International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

A Big Data Analysis on Distributed File Storage System

C. Yosepu, C. Mahesh

Abstract*: Nowadays, the digital technologies and information systems (i.e. cloud computing and Internet of Things) generated the vast data in terabytes to extract the knowledge for making a better decision by the end users. However, these massive data require a large effort of researchers at multiple levels to analyze for decision making. To find a better development, researchers concentrated on Big Data Analysis (BDA), but the traditional databases, data techniques and platforms suffers from storage, imbalance data, scalability, insufficient accuracy, slow responsiveness and scalability, which leads to very less efficiency in Big Data (BD) context. Therefore, the main objective of this research is to present a generalized view of complete BD system that consists of various stages and major components of every stage to process the BD. In specific, the data management process describes the NoSQL databases and different Parallel Distributed File Systems (PDFS) and then, the impact of challenges, analyzed for BD with recent developments provides a better understanding that how different tools and technologies apply to solve real-life applications.

***Index Terms*: Big Data Analysis, Databases, Data Management, Massive Data, Parallel Distributed File Systems, Scalability, Storage.**

I. INTRODUCTION

Nowadays, data increased constantly with high speed due to the variety of new information by techniques and the rise of Internet of Things (IoT) and cloud computing. Every two years, the scale of global data increased continuously at a rate of 2 times than normal [1,2]. In addition, the complex task is the extraction of meaningful insights from the analysis and storage of large heterogeneous and unstructured datasets due to an increase in the number of transactions on World Wide Web and online social networks. Big Data is defined as the process to analyze the vast amount of unstructured datasets, where the major goal is to extract hidden information using technical aspects [3]. The traditional database methods are used to store, manage, analyze, visualize, process and extract the useful information from the diverse datasets, where these datasets are unstructured and imbalance, which leads to BD.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
C. Yosepu*, Department of Computer Science and Engineering, Veltech Rangarajan Dr. Sagunthala R&D Institutute of Science and Technology, Chennai, India.
Dr. C. Mahesh, Department of Information Technology, Veltech Rangarajan Dr.Sagunthala R&D Institutute of Science and Technology, Chennai, India.

The user behavior can be predicted by developing the new techniques, methods and statistical models in several applications of various disciplines. A new information can be obtained by analyzing the heterogeneous and diverse combination of large datasets [4,5]. To make decisions faster, the new techniques are designed to refine the large scale datasets. By using these new techniques, the performance of predictive models and data management operating efficiency are improved. According to the current trend of BDA, the emerging research areas focused by research community on the security issues, networking for BD and Multimedia BD in IoT [6-8]. The huge volume of network data is generated by complex networks from several domains namely, Professional networks, Biological or Social networks, etc. To address the vulnerabilities and/or threats of IoT environments, security measures are raised in the areas related to communication, application and infrastructure in IoT.

The researchers gained much interest from both industry and academia to process the data due to the speed of analyzing and extracting the large-scale data. When compared with other data mining algorithms, Random Forest (RF) is an effective method, which is used to construct the model by using feature sub-space and the famous cloud platform called Hadoop [9]. However, traditional data processing techniques are highly complex to use the large-scale imbalance data efficiently, even though they have attained better performance for low-dimensional and small-scale datasets. The major challenges include imbalance data and inefficient executions faced by most sub-dataset analysis without storage distribution over PDFS. The performance and accuracy of traditional algorithms are significantly reduced due to complex datasets with high dimensionality, larger size and complex structure [10]. In this paper, the main objective is to provide the analysis of BDA in a perspective way, which consists of various stages to explore the components for processing large-scale datasets. By these analyses, the process of complex, massive and heterogeneous datasets face new problems. The large-scale datasets are processed by main stages namely, data management, data analysis, data sources and computing frameworks. Along with the strengths, recent developments, and shortcomings, the comparison of available advanced tools with various components are illustrated by emphasizing the whole BDA. The BDA is used to make the effective decision for faster business growth, risk management, an increased level of customer satisfaction and a better Return-on-Investment.

This research paper is organized as Section II describes the taxonomy of BDA. The characteristics of BDA are presented in Section III. The components of BDA and tools to analyze the BD are defined in Section IV and V. The issues of BDA are represented in Section VI. The assessment on recent techniques used in parallel distributed storage system are explained in Section VII.

Finally, the conclusion of this research work with future development are described in Section VIII.

II. TAXONOMY OF BIG DATA ANALYTICS

Traditional conventional database techniques are unable to handle or process the large volume of heterogeneous data, where these large volumes of data are described by new concepts called BD. There are various kinds of digital contents presents in BD such as structured, semi-structured, imbalance and unstructured data and these concepts are described as follows:

A. Structured Data

The data can be easy to store, visualize, enter, query, process, and model is defined as structured data. These data are managed in spreadsheets or relational databases with specific types and sizes, and in general, it is also defined as pre-defined fields. The extraction of useful information is easy due to its rigid structure, since parallel techniques are not required for processing.

B. Semi-Structured Data

These kinds of data will not follow only a rigid model, which defines the hierarchical description for various fields within the data. Moreover, the certain elements are identified by defining the semi-structured data, which consists of different meta-model namely tags and markers. The samples of self-description data are JavaScript Object Notation (JSON) and Extensible Markup Language (XML).

C. Unstructured Data

The data is stored and represented without any pre-defined format is known as unstructured data, which consists of free form texts such as documents, emails, articles, books and media files. These kinds of data are difficult to define in a rigid form and also process the data, which leads to develop the new processing mechanisms namely NOSQL.

D. Imbalance Data

Imbalance data can be described as the classes had an unequal conveyance in any datasets which provides poor performance in classification accuracy. The issue can be classified as intrinsic and extrinsic, imbalance due to rare instances and relative imbalance, dataset complexity, and finally, imbalance with the small size dataset.

III. CHARACTERISTICS OF DATA

Initially, the 3V-models are used to refer to the properties of BD, which includes variety, volume and velocity. After that researchers extend these 3V into 5V model to define the characteristics of data by adding value and veracity. The following statements define the characteristics of BD as:

- The magnitude of the large-scale datasets is defined by volume, where the variation in size are highly depends on the time and structure of the data. Specifically, size of various types of data can be obtained from different data sources, which is defined as volume. The social media consist of YouTube, Twitter, Facebook, LinkedIn, etc. are huge in volume. In 2020, the total volume of these data may be more than 40 Zetta Bytes across the globe, where these reports are produced by International Data Corporation (IDC) [11].
- At which rate, the data is received can be defined as velocity and then, the analytical purposes are carried out by refining these data. According to the concept of data streams, the speed of the data can be calculated, which will work on a sliding window buffering model [12]. The data streams are continuously generated by social networks such as YouTube and Facebook using videos and photographs sharing.
- The different types of data include structured, semi-structured, unstructured and imbalance are used for representation of BD, which is also known as a variety. The relational data, electronic health record data, tabular data are the defined as structured data, where data collected from audio, images, blogs, smartphones and log data are the samples of unstructured data. Finally, semi-structured includes the HTML and XML and imbalance data defines the number of more data presents in majority class compared with minority class.
- The noises, biases and abnormality are present in data are defined as veracity, where these issues occur because of uncertainty. The data uncertainty defines the input data, model uncertainty describes the suitable model and finally the randomness in process is represented by process uncertainty. For instance, user's opinions are uncertain for particular events in social media, however, the data provides precious information. Therefore, the BD analytics and tools are used to manage this uncertainty of data [13].
- In business perspective, another dimension of BD is value, where these values should be noticed by the organizations. The business profits are increased by reducing the operational costs for providing better customer services [14]. The data contains less value, when it is in the original form, but these values will change into high-value assets, while applying the data analytics.

There are four types of models presents in BDA such as in-memory models, MapReduce (MR)-based systems, Massively Parallel Processing (MPP) systems and Bulk Synchronous Parallel (BSP) systems. Among multiple servers, the MPP systems can divide the data into partitions and require the memory to process the data locally, because there is no disk level sharing. The limitation of MPP systems for commercial purposes is the substantial amount of both hardware and software. The iterations (i.e. supersteps) are used for computations in BSP systems. The global synchronization is used for concurrent computations in each iteration, which are performed in parallel on cluster nodes for exchanging the information between processes.

When compared to disk-based systems, the memory-models achieved 100 times faster throughput and response time, due to changes in data storage by main memory.

IV. COMPONENTS OF BIG DATA SYSTEM

The process of massive datasets is presented in this section as systematic stages, where major stages includes data capturing, managing the data and analysis of data to refine the BD. In every stage, each components of BD are described in the following sub-sections.

A. Data Capture

The initial task is to generate and capture the data for the analyzing process. The heterogeneous sources include search engines, server logs and sensors used for capturing the data. The user-generated data and machine-generated data are majorly considered as sources for a generation that can able to form the massive datasets.

- The users generate the data on the daily basis is defined as user-generated data and the useful information are provided to others by using their data contribution [15]. The few samples for user-generated data include mobile data, tweets from Twitter, product evaluations, blog information and movie ratings.
- The combination of all activities in the form of server activities, customer transactions, user behavior records, and log data are defined as machine-generated data. These data contain database audit files, machine-readable object data (Barcodes), records from call center, click stream data and configuration files in several organizations. The data are described in unpredictable formats, because many organizations have their own machine-generated data.

B. Data Management

The main contribution of the paper is described in the section, where the crucial stages of data management is divided into two subtasks, after the generation of data from various sources. The two stages include data storage and data retrieval. The responsibility of large-scale data storage can be handled by PDFS, because the BDA concepts work on the basis of large-scale data. The NoSQL databases are used for the second task, which is a data retrieval process and the PDFS characteristics are described in the following sections:

1) File Systems

The storage of massive datasets is handled by the main component PDFS and some properties such as efficiency, reliability, scalability and availability are used to define the characteristics of data. The major purpose of PDFS is storage and sharing the files among the set of connected nodes. The major PDFS are described as Kosmos file system, Quant cast File System (QFS), Google File System (GFS), IBM General Parallel File System (GPFS), Haystack File System, Hadoop distributed file system (HDFS), TidyFS and Tao file system. When the system failure happens, the recovery of data can be carried out by GFS because it distributes each piece of information in three copies. However, the limitations of GFS are the high storage overhead and shortage of the feature called Portable Operating System Interface (POSIX) and these POSIX features are supported by IBM GPFS. The performance of PDFS is validated by main parameters namely replications, storage management, response time, read/write, throughput and scalability.

According to limitations of relational database management systems, unstructured and imbalanced data are not handled and parallelization is not carried out for heterogeneous datasets. When compared with highly available hardware, the non-relational databases are used to work in a parallel distributed environment and develop in the horizontal scalability view. In BD data management, a Hadoop-supported NoSQL database plays a major role those are described as:

2) NoSQL databases

The attributes of each and every entity are stored by using traditional relational databases in the form of rows. The entire record should read for retrieving the attributes from the databases, which can handle only a smaller dataset because it is majorly designed for transactional processing. The major difference between the traditional databases and NoSQL systems is that the NoSQL with key-value pairs' storage and retrieval follows the horizontal scalability of data, but traditional databases follows the relationships between primary/foreign keys. Therefore, unstructured data are efficiently handled by NoSQL databases.

C. Data Analysis

To process the BD, the final stage is the analysis of data, here two major subtasks are presents namely data visualization and data analytics. Machine Learning (ML) algorithms are used to model the data in analytics and statistical/data mining techniques are used for fining the hidden patterns from the BD. In the Spark and Apache Hadoop, two popular parallel distributed ML libraries namely Mahout and MLlib, respectively. FlinkML is an ML library used in Apache Flink, which is a parallel distributed processing framework for analyzing large-scale databases. The useful patterns are extracted from the massive datasets by Mahout using ML algorithms and stored in HDFS.

1) Data Visualization

The intuitive insights of hidden patterns are seen and some conclusions are drawn to make a faster decision by visualization, which is considered a major objective. The uncovered patterns are observed by visualizing the data patterns using pie charts, bar charts and histograms. In distinct dimensions, these techniques are used for observing the data patterns, where BD techniques are used to make the decisions for transforming the data into insights by visualizing the large datasets. The major features visualization tools are described as data projection, statistics, visual query analysis and data modeling. Some examples of statistics such as uni-variate, bivariate and multivariate, classification, predictive analysis and clustering are considered for data modeling. Finally, data projection is defined as multidimensional scaling and principal component analysis [16].

The large datasets are analyzed by using most popular visualization tools includes Micro Strategy, ADVIZOR, Tableau, SAS Visual Analytics (VA), JMP, TIBCO Spot fire and QlikView. Due to distributed nature of data, the capability of each data is varied with other, and this makes the data more imbalance to process.

Therefore, an effective tool is required to further analysis the data and the existing effective tools are discussed in next section.

V. TOOLS FOR ANALYZING THE BIG DATA

BD is processed by the vast number of tools and some important currents tools for evaluating the BD are discussed in this section, where the most important tools includes Apache Spark, MapReduce and Storm. The first and most significant tool for processing BD is known as Apache Hadoop [17]. The three types of BD processing, such as stream processing, interactive analysis and batch processing, are mainly concentrated by various important available tools. The samples tool for batch processing includes Dryad and Mahout, stream processing contains Splunk and Strom, whereas interactive analysis preferred Apache Drill and Dremel tools. Among these three processing, stream analysis is used for real time analysis, the users are able to directly interact for their own analysis by using interactive analysis process. The BD projects are developed by using these tools, where these lists of tools and techniques are described in [18]. The most important tools are discussed as follows:

A. Apache Hadoop and MapReduce

MapReduce and Apache Hadoop are considered as one of the most established software platforms for analyzing BD. There are several components presents in this tool such as apache hive, MapReduce, Hadoop kernel and HDFS and etc. According to divide and conquer method, MapReduce process the large dataset, which is also defined as a programming model. To solve the BD problems, the most powerful software tools are used, which includes MapReduce and Hadoop. In addition, it obtains high throughput and helped in fault-tolerant storage, while processing the data.

B. Apache Mahout

The main aim of the Apache mahout is to present the most scalable and commercial machine learning techniques for analyzing the intelligent and large scale data. Through MapReduce, Hadoop function are run by using batch based collaborative filtering. The mahout contains several core algorithms that include evolutionary algorithms, regression, pattern mining, classification, clustering and dimensionality reduction.

C. Apache Drill

The BD are interactively analyzed by using Apache Drill, which can support various types of data sources, formats and query languages. The batch analysis is performed by MapReduce, which is used by drill and also for storing the data, the drill uses the HDFS. In specific, the nest data are exploited by designing the drill. More than 10,000 servers are scaled up to achieve the capacity for processing the trillions of records in seconds and petabytes of data.

D. Apache Spark

The sophisticated analytics and the process are analyzed at a high speed can be carried out by one of the open source BD processing framework called apache spark. There are three components presents in Spark such as worker nodes, driver program, and cluster manager. On the spark cluster, the starting point of execution are served by driver program, where the resources are allocated by cluster manager. The data are processed in the form of tasks by using worker nodes. To execute the tasks, every application has an executor, which is defined as the set of processes. The spark applications are deployed in an existing Hadoop clusters, which is the major advantage of these apache spark.

E. Jaspersoft

The database columns are used to provide the results by using the open source software called Jaspersoft package. The data visualized on storage platforms such as Redis, MangoDB, Cassandra, etc., where this process are speeded by this package. Without loading, extraction and transformation, the BD is explored quickly by using the main property of Jaspersoft. A powerful Hypertext Markup Language (HTML) reports are built by using these package, moreover the vast amount of data is extracted from the store directly by using the package's property. Any user from inside or outside of the organization can obtain these generated reports.

VI. ISSUES IN BIG DATA

In this section, according to the discussion of BDA, the issues in this management and development are discussed. The BD faces challenges in some important features such as data quality, processing speed, data interpretation, exception handling of BD and visualization. In this research work, the major technical issues such as data security, data quality, and privacy are discussed.

A. Data Security

While adopting BD, user resistance is created by weak security, which will lead to damage and financial loss for firm's reputation. The unauthorized parties can access and transfer confidential information due to improper installation of security mechanisms. The strong security management protocols are introduced along with several security solutions such as building the firewalls, detection, encryptions and intrusion prevention systems into BD systems to avoid the above security issues.

B. Data Quality

The fitness of the data is defined by the data quality along with the specific purpose of usage, which is the main quality for making the decision. The quality of data may be reduced due to collection of unstructured data from wide array of sources. To evaluate the data quality, a new quality metrics are need to develop by data quality control process, which also used to repair the erroneous data and make a trade-off between assurance gains and costs.

C. Privacy

Due to a massive growth of BD technologies, the collected personal data faces challenges like security and unauthorized access. Therefore, governments, firms and individuals must focus on the security concern of these extensive personal data. Because of these privacy concerns, individuals found that the process of analyzing the BD is much more difficult to prevent from unauthorized access. The BD enhances the service quality and reducing the cost, even though privacy is an important issue for both firms and end customers. Hence, a balance between privacy and the usage of personal data for services should be considered by both firms and users. However, according to customer's service, data type, service types and regulatory environments, the balances are created, but none of the existing works focused on the privacy measures.

VII. RELATED WORKS

This research work presents the discussion of recent techniques for parallel distributed storage system, which are used in the BDA. Table 1 presents the various techniques used in BD, which consists of its advantage and limitations.

TABLE I. COMPARATIVE ANALYSIS OF VARIOUS TECHNIQUES IN DISTRIBUTED STORAGE OF BDA.

Author with Year	Methodology	Advantage	Limitation	Performance Evaluation
F. Zhang et al. [19] (2015)	proposed the Distributed Frequent Item set Mining Algorithm (DFIMA).	The efficiency of iterative computations are promoted by using Spark platform and also the amount of candidate itemsets are reduced by using matrix-based pruning menthod	The time cost of DFIMA algorithm is high, when the support number is small.	Running time is used to validate the performance of the DFIMA algorithm.
J. S. Kim et al. [20] (2016)	PARADISE is developed for BDA as a parallel processing method for an integrated database.	More complex queries such as Cartesian product and 3-way join queries are supported by PARADISE.	The performance of PARADISE is reduced due to three types of overheads on clustering attribute (i.e. selection, scan and aggregation query)	The three types of overhead such as network transfer, disk arm contention and network bottleneck are used as parameter metrics.
Z. Xu et al. [21] (2016)	Adopted a "front+back" pattern to solve the problem of redundant construction and developed an architecture for the next generation public security system.	To enhance the efficiency of tasks and usage of resources, the multiple optimized strategies are provided by retrieving the huge and heterogeneous data.	During routine detection for crime suspects, the services provided to users are not satisfactory.	Peak Signal to Noise Ratio (PSNR) and Feature Similarity Index (FSIM) are used as performance metrics.
J. Zhu et al. [22] (2017)	A parallel and distributed designed principal component analysis are implemented (dpPCA).	The method is very instructive and also used to isolate the fault, trouble shooting and decision making for large set of variables.	When the normal expectation in fault settings are out of control, the kinetic procedure within the reactor are very high.	The parameter metrics such as modeling time and fault detection rate are used to validate the dpPCA.
V. Baljak et al. [23] (2018)	Developed an open-source stream processing software to connect raw output with file storage system.	The flexibility are maintained by the platform itself and also has the capacity to include other data sources.	The two important requirements for sensitive medical data such as security and privacy are not concentrated by this method.	The effectiveness of the developed software are analyzed by using various case studies.
J. Wu et al. [24] (2018)	To optimize the control plane, the BDA based secure cluster management architecture is developed.	The legality of data sources and BDA are ensured by developing the authentication scheme and ant colony optimization.	The method failed to concentrate on distributed security data storage scheme for the SDN controlled cluster.	The parameter such as energy consumption, communication overhead and traffic rate are used.
F. Hu et al. [25] (2018)	The complex BDA and time-consuming computation tasks are facilitated by ClimateSpark.	ClimateSpark provides efficient, scalable and flexible computing framework for BDA. In addition, it efficiently queries the array-based climate and specified multi-dimensional data with high data locality.	The method is unable to adopt the high resolution raster datasets with chunking/tiling data structure.	Run time, correlation coefficient, standard deviation of input and reference datasets and finally RMS difference.
S. Khalifa et al. [26] (2019)	To avoid the pitfalls and handle the BD, the Label-Aware Distributed Ensemble Learning (LADEL) is implemented.	The inter-machine communication are minimized by executing in a single pass over the data.	The ensemble classifiers provides worse classification accuracies in ordered datasets due to lack of records for representing all class labels.	Training time, prediction accuracy and scoring time are used as performance evaluation metrics.

VIII. CONCLUSION

According to the smart devices and fast growth of the sensor, the rate of data production is increased. In this work, the advanced tools with developments are focused on analyzing the vast data, where these tools are used for managing, processing, storing and visualizing the massive data.

The discussion of analyzing the structured, unstructured and imbalance data are presented and their significant impact are also described. In a general view, the components of BD with the research of technical aspects are presented. The components include data sources,

data management and data analysis with visualization are stated and finally the discussion of recent developments with its advantage and limitations to process the massive data are presented in the form of comparative analysis. From this research work, its concluded that every platform of BD has its individual focus, where some tools are mainly developed for batch processing and some of them are used for real-time analytic. Different techniques are used to analyze the large-scale and imbalance data, where these techniques include quantum computing, intelligent analysis, machine learning, cloud computing, statistical analysis, data stream processing and data mining. These comparisons are helpful for utilizing suitable tools properly to process various BD applications. Furthermore, in future research work, the research will pay more attention to the aforementioned tools to solve problems of BD effectively and efficiently.

REFERENCES

1. J.Q. Li, P. Rusmevichientong, D. Simester, J.N. Tsitsiklis, and S.I. Zoumpoulis, "The value of field experiments," Management Science, vol. 61, No. 7, pp. 1722-1740, 2015.
2. J. Li, F. Tao, Y. Cheng, and L. Zhao, "Big data in product lifecycle management," The International Journal of Advanced Manufacturing Technology, vol. 81, no. 1-4, pp. 667-684, 2015.
3. S. Maklan, J. Peppard, and P. Klaus, "Show me the money: improving our understanding of how organizations generate return from technology-led marketing change," European Journal of Marketing, vol. 49, no. 3/4, pp. 561-595, 2015.
4. K. Pousttchi, and H. Yvonne, "Engineering the value network of the customer interface and marketing in the data-rich retail environment," International Journal of Electronic Commerce, vol. 18, no. 4, pp. 17-42, 2014.
5. R. Thackeray, B.L. Neiger, C.L. Hanson, and J.F. McKenzie, "Enhancing promotional strategies within social marketing programs: use of Web 2.0 social media," Health promotion practice, vol. 9, no. 4, pp. 338-343, 2008.
6. S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: A survey". IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 531-549, 2016.
7. S. Pouyanfar, Y. Yang, S.C. Chen, M.L. Shyu, and S.S. Iyengar, "Multimedia big data analytics: A survey," ACM Computing Surveys (CSUR), vol. 51, no. 1, pp. 10, 2018.
8. F.A. Alaba, M. Othman, I.A.T. Hashem, and F. Alotaibi, "Internet of Things security: A survey," Journal of Network and Computer Applications, vol. 88, pp. 10-28, 2017.
9. Md E. Islam, Md R. Islam, and ABM S. Ali, "An approach to security for unstructured big data," The Review of Socionetwork Strategies 10.2 (2016): 105-123.
10. W. Kun, L. Tong, and X. Xiaodan, "Application of Big Data Technology in Scientific Research Data Management of Military Enterprises," Procedia computer science, vol. 147, pp. 556-561, 2019.
11. F. Bajaber, R. Elshawi, O. Batarfi, A. Altalhi, A. Barnawi, and S. Sakr, "Big data 2.0 processing systems: Taxonomy and open challenges," Journal of Grid Computing, vol. 14, no. 3, pp. 379-405, 2016.
12. S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren, and D. Valerio, "A software reference architecture for semantic-aware Big Data systems," Information and software technology, vol. 90, pp. 75-92, 2017.
13. Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International journal of information management 35.2 (2015): 137-144.
14. I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," Business Horizons, vol. 60, no. 3, pp. 293-303, 2017.
15. J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," IEEE Pervasive Computing, vol. 7, no. 4, pp. 10-11, 2008.
16. L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, and D. Keim, "Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems," In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173-182, 2012.
17. C.P. Chen, and Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information sciences, vol. 275, pp. 314-347, 2014.
18. D. B. K. Kamesh, V. Neelima, and R. Ramya Priya, "A review of data mining using bigdata in health informatics." International Journal of Scientific and Research Publications, vol. 5, no. 3, pp. 1-7, 2015.
19. F. Zhang, M. Liu, F. Gui, W. Shen, A. Shami, and Y. Ma, "A distributed frequent itemset mining algorithm using Spark for Big Data analytics," Cluster Computing, vol. 18, no. 4, pp. 1493-1501, 2015.
20. J. S. Kim, K. Y. Whang, H. Y. Kwon, and I. Y. Song, "PARADISE: Big data analytics using the DBMS tightly integrated with the distributed file system," World Wide Web, vol. 19, no. 3, pp. 299-322, 2016.
21. Z. Xu, L. Mei, C. Hu, and Y. Liu, "The big data analytics and applications of the surveillance system using video structured description technology". Cluster Computing, vol. 19, no. 3, pp. 1283-1292, 2016.
22. J. Zhu, Z. Ge, and Z. Song, "Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data," IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1877-1885, 2017.
23. V. Baljak, A. Ljubovic, J. Michel, M. Montgomery, and R. Salaway, "A scalable realtime analytics pipeline and storage architecture for physiological monitoring big data," Smart Health, vol. 9, pp. 275-286, 2018.
24. J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "Big data analysis-based secure cluster management for optimized control plane in software-defined networks," IEEE Transactions on Network and Service Management, vol. 15, no. 1, pp. 27-38, 2018.
25. F. Hu, C. Yang, J. L. Schnase, D. Q. Duffy, M. Xu, M.K. Bowen, and W. Song, "ClimateSpark: An in-memory distributed computing framework for big climate data analytics," Computers & geosciences, vol. 115, pp. 154-166, 2018.
26. S. Khalifa, P. Martin, and R. Young, "Label-Aware Distributed Ensemble Learning: A Simplified Distributed Classifier Training Model for Big Data," Big Data Research, vol. 15, pp. 1-11, 2019.

AUTHORS PROFILE

Mr. C. Yosepu has more than 10+ years of experience in the field of teaching. He was awarded B.Tech in Computer Science and Engineering and M.Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad. Presently working as an Associate Professor in St.Martin's Engineering College, Hyderabad. He is a Research Scholar at Veltech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai. His area of interest includes Big Data Analytics and Data Mining.

Dr. C. Mahesh has more than 20 years of experience in the field of teaching. He was awarded B.E in Electrical and Electronics Engineering from Madras University, Chennai. He was awarded M.E in Computer Science and Engineering, Anna University, Chennai. He was awarded Doctorate in Computer Science and Engineering in the year 2016. Currently he is working as an Associate Professor in Veltech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai. His area of interest includes Neural Networks and Natural Language Processing.