

International Conference on Intelligent Computing, Communication & Convergence
(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,
Bhubaneswar, Odisha, India

A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology

Ishwarappa¹, Anuradha J²

¹*Department of Computer Engineering, JSPM'S Jayawantrao Sawant College of Engineering, Pune, Maharashtra, India*

²*School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India*

ishwar.kalbandi@gmail.com¹; januradha@vit.ac.in²

Abstract

Big data is a collection of massive and complex data sets and data volume that include the huge quantities of data, data management capabilities, social media analytics and real-time data. Big data analytics is the process of examining large amounts of data. There exist large amounts of heterogeneous digital data. Big data is about data volume and large data set's measured in terms of terabytes or petabytes. This phenomenon is called Bigdata. After examining of Bigdata, the data has been launched as Big Data analytics. In this paper, presenting the 5Vs characteristics of big data and the technique and technology used to handle big data.

The challenges include capturing, analysis, storage, searching, sharing, visualization, transferring and privacy violations. It can neither be worked upon by using traditional SQL queries nor can the relational database management system (RDBMS) be used for storage. Though, a wide variety of scalable database tools and techniques has evolved. Hadoop is an open source distributed data processing is one of the prominent and well known solutions. The NoSQL has a non-relational database with the likes of MongoDB from Apache.

* Corresponding author. Tel.: +91-7767987465

E-mail address: ishwar.kalbandi@gmail.com

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)

Keywords: RDBMS, NoSQL, Big Data.

1. Introduction

Big Data is something so huge and complex that it is impossible for traditional systems and traditional data-warehousing tools to process and work on them. Data (Big Data) is generated by machines, generated by humans, and also generated by mother nature. With the growth of technologies and services, this large data is produced that can be structured, semi-structured and unstructured from the different sources. Big data can neither be worked upon by using traditional SQL like queries nor can the relational database management system (RDBMS) be used for storage. So that a wide variety of scalable database tools and techniques have evolved. Hadoop, an open source distributed data processing system is one of the prominent and well known solutions. The NoSQL has gained prominence as a non-relational database with the likes of MongoDB, Dynamo DB from Apache^{1,2}.

The need of big data comes from the Big Companies like Google and Facebook. For the purpose of analysis of big amount of data which is in unstructured form. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web social media, images, audio and so on. The paper is divided in the following sequence: Starting with the introduction, we talk about the characteristics of big data (5V's). It is followed with a descriptive note on the various components of Big Data based on Hadoop framework. Apache Hadoop is an open source software framework for storage and large scale processing of data sets on clusters of commodity hardware. Hadoop was developed by Doug Cutting and Mike Cafarella in 2005¹.

2. Characteristics

Big data can be described by the following characteristics:

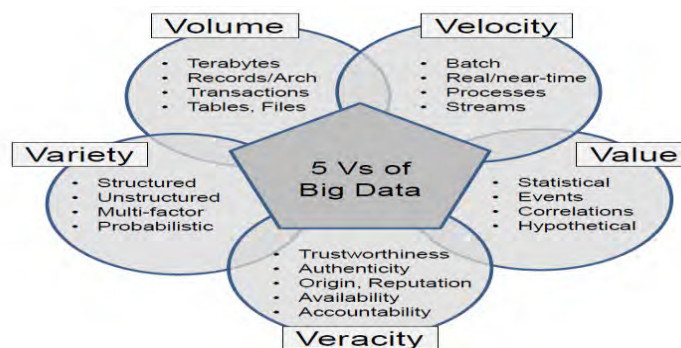


Fig1: 5Vs of Big Data

- **Volume**

This volume presents the most immediate challenge to conventional IT structures. This is the aspect that comes to most people's minds when they think of Big Data. Many companies already have large amounts of archived data in the form of logs, but do not have the capacity to process that data. The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics.

- **Velocity**

Velocity refers to the increasing speed at which this data is created, so, the increasing speed at which the data can be processed, stored and analyzed by relational databases. Velocity refers to the speed at which new data is generated and the speed at which data moves around. About social media messages going to

viral in seconds In 1999, Wal-Mart's data warehouse stored 1,000 terabytes (1,000,000 gigabytes) of data. In the year 2012, it had access to over 2.5 petabytes (2,500,000 gigabytes) of data. Every minute of every day, we upload hundreds hours of video on Youtube, We send over 200 million emails through Gmail's.

- **Variety**

The next aspect of Big Data is its variety. Big Data is not always structured data and it is not always easy to put big data into a relational database. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. Dealing with a variety of structured and unstructured data greatly increases the complexity of both storing and analyzing Big Data. 90% of data generated is data is in unstructured form.

- **Veracity**

When we are dealing with a high volume, velocity and variety of data, it is not possible that all of the data is going to be 100% correct there will be dirty data. The quality of the data being captured can vary greatly. The data accuracy of analysis depends on the veracity of the source data.

- **Value**

Value is the most important aspect in the big data. Though, the potential value of the Big Data is huge. It is all well and good having access to big data but unless we can turn it into value it is become useless. It becomes very costly to implement IT infrastructure systems to store big data, and businesses are going to require a return on investment.

3. Techniques And Technology

Since big data is not only large, but also varied and fast-growing many technologies and analytical techniques are needed in order to attempt extracting relevant information. For processing the large amount of data, the big data requires exceptional technologies. This techniques and technologies have been introduced for manipulating, visualizing and analyzing of big data.

So, to handle big data there are many solutions are available, but the Hadoop technology is one of the most widely used technologies^{3,4}.

3.1 Techniques

There are a many type of techniques that could be employed when attacking a big data project. Which ones are used depends on the type of data being analyzed, the technology available to you, and the research questions you are trying to solve. Some of the tools that came up frequently in the reviewed material are summarized here.

3.2 Technology

There are several software products and available technologies to facilitate big data analytics. Some of the most commonly used technology will discuss in this paper. Hadoop is key technology used to handle big data, its analytics and stream computing. It is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers.

4. Big Data Components On Hadoop Framework

Apache Hadoop is an open-source framework that deals with distributed computing of large datasets across clusters of computers using simple programming models. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming. Hadoop is designed to scale up from single servers to thousands of machines, each of these offering local computation and storage. Created by Doug Cutting and Mike Caferella in the year 2005, it is name

after a toy elephant. Scaling up from Single servers to thousands of machines with Local storage and computation are the advantages that Hadoop offers. This is one of the major advantages that Hadoop offers as we can use inexpensive hardware^{6, 8, 9}.

The Hadoop includes these modules

- **Hadoop Distributed File System (HDFS)**

A distributed file system that provides high-throughput access to application data. It is a distributed file system that helps to store large amounts of data in a reliable manner providing fault tolerant file system. The master node is called name node and manages the cluster metadata. It follows a Master/Slave structure, in which we have a one or more devices called as slave devices are controlled by one device known as master device. Slave node is called data node and stores data. It is a java based file system⁵.

- **Hadoop YARN / Map Reduce**

A framework for job scheduling and cluster resource management. It is a cluster resource management and has been built as a programming model in the Hadoop framework to process large amounts of data in a distributed & parallel environment on a cluster.

- **HBase**

It is scalable and distributed database that supports structured data storage for large tables. It also provides for transactional kind of capabilities by allowing updates, inserts, deletions etc. HBASE is a Hadoop database which is a non-relational (NoSQL) database that runs on top of HDFS. HBASE allows for random, real time read/write access for the big data, is columnar, and provides fault tolerant storage and fast access⁷.

- **Pig**

A high-level data-flow language and execution framework for parallel computation Apache PIG is a scripting language enabling users to write complex Map Reduce transformations including summarizing/aggregation, joining, sorting etc. One of the main features of PIG is parallel processing enabling it to handle very large datasets.

- **Hive**

A data warehouse infrastructure that provides data summarization and ad hoc querying. Hive is a Data-Warehouse software tool used for managing, querying, summarizing and analyzing large data sets. HiveQL is a SQL like language is used for querying petabytes of data in hive. It is used to analyze data in HDFS and provides full support for map/reduce. The advantage that Hive offers is that it is very similar to the traditional SQL language, fast over big datasets, it is scalable, extensible and provides for various reporting.

- **Sqoop**

Sqoop is a software tool designed to transfer bulk data between Hadoop and relational databases. Sqoop is used to im-port data from external databases into HDFS or HBASE or HIVE. It allows for data imports from and data exports to external relational databases and parallel data transferring. It uses simple SQL query as well as saved jobs that can run number of times for importing the updates regarding the data between Hadoop and relation-al databases.

- **ZooKeeper**

A high-performance coordination service for distributed applications Zookeeper offers operations services in the Hadoop framework. It is a centralized service used for maintaining configuration information, named registry, provides data synchronization and group services that are used by distributed applications. Zookeeper's architecture supports high availability through redundant services. It allows for the various distributing processes to coordinate between themselves through a shared hierarchical name space of registers called znodes.

- Avro

Avro is a data serialization system, which serializes data in a compact binary data format and provides for rich data structures and a container file for storing persistent data. It relies on schemas to read and write data. It uses JSON (Java script open notation) for defining data types & protocols. It makes use of wire format for communicating between Hadoop nodes, and between client program & services.

- Cassandra

A scalable multimaster database with no single points of failure. Apache Cassandra is a high availability, highly scalable and high performance open source distributed database management system having capability of handling huge amount of data across multiple servers. It provides for fault tolerance and is decentralized.

- Mahout

A Scalable machine learning and data mining library As per Wikipedia, the Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filter, cluster and classification.

- Tez

A generalized data-flow programming framework, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive, Pig and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop MapReduce as the underlying execution engine.

- Spark

A fast and general compute engine for Hadoop data. It provides a simple and expressive programming model that supports a wide range of applications, machine learning, stream process, and huge graph computation. Apache spark is a fast data analytics and machine learning algorithmic engine used for processing data at a large scale. Spark is integrated with Hadoop and has an advanced analytical engine which makes it 100 times faster than Hadoop map reduce by utilizing in memory processing.

- Flume

Flume is a reliable distributed service for efficiently collecting aggregating and moving large amount of LOG Data. It helps the users make most use of valuable log data. It allows for streaming data from multiple sources, collecting high volume real time web logs.

Conclusion

Big data provides an opportunity for “big analysis” leading to “big opportunities” to advance the quality of life, or to solve the mysteries of the world. We are in the development area of big data. In this paper details about Big Data have been discussed taking the Hadoop Framework as a base. We have characteristics of Big Data and provided deep information on the various components of big data from a Hadoop perspective⁷. There are various challenges and issues of big data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. In today if we see the information overloads almost everywhere, by centralizing data acquisition and consolidation in the cloud. Bigdata methods offer new insight into existing data sets. Apache Hadoop is a fast-growing data framework^{9,10}.

Apache Hadoop offers a free, cohesive platform that encapsulates data integration, data processing, monitoring and workflow scheduling etc. Future work would involve a detailed study on challenges and issues with big data various industries.

References

1. Apache Software Foundation. Official website www.apache.hadoop.org
2. University of Texas at Austin School of Information Big Data Analytics Dylan Maltby 1616 Guadeloupe, Austin, TX 78701 512-471-

3821

3. Bakshi, K, (2012),” Considerations for big data: Architecture and approach”
4. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165–1188.
5. Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks*, 16(3), 9–20.
6. Big Data basics from oreilly: <http://strata.oreilly.com/2012/01/what-is-big-data.html>
7. White, Tom. Hadoop the Definitive Guide 2nd Edition. United States : O'Reilly Media, Inc., 2010.
8. A. Vailaya, "What's All the Buzz Around "Big Data?"", IEEE Women in Engineering Magazine, December 2012, pp. 24-31,
9. S. Madden, "From Databases to Big Data", IEEE Inter-net Computing, June 2012, v.16, pp.4-6
10. Katal, A Wazid, M.; Goudar, R.H., (Aug,2013),” Big data: Issues, challenges, tools and Good practices”.