A Survey Paper on Big Data Analytics

M.D. Anto Praveena[1] Dr. B. Bharathi[2]

[1]Research Scholar, Sathyabama University, Chennai, Tamilnadu, India. Email:antopraveena@gmail.com
[2]Professor, Sathyabama University, Chennai, Tamilnadu, India. Email:bharathi.cse@sathyabamauniversity.ac.in

ABSTRACT-In recent years, the internet application and communication have seen a lot of development and reputation in the field of Information Technology. These internet applications and communication are continually generating the large size, different variety and with some genuine difficult multifaceted structure data called big data. As a consequence, we are now in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which one will be relevant to the phenomenon of interest. For example, E-commerce transactions include activities such as online buying, selling or investing. Thus they generate the data which are high in dimensional and complex in structure. The traditional data storage techniques are not adequate to store and analyses those huge volume of data. Many researchers are doing their research in dimensionality reduction of the big data for effective and better analytics report and data visualization. Hence, the aim of the survey paper is to provide the overview of the big data analytics, issues, challenges and various technologies related with Big Data.

Keywords: Big Data, Big Data Analytics

I. INTRODUCTION

Today, system and people uses the web with an exponential generation of large size of data. The size of data on the web is measured in Exabyte (EB) and Petabytes (PB). By 2025, the prediction is that the Internet will surpass the brain size of everyone living in the whole world. This firm growth of data is because of advances in digital sensors, computations, communications, and storage that have created large gatherings of data. The name Big Data had been devised, by Roger Magoulas a researcher, to describe this singularity.

Gartner Company stated that, Information or data will be the 21st century oil. In last 25 years, data has grown massively in various fields with different types. According to the statistical report of International Data Corporation (IDC), in the year 2011, the overall data volume created in the world was 1.8ZB that was enhanced by nearly nine times within next five years [1]. Now with the inclusion of marketing, smart city, the results of disease control and prevention and business intelligence applications it can be effortlessly understand that big data plays a vital role everywhere in the universe [2].

With the increase in universal data volume, the technology of big data and its analytical processes are generally used to provide the description about massive datasets. Compared with other traditional datasets and its processes, big data includes semi structured and unstructured data that need more real time analysis. Big data also gets details about new prospects for determining new values, supports us to improve an in-depth understanding of the hidden values, and also incurs new challenges, For instance, how to exceptionally organize and manipulate such big datasets. The volume of information from various sources is growing large, it also provides about some challenging issues demanding rapid resolutions. Big data visualization process is another vital process which takes an important place in big data analytics problems. Because through data visualization only the final report of data analytics will be visualized.

Since the field of Information Technology (IT) is improving a lot recently, this generates the data more easily. For instance, for every minute approximately 72 hours of video files are uploaded to YouTube by the people. This data growth challenges the field with the main problems of gathering and integrating huge volume of data from widely distributed data sources such as social media applications.

Also the unexpected growth of the cloud computing and Internet of Things (IoT) promote the growth of data. Cloud computing provides the standard for storing and accessing the enterprises data for the big data assets. In IoT, sensors are used to gather and transmit the data to be stored and processed in the cloud storage. Such data types and size are exceeds the abilities of the IT architectures and set-up of existing enterprises and its real-time requirement and its computing capacity. This increase in data volume cause many issues in storing and retrieving the massive heterogeneous datasets with the special hardware and software infrastructure.

As a result, this survey targets at providing a brief review on the big data analytics. This literature survey further organized as given below: Chapter II explains the major concepts of big data analytics and its applications. Chapter III explains the technologies

978-1-5090-6135-8/17/$31.00 ©2017 IEEE

used to implement various applications. Chapter IV explains the research challenges, related technologies. Chapter V depicts big data algorithms followed by conclusion and future enhancements.

II. BIG DATA-AN OVERVIEW

A. Big Data

Big Data increasingly benefits both research and industrial areas such as health care, finance service and commercial recommendation [1]. The Economist says, Data are becoming a new raw material of business. Economic input is almost equivalent to capital and labor. Nowadays, the data to be analyzed are dynamic and huge in volume, Also they are the group of different data types. These data come from different data sources such as Whatsapp, Twitter, Facebook, YouTube, Mobile phones GPS signals and more. Hence, the Big Data has the unique features such as heterogeneous, unstructured, semi structured, incompleteness, high dimensional.

According to industrial data analyst Doug Laney defines the big data is articulated in the year 2000's as the three V's [3]:

***1) Volume (Data in Rest*)**: Organizations collect data from a variety of data sources, including commercial transactions, social media data and information from sensors or machine-to-machine data.

***2) Velocity (Data in Motion*)**: Data streams come in at unmatched speed and should be allocated with in an appropriate manner. Different kind of IoT sensors, RFID tags and smart metering are driving the necessity to deal with data flows in real time.

***3) Variety (Data in Many Forms*)**: Data comes in different kinds of formats such as structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock and financial transactions.

But these three V's are extended as five V's later by adding two more V's such as variability and veracity. They are as follows

4) Variability (Data in Highlight): Inconsistency of the data set can hamper processes to handle and manage it.

5) Veracity (Data in Doubt): Refers to the messiness or trustworthiness of the data. The quality of captured data can vary greatly, affecting accurate analysis.

All major IT companies, including EMC, Microsoft, Google, Amazon, and Facebook, etc. already have started their big data projects. To extract information or data from big data, optimal processing power, analytics capabilities and skills are needed [5]. So, dealing the big data effectively requires generating the value against the volume, variety and veracity of big data [7].

B. Big Data Analytics Operations

To develop the knowledge discovery in databases (KDD) more clear, Fayyad and his colleagues concluded that the KDD process as shown in Fig 1 which has selection, preprocessing, transformation, data mining, and interpretation. With the above operations, it will be capable to form a complete data analytics system which is collecting the data and then find information from the data and visualize the knowledge to the user.

Fundamentally, data processing is seen as the collecting, processing, and management of data for producing new information for end users [8]. Karmasphere currently splits Big Data analysis into four steps: Acquisition, Assembly, Analyze and Action. Thus, these steps are mentioned as the 4 A's.

1) Acquisition:

Big Data architecture has to obtain high speed data from a different kind of data sources and it has to deal with different access control protocols. It is where a filter could be recognized to store only data which could be helpful or underdone data with a lesser degree of uncertainty [9]. In some applications, the conditions of generation of data are important, thus it could be interesting for further analysis to capture these metadata and store them with the corresponding data.

2) Assembly:

At this point the architecture has to deal with various data formats and must be able to parse them and extract the actual information like named entities, relation between them, etc [9]. Also this is the point where data have to be clean, put in a computable mode, structured or semi-structured, integrated and stored in the right location. Thus, a kind of Extract, Transform, and Load had to be done. Successful cleaning in Big Data architecture is not entirely guaranteed. In fact the volume, velocity, variety, and variability of Big Data may preclude us from taking the time to cleanse it all thoroughly.

3) Analyze:

Here we have running queries, modeling, and building algorithms to find new insights. Mining requires integrated, cleaned, trustworthy data. At the similar time, data mining can also be used to help enhance the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions [9].

***4) Action*:**

Valuable decisions are need to be capably interpreting the results from analysis. Consequently it is very important for the user to understand and verify outputs [9]. Further, origin of the data should be provided to help the user to know he obtains.

5) Privacy:

R. Hillard was considered it to be very significant that privacy appears in a better place in his definition about big data. Privacy can cause many problems at the analysis of data, at the creation of data [10] because if we want to aggregate data or to associate it we could have to access private data and privacy can also cause inconsistencies at the eliminating of database. To sum up handling big data implies having an infrastructure linear scalable, able to handle high throughput multi-formatted data, auto recoverable, fault tolerant, with a higher degree of parallelism and a distributed data processing [11].

C. Big Data Analytics Infrastructure

The following Fig 1 shows different layers occurs in the big data analytics.

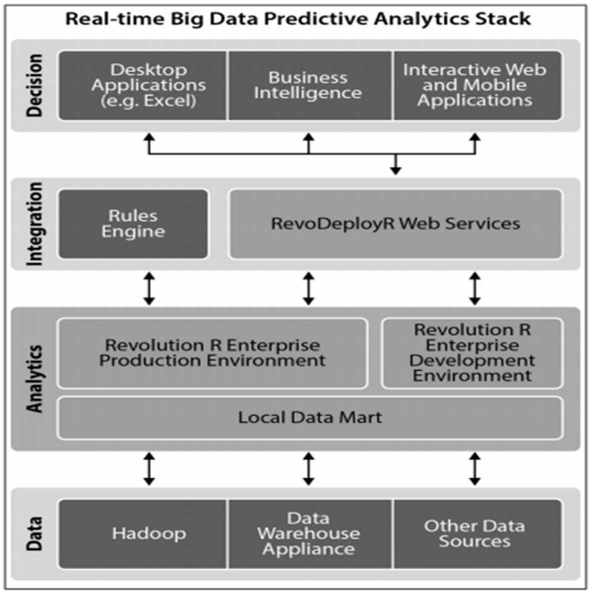
Real-time Big Data Predictive Analytics Stack
Decision
Desktop Applications (e.g. Excel)
Business Intelligence
Interactive Web and Mobile Applications
Integration
Rules Engine
RevoDeployR Web Services
Analytics
Revolution R Enterprise Production Environment
Revolution R Enterprise Development Environment
Local Data Mart
Data
Hadoop
Data Warehouse Appliance
Other Data Sources

Fig 1: Big Data Analytics Implementation Layers

The implementation Layers are as follows [12]:

1) Data Layer

This layer has RDBMS based structured data, Semi-structured and unstructured based data. NoSQL databases are used to store the unstructured data. For instance, MongoDB and Cassandra are the NoSQL databases. Streaming data from the web world, social media domain, data from IoT sensors and operational systems are the examples to unstructured and semi-structured data. Software tools such as HBase, Hive, HBase, Spark and Storm are also sitting at this layer. Hadoop and Map Reduce also support this layer.

2) Analytics Layer

Analytics layer has the environment to implement the dynamic data analytics and deploy the real time values. It has building models developing environment and modify the local data in regular interval. This also improves the performance of the analytical engine.

3) Integration Layer

This layer integrates the end user applications and analytical engine. This includes usually a rules engine and an API for dynamic data analytics.

4) Decision Layer

This layer is where the end product hits the market. It includes applications of end user such as mobile app, desktop applications, interactive web applications and business intelligence software. This is the layer where people interact with the system.

Each and every layer described above is associated with different sets of end users in real time and enables a crucial phase of real time data analytics implementation.

D. Big Data Applications

There are so many big data applications around us as shown in Fig 2. Few of them are described below:

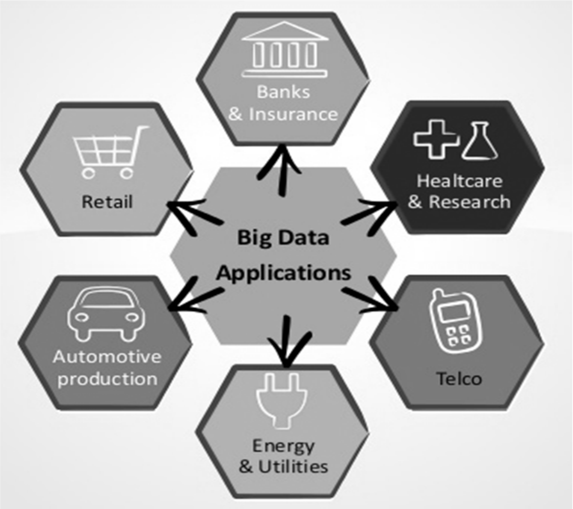

Fig 2: Big Data Application Areas

1) Fraud Recognition and Control:

Business operations face many types of fraudulent claims or transaction processing. Hence fraud recognition and control is most resounding big data application [15]. In most cases, fraud is discovered long after the fact, at which point the loss has been done and all that's left is to minimize the harm and adjust policies to prevent it from happening again. Big data platforms that can verify, analyze, claims and transactions in real time, identifying large scale patterns across so many transactions or detecting inconsistent behavior from an individual user, can change the fraud detection game.

2) Call Center Analytics:

Now we turn to the customer related big data application examples, in which call center data analytics are specifically powerful application. The current way of process in a customer's call center is often a great barometer and influencer of market sentiment, but without a big data solution, much of the awareness that a call center can provide will be ignored or revealed too late. Big data solutions can help ascertain recurring problems or customer and staff behavior patterns on the fly not only by making intellect of time or quality resolution metrics, but also by capturing and processing call content itself.

3) Log Analytic in IT:

IT departments and consultancies are generates a huge amount of logs and trace data. Without a big data solution, huge volume of the data may go unexamined. All organizations naturally do not have the source or manpower to agitate through all that information by hand, let alone in real time. With a big data solution, however both logs and trace data may be put to better use. Within this list of big data application examples, IT log analytics is the most largely applicable. Any organization with a large IT department will get assistance from the ability to quickly identify large-scale patterns to help in diagnosing and preventing problems in the field. In the same way, any organization with a large IT department will escalate the capacity to ascertain incremental performance optimization opportunities.

4) Social Media Analysis:

Of the customer-facing Big Data application examples could discuss, analysis of social media activity is one of the most important. Everyone and their mothers are on social media these days, whether they like company pages on Facebook or tweeting complaints about products on Twitter. A big data solution built to produce and investigates social media activity, like IBM's Cognos Consumer Insights, a fact solution running on IBM's big Insights big data platform, may make the sense of the chatter. Social media data can provide real time insights into how the market is responding to products and campaigns. With those insights, companies can adjust their pricing, promotion, and campaign placement on the fly for optimal results.

5) Finance Analysis

Big Data analytics can be used to analyze the financial status and prediction in enterprises. For Example, the tool is analyzing the critical stock market moves and supports in making global financial prediction and decisions. Even though this is not a fool-proof process, it is definitely advancement in the field.

6) Agriculture:

In agriculture, biotechnology centers use sensor data to enhance crop efficiency. It does test the crops and simulates to measure the plants reaction to various conditions. Its environment continuously adjusts to changes in the characteristics of various data including water level, temperature, growth, output, and gene sequencing of each and every plant in the testing environment called test bed.

III. BIG DATA TECHNOLOGIES

Big data management is the organization and manipulation of huge volumes of structured data, semi-structured data and unstructured data. The aim of big data management is to make sure the quality of high level data and availability of data for business intelligence and big data analytics applications.

There are various tools which can be used for big data management from data acquisition, data storage to data visualization. This section describes those tools and related tools. Few of the tools which are used for different purpose are described below:

A. Data Analysis

1) Hadoop:

This is an open source platform (tool) for treating big data and its analytics. It is user friendly and flexible to work with different data sources, either gathering various sources of data or accessing the data from a database in order to run processor-intensive machine learning process [21]. This tool has different types of applications such as location based data from weather, traffic sensors and social media data.

2) Map Reduce:

This is the programming environment that permits larger jobs implementation scalability against group of server. Map Reduce implementation has two main tasks: The Map task converts input dataset is into a different set of value pairs. The Reduce task combines several outputs of the Map task to form reduced tuples.

3) Hive:

Hive is the SQL-like bridges that permit predictable business applications to run SQL queries against a Hadoop cluster. It was developed earlier by Facebook, then it has been made open source software tool now, and it is a high level perception of the Hadoop which allows all to make queries against data stored in a Hadoop storage medium just as if they were manipulating a conventional data store.

4) PIG:

This is another analytical tool that attempts to make the Hadoop closer to the developers and business users. PIG contains of a Perl like language which permits the query execution over data stored on a Hadoop instead of a SQL.

5) WibiData:

Wibidata is the tool that developed for the enterprises to personalize their customer experiences. It combines the web analytics with Hadoop. It is built on top layer of the HBase. It allows web sites to explore better and process with their user data, allows real time responses to user, such as recommendations, serving personalized content, and decisions.

6) Platfora:

Platfora is the big data discovery and analytical tool. It is a platform which takes the user's queries into Hadoop jobs automatically, thus creating an abstraction layer which can exploit to simplify and organize datasets by anyone.

7) Rapidminer:

It is a software platform developed by the company of the same name that provides an integrated platform for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all procedures of the data mining process including dataset preparation, validation, results visualization and optimization.

B. Storage Technologies

As the size of data grows enormously, there is a need for efficient and effective storage techniques to handle the big data. The main advancements in this space are associated with data compression and storage virtualization [17].

1) HBase:

A non-relational (NoSQL) databases that runs on top of HDFS. Apache HBase is an open source NoSQL database environment that provides real time read and write access to those large databases. An HBase scale linearly to handle very large data sets with few billions of rows and millions of columns, and it easily combines data sources that use a wide variety of different structures and schemas [14]. HBase is natively integrated with Hadoop and works seamlessly with access engine YARN.

2) SkyTree:

It is a high-performance data analytics and machine learning platform which focuses specifically on big data analytics and handling. Machine learning, is a needed part of big data, because the high data volumes make the exploration manually. Automated data exploration approaches are too expensive.

3) NoSQL (Non- Relational Databases):

This NoSQL (Not only SQL) database, also called Not Only SQL, is an approach to data administration and database design that's useful for the big volume of data sets in distributed background.

The most popular NoSQL database is built using Apache Cassandra. Cassandra, which was once Facebook’s proprietary database, was released as open source in 2008. Other NoSQL databases implementations include SimpleDB, Google BigTable, Map Reduce, MemcacheDB, Cassandra, MongoDB and Voldemort. Companies that use NoSQL include social medias Netflix, LinkedIn, and Twitter.

C. Visualization Tools

There are so many open source visualization tools are available in market. Here very few of them are given below [18].

1) R Tool:

R is a well-known programming language and software tool for graphic and statistical computing based data visualization. This is supported by the R Foundation for Statistical Computing. The R Tool is broadly used among statistical area and data miners for developing statistical software and data analysis.

2) Tableau:

Tableau is the tool is used to visualize the result in the form of charts, maps, graphs and many other graphics form. A desktop application is available for visual analytics.

3) Infogram:

In this tool, there are easy three steps processes used to select among many visual templates, differentiate this tool with additional visualizations like charts, map, videos and ready to share your visualization. This is supporting accounts for video/audio files publishers and for the journalists of research script publisher, branded policies for businesses and classroom accounts for educational projects.

***4) ChartBlocks**:*

It is an easy-to-use free online tool that requires no complex coding, and builds visualizations from databases and spreadsheets.

5) Ember Charts:

This tool is based on the framework called Ember.js framework and it uses D3.js under the hood. Ember Charts features scatter charts, bar, pie, time series. It's very well-designed and easy to use tool.

6) Tangle:

It is a data visualization tool beyond the visualization, allows the designers and program developers to generate reactive programs that gives a better understanding of data relationships.

D. Big Data and Other Technologies

This section provides some important technologies which are closely related to big data. They are described as follows:

1) Association with Cloud Computing:

Cloud computing is the technology that can be used to store huge volume of data in web. The important target of cloud computing is to use high level computing and huge volume resources under firm management, so as to arrange for big data applications with well-defined computing capability. The improvement of cloud computing provides solutions for the storage and processing of big data. The development of big data accelerates the enlargement of cloud computing. The distributed storage technology based cloud computing can efficiently manage big data. The advancement of cloud computing can improve the efficiency of big data analytics.

2) Association with Internet of Things (IoT):

There are large number of networking sensors are implanted into various IoT devices and machines around the world. Those implanted sensors may acquire different types of data, such as network communication data, geographical data, environmental data, astronomical data, and logistic data [4]. Since the sources of data collected from different environment, IoT produced big data has different type of characteristics when it is compared with normal big data. These data has some special characteristics such as heterogeneity, variety, noise, and redundancy. An authenticated report from Intel corporation says that big data in IoT has three classic characteristics. They are as follows [16]:

(i) Plentiful terminals producing massive data.
(ii) Data produced by IoT is commonly semi-structured or unstructured
(iii) IoT data will be useful only if it is analyzed.

3) ***Association with Data Center:***

The data center is not only a paradigm for storage of data, also supports more responsibilities includes gathering of data, processing of data, organizing data, and optimizing the data values and operations in the big data paradigm. Date center has high volume of data which organizes and accomplishes data according to its objective. The development of big data provides better opportunities and challenges to data centers.

IV. BIG DATA CHALLENGES

There are many crucial challenges need to be focused while handling of Big Data and its analytical process [6]. Many big data researchers are focusing on the following challenges in their research. They are follows:

A. Storage:

The size of hard disks in the computing system nowadays is in the range of Terabytes (TB). The quantity of data produced via internet is measured in terms of Exabyte (EB). Even though the data produced in educational area is not as huge as the data produced through internet. It will get much bigger in future. So the traditional RDBMS tools such as oracle, MySQL are not able to store or process such kind of Big Data since they are not a structured data. To give the solution for this issue, databases uses NoSQL based databases such as Cassandra and MongoDB which handles unstructured and semi-structured data.

B. Data representation:

Many dataset has definite levels of heterogeneity in structure, semantics, type, organization, granularity and accessibility. Data representation aims to make data more important for data analytics and user analysis. Any improper data representation may reduce the value of the data originality and even disturbs effective data analysis process [23]. Hence if the data is represented effectively, then analysis process will be done easier.

C. Data life cycle management:

Data life cycle management process decides which data shall be stored and which data shall be discarded during the analytical process. There are challenges, one of which is that the existing storage system could not support such huge amount of data. Therefore, a principle which makes the life cycle management system effective is needed.

D. Analysis:

As big data is generated from various types of online education websites, they vary in structure and the volume. Data Analysis process may ingest time and resources. To face this issue, a special scaled out architectures are used to process the data in a distributed manner. Data are split into fragments and handled in a number of computers available in the network and the processed data is combined.

E. Reporting:

Reporting is the process which involves in displaying statistical data in the form of values. When the data size is huge, then the traditional reporting methods become challenging to understand. In these cases the statistical reports must be represented in a particular form that can be easily understood.

F. Redundancy Reduction and Data Compression:

Redundancy reduction and data compression are operational method to decrease the cost of the system by reducing the data redundancy and data compression. For example, sensor based network produced data is highly redundant. This kind of data may be filtered and redundancy may be reduced.

G. Energy Management:

The growth of data size and analytical process, storage management and transmission management of big data will certainly consume more and electric energy. So, power consumption control and management mechanism may be implemented for big data to ensure the less energy consumption.

H. Data Confidentiality:

Data confidentiality is another big blow for big data as the service providers and owners of the data could not maintain and analyze such huge datasets effectively. They depend on professionals or third-party tools to analyze such data, which improve the potential safety risks. Hence, data confidentiality is important issue for the researchers.

I. Expendability and Scalability:

The big data analytics should support present and future datasets. The big data analytics algorithm may be able to process expendability and scalability of data.

J. Cooperation:

Since big data analysis is an interdisciplinary research, it requires cooperation from domain experts to gather the possible of big data. So the big data network architecture may be established to help scientists and engineers those who involved in the process.

K. Big Data Dimensional Reduction:

Visualization is an important method in micro biome data analysis, and dimensionality reduction is a necessary procedure to achieve it. Many researchers now concentrate on the issues of big data dimensionality reduction. [24]. Few of the work have been described as follows:

Dimensionality reduction of data is the approach of decreasing the number of random variables under consideration via obtaining a set of uncorrelated principal variables. It can be divided into feature selection and feature extraction. Reducing the dimensions of data to 2D or 3D may allow us to plot and visualize it precisely. It helps in data compressing and reducing the storage space required. It fastens the time required for performing same computations. According to KNIME [1] [19] there are seven methods to reduce input dimensionality. They are as given below:

1) Eliminating data columns with missing values:

If the data values are missing in the data column then data column cannot transfer the information. So target of this process is to remove the data column which has no values or too many missing values. To overcome this issue identify the number of missing values data columns and eliminate them.

2) Low variance Mesh:

The Low Variance Filter node computes variance of each column and excludes the data columns with a variance value less than its threshold value. Variance can only be calculated for numerical value columns.

3) Reducing highly associated columns:

Suppose a data column values highly associated with other data column value, those columns are not providing any new information to the current group may be removed without diminishing the size of information available for future tasks dramatically [25]. Removal of highly associated data columns may be done by measuring the association between pairs of data columns using the method called Linear Correlation node. Then the Correlation Filter node may be applied to remove one of two highly correlated data columns.

4) Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a procedure that moves the original x coordinates of a data set into a new set of x coordinates called principal components. PCA uses an orthogonal transformation to move

5) Backward Feature Elimination:

The backward feature elimination is a loop based classification algorithm performs dimensionality reduction on data against a specific machine learning algorithms. In this algorithm, the selected classification algorithm is trained on x input features. Then it eliminates one input feature at a time and train the same model on x-*1* input features x times. At last, the input feature whose removal has fashioned the smallest increase in the error rate is removed, leaving us with x-*1* input features. The classification process is then continued using x-*2* features x-*1* times and, again, the feature whose removal produces the smallest interruption in classification performance is removed for better purpose. This gives us x-*2* input features. This classification algorithm starts with all available X input features and continues till only 1 last feature is left for classification. Each iteration *k* then creates a model trained on x-*k* features and an error rate *e (k).* Selecting the maximum tolerable error rate, we define the smallest number of features needed to attain that classification performance with the selected machine learning algorithm.

***6) Forward Feature Construction*:**

Forward feature construction process forms a number of pre-selected classifiers using an incremental number of input features. This method is based on classification algorithm which starts from one feature and enhances with other feature at a time in the continuous iterations.

7) Dimensionality Reduction via Tree Ensembles: An approach to dimensionality reduction of data is to generate an outsized and cautiously constructed set of trees against an objective attribute and then use each attribute's usage statistics [20] to find the most informative subset of important features.

V. BIG DATA ANALYSIS ALGORITHMS

Data mining algorithms and its techniques for data analysis are playing vital role in the big data analytics in terms of the dimensionality reduction, computational cost, memory requirement and management and accurateness of the end results. This section gives a brief discussion from the perspective of analysis and search algorithms [28] to explain its importance.

A. Clustering Algorithm:

One of the most popular clustering tools is CloudVista which is used in cloud computing to implement the clustering process in parallel. BIRCH and other clustering methods are used in CloudVista to show that can be handle very large scale data. GPU is another clustering tool [22] which is used to improve the performance and safety of a clustering algorithm.

B. Classification algorithms:

Like clustering algorithm for big data mining, the plan and implementation of classification algorithm learned into account the input data that are gathered by the data sources and they will be managed by a heterogeneous set of learners.

C. Frequent Pattern Mining:

Most of the time, data mining researchers on frequent pattern mining are focusing on handling huge volume of dataset at the very beginning because some initial approaches of them attempted to examine the data from the transaction data of large enterprises and shopping malls.

D. C4.5:

This tool builds a classifier in the form of decision tree. A classifier is a tool in data mining that takes a group of data which specifies the things want to classify and put efforts to predict which class the new data belongs and how it belongs to. Decision tree learning creates approximately similar to a flowchart to classify new data.

E. K-Means:

K-means algorithm creates *k* groups from a set of data or objects so that the members of a group are more similar. It's a popular data clustering and analysis technique.

F. Apriori:

The Apriori algorithm learns association rules and is implemented to a database containing a very large number of transactions and its data. Association rule learning is one of the data mining technique for learning correlations and association among variables in a database.

G.* Expectation-Maximization (*EM):

This algorithm is generally used as a clustering algorithm for knowledge discovery in mining. In statistics, the EM algorithm iterates and optimizes the likelihood of seeing experimental data until estimating the parameters or values of a statistical model with not experiment variables.

H. PageRank:

This is another analysis algorithm named PageRank which is link analysis algorithm designed to standardize the relative significance of some object linked within a network of data objects. This algorithm processes a type of network analysis looking to explore the associations among objects and rank them.

I. AdaBoost:

Adaboost algorithm constructs a classifier. It is a classifier which brings the data and attempts to predict which class a new data element belongs to. The aim of this algorithm is to make a group of weak learners and integrate them to create a single strong learner.

VI. CONCLUSION

In this literature survey, big data and its various concepts includes big data analytics, big data analytics techniques, data visualization and big data analysis algorithm have been studied. Also this survey gives overview of the possible opportunities of big data research environment. They are as follows

i) The scheduling methods are used to handle the computation resources of the cloud based platform and to make it to finish the task of data analysis as fast as possible.

ii) The other issues such as the data privacy and data security that go along with the work of data analysis are inherited inquiry topics which contain instruction to safely store and manipulate the data, how to confirm the data communication is protected, and how to prohibit someone from finding out the information about us. many problems of data security and data privacy are essentially the same as those of the traditional data analysis even if we are entering the big data age. Thus, protecting the data is the inevitable concept will also appear in the research of big data analytics.

iii) The efficient methods are used to decrease the comparison, sampling, computation time of input and a variety of reduction methods which are playing an important role in big data analyst

REFERENCES

1. Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized Travel Sequence recommendation on Multisource Big Social Media, 2016, IEEE Transactions on Big Data,Vol.2, Issue:1
2. Vallabh Dhoot, Shubham Gawande, Pooja Kanawade and Akanksha Lekhwani, Efficient Dimensionality Reduction for Big Data Using Clustering Technique, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-5, 2016, ISSN: 2454-1362
3. Gantz J, Reinsel D, Extracting value from chaos.IDC iView, 2011, pp 1–12
4. Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A survey, Mobile New Applications 2014, 171-209
5. Mayer-Schonberger V, Cukier K, Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.
6. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart.2012; 36(4):1165–88.
7. Kitchin R. The real-time city? Big data and smart urbanism. Geo J. 2014, 79(1), pp: 1–14.
8. Katrina Sin and Loganathan Muthu, Applications of big data in education data mining and learning analytics – A literature Review, ICTACT Journal on soft computing special issue on soft computing models for big data, July 2015, Vol:05, Iss: 04, pp: 1035-1049
9. Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey , Computer Science Review, 2015, Vol: 17, pp: 71-80
10. K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
11. H.V. Jagadish, D. Agarwal, P.Bernstein, Challenges and Opportunities in Big Data, The Community Research Association, 2015
12. K. Davis, D. Patterson, “Ethics of Big Data: Balancing Risk and innovation”, O’Reilly Media, 2012.
13. K. Krishnan, Data warehousing in the age of big data, in the Morgan Kaufmano series on Business Intelligence, Elsevier Science, 2013.
14. Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture, ISBN: 978-1-449-36421-2, 2013
15. http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data
16. http://hortonworks.com/apache/hbase
17. http://www.slideshare.net/infoDiagram/big-data-cloudappsvisualiconpptinfodiagramtoolbox
18. https://blog.profitbricks.com/39-data-visualization-tools-for-big-data
19. https://www.knime.org/files/knimeseventechniquesdatadimreduction.pdf
20. Seung-Hee Bae, Jong Youl Choi, Judy Qiu, Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation, HPDC,2010 Chicago
21. Cheng-Long Ma , Xu-Feng Shang , Yu-Bo Yuan, International conference on machine learning and cybernetics, 2012, vol:4
22. Jun Yan, Benyu Zhang, Ning Liu Shuicheng Yan , Effective and efficient dimensionality reduction for large scale and streaming data preprocessing, IEEE Transactions on Knowledge and data engineering, 2016, Vol:18, issue:3
23. Yetial Fan, Bon song, Yuan Ling, wei wu, A Novel Dimensionality reduction algorithm based on laplace matrix for microbiome data analysis, IEEE International Conference on Bioinformatics and Biomedicine, 2015, pp:49-54
24. Alhussein Fawzi, Bei chen, Pascal Frossard, Mathieu sinn, Structured Dimensionality reduction for additive model regression, IEEE Transactions on knowledge and data engineering, 2016, vol: 28, No:6, pp: 1589-1601.
25. Aswani Kumar, Srinivas.S. A note on the effect of term weighting on selecting intrinsic dimensionality of data, Cybernetics and Information Technologies, 2009, Vol. 9, No. 1, pp: 5-12.
26. J. MacQueen, some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symposium. Math. Statist. Probab., Berkeley, CA, 1967, pp. 281–297
27. Xu H, Li z. Guo S, Chen K,Cloudyvista, Interactive and Economical Value Cluster Analysis data in the Cloud, Proc VLDB Endowment. 2012; 5(12):1886–89.
28. Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufo Abdelaziz Bouras, A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, on Emerging Topics on Computing, IEEE, 11 June 2014.