University of Lincoln Assessment Framework
Assessment Briefing Template 2024-2025

1. Module code & title	**CMP9781M Big Data Analytics and Modelling**
2. Assessed learning outcomes	LO1 - Critically evaluate and apply the theories, algorithms, techniques and methodologies involved in Big Data Analytics and Modelling
3. Assessment title	Assessment 1
4. Contribution to final module mark (%)	50%
5. Description of assessment task	**Overview**: The objective of this assignment is to discuss on the fundamentals of distributed big data analytics and the use of the Apache ecosystem. Apache ecosystem covers a number of different things needs in the realm of big data analytics and modelling, including but not limited to distributed real-time computational systems, streaming dataflow engines, distributed analytics and machine learning platforms to name a few. The following three sections are required to expand and elaborate upon the principles of big data, different components of the Apache ecosystem and aspects on how such techniques can be used in real-life problems and what for. You are required to write and submit a report where you need to provide answers to all questions, discuss how you completed the tasks, and provide snippets of the source code you've developed for these tasks. You are expected to go into sufficient depth to demonstrate knowledge and critical understanding of the relevant processes involved. Available marks are through the completion of the written report, with clear and separate marking criteria for each required report section. **Report Guidance**: You must supply a written report containing three distinct sections that provide a full and reflective account of the processes undertaken. Information on specific marking criteria for each section is available in the accompanying CRG document. You may provide several figures that demonstrate the performance of the algorithms or theoretical aspects. For all figures – primarily for section 1.3 – please provide all necessary code snippers (including basic documentation) such that

your results are reproducible. You are expected to answer all questions in each step in each section in detail, perform all analysis on your own, and provide all files (scripts and train and test datasets) in one ZIP file. You are expected to write your own Python code using Apache Spark. You are allowed to use any library for data-management, data processing and analytics that are part of the Apache ecosystem.

Section I: Description of Distributed Big Data Processing Ecosystem (20%):

Using your own words, the lecture material and any other relevant source, explain the distributed big data processing ecosystem. Your description should cover the following points at a technical level:

- The 5 V's: Volume, Velocity, Variety, Value and Veracity
- Fault Tolerance and Resilience
- Distributed File Systems (e.g. HDFS)
- Resource Manager and Scheduler
- YARN
- Apache Spark
- Resilient Distributed Datasets
- Data Lakes
- Apache Hive
- Apache Flume
- Apache Kafka
- Spark ML

Section II: Describe a real-life use case for big data analytics (30%) :

In this section, you are asked to think of and describe a real-life use case of distributed big data analytics. Make sure to reflect upon the concept of the five V's that you have had the chance to elaborate upon in Section 1.1. Topics could include but not limited to product recommendation, customer churn analysis, customer segmentation, sales lead prioritization, fraud detection, anomaly detection, etc. How can big data technologies be used in such settings? Can distributed learning revolutionize the way companies drive their decision-making processes? Please use the knowledge obtained during the lectures, workshops and your independent learning to critically approach the above problem. You may use diagrams, plots, flow charts or any other means you deem fit to provide a complete use case.

Section III: Linear Regression Analysis (50%)

Your task here is to acquire a non-trivial research dataset from any source based on your own personal interests, which should be

	acknowledged in your report, and focus on the analysis of this data. You must submit only original work which has not been submitted for any other assignment. A few indicative examples of possible datasets would include (but are not limited to): ▪ COVID-19 prediction, ▪ Restaurant revenue prediction, ▪ House prices, ▪ Pesticide use in agriculture, ▪ etc. If still in doubt, please discuss the choice of dataset with the module instructors at the earliest available opportunity. You are asked to use the selected dataset to develop a pipeline for running a simple linear regression. Please implement the following: • Load the data and pre-process the data if necessary • Split the dataset into 70% train and 30% test set • Train a linear regression model • Evaluate its performance on the train and test sets and report the mean absolute error (MAE), the root mean squared error (RMSE) and mean squared error (MSE) • For the test set used generate a table showing the predicted vs actual values as well as a predicted vs actual plot
6. Assessment submission instructions	The submission deadline of this assignment is included in the School Submission dates on Blackboard. You must make an electronic submission of your presentation report to the Turnitin upload area for Assessment 1. The report should be no more than 4000 words (excluding figures, tables, snippets of source code, and references). The report must: • Contain your name, student number, student email address, and module name • Be in PDF • Be formatted single-spaced with 11pt font size • Do not include this briefing document If you are unsure about any aspect of this assessment component, please seek the advice of the module co-ordinator Dr Miao Yu <myu@lincoln.ac.uk>

7. Date for return of mark and feedback	Please see the **Hand In Dates.xls** spreadsheet. Note: *all marks awarded are provisional until confirmed by the Board of Examiners.*
8. Feedback format	Feedback is provided on Blackboard according to CRG criteria (see CRG file). Face-to-face feedback may also be provided to students which may require further clarifications.
9. Use of Artificial Intelligence (AI) in this assessment	No AI tools are allowed to be used.
10. Marking criteria for assessment	A Criterion Reference Grid (CRG) is used to evaluate your learning against a set of pre-defined criteria.
11. Additional information (support, advice, tips etc)	Students are encouraged to use any lecture and their own personal notes to assist them with the completion of the assessment. Also, students are allowed to use any library and/or online resource as a guide on how to solve the assessment problems.
12. Important Information on Dishonesty, Plagiarism and AI Tools	University of Lincoln Regulations define plagiarism as *'the passing off of another person's thoughts, ideas, writings or images as one's own...'*. Examples of plagiarism include the unacknowledged use of another person's material whether in original or summary form. Plagiarism also includes the copying of another student's work'. Plagiarism is a serious offence and is treated by the University as a form of academic dishonesty. For more information on examples of Academic Offences, please see the **Academic Offence Guidance**. Please note, if you use AI tools in the production of assessment work **where it is not permitted**, then it will be classed as an academic offence and treated by the University as a form of academic dishonesty. Students are directed to the University Regulations for details of the procedures and penalties involved. For further information, see www.plagiarism.org