******

***COMP3850 Project Deliverable Certificate***

| **Name of Deliverable** | *Deliverable 2* |
| --- | --- |
| **Date Submitted** | *29 / 08 / 2024* |
| **Project Group Number** | *14* |
| **Rubric stream being followed for this deliverable (highlight one)**  ***Note: the feasibility study has the same rubric for all streams.*** | *SOFTWARE Rubric*  *GAMES Rubric*  *CYBERSECURITY Rubric*  *DATA SCIENCE Rubric* |

We, the undersigned members of the above Project Group, collectively and individually certify that the above Project Deliverable, as submitted, **is entirely our own work**, other than where explicitly indicated in the deliverable documentation.

| **INITIALS** | **SURNAME** | **GIVEN NAME** | **STUDENT NUMBER** | **SIGNATURE  *(IN-PERSON OR DIGITAL)*** |
| --- | --- | --- | --- | --- |
| NM | Mahagoda | Ninuri | 46592156 | *NinuriM* |
| QN | Nguyen | Quoc Hung (Alan) | 47578130 | *AlanN* |
| AC | Chowdhury | Aasnayem Gazzali (Adam) | 46325824 | *Aasnayem* |
| NK | Khan | Noorullah | 47197404 | *N Khan* |
| TV | Vilathgamuwa | Tashiya | 46992162 | *Tvilathg* |

*NB: please write all details clearly (if handwritten).*

*© Macquarie University, 2021***List of tasks completed for the deliverable and activities since last deliverable certificate with totals for each individual team member and whole team** *(copy individual total row for each member and copy pages if more pages needed)*

| **Performed by  *(Names)*** | **Duration *(hrs)*** | **Complexity  *(L, M, H)*** | **Name of task** | **Checked by  *(Initials)*** |
| --- | --- | --- | --- | --- |
| Noorullah Khan | 1.5 | Mid | Meeting: Online Team Meeting week 3 (11 Aug) | NM |
| Noorullah Khan | 1 | Mid | Meeting: Online Team Meeting Week 4 (12 Aug) | NM |
| Noorullah Khan | 2 | Mid | Client: HDLC Options Analysis for EY | QN |
| Noorullah Khan | 1 | High | Client: Refining Options Analysis Presentation for EY | QN |
| Noorullah Khan | 1 | High | PM: Keeping PM tool updated and adding tasks | QN |
| Noorullah Khan | 0.5 | Mid | PM: Liasing with other team PM | QN |
| Noorullah Khan | 1 | High | Meeting: Weekly Meeting with Mentors at EY (13 Aug) | NM |
| Noorullah Khan | 1.5 | Mid | Meeting: Post EY weekly meeting with group 13 (13 Aug) | NM |
| Noorullah Khan | 1 | High | Meeting: with EY in EY office | NM |
| Noorullah Khan | 2.5 | High | Project: Creating database layout, researching data synthetics etc | AC |
| Noorullah Khan | 0.5 | Mid | Meeting: with Team 13 and Mentors 15th Aug | NM |
| Noorullah Khan | 0.5 | Mid | PM: Reviewing Project Management tool | QN |
| Noorullah Khan | 2 | High | Meeting: Saturday Team Meeting (17 Aug) | NM |
| Noorullah Khan | 0.5 | Low | PM: Shared PM information with other team | QN |
| Noorullah Khan | 1 | Mid | Project: Spoke with members of other team about Data | TV |
| Noorullah Khan | 1 | Mid | PM: Took documentation of ours and other teams PM charts | QN |
| Noorullah Khan | 1.5 | High | Meeting: Evening Meeting before EY (19 Aug) | NM |
| Noorullah Khan | 1 | High | Meeting: with client EY (20 Aug) | NM |
| Noorullah Khan | 4 | High | Meeting: Post-EY meeting | NM |
| Noorullah Khan | 1 | High | Research: Generating synthetic data: Computational complexity of GANs | AC |
| Noorullah Khan | 0.5 | Mid | Research: What is a Neural Network & types | AC |
| Noorullah Khan | 1 | High | Research: Big Data & R-trees | AC |
| Noorullah Khan | 1 | High | Development: Configuring Python Environment for ML modelling | AC |
| Noorullah Khan | 0.5 | Mid | Development: Trying out OpenAI's Codex for first time | AC |
| Noorullah Khan | 2 | Mid | Deliverable 2: data understanding | TV |
| Noorullah Khan | 2 | Mid | Deliverable 2: Deployment | TV |
| Noorullah Khan | 0.5 | Mid | Deliverable 2: Proof-reading sample for EY | AC |
| Noorullah Khan | 0.5 | Mid | PM: Communicating and organising with other team's PM | QN |
| Noorullah Khan | 0.5 | Mid | PM: Reading other team's Discord remaining in the loop | QN |
| Noorullah Khan | 1.5 | High | PM: Combining Gantt Charts & fixing Jira, making new sprints | NM |
| Noorullah Khan | 0.5 | Mid | Client: Powerpoint for EY week 6 | TV |
| Noorullah Khan | 0.5 | Mid | Meeting: with Eugene (Week 13) | NM |
| Noorullah Khan | 1.5 | Mid | Client: Powerpoint for EY week 6 | TV |
| Noorullah Khan | 0.5 | High | PM: Updating and Adjusting Jira | QN |
| Noorullah Khan | 1 | High | Meeting: with Team before EY (26 Aug) | NM |
| Noorullah Khan | 2 | High | Project: Working with Adam Understanding Data for Cleaning | AC |
| Noorullah Khan | 1 | High | Meeting: with EY (27 Aug) | NM |
| Noorullah Khan | 4 | High | Meeting: Post EY meeting (27 Aug) & development | NM |
| NInuri Mahagoda | 2 | Mid | Client: Created User Stories reflecting feedback from clients with Tash (10) | TV |
| NInuri Mahagoda | 2 | Mid | Client: Worked on Options Analysis for Services (11) | TV |
| NInuri Mahagoda | 1.5 | Mid | Meeting: Online Team Meeting week 3 (11 Aug) | AC |
| NInuri Mahagoda | 0.5 | Low | Research: Researched on Power BI (11) | QN |
| NInuri Mahagoda | 1 | Mid | Meeting: Online Team Meeting Week 4 (12 Aug) | AC |
| NInuri Mahagoda | 1 | High | Meeting: Weekly Meeting with Mentors at EY (13 Aug) | AC |
| NInuri Mahagoda | 1.5 | Mid | Meeting: Post EY weekly meeting with group 13 (13 Aug) | AC |
| NInuri Mahagoda | 2.5 | Mid | Meeting: Post EY weekly meeting (13 Aug) | AC |
| NInuri Mahagoda | 1.5 | Mid | Client: Worked on Updated Options Analysis Report for EY (14) | TV |
| NInuri Mahagoda | 1.5 | High | Meeting: Evening Meeting before EY (19 Aug) | AC |
| NInuri Mahagoda | 1 | Mid | Meeting: with Tash to create EY pack (19/08/24) | AC |
| NInuri Mahagoda | 2 | High | Meeting: Saturday Team Meeting (17 Aug) | AC |
| NInuri Mahagoda | 1 | High | Meeting: with client EY (20 Aug) | AC |
| NInuri Mahagoda | 4 | High | Meeting: Post-EY meeting | AC |
| NInuri Mahagoda | 2 | Mid | Deliverable 2: Worked on Scoping/Requirements Document: Data Modelling with Tash (22) | TV |
| NInuri Mahagoda | 1 | Low | Deliverable 2: Editing and Formatting Deliverable 2 Draft with Tash (22) | TV |
| NInuri Mahagoda | 2 | Mid | Deliverable 2: Project Plan - Quality Management (23) | TV |
| NInuri Mahagoda | 2 | Mid | Deliverable 2: Project Plan - Schedule | NK |
| NInuri Mahagoda | 1 | Low | Research: Understanding Jira Platforma and creating diagrams | QN |
| NInuri Mahagoda | 1 | High | Meeting: with Team before EY (26 Aug) | AC |
| NInuri Mahagoda | 1 | Mid | Meeting: with Tash to create EY pack (26 Aug) | TV |
| NInuri Mahagoda | 1 | Mid | Research: Data Cleaning Methods | NK |
| NInuri Mahagoda | 1.5 | High | Research: Data Convention Format | NK |
| NInuri Mahagoda | 1 | High | Meeting: with EY (27 Aug) | AC |
| NInuri Mahagoda | 4 | High | Meeting: Post EY meeting (27 Aug) & development | AC |
| NInuri Mahagoda | 1.5 | Mid | Deliverable 2: Project Plan - Assumptions (27) | TV |
| NInuri Mahagoda | 1 | Mid | Client: Created More User Stories for client based in feedback with Tash (27) | TV |
| NInuri Mahagoda | 1.5 | Mid | Deliverable 2: Project Plan - Update Quality Management (23) | QN |
| NInuri Mahagoda | 1 | Mid | Deliverable 2: Project Plan - Update Schedule (27) | QN |
| NInuri Mahagoda | 1 | Mid | Deliverable 2: Updated Team Manual (27) | QN |
| NInuri Mahagoda | 1 | High | Deliverable 2: Project Scope with Tash (28) | TV |
| NInuri Mahagoda | 2 | Mid | Deliverable 2: Editing and Formatting Deliverable 2 with Tash (29) | TV |
| Tashiya Vilathgamuwa | 2 | Mid | Client: Created User Stories reflecting feedback from clients with Ninuri (10) | TV |
| Tashiya Vilathgamuwa | 1.5 | Mid | Meeting: Online Team Meeting week 3 (11 Aug) | QN |
| Tashiya Vilathgamuwa | 2 | Mid | Client:Worked on Options analysis to present to EY | NM |
| Tashiya Vilathgamuwa | 1 | Mid | Meeting:Online Team Meeting Week 4 (12 Aug) | QN |
| Tashiya Vilathgamuwa | 1 | High | Meeting: Weekly Meeting with Mentors at EY (13 Aug) | QN |
| Tashiya Vilathgamuwa | 1.5 | Mid | Meeting: Post EY weekly meeting with group 13 (13 Aug) | QN |
| Tashiya Vilathgamuwa | 2 | High | Meeting:Saturday Team Meeting (17 Aug) | QN |
| Tashiya Vilathgamuwa | 2.5 | Mid | Meeting: Post EY weekly meeting (13 Aug) | QN |
| Tashiya Vilathgamuwa | 1.5 | Mid | Client:Worked on Updated Options Analysis Report for EY (14) | NM |
| Tashiya Vilathgamuwa | 1.5 | High | Meeting:Evening Meeting before EY (19 Aug) | QN |
| Tashiya Vilathgamuwa | 1 | Mid | Meeting: with Ninuri to create EY pack (19 Aug) | QN |
| Tashiya Vilathgamuwa | 1.5 | Mid | Meeting: with Eugene (Team 13) to finalise presenation (19 Aug) | QN |
| Tashiya Vilathgamuwa | 0.5 | Mid | Meeting: with Alan (18 Aug) | QN |
| Tashiya Vilathgamuwa | 1 | High | Meeting: with client EY (20 Aug) | QN |
| Tashiya Vilathgamuwa | 4 | High | Meeting: Post-EY meeting | QN |
| Tashiya Vilathgamuwa | 2 | Mid | Deliverable 2: Worked on Scoping/Requirements Document: Data Modelling with Ninuri (22 Aug) | NM |
| Tashiya Vilathgamuwa | 1 | Low | Deliverable 2:Editing and Formatting Deliverable 2 Draft with Ninur (22 Aug) | NM |
| Tashiya Vilathgamuwa | 1 | High | Meeting: Meeting with Team before EY (26 Aug) | QN |
| Tashiya Vilathgamuwa | 1 | High | Meeting: Meeting with EY (27 Aug) | QN |
| Tashiya Vilathgamuwa | 4 | High | Meeting: Post EY meeting (27 Aug) & development | QN |
| Tashiya Vilathgamuwa | 1 | Mid | Meeting: Meeting with Ninuri to create EY pack (26 Aug) | NM |
| Tashiya Vilathgamuwa | 2 | Mid | Meeting:Meeting with Eugene (Team 13) to finalise presentation (26 Aug) | QN |
| Tashiya Vilathgamuwa | 2 | Mid | Deliverable 2: Project Plan - Resource management (27) | AC |
| Tashiya Vilathgamuwa | 2 | Mid | Client: Created More User Stories for client based in feedback with Ninuri (27) | NM |
| Tashiya Vilathgamuwa | 2 | Mid | Deliverable 2: Project Plan - Risk Management (23) | AC |
| Tashiya Vilathgamuwa | 1 | Mid | Deliverable 2: Project Plan - Change management (27) | AC |
| Tashiya Vilathgamuwa | 1 | Mid | Deliverable 2: Updated Team Manual (27) | AC |
| Tashiya Vilathgamuwa | 1 | Mid | Deliverable 2: Project Scope with Niruri (28) | NM |
| Tashiya Vilathgamuwa | 2 | Mid | Deliverable 2: Editing and Formatting Deliverable 2 with Ninuri (29) | nm |
| Tashiya Vilathgamuwa | 0.5 | Mid | Research: Synthesisation methods | NK |
| Quoc Hung (Alan) Nguyen | 1.5 | Mid | Meeting: Online Team Meeting week 3 (11 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 1 | Mid | Meeting: Online Team Meeting Week 4 (12 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 1 | High | Meeting: Weekly Meeting with Mentors at EY (13 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 1.5 | Mid | Meeting: Post EY weekly meeting with group 13 (13 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 2.5 | Mid | Meeting: Post EY weekly meeting (13 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 0.5 | Mid | Meeting: Meeting | NK |
| Quoc Hung (Alan) Nguyen | 2 | High | Meeting: Saturday Team Meeting (17 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 1.5 | High | Meeting: Evening Meeting before EY (19 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 0.5 | Mid | Meeting: with Tash (18 Aug) - about presentation | TV |
| Quoc Hung (Alan) Nguyen | 0.25 | Low | Meeting: with Adam (21 Aug) - about part 4 of scoping doc | AC |
| Quoc Hung (Alan) Nguyen | 0.5 | Low | Infrastructure refinement - learn about Entra ID | AC |
| Quoc Hung (Alan) Nguyen | 1 | High | Infrastructure refinement - learn about Application Insights | AC |
| Quoc Hung (Alan) Nguyen | 0.25 | Low | Infrastructure refinement - learn about Key Vaults | AC |
| Quoc Hung (Alan) Nguyen | 0.5 | High | Infrastructure refinement - compare Terraform and Ansible | AC |
| Quoc Hung (Alan) Nguyen | 1.5 | Mid | Client: Make changes to Option Analysis after Infrastructure refinement | NM |
| Quoc Hung (Alan) Nguyen | 0.25 | High | Terraform - how to set up a terraform folder structure | NM |
| Quoc Hung (Alan) Nguyen | 2 | High | Terraform - learn to write .yaml for Terraform | AC |
| Quoc Hung (Alan) Nguyen | 3 | High | Terraform - learn about Terraform Azurerm | AC |
| Quoc Hung (Alan) Nguyen | 0.5 | High | Terraform - learn to configure inputs, outputs, variables and secrets | AC |
| Quoc Hung (Alan) Nguyen | 2 | Mid | Terraform - tutorial walkthrough on youtube to build a test project | AC |
| Quoc Hung (Alan) Nguyen | 1 | High | Terraform - write .yaml files for project infrastructure (test, not working) | AC |
| Quoc Hung (Alan) Nguyen | 2 | High | Azure - how to set up Azure Functions with Azure CLI and portal | TV |
| Quoc Hung (Alan) Nguyen | 0.75 | High | Azure - learn about Function Apps and how to configure it | TV |
| Quoc Hung (Alan) Nguyen | 0.5 | High | Azure - test Azure Functions configurations and triggers | TV |
| Quoc Hung (Alan) Nguyen | 1.5 | High | Azure - set up a whole project with Azure with Azure CLI (test) | TV |
| Quoc Hung (Alan) Nguyen | 1 | Low | Azure - play around with Application Insights, Key Vault and Entra ID | TV |
| Quoc Hung (Alan) Nguyen | 0.2 | Mid | Azure - write User Guide for Local Development of Azure Functions | TV |
| Quoc Hung (Alan) Nguyen | 1.5 | High | Github - learn about branching strategies | NM |
| Quoc Hung (Alan) Nguyen | 0.5 | Low | Github - learn about issues and project board | NM |
| Quoc Hung (Alan) Nguyen | 2 | High | Github - set up test repo to draft repo structure and workflows | NM |
| Quoc Hung (Alan) Nguyen | 2 | High | Github - learn to use Github actions (ongoing) | NM |
| Quoc Hung (Alan) Nguyen | 1 | Mid | Github - write User Guide for Github folder structure (team use only) | NM |
| Quoc Hung (Alan) Nguyen | 1 | Mid | Github - write User Guide for Github workflows (team use only) | NM |
| Quoc Hung (Alan) Nguyen | 1 | Mid | Github - set up final repo with issues and project board | NM |
| Quoc Hung (Alan) Nguyen | 1 | Mid | Deliverable 2 - writing | TV |
| Quoc Hung (Alan) Nguyen | 1 | High | Meeting: with client EY (20 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 4 | High | Meeting: Post-EY meeting | NK |
| Quoc Hung (Alan) Nguyen | 1 | High | Meeting: with Team before EY (26 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 1 | High | Meeting: with EY (27 Aug) | NK |
| Quoc Hung (Alan) Nguyen | 4 | High | Meeting: Post EY meeting (27 Aug) & development | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1.5 | Mid | Meeting: Online Team Meeting week 3 (11 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Meeting: Online Team Meeting Week 4 (12 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | Mid | Client: Worked on HDLC options analysis for EY | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Meeting: Weekly Meeting with Mentors at EY (13 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1.5 | Mid | Meeting: Post EY weekly meeting with group 13 (13 Aug) | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 2.5 | Mid | Meeting: Post EY weekly meeting (13 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Client: Updated options analysis | QN |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | High | Project: Worked on the database layout with Noor | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1.5 | High | Project: Made a sample database containing 100 rows of data | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Research: Researched data synthesization technique - Scikit-learn | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Research: Researched data cleaning techniques | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 0.5 | Mid | Research: Reviewed 'nltk' library in python | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Research: Researched data transformation techniques | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Research: Researched about Neural Networks | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Project: Created a python script to separate Fraud and Not Fraud rows into two separate files | QN |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | High | Meeting: Saturday Team Meeting (17 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1.5 | High | Meeting: Evening Meeting before EY (19 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 0.25 | Low | Meeting: with Alan (21 Aug) - about part 4 of scoping doc | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Meeting: with client EY (20 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 4 | High | Meeting: Post-EY meeting | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 2.5 | Mid | Deliverable 2: Worked on Data Preparation | NM |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | Mid | Deliverable 2: Worked on Data Evaluation | NM |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | High | Project: Tested the "make\_classification" in scikit-learn to try and synthesise data | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Meeting: with Team before EY (26 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 0.25 | Mid | Research: Reviewed excel formulas | QN |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | High | Project: Added 900 rows of data to create a dataset with 1000 rows | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Research: Researched Data Integration Methods | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 2 | High | Project: Working with Noor Understanding Data for Cleaning | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | High | Meeting: with EY (27 Aug) | TV |
| Aasnayem Gazzali Chowdhury(Adam) | 4 | High | Meeting: Post EY meeting (27 Aug) & development | QN |
| Aasnayem Gazzali Chowdhury(Adam) | 0.5 | Mid | Research: Reviewed datasets | NK |
| Aasnayem Gazzali Chowdhury(Adam) | 1 | Mid | Deliverable 2: Updated Deliverable 2 based on the feedback | NM |
| **Ninuri Total** | 49 |  |  |  |
| **Tahiya Total** | 48 |  |  |  |
| **Noorullah Total** | 47 |  |  |  |
| **Alan Total** | 48.5 |  |  |  |
| **Adam Total** | 47.5 |  |  |  |
| **TEAM TOTAL** | **240** |  |  |  |

# **Table of Contents**

[Table of Contents 11](#_zhk6r73hzqee)

**Project Plan**

[1. Introduction 15](#_ey84htrmwg41)

[1.1. Statement Of Purpose 15](#_qjhs061mrs6d)

[1.2. Scope 15](#_lkut0qrfjjxf)

[1.2.1 Justification 15](#_4uf4b0jgqo2y)

[1.2.2. Scope 16](#_rjyuljif6o41)

[1.2.3. Objectives 16](#_xyjk61ckia3q)

[1.2.4. Deliverables 16](#_54jjshzg3zw2)

[1.2.5. Exclusions (Out of Scope) 16](#_8v8472ht4ytg)

[1.2.6. Constraints 16](#_n265lwlww75s)

[1.2.7. Assumptions 17](#_6rg2mkgr1ak)

[2. Risk Management 17](#_lqjosrusuuua)

[2.1 Risk Analysis Table 17](#_g57y68clqzhr)

[2.2 Risk Monitoring 17](#_la9wnt7cbt6d)

[2.3 Risk qualification and prioritisation 17](#_g1eo1ea9bjrs)

[3. Resource Management 18](#_2y3v9jdga453)

[3.1 People: 18](#_44sxmq3i7ou8)

[3.2 Hardware: 18](#_hwh7gbdfqw4f)

[3.3 Software: 19](#_qob3qjqm8oue)

[3.4 Other Resources: 20](#_mbb3znk6zq6e)

[4. Change Management 21](#_cb5ra97yr809)

[4.1 Changes to requirements and scope 21](#_xeiv48wsckg)

[3.2 Changes to documents, code and data 21](#_to5qqghujx3x)

[3.2.1 Branching strategy: Issue-based 22](#_jedd9268oajn)

[3.2.2 Code Reviews: Enforcing Quality with Pull Requests and Github Actions 22](#_t1t9k8yz6kdg)

[3.2.3 Documentation versioning 22](#_e7o84684gu1a)

[3.2.4 Immutable storage: 23](#_cwyfyno7l3dh)

[5. Quality Management 23](#_lekybt1055ne)

[5.1. Objectives 23](#_h6i4fbli6icu)

[5.1.1. Ensuring data quality 23](#_tp7xfigdsi55)

[5.1.2. Model Accuracy 24](#_dwbw2mcsigel)

[5.1.3. Compliance with business requirements 24](#_gohx5xep14nk)

[5.1.4. Customer Focus and User Satisfaction 24](#_qtbo2vktu0pk)

[5.2. Data Quality Management: Monitoring and Assurance 24](#_d9qynmim2cp7)

[5.2.1. Data Collection Process 24](#_qy1g9rvf6euw)

[4.2.2. Data Cleaning 25](#_5zdknymy0jxw)

[5.2.3. Data Standardisation 25](#_wtehbtbxtwqy)

[5.3. Human Resource Management: Agile Delivery 26](#_fkxjyeyjsrks)

[5.3.1. Sprint Planning 26](#_71loxbsf5llm)

[5.3.2. Iteration Reviews 26](#_g28oahdueu17)

[5.3.3. Continuous Feedback Loop 26](#_etfw87v3wn8l)

[5.4. Development Management 27](#_8vusy3gogao5)

[5.4.1. Code convention 27](#_obln1aviwgdl)

[5.4.2. Prototype 27](#_fbalu9o460av)

[5.5. Documentation Management 27](#_jmrubmu6lojh)

[5.5.1. Formatting 27](#_jv1w7xz3cttp)

[5.5.2. Documentation control officers 27](#_gvcm2nvszrgf)

[6. Schedule 27](#_entroc15wpqc)

[6.1. Project Overview 27](#_hn2tdt7nve5)

[6.1.1. Deliverables 28](#_lrkiwbincapt)

[6.2. Project Phases and Key Milestones 29](#_5j8wvf4ignmv)

[6.3. Schedule - Gantt Chart 29](#_avcw20yzhg5a)

[6.4. Resource and Task Allocation 29](#_ftqw611qqx4g)

[6.5. Tools and Technologies 30](#_2ruauu2u53xh)

[6.5.1. Jira 30](#_7pkb895335j)

[6.5.2. GitHub 30](#_xekn66svh0kl)

[7. Assumptions 30](#_gxmzuce4l6nz)

[7.1. Resources 30](#_x44v8bi9ibga)

[7.2. Availabilities 30](#_26yw3gq2l76c)

[7.3. Tools 30](#_tj2rlx15qgic)

[7.4. Techniques 31](#_rv41gm4hl33o)

[7.5. Standards 31](#_wdmc9n7b1giz)

[7.6. Communication 31](#_qugceikd6nz1)

[7.7. Expectations 32](#_m9p6ygrvpycn)

[8. References 33](#_7w2zlxn99xlw)

[9. Appendices 34](#_j35rc0g36jpe)

[Appendix 1. Risk analysis table 34](#_qy7uoh9zx8ow)

[Appendix 2. Risk Indicators Table 35](#_tyldfwk9mmf4)

[Appendix 3 Probability/ Impact Matrix 36](#_u5cm5re7oncu)

[Appendix 4: Project Phases and Key Milestones 37](#_ktl5ycyng7bp)

[Appendix 5: Sprint 1 Gantt Chart 38](#_xkx4usff5wy7)

[Appendix 6: Sprint 2 Gantt Chart 39](#_rkrl6hlrdzkh)

[Appendix 7: Sprint 3 Gantt Chart 40](#_1t7oxiho8kc1)

[Appendix 8: Task Allocation Groups 41](#_e437qncfqmrd)

**Requirements/Scoping Document**

[1. Data Understanding 42](#_1yjkh1anmujp)

[1.1 Initial Data Sources 42](#_nb88drf4p74s)

[1.2 Data Collection Methods 42](#_j86cvkk8rvlu)

[1.3 Assumed Data Quality, Structure, and Completeness 44](#_31r487wyurrz)

[1.4 Mechanisms for Evaluation 45](#_wsa4qo3jwpe6)

[2. Data Preparation 45](#_78jq27ck99z1)

[2.1. Data Preparation Activities 45](#_cw9vkorct2yx)

[2.2. Data Processing Pipelines 46](#_vsglpkaap9vo)

[2.3. Different Preparation Pipelines 46](#_kjgxuavyoxm9)

[2.4. Justification for Different Models 47](#_4bqm53ndl4c2)

[2.5. Data Storage Mechanisms 47](#_zh4yyeka4p9s)

[3. Modelling 47](#_2jxsxqh)

[3.1. Modelling Techniques 47](#_ok62gylqhmrx)

[3.1.1. Neural Networks: 48](#_i534jd8at0dg)

[3.1.2. Logistic Regression Model: 48](#_vev4davqynx)

[3.1.3. Decision Tree: 49](#_7hp7gqjd5wo7)

[3.1.4. Relationship to the Chosen Model: 49](#_4702xt9sbv7y)

[3.2. How will the Effectiveness and Validity be checked? 50](#_4vi4vz8c233c)

[3.3. Test Dataset 50](#_1y810tw)

[4. Evaluation 51](#_2xcytpi)

[4.1. Evaluating Results 51](#_1ci93xb)

[4.2. Review Process 52](#_3whwml4)

[4.3. Next Steps 52](#_2bn6wsx)

[5. Deployment 53](#_qsh70q)

[5.1. Deployment Plan: 53](#_3as4poj)

[5.2. Monitoring and Maintenance: 53](#_49x2ik5)

[5.3. Final Report and Project Review: 54](#_147n2zr)

[6. References 54](#_2wvhept83mj7)

[7. Feedback and Response 54](#_o2z4lvrz5okw)

**Team Manual**

[1. Team Organisation and Structure 57](#_h51d8fyuz0vb)

[2. Team Values & ACS Code of Professional Ethics 59](#_swzr988wnlyn)

[2.1. ACS Code of Professional Conduct 59](#_elv22pnri0u7)

[2.2. Honesty 59](#_rsft776bqpik)

[2.3. Trustworthiness 59](#_yvngz9yrynap)

[2.4. Respect 59](#_p97vmuryl7h0)

[3. Communication Plan and Meet Schedule 60](#_9o4dopri62zt)

[3.1 Communication 60](#_ujlctmz1edxj)

[3.2 Meet Schedule 61](#_xp07llg59fq2)

[4. Conflict Resolution/Negotiation 62](#_a73whcyuvf9u)

[4.1. Conflict resolution 62](#_rgbgbdugtwqa)

[4.2. Escalation process 62](#_y1z2y2erfgk5)

[4.3. Preventative steps[Updated]: 62](#_8iqsxfr6qphj)

[5. Revision History Table 63](#_mk1gden1ras7)

[6. References 64](#_82tcbiqw0wgb)

**Project Plan**

# **Introduction**

## **1.1. Statement Of Purpose**

The purpose of this document is to showcase a comprehensive project plan on the formation of a fraud detection system for NRMA’s claims processing system using Generative AI (GenAI). To address this EY has split us into the AI and Data Teams to develop an end-to-end solution upgrading their current fraud detection process. The project plan was created to communicate with the stakeholders on the goals, objectives and tasks needed to complete the project throughout all phases of the project lifecycle.

The Project plan will include:

1. **Risk Management:** Managing risks effectively is crucial for the project's success. This risk management plan will outline how to identify, evaluate and address potential risks at every stage of the project.
2. **Resource Management:** Resource management aims to maximise the efficient use of people, hardware, software, and other resources. This part of the project plan will detail strategies for optimal resource allocation ensuring that the project will have enough resources to meet its objectives while adhering to budget and time limits.
3. **Change Management:** Changes to requirements, scope and updates to documents, code and data will be managed through a well defined process. This section of the project plan will delve deeper into how it will be managed and reviewed systematically.
4. **Quality Management:** Effective quality management is pivotal for the success of this project. The quality management plan outlines the approach to ensuring that all project deliverables meet the established standards and stakeholder expectations.The plan encompasses quality planning, assurance, and control processes designed to guarantee that the project adheres to high standards of quality throughout its lifecycle.
5. **Schedule:** The schedule section will follow the project overview, deliverables, task and resource allocation.
6. **Assumptions:** In this section, any assumptions made during the planning and execution phase of the project that may impact the project’s outcome will be raised.

## **1.2. Scope**

### **1.2.1 Justification**

This project was established as a way of improving NRMA’a current fraud detection system in automobile claim reports by leveraging GenAI as a means of increasing the accuracy and efficiency of the process. This helps create a robust tool for NRMA to streamline the claims process and decreases the amount of operational costs and time spent on reviewing claims.

### **1.2.2. Scope**

The scope of the project includes the development and provision of a dataset designed for training and evaluating a fraud detection model. This entails:

* **Data Acquisition:** Collecting and sourcing relevant automobile claim information
* **Data preparation:** Cleaning, organising, and enhancing the dataset to ensure it is ready for model training.
* **Dataset Delivery:** Delivering the completed dataset to the AI team, along with necessary documentation and user guides.

### **1.2.3. Objectives**

* To develop a clean, accurate and comprehensive dataset for the AI model to be trained.
* To improve EY’s existing fraud detection system by supplying high quality data for training and evaluating models.
* To complete the project within the set timeline and budget, ensuring adherence to the defined quality standards.

### **1.2.4. Deliverables**

* **Sprint 1**: Create user stories and source datasets
* **Sprint 2**: Project Plan, Requirements Document, Options Analysis Report, Test Data File.
* **Sprint 3**: Design Document, Test Cases Document, Prototype/MVP.
* **Sprint 4**: Updated Prototype, User/Training Manual.
* **Sprint 5**: Final Group Reflective Report.
* **Sprint 6**: Final Product Dataset (Bronze, Silver, Gold Layers), Final Report.

### 

### **1.2.5. Exclusions (Out of Scope)**

* Development of the AI model itself.
* Post deployment support or maintenance for the AI model or dataset
* Front end development
* Visualisation (PowerBI, Tableau etc.)

### **1.2.6. Constraints**

* **Timeline**: The project must be finalised within the designated six sprints, with deadlines established for each phase.
* **Budget**: Adherence to the project budget is required, covering all expenses related to tools, resources and any unexpected costs.
* **Data availability:** The quality and comprehensiveness of the dataset depend on the accessibility and availability of relevant data sources.

### **1.2.7. Assumptions**

* The synthesiation techniques that have been chosen are able to synthesise 7,000 rows of high quality data
* Sythesised data will be of defined standard and will be of enough quality to train the AI model
* The Azure Tools, Azure Functions and Data Lake will integrate well with each other and be enough to complete the project

# **2**. **Risk Management**

Ensuring the completion of our project relies heavily on effective risk management practices in place from the start to finish of the project journey ahead of us(Smith,2020). Outlined in the sections are the key risks we've pinpointed within the project scope arranged by type alongside a thorough examination of their likelihoods and potential impacts and strategies geared toward mitigating these risks effectively. Our proactive approach, to identifying and tackling these pitfalls early on is geared towards reducing any possible interruptions and increasing the probability of reaching our project objectives successfully.

## **2.1 Risk Analysis Table**

Appendix 1 below outlines the different risks related with different aspects of the project ranging from technology and people related risks to organisational, requirements and estimation risks. Each risk is evaluated based on probability of occurrence and the impact it can have on the project. Moreover, appropriate mitigation strategies are proposed to manage these risks effectively. Refer to Appendix 1.

## **2.2 Risk Monitoring**

Our team has identified numerous indicators to help aiding us to monitor each type of risk. These indicators help us to constantly check whether a risk is becoming more likely. High priority risks are closely monitored, but low priority risks are also being continuously tracked to ensure they do not escalate unexpectedly. The indicators are summarised in the risk indicators table (Appendix 2).

## **2.3 Risk qualification and prioritisation**

To assess the seriousness of the risks highlighted in appendix 3 a combination of probability and impact factors has been integrated into the risk register to assist the project team in ranking risks according to their likelihood and potential impact on the project's outcomes. Project managers make use of a probability and impact matrix to aid in identifying risks that demand attention and action items during project planning and execution phases (Li,2020). The matrix depicted in Appendix 3 assists the team in identifying the high priority risks marked in the red areas while categorising medium priority risks in orange zones and low priority risks, in green.

# **3. Resource Management**

To ensure the successful execution of our project, a well structured resource allocation and management strategy has been implemented(Armstrong, 2023). This strategy focuses on optimising the use of people, hardware, software and other resources while ensuring flexibility and responsiveness to project needs.

## **3.1 People:**

People resource management is important as it ensures that tasks are assigned to the most qualified individuals, optimises team performance and adapts to changes effectively.

* **Role Assignment:** Each team member has been assigned specific roles based on their expertise and strengths. The project manager (Noorullah) oversees the overall project , while specialised roles like data engineer, business analyst and cloud architect ensure that tasks are handled by the most qualified individuals.
* **Task Prioritisation:** Tasks are prioritised based on project milestones, with critical tasks assigned to team members with relevant expertise. This ensures that high impact tasks receive the attention they need for timely completion.
* **Cross training:** To mitigate risks associated with resource availability, cross training among team members is encouraged. For example, Business Analysts (Tashiya and Ninuri) are trained in python coding in order to assist Data engineers when needed.

**Management Strategy:**

* **Weekly meetings:** regular face-to-face and online meetings are held to review progress, reassign tasks if necessary and address any bottlenecks. This ensures that the team remains agile and can quickly adapt to changes.
* **Performance tracking:** Individual and team performance are tracked using Jira work management. This allows for real time monitoring of task completion and resource utilisation, enabling timely interventions of resources that need to be reallocated.

## **3.2 Hardware:**

For team members to have the essential equipment to execute tasks effectively, effective hardware management is needed.

**Laptops/ Desktop computers:** Because of the scope of our project, the only mandatory hardware the team needs is their laptops or desktop computers. These devices are essential for development, data modelling and analysis tasks. Each team member’s computer must be capable of handling the necessary processing demands, with a reliable internet connection for accessing resources, collaborating via discord and participating in online meetings. Since all critical infrastructure is cloud based no additional physical hardware is needed.

**Management strategy:**

* **Standardisation:** All team members are required to have laptops or computers that meet minimum standard in terms of processing power, memory and storage capacity to ensure smooth operations and meet Visual studio code version 1.92 requirements. Laptops used need to be able to have a working camera to join group meetings to facilitate open communication and ensure team building.
* **Maintenance and security:** Each member will be responsible for their own devices security through ensuring their devices are updated with the latest security patches and software updates. Consistent backups of important data is also encouraged to prevent any loss of progress in case of any hardware failure that may occur.

## **3.3 Software:**

Effective software management is vital as it ensures that the tools and applications used by the team are reliable, up to date and suited for our project needs.

* **Github:** All code and project related documentation are stored in Github repositories. Team members are assigned appropriate access levels to ensure secure and organised appropriate access levels to ensure secure and organised code management. Collaboration is facilitated through pull requests, code reviews, and issue tracking.
* **Jira**: Tasks and project milestones are documented and tracked in Jira. Team members are assigned tasks and responsibilities according to their roles. Jiras Gannt charts and task dependencies help visualise project timelines and manage workflows effectively.
* **Azure Data functions:** Resources are allocated based on the computational requirements of data processing tasks. Azure data functions are scaled automatically according to the workload, ensuring efficient use of resources and cost-effectiveness.
* **Azure Data lake Storage :** Data is organised and stored in Azure Data Lake Storage according to its classification and access needs. Storage costs are monitored and data is managed to optimise storage usage and ensure security and compliance.

**Management strategy:**

* **Licensing and access control:** Ensure that all software licences are upto date and that team members have the necessary permissions and access levels to perform tasks effectively.
* **Integration and compatibility:** Regurualry verify that software tools integrate seamlessly and are compatible with each other to avoid disruptions in workflow.
* **Support and maintenance:** Regularly review software performance and make adjustments as needed to meet project requirements. Utilsing support service provided by the software vendors we have used for troubleshooting and resolving any issues that may arise.
* **Training**: Our team will be trained on how to use each software tool effectively. With softwares like Github, How to guides have been created to ensure all team members are using the software correctly and ensuring no issues will occur. Moreover, User guides have been created on how to develop azure functions on local machines as well, therefore helping us on how to set this process up and what to learn to start developing.

## **3.4 Other Resources:**

In addition to people, hardware and software our project requires the management of various other resources that contribute to the successful completion of the project.

* **Documentation and knowledge repositories:** Centralised documentation that includes project plans, meeting notes, technical documentation and training materials. This is crucial for maintaining consistency, enabling knowledge sharing and ensuring that all team members are aligned. All project documentation is stored in a shared accessible location (Google Drive and Github). Team members are responsible for maintaining up to date documentation and specific roles are assigned to oversee the consistency and accuracy of the knowledge repositories.
* **Communication platforms:** Tools such as Discord and email are used for internal and external communication. These platforms are essential for daily interactions, quick updates and formal communication with mentors. Discord is the primary communication platform for more day to day interactions. While email is used for more formal communication. Specific channels and threads within discord are dedicated to different aspects of the project to maintain organised communication.

**Management strategy:**

Documentation and knowledge repositories management:

* **Regular updated:** Documentation and knowledge repositories will be regularly updated to reflect the latest project developments. Routine audits will take place to ensure all documentation is accurate, comprehensive and aligns with the project's process.
* **Version control:** Implementation of version control practises in Github and google drive to track changes and maintain the integrity of documentation. This will ensure that any previous versions can be retrieved if necessary.
* **Centralised access**: All team members will have appropriate access to these repositories and the structure will be easy to navigate with the clear guidelines on where and how to store and access documents.

Communication platform management:

* **Channel organisation:** WIthin discord, dedicated channels will be created for different aspects of the project to keep communication organisation and relevant.
* **Clear protocols**: Established protocols when to use discord versus email, ensuring that urgent or day to day updates are communicated through discord while formal communication is through email.
* **Security and privacy:** Ensured that all communication platforms are secured with appropriate privacy settings, and have encouraged the use of encrypted channels for any sensitive information sharing.

# **4. Change Management**

## **4.1 Changes to requirements and scope**

During the project, our client revised their requirement to exclude data visualisation as part of the minimum viable product (MVP) due to budget constraints. Initially the project scope included using tools like Power BI and Tableau to create visual representation of the data and several team members were undergoing research and training to utilise these tools effectively.

Impact of the change:

* **Training adjustment:** Team members who were in the process of getting trained on utilising power BI have now refocused their efforts on other critical areas of the project like python, data cleaning and data synthesising.
* **Scope Revision:** The removal of data visualisation from the MVP allows the team to reallocate resources and time to other essential aspects of the project, therefore ensuring the remaining deliverables are completed within the agreed budget and timeline.
* **Client communication:** The decision was made collaboratively with the client to help align the project outcomes with their budget constraints while maintaining the core functionality required for the MVP.

Next Steps:

* **Reallocation of resources:** The team will now redirect their focus on tasks that align with the updated project scope.
* **Ongoing communication:** Our project manager will ensure continuous communication with the client to monitor any further changes in requirements or scope and manage them effectively.

Through adapting to this change, the team remains committed to delivering a high quality final product which will meet the clients needs while also meeting the budget limitations.

## **3.2 Changes to documents, code and data**

It is crucial to implement practices that maintain the integrity of the codebase, infrastructure and data when managing a large and complex project. As previously mentioned, Github will be the platform that we will employ to perform version control for our project. Specifically, by adopting a structured branching strategy to enforce consistency in code structure, versioning documentation, and securing critical data with immutable storage, we can create an efficient and secure development environment.

##### 

### **3.2.1 Branching strategy: Issue-based**

For every issue created in Github, a corresponding branch will be named with the issue number. This ensures that each branch is clearly linked to a specific issue. For example, we could have issue number 5 that is titled “Add string validation function” with further description to explain the problem. Then the person that is assigned the issue will create a branch called “ticket-5” and add the asked functionality. In the final commit message, that person must also write “#5 add string validation function” so that when a person checks out the code repo, they can click on the “#5” which is a hyperlink to the issue that was created to read more about the problem.

This approach will also be integrated with Github Actions to automatically run tests whenever code is pushed to a specific branch. On another note, Github Actions will be used in the CI/CD pipeline to ensure infrastructure changes with Terraform are tested first before being automatically deployed and make changes to the actual solution on Azure.

### **3.2.2 Code Reviews: Enforcing Quality with Pull Requests and Github Actions**

A Pull Request (PR) workflow enforces that every new issue or feature must be developed on a separate branch and go through a PR review before merging into the main branch. GitHub's branch protection rules will prevent direct pushes to the main branch, requiring a code review before merging changes from the development branch.. Also Github Actions will be used to perform checks on the code before it is ready for code reviews and merging.

To re-iterate on the CI/CD integration with Github Actions previously mentioned, we will automate testing and deployment pipelines with Github Actions, so that every PR triggers tests for Azure Functions and infrastructure changes defined in Terraform. This ensures that the code and infrastructure are peer-reviewed before being merged to the main (production) branch.

Any new feature or fix for Azure Functions will go through the entire process, including checks for code quality and errors using GitHub Actions, before being peer-reviewed and deployed to the main branch, ensuring seamless deployment to our Azure environment.

### **3.2.3 Documentation versioning**

In this project, we’ll treat our documentation as code. This means that all documentation will be stored in the /assets/docs folder. We will use Git for version control of our documentation.While reverting changes to documentation is unlikely, we will continually update it to reflect changes in our system.

Each document will often be updated alongside the corresponding code. If documentation updates are independent of any code changes, a separate issue will be created for them. For PRs, we will not spend extensive time reviewing documentation after it has been written, as two dedicated team members (Tash and Ninuri) will ensure all documentation is up-to-date and follows the established standards.

### **3.2.4 Immutable storage:**

Our solution involves using Azure Data Lake Storage Gen 2, which is essentially Azure Blob storage with support for Immutable Policies, often used for data containers holding critical data. This is still under consideration. Another competing idea is to enable versioning in each data storage container. This means if we change our files, a version tree will allow us to revert to previous versions of the data quickly.

However, since the amount of data we will be sourcing and creating for this project is relatively small, we are also considering creating a versioning system using folders. We are also considering creating a versioning system with our folder. This means that whenever new data is added or the structure of the data changes, we can create a new folder to store the updated data. This approach will simplify our solution, but it is not scalable or realistic in a real-world scenario, as the number of folders would quickly grow and become unmanageable.

# **5. Quality Management**

Quality Management is the continuous improvement of processes and services to satisfy expectations of the client at every stage of the project (Ebrahimi and Sadeghi, 2013). This section highlights the various aspects of quality planning where we discuss the objectives of the project, quality assurance, and quality control methods that will be undertaken to ensure the quality of the project.

## **5.1. Objectives**

### 5.1.1. Ensuring data quality

With the presence of uncleaned data in the database any results we receive from queries will be inaccurate, incomplete, incorrect, and not up-to-date as they do not “represent real-world entities to which they refer” (Fan and Geerts, 2022). Thus, improving the quality of our data is crucial for this project. Key issues with data quality include data consistency, data accuracy, data currency, data deduplication, and information completeness (Fan and Geerts, 2022). Data quality management is used to effectively detect and correct errors in the data resulting in increased value and accuracy to the project results.

### **5.1.2. Model Accuracy**

The main objective of the dataset is to train an AI model on insurance claims allowing it to identify fraudulent and non-fraudulent claims. However, the poor quality of training data is known to negatively impact the performance of machine learning models with issues such as underfitting and overfitting affecting the model’s accuracy. Thus, ensuring the quality of the dataset will save the AI team training effort on their AI model, avoid disappointment from machine learning results, as well as reduce setbacks in the project's progress (Mohammed et al., 2024).

### **5.1.3. Compliance with business requirements**

This project has five main business requirements; using Azure services, Agile Delivery, Anonymised Data, User Interface, and being a Gen AI solution. Various control and assurance mechanisms must be placed to ensure that the business requirements are being maintained and adhered to at every step of the development process.

### **5.1.4. Customer Focus and User Satisfaction**

The main goal of any project is to satisfy the user and create a working product. Thus, it is crucial to involve the client in every stage of the project to ensure that what is being created is what they need. As well as gaining input and feedback on key aspects of the project to confirm that the project meets the client’s requirements whether it be budget, technology used, or procedures undertaken.

## **5.2. Data Quality Management: Monitoring and Assurance**

### **5.2.1. Data Collection Process**

The source of data represents a logical perspective that encapsulates the methodologies for data collection, data generation, and assessing organisational compliance and integrity of the source (Mohammed et al., 2024). Thus, it focuses on the notion of data provenance such as data provider, data origin, and the other organisations involved in creating and transforming data (Mohammed et al., 2024). The main source of data collection is Kaggle, an online community where data scientists and machine learning engineers can obtain datasets to train AI models as well as publish their datasets for others to use. While the data source (Kaggle) is widely reputable suggesting the quality of the data it is important to conduct assessments on the dataset to ensure that the data source is credible and adheres to the dimensions of data quality.

Some assessments that have been done to the data source during this process are (Mohammed et al., 2024):

* **Origin:** where the data was sourced
* **Transformations:** how was the data changed
* **Traceability:** which focuses on the entities involved in its history
* **Reputation:** which is when the reliability and credibility of the data source are evaluated
* **Reliability:** which reviews the collection methods and if it conforms with established practices.

Furthermore, the data in the dataset has to be evaluated in relation to the data quality dimensions to ensure their validity (Picard et al., 2020):

* **Accuracy:** measuring how correct the data values are
* Accessibility: how easily the data can be accessed and user awareness of the data being collected as well as its location
* **Consistency**: the extent key data feature provide correlating information on the same data object
* **Timeliness**: the dataset has current data and information is available on time
* **Traceability**: the data’s history and origin is verifiable
* **Usability**: the scope of which the data can be understood and utilised
* **Relevance**: the data is able to meet the need of the user and even through information may change the collected data should still be relevant

### **4.2.2. Data Cleaning**

Datasets that are collected from different sources and put in our Bronzer Layer may contain types of dirty data such as missing fields/values, contradicting data, noisy values, data integration issues and cryptic data (Swapna et al., 2016). Additionally, anonymising the data is a key requirement of the project meaning that personally identifiable information should be removed from the dataset such as names, addresses, mobile phone numbers etc. Therefore, the data set must be cleaned to ensure better quality of the dataset for AI model training.

Three main steps in data cleaning include (Swapna et al., 2016):

* Data analysis - As a dataset may contain multiple error types it is important to do a comprehensive data analysis to identify the types of errors in the dataset.
* Transformation - At this stage the identified errors are transformed to clean the dataset. This may include dropping columns with personal information, removing duplicate rows, and filling missing values in the dataset.
* Verification - After the dataset is cleaned the new dataset should be evaluated and tested on the AI model with a sample of the dataset to test its accuracy when training the model.

### **5.2.3. Data Standardisation**

Standardising data is the transformation of data into a common format focusing on interoperability standards and modelling annotations. It is usually performed by referencing a gold standard model that becomes the standard for the parameters of following datasets where non-standard values are replaced with corresponding values that comply with the standard. Data standardisation is crucial to our project

* **File type:** All datasets found should be in .csv format so that datasets can be easily integrated.
* **Data Format:** All data in a specific column must be in the same format across all datasets such as for dates (YYYY-MM-DD) or time (HH:MM:SS). This allows for uniform and clearly defined data in columns for easy understanding and integration.
* **Data Type:** float, string, int, datetime, etc. should be the same for all data in the same column

## **5.3. Human Resource Management: Agile Delivery**

Utilising Agile methodology for project delivery leads to on-time delivery, customer satisfaction, business value, and product quality (Krehbiel and Miller, 2018). Thus, is an approach for improving and managing quality in delivering a successful project. Agile practices focus on facilitating face-to-face interactions leading to better articulation of goals, encouraging collaboration, improving team dynamic, and supporting innovations and experimentation (Krehbiel and Miller, 2018).

### **5.3.1. Sprint Planning**

Sprint planning is short iterations of project stages to bring value to the customer in smaller and more frequent intervals (Krehbiel and Miller, 2018). This alongside daily stand ups promotes enhanced communication, collaboration, reflection, and continuous improvement (Krehbiel and Miller, 2018). The ability to show work and progress to the client frequently allows for the client to be involved in the process. This improves the quality of the project deliverable as changes are being made real-time to reflect the requirements of the client. The quality of the project process is met as improved communication leads to smoother project completion and less push back from stakeholders.

### **5.3.2. Iteration Reviews**

Having iteration reviews at the end of each sprint allows the team to gather all the feedback from the client, identify any areas for improvement, and re-align themselves with the project goals. This results in improved performance in the following sprint as work reflects the feedback given and inefficiencies were noticed and altered to be more streamlined enhancing the quality of the project delivery.

### **5.3.3. Continuous Feedback Loop**

Engaging in Agile Delivery allows for a continuous feedback loop with the client in which work is shown to the client and the team receives feedback. This happens in short iterations allowing for the project to change and reflect the requirements of the client. This leads to reduced need for a final inspection that may not meet the clients requirements but a more robust product made from various testing and prototypes that can be effectively integrated to the clients system.

## **5.4. Development Management**

### **5.4.1. Code convention**

A code convention is a set of guidelines made for a programming language to enhance the software structural quality. Projects that implement these conventions have improved readability of their source code and enable easier software maintenance. Aspects of a language that is included in the convention include, indentation, declarations, naming conventions, comments, etc.

### **5.4.2. Prototype**

Creating a prototype of the dataset allows the team to test the viability of the datasets capability (Burgess, 2024). For this project our team will mainly engage in vertical prototyping where we create test data sets for a specific functionality or requirement (Burgess, 2024). This allows the team to test clean datasets that have been created and see whether they are enough to train the AI model. In cases of over- and under-fitting the dataset can then be modified and new prototypes can be created. Once a prototype of good training capacity is created then the dataset can be scaled up and the model and be trained more robustly.

## **5.5. Documentation Management**

### **5.5.1. Formatting**

Having standardised formatting across all documents increases the professionality and quality of the project documentation as it is uniform.

* Font: Times New Roman
* Size: Heading 1 - 20, Heading 2 - 16, Heading 3 - 14, Normal Text -12
* Spacing: 1.15

### **5.5.2. Documentation control officers**

Document control officers are put in place as a control mechanism to make sure that all documentation is written according to the defined documentation standards. These officers are further explored in the Team Manual where the further roles and responsibilities are discussed.

# **6. Schedule**

## **6.1. Project Overview**

For our project, we are working with our client EY to create a solution using GEN AI to examine automobile claim reports and detect whether they are fraud or not fraud. The main aim of this project is to simplify their claims process and enhance their current fraud detection capabilities. To do this we are split into AI and Data Teams, as the Data Team our main project goal is to create a dataset that is clean and accurate to be used by the AI to train and test their AI model. As we follow the agile methodology when undertaking this project we have decided to call our phases sprints as we carry out the project as outlined below:

* Sprint 1: Planning Stage
  + Forming the team, getting to know each others skills and proficiencies, and creating the communication channel
* Sprint 2: Building Stage 1
  + Brainstorming ideas for our MVP taking into consideration alternatives and feasibility
* Sprint 3: Building Stage 2
  + Finalising the project plan and creating a test dataset and synthesised data for training
* Sprint 4: Building Stage 3
  + Hand over prototype dataset to AI team to train their AI model
* Sprint 5: Testing Stage 1
  + Revise the dataset in relations to accuracy results from AI and reduce any under/over-fitting issues
* Sprint 6: Testing Stage 2
  + Hand revised dataset to the AI team for further testing
* Sprint 7: Execution Stage
  + Create front-end of the fraud detection project
* Sprint 8: Deployment Stage
  + Hand over the final deliverable to the client with a user manual and final report

### **6.1.1. Deliverables**

For our deliverables, they have been split into project-related deliverables for the items we present to our project sponsor Deborah, and product-related deliverables for the items presented to our client Joshua Falanga from EY.

**Project-related deliverables:**

* Deliverable 1: Feasibility Report and Team Manual
* Deliverable 2: Project Plan, Requirements/Scoping Document and Update Team Manual
* Deliverable 3: Updated Deliverable 2, Design Document, Test Cases, Prototype/MVP
* Deliverable 4: Updated Increment 1 (Deliverable 3), User/Training Manual
* Deliverable 5: Final Group Reflective Report
* Deliverable 6: Project Presentation/Demonstration
* Individual Contributions forms

**Product-related deliverables:**

So far our deliverables to the client have been:

* User Stories
* High-Level Solution Design
* Options Analysis Report
* Deliverable 3: MVP/Prototype
* Deliverable 7: Final product, the dataset

However, for our project, we have split it into 3 main stages:

* Bronze Layer: Raw datasets
* Silver Layer: Cleaned & enhanced datasets with synthetic data
* Gold Layer: Combined dataset for training AI model

## **6.2. Project Phases and Key Milestones**

See Appendix 4 for a breakdown of the project phases and the key milestones in each phase.

## **6.3. Schedule - Gantt Chart**

The figures below showcase the tasks required for the completion of the project using our project management tool Jira to create Gantt charts to visualise this timeline. The Gantt chart includes the summary of the task, estimated starting and ending dates of the tasks and Sprints, if the task has been completed or not and the dependencies tasks have with other tasks. As we progress through the project and each Sprint, the schedule will be subject to change such as dependencies, unforeseen factors, changes from clients and setbacks from team members. Thus, the schedule will be modified to reflect these changes throughout the project. See Appendix 5, 6, and 7 for the Gantt chart figures

## 

## **6.4. Resource and Task Allocation**

Resources for this project include, team members, skills, equipment, and tools and software. Proper resource allocation leads to a “streamlined workflow, mitigation of bottlenecks, simplified project management and resource optimisation” (Martins, 2024). This allows us to improve our project management ensuring that resources are used efficiently, projects are completed on time, finished to the desired quality standards and within budget resulting in a successful project (Martins, 2024). To properly allocate our resources and skills we have split ourselves into four main groups, Individual Member, Business Intelligence Team, Development Team, and Whole Team. These tasks and resources are allocated by the Project Manager to specific groups and assigned to one team member to maintain accountability and traceability of the project. However, as we progress with the project these teams and members will change to reflect the current tasks and project stage. Refer to Appendix 8 for the Task Allocation groups.

## 

## **6.5. Tools and Technologies**

### **6.5.1. Jira**

Our project management tool is Jira project management software that allows for the tracking, assigning, and managing of tasks. Our project is split into 2-week sprints and the team is assigned tasks by the Project Manager. We have combined our tasks with Group 13 using this tool so their tasks are also placed on our Jira creating a combined timeline and Gantt chart. This allows us and the client to see the progress and timeline of the project as a whole as well as add dependencies to tasks of the other team.

### **6.5.2. GitHub**

GutHub is a version control system that allows for versions of files to be tracked and for source code to be worked on collaboratively when developing. It also allows for teams to discuss and manage tasks within the GitHub allowing for more effective collaboration. When using GitHub each branch will be created by identifying an issue and numbering it. This allows for other members to work on issues and commit work to a certain issue number so each push is organised and sorted. Once an issue has been completed the branch will be closed and the issue number will be removed to reduce misunderstandings and proceed to other issues.

# **7. Assumptions**

## **7.1. Resources**

* **Skillset** - team member have the necessary skills and background to complete the project
* Training - training for new tools and technology introduced will be provided during the project
* **Budget** - the project budget is sufficient to to cover the tools needed for the project
* **Funding** - financial support will be provided for the project with no delays in the acquisition tools and services needed for the project

## **7.2. Availabilities**

* **Team availability -** all team members will be available throughout the entirety of the project with no major changes to the team
* **Stakeholder availability** - stakeholders will be available for weekly stand-up meetings and subject matter experts will be present in weekly meetings and feedback sessions.

## **7.3. Tools**

* **Accessibility -** all necessary tools will be available for editing to all team members with no restrictions.
* **Compatibility -** the Azure services used will be compatible with each other and with the services used by the AI team for their AI model.
* **Azure services -** the costs to acquire Data Lake, Azure Functions and Application Insights will not exceed $0.45 per month for the duration of the project.
* **Github -** the team members will use GitHub to commit and push tasks. They will follow the
* **Jira -** will be used by each team member to update task progress. The project manager will assign tasks to each team member and actively update the Jira to reflect the current and future status of the project.
* **Google Drive -** will be used as a central hub for storing documents and presentations so each team member has access and is able to edit it.

## **7.4. Techniques**

* **Data synthesisation techniques** - will create 7,000 rows of clean and good quality synthetic data
* **Data cleaning method -** will clean the raw dataset to remove or replace missing values, duplicate values, and unrequired data columns.
* **Testing techniques** - testing the dataset with the AI model will provide valuable insights on the quality and performance of the dataset on the model and allow for refinement of the dataset to increase accuracy.

## **7.5. Standards**

* **Data Quality Standard** - datasets are chosen in relation to the data quality standards defined and of sufficient quality to develop a reliable dataset.
* **Documentation Standards** - all documents created for the project are created following the defined documentation standards to ensure the professionality and quality of documents created.
* **Code Conventions** - when writing code for the project the team members will adhere to the established coding standards to ensure the consistency and maintainability of the code.

## **7.6. Communication**

* **Communication channels -** team will maintain regular communication with each other to discuss about the project on the Discord server and always react or engage in discussions
* **Regular meetings -** the team will hold regular weekly meetings to discuss task status, future tasks and roadblocks.
* **Issue Communication -** when any issues or problems arise that may impact the progress of the project, team members will promptly discuss the issue with the team to reduce further delay or misunderstandings.
* **Client engagement -** the client will answer questions asked on the discord as well as provide feedback on Deliverable tasks.
* **Sponsor communication -** the project sponsor (Deborah) will reply to any issues regarding the project in a timely manner

## **7.7. Expectations**

* **Project scope** - will be clearly defined and business requirements will be established and communicated to the team.
* **Quality expectations** - project will be completed to a high quality standard from coding to documentation adhering to predefined standards.

# **8. References**

Armstrong, M. (2009). *Armstrong’s Handbook of Human Resource Management Practice*. <http://196.43.179.3:8080/xmlui/handle/123456789/1068> [15/08/2024]

Ebrahimi, M., & Sadeghi, M. (2013). Quality Management and performance: An Annotated Review. *International Journal of Production Research*, *51*(18), 5625–5643. <https://doi.org/10.1080/00207543.2013.793426> [12/08/2024]

Fan, W., & Geerts, F. (2022). *Foundations of Data Quality Management*. Springer Nature. [13/08/2024]

Martins, J. (2024). *Resource Allocation Tips for Project Managers • Asana*. Asana. <https://asana.com/resources/resource-allocation> [22/08/2024]

Li, S. Q., & Chen, Y. S. (2020). Analysis of the probability matrix model for the seismic damage vulnerability of empirical structures. *Natural Hazards*, *104*(1), 705–730. <https://doi.org/10.1007/s11069-020-04187-2> [15/08/2024]

Mohammed, S., Harmouch, H., Naumann, F., & Srivastava, D. (2024). Data Quality Assessment: Challenges and Opportunities. *Data Quality Assessment: Challenges and Opportunities [Vision]. PVLDB*, *17*(1). <https://arxiv.org/pdf/2403.00526> [10/08/2024]

Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefevre, B., & Soumarmon, T. (2020). *Ensuring Dataset Quality for Machine Learning Certification*. IEEE Xplore. <https://doi.org/10.1109/ISSREW51248.2020.00085> [21/08/2024]

Smith, P. G., & Merritt, G. M. (2002). *Proactive Risk Management: Controlling Uncertainty in Product Development*. <https://www.strategy2market.com/Preston-Smith/Proactive-Risk-Management/PMJ12-03.pdf> [14/08/2024]

Swapna, S., Niranjan, P., Srinivas, B., & Swapna, R. (2016). *Data cleaning for data quality*. IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/7724284> [16/08/2024]

Timothy, C., & Miller, D. (2018). Article 1 1-1-2018 Part of the Technology and Innovation Commons Recommended Citation Recommended Citation Krehbiel. *Journal of International Technology and Information Management Journal of International Technology and Information Management*, *27*(1). <https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1342&context=jitim> [20/08/2024]

# **9. Appendices**

## **Appendix 1. Risk analysis table**

| **Risk Category** | **Risk** | **Description** | **Probability** | **Effect/consequence** | **Impact** | **Mitigation Strategies** |
| --- | --- | --- | --- | --- | --- | --- |
| Technology Risks | System Downtime | Azure Data Functions or GIthub experiencing outages | Medium | Delays in data processing and version control | Medium | Regular backups, use of alternative tools temporarily |
|  | Data Security Breach | Unauthorised access to sensitive data during processing or storage | Low | Potential legal issues, Loss of client trust | High | Implement encryption, regularly update security protocols |
|  | Inadequate Software integration | Difficulty integrating Jira, Github, Azure Data Functions. | Medium | Reduced efficiency, increased project delays. | Medium | Conduct regular integration tests, use middleware to ease integration |
|  | Scalability Issues | Azure functions or data storage limits, slowing down processes as data volumes increase. | Low | Inefficiency in processing large datasets | Medium | Optimise code, plan for scalability. |
| People Skills | Skill Gaps | Team members lacking expertise in key areas (e.g AI | Medium | Project delays and suboptimal outcomes. | High | Provide training, hire consultants if needed. |
|  | Team collaboration | Poor communication or conflicts among team members. | Medium | Misunderstandings, mistakes, and delays. | High | Foster open communication, regular team meetings. |
|  | Resource availability | Unavailability of key team members during critical phases | Medium | Stalled progress delayed tasks | Medium | Cross train team members, have backup plans. |
|  | Role clarity | Overlapping or unclear responsibilities leading to confusion | Medium | Duplication of efforts or missed tasks | Medium | Clearly define roles and responsibilities |
| Organisation risks | Budget constraints | Project exceeded the allocated budget, causing resource shortages. | Medium | Compromise in project scope or quality | High | Regular budget reviews, adjust scope as needed. |
|  | Policy Changes | Changes in organisational policies or priorities disrupting the project. | Low | Delays,need for significant project adjustments. | Medium | Stay informed about policy updates, plan for flexibility. |
| Requirements risks | Ambiguous Requirements | Incomplete or unclear project requirements leading to confusion. | Medium | Scope creep,rework, project delays. | High | Detailed requirements gathering, regular reviews. |
|  | Changing requirements | Frequent changes in project scope or objectives disrupting the project flow. | Medium | Increased costs , extended timeline. | High | Establish a change management plan. |
| Estimation Risks | Underestimation complexity | Understanding the complexity of tasks. | Medium | Missed deadlines, budget overruns | High | Break down tasks, seek mentor opinions on estimates |
|  | Timeline overruns | Project tasks longer than estimated, affecting the overall project timeline | Medium | Delays in project completion, Potential scope reduction | High | Regular progress tracking, adjust timeline as needed |
|  | Resource Overcommitment | Over Committing resources to asks leasing | Low | Decreased productivity, potential turnover | Medium | Monitor Workload, balance task distribution. |

## **Appendix 2. Risk Indicators Table**

| Risk Type | Indicators |
| --- | --- |
| Technology Risks | * Slow or unpredictable performance of project related software. * Delays in obtaining necessary software licences * Frequent complaints about software performance from team members * Increased frequency of system crashes or data loss |
| People Risks | * Decline in team morale * Team members missing or not participating in meetings * Deterioration in team communication and relationships * Overload or burnout of key team members due to poor task distribution |
| Organisational Risks | * Weak leadership or lack of clear direction from the project manager * Insufficient support from mentors * Delays in decision making or approvals from stakeholders |
| Requirements risks | * Negative or in insufficient feedback from the client * High frequency of requirement changes * Misalignment between client expectations and project deliverables * Unclear requirements leading to rework |
| Estimation Risks | * Missed development milestones * Project lagging behind schedule as per the Gantt chart * Repeated failure to conduct regular team meetings. * Underestimation of time or resources needed for key tasks |

## **Appendix 3 Probability/ Impact Matrix**

|  | **Impact** | | | |
| --- | --- | --- | --- | --- |
| **Probability** |  | **Low** | **Medium** | **High** |
| **High** | T2: Data Security Breach | P1: Skills Gap | R1:Ambiguous Requirements  R2:Changing Requirements |
| **Medium** | T4: Scalability Issues  E1: Resource Overcommitment | P2:Team collaboration  P3: Resource Availability | O1: Budget constraints  E2: Underestimation of Complexity  E3:Timeline Overruns |
| **Low** | T1: System downtime | T3:Inadequate Software integration  E4: Missed Milestones | O2: Policy Changes |

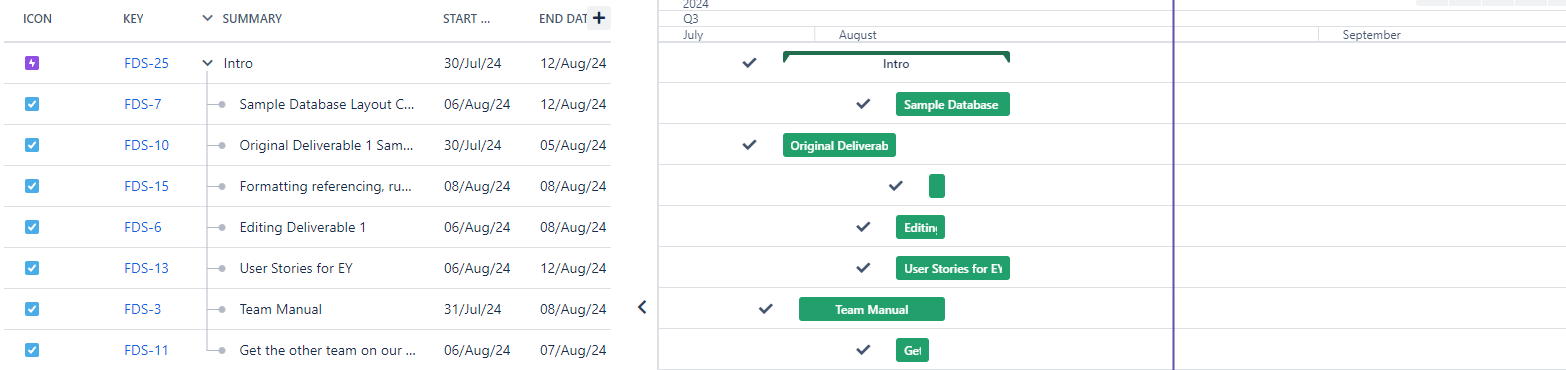
## **Appendix 4: Project Phases and Key Milestones**

| **Phase** | **Activities** | **Milestones** | **Deliverables** |
| --- | --- | --- | --- |
| Sprint 1  30/07 - 12/08  2024 | * Create discord * Sample database layout creation * Create user stories * Submit Deliverable 1 draft to the client * Complete Deliverable 1 | * Combined discord with mentors and Team 13 * Finalising required columns for the dataset | Project:   * Feasibility Report * Team Manual   Product:   * User Stories |
| Sprint 2  13/08 -  27/08  2024 | * Start sourcing data * Create High-Level Solution Design (HLSD) * Create Options Analysis * Set up GitHub * Set up Azure Infrastructure * Start loading data to the bronze layer * Prepare test data file * Submit Deliverable 2 draft to client * Complete Deliverable 2 | * Obtaining 10,000 rows of data * Github set-up * Feedback from mentors for Deliverable 2 | Project:   * Project Plan * Requirements Document * Updated Team Manual   Product:   * Options Analysis Report * HLSD presentation * Test data file |
| Sprint 3  28/08 - 10/09  2024 | * Obtain more datasets * Create the Code Convention * Research on data synthesization methods * Produce Synthetic Data * Build Azure Infrastructure setup * Build CI-CD pipeline * Clean the datasets | * Cleaned dataset created * Synthesise 7,000 rows of data | Project:   * Individual Contribution form   Product:   * MVP options document * GitHub documentation |
| Sprint 4  11/09 -  24-09  2024 | To be assigned | * Creation of MVP | Project:   * N/A   Product:   * MVP presentation |
| Sprint 5  25/09/- 8/10  2024 | To be assigned | * Handing a test dataset to AI Team | Project:   * Updated Deliverable 2 * Design Document * Test Cases Document * Prototype/MVP   Product: |
| Sprint 6  9/10 - 22/10 | To be assigned | * Revise Dataset with AI team feedback | Project:   * Updated Increment 1 (Deliverable 3) * User/Training Manual   Product: |
| Sprint 7  23/10 -  5/11  2024 | To be assigned | * Create Front end solution | Project:   * Final Group Reflective Report * Project Presentation * Individual Contribution form   Product: |
| Sprint 8  6/11 -  14/11  2024 | To be assigned | * Handover final project to client. | Project:   * Final Delivery of Project to Client   Product: |

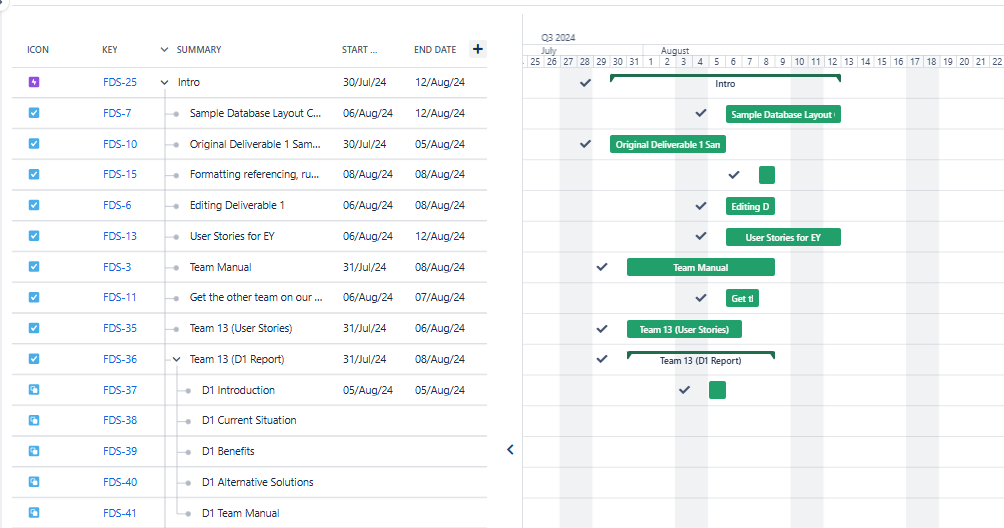
## 

## **Appendix 5: Sprint 1 Gantt Chart**

Gantt Chart for Team 14



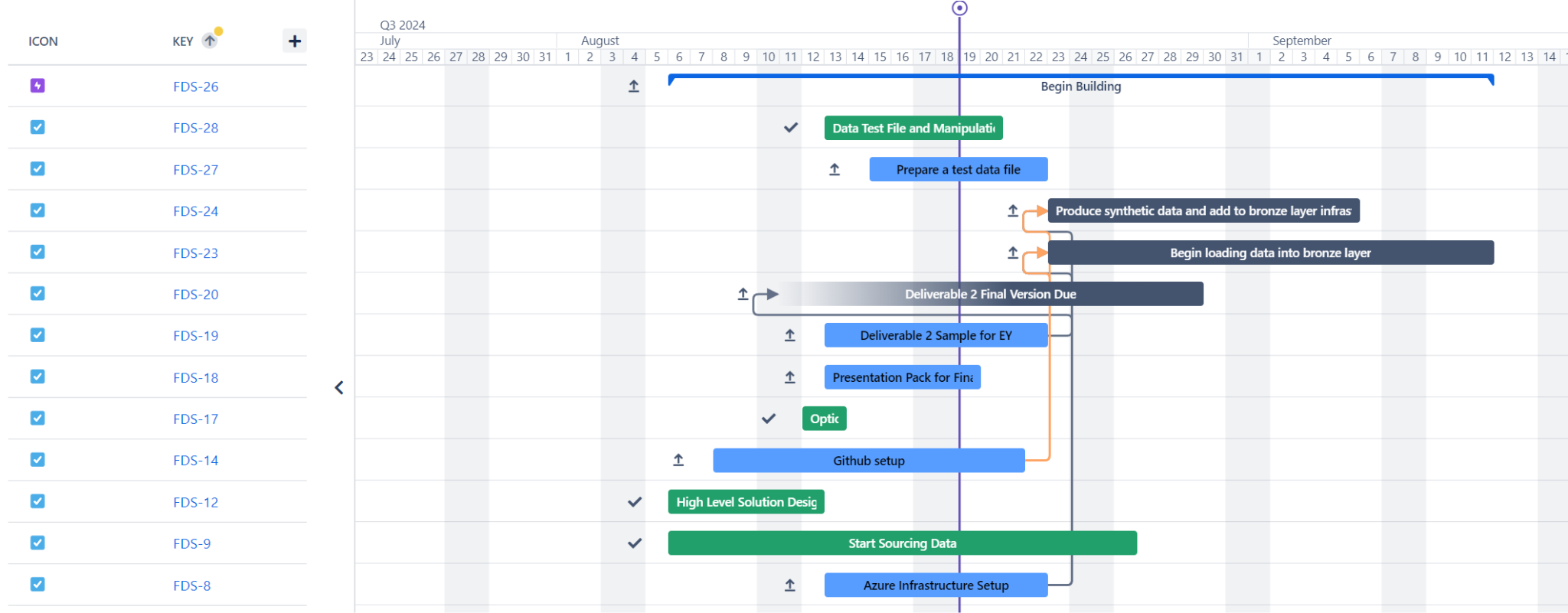
Gantt Chart combined with Team 13



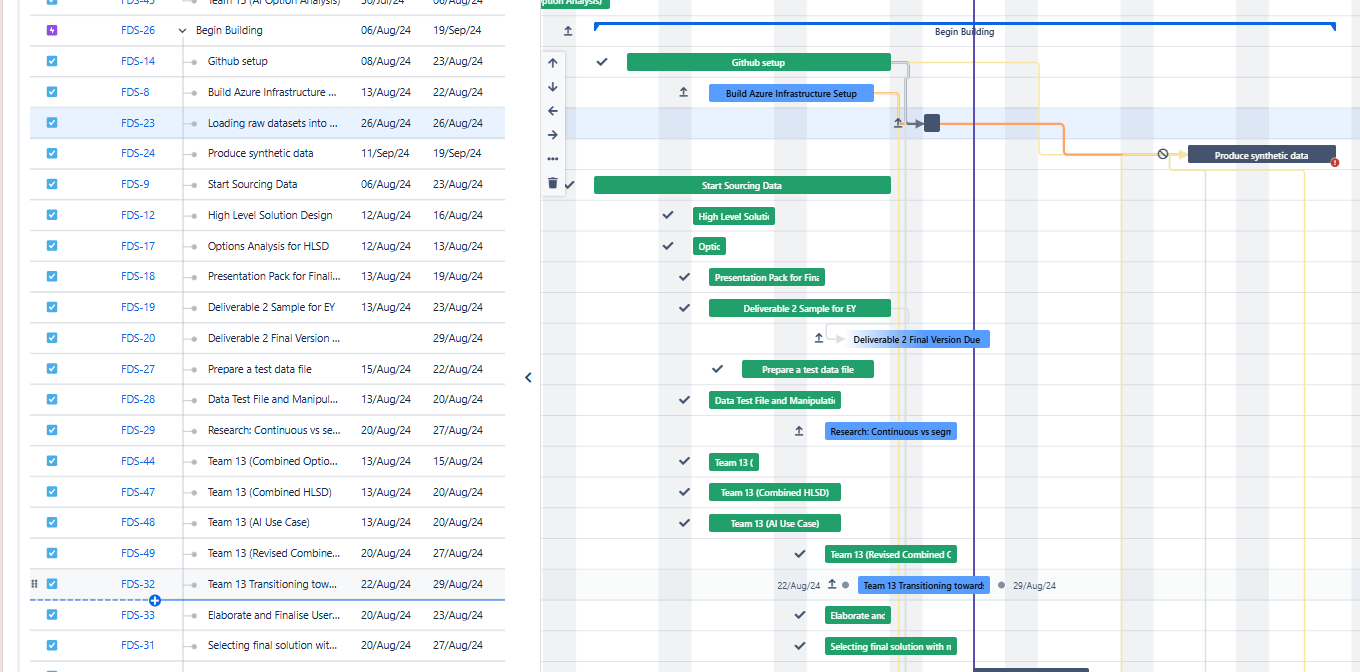
## 

## **Appendix 6: Sprint 2 Gantt Chart**

Gantt Chart for Team 14



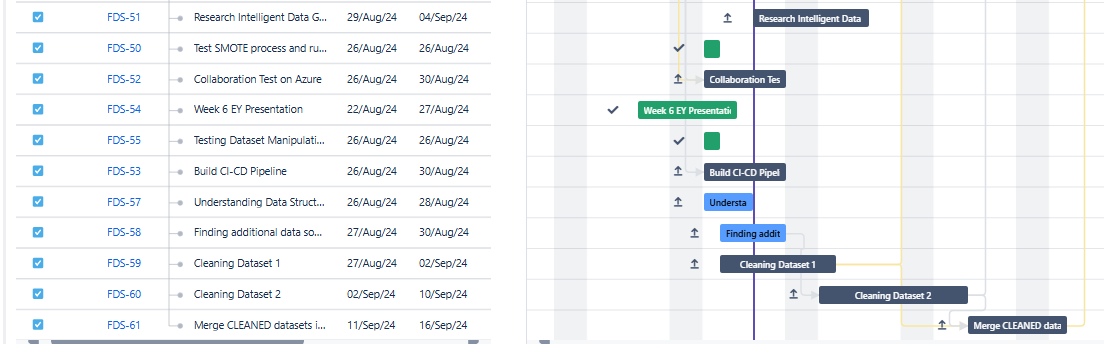
Gantt Chart combined with Team 13



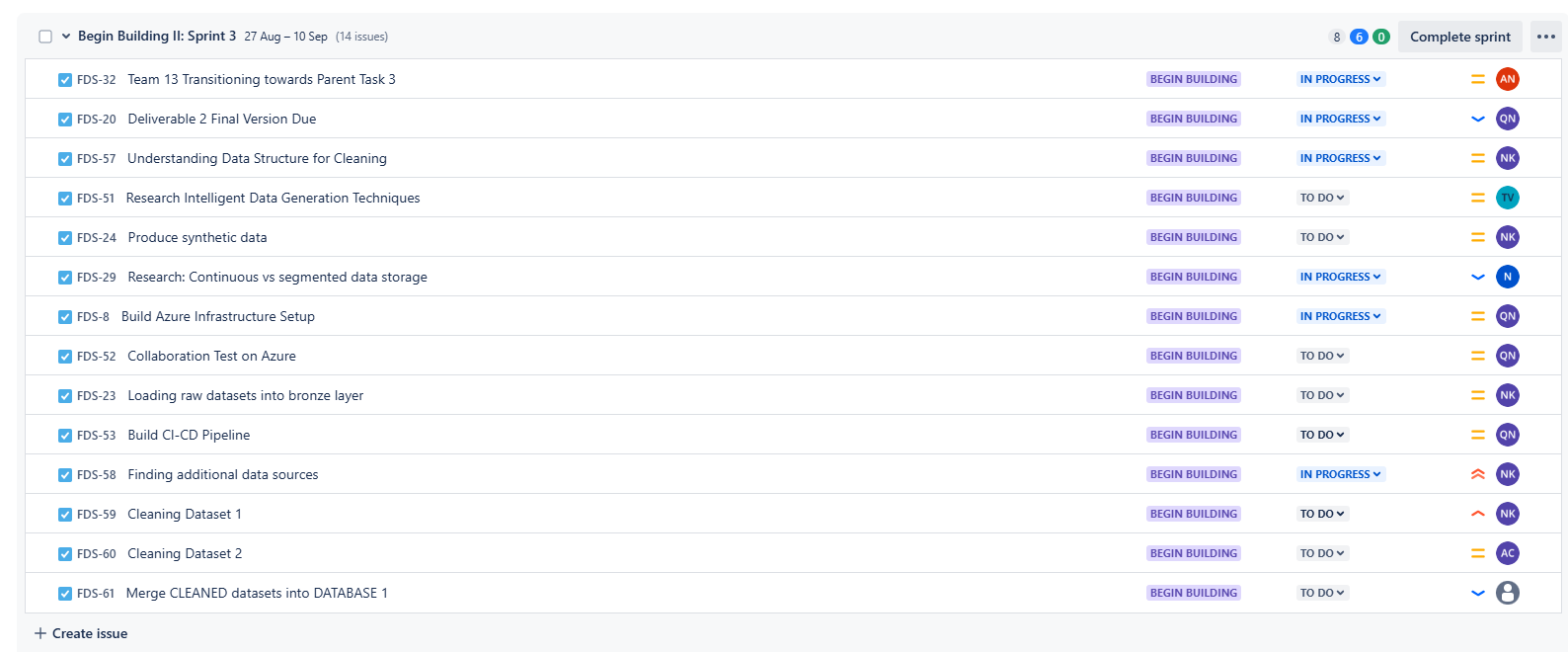
## 

## **Appendix 7: Sprint 3 Gantt Chart**

Gantt Chart combined with Team 13



Sprint 3 Backlog



## 

## **Appendix 8: Task Allocation Groups**

| Team | Members | Tasks |
| --- | --- | --- |
| Individual Member | Ninuri, Tashiya, Alan, Noorullah, Aasnayem(Adam) | Relating to the skill set of each member which is further explored in our Team Manual. |
| Business Intelligence Team | Ninuri and Tashiya | Research on aspects of the business to integrate with the product. |
| Dev Team | Aasnayem(Adam), Alan, and Norullah | Mapping dataset structure, pseudocode for cleaning, designing development pipeline |
| Whole Team | Ninuri, Tashiya, Alan, Noorullah, and Aasnayem(Adam) | Project-related tasks such as the Deliverables handing to the University. |

**Requirements Document/ Scoping Document**

# **1. Data Understanding**

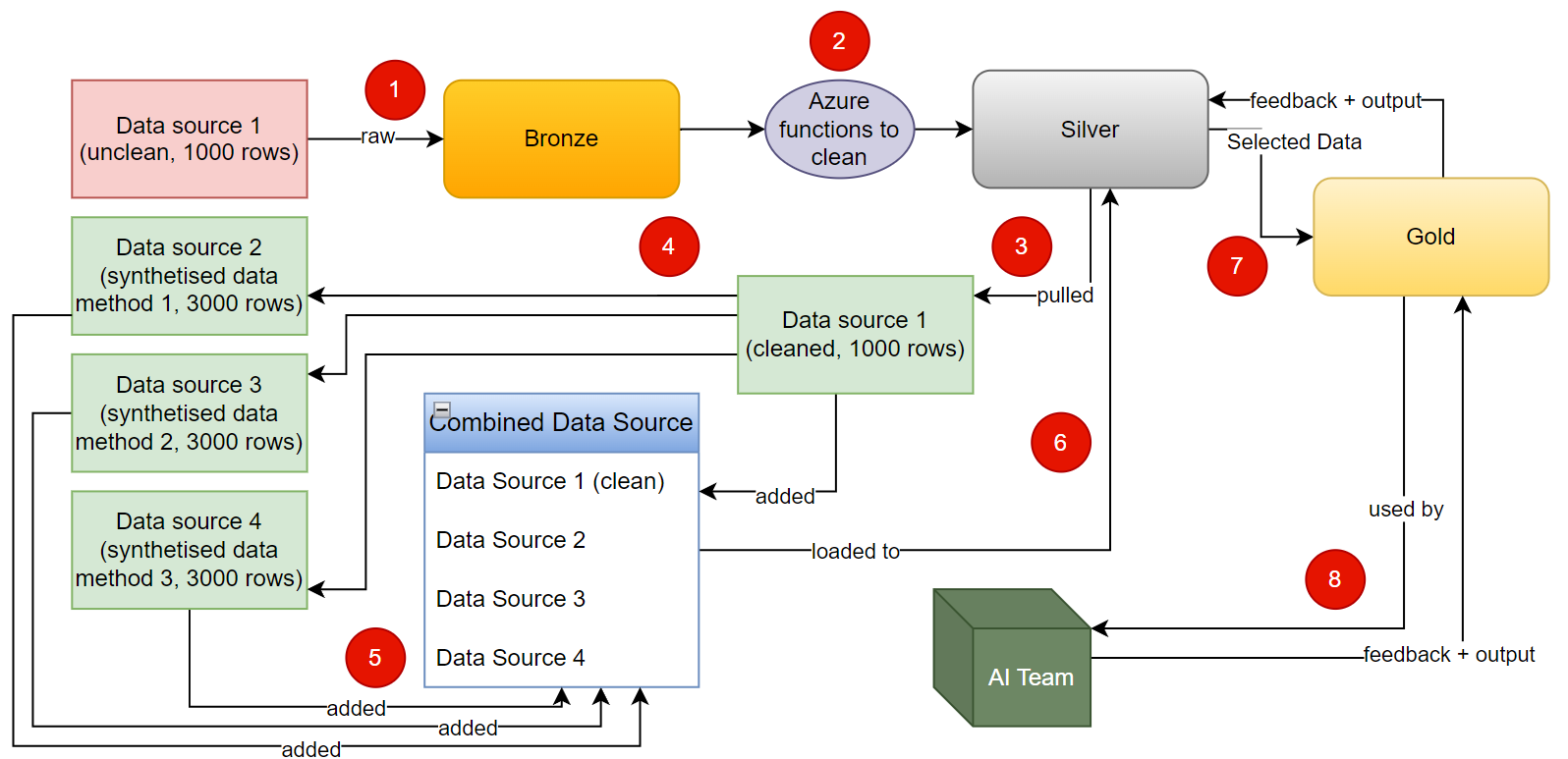
## **1.1 Initial Data Sources**

* Primary Dataset:
  + **Data Source:** Kaggle Dataset with 1,000 rows.
  + Content: Dataset includes a variety of columns relevant to insurance claims including months\_as\_customer, age, policy\_number, policy\_state, incident\_type, total\_claim\_amount, and a crucial column, fraud\_reported, which indicates whether each entry is fraudulent.
  + **Widespread Availability:** It has been discovered that this dataset is identical across multiple sources, websites, and forums, suggesting that it is widely used and well-vetted (explained more in 1.3).
  + Despite the dataset being a good starting point, it needs to be evaluated and prepared properly through being investigated, understood, mapped out and cleaned. Data cleaning would entail dropping those columns not needed, converting some metadata, and creating/augmenting new/existing columns. These processes might change over the course of the project. The data preparation will be to ensure it meets and services the project’s needs and can assist in fraud detection purposes.
* Additional Data:
  + **Synthetic Data Generation:** To expand the dataset, approximately 9,000 additional rows will be generated using three different data synthesis methods (Scikit-learn, GAN, and AI-based generation). This will then be combined to give a dataset with a total of about 10,000 rows.

## 

## **1.2 Data Collection Methods**

* Kaggle Dataset:
  + The initial dataset is already collected and readily available for analysis (after preparation). The dataset was sourced (Mwitiderrick,2018)
  + Columns and Structure: The dataset includes the following columns:
    - Months\_as\_customer, age policy\_number, Policy\_bind\_date, Policy\_state, Policy\_csl, Policy\_deductable, Policy\_annual\_premium, Umbrella\_limit, Insured\_zip, Insured\_sex, Insured\_education\_level, Insured\_occupation, Insured\_hobbies, Insured\_relationship, Capital-gains, Capital-loss, Incident\_date, Incident\_type, Collision\_type, Incident\_severity, Authorities\_contacted, Incident\_state, Incident\_city, Incident\_location, Incident\_hour\_of\_the\_day, Number\_of\_vehicles\_involved, Property\_damage, Bodily\_injuries, Witnesses, Police\_report\_available, Total\_claim\_amount, Injury\_claim, Property\_claim, Vehicle\_claim, Auto\_make, Auto\_model, Auto\_year, Fraud\_reported
* Bronze Layer:
  + **Raw Data Storage:** The unclean original dataset (1,000 rows) is stored in the Bronze layer in its raw form. The bronze layer is the initial repository for all data before any cleaning or processing.
  + **Purpose:** Keeping the unclean data in Bronze ensures that the original dataset is always accessible for reference or reprocessing if needed. It also makes the project traceable and reproducible.
* Silver Layer:
  + **Data Cleaning:** The raw dataset from the Bronze layer is cleaned using Azure Functions. This cleaned dataset (1,000 rows) is then stored in the Silver layer as a cleaned version of the original data.
  + **Synthetic Data Combination:** The cleaned dataset from Silver is used as the basis for generating synthetic data (~ 9,000 rows). The combined dataset, now totaling 10,000 rows, is also stored in the Silver layer.
  + **Data Organization:** The Silver layer will contain different folders to organise the original cleaned data and the combined dataset.
* Gold Layer:
  + **Final Data Preparation:** The data in the Silver layer is further refined and stored in the Gold layer. This final data is fully prepared and ready for use by the AI team. It will contain columns relevant and useful to the AI model and will have a feedback loop from the AI Team which will trickle down to the silver layer if our data quality needs to be improved by more cleaning and refining.



***Figure 1: Data Layers Layout Diagram***

## **1.3 Assumed Data Quality, Structure, and Completeness**

* Kaggle Dataset:
  + **Quality**: The dataset is widely available, suggesting that it is well-vetted. The exact same dataset is found from various sources and different websites. A lot of fraud detection and insurance data websites also link back to this exact same kaggle database. It also is high quality as it contains a large number of columns (39 columns) and these columns have diverse categories and types of data. There are also 1000 records meaning it’s a large dataset with variation. This also helps establish its data quality.
  + **Structure**: The dataset’s standardised structure simplifies integration but requires careful examination to ensure that no nuances are lost, and all necessary elements for fraud detection are present.
  + **Completeness**: While the dataset is likely complete in terms of the data points it intends to capture, it may not cover all scenarios or edge cases specific to the project, making the generation of synthetic data crucial.
* Synthetic Data:
  + **Quality**: The synthetic data is expected to be clean, as it is generated from the cleaned source data. It will be evaluated for quality, ensuring it introduces valuable variance and addresses potential gaps in the original dataset.
  + **Structure**: The synthetic data will match the cleaned original dataset’s structure, ensuring seamless integration.
  + **Purposeful Errors:** Some synthetic data might include purposeful errors or variations to test and validate the robustness of the AI models, helping to identify and rule out potential biases.
  + **Imbalances and biases:** As fraudulent claims themselves are a minority the data cleaning or machine learning models might skip over learning fraudulent records. To combat this use of SMOTE in synthetic data methods can increase the number of fraud records making the machine learning models learn them better.

## **1.4 Mechanisms for Evaluation**

* Data Quality Assessment:
  + **Initial Dataset:** Perform a detailed exploratory data analysis on the initial dataset to identify biases, potential overfitting risks, and any embedded assumptions that might affect the project’s outcomes.
  + **Synthetic Data:** Evaluate the synthetic data against the original dataset to ensure it adds diversity and robustness. The synthetic data should not merely replicate common patterns but should introduce valuable variance. Any purposeful errors will also be evaluated to assess their impact on model training.
* Data Structure Verification:
  + Confirm that the dataset’s structure fully supports the analyses planned. Look for any structural limitations in the dataset and address these with synthetic data as needed.
* Completeness Check:
  + Ensure that the dataset captures all relevant scenarios, especially for fraud detection. The incorporation of SMOTE elements in the synthetic data generation will help ensure a balanced representation of fraud cases.
* Additional Data Collection Requirements:
  + As the project progresses, regularly assess whether more unique or tailored data is needed. Stay flexible to incorporate supplementary data that could provide additional insights or cover overlooked scenarios.

# **2. Data Preparation**

## **2.1. Data Preparation Activities**

* **Data Selection**: In order to select the data, we looked into various data sources and identified the relevant features to include in our dataset. In our case, we went ahead with one dataset and created 100 rows of data using information from that dataset. As a consultant, it is essential to follow the requirements of the client. Therefore, this dataset was chosen according to the problem statement relating to detecting fraudulent insurance claims and according to the requirements of the AI team and the client. We then increment from 100 rows to create 1000 rows of data using this dataset. We are using only one dataset currently as it is much more efficient to just use one base dataset as it will ensure all data is consistent. However, this can change in the future and we are open to include other datasets.
* **Data Cleaning:** For data cleaning, we plan to fill in any missing data. We also plan to drop the columns which we will not need and possibly convert the categorical values into numerical ones for example: True will be converted to 1 and False will be converted to 0. We plan to clean the data and store it in the silver layer of our data structure.
* **Data Synthesization:** We plan to synthesise around 9000 rows of data using data synthesization techniques. We plan to use 3 different methods which will generate 3000 rows of data each. The methods are the “make\_classification” function in the scikit-learn library in python, GAN model, and AI based generation.
* **Data Integration:** While we are not currently planning to integrate different data sources, we can still do that if we find another valid data source. If we had to merge, we would identify the common keys such as the ClaimID and merge the datasets. We would have to be consistent with the data. If we are using datasets with different formats, we would have to convert the separate formats into a unified structure. Therefore, JSON files will be converted to tables, XML files will be converted to CSV files.
* **Data Formatting:** Data formatting ensures one particular format for our dataset. For our case we will have a CSV file. We select a CSV file as our dataset is relatively small and CSV is useful for small datasets. We have to ensure that all of the rows in the columns are consistent. For example, the date and time format should be the same for all the rows.

## **2.2. Data Processing Pipelines**

The data processing pipeline is mentioned above. We will have three different layers which are bronze, silver and gold. The bronze layer will have 1000 rows of data which will not be cleaned. While we are going from the bronze to the silver layer, the data will be cleaned using azure functions. While cleaning the data, we want to drop any columns we will not need and fill in any missing values. In the silver layer, we will also synthesise data using three different methods. Each method will synthesise 3000 rows of data. The methods will be Scikit-Learn, GAN(Generative Adversarial Networks), and AI models. This will be stored back in the silver layer and then it will be stored in the gold layer where there will be a feedback loop. The feedback loop will also move to the silver layer where the changes will be made. The AI team(Team 13) will review and give suggestions on the dataset and the Data team (Team 14) will adjust the dataset accordingly.

## **2.3. Different Preparation Pipelines**

Different models will require different preparation pipelines. The pipeline will ensure that the data is not mishandled and is optimised. This will lead to an improved performance of the model. If a Decision Tree model was chosen, the data would not need to be cleaned that much as that model can handle raw data well. Outliers would need to be removed if they impact the model negatively. However, if a Neural Network model was chosen, the data would need to be normalised and high quality data would need to be provided. Neural Networks are more sensitive therefore, transformations need to be performed to remove outliers and categorical values would need to be converted to numerical values.

## **2.4. Justification for Different Models**

Having different pipelines/models will enhance efficiency. Unnecessary steps can be avoided if there are separate pipelines. For example, in a single pipeline, cleaned data might still go through a data synthesis step that's really only needed for raw data, which wastes time and resources. With separate pipelines, the cleaned data can skip this step altogether, making the process more efficient. Therefore, less resources and time will be required. Moreover, separate pipelines allow for easy adjustments if new methods are introduced later on. A tailored pipeline will be beneficial as it will optimise the performance of the model. This will improve the accuracy of the model and reduce the risk of biases.

## **2.5. Data Storage Mechanisms**

Our team has decided on using Azure Data Lake Storage Gen2 (ADSL2) to store all of our data. The data will be stored in layers as mentioned above. Azure Data Lake Storage Gen2 is being used as it is cost effective, secure, and scalable. Our dataset will not be that large therefore, the cost for ADSL2 will be cheap. Even if our dataset increases, that will not be a problem as ADSL2 can handle large volumes of data. Therefore, this setup ensures high performance and provides a good solution for data storage.

### 

# **3. Modelling**

For the project, we are working in collaboration with Group 13 for EY. In this collaboration, we are split into two teams, Team 14 who focuses on the data side with our main deliverable being a comprehensive dataset and cloud infrastructure, and Team 13 which focuses on the AI side and will be using our dataset to train their Generative AI model. Therefore, modelling does not fall within the scope of our project with EY however, below we will be discussing some aspects of modelling that will be taking place in relation to the AI team's requirements. As this section is not part of our final deliverable to our client we will only be touching briefly on some of these modelling techniques and the AI team will be expanding further on this specific topic.

## **3.1. Modelling Techniques**

The main outcome of using these AI modelling techniques is to find out if an insurance claim being made to NRMA is fraud or not fraud. This is a classification problem as it aims to categorise the data into predefined classes of fraud and not a fraud. Therefore, as a classification issue, we will be using the following models to predict classes: logistic regression model, neural networks, and decision tree.

### **3.1.1. Neural Networks:**

* **Reason it was Chosen:**Neural Networks are chosen for its capability of modelling high nonlinear interactions between variables with several layers of interconnected nodes. This is quite valuable, especially in detecting nonlinear patterns that usually characterise fraud.
* **Intended Outcome:** Neural Networks will be to explore the AI model’s ability to capture subtle, non-linear patterns in the data that may not be found when using other traditional methods such as Random Forest.

Benefits of Neural Networks

* **Capture Complex Features:** Form the base of popular models and can be formed into elaborate architectures that productively capture complex features considering we have large amounts of data and computation.

limitations of Neural Network

* **Significant resources needed:** more data, customisation and training time is needed for larger, more complex models.
* **Need for Preprocessing:** The data is needs careful preprocessing
* **Feature Types:** the method works better when features of similar type are chosen rather than features of different types.

### **3.1.2. Logistic Regression Model:**

* **Why Chosen**: Logistic regression is a very effective yet quite simple technique when used in binary classification tasks. It quantifies the likelihood of an instance to be from a specific class, which proves suitable for determining if a claim instance is fraudulent or not (Class 1) or not fraudulent (class 0).
* **Intended Outcome**: The logistic regression model will be used to create probabilities in insurance claims that gravitate to the “fraud” class. If the probability obtained is above a certain benchmark then the claim will be classified as a “fraud”.

Benefits of Logistic Regression:

* **Simplicity:** Can be easily implemented and the results can be easily understood.
* **Efficiency:** compliments smaller datasets and is able to provide quick results.
* **Probabilistic Output:** Offers probabilities for each class of “fraud” and “non-fraud”, assisting NRMA in detecting fraudulent claims.

Limitations of Logistic Regression:

* **Linearity**: Logistic regression accepts a linear relationship between the features and the log odds, which may possibly not be able to capture complex patterns.
* **Limited Complexity:** It is less functional when used with complex or non-linear data in comparison to other advanced models such as the discussed neural networks.

### **3.1.3. Decision Tree:**

* **Why Chosen**: Decision trees provide a comprehensible method for classification, where decisions are made by exploring a tree structure made from feature values. Hence, it is easy for the user to understand and explain the prediction made by the model.
* **Intended Outcome**: Clear decision rules for identifying claims as fraud or not fraud will be provided by decision trees through the decision rules. This will be useful for explaining the reasoning behind each classification.

Benefits of Decision Tree

* **Usability:** It is easy to understand due to its tree structure
* **Usefulness**: Decision trees will be handy in data exploration
* **Efficiency:** Less data cleaning required. Moreover, outliers and missing values will not be influenced by this method.
* **No Constraints:** This method can handle numerical and categorical variables therefore data type will not be a constraint.

Limitations of Decision Tree

* **Over-fitting:** there is a chance of the model overfitting meaning that the model isn’t trained properly
* **Not fit for continuous variables:** When categorising continuous numerical variables in different categories, the decision tree loses information.

### **3.1.4. Relationship to the Chosen Model:**

The mentioned modelling techniques were chosen due to their abilities to be able to handle large and complex datasets, along with its established history in fraud detection cases. The desired result is to ensure accuracy, precision and recall of fraudulent transaction detection while reducing false positives. Each chosen model is projected to contribute unique strengths:

* **Neural Networks:** For detecting indicators of fraud by finding complex and nonlinear patterns in the data.
* **Logistic regression Model:** Offers a straightforward and understandable model that efficiently manages binary classification tasks, delivering clear probability scores for detecting fraud.
* **Decision tree:** Provides a clear, rule based classification approach which ensures stakeholders to be able to quickly comprehend and trust the decision making process in fraud detection.

## **3.2. How will the Effectiveness and Validity be checked?**

Some issues occur when training models:

* **Overfitting/Underfitting** - accuracy of testing and training data
  + Good in training but poor on tests then overfitting
  + Poor on both training and test then underfitting
* **Curse of Dimensionality** - when you collect features, especially irrelevant features you cannot get any useful information

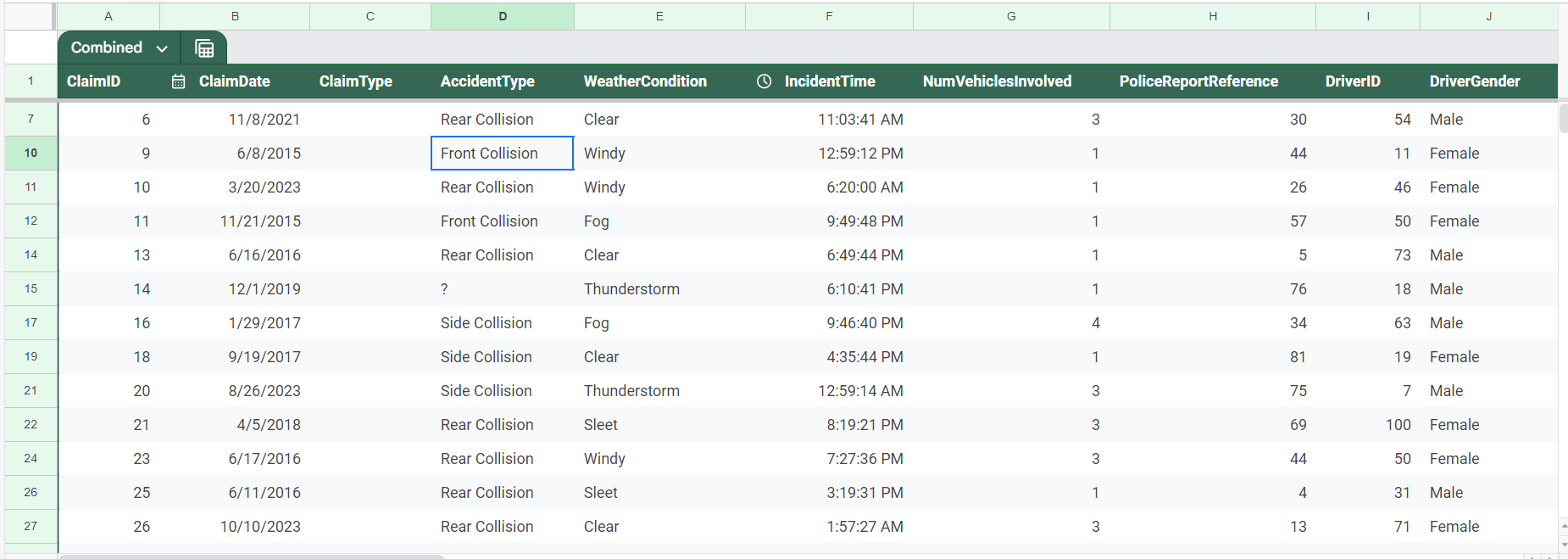
To overcome these issues and to check the validity and effectiveness of the model the following tests should be done; cross-validation, evaluation metrics, hyperparameter tuning, and external validation:

* Cross-validation: Is a prominent method used for estimating errors in tests that can then be used to choose the best model. This is done by randomly dividing data into k equal-sized parts and then fitting the model into k - 1 parts while obtaining predictions for the left-out part. This is done for each part and then combined, commonly a 10-fold or 5-fold manner. This allows for the assessment of the model's performance across various data sunsets resulting in the minimisation of overfitting
* **Evaluation metrics**: Recall, Accuracy, F1-score, and Precision, will be used to evaluate the performance metrics of each model focusing on reducing false positives and negatives.
* **Hyperparameter-tuning:** tuning K in the algorithms we are controlling the complexity of the model thus, avoiding overfitting. This will be done using grid search or random search to fine-tune the hyperparameters of each model for optimal performance.

## **3.3. Test Dataset**

The mentioned modelling techniques above all need test a dataset with the following industry standards:

* **Balanced Dataset:** A dataset with an even distribution of fraudulent and non fraudulent claims to ensure the model can effectively differentiate between the two classes. To validate if the claim is fraudulent or not the data set will contain a column that specifies whether the claim made was a “fraud” or ”not fraud” enabling the AI model to identify patterns in these sets of data.
* **Representative Features:** The features included in the test dataset should be relevant in helping the model detect fraudulent claims, such as police report, claim type, accident type etc.
* **Labelled Data:** The test dataset should be labelled as “Fraud” or “ not fraud” in each instance to evaluate the models accuracy, precision, recall and overall performance.
* **Recent Data:** The Test data should be up to date to reflect current fraud patterns ensuring the model stays relevant and effective,



***Figure 2: Example of test data given to AI Team***

# **4. Evaluation**

## **4.1. Evaluating Results**

**Null values:**

While reviewing our results, we should ensure that any rows missing any data should be null and the number of null values should also be counted. This will provide a clear understanding about the context of missing data. To solve the case of the null values, we can drop the observations with missing values. We can also use other data imputation methods such as mean, median, KNN to fill in the null values.

**Volume of Data:**

When we evaluate the volume of data, several details would need to be considered. The overall volume of data can impact the performance of the system. However, for our case, the volume of data is not extensive enough so performance should not be affected. While considering the volume of data, we have to ensure that the infrastructure is scalable. The number of rows can be increased, therefore, it has to be ensured that our system is scalable which it is in our case.

**Data Diversity:**

We have to check and analyse the variety of data. If there is a vast range of data, we can conclude that the data is diverse which will help while training the model as there is more to consider and make a decision. A low diversity of data will not be ideal as it will not consider most situations. A diverse dataset will help to avoid bias that can result in inaccurate outcomes while training the models. It is critical that the data is diverse as diverse data will provide a clearer picture which will lead to accurate outcomes.

**Gold Layer Standards:**

The Gold Layer is our most important layer. In the gold layer, the dataset will be well refined so that it can be trained by the AI team’s model. So, while evaluating the results, it is important to ensure that the data in the gold layer is consistent and organised. So, the data in the gold layer should be cleaned and there should be any biases. The gold layer data should be ready for analytical processes as the machine learning models will be trained on this dataset. So, we have to assess the performance of the gold layer by checking the data retrieval time and query speed. Moreover, the gold layer will also store the output from the AI team. So, it is crucial that the gold layer is capable of handling that process.

**Quality of Each Column:**

It should be ensured that each column represents legitimate data. For example, the dates should be actual calendar dates, times should be in AM or PM according to the clock. Most importantly, the columns should represent relevant data. While cleaning the data, irrelevant columns should be dropped. Moreover, the data should be consistent. The rows in the columns should follow the same method. For example, the dates should be in DD/MM/YYYY format. Each column should be documented which can explain the data types, ranges, transformations, etc and the columns should be clearly explained so that it is easy to understand.

## **4.2. Review Process**

While reviewing the process, it is important to ensure that the process follows the business requirements set by the client. The business rules set by our client were using a GenAI Solution, using Azure Services, Anonymise personal data, a front-end user interface, and Agile Delivery. The dataset will be reviewed by the Data team and then by the AI team who will give feedback and then a feedback loop will form to enhance the dataset. There should be data validation checks to ensure that the data is accurate. The whole process should also be documented so that each and every step in the process can be explained. The code used should also be reviewed to ensure that the process is not error prone. Finally, there should be clear communication between the Data team and the AI team and also with the client to clear any confusion and streamline the process.

## **4.3. Next Steps**

If the results are positive after the review process, we can scale up and add more rows to the dataset. Adding more rows will provide a more accurate solution as there are more rows of data to consider. As our current infrastructure is made to handle larger volumes of data, this should not be a problem and the process will be smooth. However, if the results are not positive after the review process, it may be worth revisiting the strategy and try to fix it or use a different methodology. Collaborating with the AI team and the clients will be crucial to discuss the next steps and setting up a plan for the methodology will be essential.

# **5. Deployment**

## **5.1. Deployment Plan:**

* **Deliverables to the Client:**
  + **Gold Layer Dataset:**
    - A fully processed and refined dataset stored in the Gold layer, ready for AI model training.
  + **Documentation**:
    - Comprehensive documentation detailing the data preparation process, including the steps taken to clean, synthesise, and integrate the data. This will include guidelines for accessing and using the datasets.
  + **Training:** 
    - Training sessions for the client’s team on how to interact with the Gold layer dataset, how to use the provided documentation, and how to collaborate effectively with the AI team during model development.

#### 

## **5.2. Monitoring and Maintenance:**

* **AI Model Performance Monitoring:**
  + The AI model, expected to be trained using OpenAI’s GPT-4o mini, will be closely monitored for performance metrics such as test accuracy. With the threshold acceptability being around 85-90%, if the model's accuracy falls below it, the team will work together to assess this issue.
  + This process will include examining the AI models logs and parameters while also reviewing the quality of the underlying data.
* **Data Re-evaluation and Refinement:**
  + **Inaccuracy Handling:**
    - If the AI model’s accuracy is found to be suboptimal, the data team will revisit the Silver layer, where a version of the combined dataset can be pulled for further cleaning and refinement. This refined dataset will then be stored in a new folder within the Silver layer and subsequently moved to the Gold layer for retraining the AI model.
  + **Overfitting Prevention:**
    - In cases where the AI model appears to be overfitting or performing too accurately, indicating a lack of diversity in the data, the data team will generate additional synthetic data. This may involve deliberately introducing errors or variations during the synthesis process to enhance the dataset's diversity.
* **Data Expansion:**
  + To improve the robustness of the AI model, the dataset may need to be expanded. This could involve generating more synthetic data or employing different synthesis techniques to create a larger, more varied dataset. If a synthesis method is found to be insufficient, alternative techniques will be explored to ensure the highest quality of synthetic data.

#### 

## **5.3. Final Report and Project Review:**

* **Final Report:**
  + A comprehensive final report will be produced, summarising the entire project, including data preparation, synthesis processes, AI model training results, and any refinements made during the deployment phase.
* **Project Review:**
  + A Comprehensive review of the project will be carried out, which will include assessing the effectiveness of the deployment, if the data is of quality and the overall performance of the AI model. Lessons learned will be documented to inform future projects.

# **6. References**

Mwitiderrick. (2018). insurancedata/insurance\_claims.csv at master ·mwitiderrick/insurancedata. GitHub. <https://github.com/mwitiderrick/insurancedata/blob/master/insurance_claims.csv>

# **7. Feedback and Response**

On Friday the 23rd of August we sent our Requirements/Scoping Document to the client (EY). following this we had a face to face meeting with the client on the 28th of August from 8am to 9am at the EY building to have our weekly stand-up and discuss the feedback for the document

| **Section** | **Feedback** | **Response/Action Points** |
| --- | --- | --- |
| **Data Understanding** | 1. Data Preparation steps were “a little vague” 2. Including a diagram of the layers might be handy to break up large amounts of text. 3. Needed to be specific about why the data is data “well-vetted” and why it has good data quality. | 1. Added more details into the data preparation steps and what it will entail, specifically data cleaning. Also noted that this process might change as the course of the project. 2. Added a detailed, colourful diagram of our data pipeline project idea along with steps which guide viewers on how to follow the pipeline. 3. Added details into how our data is well-vetted and what characteristics and attributes of our data make it high data quality. Including it has 39 columns and 1000 records. |
| **Data Preparation** | 1. Justify the use of a single dataset. 2. Mention that the dataset was chosen according to the specific requirements. 3. Explain the data synthesization techniques in detail for the future deliverables. 4. Specify the certain azure functions. 5. Explain the unnecessary steps avoided if separate pipelines are used. | 1. Explained why we are using a single dataset as it is more consistent. However, this is likely to change in the future. 2. Mentioned that the dataset was chosen following the requirements. 3. Will discuss more about the techniques in the later deliverables. 4. Will be discussed later as it is not decided yet. 5. Explained the unnecessary steps along with an example. |
| **Data Modelling** | 1. Emphasise that this section is not part of our project scope 2. Fix the formatting for the benefits and limitation so it is uniform across all techniques 3. Explain the reasoning behind the test data requirements | 1. Expanded why the modelling section is not part of our project scope and mentioned that the AI team is responsible for this so markers have a better understanding of our project. 2. Fixed any formatting issues so everything is uniform. 3. Expanded on our test dataset requirements to better improve our explanation on how we are validating whether a claim is fraudulent in your dataset? |
| **Evaluation** | No suggestions were received the mentors were satisfied with the contents of this section | Based on the feedback given above we expanded on some of the parts that seemed vague and fixed some formatting errors so the document was uniform. |
| **Deployment** | No suggestions were received the mentors were satisfied with the contents of this section | Based on the feedback given above we expanded on some of the parts that seemed vague and fixed some formatting errors so the document was uniform. |

**Updated Team Manual**

# 

# **1. Team Organisation and Structure**

| **Name** | **Role** |
| --- | --- |
| **Noorullah Khan** | Project Manager / Data Engineer |
| **Tashiya Vilathgamuwa** | Document Control Coordinator / Data Engineer |
| **Ninuri Mahagoda** | Document Control Coordinator / Data Engineer |
| **Aasnayem Gazzali Chowdhury** | Data Engineer |
| **Quoc Hung (Alan) Nguyen** | Devops Engineer / Data Engineer |



As Sprint 2 wraps up, our team has made some changes to the roles of each member moving forward. Despite these changes, we are maintaining a flat hierarchy due to only having 5 members. This structure will simplify our communication pattern and allow for easier management. The key difference this time is that roles have been simplified to align more closely with what each member will be doing. Notably, Tashiya and Ninuri will now be involved in the technical work. Since the project focuses on building a data pipeline, everyone will be working primarily as a Data Engineer, along with any additional responsibilities they may take on. Given the nature of the solution we’re developing, it makes sense for all team members to take on Data Engineering tasks.

**Noorullah Khan:**

1. Noorullah still holds the role of **Project Manager**. He will be responsible for assigning tasks and setting deadlines with other members.
2. Additionally, Noorullah will now become a **Data Engineer** who will be working collaboratively with other team members to build a data pipeline on Azure to build a dataset to train a generative AI model for the group 13 who also work with EY in this program. Noorullah, currently, is working on collecting a large enough raw dataset before we can work with it.

**Tashiya Vilathgamuwa and Ninuri Mahagoda:**

1. Business Analyst’s related work is mostly done, it will just come back every now and then. Therefore, both Tashiya and Ninuri have decided to transition into the technical part and become **Data Engineers**. As above, they will be working collaboratively with other team members to build a data pipeline on Azure to build a dataset. Currently, Tashiya is working on synthesising data to create an even larger dataset or fill missing data. Ninuri, on the other hand, is working on cleaning the data and working on the code convention for the team to follow.
2. They will still be working jointly as **Document Control Coordinators**. This role emphasises that they will be in charge of ensuring that our Deliverables are written in accordance with the definition given by the University. Also, they will be checking to ensure that we are not plagiarising or using AI-generated content.

**Aasnayem Gazzali Chowdhury:**

1. Aasnayem still holds the role of a **Data Engineer** which will be to build a data pipeline on Azure. Aasnayem, at the moment, is working closely with Noorullah to gather raw data sources and decide on methods to clean the data. More importantly, Aasnayem will be spending most of his time researching on data quality standards which would act as the goal for our dataset, so that we can come up with effective solutions to clean the data.

**Quoc Hung (Alan) Nguyen:**

1. Similar to everyone else, Quoc will be working as a **Data Engineer** also which would mean writing scripts to process the data gathered before.
2. Additionally, Quoc will act as a **DevOps Engineer** who will be responsible for managing the infrastructure of the solution on Azure as code and write scripts to automate the process of testing and continuously deploy code on Azure. Lastly, a DevOps Engineer will be responsible for managing the GitHub repository.

## 

# **2. Team Values & ACS Code of Professional Ethics**

## **2.1. ACS Code of Professional Conduct**

As we progress through our project with EY it is important to uphold the ACS Code of Professional Ethics to guide the behaviour and decision-making showcasing our professional conduct to the client. In this updated team manual we will discuss the core values of honesty, trustworthiness, respect for others and respect for the profession in correlation to our current ICT environment (ACS, 2023). This highlights our commitment to professionalism as we engage and support our team values and ethics throughout the course of this project.

## **2.2. Honesty**

Honesty remains one of our key values in this project as we continue to interact within our team and with the client and facilitate open and professional communication (ACS, 2023). As our project advances we have become more entangled with Team 13 as we have started to collaborate more with each when working with the client. To facilitate open and truthful communication with both teams we have opened up our private team channels on the discord so both teams are able to see what is discussed by the other team and what they are currently working on or discussing. Additionally a combined discord chat was created so that both teams can talk to each other on problems and discuss tasks without disrupting the individual team chats. Furthermore, both teams have combined our Gannt charts on Jira so that all team members know the progress of the tasks and any roadblocks we may face as well as an overview of the project status as a whole.

## **2.3. Trustworthiness**

Maintaining trust between the team members and with the client is vital for the success of the project (ACS, 2023). As we have undertaken tasks and roles in the project in line with our communicated abilities we have strived to maintain the trust of our members as we complete tasks to the best of our abilities and knowledge. We have further developed our capabilities to include GitHub, python and data synthesisation techniques as we engage in continuous development. As we have reached the building stage of our project we have cleaned our dataset to remove personal information from the dataset such as Names, Addresses and Phone numbers ensuring the data is anonymised.

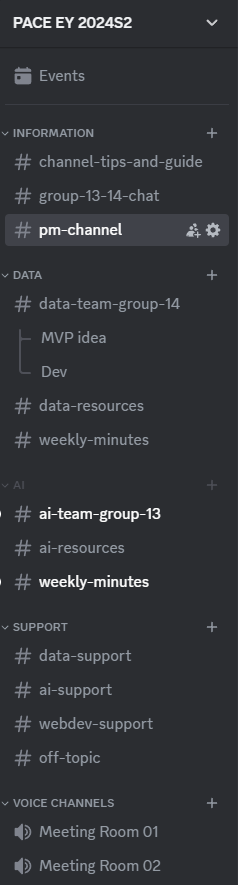
## **2.4. Respect**

Having respect for individuals around us and the profession we do is a fundamental part of collaborative success (ACS, 2023). We continue to have weekly meetings on the project to discuss our ideas and findings to the group and hear the various perspectives and opinions each member has on various topics. We have also started to have whole team meetings with Team 13 on the project as a whole to encourage communication between peers and for more ideas and perspective to be brought on ideas and issues. We have supported our members' advance in “ICT knowledge and competence” as we push our members to learn about data synthesisation techniques, and Terraform (ACS, 2023). Ultimately, our project is based on the notion of “advancing ICT capabilities and systems” for EY as we incorporate Gen AI to improve an existing system for the client (ACS, 2023).

# **3. Communication Plan and Meet Schedule**

## **3.1 Communication**

Our Discord server seems to have reached a stable stage and this structure will likely be kept moving forward. It has been set up as follows:



There are 5 categories including:

1. **INFORMATION:** channels inside this category is mainly used to make announcements and relay information from weekly meeting with sponsor to both teams
   1. channel-tips-and-guide: as the name suggests
   2. group-13-14-chat **(NEW)**: this channel will be used for cross-talk between the teams.
   3. pm-channel **(NEW)**: this channel is public but reserved for the two Project Managers to communicate with each other.
2. **DATA:** channels inside this category will be data related
   1. data-team-group-14 **(CHANGES)**: is now a public chat channel that can be viewed by anyone in this server. However, only the mentors and group 14 can send messages in this channel. We are assigned to work on the Data team. This chat will be our main mode of communication. The reason for the change is to improve the transparency which will help the other team understand the work we are doing better.
      1. MVP idea: is an example of a thread, which is a side discussion we want to have with a few selected members. The usage of a thread will help not clutter the main chat channel
   2. data-resources: is a public chat channel where anyone can post any data related resources, including data sources and blogs about cleaning data.
   3. weekly-minutes: is a public chat channel where we will post meeting minutes for any meeting we might have.
3. **AI:** channels inside this category will be AI related.
   1. data-team-group-14 **(CHANGES)**: is now a public chat channel that can be viewed by anyone in this server. However, only the mentors and group 13 can send messages in this channel. This chat will be their main mode of communication. The reason for the change is to improve the transparency which will help our team understand the work they are doing better.
   2. data-resources: is a public chat channel where anyone can post any AI related resources, including where to find open-sources models and how to fine-tune a model.
   3. weekly-minutes: is a public chat channel where group 13 will post meeting minutes for any meeting they might have.
4. **SUPPORT:** channels inside this category will mainly be used to ask mentors for support in specific problems
   1. data-support: is a public chat channel to post data related questions
   2. ai-support: is a public chat channel to post AI related questions
   3. webdev-support: is a public chat channel to post web development related questions
   4. off-topic: is a public chat channel to post any miscellaneous questions
5. **VOICE CHANNELS:** channels inside this category will mainly be used as virtual online meeting rooms where we can organise our weekly meeting.

## **3.2 Meet Schedule**

* Face-to-face meeting with sponsor: 8am to 9am weekly on Tuesday at EY office 200 George street, Sydney 2000
* Face-to-face team meeting & work together: 9:30 to 1pm weekly on Tuesday after the meeting with the sponsor.
* End of week online meeting: 11am to 12:30am weekly on Saturday using Discord

Compared to Deliverable 1, there has been no change to our meeting schedule as it has proven to work very well for our team. The only note is that for the post-EY meeting that we have, we will dedicate roughly 30 minutes to discuss some ideas with group 13. After that, the two teams will split and our team will continue our internal meeting and work together.

In summary, every week, we will have a meeting with our sponsor at their office to showcase what we have done throughout the week and decide on the direction for the next step. This meeting will be our formal way of demonstrating our work to the sponsor besides our informal chat on Discord with them. Following this meeting will be a 90-minute meeting for our team. The purpose of this meeting is to build team rapport and also discuss our plans and work for the week. Lastly, at the end of every week, we will have an online meeting on Saturday to check on the work we have done as well as planning for our next meeting with the sponsor. The purpose of these meetings can be overlapping depending on the tasks that we have at hand. However, we aim to meet with each other 3 times a week to keep everyone accountable and build trust in each other.

# **4. Conflict Resolution/Negotiation**

## **4.1. Conflict resolution**

**Approach to addressing conflict**

In the execution stage, conflicts can occur due to increased pressure, overlapping responsibilities, or differences in work styles. Our approach is to address these conflicts promptly and constructively to avoid any disruption to project progress. [updated]

**Conflict resolution procedure**

1. Recognise the conflict among team members
2. Organise a meeting involving the team members
3. Give each member an opportunity to share their viewpoint
4. Pinpoint the cause of the disagreement
5. Brainstorm potential solutions collectively as a team
6. Reach an agreement on a resolution
7. Put the solution into action and do a follow up

**Guidelines for Constructive Conflict Resolution**

* Direct attention to the problem, not personal characteristics
* Use "I" statements to communicate emotions and concerns
* Engage in active listening
* Strive to grasp all viewpoints before suggesting solutions
* Be willing to negotiate
* Maintain professionalism and respect consistently

## **4.2. Escalation process**

If conflicts cannot be resolved among team members:

1. Approach the team leader or project manager to address the matter.
2. If a resolution is not reached, escalate the concern to MQ’s leaders.

#### 

## **4.3. Preventative steps[Updated]:**

To minimise conflicts within our team:

* Encourage regular team-building activities
* Provide opportunities for open feedback
* Address potential issues early before they escalate
* Regular check ins to identify potential conflicts early on
* Clearly defined processes to ensure all team members understand the workflow to avoid overlap and confusion
* Real time feedback: Encourage team members to provide feedback on issues as they arise rather than waiting for formal meetings.
* Stress management: Note to be mindful of the pressure during execution and team members can help each other through offering support to manage stress.

By applying these updated strategies to overcome any conflicts in the execution stage the team can effectively manage conflict, ensuring that the project stays on track and that the working environment remains positive and productive. [**Updated**]

# **5. Revision History Table**

| **Version** | **Last Update** | **Author(s)** | **Changes** |
| --- | --- | --- | --- |
| V2\_S0 | 08.08.24 | Noorullah Khan | Removed from the team manual |
| V2\_S1 | 27.08.24 | Quoc Hung (Alan) Nguyen | * Ninuri, Tashiya, Noorullah will transition to the role of Data Engineer.while remaining their previous roles * Alan will act as a Dev Ops Engineer while remaining Data Engineer |
| V2\_S2 | 22.08.24 | Ninuri | * Honesty: Opened communication with AI team * Trust: Completed tasks and engaged in continuous learning * Respect: Whole team meeting with Team 13 and push others to learn new concepts relevant to the current stage |
| V2\_S3 | 28.08.24 | Quoc Hung (Alan) Nguyen | * Updated communication through new combined group chat. * Updated communication through PM channel being public for all to view. * Both teams private chats are now open for viewing to see discussions |
| V2\_S4 | 25.08.24 | Tashiya Vilathgamuwa | * Updated the approach to addressing conflict resolution to reflect the execution stage. * Updated preventative steps to reflect execution stage. |

# 

# **6. References**

ACS. (2023). *ACS Code of Professional Ethics*. ACS; Australian Computer Society. https://www.acs.org.au/content/dam/acs/CodeOfProfessionalEthics\_Mar\_2023.pdf