Challenges and Opportunities with Big Data Visualization

Rajeev Agrawal, Anirudh Kadadi
Department of CST
North Carolina A&T State University
Greensboro, USA.
ragrawal@ncat.edu,
akadadi@aggies.ncat.edu

Xiangfeng Dai
Department of CSE
North Carolina A&T State University
Greensboro, USA
colin.sunset@gmail.com

Frederic Andres
Digital Content and Media Sciences Research Division
National Institute of Informatics
Tokyo, Japan
andres@nii.ac.jp

ABSTRACT
In this big data era, huge amount data are continuously acquired for a variety of purposes. Advanced computing, imaging, and sensing technologies enable scientists to study natural and physical phenomena at unprecedented precision, resulting in an explosive growth of data. It is a huge challenge to visualize this growing data in static or in dynamic form. Most traditional data visualization approaches and tools can't support at "big" scale. In this paper, we identified the challenges and opportunities in big data visualization and review some current approaches and visualization tools.

Categories and Subject Descriptors
Big data visualization

General Terms
Design, Performance

Keywords
Visualization, Scalability, Data reduction, Big Data.

1. INTRODUCTION
Data visualization is very useful for people to understand data in a graphical manner. However, in this big data era, data grows constantly bigger and bigger. "Normally, a rough threshold is one million or more data cases for big data visualization [1]." Big volume size data is extremely difficult to be presented meaningful and valuable. Traditional data visualization is inadequate to handle big data at this point. For example, many data sets are too large to fit in memory and may be distributed across a cluster [2].

The new data visualization must come up with better ways to process, analyze and visualize huge amount of complicated data. Big data visualization brings new research challenges and opportunities. Some data is structured, organized and stored in traditional relational databases, but some other big volume size data, including streaming data, documents, images, videos, music, emails, and messages, etc, is unorganized and unstructured. According to Gartner 3Vs definition in 2001, big data has three characteristics: volume, velocity and variety. Volume means amount of data. The amount of data for big data should be extremely large. Velocity is about speed of data in and out and variety is the range of data types and sources. Later in 2012, Gartner updated the definition of big data as high volume, high velocity and high variety.

Recently, real-time analysis of big complex data should be considered as well. There are two more characteristics that are also important to big data, which are value and veracity. Having access to big data is good but unless we can turn it into value it is useless. Veracity is about the trustworthiness of the data. Since data grows constantly bigger and bigger, the traditional ways of presenting data has reached its limitations [3]. For example: how to present the whole of imaginable data with big volume size on limited screens (think about examining zettabytes rather than gigabytes). Moreover, it disrupts fluent interaction to query large data. It is also important to display big volume data in real-time. These limitations bring challenges such as ***perceptual scalability***, ***real-time scalability***, and ***interactive scalability***. When it comes to today's big data, how it looks can help convey information but it needs to be more than just beautiful and superficial. It has to work, show multiple dimensions, and be useful. Big data visualization also brings opportunities of presenting better ways to visualize big data such as ***data reduction***, ***reducing latency***, etc. In Health InfoScape, 7.2 million patient records have been collected from GE's proprietary database, and is represented under some conditions that commonly affect Americans today. The numbers and percentages represent general trends. By visualizing the relationship among different ailments, one may gain various insights about condition associations. This example can clearly give us a new insight in visualization to understand health and take better care of ourselves. Rest of the paper is organized as follows: in section 2, we discuss the challenges faced in big data visualization; section 3 covers the opportunities provided by these challenges. Some representative big data visualization tools are covered in section 5 and conclusions are mentioned in section 6.

2. CHALLENGES
Visualization of a large data set is a demanding task. The traditional ways of presenting data reached a few limitations along with data grows constantly extremely large. Visualization tools and techniques should also able to help the users to identify missing, erroneous or duplicate values. It is challenging to solve the limitations such as perceptual scalability, real-time scalability, and interactive scalability. In this section, we discuss these challenges.

MEDES '15, October 25-29, 2015, Caraguatatuba, Brazil
 ISBN 978-1-4503-3480-8/15/10…$15.00
http://dx.doi.org/10.1145/2857218.2857256

2.1 Perceptual Scalability

Human perception: Human eyes have difficulty to extract meaningful information when the data becomes extremely large. Not many existing visualization systems are designed to scale nicely to present meaningful and quality information for human perception.

Limited screen: Data is becoming simply larger and larger, it is challenging when visualization displays too many data items or features on limited screen, especially a dataset with a billion entries. Too many data to present on the limited screen that the resulting of which the visualization is too dense to be useful to the users [4]. 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 users' perceptual and cognitive capacities [5]". The limitation of screen resolution forces us to explore novel ways to display and visualize information using various abstraction techniques. It is even more challenging to present huge data on mobile devices because of smaller screens and resolutions.

2.2 Real-time Scalability

It is important to provide users with visual real-time information and it is also important to make real-time decisions based on available data [9]. However, huge amounts of data would be too large to process in real-time. Most visualization systems are only designed to handle data under a certain size because many data sets are too large to fit in memory and query large data could incur high latency. It is challenging to overcome limitations like data connectivity and limited storage and data processing capabilities in real time. For example, in recent years, social media has become ubiquitous and important for social networking and content sharing. Twitter is a popular micro blogging service [6]. It has 645 million users and they are generating more than 58 million tweets and 2.1 billion queries every day per day (reported in July, 2014) [7]. The twitter has attributes like: latitude and longitude of the device, the time the tweet occurred, the client application used, the type of device, and the language of the tweet. It is time consuming to visually explore twitter datasets and query information with those attributes, even display the simple scatterplots of tweets is not straightforward in real time.

2.3 Interactive Scalability

Interactivity is amplifying the benefits of data visualization. Interactive data visualization can help us understand the insight of data faster and better. However, it takes time to process and analyze data before visualization, especially huge amounts of data. For example, if actions taken involve querying large data and complex algorithms, it might disrupt fluent interaction. Even with data reduction strategies, data cubes can still be too large for smooth interaction [5, 8, 9]. And, the visualization system may even freeze for an extended period of time or crash while trying to present huge amounts of data. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

3. OPPORTUNITIES

Challenges bring opportunities. This section covers some opportunities and strategies to address the above challenges of big data visualization. Using ***data reduction*** techniques which include ***Sampling, Filtering and Aggregation*** to reduce the big data to a smaller amenable data before visualization. They could address the problems such as: large-scale data by performing resolution reduction on query results and display all the necessary information in a limited display. We will also mention **reducing latency** techniques, which are ***Parallelize Data Processing and Rendering, Hiding disk latency*** and ***Pre-compute Multivariate Data Tiles.*** We will discuss these approaches in more detail in the following sections.

3.1 Data Reduction

Data reduction strategies include sampling, filtering and binned aggregation, which reduce big data to smaller amenable data before visualization [10].

***Sampling*:** Reduction techniques are based on sampling. Every dataset is a sample. When given a probability value, it returns roughly that fraction of data as the result.

Filtering: Filtering techniques are necessary to explore and query large datasets. Given a set of desired conditions of querying the data, return the elements that meet these conditions.

Binned Aggregation: Data is grouped into subsets, and summaries of the subsets are returned as the result. Binning aggregates data by counting the number of data points falling within each predefined bin.

3.2 Reducing Latency

Pre-computed Data: To improve ***interactive scalability*** such as panning, zooming, dragging and dropping. Pre-computed data is a popular strategy in visualization. It can support quick exploration rather than generate image tiles intended for direct display. For example: Google Maps, Hotmap and Data Cubes.

Parallelize Data Processing and Rendering: Data tiles could be very large in the process of aggregation. It depends on binning resolution. If the data tiles have more than millions values, it will increase latency of aggregation. To speed up, visualization system could use a dense indexing scheme that simplifies parallel query processing. In web browsers, the web application could use WebGL to leverage parallel processing on the GPU. For example: imMens system [5].

A predictive middleware: A predictive middleware that will reside between the frontend visualization interface and the backend data store and will predict, pre-fetch and cache relevant data. A predictive middleware improves ***real-time scalability*** and ***interactive scalability***. It hides the latency of the backend data store by prefetching and caching the data that will be needed in the near future.

4. RELATED WORK

We will discuss some related works in this section. This is by no means an exhaustive review of all the research work in data visualization. We are presenting here few examples that address the data visualization challenges as discussed previously.

4.1 Data Cube & Nanocube

Data cubes are multidimensional extensions of 2-D tables [14]. There are a few visualization systems are built on top of data cubes. Data cubes are structures that perform aggregations across

every possible set of dimensions of a table in a database for perceptual scalability. And it also uses pre-computed data strategy to address interactive scalability, which could support quick exploration [15]. Nanocube does good job on **real-time scalability** to explore multidimensional, spatiotemporal datasets. It improved from data cubes technology, which enables real-time exploratory visualization. It supports queries such as: counting events in a spatial region (e.g. Chicago); counting events by hour, day, week, or month over a period of years. Most importantly, query times average under a millisecond for a single thread running on a computer that ranges from a laptop, to a workstation, to a server-class computing node [11]. Figure 1 shows an example of visualizing which device is more popular for tweeting. Blue color indicates iPhone and orange color indicates Android.

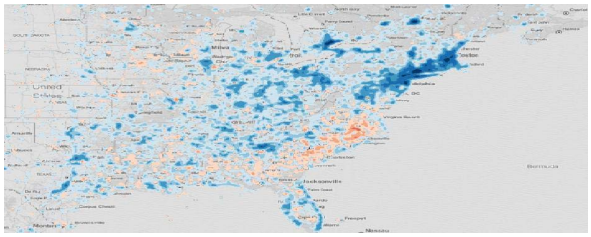

Figure 1. US-wide map of relative device popularity.

4.2 imMens

imMens is a browser-based visualization system, which is an example of ***data reduction*** strategy [5]. It takes into account data size in the visualization frontend to scale using visual summaries of the underlying data. The ***binned aggregation*** is the primary data reduction strategy of imMens. A design space of binned plots is described for the variables like: numeric, ordinal, temporal and geographic. Then binning schemes for different data types are described and stored in the databases. At the end, the data are grouped into adjacent intervals over a continuous range or binned at various levels of granularity [12]. Figure 3 shows a visualizing a dataset of Brightkite user check-ins. We can see the density of check-ins by aggregation. It also uses ***reducing latency*** strategy, which enable WebGL for data processing and rendering on the GPU."

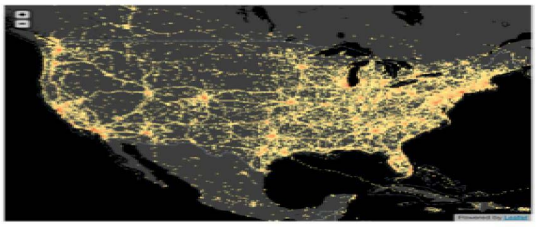

Figure 2. Visualizing a dataset of Brightkite user check-ins

4.3 ScalaR

ScalaR is a three-tiered data visualization system that provides a web-based, map-style interface for viewing large data sets. It takes SQL queries as input, and returns a visualization of the query results. Instead of running the original query, ScalaR uses ***data reduction*** strategy (***aggregation***, ***sampling*** and ***filtering***) to reduce the size of the results, which improves ***interactive scalability*** [4]. When the expected result is too large to be rendered, it dynamically performs resolution reduction. Users can see the general shape of the western coast of California/Northern Mexico on Figure 4. Figures 5 show the results of increasing the resolution to 10000 points.

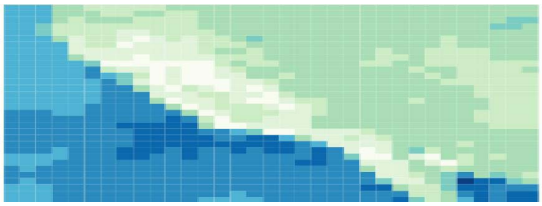

Figure 3. 1000 points resolution.

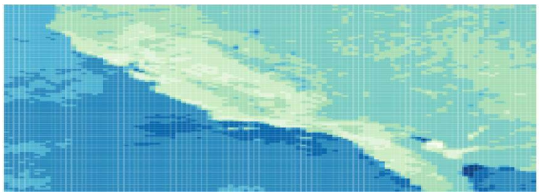

Figure 4. 10,000 points resolution.

5. VISUALIZATION TOOLS

This section covers some related visualization tools of big data. It includes service, platform, widgets and library. These tools could help us visualizing big data. Table 1 shows the list of popular visualization tools, categorized as a service, a library or a platform [8].

Table 1. Visualization tools comparison

Tools	Category	Open Source
CartoDB	service	no
Processing.js	library	yes
Gephi	platform	yes
D3.js	library	yes
Weave	platform	yes
Dundas Dashboard	service	no
SIMILE Widgets	widgets	yes
Datawatch	platform	no

5.1 CartoDB

CartoDB (https://cartodb.com) is a web service for mapping, analyzing and building applications with data. An interactive map communicates quickly and to the point complex information. Figure 6 shows Los Angeles Times creates an interactive visualization map about pollution information in the city of Los Angeles by using CartoDB service.

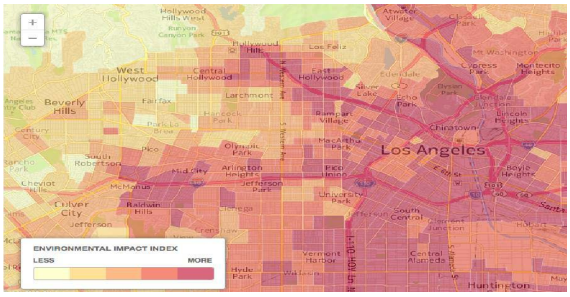

Figure 6. Los Angeles Times pollution map

5.2 Processing.js

Processing.js designed for web. It is a JavaScript library, which is the sister project of the popular Processing visual programming language. Processing.js uses web standards and doesn't any plug-ins to visualize data. Figure 7 shows the Letter-pairs analysis project. This project uses Processing.js library to read a text and to calculate in real time the number of times each pair of letters appears in a given text.

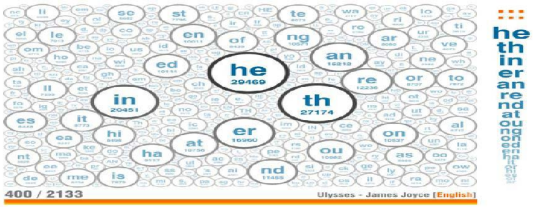

Figure 7. Letter-pairs analysis project

5.3 Gephi

Gephi is designed for complex and large number datasets of networks, systems, and graphs. It is open-source and it provides interactive visualization and dynamic data exploration [13]. Figure 8 shows the brain network of the C. Elegans worm exported from Gephi.

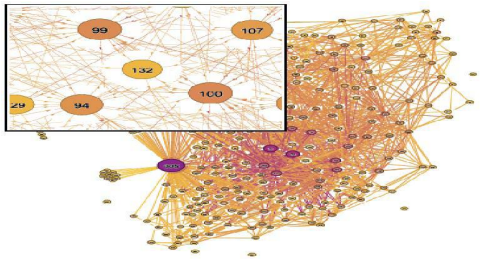

Figure 8. Brain network of the C. Elegance worm exported from Gephi

5.4 Data-Driven Documents (D3)

It is a small, free JavaScript library for manipulating HTML documents based on data. D3 can quickly visualize data as HTML or SVG, handle interactivity, and incorporate smooth transitions and staged animations into the web pages. Figure 9 shows the real-time interactive logo for the 2012 Open Knowledge Festival in Helsinki, which created with D3 JavaScript library.

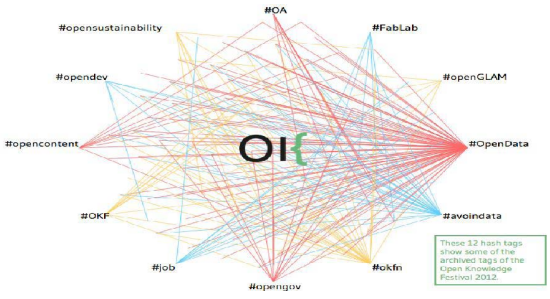

Figure 9. Real-time interactive logo for the 2012 Open Knowledge Festival

5.5 Weave

Weave (https://www.oicweave.org/) is an application development visualization platform. It could take many types of datasets, and then generate customized web-based visualizations. It supports the actions like: integrate data, analyze data and visualize data. However, it takes large amounts of memory. Figure 10 shows the visualization of foreclosures in Lowell, Mass.

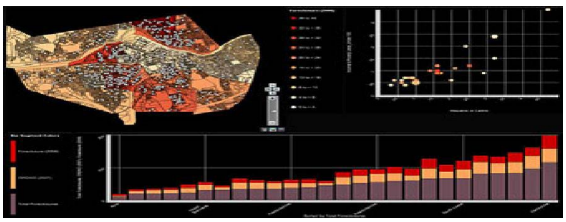

Figure 10. visualization of foreclosures in Lowell, MA

5.6 Dundas Dashboard

Dundas Dashboard (http://www.dundas.com/dashboard/) is a web-based platform. It lets users customize interactive dashboards and the dashboards can connect with users' Property Management System (PMS) databases directly. Therefore, users can visualize and analyze data from across the organizations. It also provides more than 50 powerful data visualization tools such as: interactive charts, gauges, maps, scorecards, etc. Figure 11 shows the hospitality industry style tool to provide a visual breakdown and interactive analysis of important KPIs (Key Performance Indicators).

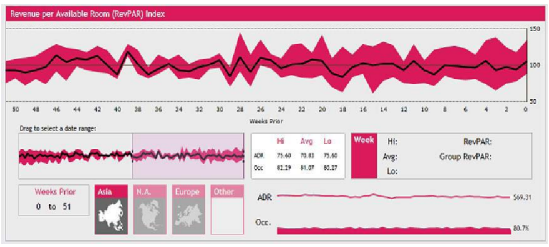

Figure 11. Hospitality industry style

5.7 SIMILE Widgets

SIMILE (http://www.simile-widgets.org) is developed by MIT, which includes widgets such as: exhibit, timeline and timeplot. SIMILE is focused on developing robust, open source tools that empower users to access, manage, visualize and reuse digital assets. SIMILE is open-source as well.

5.7.1 Exhibit, Timeline & Timeplot:

Exhibit lets you easily create web pages with advanced text search and filtering functionalities, with interactive maps, timelines, and other visualizations. Figure 12 shows the visualization of US Cities by Population that created by Exhibit Widget. Timeline is use to make Interactive timelines like the one of the JFK assassination timeline in Figure 13. It is designed for plotting time series and overlay time-based events over them. This timeplot is good for visualizing data like price, stock, and cost. An example is shown in figure 14.

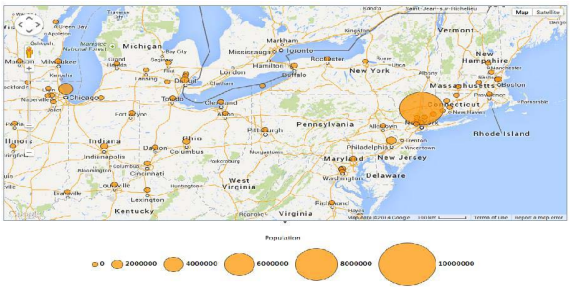

Figure 12. US Cities by Population

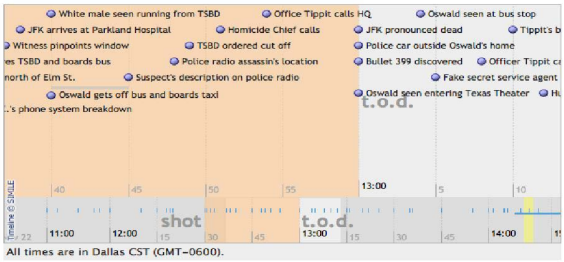

Figure 13. The JFK Assassination Timeline

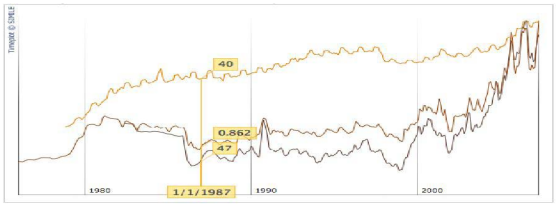

Figure 14. Energy Prices in the U.S. since 1975

5.8 Datawatch Desktop

Datawatch Desktop (http://www.datawatch.com) provides a few strong visualization tools, such as: analysis of real-time streams, visual data discovery, data visualizations and time series data analysis. By using Datawatch Desktop, we could explore different types of datasets, build a customized view, set up filters of datasets and generate customized visualizations in just minutes. It tackles data in three steps: transformation, distribution and optimization.

6. CONCLUSION

In this big data era, the traditional ways of presenting data reached a few limitations and traditional data visualization is inadequate to handle big data at this point. In this paper, we have identified some challenges of data visualization in big data era. According to the challenges, we discussed some opportunities and strategies to address the above challenges. We also reviewed some related works, current research approaches and related tools. These approaches and tools could provide new ways for visualizing big data. As big data sets are generated through collaboration of multiple researchers, it is expected that visualization tools would have the capabilities to enable this feature.

7. ACKNOWLEDGMENTS

This research is supported by National Consortium of Data Science (NCDS), Chapel Hill and The Japan Society for the Promotion of Science (JSPS) under the Grants-in-Aid for Scientific Research Challenging Exploratory Research Project.

8. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825.

[2] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[3] Jin, X., et al. (2015). "Significance and Challenges of Big Data Research." Big Data Research 2(2): 59-64.

[4] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM, New York, NY, 526-531.

[5] Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[6] Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[7] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[8] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10.

[9] Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590.

[10] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33.

[11] Lins, L., Klosowski, J.T. ; Scheidegger, C. 2013. "Nanocubes for Real-Time Exploration of Spatiotemporal Datasets", Visualization and Computer aGraphics, IEEE.

[12] Liu, Z.,Jiangz, B., and Heer, J. 2013. "Real-time Visual Querying of Big Data" Eurographics Conference on Visualization (EuroVis).

[13] Bastian M., Heymann S., Jacomy M. 2009. “Gephi: an open source software for exploring and manipulating networks”, International AAAI Conference on Weblogs and Social Media.

[14] Mukherjee, R. and Seeja K.R. 2009. “Building Parallel Rolap Data Cube for Shared Nothing Architecture” , Conference on Recent Developments in Computing and Its Applications elopments in computing and it's applications.

[15] Gray, J., et al. 1997. "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals." Data Min. Knowl. Discov. 1(1): 29-53.