**"The New World of Data- Big Data Using Framework Hadoop"**

**Rutuja Nikam**

**Master of Engineering Ist Year**

**(Computer Science & Engineering,Darapur)**

**The New World of data - Big data**

**Using Framework Hadoop**

**Rutuja Nikam**

**Master of Engineering Ist Year**

**(KGIT'S Computer Science & Engineering,Darapur)**

**Guide: Prof. Meghali Kalyankar**

**ABSTRACT**

Technology is evolutionary field, what seems new today became old by time. Now a day's concept of big data is in news, on the page of news paper, and it is a topic of research and enthusiasm in world of data. The term Big Data is new but technologies incorporate into it are old like high-speed networks, high-performance computing, task management, thread management, and data mining. People always have attraction and enthusiasm whenever new technologies come in market. If today's organizations do not adopt new technologies then they will be left far behind in their market position. But it would not be wise if we are blindly adopting new technologies without knowing its concept and values.

The term Big Data is introduced in data world to process, mange and support massive amount of data. Many organizations are using big data to handle their large amount of data chunks and to gain some meaningful result set from it.Big Data is not just about lots of data, it is actually a concept providing an opportunity to find new insight into your existing data as well guidelines to capture and analysis your future data. It makes any business more agile and robust so it can adapt and overcome business challenges.

Hadoop is the core platform for structuring Big Data, and solves the problem of formatting it for subsequent analytics purposes. Hadoop uses a distributed computing architecture consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale and support extremely large data stores.

**KEY WORDS**

Big data, Robust, Transformation, Hadoop, Analytics, High speed Network

## 1. Introduction

In Technical word Big data is data which is beyond the capacity of conventional databases. The data is too vast, fast, or doesn't compatible the structures of your database architectures. If you want to gain value from this data, you have to choose some alternative for the same.

Big data includes 3 V's which are volume, velocity and variability of large data. In this massive amount of some valuable data pattern and information lie in the hidden form because of the amount of work and time required to extract them. To leading organizations, like Google, Amazon or Face book, this power has been in reach for some time, but at amazing cost. Processing of Big data is at ease because availability of commodity hardware, cloud architectures open source. It is also feasible for small scale organization to process big data as server time in the cloud can be cheaply rented.

Organization uses value of Big data into two categories: analytical use, and enabling new products. Big data analytics can be useful to find out hidden patterns in data such as peer influence among customers, revealed by analyzing shoppers' transactions, social and geographical data, hidden previously because of process cost and time. Every single item of data being able to process in reasonable time removes the trouble need for sampling and promotes an investigative approach to data.

The past decade's successful web start-ups are prime examples of big data used as an enabler of new products and services. For example, by combining a large number of signals from a user's actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It's no coincidence that the lion's share of ideas and tools underpinning big data has emerged from Google, Yahoo, Amazon and Facebook.

The emergence of big data into the enterprise brings with it a necessary counterpart: agility. Successfully exploiting the value in big data requires experimentation and exploration. Whether creating new products or looking for ways to gain competitive advantage, the job calls for curiosity and an entrepreneurial outlook.

## 2. What is BIG DATA

BIG Data the name justify itself, a big amount of data. Big data is a concept which is used to store, manage and process large amount of data from different types of sources. In traditional systems we can only store structured data but in big data we can store both structured and non-structured data such as videos, pictures, emails, customer transactional histories, production databases Big data processes information which is supports decision making .

Concept of Big Data includes three V's: Volume, velocity and variety :

Volume – Amount of data

Velocity – Speed of processing

Variety – Source and Types of data

**2.1 Volume**

As we can currently see the exponential growth in the data storage as the data is now more than text data. Data may be in the format of videos, musics and photos, emails on our social media channels. It is very common to have Terabytes and Pet bytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents Big Data.

**2.2 Velocity**

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data.

**2.3 Variety**

Data can be stored in multiple format. For example database, excel, CSV, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, PDF or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the Big Data. This variety of the data represents Big Data.

**3. Uniqueness of Big Data**

Companies have sought for decades to make the best use of information to improve their business capabilities. However, it's the structure (or lack thereof) and size of Big Data that makes it so unique. Big Data is also special because it represents both significant information - which can open new doors - and the way this information is analyzed to help open those doors. The analysis goes hand-in-hand with the information, so in this sense "Big Data" represents a noun - "the data" - and a verb - "combing the data to find value."

The days of keeping company data in Microsoft Office documents on carefully organized file shares are behind us, much like the bygone era of sailing across the ocean in tiny ships. That 50 gigabyte file share in 2002 looks quite tiny compared to a modern-day 50 terabyte marketing database containing customer preferences and habits. How can we possibly comb through all that material to spot trends suggesting which way consumer tastes are headed or what climate changes are occurring? That's where the interpretive process comes in.

**4. Big Data and Analytics in Financial Services**

Financial organization moving towards Big data. But it seems like a long journey. Much organization sees this as organization transformation and great opportunity in the field of data. Such organizations will implement more and bigger data projects in future also to retrieve business potential, vital values and talent from it.

To make applications work in context of big data and to gain full benefits from it originations should follow below guidelines:

**Big data is not solely a technology issue.**

Financial executives must recognize that big data is more about how to use the growing velocity, variety, and volume of information to make material change and improvements in the way financial enterprises interact with customers, partners, regulators, and employees; manage risk and run efficient operations.

**Big data is a journey.**

The big data journey is unique to each organization, dependent on many factors, including the maturity of an organization's current data infrastructure, the type of business, and the availability of skills within the organization.

**Big data is an incremental process.**

Big data deployments need to start small and build out incrementally. Programs should be incorporated within business-as-usual activities and be part of an overall data management and analysis road map.

**Big data competencies need to be built within an analytics strategy.**

Analytics excellence is core to innovation across the financial industry. Business executives in the financial industry must view analytics as an important capability that will ultimately distinguish those that thrive in uncertain and uneven markets from those that fumble.

**5. The Apache Hadoop framework**

**5.1 Apache Hadoop**

Big data is concept, in market different kind of tools are available to handle big data. Apache hadoop is used to deal with big data. It is an open source software framework for storage, managing and processing of big data. Apache's top level project being built and used by global technical community and it comes under the Apache License 2.0

**5.2 Modules of Apache Hadoop framework**

- Hadoop Common: contains libraries and utilities needed by other Hadoop modules
- Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster

- Hadoop YARN: a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications

- Hadoop MapReduce: a programming model for large scale data processing

- All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers.

- Beyond HDFS, YARN and MapReduce, the entire Apache Hadoop "platform" is now commonly considered to consist of a number of related projects as well: Apache Pig, Apache Hive, Apache HBase, and others.

**5.3 High level Architecture of Hadoop**

5.3.1 HDFS and MapReduce

There are two primary components at the core of Apache Hadoop 1.x: the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework. These are both open source projects, inspired by technologies created inside Google.

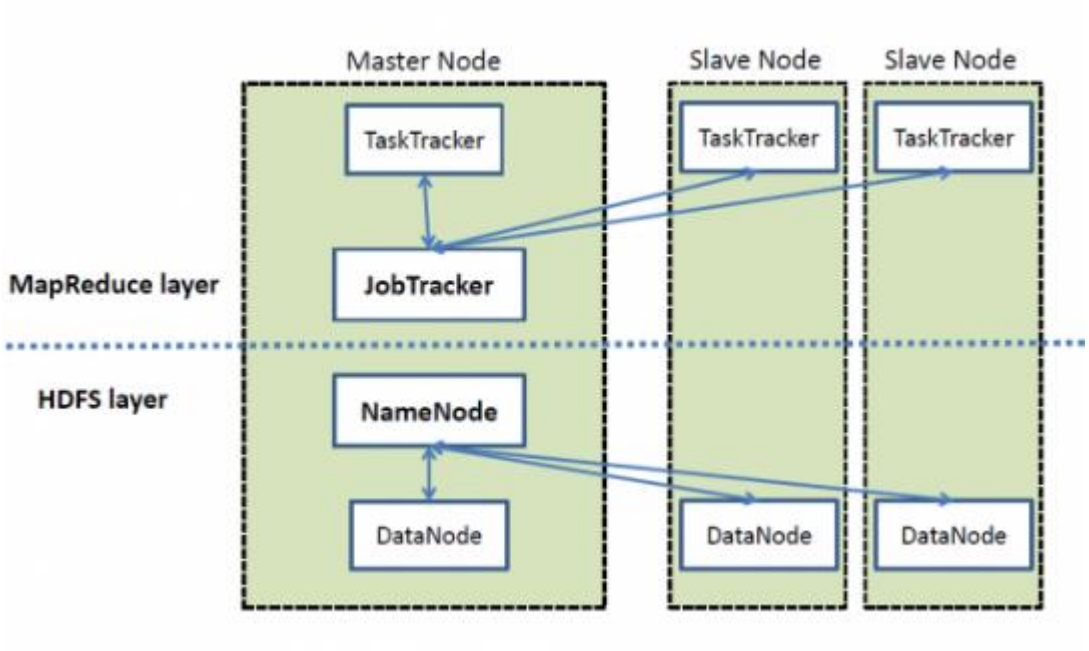


**Figure 1: High level architecture diagram of Hadoop**

5.3.2 Hadoop distributed file system

The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single name node, and a cluster of data nodes form the HDFS cluster. The situation is typical because each node does not require a data node to be present. Each data node serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other.

5.3.3 Apache Hadoop NextGen MapReduce (YARN)

Apache Hadoop YARN is a sub-project of Hadoop at the Apache Software Foundation introduced in Hadoop 2.0 that separates the resource management and processing components. YARN was born of a need to enable a broader array of interaction patterns for data stored in HDFS beyond MapReduce. The YARN-based architecture of Hadoop 2.0 provides a more general processing platform that is not constrained to MapReduce.

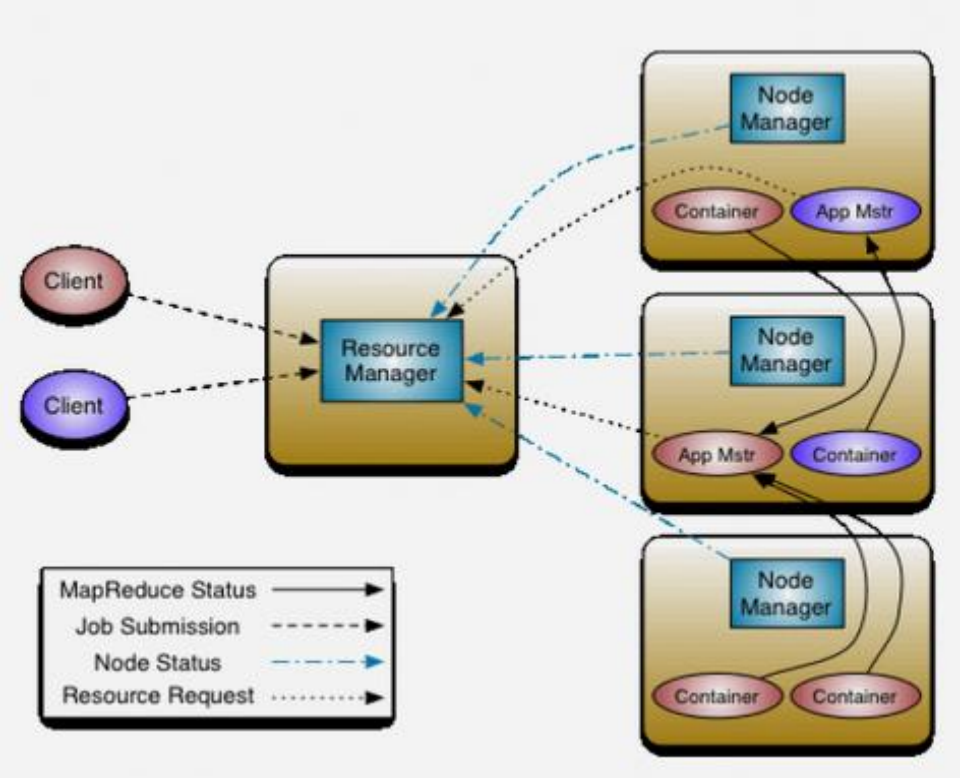


**Figure 2: Hadoop MapReduce(YARN)**

5.3.3.1 What YARN does

YARN enhances the power of a Hadoop compute cluster in the following ways:

1. Scalability: The processing power in data centers continues to grow quickly. Because YARN ResourceManager focuses exclusively on scheduling, it can manage those larger clusters much more easily.

2. Compatibility with MapReduce: Existing MapReduce applications and users can run on top of YARN without disruption to their existing processes

3. Agility: With MapReduce becoming a user-land library, it can evolve independently of the underlying resource manager layer and in a much more agile manner.

**6. Hadoop is Important**

Now a day's hadoop is a hot top in technical world. There are many reasons behind its success such as it handling big amount of data very quiqly.It is most vital point in world of data.

Other reasons include:

1. Low cost. The open-source framework is free and uses commodity hardware to store large quantities of data.
2. Computing power. Its distributed computing model can quickly process very large volumes of data. The more computing nodes you use the more processing power you have.
3. Scalability. You can easily grow your system simply by adding more nodes. Little administration is required.
4. Storage flexibility. Unlike traditional relational databases, you don't have to preprocess data before storing it. And that includes unstructured data like text, images and videos. You can store as much data as you want and decide how to use it later.
5. Inherent data protection and self-healing capabilities. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. And it automatically stores multiple copies of all data.

**7. Commercial use of Hadoop**

- It was claimed that Yahoo! Inc. was world's biggest Hadoop production application. Yahoo's Web map is hadoop application which runs on a Linux cluster and it generates data that is useful for yahoo search query. All the work done by Yahoo is contributed to the open-source community.

- Later on as the users and data increases, facebook claimed that it had the largest cluster on hadoop. They declared that data storage in the warehouse increases per day as half pet bytes.

- There are many more an organization in which hadoop is deployed in traditional sites as well as in clouds. Hadoop is easy to deploy even on a system not having hardware to acquire or any specific setup on it.

**8. Conclusion**

In simple term Big data is the aspiration for a data world to build platforms and tools to store, process and analyze data to retrieve important data value set to gain benefits in world of data transformation. When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

Big data and Hadoop is playing very crucial role in process of transforming organizational data architecture. It's a gold-rush market with pure-plays, enterprise software vendors and cloud vendors are all competing to stake a claim. The open source Apache Hadoop project includes the core modules — Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN and Hadoop MapReduce — but without the support or packaged solutions of a commercial vendor. All of the leading commercial distributions are compatible with Apache Hadoop.

**9. Acknowledgements**

This research paper provides description of the logical architecture and the layers of a big data solution. It also specifies the organizational guidelines for the use of big data and hadoop and how to retrieve important business solution from it.

**10. References:**

1. Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7: 1–5.

2. Hellerstein, Joe (9 November 2008)."Parallel Programming in the Age of Big Data". Gigaom Blog.

3. "FICO® Falcon® Fraud Manager". Fico.com. Retrieved 2013-07-21.

4. "eBay Study: How to Build Trust and Improve the Shopping Experience". Knowwpcarey.com. 2012-05-08. Retrieved2013-03-05.

5. Leading Priorities for Big Data for Business and IT. eMarketer. October 2013. Retrieved January 2014.

6. Wingfield, Nick (2013-03-12). "Predicting Commutes More Accurately for Would-Be Home Buyers - NYTimes.com". Bits.blogs.nytimes.com. Retrieved

7. Laney, Douglas. "The Importance of 'Big Data': A Definition". Gartner. Retrieved21 June 2012.

8. "What is Big Data?". Villanova University.

9. http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4

10. Billings S.A. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains". Wiley, 2013

11. Delort P., Big data Paris 2013http://www.andsi.fr/tag/dsi-big-data/

12. "Data, data everywhere". The Economist. 25 February 2010. Retrieved9 December 2012.

13. "Community cleverness required".Nature 455 (7209): 1. 4 September 2008.doi:10.1038/455001a.

14. "Sandia sees data management challenges spiral". HPC Projects. 4 August 2009.

15. Reichman, O.J.; Jones, M.B.; Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology".Science 331 (6018): 703–5.doi:10.1126/science.1197962.PMID 21311007.

16. "Data Crush by Christopher Surdak". Retrieved 14 February 2014.

17. Hellerstein, Joe (9 November 2008)."Parallel Programming in the Age of Big Data". Gigaom Blog.

18. Bamford, James (15 March 2012). "The NSA Is Building the Country's Biggest Spy Center (Watch What You Say)". Wired Magazine. Retrieved 2013-03-18.

19. "Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center". National Security Agency Central Security Service. Retrieved 2013-03-18.

20. Hill, Kashmir. "TBlueprints Of NSA's Ridiculously Expensive Data Center In Utah Suggest It Holds Less Info Than Thought". Forbes. Retrieved 2013-10-31.

21. "News: Live Mint". Are Indian companies making enough sense of Big Data?. Live Mint - http://www.livemint.com/. 2014-06-23. Retrieved 2014-11-22.

22. "Hadoop Releases". apache.org. Apache Software Foundation. Retrieved2014-12-06.

23. "Hadoop Releases". Hadoop.apache.org. Retrieved 2014-12-01.