Style Synth:

ML Fashion Recommendation System

Courtney Green, Li-Wen Hu, Satomi Ito, Nandini Kodali, Sophia Rutman

Agenda

1 Introduction

2 Computer Vision

3 Recommendation Engine

4 Backend Web Server

5 Frontend

6 Monitoring

7 Demonstration

1

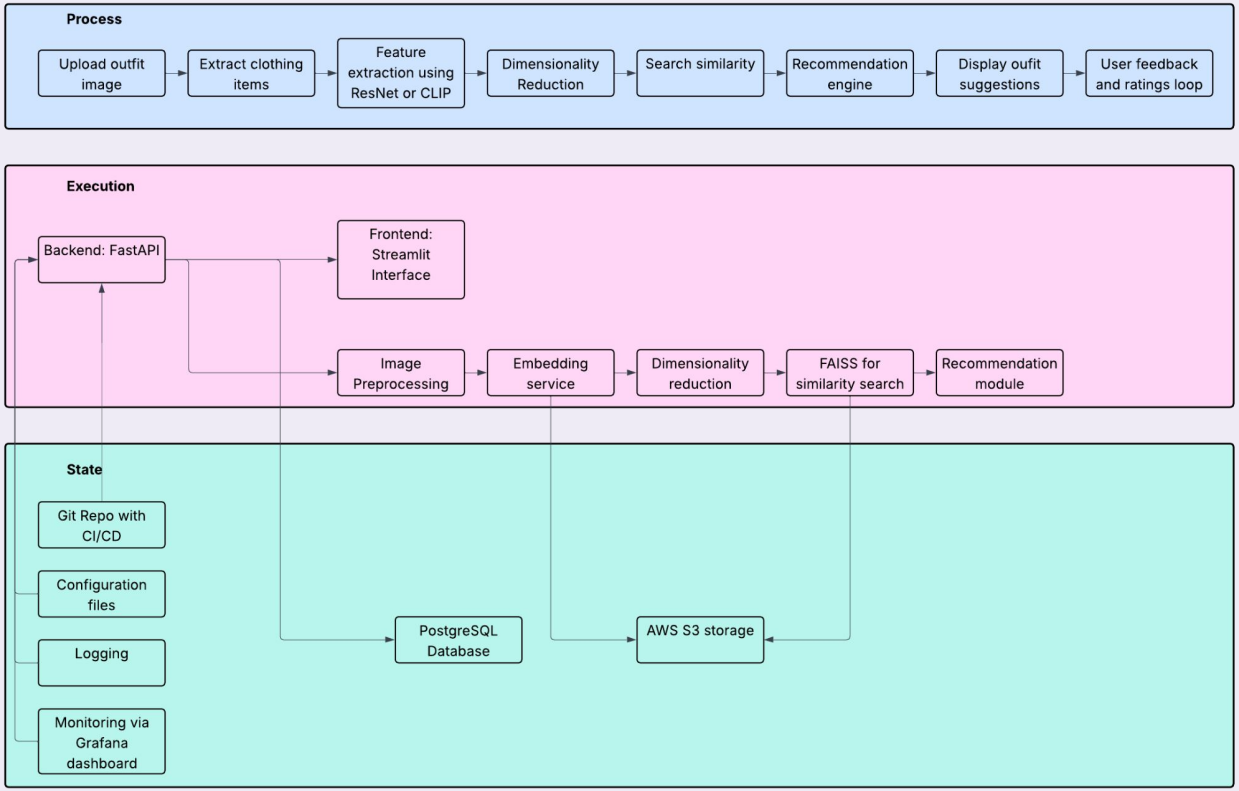
Introduction

2

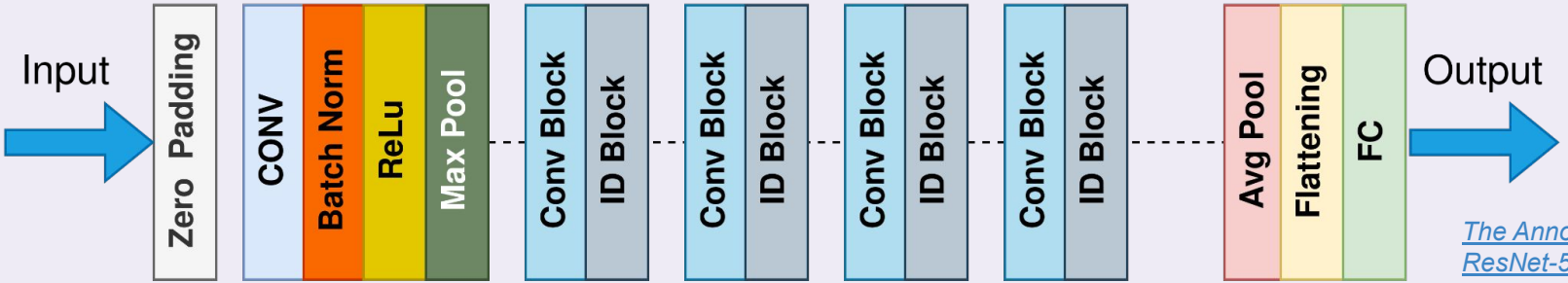
Computer Vision

Solution Architecture

End-to-end ML pipeline with FastAPI backend and Streamlit frontend. Uses ResNet for feature extraction, PCA for dimensionality reduction, and FAISS for similarity search. Data persisted in PostgreSQL and S3, with Grafana monitoring.



ResNet-50 Feature Extractor



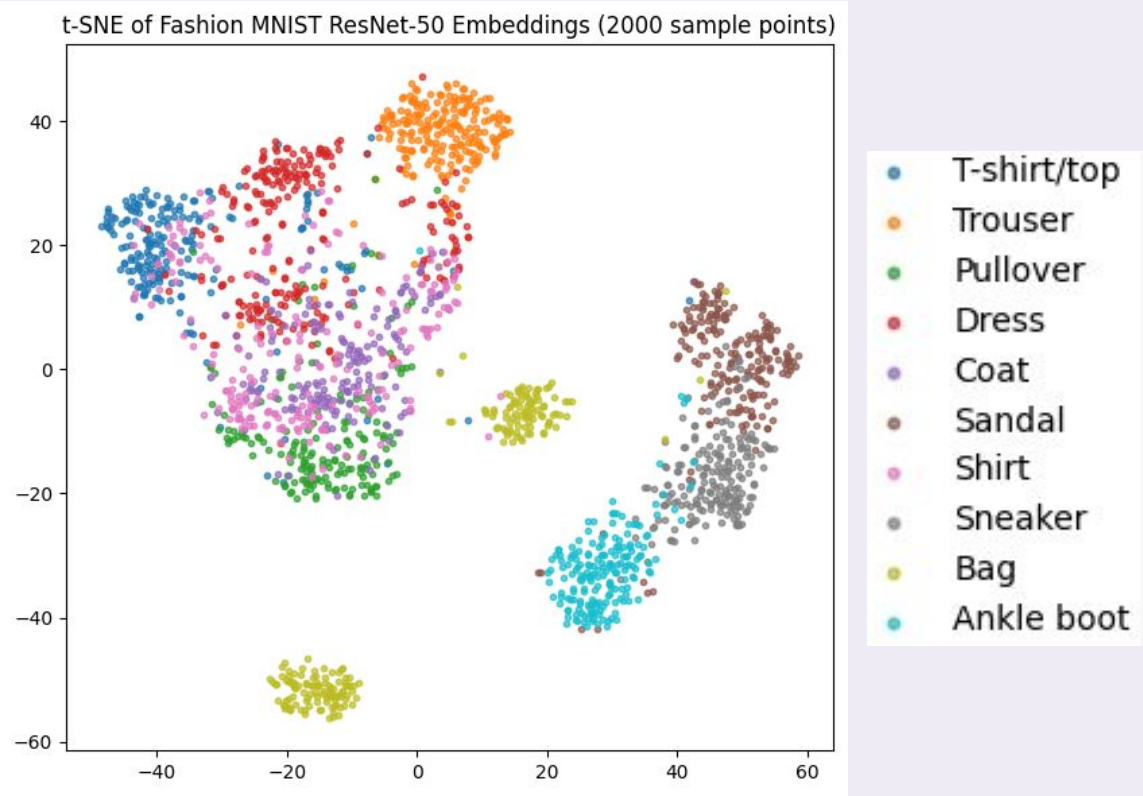
[The Annotated ResNet-50, Towards Data Science]

- Captures low-level visual features like edges
- Batch normalization stabilizes training and inference
- Max pooling reduces spatial size for efficiency

- Learn mid- to high-level visual features of images
- Residual connections preserve information across layers
- Form the core representation used for similarity and clustering

- Aggregates spatial information across the entire image
- Convert feature maps into a fixed-length vector
- Produces a 2048-dimensional image embeddings

Embedding Quality Check



3

Recommendation Engine

Fashion Recommendation Engine

- Uses ResNet-50 embeddings from the CV pipeline (2048D)
- Applies PCA for dimensionality reduction (2048D → 128D)
- Uses FAISS for fast similarity search
- Implements ranking, filtering, and metadata-aware recommendations
- Designed to be production-ready (save/load, tests, benchmarks)

Dimensionality Reduction with PCA

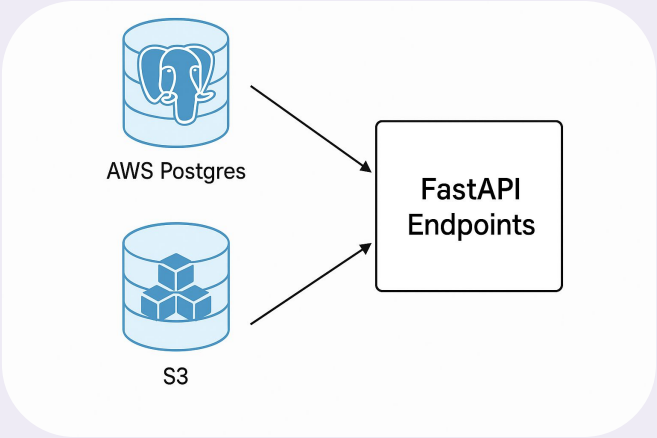
- Input: 2048D embeddings from ResNet-50
- PCA reduces to configurable dimensions
- ~95% variance retained at 128D
- Memory: ~500MB → ~30MB (≈17× reduction)
- Enables faster search and lower resource usage

FAISS & Recommendation Logic

- FAISS index with L2 or cosine similarity
- Index build time: < 1 second
- Search speed: ~10,000+ queries/second
- Recommendation logic:
    a. Distance-based ranking
    b. Class filtering (e.g., only tops, only trousers)
    c. Exclusion filtering (avoid specific items)
    d. Optional metadata-aware filtering

4

Backend Web Server



FastAPI endpoints:

1. wardrobe/upload
2. wardrobe/items
3. outfits/saved
4. predict
5. outfits/generate
6. outfits/save
7. outfits/{outfit_id}
8. wardrobe/item/{item_id}
9. wardrobe/clear-all

PostgreSQL Schema

wardrobe_items
item_id | **PK**
image_url
category
metadata

image_url: Publicly Accessible S3 URL
category: top, bottom, shoes, outerwear
Metadata: user inputs of brand, occasion, and any other notes they chose to include

embeddings
id | **PK**
item_id | **FK**
embedding
created_at

embedding: /predict generates entry into embeddings, using pretrained outfit recommender

saved_outfits
id | **PK**
outfit_id | **PK**
items
occasion
season
created_at

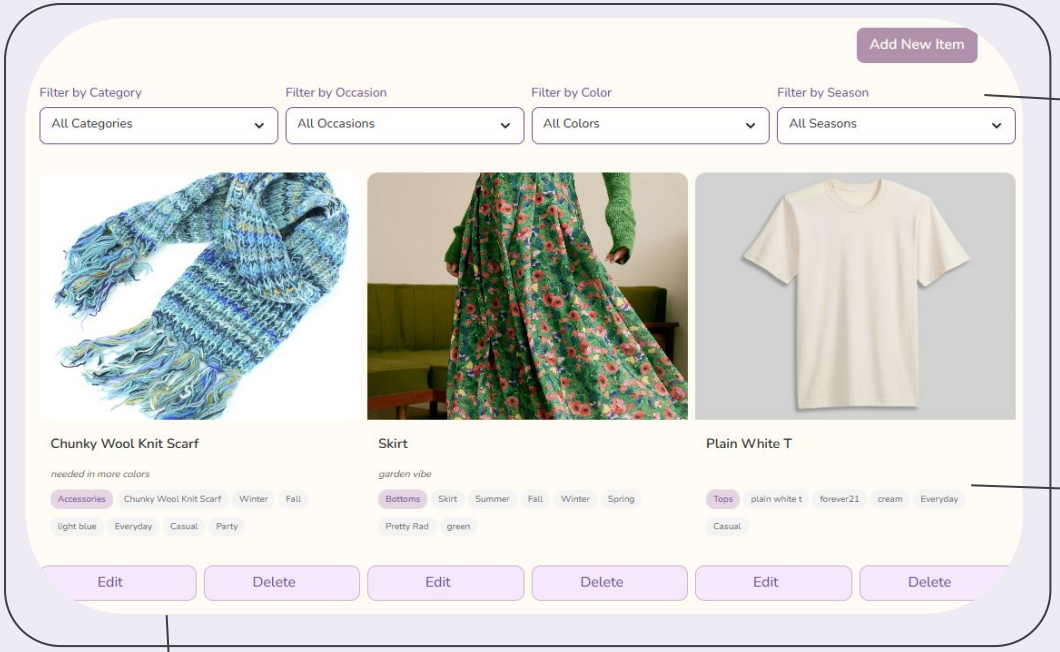
items: list of wardrobe items

5

Frontend

My Wardrobe

Add New Item
Allows user to add new items to their wardrobe.



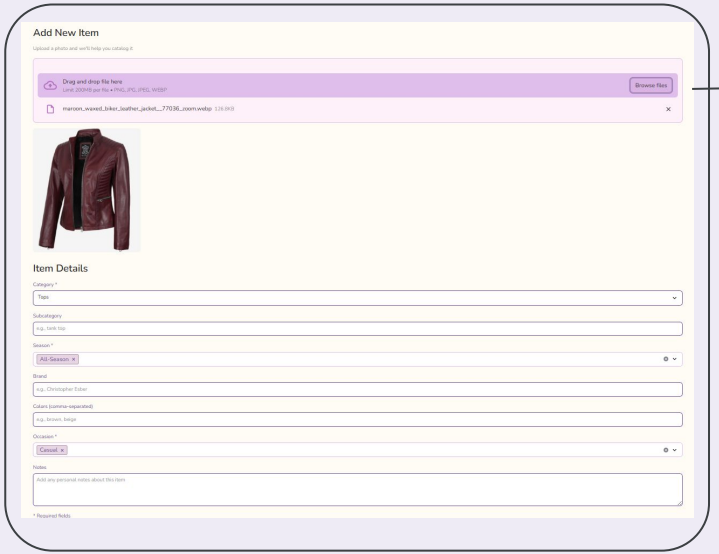
Filtering
Filter wardrobe items by category, occasion, color, or season for easy browsing.

Tags
Label items with category, season, occasion, colors, and brand for easy sorting and outfit matching.

Edit & Delete
Modify and/or remove uploaded clothing items.

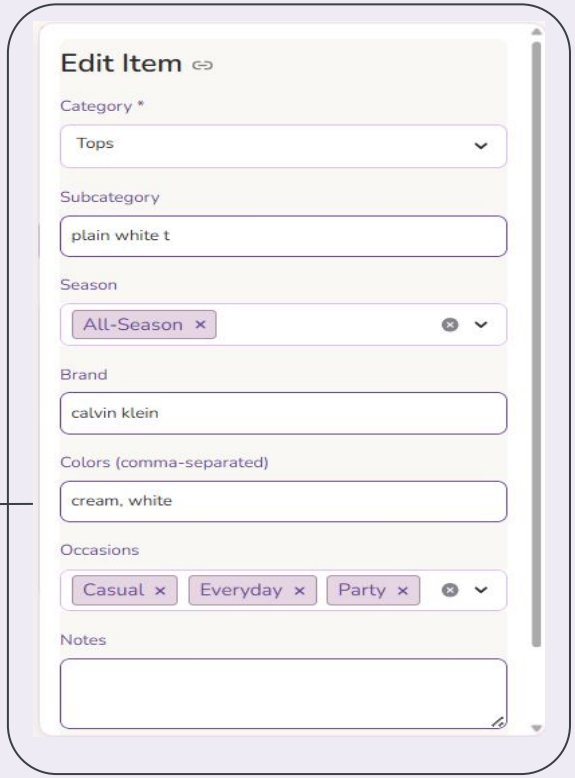
Upload New Items

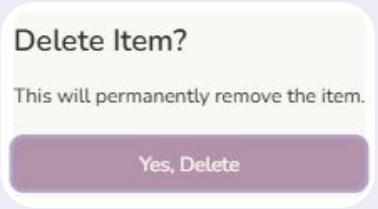
Upload clothing photos with metadata (category, season, brand, colors, occasion). Images are processed to generate embeddings for AI-powered outfit recommendations.



Wardrobe Management

Full control over wardrobe items. Users can edit metadata like category, brand, colors, and occasions to improve outfit recommendations or safely remove items with deletion confirmation.

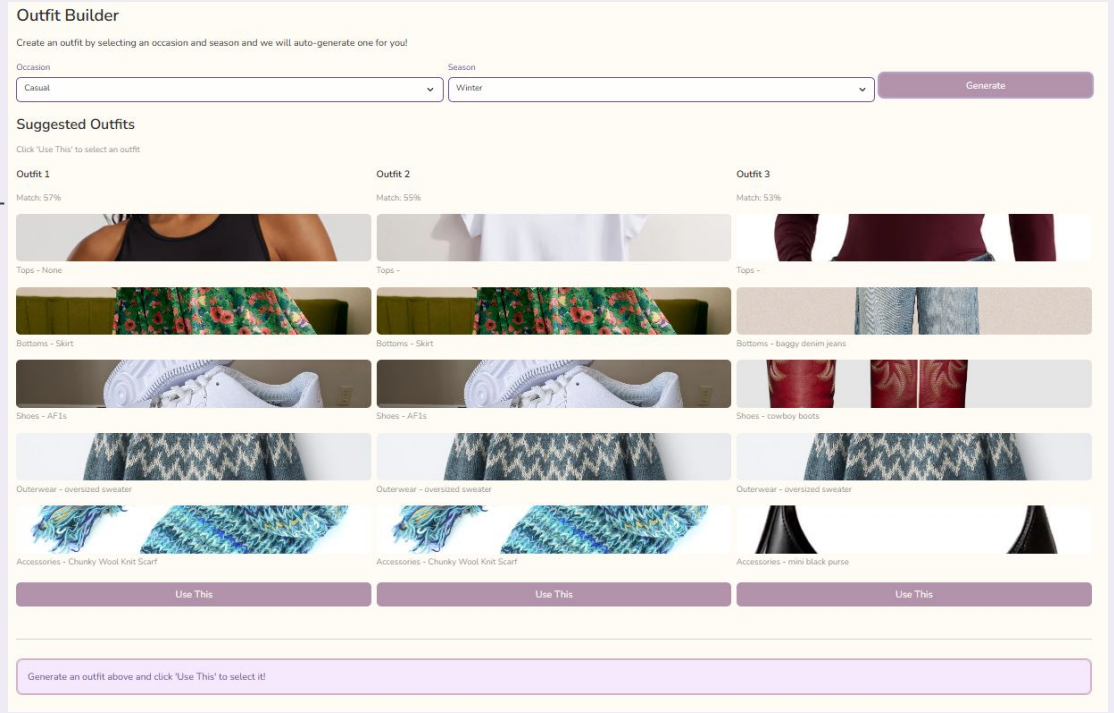




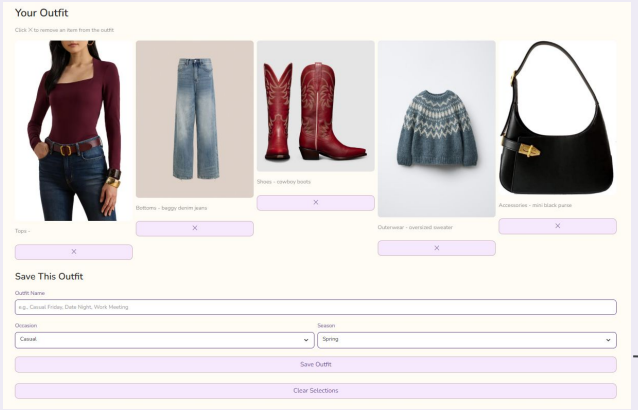
Outfit Builder

Suggested Outfits
3 ML-ranked outfit options based on occasion & season

Match Score
Embedding-based compatibility score using FAISS similarity search

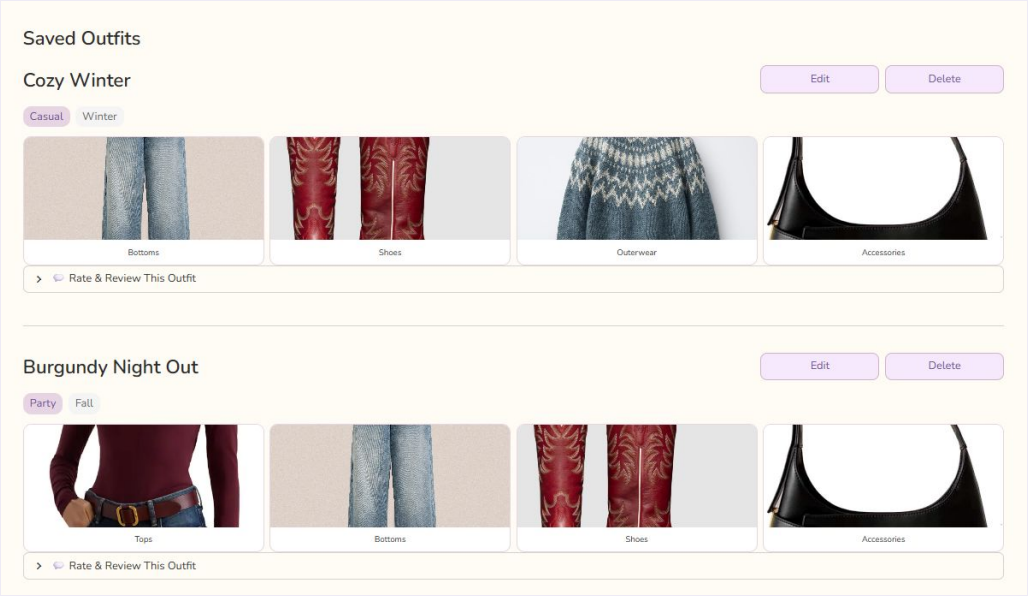


Your Outfit
Preview with item removal.



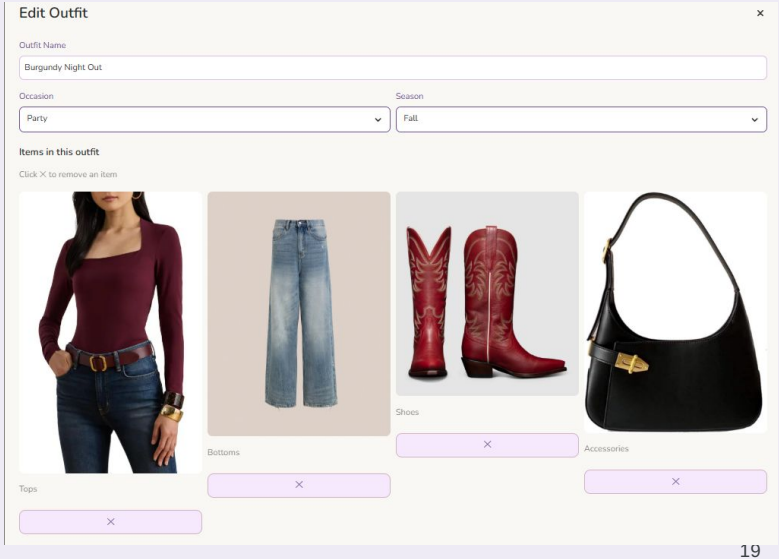
Save Outfit
Persist outfits with name, occasion & season tags

Saved Outfits



Edit Auto-Generated Outfit

Users can customize ML-generated outfits by editing the name, occasion, season, and removing individual items. This allows personalization of recommendations before or after saving.



Rate & Review Saved Outfits

Users can rate outfits (1-5 stars) and leave feedback comments. This feature captures user preferences for future model improvements. (Feedback collection in place; personalization coming soon)

6

Monitoring the app

Prometheus - Metrics Collection

- Prometheus: tracks our metrics
  - FastAPI application metrics
    - Request counts
    - Response times
  - Underlying ML metrics
    - Computer vision processing time
    - Recommendation engine query performance
  - Data-layer metrics
    - Postgres
    - AWS operations
  - Errors

Grafana: Visualization & Alerting

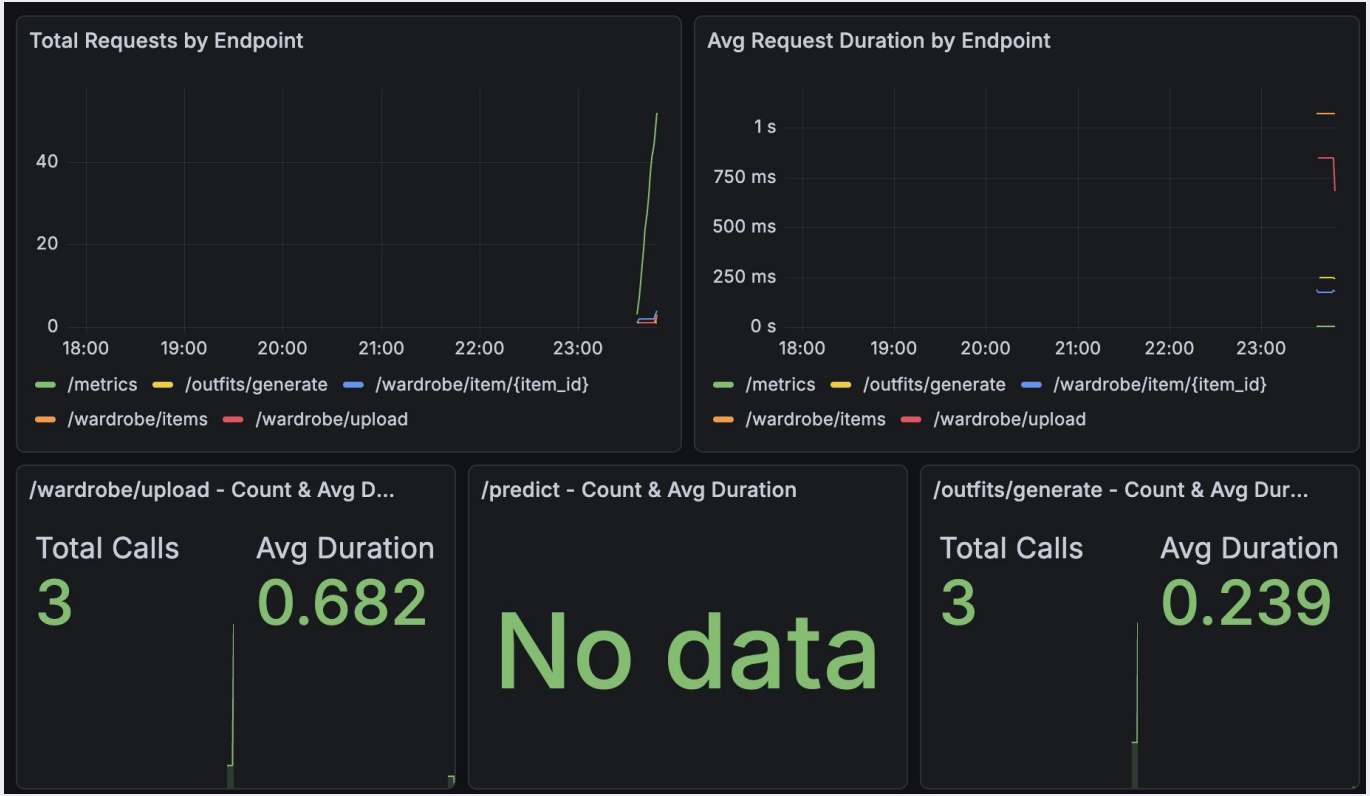
- Create custom dashboards to see API health
- Monitor recommendation engine performance
- Track user activity patterns
- Expose errors

Prometheus + Grafana: Working Together

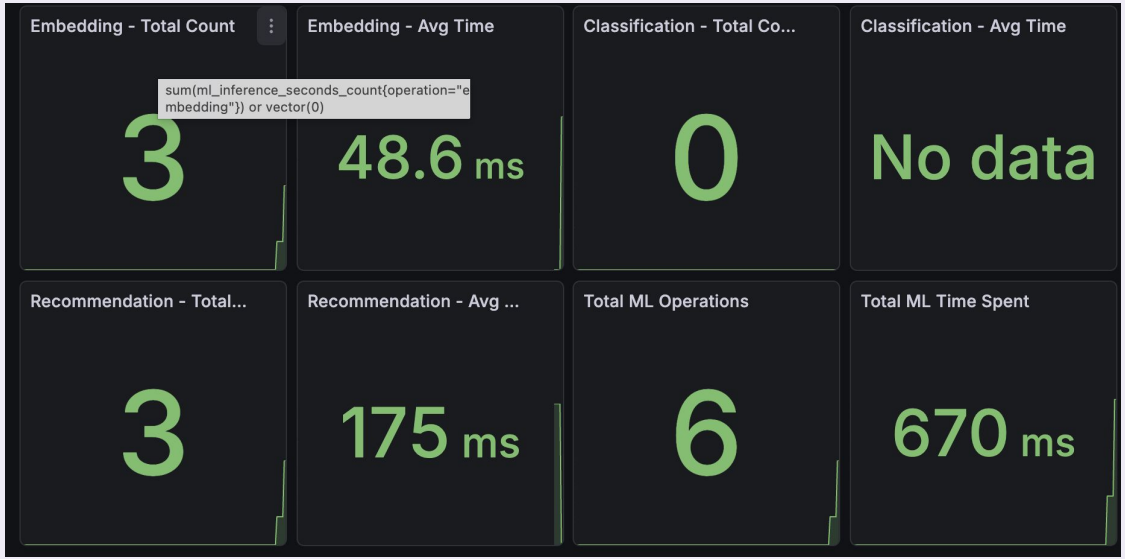
- Monitoring Stack Architecture:
  - FastAPI -> Prometheus -> Grafana -> Users/DevOps
- Prometheus: The Data Collector & Storage
  - Scrapes metrics from FastAPI backend
  - Stores time-series data
  - Provides query language
- Grafana: The Visualization Layer
  - Connects to Prometheus as a data source
  - Transform raw metrics into visual dashboards

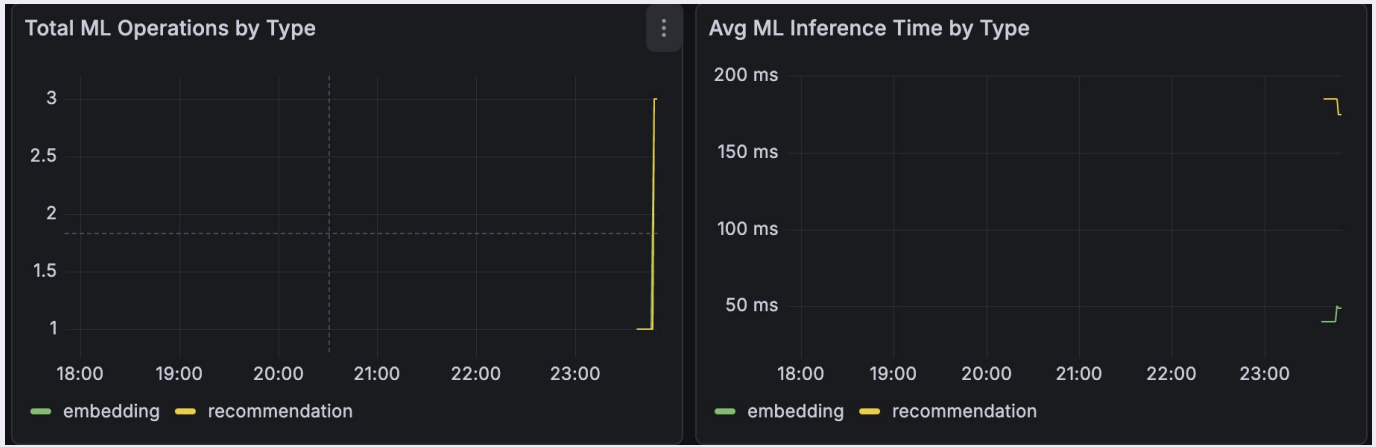
- **Why Use Both Together?**
  - **A standard monitoring stack that allows us to modify dashboards without disrupting data collection and scale both tools independently for production workloads**

1. API Metrics

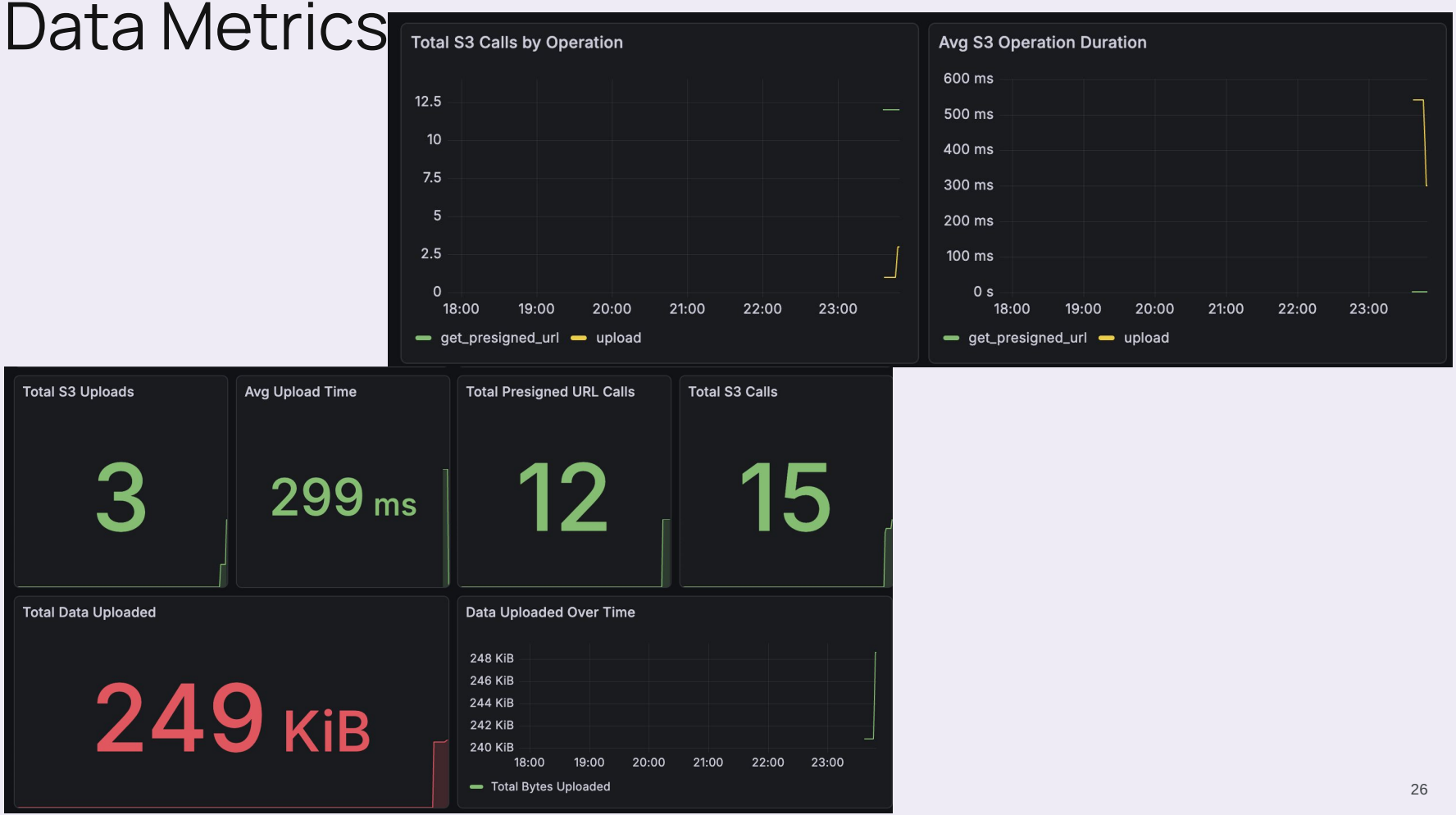


2. ML Metrics

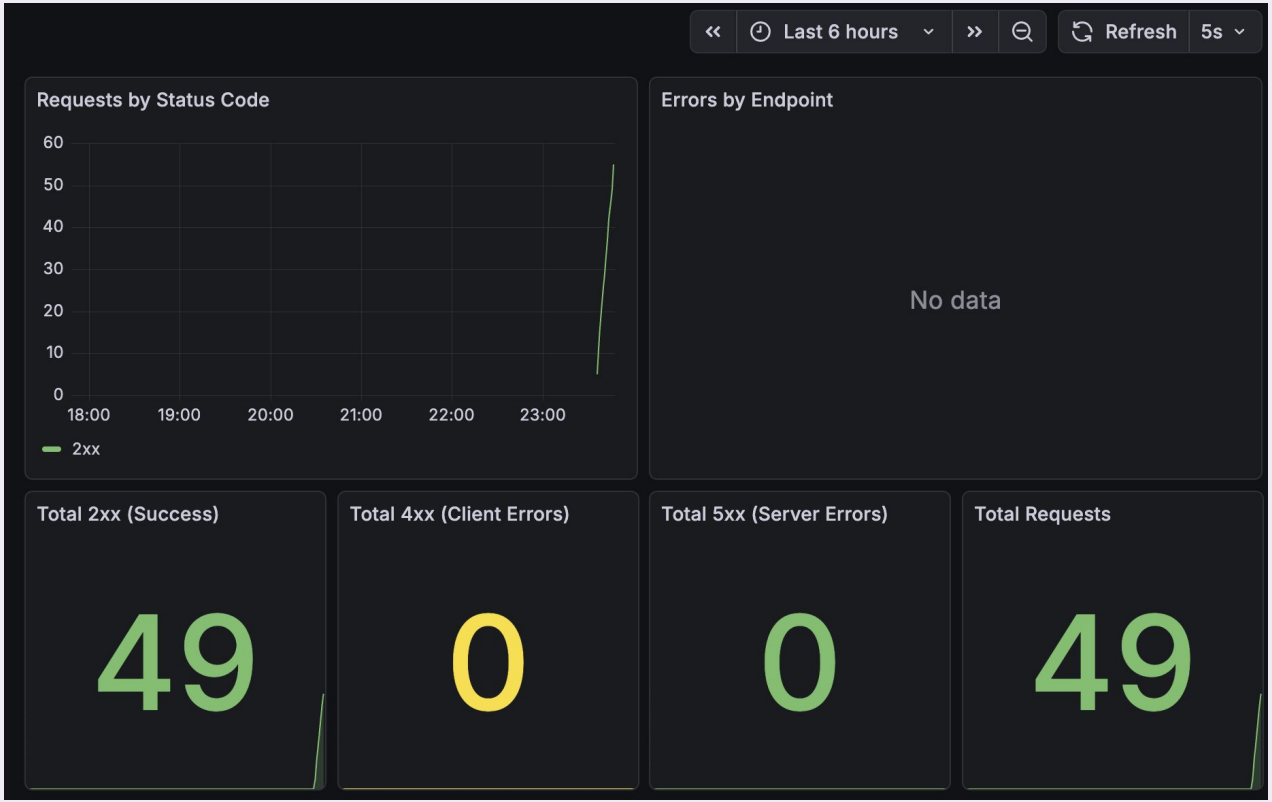

Embedding - Total Count
sum(ml_inference_seconds_count{operation="embedding"}) or vector(0)
3
Embedding - Avg Time
48.6 ms
Classification - Total Co...
0
Classification - Avg Time
No data
Recommendation - Total...
3
Recommendation - Avg ...
175 ms
Total ML Operations
6
Total ML Time Spent
670 ms


Total ML Operations by Type
3
2.5
2
1.5
1
18:00
19:00
20:00
21:00
22:00
23:00
embedding
recommendation
Avg ML Inference Time by Type
200 ms
150 ms
100 ms
50 ms
18:00
19:00
20:00
21:00
22:00
23:00
embedding
recommendation

3. Data Metrics



4. Errors


Last 6 hours
Refresh
5s
Requests by Status Code
60
50
40
30
20
10
0
18:00
19:00
20:00
21:00
22:00
23:00
2xx
Errors by Endpoint
No data
Total 2xx (Success)
49
Total 4xx (Client Errors)
0
Total 5xx (Server Errors)
0
Total Requests
49

7

App Demonstration

Thank You!