Assessment Brief 2024-25

Module title	**Big Data Tools & Techniques**
CRN	41141 / 50194
Level	7
Assessment title	**Big Data Analysis & Recommender System Project**
Weighting within module	This assessment is worth 60% of the overall module mark.
Module Leader/Assessment set by	Dr Taha Mansouri, Dr Kaveh Kiani, Nathan Topping
Submission deadline date and time	**24th April 2025 4pm** The submission deadline is 24/04/2025 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. Your submission should consist of HMTL versions of your Databricks notebooks. The submission must contain adequate explanation of your coding solution within markdown cells (as well as adequate use of commenting within the code cells themselves). We have provided a sample file which shows you how to structure your submission. Your submission will be two HTML files: one for Task 1 and a second for Task 2.

Assessment task details and instructions

Task 1: Analysis of Clinical Trial Data Using Spark SQL (35 marks)

For this task, you are working for a pharmaceutical company, and your manager has asked you to perform some analysing of clinical trials registered in the USA to better understand the market.

The data necessary for this assignment is provided in a CSV file named Clinicaltrial_16012025.csv. The .csv file has a header row containing column names for each column. Every row in the dataset corresponds to an individual clinical trial and contains a variety of information on clinical trials, including the trial name, study type, study status, funder, conditions, start date and end date. The source is ClinicalTrials.gov (note, however, you should use the file we have shared on Blackboard for the analysis). The dataset contains over 500,000 rows so can be considered as having some of the characteristics of Big Data.

You need to use Spark SQL for this implementation. You have been asked to answer the following questions:

1. List all the clinical trial types (as contained in the Type column of the data) along with their frequency, sorting the results from most to least frequent **(5 marks)**
2. The top 10 conditions along with their frequency (note, that the Condition column can contain multiple conditions in each row, so you will need to separate these out and count each occurrence separately) **(5 marks)**
3. For studies with an end date, calculate the mean clinical trial length in months. **(5 marks)**
4. From the studies with a non-null completion date and a status of ‘Completed’ in the Study Status, calculate how many of these related to Diabetes each year. Display the trend over time in an appropriate visualisation. (For this you can assume all relevant studies will contain an exact match for ‘Diabetes’ or ‘diabetes’ in the Conditions column.) **(10 marks)**

Answers which are correct, with an accurate accompanying explanation of the implementation will receive full marks. Answers meeting some of these criteria will achieve partial marks.

The above questions account for 25 of the 35 marks for Task 1. An additional 10 marks is available for the quality of your submission, based on providing clean, well-commented code which follows best practice along with clear explanations of your solution.

Task 2: Recommender System (65 Marks)

For your second task, you are working with a dataset extracted from Steam, an online video game distribution service. This dataset is available on Blackboard and named steam-200k.csv. It provides details on the games different members have purchased

and played, along with the number of hours they have played each game. It contains four columns:

- The first column contains a unique identifier for each member
- The second column contains the name of the game they purchased or played
- The third column contains details of the member behaviour, either ‘purchase’ or ‘play’. Because a game has to be purchased before it can be played there will be two entries for the same game / member combination in some instances
- The fourth is set to 1 for rows where the behaviour is ‘purchase’. For rows where the behaviours is ‘play’ the value in the fourth column corresponds to the number of hours of play

We can use both purchase and play behaviours as implicit user feedback, which is useful for training a recommender system.

Your task as a data scientist is to do the following:

- Load the dataset into a Spark DataFrame. You may want to consider carrying out some initial exploratory analysis of the data, which you are welcome to do using DataFrames, Spark SQL, Databricks visualisations, another visualisation library etc.
- Use MLlib to train a collaborative filtering recommender system on the provided data, evaluate its performance and explore some of the resulting recommendations. You will need to carry out all pre-processing steps, such as splitting the data into training and test sets. It is your decision whether to include both ‘purchase’ and ‘play’ behaviours or to choose one of these as more suitable for your purposes. You may wish to experiment with more than one approach.

Note: To run Alternating Least Squares (ALS) matrix factorization using MLlib, we need to have integer ID values for both users and items, and this dataset does not contain IDs for the games. You will need to find a way to generate a unique integer ID for each game and add this into the DataFrame. There is an additional file, games.csv, which you can use to do this, but you will receive more marks if you are able to complete this within Databricks without using this additional csv file.

Higher scoring submissions will:

- Conduct multiple runs as part of the experiment (for example, using different hyperparameters)
- Track the experiment with MLflow

Assessment Criteria	A separate rubric is provided accompanying this assessment brief.
	Assessed intended learning outcomes On successful completion of this assessment, you will be able to:

Knowledge and Understanding

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 submission will be the form of two HTML files. You should aim for a total submission length of up to 3,500 words, counting text in the markdown cells only. This is likely to consist of between 8-12 markdown cells in each notebook of around 150-200 words. However, this is a guideline only and there is no direct penalty for being under or over this figure. **However,** your aim is to provide adequate explanation to someone seeking to understand your solution. Solutions without adequate explanation will be unlikely to score well.
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 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.