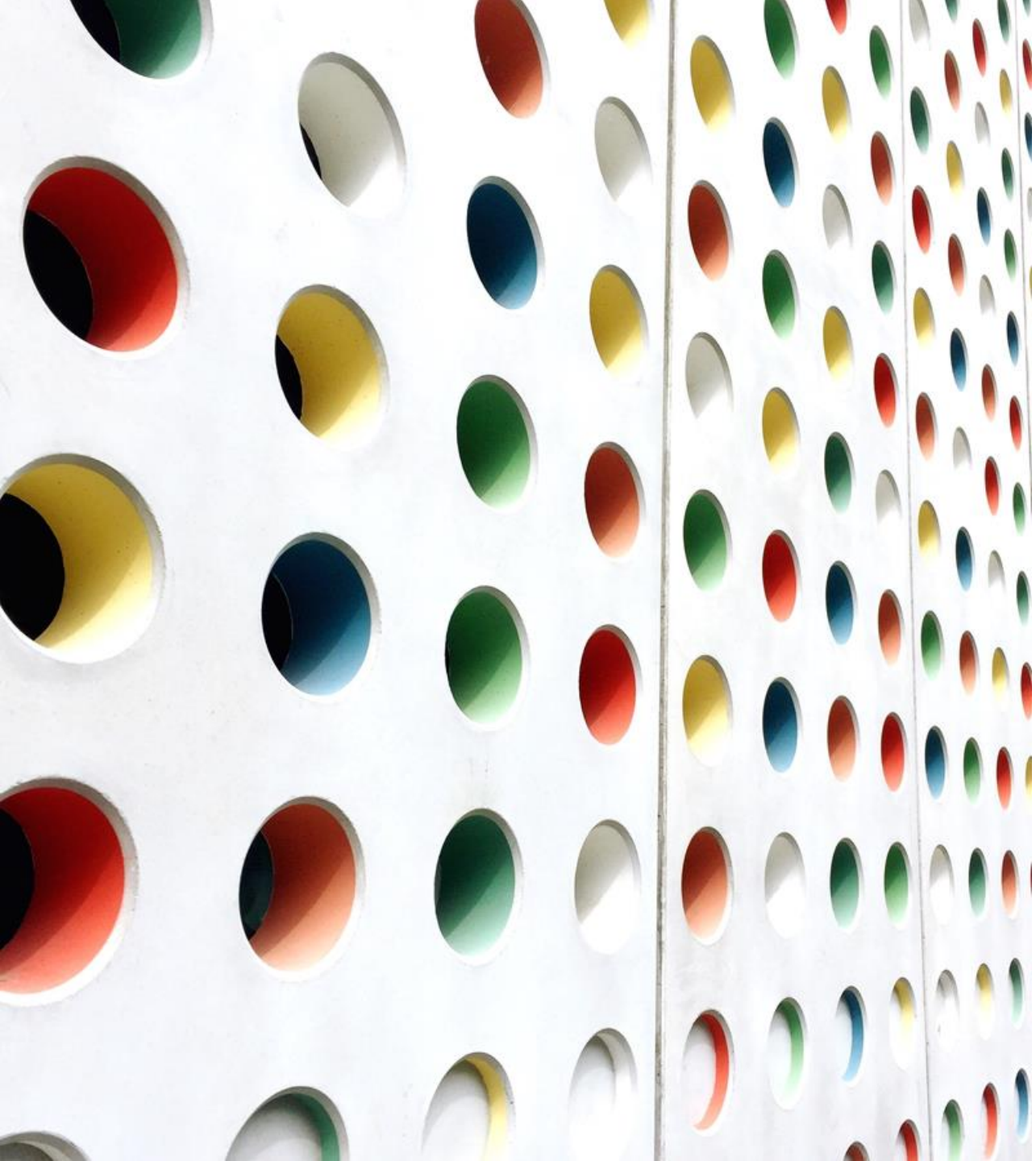


LEVERAGING DATA ANALYTICS FOR SALES OPTIMIZATION AND REVENUE GROWTH

AMAZON SALES DATASET

Prepared by:
Rabia Danish

Problem Description

- Businesses like Amazon need to identify key factors influencing product sales in the competitive online marketplace.
- Attributes such as price, rating, product category, and discounts play a crucial role in sales performance.
- The challenge lies in leveraging these insights to make accurate decisions that boost sales and customer satisfaction.
- Objective: Identify product attributes that impact sales success and optimize discounts to maximize sales volume and revenue.

Data source with information

Source: Kaggle (Amazon Sales Dataset).
File Format: CSV.
Key Attributes: Product ID, Category, Actual_price, Discounted_price, discount_percentage, Rating, and Rating Count.
Derived Attributes: Revenue, Sales_Volume, NormalizedSalesVolume, NormalizedRevenue
Dataset Size: 1465 products, ~ 4.53 MB file.

Work Distribution

Rabia Danish: Apache Nifi, Kibana
Ahnaf Shahriyar Chowdhury: Pyspark
Syed Ali Javed: Hive
Sameer Ul Haq: Kibana

Workflow

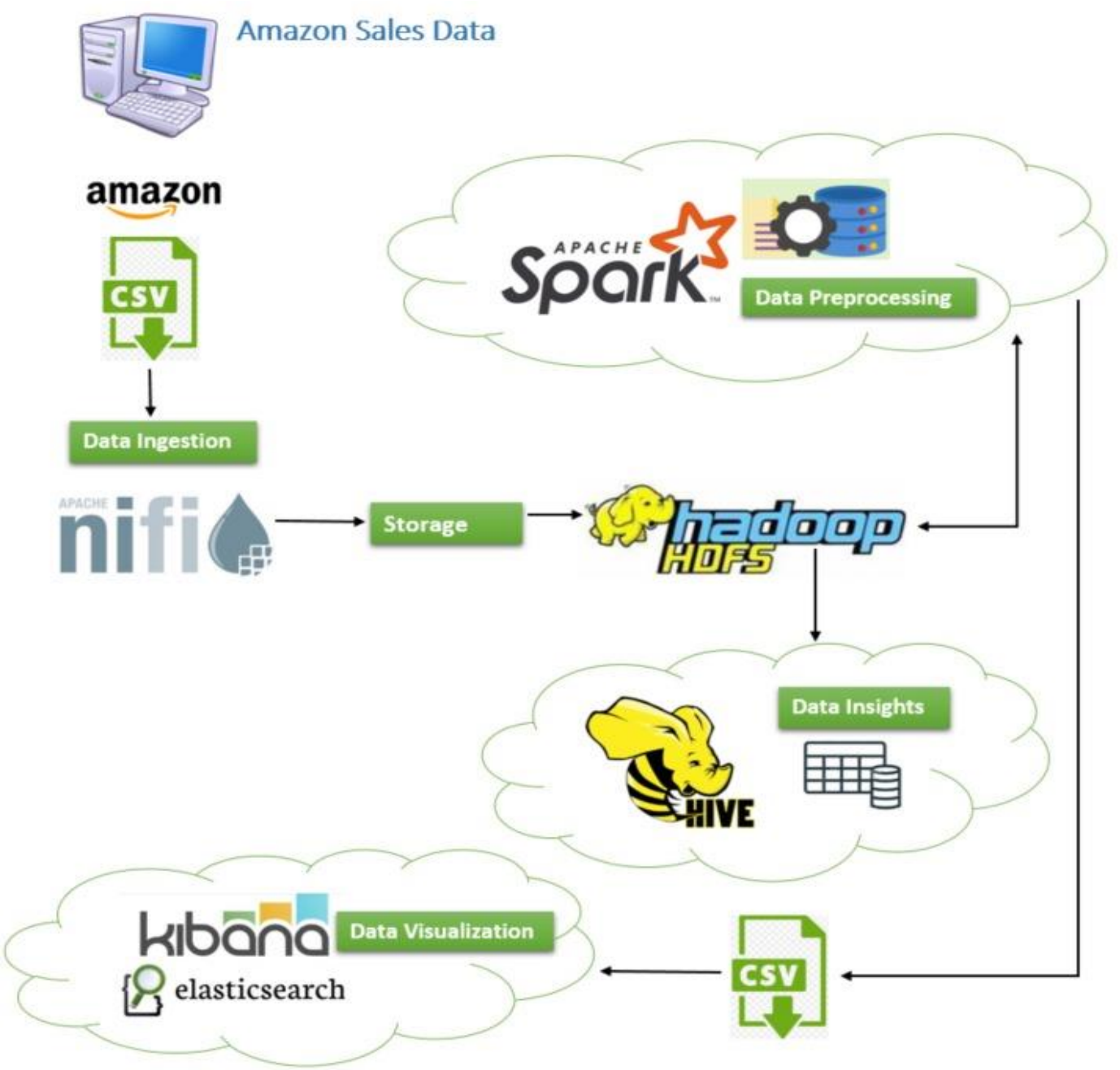
NiFi: For ingesting data from the local system to HDFS.

HDFS: Storage for scalable and fault-tolerant data retrieval.

PySpark: Data preprocessing, deriving attributes, and exploratory analysis.

Hive: Querying data to extract actionable insights.

Kibana: Visualizing key trends and patterns

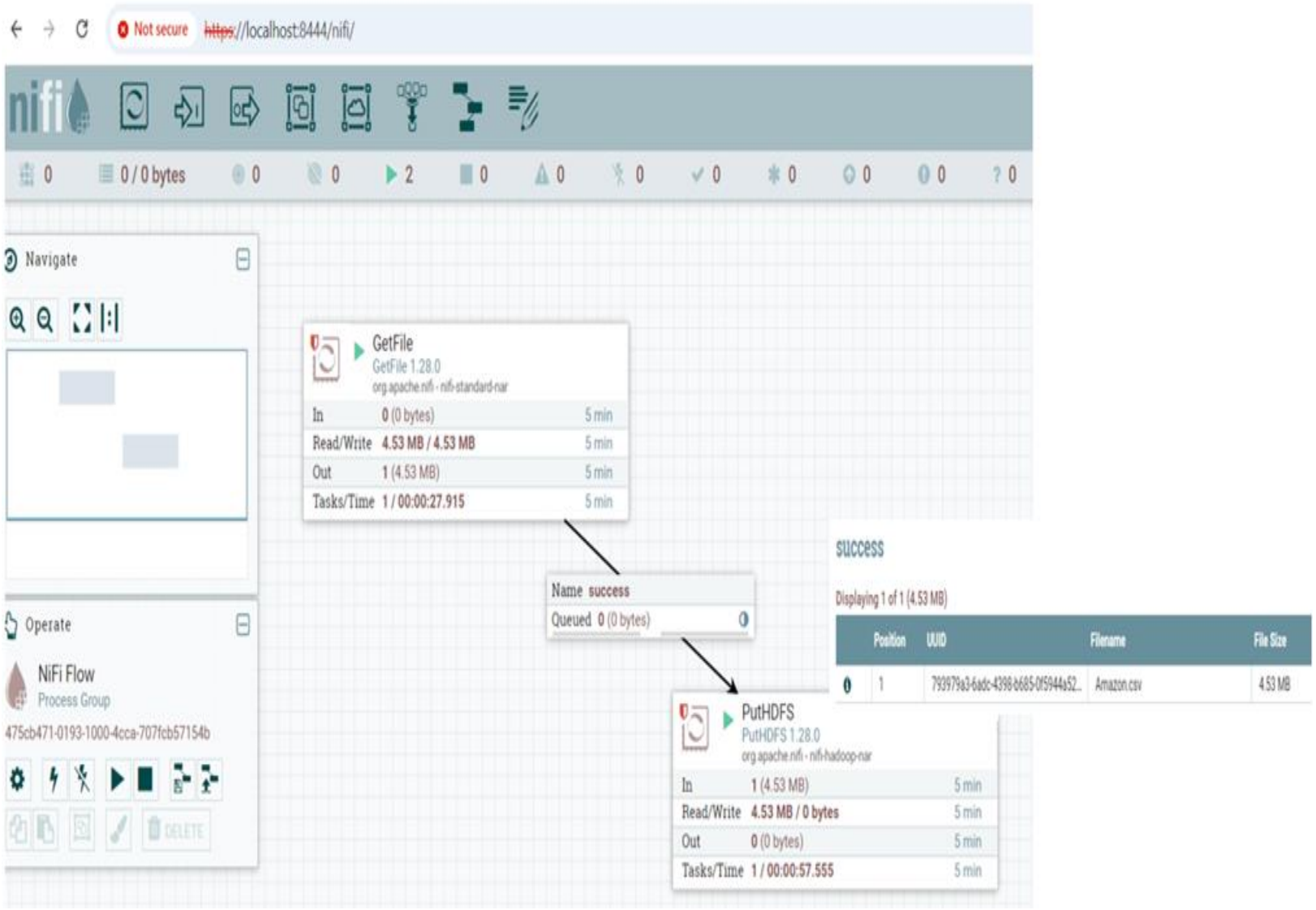


Apache Nifi – Data Ingestion

GetFile: Reads a file from the local system and retrieves the file Amazon.csv.

putHDFS: Writes the file to HDFS for storage and further processing.

Success Log: Confirms the successful transfer



PySpark - Data Preprocessing

Data Cleaning → Feature Creation → Data Normalization → Analysis → Final dataset

- Removed currency and comma symbols from actual price, discounted price and rating count column
- Convert the column from string to numerical formats.

Example: ₹1,208 to 1208

- Two new features added for better comparison and insights
- These are revenue and sales volume

- Applied to fields like sales volume and revenue and scaled to a range of 0 to 1
- Ensured fair weighting in analysis and insights

Example: 754706.4 to 0.401

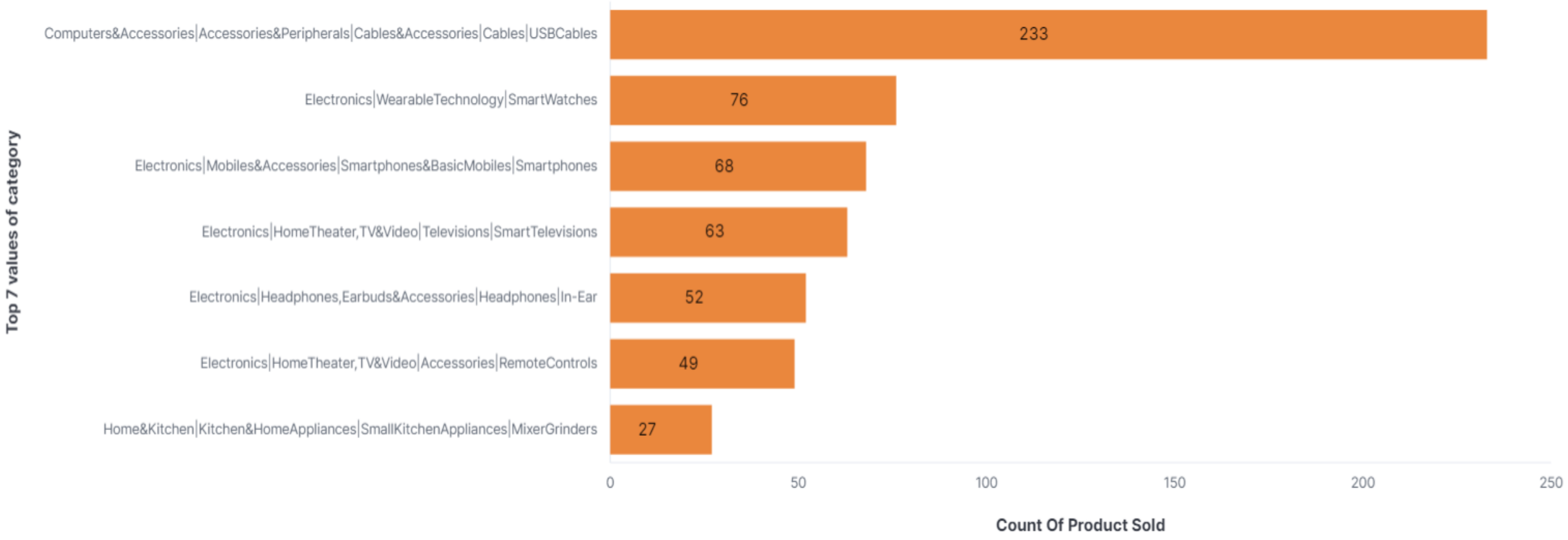
- Category analysis
- Discount analysis
- Rating analysis

Final normalized dataset ready for:

- Hive Query application
- Visualization using Kibana

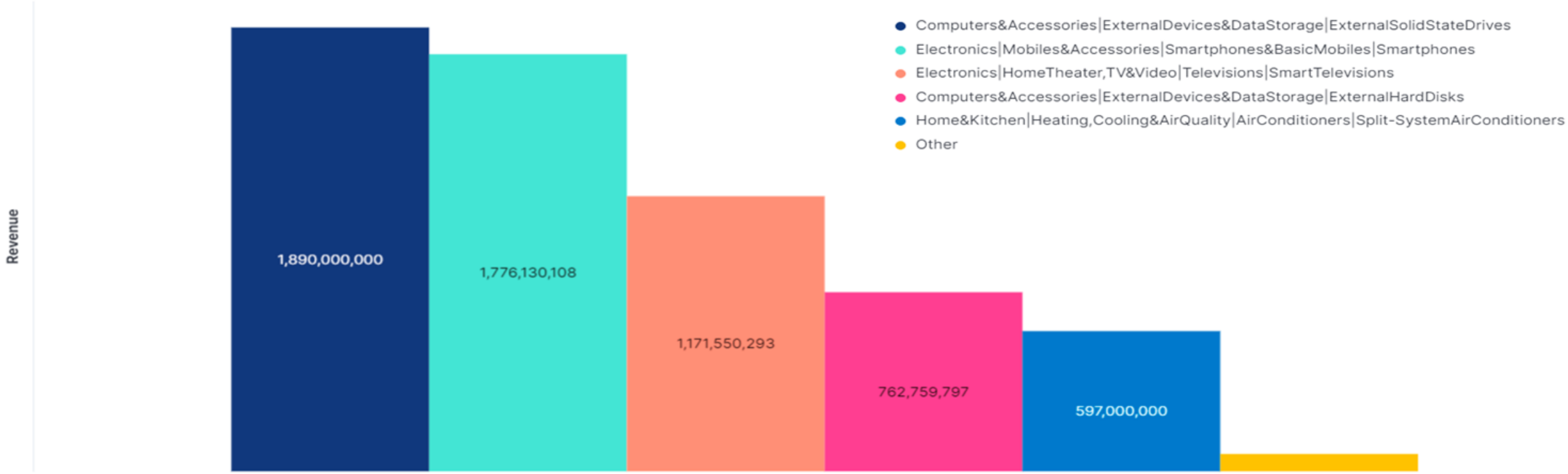
Insight 1

Product Popularity by Category



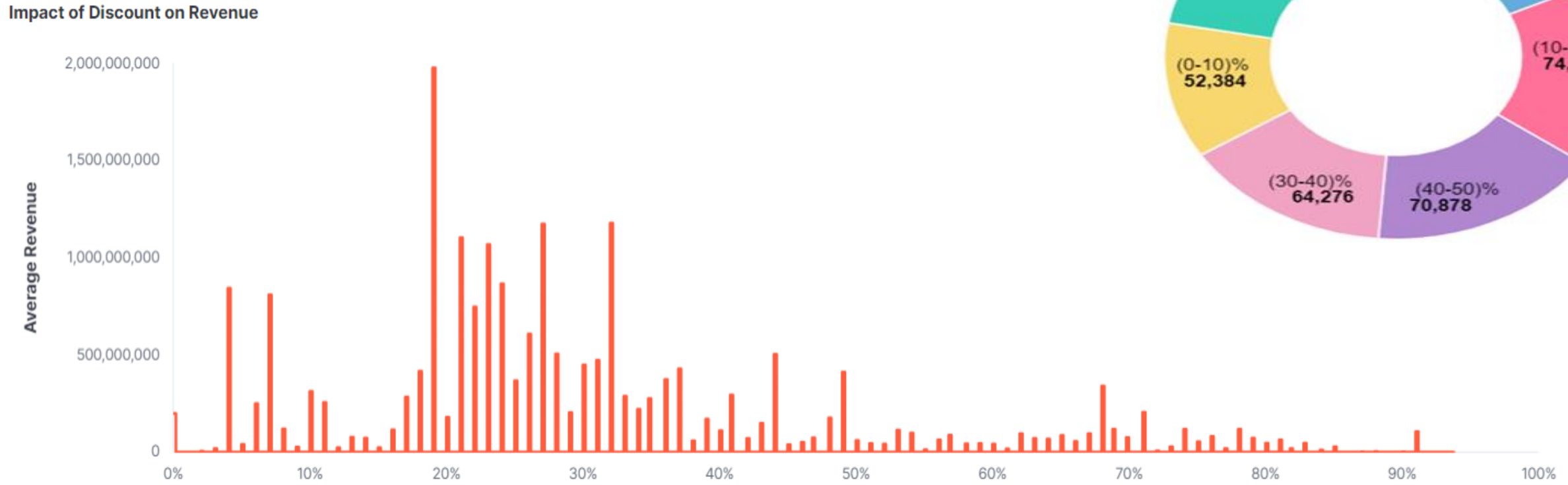
Insight 2

Revenue Distribution Across Categories



Insight 3

Impact of Discounts on Sales Volume and Revenue

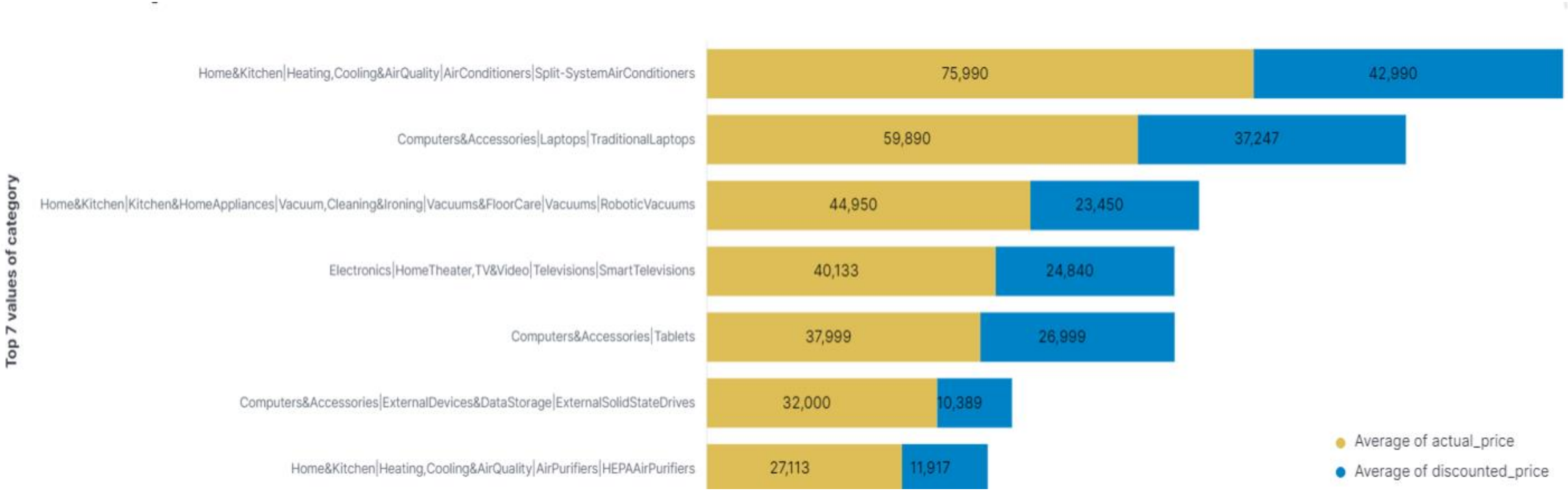


Insight 4

Effect of Rating on Revenue

Insight 5

Premium Product Categories and Discount Trends



Learning Curve

Data Insights: Analyzing product features like price, discount, and ratings is crucial for understanding sales performance.

Tool Integration: Combining tools like NiFi, PySpark, Hive and Kibana streamlined our data processing, but integration challenges were encountered.

Real-Time Analytics: Moving towards real-time data pipelines can provide dynamic insights for improved decision-making.

Future Scope: Enhance predictive analytics, real-time insights with Apache Kafka, and explore customer sentiment through NLP can uncover deeper insights into product performance and areas for improvement.

Any Questions?