A Survey on Tools and challenges in Big data analytics

Dr. P. Prabhusundhar[1], A. Roselin[2]
[1]*M.C.A., Ph.D., Assistant professor, Gobi Arts & Science College*
[2]*Research scholar, Gobi Arts & Science College*

Abstract – In our everyday life huge amount of data is generated from modern information systems and digital technologies. It's a non separable part in our daily life. A massive amount of data is generated in this digital age. But traditional data analysis may not able to handle to this The extraction of knowledge for decision-making from these enormous data sets necessitates a lot of effort at various levels. The primary goal of this study to explore the impact of big data challenges, research issues and related tools. As a result, this article can be used as a starting point for looking at big data at different phases.

Keywords: Big data, Hadoop, Massive Data.

I.INTRODUCTION

Data is generated from a variety of sources in the digital world, and the rapid growth of digital technology has resulted in the creation of big data. It enables for evolving breakthroughs in a broad variety of fields. Naturally Big Data is classified into structured, semi-structured or Unstructured. Traditional data management, warehousing, and analysis solutions are incapable of analyzing the massive amounts of data due to the difficult scenario. Distributed architecture file systems are used to store Big Data.

A. Big data characteristics

Big data is a collection of data from various sources often its characterized by what becomes known as the V's.[1]

Volume: Volume is the huge amount of data generated every second in social media, cell phones, cars, images, video and whatnot.

Variety: Variety means different kinds of structured, semi structured and unstructured data.

Velocity: Velocity refers the speed of data generated distributed and collected.

Veracity: In terms of accuracy, the data's credibility.

Value: It's actually stores the valuable, relevant and trusted data need to be saved, processed and analyzed.

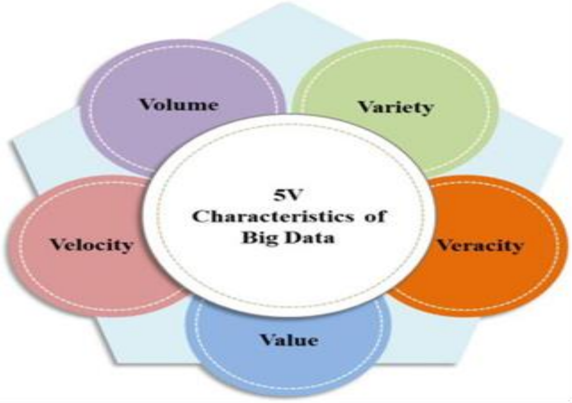

Figure [1]: Characteristics of Big Data

II.DATA ANALYTICS AND LIFE CYCLE

Big data analytics is a technique for analyzing and interpreting digital data and information. The functional features of digital products and services are determined by technological and analytical advancements in big data analytics (BDA). With the fastest and growing data generations, it's critical to streamline and understanding the method and mechanism by which big data analytics may provide value to industries in a variety of ways.

Data life cycle management is extremely beneficial to any company or application that uses and processes data to generate results. Data is created from a variety of sources and is accessible in a variety of formats [2]. It contains number of phases.

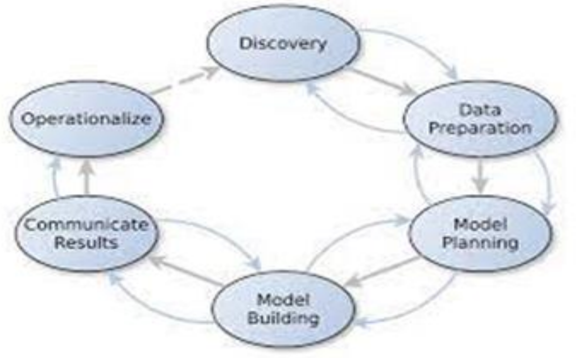

Figure [2]: Data Analytics life cycle

Phase 1:
Discovery: This is the first phase of the data analytics life cycle, and it outlines the data's purpose and how to accomplish it. To begin, establish all of the essential objectives and have a thorough understanding of the business area.

Gathering resources by examining the models that will be produced and evaluating the data sources that will be require.

The most important activities in this phase will be framing the business problem, generating an initial hypothesis that can be tested later using data, and starting data learning.

Phase 2:
Data Preparation: Prior to modeling and analysis, this step will encompass gathering, processing, and filtering the data. Assuring data availability for processing is one of the most important factors.

Various data sources are identified, and the amount of data that can be acquired in a given time frame is calculated. In this phase, the following data collection methods are used:

Data acquisition: gathering information from outside sources.

Data Entry: Manually enter data points or use computerized systems.

Signal reception: Data collection from digital devices, such as IoT devices and control systems, is referred to as signal reception.

Phase 3:
Model Planning: During this stage, the group decides on the tactics, processes, and work process it will use for the subsequent model structure stage. The group examines the data to determine the relationships between variables and, as a result, selects crucial factors and the most logical models.

Phase 4:
Model Building: The team has developed datasets for testing, preparation, and construction. Furthermore, based on the work done in the model planning stage, the group constructs and executes models in this step. The team also considers if its present equipment would suffice for running the models, or whether a more powerful environment will be required for executing models and work processes.

Phase 5:
Communicate Result: This stage determines if the outcomes were successful or not. The results of the data analysis are analyzed, and suggestions are made as to how to present the findings and conclusions to various team members and stakeholders, taking into account the risks and assumptions. Key findings will be discovered, business value will be validated, and a narrative will be developed to summarize and convey the findings to stakeholders. Make recommendation for the future work or enhancements to current processes as well. The model's impact on stakeholders' processes must be understood.

Phase 6:
Operationalize: In the final phase, the team will present the stakeholders with the full in-depth report, including briefings, coding, important results, and all technical documents and papers.

This method allows you to learn about the model's performance and restrictions in a live setting on a small scale and make the necessary improvements before deploying it. The outcomes are regularly checked to ensure that they are in line with the predicted outcomes. The report can be completed if the findings are precisely aligned with the goal. The model can then be implemented and integrated into the company.[3]

III.BIG DATA PROCESSING TOOLS

There are many different technologies available to process big data. In this section, we go over some of the current methods for analyzing big data with a focus on Map Reduce, Apache Spark, and Storm, three crucial new tools. Batch processing, stream processing, and interactive analysis are the main foci of the majority of the technologies that are now accessible. The most of batch processing technologies, like Mahout and Dryad, are built on the Apache Hadoop architecture. Real-time analytics is where stream data applications are most frequently deployed. Storm and Splunk are two examples of large-scale streaming platforms. Users can immediately interact in real time during the interactive analysis process to do their own analysis.[7]

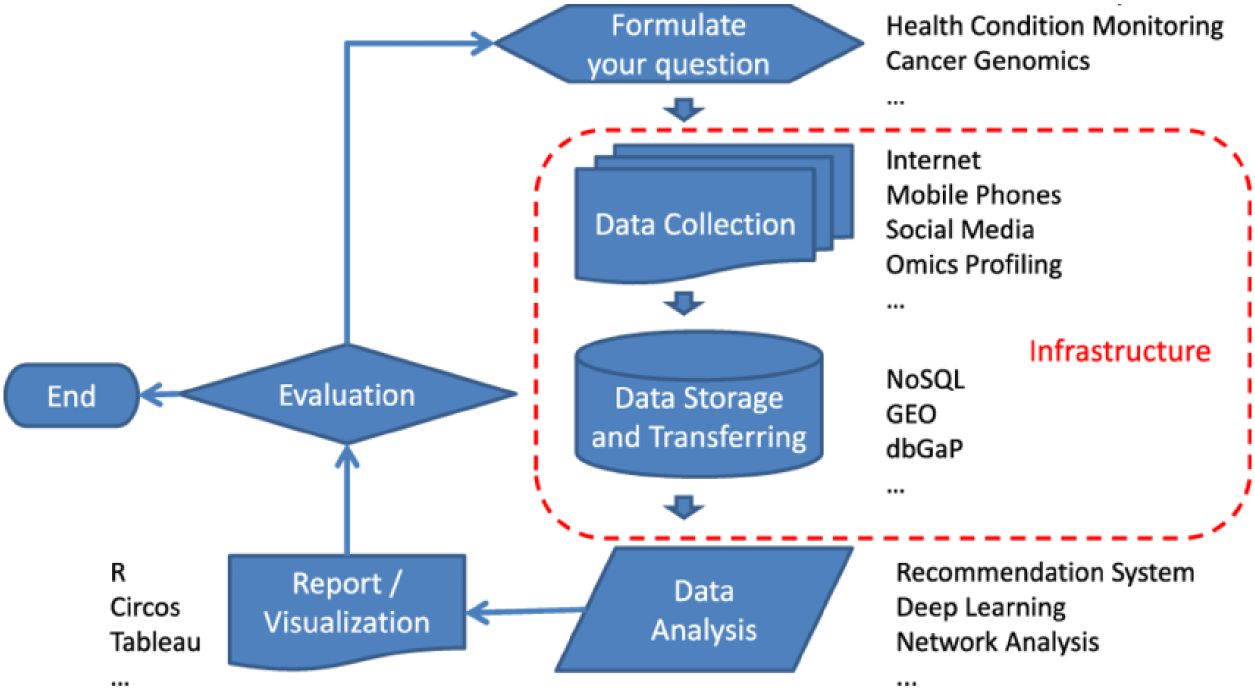

Figure [3]: Workflow diagram for Big data

(i) Apache Hadoop and MapReduce

Apache Hadoop and MapReduce are the most well-established software platforms for big data analysis. Java-based MapReduce is a processing method and a model for a distributed computing program. Map and Reduce are two vital functions that make it up the MapReduce algorithm. A set of data is transformed into another set through a map, where each element is separated into tuples (key/value pairs). The second action is a reduce task, which takes a map's output as input and concatenates the data tuples into a smaller collection of tuples. The reduction work is often dispenced following the map job because the name MapReduce implies.

The main benefit of MapReduce is that processing of data can be scaled easily over the several computing nodes. The data processing primitives used in the MapReduce model are referred to as mappers and reducers. Typically it’s tough to divide a data processing application into mappers and reducers. However, scaling an application to run over hundreds, thousands, or even tens of thousands of servers in a cluster is just a configuration modification after it has been written in the MapReduce manner. The MapReduce approach has gained popularity among programmers because of its simple scalability.

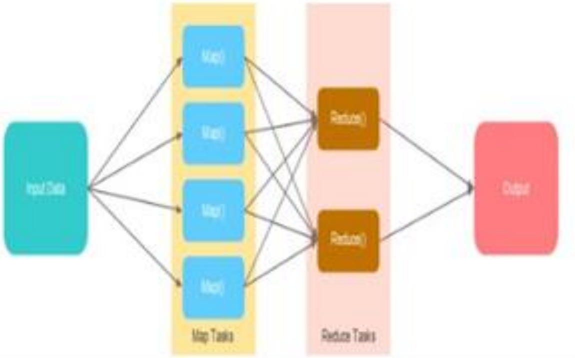

Figure [4]: Hadoop MapReduce

(ii) Apache Spark

An open-source large data processing framework called Apache Spark was created for quick processing and smart analytics. It was created in 2009 in UC Berkeley's AMPLab and is simple to use. As an Apache project, it was released as open source in 2010. Java, Scala, or Python applications can be swiftly written with Spark. It enables SQL queries, streaming data, machine learning, and graph data processing in addition to map-reduce operations. Spark provides enhanced and additional capability on top of the HDFS architecture already existing in Hadoop. Spark has a well-defined layered architecture every components and layers are loosely

Coupled. This architecture implemented with several extensions and Libraries. Apache spark architecture mainly classified into two main components.

1. Resilient Distributed Dataset (RDD)
2. Directed Acyclic Graph (DAG)

- The main focus of spark is Resilient Distributed Datasets(RDD) ,It stores data in memory and gives fault tolerance without replication. It improves speed, resource utilization. Directed Acyclic Graph is a finite graph that performs a sequence of Computations on data.
- Applications run in Hadoop cluster, up to 100 times faster in memory and 10 times faster in disk. It is possible because less amount of read and write operation done in disk.
- Spark is Written in scala Programming Language and runs in Java Virtual Machine (JVM) environment also it's supports Java, Python and R for developing applications using Spark.

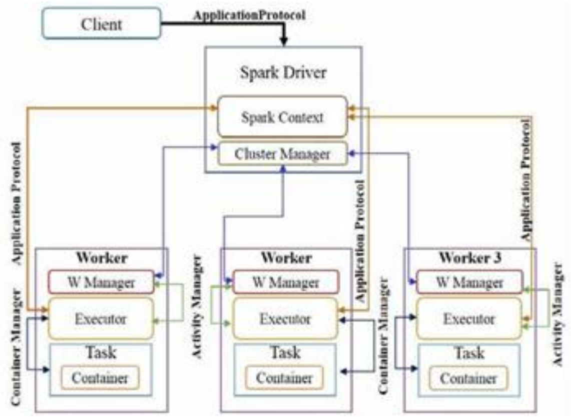

Figure [5]: Apache Spark Architecture

(iii) Storm

Apache storm main highlight is a fault tolerant, Fast with no Single Point of Failure (SPOF) distributed application. We can install it in many systems as needed to increase the capacity of the application. Apache storm internal architecture divided into two kinds of nodes, Nimbus and Supervisor. Nimbus role is Master node is the central component of apache Storm topology. it's analyzes the topology and gathers the task to be executed after that it will distributed the task to an available supervisor. A Supervisor role is a worker node. It has a one or more worker processes and it governs worker processes to complete the tasks assigned by the Nimbus. Work process is the executed the specific task. A Worker processor not executes a task itself. It has multiple executors. An Executor is a single thread originate by a worker process. A task performs the actual data it's either spout or a bolt. Zookeeper maintains the shared data with robust synchronization techniques. It is the responsible for maintaining the state of nimbus and supervisor. Storm is stateless in nature. It helps Storm to process real time data in best possible and fastest way. It is also having advanced topology called Trident Topology with state Maintains and High-level API like Pig. Apache storm uses own distributed messaging system for communication internally between nimbus and supervisor.

(iv) Apache Drill

Apache Drill is a low latency Distributed query System for interactive analysis of big data. It has particularly designed for nested data. This has more flexibility to support multiple types of query languages, data formats and data sources. It has able to scale up to 10000 servers or more and reaches the capability to process petabytes of data and trillions of records in seconds. Apache Drill uses the storage for HDFS, and map reduce for batch analysis. Drill extremely user friendly and developer friendly.

Challenges in Big data Analytics

Now a days we find big data in marketing Media and Entertainment, Health care, Government Sector, Biochemistry, Transportation, Weather patterns and Education.

Social computing and internet search indexing are some are the web-based usage of big data. We can have the help of internet search indexing like ISI, Scopus, IEEE, Reuters to do prediction of markets and social network analysis. These are some of the areas where extensive research is a must of course opportunities and challenges are inseparable. [17]

Analyzing big data in the higher level is a challenge. The methods of analysis of data such as statistical methods and computation techniques are suitable with the small data size and not for big data.

- Data storage and Analysis
- Information Security
- Discovery of knowledge and Computational complexity
- Scalability and Visualization of data are some of the challenges we take in big data analytics.

A) Data Storage and Analysis

Now a days we live more interconnected world that generates a huge amount of data by various sectors such as mobile devices, aerial sensory technologies, remote sensing, radio frequency, identification readers etc. These kinds of data spending much cost to storing them whereas they ignored or deleted finally because there is no space to store so the first challenge for big data analysis is storage medium and higher input and output speed. In that case data accessibility should be the top priority. In Past years we are storing the data in hard disk it slower the input and output performance. Solid state drive(SSD) and Phrase change memory(PCM) was introduced overcome these kinds of problems. Another challenge with big data analysis is attributed to diversity of data with the growing datasets. Data reduction, data selection, feature selection is a major task when dealing the large datasets. Hadoop and map Reduce technologies make it possible collect huge amount of structured and unstructured data in a desired time. A key process is to analyze the data in effective manner. Semi Structured and unstructured data into structured data is the standard process. [36]

The major task is to develop storage systems and to elevate productive data analysis tool that ensure guarantees on the output when we get data from different sources. To improve and scalability design of machine learning algorithm for big data analysis needs serious attention.

Figure [6]: Big Data Sources

B) Information Security

Meaningful trends are mixed by correlating and analyzing massive amount of data. Every organization has sensitive information, Information security is a major challenge in big data analysis. Authentication, Authorization and encryption are useful for upgrading security of big data. However lack of intrusion system, real time security monitoring and scale of network multiplicity of devices are some of the security measures in big data applications.

Despite extensive work done to secure big data, but it demands lot of upgrading development of multilevel security, where privacy in a prime concern is a nut to crunch. [4]

C) Discovery of knowledge and Computational complexity

The list of challenges is without incomplete without knowledge discovery and representation. Fuzzy set, Rough set, Soft set, near set, Formal set concept analysis of concept analysis are some of the tools for knowledge discovery and representation. Real life problems are processed using hybridized techniques. These techniques are not only for problem dependent but also not suitable in large data set in a sequential computer. As the size of the big data is exponentially large efficient use of available tools to process these data for meaningful information is a question mark. Data warehouse and data marts are popular approaches in this context. Data sourced from operational systems are stored in data warehouse but data marts facility analysis.

Inconsistence and uncertainty found in the dataset demand more computational complexities here use of systematic modeling of the computational complexity is not worthy. The comprehensive mathematical system may be establishing difficult but not impossible.

By comprehending the particular complexities, a domain related analytics can be done. Machine learning techniques are least memory requirements user extensive research and survey in this direction. Attempts are made for cost reduction in computational complexities performance of current big data analysis tools in dealing with computational complexities, uncertainty and inconsistencies are not encouraging. It is hue we have a big task to develop techniques and technologies to handle computational complexities with remarkable performance.

D) Scalability and visualization of data

Big data analysis technologies suffer from scalability and security. In the past moore's law helped to accurate data analysis to some extent increment techniques come to handy to solve scalability problem in big data analytics. Since there is a mismatch between data size scaling and CPU speeds, processor technology is extended with intensive number of cores.

This latest development in processor leads to a new domain called parallel computing finds application in navigation, finance, internet scratch and social networks. There comes visualization of data, when we present data more adequate using some techniques of graph theory.

Graphical visualization is to support data with proper interpretation. E-commerce companies like Alibaba, Amazon, Flip cart, Big basket and E-bay generate a lot of data because they have millions of users and billions of goods. A tool called tableau used here for big data visualization. Tableau does a job to transform large and complex data into intuitive pictures. It is not up to the mark to meet the latest expectation in functionalities response in time.

Big data are completing us for designing of hardware and software this hardware and software will take the que to cloud computing, parallel computing visualization process and scalability. Here comes the correlation of mathematical models to computer science.

IV.CONCLUSION

Big data provides more choices because of related technologies and tools. Now a days vast amount of data generated in years. Analyzing these data is very challenging task to us. End of this paper , we survey the various tools ,life cycle and different kinds of challenges used to analyze these big data. From this survey, it’s understood every big data has individual focus. Some of the big data tools designed for batch processing and some of them for real- time analytic. Every big data platform for developed for unique functionalities. Various kinds of analytics method used this like machine learning, data mining, intelligent analysis, cloud computing, data stream processing. Hopefully it has provide few useful discussion for researchers.

REFERENCE

[1] 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.

[2] Kumar Rahul and Rohitash Kumar Banyal: Data Life Cycle Management in Big Data Analytics Procedia Computer Science 173 (2020) 364–371.

[3] Manisha R Gupta: Data Modeling and Data Analytics Lifecycle International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 5, Issue 2, May 2021.

[4] Maddhi Sunitha: A SURVEY ON BIG DATA ANALYTICS: CHALLENGES, RESEARCH ISSUES AND PLATFORMS International journal of Advanced Technology in Engineering and Science Volume 5, Issue 3, May 2017.

[5] 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.

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

[7] Katal, A.; Wazid, M. ; Goudar, R.H “Big data: Issues, challenges, tools and Good practices, IEEE “Contemporary Computing (IC3), 2013 Sixth International Conference,pp:404 – 409.

[8] 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.

[9] 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.

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

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

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

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

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

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] P.Prabhusundhar, "*Optimized Regression Neural Network for Classification of Diabetes in Big Data Environment*", Natural Volatiles & Essential Oils (NVEO) (SCOPUS Journal), ISSN NO. 2148-9637, Vol. 8, No. 4, DEC - 2021, pp. 14105 - 14123.

[22] P.Prabhusundhar, "*An Adaptive Graph Based Feature Selection and Deep Learning Classification Framework for Rice Disease Prediction*", International Journal of Aquatic Science (IJAS), ISSN NO. 2008-8019, Vol. 12, No. 3, DEC - 2021, pp. 3058 – 3067.

[23] P.Prabhusundhar, "*Evolution of Big Data Analytics in Cloud Computing Environment*", International Journal of Research in Engineering and Science (IJRES), ISSN NO. 2320-9356, Vol. 10, No. 3, MARCH - 2022, pp. 62 – 67.

[24] P.Prabhusundhar, "*Ensemble Machine Learning and Hybrid Neutrosophic Cognitive Maps Based Feature Selection for Rice Disease Prediction*", International Journal of Mechanical Engineering (SCOPUS Journal), ISSN NO. 0974 - 5823, Vol. 7, No. 4, APRIL - 2022, pp. 671 – 686.

[25] P.Prabhusundhar, "*Real Time Model for Human Faces Recognition Biometrics Tracking from Surveillance Videos*", International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), ISSN NO. 2582-5208, Vol. 4, No. 7, JULY - 2022, pp. 82 – 89.

[26] 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.

[27] 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.

[28] 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

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

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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. *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016*