ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2015

Hadoop: A Challenging Approach to Mine Big Data

D.Kerana Hanirex , Dr.K.P.Kaliyamurthie

Assistant Professor, Dept. of CSE, Bharath University, Chennai, India

HOD, Dept. of CSE, Bharath University, Chennai, India

ABSTRACT: The most efficient way to manage and obtain useful information from the database is one of the emerging trends nowadays. The quantity of data stored in the database is getting increased daily.We need some method to manage all the information. In this paper we discuss the current and future trends of mining the data and the challenges to overcome in future

I.INTRODUCTION

The quantity of data created everyday is more than a Exabyte.So it is very difficult to manage the data and it opens a new challenging area. It is very difficult to manage the data in real time such as sensor networks,network monitoring etc.In the data stream model, data arrive at high speed, and we need efficient algorithms that process data.In addition due to this size and level of structure it is not possible to analyze traditional databases and methods in an efficient way.These kinds of big data problems require new tools and technologies to store ,manage and realize the benefit.The high volume of data is one of the characteristics of big data.Since most of the bigdatais unstructured and semi structured it requires techniques and tools to process and analyze. This paper proposes various algorithm design based on their space and time.We need to deal with resources in an effcient and low-cost way. Data stream mining consider the factors of amount of accuracy,space and the time required to predict.By storing more information in the look-up tables and adjusting the time and space we can get the efficieny algorithm which works very faster.

One more issue in managing data stream information is cost in terms of space and money.Many rental cost options are provided by various web services such as Amazon Elastic Compute Cloud (Amazon
EC2) which gives resizable capacity of memory. GoGrid is a web service similar to Amazon EC2, but it charges by RAM-Hours.In [1, 2] the use of RAM-Hours was introduced as an evaluation measure of the resources used by streaming algorithms. very GB of RAM deployed for 1 hour equals one RAM-Hour.

One of the challenging task is the structured pattern classification problem. Patterns may be itemsets,sequences, trees and graphs. The structured pattern classification predict to a particular class and develop a model to identify the future pattern of the input.Classification problem deal the vector data also.To apply the data to the different pattern such as trees and graphs we have to convert the standard classification problem into vector classification problem by applying various transforming techniques.

While classifying the data the dataset may have more than a number of attributes.We can reduce the number of attributes by performing feature selection. One way to deal with a structured output classification problem is to convert it to a multilabel classification problem, where the output pattern y is converted into a set of labels representing a subset of its frequents subpatterns.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2015

II. HADOOP

Cloud computing is one of the emerging technique which allows consumers and businesses to use applications without any installation and access their files at any computer with internet access. We can do efficiently by centralizing storage, memory, processing and bandwidth. Since we have to pay only for service and its usage it is a reliable and inexpensive option for association rule mining.

Hadoop is a technology which provide scalability and reliability option for to a distributed system. Hadoop provides tools for analyzing both structured and unstructured data. MapReduce and HDFS of Hadoop use simple, robust techniques on inexpensive computer systems to deliver very high data availability and to analyze enormous amounts of information quickly[3]. The reliable part of mining is taken care by the Hadoop framework. Thus the association rule mining is done on Hadoop and cloud so that outcome is a reliable and efficient. Mining association rules may require iterative scanning of large transaction or relational distributed databases, which is quite costly in processing. It aims to show that Hadoop along with cloud computing can be considered as an option for association rule mining. [3]As with increase of data set and increase of items in the candidate set all the data cannot be kept and managed on a single computer. It also shows that the scalability of Hadoop system with respect to increase in number of candidate sets and input data.

One way to increase the speed of mining is distributing the work into several machines. A MapReduce technique splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The map, maps the job into key and value. The framework sorts the outputs of the maps, which are then input to the reduce tasks. The input of reduce and output of map must have same type. Typically both the input and the output of the job are stored in a file-system. The output from the 'map' is stored in the temporary file in the HDFS, after completion of the all the map reduce task the file is converted into the permanent one. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

MapReduce reflects the pattern for preprocessing Bigdata and extracting only the data we need.

The MapReduce [3] framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job. It is not necessary that the output of map and output of reduce both be of the same type. They can be of different data types.Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Hadoop is designed to run on a large number of machines that don't share any memory or disks. The MapReduce framework sorts the outputs of the maps, which are then given as input to the reduce tasks. Both the input and output of the job are stored in the filesystem[9]. This paper uses a Newman parallel search algorithm based on MapReduce, to improve the processing speed of massive data[10]. The test proves that the parallel algorithm is scalable to large data sets in this paper. It creates clusters of machines and coordinates work among them. Hadoop consists of two key services: reliable data storage using theHadoop Distributed File System (HDFS) and high-performance parallel data processing using a technique called Map/Reduce. It was designed for clusters of commodity, shared-nothing hardware.Even if a machine fails, Hadoop continues to operate the cluster by shifting work to the remaining machines. It automatically creates an additional copy of the data from one of the replicas it manages.

III.CONCLUSIONS

This paper proposes various technique that manages real time big data.We have discussed the challenges that can be made in future for dealing with the big data. This paper also proposes techniques to overcome structured classification. This paper proposes map/reduce technique to handle the real time applications.Hence hadoop technology provides scalabilty and reliability to manage large amount of data in real time applications.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2015

REFERENCES

[1] A. Bifet, G. Holmes, B. Pfahringer, and E. Frank. Fast perceptron decision tree learning from evolving data streams.In PAKDD, 2010.
[2] Sathish Kumar M., Karrunakaran C.M., Vikram M., "Process facilitated enhancement of lipase production from germinated maize oil in Bacillus spp. using various feeding strategies", Australian Journal of Basic and Applied Sciences, ISSN : 1991-8178, 4(10) (2010) pp. 4958-4961.
[3] A. Bifet, G. Holmes, R. Kirkby,and B. Pfahringer." MOA: Massive Online Analysis" Journal of Machine Learning Research (JMLR), 2010.
[4] Kaliyamurthie K.P., Parameswari D., Udayakumar R., "QOS aware privacy preserving location monitoring in wireless sensor network", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S5) (2013) pp.4648-4652.
[5] PallaviRoy ,"A Thesis on Mining Association Rules in Cloud ",August 2012
[6] Sharmila D., Muthusamy P., "Removal of heavy metal from industrial effluent using bio adsorbents (Camellia sinensis)", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384, 5(2) (2013) pp.10-13.
[7] Albert Bifet,"Mining Big Data in Real Time,p1-4.
[8] Udayakumar R., Khanaa V., Saravanan T., Saritha G., "Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction", Middle - East Journal of Scientific Research, ISSN : 1990-9233, 16(12) (2013) pp.1781-1785.
[9] Cloud Computing on Wikipedia,en/ wikipedia.org /wiki/Cloudcomputing.
[10] Kalaiselvi V.S., Prabhu K., Ramesh M., Venkatesan V., "The association of serum osteocalcin with the bone mineral density in post menopausal women",Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 7(5) (2013) pp.814-816.
[11]NIST Definition of Cloud Computing v15,csrc.nist.gov/groups/SNS/cloudomputing/clouddef-v15.doc
[12] Hadoop Distributed File System, hadoop.apache.org/hdfs
[13]HadoopMapReduce, hadoop.apache.org/ mapreduce.
[14]. DR. A. N. Nandakumar, Nandita Yambem ,International Journal of Emerging Technology and Advanced Engineering ,"A Survey on Data Mining Algorithms on Apache Hadoop Platform", , Volume 4, Issue 1, January 2014.
[15]. Xianfeng Yang 1 and Liming Lian ,"A New Data Mining Algorithm based on MapReduce and Hadoop ",International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.7, No.2 (2014), pp.131-142.
[16] Sangeetha Rajagurusamy, Analysis of Work study in An Automobile Company ,International Journal of Innovative Research in Science, Engineering and Technology, ISSN: 2319-8753 , pp 5622-5631,Vol. 2, Issue 10, October 2013.
[17] V.G.Vijaya,Analysis of Rigid Flange Couplings ,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2319-8753 , pp 7118-7126,Vol. 2, Issue 12, December 2013.
[18] V.G.Vijaya ,DESIGN OF HUMAN ASSIST SYSTEM FOR COMMUNICATION ,International Journal of P2P Network Trends and Technology(IJPTT),ISSN: 2319-8753 ,pp 3687-3693,Vol. 2, Issue 8, August 2013.
[19] V.G.Vijaya, V.Prabhakaran ,Design of Human Assist System for Communication ,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2249-2651,pp 30-35, Volume1 Issue3 Number1–Nov2011.
[20] V.Krishnasamy, R.Kalpana devi ,Isomorphous Salts with Abnormal Water Of Hydration ,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753 , pp 3500-3509,Vol. 2, Issue 8, August 2013.
[21] Veera Amudhan R ,Tracking People in Indoor Environments ,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753 , pp 15996-16003,Vol. 3, Issue 9, September 2014.