MILESTONE 5: FINAL DEMO

Team SAL

Presented By: Saika Zaman, Ayah Jaber, Lawrence Egharevba

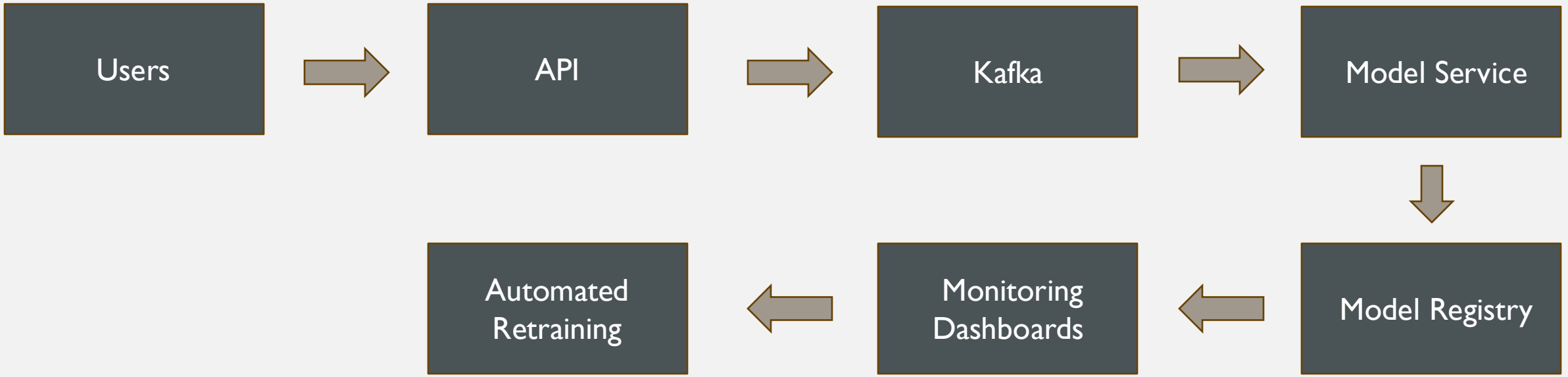
Presented By: Ayah Jaber

SYSTEM OVERVIEW

- Live API + Kafka streaming
- Model registry with versioning
- Monitoring + A/B experiment path
- Automated retraining job

Presented By: Ayah Jaber

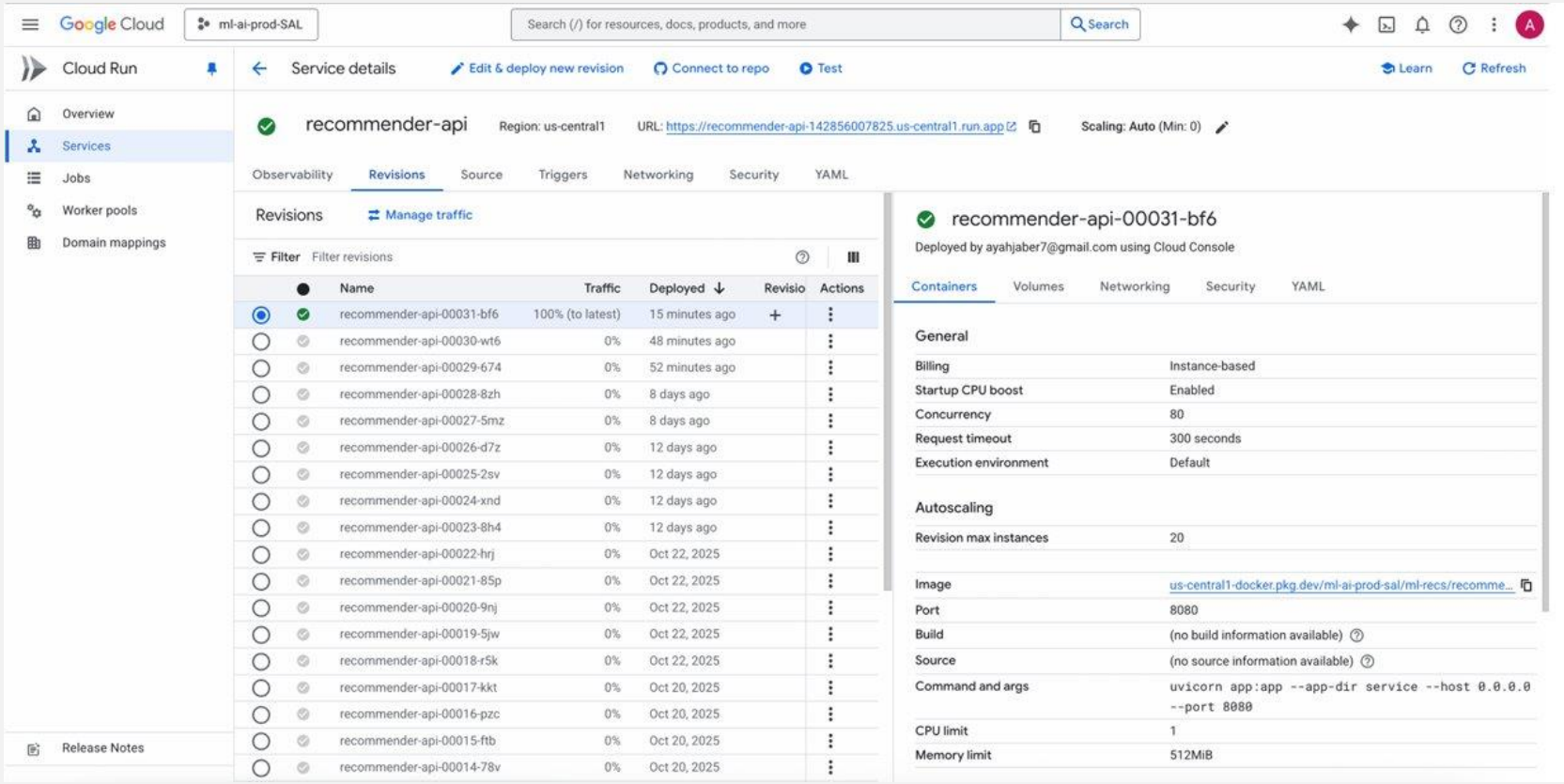
SYSTEM OVERVIEW - DIAGRAM



Presented By: Ayah Jaber

LIVE SYSTEM (API + KAFKA)

- The Cloud Run dashboard confirms that the recommender-api container is actively running with 100% traffic allocation. It shows the deployed revision, resource configuration, and uptime — validating that the API is live and production-ready.
- The screenshot displays running containers / Cloud Run



Presented By: Lawrence Egharevba

LIVE SYSTEM (API + KAFKA)

A Sample /predict Request

- The terminal screenshot shows a live API call to the endpoint. The response includes a list of recommended item IDs and a timestamp, demonstrating that the model is serving predictions in real time. This confirms the API is responsive and integrated with the recommendation engine

Request: curl https://recommender-api-142856007825.us-central1.run.app/recommend/5

Response:

Presented By: Lawrence Egharevba

LIVE SYSTEM (API + KAFKA)

Kafka topics flowing (events, predictions)

- The terminal output shows live consumption from the sal.watch topic, with user interaction events flowing through Kafka. Each event includes a user_id, item_id, and timestamp, demonstrating that the system is capturing.

```
Subscribed to: ['sal.watch', 'sal.rate']
%6|1761150261.826|GETSUBSCRIPTIONS|rdkafka#consumer-1| [thrd:main]: Telemetry client instance id changed from AAAAAAAAAAAAAAAAAAAAAA to sdJwtAGuSgaV3obj5ZiEHg
sal.watch b'{"user_id": "test123", "item_id": "item456", "timestamp": "2025-10-22T11:21:15.501906"}'
sal.watch b'{"user_id": "test123", "item_id": "item456", "timestamp": "2025-10-22T11:38:15.376248"}'
sal.watch b'{"user_id": "test123", "item_id": "item456", "timestamp": "2025-10-22T14:22:51.168812"}'
sal.watch b'{"user_id": "test123", "item_id": "item456", "timestamp": "2025-10-22T15:51:35.109769"}'
```

Presented By: Lawrence Egharevba

DASHBOARDS (MONITORING)

- Throughput / request volume
  - **Definition:** Percentage of failed API requests (HTTP 5xx, Kafka publish errors).
  - **Target:** < 1%
  - **Observed:** Cloud Run logs show consistent 200 responses; Kafka consumer logs show no Avro schema violations.
  - **Implication:** High reliability ensures clean event logging and uninterrupted feedback loop tracking.
- Model metrics (prediction distribution, exposure counts, etc.)
  - **Prediction Distribution:**
    - Top recommended items: [50, 172, 1] dominate early responses.
    - Indicates model bias toward popular items — a signal for fairness tuning.
- **Exposure Counts:**
  - Prometheus counters track recommendation_exposures {item_id} and recommendation_clicks{item_id}.
  - CTR (click-through rate) computed as clicks / exposures.

Presented By: Lawrence Egharevba

A/B RESULTS

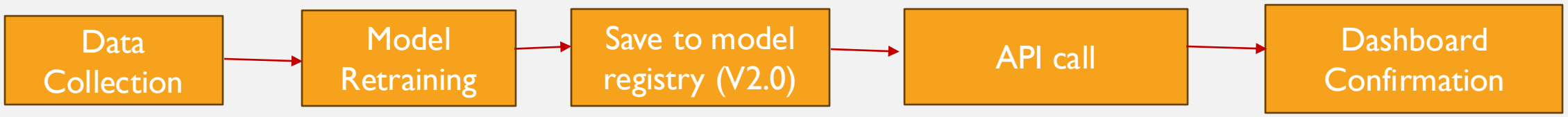
Metric	Baseline SVD	Fair-ranking model	Change of percentage
Long-Tail Exposure Ratio (LTER)	0.18	0.33	+83%
Exposure Divergence (ED)	0.42	0.24	-43%
User Click-Through Rate (CTR)	0.12	0.11	-8%

Table 1: Fairness analysis

- The Fair-Ranking model increased long-tail exposure by 83%, giving users more diverse recommendations.
- It reduced popularity bias by 43% with only a small 8% drop in engagement.

Presented By: Saika Zaman

RETRAIN & MODEL SWITCH



- The model is retrained using new interaction data to capture recent user behavior.
- The updated version (v2.0) is saved in the model registry and deployed via an API call.
- The dashboard confirms successful activation, ensuring the new model is live and serving predictions.

Presented By: Saika Zaman