ROBERT GORDON UNIVERSITY ABERDEEN

Assessment Brief

Academic Year	**2025/26**
Semester	**01**
Module Number	**CMM705**
Module Title	**Big Data Programming**
Assessment Method	**Coursework**
Deadline (time and date)	**December 16th 11:00PM IST**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle.** Failing to participate in the Viva will be considered as a non-submission
Word Limit **(see Assessment Word Limit Statement)**	**No limit**
Module Co-ordinator	**Ramindu De Silva**

What knowledge and/or skills will I develop by undertaking the assessment?

This coursework examines students' ability to design, implement, and deploy a big data analytics solution. This exercise encourages students to engage in activities such as designing the solution architecture, selecting the required technologies/products to realise the solution architecture and achieve systems functional and non-functional goals, implementing and deploying the solution, and presenting big data analytics results. The course work artefacts are uploaded to the course work as well

On successful completion of the assessment students will be able to achieve the following Learning Outcomes:

1. Critically assess a big data problem and provide a robust solution

2. Employ batch analytics techniques effectively

3. Using existing machine learning algorithms to solve a problem

4. Being able to present data/analysis findings using appropriate means

Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Task(s) - content

Dataset

You are given a comprehensive dataset on Sri Lanka's historical weather, featuring meteorological data from 2010 to June 2024 for all districts of the island. The dataset is provided in two separate files: one containing the historical weather observations and another with the geographical and location details. The weather data file includes a wide array of parameters such as temperature, precipitation, and wind speed. The separate location file provides a unique identifier to link the weather data with specific city names, latitudes, longitudes, and elevations. This data provides a detailed record of weather observations for each district over a significant period. Discover and analyze trends from this data to understand how climate patterns and meteorological events vary across different regions and over time.

Task 1 - Designing a Solution Architecture

The project's architecture needs to be designed to handle both continuous data streams and historical, batch data. It is needed to optimize efficiency and avoid data duplication. Assume that the systems receive data in real time.

[1] System diagram with proposed big data tools, to collect data from the systems in real-time, store them in scalable storage, and process periodically to produce summarised information to visualise in a dashboard.

[2] Describe the role of each component and how the overall architecture operates.

Task 2 - Data Analysis

In this section, you are developing an analytics app for weather analysts to understand how weather and locations have been behaving. This analysis provides high-level summarizations and insights allowing the weather analyst and the authorities to understand how the weather has changed over period and identify areas which needed attention and during which period. This section requires the application of three distinct frameworks for different analytical tasks

[1] Analyse the following using Hadoop MapReduce

1. Calculate the total precipitation and mean temperature (take temperature_2m_mean) for each district per month over the past decade.

   Example:

   Gampaha had a total precipitation of 30 hours with a mean temperature of 25 for 2nd month

   Colombo had a total precipitation of 24 hours with a mean temperature of 30 for 3rd month
2. The month and year with the highest total precipitation in the full dataset

   2nd month in 2019 had the highest total precipitation of 300 hr

[2] Analyse the following using Hive or Pig

1. Rank the top 10 most temperate cities across the dataset (use temperature_2m_max (°C))

What is expected of me in this assessment?

2. Calculate the average evapotranspiration for each major agricultural season (September to March and April to August) in each district over the years..

[3] Analyse the following using Spark

1. Calculate the percentage of total shortwave radiation which is more than 15MJ/m² in a month across all districts
2. The weekly maximum temperatures for the hottest months of an year

Task 3 - Performing Machine Learning model using Spark MLlib

determine the expected amount of precipitation_hours, sunshine, and wind_speed that would lead to a lower amount of evapotranspiration for the month of May. The model must be trained and validated on a large dataset split, typically 80% for training and 20% for validation.

Clearly state the steps you have followed for your analysis, including data preparation, feature selection, and model evaluation.

Example: Predict the mean precipitation_hours, sunshine, and wind_speed during month of May on 2026 to have evapotranspiration lower than 1.5mm

Use 80% of the data for training and 20% of the data for validation. Clearly state the steps you've followed for your analysis.

Task 4 - Presentation of the analysis

A static web page or a presentation of data using some other visualisation tool to view the results gathered from your analysis

1. Most precipitous month/season for each district across different periods of the year.
2. Top 5 districts based on the total amount of precipitation.
3. Percentage of months that had a mean temperature above 30°C in a single year.
4. The total number of days with extreme weather events, defined by a combination of high precipitation and high wind gusts.

This can be a static web page/dashboard with hard-coded values obtained from your analysis, it's NOT required to dynamically fetch data and display.

The main assessment criterion is how well the data is presented in an understandable manner.

Also address UX, and UI aspects in your design when presenting the data.

Task(s) - format

Deliverables

You will be required to submit the following two deliverables to Campus Moodle.

- A report compiled including the details mentioned above and with a properly completed cover sheet should be submitted in **PDF format** (please DON'T zip your report or upload any other format).
- The report can contain code segments and explanations

What is expected of me in this assessment?

- A .zip archive covering the **source code** of all your implementations, including your
  - Java files (Map Reduce)
  - Text file (Hive/Pig)
  - Zeppelin/ Jupyter/ etc. Notebooks or scripts (Spark and ML)
  - HTML, CSS, and JavaScript files. (Dashboard)

This should be submitted to **.zip** format (please DON'T use other archive formats). Don't include the data set, built jar files, or any other artefacts.

Report Format

Your report should include,

- For part 1:
  - Deployment architecture to collect, analyse the data and present the results to end-users (show the software/tools that can be used for implementing the solution)
  - And reasoning on the proposed architecture and why each technology tool was selected
- For part 2, for each question:
  - The final result/output. In case the result consists of many rows add only the topmost part of the output.
  - And the code listings of the implementation.
- For part 3:
  - Step-by-step breakdown of the steps followed.
  - For each step include code listings and screenshots.
- For part 4:
  - Screenshots of the dashboard

How will I be graded?

A grade will be provided for each criterion on the feedback grid which is specific to the assessment.

The overall grade for the assessment will be calculated using the algorithm below.

A	At least 50% of the feedback grid to be at Grade A, at least 75% of the feedback grid to be at Grade B or better, and normally 100% of the feedback grid to be at Grade C or better.
B	At least 50% of the feedback grid to be at Grade B or better, at least 75% of the feedback grid to be at Grade C or better, and normally 100% of the feedback grid to be at Grade D or better.
C	At least 50% of the feedback grid to be at Grade C or better, and at least 75% of the feedback grid to be at Grade D or better.

How will I be graded?	
D	At least 50% of the feedback grid to be at Grade D or better, and at least 75% of the feedback grid to be at Grade E or better.
E	At least 50% of the feedback grid to be at Grade E or better.
F	Failing to achieve at least 50% of the feedback grid to be at Grade E or better.
NS	Non-submission.

Feedback grid *Add more rows/criteria if necessary, up to a maximum of 8.*

GRADE	A	B	C	D
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance
Task 1 (2 subgrades)	The system architecture covers all realtime, batch, interactive, and predictive analytics aspects. Non-functional requirements such as scalability, availability, and security have been taken into consideration when designing the system Correct and optimal application of technologies and tools to implement different functional areas Role of each component of the system architecture is well defined.	The system architecture covers all realtime, batch, interactive, and predictive analytics aspects. Non-functional requirements are taken into consideration with some space for improvement Correct application of technologies and tools to implement different functional areas with the possibility of optimising Role of each component of the system architecture is well defined.	The system architecture covers all realtime, batch, interactive, and predictive analytics aspects. Non-functional requirements are not addressed adequately Correct application of technologies and tools to implement some functional areas while other functional areas could have been implemented better with different technologies/tools Role of each component of the system architecture is well defined.	The system architecture covers only real-time and batch analytics, while less focus on interactive, and predictive analytics aspects. Non-functional requirements are not addressed adequately Correct application of technologies and tools to implement some functional areas while other functional areas could have been implemented better with different technologies/tools
Task 2 (4 subgrades)	All steps followed in the data analysis are provided with clear evidence Logics implemented for the data analysis are correct and have taken an efficient/optimal approach to reaching the answer	All steps followed in the data analysis are provided with clear evidence All logic implemented for the data analysis is correct but has not taken an efficient approach to reach the answer All Produced final outputs are accurate and complete.	All steps followed in the data analysis are provided with clear evidence Majority of the Logics implemented for the data analysis are correct while some have flaws Majority Produced final outputs are accurate and complete.	All steps followed in the data analysis are provide with clear evidence Logics are implemented for the data analysis but most of the logics have flaws Outputs are produced bu not accurate mostly

GRADE	A	B	C	D
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance
	All Produced final outputs are accurate and complete.			
Task 3 (1 subgrades)	All steps including data extraction, encoding, training and validating are well documented and followed Has used correct and efficient ML algorithms to produce an ML model after evaluating multiple different algorithms The ML model produced is cross-validated against the data	All steps including data extraction, encoding, training and validating are well documented and followed Has used correct but sub-optimal algorithms to produce an ML model The ML model produced is cross-validated against the data	All steps including data extraction, encoding, training and validating are well documented and followed Has used correct but sub-optimal algorithms to produce an ML model Model cross-validation is not done.	Some steps out of data extraction, encoding, training and validating ar documented and followe Has produced a ML mod that is sub-optimal or no performing the expected functionality
Task 4 (1 subgrades)	All the given data are presented following the correct UI and UX concepts All graphs/widgets that are used to represent each data item make it easy for the user to easily consume the information	All the given data are presented following the correct UI and UX concepts mostly Majority of graphs/widgets that are used to represent each data item make it easy for the user to easily consume the information	Some of the given data are presented following the UI and UX concepts Some of graphs/widgets that are used to represent each data item make it easy for the user to easily consume the information	UI and UX concerns are not addressed Some of graphs/widgets that are used to represer each data item make it easy for the user to easil consume the informatior

Coursework received late, without valid reason, will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.

ROBERT GORDON UNIVERSITY ABERDEEN

What else is important to my assessment?

What is plagiarism?

"Plagiarism is the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student's work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source" (RGU 2022).

What is collusion?

"Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately" (RGU 2022).

For further information please see Academic Integrity.

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement lists what is included and excluded from the word count, along with the penalty for exceeding the upper limit.

What if I'm unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What else is important to my assessment?

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation 'A4: Assessment and Recommendations of Assessment Boards' sets out important information about assessment and how it is conducted across the University.