International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at :www.ijsrcseit.com
doi : https://doi.org/10.32628/IJSRCSEIT

Enhanced Resource Efficiency for Association Rule Mining in Cloud Environments via Apache Spark

Md Sohrab Ansari[1*], Prof. Vinod Mahor[2]

[1]Research Scholar, Department of Computer Science & Engineering, Millennium Institute of technology and Science, Bhopal, India

[2]Assistant Professor, Department of Computer Science & Engineering, Millennium Institute of technology and Science, Bhopal, India

ARTICLEINFO

Article History:

Accepted: 01 Dec 2023
Published: 10 Dec 2023

Publication Issue
Volume 9, Issue 6
November-December-2023
Page Number
369-381

ABSTRACT

Data from various sources, including mobile devices, sensors, and web cams, constantly accumulates and is evaluated in Big Data. These processed data are crucial in various fields, such as research, business, and industry. Apache Spark is a versatile platform for processing both batch and real-time data. Cloud computing provides resources for real-time processing of applications. Association Rule Mining (ARM) is a technology that analyzes the link between objects to identify comparable groupings. FP-Growth is the most widely used algorithm for finding common patterns and locating mining pieces quickly. The aim of this research is to enhance the efficiency of Association Rule Mining by creating rules for big data sets in Big Data environments. The proposed solution enhances association rule efficiency by utilizing the FP-Growth algorithm in a Hadoop Map Reduce setting. FP-Growth is the most used method for discovering and mining frequent patterns.

This research introduces the FP-Growth parallel method in Spark Framework. The efficient use of Spark resources through heterogeneous allocation reduces runtime and costs. Apache Spark is a versatile Big Data platform for real-time streaming and batch processing. Cloud computing is used in streaming applications to address real-time processing needs by supplying necessary resources. Using big data apps in a virtualized cloud environment may cause performance issues that impact streaming workloads. The ARM approach identifies highly associated models in item sets. FP expanding is the most common ARM algorithm.

The FP-Growth algorithm is implemented in Spark using OpenStack. This article covers OpenStack architecture, needs, configuration, and problems. The analysis evaluates resource utility at full load and no load, and evaluates performance using virtual resource allocation. Using Spark resources efficiently reduces turnaround time and optimizes costs due to their diverse distribution.

Keywords: Association Rule Mining, Cloud System, FP, Big data.

I. INTRODUCTION

The phrase "cloud storage" describes a range of operations that entail the provision of resources via the internet, such as online services. Software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) were the three primary types of these capabilities. A mist may have either internal or exterior properties. Anyone with an online connection could access a communal cloud. A private cloud is a proprietary protocol or database server that provides internet-based applications to a limited number of users who are privy to each other and the internet or computer system. Whether internal or external, cloud technology was created to offer simple, adaptable access to systems, including technical support services. The term "cloud platform" describes the hardware and software needed for a cloud computing technology to function properly. Another way to think about cloud technology is as a virtualization technology or even as a utility technology [1-4].

The concept of cloud technology refers to the ability of consumer devices to establish a connection to the information that is held on corporate networks, programs, and laptops through the use of the internet, as seen in figure 1.

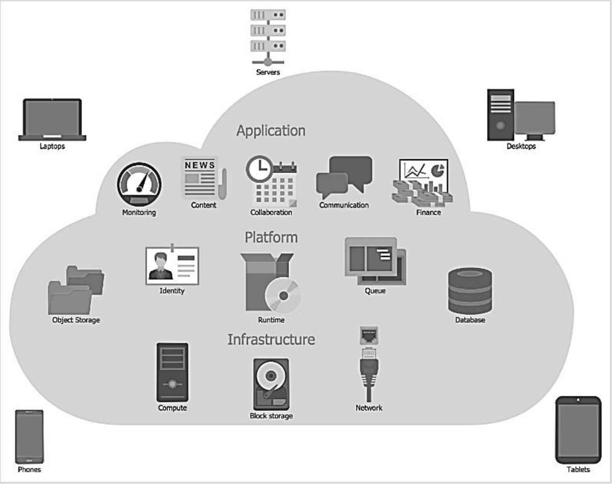

Figure .1: Architecture in the Cloud"

Along with the front endpoint, which includes the requesting consumer equipment, website, networking, and cloud software platforms, a link has been discovered. Additionally, the leading edge, which also includes the network infrastructure, which includes databases, servers, and computers, has also been discovered. In addition to its role as a warehouse for information, the leading edge is responsible for carrying out tasks that are accessible from the front finish [5-6].

There was a centralized computer that was responsible for managing the connectivity in both the front and rear endpoints. The database controller makes use of conventions in order to facilitate the flow of information in a very accessible manner. The centralized controller is responsible for managing connections between an assortment of consumer users and cloud providers. This is accomplished through the utilization of either application or firmware. Every application will occasionally have its own physical servers, which will need to be accessible from this location [7].

A. *The Big Data Environments*

Information from the huge is a term that is used to describe a collection of data that is not only enormous in size but also expands at a quick rate as it grows. Due to the fact that it is a collection of items that is both so extensive and so complex, neither of the conventional data management methods could adequately preserve or handle it. The term "big data" refers to information that is more diverse, that is collected in higher quantities, and that is transmitted at a faster rate than statistical research was previously doing. It is also possible to discuss the three Vs.

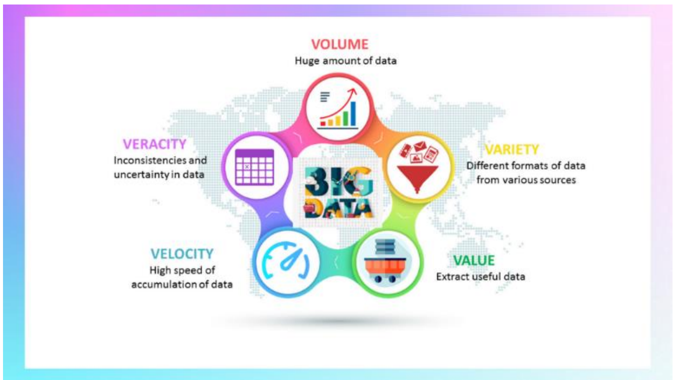
VOLUME
Huge amount of data
VERACITY
Inconsistencies and uncertainty in data
VARIETY
Different formats of data from various sources
BIG DATA
VELOCITY
High speed of accumulation of data
VALUE
Extract useful data

Figure 2 : Qualities of Big Data Environments

The goal of this process is to accomplish analysis. Doug Laney, a veteran of the industry, provided the definition of the phrase in the 2000s, which was that it should consist of the three Vs: speed, volume, and diversity. One of the many resources included things like trades, Internet of Things (IoT) devices, industrial facilities, films, and internet networks, in addition to other information sources that were comparable. In the past, storing information would have been a challenge; however, low bandwidth solutions such as data warehouses and Hadoop have made the burden of information storage more manageable [8].

a) *Speed*

The Internet of Things is gaining steam, which means that organizations are receiving information at an amazing rate. All of this information needs to be managed as fast as possible. The management of massive amounts of information in real time has become increasingly important as a result of the growing use of RFID tags, monitors, and other monitoring devices.

b) *The availability*

The analysis of information can be performed in a variety of formats, including organized, empirical data stored in conventional relational databases, as well as unorganized formats, communications, videoconferencing recordings, and information from stock exchanges, including corporate accounts.

Information that provides businesses with significant discoveries that, among many other things, assist to the formulation of marketing plans and opportunities that are virtually limitless in scope. These are the three significant actions that can be taken, and they are as follows:

c) *Integrated*

When we talk about "information from the large," we are referring to the compilation of data that comes from a wide range of sources and activities. Methods of integrating information that have been used traditionally, such as the extractor, convert, and pack operation, were frequently insufficient for the task at hand. The investigation of enormous data sets on a terabyte, or possibly even a petabyte, scale entails investigating the

new methods and materials are being developed. The process of incorporating information involves not only taking in the information but also analyzing it and offering access to it in a format that professionals working in the corporate industry might use to get things up and running more efficiently.

d) *Managing*

Using RAM was necessary because of the massive datasets that were being used. For the purpose of storing information, we have the option of using either the internet or corporate facilities, or even both. Visitors are able to record their information in whatever format that they like, and they may subsequently implement their desired information needs as well as every required system technique to any of those big datasets in an ad hoc way. The majority of individuals make their choice of storage technique based on the location with which their information is stored at the time of the transaction. The internet was quickly gaining popularity due to the fact that it can meet the modern computing requirements of businesses while also enabling them to rapidly deploy more resources on demand [9].

e) The analysis clouds

Users would be able to make a profit off of their substantial data commitment anytime businesses evaluate and take action based on relevant information. A fresh perspective would be provided to the visitor if they were to visually examine their numerous sources of data. We have made the decision to investigate the facts in order to arrive at new conclusions. The information that you have gathered should be made available to other people. Through the use of cognitive computing, information models can be constructed.

II. LITERATURE REVIEW

As expected, resource use, another natural strategy was used. This ratio of processed assignments to overall capabilities used for such cases is projected. Any flexible systems that might reorganize to use resources most efficiently used these projected percentages. That method can anticipate and balance production phase staffing.

Thus, such distributions were employed in its management loops to monitor secondary processes capable of adapting to altered settings while satisfying the same application requirements. This assessment used another computation activity where the amount of information transported was minor compared to the processing duration, but information localization was considered. This can significantly minimize information transmission efficiency losses when computations are near enough to actual information. Another inevitable necessity was that heterogeneous capabilities send information to the same system [10-12].

Xu et al. say this clouds computers architecture provides vast memory and computing capabilities for system design and implementation, allowing its potential to increase. However, data-intensive processing tools like meteorological forecasting and economic analytics were prevalent, thus large database streams with sensitive individual data were available during workflow implementation. Creating a records positioning procedure for workload development in virtualized facilities requires balancing many achievement metrics, such as commodity consumption records acquisition duration and electricity cost, since ignoring security conflicts of data points is difficult.

Ramamurthy et al. studied cloud computer connection applications, however capacity management remained still open. Capacity management often involves attaining numerous goals, but it was often portrayed as a single problem with immediate consequences. Internet capacity management includes fundamentally sensitive computing concerns, therefore optimal strategies can help design new novel cross-issue solutions. Cyber commodity management aimed to reduce customer billing costs and increase internet services provider revenue [13].

Müller et al.'s virtualized compute platforms have recently attracted academics and industry due to their promise of maximum adaptability and practical efficiency. However, there was no unanimity on whether such technique constituted sufficient business information handling. The research report examined if virtualized compute systems could do database analysis using lambada.

Compared to Li et al., modern internet networks have several components that were constantly improved. It's challenging to keep up with such frequent modifications while leading their equipment to fail. The Gandalf analytical services discussed in this article are built for use in this study's vast network architecture. Gandalf enables management quickly and thoroughly assess the consequences of every technology deployment, allowing them to identify and mitigate problems before they cause major disruptions. Gandalf mostly monitored and assessed trouble signs. The corporation would use statistical correlations rather than geographical connections to link every signal utilizing any other deployment [14-16].

Cloud Resource

This same distribution of many simulated machines ICT assets due to homogenous implementation workflows (Machine learning, information distribution, but rather connections internet apps) to confrontational distribution requisites throughout definitions of ICT commodity capabilities, about Fardet al., would be a complicated concern established in a droplet virtualization framework.

According to authors Abid et al., the study sought to uncover internet services firms' commodity distribution difficulties, such as processors, memory, and networking in internet computers. Over 75 publications on capacity management in cloud computing from 2008 to 2019 were selected methodically and assessed according to clear goals. Finally, both evolution and resource distribution studies are important since they have considered more relevant future possibilities. described a subsystem as multiple identification machines connected by network transportation software [17].

Cloud Environment Load Prediction Approach

Bhattacharjee et al. defined clouds computers as any approach that was widely employed because to its capabilities and speed. Cloud computing led to a significant expansion in broadband technology. Cloud technology offers many benefits, but it also increases greenhouse gas emissions. This was because internet information facility hardware requires far more power than PCs. Many cloud scientists' top concerns were resource conservation due to such effects. This paper thoroughly examines numerous ways to reduce server farm power use.

Association Rule Mining Approach

In this wide range of experimental testing, scientists have applied this method in both simultaneous and scattered contexts. Meanwhile, most computer projects centered on operating ARM in such a distributed environment. Its mathematical design for mining used Sparks Foundation and RP-Tree computational foundation. Horizontally organized information groups based on transactional identification handle scanner problems throughout its full information set. This scanner screens horizontally oriented documents to evaluate information support. Association Regulation Miner (ARM) defined related connections amongst varied things to find this least often repeating object group.

Frequently used objects on Hadoop

Many academics, like Raj et al., use clustering extraction to locate knowledge in transaction databases. It underpinned connection regulation mining, a sort of information extraction. Many methods for discovering common themes were proposed, including the apriority methodology, which was increasingly used and encouraged. Algorithms had two main slowdowns: searching incoming data continuously and creating every possible substring while computing top reasons from contender items. Such constraints drastically limit apriority's effectiveness when working with massive data. Acceptable efforts have been made to reduce large obstructions to increase throughput [18].

Apache Spark for Bigdata

Hadoop was another distributed computing system that used fundamental concepts to spread computation from big data volumes among numerous groups or processors. This has several computational uses, including solving complicated issues involving enormous amounts of computer data. Sequential database management was least efficient using

Hadoop. Hadoop's major components were MapReduce and HDFS. The MapReduce computing approach was used for projects that required massive amounts of data to be processed simultaneously across large groups of devices. The gadget is reliable but has flaws [19-21].

Hadoop Bigdata Distributed File System Environment

PrePost N-list-based techniques are used to extract connection rules. An N-list database structure for linking rules extraction is created by this technique. PrePost uses two database queries to generate a tree with an N-list of common 1-itemsets. A dataset can be extracted without rescanning by intersecting it with the integrating N-list. That Hadoop-based regular subsets extraction approach requires no intermediary data and few devices to connect to. FiDoop uses the Voronoi diagram-based information division approach with the Locality Sensitive Hashing mechanism to divide massive volumes to information and improve Hadoop parallelism and FIM efficiency [22-23]

Apriority algorithms for association rule mining survey

In this wide range of experimental testing, scientists have applied this method in both simultaneous and scattered contexts. Meanwhile, most computer projects centered on operating ARM in such a distributed environment. Its mathematical design for mining used Sparks Foundation and RP-Tree computational foundation. Horizontally organized information groups based on transactional identification handle scanner problems throughout its full information set. This scanner screens horizontally oriented documents to evaluate information support. Association Regulation Miner (ARM) defined related connections amongst varied things to find this least often repeating object group [24-25].

III.PROBLEM IDENTIFICATION AND OBJECTIVES

In the expansive field of data mining, association rule mining (ARM) stands out as a critical method for discovering interesting correlations, frequent patterns, associations, or causal structures among sets of items in transaction databases or other data repositories. As data volumes continue to grow exponentially, particularly in cloud environments, traditional ARM tools struggle with scalability and resource efficiency. The utilization of Apache Spark for ARM in cloud environments promises significant improvements in performance due to its in-memory cluster computing capabilities. However, despite its potential, several inefficiencies persist that must be addressed to optimize resource usage and operational speed.

Current Challenges in ARM Using Apache Spark:

ARM processes often deal with large, sparse datasets that are typical in many real-world applications like market basket analysis, web usage mining, and bioinformatics. Apache Spark, while proficient at handling large-scale data, often encounters performance bottlenecks due to the sparse nature of transactional datasets, which can lead to excessive shuffling of data across the cluster nodes.

Efficient management of computational resources in Spark-based ARM implementations is a persistent issue. The static allocation of resources often leads to underutilization or, conversely, resource contention, particularly under varying workloads typical in cloud environments.

Objectives

The primary goal of this project is to enhance the efficiency of resource utilization in cloud environments specifically for association rule mining using Apache Spark. The objectives to achieve this goal are:

- Develop a Scalable Framework: Design and implement a scalable Apache Spark-based framework that optimizes resource allocation and processing power for association rule mining tasks in cloud environments. This framework should adaptively manage resources depending on the workload characteristics to minimize overheads and operational costs.

- Optimize Data Processing Workflows: Refine data processing workflows in association rule mining to exploit Apache Spark's in-memory computing capabilities, which can drastically reduce the data processing time by minimizing disk I/O and network congestion.

IV.PROPOSED METHODOLOGY

In the context of large-scale task processing, such as association rule mining, where jobs may be compute-intensive or data-intensive, it appears that resource consumption is of utmost importance. When applied to big datasets, the present FP Growth method for generating frequency item sets requires significantly more time to implement, which results in a decrease in network performance. It is possible that the work being done to identify the design may require a variety of resources; hence, the problem of resource consumption is a challenging one to overcome. By employing cloud resources, which enables the creation of adaptive resource groups with the capability to adjust the quantity of resources on the fly, it is feasible to establish resource groups that are flexible. Despite the fact that the cloud has the ability to offer resource provisioning that is both effective and adaptable, the overall performance disruptions that may occur for a variety of potential applications may have an impact on the productivity of efficient material usage. As a consequence of this, the approach that has been proposed intends to emulate cloud-based efficient resource control for Association Rule mining applications that include Apache Spark. Additionally, it is capable of rapidly calculating frequent item sets on massive quantities of data [57].

Through the use of the task aggregation technique, it is possible to allocate many projects to virtual machines (VMs) depending on the processing power of those VMs, which is determined by MIPS software. Because of the high consumption of the server's central processing unit (CPU), the job consolidation strategy, which encompasses mapping the customer's requirements, was suggested. As a result of delegating all work areas to the virtual machine (VM), the utilization of the CPU is improved while the amount of energy consumed is reduced.

4.2 Generation of K Itemsets on a Regular Basis

The third MapReduce task includes a stage that is dedicated to the mining of frequent itemsets, the creation of k-FIU trees, and the decomposition of frequent itemsets. This stage requires a significant amount of calculations. Two primary goals are incorporated into each design: In the first step, each k-itemset that is produced by the second MapReduce job is broken down into a list of small-sized sets, with the number of sets ranging from 2 to k 1. In the second step, a DBFIU (Density-Based Frequent Itemset Ultrametric) tree is constructed by integrating the local breakdown results of the same duration. Considering that each, Unlike the previous mappers,

The following is a synopsis of the approach that was recommended.

- The FP-Growth model incorporates heterogeneous resource planning in addition to flexible group deployment.
- A decrease in expenses associated with the variability of resource provisioning

Figure 3 illustrates the general framework of the technique that was provided there. A.F. The Custom Resource Allocator is the recipient of development Association Rule mining programs, and the cloud environment is the supplier of resources for job execution. Spark master has the capability to automatically assign resources to the spark processor, and once the tasks have been completed, the Custom Resource Allocator will release the resources to the cloud. The MapReduce framework that is part of Hadoop makes it possible to spread the processing of massive volumes of data over large clusters. The methodology proposed for enhancing resource efficiency in association rule mining (ARM) within cloud environments leverages Apache Spark due to its robust in-memory processing capabilities and ability to scale across multiple nodes. The central aim is to optimize the mining of association rules from large datasets, thereby reducing the resource consumption and computational overhead typically associated with such tasks. The methodology is structured into several key phases: setting up the Spark environment, data pre-processing, parallel rule mining, optimization of rule generation, and evaluation. Each phase is detailed below [60-62].

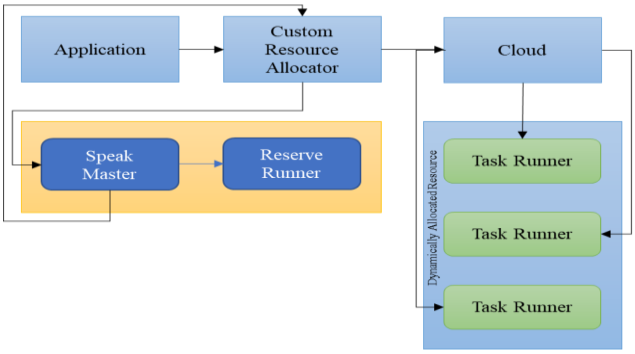

Figure 3 : Proposed Architecture

Proposed Algorithms

Algorithm Steps for Proposed System

The development of the proposed method involves the following detailed steps:

Step 1: Read Dataset

- Encoding the Database:
- The database is encoded using dictionary encoding, where strings are converted into integers. This reduces the file size significantly.
- Input Preparation for MapReduce:
- The encoded data is then formatted as input for the MapReduce job.

Step 2: Parallel Counting and Sorting of Items

- Calculation of Support Values:
- Support values for all items are calculated to determine item popularity. These values are then sorted to facilitate efficient processing.
- Use of Combiners:
- Combiners are utilized to optimize time, memory usage, and computation costs during the MapReduce job.
- Data Compression:
- During the MapReduce job, the map output transferred over the network is compressed using the Bzip2 Codec, reducing network I/O demands.

Step 3: FP-Growth Processing

- FP-Growth Algorithm Implementation:
- The FP-Growth algorithm is executed within the MapReduce framework to construct the FP-Tree.
- Data Partitioning with Modified LSH:

Data partitioning is performed using a Modified version of Locality-Sensitive Hashing (LSH) that is tailored for integer data

Step 4: Aggregation

- Aggregation of Results:
- In the final phase, the outputs from the FP-Growth process are aggregated to produce the final set of association rules.

This structured approach ensures clarity and efficiency in implementing the proposed FP-Growth based association rule mining using Hadoop MapReduce. The method is designed to optimize computational resources and minimize processing time, making it suitable for large datasets.

V. RESULT ANALYSIS

When it comes to large-scale task processing, such as association rule mining, where jobs may be compute-intensive or data-intensive, resource consumption appears to be of the utmost importance.

When applied to big datasets, the present FP Growth method for generating frequency item sets requires significantly more time to implement, which results in a decrease in network performance. It is possible that the work being done to detect the design may require a variety of resources, which will make the problem of resource consumption a challenging one to tackle..

Evaluation Based on Experiments

Using Apache Spark, the Scala programming language, and Ubuntu Server edition 18.04, the application that was recommended was developed and implemented in a cloud-based architecture. The FP-Growth approach, which is an iterative specialized application with a model developed, is utilized in order to assess the performance of the proposed system with consideration. The numerous input needs that were utilized in FP-Growth included 4.7 GB as well as 10 GB for microprocessors with 1-5 cores and RAM capacities that ranged from 1 to 12 GB.

For the sake of demonstrating the suggested method, let's assume that they have four functions. There are four meta-tasks: T1, T2, T3, and T4. The planning supervisor is responsible for two problem sets, which are R1 and R2. Both the velocity and capacity ranges for each resource are included in Table 4.1. Table 4.1 allows for the calculation of the (T) and (E) of such work on each of the resources through the use of the formula.

Table 5.1: Time spent on the assignment and overall task.

Sample Problem	Resources	Processing Speed (MIPS)		Throughput (MBPS)	
		Max-Min	Modified Max-Min	Max-Min	Modified Max-Min
P1	R1	29	50	87	100
	R2	73	100	300	500
P2	R1	128	150	279	300
	R2	283	300	130	150
P3	R1	272	300	282	300
	R2	17	30	30	150

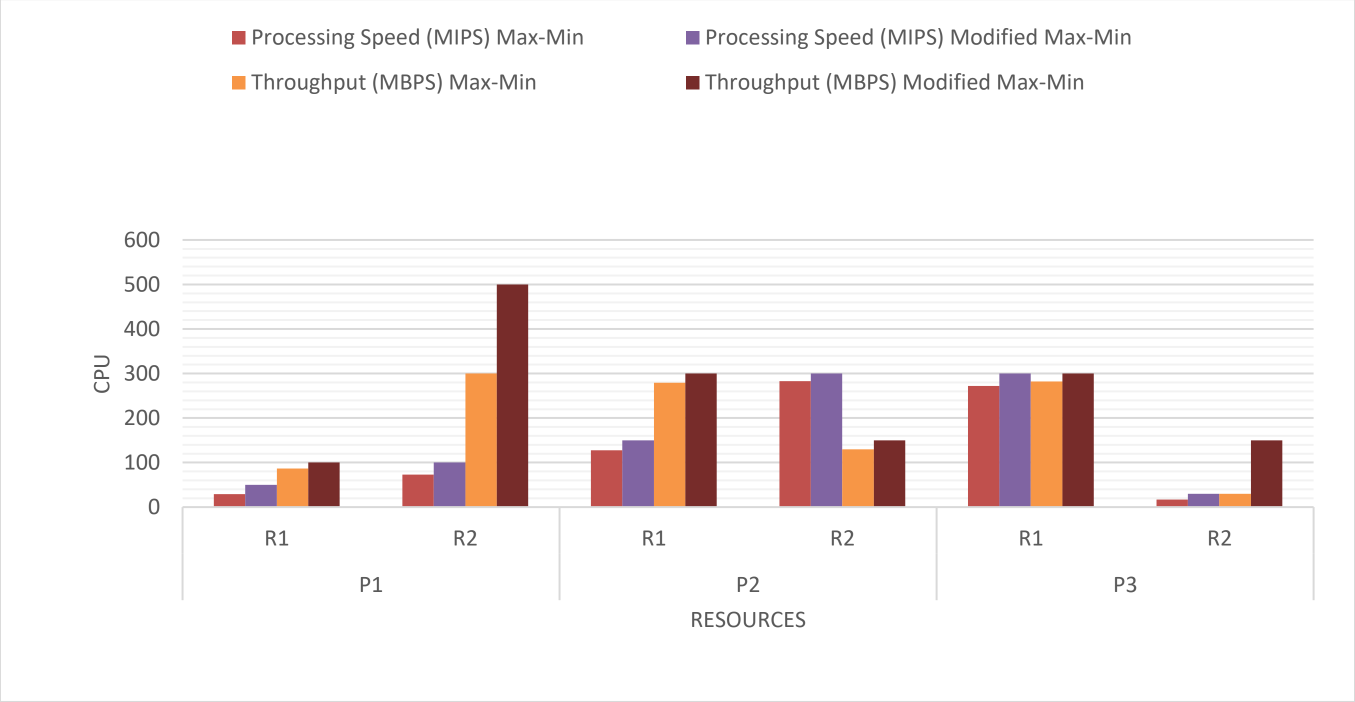

Figure 4 : Overall compression of CPU and Resources

A comparison of the resources of the CPU and RAM to the running number A model of machine learning that is based on Fp-Growth is depicted in Figure 4 respectively. With the result of task running time presented alongside container size, the chart demonstrates the outcomes of various CPU and RAM ratios, as well as a different product package. Additionally, the chart indicates the results of a different product package. FP-Growth applications require a minimum container size of one core with one gigabyte of random-access memory (RAM), and three cores with six gigabytes of RAM.

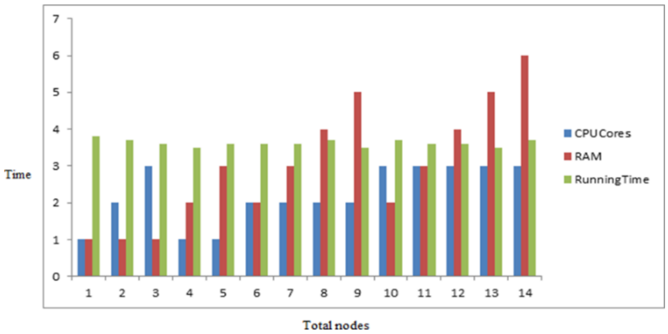

Figure 5: Machine Learning Program Running Time correlates CPU and RAM resources.

Figures 5 through 6 demonstrate the influence that the resources of the central processing unit (CPU) and random-access memory (RAM) have on the amount of time that the system is operational. The graphic illustrates the optimum resource level, along with the elbow curve that results from extra resource distribution beyond the optimal levels, which either results in an increase in the amount of time required to complete the task or results in the completion time remaining the same.

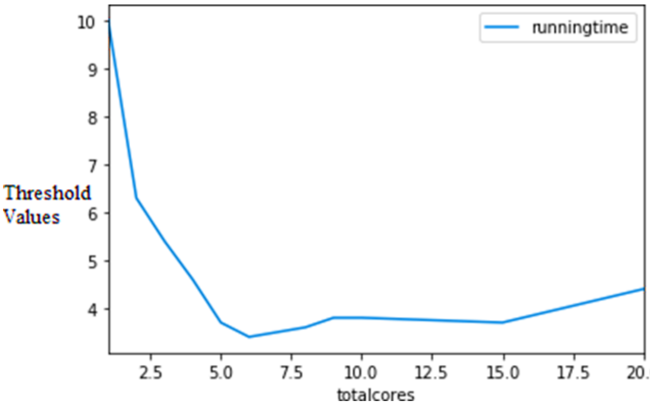

Figure 6 : The L-bow Curve of Machine Learning for Data Sizes of 23 GB

In the Spark environment, the diagram indicates that a maximum of four megabytes of data might be allocated to each core for the implementation of the FP-Growth technique. The results of the trials indicate that the optimal ratio of central processing unit (CPU) to random access memory (RAM) for carrying out the suggested technique is 750 megabytes of RAM per core for overall optimal functioning. Humans utilize the Fp-Growth approach with the quantity of information that has to be evaluated in order to determine the resources that are necessary to carry out the work successfully.

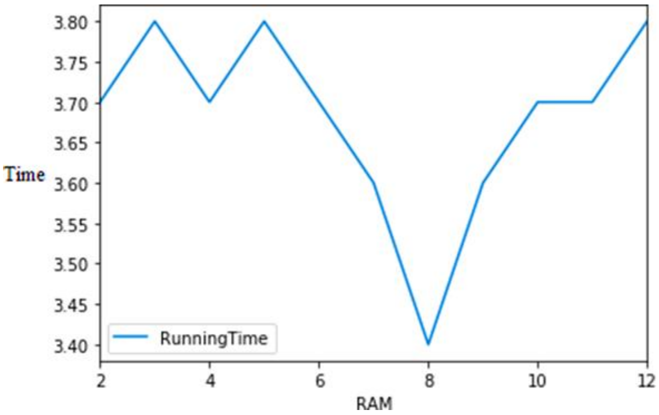

Figure 7: FP-Growth Running Time in Relation to Cpu and RAM Ratio

The utilization of resources for large-scale job processing operations that need a significant amount of computational power, such as FP-Growth, is of the utmost importance. When using cloud resources, which enable the use of flexible groupings of resources, the quantity of resources might be dynamically modified.. It may be possible in the future to do interactive forecasting with the help of FP-Growth in order to predict the resources that will be the most effective. The purpose of this is to ensure that the resource distribution can automatically adjust the allocation of resources depending on the unique nature of the characteristics that are included in the information after a few executions on the data.

VI. CONCLUSION AND FUTURE WORK OF SCOPE

ARM is used to identify shared models and their relationships between element sets. Association rule mining algorithm FP-Growth. Traditional association rule mining algorithms are ineffective for huge datasets as they require extensive scanning to identify frequent items. ARM algorithms are built for parallel execution on a distributed network to address this issue. The utilization of big data can enhance its performance. Hadoop MapReduce is utilized for an efficient FP-Growth rule exploration technique in this search job. Dictionary encoding with compression, combinatory, and partitioning improves efficiency for frequent concurrent item set exploitation in Hadoop clusters. The ARM method is evaluated for runtime across dimensions and cluster nodes.

Effective resource utilization is crucial for large-scale, compute-intensive tasks like FP-Growth. Using cloud resources with dynamic clusters enables dynamic resource size adjustments. While the cloud offers flexible and efficient resource supply, global performance interference can impact resource efficiency across applications. To minimize job execution time and resource consumption, assess the amount of data to be processed to determine suitable resources. Thus, more asymmetric container types are allocated with significant CPU resources. Additionally, the resource allocation technique determines the number of containers based on input data size, with optimal values determined by experimental evaluation. Traditional ARM algorithms are inefficient for huge datasets, as they require a significant amount of time to scan.

ARM is developed to solve the problem of parallel execution over a dispersed network. According to big data, its performance improves. For efficient FP growth, Hadoop Map Reduce, based on the 157 ARM algorithm, is a preferred option, integrating with Open Stack framework. Encoding with compression, combiner, and partitioning improves the performance of simultaneous frequent item set mining in Hadoop clusters. The ARM Algorithm is assessed for execution time across dimensions and cluster nodes. The proposed technique outperforms existing methods.

6.2 Future Work of Scope

A dynamic prediction that makes use of FP-Growth can be carried out in order to forecast the resources that are the most optimal. Once the data has been processed a few times, the resource allocation can automatically adjust itself to the particular nature of the features that are included in the data. There is potential for the system's performance to be improved in the future by reducing the amount of fault detection. The existing research work can be improved upon by developing some more sophisticated or enhanced algorithms, which will result in work that is both more accurate and more efficient.

VII. REFERENCES

[1]. Aditiya, R., Defit, S., & Nurcahyo, G. W. (2020). Prediksi Tingkat Ketersediaan Stock Sembako Menggunakan Algoritma FP-Growth dalam Meningkatkan Penjualan. Jurnal Informatika EkonomiBisnis, 67-73.

[2]. Aditya, C., Akash, M., Akash, P., Amitkumar, M., Nagarathna, K., Suraj, D., ... & Meena, S. M. (2020). Claims-Based VM Authorization on OpenStack Private Cloud using Blockchain. Procedia Computer Science, 171, 2205-2214.

[3]. Ahmed, N., Barczak, A. L., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. Journal of Big Data, 7(1), 1-18.

[4]. Alnasir, J. J., & Shanahan, H. P. (2020). The application of hadoop in structural bioinformatics. Briefings in bioinformatics, 21(1), 96-105.

[5]. Alotaibi, S., Mehmood, R., Katib, I., Rana, O., & Albeshri, A. (2020). Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and Machine Learning. Applied Sciences, 10(4), 1398.

[6]. Anbarasan, M., Muthu, B., Sivaparthipan, C. B., Sundarasekar, R., Kadry, S., Krishnamoorthy, S., & Dasel, A. A. (2020). Detection of flood disaster system based on IoT, big data and convolutional deep neural network. Computer Communications, 150, 150-157.

[7]. Anilkumar, C., & Subramanian, S. (2020). A novel predicate based access control scheme for cloud environment using open stack swift storage. Peer-to-Peer Networking

[8]. Banchhor, C., & Srinivasu, N. (2020). FCNB: Fuzzy Correlative naive bayes classifier with mapreduce framework for big data classification. Journal of Intelligent Systems, 29(1), 994-1006.

[9]. Beňo, P., Schauer, F., Šprinková, S., Šimko, M., & Komenda, T. (2020). Road to Strengthen of Virtual Infrastructure and Security of Remote Laboratories on Trnava University in Trnava.

[10]. Chengyan, L. I., Feng, S., & Sun, G. (2020). DCE-miner: an association rule mining algorithm for multimedia based on the MapReduce framework. Multimedia Tools and Applications, 79, 16771-16793.

[11]. da Rosa Righi, R., Correa, E., Gomes, M. M., & da Costa, C. A. (2020). Enhancing performance of IoT applications with load prediction and cloud elasticity. Future Generation Computer Systems, 109, 689-701.

[12]. Daghistani, T., AlGhamdi, H., Alshammari, R., & AlHazme, R. H. (2020). Predictors of outpatients' no-show: big data analytics using apache spark. Journal of Big Data, 7(1), 1-15.

[13]. Dolores, M., Fernandez-Basso, C., Gómez-Romero, J., & Martin-Bautista, M. J. (2023). A big data association rule mining based approach for energy building behaviour analysis in an IoT environment. Scientific Reports, 13(1), 19810.

[14]. Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2023). New spark solutions for distributed frequent itemset and association rule mining algorithms. Cluster Computing, 1-18.

[15]. Gupta, Y. K. (2020). Aspect of Big Data in Medical Imaging to Extract the Hidden Information Using HIPI in HDFS Environment. In Advancement of Machine Intelligence in Interactive Medical Image Analysis (pp. 19-40). Springer, Singapore.

[16]. Gupta, Y. K., & Choudhary, S. (2020). A Study of Big Data Analytics with Two Fatal Diseases Using Apache Spark Framework. International Journal of Advanced Science and Technology (IJAST), 29(5), 2840-2851.

[17]. Haiyun, Z., & Yizhe, X. (2020). Sports performance prediction model based on integrated learning algorithm and cloud computing Hadoop platform. Microprocessors and Microsystems, 79, 103322.

[18]. Haji, L. M., Zeebaree, S., Ahmed, O. M., Sallow, A. B., Jacksi, K., & Zeabri, R. R. (2020). Dynamic resource allocation for distributed systems and cloud computing. TEST Engineering & Management, 83, 22417-22426.

[19]. Hosamani, N., Albur, N., Yaji, P., Mulla, M. M., & Narayan, D. G. (2020, July). Elastic provisioning of Hadoop clusters on OpenStack private cloud. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[20]. Kazemi, A., Keshtkar, A., Rashidi, S., Aslanabadi, N., Khodadad, B., & Esmaeili, M. (2020). Segmentation of cardiac fats based on Gabor filters and relationship of adipose volume with coronary artery disease using FP-Growth algorithm in CT scans. Biomedical Physics & Engineering Express, 6(5), 055009.

[21]. Kelvin, K., Cindy, C., Charles, C., Leonardo, D. P., & Yennimar, Y. (2020). Customer Churn's Analysis InTelecomunications Company Using Fp-Growth Algorithm: Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm. JurnalMantik, 4(2), 1285-1290.

[22]. Khader, M., & Al-Naymat, G. (2020). Density-based Algorithms for Big Data Clustering Using MapReduce Framework: A Comprehensive Study. ACM Computing Surveys (CSUR), 53(5), 1-38.

[23]. Koulouzis, S., Martin, P., & Zhao, Z. (2020). Virtual Infrastructure Optimisation. In Towards Interoperable Research Infrastructures for Environmental and Earth Sciences (pp. 192-207). Springer, Cham.

[24]. Koulouzis, S., Martin, P., Zhou, H., Hu, Y., Wang, J., Carval, T., ...& Zhao, Z. (2020). Time-critical data management in clouds: Challenges and a Dynamic Real-Time Infrastructure Planner (DRIP) solution. Concurrency and Computation: Practice and Experience, 32(16), e5269.

[25]. Kristiani, E., Yang, C. T., Huang, C. Y., Wang, Y. T., & Ko, P. C. (2020). The implementation of a cloud-edge computing architecture using OpenStack and Kubernetes for air quality monitoring application. Mobile Networks and Applications, 1-23.