iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

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

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Mining of Frequent Itemset in Hadoop

Kavitha Mohan[1], Talit Sara George[2]
[1,2] PG Scholar, Asst Prof, Computer Science and Engineering, Kerala Technological University

Abstract*: Information are produced from different sources, the quick move from computerized advances has prompted development of huge information. Terabytes of information is produced every day by modern information systems and digital technologies like Internet of Things and distributed computing. In Data Mining, it takes out information from large sets of data or it is the procedure of knowledge mining from large collection of data. The large amount of data is available in the Information Industry. FIM is the most important concept in mining. Association Rule mining and Frequent item set mining are well known techniques for data mining which needs whole dataset into fundamental memory (main memory) for processing, however extensive datasets do not fit into fundamental memory. To defeat this limitation MapReduce is utilized for parallel processing of Big Data having elements such as high scalability and robustness which helps to handle problem of large datasets. The Big Data mining is essential in order to take out value from massive amount of data which give better insights using proficient techniques. This paper investigates the potential effect of FIM in big data and data mining.
Keywords— Big Data, Data Mining, Frequent Item Set Mining, Association Rule mining, MapReduce.

I. INTRODUCTION

A. Concept of Big Data

Terabytes of information originates from all over the places: sensors used to accumulate atmosphere data, posts to social media sites, advanced pictures and videos, purchase transaction records, and mobile phone GPS signals to name a few. This tremendous amount of the information is known as “Big data". Big data is the term for any collection of datasets so extensive and complex that it becomes hard to process using traditional data processing applications [1]. There are distinctive methods for characterizing and comparing Big Data with the conventional information for example, size of data, content, collection and processing. Big data has been characterized as large data sets that cannot be processed using traditional processing techniques, such as Relational Database Management Systems. BigData is either a relational database (Structured), such as stock market data or non-relational database (Semistructured or Unstructured), such as social media data or DNA data sets [2]. The Big Data characterized by 4 V’s: 1)Velocity, 2)Veracity, 3)Volume, 4)Variety.

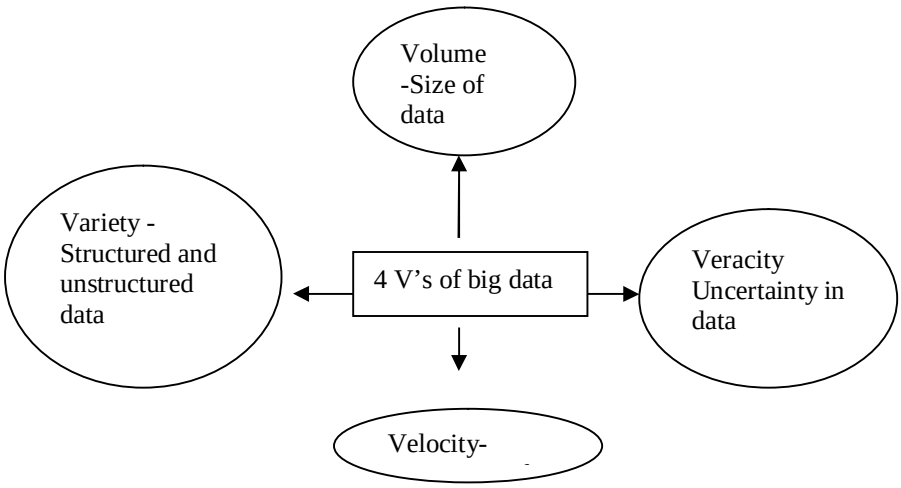


Fig. 1. 4V’s of Big Data

1) *Velocity:* Velocity is the frequency or recurrence of incoming data that should be processed. Big Data Velocity manages with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, online networks, cell phones, etc. It means the speed at which the information is produced.
2) *Veracity:* Big Data Veracity refers to the biases, noise and abnormality or uncertainty in data. Is the information that is being stored, and mined significant to the issue being analyzed. Veracity in information analysis is the greatest challenge when

compares to things like volume and speed.

3) *Volume:* Volume of the data, which means the data size or amount of data. Some of companies' data storage is about Zetabyte.
4) *Variety:* Variety refers to the many sources and types of data both structured and unstructured. It means the data forms that distinctive or different applications manage such as sequence data, grouping information numeric data or binary information.

B. Concept of Hadoop and Map Reduce in Big Data

Hadoop was established by Apache. It is an Open source software project written in Java. It is utilized to optimize huge volume of information. It is a software framework for distributed processing of large dataset across huge groups of item servers [3]. It shared the records or files among the nodes and permits to framework proceed with work if there should be an occurrence of a node failure. This approach diminishes the danger of catastrophic system failure. In which application is broken into smaller parts (pieces or blocks or fragments). Scalability, flexibility, fault tolerance etc are the characteristics of hadoop. Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System (HDFS) consists of three Components: the Name Node, Secondary Name Node and Data Node [4].

MapReduce is a programming model used by Google to process large amount of data in a distributed computing environment. It is usually used to perform distributed computing on clusters of computers [5]. Mapper and reducer are the two phases, it is sorting and filtering the input data. In Map phase data or information is shared to mapper machines and by parallel processing it produces <key, value> pairs for each record. Next shuffle phase is used for repartitioning and sorting that pair within each partition [6]. So the value corresponding same key grouped into {v1, v2,….}values. Reduce phase reducer machine process subset pairs parallel in the final result is written to distributed file system.

C. Brief idea about data mining

Data Mining is a procedure of extracting valuable knowledge from different perspectives or massive amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or 'Knowledge mining' or 'Knowledge Discovery from Data' [7]. It is an interdisciplinary subfield of computer science which involves computational process of large data sets patterns discovery. The goal of this advanced analysis process is to extract information from a data set and transform it into an understandable structure for further use. Seven important steps are used for KDD process; they are Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, Knowledge Representation.

1) *Data Cleaning:* Cleaning the unwanted data or remove irrelevant data.
2) *Data Integration:* In this step, combining the different data sources into single data store. ie, target data.
3) *Data Selection:* In this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4) *Data Transformation:* In this step, the selected data is transforms into appropriate form.
5) *Data Mining:* It is the important step, different techniques are applied for extract data patterns and rules.
6) *Pattern Evaluation:* Strictly identify tree patterns representing knowledge.
7) *Knowledge Representation:* This is the final stage, discovered knowledge is visually represented to the user.

II. DATA MINING TECHNIQUES

Several core techniques that are used in data mining describe the type of mining and data recovery operation. Association Rule, Frequent Itemset mining, Classification, Clustering, Prediction, Sequential Patterns, etc. Here, explained only about Association Rule Mining and Frequent Itemset mining.

A. Association Rule Mining

Association rule mining find frequent patterns, correlations among the items or transactional database, or relational database. Association rule can create analyzing data for frequent pattern using the criteria Support & Confidence to identify the relationship. Support is indicating of how frequently the item appears in the database. Confidence indicates the number of time has been found. There are many algorithms used in association rule mining. FP Growth Algorithm, Apriori Algorithm, ECLAT Algorithm, ASPMS Algorithm, RARM Algorithm etc.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

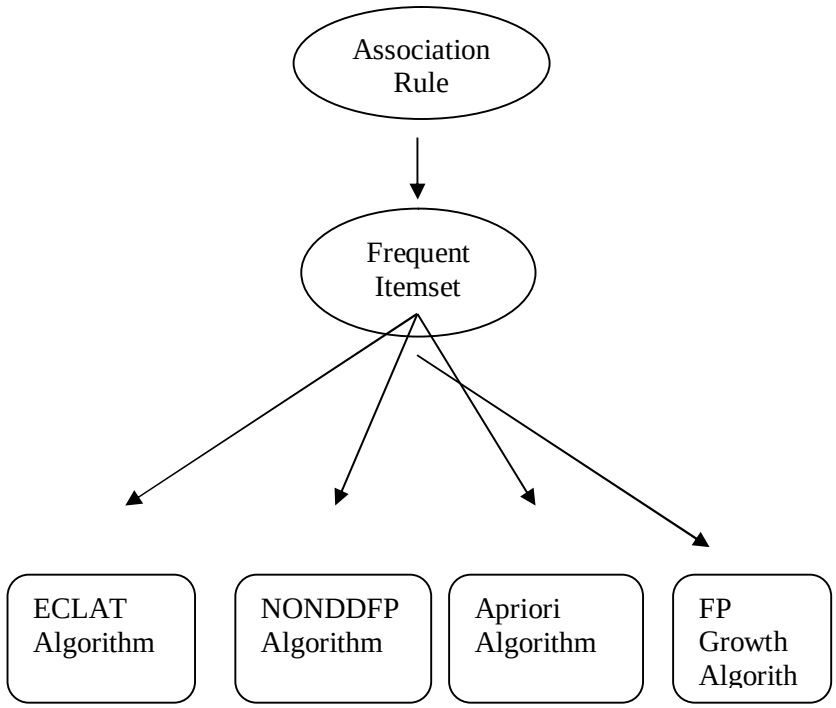


Fig.2. Different types of Algorithm used in Association Rule Mining

B. Frequent Itemset Mining

Frequent itemset mining is the major problem in data mining. Speeding up the process of FIM is critical task because mining time is high. In modern data mining application datasets are very large, so FIM tolerate performance issues [9]. A remarkable progress in this field has been made and lots of efficient algorithms have been designed to search frequent patterns in a transactional database. Frequent itemset mining can be used in a variety of real world applications. It can be used in super markets for selling, product placement on shelves etc [9].

The papers related to frequent itemset mining in big data and data mining are discussed in the following section:

M.-Y. Lin *et al.* [10] proposed three algorithms named SPC, FPC, and DPC. By using this algorithm investigate effective implementations of the Apriori algorithm in the MapReduce framework. The Apriori algorithm that mines frequent item sets is one of the most popular and widely used data mining algorithms. For improving the performance of this algorithm different parallelization techniques are used. DPC features in dynamically combining candidates of various lengths and outperforms both the straight-forward algorithm SPC and the fixed passes combined counting algorithm FPC. Extensive experimental results of this paper also shows that all the three algorithms scale up linearly with respect to dataset sizes and cluster sizes.

Sheetal Labade *et al.* [11] analyzed different methods for frequent itemset mining and also explained different privacy handling in present systems. The methods are horizontal data format, vertical data format, projected database techniques. In horizontal data format defined as mines frequent itemsets from the set of transactions in Transaction Identifier-itemset arrangement.TID is the ID of transaction and itemset is the set of items accepted in transaction TID. Apriori Algorithm, Sampling algorithm, DIC technique are work in horizontal data format. In vertical data format, item-TID_set, where item is an item name, and group of TIDs indicates the set of transactions in which item is present. Eclat algorithm is work in this method. In projected database techniques uses divide and conquer method to mine frequent itemsets. FP-Growth algorithm, CT-PRO Algorithm, H-mine algorithm are comes under this method.

Yaling Xun *et al.* [12] proposed a data distribution scheme method for partitioning the data. The effect of partitioning is balancing the loads in the clusters. FIUT method enhance the efficiency of frequent itemset mining. In this paper, Apriori and FP-Growth algorithms are used. The drawback of Apriori is that it generates large number of candidate itemsets and scanning the database repeatedly. To overcome the limitataion of Apriori, FP-Growth is used. Map Reduce programming model is applied for mining itemsets. This model can solve the scalability and performance issues. The Map Reduce job act an important role in mining frequent itemsets.

III. CONCLUSION

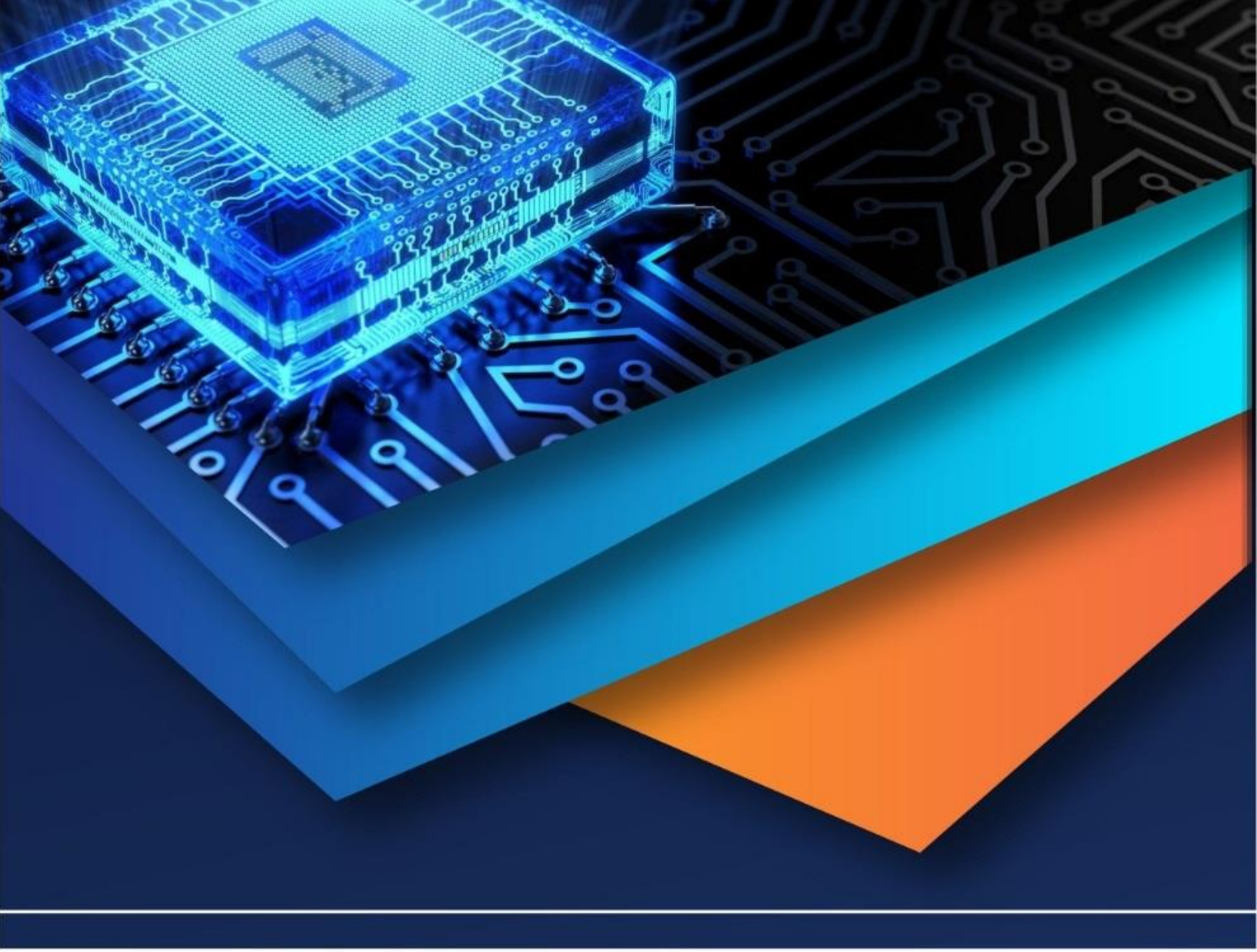
Big data is currently evolving technology. Big Data typically refers to any large quantity of raw data that cannot be collected, stored, and analyzed by normal means. Hadoop framework is used to process big data. The frequent itemset mining algorithm is one of the most commonly used algorithm in datamining which has a larger running time. By implementing FIM in hadoop its running time can be reduced and its efficiency can also be improved. This paper analysis various approaches for FIM in big data.

IV. ACKNOWLEDGMENT

I am extremely grateful to my PG Coordinator Asst. Prof. Salitha M K, Assistant Professor of Computer Science and Engineering department and my guide Asst. Prof. Talit Sara George, Assistant Professor of Computer Science and Engineering department Caarmel Engineering College, Pathanamthitta for her valuable guidance, timely suggestions and for providing all the vital facilities like providing the Internet facilities and important books, which were essential in the completion of this work. Finally, I would like to thank every individual who gave me even the slightest support to make my paper a successful one.

REFERENCES

[1] https://en.wikipedia.org/wiki/Big_data.
[2] Hamoud Alshammari, Jeongkyu Lee and Hassan Bajwa, "H2Hadoop: Improving Hadoop Performance using the Metadata of Related Jobs",IEEE Transactions,2015.
[3] D.Usha and Aslin Jenil A.P.S, "A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce", International Journal of Current Engineering and Technology,vol.4, no. 2, April 2014.
[4] Tripti Mehta, Neha Mangla, GITM Guragon, "A Survey Paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework", International Journal of Recent Advances in Engineering & Technology, February 2015.
[5] Liya Thomas, Syama R, "Survey on MapReduce Scheduling Algorithms", International Journal of Computer Applications, Vol. 95, No. 23, pp. 0975 – 8887, June 2014.
[6] Shital Suryawanshi, Prof. V.S.Wadne, "Big Data Mining using Map Reduce: A Survey Paper", IOSR Journal of Computer Engineering, Vol. 16, No. 6, PP 37-40, Nov – Dec. 2014.
[7] Nikita Jain, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER", International Journal of Research in Engineering and Technology, Vol. 02, Nov-2013.
[8] Rana Ishita, Amit Rathod, "Frequent Itemset Mining in Data Mining: A Survey", International Journal of Computer Applications, vol. 139, no. 9, April 2016.
[9] Dr. R Nedunchezhian, K Geethanandhini, "Association Rule Mining on Big Data – A Survey", International Journal of Engineering Research & Technology, vol. 5, no. 5, May 2016.
[10] Ming-Yen Lin, Pei-Yu Lee, Sue-Chen Hsueh, "Apriori-based Frequent Itemset Mining Algorithms on MapReduce", ACM, ICUIMC'12, February 2012.
[11] Sheetal Labade, Srinivas Narasim Kini, "A Survey Paper on Frequent Itemset Mining Methods and Techniques", May 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY