10/14/2021

Distributed Association Rule Mining.

Student Name: Smit M Chaudhary.

Student ID: 2232612.

Fan ID: chau0221

Name Of Supervisor: PROF John F Roddick

Submitted to the College of Science and Engineering in partial fulfilment of the requirements for the degree of Masters Of Science (Computer Science) at Flinders University – Adelaide Australia.

"I certify that this work does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text."

Signed By:

Smit M Chaudhary Date: 14th Oct2021

Abstract

Currently, several sectors are using the advanced analytical approach to extract meaningful insight from the data. The major aspect of this research is to analyze that how communication infrastructure overhead affects the key of data mining processes. This research describes the rules of association that help to build the item set for the data modeling. These extracted meaningful insights can get the company a critical competitive advantage in the market and enable the company to become a more significant player in the market. The companies are, hence using various data analytics techniques such as data mining, data preprocessing, data wrangling, data warehouse, and data visualization. All these techniques enable the company to extract meaningful insight and enable them to improve their existing operating framework. Now, the core area of study in the current paper is to analyze how communication overhead affects a key data mining process. The process under consideration is the association rule-based approach. The critical analysis is focused on the concept of how it is possible to be reduced without causing any significant reduction inefficiency. This paper would hence help future analysts and data scientists to develop the best-optimized data mining approach.

Acknowledgment

This project would not have been possible without the support of many people. Many thanks to our mentor, who read my numerous revisions and helped make some sense of the confusion. Also, thanks to my committee tutors and faculty members, who offered their invaluable guidance and support.

Thanks to the institute for letting me participate in preparing a Dissertation like this one, providing me with the best possible academic means to complete this project. And finally, thanks to my parents and numerous friends who endured this long process with me, always offering support and love.

Contents

List of Figures

CHAPTER 1: INTRODUCTION

1.1 Introduction

The progress of science and technology has revolutionized human lives. It completely changed the social dynamics and the business domain. The advent of "information technology" has accelerated society and it has altered human behavior towards technology and along with that, IT has had an enormous impact on the business domain. Almost every business is currently adopting some form of technology to become more effective and efficient in the market (Agapito et al., 2018). The rise of IT and the evolution of digital technology have led to massive data consumption in the market. Currently, consumer data has become an asset for the business. "The data is the new oil in the 21st century". Almost every sector is using the advanced analytical approach to use the data and extract meaningful insight from which will enable the company to get a competitive advantage in the market and enable the company to become a more significant player in the market. The companies are using various data analytics techniques such as data mining, data preprocessing, data wrangling, data warehouse, and data visualization. All these techniques enable the company to extract meaningful insight and enable them to improve their existing operating framework. The strategy has been the initial building block of the interaction over the last few years. To achieve the success of the business, business models have to be very effective and constant. Effective business models tend to organize the system (Telikani et al., 2020). A good path in managing business processes is to generate a to-do list. Always maintain the improvement that has been presented in the previous business models.

This paper will discuss the data mining algorithm which is widely used in various domains to improve the existing operating framework of the company. The algorithm is known as the *Association rule mining*. This algorithm has a wide range of applications and this algorithm is mainly used in the retail industry and this algorithm is also known as the market basket analysis. This paper will discuss the various aspects of the algorithm and along with that, this paper will also discuss the practical implementation of the algorithm. This paper will discuss the effective ways that businesses can be produced using the algorithm.

1.2 Background of the study

Association rule mining algorithm" is a popular data mining technique in the industry that allows the company to look for the pattern in the collected data. Identifying the pattern enables the company to take proper measures for upcoming events and enables the company to stay ahead in the market. This algorithm has a versatile utility (Agapito *et al.*, 2018). This algorithm allows the company to look for the dependencies in the collected data and enables the company to identify the repeated data. If this repeated data exceeds the threshold number, then it is considered a valuable finding from the analysis. This algorithm is applicable in a wide variety of domains including health care, security, sports, E-commerce, fraud detection, database marketing, advertisement, bioinformatics, telecommunication, web, weather forecasting, and financial forecasting (Biswas *et al.,* 2018). Since 1994 the researchers are constantly working on this particular problem to improve the productivity of the organization using this algorithm. The first algorithm proposed was the Apriori algorithm to minimize communication and increase productivity in the company (Agapito et al., 2018). This was one of the popular algorithms. However, due to rapid growth in the distributed environment of computing, the researchers are constantly iterating there to improve the algorithm or create a new more effective algorithm. In 1996, the researchers created another algorithm. The main objective of this algorithm is to reduce the communication overhead. The algorithm is known as the CD algorithm which is also known as the count-distribution algorithm. The communication overhead is the proportion of time that users want to spend on communicating with the other instead of getting qualitative work. Communication is significantly required but as the size of the members increases, communication will also be overhead.

This algorithm is based on the distributed and parallel type association-rule mining method (Chahar *et al.*, 2017). To meet the objective of the algorithm, the researchers proposed the algorithm will reduce the cost of the infrastructure and it will eliminate duplicate computational infrastructure and they also indicated that this type of algorithm is suitable for high computational infrastructure. In the same year, another algorithm was proposed, the proposed algorithm is known as the First-distribution algorithm. It uses a new and innovative method to eliminate duplicate infrastructure. It uses global and local pruning techniques (Pang and Wang, 2020). Apart from all these algorithms, various algorithms have been proposed such as RSI and WARM. All these algorithms have demonstrated a positive result in reducing the communication overhead in the network and enabling the network to be more effective and efficient.

1.3 Research aim

This research aims to eliminate the duplicate computational There are two types of data mining processes, the first one is centralized and the second one is distribution-based. The first one is very simple, it has a centralized structure where the data is gathered from one particular resource. Due to this reason, this is a straightforward technique. This technique does not involve various forms of communication to collect the data. The second method of the data mining process is more complex and requires a communication approach to collect the data (Chengyan *et al.*, 2020). As the data is distributed and various resources are critical to understanding the consumer behavior and the market pattern. Hence to collect the data in a distributed framework requires an effective communication path. The distributed network mainly consists of the Sensor, devices, terminals, equipment, and computers (Rani and Pushpalatha, 2018). All these different devices are interconnected. Due to this reason, it requires high bandwidth to operate these devices, and minimizing the usage is a massive concern for the company. Hence the research aim is to minimize the usage of bandwidth in the distributed framework using the computational algorithmic approach.

1.4 Research objective

The computational framework contains the models, systems, and applications that give an overview of advanced systems that connect the gap among the practical efforts and frontline research. In addition, these objectives will enable the researchers to evaluate the effectiveness of research and enable them to form decisions on whether the research has been successful or not (Dong, 2020). The following segment will discuss the objectives of the research.

- To eliminate duplicate computational frameworks from the network.
- To minimize the "communication overhead" in distributed computation infrastructure.
- To reduce the bandwidth usage in the distributed computational framework of the company.
- To generate the most effective "association rule mining algorithm" to improve the efficiency in the data collection.

1.5 Research Questions

The research question will guide the readers and enable them to have more clarity on the information that they are looking for in this research paper. This subsection of the chapter is based on the research objective (Han et al., 2019). All the research questions of this segment are based on the research objective that has been mentioned in the previous section. All the research questions are mentioned in the following segment of this section.

1 What are ways to minimize the "communication overhead" in the distributed computational infrastructure in

the company?

2 Which algorithm is the most effective in improving the efficiency of data collection?

3 What is the most effective approach in implementing the "Association rule mining algorithm" in the network?

4 What are ways to eliminate duplicate "computational infrastructure" from the network?

5 What are the ways to minimize the bandwidth usage in "the distribution computational infrastructure" in the company?

1.6 Research Hypothesis

This hypothesis is based on the research problem that is going to be solved in this research paper. The hypothesis segment is going to discuss the problem in the smaller part which will enable the readers to actively engage in the research paper and enable the researchers to get more accurate information to form a critical understanding of the subject.

- ➢ H1: Communication overhead is a major concern in the distributed computational infrastructure
- ➢ H0: Communication overhead is not a major concern in the distributed computational infrastructure
- ➢ H1: Effective usage of association rule mining algorithms can minimize the communication overhead in distributed computational infrastructure.
- ➢ H0: Effective usage of association rule mining algorithms cannot minimize the communication overhead in distributed computational infrastructure.

1.7 Research Rational

With the rise of distributed computing and cutting-edge technology. It increases the massive data consumption among the consumers, which has led to lots of new opportunities in the industry. The rise of IoT technology and machine learning technology has enabled companies to become more effective and efficient in the market and enable them to improve the customer experience (Agapito et al., 2018). However, with all these positive aspects, these new methodologies and technology have also brought new challenges in the business domain. As the technology is almost available to everyone and the business can no longer claim that adopting IoT and AI technology will give them a competitive advantage (Raj *et al.,* 2020). However, this implementation and using optimized algorithms to improve the consumer experience have the potential to increase the effectiveness and efficiency of the company and enable the company to become a dominant player in the market.

This particular aspect is coming into the critical discussion since communication is the major concern among the components in technologies like IoT and AI (Telikani et al., 2020). It is not a question of getting the based autonomous computation but also getting the best real-time analytics. This is because the applications of these are also among the fields like healthcare, natural disaster, defense, construction safety, and so on (Raj *et al.*, 2021). All of these require an accurate prediction and action based on real-time data from a large collection of datasets. Thus, synchronization among various computational tasks and components becomes a dire need. But in this regard, communication stress on the computational resources becomes a critical thing to be ignored or reduced as much as possible. Otherwise, both the energy and cost spent in the process would enhance drastically with the increase in complexity of tasks.

This research significantly focuses on implementing advanced algorithms to analyze the business processes and their effectiveness. The effectiveness of the business makes it very productive. This research will analyze the business modules that overhead the previous network and permit them to retrieve effective data for the various business approaches. By measuring the bandwidth usage in the many decentralized and distributed technologies like the internet of things and computational infrastructure. These technologies will help to increase the organizing the business models.

1.8 Research significance

The progressing mindset is the basis of all scientific discovery and technological advancement. All the research has been based on the necessity to find out the more optimal solution for the existing problem in the market. This research paper has a similar intention. The technological advancement has given various new opportunities in the business domain various new industries have been forming which are based on the advancement of technology (Leung *et al.*, 2017). The increased opportunities of the systemic behavior for business models, the technological advancement generally used this for increasing computational infrastructure.

However, with the rise in opportunities, there are various challenges as well that have the potential to hinder the progress of scientific discovery and technological advancement (Telikani et al., 2020). Due to these potential challenges, it is an essential step to find out the root cause of those challenges in the market and improve the existing methodology which will enable the company to further improve its consumer experience.

This research signifies the constant improvement to increase the efficiency in the network and enable the users to have a more effective data collecting system which further enables them to get accurate meaningful insight into the data (Liu *et al.*, 2018). This research will enable the company to become more effective in data mining methodology and in addition, this research result will enable them to increase complexity in collecting data which will enable more transparent and simple techniques and enable the company to increase their productivity.

1.9 Research Framework

The key points needed to be covered in the context of the thesis paper are as explained in the framework below:

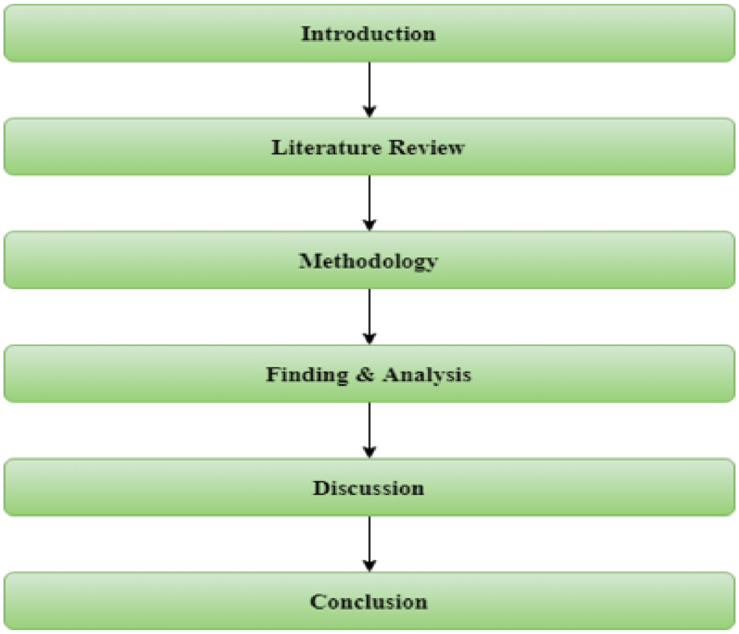

Figure 1: Research Framework

(Source: Song *et al.*, 2017)

1.10 Research structure

In this chapter, a detailed background analysis is followed in the field of the research so that tactical objectives, research questions, and hypotheses are possible to be achieved.

- **Literature Review:** In this chapter, the key aim is to explore several key literature and past studies to reach a conceptual framework for the study.
- **Methodology:** The key aspects of the method being adopted by the research to reach a conclusive analysis are addressed in this chapter.
- **Finding & analysis:** In the finding & analysis part the key findings from the study are explored.
- **Discussion:** In this chapter, the key understandings developed from the findings in the previous chapter are explored
- **Conclusion:** In the conclusion, the alignment of the research with the objectives set, the key recommendations, and further scopes of future research are explored apart from the key conclusion of the research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The literature review chapter will enable readers to get detailed information about the various aspects of the research on distributed association rule mining which will form a critical understanding of the subject. This chapter will also provide information about the various aspects of the research on data mining and association rules for the business models which play a critical role in conducting the entire research for communication overhead and computational framework (Agapito et al., 2018). The entire research is based on the existing literature on a similar subject. This existing literature is published by various authors. All these relevant and existing literature will be studied to make the initial foundational knowledge on the subject. Along with that, this chapter will discuss various relevant theories and models which will play a critical role in creating the subjective evaluation of the gathered information based on the evaluation the information will be documented in this chapter. In addition, the gaps in literature will be documented from this research. Furthermore. The conceptual framework will enable the readers to understand the framework and enable them to get the insight of the dependencies about variables and finally the conclusion section of this chapter will highlight the findings of this chapter.

2.2 Empirical study

This segment of this chapter will review the various existing literature on this subject. This segment will discuss the basics of the research, the method of the research, and the findings of the research, this approach is essential to form an initial foundational understanding of the topic. This segment will review the various relevant segments of the subject which will form a foundation of critical understanding of the subject.

According to Sashi, "*A spark-based Apriori algorithm can be an efficient choice to reduce the communication overhead*". According to the author, this approach is one of the earliest approaches to data mining. This approach was proposed in 1994. Adopting the Apriori technique for massive data has been attempted numerous times. "*However, Apriori's inefficiencies, such as continuous scanning of the input pattern, a compilation of all potential frequent item set before measuring their support value, and so on, diminish Apriori's efficiency for massive data*" (Prajapati *et al.*, 2017). Even decentralized and simultaneous Apriori solutions employing the MapReduce architecture perform poorly when the data size is huge. This is owing to the algorithm's exploratory approach, which results in significant disc latency.

2.2.1 AI-based communication networks

The input data, which is stored on storage, is examined for each cycle, resulting in significant disc I/O. Due to the "*in-memory computational approach*", Apache Spark implementations of Apriori have greater speed. It accelerates repeated database scans by storing the data in a storage layer called resilient dataset (RDD). Datasets are stored in RDDs as key-value pairs dispersed among clusters (Patel *et al.*, 2018). During the execution of RDD activities, such key-value pairs must be redistributed across clusters. "*The shuffling or redistribute procedure has overhead in terms of communication and synchronization*".

In the studies of Dasgupta and Saha (2020), it is explained that data mining is the autonomous extraction process of large datasets. Now, data mining is a key technique and area of study when it comes to the application of AI-based communication networks. In the literature association rule is a special kind of data mining technique that involves the mining of data that are associated with each other through any kind of relationship (Sinaei and Fatemi, 2018). Relationships between different sets of data can be found and extracted through the application of this particular technique. The fundamental aspect found to be important in the basic concept of association by the authors are the steps of the association algorithm which are:

- As per the user's defines the minimum confidence and support is to be set.
- Constructing the C1 (candidate 1-itemsets) and then pruning the item sets based on lower values of support than the set value.
- Join item sets to create C2 (candidate 2-itemsets) and prune the infrequent item sets.
- Repeat the 3rd step to Ck item sets until there is no possibility to create C(k+1) item sets.

Now, the key aspect of this research is the various techniques that are involved in the field of association rules. The literature itself has explored one of such algorithms: the Apriori algorithm. It is a kind of parallel association technique where a few parallel algorithms are used together. The implementation is applied to Google's MapReduce model which is a paradigm of distributed programming and implementation association (Sawan and Shah, 2018). Without much experience, programmers can use large distributed resource systems here. The tool that the researchers used is the Hadoop platform which is a distributed file system known as HDFS. It helps in distributing the data files across several servers along with distributing the running job itself near the associated data. Finally, a detailed view of the efficiency and effectiveness of the algorithm is explored in the study in detail. It is found that in terms of association rule Apriori is a simple technique that requires a minimum resource allocation. But it has a crucial drawback in particular. It is that whenever the complexity of the dataset increases with the increase in associated item sets the accuracy decreases. With the large number of distribution algorithms used in parallel (data distribution, candidate distribution, data distribution, and so on) that communication becomes a key concern here. So, to achieve an effective synchronization among the nodes more stress on communication is required to be given (Song *et al*., 2017). For instance, the communication data packets being transmitted among the nodes and file systems would be pretty much effective in achieving this. Thus, a critically synchronized array of networks is possible to achieve that would help in getting a more accurate association even if the number of associated datasets increases.

According to Khedr *et al*., (2020), communication technology being at its peak of rapid advancements IoT has become the most prevalent innovation of all. It is seen to apply to a vast range of sectors in daily life. The applications are much diverse from medical applications and communication to industrial applications like safety applications in construction or cyber-physical systems in production sectors and even in defense. The entire system of sensors acts as a standalone system which is the main framework of IoT (Wang and Zheng, 2020). Now, for this, a continuous flow of real-time data is the dire need of the system from respective sources like natural disasters, weather, battlefields, construction sites, health monitoring, and so on. As the scale of operations become larger and larger and more complicated, a wireless sensor network (WSN) becomes the only option for this network infrastructure.

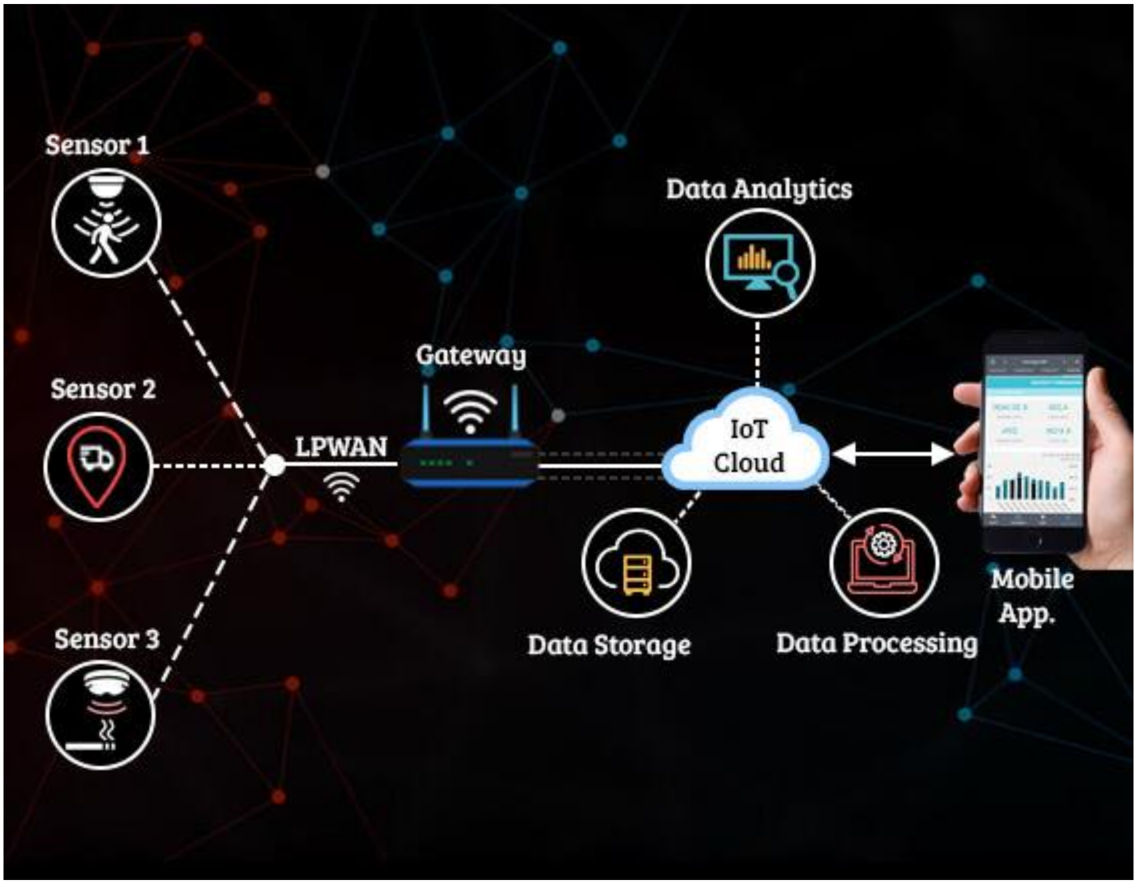

Figure 2:: WSN integration in the IoT cloud architecture

(Source: Song *et al.*, 2017)

2.2.2 Global association rule

As per the study, IDC (International Data Corporation) has pointed out that by 2025 real-time data being communicated would reach a value of 175 Zettabytes which would be a rise of 30% in terms of the amount in 2018. This is exactly where the need for the reduction of communication overhead comes into play. Regarding this, the Association rule-based mining algorithms have achieved a major interest in the field. In the studies of Khedr *et al.*, a distributed association rule is explored to analyze the implementation of WSN. The key schematic flowchart of the process can be shown as follows:

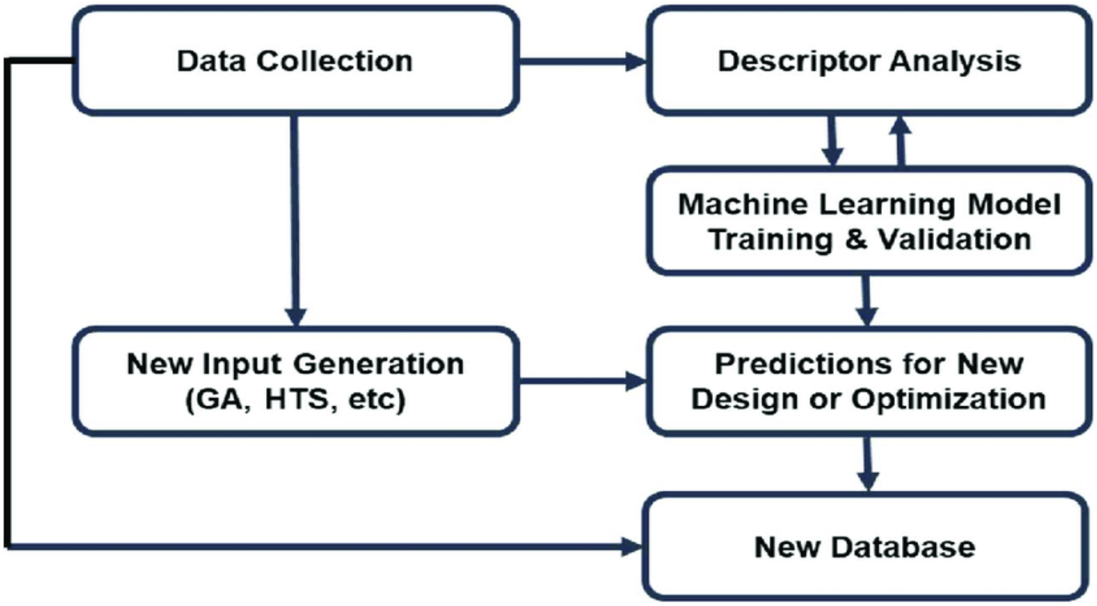
Data Collection
Descriptor Analysis
Machine Learning Model Training & Validation
New Input Generation (GA, HTS, etc)
Predictions for New Design or Optimization
New Database

Figure 3: Schematic flowchart of the process

(Source: Zhang *et al.,* 2019)

The main concentration of the scheme used in the study is focused on a global association rule where implicit global database D is the key concern. The entire global task of computation is distributed over several sensor nodes. The statistical outcomes are then collected and communicated through the association rule. Each of the Cluster heads collects requests from the base station which are then communicated to the cluster heads then. A shared association is created between the cluster heads and cluster members that promote local computations. This entire aspect helps in assuring that the number and size of messages in between nodes are minimal (Zhang *et al.*, 2019). Thus, the utilized energy is at a minimum. The researchers have focused on keeping the computational association confined to the receiving or sensor nodes only. The key aim is to achieve the best possible solution for the achievement of maximum computational resources as compared to communication overhead. With the help of the confinement of the critical computational tasks to the sensor nodes, it is possible to reduce the amount or size of transmission packets in between nodes (Agapito et al., 2018). The only communicated data are the statistical or numeric ones that help in reducing the complexity of data and, in turn, the size of the packets as well. This can even optimize the energy being used by the entire network system as well.

2.2.3 Demand for extracting patterns

According to Garg, the demand for extracting patterns from big data is constantly rising, due to this landscape change in the industry. The data mining methodology has become a relevant practice in the industry. This methodology has been proved to be an effective addition to the existing business framework which will allow companies to get a competitive advantage in the market. However, several challenges are found in the implementation of this algorithm and the rise of distributed computing has become a major matter of concern which proves to be a continuous bottleneck. Currently, the company will face two challenges while implementing the first one is the memory allocation while collecting the data and another one is the computational time for each computational system.

As the distributed computed infrastructure consists of various devices (Varma, and LijiP, 2017). It increases the average communication time and decreases the effectiveness of the system and the entire network performs poorly. According to the author, the poor performance of the system and the network can be improved by using a more effective approach. The author stated using the CD algorithm and that using the FDM, the algorithm can be an effective solution for this problem and enable the distributed system to become more effective for the company (Telikani et al., 2020). Combination of these two algorithms and calculating the frequency of each computing system will enable the stakeholders to get accurate data and help them to make a more optimized system which will help them to get more benefit. In addition, using the MapReduce algorithm. This algorithm is the most effective way to break down the bigger task into a smaller task and enable the users to complete the smaller task with ease and along with that it also helps the users to reduce the redundancy from the system.

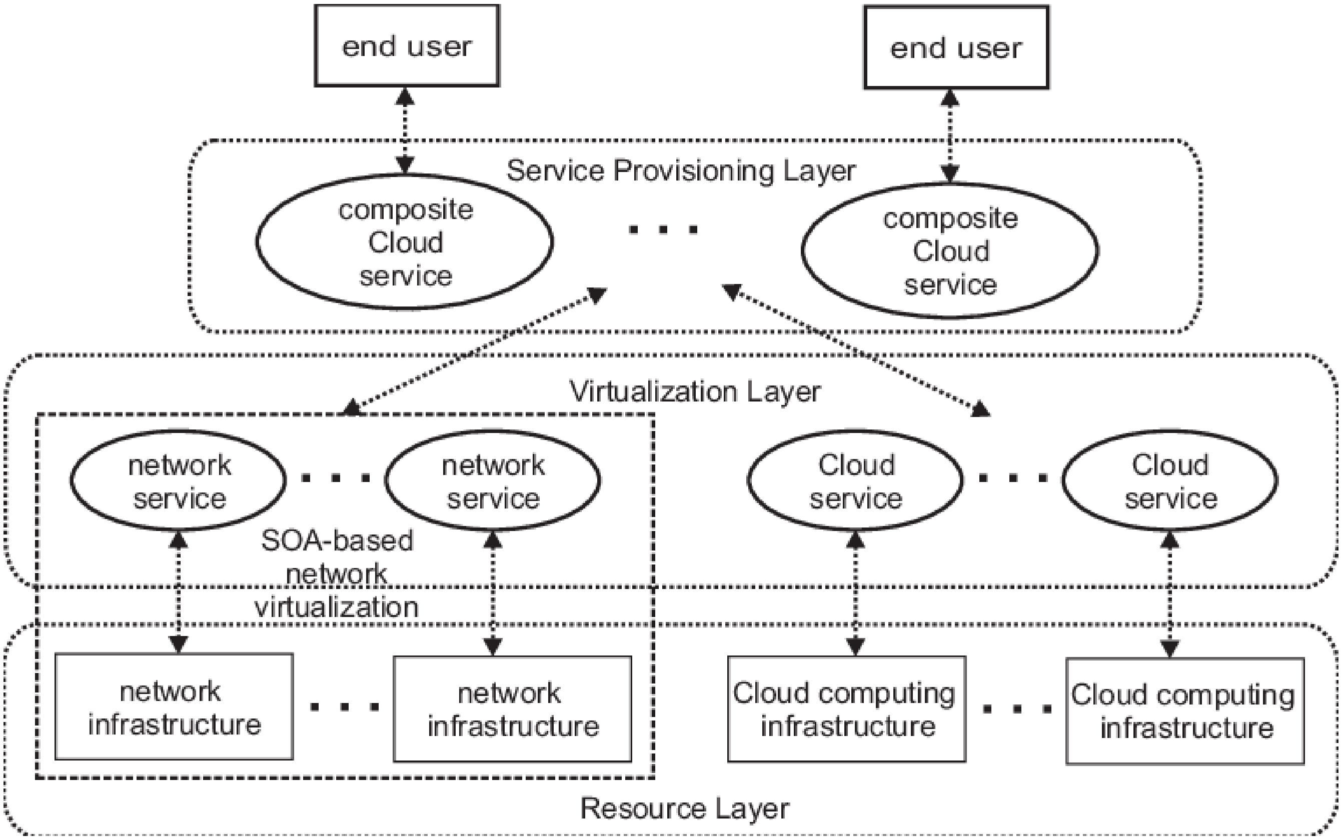

Figure 4: Proposed Framework

(Source: Sornalakshmi, *et al.* 2021)

The above figure demonstrated the proposed framework by the author which will enable the users to reduce the communication overhead in the distribution computational infrastructure. Due to the nature of data and as the data consumption is constantly increasing the framework will work in two phases. The above figure is the first phase of the entire system. The first segment of the framework is data preprocessing. Data preprocessing is an essential element of the evaluation process (Sornalakshmi *et al.,* 2021). The preprocessing enables the researchers to eliminate the null value and enables the researchers to create a more cleaned dataset that is suitable for further calculation and enables the researchers to get quality insight from the dataset. The large dataset often contains incomplete data and along with that it also contains the various null values in the dataset. Using the MapReduce algorithm the dataset will be preprocessed and the dataset will be prepared for further calculation (Agapito et al., 2018). The MapReduce algorithm contains various mathematical logic such as sorting, searching, and indexing. This will allow the researchers to organize the data before sending it for the calculation and the other method of the MapReduce algorithm enables the researchers to get a more meaningful dataset that issuitable for further calculation and enables the researchers to extract the hidden pattern in the dataset. The MapReduce algorithm will work in multiple phases.

2.2.4 Phases of the MapReduce algorithm

The first phase of the MapReduce algorithm is known as the “Mapper setup” and the second phase of the algorithm is known as the “Reducer setup”. The first phase of the algorithm is responsible for creating map steps. The map step will be created using the Hadoop framework. This feature is a popular tool for handling big data (Telikani et al., 2020). The entire database will be split into smaller chunks of data. The segmented dataset will be given to the “Hadoop distributed file system” which is shortly known as HDFS. The size of the segmented database is based on the configuration of the software framework and the entire size of the database. The framework will enable the users to build the various clusters and based on the cluster the users can make further calculations. The main objective of this first phase is to identify the key elements of the data for example if the retail industry data is obtained, then a segmented database will segregate the important detail such as “Distributor detail”, “Sales date”, “Retailer code” and “Area code” (Urmela, and Nandhini, 2017). Once this list is collected then this phase lists all this important data as per the relevance.

The second phase of the algorithm is the “reducer setup”. Once the mapper phase is completed, then the reducer phase takes place, this is the in-built function of the Hadoop framework. The output of the first phase is provided to this phase. The “key-value” pair output from the mapper phase has fed to this phase to obtain the new output. In the new output, the key-value pair is created autonomously (Mlambo *et al.,* 2018). This method of autonomous generation of the key value is also known as the shuffle step. This step enables the DB to split the key value into two parts and further enables the users to transform the normal DB into the “transactional DB”. Once a DB is converted then the null value is eliminated from the new DB to provide a more complete DB.

The next step is to look for the null transaction within DB, once the DB is converted then it is important to eliminate all the null values from the DB and enable the users to have a more valuable and authenticated DB which is more suitable for further analysis. Otherwise, the null value can hinder the performance of the analysis and provide false results which will have a drastic effect on the forecast and lead to catastrophic results in the market. The next step is an implementation of the combination of the two algorithms. These algorithms are “CD and FDM” (Hu, and Chen, 2018). The objective of both these algorithms is to provide insight into the frequent data from DB. According to the author, both these algorithms are an excellent choice to minimize the communication overhead.

However, both these algorithms have a different approach to meet their objective. Due to this reason, this hybrid approach has the potential to mitigate the existing problem and has the potential to establish itself as an innovative solution in the computational world. The CD algorithm mainly works on the subset of the entire dataset. Each of the subsets is allocated to the specific node. Each node communicates with each others. The objective of the nodes is to count the allocated key value of the segregated dataset. Each node has the individual count and the global count. The global count enables the algorithm to perform the next iteration. On the other hand, the FDM has a similarity with the apriori algorithm (Rochd, and Hafidi, 2018). In this algorithm, the nodes are used as the communication medium. The nodes are responsible for communicating the information to different nodes. To reduce the size of the resultant dataset, the “global and local” itemset are frequently used at every iteration. Once the resultant transactional dataset is created then reduction methodology is applied to the resultant DB and it further eliminates all the duplicate values from the DB.

2.2.5 Apriori algorithm

The name Apriori is taken because it uses the previous information of the consecutive properties of the item set. Users can apply the Apriori algorithms in developing the many business models as a level-wise approach or iterative approach. To enhance the effectiveness of the level-wise approach, a significant property is used that is called Apriori algorithms. This algorithm defines that all non-empty item sets of consecutive item sets should be frequent. The key aspect of the Apriori algorithms is their anti-monotonicity to help to increase the measures.

2.2.6 Data mining approach

According to Chauhan, data mining is the essential approach to stay relevant in the market. Data mining and analysis have become an integral part of the business and it completely changes the landscape of every sector. According to the author, the CD and FDM algorithm is useful however, the author in his studies have observed their major drawbacks while using these two algorithms. These drawbacks are, both the algorithm creates enormous resultant Db as the large DB is created it also has various nodes. Due to this reason the communication delay occurs, as a large number of the nodes are present, so to communicate with each node, it requires frequent massage to maintain the entire communication structure among the other nodes. Due to this reason, the execution time increases and it fails to minimize communication overhead (Shunmugam *et al.,* 2018). However, this is the foundation of an optimized solution (Agapito et al., 2018). According to the author, the problem can be solved using the DFPM algorithm, this algorithm is also known as the "Distributed frequent Pattern mining" algorithm. This algorithm has similar steps to implement.

All the data preprocessing steps are similar as this algorithm also uses the MapReduce algorithm. MapReduce uses a similar multiple-phase approach to eliminate the null value and convert the entire dataset to a transaction dataset. However, the main advantage of using this approach is that the nodes within the algorithm do communicate with others, the main objective of the nodes is to count value and exchange the count value within the other nodes. Due to this reason, this approach is more effective and efficient than the other technique. C denoted as the candidate or the resultant DB (Telikani et al., 2020). L here in this algorithm is denoted as frequent occurrence (Miholca, 2018). K denotes the step requirement. The algorithm will start its operation from an itemset and it will generate k+1 resultant DB. In addition, the "transactional data" will be allocated a specific input. The reduction job will end the calculation when it is unable to create larger candidates or resultant items. Variations of the minimal support criterion are used to generate frequent items. For such purposes of producing "*association rules*", the "decision-system" checks for the proper support to collect the data. Once the proper support is gathered from the dataset then the algorithm can be created using the following steps.

- Construct all non-empty subsets of l, for every frequent pattern, l.
- For the output of each non-empty subset, the rule is "s → (l - s)", "(*Support(l)/Support(s)) ≥ min_conf*". The "*min_conf"* is known as the "*minimum confidence threshold*".

According to the author, if the required threshold is achieved then it will create more algorithms that will enable the users to reduce the communication overhead, enable the company to have the more optimal solution for the company to get the desired result, and further enable the company to become more impactful in the industry (Puścian, 2019). The rules contain all the relevant parameters which are crucial for business operation. Based on the relevancy and priority set by the users the rule will enable the users to get more optimum results and the solution which will allow the company to become more effective and further enable them to become the significant player in the market.

2.2.7 Deep learning module

The technical experts have proposed the inclusion of the deep learning module to optimize the result enable the company to get more advantages using this algorithm, the deep learning algorithm will enable the company to automate the entire process which will help the company to get a slight competitive advantage from the rival company which does not use the deep learning module. Once the first phase of the framework has been implemented then the phase will be a useful option for the company as this will allow the company to automate the task and enable the company to look for the loopholes in the framework which will enable the company to improve their existing operational framework. The deep learning module will consist of the three different parts which will allow the operator to meaningful insight from the existing clean dataset and enable the company

(Nadimi-Shahraki, and Mansouri, 2017). The first part or the layer is the input layer this layer is responsible for collecting data. The dataset will be collected from the operator. The operator will feed to the network preprocessed dataset. Once the preprocessed dataset is given to the input then the dataset is sent to the hidden layer or the computational layer. This layer is hidden and this is an internal layer. This layer is responsible for implementing the appropriate mathematical logic or the advanced algorithm which will enable the operators to extract the meaningful information from the dataset enable them to understand the certain behavior of the element and help them to understand the root cause of the "*communication overhead in the distributed computational infrastructure"* which will further help them to come up with a more robust solution using this approach and enable the company and industry to be more effective. Once the advanced algorithm is applied and all the computation is performed on the dataset. The resultant dataset is sent to the outer layer (Qian *et al.*, 2021). The outer layer is known as the output layer (Telikani et al., 2020). This layer is an external layer. This layer is responsible for presenting the data and enabling the researchers to get meaningful insight from the dataset. Once the resultant model and the dataset are collected then the resultant dataset is sent to the various data visualization tools. These tools enable the key stakeholders to understand the visual representation of the data which further enables them to make proper decisions that will enable the company to become more effective in their operational framework and enable them to become a significant player in the market (Agapito et al., 2018). Within the neural network analysis, the researchers will apply the MapReduce algorithm before sending the data to the input layer. This algorithm will use multiple functions to provide a more effective dataset. Then this dataset will be sent for further analysis.

According to Wang, "*association rule mining*" or ARM is the most important algorithm to get meaningful information from the data. The author has emphasized the quantitative analysis approach to find out the communication among the other nodes and an innovative approach to minimize the computation. Further, the authors have started using the apriori algorithm can be useful. However, the author has emphasized using a more enhanced version of the algorithm to mitigate the drawback of the algorithm (Tehreem *et al.*, 2017). According to the author, the enhanced version of the apriori algorithm will be combined with the machine learning algorithm for implementation and to understand the behavior between other nodes.

This approach will enable to identify the relationships among the nodes and the communication pattern and based on the result the researchers can make effective steps that will allow them to mitigate all the problems in the previous algorithm. The authors have emphasized using the DBSCAN algorithm, this algorithm is known as "*Density-based spatial clustering of applications with noise*". This algorithm will enable the researchers to extract the meaning information and enable researchers to look for the element which is influencing the certain behavior in the communication. One of the benefits of using this particular algorithm is its adaptiveness to the different types of datasets. Due to the versatility, it enables researchers to have the flexibility to conduct the entire analysis and enables researchers to make the best decision which will increase the chances of the successful analysis and efficient result. The entire analysis will work in three distinctive steps. These three steps are elaborate and distinctive which allow the researchers to extract meaningful information from the dataset. The first step is data preprocessing, this step consists of the "cleaning and normalization" of the dataset (Jiang, 2020). Once the dataset is cleaned and preprocessed then the dataset then the DBSCAN algorithm will be applied to extract the communication data among other nodes. In the subsequent steps, the clustering algorithm will be applied to generate the cluster from the database. This cluster will enable the researchers to identify the pattern from the database. The "*K-means algorithm*" will be applied in the database to understand the frequency of the data pattern. This will allow the researchers to frequent data and enable the researchers to notice the pattern. Based on the result, this will allow researchers to implement a regression algorithm once the regression algorithm is applied (Telikani et al., 2020). It will enable the researchers to look for relations among the various elements in the database which influence the several factors in the communication among all the nodes and enable the researchers to extract the root cause of the communication delay which will enable them to make the final steps of the proposed framework.

The final step is involved in applying the "*advanced apriori algorithm*" (Luna *et al.*, 2019). This algorithm will allow researchers to create more optimum pathways to reduce the communication delay and enable the entire system to become more robust and resilient while operating and enable the company to become more effective in its operational framework.

2.3 Theories and models

2.3.1 Internet of things

The Internet of Things is the system or network of different physical objects that are embedded with several sensors and software systems to communicate via data exchange among each other. It is a consequence of the evolution of several fields of technology like real-time data analytics, embedded systems, machine learning, sensors, and communication technology. The convergence of all these arrays lets wireless sensor networks, embedded systems, automation, and control system technologies converge into creating the technology of IoT (Rong *et al.*, 2018). Currently, it has a wide range of applications like construction project monitoring, health monitoring, defense applications, process monitoring, and so on. This has caused a large array of real-time data analytics to be brought into the field of applications of IoT. This is exactly where effective algorithms are needed to be considered for the enabling of the IoT systems to accurately analyze data with effective integration of rules.

2.3.2 Association rule

Association rule is a data mining algorithm that helps in setting the parameters or measures of association in between multiple datasets. Any two or more non-related or independent datasets can be converted into associated ones with such an algorithm. Simple conditional operators like if-then can be used for the association purpose effectively when it comes to the context of association rule. Usually, datasets explored by machine learning systems are statistical or numerical ones. But the sheer advantage of association rules is that they can even handle categorical or non-numerical data sets (Wang *et al.*, 2020). Association rule-based mining simply focuses on the association and correlational pattern in different databases. These databases can be transactional, relational, or of any other form. There are two crucial parts of such a mining algorithm which are:

1. Antecedent or the if part
2. Consequent or then part

The association rules are, hence, to be created or formed based on minute observation of dataset patterns and frequent observations of the Antecedent-Consequent association. As the minute observation is done two particular functions are used for allocating the association rule. These two are:

- **Support:** this function is the measure of how frequently the association is there in between the datasets
- **Confidence:** this function is the measure of how many times the relationships are observed to be true

Now, in the context of IoT communication strategies this data mining strategy helps in associating different databases from different nodes. In the context of IoT-based communications, there are several sensor or input nodes that help in collecting real-time data at different input points (Telikani *et al.*, 2020). If proper association rules are possible to be put in place in case of the mining process of IoT data this will help in the embedding of these separate nodes in the collective system. Acting as a standalone system is the main framework that IoT models follow in particular.

2.3.3 Communication overhead

The communication overhead is another concept that is a must to be explored in the context of the current study.

This particularly involves not only technical but also non-technical concepts. In terms of simple communication theories, a communication overhead refers to the proportion or fraction of effort or time being spent for communication in a team concerning the productive time (Zhan *et al.*, 2019). Now, for a well-organized collective function to be done communication plays a crucial role. But as the complexity or size of the team and tasks increase, communication overhead gets larger, otherwise, proper centralized control is impossible to be achieved. In the context of communication technology also this aspect is very much similar (Verma *et al.*, 2020). be transferred in between two nodes in a communication network is known to be communication overhead (Telikani et al., 2020). This can include the overhead of packet preparation, routing process, and routing table. In network-based computing systems like IoTs communication overheads are crucial as they maintain the association among the nodes which are required to be kept embedded. In an efficient communication network to achieve an effective synchronization among the nodes, more stress on communication is required to be given. For instance, the communication data packets being transmitted among the nodes and file systems would be pretty much effective in achieving this (Wang *et al.*, 2018). Thus, a properly synchronized array of nodes is possible to achieve that would help in getting a more accurate association even if the number of associated datasets increases. But with an increase of these overheads the available energy and computing resources associated with the key computational tasks which are the main productive tasks would get reduced significantly. So, an optimized state where the communication overhead is kept as minimal as possible without hampering the association between nodes is critical to be achieved.

2.4 Literature gap

The research is based on adequate information. All this information in this research paper has been documented after extensive research. The research has demonstrated the three particular trends in using cutting-edge technology such as “machine learning and AI technology” to implement and extract the between various nodes and reduce the communication from the system network which will allow researchers to improve the communication delay. However, while conducting this research it has been observed, Implementation of this cutting-edge technology requires extensive computational power and high expertise to conduct the analysis. This is a bottleneck for various companies in the industry. Using cloud technology to mitigate this problem can be a solution (Yan *et al.*, 2017). However, there are few research papers that have been published to verify this claim. Hence, it can be a potential solution in the market that will allow the company to remove the bottleneck from the existing solution.

2.5 Conceptual Framework

The conceptual framework plays a critical role in the case of any study apart from the research objectives, questions, and hypotheses. This is a graphical representation of the association of variables involved in the study. Now, the detailed literature study has helped in finding a few key variables (both dependent and independent or DV and IV) based on the key objectives of the research. The main objective of the study is to reach the optimization of communication overhead. So, the DV being associated with the current study would be the optimization of communication overhead. The next aspect is to identify the independent variables that are associated with the study (Zhang *et al.*, 2019). The first crucial variable that is to be explored for the research is already clear from the key objectives. It is the association rule itself that would be needed for an optimized state to be achieved. Now, it is to be remembered that the rules must be so efficient that the communication overhead is kept as minimal as possible without hampering the association. The next crucial parameter involved in the study is the allocated resources for the association.

For an association to be in place the minimum possible number of nodes that are allocated in the context is two.

But here a critical aspect is the number of computing resources that are used for the communication (Telikani et al., 2020). It is already clear that with the increase of the complexity of the network the number of resources allocated also increases. But then again, the more resources are allocated the less the available resource for key computation of available datasets. So, the consideration of allocated resources is also a key area of analysis in the study (Agapito et al., 2018). Finally, another key variable to be considered is the energy being utilized for the purpose. An optimized state would not only utilize the least number of resources but also the least amount of energy. All these variables are crucial if the conditions of an optimized state are to be explored in the study.

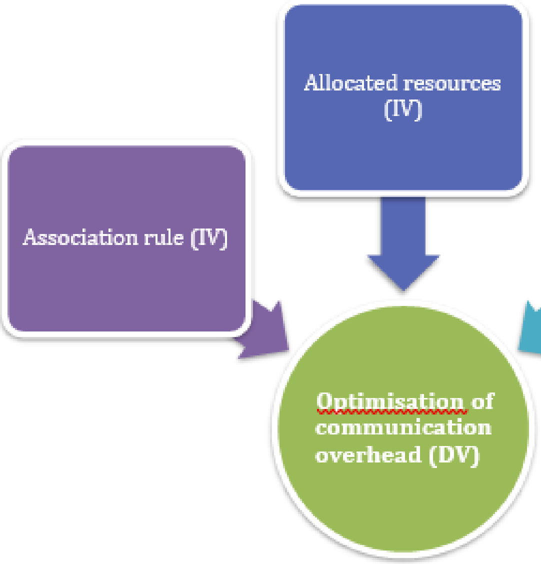
Allocated resources (IV)
Association rule (IV)
Energy being utilised (IV)
Optimisation of communication overhead (DV)

Figure 5: Conceptual Framework

(Source: Zhang *et al.*, 2019)

2.6 Conclusion

The literature review chapter has shed light on the various aspects of the research and along with that, this chapter has been based on various research papers that have been published in the past. All the relevant research papers have been studied to form an initial understanding of the subject which will enable the readers to understand the various aspects of the paper and enable them to become more aware of the subject and help the readers to form a critical understanding of the subject (Agapito et al., 2018). Apart from the empirical study section which will enable the readers and the key stakeholders to form a foundational logic. This chapter also consists of the various subsections which will enable the readers to get knowledge about the various other related fields such as IoT, machine learning, and AI. Which will help the readers to connect the entire solution and increase their understanding of the subject. Furthermore, the literature gap section will enable the readers to understand the current gap in the market which has been observed while conducting this research. This subsection will help other researchers to look at the entire research paper from different points of view and enable them to get new and fresh ideas to mitigate the existing problem.

CHAPTER 3: METHODOLOGY

3.1 Introduction

The methodology is always a critical part of any kind of research. This is because it defines the key conceptual aspects that are required to be considered by the researcher while conducting the research. The major virtues, procedures, and assumptions, all are covered in the paradigm of methodology. Now a research onion can be considered asthe key framework that can elaborate how to define the methodology layer by layer. So, in particular, the key areas that are going to be covered in this chapter are philosophical assumption of the research, research approach, research design, data collection, and analysis method. Apart from these, ethical considerations also play a major role in the methodology of research.

3.2 Research Philosophy

The first critical layer of the methodology is the research philosophy that is required to be evaluated for the study. Now regarding this, there is a total of three philosophical paradigms that are possible to be considered for a tactical research methodology. These three are positivism, interpretivism, and pragmatism. The interpretivism philosophy considers an observant to be the interpreter of the outcome that is reached through the analysis of the data collected. But in the case of the positivism philosophy, the data analysis is the key result itself and no observant interpretation is necessary for the meeting of the objectives (Qiu *et al.*, 2017). In case of the pragmatism, the suitable action is according to the need of the research i.e., whatever the research objective demands are the key concern of the philosophy. It is kind of a practical way out for concluding.

Now, the key concern here would be how to reach the philosophical assumption that the current research would require. Which philosophy should the current research study follow for meeting its objectives will be decided by the key aim of the study? The key aspect that is required to be covered in the critical discussion is the area of communication. So, it would be best if a rather interpretive way is kept that would help in overviewing the previous studies in the field with a more critical approach (Djenouri *et al., 2018*). This means a tactical interpretivism philosophy would be the best choice for the study. Thus, the selected philosophy for the study would be an interpretivism philosophy.

3.3 Research Approach

The approach of the research is the next critical layer of the methodology. It defines how the conclusion of the study is to be reached or the path that it follows. Now, two approaches, in particular, are, to some extent, opposite to each other. These two are the deductive and inductive approaches. The deductive approach requires the study to presume some hypotheses that may or may not be accepted as per the research. These hypotheses are established from the research questions, objectives, and the Conceptual variables of the study. Also, there are null hypotheses assumed for the case if the key hypotheses turn out to be unacceptable (Khedr *et al.,* 2019). In this case, the data analysis is done to test these hypotheses to be acceptable or not. But in the case of an inductive approach, no presumptions are made and the conclusion is reached directly through the analysis. No presumed hypotheses are required to be considered by the researcher hence.

For reaching the most suitable choice of research approach the overview of the objectives is important. Now, in this research communication overhead is the major consideration among several components in technologies like IoT and AI. It is not about getting the best autonomous computation but also getting the best real-time analytics

(Agapito et al., 2018). This is because the applications of these are also among the fields like healthcare, natural disaster, defense, construction safety, and so on (Varshney, 2017). All of these require an accurate prediction and action based on real-time data from a large collection of datasets. This indicates that presumed hypotheses regarding the aspects of communication overhead are possible to be made. Now, as the hypotheses are possible to be made assuming these would help the research to get a particular direction. This is the reason the research is following a deductive approach where particular hypotheses are set in the research to be followed.

3.4 Research Design

Research design is critically important is one of the key core concerns in the research methodology. As the philosophy and approach help in assuming the key conceptual parameters of the study method, now is the time to ascend to the main core design of the study method. In the case of the research design, the main method that is to be followed by the research is to be explained. Now, there are various kinds of research designs that are possible for any researcher to follow. This can be a survey, an interview, a case study, cross-sectional research, a systematic review, a ground action theory. Now, in the current research area, the researcher has chosen that a ground action design would help in ascending to the key conclusion. Major hypotheses that are to be tested are already explored in the study. Now, based on these the study is to critically review various previous studies regarding the association rule algorithm and communication overhead. The core correlation between these two is very important to be analyzed so that the right direction can be given to the field (Zheng and Zhang, 2017). Several algorithms are already tested in the field to apply the association rule-based data mining. But the communication overhead is required to be compared for these algorithms. This would help in reaching the best possible answer to the question of how to reduce the overhead while using the association rule.

3.5 Data collection methods

An indication of the key data collection method to be used in the study is somewhat clear in the previous section. The previous studies in the area are required to be explored in the study. Now, there are two types of data collection methods in a broad sense. These two are the primary method of data collection and the secondary method of data collection. The first one requires the researcher to collect the data based on some direct field works. The data would be directly collected by the researcher through methods like surveys, cross-sectional studies, or interviews. Then these data are to be analyzed for meeting the research hypotheses. In the case of a secondary method, the source of data is some indirect process and not direct field work. These sources can be various previous surveys, organizational data, or some previous research works. Now, as it is mentioned the current study is aiming towards critically analyzing previous studies. Thus, it means the source of the data being collected would be a secondary one (Sohrabi, 2018). In the current study, the communication overhead faced by various association rule-based algorithms is the key focus of the study. Also, the study is aiming towards finding out the best way to minimize or optimize this without hampering efficiency. The synchronization among various computational tasks and components becomes a dire need (Telikani et al., 2020). Otherwise, both the energy and cost spent in the process would enhance drastically with the increase in complexity of tasks. This research will enable the company to implement the advanced algorithm in their business module. With the rise in opportunities, there are various challenges as well that have the potential to hinder the progress of scientific discovery and technological advancement. Due to these potential challenges, it is an essential step to find out the root cause of those challenges in the market and improve the existing methodology which will enable the company to further improve its consumer experience. The key sources of data that are chosen here are the peer-reviewed journals and articles from the reliable database of Google Scholar. Apart from these, several organizational web pages and articles are also used as the source of data to be collected here.

3.6 Data analysis methods

The data analysis method to be used in the research is the ultimate core of all the methodology layers. Now, based on the nature of the data, the analysis can be of two broad categories. These two are quantitative data analysis and qualitative data analysis. In the case of quantitative analysis, the data being collected is either of a numerical value or are to be converted into numeric scales (like the Likert scale in research). Thus, this data can be analyzed through a statistical or a numeric approach. But the collected data can also be of the conceptual type where a broader conceptual analysis is more apt (Wang *et al.,* 2020). This type of conceptual analysis is usually known to be qualitative data analysis. Now, sometimes there is a need for both types to be used in a synchronous way to complement each other. Such a method is known to be the mixed method of analysis.

Now, in the current research, it is to be kept in mind that this research signifies the constant improvement to increase the efficiency in the network and enable the users to have a more effective data collecting system which further enables them to get accurate meaningful insight of the data. This research will enable the company to become more effective in data mining methodology and in addition, this research result will enable them to increase complexity in collecting data which will enable more transparent and simple techniques and enable the company to increase their productivity. This research will also enable companies to find out the best alternative solution to minimize the communication overhead in their existing network and enable them to extract the meaningful information more effectively which will enable the company to improve its consumer experience, thus generating more revenue in the market. So, this means that going for a mere quantitative analysis would result in a lesser dive in the key issue to be targeted. It would not be possible to reach a meaningful choice through analyzing just quantitative data. So, the research is going for a rather qualitative approach. Several data collected from the articles, journals, webpages are critically reviewed here to reach the best possible answers to the research questions.

3.7 Research ethics

Research ethics is another key concern to be addressed for any research and researcher. It helps the research to establish the reliability of the study. Ethical considerations put the limit to how much data can be manipulated without crossing the moral boundary (Agapito et al., 2018). It is true that for any successful research the data analysis must manipulate the data in the best possible way to reach the research objectives. But there are chances when such manipulation may deviate the moral alignment causing loss the reliability. Now, an optimistic aspect in the case of the current research is that the research does not require any human participant to take participate directly. Hence unethical possibilities like unalignment with the requirement of confidentiality or unbiased approach are not much of a concern. But yet some ethical considerations are must here be addressed. The first one is the requirement of giving true credit to the owner of the data. As the data is collected from secondary resources it is a must to cite and refer to the author or source links of the data being addressed (Wu and Zhang, 2020). Another key aspect here is to avoid any kind of over-manipulation of the data. This refers to the consideration that when data is being cited it must not be deviated by any means from the source data before any analysis. Thus, would change the input data itself and will cause an unnecessary change in the results. So, this is also needed to be avoided.

3.8 Conclusion

So, layer by layer the methodology is explored in the current chapter. It was critically important to be done as the research analysis part is otherwise not possible to be proceeded with. All the layers of the research onion have been explored in the chapter to reach the final core of it. The research is following a critical review of the data collected from secondary sources with a qualitative approach being followed. Now, the thesis can proceed with the next chapter i.e. the finding and analysis.

CHAPTER 4: FINDINGS AND ANALYSIS

4.1 Introduction

This chapter will discuss the findings research and along with that and an analysis of the findings will be discussed to provide the relevance of this research and enable the researchers to understand the various meaningful insight from the research and along with this chapter will enable the raiders to form a critical understanding of the topic which will allow the readers to become more aware of the subject (Telikani et al., 2020). This research is based on secondary data. Based on the secondary data this research has been able to gather crucial insight from the previous researcher's work which will enable the readers and the key stakeholders to get meaningful insight from research which will further help the company to implement the potential solution for minimizing the communication delay using the association rule mining algorithm (Agapito et al., 2018). The empirical study has shown that there are various methods to implement the "*association rule mining algorithm to reduce the communication overhead"*. This chapter will discuss the result of the implementation of the various approaches in minimizing the "*communication overhead*" from the system's network. All these results have been achieved by combining and implementing advanced analytics algorithms in a system's network. The following section will discuss the results that have been found from this research.

4.2 Result Findings.

This section will discuss the result that is found from the research, and enable the researchers to establish the connection between the empirical study and the result. This section will be segmented into multiple parts which will allow the readers and researchers to understand the efficiency of the result in each part and based on the efficiency, this section will enable to understand whether the research has been successful in minimizing the communication delay or not and further enable the readers and researchers to understand the various elements that influencing the particular problem in the distributed computing system, in addition, this section will discuss a suitable algorithm to mitigate the existing problem and enable the company to become more effective in preventing the communication delay from the network (Gutierrez-Rojas *et al.*, 2020). Before discussing the result of this research paper, the following segment will discuss the minimum hardware requirement to conduct the study.

As this cutting-edge technology is highly sophisticated, due to this reason, to implement this type of technology requires high computational power (Leung et al., 2017). There are two specific ways to obtain high computational power. The first methodology is where the users and the researchers can use cloud technology to implement this study. In cloud technology, the users or the researchers will have the readymade infrastructure and they can use the existing infrastructure to implement the machine learning, deep learning algorithm to extract meaningful information. One initiative is "Google COLAB" where the users can use this cloud infrastructure to implement this algorithm and run those sophisticated machine learning models. This approach allows the users and researchers to leverage the already built-in setup from the third-party companies for the analysis. It requires only a high-speed internet connection. However, there is one problem. the control of the data can be accessible to the third party which will be a matter of concern for the researchers and the various hackers might exploit those datasets to fulfill their objectives (Kuriakose and Nedunchezhian, 2017). Another approach to conducting the entire research is in the custom-built set. In a custom-built setup. It requires a minimum of 8 GB, DDR4 Ram, to handle the data and along with that it minimum i5 processor or even the Ryzen 5 processor that will work as well. The processor must have a quad-core to efficiently function and enable the researchers to conduct the research smoothly. Researchers can choose "Windows 10 or Linux" based on their preference. The result of this research paper is discussed in the following segments.

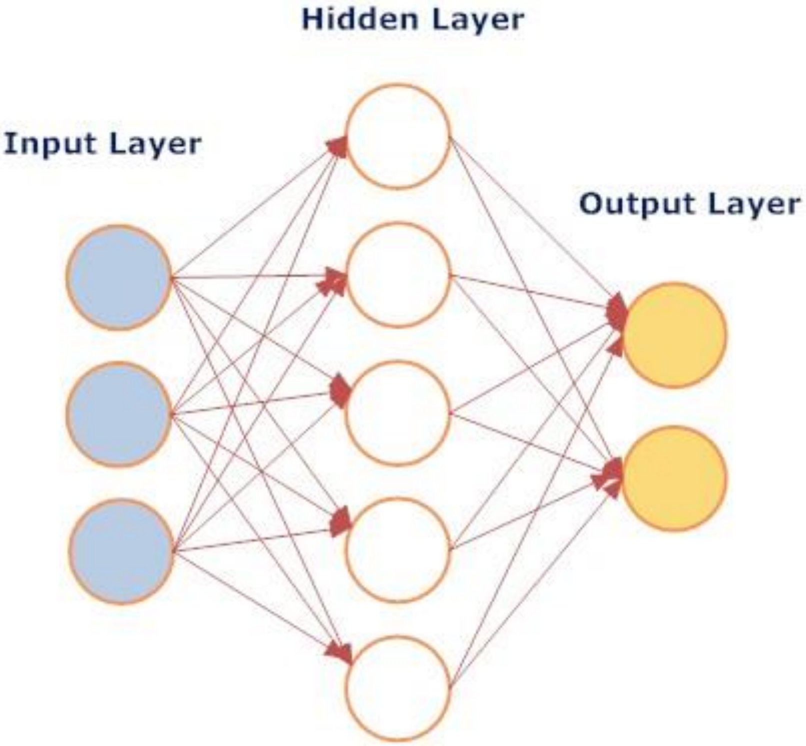

Figure 6: Neural Network analysis

(Source: Aditya *et al.,* 2017)

The above figure demonstrated the framework of the "artificial neural network or ANN". The neural framework has shown there are three different layers in the entire framework. The results show that in the input their different zone data have been added (Leung et al., 2017). These data types are extracted from the raw dataset, in this stage, the map reduction algorithm has been implemented to find out the more relevant data from the entire dataset. This is the first phase of the MapReduce algorithm which uses the mapper function to determine the most relevant data and based on the relevancy and this key-value pair has been assigned to the segmented data. The segmented data then transmitted to the hidden layer of the neural network or the computational layer is responsible for implementing all the sophisticated algorithms which will allow the researchers to perform the computation on the dataset. This step will perform the interest measures (Aditya *et al.,* 2017). This is a special part of the deep learning module which will allow the researchers to become more effective and enable them to

get a wide insight of the research which will help them to get a proper understanding of the certain behavior among the nodes and this enables researchers to manipulate the algorithm to become more efficient. The neural network module will examine the three parameters (Agapito et al., 2018). These three parameters will enable the network to further segment the segregated dataset. The three segmentations have been done on the percentage of the interest measures of the dataset. The first segment is lower than 30%. The second segment will be based on the 30% to the 60% accuracy result and the last one will be based on the result of more than 60%. This stage will also adapt the reducer stage of the MapReduce algorithm, the reducer stage has reduced all the duplicate values and enabled the researchers to get a more meaningful DB which will allow the researchers to conduct the research more effectively and efficiently. The conversion of the DB occurs during this stage. The raw DB will be converted into the transactional DB which will be appropriate for further analysis (Wainakh *et al.,* 2019). The conversion will occur when the key-value pair from the initial stage of the mapper function has been faded into the reducer function. The key-value pair will transform the entire dataset into a transaction dataset. The next step of the network is to feed the network layer which is known as the output layer, in this layer, there are a total of three blocks identified. One block is the major part, two blocks are part of this one major block. The major block is the rule comparator. In this stage the mapper stage part 2 algorithm has been applied to determine the two minor blocks. These two minor blocks are “inconsistent rule” and “consistent rule”.

The above figure demonstrated the comparison among the three algorithms. Based on the execution time the evaluation of the three algorithms has been done. These three algorithms are the “FDM, CDA algorithm, and DFPM algorithm”. The above graph demonstrates that the CDA algorithm has a high execution time. The CDA algorithm takes a long time to execute and start its function in the next step is the FDM algorithm. The FDM algorithm takes moderate execution. One of the reasons for taking such a long time for execution is that these two algorithms generate a massive dataset, and the nodes within the dataset send continuous messages to communicate with others. Due to this reason, the execution time increases for the algorithm and it takes lots of time to implement and it fails to reduce the communication time (Braun et al., 2018). Hence it proves to be an inefficient choice of the algorithm. However, it has been observed for the simple implementation of this algorithm, this algorithm can perform the initial analysis. Nevertheless, to get the optimum result and reduce the execution time DFPM algorithm is the suitable choice. The algorithm execution time is less than the other two algorithms. A comparison of the six different iterations has been observed, from the observation it has been identified the algorithm consistently has less execution time. Hence the algorithm is reliable and can be an effective choice for further analysis. Along with the execution time, this algorithm has another advantage that makes it a suitable choice for the analysis (Agapito et al., 2018). Nodes within the algorithm are connected, and all the nodes are responsible for counting and not sending information to each node. Due to this reason, it drastically reduces communication time and becomes an effective choice for the large dataset.

The figure is based on the six iterations and based on the six iterations it demonstrated the consistent rules outperformed the inconsistent rules (Telikani et al., 2020). The consistent rules refer to the combination of association rules which contains an itemset which that can be local and global frequent in the massive dataset. If the large dataset contains local and global items set in frequent manners, that means the dataset adheres to the consistent rules. Another rule is the inconsistent rules. This rule refers to the condition where the large dataset does not continually have local and global itemset, due to this inconsistency in the large dataset. It refers to an inconsistent rule. Raw dataset mostly follows the inconsistent rule as the data is often incomplete in real-life scenarios (Narayana and Vasumathi, 2018). Due to this reason, it is hard to further apply any computational logic in the dataset. Hence to obtain meaningful insight the dataset must be converted in such a way that it follows the consistent rule and enables the researchers to get the meaningful information. To get meaningful information researchers have used the MapReduce algorithm to convert the raw data set into more suitable databases which are responsive in providing effective results.

CHAPTER 5: DISCUSSION

5.1 Discussion of the findings

This section will provide an important insight of findings which enable the readers to understand the relevance of the result in the context of the research paper and in addition, this segment will enable the other companies to look for an alternative way to conduct the research and find the potential solution to mitigate the existing problem which will enable the company to improve their existing operational framework and enable the companies to become a more effective player in the market (Agapito et al., 2018). From this research various results have been observed, this observation from the research will be discussed in this segment which will help the readers and researchers to understand whether this research is successful or not.

The observations have demonstrated that the first algorithm which was proposed was an apriori algorithm in the data mining section to improve communication. This algorithm is one of the first algorithms that was proposed by researchers in 1994. However, this algorithm has a few drawbacks due to this reason this algorithm was not reliable and this algorithm is not an effective and efficient choice for reducing the communication time for the distributed computational infrastructure. Nevertheless, this algorithm was proved to be a useful choice for a single computational device. But due to the rise in data consumption and the rise of distributed devices for computational work. This algorithm is no longer relevant in the industry.

The researchers have used various algorithms such as CD, FDM, and DPFM, all these algorithms have quite useful functionality to improve the communication time. However, all the algorithms have failed to provide a satisfactory result. For example, the CD algorithm has useful functionally in counting the candidate dataset, however, this algorithm is not effective while working on the large dataset and as the large dataset does not follow the consistent rule (Telikani et al., 2020). Hence it becomes much more complicated to compute the entire dataset and it takes a longer time than usual and often the results are not as reliable as is expected. Along with that, this algorithm has a high execution time. Hence this approach is suitable for reducing communication in a large dataset. Another instance using the FDM algorithm, the FDM algorithm is efficient in handling the large dataset and this algorithm has improved execution time than the CD algorithm which makes it a suitable choice to replace the CD algorithm, However, this algorithm has interconnected nodes within the resultant dataset. All these interconnected nodes communicate with each other, all these nodes send information across the network. The constant communication over the network always causes massive traffic in the network. Due to this massive traffic, it takes new messages to travel to one node another it takes massive time and it increases the communication overhead in the distributed network. Hence this algorithm has failed to reduce the communication problem in the network. FDM algorithm is another choice to implement a machine learning model and extract meaningful information from the dataset. This algorithm is an effective choice for handling the large dataset and along with that nodes within the algorithm do not communicate with each other due to this reason the traffic is less in the network and the data from the distributed network can travel to the server in a very fast manner.

5.1.1 Contradiction among CD and FDM

Following are the contradictions among the CD and FDM-

- FDM is the easy and simple demodulation whereas the CD has complex demodulation.
- FDM gives more latency than the CD.

- FDM is more reliable than the CD.
- CD is less expensive than the FDM.
- CD has dynamic coordination.

The nodes within the FDM algorithm only exist to count global and local itemsin the dataset. Based on counting the global and local items in the data set it can enable the researchers to understand whether the dataset is following the inconsistent rule or the dataset is following the consistent rules. Based on the observation this will enable the researchers to convert the dataset and enable the dataset for further analysis to test the potential solution for the problem (Agapito et al., 2018). Along with the FDM algorithm and the MapReduce algorithm. Implementation of these two algorithms has enabled the researchers to identify the candidate dataset which is following "consistent rule and inconsistent rule". Based on further analysis it has been observed using this detection principle has allowed the researchers to get valuable insight which is crucial for reducing the communication overhead from the distributed computational network (Leung et al., 2017). Due to the multiple phases)of the MapReduce algorithm it will allow the researchers to get the authentic data to work which enables the researchers to become more effective in their analysis and enable the research to be more impactful and enable this algorithm to be more efficient in the various industries. Especially, this algorithm is used in the retail industry, hence optimizing this algorithm will enable the retail industry to be more effective at forecasting the demand and enable the industry to improve their customer experience.

5.2Conclusion

To conclude the entire chapter, this chapter will play an instrumental role for the readers to form a critical understanding and enable them to understand the relevance of the research and the relevance of the findings. This chapter will help the readers to understand the various aspects of the findings and enable the readers to find out why this research is still relevant and enable the readers to understand the effective ways the businesses and the other researchers can benefit from the research (Leung et al., 2017). This research will act as a filter which has been based on critically analyzing all the aspects of the existing algorithm in the market to reduce the communication problem in the distributed system and enable researchers to understand various aspects of the algorithm and enable them to look for the performance of the algorithm which will enable the researchers and the key stakeholders to choose from the best possible algorithm to implement and eliminate the communication problem in the distributed network system (Agapito et al., 2018). Furthermore, the use of MapReduce algorithm to detect and identify the database will enable the researchers to focus on the data preprocessing for getting effective results from the analysis. Using this innovative approach to solve the problem will enable other researchers to look for more optimized ways to conduct the study and more effective results which will help the industry to grow and advance the current practices in the industry.

CHAPTER 6: CONCLUSION

6.1 Introduction

This is the final chapter of this dissertation. This chapter will enable the readers to understand whether the entire research has been able to fulfill its objective and enable the readers to understand whether implementation of the result is viable or not. This chapter consists of several subchapters which will enable the readers to have the various aspects of the research, this chapter will discuss the three main subsections, the first subsection is linking with objectives, in this subsection, all the steps that are taken in this research will be discussed and along with that this will discuss the effective ways that are relevant in the context of the research objective. The next subsection is recommendations, in this subsection will provide better and effective ways to conduct the study in the future time and enable the readers to understand the scope of the improvement in the research framework, The subsequent section will discuss the limitation of the research and future scope of the research which enables the readers to get the general overview of this research.

6.2 Linking with the objectives

This section will enable the readers to understand the way this research paper meets the objective of the research and enable readers to understand the relevancy of the research paper and the objective of this research. The following section will discuss the points

- **To eliminate duplicate computational data from the database.**

Duplicate data is very common in databases, the databases are often filled with duplicate data. The duplicate data is often a matter of concern for the companies and the researcher. The duplicate data enable the researchers to get ineffective result. Hence the data needs to be cleaned by removing the duplicate data. Along with the duplicate data the database consists of various incomplete data. Due to this reason, it is difficult for the researchers to conduct any analysis of the raw data (Telikani et al., 2020). Hence it is very important to eliminate the duplicate data from the entire database and along with that, all the incomplete data must be eliminated from the entire database to get effective results from the analysis. Hence, to obtain accurate data from the analysis, first, the data preprocessing is implemented, a various sorting algorithm is implemented to obtain the data and further, the algorithm will be applied to identify the nature of the database, this approach will allow the researchers to obtain the understanding of whether the database is following the consistent rule or the database is following the inconsistent rule and based on the identification the dataset. It will enable researchers to establish further analysis and help them to understand whether the analysis is accurate or not.

- **To minimize the "communication overhead" in distributed computation infrastructure**

"Communication overhead" is one of the biggest problems in the "distributed computational infrastructure". The data revolution and the evolution of digital technology have increased data consumption. Due to this reason, the demand for distributed devices has increased. One of the main challenges of the distributed computing device as the devices are interconnected and the data is coming from all the various sources creates massive traffic on the network. Due to this reason, it is essential to reduce the communication overhead to minimize traffic in the network to ensure smooth communication and enable the data to be sent to the server without any communication delay. To minimize this problem of communication in the network there several algorithms have been applied to this research paper, once the algorithm is applied then all the algorithms had been compared and evaluated to understand the efficiency in the several aspects of the data which will enable the researchers to get the meaningful insight of element which are causing the problem and along with that the researchers will get the valuable information from the evaluation and comparison of the various algorithm which will allow the researchers to get a most suitable algorithm that has the effective result. This research paper has evaluated three algorithms, which are “FDM, CD, and DPFM”. All these algorithms have been examined to evaluate their effectiveness in providing favorable results. The DPFM has been considered as one of the most effective than the other two algorithms and along with using the MapReduce algorithm to perform the analysis will help the researchers to get the most optimum result and enable the users to get the most effective and optimized result which will meet the objective of this research paper. Using MapReduce to identify the nature of the dataset will further enable the researchers to get more insightful information about the large dataset and further enable the researchers to get a more impactful algorithm combination to mitigate the existing problem.

- **To reduce the bandwidth usage in the distributed computational framework of the company**

As distributed computing is on the rise that means the multiple devices will be connected in the network or they can be connected in different networks. As the number of devices is increasing it is also increasing the bandwidth usage in the network. the bandwidth usage comes at a cost. It increases the entire operational cost and along with that it also increases the latency in the network. Which is responsible for creating massive traffic in the network and creating inefficiency results which increase the cost and the company often has to face the loss. This type of algorithm has a wide range of applications in the retail industry and this type of algorithm is based on conditional probability. Hence if the probable statement is not clear then it will predict the wrong result and enable the company to have a wrong forecast which eventually leads to catastrophic results in the market. To minimize bandwidth usage from the distributed setup. The algorithm calculation will find out the duplicate setup which will enable the researchers to get insight into the duplicate infrastructure and based on the analysis the researchers can mitigate those structures which will enable them to get more optimized results and clearer insight into the effective bandwidth usage from the network.

- **To generate the most effective “association rule mining algorithm” to improve the efficiency in the data collection**

The data collection technique is one of the most important techniques which enable the researchers to conduct the analysis in a better way and enable the researchers to get meaningful insight from the dataset. There are various algorithms currently and the researchers are conducting various algorithms to get meaningful information from the dataset. The first algorithm proposed for the association rule mining was an apriori algorithm since then this field has come a long way based on conditional probability. This will enable the researchers to conduct more robust and resilient algorithms which can provide more effective and efficient algorithms and enable the researchers to achieve more effective results. These results are essential in forecasting and predicting the future and enabling the network to have a smooth and efficient implementation algorithm that reduces the communication traffic from the network. To solve this problem this paper has used a hybrid approach as most of the algorithm has their advantages and disadvantages which will create a problem for the researchers to get the optimum result. Hence, to optimum result, the hybrid approach or combination of two algorithms has been implemented in the network. The combination of both algorithms will leverage the benefit of these two algorithms and minimize the drawback of the algorithm and enable the researchers to get a more optimum result.

- **To democratize effective computation of the various frequencies**

The increasing demand for computing devices has led to massive data consumption and enables users to spend massive time on their devices. Hence it is an important step to democratize the computational infrastructure using an effective algorithm that will improve the user's experience and enable them to have more interest in using the device and as the most distributed and wearable devices are invented it is creating more data. Hence a necessary adequate framework is needed to implement data analysis. Along with that algorithm needs to be democratized to understand and evaluate. the impact on the several computational scenarios will allow the researchers to get the exact meaningful information from the dataset and further the researchers are open-minded while using the various approaches in the algorithm is will allow the researcher to get the new and innovative solution to solve the problem.

6.3 Recommendation

This research is based on secondary data. That means the data is collected from various other sources and the research work is based on the work of the previous researchers. Hence it is hard to find the relevant data which satisfied the objective of this paper. Hence the primary analysis can be done to improve the result of the research. The primary result can be expensive but it will form a critical insight that allows the researchers and readers to extract the meaningful information from the dataset and along various other algorithms can be used to the more robust and resilient dataset and along with this resultant algorithm from the research has no application “relational database and non-relational database”. The impact of the algorithm has not been assessed in those types of datasets due to the full impact of the resultant algorithm yet to be measured.

6.4 Research Limitation and Future scope

The secondary research has several limitations, those limitations are going to be discussed in this segment. The data collected is based on the previous researchers' work due to this reason. Oftentimes data is vaguely old and not relevant to current research needs due to this reasoning. Oftentimes it is impossible to extract meaningful information from this type of research. However, this type of research is useful in a few scenarios. First, this type of research is less expensive and the data is already available in the form of previous research and along with that this type of research does not require massive time. This type of research can be done in a short period. However, there are few ways which can be adopted to implement in the research which will minimize its limitation and enable the researchers to get the optimum result that they are looking for, the recent study will be collected to form the initial understanding which will create the base foundation of the research subject and along with that, all relevant and related studies will be selected to conduct the study.

The future of this research looks promising. As computing is on the way to increase and it is more likely to increase further. Hence the industry will require more robust and resilient algorithms which can provide effective and efficient results and enable the industry to have more sustainable growth. With that being said, implementation of the cloud technology looks like a very priming choice for further implementation of the study and it has some potential to be a great addition for researchers to obtain more effective results from the analysis.

Reference list

Aditya, S.P., Hemanth, M., Lakshmikanth, C.K. and Suneetha, K.R., 2017, August. Effective algorithm for frequent pattern mining. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 704-708). IEEE.

Agapito, G., Guzzi, P.H. and Cannataro, M., 2018. Parallel and distributed association rule mining in life science: A novel parallel algorithm to mine genomics data. *Information Sciences*.

Ahmed, S.A. and Nath, B., 2021. Identification of adverse disease agents and risk analysis using frequent pattern mining. *Information Sciences*, *576*, pp.609-641.

Bhukya, R. and Gyani, J., 2020. Survey on Fuzzy Associative Classifications Techniques and Their Performance Evaluation with Different Fuzzy Clustering Techniques Over Big Data. In *ICDSMLA 2019* (pp. 420-431). Springer, Singapore.

Biswas, S., Biswas, N. and Mondal, K.C., 2018, November. Parallel Apriori based distributed association rule mining: A comprehensive survey. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* (pp. 202-207). IEEE.

Bouraoui, M., Bouzouita, I. and Touzi, A.G., 2017, December. Hadoop based mining of distributed association rules from big data. In *2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)* (pp. 185-190). IEEE.

Braun, P., Cuzzocrea, A., Leung, C.K., Pazdor, A.G., Tanbeer, S.K. and Grasso, G.M., 2018, May. An innovative framework for supporting frequent pattern mining problems in IoT environments. In *International Conference on Computational Science and Its Applications* (pp. 642-657). Springer, Cham.

Chahar, H., Keshavamurthy, B.N. and Modi, C., 2017. Privacy-preserving distributed mining of association rules using Elliptic-curve cryptosystem and Shamir's secret sharing scheme. *Sādhanā*, *42*(12), pp.1997-2007.

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

Chon, K.W. and Kim, M.S., 2018. BIGMiner: a fast and scalable distributed frequent pattern miner for big data. *Cluster Computing*, *21*(3), pp.1507-1520.

Dasgupta, S. and Saha, B., 2020. Study of various parallel implementations of association rule mining algorithm. *American Journal Of Advanced Computing*, *1*(3), pp.1-7.

Dhanalakshmi, K.S. and Kannapiran, B., 2017. Analysis of KDD CUP dataset using multi-agent methodology with effective fuzzy based intrusion detection system. *Journal of Applied Security Research*, *12*(3), pp.424-439.

Dhanalakshmi, K.S. and Kannapiran, B., 2017. Analysis of KDD CUP dataset using multi-agent methodology with effective fuzzy based intrusion detection system. *Journal of Applied Security Research*, *12*(3), pp.424-439.

Djenouri, Y., Belhadi, A., Fournier-Viger, P. and Fujita, H., 2018. Mining diversified association rules in big datasets: A cluster/GPU/genetic approach. *Information Sciences*, *459*, pp.117-134.

Dong, Z., 2020, February. Big Data Oriented Mining and Implementation Analysis for Online Education Information. In *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)* (pp. 856-860). IEEE.

El‐Hasnony, I.M., Mostafa, R.R., Elhoseny, M. and Barakat, S.I., 2021. Leveraging mist and fog for big data analytics in IoT environment. *Transactions on Emerging Telecommunications Technologies*, *32*(7), p.e4057.

El-Hasnony, I.M., Mostafa, R.R., Elhoseny, M. and Barakat, S.I., 2021. Leveraging mist and fog for big data analytics in IoT environment. *Transactions on Emerging Telecommunications Technologies*, *32*(7), p.e4057.

Fu, C., Wang, X., Zhang, L. and Qiao, L., 2018, April. Mining algorithm for association rules in big data based on Hadoop. In *AIP Conference Proceedings* (Vol. 1955, No. 1, p. 040035). AIP Publishing LLC.

Gan, W., Lin, J.C.W., Chao, H.C. and Zhan, J., 2017. Data mining in distributed environment: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), p.e1216.

Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C. and Yu, P.S., 2019. A survey of parallel sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(3), pp.1-34.

Garach, P. and Patel, D., 2019, March. Privacy Protection of Class Association Rules produced by medical datasets. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* (pp. 1-4). IEEE.

Gutierrez-Rojas, D., Ullah, M., Christou, I.T., Almeida, G., Nardelli, P., Carrillo, D., Sant'Ana, J.M., Alves, H., Dzaferagic, M., Chiumento, A. and Kalalas, C., 2020, June. Three-layer approach to detect anomalies in industrial environments based on machine learning. In *2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS)* (Vol. 1, pp. 250-256). IEEE. Kuriakose, S. and Nedunchezhian, R., 2017. Efficient adaptive frequent pattern mining techniques for market analysis in sequential and parallel systems. *Int. Arab J. Inf. Technol.*, *14*(2), pp.175-185.

Hammami, H., Brahmi, H., Brahmi, I. and Yahia, S.B., 2017, September. Using homomorphic encryption to compute privacy preserving data mining in a cloud computing environment. In *European, Mediterranean, and Middle Eastern Conference on Information Systems* (pp. 397-413). Springer, Cham.

Han, Q., Lu, D., Zhang, K., Song, H., and Zhang, H., 2019. Secure Mining of Association Rules in Distributed Datasets. *IEEE Access*, *7*, pp.155325-155334.

He, Q., Zhou, W., Xu, H., Cui, L., Li, X. and Liu, J., 2018. A distributed network alarm correlation analysis mechanism for heterogeneous networks. *Journal of Circuits, Systems and Computers*, *27*(01), p.1850012.

Hu, H. and Chen, Y., 2018, May. Research on the Factors of Frequent Itemset Mining Based on Dynamic Hashing. In *2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)* (pp. 215-220). Atlantis Press.

Islabudeen, M. and Devi, M.K., 2020. A smart approach for intrusion detection and prevention system in mobile ad hoc networks against security attacks. *Wireless Personal Communications*, *112*(1), pp.193-224.

Jia, K. and Liu, H., 2017, September. An improved FP-growth algorithm based on SOM partition. In *International Conference of Pioneering Computer Scientists, Engineers and Educators* (pp. 166-178). Springer, Singapore.

Jiang, Y., 2020, June. Network big data mining algorithm based on association rules of computer technology. In *Journal of Physics: Conference Series* (Vol. 1574, No. 1, p. 012084). IOP Publishing.

Jiang, Y., Zhao, M., Hu, C., He, L., Bai, H. and Wang, J., 2019. A parallel FP-growth algorithm on World Ocean Atlas data with multi-core CPU. *The journal of Supercomputing*, *75*(2), pp.732-745.

Keerthi, K. and Saritha, S.J., 2017, August. ECLAT: Frequent itemset using MapReduce. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 3744-3748). IEEE.

Khedr, A.M., Osamy, W., Salim, A. and Abbas, S., 2020. A Novel Association Rule-Based Data Mining Approach for Internet of Things Based Wireless Sensor Networks. *IEEE Access*, *8*, pp.151574-151588.

Khedr, A.M., Osamy, W., Salim, A. and Salem, A.A., 2019. Privacy preserving data mining approach for IoT based WSN in smart city. *International Journal of Advanced Computer Science and Applications*, *10*(8), pp.555-563.

Kim, H.J., Shin, J.H., Song, Y.H. and Chang, J.W., 2019, July. Privacy-preserving association rule mining algorithm for encrypted data in cloud computing. In *2019 IEEE 12th International Conference on Cloud Computing*

(CLOUD) (pp. 487-489). IEEE.

Leung, C.K., Jiang, F. and Pazdor, A.G., 2017, August. Bitwise parallel association rule mining for web page recommendation. In *Proceedings of the International Conference on Web Intelligence* (pp. 662-669).

Li, L., 2020. Real time auxiliary data mining method for wireless communication mechanism optimization based on Internet of things system. *Computer Communications*, *160*, pp.333-341.

Liu, L., Su, J., Chen, R., Liu, X., Wang, X., Chen, S. and Leung, H., 2018, July. Privacy-preserving mining of association rule on outsourced cloud data from multiple parties. In *Australasian Conference on Information Security and Privacy* (pp. 431-451). Springer, Cham.

Luna, J.M., Fournier‐Viger, P. and Ventura, S., 2019. Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(6), p.e1329.

Miholca, D., 2018. An adaptive gradual relational association rules mining approach. *Studia Universitatis Babe-Bolyai Informatica*, *63*(1), pp.94-110.

Mlambo, M.N., Gasela, N. and Esiefarienrhe, M.B., 2018, August. Implementation and Analysis of Enhanced Apriori Using MapReduce. In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)* (pp. 1-6). IEEE.

Nadimi-Shahraki, M.H. and Mansouri, M., 2017, March. Hp-Apriori: Horizontal parallel-apriori algorithm for frequent itemset mining from big data. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (pp. 286-290). IEEE.

Narayana, G.S. and Vasumathi, D., 2018. An attributes similarity-based K-medoids clustering technique in data mining. *Arabian Journal for Science and Engineering*, *43*(8), pp.3979-3992.

Nikam, P.V. and Deshpande, D.S., 2018, April. New approach in Big Data Mining for frequent itemset using mapreduce in HDFS. In *2018 3rd International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE.

Pang, H. and Wang, B., 2020. Privacy-preserving association rule mining using homomorphic encryption in a multikey environment. *IEEE Systems Journal*, *15*(2), pp.3131-3141.

Patel, B.M., Bhemwala, V.H. and Patel, A.R., 2018. Analytical Study of Association Rule Mining Methods in Data Mining.

Prajapati, D.J., Garg, S. and Chauhan, N.C., 2017. Map reduce based multilevel consistent and inconsistent association rule detection from big data using interestingness measures. *Big data research*, *9,* pp.18-27.

Pugh, S., Binkley, D. and Moonen, L., 2018, September. The case for adaptive change recommendation. In *2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation (SCAM)* (pp. 129-138). IEEE.

Puścian, M., 2019, November. Reliable frequent itemsets mining with actor-based Apriori algorithm. In *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019* (Vol. 11176, p. 1117621). International Society for Optics and Photonics.

Qian, K., Gao, S. and Yu, L., 2021. Marginal frequent itemset mining for fault prevention of railway overhead contact system. *ISA transactions*.

Qian, K., Yu, L. and Gao, S., 2021. Fault Tree Construction Model Based on Association Analysis for Railway Overhead Contact System. *Int. J. Comput. Intell. Syst.*, *14*(1), pp.96-105.

Qiu, S., Wang, B., Li, M., Liu, J. and Shi, Y., 2017. Toward practical privacy-preserving frequent itemset mining on encrypted cloud data. *IEEE Transactions on Cloud Computing*, *8*(1), pp.312-323.

Raj, S., Ramesh, D. and Sethi, K.K., 2021. A Spark-based Apriori algorithm with reduced shuffle overhead. *The Journal of Supercomputing*, *77*(1), pp.133-151.

Raj, S., Ramesh, D., Sreenu, M. and Sethi, K.K., 2020. EAFIM: efficient apriori-based frequent itemset mining algorithm on Spark for big transactional data. *Knowledge and Information Systems*, *62*(9), pp.3565-3583.

Rani, R.M. and Pushpalatha, D.M., 2018. Discovery of Knowledge Using Association Rules in Wireless Sensor Epocs-a Survey. *International Journal of Engineering & Technology, 7*(4.10), pp.436-439.

Rani, R.M. and Pushpalatha, M., 2019. Generation of Frequent sensor epochs using efficient Parallel Distributed mining algorithm in large IOT. *Computer Communications*, *148*, pp.107-114.

Ranjith, K.S., Zhenning, Y., Caytiles, R.D. and Iyengar, N.C.S., 2017. Comparative Analysis of Association Rule Mining Algorithms for the Distributed Data. *International Journal of Advanced Science and Technology*, *102*, pp.49-60.

Rochd, Y. and Hafidi, I., 2018. Performance Improvement of PrePost Algorithm Based on Hadoop for Big Data. *International Journal of Intelligent Engineering and Systems*, *11*(5), pp.226-235.

Rong, H., Wang, H., Liu, J., Tang, F. and Xian, M., 2018, May. Verifiable and Privacy-Preserving Association Rule Mining in Hybrid Cloud Environment. In *International Conference on Green, Pervasive, and Cloud Computing* (pp. 33-48). Springer, Cham.

Sawant, V. and Shah, K., 2018, April. A System that Performs Data Distribution and Manages Frequent Itemsets Generation of Incremental Data in a Distributed Environment. In *International Conference on Advances in Computing and Data Sciences* (pp. 104-113). Springer, Singapore.

Shunmugam, S., Selvakumar, D.R. and Kavitha, P., 2018. A Virtual Coordinator based Privacy-Preserved Distributed Data mining Using Association Rule. *International Journal of Pure and Applied Mathematics*, *119*(16), pp.1535-1540.

Sinaei, S. and Fatemi, O., 2018. Run-time mapping algorithm for dynamic workloads using association rule mining. *Journal of Systems Architecture*, *91*, pp.1-10.

Sohrabi, M.K. and Roshani, R., 2017. Frequent itemset mining using cellular learning automata. *Computers in human behavior*, *68*, pp.244-253.

Sohrabi, M.K., 2018. A gossip based information fusion protocol for distributed frequent itemset mining. *Enterprise Information Systems*, *12*(6), pp.674-694.

Song, J., Xie, H. and Feng, Y., 2017. Fast association rule mining algorithm for network attack data. *Journal of Discrete Mathematical Sciences and Cryptography*, *20*(6-7), pp.1465-1469.

Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Krishnan, M.N., Ramasamy, L.K., Kadry, S. and Lim, S., 2021. An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data. *Bulletin of Electrical Engineering and Informatics*, *10*(1), pp.390-403.

Suthar, S.R., Dabhi, V.K. and Prajapati, H.B., 2017, April. Machine learning techniques in Hadoop environment: A survey. In *2017 Innovations in Power and Advanced Computing Technologies (i- PACT)* (pp. 1-8). IEEE.

Tehreem, A., Khawaja, S.G., Akram, M.U., Khan, S.A. and Ali, M., 2017, May. Parallel architecture for implementation of frequent itemset mining using FP-growth. In *2017 International Conference on Signals and Systems (ICSigSys)* (pp. 92-98). IEEE.

Telikani, A., Gandomi, A.H. and Shahbahrami, A., 2020. A survey of evolutionary computation for association rule mining. *Information Sciences*, *524*, pp.318-352.

Tribhuvan, S.A., Gavai, N.R. and Vasgi, B.P., 2017, August. Frequent Itemset Mining Using Improved Apriori Algorithm with MapReduce. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

Urmela, S. and Nandhini, M., 2017. Approaches and techniques of distributed data mining: A comprehensive study. *International Journal of Engineering and Technology (IJET)*, *9*(1), p.69.

Varma, S. and LijiP, I., 2017. Secure Outsourced Association Rule Mining using Homomorphic Encryption. *International Journal of Engineering Research & Science (IJOER)*.

Varshney, P., 2017. *Cloud Framework for Association Rule Hiding* (Doctoral dissertation).

Verma, N. and Singh, J., 2017. A comprehensive review from sequential association computing to Hadoop-Map reduce parallel computing in a retail scenario. *Journal of Management Analytics*, *4*(4), pp.359-392.

Verma, N., Malhotra, D. and Singh, J., 2020. Big data analytics for retail industry using Map reduce- Apriori framework. *Journal of Management Analytics*, *7*(3), pp.424-442.

Wainakh, A., Grube, T., Daubert, J. and Mühlhäuser, M., 2019, September. Efficient privacy-preserving recommendations based on social graphs. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 78-86).

Wainakh, A., Grube, T., Daubert, J. and Mühlhäuser, M., 2019, September. Efficient privacy-preserving recommendations based on social graphs. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 78-86).

Wang, C. and Zheng, X., 2020. Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. *Evolutionary Intelligence*, *13*(1), pp.39-49.

Wang, C., Bian, W., Wang, R., Chen, H., Ye, Z. and Yan, L., 2020. Association rules mining in parallel conditional tree based on grid computing inspired partition algorithm. *International Journal of Web and Grid Services*, *16*(3), pp.321-339.

Wang, H., Ma, S. and Dai, H.N., 2019. A rhombic dodecahedron topology for human-centric banking big data. *IEEE Transactions on Computational Social Systems*, *6*(5), pp.1095-1105.

Wang, X., Chen, S. and Leung, H., 2018, July. Privacy-Preserving Mining of Association Rule on Outsourced Cloud Data from Multiple Parties. In *Information Security and Privacy: 23rd Australasian Conference, ACISP 2018, Wollongong, NSW, Australia, July 11-13, 2018, Proceedings* (Vol. 10946, p. 431). Springer.

Wazir, S., Beg, M.M. and Ahmad, T., 2020. Comprehensive mining of frequent itemsets for a combination of certain and uncertain databases. *International Journal of Information Technology*, *12*(4), pp.1205- 1216.

Wu, Y. and Zhang, J., 2020. Building the electronic evidence analysis model based on association rule mining and FP-growth algorithm. *Soft Computing*, *24*(11), pp.7925-7936.

Xia, D., Lu, X., Li, H., Wang, W., Li, Y. and Zhang, Z., 2018. A MapReduce-based parallel frequent pattern growth algorithm for spatiotemporal association analysis of mobile trajectory big data. *Complexity*, *2018*.

Yahyaoui, H., Aidi, S. and Zhani, M.F., 2020, January. On using flow classification to optimize traffic routing in SDN networks. In *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1-6). IEEE.

Yan, X., Zhang, J., Xun, Y. and Qin, X., 2017. A parallel algorithm for mining constrained frequent patterns using MapReduce. *Soft Computing*, *21*(9), pp.2237-2249.

Yimin, M., Junhao, G., Mwakapesa, D.S., Nanehkaran, Y.A., Chi, Z., Xiaoheng, D. and Zhigang, C., 2021. PFIMD: a parallel MapReduce-based algorithm for frequent itemset mining. *Multimedia Systems*, pp.1-14.

Zhan, F., Zhu, X., Zhang, L., Wang, X., Wang, L. and Liu, C., 2019, April. Summary of Association Rules. In *IOP Conference Series: Earth and Environmental Science* (Vol. 252, No. 3, p. 032219). IOP Publishing.

Zhang, L., Wang, W. and Zhang, Y., 2019. Privacy preserving association rule mining: Taxonomy, techniques, and metrics. *IEEE Access*, *7*, pp.45032-45047.

Zhang, T., Shi, M., Wang, J. and Yang, G., 2019, April. P-EAARM: A generic framework based on spark for eas-based association rule mining. In *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* (pp. 99-104). IEEE.

Zheng, J. and Zhang, J., 2017, April. Improvement of Apriori algorithm based on matrix compression. In *7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)* (pp. 131-135). Atlantis Press.