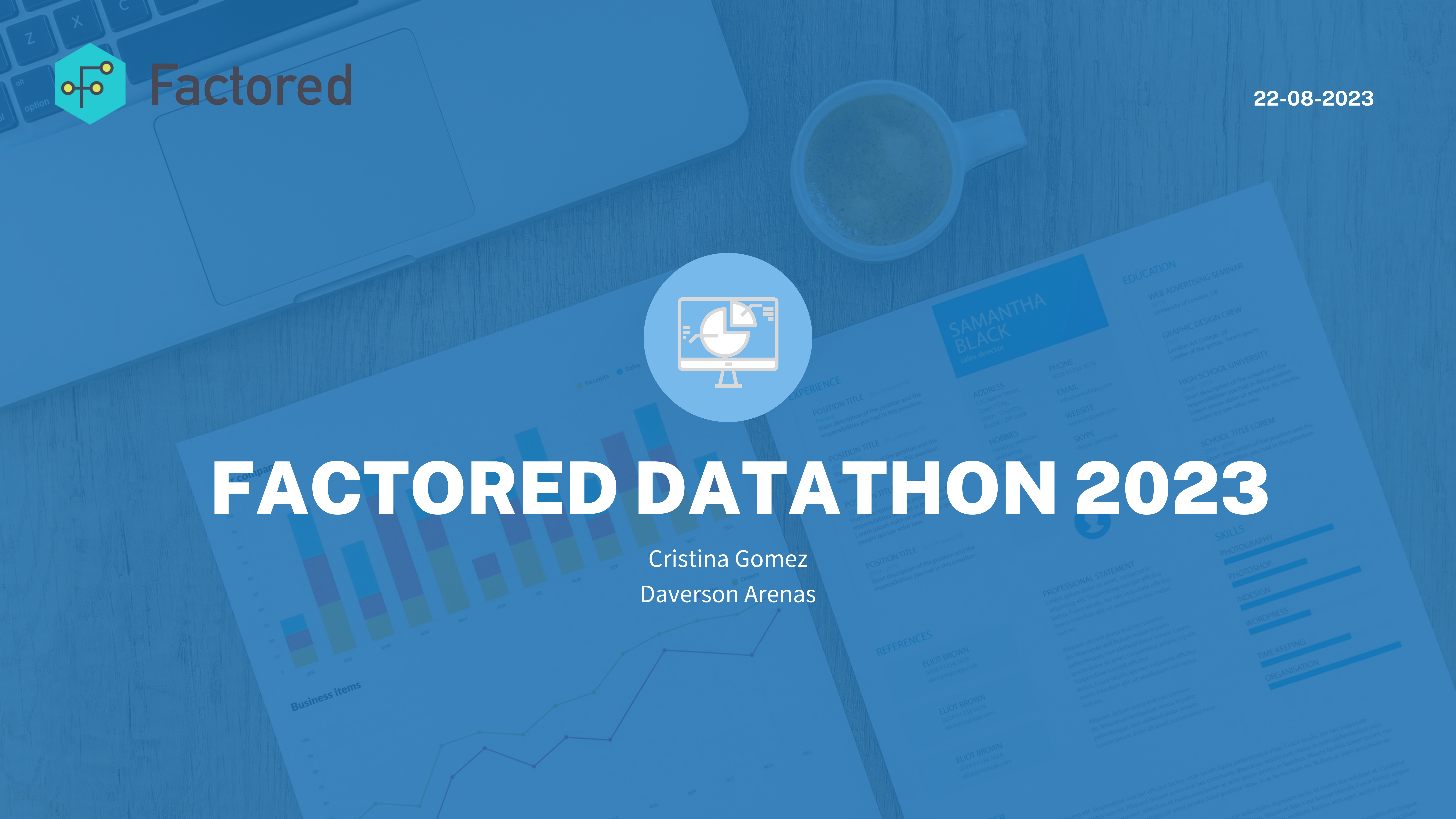
Factored

22-08-2023



FACTORED DATATHON 2023

Cristina Gomez
Daverson Arenas

PAISA GENIOUS TEAM

CRISTINA GOMEZ

DAVERSON ARENAS

Bioengineer and Data Scientist:

As the lead data scientist in our project Cristina:

- Managed feature engineering, data cleansing, and conducted data mining to create insightful visualizations
- Constructed the gold data set, essential for visualization and ML models
- Executed machine learning models for sentiment analysis, enhancing project insights

Mechanical Engineer and Data Engineer

As the lead data engineer in our project Daverson:

- Orchestrated Databricks workspace deployment on AWS, integrating Github for efficient CI/CD.
- Designed lakehouse data architecture, unifying data warehouses and data lakes.
- Created ELT pipelines for batch and streaming incremental data ingestion

Factored

The Challenge

Create an innovative data solution (web apps, chatbots, dashboards, model interfaces...) to empower businesses with insights from product reviews.

Datathon Challenge Roadmap

Stage 1

Understand the problem and define a solution approach.

Day 1 ● Day 2

Stage 2

Choose technologies, deploy services, set up workspace and create GitHub repo. First batch ingestion

Day 3 ● Day 5

Stage 3

Data architecture design, data engineering for batch and stream ingestion. Exploratory data analysis

Day 6 ○ Day 8

Stage 4

Data engineering for streaming, data cleansing, feature engineering and data visualization. Combine the data from both sources

Day 9 ○ Day 11

Stage 5

Machine learning models, frontend design, final design dashboard, documentation and presentation

Day 12 ○ Day 13

MAIN TECHNOLOGIES

Scalability, Parallelism and Speed

DATABRICKS + APACHE SPARK

- Databricks combines data warehouses & data lakes into a lakehouse architecture.

heterogeneous hardware

AWS + DELTA LAKE

- Amazon S3 serves as the data lake, coupled with Delta Lake, which functions as the storage layer
- AWS EC2 instances as the compute resources for Databricks clusters

Optimized resource utilization

MLFLOW + TEXTBLOB + NLTK

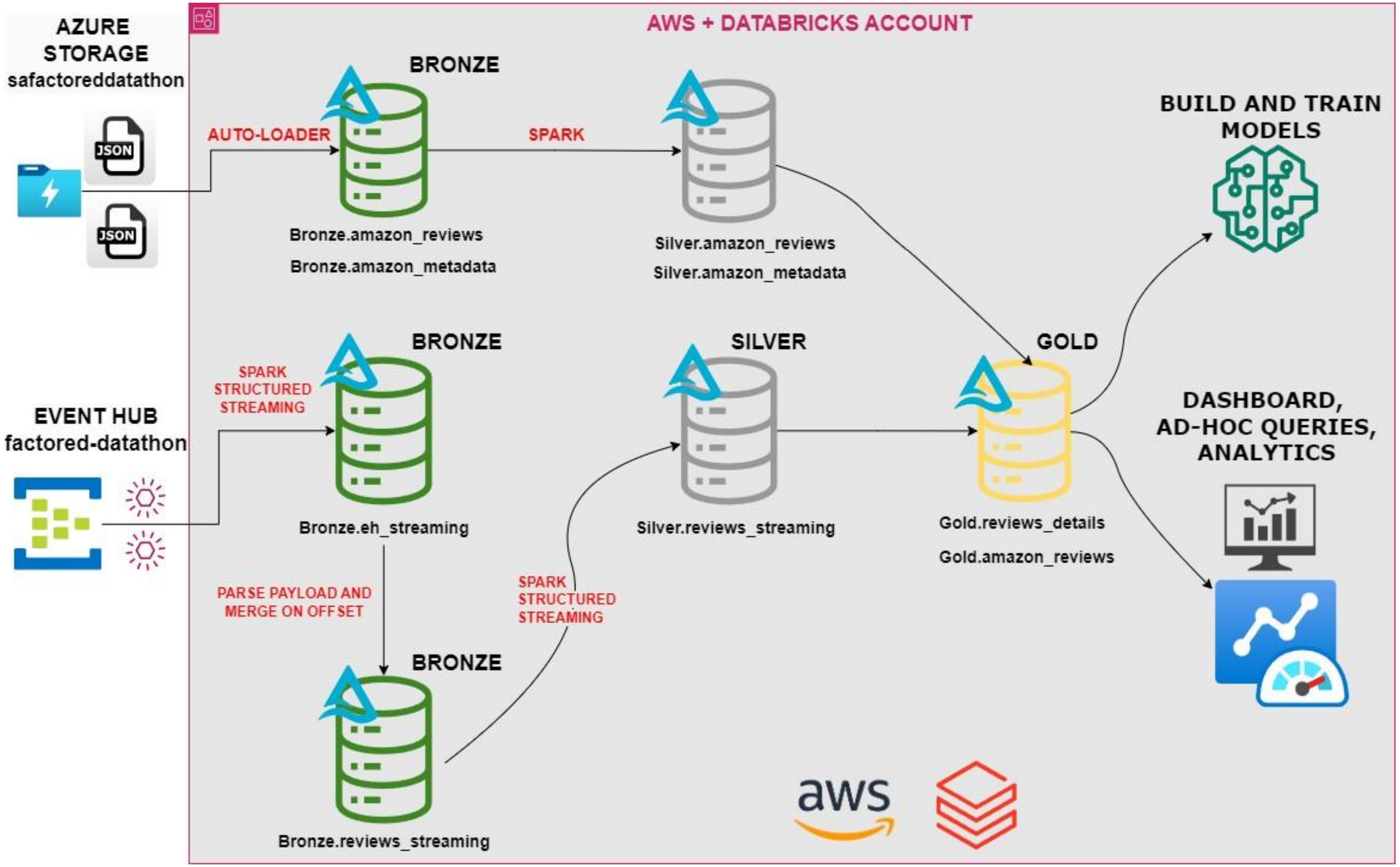
- Distributed capabilities for large-scale experiments
- Efficient experiment tracking, reproducibility and scalability

DATA ARCHITECTURE

For this challenge, Team Paisa Genious proposes a Delta Lakehouse architecture using Databricks with AWS Cloud, which offers an integrated solution for both batch and streaming data processing

- **Databricks:**
  Big data platform

- **AWS:**
  Storage
  Compute resources

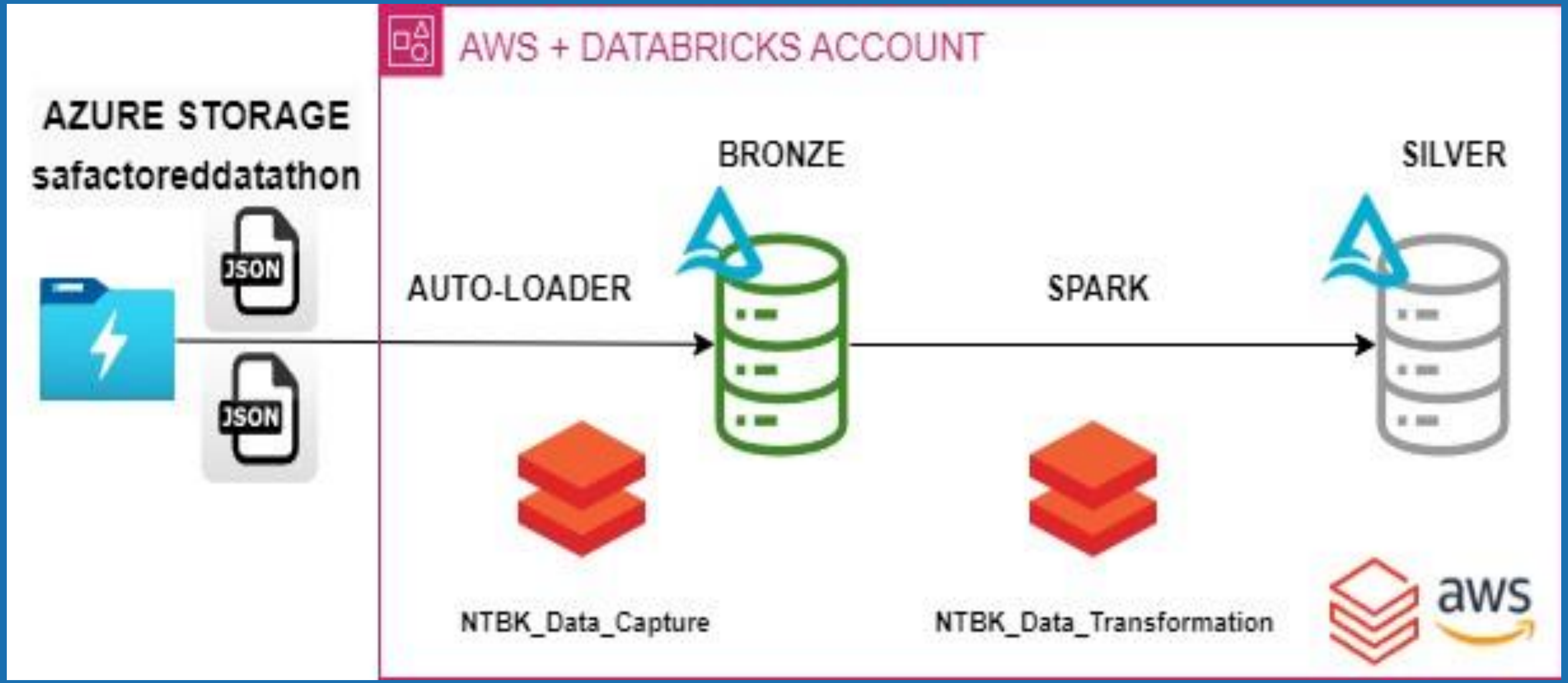


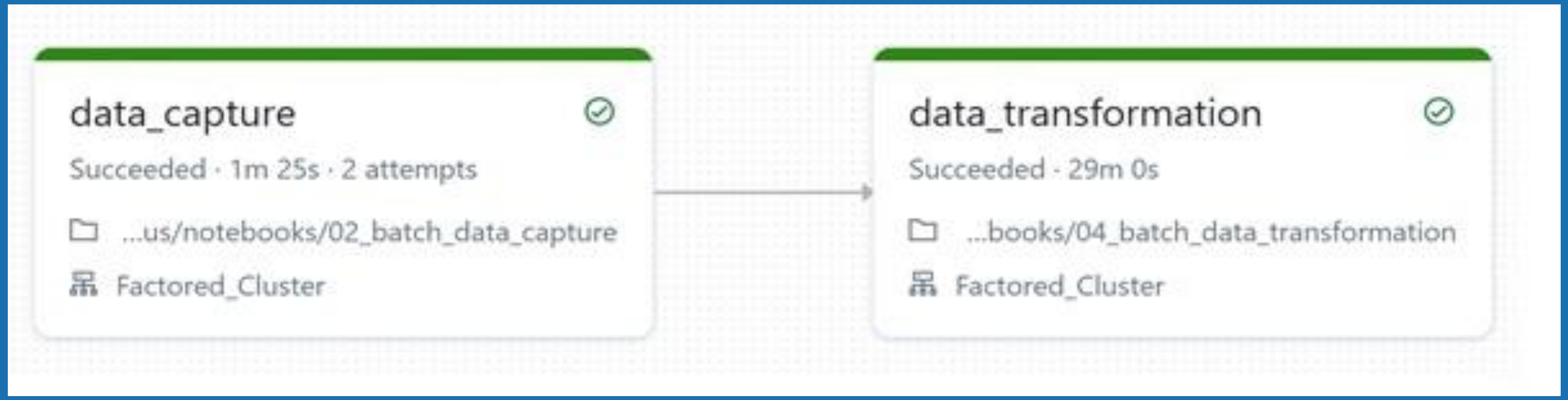
DATA ENGINEERING WITH DATABRICKS

BATCH

1. Extract and Load data to the Lake house
Auto-loader: Incrementally load new data files as they arrive

2. Transform, Clean and Filter
Spark structured streaming: Processing and transformation tasks with a batch-like behavior





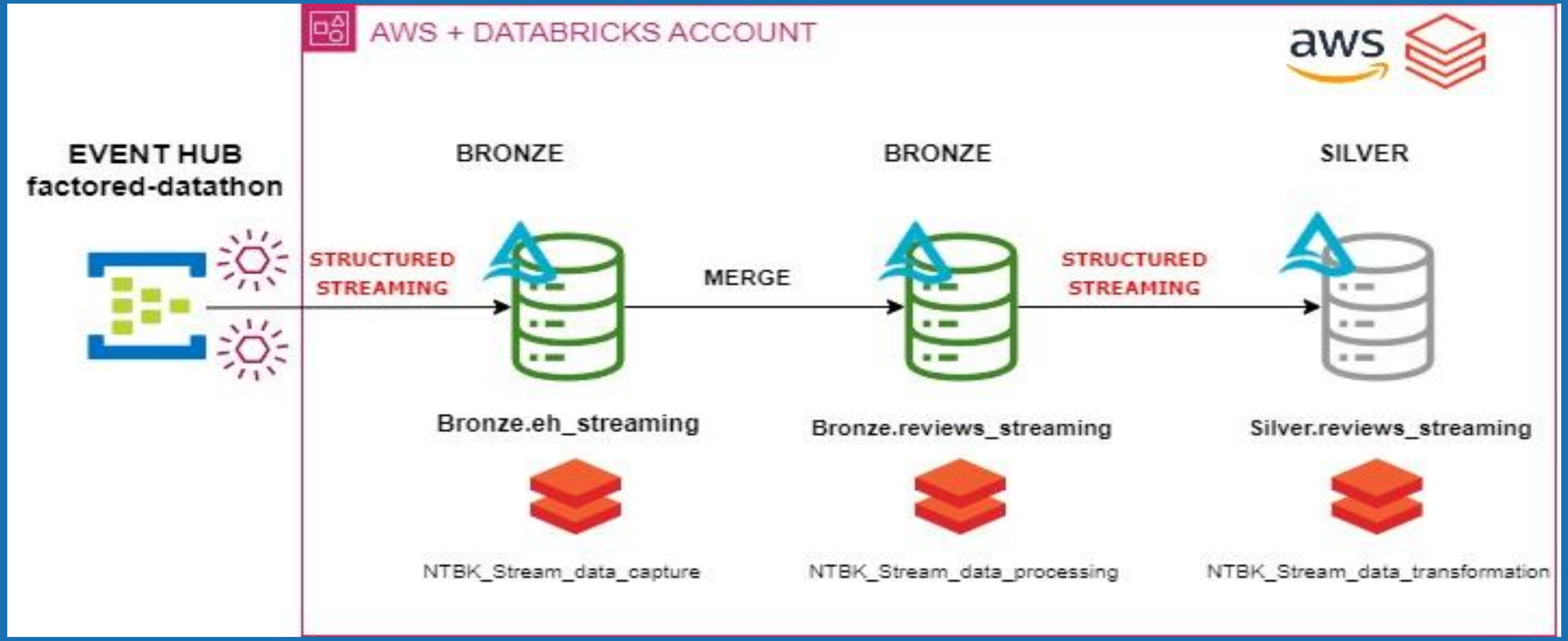
data_capture
Succeeded · 1m 25s · 2 attempts
...us/notebooks/02_batch_data_capture
Factored_Cluster

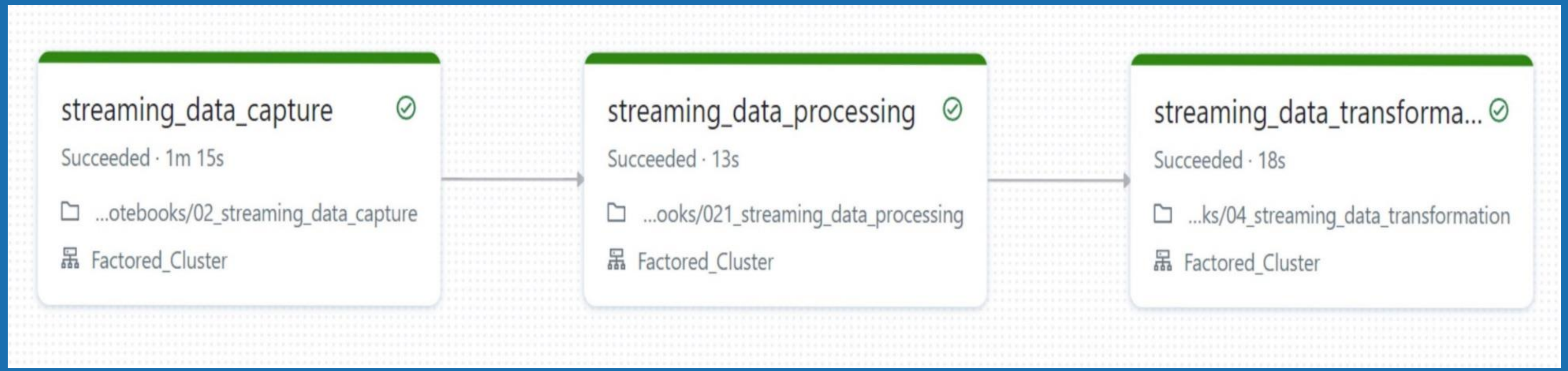
data_transformation
Succeeded · 29m 0s
...books/04_batch_data_transformation
Factored_Cluster

STREAMING

1. Extract and Load data to the Lake house
Spark structured streaming: Stream data from event hub

2. Transform Data, Clean and Filter
Spark structured streaming: Processing and transforming tasks as data becomes available





streaming_data_capture
Succeeded · 1m 15s
...otebooks/02_streaming_data_capture
Factored_Cluster

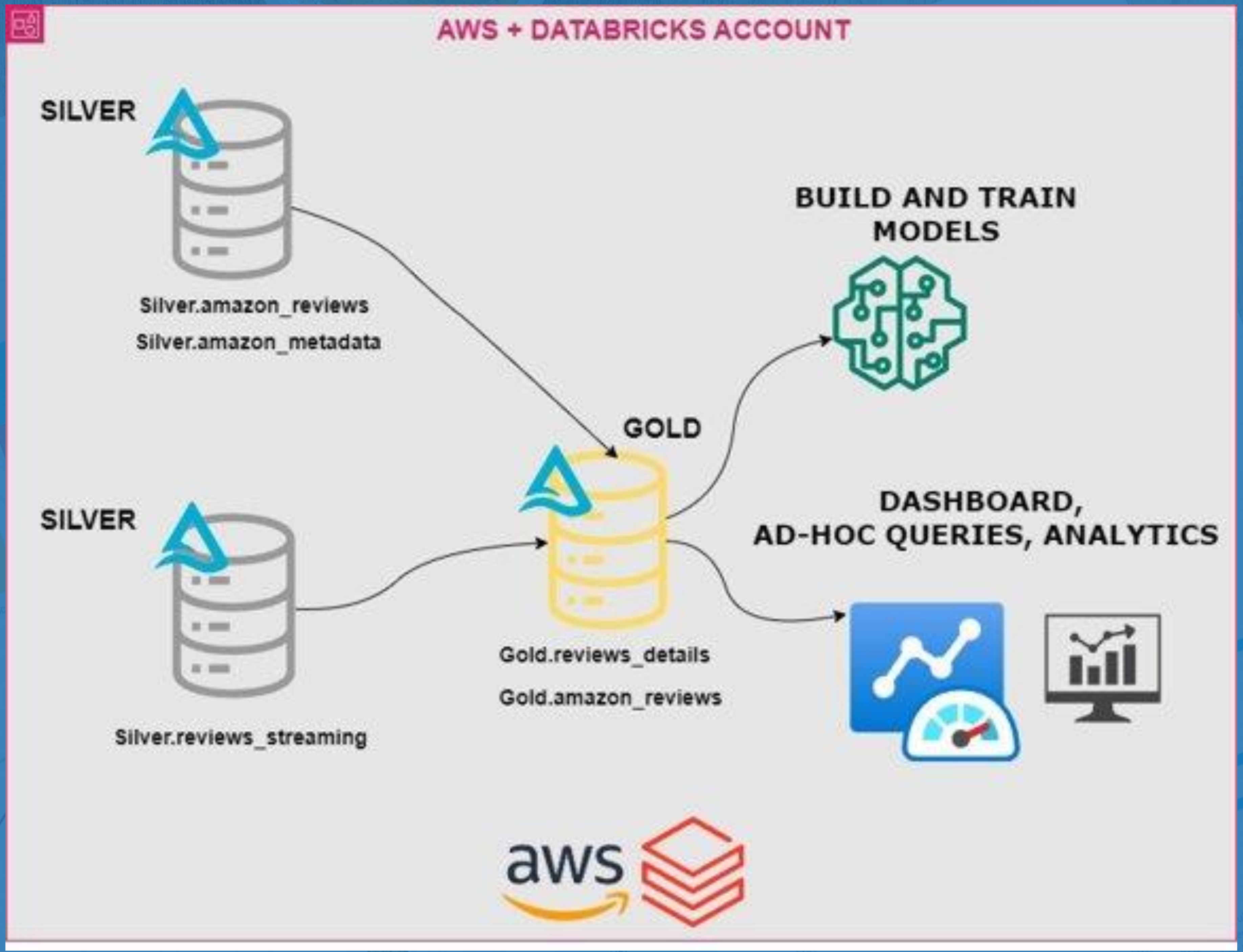
streaming_data_processing
Succeeded · 13s
...ooks/021_streaming_data_processing
Factored_Cluster

streaming_data_transforma...
Succeeded · 18s
...ks/04_streaming_data_transformation
Factored_Cluster

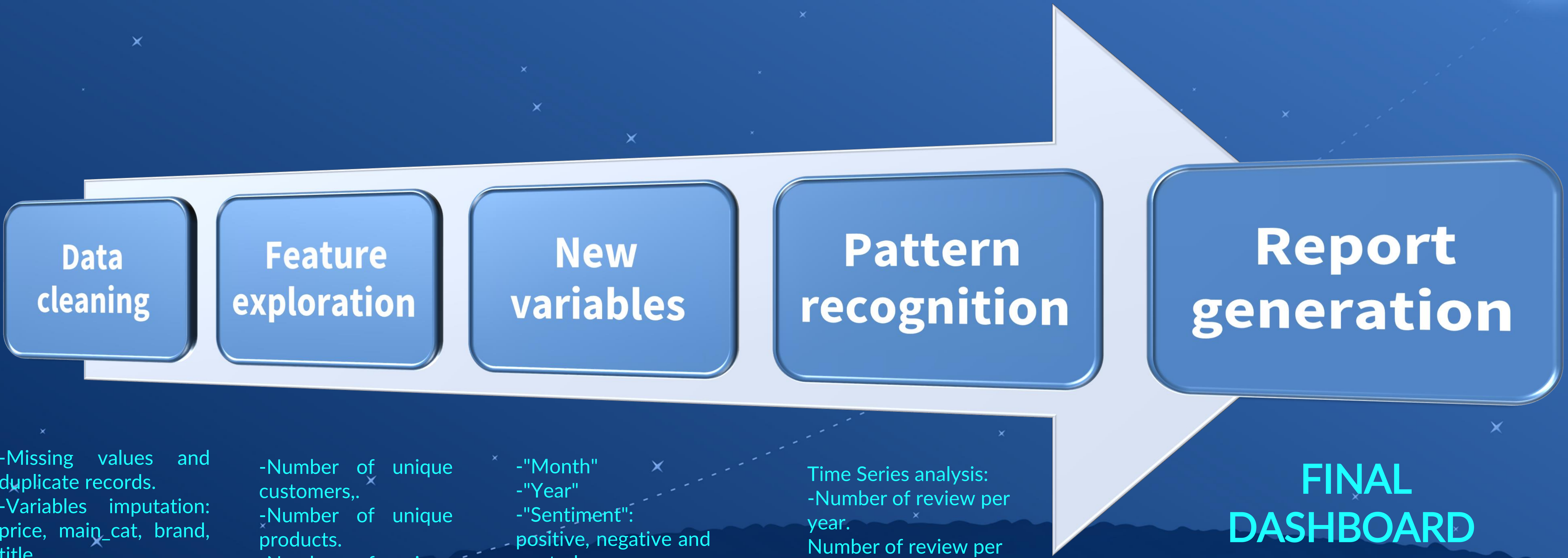
DATA ENGINEERING WITH DATABRICKS

Gold Layer: Business level Aggregates

The Gold layer aims to deliver continuously updated, clean data to downstream users and applications, including machine learning models, ad-hoc queries, and analytics tools.



EXPLORATORY DATA ANALYSIS



Data cleaning

Feature exploration

New variables

Pattern recognition

Report generation

-Missing values and duplicate records.
-Variables imputation: price, main_cat, brand, title.
-Removing unwanted characters, converting to lowercase, and handling special cases: reviewText, title, main_cat, brand

-Number of unique customers,.
-Number of unique products.
-Number of unique reviews

-"Month"
-"Year"
-"Sentiment": positive, negative and neutral
-Number of words per review

Time Series analysis:
-Number of review per year.
Number of review per month.
-average of overall ratings per year
-Setiment analysis per year

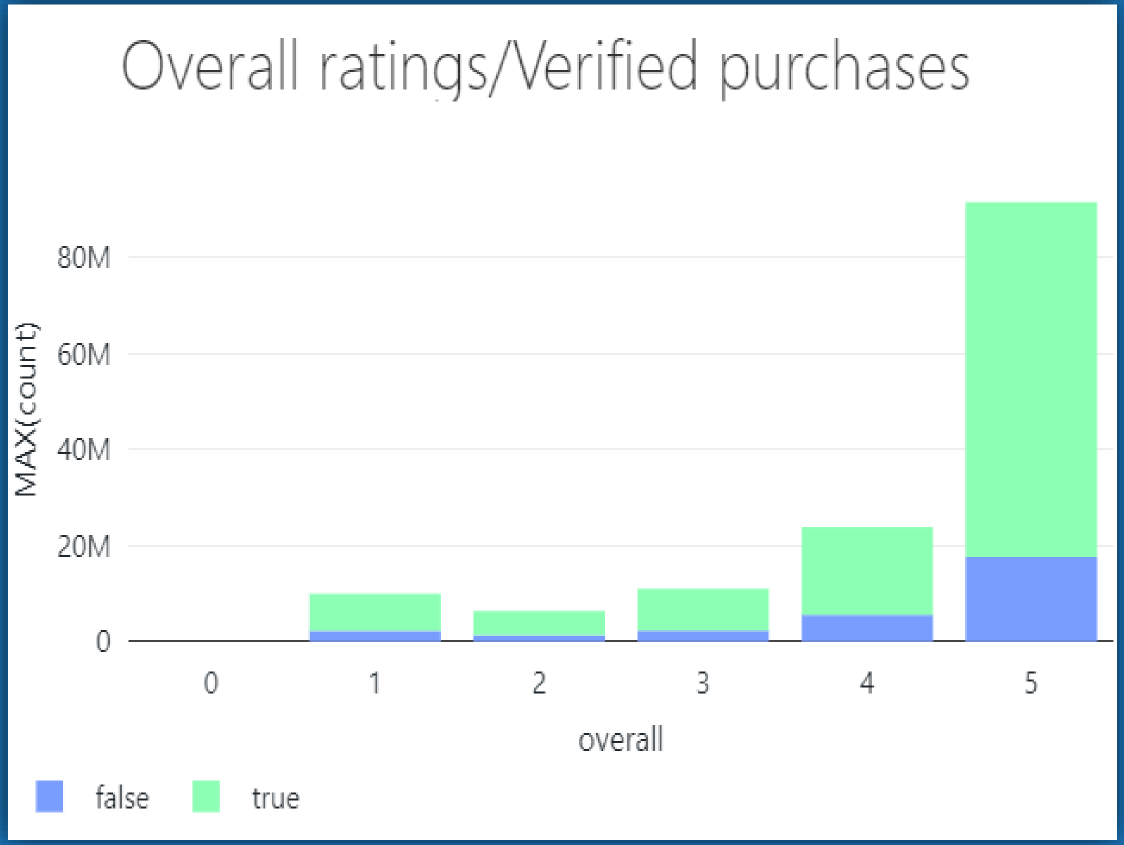
FINAL DASHBOARD

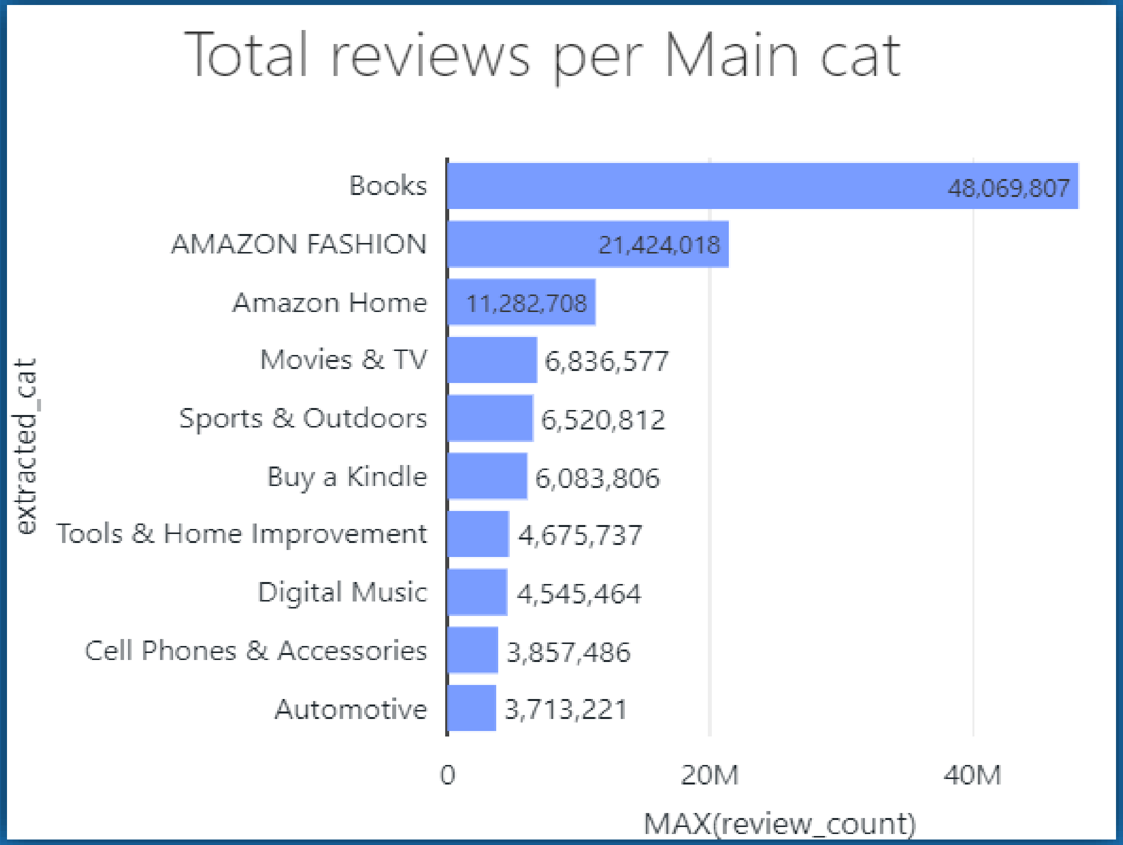
REPORT FROM HISTORICAL DATA

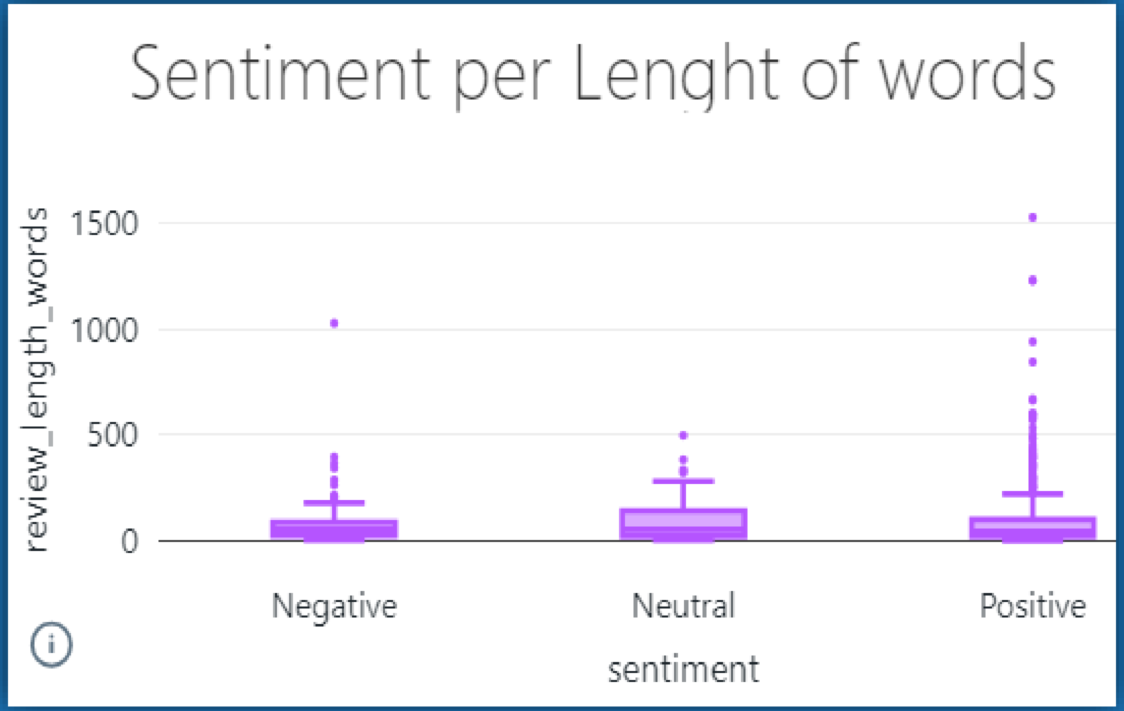
Review Verification Impact: Does sentiment differ between verified and non-verified reviews, and do verified reviews tend to be more credible?

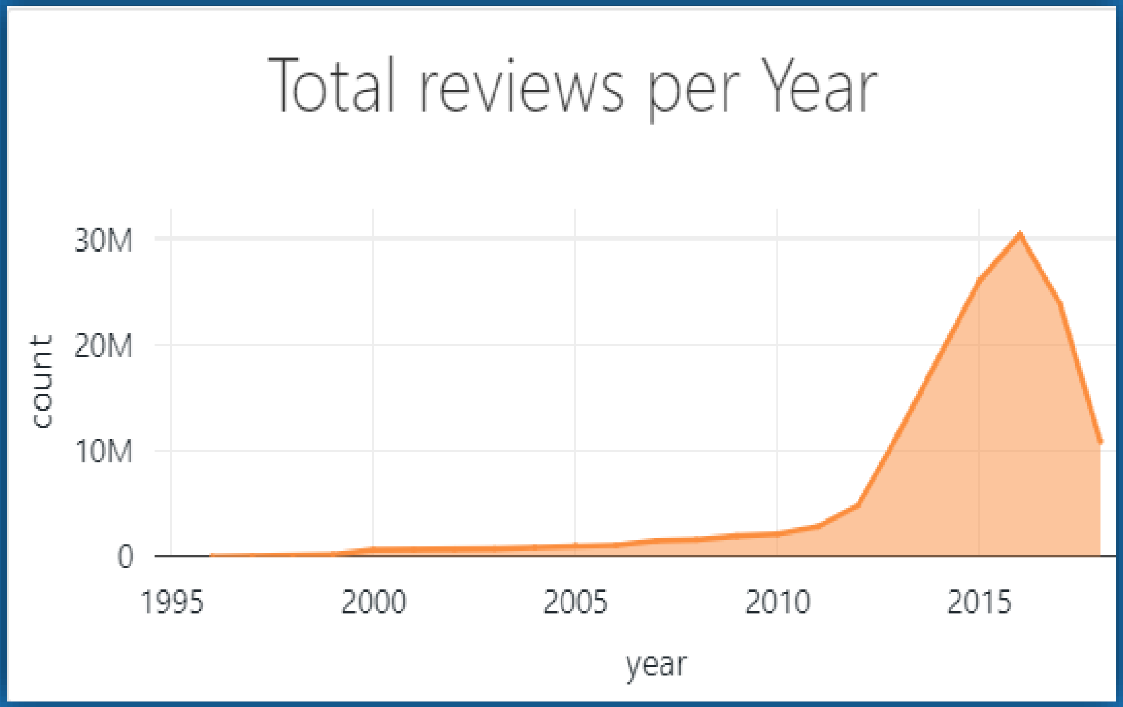
Review Length and Sentiment: Are longer or shorter reviews more likely to have a positive or negative sentiment?

Product Prioritization: Which product categories receive the most feedback, and are they also the ones with higher satisfaction?


Overall ratings/Verified purchases


Total reviews per Main cat


Sentiment per Lenght of words

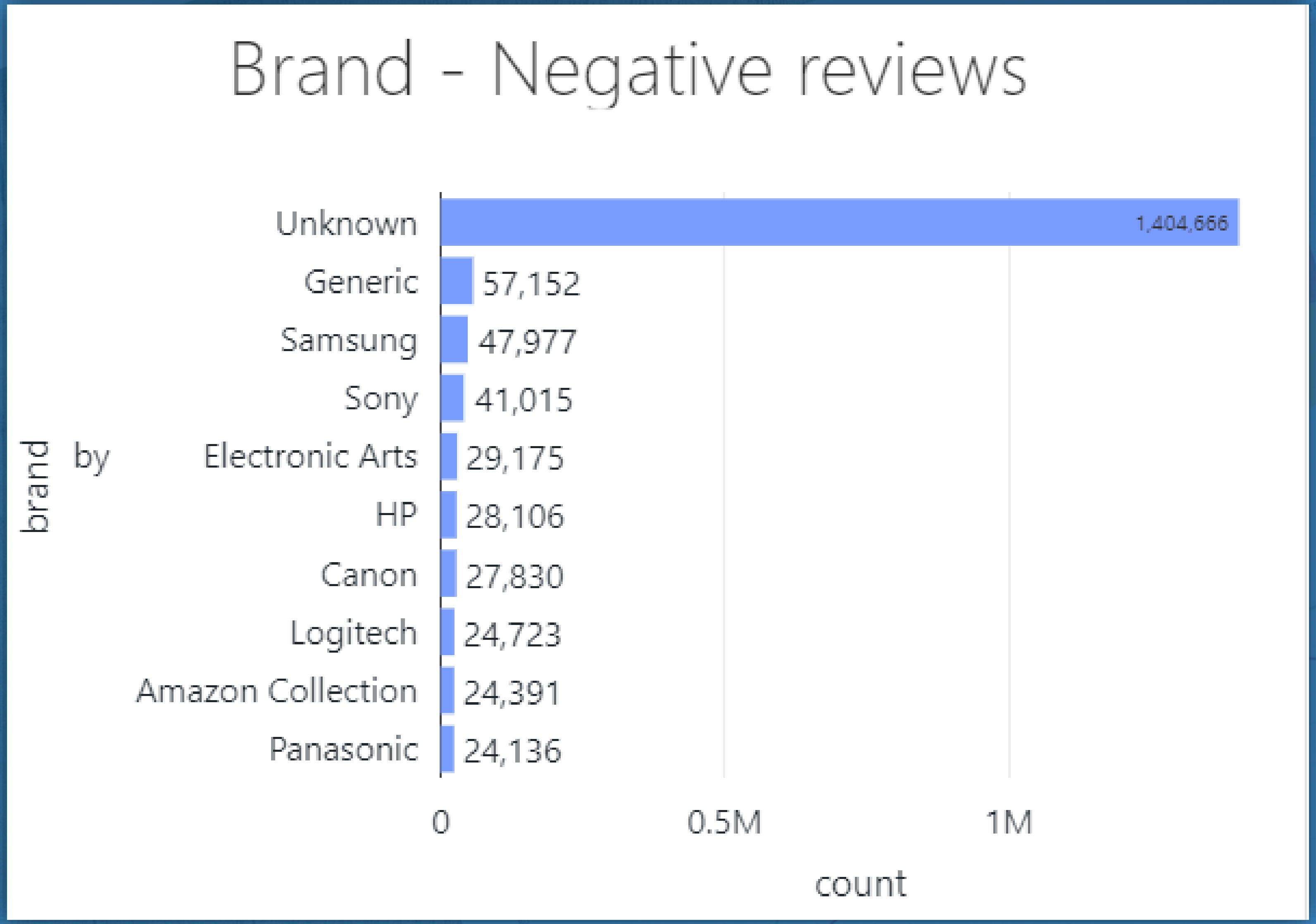

Total reviews per Year

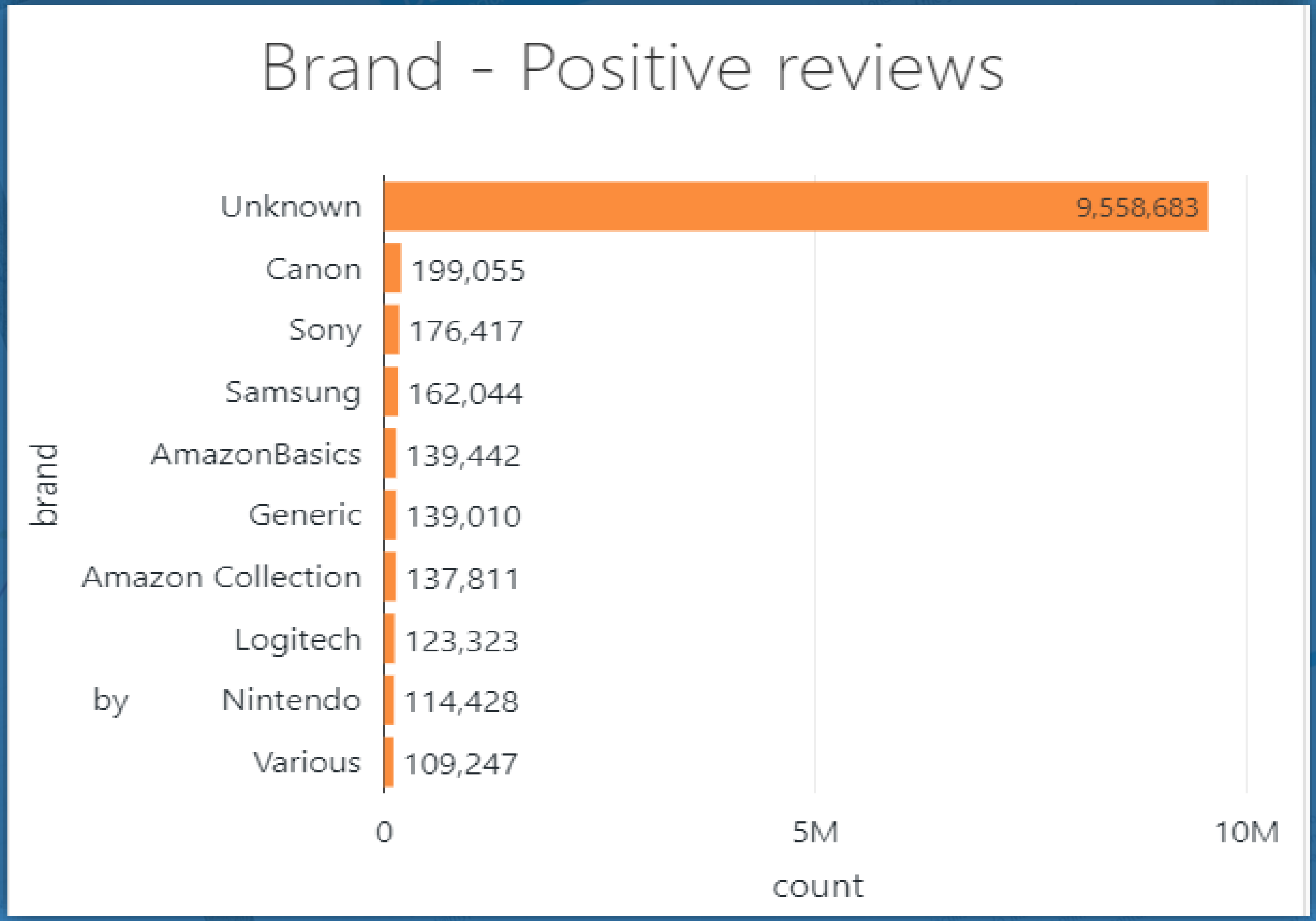
SENTIMENT ANALISYS FROM HISTORICAL + STREAMING DATA

BRAND REPUTATION ASSESSMENT

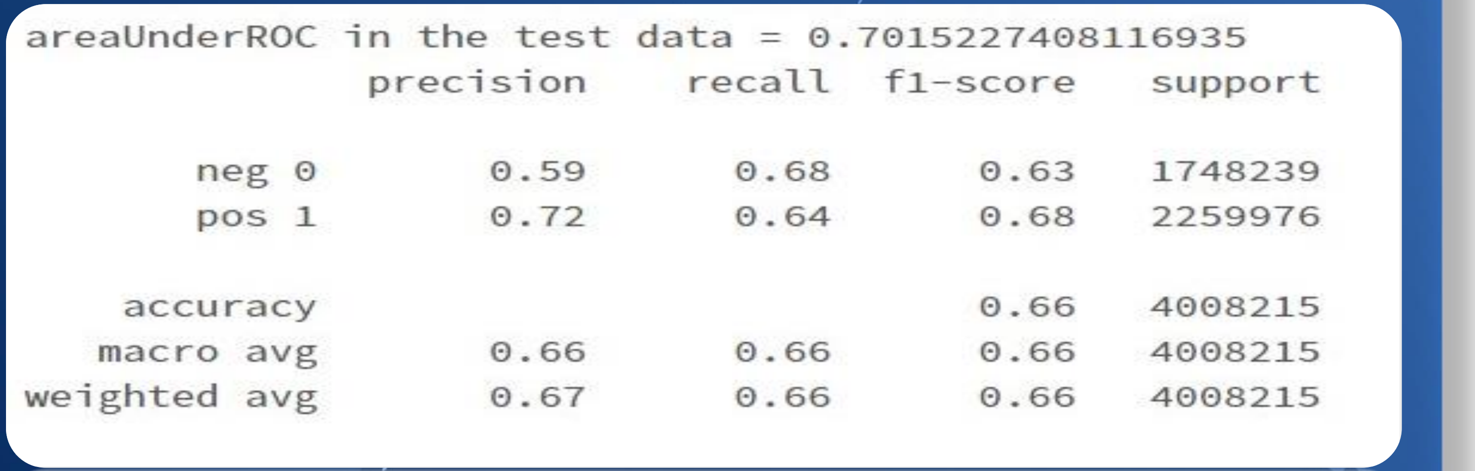
STRATEGIC DECISION MAKING

CUSTOMER PERCEPTION


Brand - Negative reviews


Brand - Positive reviews

SENTIMENT ANALYSIS ML MODEL



```
areaUnderROC in the test data = 0.7015227408116935
              precision    recall  f1-score   support

       neg 0       0.59      0.68      0.63   1748239
       pos 1       0.72      0.64      0.68   2259976

    accuracy                           0.66   4008215
   macro avg       0.66      0.66      0.66   4008215
weighted avg       0.67      0.66      0.66   4008215
```

DELTA LAKE

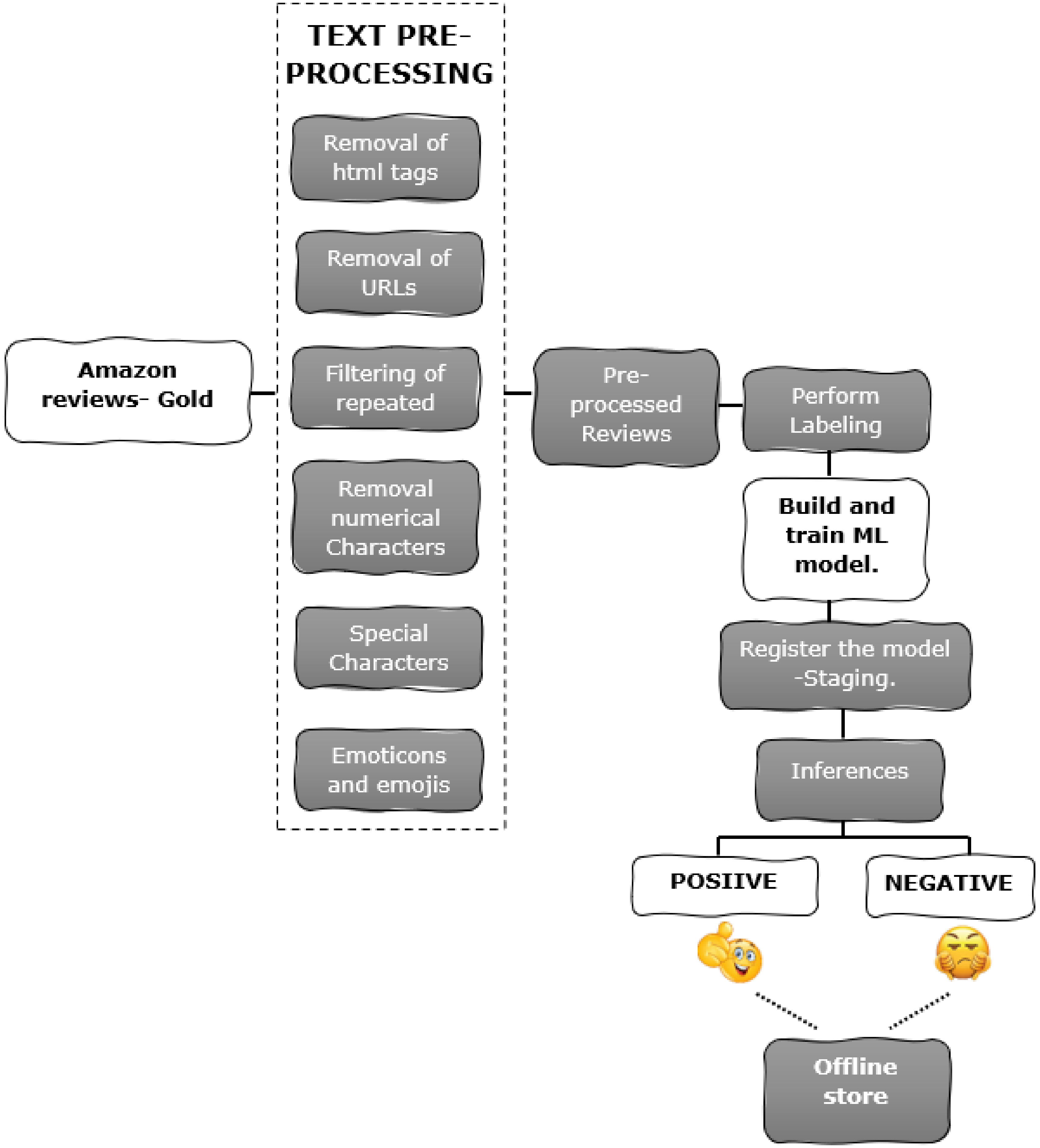
databricks

APACHE Spark

TextBlob

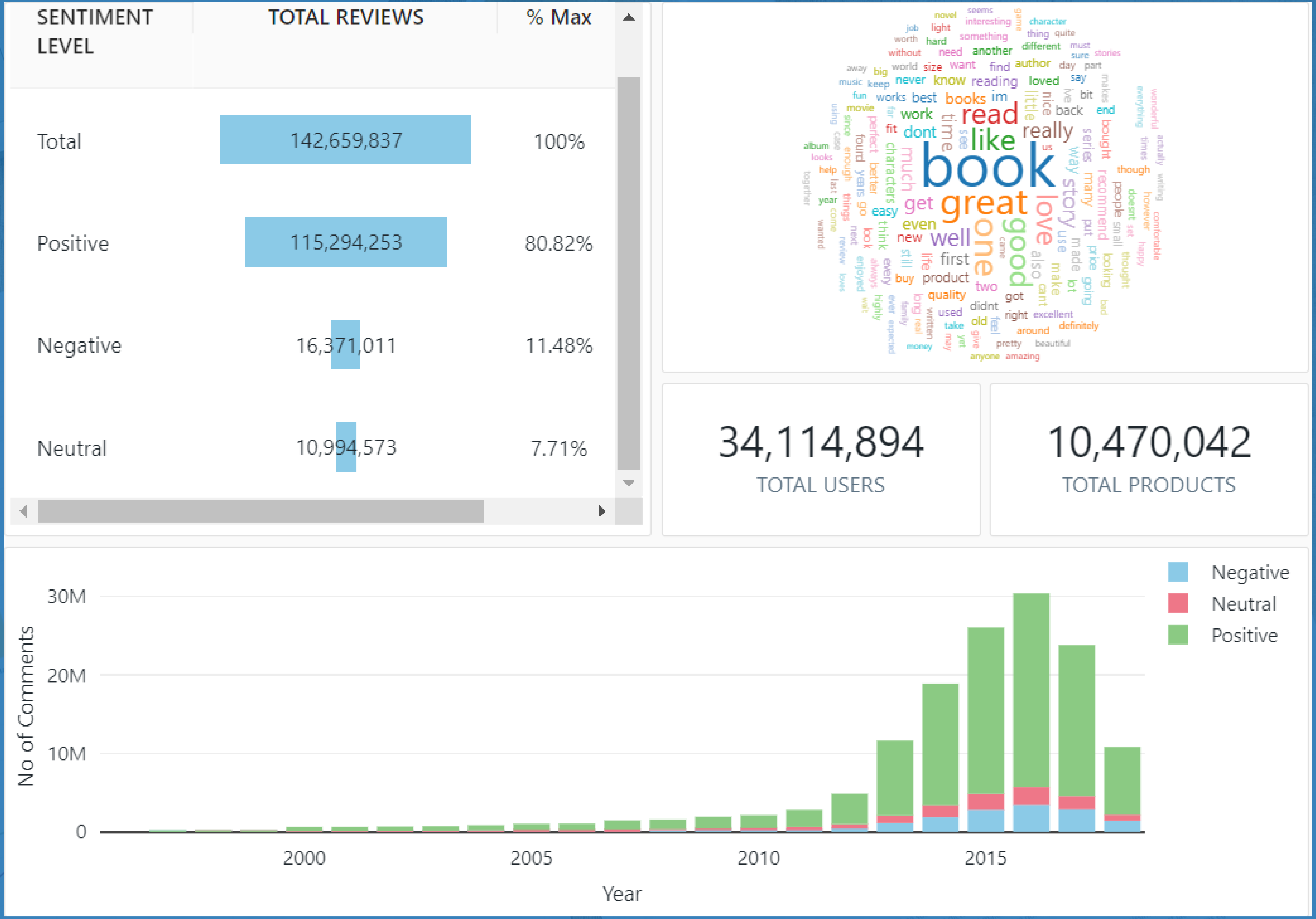
NLTK

mlflow



SENTIMENT ANALISYS DASHBOARD

- □ Identifying Key Themes
- □ Prioritizing Focus Areas
- □ Sentiment Overview
- □ Monitoring Trends
- □ Customer Engagement
- □ Historical Insights
- □ Seasonal Patterns

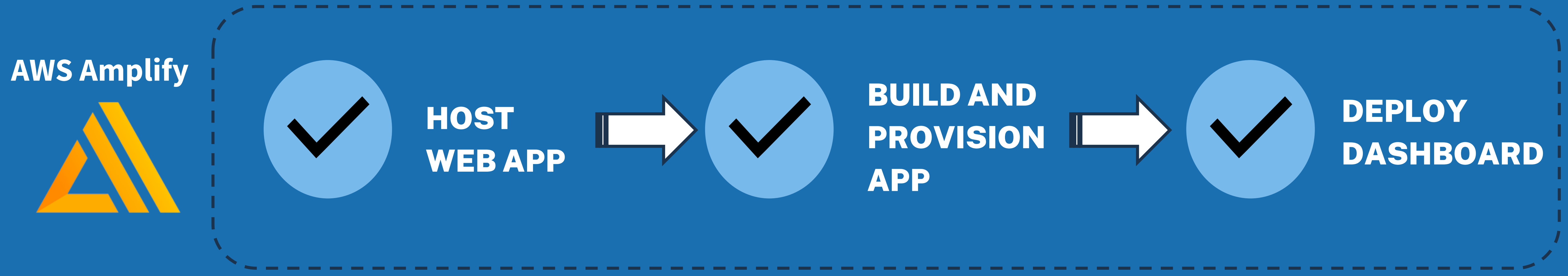


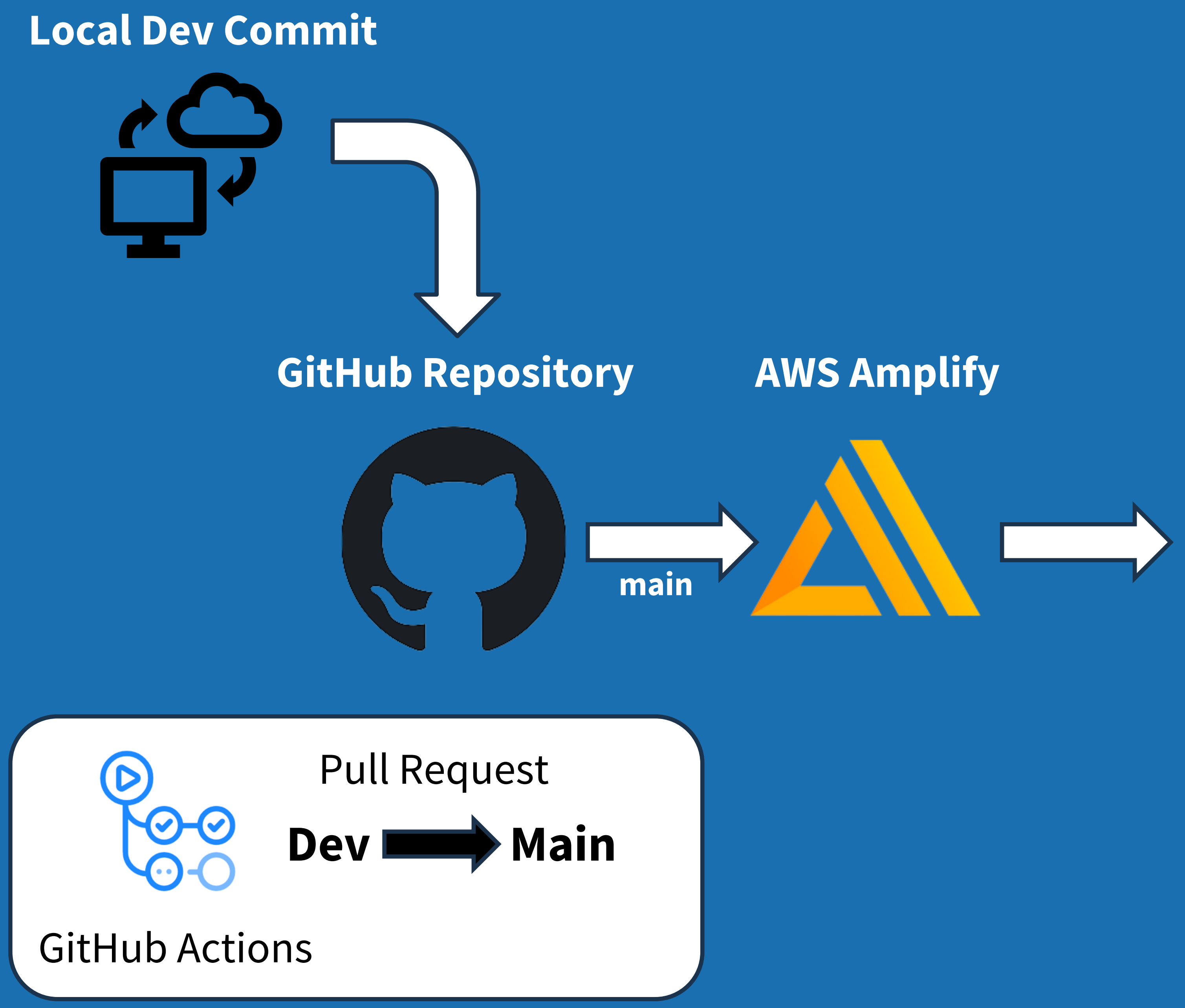
SENTIMENT LEVEL	TOTAL REVIEWS	% Max
Total	142,659,837	100%
Positive	115,294,253	80.82%
Negative	16,371,011	11.48%
Neutral	10,994,573	7.71%

34,114,894
TOTAL USERS

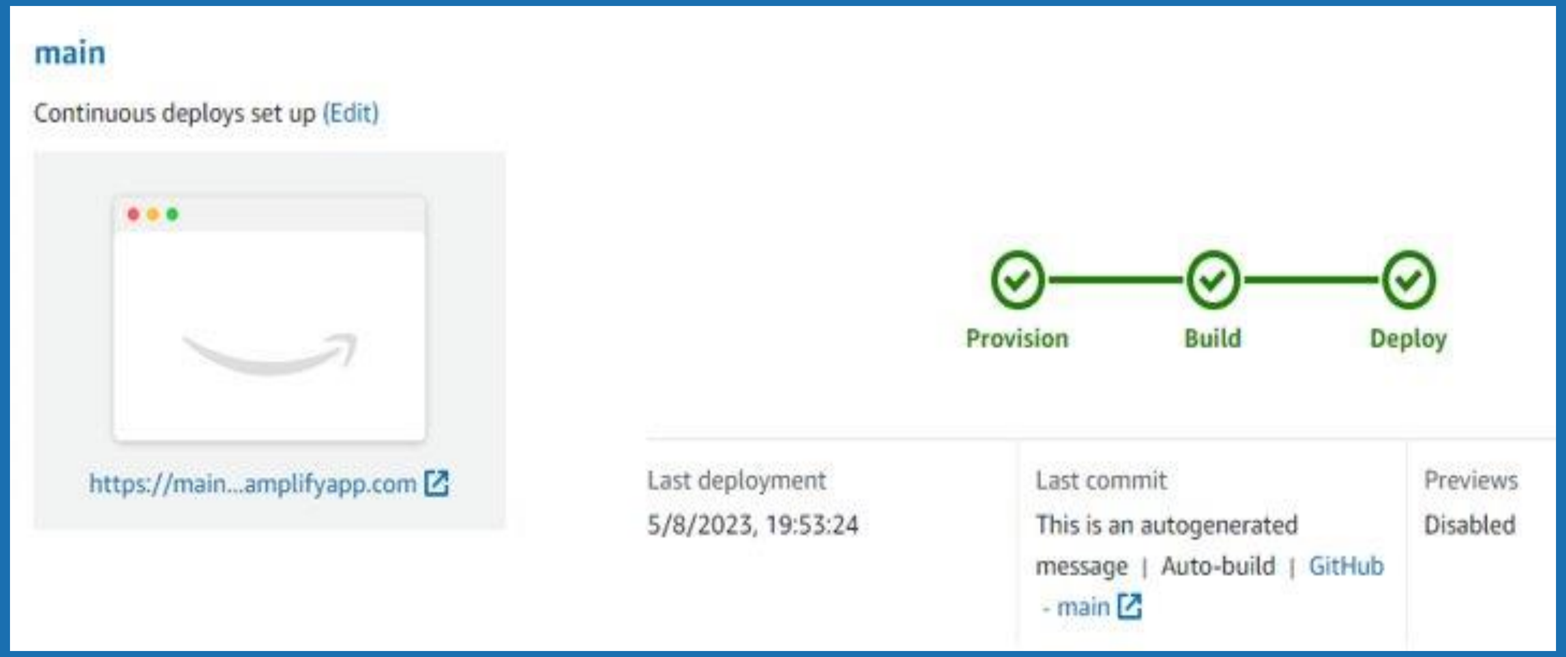
10,470,042
TOTAL PRODUCTS

FRONTEND AND CONTINOUS DEPLOYMENT





Automatic Deployment



main

Continuous deploys set up (Edit)

https://main...amplifyapp.com

Provision — Build — Deploy

Last deployment
5/8/2023, 19:53:24

Last commit
This is an autogenerated message | Auto-build | GitHub - main

Previews
Disabled

Frontend and Dashboard

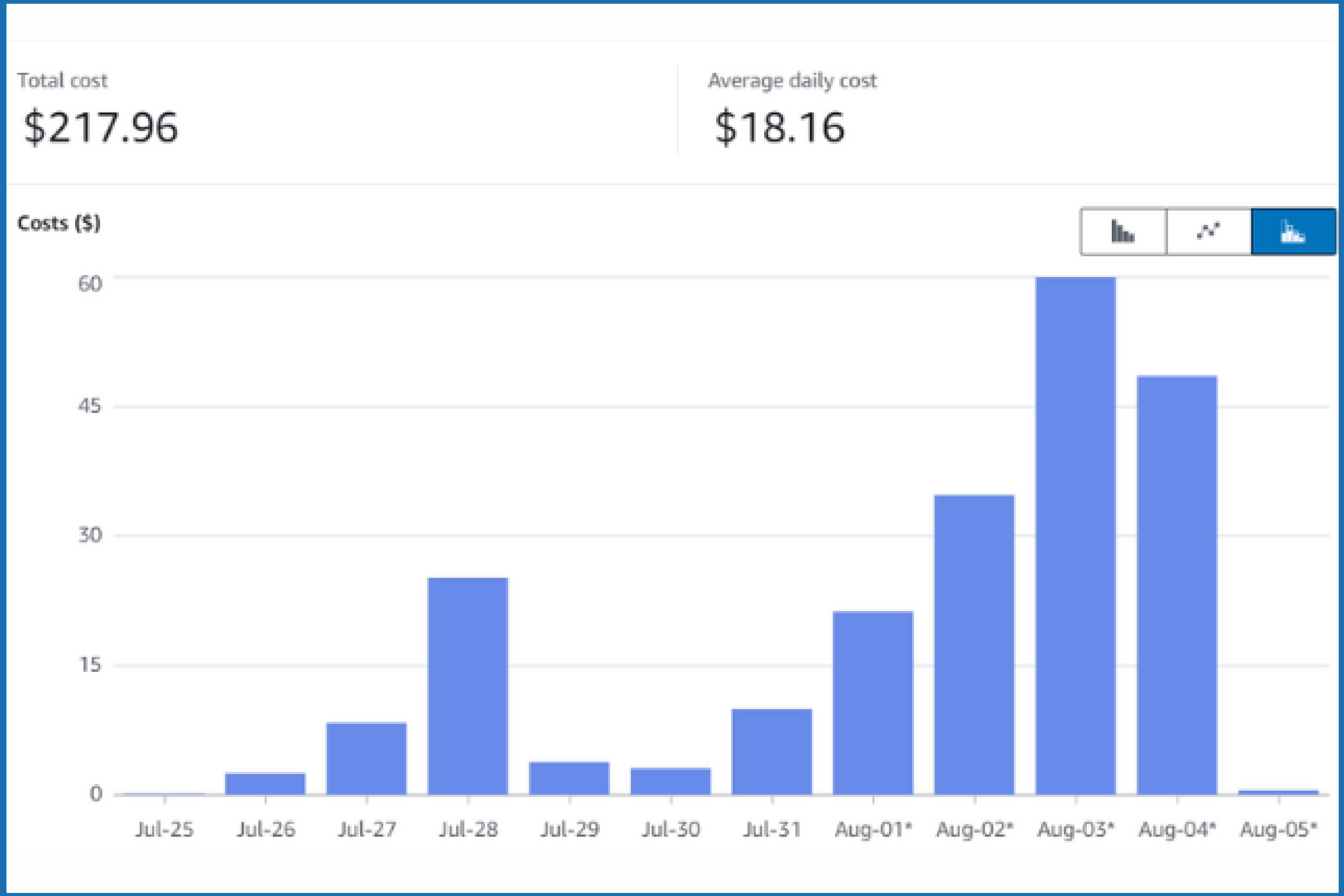
CHALLENGES AND CONCLUSIONS

SMALL TEAM

LIMITED TIME FRAME

LIMITED RESOURCES

AWS DAILY COST RESOURCES



Total cost
$217.96

Average daily cost
$18.16

PAISA GENIUS TEAM

22-08-2023

THANKS FOR YOUR ATTENTION