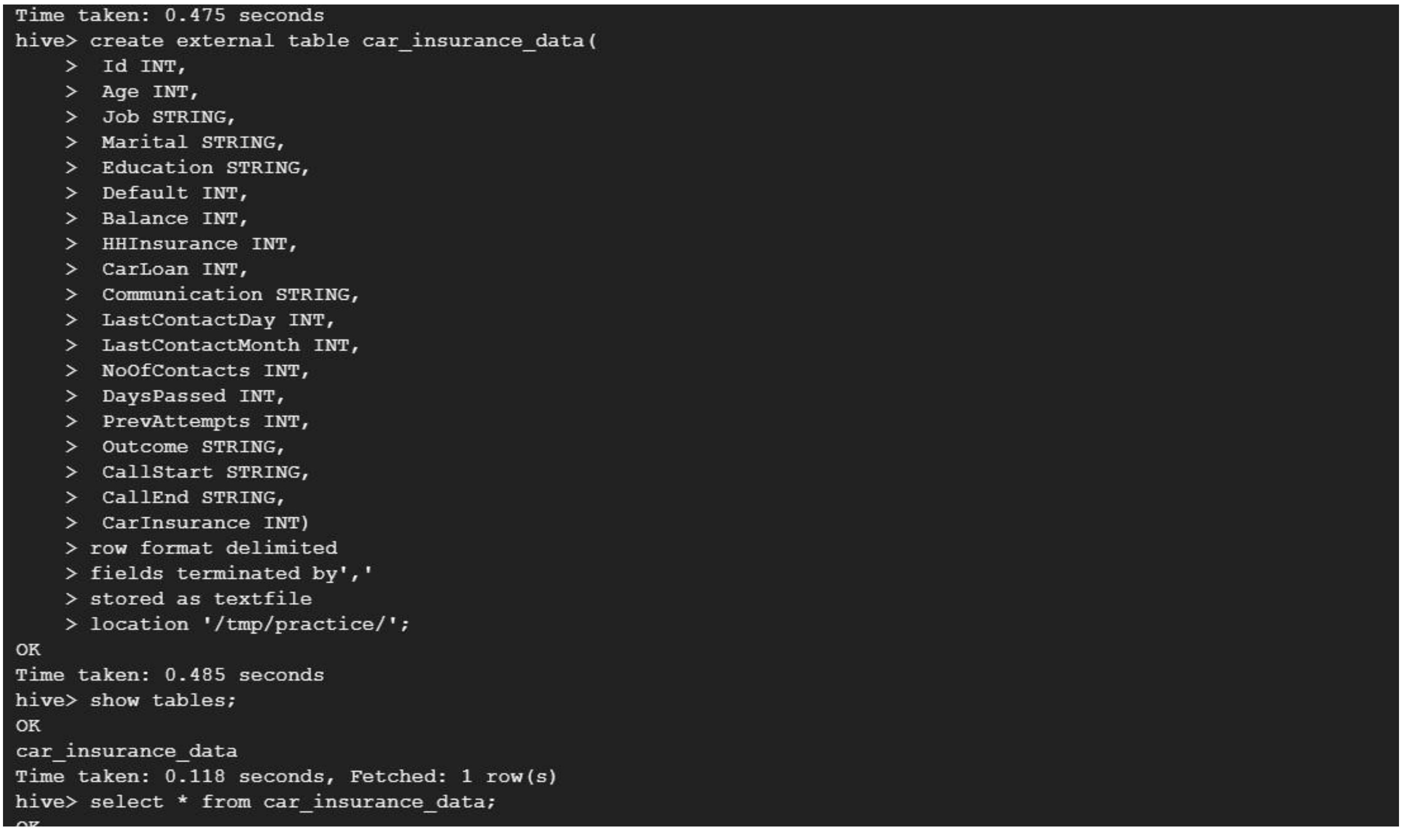
Hive Assignment - 1

Car Insurance Cold Calls Data Analysis

Problem 1 : Data Loading

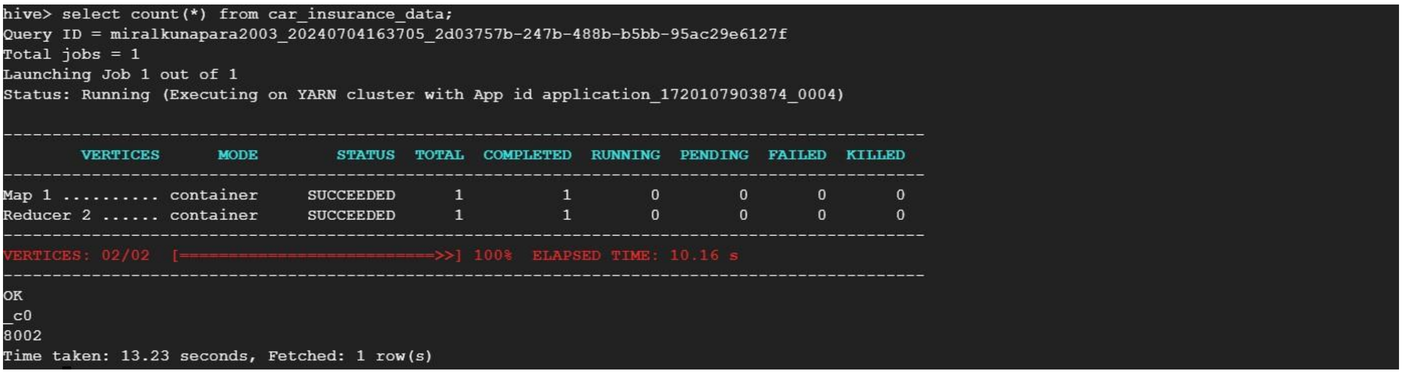
1. Create an external table with the given schema and table should store data as text file from hdfs path.



```
Time taken: 0.475 seconds
hive> create external table car_insurance_data(
    >  Id INT,
    >  Age INT,
    >  Job STRING,
    >  Marital STRING,
    >  Education STRING,
    >  Default INT,
    >  Balance INT,
    >  HHInsurance INT,
    >  CarLoan INT,
    >  Communication STRING,
    >  LastContactDay INT,
    >  LastContactMonth INT,
    >  NoOfContacts INT,
    >  DaysPassed INT,
    >  PrevAttempts INT,
    >  Outcome STRING,
    >  CallStart STRING,
    >  CallEnd STRING,
    >  CarInsurance INT)
    > row format delimited
    > fields terminated by','
    > stored as textfile
    > location '/tmp/practice/';
OK
Time taken: 0.485 seconds
hive> show tables;
OK
car_insurance_data
Time taken: 0.118 seconds, Fetched: 1 row(s)
hive> select * from car_insurance_data;
OK
```

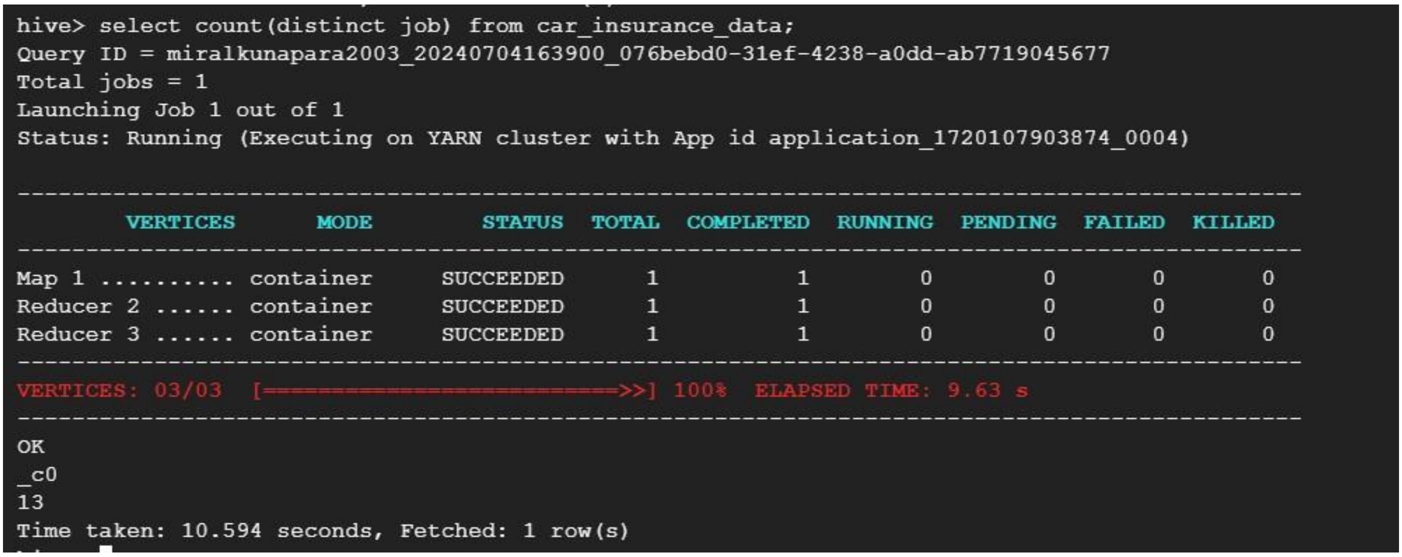
Problem 2 : Data Exploration

1. How many records are ther in database?



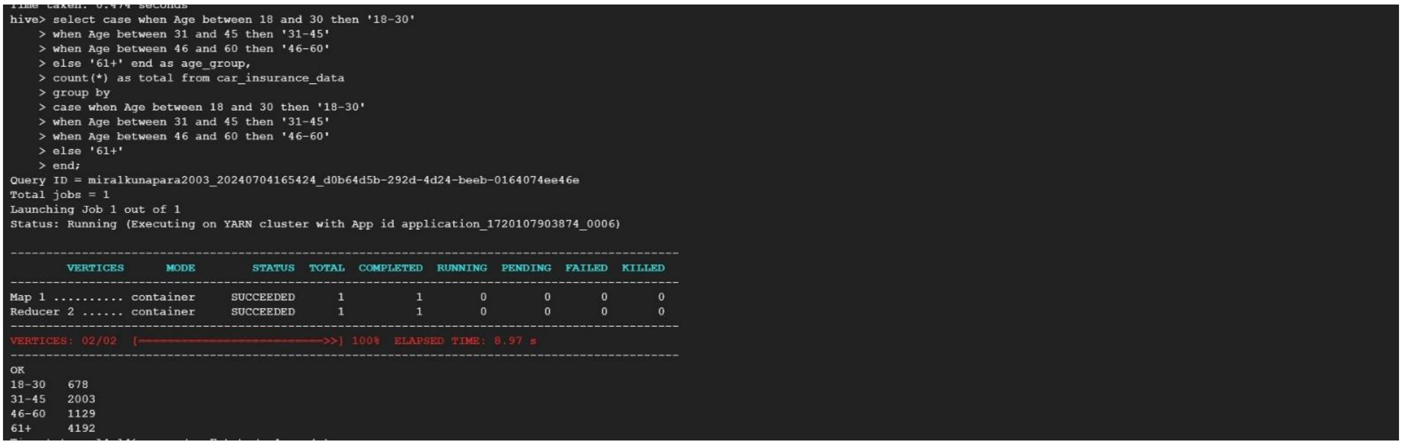
```
hive> select count(*) from car_insurance_data;
Query ID = miralkunapara2003_20240704163705_2d03757b-247b-488b-b5bb-95ac29e6127f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.16 s
----------------------------------------------------------------------------------------------
OK
_c0
8002
Time taken: 13.23 seconds, Fetched: 1 row(s)
```

2. How many unique job categories are there ?



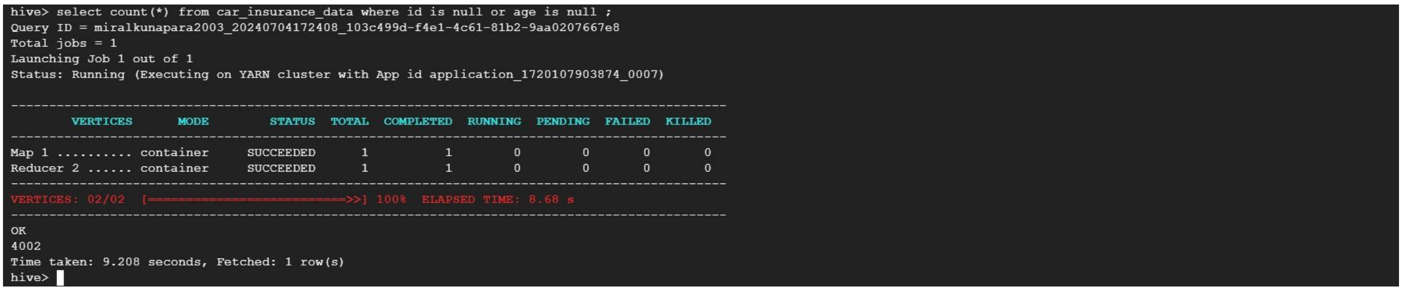
```
hive> select count(distinct job) from car_insurance_data;
Query ID = miralkunapara2003_20240704163900_076bebd0-31ef-4238-a0dd-ab7719045677
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.63 s
----------------------------------------------------------------------------------------------
OK
_c0
13
Time taken: 10.594 seconds, Fetched: 1 row(s)
```

3. What is the age distribution of customers in the dataset ? Provide a breakdown by age group :18-30,31-45,46-60,61+



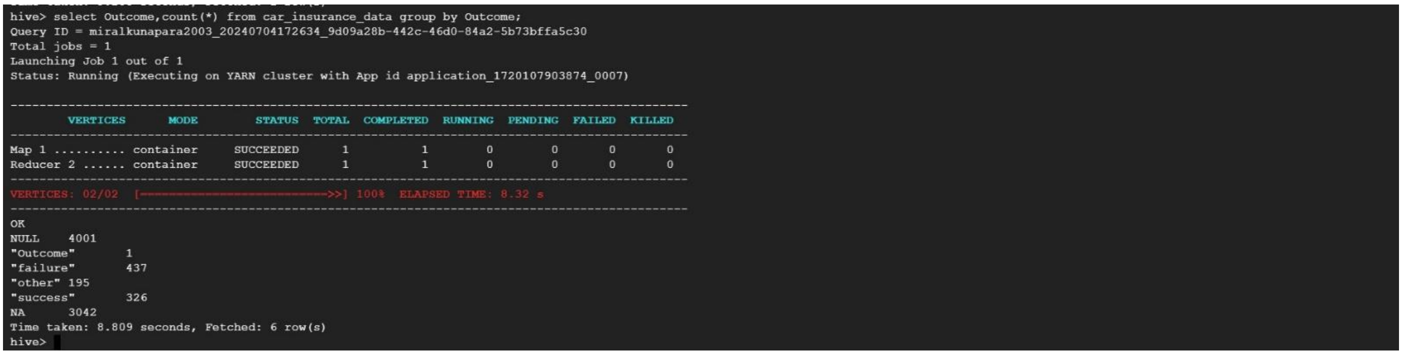
```
hive> select case when Age between 18 and 30 then '18-30'
    > when Age between 31 and 45 then '31-45'
    > when Age between 46 and 60 then '46-60'
    > else '61+' end as age_group,
    > count(*) as total from car_insurance_data
    > group by
    > case when Age between 18 and 30 then '18-30'
    > when Age between 31 and 45 then '31-45'
    > when Age between 46 and 60 then '46-60'
    > else '61+'
    > end;
Query ID = miralkunapara2003_20240704165424_d0b64d5b-292d-4d24-beeb-0164074ee46e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.97 s
----------------------------------------------------------------------------------------------
OK
18-30   678
31-45   2003
46-60   1129
61+     4192
```

4. Count the number of records that have missing values in any fields .



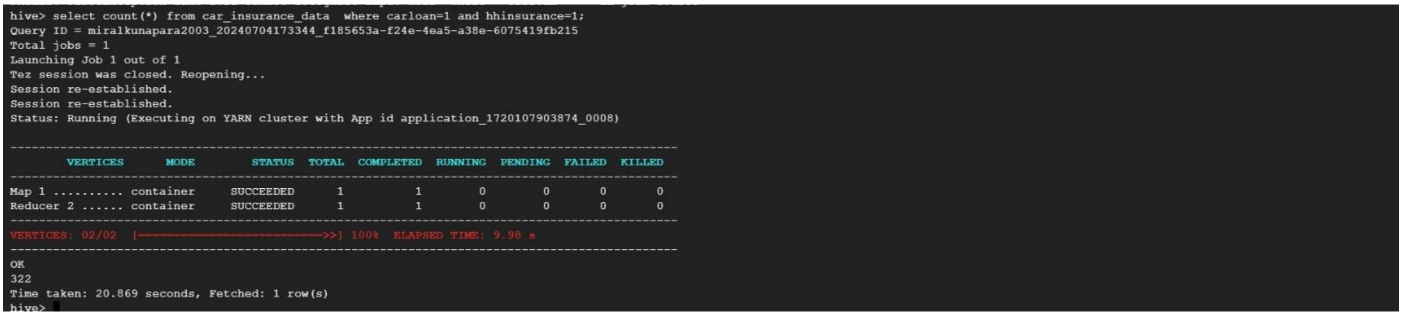
```
hive> select count(*) from car_insurance_data where id is null or age is null ;
Query ID = miralkunapara2003_20240704172408_103c499d-f4e1-4c61-81b2-9aa0207667e8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.68 s
----------------------------------------------------------------------------------------------
OK
4002
Time taken: 9.208 seconds, Fetched: 1 row(s)
hive>
```

5. Determine the number of unique ‘outocme’values and their respective counts .



```
hive> select Outcome,count(*) from car_insurance_data group by Outcome;
Query ID = miralkunapara2003_20240704172634_9d09a28b-442c-46d0-84a2-5b73bffa5c30
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.32 s
----------------------------------------------------------------------------------------------
OK
NULL    4001
"Outcome"       1
"failure"       437
"other" 195
"success"       326
NA      3042
Time taken: 8.809 seconds, Fetched: 6 row(s)
hive>
```

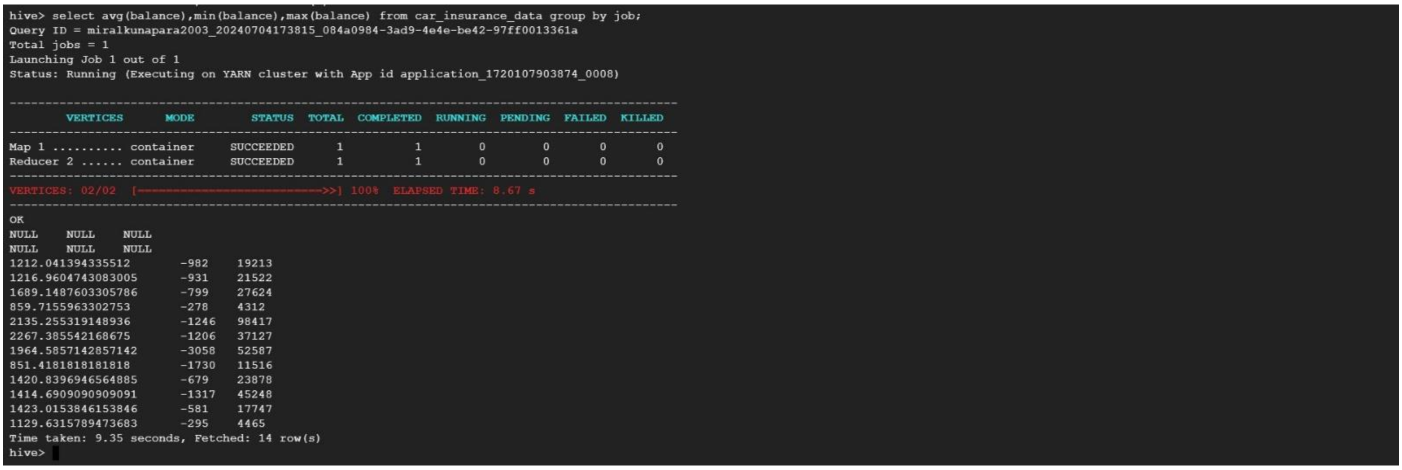
6. Find the number of customers who have both a car loan and home insurance.



```
hive> select count(*) from car_insurance_data  where carloan=1 and hhinsurance=1;
Query ID = miralkunapara2003_20240704173344_f185653a-f24e-4ea5-a38e-6075419fb215
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720107903874_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.98 s
----------------------------------------------------------------------------------------------
OK
322
Time taken: 20.869 seconds, Fetched: 1 row(s)
hive>
```

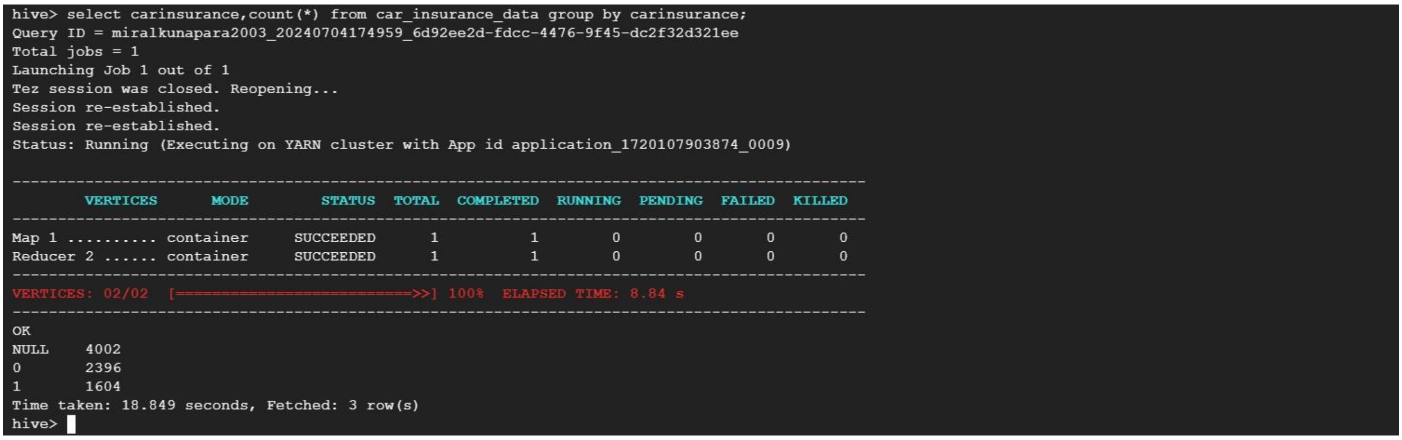
Problem 3 : Aggregations

1. What is the average,minimum and maximum balance for each job category?



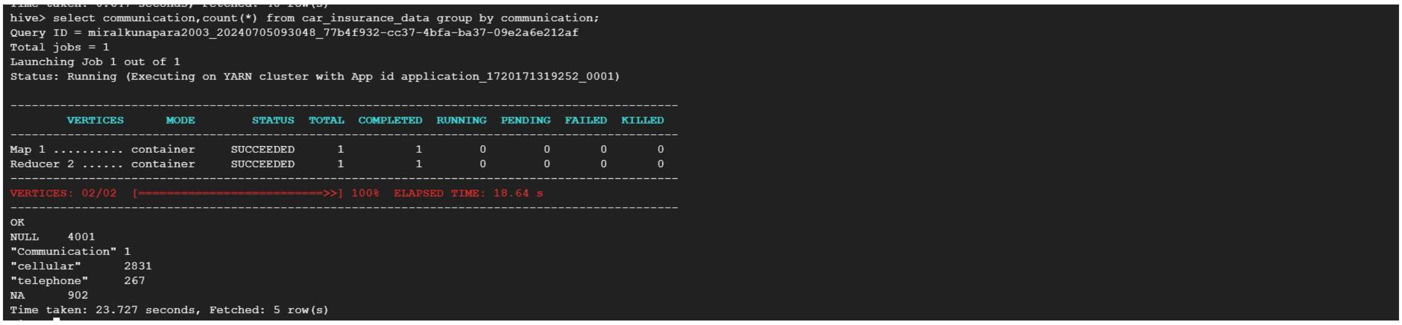
```
hive> select avg(balance),min(balance),max(balance) from car_insurance_data group by job;
Query ID = miralkunapara2003_20240704173815_084a0984-3ad9-4e4e-be42-97ff0013361a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720107903874_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.67 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL
NULL    NULL    NULL
1212.041394335512       -982    19213
1216.9604743083005      -931    21522
1689.1487603305786      -799    27624
859.7155963302753       -278    4312
2135.255319148936       -1246   98417
2267.385542168675       -1206   37127
1964.5857142857142      -3058   52587
851.4181818181818       -1730   11516
1420.8396946564885      -679    23878
1414.6909090909091      -1317   45248
1423.0153846153846      -581    17747
1129.6315789473683      -295    4465
Time taken: 9.35 seconds, Fetched: 14 row(s)
hive>
```

2. Find the total number of customers with and without car insurance .



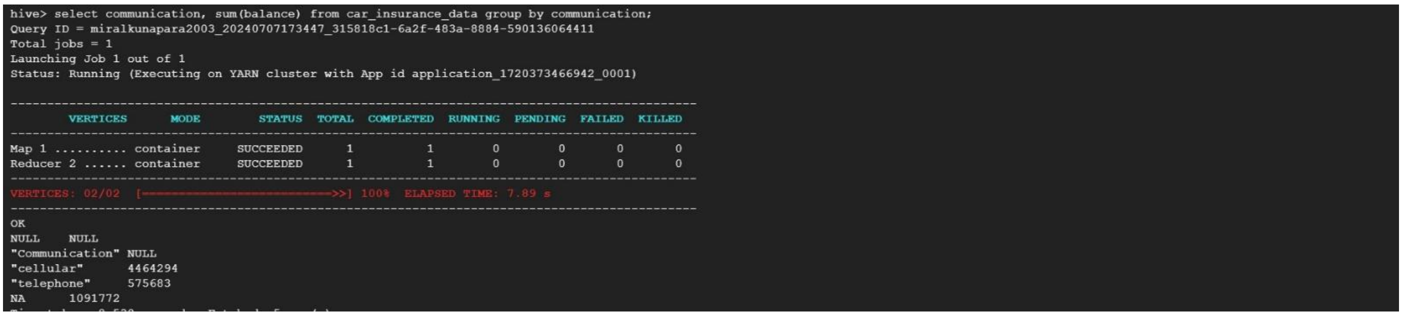
```
hive> select carinsurance,count(*) from car_insurance_data group by carinsurance;
Query ID = miralkunapara2003_20240704174959_6d92ee2d-fdcc-4476-9f45-dc2f32d321ee
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720107903874_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.84 s
----------------------------------------------------------------------------------------------
OK
NULL    4002
0       2396
1       1604
Time taken: 18.849 seconds, Fetched: 3 row(s)
hive>
```

3. Count the number of customers for each communication type .



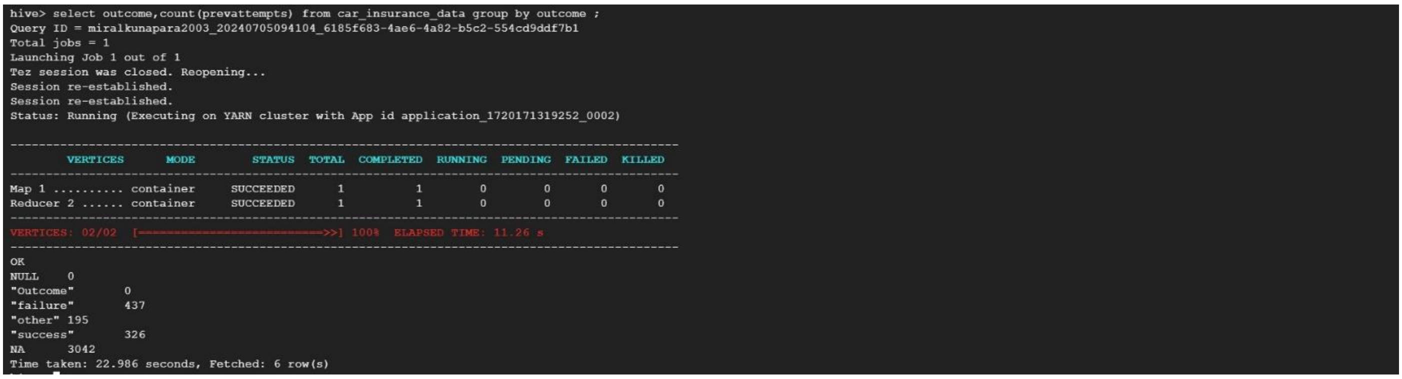
```
hive> select communication,count(*) from car_insurance_data group by communication;
Query ID = miralkunapara2003_20240705093048_77b4f932-cc37-4bfa-ba37-09e2a6e212af
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720171319252_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.64 s
----------------------------------------------------------------------------------------------
OK
NULL    4001
"Communication" 1
"cellular"      2831
"telephone"     267
NA      902
Time taken: 23.727 seconds, Fetched: 5 row(s)
```

4. Calculate the sum of ‘balance’ for each ‘communication’ type



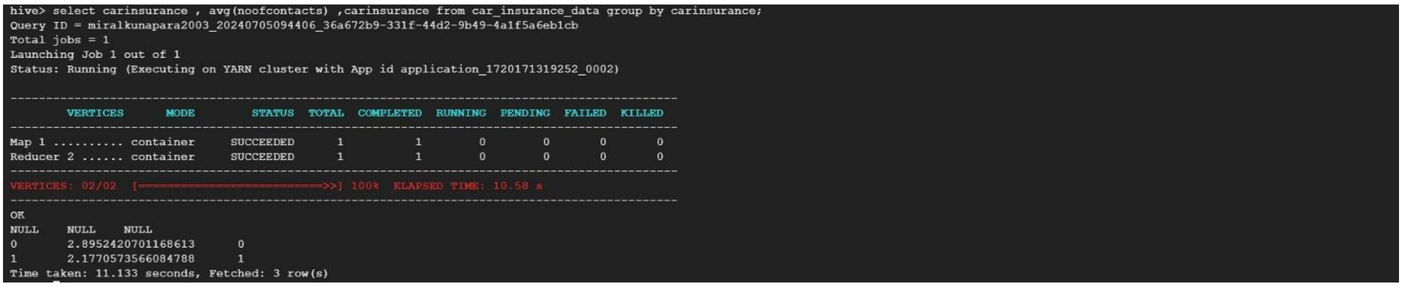
```
hive> select communication, sum(balance) from car_insurance_data group by communication;
Query ID = miralkunapara2003_20240707173447_315818c1-6a2f-483a-8884-590136064411
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720373466942_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.89 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL
"Communication" NULL
"cellular"      4464294
"telephone"     575683
NA      1091772
```

5. Count the number of ‘PrevAttemps’ for each ‘Outcome’ type .



```
hive> select outcome,count(prevattempts) from car_insurance_data group by outcome ;
Query ID = miralkunapara2003_20240705094104_6185f683-4ae6-4a82-b5c2-554cd9ddf7b1
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720171319252_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.26 s
----------------------------------------------------------------------------------------------
OK
NULL    0
"Outcome"       0
"failure"       437
"other" 195
"success"       326
NA      3042
Time taken: 22.986 seconds, Fetched: 6 row(s)
```

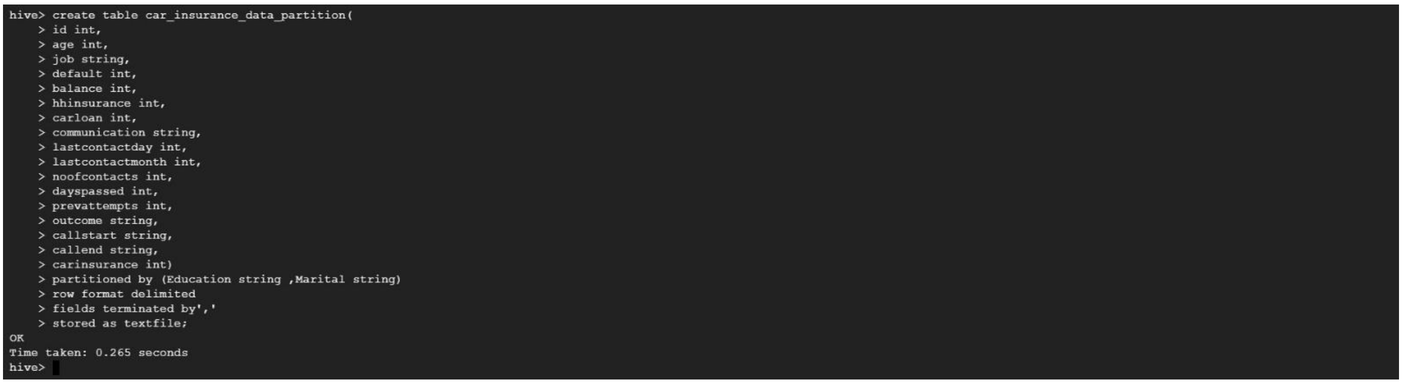
6. Calculate the average ‘Noofcontacts’ for people with and without ‘carinsurance’.



```
hive> select carinsurance , avg(noofcontacts) ,carinsurance from car_insurance_data group by carinsurance;
Query ID = miralkunapara2003_20240705094406_36a672b9-331f-44d2-9b49-4a1f5a6eb1cb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720171319252_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.58 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL
0       2.8952420701168613      0
1       2.1770573566084788      1
Time taken: 11.133 seconds, Fetched: 3 row(s)
```

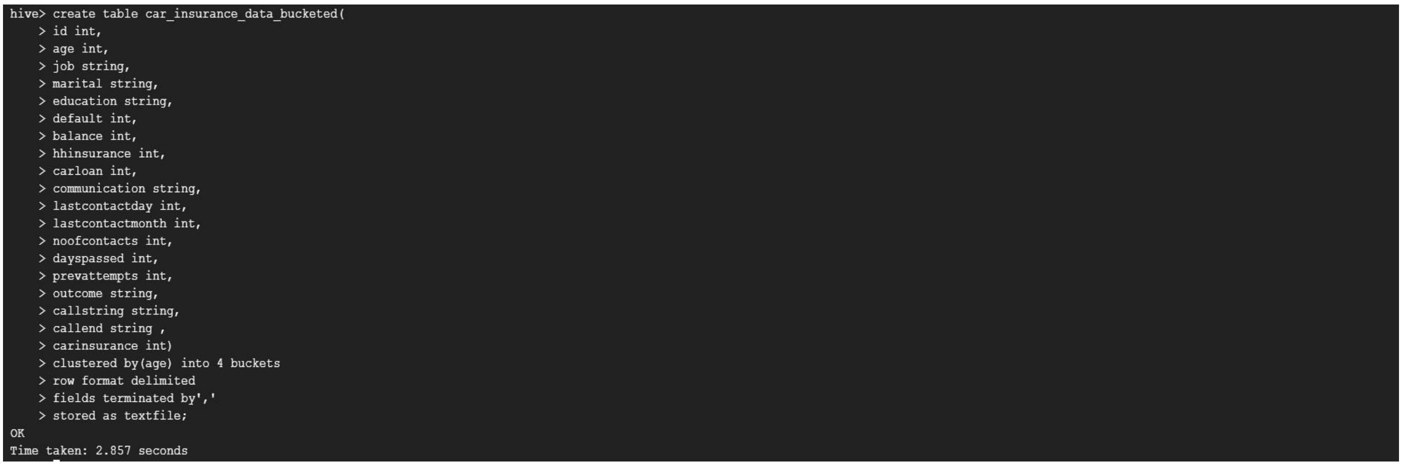
Problem 4 : Partitioning and Bucketing

1. Create a partitioned table on ‘Education’ and ‘Marital’ status . Load data from the original table to this new partitioned table



```
hive> create table car_insurance_data_partition(
    > id int,
    > age int,
    > job string,
    > default int,
    > balance int,
    > hhinsurance int,
    > carloan int,
    > communication string,
    > lastcontactday int,
    > lastcontactmonth int,
    > noofcontacts int,
    > dayspassed int,
    > prevattempts int,
    > outcome string,
    > callstart string,
    > callend string,
    > carinsurance int)
    > partitioned by (Education string ,Marital string)
    > row format delimited
    > fields terminated by','
    > stored as textfile;
OK
Time taken: 0.265 seconds
hive>
```

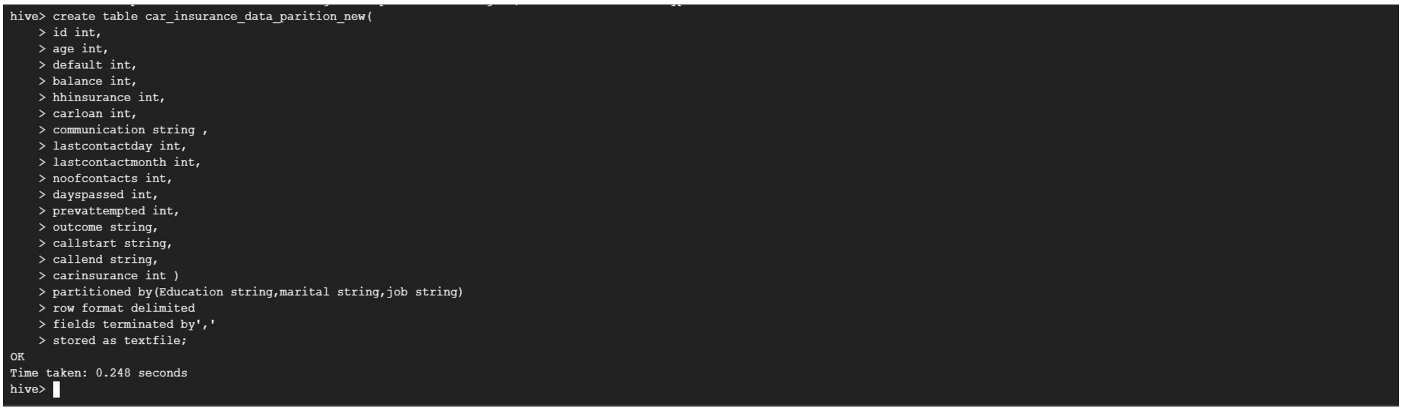
2. Create bucketed table on ‘age’,bucketed into 4 groups(as per the age group mentioned above).Load data from orginial table into bucketed table .



```
hive> create table car_insurance_data_bucketed(
    > id int,
    > age int,
    > job string,
    > marital string,
    > education string,
    > default int,
    > balance int,
    > hhinsurance int,
    > carloan int,
    > communication string,
    > lastcontactday int,
    > lastcontactmonth int,
    > noofcontacts int,
    > dayspassed int,
    > prevattempts int,
    > outcome string,
    > callstring string,
    > callend string ,
    > carinsurance int)
    > clustered by(age) into 4 buckets
    > row format delimited
    > fields terminated by','
    > stored as textfile;
OK
Time taken: 2.857 seconds
```

3. Add an additional partition on ‘job’ to the partitioned table created earlier and move the data accordingly .

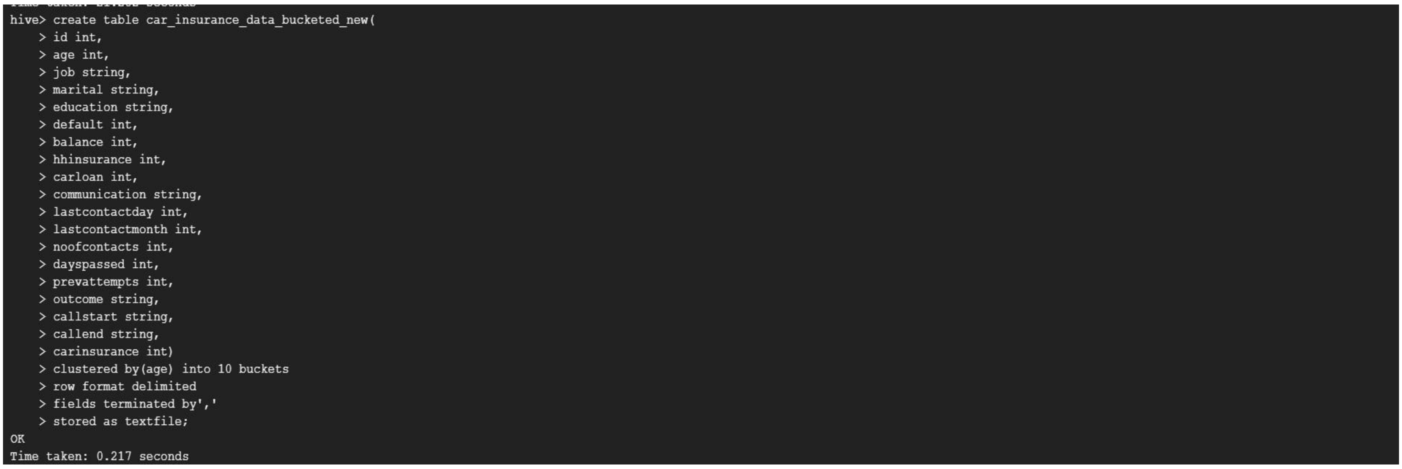
- Once we created a partitioned table ,after hive does not allow altering the partitioning of existing tables.



```
hive> create table car_insurance_data_parition_new(
    > id int,
    > age int,
    > default int,
    > balance int,
    > hhinsurance int,
    > carloan int,
    > communication string ,
    > lastcontactday int,
    > lastcontactmonth int,
    > noofcontacts int,
    > dayspassed int,
    > prevattempted int,
    > outcome string,
    > callstart string,
    > callend string,
    > carinsurance int )
    > partitioned by(Education string,marital string,job string)
    > row format delimited
    > fields terminated by','
    > stored as textfile;
OK
Time taken: 0.248 seconds
hive> 
```

4. Increase the number of buckets in the bucketed table to 10 and redistribute the data.

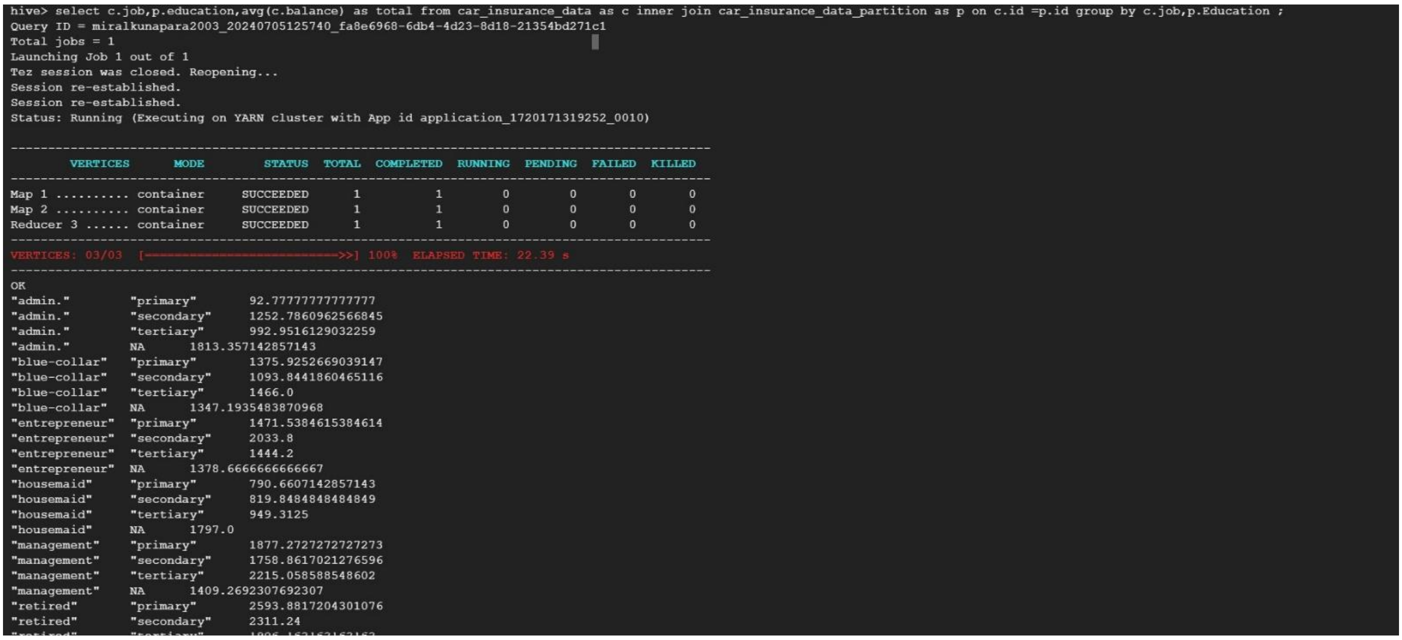
- Once we created a bucketed table ,after hive does not allow altering the bucketing of existing tables.



```
hive> create table car_insurance_data_bucketed_new(
    > id int,
    > age int,
    > job string,
    > marital string,
    > education string,
    > default int,
    > balance int,
    > hhinsurance int,
    > carloan int,
    > communication string,
    > lastcontactday int,
    > lastcontactmonth int,
    > noofcontacts int,
    > dayspassed int,
    > prevattempts int,
    > outcome string,
    > callstart string,
    > callend string,
    > carinsurance int)
    > clustered by(age) into 10 buckets
    > row format delimited
    > fields terminated by','
    > stored as textfile;
OK
Time taken: 0.217 seconds
```

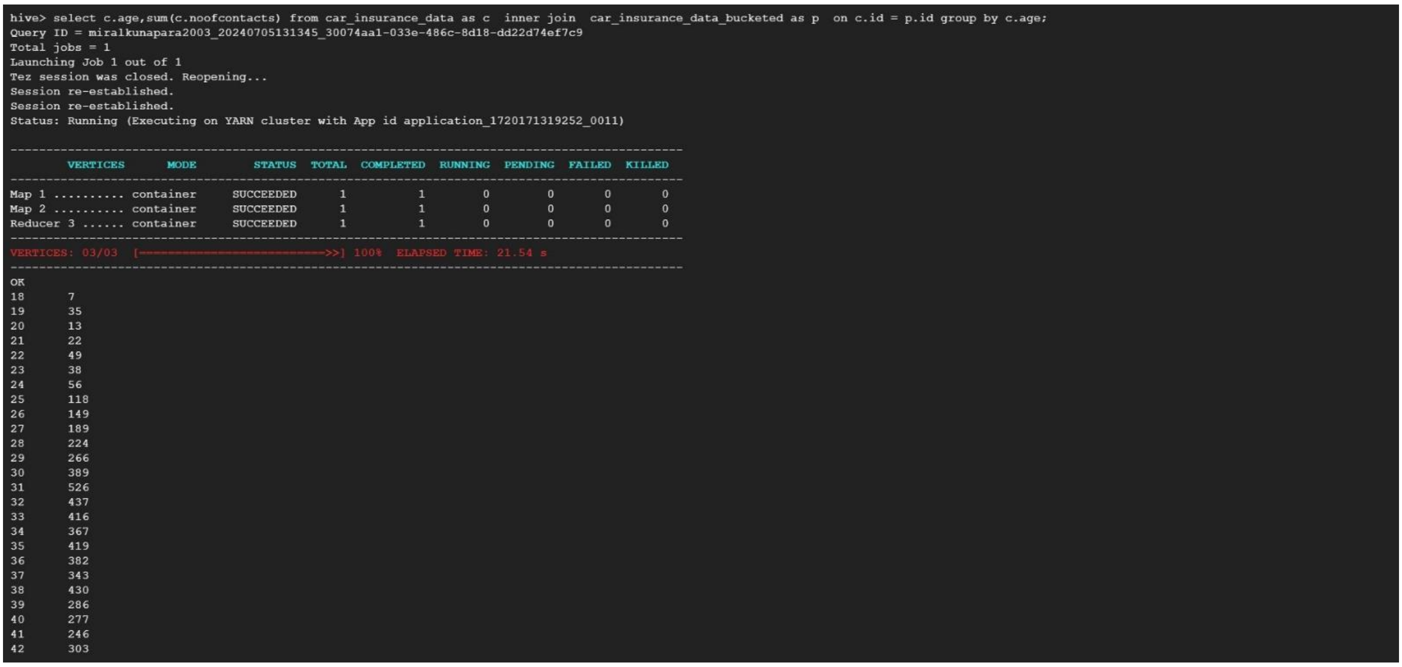
Problem 5 : Optimized join

1. Join the original table with the partitioned table and find out the average ‘Balance’ for each ‘job’ and ‘Education’ level .



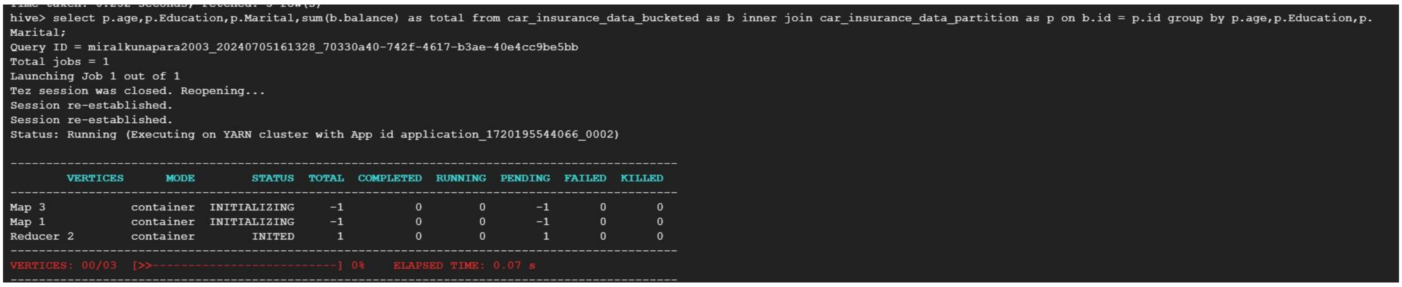
```
hive> select c.job,p.education,avg(c.balance) as total from car_insurance_data as c inner join car_insurance_data_partition as p on c.id =p.id group by c.job,p.Education ;
Query ID = miralkunapara2003_20240705125740_fa8e6968-6db4-4d23-8d18-21354bd271c1
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720171319252_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 22.39 s
----------------------------------------------------------------------------------------------
OK
"admin."        "primary"       92.77777777777777
"admin."        "secondary"     1252.7860962566845
"admin."        "tertiary"      992.9516129032259
"admin."        NA      1813.357142857143
"blue-collar"   "primary"       1375.9252669039147
"blue-collar"   "secondary"     1093.8441860465116
"blue-collar"   "tertiary"      1466.0
"blue-collar"   NA      1347.1935483870968
"entrepreneur"  "primary"       1471.5384615384614
"entrepreneur"  "secondary"     2033.8
"entrepreneur"  "tertiary"      1444.2
"entrepreneur"  NA      1378.6666666666667
"housemaid"     "primary"       790.6607142857143
"housemaid"     "secondary"     819.8484848484849
"housemaid"     "tertiary"      949.3125
"housemaid"     NA      1797.0
"management"    "primary"       1877.2727272727273
"management"    "secondary"     1758.8617021276596
"management"    "tertiary"      2215.058588548602
"management"    NA      1409.2692307692307
"retired"       "primary"       2593.8817204301076
"retired"       "secondary"     2311.24
"retired"       "tertiary"      1906.162162162162
```

2. Join the original table with the bucketed table and calculate the total ‘noofcontacts’ for each ‘age’ group.



```
hive> select c.age,sum(c.noofcontacts) from car_insurance_data as c  inner join  car_insurance_data_bucketed as p  on c.id = p.id group by c.age;
Query ID = miralkunapara2003_20240705131345_30074aa1-033e-486c-8d18-dd22d74ef7c9
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720171319252_0011)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 21.54 s
----------------------------------------------------------------------------------------------
OK
18      7
19      35
20      13
21      22
22      49
23      38
24      56
25      118
26      149
27      189
28      224
29      266
30      389
31      526
32      437
33      416
34      367
35      419
36      382
37      343
38      430
39      286
40      277
41      246
42      303
```

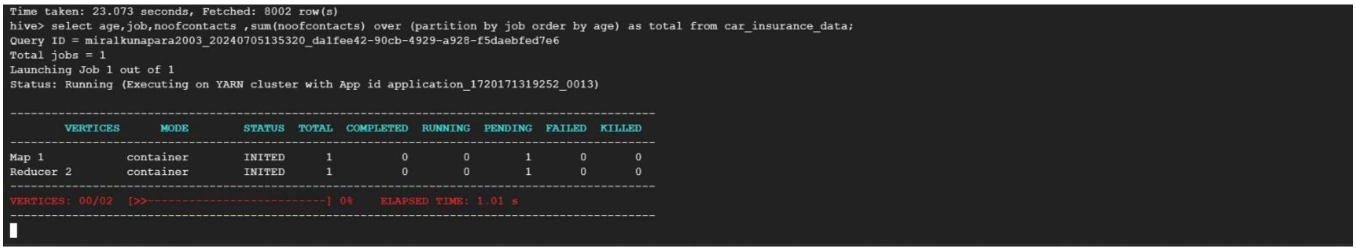
3. Join the partitioned table and the bucketed table based on ‘id’ field and find the total balance for each education level and marital status for each age group.



```
hive> select p.age,p.Education,p.Marital,sum(b.balance) as total from car_insurance_data_bucketed as b inner join car_insurance_data_partition as p on b.id = p.id group by p.age,p.Education,p.
Marital;
Query ID = miralkunapara2003_20240705161328_70330a40-742f-4617-b3ae-40e4cc9be5bb
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 3            container  INITIALIZING     -1          0        0       -1       0       0
Map 1            container  INITIALIZING     -1          0        0       -1       0       0
Reducer 2        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 0.07 s
----------------------------------------------------------------------------------------------
```

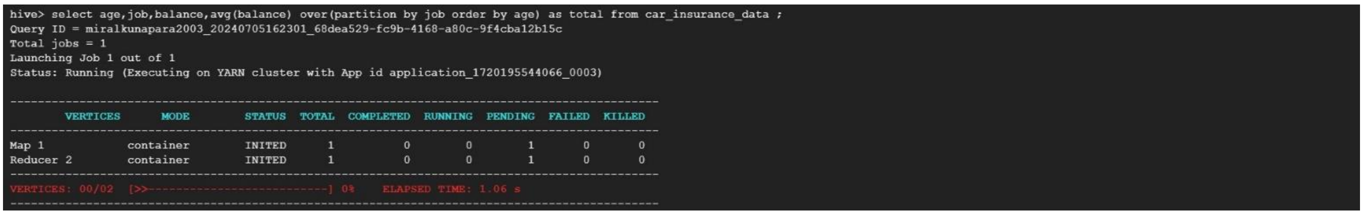
Problem 6 : Window Function

1. Calculate the cumulative sum of ‘noofcontacts’ for each ‘job’ category ,ordered by ‘age’



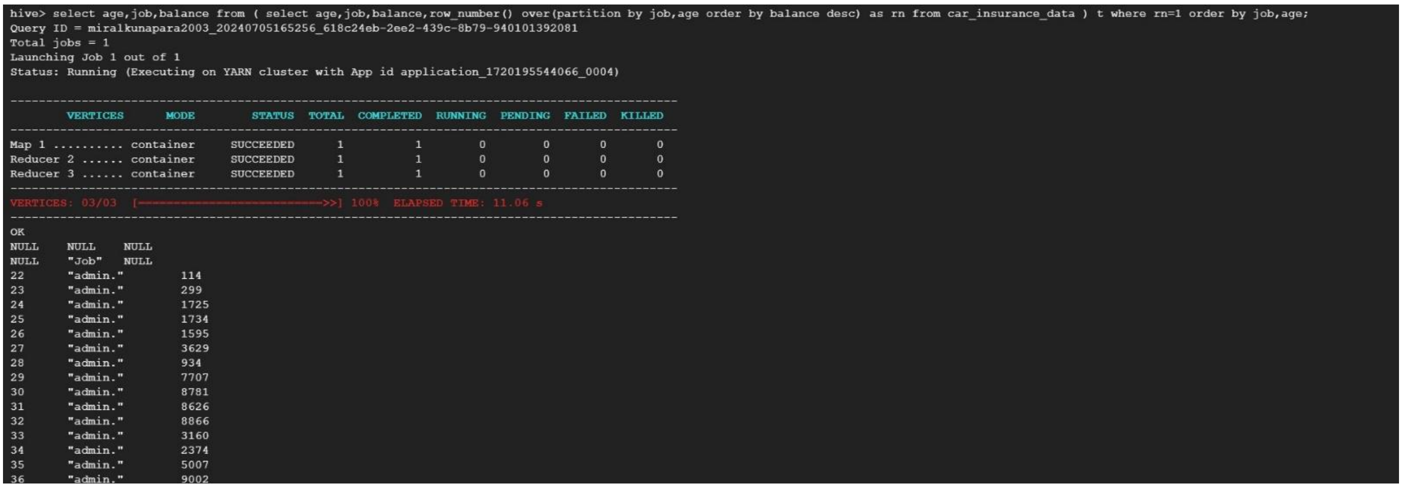
```
Time taken: 23.073 seconds, Fetched: 8002 row(s)
hive> select age,job,noofcontacts ,sum(noofcontacts) over (partition by job order by age) as total from car_insurance_data;
Query ID = miralkunapara2003_20240705135320_da1fee42-90cb-4929-a928-f5daebfed7e6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720171319252_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container        INITED      1          0        0        1       0       0
Reducer 2        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 1.01 s
----------------------------------------------------------------------------------------------
█
```

2. Calculate the running averange of ‘balance’ for each ‘job’ category ,order by ‘age’.



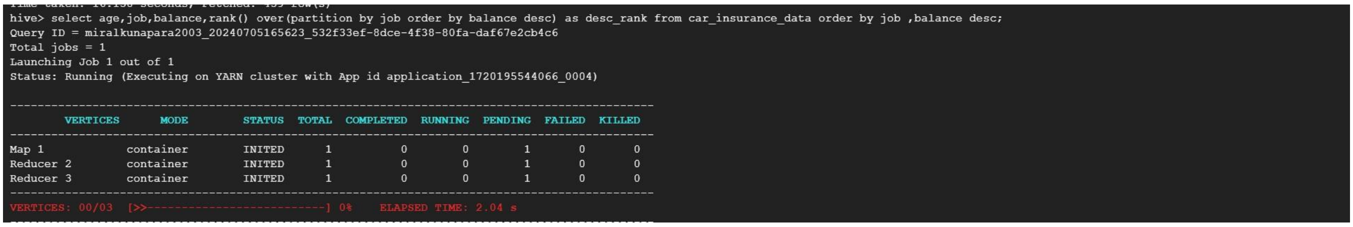
```
hive> select age,job,balance,avg(balance) over(partition by job order by age) as total from car_insurance_data ;
Query ID = miralkunapara2003_20240705162301_68dea529-fc9b-4168-a80c-9f4cba12b15c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container        INITED      1          0        0        1       0       0
Reducer 2        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 1.06 s
----------------------------------------------------------------------------------------------
```

3. For each ‘job’ category, find the maximum ‘Balance’ for each ‘age’ group using window functions.



```
hive> select age,job,balance from ( select age,job,balance,row_number() over(partition by job,age order by balance desc) as rn from car_insurance_data ) t where rn=1 order by job,age;
Query ID = miralkunapara2003_20240705165256_618c24eb-2ee2-439c-8b79-940101392081
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.06 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL
NULL    "Job"   NULL
22      "admin."        114
23      "admin."        299
24      "admin."        1725
25      "admin."        1734
26      "admin."        1595
27      "admin."        3629
28      "admin."        934
29      "admin."        7707
30      "admin."        8781
31      "admin."        8626
32      "admin."        8866
33      "admin."        3160
34      "admin."        2374
35      "admin."        5007
36      "admin."        9002
```

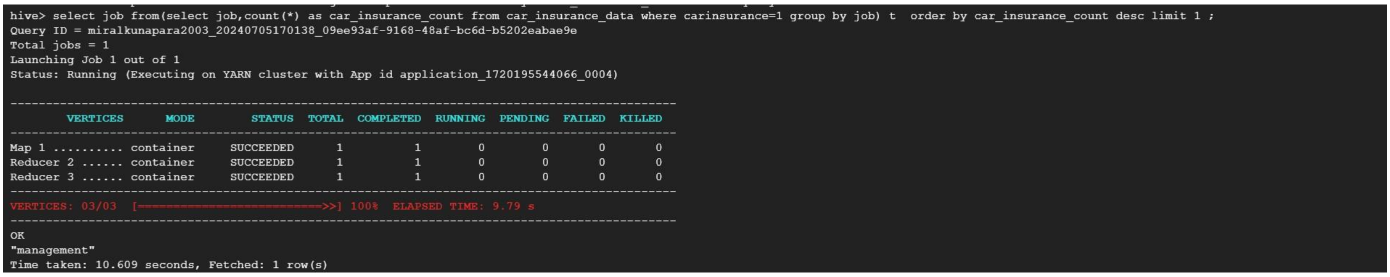
4. Calculate the rank the ‘balance’ within each ‘job’ category , ordered by ‘balance’ descending.



```
hive> select age,job,balance,rank() over(partition by job order by balance desc) as desc_rank from car_insurance_data order by job ,balance desc;
Query ID = miralkunapara2003_20240705165623_532f33ef-8dce-4f38-80fa-daf67e2cb4c6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1            container        INITED      1          0        0        1       0       0
Reducer 2        container        INITED      1          0        0        1       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
----------------------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 2.04 s
----------------------------------------------------------------------------------------------
```

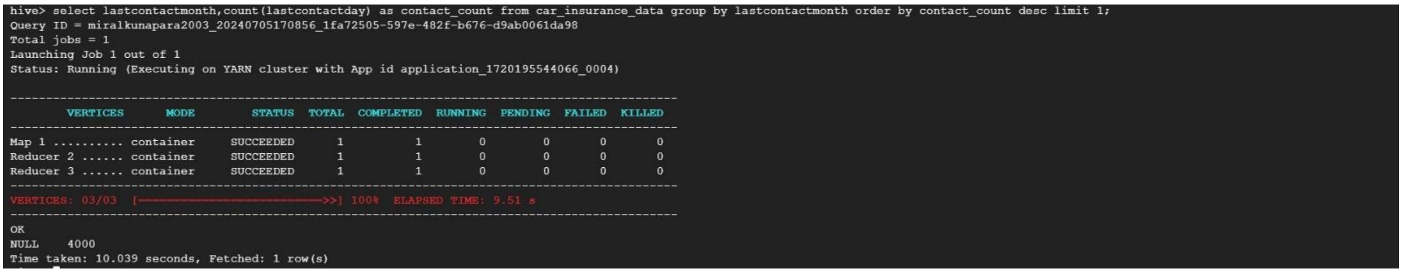
Problem 7 : Advanced Aggregations

1. Find the job category with the highest number of car insurance.



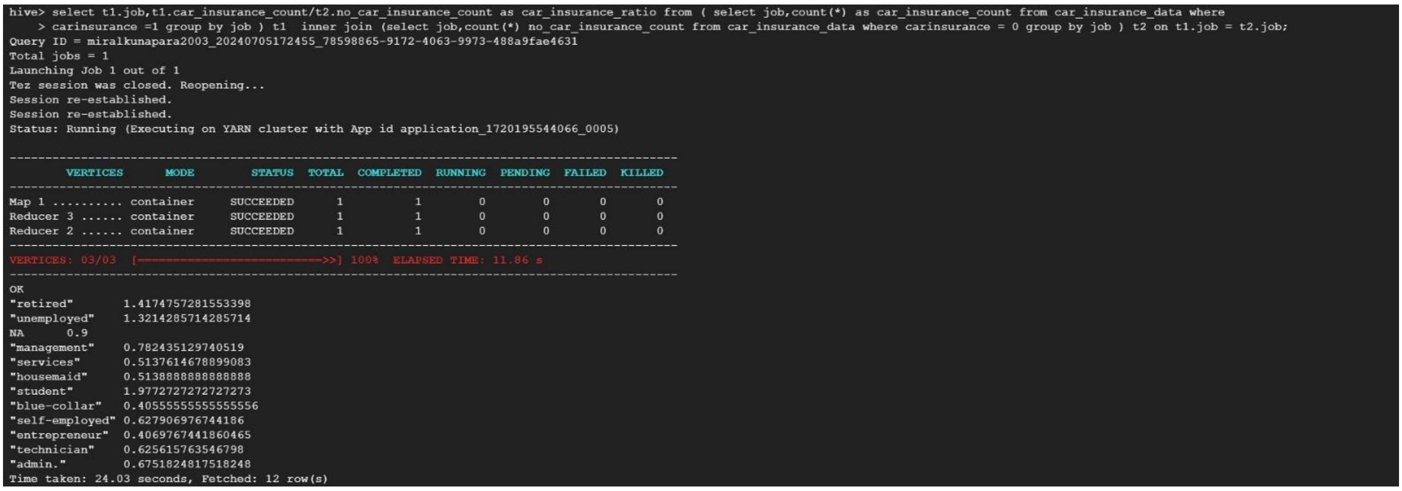
```
hive> select job from(select job,count(*) as car_insurance_count from car_insurance_data where carinsurance=1 group by job) t  order by car_insurance_count desc limit 1 ;
Query ID = miralkunapara2003_20240705170138_09ee93af-9168-48af-bc6d-b5202eabae9e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.79 s
----------------------------------------------------------------------------------------------
OK
"management"
Time taken: 10.609 seconds, Fetched: 1 row(s)
```

2. Which month has been the highest number of last contacts?



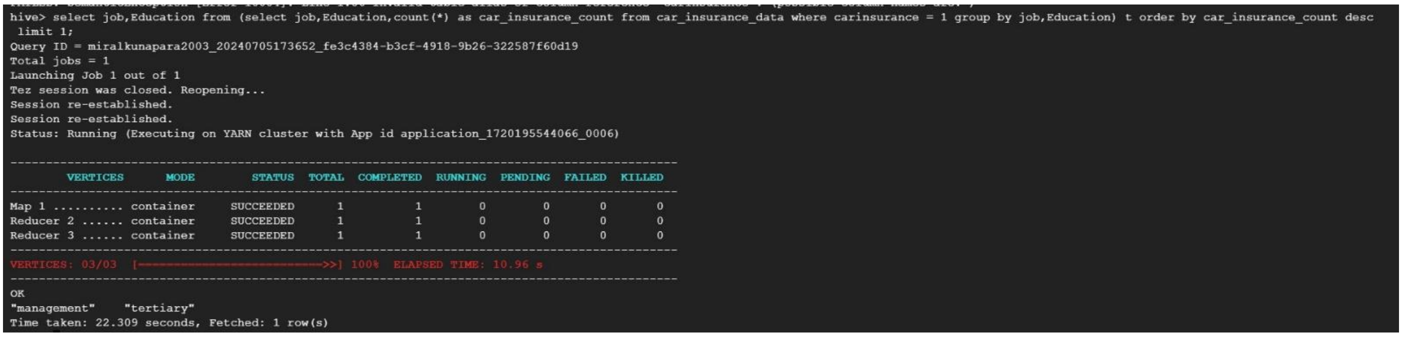
```
hive> select lastcontactmonth,count(lastcontactday) as contact_count from car_insurance_data group by lastcontactmonth order by contact_count desc limit 1;
Query ID = miralkunapara2003_20240705170856_1fa72505-597e-482f-b676-d9ab0061da98
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.51 s
----------------------------------------------------------------------------------------------
OK
NULL    4000
Time taken: 10.039 seconds, Fetched: 1 row(s)
```

3. Calculate the ratio of the number of customers with car insurance to the number of customers without car insurance for each job category.



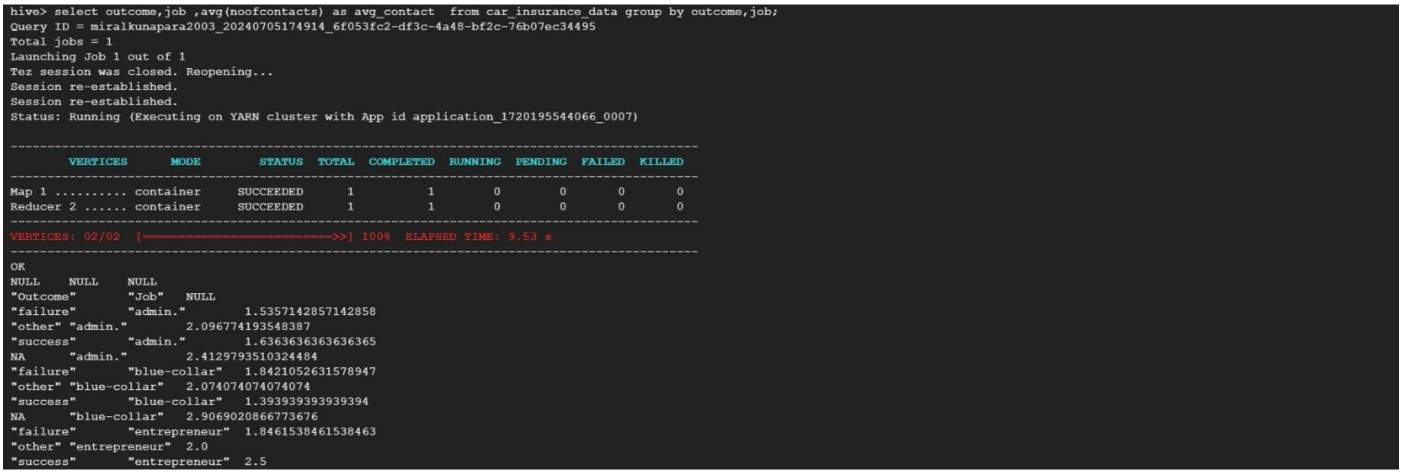
```
hive> select t1.job,t1.car_insurance_count/t2.no_car_insurance_count as car_insurance_ratio from ( select job,count(*) as car_insurance_count from car_insurance_data where
    > carinsurance =1 group by job ) t1  inner join (select job,count(*) no_car_insurance_count from car_insurance_data where carinsurance = 0 group by job ) t2 on t1.job = t2.job;
Query ID = miralkunapara2003_20240705172455_78598865-9172-4063-9973-488a9fae4631
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.86 s
----------------------------------------------------------------------------------------------
OK
"retired"       1.4174757281553398
"unemployed"    1.3214285714285714
NA      0.9
"management"    0.782435129740519
"services"      0.5137614678899083
"housemaid"     0.5138888888888888
"student"       1.9772727272727273
"blue-collar"   0.40555555555555556
"self-employed" 0.627906976744186
"entrepreneur"  0.4069767441860465
"technician"    0.625615763546798
"admin."        0.6751824817518248
Time taken: 24.03 seconds, Fetched: 12 row(s)
```

4. Find out the ‘job’ and ‘education’ level combination which has the highest number of car insurances.



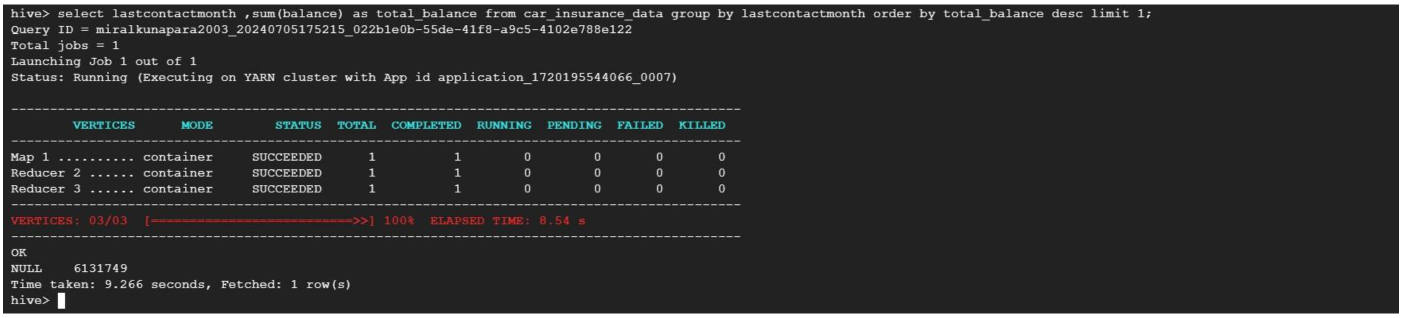
```
hive> select job,Education from (select job,Education,count(*) as car_insurance_count from car_insurance_data where carinsurance = 1 group by job,Education) t order by car_insurance_count desc
 limit 1;
Query ID = miralkunapara2003_20240705173652_fe3c4384-b3cf-4918-9b26-322587f60d19
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 10.96 s
----------------------------------------------------------------------------------------------
OK
"management"    "tertiary"
Time taken: 22.309 seconds, Fetched: 1 row(s)
```

5. Calculate the average ‘noofcontacts’ for each ‘Outcome’ and ‘job’ combination



```
hive> select outcome,job ,avg(noofcontacts) as avg_contact  from car_insurance_data group by outcome,job;
Query ID = miralkunapara2003_20240705174914_6f053fc2-df3c-4a48-bf2c-76b07ec34495
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.53 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL
"Outcome"       "Job"   NULL
"failure"       "admin."        1.5357142857142858
"other" "admin."        2.096774193548387
"success"       "admin."        1.6363636363636365
NA      "admin."        2.4129793510324484
"failure"       "blue-collar"   1.8421052631578947
"other" "blue-collar"   2.074074074074074
"success"       "blue-collar"   1.393939393939394
NA      "blue-collar"   2.9069020866773676
"failure"       "entrepreneur"  1.8461538461538463
"other" "entrepreneur"  2.0
"success"       "entrepreneur"  2.5
```

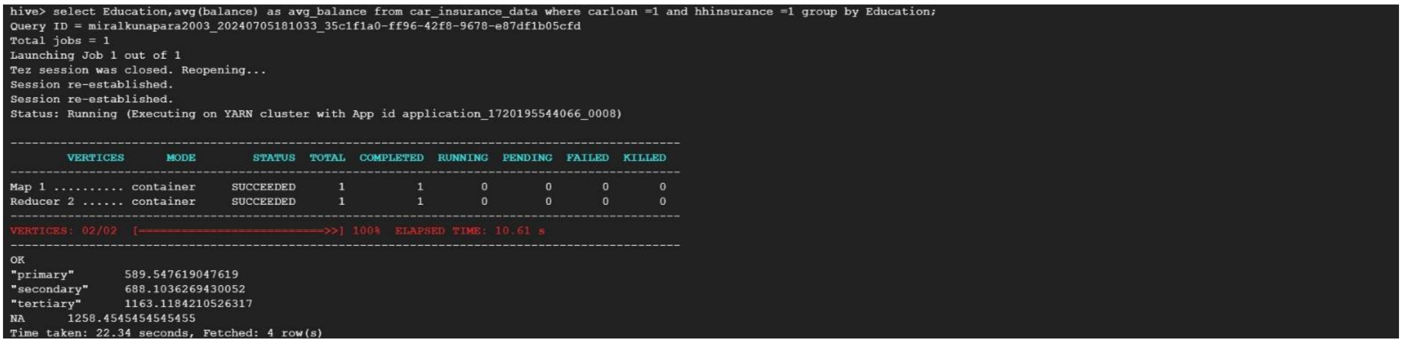
6. Determine the month with highest total ‘Balance’ of customers



```
hive> select lastcontactmonth ,sum(balance) as total_balance from car_insurance_data group by lastcontactmonth order by total_balance desc limit 1;
Query ID = miralkunapara2003_20240705175215_022b1e0b-55de-41f8-a9c5-4102e788e122
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.54 s
----------------------------------------------------------------------------------------------
OK
NULL    6131749
Time taken: 9.266 seconds, Fetched: 1 row(s)
hive> 
```

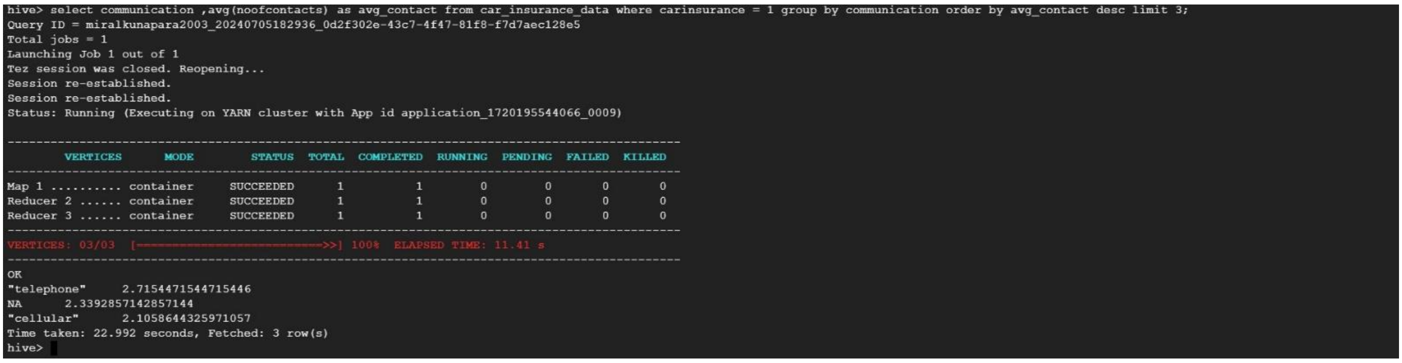
Problem 8 : Complex joins and aggregations

1. For customer who have both carloan and home insurance ,find out the averange ‘balance’ for each ‘education’ level.



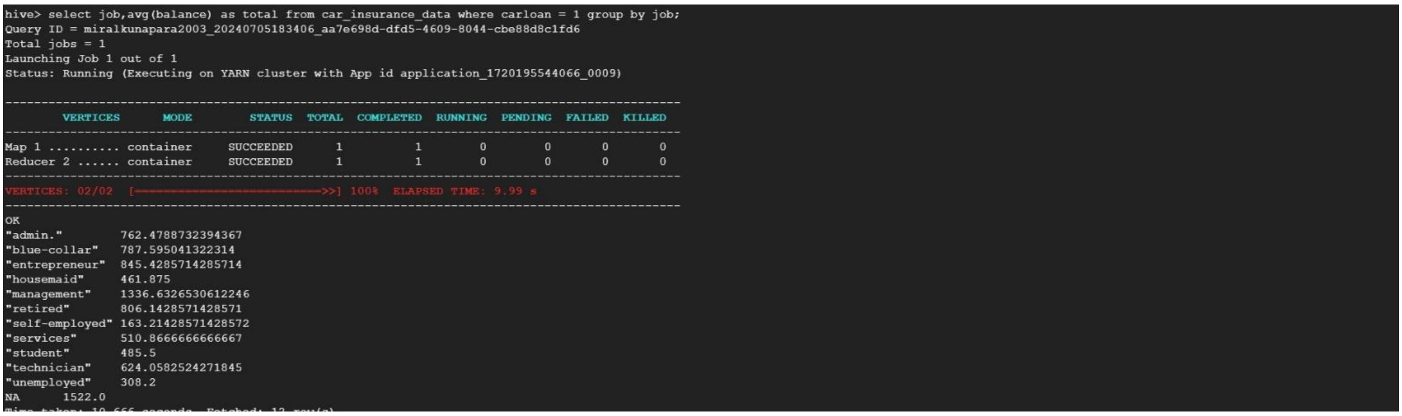
```
hive> select Education,avg(balance) as avg_balance from car_insurance_data where carloan =1 and hhinsurance =1 group by Education;
Query ID = miralkunapara2003_20240705181033_35c1f1a0-ff96-42f8-9678-e87df1b05cfd
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.61 s
----------------------------------------------------------------------------------------------
OK
"primary"       589.547619047619
"secondary"     688.1036269430052
"tertiary"      1163.1184210526317
NA      1258.4545454545455
Time taken: 22.34 seconds, Fetched: 4 row(s)
```

2. Identify the top 3 ‘Communication’ types for customers with ‘carinsurance’ ,and display their average ‘noofcontacts’.



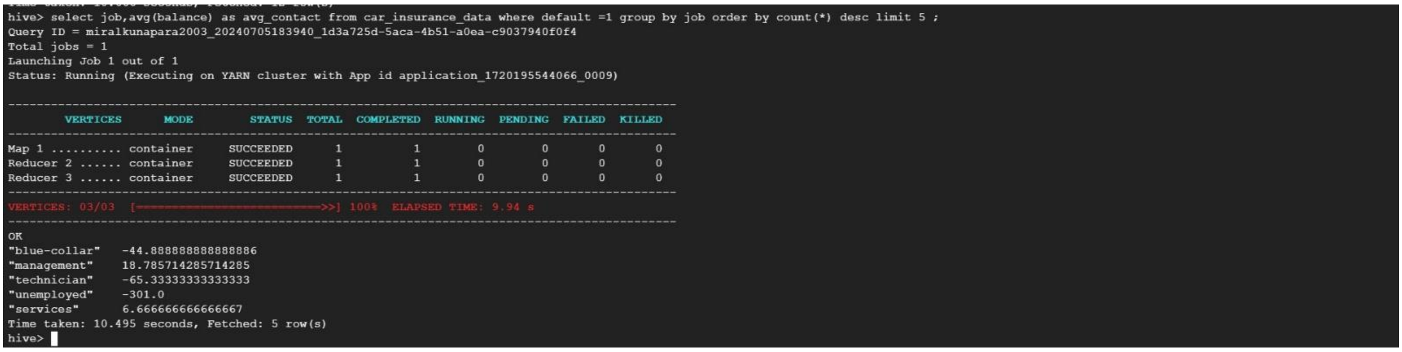
```
hive> select communication ,avg(noofcontacts) as avg_contact from car_insurance_data where carinsurance = 1 group by communication order by avg_contact desc limit 3;
Query ID = miralkunapara2003_20240705182936_0d2f302e-43c7-4f47-81f8-f7d7aec128e5
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720195544066_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.41 s
----------------------------------------------------------------------------------------------
OK
"telephone"     2.7154471544715446
NA      2.3392857142857144
"cellular"      2.1058644325971057
Time taken: 22.992 seconds, Fetched: 3 row(s)
hive> 
```

3. For customers who have a car loan ,calculate the average balance for each job category.



```
hive> select job,avg(balance) as total from car_insurance_data where carloan = 1 group by job;
Query ID = miralkunapara2003_20240705183406_aa7e698d-dfd5-4609-8044-cbe88d8c1fd6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.99 s
----------------------------------------------------------------------------------------------
OK
"admin."        762.4788732394367
"blue-collar"   787.595041322314
"entrepreneur"  845.4285714285714
"housemaid"     461.875
"management"    1336.6326530612246
"retired"       806.1428571428571
"self-employed" 163.21428571428572
"services"      510.8666666666667
"student"       485.5
"technician"    624.0582524271845
"unemployed"    308.2
NA      1522.0
Time taken: 10.666 seconds, Fetched: 12 row(s)
```

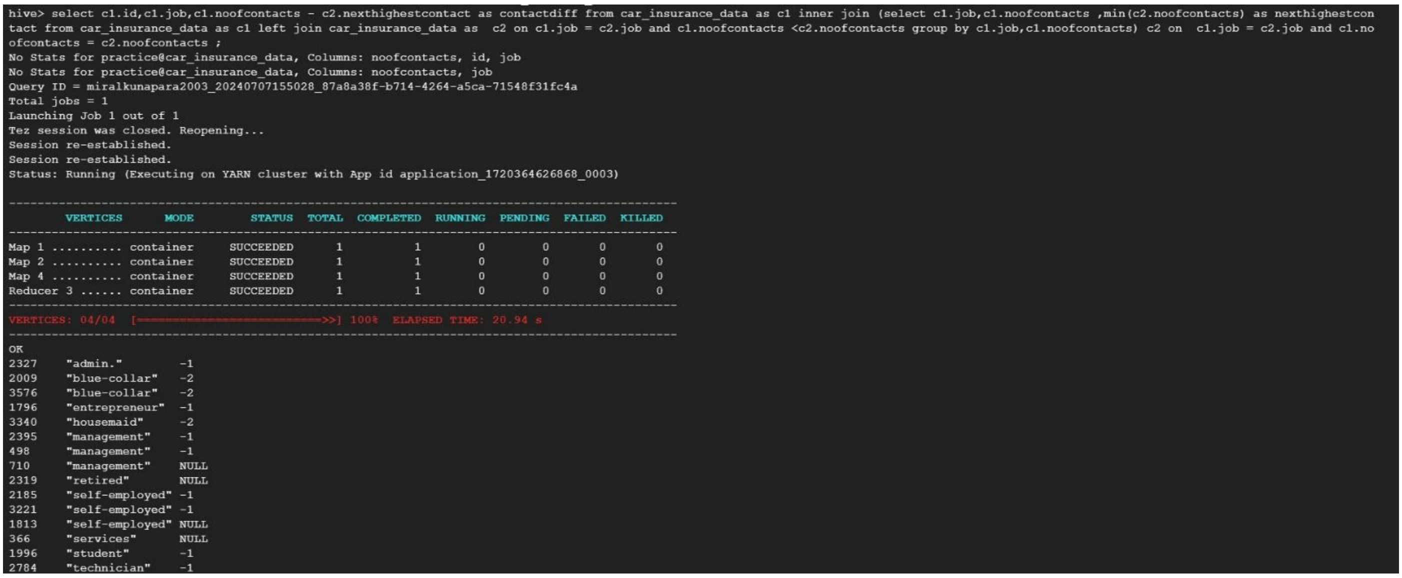
4. Identify the top 5 categories that have the most customers with a ‘default’,and show their average ‘balance’.



```
hive> select job,avg(balance) as avg_contact from car_insurance_data where default =1 group by job order by count(*) desc limit 5 ;
Query ID = miralkunapara2003_20240705183940_1d3a725d-5aca-4b51-a0ea-c9037940f0f4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720195544066_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 9.94 s
----------------------------------------------------------------------------------------------
OK
"blue-collar"   -44.888888888888886
"management"    18.785714285714285
"technician"    -65.33333333333333
"unemployed"    -301.0
"services"      6.666666666666667
Time taken: 10.495 seconds, Fetched: 5 row(s)
hive> █
```

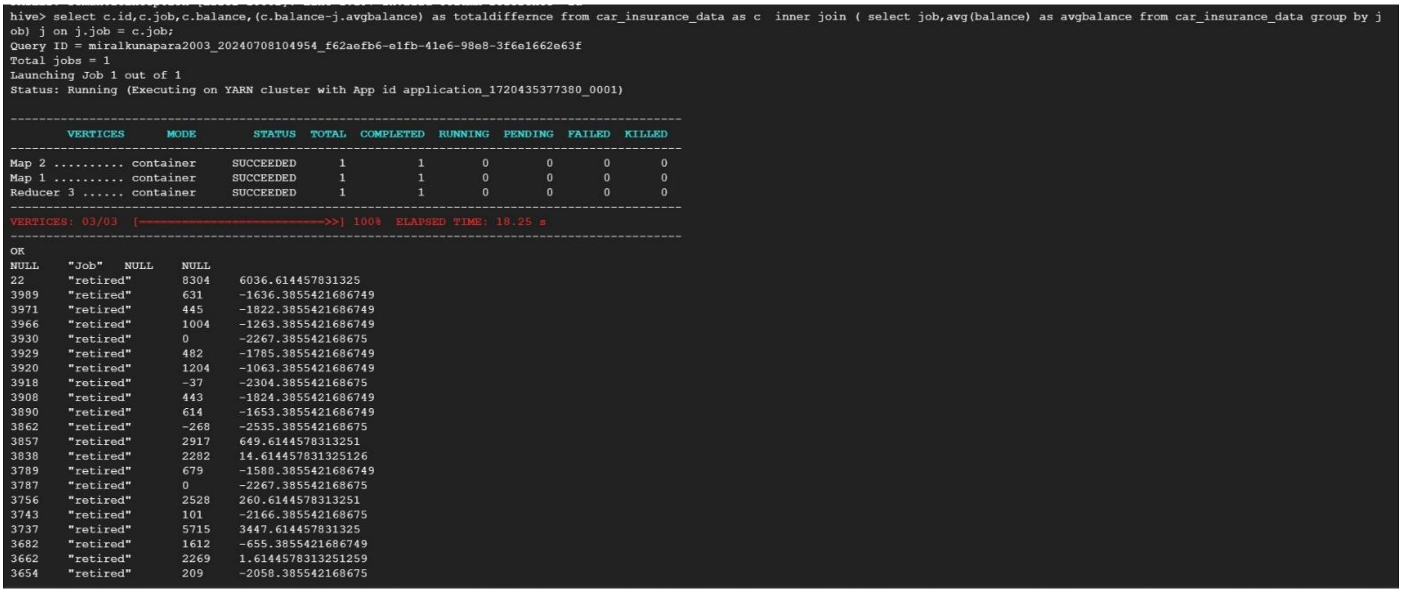
Problem 9 : Advanced window functions

1.Calculate the difference in ‘noofcontacts’ between each customer and the customer with the next highest number of contacts in the same ‘Job’ category.



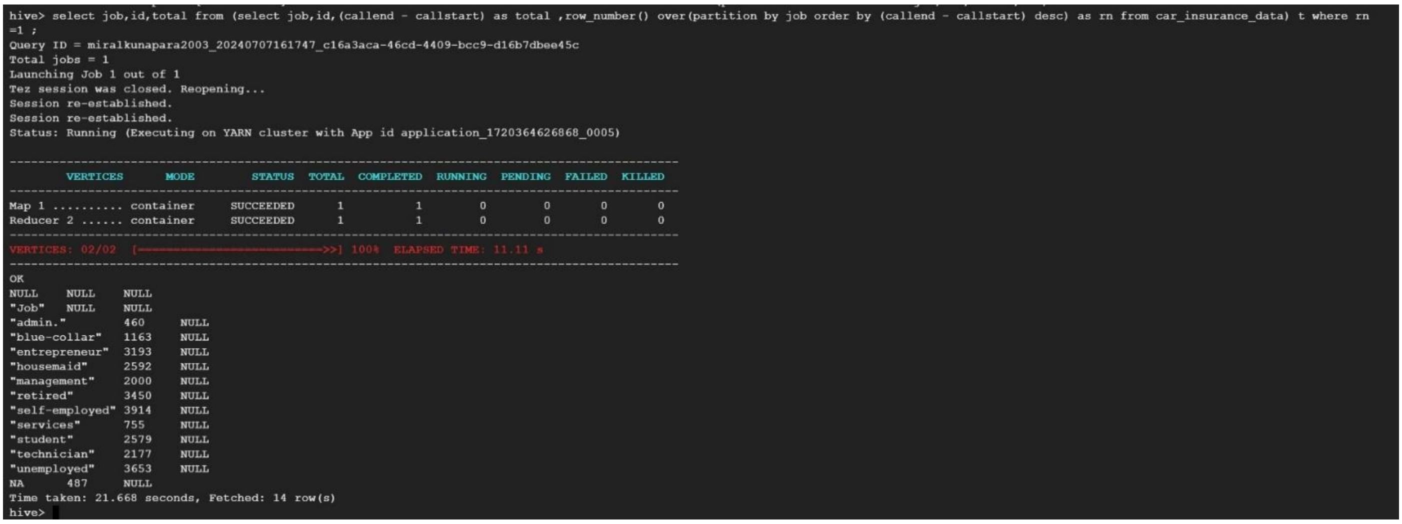
```
hive> select c1.id,c1.job,c1.noofcontacts - c2.nexthighestcontact as contactdiff from car_insurance_data as c1 inner join (select c1.job,c1.noofcontacts ,min(c2.noofcontacts) as nexthighestcon
tact from car_insurance_data as c1 left join car_insurance_data as  c2 on c1.job = c2.job and c1.noofcontacts <c2.noofcontacts group by c1.job,c1.noofcontacts) c2 on  c1.job = c2.job and c1.no
ofcontacts = c2.noofcontacts ;
No Stats for practice@car_insurance_data, Columns: noofcontacts, id, job
No Stats for practice@car_insurance_data, Columns: noofcontacts, job
Query ID = miralkunapara2003_20240707155028_87a8a38f-b714-4264-a5ca-71548f31fc4a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720364626868_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 20.94 s
----------------------------------------------------------------------------------------------
OK
2327    "admin."        -1
2009    "blue-collar"   -2
3576    "blue-collar"   -2
1796    "entrepreneur"  -1
3340    "housemaid"     -2
2395    "management"    -1
498     "management"    -1
710     "management"    NULL
2319    "retired"       NULL
2185    "self-employed" -1
3221    "self-employed" -1
1813    "self-employed" NULL
366     "services"      NULL
1996    "student"       -1
2784    "technician"    -1
```

2. For each customer ,calculate the difference between their ‘balance’ and the average ‘balance’of their ‘job’ category.



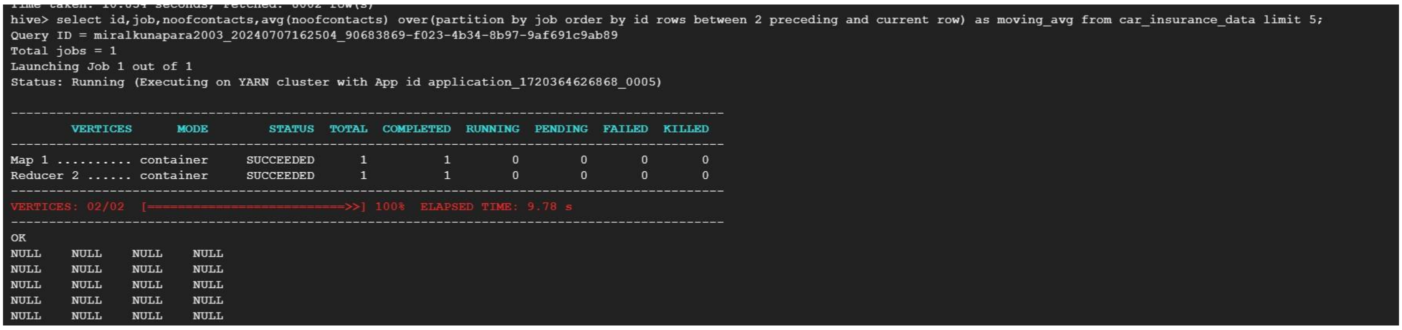
```
hive> select c.id,c.job,c.balance,(c.balance-j.avgbalance) as totaldiffernce from car_insurance_data as c  inner join ( select job,avg(balance) as avgbalance from car_insurance_data group by j
ob) j on j.job = c.job;
Query ID = miralkunapara2003_20240708104954_f62aefb6-e1fb-41e6-98e8-3f6e1662e63f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720435377380_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 18.25 s
----------------------------------------------------------------------------------------------
OK
NULL    "Job"   NULL    NULL
22      "retired"       8304    6036.614457831325
3989    "retired"       631     -1636.3855421686749
3971    "retired"       445     -1822.3855421686749
3966    "retired"       1004    -1263.3855421686749
3930    "retired"       0       -2267.385542168675
3929    "retired"       482     -1785.3855421686749
3920    "retired"       1204    -1063.3855421686749
3918    "retired"       -37     -2304.385542168675
3908    "retired"       443     -1824.3855421686749
3890    "retired"       614     -1653.3855421686749
3862    "retired"       -268    -2535.385542168675
3857    "retired"       2917    649.6144578313251
3838    "retired"       2282    14.614457831325126
3789    "retired"       679     -1588.3855421686749
3787    "retired"       0       -2267.385542168675
3756    "retired"       2528    260.6144578313251
3743    "retired"       101     -2166.385542168675
3737    "retired"       5715    3447.614457831325
3682    "retired"       1612    -655.3855421686749
3662    "retired"       2269    1.6144578313251259
3654    "retired"       209     -2058.385542168675
```

3. For each ‘job’ category,find the customer who had the longest call duration .



```
hive> select job,id,total from (select job,id,(callend - callstart) as total ,row_number() over(partition by job order by (callend - callstart) desc) as rn from car_insurance_data) t where rn
=1 ;
Query ID = miralkunapara2003_20240707161747_c16a3aca-46cd-4409-bcc9-d16b7dbee45c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720364626868_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.11 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL
"Job"   NULL    NULL
"admin."        460     NULL
"blue-collar"   1163    NULL
"entrepreneur"  3193    NULL
"housemaid"     2592    NULL
"management"    2000    NULL
"retired"       3450    NULL
"self-employed" 3914    NULL
"services"      755     NULL
"student"       2579    NULL
"technician"    2177    NULL
"unemployed"    3653    NULL
NA      487     NULL
Time taken: 21.668 seconds, Fetched: 14 row(s)
hive>
```

4. Calculate the moving average of ‘noofcontact’ within each ‘job’ category,using a window frame of the current row and the two preceding rows.

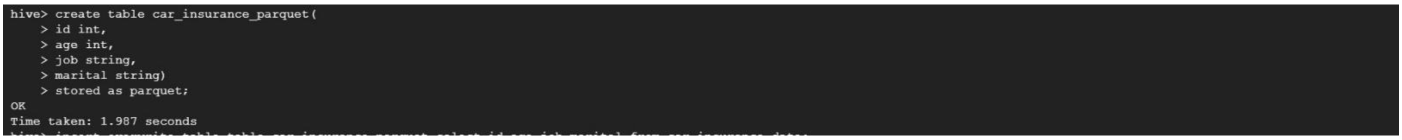


```
hive> select id,job,noofcontacts,avg(noofcontacts) over(partition by job order by id rows between 2 preceding and current row) as moving_avg from car_insurance_data limit 5;
Query ID = miralkunapara2003_20240707162504_90683869-f023-4b34-8b97-9af691c9ab89
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720364626868_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.78 s
----------------------------------------------------------------------------------------------
OK
NULL    NULL    NULL    NULL
NULL    NULL    NULL    NULL
NULL    NULL    NULL    NULL
NULL    NULL    NULL    NULL
NULL    NULL    NULL    NULL
```

Problem 10 : Performance Tuning

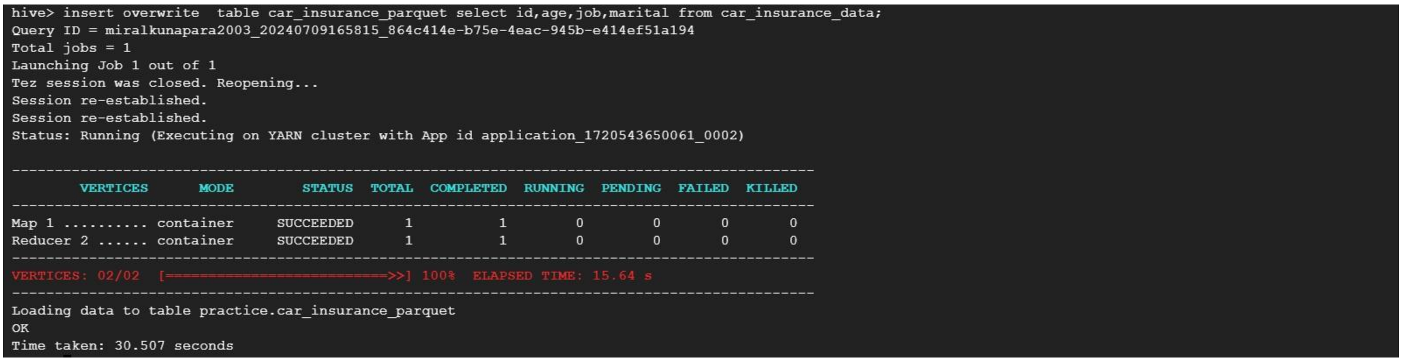
1. Experiment with different file formats(like orc,parquet) and measure their impact on the performance of your Hive queries.

- Create parquet table :



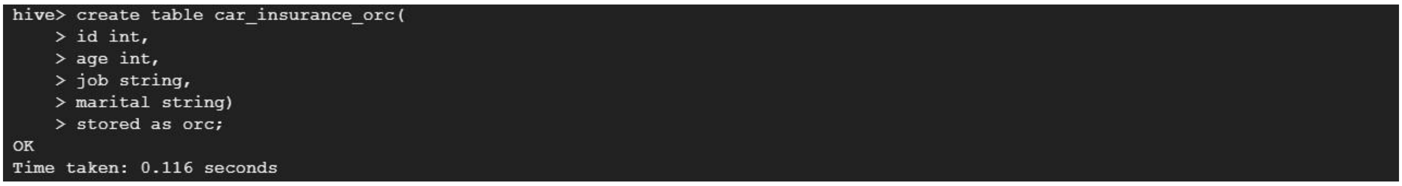
```
hive> create table car_insurance_parquet(
    > id int,
    > age int,
    > job string,
    > marital string)
    > stored as parquet;
OK
Time taken: 1.987 seconds
```

- Loading the data into this table :



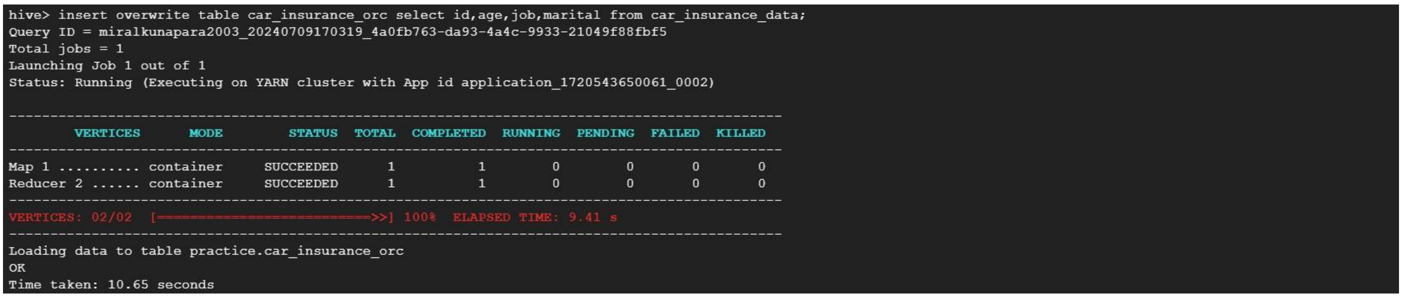
```
hive> insert overwrite  table car_insurance_parquet select id,age,job,marital from car_insurance_data;
Query ID = miralkunapara2003_20240709165815_864c414e-b75e-4eac-945b-e414ef51a194
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720543650061_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.64 s
----------------------------------------------------------------------------------------------
Loading data to table practice.car_insurance_parquet
OK
Time taken: 30.507 seconds
```

- Create orc table :



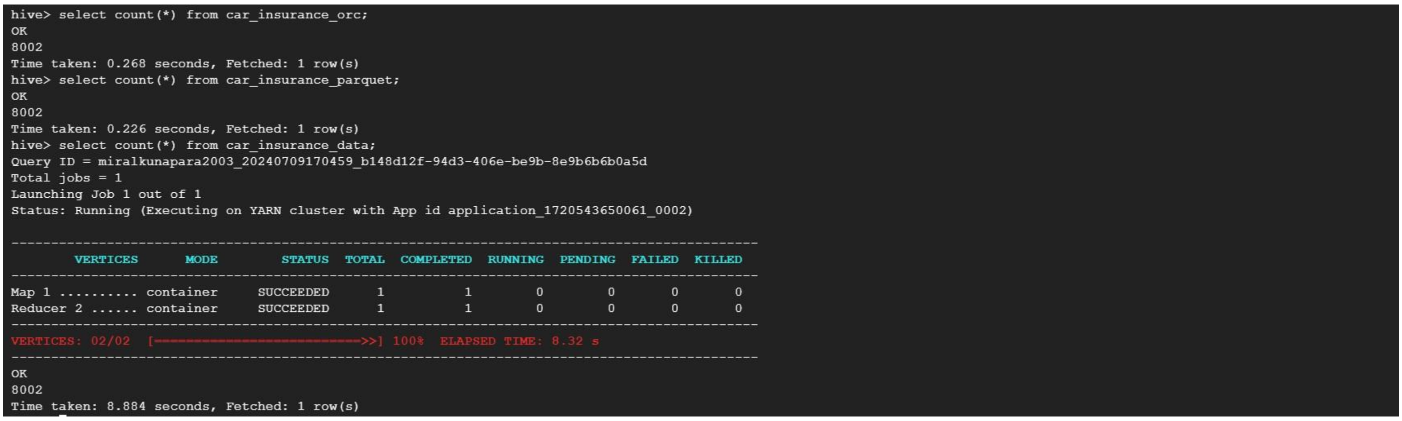
```
hive> create table car_insurance_orc(
    > id int,
    > age int,
    > job string,
    > marital string)
    > stored as orc;
OK
Time taken: 0.116 seconds
```

- Loading the data into this table :



```
hive> insert overwrite table car_insurance_orc select id,age,job,marital from car_insurance_data;
Query ID = miralkunapara2003_20240709170319_4a0fb763-da93-4a4c-9933-21049f88fbf5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720543650061_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.41 s
----------------------------------------------------------------------------------------------
Loading data to table practice.car_insurance_orc
OK
Time taken: 10.65 seconds
```

- ➢ Difference between quering all differnet file format and analyse it :



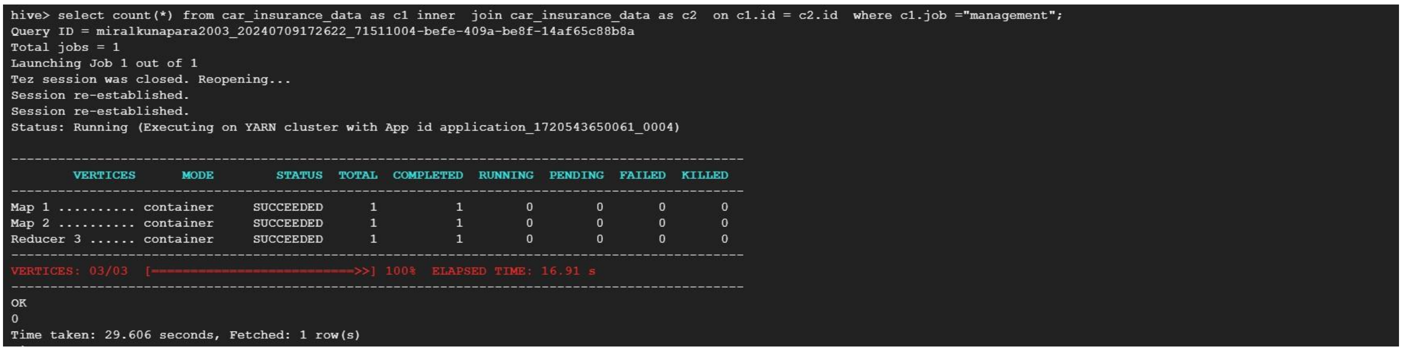
```
hive> select count(*) from car_insurance_orc;
OK
8002
Time taken: 0.268 seconds, Fetched: 1 row(s)
hive> select count(*) from car_insurance_parquet;
OK
8002
Time taken: 0.226 seconds, Fetched: 1 row(s)
hive> select count(*) from car_insurance_data;
Query ID = miralkunapara2003_20240709170459_b148d12f-94d3-406e-be9b-8e9b6b6b0a5d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720543650061_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.32 s
----------------------------------------------------------------------------------------------
OK
8002
Time taken: 8.884 seconds, Fetched: 1 row(s)
```

Conclusion :   running query on orc and parquet file format ‘s table is required less time compare to running query on sjoimple created csv table .

2. Use different levels of compression and observe their effects on storage and query performance.

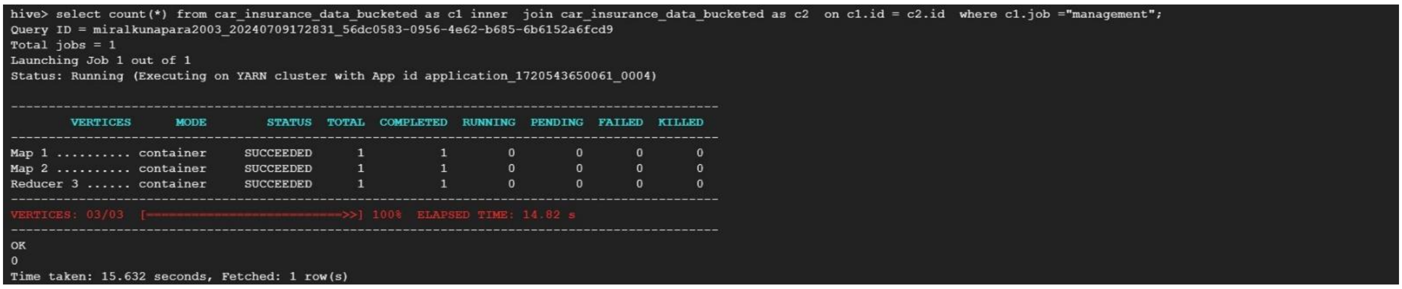
3. Compare the execution time of join queries with and with out bucketing.

- ➢ Joined tables without bucketings



```
hive> select count(*) from car_insurance_data as c1 inner  join car_insurance_data as c2  on c1.id = c2.id  where c1.job ="management";
Query ID = miralkunapara2003_20240709172622_71511004-befe-409a-be8f-14af65c88b8a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720543650061_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.91 s
----------------------------------------------------------------------------------------------
OK
0
Time taken: 29.606 seconds, Fetched: 1 row(s)
```

- ➢ Joined tables with bucketings

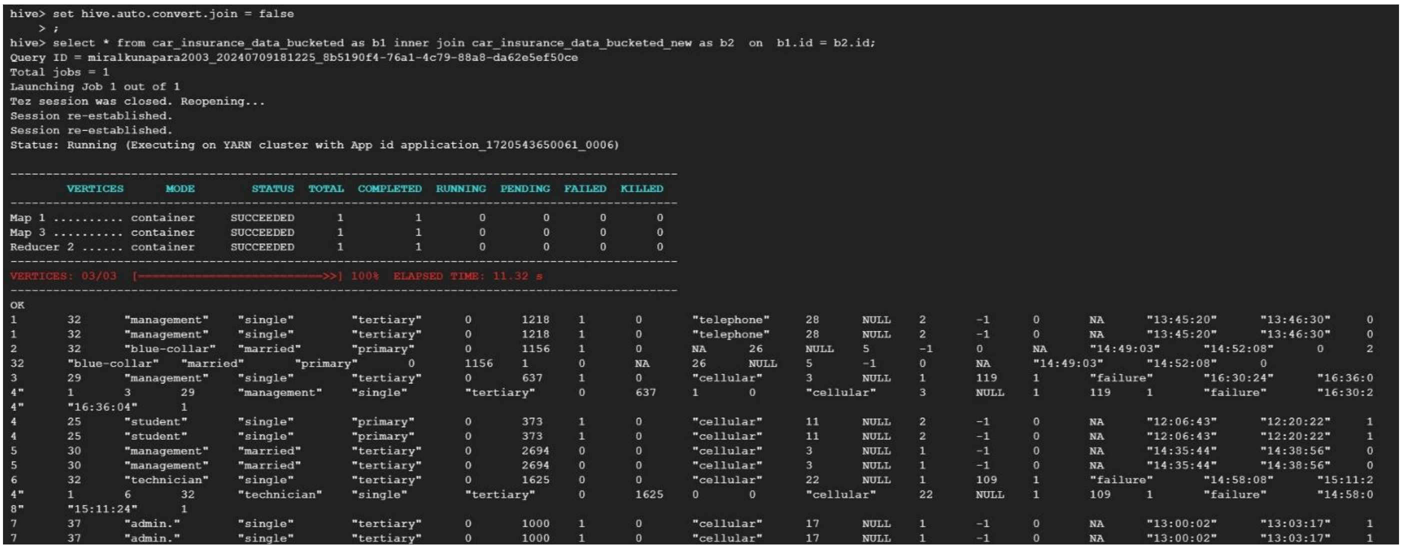


```
hive> select count(*) from car_insurance_data_bucketed as c1 inner  join car_insurance_data_bucketed as c2  on c1.id = c2.id  where c1.job ="management";
Query ID = miralkunapara2003_20240709172831_56dc0583-0956-4e62-b685-6b6152a6fcd9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720543650061_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.82 s
----------------------------------------------------------------------------------------------
OK
0
Time taken: 15.632 seconds, Fetched: 1 row(s)
```

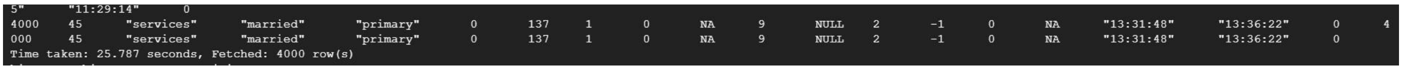
Conclusion : joining bucketed tables requires less time compare to unbucketed table .

4. Optimize your Hive queries using different Hive optimization techniques.

- Setting   hive.auto.convert.join = false

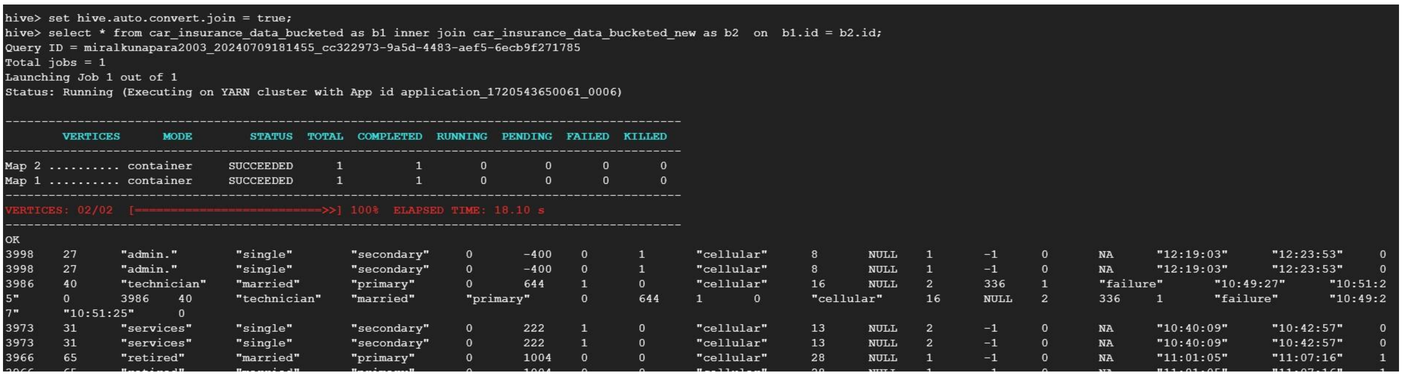


```
hive> set hive.auto.convert.join = false
    > ;
hive> select * from car_insurance_data_bucketed as b1 inner join car_insurance_data_bucketed_new as b2  on  b1.id = b2.id;
Query ID = miralkunapara2003_20240709181225_8b5190f4-76a1-4c79-88a8-da62e5ef50ce
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1720543650061_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.32 s
----------------------------------------------------------------------------------------------
OK
1       32      "management"    "single"        "tertiary"      0       1218    1       0       "telephone"     28      NULL    2       -1      0       NA      "13:45:20"      "13:46:30"      0
1       32      "management"    "single"        "tertiary"      0       1218    1       0       "telephone"     28      NULL    2       -1      0       NA      "13:45:20"      "13:46:30"      0
2       32      "blue-collar"   "married"       "primary"       0       1156    1       0       NA      26      NULL    5       -1      0       NA      "14:49:03"      "14:52:08"      0       2
32      "blue-collar"   "married"       "primary"       0       1156    1       0       NA      26      NULL    5       -1      0       NA      "14:49:03"      "14:52:08"      0
3       29      "management"    "single"        "tertiary"      0       637     1       0       "cellular"      3       NULL    1       119     1       "failure"       "16:30:24"      "16:36:0
4"      1       3       29      "management"    "single"        "tertiary"      0       637     1       0       "cellular"      3       NULL    1       119     1       "failure"       "16:30:2
4"      "16:36:04"      1
4       25      "student"       "single"        "primary"       0       373     1       0       "cellular"      11      NULL    2       -1      0       NA      "12:06:43"      "12:20:22"      1
4       25      "student"       "single"        "primary"       0       373     1       0       "cellular"      11      NULL    2       -1      0       NA      "12:06:43"      "12:20:22"      1
5       30      "management"    "married"       "tertiary"      0       2694    0       0       "cellular"      3       NULL    1       -1      0       NA      "14:35:44"      "14:38:56"      0
5       30      "management"    "married"       "tertiary"      0       2694    0       0       "cellular"      3       NULL    1       -1      0       NA      "14:35:44"      "14:38:56"      0
6       32      "technician"    "single"        "tertiary"      0       1625    0       0       "cellular"      22      NULL    1       109     1       "failure"       "14:58:08"      "15:11:2
4"      1       6       32      "technician"    "single"        "tertiary"      0       1625    0       0       "cellular"      22      NULL    1       109     1       "failure"       "14:58:0
8"      "15:11:24"      1
7       37      "admin."        "single"        "tertiary"      0       1000    1       0       "cellular"      17      NULL    1       -1      0       NA      "13:00:02"      "13:03:17"      1
7       37      "admin."        "single"        "tertiary"      0       1000    1       0       "cellular"      17      NULL    1       -1      0       NA      "13:00:02"      "13:03:17"      1
```

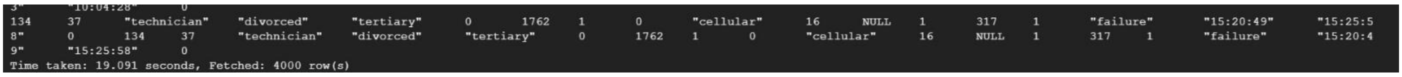


```
5"      "11:29:14"      0
4000    45      "services"      "married"       "primary"       0       137     1       0       NA      9       NULL    2       -1      0       NA      "13:31:48"      "13:36:22"      0       4
000     45      "services"      "married"       "primary"       0       137     1       0       NA      9       NULL    2       -1      0       NA      "13:31:48"      "13:36:22"      0
Time taken: 25.787 seconds, Fetched: 4000 row(s)
```

- Setting hive.auto.convert.join = true for map join



```
hive> set hive.auto.convert.join = true;
hive> select * from car_insurance_data_bucketed as b1 inner join car_insurance_data_bucketed_new as b2  on  b1.id = b2.id;
Query ID = miralkunapara2003_20240709181455_cc322973-9a5d-4483-aef5-6ecb9f271785
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1720543650061_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.10 s
----------------------------------------------------------------------------------------------
OK
3998    27      "admin."        "single"        "secondary"     0       -400    0       1       "cellular"      8       NULL    1       -1      0       NA      "12:19:03"      "12:23:53"      0
3998    27      "admin."        "single"        "secondary"     0       -400    0       1       "cellular"      8       NULL    1       -1      0       NA      "12:19:03"      "12:23:53"      0
3986    40      "technician"    "married"       "primary"       0       644     1       0       "cellular"      16      NULL    2       336     1       "failure"       "10:49:27"      "10:51:2
5"      0       3986    40      "technician"    "married"       "primary"       0       644     1       0       "cellular"      16      NULL    2       336     1       "failure"       "10:49:2
7"      "10:51:25"      0
3973    31      "services"      "single"        "secondary"     0       222     1       0       "cellular"      13      NULL    2       -1      0       NA      "10:40:09"      "10:42:57"      0
3973    31      "services"      "single"        "secondary"     0       222     1       0       "cellular"      13      NULL    2       -1      0       NA      "10:40:09"      "10:42:57"      0
3966    65      "retired"       "married"       "primary"       0       1004    0       0       "cellular"      28      NULL    1       -1      0       NA      "11:01:05"      "11:07:16"      1
3966    65      "retired"       "married"       "primary"       0       1004    0       0       "cellular"      28      NULL    1       -1      0       NA      "11:01:05"      "11:07:16"      1
```



```
3"      "10:04:28"      0
134     37      "technician"    "divorced"      "tertiary"      0       1762    1       0       "cellular"      16      NULL    1       317     1       "failure"       "15:20:49"      "15:25:5
8"      0       134     37      "technician"    "divorced"      "tertiary"      0       1762    1       0       "cellular"      16      NULL    1       317     1       "failure"       "15:20:4
9"      "15:25:58"      0
Time taken: 19.091 seconds, Fetched: 4000 row(s)
```