Big Data Analytics for Data Visualization: Review of Techniques

Geetika Chawla
Assistant Professor Dept. of CSE
FET, MRIIRS

Savita Bamal
Assistant Professor
Dept. of CSE
DPGITM GURUGRAM

Rekha Khatana
Assistant Professor
GITMGURUGRAM

ABSTRACT
This is information era, where information is increasing exponentially. Extracting information in a way that the human mind can comprehend is a big challenge. Visualization plays a key role in data discovery process and in better decision making. A question still arises how to visualize interesting structures and patterns that are in hyper-dimensional data spaces. We can sketch a convincing visual story from raw data with the use of right tool. This paper focuses on Big Data visualization, its challenges, various tools. Researchers have explored a new way to visualize and analyze complex and dynamic datasets using virtual reality. We have also investigated how virtual reality has radically changed the world of Big Data Visualization.

Keywords
Big Data, Big Data Visualization, Virtual reality

1. INTRODUCTION
Big data is a term that depicts the substantial volume of information – both structured and unstructured – that immerses a business on an everyday premise. This huge amount of information is not beneficial until the organizations get insights from this data. The information can be analyzed that prompt business for better decisions and key business moves.Huge amount of data is processed every day. For example, Facebook's Hive data warehouse holds 300 PB data with an incoming daily rate of about 600 TB in April, 2014 [3]. This example also illustrates that big data is generated very fast. Accumulating this large amount of data is very complex task. Unstructured data is another feature of big data. It is the information that does not have a pre-defined data model or format. Large amount of data is generated now days from social networks, traffic sensors, satellites imagery; broadcast audio streams, banking transactions, financial market, and many more. 3Vs (Velocity, Volume and Variety) nicely describe major aspects of big data. Current data processing systems (e.g. relational data warehouse) can process a large amount of relational data but they are not flexible to deal with semi-structured or unstructured data. So, new technologies have to be developed to handle data from various sources, e.g. social networks, stock market, data collected from multiple sensors, etc. The general categories of activities involved with big data processing are:

- Data Ingestion
- Storage
- Processing and Analyzing
- Data Insight

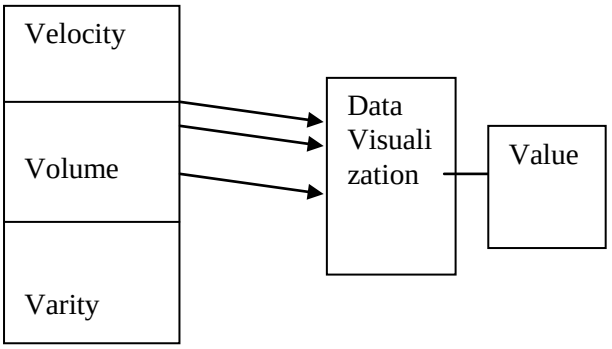

Fig 1: 4Vs.

2. BIG DATA AND VISUALIZATION
Data visualization is the presentation of data in a pictorial or graphical format [21]. To analyze and get better insights from big data, it has to be represented in some systematic format.
Data visualization is important to group together many data points, to understand relationships in data, debate questions in real time and more quickly decide where to focus research. It helps data scientists to identify hidden data patterns and how it is processed. Also, business analysts can apply data visualization techniques to identify areas that need attention or improvement, focus on factors influencing customer behavior and predict sales volumes. The visualization-based discovery tools take the challenges presented by the Three V's of big data and turn them in to opportunities for growth.

2.1 Big Data Visualization Process
The Visualization process consists of the following steps as shown in figure 2:

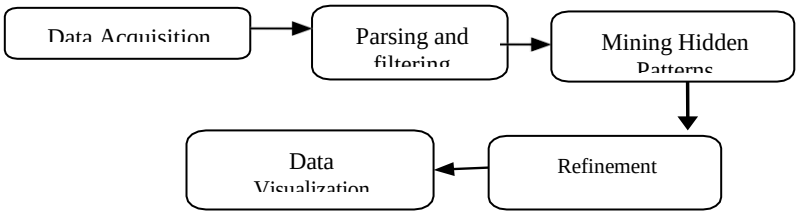

Fig 2: Big Data Visualization Process

The first step in visualization process is to acquire the data from various sources. Data collected from heterogeneous sources may be unstructured/semi-structured, so it has to be parsed into structured format. All the data may not be necessary for visualization; the next step is to filter out the unimportant data. Useful patterns are then extracted and represented in a form of charts and graphs so as to reveal hidden information easily interpreted by the user.

2.2 Challenges of Big Data Visualization

Big data visualization is challenging because of the volume, variety and velocity of data. Biggest challenge when working with big data is how to handle large data volumes and effectively display results of data visualization and analysis that is meaningful and useful. New mechanism has to be developed to look at the data so as help decision makers in getting insight from it in a very simple and easy manner by using graphs and charts. Traditional visualization tools are not capable to handle very large data sets. The tool for visualization should be able to provide us visualization with as low latency as possible. To cope with such huge amount of data, parallelization is also required, which is a challenge in visualization. The major functionality of big data visualization is to identify interesting patterns. For pattern mining data dimensions have to be carefully chosen. If we select few dimensions then it can make our visualization low and many interesting patterns may lost, similarly if we select all the dimensions that may result in dense visualization which may not be useful to the users. For example: "Given the resolution of 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" [5]. Most of the current visualization tools have low performance in scalability, functionality and response time [6].

3. BIG DATA VISUALIZATION METHODS

A number of approaches have been in use for big data visualization. These approaches have been classified according to (1) data volume, (2) data variety, and (3) data dynamics. Various data visualization methods are:

3.1 Tree map

It is a method to represent hierarchical data as a set of nested rectangles. It uses a tiling algorithm to divide the parent rectangle into sub-rectangles. Generally, qualified algorithm is used. Area of a rectangle defines the quantity assigned to a category. However, tree maps have the limitation of representing zero and negative values. Also, when more pixels are used to show the hierarchy the size distorts.

3.2 Circle Packing

It is an alternative method of tree map in which circles are used to represent different hierarchical levels. Area of a circle defines the quantity of a particular category. Like tree map it also uses different colors to categories different levels. Unlike, tree map this method is not space efficient.

3.3 Parallel Coordinates

It is a way to visualize high-dimensional data. Data elements can be plotted across many dimensions individually; parallel coordinates can display both the forest and the tree. Pattern of lines are drawn to get a clear picture of data. To see detailed performance of specific data elements individual lines can be highlighted. However, large number of data items results in over-plotting. This technique is not suitable for categorical data.

3.4 Stream Graph

It is used to show value displacement around a varying central time line. It displays the changes in data over time of different categories. In a Stream Graph, the size of each individual stream shape is proportional to the values in each category. They are ideal to represent high-volume datasets.

It is easy to acquire knowledge from a mass of information through data visualization tools. People can find things (outliers, hidden trends or clusters) that they don't know through the ideal data visualization tool. These tools also help to dive deep into fast changing data sets. The main features of Big Data Visualization tools has been discussed in the following table.

Another Example is Titans of Space 2.0 that helps to deep- dive tour through our Solar System. The goal was to gain a new perspective on what our universe actually looks like by taking advantage of the increased spatial awareness made possible by modern VR. From the perspective of Big Data Visualization, scaling is a significant issue caused by multidimensional systems, needed to seek a branch of information in order to obtain some specific value or knowledge [20]. Other important issues focused by researchers are how to merge virtual objects to real scene view. This mapping may result into distortion of real scene and also system delay. Also, there is mismatch between the virtual and physical distances. So, to help in better interaction there is a need to create an effective framework.effectively investigate paleontology, shape perception, MRI, and physics.

Table I: Characteristics of Big Data Tools

Tools	Characteristics	Applications
Plotly	Modern open source platform for agile business intelligence and data science.	Online graphing, analytics, and static tool for individuals as well as scientific graphing libraries for Python, R, MATLAB, Perl, J Arduino, and REST.
Tableau	Can Handle Large Volumes of Data, Filter Simultaneously Multiple Data Sets, Users can create and distribute interactive and shareable, dashboards depicting trends and variations, Build interactive dashboards, Built-in support for R, Google Big Query API, Allows users to analyze what has happened, does not have any statistical or predictive capabilities	Business intelligence tool used by academic researchers and government to do visual data analytics.

SAS Visual Analytics	Complete platform for analytics visualization, enabling users to identify patterns and relationships in data that weren't initially evident.	Tool for design; distribute reports, dashboards and analytics.
Microsoft Power BI	Power BI dashboards provide a 360-degree view for business users with their most important metrics in one place, updated in real time, and available on all of their devices.	Create interactive visualizations, reports and dashboards, Use natural language queries on a dashboard.
D3.js	A JavaScript library for creating dynamic, interactive data visualizations in web browsers framework.	Make use of the widely implemented SVG, HTML5, and CSS standards.

4. COCLUSION

Information is crucial in one way or the other, so to find useful patterns one has to rely on the visual representation. Huge volume of heterogeneous data can not be handled efficiently, so the tools which give us result without giving up performance and response time are required. In this paper importance of big data visualization has been identified and discussed about challenges and issues related to this. To get deep insight and 360 degree view of big data researchers are now focusing on using Virtual reality as a new technique for big data visualization.

5. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template

6. REFERENCES

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

[2] Agrawal R, Kadadi A, Dai X, and Andres F, 2015 "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.

[3] Pamela Vagata and Kevin Wilfong, 2014. Scaling the facebook data warehouse to 300PB.https://code.facebook.com/posts/229861827208629/ scaling-the-facebook-data-warehouse-to-300-pb/, April 2014.

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

[5] T. A. Keahey, Using visualization to understand big data, Technical Report, IBM Corporation, 2013, 1-16.

[6] Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, Rakesh Kumar Lenka, "Big Data Visualization: Tools and Challenges." 2nd International Conference on Contemporary Computing and Informatics (ic3i).

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

[8] John N. W., Leng J. 2001. "Scientific Applications of Visualization, Virtual Reality and High Performance Visualization Computers", and "Scientific Examples of Virtual Reality and Visualization Applications", Two associated UKHEC reports, 2001.

[9] C. Cruz-Neira, 1993. "Virtual Reality Overview". SIGGRAPH' 93, 1.1-1.18 (1993).

[10] C.L. P. Chen, C.-Y. Zhang, 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, 275 (10), August 2014, 314-347.

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

[12] Andrew Tran, 2015. Introduction to plotly, Sep,2015. http://andrewbtran.github.io/JRN- 418/class3/plotly.html#/

[13] https://www.sas.com/en_in/software/business-intelligence/visual-analytics.html.

[14] B. Laha, D.A. Bowman and J.J. Socha, 2014. "Effects of VR System Fidelity on Analyzing Isosurface Visualization of Volume Datasets," IEEE Transactions on Visualization and Computer Graphics, 20(4) 513- 522, 2014.

[15] C. Demiralp, C. D. Jackson, D. B. Karelitz, S. Zhang, and D. H. Laidlaw, 2006. "CAVE and Fishtank Virtual- Reality Displays: A Qualitative and Quantitative Comparison," IEEE Transactions on Visualization and Computer Graphics, vol. 12, pp. 323-330, 2006.

[16] T.V. Williams, 1982. A man-machine interface for interpreting electron-density maps. Ph.D. dissertation, University of North Carolina, Chapel Hill, 1982

[17] J. Chen, H. Cai, A. P. Auchus, and D. H. Laidlaw, 2012 "Effects of Stereo and Screen Size on the Legibility of Three-Dimensional Streamtube Visualization," IEEE Transactions on Visualization and Computer Graphics, vol. 18, pp. 2130-2139, 2012.

[18] C. Donalek, S. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake, S. Davidoff, J. S. Norris, G. Longo, 2014. "Immersive and collaborative data visualization using virtual reality platforms", Proceedings of the IEEE International Conference on Big Data ser. IEEE BigData 2014, pp. 609-614, 2014.

[19] https://www.wired.com/2014/5/darpa-is-using-oculus- rift-to-prep-for-cyberwar/

[20] Ekaterina Olshannikova1, Aleksandr Ometov1, Yevgeni Koucheryavy1 and Thomas Olsson2, 2015 " Visualizing Big Data with augmented and virtual reality: challenges and research agenda", Journal of Big Data (2015) 2:22

[21] https://www.sas.com/en_us/insights/big-data/data-visualization.html