SCHOOL OF ARCHITECTURE, COMPUTING & ENGINEERING

Submission instructions

- Cover sheet to be attached to the front of the assignment when submitted
- Question paper to be attached to assignment when submitted
- All pages to be numbered sequentially
- All work has to be presented in a ready to submit state upon arrival at the ACE Helpdesk. Assignment cover sheets or stationery will **NOT** be provided by Helpdesk staff

Module code	CN6022		
Module title	Big Data Infrastructure and Manipulation		
Module leader	Fahimeh Jafari		
Assignment tutor	Fahimeh Jafari, Mustansar Ali Ghazanfar		
Assignment title	Coursework		
Assignment number	1		
Weighting	100%		
Handout date	Week 5		
Submission date	Presentation: Tuesday 16th and Thursday 18th Dec 2025 Turnitin Submission: Monday 15th Dec 2025 (midday)		
Learning outcomes assessed by this assignment	5, 6, 7, 8, and 9		
Turnitin submission requirement	Yes	**Turnitin GradeMark feedback used?**	No
UEL Plus Grade Book submission used?	No	**UEL Plus Grade Book feedback used?**	
Other electronic system used?	Yes	**Are submissions / feedback totally electronic?**	Yes
Additional information			

Form of assessment:

☐ Group work ☒ Individual work

For **group work** assessment, which requires members to submit both individual and group work aspects for the assignment, the work should be submitted as:

☒ Consolidated single document ☐ separately by each member

Number of assignment copies required:

☒ 1 ☐ 2 ☐ Other

Assignment to be presented in the following format:

☒ On-line submission
☐ Stapled once in the top left-hand corner
☐ Glue bound
☐ Spiral bound
☐ Placed in a A4 ring bound folder (not lever arch)

Note: To students submitting work on A3/A2 boards, work has to be contained in suitable protective case to ensure any damage to work is avoided.

Soft copy:

☐ CD (to be attached to the work in an envelope or purpose made wallet adhered to the rear)
☐ USB (to be attached to the work in an envelope or purpose made wallet adhered to the rear)
☒ Soft copy not required

Coursework Project

The coursework consists of two sections: (1) Data Analytics using Spark SQL and (2) Machine Learning using Spark MLlib. The project should be carried out individually. The delivery of the project would be as follows:

1- **Project Presentation (35%):**
You must attend the online presentation through Microsoft Teams.
Note: If you fail to attend the presentation, your mark will be zero.
Date: Tuesday 16th and Thursday 18th Dec 2025

2- **Report Submission (65%):**
The project report (around 3,000 ± 10% words) is in the HTML format.
You must submit the report through the Turnitin link on Moodle.
Due Date: Monday 15th Dec 2025 (midday)

Assessment criteria:

Marking Scheme for Main Sit

Topic	Total Mark	Sub-marks
Big Data Analytics using Spark SQL	25 marks	• Implementation of three Spark SQL queries. [12 marks] • A short explanation of the queries. [3 marks] • Visualisation of your findings as plots or tables. [10 marks]
Machine Learning using Spark MLlib	35 marks	• Implementation of two ML algorithms. Follow the complete process of machine learning involving feature selection, preprocessing, class imbalance, etc. [16 marks] • Explain the algorithms and configurations. [3 marks] • Evaluate and compare the performance of both models [6 marks] • Visualisation of your findings as plots or tables. [10 marks]
Legal, Social, Ethical, and Professional (LSEP) considerations	5 marks	• Analyse the legal, social, ethical, and professional implications associated with the data and the analysis. You can consider factors such as data privacy, data protection, bias, fairness, transparency, and the potential impact of the analysis on individuals or society. You can work on this as a team and provide 2-3 paragraphs explaining your perspectives and insights regarding LSEP.
Presentation	35 marks	• Express understanding and knowledge of Big Data [8 marks] • Demonstrate relevant and supportive statements of their work [15] • Respond clearly and correctly to the questions [12]
Total	**100 marks**	

Assessment for resit (2nd attempt):

We will have only the project report (100%) in the HTML format. Students must develop new solutions for the same tasks. If students copy their solutions from main-sit, it will be considered as self-plagiarism and the mark will be zero. The marking scheme is the same as the main-sit.

Report Submission (100%):

The project report (around 3,000 ± 10% words) is in the HTML format.

You must submit the project through the Turnitin link on Moodle.

Due Date: TBC

Marking Scheme for Resit

Topic	Total Mark	Sub-marks
Big Data Analytics using Spark SQL	40 marks	• Implementation of three Spark SQL queries. [20 marks] • A short explanation of the queries. [10 marks] • Visualisation of your findings as plots or tables. [10 marks]
Machine Learning using Spark MLlib	55 marks	• Implementation of two ML algorithms. Follow the complete process of machine learning involving feature selection, preprocessing, class imbalance, etc. [20 marks] • Explain the algorithms and configurations. [15 marks] • Evaluate and compare the performance of both models [10 marks] • Visualisation of your findings as plots or tables. [10 marks]
Legal, Social, Ethical, and Professional (LSEP) considerations	5 marks	• Analyse the legal, social, ethical, and professional implications associated with the data and the analysis. You can consider factors such as data privacy, data protection, bias, fairness, transparency, and the potential impact of the analysis on individuals or society. You can work on this as a team and provide 2-3 paragraphs explaining your perspectives and insights regarding LSEP.
Total	**100 marks**	

Big Data Analytics using Apache Spark

CN6022 – Big Data Infrastructure and Manipulation

Task 1: Understanding the Dataset – Air Flight Status

This dataset contains all flight information, including cancellations and delays by airline, for dates back to January 2018. For your convenience, we use the *Combined_Flights_2022.csv* to access the combined data for the entire year from 2022 back to 2018. This file also has filtered out columns that are mostly null in the original dataset. To access the dataset and other supporting files, you need to download *flights.zip* from https://tinyurl.com/mwxnxnkb

Then, unzip it to get three files: *Airlines.csv*, *Combined_Flights_2022.csv*, *readme.html*

a) The features are described in *readme.html*.

b) The codes of airlines are described in *Airlines.csv*.

c) In this coursework, we use data stored in *Combined_Flights_2022.csv* with a total size of 1.31GB, which is big enough to employ big data methodologies for analysis. As big data specialists, firstly, we would like to read and understand its features and then apply modelling techniques. To view a few records from this dataset, load it into a PySpark dataframe and use a SQL query to print the first 5-10 records for your understanding.

Task 2: Big Data Analysis using Spark SQL [25 marks]

This task uses Spark SQL for converting big raw data into useful information. You need to implement **three** complex SQL queries. Apply appropriate visualisation tools to present your findings numerically and graphically. Interpret your findings shortly.

What information do you need to include in the HTML report for each student?

1. Implementation of **three** Spark SQL queries. [12 marks]
2. A short explanation of the queries. [3 marks]
3. Visualisation of your findings as plots or tables. [10 marks]

Tip: The mark for this section depends on the level of complexity of the SQL queries; for instance, using a simple select query is not supposed to get a full mark.

Task 3: Machine Learning using Spark MLlib [40 marks]

Design and build **two** ML algorithms to predict **ONE** of the following items:

- Predict which flights will be cancelled or delayed.
- Predict the delay time of flights?

What information do you need to include in the HTML report for each student?

1. Implementation of two ML algorithms. Follow the complete process of machine learning involving feature selection, preprocessing, class imbalance, etc. [20 marks]
2. Explain the algorithms and configurations. [3 marks]
3. Visualisation of your findings as plots or tables. [7 marks]

4. Evaluate and compare the performance of both models [10 marks]

Presentation [35 mark]

1. Express understanding and knowledge of Big Data [8 marks]
2. Demonstrate relevant and supportive statements of their work [15]
3. Respond clearly and correctly to the questions [12]

THE FORMAT OF FINAL SUBMISSION

Your final report must be based on the following format:

1- You can use either Google Colab (https://colab.research.google.com/) or Jupyter Notebook in Ubuntu VMWare machine for this CRWK.

3- You have to convert the source code (*.ipynb) to HTML as you did in the lab sessions.

4- Upload ONLY one single HTML file into Turnitin in Moodle. The name of the file must be in the format of “Your_ID”, such as U17672345.html

5- The submission link will be available from week 10, and you are free to amend your submitted file several times before the submission deadline. Your last submission will be saved in the Moodle database for marking.

PLAGIARISM

If there are copied Spark SQL and PySpark codes from somewhere or someone else, you will get zero with a plagiarism flag in your transcript, and should attend the “breach of regulation” committee for further explanations and the probable additional penalties.

FEEDBACK TO STUDENTS

Feedback is central to learning and is provided to students to develop their knowledge, understanding, skills and to help promote learning and facilitate improvement.

- Feedback will be provided as soon as possible after the student has completed the assessment task.
- Feedback will be in relation to the learning outcomes and assessment criteria.
- It will be offered via Turnitin GradeMark or Moodle post.

As the feedback (including marks) is provided before the Award & Field Board, marks are:

- Provisional
- available for External Examiner scrutiny
- subject to change and approval by the Assessment Board