International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4 (2016) pp 2801-2804

Data Visualization: Enhancing Big Data More Adaptable and Valuable

A. S. Syed Fiaz
Assistant Professor, Department of Computer Science and Engineering,
Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India.

N. Asha
Assistant Professor Senior, SITE, VIT University, Vellore, Tamil Nadu, India.

D. Sumathi
Assistant Professor, SITE, VIT University, Vellore, Tamil Nadu, India

A.S. Syed Navaz
Assistant Professor, Department of Computer Science,
Muthayammal College of Arts & Science, Namakkal, Tamil Nadu, India.

Abstract
The main focus of this paper is on Big Data and Data Visualization techniques which together make the usage of data analytics more efficient and valuable. The term 'Big Data', is a phenomenon that is characterized by the rapid expansion of raw data. Big Data is often voluminous and tends to rapidly change, making it challenging to get a handle on and difficult to access. The majority of tools available to work with Big Data are complex, and most enterprises don't have the sufficient experts to perform the required data analysis. Thus Data visualization techniques simplify this challenge and create an opportunity in analyzing and controlling the data in much more efficient way.

Keywords: Big Data, Data Visualization, Data Analytics, Business Intelligence (BI), Visual Analytics.

Introduction

Big data is a technology to transform analysis of data-heavy workloads, but it is also a disruptive force. Big data is about wider usage of existing data, addition of new sources of data, and analytics that examine deeper by using new tools in a more timely way to increase efficiency or to enable new business models [5]. Today, big data is becoming a business imperative because it enables organizations to accomplish several objectives:

- Apply analytics beyond the traditional analytics use cases to support real-time decisions, anytime and anywhere
- Used in data-driven decision making
- Allow people in all roles to explore and analyze information and offer insights to others
- Optimize all types of decisions, that are made by experts or automated systems
- Improve big business outcomes and manage risk throughout the period

In short, big data [9] provides the capability for an organization to reshape itself into a contextual enterprise, an organization that dynamically adapts to the changing needs of its individual customers by using information from a wide range of sources [1]. Although it is true that many businesses use big data technologies to manage the growing capacity requirements of today's applications, the contextual enterprise uses big data to enhance revenue streams by changing the way that it does business. Big data applications are available in mobile phones, telecommunications, social networking sites, sensor networks, scientific research, astronomy, atmospheric science, life sciences, medical science, government data, natural disaster and resource management, web logs [10].

Big data means for performance and capacity just because of the 3 V's as listed below: Volume-the size of data now is larger than terabytes and peta bytes. The magnificent scale and rise of size makes it difficult to store and analyze using traditional tools. For example, Face book ingests 500 terabytes of data every day [10]. Velocity-It indicates the time duration for analyzing the data, therefore big data should stream data as efficient as it maximizes the entire data value. Variety-Big data comes from a variety of sources. Traditional database systems were designed to address smaller volumes of structured data and fewer data structures whereas Big Data also deals with geospatial data, 3Dimension data, audio files, video files and unstructured data, including log records and social media.

The main differences between big data [4] use cases and traditional data warehouse or business intelligence (BI) applications are the nature and speed of the data under consideration. Typically, big data applications are thousands of times larger and require faster response time than traditional BI applications.

Table 1: Big data use cases by dimension.

Big Data		
Dimension	**Use Cases**	**Capabilities**
1. Volume	1. Data warehouse augmentation (Integrate big data and data warehouse capabilities to increase operational efficiency.)	1. Data warehousing (Deliver deep operational insight with advanced in-database analytics.)
2. Velocity	2. Big data exploration (Find, visualize, and understand all big data to Improve business knowledge.) 3. Enhanced 360° View (Achieve a true unified view, incorporating internal and external sources.)	2. Stream computing (Drive continuous analysis of massive volumes of streaming data with sub milli second response times.)
3. Variety	4. Operations analysis (Analyze various machine and operational data for improved business results.)	3. Apache Hadoop-based analytics (Process and analyze any data type across commodity server clusters.)
4. Veracity	5. Security and intelligence extension (Lower risk, detect fraud and monitor cyber security in real time.)	4. Stream computing (Monitor for truth)

HDFS

The concept of Big Data is incomplete without these two terminologies; Hadoop Distributed File System (HDFS) and Map reduce Framework. Meanwhile HDFS is the first building block of a Hadoop cluster which eventually distributes the massive data into clusters. Moving the computing to the location where the data resides, has created an greater impact in processing the huge volume of data efficiently using a distributed file system instead of using the traditional central system for accessing the data [17]. A single large file is split into blocks, and the blocks are distributed among the nodes of the Hadoop cluster [17].

MapReduce Framework

Whereas Map Reduce [17] is a software framework which in turn processes vast amount of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a consistent, fault-tolerant manner. Map Reduce processes in such a way that it splits the input data into multiple independent chunks which are moved to the map tasks for parallel processing. The map task which in turn forwards the processed data to the reduce tasks as the inputs. The processed data of both the map task and the reduce task are stored in a file system [5]. The framework is concerned of the scheduling tasks, monitoring them and re-executes the failed tasks. The Map Reduce framework and the Hadoop Distributed File System are running on the same set of nodes since both the Storage nodes and the computing nodes are same. This pattern allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high cumulative bandwidth across the cluster.

Issues in Big Data

For decades, businesses have collected data, analyzed it using a variety of Business Intelligence (BI) tools [2], and generated reports. The process continued for a long duration of time, but eventually a few highly trained data analysts were able to pull the desired results and figures. Businesses are finding that this traditional reporting process does not work nearly as well for big data, and certainly is not sufficient to capture the potential value that big data represents. Therefore, it represents both a challenge and an opportunity. The challenge [7, 8] is related to how this volume of data is controlled, and the opportunity is related to how the effectiveness of this data is enhanced by properly analyzing the information.

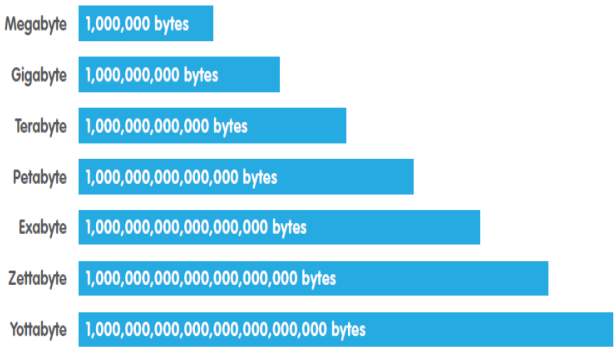

Figure 1: Data Sizes

Data analytics and visualization [3] are not new. The primary challenges stem from what are commonly termed the "three Vs" of big data: volume, variety, and velocity. Most traditional reporting and data mining tools cannot handle the vast volume of big data-although the variety and velocity of the data often present even greater challenges [15, 13]. Thus visually representing the data have made the big data much more adaptable and valuable.

Data Visualization

Data visualization is becoming an increasingly important component of analytics in the age of big data. As the volume and variety of big data grows, data visualization becomes even more important to instigate a collaborative dialogue between these groups [10].

Dealing with large volumes of data, often leads to chaos and misinterpretations between the data statistics, thus visually grouping together with many data points significantly enhances the quality of the data analytics and provides a much convenient way for the business executives, data scientists in understanding the relationships between the data. Interpreting the data visually helps in understanding the data and to quickly decide where to focus the research [10].

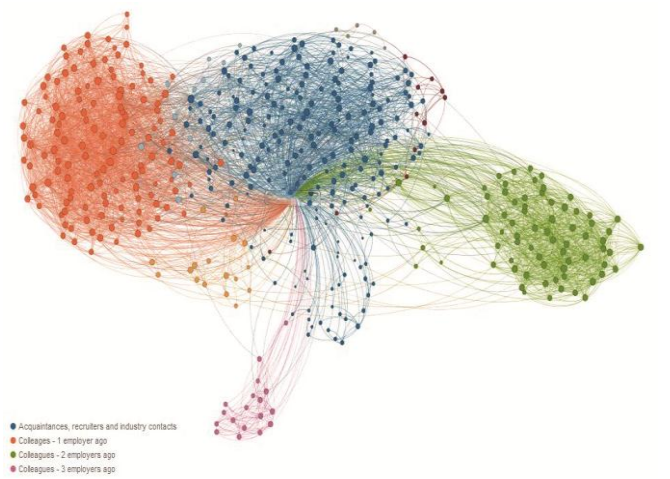

Figure 2: Data Visualization of Linked In Profile

The above figure 2 describes a user's personal LinkedIn network in a stunning visual while the links starts from the centre. It coordinates the entire datasets from acquaintances to industry contacts from various professional stages in career [12, 14]. Handling and presenting millions of rows of data creates a big headache for the data analysts. Thus major organizations have approached this problem in one of the two ways:

- Building 'samples' of the data that provides an easier and convenient way in analyzing and presenting the data.
- Creating template charts and graphs that can easily manipulate certain types of information's.

But both these techniques have failed the potential for big data. Instead of building samples or creating template charts and graphs that deals only with part of the data, we consider pairing big data with visual analytics which can use entire data and provides an automated help in selecting the best ways to present the data [16, 18]. Thus visual analytics have made the data scientists to easily deploy the entire data in a much approachable way of presentation. For an easier understanding, consider your data to be great but messy. Visual analytics acts as the master filmmaker and the gifted editor who makes your data much more efficient and acceptable, who brings the story to life. Spending hours of time in analyzing the orphan database would lead to a hectic task for the analyst's teams and also creates a greater impact on the investments. To overcome these difficulties, make data as ubiquitous as possible.

Data visualization can be adaptable in the organizations in any such of the following ways:

- Making the analytics team familiar with the collaboration tools to share, leverage and build on the data. It should create a free working space that encourages both the business users and the organization executives to participate.
- Creating an awareness of the various data visualization tools to the analysts but which in turn will lead to the use of legacy software's, the business intelligence tools or visual analytics tools. All these will significantly pay dividends in efficiency.
- The employees with design backgrounds can be embedded with the data visualization skills that transform them into analytics teams. Deploying low cost open source tools that help visualizing large data sets easier [7].

Conclusion

Data visualization may not be a perfect solution for the data analytics but on the other side it provides the best advantage of having a common language between the executives, functional leads and data scientists to share a space to discuss about the data together. All these three groups' works independently in a thriving enterprise, but data visualization supersede these differences and provides a single platform by creating a pictorial story that makes everyone understand in a timely manner. Through data visualization, organizations can control and analyze the real value of big data by accelerating the thorough understanding of the data, speeding insights and enabling the organization executives to make quicker decisions on the advantageous business opportunities.

References

[1] Accenture, "Why Big Data needs Visualization to succeed", 2014.

[2] Apache Software Foundation, "MapReduce Tutorial", 2008,https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.pdf.

[3] Accenture, "Why Big Data needs Visualization to succeed", 2014.

[4] "Big Data Business Benefits Are Hampered by 'Culture Clash'," Gartner, September 12, 2013.

[5] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly-Business Intelligence Research, Vol. 36 No. 4, pp. 1165-1188, December 2012.

[6] A.S.Syed Navaz & A.S.Syed Fiaz, "Load Balancing in P2P Networks using Random Walk Algorithm" March-2015, International Journal of Science and Research, Vol No-4, Issue No-3, pp.2062-2066.

[7] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges", The Scientific World Journal, Volume 2014.

[8] Peng Hu and and Wei Dai, "Enhancing Fault Tolerance based on Hadoop Cluster", International Journal of Database Theory and Application, Vol.7, No.1, pp.37-48, ISSN: 2005-4270, 2014.

[9] "Performance and Capacity Implications for Big Data", IBM-International Technical Support Organization, January 2014.

[10] "Performance and Capacity Implications for Big Data", IBM-International Technical Support Organization, Jan 2014.
[11] A.S.Syed Navaz, M.Ravi & T.Prabhu, "Preventing Disclosure of Sensitive Knowledge by Hiding Inference" February 2013, International Journal of Computer Applications, Vol 63-No 1. pp. 32-38.
[12] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer, "Enterprise Data Analysis and Visualization: An Interview Study", IEEE Transactions on Visualization and Computer Graphics, October 2012.
[13] Shilpa, Manjit Kaur, "Big Data Visualization tool with Advancement of Challenges", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3, ISSN: 2277 128X, March 2014.
[14] Seref Sagiroglu, Guygu Sinanc, "Big Data: A Review", IEEE Transactions, 2013.
[15] "Special Issue on Big Data",http://www.researchtrends.com/wpcontent/uploads/2012 /09/Research_Trends_ Issue30.pdf, September 2012.
[16] Suman Arora, Dr.Madhu Goel, "Survey Paper on Scheduling in Hadoop", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, ISSN: 2277 128X, May 2014.
[17] Thirumala Rao B., N.V.Sridevi, V.Krishna Reddy, L.S.S.Reddy, "Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing", Global Journal of Computer Science and Technology, Volume 11, Issue 8 Version 1.0, ISSN: 0975-4172, May 2011.
[18] UzmaShafaque, Parag D. Thakare, Mangesh M. Ghonge, Milindkumar V. Sarode, Ph.D, "Algorithm and Approaches to Handle Big Data", International Journal of Computer Applications (0975-8887), X-PLORE 2014.
[19] Venkata Narasimha Inukollu, Sailaja Arsi1 and Srinivasa Rao Ravuri, "Security Issues Associated With Big Data In Cloud Computing", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.
[20] A.S.Syed Navaz, C.Prabhadevi & V.Sangeetha "Data Grid Concepts for Data Security in Distributed Computing" January 2013, International Journal of Computer Applications, Vol 61-No 13, pp 6-11.

Author's Biography

A.S. Syed Fiaz received his ME (CSE) from Sona College of Technology, Anna University Chennai and BE (CSE) from Sona College of Technology, Anna University Chennai. He has researched and published in International journals and working as Editor Board Member & Reviewer for International journals. Currently he is working as an Assistant Professor in the Department of Computer Science and Engineering at Dhirajlal Gandhi College of Technology, Salem. His areas of interest are Cloud Computing, Computer Networks.

Asha N. working as Assistant Professor (Senior) in School of Information Technology and Engineering (SITE), VIT University, Vellore District, Tamil Nadu, India. She completed her master degree-(M.E) in Computer Science Engineering. She is pursuing Ph.D in the area of software testing. Her main research interests include Software Engineering, Software Testing, Re-engineering, Computer Networks, Data Mining and Big Data. She has many publications in national and international journals and conferences to her credit. Asha. N can be contacted by e-mail at nasha@vit.ac.in

D. Sumathi working as a Assistant Professor in School of Information Technology and Engineering (SITE), VIT University, Vellore District, Tamil Nadu, Indiasince 2006.Her area of interest includes Big Data, Mobile Networks, Next generation networks,Wireless Netwoks and Network Security. She completed her master degree-(M.Tech) in Information Technology. She is pursuing Ph.D in the area of Next Generation Networks. Sumathi.D can be contacted by e-mail at dsumathi@vit.ac.in

A.S. Syed Navaz received M.Sc in Information Technology from K.S.Rangasamy College of Technology, Anna University Coimbatore, M.Phil in Computer Science from Prist University, Thanjavur, M.C.A from Periyar University, Salem, PGDCA in Erode and Pursuing Ph.D in the area of Wireless Sensor Networks. He has researched and published papers in International journals and working as an Editorial Board Member in 7 International Journals & Reviewer for 15 International journals including Springer. He is a Member of 14 International Social Bodies. His biography is listed in "Marquis Who's who in the World" (32nd & 33rd Edition) USA. Currently he is working as an Assistant Professor in the Department of Computer Science at Muthayammal College of Arts & Science, Namakkal, India. His Research areas are Wireless Sensor Networks, Mobile Computing & Image Processing.