DEBS 2015 - Problem Description

from http://www.debs2015.org/call-grand-challenge.html

The data for the 2015 Grand Challenge is based on a data set released under the FOIL (The Freedom of Information Law) and made public by Chris Whong (http://chriswhong.com/open-data/foil_nyc_taxi/). Details about the data, the queries for the DEBS 2015 Grand Challenge, and information about how the challenge will be evaluated is provided below.

Data

Provided data consists of reports of taxi trips including starting point, drop-off point, corresponding timestamps, and information related to the payment. Data are reported at the end of the trip, i.e., upon arrive in the order of the drop-off timestamps.

The specific attributes are listed below:

medallion	an md5sum of the identifier of the taxi - vehicle bound
hack_license	an md5sum of the identifier for the taxi license
pickup_datetime	time when the passenger(s) were picked up
dropoff_datetime	time when the passenger(s) were dropped off
trip_time_in_secs	duration of the trip
trip_distance	trip distance in miles
pickup_longitude	longitude coordinate of the pickup location
pickup_latitude	latitude coordinate of the pickup location
dropoff_longitude	longitude coordinate of the drop-off location
dropoff_latitude	latitude coordinate of the drop-off location
payment_type	the payment method - credit card or cash
fare_amount	fare amount in dollars
surcharge	surcharge in dollars
mta_tax	tax in dollars
tip_amount	tip in dollars
tolls_amount	bridge and tunnel tolls in dollars
total_amount	total paid amount in dollars

Following are the first ten lines from the data file:

```
07290D3599E7A0D62097A346EFCC1FB5,E7750A37CAB07D0DFF0AF7E3573AC141,2013-01-01 00:00:00,2013-01-01 00:02:
00,120,0.44,-73.956528,40.716976,-73.962440,40.715008,CSH,3.50,0.50,0.50,0.00,0.00,4.50

22D70BF00EEB0ADC83BA8177BB861991,3FF2709163DE7036FCAA4E5A3324E4BF,2013-01-01 00:02:00,2013-01-01 00:02:
00,0,0.00,0.000000,0.000000,0.000000,0.000000,CSH,27.00,0.00,0.50,0.00,0.00,27.50

0EC22AAF491A8BD91F279350C2B010FD,778C92B26AE78A9EBDF96B49C67E4007,2013-01-01 00:01:00,2013-01-01 00:03:
00,120,0.71,-73.973145,40.752827,-73.965897,40.760445,CSH,4.00,0.50,0.50,0.00,0.00,5.00

1390FB380189DF6BBFDA4DC847CAD14F,BE317B986700F63C43438482792C8654,2013-01-01 00:01:00,2013-01-01 00:03:
00,120,0.48,-74.004173,40.720947,-74.003838,40.726189,CSH,4.00,0.50,0.50,0.00,0.00,5.00

3B4129883A1D05BE89F2C929DE136281,7077F9FD5AD649AEACA4746B2537E3FA,2013-01-01 00:01:00,2013-01-01 00:03:
00,120,0.61,-73.987373,40.724861,-73.983772,40.730995,CRD,4.00,0.50,0.50,0.00,0.00,5.00

5FAA7F69213D26A42FA435CA9511A4FF,00B7691D86D96AEBD21DD9E138F90840,2013-01-01 00:02:00,2013-01-01 00:03:
00,60,0.00,0.000000,0.000000,0.000000,0.000000,CRD,2.50,0.50,0.50,0.25,0.00,3.75

DFBFA82ECA8F7059B89C3E8B93DAA377,CF8604E72D83840FBA1978C2D2FC9CDB,2013-01-01 00:02:00,2013-01-01 00:03:00,60,0.39,
-73.981544,40.781475,-73.979439,40.784386,CRD,3.00,0.50,0.50,0.70,0.00,4.70

1E5F4C1CAE7AB3D06ABBDDD4D9DE7FA6,E0B2F618053518F24790C7FD0264E302,2013-01-01 00:03:00,2013-01-01 00:04:00,60,0.00,
-73.993973,40.751266,0.000000,0.000000,CSH,2.50,0.50,0.50,0.00,0.00,3.50
```

```
468244D1361B8A3EB8D206CC394BC9E9,BB899DFEA9CC964B50C540A1D685CCFB,2013-01-01 00:00:00,2013-01-01 00:04:
00,240,1.71,-73.955383,40.779728,-73.967758,40.760326,CSH,6.50,0.50,0.50,0.00,0.00,7.50
```

```
5F78CC6D4ECD0541B765FECE17075B6F,B7567F5BFD558C665D23B18451FE1FD1,2013-01-01 00:00:00,2013-01-01 00:04:
00,240,1.21,-73.973000,40.793140,-73.981453,40.778465,CRD,6.00,0.50,0.50,1.30,0.00,8.30
```

The data file is sorted chronologically according to the dropoff_datetime. Events with the same dropoff_datetime are in random order. Please note that the quality of the data is not perfect. Some events might miss information such as drop off and pickup coordinates or fare information. Moreover, some information, such as, e.g., the fare price might have been entered incorrectly by the taxi drivers thus introducing additional skew. Handling of data quality issues is outside of the scope of this year's Grand Challenge.

Please follow this link to access the data file comprising the first twenty days (roughly 2 million events) of data. The file is approximately 130 MB in size.

Please follow this link to access the full data file containing data for the whole year 2013. The file is approximately 12 GB in size and contains approximately 173 million events.

Query 1: Frequent Routes

The goal of the query is to find the top 10 most frequent routes during the last 30 minutes. A route is represented by a starting grid cell and an ending grid cell. All routes completed within the last 30 minutes are considered for the query. The output query results must be updated whenever any of the 10 most frequent routes changes. The output format for the result stream is:

```
pickup_datetime, dropoff_datetime, start_cell_id_1, end_cell_id_1, ... , start_cell_id_10, end_cell_id_10, delay
```

where pickup_datetime, dropoff_datetime are the timestamps of the trip report that resulted in an update of the result stream, start_cell_id_X the starting cell of the Xth-most frequent route, end_cell_id_X the ending cell of the Xth-most frequent route. If less than 10 routes can be identified within the last 30 min, then NULL is to be output for all routes that lack data.

The attribute “delay” captures the time delay between reading the input event that triggered the output and the time when the output is produced. Participants must determine the delay using the current system time right after reading the input and right before writing the output. This attribute will be used in the evaluation of the submission.

The cells for this query are squares of 500 m X 500 m. The cell grid starts with cell 1.1, located at 41.474937, -74.913585 (in Barryville). The coordinate 41.474937, -74.913585 marks the center of the first cell. Cell numbers increase towards the east and south, with the shift to east being the first and the shift to south the second component of the cell, i.e., cell 3.7 is 2 cells east and 6 cells south of cell 1.1. The overall grid expands 150km south and 150km east from cell 1.1 with the cell 300.300 being the last cell in the grid. All trips starting or ending outside this area are treated as outliers and must not be considered in the result computation.

Query 2: Profitable Areas

The goal of this query is to identify areas that are currently most profitable for taxi drivers. The profitability of an area is determined by dividing the area profit by the number of empty taxis in that area within the last 15 minutes. The profit that originates from an area is computed by calculating the median fare + tip for trips that started in the area and ended within the last 15 minutes. The number of empty taxis in an area is the sum of taxis that had a drop-off location in that area less than 30 minutes ago and had no following pickup yet.

The result stream of the query must provide the 10 most profitable areas in the subsequent format:

```
pickup_datetime, dropoff_datetime, profitable_cell_id_1, empty_taxies_in_cell_id_1, median_profit_in_cell_id_1,
profitability_of_cell_1, ... , profitable_cell_id_10, empty_taxies_in_cell_id_10, median_profit_in_cell_id_10,
profitability_of_cell_10, delay
```

with attribute names containing cell_id_1 corresponding to the most profitable cell and attribute containing cell_id_10 corresponding to the 10th most profitable cell. If less than 10 cell can be identified within the last 30 min, then NULL is to be returned for all cells that lack data. Query results must be updated whenever the 10 most profitable areas change. The pickup_datetime, dropoff_datetime in the output are the timestamps of the trip report that triggered the change.

The attribute “delay” captures the time delay between reading the input event that triggered the output and the time when the output is produced. Participants must determine the delay using the current system time right after reading the input and right before writing the output. This attribute will be used in the evaluation of the submission.

Note: We use the same numbering scheme as for query 1 but with a different resolution. In query two we assume a cell size of 250m X 250m, i.e., the area to be considered spans from cell 1.1 to cell 600.600.