# Final Project

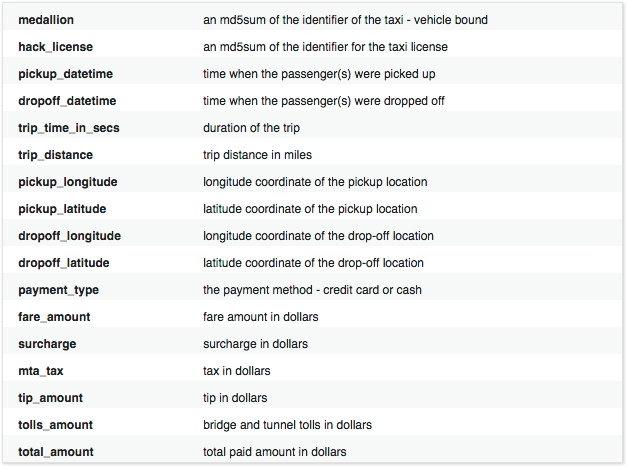
## Objective

Implement an application that is able to “answer” queries on both real-time and historical data.

You have to answer only 1 of the 3 questions asked below!!!

## Data Model

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:



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.

#### 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.

#### Query 3: Frequent Routes by week day

The goal of the query is to find the top 10 most frequent routes by week day. As query 1 a route is represented by a starting grid cell and an ending grid cellThe output format for the result stream is:

Weekday, pickup\_datetime, dropoff\_datetime, start\_cell\_id\_1, end\_cell\_id\_1, ... , start\_cell\_id\_10, end\_cell\_id\_10,

Where Weekday can be: Monday, Tuesday, etc.

Offer an option to enter the date range to query.

Download data from <https://www.dropbox.com/s/o1wqgv713n2k1kn/sorted_data.csv.gz?dl=0>

**FAQ**

**Question 1:** Is the earth flat or how to map coordinates to cells?

Answer: For the challenge we allow a simplified flat earth assumption for mapping coordinates to cells in the queries. You can assume that a distance of 500 meter south corresponds to a change of 0.004491556 degrees in the coordinate system. For moving 500 meter east you can assume a change of 0.005986 degrees in the coordinate system.

**Question 2:** What is a change of the top 10 list? Only when there is a new entry or also when the order changes?

Answer: Also when the order changes

**Question 3:** What happens to cells that do not have empty taxies and cause a decision by zero for query 2?

Answer: Results for this situation are undefined and should not be considered for calculating the output.

**Question 4:** Can you give us a concrete example of how a correctly formatted output event should look like?

Yes:

For Query 1:

2013-01-01 00:01:00,2013-01-01 00:02:00,1.7,2.9,10.3,20.27,40.3,30.27,40.3,30.28,1.1,1.2,120.1,150.2,200.1,150.2,120.1,170.2,10.1,15.2,11.1,15.7,1001

For Query 2:

2013-01-01 00:01:00,2013-01-01 00:02:00,1.1,2,50.10, ,10.1,2,40.10,,101.11,4,50.10,111.1,4,40.00,51.1,7,30.00,,102.1,10,31.10,12.1,10,31.09,1.31,10,30.00,1.122,10,30.00,211.11,11,30.00,1002

Note: the concrete numbers are toy examples to illustrate the formatting

**Question 5:** Should the queries run simultaneously during the evaluation or one after the other?

Answer: The queries must run simultaneously. However, the output must got two separated streams (i.e. end up in two different files for evaluation)

**Question 6:** In query 2 we must factor in the number of empty taxis in an area. This is defined as taxies that recently dropped someone off in that area and had no following pickup yet. However, since a trip is reported at its end, how should we know about taxies that had a pickup but are still driving with that customer?

Answer: It is correct to not account for trips that have not been reported yet. The system cannot know about these events and hence the queries cannot factor them in.

**Question 7:** How should we order elements in a list that have the same value for the ordering criterion?

Answer: You should always put the freshest information first. E.g. if route A and B have the same frequency, put the route with the freshest input information fist (i.e. the one which includes the freshest event).

**Question 8:** There are some negative values for fare\_amount and tip\_amount. Are these errors? How should we handle these values?

Answer: The negative values are errors. They should not be considered in any calculation. However, you can still use the corresponding events for calculations that do not need the fare\_amount and tip\_amount.

**Question 9:** Should we literally print the ASCII word "NULL" or should it be the NULL string (i.e., zero characters)?

Answer: Please write the ASCII word "NULL".

**Question 10:** How to update the top 10 list when it is changed due to events that leave the window?

Answer: For updates that are caused by events leaving the window, you should use the timestamp that denotes when the result changes. E.g. in a 15 min window, that would be the timestamp of the relevant leaving event + 15min.

**Question 11:** About the delay measurements. Should we At which exact point should the "delay" attribute start "counting"?

(a) Since the moment the line is read from disk?

(b) Since the moment the line \*starts\* being parsed?

(c) Since the moment the line \*has been\* parsed?

Answer: (b) is the correct option. We do not want to capture the disk I/O in the delay and want to be agnostic to all buffering that might take place before the actual event processing starts. However, we consider parsing as part of the processing, because it and may be done in different ways. Hence, parsing should be captured by the delay measurements. Your system should have a clearly defined "entry point" where it receives the events as they are encoded in the file. The measurement is to be taken once the raw string encoded event passed that point. Please specify your approach to delay measurements in your corresponding paper submission.

**Question 12:**

At which exact point should the "delay" attribute stop "counting"?

(a) At the moment the top-10 list has been \*computed\* (but before being converted to an ASCII string)?

(b) At the moment the top-10 list has been converted to ASCII (but before the actual call to fprintf/write)?

(c) After the fprintf/write has returned?

Answer: Again, it is not our intention to capture disk I/O. However, creating the output in the right format is considered part of the processing. Thus, (b) is the right option.

**Question 13:** How do you ensure that the delay measurements are implemented in an appropriate way?

Answer: Please describe how you implemented the delay measurements in the corresponding paper.

**Question 14:** How are you going to test correctness, given that the precise output is very sensitive to minor things, like floating point precision errors in coordinate-to-cell conversion, etc.

Answer: Our goal is to have solutions that are similar enough to compare their performance. We have not seen the result files yet, thus it is too early to give a final answer. However, it is not our intent to disqualify solutions due to minor deviations that e.g. might be the result of rounding errors. What we want to avoid with the correctness requirement is that solutions use approximations or other ways of simplifying the actual computational problem.

**This project is based on the challenge below.**

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