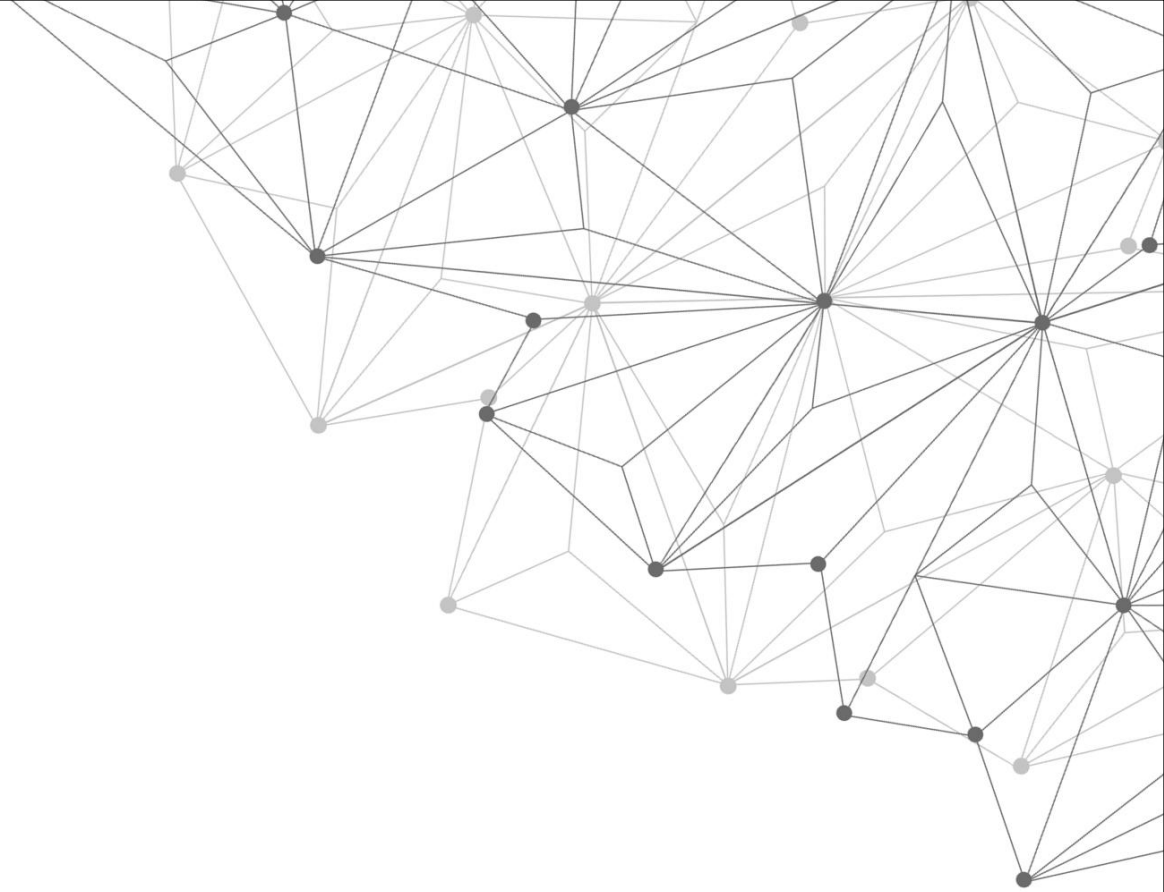
Hive Assignment



CONTENTS

Examine the Data

Create Database

```
hive>
    >
    >
    > create database hiveasssign;
OK
Time taken: 0.19 seconds
hive>
    >
    >
    >
    >
```

Create Table

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS hiveasssign.nyc2017(Summons_Number
    > BIGINT,Plate_ID STRING,Registration_State STRING,Plate_Type
    > STRING,Issue_Date STRING,Violation_Code BIGINT,Vehicle_Body_Type
    > STRING,Vehicle_Make STRING,Issuing_Agency STRING,Street_Code1
    > BIGINT,Street_Code2 BIGINT,Street_Code3
    > BIGINT,Vehicle_Expiration_Date BIGINT,Violation_Location
    > STRING,Violation_Precinct BIGINT,Issuer_Precinct BIGINT,Issuer_Code
    > BIGINT,Issuer_Command STRING,Issuer_Squad STRING,Violation_Time
    > STRING,Time_First_Observed STRING,Violation_County
    > STRING,Violation_In_Front_Of_Or_Opposite STRING,House_number
    > BIGINT,Street_Name STRING,Intersecting_Street
    > STRING,Date_First_Observed BIGINT,Law_Section BIGINT,Sub_Division
    > STRING,Violation_Legal_Code STRING,Days_Parking_In_Effect
    > STRING,From_Hours_In_Effect STRING,To_Hours_In_Effect
    > STRING,Vehicle_Color STRING,Unregistered_Vehicle
    > STRING,Vehicle_Year BIGINT,Meter_number BIGINT,Feet_From_Curb
    > BIGINT,Violation_Post_Code STRING,Violation_Description
    > STRING,No_Standing_or_Stopping_Violation STRING,Hydrant_Violation
    > STRING,Double_Parking_Violation STRING,Latitude BIGINT,Longitude
    > BIGINT,Community_Board BIGINT,Community_Council
    > BIGINT,Census_Tract BIGINT,BIN BIGINT,BBL BIGINT,NTA STRING)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES (   'separatorChar'=','  ,'quoteChar'    ='\"' )
    > STORED AS TEXTFILE
    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.591 seconds
hive>
```

Q 1.1: Find the total number of tickets for the year.

```
    >
    >
    >
    > select count(*) from hiveasssign.clean_nyc2017 where  year(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd'))=2017;
Query ID = hadoop_20240125054848_2af2f172-7174-41f4-9468-9da5e1389ef9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 34.54 s
----------------------------------------------------------------------------------------------
OK
5431903
Time taken: 35.002 seconds, Fetched: 1 row(s)
hive>
```

: - There are 5,431,903 parking violations in the year 2017

Q 1.2: Find out the total number of states to which the cars with tickets belong. The count of states is mandatory here; providing the exact list of states is optional.

```
    >
    >
    > select count(distinct(Registration_State)) from hiveasssign.clean_nyc2017 where issue_date like '%2017' and Registration_State rlike '^([A-Z])';
Query ID = hadoop_20240125055242_4c724efe-3632-4b74-8f40-e1e547cf7504
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.50 s
----------------------------------------------------------------------------------------------
OK
64
Time taken: 26.238 seconds, Fetched: 1 row(s)
hive>
```

Output: Cars belong to 64 States with ticket

Q 1.3: Find out the number of such tickets which have no addresses.

```
    >
    >
    > select count(*) from clean_nyc2017 where Street_Code1='0' or Street_Code2='0' or Street_Code3='0';
Query ID = hadoop_20240125055413_d0293984-637c-4c79-a246-99d7ac654c64
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 23.79 s
--------------------------------------------------------------------------------
OK
1816814
Time taken: 24.079 seconds, Fetched: 1 row(s)
hive>
```

Output :1816814 tickets have no addresses (either of StreetCode1, 2 and 3 is null)

Aggregation Tasks

Q 2.1: Find out the frequency of parking violations across different times of the day

```
    >
    >
    > select substring(Violation_Time, 1,2), count(*) as violationsCountINAM from clean_nyc2017 where upper(substring(Violation_Time, -1)) ='A' group by substring(Violation_Time, 1, 2) orde
r by violationsCountINAM desc;
Query ID = hadoop_20240125055524_12373b30-f053-46b6-872d-28c67f645295
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 30.17 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
A       1
.9      1
0.      1
Time taken: 30.555 seconds, Fetched: 16 row(s)
hive>
```

Output : 9 AM and 1 PM are the hours with maximum parking violations.

Q 2.1: Find out the frequency of parking violations across different times of the day

```
    > select substring(Violation_Time, 1,2), count(*) as violationsCountINPM from clean_nyc2017 where upper(substring(Violation_Time, -1)) ='P' group by substring(Violation_Time, 1, 2) orde
r by violationsCountINPM desc;
Query ID = hadoop_20240125055739_a761b251-5a71-4a75-9aa1-7066d2f4415f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED     15         15        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.66 s
----------------------------------------------------------------------------------------------
OK
01      549285
12      510012
02      466068
03      314467
04      295983
05      211173
06      104284
09      55322
08      49221
10      42540
11      29277
07      26100
00      119
87      5
76      4
16      3
15      3
56      3
54      2
59      2
13      2
22      2
65      2
78      2
27      1
28      1
30      1
41      1
50      1
17      1
21      1
36      1
37      1
51      1
68      1
```

Output : Traffic violations 1pm onwards

Q 2.2: 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.

```
    >
    > select * from (
    > select violation_bin,violation_code,Violation_Count, dense_rank() over (partition by violation_bin order by Violation_Count desc) as rank
    > from
    > ( Select violation_bin, violation_code, count(*) as Violation_Count from
    > ( select case
    > when substring(violation_time,1,2) in ('00','12','01','02','03') and upper(substring(violation_time,-1))='A' then 'MidNight_12AM_3AM'
    > when substring(violation_time,1,2) in ('04','05','06','07') and upper(substring(violation_time,-1))='A' then 'EarlyMorning_4AM_7AM'
    > when substring(violation_time,1,2) in ('08','09','10','11') and upper(substring(violation_time,-1))='A' then 'Morning_8AM_11AM'
    > when substring(violation_time,1,2) in ('12','01','02','03') and upper(substring(violation_time,-1))='P' then 'AfterNoon_12PM_3PM'
    > when substring(violation_time,1,2) in ('04','05','06','07') and upper(substring(violation_time,-1))='P' then 'Evening_4PM_7PM'
    > when substring(violation_time,1,2) in ('08','09','10','11') and upper(substring(violation_time,-1))='P' then 'Night_8PM_11PM'
    > else null end as violation_bin, violation_code from clean_nyc2017
    > )temp1
    > where violation_bin is not NULL group by violation_bin,violation_code
    > ) temp2
    > ) temp3 where rank <= 3 ;
Query ID = hadoop_20240125055954_dd66a79e-211a-4f21-accb-ee388b2243eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        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: 34.37 s
----------------------------------------------------------------------------------------------
OK
AfterNoon_12PM_3PM      36      286284  1
AfterNoon_12PM_3PM      38      240721  2
AfterNoon_12PM_3PM      37      167025  3
EarlyMorning_4AM_7AM    14      74114   1
EarlyMorning_4AM_7AM    40      60652   2
EarlyMorning_4AM_7AM    21      57896   3
Evening_4PM_7PM 38      102855  1
Evening_4PM_7PM 14      75902   2
Evening_4PM_7PM 37      70345   3
MidNight_12AM_3AM       21      36957   1
MidNight_12AM_3AM       40      25866   2
MidNight_12AM_3AM       78      15528   3
Morning_8AM_11AM        21      598060  1
Morning_8AM_11AM        36      348165  2
Morning_8AM_11AM        38      176570  3
Night_8PM_11PM  7       26293   1
Night_8PM_11PM  40      22337   2
Night_8PM_11PM  14      21045   3
Time taken: 34.91 seconds, Fetched: 18 row(s)
hive>
```

Output : These are the 3 most commonly occuring violations across the time intervals

Q 2.3: For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part).

```
    >
    >
    >
    > select Violation_Code, ViolationTime_bin , count(*) as countByViolation from (
    > SELECT  Violation_Code,
    > case
    > when substring(Violation_Time,1,2) in ('00','01','02','03','12') and upper(substring(Violation_Time,-1))='A' then 'MidNight_12AM_3AM'
    > when substring(Violation_Time,1,2) in ('04','05','06','07') and upper(substring(Violation_Time,-1))='A' then 'EarlyMorning_4AM_7AM'
    > when substring(Violation_Time,1,2) in ('08','09','10','11') and upper(substring(Violation_Time,-1))='A' then 'Morning_8AM_11AM'
    > when substring(Violation_Time,1,2) in ('12','00','01','02','03') and upper(substring(Violation_Time,-1))='P' then 'AfterNoon_12PM_3PM'
    > when substring(Violation_Time,1,2) in ('04','05','06','07') and upper(substring(Violation_Time,-1))='P' then 'Evening_4PM_7PM'
    > when substring(Violation_Time,1,2) in ('08','09','10','11') and upper(substring(Violation_Time,-1))='P' then 'Night_8PM_11PM'
    > else null
    > end as ViolationTime_bin
    > from clean_nyc2017
    > where (length(Violation_Time)=5 and upper(substring(Violation_Time,-1)) in ('A','P') and substring(Violation_Time,1,2)  in ('00','01','02','03','04','05','06','07', '08','09','10','11
','12') )) ViolationTable
    > group by Violation_Code, ViolationTime_bin
    > order by countByViolation desc
    > limit 3 ;
Query ID = hadoop_20240125060201_2359a961-b82d-462f-9324-243ba7406f34
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        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: 32.65 s
----------------------------------------------------------------------------------------------
OK
21      Morning_8AM_11AM        598060
36      Morning_8AM_11AM        348165
36      AfterNoon_12PM_3PM      286284
Time taken: 33.082 seconds, Fetched: 3 row(s)
hive>
```

Q 2.4.1: First, divide the year into seasons, and find the frequencies of tickets for each season.

```
    >
    >
    > select seasonbin, count(*) as countByViolation from
    > (
    > SELECT  Violation_Code,
    > case when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd'))  in (3,4,5) then 'SPRING'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in (6,7,8) then 'SUMMER'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in(9,10,11) then 'FALL'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in (1,2,12) then 'WINTER'
    > else 'unknown' end as seasonbin
    > from clean_nyc2017
    > ) ViolationTable
    > group by seasonbin order by countByViolation desc;
Query ID = hadoop_20240125060328_ac2edc25-68e5-41de-9b52-3bc4c44e063e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        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: 43.55 s
--------------------------------------------------------------------------------
OK
SPRING  2873380
WINTER  1704680
SUMMER  852864
FALL    979
unknown 15
Time taken: 44.135 seconds, Fetched: 5 row(s)
hive>
```

Q 2.4.2: Find the 3 most common violations for each of these seasons

```
    > select * from (
    > select seasonbin,Violation_Code,ViolationCount, dense_rank() over (partition by seasonbin order by ViolationCount desc) as rank
    > from (
    > Select seasonbin, Violation_Code, count(*) as ViolationCount from
    > (
    > SELECT
    > case when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd'))  in (3,4,5) then 'SPRING'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in (6,7,8) then 'SUMMER'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in(9,10,11) then 'FALL'
    > when month(from_unixtime(unix_timestamp(issue_date,'MM/dd/yyyy'),'yyy-MM-dd')) in (1,2,12) then 'WINTER'
    > else 'unknown' end as seasonbin ,
    > Violation_Code
    > from clean_nyc2017
    > ) temp1
    > group by seasonbin,Violation_Code
    > ) temp2
    > )temp3
    > where rank <= 3 ;
Query ID = hadoop_20240125060447_f287a879-c596-4fe7-a8cc-7fc4eb20fbe7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706159571324_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     15         15        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: 46.56 s
----------------------------------------------------------------------------------------------
OK
FALL    46      231     1
FALL    21      128     2
FALL    40      116     3
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
unknown 01/23/2017      3       1
unknown 01/30/2017      2       2
unknown 06/12/2017      1       3
unknown 03/13/2017      1       3
unknown 02/17/2017      1       3
unknown 01/24/2017      1       3
unknown PAS     1       3
```

THANK YOU

This Report was made by Kaustubh Nitin Patil