Recipe Analytics Pipeline

Project Documentation

Prepared by: Prasad Bari

E-mail - prasadbari1515@gmail.com

Batch : Data Engineer

November 21, 2025

Abstract

This document describes the Recipe Analytics Pipeline: an end-to-end data engineering mini-project using Firebase Firestore as the source, Python-based ETL, validation, and analytics scripts producing CSV exports and PNG visualizations. The document includes architecture data-flow diagrams, data model, execution steps, and example schemas.

---

1. Introduction

The **Recipe Analytics Data Engineering Project** is designed to demonstrate a complete, end-to-end data engineering workflow using **Firebase Firestore**, **Python**, and **data analytics techniques**.
The system processes recipe-related data (users, recipes, and interactions) and converts it into clean, validated datasets suitable for downstream analytics.

This project showcases:

- Real-time NoSQL database modelling.
- Programmatic data ingestion using Firebase Admin SDK.
- ETL (Extract–Transform–Load) pipeline.
- Data normalization and clean CSV generation.
- Automated data validation.
- Exploratory data analysis and insights.
- Visualization through charts.
- Professional documentation of the process.

---

2. Project Overview

The Recipe Analytics System is a complete end-to-end data engineering pipeline built using Firebase Firestore, Python, ETL, validation, and analytics.
The system simulates a real-world recipe platform where users upload recipes and interact with them through views, likes, ratings, and cooking attempts.

The goal of this project is to design:

- A scalable NoSQL data model
- Automated data upload scripts
- ETL export pipeline
- Data validation workflow
- Analytics and insight generation
- Final documentation & visual diagrams

This ensures that the system is production-ready, analytics-friendly, and well-structured.

---

3. Explanation of the data model

Collections and Fields

Users : user_id, name, email, location, preferences (array), signup_date.

Recipes : recipe_id, title, difficulty, servings, prep_time_min, cook_time_min, tags[], ingredients[] ({name,quantity,unit}), steps[], created_at, updated_at.

Interaction : Fields used (aggregate per recipe): recipe_id, avg_rating, rating_count,

likes, cook_attempts, views.

---

4. Data Flow Diagram

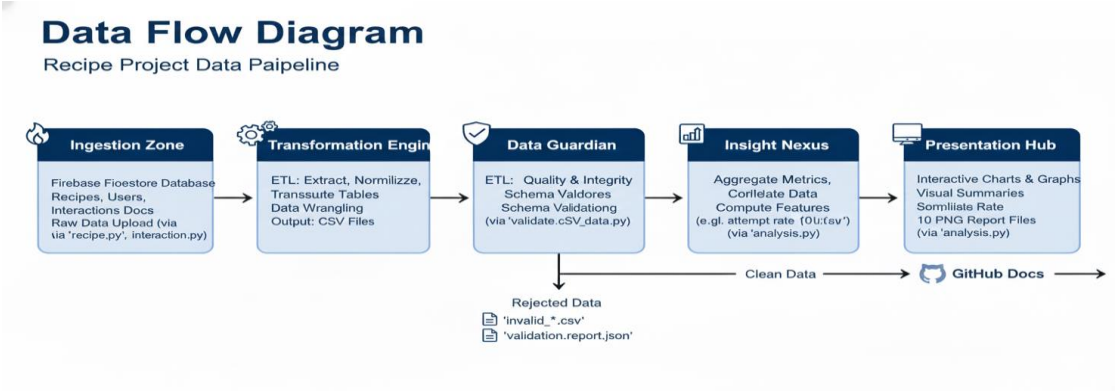


Fig 1. Data Flow Diagram

5. ER Diagram

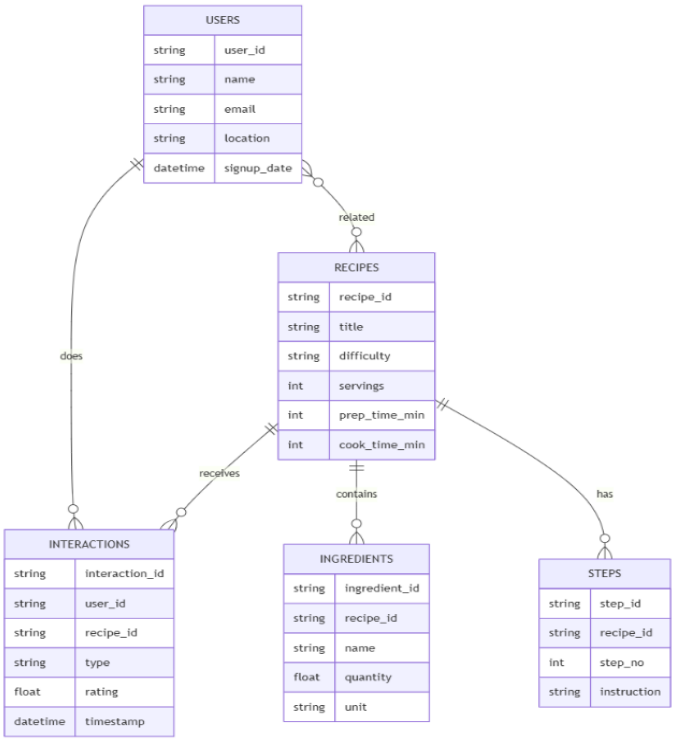
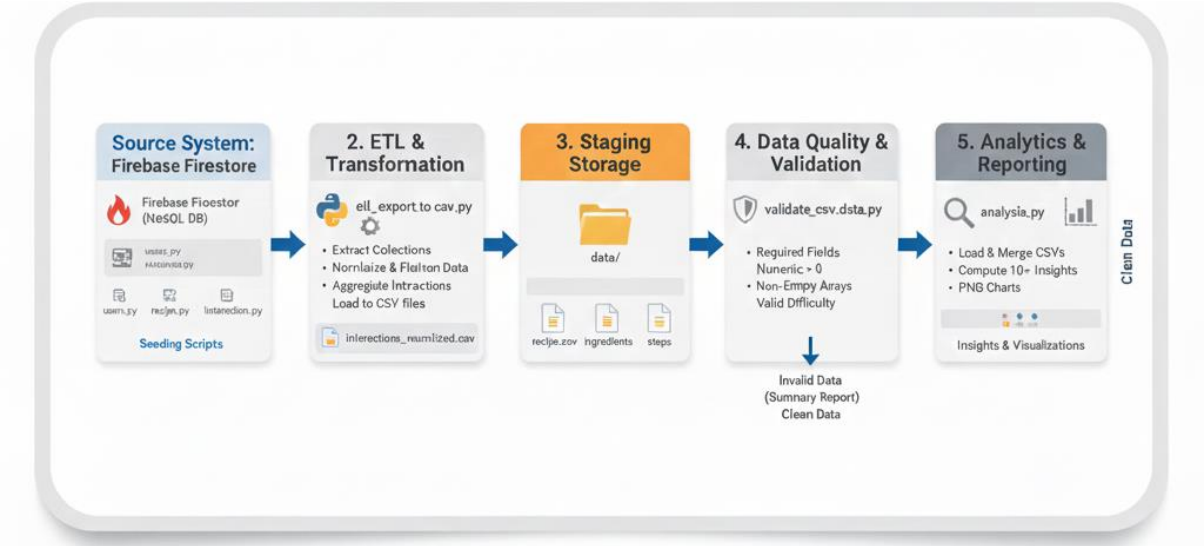


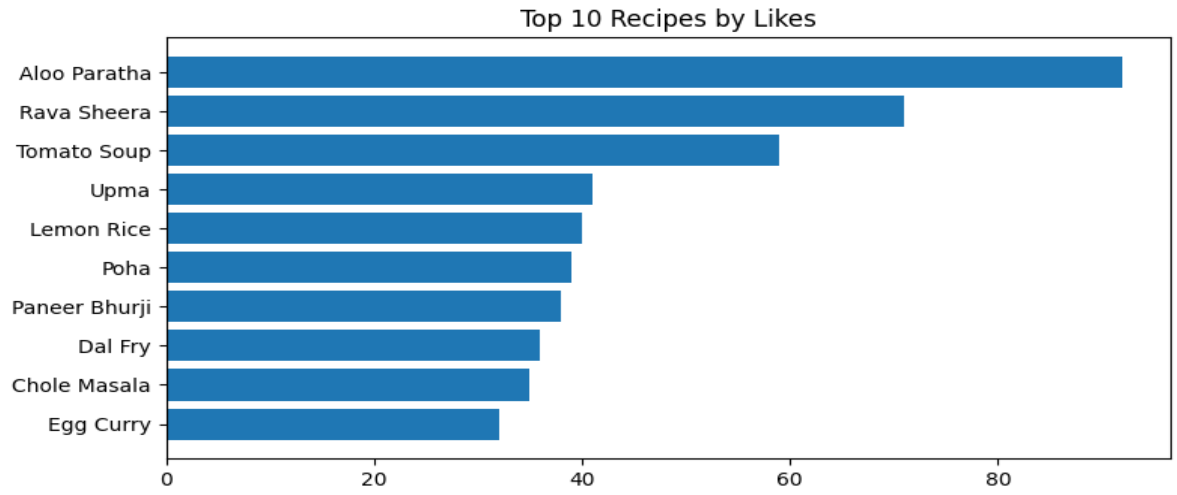
Fig 2. ER Diagram

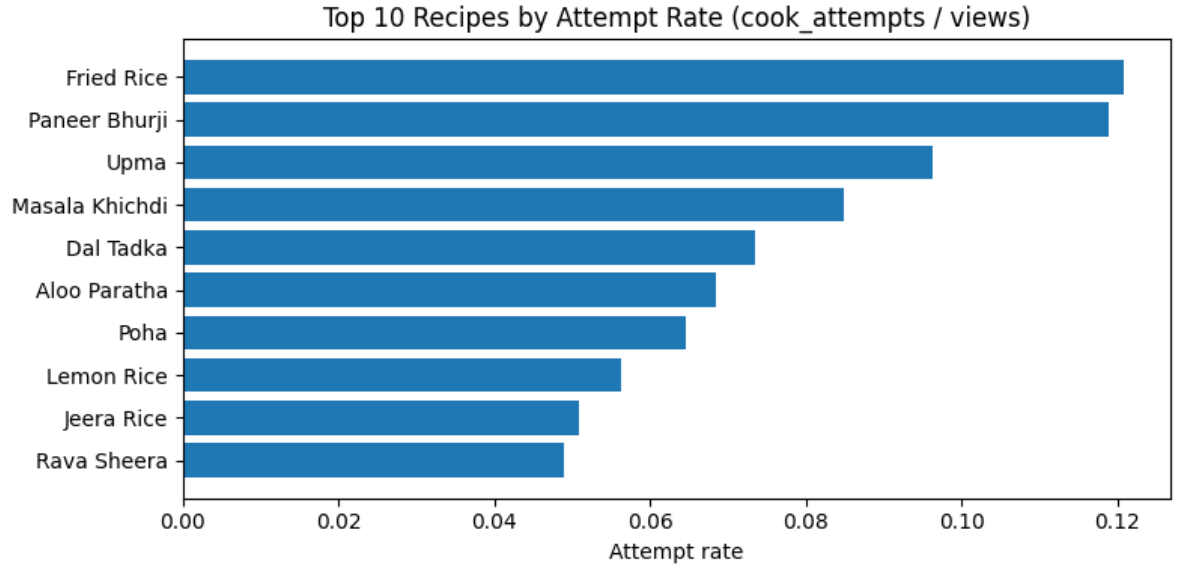
6. Architecture Diagram

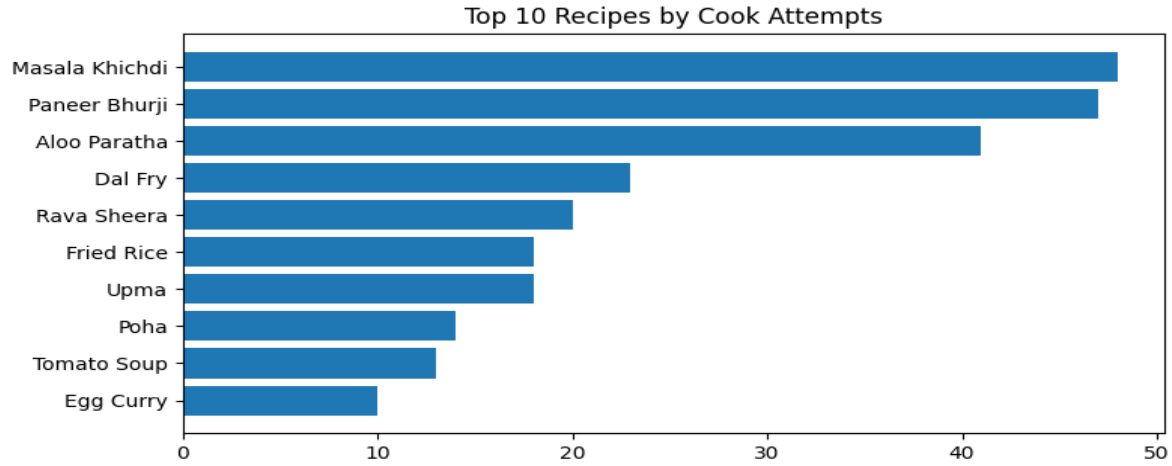


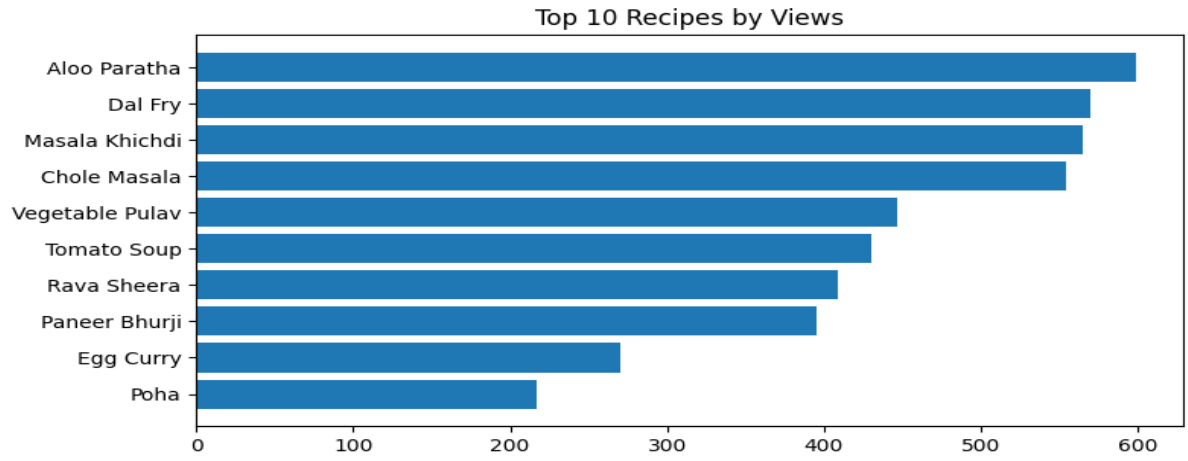
8. Analytics and Visualizations

Top Recipes (by Likes,Cook Attempt Rate, Cook attempts, views)

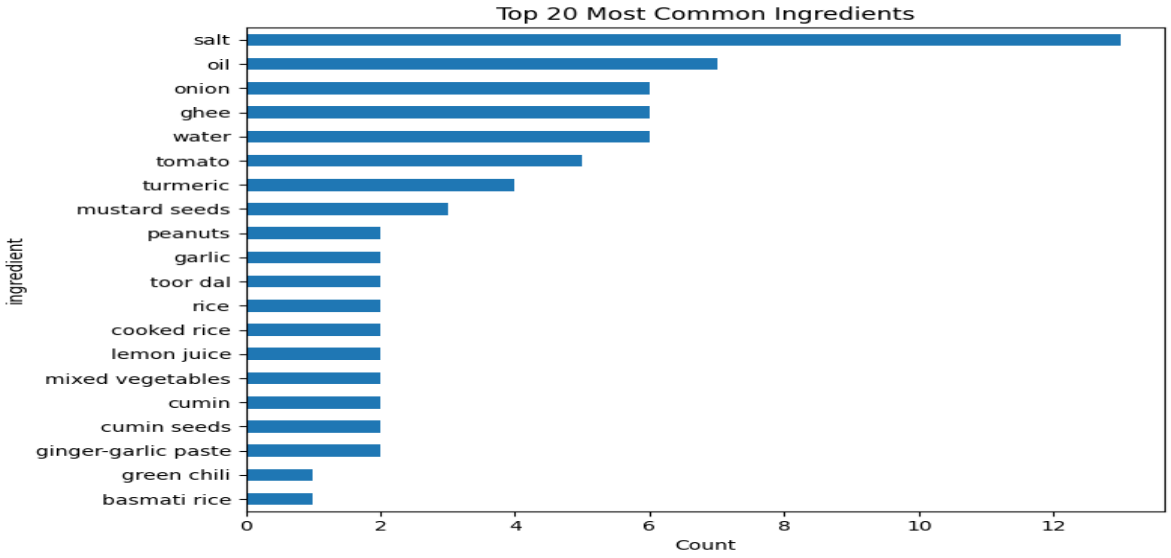


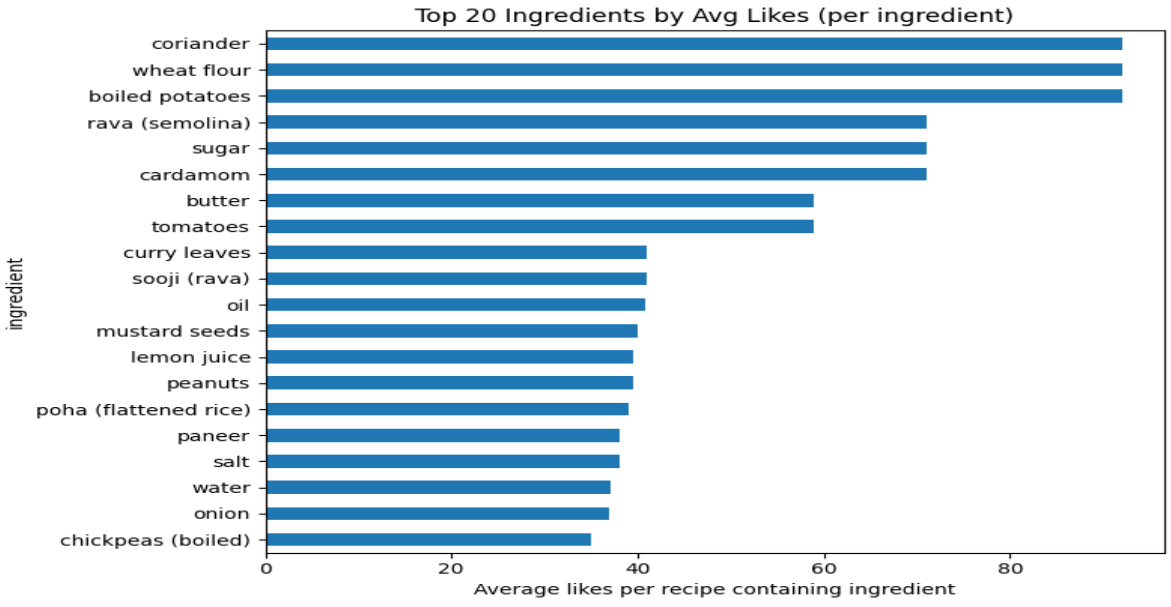




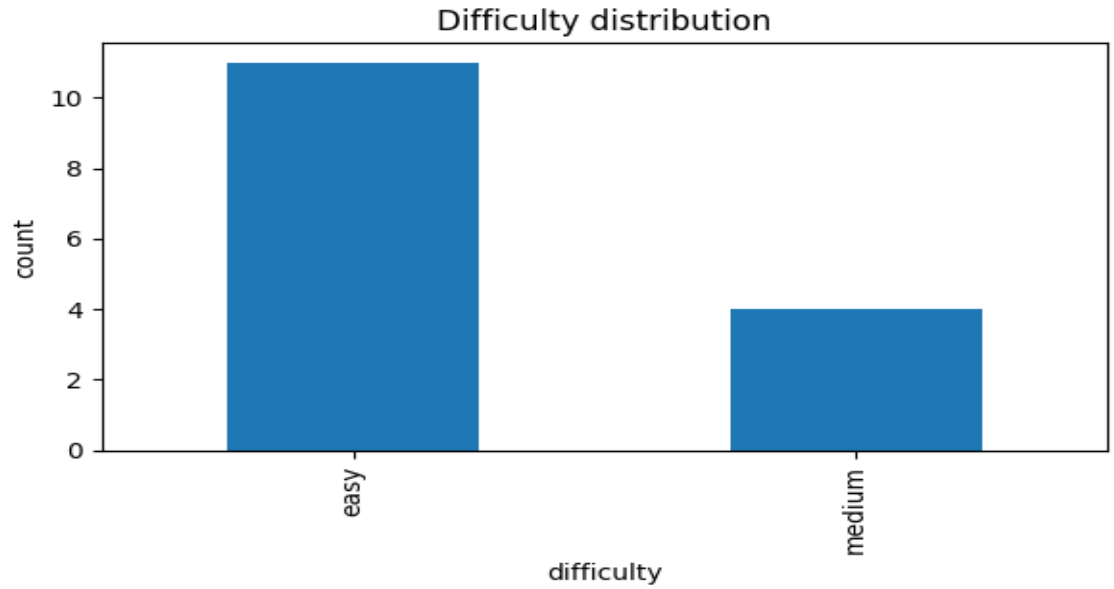


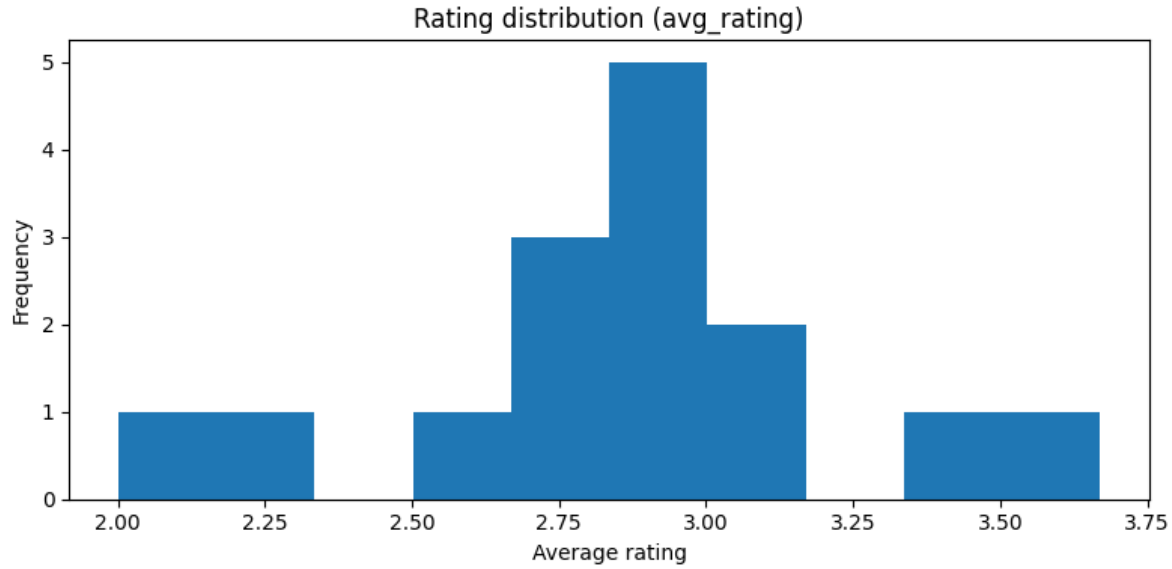
Top Ingredients





Difficulty Insights





---

9. Overall Findings

The analysis of the Recipe Analytics dataset reveals clear patterns in user engagement, recipe composition, and ingredient behavior. Easy and familiar dishes dominate the platform, drawing the highest views, likes, and cooking attempts, while medium-difficulty recipes tend to earn slightly better ratings. Core Indian staples such as salt, oil, onion, and tomato appear across most recipes, forming the foundation of the dataset, whereas ingredients like boiled potatoes, wheat flour, coriander, and butter are linked to the most liked dishes, highlighting their influence on recipe popularity. Top-performing recipes—including Aloo Paratha, Masala Khichdi, Paneer Bhurji, and Dal Fry—consistently excel across all major engagement metrics, showing strong user preference for comfort foods and everyday meals. Additionally, tag-based insights indicate that breakfast items, desserts, and quick recipes receive above-average user

interest. Together, these findings provide a comprehensive understanding of what users value most, forming a strong analytical base for future improvements, personalized recommendations, and enhanced content strategy.

---

10. Final Conclusion

The Recipe Analytics Project successfully demonstrates a complete end-to-end data engineering workflow, transforming raw Firestore data into validated datasets and meaningful insights. Through ETL processing, data validation, and analytical visualization, the system reveals clear patterns in user preferences, recipe performance, and ingredient influence. The project proves the effectiveness of a scalable NoSQL data model and automated Python pipelines, while also establishing a solid foundation for future enhancements such as recommendation systems, advanced analytics, and interactive dashboards.
