iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** X **Month of publication:** October 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue X, Oct 2018- Available at www.ijraset.com

Enhancing Data Performance using Data Mining Techniques

Ranjani Hariharan[1], K. Mythili[2]
[1]*Scholar, Dept. of Computer Science and Applications, SCSVMV University, Kanchipuram, Tamil Nadu*
[2]*Assistant Professor, Dept. of Computer Science and Applications, SCSVMV University, Kanchipuram, Tamil Nadu*

Abstract: Collection of Data and Information has a significant role on developing any technology. With the requirement of wide development in information technologies, Data Mining and Cloud Computing are some of the major technologies that support proper resource sharing. The purpose of this research is to improvise the performance of the cloud database using data mining techniques such as association rule mining with specific reference to the single cache system. One of the most significant data mining process is association rule mining which is used to find frequent patterns in the given dataset.
Keywords: Apriori Algorithm, Cloud Computing, Association Rule Mining, Support, Confidence, Dataset, Frequent Patterns.

I. INTRODUCTION

Data mining involves identification of important trends or patterns through huge amounts of data (Han & Kamber, 2006).Data mining can be defined as a form of database analysis that aims to discover vital relationships as well as patterns in a given group of data ("Anderson," 2014). Advanced statistical techniques like cluster analysis, and in some instances, artificial intelligence and neuronal network techniques are used in the data analysis processes (Han &Kamber, 2006). Discovery of relationships among the data that were previously unknown especially when the data originates from different databases is a principal objective of data mining ("Anderson," 2014). [1]

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. [2]New trends in Internet services that are dependent on clouds of servers to handle tasks are denoted by Cloud computing. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. It allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. [1] This research proposal proposes the concept of data mining association rules for a cloud environment.

A. Problem Definition

Whenever an end user gives information to a cloud, large data blocks are created and stored by various cloud providers of different users. If the user needs to perform various global challenges involving all data blocks frequently, this will result in performance overflow. This degrades the performance as system access data. In order to regularize the occurred overhead, the proposed system model checks the data blocks using Apriori Algorithm.

1) Existing System and Limitations

a) Cloud technology gives the client multirole facilities, it also brings challenges in retrieval of cloud data store.
b) The performance of data retrieval is not efficient using traditional methodologies.
c) The tools and techniques used such as SPSS modeller is less user- friendly and more complex.

2) Proposed System

a) In order to regularize the overflow, the proposed model checks the data blocks.
b) The information is transferred to a data mining tool to develop association rules that guide in determining frequent sets of different items that have 100% accuracy.
c) A java application is created to apply Apriori algorithm to the dataset imported.
d) This can be achieved using association rule mining techniques and other methodologies.

©IJRASET: All Rights are Reserved

3) Objectives of the Study: The objective of the study is to propose "the use data mining association rule in a cloud environment". Thus the objectives are:

a) The Web applications are combined with the cloud services. Therefore it requires frequent access to the cloud databases.

b) Research shows that e-commerce sites require a cloud server and store multiple copies of data sets related to the user queries.

c) To provide suggestions to enhance the performance of cloud service using data mining techniques

II. LITERATURE REVIEW

N.K. Senthil Kumar, M. Uvaneshwari, M. Viswanathan, K. Amsavalli, "One-tier Cache System Applied To Data Mining Techniques To Enhance The Information Security In The Cloud", pp. 1709–1717, 2017: Developed technique focuses on maintaining security of the cloud through data mining techniques. The cloud exports the contents of the database to the SPSS modeler, applies Carma modeling, identifies support and confidence, and filters data or information. It was observed that, with the single cache system, the security of the cloud application could be enhanced. [9]

Usama M. Fayyad (Data Mining and Knowledge Discovery: Making Sense Out of Data, 1996): KDD has grown significantly in the past few years. This growth is driven by a mix of daunting practical needs and strong research interest. The technology for computing and storage has enabled people to collect and store information from a wide range of sources at rates that were, only a few years ago, considered unimaginable. Although modern database technology enables economical storage of these large streams of data, we do not yet have the technology to help us analyze, understand, or even visualize this stored data. [10]

P. Aggarwal and M. M. Chaturvedi, "Application of data mining techniques for information security in a cloud: a survey", pp. 11–17, 2013: Reviewed how data mining techniques and relevant algorithms could play a significant role in ensuring data security in the cloud. The authors noted that problems with data security became crucial. This includes users' authentication services and data encryption and protection. Data mining algorithms offer solutions for identifying and isolating data security attacks. Such attacks may range from information leakage to fraud and infringement. [11]

A. S. Patil, "A review on data mining based cloud computing", pp. 1–14, 2014: Reviewed data mining on the basis of cloud computing, which is a significant characteristic of infrastructure. It aids in making better and more efficient knowledge-driven decisions. It was noted that mining the data in cloud computing permits organizations to centralize software management and the storage of data. This resulted in the assurance of secure, reliable, and efficient client servicers. [8]

III. TOOLS AND TECHNIQUES

Cloud Computing combined with data mining can provide powerful capacities of management. Due to the explosive data growth and amount of computation involved in data mining, an efficient and high performance computing is an excellent resource necessary for a successful data mining application. [3] Data mining techniques and applications are very essential in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. [4] The data mining in Cloud Computing allows organizations to centralize the management of software and data storage. Using data mining through Cloud Computing reduces the barriers that keep small companies from benefiting of the data mining instruments. [5]

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users. [6] The implementation of data mining techniques through Cloud Computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the cost of infrastructure and storage. [5]

A. Mining Association Rules in Large Databases

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

Association rules are of form X -> Y, where X and Y are collection of items and intersection of X and Y is null. Every rule must satisfy two users specified constrains: one is measure of statistical significance called support and other is measure of goodness called confidence.

©IJRASET: All Rights are Reserved

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Figure 1. Association Rules

The problem of mining association rules can be stated as follows: Let I = {i 1 , i 2 , …, i m} be a set of items. Let T = (t 1 , t 2 , …, t n) be a set of transactions (the database), where each transaction t i is a set of items such that t i ⊆I.

A large number of association rule mining algorithms have been developed with different mining efficiencies. Any algorithm should find the same set of rules though their computational efficiencies and memory requirements may be different. The best known mining algorithm is Apriori algorithm. The Apriori algorithm works in two steps :

1) Generate all frequent itemsets: A frequent itemset is an itmset that has transaction support above minimum support.

2) Generate all confident association rules from frequent itemsets: A confident association rule is a rule with confidence above minimum confidence.

The Apriori algorithm relies on Apriori or downward closure property to generate all frequent itemsets. Downward Closure Property: If an itemset has minimum support, then every non empty subset of this itemset also has minimum support.

B. WEKA Tool

WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (WEKA 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of WEKA include: [7]

1) Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.

2) A comprehensive collection of data preprocessing and modeling techniques.

3) Ease of use due to its graphical user interfaces.

IV. METHODOLOGIES

Distributors of cloud provide data of cloud users in the form of a file. These files are broken into blocks/chunks, which are distributed to different cloud servers. The frequently accessed blocks of data by cloud users are stored in cache. The cloud provider uses the one tier cache to answer the queries of the distributor and provides data, rather than searching all the blocks of data.

To have a formal definition, assume that there is a number *T* which supports the threshold. If *S* is an data set, for which *S* is the basket number of subset. Assume *S* is most frequently occurred, which means its support *T* or greater value for T.

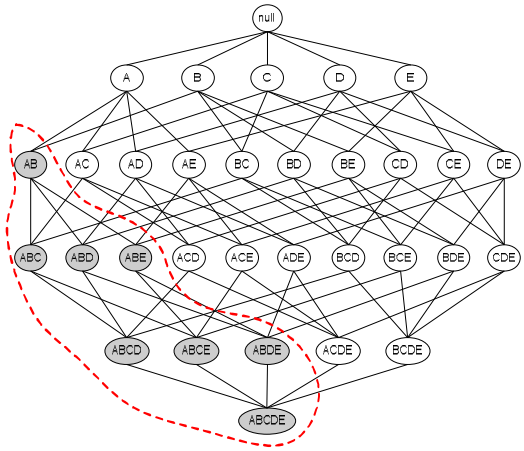


Figure 2.Pruned Supersets

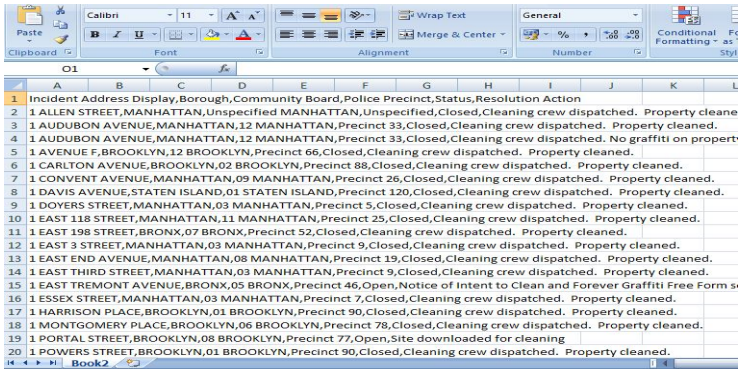
©IJRASET: All Rights are Reserved

Most occurred data sets are denoted as an "if-then" set of rules. Association rules form $S \rightarrow R$, where S is the data set and R is an item. An association rule's implies that if all the S items are in different basket, then R is likely to be seen in the frequently used basket. Whatever, a notion is likely could be formalized by explaining rule emphasizes $S \rightarrow R$ to be the support ratio for $S \cup \{R\}$. Rule of the basket's fraction with all of S that mixes with R.

A. Steps Involved

The end user information is stored in a cloud database. When a user tries to store large amount of data in cloud database the distributors of cloud provide data of users in the form of file. These files are broken into blocks, and then they are distributed to different providers in the cloud. Thus the user information is stored in chunks of data in the cloud database.

1) A Data File Was Imported In File Format: A data file is imported from the cloud database storage containing the user data. The file is imported as an excel file in .csv format where the delimiters are specified. A sample of the file can be seen in the figure given below:



	Incident Address Display,Borough,Community Board,Police Precinct,Status,Resolution Action
2	1 ALLEN STREET,MANHATTAN,Unspecified MANHATTAN,Unspecified,Closed,Cleaning crew dispatched. Property cleane
3	1 AUDUBON AVENUE,MANHATTAN,12 MANHATTAN,Precinct 33,Closed,Cleaning crew dispatched. Property cleaned.
4	1 AUDUBON AVENUE,MANHATTAN,12 MANHATTAN,Precinct 33,Closed,Cleaning crew dispatched. No graffiti on propert
5	1 AVENUE F,BROOKLYN,12 BROOKLYN,Precinct 66,Closed,Cleaning crew dispatched. Property cleaned.
6	1 CARLTON AVENUE,BROOKLYN,02 BROOKLYN,Precinct 88,Closed,Cleaning crew dispatched. Property cleaned.
7	1 CONVENT AVENUE,MANHATTAN,09 MANHATTAN,Precinct 26,Closed,Cleaning crew dispatched. Property cleaned.
8	1 DAVIS AVENUE,STATEN ISLAND,01 STATEN ISLAND,Precinct 120,Closed,Cleaning crew dispatched. Property cleaned.
9	1 DOYERS STREET,MANHATTAN,03 MANHATTAN,Precinct 5,Closed,Cleaning crew dispatched. Property cleaned.
10	1 EAST 118 STREET,MANHATTAN,11 MANHATTAN,Precinct 25,Closed,Cleaning crew dispatched. Property cleaned.
11	1 EAST 198 STREET,BRONX,07 BRONX,Precinct 52,Closed,Cleaning crew dispatched. Property cleaned.
12	1 EAST 3 STREET,MANHATTAN,03 MANHATTAN,Precinct 9,Closed,Cleaning crew dispatched. Property cleaned.
13	1 EAST END AVENUE,MANHATTAN,08 MANHATTAN,Precinct 19,Closed,Cleaning crew dispatched. Property cleaned.
14	1 EAST THIRD STREET,MANHATTAN,03 MANHATTAN,Precinct 9,Closed,Cleaning crew dispatched. Property cleaned.
15	1 EAST TREMONT AVENUE,BRONX,05 BRONX,Precinct 46,Open,Notice of Intent to Clean and Forever Graffiti Free Form s
16	1 ESSEX STREET,MANHATTAN,03 MANHATTAN,Precinct 7,Closed,Cleaning crew dispatched. Property cleaned.
17	1 HARRISON PLACE,BROOKLYN,01 BROOKLYN,Precinct 90,Closed,Cleaning crew dispatched. Property cleaned.
18	1 MONTGOMERY PLACE,BROOKLYN,06 BROOKLYN,Precinct 78,Closed,Cleaning crew dispatched. Property cleaned.
19	1 PORTAL STREET,BROOKLYN,08 BROOKLYN,Precinct 77,Open,Site downloaded for cleaning
20	1 POWERS STREET,BROOKLYN,01 BROOKLYN,Precinct 90,Closed,Cleaning crew dispatched. Property cleaned.

Figure 3.Imported data file

2) A JAVA Application Is Created: A JAVA application is created which acts an user interface to deploy and process association rule mining techniques. Apriori algorithm was selected to process and find the frequent occurring set of items and association rules from the given input data file.

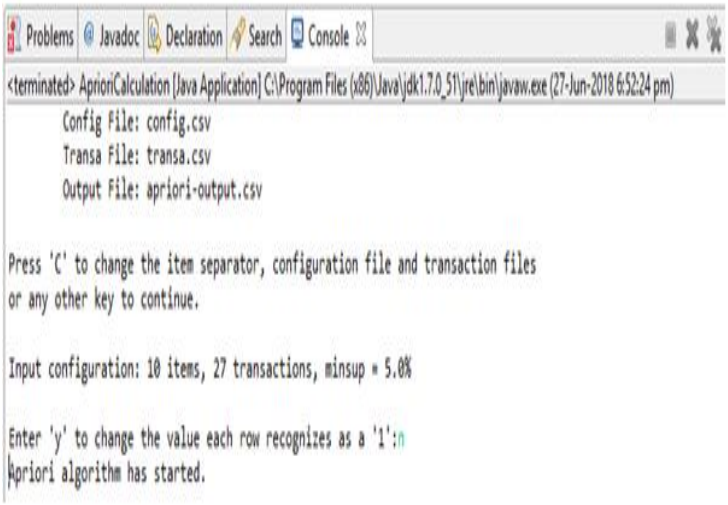


Figure 4.Java.Console Application

3) Data Filtration: The given data were filtered for acquiring the threshold values with those have 100 percent accuracy. The Apriori algorithm process and filters only those data which satisfies the rules and has 100% confidence.

4) Output Retrieved: Information with 100 percent accuracy and efficiency was retrieved. Thus the filtered data with 100% confidence is then stored in an output file.

©IJRASET: All Rights Reserved

B. WEKA Tool Process

The given input file is processed using WEKA for further analysis and comparison.

1) The input file in excel format is processed with WEKA using Apriori algorithm.

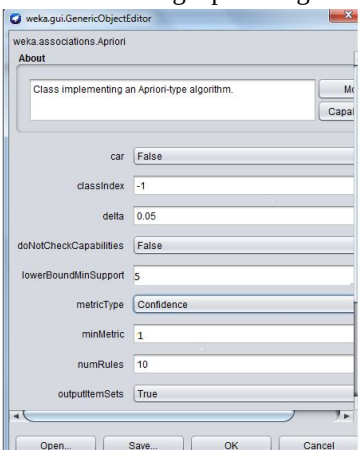


Figure 5.WEKA Rule Generator

V. RESULT AND DISCUSSION

No one can claim that it is possible to develop a hundred percent secure network that is immune to all types of attacks. As new threats and compromises are being initiated every day, system developers must improve their countermeasures to keep data private and secure. In light of the tremendous benefits offered by cloud computing, more and more organizations chose to utilize their services to improve performance and decrease cost. However, cloud providers are susceptible to attacks and security threats from insiders and outsiders.

Attackers could use multiple computing techniques to extract information about the user from the data stored in the cloud. [12] This research proposes cloud architecture to increase security and minimize the effect of such malicious attacks. However, this might result in overheads since users might require accessing certain data components very frequently. Hence cache memory concept was implemented in the proposed system by generating frequent item sets using data mining tools. [12] Even if an attacker gains access to the cloud provider's storage space, only one chunk of data will be exposed. Even though this architecture increases data security, it involves a considerable amount of overhead if the user decides to access the whole data set frequently. To overcome this drawback, data mining algorithms are used to determine the most frequently used data chunks.

A. Output Using WEKA Tool

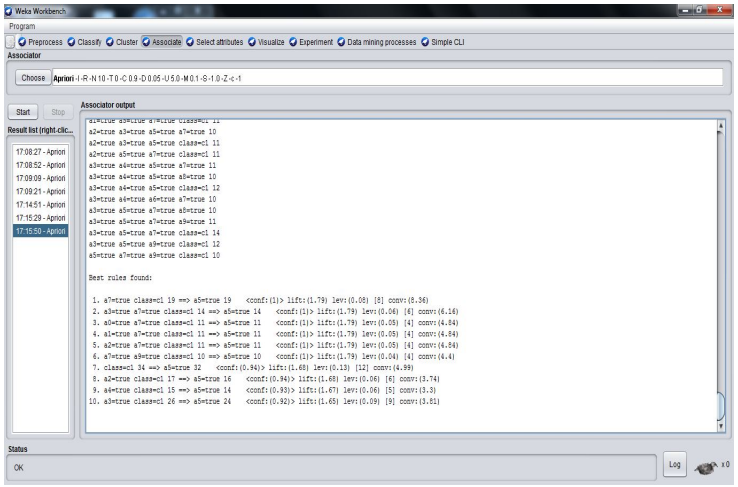


Figure 6.Output in WEKA

1) As discussed, the calibration of the inputs in both JAVA and WEKA in proposed system shows optimized results and they are thus more efficient than the existing system.

©IJRASET: All Rights are Reserved

VI. CONCLUSION

The novel approach of this research is to improve the performance through data mining techniques with major rule of using Apriori algorithm. This proposal is limited to particular cloud computing applications and the research specifies on the use of a one tier cache system. As given detail in the research, e-commerce sites require a cloud server and store multiple often data sets related to the end user who checks in their websites. The Web applications are combined with the cloud services given. In addition, the cloud databases are updated frequently by the imported user. Thus a java application was successfully created to apply association mining techniques using Apriori algorithm, which identifies information with 100 percent confidence. It is observed - the one tier cache system improves the performance of the cloud application to the greater extent.

A. Future Work

1) Setting up multiple minimum support values if various items present in the database.

2) Hash tree based Apriori algorithm can be used to increase the optimization.

REFERENCE

[1] Hamza Ahmed, “Data Mining in Cloud Computing”, International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015

[2] Omkar Singh Lodhi, “Most Commonly Used Techniques in Data Mining”. International Journal of Advances in Engineering Research2012, Vol. No. 4, Issue No. III, September

[3] CH.Sekhar, S Reshma Anjum, “Cloud Data Mining based on Association Rule”, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014

[4] K. Kala Bharathi, K. Sandhya Sree, “”,International Journal of Computer Science Engineering and Technology, Feb 2015, Vol 5, Issue 2

[5] Ruxandra, Stefania, “Data mining in Cloud Computing”, Database Systems Journal vol. III

[6] Ms. Aishwarya S. Patil, Ms. Ankita S. Patil, “A review on Data Mining based Cloud Computing”, International Journal of Research In Science & Engineering e-ISSN: 2394-8299 Volume: 1 Special Issue: 1

[7] The University of Waikato[NZ]. [Online]. Available https://www.cs.waikato.ac.nz/ml/weka/

[8] A. S. Patil, “A review on data mining based cloud computing,” International Journal of Research in Science and Engineering, vol. 1, no. 1, pp. 1–14, 2014.

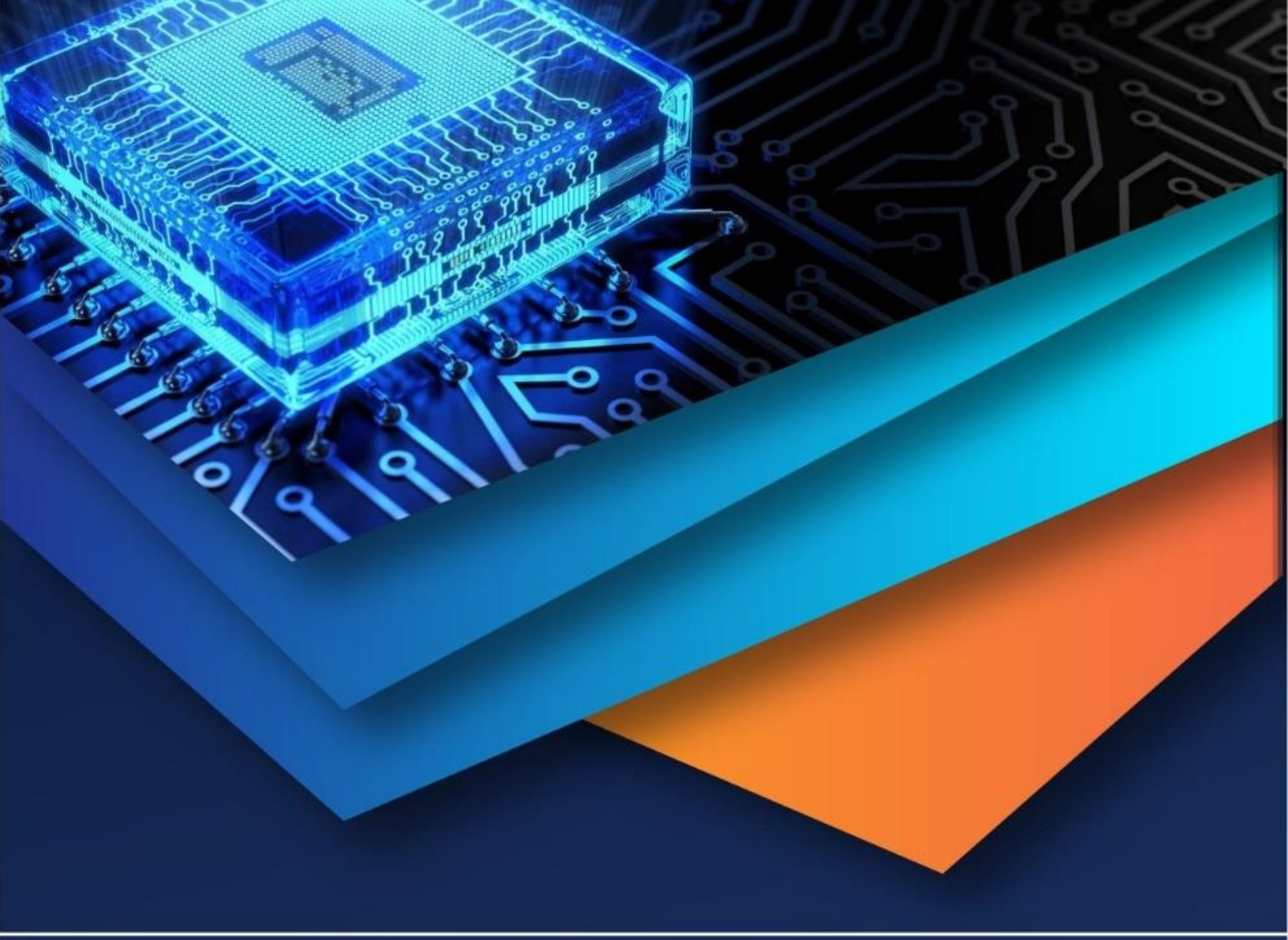
[9] N.K. Senthil Kumar, M. Uvaneshwari, M. Viswanathan, K. Amsavalli, "One-tier Cache System Applied To Data Mining Techniques To Enhance The Information Security In The Cloud", pp. 1709–1717, 2017

[10] Usama M. Fayyad (Data Mining and Knowledge Discovery: Making Sense Out of Data, 1996)

[11] P. Aggarwal and M. M. Chaturvedi, “Application of data mining techniques for information security in a cloud: a survey”, pp. 11–17, 2013:

[12] Srishti Sharma, Harshita Mehta, “Improving Cloud Security Using Data Mining”, IOSR Journal of Computer Engineering (IOSR-JCE) Volume 16, Issue 1, Ver. II (Jan. 2014)

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)