Examination of Big Dataset using LEOS, JOSE, SVM on MapReduce

Bhagyashree Patle, Vijayarajan V

Abstract*: *Data analytics (DA) is the job of reviewing datasets in order to frame conclusions about the information they have, increasingly using specialized systems and software. As with the emergence of Big Data, data analytics was needed. The problems that we are considering are going to be in a fraud detection application. Where we'll considering major aspects such application-independent format(XML/JSON) for the clusterization process based on the no label classification algorithm where we will focusing on the clusters to enhance the oversampling process and utilize the merits of parallel computing to speed up our system. We aim to use MapReduce functionality in our application and deploy it on Amazon AWS. Datasets gathered for studies often comprise millions of records and can carry hard-to-detect concealed pitfalls. In this paper, we are working on two datasets. The first one is a medical dataset and the second one is a customer dataset. Big Data Analytics is the suggested solution in this day and age, with growing demands for analyzing huge information sets and performing the required processing on complicated data structures. The problem faced at the moment is mainly, how to store and analyze the large amount of data which is generated from heterogeneous sources like social media and what to use to make data fast accessible as well as in pocket budget. To resolve all problems Map-Reduce framework is useful-by offering an integrated technique towards machine learning, it speeds up processing. In this, we will explore the LEOS algorithm, SVM, MapReduce and JOSE algorithm, their requirements, their benefits, their disadvantages, difficulties, and their corresponding solutions.

Keywords* : *Big Data Analytics, MapReduce, LEOS, Dataset, AWS, RHEEM studio, Cluster, XML, JSON

I. INTRODUCTION

This is an age of complicated, bulky information, that is to say, Big Data. Big Data plays a powerful role in changing almost all platforms of standard data analysis. Scaling hardware platforms is the necessity of the current situation to perform any sort of big data assessment. To meet all demands within a defined time frame, selecting the right or appropriate hardware as well as a software platform to analyze Big Data is a very crucial task. Miscellaneous platforms of Big Data are present with a varied set of features along with a detailed understanding of all these platform abilities is need to select the appropriate platform. The main point is the platform's adaptability to satisfy enhanced demands for information processing, particularly when building on a particular platform-based analytics solution [1]. With this in mind, the most frequently used Platforms of big data are analyzed in detail and the strengths and weaknesses of their platforms are discussed. Volume, Speed, and accuracy are the fundamental factors for handling big data [2].

We use medical and customer data sets in this project. Data set as a data collection. What precisely does this imply? A data set is a collective form of associated information sets consisting of heterogeneous items that a computer can modify as a single entity. Usually, a data set can be a single database table or a single statistical data matrix. The set of items can be made up of a few records or millions of records. Either way, they become a collection by the reality that the objects are put together. In data mining, a data analysis technique is used which searches for data trends and patterns. We are considering Medical data sets because it has large quantities of medical information, economic information, multiple measurements, demographics of particular populations and statistical data, to name but a few, insurance data collected from various sources of health care information. This complex dataset is playing a major role in data analytics.

Big Data refers to very huge data sets(volume), very high in speed(velocity) and variety. Working on Big Data is a major challenge using standard tools, strategies and hardware/software platforms. Massive information concerns huge volumes, data sets with multiple different sources that are swiftly growing and complex. Nowadays data is rapidly growing through a multimedia application on social media. As a result of that, there is rapid evolution in networking, data collection; data storage limits in almost all biological, physical, medical and engineering domains, huge amounts of data are now expanding at a high rate. These datasets can currently be used and stored using distributed systems with massive data technology, where data elements are stored and collected in multiple areas through the programming framework [3]. The current strategy or system is very smart now, it analyses the data at the time of data generation. So it reduces the effort to store data in a big data database directly [4]. Different forms of data can currently be harnessed together with photos, messages, device data, videos, social media conversations, or voice recordings and brought back with more standard and structured information with Big Data Strategy. The paper is arranged as follows: the fundamental notion of strengths, scaling and weaknesses of scaling together with different platforms are discussed in the "Scaling" section. The Big Data can be defined as whichever dataset where capability

Revised Manuscript Received on November 18, 2019.

***** Correspondence Author

Bhagyashree Patle1, School of Computer Science and Engineering, VIT Vellore, Tamil Nadu, India. Email: bhagyashreepatle@gmail.com

Vijayarajan V, School of Computer Science and Engineering, VIT Vellore, Tamil Nadu, India. Email: virtual.viji@gmail.com

presents important trouble to gather, preserve, maintain and study conventional database products [11]. This can be used to provide references for vast, massive, complex, etc. data sets that consist of a variation of semi-structured, unstructured and structured elements that can be overly vast, overly speedy, or overly complex and can not be controlled by traditional practices.

In this system, we are using two datasets. Medical dataset and customer dataset these two datasets are using in our system. This dataset is converted into XML and JSON format. In this, we are using the JOSE algorithm for proving the security to on dataset. The map-reduce algorithm divides the data into labels and no label. LEOS algorithm based on clusterization. It forms three clusters in parallelize form. Medical and customer datasets are uploaded to test and the results generated are in fixed-size ensemble decision tree which gives us a predictors value is store into the database which has been injected with indexing functionality for fast data retrieval and Results are analyzed in the Rheem Studio.

The following algorithm used in our proposed system. We are studying this algorithm in that we conclude that this is the best algorithm than the existing system.

LEOS algorithm: LEOS (Large Ensemble with Over-Sampling) that imitates random forest algorithms along with extreme target-class over-sampling to boost target-class bias. It is based on clusterization.

JOSE Algorithm: A Standards Track document specifying a presentation of integrity-protected information using JSON (JSON Object Signing and Encryption) information structures where (but not restricted to) JSON data structures are included in the information to be protected. "Integrity protection" Includes digital signatures of the public key and MACs of the symmetric key. It provides security to the dataset.

MapReduce: MapReduce is a method for processing and a program model for java-based distributed computing. There are two significant tasks in the Map-Reduce algorithm, namely the Mapper class and the Reducer class. Mapper requires a set of information and transforms it into another set of information where tuples (key-value pairs) are broken down into individual elements. The job of a reducer is to count or make a single entry of similar items.

SVM ALGORITHM: It is a standard machine called a support vector machine that works in the association of learning algorithms that analyzes information used in the analysis of regression and classification. A system that allocate new instances to one or the other category, making it a non-probabilistic binary linear classifier (although there are techniques such as Platt scaling to use SVM in probabilistic classification). This algorithm is used to make it a non-probabilistic linear binary classifier (although there are some methods to use SVM in a probabilistic classification environment).

One of the booming concepts of the end of the 20th century was the internet, as with the awareness of the internet the data storage and data processing have increasing on the cloud. This in turn gives us a lot of inefficient data in a serialized manner which consumes time and resources. Another major concern in the data analytics area is data format for processing as the application changes the data format changes which causes the increases conversion delay and it can be easily accessed by a random user. Clusterization causes unstable sets of a decision tree which contributes to the inaccuracy of the predictor's values. Database retrieval is one factor where an end-user awaits the results from the database's traditional mechanism where the retrieval rate needs to be improved.

The objective of our system is as follow:

Extreme target class sampling to increase bias and leverages Hadoop cluster parallel computing for efficient data handling (LEOS) purposes.

Clusterization based on no label classification by finding its own structure in the input.

On-demand processing for classical data service system, providing as a standard format for the data processing system, which in turn provides an Application independent data with extended security to the file format(XML/JSON).

Analysis of Results in RHEEM STUDIO in its meta processing framework.

The organization of our paper is as follows. Section II defines the evaluation of the literature and gives various information of the previous paper. Section III presents the results. Conclusions are made in Section IV.

II. LITERATURE REVIEW

This section describes the Literature review. For studying the existing system we have studied various papers. Bo Li, et al proposed a data mining model algorithm that analyzes the shortcomings of an airline's preliminary business process. The architectural framework of the flight safety monitoring platform using big data technology was introduced and explained by the functioning of the logical structure and module structure in order to achieve the need of accuracy and avoid the lacuna previously encountered. The platform was executed by splitting the system to achieve the target as 1.information acquisition, 2.information decoding, 3.information storage, 4.information analysis and last 5.visualization [2].

Nicolas Heulot, et al introduced a close screen region for selecting and highlighting data items in the current view and other linked views. Scatterplots are typical methods for visualizing the values of two components of a multidimensional information set. They had suggested a method for brushing and interactively clustering multidimensional information via a single projection of their scatterplot[5]

Huntley Parker, et al proposed a train learning algorithm. There are several new challenges in handling big data but author focused specifically on analytical difficulties. Typically, the analytics part of the entire lifecycle uses a waterfall method – finishing one phase before the next one started. We can not backtrack in the waterfall model but in agile we can. While attempts have been made to map distinct kinds of analytics to an agile approach, the steps are usually classified as splitting up operations into a small piece of assignments whereas the regular method is compatible with the waterfall model.BDA modifies a number of operations as well as their sequence in the lifecycle of analytics. The objective of agile analytics — to achieve an optimal point between information value generation and time spent getting there. They had discussed the implications in cleaning, transformation, and analytics of an agile process for BDA [6].

Atsushi Yamada, et al suggested a clustering algorithm. The clustering provides proportional project weight to their information source number. In hierarchical clustering, a project with more data sets has a more powerful pull. They had presented a structurethat can be reuse for analytics that accelerates the use of information and analytics over the firm. The Analytics Governance Framework (AGF) is a set of directing principles to streamline the work of , and data practitioners and analytics practitioners, managers. This framework enables customer firms or organizations to optimize their amount of analytical projects to accomplish company conversion and reducethe time required foran analytical project to accomplish the business effect [7].

BertjanBroeksema, et al presented a web-based application of Big Data Visual Analytics Parallel Coordinates. The system was promoted sophisticated analytics on the server while being readily accessible through web browsers as well as density-based rendering on the client with support for accelerated hardware graphics. Parallel coordinates were accessible for prototype execution [8].

PVRD Prasada Rao, et al suggested a data mining algorithm. They had summarized their suitability regarding handling real-time implementation problems for their adaptation to Big Data analysis. Due to the executionof these systematically established and tremendously used data mining algorithms effectively, Big Data processing and analytics could predict by considering different factors like the strengths of available software frameworks and platforms. Hybrid methods (concatenation of multiple platforms) might be more suitable for a specific data mining algorithm and highly adaptable and processed in real time [9].

Daniel Cheng, et al suggested the original testing hypotheses. The extensive utilization and implementation of geo maps had given a familiar set of interactions to explore incredibly huge geo data spaces and big abstract data spaces can be applied. A tile-based visual scheme (TBVA) was created based on these methods to show standard visualization for the Twitter dataset.TBVA enables John Tukey to analyze exploratory information on efficiently infinite information sets [10].

Ryan Norman, et al told about visualization methods for efficient climate analysis.After that find the application which increases the computing capacity using Quad Zeon processor's multi-thread. The article describes the design of the system, detail description ofthe design, and explains the efficiency of computation and the outcomes of multiple methods of visualization. Through this method we can visualize the huge and complex scientific data set interactively and efficiently on the screen is main factor of this system. The computational efficiency findings of the individual visualization methods and the general scheme furnish benchmarks for different large-scale development initiatives in visualization [11].

Ryan Norman, et al suggested the rule algorithm. This algorithm is used to calculates estimated probabilities for every case that occurs in the dataset after that in binary combinations, triple combinations, etc. up to a specified probability threshold avoids more laws. For each rule, confidence was calculated, that is, the conditional likelihood of a case happening. There are different and advanced technology has been developed to store and manage the operations on heterogeneous data. These types of technology allowing the existing data to go through it and processing the real-time time as it is generating. Because of these features, the department of defense is using this methodology to correct, fast and safety decisions on the basis of available data by analyzing it. [12].

Peter Triantafillou, et al developed new models, algorithms, and compositions that learn from the system's responses to particular analyst queries. To obtain new algorithms, composition, and models from which analysts can learn, their research aims to offer a new paradigm through a Data-less Big Data Analytics approach that they had coined. The new paradigm was based on three pillars like 1. understanding analyst action, 2. understanding system working in response to analyst actions, and 3. putting them all together to guess future system behavior in response to future analyst behavior [13].

Zhihao Peng, et al introduced algorithms for machine learning (ML) with applications in many areas. With various ML tools available, it had been a challenging task to decide or to choose the tool that can efficiently as well as efficiently analyze and implemented the ML algorithms. Spark offers a versatile platform to perform several [14].

Yan Huang, et al gave a brief overview of the challenges facing the creation of remote selection applications in the cloud computing environment and offers a cloud computing solution for large image data analytics. They had presented experimental cloud computing results with space, traditional film, and UAV images. They suggested analytical algorithms for imagery into a large data environment. Huge image data can become easily unmanageable and useless without the appropriate algorithms to analyze it quickly [15].

Diego F. Rueda, et al suggested iterative algorithms could be articulated and executed as tolerant pipeline data flow. Key Performance Indicators (KPIs) are used in the mobile network operators(MNO) to assess the performance of consumers with provided quality. In their system, a customized Big Data Streaming tool was introduced to find out what type of problems are facing their customers and what is their requirement for smooth communication in #G and 4G network. All these were considered as a challenge and trying to resolve them for increasing customers and business [16].

Dan Liu, et al proposed Big Data Analysis Technology for the construction of a Big Data Analysis Platform for Intelligence Transport [17]. Jeremy Debattista, et al described an Intelligent Big Data Analytics conceptual model. This model is basically based on two methods likely semantic and machine learning AI. In this model, the raw data is collected from heterogeneous sources in the form of a result. This data is governance by thevalue-driven AI technique to extract useful information [18].The following table shows the various algorithms and results of the existing system.

We have studied different types of algorithm and some protocol in which we analyzed the LEOS, MapReduce, JOSE and SVM algorithm is a much better algorithm. In recent papers for designing a more flexible Data Analytics system, they used an algorithm like the AHP algorithm, HBase, logistic regression algorithm, machine learning algorithm, etc. The detailed analysis demonstrates that none of these methods are capable of delivering highest performance, accuracy, noise, SNR, efficiency in all uncontrolled conditions such as price, safety, low complexity of execution, redundancy.The proposed algorithm shows better results as compared to the existing one.

Table 1 Analysis of the algorithm

Sr. No.	Title	Authors	Algorithm/Methods	Results
1.	Leveraging Distributed Data over Big Data Analytics Platform for Healthcare Services[19]	MandeRamesh	Map- Reduce HBase scheduling algorithm	Big Data Analytics had converted complex data sets into a well-informed decision.
2.	Decision Support Framework for Big Data Analytics[23]	Sakshi Agarwal	AHP algorithm	It identifies appropriate choices of platforms for implementation and deployment from the different solutions provided by the big data solution.
3.	HDM-MC in-Action: A Framework for Big DataAnalytics across Multiple Clusters[24]	Dongyao Wu	logistic regression algorithm	Presented the efficacy of the optimization techniques for HDM-MC.
4.	Performance Analysis and Challenges of the Map-Reduce Framework in Big Data Analytics[27]	C. Varma	MapReduce algorithm	They had provided new methods to solve issues that had been challenging us for a long time.
5.	A Methodology of Real-Time Data Fusion for Localized Big Data Analytics[28]	S. Jabbar	Transformation from RDF to XML, Transformation from RDB to XML, Transformation from XML to RDF, from XML to RDBTransformation	It displays all the features needed to make data fusion localized and viable to support change-oriented information and metadata update.
6.	Big data Analytics Using Support Vector Machine[29]	P.Amudha	MapReduce algorithm,SVM	SVM algorithm gives the better result as compared with other algorithms in Hadoop environment using Map-reduce framework.
7.	A Hybrid approach of Deep Learning with Cognitive Particle Swarm Optimization for the Big Data Analytics[30]	S. Hegde	machine learning algorithm, particle swarm optimization technique	It enhances the precision of Big Data's feature extractions and classification assignment.
8.	Cost-Effective Cloud Server Provisioning for predictable Performance of Big Data Analytics[31]	Fei Xu	iSpot: Transient server provisioning	It gives a predictable performance in a cloud environment that provides resources at an affordable rate. With the help of spark framework and DAG.
9.	A Big Data Analytics Framework for Forecasting Rare Customer Complaints[1]	Donghui Wu	LEOS	Acquired a large No. of decision trees, it selects a subset of features randomly.

10	Stocks Analysis and Prediction Using Big Data Analytics [14]	Zhihao PENG	Machine Learning	Accurately predict the stock

III. EXISTING RESULT:

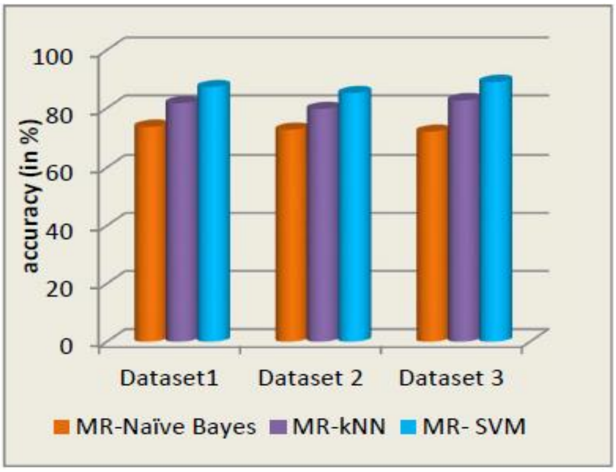
We have presented our results of a systematic review of the literature in this section. Our review results, after analyzing the literature, conclude that the Map Reduces technique, SVM Algorithm is the best algorithm for big data analytics [29].

If we compare the algorithms like MapReduce-based SVM (MR-SVM), KNN (MR-KNN), and Naive Bayes (MR-NB) with reference to the correctness, Execution duration, and error rate then from the below table2 we can say that

The SVM (MR-SVM) algorithm is better than another algorithm. The following table shows the result of the different algorithms implemented on tree different datasets for comparison purposes. Dataset1, dataset2, and dataset3 contain diabetic dataset of 8000, 10,000 and 12000 instances respectively. Dataset is the set of different categorical value which is then converted in numerical value for comparison purposes [29].

Table 2 Comparison of accuracy rate

	Accuracy in %	Accuracy in %	Accuracy in %
Algorithm	**Dataset 1**	**Dataset 2**	**Dataset 3**
MP-Naïve Bayes	72%	72%	72%
MR-KNN	80%	78%	80%
MR-SVM	88%	82%	88%

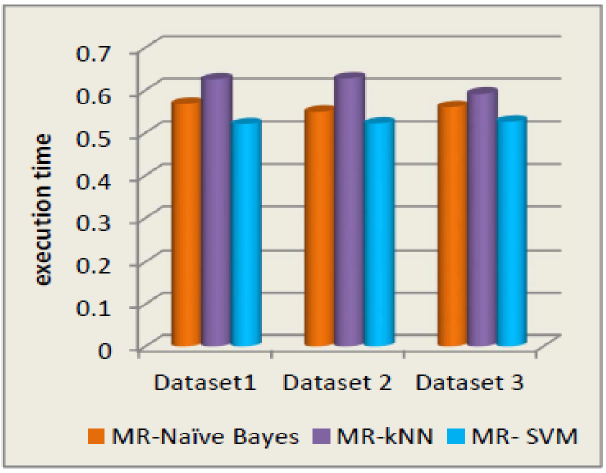


Graph 1 Graphical representation of accuracy rate

Graph 1 shows the accuracy rate given by classifiers. From the above graph, we can say that the MR-SVM classification technique gives 87% correct classification whereas MR-NB gives 73% and MR-KNN technique gives 81% for the dataset-1. Similarly, MR-SVM classifier gives 85% and 89% accuracy in classification for dataset-2 and dataset-3 respectively. Graph1 gives a clear result that SVM classifier generates comparatively better result than another classifier for any size of datasets[29].

Table 3 Comparison of execution duration

	Execution duration	Execution duration	Execution duration
Algorithm	**Dataset- 1**	**Dataset- 2**	**Dataset- 3**
MP- Naïve Bayes	0.55	0.52	0.55
MR-KNN	0.6	0.6	0.58
MR- SVM	0.5	0.5	0.5



Graph 2 Graphical representation of execution duration

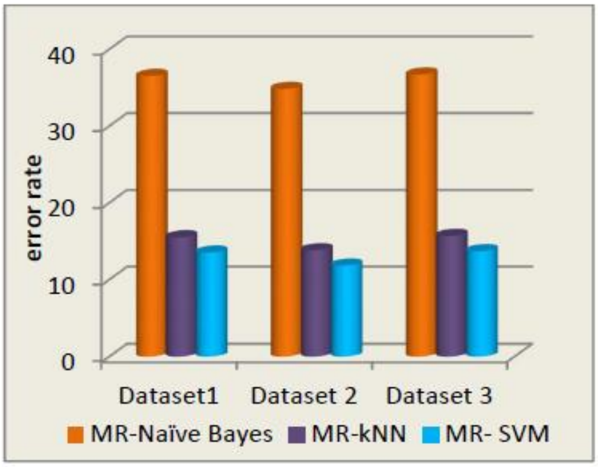
Graph 2 shows the execution duration given by 3 classifiers.Execution duration is reduced in every dataset if we compare SVM with another classifier technique.From the above graph2, we can say that SVM classifier generates comparatively better result than another classifier for any size of datasets of diabetic diseases. When analyzing the classifier with respect to error rate factor as shown in graph 3, the SVM classification strategy results in a minimum error rate compared to other classifiers [29].

Table 4 Comparison of Error rate

	Error rate	Error rate	Error rate
Algorithm	**Dataset- 1**	**Dataset- 2**	**Dataset- 3**
MR-Naïve Bayes	36	34	35
MR-KNN	15	12	15
MR- SVM	12	10	12

The experimental result demonstrated the suggested MapReduce classifier achieves greater efficiency with reference to enhanced precision, Execution duration and error rate relative to other algorithms.

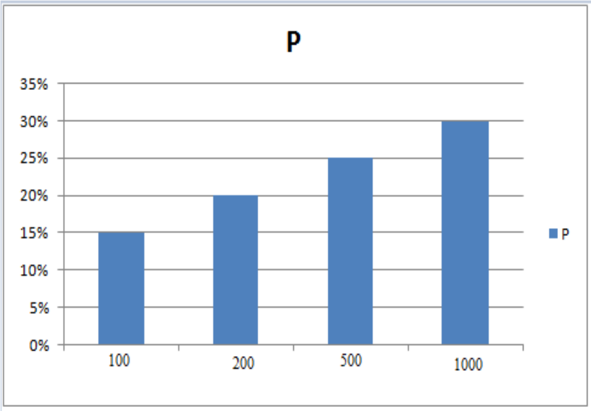
Donghui Wu [1] suggested the LEOS algorithm. MapReduce and parallelization packages were used to implement the LEOS. LEOS algorithm is applied to the different application with a different and huge amount of training set to check the complexity. Table 5 shows the parameters used in the grid search using LEOS. The training data subsets are generated according to the parameters provided:
Where n: subset size,
P: pseudo prevalence rate,
Dn: threshold



Graph 3 Graphical representation of error rate

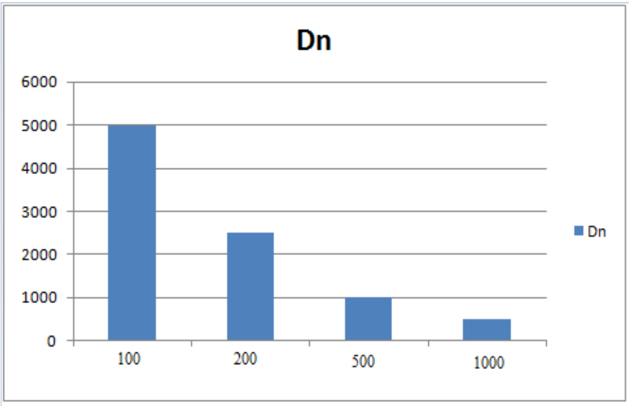
Table 5 The parameters for grid search

subset size(**n**)	pseudo revalence rate(**P**)	threshold (**Dn**)
100	15%	5000
200	**20%**	2500
500	25%	**1000**
1000	**30%**	500



Graph 4 Graphical representation of subset size and pseudo prevalence rate

The experiments showed that LEOS is the comparatively better solution for the predictions with a training subset size of 500, a pseudo prevalence rate of 20% and a minimum ensemble size of 1000 decision trees[1].



Graph 5 Graphical representation of subset size and threshold

IV. CONCLUSION

This paper's objective is to perform more flexible Data Analytics system for enhancing the Volume, security, Clusters, space and time complexity which are going to be achieved by Enhancing 1) LEOS Algorithm 2)Map Reduce Mechanism 3)SVM Algorithm 4)JOSE mechanism and eliminate major the current drawbacks in Existing System. With the literature review, we found 17 relevant papers that covered the goal we wanted to achieve in an initial analysis.This evaluation allowed us to recognize that most articles focus on classifying text and classifying images. Surprisingly, however, we found no study papers on data analytics or information classification in XML and JSON format.For this reason, more research is needed to check how Big Data Analytics can be used effectively for predictor value analysis and how Rheem Studio analyzes results.The findings of the experiment determined that the proposed Map Reduce Mechanism and SVM Algorithm relatively outperform other algorithms/techniques in terms of enhanced correctness, Execution duration, and minimum error rate.

REFERENCES

1. DonghuiWuet al,“A Big Data Analytics Framework for Forecasting Rare Customer Complaints”,IEEE,pp: 3965 - 3967,2017
2. Bo Li et al. "Big Data Analytics Platform for Flight Safety Monitoring" Big Data Technologies and Applications. Springer International Publishing, 2016. pp: 13-52
3. Elgendy et al "Big Data analytics: a literature review paper"'Industrial Conference on Data Mining', Springer, 2014.
4. Wu, X. et al "Data mining with Big Data," IEEE transactions on knowledge and data engineering, 2014.
5. Nicolas Heulot et al ” A Multidimensional Brush for Scatterplot Data Analytics”,IEEE,2014.
6. Huntley Parker et al “Agile Big Data Analytics”IEEE,2017.
7. Atsushi Yamada et al,” Governance Framework for Enterprise Analytics and Data”, IEEE, 2017.
8. BertjanBroeksema et al,” Big Data Visual Analytics with Parallel Coordinates”,IEEE,2015.
9. PVRD Prasada Rao et al,” Platforms for Big Data Analytics: Trend towards Hybrid Era”, ICECDS,2017.
10. Daniel Cheng et al,” Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis”,IEEE,2013.
11. Pak Chung Wong et al,” Visual Analytics of Large-Scale Climate Model Data”,IEEE, 2014
12. Ryan Norman et al,“Using Big Data Analytics to Create a Predictive Model for Joint Strike Fighter”,IEEE,
pp: 3590 – 3596,2018

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

13. Peter Triantafillou et al,“Data-Less Big Data Analytics (Towards Intelligent Data Analytics Systems)”,IEEE, pp: 1666 – 1667,2018
14. Zhihao Peng et al,“Stocks Analysis and Prediction Using Big Data Analytics”,ICITBS, pp: 309 – 312,2019
15. Yan Huang et al,“A CLOUD COMPUTING SOLUTION FOR BIG IMAGERY DATA ANALYTICS”,IEEE, pp: 1-4,2018
16. Diego F. Rueda et al,“Big Data Streaming Analytics for QoE Monitoring in Mobile Networks: A Practical Approach”,IEEE, pp: 1992 – 1997,2018
17. Dan Liu et al,“Big Data Analytics Architecture for Internet-of-Vehicles Based on the Spark”,ICITBS, pp: 13-16, 2018
18. Jeremy Debattista et al,“Semantic Data Ingestion for Intelligent, Value-Driven Big Data Analytics”,IEEE, pp: 1 - 8, 2018
19. Ramesh Mande et al,“Leveraging Distributed Data over Big Data Analytics Platform for Healthcare Services”,ICOEI, pp: 1115 - 1119,2018

AUTHORS PROFILE

Bhagyashree Patle received her B.E., Computer Science and Engineering from RTM Nagpur University in 2011, and M.Tech., Computer Engineering from VJTI,Mumbai in 2014. She is currently working as Assistant Professor in Department of Computer Engineering, SKNSITS, Lonavala(Maharashtra) and pursuing Ph.D. in School of Computer Science and Engineering, VIT Vellore. Her current research interests are Data Mining, Big data Analytics, Machine Learning.

Vijayarajan V has completed his B.E.(CSE)from Madras University, and M.E.(CSE) from Anna University and PhD in School of Computer Science and Engineering, VIT, Vellore. Currently, he is working as Associate Professor in School of Computer Science and Engineering, VIT Vellore. He has published more than 30 articles in various International Conferences and Journals. His research interests are Information Retrieval, Image Processing, Machine Learning.