ISSN 2320 - 2602
Volume 9 No.7, July 2020
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst16972020.pdf
https://doi.org/10.30534/ijacst/2020/16972020

A SURVEY ON COMPUTATIONAL PARALLELISM AND VISUALIZATION SERVICES OF LARGE DATA SETS

Jesmi Latheef[1], Vineetha S[2],
[1]Computer Science and Engineering RIT, Kottayam, India, jesmilatheef@gmail.com
[2] Computer Science and Engineering RIT, Kottayam, India, svineetha@rit.ac.in

ABSTRACT
Visualizing big data is a computationally intensive and important task that most users undertake on an investigative basis. The need for scalable computational power coupled with varied usage patterns makes it a strong candidate for cloud computing. The existing interactive visualization services are based on Web technologies and client-side rendering. However, these services can't make optimal use of the cloud because they rely on the client which can render any amount of visualization data that the server needs to send them and having resources to store. These demands will always be huge for big data problems and thus will swamp the client's limited memory and compute power. Visualization of large data sets in cloud is an active research topic over many years and several techniques and aspects of such system has been proposed. This describes solutions for large scale data visualization. In this paper, we have done a detailed study on big data parallel computing and Visualization in various aspects.

Key words: Big Data, Computational Parallelism, Visualization

1. INTRODUCTION

Big Data visualization allows understandable and intelligent visual representation of data patterns, thus users can more easily gather insights from data. This involves the presentation of data of almost any type in a graphical format that makes it easy to understand and interpret. But it goes far beyond from typical corporate graphs, histograms and pie charts to more complex representations like heat maps and fever charts, enabling decision makers to explore data sets to identify correlations or unexpected patterns. To provide an interesting and interactive feel to user, using multiple graphics devices and software with huge dataset, we made a study for comparing alternative architectures and techniques.

Visualization plays an important part of data analytics and helps interpret big data in a real-time structure by utilizing complex sets of numerical or factual figures. Due to the way the human brain processes information, presenting insights in charts or graphs to visualize significant amounts of complex data is more accessible than relying on spreadsheets or reports. The most effective data visualization methods on our list; to succeed in presenting your data effectively, you must select the right charts for your specific project, audience, and purpose. Together with the demand for data visualization and analysis, the tools and solutions in this area develop fast and extensively. Novel 3D visualizations and shared VR offices are getting common alongside traditional web and desktop interfaces. As the age of Big Data kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers. A good visualization tells a story, removing the noise from data and highlighting the useful information.

2. THE IMPACT OF VISUALIZATION

Main objective of visualization is to enable users to attain an understanding of key features within a set of information and thus facilitates decision making and resulting actions. As Figure 1, shows how visualization putting ideas into people's head. Each visualization produced is based on some large data set and it have one or more of the following aims.

- For activities such as learning and entertainment, describing and explaining ideas.
- Maintain regular lookup over, like IoT information streams.
- Inquire into or discuss in detail and examine information sets, such as public service data.
- Creating and testing new hypotheses, such as climate simulations.

The process of changing information gathered into understandable knowledge is both faster and easier by providing accessible visualizations to different types of users on different types of platforms and these are the challenges for visualization designers. Which is the biggest potential problem, and also the most complicated one. Any algorithm used to imply data to visual illustrations is based on human inputs, and human inputs can be fundamentally flawed, The problem with consumers is, 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.

Figure 1: Visualization conveys information to user

3. APPLICATIONS OF VISUALIZATIONS

The various applications of visualizations are discussed below:

3.1 Crisis Management

Perhaps the greatest value of real-time visualization in handling risk comes from informing decision makers who need to respond to emergent events. If a storm is on track to destroy a data center, retail outlet, or any part of a firms infrastructure or supply chain, for example, real-time visualization can be tremendously helpful. Conversely, real-time visualization of assets in a variety of geographic locations allows decision makers to allocate resources where they're needed most, which can be the difference between keeping and losing customers in industries where uptime is critical.

3.2 Sales

Real-time data visualization opens up great opportunities for firms attempting to make more sales, both in brick-and-mortar institutions and in e-commerce. Real-time analytics give firms the option to provide customers with contextual suggestions for example, a supermarket suggesting a recipe using mostly ingredients already in a customer's cart. Combine this with more efficient inventory management (restocking hot items more quickly when they sell out), and real-time visualization gives firms a tremendous amount of flexibility to get more products out to consumers.

3.3 Healthcare

The level of data generated within healthcare systems is not trivial. Traditionally, the health care industry lagged in using Big Data, because of limited ability to standardize and consolidate data. But now Big data analytics have improved healthcare by providing personalized medicine and prescriptive analytics. Researchers are mining the data to see what treatments are more effective for particular conditions, identify patterns related to drug side effects, and gains other important information that can help patients and reduce costs. With the added adoption of mHealth, eHealth and wearable technologies the volume of data is increasing at an exponential rate. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of data.

3.4 Manufacturing

Predictive manufacturing provides near-zero downtime and transparency. It requires an enormous amount of data and advanced prediction tools for a systematic process of data into useful information.

3.5 Traffic control

With the use of visualization method, a large number of traffic data can form visual traffic information with the human computer interaction, accuracy, reliable and high efficiency. Thus can understand its internal law through the transformation of graphics and image.

4. VARIOUS ASPECTS OF LARGE DATASET PROCESSING AND VISUALIZATION

Several studies have been made on large data processing and visualizations. Based on these studies, Processing and Visualization techniques are discussed:

4.1 Different Concepts on Computational parallelism

A. MapReduce Model

This is an entirely web based solution in which the end user does not want to download and install a visualization tool on his computer [1]. The data is indexed by the data server and stored in a database during the preprocessing step, and also as much information as possible about the data is collected and derived. Through an instinctive interface user can then modify the visualization parameters. As parameters changes, the rendering servers in the cloud are receiving a new rendering request. To retrieve the original parameters of a visualization they apply some algorithms and genetic algorithms are mostly used for this task. The system automatically computes and generates a new set of visualizations in the cloud that the user may prefer as the user selects a visualization from the clustered results. Figure 2, shows overall architecture of the system. The computational resources of the cloud are responsible to compute the set of suggestive visualizations. The rendering cloud utilizes MapReduce model to take full advantage of the distributed computing resources in the cloud. MapReduce is a good fit for generating suggestive visualizations.

B. Rapid Miner/Rapid Analytics

Data analytics exist in most field of study from computational fault detection to finance to climate studies.

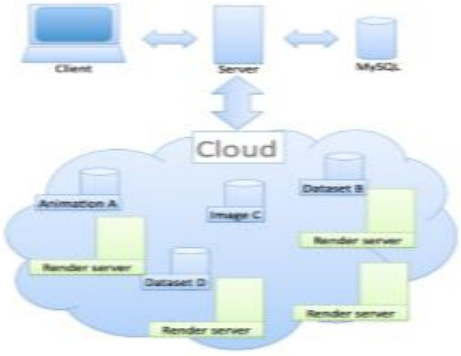

Figure 2: Overview of system architecture

There are lots of biggest challenges are facing, Sifei Lu et al [2] present a cloud-based framework that effectively utilizes data storage and cloud computing resources, a work flow management and scheduling engine, and remotely executes Rapid Analytics services to provide big data processing ability. In this, to overcome the limited support for parallelization is to integrate distributed computing framework into Rapid Miner for recognition and machine learning application. Radoop is the latest extension for Rapid Miner to execute distributed processes using MapReduce algorithms on Hadoop. Rapid Analytics is the server version of Rapid Miner that is enhanced with secured remote analysis web service scheduled remote execution and shared repositories. Custom tools were added to Rapid Miner to process spatio temporal climate data. Parts of the source code were also edited to enable the new format to be converted into native Rapid Miner/Rapid Analytics data types so that Rapid miner functions can be performed on these datasets.

C. Using MapReduce and Tree map

As information and data are increasing, existing in or as part of a tradition, some of the Web applications cannot store all the information data in cache and efficiency of data processing is clearly insufficient to address these issues. Mingyuan et al [3] solved this problem by using parallel processing in the cloud. On the Web, users submit the demands to the system. The results are returned to the Web after processing the data in the cloud. This helps to solve the problem of information and data processing of big data. Following are the steps:

- Use Hadoop to build our private cloud platform.
- Design a visualization prototype system.
- User's login our prototype system in they can apply for the released service, and use the released data to apply for a new service. They can also upload the relevant data and then apply for a new service.
- After users submitting the demands to the system, the system will analyze data in the background, and return the results to the Web.

D. Image based Vs. Model based approaches

Nick Holliman et al [4] describes about architectural differences. The requirement for scalable computational power combined with varied usage patterns makes it a strong prospect for cloud computing. There are some existing services which based on Web technologies and client-side rendering, these services can't make optimal use of the cloud because they depend on the client having the resources to store and render any amount of visualization data the server needs to send them. There are two ways of architecture pipelines through an image based approach or a model-based approach. In image-based approach, in which all rendering is undertaken in the cloud and only the pixels in the image are transmitted to the users display and all responses to interaction events are computed in the cloud. In the model-based approach, where the rendering is done on the client and the cloud sends the geometric objects making up the visualization to the client.

E. Ray-casting pipeline

The tasks of medical data visualization to cloud centers presents new security challenges. Manoranjan Mohanty et al [5], shows a framework for cloud based remote medical data visualization that protects the security of data at the cloud centers. The basic idea of framework is to securely outsource all of the client's interaction dependent rendering operations from server to the cloud data centers. To achieve this, develop a secure volume ray-casting pipeline that hides the color-coded information of the secret medical data during rendering at the data centers as they integrate the cryptographic secret sharing with pre-classification volume ray-casting. Initially, the server performs data capturing and data preprocessing operations. Then creates *n* shares of all the preprocessed information required at the data rendering step and distributes them among *n* different data centers. By using an independent secure network channel, the server also transmits the information about the data centers and the shares they are holding to the client. Upon receiving the rendering request, each data center performs the rendering operation on its share parallel with other data centers and transmits the corresponding rendered share image to the client.

F. Web based rendering

Soren Dische et al [6], present a web-based system for the collective and shared investigation and careful examination of arbitrary large 3D point clouds. This approach is based on standard WebGL on the client side. It is able to render 3D point clouds with billions of points. It uses spatial data structures and level-of detail representations to manage the 3D point cloud data and to deploy out-of core and web-based rendering concepts. Alternative 3D point based rendering techniques and post processing effects are provided to enable task specific and data specific filtering and highlighting. Figure 3, shows the web based rendering techniques includes, thick client application which uses a central server paradigm for all operations. This method will reduce the server side workload and able to serve massive number of clients. In thin client approach, the data is rendered on the server and giving only the resultant images. This will increase the server work load and each time user interaction triggers a new data request and hardware requirements are reduced.

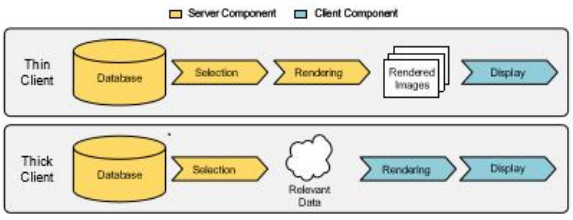

Figure 3: Comparison of web-based rendering concepts: Thin clients Vs. thick clients

G. Apache Spark cloud platform

In information explosion area big data analytics and open data mining becoming important. Wei-Yu Chen et al [7] presents the Integrated Platform for Government Open Data (IPGOD). The platform can work individually or in an integrated way and which consists of a Data System based on a cloud data warehouse and an Analytics System based on machine learning utilities. This attach or support the Apache Spark cloud platform to enhance low latency response and high performance. The Analytics System presents an expressive visualizations so that users can efficiently build machine learning models using a simple graphical user interface. Input can be divided into two categories as historical data and crawled data. The preprocessing consist of extractor, transformer and loader. Here HDFS is used for storage purpose. Users may select the features for modelling and visualizing the information for each column and the modelling data can be assigned to data frame. With the preprocessed data frame a model is created using machine learning algorithms. The default input format for spark ML is libsum rather than csv. It is difficult to visualize all the information for a huge amount of data, so pre-process the data in the back end and display much better plot for the huger data. The experimental results demonstrate that IPGOD realizes the open data and derives machine learning visualization in a user-friendly and intelligent way. Because Spark does not have its own distributed file system, they adopt HDFS for storage. In the application level, the main computing system consists of Spark MLlib packages. By providing both an open data warehouse and data visualization functionalities, users can mine their data intelligently

4.2 Different Concepts on Visualization

A. BDViewer

Here presents a web based big data processing and visualizing tool [8]. When the size of input data becomes very large then all the existing tools are not enough to do visualizations. BDViewer is a tool for processing and visualizing large data sets in a user-friendly way. Using a web interface it is able to view input csv files and output charts. Smart file loading scheduling are provided and algorithms run in Spark/Hadoop/MPI clusters to support data processing. Web browser operations are converted into MapReduce and MPI tasks. BDViewer provides software environment for distributed and parallel computing. They adopt Docklet as the cloud system in BDViewer. This also provides two categories of visualization, first for partial data and other for whole data. Because of the data to be processed is too small the whole data processing is done at front end. For whole data the following steps are followed,

1) User clicks a certain column and immediately the front end sent the column index to the back end
2) The back end processes the selected column to get the maximum and the minimum value. ,
3) The range [minval, maxval] is distributed to 1000 sections, the back end calculates the frequency at which data falls in each subinterval and sends the frequency array to the front end,
4) The front end transfers the frequency array to a heat map using the JavaScript plugin Heatmap.JS.

B. ggplot2 Visualization Package

The existing in or as part of a tradition data visualization has a sequences of a failing or shortcomings, for example, the display structure is too single and easy, the element information is not enough, the visualization function for the multi-factor data set is limited, which has become more and more impotent to meet people's needs for visualization. Wang Yang et al [9], apply Python web crawler for crawling the shipping recruitment information. Then, do some necessary data preprocessing through the R language based on reshape software package. The ggplot2 software package to do a series of visualization of shipping recruitment data. Finally, some important analysis and evaluations have been made according to the visualization results and some pertinent professional knowledge. The experimental results show that the ggplot2 drawing system can deal with the multi-factor shipping recruitment information dataset and can design the graph with high identification and large information.

C. ECharts framework

Deqing Li et al [10], contribute ECharts, an easy-to-use framework to construct interactive visualization. It provides three goals like easy to use, rich built-in interactions, and high performance. The underlying streaming architecture, together with a high performance graphics renderer based on HTML5 canvas, enables the high expandability and performance of ECharts. It is designed as a data driven streaming pipeline with stages of data processing, visual encoding and rendering, which produces graphic elements. Sometimes main UI thread updates may block caused by user interaction. So in order to solve this problem, here introduces new technique called increment rendering technique. Data can be loaded and divide in to several small chunks. Chunks are given to the pipeline one by one, and then be processed and rendered. To improve the performance of ECharts, here implement a multi-thread mode that separate data processing and canvas drawing in different threads. This is an extensional architecture that focuses on data streaming to enable direct reuse of defined interactions of existing components. This can reduce the workload when developing a new chart type for EChart. This is feasible because the handled data can be pulled into a separate module and each component has an individual processor, making the interactions on charts and components separate.

D. VisLT visualization service in Openstack

VisLT is a visualization software as a service in Openstack cloud infrastructure. The system introduced by, R Pacevic et al [11] will visualize only the results which are solved and persisted in private cloud. Apache jclouds API is used for accessing cloud services and also for managing 0penstack cloud infrastructure. User can communicate with cloud infrastructure by binding Openstack API and jclouds API. User interactivity is provided by a visual programming editor developed using the graph visualization library

JGraphX. XML documents are the inputs and which can be automatically generated in client. The visualization pipeline execute the algorithm and final image is transferred through the network and displayed in VisLT.

E. NeuViz Data Visualization Architecture

NeuViz is an architecture for processing and visualizing data produced by Neubot. Giuseppe Futia [12] shows an effective tool to navigate Neubot data to identify cases with protocol seems to recognize a distinction. NeuViz has a robust, scalable backend to support data analysis and the query is executed on network experiment dataset, then the result is stored in the NoSql database. The architecture consist of frontend and backend. Back end data and process data to allow for efficient visualization. Front end is a web interface that visualizes the data. The Raw Database accepts heterogeneous data organized in a unique manner by the Importer Stage. The Analysis Stage is a group of modules that sequentially fetch data from the Raw Database and process it to produce the data needed for the visualizations. Mostly a world-map-based visualization is used. The Analysis Databases are that store data which is ready to be visualized on the NeuViz Frontend.

5. ANALYSIS OF DIFFERENT APPROACHES

Among the various categories of computational parallelism deployment, MapReduce and Web based rendering will produce high quality output in terms of computational speed. The other approaches takes much lesser time but the result will not be sufficient. Table 1 shows the qualitative comparison of four different methods. Here, the technique used in these methods are discussed in brief and also the advantages and disadvantages of each method is given. Table 2 shows the comparison of different concepts of visualizations. From the various concepts or methodologies of visualization, NeuViz Data Visualization Architecture and Visualization using ggplot2 are most convenient as describing their advantages and disadvantages.

6. CONCLUSION

Visualizing big data is a computationally intensive task that most users undertake on an exploratory basis. The need for scalable computational power coupled with varied usage patterns makes it a strong candidate for various applications. In this paper, we have made a study on various aspects of big data visualization on web through considering computational parallelism and visualization concepts. There are different techniques and methods are used for big data processing and visualization. A detailed study on these techniques are done in this paper. Thus it has a lot of functionalities and benefits, by integrating these with visualization techniques will leads to new milestones in different areas of data analytics. From the study of gaining computational parallelism and visualization techniques, MapReduce and ggplot2, NeuViz are out performed as the best.

ACKNOWLEDGEMENT

The authors would like to acknowledge the contribution and support from the Computer Science and Engineering Department of Rajiv Gandhi Institute of Technology

Table 1: Comparison of Different Computational Parallelism - A Qualitative Approach

METHOD	**TECHNIQUE**	**ADVANTAGE**	**DISADVANTAGE**
MapReduce Model	Mapper is used to divide the data and reducer will combine it	1)Easy to implement 2)Fast implementation 3)Utilizes computational resources in cloud	1)Memory requirement is high 2)Human computer interaction is not convenient
Rapid Miner/ Rapid Analytics	Enhanced with web service, scheduled remote execution and shared repositories	1) Easy to implement 2) Increased speed up 3) Supports fault tolerance	1) It supports only spatial temporal data analysis
Image Vs. Model Based Approach	1) All rendering and interaction events are computed in cloud 2) Model based approach, rendering is done on client 3) Cloud based is the most efficient	1) Cloud rendering make optimal use of cloud 2) Able to process large amount of data	1) Waiting time for a response is high 2) Client side rendering is time consuming
Ray- Casting Pipeline	Develop a secure volume ray casting pipeline that hides the color coded information of the secret medical data during rendering at the data centers	1)Fast to produce results 2)High security of data	1)Less fault tolerant 2)High chance of failure
Web based rendering	Based on standard webGL on the client side	1)Reduce server side workload 2)Can serve massive number of clients	1)Server work load can be increased with thin client approach

Table 2: Comparison of Different Concepts of Visualization

METHOD	TECHNIQUE	ADVANTAGE	DISADVANTAGE
BDViewer	Docket as the cloud system and provides partial data and whole data visualization	1)Process large datasets in a user friendly manner 2)Fast implementation 3)Easy file loading and scheduling	1)Whole data processing is at front end required too much time 2)Heat map visualization is used only
ggplot2 visualization package	Through an R package facilitates varieties of plots and graphs	1)Easy to use 2)Efficient visualization than others	1)Requires dependency libraries
ECharts framework	Together with a high performance graphics renderer based on HTML5 canvas, enables the high expandability and performance of ECharts 2) ECharts is designed as a data driven streaming pipeline with different stages, which produces graphics elements	1) Easy to use 2) Rich built in interactions 3) High performance	1) Requires more loading time as individual processor for separate module 2) Cost is high
VisLT visualization service	A visualization software as a service in open stack cloud infrastructure	1)High performance 2)Reduced data transfer time 3)Execution time is moderate	1)Hardware requirement is high 2)High chance of failure
NeuViz Data Visualization architecture	NeuViz has a backend to data analysis and the query is executed and stored NoSql	1)Process heterogeneous data 2) Efficient and fast to data storage and retrieval	1)World map based visualization is used only 2)Less flexible and failed to raise warnings 3)Parallel processing solutions are not included

REFERENCES

1. Yuzuru Tanahashi, Cheng-Kai Chen, Stephane Marchesin and KwanLiu Ma, An Interface Design for Future Cloud - based Visualization Services, 2nd IEEE International Conference on Cloud Computing Technology and Science, Visualization and Interface Design Innovation (VIDI) Group University of California, Davis, 2010.
4. for Real Estate Information, SERVICE COMPUTATION: The Fourth International Conferences on Advanced Service Computing,Hangzhou, China , 2012.
5. Nick Holliman and Paul Watson , Scalable Real-Time Visualization Using the Cloud , IEEE CLOUD COMPUTING PUBLISHED BY THE IEEE
2. Sifei Lu, Reuben Mingguang Li, William Chandra Tjhi, Kee Khoon Lee, Long Wang, Xiarong Li and Di Ma, A Framework for CloudBased Large-Scale Data Analytics and Visualization: Case Study on Multiscale Climate Data , Third IEEE International Conference on Coud Computing Technology and Science, vol. 5, no. 3, pp. 49-55,2011
3. Mingyuan Yu, Donghui Yu, Lei Ye and Xiwei Liu, Visualization Method Based on Cloud Computing COMPUTER SOCIETY, November- December 2015.
6. Manoranjan Mohanty, P, and Pradeep Atrey, and Wei Tsang Ooi, Secure Cloud-based Medical Data Visualization , ACM, Nara, Japan, October 29November 2, 2017, .

7. Sren Discher, Rico Richter, and Jurgen Dollner, A Scalable WebGL - based Approach for Visualizing Massive 3D Point Clouds using Semantics - Dependent Rendering Techniques, The 23rd International Conference on Web3D Technology, ACM , Poznan, Poland , June 2022, 2018, .
8. Wei-Yu Chen, Peggy Joy Lu, and Steven Shiau, IPGOD: An Integrated Visualization Platform Based on Big Data Mining and Cloud Computing, ICBDC 2019, May 1012, Guangzhou, China, 2019.
9. Yan Li, Junming Ma, Bo An, and Donggang Cao BDViewer - A Web - Based Big Data Processing and Visualization Tool , 42nd IEEE International Conference on Computer Software Applications, 2018.
10. Wang Yang, Tian Ye, Li Tie-shan, Peng and Zhou Yihu, Visualization analysis of Shipping Recruitment Information Based on R, Tenth International Conference on Advanced Computational Intelligence (ICACI) , March 2931, 2018, Xiamen, China, 2018.
11. Deqing Lia, Honghui Meib, Yi Shena, Shuang Sua, Wenli Zhanga, Junting Wanga, Ming Zua, Wei Chenb, ECharts: A declarative framework for rapid construction of web-based visualization, ACM Journal on Visual Informatics, 2018.
12. R.Pacevic, A.Kaceniauska, The development of VisLT visualization service in Openstack cloud infrastructure , Advances in Software Engineering, 1-11, July 8 ,2016
13. Giuseppe Futia, Enrico Zimuel, Simone Basso, Juan Carlos De Martin, Visualizing Internet-Measurements Data for Research Purposes: the NeuViz Data Visualization Tool , National Conference AICA 2013 ..