**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Advanced Data Analytics*  *Big Data Storage and Processing* |
| **Assessment Title:** | *MSC\_DA\_BD\_ADAv5* |
| **Lecturer Name:** | *David McQuaid*  *Muhammad Iqbal* |
| **Student Full Name:** | Jose Luis Alvarado Carranza |
| **Student Number:** | 2020092 |
| **Assessment Due Date:** | 08/11/2023 |
| **Date of Submission:** | 08/11/2023 |

**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. |

ABSTRACT

INTRODUCTION

I have uploaded the Project Tweets .csv into both of the technologies used in the course. Vis a vie Hbase, SparkSQL. And I have made some comparisons in my head on the performance of each off them. I think that for the task is it more convenient to use Spark SQL than Hbase because it is visually easier to comprehend and one doesn not have to be switching between terminals to eeecute the commands.

Part I

To begin this project, we have to upload the ‘ProjectTweets.csv’ file to Hadoop. To do this, we have created a folder called ‘ca2’. There, we store the csv file.

A close up of numbers

Description automatically generated

Figure : File uploaded to hadoop

Then we have used Pyspark to manipulate the data and have store it in SparkSQL and Hbase. These to technologies are the ones that we have decided to work with. This means that we have called the data from Hadoop and stored it in the aforementioned technologies. From there we will manipulate the data to continue our assessment.

A screenshot of a computer

Description automatically generated

Figure : File read from hadoop

Spark’s shell provides a simple way to analyze data interactively.

A computer screen shot of a computer code

Description automatically generated

Figure : Upload files to Hbase

A screenshot of a computer program

Description automatically generated

Figure : Read files from Hbase

A screenshot of a computer screen

Description automatically generated

Figure : Table creation

A computer screen shot of a program

Description automatically generated

Figure :Data read from Hbase Shell

A computer screen shot of white text

Description automatically generated

Figure : YCSB for Hbase