Hive Assignment – NYC Parking Violations

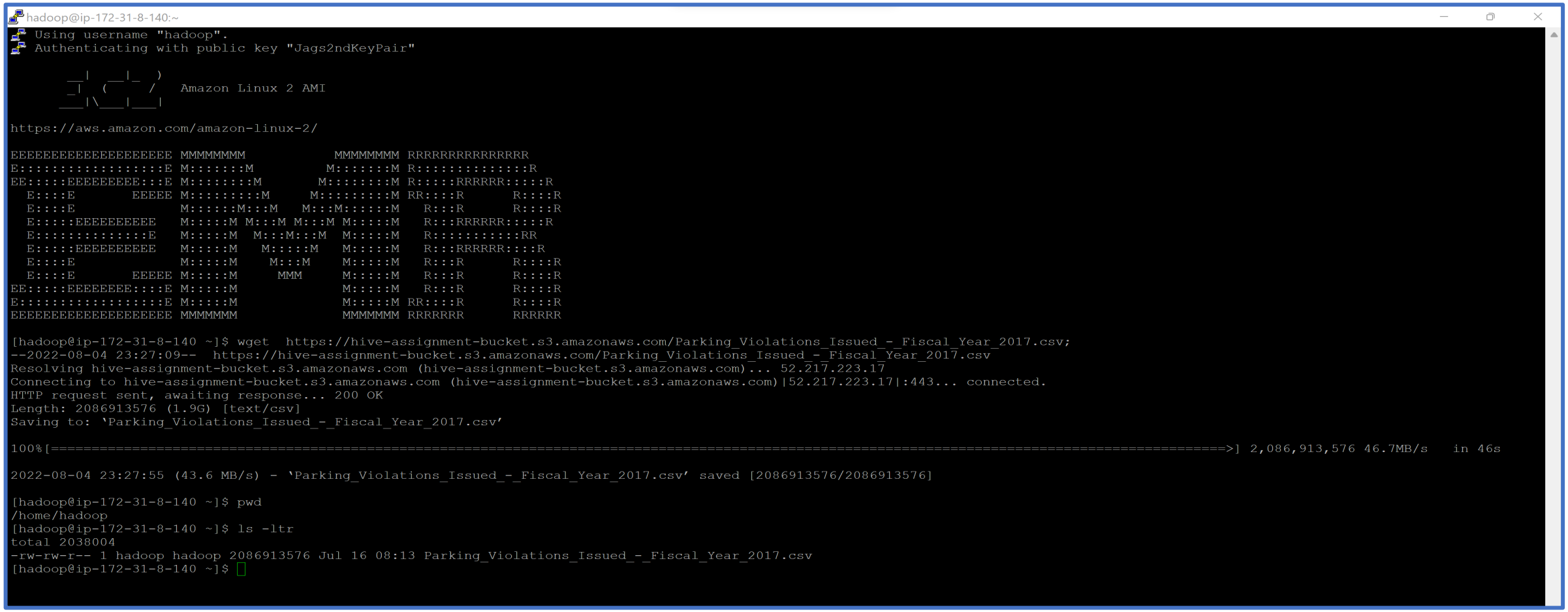
- Student : Jagadish Janakiraman
- Submission Date : 8/7/2022

Load Data to EMR

Command –

wget https://hive-assignment-bucket.s3.amazonaws.com/Parking_Violations_Issued_-_Fiscal_Year_2017.csv;

Output Screenshot -



Load Data to Table

Commands –

1. CREATE TABLE IF NOT EXISTS parkingviolations (SummonsNumber bigint, PlateID String, RegistrationState String,PlateType String, IssueDate String,ViolationCode int, VehicleBodyType String,VehicleMake String, IssuingAgency String,StreetCode1 int , StreetCode2 int ,StreetCode3 int , VehicleExpirationDate int ,ViolationLocation String, ViolationPrecinct int ,IssuerPrecinct int , IssuerCode int ,IssuerCommand String, IssuerSquad String,ViolationTime String, TimeFirstObserved String,ViolationCounty String, ViolationInFrontOfOrOpposite String,HouseNumber String, StreetName String,IntersectingStreet String, DateFirstObserved int ,LawSection int , SubDivision String,ViolationLegalCode String, DaysParkingInEffect String ,FromHoursInEffect String, ToHoursInEffect String,VehicleColor String, UnregisteredVehicle String,VehicleYear int , MeterNumber String,FeetFromCurb int , ViolationPostCode String,ViolationDescription String, NoStandingorStoppingViolation String,HydrantViolation String, DoubleParkingViolation String) COMMENT 'parkingviolations assignment' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' tblproperties('skip.header.line.count'='1’);

2. load data local inpath '/home/hadoop/Parking_Violations_Issued_-_Fiscal_Year_2017.csv' into table parkingviolations ;

Output Screenshot -

Part-I: Examine the data

Q1.1 - Find the total number of tickets for the year.

Command –
select count(*) from parkingviolations where year(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd'))=2017;
Answer – There are 5,431,903 parking violations in the year 2017

Output Screenshot -

```
hive> select count(*) from parkingviolations where  year(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd'))=2017;
Query ID = hadoop_20220807184700_15002f9f-5112-4201-bad4-f64a879cbf1c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659897309971_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: 49.80 s
--------------------------------------------------------------------------------
OK
5431903
Time taken: 52.148 seconds, Fetched: 1 row(s)
hive>
```

Q1.2 - Find out the total number of states to which the cars with tickets belong.

Command - select count(distinct(RegistrationState)) from parkingviolations where issuedate like '%2017' and RegistrationState rlike '^([A-Z])';

Answer – Cars belong to 64 States with ticket

Output Screenshot -

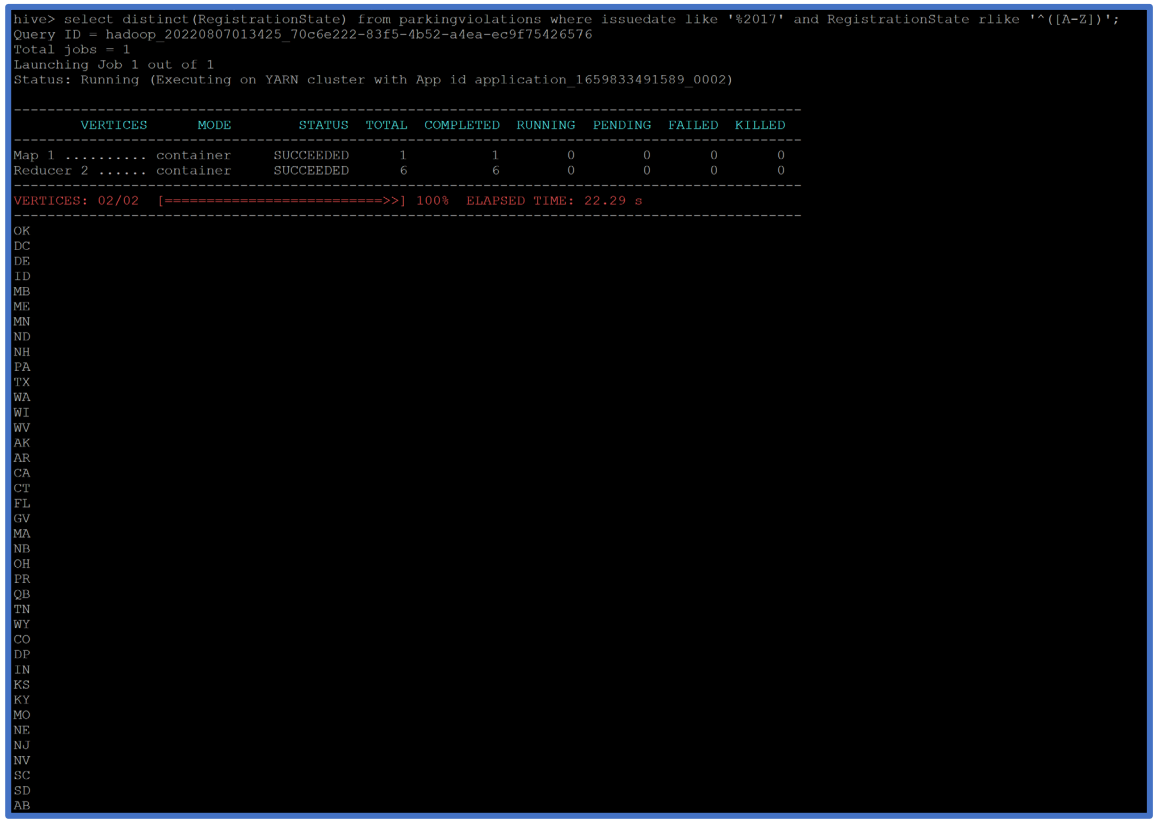
```
hive> select count(distinct(RegistrationState)) from parkingviolations where issuedate like '%2017' and RegistrationState rlike '^([A-Z])';
Query ID = hadoop_20220807013239_3ba0c0f0-900b-4314-8581-ba3d0d999205
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659833491589_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 22.33 s
----------------------------------------------------------------------------------------------
OK
64
Time taken: 22.879 seconds, Fetched: 1 row(s)
hive>
```

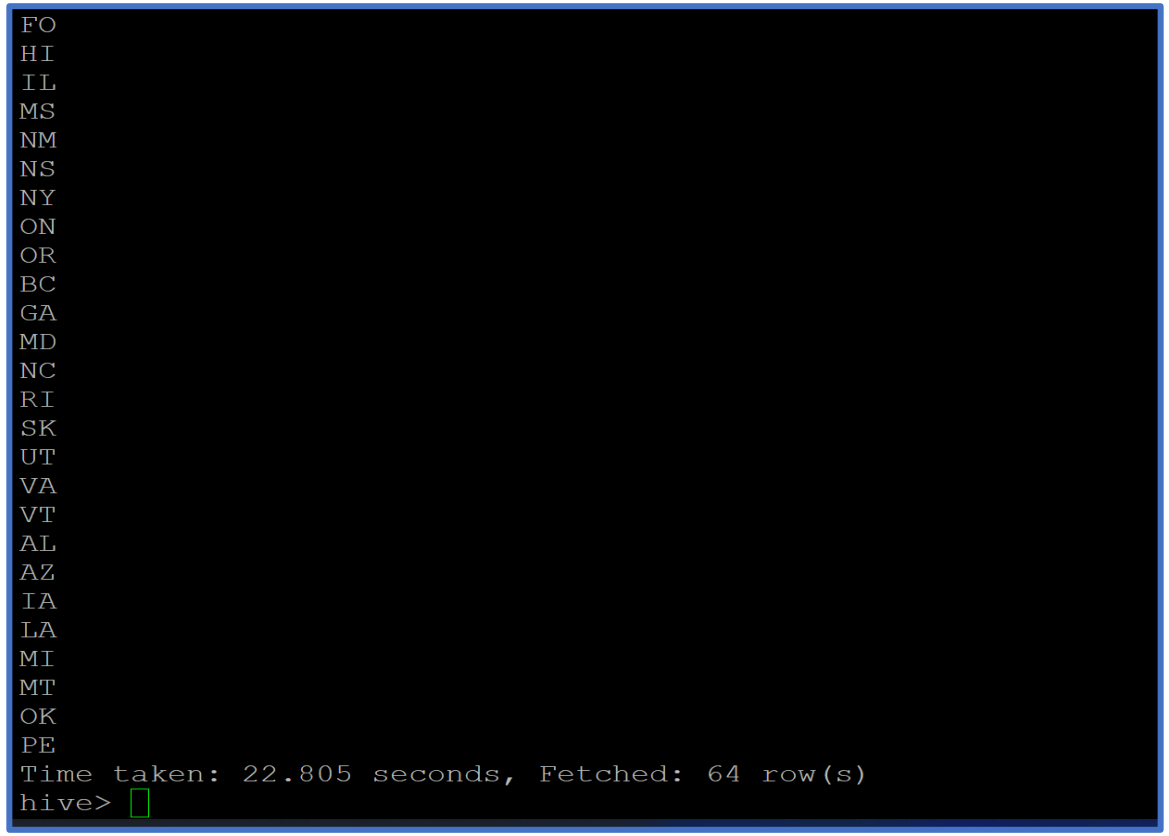
Q1.2 Optional Question – List of states with tickets

Command - select distinct(RegistrationState) from parkingviolations where issuedate like '%2017' and RegistrationState rlike '^([A-Z])';

Answer – Cars belong to 64 States with ticket

Output Screenshot -





Q1.3 – Find out the number of such tickets which have no addresses.

Command - select count(*) from parkingviolations where StreetCode1 is null or StreetCode2 is null or StreetCode1 is null;

Answer – 49 tickets have no addresses (either of StreetCode1, 2 and 3 is null)

Output Screenshot -

```
hive> select count(*) from parkingviolations where StreetCode1 is null or StreetCode2 is null or StreetCode1 is null;
Query ID = hadoop_20220805000247_d9df4b00-bef3-4612-8535-498ebf8beea3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659655376995_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: 19.51 s
--------------------------------------------------------------------------------
OK
49
Time taken: 20.087 seconds, Fetched: 1 row(s)
hive>
```

Part-II: Aggregation tasks

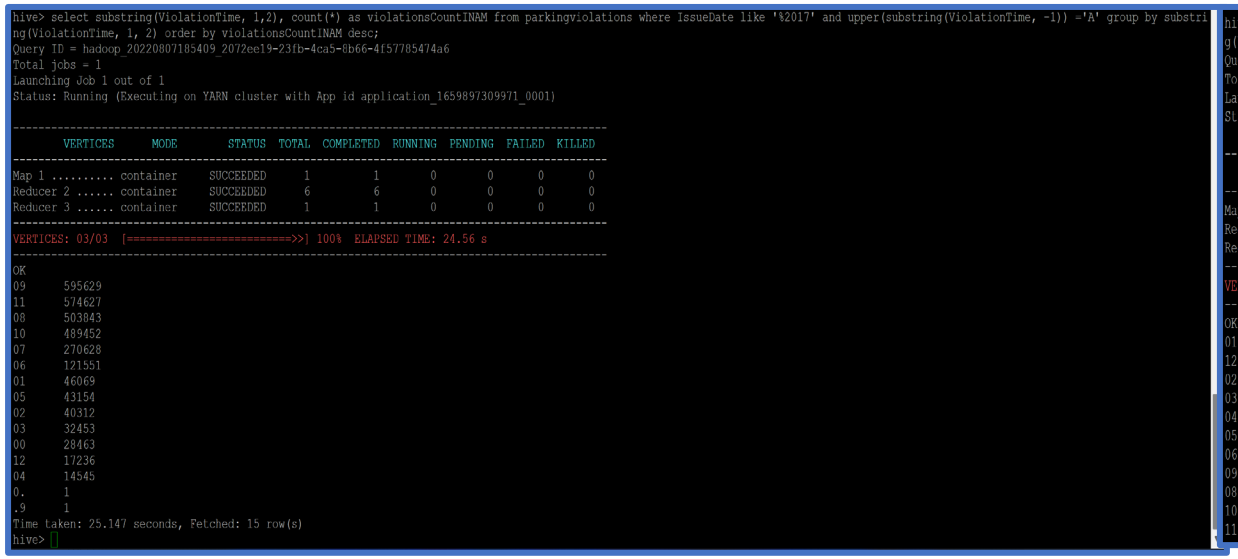
Q2.1 – Find out the frequency of parking violations across different times of the day

Commands –

- select substring(ViolationTime, 1,2), count(*) as violationsCountINAM from parkingviolations where IssueDate like '%2017' and upper(substring(ViolationTime, -1)) ='A' group by substring(ViolationTime, 1, 2) order by violationsCountINAM desc;
- select substring(ViolationTime, 1,2), count(*) as violationcountINPM from parkingviolations where IssueDate like '%2017' and upper(substring(ViolationTime, -1)) ='P' group by substring(ViolationTime, 1, 2) order by ViolationCountINPM desc;
- **Answer** – 9 AM and 1 PM are the hours with maximum parking violations.

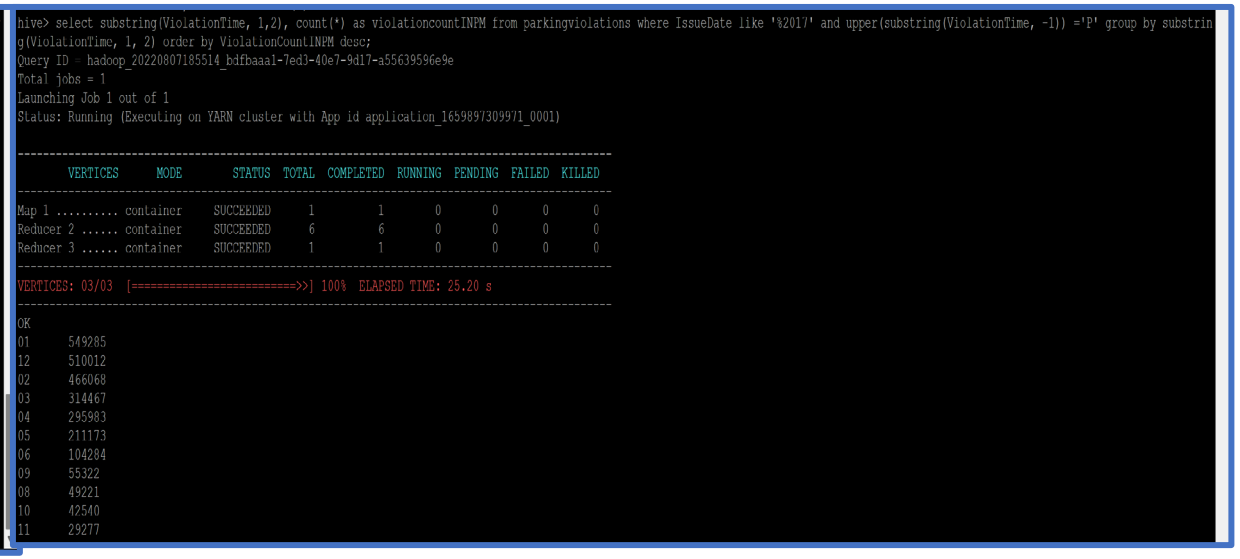
Output Screenshot -

Violations in AM (Morning)



```
hive> select substring(ViolationTime, 1,2), count(*) as violationsCountINAM from parkingviolations where IssueDate like '%2017' and upper(substring(ViolationTime, -1)) ='A' group by substri
ng(ViolationTime, 1, 2) order by violationsCountINAM desc;
Query ID = hadoop_20220807185409_2072ee19-23fb-4ca5-8b66-4f57785474a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659897309971_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 24.56 s
--------------------------------------------------------------------------------
OK
09      595629
11      574627
08      503843
10      489452
07      270628
06      121551
01      46069
05      43154
02      40312
03      32453
00      28463
12      17236
04      14545
0.      1
.9      1
Time taken: 25.147 seconds, Fetched: 15 row(s)
hive>
```

Violations in PM (Evening)



```
hive> select substring(ViolationTime, 1,2), count(*) as violationcountINPM from parkingviolations where IssueDate like '%2017' and upper(substring(ViolationTime, -1)) ='P' group by substrin
g(ViolationTime, 1, 2) order by ViolationCountINPM desc;
Query ID = hadoop_20220807185514_bdfbaaa1-7ed3-40e7-9d17-a55639596e9e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659897309971_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.20 s
--------------------------------------------------------------------------------
OK
01      549285
12      510012
02      466068
03      314467
04      295983
05      211173
06      104284
09      55322
08      49221
10      42540
11      29277
```

Q2.2 – Divide 24 hours into six equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the 3 most commonly occurring violations.

Command –

```
select * from (
select violationbin,violationcode,ViolationCount, dense_rank() over (partition by violationbin order by ViolationCount desc) as rank
from
( Select violationbin, ViolationCode, count(*) as ViolationCount from
( select case
when substring(violationtime,1,2) in ('00','12','01','02','03') and upper(substring(violationtime,-1))='A' then 'MidNight_12AM_3AM'
when substring(violationtime,1,2) in ('04','05','06','07') and upper(substring(violationtime,-1))='A' then 'EarlyMorning_4AM_7AM'
when substring(violationtime,1,2) in ('08','09','10','11') and upper(substring(violationtime,-1))='A' then 'Morning_8AM_11AM'
when substring(violationtime,1,2) in ('12','01','02','03') and upper(substring(violationtime,-1))='P' then 'AfterNoon_12PM_3PM'
when substring(violationtime,1,2) in ('04','05','06','07') and upper(substring(violationtime,-1))='P' then 'Evening_4PM_7PM'
when substring(violationtime,1,2) in ('08','09','10','11') and upper(substring(violationtime,-1))='P' then 'Night_8PM_11PM'
else null end as violationbin, ViolationCode from parkingviolations   where IssueDate like '%2017'
)temp1
where violationbin is not NULL group by violationbin,ViolationCode
) temp2
) temp3 where rank <= 3 ;
```

Output Screenshot -

```
hadoop@ip-172-31-10-189:~
    > from (
    > Select violationbin, ViolationCode, count(*) as ViolationCount from
    > (
    > select case
    > when substring(violationtime,1,2) in ('00','12','01','02','03') and upper(substring(violationtime,-1))='A' then 'MidNight_12AM_3AM'
    > when substring(violationtime,1,2) in ('04','05','06','07') and upper(substring(violationtime,-1))='A' then 'EarlyMorning_4AM_7AM'
    > when substring(violationtime,1,2) in ('08','09','10','11') and upper(substring(violationtime,-1))='A' then 'Morning_8AM_11AM'
    > when substring(violationtime,1,2) in ('12','01','02','03') and upper(substring(violationtime,-1))='P' then 'AfterNoon_12PM_3PM'
    > when substring(violationtime,1,2) in ('04','05','06','07') and upper(substring(violationtime,-1))='P' then 'Evening_4PM_7PM'
    > when substring(violationtime,1,2) in ('08','09','10','11') and upper(substring(violationtime,-1))='P' then 'Night_8PM_11PM'
    > else null end as violationbin,
    > ViolationCode
    > from parkingviolations
    > where IssueDate like '%2017'
    > )temp1
    > where violationbin is not NULL
    > group by violationbin,ViolationCode
    > ) temp2
    > ) temp3
    > where rank <= 3 ;
Query ID = hadoop_20220807152646_676be137-88c2-44a0-9ddd-053798e4fbd9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659882509877_0003)

----------------------------------------------------------------------------------------------
        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      6          6        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 39.69 s
----------------------------------------------------------------------------------------------
OK
EarlyMorning_4AM_7AM    14      74114   1
EarlyMorning_4AM_7AM    40      60652   2
EarlyMorning_4AM_7AM    21      57896   3
MidNight_12AM_3AM       21      36957   1
MidNight_12AM_3AM       40      25866   2
MidNight_12AM_3AM       78      15528   3
Night_8PM_11PM  7       26293   1
Night_8PM_11PM  40      22337   2
Night_8PM_11PM  14      21045   3
Morning_8AM_11AM        21      598060  1
Morning_8AM_11AM        36      348165  2
Morning_8AM_11AM        38      176570  3
Evening_4PM_7PM 38      102855  1
Evening_4PM_7PM 14      75902   2
Evening_4PM_7PM 37      70345   3
AfterNoon_12PM_3PM      36      286284  1
AfterNoon_12PM_3PM      38      240721  2
AfterNoon_12PM_3PM      37      167025  3
Time taken: 40.236 seconds, Fetched: 18 row(s)
hive>
```

Q2.3 – For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part).

Command –

```
select ViolationCode, ViolationTime_bin , count(*) as countByViolation from (
SELECT  ViolationCode,
case
when substring(ViolationTime,1,2) in ('00','01','02','03','12') and upper(substring(ViolationTime,-1))='A' then 'MidNight_12AM_3AM'
when substring(ViolationTime,1,2) in ('04','05','06','07') and upper(substring(ViolationTime,-1))='A' then 'EarlyMorning_4AM_7AM'
when substring(ViolationTime,1,2) in ('08','09','10','11') and upper(substring(ViolationTime,-1))='A' then 'Morning_8AM_11AM'
when substring(ViolationTime,1,2) in ('12','00','01','02','03') and upper(substring(ViolationTime,-1))='P' then 'AfterNoon_12PM_3PM'
when substring(ViolationTime,1,2) in ('04','05','06','07') and upper(substring(ViolationTime,-1))='P' then 'Evening_4PM_7PM'
when substring(ViolationTime,1,2) in ('08','09','10','11') and upper(substring(ViolationTime,-1))='P' then 'Night_8PM_11PM'
else null  end as ViolationTime_bin from parkingviolations where IssueDate like '%2017'
and (length(ViolationTime)=5 and upper(substring(ViolationTime,-1)) in ('A','P') and substring(ViolationTime,1,2)  in
('00','01','02','03','04','05','06','07',  '08','09','10','11','12') )) ViolationTable
group by ViolationCode, ViolationTime_bin
order by countByViolation desc
limit 3 ;
```

Output Screenshot -

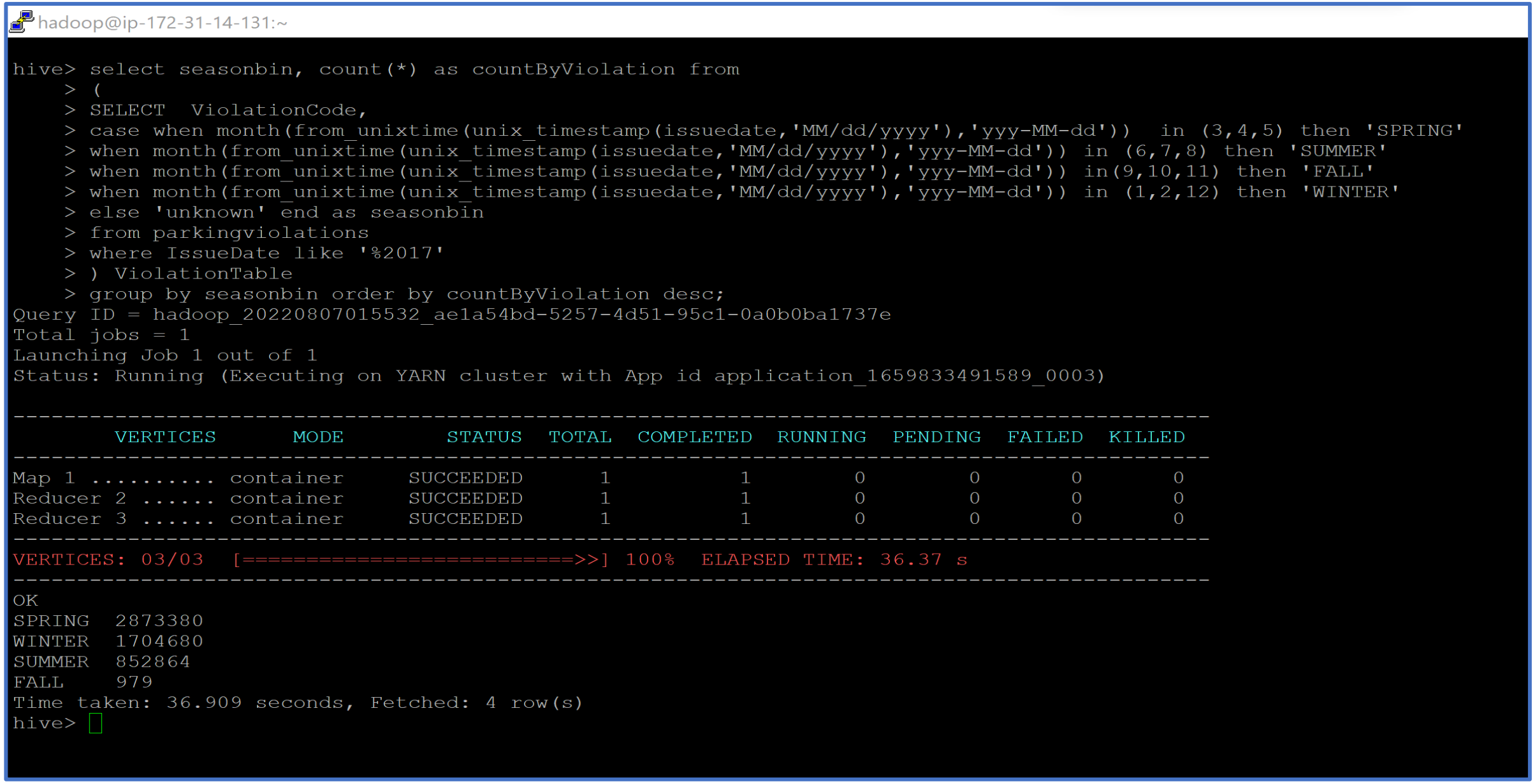
```
hive> select ViolationCode, ViolationTime_bin , count(*) as countByViolation from (
    > SELECT  ViolationCode,
    > case
    > when substring(ViolationTime,1,2) in ('00','01','02','03','12') and upper(substring(ViolationTime,-1))='A' then 'MidNight_12AM_3AM'
    > when substring(ViolationTime,1,2) in ('04','05','06','07') and upper(substring(ViolationTime,-1))='A' then 'EarlyMorning_4AM_7AM'
    > when substring(ViolationTime,1,2) in ('08','09','10','11') and upper(substring(ViolationTime,-1))='A' then 'Morning_8AM_11AM'
    > when substring(ViolationTime,1,2) in ('12','00','01','02','03') and upper(substring(ViolationTime,-1))='P' then 'AfterNoon_12PM_3PM'
    > when substring(ViolationTime,1,2) in ('04','05','06','07') and upper(substring(ViolationTime,-1))='P' then 'Evening_4PM_7PM'
    > when substring(ViolationTime,1,2) in ('08','09','10','11') and upper(substring(ViolationTime,-1))='P' then 'Night_8PM_11PM'
    > else null
    > end as ViolationTime_bin
    > from parkingviolations
    > where IssueDate like '%2017'
    > and
    > (
    > length(ViolationTime)=5
    > and upper(substring(ViolationTime,-1)) in ('A','P')
    > and substring(ViolationTime,1,2)  in ('00','01','02','03','04','05','06','07', '08','09','10','11','12')
    > )
    > ) ViolationTable
    > group by ViolationCode, ViolationTime_bin
    > order by countByViolation desc
    > limit 3 ;
Query ID = hadoop_20220807015004_ab23ff74-980e-4fbf-a9ff-a5ddb6d493ff
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1659833491589_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 43.16 s
----------------------------------------------------------------------------------------------
OK
21      Morning_8AM_11AM        598060
36      Morning_8AM_11AM        348165
36      AfterNoon_12PM_3PM      286284
Time taken: 43.71 seconds, Fetched: 3 row(s)
hive>
```

Q2.4.1 – First, divide the year into seasons, and find the frequencies of tickets for each season.

Command –

```
select seasonbin, count(*) as countByViolation from
(
SELECT  ViolationCode,
case when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd'))  in (3,4,5) then 'SPRING'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (6,7,8) then 'SUMMER'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in(9,10,11) then 'FALL'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (1,2,12) then 'WINTER'
else 'unknown' end as seasonbin
from parkingviolations
where IssueDate like '%2017'
) ViolationTable
group by seasonbin order by countByViolation desc;
```

Output Screenshot -



Q2.4.2 – Find the 3 most common violations for each of these seasons

Command –

```
select * from (
select seasonbin,ViolationCode,ViolationCount, dense_rank() over (partition by seasonbin order by ViolationCount desc) as rank
from (
Select seasonbin, ViolationCode, count(*) as ViolationCount from
(
SELECT
case when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd'))  in (3,4,5) then 'SPRING'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (6,7,8) then 'SUMMER'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in(9,10,11) then 'FALL'
when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (1,2,12) then 'WINTER'
else 'unknown' end as seasonbin ,
ViolationCode
from parkingviolations
where IssueDate like '%2017'
) temp1
group by seasonbin,ViolationCode
) temp2
)temp3
where rank <= 3 ;
```

Output Screenshot -

```
hadoop@ip-172-31-10-189:~
hive> select * from (
    > select seasonbin,ViolationCode,ViolationCount, dense_rank() over (partition by seasonbin order by ViolationCount desc) as rank
    > from (
    > Select seasonbin, ViolationCode, count(*) as ViolationCount from
    > (
    > SELECT
    > case when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd'))  in (3,4,5) then 'SPRING'
    > when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (6,7,8) then 'SUMMER'
    > when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in(9,10,11) then 'FALL'
    > when month(from_unixtime(unix_timestamp(issuedate,'MM/dd/yyyy'),'yyy-MM-dd')) in (1,2,12) then 'WINTER'
    > else 'unknown' end as seasonbin ,
    > ViolationCode
    > from parkingviolations
    > where IssueDate like '%2017'
    > ) temp1
    > group by seasonbin,ViolationCode
    > ) temp2
    > )temp3
    > where rank <= 3 ;
Query ID = hadoop_20220807145320_a80b8b15-a616-477f-a7ba-02e43efd2728
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1659882509877_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
Reducer 3 ...... container     SUCCEEDED      6          6        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 39.07 s
--------------------------------------------------------------------------------
OK
SPRING  21      402424  1
SPRING  36      344834  2
SPRING  38      271167  3
SUMMER  21      127350  1
SUMMER  36      96663   2
SUMMER  38      83518   3
WINTER  21      238180  1
WINTER  36      221268  2
WINTER  38      187386  3
FALL    46      231     1
FALL    21      128     2
FALL    40      116     3
```

Thank You