ISSN No:-2456-2165

A Informative Study on Big Data in Present Day World

Presented by: Surekha Pinnapati
Research Scholar
RNSIT, Bengaluru

Dr. Prakasha S
Associate Professor
RNSIT, Bengaluru

Abstract:- Daily, emerging information management and modern technology such as that of the Iot devices services represent a big pool of huge amounts of data. Research of all this massive volume of data involves a tremendous amount of work to generate information to decision makers at multiple levels. Advanced analytics, however, would be a current innovation field. This article's fundamental purpose is to examine the future effects of big data processing, open issues concern, and various related resources. As a consequence, this review offers a forum for various phases of development of big data. It also opened up a new horizon for researchers, based on the problems, to improve the solution.

***Keywords*:-** *Big Data Analytics; Hadoop; Massive Data; Structured Data; Unstructured Data.*

I. INTRODUCTION

Information is obtained from multiple studies mostly in internet age and indeed the gradual change from digital technology had led to the expansion of information technology. In several areas, the processing of large databases offers adaptive advances. In common, it relates to the collection of huge data sets that are difficult to handle using standard methods for managing databases or technologies for information processing. These were all collected in Terabytes of data and beyond in organized, semi-structured, and unstructured formats. Explicitly, from 3Vs to 4Vs is described. Volume, velocity, and variety apply to 3Vs. Problems pertaining to the overload of information that is already being produced.

Number of different shows information among file formats also including structured, unstructured, semi-structured, etc. The fifth V relates to truthfulness involving accessibility but transparency. The prime focus of predictive analytics will be to use various extraction and quantitative selected sources to manipulate information with high quantity, velocity, variety, and veracity. the researcher Haider addressed most of these retrieval techniques for collecting [1]

Helpful details. Figure 1 below makes reference to the concept of large datasets. The actual concept of big data, furthermore, is not specified although there is a misconception that it would be real concern[2]. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective.

Increasing production of Data Mining is projected to hit 25 billion by 2015 [3]. Data mining is a strong catalyst from both the technological viewpoint during the next age of IT sectors, which have been essentially based upon this third model, specifically related to large datasets, data storage, the Internet of Things, and social enterprise. In general, the large dataset was handled using data warehouses [4]. In this case, it is a primary problem to generate accurate information from either the big data available. Plenty of the processes described in knowledge discovery are typically only prepared to handle data sets.

That cooperation between cloud computing including with analytical methods including predictive analytics research is a serious factor also in massive data. Because they try to conduct information exploration and recognition for all of its potential implementation, these problems usually emerge [5]. Well, how explain the basic sources of information quantifiable is a current issue. In defining that spatial and frequency domain, can see the need for philosophical consequences. In comparison, the research upon this philosophy of simplicity [4] of big data that can help explain significant factors and establish complex trends in big data, refine their description, provide improved approximation of information. Much research was carried out by various researchers on big data and its trends [6], [7], [8].

It should, moreover, be remembered that almost all knowledge access to lots of big data isn't really suitable also for process of the research or judgement. In order to disseminate the results of big data, business and universities are involved. The objective of this research has been on big data problems and the strategies applicable to them. Likewise, in large datasets, we mention collaborative research question. So, each document is split into subsequent parts to illustrate this. Sections 2 addresses the problems that occur while data analytics fine tuning. Part iii open multiple research questions which could assist us just to manage and derive valuable information from large data sets. An introduction into data analytics methods and techniques is provided in part 4.

II. CHALLENGES IN BIG DATA ANALYTICS

Information technology is accrued in many fields especially, such as healthcare coverage, public policy, biochemistry, and many other intersectional studies. Big data, such as social software, network text and records, but online research scanning, is commonly used in web-based systems. Social computing requires study of social networks, online groups, recommendation systems, Integrity structures and prediction markets, where ISI, IEEE Xplorer, Scopus, Thomson are used as internet search indexing.

ISSN No:-2456-2165

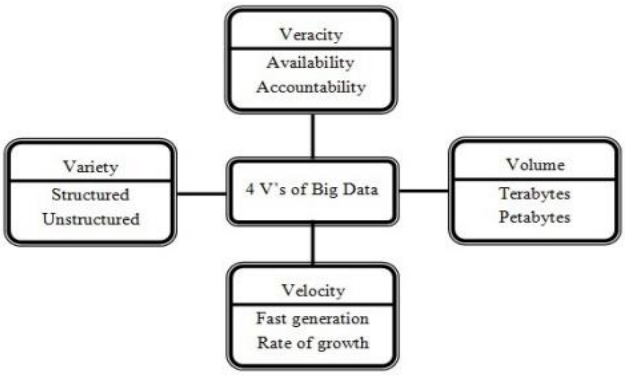

Fig. 1: Characteristics of Big Data

Designers therefore need understand the different computational speed, information security and computational techniques, to interpret big data in order to respond to these challenges. Few data sets that operate very well small size sizes, for particular, don't really scale to huge datasets. Analogously, in big data analytics, many statistical methods that work really well low bandwidth face significant issue. The difficulties of business intelligence are categorized here too in four different categories: data warehouse and research information extraction and computational capabilities [9]; data robustness and image analysis; but instead, data management. Their complexities improving healthcare business were analyzed by several investigators [10].

A. Data Storage and Analysis

The size of the dataset has greatly increased in recent years through various methods, along with portable devices, aviation creating environments, satellite imagery, monitors for radio frequency recognition, etc. These data are stored on spending a lot of expenditures, although they are ultimately neglected or removed and there is not enough space to store them. Storage media and greater input/output speed are thus the first obstacle for big data analysis. In these kinds of instances, the main priority for information exploration and recognition has to be data connectivity. The key explanation is that, for some further research, it will be viewed conveniently and efficiently. Analysts have used solid state drives to keep records in past few decades, but sporadic input/output efficiency is weaker than continuous input/output. The idea of Solid State Drive (SSD) and Phrase Change Memory (PCM) has been implemented to address this constraint. Data collection and data storage technologies, therefore, will provide the requisite output for deep learning techniques. This variability of knowledge is due to that other difficulty in Big Data research. With the ever-increasing number of datasets, information retrieval challenges have risen exponentially. In particular, image segmentation, data collection and features are extracted are an important task, notably when working with large repositories.

That was because, in interacting with all of these heavy data, conventional solutions doesn't always adapt in a sufficient time [11]. In recent decades, automating this method and designing new machine learning methods to check the accuracy has been a big challenge. In comparison to some of these, it is of primary interest to bundle data sets that aid in the knowledge discovery. Advanced developments, along with hadoop and mapReduce, allow vast amounts of semi-structured and unstructured data to be collected in a reasonable period [12]. Where to test certain details efficiently to gain better information is your main technological challenge. Similarly detail explanation of data analysis for public tweets was also discussed by Das et al in their paper [13]. The major challenge in this case is to pay more attention for designing storage sytems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machine learning algorithms to analyze data is essential for improving efficiency and scalability.

B. *Knowledge Discovery and Computational Complexities*

The exploration and analysis of information is a main problem in big data. It encompasses a set of sub-fields, including security, scanning, governance, preservation, retrieval and representation of information. Information development and interpretation methods along with linguistic variables [14], fuzzy sets [15], soft set, close set, importance - performance analysis [16], key component analysis, to identify only some [17], are available. In comparison [18], several hybridized approaches are often produced to deal with real-life issues [19]. These methods are all problem-dependent. In addition, some of these approaches may not be ideal for large sequential machine datasets. Until the amount of big data continues to grow increasingly, the opportunities available might not have been successful in analyzing such data to gain useful information. In the case of large dataset management, data warehouses and data marts are the most common solution. The database engine is primarily ideal for maintaining data from operating systems, while data mart is founded on a data warehouse and allows analysis. More computational complexities include study of large datasets. The key challenge is the management of the contradictions and uncertainties in the datasets. Structured simulation of the performance of computations is commonly used [20] [21] [22].

C. Scalability and Visualization of Data

Very significant task is its user friendliness as well as data protection for big analytical tools. In recent years, significant attention has been paid to accelerating the data collection and speeding up processing in compliance with Moore's Law. For the long - time, sequencing, on-line, and multiresolution data analysis need to be developed [23]. In the area of big data analytics, accumulative methodologies have better connectivity properties. As the data size scales an even way quicker unlike CPU speeds, with only an expanding clock frequency, there is a natural dramatic shift in consumer electronics being encoded. This change in cpus is leading to the creation of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing.

The aim of data visualization is to use those complex mathematical techniques to show them more adequately. The relation with data and proper understanding is given by graphical visualization. Every quarter, therefore, digital services such as flipkart, amazon, provide millions of

ISSN No:-2456-2165

subscribers and billions of items to sell. A lot of stuff is generated by this. To just this end, some companies use a Tableau platform for the advanced analytics. It would have the potential to translate broad and complicated data into intuitive images. This allows an industry's workers to imagine the importance of the hunt, monitor recent customer reviews, and evaluate their feelings.

We will see that machine learning has provided numerous hurdles for hardware and software which really relate to cloud processing, cloud computing, distributed computing, phase of visualization, optimization. We will need associate more mathematical models with computer science to tackle this disadvantage.

C Information Security

Huge quantity of information in predictive analytics are related, examined, and explored for meaningful means. To preserve their confidential data, all organizations have multiple strategies [24]. A real problem in big data analytics is the security of confidential data. Big data has a serious security costs involved with it. Data privacy is, however, and became a advanced analytics concern. Big data protection can be strengthened while using verification, permission, and password protection. Network size, myriad of alternative devices, [25] [26] real time security monitoring, and shortage of interference method are multiple countermeasures faced by big data processing. Therefore, attention has to be given to develop a multi-level security policy model and prevention system.

While a great deal of research work has been done to protect predictive analytics [25], it needs a lot of development. A multi-level protection, sovereignty data model for big data is just the significant challenge.

III. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Information technology and machine learning in academia and business have now become the focusing area of consideration. The goal of data science is to study big data and the analysis of features from data. Data mining and data science implementations comprise interest in scientific, simulation of uncertainty, uncertain data analysis, machine learning, statistical learning, identification of patterns, data warehousing, and signal processing. In order to forecast the potential drift of events, successful incorporation of technology and research would result. In Big Data Analytics, the main purpose of this segment is to address open research issues.

Big data processing research problems are categorised into three broad categories, such as the Internet of Things (IoT), cloud services, nano computing, and artificial intelligence. It is not restricted to these problems, nevertheless. Further analysis commonly described to healthcare data can be located in the research article given by Husing [9] .

A. IoT for Big Data Analytics

Global interactions, business art, cultural movements and an enormous amount of self qualities have been reformed by the Internet. Systems are actually taking action to monitor countless interactive gadgets through the internet and build the Wearable technology (IoT). As a result, computers are becoming internet users, much like people with web applications. The Internet of Things is drawing recent researchers' exposure to one of the most exciting advantages and limitations. For the planning purposes of technology, network and communication infrastructure, this has an imperative economic and social impact [27].

Ultimately, this new regulation of its coming years will indeed be linked but instead constructively monitored. Because of the advancement of digital devices, encapsulated and omnipresent multimedia applications, cloud services and business intelligence, the iot concept becomes more and more relevant to both the believable way. In addition, IoT poses problems in quantity, speed and range combinations.

Just as the Internet has Allowed devices to take place in a plethora of parts of the world in a wider context, and promotes numerous applications from insignificant to vital. Contrary, understanding the IoT well, which include meanings, quality and discrepancies from several other related constructs, still is mystifying [28].

In terms of improving software development and information extraction of vast automation and control systems, many other varied systems including quantum computing but also cloud computing can also be implemented around each other. A lot of research has been conducted out in such a way by Mishra, Lin and Chang [27].

An development of information through Data sources is the toughest achievement which big data professionals face. The creation of resources for the analysis of IoT data is indeed important. Utilizing data mining techniques, an IoT system produces continuously network traffic and researchers may create instruments to extract useful knowledge from all of these data. Comprehending certain network traffic produced through Embedded applications and analyzing them to collect accurate data is a difficult problem and leading to optimization of big data.

Only one approach for managing large IOT devices is machine learning methods and computer intelligence techniques. e. Key technologies that are associated with IoT are also discussed in many research papers [28]. Figure 2 depicts an overview of IoT big data and knowledge discovery process.

ISSN No:-2456-2165

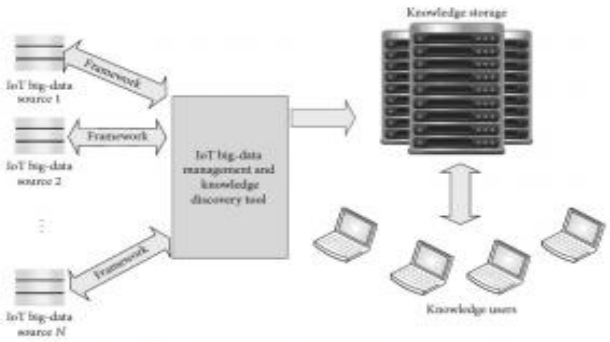

Fig. 2: IoT Big Data Knowledge Discovery

The method of information retrieval emerged from principles of transmitting individual content, including frames, laws, tagging, and neural web. It comprises of four divisions in specific, along with information exchange, knowledge base diffusion, and application of knowledge. Learning is explored in the process of knowledge discovery by the use of different conventional and computer intelligence techniques. Information found is preserved in data warehouses and expertise programs are typically built armed with the knowledge acquired.

To gain important data from either the base of knowledge, knowledge dissemination is essential. Production of knowledge is a method that explores for records. In diverse uses, the ultimate reason for applying realized information. It is the primary purpose of exploring information. The method of knowledge discovery is inherently flexible with knowledge acquisition judgment. In this field of information discovery, there are several topics, debates, and studies. It would be beyond this survey paper's reach. The information discovery method is represented in Figure 3 for visual analysis.

Fig. 3: IoT Knowledge Exploration System

B. Cloud Computing for Big Data Analytics

Supercomputing has been allowed more available and feasible by the advancement of cloud applications. In virtual machines, computing infrastructural facilities that become secret allow computers function like either a real machine, however with the versatility of requirement specifics also including number of computers, virtual memory, ram, and user interface. The use of certain virtual machines is defined as virtualization, amongst the most strategy revolves for big data [29] [30]. With the goal of creating a flexible and then on resource availability and computing, usage of cloud computing technologies are being developed. Cloud computing harmonizes vast data by offering application accessibility on requested.

The advantages of using cloud services including supplying services if there's a need and only paying for the assets add value to the product. It also increases availability and cost savings at the same time. Multiple continue to highlight open problems and conceptual framework in cloud technology in depth, enjoying the convenience in managing data, data diversity and speed, cloud computing, signal processing, and risk mitigation. So, public cloud helps develop the business strategy with technology and tools for all types of applications.

Data processing and production can be supported by big data applications using cloud services. The community cloud can provide features that allow software developers and business analysts to investigate knowledge extraction transmission technologies and collaboratively for some further evaluation and generating fruitful results. It could work to alleviate broad uses in numerous domains that really can arise. In comparison, information technology must also allow the deployment of virtual technology resources into technological developments such as spark, R, and other types of techniques for handling big data.

Data mining establishes a context to explore other options for cloud computing. Users can go to the platform and purchase critical infrastructure including cloud provider such as Google, Amazon, IBM, Software as a Service (SaaS) from the whole team of companies such as NetSuite, Cloud9, Jobscience, etc., based on special needs. Cloud storage, which provides a potential ways of processing big data, would be another benefit of cloud computing. The noticeable the first is the duration and expense necessary in the cloud environment for the data transfer of big data. Otherwise, the propagation of computing and the installed applications is hard to observe.

C. Bio-inspired Computing for Big Data Analytics

Nanoparticle programming is a methodology that is influenced by evolution to handle difficult real - world problems. Without a central command, natural mechanisms are self-organized. Check for a micro revenue management strategy and find the best data service approach for considering data management and performance maintenance costs. Biomolecules including Dna bases are developing these technologies to perform quantitative economic disputes data storage, retrieval, and processing. An essential aspect of these kind of processing was whether complex biological structures are incorporated in able to develop numerical functions and obtain smart results.

After computerization, huge quantities of data have been created from such a range of options throughout the network. A variety of intelligent analytics including data scientists and

ISSN No:-2456-2165

big data practitioners would be needed to analyze certain content and categorize them into text, image and video, etc. Proliferations of innovations such as big data, IoT [31] [32], cloud services, bio-inspired computing, etc. are evolving, although data balancing can only really be accomplished by choosing the best medium to examine broad and provide cost-effective performance.

In sophisticated data gathering and its connection to cloud computing, nanoparticle computation technique serve as an essential role. Thanks to it's own automation framework, certain techniques aid in undertaking data collection for large amounts of data. The greatest benefit is its elegance and its quick consolidation to the accurate value when solving a problem with service provision. Towards this end, Cheng explored certain developments using bio-inspired computing in detail. We may note from discussions that bio-inspired computing models have intelligent interactions, unavoidable failures of results, and help handle ambiguities. Therefore, bio-inspired computation is believed to better handle big data as possible.

D.Quantum Computing for Big Data Analysis

A hypothetical machine must have an infinitely larger capacity than some of its sheer structure and can similarly modify an expanding range of inputs. It may be important to execute this accelerated growth in computer systems. If a realistic quantum computer was already functional, it may well have produced results that seem to be exceedingly hard on contemporary computers, the big data concerns of nowadays, of particular. The biggest technological challenge in the design of quantum computers may soon be feasible. Quantum computing offers the opportunity for quantum theory to be combined to process data. Knowledge is provided by large blocks of bits in conventional computers that transmit whether its a zero or a single one [33].

The distinction here between qubit and a bit is that a qubit is a quantum device encoding the zero because the one across two distinct quantum fluctuations. The phenomenon of superposition can, hence, be capitalised on. It's because qubits include quantum behaviour. In quantum systems, for instance, 100 qubits require 2100 measurement results to be preserved in a traditional computing device. This means which, compared to conventional computing, several big data problems could be solved much more easily by bigger computing power. Hence, building a super computer and facilitating quantum computing to answer big data problems is now a problem for all of this age.

IV. TOOLS FOR BIG DATA PROCESSING

In order to accept predictive analytics, massive numbers of resources is very important. In this chapter, we address some current methods for analysing predictive analytics, focusing on three largest emerging instruments, namely MapReduce, Apache Spark, and Storm. Many of the resources provided focus on stream aggregation, filtering of streams, and collaborative analysis. The Apache Hadoop infrastructure, including Mahout and Dryad, is based on most batch processing tools. Stream data apps are mainly used for analytics in clear text. Strom and Splunk were other examples of large-scale network streaming. Toward their own analysis, the collaborative analysis method enables users to communicate continuously in real time.

The big data systems that support collaborative analysis are Dremel and Apache Drill, for instance. In the creation of big data initiatives, these tools support us. Many researchers are now discussing a fabulous list of big data methods and techniques [34]. In this section, the typical work flow of the big data project addressed by Huang et al is highlighted and shown in Figure 4 [35].

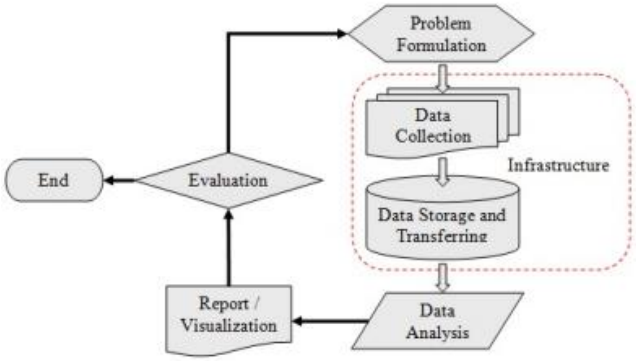

Fig. 4: Workflow of Big Data Project

A. Apache Hadoop and MapReduce

Data Analytics and Mapreduce are perhaps the most specific software frameworks also for advanced analytics. This consisting of kernel hadoop, mapreduce, distributed processing hadoop (HDFS) and apache hive, etc. Map Reduction is now a computing figure based on the divide - and - conquer approach for computing huge data. In 2 phases, such as Map Step and Reduce Step, the media manipulation method is presented. Hadoop runs on two sensor nodes, namely the master node and also the worker node. Within map stage, the central server splits the feedback into smaller community and afterwards distributes them to multiple processors. The central server then incorporates the contributions for all of the reduction stage problem instances. Moreover, Hadoop and MapReduce works as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

B. Apache Mahout

Apache Mahout seeks to provide massive and advanced analytics systems with flexible and operational data mining algorithms. Mahout's main optimization, namely cluster analysis, classification, pattern mining, regression, feature extraction, genetic algorithm, and distributed sample filtering, run on top of the Hadoop platform via the map reduced system. Mahout's aim is to develop a vibrant, open, diverse population in order to promote project discussion and possible use instances. The fundamental aim of Apache mahout would provide a method for solving serious hurdles [36].

ISSN No:-2456-2165

C. *Apache Spark*

Apache spark is an open source system also for analysis of big data designed for efficiency and technical analysis. It is simple and use and was actually designed at UC Berkeley's AMPLab in 2009. This was open-sourced as either an Apache project in 2010. Spark helps anyone to write deploy applications, Scala, or Python easily. It enables Database structure, semi structured, predictive analytics, and graph data processing, in comparison to chart reduction processes. That provide improved and external features, Spark runs on top of the existing Hadoop distributed file system (HDFS) infrastructure. Spark contains elements such as the operating system, the cluster manager and the worker nodes.

The leader software provides a basis also for spark cluster execution of programs. Within shape of activities, the cluster property acquired the services and also the server nodes to be doing the data management. Each system will now have a range of methods that really are able to execute the tasks, called executors. The key benefit has been that it provides help for the deployment of spark applications in current hadoop clusters. The Apache Spark architecture diagram is represented in figure 5. Apache Spark's various features are described below:

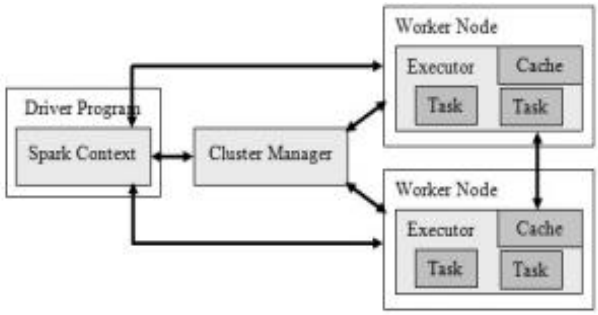

Fig. 5: Architecture of Apache Spark

D. *Dryad*

That would be another common programming paradigm also for implementation of parallel processing programmes on the set of instructions to manage broad context bases. It contains a collection of processing cores, and a user uses the network cluster's resources to run their programme in a distributed system. A dryad consumer actually uses thousands of computers, along with several processors or components. The key benefit is that there is no must for members to read anything whatsoever about parallel processing. A dryad system starts a directed computational graph consisting of computational vertices and channels of communication. Therefore, dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invokinguser defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices [37].

E. *Apache Drill*

Some other distributed framework for collaborative advanced analytics is Apache Drill. Supporting numerous varieties of distributed systems, data formats, and data sources, and has more versatility. Also it is made primarily to penetrate nested data. Thus it aims to scale up to 10,000 servers or more and has the capacity to process data bytes and trillions of records in seconds. To conduct batch analysis, drills use HDFS for storage and map reduction.

F. *Jaspersoft*

Jaspersoft is an open source package which generates notifications with database tables. It is a versatile research framework for big data and has the ability to easily visualise data on specific storage systems, including MangoDB, Cassandra, Redis, etc. One interesting component of Jaspersoft is that through selection, transformation, and loading, these can easily explore big data (ETL). In additional to this, it also has the capability to collaboratively and instantly generate powerful hypertext markup language (HTML) reports and integrations without ETL specifications from the big data store. It is way to access these produced reports with someone within or out of the user organisation.

H. Splunk

In latest days, a lot of stuff has been developed from economic sectors via machines. Splunk is a smart, real-time application designed to manipulate big data-generated machines. This incorporates the up-to-the-moment cloud and big data technology. In addition, it allows users to scan, monitor, and analyse their computer-generated data through a web interface. The findings, such as graphs, papers, and warnings, are seen in an intuitive way. Splunk is distinct from other methods for manipulating streams. Filtering hierarchical, disorganised digitally altered data, legitimate browsing, displaying numerical solutions, and dashboards are its peculiarities.

V. CONCLUSION

Info has been produced in latest days at a tremendous rate. It is difficult for a general individual to evaluate such data. In either study, we research the different research problems, obstacles, and tools used to analyse these big data towards this end. It is known from that kind of survey that every database has an individual emphasis. Some are optimized for data entry, while some are good for analytics in real time. There is also specific functionality for every database system. Analysis of the data, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream filtering are different methods used mostly for observation.

ISSN No:-2456-2165

REFERENCES

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

[2]. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.

[3]. C. Lynch, Big data: How do your data grow?, Nature, 455 (2008), pp.28-29.

[4]. 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.

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

[6]. 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.

[7]. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

[8]. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.

[9]. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[10]. R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.

[11]. Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

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

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

[14]. L. A. Zadeh, Fuzzy sets, Information and Control, 8 (1965), pp.338- 353.

[15]. Z. Pawlak, Rough sets, International Journal of Computer Information Science, 11 (1982), pp.341-356.

[16]. D. Molodtsov, Soft set theory first results, Computers and Mathematics with Aplications, 37(4/5) (1999), pp.19-31.

[17]. J. F.Peters, Near sets. General theory about nearness of objects, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

[18]. R. Wille, Formal concept analysis as mathematical theory of concept and concept hierarchies, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.

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

[20]. O. Y. Al-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.

[21]. Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, International Neurourology Journal, 18 (2014), pp.50-57.

[22]. P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.

[23]. A. Jacobs, The pathologies of big data, Communications of the ACM, 52(8) (2009), pp.36-44.

[24]. 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.

[25]. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congresso da sociedada Brasileira de Computacao, 2014, pp.1-6.

[26]. I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives, BioMed Research International, 2014, (2014), pp.1-13.

[27]. 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

[28]. X. Y.Chen and Z. G.Jin, Research on key technology and applications for internet of things, Physics Procedia, 33, (2012), pp. 561-566.

[29]. M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.

[30]. I. A. T. Hashem, 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.

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

[32]. C. Shi, Y. Shi, Q. Qin and R. Bai Swarm intelligence in big data analytics, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, 2013, pp.417-426.

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

[34]. M. Herland, T. M. Khoshgoftaar and R. Wald, A review of data mining using big data in health informatics, Journal of Big Data, 1(2) (2014), pp. 1-35.

ISSN No:-2456-2165

[35]. T. Huang, L. Lan, X. Fang, P. An, J. Min and F. WangPromises and challenges of big data computing in health sciences, Big Data Research, 2(1) (2015), pp. 2-11
[36]. G. Ingersoll, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White Paper, IBM Developer Works, (2009), pp. 1-18.
[37]. H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 2012, pp.675-683.
[38]. 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.