EAI Endorsed Transactions
on Scalable Information Systems Research Article EAI.EU

Empirical Analysis of Recent Advances, Characteristics and Challenges of Big Data

Burhanullah Khattak[1], Aurangzeb Khan[1], Khairullah Khan[1], Wahab Khan[2], Muhammad Kamran[3], Muhammad Fahad[1]

[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science & Software Engineering, IIU, Islamabad 44000, Pakistan
[3]Department of Electronics, University of Peshawar, Pakistan

Abstract

Here in this study, we provide an empirical analysis of recent advances, characteristic and challenges of big data. Initially, we acquaint the readers with the general background, history, and characteristics of big data including volume, velocity, value and variety etc. The scope of applications for big data including political services and government monitoring, enterprise management, scientific research, public utilities, public administration and internet of things are illustrated. A detailed analysis is presented regarding opportunities and challenges faced by the public and private sectors during analysis phase of big data management such as storing, visualizing, capturing and so on. In addition, we investigated and reported a detailed empirical analysis of the most recent management tools like Hadoop and MapReduce, along with their different components, usage, and limitation. Finally, open issues and future directions for this new and dynamic area of research are provided. The primary objective of this empirical analysis is to present a broad-spectrum perspective of this emerging research area with the goal to present big data related concepts in a coherent manner to the beginners.

Keywords: Hadoop, velocity, big data, variety, volume

Received on 07 November 2018, accepted on 15 July 2019, published on 24 July 2019

doi: 10.4108/eai.13-7-2018.159621

1. Introduction

Typically, big data consist of data sets that are very huge in size which is quite difficult for the commonly available software tools to manage, process, capture, and organize it in a tolerable elapsed time[1]. Big data is comparatively fresh in information technology and industry. The term also used in the past by numerous researchers and practitioners. Such as, the authors of [2] defined it as a huge amount of scientific data for visualization while the authors of the research work [3] described the same as "the huge quantity of data that are just away from current technology, scalability to save, control, and process well." John Mashey has been using the term since the 1990s and credit goes to him for coining or if nothing else making it famous[4, 5]. Big Data, which is also named data intensive technologies, has been developed as a modern scientific, business and industrial approach [5, 6]. Now a day's analyses of big data are the focus of business and science. Unlike conventional data, the term, Big Data is used to allude huge amount, complex, increasing data sets that incorporate various formats: semi-structured, structured and unstructured data. Big Data mainly focus to cover up unstructured data.[5]. The constantly moving target of big data is its "size", which is increased from only some dozen Terabytes (TB) to several Petabytes (PBs) of data, by 2012 [7].

The current international population goes above 7.2 billion [14, 49], and 51% of the world's population has internet access, by June 2017, [8][50]. According to news reports released by Global System for Mobile Communications, (GSMA) Intelligence, The amount of cell phone users worldwide has reached to 5.035 billion, out of which the latest billion users being added in the last

[*]Corresponding author. Email:Wahab.phdcs72@iiu.edu.pk

04 years [51], huge volumes of data are producing by these millions of individuals using devices like remote sensors constantly generate a lot of diverse types of data that are either unstructured or structured, this huge volume of data is referred as Big Data [9]. Now a day's Big Data are nearly associated with every field of individual activity ranging from just production design, digital facilities or commodities deliverance to the end customer and recording events to research. Such types of data are created from search queries, emails, click streams, sensors, posts, logs, images, audios, videos, social network interactions, cell phones and their uses health records, science data, and online transactions [10, 11]. Large data presenters like Google, Yahoo Flickr, and Facebook, run a hyper-scale computing environment. Only YouTube is accessed by more than 30 million visitors every day, and three (300) hrs of video are uploaded to YouTube, per minute[12]. Scalability and storage capacity has become a big problem, due to this huge amount of data being generated, At the organizational level storage demand for big data is reaching PB and even beyond[13]. Maintaining and storing such a huge amount of data at the same pace can be difficult. Different factors like performance, cost, capacity, scalability, and throughput are involved in every perfect storage solution system. Additionally, storage devices perform a significant function in justifying big data challenges. Reliability must be given an equal consideration to store big data. The retrieval of data in its original shape without any loss is known as reliability. The issues of reliability consider both external and internal system failures and vulnerabilities. As the size of data increases, the possibility of losing some data in retrieval can be extremely high. Large data-intensive applications such as Google map and Facebook requires high Input-Output-Operations-Per Second (IOPS) to maintain performance in order to stay in business [13].

The amount of unstructured data is rapidly increasing due to daily upload of high-definition videos and pictures with our mobile devices. There is a convincing need for intensive research on management and storing big data. Given the current volume of data, large organizations are employing techniques such as data compression, deduplication, object storage, and cloud storage [13].

They are saved in databases which develop enormously and become hard to handle, save, capture, visualize, analyze and share through usual database management system. Data created by a human until 2003 were five (5) EXABYTE (EB), while now a day's such a huge amount of information is being generated in a couple of days. Data of digital world stretched to 2.72 ZETABYTE (ZB) in 2012, in 2013, the volume of data recorded worldwide was 4.4 ZB, while in 2015, the size of data was reached to about eight (8) ZB i.e. double after every two years[7]. That is set to mount obliquely to forty-four (44) ZB by 2020[14]. The volume of data is predicted to be increased three hundred (300) times from 2005 to 2020 i.e. reached to 40,000 exabytes from 130 exabytes.[1] IBM points out that daily 2.5 EB of data is being generated and 90% of the data generated in a preceding couple of years [15]. About 500 GIGABYTE (GB) of data can store by a single digital computer, hence about twenty (20) billion Computers would need to hold the entire world's data. Human genome decryption procedure gets just about ten (10) years in the past, while now a day's the same process takes a week or two [16]. On the web backbone traffic, audio, video and graphical data have huge mass and is estimated to raise seventy percent (70%) by 2013 [3]. More than one million servers have owned by Google worldwide. In the world, the total numbers of mobile subscribers are six (6) billion and daily only text messages delivered by these mobile subscribers are ten (10) billion. Moreover Fifty (50) billion devices will be linked to Web, by the year 2020 [15]. The daily number of Face book's users for September 2017 was approximately 1.37 billion and monthly active users are 2.07 billion since September 30, 2017, using seventy (70) different languages, uploaded 140 billion photographs, and buddy links are 125 billion, 30 billion pieces of content each day and 2.7 billion comments and likes have been received. All day long forty-eight (48) hrs of video are uploaded per minute and Four (4) billion views executed on YouTube [17]. Google props numerous services like both monitories 7.2 billion pages each day and twenty (20) PB of data processes per day also translates into sixty-six (66) different languages. Above 140 million active users tweets 1 billion tweets on Twitter after every 72 hrs. Daily per minute 571 new websites is generated[18]. The Graph of information will soar up to fifty (50) times, within the next decade, but the Graph of information technology specialists will get 1.5 times advancement in all that data which they receive [24]. Advanced algorithms and powerful technologies are required for Big Data due to its complex nature. The usual tools of business intelligence cannot be effective anymore, as for the Big Data applications are concerned [19].

The main contributions of this study are listed below:

- To present a broad-spectrum perspective of this emerging research area
- To describe the concept of big data in coherent way such that it might be helpful in coming for beginners in this domain
- To highlight scope of applications for big data
- To present a detailed analysis regarding opportunities and challenges of big data

This paper will represent Big Data concepts and would enrich the concept given in [20] that comprises the 5Vs : Velocity, Variety , Veracity, Volume, Value and propose other elements for Big Data classification and analysis, especially evaluating and comparing big data technologies in social media, business, society administration, industry, science, public utilities and so on. The remaining portion of the paper is set as: Section II describes Characteristics of big data, Section III shows different Big Data application, and Section IV presents big data challenges, Section V presents Management Tools, and Section IV presents Hadoop usage and limitation of Hadoop in detail. While section VII concludes the work.

2. Characteristics of Big Data

Majority of the data scientists and professionals have expounded big data with 3Vs: volume, variety, and velocity (called the 3Vs) [21]. Gartner initially introduced the terms velocity, variety, and volume to show the components of big data challenges. Following are the different characteristics to describe big data [22, 23]

2.1 Volume

Volume stands for the collection of all kinds of information created from various sources and keeps on growing [24]. Huge amounts of digital data are producing constantly from millions of applications and devices (smart phones, logs sensors, social networks etc.). As stated by McAfee et al. (2012), that in 2012 daily 2.5 EB approximately were produced. This figure is doubling approximately after every 40 months., the entire digital data produced, simulated, and consumed was 4.4 ZB (ZB) in 2013, as projected by the international data corporation research reports, this amount is doubling after every two (2) years and progressed to eight (8) ZB by 2015. IDC reported that the amount of data will get in touch with Forty (40) Zeta bytes and an increase of 400 times by now. IDC projected that there will be 163 ZB of data, by 2025[25]. A question for large organizations is seminal who ought to be the owner of big data ideas that have an effect on the whole organization[26].

2.2 Variety

Variety stands for the different kinds of data gathered through Mobile devices, sensors or the Internet. Variety presents information regarding the category of data, for instance, semi-structured, structured, unstructured [1]. We no longer simply have structured data that fits decent and flawlessly into an information table. The current information is unstructured. If truth be told, eighty percent (80%) of the whole world's information comes in this types, comprising video successions, photographs, online networking refreshes, and so forth [27]. Data of such kinds incorporate audio, video, text, image, and data logs, in any format. Huge data sets comprise of unstructured and structured data, private or public, shared or confidential, distant or local, incomplete or complete, and so on. Data created from Smart Phones appliances are mostly not in structured format. Such as online games, text messages, social media and blogs produce various categories of unstructured data via sensors and smartphones. A tremendously different set of unstructured and structured data are also produced by users of the internet[28].

2.3 Velocity

The pace at which new data is being generated, accumulated and analyzed is known as Velocity. Data are regularly changed because of the continuous production of data from various sources [29]. Data are produced rapidly and ought to be processed quickly to extract valuable data and pertinent bits of knowledge. The amount of electronic mails, video clips, social media posts and new text produced daily is more than several million entries. For example, Wal-Mart produces in excess of 2.5 PB information consistently from its client's exchanges. Another great case that shows the rapid pace of Big Data is YouTube. Since new data is added so quickly, it is essential to analyze the data in real-time. These days technology for big data has the capability to analyze data as soon as it is created Apart from the above mentioned three (3) important Vs. To better define big data, additional Vs and other features have been added by some actors, these features/characteristics are discussed one by one below:

2.4 Veracity

The nature of captured data can differ very much, influencing the precise analysis. Veracity is the assorted variety of quality or dependability of the data [27], exactly how precise is this information[1]? For instance, consider all the Twitter posts with hash labels, contractions, grammatical errors, and so on, and the trustworthiness and precision of all that content. Gathering burdens and heaps of information is of no utilization if the quality or reliability isn't precise. Another great case of this identifies with the utilization of Global Positioning System (GPS) [30] information. Frequently the GPS will "float" off track when you scrutinize via city vicinity. Satellite signals vanish as soon as they bounce off high buildings or some other structures. At the point as soon as these occur, area information must be combined with another information source like street information, or information from an accelerometer to give exact information.

2.5 Value

One of the most imperative characteristics of big data is Value; it stands for the process of extracting/finding pertinent data from huge data sets used for several sectors [31]. When we discuss value, we are referring to the significance of the data being extorted[1]. If we have an unlimited quantity of data it is one thing, but it is useless until and unless it can be changed into value. Although there is a clear connection between data and insights, these do not generally imply that there is value in Big Data. To set out on a big data project, the most essential part is to figure out the expenses and advantages of analyzing and gathering the information to guarantee that eventually the information that is gathered can be monetized [27].

2.6 Variability

Variability refers to the inconsistent pace as a result of which information is stacked into your database. In big data's circumstance Variability stands for a couple of various things. First is the amount of inconsistencies in the information. These should be located through anomaly and outlier detection techniques sequentially for every substantial analytics to crop up[32]. Big data is also changeable due to a large number of data dimensions resulting from a variety of distinct sources and data types.

2.7 Validity

Like veracity, validity stands for how precise and accurate the data is for its planned utilize. As indicated by Forbes, approximately sixty percent (60 %) of data scientist's time is exhausted cleaning their data before being capable to perform any analysis [33]. The advantage from big data analytics is just as excellent like its fundamental data, thus we have to receive great information administration practices to guarantee reliable data quality, basic definitions, and metadata [32].

2.8 Vulnerability

Big data brings new issues regarding security. In any case, with big data, a data breach is a major violation. Does anybody memorize the villainous Ashley Madison hack in 2015? [108] Unluckily, numerous big data breaches are there. One more instance, as announced by Collaborative Research Network (CRN) in the month of May, 2016 "one of the hacker known as Peace posted information for sell on the dark web, which integrated information allegedly on 360 million emails and passwords for MySpace users and 167 million LinkedIn accounts." Information on numerous others can be found at Information is Beautiful.

2.9 Volatility

How much the data should be old prior to it is viewed as historical, immaterial or not valuable anymore? To what extent does information should be kept for? Prior to big data, organizations have a tendency to save data indefinitely -- a small TB of data might not produce much storage expenditures; even it can be reserved in the live database without facing any performance problems. There may not even be information archival planes set up, in a traditional information setting [32]. Volatility requires to be cautiously measured owing to the volume and velocity of big data. You now require setting up rules and regulations to access the data easily and to give assurance of quick retrieval of information as and when required. Ensure these are plainly fixing to your company requirements and procedures. The complication and costs of a storage and retrieval process are magnified with big data.

2.10 Visualization

Visualization is a new challenging characteristic of big data i.e. How to visualize the big data using current visualization tools. Because of the limitations of in-memory technology [34] and low response time, functionality, and poor scalability existing big data visualization tools visage technical challenges. when annoying to plot a billion data points you cannot depend on conventional charts/graphs[35], so you require altered ways of presenting data, for example, circular network diagrams, sunbursts[36], cone trees[37] and parallel coordinates[38]. These days' text data exist abundantly in the form of e-books, news article, social networks, and financial analysis etc. As predicted there exist 80% of the world data in an unstructured format. Conventional database query methods are not suitable to obtain useful information from this large collection of data. Documents clustering is an unsupervised categorizing of a set of documents into self-similar clusters such that each document is more identical to one another in the same cluster than with a document of other clusters [39] as shown in Figure 1 [40].

Volume
Velocity
Variety
Variability
Vulnrerab ility
Validity
Value
Veracity
Volatility
Visualizati on
BIG DATA WITH 10 Vs

Figure 1. Schematic Representation of Big Data in terms of 10 Vs

3. Application Areas

Big data are used more or less in every field of our lives. A few imperative applications of big data are listed below:

3.1 In Business and Commerce

In proportion to ballpark figure, after every 1.2 years the volume of business data throughout the world doubles, approximately across all businesses [3]. Taking the retail industry, for instance, we attempt to offer a succinct presentation in favour of the functionalities of Big Data in business aerobics. Daily in the 6000 stores of Wal-Mart's, there are approximately 267 million businesses throughout the world[41]. For searching higher competition in sell, Wal-Mart in recent times worked in partnership with HP (Hewlett Packard) to ascertain a data warehouse that has a capacity to save four (4) PB of data [24]. In the age of information, relatively every huge organization experiences Big Data issues, particularly for multinational corporations. From one perspective, those organizations generally have a large number of clients around the globe; Then again, there are extensive volume and velocity of their trade data. For example, more than 2.1 billion legal accounts worldwide managed by Fair Isaac Corporation's (FICO) Falcon credit card fraud detection system. Each day Facebook generated over and above three (3) billion pieces of content [24]. A similar issue occurs in each internet organizations. The rundown could continue forever, as we witness the future organization's war zones concentrating on Big Data.

3.2 In Populace Administration

Big Data issues also engage Public administration [42]. Normally, on one hand, the general population of a nation is extremely huge, on another hand; individuals at every age level require diverse public services. Such as, children and youth require much learning while senior citizens need much health care [24]. Everyone in the general society produces a large amount of data in all public sectors; as a result, the overall amount of data regarding civic administration in a country is incredibly enormous. For example, there are just about 3 TB of information gathered by the Library of Congress of the United State by 2011. Big Data Research and Development program was launched by the administration of Barak Obama in March 2012 with a financial plan of $200 million [43], which explore tending to essential issues facing the administration by making use of Big Data. Big Data development, in Japan, became a significant weapon, of the nationwide scientific plan [19]. In July 2012, The United Nations released a report, with the title Big Data for Development: Challenges and Opportunities. It aspires to highlight the major issues relating to Big Data challenges and to promote the conversation about how global progress can be served using Big Data. As indicated by McKinsey's report [3], Big Data functionalities, like storing information and informative samples, offer the civic zone an opportunity to enhance production and high degrees of performance. The public sector of Europe decreases the costs of administrative activities by 15–20 percent approximately and growing values from 223 billion to 446 billion, or greater than this [24]. This approximation is underperformance achievements and a fall in the diversity between a real and latent collection of tax income. These functionalities can accelerate year production enlargement by up to 0.5 rate focuses throughout the following decade.

3.3 Scientific Research and Big Data:

By the advancement of computer sciences, various scientific areas have already turned out to be extremely data-driven [44, 45]. For example, bioinformatics [46], social computing [47], astronomy, computational biology [48] and meteorology are significantly founded on data-intensive scientific detection since a huge amount of data with diverse kinds created from these areas of science [49]. Searching for information from the data generated via extensive scientific simulation is a big question to solve properly, this is a clear Big Data issue about which the appropriate response is still obscure or insatiable.

3.4 Administration Monitoring and Political Services

Several governments like the United States and India are extracting data to observe political tendencies and analyze public opinions[19]. Several applications are there that merge various data sources: voter compositions, individual interviews, and social network communications. In addition to national issues systems like these also facilitate to identify local problems. Moreover, to optimize the use of important utilities and resources, governments could apply Big Data systems such as, to check the flow of water in big systems of water supply chains. Sensors can be set in the pipelines of water. Thus, for many countries, it is feasible to be dependent on real-time checking system to identify illicit connections, seepage and manages the valves remotely to guarantee equal distribution of water to various spots of the urban[19].

3.5 Internet of Things

The internet has updated worldwide cultural upheavals, interrelations, the craft of businesses and a mind-blowing amount of individual attributes. One major and important marketplace of big data applications is Internet of Things (IoT)[50]. Due to the large collection of objects, IoT applications are constantly developing these days [19], different applications of big data are there for supporting logistic ventures. Actually, it is feasible to follow automobiles location with GPS, wireless adapters, and sensors. In this manner, data-driven applications like these allow organizations to handle, supervise workers and to optimize the routes for delivery. Right now, machines are getting in on the act to manage countless self-governing devices through the web and produce IoT. Hence, machines are turning into the user of the internet, much the same as people with the web browsers. IoT is magnetizing the interest of current researchers for the majority of its encouraging challenges and opportunities. IoT has a very important financial and societal effect for the future information structure of information, communication technology and network [51]. The new direction of the time to come shall be at last; everything will be linked and managed wisely. IoT idea is much appropriate to the practical world because of the advancement of cell phones, cloud computing, data analytics, embedded and ubiquitous communication technologies. Furthermore, challenges are presented by IoT in a combination of velocity, variety, and volume. Smart City [52] is a burning field for research founded on the use of IoT data. Such as the Smart City venture collaboration among the IBM and Miami-Dade County in Florida nearly [62] connects 35 sorts of core region government offices and Miami city and assists administration to get improved knowledge support in taking verdict for improving public safety, supervision of water resources, and decreasing bottleneck. The utilization of smart city achieves benefits for Dade County in numerous angles. Such as, the Park Management Department of Dade County secures one (01) million US Dollar this year in water charges because of well-timed recognizing and settling water pipes that were leaking [53].

3.6 Populace Utilities

Various utilities like an organization for the supply of water is introducing sensors in the pipelines, in the complicated systems of water supply, to check water flow. According to a press release that the Board for Sewerage and Water Supply of Bangalore is applying a real-time checking system to identify illicit connections, seepage and manage valves distantly to guarantee fair provision of water to various regions of the municipality [19]. It assists to cut the need for the operators of the valve and to identifying and fixing the leaking of water pipes timely.

3.7 Logistics and Transportation

Several main road transportation corporations are utilizing GPS and RFID (Radio-frequency Identification) [54] to follow vehicles and find out attractive data to get better their services. Such as, data gathered regarding the number of passengers using the vehicles in various directions are used to advance bus paths and the number of visits. Diverse real-time systems are in use to provide travellers with suggestions as well as to present priceless information about the expected time of the next motor vehicle to carry a passenger to the required location.

Big Data extraction also assists to advance the traveling industry by predicting requirement regarding populace or personal networks. Such as the total amount of booked seats issued in India daily is approximately 250,000 and booking can be done Two (02) months prior[19]. Making prophecies from such a large quantity of data is a complex problem as it relies on numerous issues for instance festivals, night train, weekends, intermediate or starting station. It is likely to extract and implement improved analytics on old and novel big data set using the machine learning algorithms. Actually, improved analytics can make sure high precision of results concerning several problems.

4. Big Data Challenges

Several gorgeous opportunities are provided by the mining of Big Data. Nonetheless, few difficulties are faced by the professionals and researchers when searching huge Data collections, mining value, and information from the mines of data like that. Challenges lie at various stages constituting: capturing of data, storing data, searching for information, data sharing, data analytics, managing and visualizing the data. Moreover, privacy and security problems are also there, particularly in distributed data-driven applications.

4.1 Data Capture and Storage

The size of datasets raises as they are constantly being collected through omnipresent data-sensing Cell phones, in-flight sensory technologies, S/W logs, digital cameras, RFID, microphones, wireless sensor systems, etcetera. Daily 2.5 quintillion bytes of data are produced and the figure continues expanding exponentially [55]. After each three (3) years the world’s technological capability to store data has nearly doubled since the 1980s. In various areas, as economic-related and medicinal related data regularly be erased on the grounds that there is not enough space to store this information. This significant information is produced and captured at a high cost, yet unnoticed at last[56].

The approach we store and capture data has modified by Big Data, constituting data storage architecture, data access mechanism, and data storage device [57]. Since we need additional storage devices and advanced I/O pace to address the challenges, in fact, we require extraordinary developments. Primarily, Big Data accessibility is the peak concern of the information finding the procedure. To save the persistent data hard disks were used in the previous decades[58]. We know that the random I/O performance of Hard Disk Drives (HDD) is slower a lot as compared to sequential I/O performance, however, HDDs are replaced gradually with Sequentional Storage Devices (SSDs) nowadays, and other technologies are also expected to come such as Pulse-code modulation (PCM). These existing storage technologies cannot have alike high performance at the same time for random and sequential Input/output (I/O) as well, thus for the processing systems of big data we need to think again about the design storage subsystems i.e. Direct Attached Storage (DAS), Storage Area Network (SAN) and Network Attached Storage

(NAS) are the venture storage structural design that was usually used [59]. Nevertheless, as it comes to large-scale distributed systems all these current storage architectures have serious limits and disadvantages [24].

4.2 Data Curation

Data Curation is intended for the retrieval of data and discovery, value calculation, data quality guarantee, reuse, and protection ultimately. This field particularly includes numerous sub-fields consisting of management, archiving, authentication, retrieval, representation and preservation [24]. The current tools for the management of the database are not able to process Big Data. This position will carry on because the advantages of developing Big Data enabling researchers to avert sicknesses fight offense and explore trade tendencies. Although the volume of Big Data continues expanding exponentially, existing ability to function with is merely in the comparatively lower stages of PB, EB, and ZB of data. NoSQL database [60], is a recent approach for distributed and huge database design and for data management. For the management of structured data, the standard approach consists of 02 elements, the first element is known as a schema, which is used for the data store, and other one element, used for the retrieval of data, is a relational database.

4.3 Big Data Aggregation

Synchronizing distributed Big Data platforms and exterior data sources with the interior infrastructures of a business is one more challenge [24]. Generally, analyzing the data created within the businesses is insufficient. For the mining of important information, it is significant to aggregate interior data with exterior data sources.

4.4 Transmission of Big Data:

Cloud data storage space is highly utilized as the improvement of cloud technologies. It is known to us that the bandwidth capacity of the network is the bottleneck in distributed systems and cloud, particularly for the communication that is large in size [24]. On another hand, cloud storage as well guides to data security issues [61] because of the needs of data integrity checking several proposals were planned under various security models and systems[62, 63].

4.5 Cleaning of Big Data

The practice of identifying and correcting or eliminating incorrect or damage records from a database, table or record set and refers to recognizing deficient, imprecise or inappropriate data parts and then substituting, amending, or erasing the dirty or uncouth data[64]. How to cope with the intricacy of big data nature and process it with a blend of applications in a distributed environment is a challenge in Big Data [24]. Nonetheless, data sources might have noises, mistakes or deficient data. Cleaning and making a decision on such huge data sets, regarding the reliability and usefulness of data, is the challenge.

4.6 Imbalanced Big Data

Classifying imbalanced dataset is one more challenge in the management of Big Data. In the previous year's, this problem has caught plenty of interest. Actually, classes with diverse distributions might create by real-world applications. Positive or minority is the first kind of class with a small number of instances, negative or majority is another class with a huge number of instances [24]. In a variety of areas for example bioinformatics, diagnosis [65]or drug discovery, classifying the Positive classes is significant. In fact, several problem areas contain more than two classes with irregular distributions, for instance, weld flaw and protein fold categorization. New challenges were introduced from these multi-class imbalance problems that were not experienced in two-class problems. Indeed, it is difficult to handle multi-class tasks than two-class tasks. Various techniques have been developed for the solution of this issue and generally classified into two categories. One category contains few binary classification methods to be suitable for multi-class classification problems, for example, Naive Bayes, K-NN, SVM, discriminant analysis, decision tree, and neural networks. Another type is identified as DEM (Decomposition and Ensemble Methods). It will be composed of decomposing a multi-class classification problem into a band of binary classification problems that can be solved by BCs (Binary Classifiers), and after that, by means of implementing an aggregative strategy on the BCs' prophecies, classifying a new observation [66].

4.7 Data Integrity

The integrity of Data [67] is crucial for attaining the fifth V (value) of Big Data by means of cross-domain relationships and integrative data analysis[68]. The challenges for data integration of data fusion, schema mapping, and record linkage is summarized [68]. Metadata is necessary for tracing these mapping to build the integrated data sources 'automatically' decidable and to facilitate large-scale analysis [69]. Nevertheless, producing metadata automatically and efficiently from Big Data is still a difficult task. In the geospatial domain, geo-data integration has glinted innovative chances determined by ever progressively mutual research atmosphere. For example, Earth Cube [94] program which gives unparalleled analysis and integration of geospatial data from a range of geoscience areas.

4.8 Data Analysis

When we talk about Big Data analysis tasks, the major and most significant challenge is scalability because the initial influence of big data is volume. Data analysis is a significant stage in the value series of big data for data mining as well as for predictions. Over the most recent couple of decades, researchers gave special treatment to speed up analysis algorithms to accelerate processors following the Moore's Law and to deal with growing volumes of data. It is essential to building up sampling, online and multi-resolution analysis techniques [70].

Since the volume of data is scaling more rapidly as compared to CPU speed, there is a usual spectacular change in processor technology [69] the clock speed still extremely drop behind, Although, following Moore's Law, the clock cycle frequency of CPUs is doubling. On the other side, CPUs are being implanted with growth figures of cores. This change in CPUs results in the development of parallel computing [71]. Moreover, the majority of the current analytical algorithms need uniform structured data and face troubles in processing the heterogeneity of Big Data. To cover this space, either we need the latest devices designed for pre-processing data to build them structured to suit current algorithms or new algorithms that handle heterogeneous data. Timeliness is the main concern for those real-time Big Data applications, like astronomy, finance, biomedicine, internet of thing, social networks, navigation, and intelligent transport systems, if the size of data that will be processed is very large how can we guarantee the timeliness of response? For stream processing, it is still a big challenge concerned to Big Data. It is all in all correct to state that Big Data has not just created various challenges and modified the ways of hardware improvement, but in the architecture of the software also. That is the veer to cloud computing [71, 72], that combines a variety of different workloads into a huge clump of processors.

4.9 Data Quality

Data quality consists of four(4) features: redundancy, consistency, completeness, and precision [68]. The inherent character of heterogeneity and difficulty of Big Data makes data correctness and completion hard to recognize and follow[73], hence 'false discoveries' threat is growing. For instance, data of social media are extremely slanted in space, demographics and time, and accuracy of the location changes to 100s (hundreds) of KMs from meters. Moreover, filtering and data redundancy control ought to be carried out at the point where the data are collected in real-time for example through sensor networks [68]. Eventually, assuring data integrity and consistency with big data is challenging particularly when the data regularly vary and are shared with numerous co-workers [74].

4.10 Data Privacy Challenges

Although big data provided ease and people enjoying it but at the same time user also faces lots of troubles. If the data is not sheltered well and while using ,it will directly threaten directly security of the data and user privacy, leading to privacy risks[75] The unparalleled networking amongst calculating platforms and smart devices contributes to Big Data but faces safety issues where a person's position, transactions, and conduct are digitally recorded. For instance, personal medical records and social media consist of individual health information elevating secrecy issues[1] [76]. Massive amount of cell phone internally stores lots of personal data. Currently, users of smart cell phones have serious issues with the security and privacy of not just big data but also security of their smart cell phones has become a serious issue. As users operate and controls their smart intelligent product of their smart homes through smart cell phones. Hence if the personal phone controlled by others or lost, it will bring grave security issues to the smart home user [75]. Another case is that organizations are using big data to check employees' efficiency by following the worker's development and movement. These security concerns uncover a space between the conference strategies/regulation and Big Data and identify new policies to tackle comprehensively security issues [68].

4.11 Data Security

The growing reliance on web and computers is now a day's world makes businesses and folks exposed to data violation and misuse[1]. Big Data faces new security challenge for algorithms, Methodologies and conventional data encryption standards [77]. Earlier investigations of data encryption concentrated on the medium and small size of data, which don't function fine for big data because of the problems of scalability and efficiency. Additionally, data security strategies and schemes to work with the structured data saved in traditional Database Management System are not efficient in dealing with extremely diverse, unstructured data. Hence, well-organized strategies for security management and data access control requirements to be examined in Big Data and these require to integrate new storage construction and new data managing systems [5]. In the era of cloud the data has become more exposed to cyber-attacks, almost every sector is affected from data breach and threats like presidential election of US, recently Unique Identity Development Authority of India (UIDAI) attack, shocking incident of Ransomware attack, Wikileaks, panama papers etc as data holders have partial control of virtualized storage, assuring data secrecy, availability and integrity turn into a key concern [1] Currently, different countries like China yet lack rules and regulation for the management and security of user's information. The personal data of people does not have good and proper supervision because of this weak and not perfect supervision system, be short of technological support and susceptibility of information loss, data is decreased and the use of data information is not of high value.[75]

4.12 Data Architecture

Big Data is increasingly changing the style scientific research carried out as witnessed by the progressively the open science and data-driven approach. Such changes lead to challenges for system architecture [68]. For instance, faultlessly combining various tools and geospatial services stay a prime concern. Other priority problems comprise of combining these tools into workflows that can be used again, to promote functionality integrating data with the tools and among communities sharing data and analysis. A perfect design would flawlessly produce and share data, models, tools, calculating resources, network and most prominently, folks. Geospatial cyber infrastructure is agilely looked into in the geospatial sciences. Although, Earth Cube program which is still in an initial progressive phase, is a fine paradigm of such cyber infrastructure in the geospatial area. Creating an analogous cyber infrastructure for other science areas is just as significant and challenging.

4.13 Data Visualization

Big data visualization exposes concealed patterns and finds out anonymous correlations to get better decision-making [68]. The central goal of data visualization [78, 79] is to display the information in an effective and efficient way by applying various graphs, to impart information simply by presenting information concealed in the complicated and huge datasets. The type of Big Data is mostly heterogeneous, semantics and structure, visualization is crucial to formulating the logic of Big Data. But to provide human communication and visualization in real-time for visually analyzing and searching big data is hard to achieve[35].

Every month online souk eBay sold billions of goods and have hundreds of millions active clients and they produce plenty of data and for this purpose, eBay shifted to Tableau[80]. A tool is set for the visualization of big data, to make all that data comprehensible. Tableau has the capability to change huge, complicated data sets into instinctive photos. The outcomes are interactive also. Realizing the importance of Tableau, the employee of eBay is able to visualize search relevancy and value to check the up-to-date client remarks and do sentiment analysis. It is specifically complex for Big Data applications to carry out the visualization of the data due to the high dimension and huge amount of big data. Nonetheless, existing tools for the visualization of Big Data generally have low efficiency in scalability, functionalities and response time. We have to think again about the technique used for the visualization of big data, not similar to the way we take up earlier than. For instance, the history mechanisms for the visualization of information are data-intensive also and require approaches that are effective to a greater extent [81].

4.14 Big Data Management

Management of Big Data is another challenge. Big data management in an effective way is critical to facilitate the mining of reliable insight and to optimize costs. In fact, the basis for the analytics of big data is a good quality data management [19]. Data is progressively sourced from different fields that are disordered and mixed-up, for example, information from sensors or machines and huge sources of private as well as public data. In the past, the majority of organizations were not capable to either store or capture these data and the data cannot be dealing with the existing tools in a sensible measure of time [82]. Nonetheless, the latest technology enhances efficiency, assists novelty in the services and goods of big business models, and endorse us to make a decision [3]. Otherwise stated, the objective of big data management is to assure consistent data that is easily accessible, secured, manageable and stored appropriately. The goal of Big Data technology is to lessen processing and hardware expenses and to confirm the value of Big Data prior to assigning important company resources. Appropriately managed big data are secure, easily accessible and reliable. Thus, applications of big data can be implemented in a variety of complicated scientific areas, like biogeochemistry, astronomy, atmospheric science, Genomics, medication, and biology.

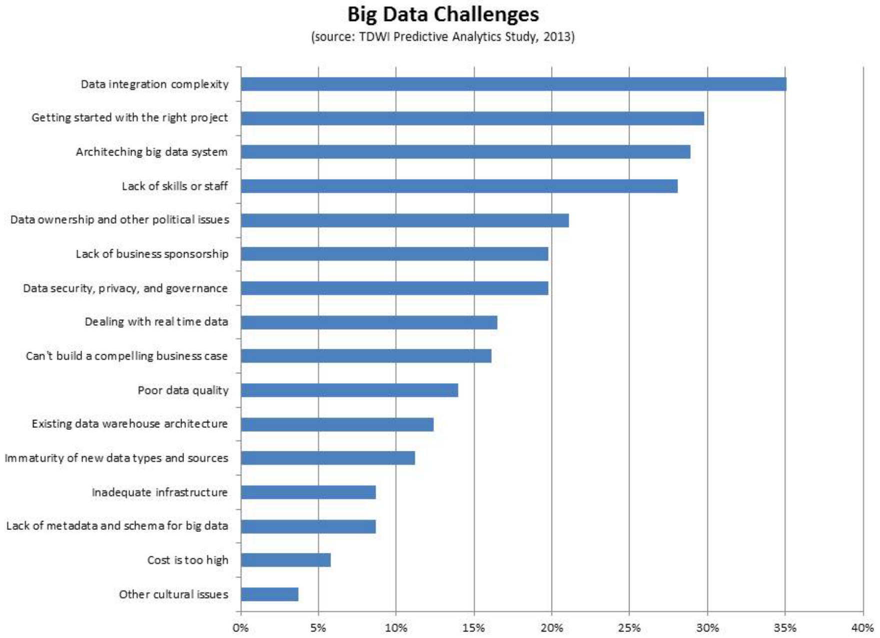

Figure 2. Big data Challenges (Source: TDWI predictive analytics study, 2013)

5. Big Data Management Tools

With the advancement of computing technology, a huge quantity of data can be managed with no need for supercomputers and high expense. Numerous methods and devices are available for the management of data, like Voldemort [53], Simple DB, Google Big Table, Data Stream Management System (DSMS), Not Only SQL (NoSQL) and Memcached DB. Nevertheless, companies should develop especial technologies and devices that can save, analyze and access in near-real time huge volumes of data as big data varies from the conventional data and can't be stored on one machine. Few of the most common tools and techniques used for big data are Big Table, Hadoop, and MapReduce. These modernizations have once again defined data management as they powerfully process huge volumes of data economically, efficiently and in a well-timed way. The subsequent part explains MapReduce and Hadoop in more detail.

6. Hadoop

Apache Hadoop [83] written in JAVA is a renowned top-level open source Big Data project that started in 2006. Hadoop has been planned to avoid/minimize the complication and poor performance faced by traditional technologies while analyzing and processing Big Data. Its capability to quickly process huge data sets is one of the major benefits of Hadoop, Actually, Hadoop does not copy the entire remote data in memory to carry out calculations like conventional technologies, Alternatively, Hadoop carries out tasks anywhere data are saved. Hence, a substantial load of communication has been reduced by Hadoop from servers and network [19]. For instance, Hadoop consumes only a few seconds to query TB of data rather than twenty (20) minor more on classical Security Information and Event Management (SIEM). One more benefit of Hadoop is its capability to execute programs as assuring fault-tolerance, generally met in a distributed environment. To make this sure, it copied data on servers to avoid the loss of data.

Hadoop power is founded on 02 main subparts: Hadoop Distributed File System (HDFS) and MapReduce[41]. Additionally, modules can be added by users on top of Hadoop when required according to their application needs and goals (e.g. reliability, capacity, security, scalability). To improve Hadoop ecosystem, the community has contributed with several open source modules. Around 63% of organizations used Hadoop to control the massive number of unstructured events and logs [84]. Particularly, very huge sizes of data can be processed by Hadoop with variable structures or at all no structure. Hadoop includes: HCatalog, Avro, Oozie, Mahout, Zookeeper, Flume, HBase, Chukwa, JAQL, Hive, Sqoop and Pig; but, the most general parts and famous models for big data are HDFS and MapReduce. Figure 3 demonstrates the ecosystem of Hadoop along with the association of different parts to each other.

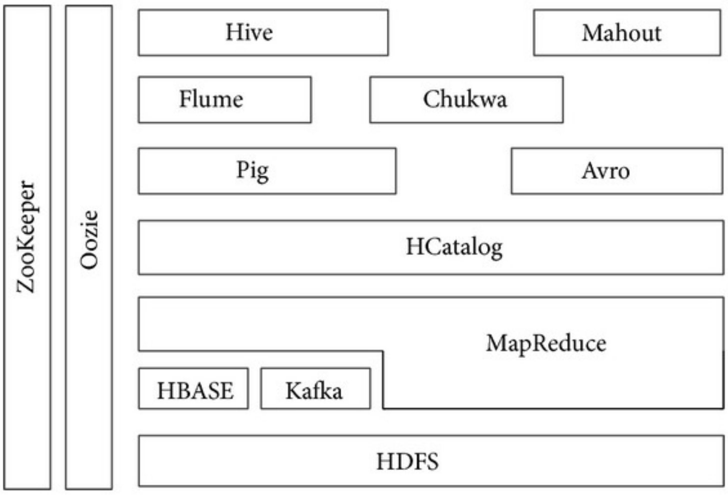

Figure 3. Hadoop ecosystem [77]

6.1 HDFS

The data storage layer of Hadoop is HDFS [85]. HDFS gives a consistent and cost-efficient storage capacity and supports, in a cluster, up to hundreds of nodes. It stores massive volumes of data (i.e. Saved files may be larger than a TB) and can manage both types of data i.e unstructured and structured. Portability crosswise diverse software and hardware platform is a key benefit of HDFS. Additionally, HDFS assists to enhance the system performance and to decrease network traffic by transferring calculations near to data storage. For fault tolerance, it also guarantees data duplication. Such characteristics describe its broad adoption. The architecture of HDFS[41] is master-slave. Huge data is distributed crosswise the cluster. The cluster has several slaves and a distinctive master. The slaves also known as DataNodes coordinate and manage data storage on single compute nodes while file system functions are managed by the master also called NameNode[82]. To provide data availability, Hadoop lies on data replication. The HDFS files are saved in blocks and the size of each block is 64 MB by default. To support the parallel processing of huge volumes of data, all HDFS files are replicated in multiples[41].

6.2 HBase

HBase[86] is an open source, a non-relational distributed database built on top of HDFS. HBase is planned for the operations having low-latency. It is founded on column-oriented key/value data model, which speeds up the efficiency of operations on identical values for huge datasets. For instance, the entire tuples are involved in write and read operations but merely a small subset of the entire attributes. It has the potency to prop large table-update rates and to horizontally scale out in distributed clusters. For extremely huge tables in a format like BigTable, HBase gives a flexible structured hosting. Tables save data in fields (columns) and tuples (rows) logically [19]. The advantage of such tables is that they can manage millions of fields and billions of tuples. Several attributes are allowed by HBase to be gathered into column families with the goal that all the column family elements are saved collectively. This approach is unique from a row-oriented relational database, where the entire columns of a tuple are saved collectively. Hence, to compare with relational databases HBase is more flexible. Alternatively, it has the benefit of permitting users to acquaint updates to deal better application’s that needs changes. Nonetheless, it has the drawback of not propping a structured query language such as SQL. HStore is known as the HBase tables and single or multiple Map-Files have saved in HDFS by every Hstore. Every table should contain a defined schema have a primary key which is used for the accessing of table. The rows are recognized through start key and name of the table while for the identical row key, columns might contain many versions. HBase offers several characteristics like to access consistently sources of Big Data, configurable and automatic sharding of tables, modular and linear scalability, natural language search and real-time queries [87]. Hbase is incorporated in several big data solutions and data-driven websites like Platform of Facebook for Messaging. Hbase can be accessed through different APIs (application programming interfaces) for instance JAVA, Representational State Transfer (REST) and Thrift. These APIs do not contain their own scripting or query languages. As a matter of course, HBase relies totally on a ZooKeeper instance. HBase has a MasterNode likewise to HDFS that deals with the slaves and cluster that save tables elements and carry out operations on data [19].

6.3 MapReduce

MapReduce is the center of Hadoop [88] and is a popular data-parallel processing model that allows mass scalability crosswise several servers in a Hadoop cluster[41]. Every server within the cluster includes a group of inner disk drives that are economical. It is a parallel programming paradigm, with high scalability and fault tolerance, for Big Data [15]. The smart design has instigated the implementation of MapReduce in various architectures of calculating, Such as Graphics Processing Units (GPUs), clouds, hybrid clusters and multi-core clusters [89]. To convey data analytical services, MapReduce also has become a principal option for cloud suppliers [90, 91] has been described in detail the functionality of MapReduce. MapReduce makes simpler the processing of huge amount of data by means of its cost-efficient and effective mechanisms. It allows writing programs that can prop up parallel processing. There are two (02) subsequent functions that are used by MapReduce to manage calculations of data. These two functions are the Reduce function and the Map function. More specifically, following are the steps that are carried out by the MapReduce to perform operations[92]:

1. Primary, the input data, for example, a long text file, is divided by the Map function into data partitions that are not dependent to make up key-value pairs.

2. After that, the entire key-value pairs are sent by the MapReduce framework into the Mapper to processes all key-value pairs independently, during numerous collimate (parallel) map tasks crosswise the cluster. A distinctive compute node is allocated to every data partition. The outcomes of the Mapper are multi intermediate or single key-value pairs. The framework is charged at this phase, to gather the entire intermediate key-value pairs, makes a group and arrange them by means of a key. Thus, the result is several keys with a catalog of the entire related values.

3. Then, to process the intermediate output data, the Reduce function is used. For each distinctive key the values related to the key is aggregated by the Reduce function in accordance with an already defined program (i.e., sorting, taking the average, summarizing, finding the maximum, filtering, hashing). And then, it creates one output key-value pairs or more.

4. In the end, the entire output Key-value pairs are stored in an output file by the MapReduce framework.

In the MapReduce paradigm, a Job-Tracker instance is running by the NameNode in order to schedule the various jobs and distribute tasks over the slave nodes. The JobTracker checks the status of the slave nodes and assigns the tasks again to assure reliable execution as soon as they stopped working. For the assigned tasks, a TaskTracker instance run by the entire slave nodes. As specified by the JobTracker, a TaskTracker instance runs the tasks and checks their execution. To execute, in parallel, numerous reduce or maps tasks, a number of JVMs [93] (Java Virtual Machines) can be used by every TaskTracker.

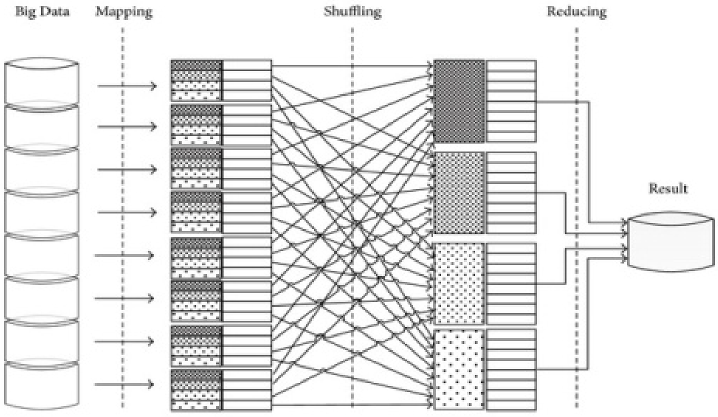

Figure 4. MapReduce architecture [77]

6.4 Pig

Apache Pig[94] is an open source, High-level platform for analyzing huge sets of data. The language layer of Pig presently includes a textual language known as Pig Latin. It supports parallel execution of MapReduce jobs and workflows on Hadoop to decrease the complication of MapReduce[19]. Using HDFS, like Hive, Pig simplifies processing in parallel and exploring huge data sets with its interactive environment. Like binaries, shell scripts and other languages for programming, Pig also allows interaction with exterior programs. Pig has its own data model called Map Data (a map is a set of key-value pairs) [95].

Pig is based on an instinctive syntax to prop up a simple progress of MapReduce workflows (nested or simple) and jobs. While supporting parallelism it decreases the development time. Hence, users can trust Pig Latin language and numerous operators to upload data and process that data[19]. Contrarily to SQL, a schema is not required by pig and can process unstructured and semi-structured data. As compare to Hive, Pig supports additional formats of data. It can execute on both the distributed environment on a Hadoop cluster and the local environment in a single JVM.

6.5 JAQL

JAQL [96] is a declarative language on top of Hadoop that presents a query language and are effectively used for huge data processing. It translates top-level queries into MapReduce jobs. JAQL was planned to query semi-structured data based on Java-Script Object Notation (JSON) format. Nonetheless, other data formats and various data types for example flat files, Comma-Separated Values (CSV) and Extensible Mark-up Language (XML) can also be queried using JAQL. Therefore, a data schema is not required by JAQL like Pig. It offers many built-in functions, I/O adapters, and core operators. Such Characteristics make sure translating, storing, data processing and data transferring into JSON format.

6.6 Hive

The data warehouse system of Hadoop is Apache Hive [19]. It is planned to make simpler Hadoop use. On the contrary to MapReduce, that handles data in files using HDFS, Hive allows showing data in a structured database which is more common for users. Actually, the data model of Hive is mostly based on tables. Tables like those represent HDFS directories and are divided into partitions. Every partition is subsequently divided into buckets. Furthermore, language like SQL was known as HiveQL[19] is provided by Hive so as to facilitate users to manipulate and access Hadoop-based data saved in HBase or HDFS. For that reason, Hive is appropriate for lots of big business applications.

Hive is unsuitable for real-time businesses. Hive is based on a low-latency operation. It is planned for large-scale processing Like Hadoop hence even little jobs might take more than a minute. In fact, HiveQL fairly translates queries (for example. Summarization, ad hoc queries, and joins) to MapReduce jobs that are processed like batch tasks.

6.7 Sqoop

Apache Sqoop is an open source software[97]. Command-line interface (CLI) is provided by Sqoop that guarantees an effective transfer a large amount of data between structured data stores (for example NoSQL, Relational databases and enterprise data warehouses) and Apache Hadoop. Sqoop has several benefits, such as it gives fault tolerance, rapid performance and use of the best possible system to decrease loads of processing to exterior systems. Like JAQL, Hive or Pig the conversion of the imported data is made with any high-level language or with MapReduce[19]. Sqoop permits easy combination with Hive, Oozie, and HBase. As Sqoop brings data from HDFS, the outcome shall be in different files. The files might be bordered sequence files, Text Files or binary Avro including sequential data. Sqoop Export process will take a group of bordered text files in parallel from HDFS, parse these files into records and add them in a database table as new tuples.

6.8 Flume

Flume [19]) is planned to aggregate, accumulate and convert data to HDFS from exterior machines. The architecture of Flume is straightforward, flexible and manages the flow of data. It is based on an easy extensile data model to manage enormous distributed sources of data. It offers different characteristics comprising tuneable reliability mechanism failure recovery service and fault-tolerance. In addition, it offers a query processing engine that can convert every new batch of data prior to it is transferred to the sink.

6.9 Chukwa

Chukwa[98] is a system for the collection of data built on the top of Hadoop. To check huge distributed systems is the goal of Chukwa. To gather data from different data sources it uses HDFS and for the analysis of collected data, it uses MapReduce. It inherits the robustness and scalability of Hadoop. It offers an interface to supervise, analyze and show the outputs.

Chukwa proffers for Big Data, a powerful and flexible platform. It facilitates analysts to analyze and collect big sets of data, in addition, to check and present outputs. To make sure flexibleness, it is structured as a pipeline of collection; define interfaces among stages and processing stages.

6.10 HCatalog

For Hadoop users, Apache HCatalog[99] offers storage management service and a table. It makes the easy exchange of data crosswise tools for data processing (for example MapReduce, Hive and Pig). This is accomplished by means of data type mechanism and a shared schema. An interface is provided by HCatalog to make simpler write and read data operations for every format of data (e.g., Sequence Files, JSON, CSV, and RCFile) for which a Hive SerDe (serializer-deserializer) can be written. To perform this, the SerDe, output format and input format are provided by the system administrator. In HDFS a relational view of data is provided by the abstracted table of HCatalog and permits to show different formats of data in a table form. Hence users don't have to aware how and where data is saved. Moreover, it props users with other services. HCatalog gives alerts about the availability of data and a REST interface is provided to allow access to Hive Data Definition Language (DDL) operations. A notification service is also provided like Oozie to notify workflow tools when in the warehouse new data becomes available.

6.11 Mahout

Apache Mahout ([94] is a machine learning, open source software library. It can be added on Hadoop's top to run algorithms using MapReduce. Mahout is planned in a way that it can also work on other platforms. Basically, it is a Java library set. Mahout has the advantage of guaranteeing effective and scalable implementation of large-scale machine learning algorithms and applications over huge sets of data. In fact, multiple optimized algorithms and analytical capabilities are provided by the Mahout Library. For example, it provides libraries for classification, clustering (such as Mean Shift, fuzzy K-means, and K-means), text mining and frequent pattern mining (for assigning contextual data and scanning text), and collaborative filtering (for comparisons and predictions). Facebook, Yahoo, Twitter, Amazon, International Business Machine (IBM), and Google are the various companies that have implemented scalable machine learning algorithms [19]. It converts machine learning tasks expressed in Java into MapReduce jobs similar to Apache Hive that supports an interface similar to SQL to query data in Hadoop distributed file system.

6.12 Zookeeper

Zookeeper is an open source service planned to manage clusters and applications within the environment of Hadoop. Zookeeper offers many advantages. For example, it props availability of data and lofty performance it also makes simpler distributed programming and guarantees consistent distributed storage. Zookeeper supports Application Programming Interfaces (APIs) for C-based programs and JAVA and is implemented in Java. It is a distributed application which is based on client-server architecture. Its server can execute crosswise many clusters. The file system structure of Zookeeper reflects typical file system tree architectures. Zookeeper also enables for distributed systems to implement scalable, reliable and fast cluster coordination services via its easy interface. To give high availability for the Resource Manager, Zookeeper is used increasingly inside Hadoop. Zookeeper is also used in HBase to make sure coordination, servers management and bootstrapping. It can be used outside the Hadoop platform, not like other components. Yahoo, Twitter, and other companies used ZooKeeper in their distributed systems for locking, sharing, configuration management, and other functions. ZooKeeper is also used by Apache Flume and In IBM's Big Insights[19].

6.13 Avro

A framework for creating Remote Procedure Calls (RPC), serializing and modeling is Apache Avro [100]. To make easy data-intensive applications, Avro specifies a fast and compact binary data format and gives prop for this format in a diversity of programming languages, for instance, Scala, C, Java, Python and C++ [20]. Avro assures effective data storages and compression at different

Hadoop nodes. It passes data from one language or program to other in Hadoop, for example from C to Pig. It is compatible with scripting languages, as data is saved by its schema [19]. The alike functionality of systems is offering by Avro, for example, Protocol Buffers, Thrift and so on. Nonetheless, It is different from those systems by assuring: (a) unlabeled data (b) as soon as a schema alters, assigned field Ids automatically and (c) Typing dynamically (processing of data without the generation of code) [19].

6.14 Oozie

Apache Oozie[101] is a workflow scheduler system planned to execute and supervise jobs in Hadoop clusters. Oozie is a scalable, dependable and extensible managing system that can effectively handle the running of a huge number of workflows. The workflow jobs obtain the shape of a Directed Acyclical Graphs (DAGs). It can prop different sorts of Hadoop jobs comprising Hive, Distcp jobs, Sqoop, Pig and MapReduce[101]. Oozie server is one of the key components of Oozie which is founded on two (02) key components: a coordinator engine which executes repeated workflow jobs triggered by a predefined schedule, and a workflow engine which saves and executes diverse kinds of workflow jobs [20]. The workflows execution is tracking by Oozie. Oozie can be customized by users to inform the client regarding the execution status and workflow through Http call backs for example workflow exits or enters an action node, the workflow is complete. Presently, by default derby is supported by Oozie besides other databases, For instance, Oracle, PostgreSQL, MySQL, and HSQL. A set of APIs library and a CLI, that is based on a client component, is provided by Oozie[99].Table 1 sum up the functionality of the different Hadoop components described above.

Table 1. Hadoop components and their Functionalities

Hadoop Component	Functionalities
HDFS	Offers a reliable and commercial storage capacity, Replicate the data.
MapReduce	Data-parallel processing model allows mass scalability and fault tolerance Nonrelational distributed
HBASE	database, Fast read/write access, designed for low-latency operations.
HCatalog	Facilitates interoperability across data processing tools, Metadata.
Pig	A high-level Scripting language, supporting parallel execution.
Flume	Planned to aggregate, Collect and transmit data from exterior machines to HDFS.
Hive	Data Warehouse system; represent data in a structured database, SQL.
Oozie	Workflow scheduler system planned to execute and control jobs.
ZooKeeper	Coordinate applications and clusters.
Chukwa	Data collection system monitor large distributed systems.
Avro	Structure for creating RPC, serializing and modeling
Mahout	Machine learning software library.

6.15 Hadoop usage

Hadoop is generally used as a part of industrialized applications by Big Data, consisting of clickstream analysis, network searching, social recommendation, and spam filtering. Since June 2012, Yahoo has run Hadoop in Forty Two Thousand (42,000) servers at four (04) data centers delivers its services and products, for example, spam searching and filtering.[82]. At the moment, the biggest Hadoop cluster includes Four Thousand (4,000) nodes that will likely to be improved to Ten Thousand (10,000) by the release of Hadoop 2.0 [53]. Concurrently, Facebook reported that their Hadoop cluster processed one hundred (100) PB of data, that daily expended at the speed of 0.5 PB, since November 2012 [82]. As indicated by Wiki, 2013, some renowned companies and agencies as well use Hadoop to hold distributed computations [Wiki, 2013]. Moreover, different organizations run Hadoop commercially to facilitate, including IBM, MapR, Oracle, Cloudera, and EMC. About 94% of clients can analyze huge volumes of data using Hadoop while 88% of clients thoroughly analyze data and 82% of clients could preserve additional data[82]. Despite the fact that Hadoop has different projects every organization applies a particular Hadoop product as per its requirements. Consequently, Facebook stores one hundred PB (100 PB) of both unstructured and structured data using Hadoop. On the other hand, IBM principally plans to create a platform of Hadoop that is efficient extremely scalable, user-friendly and accessible. Table 2 shows the particular use of Hadoop in different top industries of the world and their functions.

Table 2. Hadoop Usage

Companies	Specific usage
Facebook	Use for Data warehouse, Log file processing and as a medium for machine learning and reporting.
Twitter	Twitter use Hadoop since 2010 to save and for the processing of tweets and Log files.
LinkedIn	The flow of data via Hadoop clusters, the transaction of the stored log file used by business analytics for the analysis of data.
Alibaba	Use for the analysis of vertical search engine.
eBay	Use for browsing the products, Searching and Research purpose.
Yahoo	Use for Searching, Scaling test and Log file processing.
Eyelike, New York Times	Use for Images and videos Analysis

6.16 Hadoop Limitations

Using Hadoop, very huge amounts of data with either changing structure or no structure whatever can be analyzed, processed or handled. Nevertheless, Hadoop also contains a few shortcomings as mentioned below.

- **The Generation of Multiple Copies of Big Data:**

Data is copied in multiples because HDFS was developed for efficiency. In general, data are produced in triplicate at least. Nonetheless, six copies should be produced to maintain efficiency via data locality. Consequently, the Big Data is additionally increased in size [82].

- **Challenging structure:**

The MapReduce structure is complex in particular when complicated transformational logic should be used to get as much advantage as possible. To make simple these framework efforts have been made by open-source modules, but these modules also use registered languages [82].

- **SQL Support is extremely Limited**

Hadoop puts together programming frameworks and open-source projects crosswise a distributed system. As a result, it obtains restricted prop for SQL and be short of fundamental SQL functions, like grouping by analytics and sub-queries.

- **Require Necessary Skills**

Challenging data mining libraries are applied erratically as a component of the Hadoop project. Hence, with respect to distributed MapReduce, knowledge about algorithm and ability to improve are essential.

- **Inefficient Execution**

HDFS doesn't think query optimizers. As a result, it can't run an effectual cost-based scheme. Thus, the volumes of Hadoop clusters are mostly considerably bigger than required for a parallel database.

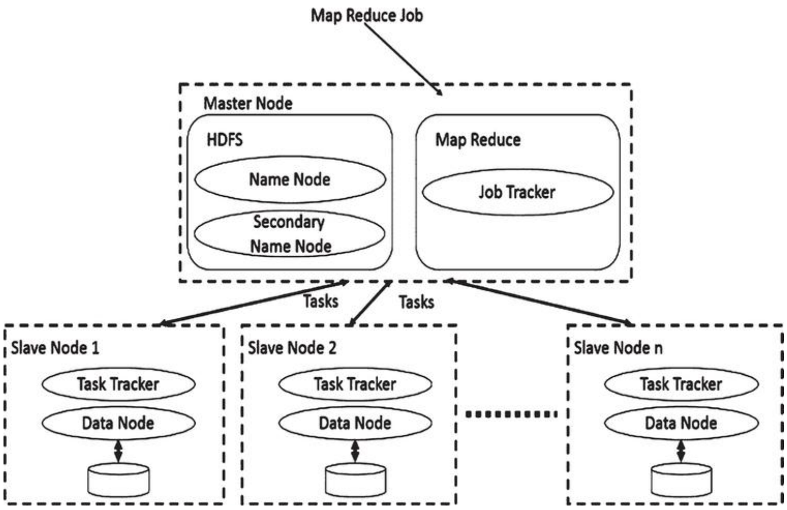

Figure 5. System architectures of MapReduce and HDFS [77]

7. Conclusion

Data is king in today's rising digital world. Data are created at a spectacular speed, daily more than 294 billion electronic mails are sent and each day above one billion different Google searches are performed. The increase and significance of big data have not been minimized till now, And, it probably should not be in the future too. For common people managing, storing and analyzing these data is difficult. Especially data privacy and security are a big challenge and a lot of research is needed to make data secure and private. This has got the shape of a gigantic challenge for businesses in 2017 and, in the future, we expect the same. As a new sign of scientific revolution is just going to commence so Big Data is the coming front line for production, novelty, and competition. This paper reflects different research problems, Tools and Challenges used to analyze these big data. In this analysis, we give a brief picture of the different issues of big data, consisting of big data challenges and opportunities, existing technologies and techniques. We are certain that soon we shall be able to get several types of development and advancement in those fields. No doubt, big data issues will seek solutions from the advancements made today and in future as well. It is crystal clear that the analysis of big data is still in the early phase of progress, as current big data tools and techniques are confined only to the full solution of the real Big Data issues. Hence, extra-scientific reserves from both public and private sectors must be transferred into this scientific model to extract enormous knowledge from big data. From software to hardware, we imminently need additional sophisticated storage and I/O techniques, more professional data-intensive techniques (social, biological and cloud computing etc.), further advance technologies (Big Data platforms with a sound approach, infrastructure, properties, and architecture), and more favorable computer architectures. big data also implies big systems, big revenue, and big challenges, thus additional research efforts in these sub-areas are essential to solve it. We are lucky to see the birth and improvement of big data, and no one can resolve it on its own ends. Human channels, capital reserves, and novel thoughts are the basic ingredients for the improvement of big data.

References

[1] K. Ahmed, *et al.*, "Security and Privacy in Big Data Environment," 2018.

[2] M. Cox and D. Ellsworth, "Managing big data for scientific visualization," in *ACM Siggraph*, 1997, pp. 21-38.

[3] J. Manyika, *et al.*, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[4] S. Lohr, "The origins of 'Big Data': An etymological detective story," *New York Times,* vol. 1, 2013.

[5] N. Dedić and C. Stanier, "Towards differentiating business intelligence, big data, data analytics and knowledge discovery," in *International Conference on Enterprise Resource Planning Systems*, 2016, pp. 114-122.

[6] F. Diebold, "A Personal Perspective on the Origin (s) and Development of'Big Data': The Phenomenon, the Term, and the Discipline, Second Version," 2012.

[7] Y. Wang, "The Challenges and Promises of Big Data—An Engineering Perspective," in *International Workshop of Advanced Manufacturing and Automation*, 2017, pp. 599-604.

[8] M. Group, "World Internet Users Statistics and 2015 World Population Stats," ed, 2015.

[9] D. Che, *et al.*, "From big data to big data mining: challenges, issues, and opportunities," in *International Conference on Database Systems for Advanced Applications*, 2013, pp. 1-15.

[10] P. Zikopoulos and C. Eaton, "D. deRoos, T. Deutsch and G. Lapis (2012). Understanding big data. Analytics for enterprise class hadoop and streaming data," ed: New York, McGraw-Hill.

[11] Y. D. Seo and J. H. Ahn, "Hadoop-Based Integrated Monitoring Platform for Risk Prediction Using Big Data," in *Applied Mechanics and Materials*, 2016, pp. 113-117.

[12] S. Ghosh and S. Kumar, "Video popularity distribution and propagation in social networks," *Int. J. Emerg. Trends Technol. Comput. Sci.(IJETTCS),* vol. 6, pp. 001-005, 2017.

[13] B. E. Şakar, *et al.*, "Global Journal of Information Technology," *Information Technology,* vol. 6, pp. 94-106, 2015.

[14] M. Khoso, "How Much Data is Produced Every Day?," *Northeastern University,* 13 May 2016 2016.

[15] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, 2013, pp. 42-47.

[16] A. K. Tiwari, *et al.*, "A review on Big Data and its security," in *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*, 2015, pp. 1-5.

[17] A. T. Stephen, "The role of digital and social media marketing in consumer behavior," *Current Opinion in Psychology,* vol. 10, pp. 17-21, 2016.

[18] A. Bhagat, "Big Data: What's Big About It."

[19] A. Oussous, *et al.*, "Big Data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences,* 2017.

[20] Y. Demchenko, *et al.*, "Addressing big data issues in scientific data infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, 2013, pp. 48-55.

[21] B. Furht and F. Villanustre, *Big data technologies and applications*: Springer, 2016.

[22] P. Zikopoulos, *et al.*, *Harness the power of big data: The IBM big data platform*: McGraw-Hill New York, NY, 2013.

[23] M. Virk and V. Chauhan, "Big Data and Shipping-managing vessel performance," *JOIV: International Journal on Informatics Visualization,* vol. 2, pp. 73-75, 2018.

[24] I. A. T. Hashem, *et al.*, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems,* vol. 47, pp. 98-115, 2015.

[25] D. Reinsel, *et al.*, "Data age 2025: The evolution of data to life-critical don't focus on big data," *Focus on the Data That's Big Sponsored by Seagate The*

Evolution of Data to Life-Critical Don’t Focus on Big Data, 2017.

[26] D. K. Srivastava, "Big challenges in Big Data research," *Data mining and knowledge engineering,* vol. 6, pp. 282-286, 2014.

[27] J. Cano, "The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity," *xsi,* vol. 1st, 11 march 2014.

[28] D. E. O'Leary, "Artificial intelligence and big data," *IEEE Intelligent Systems,* vol. 28, pp. 96-99, 2013.

[29] E. Onukwugha, "Big data and its role in health economics and outcomes research: a collection of perspectives on data sources, measurement, and analysis," ed: Springer, 2016.

[30] A. Kleusberg and P. J. Teunissen, *GPS for Geodesy*: Springer Berlin, 1996.

[31] [31] M. Dave and J. Kamal, "Identifying big data dimensions and structure," in *Signal Processing, Computing and Control (ISPCC), 2017 4th International Conference on*, 2017, pp. 163-168.

[32] G. Firican, "The 10 Vs of Big Data," *upside,* vol. 1st, 2017.

[33] G. Press, "Cleaning big data: Most time-consuming, least enjoyable data science task, survey says," *Forbes,* 2016.

[34] D. Woods, "Understanding the Value of In-Memory Technology: A Comparative Approach," *Forbes article dated Feb,* vol. 28, 2012.

[35] Q. Fu*, et al.*, "A big data processing methods for visualization," in *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, 2014, pp. 571-575.

[36] N. Ramsay and T. Wampler, "Visualization and interaction with financial data using sunburst visualization," ed: Google Patents, 2015.

[37] J. Carriere and R. Kazman, "Interacting with huge hierarchies: Beyond cone trees," in *Information Visualization, 1995. Proceedings.*, 1995, pp. 74-81.

[38] A. Inselberg, "Parallel coordinates," in *Encyclopedia of Database Systems*, ed: Springer, 2009, pp. 2018-2024.

[39] H. Xia*, et al.*, "A modified ant-based text clustering algorithm with semantic similarity measure," *Journal of systems science and systems engineering,* vol. 15, pp. 474-492, 2006.

[40] S. M. Weiss*, et al.*, *Text mining: predictive methods for analyzing unstructured information*: Springer Science & Business Media, 2010.

[41] V. N. Inukollu*, et al.*, "Security issues associated with big data in cloud computing," *International Journal of Network Security & Its Applications,* vol. 6, p. 45, 2014.

[42] R. E. Bryant, "Data-intensive supercomputing: The case for DISC," 2007.

[43] G.-H. Kim*, et al.*, "Big-data applications in the government sector," *Communications of the ACM,* vol. 57, pp. 78-85, 2014.

[44] A. Szalay, "Extreme data-intensive scientific computing," *Computing in Science & Engineering,* vol. 13, pp. 34-41, 2011.

[45] R. E. Bryant, "Data-intensive scalable computing for scientific applications," *Computing in Science & Engineering,* vol. 13, pp. 25-33, 2011.

[46] P. K. Akalın, "Introduction to bioinformatics," *Molecular nutrition & food research,* vol. 50, pp. 610-619, 2006.

[47] F.-Y. Wang*, et al.*, "Social computing: From social informatics to social intelligence," *IEEE Intelligent systems,* vol. 22, 2007.

[48] J. McDermott*, et al.*, *Computational systems biology*: Springer, 2009.

[49] D. Howe*, et al.*, "Big data: The future of biocuration," *Nature,* vol. 455, p. 47, 2008.

[50] C. Perera*, et al.*, "A survey on internet of things from industrial market perspective," *IEEE Access,* vol. 2, pp. 1660-1679, 2014.

[51] D. P. Acharjya and P. Kauser Ahmed, "A survey on big data analytics: challenges, open research issues and tools," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 7, pp. 511-518, 2016.

[52] R. P. Dameri and C. Rosenthal-Sabroux, "Smart city and value creation," in *Smart City*, ed: Springer, 2014, pp. 1-12.

[53] M. Chen*, et al.*, "Big data: A survey," *Mobile networks and applications,* vol. 19, pp. 171-209, 2014.

[54] E.-R. Info, "What is RFID," *www. epc-rfid. info/rfid,* 2014.

[55] M. Hilbert and P. López, "The world’s technological capacity to store, communicate, and compute information," *science,* vol. 332, pp. 60-65, 2011.

[56] W. Worlton, "Bulk storage requirements in large-scale scientific calculations," *IEEE Transactions on Magnetics,* vol. 7, pp. 830-833, 1971.

[57] S. F. Oliveira*, et al.*, "Trends in computation, communication and storage and the consequences for data-intensive science," in *High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on*, 2012, pp. 572-579.

[58] V. Kasavajhala, "Solid state drive vs. hard disk drive price and performance study," *Proc. Dell Tech. White Paper,* pp. 8-9, 2011.

[59] D. Leong, "A new revolution in enterprise storage architecture," *IEEE Potentials,* vol. 28, 2009.

[60] M. A. Mohamed*, et al.*, "Relational vs. nosql databases: A survey," *International Journal of Computer and Information Technology,* vol. 3, pp. 598-601, 2014.

[61] Q. Wang*, et al.*, "Enabling public auditability and data dynamics for storage security in cloud computing," *IEEE transactions on parallel and distributed systems,* vol. 22, pp. 847-859, 2011.

[62] Q. Wang*, et al.*, "Dependable and secure sensor data storage with dynamic integrity assurance," *ACM Transactions on Sensor Networks (TOSN),* vol. 8, p. 9, 2011.

[63] A. Oprea*, et al.*, "Space-Efficient Block Storage Integrity," in *NDSS*, 2005.

[64] S. Wu, "A review on coarse warranty data and analysis," *Reliability Engineering & System Safety,* vol. 114, pp. 1-11, 2013.

[65] J. Nahar*, et al.*, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications,* vol. 40, pp. 96-104, 2013.

[66] L. Zhou and H. Fujita, "Posterior probability based ensemble strategy using optimizing decision directed acyclic graph for multi-class classification," *Information Sciences,* vol. 400, pp. 142-156, 2017.

[67] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233-246.

[68] C. Yang*, et al.*, "Big Data and cloud computing: innovation opportunities and challenges," *International Journal of Digital Earth,* vol. 10, pp. 13-53, 2017.

[69] D. Agrawal*, et al.*, "Challenges and Opportunities with big data 2011-1," 2011.

[70] J. Han*, et al.*, *Data mining: concepts and techniques*: Elsevier, 2011.

[71] M. A Vouk, "Cloud computing–issues, research and implementations," *Journal of computing and information technology,* vol. 16, pp. 235-246, 2008.

[72] A. Adamov, "Distributed file system as a basis of data-intensive computing," in *Application of Information and Communication Technologies (AICT), 2012 6th International Conference on*, 2012, pp. 1-3.

[73] P. Zhang*, et al.*, "A survey on quality assurance techniques for big data applications," in *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, 2017, pp. 313-319.

[74] F. Khan*, et al.*, *Khan et al. 2014*, 2015.

[75] D. Zhang, "Big data security and privacy protection," in *8th International Conference on Management and Computer Science (ICMCS 2018)*, 2018.

[76] J. Pavolotsky, "Privacy in the age of big data," *The Business Lawyer,* vol. 69, pp. 217-225, 2013.

[77] M. E. Smid and D. K. Branstad, *Data Encryption Standard: past and future* vol. 76, 1988.

[78] S. Simoff*, et al.*, *Visual data mining: theory, techniques and tools for visual analytics* vol. 4404: Springer Science & Business Media, 2008.

[79] D. A. Keim*, et al.*, "Visual data mining in large geospatial point sets," *IEEE Computer Graphics and Applications,* vol. 24, pp. 36-44, 2004.

[80] D. G. Murray, *Tableau your data!: fast and easy visual analysis with tableau software*: John Wiley & Sons, 2013.

[81] J. Heer*, et al.*, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE transactions on visualization and computer graphics,* vol. 14, 2008.

[82] N. Khan*, et al.*, "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal,* vol. 2014, 2014.

[83] A. Hadoop, "Welcome to apache hadoop," *Welcome to Apache Hadoop,* 2016.

[84] S. c. Media, "Hadoop," 2011.

[85] D. Borthakur, "HDFS architecture guide," *Hadoop Apache Project,* vol. 53, 2008.

[86] D. Carstoiu*, et al.*, "Hadoop hbase-0.20. 2 performance evaluation," in *New Trends in Information Science and Service Science (NISS), 2010 4th International Conference on*, 2010, pp. 84-87.

[87] N. Dimiduk*, et al.*, *HBase in action*: Manning Shelter Island, 2013.

[88] D. Miner and A. Shook, *MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems*: " O'Reilly Media, Inc.", 2012.

[89] H. Jiang*, et al.*, "Scaling up MapReduce-based big data processing on multi-GPU systems," *Cluster Computing,* vol. 18, pp. 369-383, 2015.

[90] A. Azzini and P. Ceravolo, "Consistent process mining over big data triple stores," in *Big Data (BigData Congress), 2013 IEEE International Congress on*, 2013, pp. 54-61.

[91] A. O'Driscoll*, et al.*, "'Big data', Hadoop and cloud computing in genomics," *Journal of biomedical informatics,* vol. 46, pp. 774-781, 2013.

[92] Y. Wang*, et al.*, "MtMR: Ensuring MapReduce Computation Integrity with Merkle Tree-based Verifications," *IEEE Transactions on Big Data,* vol. 4, pp. 418-431, 2016.

[93] T. Lindholm*, et al.*, *The Java virtual machine specification*: Pearson Education, 2014.

[94] S. Mazumder, "Big data tools and platforms," in *Big Data Concepts, Theories, and Applications*, ed: Springer, 2016, pp. 29-128.

[95] K. Krishnan, *Data warehousing in the age of big data*: Newnes, 2013.

[96] K. S. Beyer*, et al.*, "Jaql: A scripting language for large scale semistructured data analysis," in *Proceedings of VLDB conference*, 2011.

[97] D. Vohra, "Using apache sqoop," in *Pro Docker*, ed: Springer, 2016, pp. 151-183.

[98] R. Shireesha and S. Bhutada, "A study of tools, techniques, and trends for big data analytics," *IJACTA,* vol. 4, pp. 152-158, 2016.

[99] A. Oussous*, et al.*, "Computer and Information Sciences," 2017.

[100] M. Grover*, et al.*, *Hadoop Application Architectures: Designing Real-World Big Data Applications*: " O'Reilly Media, Inc.", 2015.

[101] M. K. Islam and A. Srinivasan, *Apache Oozie: The Workflow Scheduler for Hadoop*: " O'Reilly Media, Inc.", 2015.