Carnegie Mellon University

Team 05

WESTWORLD

Khushi Bhuwania

Piyush Talreja

Tian Xie

Tirth Parekh

Varsha Reddy

Milestone: 1

- Main takeaways/learnings : collecting data, building ML models, building inference service, teamwork.
- Design choices in the milestone :

  1. What data features do we use and store. We had to keep in mind the storage and also what features are important for the model. We ended up storing features user_id, movie_id, rating, gender, age and occupation. We used 30M log entries for initial model. How we split the data for training and validation. Assumption was that the data is sequential in the kafka pipeline.

  2. Learned about various recommendation algorithms. Tried Collaborative filtering (user and item) and SVD. We went on with SVD based model as it performed better, had lower inference cost, fast training time and smaller model size.

  3. Team work: Learned about each other expertise and decided on where and when meetings will be held and what mode of communication will be used in the future.

What would we have done better? - Explore various other recommendations algorithms and stored data in databases instead of files. This would have made things easier in the future.

Carnegie Mellon University

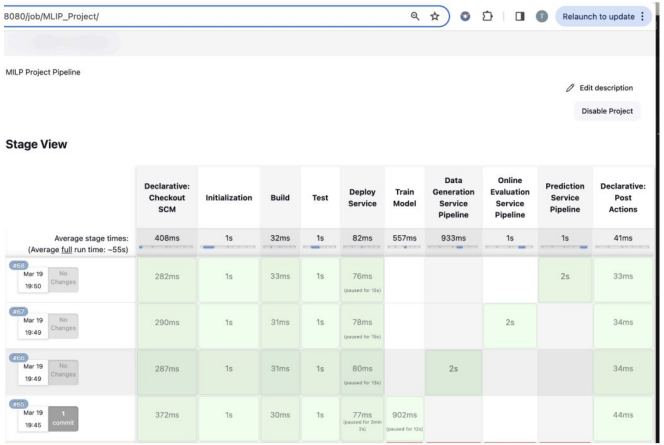
Milestone 2

Main takeaways:

- Importance of CI/CD pipeline, testing, evaluation, software quality, system availability and stability

Design choices:

- Post every commit, Jenkins pipeline was auto triggered, built and test cases were run to ensure sanity
- Since we used a microservices architecture, our Jenkins pipeline was configured to deploy the said service
- Our pipeline also automated the model updates, training model on latest data once a day



Could have done better:

- Include a rollback step in the pipeline to move back to an older version of the code with just one click

Carnegie Mellon University

Milestone 3

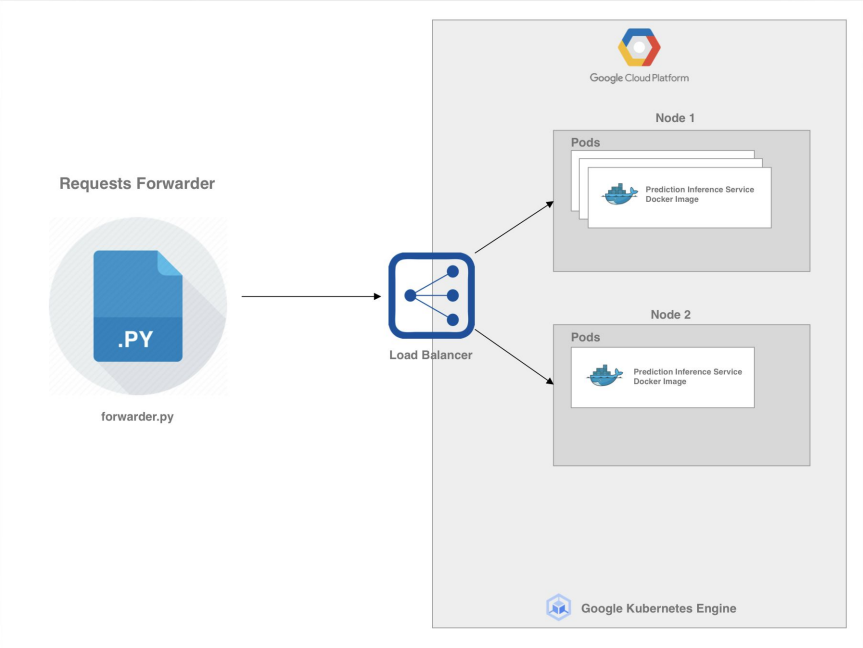
Problems with Local Deployment

- **Resource Limitations**
  - Limited CPU (utilization was above **95%**)
  - Storage
- **Maintenance Overhead**
  - Manual efforts (model updates, model switching)
- **Availability Concerns**
  - Not scalable

Solution?

Carnegie Mellon University

Deployed Inference Service on Cloud



- **Scalability**
  - Deployed our service across 2 nodes
  - Each node was configured to host multiple pods
  - Load balancer to evenly distribute load.
- **Zero-Downtime Deployment**s
  - Utilizing *Kubernetes* for zero downtime
  - Enabled us to maintain uninterrupted service availability even during updates, with the *ability to roll back* instantly to a stable release if necessary.
- **Availability: 99.8%**

Carnegie Mellon University

Monitoring

- **Decided to set up telegram bot**

Carnegie Mellon University

Milestone 4 - Fairness

- Fairness Focus
  - Prevent adult movie recommendations to children under 21.
- Dataset Filtering
  - Exclude underage viewers of adult-rated movies.
- Validation Checks
  - Ensure recommendations for children are age-appropriate.
- Design Improvements
  - Monitor and refine model to reduce unfair recommendations.
- Handling Exceptions
  - Automatically replace adult movies with suitable alternatives for underage viewers.

Carnegie Mellon University

Milestone 4 – Security

1. **STRIDE** threat modeling

2. Security Issue 1: Dos
   a. attackers send thousands of simultaneous requests to recommendation API
   b. retrieve user request frequency from Kafka stream log
   c. introduce a rate limiter, 60 requests per minute for each client

3. Security Issue 2: Review Bomb
   a. large number of users deliberately post negative reviews or low ratings for a particular movie to manipulate its perceived popularity
   b. monitor the `/rate` requests, detecting abnormal ratings for a certain movie
   c. Ratings and users involved during a review bomb will not be included in future model training

Carnegie Mellon University

Collaboration

- Bi-weekly meetings (Hybrid mode)
- Assigned tasks based on the strengths
- Openly communicated when stuck on a task for long time
- Sometimes we didn’t plan things well in advance, so sometimes we had to overload as we came closer to deadlines
- Last minute submissions

Carnegie Mellon University

THANK YOU!

Carnegie Mellon University