RESEARCH IN BIG DATA – AN OVERVIEW

Dr. S.Vijayarani[1] and Ms. S.Sharmila[2]

[1]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore
[2]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore

ABSTRACT

Big data is a prominent term which characterizes the improvement and availability of data in all three formats like structure, unstructured and semi formats. Structure data is located in a fixed field of a record or file and it is present in the relational data bases and spreadsheets whereas an unstructured data file includes text and multimedia contents. The primary objective of this big data concept is to describe the extreme volume of data sets i.e. both structured and unstructured. It is further defined with three "V" dimensions namely Volume, Velocity and Variety, and two more "V" also added i.e. Value and Veracity. Volume denotes the size of data, Velocity depends upon the speed of the data processing, Variety is described with the types of the data, Value which derives the business value and Veracity describes about the quality of the data and data understandability. Nowadays, big data has become unique and preferred research areas in the field of computer science. Many open research problems are available in big data and good solutions also been proposed by the researchers even though there is a need for development of many new techniques and algorithms for big data analysis in order to get optimal solutions. In this paper, a detailed study about big data, its basic concepts, history, applications, technique, research issues and tools are discussed.

KEYWORDS:

Big data, Technologies, Visualization, Classification, Clustering

1. INTRODUCTION

Big data is associated with large data sets and the size is above the flexibility of common database software tools to capture, store, handle and evaluate [1][2]. Big data analysis is essential for analysts, researchers and business people to make better decisions that were previously not attained. Figure 1 explains the structure of big data which contains five dimensions namely volume, velocity, variety, value and veracity [2][3]. Volume refers the size of the data which mainly shows how to handle large scalability databases and high dimensional databases and its processing needs. Velocity defines the continuous arrival of data streams from this useful information's are obtained. Furthermore big data has enhanced improved through-put, connectivity and computing speed of digital devices which has fastened the retrieval, process and production of the data.

Veracity determines the quality of information from various places. Variety describes how to deliver the different types of data, for example source data includes not only structured traditional relational data but it also includes quasi-structured, semi-structured and unstructured data such as text, sensor data, audio, video, graph and many more type. Value is essential to get the economic

value of different data which varies significantly. The primary challenge is to identify which are valuable and the way to perform transformation and the technique to be applied to perform data analysis [1].

Big data has three types of knowledge discovery; they are novelty discovery, class discovery and association discovery. Novelty discovery is used to find a new, rare one, previously undiscovered and unknown from a billion or trillion objects or events [2]. Class discovery finds new classes of objects and behavior and association discovery is used to find an unusual co-occurring association. This data by its innovative method is changing our world. This innovative concept is being driven by various aspects: A proliferation of sensors, creation of almost all information in digital form, dramatic cost reductions in storage, remarkable increase in network bandwidth, impressive cost reductions and scalability improvements in computation, efficient algorithmic breakthroughs in machine learning and other areas [2]. Analysis of big data is used to reduce fraud, helps to improve scientific research and field development. Figure 1 illustrates the structure of big data [1].

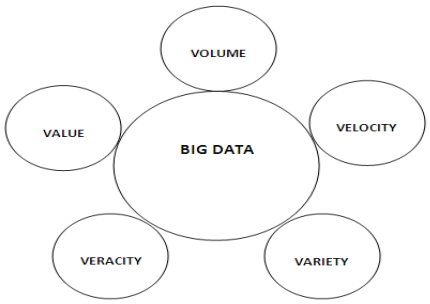

Figure.1.Structure of Big Data

Few typical characteristics of big data are the integration of structured data, semi-structured data and unstructured data. Big data addresses speed and measurability, quality and security, flexibility and stability. Another important advantage of big data is data analytic. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Table 1 shows the comparative study of different types of data based on its size, characteristic, tools and methods [1] [3].

Table 1: Comparative Study on Types of Data

Data Types	Data Sizes	characteristics	Tools	Analytical methods	Examples
Small Data	Mega bytes	Hundred – thousand records	Personal computers, excel, R	Simple statistics	Sales records, customer Database for small companies
Large data	Giga bytes, Tera bytes	Millions of records - structured data	RDBMS, Data warehouses	Advance statistics, Data mining, business intelligence	customer Database for big companies
Big data	Giga bytes , Peta bytes	Over millions of records – distributed and unstructured	Cloud ,Data centre, NoSQL, Hadoop	Map reduce , Distributed file systems	Customer interaction-social network, mobile, multimedia

The remaining portion of the paper is systematized as follows. Section 2 gives the need for big data, applications, advantages and characteristics. Big data tools and technologies are discussed in Section 3. Section 4 provides the detailed description about big data. Section 5 presents big data challenges. Finally Section 6 concludes and discussed about recent trends

2. NEED FOR BIG DATA

The massive volume of data could not be expeditiously processed by traditional database strategies and tools and it mainly focused and handled structured data [1]. At the time of development of computers the amount of data stored in the computers are very less due to its minimum storage capacity. After the invention of networking, the data stored in computers are increased because the improved developments in the hardware components. Next, the arrival of an internet creates a boom to store vast collections of data and it can be used for various purposes [2]. This situation raised concerns about the introduction of new research related concepts like data mining, networking, image processing, grid computing, cloud computing etc are used for analyzing the different types of data which are used in various domains. Many new techniques, algorithms, concepts and methods have been proposed by the researchers for analyzing the static data sets. In this digital era, after the development of mobile and wireless technologies provides a new platform in which people may share their information through social media sites for e.g. face book, twitter and google+ [3]. In these places, the data may be arrived continuously and it cannot be stored in computer memory because the size of the data is huge and it is considered as “Big Data”. This situation also created a problem about how to perform data analysis for this dynamic datasets since the existing algorithms and their solutions are not suitable for handling the big data. This situation has raised concerns about the requirement of development of new techniques, methods and algorithms [1][2].

The term 'Big Data’ came into view for first time in 1998 in a Silicon Graphics (SGI) by John Mashey The growth of big data needs to increase the storage capacity and processing power. Frequently large amounts of data (2.5quintillion) are created through social networking [1]. Big data analytics are used to examine these large amounts of data and identifies the hidden patterns and unknown correlation. Two technologies are used in big data analytics are NoSQL and Hadoop. NoSQL is a non- relation or non SQL database solution, examples are HBase, Cassandra and mongoDB [2]. Hadoop is an eco software package which includes HDFS and MapReduce. Big data rely on structured data, unstructured data and semi-structured data to back up their decisions. Tools like SAS, R, and Matlab which supports the decisive analysis but these are not developed for the large datasets and either DBMS or Map Reduce can manage the data and which arrived at high rates. [2]

Big data applications have introduced the large scale distribution applications which work with large data sets. Data analysis problem plays a vital role in many sectors [1]. The existing software for big data applications like Apache Hadoop and Google’s map reduce framework, in which these applications generates a large amount of intermediate data [2]. There are many applications of big data such as manufacturing, bioinformatics, health care, social network, business, science and technology and smart cities [3]. Big data provides an infrastructure for Hadoop in bioinformatics which incorporates sequencing next generation, large scale data analysis and other biological domains. Parallel distributed computing framework and cloud computing combines with clusters and web interfaces. [1][3].

3. BIG DATA TECHNOLOGIES

Column-oriented databases

In column-oriented database stores data in columns rather than rows, which is used to compresses massive data and fast queries [3].

Schema-less databases

Schema-less databases are otherwise called as NoSQL databases. Database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. There are two types of database such as document stores and key value stores that stores and retrieves massive amount of structured, unstructured and semi structured data [3].

Hadoop

Hadoop is a popular open source tool for handling big data and implementated in MapReduce. It is java-based programming framework which supports large data sets in distributing computing. Hadoop cluster uses a master/slave structure. Distributed file system in hadoop helps to transfer data in rapid rates. In case of some node failure a distributed file system allows the system to continue the normal operation. Hadoop has two main sub projects namely Map Reduce and Hadoop Distributed File System (HDFS) [4].

Map Reduce

This is a programming paradigm which allows execution scalability against thousands of servers and server clusters for large task. Map reduce implementation consists of two tasks such as map task and reduce task. In the map task the input dataset is converted into different key/value pairs or tuples where as in reduced tasks several forms of output of map task is combined to form a reduced set of tuples.

HDFS

Hadoop distributed file system is a file system which extends all nodes in hadoop clusters for data storage. It links all the file system together on local node to make into a large file system. To overcome the node failures HDFS enhances the security by depicting data across multiple sources [4].

Hive

Hive is a data warehousing infrastructure which is built on hadoop.It has different storage types such as plain text, RC file, Hbase, ORC etc. Built-in user-defined functions are used to handle dates, strings and other data mining tools It is SQL-like Bridge that allows BI application to run queries against Hadoop clusters [4].

Storage Technologies

To store huge volume of data, efficient and effective techniques are required. The main focus of storage technologies are data compression and storage virtualization [5].

HBase

HBase is a scalable distributive database which uses Hadoop distributed file system for storage. It supports column-oriented database and structure data [5].

Chukwa

Chukwa analysis monitors large distributed system and it adds required semantics for log collections and it uses end to end delivery model [5] .

4. RESEARCH ISSUES IN BIG DATA

Big data has three fundamental issue i.e. storage issues, management issues and processing these issues exhibits a massive set of technical research problems whereas storage issue deal with when a quality of data is exploded, each and every time it creates new storage medium. Moreover data is being created generally in every place, for example, social media, 12+ Tbytes of tweets are growing every day and typically re-tweets are 144 per tweet. The next issue is management issues, which are difficult problem in big data domain. If the data is distributed geographically it can be managed and owned by multiple entities. Digital data collection is easier than manual data collection where digital data represents the methodology for data collection. Data qualification focuses on missing data or outliers rather on validating each item [5]. Hence new approaches are needed for data qualification and data validation. In processing issue concerns about how to process 1K petabyte of data which requires a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely information. Many issues in big data can be resolved by e-science which requires grid and cloud computing.

4.1 Big Data Classification

Data classification is the process of organizing data into categories for its most effective and efficient use. A well-planned data classification system makes essential data easy to find and retrieve. There are three primary and these aspects of data classification namely methods, domains and variations. Methods describes common techniques used for classification examples are probabilistic methods, decision trees, rule-based methods, instance-based methods, support vector machine methods and neural networks [2]. Domains examine specific methods used for data domains such as multimedia, text, time-series, network, discrete sequence and uncertain data. It also covers large data sets and data streams due to the recent importance of the big data paradigm [4]. Variations in classification process discusses ensembles, rare-class learning, distance function learning, active learning, visual learning, transfer learning, and semi-supervised learning as well as evaluation aspects of classifiers [5].

Classification of types of big data is divided into three categories namely Social Networks, Traditional Business systems and Internet of Things [6]. Social Networks (human-sourced information) contains information which is the record of human experiences, previously recorded in books and works of art and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned such as social networks (Facebook, Twitter etc), blogs and comments, personal documents, pictures, instagram, flicker, picasa, videos, YouTube, internet search engine, mobile data content, text messages, user-generated maps and e-mail. Traditional business systems are process-mediated data, these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context [4].

Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured data are stored in relational database systems.

Some sources belonging to this class may fall into the category of "Administrative data", i.e. data produced by Public Agencies, Medical and health records [5]. Data produced by businesses are Commercial transaction data, Banking/stock records, E-commerce, Credit cards, etc. The last classification is Internet of Things (machine-generated data)**:** derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured [6]. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed beyond traditional approaches. Data from sensors are divided into fixed sensors, home automation, weather/pollution sensors, traffic sensors/webcam, scientific sensors, videos, mobile sensors (tracking) like mobile phone location, cars, satellite images and data from computer system logs and web logs [5][6].

Big data are classified into different categories to understand their characteristics. The classification is based on five aspects: data sources, content format, data stores, data staging and data processing. This is represented in Figure 2 [5]. Each classification requires new algorithms and techniques for performing classification tasks efficiently in big data domain.

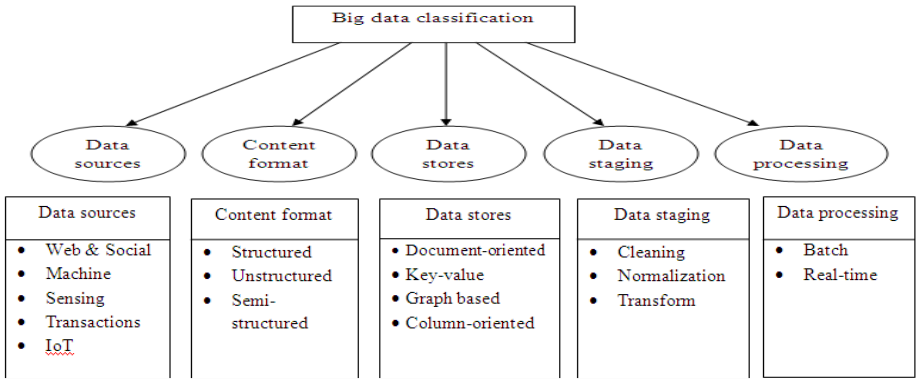

Figure 2. Big Data Classification

Data source is nothing but data is collected from different sources. Some of the important data sources are web and social media, machine generated data, sensor data, transaction data and internet of things (IoT). Social media contains volume of information which is generated using URL (Uniform resource language) to share or exchange information in virtual communities and network for example face book, twitter, and blogs. In Machine generated data information are automatically generated from both hardware and software, for example computers and medical devices. Sensor data are collected from various sensing devices and these are used to measure physical quantities [7]. Transaction data involves a time dimension to illustrate the data, for example, financial and business data. Finally IoT represents set of objects they are identified uniquely as a part of internet i.e. smart phones and digital cameras.

Content format has three formats namely structured, unstructured and semi-structured. Structured format is often managed by SQL and data resides in affixed field within a record or a file. Unstructured format is often includes text and multimedia content, it is opposite to structured data. Semi-structure format does not reside in a relational database [7]; it might include XML documents and NOSQL database. Data stores classified into four categories such as document-oriented, key-value, column-based and graph based. Document-oriented data are designed to store

and collect information and supports complex data whereas column –based data stores data in row and column format. Key-value data store is an alternative to relational database which is designed to scale very large data set and it can be accessed and stored easily. Finally graph based data stores are designed to represent the graph model with edges, nodes and properties and these are related to one another [8].

Data staging is classified into three forms; cleaning, transforming and normalization. Cleaning identifies the incomplete data. Normalization is a method which minimizes redundancy. Transform data staging which transfers data into suitable form. Finally data processing is based on two types namely batch and real-time [9]. From the above analysis it is observed that content format is suitable for all types of data like structure ,un structure and semi structured

4.2 Clusters in Big Data

A group of the identical elements closely together is known as clustering. Data clustering are also known as cluster analysis or segment analysis which organizes a collection of n objects into a partition or a hierarchy. The main aim of clustering is to classify data into clusters such that objects are grouped in the same cluster when they are "similar" according to similarities, traits and behavior. The most commonly used algorithms in clustering are partitioning, hierarchical, grid based, density based, and model based algorithms. Partitioning algorithms is called as the centroid based clustering. Hierarchical algorithms also called as the connectivity based clustering. Density based clustering is based on the concept of data reachability and data connectivity. Grid based clustering is based on the size of the grid instead of the data. Model based clustering depends upon the probability distribution. Figure 2 represents the processing of data clustering [4],[5],[6].

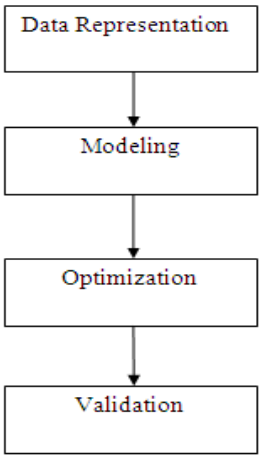

Figure.2.Processes of Data Clustering

The Clustering algorithm deals with a large amount of data. It is the most distinct feature that demands specific requirements to all classical technologies and tools used. To guide the selection of a suitable clustering algorithm with respect to the Volume property, the following criteria are considered: size of the dataset, handling high dimensionality and handling outliers/noisy data. Variety: refers to the ability of a clustering algorithm to handle different types of data (numerical, categorical and hierarchical). It deals with the complexity of big data [7]. To guide the selection of a suitable clustering algorithm with respect to the Variety property, the following criteria are considered: Type of dataset and clusters shape. Velocity: refers to the speed of a clustering

algorithm on big data. Big Data are generated at high speed. To guide the selection of a suitable clustering algorithm with respect to the Velocity property shows the criteria and Complexity of algorithm. Many clustering algorithms are available few are listed below. [5][6][7].

- K-means
- Gaussian mixture models
- Kernel K-means
- Spectral Clustering
- Nearest neighbor
- Latent Dirichlet Allocation.

Figure 3 shows the various existing clustering algorithms.

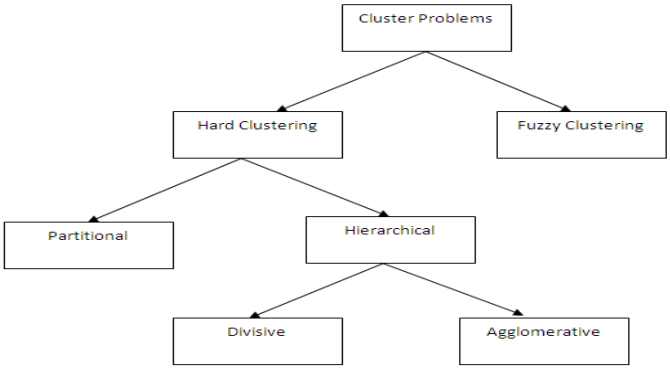

Figure 3.Clustering Algorithm

4.3 Association Rules

Association rules are (if/then) statements that help to uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then) [8]. An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting customer behaviour [9]. Association rule mining finds the frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases and other information repositories. Association rules are used for market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc. Some of the properties of association rules are how items or objects are related to each other and how they tend to group together, simple to understand (comprehensibility), provide useful information (utilizability), efficient discovery algorithms (efficiency). Different types of association rules are based on types of values handled i.e. Boolean association rules and Quantitative association rules. Levels of abstraction are divided into either single-level association rules or multilevel association rules. Dimensions of data involved into single-dimensional association rules and multidimensional association rules [10].

4.4 BIG DATA VISUALIZATION

Big Data visualization is a processing by which numerical data are converted into meaningful 3-D images. It is a presentation of pictorial or graphical format and which depends upon visual representation such as graphics, tables, maps and charts which helps to understand more quickly and easily. There are many tools in big data visualization namely polymaps, nodebox, flot, processing, tangle, SAS visual analytics, linkscape, leafelet, crossfilter, openlayer [9]. Visualization techniques are classified into three different ways (i.e.) based upon the task, based upon the structure of the data set or based on the dimension. Visualization can be classified as whether the given data is spatial or non spatial or whether the displayed data to be in 2D or 3D.Visualization components can be either static or dynamic [10]

Visualization is used for spatial data and non-spatial data. For representing 2D or 3D data also various visualization mechanisms are applied. The processing of data in visualization system can be batch or interactive. The batch processing is used for analysis of set of images. In data visualization interaction the user can interact in variety of ways which includes browsing, sampling, querying and associative..Various methods are available in data visualization and it is based on type of data, there are three types of data: Univariate, Bivariate and Multivariate. Univariate measures the single quantitative variable, it characterizes distribution and it is represented by two methods they are histogram and pie chart. Bivariate constitutes the sample pair of two quantitative variables, they are related with each other. They are represented using scatter plots and line graph methods .Multivariate data represents multidimensional data and it is represented by icon based method, pixel based method and dynamic parallel coordinate system.[8][9][10]

5. TOOLS FOR DATA VISUALIZATION

5.1. Dygraphs

Dygraphs is a fast, versatile open source JavaScript charting library .It is highly personalized and designed to interpret dense data sets, it works in all browsers and it can be zoomed on mobile and tablet devices. Few characteristics of Dygraphs they are used to handle huge datasets, highly customizable; highly compatiable and they give strong support for error bar or confidence interval. Dygraph chart is displayed in figure 2 [22].

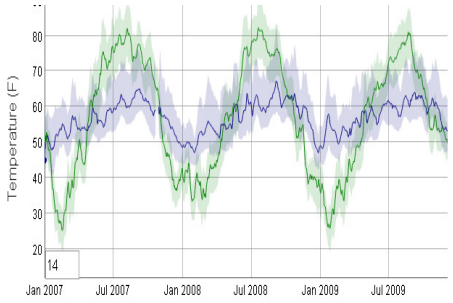

Figure.2. Dygraph Chart

5.2. ZingChart

ZingChart is a powerful charting library and they have ability to create charts dashboards and infographics. It is featured with -rich API set that allow user to built interactive Flash or HTML5

charts. It provide hundreds of chart variation and many methods For Example Bar, Scatter, Radar, Piano, Gauge, Sparkline, Mixed, Rank flow and word cloud. Figure shows Zing chart.[22]

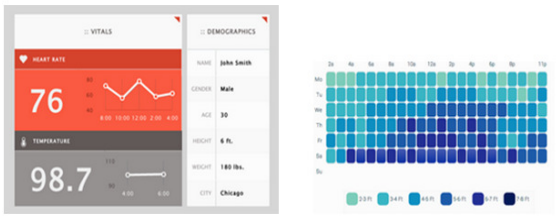

Figure.3. Zing Chart

5.3. Polymaps

Polymaps is a free java script charting library for image and vector- tiled maps using Scable Vector Graphics (SVG).They provide dynamic and interactive maps in web browsers. Complex data sets can be visualized using polymaps and offers multi-zoom functionality. The characteristics of polymaps are it uses Scalable Vector Graphics (SVG) and the Basic CSS rules are used and its imagery in spherical Mercator tile format. Figure 4 shows the layout of Polymaps. [22]

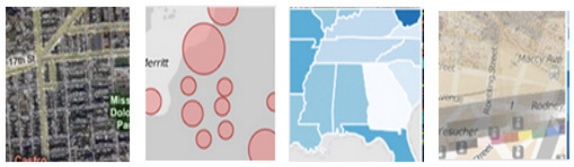

Figure.4. Polymaps

5.4. Timeline

Timeline is **a** different tool which delivers an effective and interactive timeline that responds to the user's mouse, it delivers lot of information in a compressed space. Each element can be clicked to reveal more in-depth information; it gives a big-picture view with full detail. Timeline is demonstrated in figure 5.[22]

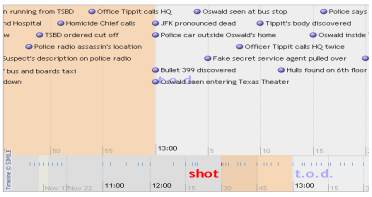

Figure.5. Timeline

5.5. Exhibit

Exhibit is an open-source data visualization and it is developed by MIT, and Exhibit makes it easy to create interactive maps, and other data-based visualizations measure oriented towards teaching or static/historical based mostly knowledge sets like birth-places of notable persons. Sample model is shown in figure 6.[22]

Figure .6.Sample Model of Exhibit.

5.6. Modest Maps

Modest Maps is a lightweight, simple mapping data visualization tool for web designers. It has highest performance and compatibility with new technology and has well designed codes which are tested and deployed widely. Modest Map is given in figure 7 [22].

Figure.7 Modest Map

5.7. Leaflet

Leaflet is an open source java script tool developed for interactive data visualization in an HTML5/CSS3.Leaflet tool is designed with clarity, performance and mobilization. Few visualizing features are given zooming and planning animation such as multi touch and double tap zoom, hardware acceleration on IOS and utilizing CSS3 features. Figure 8 shows Leaflet structure [22].

Figure.8 Leaflet structure

5.8. Visual.ly

Visual.ly is a combined gallery and infographic generation tool. It provides simple toolset for building data representations and platform to share creations. This goes above pure data visualisation, representation of visual.ly is displayed in figure 9[22].

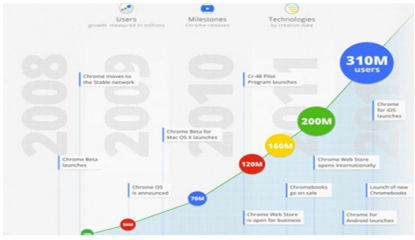

Figure.9 Visual.ly

5.9. Visualize Free

Visualize Free is a free visualize analysis tool that allows user publicly available datasets, or upload own, and build interactive visualizations to define the data. Visualize free works with Flash; HTML5.Visualization is a perfect tool for sifting with multi dimensional data. Figure 10 shows the representation of visual free [22].

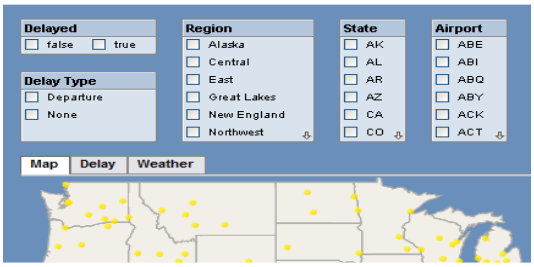

Figure.10 Representation of Visual Free

5.10. jQuery Visualize

jQuery Visualize Plugin is an open source charting plugin and creating accessible charts and graphs. It uses HTML5 canvas elements and generates bar, line, area, pie charts for visualization. Figure 11 demonstrates the jQuery Visualize [22].

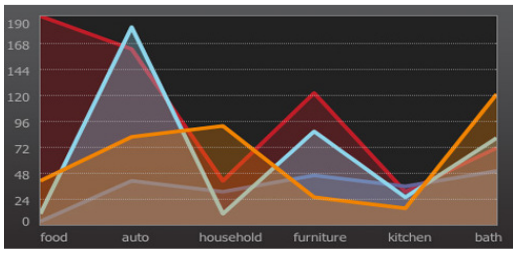

Figure.11 Demonstration of jQuery Visualize

5.11. JqPlot

JqPlot is a good tool for line and point charts. Features of JqPlot provide different style options, customized formatting, and automatic line computation, highlighting tooltips and data points. It has the ability to generate automatic trend lines and interactive points according to dataset. Figure 12 represents JqPlot [22].

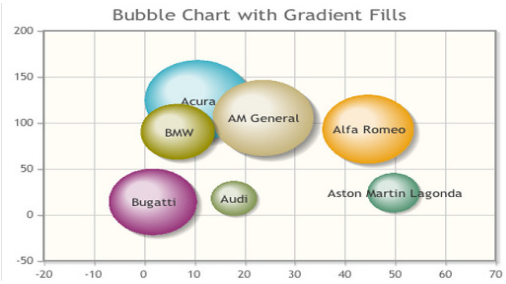

Figure.12 jqPlot

5.12. Many Eyes

Many Eyes was developed by IBM It is a web based tool and it is used for structure and unstructured data analysis, Tool data set is in text file and uploaded from spreadsheet Many Eyes tool allows user to build visualizations quickly from available or uploaded data sets. Demonstration of Many eyes is shown in figure 13 [22].

Figure.13 Many Eyes

5.13. JavaScript InfoVis Toolkit

JavaScript InfoVis is a open source Toolkit which includes a compatible structure. It allows user to download absolutely necessary data and displays chosen data for visualizations. The toolkit has a number of different styles and classy animation effects. JavaScript InfoVis is demonstrated in figure 14 [22].

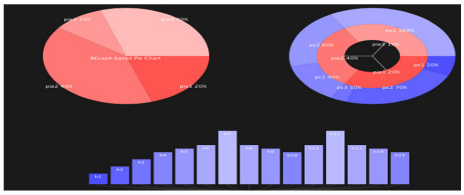

Figure.14 jqPlot

5.14. JpGraph

JpGraph is an object-oriented graph creating library for PHP-based data visualization tool. It Generates drill down graphs and large range of charts like pie, bar, line, scatter point and impulse . Some features of JpGraph are web friendly; automatically generates client-side image maps. It supports alpha blending, flexible scale, support integer, linear, logarithms and multiple Y- axes. Figure 15 shows the JpGraph representation [22].

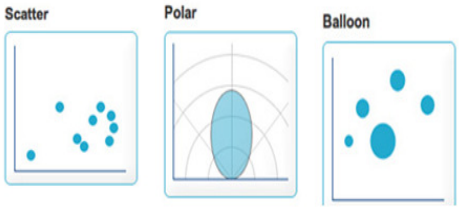

Figure.15 Representation of JpGraph

5.15. Highcharts

Highcharts is an open source JavaScript charting library and has a massive range of chart options. The output is performed using SVG and VML. The charts are animated, displayed automatically and live data streams are supported by framework. Figure 16 displays Highcharts [22].

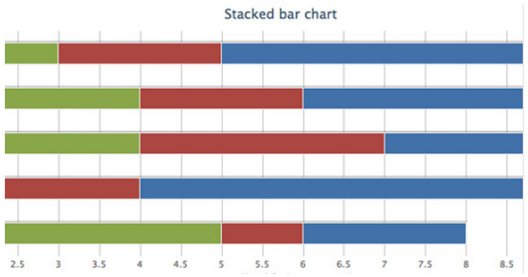

Figure.16 Representation of Highcharts

5.16. R

R is an effective free software tool for statistical computing and graphics and integrated with data handling, Calculation and graphical display. R is similar to S language which handles effective data and storage. R has its own documentation like LaTeX format. Figure 17 shows the layout of R [22]

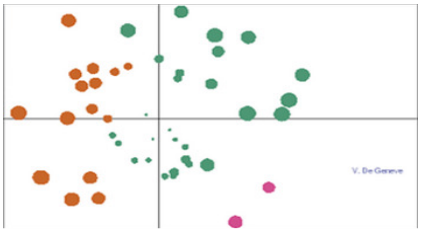

Figure.17 Layout of R

5.17. WEKA

WEKA is an open source software and collection of machine-learning algorithms assigned for data-mining, Weka is a excellent tool for classifying and clustering data using many attributes. It explores data and generates simple plots in a powerful way. Figure 18 explains the representation of WEKA [22].

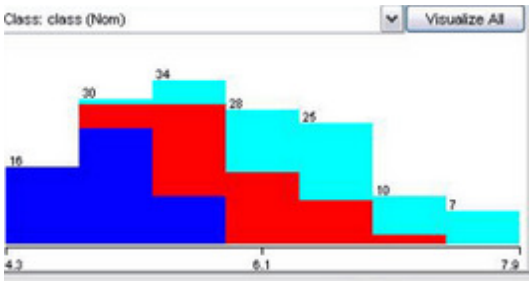

Figure.18 Representation of WEKA

5.18. Flot

Flot is a specially designed for plotting library for jQuery, it works with all common web browsers and has many handy features. Data is animated and fully controlled in all the aspects of animation, presentation and user interaction. Interactive charts can be created using Flot tool. Figure 19 shows the demonstration of Flot [22].

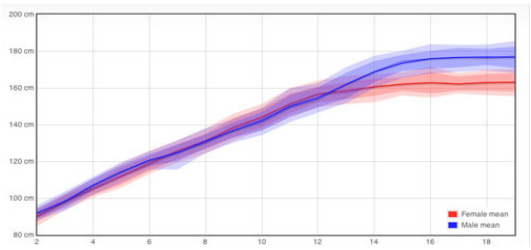

Figure.19 Demonstration of Flot

5.19. RAPHAEL

RAPHAEL is a tool which provides a wide range of data visualization options rendered using SVG.It works with vector graph on web. RAPHAEL tool can be easily integrated with own web site and codes. The supporting web browsers for RAPHAEL tools are Internet Explorer6.0+, firefox 3.0+, Safari 3.0+, Chrome 5.0+ and Opera 9.5.Model of RAPHAEL is shown in figure 20 [22].

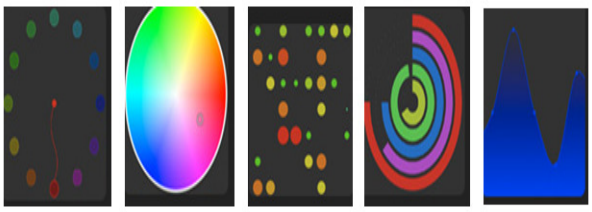

Figure.20 Models of RAPHAEL

5.20. Crossfilter

Crossfilter is an interactive GUI tool for massive volume of data and it reduces the input range on any one chart. This is a powerful tool for dashboards or other interactive tools with large volumes of data .It displays data, but at the same time, it restricts the range of the data and displays the other linked charts. Representation of Crossfilter is shown in figure 21 [22].

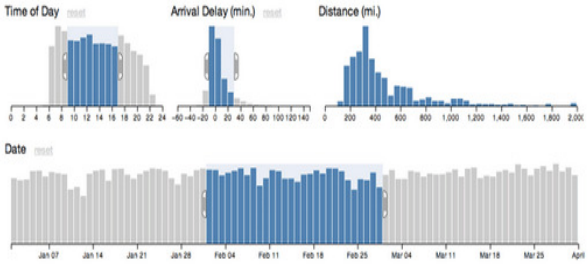

Figure.21 Representation of Crossfilter

Table 2 represents the characteristics, advantages and disadvantages of the data visualization tools.

Table 2: Comparison of data visualization tools

S.No	Tool Name	Characteristics	Advantages	Dis-Advantages
1.	Crossfilter	Exploring large multivariate datasets	Extremely fast	Restricts the range of the data and displays the other linked charts
2.	Dygraphs	Handles huge data sets Zoom able charts	Interactions are easily discoverable	Supports only limited web browsers
3.	Exhibit	easily create Web pages with advanced text search and filtering functionalities, with interactive maps, timelines and other visualization	Can easily sort out data and present them any way we like	Newcomers unused to coding visualizations, it takes time to get familiar with coding and library syntax
4.	Flot	Plot categories and textual data	Supports lines, plots, filled areas in any combination	Not applicable for large dataset
5.	Highcharts	It is a very powerful tool	It supports backward compatibility with IE8	Output is performed using only SVG and VML
6.	JavaScript InfoVis Toolkit	Uses high polished graphics for data analysis	Unique chart types Ability to interact with animated charts and graphs	This might not be a good fit for users in an organization who analyze data but don't know how to program.
7.	JpGraph	It support various types plot types like, filled line, line error, bar ,box, stock plots	Supports alpha blending, advanced gantt-charts	Difficult for new users .It takes time to get familiar with coding
8.	JqPlot	It support line pie and bar charts	It is the extension of JQuery which meets all data visualization needs	Not suitable for rapid visualization

9.	jQuery Visualize	Focus on ARIA support, user friendly to screen readers	Developers can completely separates java script code from HTML.	It uses only HTML5 for designing.
10.	Leaflet	Eliminates tap delay on mobile devices	Works on all major desktop and mobile browsers	Difficult for new users.
11.	Many Eyes	Multiple ways to display data	Upload data sets for public use	It is difficult to use in large dataset
12.	Modest Maps	Used with several extensions, such as MapBox.js, HTMAPL, and Easey	Designed to provide basic controls and building mapping tools	Support only limited applications.
13.	Polymaps	display complex data sets	Uses Scalable Vector Graphic	It is ideal only for zooming in and out of form levels
14.	R	R is a general statistical analysis platform	R also results in graphs, charts and plots	It is difficult to use in large dataset
15.	RAPHAEL	Multi-chart capabilities	Create a variety of charts, graphs and other data visualizations	It is not easy to customize
16.	Timeline	Display events as sequential time lines	Embed audio and video in timelines from 3rd-party apps	Build timelines using only Google Spreadsheet data
17.	Visual.ly	Infographic generation tools	It is specially designed to develop simple toolset representation	Difficult for new users
18.	Visualize Free	Upload data in Excel or CSV formats	Drag-and-drop components to build visualizations Uses Sandboxes for data analysis	Not applicable for large dataset
19.	WEKA	WEKA is a collection tools for data pre-processing, classification, regression, clustering, association, and visualization	Free availability Portability Comprehensive collection of data preprocessing and modeling techniques	Sequence modeling is not covered by the algorithms included in the Weka distribution. Not capable of multi-relational data mining. Memory bound

20.	ZingChart	It supports larger dataset ranging from 10k to 5000k+ High performance	It uses more than 100 types charts to fit the data	Difficult to customize

6. VISUALIZATION ALGORITHMS

Digital data are visualized in digital form with the help of visualization concept. There are various data source to display the digital data with the help of equipment for example Antennas. Visualization has issues in digital signal to overcome this problem algorithms are applied to raw data, digital 3D data and various digital equipments produces digital dataset [19]. Data should be represented in discrete form; data objects are classified into two categories organizing structure and data attribute whereas organizing structure determines the spatial location of the data and describes the topology and geometry on which data is illustrated and they are specified as cells and points. Cell is defined as an ordered sequences of points and the type of cells namely vertex, poly vertex, triangle, line, poly line, pixel, voxel and tetraheder which characterizes the sequence of points and number of points which specifies the size of the cell. Data attributes determines the format of the data [20]. Data attributes are commonly described as scalar, vector and texture coordinates etc. The Visualization algorithm has two sets of transformations. In Figure 3 visualization algorithm is explained. First set of transformation converts the data or sub data into virtual scene. These transformations are applied to structure and data type of the data. The second transformation is done when the virtual scene is created it consists of geometrical objects, textures and computer graphics. Transformations are applied to form images [21].

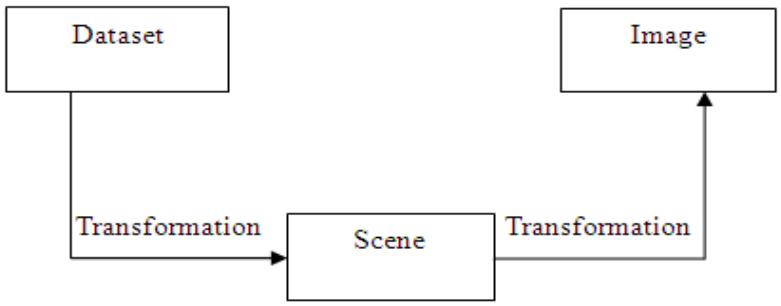

Figure 3. Visualization Algorithm

The main objectives of visualization are understanding data clearly which Is recorded, Graphical representation, Placing the search query to find the location of the data, Discovering hidden patterns, perceptibility of data items [22].The characteristics of transformation algorithm are explained in Figure 4, transformation algorithms are characterized by structure and type. The Structure transformation has two formats such as topology and geometric. Topology represent changes in topology for example conversion of polygon data into unstructured format.Geometric format represent changes in coordinate , scaling,rotation,translation and geomentry[22]. Few examples for transformations are scalar algorithm, vector algorithm, and modeling algorithm.

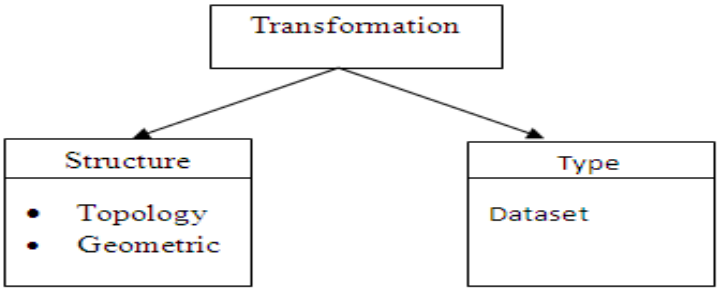

Figure 4. Characteristics of Transformation Algorithm

Big data visualization has overcome the five massive challenges of Big Data:

1. Increasing data acceleration.
2. Understanding the data
3. Delivering the data quality
4. Significant results are displayed
5. Handling outliers

Challenges of big data

- Issues related with storage and data processing
- Data obtaining and information sharing
- Human recognition and restricted screen space
- Acquire useful information
- Storing massive amount of data
- Combining data to extract meaningful information
- Querying ,data modeling and analysis
- Elaboration of data

7. CONCLUSION AND RESEARCH TRENDS

This paper is envisioned with big data tools, techniques, issues related with big data. It also focused and provided the information about how to perform big data visualization. Research trends in big data, operations of big data such as storage, search and retrieval, big data analytics and computations on big data are discussed, where storage requires managing capacity, finding out best collection and retrieval methods and synchronizes both IT and business team, it also focuses on complex security and privacy issues. Big data analytics focuses on tools, algorithm, and architecture which perform proper analysis and transfer large and massive volume of data. Computing deals with processing, transforming, handling and information storage. This paper has reviewed basic concepts of big data, its applications and research issues.

REFERENCES

1. Neelam Singh, Neha Garg, Varsha Mittal, Data – insights, motivation and challenges, Volume 4, Issue 12, December-2013, 2172, ISSN 2229-5518 2013.
2. Karthik Kambatlaa, Giorgos Kollias b, Vipin Kumarc, Ananth Gramaa, Trends in big data Analytics, (2014) 74 2561–2573
3. Francis X. “On the Origin(s) and Development of the Term \Big Data"_ Francis X., 2012
4. Venkata narasimha inukollu1, sailaja arsi1 and srinivasa rao ravuri3 Security issues associated with big data in cloud computing Vol.6, No.3, May 2014
5. Matzat1, Ulf-Dietrich Reips2,3 1 Eindhoven “Big Data” 2012, 7 (1), 1–5 ISSN 1662-5544
6. Hong Kong, Park Shatin, Mining Big Data: Current Status, and Forecast to the Future
7. Anil K. Jain Clustering Big Data, 2012
8. Daniel Keim Big-Data Visualization.
9. Hsinchun Chen Business Intelligence And Analytics: From Big Data To Big Impact AZ 85721, OH 45221-0211 U.S.A. Mack Robinson, GA 30302-4015.
10. Ibrahim Abaker Targio Hashema,n, Ibrar Yaqooba, Nor Badrul Anuara, Salimah Mokhtara, Abdullah Gania, Samee Ullah Khanb, The rise of “big data” on cloud computing: Review and open research issues. 2014
11. Edd Dumbill, Making Sense of Big Data
12. Silva Robak , prof. Z. Szafrana, Zielona Góra Uniwersytet Zielonogórski Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management 2014 249 DOI: 10.15439/2014F472
13. C.L. Philip Chen , Chun-Yang Zhang Data-intensive applications, challenges, techniques and technologies: A survey on Big Data 275 (2014) 314–347
14. Chaitanya Baru,1 Milind Bhandarkar,2Raghunath Nambiar,3 Meikel Poess,4and Tilmann Rabl Survey of Recent Research Progress and Issues in Big Data 2013.
15. Tackling the Challenges of Big Data 2014.
16. Stephen Kaisleri_SW. Alberto Espinosa Big Data: Issues and Challenges Moving Forward Stephen Kaisleri_SW. Alberto Espinosa 013 46th Hawaii International Conference on System Sciences
17. Challenges and Opportunities with Big Data A community white paper developed by leading researchers across the United States 1819
18. Danyang Dua , Aihua Lia*, Survey on the Applications of Big Data in Chinese Real Estate Enterprise 1st International Conference on Data Science,2014
19. Shilpa, Manjit Kaur challenges and issues during visualization of big data , International Journal For Technological Research In Engineering Volume 1, Issue 4, December - 2013 ISSN (Online) : 2347 - 4718
20. http://fellinlovewithdata.com/research/the-role-of-algorithms-in-data-visualization
21. http://prosjekt.ffi.no/unik-4660/lectures04/chapters/Algorithms2.html
22. http://www.creativebloq.com/design-tools/data-visualization-712402