Objective:

The assignment is meant for you to apply learnings of the module on Hive on a real-life dataset. One of the major objectives of this assignment is gaining familiarity with how an analysis works in Hive and how you can gain insights from large datasets.

Problem Statement:

New York City is a thriving metropolis and just like most other cities of similar size, one of the biggest problems its residents face is parking. The classic combination of a huge number of cars and a cramped geography is the exact recipe that leads to a large number of parking tickets.

In an attempt to scientifically analyze this phenomenon, the NYC Police Department regularly collects data related to parking tickets. This data is made available by NYC Open Data portal. We will try and perform some analysis on this data.

Dataset: https://data.cityofnewyork.us/browse?q=parking+tickets

Step 1: Create a table which stores the above downloaded dataset using the below command.

```
create table parking_violations_details
(
    Summons_Number bigint,
    Plate_ID string,
    Registration_State string,
    Plate_Type string,
    Issue_Date string,
    Violation_Code int,
    Vehicle_Body_Type string,
    Vehicle_Make string,
    Issuing_Agency string,
    Street_Code1 int,
    Street_Code2 int,
    Street_Code3 int,
    Vehicle_Expiration Date,
    Violation_Location int,
    Violation_Precinct int,
    Issuer_Precinct int,
    Issuer_Code int,
    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 string,
    Street_Name string,
    Intersecting_Street string,
    Date_First_Observed int,
    Law_Section int,
    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 int,
    Vehicle_Year string,
    Meter_Number string,
    Feet_From_Curb int,
    Violation_Post_Code string,
    Violation_Description string,
    No_Standing_or_Stopping_Violation string,
    Hydrant_Violation string,
    Double_Parking_Violation string
)
row format delimited
fields terminated by ','
tblproperties ("skip.header.line.count" = "1");
```

```
hive> create table parking_violations_details
    > (
    > Summon_number bigint,
    > Plate_ID string,
    > Registration_state string,
    > Plate_type string,
    > Issue_date string,
    > Violation_code int,
    > Vehicle_body_type string,
    > Vehicle_make string,
    > Issuing_agency string,
    > Street_code1 int,
    > Street_code2 int,
    > Street_code3 int,
    > Vehicle_expiration Date,
    > Violation_location int,
    > Violation_precint int,
    > Issuer_precinct int,
    > Issuer_code int,
    > Issuer_command string,
    > Issuer_squad string,
    > Violation_time string,
    > Time_first_observed string,
    > Violation_country string,
    > Violation_in_front_of_or_opposite string,
    > House_number string,
    > Street_name string,
    > Intersecting_Street string,
    > Date_first_observed int,
    > Law_section int,
    > 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 int,
    > Vehicle_year string,
    > Meter_number string,
    > Feet_from_curb int,
    > Violation_post_code string,
    > Violation_description string,
    > No_standing_or_stopping_violation string,
    > Hydrant_violation string,
    > Double_parking_violation string
    > ) row format delimited
    > fields terminated by ','
    > tblproperties ("skip.header.line.count" = "1");
OK
Time taken: 3.528 seconds
```

Step 2: Load the data into the above table

From local:

load data local inpath '/home/cloudera/hive_project_2/Parking_Violations_Details_2017.csv' into table parking_violations_details;

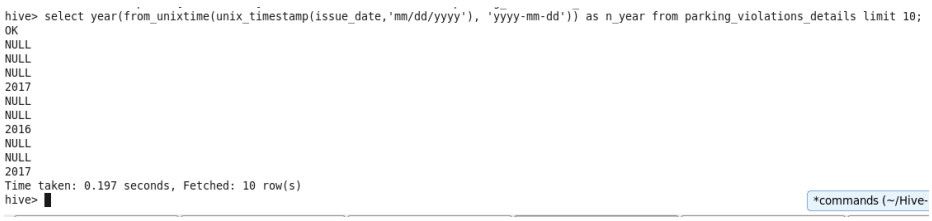
Step 3: Queries to check if the data is uploaded properly or not

select * from parking_violations_details limit 20;

```
hive> select * from parking_violations_details limit 20;
OK
5092469481      GZH7067 NY      PAS     07-10-2016      7       SUBN    TOYOT   V       0       0       0       NULL    NULL    0       0       0                       0143A           BX                      ALLERTON AVE (W/B) @    BARNE
S AVE   0       1111    D       T                                       GY      NULL    2001            0               FAILURE TO STOP AT RED LIGHT
5092451658      GZH7067 NY      PAS     07-08-2016      7       SUBN    TOYOT   V       0       0       0       NULL    NULL    0       0       0                       0400P           BX                      ALLERTON AVE (W/B) @    BARNE
S AVE   0       1111    D       T                                       GY      NULL    2001            0               FAILURE TO STOP AT RED LIGHT
4631633384      AVM7975 NY      PAS     03-09-2017      36      SUBN    GMC     V       0       0       0       NULL    NULL    0       0       0                       1211P           BK                      WB LINDEN BLVD @ LIN    COLN
AVE     0       1180    B       T                                       GY      NULL    2010            0               PHTO SCHOOL ZN SPEED VIOLATION
8196557280      GWB7054 NY      PAS     01/18/2017      70      SUBN    TOYOT   T       59590   8590    57790   NULL    109     109     109     364933  T401    J       1217P           Q       F       35-11   Prince St               0   4
08      j3              YYYYYYY         BL      NULL    2015            0       5       70A-Reg. Sticker Expired (NYS)
4631184358      EXZ9820 NY      PAS     03-02-2017      36      4DSD    HONDA   V       0       0       0       NULL    NULL    0       0       0                       1207P           BK                      WB FLATLANDS AVE @ E    100 S
T       0       1180    B       T                                       GR      NULL    1997            0               PHTO SCHOOL ZN SPEED VIOLATION
4007039033      GZE1511 NY      PAS     03-06-2017      5       4DSD    TOYOT   V       0       0       0       NULL    NULL    0       0       0                       1037A           BK                      NB UTICA AVE @ CHURC    H AVE
0       1111    C       T                                       WH      NULL    2001            0               BUS LANE VIOLATION
7662736064      83485MH NY      COM     07/20/2016      48      VAN     FRUEH   T       0       40404   40404   NULL    26      26      26      359294  T103    L       0101P           NY      F       2164    Frederick Douglass B        0
408     e9                              BROWN   NULL    2016            0       01 7    48-Bike Lane
8539360652      GEH9367 NY      PAS     05-04-2017      70      4DSD    DODGE   T       31830   5430    5580    NULL    78      78      78      347470  T301    E       0602P           K       F       433     Dean St         0       408 j
3               YYYYYYY                 GY      NULL    2011            0       19      70A-Reg. Sticker Expired (NYS)
8293544302      FYP7892 NY      PAS     10-03-2016      14      SUBN    TOYOT   T       14380   35980   36030   NULL    61      61      61      357808  T302    Q       0814A           K       O       2623    Avenue P                0   4
08      c               YYYYYYY         WH      NULL    2016            0       CC1     14-No Standing
8525962235      HHM1058 NY      PAS     04/20/2017      21      4DSD    ME/BE   T       35720   34020   22020   NULL    43      43      43      363937  T201    I       1238P           BX      O       1684    Gleason Ave             0   4
08      d1              Y Y     1130A   0100P   WH      NULL    2015            0       3       21-No Parking (street clean)
7405998999      HDM2727 NY      PAS     08-11-2016      70      VAN     FORD    T       73120   26580   73890   NULL    42      42      42      352973  T801    B       0246A           BX      F       1011    Washington Ave          0   4
08      j3                              WHITE   NULL    0               0       K 21    70-Reg. Sticker Missing (NYS)
8463518175      56989MB NY      COM     12/20/2016      38      VAN     CHEVR   T       59990   12790   12840   NULL    112     112     112     361797  T402    B       1130A           Q       F       93-54   Queens Blvd             0   4
08      h1              Y       0830A   0700P   WHITE   NULL    2011            0       03 4    38-Failure to Display Muni Rec
4635298826      KKG4708 PA      PAS     06/19/2017      36      SW      CHEVR   V       0       0       0       NULL    NULL    0       0       0                       1107A           BK                      SB FLATBUSH AVE EXT     @ CHA
PEL ST  0       1180    B       T                                       NULL    2002            0               PHTO SCHOOL ZN SPEED VIOLATION
8520357982      FXG2663 NY      PAS     05-04-2017      21      4DSD    HYUND   T       18820   10110   10010   NULL    19      19      19      358643  T103    X       0907A           NY      F       325     E 92nd St               0   4
08      d1              Y Y     0900A   1030A   GY      NULL    2008            0       1       21-No Parking (street clean)
7811876425      655790  RI      PAS     10-07-2016      21      SUBN    FORD    T       18910   27790   24890   NULL    23      23      23      341195  T103    B       0953A           NY      F       114     E 97th St               0   4
08      d1              Y Y     0930A   1100A   OTHER   NULL    0               0       CCA     21-No Parking (street clean)
7722852804      BAS3436 NY      PAS     01/31/2017      31      SUBN    HONDA   T       34350   10610   10810   NULL    14      14      14      359588  T108    A       0541P           NY      F       252     W 29th St               0   4
08      l3              YYYYY   0800A   0600P   GY      NULL    2017            0       1       31-No Stand (Com. Mtr. Zone)
4634732270      FTW2676 NY      PAS     06-05-2017      36      PICK    CHEVR   V       0       0       0       NULL    NULL    0       0       0                       1223P           QN                      NB FRANCIS LEWIS BLV    D @ 4
2ND RD  0       1180    B       T                                       MR      NULL    2008            0               PHTO SCHOOL ZN SPEED VIOLATION
8483805662      41006PC NY      IRP     01/21/2017      47      TK      FRUEH   T       10610   34470   34490   NULL    14      14      14      362611  T102    I       0709A           NY      F       462     7th Ave         0       408 l
2               Y       0700A   0700P   MR      NULL    2015            0       99      47-Double PKG-Midtown
8530700028      26796ME NY      COM     02/22/2017      78      VAN     FORD    T       30620   18520   38620   NULL    43      43      43      358860  T801    A       1114P           BX      O       2262    Ellis Ave               0   4
08      k6              YYYYYYY 0900P   0500A   WH      NULL    2011            0       A       78-Nighttime PKG on Res Street
8520272381      ZWJ55R  NJ      PAS     06/17/2017      24      SUBN    HONDA   T       35310   15710   11710   NULL    20      20      20      362195  T103    S       0447P           NY      F       172     W 77th St               0   4
08      d5              YYYYYYY                 GREY    NULL    0               0       C       24-No Parking (exc auth veh)
Time taken: 0.772 seconds, Fetched: 20 row(s)
hive>
```

Step 4: Extracting year from issue_date

select year(from_unixtime(unix_timestamp(issue_date,'mm/dd/yyyy'), 'yyyy-mm-dd')) as n_year from parking_violations_details limit 100;



Step 5: Creating table to store only the tickets issued in year 2017

create table parking_violations_details_2017

(

Summons_Number bigint,

Plate_ID string,

Registration_State string,

Plate_Type string,

Issue_Date string,

Violation_Code int,

Vehicle_Body_Type string,

Vehicle_Make string,

Issuing_Agency string,

Street_Code1 int,

Street_Code2 int,

Street_Code3 int,

Vehicle_Expiration Date,

Violation_Location int,

Violation_Precinct int,

Issuer_Precinct int,

Issuer_Code int,

Issuer_Command string,

Issuer_Squad string,

Violation_Time string,

Time_First_Observed string,

Violation_In_Front_Of_Or_Opposite string,

House_Number string,

```
    Street_Name string,
    Intersecting_Street string,
    Date_First_Observed int,
    Law_Section int,
    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 int,
    Vehicle_Year string,
    Meter_Number string,
    Feet_From_Curb int,
    Violation_Post_Code string,
    Violation_Description string,
    No_Standing_or_Stopping_Violation string,
    Hydrant_Violation string,
    Double_Parking_Violation string
)
COMMENT 'A bucketed sorted parking_violations_details_2017'
partitioned by (Violation_County string)
CLUSTERED BY (Violation_Code) sorted by (Violation_Code) INTO 8 BUCKETS
row format delimited
fields terminated by ','
tblproperties ("skip.header.line.count" = "1");
```

Step 6: To load data into partition and bucket table we need to set few properties to enable bucketing and dynamic partition

```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
set hive.enforce.bucketing = true;
```

Step 7: Load data from parking_violations_details to parking_violations_issued_2017

```
Insert into parking_violations_details_2017 partition (Violation_County) select
```

Summons_Number,

Plate_ID,

Registration_State,

Plate_Type,

Issue_Date,

Violation_Code,

Vehicle_Body_Type,

Vehicle_Make,

Issuing_Agency,

Street_Code1,

Street_Code2,

Street_Code3,

Vehicle_Expiration,

Violation_Location,

Violation_Precinct,

Issuer_Precinct,

Issuer_Code,

Issuer_Command,

Issuer_Squad,

Violation_Time,

Time_First_Observed,

Violation_In_Front_Of_Or_Opposite,

House_Number,

Street_Name,

Intersecting_Street,

Date_First_Observed,

Law_Section,

Sub_Division,

Violation_Legal_Code,

Days_Parking_In_Effect,

From_Hours_In_Effect,

To_Hours_In_Effect,

Vehicle_Color,

Unregistered_Vehicle,

```
Vehicle_Year,

Meter_Number,

Feet_From_Curb,

Violation_Post_Code,

Violation_Description,

No_Standing_or_Stopping_Violation,

Hydrant_Violation,

Double_Parking_Violation,

Violation_County

from parking_violations_details where
year(from_unixtime(unix_timestamp(issue_date,'mm/dd/yyyy'), 'yyyy-mm-dd')) = 2017;
```

The analysis can be divided into two parts:

Part-I: Examine the data

1) Find the total number of tickets for the year.

```
select count (distinct summons_number) No_Tickets, year(issue_date) as year from
parking_violations_details_2017 group by year(issue_date);
```

2) Find out how many unique states the cars which got parking tickets came from.

```
select count (distinct Registration_State) as Reg_state_count from parking_violations_details_2017;

select distinct (Registration_State) as Reg_state from parking_violations_details_2017;

select Registration_State, Count(1) as Number_of_Records from parking_violations_details_2017
group by Registration_State order by Number_of_Records;
```

3) Some parking tickets don't have addresses on them, which is cause for concern. Find out how many such tickets there are (i.e. tickets where either "Street Code 1" or "Street Code 2" or "Street Code 3" is empty).

```
select count (distinct summons_number) as No_Tickets_without_address from
parking_violations_details where Street_code1 = 0 or Street_code2 = 0 or Street_code3 = 0;
```

Part-II: Aggregation tasks

1) How often does each violation code occur? (Frequency of violation codes - find the top 5).

```
select count (Violation_Code) as frequency_of_violation, Violation_Code from
parking_violations_details_2017 group by Violation_Code order by frequency_of_violation desc limit 5;
```

2) How often does each vehicle body type get a parking ticket? How about the vehicle make? (Find the top 5 for both).

```
select Vehicle_Body_Type, count (summons_number)as frequency_of_getting_parking_ticket from
challenge.parking_violations_details_2017 group by Vehicle_Body_Type order by
frequency_of_getting_parking_ticket desc limit 5;

select Vehicle_make, count(summons_number) as frequency_of_getting_parking_ticket from
challenge.parking_violations_details_2017 group by Vehicle_make order by
frequency_of_getting_parking_ticket desc limit 5;
```

3) A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

a) Violating Precincts (this is the precinct of the zone where the violation occurred)

select Violation_Precinct, count (*) as IssuedTicket from challenge.parking_violations_details group by Violation_Precinct order by IssuedTicket desc limit 5;

b) Issuer Precincts (this is the precinct that issued the ticket)

select Issuer_Precinct, count (*) as IssuedTicket from challenge.parking_violations_details group by Issuer_Precinct order by IssuedTicket desc limit 5;

4) Find the violation code frequency across 3 precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes?

select Issuer_Precinct, Violation_Code, count (*) as TicketsIssued from challenge.parking_violations_details_2017 group by Issuer_Precinct, Violation_Code order by TicketsIssued desc limit 5;

We will not be considering 0. Therefore 18,19,14 are the three issuer precincts which have the maximum number of violations. Let's analyze the Issuer Precincts one by one.

Issuer Precinct 18

select Violation_Code, count(*) as TicketsIssued from challenge.parking_violations_details_2017 where Issuer_Precinct=18 group by Violation_Code order by TicketsIssued desc limit 5;

Issuer Precinct 19

select Violation_Code, count(*) as TicketsIssued from challenge.parking_violations_details_2017 where Issuer_Precinct=19 group by Violation_Code order by TicketsIssued desc limit 5;

Issuer Precinct 14

select Violation_Code, count(*) as TicketsIssued from challenge.parking_violations_details_2017 where Issuer_Precinct=14 group by Violation_Code order by TicketsIssued desc limit 5;

Common codes across precincts

select Issuer_Precinct, Violation_Code, count(*) as TicketsIssued from challenge.parking_violations_details_2017 where Issuer_Precinct in (18,19,14) group by Issuer_Precinct, Violation_Code order by TicketsIssued desc limit 10;

5) Find out the properties of parking violations across different times of the day: The Violation Time field is specified in a strange format. Find a way to make this into a time attribute that you can use to divide into groups.

select from_unixtime(unix_timestamp(regexp_extract(violation_time,'(.*)[A-Z]',1),'HHmm'),"HH:mm") as date_data from parking_violations_issued limit 2; --> converted to time format 01:43

select from_unixtime(unix_timestamp(concat(violation_time,'M'), 'HHmmaaa'),"HH:mmaaa") as date_data from parking_violations_issued limit 2; --> working 01:43AM

6) Divide 24 hours into 6 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

Partitioned view :

create view vw_parking_violations_2017_partitioned_bins partitioned on (Violation_Code) as

SELECT Summons_Number, Violation_Time, Issuer_Precinct,

case

when substring(Violation_Time,1,2) in ('00','01','02','03','12') and upper(substring(Violation_Time,-1))='A' then 1

when substring(Violation_Time,1,2) in ('04','05','06','07') and upper(substring(Violation_Time,-1))='A' then 2

when substring(Violation_Time,1,2) in ('08','09','10','11') and upper(substring(Violation_Time,-1))='A' then 3

when substring(Violation_Time,1,2) in ('12','00','01','02','03') and upper(substring(Violation_Time,-1))='P' then 4

when substring(Violation_Time,1,2) in ('04','05','06','07') and upper(substring(Violation_Time,-1))='P' then 5

when substring(Violation_Time,1,2) in ('08','09','10','11') and upper(substring(Violation_Time,-1))='P' then 6

else null end as Violation_Time_bin, Violation_Code

from parking_violations_details_2017

where Violation_Time is not null or (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'));

bin1

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 1 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

21

3660

40

2584

14

1574

bin2

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 2 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

14

7250

40

6403

21

5669

bin3

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 3 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

21

59465

36

37767

38

17587

bin4

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 4 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

36

28600

38

23877

37

16777

bin5

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 5 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

38

10148

14

7609

37

6944

bin6

select Violation_Code,count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Time_bin == 6 group by Violation_Code order by TicketsIssued desc limit 3;

Violation_code

TicktesIssued

7

2602

40

2159

14

2091

7) Now, try another direction. 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_Time_bin, count(*) TicketsIssued from vw_parking_violations_2017_partitioned_bins where Violation_Code in (21, 37, 38,36) group by Violation_Time_bin order by TicketsIssued desc limit 3;

Violation_Time_bin

TicketsIssued

3

116785

4

76701

5

18437

8) Let's try and find some seasonality in this data

a.) First, divide the year into some number of seasons, and find frequencies of tickets for each season. (Hint: A quick Google search reveals the following seasons in NYC: Spring(March, April, March); Summer(June, July, August); Fall(September, October, November); Winter(December, January, February))

Season Month interval

spring March, April, May

summer June, July, August

autumn September, October, November

winter December, January, February

Normal view :

create view vw_tickets_details_2017_bins as select Violation_Code , Issuer_Precinct,

case

when MONTH(Issue_Date) between 03 and 05 then 'spring'

when MONTH(Issue_Date) between 06 and 08 then 'summer'

when MONTH(Issue_Date) between 09 and 11 then 'autumn'

when MONTH(Issue_Date) in (1,2,12) then 'winter'

else 'unknown' end as season from parking_violations_details_2017;

Partitioned view :

create view vw_tickets_details_2017_partitioned_bins partitioned on (Violation_Code) as

select Issuer_Precinct,

case

when MONTH(Issue_Date) between 03 and 05 then 'spring'

when MONTH(Issue_Date) between 06 and 08 then 'summer'

when MONTH(Issue_Date) between 09 and 11 then 'autumn' select

when MONTH(Issue_Date) in (1,2,12) then 'winter'

else 'unknown' end as season, Violation_Code from parking_violations_details_2017;

select season, count(*) as TicketsIssued from vw_tickets_details_2017_partitioned_bins group by season order by TicketsIssued desc;

Season

TicktesIssued

Spring

285875

Winter

169466

Summer

84560

autumn

0

b) Then, find the 3 most common violations for each of these seasons.

 # spring season

select Violation_Code, count(*) as TicketsIssued from vw_tickets_details_2017_partitioned_bins where

season = 'spring' group by Violation_Code order by TicketsIssued desc limit 3;

Violation_Code

TicketsIssued

21

40045

36

34354

38

27001

 # winter season

select Violation_Code, count(*) as TicketsIssued from vw_tickets_details_2017_partitioned_bins where season = 'winter' group by Violation_Code order by TicketsIssued desc limit 3;

Violation_Code

TicketsIssued

21

23684

36

22084

38

18450

 # summer season

select Violation_Code, count(*) as TicketsIssued from vw_tickets_details_2017_partitioned_bins where season = 'summer' group by Violation_Code order by TicketsIssued desc limit 3;

Violation_Code

TicketsIssued

21

12565

36

9655

38

8331

 # autumn season

select Violation_Code, count(*) as TicketsIssued from vw_tickets_details_2017_partitioned_bins where season = 'autumn' group by Violation_Code order by TicketsIssued desc limit 3;

Violation_Code

TicketsIssued