IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

A General Perspective of Big Data Analytics: Algorithms, Tools and Techniques

P. Pandeeswary[1*], M. Janaki[2.]

[1,2]Department of Computer Science, Umayal Ramanathan College for Women, Karaikudi.

[*]*Corresponding Author: pandeeswary11133@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i7.129137 | Available online at: www.ijcseonline.org

Accepted: 19/Jul/2019, Published: 31/Jul/2019

***Abstract*—** Big data is the term representing any collection of datasets so large and complex which is difficult to process using traditional data processing applications. The challenges comprise of analysis, capture, search, sharing, storage, transfer, visualization, and privacy violations. Big data is a set of techniques and technologies that need new forms of integration to uncover large hidden values from large datasets which is diverse, complex, and of a massive scale. Big data environment is used to acquire, organize and analyze a variety of data. The main objective of this paper is to give a general perspective of big data analytics, its process, tools and techniques used. There is an immense need for the construction of algorithms to handle Big Data. Many algorithms are defined in the analysis of large data set. A review of various techniques and algorithms are also discussed in this paper. The massive volume of both structured and unstructured data which is so large, it is difficult to gather and analyze for getting the required solution. It is better to have some tools which help in processing the complex data sets. This paper is also focused on various tools available to extract required data from big data.

***Keywords*—** *Big data, Data extraction, Data cleansing, Decision making, Visualization, Predictive analysis*

I. INTRODUCTION

Big data is a rapidly enlarged research area spanning the fields of computer science and information management, and has become a ubiquitous term in understanding and solving complex problems in different fields such as in, applied mathematics, medicine, engineering, computational biology, healthcare analysis, social networks, finance, business management, government, education system , transportation and communication. Big data is a collection of Datasets so large and complex that it becomes difficult to process using on-hand Database Management Tools or Traditional Data processing applications [1]. It's a technique which discovers value from the diverse and massive scale of data sets. This term is loose because what is considered big data today will be considered very small in the future because of the continuous development of the storage methods [2]. The Data might be Structured, Unstructured and Semi-Structured. Big Data faces a series while exploring Big Data sets, extracting value and knowledge from such mines of information. The difficulties that are found at different levels include data capture, storage, searching, sharing, analysis, management, and visualization. The challenges are with Big Data management for ensuring a high level of data quality and accessibility. Big Data cleaning which detects and correct the inaccurate records from a recordset. In Big Data aggregation, the data is widely searched gathered and presented as a record. Analytical challenges may be at search, sharing, transfer, visualization, querying, updating, information privacy and data source. Privacy and security limitation is the ability to determine which data can be presented publicly and privately, data should not be accessed or shared by the third parties. Misinterpretation of Data means redundant or missing information within datasets which leads to false results. Such that inherent potentials are excavated to improve Decision making and obtain further advantages Data Generation, Data Acquisition, Data Storage, Data Analytics are four modules of Big Data value chain. Data Generation is the collection of data, being routed and procedures by which data reach a database.

Data Acquisition is a Process of gathering, filtering and cleaning the data before it is put into a Storage solution. Acquisition of Big Data is controlled by the 10 Vs Data Storage is a Storage infrastructure designed specifically to store, manage and retrieve the massive amount of data. It allows data to be stored in a manner where they can be easily accessed and used. It is Flexible and Scalable as required. Data Analytics is a complex process to inspect large and varied Data Sets to expose hidden patterns and unknown Correlations. The foremost objective of Big Data Analysis is to process data of high Volume, Velocity, Variety, and Veracity. Using various intelligent techniques

also it has other six Vs, so Big Data has 10 Vs of characteristics. They are: Volume refers to the vast amount of Data Provoked every day. Velocity is a measure of growth and how fast the data are converged for being analyzed. Variety contributes information about types of data as Structured, unstructured and semi-structured. Veracity covers availability and liability. Variability is Big data often models various data sources. Also referred to as inconsistent speed where big data is loaded into the database. Validity shows how accurate and correct the data is for the intended use. It gives a valid prediction. The vulnerability has got along with its new security concerns. It means anything that leaves information security exposed to a threat. Volatility specifies how long the data is valid and determines how long it can be stored. Visualization hard part of Big Data that makes overall huge data easy to understand and to visualize, pattern, trend and correlate which are undetected in text-based data can be exposed and recognized easily with visualization. Big data has a life cycle which is creation, pre-processing and output [3]The value measures the value of data, it is one of the boundless processes without end it may be structured or unstructured. Data continues to provide ever-increasing data which are available and new techniques are introduced. An ML issue is referred to as the issue of learning from past experience with respect to some tasks and performance measure[4]. The further recent trend of big data analytics technology has been towards the use of cloud in conjunction with data. Big data analytics is a significant function of big data, which discovers unobserved patterns and relationships among various items and people interest on a specific item from the huge data set [5].In this paper section I is the introduction about big data .section II contains the literature survey, section III gives the overall processing in big data analytics, section IV is about various techniques used in big data analytics, section V is about the tools available in big data, section VI gives a brief explanation about the algorithms in Big data analytics and the last section VII is the final conclusion about the paper.

II. LITERATURE SURVEY

A Rytsarev1, A V Kupriyanov1, discussed an approach that the social media clustering based on class annotation, with Big Data technology – a modern and effective tool to handle the described difficulties. And collected a huge sample of images from real profiles of Twitter users to carry out computational experiments. Based on Google Net and k-Means clustering they represented the technique for clustering. Further research will be aimed at a more detailed analysis of media content from social networks with the use of the developed algorithms [6]. Ajay Kumar Pal and Saurabh Pal concluded that they have got their objective which is to evaluate the performance of a student by the three selected classification algorithms based on Weka. These results imply that among the machine learning algorithm tested, ID3 classifier has the potential to significantly improve the conventional classification methods for use in performance [7].

Brijesh Kumar Baradwaj and Saurabh Pal discussed classification task is used on student database to predict the student's decision on the basis of the previous database. As there are many approaches that are used for data classification, and here the decision tree method is used. This study will also work to categorize those students who need special attention to reduce fail ration and taking appropriate action for the next semester examination [8] Vinayak A. Bharadi1, Prachi P. Abhyankar2, Ravina S. Patil3, Sonal S. Patade4, Tejaswini U. Nate5, Anaya M. Joshi6 discussed certain data mining algorithms which are adapted to cluster the data that shows relevance with desired attributes. K-means clustering algorithm is adopted as the base algorithm. DBSCAN and EM algorithms were also applied to data. And their future work is aimed at applying advanced mining techniques to a larger dataset such as one of the big data techniques Map-Reduce [9].

Senthil Vadivu, Dr. Prasanna Devi, Vishnu Kiran, and Manivannan analyzed the attributes of farmland and farming methods and applied predictive analytics methodology over vast data set to forecast the outcome of the commodity. The purpose of this study is to collect information on the current and foreseen land-use practices with particular emphasis on the role of cropping/farming systems to predict the outcome of the crop quality (Rice) [10]. Ashwani Kumar Kushwaha and Sweta Bhattacharya proposed how to make agriculture well organized by predicting and thus improve the crop yields by using soil information. They introduced a new Agro algorithm which is used to predict the suitability of a crop for a particular soil type and enhances the overall quality of agricultural production. This also helps the farmers to choose a particular crop to sow depending on the climatic condition and provides necessary information to select the best weather to do quality farming. They used big data using Hadoop platform which helps to deal with a large number of datasets in the agricultural domain [11].

III. PROCESS OF BIG DATA

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of finding useful information, informing the conclusions, and then supporting decision-making. Data analysis has many surface and approaches, encircling diverse techniques in a variety of names while being used in different business, science, and social science domains. Data analysis collects raw data and converts it into information useful for decision-making by users. Data are gathered and analyzed to respond to questions, test

hypotheses or disprove theories. It has several phases of' data requirements, data collection, data processing data cleaning, and exploratory data analysis standard for big data analytics.

A. Extraction of Information and Integration in Big Data

Extraction is a process of acquiring knowledge from structured and unstructured data sources. Integrating the information from several sources and extracting the meaningful and relevant information from massive data is a significant challenge. The knowledge obtained must be ina machine-readable and machine-interpretable format which must give a conclusion. There are several stages in extraction; they are data selection and gathering, data cleaning, integration and storage, feature extraction, knowledge extraction, visualization.

B. Analysis of Information in Big Data

It involves the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. The analysis process has certain challenges like storage medium diversity of data and designing machines which improves efficiency and scalability, computational complexities, visualization of data, security threats. It has certain tools such as Apache, Hadoop, and Map Reduce, ApacheMahout, Spark, Dryad, Storm, Apache Drill, Jasper soft and Spunk. It is categorized as descriptive analysis, predictive analysis, and prescriptive analysis. Descriptive Analysis utilizes the historical data to learn from the past and helps to influence it in future outcomes. Predictive Analysis focuses on predicting future possibilities and trends from historical data. From current data, it will predict what happens next or in the future. It's associated with business application. Prescriptive Analysis addresses decision making and efficiency. It's related to both description and Predictive Analysis.

C. Decision Making in Big Data

Information is the key factor of success, which has an impact on performance and decision making, precisely saying it is in the quality of making decisions. It includes structured, semi-structured and unstructured real-time data, constituting of a data warehouse, OLAP, ETL, and information Business firms and academicians have designed some unique ways of tapping value from big data. There is a considerable scope of using large datasets as an additional input in making decisions. Information and knowledge are the most valuable assets for most of the organizations in decision-making processes. It needs a medium to process data into information, loaded with value and relevance for use in organizational processes. Information Systems (IS) represent these media. Decision-making ability was enhanced by various tools and techniques. The environment changes dynamically, with the convergence of IT, telecommunications and electronic media, and has a large impact on the decision-making processes of enterprises. Decision-making is accompanied by a massive amount of data and software for their analysis.

IV. TECHNIQUES USED IN BIG DATA ANALYTICS

There is a number of techniques used in big data analytics. Big data is an application of specialized techniques and technologies that process very large data sets. Data sets are often so large and difficult that it becomes complicated to process using on-hand database management tools. By technique simply we mean some method to accomplish a goal or a task. The goal is to adequately store, access, and process a massive amount of information which constitute big data. Indeed researchers continue to develop new techniques and improve from the existing ones, particularly in response to the need to analyze new combinations of data.

A. Association Rule Learning

The Association Rule Mining is a popular approach which is used for analyzing the given dataset to discover the new interesting patterns or relationships between the various items in the dataset. The problem with finding a relation between items is generally termed as market basket analysis. In that problem, the presence of items within baskets is distinguished so that the customers buying habits can be analyzed. This technique is used in inventory management, sales promotion, etc. The discovery of association rules is primarily depending on detecting the frequent sets.

B. Classification Tree Analysis

Classification is a machine learning technique. Classification tree analysis is when the predicted outcome is the class to which the data belongs. Classification techniques are capable of processing a large volume of data. It can predict categorical class labels and classifies the data based on the training set and class labels thus it can be used for classifying newly available data.

C. Genetic Algorithms

A genetic algorithm is a heuristic search method used in Artificial intelligence and computing process. It is used for finding the most optimized solutions in search problems based on the theory of natural selection and evolutionary biology. It's excellent in handling search through large and complex data sets.

D. Machine Learning

Machine learning is one of the applications of Artificial Intelligence (AI). AI is the ability to learn automatically and improve from its experience without being explicitly programmed. It focuses on the development of computer programs which can access data and make use of it by learning themselves. It is used in the large volume of

data sets to discover useful hidden patterns and other useful information like customer choices, market trends that can help organizations.

E. Regression Analysis

Regression analysis is a statistical technique used to examine relationships between variables. It is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of their relationships. From the existing data, the regression model analyzes it and then builds its knowledge base. A data analysis technique allows answering questions about associations between.

F. Sentiment Analysis

Sentiment analysis is a process of using text analytics to my various sources of data for their opinions. It determines whether a piece of writing is positive, negative or neutral. Often, sentiment analysis is done on the data which is collected from the source of the internet and from various social media platforms. It has three classification levels; they are i) document level ii) Sentence level and iii) Aspect level. It is also said to be opinion mining or Subjective analysis or opinion mining. It refers to the natural language processing and text analysis to identify and extract useful information from the large datasets.

G. Social Network Analysis

Social Network Analysis is a process of quantitative and qualitative analysis of a social network. The evolution of the social networking concept created a form of highly complex and graph-based networks with many types of nodes and ties. It characterizes the structures in terms of node and the ties, edges, or links that connect them.

V. TOOLS USED FOR DATA SCIENCE

Though Big Data principles and approaches are repeatedly discussed, there are not many technologies which are suitable to deal with such data. Due to the definitions of the volume and the velocity, the tools which are supposed to deal among big data have to suggest a distributed computing approach. There are subsequent approaches: multiple data and a single program, and then single data and multiple programs. In the first case, a single program, which is run on more nodes, where all nodes processing different data. With the contrary, the second case is considered as having only one dataset, which is processed by a program separated on small tasks that are run on different nodes in parallel. Due to it, there are some tools that try to abstract from the physical distribution as much as probable. Hadoop as the first publicly known and discussed the technology of Big Data processing has been used as the base of open source and commercial extensions. Further, it is referred to as; most of the set of big data tools are based on the Hadoop solution.

A. Hadoop

Apache Hadoop is one of the most well-known and widely used tools in the big data industry with its enormous competence of large-scale processing data. Itis a 100% open-source framework which runs on service hardware in an already existing data center. Further, it can run on a cloud infrastructure. It consists of four parts such as Hadoop distributed file system,map-reduce, YARN, and libraries. Hadoop framework runs in parallel base on a cluster and has the capability to allow us to process data across all the nodes. Also, it replicates the data in a cluster hence providing the high availability.

B. Apache Spark

ApacheSpark open-source framework for big data processing, it's built around speed, ease of use, and sophisticated analytics. Apache Spark is the next developing tool among tools in the industry of big data. It fills up the gaps of Apache Hadoop regarding data processing. Amusingly, Spark can switch both batch data and real-time data. It can run jobs 100 times faster than Hadoop's Map Reduce. As Spark does in-memory data processing, than traditional disk processing it processes data much faster. In fact, this is a plus point for data analysts to handle certain types of data to achieve a faster outcome. Apache Spark is stretchy to work with HDFS as well as with other data stores. It's also relatively easy to run Spark on a single local system to make improvement and testing easier. Distributed task transmissions, scheduling, I/O functionality are facilitated in apache spark.

C. Apache Storm

Apache Storm is a free, open-source computation system in big data. It offers distributed real-time, fault-tolerant processing system. With real-time computation capabilities, Apache Storm is a distributed synchronized framework for reliably processing the boundless data stream. In Apache Storm, each unit of data will be processed at least once or exactly once. The framework chains any programming language. It follows the "fail fast, auto restart" approach. It is written in Closure and runs on the JVM. Storm topologies can be considered as related to Map Reduce job. Conversely, in case of Storm instead of batch data processing, it is real-time stream data processing. Based on the topology constitution, Storm scheduler distributes the workloads to Nodes.

D. Cassandra

A database which is widely used today to offer effective management of huge amounts of data that's across the servers. A best big data tool that primarily processes structured data sets. It provides vastly available service with no single point of failure. Furthermore, it has certain capabilities which are not provided by other relational database and any SQL database. Its capabilities are

continuous availability as a data source, linear scalable performance, simple operations, scalability, and performance. Apache Cassandra architecture does not pursue master-slave architecture, and all nodes play the same role. It assists in replicating across multiple data centers by providing lower latency for users. Thus, adding a new node is no matter in the existing cluster even at its uptime.

E. Rapid Miner

Rapid Miner is used for data preparing, machine learning, deep learning model deployment, and prototyping. It offers a suite of products to build new data mining processes and setup predictive analysis. It has certain features on big data predictive analytics, remote analysis processing, filtering the data, merging it, joining and aggregating the data, it builds trains and validates predictive models, also reports and triggers the notifications. The server is located on-premise, or in a cloud infrastructure follows a client/server model. It provides a GUI to design and execute workflows and it's written in java. It provides 99% of an advanced analytical solution.

F. Mongo DB

Mongo DB is an open-source No SQL database which is cross-platform companionable with many built-in features. It is best for a business that requires fast and real-time data for instant decisions. Also ideal for the users who desire data-driven experiences. It runs in the MEAN software stack, NET applications and, Java platform. Some prominent features of Mongo DB are:

- It stores any type of data like integer, string, array, object, Boolean, date, etc.
- It provides a flexible cloud-based infrastructure.
- It is flexible and data across the servers are easily partitioned in a cloud structure.
- Dynamic schema is used in Mongo DB.
- Therefore, data can be prepared on the fly and quickly .which is another way of cost-saving.

G. R Programming Tool

This is one of the extensively used open-source big data tool in the big data industry for statistical analysis of data. The most affirmative part of this big data tool is – although used for statistical analysis, as a user we need not be a statistical expert. R has its own individual public library CRAN (Comprehensive R Archive Network) which consist of more than 9000 modules and algorithms used for statistical analysis of data. It runs on Windows and Linux server as well inside SQL server. R also supports Hadoop and Spark. Using the R tool one can work on distinct data and try to give out a new analytical algorithm for analysis. R is a portable language. Therefore, an R model is built and the tested data on a local data source can be easily implemented in other servers or even against Hadoop data.

H. Neo4j

Hadoop may not be a wise alternative for all bigdata-related problems. For example, when you need to deal with a huge volume of network data or graph associated issue like social networking or demographic pattern, a graph database may be an ideal choice.Neo4j is one of the big data tools that is broadly used graph database in the big data industry. It follows the essential structure of the graph database which is interconnected node-relationship of data. In storing the Data it maintains a key-value pattern. It is known for its high availability, scalability, and reliability. Significant features of Neo4j are,

- It supports the ACID transaction and query language for graphs which is generally known as Cipher.
- It is flexible because it does not require a schema or data type to store data.
- It can integrate with other databases.

I. Apache Flink

Apache Flink is a framework with the distributed processing engine for stateful computations above *unbounded and bounded* data streams. It is designed in the way such that they have to run in *all general cluster environments*, by executing computations at *in-memory speed* and with *any scale.* **The Unbounded streams** don't have a defined end it has the start only. They do not terminate until providing the data as it is generated. It must be processed endlessly, i.e., after ingesting the events must be handled rapidly. It is not possible to wait for all input data to arrive because the input is unbounded and will not be complete at any point in time. **Bounded streams** have both defined start and end. Before performing any computations it can be processed by injecting all data. Bounded stream processing is also referred to as batch processing.

J. HPCC

The High-Performance Computing Cluster (HPCC) is another best big data tool. HPCC is one of the competitors of Hadoop in the big data market. It provides a single platform, with single architecture and in single programming language for data processing. An open-source big data tools in the Apache 2.0 license. Some of the interior features of HPCC are, it accomplishes big data tasks in a highly efficient manner with far less code, offers enhanced scalability, performance, high redundancy and availability, for parallel processing it optimizes the code automatically, big data management supports an end to end workflow. It is extensible and highly optimized. It helps in building graphical execution plans.

VI. ALGORITHMS USED FOR BIG DATA ANALYTICS

Big data is data so large that it does not fit in the main memory of a single machine, and require to process big data

by efficient algorithms arise during internet searching, monitoring network traffic, machine learning, scientific computing, processing signal, and several other areas. There are varieties of Machine Learning and data mining algorithms are available for creating valuable analytic platforms. Conventional goals will determine which algorithms are used to sort out and process the information available. Various algorithms have been developed to deal distinctively with business problems. Further algorithms were designed to augment the current existing algorithm or to perform in new ways. Algorithm models acquire different shapes, which depends on their purpose. Using different algorithms to provide comparisons can suggest some astonishing results about the data being used. Unsupervised clustering algorithms are used to discover relationships Within an Organization's Dataset.

A. K-Means Clustering Algorithm

It is one of the unsupervised learning algorithm, which is used when data is unlabeled (i.e.. groups or clusters which has undefined data). The goal of the Algorithm is to discover groups in data, with the number of groups represented by variable k. The aim of K means algorithm is to divide M points within N dimensions into K clusters so that the precision rate and the recall rate are maximum [12]. Based on the similarity Data points are Clustered. It works iteratively based on provided features to allocate each data point to one of the Data points to one of the k-groups. Results of clustering are centroids of k-cluster which is used to label new data. Training data for labels (similar Data points are assigned to the single cluster). Based on centroids the cluster is formed. Two clusters of data are formed on applying algorithm. Clusters and their centroids are attributes. Also, it aims the k-means algorithm to minimize the squared error function,

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} d\left(\overline{x}_i, \overline{c}_j\right)^2$$

Benefits obtained by using algorithm are: Easy to implement and produce a higher cluster. It produces higher clustering even though it has a large number of variables. It can change between cluster when centroids are recomputed. And also have some drawbacks: predicting cluster count is difficult, error at the initial stage has an impact on the final results of data points also have an impact on the final result, sealing is sensitive.

B. Linear Regression Algorithm

It is a statistical Analytics which finds a relationship between the variables. Regression is a way to initiate a relationship between the independent and the dependent variables. Dependant has continuous waves while the Independent is either continuous or categorized values. A regression model which have more than one predictor variable is called the Multiple Regression Model. Multiple Linear Regression technique is applied to existing data. The results so obtained are verified and analyzed using the data mining technique that is Density-based clustering technique [13]. LR is about the most basic econometric method in the research field of prediction and can be divided generally into univariate regression and multivariate regression [14]. This algorithm reduces space complexity and simple to understand, good interoperability. But it's prone to outliers, multi co-linearity must be avoided before building, they are not really distributed.

C. Logistic Algorithm

This is one of the statistical analysis methods. By analyzing previous observation in data it predicts the dependent data variable. Prediction is done by analyzing the relationship between the independent variables. Logistic regression is used to predict the discrete outcomes based on variables which can be discrete, continuous or mixed. It finds the relationship between one dependent binary variable and one or more independent variables. Every independent variable is multiplied with weights and summed up. This result will add to the sigmoid function to find the result between 0 and 1. The values above 0.5 are considered as 1, and the values below 0.5 are considered as 0[15]. Hence optimization techniques are used to find the best regression coefficients and weights. Thus, when the dependent variable has two or more discrete outcomes, the logistic regression is a commonly used technique. The outcome might be in the form of Yes / No, 1 / 0, True / False, High/Low, given a set of independent variables. It is important to discover the best weights or regression coefficients. It reduces space complexity and its simplest one, good interoperability, feature importance is generated. But they are not distributed really; multi-co linearity must be avoided before building, prone to outliers.

D. C4.5 Algorithm

It is used to generate a tree. It's developed by Ross Quinn in 1993. It's an extension of the ID3 algorithm. It's output from that. It has continuous attributes which are scanned to find distinct value. It's repeatedly done for every attribute. The main features of the C4.5 algorithm are as follows- C4.5 algorithm can be interpreted effortlessly. C4.5 algorithm is very simple to implement. C4.5 algorithm uses both discrete and continuous values. Some restrictions of the C4.5 algorithm are as follows- Little deviation in data can lead to a different decision tree. It works well with large datasets. One should not apply the C4.5 algorithm to a small dataset on it. C4.5 and QUEST have the finest combinations of error rate and speed, but the C4.5 tends to generate trees with twice as many leaves as those from Quest. QUEST is a binary-split decision tree algorithm for classification and regression[16]. It is used to form decision tree, a decision tree is a methodology and prediction that is a very strongest

and famous method of decision tree transforming a very big fact into decision tree representing rule, the rule can be easily understood with natural language, can also be expressed [17]. C4.5 Decision tree algorithms are well known for rules construction and commonly applied in the machine learning process [18]. It's Continuous and discrete attributes are handled, missing attributes are not calculated and removes unnecessary branches. In some cases empty branches are found, Susceptible to noise and overfitting is found. Generate small decision tree rules quickly. Here the entire dataset is tested to create the decision tree. Easily indicates the essential attributes for classification. Also uses leaf nodes to replace the unwanted branches. It accepts both continuous domain and discrete domain values.

E. Support Vector Machine

It is one of the supervised learning algorithms, generally used in classification and regression challenges. It's used to classify two datasets; support vector and hyperplane. Support vector is Data points found around the Hyperplane. The hyperplane is a line which supports Datasets. The major work of the SVM algorithm is to find which Dataset the new Data we add belongs to. Support Vector Machine (SVM) is playing a decisive role and it provides various techniques which are well suited to obtain results in an efficient way also with a good quality level. The SVM is a generalization of the separating hyperplane classifier. This generalization combines the notions of the optimal separating hyperplanes, soft margins, and the enlargement of the input attribute space with a nonlinear mapping to a feature space [19]. The classification capabilities of traditional SVMs can be significantly enhanced through the transformation of the original feature space into a feature space of a higher dimension by using the "kernel trick". It has good accuracy [20]. Once the hyperplanes are constructed, it classifies the new examples according to the previously specified decision boundaries [21]. The model produced by SVC depends only on the training data because the factor of the cost of the model building does not care about training points that lie outside the margin [22]. with small Dataset works Better and efficient because of training points. But it's not used with large Datasets because it creates high latency and less effective in overlapping classes.

F. Apriori Algorithm

It uses a bottom-up approach. It's mostly used in Database which contains transaction records or records containing many fields. It is mostly used to find association rules. It contains repeated steps. It terminates when no more extensions are found. The Apriori algorithm is a data mining algorithm and it is used for the suggestion of frequently purchased itemsets [23]. This algorithm can handle large datasets, parallelization, and implementation are the simplest one. Its robustness and interpretability make it possible to obtain reliable results that can be interpreted by non-technical personnel [24]. Since it has some drawbacks the scanning rate of the database is high and presumes that memory is the transactional database.

G. Association Rule Algorithm

In the case of big data having large volumes of data that makes it impossible to generate rules at a faster rate. By making the use of parallel execution, with the map-Reduce framework in Hadoop, the rules can be generated much faster and also in an efficient way. Association rules mining could be limited to the problem of finding large itemsets, where a large item set is a collection of items existing in a database transactions equal to or greater than the support threshold[25]. In 1998, Liu et al connected association rule mining (ARM) and classification rule mining to give rise to the task known as Associative Classification (AC) [26]. The concept of strong association rules was first used by Agarwal et al to identify the various association rules between the items that are sold during a large scale transaction database collected from a supermarket using a point system [27]. Association rule mining finds frequent patterns, associations, correlations, or causal structures among various sets of items or objects in the transaction databases, relational databases, and other information repositories.

H. EM (Expectation-Maximization)

The Expectation-Maximization (EM) algorithm is a tool which is used to compute a maximum likelihood estimation of incomplete data, with latent or unknown variables. A clustering algorithm used in knowledge discovery. The GMM is one of the cluster algorithms and it classifies data by describing them using multiple Gauss distributions [28]. It is usually used with the Expectation-Maximization (EM) algorithm which optimizes the parameters of the GMM. It uses the clustering method for predicting data models that can be used in other statistical analysis methods.

I. Ada boost

Ada Boost is an algorithm which constructs a classifier and then boosts. It means that among a number of machines it looks for the best learning algorithm and then chooses the most effective one, and then it propagates the improved information to the other machines. With this manner, it helps to optimize the ability to learn about participating machines. It starts from the weak learning algorithm and learns frequently to attain a series of weak classifiers, and then it combines these weak classifiers to form a strong classifier. The Stage-wise Additive Modelling using a Multiclass Exponential loss function (SAME) is a multi-class Ada Boost algorithm that fits a classifier to a training set then compute the generalization error using an equation similar to Eq.1 [29] The Ada Boost algorithm adaptively adjusts the weak error classifiers to allow for superior accuracy compared to other powerful classifiers, with much

less tuning or setup [30]. Ada Boost is sensitive to noisy data and outliers in data.

J. Naïve Bayesian

Named for Thomas Bayes, an English statistician who gave his name to Bayes' Theorem. Naïve Bayes is not a single algorithm, but it's a family of classification algorithms. It is called "naïve" because as it learns, it assumes that all attributes of an item are independent of each other. It uses less training data for its calculation of parameters in the prediction phase. Based on other known features the algorithm learns to predict an attribute. In simple words, according to Naïve Bayes classifier, the presence of a particular feature of a class is not related to the presence of any other feature [31]. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes

VI. CONCLUSION

Big data with predictive analytics, high-performance computing systems, machine learning techniques, and other strategies are used in the past and will be used heavily in the future too. Across multiple fields, it will explore several applications of big data. Big data analytics is a security-enhancing tool in the future. It's a known fact that the integration of digital data and its implementation through analytics is fetching humongous rewards to brands and businesses around the world. To balance this technology further, dark data is emerging and marks the future of big data. The data sets are generally untapped, unstructured and untagged and so they are referred to as dusty data. The Big Data future scope will predict such data sets will come into the limelight and further modernize the technology further. This paper has an overview of the big data Tools and Algorithms. There are excessively a lot of future important challenges in Big Data management and analytics which arise from the nature of data: large, diverse, and evolving.

REFERENCES

[1] .CH.Chandra Sekhar, CH.Sekhar, 'productivity Improvement in Agriculture Sector Using Big Data Tools', publication a www.researchgate.net, 978-1-5090-6399-4/17/$31.00_c 2017 IEEEConference Paper · March 2017.

[2]. Mohammed K. Hassan, Ali I. El Desouky, Sally M. Elghamrawy, and Amany M. Sarhan, 'Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals', © Springer International Publishing AG 2017.

[3]. Ashish Bajpai, Dayanand, "Big Data Analytics in Cyber Security" International Journal of Computer Sciences and Engineering, Vol:6, Issue;7, 2018.

[4]. K. Sree Divya, P.Bhargavi2, "Machine Learning Algorithms in Big data Analytics", International Journal of Computer Sciences and Engineering, Vol:6, Issue;1,2017.

[5]. Mohamed Abdel-Basset, Mai Mohamed, Florentin Smarandache, and Victor Chang,' Neutrosophic Association Rule Mining Algorithm for Big Data Analysis', Symmetry 2018, 10, 106; DOI: 10.3390/sym10040106

[6]. A Rytsarev, A V Kupriyanov, Clustering of social media content with the use of Big Data technology' The IV International Conference on Information Technology and Nanotechnology.

[7]. Ajay Kumar Pal, Saurabh Pal, 'Analysis and Mining of Educational Data for Predicting the Performance of Students', International Journal of Electronics Communication and Computer Engineering Volume 4, Issue 5.

[8]. Brijesh Kumar Baradwaj, Saurabh Pal,' Mining Educational Data to Analyze Students" Performance', *(IJACSA)* International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[9]. Vinayak A. Bharadi, Prachi P.Abhyankar, 'Analysis And Prediction in Agriculture Data using Data Mining Techniques'International Journal of Research In Science & Engineering Special Issue 7ICEMAN March 2017 IJRISE

[10]. Senthil Vadivu, Dr. Prasanna Devi, Vishnu Kiran, Manivannan,' Modelling A Predictive Analytics Methodology For Forecasting Rice Variety And Quality On Yield On Farm And Farming Attributes Using Bigdata', International Journal of Pure and Applied Mathematics Volume 116 No. 5 2017, 61-65.

[11]. Ashwani Kumar Kushwaha, SwetaBhattachrya,' Crop yield prediction using Agro Algorithm in Hadoop', IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 5, No2, April 2015

[12]. Ekta Joshi1, Dr. D. A.Parikh, 'An Improved K-Means Clustering Algorithm', International Journal of Scientific Research in Science, Engineering and Technology, 2018IJSRSET, Volume 4, Issue 1.

. [13]. D Ramesh, B Vishnu Vardhan, 'Analysis Of Crop Yield Prediction Using Data Mining Techniques', IJRET: International Journal of Research in Engineering and Technology, Volume: 04, Issue: 01 Jan-2015.

[14]. Lean Yu, Yaqing Zhao, Ling Tang c, Zebin Yang d, 'Online big data-driven oil consumption forecasting with Google trends ', International Journal of Forecasting 35 (2019) 213–223, 2017 Elsevier.

[15]. Gunasekaran Manogaran and DaphneLopez, 'Health data analytics using scalable logistic regression with stochastic gradient descent', Int. J. Advanced Intelligence Paradigms, Vol. 10, 2018.

[16]. Abdallah, V. Detriments, H. Mylonas, K. Tatsis & E. Chatzi,' Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data',2018 Taylor & Francis Group, London.

[17]. Eka Sugiyarti, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda,

Shankar K., Andino Maseleno,' Decision Support System Of Scholarship Grantee Selection Using Data Mining ', International Journal of Pure and Applied Mathematics, Volume 119 No. 15 201,

[18]. Michelle Cristina Araujo Picolia, Gilberto Camara, Ieda Sanchez,' Big earth observation time series analysis for monitoring Brazilian Agriculture ', ISPRS Journal of Photogrammetry and Remote Sensing · August 2018, isprsjprs.

[19]. Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson and Dionysis Bochtis,' Machine Learning in Agriculture: A Review ', Sensors 2018, 18, 2674.

[20]. Miguel Molina-Solana, Mar´ıa Rosa,↤, M. Dolores Ruiz, Juan Gomez-Romero, M.J. Martin-Bautista, 'Data Science for Building Energy Management: a review', Article *in* Renewable and Sustainable Energy Reviews, April 2017 DOI: 10.1016/j.rser.2016.11.132.

[21]. Nirbhey Singh Pahwa, Vidhi Soni, Neeha Khalfay, Deepali Vora, Stock Prediction using Machine Learning a Review Paper 'International Journal of Computer Applications (0975 – 8887) Volume 163 – No 5, April 2017.

[22].Mansi Shinde, Kimaya Ekbote, Sonali Ghorpade, Sanket Pawar, Shubhada Mone, 'Crop Recommendation and Fertilizer Purchase System', Mansi Shinde et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (2), 2016, 665-667.

[23]. Magdalena Cantabella, Raquel Martínez-España ∗, Belén Ayuso, Juan Antonio Yáñez, Andrés Muñoz,' Analysis of student behavior in

learning management systems through a Big Data framework', Future Generation Computer Systems 90 (2019)262272,2 2018 Elsevier B.V.
[24]. Mohamed Abdel-Basset 1,* ID, Mai Mohamed 1, Florentin Smarandache 2,* ID and Victor Chang 3, 'Neutrosophic Association Rule Mining Algorithm for Big Data Analysis', Symmetry 2018, 10, 106; DOI:10.3390/sym10040106.
[25]. F. Padillo · J.M. Luna · S. Ventura, 'A grammar-guided genetic programming algorithm for associative classification in Big Data', Article in Cognitive Computation · November 2018.
[26]. J. Jenifer Nancy, Jhansi Rani, Dr. D. Devaraj, 'Association Rule Mining in Big Data using MapReduce Approach in Hadoop', GRD Journals | Global Research and Development Journal for Engineering | International Conference on Innovations in Engineering and Technology (ICIET) - 2016 | July 2016.
[27]. Daxin Tian, Yuki Zhu, Xuting Duan, Junjie Hu, Zhengguo Sheng, Min Chen, IEEE, Jian Wang, Yunpeng Wang,' An Effective Fuel Level Data Cleaning And Repairing Method for Vehicle Monitor Platform', - IEEE Transactions, 2019 - ieeexplore.ieee.org.
[28]. Mohamed Alleghany, Dhiya-Jumeily, Thar Baker, Abir Hussain, Jamila Mustafin, and Ahmed J. Aljaaf,' Applications of Machine Learning Techniques for Software Engineering Learning and Early Prediction of Students' Performance ',© Springer Nature Singapore Pte Ltd. 2019.
[29]. Shili Lian, Liangliang Wan, Ling Zhan, and Yansheng Wu, 'Research on Recognition of Nine Kinds of Fine Gestures Based on Adaptive AdaBoost Algorithm and Multi-Feature Combination', IEEE, VOLUME 7, 2019.
[30]. Rishabh Kapoor, Shiva Verma & M L Sharma,' Multi-Lingual Comparative Sentiment Analysis Of Twitter Data Using Ann Algorithm', R Kapoor, S Verma, ML Sharma - Digitalization,2018researchgate.net, ISBN: AR No.-3093/ISBN/2018/A.
[31]. A. Sankari Karthiga, M. Satish Mary, M. Yogini,' Early Prediction of Heart Disease Using Decision Tree Algorithm', International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST) Vol.3, Issue.3, March 2017.

Authors Profile

P.Pandeeswary is currently pursuing the M.Phil Degree in Computer Science in the Department of Computer Science, Dr. Umayal Ramanathan College ForWomen, Karaikudi. Her research Interest includes Diabetic disease prediction and big data analytics.

Second Author Dr. M.Janaki is working as an Associate Professor in the Department of Computer Science, Dr. Umayal Ramanathan College For women, Karaikudi. She has 14 years of teaching experience. She has Published 12 International Journals. Area of research includes Cloud Computing Security, Big data Analytics. She has delivered Lectures in various colleges and conferences. Her main focus is to teach with passion, to make innovative research, to nurture the young minds to build a better society.

