Big Data Analysis for Data Visualization

Abstract: One of the main characteristics of scaling data is complexity. Heterogeneous data contributes to data integration and the process of big data problems. Both of them are essential and difficult to visualize and interpret large-scale databases since they require considerable data processing and storage capacity. The data age, where data grows exponentially, is a significant struggle to extract data in a manner that the human mind can grasp. This paper reviews and provides data visualization and the Heterogeneous Distributed Storage description and their challenges using different methods through some previous researches. Besides, the results of reviewed research works are compared, and the fundamental shift in the world of large data visualization of virtual reality is discussed.

Keywords: Big Data, Heterogeneous, Visualization, Distribute Data Value.

Introduction

Data analytics and visualization are joining the Big Data age with the ever-growing amount of data generated by computers, social media, mobile devices, etc (Mustafa et al., 2020; Obaid et al., 2020; Zebari et al., 2020). It is both essential and complex to visualize and interpret largescale databases since they require significant data processing and storage capacity (Dino et al., 2020; Mahmood et al., 2021; Zhu et al., 2015). According to Science Daily, the pace of data development in recent years is impressive; 90% of all technology in the world has been developed over the last two years (Dragland, 2013; Zebari et al., 2018; Zeebaree et al., 2020). All of this is a real flood that needs a paradigmatic change to the past regarding data processing philosophies, technologies, or techniques and further focus to withstand it (Alzakholi et al., 2020; Caldarola et al., 2014; Zeebaree et al., 2019). A new word, Big Data, which has received a lot of buzz in recent years, has been coined to efficiently detect this data boom and spread revolutionary technical technologies capable of dealing With this massive amount of data (Franks, 2012; Jader et al., 2019; Zebari et al., 2019). In reality, a look at Google Trends reveals that the word Big Data has been increasingly popular over time, from 2011 until today (Saeed et al., 2020a; Weinberg et al., 2013; Zeebaree et al., 2020). Big data can be described in many ways, based on the various viewpoints from which handling massive data sets is viewed. From a technical view, Big Data illustrates " Collection of data that are not capable of recording, store, handle and review traditional computing resources in database" (Dino and Abdulrazzaq, 2019; Manyika et al., 2011; Zeebaree et al., 2020). Big data is an

internal and decision-making concern from the advertisers' perspective rather than a technical challenge (Dino & Abdulrazzaq, 2020; Weinberg et al., 2013; Zeebaree et al., 2020). It can also apply to " data that goes beyond the spectrum of hardware and software resources widely used in its recording, control, and processing by the user in a tolerable period" (Caldarola and Rinaldi, 2017; Haji et al., 2020; Saeed et al., 2020b). Finally, Big Data should be understood from a consumer viewpoint as new exciting, complex computing technologies that supplement the old ones (Osanaiye et al., 2019; Zeebaree et al., 2017; Zeebaree et al., 2019). Considerable data are now produced by social networks, traffic sensors, satellites' imagery, the transmission of audio, banking, the stock market, etc. 3Vs explain attractively significant facets to extensive data (Velocity, Volume, and Speed) and presents data management structures such as connection database servers, which can handle multiple relationship records but is not versatile in managing unstructured or semi-structured data (Shukur et al., 2020; Shukur et al., 2020; Zeebaree et al., 2020). Therefore, it's essential to create new technology to collect data from diverse channels such as social networks, stock exchange, multi-sensor data, etc. The general operation groups included are (Chawla et al., 2018): Data Ingestion, Storage, Processing and Analyzing and Data Insight. The rest of the paper's contents are structured as follows: Section two gives the Big Data Visualization and visualization process background, and their methods are explained, then Big data visualization challenges are described. Section three reviews the previous literatures on big data analysis for data visualization. Section four discusses and compares the results and the techniques utilized in each related works. Finally, section five concludes the paper.

Background Theory

Visualization and big data and characteristics

Visualization seems to be a picture or graphic display of data. Data visualization has to be interpreted formally to evaluate and extract more in-depth perspectives from big data. Visualization of data helps to pull multiple data points together, understand data relations,discuss problems in real-time, and determine more easily where to concentrate analysis (Abdullah et al., 2020; Haji et al., 2020; Khalifa et al., 2019). It allows data scientists to find secret data patterns and how they are stored. Business analysts may also use techniques of data visualization to define areas that require change or enhancement, concentrate on variables that affect consumer behavior, and forecast revenue volumes

Big Data Visualization Process

As seen in figure 1, the method of visualization consists of the following steps (Chawla et al., 2018; Chen and Zhang, 2014)

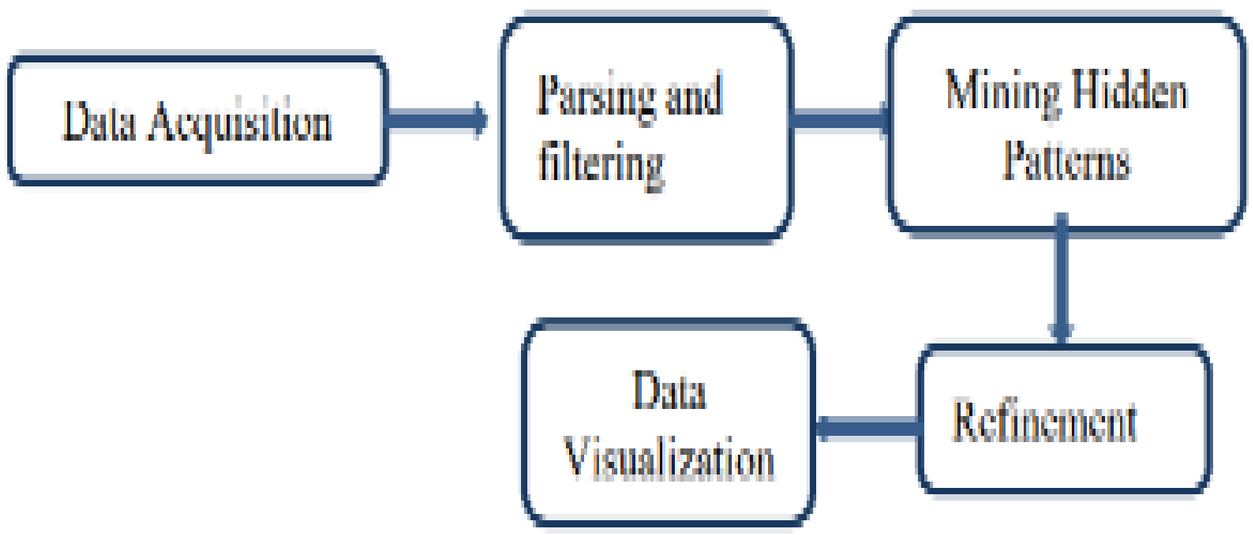

Figure1: Big Data Visualization Process

The first step in the method of visualization is the retrieval of data from multiple sources. There could be unstructured/semi-structured data obtained from heterogeneous sources, so it needs to be parsed in a structured format (Zeebaree, 2020). For visualization, all the data might not be necessary; the next move is to strip out the unimportant data. In the form of diagrams and charts, useful patterns are then derived and represented. Useful patterns are then extracted and depicted in charts and graphs in order to expose the user's simple understanding of secret knowledge (Chawla et al., 2018; Sallow et al., 2020; Zebari et al., 2017).

Big data visualization Methods

Several approaches to big data visualization have been used. These methods are graded by (1) data size, (2) variety for data, and (3) dynamics of data. Different methods for visualizing data are:

Treemap

This method considers a way to view hierarchical data as a collection of nested rectangles. The parent rectangle is separated into sub-rectangles by a tiling algorithm. A trained method is usually used. The rectangular region defines the number allotted to a category. The constraint of zero and negative values are therefore restricted to treemaps. Also, the hierarchy is skewed with more pixels (Khalifa et al., 2019; Tennekes and Jonge, 2011).

Circle Packing

It is an alternative treemap approach that uses circles to represent various hierarchical layers. The circle region determines the number of a type. It also uses multiple colors in different groups,

including the treemap. This approach is not space-efficient, as opposed to the treemap (Mohammad et al., 2017; Tedesco et al., 2013).

Parallel Coordinates

This method is a means of showing big data. Data components can be mapped separately through many sizes; both the forest and the tree can be seen in parallel coordinates. Line trends are drawn to collect consistent results. Person lines may be outlined to see the precise output of individual data items. Numerous data objects contribute to overplotting, though. This method is not used for data categorical (Johansson et al., 2008).

StreamGraph

This method is used to show the displacement of values along a different central timeline. It indicates improvements in data from multiple categories over time. The size of each stream form is equal to each category's values in a stream graph. Ideal for presenting a big dataset (Byron and Wattenberg, 2008). Data visualization tools will quickly gain awareness from a mass of information. People may discover things they do not know (outliers, secret patterns, or groups) with the perfect tool to visualize data. These instruments also let you dig into quickly shifting data sets. The main features for big data visualization applications are addressed in the table 1 (Alzakholi et al., 2020; Byron and Wattenberg, 2008). Another example is Space Titans 2.0, which helps to explore the System of Solar thoroughly. The goal was to get a new attitude on how our world looks from the benefit of modern VR's expanded space awareness. Skilling is a significant problem created by multidimensional structures from a Large Data Visualization perspective, which is required to scan an information branch to acquire any particular meaning or knowledge (Olshannikova et al., 2015; Zeebaree et al., 2019). Researchers are also involved in how simulated objects are combined with actual scene vision. This mapping could misrepresent the real scene as well as slow the device. Even physical and virtual distances are different; for that reason, an appropriate structure system has been developed to strengthen the relationship. Also, the paleontology, type interpretation, MRI, and physics must be studied effectively (Chawla et al., 2018).

Table 1: Big Data Tool Characteristics

Tools	Applications	Characteristics
Tableau	Market intelligence platform for the visual data collection used by scholars and public bodies	Can manage huge amounts of data, filter several data sets concurrently, users can generate and share dynamic and sharingable, dashboards depicting patterns and variants, develop interactive dashboards, built-in R support, Google Big Data Query API.
Plotly	online graphing, analysis, and static tools in both Python, R, MATLAB, Perl, J Arduino, and Restate graphics libraries	New open-access agile framework for data analytics and market research.
SAS Visual Analytics	Design tool; report, dashboard, and analytical distribution	Full research tool to allow users to recognize trends and relationships in data that are not clear initially
Microsoft Power BI	Using natural language questions on a dashboard to create immersive graphics, graphs and dashboards	For business users with their most important measurements in a single place, updated in near real - time, and available on all of their devices, power dash boards include a 360 ° view
D3.js	Using SVG, CSS specification, and HTML5 that are commonly applied	JavaScript library for immersive, collaborative web browser visualization

Big Data Visualization Challenges

Big data visualization is complicated based on the number, variety, and speed of data. The biggest problem when dealing with big data is how to manage huge data volumes and efficiently show the practical and usable outcomes of data visualization and analysis. A new mechanism needs to be built to look at the data in such a way as to tools are not capable of handling extensive data sets. The presentation tool will give us the lowest possible latency for display. Parallelization is often required for processing such vast volumes of data, which is a visualization task. Interesting trends may be described as the central aspect of large data visualization. The measurements of the data must be carefully selected for pattern mining. If we choose just a few dimensions, our visualization can below, and several fascinating patterns can be lost; likewise, if we select all the measurements, this can contribute to a complex view that is not usable for the users. E.g., visioning any points of data will lead to overplotting, overlapping, and the sheer perceptive and cognitive capability of the user, given the resolution of standard displaying (1.3 million pixels)"(Keahey, 2013). In terms of scalability, accessibility and response time, most existing visualization methods are poorly efficient (Ali et al., 2016).

Literature Review

Zhu et al. (2015) proposed Visualization by the Heterogeneous Distributed Storage Infrastructure (VH-DSI) solution to improve the speed of I/O and accelerate comprehensive visualization performance. Their proposed solution replaces the conventional parallel type file system with the distributed type file system type for supporting the visualization applications. Furthermore, the authors proposed a novel scheduling algorithm called HetetShe in VH-DSI for computing task assignment to data nodes regarding data locality and cluster heterogeneity. Also, VH-DSI contains a design for supporting the POSIX-IO of a distributed file system. The experimental results showed the importance of the proposed VHDSI solution and HeteSchi algorithm for visualization applications in achieving improved performance in both the response time reduction and visualization accelerating. Ali et al. (2016) proposed a novel processing algorithm named Scalable Uniform Storage (SUORA) through Addressing Optimally Adaptive and Randomized Numbers for heterogeneous devices. Their proposed algorithm is a random spurious algorithm that equally distributes data through a tiered and hybrid storage cluster. It separates and maps heterogeneous devices onto various buckets and allocates them to different sections in each bucket. Furthermore, the authors produced a deterministic and pseudo-random number sequence for data mapping among devices and segments. Data movement is also implemented for better read throughput achievement while retaining load balance regarding bucket threshold and data hotness. The evaluation performance results showed that the SUORA algorithm obtains effective adaptive data distribution for the heterogeneous storage system and data centers. Zhou et al. (2016) proposed HiCH approach to handle the distribution of improved data in a heterogeneous object-based storage structure and better leverage heterogeneous computers' ability. HiCH separates heterogeneous devices into separate buckets depending on the Sheepdog assessment and applies different consistent hashing rings to each bucket. According to hotness, data access, access time, and habits, it brings data into separate hashing rings. The results showed that the HiCH algorithm could increase storage systems' efficiency and make better use of heterogeneous storage devices. Kaneko et al. (2016) analyzed a storage system configured by the implemented guideline through using sysstat and to compare read/write data throughput among traditional data placement. This guideline is that data accessed by a client should be located on all servers equally. The results showed that the suggested approach increases the overall data throughput rate while increasing the number of access sources. Yu and Yu (2016) proposed technical visualization of heterogeneous processors with Legion runtime framework. The essential functions for conducting science visualization that can consist of several operations with various data criteria have been outlined. This approach will help users optimize storage partition programming, data visualization, and data movement for heterogeneous distributed-memory architectures, allowing multiple operations to be carried out concurrently on current and future supercomputers. Fiaz et al. (2016) provided techniques for Big Data and Data Visualization that make the use of data analytics more powerful and useful. The authors mentioned that any of the methods used to deal with Big Data

are complicated, and most organizations do not have enough experts to conduct the requisite data analysis. Data visualization methods simplify this problem and provide an ability to interpret and control data efficiently.

Malik et al. (2016) created a method that translates data in a way that will not require information leakage. This includes data and metadata such that they do not boost sophistication and retain a clear relationship between them. Metadata is extended to make it more useful for some kinds of data source. A case demonstrates textual material translation into RDF in a relational database. These methods may help the comprehensive coverage of the audio, video, image, and text formats data model. Loorak et al. (2017) presented the heterogeneous storage architecture using data value. Dynamically account of data value chose the convenience store according to the different data value. The high-level data value choice SSD strategy, the low-level data value choice HHD strategy second, makes the system's better performance. The research and assessment basis on Hadoop's distributed file system was also proposed. Li et al. (2017) presented the heterogeneous storage architecture using data value. Dynamically account of data value chose the convenience store according to the different data value. The high-level data value choice SSD strategy, the low-level data value choice HHD strategy second, makes the system's better performance. The research and assessment basis on Hadoop's distributed file system was also proposed. Wang (2017) introduced methods of data analysis for heterogeneous data and analytics of big data, Big Data techniques, some conventional methods of data mining (DM), and machine learning (ML). There is an overview of in-depth knowledge and its ability in Big Data analytics. The benefits of Big Data Analytics, High-Performance Computing (HPC), Deep Learning, and Heterogeneous Computing Integration are presented. Problems in dealing with heterogeneous data and research in big data are also discussed in dealing with heterogeneous data and massive data collections. Liu et al. (2017) proposed an improved method of visualization. The graphic framework is dynamically modified according to the user requirements based on the original visual structure. Also, based on the change in data, the relationships between entities can change dynamically. At the same time, they use the improvement approach instead of practicing SQL to query the database. The data set does not constrain the process suggested in this paper, and any data set can use the method. The study demonstrates that exploring the data interaction can be profoundly streamlined for users to visually explore and appreciate the data while the user has little or no comprehension of the data set structure. Zhi (2017) introduced the distributed optimization storage model based on hash distribution suggested by studying data processing characteristics in the background cloud computing. Furthermore, compared to the sequential storage distribution strategy, the distributed optimization storage model improves by 12.2 percent in terms of throughput, and the response latency decreases by 9.8 percent. The random pace of writing is around 8Mb/s. The simulation results demonstrate that the distributed storage device model architecture based on hash distribution is based on cloud computing. Iturbe et al. (2017) introduced three key contributions: (1)

Large Data Anomaly Detection Systems (ADSs) that could apply to Industrial Networks (INs) are surveyed and compared. (2) A novel taxonomy was developed to identify current ADSs based on IN. (3) A debate was addressed on transparent topics in large Data ADSs for INs that can further grow. Detection of Big Data abnormalities in Industrial Networks is still an emerging field, finally promising some potential areas of study work on these open problems. Kammer et al. (2018) proposed a systematic tool that makes it easier to evaluate and build ML-based clustering algorithms using various visualization features such as glyphs, semantic zoom, and histograms. Machine learning (ML) provides data discovery, e-commerce, or adaptive learning environments to create structures through clustering and classification. The result of the study developed the concept of interactive Big Data Landscapes.

Mahfoud et al, (2018) proposed an immersive visualization platform using Microsoft HoloLens to investigate heterogeneous data from different sensors. Their methodology discusses the core components for an observer to imagine dynamic data and discover hidden similarities in mixed reality; it also introduces automatic algorithms for event identification to identify suspect data. The demonstration framework illustrates the interactive mixed-reality analytics features, which liberates analysts from conventional computing environments and allows them to track and interpret data from time series anywhere on site. Zhou et al. (2019) applied modern PRS data duplication scheme to achieve effective data aggregation for heterogeneous storage structures. In compliance with data access trends, the PRS groups ob ject and distribute replicas with their functionality to heterogeneous computers. PRS uses a pseudo-random algorithm to refine counterparts' design by considering the efficiency and capacity of storage systems. The experimental results indicated that PRS is a highly effective replication mechanism for heterogeneous systems. Liang and Zhou (2019) proposed a broad data storage scheme based on HBase for remote sensing images. The approach uti lizes distributed storage and column-oriented open-source database (HBase) as the big data remote sensing image storage model. It uses tile pyramid technology and a parallel processing system (MapReduce) to create the remote sensing image tile pyramid. Finally, in the distributed database HBase, the remote sensing image data blocks are stored. This technique can effectively boost the storage problem of large data image remote sensing and has good reliability, scalability, and quality of processing.

Mehmood et al. (2019) proposed using big data Analysis technology to collect all information for further analysis. This interface allows data processing, retrieval, incorporation and further study and viewing of findings. This approach is the first effort to integrate various data points from four pilot areas in the CUTLER pro ject. Carranza et al. (2020) introduced a framework for higher-order spectral clustering by typing graphs and typing graph behavior in heterogeneous networks. The suggested approach constructs clusters that retain connectivity from typed graph lets to higher-order structures set up. The technique generalizes prior studies on higher-order clustering

of spectral. Authors technically illustrated various substantial consequences, including a Cheeger-like inequality for typed-graph let activity that indicates near-optimal limits for the technique. The theoretical findings significantly simplify previous work while offering a unifying theoretical basis for studying a higher order's spectral methods. Scientifically, three major implementations, including clustering, compression, and relation estimation, illustrate the efficacy of the method quantitatively. Woolsey et al. (2020) designing a novel calculation allocation method based on a Maximum Distance Separable (MDS) storage assignment for heterogeneous Coded Elastic Computing (CEC) network in addition to decrease time of the computation. In order to find optimum computational load and a "filling problem," recommend a novel formulation for optimization of combinatorics and solve it precisely by decomposing it into a problem of optimization.

After reviewing some research works related to big data visualization. The researchers used many methods and algorithms to get a significant way of extensive data. Those algorithms differ in satisfaction. For that reason, the challenges and the methods of the proposed approaches in related works using virtual reality based on visualization big data discovered a path of observing and analyzing diverse and complex data structures.

Discussion

It is evident from previously stated literature reviews that numerous research studies have concentrated on their importance. This study showed that researchers used different methods of solution for Big Data visualization. It enables distributed processing of large amounts using simple datasets through clusters of a computer model for programming. It has many essential features like fault tolerance, reliability, high availability, scalable, and costeffectiveness. In Table 2, the statistical assessment between these studies is shown.

Table 2: COMPARISON OF THE PROPOSED APPROACHES BY PREVIOUS RESEARCHES.

Author(s)	Algorithm	Objectives	Significant Results
Zhu et al. (2015)	HeterSche	Speeding Up Data Visualization Via A Heterogeneous Distributed Storage Infrastructure	reduces the response time by at least 5 times
Ali et al. (2016)	(SUORA)	Heterogeneous storage algorithms for flexible and uniform application distributors	Profoundly efficient adaptive data distribution for data centers and heterogeneous storage systems.
Zhou et al. (2016)	HiCH	Hierarchical Consistent Hashing for Heterogeneous Object-based Storage	Increase storage systems output and make heterogeneous computers more available.
Kaneko et al. (2016)	systat and compare read or write data throughput	A guideline for data placement in heterogeneous distributed storage systems	guideline get better the rate of aggregate data throughput while increment the number of access streams
Yu and Yu (2016)	Legion runtime system	A Study of Scientific Visualization on Heterogeneous Processors Using Legion	demonstrated the distribution scheme for various data and scalable execution and the easy utilization by a hybrid data partitioning
Fiaz et al. (2016)	Techniques for Big Data and Data Visualization together	Data Visualization: Enhancing Big Data More Adaptable and Valuable	provide an ability to interpret and control data in a more efficient manner.
Malik et al. (2016)	Relationship between data and metadata	Semantically improved simplified data transformation to heterogeneous data	useful for the wide coverage of the audio, video, image and text formats data model
Loocak et al. (2017)	Heterogeneous Embedded Data Attributes (HEDA)	Exploring the possibilities by integrating heterogeneous data attributes	new ways of utilizing familiar visualization techniques
Li et al. (2017)	data value and hedoop distributed file system	distributed heterogeneous storage based on data value	make better performance of the system
Wang (2017)	----------	Heterogeneous and big data processing	----------
Liu et al. (2017)	----------	A new method for representation of data and a heterogeneous data support scheme	method of exploring the data interaction can be profoundly streamlined
Zhi (2017)	hash distribution	Research of Distributed Data Optimization Storage Model In the Cloud Computing Environment	The hash distribution-based storage system template architecture is cloud based.
Iturbe et al.	----------	Heterogeneous Anomaly Detection	Promising some future fields of

(2017)		Systems in Industrial Networks:	research on the already evolving manufacturing networks
Kammer et al. (2018)	Clustering algorithm for machine learning	Big Data Landscapes: Improving the Visualization of Clustering algorithm for machine learning	Created the concept of interactive Big Data Landscapes.
Mahfoud et al. (2018)	Microsoft HoloLens	Immersive visualized heterogeneous on-site decision-making for irregular identification	interactive mixed-reality analytics features, which liberates analysts from conventional computing environments and allows them to track and interpret data from time series anywhere on site.
Zhou et al. (2019)	pseudo random algorithm	A Pattern-Directed Replication Scheme for Heterogeneous Object-based Storage	extremely effective replication scheme for heterogeneous store systems
Liang and Zhou (2019)	HBase and parallel processing system MapReduce	Research on Distributed Storage of Big Data Based on HBase Remote Sensing Image	effectively boost the storage problem of large data image remote sensing, and has good reliability, scalability and quality of processing
Mehmood et al. (2019)	CUTLER	Large data lake deployment for heterogeneous information sources	The first effort to integrate a variety of data points from four pilot fields is this approach in the CUTLER project.
Carranza et al. (2020)	clustering, compression, and relation estimation	Higher-order Clustering in Complex Heterogeneous Networks	Simplifying prior studies considerably Thus giving way to theoretical framework
Woolsey et al. (2020)	novel calculation allocation	Elastic Coded Computing on Heterogeneous Speed and Speed Store	novel formulation for optimization of combinatorics and Solve it by decomposing it into an optimization method.

5. Conclusion

Heterogeneous data contributes to data integration and big data process problems. Both of them are essential and difficult to visualize and interpret large-scale databases since they require considerable data processing and storage capacity. This paper reviews some research works on big data analysis for data visualization. It also compares their results according to their algorithms and methods. For that reason, the challenges and the methods of the proposed approaches in related works using virtual reality based on visualization big data discovered a path of observing and analyzing diverse and complex data structures.