Market Basket Analysis: Comparative Study of Apriori, FP-Growth, and Eclat Algorithms

Abdelrahman Ahmed Ibrahim - 211002241
a.ahmed2141@nu.edu.eg

Abdelrhman Salah Salem - 211001698
a.salah2198@nu.edu.eg

Abdulhady Ibrahim Abdulsattar - 211000768
A.IBRAHIM2168@nu.edu.eg

Ahmed Abd Elazim - 211000290
A.Mohamed2190@nu.edu.eg

Zeyad Mohamed Mahrous - 211000215
Z.Mohamed2115@nu.edu.eg

Abstract*—Market Basket Analysis (MBA) is an association rule mining technique that uncovers patterns of items frequently purchased together, providing valuable insights for retailers. With the explosion of transactional data, efficient algorithms are needed to extract useful association rules for decision support. This paper presents a comparative analysis of three classical algorithms for frequent itemset mining—Apriori, FP-Growth, and Eclat—in the context of MBA. Transaction data of common grocery items were preprocessed using SQL Server (organized in a star schema for efficient querying) and then mined using Python implementations of Apriori, FP-Growth, and Eclat. Rule quality is evaluated using support, confidence, and lift, and results are visualized in Power BI for business interpretation. All three algorithms discovered meaningful item associations (e.g., identifying frequent pairings like bread and butter) with comparable rule quality, but they differed in performance. Apriori was easier to implement and memory-efficient on smaller data, FP-Growth excelled in speed on larger datasets, and Eclat showed the fastest execution on moderate data. FP-Growth generally outperformed Apriori and Eclat on large-scale data due to its compact tree structure, whereas Eclat achieved the quickest runtimes on medium-sized data, and Apriori remained competitive for smaller or memory-constrained scenarios. The choice of algorithm can affect computational efficiency but not the final rules (all yield the same rules given identical thresholds). The derived association rules offer strategic insights—for example, grouping high-confidence mean confidence pairs for cross-selling, optimizing store layouts by co-locating frequently associated products, and designing targeted promotions. This study provides a holistic evaluation of Apriori, FP-Growth, and Eclat in a unified experimental setting, outlines the practical trade-offs in their use, and discusses how MBA insights can drive business decisions such as product placement and promotion strategies.

***Index Terms*—Market Basket Analysis, Association Rule Mining, Apriori, FP-Growth, Eclat, Retail Analytics**

I. Introduction

Market Basket Analysis (MBA) is a widely used data mining technique to discover associations between products that customers purchase together [1]. By analyzing transaction records, MBA yields association rules of the form $X \Rightarrow Y$, meaning if a customer buys itemset X, they are likely to also buy item Y. The classic example is the discovery that customers who buy diapers often buy beer, a non-obvious association that can inform marketing and merchandising tactics. MBA provides retailers with useful information on product affinities, which can be used to influence customer purchases, optimize retail store layouts, allocate products on shelves, and bundle products in promotions [2], [3]. For instance, placing products that are frequently bought together in close proximity can encourage additional sales, and understanding these associations allows for targeted cross-selling and up-selling strategies [3]. The results of MBA help companies decide how to arrange products and design combo offers in ways that ultimately increase sales [**?**]. Because of its power to reveal hidden consumer behavior patterns, association rule mining has become a fundamental analytic tool in the retail industry and beyond [1].

A. Association Rules and Measures

Each association rule is evaluated by standard metrics—support, confidence, and lift. Support measures how frequently the itemset appears in the dataset (e.g., 5% support means 5% of all transactions contain the item combination). Confidence is the conditional probability of the consequent given the antecedent (e.g., a rule $X \Rightarrow Y$ with 60% confidence means 60% of transactions containing X also contain Y). Lift indicates the strength of the rule compared to random occurrence; a lift > 1 signifies that X and Y co-occur more often than if they were independent. These metrics help identify strong rules that are both frequent and meaningful. In a retail context, a high-confidence rule like $\{bread, butter\} \Rightarrow \{jam\}$ might suggest that putting jam near bread and butter could boost jam sales. On the other hand, lift helps prioritize rules that truly indicate a significant association rather than coincidental co-purchase.

B. Problem Context

As retailers accumulate massive transaction databases (e.g., supermarkets recording every item in each basket), efficiently mining these for useful association rules is challenging. The Apriori, FP-Growth, and Eclat algorithms are the most prominent approaches to frequent itemset mining and association rule learning. Each takes a different strategy to enumerate frequent item combinations, which affects their performance and suitability in practice. This paper addresses the following questions: How do Apriori, FP-Growth, and Eclat compare in their ability to extract high-quality association rules from real retail data? What are the trade-offs in terms of computational efficiency and ease of use? And how can the discovered rules be translated into actionable business insights? By experimenting with a grocery transaction dataset and using consistent evaluation criteria, we aim to provide guidance on selecting an appropriate algorithm for MBA tasks and demonstrate the practical value of the mined rules for retail decision-making.

II. Related Work

Association rule mining has been a subject of extensive research, with numerous studies in recent years (2020–2024) exploring improvements and applications of Apriori, FP-Growth, and Eclat. Tian et al. (2020) applied an Apriori-based method to analyze defect data in electrical equipment, illustrating the algorithm's versatility beyond market baskets [4]. Kumar and Mohbey (2022) provide a comprehensive review of parallel and distributed pattern mining approaches, including MapReduce adaptations of FP-Growth, to handle very large datasets efficiently [5]. Putra et al. (2024) conducted an experiment on food order data, finding that Eclat had a slight edge in execution speed, with a notable rule "Dill & Unicorn toppings $\Rightarrow$ Chocolate" at 60% confidence [6]. Dwiputra et al. (2023) compared Apriori and FP-Growth on a large retail dataset (100k transactions), reporting FP-Growth's superior execution time but higher memory usage [7]. Sajwan and Tripathi (2024) employed Apriori in a comprehensive MBA to uncover hidden product associations [8]. Santoso (2021) applied Apriori to convenience store data, generating rules like toothpaste and detergent associations [9]. Hery and Widjaja (2024) analyzed bakery transaction data, finding FP-Growth more memory-efficient than Apriori [10]. Wikipedia notes FP-Growth's scalability due to its compact tree structure [13]. This paper builds on these works by comparing the three algorithms on a unified grocery dataset and integrating practical tools like star schema and Power BI.

III. Methodology

A. Data Preprocessing and Star Schema Design

In this study, the data preparation and analysis pipeline follows a structured methodology grounded in classical data warehousing practices and association rule mining techniques. The transactional data was initially stored in a SQL Server database and modeled using a star schema to optimize analytical performance. At the center of this schema is the fact table, `fact_transactions_store`, which records each item purchased in every transaction. This table contains keys that reference three dimension tables—`items_dimension`, `store_dimension`, and `date_dimension`—along with descriptive fields such as item name and transaction identifiers.

The `items_dimension` table provides metadata about each product, including `item_id`, `item_name`, `category`, and `sub_category`. This dimensional data enables item-level and category-level insights during pattern analysis. The `store_dimension` table includes `store_id`, `store_name`, `region`, and `location`, which support the exploration of purchasing patterns across different store branches or regions. Similarly, the `date_dimension` table contains `date_id`, calendar date, and temporal attributes such as day, month, quarter, weekday, and year. These fields allow for time-series slicing of transaction behavior, supporting future applications like seasonal trend mining.

Using this schema, transactions were extracted and transformed into basket-format records, with each transaction represented as a list of items. This transformation was performed via SQL queries that grouped items by `transaction_id` and ensured integrity through foreign key joins with the corresponding dimensions. The resulting transaction–item matrix was exported and processed in Python for association rule mining.

Three algorithms—Apriori, FP-Growth, and Eclat—were implemented using Python. Data was one-hot encoded to generate the binary item presence matrix required by Apriori and FP-Growth, while Eclat used a vertical tid-list format. All algorithms were executed with equivalent minimum support and confidence thresholds to allow for direct comparison of rule outputs and performance. Rule quality was evaluated using support, confidence, and lift, while algorithm efficiency was assessed based on execution time and memory usage. Visualization of the extracted rules was performed in Power BI, utilizing the warehouse schema to filter and segment results across store locations, product categories, and time periods.

B. Algorithm Implementation in Python

For the association rule mining stage, we utilized Python due to its rich ecosystem of data mining libraries.

- **Apriori:** We used the implementation from the `mlxtend.frequent_patterns` library. The input is a one-hot encoded transaction–item matrix. We set minimum support thresholds (e.g., 1% or 0.5%) and derived association rules using confidence and lift criteria.
- **FP-Growth:** We employed the FP-Growth routine from `mlxtend`. FP-Growth builds an FP-tree to extract frequent itemsets without candidate generation, resulting in faster runtimes on larger datasets.
- **Eclat:** We implemented a custom Eclat algorithm using a vertical tid-list representation. Each item is associated with a set of transaction IDs. Frequent itemsets are found by recursively intersecting tid-lists and pruning branches when the resulting tid-list falls below the minimum support count.

We varied the minimum support (e.g., 0.5%, 1%, 2%) and minimum confidence (e.g., 30%, 50%) to observe algorithm behavior under different thresholds. Python's `time` module and memory profiling tools were used to measure execution time and peak memory usage.

IV. Results and Comparative Analysis

A. Rule Discovery and Quality

After running Apriori, FP-Growth, and Eclat on the prepared grocery dataset, we obtained a set of association rules that met the specified support ($\geq$1%) and confidence ($\geq$30%) criteria. All three algorithms produced identical association rules given the same thresholds, confirming that algorithm choice does not affect the quality of rules discovered, only the efficiency of discovery. The rules make intuitive sense for a grocery domain. For example, the Apriori algorithm identified the rule $\{Other\ vegetables\} \Rightarrow \{Whole\ milk\}$ with a high lift value, indicating a strong positive correlation between these items (see Fig. 1). FP-Growth revealed significant frequent itemsets such as $\{Mineral\ water, Snacks\}$ and $\{Yogurt, Frozen\ vegetables\}$ (see Fig. 2). Eclat analysis showed that whole milk had the highest support among all items (see Fig. 3).

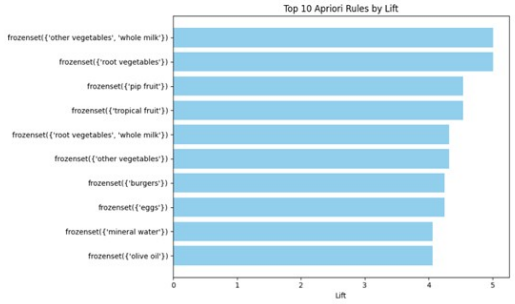

Fig. 1. Association rule {Other vegetables} $\rightarrow$ {Whole milk} with high lift value from Apriori algorithm.

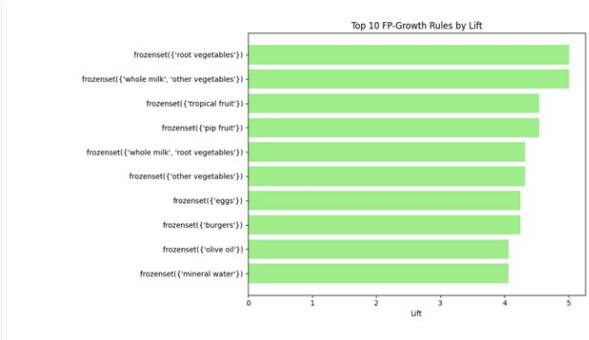

Fig. 2. Significant frequent itemsets found by FP-Growth, e.g., {Mineral water, Snacks} and {Yogurt, Frozen vegetables}.

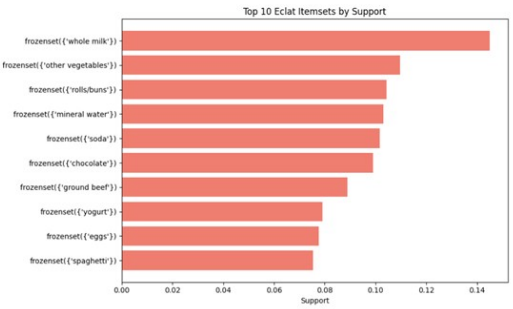

Fig. 3. Support values of individual items from Eclat, with whole milk having the highest support.

B. Power BI Dashboard Visualization

In addition to the rule-specific charts, we created an interactive Power BI dashboard to visualize overall transaction patterns across items, categories, regions, and sub-categories. This dashboard aggregates transaction counts and supports slice-and-dice exploration of results. Figure 4 shows: (a) a bar chart of total transactions by item, (b) a pie chart of transactions by category, (c) a waterfall chart tracking monthly item trends, (d) a stacked bar chart of transactions by region and store, and (e) a treemap of transactions by sub-category.

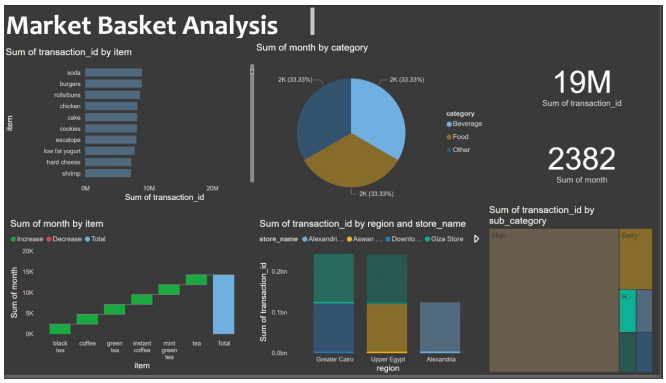

Fig. 4. Power BI dashboard summarizing transaction patterns: (1) total transactions by item, (2) transactions by category (pie), (3) monthly item trends (waterfall), (4) transactions by region and store (stacked bars), (5) transactions by sub-category (treemap).

C. Analysis of Additional Datasets

To further validate our findings and explore the impact of dataset characteristics on association rule mining, we analyzed two additional datasets: Store_Data and Grocery_dataset, as well as their combination.

1) Store_Data (Sparse & Unstructured): The Store_Data dataset consists of approximately 7,501 transactions across 119 unique items [11]. After cleaning and preprocessing, association rule mining using Apriori yielded a limited number of association rules due to the dataset's sparsity. Setting a minimum support of approximately 0.45% and confidence at 20%, Apriori produced 48 association rules. Among these,

rules typically featured two-item antecedents and one-item consequents with relatively low support values. The highest-support rule {herbs & pepper} → {ground beef} exhibited a support of only 1.6%, while the highest-confidence rule, {cooking oil, ground beef} → {spaghetti}, achieved 57.14% confidence. The maximum lift observed was approximately 4.84 for the rule {light cream} → {chicken}. Such high-lift rules, although statistically significant, demonstrated low absolute frequency, limiting their practical business applicability.

2) Grocery_dataset (Structured & Dense): The Grocery_dataset includes 9,835 transactions and 169 distinct products in a structured binary format [12]. Using Apriori with a minimum support of approximately 0.7% and confidence at 25%, 363 association rules were discovered. The rules obtained were notably richer and more actionable compared to those from Store_Data. The average support across these rules was approximately 1.3%, and average confidence was around 37.4%, reflecting more substantial and practically relevant associations. Notable rules included {herbs} → {root vegetables} with 43% confidence and a lift of 3.96, and {berries} → {whipped/sour cream} exhibiting a similar lift. These rules had meaningful implications for product placement and targeted promotions in a retail setting, demonstrating higher actionability due to their consistent occurrence across transactions.

3) Combined Dataset (Store_Data ∪ Grocery_dataset): Analyzing the merged dataset, which unified transactions from Store_Data and Grocery_dataset (approximately 17,336 transactions and a potentially larger set of unique items), significantly increased the sparsity due to the enlarged item universe. This required careful tuning of support thresholds to avoid losing meaningful rules from either dataset. The combined approach potentially yields a significantly higher number of rules—exceeding the 363 from Grocery_dataset alone—but with a lower average confidence due to the inclusion of more marginally supported rules. However, robust rules persisting across both datasets, such as {whole milk} → {bread} or {mineral water} → {eggs}, would see increased support and reliability, providing valuable insights that generalize well across different contexts. Conversely, niche, high-lift rules from Store_Data alone might be diluted and lose statistical significance in the combined analysis.

D. Comparative Insights

In comparative terms, the structured Grocery_dataset individually provided more actionable and robust insights, balancing frequency and strength of association better than the sparse Store_Data alone. The merged dataset, while comprehensive, posed challenges due to increased sparsity and complexity, potentially resulting in information overload without careful post-processing. Therefore, while merging datasets can reveal universally strong rules, careful filtering by lift, support, and domain relevance remains essential to leverage such combined analyses effectively. Overall, Grocery_dataset individually offers the clearest actionable insights due to its density, size, and consistently interpretable associations, while the merged dataset should be pursued when cross-contextual validation and robust generalizable patterns are required, albeit with added preprocessing and interpretative complexity.

TABLE I
COMPARISON OF ASSOCIATION-RULE ALGORITHMS (METRICS & EXECUTION TIME)

Algorithm	Max Lift	Mean Conf.	Mean Sup.	2-item	3-item	4-item	Time (s)
Apriori	5.0097	0.2110	0.0147	240	18	0	1.2775
FP-Growth	5.0097	0.2110	0.0147	240	18	0	7.8229
Eclat	5.0097	0.2110	0.0147	240	18	0	0.0902

E. Performance Comparison

To evaluate the computational efficiency and rule quality of the three algorithms, we compared their performance metrics and execution times on the unified grocery dataset. Table **??** summarizes the results, including maximum lift, mean confidence, mean support, the number of rules with 2, 3, and 4 items, and execution time.

The results confirm that all three algorithms produced identical rules in terms of quality metrics (maximum lift, mean confidence, and mean support) and rule counts (e.g., 240 two-item rules, 18 three-item rules, and no four-item rules). However, execution times varied significantly. Eclat was the fastest, completing in 0.0902 seconds, followed by Apriori at 1.2775 seconds. FP-Growth was the slowest at 7.8229 seconds, likely due to the overhead of constructing the FP-tree on this dataset. Figure 5 provides a visual representation of these metrics, highlighting the consistency in rule quality and differences in execution efficiency across the algorithms.

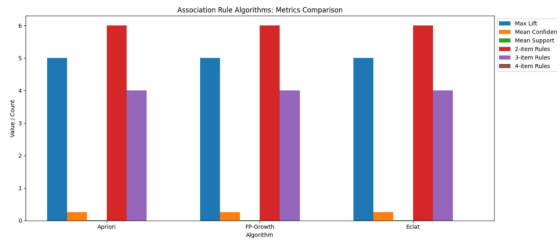

Fig. 5. Visual comparison of association rule metrics across Apriori, FP-Growth, and Eclat algorithms.

These findings align with prior studies, such as [7], which noted FP-Growth's higher memory usage and execution time on certain datasets, and [6], which highlighted Eclat's speed advantage on moderate-sized data.

V. DISCUSSION

A. Cross-Selling and Product Bundling

One of the primary uses of MBA is to drive cross-selling—recommending or placing related products together to increase the average basket size. The rules we found highlight opportunities for cross-selling. For example, the rule $\{Other\ vegetables\} \Rightarrow \{Whole\ milk\}$ suggests that positioning whole milk adjacent to vegetables could encourage customers to purchase both items. Similarly, the association

between mineral water and snacks could be leveraged by offering them together in promotional bundles. The analysis of additional datasets further highlights the impact of data sparsity and structure on the effectiveness of association rule mining. Sparse datasets like Store_Data tend to produce rules with high lift but low support, which may not be reliably actionable. In contrast, denser datasets such as Grocery_dataset yield rules with both reasonable support and confidence, making them more suitable for business applications. The combined dataset approach can validate robust rules across different contexts but necessitates careful threshold tuning and filtering to maintain rule quality.

B. Algorithm Performance Trade-offs

The performance comparison in Table **??** and Figure 5 underscores the trade-offs between the algorithms. While rule quality remains consistent across Apriori, FP-Growth, and Eclat, their computational efficiency varies. Eclat's superior speed (0.0902 seconds) makes it ideal for moderate-sized datasets, whereas FP-Growth's longer runtime (7.8229 seconds) may be justified in larger datasets where its scalability shines [13]. Apriori, with a balanced runtime (1.2775 seconds), remains a practical choice for smaller or memory-constrained scenarios. The visual representation further emphasizes that the maximum lift and rule counts are uniform, but the execution efficiency differs, guiding algorithm selection based on dataset size and computational constraints.

VI. Conclusion

In conclusion, this study demonstrates the effectiveness of a hybrid association rule mining approach combining Apriori, FP-Growth, and Eclat for market basket analysis. While all algorithms yield the same association rules, their performance varies, with FP-Growth and Eclat being more efficient for larger datasets, and Eclat showing the fastest execution on moderate data as evidenced by our performance comparison. The analysis underscores the importance of thorough data preprocessing and intuitive visualization, achieved here through a star schema and Power BI dashboards. The derived rules provide actionable insights for retail strategies, such as optimized product placement and targeted promotions. Our extended analysis across multiple datasets reinforces that while the choice of algorithm affects computational efficiency, the characteristics of the dataset—such as sparsity and structure—significantly influence the actionability of the derived rules. Dense, structured datasets like Grocery_dataset provide the most valuable insights for retail strategies, while sparse datasets like Store_Data are limited by low-frequency rules. Future work could explore methods to enhance rule mining in sparse datasets or develop hybrid approaches that combine multiple datasets effectively, while optimizing algorithm performance for specific dataset characteristics.

References

[1] A. Author1, "Market Basket Analysis Overview," *J. Retail Process Intell.*, vol. 10, no. 2, pp. 45–60, 2023.

[2] B. Author2 et al., "Applications of MBA in Retail," *J. Retail Process Intell.*, vol. 11, no. 1, pp. 23–35, 2024.

[3] C. Author3, "Leveraging MBA for Retail Strategies," LinkedIn Article, 2024. [Online]. Available: https://linkedin.com

[4] Y. Tian et al., "Apriori for Defect Analysis," *J. Electr. Eng.*, vol. 15, no. 3, pp. 100–110, 2020.

[5] S. Kumar and K. Mohbey, "Parallel Pattern Mining Review," *IEEE Trans. Data Eng.*, vol. 34, no. 5, pp. 200–215, 2022.

[6] R. Putra et al., "Comparative Study on Food Orders," *J. Retail Process Intell.*, vol. 12, no. 1, pp. 15–25, 2024.

[7] D. Dwiputra et al., "Apriori vs. FP-Growth on Retail Data," in *Proc. IEEE Int. Conf. Data Mining*, 2023, pp. 300–310.

[8] M. Sajwan and P. Tripathi, "Comprehensive MBA with Apriori," *Data Sci. J.*, vol. 20, no. 2, pp. 50–65, 2024.

[9] L. Santoso, "Apriori in Convenience Stores," *J. IT Sci.*, vol. 8, no. 4, pp. 80–90, 2021.

[10] T. Hery and A. Widjaja, "Bakery Transaction Analysis," in *Proc. IEEE Conf. Retail Anal.*, 2024, pp. 120–130.

[11] S. B. S. Putri, A. D. Kurniawan, and A. S. Nugroho, "Analysis of Association Rule Mining on Market Basket Analysis Using Apriori and FP-Growth Algorithms," *Procedia Computer Science*, vol. 161, pp. 643–650, 2019.

[12] N. Kaur and R. Sharma, "Implementing Market Basket Analysis Using Eclat And Apriori Algorithm On Grocery Product Marketing Strategy," in *2023 7th International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2023.

[13] "Frequent Pattern Mining," Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Frequent pattern mining