International Journal of Operations Research and Information Systems
Volume 11 • Issue 3 • July-September 2020

Big Data and Cloud Computing:
A Critical Review

Akansha Gautam, Department of Computer Science, University of Delhi, Delhi, India

Indranath Chatterjee, Department of Computer Science and Engineering, J.K. Lakshmipat University, Jaipur, India & Department of Computer Engineering, Tongmyong University, Busan, South Korea

https://orcid.org/0000-0001-9242-8888

ABSTRACT

With evolving technology, a huge data is being generated from everywhere in various forms. The driving factors for the evolution of data, such as retail, media, banking, healthcare, and education, leads to a very large and complex collection of data popularly coined as big data. Handling, management, and analysis of big data seem to be a complicated process. Utilising cloud environment for analysing big data is a recent research trend. Big data analytics can provide cost-effective ways to analyse information quickly and helps in decision making, improvement in services or products. This paper aims to critically review the literature to find current issues and research gaps. This study illustrates the existing solutions and methods provided for big data and its rise in cloud computing technology. Furthermore, this paper throws light on the open research challenges in this domain, stating the scope of future work.

KEYWORDS

Apache Spark, Big Data, Big Data Analytics, Cloud Computing, Data Mining, Hadoop

1. INTRODUCTION

Data being the raw material for information is a collection of symbols, text, image, video or any other form. It is further processed by applying certain operations or set of instructions for removing all types of ambiguity to retrieve concrete information. Data can be of any type like structured, unstructured and semi-structured that can be generated by both humans and machines. Structured data is one that can be stored, accessed and processed in a well-defined structured format. A student or an employee table in a database is an example of structured data. Unstructured data is present in an unknown form and has an unstructured classification - for example, a combination of images, text files, audio, videos, etc. And the last one is a semi-structured type of data. This form can be in a structured and unstructured format as well. Data represented in an XML file, JSON documents, and NoSQL databases are examples of a semi-structured state.

Big data is a concept or form of data having a very large volume, which keeps growing exponentially with the time. It deals with the enormous and complex data sets which cannot be processed or managed with the conventional data processing software or any other tools. Big data

DOI: 10.4018/IJORIS.2020070102

analytics has the potential to transform and provide insight to every business or field. Earlier, data was stored in Megabytes or Gigabytes, but today, a large amount of data is producing ceaselessly in terms of Petabytes (PB) or Zettabytes (ZB) which eventually requires huge storage area and management.

According to some facts about data generation, videos of 300 hours in length are uploaded to YouTube in every minute and views performed every single day are almost 5 billion (Sagiroglu and Sinanc 2013). Each day, 400 hundred million tweets are sent (Jannapureddy et al. 2019). On average, people post about 500 million tweets per day (Sagiroglu and Sinanc 2013). Data processed by Google is in terms of petabytes (Chen, Mao, and Liu 2014). Over 10 PB of log data is generated by Facebook every month (Chen, Mao, and Liu 2014). Within the next decade, an increase in the amount of information will be around 50 times of current number and the information technology specialists' number will go up by 1.5 times (Sagiroglu and Sinanc 2013). Ninety percent of the whole data currently available is the amount of big data created from its different sources in the last two years (Jannapureddy et al. 2019). Social media, banks, business processes, web servers, instruments, websites, stock markets, emails, health records, medical image data (Chatterjee et al. 2018), search queries, logs, sensors, scientific data, online transactions, videos, audios, images, financial services, retail, text document, photography, mobile phones are some of big data sources in this electronic era. Big data trend is leading in many areas such as improving healthcare and public health, understanding and targeting customers, understanding and optimizing business processes and many other fields like improving science and research, optimizing machine and device performance, improving security and law enforcement.

The significance of big data does not spin around the size of the data. Retrieval of useful and accurate information by processing and analysing the data for better outcomes, smart decision making, early detection of errors, cost, and time reductions make it significant. To manage it precisely, we need some other platform. One of the optimal solutions to this big data problem, we will be addressing in this paper is the 'Hadoop' cluster and Hadoop Distributed File System (HDFS) as its storage platform on 'cloud' platform to manipulate the data. Hadoop can be used to manage the big data using cloud servers, but without Hadoop, cloud alone cannot handle big data.

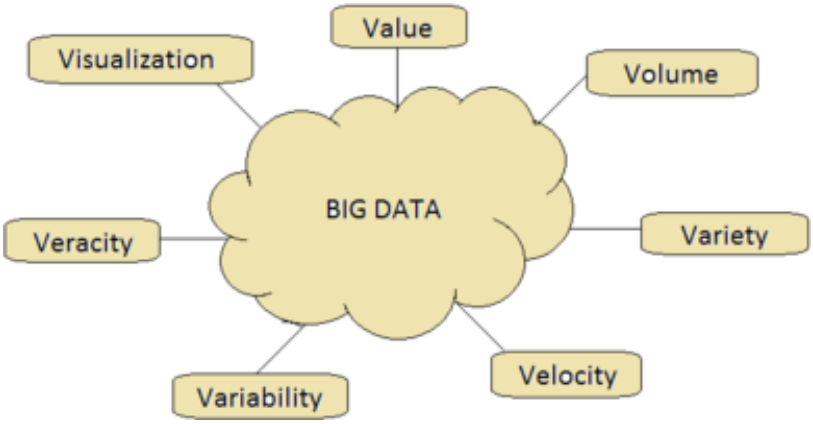
This review paper provides an elaborative overview of the concepts, including big data, cloud computing, Apache Hadoop, and ApacheSpark. Section II condenses the conceptual overview of fundamental terms; Section III illustrates the literature review work done on the mentioned topics until now. Section IV following presents the discussion part where we mention the gaps that we observed during literature review work and propose the open research issues. Finally, the paper concludes by mentioning the inference of this study and the scope of future works.

2. CONCEPTUAL OVERVIEW

2.1. Big Data

Big data can be commonly characterized as 7 V's – Volume, Variety, Velocity, Variability, Veracity, Visualisation, and Value (as shown in Figure 1). The figure depicts seven dimensions of big data to determine which kind of data is hard to process. Volume defines the size of the data, which is enormous undoubtedly. Earlier, it used to be measured in Gigabytes but now is measured in Zettabytes (ZB) or even in Yottabytes (YB). Variety means the different forms of data like a structured, unstructured, semi-structured and mixed form of acquired data. Velocity describes the speed at which the data is being generated or processed. Variability is different from the variety. Variability focuses on properly understanding and interpreting the exact meanings of raw data depending on its context. Big data always exists in one kind of a state. These states can be clear, consistent, connect, or confirmed state. This is represented by an entity known as veracity. Visualisation plays a vital role in today's world. It refers to the representation of a large amount of complex data in a much more effective and meaningful way using charts and graphs instead of long excel files with rows and columns of data

Figure 1. Seven V's of big data



or word doc. Data can be easily readable in graphs or charts rather in spreadsheets. After addressing volume, variety, velocity, variability, veracity, visualisation factors, we make sure data provides an immense value if correctly processed.

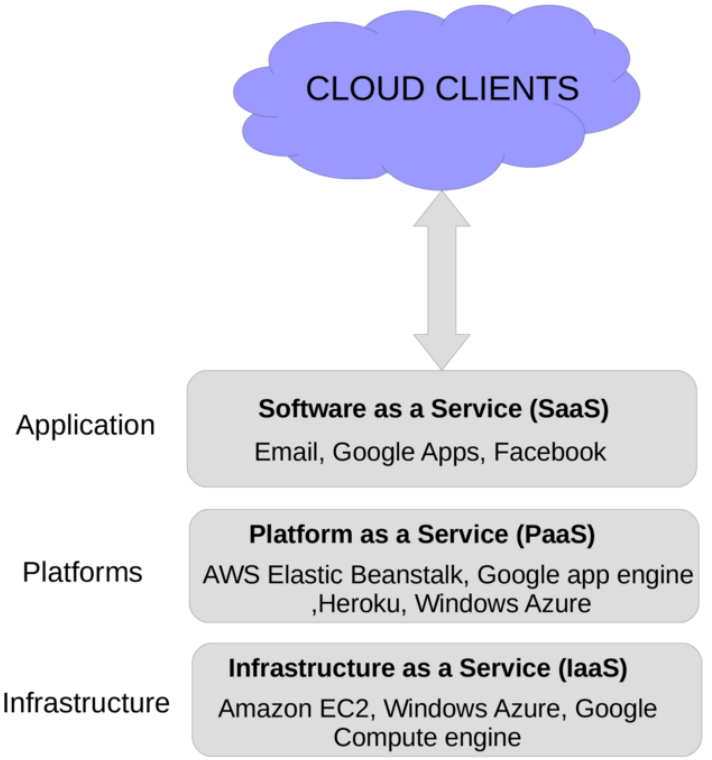
2.2. Cloud Computing

Cloud computing is a concept of using servers on the internet at different remote locations to store, manage and process the data instead of doing it on a local server or on a personal computer. Data management can be done using databases and process it by renting a server that has larger processing power so that work can be done faster, rather than doing it on a local computer which could be slow. On the other hand, for doing large batch processing work, a server can be bought with large processing power, but it is useless after the work is done. Cloud computing is all about delivering hosted services over the internet. So, all of these problems can be solved with cloud computing. Cloud computing is further divided into cloud models. There are two types of models which are service models and deployment models. Service models mean the kind of service that the cloud offers. These services can be divided into further modules as IaaS (Infrastructure as a service), PaaS (Platform as a service) and SaaS (Software as a service). Deployment models state how applications can be deployed or uploaded on the cloud. It is further divided into public cloud, private cloud, community cloud, and hybrid cloud.

IaaS is providing virtualised computing resources to customers over the internet without worrying about the underlying physical machine as everything is managed by the cloud provider. It abstracts the usage of a physical machine. Amazon EC2, Google Compute Engine, Windows Azure provides IaaS. In PaaS, customers do not have control over the underlying architecture, including operating systems, storage, servers, etc. This service provides access to the only user interface, i.e., application development platforms and tools. Example of such PaaS providers is AWS Elastic Beanstalk, Google app engine, Windows Azure and Heroku. In SaaS, neither any server nor any user interface is provided but only software. Cloud providers lease applications or software which is owned by them to their clients where customers can use the software as they want. Figure 2 shows the broad level architecture of data-related cloud computing framework. The figure manifests three types of cloud computing services along with few examples of service providers of each service.

The public model is sharing a server with a host of different people over the internet. A private cloud can act as a standalone system where the server is given with private access, and only the customer's operating system is deployed on it and data is entirely safe. Hybrid cloud is a mixture of both public cloud and private cloud, for example, if in a data science company, a website has to be the

Figure 2. Cloud computing architecture



host. Sharing a common server and keeping the data confidential at the same time is a situation where a hybrid cloud can be used. In a community cloud, architecture is shared by a group of individuals or organisations. Figure 3 shows the different deployment models of cloud computing. It summarises the number of ways in which the cloud can be provided to us.

Cloud computing has several benefits over accessing personal or expensive systems. In cloud computing technology, customers pay for what they use whereas on personal computers or single organisation establishments, the pay is higher, and scalability is less. Migration flexibility service is also available in cloud computing where organisations can move their workloads to or from the cloud. Cloud service providers help their clients to access and share the data from anywhere over the internet without worrying about the server space requirement or experts to handle the hardware and software.

2.3. Hadoop

Hadoop, developed by Doug Cutting (Jannapureddy et al. 2019), is an open-source Java-based framework that allows distributed processing of large data sets on the cluster of commodity hardware. Commodity hardware means one need not purchase expensive servers but can work on simple and cheap servers. Hadoop mainly consists of a storage part known as Hadoop Distributed File System (HDFS) and a processing part, which is the MapReduce programming model (Patel, Birla, and Nair 2012). Figure 4 shows the Hadoop file system. The figure signifies two main components: HDFS as a distributed storage system and MapReduce as data processing framework in Hadoop. NameNode and Secondary Node daemon act as Master node and DataNode daemon is a Slave node in HDFS. Job Tracker is a master process and task tracker is a slave process of MapReduce engine. There can be multiple DataNode and Task tracker daemons. Hadoop uses distributed storage as well as distributed processing. Its ecosystem includes various components like Sqoop, Zookeeper, Hive, Pig, Hadoop

Figure 3. Cloud computing deployment models

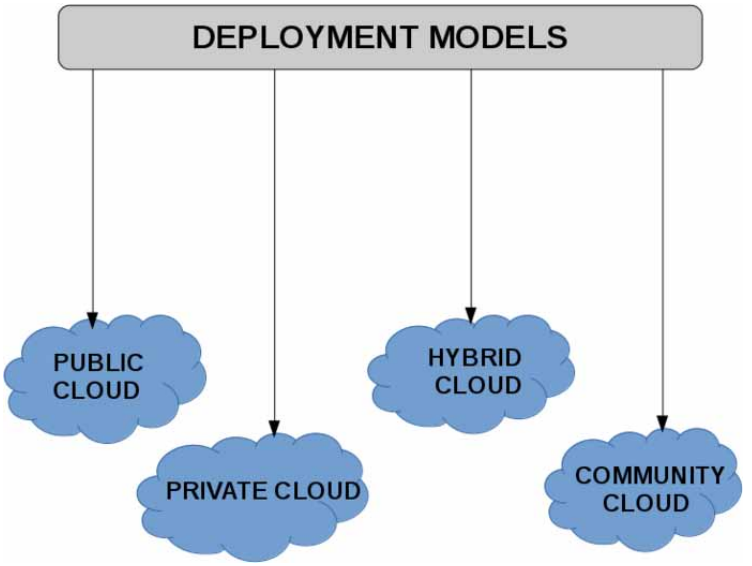
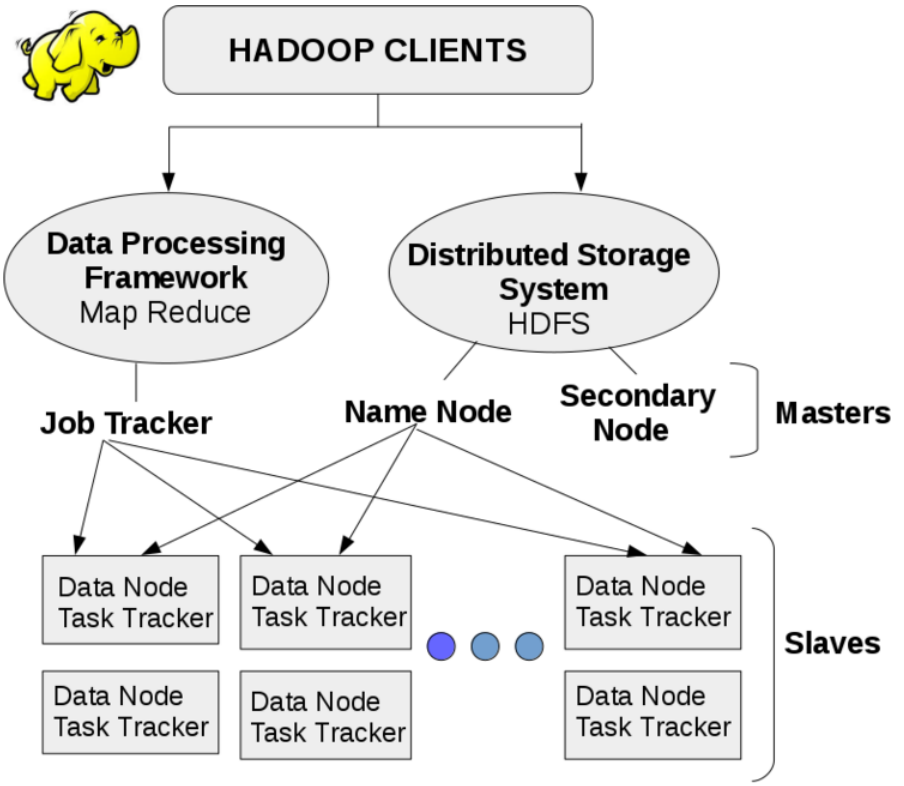


Figure 4. Hadoop cluster architecture



Distributed File System (HDFS), HBase, and MapReduce (Devi and Kasireddy 2019). It is a data management tool and uses scale-out a storage facility that means nodes can be added at any point, whenever storage seems to be less. Hadoop cluster is a group of systems having Hadoop installed and connected to each other. Data size is the most important factor while defining a Hadoop cluster. Hadoop has a master-slave architecture which means only one system will act as a master daemon which will have full control and other nodes will be known as slave daemons. In Hadoop, the place where all the data is stored can be called as Hadoop Distributed File System (HDFS).

2.4. Apache Spark

Apache Spark is an open-source, scalable, massively parallel in-memory data processing engine for any analytics application. To process big data, like MapReduce, spark also distributes the data on the cluster, and that data is further processed parallelly. It offers real-time computation and low latency because of in-memory computations. High-level APIs are provided in Java, Scala, Python, SQL, and R by Apache Spark. Spark code can be written in any of these four languages. Apache Spark provides in-memory data processing that eventually makes it much faster. It is a hundred times faster in memory and ten times faster on the disk comparing with Hadoop.

3. LITERATURE REVIEW

This section provides a deep review of various research works done so far in the domain of cloud computing, big data, and Hadoop.

Zhang et al. (2010) studied the concept of cloud computing and addressed some research challenges in this field. They discussed various attractive features of cloud computing, which includes no up-front investment, highly scalable, lowering operating cost, easy access, reduction in business risks and maintenance expenses. The authors further compared cloud computing with other technologies such as grid computing, utility computing, virtualisation, and autonomic computing. They further described the cloud computing architecture into different layers: Hardware layer, infrastructure layer, platform layer, and application layer. According to the categories of cloud services, there are three types of clouds which are public cloud, private cloud and hybrid cloud each of its having both benefits and drawbacks. The authors also mentioned another type of cloud that is a virtual private cloud, which is an alternate solution to handle the limitations of both public and private clouds. The authors mentioned that cloud computing has several characteristics like multi-tenancy, shared resource pooling, geo-distribution, and ubiquitous network access, service-oriented, dynamic resource provisioning, self-organizing, and utility-based pricing. The authors further provided a review of technologies used in cloud computing like architectural design of data centers, distributed file systems over clouds, distributed application framework over clouds. They also presented some cloud computing products through a survey, which are Amazon EC2, Microsoft Windows Azure platform, and Google App Engine. Further, authors focused on various research challenges in cloud computing technologies such as automated service provisioning, virtual machine migration, server consolidation, energy management, traffic management and analysis, data security, software frameworks, storage technologies and data management, novel cloud architectures.

Dillon et al. (2010) defined cloud computing technology and pinpointed the challenges and issues related to it. They presented five major elements of cloud computing as on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Service models and deployment models are further identified and discussed in brief. Four deployment models in the cloud community are public cloud, private cloud, community cloud, hybrid cloud. The authors further discussed two computing technologies: service-oriented computing and grid computing technology and articulated their relationships with cloud computing. They also highlighted various challenges that prevent the adoption of cloud computing technology: security, performance, availability, raise in the cost of data communication, a feasible charging model for SaaS provider, service level agreement,

and privacy concerns in moving data on to the cloud. Furthermore, cloud interoperability issue is identified and presented solutions like providing an intermediary layer between cloud consumers and cloud-specific resources, standardisation, open cloud API, SaaS, and PaaS interoperability.

Marston et al. (2011) studied the concept of cloud computing from the perspective of the business. They discussed key advantages such as lowering the cost and easy hardware access. The authors mentioned strengths, weaknesses, opportunities, and threats for the cloud computing industry. They further identified the various issues that can affect the stakeholders of cloud computing like consumers, providers, enablers, and regulators. They provided a set of recommendations for the practitioners to manage this technology and outlined different areas of research in this domain that need attention for the future purpose. They mentioned the scopes to pursue research in identifying the applications to move to the cloud, the cloud providers' strategy, development of standards. Lastly, they focused on key issues facing governmental agencies who will need to get involved in the regulation of cloud computing which includes cloud computing economics, strategy issues, IS policy issues in cloud computing, and regulatory issues.

Subashini and Kavitha (2011) surveyed on security issues in service delivery models of a cloud computing system. Cloud computing offers several advantages like better utilisation of resources, lower costs, scalability, ubiquitous network access and many more but still suffers from some major security risks. They described security threats in Software as a Service (SaaS) model. Data security, data integrity, network security, data locality, data access, data segregation, authentication and authorisation, data confidentiality, availability, backup, web application security, virtualization vulnerability, data breaches, identity management and sign-on process are the key security elements that needs to be considered as a vital part of SaaS application development and deployment process. The authors further mentioned the security risks in Platform as a Service (PaaS) service in which developers are offered a complete software development lifecycle management, from planning to design to build an application to deployment to testing to maintenance. But these advantages itself can help hacker to use the cloud infrastructure for malware attacks. They addressed security issues in Infrastructure as a service (IaaS). IaaS provides a virtual server to the customers and needs to pay only for the resources they use. But IaaS only provides basic security like load balancing, perimeter firewall, etc. A higher level of security is required for the applications moving to the cloud environment. The authors also identified current security solutions presented by various researchers in this field such as to develop a development framework having tight security architecture, resource isolation during data processing, avoid IP spoofing and use encrypted protocols, to use in-house private clouds.

Zissis et al. (2012) studied the concept of cloud computing and addressed all the security issues related to this technology (Jadeja and Modi 2012). They discussed cloud service models: IaaS, PaaS, SaaS, and deployment models of cloud architecture, which are a private cloud, public cloud, hybrid cloud, community cloud. They further identified the important characteristics of cloud computing, which include broad network access, improvement in reliability factor, infrastructure scalability, location independence, flexibility/elasticity, economies of scale, cost-effectiveness and sustainability. The authors further illustrated various security issues of using cloud computing technology: Trust, security identification of threats like confidentiality and privacy, integrity, availability. They introduce the Trusted Third Party (TTP) services, which can assure specific security characteristics like confidentiality, integrity, and authenticity of data within a cloud environment. They claim that a trusted third party can act as an entity that provides secure interaction between the two parties. TTP can be connected to public key infrastructure, which further provides strong authentication, authorisation, data confidentiality, data integrity, non-repudiation. TTP can be relied upon for various characteristics like low and high confidentiality, server and client authentication, creation of security domains, cryptographic separation of data, certificate-based authorisation. This approach uses public key infrastructure, Single-Sign-On technology, and LDAP directories to securely authenticate the entities and deals with the threats in cloud computing such as integrity, confidentiality, authenticity, and availability of data.

Boyd and Crawford (2012) illustrated some critical questions based on the issue of big data. They defined big data in three perspectives that are a cultural, technological and scholarly phenomenon that further rests upon technology that deals with large data sets, and analysis that identifies patterns from the data sets and lastly mythology which shows the belief of getting intelligence and knowledge from large data sets. The authors discussed and interrogated six big data-related questions: 1. big data changes the definition of knowledge; 2. claims to objectivity and accuracy are misleading; 3. bigger data are not always better data; 4. big data loses its meaning if taken out of context; 5. accessibility of data does not make it ethical; 6. limited access to big data creates new digital divides. The authors aimed to interrogate the assumptions, values, and biases of this big data field, which could further help the scholars and researchers as an essential component.

Patel et al. (2012) studied the big data problem and addressed its technologies such as massively parallel processing (MPP) databases, data mining grids, distributed file systems, distributed databases, cloud computing platforms, internet, and scalable storage systems. They defined its optimal solution using the Hadoop cluster, HDFS for storage and parallel processing to process large data sets using Map-reduce programming framework and described their architecture in detail. The authors performed the big data experiments and setup the Hadoop data cluster consisting of four nodes with one node as master daemon and others as slave daemons using HDFS for storage. The first configured and tested a single node cluster. Eight concurrent tasks per server were assigned to run on a cluster. Each MapReduce program is further divided into M number of map tasks, and R number of reduce tasks and both the input and output are stored in HDFS. After the experimental setup, they performed two experiments. The first experiment was processing text by counting several words that occurred in a large data sized document. The second experiment was analysing Earthquake data set based on a U.S geology survey. The experiment was to test whether Hadoop behaves optimally in the two conditions: if the number of nodes is increased or if the size of the data set and nodes are increased. The results from experiments indicated the favorable results of the given approach. They inferred that the Hadoop cluster is scalable and takes less execution in both the aforementioned conditions. They suggested evaluating performance and using Hadoop on cloud platforms as future work.

Vavilapalli et al. (2013) illustrated the design, development, and current state of Apache Hadoop Yarn. Scalability, multi-tenancy, serviceability, locality awareness, high cluster utilization, reliability, secure and auditable operation, Support for Programming Model Diversity (SPMD), flexible resource model, backward compatibility are some of the requirements of Hadoop, all the way to YARN. Hadoop on Demand (HoD) was developed and deployed by Yahoo! to address some of its multi-tenancy issues (Vavilapalli et al. 2013). The authors introduced a new architecture which separated the programming model from resource management infrastructure and assigned many scheduling functions like fault-tolerance to per-application component o address the mentioned requirements. They further defined the main components of the Yarn architecture, which are resource manager, application manager, node manager. They provided experimental evidence to demonstrate the improvements in terms of efficiency by running YARN on production environments and confirmed the flexibility by describing the porting of other programming frameworks like Giraph, Hoya, Hadoop MapReduce, Spark, Storm onto YARN. They also identified various benefits of YARN such as greater scalability, higher efficiency, allows different frameworks to share a cluster efficiently.

Kaisler et al. (2013) discussed the issues and challenges of big data in this paper. They defined volume as a critical characteristic of big data and considered value, velocity, variety, and complexity as other characteristics. Value is created by analysing the big data and retrieving useful information from it. The authors illustrated three major issues while dealing with the big data problem: storage issues, management issues, and processing issues and each of these fields consist of a massive set of research problems. They presented two solutions for storage problems. The first is to process the data and transmit only resultant piece of information, and second is to analyse the data in order and transmit only that data which is evaluative to downstream analysis. Integrity and source of metadata should also be transferred along with the actual data in either of the case. They highlighted that

there is no perfect big data management solution exists yet. This shows an important gap in the research literature that needs to be filled. Processing issues require extensive parallel processing and new algorithms that can help in providing timely and required information. They further discussed various dynamic design challenges in the big data field. These include quality versus quantity, speed versus scale, compliance and security, data input and output processes, data ownership, data growth versus data expansion, unstructured data versus structured data, all data versus some data value, approach of distributed data and distributed processing which includes cloud computing along with the MapReduce programming as one of its solution but it can still have performance degradation problem. The authors further addressed analytical challenges under big data processing which are scaling, finding key data, processing discrete large data points set into high valued data, analysing a large heterogeneous set of data, and lastly modeling the world in order of our desire with abundant of data available in hand.

Sagiroglu and Sinanc (2013) presented a review of big data. They discussed 3 V's of the big data as volume, velocity, and variety. Further elaborating, they defined verification as another component of it. They further derived from a TDWI survey that more attention is required towards big data analysis to optimise decision making. They described the potential of big data in five main topics including the public sector, healthcare, manufacturing, retail and personal location data specified by McKinsey Global Institute (Sagiroglu and Sinanc 2013). The authors illustrated two methods for big data processing. One is a Hadoop cluster using the HDFS and MapReduce framework and the other is a High-Performance Computing Cluster (HPCC) systems platform. They defined differences between the two methods. The authors addressed privacy and security issues.

Wu et al. (2013) proposed a HACE theorem which specifies the features of big data, and suggests a big data processing model, from the perspective of data mining. Here, H stands for heterogeneous data; A is autonomous sources with distributed and decentralized control, C and E indicate the complex and evolving relationships among data. The authors further presented a big data processing framework which consists of three tiers from inside out that are data accessing and computing procedures (tier 1) which focuses on the effective computing platform such as MapReduce or Enterprise control language (ECL) to access such large amount of data, data privacy, and domain knowledge (tier 2) illustrates two most pressing issues. First is data sharing and privacy, and second is domain and application knowledge, which further helps in providing information for designing big data mining algorithms (Kumar and Chatterjee 2016). Tier 3 is big data mining algorithms that concentrate on algorithms that can help in handling the complexity and large volume of big data. They have analysed various challenges in this big data field which are required of high-performance computing platforms. At the data level, variety of data collection environments and information sources can lead to noise and error in data. To avoid this situation, a safe information sharing protocol should be developed. A careful designing of algorithms is required to have the best model out of the big data, and big data mining framework should be more carefully designed to deal with unstructured data.

Katal et al. (2013) discussed issues, challenges, tools and good practices of big data. They illustrate the big data concept and its properties like variety, volume, velocity, variability, complexity, and value. They introduced the importance of big data and its various projects, which play an important role in changing the concept of science into big science. They further mentioned big data challenges and issues such as privacy and security, data access and information sharing, analytical challenges, issues related to storage and processing, skill requirements. Some technical challenges include fault tolerance, quality of data, data scalability and heterogeneous data. They discussed the tools and techniques available for big data handlings such as Hadoop and its components including HDFS and MapReduce, comparison of Hadoop technique with other system techniques like HPC and grid computing tools, volunteer computing technique, and RDBMS and provided MapReduce as the best solution to this problem. They further focused on big data good practices which include generality of technology is required, structured and unstructured both data should be analysed properly, data

quality should be improved, scalability of the data stored should be handled, to reduce processing time, investment in data quality and metadata should be kept in mind.

Chen et al. (2014) surveyed big data and review related technologies such as cloud computing, Hadoop, Internet of Things, data centers. They mentioned four features of big data: volume, variety, velocity, and value. They listed some challenges of big data, which includes data representation, redundancy reduction and data compression, data life cycle management, analytical mechanism, data confidentiality, energy management, expandability, and scalability. The authors further discussed four phases of big data chain, i.e., data generation, data acquisition, data storage, and data analysis. They also focused on big data storage and illustrated essential issues such as large storage systems, distributed storage systems, and big data storage mechanisms. Big data analytic methods, including bloom filter, hashing, index, trial, parallel computing, are discussed in brief. Moreover, different analytical architectures for different application requirements are identified, which are real-time vs offline analysis, analysis at different levels, analysis with different complexity. Tools for big data mining and analysis include R, Excel, Rapid-I Rapidminer, KNMINE, Weka/Pentaho. The authors also highlighted key application fields of big data such as application evolutions, structured data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis, mobile data analysis. They also focused on open issues such as technology development, theoretical research, practical implications, and data security. Future opportunities in this field include handling data of larger scale, higher diversity, and more complex structures, mastering data resource performance, exploring new innovative technologies should be explored in terms of data acquisition, storage, processing, analysis, information security through the development of big data, visualisation of data.

Raghupathi and Raghupathi (2014) illustrated the problem of big data in the healthcare industry. As the industry produces a large amount of data, there should be a methodology to analyse and improve the outcomes and reduce the cost. Health data sets are so complex and large that they are difficult to process by traditional data handling approaches. They defined the four V's of big data healthcare analytics as volume, velocity, variety, and veracity. This large data set includes electronic patient records consisting of patient data; physician's written notes and prescriptions, medical images, laboratory, pharmacy, insurance, and other administrative related data. They have addressed various benefits of effectively using big data in this field like early detection of patient's disease, managing population health, and detection of health- care fraud quickly, predicting patient's length of stay, elective surgery to be chosen by various patients and patients who will not be benefited by the surgery, and various complications. They defined Hadoop and its components like HDFS, MapReduce, HBase, Oozie, Hive, Zookeeper as the most remarkable platform to handle the problem of big data. As the partitioned data sets are allocated to numerous servers (nodes) by it, and each node solves the particular part of larger problems, thus the final result is achieved by combining them together. The authors also focused on various challenges that need to be addressed such as data needs to be menu-driven, user-friendly, transparent, privacy, security enablement, establishing standards and governance, the lag between data collection and processing needs to be fixed.

Gandomi and Haider (2015) defined big data along with their characteristics. They focused on the analytics used for big data, especially related to unstructured data, which constitute, 95% of big data. Along with the three V's of big data, i.e., volume, variety, velocity, other dimensions have also been considered like veracity, variability, and value. The authors presented various big data analytical techniques for both structured and unstructured data. This list includes text analytics which extracts information from textual data, audio analytics which analyse and extract information from unstructured data presented in audio format; video analytics known as video content analysis (VCA) which involves various techniques to monitor and extract data from video streams; social media analytics refers to the structured and unstructured data analysis from social media platforms; and lastly, predictive analysis which predict future outcomes based on historical and current data using variety of techniques. They also highlighted few fields for future research like real-time analytics

because of the growth in location-aware social media and mobile apps. Noise and error in big data can lead to the development of statistical techniques required for mining data.

Stephens et al. (2015) compared the genomics field with other three major generators of big data.: Astronomy, Twitter, and YouTube. These domains are further analysed in terms of acquisition of data, secondly storing the data, distribution of data, and lastly analysing the data that comprises the life cycle of data sets. The authors further discussed the technological needs for big data genomics. For data acquisition, the pairing of genome sequence with automated methods is required so that metadata and phenotype data can be collected from different environments to compare them. They addressed efficient storage technologies like 3-D memory, overcoming the problem of I/O bottleneck and network speed by using integrated computing technologies. They mentioned that algorithmic developments that can represent genomes as a graph could also help in comparing one genome to another. The only solution specified by the authors for genome sequences distribution is the cloud computing technique to deal with a population scale. They stated that technologies like Hadoop and other highly scalable systems are a good start to analyse genomic sequences, mutation changes or any development and evolution but at the same time, these technologies are difficult and expensive to use. They noticed that genomics has various challenges in terms of the acquisition, storage, data distribution and most importantly, data analysis, which needs to be faced in the future.

Hashem et al. (2015) studied the big data concept, its characteristics, classifications and its rise in cloud computing technology. They defined the characteristics like volume, velocity, variety, and value. They further classified the big data in five aspects: various sources of data, content format, storing data, staging, and processing of data. They also described the characteristics of each aspect in a tabular form. The authors further discussed cloud computing technology and its service models: PaaS, SaaS, and IaaS. They illustrated the relationship between cloud computing and big data. Cloud computing provides a facility to process big data and serves as a service model. They mentioned MapReduce as a good big data processing framework in a cloud environment. Various case studies are provided by the authors on big data using cloud computing technology. The authors further considered the problem of storage in big data, as earlier structured RDMS was used to store data but was incapable of managing a massive amount of data. They highlighted a few existing storage technologies: Network-attached storage (NAS), direct-attached storage (DAS) and storage area network (SAN). They discussed the Hadoop background and its two components HDFS and MapReduce in respect of cloud technology. The authors also focused on research challenges in big data processing such as availability, data quality, scalability, data integrity, the transformation of data, privacy issues and data heterogeneity. They also presented open research issues in big data in clouds such as data staging, data security, data analysis, and distributed storage systems.

Ta et al. (2016) studied big data streaming in healthcare analytics. Multiple processes can be executed smoothly within the healthcare sector by taking advantage of big data analytics. Healthcare data sources can be collected in both non-clinical and clinical domains, including sensing data, social media, biomedical images, genomic data, biomedical signals, clinical text, and electronic healthcare records. Big data stream computing is studied to bring improvement in quality services and cost reduction in the healthcare industry. Big data analytics in healthcare can help in retrieving the required piece of information, and further use it to predict the diseases and cure them, raise the quality of life and preventable deaths can be avoided. The authors proposed a generic architecture for analysing big data healthcare using open sources such as Hadoop, Apache Storm, Kafka, and NoSQL Cassandra. They mentioned that Apache Hadoop is batch-oriented computing, whereas Apache storm, NoSQL Cassandra, Kafka are the prime technologies in the field of Big Data Streaming Computing (BDSC). They further described Kafka messaging system and Kafka system architecture. They also discussed Apache storm technology, its topology and storm cluster architecture. To improve the big data computing in the domain of healthcare, a generic architecture of combined advantages of both batch computing and stream computing is provided by the authors. They also focused on related future work, which includes the requirement of power tools such as machine learning and the use of

data mining techniques for the efficiency improvement of data analytics. As the healthcare data is growing exponentially, achieving the efficient result from this industry is still questionable, and it can be treated as a base for future works.

Asha et al. (2018) presented an algorithm for retrieving the associated patterns in big data using Hadoop cluster. They stored data (structured or unstructured) in HDFS and performed the Map-Reduce technique. Apriori algorithm was used as a mining procedure to fetch the patterns in a Hadoop environment. They mentioned the advantages of their proposed system, which includes extracting frequent itemset, performance improvement, associated models existing infrequent itemsets, performance evaluation of the Hadoop single node cluster. The authors suggested future work as a replacement of data node by multi-nodes, that can be used to support applications such as medical applications, stock market, finance, weather forecasting ad business.

Hirve and Reddy (2019) surveyed the techniques related to visualisation of data that can be used for the analysis of big data. Data representation plays an important role in the field of big data. A better visual representation can provide a better understanding, exploration of data and a rapid increase in the growth of innovation. Visualisation can be done in any form such as reports, bar charts, tree diagrams, lines graphs, or any other diagram-based representation. The authors proposed a case study to compare all the data visualisation tools and techniques existing so far, which is appropriate for handling big data. They highlighted all the advantages and disadvantages of data visualisation tools to recommend the best suited. The literature survey part, as mentioned showed up all the previous work based on visualisation tools and techniques. They mentioned various tools such as Google charts tool for all type of browsers; Wordle, a web-based tool developed for visualising formats; Data wrapper, texts and different layouts of the text; Tableau for charts, graphs and maps production; e-charts of Baidu includes pie charts, maps, line graphs, scatter plots, and others. They further explained the methodology for using big data Hadoop, Hue, CDH's Apache Impala, and Hive. Cloudera is used for processing queries and generate reports of data set. Further, for 3D visualisation and QR code generation, the generated 2D output is treated as an input using the Unity 3D engine. They categorized frequently used data such as population-based data, e-commerce, and retail data, data related to the environment. The authors also illustrated visualisation based on locations such as visualisation based on points, regions, lines. After comparing various visualisation tools like Baidu e-charts, Wordle, Google charts, Data Wrapper, the authors concluded that Google charts performed best as compared to other tools in terms of ease in visualisation, reliability and other aspects. They also suggested future work to extend this concept to a specific level. Using PHP host file, the visualized content of code in 2D format can be provided to unity AR (Augmented reality) engine which will be treated as input data, and it will help in converting it into a mobile application. The generated statistics will be transfigured to small QR codes, which will allow the users to see it using smartphones.

Jannapureddy et al. (2019) studied big data analysis using the auto-scaling framework in the cloud environment. They mentioned that pre-configuration is required to set up on Apache Hadoop with sufficient resources to compute the data load at the peak, but this may cause wastage of computing resources in case of lower usage levels than the predetermined load. The authors investigated an auto-scaling framework in consideration of this problem, which can automatically adjust the virtual nodes when real-time data load is maximum or minimum, which can eventually minimize the cost of resource use. They considered a cost-effective auto-scaling (CEAS) framework via MapReduce framework using Hadoop, which was first presented for an Amazon Web Services (AWS) in the cloud environment. The authors presented a case study of real-time sentiment analysis of Universities' twitter data with various functionalities using the MapReduce technique to validate the effectiveness of the proposed framework.

Hajdar and Jedidi (2019) proposed a new approach based on tasks or job scheduling in a big data cluster using Hadoop as a technology. They developed the approach for scheduling the tasks to improve the efficiency of the scheduler inside the DataNodes presents in a big data cluster. They focused on optimizing the assignment of tasks by the NameNode to the data nodes. The authors

mentioned that the default scheduler of Hadoop is inefficient for non-homogeneous components and faces failure in identifying slow tasks. Hadoop's traditional task scheduler includes three schedulers: FIFO scheduler, Hadoop Fair Scheduler (HFS) and capacity scheduler. They mentioned that their proposed algorithm also considers the resource utilization in terms of CPU, I/O, RAM, Network and scheduled job types. They experimented with four machines, including one NameNode and rest as DataNodes machines. The authors considered a homogeneous cluster and different types of jobs are being initiated into that cluster to test the performance of the proposed algorithm. The results showed that their task scheduler performed better when compared to FIFO scheduler and Capacity Scheduler. The authors also focused on testing and evaluating the performance of the algorithm above in a heterogeneous cluster as the future works.

4. DISCUSSIONS

In this paper, several kinds of literature have been reviewed and discussed to identify the state-of-the-art scenarios of the topics being investigated, which includes big data analysis, Hadoop and cloud computing. We have found various gaps in the existing literature that require new understanding or fresh insights. This paper addresses the important issues and gaps in a single article to broaden up the research domain to facilitate the researcher to enhance the technology and to formulate a deeper, more profound understanding.

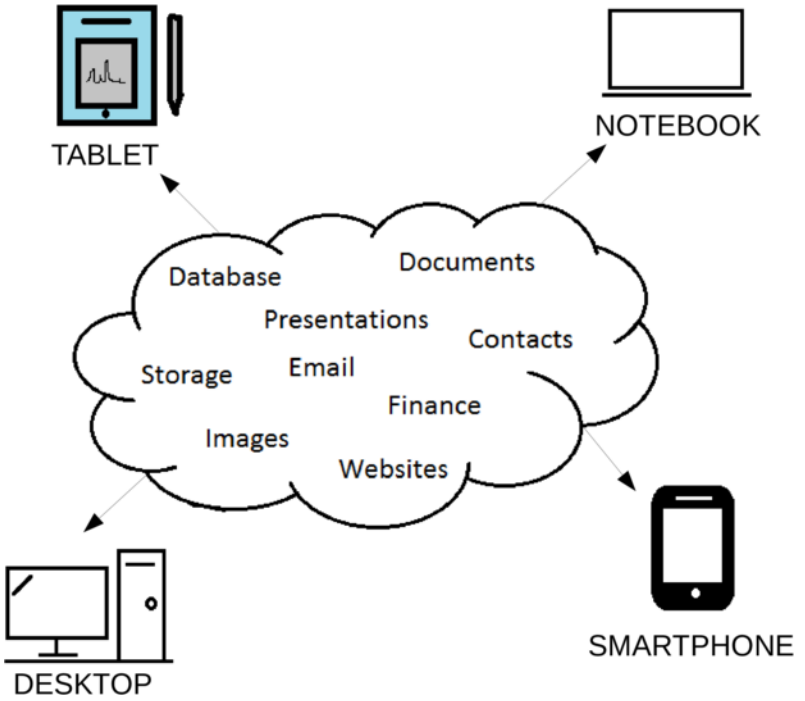
After reviewing papers, as stated in Section III, and considering findings, we found that there still exist many gaps in the research domain of big data, Hadoop and cloud computing. They need to be filled which had not been previously exposed or satisfactorily described. Following are the glitches that need to be considered:

- Privacy and data security issues in the cloud environment;
- Visualisation of large data sets for better understanding;
- Analysing large heterogeneous or unstructured data sets.

4.1. Privacy and Data Security Issues in the Cloud Environment

Security and privacy are the two biggest sticking points when it comes to moving data into the cloud. Out of the two issues, security is obviously the most important as the data can be securely stored in a way that does not ensure privacy but privacy cannot be maintained if supporting systems are not secure. Cloud computing technology provides the advantage of accessing all our data and applications from any network device. Figure 5 shows the ubiquitous network access using cloud computing. Data present in any form like documents, images, email, etc. can be moved to the cloud and users can retrieve information using tablet, desktop, smartphone or any other network device from anywhere and anytime. Moreover, cloud analytic applications have noticeable benefits for big data processing, making it scalable, easy and cost-effective (Elsayed and Zulkernine 2018). Cloud serves many facilities to the users such as pay per use, no hardware setup and easy access from any network, etc. Despite all these advantages, still there exist privacy and data security risks in the cloud environment that are holding back cloud projects. These risks include invisibility of data present on cloud applications, not having full control over access rights of sensitive data, incapability of preventing intruder or misuse of the data, etc. Elsayed and Zulkernine (2018) considered security breaches related to malicious, vulnerable and misconfigured analytic applications that can harm data integrity and confidentiality. They presented a real-time security monitoring as a service (SmaaS) framework that detects oddities in the cloud analytical applications which are running on Hadoop clusters. Mo (2019) introduced two distributed storage systems in the cloud computing environment, which are Google's GFS and Hadoop's HDFS. Furthermore, the author proposed an enhanced ant colony algorithm to tackle the problem of the internet of things (IoT) data security storage in a cloud

Figure 5. Ubiquitous network access using cloud computing



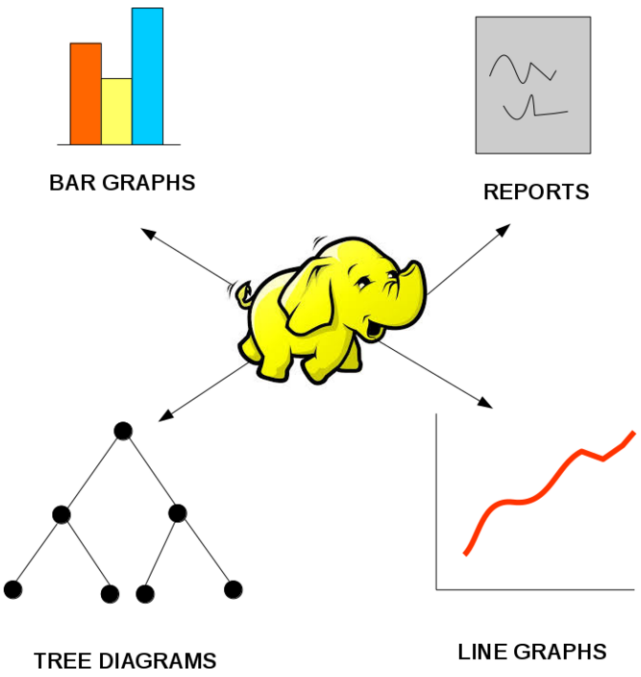
environment. The algorithm inspects the shortest completion time and load balancing of data storage. Dey et al. (2018) studied mobile cloud computing and secured the transferring of data between the different parties. They mentioned some existing authentication schemes such as Secure Shell (SSH) password-based authentication and token-based authentication. But these existing schemes are either expensive in terms of computation or not appropriate for the mobile cloud computing environment in terms of resource utilization and security. The authors proposed a new scheme, known as a mutual authentication scheme based on mobile cloud computing. In this AMLT scheme, Authentication is based on Message digest Location and Timestamp. They stated that AMLT utilizes symmetric keys to decrease the complexity of computation, and dynamic keys are utilized to minimize identification of encryption keys. To validate the AMLT security, they used protocol analyzer Scyther. Wu et al. (2019) presented a solution to cloud storage security through equilibrium analysis. They performed evaluations on several game models between users and public cloud storage providers to examine the security of different service scenarios. They highlighted that users could check the risk of private data hacking by the cloud service providers, and the cloud service providers can also plan their strategies to make their services more reliable. They inspected and verified this approach by performing an experimental study of 32 users allowing to validate the method for real-time service improvement. Zhang et al. (2019) provided a hybrid encryption algorithm to deal with the data security issue in the hospital cloud database. The authors first considered the AES algorithm. They improved it and called it a P-AES algorithm. Further, this algorithm is combined with the RSA algorithm. They performed experiments to verify the algorithm, and the results showed that this encryption algorithm has various advantages such as better security, good processing ability for longer data, fast encryption and decryption speed, and can solve the issue of data security in a cloud database to a certain extent. There can be some measures which can counter these threats and security issues in cloud computing such as enhanced security policy should be introduced and implemented, improvement in access control management by allowing only authorized user to access data, different tools for data protection

and security techniques should be used. These courses of action can help to reduce the risk of abuse of cloud and ensure the integrity of the data.

4.2. Visualization of Large Data Sets

The second most important gap in cloud computing and big data domain is data visualisation. It is a process of taking raw data and transforming it into graphs, charts, images, and even videos representing the purpose and allow us to gain insights from it. Visual representation of data patterns plays a vital role when you have an enormous amount of data with you. Visualisation techniques help users to easily obtain insights plays a vital role when you have an enormous amount of data with you. Data visualization can be conveyed in the form of three basic types: Relationship, Comparison, and Composition. A relationship shows the connection or correlation between two or more variables in a dataset. Scatterplot technique or bubble charts can be used here. Comparison sets one set of variables apart from another and displays the interaction between them. Bar graphs and line charts can help in performing the comparison. The third type of relationship is composition. It collects different types of data that make up a whole and displays them together. The canonical example of this can be a pie chart. Distribution can act as a fourth type of visualization. It can layout a collection of data to show how it correlates and help to understand any kind of interaction between the variables. Visualisation techniques help users to easily obtain insights from data and make their information more actionable. These techniques illustrate relationships within data and discover data values. Visualisation can be attained in pictorial or graphical formats such as bar graphs, charts, and spreadsheets. Figure 6 illustrates different visualisation formats of big data such as line graphs, tree diagrams, reports, bar graphs. Visualisation of data is important because it helps in decision making, simplifying the complex quantitative information and supports in analysing and exploring big data easily. It identifies the areas that need attention or improvement and recognizes the correlations and unexpected relationship between data points and variables. It also explores new patterns, spot trends and unveils hidden patterns in the data. Clarity, efficiency, and accuracy are the three major

Figure 6. Visualisation of big data



considerations for data visualisation. Clarity makes sure that the data set is complete and relevant. Accuracy ensures that appropriate graphical representation is used to convey right messages. The use of efficient visualisation technique can help in highlighting all the data points. Before visualizing the data, some basic factors need to be taken care which are visual effects, coordination system, data types and scale, and informative interpretation. Visual effects include the usage of appropriate shapes, colors, and sizes to represent the analysed data. The coordinate systems help to organize the data points within the provided coordinates efficiently. The data types and scale choose the type of data such as numeric or categorical. The informative interpretation creates the visuals in an effective and easily decipherable manner using labels, titles, and pointers. There are various multi-dimensional visualisation tools such as Google charts, Tableau Public, Weave, Wordle, Kibana. Devi et al. (2019) illustrated graph analysis and visualisation of social network big data. They considered graph as a useful representation to visualize the hidden relationship between unstructured data sets. The authors showed the process of creating, transforming, visualizing, and analysing large-scale graphs from sample data obtained from Amazon social networking website. Nazir et al. (2019) presented a review on visualization of big data in cardiology. The author proposed a study on existing literature related to visualisation of big data in cardiology and outline different visualization techniques. They highlighted the various benefits of visualization which can help to achieve important information from patient's data.

4.3. Analyzing Large Unstructured Data Sets

The third gap that we consider is analysing heterogeneous or unstructured data. Heterogeneous data is a kind of data that is available in many formats, and it has a high variability of different data types. Medical imaging data pose a huge volume in terms of data size (Chatterjee 2018; Chatterjee et al. 2019). Due to high data redundancy and missing values, they are ambiguous and persist low quality. As mentioned above, unstructured data refers to the piece of information that does not reside in a traditional row-column database. They lack proper format in storage. It is quite a cumbersome task to retrieve valuable information from unstructured data. Most of the big data encountered is in textual format which is highly unstructured in nature. Most of the big data encountered is in an unstructured format. Unstructured data can be characterized based on its source, including human-generated data or machine-generated data. Human-generated data is found in huge amounts across the internet such as website content, social media data, and mobile data, whereas machine-generated data is scientific data. Yue (2018) studied an ecosystem based on a healthcare hybrid Hadoop for unstructured healthcare data logs. The author mentioned that the proposed ecosystem consists of different components such as Sqoop and Zookeeper, Hive, Pig, Hadoop Distributed File System (HDFS), HBase and MapReduce. Unstructured healthcare data can also be retrieved using the Apache Drill. The author identified that the tools like Hadoop ecosystem and Apache Drill are authentic enough to gain access to a large amount of complicated data in the healthcare domain. Natural language processing (NLP), data mining, text mining are some techniques that can help to produce significant and actionable insights from highly unstructured data. Text mining or text analytics is the process of deriving information from natural language text. It involves the process of structuring the given input text and further obtain patterns within the structured data and lastly evaluate and interpret the output. It gains high-quality information from the text. On the other hand, natural language processing is the artificial intelligence method which communicates with an intelligent system using the natural language. The overall goal is to turn the unstructured data present in the form of text into data analysis via the application of natural language processing. That's why text mining and NLP go hand-in-hand. NLP can be used in speech recognition, sentiment analysis, information extraction, advertisement matching, and machine translation.

Finally, we can throw light on the latest and upcoming research trend in the domain mentioned above and also point out the potential area of investigation, which includes:

- Scaling and finding key data from big data;
- Evaluating performance and using Hadoop technology on a cloud platform.

5. CONCLUSION AND FUTURE WORK

In this highly data-driven world, storing and processing this data for extracting valuable information becomes a challenging task. Traditional computing systems cannot accomplish the task of handling big data within the given time frame as the computing resources of these systems is not sufficient to process and store such a huge volume of data. This is where various tools such as Hadoop and Apache Spark come into the picture. In this study, we have examined the concept of big data, its characteristics, sources and some facts related to it. We have provided a conceptual overview of Hadoop technology, cloud computing, and Apache Spark in detail. Here, we have presented a critical broad review of state-of-the art-related literature. Privacy and data security issues in cloud computing require more intentness as various data-related issues like data breaching, data removal, data recovery, data locality can lead it to security threats. Guest hopping attack, malicious insider, side-channel attack are some Content Security Policy attacks. There also exist network level and application related security attacks such as DNS attack, IP spoofing, cloud malware injection attack, cookie poisoning, hidden field manipulation that can act as threatening remarks. Another gap that demands attention is visualization of incredible amounts of data that bring the value of the information to light. Raw details of data can remain obscure if not analyzed properly. That's where data visualization comes into play. It can change the way we make sense of the information to create value out of it. The visual representation can help in discovering new patterns and spot trends from large data sets. Analyzing large heterogeneous or unstructured datasets is also an exacting issue that requires further research. Furthermore, current research gaps and open challenges are identified and discussed that require further study.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The corresponding author of this paper is Indranath Chatterjee and can be reached at indranath.cs.du@gmail.com.

REFERENCES

Asha, P., Prem Jacob, T., Pravin, A., & Asbern, A. (2018). Mining the Associated Patterns in Big Data Using Hadoop Cluster. In *International Conference on Intelligent Data Communication Technologies and Internet of Things*, (pp. 1255–1263). Springer.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Chatterjee, I. (2018). Mean deviation based identification of activated voxels from time-series fMRI data of schizophrenia patients. *F1000 Research*, *7*(1615). doi:10.12688/f1000research.16405.2 PMID:30687497

Chatterjee, I., Agarwal, M., Rana, B., Lakhyani, N., & Kumar, N. (2018). Bi-objective approach for computer-aided diagnosis of schizophrenia patients using fMRI data. *Multimedia Tools and Applications*, *77*(20), 26991–27015. doi:10.1007/s11042-018-5901-0

Chatterjee, I., Kumar, V., Sharma, S., Dhingra, D., Rana, B., Agarwal, M., & Kumar, N. (2019). Identification of brain regions associated with working memory deficit in schizophrenia. *F1000 Research*, *8*(124), 124. doi:10.12688/f1000research.17731.1 PMID:31069066

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. doi:10.1007/s11036-013-0489-0

Dey, S., Ye, Q., & Sampalli, S. (2018). AMLT: A Mutual Authentication Scheme for Mobile Cloud Computing. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (Smart- Data)*, (pp. 700–705). IEEE. doi:10.1109/Cybermatics_2018.2018.00140

Dillon, T., Wu, C., & Chang, E. (2010). Cloud computing: issues and challenges. In *2010 24th IEEE international conference on advanced information networking and applications*, (pp. 27–33). IEEE. doi:10.1109/AINA.2010.187

Elsayed, M., & Zulkernine, M. (2018). Towards security monitoring for cloud analytic applications. In *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, (pp. 69–78). IEEE. doi:10.1109/BDS/HPSC/IDS18.2018.00028

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

Hadjar, K., & Jedidi, A. (2019). A New Approach for Scheduling Tasks and/or Jobs in Big Data Cluster. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, (pp. 1–4). IEEE. doi:10.1109/ICBDSC.2019.8645613

Hashem, Yaqoob, Anuar, Mokhtar, Gani, & Khan. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.

Hirve, S., & Pradeep Reddy, C. H. (2019). A Survey on Visualization Techniques Used for Big Data Analytics. In *Advances in Computer Communication and Computational Sciences* (pp. 447–459). Springer. doi:10.1007/978-981-13-6861-5_39

Jadeja, Y., & Modi, K. (2012). Cloud computing-concepts, architecture and challenges. In *2012 International Conference on Computing, Electronics and Electrical Tech- nologies (ICCEET)*, (pp. 877–880). IEEE. doi:10.1109/ICCEET.2012.6203873

Jannapureddy, R., Vien, Q.-T., Shah, P., & Trestian, R. (2019). An auto- scaling framework for analyzing big data in the cloud environment. *Applied Sciences (Basel, Switzerland)*, *9*(7), 1417. doi:10.3390/app9071417

Kaisler, Armour, Espinosa, & Money. (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences*, (pp. 995–1004). IEEE.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, (pp. 404–409). IEEE. doi:10.1109/IC3.2013.6612229

Kumar, A., & Chatterjee, I. (2016). Data Mining: An experimental approach with WEKA on UCI Dataset. *International Journal of Computers and Applications*, *138*(13).doi:10.5120/ijca2016909050

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing—The business perspective. *Decision Support Systems*, *51*(1), 176–189. doi:10.1016/j.dss.2010.12.006

Mithili & Kasireddy. (2019). Graph Analysis and Visualization of Social Network Big Data. In Social Network Forensics, Cyber Security, and Machine Learning, (pp. 93–104). Springer.

Mo, Y. (2019). A Data Security Storage Method for IoT Under Hadoop Cloud Computing Platform. *International Journal of Wireless Information Networks*, *26*(3), 1–6. doi:10.1007/s10776-019-00434-x

Nazir, S., Khan, M. N., Anwar, S., Adnan, A., Asadi, S., Shahzad, S., & Ali, S. (2019). Big Data Visualization in Cardiology—A Systematic Review and Future Directions. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 115945–115958. doi:10.1109/ACCESS.2019.2936133

Patel, A. B., Birla, M., & Nair, U. (2012). Addressing big data problem using Hadoop and Map Reduce. In *2012 Nirma University International Conference on Engineering (NUiCONE)*, (pp. 1–5). IEEE. doi:10.1109/NUICONE.2012.6493198

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*(1), 3. doi:10.1186/2047-2501-2-3 PMID:25825667

Sagiroglu, Seref, & DuyguSinanc. (2013). Big data: A review. In *2013 International Con- ference on Collaboration Technologies and Systems (CTS)*, (pp. 42–47). IEEE.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7), e1002195. doi:10.1371/journal.pbio.1002195 PMID:26151137

Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, *34*(1), 1–11. doi:10.1016/j.jnca.2010.07.006

Ta, V.-D., Liu, C.-M., & Nkabinde, G. W. (2016). Big data stream computing in healthcare real-time analytics. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, (pp. 37–42). IEEE.

Vavilapalli, Murthy, Douglas, Agarwal, Konar, Evans, & Graves. (2013). Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing*. ACM. doi:10.1145/2523616.2523633

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107.

Wu, Y., Lyu, Y., & Shi, Y. (2019). Cloud storage security assessment through equilibrium analysis. *Tsinghua Science and Technology*, *24*(6), 738–749. doi:10.26599/TST.2018.9010127

Yue, H. (2018). Unstructured Healthcare Data Archiving and Retrieval Using Hadoop and Drill. *International Journal of Big Data and Analytics in Healthcare*, *3*(2), 28–44. doi:10.4018/IJBDAH.2018070103

Zhang, F., Chen, Y., Meng, W., & Wu, Q. (2019). *Hybrid Encryption Algorithms for Medical Data Storage Security in Cloud Database. International Journal of Database Management Systems*, 11.

Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and re- search challenges. *Journal of Internet Services and Applications*, *1*(1), 7–18. doi:10.1007/s13174-010-0007-6

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, *28*(3), 583–592. doi:10.1016/j.future.2010.12.006

Akansha Gautam has completed M.Sc. in Computer Science from Department of Computer Science, University of Delhi, Delhi, India. Prior to this, she accomplished B. Sc. in Computer Science from Indraprastha College for Women, University of Delhi, India. Her research interest lies primarily in the areas of Big data Analysis and Cloud computing.

Indranath Chatterjee (PhD) is currently working as a Professor in the Department of Computer Engineering at Tongmyong University, Busan, South Korea. Previously, he was working as an Assistant Professor at JK Lakshmipat University, Jaipur, India. He did Ph.D. in Computer Science from Department of Computer Science, University of Delhi, Delhi, India. Prior to this, he did his B.Tech in Computer Science and Engineering from West Bengal University of Technology, Kolkata, India. He is currently serving as a member of the advisory/editorial board of various international journals and Open Science organizations. His research areas include Data Analytics, Machine Learning, Computational Neuroscience, Medical Imaging and Data Science.