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, Data Visualization Techniques		
Assignment Type:	Individual	**Weighting(s):**	40%, 40%
Assignment Title:	Integrated Continuous Assessment 2		
Lecturer(s):	Muhammad Iqbal, James Garza		
Issue Date:	19th October 2023		
Submission Deadline Date:	26th Nov 2023		
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 separate files based on your work, for example word file, jupyter notebook, dataset and any supporting information.		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	N/A		

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:

MLOs SSBD

MLO 1 - Articulate / present the necessity for big data storage solutions and their integration in modern commercial Artificial Intelligence systems.

(Linked to PLO 1)

MLO 5 - Develop a proposal utilising current cloud-based architectures to implement solutions for a given big-data processing project. Present to team members and stakeholders

(Linked to PLO 2, PLO 6)

MLOs Data Visualisation

MLO 1 - Discuss the concepts, techniques and processes underlying data visualisation (Linked to PLO 1)

MLO 3 - Select appropriate data visualisation techniques for a given use case, data characteristics and multiple transmission media (Linked to PLO 3, PLO 4)

MLO 5 - Present a detailed visualisation of a data analysis to peers, team members and project stakeholders. (Linked to PLO 3, PLO 6)

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
90% +	Exceptional	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
50 – 59%	Good	Achievement includes that required for a Pass and in **some** respects is significantly beyond this
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
0 – 34%	Fail	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 of 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

Question 1:

Can you define Big Data? Explain major characteristics of Big Data. Can banks enhance their profits with the support of big data processing and analysis? Research and name top three businesses that have obtained the benefits of big data storage solutions in the recent past.

(20 Marks)

Question 2:

What is HDFS, and how does it differ from a traditional file system? Describe and explain different layers of Hadoop framework. Explain five important characteristics of the Hadoop framework. Show the deployment of Hadoop on your virtual machine (VM) by providing the screenshots of (Namenode, Datanode etc.) and your username clearly shows your VM.

(20 Marks)

Question 3:

Demonstrate a comparison of MySQL and Apache Hive based on the architecture and performance. Consider a dataset and perform a query on both systems with at least 5,000 rows and at least 5 features. Show the duration of query execution by displaying screenshots obtained from a virtual machine (VM).

(20 Marks)

Question 4:

Explain Apache Flink architecture and illustrate with your own conceptual diagram (Use of online/ book images is prohibited, Use draw.io to create the image). What is Apache Storm, and how does it differ from other distributed computing systems? Consider a text file comprising at least 20,000 words and write a wordcount program (Java/ Python) to count the frequency of words and related aggregation functions.

(20 Marks)

Question 5:

Why is Apache Storm useful for Stream processing specifically? Distinguish the characteristics of Apache storm as compared to Hadoop. What is the role of Apache Zookeeper in Apache Storm deployment. Provide the screenshot of your VM to show working of Storm UI including Cluster, Nimbus and Owner summary.

(20 Marks)

Data Visualisation

Create an interactive dashboard using your chosen dataset. Create a wireframe proposing your design before the implementation of the dashboard. The dashboard will include at least two rows and three columns of six sections in total. There should be at least four plots in the sections, and the remaining sections could be text, tables or any other relevant information you deem necessary to give critical insights to the viewer of the dashboard. The dashboard will include a range of visualisations that effectively communicate the key insights derived from the exploratory data analysis.

The dashboard could include any of the following visualisations:

A heatmap showing the correlation matrix between all continuous variables. A heatmap could help the viewers understand the strength and direction of the relationships between variables.

A scatter plot matrix will show pairwise relationships between all continuous variables. A scatter plot matrix could enable the viewer to identify visually potential outliers or non-linear relationships between the variables visually.

A bar chart showing the distribution of the target variable could help the viewers understand the range and distribution of the target variable.

Histogram for each continuous variable could help the viewers understand the distribution of the continuous variables.

A stacked bar chart showing the distribution of each categorical variable could help viewers understand the distribution of the categorical variables and how they relate to the target variable.

A box plot showing the distribution of the target variable for each category of each categorical variable could enable viewers to understand how the target variable varies across the different categories of categorical variables.

A line chart showing any patterns or trends that change over time across a continuous variable.

A scatter plot showing the relationship or association between two continuous variables. Particularly useful for identifying patterns, trends, clusters or correlations in the data.

The dashboard will be designed to be interactive, allowing the audience to filter and explore the data in more detail. For example, the audience can filter the categorical variables by category or select specific data points in the scatter plots to explore and understand the underlying data. The dashboard will also briefly describe the dataset, its provenance and domain, and a summary of the key insights derived from the exploratory data analysis. There will need to be at least four interactive plots in the dashboard. You can use any Python visualisation libraries such as Plotly Dash or Altair to create the interactive dashboard.

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 and datasets should be provided and uploaded as separate files on Moodle.
- Must be clearly specified the number of words used in the report.
- Number of Words for the report (Min: 2000 words +- 5%) excluding diagrams and code.
- Use any version control system (for example Github) to show the weekly progress of your integrated CA2 and there should be at least 5 commits. You should provide access to the Github repository to your lecturers.
- Use Harvard Referencing when citing third party material
- Be the student’s own work.
- Be submitted by the deadline date specified or be subject to late submission penalties
- Must be clearly specified the number of words used after each section in the report.
- Include the CCT assessment cover page.
- We can check the authenticity of practical work on your Virtual machine (Do not destroy your VM at least 8 weeks after the submission of this CA).

Marking Schedule Data Visualisation:

Description	Weighting
Explain what was done with the data and the pre-processing and cleaning.	10%
Creativity and originality of the dashboard. Explain the colour selection and Python visualisation library use. Effectiveness of the dashboard in communicating key insights. Explain the selection of the plots and key insights found.	30%
Suitability of visualisation techniques for the given use case and data characteristics. Explain the plots and why you chose the specific visualisations.	40%
Clarity of dashboard design and user interface. An initial wireframe of the dashboard and rationale on the six sections giving key insights to the viewer of the dashboard. The plots are interactive and easy to use.	20%
Poor referencing, spelling, grammar and layout will incur marking penalties.	
Total	**100%**