School of Information Technology

Diploma in Applied AI & Analytics

AY2023/2024 S1

IT3388

Big Data Management Project

Project Guide

Contents

Contents

1 Introduction 3

2 Project Schedule & Submissions 4

3 Roles and Responsibilities 5

4 Project Management 5

5 Project Scope & Scenario 6

6 Module Assessment 7

APPENDIX A – FORM FOR SUBMITTING PROJECT TEAM ORGANISATION 13
(TO BE COMPLETED AS A TEAM) 13

APPENDIX B – WEEKLY PROGRESS UPDATES 14
(TO BE COMPLETED AS A TEAM IN BLACKBOARD GROUP BLOG) 14

APPENDIX C – PROJECT PROPOSAL 15
(TO BE COMPLETED AS A TEAM) 15

APPENDIX D – EXPLORATORY DATA ANALYSIS 16
(TO BE COMPLETED INDIVIDUALLY) 16

APPENDIX E – FINAL PROJECT REPORT 17
(TO BE COMPLETED INDIVIDUALLY) 17

APPENDIX F – SELF AND PEER REVIEW 18
(TO BE COMPLETED INDIVIDUALLY) 18

APPENDIX G – CHECKLIST OF QUALITY ATTRIBUTES FOR FINAL PRESENTATION 19

1 Introduction

1.1 Aims

With more and more industry storing their data on the cloud, the skill and knowledge to analyse these data from the cloud directly is getting more important. Through this WIU, learners will demonstrate their competencies in analysing Big Data to arrive with an outcome. Learners will be able to identify and access big data analytics related business problems and propose a viable big data analytics solution to address the business issues. Learners will also be able to evaluate the effectiveness of the analytics solutions in terms of predictive accuracies as well as system scalability.

1.2 Objectives

At the end of this unit, learners will be able to:

1. Collect and analyse big data using research frameworks and historical data that comply with data and privacy and ethics.
2. Perform big data processing and aggregation operations on big data in the cloud platform using data manipulation language to address the business needs.
3. Perform data visualisation and data analysis on the dataset from the cloud.
4. Deploy their project on a cloud platform.
5. Work collaboratively in a team to develop dashboards using data storytelling approach for an effective narrative and visual representation.

2 Project Schedule & Submissions

This project will be completed over 17 weeks. The detailed lesson plan can be found in Module Guide.

Week	Activities	Deliverables/ Resources	In-Course Assessment
11	Big Data Analytics Project Overview Project Briefing + Team Formation - Form project team - Understand project scope and expectation - Clarify project requirements	*Course materials & project guide*	
12	Exploratory Data Analysis	*Course materials & project guide*	
13	Project Proposal - Business problem/Use case - Project work assignment - Exploratory Data Analysis	*Presentation slides, notebooks & Annex C, D*	*Project Proposal 10% + Exploratory Data Analysis 20%*
14	Project Implementation		
15	SDL Week		
16	Project Implementation		*30% Interim Progress Review*
17	Project Implementation		
18	Final Project Presentation Report Submission	*Final Project Presentation Report*	*Presentation - 30%* *Report - 10%*

All submissions are to be submitted via Brightspace. Late submissions will be penalised.

3 Roles and Responsibilities

Project groups consist of **4 students**. Each group must elect a group leader.

3.1 Project Supervisor / Module Tutor

- Provide feedback to project teams' proposals and designs.
- Advise the project group during project development, including the task allocation, development plan, tools used, etc.
- Provide inputs from the user perspective to the project team.
- Act as technical advisors (supervisors are **NOT** expected to debug programs or solve problems for students).

3.2 Group Leader

- Maintain the project schedule and documentation, and record attendance at meetings.
- Ensure documentations of the weekly team and progress update meeting is captured in the Group Blog in Blackboard every week.
- Coordinate and assign work among members.
- Arrange meetings with the project supervisor. Leader is required to book the meeting rooms if meeting is held outside of the lab/tutorial hours.
- Communicate with the supervisor on behalf of the team to discuss any project-related matters.

3.3 Group Members

- Carry out allocated tasks.
- Be punctual and attend all lessons and project meetings.
- Cooperate with the rest of the group during project development.

4 Project Management

4.1 Weekly Team Meeting

Each team is required to make use of the practical time to conduct a weekly team and progress update meeting. The weekly progress updates should be captured in the Group Blog in Blackboard (refer to Appendix B).

4.2 Project Plan

Each team is required to distribute the workload evenly among members. A project schedule Gantt chart detailing assignment by tasks, members' responsibility and timeframe must be created. This chart must be updated to reflect task completion and changes on a regular basis during the weekly progress meeting. The chart, and all revisions to the chart, should be captured in the Group Blog in Blackboard (refer to Appendix B).

4.3 Project Control

Milestones are used to monitor progress. At each milestone, there are deliverables to be submitted. Students are expected to meet the milestones as the project progresses, and not delay all work towards the end of the project.

5 Project Scope & Scenario

5.1 Background

In this project, you have to identify and describe business problems or areas of interest that are related to the given data set. The big data pipeline should involve investigation and exploration of the big data before performing analysis and developing a set of conclusions.
At the end of the big data pipeline, you will evaluate the effectiveness of the analytics solutions in terms of predictive accuracies as well as system scalability.
The evaluation of the results or conclusions should be supported by some metrics collected during the machine learning flow.

5.2 Minimum Feature Requirements

- Must convey a conclusion of the research
- Data must be uploaded into a cloud platform and operated from the cloud platform.
- The interface will be rich and intuitive in presenting the results of the research.
- Should be highly interactive, attractive, and adhere to the principles of good data visualization.
- **Please go thru the rubrics**

6 Assessment

6.1 Assessment Components

Assessment Component	Group	Individual	Total
Project proposal and Exploratory Data Analysis	20% 0%	0% 10%	30%
Interim Progress Review	0%	30%	30%
Final Presentation Report	10%	20% 10%	40%
TOTAL	**30%**	**70%**	**100%**

The assessment will be done through:

- Continuous assessment by Supervisors throughout the duration of the project.
- Assessment by independent markers during the presentations.

6.2 Assessment Rubrics

6.2.1 Exploratory Data analysis (Individual 10%)

Main Component	Sub Component	Achievement Level				
		Not Competent	Developing	Functional	Competent	Proficient
Exploratory Data Analysis (Individual 10%) Base mark: 20	**Interpretation** Marks: 5	No evidence that experiences of users were examined using an empathy tool/process	Identified key stages experienced by users using an empathy tool/process	Described thoughts or feelings of users at the different stages of the experience using an empathy tool/process	Identified pain points experienced by users using an empathy tool/process	Determined the needs of users using an empathy tool/process
	Interdisciplinary Thinking Marks: 5	No evidence that ideas were developed using perspectives from a different discipline	Adopted ideas from another discipline for users	Adapted ideas from another discipline to suit the needs of users	Used perspectives from another discipline to develop ideas for users	Combined perspectives from different disciplines to develop ideas for users
	Business Understanding Marks: 5	No evidence of data exploration on scenario datasets	Little or incorrect data exploration on scenario datasets	Basic data exploration on scenario datasets	Detailed data exploration on scenario datasets	Extensive data exploration on scenario datasets with interesting discoveries
	Data Understanding Marks: 5	No evidence of data exploration on scenario datasets	Little or incorrect data exploration on scenario datasets	Basic data exploration on scenario datasets	Detailed data exploration on scenario datasets	Extensive data exploration on scenario datasets with interesting discoveries

6.2.2 Project Proposal (Group 20%)

Main Component	Sub Component	Achievement Level				
		Not Competent	**Developing**	**Functional**	**Competent**	**Proficient**
Project Proposal (Group 20%) Base mark: 40	**Business Understanding** Marks: 10	No evidence of overall understanding & descriptions on business objectives, hypothesis, and project goal	Inadequate overall understanding & descriptions on business objectives, hypothesis, and project goal	Limited overall understanding and descriptions on business objectives, hypothesis, and project goals	Clear overall understanding and descriptions on business objectives, hypothesis, and project goals	Comprehensive overall understanding and descriptions on business objectives, hypothesis and project goals
	Proposal Report Marks: 10	Did not fully address any parts of the proposal (i.e. many missing content/parts) Content is not organized and hard to follow	Did not fully address some parts of the proposal (i.e. some missing content/parts) Content is not well organized and hard to follow	Fully addressed all parts of the proposal but could have done better Content is organized but could have done better	Fully addressed all parts of the proposal Content is well organized and clearly conveyed	Fully addressed all parts of the proposal with additional details/ insights Content is well organized and clearly conveyed with interesting discoveries/ insights
	Project Planning Marks: 10	No evidence that a project plan with key milestones was formulated	Identified desired outcome for each stage of an idea development	Determined milestones with consideration of desired outcomes for each stage of an idea development	Prepared a roadmap to deploy an idea based on key milestone	Prepared an action plan to achieve key milestones for the development of an idea
	Pitching Marks: 10	No evidence that persuasion technique was used to express the value of an idea to resource owners	Described an idea to resource owners using basic approaches from a persuasion technique	Described an idea to resource owners according to a persuasion technique	Explained the usefulness of an idea to resource owners using a persuasion technique	Expressed the value of an idea to resource owners in an impactful manner using a persuasion technique

6.2.3 Interim Progress Review (Individual 30%)

Main Component	Sub Component	Achievement Level				
		Not Competent	Developing	Functional	Competent	Proficient
Interim Progress Review (Individual 30%) **Base mark: 60**	**Data Collection** Marks: 10	No evidence of data collection	Inadequate collection of data	Limited/basic collection of data Obtain from some (1 to 2) data sources including structured and unstructured data sources.	Used various means of collecting & storing necessary data. Obtain from multiple data sources including structured and unstructured data sources.	Used extensive means of collecting & storing necessary data Obtain from multiple data sources including structured, unstructured and streaming data sources
	Data Management Marks: 10	No evidence that data on users were analyzed using a systematic process	Compiled data on users using a spreadsheet to begin a systematic process	Arranged similar data on users into groups according to a systematic process	Identified connections or correlations between groups of data on users through a systematic process	Determined key themes from the data on users through a systematic process
	Data Preparation Marks: 15	No evidence of data preparation	Irrational inclusion or exclusion of selected data Incorrect data cleaning and transformations	Simple rationale on inclusion or exclusion of selected data Complete most of the necessary data cleaning and transformation	Clear rationale on inclusion or exclusion of selected data Complete all necessary data cleaning and transformations	Thorough rationale on inclusion or exclusion of selected data Comprehensive data cleaning and transformations
	Data Visualization Marks: 15	No evidence of any visualization used	At least 1 visualization Visual representation is unable to and effectively show the characteristics of the data The choice of visual representation is unsuitable for the data type	At least 1 visualization to show the characteristics of the data The choice of visual representation is suitable but there are other better approaches	Multiple visualizations used to effectively show the characteristics of the data The choice of visual representation provides accurate message	Multiple visualizations used to clearly and effectively show the characteristics of the data Visual representation is intuitive The choice of visual representation provides accurate message
	Co-creating Marks: 10	No evidence that the project was co- developed with others	Gathered inputs from others on the idea	Exchanged information with others on the idea	Cooperate with others to develop the idea	Partner others in a mutually beneficial arrangement to jointly develop the idea.

6.2.4 Final Presentation (30%)

A) Group (10%)

Main Component	Achievement Level				
	Not Competent	**Developing**	**Functional**	**Competent**	**Proficient**
Integration (Group 10%) Base mark: 10	None of the notebooks are integrated.	Some of the notebooks are integrated but not functional.	Some of the notebooks are integrated and functional.	Majority of the notebooks are integrated and functional.	All notebooks are properly integrated and functional.

B) Individual (20%)

Main Component	Sub Component	Achievement Level				
		Not Competent	**Developing**	**Functional**	**Competent**	**Proficient**
Final Presentation (Individual 20%) Base mark: 20	**Iteration** Marks: 5	No evidence that feedback from users was used to modify the selected idea	Used feedback from users to improve the appearance of the selected idea	Used feedback from users to improve the concept of the selected idea	Reviewed feedback from users to make changes according to expectations of the selected idea	Analyzed feedback from users to improve user experiences with the selected idea
	Implementation Marks: 10	No evidence of predictive modelling techniques used, no models were built, and no recommendations made	Irrational selection of predictive modelling techniques At least 1 relevant predictive model were built with inappropriate test design and considerations Unable to relate model results to hypothesis and make recommendation	Simple rationale on the selection of predictive modelling techniques At least 2 relevant predictive models were built with basic test design and considerations Able to relate model results to hypothesis and make simple recommendations	Clear rationale on the selection of predictive modelling techniques At least 3 relevant predictive Models were built with appropriate test design and considerations Able to relate model results to hypothesis and make useful recommendations	Comprehensive rationale on the selection of predictive modelling techniques At least 4 relevant predictive **m**odels were built with detailed test design and considerations Able to relate model results to hypothesis and make novel recommendations
	Communication Marks: 5	No evidence that communication techniques were used to discuss an idea to stakeholders	Used a communication technique to explain an idea with stakeholders	Used multiple communication techniques to explain an idea with stakeholders	Applied an assortment of communication techniques to gather responses from stakeholders on an idea	Applied an assortment of communication techniques to engage stakeholders in a discussion on an idea

6.2.1 Final Report (Individual 10%)

Main Component	Achievement Level				
	Not Competent	Developing	Functional	Competent	Proficient
Final Report (Individual 10%) Base mark: 10	None of the notebooks are integrated. Content is badly organised and hard to follow	Some of the notebooks are integrated but not functional. Content is not well organized and hard to follow	Some of the notebooks are integrated and functional. Content is organized but could have done better	Majority of the notebooks are integrated and functional. Content is organized with recommendations clearly conveyed	All notebooks are properly integrated and functional. Content is well organized with recommendations clearly conveyed

The following is the assessment criteria for merit/demerit:

- Exemplary attitude, resourcefulness, initiative and additional enhancements, good team member/leader, team collaborates extensively and work effectively together
- Fails to meet deadlines, has poor attendance records or shows bad conduct, poor team spirit.

APPENDIX A – FORM FOR SUBMITTING PROJECT TEAM ORGANISATION

(TO BE COMPLETED AS A TEAM)

Module Group: _____

Project Group Number:_____ **Team Name (optional):** _____

S/n	Name	Admin No.	Handphone	Personal email
1*				
2				
3				
4				

***denotes the Team Leader

APPENDIX B – WEEKLY PROGRESS UPDATES

(TO BE COMPLETED AS A TEAM IN BLACKBOARD GROUP BLOG)

Each team is required to make use of the practical time to conduct a weekly team and progress update meeting.

Team leader is to ensure that the weekly progress updates and revised Gantt chart (if any) is captured in the Group Blog in Blackboard in the following format:

Blog Title:	**Progress Updates for <Week #>**
Blog Post:	**Team Member's Progress** 1. <Team Member Name> - <Work done for the week> 2. <Team Member Name> - <Work done for the week> 3. <Team Member Name> - <Work done for the week> 4. <Team Member Name> - <Work done for the week> **Gantt Chart (if any)** <Upload/ paste image of revised Gantt Chart> **Miscellaneous (if any)** <Upload/ insert any relevant info/ documents found or done>

Your supervisor will review the progress of the team based on the updates in the Group Blog.

APPENDIX C – PROJECT PROPOSAL

(TO BE COMPLETED AS A TEAM)

Module Group: _____

Project Group Number:_____ **Team Name:** _____

Description of Business

- Background

(introduce the business/organization that you're creating the analysis/solution for)

- Problem Statement

(describe in detail the targeted problem, the proposed hypotheses and projected outcome of the analysis. ***Each team needs to identify 1 issue/problem statement that they want to work on and gather feedback from at least 1 stakeholder****. Include news articles, statistics, stakeholder(s) feedback, etc. to support your problem statement.)*

- Hypothesis

(Each team needs to identify 1 issue/targeted problem and ***conjecture a set of hypotheses****, which will be verified or disproven by the analysis.* ***Each team member needs to identify at least 1 (sub) issue/ hypothesis*** *that he/she will try to address based on the issue/ problem statement that the team is trying to analyse.* ***Indicate Team Member name for each hypothesis listed.****)*

Proposed Solution

- Features

(give a complete description of the features you plan to provide in your analysis/solution. Include data sets, dashboard functions and possible insights. ***Each team member*** *needs to list down the names and URLs of* ***ALL*** *the possible data sources/ datasets that he/she has found and describe how he/she intends to use it and what kind of possible insights can be derived****. Indicate Team Member name for each feature listed.****)*

Miscellaneous

(anything else you wish to add about your project can be included here)

APPENDIX D – EXPLORATORY DATA ANALYSIS
(TO BE COMPLETED INDIVIDUALLY)

Module Group: _____

Project Group Number:_____ **Team Name:** _____

Description of Business

- Problem Statement / Hypothesis

(similar to Appendix C - describe in detail the targeted problem, the proposed hypotheses and projected outcome of the analysis)

Details of Datasets

- Scope of datasets
    - List down all possible data sources that *have been studied* and explain why a dataset is included for further analysis or excluded
      *(List down the names and URLs of **ALL** the possible data sources that you have **found** – and for each URL, give reasons why you have decided to include OR exclude them)*
- Common or individual datasets
  *(For this section, you will give details of **ALL** the datasets **used** in the entire project)*
    - Description of the dataset
      (Describe the dataset and the kind of data it contains. For example, this dataset contains NYP Student enrolment data from Year 2010 – 2020)
    - Source of the dataset, e.g., URLs
      (Include the URL for each dataset)
    - Type of the dataset, e.g., XML, CSV?
      (Format of your dataset)
    - Is the dataset shared or common?
      (Is this dataset shared/common? If not, list which team member(s) are using it.)
    - How is the dataset being used? e.g., selected fields or all fields?
      (Will the entire dataset be used? If not, list out the names of the column that you will be using.)
- Individual contribution *(Indicate Team Member name)*
    - List datasets being used
      (as specified in the previous section under Common or individual datasets)
    - List datasets being used
      (as specified in the previous section under Common or individual datasets)
    - Describe the data cleaning and transformation done
      (Explain briefly the steps required to clean your data source. For instance, you can state points like (1) Striped symbols and empty spacings from ALL columns, (2) Wrote Excel formulas to transform abbreviations into actual description, etc.)
    - How is the dataset related to problem / hypothesis?
      (Justify how your dataset is relevant to your team's hypothesis, for example, this dataset contains environmental data collected by NEA which helps us gather insights on… .....etc.)

Data Catalogue

(Describe the structures and fields of your datasets/ data sources. Draw ER diagrams to show the relationship between the various datasets if the datasets that you are using is related to other datasets. For example, if dataset A contains column B whose meaning can only be understood by interpreting column Z in the dataset)

Miscellaneous

(anything else you wish to add about your project can be included here)

APPENDIX E – FINAL PROJECT REPORT

(TO BE COMPLETED INDIVIDUALLY)

This report should follow the guidelines stated below. As a general rule, your report (excluding the cover page, contents page and the appendixes) **should not exceed 10 pages**. The report should be prepared using **one and a half line spacing**.

Header and/or footer should be included to show at least the project name, the chapter name, the file name of the document and the page number. Each section/chapter should begin on a new page with an appropriate heading. **Name of student** who has documented the section **should appear at the header of the section**.

You may include additional sections such as preface, acknowledgement, glossary of terms and symbols used, etc.

Recommended content outline:

i) **Cover Page**, this should reflect:
- Project Module Name, Document Title and Supervisor's Name
- The Project Team Number, Team Member's Name and Admin No. (stating roles played, e.g., Team Leader, Secretary etc.)
- The Latest Report Amendment Date

ii) **Contents Page**, this should list all the chapters/sections found in your report with their respective page number.

iii) **Executive Summary**
- Summary of business objectives, problem statement, hypothesis, or goals
- Describe the business scenarios, recommendation and insights discovered

v) **Problems Encountered**

vi) **Conclusion/Recommendations**
- Summarize the works, provide results interpretation, and conclude whether the outcome of the analysis fit the project expectation and propose a recommendation.

APPENDIX F – SELF AND PEER REVIEW
(TO BE COMPLETED INDIVIDUALLY)

This evaluation is designed to assist in the equitable distribution of assessment for each student in the project teams. Please record your own evaluation of the level of contribution (from 1 to 10) for yourself as well as other team members. Please put down relevant comments and reasons for justification purposes.

This evaluation is to protect your interest against any unfairness and thus you **DO NOT HAVE TO DISCUSS** with any of your team members. This evaluation is **STRICTLY CONFIDENTIAL**.

Module Group:_____ **Project Team**: _____

Supervisor Name: _____

All Team Members' Name (Including yourself)	Level of Contribution (1 to 10)	Comments / Reasons

Student Name:_____**Admin No**: _____

Signature:_____**Date**: _____

APPENDIX G – CHECKLIST OF QUALITY ATTRIBUTES FOR FINAL PRESENTATION

1. Delivery

- **Presentation Materials***
  - ☐ Content – Quality of info, Well Organized, Logical
  - ☐ Vocal – Clear and Audible
  - ☐ Visual – Effective use of visual aid
  - ☐ Verbal – Standard of English
  - ☐ Impact-full (Convincing and lasting impression)

- **Quality of Presentation**
  - ☐ Attire#
  - ☐ Eye Contact
  - ☐ Body Language (e.g. posture & mannerism)
  - ☐ Level of Confidence & Enthusiasm
  - ☐ Time Management

2. Q & A

- ☐ Ability to understand & respond
- ☐ Quality of response
- ☐ Level of confidence

 # Attire

*Dress code is **Smart Casual*** (no torn tops).

<u>For Ladies:</u>

- Top: polo tee / blouse *(no spaghetti strap / tube top please)*

<u>For Gentlemen:</u>

- Top: polo tee / shirts