Perspective on Big Data Hadoop Tools and Technologies

[1]Shital P. Adkine, [2]Dr. Manish T. Wanjari, [3]Dr.Keshao D. Kalaskar

[1]Dept.of Computer Science,Sardar Patel Mahavidyalaya, Chandrapur, Maharashtra-India
[2]Dept. of computer science, Shivaji Science College, Nagpur, Maharashtra-India
[3]Dept. of computer science, Dr, Ambedkar College, Chandrapur, Maharashtra-India

ABSTRACT

The word 'big data' itself tells everything is big like huge volume of data which can be in a structured, semi-structured, unstructured form. As it is Big so generated in a very large amount making it difficult to process using traditional techniques. To process these generated data traditional big data management system is incompetent to manage the large amount of data with different structures, thus Hadoop the framework which is designed to process the large data sets and provides high performance and fault tolerance from a single server to thousands of machines with different. In this paper we describe a study of big data, Hadoop along with a comparison of various tools and technologies used in big data management.

Keywords-component, Privacy, Unstructured Big Data, Big Data Classification, Big Data Tools

1. Introduction

Big data the term itself indicating its meaning a massive pool of data. Now a day's valuable asset means data. The usage of big data spread due to commercialization and digitization in each area. Mainly big data lies on 6 pillars of v's, volume variety, velocity, veracity, value and variability.

1.1. Big data is characterized with the help of Six Vs

1. Volume: Volume describes as a large quantity of data produced by every day in sets, tables and files by any organization like healthcare, education institutes & commercial business in terms of petabytes & zettabytes.
2. Variety: It refers to types of data which deals with may be structured or unstructured, semi-structured.
3. Velocity: It refers to the rate of growth; the speed of big data is generated Telecommunication produces 35TB of data on per day.
4. Veracity: Veracity defines availability and accountability biases, noise and anomaly in data.
5. Value: Having endless amounts of data but it can be move into value.
6. Variability: Variability deals with the data whose meaning is constantly changing

1.2. Sources of Big Data:

Big data are coming from several different sources. The three main primary sources of big data which includes the organization in

1. Social networks.
2. Traditional business system.
3. Internet of Things (IoT).

The data from these sources can be structured, semi-structured, or unstructured, or any combination of these varieties. Social Networks includes the data, human-sourced information from LinkedIn whatsup Twitter and Facebook, Instagram, Flickr, Pinterest, etc. Traditional Business Systems deals with customersservices likeCommercial transactions, E-commerce like Alibaba, Amazon, Flipkart generates huge amount of logs from which users buying trends, Banking records, Credit cards, healthcare records and Internet of Things include Sensors, traffic, weather, mobile phone location, etc. Security, surveillance videos, and images Satellite images, Data from computer systems (logs, weblogs, etc.) The connectivity of large number of heterogeneous devices produces huge data [3], which includes features such as heterogeneity, variety, unstructured feature, noise, and high redundancy.

1.3 Behavioral types of big data

Different types of big data based on content format are as follows:

Structured Data

The data stored in relational databases table in the format of row and column. Structured data include numbers, text, and dates; in terms of database, it is called *strings*. Data have fixed structures and these structures used for organizations to creating a perfect model. Data model permission to store, process and operate on data. Analysis and storing of structured data is very easy. Because of high cost, limited storage space and techniques used for processing, causes RDBMS the only path to store and process the data effectively. Programming language called Structured Query Language (SQL) is used for managing this type of data.

Unstructured Data

Without any specific structure and due to this could not be stored in a row and column format is unstructured data. The data is contradictory to that of structured data. It cannot be stored in a databank. Volume of this data is growing extremely fast which is very tough to manage and analyze it completely. To analyze the unstructured data advanced technology knowledge is needed.

Semi-structured Data

Data which is in the form of structured data but it does not fit the data model is semi-structured data. It cannot be stored in the form of data table, but it canbe stored in some particular types of files which hold some specific marker or tags. These markers are distinguished by some specific rule and the data is enforced to be stored with a ranking. This form of data increased rapidly after the introduction of the World Wide Web where various form of data need medium for interchanging the information like XML and JSON.

1.4 Margins of Existing Systems

The existing systems have major restrictions preventing their use in applications. The limitations are:

1. Lack of integrity
2. Lack of availability and continuity of service
3. Lack of accuracy
4. Existing systems provide vertical scalability.
5. Inconsistency in data format
6. Risk of mismanagement

Big Data deals with modern tools and techniques, and to process this huge data set the previously work traditional data management system is not work properly to handle this enormous amount of data. Traditional relational databases are obsolete and cannot store and process the data generated from recent business applications [4]. Traditional computational frameworks, system architectures and processing systems are designed to handle structured data [5]. One solution to this is Hadoop which work to solve the problems in existing big data management system, which is design to process effectively by providing scalability fault tolerance h and high performance

Table 1:Comparison of traditional and big data [1]

	Traditional data	Big data	Pros of big data
Types of data	Structured data	Structured, Unstructured and semi-structured	develop variety
Volume	Small amount of data. Range- Gigabyte - terabytes	Large amount of data. Range-<petabytes.	Cost reduces and help business intelligence
Data schema	Fixed schema	Dynamic schema	Preserves the information in data.
Data Relationship	Relationship with data is explored easily	Difficulty in relationship betweendata items.	-
Scaling	More than one server for computing	Single server for computing	Cost effective
Accuracy	Less accurate results	High accurate results	Confident results and reliable

Why Hadoop?

The key features and ability to process enormous amount of data with effective storage, computation and analysis has been a great impulseto take a look into the structural design of the industry leading big data processing framework byApache, Hadoop. Earlier days due to the less advanced technology to deals with unstructured data is not handling by several industries. Hadoop is a solution for big data, change the way and decision-making process be used for unstructured data. Hadoop provides a reliable and scalable platform which is used to solve problems caused by massive amount of heterogeneous data. Hadoop technology accepted because of the features like flexibility, scalability, performance, and cost effective. The Hadoop consists of Hadoop kernel, MapReduce, Hadoop distributed file system (HDFS) and Apache hive etc. MapReduce is a programming model which is used to processing large datasets and analyzing it in a cost-effective manner based on divide and conquers method. The divide and conquer method are implemented by two steps such as Map step and Reduce Step. Hadoop data analytics environment deals with data storage, data processing, data access, data management, privacy data protection.

Hadoop distributed file system (HDFS) for the analysis of massive type of data sets using theMapReduce programming model. Hadoop work on three master points scalability, computation capacity, and storage. Hadoopstores file system metadata known as block**s.** It contains the name node and data nodes. HDFS work on master-slave architecture. An HDFS cluster contains a single name node, a master server that manages the file system namespace and directories in the form of hierarchy structure. Data node divides in totwo files, the first one contains data itself and the second deals with block’s generation stamp. Hadoop Distributed File System is designed to tackle fault tolerant and effective-cost hardware. HDFSHadoop stands for “YET ANOTHER RESOURCE NEGOTIATOR”. It provides different processing techniques, like batchprocessing, interactive processing, stream processing graph processingetc. Hadoop common contains libraries and directories.

Table 2 – List of latest tools available to handle big data [2]

Features	Hadoop	MapReduce	HBase	Hive	Spark	Pig
Data flow	Hadoop is a chain of stages.	MapReduce is bas on distributed programming moc that was designed for processing of huge volumes of datasets in paralle such that it is independently wo without bothering sub work.	HBase is run on top of HDFS and it stores data in the key / value form.	Data flow in Hive behaves at the query execution level right from the UI. Meta store sends metadata info back to the compiler.	Spark represents a data flow in a form of a direct acyclic graph (DAG).	The Pig is used to analyze larger sets of data that presents them as data flows
Data processing	Hadoop uses batch processing system.	MapReduce is a tool which is suitable for parallel processing of huge data.	HBase is used to store data into a column-oriented database format.	Hive is used to summarization of data, query, and analysis.	Spark is micro-batch processing and system	Pig is a tool used for analyzing of huge data sets.
Streaming engine	Hadoop deals with large data sets as input, processes it and produces the output.	MapReducestreaming is a type of native batch processing engine.	The batch load is optimized to run on the Spark execution engine	Hive streaming provides for Software based enterprise content delivery that is done behind the firewall for efficiency and security.	Spark streams data in to micro-batches.	Pig provides a parallel architecture-oriented streaming engine that can update Hadoop data over small portions.
Scalability	Hadoop provides scalable, flexible data storage and analysis.	MapReduce provides scalability means that single server to thousands of different machines	HBase provides extreme scalability, reliability, and flexibility for data.	Hive is much familiar, fast, scalable and extensible.	Spark provides linear scalability in the distributed environment	Pig provides high level scalability
Latency	Hadoop gives higher latency than both Spark and Fink.	MapReduce gives low latency.	HBase is fast and used for low latency data access. It stores data in - memory table	Hive has high Latency as compared to HBase.	Spark gives low latency than Hadoop	Pig is streaming writes, just like Map Reduce. Low latency queries are not supported in

			known as MemStore.			Pig;thus it is not suitable for OLAP and OLTP.
Scheduler	Hadoop provides two types of schedulers. fair scheduler and Capacity scheduler in Hadoop. The scheduler in Hadoop becomes the pluggable component.	MapReduce provides the Fair Scheduler, which provides a way to share large clusters.	HBase Scheduler uses a polling to change state at controlintervals . if required based on configuration it can trigger jobs.	Hive schedules table every hour by use of Oozie schedule.	Spark deals With its own flow scheduler due to in-memory computation.	Oozie is the tool for workflow scheduler in Hadoop for Apache Pig – Secondly, writes a brief Pig script for each data file to extract the required data fields.
Cost	A mid-range Intel server is recommended an enterprise-class Hadoop cluster.	Map reduce Cost is high but Hadoop cluster, a mid-range Intel server is recommended for it.	The cost of HBaseis depends on your usage pattern; S3 listing and file transfer might cost money.	Hive is also open source, and built on top of Hadoop for data querying.	Spark is very costly	Pig is lower in cost to write and maintain compared to MapReduce
Development	Hadoop is developed by Apache Software Foundation.	MapReduce is developed by Google for a new style of large data processing	HBaseis an open-source project that was incubated by Apache Software Foundation.	Hive was initially developed by Facebook, but also some other companies develop and use it.	. Spark is developed in the University of California after some time it's codebase donated to Apache Software Foundation	. Pig is originally developed by Yahoo & Facebook

5. Conclusion

In this paper we concentratedon Big Data &Hadoop along with six V's and big data tools. Traditional big data management systemsare not up to the mark to handle massive data sets. Many issuesarise while handling the big data due to some sort of lack of accuracy, lack of integrity, lack of privacy, etc. Some major issues have to come with the traditional big data management system. To overcome these types of issues Hadoop is the solution for the processing of large data sets very effective manner. It is a future technology which provides excellent scalability as from a single server to thousands of machines according to our requirements. We have discussed a big data tool to handle heterogeneous data. This review is conducted to give academics an appropriate guideline in determining the promising region regarding the Hadoop. Hadoop is indeed a technology to store and process the huge data sets. Major concern that is associated with big data is ensuring its security and integrity. Apache Spark is another tool used in analytics of big data.It is faster performance than Map Reduce. There are almost all tools covered which deals with big data.

References

[1]. *Chunarkar-Patil P, Bhosale A. Big data analytics. Open Access J Sci. 2018;2(5):326–335. DOI: 10.15406/oajs.2018.02.00095*

[2]. *Toshifa, Aniruddh Sanga, Shweta Mongia, International Conference on Advancements in Computing & Management (ICACM-2019)*

[3]. *P. R. B. B, P. Saluja, N. Sharma, A. Mittal, and S. V. Sharma, "Cloud Computing for Internet of Things & Sensing Based Applications," Sensing Technology (ICST), pp. 374–380, 2012.*

[4]. *M. Junghanns, M. Neumann, and E. Rahm, "Management and Analysis of Big Graph Data: Current Systems and Open Challenges," in Handbook of Big Data Technologies, 2017, pp. 457–505.*

[5]. *S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," Electronic Markets, pp. 173–194, 2016.*

[6]. *S. Srivastava, "Appraising a Decade of Research in the Field of Big Data 'The Next Big Thing,'" Computing for Sustainable Global Development (INDIACom), no. 2014, pp. 2171–2175, 2016.*

[7]. *J. Quackenbush, "Microarray data normalization and transformation," Nature Genetics, vol. 32, no. December, pp. 496– 501, 2002.*

[8]. *J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.*

[9]. *R. Casado and M. Younas, "Emerging trends and technologies in big data processing," 2014.*

[10]. *S. Venkatraman, S. Kaspi, Kiran Fahd, and R. Venkatraman, "SQL Versus NoSQL Movement with Big Data Analytics," International Journal of Information Technology and Computer Science, vol. 8, no. 12, pp. 59–66, 2016.*

[11]. *D. J. Abadi, P. A. Boncz, and S. Haritopoulos, "Column-oriented Database Systems," Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1664–1665, 2009.*

[12]. *F. Naumann, "Data Profiling Revisited," vol. 42, no. 4, 2013.*

[13]. *M. Seeger, "Key-Value stores: a practical overview," pp. 1–21, 2009.*

[14]. *B. Baesens, R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao, "Transformational issues of big data and analytics in networked business," MIS quarterly, vol. 38, no. 2, pp. 629–631, 2014.*

[15]. *S. Amini and C. Prehofer, "Big Data Analytics Architecture for Real- Time Traffic Control," Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017.*

[16]. *S. Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data," IEEE Access, vol. 4, pp. 2751–2763, 2016.*

[17]. *B. Zhou and J. Pei, "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," Knowledge and Information Systems, vol. 28, no. 1, pp. 47– 77, 2011.*

[18]. *A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, "L-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, 3–es, 2007.*

[19]. *N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in2007 IEEE 23rd International Conference on Data Engineering, IEEE, 2007, pp. 106–115.*