316584960 Liz Aharonian בס"ד
311492110 Ori Ben-Zaken

Project Design

Background:

Goal: demand per time interval and area for Uber Taxi Ride.

In cases of very high demand, fares may increase to help ensure those who need a ride can get one. Some riders will choose to pay, while some will choose to wait.

Our goal is to predict demand per time interval for Uber Taxi Ride, using ML and big data processing tools.

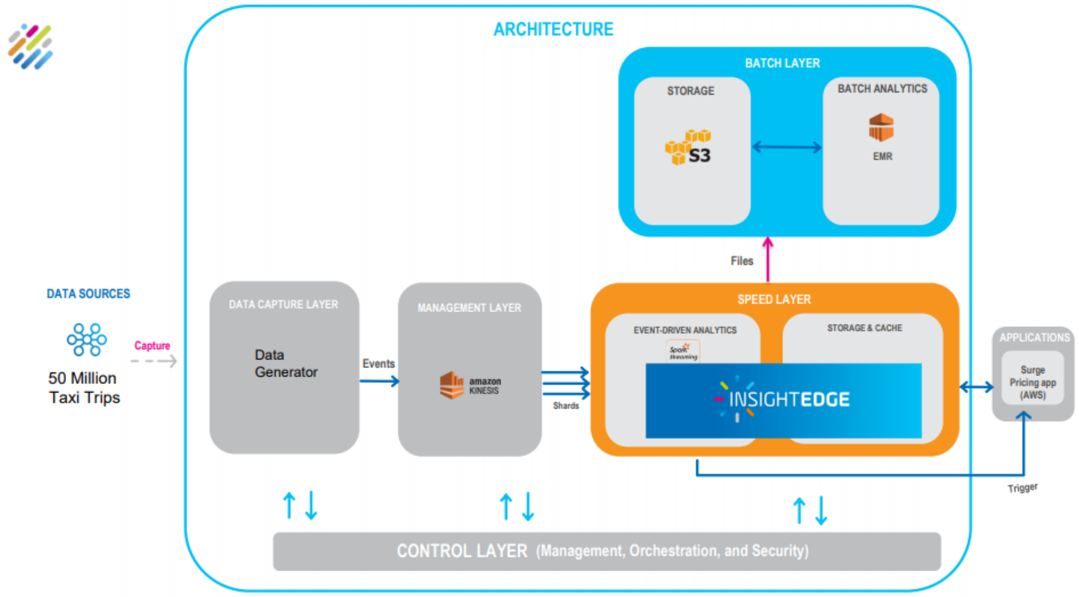
The demand parameter is calculated according to time intervals and area, this gives two major advantages to our predicting model:

1. Help the model's users to know which area they should give service.
2. Using a simple function based on the demand to predict and calculate the fare amount.

Technologies:

- Amazon SageMaker (Data Science)
- Amazon Kinesis (Streaming)
- Amazon S3 (Storage)
- Machine Learning (Apache Spark)
- Interactive Development (Apache Zeppelin)
- Visualization (Grafana)
- BI (Tableau)

Architecture:



Phase 1 – find relevant data – Done.

Relevant data sources:

- Google dataset search - https://toolbox.google.com/datasetsearch
- Kaggle - https://www.kaggle.com/datasets
- https://github.com/toddwschneider/nyc-taxi-data/blob/master/setup_files/raw_2014_uber_data_urls.txt

Phase 2 – data cleansing – Done.

Using pandas package in python for cleaning the data and add features:

- Remove rows which contain null values.
- Remove cols which more than half of their values are nulls or empty.
- Adding relevant cols\features:
    - Day\month\year, is_holiday, is_weekand,demand_per_time_interval_and_area.
- Convert string values to numerical values.
- Drop irrelevant cols.
- Recognize features with high correlation and use one of them using sklearn tool kit.

Phase 3 – training model – due date: 01\01\19.

- Use sklearn package in order to train regression model. Our goal is to predict demand per time interval and area.
- Evaluate the model – calculate the loss, change hyper params and examine different models to find the best model.
- Saving endpoint of the model so that we could use it to make predictions in real time.

16\01\19 – submit report 1.

Phase 4 – data streaming - due date: 01\02\19.

Using Apache Kafka and InsightEdge (Gigaspaces product) in order to stream the data.

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.

Phase 5 – running the model - due date: 01\03\19.

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

Phase 6 – visualization of the model - due date: 1\04\19.

Using Grafana and Tableaue (BI) in order to present the model.
Grafana is an open-source, general purpose dashboard and graph composer.

Tableaue is an interactive data visualization product focused on business intelligence.

We are going to use these products in order to present the model predictions statistics and achievements.

Phase 7 – Packaging the model - due date: 01\05\19.

Package our model and all its dependencies using Docker, a computer program that performs operating-system-level virtualization.

We are going to use it, so that our project would not be depend on specific operating system.

03\05\19 – submit report 2.

21\06\19 – submit final report.