DAT7302
BIG DATA ANALYTICS
ASSESSMENT 1

ASSESSMENT BRIEF

Assessment Number	1
Assessment Type (and weighting)	*Portfolio- Coursework (100%)*
Assessment Name	Portfolio
Assessment Submission Date	*5th January 2025 (10 PM)*

Learning Outcomes Assessed:

LO1: Formulate appropriate queries to interrogate data based on a given brief.
LO2: Devise an effective process to capture, organise and store data for analysis.
LO3: Effectively utilise Cloud technologies for processing and analysing Big Data.

Use of Generative Artificial Intelligence (GAI) Applications in this Assessment

AI Status	Application	Notes
Category A	No GAI tool is permitted.	While grammar and/or spell checkers may be used to correct individual words and sentences, the use of GAI is not allowed. This is because the learning outcomes require you to produce original assessment work without any GAI assistance Any GAI generated content which is presented as your own original work and is not acknowledged will be assessed for academic misconduct.

Assessment Task:

Based on the given dataset/s, clearly define the business problem and identify specific business questions that can be addressed using data analysis. You are required to use appropriate data analysis methods and visualisation techniques to answer those questions. You are required to create a data pipeline to prepare data for analysis. Based on your analysis, draw conclusions, provide actionable insights, and suitable recommendations to the business.

The broad set of sub-tasks that you are required to perform (not in specific order):

- Define the business problem and develop at least **eight** business questions
- Use SQL to join/merge different datasets to create a single file/dataset
- Use AWS to create a data pipeline using AWS S3 and AWS Glue
- Use AWS Athena to query data
- Use IAM role/policies if required
- Use AWS Cloud9 and AWS Cloudformation if required
- Use appropriate data wrangling techniques for transformation
- Use Python and Spark to create a data pipeline for data analysis, perform exploratory data analysis, and data visualisation
- Use MongoDB to query dataset (JSON format) for data analysis
- Review previous scholarly literature/articles on similar datasets or business problems. The intent should be to identify various business problems/questions have been raised and addressed using the data analytical and visualisation techniques
- Reflect how different data analysis and visualisation techniques have addressed the business questions
- Derive insights from the data analysis performed
- Provide recommendations to the business based on your data analysis and derived insights

Dataset Description

File name: title.akas.tsv.gz

- Contains the following information for titles

Attributes Description

- titleId (string) - a tconst, an alphanumeric unique identifier of the title.
- ordering (integer) – a number to uniquely identify rows for a given titleId.
- title (string) – the localized title.
- region (string) - the region for this version of the title.
- language (string) - the language of the title.
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning.
- attributes (array) - Additional terms to describe this alternative title, not enumerated.
- isOriginalTitle (boolean) – 0: not original title; 1: original title.

File name: title.basics.tsv.gz

- Contains the following information for titles

Attributes Description

- tconst (string) - alphanumeric unique identifier of the title.

- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc).
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release.
- originalTitle (string) - original title, in the original language.
- isAdult (boolean) - 0: non-adult title; 1: adult title.
- startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year.
- endYear (YYYY) – TV Series end year. for all other title types.
- runtimeMinutes – primary runtime of the title, in minutes.
- genres (string array) – includes up to three genres associated with the title.

File name: title.principals.tsv.gz

– Contains the principal cast/crew for titles

Attributes Description

- tconst (string) - alphanumeric unique identifier of the title.
- ordering (integer) – a number to uniquely identify rows for a given titleId.
- nconst (string) - alphanumeric unique identifier of the name/person.
- category (string) - the category of job that person was in.
- job (string) - the specific job title if applicable, else.
- characters (string) - the name of the character played if applicable, else.

File name: title.ratings.tsv.gz

– Contains the IMDb rating and votes information for titles

Attributes Description

- tconst (string) - alphanumeric unique identifier of the title.
- averageRating – weighted average of all the individual user ratings.
- numVotes - number of votes the title has received.

File name: name.basics.tsv.gz

– Contains the following information for names

Attributes Description

- nconst (string) - alphanumeric unique identifier of the name/person.
- primaryName (string)– name by which the person is most often credited.
- birthYear – in YYYY format.
- deathYear – in YYYY format if applicable, else .
- primaryProfession (array of strings)– the top-3 professions of the person.
- knownForTitles (array of tconsts) – titles the person is known for

Notes:

- Attempt should be made to use the best practices to create data pipeline, perform data wrangling and data analysis, executing queries, choosing and apply appropriate data visualisation techniques. For example, using query optimisation techniques, use of modular programming
- The steps are iterative and might not necessarily follow each other in the given sequence
- In your report, give proper justification for each of the actions being performed
- In your report, number and label each figure and table. Figure captions should be placed below the figure. The table captions should be placed above the table. Cross-reference figures and tables
- Your report should be produced in a professional report style and must not exceed 7500 words, excluding the code, references, and appendices sections. There is a penalty if you exceed the word limit. Similarly, there is a penalty if you have lesser number of words than the word limit
- You should include a word count at the end of the assessment (excluding code, references, figures, tables and appendices)
- The report must be well-structured with a title page (student name, student number, name of module instructor, course name, date, and time of the workshop session), table of contents, appropriate headings, and page numbering
- The word document must be in Arial or Calibri Light font size 12. There should be double-spacing, and each page should be numbered
- The code snippets must be numbered
- All the learning outcomes should be covered
- The file name should be saved as DAT7302_1_Student-number. The pdf version of the report must be submitted only through Turnitin Link, and no report is to be emailed or sent on Teams
- Only one-time Turnitin submission is allowed
- Students will be asked to present or reflect on the submitted report
- It is expected that the Reference List will contain between fifteen to twenty sources. As a MINIMUM, the Reference List should include five refereed academic journals and three academic books
- Harvard Referencing should be used for this assignment. All written work should be referenced using the standard University of Bolton referencing style– see: https://libguides.bolton.ac.uk/resources/referencing/
- Please note: the deadline for submitting your report is 5th January 2025 (10:00 PM). Please submit the correct and completed report well before time. We suggest submitting your work 10-minute before the deadline. Extension requests won't be taken into consideration. Also, you are responsible for uploading the correct report on the Turnitin submission link.
- Requests from students to delete submissions on the grounds of high plagiarism or submission of incorrect reports will not be considered.
- A few tips to avoid chances of academic misconduct
    - Do not share your assignments with your classmates

    - Do not lend your laptop to your classmates. In case, you must make sure that the assignment and directory are password protected. Also, you are logged out of the AWS account, Google colab, and MongoDB accounts.
    - If working in the library or the University computer labs, do not leave your system unattended. Always log off or lock it before leaving the desk even for few minutes
- Declaration: At the end of the assessment, you should also include a declaration of any software tools including Generative AI (GAI) applications that you used in developing and completing the assessment.
- Feedback dates are based on the University of Bolton's fifteen working day feedback policy. Some dates may have been adjusted to consider holiday periods.

Structure of the Report (for reference)

1. Title Page
2. Declaration
3. List of Figures
4. List of Tables
5. Table of Contents (with page numbers)
6. Introduction
    a. Business Problem
    b. Business Questions
7. Review of Literature
8. Methodology
    a. This section will explain the techniques that will be used to address the business problem and business questions.
9. Implementation
10. Results
11. Discussion
    a. Derived Insights
    b. Recommendations
12. Conclusions
13. Personal Reflection (maximum of 500 words)
    a. Explain what you learned while completing this assignment, the challenges you faced, your future action plans
14. References
15. Appendices (if any)

Presentation:

All submitted work must be accompanied by a formal presentation, and attendance is obligatory. Students who do not participate in the presentation will receive a failing grade for the assignment. The evaluation panel will ask questions during the presentation to assess each student's understanding of the project and to verify the originality of their work. Successfully passing the presentation is a prerequisite for passing the Assessment.

Late work:

Late work will be subject to the following penalties:

- Up to 7 calendar days late = 10 marks subtracted, but if the assignment would normally gain a pass mark, then the final mark is no lower than the pass mark for the assignment.

- More than 7 calendar days late = This will be counted as non-submission, and no marks will be recorded.
- Late submission of assessments on refers and those which are graded Pass/Fail only is not permitted unless an extension is approved. See below.

Extensions

- In the case of exceptional and unforeseen circumstances, an extension of up to 14 days after the assessment deadline may be requested using the standard University Extension Request Form. For approval, there would need to be an explanation and evidence of relevant circumstances. Longer extensions for individual projects and artefacts may be granted at the discretion of the Programme Leader.
- Requests for extensions which take a submission date past the end of the module (normally week 15) must be made using the Mitigating Circumstances procedure.
- Some students with registered disabilities will be eligible for revised submission deadlines. Revised submission deadlines do not require the completion of extension request paperwork.
- Please note that the failure of data storage systems is not considered to be a valid reason for an extension. It is, therefore, important that you keep multiple copies of your work on different storage devices before submitting it.

Academic misconduct:

Academic misconduct may be defined as any attempt by a student to gain an unfair advantage in any assessment. This includes plagiarism, collusion, commissioning (contract cheating) amongst other offences. In order to avoid these types of academic misconduct, you should ensure that all your work is your own and that sources are attributed using the correct referencing techniques. You can also check originality through Turnitin. Please note that penalties apply if academic misconduct is proven. See the following link for further details:
https://www.bolton.ac.uk/student-policy-zone/student-policies-2024-25/academic-misconduct-regulations-and-procedures-2024-25

Minimum Secondary Research Source Requirements:

Level HE7 - It is expected that the Reference List will contain between **fifteen to twenty sources**. As a MINIMUM the Reference List should include **four refereed academic journals and five academic books.**

Specific Assessment Criteria/Marking Scheme:

Distinction (70% and above)

An excellent data analysis and visualisations would be presented that are appropriate to the business problem and business questions. The business problem and questions are clearly defined and relevant. The insights drawn from analysis are relevant and impactful. The data pipeline is functional, efficiently and accurately implemented. Queries are accurate, efficient, and optimised for performance. There is a use of some advanced queries. The visualisations are clear and appropriate to support the analysis and addressing business questions and problem. The visualisations are well-designed, have clear labels, and use of colors to enhance the understanding and readability. The code is clean, readable,

and follows the best practices. The documentation is supported by appropriate comments. The results are supported by appropriate and adequate justifications.
Personal Reflections will be succinct, insightful and original. Extensive research demonstrating use of a wide range of contemporary and seminal sources will be evident. Academic writing style, English and referencing will be excellent.

Merit (60% -69%)
Good data analysis and visualisations would be presented that are appropriate to the business problem and business questions. The business problem and questions are mostly clearly defined and relevant. The insights drawn from analysis are mostly relevant and impactful. The data pipeline is functional and accurate, and mostly efficiently implemented. Queries are mostly accurate, efficient, and optimised for performance. There is a use of some advanced queries. The visualisations are clear and appropriate to support the analysis and addressing business questions and problem. The visualisations are designed with some design issues. The code is mostly clean, readable, with minor style issues. The documentation is supported by comments. The results are mostly supported by adequate justifications.
Personal Reflections will be succinct and original. Research demonstrating use of a wide range of relevant research sources will be evident. Academic writing style, English and referencing will be good.

Pass (50%-59%)
Data analysis and visualisations would be presented with an attempt to support the business problem and business questions. The business problem and questions are defined. The insights drawn from analysis. The data pipeline is functional. Queries are accurate but not optimised for performance. Attempt has been made use of some advanced queries. The visualisations are not clearly understandable. The code is readable but lack structure. The documentation is mostly supported by comments. The results are not supported by adequate justifications some places.

Some original reflections will be presented. Research demonstrating use of a range of relevant research sources will be evident. Academic writing style, English and referencing will be satisfactory.

Fail (Below 50%): Students who do not meet the requirements of the Pass criteria will not successfully complete the assessment activity.

General Assessment Criteria for Written Assessments

GENERAL ASSESSMENT GUIDELINES – LEVEL HE7

		Relevance Learning outcomes must be met for an overall pass	Knowledge and Understanding	Analysis, Creativity and Problem-Solving	Self-awareness and Reflection	Research/ Referencing	Written English	Presentation and Structure
DISTINCTION	Exceptional Quality **85-100%**	Work is directly relevant and expertly addresses the requirements of the brief. **Learning outcomes are met.**	Demonstrates an exceptional breadth and depth of knowledge and understanding of theory and practice which is beyond the threshold expectation for the level. Produces exceptional work which makes a contribution to knowledge in the subject area.	Presents an exceptional critique of advanced research material resulting in clear, original and illuminating conclusions. Expertly interprets complex matters and ideas and makes sound judgments in the absence of complete data. Demonstrates exceptional creative flair and a high level of originality. Develops distinctive, insightful and creative solutions to complex problems.	Provides insightful reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility.	An extensive range of advanced research sources critically evaluated and selected. Sources cited accurately in both the body of text and in the Reference List/ Bibliography.	Writing style is clear, succinct and relevant to the requirements of the assessment. An exceptionally well written answer with only rare errors in spelling, grammar and punctuation. Uses a wide range of less common and more advanced vocabulary and sentence types fluently and flexibly to convey precise meanings. Paragraphs are expertly structured with linking and signposting.	The presentational style and layout are correct for the type of assignment. Evidence of planning and logically structured. Where relevant, there is effective inclusion of, and reference to, figures, tables and images.
	Excellent Quality **70-84%**	Work is relevant and comprehensively addresses the requirements of the brief. **Learning outcomes are met.**	Demonstrates an excellent breadth and depth of knowledge and understanding of theory and practice for this level. Clearly demonstrates originality in the application of knowledge.	Presents an excellent critique of advanced research material resulting in clear, original and illuminating conclusions. Comprehensively interprets complex matters and ideas and makes sound judgments in the absence of complete data. Demonstrates excellent creative flair and a high level of originality. Develops distinctive and creative solutions to complex problems.	Provides excellent reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility	A wide range of advanced research sources critically evaluated and selected. Sources cited accurately in both the body of text and in the Reference List/ Bibliography.	Writing style is clear, succinct and relevant to the requirements of the assessment. A very well written answer with very few errors in spelling, grammar and punctuation. Uses a wide range of vocabulary and sentence types fluently and flexibly to convey precise meanings. Paragraphs are very well structured with linking and signposting.	The presentational style and layout are correct for the type of assignment. Evidence of planning and logically structured Where relevant, there is effective inclusion of, and reference to, figures, tables and images.
MERIT	Good Quality **60-69%**	Work is relevant and addresses most of the requirements of the brief well. **Learning outcomes are met.**	Demonstrates a thorough breadth and depth of knowledge and understanding of theory and practice for this level. Demonstrates originality in the application of knowledge.	Presents a comprehensive critique of advanced research material resulting in clear and original conclusions. Systematically interprets complex matters and ideas and makes sound judgments in the absence of complete data. Demonstrates creative flair and originality. Develops creative solutions to complex problems.	Provides good reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility, as required by the assessment.	A range of advanced research sources critically evaluated and selected. In the main, sources cited accurately in both the body of text and in the Reference List/ Bibliography.	Writing style is clear, succinct and relevant to the requirements of the assessment. A well written answer with some minor errors in spelling, grammar and punctuation. Uses a range of vocabulary and sentence types to convey precise meanings. Paragraphs are well structured with linking and signposting.	The presentational style and layout are correct for the type of assignment. Evidence of planning and logically structured in the main. Where relevant, there is effective inclusion of, and reference to, figures, tables and images.

PASS	Satisfactory Quality **50-59%**	Work addresses the requirements of the brief, although superficially in places. Minor irrelevant content. **Learning outcomes are met.**	Demonstrates a sufficient breadth and depth of knowledge and understanding of theory and practice for this level. Demonstrates originality in the application of knowledge in places.	Presents some critique of advanced research material resulting in original conclusions. Shows the ability to systematically interpret complex matters and ideas and makes sound judgments in the absence of complete data. Demonstrates creative flair and originality. Develops some creative solutions to complex problems.	Provides justified reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility, as required by the assessment.	Advanced research sources critically evaluated and selected. Some errors evident in relation to referencing the text and Reference List/Bibliography	Writing style is readable and relevant to requirements of the assessment in the main. Competently written with minor lapses in spelling, grammar and punctuation. For example, paragraphs are structured and include some linking and signposting. Sentences are complete. A range of appropriate vocabulary is used.	The presentational style and layout are correct for the type of assignment. Logically structured in the most part. Inclusion of figures, tables and images but not all relevant or referred to.
FAIL	Borderline Fail **45-49%**	Work addresses some of the requirements of the brief. Irrelevant and superficial content. **One or more learning outcomes have not been met.**	Some omissions evident in knowledge and understanding of theory and practice for this level. Demonstrates limited originality in the application of knowledge.	A limited critique of research material presented, with simplistic conclusions drawn. Some complex matters and ideas interpreted but not systematically, resulting in flawed judgements. Limited creativity and originality evident. Demonstrates insufficient problem-solving skills and initiative.	Provides limited reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility, when required.	Sources selected are limited and lack validity/relevance. Poor referencing technique employed.	Writing style is generally readable but lacks relevance to requirements of the assessment in places. Intermittent lapses in grammar, spelling and punctuation pose obstacles for the reader. For example, some paragraphs may lack structure and there is limited linking and signposting. Some appropriate vocabulary is used	For the type of assignment the presentational style, layout and/or structure are lacking. Inclusion of figures, tables and images but not clear, relevant and/or referred to.
	Fail **0-44%**	Work does not address the requirements of the brief. Irrelevant and superficial content. **One or more learning outcomes have not been met.**	Demonstrates inadequate knowledge and understanding of theory and practice for this level. A lack of originality in the application of knowledge is evident.	An absence of critique of research material evident with unjustified conclusions. Complex matters and ideas not interpreted correctly resulting in flawed judgements. Creativity and originality are absent. Insufficient problem-solving skills and initiative demonstrated.	Provides inadequate reflection and critical self-awareness in relation to the outcomes of own work and personal responsibility, when required.	Relevant and valid sources are not selected. Poor referencing technique employed.	Writing style is unclear and does not match the requirements of the assessment in question. Deficiencies in spelling, grammar and punctuation makes reading difficult and arguments unclear. Unstructured paragraphs	For the type of assignment the presentational style, layout and/or structure are lacking. Inclusion of figures, tables and images but not clear, relevant and/or referred to.