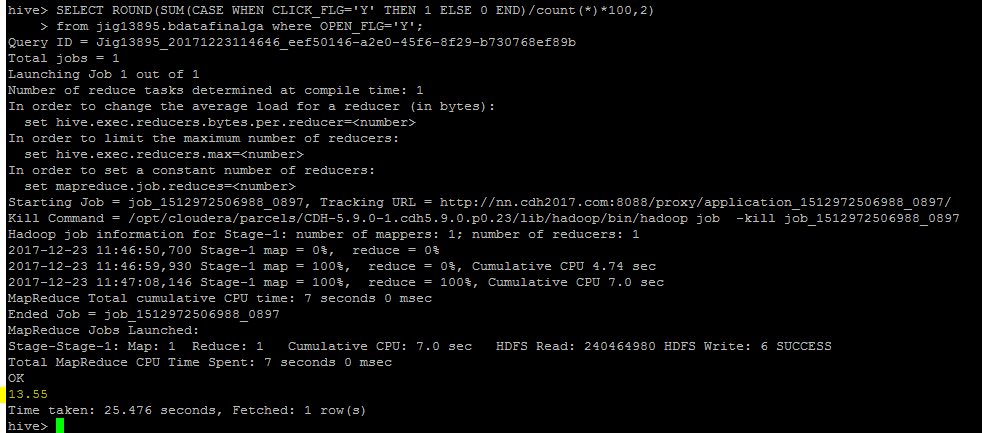
**Questions:**

**1.Find the Click To Open Rate i.e overall CTOR**

SELECT ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

from jig13895.bdatafinalga where OPEN\_FLG='Y';

o/p: 13.55



**--Find CTOR by gender**

SELECT I1\_GNDR\_CODE,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY I1\_GNDR\_CODE ORDER BY I1\_GNDR\_CODE;

o/p:

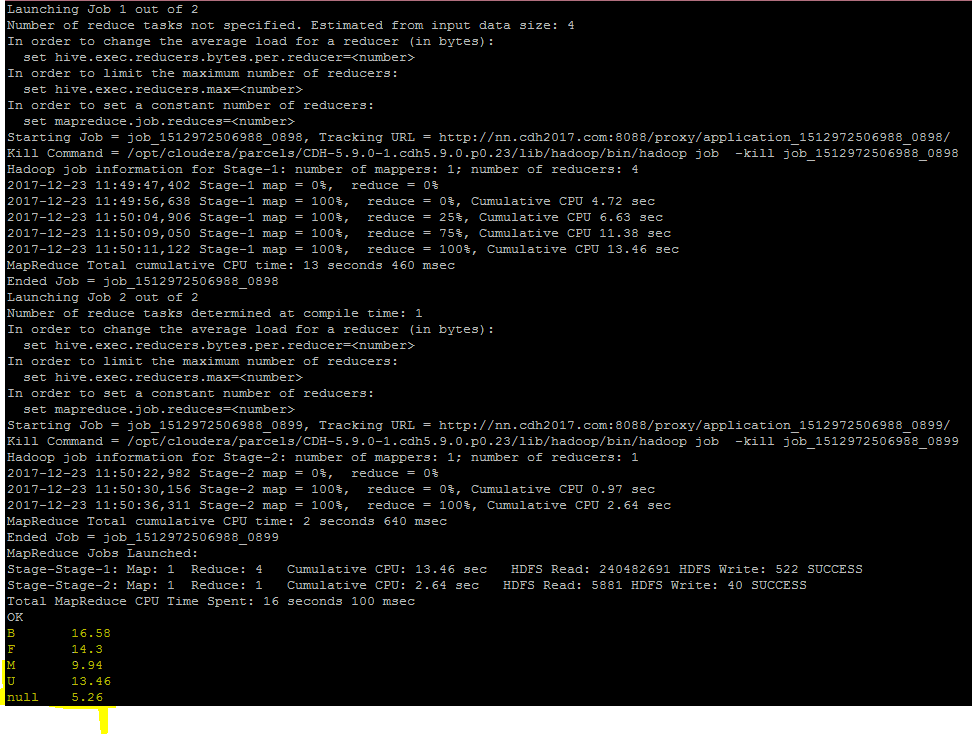
B 16.58

F 14.3

M 9.94

U 13.46

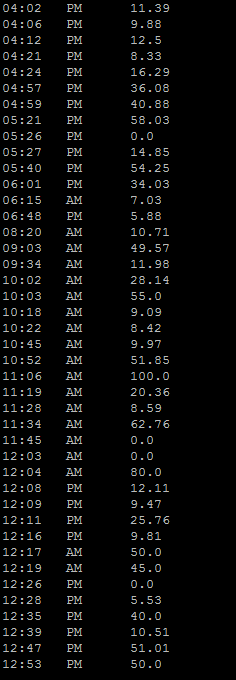
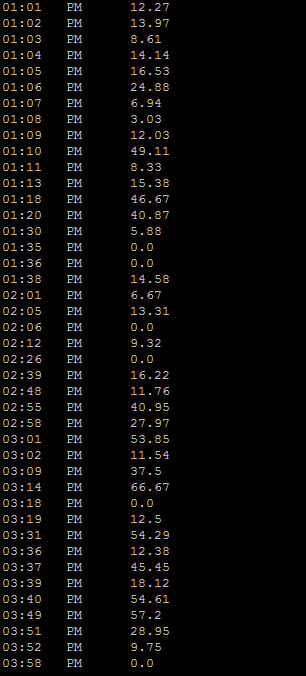
null 5.26



**--Find CTOR by time of the day**

SELECT TimeStamp,Am\_Pm,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

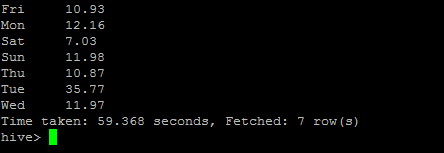
from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY TimeStamp,Am\_Pm ORDER BY TimeStamp,Am\_Pm;



**--Find CTOR by Day of the week**

SELECT Day,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

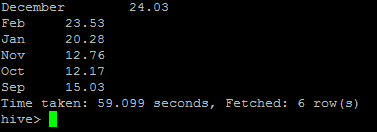
from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY Day ORDER BY Day;



**--Find CTOR by Month**

SELECT Month\_Desc,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

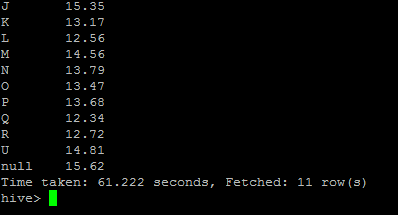
from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY Month\_Desc ORDER BY Month\_Desc;



**--Find CTOR by Lead's Income group**

SELECT TRW\_INCOME\_CD\_V4,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

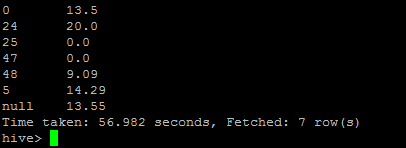
from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY TRW\_INCOME\_CD\_V4 ORDER BY TRW\_INCOME\_CD\_V4;



**--Find CTOR by Lead's Ethnicity**

SELECT ASIAN\_CD,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

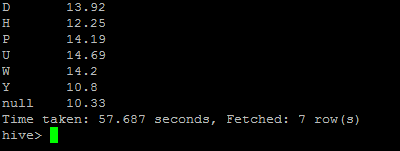
from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY ASIAN\_CD ORDER BY ASIAN\_CD;



**--Find CTOR by Lead's Household status**

SELECT I1\_INDIV\_HHLD\_STATUS\_CODE,ROUND(SUM(CASE WHEN CLICK\_FLG='Y' THEN 1 ELSE 0 END)/count(\*)\*100,2)

from jig13895.bdatafinalga where OPEN\_FLG='Y' GROUP BY I1\_INDIV\_HHLD\_STATUS\_CODE ORDER BY I1\_INDIV\_HHLD\_STATUS\_CODE;



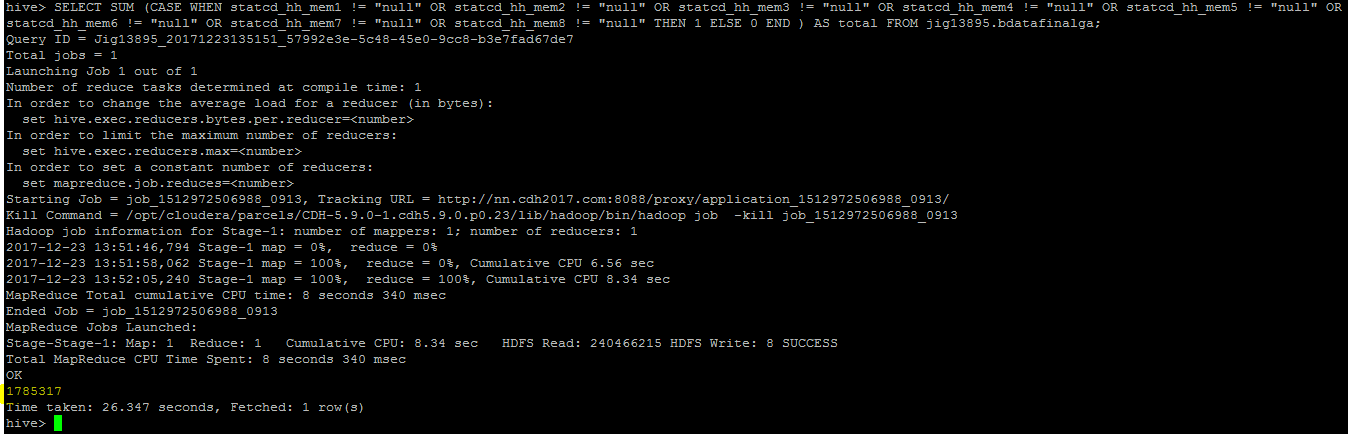
**Tableau Visualization Path**: 52.3.237.208 inside E: drive🡪users🡪jig13895🡪Documents🡪My Tableau Repository🡪workbooks🡪jig13895 Final Assignment

**2.Find Household members information**

**a.Find count of leads who has information about the members of household**

SELECT SUM (CASE WHEN statcd\_hh\_mem1 != "null" OR statcd\_hh\_mem2 != "null" OR statcd\_hh\_mem3 != "null" OR statcd\_hh\_mem4 != "null" OR statcd\_hh\_mem5 != "null" OR statcd\_hh\_mem6 != "null" OR statcd\_hh\_mem7 != "null" OR statcd\_hh\_mem8 != "null" THEN 1 ELSE 0 END ) AS total FROM jig13895.bdatafinalga;

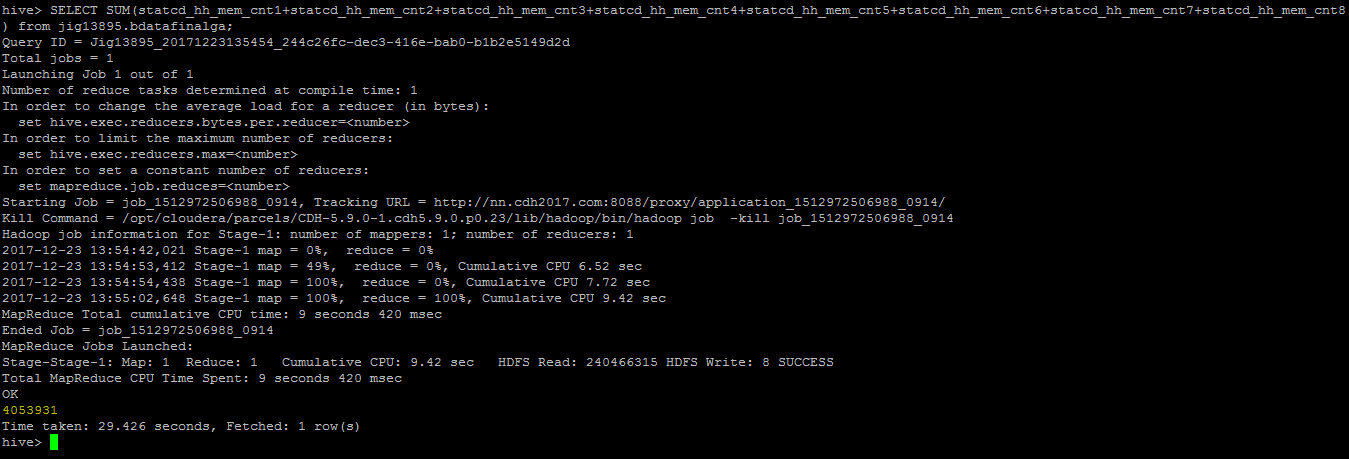
o/p-->1785317



**b.Find count of household number of household members**

SELECT SUM(statcd\_hh\_mem\_cnt1+statcd\_hh\_mem\_cnt2+statcd\_hh\_mem\_cnt3+statcd\_hh\_mem\_cnt4+statcd\_hh\_mem\_cnt5+statcd\_hh\_mem\_cnt6+statcd\_hh\_mem\_cnt7+statcd\_hh\_mem\_cnt8) from jig13895.bdatafinalgasampleNew;

o/p--> 4053931



**c.Find count of household members by type**

SELECT SUM((CASE WHEN statcd\_hh\_mem1 == "U" THEN 1 ELSE 0 END) +

(CASE WHEN statcd\_hh\_mem2 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem3 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem4 == "U" THEN 1 ELSE 0 END)+

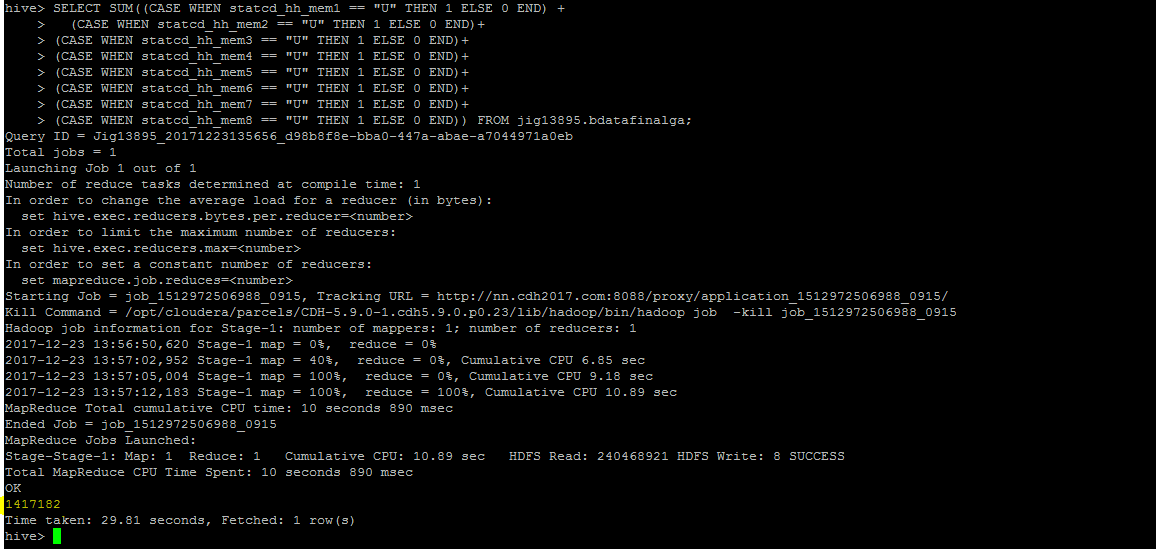
(CASE WHEN statcd\_hh\_mem5 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem6 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem7 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem8 == "U" THEN 1 ELSE 0 END)) FROM jig13895.bdatafinalga;

o/p-->1417182



**d. Find % of household members type**

SELECT ROUND((SUM((CASE WHEN statcd\_hh\_mem1 == "U" THEN 1 ELSE 0 END) +

(CASE WHEN statcd\_hh\_mem2 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem3 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem4 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem5 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem6 == "U" THEN 1 ELSE 0 END)+

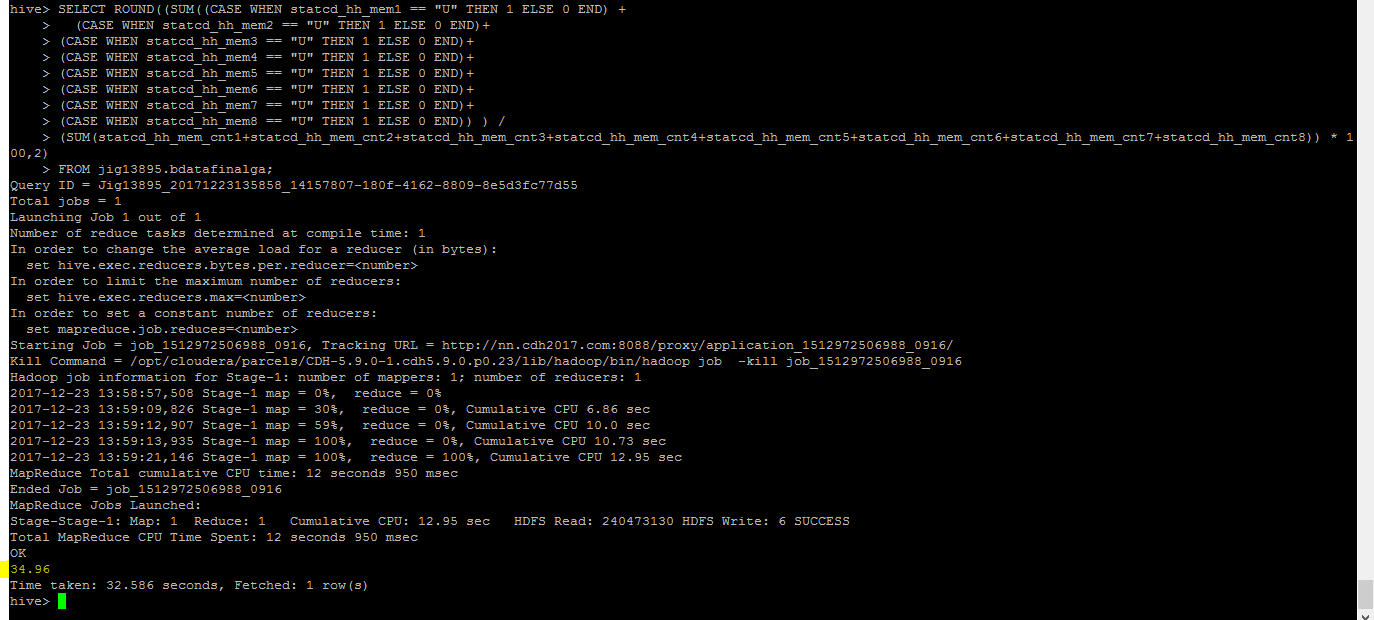
(CASE WHEN statcd\_hh\_mem7 == "U" THEN 1 ELSE 0 END)+

(CASE WHEN statcd\_hh\_mem8 == "U" THEN 1 ELSE 0 END)) ) /

(SUM(statcd\_hh\_mem\_cnt1+statcd\_hh\_mem\_cnt2+statcd\_hh\_mem\_cnt3+statcd\_hh\_mem\_cnt4+statcd\_hh\_mem\_cnt5+statcd\_hh\_mem\_cnt6+statcd\_hh\_mem\_cnt7+statcd\_hh\_mem\_cnt8)) \* 100,2)

FROM jig13895.bdatafinalga;

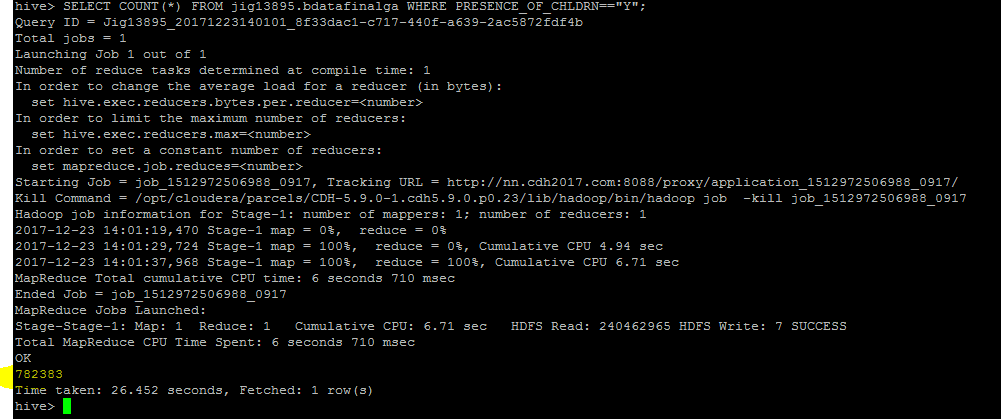
o/p-->34.96



**e. how many households have children**

SELECT COUNT(\*) FROM jig13895.bdatafinalga WHERE PRESENCE\_OF\_CHLDRN=="Y";

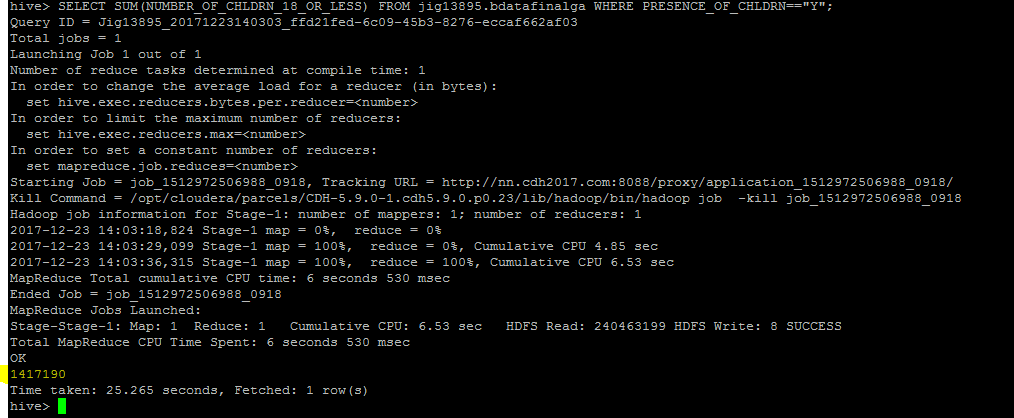
O/P-->782383



**f.Overall how many children are there**

SELECT SUM(NUMBER\_OF\_CHLDRN\_18\_OR\_LESS) FROM jig13895.bdatafinalga WHERE PRESENCE\_OF\_CHLDRN=="Y";

o/p 🡪1417190



**g. how many of the children are male and female**

SELECT SUM((CASE WHEN GNDR\_OF\_CHLDRN\_0\_3 == "M" THEN 1 ELSE 0 END) +

(CASE WHEN GNDR\_OF\_CHLDRN\_4\_6 == "M" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_7\_9 == "M" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_10\_12 == "M" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_13\_18 == "M" THEN 1 ELSE 0 END)) AS Males, SUM((CASE WHEN GNDR\_OF\_CHLDRN\_0\_3 == "F" THEN 1 ELSE 0 END) +

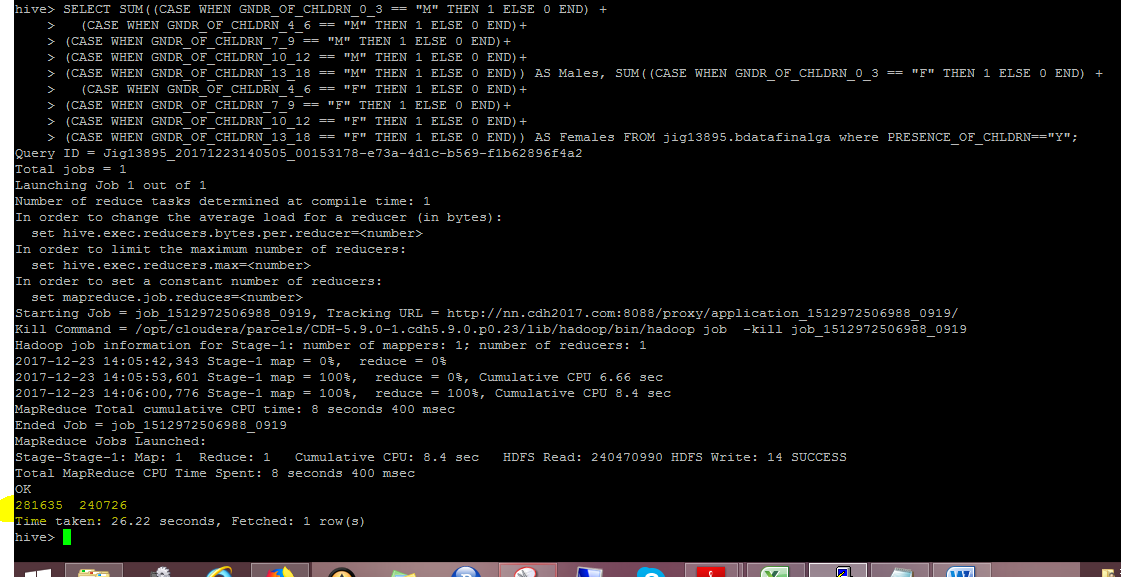
(CASE WHEN GNDR\_OF\_CHLDRN\_4\_6 == "F" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_7\_9 == "F" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_10\_12 == "F" THEN 1 ELSE 0 END)+

(CASE WHEN GNDR\_OF\_CHLDRN\_13\_18 == "F" THEN 1 ELSE 0 END)) AS Females FROM jig13895.bdatafinalga where PRESENCE\_OF\_CHLDRN=="Y" and new\_id='160033';

o/p-->Males 281635; Females : 240726



**Kindly find the scripts** 