International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

Various Platforms and Machine Learning Techniques for Big Data Analytics : A Technological Survey

Shahid Mohammad Ganie[1], Majid Bashir Malik[2], Tasleem Arif [3]
[1]Department of Computer Sciences, BGSB University, Rajouri, J&K, India
[2]Department of Computer Sciences, BGSB University, Rajouri, J&K, India
[3]Department of Information Technology, BGSB University, Rajouri, J&K, India

ABSTRACT

Data is growing drastically more and more every day and it becomes difficult task to store, analyse and interpret this data. Big data is a term that describe large volumes of high velocity, complex and variable data that cannot be stored and processed using traditional approach. Big data analytics require advanced tools and techniques in order to capture, storage, distribution, management, and analysis the data. Because of the complexity and heterogeneity of big data, various data mining and machine learning techniques are being used for big data analytics in order to develop better expert systems of real-world problems. In this paper, we have surveyed the state-of-art analysis of various platforms (software as well as hardware) for big data analytics like Hadoop ecosystem, Spark, High performance clusters (HPC), Graphical Processing Unit (GPU), etc., which are together used to collect, store, process and analyse the big data. This paper also reinforces some machine learning techniques that must be taken in account while dealing with big data lifecycle.

Keywords : Big data, Big data platforms, Hadoop, spark, HPC, GPU, Machine learning

I. INTRODUCTION

A. Big Data

Big data is defined as a collection of large and complex data sets that are generated from different sources including healthcare system, social networking sites, online transactions, sensors, smart meters, and administrative services [1]. With all these sources, the complexity of big data goes beyond the ability of typical tools for storing, analysing, and processing data[2], [3], [4], [5]. According to the Gartner IT Glossary [6] "Big data is a collection of high-volume, high-velocity and high-variety information assets that require advanced tools and techniques in order to store and process the data for building superior insight and decision making". Nowadays the problem of big data analytics is often solved through cloud computing architecture in parallel and distributed fashion [7], [8]. These frameworks are designed on the basis of various metrics like scalability, data I/O rate, fault tolerance, batch processing, real time processing so on and so forth [2]. This paper first describes the concept of big data analytics, then we deeply analyse different big data platforms and then in next section various machine learning techniques for big data analytics are described.

B. Motivation towards the study

CSEIT217242 | Received : 01 Aug 2018 | Accepted : 25 Aug 2018 | July-August-2018 [3 (6) : 679-687]

Growth of the data can be understood from the fact, Internet Users generate 2.5 quintillion bytes of data each day, 90% of all data today was created in the last two years, the world internet population has grown 7.5% from 2016 and now represents 3.7 billion people according to recent research cited by Domo [9]. 300+ hours of videos are being uploaded on YouTube every minute [10]. 40,000 search queries are being posted every second, 350000 tweets per minute and 171 million emails per minute [11]. The healthcare data is growing with the rate of 48% annually and by 2020, the Stanford University study estimates that 2,314 Exabyte's of data will be produced per year [12]. As represented in **Figure 1**, most of the data is being generated by different organisations [13]:

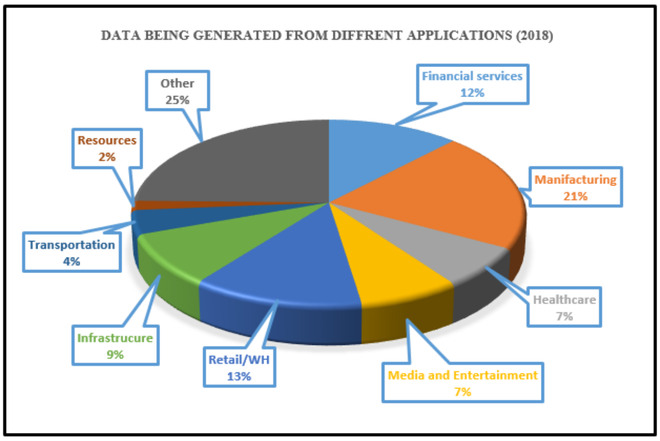

Figure 1. Generation of "Global Datasphere"

C. Characteristics of big data

Literature survey on big data in healthcare characteristics gives a brief description of **42 V's** [14], **17 V's** [15], **10 V's** [16], **7 V's** [17] so on and so forth, but it summarizes and divides the concepts into five dimensions as presented in **Table 1** [18], [19].

Table 1: The 5 V's of Big Data

V's of Big Data	Definition
Volume (Data in size)	It can be defined as the parameter that defines the quantity of the data being generated and stored from different sources.
Velocity (Data in speed)	The rate at which data is being generated and processed along with the speed at which data changes especially in case of streaming.
Variety (Data complexity)	It is type and nature of the data. Type refers to different data formats and generating sources, while nature defines different forms like structured, semi structured, and unstructured data.
Veracity (Data Quality)	It is all about quality of the data. The data that is submitted can either be incorrect, may have noise or even missing values. How can this data be confided for its truthfulness?
Value (Data usefulness)	It refers to identification and extraction of valuable information for analysis. Data incorporates information of great benefit and insight for users.

II. BIG DATA PLATFORMS

This fast growing and tremendous amount of data has far exceeded the bounds of the human interpretation and analytical capabilities [20]. The goal is to link computational power of the technology to automate the capability of ingestion, processing and visualization of results from huge volume of data for the betterment of the society [21]. Big data platforms can be used for the same. The big data platforms have been categorised on the basis of scaling [22]. Scaling is the ability of the system to adapt the increasing demands in terms of processing requests [23]. Scaling can be Horizontal Scaling and Vertical Scaling:

A. Horizontal scaling platforms

Horizontal scaling or "scale out" is distributed data processing that shares workload among multiple nodes.

This increases the processing capability up-to a great extent. The most popular horizontal scale out platforms are as follows [2]:

1. Peer-to-Peer networks [24], [25] It contains millions of nodes connected in a network. The working principle of peer-to-peer network is decentralized and distributed over network, where the nodes (known as peers) are utilizing the resources and provide services to the peers within the network. Typically, Message Passing Interface (MPI) [26] communication system is used to share and exchange the data between nodes.

2. Apache Hadoop [27] It is batch processing open source framework that allows to store and process large and complex data sets which cannot be handled by traditional approaches. Hadoop works on distributed environment across cluster of computers and is designed to scale up from the single servers to the hundreds and even thousands of machines. Apache hadoop framework is highly fault tolerant with the feature of replicating data [2]. The Hadoop ecosystem is comprised of array of related software's, that are collectively used for data collection, data storage, data processing and data analysis [28], [29].

2.1 Big Data Collection

- **Sqoop:** It is an open source framework comprised of SQL and Hadoop [29]. Sqoop provides a command line interface used to transfer (import and export) bulk data between HDFS and relational database servers.
- **Flume:** It is tool for data ingestion in HDFS, which collects, aggregates and transpose large amount of streaming data such as log files, social media, email, etc. [30]. It captures streaming data from various web servers to HDFS.
- **Kafka:** [31] Is an open source distributed publish-subscribe messaging system, It was originally developed at LinkedIn and later it becomes part of apache project and is now used by many of the organisations because of its scalability, durability and fault-tolerance.
- **Chukwa:** It is distributed data collection and rapid processing system, Chukwa is a powerful and flexible platform with different core components as Agents, Collectors, MapReduce Jobs and Hadoop Infrastructure Care Centre (HICC) [28].

2.2 Big Data Processing

- **MapReduce:** [32] Which was proposed by Dean and Ghemawat at Google. MapReduce takes care of Processing and computing data present in HDFS. The working principle of MapReduce is mainly divided into two parts, called as Mappers and Reducers [33]. The Map task takes input data from HDFS and converts into some intermediate results to the reducers. The reduce task is based on map task; it takes the output of Mappers as input and then generate the final results by aggregating the intermediate sets which are again reflected to HDFS [34].
- **YARN:** [35]Yet Another Resource Negotiator is a framework for job scheduling and cluster resource management i.e. the jobs across the cluster. It is global scheduler that handles both batch and stream processing [36]. It is also known as MapReduce version 2 and is compatible with the MapReduce, having master slave architecture with full support of virtual distributed system.
- **Storm:** It is an open source computing system used for real time processing of big data analytics [7]. Storm is distributed, reliable and fault-tolerant in nature, used for Extract Load and Transform (ETL) operations, online machine learning and continuous computation [37].
- **Flink:** It is data ingestion tool in HDFS [38], which came into picture for collecting, aggregating and transport large amount of streaming data. The main idea behind the flink is to capture large amount of data from varied sources into HDFS [39].
- **Spark:** [40] It is used for both batch and real time processing, it is a next generation concept for big

data processing which was developed by researchers at the University of California at Berkeley. Spark is an open source cluster computing framework particularly designed for the speed up (100 times more than the speed of hadoop MapReduce for data that resides in main memory and around 10 times faster when data is in disk) in terms of processing [41].

2.3 Big Data Storage

- **HDFS:** Hadoop Distributed File System (HDFS) [7] is a main component of hadoop ecosystem that is used to store large data files across various cluster of cost effective system (Commodity hardware) while providing high availability and fault tolerance [42].
- **HBase:** Apache HBase is NoSQL columnar database that allows us to store un-structured and semi-structured data easily and provide real-time read/write access [43].
- **Hcatalog:** It is a metadata and table management system for Hadoop, provides a shared schema and data type mechanism for various tools of Hadoop ecosystem [28].

2.4 Big Data Analysis

- **Pig:** [44] It is a high level data flow system developed by yahoo, used to write simple queries that are converted into MapReduce program and then executed over hadoop cluster. It has been used to overcome the complexity of Map and Reduce Stages; Pig helps to process bulk large data sets by spending less time in writing MapReduce programs.
- **Hive:** [45] It is tool for big data analytics built on the top of Hadoop, used to analyze the structured and semi-structured data. Basically, it provides an interface mechanism for performing different queries written in HQL (Hive Query Language) that are similar to SQL statements.
- **Mahout:** [7] Apache Mahout is meant for machine learning that runs on Hadoop ecosystem extracts meaningful information from large volume of data. It is used for various machine learning tasks like clustering, classification, collaborative filtering and text mining in a scalable and distributed fashion [46].

B. Vertical Scaling Platforms

Vertical Scaling or “scale up” is achieved by enhancing the processing and storage capability of a single server by increasing the number of processors or cores and enhancing the memory and other required resources. The most popular vertical scale up platforms are as follows [2]:

- **HPC clusters:** [47] The blades or supercomputers with a large number of processing cores have a dynamic memory organization, different levels of cache and communication mechanisms that are optimized for diverse user requirements. To achieve scalability of such systems is much costlier than Hadoop or Spark. A Message Passing Interface (MPI) is usually used for communication in such systems [48].
- **Graphics Processing Unit (GPUs):** They are used for accelerating the graphic operations for pictorial representation on display using frame buffer [2]. GPGPU (General-Purpose computing on Graphics Pro-cessing Units) is the result of advanced enhancements in hardware and algorithms in GPU like parallel architecture.
- **Field Programmable Gate Arrays (FPGA):** [49] Are specialized and custom built hardware units that can be optimized for high speed and throughput for a particular application. The customization and development cost is typically very high, as compared to other platforms also its feature of programming using Hardware Descriptive language (HDL) increases the overall development cost [2].

Big data platforms are comprised of bunch of tools which are used together by various organisations to handle data. The pictorial representation of Hadoop ecosystem is shown below in **Figure 2** [28].

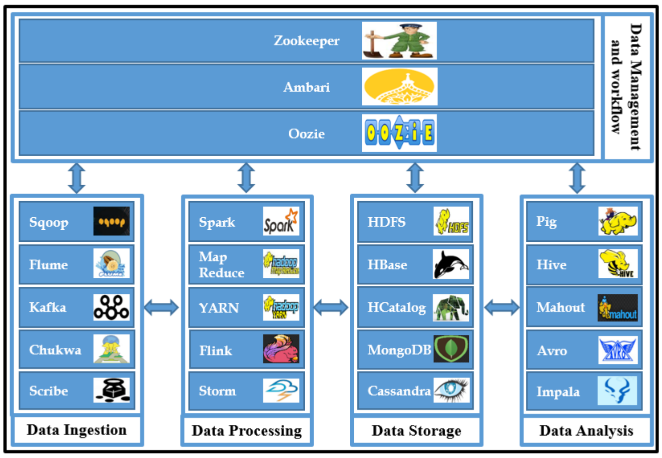

Figure 2 : Platforms for Big Data Lifecycle

III. MACHINE LEARNING TECHNIQUES FOR BIG DATA ANALYTICS

The era of machine learning has make a lot of progress in the technological field by providing a great potential of tools and techniques in order to handle the big data [50], [51]. Machine learning whenever implemented have resulted in better outputs in terms of various metrics, sometime it may even exceed human expertise. Machine learning techniques can be categorized on the bases of learning mechanism as shown in **Figure 3** [52].

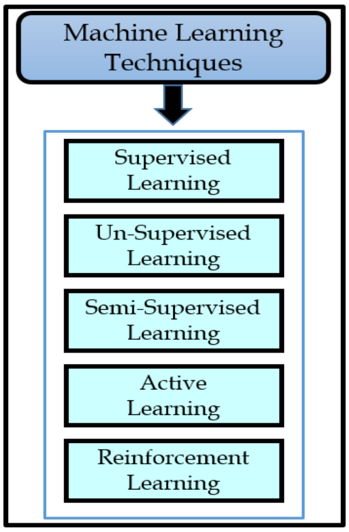

Figure 3 : Machine Learning Techniques.

Machine learning algorithms are being used successfully for big data analytics by various organizations in almost every sphere that affects human life. Some of the prominent techniques that have significant application in big data analytics are discussed below [50]:

A. **Neural Networks:** These are the algorithms which are inspired by the neurons have self-capability to produce better results on large volume of data [53]. The structure of neural network can be divided into three layers: input (disparate sources), hidden (accountable for internal processing) and output (desired results). Some of the networks are Multilayer Perceptron, Convolutional Neural Networks, Recurrent Neural Networks, etc.

Supervised Learning: Classification and regression algorithms are part of supervised learning technique. These algorithms are trained using labelled data along with targeted output for big data processing [54]. Some of the algorithms are Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest, Linear Regression, Logistic Regression, Polynomial regression, etc.[55].

D. **Un-supervised learning:** Clustering and association algorithms are types of un-supervised learning. These models does not require any labelled training, they are being used to find the previously unknown patterns in large datasets [56]. K-Means clustering, DBSCAN, Hierarchical Clustering are the examples of un-supervised learning technique.

E. **Reinforcement Learning:** One of the most important learning aspect, which enable the systems (model free or model based) to learn through the feedback received from external environment [50]. The learning mechanism is based on the trial-and-error method by using different parameters like agent, environment, action and state. Some of the algorithms are Q-learning, R-learning, TD-learning, etc.

IV. CONCLUSION AND FUTURE SCOPE

The massive amount of data that is being generated from multiple sources with rapid pace is found almost in every field especially in science and engineering domains. The overall process from data acquisition up to the generation of results required an attention to develop a sophisticated framework for big data analytics based on machine learning paradigms. This paper began with the brief introduction of big data, its classification and characteristics. Then, a brief introduction to various platforms for big data and big data analytics has been given, through which we can store, manage, analyse, filter and distribute data. We have also reviewed various machine learning techniques for big data analytics.

But still there are different types of issues which could be solved in future. Some of the issues that are related with the machine learning for big data analytics are learning for huge volume of data, data formats, streaming data, uncertainty and incomplete data so on and so forth. The privacy and security issues in big data are the important topics of research. Similarly, new machine learning tools and techniques should be developed so that important information can be extracted from the complex and heterogeneous unstructured data.

V. REFERENCES

[1]. Oracle, “Oracle: Big Data for the Enterprise Oracle White Paper—Big Data for the Enterprise,” An Oracle White Pap., no. June, 2013.

[2]. D. Singh and C. K. Reddy, “A survey on platforms for big data analytics,” J. Big Data, vol. 2, no. 1, pp. 1–20, 2015, doi: 10.1186/s40537-014-0008-6.

[3]. C. A. Technology, “Batch processing 11/28/2017 6,” pp. 1–4, 2018.

[4]. M. Docs, “Big data architectures 11/28/2017 10,” pp. 1–7, 2018.

[5]. B. C. Big, A. B. Big, and V. Machines, “Big compute architecture style 08/30/2018 3,” pp. 8–11, 2018.

[6]. A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[7]. S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, “A survey of open source tools for machine learning with big data in the Hadoop ecosystem,” J. Big Data, vol. 2, no. 1, pp. 1–36, 2015, doi: 10.1186/s40537-015-0032-1.

[8]. T. R. Rao, P. Mitra, R. Bhatt, and A. Goswami, The big data system, components, tools, and technologies: a survey, vol. 60, no. 3. Springer London, 2019.

[9]. “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.” Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#45381faf60ba. Accessed: 22-May-2019].

[10]. “300 Hours of Video are Uploaded to YouTube Every Minute.” Online]. Available: https://tubularinsights.com/youtube-300-hours/. Accessed: 19-Feb-2019].

[11]. “Google Search Statistics - Internet Live Stats.” Online]. Available: https://www.internetlivestats.com/google-search-statistics/. Accessed: 22-May-2019].

[12]. “Infographic: How Big Data Will Unlock the Potential of Healthcare.” Online]. Available: https://www.visualcapitalist.com/big-data-healthcare/. Accessed: 23-May-2019].

[13]. R. Saracco, “Another shift in content production ,” pp. 2019–2020, 2020.

[14]. T. Shafer, "The 42 V ' s of Big Data and Data Science," kdnuggets.com Elder Res., pp. 1–3, 2017.

[15]. V. K. A. -Arockia Panimalar. S, Varnekha Shree. S, "The 17 V's of Big Data," Int. Res. J. Eng. Technol., vol. 4, no. 9, pp. 3–6, 2017.

[16]. N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian, "The 10 Vs, Issues and Challenges of Big Data," pp. 52–56, 2018, doi: 10.1145/3206157.3206166.

[17]. M. A. U. D. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ. - "Engineering Educ. Ind. Involv. Interdiscip. Trends", ASEE Zo. 1 2014, 2014, doi: 10.1109/ASEEZone1.2014.6820689.

[18]. V. C. Storey and I. Y. Song, "Big data technologies and Management: What conceptual modeling can do," Data Knowl. Eng., vol. 108, no. February, pp. 50–67, 2017, doi: 10.1016/j.datak.2017.01.001.

[19]. H. Jasim Hadi, A. Hameed Shnain, S. Hadishaheed, and A. Haji Ahmad, "Big Data and Five V'S Characteristics," Int. J. Adv. Electron. Comput. Sci., no. 2, pp. 2393–2835, 2015.

[20]. S. Mazumdar, D. Seybold, K. Kritikos, and Y. Verginadis, A survey on data storage and placement methodologies for Cloud-Big Data ecosystem, vol. 6, no. 1. Springer International Publishing, 2019.

[21]. D. Blazquez and J. Domenech, "Big Data sources and methods for social and economic analyses," Technol. Forecast. Soc. Change, vol. 130, no. September 2017, pp. 99–113, 2018, doi: 10.1016/j.techfore.2017.07.027.

[22]. D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," J. Big Data, vol. 2, no. 1, pp. 1–20, 2015, doi: 10.1186/s40537-014-0008-6.

[23]. M. Merrouchi, M. Skittou, and T. Gadi, "Popular platforms for big data analytics: A survey," 2018 Int. Conf. Electron. Control. Optim. Comput. Sci. ICECOCS 2018, pp. 1–6, 2019, doi: 10.1109/ICECOCS.2018.8610652.

[24]. M. Irestig, N. Hallberg, H. Eriksson, and T. Timpka, "Peer-to-peer computing in health-promoting voluntary organizations: A system design analysis," J. Med. Syst., vol. 29, no. 5, pp. 425–440, 2005, doi: 10.1007/s10916-005-6100-x.

[25]. P. Kisembe and W. Jeberson, "Future of Peer-To-Peer Technology with the Rise of Cloud Computing," Int. J. Peer to Peer Networks, vol. 8, no. 2/3, pp. 45–54, 2017, doi: 10.5121/ijp2p.2017.8304.

[26]. O. Sievert and H. Casanova, "A simple MPI process swapping architecture for iterative applications," Int. J. High Perform. Comput. Appl., vol. 18, no. 3, pp. 341–352, 2004, doi: 10.1177/1094342004047430.

[27]. S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," J. Big Data, vol. 2, no. 1, pp. 1–36, 2015, doi: 10.1186/s40537-015-0032-1.

[28]. S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares, "BIG DATA for Healthcare: A Survey," IEEE Access, vol. 7, pp. 7397–7408, 2019, doi: 10.1109/ACCESS.2018.2889180.

[29]. S. Mehta and V. Mehta, "Hadoop Ecosystem: An Introduction," Int. J. Sci. Res., vol. 5, no. 6, pp. 557–562, 2016, doi: 10.21275/v5i6.nov164121.

[30]. V. S. N. Bhagavatula and S. S. Raju, "A SURVEY OF HADOOP ECOSYSTEM AS A HANDLER OF BIGDATA," no. August 2016, 2017.

[31]. B. Leang, S. Ean, G. A. Ryu, and K. H. Yoo, "Improvement of kafka streaming using partition and multi-threading in big data environment," Sensors (Switzerland), vol. 19, no. 1, 2019, doi: 10.3390/s19010134.

[32]. J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," OSDI 2004 - 6th Symp. Oper. Syst. Des. Implement., pp. 137–149, 2004, doi: 10.21276/ijre.2018.5.5.4.

[33]. P. Sun and Y. Wen, "Scalable Architectures for Big Data Analysis," Encycl. Big Data Technol., vol. c, pp. 1446–1454, 2019, doi: 10.1007/978-3-319-77525-8_281.

[34]. I. Kaur, N. Kaur, A. Ummat, J. Kaur, and N. Kaur, "Research Paper on Big Data and Hadoop," vol. 8491, no. 1, pp. 50–53, 2016.

[35]. B. J. Mathiya and V. L. Desai, "Apache Hadoop Yarn Parameter configuration Challenges and Optimization," Proc. IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS 2015, 2015, doi: 10.1109/ICSNS.2015.7292373.

[36]. Y. Perwej, B. Kerim, M. S. Adrees, and O. E. Sheta, "An Empirical Exploration of the Yarn in Big Data," Int. J. Appl. Inf. Syst., vol. 12, no. 9, pp. 19–29, 2017, doi: 10.5120/ijais2017451730.

[37]. S. Alkatheri, S. A. Abbas, and M. A. Siddiqui, "Big Data Frameworks: A Comparative Study," Int. J. Comput. Sci. Inf. Secur., vol. 17, no. 1, 2019.

[38]. D. Y. Perwej, M. Omer, and B. Kerim, "A Comprehend The Apache Flink In Big Data Environments," IOSR J. Comput. Eng. (IOSR-JCE), e-ISSN 2278-0661,p-ISSN 2278-8727,www.iosrjournals.org, vol. Volume 20, no. March, p. Page 48-58, 2018, doi: 10.9790/0661-2001044858.

[39]. T. Rabl, J. Traub, A. Katsifodimos, and V. Markl, "Apache Flink in current research," it - Inf. Technol., vol. 58, no. 4, pp. 2–9, 2016, doi: 10.1515/itit-2016-0005.

[40]. H. Benbrahim, H. Hachimi, and A. Amine, "Comparison between Hadoop and Spark," Proc. Int. Conf. Ind. Eng. Oper. Manag., vol. 2019, no. MAR, pp. 690–701, 2019.

[41]. N. M. Faseeh Qureshi et al., "Dynamic Container-based Resource Management Framework of Spark Ecosystem," Int. Conf. Adv. Commun. Technol. ICACT, vol. 2019-Febru, no. February, pp. 522–526, 2019, doi: 10.23919/ICACT.2019.8701970.

[42]. P. Basu, "HDFS for Big Data," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2013, doi: 10.1017/CBO9781107415324.004.

[43]. C. Jin and S. Ran, "The research for storage scheme based on Hadoop," Proc. 2015 IEEE Int. Conf. Comput. Commun. ICCC 2015, pp. 62–66, 2016, doi: 10.1109/CompComm.2015.7387541.

[44]. S. rna C and Z. Ansari, "Apache Pig - A Data Flow Framework Based on Hadoop Map Reduce," Int. J. Eng. Trends Technol., vol. 50, no. 5, pp. 271–275, 2017, doi: 10.14445/22315381/ijett-v50p244.

[45]. A. Fuad, A. Erwin, and H. P. Ipung, "Processing performance on Apache Pig, Apache Hive and MySQL cluster," Proc. 2014 Int. Conf. Information, Commun. Technol. Syst. ICTS 2014, pp. 297–301, 2014, doi: 10.1109/ICTS.2014.7010600.

[46]. V. R. Eluri, M. Ramesh, A. S. M. Al-Jabri, and M. Jane, "A comparative study of various clustering techniques on big data sets using Apache Mahout," 2016 3rd MEC Int. Conf. Big Data Smart City, ICBDSC 2016, pp. 374–377, 2016, doi: 10.1109/ICBDSC.2016.7460397.

[47]. D. Kumar, L. Ali, and S. Memon, "Design and Implementation of High Performance Computing (HPC) Cluster Design and Implementation of High Performance Computing (HPC) Cluster," no. January, 2018.

[48]. C. S. Yeo, R. Buyya, R. Eskicioglu, and P. Graham, "Handbook of Nature-Inspired and Innovative Computing," Handb. Nature-Inspired Innov. Comput., no. June 2014, pp. 0–24, 2006, doi: 10.1007/0-387-27705-6.

[49]. J. Ruiz-Rosero, G. Ramirez-Gonzalez, and R. Khanna, “Field Programmable Gate Array Applications—A Scientometric Review,” Computation, vol. 7, no. 4, p. 63, 2019, doi: 10.3390/computation7040063.

[50]. J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, “A survey of machine learning for big data processing,” EURASIP J. Adv. Signal Process., vol. 2016, no. 1, 2016, doi: 10.1186/s13634-016-0355-x.

[51]. K. S. Divya, P. Bhargavi, and S. Jyothi, “Machine Learning Algorithms in Big data Analytics,” Int. J. Comput. Sci. Eng., vol. 6, no. 1, pp. 63–70, 2018, doi: 10.26438/ijcse/v6i1.6370.

[52]. D. Fumo, “Types of Machine Learning Algorithms You Should Know,” Towar. Data Sci., pp. 1–7, 2017.

[53]. C. Mamatha, P. Buddha Reddy, M. A. Ranjit Kumar, and S. Kumar, “Analysis of big data with neural network,” Int. J. Civ. Eng. Technol., vol. 8, no. 12, pp. 211–215, 2017.

[54]. M. Vennapusa and S. Bhyrapuneni, “A comprehensive study of machine learning mechanisms on big data,” Int. J. Recent Technol. Eng., vol. 7, no. 6, pp. 773–779, 2019.

[55]. O. Obulesu, M. Mahendra, and M. Thrilokreddy, “Machine Learning Techniques and Tools: A Survey,” Proc. Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2018, no. Icirca, pp. 605–611, 2018, doi: 10.1109/ICIRCA.2018.8597302.

[56]. S. L. Nita, “Machine Learning Techniques Used in Big Data,” Sci. Bull. Nav. Acad., vol. 19, no. 1, pp. 466–471, 2016, doi: 10.21279/1454-864x-16-i1-078.

Cite this Article

Shahid Mohammad Ganie, Majid Bashir Malik, Tasleem Arif, "Various Platforms and Machine Learning Techniques for Big Data Analytics : A Technological Survey", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 3 Issue 6, pp. 679-687, July-August 2018.

Journal URL : https://ijsrcseit.com/CSEIT217242