eCommerce Analytics and Recommendations

Team Name - "FourYottabytes"

Amogh Hassija (amoghhassija@iisc.ac.in)
Ashwin Korwarkar (kashwin@iisc.ac.in)
Henna Arora (hennaarora@iisc.ac.in)
Sai Harish Pathuri (saipathuri@iisc.ac.in)

PROBLEM STATEMENT

Definition

Analyzing e-Commerce customer behavior to provide recommendations for a better experience using Apache Spark Dataframes. The data being used is eCommerce user behavior data from kaggle which consists of user's event history on the platform.

Data: [eCommerce behavior data from multi category store | Kaggle](#)

Motivation

User behavior data refers to the data collected by e-commerce platforms about the actions and activities of their users on their website or app. This data can include information such as the pages a user visits, the products they view, the searches they perform, and the actions they take, such as adding items to their cart or making a purchase.

Having access to this type of data can be incredibly useful for e-commerce platforms, as it can help them better understand their users and tailor their services to meet their needs. For example, by analyzing user behavior data, an e-commerce platform can identify the most popular products among its users, as well as the factors that influence a user's decision to make a purchase. This information can then be used to optimize the platform's product offering and improve the user experience. User behavior data can also be used to personalize the shopping experience for individual users. By analyzing a user's behavior data, an e-commerce platform can make recommendations for products that are likely to be of interest to that user based on their previous actions and interactions with the platform. This can help to increase engagement and drive sales, as users are more likely to make purchases when they are presented with personalized recommendations that are relevant to their interests and preferences.

In addition to helping e-commerce platforms understand and serve their users, user behavior data can also be used to identify potential issues and improve the overall performance of the platform. For example, by analyzing user behavior data, an e-commerce platform can identify areas of the website or app that are causing users to become confused and make changes to improve the user experience. This can help to reduce user churn and increase customer satisfaction, which are crucial factors for the success of any e-commerce platform.

Design Goals

E-commerce data is a prime example of big data, as it includes many transactions, customer interactions, and other types of data. This data can be challenging to analyze using traditional methods, so big

data processing tools like Apache Spark are needed to process and analyze it effectively. Also, in an ideal setup, there can be many teams working in the Analytics department, each formulating a way of recommending products to the users which have the highest chance of purchase. Each of these teams would need to have access to some kind of data. Having a centralized data pipeline is very much required in this situation. The ETL pipelines would have to strategically process the data in a way that they meet every team's requirements. All these factors have been taken into consideration while developing the modules.

Features Required

The following features have been built to support the analysis we have performed for the last four weeks.

1. Centralized ETL Pipeline
2. Personalized Product Recommendations
3. Price Point Analysis Recommendations
4. Market Basket Analysis Recommendations
5. Promoting Underdogs and Rank personalization

Scalability and Performance Goals

There are several ways a big data platform can scale, depending on the specific needs and goals of the organization using it. One approach is to use a distributed architecture, which involves dividing the data and processing it across multiple machines or nodes. This can help improve the performance and scalability of the platform, as it allows for parallel processing of data and enables the system to handle large amounts of data more efficiently.

This strategy perfectly fits for the current use-case because the number of customers and events on the e-Commerce platform can dramatically increase. In case of a rise in the number of events/transactions, the platform must ensure that it is able to meet the demands of all the customers. From the platform point of view, this includes that all the features that were discussed in the above section are able to loosely scale even in the event of sudden rise in traffic. Since we have used Apache Spark as the base processing tool, it provides access to certain parameters like number of cores, executors, and memories, among many other tunings which can help scale the platform in such cases.

SOLUTION APPROACH

High Level Design

We build a centralized ETL store to decouple the data processing from the data analysis. We have followed a modular design for developing a solution to the problem statement. Modularity is an important concept in software design, as it allows multiple teams to work on different parts of a project without interfering with each other. By creating a centralized ETL (extract, transform, load) design, we have effectively decoupled the data processing from the data analysis, allowing each team to focus on their own tasks without impacting the work of the other team. This has improved efficiency and collaboration within the team.

The solution comprises of two key modules:

1. ETL Engine
2. Recommendation Engine

The ETL Engine: The extract step involves extracting data from the raw dumps. The transform step involves cleaning, filtering, and transforming the data to a consistent format and structure. The load step involves loading the transformed data into a data warehouse or database (in suitable formats such as transaction logs, customer profiles, and product catalogs) for further analysis and reporting. ETL is typically performed using the specialized ETL tool: Apache Spark, to ensure efficient and reliable data processing.

Recommendation Engine: The data generated from the ETL is analyzed to identify patterns and trends in customer and product behavior, such as preferences (also product preferences) and interests. Based on these insights, the recommendation engine generates personalized recommendations, such as product or content suggestions, to individual customers. These recommendations are typically displayed on the website or app, and can improve customer engagement and conversion rates. We have used various methods of recommendations for various types of customers.

Architecture Diagram
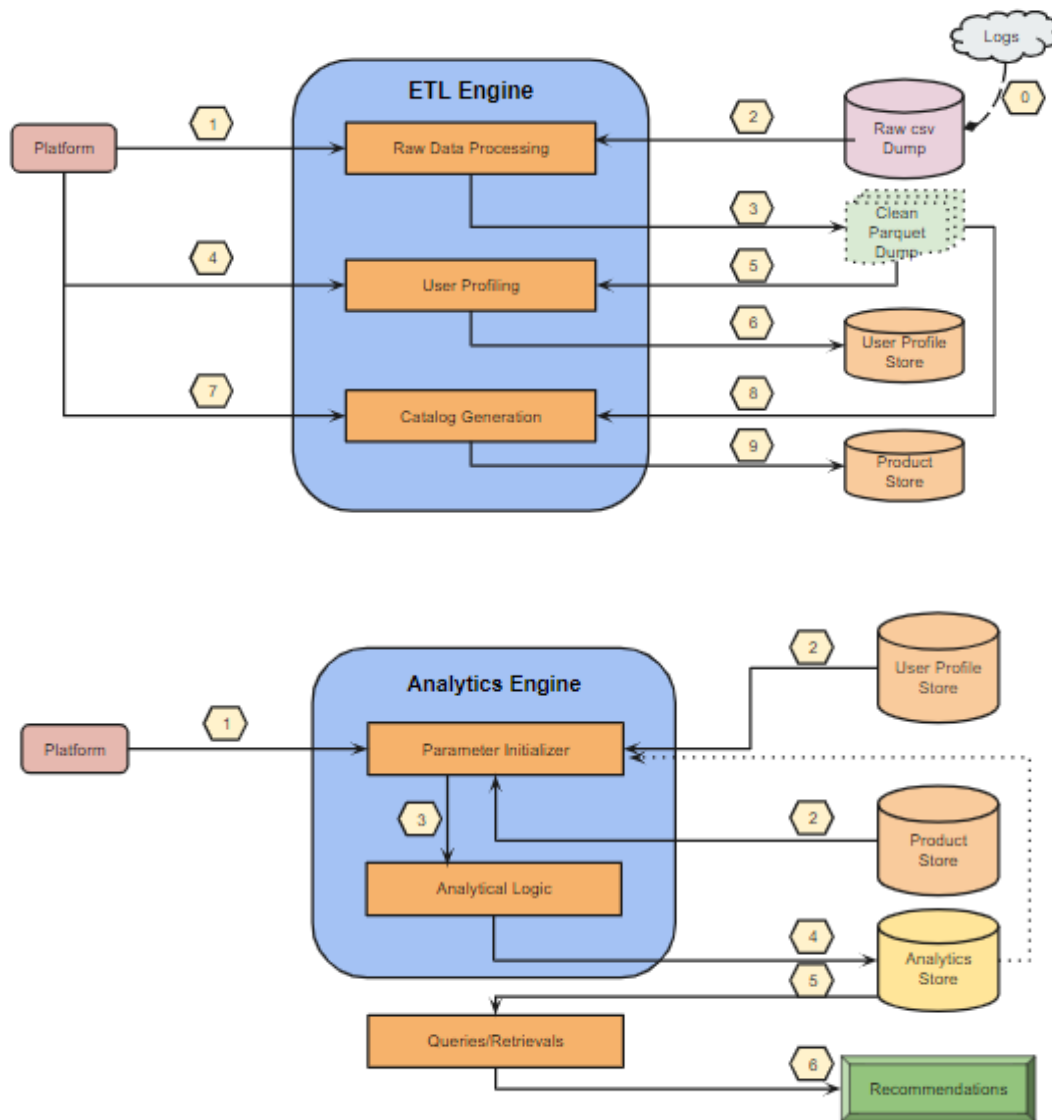
Big Data Platforms

This project has used Apache Spark Dataframes for processing and analyzing the data. We have also used Spark ML algorithms like K-Means Clustering to understand the user groups and segment the users as part of EDA. As part of Market Basket Analysis, we have used the FP Growth Algorithm from Spark ML-lib.

EVALUATION

Experiment Design

1. Personalized Product Recommendation: To provide product recommendations to the users based on their average spendings or views on the portal

Motivation: We need a more robust engine that could provide personalized recommendations rather than simply recommending products that sold the most.

Analytics:

- Using the given the threshold values, the engine categorizes the users into **purchase, view or cold-start**
- Top products are picked based on the average price of the user.
- This ensures that the probability of conversion of the event from "view" to "purchase" is high.
- The recommendations are done for Homepage (All categories) or Category page(Example: Electronics)

2. Underdog Product Analysis: To find out products with low views but high conversion rate.

Motivation:

- Push "Underdog" products – Products with lesser views but high conversion rate

- Boost products which are unexplored but when visible mostly leads to a sale

- Product visibility is an important analytics tool to drive business sales.

Analytics

- Generate product conversion rate (PCR) & product view count (PVC)

- Generate product's category average conversion rate (CCR) and average category view count (CVC)

- If (PCR > x% of CCR) && (PVC < y% of CVC) are underdog products

3. Personalized Placement Analysis:
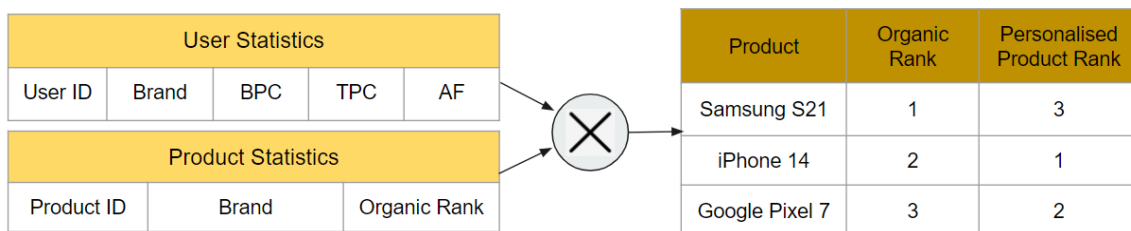
Definition

- Product Rank - Ranking position of any product in the organic eCommerce search results.
- Personalized Product Rank - Incorporate user preference to get revised rank

Motivation

- Enhance user experience by giving a personalized touch
- Higher conversion rate to drive enhanced sales

Analytics

- Generate organic product rank based on conversion rates
- Generate user preference based on adjustment factor (brand purchase count/total purchase count)
- Adjustment factor multiplier on organic product rank to get revised product rank

User Statistics				
User ID	Brand	BPC	TPC	AF

Product Statistics		
Product ID	Brand	Organic Rank

| Product	Organic Rank	Personalised Product Rank
Samsung S21	1	3
iPhone 14	2	1
Google Pixel 7	3	2

4. Market Basket Analysis

Definition: A set of measures that show what combination of products most frequently occur together in orders.

Motivation
- A standard technique to uncover association between products.
- Improvising on top of the basic analytics stores, we provide recommendations of the form "Frequently bought together" or "Customers also bought."
- It fit well to our need and was MUST to build a good recommendation platform.
- Association made between products based on their -
  - Product Purchases - product purchases per user, product purchases per user session
  - Product Views - product views per user, product views per user session

Analytics: We built below analytics from the clean parquet dumps :-
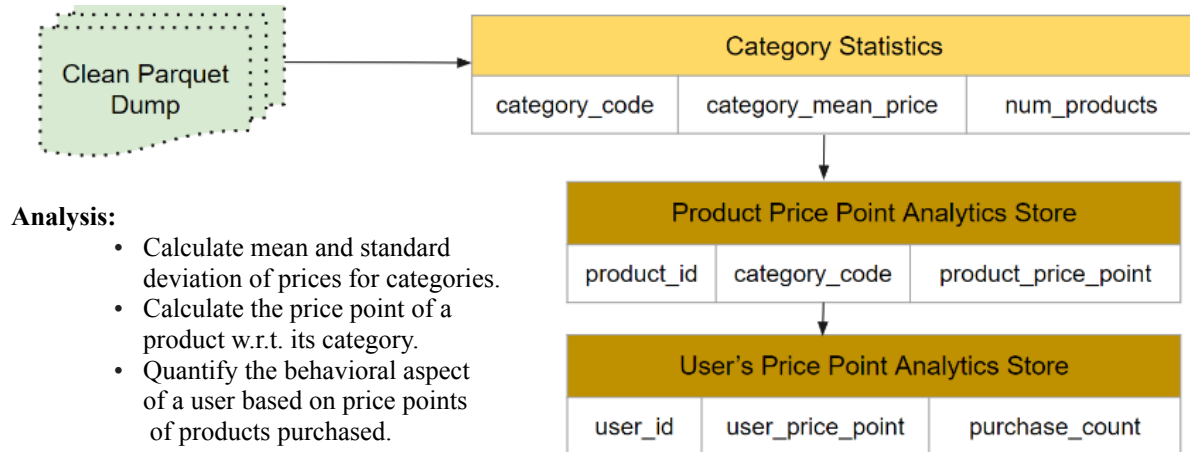- Our own associative products confidence scores for two products
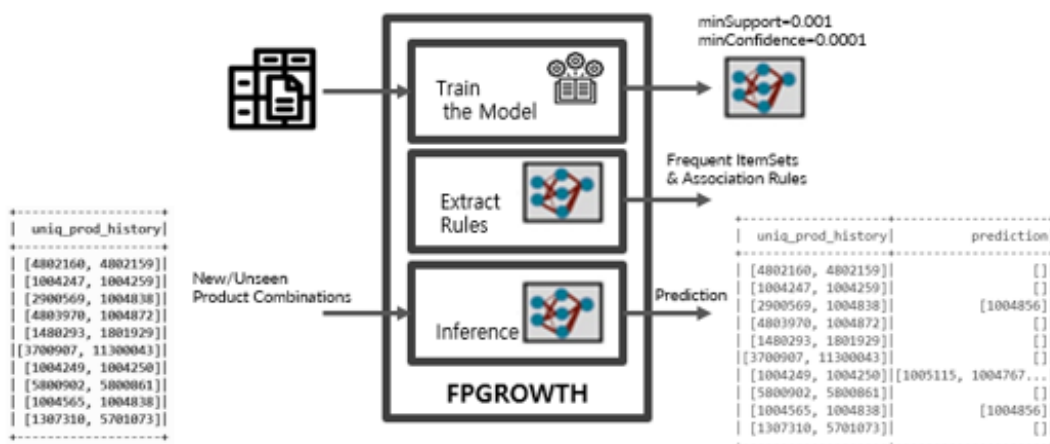
- Spark ML based FPGrowth data mining model to build rules and predict the product that can be added to a new/unseen product combination(s).

5. Price Point Analysis:

The Aim: Quantitative estimation of the user's *purchasing power*.

Motivation:

- Different users purchase products belonging to different classes (cheapest/budget/mid-range/top-line) based on their capacity and preference.
- We regard purchase history as a strong metric of the user's behavioral aspects and try to quantify the same (price_point).

Analysis:

- Calculate mean and standard deviation of prices for categories.
- Calculate the price point of a product w.r.t. its category.
- Quantify the behavioral aspect of a user based on price points of products purchased.

```
query_PPA2(["568782581", "512480149"], "electronics.smartphone")
```

```
User-568782581
Cold Start User
User-512480149
+----------+---------------------+
|product_id|category_code        |
+----------+---------------------+
|1003532   |electronics.smartphone|
|1003705   |electronics.smartphone|
|1004777   |electronics.smartphone|
|1004732   |electronics.smartphone|
|1003989   |electronics.smartphone|
|1004740   |electronics.smartphone|
|1004819   |electronics.smartphone|
|1003548   |electronics.smartphone|
|1003707   |electronics.smartphone|
+----------+---------------------+
```

Scalability Metrics

Since the data we have is for around 9 months, we ran the ETL jobs for various counts of months to generate the analysis for the last 'k' months. The ETL module and Recommendation module scaled as follows:

User Profiling

# Months	Ram (GB)	# Cores	Time (min)
1	8	2	13
2	8	2	28
3	8	2	42
4	8	4	66
4	40	12	8

Recommendation Profiling

| # Months	Ram (GB)	# Cores	Purchase Threshold	View Threshold	# Users	Time (min)
1	2	2	4	30	20	3
2	12	2	10	200	120	3
2	40	12	30	100	120	1
4	40	12	10	200	500	0.8
4	40	12	10	200	2050	2.5
6	40	12	10	200	4000	6.5

Plots and Analysis - EDA

As part of EDA, we have thoroughly studied the data to understand the columns and their related information. Below are some of the charts that showcase the preliminary analysis of the dataset.
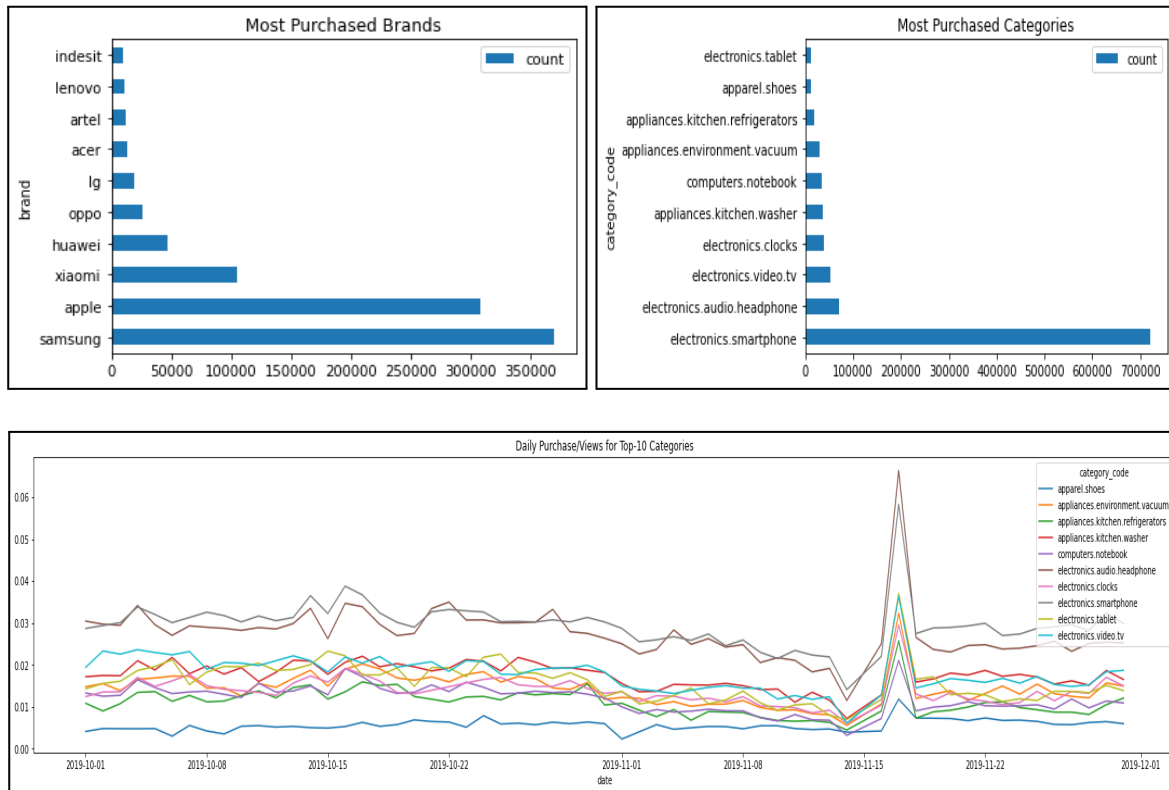
Customer Segmentation: (K means)
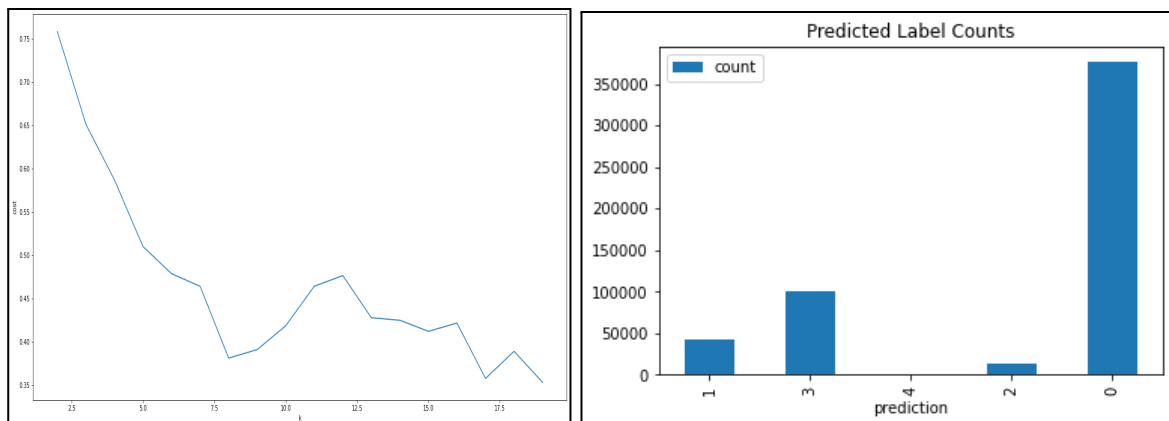
Above is the plot of k (Number of clusters) vs cost (silhouette score) for customer segmentation. User distribution for k=5 is shown.
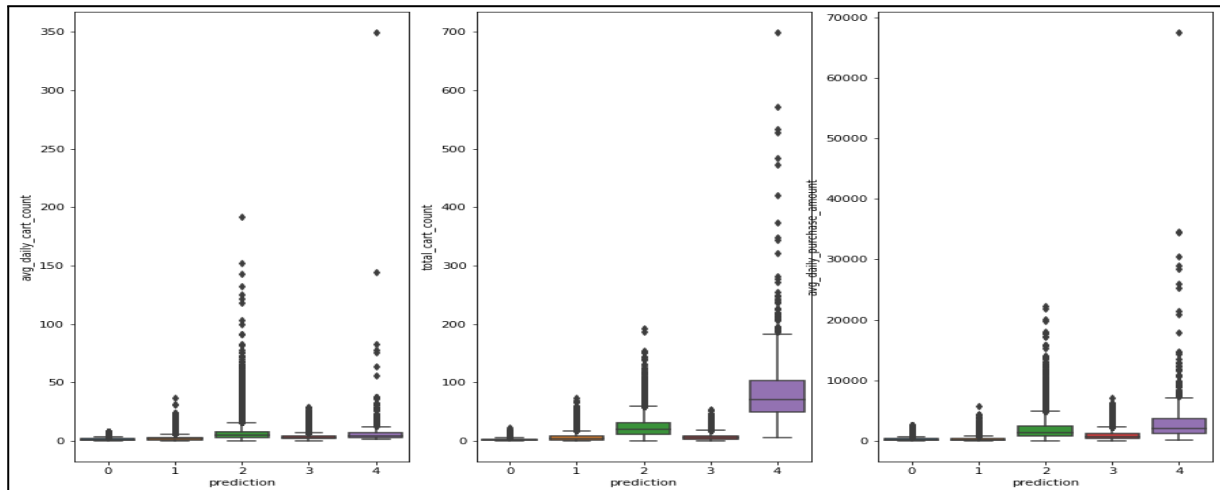
SUMMARY

Through this project, the team has understood how recommendation systems are an important feature of ecommerce platforms, and how they are designed to help users discover new products that they might be interested in purchasing. Our recommenders analyzed a user's past purchasing history, browsing behavior, to make personalized recommendations for products that the user might like.

The major goal in the project proposal was to do some compound analysis on the behavior data to bring out some interesting insights. Through these recommender systems and the analysis, we believe we have fared well. Our solutions showcased some scaling as well. We have also used Spark ML packages to use some algorithms like K-Means for customer segmentation, FPGrowth for mining the associations between products for each user. These algorithms have helped us do the Market Basket Analysis, which gives an insight into how the association rules can help recommend products which have the highest probability of being bought together.

As a future study, we can build some ML based recommenders using techniques like collaborative filtering, and content based similarity recommenders. A comparison between current recommenders and the ML recommenders can also be taken up to gauge the performance of the models.