An effective approach for Visualizing Big Data

G.L.Anand Babu
Department of Information Technology, CVSR School of Engineering, Anurag Group of Institutions, Hyderabad ,Telangana, India

G.Sekhar Reddy
Department of Information Technology, CVSR School of Engineering, Anurag Group of Institutions, Hyderabad ,Telangana, India

Swathi Agarwal
Department of Information Technology, CVSR School of Engineering, Anurag Group of Institutions, Hyderabad ,Telangana, India

Abstract - Visualization is an important approach in helping Big Data to get a complete view of data and discover data values. Big Data analytics plays a key role for reducing the data size and complexity in Big Data applications. Big Data analytics and visualization should be integrated effortlessly so that they work best in Big Data applications. The main aim is to consider an effective approach in visualization methods for existing Big Data, as well as to offer new solutions for issues related to the current state of Big Data Visualization. This paper provides a classification of existing analytical methods, visualization techniques and tools, with a particular emphasis placed on surveying the evolution of visualization methodology over the past years. Despite the technological development of the modern world, human involvement (interaction), judgment and logical thinking are necessary while working with Big Data. Therefore, the role of human perceptional limitations involving large amounts of information is evaluated. Based on the results, a non-traditional approach is proposed.

Keywords: Big Data, visualization, virtual reality, interactive visualization, Human interaction.

I. INTRODUCTION

Data visualization is representing data in some systematic form including attributes and variables for the unit of information [1]. Visualization-based data discovery methods allow business 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 desktops, laptops, or mobile devices such as tablets and smart phones [2]. Big data are high volume, high velocity, and high variety datasets that require new forms of processing to enable enhanced process optimization, insight discovery and decision making. Challenges of Big Data lie in data capture, storage, analysis, sharing, searching, and visualization [4]. Visualization can be thought of as the "front end" of big data. There are following data visualization myths [3]:

- All data must be visualized: It is important not to overly rely on visualization; some data does not need visualization methods to uncover its messages.
- Only good data should be visualized: A simple and quick visualization can highlight something wrong with data just as it helps uncover interesting trends.
- Visualization will always manifest the right decision or action: Visualization cannot replace critical thinking.
- Visualization will lead to certainty: Data is visualized doesn't mean it shows an accurate picture of what is important. Visualization can be manipulated with different effects.

Visualization approaches are used to create tables, diagrams, images, and other intuitive display ways to represent data. Big Data visualization is not as easy as traditional small data sets. The extension of traditional visualization approaches have already been emerged but far from enough. In large-scale data visualization, many researchers use feature extraction and geometric modeling to greatly reduce data size before actual data rendering. Choosing proper data representation is also very important when visualizing big data [4].

Advantages of data visualization tools

- Improved decision making

- Better ad-hoc data analysis
- Improved collaboration/information sharing
- Increased return on investment
- Time savings
- Provide self service capabilities to end users

II. BIG DATA PROCESSING METHODS

Currently, there exist many different techniques for data analysis [5], mainly based on tools used in statistics and computer science. The most advanced techniques to analyze large amounts of data include: artificial neural networks; models based on the principle of the organization and functioning of biological neural networks; methods of predictive analysis; statistics; Natural Language Processing; etc. Big Data processing methods embrace different disciplines including applied mathematics, statistics, computer science and economics. Those are the basis for data analysis techniques such as Data Mining, Neural Networks, Machine Learning, Signal Processing and Visualization Methods. Most of these methods are interconnected and used simultaneously during data processing, which increases system utilization tremendously. We would like to familiarize reader with the primary methods and techniques in Big Data processing. As this topic is not a focus of the paper, this list is not exhaustive. Nevertheless, the main interconnections between these methods are shown and application examples are given.

Optimization methods are mathematical tools for efficient data analysis. Optimization includes numerical analysis focused on problem solving in various Big Data challenges: volume, velocity, variety and veracity that will be discussed in more detail later. Some widely used analytical techniques are *genetic programming*, *evolutionary programming* and *particle swarm optimization*. Optimization is focused on the search of the optimal set of actions needed to improve system performance. Notably, genetic algorithms are also a specific part of machine learning direction. Moreover, *statistical testing*, *predictive* and *simulation models* are applied also as for Statistics methods.
Statistics methods are used to collect, organize and interpret data, as well as to outline interconnections between realized objectives. Data-driven statistical analysis concentrates on implementation of statistics algorithms. *A/B testing* [6] technique is an example of a statistics method. In terms of Big Data there is a possibility to perform a variety of tests. The aim of A/B tests is to detect statistically important differences and regularities between groups of variables to reveal improvements. Besides, statistical techniques contain cluster analysis, data mining and predictive modeling methods. Some techniques in *spatial analysis* originate from the field of statistics as well. It allows analysis of topological, geometric or geographic characteristics of data sets.
Visualization methods concern the design of graphical representation, i.e. to visualize the innumerate amount of the analytical results as diagrams, tables and images. Visualization for Big Data differs from all of the previously mentioned processing methods and also from traditional visualization techniques. To visualize large-scale data, feature extraction and geometric modeling can be implemented. These processes are needed to decrease the data size before actual rendering. Intuitively, visual representation is more likely to be accepted by a human in comparison with unstructured textual information. The era of Big Data has been rapidly promoting the data visualization market. According to Mordor Intelligence the visualization market will increase at a compound annual growth rate (CAGR) of 9.21 % from $4.12 billions in 2014 to $6.40 billions by the end of 2019. SAS Institute provides results of an International Data Group (IDG) research study in the white paper [7]. The research is focused on how companies are performing Big Data analysis. It shows that 98 % of the most effective companies working with Big Data are presenting results of the analysis via visualization. Statistical data from this research provides evidence of the visualization benefits in terms of decision making improvement, better ad-hoc data analysis, improved collaboration and information sharing inside/outside an organization.

Nowadays, different groups of people including designers, software developers and scientists are in the process of searching for new visualization tools and opportunities. For example, Amazon, Twitter, Apple, Facebook and Google are companies that utilize data visualization in order to make appropriate business decisions. Visualization solutions can provide insights from different business perspectives. First of all, implementation of advanced visualization tools enables rapid exploration of all customers/users data to improve customer-company relationships. It allows marketers to create more precise customer segments based on data from purchasing history or life stage and other factors. Besides, correlation mapping may assist in the analysis of customer/user behavior to identify and analyze the most profitable of them. Secondly, visualization capabilities allow companies opportunities to reveal correlations between product, sales and customer profiles. Based on gathered metrics, organizations may provide novel special offers to their customers. Moreover, visualization enables tracking of revenue trends and can be useful for risk analysis. Thirdly, visualization as a tool provides better understanding of data. Higher efficiency is reached

by obtaining relevant, consistent and accurate information. So, visualized data could assist organizations to find different effective marketing solutions. In this section we familiarized the reader with the main techniques of data analysis and described their strong correlation to each other. Nevertheless, the Big Data era is still in the beginning stage of its evolution. Therefore, Big Data processing methods are evolving to solve the problems of Big Data and new solutions are continuously being developed. By this statement we mean that big world of Big Data requires multiple multidisciplinary methods and techniques that lead to better understanding of the complicated structures and interconnections between them.

III. VISUALIZATION METHODS

Historically, the primary areas of visualization were Science Visualization and Information Visualization. However, during recent decades, the field of Visual Analytics was actively developing. As a separate discipline, visualization emerged in 1980 as a reaction to the increasing amount of data generated by computer calculations. It was named Science Visualization, as it displays data from scientific experiments related to physical processes. This is primarily a realistic three-dimensional visualization, which has been used in architecture, medicine, biology, meteorology, etc. This visualization is also known as Spatial Data visualization, which focuses on the visualization of volumes and surfaces. Information Visualization emerged as a branch of the Human-Computer Interaction field in the end of 1980s. It utilizes graphics to assist people in comprehending and interpreting data. As it helps to form mental models of the data, for humans it is easier to reveal specific features and patterns of the obtained information.

Visualization methods have evolved much over the last decades, the only limit for novel techniques being human imagination. To anticipate the next steps of data visualization development, it is necessary to take into account the successes of the past. It is considered that quantitative data visualization appeared in the field of statistics and analytics quite recently. However, the main precursors were cartography and statistical graphics, created before the 19th century for the expansion of statistical thinking, business planning and other purposes [8]. The evolution in the knowledge of visualization techniques resulted in mathematical and statistical advances as well as in drawing and reproducing images.

By the 16th century, tools for accurate observation and measurement were developed. Precisely, in those days the first steps were done in the development of data visualization. The 17th century was swept by the problem of space, time and distance measurements. Furthermore, the study of the world's population and economic data had started. The 18th century was marked by the expansion of statistical theory, ideas of data graphical representation and the advent of new graphic forms. At the end of the century thematic maps displaying geological, medical and economic data was used for the first time. The first methods were performed as simple plots followed by one dimensional histograms [9]. Still, those examples are useful only for small amounts of data. By introducing more information, this type of diagram would reach a point of worthlessness.

At the turn of 20–21st centuries, steps were taken in the development of interactive statistical computing and new paradigms for data analysis. Technological progress was certainly a significant prerequisite for the rapid development of visualization techniques, methods and tools. More precisely, large-scale statistical and graphics software engineering was invented, and computer processing speed and capacity vastly increased. Moreover, currently used technologies for data visualization are already causing enormous resource demands which include high memory requirements and extremely high deployment cost. However, the currently existing environment faces a new limitation based on the large amounts of data to be visualized in contrast to past imagination issue.

Modern effective methods are focused on representation in specified rooms equipped with widescreen monitors or projectors [10]. Nowadays, there are a fairly large number of data visualization tools offering different possibilities. These tools can be classified based on three factors: by the data type, by visualization technique type and by the interoperability. The first refers to the different types of *data to be visualized:*

- *Univariate data* One dimensional arrays, time series, etc.
- *Two-dimensional data* Point two-dimensional graphs, geographical coordinates, etc.
- *Multidimensional data* Financial indicators, results of experiments, etc.
- *Texts and hypertexts* Newspaper articles, web documents, etc.
- *Hierarchical and links* The structure subordination in the organization, e-mails, documents and hyperlinks, etc.
- *Algorithms and programs* Information flows, debug operations, etc.

The second factor is based on *visualization techniques and samples to represent different types of data*. Visualization techniques can be both elementary (line graphs, charts, bar charts) and complex (based on the

mathematical apparatus). Furthermore, visualization can be performed as a combination of various methods. However, visualized representation of data is abstract and extremely limited by one's perception capabilities and requests. Types of visualization techniques are listed below:

1. *2D/3D standard figure* [11]. May be implemented as bars, line graphs, various charts, etc.. The main drawback of this type is the complexity of the acceptable visualization for complicated data structures.
2. *Geometric transformations* [12]. This technique represents information as scatter diagram. This type is geared towards a multi-dimensional data set's transformation in order to display it in Cartesian and non-Cartesian geometric spaces. This class includes methods of mathematical statistics.
3. *Display icons* [13]. Ruled shapes (needle icons) and star icons. Basically, this type displays the values of elements of multidimensional data in properties of images. Such images may include human faces, arrows, stars, etc. Images can be grouped together for holistic analysis. The result of the visualization is a texture pattern, which varies according to the specific characteristics of the data.
4. *Methods focused on the pixels* [14]. Recursive templates and cyclic segments. The main idea is to display the values in each dimension into the colored pixel and to merge some of them according to specific measurements. Since one pixel is used to display a single value, therefore visualization of large amounts of data can be reachable with this methodology.
5. *Hierarchical images* [15]. These type methods are used with the hierarchical structured data.

The third factor is related to the interoperability with visual imagery and techniques for better data analysis. The application used for the visualization should present visual forms that capture the essence of data itself. However, it is not always enough for a complete analysis. Data representation should be constructed in order to allow a user to have different visual points of view.

To this end, the most effective visualization method is the one that uses multiple criteria in the optimal manner. Otherwise, too many colors, shapes, and interconnections may cause difficulties in the comprehension of data, or some visual elements may be too complex to recognize.

Using visual and automated methods in Big Data processing gives a possibility to use human knowledge and intuition. Moreover, it becomes possible to discover novel solutions for complex data visualization. Vast amounts of information motivate researchers and developers to create new tools for quick and accurate analysis. As an example, the rapid development of visualization techniques may be concerned. In the world of interconnected research areas, developers need to combine existing basic, effective visualization methods with new technological opportunities to solve the central problems and challenges of Big Data analysis.

IV. CONCLUSION

In this paper we have obtained relevant Big Data Visualization methods classification and have suggested the modern tendency towards visualization-based tools for business support and other significant fields. Past and current states of data visualization were described and supported by analysis of advantages and disadvantages. Visualizations can be static or dynamic. Interactive visualizations often lead to discovery and do a better job than static data tools. Interactive visualizations can help gain great insight from big data. Interactive brushing and linking between visualization approaches and networks or Web-based tools can facilitate the scientific process. For visualization problems discussed in this work, it is critical to understand the issues related to human perception and limited cognition. Only after that, the field of design can provide more efficient and useful ways to utilize Big Data. This will help develop new methods and tools for big data visualization. Big Data analytics and visualization can be integrated tightly to work best for Big Data applications. Immersive virtual reality (VR) is a new and powerful method in handling high dimensionality and abstraction. It will facilitate Big Data visualization greatly.

REFERENCES

[1] M. Khan, S.S. Khan, Data and Information Visualization Methods and Interactive Mechanisms: A Survey, *International Journal of Computer Applications,* 34(1), 2011, pp. 1-14.
[2] Intel IT Center, Big Data Visualization: Turning Big Data Into Big Insights, White Paper, March 2013, pp.1-14.
[3] P. Simon, The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions, *Harvard Business Review*, June 13, 2014, pp. 1-8.
[4] C.L. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, 275 (10), August 2014, pp. 314-347.
[5] Akerkar R. Big Data computing. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2013.
[6] Lake P, Drake R. Information systems management in the Big Data era. Advanced information and knowledge processing. Springer; 2015.
[7] SAS: Data visualization: making big data approachable and valuable. Market Pulse: White Paper (2013)
[8] Few S, EDGE P. Data visualization: past, present, and future. IBM Cognos Innovation Center; 2007.

[9] Tufte ER. The visual display for quantitative information. Chelshire: Graphics Press; 1983.
[10] Febretti A, Nishimoto A, Thigpen T, Talandis J, Long L, Pirtle J, Peterka T, Verlo A, Brown M, Plepys D et al. CAVE2:a hybrid reality environment for immersive simulation and information analysis. In: IS&T/SPIE Electronic Imaging (2013). International Society for Optics and Photonics
[11] Tory M, Kirkpatrick AE, Atkins MS, Moller T. Visualization task performance with 2D, 3D, and combination displays.IEEE Trans Visual Comp Graph. 2006;12(1):2–13.
[12] Stanley R, Oliveria M, Zaiane OR. Geometric data transformation for privacy preserving clustering. Departament of Computing Science; 2003.
[13] Healey CG, Enns JT. Large datasets at a glance: combining textures and colors in scientific visualization. IEEE Trans Visual Comp Graph. 1999;5(2):145–67.
[14] Keim DA. Designing pixel-oriented visualization techniques: theory and applications. IEEE Trans Visual Comp Graph. 2000;6(1):59–78.
[15] Kamel M, Camphilho A. Hierarchic Image Classification Visualization. In: Proceedings of Image Analysis and Recognition 10th International Conference, ICIAR; 2013.