National College of Ireland

Masters in Science in Data Analytics (MSCDAD_A)
Masters in Science in Data Analytics (MSCDAD_B)
Masters in Science in Data Analytics(MSCDAD_C)
Masters in Science in Data Analytics(MSCDA_JANOL23)
Postgraduate Diploma in Science in Data Analytics (PGDDA_JAN23)

Database and Analytics Programming

Team Project (70%)

Semester 2, 2022/23

Submission deadline: 25th April 2023 at 23:59

1 Introduction

This project is designed to evaluate the learning objectives of the Database and Analytics Programming module as outlined below:

LO1: Analyse, compare, contrast and critically evaluate the characteristics of programming languages, programming environments and database systems commonly utilised for data analytics solution implementation.

LO2: Critically assess the challenges associated with processing big data datasets and compare and contrast programming for big data vis-à-vis programming for conventional datasets.

LO3: Evaluate tools and techniques for managing the data pipeline and preparing data for further analysis through data wrangling, cleaning, and validation.

LO4: Critically assess methods and practices for software development in order to design and implement data programming requirements.

LO5: Evaluate, design and implement solutions for processing datasets by using key programming patterns and constructs for data analytics, relevant programming languages, and suitable database systems.

2 Objectives

The objective of this project is to identify and carry out a series of analyses on a collection of large datasets that are somehow related or complement each other, utilising appropriate programming languages, programming environments and database systems.

This project is a team project. Teams will have a **maximum of 3** members.

Your project must incorporate the following elements:

1. A minimum of two datasets should be used, of which at least one should be semi-structured. For teams of 3, a third dataset (semi-structured) will be required.
2. Datasets must be programmatically stored in appropriate database(s) prior to processing.

3. Programmatic pre-processing, transformation, analysis and visualisation of the data.
4. Programmatically storing the processed output data in appropriate databases.

For example, you could use Python to programmatically retrieve a semi-structured dataset (XML or JSON) or web-scraped or streaming data) and store this data in MongoDB.

You could then use an ETL process (for instance, using Dagster or Luigi Python libraries) to read these data from MongoDB, to process and transform it, finally creating structured datasets that you store in PostgreSQL for later usage.

Following that you could conduct further analysis on these data to find interesting patterns my applying knowledge gained in other modules (e.g., statistical analysis,machine learning), and generate visualisations to better present the results.

Each dataset should contain at least 1,000 records. Some appropriate datasets may be found at:

- https://catalog.data.gov/dataset?res_format=XML
- http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/
- https://data.gov.ie/dataset?res_format=JSON
- https://catalog.data.gov/dataset?res_format=JSON
- https://data.worldbank.org/

A list of other potential sources will be posted on Moodle.

3 Deliverables

Project Report

The objectives, methodology and results of your analysis should be presented in the form of a project report. This report should discuss the programming and data processing challenges that you encountered and the means and mechanisms you implemented to overcome these challenges.

The report should be around 3,000 words in length (excluding references), should use appropriate academic style and referencing, and be presented in the IEEE conference format. Templates for Microsoft Word and LaTeX can be downloaded from the IEEE[1].

The report should contain the following sections:

- **Abstract**
  This should provide a summary of the project objectives, methods and results. Take a look at abstracts from papers in your literature review to get an idea of what constitutes a good/bad
- **Introduction**
  Here you should provide a short motivation for the project, describe the relevance of the topic and state the objectives of the project. Note that the proposed analysis should answer a novel question, which should be clearly stated by means of appropriately formed research question(s).

[1] https://www.ieee.org/conferences_events/conferences/publishing/templates.html

- **Related Work**
  In this section, you should summarise relevant academic work that addressed similar problems or guided your decisions. Note that this should be a **critical evaluation**. It should be more than a mere summary of the works and should discuss their limitations and implications.

- **Methodology**
  This section should contain :

  - A detailed description of the underlying dataset(s) and your justification for choosing them.
  - Full descriptions and justifications of the data processing activities carried out, such as use of APIs, databases, etc.
  - Complete descriptions and justifications of the implemented data processing algorithms.
  - Justifications for the choice of technologies used, such as programming languages, libraries and databases.
  - Diagrams providing a visual overview of the data gathering, processing and analysis flow.

- **Results and Evaluation**
  Here you should present the results of your work, making appropriate use of figures, tables, etc. You should provide evidence of how the project objectives were met, ensuring that you discuss your research findings, their interpretation(s) and implications.

- **Conclusions and Future Work**
  In this section you should detail what others can/could learn from your work. You should discuss your findings in the context of the research question(s) you elicited earlier. You should present the limitations of your work, i.e. this should be a critical self-evaluation. Lastly, you should suggest potential directions for future work. Typically you would describe what you would do differently or how you would extend your work if you had more time.

- **Bibliography**
  Here you should provide a **complete list** of the academic works cited and online materials used in the project. References should be included as in-text citations **using the IEEE citation style**.

Project Presentation

As a team you should create a video presentation (maximum 10 minutes long) that will act as a discussion point for your work. It should be used to provide a discussion on what you did, how you did it, why you did it and what you discovered.

Note that although individual members will be presenting different parts of the video, each member of the team is expected to be able to present all aspects of the work individually and without assistance from other group members, if required.

Code Artefact

You should create a *zip* or *gz* archive all assets such as program code, data and system configuration details. If working as a team, there is an additional requirement for you to set up a **private** GitHub repository where each team member should commit their own code and any changes they make to code created by other team members.

Note: Having one team member be responsible for making all commits is not acceptable.

4 Submission

The project carries 70% of the total marks for the module, with a mark of 40% or greater being required to pass.

There should be only **one submission per team**, consisting of:

- A **project report** that must include the full name of each team member (as per NCI official documents) and their student number. These must be clearly visible on the front page of the report. The report should be named *teamX.pdf* where X is your team identifier and should be uploaded as a PDF document to the **Project Report** Turnitin link on Moodle.
- A **code artefact**, which should be uploaded as a *zip* or *gz* archive to the **Code Artefact** link on Moodle. This should be named *teamX.zip* or *teamX.gz*, where X is your team identifier.
- A **video presentation** that must include the full name of each team member (as per NCI official documents) and their student number. These must be clearly visible at the start of the video. This should be uploaded as a *mp4* video to the **Project Presentation** link on Moodle.
- A **work breakdown report** describing in detail the contribution of each team member. Again, this should be in PDF format and should include the full name of the team members (as per NCI official documents) as well as their student numbers. This should be named *teamXworkbreakdown.pdf* (where X is your team identifier) and should be uploaded to the **Work Breakdown Report** Turnitin link on Moodle.

Late submissions will not be accepted unless an extension has been requested through NCI360 and officially approved.

5 Marking

The project will be marked according to the grading rubric provided at the end of this document.

6 Academic Integrity

Any written work created by others must be properly cited and should be paraphrased or summarised where possible, otherwise it should be included in quotes. Figures not created by you should include an acknowledgement detailing the name(s) of the creator(s). Code found on the internet should not be claimed as your own, but instead a comment should be included in the source code indicating where you obtained it.

Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library[2].

Note: All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation. Any submission showing evidence of such misconduct will be referred to the college's academic misconduct committee for disciplinary action.

[2] https://libguides.ncirl.ie/academicintegrity

Grading Rubric - Database and Analytics Programming Project

Semester 1 - 2022/23

Criterion	**Solid H1** ≥ 80%	**H1** ≥ 70% < 80%	**H2.1** ≥ 60% < 70%	**H2.2** ≥ 50% < 60%	**Pass** ≥ 40% < 50%	**Fail** < 40%
Project Objectives **(10%)**	Very challenging project objectives are exceptionally well presented, fully met and thoroughly discussed	Challenging project objectives are well presented, are fully met and thoroughly discussed.	Reasonable project objectives are well presented, fully met and adequately discussed.	Reasonable project objectives are clear, are mostly met and adequately discussed.	The objectives are clear, if unambitious and are at least partially met and briefly discussed.	The objectives of the project are unclear, have not been discussed. It is not possible to discern if the objectives have been met.
Literature Review **(10%)**	An excellent critical analysis of substantive and highly relevant literature.	A very good critical analysis of substantive and relevant literature.	A good analysis of relevant literature. The critical analysis aspect could be somewhat stronger.	An adequate analysis of mostly relevant literature. The critical analysis aspect could be significantly stronger.	A limited analysis of some relevant literature but it lacks evidence of understanding.	Little or no relevant literature reviewed. Very limited evidence of understanding.
Data Complexity and Handling **(20%)**	The datasets have been well prepared and meaningfully explored. All datasets were stored in appropriate databases before and after processing. At least two datasets have a high degree of complexity. At least one dataset was programmatically retrieved - through an API or by web scraping.	The datasets have been well prepared and meaningfully explored. All datasets were stored in appropriate databases before and after processing. At least two datasets have a high degree of complexity.	The datasets have been well prepared and explored. At least one dataset was stored in an appropriate database. At least one dataset has a high degree of complexity.	The datasets have been appropriately prepared for analysis. At least one dataset was stored in an appropriate database. At least one of the datasets is non-trivial.	The datasets were appropriately handled given the objectives. The use of databases is very basic and some inappropriate choices may be evident. The datasets are somewhat trivial.	Only one somewhat trivial dataset was used. No database was used to store the datasets. No obvious development was carried out.
Data Processing Implementation **(20%)**	The data processing algorithms used play a well conceived and essential role in meeting the project objectives. The implementation significantly exceeds the stated minimum requirements.	The data processing algorithms used play a well conceived and essential role in meeting the project objectives. Multiple data processing techniques / languages were employed.	The use of data processing algorithms is well-thought and appropriate for the project objectives. Comprehensive use of at least one data programming language and multiple techniques.	The use of data processing algorithms is meaningful and appropriate for the project objectives. There is evidence of appropriate use of at least one data programming language and a small number of appropriate techniques.	Appropriate but basic use of data processing algorithms. Basic use of data programming languages and a limited number of techniques.	Poor or no implementation. If an implementation is provided, it demonstrates inappropriate use of data processing algorithms.

Grading Rubric (continued)

Criterion	**Solid H1** $\geq 80\%$	**H1** $\geq 70\% < 80\%$	**H2.1** $\geq 60\% < 70\%$	**H2.2** $\geq 50\% < 60\%$	**Pass** $\geq 40\% < 50\%$	**Fail** $< 40\%$
Level of Automation **(10%)**	All parts of the analysis are automated within a single process control flow. Every run of the process can result in different results as new data is extracted and subsequently included, such as data obtained through an API.	All parts of the analysis are automated within a single process control flow.	Most core components of the analysis are automated within a single process control flow.	Some components of the analysis are included within a larger process. However, some components of the analysis are run as separate processes.	Individually all components of the analysis are automated, but not necessarily connected together as part of a larger process flow.	Little or no evidence of automation.
Results and Conclusions **(20%)**	Three or more insightful findings are excellently presented and thoroughly discussed in the context of the domain using appropriate references to prior work.	Three or more interesting and non-arbitrary findings are presented and thoroughly discussed the context of the domain using appropriate references to prior work.	Three or more interesting non-arbitrary findings are presented and thoroughly discussed.	Two or more interesting non-arbitrary findings are presented and appropriately discussed.	Two or more interesting non-arbitrary findings are presented but are poorly discussed.	Little to no non-arbitrary results and/or findings are presented.
Quality of Writing **(10%)**	Exceptionally well written, with no language errors. All figures are well conceived, readable and correctly captioned. The IEEE template is strictly adhered to. The report does not exceed the length limits. All references are appropriately and correctly used.	Well written, with no significant language errors. All figures are well conceived, readable and appropriately captioned. The IEEE template is adhered to. The report does not exceed the length limits. References are appropriately and correctly used.	Well written, but has a few significant language or style errors. Figures are well presented. The IEEE template and length limit are adhered to. References are complete and correctly used.	Adequately written. but as a few significant language and/or style errors. Some figures are may be hard to read. The IEEE template and length limit are mostly adhered to. References are complete, and correctly used.	Adequately written, with some significant language and/or style errors. Figures may be hard to read or presented in a suboptimal manner. The IEEE template may not have been followed. References are mostly complete and correctly used.	Poorly written and littered with typographical errors and/or poor use of English. The IEEE template was not used. Figures may be hard to read. References (if any) are largely incomplete.