National College of Ireland
Cloud Competency Centre

Dr. Horacio González–Vélez
Associate Professor and Founding Head of The Cloud Competency Centre
National College of Ireland
Mayor Street, IFSC, Dublin 1, IRELAND
www.ncirl.ie/cloud
horacio@ncirl.ie

Assessment Guide

Data Intensive Architectures (DIA) : MSc Data Analytics : Dr. H. González–Vélez
Issued: 16th October 2020

Contents

Administrative

This module assessment guide has been designed to facilitate the understanding of the scope for the assessment of the Data Intensive Architectures (DIA) module as part of the MSc Data Analytics at the School of Computing of the National College of Ireland.
The learning outcomes of this module are:

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.

Assessment of this module is based on **TWO** coursework assignments which represent 20% and 80% of the final grade awarded respectively. As published in the School portal, the deadlines for submission of assignments are:

1. Hypothesis (20%): Report– DEADLINE: 29^{th} October 2020 9am.
2. Project (80%): Video, Code and Report– DEADLINE: 8^{th} December 2020 9am.

Should any student miss the assessment with a valid reason, (s)he can now apply for an application for coursework Extension/Re-run Form online, via NCI360. PCF forms are no longer in use and will not be accepted by the School of Computing office.
Both assignments have to be electronically submitted using the dropbox provided in the module page in Moodle. Please:

- ensure that your name in full (as per NCI official documents) and student number are clearly visible on the front page of the written reports; and
- name your files starting with the first letter of your given name followed by the first three letters of your surname, your student id, a dash, and the word “report”. No spaces or any other alphanumeric characters should be included in the filename. That is to say, when “Mary Murphy” with student id 20123456 submits her coursework report, she should name the file: `mmur20123456-report.pdf`.

N.B.

All submissions will be electronically screened for evidence of academic misconduct (plagiarism and collusion). The combined mark for the coursework represents 100% of your overall mark for this module.

Hypothesis 20%

The first assignment entails a report which provides the definition of the hypothesis to drive the final project.

(a) Deliverable

1. A short (up to 2 pages) report formatted using the IEEE double-column template. It should include the citations to the key sources using IEEE bibliographic format. Please refer to the following link for the formatting requirements and LaTeX/Word templates: `https://www.ieee.org/conferences/publishing/templates.html`.

(b) Structure

This assignment ought to include the following components:

Title Choose a title that encompasses the specific topic within the Data Intensive Architectures (DIA) area.

Question Phrase your hypothesis addressing its specifics and the measurable side.

Value Explain why this hypothesis is worth investigating and why its answer is non self-evident.

Justification Describe why the hypothesis is feasible to be pursued via a MapReduce approach, clear (measurable), and ethical. Cite the key 2-3 sources (journal articles/conference papers/patents) you are basing your hypothesis on.

References Include the citations to the key 2-3 sources using the IEEE style, as bibliography at the end of the document.

Project 80%

(a) Datasets

Programmatically analyse and interrogate 2 (or more) curated Open Data [1] datasets, i.e. they should be free to be used, re-used and redistributed by anyone and curated by a recognised European entity. Your data sets should fulfil the following minimum requirements:

1. Be related in some way.
2. Complement each other such that your study (or something very similar) could not be conducted without one of your datasets.
3. 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.
4. Be ethically employed [2].

(b) Processing

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

1. Programmatically prepare your datasets this includes:
   (a) Extracting them from well-curated Irish/European **Open Data** repositories and placing them into your own block/blob storage or similar. There is a large number of repositories with relevant data, where some examples include, but are not limited to:
      - Ireland's Open Data Portal `https://data.gov.ie/`
      - Ordnance Survey Ireland Open Data Portal `https://data-osi.opendata.arcgis.com/`
      - Dublinked: Open Data for the Dublin Region `https://data.smartdublin.ie/`
      - Central Statistics Office, Ireland: `http://www.cso.ie`
      - eHealth Ireland Open Data Portal `https://data.ehealthireland.ie/`
      - Open Data available from Fáilte Ireland `https://failteireland.ie/Research-Insights/Open-data.aspx`
      - Fingal Open Data `http://data.fingal.ie`
      - UK's open government data repository: `http://data.gov.uk`
      - European Data Portal `https://www.europeandataportal.eu/en` and the EU Open Data Portal `http://data.europa.eu/`
        Additional relevant Open Data Publishers in Ireland are listed in `https://wiki.openstreetmap.org/wiki/Ireland/Open_Data`. If additional datasets are required, you can, in exceptional circumstances, resort to public cloud providers Open Data repositories such as the Amazon's public dataset repository: `https://aws.amazon.com/datasets` or Google's Public Data Directory: `http://www.google.com/publicdata/directory`.

   (b) Clean them.

   (c) Conform/transform and combine the datasets.

   (d) Providing at least a cursory exploratory study to motivate your project focus and formally describe the data.

   (e) Prepare the data for at least one complete analysis with particular emphasis on how they interrelate.

2. Perform analysis using MapReduce [3].

3. Interrogate the combined dataset and MapReduce results to provide at least 3 interesting insights into the data you have chosen.

As discussed during lectures, some ideas about possible projects can be found through the case studies reported in [4].

N.B.

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

(c) Deliverables

1. A short video (up to 7 minutes in duration) containing a short individual oral presentation explaining the key findings of the project and a demonstration of the MapReduce code execution.

2. A zip archive of all code and datasets used to produce the results also with a (brief) explanation on how they were run in order to facilitate their rerunning if needed.

3. A 6-page report formatted using the IEEE double-column template. The report has to include all figures and any references to existing work. Please refer to the following link for the formatting requirements and LaTeX/Word templates:
   `https://www.ieee.org/conferences/publishing/templates.html`.

(d) Structure of the Report

All relevant findings should be compiled into an accompanying report, which should be submitted along with any programming code elements and the video demo. Your 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. It should be structured as follows:

Abstract: a roughly 200-word executive summary of the project and the key results

Introduction: set the scene of the project, i.e., the objectives of the project (for example what are you trying to find out)

Data: describe your datasets including metadata elements, the format the datasets are represented with, where the datasets can be retrieved from and their licensing, how they were generated, when they were generated (ideally have records from 2019 and 2020), and who generated them.

Methodology: essentially, how have you addressed Data Quality in your project?. Conceivable here, you would also discuss how other people have used the datasets you have chosen. Ethical consideration on the datasets and their intended use should also be integrated here.

Implementation and Architecture: how have you built your application workflow, what components and/or forms of analytics have you used and why?

Results: what did you find out about your data sets? e.g.: what was surprising? what was expected? what did you find out with respect to your motivational question that is presented in the introduction? Finally discuss any interesting aspects of your results or key challenges you solved in achieving your results.

Conclusions and future work: what (in general) did you learn and find out? If you were to do the project again, what would you do differently? If you had more time (e.g. in your final project) what would you do next to extend your work?

References: a complete list of academic works and/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`

(e) Marking Grid

MARKING GRID– Data Intensive Architectures (DIA) , Assessment Guide, Dr. Horacio González–Vélez Due date: 8^{th} December 2020 .

ASSESSMENT CRITERIA	*EXCELLENT / VERY GOOD*	*GOOD*	*SATISFACTORY*	*THRESHOLD*	*FAIL*
Project Objectives, Datasets, and Ethics: (40% weight)	Challenging project objectives are well presented, met, and thoroughly discussed. Datasets have been well prepared and explored. At least two datasets have a high degree of complexity. Comprehensive consideration of ethical issues	Reasonable project objectives are clear, and at mostly met. Datasets have been prepared and meaningfully explored. At least one datasets has some degree of complexity. Reasonable consideration of ethical issues.	There are clear objectives, which are at least partially met. Datasets have been somehow prepared and explored. At least one datasets is non-trivial. Some consideration of ethical issues.	There are some objectives, which are at least partially met. Datasets are prepared and probably somewhat trivial. Scant consideration of ethical issues	Cannot discern project objectives, and/or if project objectives were met. Less than 2 datasets. No obvious development conducted. No consideration to any ethical issues
MapReduce Code design, methods, and analysis. (30% weight)	Excellent/very good application of MapReduce design principles in terms of appropriate: methodology; and methods for generating and analysing data.	Good application of MapReduce design principles in terms of appropriate: methodology; and methods for generating and analysing data.	Adequate application of MapReduce design principles in terms of appropriate: methodology; and methods for generating and analysing data.	Weak application of MapReduce design principles and limited evidence of understanding of: appropriate methodology; and methods for generating and analysing data.	Poor application of MapReduce design principles and very limited evidence of understanding of: appropriate methodology; and methods for generating and analysing data.
Identified impact/outcomes, structure, proposal abstract, and referencing. (30% weight)	Excellent/very good consideration of potential research impact/outcomes. Excellent/very good abstract and structure. All referencing consistent and appropriate	Good consideration of potential research impact/outcomes. Good abstract and structure. Most referencing consistent and appropriate.	Adequate consideration of potential research impact/outcomes. Adequate abstract and structure. Adequate consistent and appropriate referencing.	Limited/weak consideration of potential research impact/outcomes. Weak abstract and structure. Frequent inconsistent and/or inappropriate referencing.	Very limited and poor consideration of potential research impact/outcomes. Poor abstract and structure. Very frequent inconsistent and/or inappropriate referencing.
	70-100	*60-69*	*50-59*	*40-49*	*<40*

THE FINAL MARK MUST BE 40% OR ABOVE TO ACHIEVE A PASS

References

[1] R. Kitchin, The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences. London: Sage, 2014. ISBN: 978-1-4462-8747-7.

[2] A. Zwitter, "Big data ethics," Big Data & Society, vol. 1, no. 2, p. 2053951714559253, 2014.

[3] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107–113, Jan. 2008.

[4] J. Kolodziej and H. González-Vélez, eds., High-Performance Modelling and Simulation for Big Data Applications - Selected Results of the COST Action IC1406 cHiPSet, vol. 11400 of Lecture Notes in Computer Science. Springer, 2019. (Open Access)–ISBN 978-3-030-16271-9.