CCT College Dublin Continuous Assessment

Programme Title:	HDip AI Concepts - Feb 2023 - SB+HCI cohort		
Cohort:	FT/ PT		
Module Title(s):	Storage Solutions for Big Data (10 ECTS)		
Assignment Type:	Individual	**Weighting(s)**:	50%
Assignment Title:	CA1		
Lecturer(s):	Dr. Muhammad Iqbal		
Issue Date:	7th March 2024		
Submission Deadline Date:	28th April 2024		
Late Submission Penalty:	Late submissions will be accepted up to **5** calendar days after the deadline. All late submissions are subject to a penalty of **10%** of the mark awarded. Submissions received more than 5 calendar days after the deadline above **will not** be accepted and a mark of 0% will be awarded.		
Method of Submission:	**Moodle**		
Instructions for Submission:	Upload one or more files composed of word file, Jupyter notebook, python files, dataset and any supporting information.		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	3 weeks after submission date		

Learning Outcomes:
Please note this is not the assessment task. The task to be completed is detailed on the next page.
This CA will assess student attainment of the following minimum intended learning outcomes:

MLO2: Critically evaluate and utilise appropriate distributed storage and processing architectures for defined business problems.
(Linked to PLO 3, PLO 4)
MLO3: Design and create programming solutions utilising relevant design patterns to extract information from large-scale data.
(Linked to PLO 2, PLO 3)
MLO4: Utilise an available big data solution to update / scale a given legacy solution.
(Linked to PLO 4)

Attainment of the learning outcomes is the minimum requirement to achieve a Pass mark (40%). Higher marks are awarded where there is evidence of achievement beyond this, in accordance with QQI *Assessment and Standards, Revised 2013*, and summarised in the following table:

Percentage Range	CCT Performance Description	QQI Description of Attainment	
		Level 6, 7 & 8 awards	Level 9 awards

90% +	Exceptional	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this
80 – 89%	Outstanding		
70 – 79%	Excellent		
60 – 69%	Very Good	Achievement includes that required for a Pass and in **many** respects is significantly beyond this	Achievement includes that required for a Pass and in **many** respects is significantly beyond this
50 – 59%	Good	Achievement includes that required for a Pass and in **some** respects is significantly beyond this	Attains all the minimum intended programme learning outcomes
40 – 49%	Acceptable	Attains all the minimum intended programme learning outcomes	
35 – 39%	Fail	Nearly (but not quite) attains the relevant minimum intended learning outcomes	Nearly (but not quite) attains the relevant minimum intended learning outcomes
0 – 34%	Fail	Does not attain some or all of the minimum intended learning outcomes	Does not attain some or all of the minimum intended learning outcomes

Please review the CCT Grade Descriptor available on the module Moodle page for a detailed description of the standard of work required for each grade band.

The grading system in CCT is the QQI percentage grading system and is in common use in higher education institutions in Ireland. The pass mark and thresholds for different grade bands may be different from what you have experience in the higher education system in other countries. CCT grades must be considered in the context of the grading system in Irish higher education and not assumed to represent the same standard the percentage grade reflects when awarded in an international context.

Assessment Task

In this CA, you are required to identify a large dataset and perform an analysis using a big data framework for the storage and use an appropriate programming language for the processing.

CA1 Specification:

The following elements should be part of the CA1 as

a) Obtain a large dataset from any online public repositories (For example, data.gov.ie, Kaggle and the datasets available at https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html), while the size of the dataset should justify the complexity level (Structured or Semi-structured or Unstructured). Source data sets can be static (file or database) or use an API to retrieve the data.
b) Utilisation of a distributed data processing environment for some part of the analysis (e.g., Apache Hadoop Map-reduce/ Apache Spark) and storage of data on the hadoop distributed platform.
c) Use any NoSQL database for the storage and processing of queries based on the chosen dataset and objectives of the CA1.
d) Programmatically accessing the source data from the chosen NoSQL database using relevant MapReduce / Spark code.

Scenario:

- Initially, the data can be stored into Hadoop/ MySQL/ NoSQL storage. Hadoop MapReduce/ Spark processing would utilize MySQL/ NoSQL database as an input. After processing the data through Mapreduce/ Spark, you can store it into a Hadoop or NoSQL database.

- You can use Python/NumPy/Pandas/Matplotlib to conduct further analysis of the MapReduce output data (e.g., data analysis), and generate data visualisation plots to explain the outcomes.

Deliverables:

The results of the data processing activities must be presented in the form of the report and programming code along with comments. This report should highlight and explain the programming and data handling challenges that you will face and the methods you employed to overcome these challenges. The report should include the following

1) A description of the source dataset(s) and the dataset should not be older than five years.
2) A description of the objective of Big data storage and analysis.
3) Details of the data processing activities carried out, including data preparation and processing in the Hadoop MapReduce/ Spark environment.
4) A discussion of the rationale and justification for the choices you have made in terms of data processing, programming language choice, and design patterns that you have implemented.
5) A report of results by making appropriate use of the figures along with captions and tables, etc.

Note that MapReduce-style processing in this instance is considered to include platforms, such as Hadoop/ Apache Spark.

The marks for CA1 are based on the following deliverables

- A clear motivation, objectives and handling of data complexity based on the considered dataset (20%).
- A precise justification of chosen distributed data processing framework as well as deployment (Hadoop or Apache Spark), coding and implementation (30%).
- The results are discussed by supporting appropriate discussion as well as visualizations and conclusions. Harvard style of citations and references should be used in the report (30%).
- The framework and data storage and processing will be demonstrated using a voice-over PowerPoint presentation that lasts no more than seven to eight minutes (20%).

Submission Requirements

All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the marks awarded.

- The code, datasets and voice over presentation should be provided and uploaded in zip format on Moodle.
- Must be clearly specified the number of words used in the report.
- Number of Words for the report (1500 words ±5%) excluding diagrams and code.
- Use Harvard Referencing when citing third party material
- Be the student's own work.
- Include the CCT assessment cover page.
- Be submitted by the deadline date specified or be subject to late submission penalties.
- User version control like Github to show the progress in CA1 and you must have 5 commits substantial before the submission of CA1.

Acceptable Use of AI for Assignment at CCT

Acceptable and Unacceptable Use of AI	• The use of generative AI tools (e.g. ChatGPT, Dall-e, etc.) is permitted in this assignment for the following activities:
	  ○ Brainstorming and refining your ideas;
	  ○ Fine tuning your research questions;
	  ○ Finding information on your topic;
	  ○ Drafting an outline to organise your thoughts; and
	  ○ Checking grammar and style.
	• The use of generative AI tools is not permitted in this course for the following activities:
	  ○ Impersonating you in classroom context
	  ○ Completing group work that your group has assigned to you
	  ○ Writing a draft of a writing assignment
	  ○ Writing entire sentences, paragraphs or papers to complete class assignments.
	• You are responsible for the information you submit based on an AI query. Your use of AI tools must be properly documented and cited.
	• Any assignment that is found to have used generative AI tools in an unauthorised way will be subject to college disciplinary procedures as outlined in the QA Manual.
	• When in doubt about permitted usage, please ask for clarification.
		This statement is useful when you are allowing the use of AI tools for certain purposes, but not for others. Adjust this statement to reflect your particular parameters of acceptable use, and your discipline context.

GRADING RUBRIC – Storage Solutions for Big Data - 2024

GRADE	**90-100%**	**80-90%**	**70-79%**	**60-69%**	**50-59%**	**40-49%**	**35-39%**	**<35%**
Performance	**Exceptional**	**Outstanding**	**Excellent**	**Very Good**	**Good**	**Acceptable**	**Fail**	**Fail**
Motivation, Objectives, and Handling of Data Complexity (20%)	Exceptional understanding and articulation of project motivation and objectives with a sophisticated approach to handling data complexity.	Outstanding clarity and insight into project motivation and objectives, demonstrating a strong understanding of data complexity.	Excellent presentation of project motivation and objectives with a clear approach to handling data complexity.	Very good explanation of project motivation and objectives, showing a competent handling of data complexity.	Good articulation of project motivation and objectives, with an acceptable approach to handling data complexity.	Adequate presentation of project motivation and objectives, with some limitations in addressing data complexity.	Limited or unclear presentation of project motivation and objectives, with significant shortcomings in handling data complexity.	Project lacks clear motivation and objectives, and data complexity is not adequately addressed.
Justification, Coding, and Implementation (30%)	Exceptional justification of the chosen distributed data processing framework, impeccable coding, and flawless implementation with a deep understanding of deployment strategies.	Outstanding justification of the chosen framework, with high-quality coding and implementation, demonstrating a strong grasp of deployment considerations.	Excellent justification for the chosen framework, with well-implemented code, showing proficiency in deployment.	Very good justification for the chosen framework, solid coding, and implementation skills with a satisfactory deployment approach.	Good justification for the chosen framework, acceptable coding, and implementation, with some weaknesses in deployment strategy.	Adequate justification for the chosen framework, basic coding and implementation, with notable shortcomings in deployment.	Limited or unclear justification for the chosen framework, inadequate coding and implementation, and significant issues with deployment.	Chosen framework, coding, and implementation are not justified, with severe deficiencies in deployment.
Results, Discussion, Conclusions, Citations, and References (30%)	Exceptional presentation of results, insightful discussion, and well-drawn conclusions. Comprehensive use of citations and references, showcasing a deep understanding of the field.	Outstanding presentation of results, thorough discussion, and strong conclusions. Effective use of citations and references, indicating a solid understanding of the subject matter.	Excellent presentation of results, clear discussion, and sound conclusions. Adequate use of citations and references, demonstrating a good understanding of the field.	Very good presentation of results, acceptable discussion, and reasonable conclusions. Some improvement needed in citations and references.	Good presentation of results, basic discussion, and conclusions. Limited use of citations and references, requiring improvement.	Adequate presentation of results, minimal discussion, and conclusions. Lacks proper use of citations and references.	Limited or unclear presentation of results, insufficient discussion, and conclusions. Serious deficiencies in citations and references.	Results, discussion, conclusions, citations, and references are either absent or extremely deficient.

Voice-over PowerPoint Presentation (20%)	Exceptional execution of the voice-over PowerPoint presentation, captivating the audience with clear and engaging content within the specified time frame.	Outstanding execution of the voice-over PowerPoint presentation, maintaining audience interest with clear and well-paced content within the time limit.	Excellent execution of the voice-over PowerPoint presentation, effectively delivering content in a compelling manner within the time frame.	Very good execution of the voice-over PowerPoint presentation, keeping the audience engaged with a generally clear and timely presentation.	Good execution of the voice-over PowerPoint presentation, with some areas needing improvement in clarity and adherence to the time limit.	Adequate execution of the voice-over PowerPoint presentation, with notable shortcomings in clarity and time management.	Limited or unclear execution of the voice-over PowerPoint presentation, with significant issues in clarity and time management.	Voice-over PowerPoint presentation is either absent or extremely deficient.