**Review Article** **Impact Factor: 0.621** **ISSN: 2319-507X**
**Pruthvika Kadu, IJPRET, 2014; Volume 2 (9): 1011-1017** **IJPRET**


**INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY**

A PATH FOR HORIZING YOUR INNOVATIVE WORK

# A REVIEW ON BIG DATA MANAGEMENT AND ITS SECURITY

**PRUTHVIKA S. KADU[1], DR. H. R. DESHMUKH[2], PROF P.G. ANGAITKAR[3], S. A. KARALE[3]**

1. M. E. First Year, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.
2. Prof and HOD, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.
3. Asst. Prof, Department of Computer Science & Engineering, IBSS College of Engineering, Amravati.

**Accepted Date: 27/02/2014 ; Published Date: 01/05/2014**

**Abstract:** Big Data is a term defining data that has three main characteristics. First, it involves a great volume of data. Second, the data cannot be structured into regular database tables and third, the data is produced with great velocity and must be captured and processed rapidly. Oracle adds a fourth characteristic for this kind of data and that is low value density, meaning that sometimes there is a very big volume of data to process before finding valuable needed information. Big Data is a relatively new term that came from the need of big companies like Yahoo, Google, Facebook to analyze big amounts of unstructured data, but this need could be identified in a number of other big enterprises as well in the research and development field. This paper reports the review on big data problem, its optimal solution using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and using parallel processing to process large data sets using Map Reduce programming framework and Big data management.

**Keywords:** Big Data, Big data problem, Hadoop, MapReduce, HDFS, Big data Security

***PAPER-QR CODE***

**Corresponding Author: MS. PRUTHVIKA S. KADU**

**Access Online On:**

www.ijpret.com

**How to Cite This Article:**

Pruthvika Kadu, IJPRET, 2014; Volume 2 (9): 1011-1017

## INTRODUCTION

### What Big data actually is?

Nowadays, we live in a more and more interconnected world that generates a great volume of information every day, starting from the logging files of the users of social networks, search engines, e-mail clients to machine generated data as from the real-time monitoring of sensor networks for dams or bridges, and various vehicles such as airplanes, cars or ships. According to an infographic made by Intel, 90% of the data today was created in the last two years, and the growth continues. It is estimated that all the global data generated from the beginning of time until 2003 represented about 5exa Bytes (1 exaByte equals 1 million giga Bytes), the amount of data generated until 2012 is 2.7 zettaBytes (1 zettaBytes equals 1000 exaBytes) and it is expected to grow 3 times larger than that until 2015[1]. For example, the number of RFID tags sold globally is projected to rise from 12 million in 2012 to 209 billion in 2021[1]. All this volume represents a great amount of data that rise challenges when talking about acquiring, organizing and analyzing it. Big Data is an umbrella term describing all these types of information mentioned above.

Michael Cox and David[2] Ellsworth were among the first to use the term big data literally, referring to using larger volumes of scientific data for visualization (the term large data also has been used). Currently, there are a number of definitions of big data. Perhaps the most well known version comes from IBM[3], which suggested that big data could be characterized by any or all of three “V” words to investigate situations, events, and so on: volume, variety, and velocity.

Volume refers to larger amounts of data being generated from a range of sources. For example, big data can include data gathered from the Internet of Things (IoT). As originally conceived[4],IoT referred to the data gathered from a range of devices and sensors networked together, over the Internet.

Variety refers to using multiple kinds of data to analyze a situation or event. On the IoT, millions of devices generating a constant flow of data results in not only a large volume of data but different types of data characteristic of different situations.

Velocity of data also is increasing rapidly over time for both structured and unstructured data, and there’s a need for more frequent decision making about that data.

## LITERATURE REVIEW

### What is Big data problem?

Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. One current feature of big data is the difficulty working with it using relational databases and desktop statistics/visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers"[5]. The various challenges faced in large data management include – scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. In addition to variations in the amount of data stored in different sectors, the types of data generated and stored—i.e., whether the data encodes video, images, audio, or text/numeric information—also differ markedly from industry to industry[6].

### How do we manage Big data?

This type of data is impossible to handle using traditional relational database management systems. New innovative technologies were needed and Google found the solution by using a processing model called MapReduce. There are more solutions to handle Big Data, but the most widely-used one is Hadoop, an open source project based on Google's MapReduce and Google File System. Hadoop was founded by the Apache Software Foundation. The main contributors of the project are Yahoo, Facebook, Citrix, Google, Microsoft, IBM, HP, Cloudera and many others. Hadoop is a distributed batch processing infrastructure which consists of the Hadoop kernel, Hadoop Distributed File System (HDFS), MapReduce and several related projects.

### Hadoop

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It enables applications to work with thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS) [7].

### HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop provides a distributed filesystem (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS

has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster. HDFS is based on the principle that Moving Computation is Cheaper than Moving Data", meaning that it is easier to move the computation where that data to be processed is, rather than moving the data to where the computation is running, this being true especially when the I/O files have a big size. HDFS offers great portability, being written in Java and designed to run on commodity hardware, usually on machines that run a GNU/Linux operating system.

**HDFS Architecture**

As show in Figure 1, an HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manages storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNode. NameNode determines the mapping of blocks to Datanodes. HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks[9].

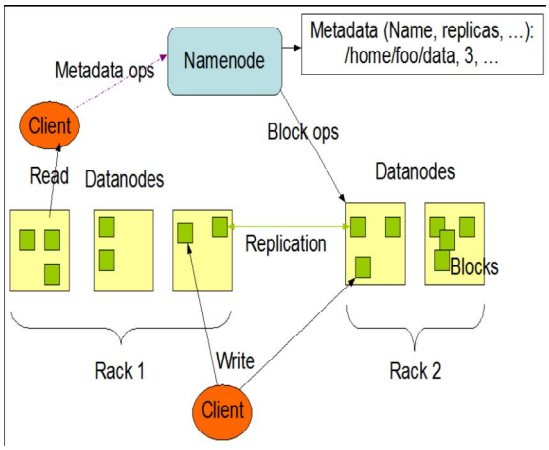


Figure 1. HDFS Architecture

**MapReduce Programming Framework**

MapReduce is a software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers. MapReduce is a programming model for processing and generating large data sets. Users specify a map function that processes a

key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key[8].

**"Map" step:** The master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain:

Map (k1, v1) → list (K2, v2)

**"Reduce" step:** The master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain[9]:

Reduce (K2, list (v2)) → list (v3)

**Security Issues with Big Data**

One of the key security issues involved with big data aggregation and analysis is that organizations collect and process a great deal of sensitive information regarding customers and employees, as well as intellectual property, trade secrets and financial information. As organisations look to gain value from such information, they are increasingly seeking to aggregate data from a wider range of stores and applications to provide more context in order to increase the value of the data – for example, to provide a clearer picture of customer preferences in order to better target them.

By centralizing data in one place, it becomes a valuable target for attackers, which can potentially leave huge swathes of information exposed, which could undermine trust in the organization and damage its reputation. This makes it essential that big data stores are properly controlled and protected.

Another potential problem relates to regulatory compliance, especially with data protection laws. Such laws are more stringent in some jurisdictions than others, particularly with regard to where data can be stored or processed. Organizations need to carefully consider the legal ramifications of where they store and process data to ensure that they remain in compliance with the regulations that they face. However, there are also security advantages to big data projects. When centralizing data stores, organizations should first classify the information and apply appropriate controls to it, such as imposing retention periods as specified by the regulations that they face. This will allow organizations to weed out data that has little value or

that no longer needs to be kept so that it can be disposed of and is no longer available for theft or subject to litigation demanding presentation of records. Another security advantage is that large swathes of data can be mined for security events, such as malware, spear phishing attempts or fraud, such as account takeovers[10].

**Big Data Security Controls**

Research firm Forrester recommends that in order to provide better control over big data sets, controls should be moved so that they are closer to the data store and the data itself, rather than being placed at the edge of the network, in order to provide a more effective line of defence. It also states that separate silos of data control and protection – such as archiving, data leakage prevention and access controls – should be brought together. In terms of access controls, they should be granular enough to ensure that only those authorized to access data can do so, in order to prevent sensitive information from being compromised. Controls should also be set using the principle of least privilege, especially for those with greater access rights, such as administrators. Products such as Vormetric bring together data encryption and its related policy management and key storage elements and link access control to the data. Therefore companies can decide who can view the data or in the case of an administrator allow them physical access: but should they try to read the data it would be useless because the process would not have allowed decryption. Such an approach is highly effective in any multi-silo environment where any form of electronic data is stored. To ensure that access controls are effective, they should be continuously monitored and should be modified as employees change role in the organization so that they do not accumulate. This can be done using existing technologies excessive rights and privileges that could be abused. in use in many organizations such as database activity monitoring tools, the capabilities of which are being expanded by many vendors to deal with unstructured data in big data environments. Other useful tools include Security Information and Event

Management (SIEM) technologies, which gather log information from a wide variety of applications on the network.

**CONCLUSION**

As data volumes continue to expand, as they take in an ever wider range of data sources, much of which is in unstructured form, organizations are increasingly looking to extract value from that data to uncover the opportunities for the business that it contains. However, traditional data storage and analysis tools are not, on their own, up to the task of processing and analyzing the information the data contains, owing not just to the volume of data, but also to the unstructured, ad hoc nature of much of the content. In addition, the centralized nature of big

data stores creates new security challenges to which organizations must respond, which require that controls are placed around the data itself, rather than the applications and systems that store the data.

**REFERENCES**

1. Big Data Infographic: Solve your Big Data Problems? http://www.intel.in/content/www/in/en/big-data/solving-big-dataproblems-infographic.html

2. M. Cox and D. Ellsworth, Managing Big Data for Scientific isualization," Proc. ACM Siggraph, ACM, 1997, pp. 5-1–5-17.

3. K. Ashton, "That 'Internet of Things' Thing," RFID J., 22 June 2009; www.rfidjournal.com/article/view/4986.

4. P. Zikopoulous et al., Harness the Power of Big Data, McGraw-Hill, 2013..

5. Thomas Herzog, Associate Commissioner, New YorkState, Thomas Kooy, IJIS Institute Big Data and the Cloud, IJIS Institute Emerging Technologies, Available:http://www.correctionstech.org/meeting/2012/Presentatio ns/Red_01.pdf, Aug, 2012.

6. Jacobs, A., The Pathologies of Big Data, ACM Queue, Available: http://queue.acm.org/detail.cfm?id=1563874, 6th July 2009.

7. Apache Software Foundation. Official apache hadoop website, http://hadoop.apache.org/, Aug, 2012.

8. Hung-Chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D.Stott Parker from Yahoo and UCLA, "Map-Reduce- Merge: Simplified Data Processing on Large Clusters", paper published in Proc. of ACM SIGMOD, pp. 1029– 1040, 2007.

9. The Hadoop Architecture and Design, Available: http://hadoop.apache.org/common/docs/r0.16.4/hdfs_design.html, Aug, 2012.

10. Big data and infosecurity'. Varonis, 2012. Accessed June 2012. http:// blog.varonis.com/big-datasecurity/.