Exploring Alcohol Dataset Using MapReduce, Aprioi And SON Algorithms In A Distributed Environment

Divyansh Anand
School of Computing
National College of Ireland
Dublin, Ireland
x22240217@student.ncirl.ie

***Abstract*—This research aims to analyze alcohol sales dataset by finding hidden trends, correlations and optimization techniques to handle inventory efficiently. Data is processed through various stages including data ingestion, data cleaning, data processing and data visualization to get the complete understanding of the dataset and techniques. This research briefly explains working of Map Reduce, Apriori, SON and Pearson correlation algorithms. This research explored seasonal variation in sales of alcohol, most preferred brands of alcohol, correlations between sales performance and factors such as brand and excise tax rate. This study finds the frequent item sets sold in the shop using Apriori and SON algorithms and explains their performance in detail. Results show SON algorithm is fast and efficient in handling huge amount of data in a distributed environment.**

***Index Terms*—Hadoop, Map Reduce, Apriori, SON, Pearson Correlation, Alcohol Sales Dataset**

I. Introduction

Retail industries processes billions of transactions on daily basis which help them earn profits to manage inventory and staff. Sales for any retail shop differs from time to time, understanding the trend can help manage the inventory more effectively. Data analysis plays a crucial role for understanding insights from large datasets. This research demonstrates a scalable system which is designed to handle alcohol sales data to identify trends, correlations and optimization opportunities. The proposed scalable system, processes and analyses data in several functions which include data ingestion, data preprocessing for ensuring data quality and data analysis based on the research questions. Later, data visualization libraries are used to make the results more presentable. To efficiently handle the large alcohol sales dataset with more than a million records this research used Map Reduce functionality within the Hadoop framework. Additionally, Apriori and SON algorithms have been applied using Map Reduce to get the frequent items bought together. With these algorithms, one can analyze the patterns which can help the companies manage their inventory in the most optimised way which can definitely lead to increase in sales for the companies. This research also highlight and compare the performance of Aprioi and SON algorithms which can help understand their advantages and disadvantages in handling the dataset.

Identify applicable funding agency here. If none, delete this.

Research questions:
1. How does distributed framework influence efficiency and scalability of Apriori and SON algorithms for mining frequent item-sets to optimize the inventory?
2. How do sales of different alcoholic beverage categories vary across different months or seasons?
3. Are there any observable correlations between sales performance and factors such as product size, brand, price point or excise tax rate?
4. Which specific products or brands within each major category (beer, wine, spirits) are the top sellers based on sales quantity, revenue, or other relevant metrics?

II. Related Work

A. Maintaining the Integrity of the Specifications

This literature review section concentrates on studies that use scalable algorithms and data stream processing to analyze big datasets. The focus will be on contrasting and comparing various methods in order to show their strengths and weaknesses for processing huge datasets in scalable and distributed environments.

When one item is paired up with another item frequently then a regular pattern in shopping can be observed. According to study by [1] states regular pattern mining is very beneficial in domains such as extracting knowledge from transactional data for market basket analysis. Since the beginning of mining, several algorithms have been created and improved to extract knowledge from transactional data. Unfortunately, with data growing at tremendous rate day by day, handling such huge datasets can be expensive in terms of storage, processing and analyzing patterns.

Research by [2] covers challenges of analyzing Big Data stream. Some of the major challenges which can help in our study are scalability, fault tolerance, timeliness and accuracy. Scalable frameworks and algorithms are needed for processing a large dataset in Hadoop environment in order to effectively manage the massive amount of data without exhausting computational resources. In a distributed environment such

as Hadoop, robust fault tolerance methods are important to guarantee continuous processing even in case of any component failures or any node failure. This challenge has been handled in the proposed research by creating multiple nodes like Name node, secondary name nodes, etc. Processing data fast is essential, especially in real time scenarios. Timeliness can be maximized by implementing scalable architectures and platforms which is essentially considered in our proposed research. For reliable results ensuring high accuracy is must, given the complexity and size of the proposed alcohol dataset, the researcher has considered stream specific requirements and implemented techniques to achieve accurate results.

Once we overcome the challenges of big data processing, considering the approach for data analysis is very important. The research by [3] highlights core functionalities and various implementations of MapReduce programming model. Map Reduce functionality breaks down large scale data processing task into smaller and manageable subtasks. These subtasks are then distributed and processed concurrently across clusters of computers, which increases the speed of data analysis. Additionally, the article mentions Hadoop as a widely used open-source implementation of MapReduce. The researcher has used map reduce for the analysis as it offers a powerful approach to handle massive dataset efficiently in a distributed environment. Framework to work with Map reduce programming and to handle large data processing is an important choice to make while doing the analysis. This research [4] on comparing Apache Spark and Hadoop with map reduce is very insightful. They have used four datasets of different sizes to simulate different workloads. They have compared the frameworks based on execution time, accuracy and scalability. This study founds out, Spark is significantly faster than Hadoop during training phase. Spark's ability to scale well with more processing nodes in a cluster is responsible for this speedup. However, Spark's performance degrades as the data volume increases. On the other hand, Hadoop maintains consistent performance regardless of workload size. For accuracy, Hadoop provides slightly more accuracy for classification than Spark. Proposed research has used Hadoop framework with Map reduce to process alcohol sales dataset which provides consistent accuracy regardless of workload size.

This study [5] shares details regarding Apriori alogorithm focuses on its application in mining frequent itemsets and association rules. Apriori algorithm is a useful tool for mining association rules, which can help in making marketing strategies and decision making process. This research discusses the development of the Apriori algorithm and its variants both in China and abroad. The growth of research in this field is illustrated with a focus on algorithm's efficiency and applications in different sectors. The fundamental concept of Apriori algorithm is to effeciently mine association rules from transactional databases. Relationship between elements in a database are described by association rules which shows which items appear together frequently. Apriori algorithms aims to find frequent itemsets or grouping of items that occur together frequently. There are few adantages of apriori such as it's easy to understand and implement. It works well on datasets with low density of frequent items. Pruning mechanism of apriori algorithm makes it more effecient and it's widely used and applied to different industries. However, there are few drawbacks of using this algorithm. It requires large amount of memory especially for datasets with a large number of unique transactions. It is time consuming as it scans the database multiple times.

Considering drawbacks of Apriori algorithm, Son algorithm has better performance than Apriori algorithm, as per study in [6] SON algorithm is designed to reduce input/output overhead by dividing the dataset into non-overlapping partitions and processing them separately. It uses a two step process to identify frequent itemsets. It is more scalable than Apriori algorithm for large datasets, as it can handle huge datasets by parallelizing the mining process. It only requires two scans of the dataset, making it more efficient in terms of consuming resources. The SON algorithm is inherently parallelizable, making it well-suited for distributed computing frameworks like MapReduce.

Overall, in this proposed research, huge dataset of alcohol sales will be used to test the performance of Apriori and SON algorithms. Map reduce and hadoop framework will be used for analysis to keep the system distributed and more efficient. Further sections will discuss the Methodology, Results, Conclusion and Future work for this study.

III. Methodology

A. Dataset Details

Alcohol sales dataset is a public dataset, sourced from Kaggle website. [7] This dataset has 14 columns with more than a million records. Summary of columns is given in below fig1.

```
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 14 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   InventoryId     1048575 non-null  object
 1   Store           1048575 non-null  int64
 2   Brand           1048575 non-null  int64
 3   Description     1048575 non-null  object
 4   Size            1048575 non-null  object
 5   SalesQuantity   1048575 non-null  int64
 6   SalesDollars    1048575 non-null  float64
 7   SalesPrice      1048575 non-null  float64
 8   SalesDate       1048575 non-null  object
 9   Volume          1048575 non-null  int64
 10  Classification  1048575 non-null  int64
 11  ExciseTax       1048575 non-null  float64
 12  VendorNo        1048575 non-null  int64
 13  VendorName      1048575 non-null  object
dtypes: float64(3), int64(6), object(5)
memory usage: 112.0+ MB
```

Fig. 1. Column Details

B. Workflow Diagram

The workflow diagram given below in fig2 outlines the steps involved in processing, from data intake to deriving actionable insights from the analysis. Each step is designed to achieve the research objective. 1. Alcohol dataset is the primary dataset

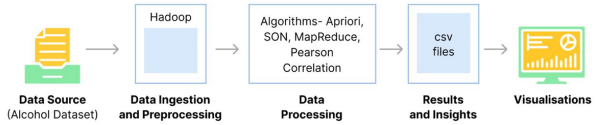

Fig. 2. Workflow Diagram

used for this research which contains information about sales of different alcoholic beverages.
2. Data Ingestion and preprocessing phase involves tasks such as data cleaning checking for null values and formating the columns. This phase has handled the columns having different date formats, in the raw file, some columns have dates in the format 'mm-dd-yyyy' while others are in 'dd/mm/yyyy' format. Pre-processing script snapshot is shared below in fig3.

```
hduser@Divyansh:~/project$ python3 preprocessingScript.py
No null values found in the dataset
SalesDate column has been handled
New CSV file has been Created with Correct SalesDateValues /home/hduser/project/
cleaned_data.csv
```

Fig. 3. Pre-processing Script Snapshot

3. Data Processing phase used different algorithms like Apriori, SON, MapReduce and Pearson Correlation to answer respective research questions.
4. All the results from above data processing are stored in different csv files, using python libraries different visualizations are produced to understand the insights.

C. Algorithms Used For Analysis

Below are the details of algorithms used in this research with details regarding how it processed the data.

1. Apriori Algorithm: The overview of the Apriori algorithm for this research has been highlighted below. It starts by finding individual items that meet a minimum support threshold. These items are considered frequent itemsets of size 1. It then iteratively generates candidate itemsets of larger size based on the frequent itemsets based on itemsets found in previous iteration. This process keeps repeating until no new frequent itemsets can be generated. The algorithm uses the property of apriori which says that any subset of the frequent itemset must also be frequent to generate itemsets of larger size. This feature prunes the search by avoiding the generation of candidate itemsets that contain infrequent subsets. For every candidate itemset, the algorithm goes in the entire dataset to count the number of transactions that contain the itemset. The count is then used against the minimum support threshold to indicate whether the itemset is frequent or not. Some association rules are applied in the algorithm to enhance the analysis.

2. SON(Savasere, Omiecinski, and Navathe) Algorithm: The overview of the SON algorithm for this research has been discussed below. In the first phase, data gets divided into chunks which are processed independently by different mappers. Each mapper processes a subset of the data. Each mapper uses a subset of data to generate local frequent itemset which used apriori algorithm's core fundamental. Following local processing, a reducer receives the candidate frequent itemsets from each mapper. In order to determine which itemsets are actually frequent throughout the dataset, the reducer aggregates the candidate frequent itemsets from each mapper and founds a global support count. The gloablly frequent itemsets are emitted as the final output of this algorithm.

3. Map Reduce Algorithm: It can be broken down into three phases namely Mapper, Shuffle - Sort and Reducer phase. In this research, during Map phase input data was divided into small chunks and each chuck was processed independently by multiple mapper tasks. Each mapper task applies a mapper function to the data and generates intermediate key-value pairs. In Shuffle and Sort phase, the intermediate key-values generated by mapper tasks are sorted based on their keys and partitioned across the reducer tasks. All key-value pairs with the same key are grouped together, ensuring that each reducer task receives all the values associated with a particular key. In reducer phase, each reducer processes a subset of key-value pairs. It applies reducer function to each group of values associated with the same key, giving final results.

4. Pearson Correlation Algorithm: To calculate Pearson correlation mathematically, formula is -

$$r = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{\sqrt{(n \cdot \sum x^2 - (\sum x)^2) \cdot (n \cdot \sum y^2 - (\sum y)^2)}}$$

In this formula, n denotes number of records, $\sum xy$ is the sum of the product of paired values, $\sum x$ and $\sum y$ are the sums of each column, and $\sum x^2$ and $\sum y^2$ are the sums of squares of each column. In this research, map reduce functions are used to calculate Pearson correlation coefficients for pairs of columns by distributing the computations across multiple mappers and reducers which makes it suitable for analyzing big data. Mapper phase breaks down sales records and create key-value pairs. It uses four columns namely brand, excise tax, sales dollar, sales quantity. Key idetifies a column pair for example 0-1 for Sales dollar vs excise tax and values hold the actual data points. Then a combiner, combines values for each key pairs (Sales dollar vs excise tax). It calculates sums and products needed for correlation then sends the out put to reducer phase. Reducer phase takes pre-processed data and computes Pearson correlation coefficient for each column pair. It gives key-value pair as output where key is column

pair identifier and value is correlation coefficient.

IV. Results

This section will discuss about the observations of data analysis in Hadoop.

A. Q1. How does distributed framework influence efficiency and scalability of Apriori and SON algorithms for mining frequent item-sets to optimize the inventory?

Output of Apriori Algorithm process is given below in fig4:

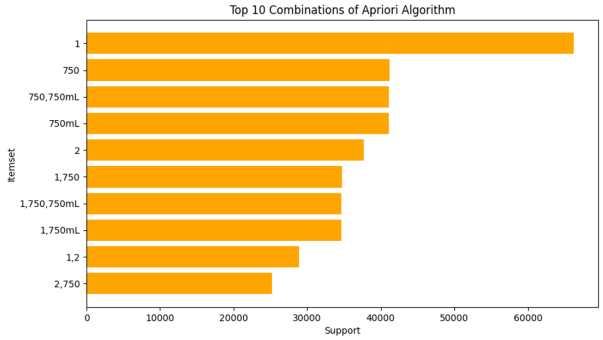

Fig. 4. Top 10 Item sets of Apriori

Observations from the output of algorithms are identical which is represented in fig4 and fig5-
1. Majority of the people prefer buying a single bottle(quantity) of alcohol as compared to two bottles.
2. The most preferred size of bottle is 750 ml(single quantity).
3. Fewer people prefer buying two quantities of bottles with 750 ml of volume.

Output of SON Algorithm process is given below in fig5:

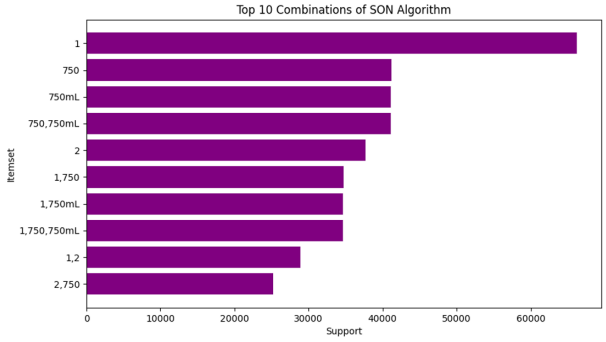

Fig. 5. Top 10 Item sets of SON

Based on the outputs, inventory can be aligned properly according to customer demand.

The most popular choice is single units of alcohol, stocking sufficient single units can keep up with the customer demands. Storage size should be considered as most optimal choice is 750 ml bottle size. Increasing the space for bigger bottles can help meet the customer needs. Limited stocking of small multi-pack size alcohol can help save the valuable storage space. These considerations for inventory management can help increase the sales of the stores.

Performance of both the algorithms are discussed below-

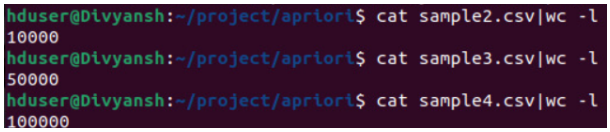

Fig. 6. Sample file size

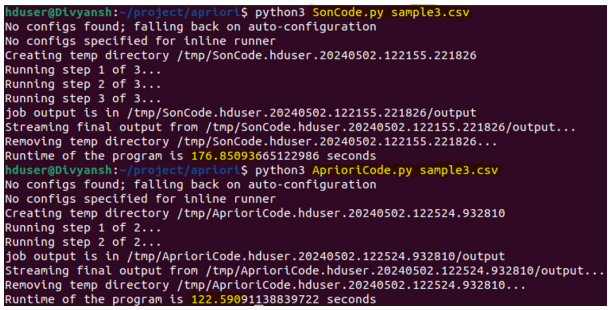

Fig. 7. Processing Time for 50,000 records

1. Processing Time: Processing time of Apriori and SON varies with dataset size. From fig6 and fig7, it is clear that Apriori algorithm took 132.48 seconds which is longer time than SON 127.87 seconds for 50,000 records. Both the codes use Map reduce functionality to do the analysis, Apriori uses 2 steps, in step 1 it generates frequent item sets and in step 2, it finds association rules. However, SON algorithm uses 3 steps one for generating local frequent item sets, second for collecting item sets and last step to find global frequent item sets.
2. Scalability: The scalability of Apriori algorithm is limited as it scans the dataset multiple times to generate candidate keys. With huge dataset size, the complexity increases due to which it's processing time increases. SON algorithm is designed to handle large datasets efficiently. Due to its distributed global aggregation and chunk size data processing, it works better than Apriori algorithm.
3. Resource Utilization: As compared to SON, Apriori algorithm uses significant memory to store candidate item sets and support counts for input data. Apriori algorithm could not handle dataset with records greater than 1 lac in the virtual machine while SON algorithm could handle it.
4. Algorithm complexity: Apriori has a relatively simple implementation but it becomes computationally expensive when running for large dataset. SON is a bit complex provided multiple steps are involved in local and global aggregation. The chunk size and minimum support for the

SON algorithm in this research has been kept 0.01 and 2500 respectively. For Apriori the minimum support values is same as SON, 0.01.

5. Output Accuracy: Both the algorithms performed well in terms of accuracy, output records are identical for both the algorithms, resulting in same size and count for the csv files as shown in below fig8.

```
hduser@Divyansh:~/project/apriori$ cat SonOutput.csv |wc -l
319228
hduser@Divyansh:~/project/apriori$ cat aprioriOutput.csv |wc -l
319228
```

Fig. 8. Enter Caption

Overall, both algorithms have their advantages and limitations while processing the data. For this research, SON performed better that Apriori algorithm for finding the frequent item sets from alcohol sales dataset.

B. How do sales of different alcoholic beverage categories (beer, wine, spirits) vary across different months, can we identify any seasonal patterns or peaks in demand for certain product types

As observed in below fig9, the graph shows seasonal pattern in sales for all three categories of alcohol namely beer, wine and spirits. Wine and Spirit sales peak during winter season, which shows consumption of alcohol increases during winter season and winter holidays.

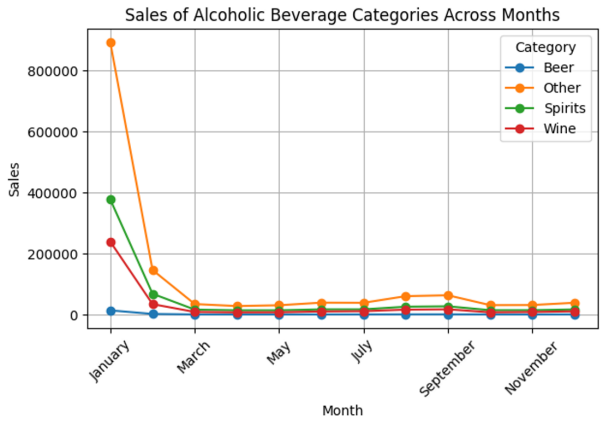

Fig. 9. Line Graph of Alcohol Sales throughout the months

C. Are there any observable correlations between sales performance and factors such as sales quantity, brand, price point, or excise tax rate?

Observations:

1. From the fig10, one can observe that there is extremely strong positive correlation between Brand and SalesDollar of around 0.98 which depicts, when Brand value increases the Sales tend to increase as well.

2. Another observation, Brand vs Excise tax shows a strong positive correlation with value around 0.99, which means when excise tax of any alcohol increases the sales also increases.

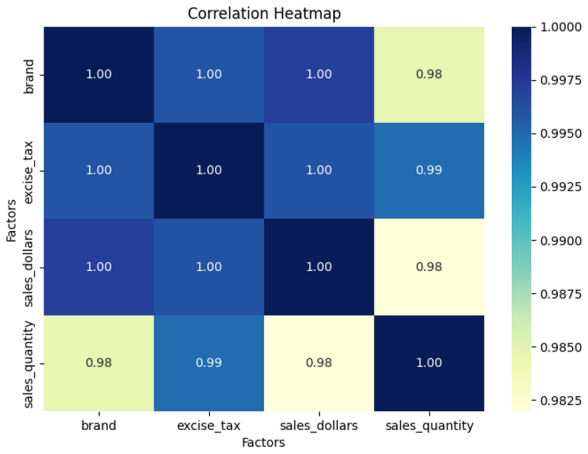

Fig. 10. HeatMap of Sales Performance Vs factors

D. Which specific products or brands within each major category (beer, wine, spirits) are the top sellers based on sales quantity, revenue, or other relevant metrics??

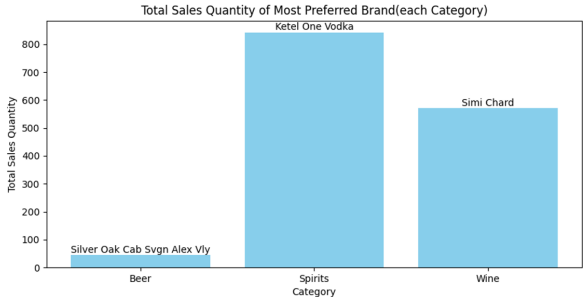

Fig. 11. Histogram of Sales Quantity of most preferred brands

Observations:

1. The above fig11 represents total sales quantity of best selling brand in each category of alcohol namely Beer, Spirits and Wine. Spirit with Brand "Ketel One Vodka" has the highest sales quantity with 843 quantities sold, followed by "Simi Chard" in Wine with 570.

2.Silver Oak Cab Savgn Alex Vly has the most preference in Beer category.

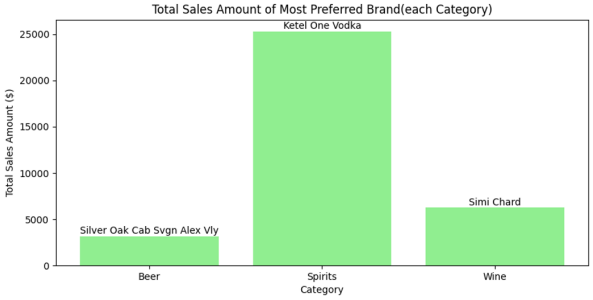

Fig. 12. Histogram of Sales Dollar of most preferred brands

Observations:
1. Fig12 represents total sales of most preferred brands in each category. For the Spirits category, the most preferred brand "Ketel One Vodka" has the highest sales of more than 25000 Dollars.
2. SimiChard Brand from wine category earns slightly above 5000 dollars whereas Silver Oak Cab Savgn Alex Vly earns total sales of more than 3000 dollars.

V. Conclusion and Future work

This research was designed to carry out a detail examination of alcohol dataset for the purpose of revealing trends, correlations and algorithms that are crucial for optimizing inventory management and can increase sales. The findings of this research can offer better understanding of customer preferences. The availability of data processing capabilities and frameworks allowed this research to utilize scalable algorithms and distributed computing concepts through Hadoop for analyzing large dataset and extract meaningful insights.
The comparison between Apriori and SON algorithms suggest that although both are good at mining frequent datasets, SON performs better in terms of scaling and utilization of resources.
Moreover, the analysis of consumption patterns revealed that Winter season was a season for wine, spirits, and beer. For inventory optimization purposes, retailers can understand customer preference. From the analysis, research represents that most of the customer preferred single unit of alcohol with 750 ml capacity. Overall, this investigation is significant as it highlights information regarding inventory, trends, customer behaviour in retail industry.

While data processing and scalable algorithms like Map Reduce, SON and Aprioi shows effective performance in analyzing big data, there's always room for improvement. To compare their performance, Apriori and SON algorithms have been tested with 50k records for this research due to lack of storage and limited computing capacity of the system. This can be improved with better machines, which can take this research at a higher level. Understanding these algorithms are complex, more tuning and checks can be added for better performance. More advance and efficient predictive models can be used to analyze the alcohol dataset to uncover more hidden insights.

References

[1] M. A. Hasan, N. Hassan, M. Hasibuzzaman and M. R. Huq, "A Scalable Approach for Improving Implementation of a Frequent Pattern Mining Algorithm using MapReduce Programming," 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 2019, pp. 106-111, doi: 10.1109/ICSITech46713.2019.8987446. keywords: Frequent Itemset Mining;MapReduce;Hadoop;CATS-Tree.

[2] Kolajo, Taiwo Daramola, Olawande Adebiyi, Ayodele. (2019). Big data stream analysis: a systematic literature review. Journal of Big Data. 6. 47. 10.1186/s40537-019-0210-7.

[3] R, Bharath. (2020). Map Reduce: Data Processing on large clusters, Applications and Implementations. 05. 214-220.

[4] Tekdoğan, Taha Cakmak, Ali. (2021). Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification. 15-20. 10.1145/3481646.3481649.

[5] Xie, Haoyu. (2021). Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets. Open Journal of Social Sciences. 09. 458-468. 10.4236/jss.2021.94034.

[6] T. Xiao, C. Yuan and Y. Huang, "PSON: A Parallelized SON Algorithm with MapReduce for Mining Frequent Sets," 2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming, Tianjin, China, 2011, pp. 252-257, doi: 10.1109/PAAP.2011.38. keywords: Itemsets;Partitioning algorithms;Algorithm design and analysis;Data mining;Distributed databases;frequent sets mining;parallelized SON algorithm;MapReduce;Hadoop,

[7] Inventory Analysis Case Study . (2023, July 13). Kaggle. https://www.kaggle.com/datasets/bhanupratapbiswas/inventory-analysis-case-study/data?select=SalesFINAL12312016.csv