

Computer Science Department  
Columbia University

Curricular Practical Training Report


Name: Sankalp Prakash Apharande

UNI: spa2138

Company Name: Walmart Global Tech

Supervisor Name: RICHARD WILLIAMSON

Supervisor's Email: [Richard.Williamson@walmart.com](mailto:Richard.Williamson@walmart.com)

Supervisor's Signature: 

# Internship Report

**Role:** SDE-III Intern

**Company:** Walmart Global Tech

**Location:** Bentonville, AR

**Team:** End-to-End (E2E) Applied AI Architecture Team

## Summary:

### About the Company and my team:

I worked as an SDE-III intern at Walmart Global Tech at their Headquarters in Bentonville, AR, from May 31, 2022 to Aug 12, 2022. Walmart Global Tech uses cutting-edge technologies that create unique and innovative experiences for our associates, customers, and members across Walmart, Sam's Club, and Walmart International.

I was a part of the E2E Applied AI Architecture team, whose primary responsibility included enabling AI across the company focusing on ML engineering to support data scientists by developing end-to-end ML Pipelines.

### My role in the team:

My two projects were designed to carry out the following requirements:

1. Extending our ML Framework to the Pricing team of data scientists
2. Development of a scheduled job to keep track of unused long-running GCP resources to eliminate unnecessary daily server cost

### Project Scope:

I worked on two projects during my internship;

1. Development of ML Model deployment service for Pricing team of data scientists
  - a. For seamless Continuous Integration and Continuous Deployment (CICD) of their models
  - b. Developing a new model deployment service with Google's Vertex AI cloud platform.
2. Developing the module to keep track of unused long-running GCP resources to reduce excessive Cloud Spend.

### Technologies Used:

I primarily worked on developing modules in Python using GCP Vertex AI, GCP Cloud Storage, GCP BigQuery, and GCP DataProc and Compute Engine. I also used Docker to deliver software in packages called containers.

## Project 1: Trained Model Deployment Service

### Problem Statement:

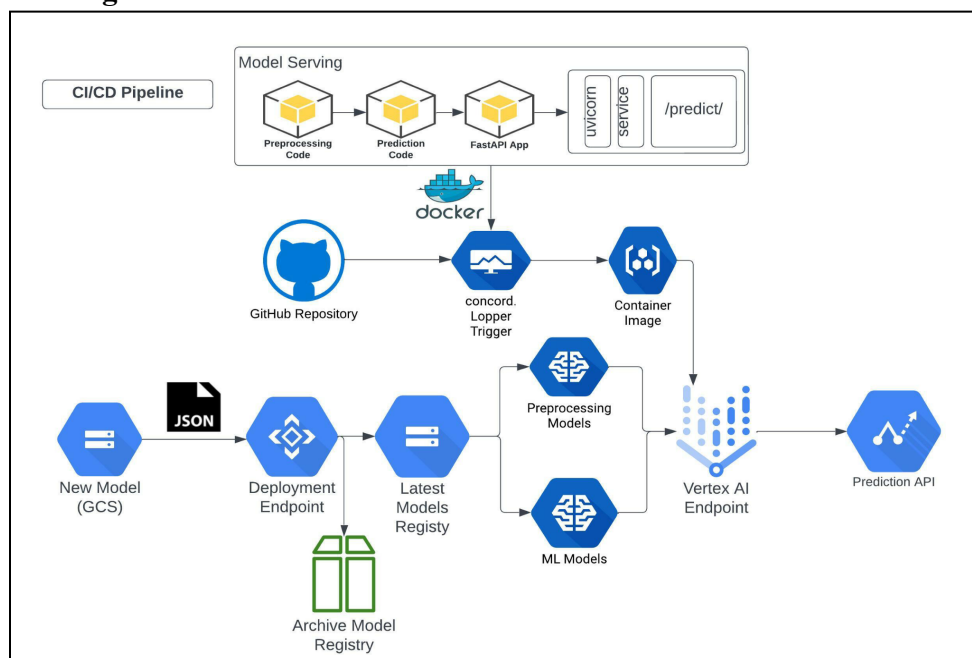
1. To expand the coverage of price anomaly detection of items on Walmart Marketplace, a team of data scientists develops and frequently trains the anomaly detection models for items without sufficient price history.
2. But current infrastructure does not support the rapid deployment of these models.

As an ML engineer, my task was to develop a service to deploy these models on a highly available, highly scalable, low latency endpoint for online predictions.

### Requirements:

1. Service to easily deploy Real-Time custom-trained models on GCP Vertex AI Endpoint.
2. Ensure efficient and automated CICD of this endpoint.
3. Ensure low latency, high availability, and high scalability of this endpoint.
4. Support custom preprocessing on input requests.
5. The endpoint should support 100,000 requests per day.

### Architecture Diagram:



*Fig. 1: Model Deployment Service Architecture Diagram*

1. CI/CD Pipeline to create a custom container image containing the code for preprocessing and prediction
2. A GitHub trigger to automate container image creation and Vertex AI Endpoint creation with custom container image.
3. Vertex AI Endpoint fetches two newly trained ML models from the latest models GCS bucket used for preprocessing model and prediction.
4. Deployment Endpoint takes input GCS bucket containing new models and deploys these models into two GCS Buckets; latest models registry and archive model registry.

## Results:

GCP Vertex AI Endpoint ensures High Availability, High Scalability, Low Latency, Low Cost

**High Scalability:** The following graphs show the performance of Vertex AI Endpoint handling 1000 concurrent requests

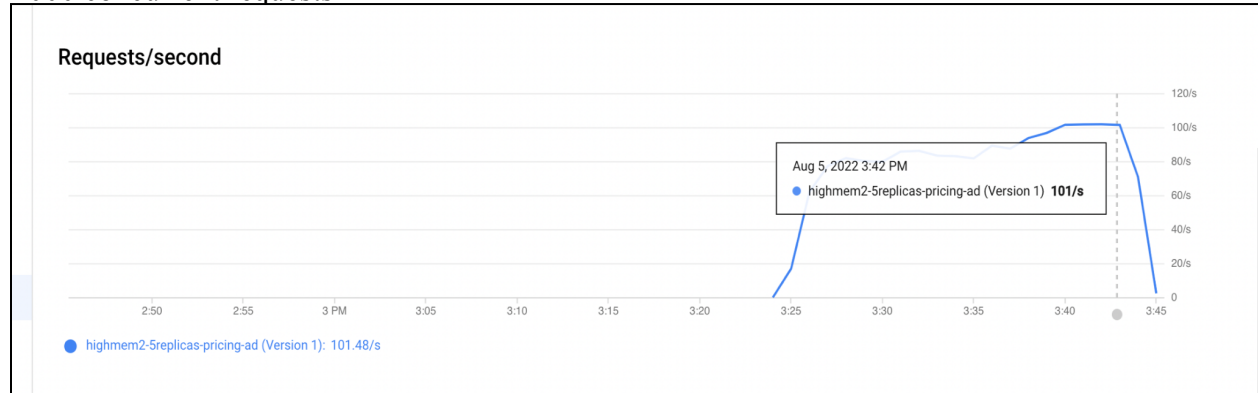


Fig. 2: Requests/second handled by Vertex AI Endpoint

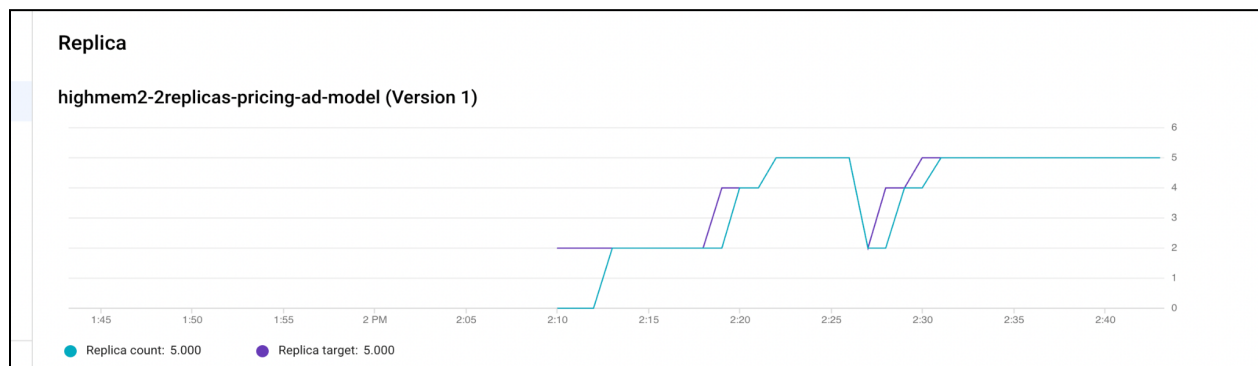


Fig. 3: Autoscaling: Number of replicas as per the incoming traffic

**Low Latency (Time/Req) Using Vertex AI Endpoint:**

- 50% of 100,000 requests: 131.81ms
- Time taken for 100,000 requests: 19.18 mins

**Low Cost: \$2.91 per day**

n1-highcpu-32 machine: **\$1.303** per node per hour

Peak replicas used: 7; 100,000 requests per day

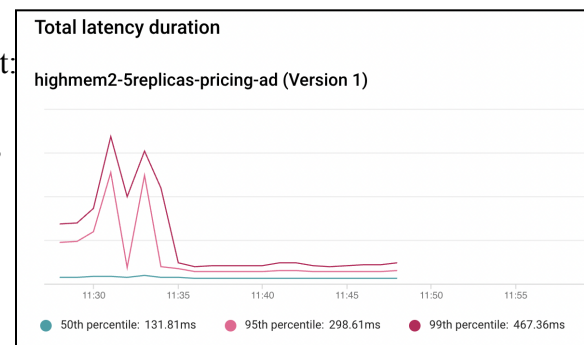


Fig.4: Total Latency for 100,000 requests

$$(\$1.303/\text{node hr}) \times (1 \text{ hr}/60 \text{ minutes}) \times (19.18 \text{ mins}) \times (7 \text{ nodes}) = \$2.91 \text{ per day}$$
  
i. e., \$1064.72 per year

## Project 2: Anomaly Detection of GCP VMs

### Problem Statement:

1. Most of the time, GCP resources, such as Compute Engine Instances, DataProc Clusters, DataProc Batch jobs, and DataProc Spark Sessions, aren't terminated after the use
2. These long-running GCP resources are often underutilized

In my second project, my task was to develop a Python module that could track underutilized resources and notify the team with results.

### Requirements:

1. Gather hourly average CPU usage time series data for all currently running Compute Engine Instances and DataProc Clusters using Google Cloud APIs.
2. Collect avg maximum executors usage hourly time series data for DataProc Batch Jobs.
3. Gather instance names, IDs, and the running duration for each resource
4. Detect underutilized GCP resources using a Simple Rule-Based Model.
5. Develop an Email alert module to send results of detected anomalies to the entire team

### Architecture Diagram:

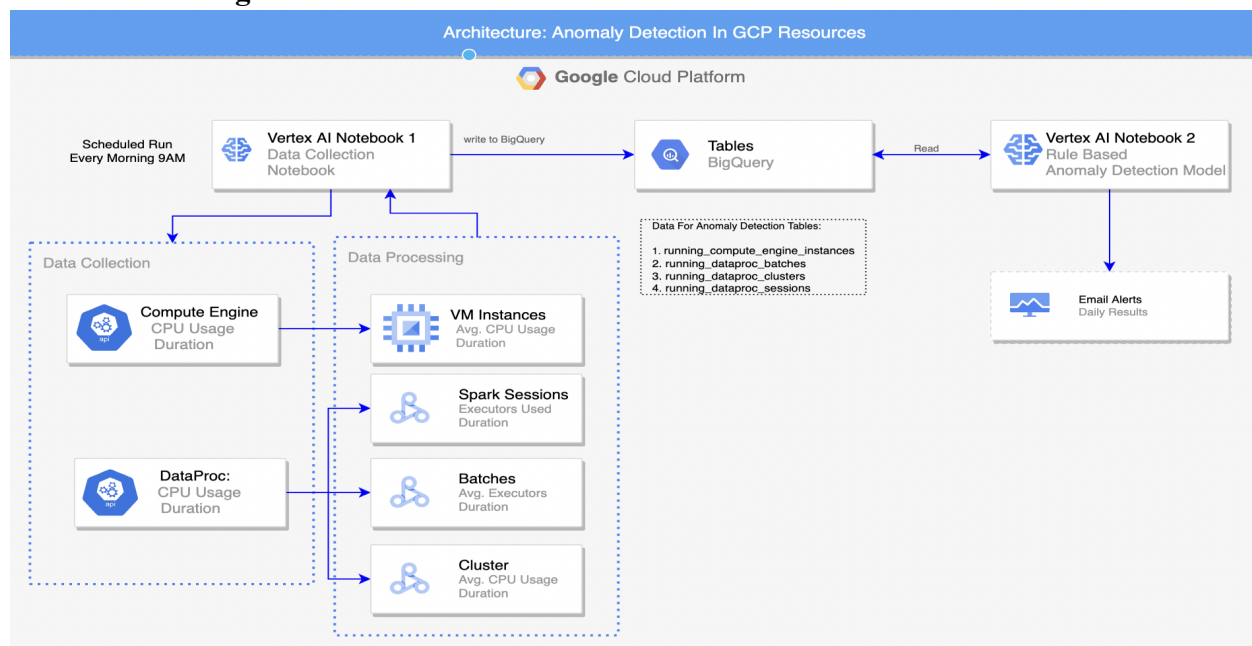


Fig.4: Anomaly Detection of GCP VMs Architecture Diagram

1. Leveraged GCP Vertex AI Infrastructure for data collection and data preprocessing of running GCP resources.
2. Processed results are written into BigQuery tables partitioned by date
3. VertexAI Notebook for detecting anomalous long-running GCP resources using a rule-based model
4. Vertex AI Notebook sends email alerts with detailed daily summary

### Results:

The developed notebook detected at least 120+ anomalous long-running GCP resources daily.

One Anomaly Example: A 12 node n1-standard-64 compute engine cluster was detected, which saved **\$875 daily**.

# **Future Work, Learnings, and Impact**

## **Future Scope of My Projects:**

### **Project 1: Trained Model Deployment Service**

1. Model Deployment Service can be extrapolated to serve other Data Science teams.
2. By further benchmarking, it's possible to support up to 10,000 concurrent requests.

### **Project 2: Anomaly Detection of GCP VMs**

1. Based on the type of the machine and the worker nodes in each cluster and hourly price, we can rank the anomalous long-running GCP resources from most to least expensive.
2. Implementing an unsupervised ML model on GCP resources data.

## **Learnings:**

1. Hands-on experience with cutting-edge cloud technologies:  
I used GCP Vertex AI Endpoint, Google Cloud Storage, and BigQuery, Compute Engine.
2. System Design and Architecture  
Both of my projects required me to develop end-to-end solutions from scratch. I learned to design end-to-end services, which I enjoy the most as Machine Learning Engineer.
3. Soft Skills
  - a. Stand-up Meetings and Weekly updates:  
I also learned how to work with remote team members effectively. To keep myself aligned with deliverables, I scheduled weekly update meetings.
  - b. Express Ideas concisely:  
As a software engineer, it's extremely important to express your engineering ideas and their implementation concisely. I learned how to communicate my ideas and questions to my team members effectively.
  - c. Networking events:  
Over ten weeks of my internship, I met many new people working for Walmart. I realized the importance of making connections with other people.

## **Relation with coursework at Columbia University and prior work experience**

1. E2E Applied AI team's work was highly aligned with my coursework as a computer science graduate student with a Machine Learning track
2. During my internship, I used a lot of cloud technologies, and my learnings from Cloud Computing and Big Data course were very useful
3. There debugging and documentation reading skills I learned while working as a Machine Learning Engineer for 2.5 years were immensely helpful in tackling the problems

## **Impact of my work on the organization as a whole**

1. During my project of Model Deployment Service, our team became one of the very first to use the full potential of GCP Vertex AI Endpoint for online prediction. This work can be a stepping stone to onboard many Data Science teams on Vertex AI.
2. Anomalous long-running GCP resources project directly helps the organization to reduce unnecessary daily server costs. This work will also serve as a blueprint for other organizations to reduce their daily server costs.