

# **BIG DATA, HADOOP AND ITS CHALLENGES**

**BY**

**WAHEEDA AFREEN**

**ID: 182-15-11581**

This Thesis Base Project Report Presented in Partial Fulfillment of the All Requirements for the Degree of Bachelor of Science in Computer Science and Engineering at Daffodil International University.

Supervised By

**Asma Mariam**

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

**Md. Juel Mia**

Sr. Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH.**

**September 2022**

## APPROVAL

This Project/internship titled “**BIG DATA, HADOOP AND ITS CHALLENGES**”, submitted by Waheeda Afreen, ID No: 182-15-11581 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-Sep-22.

### BOARD OF EXAMINERS

  
13.09.2022

**Chairman**

**Dr. Sheak Rashed Haider Noori**

**Professor and Associate Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

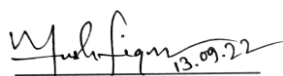


**Internal Examiner**

**Sazzadur Ahmed (SZ)**

**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology

  
13.09.22

**Internal Examiner**

**Mushfiquir Rahman (MUR)**

**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

  
23.9.22

**External Examiner**

**Dr. Md Sazzadur Rahman**

**Associate Professor**

Institute of Information Technology  
Jahangirnagar University

## DECLARATION

We hereby declare that this project has been done by us under the supervision of **Asma Mariam, Lecturer - Computer Science and Engineering, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**



---

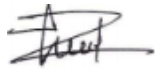
**Asma Mariam**

Lecturer

Department of CSE

Daffodil International University

**Co-Supervised by:**



**Md. Juel Mia**

Sr. Lecturer

Department of CSE

Daffodil International University

**Submitted by:**



---

**Waheeda Afreen**

ID: -182-15-11581

Department of CSE

Daffodil International University

## ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project successfully.

We are really grateful and wish our profound indebtedness to **Asma Mariam, Lecturer - Computer Science and Engineering, Department of CSE**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Big Data, Hadoop and Its Challenges*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Asma Mariam, Md. Juel Mia**, and Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

## **ABSTRACT**

Big data approaches are widely employed today. One of the pillars is big data analysis. With the nation's development and technological growth, it gets increasingly harder for us to organize them as our vast jumbled data grows. When dealing with terabytes, gigabytes, and a great deal of complicated and unstructured data, Big Data is what we refer to as. Real-time analysis is required. Every day, we depend on data to get by in our everyday lives. On social media, we hunt for random, unstructured material. The largest difficulty is dealing with these random data. Depending on the size of the dataset and the cluster's number of nodes, Hadoop Map Reduce can execute tasks in a matter of minutes. The second problem is undesirable for networks with a lot of data and for online transaction processing. It is also unsuitable for iterative execution. The purpose of this study is to explain a method to deal with issues like real-time processing, simple operations, and handling enormous datasets on the network, offering machine learning methods and 100 times quicker performance within memory primitives.

## **TABLE OF CONTENTS**

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv

<b>CHAPTER</b>	<b>PAGE</b>
<b>CHAPTER 1: INTRODUCTION</b>	<b>01 - 06</b>
1.1 Introduction	01
1.2 Motivation	03
1.3 Rational of the study	03
1.4 Research questions	03
1.5 Expected output	04
1.6 Project Management and Finance	04
1.7 Report layout	06

<b>CHAPTER 2: BACKGROUND</b>	<b>07 – 08</b>
2.1 Preliminaries/Terminologies	07
2.2 Related Works	07
2.3 Comparative Analysis and Summary	08
2.4 Scope of the Problem	08
2.5 Challenges	08
 <b>CHAPTER 3: RESEARCH METHODOLOGY</b>	 <b>09 – 15</b>
3.1 Research Subject and Instrumentation	09
3.2 Data Collection Procedure/Dataset Utilized	12
3.3 Statistical Analysis	12
3.4 Proposed Methodology/Applied Mechanism	14
3.5 Implementation Requirements	14
 <b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	 <b>16 - 19</b>
4.1 Experimental Setup	16
4.2 Experimental Results & Analysis	16
4.3 Discussion	19

<b>CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY</b>	<b>20 - 22</b>
5.1 Impact on Society	20
5.2 Impact on Environment	20
5.3 Ethical Aspects	21
5.4 Sustainability Plan	21
 <b>CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH</b>	 <b>23 – 24</b>
6.1 Summary of the Study	23
6.2 Conclusions	23
6.3 Implication for Further Study	24
 <b>APPENDIX</b>	 <b>24</b>
 <b>REFERENCES</b>	 <b>25 - 26</b>



## LIST OF FIGURES

FIGURE	PAGE NO
Figure 1.1.1: Understanding the 2 Vs of Big Data	02
Figure 1.6.1: Management about project finance	05
Figure 3.1.1: The integration of dataset pipeline.	10
Figure 3.1.2: Accessing the flowchart of data.	11
Figure 3.3.1: Statistical analysis and show with bar chart	13
Figure 4.2.1: Product sell on the city basis	16
Figure 4.2.2: Order based picked hour	17
Figure 4.2.3: Product demand and total sell.	18
Figure 4.2.4: Total sell calculation for monthly basis.	18
Figure 4.2.5: Product sold of return Prediction using Naive Bayes	19

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 3: Total order of product set and monthly amount.	12

# CHAPTER 01

## Introduction

### 1.1 Introduction

When people's reliance on Google, Facebook, YouTube, and Twitter increased significantly, the idea for Hadoop emerged. Terabytes and gigabytes of data are to be transformed using Hadoop so that they are usable by people. Our lives have become a lot simpler because to Hadoop. Data analysis for the benefit of people. We are dependent on all of the info on the internet nowadays and cannot function without it. Every day, we rely on some kind of information. Our time will be saved if they are sorted nicely. We have begun working on data analysis with all of this in mind.

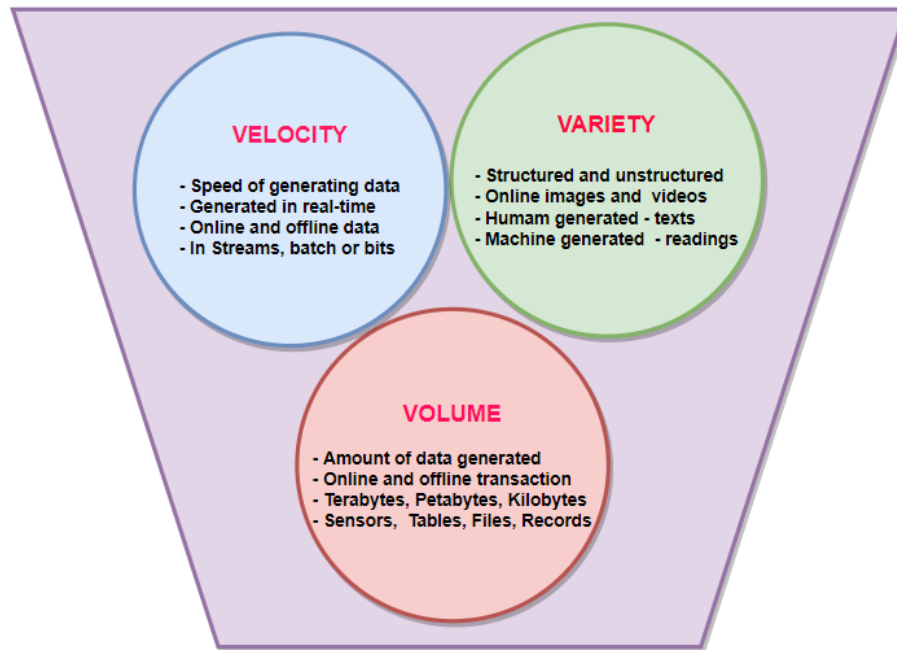
**What Is Big Data:** Big data is an order or combination of organized, complicated word processing file. Exceeds the capabilities of a conventional enormous amount of data coming from numerous data sources and huge data that are in diverse formats. Utilizing conventional analysis techniques is challenging or nearly impossible

**Hadoop:** A number of open source programs have been developed by the Apache team as part of the Hadoop project for reliable and scalable distributed computing. With a simple programming style, the Hadoop software library is a platform that enables distributed processing of large datasets throughout clustering. The goal is to scale from a small number of servers to millions of devices, each providing local computing and storage. The library is designed to detect and address application-level issues.

**Why big data Analysis:** Enormous businesses like Facebook, Google, and Microsoft today cannot provide better customer service without managing big data. They are unable to recognize what is best for their company and what is advantageous for themselves. Organizations can better utilize their data by using big data analytics to find new opportunities and patterns. This therefore results in wiser business decisions, happy customers, efficient operations, increased profitability, and smarter business decisions. Big data may improve judgments, provide confidence, and provide insights into important

situations. Big data is really beneficial to many different types of people and can cover a lot of information.

**Describe the three V's of big data:**



**Figure 1.1.1: Understanding the 3 Vs of Big Data**

**Volume:** Big data is the enormous amount of the word processing file that is collected and kept, including those on social networking sites like Facebook, Messenger, and Hospital. Terabytes of records are uploaded daily to Twitter, which contains a large number of message search options, postings, and comments. Big data can be handled with ease

**Variety:** We are all aware that daily data will be in the terabyte or gigabyte range and will not be structured. Our unstructured and probabilistically linked/dynamic data, which includes multiple factors, comes from a variety of sources.

**Velocity:** Due to its ability to compare with other data, velocity is crucial. The pace at which typical big data operates and the source of the data we are working with are both factors in velocity. Real-time streams of data are produced by big data at high speeds. processes for transactions.

## **1.2 Motivation**

Word processing file is growing every day. Word processing file first began to be collected when we started utilizing social media like Facebook, YouTube, and other internet services about 2005, data usage quickly soared. Because of this, the demand for big data analysis is rising steadily these days. Big data is being led by the computer thanks to enhanced digital sensing. For more than 20 years, computers have operated. Big data is one of the best things about the current period. Many thesis papers on big data, Hadoop, and characteristics have been published. Hadoop MapReduce is superior in some situations, whereas a different environment is superior in others. In some circumstances, the other ecosystem is preferable, whereas in others, Hadoop map-reduce is superior. We selected the subject because we will employ an already-in-use way to obtain the analysis result using a real-time analysis method. When there is a great loss of time, real-time data analysis is crucial.

For instance:

Detection of fraud involving ATM cards.

Business: Information about customers, reviews, and purchased goods.

Hospital: Analysis of patent issues and doctor information

## **1.3 Rationale of the Study**

Since 1990, big data has been employed. For a lot of business requirements and technological advancements during the past three decades, the term "big data" has been too significant. Big data helps people achieve much greater success. Due to the requirement for organized data, every large firm nowadays needs a data professional. Big data as we know it is possible. There is a lot of work being done worldwide involving big data. We'll talk about Hadoop map-reduce here. A larger amount of datasets will increase the processing time. We'll go over a few of the problems with higher numbers in this section. The main debate centers on it. Spark will make this process go more quickly than it did before. I'll also go through the technologies we'll employ to get around the Hadoop map-reduce problem.

## **1.4 Research Questions**

- How can we get facts regarding Hadoop, big data, and their features?
- What feature extraction strategies would be most beneficial?
- What are the advantages of reading this essay in one's life?
- What algorithms ought to be applied to produce the most precise results?

## **1.5 Expected Output**

Future generations will have a greater understanding of big data, Hadoop, and its properties than they did in the past, according to the thesis statement. They are also familiar with the real-time analytic Ecosystem. The ecosystem is the name given to the networked system. There are numerous datasets that require as like there are numerous types of social media. This topic's objective is to offer a framework, some tools, and a technique. We have used the Python programming language and certain spark libraries to import some spark libraries for analysis. Spark Session, StructType, StructField, and StringType from PySpark are examples of a library.

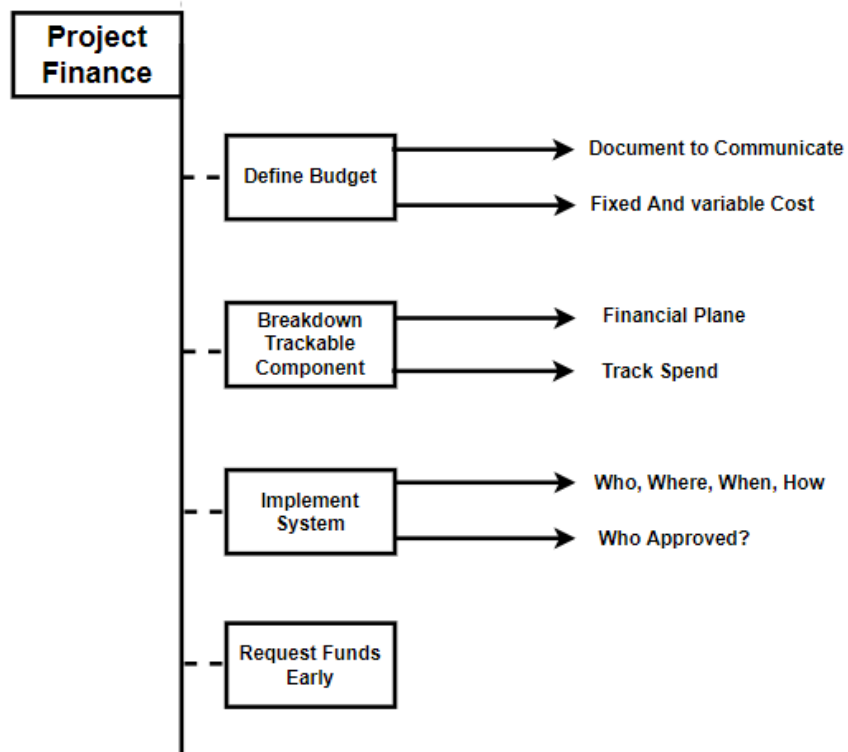
.

## **1.6 Project Management and Finance**

Without the support of the group organization have to left this environment and participated in a project where we had to solve the problem ourselves. These four guidelines, however, helped us avoid trouble or better comprehend the situation.

The first step in creating the budget is understanding that it must be documented, shared with all parties, and defined. the members of the project team who are responsible for approving it and allocating it to have

Second, we could list names of individuals next to businesses or various project-related objects that need to be tracked so that trackable elements might be integrated get in particular estimated expense.



**Figure 1.6.1: Management about project finance**

Implementing systems be the third item. Having a structure where things are monitored and approved is just as important as having a plan. first determining who will be performing this. The second is the location of this. Third: the time for modifications. Fourth, the method of execution. The authority to approve these budget alterations for these financial resources is the fifth and most crucial point.

The fourth tip is to submit your request for funding as soon as possible because it takes time for it to be approved and distributed. Depending on the organization or setting, this period may also require interacting with the government or other organizations, which can add to the delay.

## **1.7 Report Layout**

In chapter 1, we centered this paper on a few key issues. What we are working on, why we are working on it, and what we will learn from this paper. the primary justification for this investigation, the anticipated results.

Chapter 2: We will talk about our related work in this chapter. What obstacles must we overcome in order to succeed? Issues with comparative analysis and breadth. Throughout this section, we have used vocabulary that is familiar to us.

Chapter 3 is crucial to understanding how the entire project was implemented. In this chapter, many methodologies are discussed. Here, we explain the data collection process, dataset usage, statistical analysis of our data, recommended technique, and implementation needs.

Chapter 4 discusses the steps that must be taken before to the experiment. The second step is to use several diagram kinds to assess our results.

Impact on society, the environment, and other organizations is covered in chapter 5 in an unending loop. The topic of ethics is handled here as well. Here is information on how to plan a project for sustainability.

The paper is finally summarized in chapter 6, which comes last. Also included here is our strategy for subsequent research using this publication.



## CHAPTER 02

### Background

#### 2.1 Preliminaries/Terminologies

Every research paper requires choosing a topic. It is referred to as the preliminary research phase's first stage. It aids in our better understanding of the types of studies that are available and what has already been said on a subject. Let's choose a topic for our paper about sustainability in order to grasp it better. Topics, we needed to put them together in a survey question, so we did our first survey on big data, Hadoop, and how to analyze them. Is Hadoop Map-Reduce an ideal choice for big data analytics performed in real time?

**Is Hadoop map reducing the ideal choice for big data analysis that is done in real time?**

We now wish to use our paper to discover the answer to the question.

**Thesis: When analyzing real-time data, Hadoop map reduce is not the best option. So we talk about Apache Spark, another ecosystem.**

#### 2.2 Related Works

In this study, “Ishwarappa and Anuradha” describe large amount of data, its feature, the method and automation employed play with large amount of data. The term "big data" refers to the use of massive volumes of unstructured data by major corporations like Facebook, Google, YouTube, Twitter, and many other online social networking sites. The Hadoop framework-based big data approach deals with a sizable amount of datasets. Big data isn't just a lot of data; it's also constantly working on the technology required to make it run more quickly.

## **2.3 Comparative Analysis and Summary**

Now, a lot more research is being done on big data-related problems. This is the main goal of how individuals may work quickly and efficiently by learning technology. Since there is so much data now, it has become crucial to sort it out and use it for the benefit of the public. It should be simple for individuals to find what they're looking for online. Finding any information, such as doctors' names, product names, etc., appears to be simple for everyone. Furthermore, those showers also need to be developed because there is a lot of data.

## **2.4 Scope of the Problem**

Technology has already been successful with big data in several industrial applications. They had to cope with complicated data, including sensor-enabled robots that might use social media and mobile devices. In its own office, the Indian government has recently started working on big data analysis to better understand people. By performing big data analysis, we can find solutions to a wide range of such issues.

## **2.5 Challenges**

When discussing distributed file systems, cluster computing is necessary to implement our work for verification and analysis. It can be difficult for a university student to use AWS or Google to build a meaningful cluster computing system without a lot of funding. AWS or Google charge a lot of money for their service. Therefore, we chose to utilize the Databricks cluster created by the Apache Foundation. With some restrictions, it provides us with a genuine cluster system.

## **CHAPTER 03**

### **Research Methodology**

#### **3.1 Research Subject and Instrumentation**

Working without data is useless for improved explanations or visualizations. We will describe how the data was gathered for this study as well as the tools and techniques we employed to make the task move more quickly.

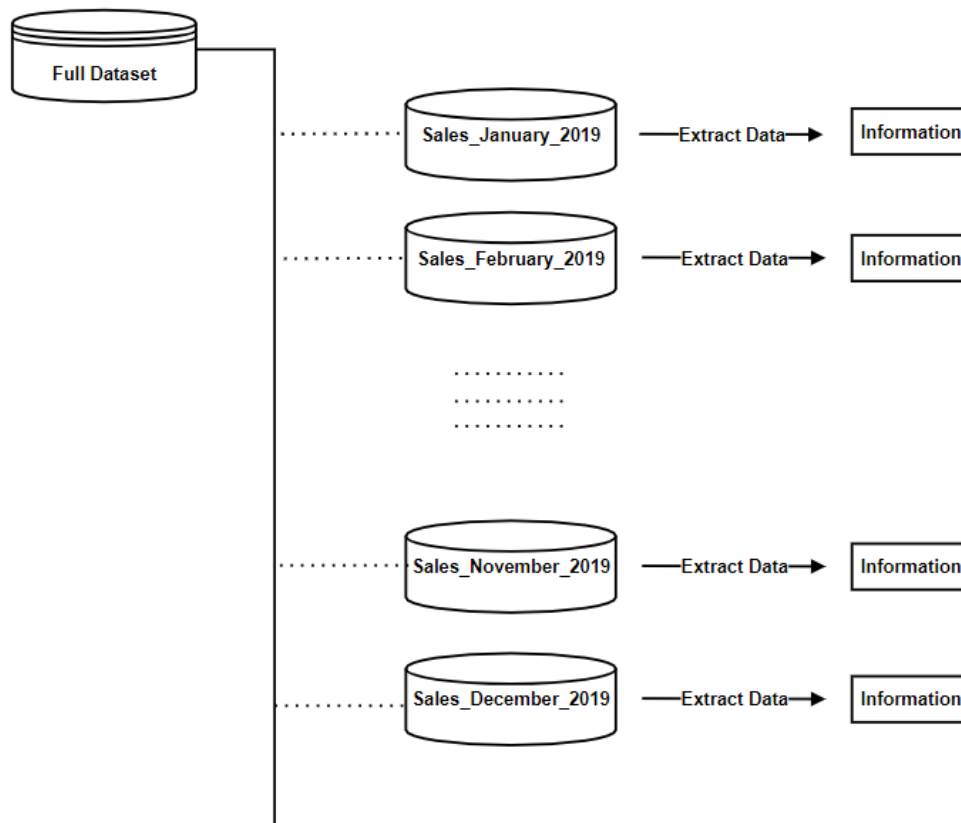
In order to examine several situations for our job, we selected an industry sales report for the entire year Sales in

**Table 1: Sales report**

<b>Month</b>	<b>Year</b>
January	2019
February	2019
.....	.....
.....	.....
November	2019
December	2019

The survey collected a total of twelve records.

The most vital part of our first step is reading the dataset in order to collect data and analyze it all. There were various procedures to take in order to work with the dataset. Figure 3.2 shows a diagrammatic explanation of how we read the dataset from our storage.



**Figure 3.1.1: The integration of dataset pipeline.**

Reading datasets to acquire data and analyze it all is the most crucial phase in our first process. There were various procedures to take in order to work with the dataset. Figure 3.1.1 shows a diagrammatic explanation of how we read the dataset from our storage.

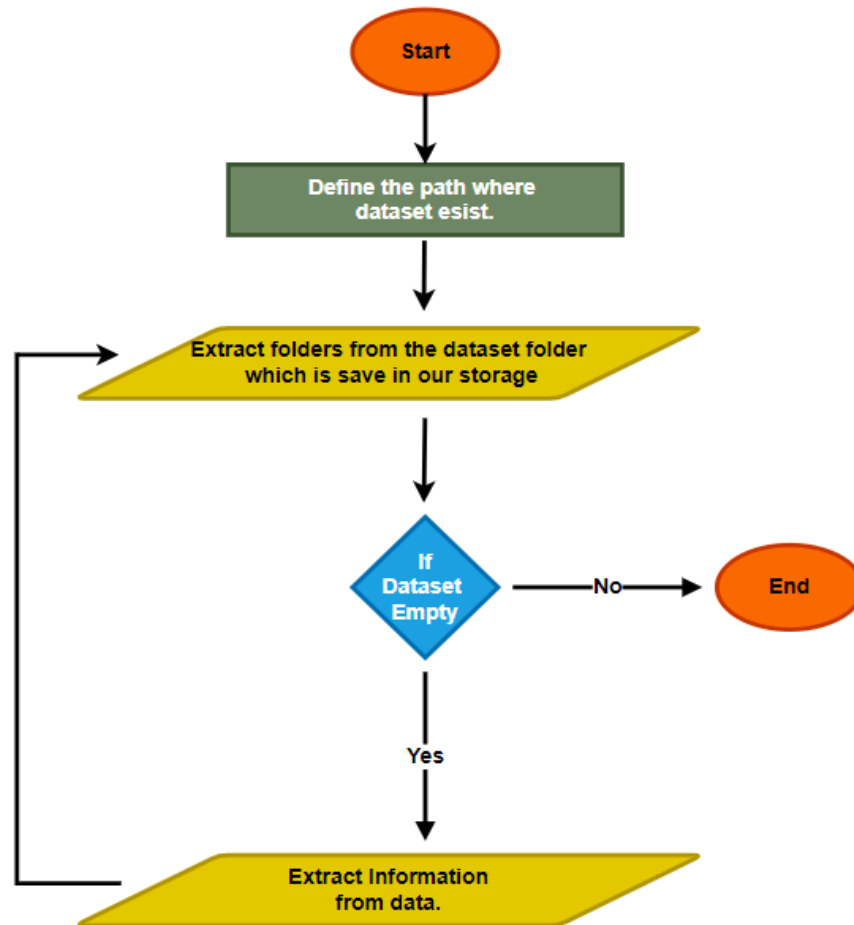


Figure 3.1.2: Accessing the flowchart of data.

We can see how data will arrive from our local storage in this flowchart for accessing data. We should specify the local storage path to access the dataset after starting the access. We extract the folder from the entire dataset once it has loaded completely on our PC. Next, we determine whether or not our dataset is empty. We repeat the procedure from the extracting folder if our dataset is empty. If it is not empty, the process is ended.

### 3.2 Data Collection Procedure/Dataset Utilized

We have opted to use qualitative methodologies in this study to gather some pertinent secondary data. Face-to-face data collecting is difficult, according to the current circumstances of the corvid-19. We choose to gather our dataset from a secondary source because of this. Secondary Data: In order to assure that I will correctly address the research problem, we think that certain secondary data is crucial. The following is a list of some informational sources: (1) The web and pertinent websites (2) Literature and News (3) Academic Papers

### 3.3 Statistical Analysis

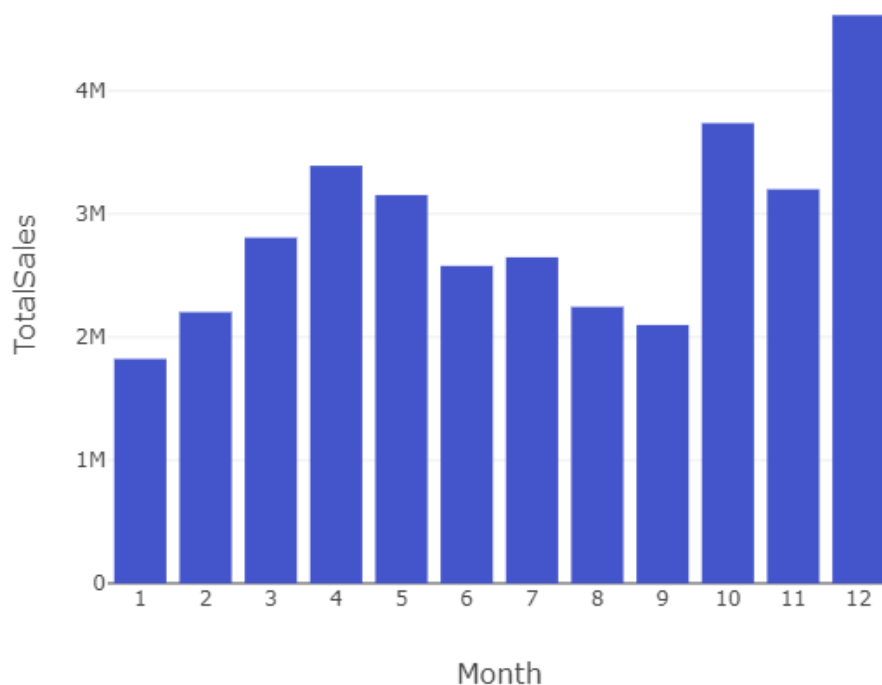
In this section, we go over the variety of our dataset and other data that is included in it. Through a chart will clarify and demonstrate here.

- Name of the month: **January**, Sell: 1822256.73, Product: USB-C Charging Cable12% & AAA Batteries (4-pack)11%, Total Order: 141234
- Name of the month: **February**, Sell: 2202022.42, Product: USB-C Charging Cable13% & Lightning Charging Cable12%, Total Order: 150502
- Name of the month: **March**, Sell: 2807100.38, Product: USB-C Charging Cable12% & Lightning Charging Cable12% , Total Order: 14551
- Name of the month: **April**, Sell: 3390670.24, Product: Lightning Charging Cable12% & USB-C Charging Cable11%, Total Order: 17539
- Name of the month: **May**, Sell: 3152606.75, Product: Lightning Charging Cable12% & AAA Batteries (4-pack)11%, Total Order: 15828
- Name of the month: **June**, Sell: 2577802.26, Product: Lightning Charging Cable11% & AA Batteries (4-pack)11%, Total Order: 12991
- Name of the month: **July**, Sell: 2647775.76, Product: Lightning Charging Cable12% & USB-C Charging Cable12%, Total Order: 13762
- Name of the month: **August**, Sell: 2244467.88, Product: AA Batteries (4-pack)11% & Lightning Charging Cable11%, Total Order: 11483
- Name of the month: **September**, Sell: 2097560.13, Product: USB-C Charging Cable12% & Lightning Charging Cable11%, Total Order: 248151

- Name of the month: **October**, Sell: 3736726.88, Product: USB-C Charging Cable12% & Lightning Charging Cable12%, Total Order: 19441
- Name of the month: **November**, Sell: 3199603.2, Product: USB-C Charging Cable12% & Lightning Charging Cable12%, Total Order: 278797
- Name of the month: **December**, Sell: 4613443.34, Product: USB-C Charging Cable12% & Lightning Charging Cable12%, Total Order: 24008

Here, we can observe that December had the largest product sales, with a total order of 24,008 and sales of USB-C and Lightning charging cables of 13% each. The lowest selling month order is 141234 for January. Additionally, 12% of USB-C Charging Cables and 11% of AAA Batteries (4-pack) were sold.

The figure 3.3 bar chart explains how to validate the dataset. Here, the total sales for the various months are seen in vivid detail. The month is shown on the X-axis, while the sale is shown on the Y-axis.



**Figure: 3.3.1 Statistical analysis and show with bar chart**

### **3.4 Proposed Methodology/Applied Mechanism**

Perform a descriptive study of this essay. The analysis tool we are utilizing is Apache Spark. And the technique we implemented through the python programming language. We are using various charts to display various results. Once the research is complete, we may discover the solution and understand clearly benefits how to improving company.

.

### **3.5 Implementation Requirements**

Using the phrase "Big Data Analysis" is used, be aware of Hadoop cluster. A Hadoop cluster is essentially the collection of nodes, which are groups of connected computers that work together to process large amounts of data in a parallel fashion.

We do this by using data bricks. With our dataset, Databricks Lakehouse is an easy-to-use, open, multi-cloud platform that makes it simple to establish clusters and conduct analytics. We established a cluster, which provides us with 1 Drive, 15.3 GB of Memory, 2 Cores, and 1 DBU for free.

We also made a notebook, to which we attached our cluster, and used it to examine our data. Scala and Spark 3.1.2 are already installed in the notebook. But for further in-depth analysis, we're using Python.

.

#### **Apache Spark 3.1.2**

The framework an open-source, multi-language engine for running and analyzing data engineering, data science, and machine learning on single-node workstations or clusters. It was initially created at the University of California Berkeley's AMPLab. The Spark codebase was later transferred to the Apache Software Foundation after its initial development, and they have continued to maintain it to this day.



### **Python 3.8**

Python is a well-liked high-level programming language today. Nearly every area of Python has contributed. Python makes it much simpler to construct applications for data science, machine learning, web development, and mobile devices. For the preparation of our sales data, Python is used. To do that, we need a pyspark library that can be used with both Python and Spark. We read the dataset, define the path to it, and then use Python to run some SQL queries.

## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Experimental Setup

We must pre-process our dataset after setting up the entire ecosystem and before performing the study. similar to deleting a null value, duplicating a value, and removing useless data. Verify that all of the spark python libraries have been correctly imported. This part will demonstrate how we conduct experiments and evaluate the dataset with clarity. Following the experiment, we go over the analyses' findings.

#### 4.2 Experimental Results & Analysis

The city with the highest volume of sales at the period is depicted below in figure 4.1. Austin has the lowest sales and San Francisco has the most. Aside from Dallas, sales in Atlanta and Seattle were nearly equal.

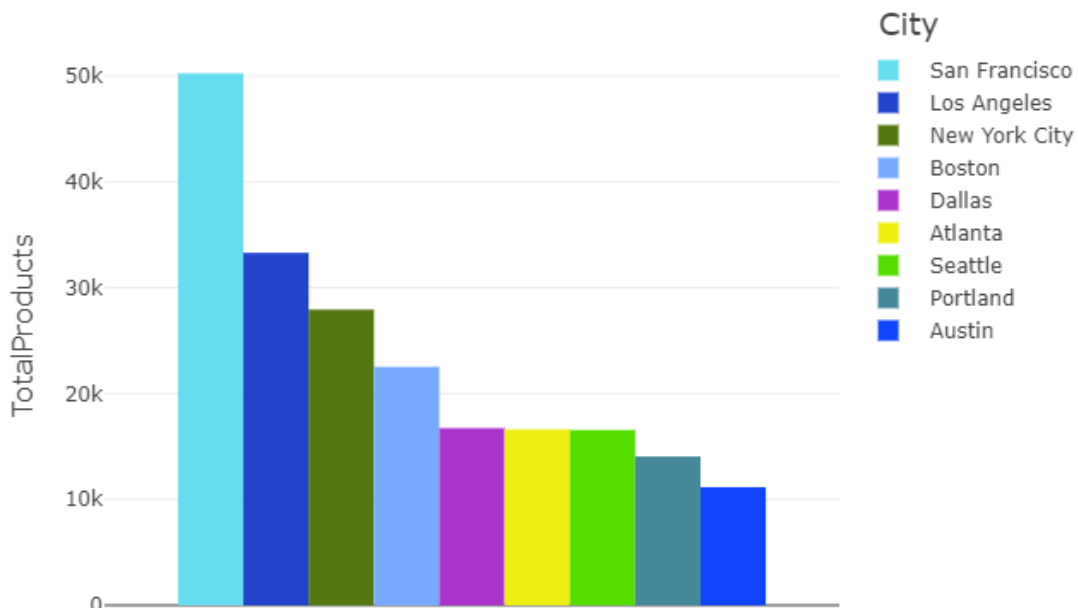
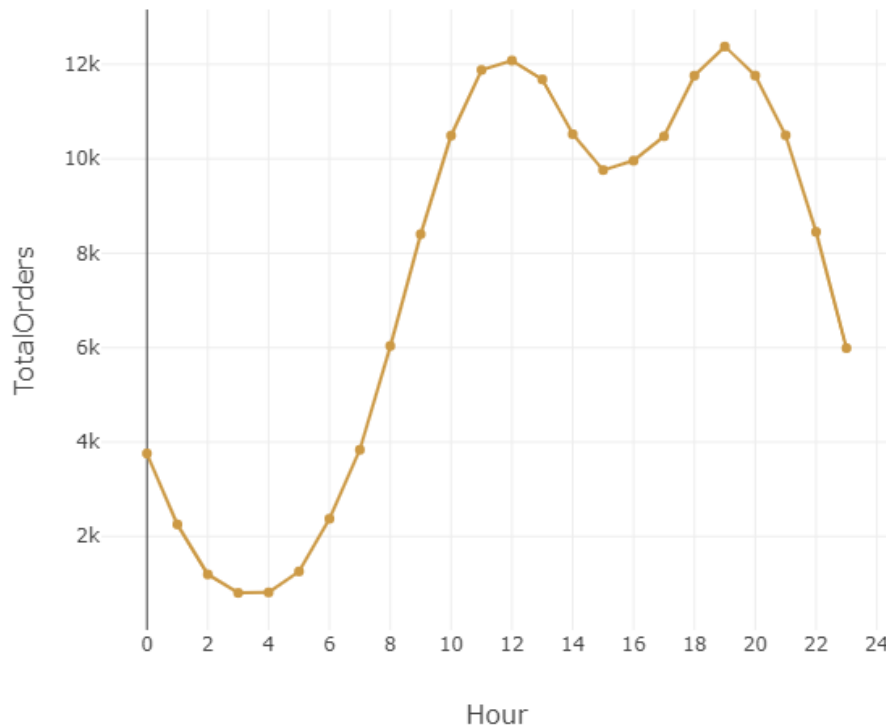


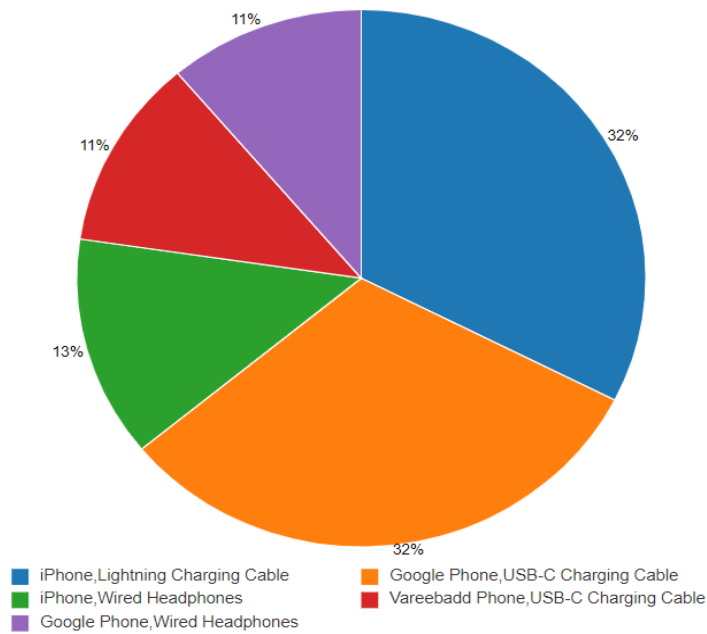
Figure 4.2.1: Product sell on the city basis

The hourly order ratio is displayed in a row. Hourly order ratio is displayed on the line. The order rate is highest between the hours of 10 a.m. and 10 p.m., and it decreases after that period. Additionally, Also, the late-night order rate is relatively low. From the pie chart below



**Figure 4.2.2 Order based picked hour**

We can observe from the pie chart below that the category of products are popular and have a large sales volume. The product categories with the most sales include the Google phone, USB-C Charging Cable, and iPhone, Lighting Charging Cable. These 2 product classes accounted for 32% of total sales. Then, 13% of all sales were made by the group of Wired Headphones. category had the lowest sales proportion.



**Figure 4.2.3 Product demand and total sell.**

Figure 4.4 shows the total sales per month. The previous month was the month with the highest sales and the first month was the month with the lower sales.



**Figure 4.2.4 Total sell calculation for monthly basis.**

The reason we are using Naive Algorithm is so that we can outsource our product predictions. We can use this algorithm to find out how many of the products will be sold and how many will not be sold.

The screenshot shows a web application window titled "Predictive Analysis". Inside, the heading "Probability of Product Return" is centered. Below it, a green status message reads "Model found. Ready for prediction." There is a horizontal input section with three dropdown menus: the first is set to "USB-C Charging Cable", the second to "December", and the third to "11.95". To the right of these is a "Show Prediction" button. Below the input fields, the results are displayed: "Probability of Product return:- 20.78%", "Probability of Product not returning:- 79.22%", and a final line stating "The Product will not be returned".

**Figure 4.2.5 Product sold of return Prediction using Naive Bayes**

### 4.3 Discussion

Throughout this entire section, we've been talking about the findings from various experiments applied to the dataset. After receiving the results, we displayed them all using figure.

## **CHAPTER 5**

### **Impact on Society, Environment and Sustainability**

#### **5.1 Impact on Society**

Big Data has an impact on businesses like Google, Facebook, Yahoo, Twitter, and other major corporations. The organization then has an immediate effect on our entire economy. Both society and the economy have an impact on technology. The entire procedure repeats itself in an infinite cycle. Manufacturing and the healthcare sector employ big data to raise their degree of industrial privacy, automation, security, and customer pleasure, among other things.

Big data is being used in so many different ways, including autonomous vehicles, smart homes, weather forecasting, natural disasters.

#### **5.2 Impact on Environment**

The biggest problem we currently face as a species and environment is climate change. Big data is assisting us in fully comprehending all of its intricate interconnections. Information on global warming also uses big data. Big data applied to reducing global warming is referred to as green data.

Different green data generation models exist throughout Europe. Copernicus is one of the role models. It is a satellite-based initiative for monitoring the planet. It is extremely effective in calculating, among other things, how rising temperatures affect river flows. Additionally, Copernicus is already supplying crucial data to improve the management of biodiversity, air quality, fishery, and agriculture as well as water resources.

Other multinational initiatives including Aqueduct, Global Forest Change, and Danger Map employ environmental data to slow down global warming.

**Aqueduct:** Conducting water quality and quantity analyses and providing the public with interactive risk maps.

**Global Forest Change:** Counting trees individually using high-resolution satellite photos over an extended period of time to determine the total amount of deforestation.

**Danger Map:** Using information from millions of citizens, identifying pollution.

### 5.3 Ethical Aspects

In our project communication plan, we have a declaration concerning ethical communication. The strategy can outline the ideals and principles our group has decided to uphold in regards to ethical communication.

- 1) Have to be in touch openly, truthfully, frankly, honestly finally honestly;
- 2) Have to be guarantee that all communications are timely and that all communications are short and sweet.
- 3) Have to ensure that all of our communications are succinct.
- 4) We promise to keep lines of communication open with all of our stakeholders

### 5.4 Sustainability Plan

For our project, sustainability planning is underway. Our project will be sustained in order to continue operating in the long run in this way. Here, we outline the long-term viability of our proposal. By doing this, it ensures that resources used for the project are not lost in the future.

Three sections typically make up a sustainability plan. The three components of our project that we should adhere to are: (A) Community Sustainability; (B) Financial Sustainability; and (C) Organizational Sustainability.

First off, carefully involving beneficiaries in planning and implementation from the start is an excellent method to guarantee Community sustainability. To ensure that the community feels invested in the project and that their preferences are implemented with our project, we consult with all stakeholders whenever feasible.

Second, depending on the type of project we intend to accomplish, financial sustainability will appear very differently. Applying for a one-time project that is expected to have nearly no subsequent expenditures requires far less thought than applying for a project with ongoing expenses that must be paid over time in order to be financially sustainable.

Third, organizational sustainability refers to our company's ability to survive as a whole. Whether it's through internal resources like income-generating activities or membership fees or external ones like grants and long-term support, we wish to accomplish this. It is crucial to demonstrate to the donor that we are an excellent partner, so this component of the sustainability strategy shouldn't be overlooked. things that they can take into account going forward.



## **CHAPTER 6**

### **Summary, Conclusion, Recommendation and Implication for Future Research**

#### **6.1 Summary of the Study**

When using Hadoop MapReduce for massive data processing, there are various difficulties that arise. issues such an excessive amount of network data, OLTP appropriateness issues, iterative execution issues, and so forth. We are talking about a method to effectively deal with those issues. We used an actual data set in order to improve visualization. having 0.31 percent or less of our entire dataset be null and bad. Following preprocessing, we examine a few questions to determine the response from our data set, which is displayed.

the main purpose of Apache Spark is to solve the challenges of real-time processing of large amounts of data with MapReduce. All large organizations, businesses, and e-commerce need to leverage data for commercial purposes in order to improve consumer satisfaction. Every large organization, firm, and e-commerce must evaluate its data for its commercial purposes in order to improve consumer happiness. Our study is based on real-time using spark. so that we don't have to wait and may receive our results immediately.

#### **6.2 Conclusions**

We investigated the Spark ecosystem before using Spark and the Python programming language to examine big data. Using spark analysis, several questions' answers were defined. Here, we present some statistical diagrams illustrating the top-selling month, peak advertising period, and product category. Sales of the top-selling product are 32% and those of the lowest-selling product group are 11%. According to our projection, the ideal times for advertising are from 10:00 am to 12:30 pm and from 6:00 pm to 8:00 pm. E-commerce can assist any firm by creating a stable framework by expanding the data size.

## **6.3 Implication for Further Study**

Our next objective is to change how we are currently improving our work. As it is very beneficial for the area in many ways, including helping large firms, e-commerce, the banking system, government projects, the analyst sector, business analysts, as well as regular individuals to forecast and evaluate their information. Therefore, we will endeavor to enhance this work to meet all needs associated with our project and study. Therefore, we have further plans to enhance our work for upcoming work on this project.

- Create a straightforward user interface (UI) system that can easily take information and provide a basic analysis based on our needs.
- We want to use the APIs of many organizations to obtain their real-time open-source data.
- It's crucial to spot fraud right away, before the offender commits an offense like online banking fraud or social media fraud, for example.
- Our next focused endeavor is accuracy and obtaining results immediately.
- Since we are working with large data, we can refer to it as such if the size of our data exceeds a certain threshold, such as billions or trillions. Therefore, building a real cluster with numerous computers and working with real huge data is our next goal.

## **APPENDIX**

### **Creating cluster, maintaining cluster and pipeline integration.**

The biggest problem we had while working on this project was setting up and maintaining clusters with pipelining. Databricks, which is maintained by the Apache Foundation, offers a complete pipeline integration, cluster creation, and maintenance procedure. We have inexpensive access to all of the processes thanks to Databricks. We made use of a free Databricks capability. We are unable to integrate many functionalities in our project as a result. After much time, effort, and commitment, we finally succeed in achieving our goals.

## **References:**

### ***Conference/Journal Papers:***

- [1] Li, J.; Zeng D., Big data meet green challenges: Greening big data. IEEE Syst., vol. 10, pp. 873–887 January 2016.
- [2] Nicolas Poggi; Josep Ll. Berral; Thomas Fenech, and David Carrera, The state of SQL-On-Hadoop in the Cloud, European Research Council (ERC), 2020
- [3] Bijesh Dhyani; Anurag Barthwal, Big Data Analytics using Hadoop, International Journal of Computer Applications (0975 – 8887), Vol 108, December 2014
- [4] R. S. Sandeep, C. Vinay, S. M. Hemant, Strength and Accuracy Analysis of Affix Removal Stemming Algorithms, International Journal of Computer Science and Information Technologies, vol. 4, no. 2, pp. 265-269, April 2013.
- [5] Jasmine Zakir; Tom Seymour; Kristi Berg, BIG DATA ANALYTICS, International Association for Computer Information Systems, Vol. 16, pp. 81-90, 2015
- [6] Azeroual, O.; Fabre, R. Processing Big Data with Apache Hadoop in the Current Challenging Era of COVID-19. Big Data Cogn. International Association for Computer Information Systems, March 2021
- [7] Shaina; Dr. Sushil Kumar, Big Data Analytics using Apache Hadoop, Turkish Journal of Computer and Mathematics Education, Vol. 12, 10 May 2021
- [8] K. Manasa; Md. Asim, Big Data Analytics: Predicting Academic Course Preference Using HADOOP Inspired Map Reduce, International Journal of Research (IJR), Vol. 8, June 2021
- [9] T.K. Das; P. Mohan Kumar, BIG Data Analytics: A Framework for Unstructured Data Analysis, International Journal of Engineering and Technology (IJET), Vol. 5, March 2013
- [10] Jorge L. Reyes-Ortiz; Luca Oneto; Davide Anguita, Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf, INNS Conference, Vol. 53, pp. 121–130, 2015,
- [11] Ishawarappa; Anuradha, A Brief Introduction on Big Data 5Vs Characteristic and Hadoop Technology, International Conference on Computers Communications and Control (ICCCC), pp. 319-324, 2015

### ***Books:***

- [12] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to Algorithms, 3rd Edition, The MIT Press, 2009, pp. 120-122.
- [13] Vignesh Prajapati, Big Data Analytics with R and Hadoop, 1st Edition, Packt Publishing Ltd, 2013, pp. 149-173
- [14] Ragib Hasan, Gobeshonay Hatekhori, Adarsha Prokashak, 3rd, 2020, pp. 22-80

**Websites:**

- [15] Learn about Weather Data Analysis using Hadoop to Mitigate Event Planning Disasters, available at << <https://scholarworks.bridgeport.edu/xmlui/handle/123456789/1105>>>, last accessed on 24-03-2022 at 10:31 AM.
- [16] Learn about Cluster in big data, available at << <https://www.quora.com/What-is-cluster-in-big-data>>>, last accessed on 24-03-2022 at 10:13 AM.
- [17] Learn about Hadoop Challenges, available at << <https://www.researchgate.net>>>, last accessed on 10-02-2022 at 19:37.
- [18] Learn about Different Dataset, available at << <https://www.scholar.google.com>>>, last accessed on 10-02-2022 at 19:37.
- [19] Learn about Different Dataset, available at << <https://www.scholar.google.com>>>, last accessed on 10-02-2022 at 19:07.
- [20] Learn about Hands-On Big Data Analysis with Hadoop 3, available at << <https://www.packtpub.com>>>, last accessed on 02-02-2022 at 12:07 AM.
- [21] Learn about Introduction to Big Data Analysis & Machine Learning in Python with PySpark, available at << <https://www.crowdcast.io>>>, last accessed on 01-02-2022 at 23:58.
- [22] Learn about BigQuery: Cloud Data Warehouse, available at << <https://www.cloud.google.com>>>, last accessed on 01-02-2022 at 12:47 AM.

## pre defense 2

### ORIGINALITY REPORT

<b>26%</b>	<b>20%</b>	<b>6%</b>	<b>24%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>21%</b>
<b>2</b>	<b>dspace.daffodilvarsity.edu.bd:8080</b> Internet Source	<b>3%</b>
<b>3</b>	<b>www.rroij.com</b> Internet Source	<b>&lt;1%</b>
<b>4</b>	<b>www.thunderheadalliance.org</b> Internet Source	<b>&lt;1%</b>
<b>5</b>	<b>Submitted to University of Zagreb - Faculty of Economics</b> Student Paper	<b>&lt;1%</b>
<b>6</b>	<b>www.unilibre.edu.co</b> Internet Source	<b>&lt;1%</b>
<b>7</b>	<b>Kari Venkatram, Mary A. Geetha. "Review on Big Data &amp; Analytics – Concepts, Philosophy, Process and Applications", Cybernetics and Information Technologies, 2017</b> Publication	<b>&lt;1%</b>
<b>8</b>	<b>www.analyticsvidhya.com</b> Internet Source	<b>&lt;1%</b>