Visualization of Big Data: Tools and Techniques for Data-Driven Decision Making

Deeksha Sharma
Department of Computer Science
Shaheed Rajguru College of applied sciences for Women, University of Delhi
deeksha.sharma177@gmail.com

Asha Yadav
Department of Computer Science
Shaheed Rajguru College of applied sciences for Women, University of Delhi
yadav.asha26@gmail.com

Abstract

In this paper, we review about one of the challenges of big data that is big data visualization, which is referred to presentation of data in a pictorial or graphical format and how it simplifies complex issues and develops mutual understanding.
As the amount of data increases, visualizing data and making sense from it becomes difficult. Data visualization software enable the decision makers to see results visually and extract important information easily as human mind can easily relate to charts and graphs than reading pages and pages. Also using visualization we can easily compare results based on big datasets which is not that feasible otherwise.

We have first discussed about what basically big data is and why working with it is a major problem. Then we mentioned the importance of visualization-based data discovery tools and how these help in deriving value from big data.

Key Words: Big Data, Visualization, Para View, Hadoop.

7. Introduction

These Big data is watching you each time you go to a website, every minute you spend on the internet and a huge amount of data is captured and analyzed every second. Thus, Big Data plays crucial role for increasing productivity growth in the whole world. Since, it is affecting many diverse domains like education, health field, administrative sectors, etc apart from software intensive industry. When the size of data grows beyond Exabyte (10^{18}) and it surpasses the available technologies capability to be stored, managed and processed efficiently, it enters the world of Big Data. "Big Data" is a term encompassing the use of techniques to capture, process, analyze and visualize potentially large data sets in a reasonable time frame not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies" [1].

Big data help businesses to achieve deeper and faster insights of the large valuable data, thus improving customer experience and increasing the overall throughput of the business. But, with the ever increasing size and variety of data gathered Big Data is proving to be a tricky challenge for many organizations to achieve desired outcome. The velocity with which data flows in makes it difficult to handle and access simultaneously.

Big Data represents a fast-growing multibillion-dollar worldwide market. The adoption of Big Data solutions outside of high-performance computing (HPC) is continuing at a rapid pace. IDC expects the Big Data technology and services market to grow from $6 billion in 2011 to $23.8 billion in 2016. This represents a compound annual growth rate (CAGR) of 32%, or about seven times that of the overall information and communication technology (ICT) market [2].

The solutions to such problems are the visualization-based data discovery tools. These tools promote self-service business intelligence making it easier for users to integrate data, analyze and present it in a way which can be easily understood. So, visualization-based data discovery tools are worth exploring by businesses that seek to derive more value from big data

8. Visualization Challenges

All The well-known 3v's of big data Volume, Variety and Velocity pose a great challenge on the visualization forms that can be used to depict the insights of the data. While the extraction of valuable information for decision making via Big Data majority depends on reducing the latency time from data capture to action that explain the data to the management. The Big Three V's of Big Data are explained as below:

Volume: Using visualization-based data discovery tools, businesses can work with an immense number of datasets turning their attention from managing data to gaining rich insights that is, enabling businesses to

derive meaning from large and growing, volumes of data.

Velocity: With visualization-based data discovery tools, businesses can replace batch processing with real-time processing of continually updated data streams, making more people to access, analyze and view real-time data.

Variety: Using visualization-based data discovery tools, we can combine as many data sources as needed helping businesses derive more meaning from structured data as well as semi-structured and unstructured data.

The data tend to become unstructured as business activities and complexity of big data increases (see Figure 1). Increasing availability of mobile devices is another challenge leading to rise of visualization-based data discovery tools. Businesses that depend on centralized creation of reports are missing the opportunity to adapt a faster, cost-effective and more democratized Business Intelligence Model which combines the advantages of big data and mobile workforce to speed insights and improve collaboration.

According to big data- the next Big thing, a joint report by NASSCOM and CRISIL Global Research and Analytics, "The Indian Big Data Industry is expected to grow from US$200 million in 2012 10 US$1 billion in 2015 at a CAGR of 83%...In India, Big Data analytics and related IT services will create an estimated 15,000 to 20,000 specialist jobs by 2015." [3].

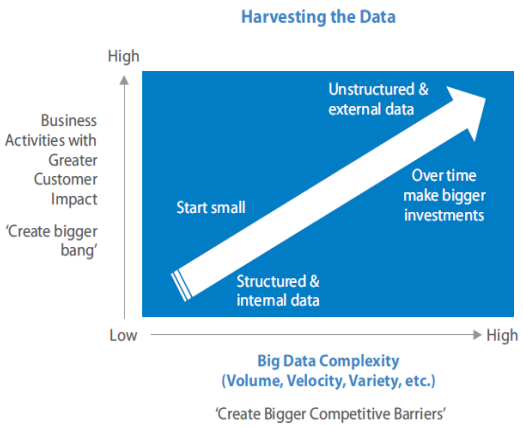

Figure 1. Big Data Complexity [4]

Thus, the smart data intensive processes require tools that can provide quick and deep access to the hidden ice-berg of information.

9. Visualization Based Data Discovery Tools

The "By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in information, an information map is kind of useful."- David McCandless (author, data journalist & information designer). [5]

The tools that can be used to visualize the information for fast and accurate decision making used the data heuristic available. Some technologies are supporting back-end concerns such as storage and processing but visualization-based data discovery tools focus on the front-end of big data such as helping explore the data easily and understand it completely.

Visualization-based tools used for data discovery allow the business users to crush frantic data sources to produce routine analytical views that can be customized with flexibility and ease of use. These tools have a democratizing effect on business because of their ease of use and intuitive interfaces. Also data analysis and visualization can be done by a large number of users with minimal training.

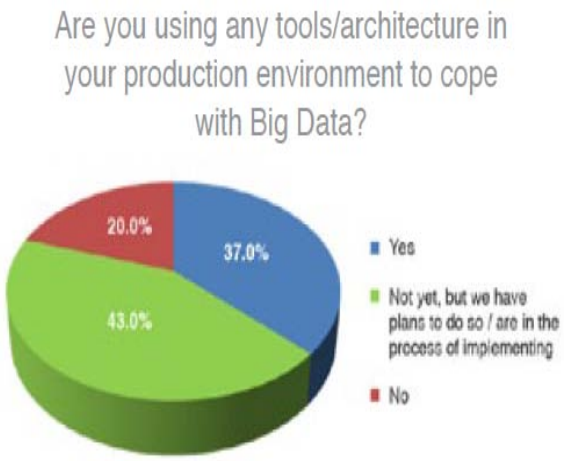

Figure 2 Industry-Based Usage of Big Data Tool [6]

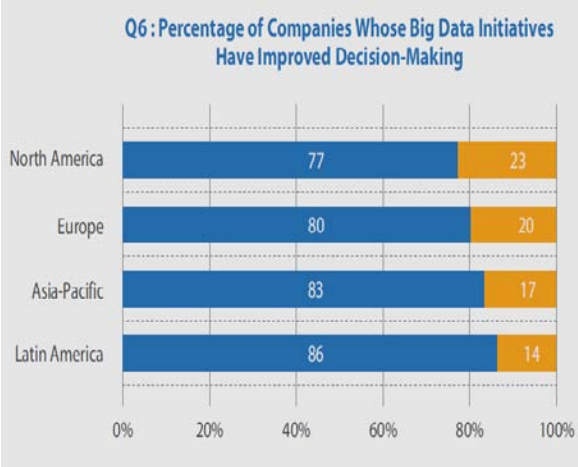

Figure 3 Percentage of companies whose big data initiatives have improved decision making [4]

80% of the companies are already using or planning to use visualization-based tools to cope with big data (see Figure 2), and how these tools have improved their decision making (see Figure 3).

10. Features of Visualization Tools

Below are some of the major features the visualization tools must have so that the end users can analyze and visualize data with ease:

- Interactive graphics (including charts, graphs, tree maps, maps, etc.) so that the data can be visualized in the best possible way.
- Should require little or no any programming skills because to visualize data, code does not play any role.
- Data should be presented using data structures easy-to-understand by non-technical users too.
- Should produce a quick summary for each variable and display it in an easy-to-read format.
- Should support analytics of the big data.
- It should support creation of reports in real-time and with ease.
- Should have the ability to isolate errors or fix simple data problems.
- Should be able to filter data by values or categories.
- For fast access to big data, it should hold data in-memory.
- It should allow users to share answers and insights easily.

Along with these features, the tools should be easily brought in action by the users in line with their existing tasks, and should maintain control and security over the data.

11. Overview of Existing Tools

Wherever ParaView: This is an open-source, multi-platform data analysis and visualization application. The application is build on the Visualization Tool Kit (VTK) libraries, which provide visualization services for data so that data can be analyzed using qualitative and quantitative techniques [7]. It can run on supercomputers to analyze big datasets and also on personal computers for small amount of data and is widely used in many organizations, universities, and industries. The application is very easy to use, the users just need to open data files and start visualizing them.

Some of the features of ParaView are: the capability to visualize large datasets, supports a variety of input/output and file formats, have interactive user interface, can run on different machines using client-server approach, and it is scripted in the python language.

Hadoop: Hadoop is a framework of tools which supports running of application on big data. It uses the MapReduce algorithm, where the data is processed on different computers connected over a network, to run the applications and a storage part known as Hadoop distributed file system (HDFS) [8]. There is a master computer, consisting of job tracker, name node, task tracker and data node and salve computers, consisting task tracker and data node. Job tracker divides the big task into smaller once and gives these tasks to different task trackers. And the name node keeps track of which data is residing where.

The application can analyze huge amounts of data. Some of the key features are: scalable storage platform as the number of slave computers can be increased, cost-effective storage as the slaves are not the supercomputers, can derive useful insights from data, fast data processing, and fault tolerance as different copies of data reside in the same computer and also keep a backup of the table maintained by the name node. 56% companies are using NoSQL data store instead of RDBMS, 30% use Hadoop processing platform and only 12% use real-time event processing tools (see Figure 4).

Figure 4 Products used in companies for big data. [6]

12. Conclusion

The paper discusses about big data and big data visualization. The paper also provides a deep understanding of various challenges that prove as a hindrance to data visualization. Further we have discussed about visualization tools and the desirable features of visualization tools for data discovery and how these tools help in deriving value from big data. At the end two famous visualization tools ParaView and Hadoop are discussed, listing some of their features too.

13. References

[1] Erdman, Arthur G., Daniel F. Keefe, and Randall Schiestl. "Grand challenge: Applying regulatory science and big data to improve medical device innovation." Biomedical Engineering, IEEE Transactions on 60, no. 3 (2013): 700-706.

[2] Dan Vesset, Henry D. Morris, "Unlocking the Business Value of Big Data: Infosys BigDataEdge", February 2013.

[3] HT education times, July 01 2015.

[4] The Emerging Big returns on big data, TCS-Big-Data-Global-Trend-Study-2013, Mar 21, 2013.

[5] Montague, John Joseph, et al. "seeing the trees & understanding the forest." 2014.

[6]http://www.gigaspaces.com/sites/default/files/product/BigDataSurvey_Report.pdf.

[7] https://en.wikipedia.org/wiki/ParaView, last access on 16.01.2016.

[8] https://en.wikipedia.org/wiki/Apache_Hadoop, last access on 16.01.2016.