A Project Report

On

SCALABLE BIG DATA ANALYSIS OF AMAZON REVIEWS USING AZURE CLOUD

By

Meet A. Patel (Student # 200513407)

And

Rishi B. Patel (Student # 200529611)

Under the esteemed guidance of

Prof. (Dr.) Lisa Fan

CS 714 – Big Data Analytics & Cloud Computing *(Summer 2025)*

M.Sc. Computer Science (Data Science)

ACKNOWLEDGEMENT

We would also like to extend our heartfelt thanks to **Dr. Lisa Fan** for providing the opportunity to undertake this challenging and fascinating project. We appreciate her guidance and encouragement throughout. Her teachings have continually inspired us, and we deeply appreciate all she has done for the CS-714 class.

We also offer heartfelt thanks to our parents, relatives and friends, our well-wishers, whose constant support and many sacrifices make our work possible.

ABSTRACT

Our project applies the concepts of big data analytics and cloud computing techniques to transform large-scale Amazon product reviews into actionable insights. Using the UCSD Amazon Reviews 2023 dataset, we analyse two distinct product categories, All_Beauty and Electronics, which has millions of reviews collected over a timespan of 2 decades. Our solution is built using Microsoft Azure and follows the medallion Architecture. Raw JSON data files are ingested into Azure Data Lake Storage Gen2 (ADLS Gen2) (Bronze tier), further the data is cleaned and standardized using Azure Databricks (Silver tier), and is delivered as table format (Parquet) to Azure Synapse (Gold tier) for fast and efficient SQL access to design interactive dashboards. The dashboards highlight key performance indicators such as review volume, average rating, and long-term sentiment trends. The outcome is a reproducible, scalable end-to-end big data pipeline and visualization dashboards that enables businesses, sellers, and all related stakeholders to monitor changing customer preferences and derive meaningful decisions from complex, messy raw data.

ABBREVIATIONS

RG	**Resource Group**
ADF	**Azure Data Factory**
ADLS Gen2	**Azure Daat Lake Storage Gen2**
POWER BI	**Power Business Intelligence**
TD-IDF	**Term Frequency – Inverse Document Frequency**
SVM	**Support Vector Machine**
MB	**Megabyte**
RBAC	**Role Based Access Control**
SFTP	**SSH File Transfer Protocol**
NFS	**Network File System**
CLI	**Command-Line Interface**
SDK	**Software Development Kit**
REST API	**Representational State Transfer Application Programming Interfaces**
HTTPS	**Hypertext Transfer Protocol Secure**
HTTP	**Hypertext Transfer Protocol**
ETL	**Extract, Transform, Load**
ELT	**Extract, Load, Transform**
ML	**Machine Learning**
ACID	**Atomicity, Consistency, Isolation, Durability**
AI	**Artificial Intelligence**
UCSD	**University of California, San Diego**

TABLE OF CONTENTS

1. INTRODUCTION

1.1 MOTIVATION

In the fast-changing world of e-commerce and online shopping, customer generated data plays a vital role in shaping product perception and trustworthiness. Amongst the different available types of data, we are of the opinion that product reviews are certainly most influential in shaping the product perception, as millions of the customers rely on these reviews to make informed decisions about buying a specific product. A very popular e-commerce website used by millions of people across the globe, Amazon, has a massive corpus of reviews spanning over 2 decades and covering a vast range of product categories. These reviews, when used properly, has the potential to offer great insights about the consumer sentiment, changing preferences, and the market trends.

Through the medium of this project, we plan to use big data and cloud computing technologies to process and analyse the **Amazon Reviews Dataset 2023** (prepared and published by the McAuley Lab, University of California, San Diego). The dataset has over **571 million reviews** spanning across over 30 categories (and each category has hundreds of thousands of different products in it). As a result, this dataset offers both breadth and depth, making it a good choice for exploring the capabilities of big data tools and cloud services.

1.2 PROJECT OBJECTIVE

Our project builds a cloud-based big-data pipeline on Microsoft Azure that can ingest data, transform it and visualize insights from extremely large volumes of product reviews. To restrict the scope, we focus on two categories – All_Beauty and Electronics – and organize the workflow following the Medallion Architecture approach. The raw data files are in ADLS Gen2 (Bronze). Afterwards, the data undergoes transformation in Azure Databricks (Silver), and finally the analysis tables publish outputs (Gold). From these tables, we analyse how sentiment changes over time using star-rating buckets (1-2 stars negative, 3 stars neutral, and 4-5 positive) and additionally we do sentiment analysis on the data as well. Using dashboards, we compare how trends evolve and discover insights.

This project ensures raw, massively sized review data is not only collected but also transformed into clean and reliable insights. The focus on the two categories All_Beauty and Electronics categories provides perspectives on consumer sentiment, and the dashboards make the results accessible and interpretable. All in all, the objective is to create a reproducible and low-cost system that shows how Azure cloud services can be orchestrated to support big-data analytics.

2. PROBLEM SPECIFICATION

2.1 PROBLEM STATEMENT

Business organizations and its stakeholders collect product reviews at web scale, but the raw data is very noisy, heterogeneous, and extremely large for ad-hoc processing on any local computer. E-commerce websites, and different marketplaces, accumulate millions and billions of reviews. But in the raw, they are not much useful as they are inconsistent, and hard to compare across time. Absence of a dedicated data pipeline for solving such an issue, causes analysts to either overfit to small samples of data or spend access amounts of time to clean data instead of learning insights from it.

The problem we address through this project is to turn massive volumes of Amazon reviews data (raw records) into clean, analysis-ready, and decision-focused insights on student budget and tight timeline, using a cloud-native approach. The main challenges include reliable ingestion of raw data, scalable transformation of raw data into analysis-ready information and discovering different trends in a way that are reproducible and explainable. Hence, our approach relies on utilizing the services offered by the Microsoft Azure platform to make a dedicated pipeline to efficiently transform large amount historical data spanning over 2 decades into reliable insights.

2.2 PROJECT GOAL

The goal of this project is converting the large amount of raw review data into decision-ready data which highlights the customer preferences and sentiments, so businesses and online e-commerce organizations can utilize big data to take decisions accordingly in the best interest of the business. The project uses Azure-based data pipeline which follows the Medallion architecture.

The raw data lands in ADLS Gen2, afterwards the transformations take place in Databricks and Synapse and the cleaned data is published. This cleaned data is passed on to Power BI to make meaningful visual dashboards for the discovered hidden insights.

The goal is met when the pipeline reliably ingests and transforms the data end to end, such that the clean data supports making interactive dashboards.

2.3 LIMITATIONS TO EXISTING WORK

There is abundant literature in the fields of sentiment analysis, opinion mining, and big data analysis. Sentiment analysis classifies review text by attitude (positive, negative, neutral). Opinion mining, on the other hand, goes a step ahead by fetching what the opinion is about. Big data analysis refers to storing and processing very large amounts of datasets using distributed systems so that ingestion, transformation, and querying scale horizontally.

Despite this prior work, we do not find peer-reviewed studies that implement an end-to-end Azure based medallion architecture data pipeline for Amazon reviews. This gap further motivates us to take this project.

T. U. Haque et al., 2018 [1] builds text classifiers to polarize Amazon reviews into positive and negative classes. The authors combine the concept of active learning to label data with bag of words, TD-IDF, and evaluate various models like SVM, Naïve Bayes, etc. The work done is model-centric as it optimizes text classification accuracy on just a static sample. This work does not consider a cloud-based pipeline, which addresses sentiments across different time periods and publish visual dashboards.

M. Mishra et al., 2019 [2] conducts descriptive analysis on roughly 5GB of Amazon reviews using Hadoop and Hive. They then go on to make recommendation model in Azure ML Studio but on a sample subset (whose size is just roughly 111 MB). This approach is cloud-focused but the processing is limited to roughly just 100MB (which does not qualify for big data). Additionally, the architecture does not follow Medallion architecture and relative less information is available for insights as the authors sampled down the data to 111 MB.

3. BACKGROUND

3.1. BIG DATA

Big Data refers to datasets whose scale, speed, and complexity exceed the capabilities of traditional computers. In other words, big data arrives quickly, comes in many different formats, rapidly changes with time, and requires distributed computation and storage so that it can be transformed into reliable and decision-ready information.

A common characterization is the Five Vs of big data. The first V, volume denotes the massive amount of data that requires distributed storage and parallel processing. The second V, velocity captures the speed at which new records of data are generated and are ingested. The third V, variety refers to the different type of data available (heterogeneous data), like relational tables, semi-structured logs, and unstructured data. The fourth V, veracity emphasizes uneven data quality like duplicates, missing fields, etc., which needs cleaning and preprocessing. The fifth V, value is the final goal, by transforming the raw data into actionable metrics that inform decisions.

3.2 CLOUD & MICROSOFT AZURE

3.2.1 AZURE RESOURCE GROUP

Azure Resource Group is a logical folder that groups related Azure resources. This setup allows for easy deployment, updates, monitoring, tagging, security, and removal as a single unit. Governance, including RBAC (Role-Based Access Control) roles, policies, locks, and tags, is established at the Resource Group level and applies to everything inside it. Each resource belongs to only one Resource Group at a time, but we can move it if necessary. When we delete a Resource

Group, all resources within it are deleted as well. We can use locks to avoid mistakes. The Resource Group region holds its management metadata, while the resources can exist in other regions. However, keeping them in the same region is usually easier and more reliable.

3.2.2 AZURE BLOB STORAGE

Large, unstructured data, including text, photos, videos, logs, and backups, can be stored in Microsoft's Azure Blob Storage cloud object storage. Blobs are kept in storage accounts in containers. SFTP/NFS, CLI, PowerShell, SDKs, and REST APIs can all be used to safely access them via HTTPS. For landing and long-term storage, it is frequently used as an inexpensive and long-lasting solution. To help control expenses, it has built-in encryption, high availability, and tiering (Hot, Cool, Archive).

3.2.3 AZURE BLOB CONTAINER

The essential, folder-like area in a storage account that contains blobs files is called an Azure Blob container. It establishes boundaries for organization and access; hide anonymous access. Containers are called file systems in ADLS Gen2. Under it, directories can be created, such as /bronze, /silver, and /gold.

3.2.4 AZURE DATA LAKE STORAGE Gen2 (ADLS Gen2)

ADLS Gen2 is Azure Blob Storage enhanced with a hierarchical namespace like directories or sub directories, so big data engines read, write and manage data efficiently, it integrates natively with Synapse and Databricks and uses Role-Based Access Control (RBAC) and Access Control Lists (ACLs) plus encryption for security. Unlike standard Azure Blob Storage (general-purpose object storage), ADLS Gen2 adds file-system semantics and POSIX-style ACLs, making it better for analytics and Lakehouse workloads. For durability, we can select a redundancy level: Locally Redundant Storage (LRS) keeps three copies in one datacenter (lowest cost), Zone-Redundant Storage (ZRS) keeps three copies across availability zones in one region (high availability), Geo-Redundant Storage (GRS) also copies data to a paired secondary region (disaster recovery), and Geo-Zone-Redundant Storage (GZRS) combines ZRS in the primary region with copies in a secondary region (the strongest protection).

3.2.5 AZURE DATA FACTORY

Azure Data Factory is a fully managed, serverless data-integration service to build and run pipelines that ingest, transform, and orchestrate data on schedules or events. It provides 90+ built-in connectors for on-premises, Azure, and multicloud sources; supports code-free ETL/ELT with Mapping Data Flows or code-centric steps (e.g., calling Databricks/SQL); and includes triggers, dependencies, monitoring, and alerts for reliable operations—all on a pay-as-you-go model.

3.2.6 AZURE DATABRICKS

Azure Databricks is Azure’s managed Apache Spark platform for data engineering, analytics, and machine learning. It provides a collaborative notebook workspace and managed, autoscaling compute to run ETL and ML quickly. It integrates tightly with ADLS Gen2, Synapse, Power BI, and Azure ML, and uses Delta Lake for reliable (ACID) tables and time travel. Security is handled through Microsoft Entra ID (managed identities) with secrets in Key Vault, and governance via Unity Catalog. Jobs can be scheduled, clusters can auto-terminate to control cost, and everything runs within our Azure subscription.

3.2.7 AZURE SYNAPSE ANALYTICS

Azure Synapse is Microsoft's cloud analytics platform that unifies SQL (serverless and dedicated), Apache Spark, and Data Explorer in a single workspace (Synapse Studio) to ingest, explore, transform, and secure data end to end. It's "lake-first": We can query files in ADLS Gen2 directly or build governed tables and pipelines (built on the same engine as Azure Data Factory) and connect easily to Power BI for reporting. This lets teams choose the right engine per task while keeping operations, security, and monitoring in one place.

3.2.8 POWER BI

Power BI is Microsoft's business intelligence platform for connecting to data, modeling it, and creating interactive reports and dashboards that can be shared securely. It integrates with Microsoft 365 (e.g., Teams, PowerPoint, Excel), supports AI-assisted analysis, scales from self-service to enterprise use and can be embedded in custom apps helping organizations turn data into actionable insights.

4. CONNECTION TO CS-714 COURSE

Our project uses big data as it works with large amounts of heterogeneous Amazon review records that mix structured fields with text fields. Instead of relying on a single computing resource, we employ the services offered by a cloud service provider, Microsoft Azure.

Amazon product reviews dataset exhibits all five Vs, high volume across different years, velocity as new arrive with every passing year, variety across text, ratings and other fields, veracity in the user-generated content, and the potential value for understanding the insights which the data has to offer.

The designed pipeline transforms the raw data, manages data in a distributed file system, and gives analysis ready output that supports making dashboards with insights. We apply resource provisioning, access control and quota management from the cloud foundations. From the big data analytics, we handle schema, portioning of data, and suitable file formats for parallel reads.

This project is relevant to CS714 as it translates the course's ideas into an end-to-end cloud-based data pipeline capable of turning messy, high-volume data into reliable insights.

5. APPROACH

5.1 DATASET

The project uses the Amazon Reviews Dataset 2023 [4] released by the McAuley of the University of California, San Diego. The most recent release of the dataset was in 2023. This release provides large-scale user reviews and item metadata across 33 different categories. The data spans across over 2 decades, ranging from May 1996 to September 2023.

Data from two categories is taken into consideration: All_Beauty and Electronics. These categories were selected because they fitted the definition of big data. Official per-category statistics from the dataset website are summarized in the following table:

Category	#Users	#Items	#Reviews	#R_Token	#M_Token
All_Beauty	632.0K	112.6K	701.5K	31.6M	74.1M
Electronics	18.3M	1.6M	43.9M	2.7B	1.7B

Table 5.1 Per-category statistics table.

#Users – Total count of unique reviewers.

#Items – Total count of unique products.

#Reviews – Total number of user-item interactions with a star rating.

#R_Token – Total count of review-text tokens across the user review data

#M_Token – Total count of metadata tokens across items

5.2 ARCHITECTURE

5.2.1 ARCHITECTURE (ALL_BEAUTY CATEGORY)

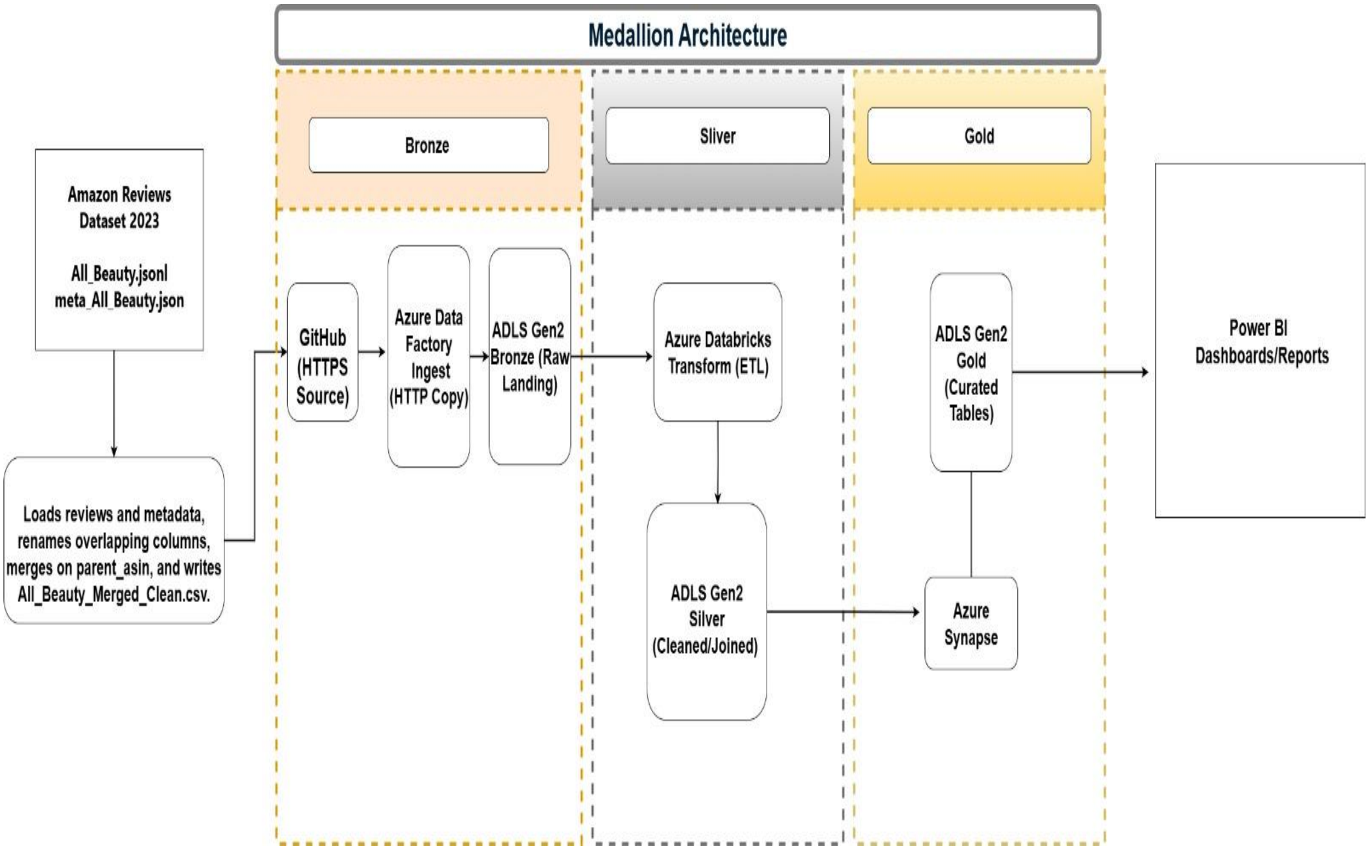

Medallion Architecture
Bronze
Sliver
Gold
Amazon Reviews Dataset 2023
All_Beauty.jsonl
meta_All_Beauty.json
Loads reviews and metadata, renames overlapping columns, merges on parent_asin, and writes All_Beauty_Merged_Clean.csv.
GitHub (HTTPS Source)
Azure Data Factory Ingest (HTTP Copy)
ADLS Gen2 Bronze (Raw Landing)
Azure Databricks Transform (ETL)
ADLS Gen2 Silver (Cleaned/Joined)
ADLS Gen2 Gold (Curated Tables)
Azure Synapse
Power BI Dashboards/Reports

Fig. 5.2.1 All_Beauty Data Pipeline Architecture

End to End Data Pipeline (Medallion Approach)

This solution sets up a repeatable data pipeline on Azure for the Amazon Reviews – All Beauty dataset. It uses the Medallion pattern: bring data in reliably, clean and standardize it, store it efficiently, and serve it to Power BI for reporting.

Before ingestion, a one-time local step merges reviews (*All_Beauty.jsonl*) and metadata (*meta_All_Beauty.jsonl*), fixes duplicate column names where both files use the same names for different data by renaming the metadata fields (for example, *title* → *product_title, images* →

product_images),and then joins on *parent_asin* as the common field and publishes *All_Beauty_Merged.csv* to GitHub. In the Bronze layer, this raw merged CSV is copied from GitHub over HTTPS into Azure Data Lake Storage Gen2 without changes using Azure Data Factory. In the Silver layer, Azure Databricks with PySpark cleans and standardizes the data, columns not needed for analysis, like images, *product_images*, *product_videos, product_categories* (kept main_category instead), and *bought_together*, are removed to reduce size and improve performance because they are not used in the Gold metrics and *bought_together* is 100 percent null, numeric fields such *as helpful_vote, rating_number*, and *product_price* are correctly typed, *verified_purchase* is cleaned so it is always True or False and also 1 or 0, which makes the verified-purchase percentage accurate, a real timestamp *event_ts* is derived from UNIX milliseconds, *review_year* is extracted, and the cleaned data is written as Parquet, partitioned by *review_year*. In the Gold layer, Azure Synapse Serverless SQL reads Silver with OPENROWSET and writes curated Parquet datasets using CETAS, which also creates external tables, the Gold layer contains *year_summary, brand_year, category_year, month_trend, price_rating, product_year,* and *top_reviews*, and includes the business metric *pct_verified,* which is the share of verified purchases in the aggregated tables. Together, these curated tables provide a consistent, high-performance foundation for reliable reporting in Power BI and other analytics tools.

Data is stored in the account *amazondatastorageaccount* (hierarchical namespace enabled) within the container *amazon2023data*. Bronze files reside under bronze/All_Beauty, Silver Parquet (partitioned by *review_year*=YYYY) under silver, and Gold datasets under gold with clear subfolders matching each curated table (*year_summary, brand_year, category_year, month_trend, price_rating, product_year, top_reviews*). This consistent structure makes discovery and lifecycle management straightforward.

All services use Azure-native identities and roles. Data Factory ingests over HTTPS and authenticates to storage using an access key or, preferably, Managed Identity. Databricks connects to ADLS Gen2 over ABFSS using a Service Principal, its client secret is stored securely in Azure Key Vault or a Databricks secret scope. Synapse Serverless uses the workspace Managed Identity with the Storage Blob Data Contributor role on the storage account, so no keys are embedded in SQL. Network access can be restricted with the storage firewall and Private Endpoints when needed.

5.2.2 ARCHITECTURE (ELECTRONICS CATEGORY)

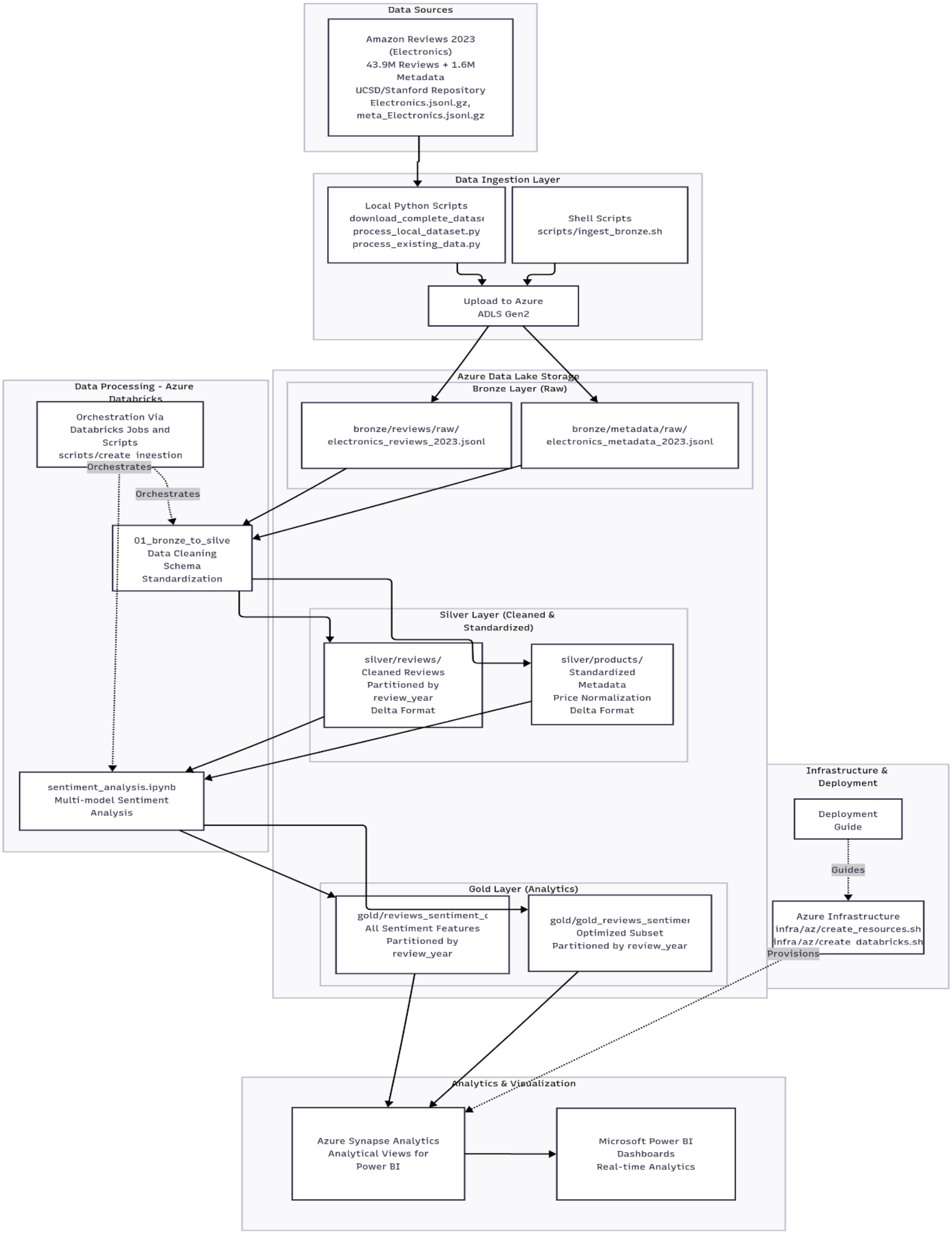

Data Sources
Amazon Reviews 2023 (Electronics) 43.9M Reviews + 1.6M Metadata UCSD/Stanford Repository Electronics.jsonl.gz, meta_Electronics.jsonl.gz
Data Ingestion Layer
Local Python Scripts download_complete_datase process_local_dataset.py process_existing_data.py
Shell Scripts scripts/ingest_bronze.sh
Upload to Azure ADLS Gen2
Data Processing - Azure Databricks
Orchestration Via Databricks Jobs and Scripts scripts/create_ingestion
Orchestrates
Orchestrates
01_bronze_to_silve Data Cleaning Schema Standardization
sentiment_analysis.ipynb Multi-model Sentiment Analysis
Azure Data Lake Storage
Bronze Layer (Raw)
bronze/reviews/raw/ electronics_reviews_2023.jsonl
bronze/metadata/raw/ electronics_metadata_2023.jsonl
Silver Layer (Cleaned & Standardized)
silver/reviews/ Cleaned Reviews Partitioned by review_year Delta Format
silver/products/ Standardized Metadata Price Normalization Delta Format
Gold Layer (Analytics)
gold/reviews_sentiment_c All Sentiment Features Partitioned by review_year
gold/gold_reviews_sentimer Optimized Subset Partitioned by review_year
Infrastructure & Deployment
Deployment Guide
Guides
Azure Infrastructure infra/az/create_resources.sh infra/az/create_databricks.sh
Provisions
Analytics & Visualization
Azure Synapse Analytics Analytical Views for Power BI
Microsoft Power BI Dashboards Real-time Analytics

Fig. 5.2.2 Electronics Data Pipeline Architecture

The solution begins with very large and mostly unstructured datasets: Amazon's 2023 review collection for the electronics category. This source holds millions of customer comments along with supporting product details such as brand, price, and identifiers. Since the raw files are distributed in JSON format, the first task is to fetch and stage them. Local Python scripts handle downloading and light preprocessing, while a simple shell routine moves the material into Azure Data Lake Storage Gen2. At this stage, often referred to as the *bronze zone*, the information remains in its original condition, so nothing is lost.

Once in the lake, the material is not yet usable for reporting. To make it consistent, Databricks jobs run transformation notebooks that check schema alignment, rename or standardize fields, and apply price normalization. Reviews are also partitioned by year, which makes historical analysis much faster later on. This produces what is called the *silver layer*—data that is cleaned, organized, and ready for enrichment.

An important step follows, sentiment enrichment. Using a dedicated notebook, reviews are passed through several natural language processing models—such as VADER, TextBlob, and a rule-based scorer. Each of these provides a slightly different perspective on the tone of a review, and their outputs are blended into an ensemble score. In addition to polarity, other linguistic signals like volatility, readability, and emotional strength are extracted. The result is a richer representation of customer opinion than ratings alone can provide.

The curated outcomes are stored in the *gold layer*. Here the reviews carry both standardized structure and advanced features, and the metadata has been aligned with them. The gold zone is intentionally optimized for analytics: subsets are partitioned by year, and lightweight tables are created to support common queries without scanning the entire dataset. This makes it possible to calculate key performance indicators quickly—whether that be the share of positive reviews for a brand, shifts in sentiment during holiday seasons, or the level of alignment between written comments and star ratings.

To make these insights consumable, the gold outputs are published into Azure Synapse Analytics, where SQL-based analytical views are defined. These views serve as the bridge to visualization, exposing the metrics in a form that Power BI can easily consume. Business dashboards then combine multiple perspectives: product health, brand reputation, verified purchase rates, review quality, seasonal or pandemic-related shifts, and more. Executives and analysts can drill into detail, compare periods, or segment by price tier—all without interacting with the raw files.

In short, the architecture steadily refines noisy raw text into structured intelligence. It begins with simple storage, moves through cleaning and sentiment analysis, and ends with interactive dashboards. The progression from bronze to silver to gold ensures that each stage adds value: preserving fidelity, imposing order, enriching with advanced features, and finally transforming it into strategic knowledge for decision makers.

6. RESULTS

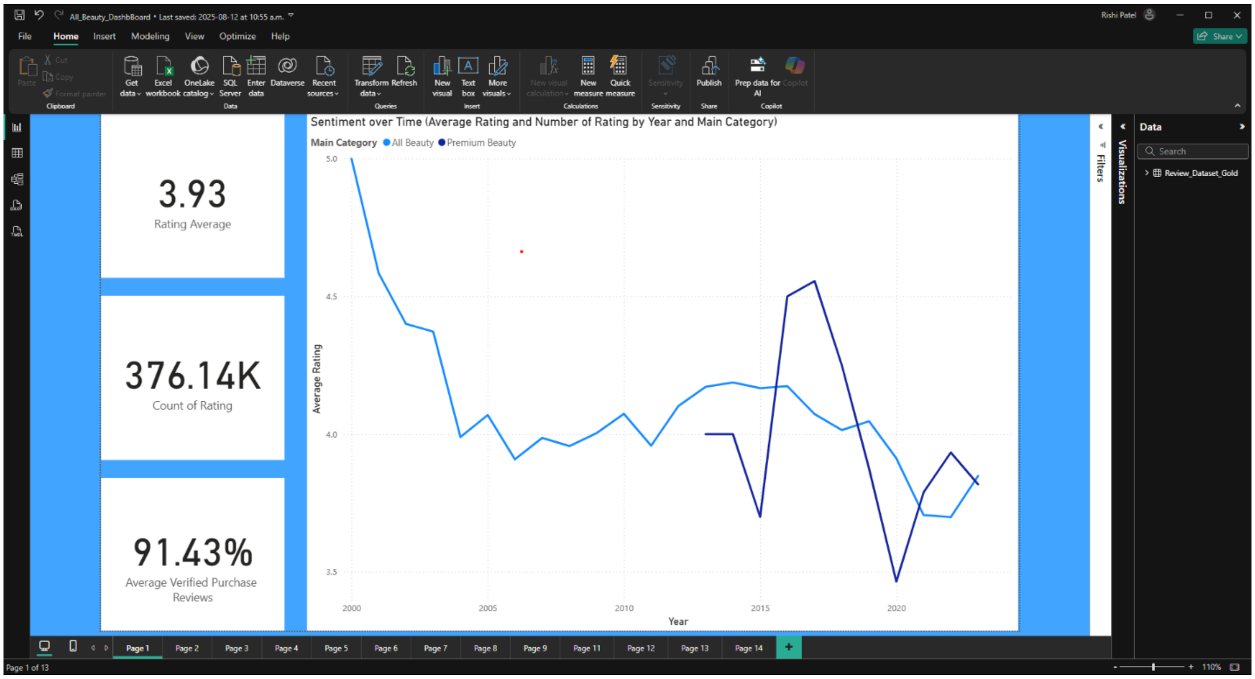


Fig. 6.1 Sentiment over Time (All_Beauty Category)

In Fig 6.1, the line chart shows the average rating trend from 2000-2023 for both *All Beauty* and *Premium Beauty* category.

Ratings in early 2000s were close to 5 but declined steadily to around 3.8 in last few recent years.

Premium Beauty category follows a similar trend with sharp fluctuations around the period 2016-2019.

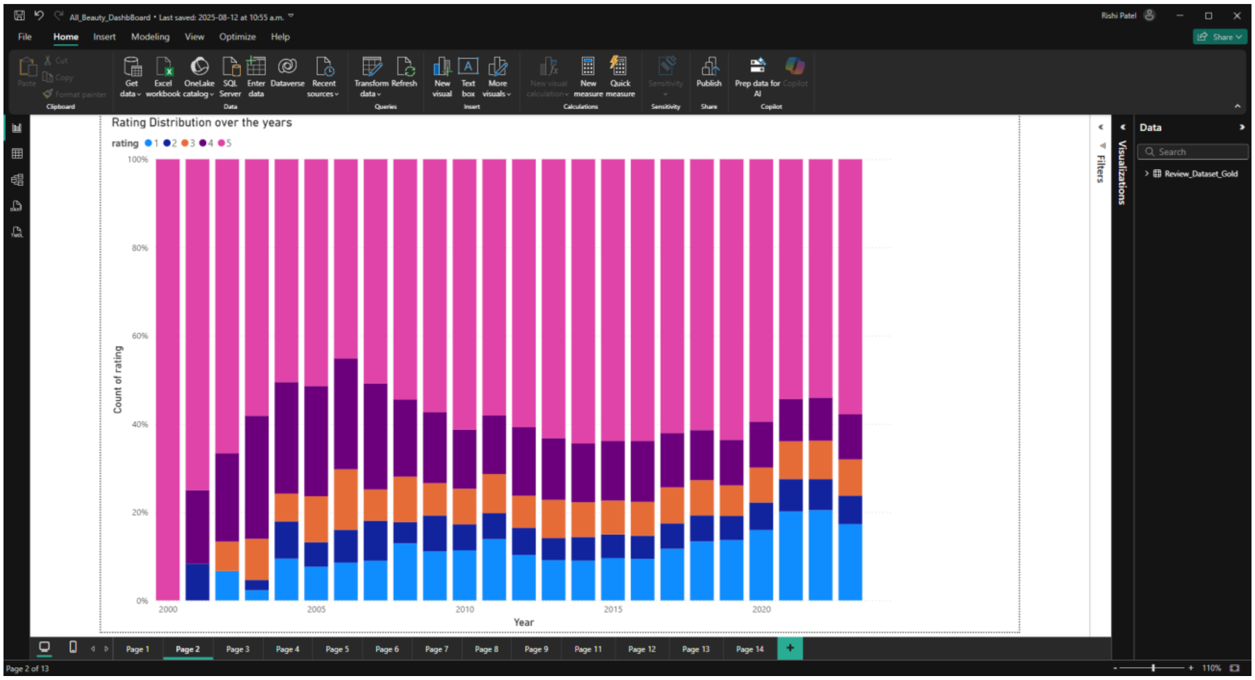


Fig. 6.2 Rating Distribution over the years (All_Beauty Category)

In Fig. 6.2 stack bar chart distributions of ratings ranging from 1-5 stars.

We observe that there is gradual rise lower star rating (1-3 stars) after 2005. This indicates that there is more mixed customer feedback.

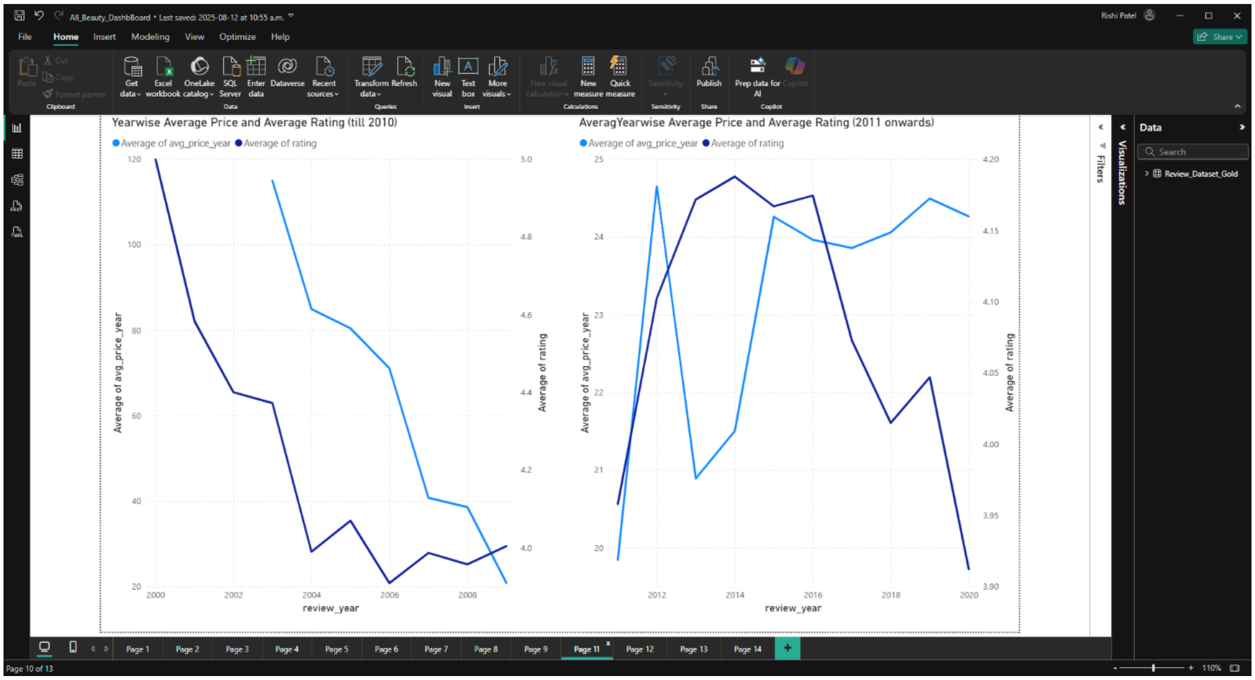
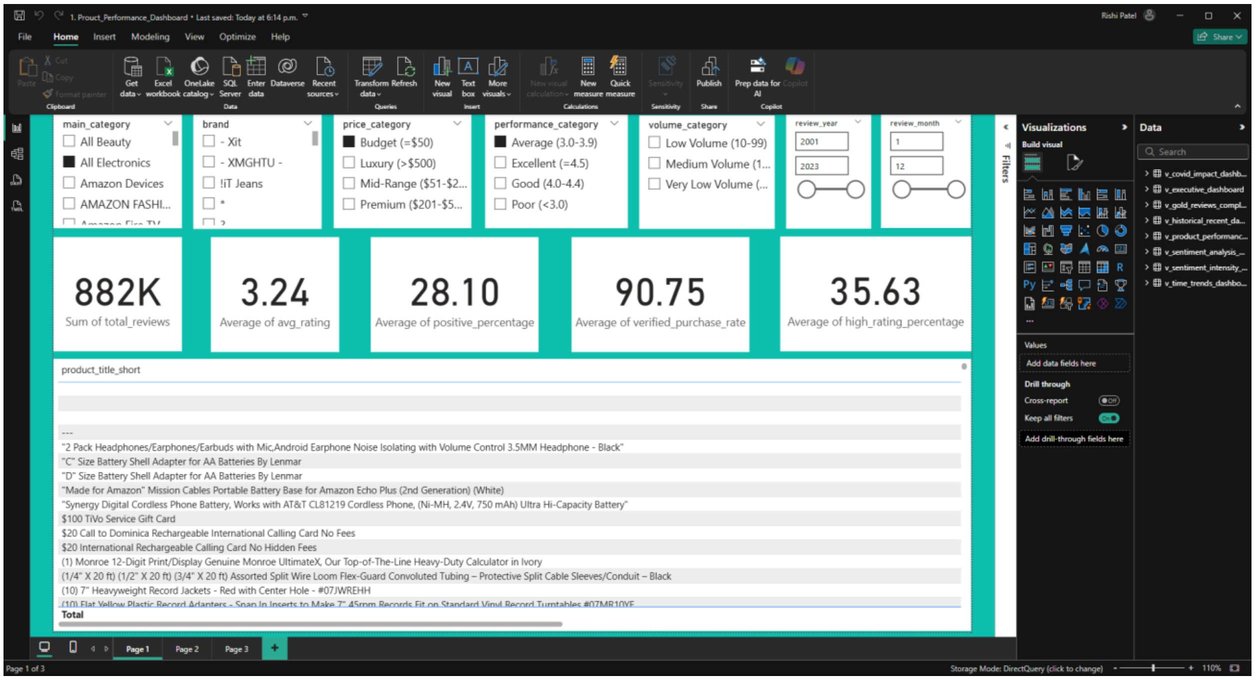


Fig. 6.3 Year-wise average price vs average rating till 2010 and 2011 onwards (All_Beauty Category)

In Fig. 6.3, we see that for the period of 2000-2010 prices dropped drastically from $100 to roughly $20-25 and so did the ratings (fell from 5 to approximately 4 stars).

We also observe that for the period 2011-2020, the prices were stabilized at roughly $25, and the rating fluctuations were also severe.



1. Prouct_Performance_Dashboard • Last saved: Today at 6:14 p.m.

main_category
☐ All Beauty
■ All Electronics
☐ Amazon Devices
☐ AMAZON FASHI...

brand
☐ - Xit
☐ - XMGHTU -
☐ !iT Jeans
☐ *

price_category
■ Budget (=$50)
☐ Luxury (>$500)
☐ Mid-Range ($51-$2...
☐ Premium ($201-$5...

performance_category
■ Average (3.0-3.9)
☐ Excellent (=4.5)
☐ Good (4.0-4.4)
☐ Poor (<3.0)

volume_category
☐ Low Volume (10-99)
☐ Medium Volume (1...
☐ Very Low Volume (...

review_year: 2001 – 2023

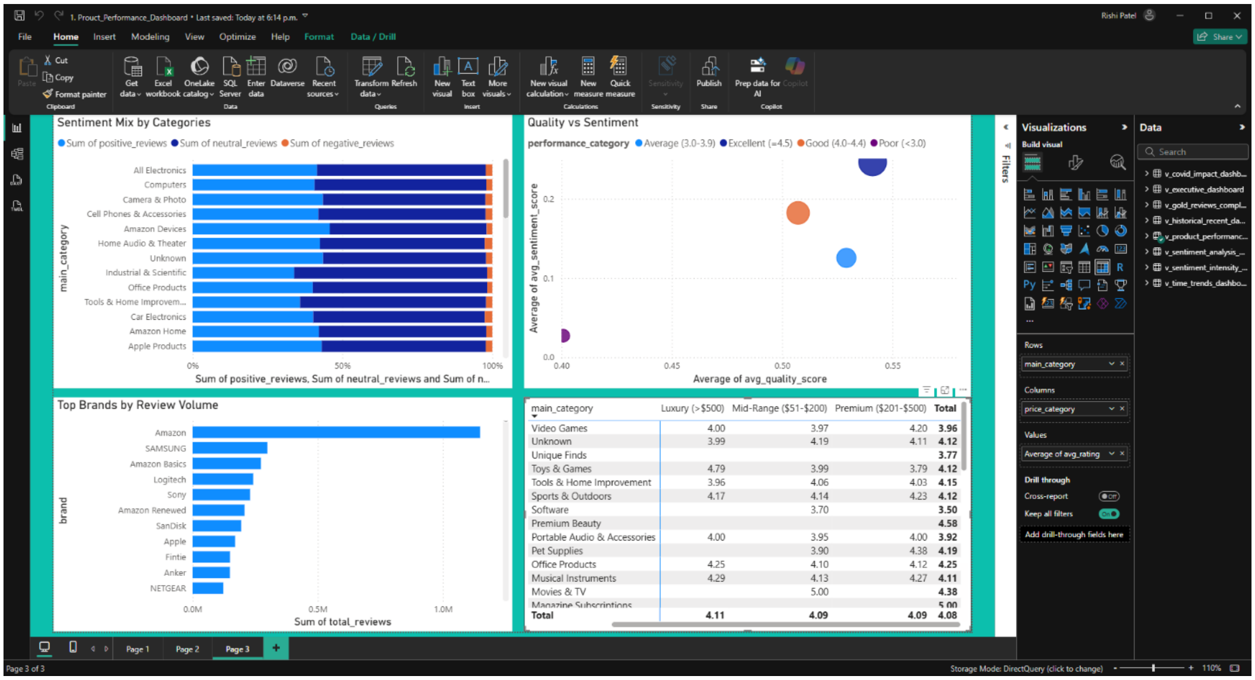
review_month: 1 – 12

882K	3.24	28.10	90.75	35.63
Sum of total_reviews	Average of avg_rating	Average of positive_percentage	Average of verified_purchase_rate	Average of high_rating_percentage

product_title_short

"2 Pack Headphones/Earphones/Earbuds with Mic,Android Earphone Noise Isolating with Volume Control 3.5MM Headphone - Black"
"C" Size Battery Shell Adapter for AA Batteries By Lenmar
"D" Size Battery Shell Adapter for AA Batteries By Lenmar
"Made for Amazon" Mission Cables Portable Battery Base for Amazon Echo Plus (2nd Generation) (White)
"Synergy Digital Cordless Phone Battery, Works with AT&T CL81219 Cordless Phone, (Ni-MH, 2.4V, 750 mAh) Ultra Hi-Capacity Battery"
$100 TiVo Service Gift Card
$20 Call to Dominica Rechargeable International Calling Card No Fees
$20 International Rechargeable Calling Card No Hidden Fees
(1) Monroe 12-Digit Print/Display Genuine Monroe UltimateX, Our Top-of-The-Line Heavy-Duty Calculator in Ivory
(1/4" X 20 ft) (1/2" X 20 ft) (3/4" X 20 ft) Assorted Split Wire Loom Flex-Guard Convoluted Tubing – Protective Split Cable Sleeves/Conduit – Black
(10) 7" Heavyweight Record Jackets - Red with Center Hole - #07JWREHH
(10) Flat Yellow Plastic Record Adapters - Snap In Inserts to Make 7" 45rpm Records Fit on Standard Vinyl Record Turntables #07MR10YE
Total

Fig. 6.4 Product Performance KPIs (Electronics Category)



1. Prouct_Performance_Dashboard • Last saved: Today at 6:14 p.m.

main_category	Luxury (>$500)	Mid-Range ($51-$200)	Premium ($201-$500)	Total
Video Games	4.00	3.97	4.20	**3.96**
Unknown	3.99	4.19	4.11	**4.12**
Unique Finds				**3.77**
Toys & Games	4.79	3.99	3.79	**4.12**
Tools & Home Improvement	3.96	4.06	4.03	**4.15**
Sports & Outdoors	4.17	4.14	4.23	**4.12**
Software		3.70		**3.50**
Premium Beauty				**4.58**
Portable Audio & Accessories	4.00	3.95	4.00	**3.92**
Pet Supplies		3.90	4.38	**4.19**
Office Products	4.25	4.10	4.12	**4.25**
Musical Instruments	4.29	4.13	4.27	**4.11**
Movies & TV		5.00		**4.38**
Magazine Subscriptions				**5.00**
Total	**4.11**	**4.09**	**4.09**	**4.08**

Fig. 6.5 Sentiment-based analysis (Electronics Category)

From fig 6.5, we can see sentiment (+ve, -ve, neutral) for different categories. We also see top brands for the selected category, and quality versus sentiment comparison.

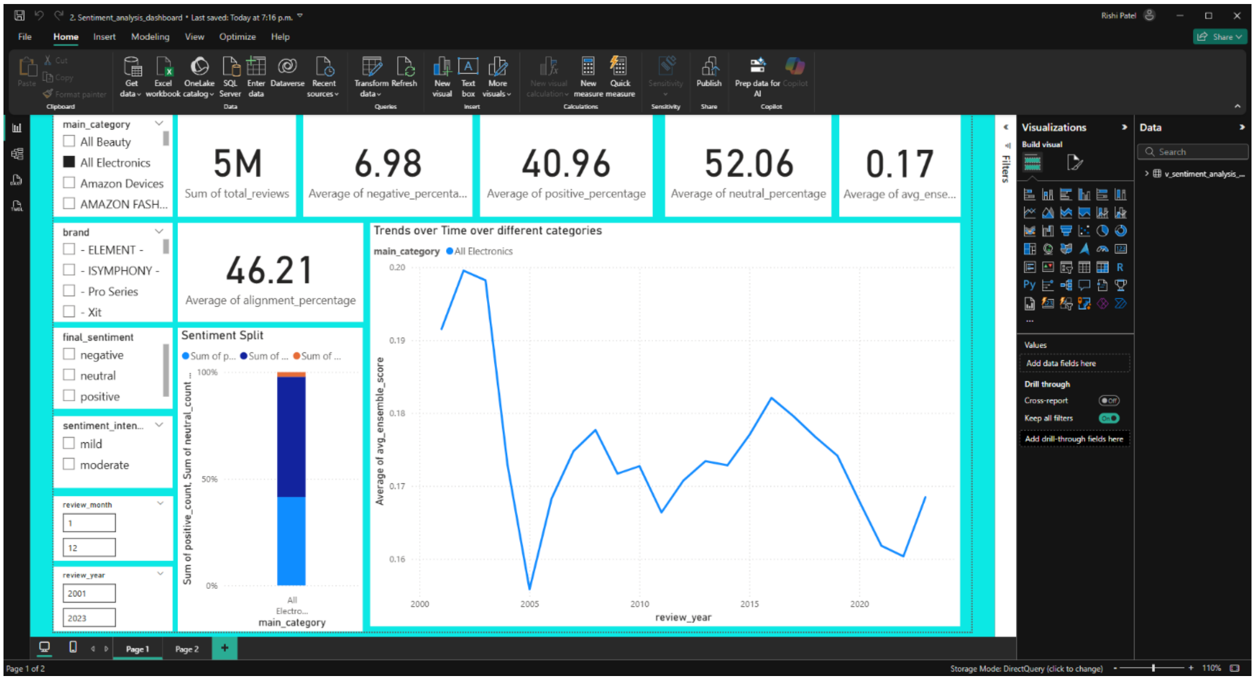


Fig. 6.6 Sentiment Analysis KPIs

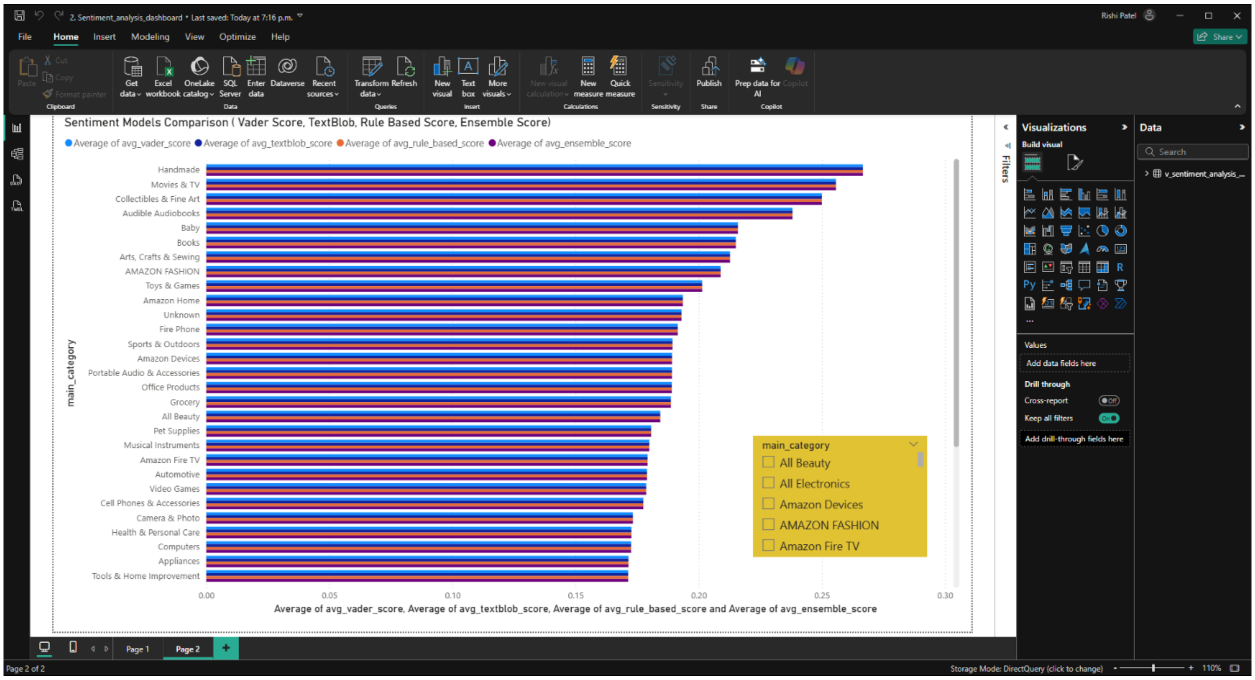


Fig. 6.7 Comparison of different Sentiment Models

In fig. 6.7, we see the clustered bar chart comparing the different sentimental models like Vader, TextBlob, and Ensemble Scores across multiple categories. Handmade category shows highest average scores while categories like Appliances have lower scores.

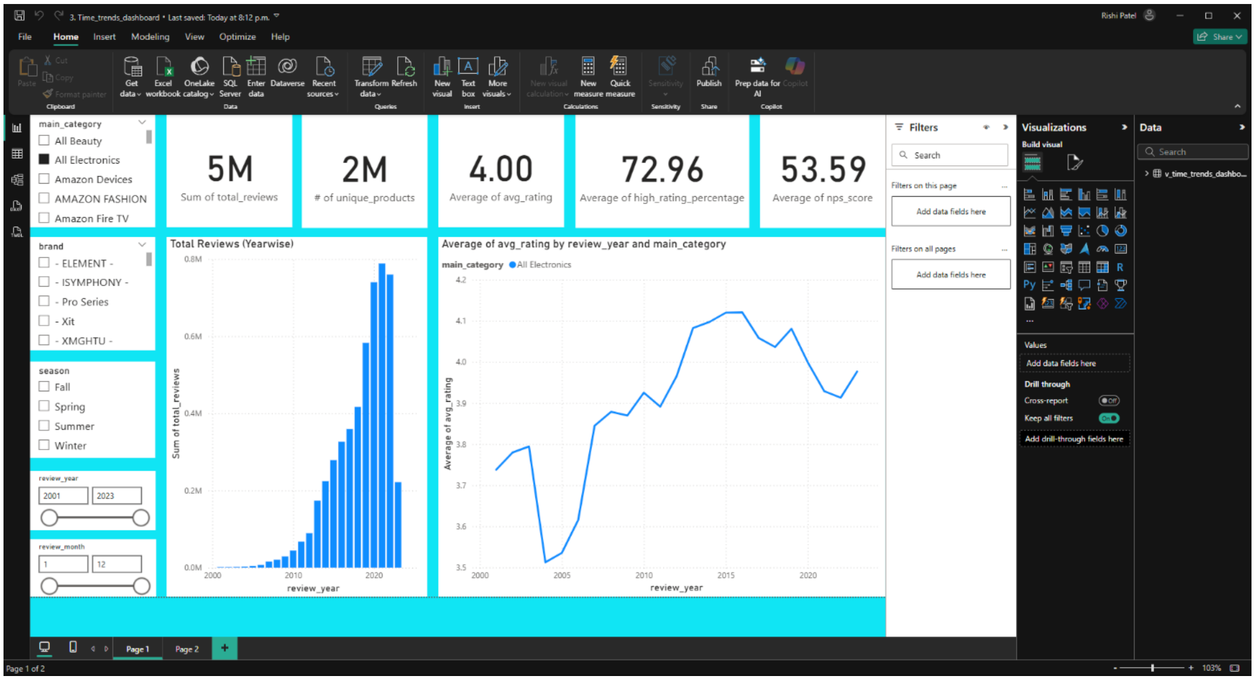


Fig. 6.8 Time Trends KPIs

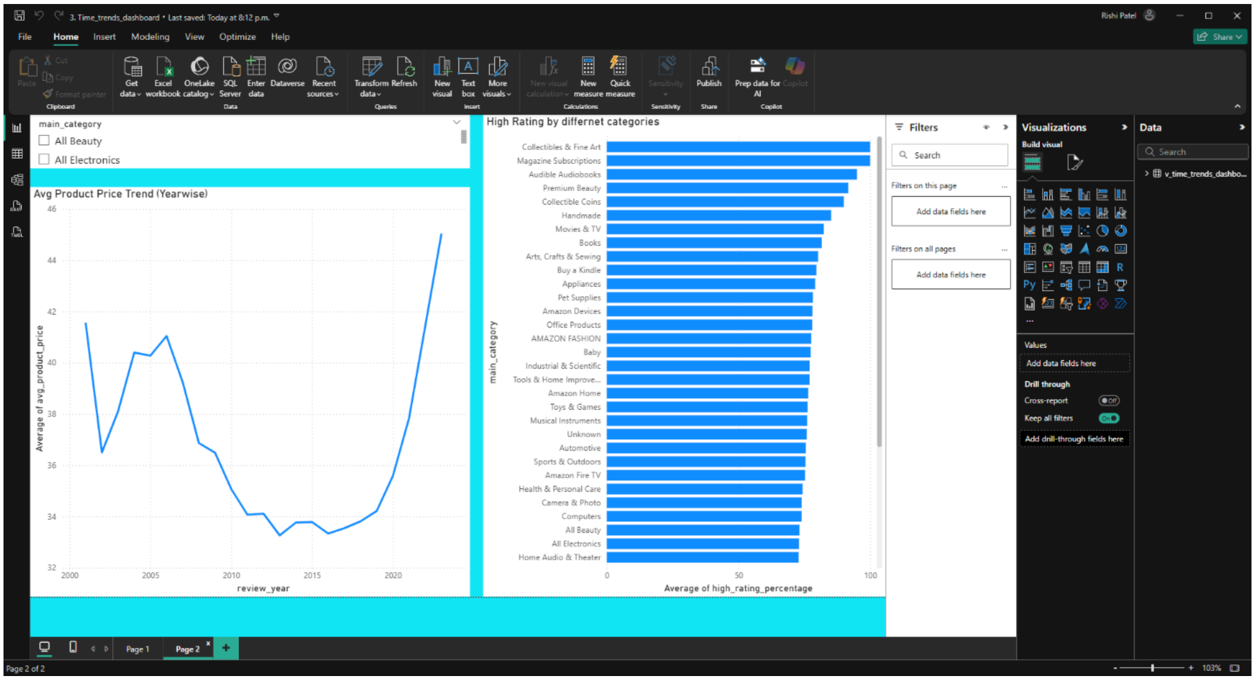


Fig. 6.9 Average Product Price and High Rating by different categories

From the fig 6.9, we observe that the average prices across the categories were close to $40 in early 2000s which dipped to mid 30s in 2010s. Additionally, we see different categories with high ratings like collection & fine art, and magazine subscriptions.

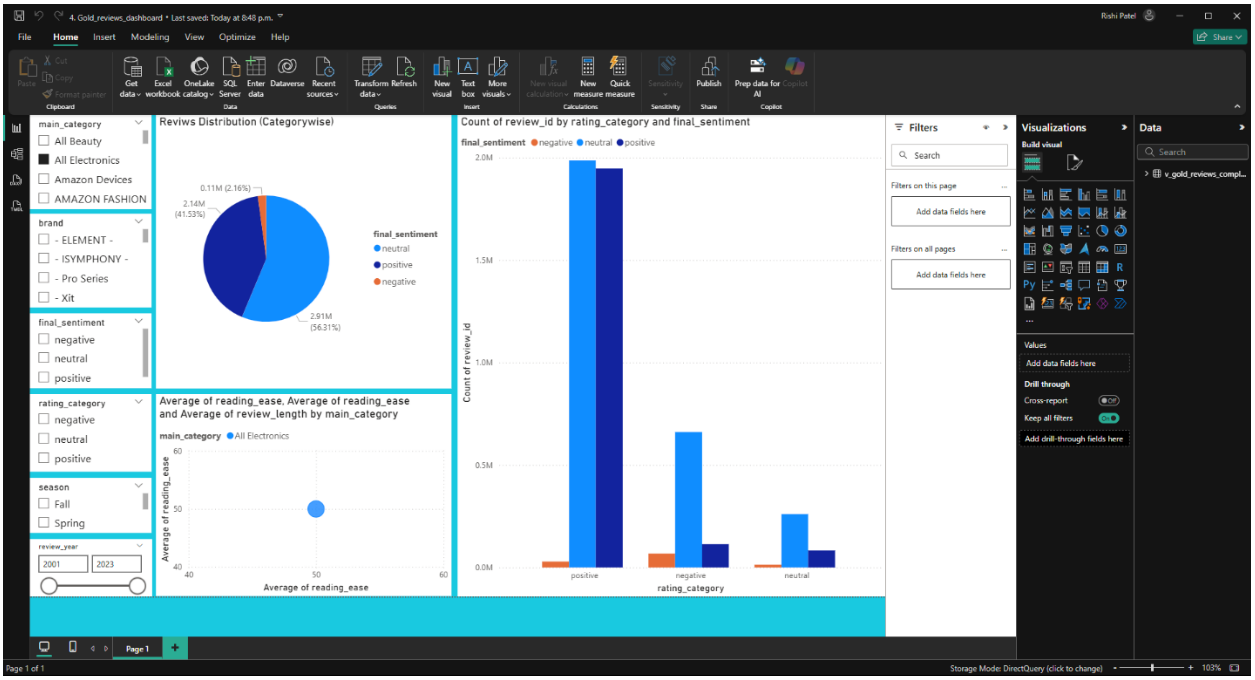


Fig. 6.10 Category wise KPIs

In fig. 6.10, we KPIs that can be accessed for different categories, as well as count of review_id and the respective sentiment.

From the above figures 6.1 to 6.10, we see that the dashboards made for both, *All_Beauty* and *Electronics* categories provide comprehensive results across multiple dimensions, including sentimental trends, rating distributions, pricing effects, and categoty specific key performance indicators.

These results highlight patterns in customer behaviour and their changing preference across different categories, uncovering insights that set the foundation for deeper and better analysis in this field.

7. ANALYSIS

The project shows how a cloud-based big data pipeline can transform large, humongous volume of Amazon reviews data into meaningful actionable insights. The dashboards for both, *All_Beauty* and *Electronics*, revealed broad trends and many category-specific patterns in product rating, and brand performance.

Across both the categories, there is a gradual drop in average ratings over time. Early years showed higher positive feedback, while in the recent years there has been a balance, and customers have been critical of products they review.

Upon reviewing the rating distribution, we notice that although 5-star dominate overall, the lower ratings has steadily risen over the years, highlighting the customers critique nature. It also shows that the customers are comparatively more willing to report dissatisfaction. Additionally, sentimental analysis reveals the presence of large of neutral reviews. The relative low alignment between the star ratings and written sentiments show that depending solely on ratings may overlook the different customer perspectives.

Overall, the project analysis demonstrates that combining ratings, sentiments, pricing, and different dimensions, we get a holistic view of customer behavior. Additionally, the project shows the capability of the Azure-based end to end data pipeline to deliver insights across millions of records.

8. CONCLUSION

This project demonstrated how a cloud-native big data pipeline transforms massive, large-scale, noisy and heterogeneous datasets into meaningful, decision-ready insights. Using the Amazon Reviews Dataset 2023, we have focused on two categories, *All_Beauty* and *Electronics*, to show the scalability of our approach.

We implemented the Medallion architecture in Microsoft Azure, moving data through bronze, silver, and gold tiers. This is how raw JSON data was ingested in ADLS Gen2, cleaned and transformed in Azure Databricks, curated into table format using Azure Synapse Analytics, and finally visualized with the help of Power BI. This ensured that the data was reliable, standardized, and completely optimized for analytical queries, all while making sure that the solution stays cost-effective.

The dashboards as shown in Figs. 6.1 to 6.10, highlights various valuable insights into customer sentiment, product ratings, pricing behaviour, and changing preference over the period.

Through the medium of this project, we successfully obtained built and deployed a scalable, cost-efficient, and reproducible Azure-based end-to-end big data pipeline for converting raw data into actionable insights.

9. LIMITATION TO OUR WORK

This project is completed as a part of the academic setting, and as such, has few practical constraints which we must consider. The first limitation is that of data availability and timeliness. The study uses historical data spanning over 2 decades rather than live or continuously updated sources. As a result, seasonal spikes and abrupt shifts in customer behaviour outside the recorded period are not reflected in the findings.

Another constraint is access to real-time data. Because of legal IT considerations, we do not perform web scrapping or automated crawling of Amazon reviews. As some of the review content is not publicly owned, automated collection may breach usage terms. So, we have any such activity and limited the project to datasets made available for academic use by the UCSD's McAuley Lab.

Other obvious constraints for the implementation of this project include limited time frame to work on the project, extremely tight student budget leading to platform and configuration constraints.

However, these constraints do not diminish the contribution of our, rather they frame it. As resources, time, and permissions expand, the same design would extend to broader category coverage, turning this academic project into production-grade pipeline fully capable of deriving insights from the raw data.

10. FUTURE WORK

Going forward, we plan on extending the project in three areas: integrating fresher, newer data, deeper and better analytics, and broader coverage across other categories, all while trying to preserve the cost-effectiveness of medallion-based data pipeline.

Firstly, we can integrate newer data on a periodical basis (like weekly, monthly, etc.). Instead of reloading the entire data, the pipeline would track newer ingestions. However, we should have access to real-time to the data to make this feasible.

Additionally, we can integrate the services offered by Azure Cognitive Services – Text Analytics to move forward from traditional sentimental analysis. However, here the trade-off would be budget as the budget tend to increase with increase in the number of records.

Future work would also broaden the category coverage. This would require revisiting the partitioning strategy and making sure that the dashboards remain responsive at higher volumes.

In summary, future work strengthens freshness, depth, and robustness of the existing work: periodic IT laws complaint data integration, better sentiment and opinion mining, and measured scale-out to include more categories, all built with clear costs and repeatable steps to support decisions over time.

11. REFERENCES

[1] Tanjim Ul Haque., et al., *“Sentimental Analysis on Large Scale Amazon Product Reviews”* 2018

[2] Monika Mishra, et al., *“Big Data Predictive Analysis of Amazon Product Review”* 2019

[3] Blend Berisha, et al., *“Big Data Analytics in Cloud Computing: an overview”* 2022

[4] Amazon Reviews 2023 Dataset *“ https://amazon-reviews-2023.github.io/”*