University of Lincoln
School of Computer Science
Assessment Briefing 2023-2024

The use of AI tools to generate all or part of your assessment submission is **not** permitted unless specifically mentioned below.

Module Code and Title: CMP3749M Big Data Assessment Brief

Contribution to Final Module Mark: 100%

Description of Assessment Task and Purpose:

This assessment is an **individual** assignment. As a data scientist, your main objective is to organise and analyse data regardless how big or small the data is, often employing various data science software/tools/algorithms. The analysis made by a data scientist must be easy enough to understand by all stakeholders including those who have no knowledge of data science.

The objective of this assessment is to show that you can perform an analysis over a dataset to guide the stakeholders to understand the data. The data can be downloaded from Blackboard in the assessment documents area. The data needs to be analysed using the data science tools and techniques you were taught in class.

You are required to write and submit a report where you need to provide answers to all questions, discuss how you completed the tasks outlined in the **Report Guidance** (see below). In addition, you are to provide the Python code using Pyspark to accomplish the given tasks. You are expected to go into sufficient depth to demonstrate knowledge and critical understanding of the relevant processes involved.

Note: most of the marks will stem from the clarity of your report, with the source code used as evidence.

Report Guidance:

You are to submit a single report (pdf format) and associated python source code (single zip file) containing the following two tasks:

- Task 1 – PySpark Analysis of Nuclear Plants dataset (strict max 1000 words) (40%)
- Task 2 – MapReduce for Margie Travel dataset (strict max 1000 words) (40%)
- Task 3 – Big Data Tools and Technology Appraisal (strict max 750 words) (20%)

You must split the report into three distinct sections (see tasks requirements) providing a full and reflective account of the processes undertaken to conduct the above tasks. You are expected to answer all questions in each task in detail, perform all analysis on your own (i.e., individual work). Provide all Python scripts (PySpark and MapReduce tasks) in one ZIP file uploaded to the supporting document area on Blackboard as supporting evidence of accomplishing the tasks.

The use of AI tools is not permitted in the generation of the final report for this assessment.

Task 1 – PySpark Analysis of Nuclear Plants dataset (strict max 1000 words) (40%)

You must clearly identify this part of the report as "Task 1 – PySpark Analysis of Nuclear Plants dataset".

The dataset is of pressurised water reactors (a type of nuclear reactors) with various measurements in different parts of the reactor, including vibration, pressure and power levels. The first column in the spreadsheet indicates the status of the reactor, i.e., 'normal' or 'abnormal'. All the other columns are features which could help us to gain insights into the status of each reactor. You are asked to provide an analysis over this data to discuss if these features could be potentially used to predict whether a reactor is normal or abnormal.

Download the dataset named `nuclear_plants_small_dataset.csv' from Blackboard on Blackboard under **Task_1_dataset**, then write Python code using PySpark to accomplish the following:

1- As a first step, you need to load the data from the file `nuclear_plants_small_dataset.csv' into a Pyspark DataFrame. Before making any analysis, it is required to know if there are missing values in the data. Are there any missing values? Discuss how you will deal with missing values, even if there are no missing values in this dataset. (12 marks)

2- It is beneficial to understand the data by looking at the summary statistics. There are two groups of subjects (i.e., the normal group and the abnormal group) in this dataset. For each group, show the following summary statistics for each feature in a table: minimum, maximum, mean, and median values. For each group, plot the box plot for each feature. (14 marks)

3- To understand the relationship between features. If two features have high correlations, using only one of them could be enough for analysis? Present a correlation matrix of the features and report on your observations to the correlation matrix. In addition, discuss the highly correlated features, if any, and their usefulness for further processing (e.g., data classification)? (14 marks)

Note: allocated marks are awarded for providing the required report section "Task 1 – PySpark Analysis of Nuclear Plants dataset" and accompanying source code. Check the Criterion Reference Grid for details of how the presentation will be graded.

Task 2 – MapReduce for Margie Travel dataset (strict max 1000 words) (40%)

You must clearly identify this part of the report as "Task 2 – MapReduce for Margie Travel dataset".

The dataset is for Margie's Travel (MT) provides concierge services for business travellers. In an increasingly crowded market, they are always looking for ways to differentiate themselves and provide added value to their corporate customers. There are two files containing lists of data. These are located on Blackboard under **Task_2_dataset**. The coursework data folder includes the files:

- AComp_Passenger_data_no_error.csv: this data file contains details of passengers that have flown between airports over a certain period. The data is in a comma delimited text file, one line per record, using the following format:

Col. #	Field	Format
1	Passenger id	$XXXnnnnXXn$
2	Flight id	$XXXnnnnX$
3	From airport IATA/FAA code	XXX
4	Destination airport IATA/FAA code	XXX
5	Departure time (GMT)	n [10] (Unix 'epoch' time)
6	Total flight time (mins)	n [1. .4]

X is Uppercase ASCII

n is digit 0...9

$[n. . m]$ is the min/max range of the number of digits/characters in a string.

- top30_airports_LatLong.csv: The second data file is a list of airport data comprising the name, IATA/FAA code, and location of the airport. The data is in a comma delimited text file, one line per record using the following format:

Col. #	Field	Format
1	Airport Name	X [3. .20]
2	Airport IATA/FAA code	XXX
3	Latitude	$n.n$ [3. .13]
4	Longitude	$n.n$ [3. .13]

There are two additional data input files which can be used for analysis and validation however should not be used for the final execution of the implementation:

- AComp_Passenger_data.csv: there are various errors in this data file, which illustrate a range of potential errors that could occur when handling large scale data from multiple, sometimes unreliable, sources. It is not necessary to handle this file and address these errors in your application. This file is instead provided to highlight the requirement of error handling in MapReduce applications.

- AComp_Passenger_data_no_error_DateTime.csv: it may use to convert date/time data from Unix epoch time to a human readable format for use when debugging and for validation purposes.

Write a Python code, must use MapReduce, to accomplish the following:

1- Determine the number of flights from each airport in a table. Then provide a list of any not used airports. (12 marks)

2- Create a list of flights based on the Flight id; including number of passengers, relevant IATA/FAA codes, and departure and arrival times (times converted to HH:MM format). (14 marks)

3- Calculate the line-of-sight (nautical) miles for each flight and the total travelled by each passenger, then output the passenger having earned the highest air miles. (14 marks)

Note: allocated marks are awarded for providing the required report section "Task 2 – MapReduce for Margie Travel dataset" and accompanying source code. Check the Criterion Reference Grid for details of how the presentation will be graded.

Task 3 – Big Data Tools and Technology Appraisal (strict max 750 words) (20%)

You must clearly identify this part of the report as "Task 3 – Big Data Tools and Technology Appraisal".

In this appraisal task, you are required to critically evaluate the Big Data analytics concepts, tools, and techniques employed to solve and answer the previous distinct tasks: Task 1, which involves a PySpark analysis of a Nuclear Plants dataset, and Task 2, focused on utilizing MapReduce for a Margie Travel dataset.

Your evaluation and discussion should provide an in-depth analysis and critical reflection of the methodologies used, their effectiveness, limitations, and potential areas for improvement. Your appraisal might also assess the scalability of both PySpark and MapReduce for handling large-scale data processing. Discuss whether it demonstrated efficiency in scalability or had limitations.

Learning Outcomes Assessed:

On successful completion of this component a student will have demonstrated competence in apply data science toolkits in a range of applications and solve real-world problem.

- LO1: Critically appraise and evaluate Big Data Analytics concepts, tools and techniques.
- LO2: Apply data science toolkits in a range of applications and solve real-world problem.

Knowledge & Skills Assessed:

Subject Specific Knowledge, Skills and Understanding: Literature searching, Referencing, Numeracy, Project Planning, Techniques and Skills in Data Science, Subject-specific knowledge.

Professional Graduate Skills: Independence and personal responsibility, adaptability, written communication, creativity, critical thinking, IT skills, self-reflection and life-long learning, problem solving, effective time management, working under pressure to meet deadlines.

Emotional Intelligence: Self-awareness, self-management, motivation, resilience, self-confidence.

Career-focused Skills: Big Data tools, techniques, skills and attributes required by employers, a range of problem strategies to present skills and attributes to employers.

Assessment Submission Instructions:

The final deadline for submission of this work is included in the School Submission dates on Blackboard. The final submission will must be:

1. **Report** (Tasks 1, 2, and 3)

You must make an electronic submission of your report to the Turnitin upload area.

The report must:

- Contain your name, student number, student email address, and module name.
- Be in single PDF with
  - no more than 1000 words for the section "Task 1 - Analysis of Nuclear Plants dataset"
  - no more than 1000 words for the section "Task 2 - MapReduce for Margie Travel dataset"
  - no more than 750 words for the section "Task 3 - Big Data Tools and Technology Appraisal"
- Be formatted single-spaced with 12pt font size; Do not include this briefing document.

2. Source Code

Your python (Pyspark) code should be submitted as a single zip archive, to the assessment supporting documents area on blackboard. This zip archive should contain your python code for all tasks and include code comments where appropriate to aide understanding.

All tasks elements are individually assessed. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work. Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid (CRG). Your citations and referencing should be in accordance with the University guidelines.

If you are unsure about any aspect of this assessment component, please seek the advice of the module lecturers (contact details are available on blackboard).

Date for Return of Feedback:

Please see the SoCS assessment dates spreadsheet.

Format for Assessment:

This assessment is **individual**. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work.

Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid.

Feedback Format:

Summative feedback will be provided on Blackboard according to CRG criteria (see CRG file).

Additional Information for Completion of Assessment:

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 relevant online resource as a guide on how to solve the assessment problems.

Assessment Support Information:

Students are encouraged to seek assistance from any member of the delivery team.

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.

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