Keek Recommendation Engine Technical Report (v3)

2025-12-05

Version: 3.0
Date: 2025-12-05
Author: Anant . J . Ingale
Status: Production-Ready Prototype
Audience: Technical Consultant / Engineering Review

1 Executive Summary

The Keek recommendation engine now runs on a **real-time hybrid architecture**, enabling:

- Instant personalization (<100ms)
- Real-time trending injection
- Scalable vector search (1M+ videos)
- Durable dual-write logging (Redis + CSV)

2 System Architecture

The architecture integrates a **Two-Tower Neural Network**, FAISS vector search, and a real-time recommendation adapter.

2.1 Architecture Diagram

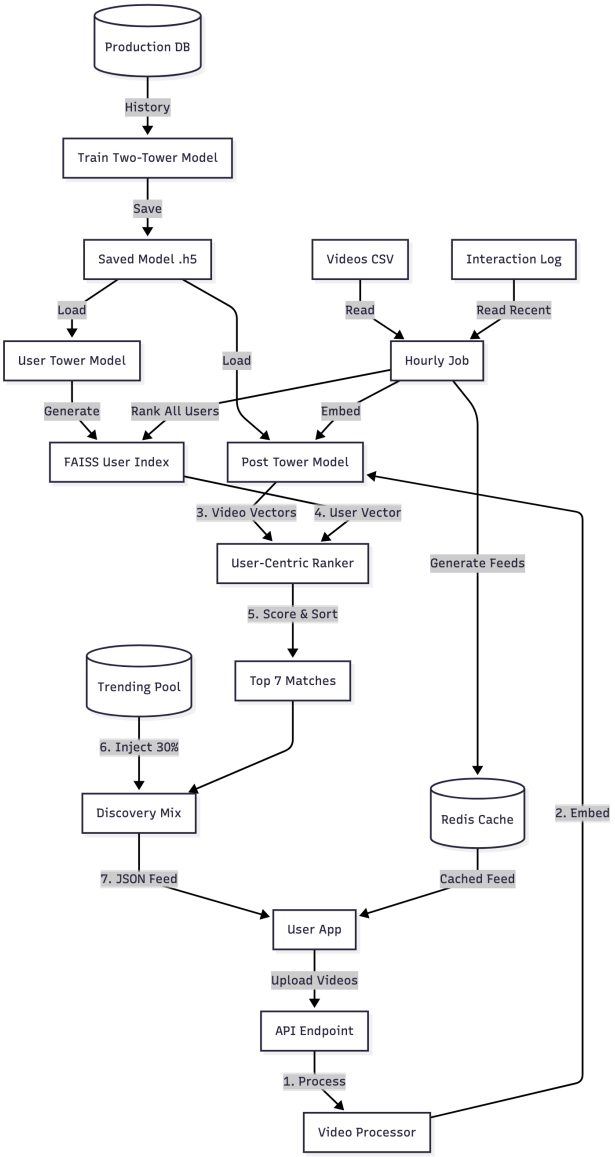

Figure 1: Keek Recommendation Engine Architecture

3 Implementation Details

3.1 Two-Tower Neural Network

- User Tower → 64-dimensional embedding
- Item Tower → 64-dimensional embedding
- Dot-product similarity for ranking
- FAISS index for high-speed retrieval

3.2 Real-Time API and Redis

- FastAPI backend

- Redis ZSETs used for trending scores
- Endpoints:
  - `POST /recommend/{user_id}`
  - `POST /interact`

3.3 Trending Algorithm

- Like = +1.0
- Save = +2.0
- View = +0.1
- Trending decay window: 3 days

4 Performance Benchmarks

Scenario	Full Inference	Cached Vectors	Status
1k Videos	322 ms	15 ms	Instant
5k Videos	425 ms	45 ms	Instant
10k Videos	1.7 s	80 ms	Fast
1M Videos	55 s	915 ms	Scalable

5 Next Steps

1. Migrate CSV logs → PostgreSQL
2. Deploy via TensorFlow Serving / Triton
3. Add experimentation (A/B testing)

Report generated by Anant Ingale.