Assessment Brief 2024-25

Module title	**Big Data Tools & Techniques**
CRN	41141 / 50194
Level	7
Assessment title	**Big Data Analysis Group Work**
Weighting within module	This assessment is worth 40% of the overall module mark.
Module Leader/Assessment set by	Dr Taha Mansouri, Dr Kaveh Kiani, Nathan Topping
Submission deadline date and time	**21st March 2025 4pm** The submission deadline is 21/03/2025 by no later than 16:00. Any submission received after 16:00 (even if only by a few seconds will be considered as late).
How to submit	You should submit your assessment using the submission area on Blackboard. This only needs to be submitted once as a group. Your submission should consist of your screen recording as an mp4 file, and your Databricks notebook as both an html and an ipynb file. This notebook should include all your code and must be included in your submission. Submissions without a supporting Databricks notebook will receive a score of zero.

Assessment task details and instructions

This is a group assessment task and should be completed with your allocated group members, within the groups shared with you via Blackboard. In this assessment, you will be working together as a group to complete some data analysis tasks using Databricks and Spark, on a dataset which has Big Data characteristics. Your submission will take the form of a screen recording video in mp4 format. In this screen recording, you should take us through the details of the technical implementation of your data analysis in Databricks / Spark and the results. Alongside this, you will also need to submit your Databricks notebook so we can verify your solution. We provide some guidance on what to include in your screen recording video and tools for doing this below.

For this task, you will be working with a dataset of emails sent and received by senior executives at Enron, an energy company which collapsed in 2001. The dataset was originally made public by the Federal Energy Regulatory Commission during its investigation into the company. It has since been widely used for research purposes. It has many of the characteristics of Big Data; it contains over 500,000 emails and is semi-structured. The dataset is provided in the assessment folder alongside this brief.

For this scenario, you are working as a researcher using the dataset, and you want to start by answering some data analysis questions on the data. We have provided below a list of questions, and you should pick from these. You should pick a total of **four:** one from Group A (Easy), two from Group B (Medium) and one from Group C (Hard). So for example, you could pick A1, B3, B4 and C3.

Question Group	Question Difficulty	Question
Group A	Easy	1. What is the mean number of emails sent by each sender in the dataset, i.e., the number of emails sent divided by the number of unique senders in the dataset (based on the 'From' field) 2. What is the total number of unique recipients in the dataset (based on the 'To' field, you can filter out emails which don't contain this field, or which don't appear to contain a valid email address in this field) 3. What is the number of emails with a subject line (i.e., the 'Subject' field is not missing or empty)?
Group B	Medium	1. Who are the top 10 senders by number of emails sent (based on the 'From' field), along with the number of emails they each sent 2. Who are the top 10 recipients by number of emails received (based on the 'To' field, filtering out emails not containing this info), along with the number of emails they received 3. What is the number of emails sent internally within Enron (based on both 'From' and 'To' fields containing @enron.com) 4. What are the top 10 domains by email count in the 'From' field, excluding emails with the Enron domain (e.g., @hotmail.com, @yahoo.com, etc.)? Include the number of emails from each domain in the answer. 5. What are the top 10 pairs of sender-recipient combinations (based on 'From' and 'To' fields and treating each recipient separately) along with the number of emails exchanged by them
Group C	Difficult	1. What are the word frequencies of the top 100 words in the subject lines (using the separately provided stopwords.txt document to remove stop words)? 2. Who are the top ten senders (based on the 'From' field) who received no emails themselves (based on the 'To' field) 3. Enron collapsed in December 2001. Calculate the number of emails exchanged each day of the year 2000 based on the 'Date' field and visualise this in a line graph.

You should complete the required analysis as a group using Databricks and Spark. You are free to use both RDDs and Spark DataFrames to complete the analysis, although because of the semi-structured nature of the data you may find it easier to do some of the data processing using RDDs. Once you have completed this analysis, you should create a screen recording as a group. In this scenario, the purpose of the screen recording is to share your research with other researchers.

Note, **all group members must take part in this aspect of the assignment**, and any group member not taking part will receive zero marks.

In this recording, you should:

- **Briefly** start by each stating your name so it is audible on the recording (this is so we know all group members have participated)
- With your code solution on screen, talk us through your implementation for the four questions you selected. You don't need to run the code while doing the screen recording, but you should explain what each part does and what the output means. You can assume you're explaining this implementation to someone with a similar level of technical knowledge (so, for example, you don't need to explain concepts such as RDDs, DataFrames, etc) but you should explain what each step does and why you have used the methods and functions you have.
- Throughout, make clear any assumptions you have made about the data. You are fine to make reasonable assumptions based on your own initial exploration of the data, as long as you make this clear and make some attempt to verify if the assumption seems to hold in the majority of cases. (For example, you might assume that the recipient / sender email address always includes an '@' sign and filter out any elements where this is not the case. This is a reasonable assumption to make, but you should also demonstrate that you have investigated the impact this has on the analysis, i.e., in how many instances is this not the case? Does this assumption therefore significantly impact your final result?)
- You should state and explain the result you got for all four questions you selected.
- For the final 1-2 minutes of the screen recording, each member of the group should briefly critically reflect on their contribution to the task and any learnings from working as a team on this activity.
- The overall screen recording should be around 15 minutes, and **no longer** than 20 minutes.

There are two tools provided by the university which you can use for screen recording; PowerPoint or ScreenPal. We have provided a separate document with guidance on these tools along with an example screen recording, similar to what you might provide for this assignment.

Note on reading data into a Spark DataFrame

Each row of the csv contains two columns, the second of which contains the email message. This column can contain newline characters, which can cause issues if read into an RDD or DataFrame in Spark using the default options, as each newline character

will be read as the start of a new row. You can use the following code to read the data into a DataFrame with multiline support enabled:

```
df = spark.read.csv(
    "/FileStore/emails.csv",
    header=True,         # Use the first row as the header
    inferSchema=True,    # Infer data types
    quote='"',           # Define the quote character
    escape='"',          # Escape quotes inside quoted fields
    multiLine=True       # Enable multiline support
)
```

If you subsequently want to work with the data using RDDs, you can use the below to create a RDD from the message column of the DataFrame:

```
rdd = df.select("message").rdd
```

Final Hints

- Start by investigating the structure of the email messages. You probably want to view a randomly selected sample of them to get an understanding of their general structure.
- You should be familiar with lambda functions (as used in the workshops) and also some general Python functions. In particular, len(), strip(), replace() and split() may all be useful.
- Be aware of fields which can hold more than one value. For example, the 'To' field can hold multiple comma-separated email addresses, and each one should be treated as a distinct recipient in your analysis.

Assessment Criteria	A separate rubric is provided accompanying this assessment brief. However, please note that you are **not** being assessed on your presentation skills or on the quality of your spoken English, but on the technical aspects of your implementation and your explanation / understanding of the tools and libraries you are using. There will be **one** group mark which will be applied to all members of the group present during the screen recording. Any members not participating in this session will be graded 0 for this assessment component.
Knowledge and Understanding	**Assessed intended learning outcomes** On successful completion of this assessment, you will be able to: 1. List the 5Vs which characterise Big Data 2. Discuss the application of industry standard tools and systems for working with Big Data, such as Apache

Spark and MongoDB, and describe their use in processing or storing Big Data
3. Apply your knowledge of industry standard tools to develop strategies for working with complex datasets

Practical, Professional or Subject Specific Skills

4. Use industry standard tools, such as Apache Spark, to process and analyse Big Data
5. Plan and implement a data pipeline, using appropriate techniques to process, manipulate, analyse and visualise large, complex datasets
6. Analyse messy, real-world Big Data which is semi-structured or unstructured

Employability Skills developed / demonstrated

You will develop a range of employability skills sought by employers through each assessment.

Through this assessment will have an opportunity to develop and demonstrate the following employability skills:

Skill	I	U	A	D
Communication				X
Critical Thinking and Problem Solving				X
Data Literacy				X
Digital Literacy				X
Industry Awareness				
Innovation and Creativity				
Proactive Leadership		X		
Reflection and Life-Long Learning				X
Self-management and Organisation			X	
Team Working				X

I = You will have been introduced to this skill

U = You will have developed an understanding of this skill in the context of your subject

A = You will be able to apply this skill in the context of your subject

D = You will have demonstrated an enhanced understanding and application of this skill in a wider context

Using Artificial Intelligence (AI) Tools	You can use AI tools to help you understand the concepts and theory, and as a mentor to help you develop the coding skills necessary to complete this task. However, this submission must be your own work and your own explanation and you should also briefly state at the end if you have made use of any AI tools in this task.
Word count/ duration (if applicable)	Your screen recording should be around 15 minutes, and **no longer** than 20 minutes. If your submission is longer than 20 minutes only the first 20 minutes of the video will be considered when grading your submission.
Feedback arrangements	You can expect to receive feedback 15 working days after the submission date.
Academic Integrity and Referencing	Students are expected to learn and demonstrate skills associated with good academic conduct (academic integrity). Good academic conduct includes the use of clear and correct referencing of source materials. Here is a link to where you can find out more about the skills which students need: Academic integrity & referencing Referencing Academic Misconduct is an action which may give you an unfair advantage in your academic work. This includes plagiarism, asking someone else to write your assessment for you or taking notes into an exam. The University takes all forms of academic misconduct seriously.
Assessment Information and Support	**Support for this Assessment** You can obtain support for this assessment by attending one of the weekly Drop In sessions of emailing the module leaders. You can find more information about understanding your assessment brief and assessment tips for success here. **Assessment Rules and Processes** You can find information about assessment rules and processes in the Assessment Support module in Blackboard. **Develop your Academic and Digital Skills** Find resources to help you develop your skills here. **Concerns about Studies or Progress** If you have any concerns about your studies, contact your Academic Progress Review Tutor/Personal Tutor or your Student Progression Administrator (SPA). **askUS Services**

	The University offers a range of support services for students through askUS including Disability and Inclusion Service, Wellbeing and Counselling Services. **Personal Mitigating Circumstances (PMCs)** If personal mitigating circumstances (e.g. illness or other personal circumstances) may have affected your ability to complete this assessment, you can find more information about the Personal Mitigating Circumstances Procedure here. Independent advice is available from the Students' Union Advice Centre about this process: https://www.salfordstudents.com/advice/centre
In Year Retrieval Scheme	Your assessment is/is not (please delete as appropriate) eligible for in year retrieval. If you are eligible for this scheme, you will be contacted shortly after the feedback deadline. You can find more information about this scheme in the Assessment Support module in Blackboard.
Reassessment	If you fail your assessment, and are eligible for reassessment, you will be able to find the date for resubmission on your module site in Blackboard. For students with accepted personal mitigating circumstances for absence/non submission, this will be your replacement assessment attempt. The reassessment task will be the same as the original assessment. We know that having to undergo a reassessment can be challenging however support is available. Have a look at all the sources of support outlined earlier in this brief.