

15-319 / 15-619

Cloud Computing

Recitation 14

April 21st 2020

Overview

- **Last week's reflection**
 - Project 4.2
- **This week's schedule**
 - **Twitter Analytics: The Team Project**
 - Phase 3 Live Test

P4.2 - Taxi Fare Prediction Application

Accepts speech queries to get the cab fare estimate for going from point A to point B (based on historical data), and returns the result as speech



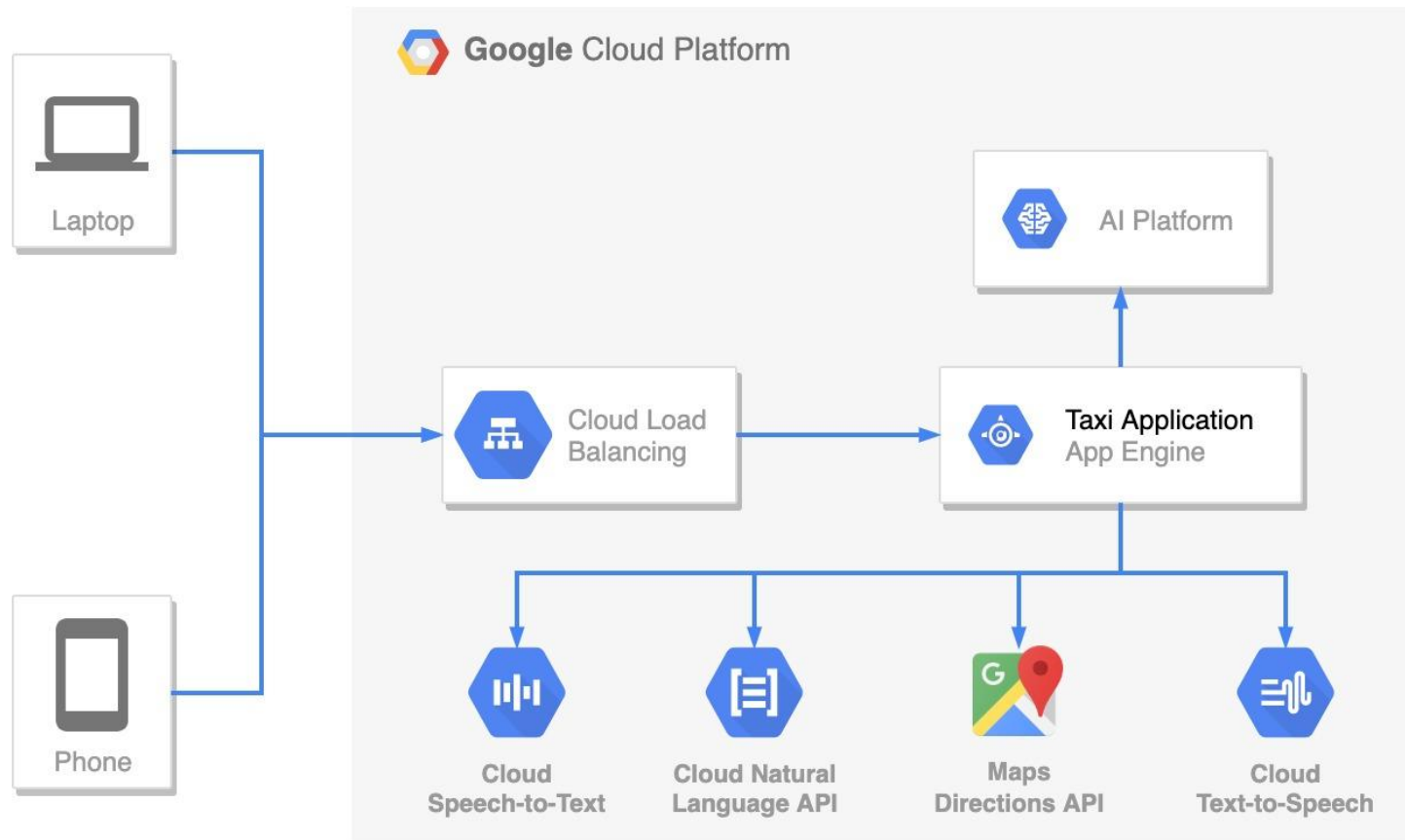
*I would like to get from Central Park
Zoo to Grand Central Terminal*



*Your expected fare from Central Park
Zoo to Grand Central Terminal is \$29.69*

P4.2 - ML Application Pipeline

Build an end-to-end application pipeline to predict car fare requests using the following architecture.



P4.2 - Reflection

- The quality of the raw data used to train machine learning models impacts the accuracy of the models.
- It is important to clean the raw data to remove the outliers from a dataset before training a model.
- Feature engineering is critical to the accuracy of ML models.
 - Inspect and visualize the dataset to select discriminating features.
 - Constructing new features improves the accuracy of an ML-based predictor.
 - Transform your domain knowledge to the feature engineering.
- Cloud-based Machine Learning services, Google AI Platform, provide an easy and affordable option to train and deploy your own ML model.
- Hyperparameter tuning on Google AI Platform improves predictor accuracy.
- Cloud ML services enable developers to build and deploy sophisticated AI-based applications without having the underlying ML knowledge.

TEAM PROJECT

Twitter Data Analytics



Phase 3 Live Test on Sunday 4/26

Time	Task	Description
4:00 pm	DNS Submission	Submit your DNS for the Live Test before the deadline
5:30 pm - 5:31 pm	DNS Validation	Validate your DNS. This is the last chance to update your DNS for the Live Test

Time	Value	Target	Weight
6:00 pm - 6:30 pm	Warm-up (Q1 only)	0	0%
6:30 pm - 7:00 pm	Q1	35000	15%
7:00 pm - 7:30 pm	Q2	12000	25%
7:30 pm - 8:00 pm	Q3	6000	25%
8:00 pm - 8:30 pm	Mixed (Q1,Q2,Q3)	12000/3000/2000	5+5+5 = 15%

Phase 3: Summary

- No new queries (we test Q1, Q2, Q3)
- Use AWS Managed Services
- Hourly budget is \$1.28
- 10% penalty if total spending > \$100
- 100% penalty if total spending > \$150
- EC2 / EBS not allowed
 - See resources in the EC2 dashboard? NOT ALLOWED
<https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Instances:>
<https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Volumes:>
 - Install database via command line? NOT ALLOWED

Phase 3: Summary

- RPS targets have been increased →
 - Q1: 35000
 - Q2: 12000
 - Q3: 6000
 - Mixed: 12000/3000/2000
- Terraform is required if possible (on Github)

Hints:

- Do mock Live Test by yourselves, including mix queries
- Monitor CPU credits and EBS credits

Upcoming Deadlines



- Team Project Phase 3
 - **Live-test:** Sunday, Apr. 26, 2020 5:59 PM ET
 - **Code and report due:** Tuesday, Apr. 28, 2020 11:59 PM ET