International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

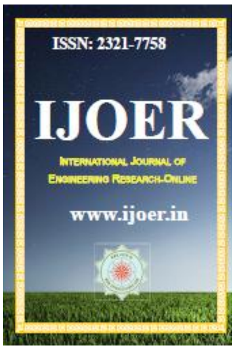
Vol.4., Issue.4., 2016
(July-August)

REVIEW ARTICLE

ISSN: 2321-7758

BIG DATA VISUALIZATION

MOHSIN L. SHAIKH
Navinchandra Mehta Institute of Technology and Development,
Dadar (W), Mumbai.



ABSTRACT

With growing technologies in the world, the amount of data being handled and processed has increased tremendously. Big Data analytics plays a very significant part in reducing the size of the data as well as the complexity in applications that are being used for Big Data. Big Data Visualization is an important approach in creating meaningful visuals and graphical representations from the Big Data that help in better decision making and that give a clear insight into the data. Visualization, Big Data, Big Data Visualization, data visualization techniques are some of the topics that are discussed in this paper and examples for visualizations have been presented as well.

Keywords—Visualization, Data processing, Data analytics, Big Data, Interactive visualizations.

I. VISUALIZATION

A. *WHAT DOES VISUALIZATION MEAN?*

Visualization is any technique for creating images, diagrams or animations to communicate a message. Visualization is one of the most primitive forms of communication known to man

B. *WHY DO WE USE VISUALIZATION?*

Visualization through pictorial imagery has been an effective medium to communicate abstract and concrete ideas since ancient times. Today Visualization plays a significant role in applications in the field of science, engineering, interactive media, medicine, education etc. The field of computer graphics is a typical example of application of visualization. A visual can communicate more information than a table in a much smaller space. This characteristic of a visual makes it more effective than typical tables used for representing data.

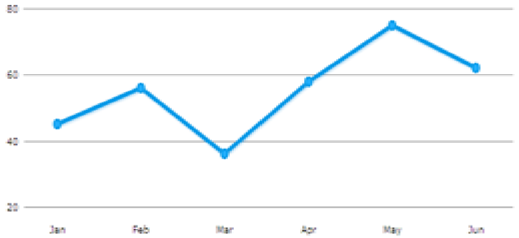
Edward Tufte, a data visualization expert, says "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space". This characteristic of visualizations is what makes it vital to businesses.

Ex: The table below shows a sample sales data of a particular brand of sports bikes for the months of January to June.

Month	Jan	Feb	Mar	Apr	May	Jun
Sales	44	58	37	59	74	61

The same information can be visualized in a second or two using a visualization chart which is much more intuitive compare to a tabular piece of information.

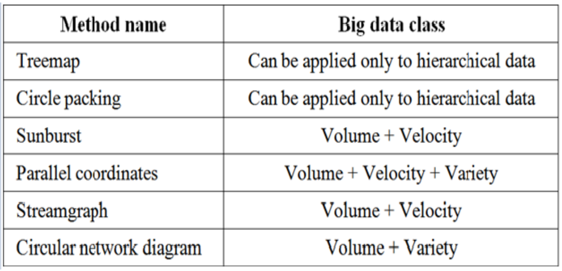
C. *APPLICATIONS OF DATA VISUALIZATION*



Data Visualization is representing data in some systematic form including attributes and variables for the unit of information. Data Visualization is a subcategory of visualization dealing with statistical graphics and geographical data that is abstracted in schematic form. Visualization based data discovery method allows business users to mix up data from different data sources that cannot be easily compared which in turn help in generating custom views.

Visualizations can be of various types depending on the type of data representation and domain. Scientific visualization, product visualization, information visualization, visual analytics, knowledge visualization are a few examples of different type of visualizations.

The following table shows the benefits of data visualization according to the respondent percentages of a survey.



Method name	Big data class
Treemap	Can be applied only to hierarchical data
Circle packing	Can be applied only to hierarchical data
Sunburst	Volume + Velocity
Parallel coordinates	Volume + Velocity + Variety
Streamgraph	Volume + Velocity
Circular network diagram	Volume + Variety

II. BIG DATA

I. *WHAT IS BIG DATA?*

Big Data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. The term often refers simply to the use of predictive analytics or certain advanced methods to extract value from data and often to a particular size of data set.

II. *WHAT DOES BIG DATA INCLUDE?*

The term Big data includes data produced by various applications, devices and systems. The following mentioned topics come under the umbrella of Big Data.

A. *Social Media Data*

Social Media such as Whatsapp, Twitter and Facebook collect data in the form of information and views posted by the large number of its users across the globe.

B. *Search Engine Data*

These are the largest sources of information collection as they retrieve tremendous amount of data from different databases.

C. *Stock Exchange Data*

Stock Exchanges across the world hold data about the bought and sold decisions made on a share of different companies made by the stakeholders/customers.

D. *Black Box Data*

Black box is a component of aerial transport devices such as airplanes, helicopters, jets etc. It captures various details such as voices of the flight crew, aircraft performance information, recording of the earphones as well as microphones.

Hence, Big Data includes huge volumes of an extensible variety of data. The data in it mostly belongs to the following three types.

A. *Structured Data : Relational Data*
B. *Semi-Structured Data : XML Data*
C. *Unstructured Data: Word, Excel Sheets PDF's, Media Logs etc.*

III. *CHARACTERISTICS*

Big data can be described by the following characteristics:

A. *Volume of data*

In case of Big Data the amount of data (information) that is produced is very significant. It is the size of the data which determines the worth and prospects of the information beneath that can be contemplated and it helps in determining whether it can actually be considered Big Data. The name 'Big Data' itself encloses a word that is related to size and hence the characteristic.

B. *Velocity of data*

The word 'velocity' in the case speaks of the speed of generation of data or how fast the data is generated and administered to come across the needs and the experiments which lie ahead in the path of growth and development.

C. *Variety of data*

The next characteristic of Big Data is its variability. This means that the domain or the category to which Big Data belongs to is also a very crucial detail which needs to be recognized by the data analysis. This helps the people, who are closely analyzing the information and are related with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

D. Variability of data

Variability of data refers to the inconsistency that can be crucial by the details at period intervals, thus affecting the process of managing and handling the data effectively. This is a scenario that can be an issue to those who are examining and processing the data.

E. Veracity of data

The amount of the information actuality captured can vary greatly as the data can span across various domains as well as sources. Accuracy of analysis depends on the veracity of the source data. Thus making it important to understand the sources of data.

IV. BENEFITS OF USING BIG DATA

Big Data is emerging as one of the most important technologies in the world. A few well known benefits of Big Data are listed below.

1. Using information from previous medical record of a patient hospital's are providing a quicker and better services.
2. Using information from social media, product companies and retail organizations get to know the consumer preferences and perceptions about the products.
3. Marketing agencies use information from social networking sites such as Facebook, Twitter to understand response of the people towards campaigns, advertisements and promotions.
4. Search engines record information from the users to filter the searches based on the previous choices and to show results that are most relevant.

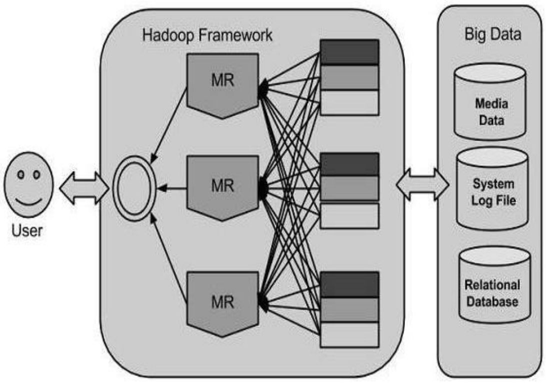
V. BIG DATA TECHNOLOGIES

Big data technologies are crucial in providing accurate analysis which may lead to better decision making in greater operational efficiencies, reducing costs and reducing the risk for business.

To fully harness the power of Big Data , infrastructure that can manage and process huge volumes of structured and unstructured data in real time as well as systems that provide data privacy and security are required.

Hadoop: Hadoop is an Apache open source framework written in java that allows distributed processing of large data sets across clusters of computers using simple programming models. A Hadoop framework based application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed such that it can scale from a single server to thousand of machines, each offering local computation and storage.

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data.



III. BIG DATA VISUALIZATION

I. NEED FOR BIG DATA VISUALIZATION

As we are all aware the benefits of visualization over traditional reports, following are the few reasons which specify the need for Big Data Visualization.

1. Visually represented information is easy to process.
2. Visualizing the Big Data helps in looking at information from different perspectives.
3. Visualization helps in drilling down the Big Data to a very minute level where in information can be analyzed.
4. Patterns in data can be identified and can be used to make predictive analysis.
5. The more dimensions of data visualized, the higher are the chances of recognizing potentially interesting correlations, or outliers.

II. CHALLENGES IN BIG DATA VISUALIZATION

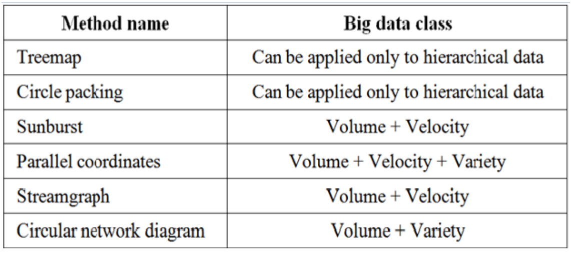
1. Information/Data Loss: Data can be lost during the data harmonization (creating meaningful data from Big Data) process, this could lead to shortcomings in representing information properly.
2. Large image perception: Data visualization techniques are not only limited by the resolution and the aspect ratio of the devices but also the physical perception limits.
3. Visual noise: dataset objects are mostly relative to each other. Hence, they cannot be divided as separate objects on the screen.
4. High performance requirements: querying large data sets can result in high latency and may cause disruption in fluent interaction. Thus the high performance requirements.

III. BIG DATA VISUALIZATION METHODS

In the current Big Data scenario, visualization techniques should provide an overview first, then allow zooming in and filtering information to provide insights. Visualization plays an important role in understanding the consumers and getting a complete view of the choices and preferences of consumers with the help of Big Data. Designing a visualization tool which handles indexing efficiently in Big Data is quite challenging. Unstructured data forms such as tables, texts, graphs and other metadata must be dealt with by the visualization systems. Due to the limitations in bandwidth visualizations should move closer to data to extract meaningful information from it.

Big data visualization can be performed using a varied number of approaches such as one or more views per representation, dynamic changes in the number of factors and filters.

Visualization techniques can be classified based on data variety, data dynamics and volume of data. The following table shows the classification of the visualization techniques.



Method name	Big data class
Treemap	Can be applied only to hierarchical data
Circle packing	Can be applied only to hierarchical data
Sunburst	Volume + Velocity
Parallel coordinates	Volume + Velocity + Variety
Streamgraph	Volume + Velocity
Circular network diagram	Volume + Variety

A lot of the Big Data Visualization tools run on the Hadoop platform. Hadoop Distributed File System(HDFS), Hadoop YARN, Hadoop Common, Hadoop MapReduce are the various modules in Hadoop, these are efficient in analyzing Big Data but lack in the visualization aspect of the it. The following are the various tools that support visualization:

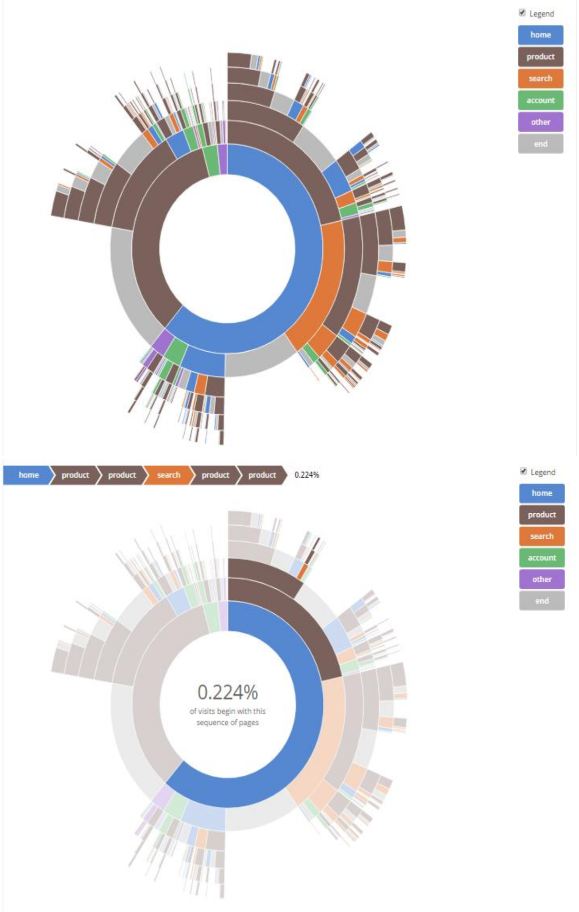
1. Tableau: It is a BI (Business Intelligence) tool that supports interactive and visual analysis of data. It has an in-memory data engine to accelerate visualization.
2. Flare: It is an ActionScript library for creating data visualization that runs in Adobe Flash Player.
3. Pentaho: It supports a variety of BI functions such as analysis, dashboard, enterprise-class reporting, and data mining.
4. Jasper Reports: It has a software layer that generates reports from Big Data storages.
5. ManyEyes: It is a visualization tool launched by IBM. Many Eyes is a public website where users can upload data and create interactive visualization.

Big Data processing tools often process ZB (Zeta Bytes) and PB (Peta Bytes) of data , but it is not easy to visualize ZB and PB data. Following are the Big Data Visualization tools that are currently used NodeBox, Google Chart API's, D3, visual.ly etc.

Following example is a Sequential Sunburst Chart created using D3.js library.

Ex: A good example is to summarize navigation paths through a web site. The visualization makes it easy to understand visits that start directly on a product page (e.g. after landing there from a search engine), compared to visits where users arrive on the site's home page and navigate from there. Where a funnel lets you understand a single preselected path, this allows you to see all possible paths .Interactive breadcrumb trail helps to emphasize the sequence, so that it is easy for a first-time user to understand what they are seeing.

A Sequential Sunburst Chart created using D3.JS.



When traced by the sequences that are rendered we can understand the various different patterns or scenarios in which a typical user passing through the websites navigates.

The sequence that is traced in the image given below is that of a user starting at the home page of the website moves on to a product then navigates further to a product, makes a search after that, then navigates further to products and different product makes up of 0.224% i.e. 0.224% of all the users navigating through the website follow this navigation pattern.

These kinds of patterns help in identifying the behavior of the users accessing and navigating across the website and can further be directed to products based on their previous choices. Such techniques are already used by search engines to refine the searches for users based on their previous browsing and search patterns.

V. CONCLUSION

Visualizations have been one of the most successful medium in conveying information for centuries and still has a great impact. The era of Big Data has begun. Visualizations not only help in gaining insight from the Big Data but lead to better decision making. The extension of some conventional visualization techniques to understand Big Data is not enough. With the increase in technology newer methods for data visualization should be developed so that the real power of Big Data can be harnessed to a much greater extent. Big Data Visualization and Analytics should be integrated to enhance the Big Data applications. Big Data has a lot of potential unrecognized piece of information that needs to be explored and visualizations will help in understanding it better.

Acknowledgment

I thank my colleagues from NMITD who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

I thank Mrs. Sulakshana Vispute and Fractal Analytics Pvt. Ltd. for all the guidance and assistance provided with my Research topic and the IEEE formats.

References

[1]. Intel IT Center, Big Data Visualization: Turning Big Data Into Big Insights, White Paper, March 2013, pp.1-14.

[2]. Using Visualization to understand Big Data By T. Alan Keahey, Ph D., IBM Visualization Science and Systems Expert.

[3]. Lidong Wang, Guanghui Wang, Cheryl Ann Alexander. Big Data and Visualization: Methods, Challenges and Technology Progress. Digital Technologies. Vol. 1, No. 1, 2015, pp 33-38. http://pubs.sciepub.com/dt/1/1/7

[4]. https://www.wikipedia.org/

[5]. Principles of Data Visualization – What we see in a visual White Paper

[6]. http://www.fusioncharts.com/whitepapers/principles-of-data-visualization/

[7]. NASSCOM Big Data Report 2012.

[8]. http://www.slideshare.net/AllAnalytics/data-visualization-techniques

[9]. http://www.tutorialspoint.com/

[10]. http://www.google.com