ISSN:0975-9646

N.Kumar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 4141-4144

A Big Data Mining – A Review

N.Kumar[1], Dr.T.Christopher[2]

[1] *Research Scholar, PG and Research Department of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, India.*
[2] *Assistant Professor, PG and Research Department of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, India.*

***Abstract*- Big Data is a new term used to mention large scale data set that due to their large in size and complexity. We cannot able to processes those datasets with traditional relational database management systems or data mining software. Big data mining is the capability of extracting useful information from those large datasets or real time streams of data, that due to its size and speed. The Big data challenge is becoming one of the most existing opportunities. In this paper, we explore the issue and a broad overview of Big data and its current state.**

***Keywords*- Big Data, Big Data mining, Hadoop, Map Reduce.**

I. INTRODUCTION

In recent years we have witnessed a dramatic increase in our ability to collect data from various sensors, devices in different formats, from independent and connected application. The is data flood has outpaced our capability to process, analysis, store and understand those data sets. Consider the internet data the web page indexed by Google were around one million in 1998, but quickly reaches 1 billion in 2000 and have already exceeded 1 trillion by 2008. This rapid expansion is accelerated by dramatic increase in acceptance of social networking applications such as face book, twitter, viber etc, that allows user to create contents freely and amplify huge data volumes. Furthermore, with Smartphone becoming the sensory gateway to get real time data from people. The huge amount of data that mobile carrier can potentially process to improve our daily life. It can be foreseen that Internet of Things (IoT) applications will raise the scale of data to an unprecedented level. People and device are loosely connected. These connected components generate huge amount of data and becomes necessary to extract valuable information can be discovered from those data to improve our life and makes our world better place to live. The Big data applications are facing the following challenges are (i) System Capacity (ii) Design of Algorithm and (iii) Process model. We further introduce Big data in section II, Big data mining and its applications in section –III, Big data processing platform in section IV , some of the related research work on Big data section V, we discuss importance of Open source in section VI, and finally conclusion in section VII.

II. BIG DATA

Big data is the new term used to identify the datasets that due to their large in size, we cannot manage them with using typical data mining software tools. Big data is currently defined using three data characteristics such as volume, variety and velocity.[1] It is meant that, when the volume , velocity of the data are increased, the existing techniques and technologies may not able to store or process those datas. At that point that type of data is defined as Big data.

According to IDC 'Digital Universe study' the world information is doubling every two years and is predicted to reach 40 ZB by 2020[2]. The increase in data also referred as "data tsuanami" is driven by social media and mobile networked devices (IoT), finance and online retail transactions as well as advances in the physical and life sciences. As evidence the social media micro blogging site twitter processing approximately 12 TB of data per day, while face book receives five hindered likes per day[3].

Big data characterized not only by its volume, but rich mix of data types and formats, (variety) and its time sensitive nature which marks a deviation from traditional batch processing(Velocity). These characteristics are commonly refereed as 3Vs. Traditional distributed systems and database are no longer suitable to effectively capture , store , manage analysis this data and exhibit limited scalability.

A. Characteristics of Big data

1) Volume: The amount of data. Data volume continuous to increase on unprecedented rate.

2) Velocity: Data in motion. The Speed at which the data is created, processed and analyzed continuous to accelerate.

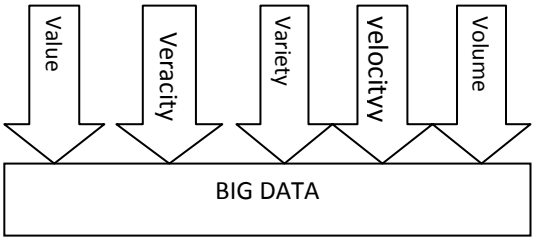

Figure 1: The Big Data Characteristics

3) Variety: Different types of data and data sources included structured, semi structured and un structured data.

4) Veracity : Data Uncertainty. Veracity refers to the level of reliability associated with certain type of data.

5) Value: Business value. The organization benefits from those data sets.

Gartner[4] summarizes this in their definition of Big data in 2012 as high volume, velocity and variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

III. BIG DATA MINING AND ITS APPLICATIONS.

The data produced nowadays is estimated in the order of zetta bytes and it is growing around 40 % every year. With the above facts, we define Big Data mining is a processing of extracted information from very large data sets and data streams. We need a new algorithms and new tools to deal with all of these large data sets. Some of the few Big data applications are as follows.

A. Health

Mining DNA of each person to discover, monitor and improve health.

B. Smart cites

Cities focussed on sustainable economic development and high quality of life, with wise management and natural resources.

C. Global Pulse[5]:

Is an United National initiative launched in 2009, that functions as an innovative lab and that is based on mining Big data for developing countries in their white paper ' Big Data for development , Challenges and Opportunities"[6].

IV. BIG DATA PROCESSING PLATFORMS.

The big data platforms can be categorized into the following two types of scaling:

Horizontal Scaling: Horizontal scaling involves distributing the workload across many servers which may be even commodity machines. It is also known as "scale out", where multiple independent machines are added together in order to improve the processing capability.

Vertical Scaling: Vertical Scaling involves installing more processors, more memory and faster hardware, typically, within a single server. It is also known as "scale up" and it usually involves a single instance of an operating system.

TABLE 1
SUMMARY OF BIG DATA MINING PLATFORMS

S.No	Scaling Type	Platforms
1	Horizontal	Peer to Peer
2		Hadoop (HDFS)
3		Map Reduce
4	Vertical	High Performance Computing (HPC)
5		Multicore CPU
6		Graphics Processing Unit (GPU)

A Horizontal scaling platforms

Some of the prominent horizontal scale out platforms includes peer-to-peer networks and Apache Hadoop. We will now discuss each of these platforms in more detail in this section.

1) Peer-to-peer networks[7] involve millions of machines connected in a network. It is a decentralized and distributed network architecture where the nodes in the networks (known as peers) serve as well as consume resources. It is one of the oldest distributed computing platforms in existence. The major bottleneck in such a setup arises in the communication between different nodes. Broadcasting messages in a peer-to-peer network is cheaper but the aggregation of data/results is much more expensive. In addition, the messages are sent over the network in the form of a spanning tree with an arbitrary node as the root where the broadcasting is initiated.

2) Apache hadoop[8] is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant. The Hadoop platform contains the following two important components:

3) Distributed File System (HDFS) [9] is a distributed file system that is used to store data across cluster of commodity machines while providing high availability and fault tolerance.

4) Hadoop YARN [10] is a resource management layer and schedules the jobs across the cluster.

5) MapReduce [11] The programming model used in Hadoop is MapReduce which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. At a high-level, mappers read the data from HDFS, process it and generate some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS.

B. Vertical scaling platforms

The most popular vertical scale up paradigms are High Performance Computing Clusters (HPC), Multicore processors, and Graphics Processing Unit (GPU). We describe each of these platforms and their capabilities in the following sections.

1) High performance computing (HPC) clusters [12], also called as blades or supercomputers, are machines with thousands of cores. They can have a different variety of disk organization, cache, communication mechanism etc. depending upon the user requirement. These systems use well built powerful hardware which is optimized for speed and throughput.

2)Multicore CPU: Multicore refers to one machine having dozens of processing cores [13]. They usually have shared memory but only one disk. More recently, the number of cores per chip and the number of operations that a core can perform has increased significantly.

3) Graphics processing unit (GPU) is a specializd hardware designed to accelerate the creation of images in a frame buffer intended for display output [14]. Until the past few years, GPUs were primarily used for graphical operations such as video and image editing, accelerating graphics-related processing etc

V. RELEATED RESEARCH WORK

We selected few contributions that together shows very significant state-of-the-art research in Big Data Mining, and that provides a broad overview of the field and its forecast to the future.

Scaling Big Data Mining Infrastructure: The Twitter Experience by Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.). This paper presents insights about Big Data mining infrastructures, and the experience of doing

analytics at Twitter. It shows that due to the current state of the data mining tools, it is not straightforward to perform analytics. Most of the time is consumed in preparatory work to the application of data mining methods, and turning preliminary models into robust solutions.

Mining Heterogeneous Information Networks**:** A Structural Analysis Approach by Yizhou Sun (North- eastern University) and Jiawei Han (University of Illinois at Urbana-Champaign). This paper shows that mining heterogeneous information networks is a new and promising research frontier in Big Data mining research. It considers interconnected, multi-typed data, including the typical relational database data, as heterogeneous information networks. These semi-structured heterogeneous information network models leverage the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data.

Mining Large Streams of User Data for Personalized Recommendations by Xavier Amatriain (Netix). SIGKDD Explorations Volume 14, Issue 2 Page 2. This paper presents some lessons learned with the Netix Prize, and discuss the recommender and personalization techniques used in Netix. It discusses recent important problems and future research directions. Section 4 contains an interesting discussion about if we need more data or better models to improve our learning methodology.

VI. Open Source Tools For Big Data Mining

The Big Data phenomenon is intrinsically related to the open source software. Large companies as Facebook, Yahoo!, Twitter, and LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

A.Apache Pig [15]: software for analyzing large data sets that consists of a high-level language similar to SQL for expressing data analysis programs, coupled with infrastructure for evaluating these programs. It contains a compiler that produces sequences of Map Reduce programs.

B.Cascading [16]: software abstraction layer for Hadoop, intended to hide the underlying complexity of MapReduce jobs. Cascading allows users to create and execute data processing workflows on Hadoop clusters using any JVM-based language.

C.Scribe [17]: server software developed by Facebook and released in 2008. It is intended for aggregating log data streamed in real time from a large number of servers.

D.Apache Mahout [18]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.

C.R [19]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.

D.Vowpal Wabbit [20]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VWis able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning.

E.PEGASUS [21]: big graph mining system built on top of MAPREDUCE. It allows to find patterns and\ anomalies in massive real-world graphs.

F.GraphLab [22]: high-level graph-parallel system built without using MAPREDUCE. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighbouring vertices.

VII. Conclusion

The Big data and Big data mining is going to continue during the nest years, and the role of data mining researches will have to manage more amount of data. This data is going to be more diverse, larger, and faster. We discussed some insights about Big data , Big data mining We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before.

References

[1] Shan Suthaharan, Big data classification: problems and challenges in network instruction.

[2] Digital Universe study (on behalf of EMC Corporation)(2012) Big data, Bigger Digital Shadows. And biggest growth, http://idcdocserve.com/1414/

[3] Mckinsey Global Institute (2011) Big data: the next frontier for innovation, competition and productivity. In http://www.mckinsey.com /frontier_for_innovation.

[4] Garnter,http:// www.gartner.com/it_glossary /Big _Data

[5] UN Global Pluse, http://www.unglobleplus e.com

[6] E.Letoate, Big Data for development,Oppurtunties & Challenges, may 2011.

[7] Steinmetz R, Wehrle K (2005) Peer-to-Peer Systems and Applications. Springer Berlin,Heidelberg

[8] Hadoop. http://hadoop.apache.org/

[9] Borthakur D (2008) HDFS architecture guide. HADOOP APACHE PROJECT. http://hadoop.apache.org/docs/r1.2.1/ hdfs_design.pdf

[10] Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S (2013) Apache hadoop yarn: Yet another resource negotiator. In: Proceedings of the 4th annual Symposium on Cloud Computing., p 5 Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113

[11] Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B (2012) Parallel data processing with MapReduce: a survey. ACM SIGMOD Record 40(4):11–20

[12] Buyya R (1999) High Performance Cluster Computing: Architectures and Systems (Volume 1). Prentice Hall, Upper SaddleRiver, NJ, USA

[13] Bekkerman R, Bilenko M, Langford J (2012) Scaling up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press.
[14] Owens JD, Houston M, Luebke D, Green S, Stone JE, Phillips JC (2008) GPU computing. Proc IEEE 96(5):879–899.
[15] Apache Pig, http://www.pig.apache.org/.
[16] Cascading, http://www.cascading.org/.
[17] Facebook Scribe, https://github.com/ facebook/scribe.
[18] Apache Mahout, http://mahout.apache.org.
[19] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.
[20] J. Langford. Vowpal Wabbit, http://hunch.net/˜vw/,2011.
[21] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.
[22] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2010