Postgraduate Diploma in Data Analytics
Data Intensive Architectures

Project (100%)

Semester 3, 2021/22

1 Overview

You are required to programmatically acquire, analyse and interrogate two or more datasets using a distributed or parallel processing environment.

1.1 Datasets

The datasets should meet the following minimum requirements:

1. They should be relatable in some way.
2. They should complement each other such that your study, or a study similar to yours, could not be conducted without one of the datasets.
3. They should be sufficiently large to be considered "data-intensive". Although there is no upper bound on the size of these datasets, you should take into account the capacity and capabilities of your cloud instances and likely processing times.
4. They should be both legally and ethically sourced and employed.

1.2 Processing

The following are minimum requirements for the processing phase of your analysis:

1. Your data sets should be sourced from open/free data repositories. A non-exhaustive list of such repositories will be provided on Moodle. You should programmatically extract the data the where possible, preferably using an API where one is available, otherwise by directly downloading the data from the source websites.
2. They should then be cleaned, transformed and conformed.
3. Appropriate joining approaches should be used to combine the datasets.
4. At a very minimum, the data should be completely prepared for at least one full and substantive analysis.

Your project should address **data quality**, ideally in the context of the ISO/IEC 25012 standard.

1.3 Analysis

You must perform at least one complete, substantive analysis:

1. The analysis must be performed using a distributed or parallel processing approach.
2. The processing must be oriented towards extracting at least three interesting, non-trivial insights into the combined datasets. Your research questions should be your starting point here.
3. Where appropriate, you may use tools such as Tableau or PowerBI to visualise the results. You may also use R with libraries such as *ggplot2*/*plotly* or Python with *seaborn*/*plotly* to produce such visualisations.

2 Learning Outcomes Assessed

This project is designed to evaluate all learning outcomes for this module, namely:

LO1 Critically compare and contrast multiple distributed system models and their associated enabling technologies.

LO2 Demonstrate in-depth knowledge of different types of processing on different data-intensive computational resources.

LO3 Identify and categorise platforms and software environments for cloud and cognitive computing.

LO4 Critically analyse the features of high performance computing platforms and how they enable parallel and distributed programming paradigms.

3 Deliverables

3.1 Report

The results of the analysis must be presented in the form of project report. The report should be 6-8 pages in length (including figures and references), and must follow the IEEE format[1] in addition to be employing appropriate referencing methods and academic writing style. The report should include the following:

1. A full description of the source datasets;
2. A statement of the objective(s) of the analysis. Note that the analysis should attempt to answer a novel question;
3. Details of the data processing activities carried out, including preparation of the data and processing the data in a distributed or parallel processing environment;
4. A discussion on scalability/speedup issues, particularly those related to dataset size.
5. A presentation of the rationale for and justification of the choices you have made;
6. A brief discussion on the ethical considerations involved in the sourcing and processing of data; and
7. A presentation of results by making appropriate use of figures, tables, etc.

Do not include program code in your report. Algorithms expressed in pseudocode may be included if they significantly aid the understanding of your work.

3.2 Code Artefact

A *zip* or *gz* archive of all code and datasets used to produce the results. The directory structure and file naming should follow a logical and consistent naming scheme. The root directory of the archive should also contain a plain text file providing a brief explanation on the processing and analysis can be re-run if required.

3.3 Presentation

Presentations will be submitted as a video recording, with the following mandatory requirements:

1. Duration: Max 10 minutes (approx. 7 slides).
2. The presentation should give a summary of the methodology applied.
3. You should also demonstrate and briefly discuss your code as well as showing your results.

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

4 Project Report

The report should have the following structure:

Abstract
150-200 words

Provide a summary of the motivation for and objectives of your research, the approach taken and key findings.

Introduction
The remainder of the 1st page and up to 1 column of 2nd page

Here you should present a statement of the objective(s) of the analysis, a motivation of the problem and a discussion of the relevance of chosen topic. You should also describe your chosen datasets and your rationale for choosing them. Finally, you should present appropriately formed research question(s).

Related Work
1 page with citations into 10 or more peer-reviewed works

Present an analysis of relevant peer-reviewed work that addresses similar problems or guided your decisions. This should focus on works that have used distributed or parallel analyses on the datasets you have chosen and/or those performing a similar distributed or parallel analysis on other datasets. Your review should not include general papers on the topic of big data processing approaches but instead should focus on those works where such approaches have been used to tackle problems in the same or a similar domain. Provide a critical evaluation of the cited works (i.e. you should go beyond providing a summary of each work.

Methodology
2-3 pages

In this section you should provide an overview of the architecture and application workflow of your analysis. You should also discuss (in the order they are carried out) the data processing techniques and patterns used to ingest, process and export the data, and the justifications for employing them.

Results & Evaluation
$1\frac{1}{2}$-2 pages

Here you should present your results, making appropriate use of figures, tables, etc. Focus on those findings that were unexpected and detail how your findings (partially) answer the research question(s).

Conclusions & Future Work
$\frac{1}{2}$ page

In this section you should discuss your research findings as well as their implications and limitations. You should also describe options for extending the work that could be explored.

References
no page limit

Here you should provide a complete list of the academic works and/or online materials used in the project. References should be included as in-text citations using to the IEEE citation style. For citation style guidelines, please refer to the NCI Library guide for Data Analytics[2]

Space-saving tips

- Never have a line less than half-full at the end of a paragraph. Almost any paragraph can be rewritten so that this is not the case!
- Graphs, flow diagrams and tables are easy to do sub optimally– draw them properly and decide if they really need to be as big as they are, or if they really should span both columns.
- Sub figures (e.g. 3 graphs as one figure prefixed a, b c that span both columns) are usually fairly space efficient.
- The LaTeX template is significantly cleverer than the Word one, and will do more work to save space.
- In LaTeX, paragraph spacing is heavily optimised. This also means that cutting out a line or two before a new section can cause paragraph spacing to be recalculated thus saving significant space.

[2] http://libguides.ncirl.ie/dataanalytics

5 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). In general your code should be your own. Small snippets of code found on the internet should not be claimed as your own, but instead a comment should be included in the source code indicating who wrote it and where you obtained it.

Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library[3].

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.

6 Submission

Your submission must include your project report document along with any programming code and system configuration elements. The final report and code artefact must be submitted to Moodle before 23:59 on 12th August 2022. Late submissions will be subject to the usual penalties. If you miss the deadline for any valid reason you can submit an application for coursework Extension/Re-run using the NCI360.

Ensure that your name in full (as per NCI official documents) and student number are clearly visible on the front page of the report and in the filename of any artefacts uploaded.

7 Marking

The project carries 100% of the total marks for the module, with a mark of 40% or greater being required to pass. These are awarded according to the grading rubric on the following page.

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

GRADING RUBRIC – Data Intensive Architectures Project - Semester 3, 2021-22

CRITERION	*H1* (70%+)	*H2.1* (60% - 69%)	*H2.2* (50% - 59%)	*PASS* (40% - 49%)	*FAIL* (0% - 39%)
Abstract 5%	An excellent abstract that succinctly but comprehensively summarises the objectives and key findings of the analyses.	A very good abstract that comprehensively summarises the objectives and key findings of the analyses.	A good abstract that to a large extent summarises the objectives and key findings of the analyses.	A reasonable abstract that offers an incomplete summary of the objectives and/or key findings of the analyses.	A poor abstract that does not adequately summarise the objectives or findings of the analyses.
Introduction 10%	An excellent introduction that provides a compelling case for the proposed analyses.	A very good introduction that offers a very convincing case for the proposed analyses.	A good introduction that furnishes a largely convincing case for the proposed analyses.	A adequate introduction that offers a somewhat weak case for the proposed analyses	A poor introduction that fails to motivate the problem or provide a case for the proposed analyses.
Related Work 15%	An excellent critical analysis of substantive and relevant literature leading to compelling rationale for the proposed analyses.	A very good critical analysis of substantive and relevant literature leading to convincing rationale for the proposed analyses.	A good analysis of relevant literature leading to clear rationale for the proposed analyses.	An adequate analysis of mostly relevant literature leading to an adequate rationale for the proposed analyses.	A review of some relevant literature but limited evidence of understanding and a weak rationale for proposed research.
Methodology 40%	An excellent application of distributed or parallel design principles in terms of appropriate methodology as well as the methods for generating and analysing data. A very thorough discussion on scalability and ethical issues is provided.	A good application of distributed or parallel design principles in terms of appropriate methodology as well as the methods for generating and analysing data. A good discussion on scalability and ethical issues is provided.	An adequate application of distributed or parallel design principles in terms of appropriate methodology as well as the methods for generating and analysing data. A very thorough discussion on scalability and ethical issues is provided.	A weak application of distributed or parallel design principles and limited evidence of understanding of appropriate methodology or methods for generating and analysing data. There is limited discussion on scalability or ethical issues.	A poor application of distributed or parallel design principles and limited evidence of understanding of appropriate methodology or methods for generating and analysing data. There is little or no discussion on scalability or ethical issues.
Results 20%	An excellent presentation of the results using clear and appropriate visualisations.	A very good presentation of the results using clear and largely appropriate visualisations.	A good presentation of the results, using largely appropriate visualisations. Some issues with legibility of parts of the visualisations.	An adequate presentation of the results. Some inappropriate choices of visualisations and/or major issues with legibility of parts of the visualisations.	A poor presentation of the results, with inadequate choicesof visualisation types and poor implementation.
Conclusions and Future Work 10%	An excellent discussion of the implications and limitations of the results. An excellent consideration of potential research impact/outcomes.	A very good discussion of the implications and limitations of the results accompanied by a considerable discussion of potential research impact/outcomes.	A good consideration of the implications and limitations of the results accompanied by a reasonable discussion of potential research impact/outcomes.	Adequate but incomplete consideration of the implications and limitations of the results accompanied by a passable discussion of potential research impact/outcomes.	Little or no consideration of the implications and limitations of the results. Scant discussion of potential research impact/outcomes.