

## **Interactive Visualization of Datasets Using Data Processing and Data Presentation Challenges**

**Bharath Simha Reddy Maram**

School of Computer ,Science Engineering  
Lovely Professional University  
Phagwara,India  
[Bharathsimha.maram@gmail.com](mailto:Bharathsimha.maram@gmail.com)

**Ranbir Singh Batth**

School of Computer,Science Engineering  
Lovely Professional University  
Phagwara,India  
[Ranbir.21123@lpu.co.in](mailto:Ranbir.21123@lpu.co.in)

**Jyoti**

School of Computer Applications  
Lovely Professional University  
Phagwara,India  
[Dhanjujyoti1080@gmail.com](mailto:Dhanjujyoti1080@gmail.com)

**Abstract**—For Data analysis, the Visualization of data provides a powerful meaning. It is about the different origins of Interactive Data Visualization that have been created, over the last three hundred years ago and the legacy of employing Bertin's idea of visual representations that were mapped to their data objects. Differentiates the Raw data and Preprocessed data (Static, Dynamic) and the challenges faced in Data pre-processing and Data presentation over the last decade.

**Keywords**— Data processing, Data presentation, static, Dynamic, interactive visualization

### **I. INTRODUCTION**

A single image is the value of a thousand words. Understanding text or, information is extracted from data, could mean various things to completely different individuals. However altogether such considerable ways become visualization, as it is an essential tool. The method used to communicate information or data by turning the data into visual images called Data Visualization. Data Visualization is related to statistical graphs and, information visualization.

A proper data visualization is to be achieved by interactive data analysis. The process, that undergoes data visualization is of extracting data or a set of data into the disk memory and then selecting and applying the algorithms of Data visualization. It concentrates on the interactive visual data analysis of modern Visualization systems. Natively, most of the data of different businesses and companies are in the format of XML and transformed, stored in a relational format. The difference between interactive visualization and non-interactive information visualization could seem a minor one. The method of interchangeable visualization is simply a continuation of idle visual image steps with various datasets given in numerous possibilities. In reality, the datasets which are introduced throughout the method are unique. The main purpose of an interactive method is to make your mind up

what information one desires to know within the further steps supported by what someone has learned within the last step. Generally, someone might seek to envision simply a set of the previously used information or its summarization. In exploration steps, a number of operations became common within the data analytics that they have nonheritable tags: for instance, pivot table, slicing, dicing, roll-up, drill-down. The datasets retrieved throughout the observation method area unit connected is to be a basic priority for the planning of an interactive visual image tool.

If someone needed to indicate all knowledge patterns on the monitor (even it is created a very little sense for the purpose of understanding), there have been enough picture elements to try and do so. By the appearance of enormous datasets, whether within the data warehouse types or scientific databases, within which the style of information visual image tools must be created. Firstly, we tend to might not assume that data is shown on the monitor. The amount of information points that are extracted from data is currently larger in magnitude than the number of accessible picture elements. Data must be compressed before normal visualization techniques are requested. We want scalability in visualization. The time taken for data retrieval dominates the act of the processing of visualization, thus can't be neglected. Interactive visualization is not possible with an inefficient database. So, we tend to additionally want processing scalability.

## **II. LITERATURE REVIEW**

Godfrey, P. et al. 2016 [1] discussed the different origins of Interactive Data Visualization that have been created, over the last three hundred years ago. It concentrates on the interactive visual data analysis towards modern Visualization systems and the legacy of employing Bertin's idea of visual representations that were mapped to their data objects. Differentiates the Raw data and Pre-processed data (Static, Dynamic) and the challenges faced in Data pre-processing and Data presentation over the last decade.

Luo, Q. 2008 [2] explains about Data mining tasks and applications which are based on successful knowledge discovery that turns data into knowledge. Data mining tasks that help in knowledge discovery includes Summarization, classification, Clustering, Trend analysis. Data mining adopted its techniques from many research areas which further helps in Data mining applications.

Armstrong, R. 2001 [3] gave a detail introduction about the way to optimizing the performance of Data Warehouse. As decision making based on data became less time for the users, the data must be available 24/7 to the users in an optimized way. The neutral data model, Rational and Irrational Summaries help to maintain the optimized performance of data warehouse, which includes exporting data into a different platform.

Streit et al. 2008 [4] explain the general, uncertain data in different ways: possibility (values may be represented as possibility curves), estimates (values are known to be unreliable with unknown accuracy) and intervals (values are known to fall within a specific range).

Olsson T et al. 2016[5] Visualization issues discussed in this work are critical to understanding human perception and cognition problems. The development area can only provide more effective and practical ways of using Big Data after that. It can be concluded

that by considering fundamental cognitive psychological principles and implementing most natural interaction with visualized virtual objects, the methodology of data visualization can be improved.

Lidong Wang et al. 2015[6] Interactive visualizations can benefit from big data to gain great insight. The scientific process can be improved by collaborative brushing and connecting visualization approaches to networks or web-based resources. Web-based visualization helps keep complex data up-to-date and up-to-date views.

### **III. HISTORY**

The subject area Data visualization has been around centuries, it is not new. The arrival of DBMS brought automation to accumulating and filter through digital information. The created potentialities to picture the massive size of information with efficiency. For primary systems that are designed for envisioning the data, Bertin's work and therefore the plan of aligning dataset into visual constants was helpful. Columns and Rows might be re-arranged by suggesting that of various methods like grouping, ranking, sequencing. The other system that used Bertin's framework to figure out the data that was in an RDBMS, which was developed in 1986 [2]. The aim of the tool known as A Presentation Tool (APT) which was to advance an application freelance system that might be capable of accordingly making visual images for data in RDMS. So as to attain smart output, the APT authors have written an explicit design component and created the expectation that graphical images (visualizations) are statements in a graphical representation. They outlined to boot an idea of fluency and capability, that were probably encouraged by Bertin's plan of quality to a graphical illustration of information. Expressiveness is often implied as a capability to precise a group of facts that of a language. Efficiency is relevant to the flexibility of a user to know a given explicit illustration. Formalization of specific sentences was allowed by APT to see to what limit the input of visual image meets the efficiency expressiveness criteria. This tool used a task- analytic method that meant that viewers might say a logical depth of a visualized image work to be done. The system that analyzed the given task and selected an optimized path to give the output.

### **IV. LATEST VISUALIZATION SYSTEMS**

Tools that were designed in the earlier 2000's place has a lot of stress on analyzing the bilateral visual data. For instance, just in the case of the Rivet [3], the internal DBMS design was made to enhance the fast development of easily interactable visualization of huge datasets. Data was exported to the Rivet interface and keep it in an exceedingly kind of rows, that Rivet thought of as an in-ordered collection of characters that might be classified into tabular form if the data was of a similar format. Rivet has the capacity of supporting various kinds of dataset sources like files and databases. When comes to design, Rivet was working on homogeneous model and data objects separation from visual objects that have been making a single visual image having multiple distinct views of similar data. Additionally, Rivet used a mechanism of selectors (subsets of distinguishing data) and visuals. Experience gained while the developing of Rivet leads to the word of VisQL. It is a Polaris system language [4] that used in an idea of blocks comparable to attributes of tables. It is organized to generate the exact datasets of related queries that were directly described by a given field in terms of visual parameters. Additionally, it was supported for explicit visualization by exploitation methods like deriving further sub-fields, sort and filtering, tooltips and brushing, and redo, undo functions. Further fields were made in Polaris to allow viewers to visualize further information which was derived from a dataset like quantitative

aggregated measures, count of unique values, ad-hoc gathering, and the threshold aggregations.

## **V. DATA PROCESSING CHALLENGE**

Data processing is to convert the raw data into related information. It follows a method in which original data is stored in the machine and output is gained as information after continuously processing the data and the analysis of data. The data processing life cycle involves many various steps. Scalability became the most important test for visual analytics. Instead, maintaining with endlessly increasing massive data sets became a continuous effort in several different subjects of computer science. It's been several years that, the database community is working hard to seek out the answers to the challenges. Most of the challenges were hardware-based, however, most concerned about the latest algorithms for the query process. In this paper, I'm going to re-discuss about major latest trends in the area of data processing.

### **A. Exact Answers**

#### **1) Exact Answers to be extracted from Raw Data**

Several tools were made over the past thirty years for making up the less query time, however, the individual data records keep in RDMS tables have been processed by queries against the original data. Not amazingly, when we consider latency, it was probably the least efficient method for providing a dataset to a visual image tool. Over the past years, most of the query processing analysis in the database has been moved too far from that of the model either by giving a viewer with nearer answers solely or by data preprocessing to enhance the performance of the query. A recent trend in superior systems involves the utilization of the main memory of databases. When Hyper [5] collaborates with the main memory, the capability of many tabular data with huge multi-core parallel power processing can be achieved. To release the new computing capability of systems, a radical re-engineering of the database was necessary. Low-overhead of data illustration was compiled instead of understanding the code, classification of high performance, and low-priced synchronization of transaction. High-priced systems like Hyper have TB of information that is kept and then processed. However, it is possible to be affordable among the years.

#### **2) From pre-processed data, exact Answers to be extracted**

The concept of a materialized view was introduced inside a database association in the earlier 1980s. As a cache, the materialized picture was provided a quick response to the dataset. The performance distinction could also important in applicable areas, wherever queries were advanced or have to be compelled to collect a great number of datasets. In a very large data-warehouse, wherever just about all the queries involved in aggregation, such materialized pictures were known as cubes [6]. Instead of processing the queries as they get executed, typical aggregations like total sales price classified by the location, time, and product; should be pre-processed prior to and keep in the warehouse together with the original datasets. At runtime, similar aggregations will be connected directly to the display screen for a lot of efficient interactive sessions.

### **B. Approximate Answers**

The size of data which is to be pictured usually exceeded by magnitude orders the pixel numbers out there on the screen. The dataset needs to be deduced by aggregation, filtering,

principal component analysis as solely a little amount of original data will be shown on the screen. As every compression technique leads to detail loss, the “pictured answers” will solely be nearer. However, if you are not able to show precise solutions to the queries, maybe the answers collected from the database might be nearer in addition. Most of the analysis in the systems over the previous years targeted on calculating approximate solutions instead of precise solutions. There were 2 primary ways that to attain: statically, once queries were calculated over pre-processed data samples or through the progressive query process.

### **1) Sampling**

Data sampling could be a statistical analysis technique implemented to choose the data model, manipulate the data and analyze the data for a representative set of data points to recognize the patterns and trends within the models quickly data set being examined. It permits data scientists, predictive modelers, and different data analysts to figure with a little, manageable quantity of data about a statistical population to create and run analytical models quickly, whereas still producing exact findings.

There are five types of sampling techniques. Techniques include simple random sampling, Stratified, clustering, multistage, systematic sampling. Sampling can even be based on non-probability, an approach within which a data sample is decided and extracted based on the judgment of the analyst. As inclusion is decided by the analyst, it may be harder to generalize whether the sample accurately represents the larger population than once probability sampling is employed. A non-probability sampling of data methods includes Convenience sampling, consecutive, purposive or judgmental, quota sampling. When a sample is generated, it can be used for predictions.

### **2) Approximate Incremental Answers**

Generally, people start by asking vast, large-variety questions and then step-by-step filters their questions based on review or feedback. For people, they expect a very fast result rather than they don't need perfect answers for such questions. Interactive systems must generate continuously filtering estimated answers and that solutions are good enough for the client. To implement the incremental visualization in a straightforward approach is to implement in VisReduce [7]. This system uses columnar storage integrated with a slightly updated MapReduce technique. But instead of calculating the exact answers, VisReduce returns incomplete answers incrementally and, computed the larger data samples progressively.

### **3) Static Approximate Answers**

The concept of precomputed samples was introduced, to solve the issue of the unpredictability of SQL operators for sampling. We precomputed the sample data and store the sample for any further use, instead of sampling the data tables at runtime. These sampled data were created to use with particular queries, such that, queries return the guaranteed answers. When compared to traditional materialized views, the cost of storing sampling is negligible. Even, the materializing samples could be developed to build an entire database.

## **C. Dedicated and Tightly-coupled Systems**

The concept of integrating a visualization system with a database system is to be well defined, but the same level and integration stage have been understood to various people in

different ways. Devise system might be the first attempt to build a strongly joined system with visualization. The concept of plotting pictorial actions to database requests has been sightseen and enforced in numerous methods in several systems. However, communicating to data storage that the user desires to visualize could facilitate in different ways than simply query performance. Visualization tools can fetch the data from the database about what quantity of data required to visualize an image, therefore, restricting the quantity of data which has been sent from a data storage. This enhances the performance at 2 stages: it compromises the communication prices and removes the necessity for the reduction of data at a later stage.

It is fascinating to visualize the data; however, the challenge of data processing was addressed during a business conception organization. The Tableau software system is probably going the foremost productive between them. The corporate distributes a variety of visualization tools fitted to business intelligence such as Tableau Server, Tableau Public, and Tableau Online. The first design of the Desktop version of Tableau was planned that it has the capability to attach to a totally distinct relative still as a gradable database. So as to cut back the excess data, it uses data extracts that were formerly processed implementing Firebird open origin DBMS. Though, even with generated extracts supporting filtered, tested or moved up subgroups of the first dataset. Firebird became not to be economical and Tableau software system determined to form its own read-only and optimized column-based data engine for data conception.

## **VI. DATA PRESENTATION CHALLENGE**

In most of the visualization systems, the quantity of data that is to be pictured exceeds the number of picture elements of the screen by the magnitude orders. The data must be deduced or compressed in such a way that it will be shown. As increasing the compression rate, a lot of and additional information on the particular dataset might be lost. So, the process of data reduction should be done by a relevant demonstration of the changed dataset. Visual scalability was that of the system capability to display massive datasets effectively in terms of the dimension of data points or the sheer variety of those points. Filtering, Aggregation, Sampling are the main techniques for compressing the data inside the Database.

### **A. Filtering**

Data filtering is for eliminating the rows to get rid of unessential data. This can be done to confirm the single variable used while attempting to visualize the data. Removing unessential data reduces the nosiness below the amount that will confuse the visually analyzed. The simplest way/technique of filtering is the outlier removals. Outliers are the abnormal values in the dataset. In certain conditions, we would like to keep outliers. If we permit the outlier values within the dataset, they'll simply produce the noise, which can lead to wrong or false predictability of the model.

### **B. Aggregation**

Data aggregation is the procedure within which the info is collected and summarized, for the purpose like statistical analysis. Online analytic processing could be an easy kind of aggregation of data within which the merchandiser uses an internet coverage mechanism for processing the dataset. Data aggregation is user-based: Individual-based aggregation

services provide the customer with a point for collecting their private info from different internet pages. The individual can use one master's personal identification number (PIN) to grant access to their numerous accounts.

### **C. Sampling**

Sampling returns the fraction of initial data patterns which were extracted from datasets were given some specific probable value. In the context, the solution created by sampling must be nearly approximate and it represents unsure information. Generally, ensure data is often mentioned in 3 other ways: estimating the values with unknown precision, intervals (to fall in particular range), and the probabilities. It's a summon to show unsure data in a very easy manner legible to users. Data presentation was not the only efficient in terms of a clock and the storage required to generate a visualization. Whenever, dealing through the approximation—filtering, sampling or with aggregating data, and also more important to produce picturing in such a method, it contains scalability. It means the inputs specified for two interrelated queries in visualization must be generated similarly.

## **VII. CONCLUSION**

Many researchers had expected, and some were still saying, “Visual analytics tools must support the flexible and fluent visualizations uses at certain rates of resonant with the pace of human thought”. In a different context, for exploring the data to be really cooperative, the queries must be more established within a certain 1-5 seconds of latency. In the future, we might prefer to come up with new visualization tools and perform more researches, probably together with algorithms of data mining. Moreover, the approach to the new algorithm lends itself to applying to different applicable fields like e-commerce.

## **REFERENCES**

- [1] Hamad, Mortadha M., and Alaa Abdulkhar Jihad. "An enhanced technique to clean data in the data warehouse." In 2011 Developments in E-systems Engineering, pp. 306-311. IEEE, 2011.
- [2] Armstrong, Robert. "Seven steps to optimizing data warehouse performance." *Computer* 34, no. 12 (2001): 76-79.
- [3] Luo, Qi. "Advancing knowledge discovery and data mining." In First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pp. 3-5. IEEE, 2008.
- [4] Bosch, Robert, Chris Stolte, Diane Tang, John Gerth, Mendel Rosenblum, and Pat Hanrahan. "Rivet: A flexible environment for computer systems visualization." *ACM SIGGRAPH Computer Graphics* 34, no. 1 (2000): 68-73.
- [5] Olshannikova, Ekaterina, Aleksandr Ometov, Yevgeni Koucheryavy, and Thomas Olsson. "Visualizing big data." In *Big Data Technologies and Applications*, pp. 101-131. Springer, Cham, 2016.
- [6] Wang, Lidong, Guanghui Wang, and Cheryl Ann Alexander. "Big data and visualization: methods, challenges and technology progress." *Digital Technologies* 1, no. 1 (2015): 33-38.
- [7] Gray, Jim, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals." *Data mining and knowledge discovery* 1, no. 1 (1997): 29-53.

- [8] Im, Jean-François, Felix Giguere Villegas, and Michael J. McGuffin. "Visreduce: Fast and responsive incremental information visualization of large datasets." In 2013 IEEE International Conference on Big Data, pp. 25-32. IEEE, 2013.
- [9] Olshannikova, Ekaterina, Aleksandr Ometov, Yevgeni Koucheryavy, and Thomas Olsson. "Visualizing Big Data with augmented and virtual reality: challenges and research agenda." *Journal of Big Data* 2, no. 1 (2015): 22.
- [10] Donalek, Ciro, S. George Djorgovski, Alex Cioc, Anwell Wang, Jerry Zhang, Elizabeth Lawler, Stacy Yeh et al. "Immersive and collaborative data visualization using virtual reality platforms." In 2014 IEEE International Conference on Big Data (Big Data), pp. 609-614. IEEE, 2014.
- [11] Gorodov, Evgeniy Yur'evich, and Vasiliy Vasil'evich Gubarev. "Analytical review of data visualization methods in application to big data." *Journal of Electrical and Computer Engineering* 2013 (2013): 22.
- [12] Liu, Zhicheng, Biye Jiang, and Jeffrey Heer. "imMens: Real-time Visual Querying of Big Data." In *Computer Graphics Forum*, vol. 32, no. 3pt4, pp. 421-430. Oxford, UK: Blackwell Publishing Ltd, 2013.
- [13] Shull, Forrest. "Getting an intuition for big data." *IEEE software* 30, no. 4 (2013): 3-6.
- [14] Kim, Yong-Il, Yoo-Kang Ji, and Sun Park. "Social network visualization method using inference relationship of user based on cloud." *International Journal of Multimedia and Ubiquitous Engineering* 9, no. 4 (2014): 13-20.