ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

FP Tree Algorithm and Approaches in Big Data

T.Rathika[1], J.Senthil Murugan[2]

Assistant Professor, Department of CSE, SRM University, Ramapuram Campus, Chennai, Tamil Nadu,India[1]

Assistant Professor, Department of MCA, VelTech HighTech Engineering College, Chennai, Tamil Nadu, India[2]

ABSTRACT: Data mining is process of extracting a large amount of data that need to be analysed patterns have to be extracted from that to share knowledge. In this new era with rumble of data together structured, semi-structured and unstructured, in the field of machine learning, educational data mining, web mining and text mining, environmental research areas, it has become difficult to process, manage and analyzed patterns using traditional databases and relational databases So, a appropriate architecture should be understood to gain and share knowledge about the Big Data. This paper presents an analysis of various algorithms from for handling such big data set. These algorithms define various structures and methods implemented to handle Big Data, also in the paper are listed various data mining tools that were developed for analyzing them.

KEYWORDS: Data Mining, Big Data, Clustering. Frequent Pattern, Association Rule,Hadoop and MapReduce framework.

I. INTRODUCTION

Data Mining (DM) is known as the process of extracting useful patterns or information from huge amount of data. Most of the research people think about data mining as a knowledge discovery databases. Data mining task is the automatic or semi-automatic analysis of large quantities of data to extract unknown interesting patterns such as groups of data records (cluster analysis), unusual records and dependencies (association rule mining).Data Mining is the new technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data. The data mining environment produces a large volume of the data. The information retrieved in the data mining step is transformed into the structure that is easily understood by its user. Data mining involves various methods such as frequent pattern tree algorithm, association rule and cluster analysis, to disclose the hidden patterns inside the large data set.

Big data are the large amount of data being processed by the data mining environment. In other words, it is the collection of data sets massive and complex that it becomes difficult to process using on hand relational database management tools or traditional data processing applications, so data mining tools were used. Big data are about turning unstructured, invaluable, imperfect, complex data into usable information. Data have hidden information in them and to extract this new information; interrelationship among the data has to be achieved. An information may be retrieved from a hidden or a complex data set. The standard data analysis method such as exploratory, clustering, factorial, analysis need to be extended to get the information and extract new knowledge treasure. Big data can be measured using the following that is categorized by four dimensions of V's:, volume, velocity , variety and veracity;

1. Volume –The amount of data is at very large scale. The amount of information being collected is so huge that modern database management tools are unable to handle it and therefore become obsolete.

2. Velocity -We are producing data at an exponential rate .It is growing continuously in terabytes and peta-bytes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

3. Variety -We are creating data in all forms -unstructured, semi structured and structured data. This data is heterogeneous is nature. Most of our existing tools work over homogenous data, now we require new tools and techniques which can handle such a large scale heterogeneous data.

4. Veracity-The data we are generating is uncertain in nature. It is hard to know which information is accurate and which is out of date.

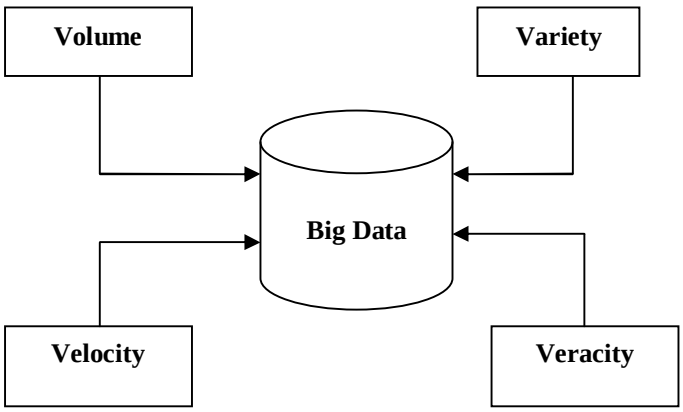

Fig.1. Categorized of big data in 4 V's.

The paper reviews different aspects of the big data. The paper has been described as follows, Section I: Introduction about Big data. Section II. Related Work for big data .Section III: deals with the architecture of the big data. Section IV: describes the various algorithms used to process Big Data. Section V: describes different big data technology and tools.

II. RELATED WORK

R.Agarwal proposes data mining association rule for big data sets .Various fast algorithm for scattered system is proposed Cheun.D.W proposes a rapid distributed algorithm for data mining association rules by reducing the number of messages passed.Thabet Slimani proposes current trend in association rule mining and compare the huge performance of different algorithms. The reveals challenges and opportunities in databases in existence of big data. As a recent effort introduces contributions on optimizing big data processing efficiency in Hadoop and MapReduce. Sam Madden discusses about the traditional databases and the databases required with Big data concluding that the databases don't solve all aspects of the Big data problem and the association rule mining algorithms need to be more robust and easier for unsophisticated users to apply.An architectural considerations for Big data are discussed concluding that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability.

III. BIG DATA ARCHITECTURE

Big data for development is about turning imperfect, complex, often unstructured data into actionable information. This implies leveraging advanced computational tools which have developed in other fields. Big data means enormous amounts of data, such large that it is difficult to collect, store, manage, analyze, predict, visualize, and model the data. Big data analytics refers to tools and methodologies that aim to transform massive quantities of raw data into "data about data" for analytical purposes. Big data are the collection of large amounts of unstructured data. Big data means

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

enormous amounts of data, such large that it is difficult to collect, store, manage, analyze, predict, visualize, and model the data. Big data architecture typically consists of three segments: storage system, processing and analysis.

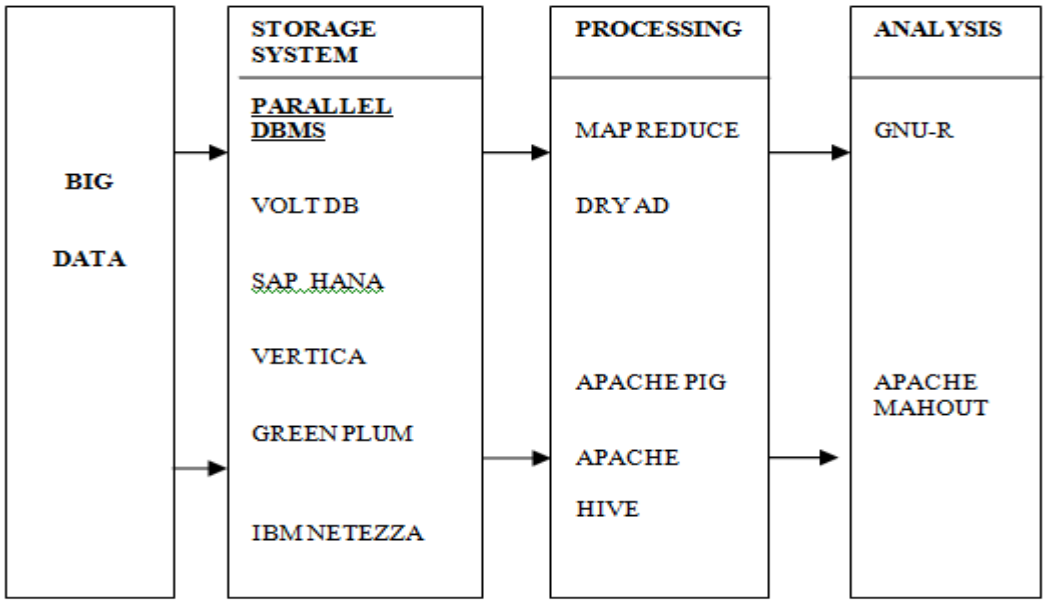

Fig. 2. Architecture of Big Data

IV. ALGORITHM

Many algorithms were defined earlier in the analysis of large data set. Will go through the different work done to handle Big data. First one is called as association rule mining, second is clustering and finally is called as frequent item set.

1. Association Rule Mining: Association is about to identify the interrelation between different attributes as well as attribute values. Association mining is most useful mining approach used as an individual process as well as a stage of some data mining model to improve the accuracy in results. It basically states about the relationship between the data values. It is associated with different mining operations such as data cleaning, classification, filtration approach etc. Association mining can be done horizontally or vertically. The levels of association mining are also considered by defining the number of associated attributes

2. Clustering: It is never effectual to process on large dataset at one time. To obtain the quick running results from the large datasets, complete dataset is divided in smaller datasets. There are number of existing clustering approaches to collect the similar dataset in one cluster. Clustering is basically performed on distance based analysis such as Euclidian distance analysis. There are number of existing clustering approaches such as Kmeans approach, Fuzzy CMeans approach, Hierarchical Clustering etc.

3. Frequent Itemset: The item set which satisfies the criteria of being greater or equal to minimize support are frequent item set. It is denoted by L i, i= itemset. If itemset does not satisfy the criteria it is non frequent itemset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

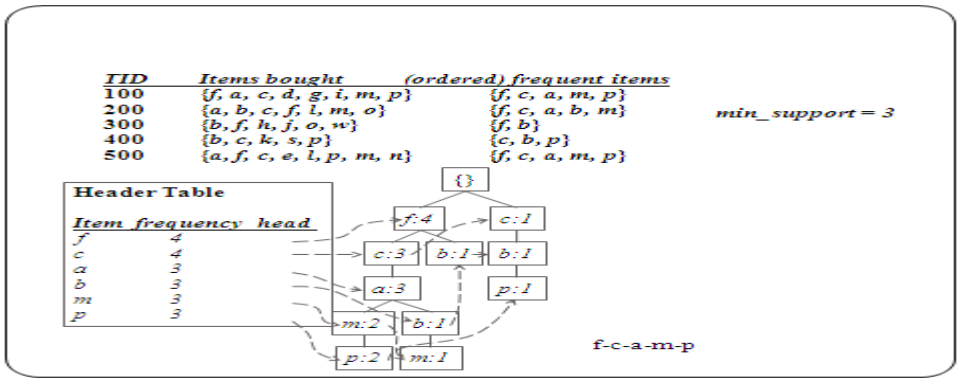

Table.1.Frequent Itemset criteria

V. BIG DATA TECHNOLOGY AND TOOLS

There are varieties of applications and tools developed by various organizations to process and analyse the big data. The Big data analysis applications support parallelism with the help of computing clusters. These computing clusters are collection of hardware connected by Ethernet cables. The following are major applications in the area of big data analytics.

I. MapReduce: MapReduce is a programming model for computations on massive amounts of data and an execution framework for large-scale data processing on clusters of commodity servers. It was originally developed by Google and built on well -known principles in parallel and distributed processing. MapReduce program consists of two functions are as Map function and Reduce function. MapReduce computation executes as follows;

1. Each Map function is converted to key-value pairs based on input data. The input to map function is tuple or document. The way key -value pairs are produced from the input data is determined by the code written by the user for the Map function.

2. The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.

3. The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by the user for the Reduce function.

Major Advantages of MapReduce:

1. The MapReduce model hide details related to the data storage, distribution, replication, load balancing and so on.

2. Furthermore, it is so simple that programmers only specify two functions, which are map function and reduce function, for performing the processing of the Big Data.

3. MapReduce has received a lot of attentions in many fields, including data mining, information retrieval, image retrieval, machine learning, and pattern recognition.

II. Hadoop : Hadoop is a free, Java based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was inspired by Google's MapReduce Programming paradigm.Hadoop is a highly scalable compute and storage platform. But on the other hand, Hadoop is also time consuming and storage-consuming. The storage

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

requirement of Hadoop is extraordinarily high because it can generate a large amount of intermediate data. To reduce the requirement on the storage capacity, Hadoop often compresses data before storing it. Hadoop takes a primary approach to a single big workload, mapping it into smaller workloads. These smaller workloads are then merged to obtain the end result. Hadoop handles this workload by assigning a large cluster of inexpensive nodes built with commodity hardware. Hadoop also has a distributed, cluster file system that scales to store massive amounts of data, which is typically required in these workloads. Hadoop has a variety of node types within each Hadoop cluster; these include DataNodes, NameNodes, and EdgeNodes. The explanations are as follows:

A.NameNodes: The NameNodes the central location for information about the file system deployed in a Hadoop environment. An environment can have one or two NameNodes, configured to provide minimal redundancy between the NameNodes. The NameNodes contacted by clients of the Hadoop Distributed File System (HDFS) to locate information within the file system and provide updates for data they have added, moved, manipulated, or deleted.

B..DataNodes: DataNodes make up the majority of the servers contained in a Hadoop environment. Common Hadoop environments will have more than one DataNodes, and oftentimes they will number in the hundreds based on capacity and performance needs. The DataNodes serves two functions: It contains a portion of the data in the HDFS and it acts as a compute platform for running jobs, some of which will utilize the local data within the HDFS.

C.EdgeNodes: The EdgeNodes the access point for the external applications, tools, and users that need to utilize the Hadoop environment. The EdgeNodes sits between the Hadoop cluster and the corporate network to provide access control, policy enforcement, logging, and gateway services to the Hadoop environment. A typical Hadoop environment will have a minimum of one EdgeNodes and more based on performance needs.

III. IBM InfoSphere: It is an Apache Hadoop based solution to manage and analyze massive volumes of the structured and unstructured data. It is built on an open source Apache Hadoop with IBM big Sheet and has a variety of performance, reliability, security and administrative features.

IV.Spreadsheet-Style Analysis(SSA): Web-based analysis and visualization .Define and manage long running data collection jobs and analyze content of the text on the pages that have been retrieved.

V.RHadoop: Statistical tools for managing big data.

VI.Mahout: Data mining and machine learning tools over big data.

VI. CONCLUSION

The paper is a study of different applications and tools of Big Data analytics. Big Data is a very challenging and recent trend research area in the IT industry. Data is too big to process using traditional tools of data processing applications. Academia and industry has to work together to design and develop new tools and technologies which effectively handle the processing of Big Data. Big Data is an emerging trend and there is instant need of new machine learning and data mining techniques to analyze massive and complex amount of data in near future. Hadoop and Map Reduce tool for big data is described in detail focusing on the areas where it needs to be improved so that in future Big data can have technology as well as skills to work with.

REFERENCES

[1].Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012
[2].Joseph McKendrick, "Big Data, Big Challenges,Big Opportunities: 2012 IOUG Big Data Strategies Survey", IOUG, Sept 2012.
[3] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE, 2012
[4].Challenges and Opportunities with Big Data", 2012.
[5]Big Data Survey Results", Treasure Data, 2012
[6]. MVijayalakshmi, M.Renuka Devi, "A Survey of Different Issues of Different Clustering Algorithms used in Large Data Sets", International Journal of Advanced Research in Computer Science and Software Engineering, March 2012.
[7].Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,"Data mining with Big Data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

[8]. Girola11] Girola, Michele, et al. "IBM Data Center Networking: Planning for virtualization and cloud computing." GOOGLE/IP. COM/IBM Redbooks (2011).
[9]. Dittrich, Jens, and Jorge-Arnulfo Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." Proceedings of the VLDB Endowment 5.12 (2012): 2014-2015.
[10]. Sam Madden, " From Databases to Big Data", IEEE, Internet Computing,May-June 2012.
[11]Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012
[12]Joseph McKendrick, "Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey", IOUG, Sept 2012