Data Visualization Digital Assignment 2

Name: Om Ashish Mishra

Registration Number: 16BCE0789

Slot: C2

# The Question:

Discuss about the challenges involved in building visualization tools for Big data sources like Hadoop, Streaming and NoSQL.

# The Answer:

# **The Problems With Visualization**

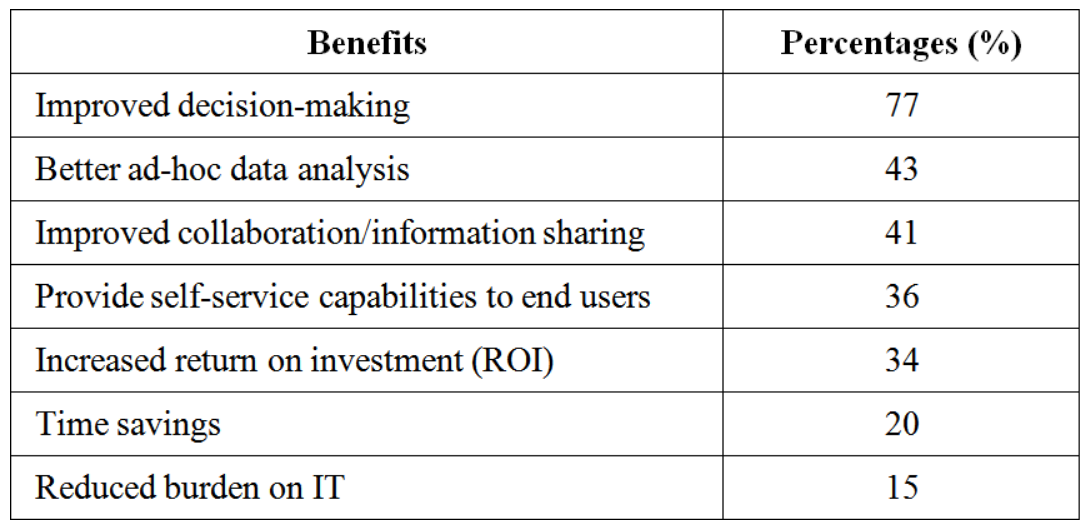
Unfortunately, there are a few current and forthcoming problems with the concept of data visualization:

1. **The oversimplification of data.** One of the biggest draws of visualization is its ability to take big swaths of data and simplify them to more basic, understandable terms. However, it’s easy to go too far with this; trying to take millions of data points and confine their conclusions to a handful of pictoral representations could lead to unfounded conclusions, or completely neglect certain significant modifiers that could completely change the assumptions you walk away with. As an example not relegated to the world of data, consider basic real-world tests, such as alcohol intoxication tests, which try to reduce complex systems to simple “yes” or “no” results—as Monder Law Group points out, these tests can be unreliable and flat-out inaccurate.
2. **The human limitations of algorithms.** This is the biggest potential problem, and also the most complicated. Any algorithm used to reduce data to visual illustrations is based on human inputs, and human inputs can be fundamentally flawed. For example, a human developing an algorithm may highlight different pieces of data that are “most” important to consider, and throw out other pieces entirely; this doesn’t account for all companies or all situations, especially if there are data outliers or unique situations that demand an alternative approach. The problem is compounded by the fact that most data visualization systems are rolled out on a national scale; they evolve to become one-size-fits-all algorithms, and fail to address the specific needs of individuals.
3. **Overreliance on visuals.** This is more of a problem with consumers than it is with developers, but it undermines the potential impact of visualization in general. When users start relying on visuals to interpret data, which they can use at-a-glance, they could easily start over-relying on this mode of input. For example, they may take their conclusions as absolute truth, never digging deeper into the data sets responsible for producing those visuals. The general conclusions you draw from this may be generally applicable, but they won’t tell you everything about your audiences or campaigns.
4. **The inevitability of visualization.** Already, there are dozens of tools available to help us understand complex data sets with visual diagrams, charts, and illustrations, and data visualization is too popular to ever go away. We’re on a fast course to visualization taking over in multiple areas, and there’s no real going back at this point. To some, this may not seem like a problem, but consider some of the effects—companies racing to develop visualization products, and consumers only seeking products that offer visualization. These effects may feed into user overreliance on visuals, and compound the limitations of human errors in algorithm development (since companies will want to go to market as soon as possible).

# **Big Data**

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. The challenges of Big Data visualization are discussed. New methods, applications, and technology progress of Big Data visualization are presented.

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 smartphones. The table below shows the benefits of data visualization according to the respondent percentages of a survey.



There are some points of advice for visualization: (1) Do not forget the metadata.Data about data can be very revealing. (2) Participation matters.Visualization tools should be interactive, and user engagement is very important. (3) Encourage interactivity.Static data tools don’t lead to discovery as well as interactive tools do.

Big data are high volume, high velocity, and/or high variety datasets that require new forms of processing to enable enhanced process optimization, insight discovery and decision making. Challenges of Big Data lie in data capture, storage, analysis, sharing, searching, and visualization. Visualization can be thought of as the “front end” of big data. There are following data visualization myths:

• All data must be visualized:It is important not to overly rely on visualization; some data does not need visualization methods to uncover its messages.

• Only good data should be visualized:A simple and quick visualization can highlight something wrong with data just as it helps uncover interesting trends.

• Visualization will always manifest the right decision or action:Visualization cannot replace critical thinking.

• Visualization will lead to certainty:Data is visualized doesn’t mean it shows an accurate picture of what is important. Visualization can be manipulated with different effects.

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

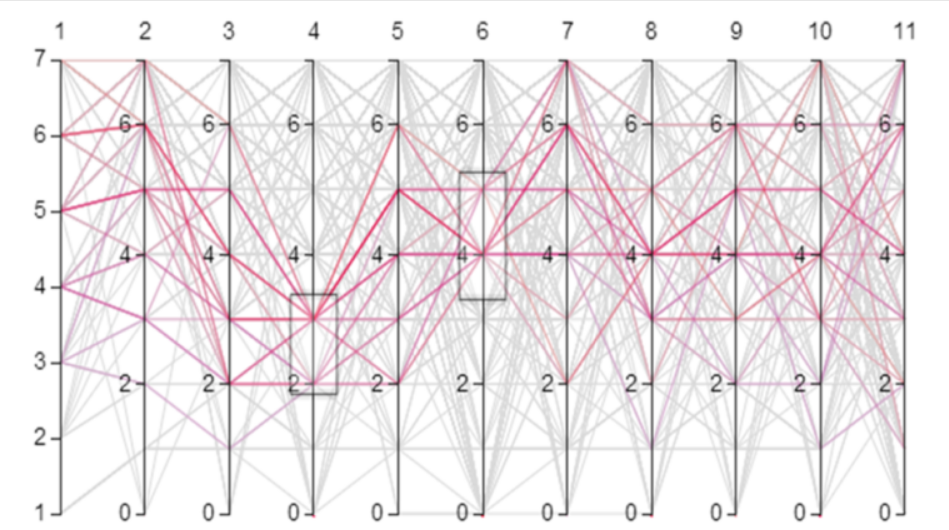
The goal and the objectives of this paper are to present new methods and advances of Big Data visualization through introducing conventional visualization methods and the extension of some them to handling big data, discussing the challenges of big data visualization, and analysing technology progress in big data visualization.

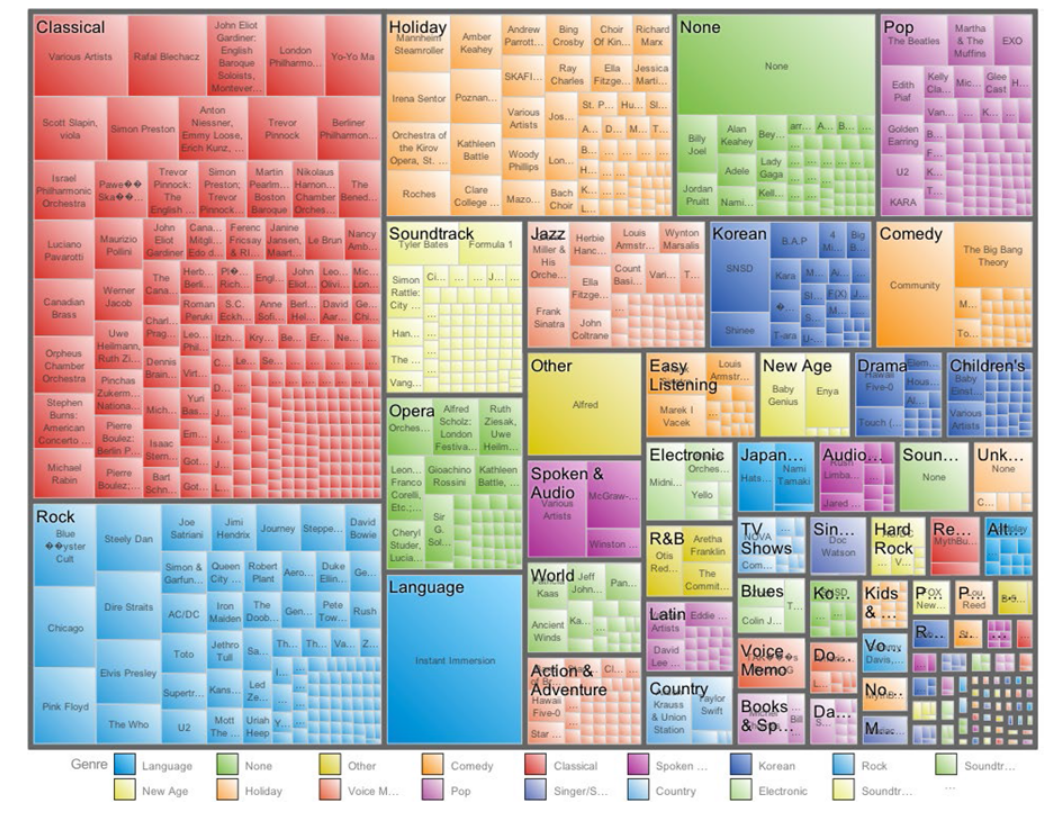
In this study, authors first searched for papers that are related to data visualization and were published in recent years through the university library system. At this stage, authors mainly summarized traditional data visualization methods and new progress in this area. Next, authors searched for papers that are related to big data visualization. Most of these papers were published in the past three years because big data is a newer area. At this stage, authors found that most conventional data visualization methods do not apply to big data. The extension of some conventional visualization approaches to handling big data is far from enough in functions. The authors focused on big data visualization challenges as well as new methods, technology progress, and developed tools for big data visualization.

# 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, multiple data series or combination of charts, 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.

Parallel coordinates is used to plot individual data elements across many dimensions. Parallel coordinate is very useful when to display multidimensional data. The first Figure shows parallel coordinates. Treemap is an effective method for visualizing hierarchies. The size of each sub-rectangle represents one measure, while colour is often used to represent another measure of data. The Second Figure shows a treemap of a collection of choices for streaming music and video tracks in a social network community. Cone tree is another method displaying hierarchical data such as organizational body in three dimensions. The branches grow in the form of cone. A semantic network is a graphical representation of logical relationship between different concepts. It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge.





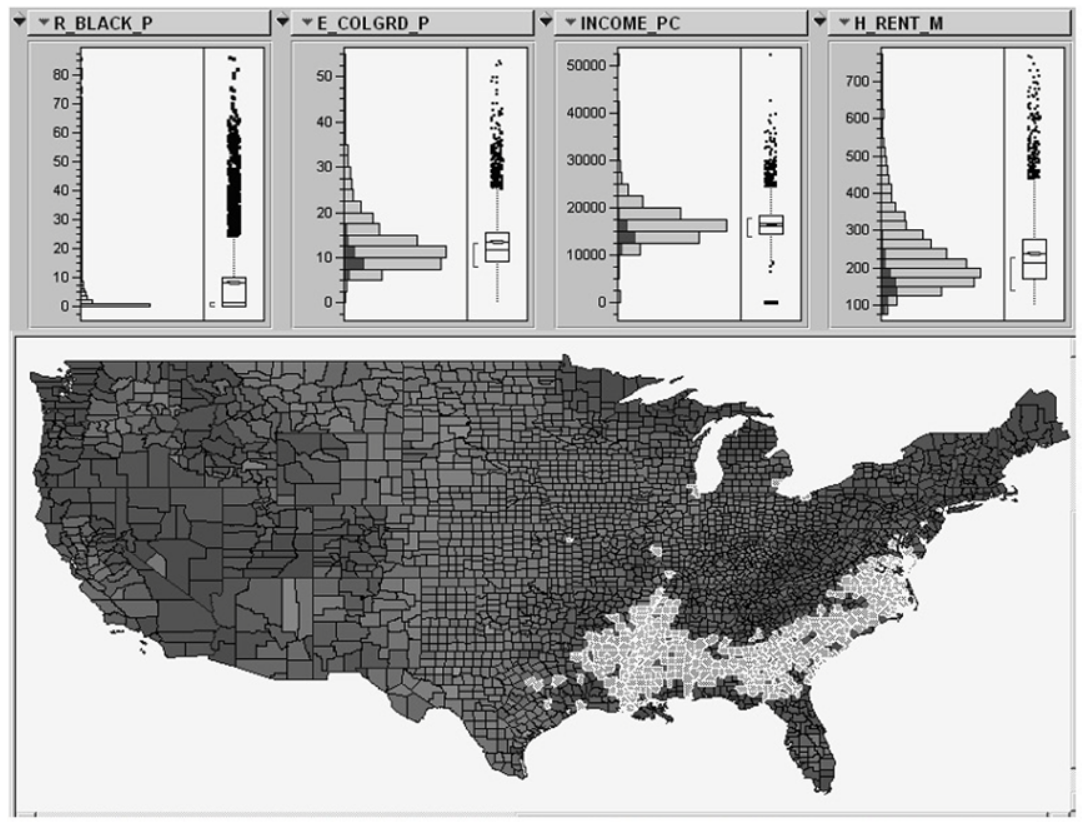
Visualizations are not only static; they can be interactive. Interactive visualization can be performed through approaches such as zooming (zoom in and zoom out), overview and detail, zoom and pan, and focus and context or fish eye. The steps for interactive visualization are as follows:

1. *Selecting:*Interactive selection of data entities or subset or part of whole data or whole data set according to the user interest.

2. *Linking:*It is useful for relating information among multiple views. An example is shown in the figure below.

3. *Filtering:*It helps users adjust the amount of information for display. It decreases information quantity and focuses on information of interest.

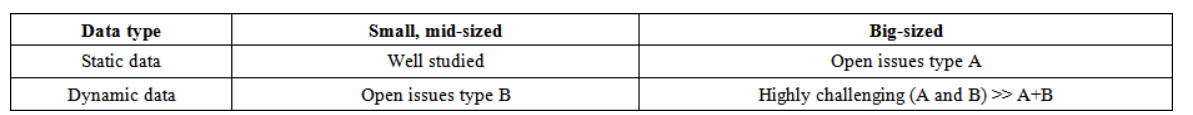
4. *Rearranging or Remapping:* Because the spatial layout is the most important visual mapping, rearranging the spatial layout of the information is very effective in producing different insights.



New database technologies and promising Web-based visualization approaches may be vital for reducing the cost of visualization generation and allowing it to help improve the scientific process. Because of Web-based linking technologies, visualizations change as data change, which greatly reduces the effort to keep the visualizations timely and up to date. These “low-end” visualizations have been often used in business analytics and open government data systems, but they have generally not been used in the scientific process. Many visualization tools that are available to scientists do not allow live linking as do these Web-based tools.

# Challenges of Big Data Visualization:

Scalability and dynamics are two major challenges in visual analytics. The table given below shows the research status for static data and dynamic data according to the data size. For big dynamic data, solutions for type A problems or type B problems often do not work for A and B problems.



The visualization-based methods take the challenges presented by the “four Vs” of big data and turn them into following opportunities .

• *Volume*: The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.

• *Variety*: The methods are developed to combine as many data sources as needed.

• *Velocity*: With the methods, businesses can replace batch processing with real-time stream processing.

• *Value*: The methods not only enable users to create attractive infographics and heatmaps, but also create business value by gaining insights from big data.

Visualization of big data with diversity and heterogeneity (structured, semi-structured, and unstructured) is a big problem. Speed is the desired factor for the big data analysis. Designing a new visualization tool with efficient indexing is not easy in big data. Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability.

Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata. Big data often has unstructured formats. Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently. Visualization software should be run in an in situ manner. Because of the big data size, the need for massive parallelization is a challenge in visualization. The challenge in parallel visualization algorithms is decomposing a problem into independent tasks that can be run concurrently.

Effective data visualization is a key part of the discovery process in the era of big data. For the challenges of high complexity and high dimensionality in big data, there are different dimensionality reduction methods. However, they may not always be applicable. The more dimensions are visualized effectively, the higher are the chances of recognizing potentially interesting patterns, correlations, or outliers.

There are also following problems for big data visualization:

• *Visual noise:*Most of the objects in dataset are too relative to each other. Users cannot divide them as separate objects on the screen.

• *Information loss:*Reduction of visible data sets can be used, but leads to information loss.

• *Large image perception:*Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.

• *High rate of image change: Users*observe data and cannot react to the number of data change or its intensity on display.

• *High performance requirements:*It can be hardly noticed in static visualization because of lower visualization speed requirements--high performance requirement.

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.

In Big Data applications, it is difficult to conduct data visualization because of the large size and high dimension of big data. Most of current Big Data visualization tools have poor performances in scalability, functionalities, and response time. Uncertainty can result in a great challenge to effective uncertainty-aware visualization and arise during a visual analytics process.

Potential solutions to some challenges or problems about visualization and big data were presented:

1. Meeting the need for speed: One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.

2. Understanding the data: One solution is to have the proper domain expertise in place.

3. Addressing data quality: It is necessary to ensure the data is clean through the process of data governance or information management.

4. Displaying meaningful results: One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized.

5. Dealing with outliers: Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

# Problem with Hadoop:

### **1: You never get to production**

Moving from proof of concept (POC) to production is a significant step for big data workloads. Scaling Hadoop jobs is fraught with challenges. Sometimes large jobs just won’t finish. A job that ran in testing won’t run at production scale. Data can also be an issue: the POC often uses unrealistically small or uniform datasets.

Before you go into production, perform realistic scale and stress testing. Such testing will exercise the scalability and fault-tolerance of your applications. It will also help you develop a model for capacity planning so you can stay ahead of the curve.

### **2: You start missing deadlines**

Your first application made it into production. Congratulations! Initially, you easily hit your SLAs, but as use of the Hadoop cluster grows, the run times become unpredictable. At first deadlines are missed sporadically, so the problem is ignored. Over time it gets worse, until a crisis emerges.

Don’t wait for a crisis to take action. As comfortable margins start to erode, add capacity or optimize your applications to keep pace. Adjust your capacity-planning model, with particular attention on worst-case performance, so that it matches what you’re seeing in reality.

### **3: You start telling people they can’t keep all that data**

Another symptom of impending crisis is shrinking data retention windows. Initially, you hoped to keep 13 months of data for year-over-year analysis. Because of space constraints, you find yourself cutting that number. At some point, you lose the ability to do the type of big data analysis that justified your Hadoop investment to begin with.

Shrinking retention windows are the storage equivalent of missed deadlines. The dynamic is also the same: a margin that initially seems comfortable becomes “just enough” and then “not enough.” Act early. As margins erode, revisit your capacity models to see why your predictions didn’t hold, and adjust to better track what’s happening.

### **4: Your data scientists are starved**

An over-utilized Hadoop cluster can stifle innovation. There’s not enough compute capacity for data scientists to launch large jobs. There’s insufficient space for them to store large, intermediate results.

Capacity planning routinely omits or underestimates the needs of data scientists. That omission, compounded with inadequate planning for production work, means the needs of data scientists often become marginalized. Be sure your planning includes data scientists’ requirements, and act early when you see signs that capacity is falling short.

### **5: Your data scientists are reading Stack Overflow**

In the early days of your Hadoop implementation, your ops team and data scientists worked hand in hand. If the data scientists ran into problems, the ops team would jump in to help. But as your Hadoop implementation became successful, the stresses of maintaining and growing it consumed your operations team. Your data scientists now troubleshoot Hadoop themselves, often by trawling through questions posted to [company]Stack Overflow[/company].

As Hadoop expands and becomes more mission critical, the effort to maintain it increases. If you want to keep your data scientists focused on data science (and off of Stack Overflow), you may have to revisit the size of your Hadoop operations team.

### **6: It starts getting really, really hot**

Your hair might not be on fire, but your data center could be! When servers are provisioned for power, there’s often an assumption they won’t run at capacity. But a large Hadoop job can red-line racks of machines for hours, blowing under-provisioned circuits. (Similar problems arise on the cooling side.) Make sure your Hadoop cluster can run at full power for extended periods of time.

### **7: You get sticker shock**

The number one “success disaster” with Infrastructure-as-a-Service based deployments of Hadoop (such as AWS) is out-of-control spending. You suddenly find yourself with a bill that is three times last month’s, blowing your budget.

Capacity planning is as important for IaaS-based Hadoop implementations as it is for on-premise ones—not for managing capacity, but for managing costs. But good capacity planning is just the start. If you plan on growing an IaaS-based Hadoop implementation, expect to invest heavily in systems to track and optimize costs, as Netflix has done.

## **Smooth(er) Hadoop scaling**

Hadoop plans typically underestimate the effort required to keep a Hadoop cluster running smoothly. It’s an understandable miscalculation. With classic enterprise applications, the initial implementation effort is orders of magnitude larger than ongoing maintenance and support. People assume Hadoop follows a similar pattern, but it doesn’t. Hadoop gets harder to maintain as it scales, and it requires a lot of work from your ops team.

Good capacity planning is essential to promote sanity. That means not only having a good capacity model, but updating it as it starts to diverge from reality (and it will). Don’t support innovation as an afterthought: provide data scientists with a guaranteed level of support. Adding capacity is not always the answer: managing usage is equally important. Get your users (and the business owners driving them) to plan time to optimize their jobs between bursts of new development. Just a bit of optimization can significantly reduce your ongoing costs.

# Problems with Streaming

The amount of data that we can transfer from SAMI to the browser depends mostly on the number of devices that the user has connected to his/her account, and the ones that are currently sending data to the platform. With this in mind, and having Simband as a first device, we knew that we could get a really high volume of data that needed to be transmitted in short periods of time—so short that making AJAX call to poll the server wasn’t an option at all. The AJAX call would’ve taken longer than the time we’d have to process all the data received. The next possible solution was obvious, given the current technology that exists: WebSockets.

# Problems with NoSQL

Data is growing very rapidly and becoming more complex in variety, velocity, and volume. The notion of big data is associated with opportunities as well as challenges for existing computing techniques. Traditional data management tools and techniques are typically designed for structured data management and not sufficiently process and analyze large data volumes. NoSQL provides set of storage alternatives with the various characteristics that are intended to manage and process big data. There are different types of NoSQL stores such as Key-Value, Document-Oriented, Column-Oriented and GraphOriented. It is necessary to set evaluation criteria for these databases. This work presents evaluation criteria for existing NoSQL stores cope with the current data challenges.

# 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.