

Department of Computing

School of Mathematical, Physical and Computational Sciences Assessed Coursework Set

Module Title: Big Data and Cloud Computing

Lecturers responsible: Prof. Atta Badii, Dr Xiang Li

Assistant Lecturer: Aadil Sattar

Type of Assignment: Coursework

Individual/group Assignment: Individual

Total Weighting of the Assignment: 50% comprising of 25% for each of Task A and Task B

Expected time spent on this assignment: 20 hrs

Page limit/Word count: Approximately 3000 words max, consisting of two sections of 3.5 (max) pages each, to report on the implementation of two tasks (Task A and Task B); Section A to report on Cloud Computing Task implementation (Task A) and Section B to report on the Big Data task implementation (Task B) – Maximum of 7 pages excluding appendices (for screen-shots and/or programming code segments) and the report should follow the School Style Guide.

Items to be submitted: Two PDFs to be submitted via BB, each of 3.5 pages max, one for Section A (Task A) and one for Section B (Task B). The PDFs are to include, on the first page, a link of GitLab or similar repository to the code. For Section A, as the solution is to be provided within your free Azure account, please provide a temporary username and password to your Azure account; for Section B the link can be a GitLab.

Work to be submitted online via Blackboard Ultra by: 12:00 Hrs. Monday 19th May 2025.

Artificial Intelligence Tools (select one of these): May be used to support work, are actively encouraged

NOTES

By submitting this work, you are certifying that you have read the assessment guidelines, which are displayed in the folder of Assessment on the Blackboard course for this module, and that you have conformed to and understand the associated policies and practices, including those on:

- Submitting your own work, not that of other people or systems (including those using artificial intelligence), and the associated penalties for Academic Misconduct
- Submitting by the specified deadline, and the penalties associated with late submission (if allowed)
- The exceptional circumstances system
- For students with relevant needs, attaching with a green sticker

Assignment Tasks based on the explanatory notes in the Appendix to this document.

Section A (The Cloud Computing Task)

Task A: Implement a MapReduce solution to determine the passenger(s) having had the highest number of flights based on flights and passenger data provided in the Assignment Folder of the Module on Blackboard.

Section B (The Big Data Task):

Task B: Implement a solution to predicting flight delays based on historical weather and airline data as provided in your free azure account and explain the reason for your preferred Machine Learning (ML) model.

If you face any difficulties, make clear, in your submission, how far you were able to proceed with the implementation and explain the challenges you faced.

Marking Criteria and Overall Assessment Grading

- Total marks for each of the two tasks (A & B) will be normalised for 25% credit towards the overall coursework.
- The table below indicates the level of performance expected for each range of assessments:

Classification Range	Typically, the work should meet these requirements
Distinction (>= 70%)	The assignment demonstrates: <ul style="list-style-type: none">• Excellent technical skills in implementing the system, possibly also suggesting any other solution deemed viable, including reasons for the preferred solution.• Professional technical writing skills and style.
Merit (60-69)	The assignment demonstrates: <ul style="list-style-type: none">• Excellent technical skills in implementing the solution.• Appropriate technical writing skills and clear presentation; including reasons for the preferred solution.
Good standard (Pass) (50-59)	The assignment demonstrates: <ul style="list-style-type: none">• Excellent technical skills in implementing the system.• Moderate technical writing skills and clear presentation; including reasons for the preferred solution.
Failing categories (40-49)	The assignment demonstrates: <ul style="list-style-type: none">• Technical skills in implementing the system have fallen below the basic satisfactory level.• The technical report content and presentation overall has fallen short of basic satisfactory level.
Unsatisfactory work (0-39)	The coursework fails to demonstrate technical skills to implement technical writing and clear presentation; inadequate or non-existent reasoning for the preferred solution.

Marking Scheme and feedback template for **Task A** (Cloud Computing Task)

Total marks for this Task B will be normalised for 25% credit towards the overall coursework assessment.

MapReduce Concepts		
Concept	Example	Max.
Map Phase	Inputs and Outputs	5
Reduce Phase	Inputs and Outputs	5
Segmentation of Roles	Split of work	2
File Handling	Use of Files and Buffers	3
Distributed parallelism	Advantages, fault tolerance etc.	3
Explanation of additional process	Combining/Shuffling/partitioning etc.	1
Flowchart	Illustration of MapReduce problem solving	1
		20

Software Prototyping		
Concept	Example	Max
Project Structure	Object-Orientation/class hierarchy	7
Code Re-usability	Generics, Templating	7
Solution Elegance	Design Optimality	6
		20

Implementation		
Aspect		Max.
Task Implementation	Key/Value Selection	6
	Correct Result	4
	Output Format	4
Parallelisation	Multi-threading	6
		20

Documentation		
Aspect		Max.
Report Structure	Abstract, Sections, Length, References, etc.	2
Section Content	Description of development	4
	Evidence of use of Version Control	5
	Evidence of understanding MapReduce	7
	Conclusions	5
Report Quality	Overall Quality of Report	5
Code Commenting	Use of comments in code	12
		40
Total		100

Marking Scheme and feedback template for **Task B** (The Big Data Task)

- Total marks for this Task B will be normalised for 25% credit towards the overall coursework assessment.

The key criteria for the assessment of the submitted coursework	Contribution to Mark in %
Introduction <ul style="list-style-type: none"> • Brief description of the background of the case study. • Description of the tools and techniques deployed, including “Data Factory”, “Data Bricks”, “Power BI” as used to analyse this solution (explaining the solution architecture). 	5 10 15
Solution Implementation Implementation of Solution: <ul style="list-style-type: none"> • Creating the Data Bricks cluster • Load sample data • Setup the Data Factory • Data factory pipeline • Operation of ML • Summarizing data • Visualisation of data 	7 5 8 5 15 5 15 60
Evaluation: Your personal reflections on: <ul style="list-style-type: none"> • Stating reasons for your preferred solution and recommendations for how the solution could be improved including design steps for data privacy and security risks safeguarding and minimisation of at least one potential adverse impacts of the present solution. 	10 10
Presentation of the report: <ul style="list-style-type: none"> • Structure and layout of the report • Professional writing style • Use of figures, tables, references, citations, and captions 	5 5 5 15
Total	100

Assignment Case Study Description and Data Access Details for Task A

For this coursework there are two files containing lists of data. These are located on the Blackboard system in the Big Data and Cloud Computing assignments directory – download them from:

Blackboard → Enrolments → [CSMBD24-25MOD: Big Data and Cloud Computing \(24-25\)](#)

In the Assessment tab, select as follows:

[Assessment](#) → [Cloud Computing](#) → [Coursework Data](#)

The coursework data folder includes the following two files:

[AComp Passenger data no error.csv](#)

[Top30 airports LatLong.csv](#)

The first 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:

Passenger id	Format: <i>XXXnnnnnXXn</i>
Flight id:	Format: <i>XXXnnnnnX</i>
From airport IATA/FAA code	Format: <i>XXX</i>
Destination airport IATA/FAA code	Format: <i>XXX</i>
Departure time (GMT)	Format: <i>n [10]</i> (Unix 'epoch' time)
Total flight time (mins)	Format: <i>n [1. .4]</i>

Where *X* is Uppercase ASCII, *n* is digit 0..9 and [*n. . m*] is the min/max range of the number of digits/characters in a string.

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

Airport Name	Format: <i>X [3. .20]</i>
Airport IATA/FAA code	Format: <i>XXX</i>
Latitude	Format: <i>n. n [3. .13]</i>
Longitude	Format: <i>n. n [3. .13]</i>

There are two additional data input files which can be used for analysis and validation. However, these should not be used for the final execution of the implemented jobs, these. n be downloaded from this directory and are as follows:

[AComp Passenger data.csv](#)

[AComp Passenger data no error DateTime.csv](#)

Once the above data files are accessed, the next steps are to implement Task A as described in the next section which describes Task A.

Task A Description

Determine the passenger(s) having had the highest number of flights.

For this task you are expected to implement a MapReduce-like executable prototype, (in Java, C, C++, or Python). The objective is to develop the basic functional 'building-blocks' that will address the Task above, in a way that emulates the MapReduce/Hadoop framework i.e. executing MapReduce without deploying the Hadoop Cluster.

Write a brief report (no more than 3 pages, excluding any appendices), describing:

- The high-level description of the development of the prototype software.
- A simple description of the version control processes undertaken.
- A detailed description of the MapReduce functions implemented.
- The output format of any reports that the job is to produce.

The solution may use multi-threading as required. The marking scheme reflects the appropriate use of coding techniques, succinct code comments as required, data structures and overall program design. The code should be subject to version control best-practices using a hosted repository under your university username.

Although students at the MSc level would be expected to have sufficient programming skills in Python/Java to be able to implement the required code for this task which is typically not complex, there are however ample tutorial resources available on the course site on BB, including worked examples and videos of walkthrough of typical code in both Python and Java as well as plenty of text and links to external resources to support the required programming for this task, including tutorial resources and support for threading control.

Assignment Case-Study for Task B


Margie's Travel (MT) provides concierge services for business travellers. They need to modernise their system. They want to focus on web apps for their customer service agents who provide flight booking information to the travellers. This could, for example, include features such as a prediction of flight delay of 15 minutes or longer, due to weather conditions.

You are expected to analyse the design of a system to predict the flight delay by processing the data provided. Your solution will need to be responsive to the customers' needs as specified. Your report is to describe your progress on the objectives of the task, including the aspects set out below:

1. A brief description of the background of the case study.
2. A description of the implementation of the solution supported by your free Azure account including tools and techniques deployed such as Data Factory, Data Brick, and Power BI.
3. The reason for your preferred solution.

By clicking on the link below, you can take the first step to creating your free MS Azure account and should be able to complete all the steps for Task A using the student's free \$100 allowance. Please ensure that you close all files and all applications processing before you exit the system at the end of each session of usage to avoid wasted time-on-the system being metered up, and thus wasted credit and running out of your \$100 credit before being able to complete your solution. **Save & Close All Running Instances:** i.e. save away your files and finally close any applications running and close

your account before your credit runs out so that you will not become liable for payment of charges. Accordingly, to ensure that you make the most efficient use of your credit, plan your offline preparatory and online pipeline configuring and execution sessions carefully; keep track of your usage so that it will not exceed the free credit limit, and you will not incur charges.

Azure for Students – Free Account Credit Microsoft Azure	 <p>With Microsoft Azure for Students, get a \$100 credit when you create your free account. There is no credit card needed and 12 months of free Azure services.</p> <p>Please register for this free account through https://azureforeducation.microsoft.com/devtools In this way you will not be asked for a credit card number and risk subsequently being billed because there is no need for you to incur charges in attempting to use the MS Azure under this scheme; so please follow this link and register correctly so that you will not risk incurring charges for which you will be liable.</p>
---	--

Coursework description for Task B: Analysis of Big Data solution Architecture

Implement and evaluate the big data solution to be provided to a customer, a VIP Flights Booking Agency, who needs to improve their system to provide ranked recommended booking options for their VIP customers such that the likelihood of travel delays would be minimised - as described below.

Task B Description:

Implement a solution to predict flight delays based on historical weather and airline data

In order to deal with big data, it is required to process data in a distributed manner. Azure Synapse Analytics in Azure Machine Learning (ML) provides a platform for data pre-processing, featurisation, Machine Learning model training and deployment. It can connect Spark pools within the MS Azure Synapse Analytics platform, at your disposal, where PySpark helps with pre-processing the data in an interactive way. This environment provides powerful Bigdata Analytics tools such as **Data Factory**, **Data Brick** and **Power BI** which you will need to learn to use in configuring your own pipeline and developing your own solution, from data ingestion to ML results display for this coursework using the data for the case study available on Azure Synapse Case Study as described below. You will then be able to write a short report as specified in this document to **present and justify your choices for the deployment specific components including, crucially, the Machine Learning (ML) model and Results display using** the configuring, selection and execution of appropriate building blocks including the particular ML as selected from a menu of available models. This task is to provide you with the requisite skills to become a data scientist cable of efficient configuration and deployment of optimal ML solution stacks using cloud-based AI Engineering environments as required in practical applications of AI. The justification of your choice of ML could be supported by performance benchmarking of solutions as well as consideration of potential risks such as data privacy and security of the solution. Your reflections should include suggestions on how the present solution could be improved for data privacy and security risks safeguarding and minimisation of at least one potential adverse impacts of the present solution.

Please download the case study from the Big Data assignment guide in the assessment area:

Blackboard → Enrolments → CSMBD24-25MOD: Big Data and Cloud Computing (24-25)

In the Assessment tab, select as follows:

Assessment → Big Data → Case Study

And develop your solution accordingly.
