**COURSEWORK**

Table of Contents

[Introduction 4](#_Toc101315297)

[Part A.1 4](#_Toc101315298)

[Characteristics 4](#_Toc101315299)

[Task A.2 5](#_Toc101315300)

[Comparison Hadoop vs Relational Database Systems. 5](#_Toc101315301)

[Task B.1 6](#_Toc101315302)

[Task B.2 9](#_Toc101315303)

[Part C: Big Data Project Analysis 10](#_Toc101315304)

[Task C.1: 10](#_Toc101315305)

[Task C.2: 10](#_Toc101315306)

[Task c.3 11](#_Toc101315307)

[Task c.4 11](#_Toc101315308)

[Reference: 13](#_Toc101315309)

**Table OF Figures**

[**Figure 1 Student Class 6**](#_Toc101315570)

[**Figure 2 main class 7**](#_Toc101315571)

[**Figure 3 Reducer Class 8**](#_Toc101315572)

[**Figure 4 Mapper Class 8**](#_Toc101315573)

[**Figure 5 result 9**](#_Toc101315574)

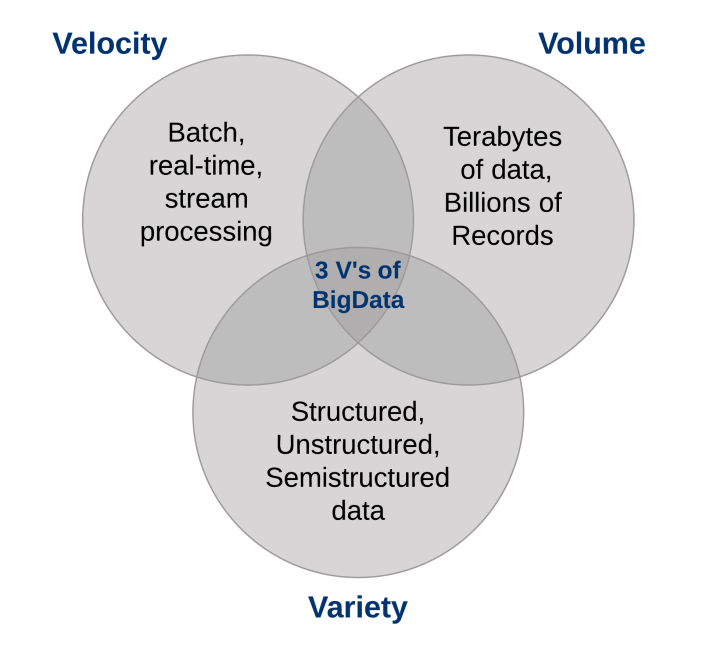
# Introduction

Big Data is referred to large amount of data which is being gathered from different resources and stored for further processing. The data which is being gathered will go through different process to extract suitable information from it. There are different types of big data which are structured, unstructured and semi-structured.

# Part A.1

## Characteristics

There are primarily three characteristics of big data which are velocity, volume and variety. But it is further which are as follows.



1. Velocity

Velocity is how efficiently and fast the data is being processed. It is important because the huge amount of data needs high speed of processing power to perform tasks on it otherwise it will take ages to perform any task on it.

1. Volume

As name says the volume of data refers large amount of data which might be in Gigabytes or Yottabytes depends on the industry to industry. But this data doesn’t stop at any point in big data the data is continuously gathered from different resources.

1. Value

Value of a data is depending on how much a data is beneficial for the organisation. Does the data stand on the goals the organisations want to achieve?

1. Variety

There is different type of data which is collected from different resources such as social media, Weather forecast or written records. It is a challenge in big data to get and store these types of data which effects the performance because of different types and resources.

1. Veracity

Veracity is the accuracy of data which is a curtail character of big data. Low veracity can damage the accuracy of the result.

1. Validity

Validity defines how authentic a data is for the specific purpose.

1. Volatility

Volatility in big data is the data is continuously changing. The data received previously is different from received today.

1. Visualization

Visualization is the graphical representation of data which helps to share insights with non-technical members.

# Task A.2

## Comparison Hadoop vs Relational Database Systems.

The basic comparison between RDBMS and Hadoop is that the RDBMS can only store structured data whereas the Hadoop can store three types of data that could be structured, un-structured and semi-structured. Hadoop has two main components first is HDFS and MapReduce. HDFS is a Distributed File System which stores data in a distributed manner and MapReduce is there for processing data. Hadoop Stores data in a specific file size which can be extended then these files will store in a cluster of computers which store that file in a distributed manner. RDBMS stores data in a table format that could contains relations structure between tables using rows and columns. (Aditi Malhotra. 2018).

Hadoop doesn’t support Real time data processing (OTLP). It supports large scale batch processing (OLAP) which is mainly used for data mining. Hadoop can run complex queries but the time to run the queries depends on the size of data. On the other hand, RDBMS supports Real time data processing due to highly normalised data (Aditi Malhotra. 2018).

**Scenario:**

Business Analyses:

Hadoop plays an important role in the commercial industry it provides a batter understanding of the market by looking the sales and reviews of the products. It helps the company to improve the performance of the company. In retails company can manage their stock by predicting the products in high demand. These predictions can be conducted after processing the data received from different resources which are social media, google searches and geographical positioning. Hadoop helps to understand where the specific product is needed at what time at a specific location. The big stores like Walmart uses the big data for the prediction of the product which boosts their sales by understanding the need of the customers.

# Task B.1

I have attached couple of screen shots of the code I build for the MapReduce the code is being implemented using java, screenshots are as follows.

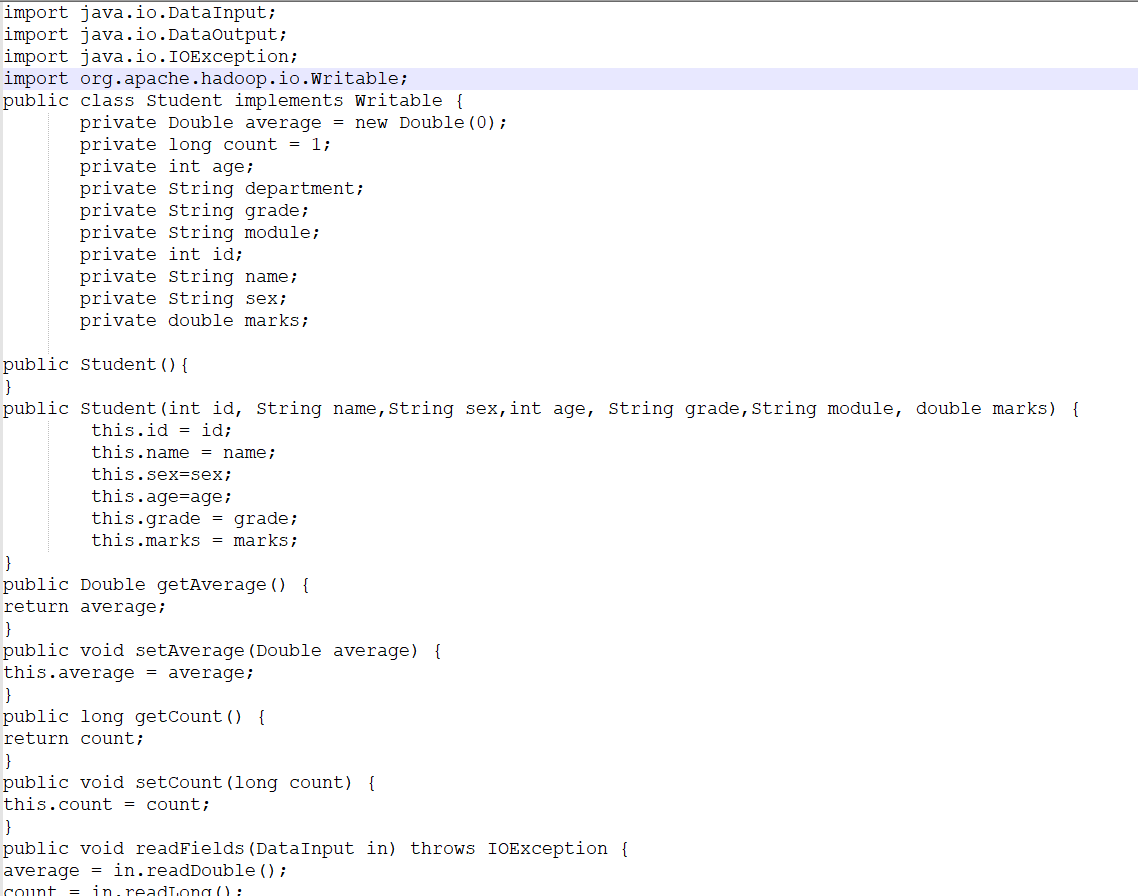
****

Figure 1 Student Class

****

Figure 2 main class

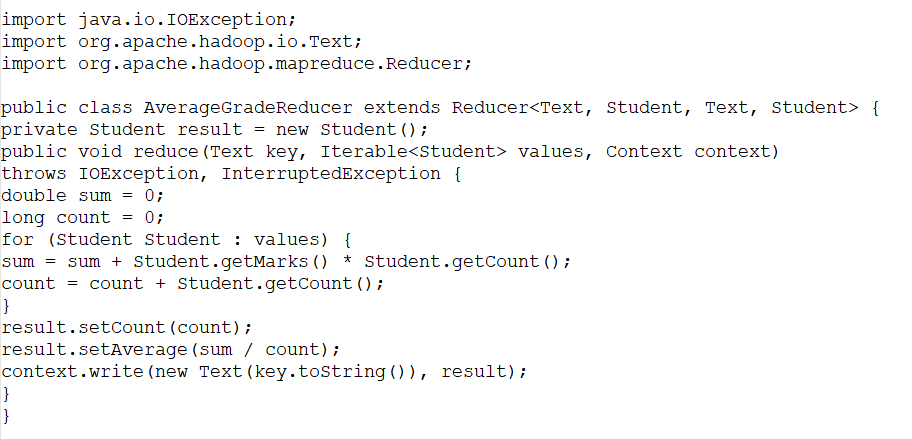
****

Figure 3 Reducer Class

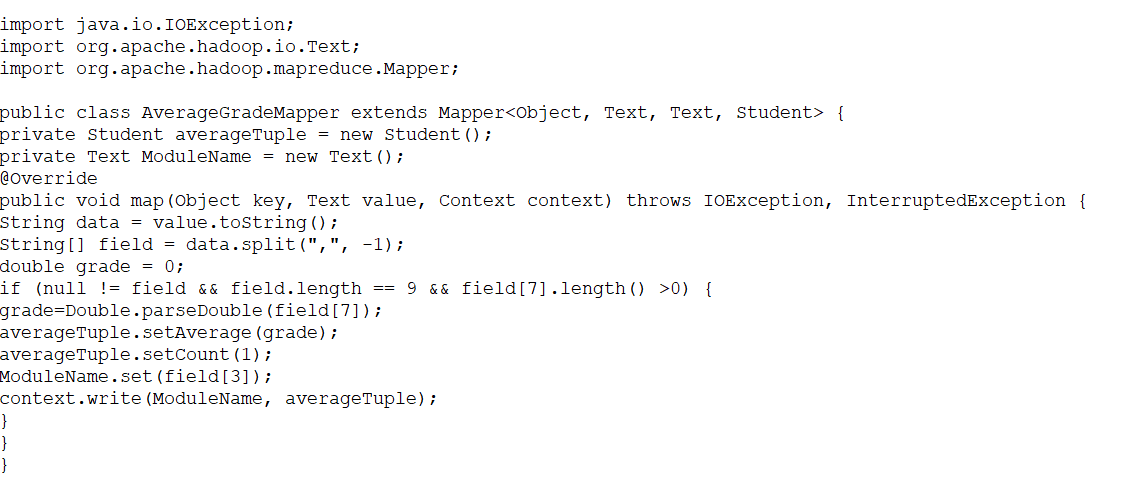
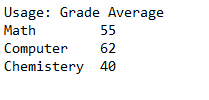
****

Figure 4 Mapper Class

****

**Figure 5 result**

# Task B.2

In the above figure 2 and figure 3 I have shown the code for the MapReduce which is being implemented in java. As mentioned in above figures I have implemented student class which is a writeable object. This class will be used to output the values from the mapper class. In the mapper class the data is being separated by defining comma as delimiter. Marks of the students were stored on the 7th index of the array and we are counting in the second column will be used to count for further use in reducer class. The AverageGradeReducer class extends from the reducer and iterate through all the students After iterating on first student is shift its iteration to second student as we are using extended for loop for the iteration. This loop manages two value sum and count sum will add all the marks we get from the student object and count will count the number of students. This data will be used after to get an average using sum divided by count.

Finally, as we can see in the above figure, we have GetAverage class which contains main function. This function is used to combine the result of average to the operators associated to it.

# 

# Part C: Big Data Project Analysis

## Task C.1:

Before choosing the methodology, we need to look at the differences between data lake and data warehouse. Data warehouse stores processed data in structured form for example it processes data before storing and store a processes data inti. Whereas data lake stores Raw / unstructured data which will be processed after so we can say they stored data is in a raw form. As per the given scenario the company cropY is going to store 500 petabytes of data which needs a huge storage and high processing power. There are couple of factors considered during the storage of data which are as follows.

Processing:

Data warehouse stores structures data so it need to be in a specific format before storing on the other hand data lake stores data in raw form it stores data as it is received and then process it at the time of need so the speed of data processing will depend on the size of data.

Cost:

It is one of the main aspects for storing big data. It needs to store hundreds of gigabytes of data so the cost of storing suppose to be high but as compared to data lake and data warehouse, data lake is a big data technology which is an open-source platform so the cost of software licencing is free and it installs on low-cost commodity hardware. Storing data on data warehouse can be much costly depending on size and demand.

Conclusion:

As per the above scenario the company cropY should use the data lake technology because it needs to store huge amount of data which is being received from different resources which are various sensors, drones, social media, online data etc. so if we go for data warehouse, we need to process it and put it in a specific format before storing so it which will cost a lot of resources to do that and it will not be efficient for gathering the data. So, it is better to store it first using data lake technology and then process it after storing which saves a lot of time in collection of data.

## Task C.2:

The CropY company has a large amount of data collected of plants, crops, diseases, pests, etc. this data is being collected from different sources so we used big data technology to store it. For processing and querying on it. It is suggested to use Apache spark an open-source technology which uses for big data workloads. It operates on in memory caching for optimising the query to run fast on any size of data. Hadoop MapReduce works on a programming model for processing on big data sets which is a distributed parallel algorithm. MapReduce reads data from the cluster of computers and then perform operation on it and after writes the results back on it. Which leads slow process due to latency of disk input/output (AWS. 2022).

Apache Spark is an opensource platform works on interactive queries, machine learning and real time workloads. It doesn’t have its own storage system so it runs analytic operations on other storage system like Hadoop HDFS storage system which is a distributed file system. Through its in memory caching technique it optimises the query to execute and run fast analytic query against any size of data (AWS. 2022).

Apache spark has the ability to run multiple workloads, which includes interactive queries, machine learning, graph processing and real time analytics. Single application can association with multiple workloads (AWS. 2022).

So, it is recommended to company CropY to use Apache spark technology which is efficient for working on bigdata due to its multiple workload’s strategy and optimizing query it will be suitable for the company.

## Task c.3

The company CropY provides predictions and analytic services for some tasks in a real time or we can say near real time in a few seconds after the arrival of new data from the resources. It is being suggested by some IT members to use MapReduce framework to implement and achieve the real time prediction. First, I will define the MapReduce and then suggest my answer.

MapReduce is a batch processing technique in Hadoop. Batch Processing collects data and store it on disk and then process it. It is efficient to process large amount of data. Hadoop then use MapReduce to process the data in Batch jobs. Batch jobs run automatically without any manual involvement. The time to run the task depends on the size of the data and we can set it to run every hour to run. But as per our requirement we need real time prediction for the data continuously receiving from the resources which could be receiving from different sensors so in batch it will store first and then after it will send to the batch jobs for processing using MapReduce.

Whereas, as per the requirement it is suggested to use Stream processing to achieve the real time prediction. Spark stream watch the data as it comes and respond on it in milliseconds after processing. It is process data hundred times faster than the MapReduce. Spark stream can also handle large file sometime data received is in a huge file and it is not possible to store it so the spark stream handles it and after process returns the result in an important bit. So, for conclusion the above discussion gives a clear idea about the scenario. I don’t agree with the suggested MapReduce framework. We should go for Stream processing rather than the batch processing (Abhishek Ghosh May 25, 2019).

## Task c.4

Businesses these days are moving to cloud computing because of several reasons. The main of all is to attain top enterprise level technology without taking worry of handling and managing the hardware by ourself which always needs maintenance to keep it working. Cloud makes the customer free from that hurdle and allow them to focus on their product rather then how to manage it.

Cloud also helps to scale the product especially based on big data which need huge number of storage space we can manage the resources according to the need of the system. Which also reduces cost by decreasing the resources when needed less. Cloud also guaranties high availability of the product through out the world through its different geographical availability and its highly efficient and fast servers which are continuously being managed by the cloud providers. Our strategy contains few steps which are as follows:

1. Lift and shift:

In the step we move our whole system to the cloud. It will be little expensive at start but when we see the out come the number of people reach the system and we save time and money in maintaining the systems which is not our headache any more.

2. Iterate rapidly:

Once we move the system to cloud, we need to focus on the system iteratively. We need to keep looking which feature cost us more resources and how we can reduce the cost by shifting services to other providers or managing the resources not in use.

3. Adopt elasticity and flexibility for system:

Once we shifted the system to cloud and started making changes iteratively, we need to scale the resources dynamically, increase and decrease according to customer needs.

For security we use encryption while transferring the data which secures the system as the servers we use are private which also provides the security for the system.

# Reference:

**Aditi Malhotra. 2018.** “*Comparing SQL Databases and Hadoop*”, Available at [Comparing SQL Databases and Hadoop - Whizlabs Blog](https://www.whizlabs.com/blog/hadoop-vs-sql-database/#:~:text=Hadoop%20vs%20SQL%20database%20%E2%80%93%20of%20course%2C%20Hadoop,works%20better%20on%20low%20volume%20of%20data%20%28Gigabytes%29.) **[Accessed 16 April 2022]**

**AWS. 2022**. “*Introduction to Apache Spark*”, **Available at**: [What is Apache Spark? | Introduction to Apache Spark and Analytics | AWS (amazon.com)](https://aws.amazon.com/big-data/what-is-spark/) [**Accessed 16 April 2022**].

**ABHISHEK GHOSH MAY 25, 2019**. “*Differences Between Batch Processing and Stream Processing*”, Available at: [Differences Between Batch Processing and Stream Processing (thecustomizewindows.com)](https://thecustomizewindows.com/2019/05/differences-between-batch-processing-and-stream-processing/) [**Accessed 16 April 2022**].