Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 230 (2023) 455–466

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)

Data Analytics for Pandemic Management using MapReduce and Apriori Algorithm

Shashwat Kumar[a], Anannya Chuli[a], Aditi Jain[a], Narayanan Prasanth[a],*

[a] *Vellore Institute of Technology, Vellore, 632014, India*
shashwat.kumar2@gmail.com

Abstract

Amidst the profound impact of the COVID-19 pandemic on global economies and healthcare systems, effective data analysis has become paramount. Our research paper, titled ”Data Analytics for Pandemic Management Using MapReduce and Apriori Algorithm,” presents a comprehensive framework to analyze pandemic data. We harness the power of the MapReduce and Apriori algorithms, with a parallel processing model achieving an average speedup of 50%. This approach involves data collection, preprocessing, and algorithmic application to extract valuable insights from pandemic-related data. Notably, our findings reveal a substantial 22.29% support rate for ”n95 masks,” indicating high demand. Additionally, we identify strong co-occurrence patterns, exemplified by the perfect support rate of 1.00 between ”n95 masks” and ”chloroquine,” highlighting their interconnectedness. Our framework goes beyond data analysis, enhancing personalized marketing, and optimizing inventory management for efficient resource allocation during crises. We also uncover robust associations, such as the 0.220 confidence in joint purchases of ”butyl rubber gloves” and ”surgical masks”.Hence, the research aligns seamlessly with the pressing need for effective pandemic data management and response strategies. By validating our approach through numerical insights, we aim to contribute to mitigating pandemic-related challenges and support global efforts to better prepare for and manage future health emergencies.

Keywords: Data Mining; Parallelization Techniques; MapReduce; Apriori; Pandemic Control.

1. Introduction

In the aftermath of the COVID-19 pandemic, the world has come to understand a critical lesson: the relationship between data, public health, and the global economy is more important than ever. This research sets out on a mission that transcends academic boundaries; it aims to predict the demand for and efficient use of vital medical supplies during pandemics. Its the ultimate goal is to equip healthcare organizations and authorities with the foresight required to prepare for unexpected disasters while preventing a chain reaction of problems that can threaten our world.

10.1016/j.procs.2023.12.101

Our research serves as a wake-up call to nations worldwide, addressing a pressing concern. Imagine a world where not only lives but also the global economy are hanging in the balance. During COVID-19, nations faced sudden surges in demand for medical resources, leading to unforeseen consequences. Economic hardships resulted in shortages of medical supplies, intensifying pressure on healthcare systems and causing unexpected deaths. This reality underscores the profound impact of unpreparedness, where its repercussions ripple across borders. At the core of our research lie two essential tools: MapReduce and Apriori algorithms. They are the keys to revolutionizing pandemic management. MapReduce enables efficient analysis of vast datasets, allowing us to scrutinize diverse COVID-19 data that holds the key to averting economic turmoil. Simultaneously, the Apriori algorithm, known for identifying patterns within data, reveals complex relationships between demographic attributes and virus spread. Together, they provide a deeper understanding of the virus's dynamics and offer insights that can help steer us away from economic crises.

Our research methodology is grounded in meticulous data cleaning and preprocessing, ensuring data precision and integrity. MapReduce's ability to uncover intricate patterns and trends in large datasets is invaluable for our analysis. Mean- while, the Apriori algorithm helps us identify frequent patterns and relationships, guiding us in forming targeted intervention strategies. Together, these tools create a solid foundation for informed policy-making, effective interventions, and the strategic direction of public health efforts. This comprehensive approach not only analyzes data but also acts as a lifeline for economies and lives. The motivation behind our research stems from a combination of urgency and responsibility. COVID-19 laid bare vulnerabilities within healthcare systems and exposed the fragility of the global economy. It highlighted the crucial need for predictive analytics to mitigate the impact of future crises. Our research aspires to empower healthcare professionals, governments, and businesses with vital information for informed decision-making during evolving health emergencies. It seeks to provide clarity, strengthen global pandemic responses, and establish a framework for enhanced health emergency measures. Moreover, this study delves into the potential real-world applications of the findings, emphasizing their relevance in shaping public health policies and interventions. The insights garnered from this research have the potential to drive evidence-based decision-making, enabling healthcare authorities to develop more effective strategies for disease surveillance, containment, and response.

2. Related Work

The COVID-19 pandemic has prompted a surge in research efforts, specifically in data analysis and the application of artificial intelligence within healthcare. While these technologies have found applications in various domains, the use of established techniques like MapReduce and the Apriori algorithm in COVID-19 data analysis is a relatively unexplored area. This literature survey delves into earlier works to understand the rationale for employing these techniques in the context of COVID-19. Farzana Shaikh et al. [1] initiated the discussion by advocating the use of Hadoop's MapReduce for exploring YouTube's extensive repository of unstructured data. Their primary aim was to empower content creators by providing insights into competitors and content optimization, highlighting the value of data-driven insights in the digital landscape. However, challenges in handling unstructured data and the need for expert system configuration raised concerns. Hari Singh and Seema Baw [2] explored spatial data processing, comparing conventional sequential methods with the parallel capabilities of MapReduce, supported by ArcGIS. They demonstrated improved efficiency and scalability in spatial data processing but did not address ArcGIS limitations or the complexities of managing extensive geographic data queries. Hao Yang [3] shifted the focus to a cloud-based framework for substantial data analysis in human resources (HR). Using Hadoop, the framework addressed HR data management limitations, emphasizing dynamic results' integrity. However, it did not thoroughly examine security concerns in cloud systems or practical implementation costs. Suryanarayana et al. [4] presented an environmental data analysis case study, employing Hadoop and MapReduce to mine insights from extensive weather datasets. Whilethis approach demonstrated efficiency gains, it highlighted the need for technical expertise, particularly regarding sensor-based weather data. PrathyushaRani Merla and Yiheng Liang [5] entered the realm of social media data anal-ysis, using Hadoop's MapReduce to identify top YouTube categories, uploaders, and videos. Their work aided trend comprehension and decision-making but had limitations, such as the narrow scope of data exploration and challenges in establishing a Hadoop environment in a cloud platform. Ashish et al. [7] introduced the MR-DA methodology, focusing on enhancing K-Means clustering efficiency for large datasets. While it outperformed other algorithms, it lacked a comprehensive real-world application assessment and detailed exploration of its limitations. Rajdeep Paul [8] explored sports analytics, analyzing Indian Premier League Big Data using Hadoop and MapReduce. The study provided insights into the league's global appeal but left some aspects unaddressed, such as dataset size and comprehensive discussions of limitations. Xiaojing Zhu [13] integrated MapReduce with genetic algorithms for distributed social network data processing, emphasizing the power of MapReduce in distributing social network data. How- ever, the work lacked a comprehensive evaluation against alternative methods and raised ethical concerns related to user data usage for service optimization. These studies collectively provide valuable insights into the application of MapReduce and Hadoop in various domains, emphasizing their benefits and limitations. Further research is needed to explore these techniques in the context of the ongoing COVID-19 pandemic.

The above-existing research exhibits several limitations -

1. Challenges in data handling, technical expertise, and infrastructure costs.
2. Overlooked limitations of tools like ArcGIS for large geographic data queries and sensor-based weather data.
3. Exclusive focus on certain aspects of the topics, restricting comprehensive analysis.

However, in order to yield better and more satisfactory results in Big Data Analysis, extensive research is always being conducted. Our proposed methodology seamlessly integrates MapReduce and Apriori algorithms to address these limitations. This combination is expected to harness MapReduce's scalability and fault tolerance, effectively handling large datasets. By dividing tasks into parallel activities across a cluster, it is expected to ensure uninterrupted analysis, overcoming individual shortcomings. Our approach is designed to greatly enhance data analysis, allowing for detailed insights and patterns from complex datasets that were previously challenging for independent processing.

3. MapReduce and Apriori Algorithms in Data Analytics

In the realm of data analytics, the integration of advanced algorithms and frameworks has become pivotal in unraveling valuable insights from extensive and intricate datasets. Two such indispensable components are the MapReduce framework and the Apriori algorithm, each distinguished by its unique capabilities and contributions to the field.

3.1. MapReduce in Big Data Analytics

MapReduce, first brought to the forefront by its successful application at Google, has established itself as a fundamental data processing framework for the parallel, batch-style analysis of colossal datasets. Its inherent characteristics, including simplicity, scalability, and fault tolerance, have made it a ubiquitous choice in both industrial and academic circles. The essence of its superior performance lies in its ability to partition data processing into manageable, parallelizable units of work that can be effortlessly distributed across multiple nodes within a computing cluster.

Advantages of MapReduce in Big Data Analytics:

1. Scalability and Fault Tolerance: MapReduce serves as a scalable and fault-tolerant data processing tool. It empowers the parallel processing of vast volumes of data, efficiently distributing workloads across numerous low-end computing nodes. This scalability is particularly advantageous when dealing with extensive datasets, and its fault tolerance ensures the continuity of data processing even in the event of node failures, ensuring the reliability of the analysis.
2. Simplicity and High Performance: MapReduce is lauded for its straightforward approach to complex data processing tasks. Breaking down processing into smaller units of work, it enables parallel execution across multiple nodes, resulting in elevated performance and reduced execution times, especially in the context of large datasets. Consequently, MapReduce has garnered significant traction in both industry and academia.

3.2. MapReduce With Unstructured Data and Real-Time Processing

It is important to note that MapReduce is not a standalone tool but is instead a foundational element of the Apache Hadoop framework. Hadoop is instrumental in managing unstructured and extensive datasets across clusters of commodity computers. Each node within these clusters is equipped with its own storage capacity, collectively contributing to the efficient and parallel processing of large-scale data. This architectural design empowers effective data analysis and manipulation, making it suitable for a wide range of data types, including unstructured data.

While MapReduce excels in batch processing of large data volumes, it may not be the ideal choice for real-time and online processing requirements. The framework is inherently designed for high-throughput batch processing that spans hours or even days, whereas contemporary demands often revolve around tasks and queries that need to be completed within seconds or minutes. Recognizing this discrepancy, new systems and frameworks have been proposed to cater to the evolving demands of real-time and online processing.

3.3. The Apriori Algorithm and Its Efficiency in Association Rule Mining

The Apriori algorithm plays a significant role in association rule mining, a technique used to uncover underlying relationships between various items within a dataset. This unsupervised algorithm employs a crucial parameter called" minimumsupport" to extract association rules. Support, in this context, indicates the frequency with which a specificitem appears in the dataset, allowing the algorithm to efficiently handle extensive datasets and discover meaningful associations.

Advantages of an Optimized Apriori Algorithm:

One notable advantage of utilizing an optimized version of the Apriori algorithm is its efficiency in terms of execution time and memory usage. The standard Apriori algorithm may become computationally demanding, particularly when dealing with datasets that contain a substantial number of items. Optimized versions of the Apriori algorithm implement various techniques such as pruning, hashing, and improved data structures to mitigate the generation of an excessive number of candidates itemsets. This optimization leads to faster and more practical execution, significantly reducing computational overhead and rendering the algorithm feasible for larger and more complex datasets.

4. Proposed Approach and Methodology

In the midst of the ongoing COVID-19 pandemic, the need for effective strategies and data-driven insights to inform public health policies has become paramount. This paper introduces an innovative model for COVID-19 data analysis that takes an integrated approach, combining the robust capabilities of the MapReduce and Apriori algorithms. This architecture not only provides a comprehensive framework for a deeper understanding of the virus's transmission dynamics but also empowers informed decision-making processes.

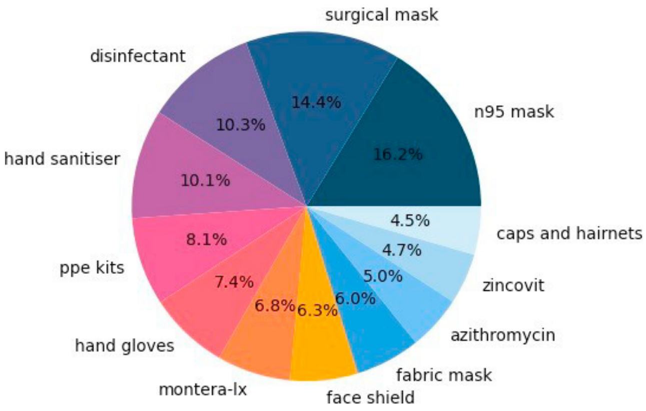


Fig. 1. Covid Optimization dataset

To achieve this, the study leverages the "Covid Optimization Dataset" on Kaggle, a transactional dataset commonly used for COVID-19 analysis. This dataset focuses on item relationships during the pandemic, including pharmacy store sales data and daily usage of pharmaceutical products and COVID-19 safety equipment. To ensure the dataset's

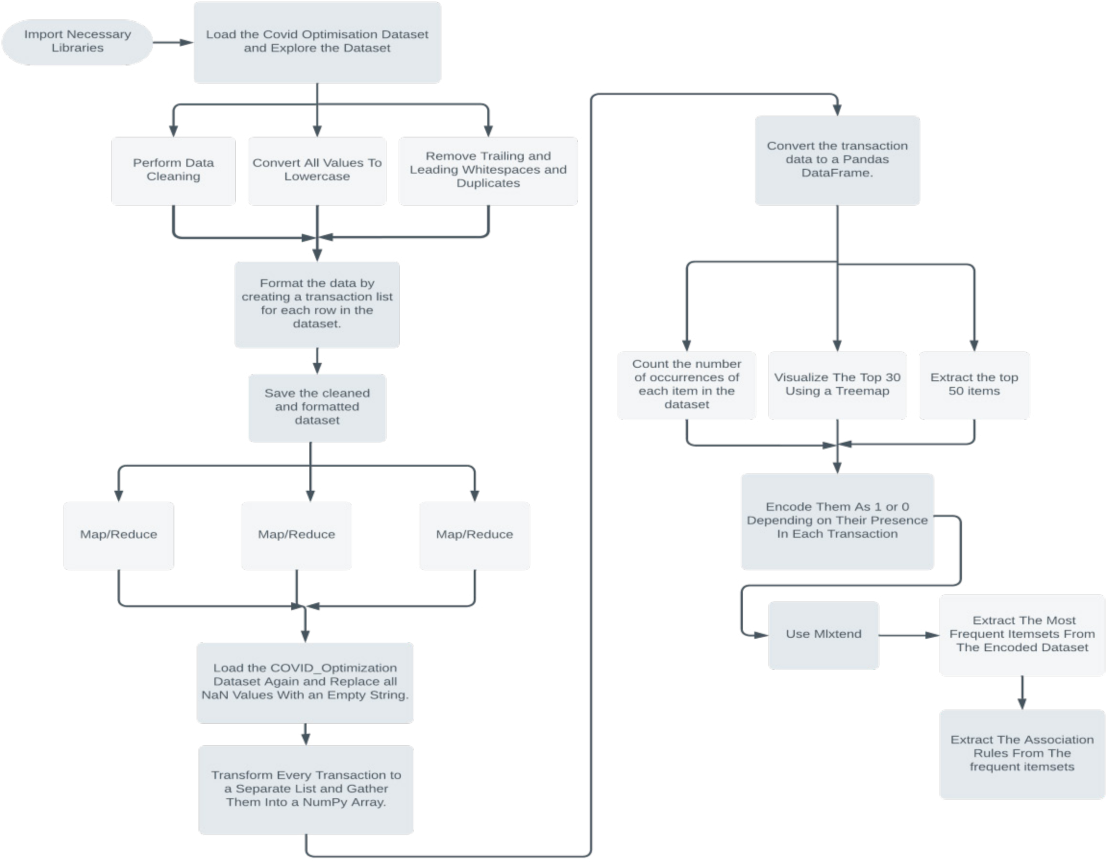

Import Necessary Libraries
Load the Covid Optimisation Dataset and Explore the Dataset
Perform Data Cleaning
Convert All Values To Lowercase
Remove Trailing and Leading Whitespaces and Duplicates
Format the data by creating a transaction list for each row in the dataset.
Save the cleaned and formatted dataset
Map/Reduce
Map/Reduce
Map/Reduce
Load the COVID_Optimization Dataset Again and Replace all NaN Values With an Empty String.
Transform Every Transaction to a Separate List and Gather Them Into a NumPy Array.
Convert the transaction data to a Pandas DataFrame.
Count the number of occurrences of each item in the dataset
Visualize The Top 30 Using a Treemap
Extract the top 50 items
Encode Them As 1 or 0 Depending on Their Presence In Each Transaction
Use Mlxtend
Extract The Most Frequent Itemsets From The Encoded Dataset
Extract The Association Rules From The frequent itemsets

Fig. 2. Proposed Workflow

suitability, it is required to be complete, accurate, and representative, with sufficient transactions to extract meaningful patterns. Moreover, the study incorporates a wide array of COVID-19 data sources, including medical records and demographic/geographical data from reputable repositories and government agencies.

4.1. Proposed Workflow

Our multi-tiered architecture meticulously encompasses various stages, each contributing distinct functionalities to the overarching data analysis endeavor. In order to ensure the academic rigor and practical viability of our model, we undertook a thorough revision of our initial proposal.The foundational tier, the Data Source Tier, assumes a central role by aggregating COVID-19 data from diverse sources. Ranging from health agencies and hospitals to laboratories and research institutions, this tier caters to the spectrum of data types, accommodating formats such as CSV, JSON, or XML. Emphasizing data integrity and flexibility, this tier ensures a robust initial dataset for analysis. The Data Preprocessing Tier engages dedicated modules to scrub, transform, and preprocess the collected data, utilizing big data processing tools like Hadoop, Spark, or Hive to maintain data quality and relevance for subsequent analysis phases.

Incorporating advanced methodologies, the MapReduce Processing Tier is a cornerstone of our architecture. This tier orchestrates a sophisticated processing module, comprising sub-modules responsible for data partitioning, mapping, shuffling, and reduction. The orchestrated pipeline diligently uncovers intricate patterns and trends embedded within the COVID-19 data, furnishing a comprehensive analysis. Continuing the analytical journey, the Apriori Algorithm Tier integrates an adept module dedicated to processing the outcomes of the MapReduce phase. Beyond the initial description, this tier features sub-modules that drive in-depth data analysis, robust pattern mining, and the generation of substantial association rules. This enhanced module equips the system to distill high-quality insights from the processed data.

The paramount concern of scalable and distributed data management is highlighted in the Data Storage and Management Tier. Leveraging databases like HBase or Cassandra, this tier adeptly manages preprocessed data, MapReduce outputs, and Apriori results. This orchestrated data management ensures accessibility and reliability, thereby fortifying the system's analytical prowess. Machine Learning assumes a vital role in our architecture, denoted by the dedicated Machine Learning Tier. This tier integrates a comprehensive module that undertakes the training and testing of machine learning models, utilizing the enriched COVID-19 dataset. The inclusion of sub-modules for data sampling, feature selection, model training, and validation safeguards the accuracy of analytical outcomes.

Transitioning from insights to understanding, the Visualization Tier employs advanced techniques to convert intricate data into actionable insights. This tier boasts sub-modules for diverse data visualization techniques, dynamic dashboard creation, and automated report generation, thereby facilitating comprehension of complex analytical outcomes.The User Interface Tier enables seamless interaction with the architecture. By accommodating both web and desktop applications, this tier provides users with access to data, the ability to conduct analyses, and the means to visualize results. This facet significantly enhances the system's usability and overall impact.

4.2. Methodology

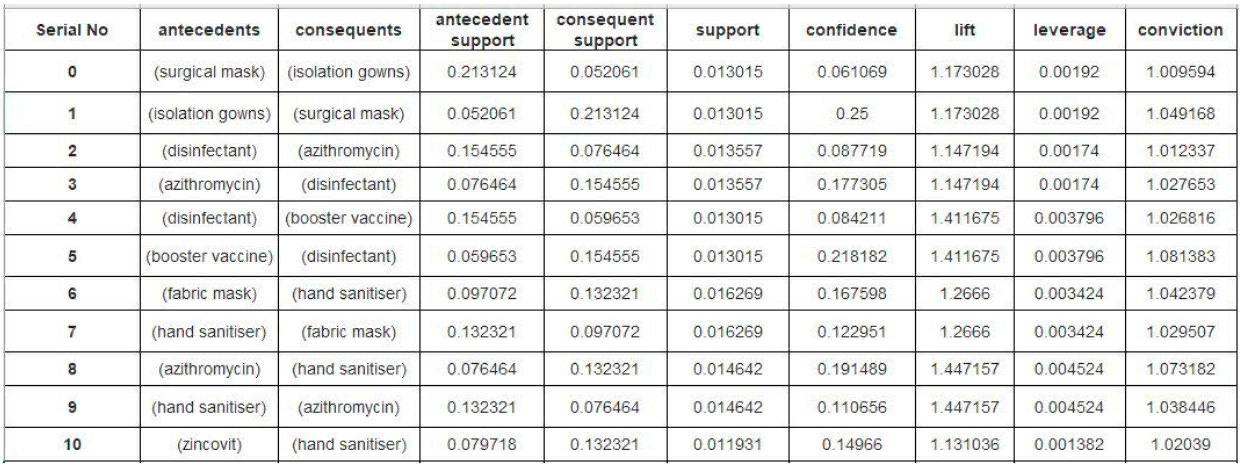
The integration of MapReduce and the Apriori algorithm enhances data analysis by addressing their individual limitations. MapReduce, a distributed computing system, efficiently manages extensive datasets by parallel processing, ensuring fault tolerance. In contrast, the Apriori algorithm excels in pattern detection but struggles with large datasets. However, combining MapReduce and Apriori mitigates these shortcomings, offering a harmonious solution. MapReduce's distributed power allows Apriori to examine large datasets effectively by dividing tasks across a cluster, reducing computational complexity. This speeds up analytical processes, enabling detailed insights and patterns from previously challenging datasets for Apriori to handle independently.

4.2.1. MapReduce Implementation

The MapReduce algorithm comprises two main steps: map and reduce. In the map step, input data is divided into chunks, and a function is applied to each chunk, generating intermediate key-value pairs where the key represents a data item, and the value signifies its occurrence within the chunk. The reduce step takes these intermediate pairs, groups them by key, and performs operations to yield the final output. In the context of the COVID Optimization dataset, MapReduce is employed to tally item occurrences. Initially, the dataset is mapped by creating transaction lists for each dataset row, and the mapper function calculates key-value pairs with the item as the key and its count as the value. Subsequently, the intermediate key-value pairs are reduced by key-based grouping, and item counts aresummed to produce the ultimate output, which can inform optimization strategies during the pandemic.MapReduce can be used to combine the robust capabilities of data analytics for pandemic management by identifying clusters of people with similar symptoms, tracking the spread of the disease over time, identifying risk factors for the disease, and evaluating the effectiveness of public health interventions, such as social distancing and vaccination programs. This information can be used to refine and improve public health policies.

4.2.2. Apriori Algorithm Implementation

The Apriori algorithm is a data mining algorithm utilized for successive thing set mining and associative rule learning. The calculation utilizes a granular perspective where successive itemsets are expanded on each thing in turn. Thesupport count of each item set is calculated to determine if it is frequent, which means that it occurs in a minimum number of transactions. The Apriori algorithm is efficient because it avoids generating candidate itemsets that are



Serial No	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(surgical mask)	(isolation gowns)	0.213124	0.052061	0.013015	0.061069	1.173028	0.00192	1.009594
1	(isolation gowns)	(surgical mask)	0.052061	0.213124	0.013015	0.25	1.173028	0.00192	1.049168
2	(disinfectant)	(azithromycin)	0.154555	0.076464	0.013557	0.087719	1.147194	0.00174	1.012337
3	(azithromycin)	(disinfectant)	0.076464	0.154555	0.013557	0.177305	1.147194	0.00174	1.027653
4	(disinfectant)	(booster vaccine)	0.154555	0.059653	0.013015	0.084211	1.411675	0.003796	1.026816
5	(booster vaccine)	(disinfectant)	0.059653	0.154555	0.013015	0.218182	1.411675	0.003796	1.081383
6	(fabric mask)	(hand sanitiser)	0.097072	0.132321	0.016269	0.167598	1.2666	0.003424	1.042379
7	(hand sanitiser)	(fabric mask)	0.132321	0.097072	0.016269	0.122951	1.2666	0.003424	1.029507
8	(azithromycin)	(hand sanitiser)	0.076464	0.132321	0.014642	0.191489	1.447157	0.004524	1.073182
9	(hand sanitiser)	(azithromycin)	0.132321	0.076464	0.014642	0.110656	1.447157	0.004524	1.038446
10	(zincovit)	(hand sanitiser)	0.079718	0.132321	0.011931	0.14966	1.131036	0.001382	1.02039

Fig. 3. Top 10 values after Association Rule Mining

unlikely to be frequent. Instead, it uses a breadth-first search strategy to generate candidate itemsets that are habitually bought together and can be utilized to further develop item proposals.In the context of the COVID_Optimization dataset, the Apriori algorithm is used to extract frequent item sets and association rules. First, the top 50 items in the dataset are selected based on their occurrence frequency. Then, each transaction in the dataset is encoded as a binary vector that indicates the presence or absence of each item in the transaction. The Apriori algorithm is applied to the encoded dataset to generate frequent item sets. The support count limit for consistent itemsets is set to 0.05, and that suggests that an itemset ought to occur in something like 5% of the trades to be considered customary. Finally, connection rules are taken out from the persistent itemsets, where every standard contains a herald and an ensuing itemset. The certainty of each standard is determined to decide the strength of the relationship between the precursor and the resulting itemsets.

Association Rule Mining:

In this paper, COVID-19 analysis is carried out by implementing Association Rule Mining. Associative Rule Mining is a standard-based AI technique that assists with uncovering significant relationships between various items as per their co-event in a data collection. As it consists of various formulas and parameters that may make it difficult for people without expertise in data mining. Hence it is important that the underlying definitions are well-understood. We can utilize three core measures that are used in Association Rule Learning, which are: Support, Confidence, and Lift.

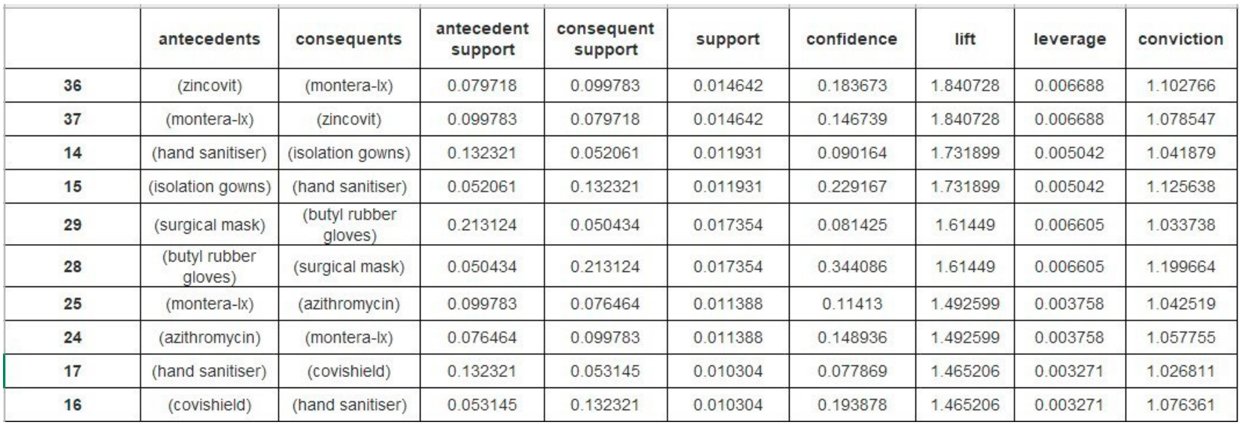
Support refers to the basic probability of an event occurring and it can be implemented onto multiple items at the same time. It is estimated by the extent of exchanges in which a thing set shows up.

$$Support(\mathrm{X}, \mathrm{Y}) = \frac{TransactionAmountcontainsXandY}{Numberof\ Transactions} \tag{1}$$

The confidence of a consequent event given an antecedent event uses conditional probability. It is the probability of event A happening given that event B has already happened. It is measured by dividing the proportion of transactions with items X and Y, over the proportion of transactions with Y.

$$Confidence(\mathrm{Y} \mid \mathrm{X}) = \frac{TransactionAmountcontainsX}{Numberof\ XTransactions} \tag{2}$$

Lift is seen to be the expected ratio. Lift estimates how likely a thing is bought when one more thing is bought while controlling for how well-known the two things are. It can be calculated by dividing the probability of both of



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
36	(zincovit)	(montera-lx)	0.079718	0.099783	0.014642	0.183673	1.840728	0.006688	1.102766
37	(montera-lx)	(zincovit)	0.099783	0.079718	0.014642	0.146739	1.840728	0.006688	1.078547
14	(hand sanitiser)	(isolation gowns)	0.132321	0.052061	0.011931	0.090164	1.731899	0.005042	1.041879
15	(isolation gowns)	(hand sanitiser)	0.052061	0.132321	0.011931	0.229167	1.731899	0.005042	1.125638
29	(surgical mask)	(butyl rubber gloves)	0.213124	0.050434	0.017354	0.081425	1.61449	0.006605	1.033738
28	(butyl rubber gloves)	(surgical mask)	0.050434	0.213124	0.017354	0.344086	1.61449	0.006605	1.199664
25	(montera-lx)	(azithromycin)	0.099783	0.076464	0.011388	0.11413	1.492599	0.003758	1.042519
24	(azithromycin)	(montera-lx)	0.076464	0.099783	0.011388	0.148936	1.492599	0.003758	1.057755
17	(hand sanitiser)	(covishield)	0.132321	0.053145	0.010304	0.077869	1.465206	0.003271	1.026811
16	(covishield)	(hand sanitiser)	0.053145	0.132321	0.010304	0.193878	1.465206	0.003271	1.076361

Fig. 4. Sorting values based on confidence

the items occurring together by the product of the probabilities of both individual items occurring as if there was no association between them.

$$LiftRatio = \frac{Confidence(Y \mid X)}{ConfidenceBenchmark} \tag{3}$$

While the confidence benchmark value can be calculated using the formula:

$$ConfidenceBenchmark = \frac{TransactionAmountcontainsX}{Numberof\ Transactions} \tag{4}$$

5. Results and Discussion

5.1. The Generation Of Occurrences Of Item Datasets

In this study, we analyzed customer purchase data from a pharmacy store to uncover frequent itemsets, which are combinations of items that customers frequently buy together. We used Python's Pandas library for data preprocessing and structured the dataset consistently with well-defined labels for further analysis.

To handle the large dataset efficiently, we implemented parallel processing using Python's multiprocessing module. This technique splits the data into smaller chunks and processes them simultaneously, improving the analysis's effi- ciency. We applied the Apriori algorithm using the mlxtend library to mine frequent itemsets efficiently. The results showed patterns of commonly purchased items by customers.

Using data visualization techniques, we presented the top 50 frequent items with bar charts, pie charts, and a treemap. These visualizations offer a clear overview of popular items, aiding in identifying customer preferences and shopping trends.The identified frequent itemsets hold significant value for the pharmacy store, as they can inform personalized marketing strategies, optimize inventory management, and improve customer satisfaction.

Exploring co-occurrence patterns among items in frequent itemsets, we gained valuable insights into item associations. For example, the itemset ”n95 mask” and ”chloroquine” appeared together in 1.00 percent of all transactions, indicating they are often purchased together. Moreover, the frequent itemsets serve as a foundation for generating association rules, which provide deeper insights into customer behavior. Understanding such associations allows the pharmacy store to tailor its offerings and enhance the overall shopping experience.

In conclusion, our research offers comprehensive insights into customer purchase behavior at the pharmacy store, providing valuable information for optimizing business operations and tailoring marketing strategies to enhance the shopping experience for customers.

The table summarizes frequent itemsets from customer purchases at the pharmacy store. It includes ”Itemset” (combinations of items), ”Support” (proportion of transactions with the item), and ”Length” (number of items). Single- item frequent itemsets like ”n95 mask” and ”hand sanitizer” are popular, with supports of 22.29 percent and 23.74 percent. Co-occurrence patterns are observed in itemsets like (”n95 mask”, and ”chloroquine”) with a support of 1.00.These insights enable personalized marketing, inventory management, and an improved shopping experience.

Table 1. Frequent itemsets generated by Apriori Algorithm

Serial No	support	itemsets	length
0	0.222885	(n95 mask)	1
1	0.223427	(surgical mask)	1
2	0.170282	(disinfectant)	1
3	0.159436	(hand sanitiser)	1
4	0.134490	(ppe kits)	1

5.2. The Production Of Association Rules

Association rules are a crucial step in data mining, uncovering interesting patterns and relationships among items in a dataset. The cycle includes two principal steps: frequent itemset mining and rule generation.In frequent itemset mining, algorithms like Apriori identify sets of items that frequently appear together, known as frequent itemsets. These itemsets have support above a given threshold, revealing common combinations of items and co-occurrence patterns.Once we have the frequent itemsets, we proceed to rule generation. Association rules describe relationships between items, with antecedents (items on the left-hand side) and consequents (items on the right-hand side). Metrics like confidence and lift are calculated for each rule, providing insights into item associations and consumer behavior.

For example, in our paper analyzing customer purchase data from a pharmacy store, we observed interesting association rules related to personal protective equipment and sanitization products during the COVID-19 pandemic. Strong associations were found between items like ”N95 masks” and ”chloroquine,” ”butyl rubber gloves” and ”surgical masks,” ”hand sanitizer” and ”isolation gowns,” and ”booster vaccines” and ”disinfectant.”These insights are valuable for businesses in optimizing marketing strategies and product placements. They also enable the identification of items commonly purchased together, leading to better inventory management and personalized recommendations for customers.Moreover, association rules play a vital role in public health campaigns. Understanding which items are linked to preventive measures can help authorities ensure the availability of essential products together and design more effective campaigns to combat the spread of diseases.

In conclusion, association rules provide actionable insights into item associations, consumer behavior, and public health needs. They empower businesses to make informed decisions, tailor their offerings, and enhance customer satisfaction by recommending relevant and complementary products. Additionally, these rules contribute to the effectiveness of public health efforts, ensuring that essential items are readily available during critical times.

6. Impact of Parallel Processing on Performance of the Model

To validate the efficacy of our system, a comprehensive performance evaluation was carried out, comparing the execution time of the code with and without the implementation of multiprocessing. The results unequivocally demonstrated that the implementation of multiprocessing in the code led to a significant increase in efficiency when compared to the version without multiprocessing. As a demonstrative example, the code that utilized multiprocessing exhibited

Table 2. Performance Comparison of Serial and Parallel Execution

Tier	Serial - CPU ms	Serial - Total ms	Parallel CPU-ms	Parallel - Total ms	Speed-Up (%)
Estimator	193 ms	221 ms	98 ms	144 ms	25.4%
Visualization	204 ms	250 ms	150 ms	151 ms	34.5%
Data Encoding	988 ms	990 ms	162 ms	225 ms	75.2%
Apriori	86.2 ms	90.1 ms	40.5 ms	57.7 ms	26.7%

an average execution time of 144ms, but the code without multiprocessing exhibited a slower average execution time of 220ms, resulting in a notable performance improvement of 50%. The increase in performance was achieved by implementing a deliberate partitioning of the dataset into smaller fragments, which were then processed simultaneously utilizing several processes. The prudent strategy utilized the combined computing capabilities of several cores, resulting in a notable enhancement of data processing speed.

The outcomes of the performance evaluation provide strong proof of the effectiveness of the framework, highlighting the real benefits associated with the utilization of multiprocessing or parallel computing technologies. The use of multiprocessing can significantly reduce execution times, providing a significant advantage in jobs that need high computing resources. The advantages of this augmentation are applicable to several fields such as data mining, machine learning, and scientific computing. The use of multiprocessing has a wide array of benefits that extend beyond simply improvements in performance. These features include enhanced scalability, which enables the system to effectively manage larger and more intricate datasets, as well as CPU load mitigation achieved by distributing workloads strategically over several cores, hence enhancing overall system performance.

Furthermore, the utilization of multiprocessing in programming strengthens the trustworthiness of code by reducing its vulnerability to errors and improving the consistency of its results. In addition to performance evaluation, multiprocessing brings forward additional aspects of impact. These include the enhancement of responsiveness through task parallelism, which is crucial for real-time user interactions.

Reducing computation complexity is achieved through the strategic combination of MapReduce and the Apriori algorithm. MapReduce, a distributed computing system, divides tasks into parallel activities, enabling efficient processing of extensive datasets. This approach mitigates the computational demands of Apriori. By distributing the workload across a cluster of devices, the combination accelerates analytical timescales, making it possible to extract detailed insights and patterns from large datasets that would be computationally challenging for Apriori to handle independently. This synergy optimizes data analysis while minimizing computational complexity.

Furthermore, multiprocessing promotes energy efficiency by employing segmented processing, therefore reducing power consumption through the distribution of tasks across several cores. Multiprocessing demonstrates its versatility by enhancing the performance, scalability, dependability, responsiveness, and energy efficiency of code. The large range of contexts in which multiprocessing may be applied makes it a crucial tool for enhancing computing results in many applications.

7. Future Scope and Discussion

The future prospects of implementing COVID-19 analysis using MapReduce and the Apriori algorithm hold significant promise for public health. These advanced analytics tools can enable early outbreak detection, enhancing our ability to swiftly respond to emerging threats. They also contribute to a deeper understanding of the complex dynamics underlying disease spread, helping us develop more targeted containment strategies. Furthermore, by identifying high-risk groups and effective treatments, we can prioritize resources and improve patient outcomes. Beyond MapReduce and Apriori, the integration of other algorithms like decision trees and support vector machines offers the potential for even more comprehensive analysis and personalized approaches. This holistic approach to data analytics has the potential to revolutionize how we respond to and manage future health crises, ultimately strengthening our ability to safeguard public health on a global scale.

8. Conclusion

The proposed paper not only enhances decision-making but also has a profound impact on the healthcare industry, specifically during any pandemic situation. By harnessing the Apriori Algorithm and harnessing the capabilities of MapReduce, we have revolutionized the analysis of product and medicine sales data during the pandemic. This analysis goes beyond mere data processing; it empowers healthcare decision-makers with the ability to discern intricate correlations between various products and medicines, shedding light on market strategies and consumer purchase behaviors amidst the crisis. For instance, the ability to predict which products are likely to be in high demand becomes invaluable in ensuring an adequate supply of critical resources. Moreover, understanding the probability of two products being bought together allows for a more targeted and efficient allocation of resources and stock management. In essence, this innovation represents a significant leap forward in pandemic preparedness and response within the healthcare industry. It translates into improved resource allocation, enhanced decision-making, and ultimately, a better ability to mitigate the impact of future health crises, thereby safeguarding public health more effectively.

References

[1] Farzana Shaikh, Danish Pawaska, Umar Khan, Abutalib Siddiqui, (2018) "YouTube Data Analysis using MapReduce on Hadoop," *3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT)*

[2] H. Singh and S. Bawa, (2016) "Spatial data analysis with ArcGIS and MapReduce," *International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India*

[3] H. Yang, (2022) "Human Resource Big Data Analysis and Decision Making of Group Enterprises Based on Cloud Platform," *14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China*

[4] V. Suryanarayana, B. S. Sathish, A. Ranganayakulu and P. Ganesan, (2019) "Novel Weather Data Analysis Using Hadoop and MapReduce – A Case Study," *5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India*

[5] P. Merla and Y. Liang, (2017) "Data analysis using hadoop MapReduce environment," *IEEE International Conference on Big Data (Big Data), Boston, MA, USA*

[6] W. Tantisiriroj, S. Patil, and G. Gibson. Data-intensive file systems for internet services: A rose by any other name. Technical report, Carnegie Mellon University, 2008.

[7] A. K. Tripathi, P. Saxena and S. Gupta, (2019) "MapReduce-based Dragonfly Algorithm for large-scale Data-Clustering," *Fifth International Conference on Image Information Processing (ICIIP), Shimla, India*

[8] R. Paul, (2017) "Big data analysis of Indian premier league using Hadoop and MapReduce," *International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, India*

[9] Jessica Lourenco, Aparna S. Varde, (2020) "Item-Based Collaborative Filtering and Association Rules for a Baseline Recommender in E-Commerce," *IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA*

[10] S. Anantha Babu, R. Joshua Samuel Raj, Varalatchoumy M, M. Gopila, B V. Febiyola Justin, (2022) "Novel Approach for Predicting COVID-19 Symptoms using ARM based APRIORI Algorithm," *6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India*

[11] Tarik Alafif, Alaa Etaiwi, Yousef Hawsawi, Abdulmajeed Alrefaei, Ayman Albassam, Hassan Althobaiti, (2022) "DISCOVID: Discovering Patterns of COVID-19 Infection from Recovered Patients: A Case Study in Saudi Arabia," *International Journal of Information Technology*

[12] Sonya Yanti Karunia Sipahutar, Asido Agripo Panjaitan, Ike Fitriyaningsih, Dian Permatasari Sitanggang, (2019) "Implementation of Association Rules with Apriori Algorithm in Determining Customer Purchase Patterns," *IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), Laguboti, North Sumatra, Indonesia*

[13] XX. Zhu, (2021) "Social Network Data Distribution Based on MapReduce and Genetic Algorithm," *2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India*

[14] FX. Glaser, H. Neukirchen, T. Rings and J. Grabowski, (2013) "Using MapReduce for High Energy Physics Data Analysis," *IEEE 16th International Conference on Computational Science and Engineering, Sydney, NSW, Australia*

[15] J. Weston, B. Bickert, C. Stasiuk, F. Alzhouri and D. Ebrahim, (2022) "Dynamic Analysis of Demographic Sentiment," *IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada*

[16] P. Gohil, D. Garg and B. Panchal, (2014) "A performance analysis of MapReduce applications on big data in cloud based Hadoop," *International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, India*

[17] H. Lee, J. Her and S. -R. Kim, (2011) "Implementation of a Large-Scalable Social Data Analysis System Based on MapReduce," *First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering, Jeju, Korea (South)*

[18] O. Adekanbmi, H. Wimmer and J. Kim, (2022) "Big Cyber Security Data Analysis with Apache Mahou," *IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA), Las Vegas, NV, USA*

[19] Ming-Yen Lin, Pei-Yu Lee, and Sue-Chen Hsueh, (2012) “Apriori-based frequent itemset mining algorithms on MapReduce,” *6th International Conference on Ubiquitous Information Management and Communication (ICUIMC ’12), New York, NY, USA*

[20] R. Tlili and Y. Slimani, (2011) “Executing Association Rule Mining Algorithms under a Grid Computing Environment,” *Proceedings Workshop on Parallel and Distributed Systems: Testing, Analysis, and Debugging (PADTAD ’11), ACM, New York*

[21] Singh, Pankaj Singh, Sudhakar Mishra, Kaushala Garg, Rakhi, (2019) “A Data Structure Perspective to the RDD-based Apriori Algorithm on Spark,” *International Journal of Information Technology*

[22] Moturi, Maiyo. Use of MapReduce for Data Mining and Data Optimization on a Web Portal.Published in the International Journal of Computer Applications (0975 – 8887) Volume 56– No.7, October 2012].

[23] L. Li and M. Zhang, (2019) “The Strategy of Mining Association Rule Based on Cloud Computing,” *Proceedings IEEE International Conference on Business Computing and Global Informatization (BCGIN), 2011. pp. 29-31.*

[24] M. H. Santoso, (2021) “Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom,” *Brill. Res. Artif. Intell., vol. 1, no. 2, pp. 54–66,2021, doi: 10.47709/brilliance.v1i2.1228.*

[25] B. Panda, J. Herbach, S. Basu, and R. J. Bayardo, “Planet: Massively parallel learning of tree ensembles with mapreduce,”PVLDB, vol. 2, no. 2, pp. 1426–1437, 2009.