Abstract

Big Data analytics plays a key role through reducing the data size and complexity in Big Data applications. Visualization is an important approach to helping Big Data get a complete view of data and discover data values. Big Data analytics and visualization should be integrated seamlessly so that they work best in Big Data applications. Conventional data visualization methods as well as the extension of some conventional methods to Big Data applications are introduced in this paper. This report presents the recent methods used in Bigdata visualization ,it's applications , technological progress and challenges involved.

Contents

1 INTRODUCTION **4**

1.1 Data Visualization Market 4

1.2 The Rise Of Data Visualization 4

2 BIG DATA ANALYTICS **6**

2.1 3Vs Model And New Vs for Bigdata 7

2.1.1 Volume 7

2.1.2 Velocity 7

2.1.3 Variety 8

2.2 The Four Additional Vs 8

2.2.1 Veracity 8

2.2.2 Variability 9

2.2.3 Value 9

2.2.4 Visualization 10

3 DATA VISUALIZATION METHODS **11**

3.1 Conventional Data Visualization Methods 11

3.1.1 Pie Chart 12

3.1.2 Bar Chart 12

3.1.3 Line Chart 12

3.1.4 Area Chart 13

3.1.5 Scatter Plot 14

3.2 Interactive Visualization 14

3.2.1 Selecting 15

3.2.2 Linking 15

3.2.3 Filtering . . . 15
3.2.4 Rearranging or Remapping . . . 15
3.3 Data Visualization Tools . . . 16

4 BIG DATA VISUALIZATION AND CHALLENGES 17
4.1 Challenges in Big Data Visualization . . . 17
4.1.1 Scalability . . . 18
4.1.2 Dynamics . . . 18
4.1.3 Need Of Massive Parallelization . . . 19
4.1.4 Application Architecture and Data Management . . . 19
4.2 Potential Solutions . . . 20
4.2.1 Meeting the Need for Speed . . . 20
4.2.2 Dealing with Outliers . . . 20
4.2.3 Displaying meaningful results . . . 21

5 PROGRESS OF BIGDATA VISUALIZATION 22
5.1 Big Data Visualization Methodology . . . 22
5.1.1 Overview First . . . 22
5.1.2 Zoom and Filter . . . 23
5.1.3 Details-on-Demand . . . 23
5.2 Big Data Visualization Approach . . . 24
5.2.1 TreeMap . . . 24
5.2.2 Parallel Coordinates . . . 25
5.2.3 Semantic Network . . . 25
5.2.4 Sunburst . . . 25
5.3 Properties Of Visualization Tools . . . 26
5.4 Virtual Reality Platform for Scientific Data Visualization . . . 27
5.5 SWOT Analysis Of Current Tools . . . 27

6 CONCLUSION 29

List of Figures

3.1 Standard Pie Chart 12
3.2 Bar Chart 13
3.3 Line Chart 13
3.4 Area Chart 13
3.5 Scatter Plot Diagram 14
3.6 Interactive brushing and linking between histogram plots (top) and geographic map (bottom) of datasets 16

5.1 Treemap 24
5.2 Parallel Coordinates 25
5.3 Semantic Network 25
5.4 SunBurst Visualization 26

Chapter 1

INTRODUCTION

1.1 Data Visualization Market

Data visualization is representing data in some systematic form including attributes and variables for the unit of information. 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. Benefits of data visualization according to the respondent percentages of a survey includes improved decision making,better ad-hoc data analysis,improved collaboration,provide self service capability to end users,time savings etc.

1.2 The Rise Of Data Visualization

Data illustration techniques have been in use since as early as 6200 BC, when the oldest known map was drawn. However, it was not until the eighteenth century when data visualizations went beyond mapping and more abstract measures were introduced, including the ever-popular pie and bar charts.The nineteenth century saw the creation of what many have argued to be the worlds best data visualization: Charles Joseph Minards 1869 visualization titled Napoleons March, which depicts the movement and losses of Napoleons army as it invaded Russia in 1812.After 1975, we witnessed the most rapid advancements in data visualization, which stemmed from the development of software and computer systems.

Data visualizations moved beyond pie and bar charts, and more complex formats began to appear and aid us in processing information. For example, through the use of mind maps, our thought patterns can now be visually organized. Apps like Flipboard and Newsmap have completely reinvented the display of news, while tag clouds have provided another way to discover and search for information. And through network graphs, we can now uncover the connectivity between any number of entities, be they our own social circles, groups of companies or globally dispersed cities.

Chapter 2

BIG DATA ANALYTICS

BigData, according to Wikipedia (Sep 2015), is the term for a collection of data set so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications. The data-sets not only contain structured databases, but also include unstructured databases such as social media data or GPS (Global Positioning System) data.

BigData comes from everywhere influence our life, and so is too big,complex and moves too fast. For example, posting pictures and writing comments on Facebook; uploading and watching videos on YouTube; sending and receiving messages through smart phones;sending voice messages through Whatspapp all count as BigData. To analyse BigData, new analytical methods have to be developed to feed business, government and organization needs.

Distributed computing and parallel processing techniques are widely used in industry for BigData applications. Hadoop (High-availability distributed object oriented platform), the most popular open-source platform for reliable, scalable, distributed computing, is often referred to by BigData researchers. Two main core frameworks in Hadoop: Hadoop Distributed File System (HDFS) and MapReduce, have being deployed in industries for the management of cluster distributed data centers such as Facebook, Google, Yahoo, Amazon. com, eBay and Twitter (hadoop.appache.org).

2.1 3Vs Model And New Vs for Bigdata

According to Gartner 3Vs definition[3], Volume, velocity and variety to characterize the concept of Big Data.

2.1.1 Volume

90% of all data ever created was created in the past 2 years. From now on, the amount of data in the world will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us. If we look at airplanes they generate approximately 2,5 billion Terabyte of data each year from the sensors installed in the engines. Self-driving cars will generate 2 Petabyte of data every year. Also the agricultural industry generates massive amounts of data with sensors installed in tractors. Shell uses super-sensitive sensors to find additional oil in wells and if they install these sensors at all 10.000 wells they will collect approximately 10 ExaByte of data annually. That again is absolutely nothing if we compare it to the Square Kilometer Array Telescope that will generate 1 Exabyte of data per day.

In the past, the creation of so much data would have caused serious problems. Nowadays, Bigdata technologies allows us to manipulate such a huge volume of data

2.1.2 Velocity

The Velocity is the speed at which the data is created, stored, analysed and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers required substantial time to process the data and update the databases. In the big data era, data is created in realtime or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created.

The speed at which data is created currently is almost unimaginable: Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million

emails are sent, around 20 million photos are viewed and 30.000 uploaded on Flickr, almost 300.000 tweets are sent and almost 2,5 million queries on Google are performed.

2.1.3 Variety

In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by organisation is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

There are many different types of data and each of those types of data require different types of analysis or different tools to use. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.

2.2 The Four Additional Vs

Now that the context is set regarding the traditional Vs, lets see which other Vs are important for organisations to keep in mind when they develop a big data strategy.

2.2.1 Veracity

Having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect. Incorrect data can cause a lot of problems for organisations as well as for consumers. Therefore, organisations need to ensure that the data is correct as well as the analysis performed on the data are correct. Especially in automated decision-making, where no human is involved anymore, you need to be sure that both the data and the analyses are correct.

IBM coined Veracity as the fourth V, which represents the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment. Yet they contain valuable information. Thus the need to deal with imprecise and uncertain data is another facet of big data,

which is addressed using tools and analytics developed for management and mining of uncertain data.

2.2.2 Variability

Big data is extremely variable. Brian Hopkins, a Forrester principal analyst, defines variability as the variance in meaning, in lexicon. He refers to the supercomputer Watson who won Jeopardy. The supercomputer had to dissect an answer into its meaning and to figure out what the right question was. That is extremely difficult because words have different meanings an all depends on the context. For the right answer, Watson had to understand the context.

Variability means that the meaning is changing (rapidly). In (almost) the same tweets a word can have a totally different meaning. In order to perform a proper sentiment analyses, algorithms need to be able to understand the context and be able to decipher the exact meaning of a word in that context. This is still very difficult. SAS introduced Variability as additional dimensions of big data. Variability refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs.This imposes a critical challenge: the need to connect, match, cleanse and transform data received from different sources.

2.2.3 Value

All that available data will create a lot of value for organisations, societies and consumers. Big data means big business and every industry will reap the benefits from big data. It is estimated that potential annual value of big data to the US Health Care is $300 billion, more than double the total annual health care spending of Spain. They also mention that big data has a potential annual value of 250 billion to the Europes public sector administration. Even more, in their well-regarded report from 2011, they state that the potential annual consumer surplus from using personal location data globally can be up to $ 600 billion in 2020. That is a lot of value.

Of course, data in itself is not valuable at all. The value is in the analyses done on that data and how the data is turned into information and eventually turning it into knowledge. The value is in how organisations will use that data and turn their organ-

isation into an information-centric company that relies on insights derived from data analyses for their decision-making.

2.2.4 Visualization

This is the hard part of big data. Making all that vast amount of data comprehensible in a manner that is easy to understand and read. With the right analyses and visualizations, raw data can be put to use otherwise raw data remains essentially useless. Visualization of course do not mean ordinary graphs or pie charts. They mean complex graphs that can include many variables of data while still remaining understandable and readable.

Visualizing might not be the most technological difficult part; it sure is the most challenging part. Telling a complex story in a graph is very difficult but also extremely crucial. Luckily there are more and more big data start-ups appearing that focus on this aspect and in the end, visualizations will make the difference.

Chapter 3

DATA VISUALIZATION METHODS

A picture is worth a thousand words especially when you are trying to understand and gain insights from data. It is particularly relevant when you are trying to find relationships among hundreds, or even thousands, of variables to determine their relative importance.

3.1 Conventional Data Visualization Methods

Many conventional data visualization methods are often used. They are: table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart, time line, Venn diagram, data flow diagram, and entity relationship diagram, etc. In addition, some data visualization methods have been used although they are less known compared the above methods. The additional methods are: parallel coordinates, treemap, cone tree, and semantic network, etc.

There are many conventional data visualization techniques which are focused in this document because these techniques have generic features and common understanding. These data presentation should be beautiful, elegant, descriptive, and interpretable in order to convey message to the reader effectively. There are new developed fascinating methods are introducing, but modern approaches have its own implementation problems and no commonality, so difficult to adopt. Data visualization represents data in the way that simplifies data interpretation and its relationship.

3.1.1 Pie Chart

A pie chart is also called circle graph. Pie chart circle is divided into number of sectors, each circle describe a proportion in a whole quantity. The pie chart control is use to determines the size of data wedge as compare to other data wedges. In pie chart a wedge represents the part of data that has common feature or characteristics. Wedges can be labeled to identify different data points. Most of the time is shown in percentage.

.

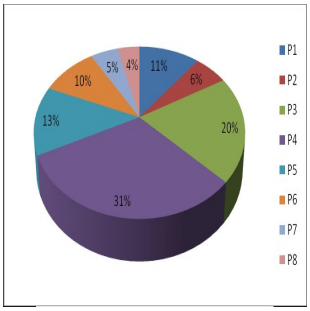

Figure 3.1: Standard Pie Chart

3.1.2 Bar Chart

One of the most commonly use data visualization method is bar chart, it is also called Bar Graph(also called Column chart). Bar chart is most of the time use for discrete data not for continuous data. Bar Chart control has been use to represent data in horizontal bars, the vertical length of the bar represent the values. Bar chart is use to represent a single data series and related data points are group in one series. For example monthly salaries, it can be mutli bar graph i.e. percentage increase per month as shown in the following figure.

3.1.3 Line Chart

Line chart is common well known graph in many fields, also label as line graph. It is a graph which is use to display information in connected points. These points are connected through continuous or straight line. Line graph is the extension of Scatter plot. Data points can be represented by icons or symbols, or can also draw simple line without icons

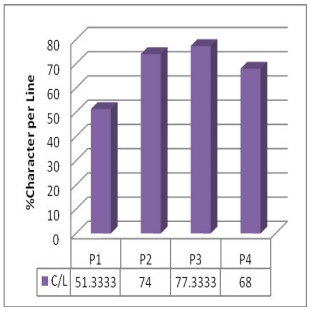

Figure 3.2: Bar Chart

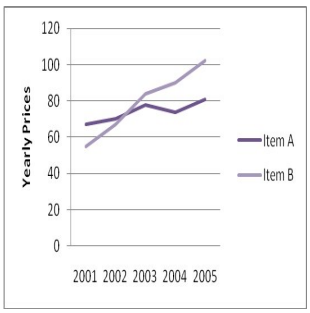

Figure 3.3: Line Chart

3.1.4 Area Chart

Area chart is also called area graph, use to display quantitative data graphically. Area chart control is use represent data in bounded area. The bounded area is based on the line graph, the line is generated and the area below is shaded with colors,different texture and hatching, which produce area graph.

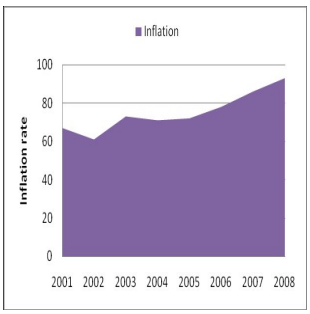

Figure 3.4: Area Chart

3.1.5 Scatter Plot

Scatter Plot is also known as plot, plot chart, scatter chart, scattergram, scatter diagram or scatter graph. Scatter plot is graphical display of set of data in Cartesian coordinate, shows the relationship between two variables, one variable represent horizontal distance (independent variable) and second variable vertical distance (dependent variable) of data point from the coordinate axis. Scatter plot shows the how strong the relationship are between the variables, and determines whether their exit any outlier in the data or not. It is use to look how the data is dispersed.

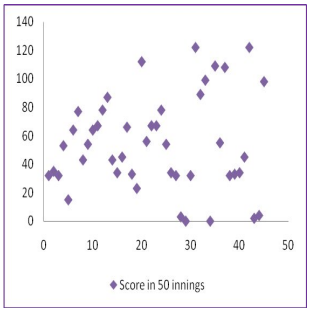

Figure 3.5: Scatter Plot Diagram

3.2 Interactive Visualization

The users are interested in the abstract data about which they have desire to understand, the user dont have sufficient pre knowledge about that data. Hence for the exploration, analysis, and for the representation of data or information visualization interactive techniques are exceptionally momentous. The challenge in information visualization is to provide data visually in order to the user effectively understand the information for which the user is looking for, for this purpose provide interaction mechanism that make is possible to manipulate visualization effectively and effortlessly as probable. Users can interact with interfaces or visualization in different ways by means of mouse over, single click, double click, or can add multiple interactive options by mouse right button click. There are many interactive techniques available to interact with charts or graphical representation to understand the drill down details. Card et al introduce the interactive mechanism of visualization in 1999.

Interactive visualization can be performed through approaches such as zooming

(zoom in and zoom out), overview and detail, zoom and pan, and focus and context or fish eye. The steps for interactive visualization are as follows:

3.2.1 Selecting

One of the most important and fundamental requirements in visualization is interactive selection of data entities or subset or part of whole data or whole data set, that is of interest to the user. This is useful to view detail information about the selected entities, highlight entities that are covered or hidden, cluster entities that are related and extracting entities that may be use in future.

3.2.2 Linking

Linking and brushing are the most common form of linking. Linking is useful to relate information among multiple views. Information can be mapped differently to different views to reveal or expose different perspectives (viewpoint) or different portions of the information. On the bases of selection criteria the user can select entities in one structure, which then shows the distribution the selected entities in another structure.

3.2.3 Filtering

Filtering enables users to dynamically adjust amount of information to display, means to decrease or increase information quantity that need to display, and focus on information of interest. For this purpose need some dynamic query values to manipulate, in this regard visual widgets can play vital role e.g. slider can be use in different ranges, field box can be use to put attribute value in specified range etc.These widgets enables one from the current query parameters and enables to quickly adjust query parameters and instantaneously view filtered results in the visualization in real time.

3.2.4 Rearranging or Remapping

As a single mapping or plotting visual form of information may be not enough, so the users must empower to customize mapping among many maps. To enable users to customize maps by its own choice provide the way to better understand the information.

As the spatial layout is the most important visual mapping, rearranging the spatial layout of the information is the most effective for producing different insights

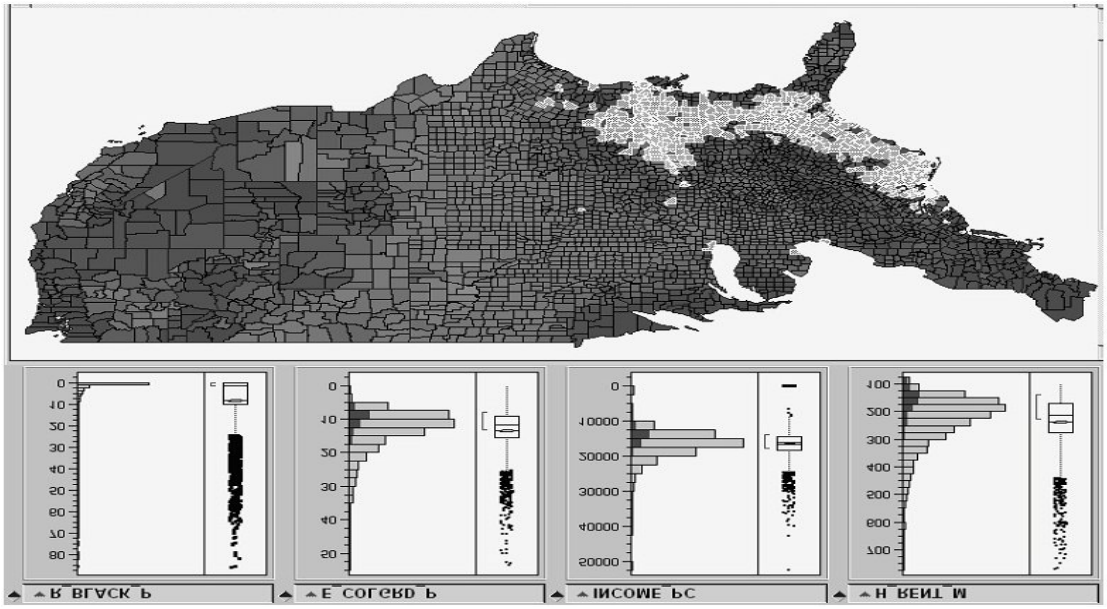

Figure 3.6: Interactive brushing and linking between histogram plots (top) and geographic map (bottom) of datasets

3.3 Data Visualization Tools

Following are the most awesome data visualization tools available on the web

- Dygraph
- ZingCharts
- InstantAtlas
- Timeline
- Exhibit
- Modest Maps
- Leaflet
- WolframAlpha
- Visual.ly
- VisualizeFree
- Better World Flux
- Fusion Charts
- jqPlot
- Dipity
- Many Eyes
- D3.js
- jpGraphs
- HighCharts
- Google Charts
- CrossFilter

Chapter 4

BIG DATA VISUALIZATION AND CHALLENGES

Visualization is an essential tool for making sense of big data.It provides a far richer view of big data than can be obtained from tables and statistics alone. However, the key to effective analysis of big data is the integration of visualization into analytics tools so that all kinds of users can interpret big data from a wide range of sources click streams, social media, log files, videos and more.

Online marketplace eBay, have hundreds of million active users and billions of goods sold each month, and they generate a lot of data. To make all that data understandable, eBay turned to Big Data visualization tool: Tableau, which has capability to transform large, complex data sets into intuitive pictures. The results are also interactive. Based on them, eBay employees can visualize search relevance and quality to monitor the latest customer feedback and conduct sentiment analysis.

4.1 Challenges in Big Data Visualization

For Big Data applications, it is particularly difficult to conduct data visualization because of the large size and high dimension of Big Data. However, current Big Data visualization tools mostly have poor performances in functionalities, scalability and response time. What we need to do is rethinking the way we visualize Big Data, not like the way we adopt before. For example, the history mechanisms for information visualization also are data-intensive and need more efficient approaches. Uncertainty can lead

to a great challenge to effective uncertainty-aware visualization and arise in any stage of a visual analytics process. New framework for modeling uncertainty and characterizing the evolution of the uncertainty information are highly necessary through analytical processes.

Major challenges in visual analytics are

4.1.1 Scalability

Scalability represents the scalability of visual representations.A visual analytics tool which can be used for lower dimensions of data may not be appropriate for visualizing higher dimensions of data.Since Bigdata applications are involving high dimensions of data we need tools that can operate on these data quickly and effectively.

Perceptual and interactive scalability are also challenges of big data visualization. Visualizing every data point can lead to over-plotting and may overwhelm users perceptual and cognitive capacities; reducing the data through sampling or filtering can elide interesting structures or outliers. Querying large data stores can result in high latency, disrupting fluent interaction

Scalability of the visualization tool should answer the following questions if it is considered scalable.

- How to run queries on distributed systems to explore big data sets?
- How to visualize a million multi-variate items on a screen?
- How to lower the time needed to run a clustering algorithm on xGbytes?
- How to design an interactive user interface loading big data in ¡ 1 sec?

4.1.2 Dynamics

Visualization tool should be answer the following questions if it is considered to be dynamic.

- How to aggregate data streams?
- How to visualize a continuously changing data structure?

- How to adapt clustering algorithms to consider dynamic data?
- How to design an interactive user interface continuously fed by data?

4.1.3 Need Of Massive Parallelization

Due to increasing data sizes and the emergence of the Big Data problem, the need for massive parallelization is a driving visualization research challenge. Supercomputing simulations regularly generate massive data sets with billions of data points per time step. Parallelization is an effective way of dealing with such data: there is more memory for storing data, there is more compute power for executing algorithms, and there is often more I/O bandwidth for reading data. The basic challenge for parallel visualization algorithms is to decompose the problem into independent tasks that can be run concurrently on all of the processing elements (i.e., the instances of the program), thus avoiding idle time.

Data parallelism is the dominant technique; data sets are decomposed into pieces and the pieces are partitioned over the processing elements. This approach has been shown to be highly scalable with results for hundreds of thousands of processing elements in research prototypes and tens of thousands of processing elements in production software .The role of visualization software, with respect to parallelization, is to provide a framework that shields algorithm developers from complexity.

4.1.4 Application Architecture and Data Management

Application architecture refers to the system design of visualization software. Data management for visualization must provide visualization techniques that integrate into the data life cycle. Although these two topics are distinct, they are treated together here, since emerging data management needs will drive application architecture. Traditionally, data management has not been a pressing concern for visualization software. Data, whether observed or simulated, was stored in the file system for processing;visualization software simply read whatever data it needed from files whenever it needed it. However, increases in data size, observed and simulated, as well as diversity of data sources, mandate new approaches in data management.

Application architectures exist to solve the simplest use model: have data, want a

picture," where the application architecture serves as a black box that consumes data and produces imagery with user-selected methods and parameters.Twenty-five years ago, the architecture for most visualization applications was a single binary that read from the local file system and produced images using local graphics. A little over a decade ago, scientific visualization applications for large data shifted to a client-server design where data was processed by a remote parallel server, producing geometry that was rendered by a local client. Today, application architectures for visualization frequently involve web clients and remote data access. In short, application architectures evolve to meet evolving data management needs.

4.2 Potential Solutions

The potential solutions of the common problems in big data visualization is presented in following section.

4.2.1 Meeting the Need for Speed

Meeting the need for speed In todays hyper-competitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed. The challenge only grows as the degree of granularity increases. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly. Another method is putting data in-memory but using a grid computing approach, where many machines are used to solve a problem.Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time.

4.2.2 Dealing with Outliers

The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text.Outliers typically represent about 1 to 5 percent of data, but when youre working with massive amounts of data, viewing 1 to 5 percent of the data is rather difficult. How do you

represent those points without getting into plotting issues? Possible solutions are to remove the outliers from the data (and therefore from the chart) or to create a separate chart for the outliers. You can also bin the results to both view the distribution of data and see the outliers. While outliers may not be representative of the data, they may also reveal previously unseen and potentially valuable insights.

4.2.3 Displaying meaningful results

Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information.For example, if we have 10 billion rows of retail SKU data that were trying to compare. The user trying to view 10 billion plots on the screen will have a hard time seeing so many data points. One way to resolve this is to cluster data into a higher-level view where smaller groups of data become visible. By grouping the data together, or binning, data can be more effectively visualized.

Chapter 5

PROGRESS OF BIGDATA VISUALIZATION

The directions that research is taking for visualizing big data are a mix: both applying existing principles such as details on demand at higher levels of scale (for example, providing more levels of drill-down between the overview and the lowest level of granularity) and also coming up with new and specialized visualizations that allow larger quantities of data to be represented intelligibly. And as such tools become more sophisticated and more mainstream.

5.1 Big Data Visualization Methodology

Visualizations should be designed in the era of big data in a way such that visualization tools should provide an overview first, then allow zooming and filtering, and provide deeper details on demand.

5.1.1 Overview First

The most important part of a dashboard is the overview section. Its the first thing a viewer sees in the dashboard, and guides the him or her to other parts of the product for further exploration.

The overview should summarize the overarching story from the entire data set without getting into the minor details. It shouldnt overload the user with too much data, which is where interactive charts, gauges, and maps serve to reduce data clutter, and

bring out the story more powerfully

At the same time it shouldnt leave out important parts of the story by using just a single pie chart, and hiding all the data a layer deeper. Often, great dashboards use a combination of chart types like the line chart, bar chart, maps, and gauges to give the viewer variety, and clarity when studying the data. The overview section should be carefully planned to highlight the important parts of the story, and give lesser weight to the not-so-critical parts. To do this, you may want to organize the entire section into many sub-sections that are clearly labeled. Of course, the important sections would be placed more prominently than the others.

5.1.2 Zoom and Filter

Since all the data is presented to the user in the overview section, the viewer will want to focus on particular areas of interest. This involves zooming and filtering the data using the dashboards interactive features: zooming, scrolling, panning, drill-down, legend, range selector, etc. For example, zooming may be drilling down from global to country-specific data while filtering may be excluding information in a specific time range.

From a design perspective, designer should aim to provide the user with plenty of control for zooming and filtering data from the overview. This will yield maximum insights and action from the information at hand.

5.1.3 Details-on-Demand

Designer have identified areas of interest from the overview section, and have dug deeper into the data using zooming and filtering, but user still may not have found what he started looking for.

A dashboard that excels at giving an overview, and allows extensive zooming and filtering, should go all the way and give the viewer access to the minute details. This would bring them as close as possible to the raw data, and equip them to find what they started looking for. This third layer of data would be less visual, and more text-heavy with a focus on accurate information rather than trends. This way the analyst gets what he or she needs, in a way that drives action.

5.2 Big Data Visualization Approach

Visualization can play an important role in using big data to get a complete view of customers. Relationships are an important aspect of many big data scenarios. Social networks are perhaps the most prominent example and are very difficult to understand in text or tabular format; however, visualization can make emerging network trends and patterns apparent. A cloud-based visualization method was proposed to visualize an inherent relationship of users on social network. The method can intentionally present the users social relationship based on the correlation matrix to represent a hierarchical relationship of user nodes of social network. In addition, the method uses Hadoop based on cloud for the distributed parallel processing of visualization, which helps expedite the big data of social network

Big data visualization can be performed through a number of approaches such as more than one view per representation display, dynamical changes in number of factors, and filtering (dynamic query filters, star-field display, and tight coupling).

5.2.1 TreeMap

It is based on space-filling visualization of hierarchical data.

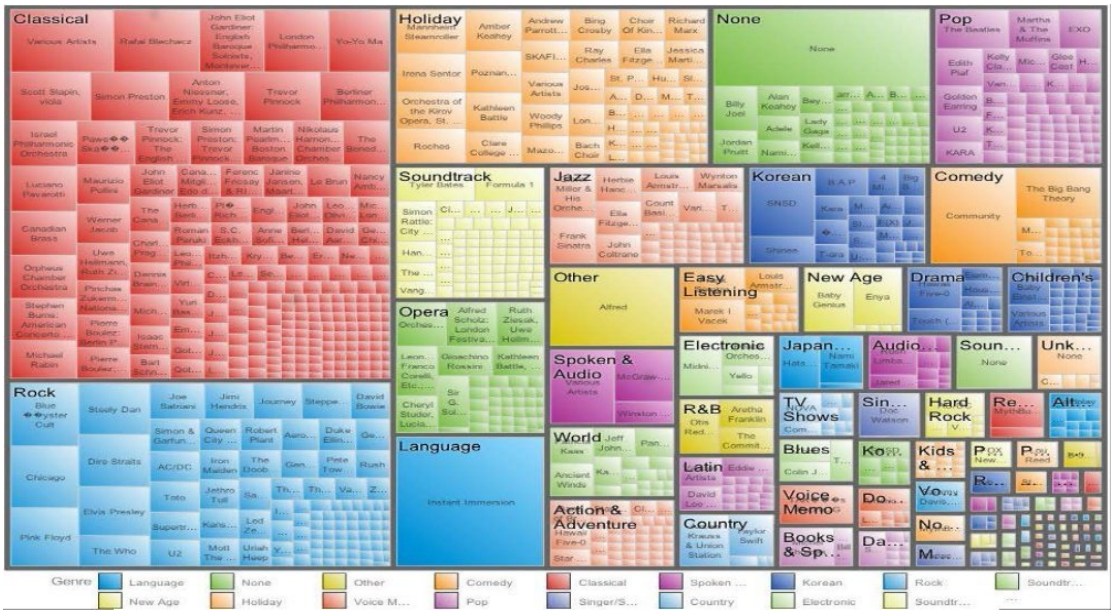

Figure 5.1: Treemap

5.2.2 Parallel Coordinates

It allows visual analysis to be extended with multiple data factors for different objects.

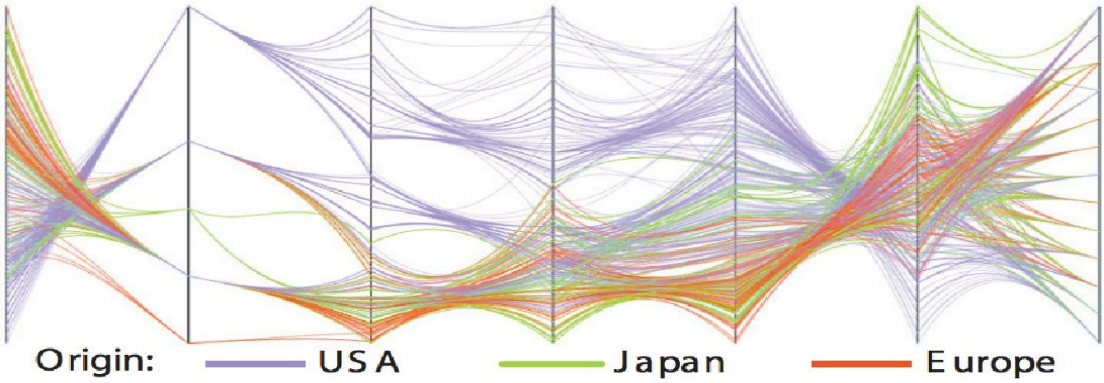

Figure 5.2: Parallel Coordinates

5.2.3 Semantic Network

A semantic network or net is a graph structure for representing knowledge in patterns of interconnected nodes and arcs.

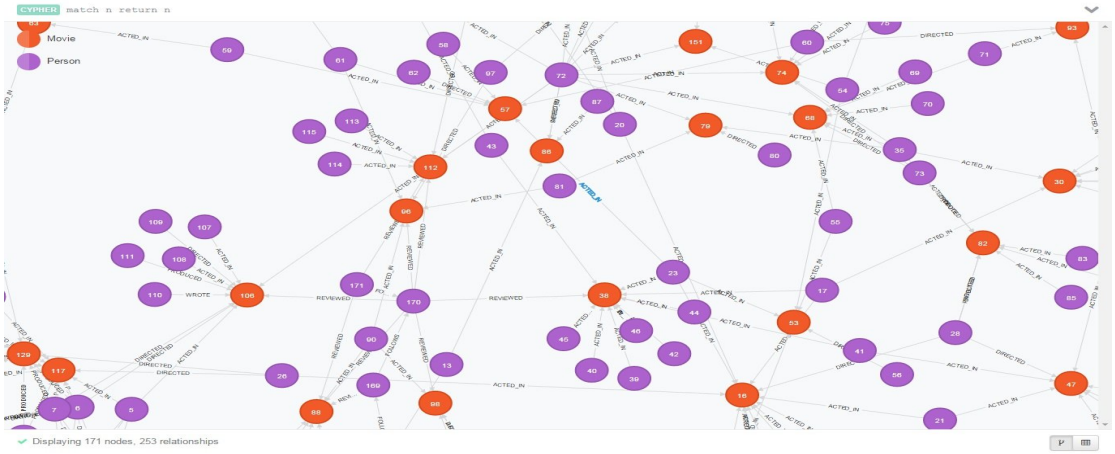

Figure 5.3: Semantic Network

5.2.4 Sunburst

It uses tree-map visualization and is converted to polar coordinate system. The main difference is that the variable parameters are not width and height, but a radius and arc length.

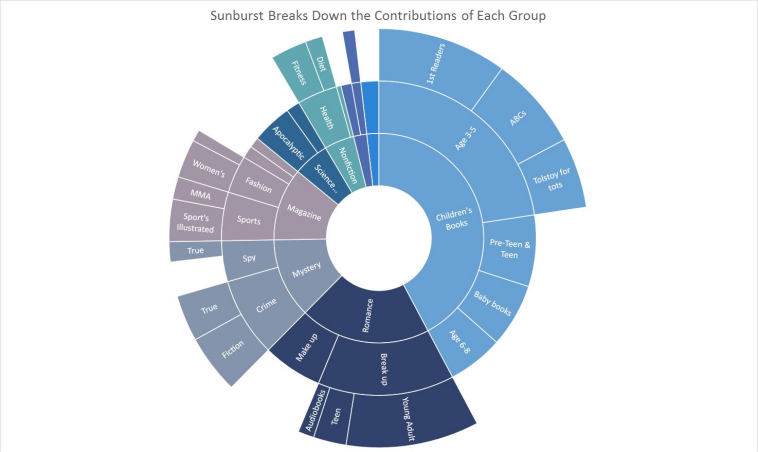

Figure 5.4: SunBurst Visualization

5.3 Properties Of Visualization Tools

Traditional data visualization tools are often inadequate to handle big data. Methods for interactive visualization of big data were presented. First, a design space of scalable visual summaries that use data reduction approaches (such as binned aggregation or sampling) was described to visualize a variety of data types. Methods were then developed for interactive querying (e.g., brushing and linking) among binned plots through a combination of multivariate data tiles and parallel query processing. The developed methods were implemented in imMens, a browser-based visual analysis system that uses WebGL for data processing and rendering on the GPU

Big data processing tools can process ZB (zettabytes) and PB (petabytes) data quite naturally, but they often cannot visualize ZB and PB data. At present, big data processing tools include Hadoop, High Performance Computing and Communications, Storm, Apache Drill, RapidMiner, and Pentaho BI. Data visualization tools include NodeBox, R, Weka, Gephi, Google Chart API, Flot, D3, and Visual.ly, etc. A big data visualization algorithm analysis integrated model based on RHadoop was proposed. The integrated model can process ZB and PB data and show valuable results via visualization. The model is suitable for the design of parallel algorithms for ZB and PB data.

Method name	Large data volume	Data variety	Data dynamics
Treemap	+	-	-
Sunburst	+	-	+
Parallel coordinates	+	+	+
Circular network	+	+	-
Circle packing	+	-	-

Table 5.1: Properties of visualization methods

5.4 Virtual Reality Platform for Scientific Data Visualization

The use of immersive virtual reality (VR) platforms for scientific data visualization has been in the process of exploration including software and inexpensive commodity hardware. These potentially powerful and innovative tools for multi-dimensional data visualization can provide an easy path to collaborative data visualization. Immersion provides benefits beyond traditional desktop visualization tools: it results in a better perception of data scape geometry and more intuitive data understanding.

Immersive visualization should become one of the foundations to explore the higher dimensionality and abstraction that are attendant with big data. The intrinsic human pattern recognition (or visual discovery) skills should be maximized through using emerging technologies associated with the immersive VR.

5.5 SWOT Analysis Of Current Tools

The SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis is a well-known method to ensure that both positive factors and negative factors are identified. A SWOT analysis of the above software tools for big data visualization has been conducted and is shown in table 5.2. In table 5.2, Strengths and Opportunities are positive factors; Weaknesses and Threats are negative factors.

Strengths	Opportunities
With the functions of visualization and interaction for visualizing data.	Immersive visualization with virtual reality (VR) result in a better perception of data scape geometry and more intuitive data understanding.
Able to visualize a variety of data types	The intrinsic human pattern recognition (or visual discovery) skills could be maximized.
Weaknesses	**Threats**
There is room to improve in visualizing big data with high velocity or the combination of three high Vs (Volume + Velocity + Variety).	Lack adequate visualization in a lot of Big Data applications.

Table 5.2: The SWOT analysis of current big data visualization software tools

Chapter 6

CONCLUSION

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. Web-based visualization helps get dynamic data timely and keep visualizations up to date.

The extension of some conventional visualization approaches to handling big data is far from enough in functions. More new methods and tools of Big Data visualization should be developed for different Big Data applications. Advances of Big Data visualization are presented and a SWOT analysis of current visualization software tools for big data visualization has been conducted in this paper. 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.

Bibliography

[1] Datafloq.com, 'Datafloq - The One-Stop Shop for Big Data', 2015. [Online]. Available: https://datafloq.com/read/3vs-sufficient-describe-big-data/166. [Accessed: 27- Oct- 2015].

[2] Mckinsey.com, 'Big data: The next frontier for innovation, competition, and productivity', 2015. [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. [Accessed: 27- Oct- 2015].

[3] Stamford, Gartner Says Solving Big Data Challenge Involcves More Than Just Managing Volumes of Data,27,2011,posted on June http://www.gartner.com/newsroom/id/1731 916

[4] E.Y. Gorodov and V.V. Gubarev, Analytical Review of Data Visualization Methods in Application to Big Data, Journal of Electrical and Computer Engineering, 013, Article ID 969458, pp.1-7.

[5] 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

[6] Intel IT Center, Big Data Visualization: Turning Big Data Into Big Insights, White Paper, March 2013, pp.1-14.

[7] V. Sucharitha, S.R. Subash and P. Prakash , Visualization of Big Data: Its Tools and Challenges, International Journal of Applied Engineering Research, 9(18), 2014, pp. 5277-5290.