# Challenges and Issues in Data Analytics

Saurabh Singhal[a], Amit Kr. Gupta[b], Ruchi Rani Garg[c]

[a] Assistant Professor, KIET Group of Institutions, India
[b] Associate Professor, KIET Group of Institutions, India
[c] Associate Professor, MIET Group of Institutions, India.

***Abstract**:* Latest advancements in the field of Digital Technologies like Cloud Computing & Internet of Things continuously generating very large amount of data due to the implementation of newly invented information systems. Lots of efforts are required for extraction and analysis of this huge amount of data in order to make effective, efficient and accurate decision making. Thus, analysis in the field of big data created a very wide domain for research and development. In this paper, our objective is to get aware about impact of big data challenges, issues and tools deployed. By this paper we try to generate a platform for exploration of big data at several stages. Besides this it also led out certain issues which encourage researchers to provide solutions against the issues and challenges.

***Keywords:*** Hadoop, Data *Analytics in Big Data, Types of Data – Massive Data, Tremendous Data, Structured Data, Unstructured Data.*

## I. INTRODUCTION

In the current scenario, a large amount of data is generated by the dynamic change and migration of digital technology. This brings innovative breakthroughs to all areas thanks to the large collection of data collected. Big Data can be defined as data that requires new architecture, techniques, algorithms, analysis, scale, diversity, complexity to manage and extract hidden knowledge and value. Data, including diversity, quantity, and speed information, requires a new form of processing that enables improved decision making, knowledge discovery, and process optimization.

Typically, this is a very large and complex data set that can not be addressed by implementing old database management tools or traditional database management tools and applications. These datasets can be used in a structured, semi-structured, and unstructured form, with data sizes larger than petabytes. It contains the concept of 3 V, which is further extended to 4 V. Figure 1 [1] provides a clear scenario for V.

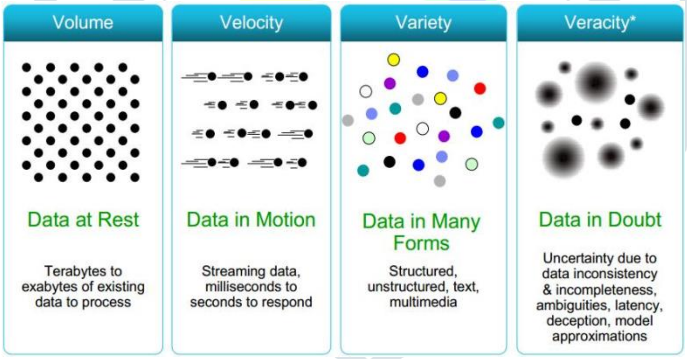
Volume
Velocity
Variety
Veracity*
Data at Rest
Data in Motion
Data in Many Forms
Data in Doubt
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

**Figure (1) Representing 4 V's in Big Data**

3 V stands for volume, speed and variety. The volume shows the huge amount of data generated each day. Speed represents the rate of growth and the speed of data collection. The variety indicates the type of data, whether structured, semi-structured or not. The fourth V refers to Veracity which emphasizes availability and responsibility. The primary goal of Big Data analysis is to process and analyze large datasets by considering all V's using traditional and computer-based intelligent techniques and tools [1].

In 2015, big data reached 25 billion [2]. According to the point of view of ICT (Information and Communication Technologies), big data is a powerful driver of the next-generation computer domain [3], which constitutes a new platform that has been proven. Previous data warehouses have been introduced to handle large data sets, but the important problem is extracting relevant information. As a result, the available data mining techniques can not handle it properly because of the lack of coordination between the DBMS and the analytical tools used for statistical analysis and data extraction or retrieval. The basic problem is to quantitatively describe the essential characteristics of big data, and an epistemological significance is needed to describe the revolution and evolution of data [4].

In addition, it is important to consider that all available data in the form of big data is not relevant for analysis and decision-making. Therefore, in the fields of computer science and education, we try to classify the results generated by large-scale data analysis into two parts. In this article, we will focus on the challenges of large data and develop several research topics. To do this, we classify this article in the following section:

**II.** Challenges encountered during fine tuning of Big Data
**III.** Research issues for processing Big Data and Extraction of Relevant Information
**IV.** Current Big Data tools and techniques
**V.** Summarizing outcomes and Conclusion

## II. Challenges encountered during fine tuning of Big Data

In our time, large-scale data collection is accepted in many areas such as retail, health, government, and other scientific research. All web applications such as social computing, Internet texts and documents, indexing Internet searches, etc. frequently encounter large data. The field of social computing includes social network analysis, the online community, the arbitration and reputation system, the forecast market. ISI, IEEE Xplorer, Scopus, Thomson, Reuters etc., but the advantages and benefits of big data will create new possibilities in the world of knowledge processing for future

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

**Figure (2) Representing various sources of Big Data**

To process a task first, you need to know the computational complexity, the information security, and the calculation method for analyzing large data. For example, several statistical methods work very well for small data blocks, but work in the worst case because the data becomes volumes. Similarly, some computer techniques give the best results when applied to small amounts of data, but face specific serious problems when processing large amounts of data [3]. Here, we have categorized the problems and tasks to classify big data into four categories.

**a).** Data Storage Analysis
**b).** Knowledge Discovery and Computational Complexities
**c).** Scalability and Visualization of Data
**d).** Information Security

### 2.1 Data Storage Analysis

For data sources such as mobile equipment, mood sensor technology, remote sensing, wireless identification reader, data size is growing exponentially. Storing this data requires high cost and space. Therefore, the biggest challenge of big data is to identify the storage medium and the speed of I / O. In this scenario, you need to make data accessibility a priority to discover knowledge and data representation. . The main reason is that you can access it whenever you need it for later reference in a practical way. In the past, the analyzer used the hard drive for storage, but had negative effects on I / O performance. To overcome it, the concept of SSD (Solid

State Drive) and PCM (phrase change memory) has been introduced. However, the available storage mechanisms do not meet the performance criteria when processing large data.

As the size of the data set grows, another problem is the diversity of the data.

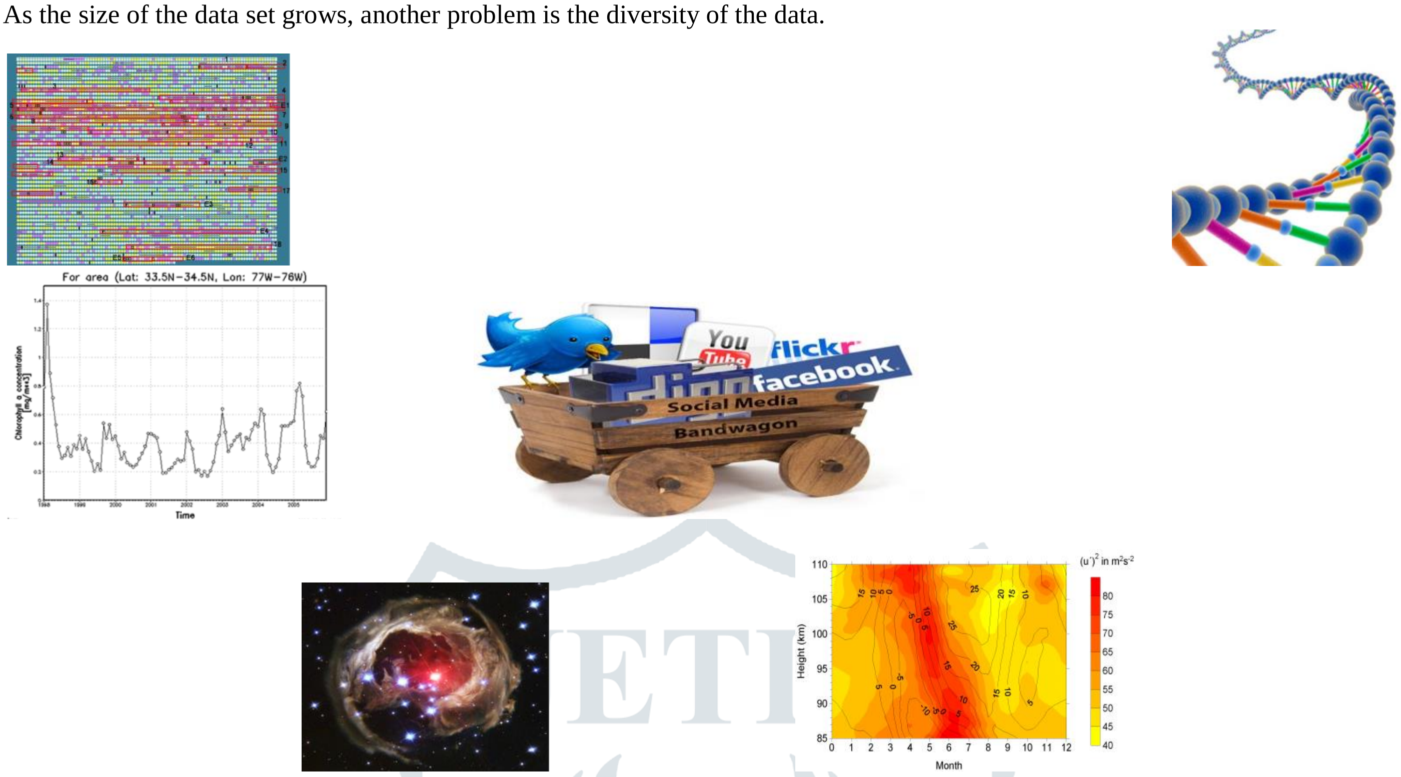

**Figure (3) Representing various forms of Big Data**

Since existing algorithms do not respond appropriately, selection and reduction of data, selection of functions is an important task while dealing with large data sets that pose unprecedented challenges to researchers. The automation of this process and the development of new algorithms are major concerns for future researchers. The latest emergence of technologies such as Hadoop and Map Reduce can collect large amounts of data at the right time, whether structured or not. We need to design a process to convert semi-structured data and unstructured data into structured data and apply a mining algorithm to extract related information [5] [6].

The main concern is the design of the storage system and an efficient tool for analyzing large data, ensuring accurate output even if the data comes from different sources. In addition, the development of specific algorithms is essential to improve the efficiency and scalability of large-scale data analysis.

### 2.2 Knowledge Discovery and Computational Complexities

The big problem with big data is knowledge discovery and expression, which includes authentication, archive management, storage, and information acquisition. So far, technologies have been developed for fuzzy sets [7], raw sets [8], soft sets [9], close sets [10], analysis of formal concepts [11], Principal Component Analysis [12] Not suitable for large data sets. Although good results are obtained with parallel computing, the size of the data set has increased considerably, so that the available tools have not been able to process the information effectively. The most common approach is the data warehouse and the data mart application, the data warehouse stores data from the feeder system, the datamart facilitates the analysis.

The first problem is the management of inconsistencies and uncertainties in large datasets. Generally, a systematic model of computer complexity is implemented. Rather, it may be difficult to develop and implement a comprehensive mathematical model applicable to the entire dataset, but rather by managing some specific complexities, a domain-specific analysis is done in a practical way. You can In order to minimize the complexity and computational costs, a lot of research and research have been done in this dimension, using the concept of machine learning with minimal use of memory [13].

However, the currently implemented Big Data tool displays very poor performance while managing complexity, uncertainty, and calculation gaps. Developing new tools and technologies to compensate for complexity, uncertainty and calculation gaps efficiently and effectively is a major challenge.

### 2.3 Scalability and Visualization of Data

Scalability and security is an absolute priority of the techniques that are implemented when analyzing large volumes of data. Moore's law, by increasing the speed of the processor, to speed up the process of data analysis, was implemented by the researchers. Sampling, online, multi-tire solution is an essential part of Moore's Law. Concern about big data analytics, excellent scalability features has been held by incremental technology. The data size is growing at a very fast pace compared to the processing speed of the CPU, must have a dynamic change of process technology embedded in the [14].

Visualizing the data, by implementing a number of techniques, such as graph theory, should be presented in a more appropriate manner. It ensures the link between data sets with the appropriate interpretation. However, today, online shopping websites, because it does not have to deal with hundreds of millions of users every day, has generated a large amount of data. Is a specific tool array is implemented for viewing large amounts of data, it has the ability to convert complex and huge data into an intuitive image. Sentimental analysis, track customer feedback, and enable the relevance of the research. In contrast, the big data tools that are currently employed, feature, and inferior in scalability, and response time.

Parallel Computing, distributed computing, in connection with the development of hardware and software in the field of Big Data to create a growth and development situation, such as cloud computing, and many challenges and problems occur , computer relations in the field of science and mathematical models.

### 2.4 Information Security

In the process of analyzing large volumes of data, the first association of large amounts of data is performed, the extraction of data is performed in a meaningful pattern, the final analysis is performed. Organization to protect sensitive information, has implemented a variety of policies. In large amounts of data analysis, the protection of confidential information is the main concern. Because there is a high risk [15]. At present, information security problem has been turned into a big data problem. Authentication, authorization, implementing technology, such as encryption, will be able to reinforce security aspects in the field of large-scale data. The security measures that large-scale data applications face, scaling the network, heterogeneous devices, real-time monitoring, lack of penetration system [16]. To remedy this, you need to design a multi-level prevention system of security policy.

## III.Research issues for processing Big Data and Extraction of Relevant Information

Industry and academic sectors have embraced the field of large-scale data science and data analysis as the final axis of research and development. The science that processes a large amount of data (big data) and allows a well-informed extraction of large data is called "data science". Data science applications can be summarized as follows:

- Information Science
- Uncertainty Modeling
- Uncertain Data Analysis
- Machine Learning
- Statistical Learning
- Pattern Recognition
- Data Warehousing
- Signal Processing

An effective and efficient fusion of technology and analysis leads to accurate prediction of future events that we call "data modeling". In this section, the most important problem is to discuss the problem generated by Big Data Analytics. We have largely categorized these problems in the following areas.

**a).** IoT for Big Data Analytics
**b).** Cloud Computing
**c).** Bio Computing
**d).** Quantum Computing

### 3.1 IoT for Big Data Analytics

The IoT, for a myriad of the myriad of autonomous gadgets is connected to the Internet, it contains the concept of MM (Machine-to-Machine) communication, which results in an increase in the structural and cultural revolution and the number of users. IoT is the center of attraction for researchers because of promising opportunities and challenges. When designing a future communication network, IoT has a vital effect. In this case, everything is connected and must be controlled intelligently. Mobile devices, ubiquitous and integrated communication technology, the development of cloud computing and data analysis, the IoT domain is increasingly important. IoT is very interested in the concept of large data of 3V and 4V. Moreover, the IoT itself is a mystery in terms of definition and structure [17].

Large-scale data experts face the greatest challenge in acquiring knowledge from IoT data. IoT device is to generate a continuous data, the researchers introducing the concept of machine learning, has developed tools and techniques to extract meaningful information from this data. Big Data analysis includes understanding IoT data and extracting meaningful information. From the point of view of IoT, the only possible solution is an implementation of machine learning algorithms and intelligent calculation techniques [18].

### 3.2 Cloud Computing

The concept of supercomputing is becoming more accessible and affordable for the development of virtualization technology. The IT infrastructure must be hidden so that the virtualization software behaves like a real computer. Such a virtual machine placement is the domain of cloud computing. In order to provide availability of scalability and on-demand resources and data, both large scale data and cloud computing technologies have been developed. Cost reduction and availability are improved [19] [20].

Big Data applications must support the analysis and development of data when deployed through cloud computing. In the cloud environment, it provides tools for interactive and collaborative environments, allowing you to acquire and extract knowledge from data and business analysts.

In the case Cloud computing can store large amounts of data, but downloading and downloading data can have a negative impact on performance, so you need to consider time and cost factors. Since the coordination between the underlying software and the hardware is inadequate, it can be difficult to spread the calculations. Here, the main concern is the security for the data exposure to the published server. All these problems constitute a new branch of research and development.

### 3.3 Bio Computing

Computer science has evolved because of the interference and intervention of biological and natural mechanisms such as DNA and proteins. Such a system is self-organized without central control. With the introduction of bioinformatics, you can achieve cost-optimized search algorithms, optimal data service solutions, and minimal cost over data management and service maintenance. DNA and proteins are used to perform computer calculations involving the storage, retrieval and processing of data. The main feature of biological informatics is the integration of biological materials to perform computational functions and achieve intelligent performance. These are ideal for large data applications.

Thanks to digitization around the world, a lot of data is generated on the Web. To analyze these data and categorize them as text, images, videos, etc., you need to implement typical analytical tools and techniques from data scientists and leading data experts. Several technologies have been developed to handle large data, but in order to analyze economically it is essential to choose an appropriate platform.

Bioinformatics plays an important role in intelligent data analysis. The algorithm designed here allows data mining from huge data sets for optimized applications. First, it is simplicity and rapid convergence towards the best solution while solving the problem of service provision [22]. By discussing bioinformatics, it helps in smarter interactions, inevitable data loss, ambiguities management, and can be viewed and implemented as future computing.

### 3.4 Quantum Computing

The memory of the quantum computer is exponentially larger than the physical memory and the exponent inputs can be used simultaneously [21]. By implementing the concept of quantum computing in a real-life scenario, it is possible to solve a complex problem of large data, but the technical difficulty lies in the development of a real quantum computer. The implementation of quantum computing includes the fusion of quantum mechanics and information processing technology. In conventional computers, information can be presented in the form of bit strings, but quantum computers can present information in the form of q bits or quantum bits. The difference between qbit and bit is that qbit is a quantum system that codes 0 and 1 into two distinct quantum states. Therefore, he can take advantage of the phenomenon of superposition and entanglement.

## IV. Current Big Data tools and techniques

The At the moment, the tools you have some to handle large amounts of data. Part of the important tool emerging Map Reduce, Spark Apache, it is the storm. Most tools, batch processing, flow processing, follows the concept of interactive analysis. Most of the batch tool, based on Hadoop infrastructure such as Mahout and Dryade. Stream processing tools such as Strom and Splunk is used for real-time applications. Interactive processing tools, such as Drill Drenel and Apache's, allows the user to interact in real time for analysis. Such a tool is useful in the development of large scale data projects.

For some of the important tools are described as follows.

**1).** Apache Hadoop and Map Reduce
**2).** Apache Mahout
**3).** Apache Spark
**4).** Dryad
**5).** Storm
**6).** Apache Drill
**7).** Jaspersoft
**8).** Splunk

### 4.1. Apache Hadoop and Map Reduce

Apache Hadoop and Map Collapse is the most established software platform for analyzing large amounts of data. HDFS (Hadoop Distributed File System), Hadoop's kernel, map collapse, was built as Apache Hive. Map Reduce is a programming model for processing large data sets according to the strategy of division and Conquer.

The implementation of the strategy and Divide Conquer is done in the next two steps.

**a).** Map stage

**b).** The reduction of the stage

Hadoop works on two types of nodes that master and worker node nodes. The master node divides the problem into small subproblems and then distributed to the worker node of the Step card. Then the master node will associate the output of all the sub-questions to reduce the step. In addition, Hadoop and Map Minimize acts as a powerful software framework to solve the big data problem. also contributes to fault tolerant data storage and processing.

### 4.2 Apache Mahout

Scalable for large-scale intelligent data applications on a large scale in aims to provide a commercial machine learning techniques. The Mahauto kernel algorithms, grouping, classification, mining model, regression, dimension reduction, scalable algorithm includes batch-based collaboration filtering performed on the Hadoop platform via a Reduce Framework card. The purpose of Mahoto, in order to facilitate the discussion of project cases and the potential use, dynamic, there is a reaction, is to build a diverse community. The basic goal is to provide a tool to solve the problems of large amounts of data. Google, IBM, Amazon, Yahoo, Twitter, several companies such as Facebook has implemented a scalable machine learning algorithm [27].

### 4.3 Apache Spark

Apache Spark was developed by UC Berkeleys AMP Labs in 2009. In open source, which was released in 2010 for large-scale data processing, the framework was built for sophisticated analysis and processing at high speed. Users, Java, Python, you can quickly describe the Scala application. In addition to reducing map, SQL queries, continuous data, machine learning, it also supports the data of the processing graph. It works on the top of HDFS. This pilot program, Cluster Manager, consists of several elements: worker nodes. The main goal is to store data in memory, a certain distributed elastic dataset that provides fault-free replication failure (RDB). It supports interactive calculations, to improve efficiency in the use of speed and resources.

### 4.4 Dryad

This is another important programming model for the implementation of parallel and distributed programs in order to deal with the large context based on the data flow graph. This includes the cluster of compute nodes, which users use the cluster resources to accomplish distributed to the program. Dryade uses thousands of machines. In each of the machines, there is a plurality of processors or cores, users do not need the concept of concurrent programming. the dryad application works in the realized calculation graph consists of calculation vertex and calculation channel. Therefore, the production of job graphics, process planning machines available, the processing of cluster transition defects, the collection of performance indicators, the visualization of jobs, calls for a user-defined policy, such as the dynamic update of the employment graph corresponding to the policy offers a wide range of functions. Take a decision [26].

### 4.5 Storm

IStorm is a distributed, fault-tolerant, real-time computer system for processing flow data. This is in contrast to the hatch (which is used for batch processing), it was specially designed for processing. In addition, in order to provide competitive performance, tuning, it is operation, simple scalable and fault-tolerant configuration. Storm group is similar to the harp cluster cluster. In the storm cluster user, but to run different topologies for various tasks, the implementation of the Hafupu platform will reduce the work of the corresponding application. Storm group is composed of two types of nodes of the master node and the nodes of workers. These two nodes, respectively, implement both roles of Nimbus and the supervisor. Two of the role of these, you can perform similar tasks, such as job tracking tool and job tracking plan. Nimbus is responsible for distributing code across the storm group, scheduling tasks, assigning work nodes, and monitoring the entire system. The supervisor compiles the task assigned by Nimbus. In addition, the workstation can start and end the process according to Nimbus instructions. All computer technology is divided and distributed among a large number of work processes, each work process implementing a part of the topology.

### 4.6 Apache Drill

This is another distributed system for performing interactive large data analysis. It adapts flexibly to query language, data format, and data source type. It is specially designed to search for nested data. Up to 10,000 servers
can be expanded and petabytes of data and billions of records can be processed in seconds. Use HDFS for storage and reduce the map to perform a batch scan.

### 4.7 Jaspersoft

It is an open source software and provides reports from the database. Scalable for large data, it also has a considerable display function for platforms such as MangoDB, Cassanndra, Redis, etc. You can perform extended discovery of datasets without extracting, transforming, or importing. In addition, you can create powerful HMPL dashboard reports and reports from large-scale data without ETLs.

### 4.8 Splunk

Recently, the business sector generates huge amounts of data via machines. It is a real-time intelligent platform designed to search for data generated by a machine. The main functions are structured unstructured machine indexing, data generation, real-time search, analysis result reporting and dashboard. It's a combination of big data and cloud technology. It allows users to process large data via the web interface. The results of the analysis are displayed intuitively, such as graphs, reports, and so on. The main objective is to provide application measures, diagnose problems, process the technological infrastructure and provide intelligent support for business operations.

## V. Summarizing outcomes and Conclusion

Collecting data from multiple applications is growing very fast each day. You can see that these data are only useful if they have been analyzed intelligently. Leading to the development of new technologies that can facilitate large-scale data analysis. The development of a powerful computer is an advantage for such a technical implementation and leads to the birth of an automated system. Converting data into well-informed information is not easy because it requires parallel computing and the processing of a complex data mining architecture. Many different models, such as fuzzy sets, raw sets, soft sets, and neural networks, are useful in representing data. In many cases, large data will be reduced to include only the critical characteristics needed for a particular search perspective or application domain. Therefore, a specific reduction technology has been developed. In many cases, the collected data has missing values and you must generate these values. Tuples with missing values are removed from the dataset before the scan. As a result, new challenges can arise in terms of the performance, efficiency, and scalability of dedicated data-intensive computing systems. In some cases, this leads to a loss of information and is not preferable. In addition, fast processing, high performance, efficient storage, programming and broadband are other combustion problems [28].

Tools and machine learning techniques are gaining popularity among researchers to facilitate meaningful results from the above concept. Research in this area is concerned with data processing, algorithm porting, optimization. There are advantages and disadvantages to each tool, but to deal with issues in the Big Data areas, you can develop more effective tools. These tools require preparation to handle noisy and unbalanced data, uncertainties and discrepancies, missing values.

Towards the end of this research article, we focused on various research topics, tasks and tools deployed to analyze large data. It is understood that to provide intelligent analysis, it is necessary to focus on any large data platform. Some are superior to batch processing, real-time analysis, statistical analysis, machine learning, cloud computing, quantum computing, data flow processing, and more. In the near future, researchers need to pay more attention to the concept of Data Veracity. Data Veracity is at both ends. At the input level, you must enter the appropriate data into the system and you must provide meaningful information at the user level. Therefore, to develop technologies that effectively solve big data problems, we need to focus on meaningful extraction concepts.

## REFERENCES

**[1].** M. K. Kakhani, S. Kakhani and S. R. Birdar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering and Management, 2(8) (2015), pp. 228 – 232.

**[2].** C. Lynch, Big Data: How do your data grow? Nature 455 (2008), pp. 28 – 29

**[3].** X. Jin, B. W. Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2 (2) (2015), pp. 59 – 64.

**[4].** R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1 (1) (2014), pp. 1 – 12.

**[5].** T. K. Das and P. M. Kumar Big data analytics : A framework for unstructured data analysis, International Journal of Engineering adn Technology, 5 (1) (2013), pp. 153 – 156.

**[6].** T. K. Das, D. P. Achariya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, International Conference on Computer Communication and Informatics, 2014.

**[7].** L. A. Zadeh, Fuzzy Sets, Information and Control, 8 (1965), pp. 338 – 353.

**[8].** Z. Pawlak, Rough Sets, International Journal of Computer Information Science, 11 (1982), pp. 341 – 356.

**[9].** D. Moldtsov, Soft Set theory first results, Computers and Mathematics with Applications, 37 (4/5) (1999), pp. 19 – 31.

**[10].** J. F. Peters, Near Sets, General theory about nearness of objects, Applied Mathematical Sciences, 1 (53) (2007), pp. 2609 – 2629.

**[11].** R. Wille, Formal Concept analysis as mathematical theory of concept and concept hierarchies, Lecture notes in Artificial Intelligence, 3626 (2005), pp. 1 – 33.

**[12].** I. T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.

**[13].** O. Y. AI – Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data : A review, Big Data Research, 2 (3), (2015), pp. 87 – 93.

**[14].** A. Jacobs, The pathologies of Big data, Communications of the ACM, 52 (8) (2009), pp. 36 – 44.

**[15].** H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, International Conference on Information Technology and Management Innovation, 2015, pp. 1041 – 1044.

**[16].** Z. Hongium, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congresso da sociedata Brasileria de Computacao 2014, pp. 1 – 6.

**[17].** N. Mishra, C. Lin and H. Chang, A cognitive adopted framework for iot big data management and knowledge discovery prospective, International Journal of Distributed Sensor Networks, 2015, (2015), pp. 1 – 13.

**[18].** X. Y. Chen and Z. G. Jin, Research on key technology and applications for Internet of Things, Physics Procedia, 33, (2012), pp. 561 – 566.

**[19].** M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. S. Netto and R. Buvya, Big Data Computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp. 3 – 15.

**[20].** I. A. T. Hshem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing : Review and open research issues, Information Systems, 47 (2014), pp. 98 – 115.

**[21].** M. A. Nielson and I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, New York, USA 2000.

**[22].** L. Wang and J. Shen, Bioinspired cost – effective access to big data, International Symposium for Next Generation Infrastructure, 2013, pp. 1 – 7.

**[23].** M. Herland, T. M. Khoshgoftaar abd R. Wald, A review of data mining using big data in health informatics, Jounral of Big Data, 1 (2), (2014), pp. 1 – 35.

**[24].** T. Huang, L. Lan, X. Fang, P. An, J. Min and F. Wang, Promises and challenges of big data computing in health sciences, Big Data Research, 2 (1), (2015), pp. 2 – 11.

**[25].** C. L. Philip, Q. Chen and C. Y. Zhang, Data – intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), pp. 314 – 347.

**[26].** H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Data flow graph runtime, 2nd International Conference on Cloud and Green Computing, 2012, pp. 675 – 683.

**[27].** G. Ingersoll, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White paper, IBM Developer Works, (2009), pp. 1 – 18.

**[28].** D. P. Acharjya, S. Dehuri and S. Sanyal Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978 – 3 – 319 – 16597 – 4, 2015.