International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1173-1176 www.ijaem.net ISSN: 2395-5252

A Comparative Study of Big Data Visualization Tools and Techniques

1Khadija Begum, 2Md Mamunur Rashid

1Student, The People's University of Bangladesh, Dhaka 1207, Bangladesh
2Student,Pukyong National University,Nam-gu, Busan, South Korea.

Submitted: 10-03-2021 Revised: 30-03-2021 Accepted: 01-04-2021

ABSTRACT: We are living in the era of Internet and Social Media where everything is recorded digitally. Petabytes of data ate getting generated and processed which needs to be explored, cleansed and analyzed to make them usable. Big data is a term that describes the large volume of data – both structured and unstructured. Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. It also leads to new opportunities with innovative ideation for solving the big-data problem via visual means. It is a tough ask to visualize such a massive amount of data and lots of challenges are associated with it. In this paper, we will discuss the importance and challenges of big data visualization and review some big data visualization tools.

KEYWORDS: Big Data, Visualization Challenges, Data Visualization Tools & Techniques.

I. INTRODUCTION

Big Data has become one of most popular topic for almost all the industries which includes Academics, IT farms, and governments. The amount of data has increased exponentially within past few years due to several factors like Internet, Social Media, Telecommunication, IoT sensors and digitalization of all offline records like our medical history etc. The massive data got produced from these sources is called the Big Data. Big Data, as mentioned by Gubarev Vasiliy Vasil'evich— is a phenomenon, which have no clear borders, and can be presented in unlimited or even infinite data accumulation. And even more, the accumulated data can be presented in various data formats, most of them are not structural data flows.[1] The main concerns of Big Data are capturing, storing, analysing, and visualizing data. Interesting patterns can be found from Big Data analysis and it becomes much easier for our brain to find meaningful patterns if represented visually which helps to take decision accordingly.

Data visualization is easy and quick way to convey messages and represent complex things and has been around for centuries. Big Data is characterized by 5Vs, i.e. huge Volume, high Velocity, high Variety, low Veracity and high Value.[2] Processing this huge amount of data is not the prime concern but to process data with high diversity. High diversity and uncertainty in data hampers the response time of the application as it has to deal with structured, semi and unstructured data. The new data visualization must come up with better ways to process, analyse and visualize huge amount of complicated data. Big data visualization brings new research challenges and opportunities.

Traditional visualization tools often fails to deliver the expected outcome from this complex heterogeneous big data. Designing a new visualization tool with efficient indexing is not easy in big data. Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability. Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata. Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently. Because of the big data size, the need for massive parallelization is a challenge in visualization. [3]

Visualization-based data discovery methods allow users to mash up disparate data sources to create custom analytical views. Advanced analytics can be integrated in the methods to support creation of interactive and animated graphics on many devices like desktops, laptops, or mobile devices such as tablets and smartphones.

Over the years big data visualization has helped organizations from various perspectives like decision making, Data analysis etc.

Table-1: shows the benefits of data visualization according to the respondent percentages of a survey.

II. CHALLENGES

It's nearly impossible for traditional visualization tools to process very large data sets which are evolving continuously. There is high latency for traditional visualization approaches due to very complex nature of big data. To provide interactive visualization with as low latency, we can do the following things:

- Use the pre-computed data
- Parallelize Data Processing and Rendering
- Use a predictive middleware

Big Data visualization tool should be able to deal with semi-structured and unstructured data and it is realized that to cope with such huge amount of data there is need for immense parallelization, which is a big ask in visualization. Breaking down problems into independent tasks to be able to process independently is the main challenge for parallelization algorithm.

The task of big data visualization is to recognize interesting patterns and correlations. We need to carefully choose the dimensions of data to be visualized, if we reduce dimensions to make our visualization low then we may end up losing interesting patterns but if we use all the dimensions we may end up having visualization too dense to be useful to the users. [4]

Due to vast volume and high magnitude of big data it becomes difficult to visualize. Most of the current visualization tool have low performance in scalability, functionality and response time [3].

Some other important big data visualization problems are as follows.

Visual Noise: Most of the objects in dataset are too relative to each other, and on the screen watcher cannot divide them as separate objects. So, sometimes, the analyst cannot get even a bit of useful information from whole data visualization without any pre-processing tasks.

Large Image Perception: There is a certain level of human being perception for different data visualization. Despite that this level for graphical data visualization is much higher, compared to table data visualization, it has its own limitations. And after achieving this level of perception, the human being just loses the ability to acquire any useful information from the data overloaded view. With growth of data volumes shown at once, human being will meet a difficulty in understanding data and its analysis. Therefore, it can be said that data visualization methods are limited not only by aspect ratio and resolution of device but also by physical perception limits.

Information Loss: These approaches operate with data aggregation and filtration, based on the relatedness of objects in concrete dataset by one or more criteria. Using these approaches can mislead the analyst, when he cannot notice some interesting hidden objects, and, sometimes, complex aggregation process can consume a large amount of time and performance resources in order to get the accurate and required information.

High Performance Requirements: The graphical analysis does not stop on only static image visualization, so the above problems become more significant in dynamic visualization.

High Rate of Image Change: And the last problem is high rate of image change. This problem becomes the most significant in monitoring tasks, when a person who observes the data just cannot react to the number of data changes or its intensity on display. The simple decrease of changing rate cannot provide the desired result, as the reaction speed of the human being directly depends on it.

III. VISUALIZATION TOOLS

To solve out the problems mentioned in previous section various tools have emerged. The single most important feature of visualization is that it should be interactive, which means that user should be able to interact with the visualization. Data visualization tools provide designers with an easier way to create visual representations of large data sets. When dealing with data sets that include hundreds of thousands or millions of data points, automating the process of creating a visualization makes a designer's job significantly easier. In the following segment we have reviewed some of the most popular visualization tools.

1) Tableau: Tableau is interactive data visualization tool which is focused on Business Intelligence. Tableau provides very wide range of visualization options. Tableau has a variety of options available, including a desktop app, server and hosted online versions, and a free public option. There are hundreds of data import options available, from CSV files to Google Ads and Analytics data to Salesforce data. It has Hundreds of data import options, Hundreds of data import options, free public version and lots of tutorials to get expert on it.

An interactive visualization of the highest-grossing actors of all time.

2) Microsoft Power BI: Power BI is a powerful cloud-base business analytics service. Visualization are interactive and rich. Power BI consists of 3 elements, Power BI Desktop, Service(SaaS), Apps. Every service is available to us that is why it makes Power BI flexible and persuasive. Since this is a Microsoft tool, it has a very strong brand integration with the other MS tools.

3) Infogram: It is a web-based data visualization and infographics tool that permits users to create and share digital charts, infographics, and maps. Finished visualizations can be exported into a number of formats: .PNG, .JPG, .GIF, .PDF, and .HTML. Interactive visualizations are also possible, perfect for embedding into websites or apps. Infogram also offers a WordPress plugin that makes embedding visualizations even easier for WordPress users.

4) Google Charts: Google Chart is a powerful, easy to use and an interactive data visualization tool for browsers and mobile devices. Data sources include Google Spreadsheets, Google Fusion Tables, Salesforce, and other SQL databases. There are a variety of chart types, including maps, scatter charts, column and bar charts, histograms, area charts, pie charts, treemaps, timelines, gauges, and many others. The best part about Google Charts is that it’s completely free, and available to use for both personal and commercial projects. But, it needs coding knowledge to build data visualizations in Google Charts. If you don’t know how to code, figuring out how to use Google Charts may have a high learning curve.

5) Jupyter: JupyteR is an open-source project enabling Big Data analysis, visualization and real-time collaboration on software development across more than a dozen of programming languages. The interface holds the field for code input, and the tool runs the code to deliver the visually-readable image based on the visualization technique chosen. The ability to interact with multiple frameworks like Spark turns Jupyter into an all-around capable solution for processing the data from large,data-intensive applications with disparate sources of input. [6]

IV. VISUALIZATION TECHNIQUES

The volume, variety and velocity of big data requires from an organization to leave its comfort zone technologically to derive intelligence for effective decisions. New and more sophisticated visualization techniques based on core fundamentals of data analysis take into account not only the cardinality, but also the structure and the origin of such data.

4.1 Kernel Density Estimation for Non-Parametric Data: If we have no knowledge about the population and the underlying distribution of data, such data is called non-parametric and is best visualized with the help of Kernel Density Function that represents the probability distribution function of a random variable. Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

4.2 Box and Whisker Plot for Large Data: A binned box plot with whiskers shows the distribution of large data and easily see outliers. In its essence, it is a graphical display of five statistics (the minimum, lower quartile, median, upper quartile and maximum) that summarizes the distribution of a set of data. The lower quartile (25th percentile) is represented by the lower edge of the box, and the upper quartile (75th percentile) is represented by the upper edge of the box. The median (50th percentile) is represented by a central line that divides the box into sections. Extreme values are represented by whiskers that extend out from the edges of the box. Box plots are often used to understand the outliers in the data.

4.3Word Clouds and Network Diagrams for Unstructured Data: A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is. They can also help business users compare and contrast two different pieces of text to find the wording similarities between the two.

Network Diagrams is another visualization technique that can be used for semi-structured or unstructured data. Network diagrams represent relationships as nodes (individual actors within the network) and ties (relationships between the individuals).

4.4 Correlation Matrices: A correlation matrix allows quick identification of relationships between variables by combining big data and fast response times. Basically, a correlation matrix is a table showing correlation coefficients between variables: Each cell in the table represents the relationship between two variables. Correlation matrices are used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

V. CONCLUSION

In this paper, we have described few basic things of big data, challenges and problems of Big Data visualization as well as some Big Data visualization tools which can be very effective for different kinds of users. Data visualization may become a valuable addition to any presentation and the quickest path to understanding any data. The process of visualizing data can be challenging as well as enjoyable some times. There are many available tools and techniques of big data visualization and to choose the most appropriate visualization technique we need to understand the data, its type and composition. This paper will certainly be of great help in choosing the best tool of interest.

REFERENCES

[1]. Evgeniy Yur'evichGorodov and Vasiliy Vasil'evich Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data," 10 October 2013

[2]. Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, Rakesh Kumar Lenka, "Big Data Visualization: Tools and Challenges" 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) – 14-17 Dec. 2016

[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]. Tavel, P. "modeling and simulation design," AK Peters Ltd. Natick, MA, 2007.

[5]. CAMERON CHAPMAN "A Complete Overview of the Best Data Visualization Tools"

[6]. Vladimir Fedak "Top 4 Popular Big Data Visualization Tools" Jan 9, 2018