# Data Intensive Architectures (H8DIA)
# MSc in Data Analytics (MSCDAD)- All Instances
# Project (100%)

**Programme:** MSc in Data Analytics
[***MSCDAD_A_JAN24I, MSCDAD_B_JAN24I, MSCDAD_B_JAN24I***]

## General Instructions

- This project is an individual assessment that is worth 100% of the final grade awarded. The deadline for submission of this assessment is divided into two stages (given below) which will be marked individually out of 100 marks and weighted as per below:

  1. **Stage I [20%]**: Tuesday, 9th April 2024 @ 23:55 hrs
  2. **Stage II [80%]**: Wednesday, 15th May 2024 @ 23:55 hrs

- The documents must be submitted as PDF document to Moodle before the deadline (mentioned above). The report should be concise, with the main part of the report (excluding references and appendix), limited to the no. of pages as mentioned in sections 2.1 and 2.2 in a typical 1-column format with a paragraph font size of 12 pt. Include student name, student ID, and course name at the top of the first page.

  Late submissions will not be penalized if the student applied for an extension through NCI360 and it was approved. Use an **IEEE Referencing Layout** for this submission.

- The *project cover sheet* (provided on Moodle) must be duly filled out (all sections) and attached to both submissions. Please note that submissions that don't have the cover sheet attached will **NOT BE MARKED**.

- **Turnitin:** All submissions will be electronically screened for evidence of academic misconduct (plagiarism and collusion).

# 1 Tasks

## 1.1 Datasets

Programmatically analyze and interrogate 2 (or more) datasets. Your datasets should fulfil the following minimum requirements:

- Be related in some way.
- Complement each other such that your study (or something very similar) could not be conducted without one of your datasets.
- Be at least moderately sized for your project to be considered "data-intensive".
- Whilst there is no upper limit on size, be realistic with respect to the capabilities of your cloud instance(s) and processing times.
- Be ethically sourced and employed [1].

## 1.2 Procesing

In terms of what to do with your datasets, please observe the following minimum requirements:

1. Programmatically prepare your datasets, which includes:
   (a) Extracting them from open/free data repositories (placing them potentially into your own block/blob storage or similar). There is a huge number of online repositories of relevant data, where some examples include, but are not limited to:
      - UK's open government data repository: https://data.gov.uk
      - Central Statistics Office, Ireland: https://www.cso.ie
      - Ireland's open government data repository: https://data.gov.ie
      - Kaggle: https://www.kaggle.com
      - Run My Code: https://www.runmycode.org
      - Amazon's public dataset repository: https://aws.amazon.com/datasets
      - Google's Public Data Directory: https://www.google.com/publicdata/directory
      - The UCI machine learning repository: https://archive.ics.uci.edu/ml/
      - Zenodo: https://zenodo.org
      - Dublinked: https://data.smartdublin.ie
      - Data.gov: https://www.data.gov
      - Quandl: https://www.quandl.com

   (b) **Data Cleaning:**Detail the specific techniques and tools used for data cleaning, and the rationale behind choosing them.

   (c) **Data Transformation and Combination:** Elaborate on the methods of data conformation and transformation to standardize and merge datasets from different sources. Highlight the challenges and strategies in combining datasets with varying structures or formats.

   (d) **Exploratory Study:** Outline how a preliminary exploratory analysis is conducted to guide the project's focus and provide a formal description of the data characteristics.

   Include visualizations that offer initial insights and identify potential areas of interest.

   (e) **Data Analysis Preparation:**
      i. Detail the preparation of the data for a complete and comprehensive analysis.

ii. Discuss any preprocessing steps, such as normalization or feature selection, that are necessary for the chosen analysis methods.

(f) **Scalability and Performance Considerations:**

i. Address the scalability challenges, especially those arising from large dataset sizes, and the approaches to handling them.

ii. Discuss the implementation of speedup techniques, possibly parallel processing or efficient data structures, to optimize performance.

2. Perform analysis using the MapReduce programming model [2]. Provide an in-depth explanation of how the MapReduce programming model is applied to analyze the data.
3. Summarize at least three significant insights obtained from interrogating the combined dataset and MapReduce results. Discuss how these insights contribute to the understanding of the data and the overarching questions the project seeks to answer.

Some ideas about possible projects can be found through the case studies reported in [3].

Your project **MUST** explicitly address Data Quality as defined by the ISO/IEC 25012 standard here.

# 2 Deliverables

## 2.1 Stage 1: Proposal/Inerim Report Progress (20%)

The proposal should pitch your project idea including the topic of interest and motivation, objectives, a brief description of the datasets, and an overview of the methods intended to be used to develop the project and extract the insights to achieve the objectives. The proposal report must follow the IEEE conference format and should be between 4-5 pages in length (this includes all figures, but not references). For this proposal IEEE referencing style must be used. Microsoft Word and LaTeX templates are available at `http://www.ieee.org/conferences_events/conferences/publishing/templates.html`. Table 1 shows a suggested structured for the proposal report.

Table 1: Proposal / Interim Progress Report sections structure.

| **Section** | **Weight** | **Length Limit** | **Description** |
|---|---|---|---|
| Motivation and Introduction | 25% | Max 500 words. | Motivate your choice of topic. |
| Research Questions & Objectives | 25% | Max 200 words. | What questions do you plan to answer? |
| Data Sources | 10% | Max 200 words. | Source of data and brief description. |
| Distributed Methods | 20% | Max 400 words | Brief description of the proposed distributed methods. |
| Evaluation Methods | 20% | Max 300 words | Brief description of the methods to evaluate the results and answer the research questions. |
| Bibliography | | Exempt from length limits. | References in IEEE referencing style |

## 2.2 Stage II: Final Report, Source-Code, and Datasets (80%)

### 2.2.1 Final Report

The project report should discuss the challenges that you encountered whilst handling your chosen datasets and the means and mechanisms you implemented to overcome these challenges.

The project report should be 10-12 pages in length in IEEE single-column format including all figures (excluding bibliography/references). Please refer to this link for the formatting requirements and LaTeX& Word templates: https://www.ieee.org/conferences/publishing/templates.html. The structure suggested for the report is:

- **Abstract:** A 150-250 words providing a high-level of the project, its core findings and key results, and the domain of the datasets (not necessarily in this order)
- **Introduction:** Begin with a brief overview of the current landscape of data-intensive architectures and their significance in handling vast amounts of data.
  - **Project Objectives:** Clearly articulate the objectives of the project. This might include understanding specific patterns within large datasets, improving data processing methodologies, or exploring the potential of data-intensive architectures to solve real-world problems. Specify what the project aims to discover or achieve.
  - **Scope and Challenges:** Outline the scope of the project, including the types of datasets being analyzed and the specific domain (such as healthcare, environmental science, social media analysis, etc.) it focuses on. Discuss the anticipated challenges in managing, processing, and analyzing these datasets, including issues related to data quality, volume, velocity, and variety.
  - **Innovation and Importance:** Highlight the innovative aspects of the project, such as the use of cutting-edge data processing frameworks, novel analytical methods, or the exploration of new datasets. Explain why this project is important for the field of data science and what new insights or advancements it hopes to contribute.
  - **Structure of the Paper:** Briefly describe the structure of the rest of the paper.
- **Data:**
  - **Dataset Overview:** Describe the chosen datasets, including their sources, sizes, and types. Mention the diversity of the data and its relevance to the project's objectives.
  - **Data Selection Criteria:** Explain the criteria used for selecting the datasets, such as relevance to the research question, data integrity, and completeness. Highlight how these datasets complement each other and are essential for the comprehensiveness of the study.
  - **Brief Literature Review:** Analyze how similar datasets have been applied in existing research. Identify methodologies, challenges, and solutions from these studies to contextualize your approach and enrich your analysis.
- **Methodology:**
  - **Data Collection & Preprocessing:** Detail the comprehensive strategy for acquiring, cleaning, and preprocessing data from multiple sources, ensuring it is suitable for analysis. Discuss the criteria for data selection, such as relevance, accuracy, and completeness, and the techniques used for cleaning and preprocessing to ensure consistency and reliability.
  - **Analytical Approach:** Elaborate on the use of the MapReduce programming model for processing vast datasets. Discuss how this model facilitates parallel processing of large data sets across distributed systems, significantly reducing computation time and increasing efficiency. Provide examples of specific algorithms or processes that will be implemented using MapReduce, such as text analysis or pattern recognition.
- **Implementation and Architecture:**
  - **System Design:** Dive deeper into the architectural design of the system, emphasizing its scalability and robustness. Describe how the architecture supports the efficient processing and analysis of large datasets, including the use of cloud-based storage and computing resources. Discuss the rationale behind the choice of architecture, focusing on its ability to adapt to varying data volumes and computational needs.
  - **Tool Selection and Workflow:** Provide a detailed overview of the open-source tools and custom scripts that form the backbone of the data processing workflows. Explain how these tools are integrated into the system architecture and how they contribute to the project's objectives. Examples might include data

ingestion tools, distributed computing frameworks like Apache Hadoop or Spark, and data analysis libraries.

- **Results:**
  - Summarize the main insights, noting any unexpected patterns versus expected outcomes. Discuss how these findings respond to the project's main objective.
  - Highlight any surprising discoveries and confirmations of initial hypotheses, emphasizing their relevance to your research.
  - Evaluate how the results address the introductory motivational question, detailing the implications of your findings.
  - Mention interesting aspects of your results and describe major challenges overcome during the analysis. Conclude with the importance of these results for your field and future research directions.
- **Conclusions and future work:** Conclude by summarizing the primary insights gained and the overall findings of your project. Reflect critically on the project's execution: consider what could have been done differently to enhance the outcomes or address challenges more effectively.

  For future work, articulate a clear plan on how to build upon your current findings with additional research or by incorporating new methodologies or data sets. Consider the implications of your work for the field and propose specific next steps that could be taken if provided with more time or resources.ject) what would you do next to extend your work?
- **References:** A complete list of academic works or online materials used in the project. References should be included as in-text citations according to the IEEE citation style. To find academic works and citation style guidelines, please refer to the NCI Library guide for Data Analytics: http://libguides.ncirl.ie/dataanalytics

### 2.2.2 Source-Code and Datasets

All source code and datasets that contributed to the findings presented in this report should be compiled into a single compressed ZIP file for submission. To enhance the comprehensibility and usability of the submitted materials, include a detailed README file within the ZIP archive. This README should cover the following:

1. **Installation Requirements:** List all necessary software, libraries, and dependencies required to execute the code, along with version numbers where applicable.
2. **Setup Instructions:** Provide step-by-step instructions for setting up the environment, installing the required dependencies, and any additional configuration steps.
3. **Execution Guide:** Detail the commands or steps needed to reproduce the results, including how to run the code and any arguments or parameters that can be modified.
4. **Troubleshooting Tips:** Include common issues that might arise during the setup or execution process and their solutions, to assist users in resolving potential problems efficiently.

### 2.2.3 Marking Grid

Worth 80% of the final mark, the final report will be graded using the marking grid shown in Table 2.

Table 2: Grading Rubric: Data Intensive Architectures (H8DIA)

| **Assessment Criteria** | **H1 (>70%)** | **H2.1 (>60%)** | **H2.2 (>50%)** | **Pass (>40%)** | **Fail (<40%)** |
|---|---|---|---|---|---|
| **Project Objectives (20%)** | Challenging project objectives are well presented, met, and thoroughly discussed. | Reasonable project objectives are clear, and at mostly met. | There are clear objectives, which are at least partially met. | There are some objectives, which are at least partially met. | Cannot discern project objectives, and/or if project objectives were met. No obvious development conducted. |
| **Datasets** (10%) | The datasets have been well prepared and explored. At least two datasets have a high degree of complexity. | The datasets have been prepared and meaningfully explored. At least one datasets has some degree of complexity. | The datasets have been somehow prepared and explored. At least one datasets is non-trivial. | The datasets are prepared and probably somewhat trivial. | Less than two datasets. |
| **MapReduce Design, Methods, Analysis, and Ethics**(40%) | Excellent/very good application of MapReduce design principles in terms of appropriate: methodology; methods for generating and analysing data; and consideration of any ethical issues. | Good application of MapReduce design principles in terms of appropriate: methodology; methods for generating and analysing data; and consideration of any ethical issues. | Adequate application of MapReduce design principles in terms of appropriate: methodology; methods for generating and analysing data; and consideration of any ethical issues. | Weak application of MapReduce design principles and limited evidence of understanding of: appropriate methodology; methods for generating and analysing data; and ethical issues. | Poor application of MapReduce design principles and very limited evidence of understanding of: appropriate methodology; methods for generating and analysing data; and ethical issues. |
| **Analysis and Identified Impact/Outcomes** (20%) | Excellent/very good analysis of the results and consideration of potential research impact/outcomes. | Good analysis of the results and consideration of potential research impact/outcomes. | Adequate analysis of the results and consideration of potential research impact/outcomes. | Limited/weak analysis of the results and consideration of potential research impact/outcomes. | Very limited or poor analysis of the results and consideration of potential research impact/outcomes. |
| **Structure, Abstract, and Referencing** (10%) | Excellent/very good abstract and structure. All referencing consistent and appropriate. | Good abstract and structure. Most referencing consistent and appropriate. | Adequate abstract and structure. Adequate consistent and appropriate referencing. | Weak abstract and structure. Frequent inconsistent or inappropriate referencing. | Poor abstract and structure. Very frequent inconsistent or inappropriate referencing. |

# References

[1] A. Zwitter, “Big data ethics,” *Big data & society*, vol. 1, no. 2, p. 2053951714559253, 2014.

[2] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] J. Kołodziej, H. González-Vélez, and H. Karatza, “High-performance modelling and simulation for big data applications,” *Simul. Model. Pract. Theory*, vol. 76, 2017.