International Journal of Research in Engineering and Applied Sciences (IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 6 Issue 11, November - 2016, pp. 10~22
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 6.573 | Thomson Reuters ID: L-5236-2015

PREDICTIVE ANALYTICS IN DATA MINING WITH BIG DATA: A LITERATURE SURVEY

V.Vignesh[1],
Research Scholar,
Department of Computer Science,
Karpagam University, Coimbatore- 641021, Tamil Nadu, India .

M. Mohanapriya[2]
Head,
Department of Computer Science and Engineering, Karpagam University, Coimbatore- 641021,

Abstract

Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big data analytics is the process of examining large amounts of data. Big Data is characterized by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. Using MapReduce programming paradigm the big data is processed. While there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. Using MapReduce programming paradigm the big data is processed. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. These technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB. In this research paper various technologies for handling big data along with the advantages and disadvantages of each technology for catering the problems in hand to deal the massive data has discussed.

KeyWords: Big Data, Parameters, Evolution, Hadoop, HDFS

I. INTRODUCTION

A. Definition

With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web sales, social media, audios, images and so on. The need of big data comes from the Big Companies like yahoo, Google, facebook etc for the purpose of analysis of big amount of data which is in unstructured form. Google contains the large amount of information .So; there is the need of Big Data Analytics that is the processing of the complex and massive data sets. Big data analytics analyze the large amount of information used to uncover the hidden patterns and the other information which is useful and important information for the use.

B. Big Data Parameters

As the data is too big from various sources in different form, it is characterized by the 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity. Variety makes the data too big. Data comes from the various sources that can be of structured, unstructured and semi-structured type. Different variety of data include the text, audio, video, log files, sensor data etc.

Volume represent the size of the data how the data is large. The size of the data is represented in terabytes and petabytes. Velocity Define the motion of the data and the analysis of streaming of the data.

C. How big is Big Data and Evolution?

In the last twenty years, the data is increasing day by day across the world .some facts about the data are, there are 277,000 tweets every minute, 2 million searching queries on Google every minute, 72 hours of new videos are uploaded to YouTube, More than 100 million emails are sent, 350 GB of data is processing on facebook and more than 570 websites are created every minute. During 2012, 2.5 quintillion bytes of data were created every day. Big data and its analysis are the center of modern science and business areas. Large amount of data is generated from the various sources either in structure or unstructured form. Such type of data stored in databases and then it become difficult to Extract, transform and load . IBM indicates that 2.5 exabytes data is created everyday which is very difficult to analyze. The estimation about the generated data is that till 2003 it was represented about 5 exabytes, then until 2012 is 2.7 Zettabytes and till 2015 it is expected to increase 3 times.

II. LITERATURE REVIEW

John A. Keane [2] in 2013 proposed a framework in which big data applications can be developed. The framework consist of three stages (multiple data sources, data analysis and modelling, data organization and interpretation) and seven layers (visualisation/presentation layer, service/query/access layer, modelling/ statistical layer, processing layer, system layer, data layer/multi model) to divide big data application into blocks. The main motive of this paper is to manage and architect a massive amount of big data applications. The advantage of this paper is big data handles heterogeneous data and data sources in timely to get high performance and Framework Bridge the gap with business needs and technical realities. The disadvantage of this paper is too difficult to integrate existing data and systems. 2. Xin Luna Dong [5] in 2013 explained challenges of big data integration (schema mapping, record linkage and data fusion). These challenges are explained by using examples and techniques for data integration in addressing the new challenges raised by big data, includes volume and number of sources, velocity, variety and veracity. The advantage of this paper is identifying the data source problems to integrate existing data and systems. The disadvantage of this paper is big data integration such as integrating data from markets, integrating crowd sourcing data, providing an exploration tool for data sources. 3. Jun Wang [17] in 2013 proposed the Data-g Rouping-Aware (DRAW) data placement scheme to improve the problems like performance, efficiency, execution and latency. It could cluster many grouped data into a small number of nodes as compared to map reduce/hadoop. the three main phases of DRAW defined in this paper are: cluster the data-grouping matrix, learning data grouping information from system logs and recognizing the grouping data. The advantage of the paper is improve the throughput up to 59.8%, reduce the execution time up to 41.7% and improve the overall performance by 36.4% over the Hadoop/map reduce. 4. Yaxiong Zhao [7] in 2014 proposed data aware caching (Dache) framework that made minimum change to the original map reduce programming model to increment processing for big data applications using the map reduce model. It is a protocol, data aware cache description scheme and architecture. The advantage of this paper is, it improves the completion time of map reduce jobs. 5. Jian Tan [6] in 2013 author talks about the theoretical assumptions, that improves the performance of Hadoop/map reduce and purposed the optimal reduce task assignment schemes that minimize the fetching cost per job and performs the both simulation and real system deployment with experimental evolution. The advantage of this paper is improves the performance of large scale Hadoop clusters. The disadvantage of this paper is environmental factors such as network topologies effect on a reduce task in map reduce clusters.

The process of the research into complex data basically concerned with the revealing of hidden patterns. Sagiroglu, S.; Sinanc, D. (20-24 May 2015),"Big Data: A Review" describe the big data content, its scope, methods, samples, advantages and challenges of Data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc.By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets. The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data. According to the Intel IT Center, there are many challenges related to Big Data which are data growth, data infrastructure, data variety, data visualization, data velocity. Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ; (17-19 Jan. 2015),"A Big Data implementation based on Grid Computing", Grid Computing offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power. Grid Computing makes the big contributions among the scientific research, help the scientists to analyze and store the large and complex data. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2014) "Shared disk big data analytics with Apache Hadoop" Big data analytics define the analysis of large amount of data to get the useful information and uncover the hidden patterns. Big data analytics refers to the Mapreduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the purpose of implementation of Google's Mapreduce Model. In this the performance of SF-CFS is compared with the HDFS using the SWIM by the facebook job traces .SWIM contains the workloads of thousands of jobs with complex data arrival and computation patterns. Aditya B. Patel, Manashvi Birla, Ushma Nair (*6-8 Dec. 2015)* "Addressing Big Data Problem Using Hadoop and Map Reduce" reports the experimental work on the Big data problems. It describe the optimal solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and Map Reduce programming framework for parallel processing to process large data sets.

Real Time Literature Review about the Big data According to 2013, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads (90 seconds max, 500MB per video), the ability to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily.

III. PROBLEM DEFINITION

Big Data has come up because we are living in society that uses the intensive use of increasing data technology. As there exist large amount of data, the various challenges are faced about the management of such extensive data .The challenges include the unstructured data, real time analytics, fault tolerance, processing and storage of the data and many more. A. *Problem Description* The size of the data is growing day by day with the exponential growth of the enterprises. For the purpose of decision making in an organizations, the need of processing and analyses of large volume of data is increases. The various operations are used for the data processing that includes the culling, tagging, highlighting, searching, indexing etc. Data is generated from the many sources in the form of structured as well as unstructured form [20]. Big data sizes vary from a few dozen terabytes to many petabytes of data. The processing and analysis of large amount of data or producing the valuable information is the challenging task. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main problems regarding big data are the storage capacity and the processing of the data.

IV.TECHNIQUES AND TECHNOLOGY

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data [20]. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. Hadoop

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's Map reduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, Mapreduce, HDFS and numbers of various com ponents like Apache Hive, Base and Zookeeper[17]. MapReduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them . MapReduce have two stages which are [18]:

Map ():- The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node.

Reduce ():- The, Master node collects the answers from all the sub problems and combines them together to form the output.

B. HDFS

HDFS is a block-structured distributed file system that holds the large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of NameNode and many DataNodes.The name node stores the metadata for the NameNode.NameNodes keeps track of the state of the DataNodes. NameNode is also responsible for the file system operations etc [5]. When Name Node fails the Hadoop doesn't support automatic recovery, but the configuration of secondary nod is possible.

HDFS is based on the principle of "Moving Computation is cheaper than Moving Data". Other Components of Hadoop [6]:

HBase: it is open source, Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce.

Pig: Pig is high-level platform where the MapReduce programs are created which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.

Hive: it is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.

Sqoop: Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.

Avro: it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.

Oozie: Oozie is a java based web-application that runs in a java servlet. Oozie uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs.

Chukwa: Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.

Flume: it is high level architecture which focused on streaming of data from multiple sources.

Zookeeper: it is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

C. HPCC

HPCC is a open source computing platform and provide the services for management of big data workflow. HPCC' data model is defined by the user.HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platfprm, a single architecture and a single programming language used for the data processing.HPCC system is based on Enterprise control language that is declarative, on-procedural programming language HPCC system was built to analyze the large volume data for the purpose of solving complex problem. The main components of HPCC are: HPCC data refinery: massively parallel ETL engine.HPCC data delivery: Massively structured query engine. Enterprise Control Language distributes the workload between the nodes

V. FUTURE SCOPE

There is nothing concealed that big data significantly influencing IT companies and through development new technologies only we can handle it in a managerial way. Big data totally change the way of organizations, government and academic institution by using number of tools to make the management of big data. In future Hadoop and NoSQL database will be highly in demand moving forward. The amount of data produced by organizations in next five years will be larger than last 5,000 years. In the upcoming years cloud will play the important role for private sectors and organisations to handle the big data efficiently.

VI. CONCLUSION

We are in the development area of big data. There are various challenges and issues regarding big data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. Big-data analysis fundamentally transforms operational, financial and commercial problems in aviation that were previously unsolvable within economic and human capital constraints using discrete data sets and on-premises hardware. By centralizing data acquisition and consolidation in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods offer new insight into existing data sets.

REFERENCES

[1] Yuri Demchenko "The Big Data Architecture Framework (BDAF)" Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.

[2] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.

[3] Margaret Rouse, April 2010 "unstructured data".

[4] Nguyen T.D., Gondree M.A., Khosalim, J.; Irvine, "Towards a Cross Domain MapReduce Framework" IEEE C.E. Military Communications Conference, MILCOM 2013, 1436 – 1441

[5] Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration" IEEE International Conference on , 29(2013) 1245–1248

[6] Jian Tan; Shicong Meng; Xiaoqiao Meng; Li ZhangINFOCOM, "Improving ReduceTask data locality for sequential MapReduce" 2013 Proceedings IEEE ,1627 - 1635

[7] Yaxiong Zhao; Jie Wu INFOCOM, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" 2013 Proceedings IEEE 2013, 35 - 39 (Volume 19)

[8] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24

[9] Minar, N.; Gray, M.; Roup, O.; Krikorian, R.; Maes, "Hive: distributed agents for networking things" IEEE CONFERENCE PUBLICATIONS 1999 (118-129)

[10] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G,"A Big Data implementation based on Grid Computing", Grid Computing, 2013, 17-19

[11] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012, 18-22

[12] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012, 6-8

[13] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,2010, 72-77.

[14] Chan,K.C.C. Bioinformatics and Biomedicine (BIBM), "Big data analytics for drug discovery" IEEE International Conference on Bioinformatics and Biomedicine 2013,1.

[15] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013. [16] Wang, J.; Xiao, Q.; Yin, J.; Shang, P. Magnetics, "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality"IEEE Transactions (Vol: 49), 2013, 2514 – 2520 [17] HADOOP-3759: Provide ability to run memory intensive jobs without affecting other running tasks on the nodes.

[16]. Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"

[17]. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[18]. Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"

[19]. Yu Li ; Wenming Qiu ; Awada, U. ; Keqiu Li,,(Dec 2012)," Big Data Processing in Cloud Computing Environments"

[20]. Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;,(17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing

[21]. Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review"

[22]. Grosso, P. ; de Laat, C. ; Membrey, P.,(20-24 May 2013)," Addressing big data issues in Scientific Data Infrastructure"

[23]. Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance"

[24]. Szczuka, Marcin,(24-28 June,2013)," How deep data becomes big data"

[25]. Zhu, X. ; Wu, G. ; Ding, W.,(26 June,2013)," Data Mining with Big Data"

[26]. Zhang, Du,(16-18 July,2013)," Inconsistencies in big data"

[27]. Tien, J.M.(17-19 July,2013)," Big Data: Unleashing information"

[28]. Katal, A Wazid, M. ; Goudar, R.H., (Aug,2013)," Big data: Issues, challenges, tools and Good practices"

[29]. Zhang, Xiaoxue Xu, Feng,(2-4 Sep. 2013)," Survey of Research on Big Data Storage"

[30]. http://dashburst.com/infographic/big-data-volume-variety-velocity/

[31].http://www01.ibm.com/software/in/data/bigdata/ [3] W. Gao, Y. Zhu1, Z. Jia, C. Luo, L. Wang, Z. Li, J. Zhan, Y. Qi, Y. He, S. Gong, Xiaona Li,
S. Zhang, and B. Qiu. BigDataBench: a Big Data Benchmark Suite from Web Search Engines. in The Third Workshop on Architectures and Systems for Big Data(ASBD 2013) in
conjunction with The 40th International Symposium on Computer Architecture, May 2013.

[32] Shweta Pandey, Vrinda Tokekar. Prominence of MapReduce in BIG DATA Processing. In Fourth International Conference on Communication Systems and Network Technologies, IEEE, pages 555-560 , 2014