Big Data Visualization: Tools and Challenges

Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, Rakesh Kumar Lenka
Department of Computer Science and Engineering
International Institute of Information Technology, Bhubaneswar, India
{syedmohdali121, noopur2827}@gmail.com
{gopal, rakeshkumar}@iiit-bh.ac.in

***Abstract*—In today's world where everything is recorded digitally, right from our web surfing patterns to our medical records, we are generating and processing petabytes of data every day. Big data will be transformative in every sphere of life. But just to process and analyze those data is not enough, human brain tends to find pattern more efficiently when data is represented visually. Data Visualization and Analytics plays important role in decision making in various sectors. It also leads to new opportunities in the visualization domain representing the innovative ideation for solving the big-data problem via visual means. It is quite a challenge to visualize such a mammoth amount of data in real time or in static form. In this paper, we discuss why big data visualization is of utmost importance, what are the challenges related to it and review some big data visualization tools.**

***Index Terms*—Big Data; visualization; dashboard; interactive visualization;**

I. INTRODUCTION

In recent years Big Data has become topic of interest for all the industries including, Academics, IT Companies, and governments [1]. The rate of growth of data has increased exponentially within few years due to several factors like Internet of Things(IoTs), sensors in our environment, and digitalization of all offline records like our medical history etc. Big Data has proved its importance to this world within such a small amount of time that today almost all IT and non-IT companies are storing all the data they produce.

Today businesses struggle to just store the massive amount of data whereas analyzing, interpreting and presenting it in meaningful ways is a thought for later [2]. The main challenge of Big Data lies in capturing, storing, analyzing, sharing, searching, and visualizing data. One of the major aspect of Big Data analysis is that we can find interesting pattern in huge data set, but actually the result of the analysis is usually raw numbers and by those numbers it is very difficult to interpret anything. But if those numbers are represented visually then it becomes much easier for our brain to find meaningful patterns and take decision accordingly. Fig. 1, shows the benefits of Big data visualization [3].

Data visualization is certainly not a new thing; it has been around for centuries. Data visualization is easy and quick way to convey messages and represent complex things [4]. We humans are adapted to find patterns in everything we see. Since the data is mounting at such a massive rate the traditional ways of presenting data is obsolete [1]. Compared to traditional data,

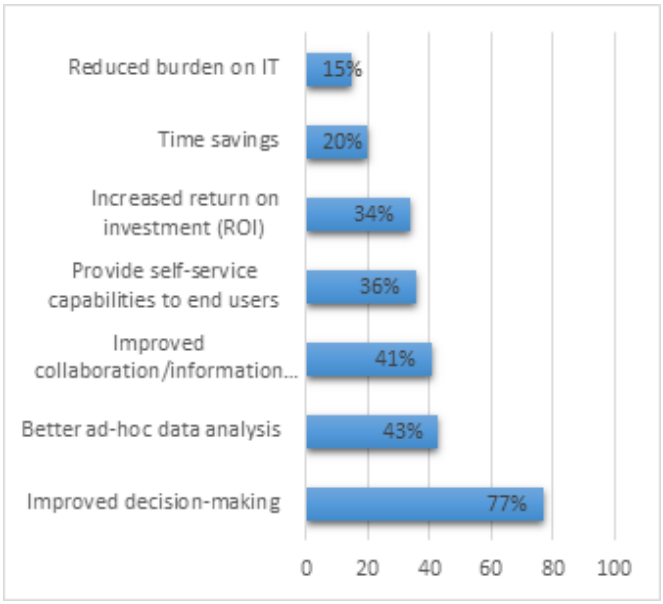

Fig. 1. Benefits of big data visualization

Big Data is characterized by 5Vs, i.e. huge Volume, high Velocity, high Variety, low Veracity and high Value. Actual challenge is not only to process this huge amount of data but to process data with high diversity. High diversity and uncertainty in data reduces the response time of the application as it has to deal with not only traditional structured data but also with semi and unstructured data [1].

II. CHALLENGES

Traditional visualization tools have reached to their limits when encountered with very large data sets and these data are evolving continuously. Though there are some extensions to traditional visualization approaches but they lag behind by miles. The visualization tool should be able to provide us interactive visualization with as low latency as possible. To reduce the latency, we can do the following things: [5]

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

Big Data visualization tool must be able to deal with semi-structured and unstructured data because big data usually have this type of format. It is realized that to cope

978-1-5090-5256-1/16/$31.00 ©2016 IEEE

with such huge amount of data there is need for immense parallelization, which is a challenge in visualization. The challenge in parallelization algorithm is to break down the problem into such independent task that they can run independently [6].

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. For example: "Given the conventional displays (1.3 million pixels), visualizing every data point can lead to over-plotting, overlapping and may overwhelm user's perceptual and cognitive capacities [7]".

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 [8]. Methods have been proposed which not only visualizes data but processes at the same time. These methods use Hadoop and storage solution and R programming language [9] as compiler environment in the model [10]. Fig 2 shows the outline of such a model.

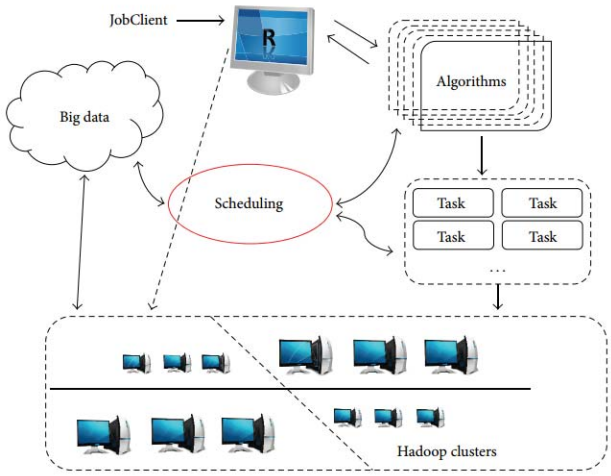

Fig. 2. The big data visualization algorithm analysis integrated model.

Some other important big data visualization problems are as follows [11]

- Visual noise: Most of the objects in dataset are too relative to each other. It becomes very difficult to separate them.
- Information loss: To increase the response time we can reduce data set visibility, but this leads to information loss.
- Large image perception: Even after achieving desired mechanical output we are limited by our physical perception.
- High rate of image change: If rate of change to image is too high it becomes impossible to react to the number.
- High performance requirements: During static visualization this factor might be ignored compared to dynamic visualization which demands more i.e. high performance

III. VISUALIZATION TOOLS

Various tools have emerged to help us out from the above pointed problems. The most important feature that a visualization must have is that it should be interactive, which means that user should be able to interact with the visualization. Visualization must display relevant information when hovered over it, zoom in and out panel should be there, visualization should adapt itself at runtime if we select subset or superset of data. We reviewed some of the most popular visualization tools.

A. Tools

1) Tableau: Tableau is interactive data visualization tool which is focused on Business Intelligence. Tableau provides very wide range of visualization options. It provides option to create custom visualization. It is fast and flexible. It supports mostly all the data format and connection to various servers right from the Amazon Aurora to Cloudera Hadoop and Salesforce. User interface is intuitive, wide variety of charts are available. For simple calculations and statistics one does not require any coding skills but for heavy analytics we can run models in R and then import the results into Tableau. This requires quite a bit of programming skill based upon the task we need to perform.

Fig. 3. Furniture sales profit and loss over globe (filled map visualization).

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. With more than 60 types of source integration you can start creating visualization in matter of minutes. Power BI combines the familiar Microsoft tools like Office, SharePoint and SQL Server. The feature that it distinguishes from other tools is that you can use natural language to query the data. You don't

require programming skills for this tool but there is option available to run your R script. You can merge multiple data sources and create models, which comes in handy. Fig. 4 represents 3 visualizations in 3 coordinates, i.e. left, bottom and right. Left represents profit by county and market, bottom represents profit by region and right coordinate represents all over sales and profit.

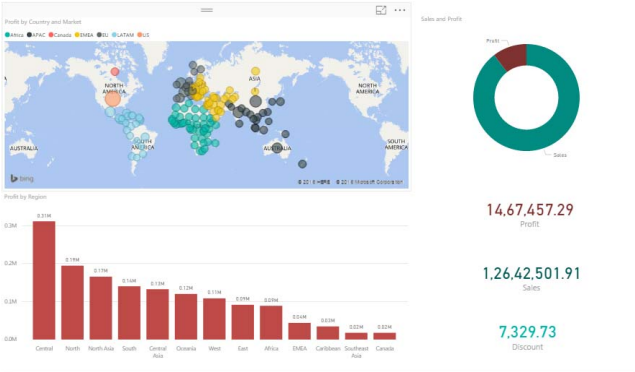

Fig. 4. Report on market analysis.

3) Plotly: Plotly is also known as Plot.ly is build using python and Django framework. The actions it can perform are analyzing and visualizing data. It is free for users but with limited features, for all the features we need to buy the professional membership. It creates charts and dashboards online but can be used as offline service inside Ipython notebook, jupyter notebook and panda. Different variety of charts are available like statistical chart, scientific charts, 3D charts, multiple axes, dashboards etc. Plotly uses a tool called "Web Plot Digitizer(WPD)" which automatically grabs the data from the static image [12].

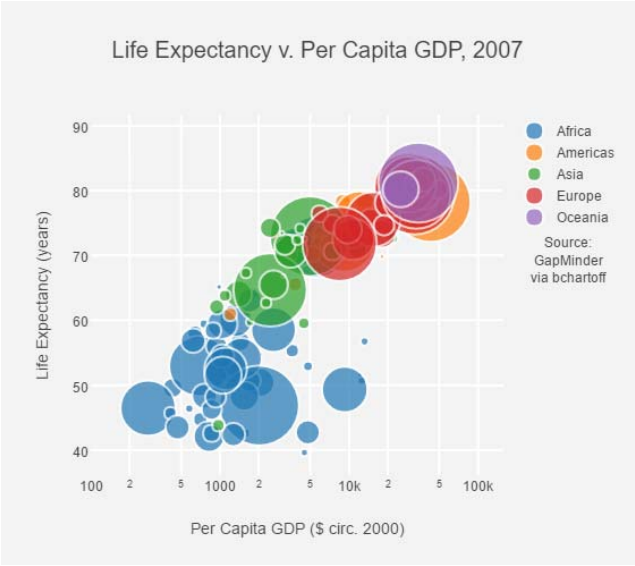

Fig. 5. Life expectancy vs per capita GDP, 2007 (bubble chart) [13].

Plotly on premises service is also available, it is like plot.ly cloud but you host data on your private cloud behind your own firewall. This for those wo have concern about the privacy of their data. Python, R, MATLAB and Julia APIs are available for the same.

4) Gephi: Gephi is open-source network analysis tool written in Java and OpenGL. It is used to handle very large and complex datasets. The network analysis includes

- Social Network Analysis
- Link Analysis
- Biological Network Analysis

With its dynamic data exploration Gephi stands out rest of its competition for graph analysis. No programming skills are required to run thin tools but a good knowledge in graphs is necessary. It uses GPU 3D render engine to accelerate the performance and give real time analysis [14].

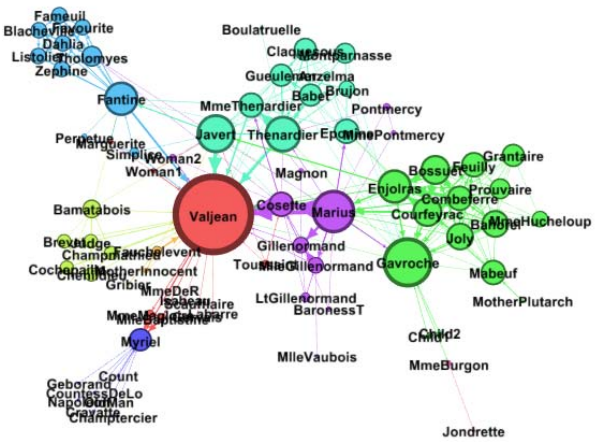

Fig. 6. Co-appearance weighted network of characters in the novel "les miserable" from Victor Hugo.

5) Excel 2016: Microsoft Excel is a spreadsheet developed by Microsoft. It can not only be used for Big Data and statistical analysis but it is also a powerful visualization tool. Using power query excel can connect to most of the services like HDFS, SaaS etc and is capable of managing Semi-Structured data. Combined with visualization techniques like "Conditional Formatting" and interactive graphs makes Excel 2016 a good contender in the ocean of Big Data visualization tools.

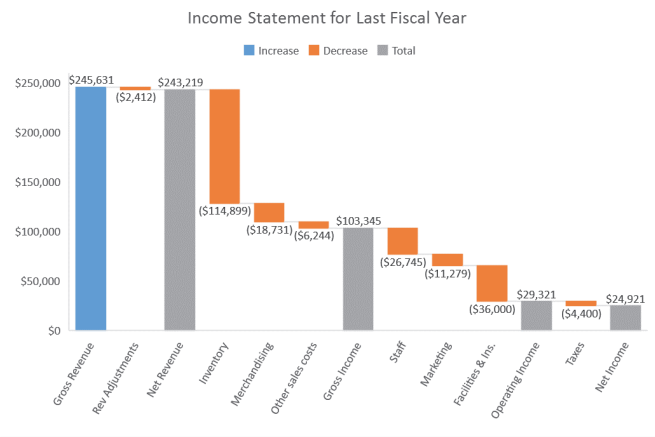

Fig. 7. : Income statement for fiscal year[15].

B. Comparison of Tools

We have compare the above tools on the basis of various attributes. These attributes are as follows:

- Open Source: If our tool is open source [16] or not.
- Integration with popular data sources: These include, MapR Hadoop Hive, Salesforce, Google Analytics, Cloudera Hadoop etc.
- Interactive Visualization: Are visualizations created by tools interactive or not.
- Client Type: What are different type of clients available for the tool, i.e. Desktop, online or Mobile App.
- MOOCS: If tutorials are available online for learning the tool.
- API: If APIs are available to embed the services of the tool.

These factors should be kept in mind while choosing for appropriate tool as per requirement. Table 1 represents the comparison of various big data visualization tools on the basis of above discussed factors.

C. Limits/Demerits

1) Tableau: Though Tableau public is available for free of cost but it provides its service online with 1GB of storage. For desktop version one has to buy license. Moreover, you need to buy the license of Server and Desktop versions separately. Coding skill is required if you need to work around R script for in-depth analysis.

2) Microsoft Power BI:

- Desktop version of the software is available for free but major drawback arises when we try to access cloud services. For that we need to have a Work Account to sign in, public account will not do the trick here.
- Workbook size is limited only to 250 MB.
- It is slow compared to Tableau

3) Plotly:

- Pro users have limited features like upload size of files are only up to 500KB.
- If you buy the professional version, you will get unlimited charts but still upload size of files will be only 5MB.
- No official offline client for Plotly is available, lots of coding skill is required if we want to work with Plotly offline.

4) Gephi: Major advantage of Gephi is its disadvantage i.e. Gephi only specializes in graph visualization, you cannot use if for other types of visualizations.

5) Excel 2016:

- API only available with the Office 365 subscription.
- Excel is not free.

D. Comparisons of Visualization Techniques

Not all visualization is applicable at all places, we need to choose wisely which techniques to use when. Table 2 [8] represents some of the popular visualization techniques and when or when not to use them. Table 3 [17] represents the classifications of the visualization methods according to big data classes.

1) Treemap: It is based on space-filling visualization of hierarchical data.
2) Circle packing: It is same as of treemap but instead of rectangles we use circles. This is not as space efficient as treemap visualization.
3) Sunburst: Hierarchy is displayed in the circular arrangement. Different layers on sunburst represents different levels of hierarchy.
4) Parallel coordinate: It represents numerous data elements for distinct objects.
5) Stream graph: It is displacement of stacked area around central axis which looks like a flowing shape.
6) Circular network diagram: Different objects are placed in the form of circle and linked to each other according to relativeness.

TABLE II
ATTRIBUTES OF VISUALIZATION TECHNIQUES

Method Name	Large Data Volume	Data Variety	Data Dynamics
Treemap	Y	N	N
Circle packing	Y	N	N
Sunburst	Y	N	Y
Parallel coordinate	Y	Y	Y
Stream graph	Y	N	Y
Circular network diagram	Y	Y	N

TABLE I
COMPARISONS OF BIG DATA VISUALIZATION TOOLS

	Open Source	Integration with popular sources	Interactive Visualization	Desktop Client	Online Client	Mobile Application	MOOCS	API
Tableau	N	Y	Y	Y	Y	Y	Y	Y
Power BI	N	Y	Y	Y	Y	Y	Y	Y
Plotly	Y	N	Y	N	Y	N	Y	Y
Gephi	Y	N	Y	Y	N	N	Y	N
Excel 2016	N	Y	Y	Y	Y	Y	Y	Y

TABLE III
CLASSIFICATION OF VISUALIZATION TECHNIQUES

Method Name	Big Data Classes
Treemap	Can be applied only to hierarchical data
Circle packing	Can be applied only to hierarchical data
Sunburst	Volume + Velocity
Parallel coordinate	Volume + Velocity + Variety
Stream graph	Volume + Velocity
Circular network diagram	Volume + Variety

IV. CONCLUSION

In the world of big data where every information is crucial in one way or the other we rely on the visual information to find useful patterns. But traditional methods of visualization do not keep up with the pace and volume of data, we require such tools which deal with all the characteristics of big data and gives us result without giving up performance and response time. In this paper we identified why big data visualization is important for and what are the challenges and issues related to this. We also noted that interactivity of visualization is of utmost importance and good visualization tools should produce interactive visualization. We also studied how people are proposing new systems to deal with these challenges.

We reviewed some of the popular visualization tools and observed their merits and demerits. These tools are quite promising, they generate rich and interactive visualizations, most of them tackle the huge volume of data and response in acceptable amount of time. It is clear from the analysis of these tool that there cannot be one winner among them. One should choose them according to their requirement. For e.g. a small business might not want to use Tableau because of its high cost. Before choosing any of the visualization tool businesses want to review what all are their requirement and which tool(s) suite the best for them. This paper will help them to choose their tool of interest.

ACKNOWLEDGEMENT

This research was supported by International Institute of Information Technology, Bhubaneswar.

REFERENCES

[1] Jin X, Wah BW, Cheng X, and Wang Y, "Significance and challenges of big data research," Big Data Research, 2015 Jun 30;2(2):59-64.

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

[3] SAS, "Data Visualization: Making Big Data Approachable and Valuable, White Paper," January 2013, pp.1-4

[4] SAS, Data Visualization: What it is and why it matters, www.sas.com/en_sg/insights/big-data/data-visualization.html

[5] Agrawal R, Kadadi A, Dai X, and Andres F, "Challenges and opportunities with big data visualization," 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, 2015 Oct 25 (pp. 169-173). ACM.

[6] Childs H, Geveci B, Schroeder W, Meredith J, Moreland K, Sewell C, Kuhlen T, and Bethel EW, "Research challenges for visualization software," Computer, 2013 May 1(5) pp:34-42.

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

[8] Lidong Wang, Guanghui Wang, and Cheryl Ann Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," Digital Technologies, vol. 1, no. 1 (2015), pp. 33-38. doi:10.12691/dt-1-1-7.

[9] The R Journal, http://journal.r-project.org/

[10] Cai L, Guan X, Chi P, Chen L, and Luo J, "Big data visualization collaborative filtering algorithm based on RHadoop," International Journal of Distributed Sensor Networks, 2015 Jan 1;2015:3.

[11] Gorodov EY and Gubarev VV, "Analytical review of data visualization methods in application to big data," Journal of Electrical and Computer Engineering, 2013 Jan 1;2013:22.

[12] Plotly, Automatically Grab Data From an Image with WebPlotDigitizer, http://blog.plot.ly/post/70293893434/automatically-grab-data-from-animage-with

[13] Plotly, Make a bubble chart, http://help.plot.ly/make-a-bubble-chart/

[14] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." ICWSM 8 (2009): 361-362.

[15] Introducing new and modern chart types now available in Office 2016 Preview, https://blogs.office.com/2015/07/02/introducing-new-and-modern-chart-types-now-available-in-office-2016-preview/

[16] Open Source Initiative, The Open Source Definition, https://opensource.org/osd-annotated

[17] E.Y. Gorodov and V.V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data," Journal of Electrical and Computer Engineering, 2013, Jan 1;2013:22