**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Storage Solutions for big data |
| --- | --- |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Kavi Patak |
| **Student Number:** | sba22391 |
| **Assessment Due Date:** | 26 May 2024 |
| **Date of Submission:** | 26 May 2024 |

****

**Declaration**

| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |
| --- |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Storage Solutions for Big Data:

Continuous Assessment 2

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

by Kavi Patak

sba22391

Storage Solutions for Big data

Lecturer: Dr. Muhammad Iqbal

CCT College, Dublin

May 26, 2024

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Table of Contents

[**Storage Solutions for Big Data: 2**](#_3iddv1ojtzjk)

[**Table of Contents 3**](#_ytyrg5mbj69k)

[Question 1: 4](#_usiluud1xtj5)

[Define Big Data and outline its key characteristics. Discuss the potential for banks to increase profits through big data processing and analysis. Identify three businesses that have successfully leveraged big data storage solutions in recent times. 4](#_l90b5kk1iqyh)

[Question 2: 4](#_6v3q9s8f4v4a)

[Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes 4](#_9r0rujli3652)

[a) Start all five processes of hadoop distributed file system 4](#_wpiu62odcn4e)

[b) Find the text dataset of size (500 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data. 4](#_o57biulv9hxn)

[c) Execute Mapper and Reducer programmes by using loaded input file. 4](#_hkzik1kyrpsj)

[d) Display the frequency of each word obtained from the dataset 4](#_gb9yzkj45w4b)

[e) Download the output file from hadoop and upload it on Moodle. If you could not provide the screenshots for all commands or related files, no marks will be awarded. 4](#_bcug5tipbkkn)

[**References 6**](#_kya26suve4eo)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

### Question 1:

#### Define Big Data and outline its key characteristics. Discuss the potential for banks to increase profits through big data processing and analysis. Identify three businesses that have successfully leveraged big data storage solutions in recent times.

Big Data may be defined in many ways, depending on who you ask. However, most definitions of big data reflect the growing technological ability to capture, aggregate, and process an ever increasing volume, velocity, and variety of data (Iqbal, n.d.). Volume, Velocity and Variety, have become known as the 3 V’s of Big Data. Big Data comes in a variety of forms including pictures, video, audio, and social media, and hence a variety of formats, whether structured, semi structured, or unstructured. Volume and variety refer to data that is too large or complex for storage and analysis by traditional methods. Since the advent of the Internet and proceeding developments in technologies like the Internet of Things(IoT), Sensors, GPS devices, Machine Learning (ML) and Artificial Intelligence (AI), the volume and variety of data being produced is on an upward and seemingly unbound trajectory (Agnellutti, 2014). This data is being produced at an exponential rate, with velocity referring to this, and the requirement for its instant processing for immediate use and inferred insight. Other interpretations of Big Data include the addition of veracity and value, with veracity referring to the messiness and trustworthiness of data, while value refers to the potential insight or gain, hidden within the data (Yaqoob et al., 2016).

### Question 2:

#### Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes

#### a) Start all five processes of hadoop distributed file system

#### b) Find the text dataset of size (500 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data.

#### c) Execute Mapper and Reducer programmes by using loaded input file.

#### d) Display the frequency of each word obtained from the dataset

#### e) Download the output file from hadoop and upload it on Moodle. If you could not provide the screenshots for all commands or related files, no marks will be awarded.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# References

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**