https://doi.org/10.55843/ivisum2382165s

EVALUATING DATA VISUALIZATION TOOLS BASED ON IMPORTING AND PROCESSING TIMES

Fehmi Skender, page 165-177

ABSTRACT

Following the rapid development of information and communication technologies, we witness the inevitable production of vast amounts of data, often referred to as Big Data. These large datasets are not just a result of technological advancements and smart devices; they are also closely tied to human activity. Data, as a concept, holds little meaning unless processed and presented in an understandable format with the power to convey information effectively. Alongside the development of software for processing Big Data, there is a parallel evolution of software for data visualization. Data visualization involves translating information into meaningful visual formats such as charts, maps, or diagrams to simplify data analysis and facilitate informed decision-making. Due to the complexity of Big Data, visualization is crucial and serves as a gateway through which the brilliance of information reaches humans. This study encompasses Big Data, recent data visualization tools based on Gartner's rankings, criteria checklists commonly used for tool comparisons, and the interpretation of results obtained from various measurements of data import and visualization times. The research employs scientific methods such as literature review, experimental methodology, and comparative analysis. This study evaluates popular data visualization tools, including Power BI, Tableau, Zoho Analytics, Qlik Sense, Domo, and TIBCO Spotfire, based on Gartner's rankings. Testing these tools across various data types and sizes, we found that they perform similarly for large datasets. The statistical analysis strongly corroborated the hypothesis that there exists a substantial disparity in data download and processing times among these tools, particularly for datasets exceeding 1 gigabyte in size. Data visualization remains crucial for effective analysis, especially with vast datasets. Optimizing visualization tools for performance and human-centric design is essential in a continuously evolving technological landscape. This study highlights the significance of understanding and fine-tuning data visualization tools to enhance data analysis quality and efficiency.

Keywords: Big data, data visualization, visualization techniques.

Mr. Fehmi Skender
International Vision University, Gostivar, N.Macedonia

e-mail: fehmi.skender @vision.edu.mk

UDK: 004.9:001.81

Date of received:
03.08.2023

Date of acceptance:
26.08.2023

Declaration of interest:
The authors reported no conflict of interest related to this article.

INTRODUCTION

In the fast-paced landscape of the modern world, one can hardly escape the omnipresence of information and computer technologies. From the indispensable smartphones in our pockets to the ever-watchful computers on our desks, and the interconnected smart devices that form the Internet of Things, technology has seamlessly woven itself into the fabric of our lives (Viktor Mayer-Schönberger, 2013). As time hurtles forward, change occurs at breakneck speed, making it increasingly challenging to keep pace. The relentless demand for acquiring new skills and knowledge is not the only hurdle; the cost of staying at the forefront of this technological wave is a burden shared by all.

With the meteoric rise of information and communication technologies, a deluge of data has become an inescapable byproduct, an immense force aptly christened as Big Data. The genesis of Big Data isn't solely attributed to technological advancements and the proliferation of smart devices; it's equally rooted in their burgeoning symbiotic relationship with humankind (Skender & Ali, 2019). The term "data" would be devoid of significance if not for its processing capabilities and the formidable power it wields to convey information. Hand in hand with the evolution of big data processing software, there's been a parallel surge in the development of tools for data visualization (Skender & Manevska, 2022).

Data visualization serves as the bridge that transforms the amorphous sea of data into coherent, comprehensible visuals like charts, maps, and diagrams. This transformation serves to simplify the otherwise intricate process of data analysis and aids in making informed decisions (Ali, 2022). Given the sheer complexity of handling vast datasets, data visualization emerges as a necessity, a portal through which the beacon of information illuminates the path to human understanding. (Nuredin, 2023) It encompasses a vast realm, ranging from the arduous task of managing copious data to the latest Gartner rankings that scrutinize visualization tools and widely accepted checklists of criteria used to compare these tools. It involves deciphering the evolving landscape of data depreciation and visualization over time.

In the realm of scientific inquiry, various methods come into play, including comprehensive literature reviews, rigorous experimental methodologies, and the method of comparative analysis. These strategies converge to illuminate the dynamic landscape of data visualization and its indispensable role in our ever-evolving digital world. Mathematica, with its powerful data visualization capabilities, stands as a beacon in this sea of data, offering tools and techniques to craft meaningful visuals that unravel the intricate tapestry of information in the age of Big Data (Selimi A. , 2019).

Harnessing the power of digital tools in education is not just about keeping up with the times; it's about propelling your learning journey to new heights. These tools serve as invaluable allies, helping you not only understand complex concepts but also boosting your academic performance in tangible ways (Selimi, Saracevic, & Useini, 2020).

Digital tools offer you the opportunity to learn at your own pace, making education more flexible and accommodating your unique learning style. Imagine having access to interactive simulations that breathe life into abstract theories, making lessons engaging and memorable. With these tools, you can turn your bedroom into a virtual laboratory, your living room into a lecture hall, and the world wide web into your library of endless knowledge (Selimi & Üseini, 2019).

Moreover, digital tools provide instant feedback, pointing out your strengths and areas that need improvement, so you can adapt and grow. Your grades are no longer static numbers; they become a dynamic reflection of your progress. These tools empower you to take control of your education, setting you on a path of continuous improvement (Selim & Ali, 2022).

As you navigate the educational landscape with digital tools as your companions, remember that knowledge is the ultimate currency of the future. Embrace these tools, and watch as your grades flourish, your understanding deepens, and your enthusiasm for learning soars to new heights. The journey may be challenging, but the rewards are boundless. You have the tools at your fingertips – use them to shape your destiny (Selimi, Saračević, & Rushiti, 2018).

1. Big data and visualization

1.1. Big data

Big Data can be defined differently without any major difference. Big data is actually a product of how information is gathered, analyzed and used.

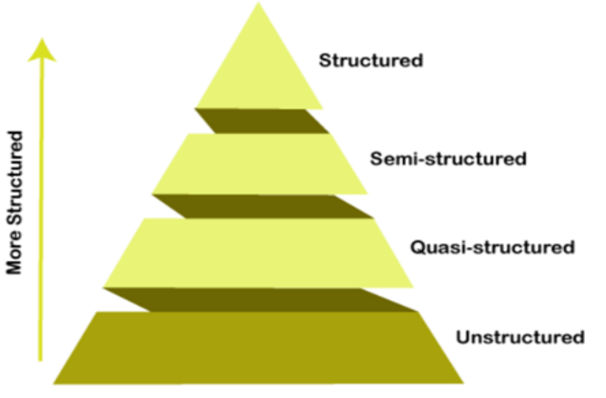


FIGURE 1. DATA TYPES
SOURCE: HTTPS://WWW.JAVATPOINT.COM/BIG-DATA-CHARACTERISTICS)

It is a huge volume of data generated every day from different sources such as social networks, various sensors, the Internet, and other sources. For example, by introducing an electronic journal in many countries, schools, in addition to the usual generated data, generate data related to students' disciples and their performance (Selim A. , 2021). Big Data contain valuable information that can be used to obtain indicators of making more efficient decisions, whether in education or in various other areas such as business, medicine, science and much more (www.javatpoint.com., 2021), (Figue 1).

- **Structured data:** In the structured scheme, together with all necessary columns, they are located in a table form. Structured data is stored in the system for managing relational databases.
- **Semi-structured: Semi-structured** data: In semi-structured data, the scheme is not adequately defined, for example, JSON, XML, CSV, TSV and e-mail. Online transaction processing systems (OLTP) are built to work with semi-structured data.

- **Unstructured Data:** Unstructured data: Unstructured data includes all unstructured files, log files, audio files, and images. Some organizations have large amounts of data, but do not know how to use data values, because the data is raw.
- **Quasi-structured Data:** The data format contains text data with inconsistent data formats that are formatted with effort and time, using some tools.

FIGURE 2. CHARACTERISTICS OF BIG DATA: TYPES AND 6 V'S

Given that semi-tructuated and structured data are much closer to processing, quasi and unstructured data, semi-structured data are more subject to qualitative analysis that is also of great benefit in analysis and decision-making (Viktor Mayer-Schönberger, 2013). Big datais characterized by several basic features, which depending on the nature of the data and time will increase. Recent scientific research and sources confirm 6 basic characteristics of large data, and if there are other sources where the number of characteristics is higher (Figure2).

1. Volume: The volume of data refers to the amount of information or data that is created, collected, and preserved by various organizations and systems. It is measured in various units of measurement such as terabytes, petabytes, exabytes, and zetabytes. The emergence of large volumes of data is called "big data" and is a significant phenomenon in today's information allocation.

2. Velocity: Speed as a characteristic refers to the speed at which data flows or generates. Given that the data is generated and circulated quickly and continuously, it is also necessary to perform the collection, processing and analysis of the data at the same speed.

3. Variety: Diversity refers to the many types of data that are generated, processed and analyzed in the context of large data. Diversity implies different formats and structures of data, i.e. structured, semi-structured and unstructured data. In many cases, this data can be in the form of text, numbers, images, sounds, videos, documents, emails and many other types of information.

4. Veracity: The veracity of big datarefers to the veracity of the data. A data consistency check is carried out, as well as an analysis of whether the data is inaccurate or authentic. Verification mostly takes into account several factors, such as data origin, collection and processing methods, as well as data security and access measures.

5. Value: The value is the fifth and last feature of big data. It is defined as the added value or usefulness that can carry the information collected in the decision-making process, business activity, or analysis. For data to be useful, it is necessary to convert to knowledge. This requires the use and combination of various technologies such as knowledge building, predictive analysis, text analysis, etc.

6. Volatility: The instability of big datarefers to the amount of changes and variations in data over a certain period of time as a result of various factors, such as changes in the way data are gathered, changes in regulatory laws and rules, (Nuredin, A; & Nuredin M., 2023) technological progress or other influential factors. Which means the data is not static and can change or grow over time.

1.2. Visualisation of big data

Visualization of big datais a display of large data, their understanding and interpretation through the prism of extensive and complex information. In a world where large amounts of data become a daily reality, visualization plays a big role in transforming raw data into meaningful and useful information. Not only does it provide an intuitive way to interact with data, but it also allows for rapid detection of shapes, trends and indicators without which a large number of information of great importance would remain hidden (Dykes, 2020). The human brain best preserves data in picturesque than in sound or written form. For this great reason, the data in any type and be, their final result would best be seen if it was in the form of a scheme, schedule or diagram (oracle.com, 2017). The different features of big datarequire users to be more engaged in researching innovation in visualization techniques when it comes to drawing indicators or making decisions (Wilke, 2019). Occasionally, new and advanced visualization techniques are developed based on the basis of data analysis, which include not only cardinality (data number) but also the structure and origin of the data (Khadija Begum, 2023).

Kernel Density Estimation (KDE), Kernel Density Estimation begins displaying data and the goal is to generate a distribution curve (X, 2023).
Box and Whistler Plot: are used to graphically represent numerical data classes according to quartile distribution of large data, allowing easy identification of outliers.
Correlation Matrix is used to quickly visualize connections between different variables by combining large amounts of data
Decision Trees: In a set of data, decision trees reflect the relationship between input and expected values. When the closeness of the link between target and input values is revealed, they are grouped into a branch of the decision-making tree.
Histograms: Histograms are a simple way to display information about large sets of data. They work by grouping data at intervals and showing how frequency (how many times they appear) of values at each interval. Each interval is represented as a segmented column of the diagram.
Connectivity Charts: Helps visualize and analyze connections between different events or phenomena and their mutual influence.
Scatterplot Matrices: is a powerful visualization technique that shows the multivariant connection between variable combinations using a network (or matrix) of scattered charts. This allows you to customize scattered charts with density ellipses for all data or only for groups of data for detailed analysis. This technique helps identify and analyze connections between multiple variables at once and is useful in finding certain trends in data (https://plotly.com/python/splom/, n.d.).
Line Charts: Are mostly used to display values or trends associated with time.
Sankey Diagrams: Are used to display flows or connections between different segments or graphics elements.
Gantt Charts: used to plan and display sequences of activities over time.

Tree Maps: are used to display hierarchical structures and divide data into their parts (Duke, 2023).
Pie Charts: are used to display the composition of the whole, where each part of the circle is part of the whole.
Word Clouds: Are used to visualize the most common words or phrases in text data.
Radar Charts: are used to display many variables compared to a specific reference standard.
Venn Diagrams: are used to display similarities and differences between different groups or multiplications.
Trend Lines and Regression: used to identify trends and regression between variables.

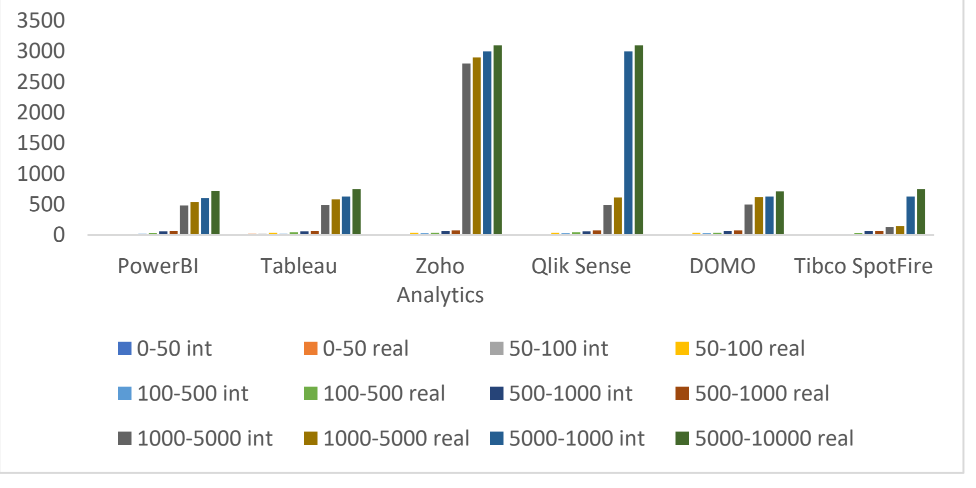
2. Methodology and materials.

In addition to reviewing literature, the study also includes an experimental part of chronometric measurement of the times of data importation and their visualization times. Comparative analysis was conducted using the SPSS software. The data is distributed in two groups, integers and real data types, at intervals from 0–50, 50–100, 100-500, 500-1000, 1000-5000 and 5000 to 1000 MB.

3. Results

According to gartner's last ranking, Power BI, Tableau, Zoho Analytics, Qlik Sense, Domo and TIBCO Spotfire tools are selected. The tools were tested with 6 different data groups of the type of integer and real data. For this purpose, data was taken from the large kaggle database. The data was divided into 7 different groups, 1-group, 0 - 50 MB, 2-group 50-100 MB, 3-group 100-500 MB, 4-group 500-1000 MB, 5-group 1000-5000 MB and sixth group more than 5000 MB. All 6 visualization tools imported the same amounts of data, with the results obtained per second (Figure 3).

FIGURE 3. GRAPHICAL REPRESENTATION OF THE RESULTS FROM THE EXPERIMENTAL PART AND FROM THE LITERATURE REVIEW



In terms of data size, a dicryptic analysis is done regarding variables,

TABLE 1. RESULTS OF THE DESCRIPTIVE ANALYSIS OF THE DATA BY TWO DATA TYPES

tools, and data downloading. Processing time shows no major concessions, about import time is a key factor when comparing tools. In terms of integer and real data types, SPSS correlation analysis shows a significant difference in data greater than 1GB. Descriptive statistics are given in Table 1.

TABLE 2. THE RESULTS OF CHI-SQUARE

Pro1000_5000mb_int

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	42.000ᵃ	36	.227
Likelihood Ratio	27.243	36	.853
Linear-by-Linear Association	.072	1	.789
N of Valid Cases	7		

pro5000_10000mb_int

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	42.000ᵃ	36	.227
Likelihood Ratio	27.243	36	.853
Linear-by-Linear Association	.472	1	.492
N of Valid Cases	7		

Pro1000_5000mb_real

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	42.000ᵃ	36	.227
Likelihood Ratio	27.243	36	.853
Linear-by-Linear Association	.039	1	.844
N of Valid Cases	7		

pro5000_10000mb_real

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	42.000ᵃ	36	.227
Likelihood Ratio	27.243	36	.853
Linear-by-Linear Association	.246	1	.620
N of Valid Cases	7		

With the support of statistical calculations of the difference in small squares, two hypotheses are placed concerning all 6 data visualization tools,

H0: There is no difference between data visualization tools at the time of data download.

H1 (Ha): There is a difference between visualization software at the time of data download.

Consequently, the hypotheses set can be concluded that the initial hypothesis is accepted to download data from 0-50MB, 50-100MB, 100-500MB and 500-1000MB in both types of data. While the alternative hypothesis H1 (Ha) is accepted in relation to the time of data download in data from 1000–5000MB and 5000-10 000MB.

4. Conclusion

Data visualization is essential for analyzing large data. Visual data presentation is important for quality data analysis, as it helps to detect connections and trends that may not have been observed by other methods of analysis. With a number of visualization tools available on the market, it can be concluded that each tool can be optimized to improve the quality of visualization and efficiency of data analysis.

Studying the time performance of visualization tools, especially in large amounts of data, is essential to improve the visualizations of large data. Visualization of Big Datafaces challenges such as visual differences, data crowding, and the need for high tool performance. Optimizing tools and visualization methods is an important challenge in researching large data. Visualizations should be human-oriented and focus on understanding and interpreting data.

Analysis of large-scale data visualization tools compares time performance in two categories: download time and data processing time. Although most tools showed similar performance, qualitative analysis was a challenge. The research involved examining time values for six different data sizes, including goals and real data, with a focus on large data sets. In today's world, visualization of large data is essential, with the need for efficient, clear and understandable visualizations. The research looked at various data visualization platforms, and the selection of the tool was made according to the prestigious gartner portal. Although techniques and methods are similar, differences in time performance in importing and processing data have been identified as a challenge. The study has opened up other issues and challenges in the field of visualization by acting on Python programming language libraries and the need for more modern and powerful computing machines.

REFERENCES

Ali, I. a. (2022). Pair Programming In Primary Education - A Macedonian Case Study. *12th International Conference on Applied Internet and Information Technologies AIIT 2022.* Zrenjanin, Serbia.

Duke, R. (2023). *https://www.gantt.com/*. (https://www.gantt.com/) Retrieved 09 01, 2023, from https://www.gantt.com/

Dykes, B. (2020). *Effective data storytelling : how to drive change with data, narrative and visuals* . New Jersey: John Wiley and Sons, Inc., Hoboken.

https://plotly.com/python/splom/. (n.d.). *https://plotly.com/python/splom/*. (https://plotly.com/python/splom/) Retrieved 08 02, 2023, from https://plotly.com/python/splom/

Khadija Begum, M. M. (2023). Comparative Study of Big Data Visualization Tools and Techniques. *Applied Informatics for Industry 4.0* (DOI: 10.1201/9781003256069-16), 188-199.

Nuredin, A. (2023). Legal Status of Artificial Intelligence and the Violation of Human Rights. Sui Generis, 2(1), 7-28.

Nuredin, A; & Nuredin M. (2023) Farklı Alanlarda Etik,International Vision University Press

oracle.com. (2017, 03 20). *oracle.com/solutions*. (oracle.com/solutions/business-analytics/,) Retrieved 08 02, 2023, from oracle.com/solutions/business-analytics/

Selim, A. (2021). Systematic Review of Big Data, Digital Transformation Areas and Industry 4.0 Trends in 2021. *International Scientific Journal Vision*, 27 - 41.

Selim, A., & Ali, I. (2022). Dijital Okuryazarlık Becerilerinin Kuzey Makedonya'da YüksekögretimÜzerindeki Etkisi. *6. International Congress of Economics & Business*, (pp. 81-91).

Selimi, A. (2019). Application of the Computational Geometry Algorithms in the Software Package Mathematica. *International Scientific Journal Vision*, 29-46.

Selimi, A., & Üseini, A. (2019). Yenilikçi eğitim ile dijital yetkinlik ve girişimcilik becerilerinin geliştirilmesi–Kuzey Makedonya örneği. *ICEB'19-International Congress of Economics and Business*, (pp. 204-213). Bursa.

Selimi, A., Saračević, M., & Rushiti, A. (2018). İnteraktif Matematik Eğitiminde Multimedya Cihazlarin Önemi – MAKEDONYA ÖRNEĞİ. In *Balkan Eğitim Araştırmaları 2018/ Balkan Educational Studies 2018* (pp. 31-41). Balkan Eğitim Araştırmaları 2018/ Balkan Educational Studies 2018.

Selimi, A., Saracevic, M., & Useini, A. (2020). Impact of using digital tools in high school mathematics: A case study in North Macedonia. *Universal Journal of Educational Research*, 3615-3624.

Skender, F., & Ali, I. (2019). A Digitalization And Industry 4.0. *International Scientific Journal Vision*, 47-62.

Skender, F., & Manevska, V. (2022). Data Visualization Tools-Preview and Comparison. *Journal of Emerging Computer Technologies*, 30-35.

Viktor Mayer-Schönberger (Author), K. C. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* New York: Houghton Mifflin Harcourt.

Viktor Mayer-Schönberger, K. C. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* New York: Houghton Mifflin Hourcart.

Wilke, C. O. (2019). *Fundamentals of Data Visualization.* Sebastopol: O'Reilly Media, Inc.

www.javatpoint.com. (2021). *https://www.javatpoint.com/big-data-characteristics*. (www.javatpoint.com.) Retrieved 08 02, 2023, from https://www.javatpoint.com/big-data-characteristics

X, E. (2023, 01 17). *https://www.aptech.com/blog/the-fundamentals-of-kernel-density-estimation/*. (www.aptech.com) Retrieved 08 2023, 02, from https://www.aptech.com/blog/the-fundamentals-of-kernel-density-estimation/