Universe International Journal of Interdisciplinary Research
(Peer Reviewed Refereed Journal)

© UIJIR | ISSN (O) – 2582-6417
SEPT. 2022 | Vol. 3 Issue 4
www.uijir.com

PRIVACY AND SECURITY IN BIG DATA

Author's Name: Dakshita Mishra

Affiliation: Student of BSc Data Science and Analytics – Essex University, Essex, Colchester, Wivenhoe Park, CO4 3SQ, United Kingdom

E-Mail ID: dakshitamishrauni@gmail.com

DOI No. – **08.2020-25662434**

Abstract

A plethora of data is available in any organisation's primary, secondary, and tertiary sectors. Structuring this data in an easy-to-use format has been a challenge for years. In this paper, we focus on and discuss the numerous ways of collecting, sorting, and analysing data while dealing with its challenging security and privacy threats.

Keywords: *Privacy, security, Big Data, organsiation*

INTRODUCTION

According to Wikipedia, big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. The slightly confusing definition with long, complicated words like "traditional data processing application software", so to make it easier, I remember it as a large variety of data collected passively from digital interactions which involve a significant volume and a high rate of velocity.

A variety of data sources, formats and infrastructures help reflect the possible vulnerabilities that can cause to a potential system. To handle compound situations, rules and regulations must be implemented and practised rigorously to avoid unauthorised access to data, mishandling, data sharing or leaks.

A secure and protected mechanism or regulation prevents third parties from interfering or seeking new methods to take advantage of the pre-existing systems.

If carefully considered, integrity, availability, and confidentiality of information will be more effective and helpful when developing a new programme.

The older systems have a different programming mechanism that suits the kind of data they were initially dealing with, so these mechanisms turn out to be ineffective in today's time until appropriate updates are made to the systems for them to handle the high volume of data.

Even the new open-source technologies tend to cause problems if they have default credentials or are not well-understood workwise.

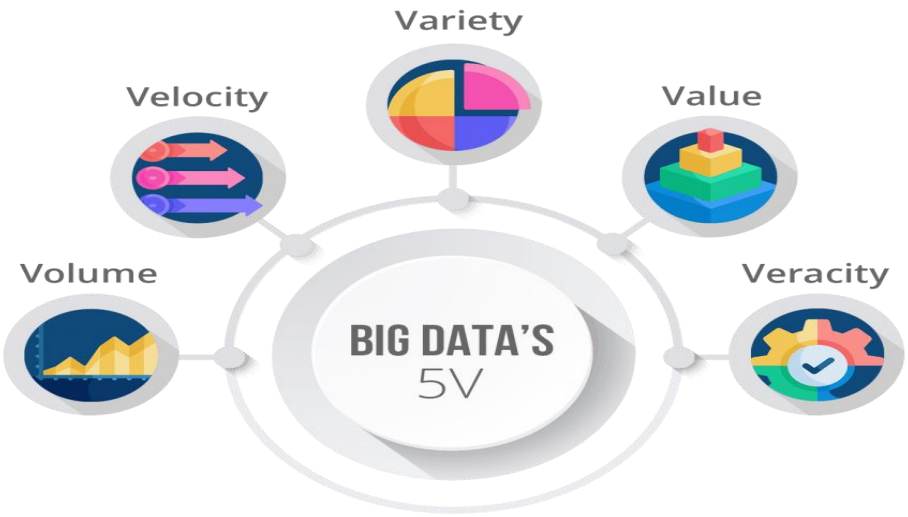
The actual use of big data has just begun, even though data has been incoming for years.

Databases are becoming increasingly important due to their ability to display massive amounts of

data in a way that makes analytics fast and comprehensive.

Although big data is relatively new, large data sets go back about 60 years.
In 2005 a few new developments for data storage programmes came into place. The result of open-source frameworks was necessary for the growth of big data because they make it easier to work with and cheaper to store. We are still generating massive amounts of data—more the devices connected to the internet, gathering data on customer usage patterns and product performance.

When we talk about big data, we talk about the main characteristics of big data. These characteristics come from the 5 V's. Having a brief and basic knowledge about these makes understanding big data much more manageable while giving it a deeper meaning. The top 3 V's include Volume, Velocity and Variety as they cover the significant chunks of characteristics when it comes to big data. Apart from these, we have two more essential V's, Value and Veracity. As we see later, each of the characteristics is defined individually to highlight each their importance and the part they play when it comes to big data.



THE 5V'S OF BIG DATA

VOLUME: Here, volume is not directly related to the word big in "big data". The quantity of unstructured/semi-structured, low-density data handled at any time makes big data interesting. Volume is the initial size and amount of data being dealt with when collected from any specific source [1]. So, to be considered big data, the volume needs to be large enough to meet the criteria of big data.

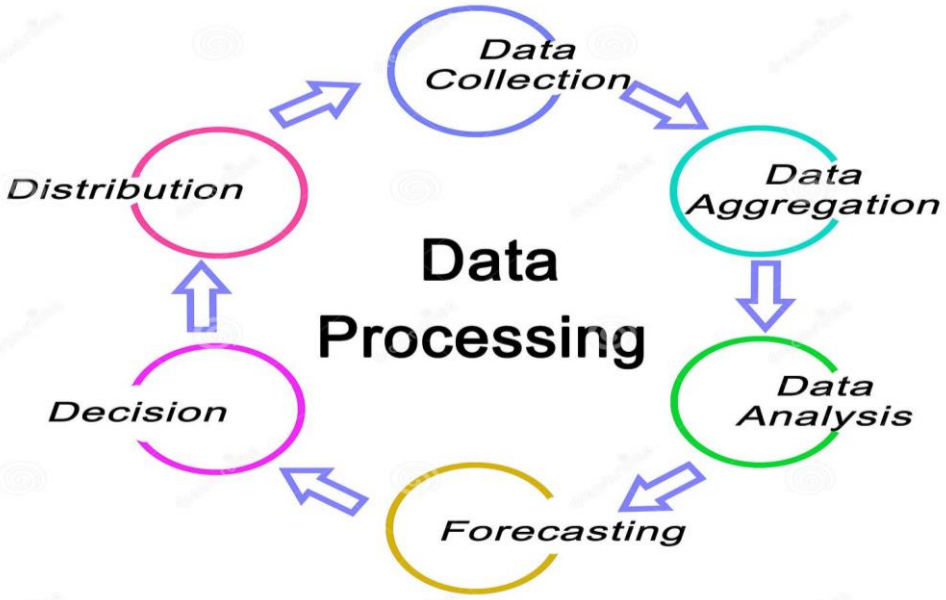
VELOCITY: The definition of velocity in physics is the speed in a particular direction. In the case of data, we refer to the speed at which data is collected, processed, analysed, and retrieved. In the last two years, the data collected was about 90% of the data generated in the world [2]. In every instance one turns to any search engine to look for answers to questions, we unknowingly add data into databases. Apart from this, some applications require real-time data and a sudden need for action and reasoning; hence having velocity only makes it smoother.

VARIETY: The significance of Variety in big data is essential as it focuses on the types of data

available. Initially, data that was collected was of the specific type called structured. However, as technology and tools have progressed, we can now work with other data types, such as unstructured and semi-structured data. At the same time, we lack flexibility in using traditional databases on the collected data types.

VALUE: Data has an ingrained value. Its value only grows through a journey of filtration and analysis to produce more valuable outputs than the original input itself. Most big tech companies can promise better products and services now, and it is because of the values that such efficiency and guarantee are promised from their data.

VERACITY: Now that we do not limit ourselves to working with only structured data types, we have got a lot more on our plate to handle and control. This whole set of veracity challenges with working on data types that traditional databases and mechanisms are not ready for makes big data so interesting.



DATA PROCESSING

Think of working in a restaurant to make any dish of liking. The first step is to collect all the ingredients and then follow the steps in the book to ensure the right spices go at the right time, and each step in the recipe book when followed to perfection. It is the same when it comes to data processing. We must follow a series of steps just like the one in the recipe book, and all we have to start with are the ingredients which in this scenario is the data. So as listed below are the steps for data processing:

DATA COLLECTION

This is the first step in the journey of data processing. Data from various sources, including data warehouses and data lakes. This step's essential in maintaining the quality of the collected data, for which the sources where the collection takes place are monitored for availability, dependability, and security.

DATA PREPARATION

This stage comes into action after correctly conducting the first stage. As the name suggests, data is prepared for further stages by cleaning, organising, and pushing the data through a quality check to avoid errors and redundancy. This step ensures that the data inputted in the next stage is error-free and high-quality.

DATA INPUT

With the newly organised and prepared data, minor changes are conducted to help convert the unusable raw data into a usable form. This step has now constructed data to allow it to complete the steps to follow effectively.

DATA PROCESSING

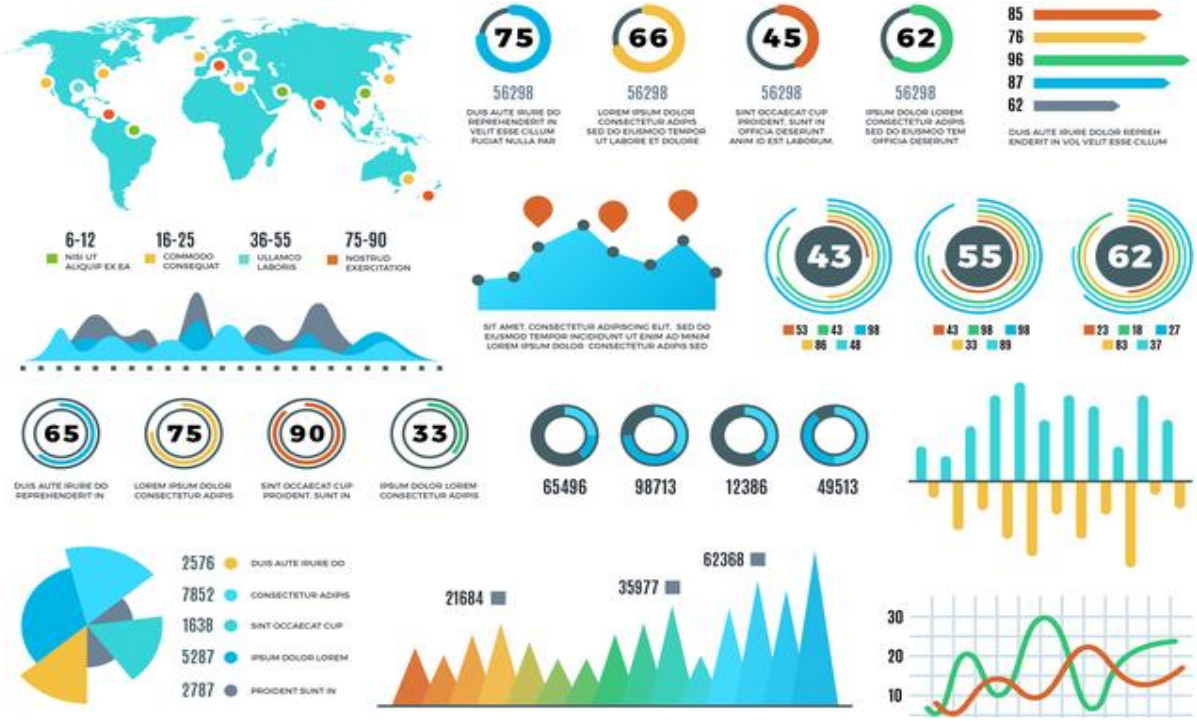
In this section of the data processing journey, we use additional tools such as machine learning algorithms to refine the data and use them in their respective categories.

DATA OUTPUT/INTERPRETATION

This is discussed in the next part to help clarify what the output can look like and why its represented in the way it is.

DATA STORAGE

The sole purpose of having to store anything is to be aware of its location when items in the storage facility are in need. Data can have two primary uses right after its processing, most of which are stored while the remaining immediately used depending on the output of the processed data. On the other hand, holding the data safely and in compliance with the data protection laws are also essential to abide by [3].



© UIJIR | ISSN (O) – 2582-6417
SEPT. 2022 | Vol. 3 Issue 4
www.uijir.com

DATA VISUALIZATION

As can be implied by the name, we visualise data. At the same time, this comes in several ways, such as maps, graphs, charts, diagrams, tables, and dashboards. As mentioned before, thinking of the restaurant scenario and heading towards a customer to serve them their meal. You are purely the carrier delivering the food to the customer, but being in that role, you know that when an individual gets their meal, something about how a simple meal is presented makes a difference.

So as a data analyst, if you can deliver the data interestingly and legibly to your client, it makes you and your work much more appreciable. The only thing that truly matters is the presented output based on the client's criteria. This client does not know the nitty-gritty of what goes into data processing and visualisation. When this output gets represented in a form that's useful to them, only then can they genuinely understand? Having fulfilled your purpose by giving meaning and character to the data to help utilise it for future and present needs.

Visualisation leads to interpretation which later leads to meaningful and informed decisions. As humans, we are attracted to patterns and colours more quickly than we are to a whole bunch of numbers. We tend to grasp and understand concepts faster when presented visually. Visualisation of data is one such art where numbers and words and combined to form colourful shapes and patterns to appeal to our eyes and brains. Our first instinct is to internalise and make conclusions from what we see at first glance. Aside from the advantages mentioned earlier, visualisation is vital to big data. The following are a few more reasons why it is essential to use visualisation tools in such circumstances. They help connect and find relations between entities while interactively aiding in exploring more opportunities. Lastly, they make sharing results and information based on a specific data set uncomplicated and unchallenging.

Shared are the general forms of visualisation with a bit about each of them [4]:

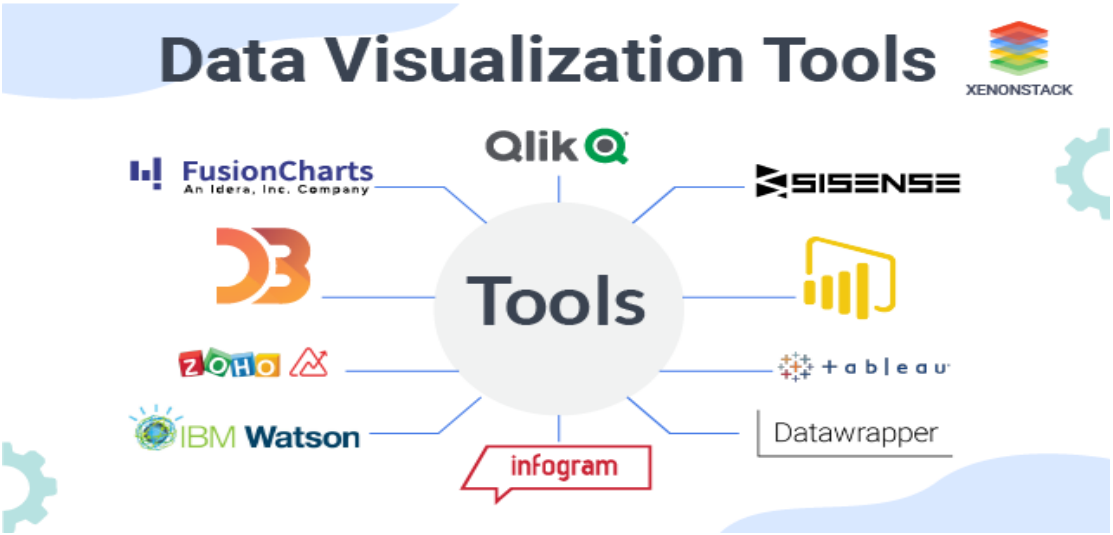
Charts: Umbrella term for the types of graphs and plot plots along an axis.

Tables: Data figures are placed alongside each other in columns and rows.

Graphs: Values plotted against another set of values on an axis. Plotting them against each other results in a trend or pattern when both values meet at intersection points.

Maps: are an example of geospatial visualisation as they show specific values on real-time maps of routes, countries, and cities.

Dashboards are a collection of multiple data sets and visualised versions; all put together in one place for convenience and comparison.



TOOLS FOR PROCESSING AND VISUALISATION OF DATA

Tools are separated based on their ability to deliver one or more functions. We currently focus on tools that only process and visualise data individually.

The main task of a data processing tool is to use the raw collected data and convert it into a valuable form for the user at hand. Data input into these tools goes through the steps detailed earlier. To summarise, the data will be cleaned, organised, and filtered based on a list of conditions (provided by the client).

On the other hand, while data visualisation tools have wholly different functions, their output goals are the same as data processing. These tools use the processed data in a more visually pleasing form to the eye while being of great assistance and understanding.

Ultimately, as they are all just tools, there tend to be a few overlapping interests, while their pros or cons also set them aside. When it comes to what they have in common, one of the first on the list is the ease of use for the user. Secondly, they can do multiple complex functions at once and in one application, making it convenient for the user. Lastly, they tend to have cost considerate attitude because, yes, there is nothing such as a free lunch, but neither do you want your customers to turn away just by a look at the cost they have to pay for the services at hand.

Moving ahead, we will examine the overall advantages and disadvantages that these tools might come with. Starting with the benefits, first and fore, these tools help add a lot of meaning and value to the data; next, they enable us to see emerging patterns and trends and take necessary actions on their basis and, finally, well-estimated analysis.

To conclude on this topic of tools and their functionalities, we must remember that with all good comes the bad. Hence there are disadvantages to using tools, but, in the end, the advantages outweigh the disadvantages. A few of the cons of using these technology-based tools are that they provide estimates, not accurate results, and they might have confusing correlations along with contrast colours that might make matters worse.

BIG DATA PRIVACY AND SECURITY ISSUES

When it comes to an organisation, its primary goal is to keep its users or customers happy and satisfied with its services. To make this possible, it does not only involve providing the services, but it is equally essential that the data/information the user shares with the organisation is kept safe. Keeping it safe requires the privacy and security rules of the organisation to be made right. A data protection plan prevents multiple fraudulent activities such as hacking, phishing and data stealing. Moving forward, we discuss privacy and security individually in a little more depth.

PRIVACY

Concerning privacy, everything is in the user's control, such as accessing, regulating, protecting, and sharing information. Having privacy regulations helps protect the user's particulars from being shared with third parties without the user's consent. Compromising one's privacy only plays in favour of the hacker. Besides this, privacy in big data is essential to keep sensitive data in the right hands and to minimise any risks. Personally identifiable information, also known as PII, requires utmost privacy regarding protection. The concerns keep growing with an increase in the pace of the data and regulations that govern it [5]. While privacy is a big subject, customer trust is equally important. With an increase in data collected from users, the chances of customer trust also elevate; as you get to know your customers better, you tend to gain a detailed profile about their preferences and hence, in turn, store more data than before. This data again requires the same privacy and data protection.

As blogged by the Truvantis website, an IAPP survey about the most demanded privacy management tools mentioned [6]:

- Data-mapping and Data-flow Tools
- Privacy Risk Assessment and Management Solutions
- Data Subject Access Rights (DSAR) and Data Subject Consent Tools
- Privacy Legal Update Tools

SECURITY

While talking about security, everything is under the control of the system. The system's job is to protect the data from falling into the wrong hands through an outside source, leak, cyber-attack, or breach. After the division of security and privacy challenges by the cloud security alliances, we get

four subcategories. Out of which, two are concerning security, namely infrastructure and reactive security.

Even though security and privacy are two very different terms, they go best when they are talked about together; the cause is their common outcome goals and preferences. Platforms do not seem to have data challenges and security limitations on location. Distributed data, Endpoint vulnerabilities, Data mining solutions and Access controls are a few of the topmost commonly caused challenges regarding big data [7].

CONCLUSION

In today's world, unstructured and unverified data are abundant; it has become a necessity to look after your security, privacy plans and rules before making any new, sudden decisions as there is an entirely different way of looking at things as we are advancing in multiple fields of technology and tools rapidly. The big data arena comprises security, storage, gathered information, transfers and privacy to access data. We must deal with threats and risks by monitoring them and obliterating channels leading to network leakages, communication, and analysed results.

In this study, the main focus is on the current concerns and challenges regarding ample data security and privacy. Security concerns will only increase as technology, tools, and software get updated and developed to deal with the simultaneously growing data. To deal with this rapid growth, we need to find solutions to the current issues and predicted upcoming ones too. Having well-developed human-system interactions to produce accurate results instead of estimated values would be a big step towards this development. I aspire that this research paper will provide you with the in-depth knowledge you need to understand big data and its surrounding ecosystems. Hopefully, any limitations you might have had earlier are now eliminated and have helped churn you towards the better development of software, systems, and solutions for your yearning future.

REFERENCES

[1] https://www.google.com/search?q=volume+in+big+data&oq=volume+in+big+data&aqs=chrome..69i57j0i512l3j0i22i30l2j0i10i15i22i30j0i22i30l3.4910j1j7&sourceid=chrome&ie=UTF-8
[2] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=38c49f6560ba
[3] https://www.talend.com/resources/what-is-data-processing/#:~:text=Data%20processing%20occurs%20when%20data,end%20product%2C%20or%20data%20output.
[4] https://www.tableau.com/learn/articles/data-visualization#:~:text=Data%20visualization%20helps%20to%20tell,data%20and%20highlighting%20useful%20information.
[5] https://www.informatica.com/hk/resources/articles/what-is-big-data-privacy.html#:~:text=What%20is%20big%20data%20privacy,the%20scale%20and%20velocit

y%20required.
[6]
https://www.truvantis.com/blog/data-privacy-tools-in-2022
[7]
https://www.dataversity.net/big-data-security-challenges-and-solutions/