15-319 / 15-619 Cloud Computing

Recitation 13

November 19th 2019

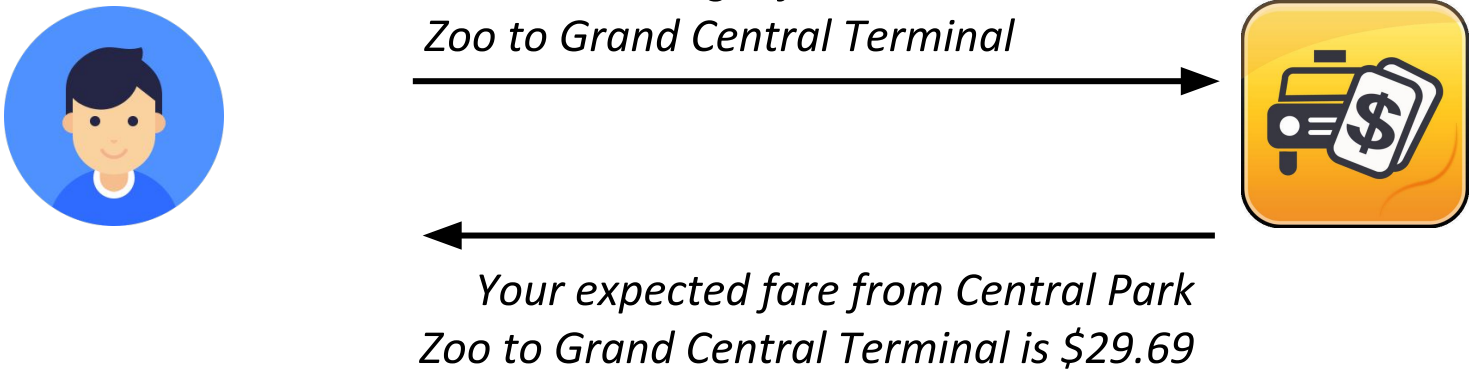
Overview

- **Last week's reflection**
  - Project 4.2
  - Quiz 11

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

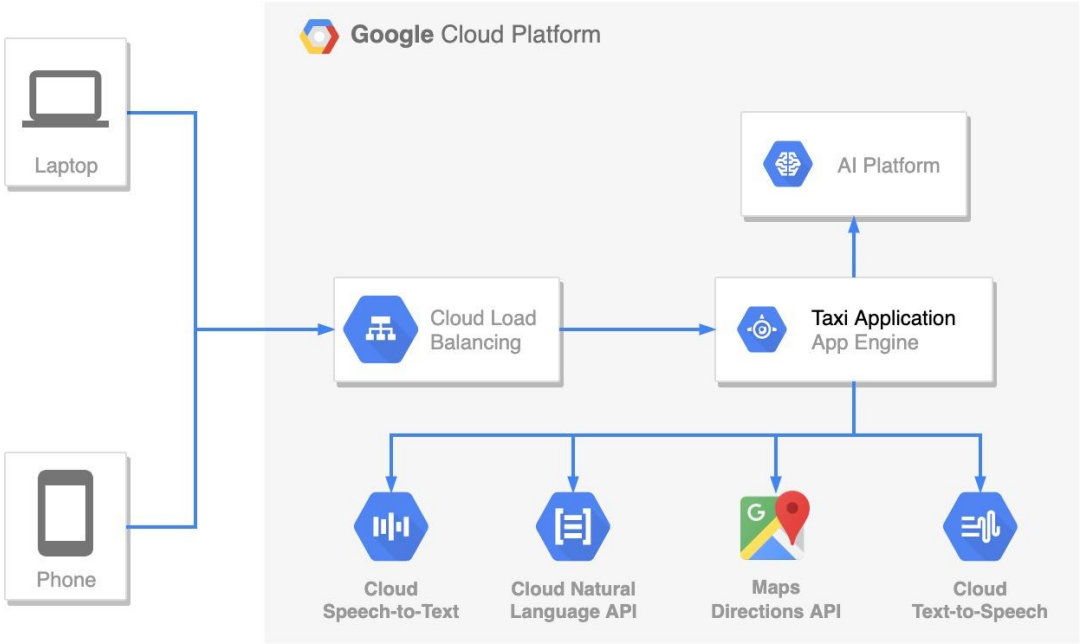
P4.2 - Taxi Fare Prediction Application

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



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 incorrect/corrupt values 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.
  - Different features impact the accuracy of the predictor.
- Cloud-based Machine Learning services, Google AI Platform, provide an easy and affordable option to run large machine learning workloads.
- 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 11/24**

Submit DNS for Live Test

Information

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 Database Tier DNS. This is the last chance to update your DNS for the Live Test

Phase 3 Live Test on **Sunday 11/24**

Live Test

Information

Time	Value	Target	Weight
6:00 pm - 6:30 pm	Warm-up (Q1 only)	0	0%
6:30 pm - 7:00 pm	Q1	38000	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 Reads(Q1,Q2,Q3)	15000/3000/2000	5+5+5 = 15%

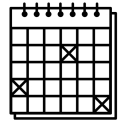
Phase 3: Summary

- No new queries (we test Q1, Q2, Q3)
  - 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:
  - sudo apt install mysql-server NOT ALLOWED

Phase 3: Summary

- RPS targets have been changed →
  - Q1: 38000
  - Q2: 12000
  - Q3: 6000
  - Mixed: 15000/3000/2000
- Terraform is required if possible (on Github)

Team Project Time Table



Phase (and query due)	Start	Deadlines	Code and Report Due
Phase 1 • Q1, Q2	Monday 10/07/2019 00:00:00 ET	**Checkpoint 1, Report: Sunday 10/13/2019 23:59:59 ET** **Checkpoint 2, Q1: Sunday 10/20/2019 23:59:59 ET** **Phase 1, Q2: Sunday 10/27/2019 23:59:59 ET**	Phase 1: Tuesday 10/29/2019 23:59:59 ET
Phase 2 • Q1, Q2,Q3	Monday 10/28/2019 00:00:00 ET	Sunday 11/10/2019 15:59:59 ET	
Phase 2 Live Test (Hbase **AND** MySQL) • Q1, Q2, Q3	Sunday 11/10/2019 17:00:00 ET	Sunday 11/10/2019 23:59:59 ET	Tuesday 11/12/2019 23:59:59 ET
Phase 3 • Q1, Q2, Q3 (Managed services)	Monday 11/11/2019 00:00:00 ET	Sunday 11/24/2019 15:59:59 ET	
Phase 3 Live Test • Q1, Q2, Q3 (Managed services)	**Sunday 11/24/2019 17:00:00 ET**	**Sunday 11/24/2019 23:59:59 ET**	**Tuesday 11/26/2019 23:59:59 ET**

Upcoming Deadlines

- Team Project Phase 3
  - **Live-test:** Sunday, November 24, 2019 3:59 PM ET
  - **Code and report due:** Tuesday, November 26, 2019 11:59 PM ET

Questions?