# A Comprehensive Overview of BIG DATA Technologies – A Survey

Muhammad Umair Raza
Southwest University of Science and Technology, Mianyang, P.R. 621010. China.
umair2007pak@gmail.com

Zhao XuJian
Southwest University of Science and Technology, Mianyang, P.R. 621010. China.
jasonzhaoxj@gmail.com

## ABSTRACT
In as much as the approaches of the new revolution, machines including transmission media like social media sites, nowadays quantity of data swell hastily. So, size is the core and only facet that leaps the mention of **BIG DATA**. In this article, an effort to touch a comprehensive view of big data technologies, because of the swift evolution of data by an industry trying the academic press to catch up. This paper also offers a unified explanation of big data as well as the analytics methods. A practical discriminate characteristic of this paper is core analytics associated with unstructured data which is more than 90% of big data. To deal with complicated Big Data problems, great work has been done. This paper analyzes contemporary Big Data technologies. Therein article further strengthens the necessity to formulate new tools for analytics. It bestows not sole an intercontinental overview of big data techniques even though the valuation according to big data Hadoop Ecosystem. It classifies and debates the main technologies feature, challenges, and usage as well.

## CCS Concepts
**• Information systems→Information systems applications**

## Keyword
Big Data Technology; Apache Hadoop; HDFS, MapReduce

## 1. INTRODUCTION
In this article, the basic concepts belong to big data technologies. The unexpected data increment has left numerously improvised. There is a fast evaluation of the data's quantity but on the other hand, willing to accept the concept of both public and private sectors as well. The binding of big data discourse to a more common outlet shows that there remains a clear knowledge of perception and their terminology[4]. For instance, the primary question is how data reached as **BIG DATA**? Thus, the cost of big data ideas and techniques needs to be documented in the academic press. Nowadays, the systematic generation of huge volume data from varying roots such as (Black box storage data, Social media, stock exchange, etc.). Earlier to the revolution of big data, organizations couldn't gather theirs archive for lengthy eras not proficiently accomplish huge data set. Traditional equipment had inadequate storage capacity. In the context, of Big Data scalability, flexibility and performance must be needed. Indeed, management of big data needs important resources, innovative methods, and technologies. On the other hand, big data required to sterilized, processing, secure, as well as provide grainy access to vast evolve data sets[30].

As the outcome of modified big data projects worldwide and dissimilar big data models, fresh technologies, the context has been developed to impart further storage, and real-time analysis and parallel processing from varied references. Meanwhile, the latest solutions for data security and privacy have evolved. Besides, due to the sustainable technological advancement, cost value of hardware storage and processing solution is incessantly descending.

To study big data different software and hardware technologies are build. The endeavor to verify the more authentic result of big data's applications. However, it may be time taken and effortful to choose among techniques in such surroundings. There are a lot of big data surveys but most of them tend to core on algorithms and manner used to storage and processing of big data than technologies.

In this paper comprehensively we talk about big data technologies. We classify and profoundly differentiate them not only according to their storage, processing, challenges, and features as well. This conception helps to comprehend the links among various big data technologies as well as functionalities.



*ICBDC 2020*, May 28–30, 2020, Chengdu, China

ACM ISBN 978-1-4503-7547-4/20/05…$15.00
DOI: https://doi.org/10.1145/3404687.3404694

## 2. PAST WORK ON BIG DATA/ REVIEW OF LITERATURE
Term big data explain an immense surge data sets that involve heterogeneous composition such as shaped data (Relational data etc.), semi-shaped (XML data) and unshaped data (Pdf, Text, Media log, etc.). The big data has a multiplex nature that really required more mighty technologies as compared to customary databases. So, in the case of big data solicitation, the standard static business intelligence techniques cannot be more efficient.

### 2.1 The Majority of Data Experts and Scientist They Describe Big Data by Some Characteristics
**Volume:** The massive volume of digital data is interminable generated from millions of computers and billions of applications such as (smartphones data, barcodes data, social media data, sensors, etc.). As stated by [46] it is approximated that 2.5 exabytes were produced in 24 hours in 2012. On the

other hand, this amount became double in 2013. In 2013 the international data corporation appraise 4.4 Zettabytes (ZB) of all digital data produced a duplicate and consumed. According to the perception of IDC the size of data will ascend to nearly 40 Zettabytes in 2020 and increase of 400 times by now.

**Velocity:** Rapidly data are generated from multiple ways and should process rapidly to obtain handy information. For Instance, more than 2.5 petabytes data are generated in each hour due to customer's transactions in Walmart (an international discount retail chain). It is good to say YouTube and Facebook also creators of Big Data.

**Variety:** Big data created in a distinct structure such as (audios, videos, comments, documents) by several distributed reservoirs. Massive data sets consist of structured, unstructured and semi-structured, which is maybe (public/personal, confined/distant, split/sensitive concluded/uncompleted, etc.)[52]. [31]the inclusion of previous defined V's some other dimensions are also mentioned below:

**Veracity:** IBM proclaimed a fourth, V, veracity that represents some information reasons for the unreliability inherent in it such as, customer thinking, social media uncertainty, since they involve mortal verdict, as yet they held worthy particulars. The necessity to conclude pacts along indefinite data therefore, another element of big data which tend to use tools and investigate to the purpose of collecting dubious information.

**Variability:** In the context of big data variability about a few different things. That's the amount of data incompatibility[15]. They must be established in ways that meaningful analysis can start with anomaly detection techniques.

**Validity:** In contrast to truthfulness, the validity indicates the correctness of data for the intended use. As stated by Forbes, there are more than 60% of data scientists spend time cleaning data before to analyze, the advantage of big data analytics is good as its rudimentary data, so excellent data governance practice is needed to guarantee data quality and shared definitions with metadata[26].

**Vulnerability:** New safety issues arise from data. After all, data breaches with big data constitute a major violation. Sadly, there are so many violations of big data that have occurred. In May 2016 a hacker called "peace posted data for sale on the dark web, which was alleged to have approximately 167 million LinkedIn accounts and 360 million E-mails and password for Myspace users".

**Volatility:** For consideration of volatility, we necessity to study the volume, variety, and velocity. Volatility refers to that data should be stored for how long. Within this word, we required that we managed that which point and when data are not relevant as well as to the modern analysis. Because of the volume, velocity, and variety, it's very obligatory to understand volatility. For some reasons, the identical data will always be there but sometimes for others, this is not will be the case[39].

**Visualization:** In the tools of big data visualization, confront some technological difficulties by some restrictions of in-memory technology and imperfect scalability, functionality and response time as well. When attempting to trace thousands of datasets, we cannot use traditional graphs and we need several distinct methods to represent the data like data clustering or three maps, sunburns, parallel coordination as well as circular network diagrams.

**Value:** value is the most salient characteristic of big data technology. The worth in the future of Big Data is enormous. It has worthy access to big data. It is very expensive to implement IT infrastructure systems for storing huge data[35].

## 2.2 Application of Big Data

Big data techniques are widely and extensively indexed. It is used for such purposes as a search engine, transportation & logistics, Data storage, videos & pictures analysis, Telecommunications, Web & Social Media, Medicines & Healthcare, Science & Research as well as Social Life. Few of them eminent applications are been discussed below[55].

**Transportation and logistics:** Publically operating carriers use RFID and GPS to track buses furthermore, to search the use of fascinating data to enhance their facilities. For example, to optimize the bus paths and the oftenness of journeys, the data collection on the number of travelers on buses in some different routes. Data Mining also helps to improve the business traveling by forecasting the public and private networks' demand[58]. For illustration, India has one of the busiest railway system, every single day nearly 250,000 seats are reserved and reservation can be done by almost 60 days in advance. To prediction about such data is a problem because of it up to some factors like the festival, weekend, etc. By the using of machine learning techniques, it's viable to mine and put on advance analysis on the previous as well as new big data technologies.

**Healthcare and Medicine:** Big Data technologies are helpful for storing of the medical record. Data can be captured from multiple sensors and equipment's that are devoted to patients and it also generates from heterogeneous sources like (Laboratory and clinical data, hospital operations and pharmaceutical data)[50][77]. The medical data set has numerous beneficial applications because the healthcare data is proficiently suited for big data processes and analytics. Recently, in several areas in healthcare have been related that can be frankly beneficial from such treatment[49][47].

**Social Media Analysis*:*** IBM introduces a spiritual analysis, to find out the invisible perceptivity from millions of web sources. It is used by a company to pick up superior understanding and calculation their clients. It catches shoppers' information from web-based life that predicts client behavior and warfare[66][18].

**Science and Research:** Science and analysis are currently compulsive by technologies. New prospectus added by Big data. [32]the sizeable and most strong practical accelerator, Large Hadron Collider (LHC) has been launched by CERN, (European Organization for Nuclear Research). Unrestrained information was produced by the Experiment. The data center of CERN has 65,000 processors analyzing 30 petabytes of information. Its computing power is spread by thousands of pcs across 150 data centers around the world.

**Politics Analysis*:*** Big data analytics helps in winning the US Presidential election by Mr. Barack Obama in 2012[73]. His strive consisted of 100 worthy analytics members to shake heaps of terabytes of data. For analytical databases, a coalition of HP Vertica is used massively parallel.

# 3. BIG DATA-APACHE HADOOP AND MAPREDUCE (THE ARCHITECTURE OF BIG DATA TECHNOLOGY)

Hadoop (Highly Archived Distributed Object-Oriented Programming) developed in 2005 by Mr. Doug Cutting and Mike Cafarella. The name of Hadoop was selected by Doug Cutting as it was the name of his son's toy elephant. It is an open-source software system that makes it reliable as well as scalable and also provides distributed computing to organizations [45]. This software knobs enormous amounts of multiple types of data from distinct sources such as pictures, videos, audios, folders, software sensor recording and communication data as well[43]. Hadoop's primary benefit is its ability to quickly process to big data set. In reality, unlikely in the traditional way, Hadoop doesn't copy the entire separate data into memory to performs computations. For instance, even the terabytes of data just take the Nano-seconds to query in Hadoop. Further superiority of Hadoop is the capability to work during the time that ensuring the fault tolerance, normally found in distributed surroundings[52][67][76].

The capability of Hadoop is standing on two major components: (i)Hadoop Distributed File System (HDFS) (ii)MapReduce (MR)[33]. Moreover, users can also build modules on the top of Hadoop, according to application requirements and their objectives. These modules are said to be a Hadoop ecosystem.

## 3.1 Hadoop Distributed File System(HDFS*)*

To store data in HDFS depends on its file system and a database(non-relational) called Hbase. HDFS entirely files oriented system which creates high performance and efficient access to data to run on commodity hardware. It has numerous replicas to easily get data and swiftly return to the user[63]. One of the main reasons for building these imitations is to offer the accessibility throughout and if some node fails to perform but nothing should be stopped. Simply, in Hadoop each block data must be replicas itself[5][74]. There are five major components of HDFS called; (i) Name node (ii) Data node (iii) Seconder Name node (iv) Job Tracker (v) Task Tracker.

### *3.1.1 Name node*

Name Node considering as the core of HDFS file system as it contains metadata information about the data of the user. while the read operation it doesn't stock physical data but it keeps all pertinent facts and figures which are essential to amalgamate the split data during the reading[33]. Hadoop cluster availability is extremely dependent on the Name Node as all the information of metadata is present only on the Name Node. On Name Node server each file and folder is portrayed as iNode consisting of processed data such as the moment of file access, amendment, authorization on file/directory and file block size, etc. The client HDFS first contacts Name Node to collect appropriate iNode information while performing read operations, and then accesses all the information nodes to acquire the actual user data. Name node is also called single point of failure.

### *3.1.2 Data node*

Data nodes in Hadoop are primarily accountable for the creation, replication, and delectation of the data file. Huge data files broken first into tiny blocks on the Name Node, and then store into the selected Data Node. Name Node tracks all the information of metadata partitioned blocks stored on data nodes[72]. Formerly data save successfully in Data Node after that it replicates on more than one backup nodes which already available in HDFS client. If there is a collapse in HDFS client to obtain file block from primary Data Node either because of Data Node is much busy to serving other clients or it is down, then it will contact to corresponding backup data node to retrieve data. Here's foremost remember the Name Node cannot directly communicate to Data Node, but via pulses that the Data Node regularly sends to Name Node.

### *3.1.3 Secondary name node*

This node used to help of the master node. when the name node performs some actions it creates a checkpoint and saves in secondary name node. Meanwhile, if the master node is dead or maybe create a problem, restart that node and pings its secondary name node to gather checkpoint to get the prior state. There is a great degree of fault tolerance by secondary name nodes[29].

### *3.1.4 Job tracker*

The job tracker speaks to the Name node to adjudicate where the data is located. The Job Tracker schedules decrease the intermediate fusion or action of individual maps. It monitors how these individual tasks have succeeded and failed. It operates to complete the whole task as well. If a job is not done, the Job Tracker restarts the task automatically, but probably at another node, to a predefined retries limit[15].

### *3.1.5 Task tracker*

The Job Tracker supervises the general execution of a MapReduce job scheduling. On each slave node, the Task Trackers handle the execution of individual scheduling. Even though, the slave node contains a single Task Tracker. The Java Virtual Machines (JVMs) can be created by each Task Tracker to handle several maps or reduce the parallel allocation. In every short time, Task Trackers also send messages to the Job Tracker, to reassure Job Tracker is still alive[64].

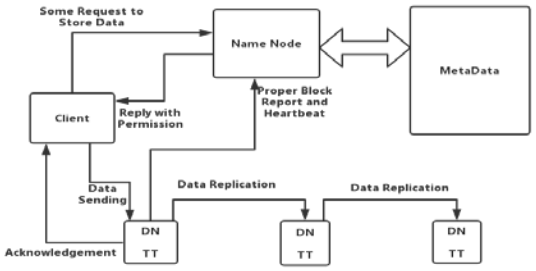


**Figure 1. Data storage in HDFS.**

Client send some request to Name Node for data storing, Name Node give proper response with permission to client. Data Node accept data from client with acknowledgement, Data Node store data and have 2 others data replication Node and Data Node send proper block report as well as Heartbeat in every short period of time to Name Node. Actually name node play a vital role it is also a single point of failure. Metadata stores all information about storage.

## 3.2 HBase

Hbase is a completely non-relational, open-source, distributed Hadoop based database. It intended exclusively for execution with low latency. Hbase is key/value pair column-oriented database[57]. It can pillar aloft table update rates, also in distributed clusters horizontally. Furthermore, it offers a flexible layout, for large tables just like BigTable format[8]. Logically data store in table format. The benefit of such tables is that

millions of rows and columns can be processed. Hbase tables are known as Hstore.

Hbase, offer numerous characteristics such as real-time queries, natural language searching, linear, modular, automatic and configurable access to table sharing[28]. It is included on many data-driven sites, just like Facebook messaging platform[2].

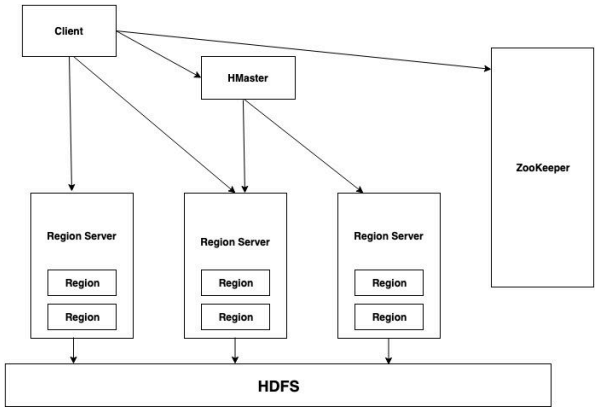


**Figure 2. Architecture of Hbase.**

Hbase architecture has 3 main components: HMaster, Region Server and Zookeeper.

**HMaster:** HBase's Master server implementation is HMaster. It is a process in which regions are allocated to server region as well as operations with DDL (Create, Delete table). It tracks all Instance of region Server present in cluster. In a distributed system, Master runs multiple background threads. HMaster has several advantages such as load balance controlling and failover etc.

**Region Server:** HBase Tables are divided into regions, horizontally by row key Selection. **Regions** are the basic building elements of the HBase cluster consisting of distribution and consisting of Column groups. Area Server operates on the HDFS Data Node located in the Hadoop cluster. Area Server Regions are responsible for multiple things, such as handling, controlling, executing as well as reading and writing HBase operations on that group of regions. A Region has default size of 256 MB.

**Zookeeper**: It is like being a Hbase leader. It provides services such as keeping information about the configuration, naming, providing distributed synchronization, notification of server failure etc. Using zookeeper, clients communicate with region servers.

## 3.3 MapReduce (MR)

The MapReduce has become omnipresent for the processing of large scale data. This application of Hadoop open source is widely accepted by organizations ranging from a two-person start-up to fortuity 500 companies[1]. It reclines at the core of a developing stack for data analytics, that supports heavyweight industries such as IBM, Microsoft, and Oracle, etc. one of the MapReduce advantages is the capacity to horizontally scale to high volume of data on thousands of commodity servers, easy-to-understand semantics for programming, and high rate of fault tolerance[41]. It is the primary crucial step for the upcoming generation to management and analysis tools for big data. MapReduce has captivating advantages for big data applications. As a matter of fact, it makes simple the gigantic size of data by its effective and cost-efficient mechanism. It enables to write, so the parallel processing is possible[52][59]. In reality, the MapReduce programming model utilizes two following features: The Map function and the Reduce function, to handle processing[34].

First of all, the map function splits input data into maverick data partitions representing pairs of key/value.

Then, through several parallel map tasks, the MapReduce framework sends all the key/value pairs independently to mapper across this cluster. The mapper produces may be multiple intermediate key/value pairs. At this level, the substructure responsible for collecting and sorting all the intermediate key/value pairs. Therefore, there are multiple keys which have the list of related values.

Now, the reduction function exerts to process the whole output of the intermediate data. The reduction function adds the key values according to the pre-defined program for every single key. (i.e., filter, summarize, sorting, hashing, take the average of maximum). Then one or more key-value pairs will be generated[20].

Finally, MapReduce stores all output (key/value) pairs within the output folder smoothly.

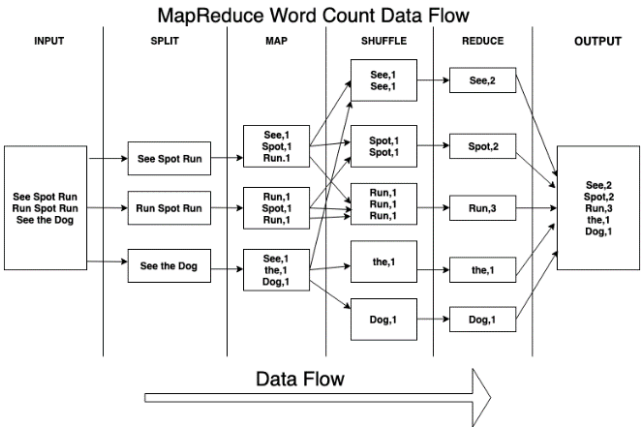


**Figure 3. Workflow/Architecture of MapReduce.**

## 3.4 Yarn

Than MapReduce, Yarn has been genetically modified. As compare to MapReduce it provides more scalability, parallelism as well as improves the management of resources. It also provides features of the operating system of big data analytics. The YARY resource manager has changed the Hadoop architecture. In general, YARN operates on the top of HDFS. This position enables different applications to be carried out in parallel[33]. It also allows the bath as well as interactive processing to be handled in real-time.

In contrast to MapReduce, YARN improves effectiveness by partitioning the Job Tracker's two primary functionalities into two different daemons[42]: (1) Resource-Manager (RM) apportion and regulates the cluster's resources. (2) Application-Master (AM) is planning to, match and monitor their process with TaskTracker[80].

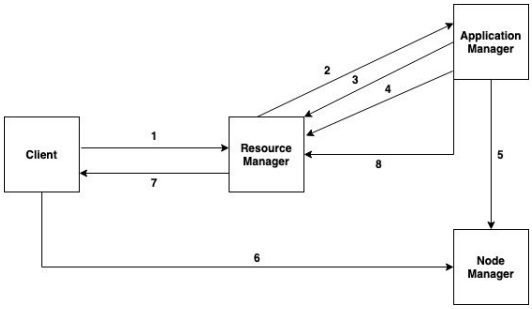


**Figure 4. Workflow/Architecture of Hadoop Yarn.**

1. Client send an application.
2. The resource manager assigns the program manager to start a container.
3. With the resource manager The program manager register itself.
4. The program manager negotiates containers from the resource manager.
5. The application notifies node manager that containers should be released.
6. Application code within the container is executed.
7. Clients contacts resource manager to monitor the status of an application.
8. Upon the completion of the processing the application manager un-registers with the resource manager.

# 4. HADOOP ECO-SYSTEM

Apache Software Foundation is supporting various other projects associated with Hadoop. A specific aspect of big data is addressed in each project and Hadoop provides supplementary services. The projects associated with Hadoop is said to be Hadoop Eco-System[75]. The description is below;

**Cassandra:** It is a scalable database that offers elevated availability as well as supports multi-master to avoid solitary points of failure. MapReduce can recoup data from Cassandra. It is a Big Data, Database, which can flee without HDFS. It supported by both Google Big Table and Google File System as well[22][7].

**Hive:** It's an infrastructure for the data storehouse that offers data summarization, ad-hoc querying and HDFS-based analysis of huge datasets[54]. It also provides structural design for this information and also a HiveQL based on SQL. It also offers flexibility for customizing mappers and reducers, if logic cannot be expressed efficiently in HiveQL[9][36].

**Pig:** Pig is a high-level programing language as well as a parallel execution framework. A program that is written in Pig able to manage large datasets through significant parallelism. The basic infrastructure of Pig comprises of a compiler which is a factor of production MapReduce sequences with parallel implementations. Pig's language, Latin express sequences, and users can also build up their function to read, write and processing for data[79].

**Tez:** It's a broad-based information stream programming framework, which is built on the top of Hadoop YARN. It offers a strong and versatile engine to perform a complicated DAG (directed acyclic graph) tasks for batch or interactive processing. It increases the power of MapReduce by expressing computations in the data flow graph. Tez adopted by Hive, Pig, and other eco-system members to substitute MapReduce job[12].

**Chukwa:** It is a mechanism of data collecting to monitoring large distributed clusters. It constructs on the top of HDFS & MapReduce to offer large-size logging and analytics. It has a pliable and strong toolbox to showing, monitoring and analyzing the outcomes on the collected data[24].

**Zookeeper:** The coordination between distributed applications provided by Zookeeper. Several projects of Hadoop use the Zookeeper for coordinating the cluster and provide distributed facilities that are extremely accessible. It provides a centralized service for maintenance, providing distributed synchronization and community services[13].

**Ambari:** It provides a step-by-step wizard with the Hadoop cluster to install services, for example, Hive, Hbase, Pig and Zookeeper, etc. To simplify Hadoop management as well as the Hadoop cluster, Ambari is a web-based tool that also handles services. It provides key management for Hadoop services to begin, stop and reset over the cluster. It controls the current status of the Hadoop cluster[6].

**Avro:** This scheme for serializing the information. It has wealthy data structures. It offers compact and binary data format for storing persistent data and remote procedure call (RPC). Code generation is not required for reading, writing data and nor to use RPC protocols[78].

**Mahout:** Mahout at the top of the MapReduce machine learning, data mining and math library. This project aims to offer scalable and rapidly machine learning and data mining algorithm[10].

**Spark:** It is a quick and general data processing engine. It provides an easier alternative to the use of MapReduce and runs programs up to 100 times quicker than MapReduce. It is a sophisticated directed acyclic graph (DAG), which allows quickly in-memory computation and cyclic data flow. Spark is running on Hadoop and can access HDFS, Hbase, and Cassandra[11].

**Sqoop:** A project intended to efficiently transfer bulk data between Hadoop and structured databases[69].

**Oozie:** Oozie is an Apache Hadoop workflow scheduler scheme. The Directed Acyclical Graph (DAGs) of actions operates in flow employment. Oozie is incorporated into the remainder of the Hadoop pile which supports several different kinds of Hadoop tasks as well as system specific jobs (e.g., Java program and shell scripts)[51].

# 5. HADOOP DISTRIBUTION

Different IT providers and communities are enhancing Hadoop infrastructure, tools, and structure. It is useful for big data technologies to share revolution through open-source modules. Anyway, it's a pitfall users can wind up with a Hadoop platform consisting of separate module from distinct sources[52]. There is a specific level of maturity for each module, a variant in the Hadoop platform is at danger of being incompatible. The integration of different techniques on a single platform also increases the same peril. Usually, every module is appraised. Even though, the multi-source coalition can mostly have concealed threats that are not fully researched nor tested.

Many IT vendors, such as IBM, Cloudera, MapR, and Hortonworks, initiate their modules and packaged them into distributions to deals with these matters.

## 5. 1 InfoSphere BigInsights-IBM

It's beginning to simplify the utilization of Hadoop. It can meet company requirements for storage, processing, advance evaluation, and visualization. The fundamental versions of IBM InfoSphere is HDFS, Hbase, MapReduce, Hive, Mahout, Oozie, Pig, Zookeeper, etc., have been released now.

Enterprise Edition provides some additional principal services: reliability features, performance capabilities, security management and optimization of fault-tolerance. It encourages sophisticated big data analytics with adaptive algorithms such as (Text processing). IBM also offers layers of data access that can also be attached to distinct sources of data (like DB2, streams, data Stage, JDBC, etc.)[44]. There are some other benefits of IBM distribution: first, the possibility of storing data streaming to BigInsights clusters directly. Second, it promotes real-time analysis data streaming as well as facilitates visualization via dashboards and big sheets in the cluster.

## 5. 2 Cloudera

Cloudera is a Hadoop distribution that is most commonly used. It allows Hadoop to deploy and manage an Enterprise Hub[56]. It offers numerous advantages including centralized management tools, unified batch processing, an interactive SQL and role-based access control[16]. IMPALA is one of the principal Cloudera module[62]. It is an interesting Hadoop compatible query language module[30]. Impala structure data on a column-based shape. It enables synergistic and real-time analysis of big data managed. Contrarily Hive, MapReduce framework doesn't use by Impala. Alternatively, it also utilizes an individual in-memory processing mechanism for quick queries over the massive amount of data. Hence, Impala is quicker as compare to Hive while fetching the query. Indeed, Impala can candidly use data from current HDFS and Hbase sources.

Cloudera also has a versatile model that is quicker than Hive, supporting both structured and unstructured data. For Example, Cloudera is 10 times more quickly than Hive and MapReduce. Cloudera confirms that approximately 5 to 47 times its performance dividend for request with at least a single join as compare to HiveQL (Hive Query Language)[57].

Although, Cloudera has some disadvantages. Such as, it's not perfect for querying streaming data (e.g., videos or uninterrupted sensor data). All join activities shall be conducted in memory restricted by the cluster's limited memory node[23].

## 5. 3 MapR

MapR is an enterprise-designed business distribution for Hadoop. The precision, efficiency, and easiness of Big Data storing, processing as well as evaluation with machine learning algorithms have been improved. It offers a broad range of components and projects to the Hadoop environment[52]. It doesn't use HDFS. Although, it generates its MapR file systems (MapR-FS) which enable simple backups to enhance the performance. The benefit of MapR-FS is NFS compatible. MapR is based on Hadoop's current programming model.

## 5. 4 Hortonworks Data Platform(Hdp)

The HDP is erect on Hadoop to the storing, querying as well as processing. It is a quick, scalable and cost-effective solution. It offers multiple management, surveillance, and integration. Furthermore, HDP offers open-source, managing instruments also promotes links with certain BI platforms[16].

# 6. CHALLENGES OF BIG DATA

Big data provides numerous appealing possibilities. Moreover, practitioners and researchers face various difficulties to explore big data sets[55]. The problem occurs at various stages of data management such as data collection, storage, search, etc.,[14]. In distributed data-driven applications, there are some security and privacy issues as well[60].

**Heterogeneousness and rawness:** The big data analytics face some difficulties from its huge size also with the presence of varied data on divergent shapes. There are several models with very distinct characteristics for complex heterogeneous mixture data, there are numerous patterns that have very different properties. Data may be both structured and unstructured. More than 80% of the data produced unstructured by organizations. It is extremely dynamic and has no particular format. It may be the multi-shaped (e.g., images, pdf documents, medical records, video, audio, etc.). Transforming this data into a structured form is a vital challenge in the mining of big data. So the latest technologies have to be adopted to deal with such kind of data[37].

**Scalability and complexity:** Management of huge and speedily expandable data is a series challenge. To manage increasing data volumes cannot be carried by traditional data management techniques. The scalability and complexity of big data to be analyzed are also major obstacles to data analysis[48].

**Big data storing and quality:** Storage and analysis huge amount of data is pivotal for a corporation to work need an extensive and multiplex hardware infrastructure. Data storage devices are becoming more and more essential with consistent data development and many companies are looking forward to high storage capacity to compete with this issue[17]. For the decision-making, accuracy and on-time availability of data are essential. Big data is at most sympathetic when an information management process is implemented to guarantee data accuracy and quality.

**Big data cleaning:** In the case of traditional databases, the following steps (Cleansing, Aggregation, Encoding, Storage and Access) are not emerging. There is a challenge to manage the processing and complex structure of Big Data in a distributed environment with the combination (Velocity, Volume, and Variety)[38]. The dependability of the source and nature of data must be verified before using resources to reliable outcomes. The problem is purifying such amount of data sets and choose which data set is accurate and helpful.

# 7. SECURITY AND PRIVACY IN BIG DATA

The organizations need to securely process and regulations to assurance their framework. For Big Data security and privacy issues, accustomed techniques are considered as ineffective[61]. However, new techniques are also hosted to unidentified back doors and default credentials[60]. It is necessary to the consideration of confidentiality, integrity, and availability of data.

**Security:** Miscellany of data source, formats, streaming as well as infrastructures might cause unique vulnerabilities to safety. The Cloud Security federation has broken down the challenges of safety and privacy of big data into distinct classifications; security of infrastructure, data protection, data management, integration and reactive security[65]. The Infrastructure of

security comprises of safe and secure distributed programming. The data security concerns to analytics, encrypted and grainy access control data centers. Data management involves secure data storage, processing, logging, auditing and data provenance[81]. Furthermore, validation, filtration, and real-time monitoring include integrity and reactive safety. Based on suggested issues, the authorization and authentication mechanisms of users are crucial also encoding and data masking are essential to implement for both states of data (rest and stream).

**Privacy:** The development of systems has led to independent collection control[25]. Recently, the National Security Agency (NSA) under the cover of defending US citizens has been wiretapping personal data from miscellaneous sources like databases of vast companies, cyberspace, and telecom companies. The eternally increasing the secrecy concerns about big data including knowing the latest and secret actuality about people, amalgamating their private details, including value their institutions with collected data from unknowing persons, threatening uneducated people by prognostic analysis by social media, finally exchanging datasets between the organizations[19]. In response to such complex matters, rules and regulations must have been clear limits for unauthorized access, data sharing, illegal use, and also duplication of personal information[60][68].

## 8. WHAT SHOULD BE HAPPEN IN FUTURE?

There are several important challenges for the future in management of Big Data technologies that arise from the nature of data such as complexity, diversity, and evolving. In the next years, researchers will have to face several difficulties in various areas.

**In medical science:** Today, the healthcare system is on an unsustainable trajectory. The volume of costs in the current system is because of the patient's having continuing diseases. Therefore, preventive care, as well as population health control, should be a priority in the future[71]. Big Data makes easier for understanding. In the future of the healthcare sector, Personalized medicine is being promoted. Nowadays, the production of medicines is for the masses not for the independent. Looking forward, with the advent of Big Data applications, further, customize medicines that use patient specifically data just like genomics and proteomics can be generated which is based on the describing of similar patients and their responses to such approaches. Social media and mobility are increasingly adopted, patients are adopted more and more aware of the alternatives accessible to them. In the future, we expect the development of new data sources and analytical technologies to change the way we practice medicine[53].

**In social media:** The term "Social media" is a wide range of online platforms for creating and exchanging content for the user. Social media classified into the following types such as Social networks (e.g., Facebook, LinkedIn, Twitter, Tumbler, Instagram, YouTube)[18] as well as some mobile apps. The research about social media analytics extends to a number of several directions including, psychology, sociology, computer science, mathematics, physics, and economics. In social media specifically, we need to enhance the predicting the future linkages between the existing nodes that underlying network. Normally, social networks structures are not static and they continuously expand. Wherefore, it is a natural objective to realize and forecast the dynamics of the network[27].

**In IoT:** Because of the rapid enlargement of IoT based applications in the cloud, the number of connected devices is increasing swiftly[55][40]. The expectation is that connected devices will be reached to 24 billion in 2020. These devices will be connected via the cloud for different kinds of applications. IoT and cloud computing work on the integration that makes a new prototype, which has been designated as a cloud of things (CoT). In CoT, the objects of IoT are expanded through the internet from sensors to all front-end objects. Furthermore, the distributed sites are attached as the entire body, just like as smart houses, smart factories, smart cities, as well as the smart planet. A logical design of the smart city is provided Based on CoT[70]. By combining the cloud platform and IoT, CoT needed to enhance the interactive and interoperability capability of smart applications. In divergent industries and research areas, CoT will take a progressively important role. There are some problems such as resource distribution that stabilize energy and efficiency, the standard of service provisioning, storage of data architecture, security, privacy and unnecessary communication of data will be associated in CoT[82][21][3].

## 9. CONCLUSION

The intention of this article to delineate, evaluation, and review of big data technologies. Firstly, this article described, what is big data means and to consolidate the divergent discourse on big data. In this article, we present varied definitions of big data, which underlying the fact that size is only one facet of big data. On the other hand, some other dimensions, such as Velocity and Variety are also foremost. The paper's mainly focused on analytics in order to gain viable and precious insights from big data. Big data is applied in almost every area ranging from the financial sector to in healthcare sector. Big Data can be handled by the implementation of several techniques. However, there is still scope for further research because of the problems of storage, processing, and management are surrounded by great issues in a broad classification. The magnitude of Data has been generated every minute which is may be structured, unstructured as well as semi-structured that need sufficient storage. Furthermore, the issues which are related to the fast-growing data but the result is still concerning and management issues related to Big Data are also still under consideration for future studies.

## 10. ACKNOWLEDGEMENTS

This research was funded by Humanities and Social Sciences Foundation of the Ministry of Education, grant number 17YJCZH260 and CERENT innovation Project, grant number NGII20180403. The authors would like to specially thanks to loving family who support in every time as well as all friends and lab members.

## 11. REFERENCE

[1] 6th Symposium on Operating Systems Design and Implementation — Technical Paper: *https://www.usenix.org/legacy/event/osdi04/tech/full_papers/dean/dean_html/*. Accessed: 2019-08-01.

[2] Aiyer, A. et al. 2012. Storage Infrastructure Behind Facebook Messages. *IEEE Data Engineering*. (2012), 1–10.

[3] Al-fuqaha, A. et al. 2015. Internet of Things : A Survey on Enabling. *IEEE Communications Surveys & Tutorials*. 17, 4 (2015), 2347–2376.

DOI:https://doi.org/10.1109/COMST.2015.2444095.

[4] Al-Sai, Z.A. et al. 2019. Big Data Impacts and Challenges: A Review. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*. (2019), 150–155. DOI:https://doi.org/10.1109/JEEIT.2019.8717484.

[5] Alam, A. and Ahmed, J. 2014. Hadoop Architecture and Its Issues. (2014). DOI:https://doi.org/10.1109/CSCI.2014.140.

[6] Ambari -: *http://ambari.apache.org/*. Accessed: 2019-08-02.

[7] Apache Cassandra: *http://cassandra.apache.org/*. Accessed: 2019-08-01.

[8] Apache HBase – Apache HBase™ Home: *http://hbase.apache.org/*. Accessed: 2019-07-31.

[9] Apache Hive TM: *http://hive.apache.org/*. Accessed: 2019-08-02.

[10] Apache Mahout: *http://mahout.apache.org/*. Accessed: 2019-08-02.

[11] Apache Spark™ - Unified Analytics Engine for Big Data: *http://spark.apache.org/*. Accessed: 2019-08-02.

[12] Apache Tez – Welcome to Apache TEZ®: *http://tez.apache.org/*. Accessed: 2019-08-02.

[13] Apache ZooKeeper: *http://zookeeper.apache.org/*. Accessed: 2019-08-02.

[14]Ardagna, C.A. et al. 2016. Big Data Analytics as-a-Service : Issues and challenges. (2016), 3638–3644.

[15] Arora, Y. Big Data Technologies : Brief Overview. 131, 9, 1–6.

[16] Azarmi, B. *Scalable Big Data Architecture*.

[17] Balachandran, M. 2017. ScienceDirect ScienceDirect ScienceDirect Challenges Deploying Challenges and and Benefits Benefits of of Deploying Big Data Data Analytics Analytics in in the the Cloud Cloud for for Business Business Intelligence Intelligence Big. *Procedia Computer Science*. 112, (2017), 1112–1122. DOI:https://doi.org/10.1016/j.procs.2017.08.138.

[18] Barbier, G. Chapter 12 DATA MINING IN SOCIAL MEDIA. DOI:https://doi.org/10.1007/978-1-4419-8462-3.

[19] Bardi, M. et al. 1926. Big Data Security and Privacy: A Review. *Journal of the Chemical Society (Resumed)*. 129, 2 (1926), 663–670. DOI:https://doi.org/10.1039/JR9262900663.

[20] Braganza, A. et al. 2017. Resource management in big data initiatives : Processes and dynamic capabilities ☆ , ☆☆. *Journal of Business Research*. 70, (2017), 328–337. DOI:https://doi.org/10.1016/j.jbusres.2016.08.006.

[21] Cai, H. et al. 2017. IoT-Based Big Data Storage Systems in Cloud Computing : Perspectives and Challenges. 4, 1 (2017), 75–87.

[22] Chang, F. et al. 2006. Bigtable: A Distributed Storage System for Structured Data (Awarded Best Paper!). *Osdi*. (2006), 205–218. DOI:https://doi.org/10.1145/1365815.1365816.

[23] Chauhan, A. 2013. *Learning Cloudera Impala*.

[24] Chukwa - Welcome to Apache Chukwa: *http://chukwa.apache.org/*. Accessed: 2019-08-02.

[25] Conference, I.I. et al. 2015. Data Confidentiality Challenges in Big Data Applications. 8, (2015), 2886–2888.

[26] Dave, M. and Kamal, J. 2017. Identifying Big Data Dimensions and Structure. (2017), 163–168.

[27] Desai, P. V. 2018. A survey on big data applications and challenges. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*. Icicct (2018), 737–740. DOI:https://doi.org/10.1109/ICICCT.2018.8472999.

[28] Dimiduk, N. and Khurana, A. *HBase in Action*.

[29] Dwivedi, K. 2014. Analytical Review on Hadoop Distributed File System. (2014), 174–181.

[30] Eldawy, A. and Mokbel, M.F. 2017. The era of Big Spatial Data. *Proceedings of the VLDB Endowment*. 10, 12 (2017), 1992–1995. DOI:https://doi.org/10.14778/3137765.3137828.

[31] Gandomi, A. and Haider, M. 2015. International Journal of Information Management Beyond the hype : Big data concepts , methods , and analytics. *International Journal of Information Management*. 35, 2 (2015), 137–144. DOI:https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

[32] Hep, T. et al. 2019. *A Roadmap for HEP Software and Computing R & D for the 2020s*. Springer International Publishing.

[33] Hurwitz, J. et al. 2013. *Bir Data for Dummies*.

[34] Industry's Next Generation Data Platform for AI and Analytics | MapR: *https://mapr.com/*. Accessed: 2019-08-01.

[35] Ishwarappa and J, A. 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. 48, Iccc (2015), 319–324. DOI:https://doi.org/10.1016/j.procs.2015.04.188.

[36] Ismail, A.S. et al. Querying DBpedia Using HIVE-QL. 102–108.

[37] Jaseena, K.U. and David, J.M. 2014. ISSUES,CHALLENGES, AND SOLUTIONS: BIG DATA MINING. (2014), 131–140.

[38] Khan, N. et al. 1990. Big Data: Survey, Technologies, Opportunities, and Challenges. *Japanese Journal of Applied Physics*. 29, 8 (1990), L1497–L1499. DOI:https://doi.org/10.1143/JJAP.29.L1497.

[39] Khan, N. et al. 2018. The 10 Vs, Issues and Challenges of Big Data. March (2018), 52–56. DOI:https://doi.org/10.1145/3206157.3206166.

[40] Li, S. et al. 2018. US CR. (2018). DOI:https://doi.org/10.1016/j.jii.2018.01.005.

[41] Lin, J. 2013. MAPREDUCE IS GOOD ENOUGH ? March (2013), 28–37. DOI:https://doi.org/10.1089/big.2012.1501.

[42] Machova, R. et al. 2016. Processing of Big Educational Data in the Cloud Using Apache Hadoop. (2016), 46–49.

[43] Manwal, M. Big Data and Hadoop -A Technological Survey.

[44] Martino, B. Di et al. 2014. Big data (lost) in the cloud. *International Journal of Big Data Intelligence*. 1, 1/2 (2014), 3. DOI:https://doi.org/10.1504/ijbdi.2014.063840.

[45] Mass, C. et al. 2013. Volume 3, Issue 12, December 2013.

3, 12 (2013), 14947.

[46] Mcafee, A. and Brynjolfsson, E. 2012. Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017. *Harvard Business Review*. October (2012), 1–9.

[47] Mehta, N. and Pandit, A. 2018. Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*. 114, January (2018), 57–65. DOI:https://doi.org/10.1016/j.ijmedinf.2018.03.013.

[48] Mishra, S. 2015. Challenges in Big Data Application : A Review. 121, 19 (2015), 42–46.

[49] Mitra, A. et al. 2016. A Novel Big-Data Processing Framwork for Healthcare Applications. (2016), 3548–3555.

[50] Nambiar, R. 2019. A look at challenges and opportunities of Big Data analytics in healthcare - IEEE Conference Publication. (2019), 17–22.

[51] Oozie - Apache Oozie Workflow Scheduler for Hadoop: *http://oozie.apache.org/*. Accessed: 2019-08-03.

[52] Oussous, A. et al. 2018. Big Data technologies : A survey. *Journal of King Saud University - Computer and Information Sciences*. 30, 4 (2018), 431–448. DOI:https://doi.org/10.1016/j.jksuci.2017.06.001.

[53] Pashazadeh, A. and Navimipour, N.J. 2018. Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. *Journal of Biomedical Informatics*.

[54] Patel, D. et al. 2017. Analyzing Network Traffic Data Using Hive Queries. 3 (2017), 3–8.

[55] Philip Chen, C.L. and Zhang, C.Y. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 275, (2014), 314–347. DOI:https://doi.org/10.1016/j.ins.2014.01.015.

[56] Pol, U. 2016. International Journal of Advanced Research in Big Data and Hadoop Technology Solutions with Cloudera Manager. September (2016).

[57] Prasad, B.R. and Agarwal, S. 2016. Comparative Study of Big Data Computing and Storage Tools: A Review. *International Journal of Database Theory and Application*. 9, 1 (2016), 45–66. DOI:https://doi.org/10.14257/ijdta.2016.9.1.05.

[58] Rajaraman, V. 2016. Big Data Analytics. August (2016), 2015–2016.

[59] Ravi, V.T. Comparing Map-Reduce and FREERIDE for Data-Intensive Applications.

[60]Raza, M.U. 2017. Big Data – Security and Privacy policy. 5, 6 (2017), 51–54.

[61] Rezaeijam, M. A Survey on Security of Hadoop.

[62] Sakr, S. *Big Data 2.0 Processing Systems A Survey*.

[63] Shafer, J. et al. 2010. The Hadoop distributed filesystem: Balancing portability and performance. *ISPASS 2010 - IEEE International Symposium on Performance Analysis of Systems and Software*. March 2010 (2010), 122–133. DOI:https://doi.org/10.1109/ISPASS.2010.5452045.

[64] Shao, Y. et al. 2018. Computers & Industrial Engineering E ffi cient jobs scheduling approach for big data applications. *Computers & Industrial Engineering*. 117, March 2017 (2018), 249–261. DOI:https://doi.org/10.1016/j.cie.2018.02.006.

[65] Sinanc, D. et al. 2015. A survey on security and privacy issues in big data. December (2015). DOI:https://doi.org/10.1109/ICITST.2015.7412089.

[66] Singh, S. et al. 2015. Big Data : Technologies , Trends and Applications. 6, 5 (2015), 4633–4639.

[67] Sogodekar, M. et al. 2016. Big data analytics: Hadoop and tools. *IEEE Bombay Section Symposium 2016: Frontiers of Technology: Fuelling Prosperity of Planet and People, IBSS 2016*. (2016). DOI:https://doi.org/10.1109/IBSS.2016.7940204.

[68] Somasekaram, P. 2016. Privacy-Preserving Big Data in an In-Memory Analytics Solution. *Luleå University of Technology*. (2016).

[69] Sqoop -: *http://sqoop.apache.org/*. Accessed: 2019-08-03.

[70] Sur, S. et al. Can High-Performance Interconnects Benefit Hadoop Distributed File System ?

[71] Taguchi, Y.H. et al. 2014. Heuristic principal component analysis-based unsupervised feature extraction and its application to bioinformatics. *Big Data Analytics in Bioinformatics and Healthcare*. i, (2014), 138–162. DOI:https://doi.org/10.4018/978-1-4666-6611-5.ch007.

[72] Tech, M.R.D. 2014. Handling Big Data with Hadoop Toolkit. 978 (2014).

[73] The real story of how big data analytics helped Obama win | InfoWorld: *https://www.infoworld.com/article/2613587/the-real-story-of-how-big-data-analytics-helped-obama-win.html*. Accessed: 2019-07-30.

[74] To, Q.C. et al. 2018. A survey of state management in big data processing systems. *VLDB Journal*. 27, 6 (2018), 847–872. DOI:https://doi.org/10.1007/s00778-018-0514-9.

[75] Uzunkaya, C. et al. 2015. Hadoop Ecosystem and Its Analysis on Tweets. *Procedia - Social and Behavioral Sciences*. 195, (2015), 1890–1897. DOI:https://doi.org/10.1016/j.sbspro.2015.06.429.

[76] Wang, H. et al. 2016. Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*. 367–368, (2016), 747–765. DOI:https://doi.org/10.1016/j.ins.2016.07.007.

[77] Wang, Y. et al. 2018. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*. 126, (2018). DOI:https://doi.org/10.1016/j.techfore.2015.12.019.

[78] Welcome to Apache Avro! *http://avro.apache.org/*. Accessed: 2019-08-02.

[79] Welcome to Apache Pig! *http://pig.apache.org/*. Accessed: 2019-08-02.

[80] White, T. *Hadoop : The Definitive Guide*.

[81] Zheng, Z. et al. 2015. Real-Time Big Data Processing Framework : Challenges and Solutions. 3190, 6 (2015), 3169–3190.

[82] Zhou, J. et al. 2013. CloudThings : a Common Architecture for Integrating the Internet of Things with Cloud Computing. (2013), 651–657.