National College of Ireland

Masters in Science in Data Analytics
(MSCDAD_A_JAN24I/MSCDAD_B_JAN24I / MSCDAD_C_JAN24I)

Scalable Systems Programming

Project (100%)

Noel Cosgrave

Release Date: 7th June 2024
Submission Date: 12th August 2024

Duration: 67 days

UN sustainability goals: Students are encouraged to address one or more of the UN Sustainability Goals[1] when choosing a topic for the project. Note that this is entirely optional and will not impact on the marks awarded for the project.

This assessment is worth 100% of the marks for this module and is designed to evaluate all the learning objectives for the module, as listed below:

LO1 - Demonstrate in-depth knowledge of parallel algorithms on large amounts of data

LO2 - Identify and categorise search techniques including similarity search and search engine technologies.

LO3 - Critically compare and contrast different data-stream processing and specialised algorithms.

LO4 - Critically analyse mining and clustering algorithms on large multi-dimensional datasets.

LO5 - Develop and implement efficient programming solutions for problems relating to processing data at scale.

Instructions

You are required to programmatically acquire a suitable data set, to perform a **cluster analysis** on that data set using a distributed processing approach and to document your analysis in a technical report.

[1] https://sdgs.un.org/goals

This data set should meet the following minimum requirements:

1. It should be large enough to warrant the use of a scalable distributed computing approach. Typically this will mean using data sets of 4GiB or greater in size. Although there is no upper limit on the size of the data sets, you should take into account the capacity and capabilities of the your system and likely processing times.

2. It should be both legally and ethically sourced and employed.

You should experiment with at least two of the clustering approaches presented in the lectures, comparing and contrasting the performance of each using appropriate metrics.

Processing

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

1. Where possible you should extract the data directly from the source(s) and place the data into your own block or blob storage.

2. They should then be cleaned, transformed and conformed as required.

3. The data should be completely prepared for a thorough and substantive analysis.

Analysis

1. The analysis must be performed using MapReduce in a true distributed processing environment. For the purposes of this assessment, this includes both Apache Hadoop and Apache Spark.

2. The MapReduce processing must be oriented towards extracting at least three interesting, non-trivial insights into the data set(s) or the performance of the algorithm on these data sets. 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/bokeh to produce such visualisations.

Project Report

The report should be structured as follows:

1. Abstract
    - Provide a summary of the objectives, methodology and results of the analysis. Note: Look at abstracts in your literature review to get an idea of what constitutes a good or bad abstract.
2. Introduction
    - Present a motivation of the problem and a discussion of the relevance of chosen topic.
    - Provide a statement of the objective(s) of the analysis and the elicitation of appropriately formed research question.
3. Related Work
    - Present an analysis of relevant (academic) works that addressed similar problems or guided your decisions. This should focus on works that have used the data sets you have chosen and/or those performing a similar analysis on other data sets.
    - The emphasis should be on comparing and contrasting different data-stream processing and other specialised algorithms, specifically concentrating on parallel algorithms relevant to your problem and suitable to be deployed on scalable processing environments.
    - This should be a critical evaluation (i.e. it should go beyond being a mere summary of the referenced works).

- Do not provide a review of general papers on the topic of big data processing approaches but instead focus on those works where such approaches have been used to tackle problems in the same or a similar domain.

4. Methodology
   - Provide a description of the data sets chosen and their attributes in a data dictionary.
   - Explicitly detail all translation rules defining data manipulation(s), such as the setting of default values, the splitting or combining of attributes and/or map values.
   - Provide an overview of the architecture and application workflow of your analysis. Here you should address the scalability approaches you have employed and your reason for choosing them.
   - Discuss (in the order they are carried out) the data processing activities used to ingest, process and export the data, and the justifications for employing them.
   - Address any ethical considerations in the sourcing and processing of data.
5. Results
   - Present your results, making appropriate use of figures, tables, etc.
   - Focus on those findings that were unexpected.
   - Detail how these findings (partially) answer the research question.
6. Conclusions and Future Work
   - Discuss your research findings as well as their implications and limitations.
   - Detail options for extending the work that could be explored.
7. References
   - 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.

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

Submission

Your report must be in IEEE format and should be uploaded as a **single document** in **PDF format only** to the Turnitin link on Moodle by the submission date shown at the top of this document.

Any supporting code should be compressed in zip format and uploaded the Code Artefact link on Moodle.

As this is a terminal assessment, late submissions will not be accepted.

Marking

Marks for the assessment will be allocated according to the rubric at the end of this document.

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 acknowledgment detailing the name(s) of the creator(s).

While the use of AI tools to help locate appropriate literature is acceptable, any other use of such tools is **strictly prohibited**. Any use of AI must be documented as per the Use of AI in Teaching and Learning: Student Guide [2].

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.

[2] https://libguides.ncirl.ie/useofaiinteachingandlearning/studentguide

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

Grading Rubric - Scalable Systems Programming Project

Semester 3 - 2023/24

Criterion	**H1** $\geq$ 70%	**H2.1** $\geq$ 60% < 70%	**H2.2** $\geq$ 50% < 60%	**Pass** $\geq$ 40% < 50%	**Fail** < 40%
Abstract **(5%)**	An excellent abstract that succinctly but comprehensively summarises the objectives and key findings of the analysis.	A very good abstract that comprehensively summarises the objectives and key findings of the analysis	A good abstract that to a large extent summarises the objectives and key findings of the analysis.	A reasonable abstract that offers an incomplete summary of the objectives and/or key findings of the analysis.	A poor abstract that does not adequately summarise the objectives or findings of the analysis.
Introduction **(10%)**	An excellent introduction that provides a compelling case for the proposed analysis.	A very good introduction that offers a very convincing case for the proposed analysis.	A good introduction that furnishes a largely convincing case for the proposed analysis	An adequate introduction that offers a somewhat weak case for the proposed analysis.	A poor introduction that fails to motivate the problem or provide a case for the proposed analysis.
Related Work **(20%)**	An excellent critical analysis of substantive and relevant literature leading to compelling rationale for the proposed analyses, demonstrating a thorough knowledge of parallel algorithms and analysis on large amounts of data.	A very good critical analysis of substantive and relevant literature leading to convincing rationale for the proposed analyses, demonstrating very good knowledge of parallel algorithms and analysis on large amounts of data.	A good analysis of relevant literature leading to clear rationale for the proposed analyses demonstrating a reasonable knowledge of parallel algorithms and analysis on large amounts of data.	An adequate analysis of mostly relevant literature leading to an adequate rationale for the proposed analyses, demonstrating a basic knowledge of parallel algorithms and analysis on large amounts of data.	A review of some relevant literature but limited evidence of understanding and a weak rationale for proposed research, demonstrating a poor knowledge of parallel algorithms and analysis on large amounts of data.
Methodology **(35%)**	An excellent application of scalable programming techniques to the problem domain with a clear and comprehensive discussion on the choices made. Excellent consideration of the ethical issues pertaining to the sourcing and analysis of data. Scalability issues are very thoroughly addressed.	A very good application of scalable programming techniques to the problem domain with a largely complete and clear discussion on the choices made. Good consideration of the ethical issues pertaining to the sourcing and analysis of data. Scalability issues are thoroughly addressed.	A good application of scalable programming techniques to the problem domain with a reasonably clear and comprehensive discussion on the choices made. A reasonable degree of consideration of the ethical issues pertaining to the sourcing and analysis of data. Scalability issues are addressed to a reasonable extent.	An adequate application of scalable programming techniques to the problem domain with some minor flaws, accompanied by an incomplete discussion on the choices made. Some consideration of the ethical issues pertaining to the sourcing and analysis of data.	A poor or non-existent methodology with scant or no discussion on the choices made. Very little consideration of the ethical issues pertaining to the sourcing and analysis of data. Scalability issues are not addressed.

Grading Rubric (continued)

Criterion	**H1** $\geq$ 70%	**H2.1** $\geq$ 60% < 70%	**H2.2** $\geq$ 50% < 60%	**Pass** $\geq$ 40% < 50%	**Fail** < 40%
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 the 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 choices of visualisation types and poor implementation.
Conclusions and Future Work **(10%)**	An excellent discussion of the implications and limitations of the work. An excellent consideration of potential research impact/outcomes.	A very good discussion of the implications and limitations of the work accompanied by a considerable discussion of potential research impact/outcomes.	A good consideration of the implications and limitations of the work accompanied by a reasonable discussion of potential research impact/outcomes.	Adequate but incomplete consideration of the implications and limitations of the work accompanied by a passable discussion of potential research impact/outcomes.	Little or no consideration of the implications and limitations of the works. Scant discussion of potential research impact/outcomes.