ISSN:0975-9646

Big Data: Technologies, Trends and Applications

Sudhakar Singh [a,*], Pankaj Singh [b], Rakhi Garg [c], P K Mishra [a]

[a] *Department of Computer Science, Faculty of Science,*
Banaras Hindu University, Varanasi 221005, India

[b] *Faculty of Education,*
Banaras Hindu University, Varanasi 221005, India

[c] *Mahila Maha Vidyalaya,*
Banaras Hindu University, Varanasi 221005, India

Abstract-Big Data is an excessive amount of imprecise data in variety of formats generated from variety of sources with rapid speed. It is most buzzed terms among researcher, industry and academia. Big Data is not only limited to data perspective but it has been emerged as a stream that includes associated technologies, tools and real word applications. The objective of this paper is to provide a simple, comprehensive and brief introduction of Big Data to the beginners in subject. In this paper, we provide an overview of Hadoop and its sub-projects and a brief review of various developed technologies for Big Data. We also discuss some recent trends and eminent applications in Big Data. Although this paper does not touch each and every dimension of Big Data as it is not possible to make it in a single paper but essential aspects are covered, which may benefit to the people new in Big Data world.

Keywords: **Big Data; Hadoop; MapReduce; Yarn; Technology; Eco-System**

1. INTRODUCTION

Big Data is one of the most buzzed and hyped phrase nowadays. Before come to define the Big Data, we would like to first explore the sources generating excessive data. Data may be generated either by human or by machine. Human generates data as documents, emails, images, videos, posts on facebook or tweeter etc. Data comes into machine generated category are sensor data and logs data i.e. web logs, click logs, email logs. Machine generated data are of larger size than human generated data. After the invention of big data technologies, machine generated data came into play in order to process them. Major sources of Big Data are purchase transaction records, web data, social media data, click stream data, cell phone GPS signals, and sensor data [1-2]. Social networking sites like facebook, Twitter, LinkedIn generates a large volume of social media data. Online advertising and E-Commerce companies always looking for user navigation data i.e. users click stream on a website. Sensors embedded in machines generate large amount of data. As the real word examples of Big Data, facebook has 40 PB data captures 100 TB of data per day, Yahoo has 60 PB data and Twitter captures 8 TB data per day [1].

Large scale data processing or analysis and mining intelligence from it is always being a centre of attraction. Typical data analytical tools cannot support large scale data. We have to use some different distributed tools and techniques to analyze such large scale data since traditional storage systems do not have analytical power and traditional data analysis tools are unable to handle Big Data. There may be a reasonable doubt that in spite of well known distributed system like MPI (Message Passing Interface), do we need another distributed system. We need different distributed system since typical distributed system has some problems as follows. First, it is highly dependent on network and requires huge bandwidth. Second, partial hardware or job failures are difficult to handle. Third, it wastes a lot of processing power in movement and distribution of data.

In case of analysis of Big Data, the complex characteristics of Big Data are the major challenges in the way of processing and managing it. A new distributed system Hadoop has been developed for processing large and excessive data in distributed and parallel fashion. We define Big Data in section 2. Section 3 discusses the evolution of Hadoop and describes various components and daemons of Hadoop. Other associated Big Data technologies are described in section 4. Section 5 discusses the recent trends in Big Data. Section 6 enumerates a number of applications of Big Data and technologies. Finally we conclude paper in section 7.

2. BIG DATA

Big Data is that extent of data, which cannot be stored and processed by a single machine. Big Data do not refers to the data only big in size. Most well known definition of Big Data jointly given by Gartner and IBM [2-4] is a four Vs concept: Volume, Velocity, Variety and Veracity. So data possesses large volume, comes with high velocity, from variety of sources and formats and having great uncertainty is referred as Big Data. Volume- represents scale of data i.e. Big Data has massive volume. Velocity- refers speed of generation and processing of data i.e. rate of entering streaming data in the system is really fast. Variety- refers different form of data i.e. unstructured or semi-structured data (text, sensor data, audio, video, click stream, log file, XML) originated from different sources. Veracity- refers uncertainty of data i.e. quality of data being captured. Data like posts on social networking sites are imprecise [5-6].

3. APACHE HADOOP

In this section, we focus on the evolution of Hadoop and architecture of Hadoop components. Daemons of Hadoop as well as its versions are also discussed.

3.1. Evolution of Hadoop

Hadoop was created by Doug Cutting in 2005 [7]. It is consequent result of Nutch search engine project of Dough Cutting. Google published two papers on GFS (Google File System) [8] and MapReduce [9] in 2003 and 2004 respectively. Nutch project was then rewritten to use MapReduce. Cutting jointly with a team at Yahoo! Started a new project and named it after his son's toy elephant. In 2006, Apache Hadoop project was started for the development of HDFS (Hadoop Distributed File System) and Hadoop MapReduce, Now Hadoop is top level project of the Apache software foundation [10]. In 2008, a Hadoop Cluster at Yahoo! has won Terabyte Sort Benchmark [10-11].

3.2. Core Components of Hadoop

Hadoop is a large-scale distributed batch processing infrastructure for parallel processing of big data on large cluster of commodity computers [12]. Hadoop consists of three core components: HDFS, MapReduce and YARN. HDFS and MapReduce design are based on Google's File System and MapReduce. YARN framework is a NextGen MapReduce also called MapReduce 2.0, was added in Hadoop-2.x version for job scheduling and resource management of Hadoop cluster. Hadoop is extremely scalable distributed system and requires minimum networks bandwidth. Hadoop infrastructure automatically handles fault tolerance, data distribution, parallelization and load balancing tasks. In traditional parallel and distributed system, data are moved to the node for computation which can never be feasible in case of Big Data. Hadoop is a joint system providing computational power i.e. MapReduce and distributed storage i.e. HDFS at one place. Its design is based on distributing computational power to where the data is; instead of moving data [13].

3.2.1. HDFS Architecture

HDFS is a distributed file system, which provides unlimited storage, scalable and fast access to stored data. It supports horizontal scalability. Thousands of nodes in a cluster hold petabyte scale of data and if there is a requirement of more storage, one needs to just add more nodes only [1]. It uses block-structured file system and stores the files in a replicated manner after breaking the file into fixed size blocks. Default block size is 64 MB and each block is replicated at three nodes by default. Storing data in this way provides high fault tolerance and availability during execution of Big Data applications on Hadoop cluster [12-13].

Hadoop is designed on Master-Slave architecture. There is single master node known as NameNode and multiple slave nodes known as DataNodes. Master node coordinates all slave nodes. DataNodes are the workhorses and stores all data. NameNode is the administrator of file system operations i.e. file creation, permissions etc. Without NameNode no one can operate cluster and write/read data. NameNode is called a single point failure [1]. Fig. 1 shows the functionality of NameNode and DataNode in HDFS. NameNode assigns a block *id* to each block of a file and stores all the metadata of the files in its memory in order to be fast accessed. Metadata are the file name, permission, replication and location of each block of the file. DataNodes store all the files as replicated blocks and retrieve them whenever required.

3.2.2. MapReduce Framework

MapReduce is an efficient, scalable and simplified programming model for large scale distributed data processing on a large cluster of commodity computers [12] [14-15]. It works on the data residing in HDFS. MapReduce is a programming framework, which provides generic templates that can be customized by programmer's requirements. It process large volumes of data in parallel by breaking the computation job into independent tasks across a large number of machines. It distributes the tasks across machines in Hadoop cluster and put together the results of computations from each machine. It takes care of the hardware and network failure. A failed task is assigned to other node to re-execute itself without re-executing other tasks. It balances the workload and increase the throughput by assigning work of slower or busy nodes to idle nodes [1] [16].

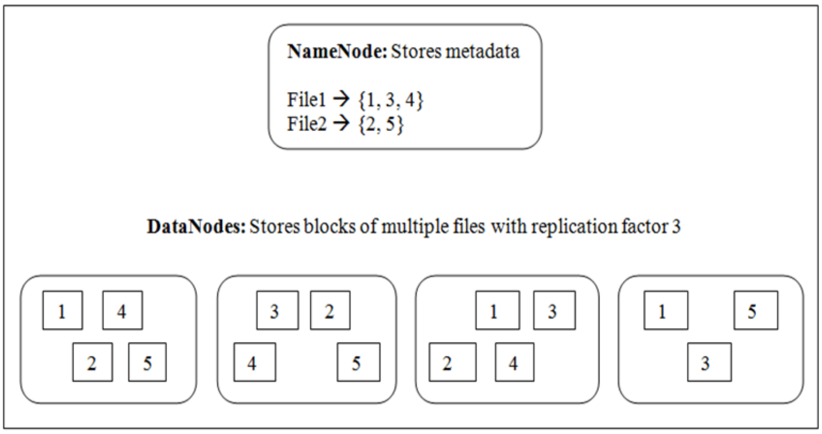
NameNode: Stores metadata
File1 → {1, 3, 4}
File2 → {2, 5}
DataNodes: Stores blocks of multiple files with replication factor 3
1
4
2
5
3
2
4
5
1
3
2
4
1
5
3

Fig. 1. NameNode and DataNodes in HDFS [14]

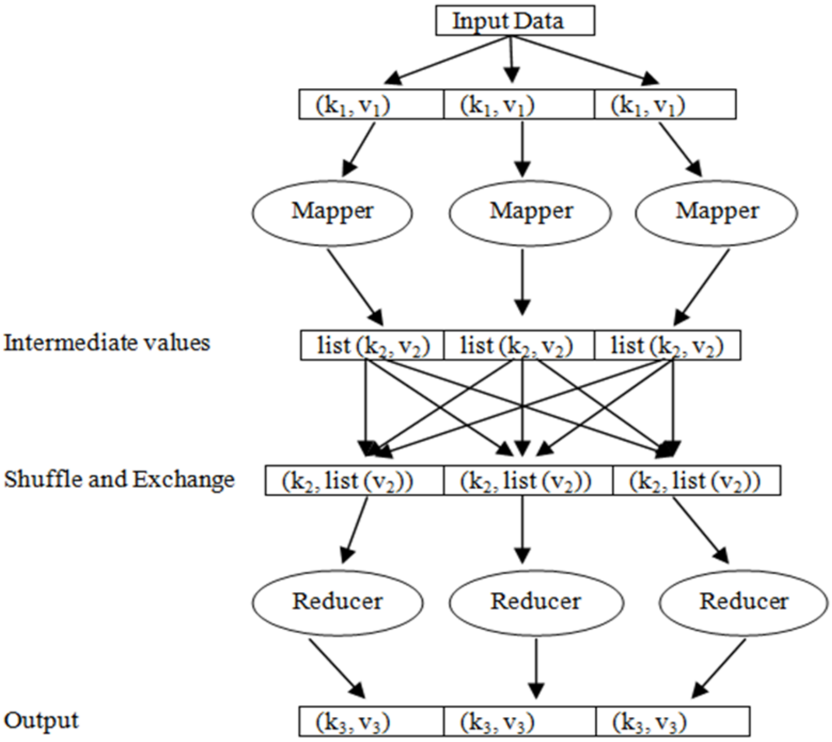

Fig. 2. Work flow in MapReduce framework [14]

Map Reduce programs can run on Hadoop in multiple languages like Java, Python, Ruby and C++ [17]. MapReduce program consists of two functions Mapper and Reducer which runs on all machines in parallel fashion. The input and output of these functions must be in form of *(key, value)* pairs. Fig. 2 illustrates the work flow in MapReduce framework.

Map Function: Mapper is applied in parallel on input data set. The Mapper takes the input *(k_1, v_1)* pairs from HDFS and produces a list of intermediate *(k_2, v_2)* pairs. Mapper output are partitioned per reducer i.e. the number of reduce task for that job.

Reduce Function: The Reducer takes *(k_2, list (v_2))* values as input, make sum of the values in *list (v_2)* and produce new pairs *(k_3, v_3)* as final result.

Combiner Function: It is optional and also known as Mini Reducer. It is applied to reduce the communication cost of transferring intermediate outputs of mappers to reducers.

Shuffle and exchange is the single point of communication in MapReduce. MapReduce framework shuffle the intermediate output pairs of mappers and exchange them between reducers to send all pairs with the same key to a single reducer [12].

3.2.3. Daemon Processes in Hadoop

Hadoop has five daemons that are the processes running in background. These are NameNode (NN), Secondary NameNode (SNN), DataNodes (DN), JobTracker and TaskTrackers and described as follows [12] [19-20].

NameNode: Each Hadoop cluster has exactly one NameNode which runs on master machine. NameNode manages metadata and access control of the file system.

Secondary NameNode: There is also a backup NameNode named as Secondary NameNode which periodically wakes up and process check points and downloads updates from NameNode. It can be used latter to restore failed NameNode, providing fault tolerance.

DataNodes: DataNode runs on each slave machines in cluster and holds file system. Each DataNode manages blocks of the file system assigned to it.

JobTracker: Exactly one JobTracker runs in a cluster. All running tasks are halted if JobTracker goes down. Initially jobs are submitted to JobTracker. Then it talks to the NameNode to determine the location of data and talks to TaskTrackers to submit the tasks.

TaskTrackers: TaskTracker runs on each slave node and accepts map & reduce tasks and shuffle operations from JobTracker.

3.3. Hadoop-1.x vs. Hadoop-2.x

Apache releases a new version of Hadoop after fixing bugs of previous releases and incorporating new functionality and performance improvements. It introduced MapReduce 2.0, an improved and optimized framework in Hadoop-2. The major difference between Hadoop-1.x and Hadoop-2.x is the computational framework, *NextGen MapReduce (YARN)* or *MapReduce 2.0 (MRv2)*. Hadoop-1.x uses *MRv1* which have two daemon process JobTracker on Master and TaskTracker on Slave. While Hadoop 2.x uses MRv2 (YARN), which has *ResourceManager (RM)* on master machine and *NodeManager (NM)* on slave machines and a Application Specific *ApplicationMaster (AM)* [21].

Hadoop YARN (Yet Another Resource Negotiator) is a framework for job scheduling and cluster resource management [10]. In Hadoop-2.x, the functionality of JobTracker of Hadoop-1.x splits into separate daemons, global ResourceManager and per-application

ApplicationMaster for resource management among all applications in the system and job scheduling/monitoring. Data-computation framework is formed by the ResourceManager and per-node slave, the NodeManager. The ApplicationMaster is a framework specific library which negotiates resources from the ResourceManager and works with NodeManagers to execute and monitor the tasks [21].

4. HADOOP ECO-SYSTEM

Apache Software Foundation supports a number of other Hadoop related projects [10]. Each project deals with a certain aspect of Big Data and provides complementary services to Hadoop. The Hadoop related projects come under umbrella of Hadoop Eco-System [22]. We describe each one by one as follows.

1) *HBase:* HBase is the Hadoop database, inspired by Google's BigTable [23]. It is a scalable, distributed and non-relational database that supports storage for big tables of structured data. It uses HDFS as its underlying storage. HBase is used when there is a need of random and real time read/write access of Big Data. It provides BigTable like capabilities on top of Hadoop [24].
2) *Cassandra:* Cassandra is a scalable database provides high availability and supports multi-master to avoid single points of failure. MapReduce can retrieve data from Cassandra. It is a BDDB i.e. Big Data Data Base, which can run without HDFS. Its supporting systems are derived from Google Big Table [23] and Google File System [8] [25].
3) *Hive:* Hive is data warehouse infrastructure that provides data summarization, ad-hoc querying and analysis of large datasets residing in HDFS. It provides a mechanism to project structure on this data and also a query language HiveQL based on SQL. It also provides flexibility to plug in custom mappers and reducers when logic could not be efficiently expressed in HiveQL [26].
4) *Pig:* Pig is a high level data-flow language and also an execution framework for parallel computation. A pig program is amenable to substantial parallelization, which enables them to handle big datasets. Pig's underlying infrastructure consists of a compiler that generates sequences of MapReduce programs whose parallel implementations already exist. Pig's language, Pig Latin express data flow sequences and also provides ability to the users to develop their own function for reading, writing and processing data [27].
5) *Tez:* Tez is a generalized data flow programming framework, currently built on top of Hadoop YARN. It provides a powerful and flexible engine for executing a complex DAG (directed acyclic graph) of tasks to process data in batch or interactive way. It makes MapReduce paradigm to more powerful by expressing computations in data flow graph. Hive, Pig and other framework of Hadoop eco-system is adopting Tez to replace MapReduce jobs [28].
6) *Chukwa:* Chukwa is a data collection system for monitoring large distributed clusters. It is built on top of HDFS & MapReduce framework and provides large scale log aggregation and analytics. It has a flexible and powerful toolkit for displaying, monitoring and analyzing the results to apply on the collected data [29].
7) *Zookeeper:* Zookeeper makes high performance coordination among distributed applications. Several Hadoop projects use Zookeeper to coordinate the cluster and provide highly available distributed services. It gives a centralized service for maintaining configuration information, naming, providing distributed synchronization and providing group services [30].
8) *Ambari:* Ambari is a web-based tool for making Hadoop management simpler. It provision the Hadoop cluster by providing a step-by-step wizard for installing services e.g. Hive, HBase, Pig, Zookeeper etc. on Hadoop cluster and also handles configuration of these services. It provides central management to start, stop and reconfigure the Hadoop services over cluster. It monitors the health and status of Hadoop cluster [31].
9) *Avro:* Avro is a data serialization system. It provides rich data structures; a compact and fast binary data format; a container file to store persistent data; and remote procedure call (RPC). It does not require code generation to read or write data nor to use or implement RPC protocols [32].
10) *Mahout:* Mahout is a machine learning, data mining and math library on top of MapReduce. The goal of this project is to provide scalable and fast machine learning and data mining algorithms [33].
11) *Spark:* Spark is a fast and general engine for processing large scale data. Spark provide an easier to use alternative to MapReduce and run programs up to 100 time faster than Hadoop MapReduce in memory or 10 time faster on disk. It has an advanced directed acyclic graph (DAG) execution engine that supports cyclic data flow and fast in-memory computation. Spark runs on Hadoop and can access HDFS, Cassandra, and HBase [34].

5. BIG DATA TRENDS

Big Data opens new opportunities in research and development and is not only limited to Hadoop and its eco-system. A number of tools and projects dedicated to customized requirements are being developed to deploy on top of Hadoop. Many enterprises are launching their own Hadoop distributions. Cloud computing is using Hadoop to provide data processing and storage services. Computation framework of Hadoop is being efficient and flexible. This section gives a brief description of some trends of Big Data.

5.1. Big Data Eco-System

Big Data Eco-system is even bigger than Hadoop Eco-System and growing rapidly. We can categorize the projects and tools of Big Data Eco-System on the basis of their core functionality for which they are developed. Table 1 summarizes the Big Data related projects.

Table 1. Summary of Big Data related Projects [1]

Sl. No.	Core Functionality	Tools/Projects
1	Getting Data into HDFS	Flume, Chukwa, Scoop, Kafka, Scribe
2	Compute Frameworks	MapReduce, YARN, Cloudera SDK, Weave
3	Querying Data in HDFS	Pig, Hive, Cascading Lingual, Stinger, Hadapt, Greenplum HAWQ, Cloudera Search
4	Real Time Data Access	HBase, Apache Drill, Citus Data, Impala, Phoenix, Accumulo, Spire
5	Big Data Database	HBase, Cassandra, Amazon SimpleDB, Redis, Voldermort
6	Hadoop in the Cloud	Amazon Elastic MapReduce (EMR), Whirr
7	Work Flow Tools	Oozie, Cascading, Scalding, Lipstick
8	Serialization Framework	Avro, Protobuf, Trevni
9	Monitoring Systems	Hue, Ganglia, Open, Nagios
10	Applications	Mahout, Giraph
11	Stream Processing	Storm, Apache S4, Samza, Malhar
12	Business Intelligence Tools	Datameer, Tableau, Pentaho, SiSense, SumoLogic

5.2. Hadoop Distributions

A distribution provides easy installation and packages multiple components to work together. It is tested and patched with works & improvements. Hadoop is an open source project of Apache. Like Linux distributions as RedHat, Ubuntu and Suse some enterprises launched their own Hadoop distributions with tools to manage and administer the cluster and also with a free/premium policy. Cloudera [35] is an oldest distribution of Hadoop. HortonWorks [36] is a newer distribution very close to Apache Hadoop. MapR [37] provides its distribution with their own file system alternative to HDFS. Intel [38] provides its distribution with encryption support.

5.3. Hadoop in the Cloud and Virtualized Environment

Hadoop is originally designed to process on cluster of physical machines but now it is also used in cloud and virtual machines [39-40]. Hadoop clusters can be set up in public and private cloud. Amazon offers on demand Hadoop cluster. Google provides Hadoop on Google Compute Engine. Hadoop can be launched as a service in the public cloud like AWS, Rackspace, MS Azure, IBM Smart Cloud etc. Amazon's EMR (Elastic MapReduce) offers a quick and easy way to run MapReduce jobs without installing Hadoop clusters on its cloud. MapR is the only commercial distribution available through the EMR service. Amazon EC2 (Elastic Compute Cloud) service also provides option to independently deploy MapR. Hadoop can be run using Amazon's S3 (Simple Storage Service) instead of HDFS [41-45]. Hadoop clusters deployed in virtual infrastructures have their own benefits. A single image can be cloned save operation costs. Cluster can be set up on demand and physical infrastructure can be reused. Also cluster size can be enlarged or reduced on demand [46].

5.4. Hadoop as a Big Data Operating System

Hadoop is turning into a general purpose data operating system. Its distributed analytic frameworks MapReduce 2.0 i.e. YARN is a now functioning as distributed resource manager. YARN provides the daemons and APIs to develop generic distributed applications of real world and also handles and schedule resources. Different data analytics operations i.e. graph analytics, streaming data analysis etc. can be plugged in with Hadoop to use storage and computation framework [47-48].

5.5. Big Data Security and Privacy Issues

Big Data characteristics volume, velocity, variety have magnified the security and privacy issues. Security and privacy issues become more critical due to data hosted in large scale cloud infrastructures, diversity of data format and sources, streaming data and high volume inter-cloud migration. Large scale cloud infrastructures use a diversity of software platforms and are spread across large networks of computers, which provide more opportunities to attackers [49]. A. C. Mora *et.* al surveyed and drafted a list of top ten Big Data security and privacy challenges

6. Big Data Applications

Big data technologies have wide and long list of their applications. It is used for Search Engine, Log Processing, Recommender System, Data Warehousing, Video and Image Analysis, Banking & Financial, Telecom, Retail, Manufacturing, Web & Social Media, Medicine, Healthcare, Science & Research and Social Life. We are discussing some of the eminent applications here.

6.1. Politics

Big Data analytics help Mr. Barack Obama to win the US presidential election in 2012 [51]. His campaign was built of 100-strong analytics staff to shake dozens of terabyte scale data. They used a combination of the HP Vertica massively parallel processing analytical database [52] and predictive models with R [53] and Stata [54] tools.

6.2. National Security

Babak Akhgar *et.* al [55] authored a book on Application of Big Data for National Security. Authors relate the Big Data technologies to national security and crime detection and prevention. They present strategic approaches to deploy Big Data technologies for preventing terrorism and reducing crime.

6.3. Health Care and Medicine

Big Data technologies can be used for storing and processing medical records. Streaming data can be captured from sensors or machines attached to patients, stored in HDFS and analyzed quickly [1]. With Big Data tools and human genome mapping, there may be a commonplace for people to have their genes mapped as the part of their medical record. Genetic determinants that cause a disease will be easy to find, which help in the development of personalized medicine [56].

6.4. Science and Research

Science and research are now driven by technologies. Big Data adds new possibilities to them. CERN, the European Organization for Nuclear Research have started the world's largest and most powerful particle accelerator, Large Hadron Collider (LHC). The experiment generated excessive amount of data. Data center at CERN has 65,000 processors, which analyzed 30 petabytes of data. Its computing powers of thousands of computers are distributed across 150 data centers worldwide [57-58].

6.5. Social Media Analysis

IBM provides a social media analytics, a powerful SaaS solution to discover hidden insights from millions of web sources. It is used by businesses to gain a better understanding of their customers, market and competition. It captures consumer data from social media, predicts customer behavior and creates customized campaign [59].

7. Conclusion

Big Data is not only concerned to data big in volume but also data with big velocity, big variety and big veracity. Big Data has introduced a new attitude in data processing and analysis and new opportunities to provide solutions of real world problems, which are considered infeasible as before. Apache Hadoop is the most revolutionary technology which opened the door of infinite possibilities in Big Data. Initially Hadoop developed with two core components HDFS and MapReduce. YARN, the NextGen MapReduce framework turns Hadoop as a general purpose data operating system. Apache supports a number of sub-projects providing specific services and works on top of Hadoop. Apache is not the only organization that develops tools and projects for Big Data, many other organizations are also contributing and some provides their own Hadoop distributions. Hadoop can also be set up and configured in cloud and virtualization infrastructures. Cloud provides Hadoop services without having our own cluster while virtualization enables us to set up on demand Hadoop clusters. Hadoop is adopted in wide areas from science & engineering to social life and has changed the way of thinking and solving problems.

References

[1] Mark Kerzner and Sujee Maniyam, "Hadoop Illuminated," https://github.com/hadoop-illuminated/hadoop-book , 2013, Accessed on Sept. 20, 2015.

[2] L. Douglas, "3d data management: Controlling data volume, velocity and variety," Gartner, Retrieved 6 (2001).

[3] IBM What is big data? - Bringing big data to the enterprise. http://www-01.ibm.com/software/in/data/bigdata/, Accessed on Sept. 20, 2015.

[4] M. A. Beyer and L. Douglas, "The importance of big data: A definition," Stamford, CT: Gartner, 2012.

[5] IBM Big Data & Analytics Hub, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, Accessed on Sept. 20, 2015.

[6] J. S. Ward and A. Barker, "Undefined By Data: A Survey of Big Data Definitions," http://arxiv.org/abs/1309.5821v1.

[7] Tom White, "Hadoop: The definitive guide," O'Reilly Media, Inc., 2012.

[8] S. Ghemawat, H. Gobioff and ST Leung, "The Google file system," in ACM SIGOPS operating systems review, vol. 37, no. 5, ACM, 2003.

[9] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in Proc. 6th Symposium on Opearting Systems Design & Implementation, 2004.

[10] Apache Hadoop, http://hadoop.apache.org

[11] Sort Benchmark, http://sortbenchmark.org/

[12] Yahoo! Hadoop Tutorial, http://developer.yahoo.com/hadoop/tutorial/index.html, Accessed on Sept. 20, 2015.

[13] HDFS Architecture Guide, https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, Accessed on Sept. 20, 2015.

[14] Sudhakar Singh, Rakhi Garg and P K Mishra, "Review of Apriori Based Algorithms on MapReduce Framework," in Proc. International Conference on Communication and Computing (ICC - 2014), Elsevier Science and Technology Publications, 2014.

[15] MapReduce Tutorial, http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html, Accessed on Sept. 20, 2015.

[16] K-H. Lee, Y-J. Lee, H. Choi, Y. D. Chung and B. Moon, "Parallel Data Processing with MapReduce: A Survey," in ACM SIGMOD Record, vol. 40, no. 4, pp. 11–20, (2011).

[17] Hadoop Tutorials, http://hadooptutorials.co.in/tutorials/hadoop/understanding-hadoop-ecosystem.html, Accessed on Sept. 20, 2015.

[18] Hadoop Architecture Overview, http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview.html, Accessed on Sept. 20, 2015.

[19] IBM developerWorks, http://www.ibm.com/developerworks/library/l-hadoop-1/, Accessed on Sept. 20, 2015.

[20] R. P. Padhy, "Big Data Processing with Hadoop-MapReduce in Cloud Systems," in International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 2, no. 1, pp. 16-27, 2013.

[21] Apache Hadoop NextGen MapReduce (YARN), http://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html, Accessed on Sept. 20, 2015.

[22] The Hadoop Ecosystem Table, https://hadoopecosystemtable.github.io/, Accessed on Sept. 20, 2015.

[23] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes and R. E. Gruber, "Bigtable: A distributed storage system for structured data," in ACM Transactions on Computer Systems (TOCS), vol. 26, no.2, (2008): 4.

[24] Apache HBase, http://hbase.apache.org/

[25] Apache Cassandra, http://cassandra.apache.org/

[26] Apache HIVE, http://hive.apache.org/

[27] Apache Pig, http://pig.apache.org/

[28] Apache TEZ, http://tez.apache.org/

[29] Apache Chukwa, http://chukwa.apache.org/

[30] Apache Zookeeper, http://zookeeper.apache.org/

[31] Apache Ambari, http://ambari.apache.org/

[32] Apache Avro, http://avro.apache.org/docs/current/

[33] Apache Mahout, http://mahout.apache.org/

[34] Apache Spark, http://spark.apache.org/

[35] Cloudera CDH, http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html

[36] HortonWorks, http://hortonworks.com/

[37] MapR, https://www.mapr.com/

[38] The Intel Distribution for Apache Hadoop Software, http://www.intel.in/content/www/in/en/big-data/big-data-intel-distribution-for-apache-hadoop.html

[39] Wen-Chung Shih, Shian-Shyong Tseng and Chao-Tung Yang, "Performance study of parallel programming on cloud computing environments using mapreduce," in Porc. International conference on Information Science and Applications (ICISA), IEEE, 2010.

[40] Maryam Kontagora, and Horacio Gonzalez-Velez, "Benchmarking a MapReduce environment on a full virtualisation platform," in Proc. International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), IEEE, 2010.

[41] Google Cloud Platform, https://cloud.google.com/solutions/hadoop/

[42] Running Hadoop in Cloud, http://www.ibmbigdatahub.com/blog/running-hadoop-cloud

[43] MapR in Cloud, https://www.mapr.com/products/hadoop-as-a-service

[44] Amazon EMR, https://aws.amazon.com/elasticmapreduce/

[45] HDInsight, http://azure.microsoft.com/en-in/services/hdinsight/

[46] Virtual Hadoop, https://wiki.apache.org/hadoop/Virtual%20Hadoop

[47] 8 big trends in big data analytics, http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html, Accessed on Sept. 20, 2015.

[48] YARN to Spin Hadoop into Big Data Operating System, http://www.datanami.com/2013/05/28/yarn_to_spin_hadoop_into_a_big_data_operating_system_/, Accessed on Sept. 20, 2015.

[49] Yong Yu, Yi Mu and Giuseppe Ateniese, "Recent advances in security and privacy in big data," (2015): 365.

[50] A. C. Mora et. al, "Top ten big data security and privacy challenges," Cloud Security Alliance (2012).

[51] InfoWorld, http://www.infoworld.com/article/2613587/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win.html, Accessed on Sept. 20, 2015.

[52] HP Vertica, https://www.vertica.com/

[53] The R Project for Statistical Computing, https://www.r-project.org/

[54] Stata: Data Analysis and Statistical Software, http://www.stata.com/

[55] Babak Akhgar, G. B. Saathoff, H. R. Arabnia, R. Hill, A. Staniforth and P. S. Bayerl, "Application of Big Data for National Security," 1st Edition, Elsevier Store, 2015.

[56] Ten Practical Big Data Benefits, http://datascienceseries.com/stories/ten-practical-big-data-benefits, Accessed on Sept. 20, 2015.

[57] CERN Computing, http://home.web.cern.ch/about/computing

[58] Bernard Marr, The Awesome Ways Big Data Is Used Today To Change Our World, https://www.linkedin.com/pulse/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world#notifications, Accessed on Sept. 20, 2015.

[59] IBM Social Media Analytics, http://www-01.ibm.com/software/analytics/solutions/customer-analytics/social-media-analytics/, Accessed on Sept. 20, 2015.