Hive Certification Project Report –Edureka

Manoj K V

Index:

Problem Statement

1. Find the list of people with grade "B" who have taken loan.

2. Find the list of people having interest more than 1000.

3. Find the list of people having loan amount more than 1000.

4. Get the highest loan amount given to grade users (A-G).

5. Highest loan amount given in that year with that Employee id and Employees annual income.

6. Get the total number of loans with loan id and load amount which are having loan status as Late.

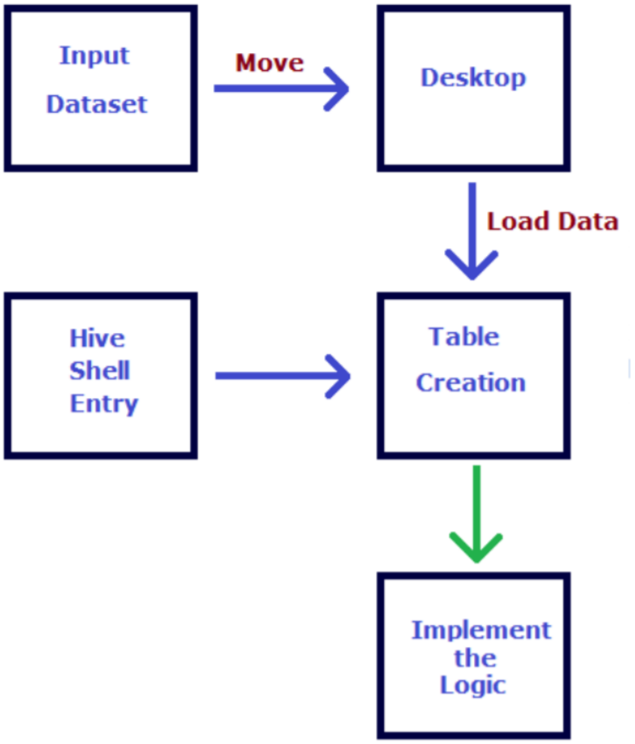
7. Average loan interest rate with 60-month term and 36-month term.

Technology/Software Used:

- Hadoop environment
  (HDP sandbox)

- Apache Hive
  Hive 1.2.1000.2.4.2.57-1

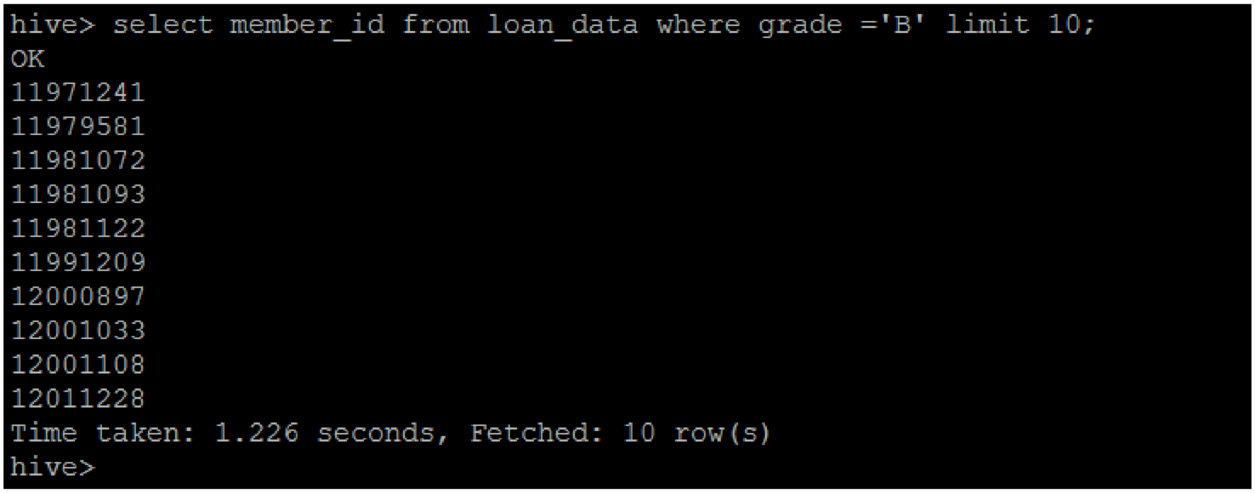
Solution Flow Diagram



Solution:

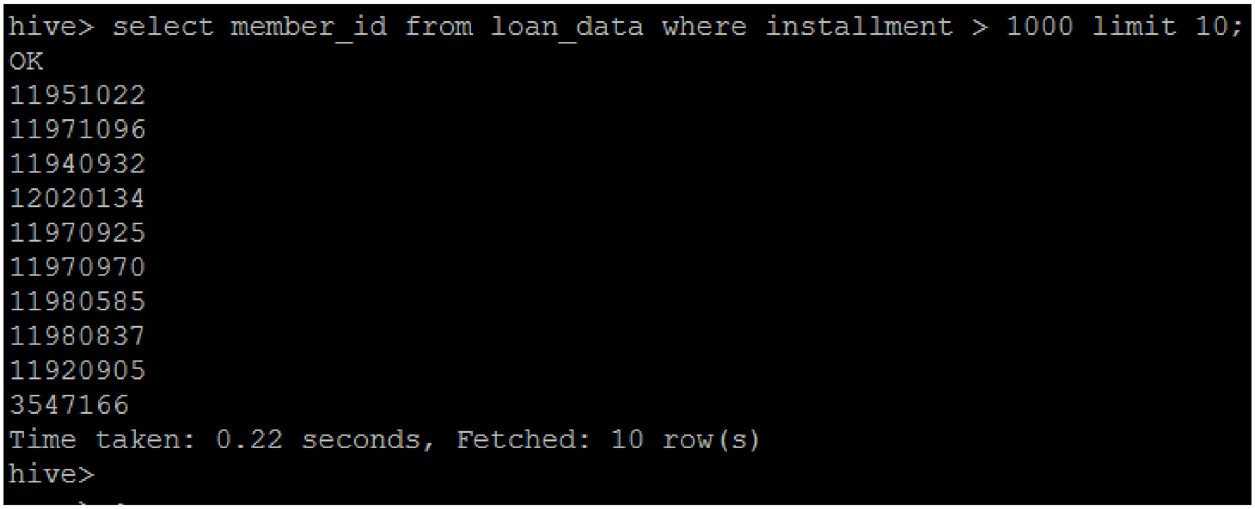
--1. Find the list of people with grade "B" who have taken loan.

```
select member_id from loan_data where grade ='B';
```



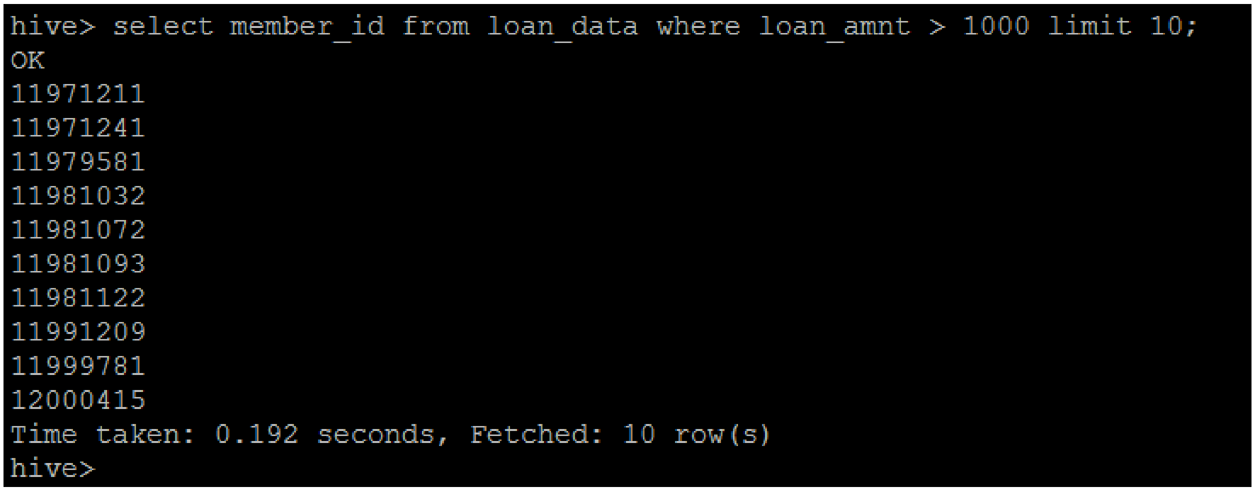
```
hive> select member_id from loan_data where grade ='B' limit 10;
OK
11971241
11979581
11981072
11981093
11981122
11991209
12000897
12001033
12001108
12011228
Time taken: 1.226 seconds, Fetched: 10 row(s)
hive>
```

```
--2. Find the list of people having interest more than 1000.

select member_id from loan_data where installment > 1000;
```



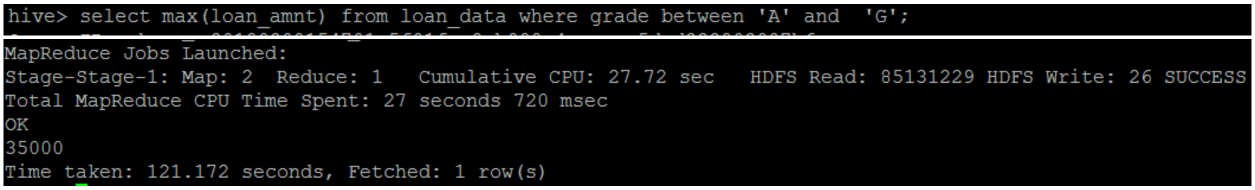
```
hive> select member_id from loan_data where installment > 1000 limit 10;
OK
11951022
11971096
11940932
12020134
11970925
11970970
11980585
11980837
11920905
3547166
Time taken: 0.22 seconds, Fetched: 10 row(s)
hive>
```

```
--3. Find the list of people having loan amount more than 1000.

select member_id from loan_data where loan_amnt > 1000;
```

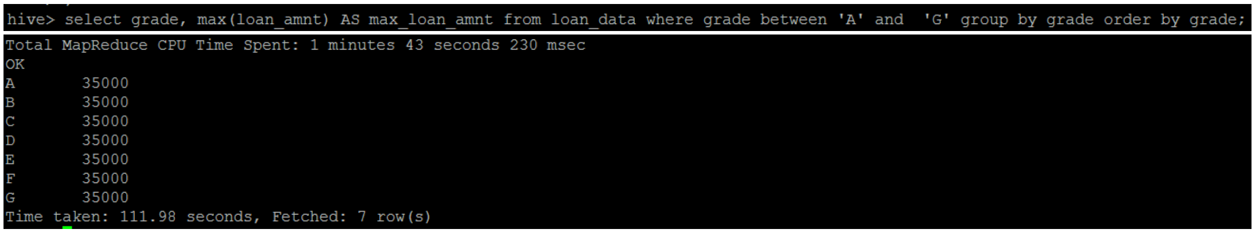


```
hive> select member_id from loan_data where loan_amnt > 1000 limit 10;
OK
11971211
11971241
11979581
11981032
11981072
11981093
11981122
11991209
11999781
12000415
Time taken: 0.192 seconds, Fetched: 10 row(s)
hive>
```

```
--4. Get the highest loan amount given to grade users (A-G).

select max(loan_amnt) from loan_data where grade between 'A' and  'G';

select grade, max(loan_amnt) AS max_loan_amnt from loan_data where grade
between 'A' and  'G' group by grade order by grade;
```

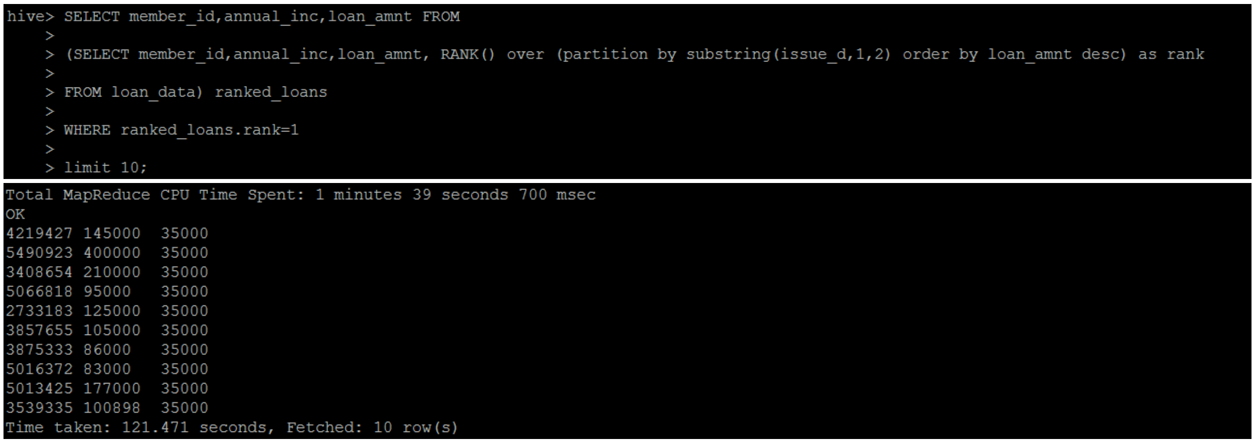


```
hive> select max(loan_amnt) from loan_data where grade between 'A' and  'G';
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 27.72 sec   HDFS Read: 85131229 HDFS Write: 26 SUCCESS
Total MapReduce CPU Time Spent: 27 seconds 720 msec
OK
35000
Time taken: 121.172 seconds, Fetched: 1 row(s)
```



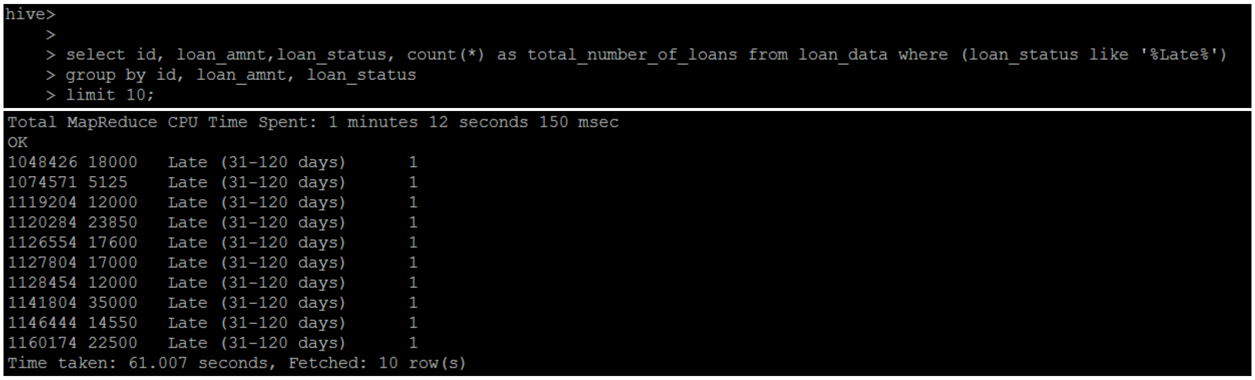
```
hive> select grade, max(loan_amnt) AS max_loan_amnt from loan_data where grade between 'A' and  'G' group by grade order by grade;
Total MapReduce CPU Time Spent: 1 minutes 43 seconds 230 msec
OK
A       35000
B       35000
C       35000
D       35000
E       35000
F       35000
G       35000
Time taken: 111.98 seconds, Fetched: 7 row(s)
```

```
--5. Highest loan amount given in that year with that Employee id and
Employees annual income.

SELECT member_id,annual_inc,loan_amnt FROM

(SELECT member_id,annual_inc,loan_amnt, RANK() over (partition by
substring(issue_d,1,2) order by loan_amnt desc) as rank

FROM loan_data) ranked_loans

WHERE ranked_loans.rank=1
;
```



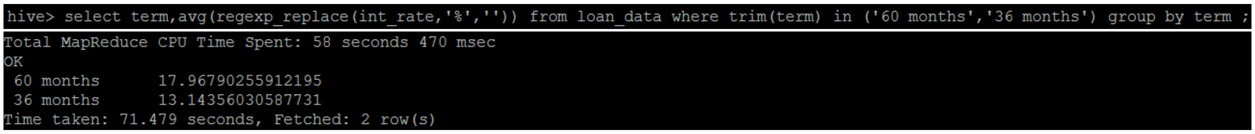
```
hive> SELECT member_id,annual_inc,loan_amnt FROM
    >
    > (SELECT member_id,annual_inc,loan_amnt, RANK() over (partition by substring(issue_d,1,2) order by loan_amnt desc) as rank
    >
    > FROM loan_data) ranked_loans
    >
    > WHERE ranked_loans.rank=1
    >
    > limit 10;
Total MapReduce CPU Time Spent: 1 minutes 39 seconds 700 msec
OK
4219427 145000  35000
5490923 400000  35000
3408654 210000  35000
5066818 95000   35000
2733183 125000  35000
3857655 105000  35000
3875333 86000   35000
5016372 83000   35000
5013425 177000  35000
3539335 100898  35000
Time taken: 121.471 seconds, Fetched: 10 row(s)
```

```
--6. Get the total number of loans with loan id and loan amount which are
having loan status as Late.

select id, loan_amnt,loan_status, count(*) as total_number_of_loans from
loan_data where (loan_status like '%Late%') group by id, loan_amnt,
loan_status;
```



```
hive>
    >
    > select id, loan_amnt,loan_status, count(*) as total_number_of_loans from loan_data where (loan_status like '%Late%')
    > group by id, loan_amnt, loan_status
    > limit 10;
Total MapReduce CPU Time Spent: 1 minutes 12 seconds 150 msec
OK
1048426 18000   Late (31-120 days)      1
1074571 5125    Late (31-120 days)      1
1119204 12000   Late (31-120 days)      1
1120284 23850   Late (31-120 days)      1
1126554 17600   Late (31-120 days)      1
1127804 17000   Late (31-120 days)      1
1128454 12000   Late (31-120 days)      1
1141804 35000   Late (31-120 days)      1
1146444 14550   Late (31-120 days)      1
1160174 22500   Late (31-120 days)      1
Time taken: 61.007 seconds, Fetched: 10 row(s)
```

```
--7. Average loan interest rate with 60-month term and 36-month term.

select term,avg(regexp_replace(int_rate,'%','')) from loan_data where
trim(term) in ('60 months','36 months') group by term ;
```



```
hive> select term,avg(regexp_replace(int_rate,'%','')) from loan_data where trim(term) in ('60 months','36 months') group by term ;
Total MapReduce CPU Time Spent: 58 seconds 470 msec
OK
 60 months      17.96790255912195
 36 months      13.14356030587731
Time taken: 71.479 seconds, Fetched: 2 row(s)
```

Thank you!