CS 348 - Homework 7: Data Warehousing (HIVE and BigQuery).
(40 Points)

Fall 2020
Due on: **12/1/2020 at 11:59 pm**
This assignment is to be completed by individuals. You should only talk to the instructor, and the TAs about this assignment. You may also post questions (and not answers) to Campuswire.

There will be a 10% penalty if the homework is submitted 24 hours after the due date, a 15% penalty if the homework is submitted 48 hours after the due date, or a 20% penalty if the homework is submitted 72 hours after the due date. The homework will not be accepted after 72 hours, as a solution will be posted by then.

Submission Instruction: Write your answers in this document. Create a pdf and upload to Gradescope.

For questions 1 and 2, consider the us_cities.csv file included with this homework (this file does not include column headers). us_cities_sample.csv includes the column headers and only a few rows. The create-table statement can be found in createTable.sql.

For more information about HIVE and BigQuery, including the supported SQL clauses, you can visit the HIVE and BigQuery documentation.
https://cwiki.apache.org/confluence/display/Hive/LanguageManual

https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax

Question 1) (10 points)
Using HIVE, list each state along with its population. The population of a state is the total population of its cities. Sort your result by population in a descending order. Write your answer (query) and insert a screenshot of your result (partial result is sufficient). Your screenshot should show the HIVE command prompt.

Expected Result:

```
+-----------------------+-----------+
|      state_name       | totalPop  |
+-----------------------+-----------+
| California            | 59608877  |
| New York              | 37847876  |
| Texas                 | 33020438  |
| Florida               | 30243610  |
| Illinois              | 18266059  |
| Pennsylvania          | 15641201  |
| Ohio                  | 13764673  |
| Georgia               | 10750970  |
| Michigan              | 10429997  |
| Washington            | 10134988  |
| Arizona               | 9581518   |
| North Carolina        | 9203228   |
| Massachusetts         | 8972816   |
| Virginia              | 8197045   |
| Colorado              | 7499152   |
| Missouri              | 7463356   |
| Minnesota             | 7275566   |
| Maryland              | 6880970   |
| Indiana               | 6161106   |
```

....

```
OUTPUT is TRIMMED
```

Question 2, HIVE) (15 points)

The explode command in HIVE can be used to flatten complex data types, such as maps and arrays. See the following example of Explode in the HIVE documentation: https://cwiki.apache.org/confluence/display/Hive/LanguageManual+LateralView

Using explode to flatten the zips column in the cities data set, find the highest zip code in the US. Write your query and insert a screenshot of your result. Your screenshot should show the HIVE command prompt.

Expected Result:

```
+--------------+
| highestzip   |
+--------------+
| 99950        |
+--------------+
```

Question 3) (5 points)

What is the MongoDB operator that is similar to Explode in HIVE?

Question 3, BigQuery) (10 points)
Choose a public data set in BigQuery and write two queries. Insert screenshots of your queries and results (partial results are sufficient).