BIG DATA ANALYTICS USING APACHE HADOOP

Vijay Gupta, Shruti Tiwari and Payal Mehta
Jagannath International Management School (GGSIPU) OCF, Pocket 9, Sector-B, Vasant Kunj, New Delhi, India
**Author for Correspondence*

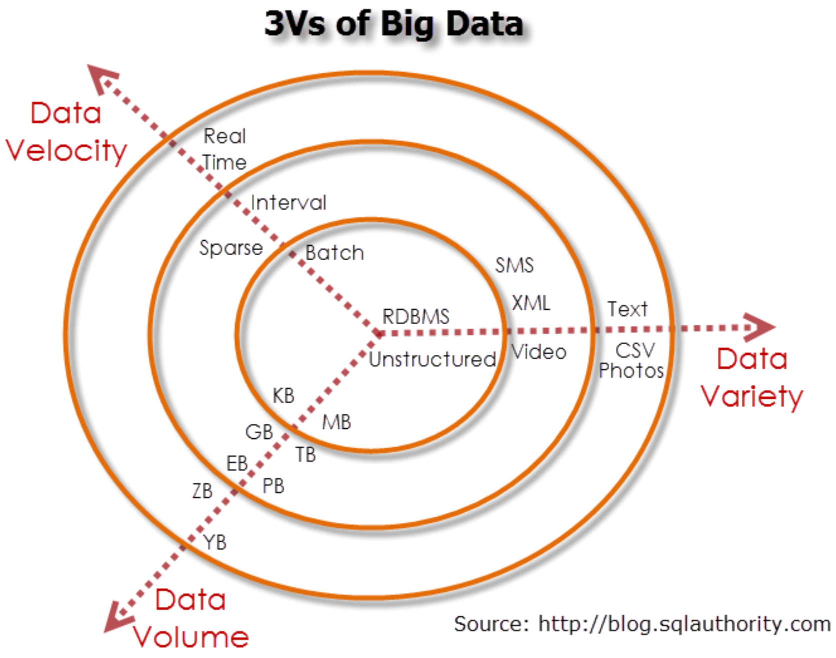
ABSTRACT

Big data is a collection of huge amount of data which is defined by its three main characteristics which are high volume, high velocity and huge variety. Big Data has enormous significance in almost all fields such as education, scientific research etc. In the present world, where users are continuously uploading data to the internet, the overall size of data that has to be stored and studied exceeds the capacity of traditional storage and analysis techniques. Hence, it is becomes mandatory to introduce new and systematic methods for the analysis of Big Data in order to extract useful information from it. Hadoop provides an efficient platform for the analysis of Big Data. It consists of a distributed storage system, HDFS for the storage of large volumes of data**.** In this work, the challenges of Big Data and how Hadoop overcomes them have been identified.

Keywords: *Big Data, Hadoop Ecosystem, HDFS, MapReduce*

INTRODUCTION

Big data is a collection of data sets so complex and huge that it becomes tough to process it by using traditional data processing apps. The challenges related to Big Data involve capturing, searching, analysis, transferring, sharing, visualization, storage and privacy violations (Importance of Big Data, 2014). Big data is a widely known term used to describe the availability of data and exponential growth, both unstructured and structured. Having more data would lead to more accurate evaluation. This leads to better decision making. Making better decisions would mean larger operational efficiencies, reduced risk and cost reductions. In 2001, an analyst named Doug Laney gave the new mainstream definition of Big data. He expressed that Big data can be described with three aspects namely volume, velocity and variety.



Source: http://blog.sqlauthority.com

Figure 1: 3 Vs of Big Data

A. Volume

There is a massive growth in the data storage as the data is now more than text data (Dave, 2013). Data can be found in the form of videos, music and images on social media. It is very common to have zetabytes of the storage system for enterprises.

As the database grows the applications and architecture built to support the data needs to be reassessed quite often. Sometimes the same data is reassessed 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*.

B. Velocity

There used to be a time when we believed that data of yesterday is recent. The matter of the fact is that newspapers still follow that logic. However, news channels and radios have changed how fast we receive the news.

Today, people depend on social media to update them with the latest happening. On social media sometimes a few seconds old message (a tweet etc.) is not something that attracts 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 decreased to seconds. This high velocity data represent *Big Data*.

C. Variety

Data can be stored in multiple forms. For example database, excel, access or in a simple text file. Sometimes the data is not even in the traditional form as we assume, it may be in the form of video, SMS, pdf or something we might have not even thought about. It will be easy to analyze the data in the same form, however, it is not the case most of the time. The real world has data in many different forms and that is the challenge we need to overcome with the *Big Data*. This variety of the data represents *Big Data*.

Categories of Big Data

Big data can be found in three forms:

A. *Structured*

Any data which can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data (Introduction to Big Data- Types, Characteristics & Benefits, no date).

Over the period of time, recent advancements in the IT field have led to accomplishments in developing techniques for working with the data where the format is well known in advance. However, now days, we are focusing on issues when size of such data grows to a huge extent- even in the range of zetabyte. Example: Employee details stored in an employee table.

B. Unstructured

Any data with unknown format or structure is defined as unstructured data. Unstructured data is not only huge in size but it also poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a data source containing a combination of simple text files, images, videos etc. Nowadays, organization have huge amount of data available to them but unfortunately they don't know how to obtain meaning out of it as this data is in its raw form or unstructured format. Example: output returned by 'Google Search' (Introduction to Big Data- Types, Characteristics & Benefits, (no date)).

C. Semi-Structured

Semi-structured data can have both the formats of data, i.e. we can perceive semi-structured data as structured in form but it is actually not defined for e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file (Introduction to Big Data- Types, Characteristics & Benefits (no date)).

Benefits of Big Data

Big data analytics helps organizations tackle their data and take advantage of it to identify new opportunities. This also leads to smarter business decisions, more efficient operations, higher profits and happier customers. In the report *Big Data in Big Companies,* IIA Director of Research Tom Davenport interviewed many businesses to analyze how they used big data (Bender, 2014). He found the following ways Big Data is useful:

1) *Cost Reduction***:** Hadoop and cloud-based analytics bring relevant cost advantages for storing large amounts of data and they identify more efficient ways of doing business.

2) *Faster, Better Decision Making*: With the speed of Hadoop combined with the ability to evaluate new sources of data, businesses are able to understand information quickly – and make decisions based on their analysis.

3) *New Products and Services*: With the ability to predict customer needs and satisfaction through analytics comes the power to give customers what they desire. Researchers point out that with big data analytics, more companies are creating new products to meet customers' needs.

4) When big data is effectively and efficiently collected, processed, and analyzed, companies are able to gain a better understanding of their business, customers, products, competitors, etc. which can lead to increased sales, lower costs, better customer service, and/or improved products and services. Example: Netflix uses big data to improve customer experience (Bender, 2014).

Hadoop

Hadoop is an Apache project founded in 2008 by Doug Cutting at Yahoo and Mike Cafarella at the University of Michigan (Polato *et al.*, 2014). Hadoop can be thought of as a computing environment built on top of a distributed clustered file system which was designed especially for very large-scale data operations (Zikopoulos *et al.*, no date).

However, Hadoop cannot be understood to be a replacement to RDBMS. It adds characteristics to RDBMS features to enhance the efficiency of DBMS technology. Moreover, it resolves the issue of different sets of data problems that the traditional database system is not able to solve (Bhatia and Gupta, 2015).

Benefits of Using Hadoop

Some of the key features that led to the popularity (Alam and Shakil, 2016) of hadoop are:

1) *Economic Benefits*: There is no initial capital investments are required to set a hadoop cluster.

2) *Scalability:* Hadoop provides support for dynamic scalability. By the addition and ejection of machines, any amount of data can be stored and processed using Hadoop framework.

3) *Reliability:* Hadoop creates three copies of a data set i.e. replication factor of a Hadoop job is three. This feature offers its users with a robust and reliable framework because even if one of the machines holding that data set goes down the system will still be running as the data will be available at other replicated locations.

4) *Flexibility:* The number of machines can easily be added and ejected from the cluster depending upon the user desires.

Hadoop Ecosystem

Hadoop ecosystem consists of the following components:

A. Oozie

It organizes the workflow to manage hadoop jobs. It is a server-based workflow Engine specialized in performing workflow jobs with actions that run Hadoop MapReduce and Pig jobs. Oozie is executed as a Java Web-Application that runs in a Java Servlet-Container (Kumbhare, no date).

B. Hive

Hive is a SQL-based data warehouse system for Hadoop that enables data summarization, impromptu queries, and the evaluation of huge datasets stored in Hadoop supported file systems (e.g., HDFS, MapR-FS, and S3) and some NoSQL databases. Hive's SQL dialect is called HiveQL (Haddop Ecosystem - Think Big Analytics, (no date)).

C. Pig

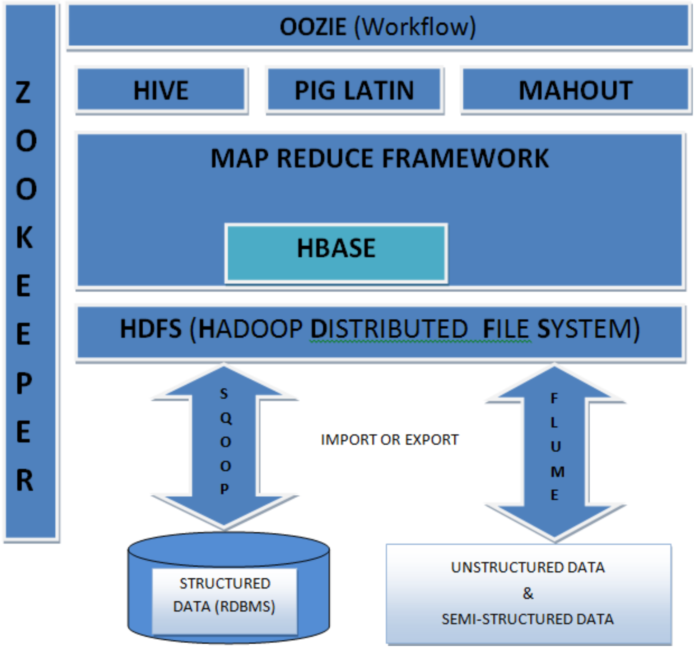
Pig is a platform for building data flows for the ETL (extract, transform, and load) processing and evaluation of large datasets (Alam and Shakil, 2016). Pig Latin, the programming language for Pig provides common data handling operations, such as grouping, joining, and filtering. Pig generates Hadoop MapReduce jobs to enforce the data flows.

D. Mahout

Mahout is an open source machine learning library from Apache written in java (Bhatia and Gupta, 2015). The algorithms it uses fall under the category of machine learning or collective intelligence. Machine learning implementations in Mahout are written in Java, and some parts are constructed upon Apache's Hadoop distributed computation project.

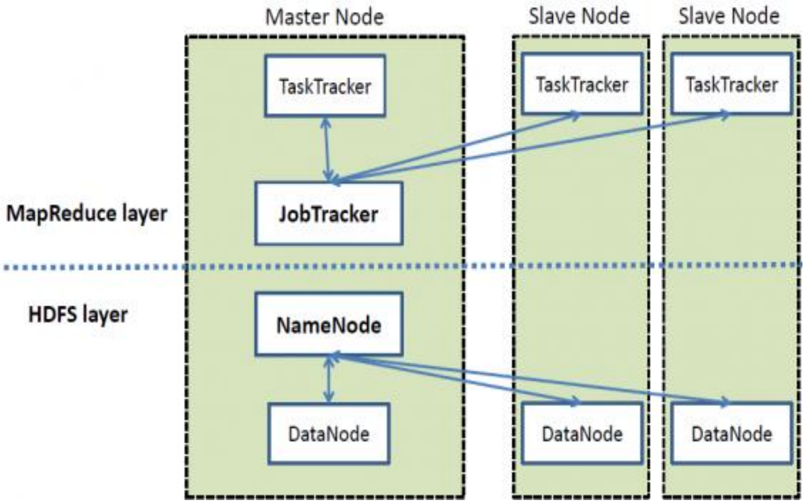
E. Hadoop MapReduce

There are two fundamental versions of Hadoop MapReduce. In the first version called MRv1, Hadoop MapReduce is basically based on two components: (1) the Task Tracker which aims to manage the execution of the Map/Reduce functions and (2) the Job Tracker which depicts the master part and allows resource management and job scheduling. The Job Tracker manages the Task Trackers (White, 2012).



Source: http://hadooptutorials.co.in

Figure 2: Hadoop Ecosystem



Source: http://a4academics.com

Figure 3: Master-Slave Architecture

In the second version of Hadoop called YARN (Yet Another Resource Negotiator), the two major functions of the Job Tracker have been divided into separate daemons: (1) a global Resource Manager and (2) per application Application Master, the Resource Manager receives and executes MapReduce jobs. The per-application Application Master acquire resources from the Resource Manager and works with the Node Manager(s) to execute and monitor the tasks. In Yarn, the Resource Manager takes place of the Job Tracker (Yang, 2015).

Steps of the map reduce algorithm are (Kharode and Deshmukh, 2015):

1) *Map step:* In this step, it takes key and value pairs of input data and changes into output intermediate list of key/ value pairs.

map (key_{in}, $value_{in}$) → list (key_{out}, $value_{intermediate}$)

2) *Reduce step:* In this step, after being rearranged and sorted, the output intermediate key pair is passed through a reduce function where these values are merged together to form a smaller sets of values.

reduce (key_{out}, list ($value_{intermediate}$) → list ($value_{out}$))

The premise is that the entire dataset- or at least a good portion of it- is processed for each query (Hadoop Illuminated).

But this is its power. MapReduce is a batch query processor, and the ability to run an ad hoc query against your whole dataset and get the results in a reasonable time is transformative. It changes the perception about data and unlocks data that was earlier archived on tape or disk. It gives people the opportunity to innovate with data.

The questions that took too long to get answered can now be answered, which in turn leads to new questions and new insights.

F. HBASE

Hadoop Database or HBASE is a non-relational (NoSQL) database which runs on top of HDFS. HBASE was built for big tables which have many rows and columns with fault tolerance ability and horizontal scalability based on Google Big Table (Hadoop Ecosystem).

G. HDFS

Hadoop Distributed File System is an open source implementation of the distributed Google File System (GFS) (Ghemawat *et al.*, 2003). It presents a scalable distributed file system for storing large files over distributed machines in a reliable and structured way (Dean and Ghemawat, 2008). As shown in Figure 3, HDFS consists of a master /slave architecture with a Name Node being master and several Data Nodes as slaves.

1) *Name Node*: Name Node acts as the master of the system and is responsible for the co-ordination of blocks on the Data Nodes (Kumbhare, no date). It runs on high quality software, as it’s responsible for the storage of meta-data. It is the only entry point for the occurrence of a failure in Hadoop cluster.

2) *Data Node*: Data Nodes acts as slaves and it is utilized on each machine in the cluster. Data Node stores the actual data. Its responsibility is to process the requests of client for read and write on the blocks.

H. Sqoop

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop (Apache Sqoop, 2016). It supports incremental loads of a single table or a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import.

I. Flume

Flume is a distributed, reliable and functional service for efficiently collecting, summarizing, and shifting large amounts of data. It has a easy and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses basic extensible data model that allows for online analytic application (Apache Flume, 2016).

J. Zookeeper

ZooKeeper is a centralized service for retaining configuration information, naming, providing distributed coordination, and providing group services (Polato *et al.*, 2014). In case of any failure users can connect to any node and ensure that they will receive the correct, up-to-date information.

Applications of Big Data in Various Fields

1) *Significance to Industrial Upgrades:* Big data is currently a common challenge faced by many industries, and it brings massive issues to these industries' digitization (Jin *et al*., 2015). Research on common problems of big data, especially on innovation of core technologies, will enable industries to tackle the complexity induced by data interconnection and to master unreliability caused by redundancy and shortage of data.

2) *Significance to Scientific Research:* The development of big data has spawned a new research paradigm; that is, with big data, researchers may only require to find or derive from it the required information, knowledge and intelligence. They even do not require to directly access the entities to be studied.

3) *Significance to Helping People Better Perceive the Present:* Big Data, specifically big networked data, contains societal information and can thus be viewed as a network mapped to society. Thus, analyzing big data and further summarizing laws it implicitly contains can help us better perceive the present.

4) *International Initiatives on Big Data:* Because of the great value of big data, many countries have introduced their plans on big data related research and applications. The initiative targets to develop core technologies to collect, store, manage, evaluate and share large-scale data, and use these technologies to accelerate the speed of discovery in science and engineering, strengthen national security, fully change the education and learning mode, and enthusiastically cultivate new skills for developing and using big data technologies (Jin *et al*., 2015).

Challenges of Big Data and How Hadoop Overcomes them

1) *Storage:* One of the major problems of regular file system is that the data is too big to be stored in one computer (Hadoop Illuminated). This eliminates lot of storage system and databases that were built for single machines. In Hadoop, the data is stored on multiple computers which provides a solution to this problem.

2) *Cost:* Another challenge for the regular file system is that very high end machines are expensive. Traditional storage machines are expensive with top-end components. To overcome this challenge, Hadoop was run on commodity hardware.

3) *Hardware Failures:* The commodity hardware fails in many cases as they don't have high end components like the main frames. The Hadoop software is intelligent enough to deal with hardware failures.

4) *Data Loss:* Hardware failure may lead to data loss in most file systems. Data is spread out all over the nodes in a network of machines. When a node fails, all the data on that node will become unavailable or lost. One approach is to make multiple copies of this data and store them on different machines so that even if one of the nodes fails, other nodes will have the data. This is called 'replication'. It has been implemented in HDFS.

5) *Co-ordination:* It is a major challenge among the distributed nodes in a network. The approach used in HDFS for overcoming this problem is to have a master to be the coordinator. It simplifies architecture, design and implementation of the system.

CONCLUSION

Big Data has proved its importance in different areas from sports, education and healthcare to large scale data analytics. If properly stored and analyzed, Big Data can provide fruitful results in almost all these areas. Apache’s Hadoop with its HDFS and MapReduce provides an efficient framework to analyze Big Data. As a future scope we can write different MapReduce programs for analysis of Big Data. We can also edit the source code of Hadoop to enhance the performance of Hadoop.

REFERENCES

Apache Flume (no date). [Online]. Available: flume.apache.org [Accessed 3 March, 2017].

Apache Sqoop (2016). [Online]. Available: sqoop.apache.org [Accessed 3 March, 2017].

Kumbhare D (no date). *Understanding Hadoop Ecosystem* [Online]. Available: www.hadooptutorials.co.in [Accessed 28 February, 2017].

Hadoop Ecosystem (no date). [Online]. Available: www.thinkbiganalytics.com [Accessed 15 February, 2017].

Hadoop Illuminated (no date). [Online]. *Hadoop Distributed File System (HDFS) – Introduction.* [Online]. Available: www.hadoopilluminated.com [Accessed 28 February, 2017].

Introduction to Big Data- Types, Characteristics & Benefits (no date). [Online]. Available: www.guru99.com [Accessed 12 January, 2017].

Dean J and Ghemawat S (2008). Map Reduce: simplified data processing on large clusters. *Communications of the ACM* **51**(1) 107-113.

Yang J *et al.*, (2015). Mapreduce parallel programming model: A state-of-the-art survey. *International Journal of Parallel Programming* **2015** 1-35.

Alam M and Shakil KA (2016). Big Data Analytics in Cloud environment using Hadoop. *arXiv preprint Available at:* https://arxiv.org/pdf/1610.04572

Bender M (2014). *What is Hadoop and how does it Compare to Relational Databases?* [Online]. Available: blog.performancearchitects.com [Accessed 11 February, 2017].

Zikopoulos PC *et al*., (no date). All About Hadoop: The Big Data Lingo Chapter. In *Understanding Big Data*, (USA, New York, The McGraw Hill Companies).

Dave P (2013). *Big Data – What is Big Data – 3 Vs of Big Data – Volume, Velocity and Variety – Day 2 of 21* [Online]. Available: www.blog.sqlauthority.com [Accessed 17 January, 2017].

Polato I *et al.*, (2014). A comprehensive view of hadoop research systematic literature review. *Journal of Network and Computer Applications* **46** 1-25.

Bhatia P and Gupta S (2015). Correlated Appraisal of Big Data, Hadoop and MapReduce. *Advances in Computer Science: an International Journal* **4**(4) 16.

Kharode RM and Deshmukh AR (2015). Study of Hadoop Distributed File system in Cloud Computing. *International Journal of Advanced Research in Computer Science and Software Engineering* **5**(1) 990-993.

Ghemawat S *et al.*, (2003). The Google File System. In: *Proceedings of ACM Symposium on Operating Systems Principles (SOSP),* Bolton Landing, NY 29-43.

White T (2012). *Hadoop: The Definitive Guide*, (O'Reilly Media, Inc., California, USA).

The Importance of Big Data (2014). [Online]. Available: www.charc-concepts.org [Accessed 15 January, 2017].

Jin X *et al.*, (2015). *Significance and Challenges of Big Data Research.* [Online]. Available: www.elsevier.com [Accessed 27 January, 2017]