International Journal of Engineering & Technology, 7 (3.1) (2018) 93-97

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Review on Latest Technologies in Big Data Analysis

Castro S[1]*, Pushpalakshmi R[2]

[1]*Assistant Professor, Department of Information Technology, Kalasalingam Institute of Technology, Krishnankoil, India.*
[2]*Associate Professor, Department of Information and Technology, PSNA College of Engineering and Technology, Dindigul, India.*
**Corresponding 22222222author E-mail: suseelcastro@gmail.com*

Abstract

In this digital world, the modern information systems have produced a large amount of data which needs huge depositary in terms of terabytes for storage. Some of the digital technologies such as cloud computing and Internet of Things (IoT) are considered as the major sources of such large data. It is necessary to extract knowledge by analyzing these huge data which needs several attempts at multiple stages for decision making. Thus, the recent researches have focused on the analysis of big data. The main aim of this paper is to investigate the challenges of big data, applications, opportunities, implantation tools and its research problems. Thus, this study presents a platform to investigate big data at various levels. Moreover, it initiates a novel perspective for researchers to provide the solutions according to the challenges and research problems.

Keywords: *Big data analysis, cloud computing, Internet of Things, Data Storage, and Knowledge Discovery*.

1. Introduction

In recent years, the penetration of Internet has increased constantly, as thousands of people use web for browsing, communicating each other through social media and email, and access some of the wireless multimedia services like mobile digital TV [1, 2]. Also, there are various necessitating mobile network services are available that need high data rates to perform particular operations like synchronization of storage device to cloud computing servers such as high resolution videos [3-5]. The access to information on global communication infrastructure with the developments in digital storage has built a huge amount of data like Internet and mobile streaming data. Furthermore, the Investigations are based on the data analysis in several knowledge oriented fields like management, engineering and science.

In mobile big data, the location based services are much important which gathers location data and these are applied to upgrade and personalize the mobile communication services. Thus, the computing and storage has become essential utilities which are pervasively available. Besides, several algorithms have been proposed for connecting datasets to authorize more complex investigations. The paradigm of big data refers to an extremely huge and complex datasets in terms of terabytes, exabytes and petabytes. As the conventional data processing techniques are insufficient for collecting, storing and analyzing the data, looking to gather knowledge from the collected data and make an interpretation of it into competitive benefits. Therefore, the system requires more computing power to process big data and the cloud computing provides storage to store such large data. In such situation, the network accessible data storages are provided by some of the global cloud providers like Amazon [6], Google [7], IBM [8], and Microsoft [9], cost by gigabyte per month where the computing cycles are cost by CPU per hour [10].

2. Problems in Big Data Analysis

Now-a-days, big data analysis has gained more concern, which is the accumulation of data from several fields such as medicinal services, public organization, marketing, biochemistry, banking, and other interdisciplinary research areas. Some of the web based applications have experienced big data as often as possible, for example, social network computation, web content and archives, and web search ordering. The social network computation contains analysis of social networks, Internet communities, prediction marketing, and recommender and reputation frameworks. Likewise, web search ordering comprises Institute of Electrical and Electronics Engineering (IEEE) explorer, Indian Standard Institute (ISI), Scopus and Thomson Reuters and so on. Through these benefits of big data, the future researchers can obtain few novel opportunities in the processing of knowledge based tasks. But those opportunities consistently arise a few challenges. Figure 1 illustrates the features of big data.

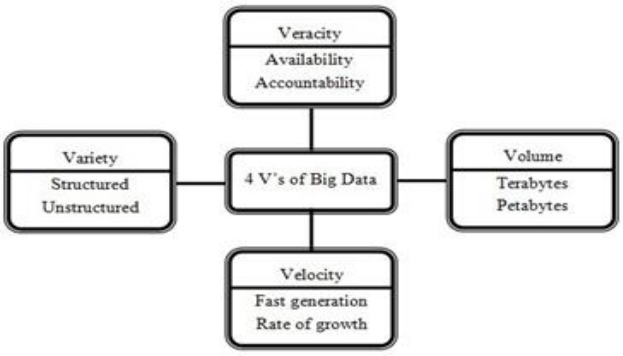


Fig. 1: Big Data Features

In order for analyzing the big data, it is essential to deal some of the challenges like data privacy and security, and computational

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

complexity and its techniques. For instance, several statistical strategies can provide better performance on small size data that never scale to large data. Correspondingly, numerous computational methods can provide good performance on small data which encounter important challenges in big data analysis. There are several researchers have developed novel techniques to encounter several challenges in healthcare services [11, 12]. In this article, the challenges in big data analysis are categorized into four wide classifications such as data analytics and its storage, scalability and data visualization, computational complexity with knowledge discovery, and data security. The following subsections have briefly discussed these issues.

Data Analysis and Storage: As of now, there is exponential growth in data size by means of several recent technologies like mobile and cellular devices, technologies of aerial sensing, remote sensing, RF id readers and so forth. In order to store these large data, one has to spend considerable amount of money whereas he/she neglected eventually because there is no adequate storage space for storing their data. Thus, the initial challenges of big data analytics are medium for storage and higher speed for input and output. In these instances, the accessibility of data must have the higher priority to accomplish representation as well as knowledge discovery. The major reason for that is, it should be accessed effectively and instantly to perform further investigation. In past few years, the researchers employ hard disk drives for storing the collected data which slower the performance of random inputs/outputs compared with sequential inputs/outputs.

In order to defeat such constraint, a few techniques such as solid state drive (SSD) and phrase change memory (PCM) was presented. Nevertheless, the present data storage approaches cannot have the expected performance to process big data. Data diversity is considered as another major challenge of big data analysis. With the consistently developing of datasets, there is a significant increase in data mining tasks. Moreover, when working with huge datasets it has to take the feature selection and data selection and reduction as an important task. This introduces remarkable challenges for the researchers. It is due to the conventional algorithms that always do not respond in a sufficient time while handling such large amount of data. In these days, data consistency is taken as a main challenge which is assured by formulating novel machine learning techniques and some automation processes. Moreover, clustering of these huge datasets helps to analyze the big data that needs major concern [13].

It makes possible by few advanced technologies like hadoop and map Reduce for gathering huge amount of semi- and unstructured data in an adequate period of time. The primary designing challenge is that the effective way of analyzing such large data to obtain best knowledge. To this end, a usual process is to perform translation of semi- and unstructured data into structured form and the knowledge is extracted by applying data mining techniques. Das and Kumar [14] have discussed a novel framework for analyzing large data. Likewise, Das et al [15] have given the detailed description of data analytics for public tweets. In such case, the primary challenge is to give more care to design data storage frameworks and for developing efficient tools for data analysis which render assurance on the data output when it originates from various sources. Besides, model of machine learning techniques for analyzing data is necessary to enhance system efficiency and scalability.

Discovery of Knowledge and its Computational Complexity: In big data, knowledge discovery and data representation are considered as the primary concerns which incorporate several sub fields like data verification, archiving, maintenance, preservation, data representation, and recovery. Various tools are available to perform such knowledge discovery and data representation such as fuzzy sets [16], near sets [17], rough sets [18], soft sets [19], principal component analysis (PCA) [20], and traditional concept analysis [21] and so on. Moreover, numerous hybridized approaches are likewise formulated for processing real time issues. Every one of these approaches is problem subordinate. Furthermore, few of these approaches are not appropriate to deal vast datasets in serial computers. Simultaneously, few of the approaches possess better features of scalability across parallel computers.

Since the exponential increase in the size of big data, the existing tools are not effective for processing such huge size data to obtain significant data. The well known techniques for the case of huge dataset management are data mart and data warehouse. This data warehouse is typically responsible for storing data originated from operational frameworks. Likewise, the data mart is depended on the data warehouse and enables data investigations. It is required more computational complexity for analyzing huge datasets and the primary concerns is to deal irregularities and vulnerability present in such datasets.

Generally, the computational complexity is modeled systematically which is commonly used. It is so hard to develop complete mathematical frameworks which are widely applicable for big data analysis. Comprehension of specific complexities is used to process the analytics on a domain specific data easily. A progression of such improvement could implement big data analysis for various fields. Towards this direction, several researches and surveys have been done by employing machine learning approaches with minimum storage requirements. The fundamental aim of those researches is to reduce processing cost and computational complexity [22-24].

But, present tools of big data analysis possess poor performance in dealing with computational complexities, vulnerability and irregularities. It results a great challenge to build novel systems and strategies which can handle computational complexities, vulnerability and irregularities efficiently.

Data Visualization and Scalability: In big data analysis, scalability and security are considered as the most critical challenges. In past few decades, the researchers have concentrated on the acceleration of data analysis and their accelerate processors operates according to the Moore's Law. For the previous years, it is essential to formulate multi-resolution, sampling, and online data analysis strategies. In the aspect of big data analytics, incremental strategies can provide good property on scalability. To scale the data size faster than CPU speeds, it is necessary to embed more number of cores with the processor technology which has dramatic shifts naturally [25]. This kind of processor shifts result the evolution in parallel computing. Some of the real time applications of parallel computing are, social networking, navigation systems, web search, financing, timeliness and so forth.

The main goal of data visualization is to show them more sufficiently utilizing a few strategies of graph theories. Graphical perception furnishes the connection between information with legitimate translation. Notwithstanding, online commercial centers such as Amazon, Flipkart, E-bay etc have a huge number of clients and billions of products to retail every month which produces a considerable amount of data. There are some organization utilizes Tableau tools to analyze the visualization of big data. Therefore, it possesses the ability to convert complex and huge amount of data into natural pictures.

These assistance workers of an organization for visualizing the search relevance observe the most recent client feedback, and their sentimental analytics. Nevertheless, present tools for big data visualization always produce poor system performance in execution, reaction time, and scalability. From this, it can be understand that the big data have generated numerous challenges for the advancements of both hardware and software that prompts to few computing technologies such as cloud computing, parallel computing, and distributed computing, and visualization and scalability challenges. In order to get over these issues, it needs to correlate more numerical models to software engineering.

Data Security: In the analysis of big data, a vast amount of data is analyzed and correlated to obtain meaning patterns. Each organization possesses different security policies to protect their private and fragile information. Protecting sensitive data needs a major

concern in the big data analytics. There is a massive security hazard related with big data analysis [26]. Thus, data security is turning into a major big data analytics issue. The big data security can be improved by utilizing the methods of validation, approval, and encryption. Different security efforts that the applications of big data encounter are, size of the network, wide range of devices, monitoring of real time security, and absence of intrusion framework [27, 28].

The security issues generated by the big data have gained the attention of data security. Hence, it is essential to give more attention to present a model of multi stage security and preservation frameworks. Though various researches have been done to provide security to big data, but it necessitates more improvements [27]. The significant challenge is to build a privacy preserving data model with multistage security for big data.

3. Research Issues and Challenges

All The growth and advancements in multimedia applications, social networking, e-commerce, distributed computing and cloud computing have significantly increase the data. Moreover, since the requirements of retail analysis are always developing, the traditional hub-and-spoke structures cannot able to fulfill the requests and, thus, novel and upgraded designs are vital [29]. In such situation, open research problems and current challenges are confronted which incorporates data storing, capturing, processing, analyzing, filtering, searching, sharing, querying, visualizing, and privacy preserving of the large amount of data. The previously mentioned problems are classified and explained as [30]:

Data maintenance with storage: The existing data maintenance techniques cannot fulfill the requirements of big data because of their limited storage space. Since the data volume grows exponentially which needs external storage capacity. Moreover, the current techniques are not capable of to storing data effectively due to the heterogeneous property of big data.

Data transfer and screening: In cloud computing, bandwidth capacity of the network is the major disadvantage. So the data transfer is a challenging task to overcome, particularly when the data size is too large. In order to manage large scale and organized datasets, some of the techniques like data marts and data warehouses are considered as better solutions. Data warehouses are said to be social database frameworks which authorize the data analysis, storage, and reporting, but the data marts depend on data warehouses and enable the data analysis. In such situation, NoSQL databases are presented as a possible technology for distributed and large scale data maintenance and its database model [31]. The schema-free alignment is considered as the major benefit of NoSQL databases and it enables the fast changes of the data structure and evades revising of tables.

Data execution and analysis: In big data analysis, query reaction time is taken as a serious issue, as sufficient time is required while crossing information in a database and executing real-time data analysis. An adaptable and reconfigured grid for the enhancement of data processing and integration of applications and data-parallelization techniques can be more efficient methodologies to extract more significant knowledge from available datasets.

Data privacy with security: The security issues are observed regarding the processing and storage of big data because third-party services can perform data hosting and some other risky operations. The static data based strategies are used in recent technologies, though the big data involves dynamic variation of present and additional data or attribute changes. Data mining of privacy preserving without revealing sensitive private data is another challenging area to be researched.

4. Tools for Analyzing Big Data

There are several tools are presented for big data processing and analysis. In this segment, we talk about some recent approaches for big data analysis with importance on three critical developing tools in particular Apache Spark, Map Reduce, and Storm. Several existing tools focus on batch execution, stream execution, and interactive investigation. Most of the batch execution tools depend on the infrastructure of Apache Hadoop for example, Mahout and Dryad. Applications of stream data are typically employed for the real time analysis. Strom and Splunk are the few models of large-scale data streaming platforms. The process of interactive analysis enables cloud users to make interaction directly at real time to perform their private analysis, for instance, Dremel and Apache are some of the platforms of big data analysis which supports interactive data analysis. To develop the projects on big data, these tools are employed. Various researchers have discussed the tools and techniques for big data analysis [32, 33]. Huang et al [34] have analyzed the regular work procedure of big data analysis.

Apache Hadoop and Map Reduce: Apache Hadoop and Map reduce are known as the most implemented software platform for the analysis of big data. It comprises of hadoop distributed file system (HDFS), map reduce, kernel of hadoop, and apache hive and so forth. To process huge datasets, the map reduce is employed which is a programming model depends on divide and conquer strategy. These divide and conquer strategies are executed in two phases, for example, Map phase and Reduce phase. There are two types of nodes are present in hadoop processing namely master nodes and worker nodes. The operation of master nodes is to split the input into number of sub issues and disseminates them to the work nodes in the system to proceed map phase. After that, the results for the entire sub issues are combined in reduce phase by the master node. In addition, Hadoop and Map Reduce functions as an effective programming structure to solve the issues of big data. It is likewise useful in fault tolerance storage and better throughput big data processing.

Apache Mahout: The main objective of apache mahout is to present a commercial based and scalable machine learning strategies for the applications of intelligent and wide scale data analysis. The mahout comprises some of the core algorithms such as data clustering, data classification, pattern mining of data, large volume data dimensionality reduction, transformative algorithms, data regression, and filtering of collaborative batches. These algorithms are operate on the platform of Hadoop via map reduce system. The aim of mahout is to construct energetic, responsive, different communities to enable conversations on the assignment and future use cases. The fundamental goal of Apache mahout is to present an effective tool to reduce big challenges. Several organizations those who established the algorithms of scalable machine learning are Amazon, Facebook, Google, IBM, Twitter, and Yahoo [35].

Apache Spark: Apache spark is defined as an open source framework for processing big data which is developed to obtain speed process, and advanced analysis. Initially, it was established in 2009 at UC Berkeleys AMP Lab and it is very simple to use. In the year of 2010, this apache spark was open sourced as an Apache assignment. Spark is needed to compose applications rapidly in Java, Python and Scala. Moreover to the map reduce process, it validates SQL queries, data streaming, processing of graph data and machine learning. The spark executes over previous infrastructure of hadoop distributed file system (HDFS) for providing improved and additional operations.

The spark comprises a few elements such as cluster organizer, driver programs, and laborer nodes in which the driver programs act as the initiating point for implementation of the application over the spark clusters. The cluster organizer assigns the network resources and the laborer nodes to perform the processing of data as tasks. Every application will possess a group of data processors named executors which are in response of implementing the tasks. The main benefit is that it offers support to deploy spark based

applications in a current hadoop clusters. The architecture model of Apache Spark is illustrated in Figure 2. The following illustrates the different features of Apache Spark.

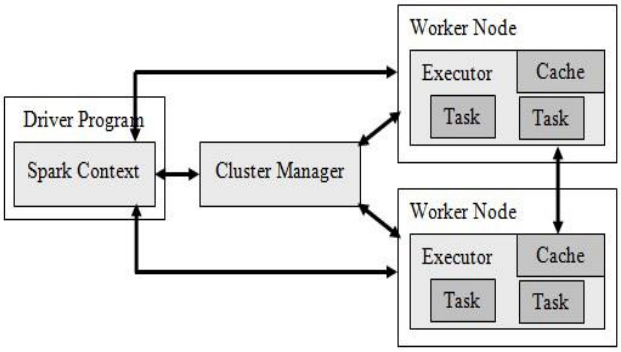


Fig. 2: Architecture of Apache Spark

- The fundamental aim of spark incorporates resilient distributed datasets (RDD) that accumulate data in its memory and renders fault tolerance with no replication. This can provide iterative computation and enhances their speed as well as resource utilization.
- Additionally, the Map Reduce model can also support data streaming, graph processing, and machine learning algorithms and these are the principal benefit of Apache spark.
- Another benefit is that, the application program can be executed by the users in various languages like Java, R, Python, and Scala. This is conceivable as it accompanies with high level libraries for sophisticated analysis. Those typical libraries enhance the developer profitability and can be consistently integrated to make complex work processes.
- In a Hadoop cluster, the spark can able to execute its application which is faster up to 100 times, and faster 10 times while executing on disks. It is conceivable due to the decrease in number of read/write activities to disk.
- The programs are written in the programming language of scala which are executed over the environment of java virtual machine (JVM). Moreover, it can support some of the programming languages like java, python and R to develop applications using Spark.

Dryad: Dryad is a well known programming model to implement parallel as well as distributed programs to handle vast context based on the dataflow graphs. It comprises of a set of processing nodes, and a client utilizes the cluster computer resources for executing their own programs in a distributed manner. Without a doubt, a dryad client utilizes a large number of machines, every one of them with different processors or data cores. The significant benefit is that a client does not require for knowing anything about simultaneous programming. The dryad applications execute computationally coordinated graphs that are made up of computational vertices with communication channels. Thus, the dryad renders a substantial number of operations comprising creating a task graph, task scheduling to allocate machines for the available tasks, handling of transition failure in clusters, gathering of execution metrics, task visualizing, imploring the user characterized strategies and progressively refreshing the activity chart in light of these policy choices without understanding the semantics of the graph vertices [36].

Storm: To process large data stream, storm is required which is a fault tolerant as well as distributed real time computation frameworks. It is particularly intended to process real time operations in contrast with Hadoop that is to execute batch processing. Furthermore, it is likewise simple to set up and control, fault-tolerant, and scalable for providing aggressive performances. The storm clusters are obviously most similar with hadoop clusters. Over storm clusters, the users can execute several topologies for various storm jobs whereas the hadoop platform executes map reduce tasks for relative applications. There are number of contrasts between map reduce tasks and their topologies. The essential contrast is that the map reduce tasks finally completes whereas the topology processing messages constantly, or until the point that user end it.

The storm clusters comprises of two sorts of nodes, for example, master node and laborer node. There are two types of operations performed by the master nodes and laborer nodes like nimbus and supervisor respectively. These two operations possess similar roles as per the job and task trackers of map reduce structure. The nimbus is responsible for code distribution over the storm clusters, task scheduling and allocating to laborer nodes, and supervising the entire framework. The tasks are complied by the supervisors which are them allocated by the nimbus to them.

Moreover, it begins and ends the operation as fundamental based on the guidelines of nimbus. The entire computational innovation is apportioned and disseminated to various labor operations and every labor process executes a bit of the topology.

5. Future Work Suggestions

Today, the data is gathered from various applications over a wide classification of fields throughout the world and the size of collected data is expected to double in every two years. This has no usage unless those are investigated to obtain effective information. This requires the advancement of strategies that can be utilized to encourage the analysis of big data. The advancement of intense PCs is an advantage to execute these strategies prompting computerized frameworks. The change of information into knowledge is in no way, form a simple task to obtain better performance in large scale data execution, comprising utilizing parallelism of present and forthcoming computer architecture to perform data mining. Additionally, these data may include vulnerability in various structures.

A wide range of models such as neural networks, fuzzy sets, rough sets, and soft sets, their speculations and mixture models acquired by joining at least two of these models have been observed to be productive in depicting information. Those models are so much useful for data analysis as well. Generally, big data are lessened to incorporate just the critical attributes essential from a specific review perspective or relying on the application territory. In this way, lessening methods have been produced. Frequently the data gathered have missing measures. These measures should be produced or the tuples possessing these missing measures are removed from the datasets before performing data analysis.

More essentially, these new difficulties may include, in some cases even disintegrate, the execution, effectiveness and scalability of the devoted data intensive computing frameworks. Once in a while, the upcoming strategy prompts loss of data and subsequently not favored. This raises numerous research problems in the business and research group in types of catching and operating data in an effective manner. Moreover, quick execution while accomplishing better performance and maximum throughput, and accumulating it effectively for later usage is another problem. Furthermore, programming of big data analysis is an essential and challenging problem. Communicating the application requirements for data access and modeling programming language theory to utilize parallelism are a quick need [37].

Furthermore, machine learning ideas and tools are getting fame among researchers to promote significant outcomes from these ideas. Researches in the territory of machine learning for big data have concentrated on data execution, implementation of algorithms, and their advancements. As of late, several machine learning tools for big data are initiated that necessitates drastic variation to choose it. We contend that while every one of the tools has their benefits and impediments, more effective tools can be built for managing issues essential to big data. The effective tools to be built must have requirements to deal variation and noisy information, vulnerability and irregularity, and missing measures.

6. Conclusion

Due to the growth of advanced digital technology, the data generation is at the dramatic pace and the manual analysis of this big data is a challenging one. In this survey, several research challenges, problems, applications and implementation tools are investigated. From this study, it can be understand that each platform for big data has its unique focus and specific functionalities in which some of them are intended to design for batch processing and some of them for real time data analysis. Using different techniques, there are several analysis are performed such as cloud computing, data mining, data stream processing, intelligent analytics, machine learning, statistical analysis, quantum analysis and so on. In future, an effective and efficient solution will develop to propose a novel technique to solve the big data problems.

References

[1] Kryftis, Y., Mavromoustakis, C.X., Batalla, J.M., Mastorakis, G., Pallis, E., Skourletopoulos, G.: Resource usage prediction for optimal and balanced provision of multimedia services. Proceedings of the 2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD 2014), pp. 255–259. Greece, Athens (2014)

[2] Skourletopoulos, G., Xanthoudakis, A.: Developing a business plan for new technologies: application and implementation opportunities of the interactive digital (iDTV) and internet protocol (IPTV) television as an advertising tool. Bachelor's Degree Dissertation, Technological Educational Institute of Crete, Greece (2012)

[3] Mavromoustakis, C.X., Mastorakis, G., Bourdena, A., Pallis, E., Kormentzas, G., Rodrigues, J.J.P.C.: Context-oriented opportunistic cloud offload processing for energy conservation in wireless devices. Proceedings of the IEEE Global Communications Conference (GLOBECOM 2014)—The Second International Workshop on Cloud Computing Systems, Networks, and Applications (CCSNA), pp. 24–30. Austin, Texas, USA (2014a)

[4] Mavromoustakis, C.X., Andreou, A., Mastorakis, G., Bourdena, A., Batalla, J.M., Dobre, C.: On the performance evaluation of a novel offloading-based energy conservation mechanism for wireless devices. In: Proceedings of the 6th International Conference on Mobile Networks and Management (MONAMI 2014), 22–24 Sept 2014. Wuerzburg, Germany (2014b)

[5] Mavromoustakis, C.X., Mastorakis, G., Bourdena, A., Pallis, E., Kormentzas, G., Dimitriou, C.: Joint energy and delay-aware scheme for 5G mobile cognitive radio networks. In: Proceedings of IEEE GlobeCom 2014, Symposium on Selected Areas in Communications: GC14 SAC Green Communication Systems and Networks, Austin, TX, USA (2014c)

[6] Amazon Web Services, Inc.: Elastic Compute Cloud (EC2). http://aws.amazon.com/ec2 (2015). Accessed 18 Oct 2015

[7] Google, Inc.: App engine—platform as a service. https://cloud.google.com/appengine (2015). Accessed 18 Oct 2015.

[8] IBM Corporation: IBM bluemix. http://www.ibm.com/cloud-computing/bluemix (2015b). Accessed 18 Oct 2015

[9] Microsoft Corporation: Microsoft azure: cloud computing platform and services. https://azure. microsoft.com (2015). Accessed 18 Oct 2015

[10] Buyya, R., Ranjan, R., Calheiros, R.N.: Intercloud: utility-oriented federation of cloud computing environments for scaling of application services. Algorithms Arch. Parallel Process. 6081, 13–31 (2010)

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

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

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

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

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] Demirkan, H., Delen, D.: Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decis. Support Syst. 55, 412–421 (2013)

[30] Chen, M., Mao, S., Liu, Y.: Big data: a survey. Mobile Netw. Appl. 19, 171–209 (2014)

[31] Han, J., Haihong, E., Le, G., Du, J.: Survey on NoSQL database. In: Proceedings of the 2011 6th International Conference on Pervasive Computing and Applications (ICPCA), Port Elizabeth, pp. 363–366 (2011)

[32] 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.

[33] 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.

[34] [34] 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.

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

[36] 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.

[37] 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, 2015.9

[38] G. Ramprabu, S. Nagarajan, "Design and Analysis of Novel Modified Cross Layer Controller for WMSN", Indian Journal of Science and Technology, Vol 8(5), March 2015, pp.438-444.