Liz Aharonian – 316584960
Ori Ben Zaken - 311492110

# Final Project – Report 2

## Milestone #1 : Find relevant data (Semester A)

- After examining different taxi trips datasets we decided to take our data from 2014_Uber_Data, can be found here.
- Our dataset contains records of ~360K uber taxis trips. Each record contains the following details: pickup date and time, pickup latitude, pickup longitude and region code.

## Milestone #2: Data cleansing (Semester A)

- We performed full data cleansing that includes: removing records with Null values and removing columns that more than half to their values are Null or empty. In our case, no column was removed.
- Our model target is to predict the demand for taxi per time interval and area. In order to do so we added the “demand” column. We counted all of the taxi trips that occurred in a certain region and certain 10-minutes time interval. The result was added to each of relevant trip records.
- Added more features in order to help the prediction. For each record we added the features: is_Holiday, is_Weekend according to the trip date.
- In addition, in order to prepare the data for machine learning model we converted all the string values to numerical values.

## Milestone #3: Training ML model (Semester A + B)

- We chose randomly 100K trip records as our dataset for the model. Splitting of course the dataset to training set and validation set.
- We tried many regression models, reached so far to the best results with Random Forest Regression model.
- These days we are also examining neural models to see if you can get even better results.
- The model is a spark model which can perform it's work and computation in distributed system.

## Milestone #4: Data streaming (Semester B)

- Using Apache Kafka in order to stream the data in real-time.
- Kafka aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds, making it highly valuable for enterprise infrastructures to process streaming data.

## Milestone #5: running the model (Semester B)

- Using trained serialized model in order to predict demand per time interval and area on new examples which hasn't been seen by the model.

Liz Aharonian – 316584960
Ori Ben Zaken - 311492110

## Milestone #6: visualization of the model (Semester B)

- Using Grafana in order to present the model.
  Grafana is an open-source, general purpose dashboard and graph composer.
- We are going to use this product in order to present the model predictions statistics and achievements.

## Reporting hours:

In semester A, our working day on the project was Sunday. We sat every Sunday, usually from 9am-7pm, sometimes with our mentors at Gigaspaces, Hertzeliya.

- Tutorials for AWS tools
  - 28/10/2018
  - 4/11/2018
  - 11/11/2018
- Find relevant data
  - 18/11/2018
  - 25/11/2018
- Data cleansing
  - 2/12/2018
  - 9/12/2018
  - 16/12/2018
- Training ML model
  - 16/12/2018
  - 23/12/2018
  - 30/12/2018
  - **06/01/2019**
  - **13/01/2019**

In semester B our working day on the project is Tuesday. We seat every Tuesday, usually from 12am-9pm.

- Data Streaming
  - 05/03/2019
  - 12/03/2019
  - 19/03/2019
  - 26/03/2019
- Running The Model
  - 09/04/2019

    - 16/04/2019
    - 23/04/2019
- Visualization Of The Model
    - 30/05/2019

In addition, we also work sometimes after our working days remotely in the beginning of the week.

Since we started our work on the project, each of us spent 240 hours.