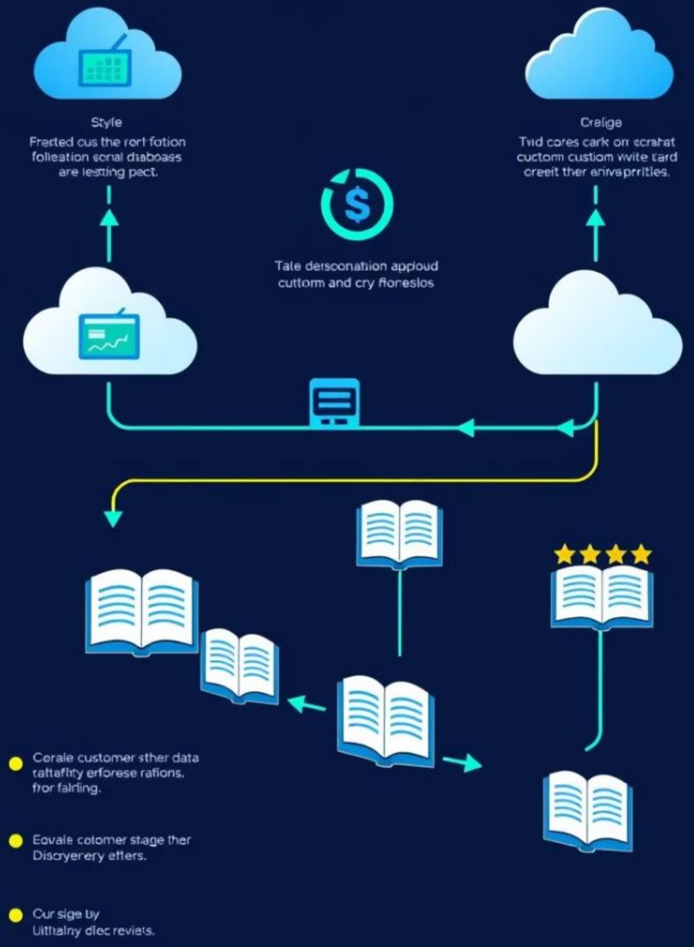

Style
Frested cus the rort fotion
folleation sonal diaboass
are lesting pect.
Cralge
Tvid cores cark on ecrahat
cuctom custiom wite tard
creeit ther anivaprrities.
$
Tale dersconation appoud
cuttom and cry florestos
Corale customer sther data
taftafity erforese rafions.
fror falrding.
Eovale colomer stage ther
Discryerery etters.
Cur sige by
Uithalny dlec reviets.

Amazon Book Reviews Recommendation System

Big Data Pipeline on Google Cloud Platform - Final Project Presentation

Introduction

Dataset(from kaggle)

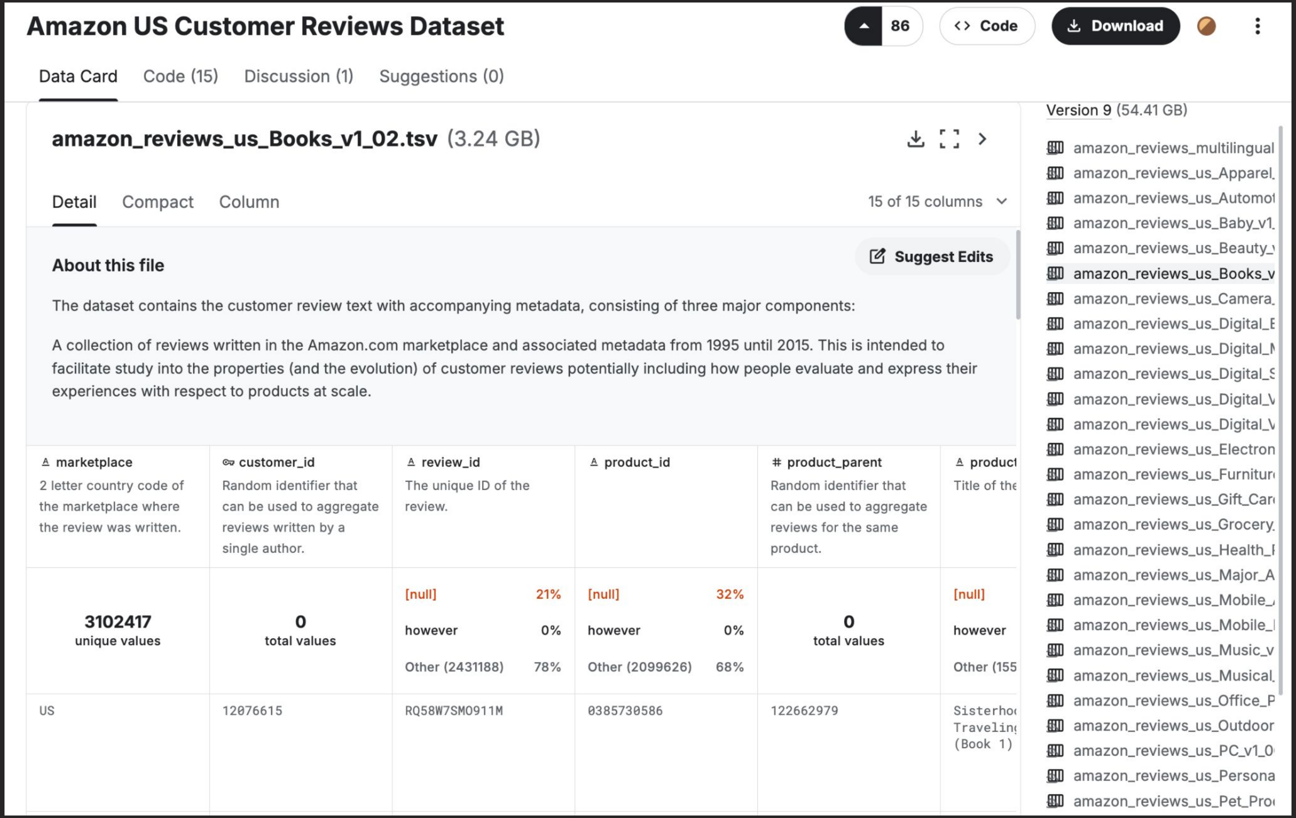
3.1M reviews, 1.5M users, 779K products with ratings and metadata. 3 GB.

Problem Statement

Build a scalable recommendation system for Amazon book reviews.

Objective

Compare different recommendation algorithms on large-scale data using GCP.



Amazon US Customer Reviews Dataset

amazon_reviews_us_Books_v1_02.tsv (3.24 GB)

About this file

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

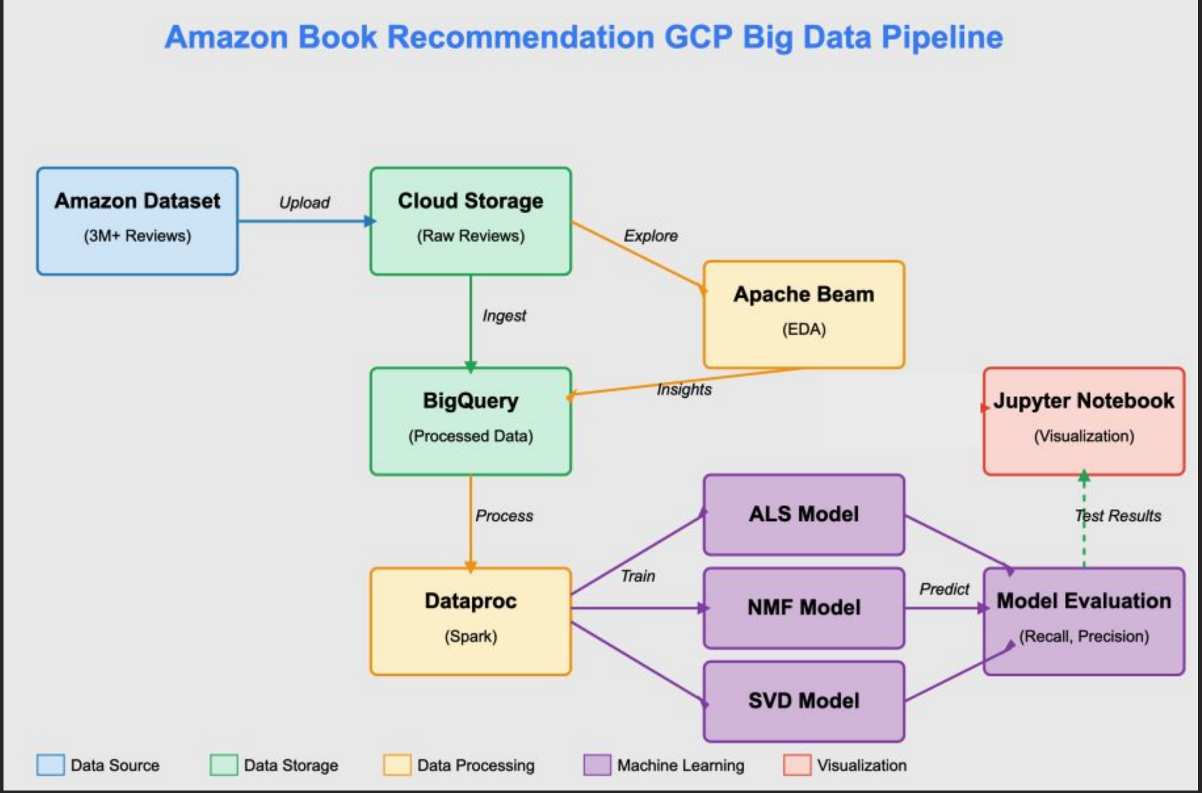
A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale.

marketplace	customer_id	review_id	product_id	product_parent	product
2 letter country code of the marketplace where the review was written.	Random identifier that can be used to aggregate reviews written by a single author.	The unique ID of the review.		Random identifier that can be used to aggregate reviews for the same product.	Title of the
3102417 unique values	0 total values	[null] 21% however 0% Other (2431188) 78%	[null] 32% however 0% Other (2099626) 68%	0 total values	[null] however Other (155
US	12076615	RQ58W7SMO911M	0385730586	122662979	Sisterhoo Travelin (Book 1)

Version 9 (54.41 GB)

- amazon_reviews_multilingual
- amazon_reviews_us_Apparel_
- amazon_reviews_us_Automot
- amazon_reviews_us_Baby_v1_
- amazon_reviews_us_Beauty_
- amazon_reviews_us_Books_v
- amazon_reviews_us_Camera_
- amazon_reviews_us_Digital_E
- amazon_reviews_us_Digital_M
- amazon_reviews_us_Digital_S
- amazon_reviews_us_Digital_V
- amazon_reviews_us_Digital_V
- amazon_reviews_us_Electron
- amazon_reviews_us_Furnitur
- amazon_reviews_us_Gift_Car
- amazon_reviews_us_Grocery_
- amazon_reviews_us_Health_
- amazon_reviews_us_Major_A
- amazon_reviews_us_Mobile_
- amazon_reviews_us_Mobile_
- amazon_reviews_us_Music_v
- amazon_reviews_us_Musical_
- amazon_reviews_us_Office_P
- amazon_reviews_us_Outdoor
- amazon_reviews_us_PC_v1_0
- amazon_reviews_us_Persona
- amazon_reviews_us_Pet_Pro

Pipeline Architecture & GCP Services



Data Ingestion

Google Cloud Storage

Data Processing

Dataflow (Apache Beam), BigQuery

Storage & Query

BigQuery

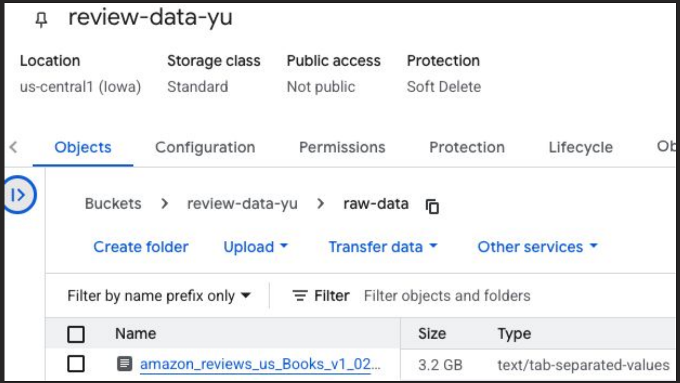
Machine Learning

PySpark ML, scikit-learn

GCP Services Used

Google Cloud Storage

Stored raw TSV data and temporary storage for processing.



review-data-yu

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	Soft Delete

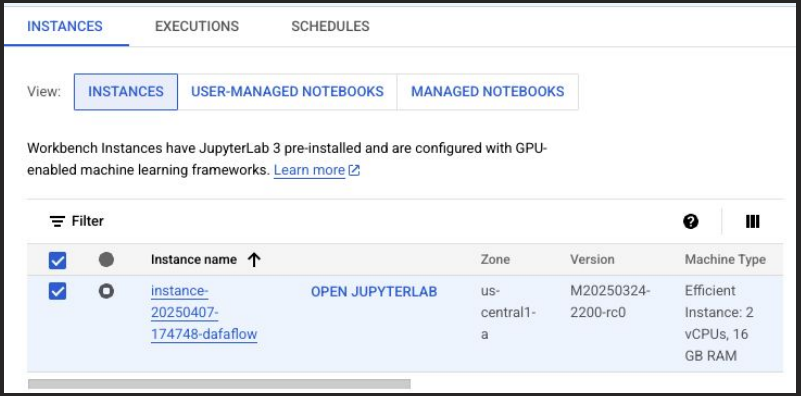
☐	Name	Size	Type
☐	amazon_reviews_us_Books_v1_02...	3.2 GB	text/tab-separated-values

BigQuery

Structured storage, SQL queries, aggregations, and train/test splits.

Dataflow

Distributed Apache Beam pipelines for EDA on 3.1M reviews.



Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

☑	●	Instance name ↑		Zone	Version	Machine Type
☑	○	instance-20250407-174748-dafaflow	OPEN JUPYTERLAB	us-central1-a	M20250324-2200-rc0	Efficient Instance: 2 vCPUs, 16 GB RAM

Dataproc Workbench

Managed notebooks running PySpark and recommendation models.

Pipeline Implementation - Part 1: EDA with Beam

Data Loading

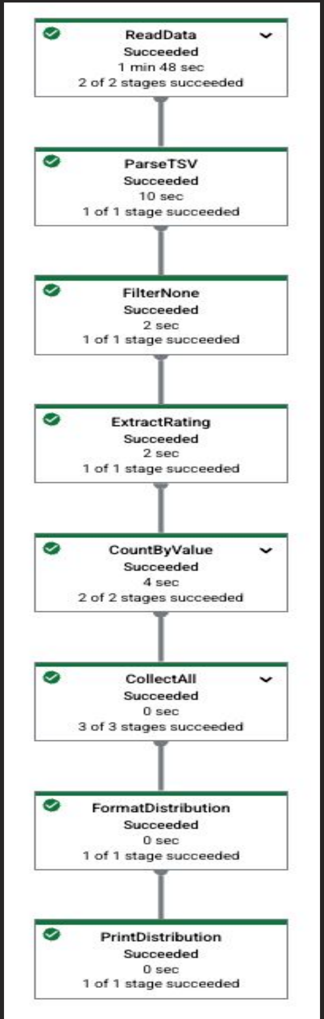
Loaded TSV data from Cloud Storage into Dataflow.

Distributed Processing

Apache Beam pipeline analyzed ratings and user activity.

Key Metrics

3,105,520 records, 1,502,380 users, 779,733 products, avg rating 4.18.



ReadData
Succeeded
1 min 48 sec
2 of 2 stages succeeded

ParseTSV
Succeeded
10 sec
1 of 1 stage succeeded

FilterNone
Succeeded
2 sec
1 of 1 stage succeeded

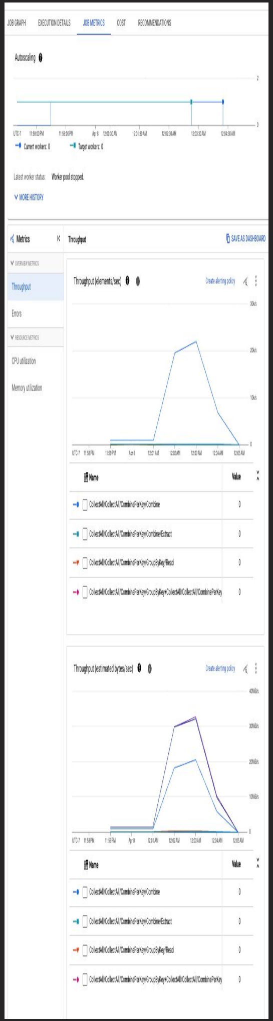
ExtractRating
Succeeded
2 sec
1 of 1 stage succeeded

CountByValue
Succeeded
4 sec
2 of 2 stages succeeded

CollectAll
Succeeded
0 sec
3 of 3 stages succeeded

FormatDistribution
Succeeded
0 sec
1 of 1 stage succeeded

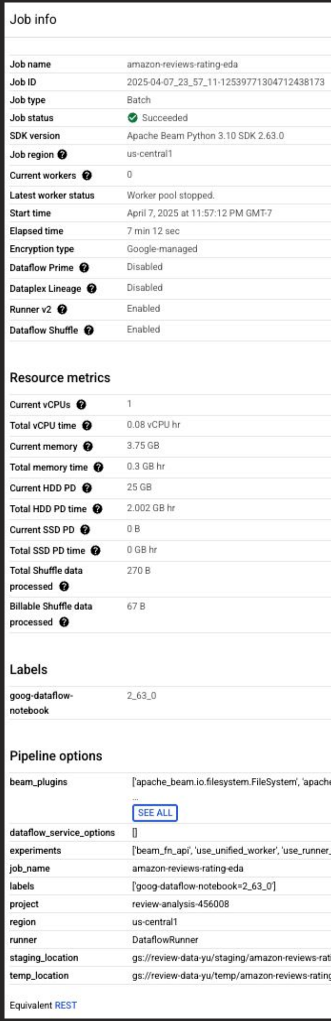
PrintDistribution
Succeeded
0 sec
1 of 1 stage succeeded



Latest worker status: Worker pool stopped.

Name	Value
CollectAll/CollectAll/CombinePerKey/Combine	0
CollectAll/CollectAll/CombinePerKey/Combine/Extract	0
CollectAll/CollectAll/CombinePerKey/GroupByKey/Read	0
CollectAll/CollectAll/CombinePerKey/GroupByKey+CollectAll/CollectAll/CombinePerKey	0

Name	Value
CollectAll/CollectAll/CombinePerKey/Combine	0
CollectAll/CollectAll/CombinePerKey/Combine/Extract	0
CollectAll/CollectAll/CombinePerKey/GroupByKey/Read	0
CollectAll/CollectAll/CombinePerKey/GroupByKey+CollectAll/CollectAll/CombinePerKey	0



Job info

Job name	amazon-reviews-rating-eda
Job ID	2025-04-07_23_57_11-12539771304712438173
Job type	Batch
Job status	✓ Succeeded
SDK version	Apache Beam Python 3.10 SDK 2.63.0
Job region	us-central1
Current workers	0
Latest worker status	Worker pool stopped.
Start time	April 7, 2025 at 11:57:12 PM GMT-7
Elapsed time	7 min 12 sec
Encryption type	Google-managed
Dataflow Prime	Disabled
Dataplex Lineage	Disabled
Runner v2	Enabled
Dataflow Shuffle	Enabled

Resource metrics

Current vCPUs	1
Total vCPU time	0.08 vCPU hr
Current memory	3.75 GB
Total memory time	0.3 GB hr
Current HDD PD	25 GB
Total HDD PD time	2.002 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr
Total Shuffle data processed	270 B
Billable Shuffle data processed	67 B

Labels

goog-dataflow-notebook	2_63_0

Pipeline options

beam_plugins	['apache_beam.io.filesystem.FileSystem', 'apache ... SEE ALL
dataflow_service_options	[]
experiments	['beam_fn_api', 'use_unified_worker', 'use_runner_
job_name	amazon-reviews-rating-eda
labels	['goog-dataflow-notebook=2_63_0']
project	review-analysis-456008
region	us-central1
runner	DataflowRunner
staging_location	gs://review-data-yu/staging/amazon-reviews-rati
temp_location	gs://review-data-yu/temp/amazon-reviews-rating

Equivalent REST

Part 2: Data Preprocessing

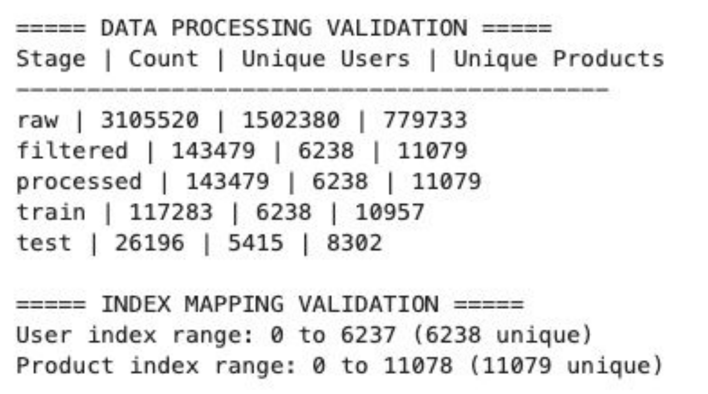
Data Loading & BigQuery Setup: Imports and organizes raw data.

High-Frequency Filtering: Focuses on active users and products.

ID Mapping & Processing: Creates numerical IDs for modeling.

Train/Test Data Split: Prepares data for model development.

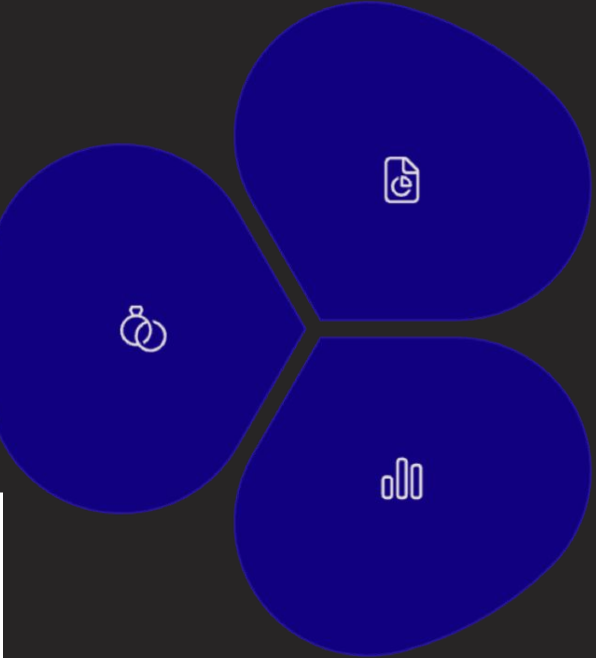
Validation: Ensures data integrity throughout the process.



```
===== DATA PROCESSING VALIDATION =====
Stage | Count | Unique Users | Unique Products
-----------------------------------------
raw | 3105520 | 1502380 | 779733
filtered | 143479 | 6238 | 11079
processed | 143479 | 6238 | 11079
train | 117283 | 6238 | 10957
test | 26196 | 5415 | 8302

===== INDEX MAPPING VALIDATION =====
User index range: 0 to 6237 (6238 unique)
Product index range: 0 to 11078 (11079 unique)
```

Part 3: Recommendation Models



ALS

Alternating Least Squares using Spark MLlib.

Distributed matrix factorization algorithm.

SVD

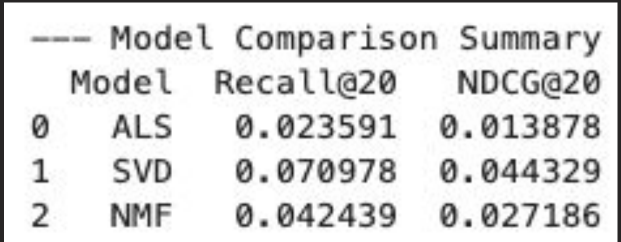
Singular Value Decomposition using scikit-learn.

Classical matrix factorization technique.

NMF

Non-negative Matrix Factorization using scikit-learn.

Optimized for sparse matrix operations.



```
--- Model Comparison Summary
  Model  Recall@20   NDCG@20
0   ALS   0.023591  0.013878
1   SVD   0.070978  0.044329
2   NMF   0.042439  0.027186
```

Challenges & Solutions

Scale of the Dataset

Problem: 3.1M reviews too large for single-machine processing.

Solution: Used Dataflow for distributed EDA and BigQuery for filtering.

Sparse User-Item Matrix

Problem: Very sparse interaction matrix.

Solution: Filtered for users/products with ≥30 interactions.

Debugging

Problem: Hard to trace issues in complex pipeline logic.

Solution: Used Jupyter Notebook as the main environment to quickly test and debug.

Future Work

- Submit Spark to Dataproc for distributed training
- Process reviews with NLP
- Hybrid recommendation models, fine tuning
- Deploy as web service with API access

THANK YOU