Performance Comparison and Analysis of a Designed Temperature Controller for a CSTR model

Isha Singh
Amity Institute of Information Technology

Abstract- *Big Data Analytics and Processing is a very popular, exciting and streaming topic. In today's digital world it plays a pivotal role and is deployed in almost every field of life such as healthcare, education, Business Intelligence (BI), Machine Learning (ML), finance and many more. In every fortnight new tools and advancements related to this field keep on coming so in this research paper also includes various popular and advance tools related to Big Data Analytics and Processing. Big data is not a new topic, it exists much before the term 'Big Data' was invented so this research paper also includes the yearly progress of Big Data Analytics from its scratch stage to the form in which it exists today. The modern tools are being developed keeping challenges in mind and improving the quality of decision making. This field provides a platform to people to procure exciting packages. There is a great requirement of Data Scientists and Data Analysts in every sector. This makes this topic even more enthral.*

Keywords- Machine Learning, Business Intelligence, Data Visualisation, Data Mining, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Relational Database Management System (RDBMS), Structured Query Language (SQL), Decision Support System (DSS), Online Transaction Processing(OLTP), Online Analytical Processing (OLAP), Internet of Things (IoT), Natural Language Processing (NLP), Deep Learning(DL), Machine Learning(ML), Dashboards, k-nearest neighbours (KNN).

I. INTRODUCTION

BIG DATA

As the name specifies 'Big Data' refers to large amount of data. We all use Smartphones and laptops; these devices are connected through internet. Such devices use various applications like Facebook, Snapchat, Instagram, Telegram, YouTube Etc. These applications deal with a variety of data and that too in large amount. This large amount of data which is generated online is referred to as 'Big Data'. Every minute around One million individuals visit Facebook, 2.1 million snaps are posted on Snapchat, 3.8 million searches are conducted on Google, 4.5 million videos are watched on YouTube, and 188 million emails are sent. [1][3]

Figure 1: Data generated on Internet in an Internet minute [2]

II. CHARACTERISTICS OF BIG DATA

5V's make up the characteristics of Big Data. The concept of 5V's of Big Data was introduced by "Doug Laney". Earlier there were only 3V's: Volume, Velocity and Variety. Later 2V's i.e., Value and Veracity were added.

2.1 Volume:

This refers to the huge quantity of data which is generated by various online sources. The volume if data is increasing day by day due to increase in smart devices, IoT, cloud computing, social media platforms. This data is stored in digital containers called database. [7]

2.2 Velocity:

This refers to rate or speed at which data is generated by online platforms. Increase in no. of social media platforms, websites, IoTs is main reason for this.
Ex: 1.8million snaps are created within one Internet minute. [7]

2.3 Variety:

Big Data can be generated by humans, machinery or organisations also it can be organised, unorganised or semi organised so by this characteristic we refer to the various kinds of data that can be generated by online platforms.[7]

2.4 Value:

As we know that Big Data refers to the plenitude of Data but it is not necessary that all the data collected is equally important. So, in this characteristic we classify the data on the basis of how much important and useful it is.[7]

2.5 Veracity:

It refers to the measure of how much reliable the data is. Since the data is collected from multiple sources, we need to check the quality, trustworthiness and correctness of data. [7]

Figure 2: 5V's of Big Data [7]

III. LITERATURE REVIEW

This research paper dives into an overview of the Big data Analytics and Processing. It starts with explaining what Big Data actually followed by the various characteristics of Big Data (the 5V's). Further it explores the various ways of generating Big Data. And the categories in which the Big Data is classified. It explains the about the various types of Data Analytics and gives an overview of the multitude of tools used for the same. This research paper also traces the Transformative journey i.e., the evolution of Data analytics since 1960's to present. It then progresses towards theArt of Big Data Processing elucidating the stages involved in the process. Additionally it also encompasses multifaceted Adversities and Deployments of the topic.

IV. MEANS TO GENERATE BIG DATA

Big Data is not a modern term though it gained popularity in past 5-6 decades. Big Data existed even before the term 'Big Data' was invented. Big Data gained popularity only after emergence of social media.

Big Data can be generated by:

1. **Humans**: when humans use smart devices, they generate large amount of data that contributes to Big Data. E.g.: Uploading stories on Instagram and Facebook, making snaps, sending messages etc. Such data can exist in many forms like photos, videos, audios.[6]
2. **Machinery**: Such data is generated directly by machines without the involvement of any machines. E.g.: Data generated by IOT devices, Temperature sensors etc. Machines generate data in large amount and at high speeds.[6]
3. **Organizations**: There are a large no. of organisations which also contribute to Big Data. Such data is highly organised and helps in taking decisions. E.g.: Making in stock tables, profit and loss tables, Budget etc. [6]

V. TAXONOMY OF BIG DATA

Big Data can be classified into three distinct categories based on a taxonomy of data types and sources:

1. Structured Data: It is highly organised form of data which data is archived in the form of rows and columns. Here we use algorithms to retrieve, analyse and process the data. Since it is highly organised, it is easy to work with it.
Ex: In University admission system the data of students such as their Enrollment no., Address, phone no., email id, marksheets etc. are stored in highly organised form.[4]

2. Semi structured Data: Semi structured data is formed by combining both structured and unstructured data. This indicates that while it does have certain characteristics of structured data, it also has information that lacks a distinct organizational structure, does not adhere to relational databases, and does not follow established models.
Ex: Semi structured data is typically seen in JSON and XML.[4]

3. **Unstructured Data**: This type of data is highly unorganised and does not follow any particular format. It is very time taking and tedious to retrieve data and work with such type of data. It cannot be analysed without the proper data analysing tools.

Ex: Posts on social media, audio, video content, images etc. [4]

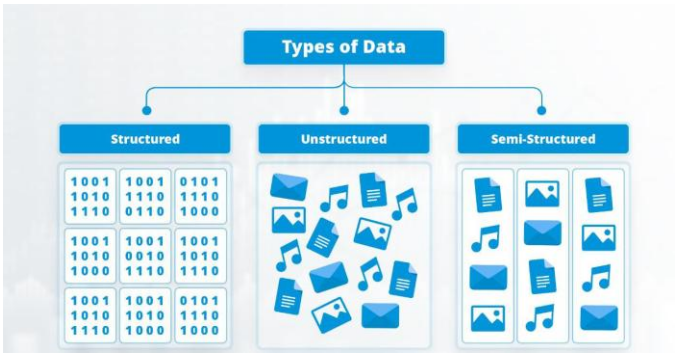

Figure 2: Types of Data [5]

VI. BIG DATA ANALYTICS

Big Data analytics is a practice of analysing a mammoth amount of data to discover useful insights from a cluster of data. It involves finding hidden patterns, future insights, correlations, changing customer preferences etc. so that organizations can take useful and wise decisions.
[8]

TYPES OF ANALYTICS USING BIG DATA:

Numerous Big Data Analytics Techniques exist, each having a specific purpose. Here are four main types:

1. **Descriptive Analytics**: It is one of the first step of the process. In this type of analytics, the past data of the company or an organisation is analysed to identify trends and know **what happened in the past**.[9]
2. **Diagnostic Analytics**: In this step the descriptive analytics is carried ahead. As the name suggests it involves diagnosing or understanding the reasons due to which a particular event took place. It involves doing data mining and root cause analysis. The goal is to identify **why did it happened in the past.**[9]
3. **Predictive Analytics**: Prediction means the act of forecasting or estimating a future event or outcome so this step involves forecasting **what will happen in future** based on past as well as present events and trends. Regression analysis, time series analysis, and machine learning are the common tools used for this.[9]
4. **Prescriptive Analytics**: This is the last step which involves prescribing or suggesting specific decisions or actions that need to be taken in order to achieve the desired outcomes. It involves identifying **how can we make it happen.** [9]

VII. SEVERAL TOOLS EMPLOYED IN BIG DATA ANALYTICS

With the rising importance of Big Data Analytics several tools have been created to streamline the process of analysing the Data. Here are some popular tools for this purpose:

1. Apache Hadoop: It is a very popular open-source Data Analytics tool released in April,2006 and is used by many companies. It supports distributed processing of data across large number of computers. It can be scaled up to thousands of computers from one server. [20]

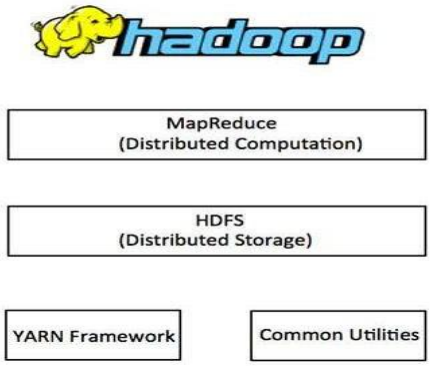

Figure 4: Hadoop [10]

2. Apache Spark:

It was open sourced in 2010.It is also a distributed computing system and in-memory processing capabilities for Big-Data Analytics. It offers advanced analytics capabilities, graph processing, machine learning, high-speed and can work with several programming languages. [20]

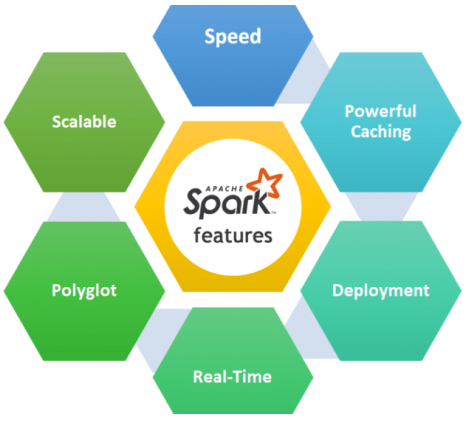

Figure 5: Apache Spark [11]

3. Apache Hive:

The fault-tolerant, distributed data warehousing system known as Apache Hive enables large-scale analytics. Data can be promptly analysed using Hive Metastore (HMS), enabling decision-making based on data. Apache Hadoop is the foundation of Hive. To read, write, and manage petabytes of data, Hive users can use SQL. [21]

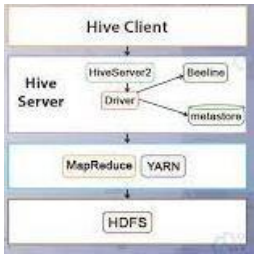

Figure 6: Apache Hive [12]

Figure 7: Tableau [13]

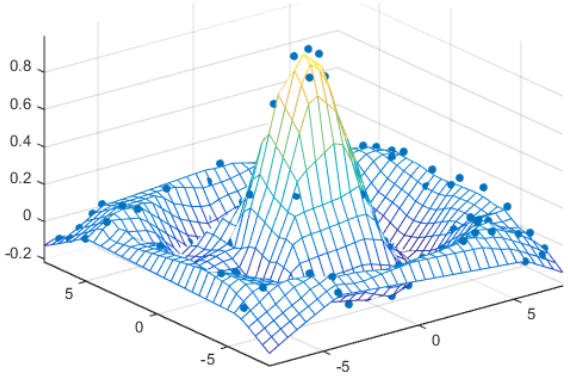

Figure 8: MATLAB [14]

4. Tableau: Business intelligence is the primary emphasis of the interactive data visualization software Tableau. California's Mountain View is where it was established in 2003. Tableau generates data visualizations in the form of graphs, dashboards, and worksheets using query relational databases, online analytical processing cubes, cloud databases, and spreadsheets It can also extract and store information from an in-memory data engine.[22]

5. **MATLAB**: MATLAB is an abbreviation for 'matrix laboratory'. It is a programming and numerical data analytics and data visualisation tool. It allows statistical analysis, signal processing, data visualisation through in- built functions and tools. [25]

6. Apache Cassandra: Many businesses rely upon Apache Cassandra because of its scalability and high availability for Data Analytics. It is used to handle tremendous volumes of data. Apache Cassandra is an open source, distributed NoSQL database. It has high fault-tolerance and thus is an ideal platform for mission- critical data. [20]

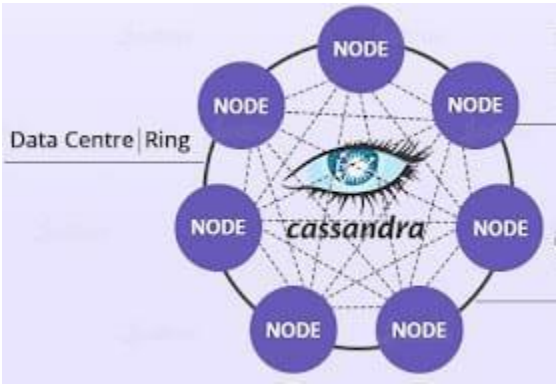

Figure 10: Power BI [16]

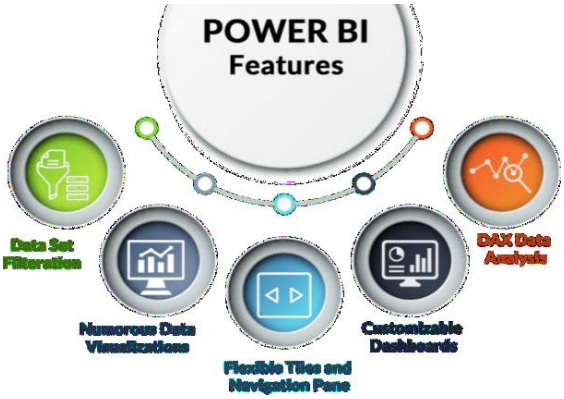

Figure 9: Apache Cassandra logo [15]

7. **Microsoft Power BI**: Microsoft developed this powerful Business analytics and Data Visualisation tool. It includes Data Connectivity, Data transformation, Data Modelling, Sharing and collaborations, creating real-time reports and dashboards and also allows mobile access. [23]

8. Python: Guido Van Rossum created the open-source programming language python in 1991, with many libraries used for analysing data.

NumPy- Scientific computing Matplotlib-Data Visualisation
Scikit-learn- Machine Learning and Data Mining
Statsmodels- Statistical analysis

Pandas- Data Manipulation and Analysis [24]

Figure 11: python libraries [17]

9. **R:** It is also an open source, platform independent programming language like Python. It is used extensively for data visualisation and Data Mining. It was created primarily to handle bulky statistical tasks.[23]

Figure 12: R programming logo [18]

10. **SQL:** SQL stands for Structured Query Language and is also known as MySQL. It is a popular Relational Database Management tool used to store, create, manipulate, retrieve, Define data. In this the user issues various queries to use the data. It is a must have skill for Data Analysts.

Figure 13: SQL Architecture [19]

VIII. EVOLUTION OF BIG DATA ANALYTICS

YEAR	DEVELOPMENT IN DATA ANALYTICS
1960's	Mainframe computers were used to store, process and analyse large volumes of data
1970's	-Structured approach such as Relational Database Management System (RDBMS) was introduced. Ex. R, Oracle etc. -Structured Query Language (SQL) was introduced to interact with data - Data was stored in Data Warehouses and used by Decision Support System (DSS) to analyse data and make strategic decisions.
1980's	-Online Transaction Processing (OLTP) systems for maintaining real time transaction records -Data mining techniques for extracting useful data
1990's	-Data warehousing was advanced with multidimensional databases and Online Analytical Processing (OLAP)tools.
2000's	-The term 'Big Data' gained popularity -3Vs of Big Data were introduced -Big Data analytics tools such as Apache Hadoop, Tableau, Cassandra, MapReduce, Apache Spark, MongoDB etc. were introduced.
2010's	-Data began to be generated by sensors due to the introduction of Internet of Things (IOT) -Real-time data analytics gained popularity -NoSQL databases were introduced -Emergence of Natural Language Processing techniques and deep learning -Data Analytics tools such as Microsoft Power BI, Looker was introduced.
2020's	-In-memory computing is focused as it is allows efficient utilisation of memory -Hybrid and Multi-cloud environments are being used by many organizations -Fast and Actionable data is gaining popularity -Increased demand for AI bases solutions -More advanced Data Analytics tools are introduced Ex. Qlik Sense [26]

Table 1: Yearly progress of big data analytics

IX. BIG DATA PROCESSING

As we all know that data is the raw information collected from any source for any specific purpose. This raw information cannot be used directly by the organizations to

make useful and important decisions. Thus, this data has to undergo through a process called Big Data Processing by Data Engineers and Data Scientists.

Big Data Processing refers to the process of collecting, integrating, transforming, cleaning, extracting, interpreting, analysing and visualising large amounts of data to extract useful data for taking wise decisions. It refers to the series of steps which makes handling Big Data easily, effectively and more efficiently. [27]

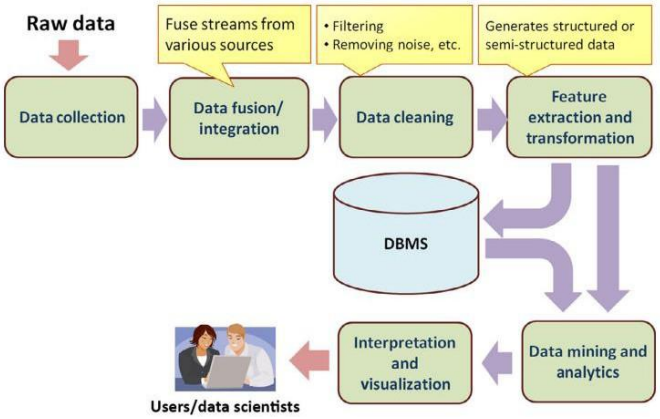

Figure 14: Stages of Data Processing [27]

STAGES INVOLVED IN BIG DATA PROCESSING:

Now we will talk about the various stages, methodologies or techniques through which the big data goes in order to become a useful information.

1. **Data collection**: This is the very first step in which data is gathered from various sources. The data ca be structured, unstructured or semi-structured. The data can be in any form like text, audios, videos, images, signals collected by sensors, surveys, interviews, social media etc. There are various ways for data collection such as log files, API calls, Web scraping, Sensors and IoT devices, surveys depending upon problem, sector or reason for which data is being collected.
2. **Data integration**: This is also referred to as data fusion. In previous step we have collected data from various sources so in this step we combine the data from various sources and make a unified format. This makes all the data available at one place and thus makes the process easier.
3. **Data Cleaning**: As the name suggests Data Cleaning refers to the process of removing garbage from the data. Here garbage refers to duplicate or redundant data, wrong data, inconsistent data and missing data. It is done to make the data accurate and reliable so that correct analysis and interpretation is done which will ultimately result in getting correct insights and wise decisions.
4. **Data Extraction and Transformation**: In this step useful data is retrieved from the whole and is converted into a format suitable for further analysis and processing. Extraction can be full or incremental. In full extraction all the data from the previous system is loaded into the current system whereas in incremental extraction only the data that has been modified since the previous extraction is loaded. Data transformation involves improving, enhancing or grouping the extracted data.
5. **Data Mining**: Data Mining is the approach of unveiling patterns, trends, correlations and other useful facts from huge data sets. It is also designated as Knowledge Discovery in Data (KDD). It is a very crucial step as it gives a competitive edge in market. It allows an organisation to focus upon their weaknesses, identify opportunities, solve problems related to business, take effective decisions and work according to current trends and scenarios.
6. **Data Analytics**: This refers to the process of discovering useful information from a cluster of data. It involves finding hidden patterns, future insights, correlations, changing customer preferences etc. so that organizations can take useful and wise decisions. In this statistical and computational techniques are used to understand the data, trend and patterns deeply. There are many modern tools for Data Analytics such as Apache Hadoop, Apache Spark, Power BI, Tableau etc.
7. **Data Interpretation**: After analysing the data the data is reviewed and evaluated to draw pertinent conclusions from it using different analytical and research techniques this is referred to as data interpretation. It is important for identifying critical findings that can be put into use, allowing businesses, organisations and researchers to get benefits from their big data related projects.
8. **Data Visualization**: As it is said that something seen through visuals is more memorable than any text in documented form so this step involves representing the conclusions, findings and outcomes of the Big Data processing in a visual format. Data visualization can be done through charts, graphs, maps, dashboards etc. These help in easy and effective understanding of data as they easily grasp the attention of the viewers. Some popular data visualization tools are: Tableau, Qlik Sense, Microsoft Power BI, Whatagraph etc.

Each step listed in the above sequence has its own importance in converting the raw information into information useful for taking wise decisions. On the addition of new data few steps need to be repeated.

X. ADVERSITIES OF BIG DATA ANALYTICS AND PROCESSING

Big Data is now a boom so with this modern topic many modern challenges are also connected which need to be handled carefully. Some of the common challenges are:

1. **Storage**: In today's date large volume of data is being generated every second whether it is structured, semi-structured or unstructured. Allocating memory space to store this huge volume, variety, velocity, veracity and value of data is a major concern. [28]
2. **Security**: Big Data also includes confidential information such as passwords, credit card and debit card information, personal information and other sensitive information. Such information is at a high risk of getting stolen by the various cyber thieves. New threats keep on evolving thus the system needs to be updated with new solutions available in the market to protect the data. Hence ensuring proper security is also an important and indispensable concern. [28]
3. **Algorithm Complexity**: Traditional algorithms involve high latency and memory requirements. With the increase in complexity of the variety of Big Data the algorithms used to process and analyse it have also become complex.
4. **Advanced softwares with high Computing Power:** The modern Big Data analysis and processing requires advanced softwares with high computational power such as Apache Spark, Apache Hadoop etc.
5. **Data Consistency**: Maintaining data consistency after processes such as data cleaning, extraction, transformation etc. is also a challenging task. This is because during such processes the data is repeatedly integrated, disintegrated and updated by multiple sources.[28]
6. **Real time processing**: In certain cases, enabling real time processing with streaming data is also complex and challenging. [29]
7. **Data extraction**: It has become difficult and time taking task to extract useful data from such huge volumes of data. It also involves choosing right samples and techniques to analyse and process data effectively and efficiently.

XI. DEPLOYMENTS OF BIG DATA ANALYTICS AND PROCESSING

Modern era has numerous applications of Big Data Analytics and Processing. It is used in almost every field and has made their works much easier than before. Its ability to collect, analyse, interpret, transform and visualise huge amount of data has transformed the businesses and research areas.

Some popular applications are as follows:

1. **Business Intelligence (BI):** Big Data analytics and processing is helping many companies in gaining useful insights from large amount of data collected through various sources. It helps in understanding customer behaviour, market trends, threats, opportunities and competitive analysis. This helps in taking better decisions and making strategic plans.[30]
2. **E-commerce**: It is helping the retailers to know about customer's likes and dislikes and their behaviour towards their goods and services. Also, it contributes in grouping customers and going with the current market trends. If done properly, then it also helps in getting an edge over competitors. It also strategizes marketing and advertising campaigns. [30]
3. **Healthcare**: The data collected in this field is related to patient's medical records, treatments done, medicines used, hereditary information and genetic records. This helps to figure out the disease and its causes. [30]
4. **Supply chain management**: Big Data Analytics and Processing provides useful information about customer demand, available stocks, stock supplied, supplier's performance, pathways through which transportation is done, costs involved in whole process. This information is useful in making future plans for improvement.[30]
5. **Finance**: In the field of finance, it helps in maintaining records related to each transaction, available balance, ongoing schemes related to interest rates these helps in guiding where to and how to invest t get more profits. [30]
6. **Government policies**: Big Data plays a very crucial role in knowing the current population, birth and death rates, food and healthcare facilities needed on the basis of this data various policies related to health, medical, education are introduced.[30]

XII. PROPOSED WORK

Big Data is a never ending and exciting topic so data scientists are continuously working to improve effectiveness

and efficiency of data analytics and data processing tools. The new tools include features such as low latency, better support to real-time processing, higher data privacy and security options, machine learning, better data visualisation, integration with AI, enabling better performance with streaming data.

- Since identifying trends in data is a very prolonged and wearisome task, we can use ML and DL algorithms to identify patterns, correlations from the data. This can help us to find the insights that can remain unnoticed by humans.
- Data Cleaning is an indispensable part of this process which includes outlier detection, handling missing, inconsistent data etc. This protracted task can also be streamlined using various AI algorithms like Rule-based cleaning, K- Nearest Neighbours (KNN) algorithm, Fuzzy Matching algorithms, Outlier Detection Treatment etc.
- Nowadays, there are various AI-powered Data Visualisation tools that can be deployed for generating collaborative dashboards with high precision even with live streaming data.

This will result in better decisions in less time and increasing the competitive spirit in big data market.

REFERENCES

[1] https://jattsingh-687.medium.com/what-is-big-data-how-to-handle-it-4383311fad16
[2] https://cloud.google.com/learn/what-is-big-data
[3] 1. 01-05.pdf (iosrjournals.org)
[4] https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between- structured-semi-structured-and-unstructured-data/?sh=5db0d4c52b4d
[5] https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics
[6] https://www.geeksforgeeks.org/data-analytics-and-its-type/
[7] https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and- applications/
[8] https://hive.apache.org/
[9] https://en.m.wikipedia.org/wiki/Tableau_Software
[10] https://careerfoundry.com/en/blog/data-analytics/data-analytics-tools/
[11] https://www.geeksforgeeks.org/top-10-python-libraries-for-data-science-in-2021/
[12] https://www.tutorialspoint.com/matlab/matlab_overview.htm
[13] Top 10 Big Data Trends of 2020 (analyticsinsight.net)
[14] https://hevodata.com/learn/big-data-processing/
[15] https://www.bornfight.com/blog/data-processing-and-biggest-big-data-processing- challenges
[16] https://www.simplilearn.com/challenges-of-big-data-article
[17] https://www.knowledgehut.com/blog/big-data/applications-for-big-data

LIST OF FIGURES

FIGURE NO.	CAPTION OF THE FIGURE	SOURCE	PAGE NO.
1	Figure 1: Data generated on Internet in an Internet minute	[2] https://medium.com/@toprak.mhmt/big-data- 7eeeadcf67e1	8
2	Figure3: 5V's of Big Data	[7]http://www.xenonstack.com/blog/what-is-big-data	9
3	Figure 2: Types of Data	[5]https://www.astera.com/type/blog/unstructured- data-challenges/	11
4	Figure 4: Hadoop	[10]https://www.tutorialspoint.com/hadoop/images/had oop_architecture.jpg	13
5	Figure 5: Apache Spark	[11]https://miro.medium.com/v2/resize:fit:1024/0*TQ4A z1sDKHMH1YtY.png	13
6	Figure 6: Apache Hive	[12] https://data-flair.training/blogs/apache-hive-architecture/	13
7	Figure 7: Tableau	[13]https://images.shiksha.com/mediadata/shikshaOnline /mailers/2021/naukri-learning/oct/28oct/What-is- Tableau.jpg	13
8	Figure 8: MATLAB	[14]https://www.mathworks.com/help/matlab/visualize/ nonuniform.png	14