International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I11P218

Volume 70 Issue 11, 170-177, November 2022

Review Article

Integration of Big Data and Cloud Computing: Tools, Issues, and Reliability

Ranjit Rajak[1], Satish Chaurasiya[2], Anjali Choudhary[3]

[1,2,3]*Department of Computer Science and Applications, Dr. Harisingh Gour Central University, Sagar (M.P.), India.*

[2]*Corresponding Author : satishchaurasiya5@gmail.com*

Received: 06 August 2022 Revised: 02 November 2022 Accepted: 16 November 2022 Published: 26 November 2022

Abstract *- Using information technology in various communication methods produces a massive amount of data. In current years, data size is increasing because of gathering business information, the Internet of Things (IoT), the enormous growth of social networks, etc. If the data size is large, it has two main issues: processing and storage. These issues can be solved using Cloud Computing. Cloud Computing provides the facility of the virtual environment. Where data is stored and processed on virtual servers. Cloud computing also provides a reliable, scalable, fault-tolerant environment to handle Big Data in Distributed Management systems. Big Data, a collection of homogeneous and heterogeneous information scaling up at very high-speed need, needs to be analysed without chance of any error and mistake, which may lead to improper and compromised evaluation from where the reliability measure needs to be adapted into the frame. This paper aims to provide detailed information regarding Big Data in Cloud Computing, such as definitions, Characteristics, technologies used, and reliability of Big Data in Cloud Computing. This paper describes the research challenges and different security aspects of Big Data. Finally, the reliability of Big data in the Cloud Computing paradigm is analysed in terms of the probability distribution*

Keywords *- Big Data, Big Data Reliabilities, Hadoop, Map Reduce, Big Data Challenges, Big Data Tools and Technologies.*

1. Introduction

This paper explores the idea of Big Data and Cloud Computing. Firstly we understand the concept of data. The raw material of information is known as data before being arranged, organised, and processed. Information cannot be used primarily until processed and without performing a preparatory operation on given raw material or data. Data is raw facts and figures obtained through some experiment or survey, and the information is extracted after performing analysis and processing. In recent years, there has been a deep interest in storing and processing a large amount of data such as science, finance, government, etc. Big Data is a dataset with a giant size and capacity of traditional software tools that are commonly used these days are insufficient to manage and process data in specified periods with variety, velocity, volume and value, and veracity[1]. The word Big Data is used to represent the large amounts of data that are organised in a structured, semi-structured, and unstructured format. Data is being warehoused and gathered from material available on the web, internet business, bank cards, and social networking sites. There is a large number of software tools have been designed to support Big Data and Cloud Computing. The Cloud computing environment provides three significant attributes, which are reliable, accessible, and scalable to big data platforms which perform data storage and processing [2,3]. The two mentioned Big Data, and current developing technologies can distinguish Cloud Computing. Big Data and Cloud Computing provide the basics for managing a broad scope of data resources. The primary intent of this paper is to discuss an extensive exploration of Big Data and Cloud Computing environment definition, characteristics, classification of Big Data, a storage system of Big Data, Hadoop, Map Reduce technology discussed besides, focus the discussion on Cloud Computing definition, characteristics, models of Cloud Computing, technologies used for Cloud Computing such as Microsoft Azure, AWS, etc. This paper also discusses the relationship and reliability of Big Data in the Cloud Computing paradigm regarding software, hardware, and data itself, respectively. Moreover, content-based reliability and system reliability are discussed along with reliability function and hazard rate.

2. Big Data

The term‘ big data’ was introduced when handling vast amounts of data generated by different sources, such as social media, was very complex in the conventional database system. The conventional system could not handle, analyse and manage these data efficiently. Big data is a new platform applicable in every field where data is a vast and significant concern. The number of authors defined big data as follows in table 1, and there are various big data resources as shown in figure 1.

Table 1. Big Data Definition by Various Authors

Definition1[4,5]	"Big data is high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimisation."
Definition2[6]	"Big Data is data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time."
Definition3[7]	"Big data refers to datasets with sizes beyond the ability of common software tools to capture, curate, manage, and process the data within a specified elapsed time."
Definition4[8]	"Big data differs from 'regular' data along four dimensions, or '4 Vs'- volume, velocity, variety, and veracity."
Definition5[9]	"Big data–that is, the sophisticated and rapid analysis of massive amounts of diverse Information."

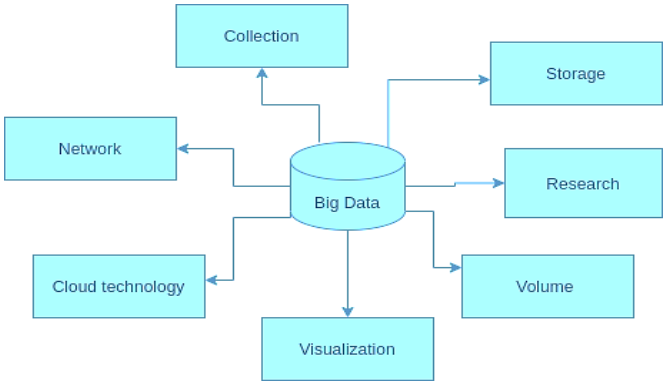

Fig. 1 Big Data Resources

Big data is presently characterised by 9 Vs[10], as shown in figure 2, and brief details are tabulated below.

- *Veracity*

Veracity refers to the meaningful data concerned with storage and mining corresponding to the challenges being analysed. It talks about the noise biases and abnormalities present in data. It mainly focuses on the quality of data.

Ex. Data from a medical experiment or trial, the degree to the accuracy of data

- *Value*

To determine the measurable and financial impact of Big Data, it goes for transformation and analysis. It includes the extraction of value from massive data

Ex. Useful information within the data field containing numeric quantity, textual characteristics, date or time measurement

- *Volume*

It is related to heavy data generated every day from various sources such as social media, research fields, advanced scientific computing fields, etc.

Ex. Social media such as Facebook, where data is uploaded and downloaded daily in different formats such as images, video, audio, etc.

- *Variety*

It refers to the various data formats, such as video, audio, text, graphics, etc., due to the era of technology. They can be structured or unstructured formats.

Ex. Social media posts, emoticons, audio, and animated messages.

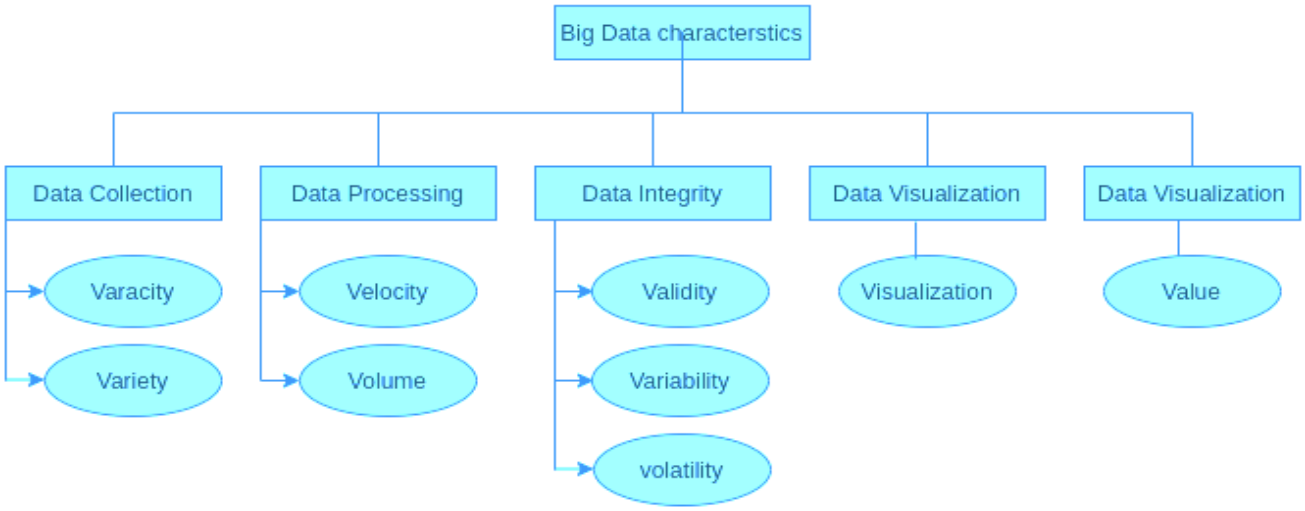

Fig. 2 Big Data Characteristics

- *Velocity*

It is concerned with data processing speed, including collecting and analysing data from different sources.

Ex. Data streaming and processing, Bluetooth sensor flooding, and the global positioning system.

- *Validity*

Validity refers to the correctness and accuracy of data intended for use, the primary concern of data validity is making the right decision from the data. It deals with the uncorrupted transmission of data.

Ex. Checking the correct data type entered, consistency check, format checking, range checking, and uniqueness check of data.

- *Volatility*

Volatility refers to the retention of data for a certain amount of time, business firms keep data secure for some time, and the same data can be destroyed after the retention period expires.

Ex. The warranty period of any product, the product's service history during the warranty period.

- *Variability*

Variability refers to the inconsistency of data within periodic change, including daily or seasonal; it deals with event triggers data and how much the data varies in the period, which is difficult to manage.

Ex. Clicks on the website after result declaration.

- *Visualisation*

Visualisation refers to the comprehensibility of a huge amount of data. Visualisation and analysis put raw data into use. Otherwise, it remains useless. It helps to make large data sets understood, able, and re adable with the help of graphs and charts.

Ex. Use scatter plots, time series charts, cartograms, and dot distribution maps.

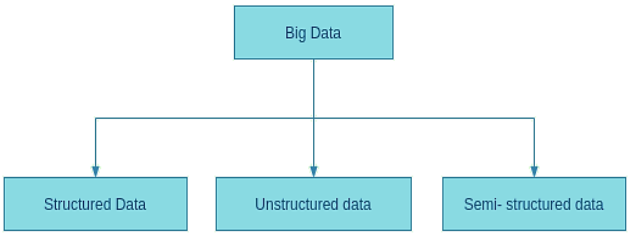

Fig. 3 Big Data Classification

2.1. Classification of Big Data

The structure of data is a benchmark in the categorisation and classification of big data, which is uniquely categorised into three distinguished categories, namely, structured data, unstructured data, and semi-structured data, as shown in figure 3 shown below and the brief description of mentioned classes of big data with examples are mentioned below.

2.1.1. Structured Data

Those types of data are represented in the well-organised form of tables and databases to be processed with a specific schema.

Ex. A relational database, excel files

2.1.2. Unstructured Data

Unstructured data cannot be represented in an organised form. This type of data neither has a pre-defined data model nor any schema.

Ex. Text, audio or files, mobile activity, social media comments and posts, and images from satellite and surveillance devices.

2.1.3. Semi-Structured Data

Semi-structured data cannot be represented by tabular structure or pre-defined data model but have tags or markers which distinguish semantic elements or fields contained within data.

Ex. ML files, TCP/IP packets, web pages, zipped files.

2.2. Big Data Tools and Technologies

The current technologies cannot satisfy the need for Big Data, and they are also not suitable and reliable for the increasing speed of data transfer. These technologies have lesser speed compared to increasing data. Several technologies have been developed to meet the requirement of a large amount of data.

Hadoop is one of the prominent and open-source tools for handling big data, which utilises the concept of distributed computing for data storage and parallel computing for processing big data, respectively. HDFS (Hadoop Distributed File System) is a specially designed file system for a cluster of interconnected commodity hardware machines. In contrast, the processing component in Hadoop is Map Reduce, later replaced by YARN (Yet Another Resource Navigator) in the second version of Hadoop.
Some primary Big Data tools are given below with their brief explanation.

2.2.1. Hadoop

Hadoop is a framework of java programming which available free of cost. Hadoop is a part of apache software and is provided by the foundation of apache. A large number of

the dataset is processed by the Hadoop platform over a cluster of servers, and numerous applications can be executed simultaneously on the system with a large number of computing and active nodes.

Hadoop provides a fast data transfer rate. It supports the processing of datasets in the environment of distributed computing. It works on the structure of the master enslaved person. Hadoop is further grouped into two parts- *Hadoop Distributed File System* is the storage component, and *Map Reduce* is the processing one. Prominent Multi-National Companies like Google, IBM, Yahoo, etc., are using Hadoop for subsidiaries; their applications consist of a huge amount of data.

2.2.2.Apache Spark

Apache Spark is also an open-source, distributed cluster computing system used to speed up the analysis process. It is a framework that provides much faster performance. Apache spark is grouped into two categories; cluster manager is the first, and distributed storage system is the second. Some prominent features make it a perfect framework and increase spark performance for Big Data applications.

2.2.3. Hive

Hive is an SQL-based data warehouse tool developed by Facebook, which was later open-sourced to apache and used SQL-like queries for accessing, updating, and managing giant data sets in distributed storage environments.

Hive can be installed on top of Hadoop and facilitates users to process structured data; more specifically, it is used to run HQL queries similar to SQL on large amounts of data. Hive is heavily used for online analytical processing(OLAP), Hive is relatively scalable, fast, and flexible.

2.2.4. Data Lake

A data lake is a Big data tool developed by a software vendor called Databricks Inc., an initiative by the developers of the Spark processing engine, which was later opened and sourced to the Spark-based technology by the Linux Foundation in the year 2019. Delta Lake is "an open format storage layer that delivers reliability, security, and performance on your data lake for both streaming and batch operations."

Delta Lake is a tool that resides on top of other tools and creates a single inhabitant for unstructured, unstructured, and semi-structured, which eliminates the data complexity of big data applications. It supports ACID transactions and includes Spark-compatible APIs. Storing data in apache parquet format

2.2.5. Kafka

Kafka is a streaming tool based on a distributed system maintained by apache and is operated by approximately 80% of Fortune companies and other multiple organisations for pipelining of high-performance data. Kafka is an extensive data framework for streaming data processing, namely reading, storing and analysing.

Kafka was developed by LinkedIn, which was later open-sourced to apache in 2011.

2.3. Big Data Research issues

Distributed Mining: Many data mining techniques cannot uncover hidden information from Big Data. So more research is required to provide new methods for both theoretical and practical analysis.

2.3.1. Statistical-Significance

Sometimes the randomness may lead to inaccurate and compromised results, so achieving significant statistical results may be primary vitals

2.3.2. Analytical Architecture

The significance of optimal architecture in the analytical system is not clear in simultaneously dealing with real-time and historical data. The Lambda architecture decomposes the problem into three distinct layers of ordinary real-time data to meet the challenge of an arbitrary function. The three mentioned layers are the batch layer, the serving layer and the speed layer.

2.3.3. Heterogeneity

Big Data typically deals with, and such Heterogeneity in Big data is the major concern as the data is collected from several sources, and the heterogeneity problem is currently under the study

2.3.4. Security

Big Data security is a significant and critical concern as this problem becomes a challenge for corporations while migrating data into the cloud. It is difficult to describe the data ownership, location, and accessibility privilege of the data.

3. Cloud Computing

Cloud Computing [22,23] may consider a shared resource that provides the facility of online services such as applications for desktop applications, virtual servers, storage, etc. Cloud Computing provides the facility of virtual resources. For example, we store photos online instead of on our home computers or social networking sites. The term Cloud Computing can be referred to as an on-demand computing resource and system capable of providing several integrated computer services without being bound to local resources. If we try to use these services, it may take a long time, but if we use that service online, then we are using Cloud Computing services[14,15].

In other words, Cloud Computing delivers services over the internet. If the network connection is available, then the

model of Cloud Computing permits access to the shared resource and information available anywhere.

3.1. Cloud computing Service Model

Cloud Computing is primarily classified based on the Cloud Computing service model [16,17]. Cloud Computing service models are described as mentioned below-

3.1.1. Software as a Service (SaaS)

The Software as a Service model is called an on-demand provider. This gives the capability of consumers to operate a variety of software that can be run without installing them on their physical machine. It provides services over the net via the application running on the cloud infrastructure.

3.1.2. Platform as a Service (PaaS)

The platform as a service model permits developers to deploy their programs on the cloud.

3.1.3. Infrastructure as a Service (IaaS)

Infrastructure as a Service is applied to developing and maintaining the organisation's infrastructure. These are virtual platforms connected to the networks.

4. Relation Between Big Data and Cloud Computing

Cloud computing is something new practice within technology development, as the development of generation has led to the rapid development of the electronic knowledge society.

It leads to a state of great realities, and the rapid growth of big data is a problem that can face in developing a digital knowledge society. Cloud computing and big data jump together, as big data is about storage capacity within the cloud environment, and cloud computing uses massive computing and storage resources.

Big data stimulate and accelerate the improvement of cloud computing by way of imparting significant information software with computing functionality. In environmental computing, distributed storage generation allows controlling huge amounts of data. Cloud Computing and big data interrelate with each other, and the rapid increase in big data always creates a challenge.

Clouds provide ideal solutions for big data, just as conventional storage cannot meet the needs for dealing with big data. Therefore, the cloud computing environment is ideal for large data. Massive data storage and processing require enlargement because the cloud presents enlargement through digital machines and helps big data to develop and grow to be reachable. This is the compatible relationship between big data and cloud computing.

5. Big Data Reliabilities

Reliability in big data refers to the accuracy and completeness of the data; as the volume, variety, and velocity of data are very large, reliability becomes the major and challenging vital. The reliability assessment reveals the data's major problems and challenges. The three major reliability assessments are:

- *Validity*
  Validity refers to the correct format and the proper storage of the data.
- *Completeness*
  Completeness refers to the presence of attributes and values of the data field.
- *Uniqueness*
  Uniqueness in data refers to duplicate and dummy entries in data fields.

Big Data reliability[25] is also one of the significant concerns in data integrity as it is also one of the factors of data security in the cloud computing environment. Unanticipated errors and faults in extensive data analyses may lead to a noticeable gap between expected predictions and recognised patterns and the actual data insights. Figure 5 demonstrates the interconnected dependency of multiple big data reliabilities in the cloud computing environment.

5.1. Big Data Reliability

In a Cloud computing environment, big data reliability can be classified into three categories, namely hardware reliability, software reliability, and data reliability, as shown in figure 5.

5.1.1. Hardware Reliability

A statement of the ability of the hardware to perform its functions for some time. It is usually expressed as MTBF (mean time between failures)

5.1.2. Software Reliability

Software reliability is the probability that software will provide failure-free operation in a fixed environment for a fixed interval of time.

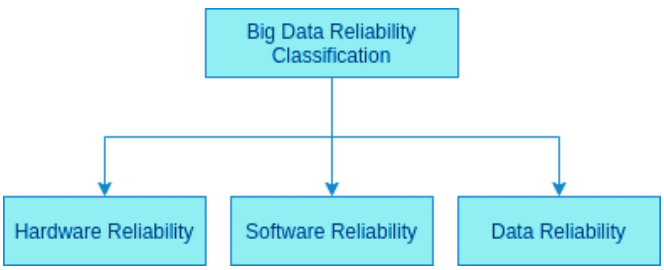

Fig. 4 Big Data Reliability Classification

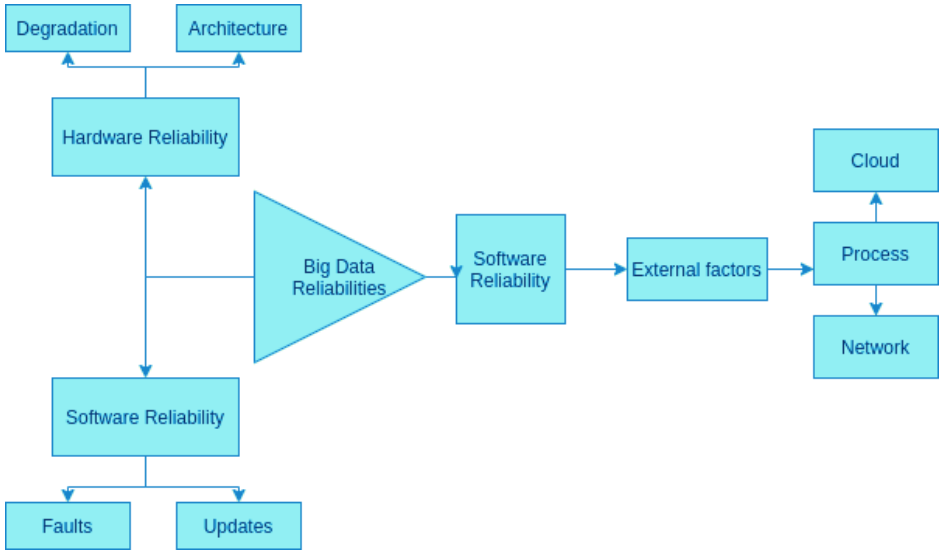

Fig. 5 Big Data Reliability Components

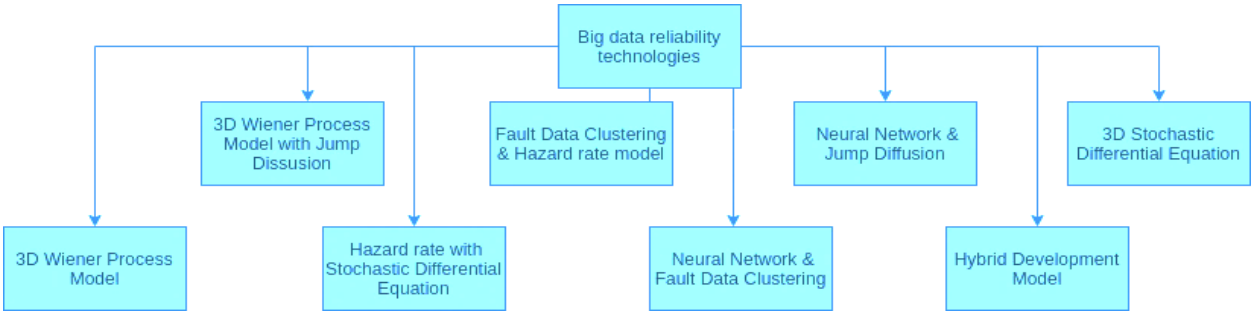

Fig. 6 Software Technologies for Big Data Reliability

5.1.3. Data Reliability

A collection of techniques ensuring data consistency in distributed settings, such as mobile environments. In this specific case, data inconsistency is due to fault or unavailability of wireless protocols and devices

5.2. Component-Based Reliability

The reliability of big data for managing the quality and development of software concerned can be accessed by applying multiple hazard-rate m and reliability growth models. The cloud computing paradigm in big data can be considered effective in assessing the reliability of big data.

Let the random variable denoted by Xk (where k=1, 2,….) represents the time interval between two consecutive software failure represented by (k −1) -th and k -th big data accession, so the function Zk(x) for the hazard rate in the phase of operation at any point of time x is given[19] by:

$$z_k(x) = \varphi\{ N - (k-1) \} \quad (k = 1, 2, , N; N > 0, \varphi > 0) \qquad (1)$$

where quantities are defined as below:

z (x) k is the rate of hazard for every database in the software and cloud,

N is the number of latent faults,

φ is the rate of hazard for every inherent fault.

Eq.(1) represents the rate of hazard for the database of software in case of occurrence of fault given by k, the time interval for every consecutive software fault in database k −1) th and k -th at the phase of operation the distribution function is written [16]as:

$$F_k(x) \equiv \Pr\{X \le x\} \quad (x \ge 0) \qquad (2)$$

As such, the probability of occurring any event X is given by denoted by Pr{X}

5.3. System Reliability

In the Big data and cloud computing paradigm, the reliability function[16] denoted by r(t) represented at the time time-interval t, before which the component of the system does not fails, i.e. until time interval t, the components of ai given by xi in case of normal operation the probability indicating the reliability function is written as:

$$r(t) = e^{-\left(\frac{t}{\alpha}\right)^{\beta}} \quad (3)$$

where α represents the parameter of scaling and β is the parameter representing the shape

case1:$\beta < 1$, in this case, the early failure period of a component is depicted by the reliability function

case 2: $\beta = 1$, the component's reliability is reliable or valuable throughout life.

Case 3: $\beta > 1$, the component is reliable, and wear-out period performance is observed.

6. Conclusion

Big data is gaining considerable impact on business decision-making by allowing policymakers to mine real-time business insights, the attributes of volume velocity and velocity are major concerns in analysing the massive amount of data. Cloud computing services, i.e. SaaS, PaaS, and IaaS, provide the major indispensable vitals to support underlying storage, processing, and analysis of big data. In this paper, the discussion begins with big data classification, characteristics of big data, tools and technologies for big data, and research issues of big data. The service models which turned to on-demand service, the concept of cloud computing and the relationship between cloud computing and big data are also discussed. Furthermore, the Big data reliabilities aspects, including software reliability along with tools and technologies, hardware reliability, and data reliability, are also highlighted in this paper. Finally, the reliability function and hazard rate for component-based and system reliability is also equated.

Acknowledgment

The authors pay tribute and dedicate this research work to Late Mr. Sunil Kumar Patel for his contribution and dedication. May his soul rest in peace.

References

[1] Charmaz K, and A. Bryant, "*The SAGE Handbook of Grounded Theory: Paperback Edition*," 2010.

[2] Neves, Pedro Caldeira, Bradley Schmerl, Jorge Bernardino, and Javier Cámara, "Big Data in Cloud Computing: Features and Issues," *International Conference on Internet of Things and Big Data*, 2016.

[3] Nabeel Zanoon, Abdullah Al-Haj, Sufian M Khwaldeh, "Cloud Computing and Big Data is there a Relation between the Two: A Study," *International Journal of Applied Engineering Research*, vol. 12, no. 17, pp. 6970-6982, 2017.

[4] Laney D, "*3D Data Management: Controlling Data Volume, Velocity, and Variety,"* Technical Report, META Group, 2001.

[5] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, & Byers A. H, "*Big Data: The Next Frontier for Innovation, Competition, and Productivity*," McKinsey Global Institute, pp. 156, 2011.

[6] Jacobs A, "The Pathologies of Big Data," *Communications of the ACM*, vol. 52, pp. 36–44, 2009.

[7] Bharadwaj A, El Sawy OA, Pavlou PA, Venkatraman NV, "Digital Business Strategy: Toward a Next Generation of Insights," *MISQ*, vol. 37, no. 2, pp. 471–482, 2013.

[8] Abbasi A, Sarker S, Chiang RH, "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems*, vol. 17, no. 2, pp. 1–32, 2016.

[9] J. Roski, G.W. Bo-Linn, T.A. Andrews, "Creating Value in Health Care Through Big Data: Opportunities and Policy Implications," *Health Affairs*, vol. 33, pp. 1115-1122, 2014.

[10] S. Sami and N. Sael, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, 2016.

[11] K.Iswarya, "Security Issues Associated With Big Data in Cloud Computing," *SSRG International Journal of Computer Science and Engineering*, vol. 1, no. 8, pp. 1-5, 2014. Crossref, https://doi.org/10.14445/23488387/IJCSE-V1I8P101

[12] G. Muneeswari, R. Surendiran, J. Jeneetha Jebanazer, P. Josephin Shermila, E. Anna Devi and A. Jeyam, "Urban Computing: Recent Developments and Analytics Techniques in Big Data," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7, pp. 158-168, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I7P217

[13] E.Kesavulu Reddy, "The Analytics of Clouds and Big Data Computing," *SSRG International Journal of Computer Science and Engineering*, vol. 3, no. 11, pp. 31-35, 2016. Crossref, https://doi.org/10.14445/23488387/IJCSE-V3I11P107

[14] Rajak, Nidhi & Rajak, Ranjit, "Performance Metrics for Comparison of Heuristics Task Scheduling Algorithms in Cloud Computing Platform," *Machine Learning Approach for Cloud Data Analytics in IoT*, 2021. Crossref, https://doi.org/10.1002/9781119785873.ch9.

[15] S Walia, Ranjit Rajak, Mohammad Sajid, "E-Commerce with Fog-Enabled Cloud Computing: Framework, Opportunities, and Challenges," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 13, 2022.

[16] S. Sharma, M. Sajid, "Integrated Fog and Cloud Computing: Issues and Challenges," *International Journal of Cloud Applications and Computing (IGI)*, vol. 11, no. 4, 2021.

[17] M. Sajid, Z. Raza, "Cloud Computing: Issues & Challenges," *International Conference on Cloud, Big Data and Trust (ICCBDT)*, RGPV, pp. 35-41, 2013.

[18] Helen Anderson Akpan, B.RebeccaJeya Vadhanam, "A Survey on Quality of Service in Cloud Computing," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 27, no. 1, pp. 58-63, 2015.

[19] Chang PC, “Reliability Evaluation and Big Data Analytics Architecture for a Stochastic Flow Network with Time Attribute,” *ANN Oper Resource*, vol. 311, pp. 3–18, 2022. https://doi.org/10.1007/s10479-019-03427-4.

[20] S. Subatra Devi, "Big Data - Benefits and its Growth," *International Journal of Computer Trends and Technology*, vol. 68, no. 5, pp. 14-17, 2020.

[21] Ramya D, Ramyashree P R, Sunaina Rashmi R, Nalina V, "Green Cloud Computing-A Review," *International Journal of Recent Engineering Science*, vol. 5, no. 6, pp. 16-18, 2018.

[22] Rajak, Nidhi & Rajak, Ranjit & Prakash, Shiv, “A Workflow Scheduling Method for Cloud Computing Platform,” *Wireless Personal Communications*, pp. 1-23, 2022. Crossref, https://doi.org/10.1007/s11277-022-09882-w

[23] Rajak R, Kumar S, Prakash S, et al., “A Novel Technique to Optimise Quality of Service for Directed Acyclic Graph (DAG) Scheduling in Cloud Computing Environment Using Heuristic Approach,” *The Journal of Supercomputing*, 2022. Crossref, https://doi.org/10.1007/s11227-022-04729-4

[24] Sharma, Shalini, Kumar, Naresh, and Kaswan, Kuldeep Singh, “Big Data Reliability: A Critical Review,” *IOS Press*, pp. 5501-5516, 2021.

[25] X. Wu, X. Liu and S. Dai, "The reliability of Big Data," *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, pp. 295-299, 2014. Crossref, https://doi.org/10.1109/ITAIC.2014.7065054.