International Journal of Engineering Science Invention Research & Development; Vol. IV, Issue VI, DECEMBER 2017
www.ijesird.com, E-ISSN: 2349-6185

STATE-OF-THE-ART BIG DATA ANALYTICS AND ALGORITHMS

Rajiv Senapati, D. Anil Kumar
Dept. of CSE, GIET Gunupur
rajiv461@gmail.com, danil@giet.edu

Abstract: **In the age of big data, the incoming data are of huge volume and they are composed of structured, semi-structured and unstructured data with various types including streaming data. Analysis of such kind of data is a big challenge. The features of Big data includes massive, multidimensional, heterogeneous, complex, semi-structured, incompleteness, noisy, and erroneous data which may change the statistical and data analysis approaches. The traditional data analytics approach may not be able to handle such kind of data. Now, developing high performance framework to efficiently analyze big data and to design appropriate mining algorithms to find the useful things from big data is highly essential. In this paper we have addressed some of the algorithms that can be used for Big data analytics with some necessary modifications.**

Keywords— *Big data; Data analytics; Algorithms; Data mining.*

I. INTRODUCTION

The term big data describes a situation where the volume, velocity and variety of data exceed the computing capacity for accurate and timely decision making. In every minute some terabytes of data is being generated from various sources such as hand held devices, social networking sites, Internet of things, multimedia, Machine-to-machine interactions, call detail records, environmental sensing and RFID systems and many other new applications. All these forms of data are expanding and coupled with fast-growing streams of unstructured and semi structured data from social media with characteristics of volume, velocity, and variety [1], [2], [3] as shown in Fig. 1. As a result, the whole data analytics has to be re-examined from the following perspectives:

Volume: In Big data the datasets are having orders of magnitude that are larger than traditional datasets and it requires more intelligent at each stage of the processing and storage life cycle. There are various sources of data that includes business transactions, social media and information from sensor or machine-to-machine data. Here the computing requirements exceed the capability of a single and simple computer.

Velocity: It refers to the rate at which information moves around the system, this is how Big data is different from other data systems. Data frequently entered into the system from various sources exponentially to gain insights and update the current understanding of the system. In Big data, data is continuously added, processed and analysed in the system to keep up with the deluge of new information and to produce valuable information as soon as possible when it is required.

Variety: In real scenario the data collected from various sources are in different formats in the form of structured, semi-structured, unstructured such as text documents, email, video, audio, stock ticker data and financial transactions etc. Now, around 85 percent of an organization's data is unstructured and not numeric but it still must be folded into quantitative analysis and decision making.

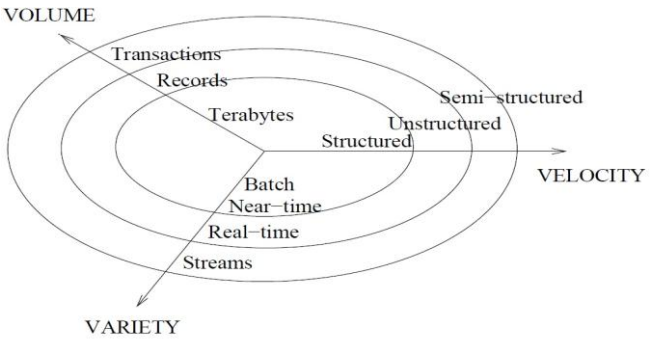

Figure 1: 3Vs of Big data

Some of the other characteristics of Big data include,

Veracity: Another issue in Big data is the variety of data sources, in addition to that the complexity of processing may also leads to a bigger challenge in evaluating the quality of the data which really influences the analysis result quality.

Variability: Variation in the data influences quality result. Hence it becomes a challenge to identify process and filter low quality data to make it more meaningful for strategy making.

Value: The objective of data analytics to derive meaningful information for decision making. It has become a challenge for big data to deliver value due to several complexities.

Rest of the paper is organized as follows. In Section 2 we have presented Literature review. In Section 3 we have discussed data analytic operations. In Section 4 we have discussed Big data analytics. Finally, Section 5 concludes this paper.

II. LITERATURE REVIEW

Big data is too difficult to process by most information systems or methods due to its variety of features such as Volume, Velocity, and Variety. Hence as per the work reported in [4] most traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to big data. The work reported in [1] presented Big data in terms of 3Vs i.e. volume, velocity, and variety. It means, data size is large, the generation of data is rapid, and it is in heterogeneous form. The work reported in [2], [3] expressed that 3Vs of Big data is still insufficient to express the Big data exactly. In the work reported in [3] expressed Big data in term of veracity, validity, value, variability, venue, vocabulary, and vagueness. As per the work reported in [6] and [7] the market of Big data will be $114 billion by early 2018. As per the forecasts reported in the work [5] and [8] the scope of big data will be grown rapidly in the forthcoming future. Apart from marketing domain, Big data has its root in many areas such as disease control and prevention [9], business intelligence [10], and smart city [11]. From this we can easily understand that big data is of vital importance everywhere. Therefore, many researchers are focusing on developing effective technologies to analyze the Big data. In this paper, we have discussed a systematic description of traditional large-scale data analytics as well as state-of-the-art data analytics algorithms. The methods used mostly for data analytics is presented in Table 1.

III. DATA ANALYTICS

As per the work reported in [12], the knowledge discovery in databases (KDD) is summarized by operations such as selection, pre-processing, transformation, data mining, and interpretation/evaluation as presented in Fig. 2.

These stages of KDD helps to gather data from various sources and then pre-processing those real world data to make it eligible for processing and then presenting the information to the relevant users. In Fig. 2, we have simplified the process of KDD into input stage, data analytics stage and output stage. During Data-in stage the operations to be performed are Data collection, gathering, selection and pre-processing. Data mining operation is performed during Data analysis stage. Evaluation and interpretation is performed to discover knowledge in Output stage.

Mechanism	Method	References
Clustering	TKM	[30]
	BIRCH	[31]
	DBSCAN	[32]
	RKM	[33]
Classification	SLIQ	[34]
	TLAESA	[35]
	FastNN	[36]
	SFFS	[37]
Association	FP-Tree	[24]
	CLOSET	[38]
	CHARM	[39]
	FAST	[40]
Sequential Pattern	SPADE	[41]
	ColSpan	[42]
	SPAM	[43]
	ISE	[44]

Table 1: Data Analysis methods for KDD.

A. DATA-IN

As shown in Fig. 2, the data collection, selection, pre-processing, and transformation operators are in

the Data-in stage. The kind of data to be analysed and selecting the relevant information from the gathered data is carried out using selection operation. Hence, the data from various sources needs to be integrated and targeted. To make the data meaningful, the pre-processing phase used to clean the noisy data, filtering the unnecessary data, inconsistent data and incomplete data. After the selection and pre-processing phase is over then the data needs to be transformed into a form that can be acceptable by data mining engine with the help of some of the methods such as dimensional reduction, sampling, coding, transformation by normalization etc.

The four phases of KDD such as data extraction, cleaning, integration and data transformation can be viewed as the pre-processing processes of data analysis [13]. In real world scenario there are chances of redundancy, incompleteness, inconsistency, noisy or outliers. The objective of the preprocessing phase is to clean them so as to make the data ready for analysis. If the data are too complex as well as too large to be handled then the preprocessing operators will try to reduce them. In case of noise the preprocessing operators are used to identify them and smooth them to make them consistent. It can be said that these operators influences the analytics results in data analysis and to reduce the complexity of data to speed up the computation time of data analysis and to improve the correctness of the analytics result.

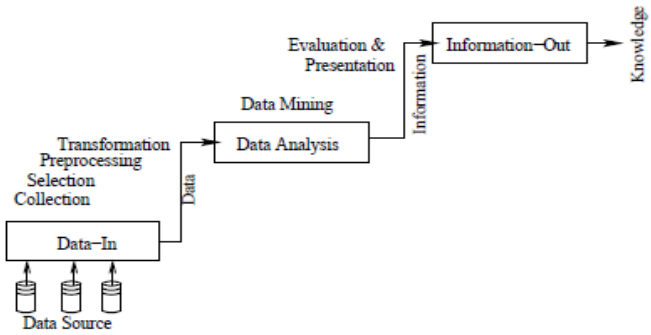

Figure 2: KDD process

B. DATA ANALYSIS

As presented in Fig. 2, the responsibility of data analysis stage is to find the hidden patterns or rules or information's from the data. The data mining methods [13] are basically used here. Apart from data mining some other traditional techniques such as machine learning and other statistical methods can also be used to analysis the data. However, the statistical methods were used in the early stage of data analysis to help in understanding the situation that we are experiencing now-a-days. Some of the domain specific algorithms such as Apriori algorithm as one of the useful algorithms designed for the association rule mining problem [14]. To accelerate the response time of a data mining operator, machine learning [15], metaheuristic algorithms [16], and distributed computing [17] were used in conjunction with the traditional data mining algorithms to provide more efficient ways for solving the real time problems.

Clustering is another data mining algorithm that can be used to understand the new input data. The basic idea of this problem is to separate a set of unknown labeled input data into different groups such as in k-means algorithm as reported in [19] and [20]. Classification is another data mining algorithm that relies on a set of labeled input data to construct a set of classifiers which will then be used to classify the unknown labeled input data to the groups to which they belongs [13]. Such problems are basically solved by using decision tree-based algorithm [21], naive Bayesian classification [22], and support vector machine (SVM) [23] techniques.

Some of the data mining techniques such as clustering and classification attempts to classify *n* number of inputted data into *k* number of groups, in which inter clusters objects behavior is highly dissimilar and intra clusters objects behavior is highly similar. Similarly, association rules mining and sequential pattern mining are the well known algorithms those are focused on finding relationships among inputted data stes. Idea behind association rule mining is to find out all the correlation between the inputted data [14]. A well known algorithm i.e. Apriori algorithm [14] is one of the most popular technique used for this purpose. Later various advanced techniques are adapted to reduce the cost of the Apriori algorithm, such as by applying the genetic algorithm with Apriori etc. If we consider the sequence or time series of the inputted data, then it will be referred to as the

sequential pattern mining problem [26]. Several Apriori-like algorithms were presented to solve such kind of problems.

C. INFORMATION-OUT

Evaluation and analysis of operations that are used for data analytics are performed in this stage. Evaluation basically measures the results obtained from the analysis. It is considered as one of the important phase of data mining. Some other techniques such as Sum of Squared Errors (SSE) is used by the selection operation of the genetic algorithm (GA) for clustering problems [18]. To solve classification problem of data analytics, two of the major goals to be set are cohesion i.e. the distance between each data point and the centroid of its cluster should be minimum and coupling i.e. the distance between data which belong to different clusters should be maximum. In most of the clustering or classification problems the SSE is used, which can be defined using the following expression.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} D(X_{ij} - C_i)$$

Where C_i can be expressed as,

$$C_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

Where k is the number of clusters which is typically given by the user, n_i is the number of data in the i^{th} cluster, x_{ij} is the j^{th} data in the i^{th} cluster, c_i is the mean of the i^{th} cluster, and n is the number of data. The most commonly used distance measure for the data mining problem is the Euclidean distance. Another approach to provide the useful information to the end user in a comprehend the meaning way is the Graphical user interface (GUI). As the work reported in [28], it is required to overview the data first, then zoom and filter the data and then finally retrieve the details on demand by the end users as per the situation [27, 29].

IV. BIG DATA ANALYTICS

Now a day, the data that need to be analyzed are not just large but they are composed of structured, semi-structured and unstructured data with various types including streaming data. The Big data has the unique features of massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous data which may change the statistical and data analysis approaches. The traditional data analytics may not be able to handle such kind of data. Now, it is very essential to develop high performance framework to efficiently analyze the Big data as well as to design appropriate data analytics algorithms to find the useful things from big data. In this section we have highlighted some of the algorithms for big data analysis.

Clustering algorithms: Clustering is the process of grouping the data into classes and clusters, so that the objects within the cluster have high similarity in comparison to one another but are very dissimilar to the objects in other clusters. This technique has its root in many areas such as machine learning, data analytics, data mining, statistical analysis, biology etc. In today's scenario the traditional algorithms are not sufficient for big data analysis. The reason is today's data analytics requires all the data are to be in the unique format and be loaded into the same machine so as to find some meaningful things from the entire dataset. Analyzing a large volume with multi-dimensional dataset has attracted many researchers from various disciplines. Reducing the data complexity is one of the major issues in for big data clustering. Data clustering can be categorized as single-machine clustering and multiple-machine clustering which is the parallel and MapReduce solution [45].

Classification algorithms: Classification is a form of data analytics that can be used to abstract models that describes important data classes. In the context of Big data analytics, the traditional algorithms can be modified so as to work under parallel processing environment to meet the bid data analytics challenge. As per the work reported in [46] the algorithm takes data inputted by various distributed data sources and they will be processed by a heterogeneous set of learners. In a distributed data classification systems one operator is to perform a classification function by itself while the other is to forward the inputted data to another learner to have them labeled. The information will be exchanged among learners during the process. This technique is called as

cooperative learning. This method can be used to improve the efficiency of classification problem in big data.

Frequent pattern mining algorithm: Basically the frequent pattern mining also called as association rule mining and sequential pattern mining is applied in many areas in past especially in case of large-scale datasets. The huge volume of transactions i.e. more than tens of thousands is the issues in Big data. Thus handling such huge scale of data were studied for several years, such as FP-tree [24]. Now, advanced techniques such as parallel computing and cloud computing attracted the researcher to think in this direction. Map-reduce is a solution used in to enhance the performance of the frequent pattern mining algorithm [47, 48] and can be used in cloud platform [49]. In observed from many studied that the performance of map-reduce model is remarkable for big data analysis as compared to other traditional data analytics methods.

Machine learning for big data mining: Machine learning algorithms are specifically design for specific problems. Typically this algorithm is used for searching algorithms. The machine learning algorithms basically used for finding approximate solution for the optimization problems. Thus it can be used for most of the data analytics problems. Machine learning algorithms are not only be used for solving the clustering problem [18] but also it can also be used to solve the frequent pattern mining problem [25]. The machine learning algorithm will potentially improve data analysis operations in KDD.

Data mining algorithm for map-reduce solution: For Big data mining most of the traditional data mining algorithms are not used directly because they are not particularly designed for parallel computing. Now, several attempts have been made by the researchers to modify the algorithm to make them applicable to parallel platform like Hadoop.

V. CONCLUSION

This paper addressed big data analytics in three different phases that are Data-in, Data analysis and Information-out for mapping the data analysis process of knowledge discovery. The major challenge of big data analytics is its size and variety. Many researchers have proposed many techniques to overcome this issue. From this paper it can be concluded that the traditional algorithms can be modified so as to make it eligible for big data analytics.

REFERENCES

[1] Laney D, "*3D data management: controlling data volume, velocity, and variety*," META Group, Tech. Rep., 2001.
[2] Van Rijmenam M, "*Why the 3v's are not sufficient to describe big data*," BigData Startups, Tech. Rep., 2013
[3] Borne K, "*Top 10 big data challenges a serious look at 10 big data v's*," Tech. Rep., 2014
[4] Fisher D, DeLine R, Czerwinski M, Drucker S, "Interactions with big data analytics,"*Interactions*.2012; 19(3):50-9.
[5] Press G, *"$16.1 billion big data market: 2014 predictions from IDC and IIA*," Forbes, Tech. Rep.,2013.
[6] Taft DK, "*Big data market to reach $46.34 billion by 2018*," EWEEK, Tech. Rep.,2013.
[7] "*Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive analytics*," ABI Research,Tech.Rep.2013.[Online].Available:https://www.abiresearch.com/press/
[8] Kelly J, Floyer D, Vellante D, Miniman S, "*Big data vendor revenue and market forecast 2012-2017*," Wikibon, Tech. Rep.,2014.
[9] Mayer-Schonberger V, Cukier K, "*Big data: a revolution that will transform how we live, work, and think*," Boston: Houghton Mifflin Harcourt; 2013.
[10] Chen H, Chiang RHL, Storey VC, "*Business intelligence and analytics: from big data to big impact*," 36(4): 2012; 1165-88.
[11] Kitchin R, "*The real-time city? big data and smart urbanism*," Geo J. 79(1), 2014;1-14.
[12] Fayyad UM, Piatetsky-Shapiro G, Smyth P, "*From data mining to knowledge discovery in databases*," 17(3), 1996, AI Mag.:37-54.
[13] Han J. "*Data mining: concepts and techniques,*" 2005, San Francisco: Morgan Kaufmann Publishers Inc.
[14] Agrawal R, Imieliski T, Swami A. "*Mining association rules between sets of items in large databases*," ACM SIGMOD , 1993;22(2):207-16.
[15] Witten IH, Frank E. "*Data mining: practical machine learning tools and techniques,*" 2005, San Francisco: Morgan Kaufmann Publishers Inc.
[16] Abbass H, Newton C, Sarker R., "*Data mining: a heuristic approach*," 2002, Hershey: IGI Global.
[17] Cannataro M, Congiusta A, Pugliese A, Talia D, Trun_o P., "*Distributed data mining on grids: services, tools, and applications*," IEEE Trans Syst Man Cyber Part B Cyber. 2004.
[18] Krishna K, Murty MN, "*Genetic k-means algorithm.*"IEEE Trans Syst Man Cyber Part B Cyber. 1999; 29(3):433-9.
[19] Jain AK, Murty MN, Flynn PJ, "*Data clustering: a review.*" ACM Comp Surveys. 1999; 31(3):264-323.
[20] McQueen JB, "*Some methods of classification and analysis of multivariate observations.*" In: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[21] Safavian S, Landgrebe D., "*A survey of decision tree classifier methodology*," IEEE Trans Syst Man Cyber. 1991; 21(3):660-74.
[22] McCallum A, Nigam K., "*A comparison of event models for naive bayes text classification*," In: Proceedings of the National Conference on Artificial Intelligence, 1998. pp. 41-48.
[23] Boser BE, Guyon IM, Vapnik VN. "*A training algorithm for optimal margin classifiers*," In: Proceedings of the annual workshop on Computational learning theory, 1992.
[24] Han J, Pei J, Yin Y. "*Mining frequent patterns without candidate generation*," In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000.
[25] Kaya M, Alhajj R., "*Genetic algorithm based framework for mining fuzzy association rules*," Fuzzy Sets Syst. 2005;152(3):587-601.
[26] Srikant R, Agrawal R., "*Mining sequential patterns: generalizations and performance improvements*." In: Proceedings of the International Conference on Extending Database Technology: Advances in Database Technology, 1996. pp 3-17.
[27] d'Aquin M, Jay N. "*Interpreting data mining results with linked data for learning analytics: motivation, case study and directions.*" In: Proceedings of the International Conference on Learning Analytics and Knowledge, pp 155-164.
[28] Shneiderman B, "*The eyes have it: a task by data type taxonomy for information visualizations*," In: Proceedings of the IEEE Symposium on Visual Languages, 1996, pp 336-343.
[29] Kopanakis I, Pelekis N, Karanikas H, Mavroudkis T. "*Visual techniques for the interpretation of data mining outcomes*." In: Proceedings of the Panhellenic Conference on Advances in Informatics, 2005. pp 25-35.
[30] Elkan C. "*Using the triangle inequality to accelerate k-means*." In: Proceedings of the International Conference on Machine Learning, 2003, pp 147-153.
[31] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "*BIRCH: an efficient data clustering method for very large databases.*" ACM Sigmod Record. Vol. 25. No. 2. ACM, 1996, pp 103-114.
[32] Ester M, Kriegel HP, Sander J, Xu X, "*A densitybased algorithm for discovering clusters in large spatial databases with noise,*" In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
[33] Ordonez C, Omiecinski E, "*Efficient disk-based kmeans clustering for relational databases*" IEEE Trans Knowledge Data Eng. 2004; 16(8):909-21.
[34] Mehta M, Agrawal R, Rissanen J, "*SLIQ: A fast scalable classifier for data mining*," In: Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology. 1996. pp 18-32.
[35] Mico L, Oncina J, Carrasco RC. "*A fast branch and bound nearest neighbour classifier in metric spaces.* Pattern Recognition Letter. 1996; 17(7):731-739.
[36] Djouadi A, Bouktache E. "*A fast algorithm for the nearest-neighbor classifier*," IEEE Trans Pattern Anal Mach Intel. 1997; 19(3):277-282.
[37] Ververidis D, Kotropoulos C,"*Fast and accurate sequential oating forward feature selection with the bayes classiffer applied to speech emotion recognition*," Signal Process. 2008; 88(12):2956-2970.
[38] Pei J, Han J, Mao R, "*CLOSET: an efficient algorithm for mining frequent closed itemsets*," In: Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2000. pp 21-30.
[39] Zaki MJ, Hsiao C-J. "*Efficient algorithms for mining closed itemsets and their lattice structure*", IEEE Trans Knowledge Data Eng. 2005;17(4):462-78.
[40] Chen B, Haas P, Scheuermann P. "*A new two-phase sampling based algorithm for discovering association rules,*" In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. pp 462-468.
[41] Zaki MJ, "*SPADE: an efficient algorithm for mining frequent sequences*," Mach Learn. 2001; 42(1-2):31-60.
[42] Yan X, Han J, Afshar R, "*CloSpan: mining closed sequential patterns in large datasets*," In: Proceedings of the SIAM International Conference on Data Mining, 2003.
[43] Ayres J, Flannick J, Gehrke J, Yiu T, " *Sequential Pattern Mining using a bitmap representation*," In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. pp 429-435.
[44] Masseglia F, Poncelet P, Teisseire M, "*Incremental mining of sequential patterns in large databases*," Data Knowledge Eng. 2003; 46(1):97-121.
[45] Shirkhorshidi AS, Aghabozorgi SR, Teh YW, Herawan T, "*Big data clustering: a review,*" In: Proceedings of the International Conference on Computational Science and Its Applications, 2014. pp 707-720.
[46] Tekin C, van der Schaar M, "*Distributed online big data classification using context information*," In: Proceedings of the Allerton Conference on Communication, Control, and Computing, 2013. pp 1435-1442.
[47] Lin MY, Lee PY, Hsueh SC, "*Apriori-based frequent itemset mining algorithms on mapreduce*," In: Proceedings of the International Conference on Ubiquitous Information Management and Communication, 2012. pp 76:1-76:8.
[48] Leung CS, MacKinnon R, Jiang F, "*Reducing the search space for big data mining for interesting patterns from uncertain data*," In: Proceedings of the International Congress on Big Data, 2014. pp 315-322.
[49] Yang L, Shi Z, Xu L, Liang F, Kirsh I, "*DH-TRIE frequent pattern mining on hadoop using JPA*," In: Proceedings of the International Conference on Granular. Computing, 2011.