

# 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 -

```
hadoop@ip-172-31-8-140:~
Using username "hadoop".
Authenticating with public key "Jags2ndKeyPair"

  ____|_  ____|_  )
  ____|_  ____|_  /   Amazon Linux 2 AMI
  ____|_  ____|_  |

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM                MMMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M                M::::::::M R::::::::::::R
EE::::::::EEEEEEEEEE:E M::::::::M                M::::::::M R::::::::RRRRRR::::R
  E::::E          EEEEE M::::::::M                M::::::::M RR::::R          R::::R
  E::::E          M::::::::M:M                M::::::::M R::::R          R::::R
  E:::::EEEEEEEEEE M::::::::M M::::::::M M::::::::M R::::RRRRRR::::R
  E:::::EEEEEEEEEE M::::::::M M::::::::M M::::::::M R::::::::::::RR
  E:::::EEEEEEEEEE M::::::::M M::::::::M M::::::::M R::::RRRRRR::::R
  E::::E          M::::::::M M::::::::M M::::::::M R::::R          R::::R
  E::::E          EEEEE M::::::::M                M::::::::M R::::R          R::::R
EE:::::EEEEEEEEEE:E M::::::::M                M::::::::M R::::R          R::::R
E:::::EEEEEEEEEE:E M::::::::M                M::::::::M RR::::R          R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM                MMMMMMMM RRRRRRR          RRRRRR

[hadoop@ip-172-31-8-140 ~]$ wget https://hive-assignment-bucket.s3.amazonaws.com/Parking_Violations_Issued_-_Fiscal_Year_2017.csv;
--2022-08-04 23:27:09-- https://hive-assignment-bucket.s3.amazonaws.com/Parking_Violations_Issued_-_Fiscal_Year_2017.csv
Resolving hive-assignment-bucket.s3.amazonaws.com (hive-assignment-bucket.s3.amazonaws.com)... 52.217.223.17
Connecting to hive-assignment-bucket.s3.amazonaws.com (hive-assignment-bucket.s3.amazonaws.com)|52.217.223.17|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2086913576 (1.9G) [text/csv]
Saving to: 'Parking_Violations_Issued_-_Fiscal_Year_2017.csv'

100%[=====>] 2,086,913,576 46.7MB/s in 46s

2022-08-04 23:27:55 (43.6 MB/s) - 'Parking_Violations_Issued_-_Fiscal_Year_2017.csv' saved [2086913576/2086913576]

[hadoop@ip-172-31-8-140 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-8-140 ~]$ ls -ltr
total 2038004
-rw-rw-r-- 1 hadoop hadoop 2086913576 Jul 16 08:13 Parking_Violations_Issued_-_Fiscal_Year_2017.csv
[hadoop@ip-172-31-8-140 ~]$
```

## 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 -

```
hadoop@ip-172-31-12-187:~
https://aws.amazon.com/amazon-linux-2/
13 package(s) needed for security, out of 26 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M          M::::::::M R::::::::::::R
EE::::::::EEEEEEEEEE E M::::::::M          M::::::::M R::::::::RRRRRR:R
E:::E          EEEEE M::::::::M          M::::::::M RR:::R          R:::R
E:::E          M::::::::M::M M:::M::M::M M:::M::M R:::R          R:::R
E:::EEEEEEEEEE M::::M M:::M M:::M M:::M M:::M R::RRRRRR:R
E::::::::::::E M::::M M::M::M M:::M M:::M R:::::::::RR
E:::EEEEEEEEEE M::::M M:::M M:::M M:::M R::RRRRRR::R
E:::E          M::::M M:::M M:::M M:::M R:::R          R:::R
E:::E          EEEEE M::::M M::M M:::M M:::M R:::R          R:::R
EE:::EEEEEEEE::E M::::M          M::::M R:::R          R:::R
E::::::::::::E M::::M          M::::M RR:::R          R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMMM          MMMMMMMM RRRRRRRR          RRRRRR

[hadoop@ip-172-31-12-187 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-12-187 ~]$ ls -ltr
total 2038008
-rw-rw-r-- 1 hadoop hadoop 2086913576 Jul 16 08:13 Parking_Violations_Issued_-_Fiscal_Year_2017.csv
-rw-rw-r-- 1 hadoop hadoop 651 Aug 7 18:34 derby.log
[hadoop@ip-172-31-12-187 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> 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');
OK
Time taken: 1.808 seconds
hive> load data local inpath '/home/hadoop/Parking_Violations_Issued_-_Fiscal_Year_2017.csv' into table parkingviolations ;
Loading data to table default.parkingviolations
OK
Time taken: 19.855 seconds
hive> select * from parkingviolations limit 5;
OK
5092469481      GZH7067 NY      PAS      07/10/2016      7      SUBN      TOYOT      V      0      0      0      0      0      0      0      0143A      BX      A
LLERTON AVE (W/B) @ BARNES AVE      0      1111      D      T      5092451658      GZH7067 NY      PAS      07/08/2016      7      SUBN      TOYOT      V      0      0      0      0      0      0      0400P      BX      A
LLERTON AVE (W/B) @ BARNES AVE      0      1111      D      T      4006265037      FZX9232 NY      PAS      08/23/2016      5      SUBN      FORD      V      0      0      0      0      0      0      0233P      BX      S
B WEBSTER AVE @ E 1 94TH ST 0      1111      C      T      8478629828      66623ME NY      COM      06/14/2017      47      REFG      MITSU      T      10610      34330      34350      20180630      0014      14      14      359594      T102      J      1120A      N
Y      O      330      7th Ave      0      408      12      Y      0700A      0700P      WH      2007      0      04      47-Double PKG-Midtown
7868300310      37033JV NY      COM      11/21/2016      69      DELV      INTER      T      10510      34310      34330      20170228      0013      13      13      364832      T102      M      0555P      N
Y      F      799      6th Ave      0      408      h1      Y      0700A      0700P      WHITE      2007      0      31 6      69-Failure to Disp Muni Recpt
Time taken: 2.014 seconds, Fetched: 5 row(s)
hive>
```

## **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 -

```
hive> select distinct(RegistrationState) from parkingviolations where issuedate like '%2017' and RegistrationState rlike '^([A-Z])';
Query ID = hadoop_20220807013425_70c6e222-83f5-4b52-a4ea-ec9f75426576
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
-----
VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 22.29 s
-----
OK
DC
DE
ID
MB
ME
MN
ND
NH
PA
TX
WA
WI
WV
AK
AR
CA
CT
FL
GV
MA
NB
OH
PR
QB
TN
WY
CO
DP
IN
KS
KY
MO
NE
NU
NV
SC
SD
AB
```

```
FO
HI
IL
MS
NM
NS
NY
ON
OR
BC
GA
MD
NC
RI
SK
UT
VA
VT
AL
AZ
IA
LA
MI
MT
OK
PE
Time taken: 22.805 seconds, Fetched: 64 row(s)
hive> █
```



**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 substring(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 substring(ViolationTime, 1, 2) order by ViolationCountINPM desc;
Query ID = hadoop_20220807185514_bifbaaal-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 -

```
hadoop@ip-172-31-14-131:~
hive> 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;
Query ID = hadoop_20220807015532_ae1a54bd-5257-4d51-95c1-0a0b0ba1737e
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    1         1         0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 03/03  [=====>>] 100%  ELAPSED TIME: 36.37 s
-----
OK
SPRING  2873380
WINTER  1704680
SUMMER  852864
FALL    979
Time taken: 36.909 seconds, Fetched: 4 row(s)
hive> █
```

## 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**