Problem Statement

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 analyse 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.

Note: We have uploaded the data for 2017 on an S3 bucket at the following link, "s3://hiveassignmentdatabde/Parking_Violations_Issued_-_Fiscal_Year_2017.csv". You need to copy this data to your own bucket and then run the required queries.
Note: Consider only the year 2017 for analysis and not the Fiscal year.
The data dictionary is available on this page.

The analysis can be divided into two parts:

Part-I: Examine the data

1. Find the total number of tickets for the year.
2. Find out how many unique states the cars which got parking tickets came from.
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)

Part-II: Aggregation tasks

1. How often does each violation code occur? (frequency of violation codes - find the top 5)
2. How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)
3. A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:
    1. Violating Precincts (this is the precinct of the zone where the violation occurred)
    2. Issuer Precincts (this is the precinct that issued the ticket)
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?
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.

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
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)
8. Let's try and find some seasonality in this data
    1. 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))
    2. Then, find the 3 most common violations for each of these seasons.

Note: Please ensure you make necessary optimizations to your queries like selecting the appropriate table format, using partitioned/bucketed tables. Marks will be awarded for keeping the performance also in mind.