International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 8, August 2013 ISSN 2319 - 4847

Research Issues in Big Data Analytics

Manish Kumar Kakhani[1], Sweeti Kakhani[2] and S.R. Biradar[3]

[1]Assistant Professor, Faculty of Engineering and Technology, MITS University, Lakshmangarh, Rajasthan
[2] Lecturer, Faculty of Arts, Science and Commerce, MITS University, Lakshmangarh, Rajasthan
[3] Professor, Faculty of Engineering and Technology, MITS University, Lakshmangarh, Rajasthan

Abstract

Recently, Big Data has attracted a lot of attention from academia, industry as well as government. It is a very challenging research area. Big Data is term defining collection of large and complex data sets that are difficult to process using conventional data processing tools. Every day, we create trillions of data all over the world. These data is coming from social networking sites, scientific experiments, mobile conversations, sensor networks and various other sources. We require new tools and techniques to organize, manage, store, process and analyze Big Data. This paper systematically presents various research issues related to Big Data analytics.

Keywords: Big Data, MapReduce, Hadoop, Analytics

1. INTRODUCTION

What is Big Data? Many Researchers and organizations have tried to define Big Data in different ways. Gartner defines Big Data are high-volume, high-velocity and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [1].Big Data is defined as the representation of the progress of the human cognitive processes, usually includes data sets with sizes beyond the ability of current technology, method and theory to capture, manage, and process the data within a tolerable elapsed time [2]. According to Wikipedia, Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools [3].
Scientists break down Big Data into many dimensions: Volume, Velocity, Variety, Veracity and Value [4, 5]

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 petabytes.
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.
5. Value-The data we are working with is valuable for society or not.

IBM estimates that every day 2.5 quintillion bytes of data are created ,out of which 90% of the data in the world today has been created in the last two years .This data comes from sensors used to gather climate information, posts to social media sites, digital pictures and videos uploaded on internet, purchase transaction records, and cell phone conversation. All this data is Big Data [6].
The International Data Corporation (IDC) study predicts that the world will generate 50 times the amount of information and 75 times the number of information containers by 2020 while IT personnel to manage it will grow less than 1.5 times. The unstructured information such as files, email and video will constitute 90% of all data created over the next decade [7].
The 2011 digital universe study: extracting values from chaos says that the digital universe is 1.8 trillion gigabytes in size and stored in 500 quadrillion files and its size gets more than double in every two years time frame. If we compare the digital universe with our physical universe then it's nearly as many bits of information in the digital universe as stars in our physical universe [8].
According to Intel, 90% of the data today was created in the last two years, and the growth continues. It is estimated that the amount of data generated until 2012 is 2.7 zettabytes and it is expected to grow 3 times larger than that until 2015 [9].
According to cnet, the initial presidential debate between U.S. president barack obama and former governor mitt romney on october 4, 2012, generated more than 10 million tweets, making it the most tweeted political event in U.S. history [10].
It is observed that social networking sites like Facebook, Twitter and LinkedIn are growing at very fast rate. Facebook has 750 million users, Twitter has 250 million users, LinkedIn has 110 million users [11].These users producing huge amount of data in form of posts, comments and other activities. Google has released new statistics for YouTube, a video

website, usage across the world. It is estimated that 60 hours of video is uploaded every minute on YouTube and over 4 billion YouTube videos are viewed everyday [12]. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 .All these examples show that enormous amount of data is generated everyday on internet by users only. All this data is Big Data and it is exploding at exponential rate.

In order to tackle the Big Data challenges, many governments, organizations and academic institutions come forward to take initiates in this direction. Recently, the US government announced a Big Data Research and Development initiative , to develop and improve the tools and techniques needed to access, organize, and analyze Big Data and to use Big Data for scientific discovery, environmental and biomedical research, education, and national security [13] .Such a federal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management (led by the University of Washington), analytical approaches for genomics based massive data computation (led by Brown University), large scale machine learning techniques for high-dimensional datasets which may be as large as 500,000 dimensions (led by Carnegie Mellon University), social analytics for large-scale scientific literatures (led by Rutgers University), and several others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures which allow us to bring the massive amounts of data down to a human manageable and interpretable scale. Other countries such as the National Natural Science Foundation of China (NSFC) are also catching up with national grants on Big Data research [14].

2. BIG DATA ANALYTICS TOOLS

There are varieties of applications and tools developed by various organizations to process and analyze 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 [15].

MapReduce program consists of two functions –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.

MapReduce has two major advantages: the MapReduce model hide details related to the data storage, distribution, replication, load balancing and so on. 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.

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 [16].

Hadoop is a highly scalable compute and storage platform. But on the other hand, Hadoop is also time consuming and storage-consuming. The storage 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. NameNode: The NameNode is 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 NameNode is 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. DataNode: DataNodes make up the majority of the servers contained in a Hadoop environment. Common Hadoop environments will have more than one DataNode, and oftentimes they will number in the hundreds based on capacity and performance needs. The DataNode 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. EdgeNode: The EdgeNode is the access point for the external applications, tools, and users that need to utilize the Hadoop environment. The EdgeNode 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 EdgeNode and more based on performance needs

iii. IBM InfoSphere BigInsights

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

3. RESEARCH DONE IN BIG DATA ANALYTICS

Big Data analytics is a hot research area today. There are several research papers published to tackle Big Data problems efficiently. Sachchidanand Singh et al. explained the concept, characteristics & need of Big Data and different offerings available in the market to explore unstructured large data [17]. Changqing ji et al. discussed the scope of Big Data processing in cloud computing environment [18]. Dan Garlasu et al. proposed an architecture for managing and processing Big Data using grid technologies [19]. Tyson Condie et al. discussed the machine learning computational models for Big Data [20]. Xindong Wu et al. presented a HACE theorem that characterizes the features of the Big Data revolution and proposed a Big Data processing model from the data mining perspective [21]. Dr. Sun-Yuan Kung proposed cost-effective design on kernel-based machine learning and classification for Big Data learning applications [22]. Kapil Bakshi explored the approaches to analyze unstructured data like imagery, sensors, telemetry, video, documents, log files, and email data files [23].Xiaoyan Gu et al. investigated energy efficient architecture for Big Data application [24].Chansup Byun et al. brought together the Big Data and Big compute by combining Hadoop clusters and MPI clusters [25].

These are few developments in the area of Big Data. Big Data analytics is a new research area and there is lot of scope of research in this area. Preliminary Research has been started but still lot to be done in future.

4. RESEARCH SCOPE IN BIG DATA ANALYTICS

Many researchers have suggested that commercial DBMSs are not suitable for processing extremely large scale data. They are suggesting new Big Data base management system which must be cost effective and scalable. The use of parallelization techniques and algorithms is the key to achieve better scalability and performance for processing Big Data.

Big Data is a new challenge for academia and industry. Researchers are defining new theories, methods and technologies for Big Data management and analysis. Advancing Machine learning, data mining and statistical techniques for processing of Big Data are key to transforming Big Data into actionable knowledge.

Current data base management systems are unable to store the increasing flood of Big Data. There is a need of hierarchical storage architecture to handle the challenge of storing the Big Data. Existing data processing algorithms are excellent at processing homogeneous and static data. But today data is continuously generating from various resources. This data is heterogeneous and dynamic in nature. New scalable data processing algorithms are required to process such data.

While processing a query in Big Data, speed is major criteria .In such case, indexing is a optimal choice for complex query processing. Parallelization and divide and conquer are good algorithmic solutions to handle Big Data effectively.

Organizations are reducing their cost by using online Big Data applications. This strategy is profitable to organizations but producing new security threats. Security of Big Data is prime concern for researchers as well as industry. Security in Big Data is mainly in the form of how to process data mining without exposing sensitive information of users. As Big Data is dynamic in nature hence producing new challenges for researcher to create algorithms to handle such situations.

The main challenges of Big Data are data variety, volume, analytical workload complexity and agility. Many organizations are struggling to deal with the increasing volumes of data. In order to solve this problem, the organizations need to reduce the amount of data being stored and exploit new storage techniques which can further improve performance and storage utilization.

The IT professionals and students looking to build a career and skills in Big Data & Apache Hadoop can take advantage of IBM's BigDataUniversity.com website where users can learn the basics of Hadoop, stream computing and open-source software development [26].

5. CONCLUSION

The paper is a systematic study of various issues of Big Data analytics. Big Data is a very challenging research area. Data is too big to process using conventional tool of data processing. 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 immediate need of new machine learning and data mining techniques to analyze massive amount of data in near future.

ACKNOWLEDGMENT

We are thankful to Mody Institute of Technology and Science, Lakshmangarh, Rajasthan, India for the valuable support in our research.

References

[1] "Big Data: science in the petabyte era," *Nature 455 (7209):1*, 2008
[2] Douglas and Laney, "The importance of 'Big Data': A definition" ,2008
[3] http://en.wikipedia.org/wiki/Big-data
[4] http://dashburst.com/infographic/Big-data-volume-variety-velocity/
[5] http://www.wired.com/insights/2013/05/the-missing-vs-in-Big-data-viability-and-value/
[6] http://www-01.ibm.com/software/in/data/Bigdata/
[7] http://www.kurzweilai.net/worlds-data-will-grow-by-50x-in-next-decade-idc-study-predicts
[8] http://www.emc.com/collateral/demos/microsites/ emc-digital-universe-2011/index.htm
[9] http://www.intel.in/content/www/in/en/Big-data/solving-Big-dataproblems-infographic.html
[10] http://news.cnet.com/8301-1023_3-57525741-93/obama-romney-debate-and-Big-bird-generate-10-million-tweets/
[11] http://www.ebizmba.com/articles/social-networking-websites
[12] http://www.labnol.org/internet/youtube-statistics-2012/20954/
[13] http://www.whitehouse.gov/sites/default/files/microsites/ostp/Big_data_press_release_ final_2.pdf
[14] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,"Data mining with Big Data", IEEE, 2012
[15] Jimmy Lin, Chris Dyer," Data-Intensive Text Processing with MapReduce", Manuscript prepared April 11, 2010.
[16] Tom White," Hadoop: The Definitive Guide", O'Reilly Media, Inc, 2009
[17] Sachchidanand Singh,Nirmala Singh," Big Data Analytics", International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India,2012
[18] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li," Big Data Processing in Cloud Computing Environments", International Symposium on Pervasive Systems, Algorithms and Networks,2012
[19] Dan Garlasu, Virginia Sandulescu,Ionela Halcu,Giorgian Neculoiu,Oana Grigoriu,Mariana Marinescu,Viorel Marinescu," A Big Data implementation based on Grid Computing"
[20] Tyson Condie , Paul Mineiro , Neoklis Polyzotis , Markus Weimer," Machine Learning on Big Data", ICDE Conference, 2013
[21] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,"Data mining with Big Data",
[22] Dr. Sun-Yuan Kung," From Green Computing to Big-Data Learning: A Kernel Learning Perspective"
[23] Kapil Bakshi," Considerations for Big Data: Architecture and Approach",
[24] Xiaoyan Gu,Rui Hou,Ke Zhang,Lixin Zhang,Weiping Wang," Application-driven Energy-efficient Architecture Explorations for Big Data",
[25] Chansup Byun, William Arcand, David Bestor, Bill Bergeron, Matthew Hubbell, Jeremy Kepner, Andrew McCabe,Peter Michaleas, Julie Mullen, David O'Gwynn, Andrew Prout, Albert Reuther, Antonio Rosa, Charles Yee," Driving Big Data With Big Compute",
[26] http://Bigdatauniversity.com/

AUTHOR

Manish Kumar Kakhani received his 5 year Integrated Bachelor of Technology and Master of Technology in Information Technology from Indian Institute of Information Technology and Management, Gwalior, India in 2010.During 2010-2012, he worked in Kalinga Institute of Industrial Technology, Bhubaneswar, India.

Currently, he is working as Assistant Professor and also a research scholar in Mody Institute of Technology and Sciennce, Lakshmangarh, India in computer science and engineering department. His research interest includes Big Data analytics, data mining and machine learning

Sweeti Kakhani received his Bachelor in computer science from Mohan Lal Sukhadia University, Udaipur, India in 2007.She completed her Master in Computer Applications from Rajasthan Technical University, Kota, India in 2010.Currently she is working as Lecturer in Mody Institute of Technology and Sciennce, Lakshmangarh, India in computer science department. His research interest includes Database systems and data mining.

S R Biradar received M.Tech in Computer Science and Engineering from Manipal University, PhD from Jadavpur University. He is currently employed as Professor in the department of Computer Science and Engineering at MITS, Lakshmangarh, India and member of professional societies - ACM, IEEE, ISTE. His current research interest includes wireless networks and image processing. He has published 40+ papers in refereed workshops, conferences, magazines and journals. He has served as a reviewer for various conferences, workshops, and journals.