Big Data Processing

Project Brief

M.Sc. in Artificial Intelligence

Department of Artificial Intelligence, Faculty of ICT

2022/23 - Semester 2

Study Unit Code	ICS5114 - Big Data Processing
Assessment	100% of marks.
Deadline	Tuesday 23rd May 2023 at 15:00 CEST
Type	Individual Project.

Read the instructions carefully, thoroughly and completely before starting to work on this project. Failure to comply with all the requirements can result in unnecessary loss of marks or potential disqualification. Any questions or requests for clarifications about this project need to posted on the dedicated forum on VLE. Before asking a question, make sure it has not already been asked by other students.

Students will work on their own individually. The work submitted for this project needs to be exclusively done for this study unit, and not reused or recycled from other study units. Make sure to submit your assignment before the specified deadline. **Late submissions will not be accepted.**

1 Introduction

Climate Change is a recurring topic that raises numerous concerns regarding droughts and desertification, the frequency and severity of natural disasters, destabilisation of weather systems and seasons, sea-level rise, and extinction of species. It also threatens to impact humans directly, putting towns, cities and in some cases countries at risk of frequent infrastructural damage, famine, economic impact and high energy costs in order to counteract extreme cold or heat spells.

For this project, you can make use of any data sets available from reputable sources, to analyse and visualise the effects of climate change. Some of the data sources you can use include:

1. Copernicus Climate Data Store[1]
2. UN's Intergovernmental Panel on Climate Change (IPCC) - Climate Change Data[2]
3. NASA's Earth Data[3]
4. NASA's Goddard Institute for Space Science Data[4]

[1] https://cds.climate.copernicus.eu
[2] https://ipcc-browser.ipcc-data.org/browser/search
[3] https://earthdata.nasa.gov
[4] https://data.giss.nasa.gov

For this project you are required to make use of the techniques and technologies discussed in the lectures to stream, process and analyse data so that you discover, demonstrate and visualise different aspects of climate change. Possible ideas include:

- Correlation between different pollutant levels and global temperature rise.
- Trends in ice sheet retraction, sea level and ocean heat.
- Water scarcity and areas suffering from drought and desertification.
- Effects on weather systems, and the frequency of extreme weather events such as hurricanes and typhoons.
- Societal and economical costs of climate change.

You are expected to **analyse historical trends** and also **predict** how the future is expected to be for any of the chosen aspects of climate change.

2 Technical Requirements

For this project you are required to use both **stream processing** and **batch processing** for the data sets you choose to use. You are expected to:

1. Feed the data through a **message broker** such as Apache Kafka.
2. **Stream process** the data using technologies such as Apache Flink, Apache Storm or Spark Streaming.
3. If applicable, use **graph processing** tools to model the relationships between data entities.
4. **Store** the data and perform analysis and modelling, using any appropriate Machine Learning techniques to generate insights into what information the data is telling us.
5. If applicable, use **Map Reduce** techniques to parallelise the processing and analysis of data using a distributed system.
6. **Visualise** the information appropriately in a clear and understandable way, showing what the effects of climate change are, and what we can expect in the coming years.

3 Deliverables

You are expected to deliver **three (3) components**:

1. A working self-contained technological artifact.
2. A report.
3. A demo presentation.

3.1 Technological Artifact

You must submit the code and a self-contained, packaged, working solution that runs easily without assuming any prerequisites or lengthy installation procedures apart from the basics. It is highly recommended to package your solution using Docker[5] to ensure your setup is reproducible with minimal effort. Provide a README file with any instructions how to run the project, including clear, tested and reproducible steps. For example, if part of your solution includes making use of a Jupyter Notebook[6] to show the code that visualises the results, state clearly which modules are needed for your project to run, and how they need to be installed.

3.2 Report

You must submit a report of around **3000 words** ($\pm 10\%$), excluding references and signed declaration forms. Follow the ACM Conference Proceedings[7] template. An Overleaf template[8] is also available. The report should include the following sections:

Abstract: Describe the challenge you are solving, the proposed solution and the achieved results. *Limit this to a maximum of 200 words.*

Introduction: Explain the challenge, the motivation behind it, and what are the objectives.

Related Work: Describe what techniques are available to address the challenge you are trying to solve, together with their strengths and weaknesses.

Methodology: Provide all the details of your solution, justifying any technological, design or implementation choices.

Testing and Evaluation: Explain how did you test the solution and how you verified the results. Include any discussion about the results you obtained. Evaluate your solution in terms of how your architecture satisfied the requirements, including any strengths and weaknesses.

Concluding Remarks: Re-establish what your objectives were and how you achieved them in this project. Include any future work you think could take this project further.

References: Make sure you include any academic references and links to material you used, giving appropriate credit. Citations and references should follow the ACM style as specified by the template.

3.3 Demo Presentation

This will be of approximately 20 minutes (15 min + questions) and will be done at a date to be established later.

4 Submission

Submission of both the technological artefact and the report will be on **VLE**.

The **technological artifact** will be submitted as a zip file. Be mindful of file size limits on VLE (maximum 100Mb). A special concession will be made if the submission exceeds this size

[5] https://www.docker.com
[6] https://jupyter.org
[7] https://www.acm.org/publications/proceedings-template
[8] https://www.overleaf.com/gallery/tagged/acm-official

limit, especially due to the size of the data. In this case, your zip file should include a README file with a Github[9] link and any relevant instructions. To avoid involuntary plagiarism or collusion, make sure your Github project is **not publicly available before the deadline**. Also make sure you **do not commit any further changes to the project after the deadline**.

The **report** will be submitted via **TurnItIn**. Make sure it has a low plagiarism score (excluding references) to avoid being penalised. You should also include the filled and signed **Plagiarism and Collusion declaration form[10]**, with your report.

The **presentation** does not need to be submitted, but will be delivered in person or remotely online via Zoom, depending on the circumstances of the time.

5 Plagiarism and Collusion

Plagiarism and collusion are taken very seriously and any suspicion of copying or claims of the work of others to be one's own will be thoroughly investigated. If confirmed, severe action will be taken according to the regulations of the University of Malta. For more details about what constitutes plagiarism or collusion refer to the University of Malta guidelines[11].

6 Marking Criteria

Component	**Marks**	**Criteria**
Technological Artifact	45%	Working project without errors. Use of appropriate technologies for both stream processing, batch processing and data storage. Delivers information that clearly communicates trends and predictions extracted from the processed data.
Report	40%	Clear and unambiguous language. Academic writing style. Proper citations and referencing. Sound motivation and objectives. Comprehensive review of existent work. Detailed methodology with sound justifications for any choices or decisions taken. Comprehensive testing and good evaluation. Concluding remarks that discuss how the results achieve the set objectives and how the work can be taken further.
Demo Presentation	15%	Concise and engaging presentation. Good visuals and explanations of what the project is about and how the objectives were achieved. Good communication skills and delivery. Live demonstration of solution. Good answers to questions that demonstrate the acquired knowledge.

[9] https://github.com

[10] https://www.um.edu.mt/ict/students/formsguidelines/

[11] https://www.um.edu.mt/__data/assets/pdf_file/0007/436651/UniversityGuidelinesonPlagiarism.pdf