CSC 343, Lab #7: Data Modeling

The *la_parking_citations.zip* file contains Los Angeles parking citation data from December 2015 and December 2016. The data included here is a subset of the >1GB data set available here: https://www.kaggle.com/cityofLA/los-angeles-parking-citations/version/7

The data is comma-delimited and includes the following:

- *2015-12.csv* and *2016-12.csv* includes citation information with the following unlabeled columns:
    - *ticket_number* – a unique parking ticket #
    - *date* – the date of the citation, in *yyyy-mm-dd* format
    - *plate* – the license plate number
    - *make* – the make of the car
    - *color* – the color of the car
    - *code* – the code identifying the citation type
    - *amt* – the citation amount, in dollars

- *codes.csv* includes the following unlabeled columns:
    - *code* – the code identifying the citation type
    - *description* – the description of the citation type

For each question, specify the SQL code required to complete each step, unless instructed otherwise.

1. Create a Hive/Impala *internal (managed)* table named *citation* that will store the citation data, using the column names above. All data should be stored as strings with the exception of *date*, which should be stored as a *timestamp*, and *amt*, which should be stored as a decimal value. The data itself will be stored in */user/cloudera/parking*.

2. Create a Hive/Impala table that stores the *code* and *description* as strings, where the data will be stored in */user/cloudera/parking_codes*.

3. Describe how to add the citation data and code/description data to the database. Your description must include any code used (e.g., if using SQL), or any buttons pressed (e.g., if using Hue). Note that your answer does not need to include SQL code.

4. In your *citation* table, change the name of the *ticket_number* column to *id.*

5. In a single query, find the number of citations issued in December of 2015 and the number of citations issued in December 2016. Note that for the *timestamp* field named

date, the *year* can be extracted by using `year(date)` in your query. Include both the SQL code and the result in your answer.

6. In 2016, which day(s) in December had the most citations, and how many citations were issued? Which days in 2016 had the least citations, and how many citations were issued? Include both the SQL code and result in your answer.

7. Which three citations were the most expensive? For these results, you should display the code, the description, and the amount, without any duplicate values. Include both the SQL code and result in your answer.

8. Delete the *citation* table from the database, but do **NOT** delete the data from HDFS!