Deliverable 2 Part 1

[List of Assumptions 5](#_kj6vxz2txa0w)

[**Revision History Table 6**](#_29u1pvk9hcpy)

[**1. Introduction 6**](#_ey84htrmwg41)

[1.1. Statement Of Purpose 6](#_qjhs061mrs6d)

[1.2. Scope 7](#_lkut0qrfjjxf)

[1.2.1 Justification 7](#_4uf4b0jgqo2y)

[1.2.2. Scope 7](#_rjyuljif6o41)

[1.2.3. Objectives 8](#_xyjk61ckia3q)

[1.2.4. Deliverables 8](#_54jjshzg3zw2)

[1.2.5. Exclusions (Out of Scope) 8](#_8v8472ht4ytg)

[1.2.6. Constraints 8](#_n265lwlww75s)

[1.2.7. Assumptions 9](#_6rg2mkgr1ak)

[**2. Risk Management 9**](#_lqjosrusuuua)

[2.1 Risk Analysis Table 9](#_g57y68clqzhr)

[2.2 Risk Monitoring 9](#_la9wnt7cbt6d)

[2.3 Risk qualification and prioritisation 9](#_g1eo1ea9bjrs)

[**3. Resource Management 10**](#_2y3v9jdga453)

[3.1 People: 10](#_44sxmq3i7ou8)

[3.2 Hardware: 10](#_hwh7gbdfqw4f)

[3.3 Software: 11](#_qob3qjqm8oue)

[3.4 Other Resources: 13](#_mbb3znk6zq6e)

[**4. Change Management 14**](#_cb5ra97yr809)

[4.1 Changes to requirements and scope 14](#_xeiv48wsckg)

[3.2 Changes to documents, code and data 14](#_to5qqghujx3x)

[3.2.1 Branching strategy: Issue-based 15](#_jedd9268oajn)

[3.2.2 Code Reviews: Enforcing Quality with Pull Requests and Github Actions 15](#_t1t9k8yz6kdg)

[3.2.3 Documentation versioning 15](#_e7o84684gu1a)

[3.2.4 Immutable storage: 16](#_cwyfyno7l3dh)

[**5. Quality Management 16**](#_lekybt1055ne)

[5.1. Objectives 16](#_h6i4fbli6icu)

[5.1.1. Ensuring data quality 16](#_tp7xfigdsi55)

[5.1.2. Model Accuracy 17](#_dwbw2mcsigel)

[5.1.3. Compliance with business requirements 17](#_gohx5xep14nk)

[5.1.4. Customer Focus and User Satisfaction 17](#_qtbo2vktu0pk)

[5.2. Data Quality Management: Monitoring and Assurance 18](#_d9qynmim2cp7)

[5.2.1. Data Collection Process 18](#_qy1g9rvf6euw)

[5.2.2. Exploratory Data Analysis (EDA) 18](#_6m2b1emo52ya)

[4.2.3. Data Cleaning 19](#_5zdknymy0jxw)

[5.2.4. Data Standardisation 19](#_wtehbtbxtwqy)

[5.3. Human Resource Management: Agile Delivery 20](#_fkxjyeyjsrks)

[5.3.1. Sprint Planning 20](#_71loxbsf5llm)

[5.3.2. Iteration Reviews 20](#_g28oahdueu17)

[5.3.3. Continuous Feedback Loop 21](#_etfw87v3wn8l)

[5.4. Development Management 21](#_8vusy3gogao5)

[5.4.1. Code convention 21](#_obln1aviwgdl)

[5.4.2. Prototype 21](#_fbalu9o460av)

[5.5. Documentation Management 21](#_jmrubmu6lojh)

[5.5.1. Formatting 21](#_jv1w7xz3cttp)

[5.5.2. Documentation control officers 22](#_gvcm2nvszrgf)

[**6. Schedule 22**](#_entroc15wpqc)

[6.1. Project Overview 22](#_hn2tdt7nve5)

[6.1.1. Deliverables 23](#_lrkiwbincapt)

[6.2. Project Phases and Key Milestones 23](#_5j8wvf4ignmv)

[6.3. Schedule - Gantt Chart 23](#_avcw20yzhg5a)

[6.4. Resource and Task Allocation 24](#_ftqw611qqx4g)

[6.5. Tools and Technologies 24](#_2ruauu2u53xh)

[6.5.1. Jira 24](#_7pkb895335j)

[6.5.2. GitHub 24](#_xekn66svh0kl)

[**7. Handover requirements 24**](#_s9jsgqm9645g)

[7.1 Software and Tools 25](#_5jm05f50vsvk)

[7.2 Data Deliverables 25](#_hhnycii7yea5)

[7.3. Testing and Validation 25](#_nlw987petja9)

[7.4. Model Deliverables: (To be done by the AI Team) 25](#_jcin8mc98mc4)

[7.5 User and Training Manuals 26](#_cm5otmkttos5)

[7.6 Final Report and Evaluations 26](#_oiv7i7i1gd7t)

[7.7 Client Consultation and Acceptance: 26](#_9v1ioya3ssvh)

[**8. Assumptions 26**](#_gxmzuce4l6nz)

[8.1. Resources 26](#_x44v8bi9ibga)

[8.2. Availabilities 26](#_26yw3gq2l76c)

[8.3. Tools 27](#_tj2rlx15qgic)

[8.4. Techniques 27](#_rv41gm4hl33o)

[8.5. Standards 27](#_wdmc9n7b1giz)

[8.6. Communication 27](#_qugceikd6nz1)

[8.7. Expectations 28](#_m9p6ygrvpycn)

[**9. References 29**](#_7w2zlxn99xlw)

[**10. Appendices 30**](#_j35rc0g36jpe)

[Appendix 1. Risk analysis table 30](#_qy7uoh9zx8ow)

[Appendix 2. Risk Indicators Table 33](#_tyldfwk9mmf4)

[Appendix 3 Probability/ Impact Matrix 34](#_u5cm5re7oncu)

[Appendix 4: Project Phases and Key Milestones 35](#_ktl5ycyng7bp)

[Appendix 5: Sprint 1 Gantt Chart 37](#_xkx4usff5wy7)

[Appendix 6: Sprint 2 Gantt Chart 37](#_rkrl6hlrdzkh)

[Appendix 7: Sprint 3 Gantt Chart 38](#_1t7oxiho8kc1)

[Appendix 8: Sprint 4 Gantt Chart 39](#_xkxdm5s4bjlb)

[Appendix 9: Sprint 5 Gantt Chart 40](#_gugdm1jztf3g)

[Appendix 10: Task Allocation Groups 41](#_e437qncfqmrd)

[Updated Requirements / Scoping Document (0.25 marks Del 3 and 0.1 marks for Del 4) 43](#_7yqapkhzhhnt)

[**Revision History Table 43**](#_2hy0duuzboz1)

[**1. Data Understanding 43**](#_1yjkh1anmujp)

[1.1 Initial Data Sources 43](#_nb88drf4p74s)

[Raw Dataset 1: 1000 rows ~ 261 KB 44](#_550boq51ih0z)

[Raw Dataset 2: 10000 rows ~ 2870 KB 46](#_c2hcic46y9mz)

[Raw Dataset 3: 10000 rows ~ 1544 KB 47](#_g5jzp4l0szj3)

[1.2 Data Collection Methods 49](#_j86cvkk8rvlu)

[1.3 Assumed Data Quality, Structure, and Completeness 50](#_31r487wyurrz)

[1.4 Mechanisms for Evaluation 51](#_wsa4qo3jwpe6)

[**2. Data Preparation 52**](#_78jq27ck99z1)

[2.1. Data Preparation Activities 52](#_cw9vkorct2yx)

[2.2. Data Processing Pipelines 53](#_vsglpkaap9vo)

[2.3. Different Preparation Pipelines 54](#_kjgxuavyoxm9)

[2.4. Justification for Different Models 54](#_4bqm53ndl4c2)

[2.5. Data Storage Mechanisms 54](#_zh4yyeka4p9s)

[**3. Modelling 54**](#_2jxsxqh)

[3.1. Modelling Techniques 55](#_ok62gylqhmrx)

[3.1.1. Random Forest Classifier: 55](#_i534jd8at0dg)

[3.1.2. Random Forest Regressor: 55](#_ou9mcaqtbq42)

[3.1.3. Support Vector Regression (SVR): 56](#_37sedja8434b)

[3.1.4. K-Nearest Neighbours (KNN): 56](#_15du7qhed2hc)

[3.1.5. Relationship to the Chosen Model: 57](#_4702xt9sbv7y)

[● Random Forest Classifier: 57](#_r0z92efcdxfk)

[● Random Forest Regressor: 57](#_4ormndlnru6p)

[● Support Vector Regression (SVR): 57](#_n0j9c4hai0g2)

[● K-Nearest Neighbours (KNN): 57](#_kr8irnmzyd5v)

[3.2. Evaluation Metrics 57](#_4vi4vz8c233c)

[3.3. Hyperparameter-tuning 58](#_hbtzssk9su7t)

[3.4. Bias and Fairness 58](#_b3aukivr97ci)

[3.5. Test Dataset 58](#_1y810tw)

[**4. Evaluation 60**](#_2xcytpi)

[4.1. Evaluating Results 60](#_1ci93xb)

[4.2. Evaluation Techniques 61](#_3ypoxuretb8u)

[4.3. Review Process 61](#_3whwml4)

[4.4. Next Steps 61](#_2bn6wsx)

[**5. Deployment 62**](#_qsh70q)

[5.1. Deployment Plan: 62](#_3as4poj)

[5.2. Monitoring and Maintenance: 62](#_49x2ik5)

[5.3. Final Report and Project Review: 63](#_147n2zr)

[**6. References 63**](#_2wvhept83mj7)

[**7. Feedback and Response 63**](#_o2z4lvrz5okw)

[**1. Team Organisation and Structure 66**](#_h51d8fyuz0vb)

[**2. Team Values & ACS Code of Professional Ethics 68**](#_swzr988wnlyn)

[2.1. ACS Code of Professional Conduct 68](#_elv22pnri0u7)

[2.2. Honesty 68](#_rsft776bqpik)

[2.3. Trustworthiness 68](#_yvngz9yrynap)

[2.4. Respect 68](#_p97vmuryl7h0)

[**3. Communication Plan and Meet Schedule 69**](#_9o4dopri62zt)

[3.1 Communication 69](#_ujlctmz1edxj)

[3.2 Meet Schedule 70](#_xp07llg59fq2)

[**4. Conflict Resolution/Negotiation 71**](#_a73whcyuvf9u)

[4.1. Conflict resolution 71](#_rgbgbdugtwqa)

[4.2. Escalation process 71](#_y1z2y2erfgk5)

[4.3. Preventative steps[Updated]: 71](#_8iqsxfr6qphj)

[4.4 Documentation of Resolutions and Agreed Changes: 72](#_ujbn8winycgh)

[**5. Revision History Table 73**](#_mk1gden1ras7)

[**6. References 73**](#_82tcbiqw0wgb)

**NOTE: when revising the document, for any changes made please add a comment to that section detailing what has been changed**

**Common sections for all groups:**

Deliverables Certificate

Updated Project plan (with revision history + new “Handover Requirements” section)

Updated Requirements/ Scoping Document (with revision history)



### 

### **List of Assumptions**

**(0.25 marks Del 3 and 0.15 marks for Del 4)**

Provide a list of assumptions (relevant to design, requirements, project plan or any of the documentation). This will help you think about them and whether they are reasonable and also help the marker/review team to understand why you have done certain things. Please review the assumptions as a group before submission. A poor assumption will not be a valid reason for poor design decisions.

**Updated Project Plan**

Revised Project Plan (0.5 mark **Del 3**)

The following is required:

1. **Include the heading in the project plan: Handover requirements.** This section can be added after your schedule to explain in detail what the client needs. Handovers will be different according to what the project is. It's about making sure the client gets what they were expecting and needed before the team disappeared. So any data, software, algorithms, scripts, test cases, testing tools/environments, outputs and documentation would be handed over together with training/manuals that lets them go forward with it - including reports or test cases showing current output and/or what is fully functional or not.

For all projects, the point is to discuss with the client what they need at the end of the project. Make sure they have everything they need and know how to use what you have given them. There is no set format or requirement for the marker.

a. For games and software projects, think about who will do play testing and/or acceptance testing, asking questions such as when will testing start, when and what devices/sites does the final product need to be deployed on, who will maintain the system, how will the system be maintained, what are the user manual/training manual/installation/installation wizard/administration parameters requirements, what database/file setup is required, what documentation is required.

1. **A revised and updated version of your project *schedule*,** that takes into account any schedule revisions, covering the balance of the project, that have arisen since you submitted the previous schedule. Anything else that has changed such as team structure, resource allocation, risk evaluation, change management processes, communication strategy, etc should also be updated in the plan and **quality manual and included in the revision history**. Resubmit the entire updated document.

# **Revision History Table**

| **Version** | **Last Update** | **Author(s)** | **Changes** |
| --- | --- | --- | --- |
| V2\_S1  Introduction | 30/09/24 | Ninuri/Tashiya | * Expanded on the scope to include information on how we will perform NRMA’s fraud detection |
| V2\_S2  Risk Management | 30/09/24 | Tashiya | * Added More technical and Business related risks on to the Risk analysis table, Risk indicators table and Probability/impact matrix |
| V2\_S3  Resource Management | 30/09/24 | Tashiya | * Added the costs for each resource |
| V1\_S4  Change Management | 28/08/24 | Tashiya | * No changes made |
| V2\_S5  Quality Management | 30/09/24 | Ninuri | * Defined quality metrics and KPIs for each aspect of managing quality * Added more data standardisation * Added information on conducting exploratory data analysis (EDA) |
| V2\_S6  Schedule | 30/09/24 | Ninuri | * Updated the project schedule to cover changes in structure, resource allocation and schedule revisions * Added new project schedule images for completed and in progress sprints |
| V1\_S7  Handover requirements | 30/09/24 | Ninuri/Tashiya | * Added a new section “Handover requirements” which outlines the methods and requirements we will have in place when handing over this project to our client. |
| V2\_S8  Assumptions | 30/09/24 | Ninuri | * Added more assumptions on design, requirements, documentation and aspects of the 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.
* **Fraud Detection Model:** Coordinating with the AI team to create a GenAI model that can effectively detect fraud through machine learning algorithms such as anomaly detection and neural networks.
* **Model Training and Testing:** Collaborating with the AI team to train their model with the provided dataset and evaluating the model’s performance through testing to improve its capabilities.

### **1.2.3. Objectives**

* To develop a clean, accurate and comprehensive dataset for the AI model to be trained.
* To improve NRMA’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.
* To improve the current fraud detection process at NRMA using GenAI to reduce review time and operation costs.

### **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 (handled by AI team)
* Post deployment support or maintenance for the AI model or dataset
* Front end development of user interface for claims officer
* Data 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
* Synthesised 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 data science project ranging from technology and people related risks to organisational/ business related risks , 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. Additionally we’ve considered data and budget resources to ensure comprehensive management and readiness.

## **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.

**People Costs:**  It's important to note that our team is providing our labour free of charge, as this project serves as a valuable opportunity for gaining practical experience. This arrangement not only enhances our skills but also allows us to contribute meaningfully to the project's success without incurring labour costs.

## **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.

**Hardware Costs:**

* **Laptops/Desktops:** If hardware needs to be replaced or upgraded, the cost per device is estimated at $1,200. Total hardware budget allocation: $6,000.
* **Maintenance:** No formal IT support budget, but members are encouraged to perform regular updates. No additional budget allocation required for this.

## **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(ADLS2) :** 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.

**Software Costs:**

* Azure Data Functions - The first 400,000 GB/s of execution and 1,000,000 executions are free. Since we will be processing 10 MB of data, it is unrealistic to use up all of the free-tier capacity. Therefore, Azure Functions will incur a total cost of $0.
* Azure Data Lake Storage (ADLS2): Data is organised and stored in Azure Data Lake Storage according to its classification and access needs.
* **Cost Structure**:
  + Storage: The first 51,200 GB/month is charged at $0.031 per GB.
  + Write Operations: $0.10942 (for every 4 MB, per 10,000 write operations).
  + Read Operations: $0.00874 (for every 4 MB, per 10,000 read operations).

Overall Software cost:

* Total Operations Cost: Write Operations ($0.10942) + Read Operations ($0.00874) = **$0.11816**.
* Since we have three layers of storage (bronze, silver, and gold), the total operations cost for all layers is: **$0.11816 \* 3 = $0.35448**.
* Adding the base storage cost to the operations cost gives us:
* Base Storage: $0.031
* Total Operations Cost: $0.35448
* **Total Monthly Cost for ADLS2**: **$0.031 + $0.35448 = $0.38548/month** (rounded to **$0.39** for simplicity).

## **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.
* Data Acquisition (Kaggle): We used Kaggle to source datasets that meet business requirements

**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.

**Data and other resource costs:**

* **Data Acquisition:** Data from Kaggle is free, with no associated costs.
* Google Drive, Github and Discord are all free tools.

# **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. To address these concerns exploratory data analysis (EDA) will be implemented in order to analyse data distributions, potential inconsistencies, missing data and outliers. Key performance indicators (KPIs) that will be used to measure data quality include:

* **Data Accuracy:** verifying data against standards or benchmarks
* **Data Consistency:** comparing contradictions with similar datasets
* **Data Completeness:** measuring the percentage of missing values to be less than 20% of the dataset

### **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). To test the accuracy of the dataset when train by the model quality metrics such as:

* **Model accuracy:** precision, recall and F1 score with a percentage high than 85%
* **Training efficiency:** number iteration taken by the model to provide good results will be less than 50

Will be used as a way of gaining feedback on the quality of the dataset to be further improved and used again by the AI team in a feedback loop.

### **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 such as:

* **Azure Usage:** percentage of overall services deployed on Azure is over 90%
* **Agile Compliance:** Assessing the progress of sprints and speed of project delivery to be over 85%
* **Anonymisation:** 100% of PII is removed in all datasets and cases

### **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. To track the feedback and satisfaction of clients the following KPIs it be used:

* **Client engagement:** number/frequency of feedback sessions with client is at least once per sprint
* **Satisfaction scores:** given after each sprint on how accurately the product meets the users needs and has to be over 75%

## **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

### **5.2.2. Exploratory Data Analysis (EDA)**

EDA will be conducted on each dataset at the start of the dataset preparation stage as it allows for the exploration of data distributions and potential issues. By identifying missing data, inconsistencies, outliers of unusual patterns these issues can be addressed before the cleaning stage.

* **Correlation Analysis:** find correlations between features that could impact the model’s performance
* **Summary statistics:** spotting irregularities by checking mean, median, mode and the standard deviation
* **Data visualisation:** creating graphs and plots to examine the distribution of the data

### **4.2.3. 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):

* **Exploratory data analysis** - As a dataset may contain multiple error types it is important to do a comprehensive exploratory 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.

To check the accuracy of the dataset after cleaning these quality metrics will be held to ensure the quality of the final dataset:

* **Model Accuracy:** using accuracy score, MSE, MAE and r^2 the accuracy of the imputed missing variables in the dataset has to be above 85%
* **Transformation Analysis:** number of duplicate rows and missing values has to be 0
* **Verification Check:** results from the AI model have to be above 90%

| **Quality Metric** | **Initial Result** | **Actions Taken** |
| --- | --- | --- |
| **Model Accuracy** | Accuracy score, MSE, MAE and r^2 of certain columns using KNN and random forest were above 95% however some only had a score of 70% | Hyper-tuning was done using GridSearchCV and RandonSearchCV to increase the scores |
| **Transformation Analysis** | At the beginning all datasets had missing values. | After imputing missing values there were no missing values in the dataset |
| **Verification Check** | Dataset with missing values resulted in an accuracy of less than 50% | Missing values were imputed or rows were dropped to create a complete dataset |

### **5.2.4. 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 and should adhere to the following metrics:

* **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. All columns should be converted to numerical values as the AI model only trains with numerical values
* **Column Name:** All column names have to be the same across datasets
* **Categorical Data Categories:** Data under categorical columns should be named the same across all datasets. E.g. for Education Level all dataset should have the categories High School, Primary School etc. not Secondary School or Tertiary.

## **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. To track the effectiveness of each sprint the following KPIs will be used:

* **Sprint completion rate**: 90% of the sprint is completed within the defined time frame
* **Team velocity:** increasing amount of tasks to be done by the team each sprint

| **Quality Metric** | **Initial Result** | **Actions Taken** |
| --- | --- | --- |
| **Sprint completion rate**: | 100% of Sprints 1, 2, 3, 4 have been complete on time | N/A |
| **Team velocity:** | Number of tasks in each sprint is greater than the previous | Work is split effectively amongst the team and deadlines are met. |

### **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.

* **Client satisfaction score:** A satisfactions score of over 85% is given after each sprint review
* **Change request rate:** Less than 3 change requests are made per stand-up
* **Feedback turnaround time:** feedback from client is given with 48 hours so adjust can be made promptly

| **Quality Metric** | **Initial Result** | **Actions Taken** |
| --- | --- | --- |
| **Client satisfaction score** | Client is highly satisfied with progress at each sprint | N/A |
| **Change request rate** | Maximum of 2 changes have been request from all the stand-ups so far | Requests are noted and implement by the next stand-up |
| **Feedback turnaround time** | Minor feedback on items present during stand-up are given immediately. However feedback needed for deliverables are not given on time and may take a week or longer. | Follow up messages and email are sent to client and may need to escalate with teaching staff if the issue continues to arise. |

### **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.

* **Client engagement frequency:** interaction with at least 2 mentors per sprint
* **Feedback incorporation rate:** 100% of feedback incorporated within the next sprint

| **Quality Metric** | **Initial Result** | **Actions Taken** |
| --- | --- | --- |
| **Client engagement frequency** | There is atleast 3-4 mentors discussing the project across all team members. | Communicate with different mentors for a variety of feedback and ask question on crucial topics. |
| **Feedback incorporation rate** | 90%-100% of feedback is incorporated within the next sprint or the following sprint. | If feedback cannot be incorporated discuss with mentors and ask questions for further clarification. |

## **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
* Table of Contents: Separate one for each document and must link to each heading
* Cover Page: At the start of each document

### **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 GenAI 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 **(completed)**
  + Forming the team, getting to know each others skills and proficiencies, and creating the communication channel
* Sprint 2: Building Stage 1 **(completed)**
  + Brainstorming ideas for our MVP taking into consideration alternatives and feasibility
* Sprint 3: Building Stage 2 **(completed)**
  + Finalising the project plan and creating a test dataset and synthesised data for training
* Sprint 4: Building Stage 3 **(completed)**
  + Hand over prototype dataset to AI team to train their AI model
* Sprint 5: Testing Stage 1 **(in progress)**
  + 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
* MVP proposition document
* 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, 7, 8 and 9 for the Gantt chart figures

## 

## **6.4. Schedule Revisions**

| Sprint | Task | Revision |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 

## **6.5. 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. As we reach the end of our sprint 4 and begin our sprint 5 we have decided to increase the collaboration with the AI team and have two members work with the AI team temporarily to ensure issues with the AI model and dataset are fixed and problems can be communicated more effectively. Refer to Appendix 10 for the Task Allocation groups.

## **6.6. Tools and Technologies**

### **6.6.1. 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.

### **6.6.2. 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. Handover requirements**

During the handover process we as the data team have to present to the client NRMA the necessary documentation, resources and tools required to effectively use the fraud detection system after the project has been completed. The following section will cover the vital documentation that is handed over to the NRMA at the end of the project. Through discussions with the client we have confirmed the necessary handover requirements that have to be met in order for them to maintain and develop the system.

## **7.1 Software and Tools**

* **Environmental Setup:** Instructions to set up the Azure environment used throughout the project such as Azure Functions and Data Lake.
* **Automation Scripts:** Details on software requirements such as libraries and packages to run scripts as well as a guide on running scripts that were created to automate model training, data processing and deployment tasks.
* **Deployment Instructions:** Clear documentation on deploying the scripts and fraud detection model with necessary configurations

## **7.2 Data Deliverables**

* **Final Dataset:** The complete and structured dataset utilised to train the fraud detection model, organised into various levels of data quality (Bronze, silver, Gold)
* **Data Documentation:** Detailed documentation outlining the structure, attributes and origins of the dataset including the preprocessing or transformations that were applied throughout the project.
* **Data Pipeline:** If any scripts were utilised to clean,load or structure the data these will be provided along with the guidelines on how to execute them.

## **7.3. Testing and Validation**

* **Validation Scripts:** Descriptions of scripts used to evaluate the accuracy of cleaning scripts or the evaluations of models used to synthesise data
* **Testing Tools:** Manual on testing process used such as environment, datasets and tools so it can be replicated.

## **7.4. Model Deliverables: (To be done by the AI Team)**

* **Fraud Detection Model:** Provide the completed and trained AI model, prepared and ready to be deployed.
* **Model Documentation**:A detailed explanation of the model, including its structure, algorithms, chosen hyperparameters and performance metrics.
* **Model code and Scripts:** All code used for the model training and evaluation, including any python scripts along with the necessary libraries and packages.
* **Model Outputs:** Examples of model outputs that demonstrate how fraudulent claims are flagged, accompanied by performance reports such as precision, recall and F1 score.

## **7.5 User and Training Manuals**

* **User Manual:** Detailed guide on how to use the fraud detection system with key directions to input data, troubleshoot issues and understanding the results.
* **Training Manual:** Documentation on how to use the model and interpret outputs with directions on how to retrain the model with current data.

## **7.6 Final Report and Evaluations**

* **Project Completion Report:** A thorough report that outlines the project’s objectives, methodologies and results, highlighting significant decisions and outcomes.
* **Evaluation Metrics:** Documentation detailing how success was measured, including performance benchmarks and how effectively the system achieved NRMA’s fraud detection objects and requirements.

## **7.7 Client Consultation and Acceptance:**

Before the final handover is conducted a meeting with the stakeholder’s at NRMA will be held to review all the required deliverables and ensure they meet the requirements of the client and project objectives. Confirmation that the NRMA team is able to use the new fraud detection system successfully. Ultimately a final testing process will be run to allow the NRMA team to test the final system in their environment and address and final adjustments before the project conclusion.

# **8. Assumptions**

## **8.1. 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

## **8.2. 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.

## **8.3. 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.

## **8.4. 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.

## **8.5. 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.

## **8.6. 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

## **8.7. 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.

## **8.8. Design**

* **Azure Services Compatibility -** assumed that the Azure services used such as Data Lake, Application Insights and Data Lake will be compatible with the infrastructure and Azure tools used by the AI team
* **Synthetic Data Accuracy -** the data synthesisation techniques used wll create 7,000 rows of synthetic data of high quality.

## **8.9. Documentation**

* **Access to documentation -** all team members will have unrestricted access to the Goodle Drive
* **Adherence to documentation standards** - all team members will follow the guidelines made for documentation and the documentation control officers will check for compliance.

## **8.10. Quality Management**

* **Dataset Quality -** the chosen datasets will be up the predefined quality standards
* **Coding Quality - t**eam members will adhere to coding standards so the codebase is consistent, readable and maintainable.

## **8.11. Risk and Contingency**

* **No delays in acquiring tools -** necessary tools, services and financial aid will be provided in a timely manner
* **Training and adaptation** - team members will be able to learn new tools and technologies introduced during the project quickly to reduce delays
* **No major technical issues -** during the project no major issues relating to tools, coding or Azure will arise.

# **9. 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]

# **10. 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. | Medium | Inefficiency in processing large datasets | Medium | Optimise code, plan for scalability. |
|  | Data Quality Issues | Incomplete or inaccurate data affecting model performance | Medium | Decreased model accuracy | High | Robust data cleaning and validation |
|  | Algorithm complexity | Overly complex algorithms increase cost and time | Medium | Higher Computational requirements | Medium | Simplify algorithms based on data needs |
|  | Infrastructure Limitations | Inadequate computational resources for large datasets | Low | Delayed model training | Medium | Plan for scalable infrastructure |
|  | Model overfitting | The model fits the training data too well but performs poorly on unseen data | Medium | Decreased model generealisationality, which leads to inaccurate predictions in production | High | Implement Cross validation, regularisation techniques and monitor performance on test data |
|  | Data Privacy Compliance | Mishandling of personal data , violating privacy laws | Low | Legal penalties, loss of client trust, project delays | High | Ensure data anonymization, adhering to privacy regulations, regularly audit data handling processes. |
|  | Version Control conflicts | Conflicting Code changes leading to loss of work or errors in the project | Medium | Delayed development progress, bugs introduced into the project | Medium | Establish version control best practises(Regular commits, branch management) |
|  | Insufficient Data for modelling | Lack of sufficient data for effective model training or poor representation of edge cases | Medium | Inadequate model performance, failure to generalise on new data | High | Use data augmentation, synthetic data generation(CTGAN) or seek additional datasets |
|  | Bias in Data | Presence of bias in training data, leading to unfair or unethical model predictions | Medium | Discriminatory model outputs, loss of trust from NRMA | High | Perform bias audits on datasets, ensure diversity in data. |
| 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 / Business 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. |
|  | Shifting priorities | Clients changing focus on deprioritizing the project | Low | Delays or project termination | Medium | Regular stakeholder communication |
|  | Market Changes | Shifts in market trends affecting the project( economic downturns, supply chain disruptions) affecting business needs or priorities. | Medium | Shift in business goals, project delays | Low | Monitor trends, maintain flexibility, maintain backup plans |
|  | Inadequate change management | Poor management of changes to project scope , objectives or timelines causing confusion and inefficiencies | Medium | Scope creep, delays or suboptimal results | Medium | Implement a structured change management process, ensure proper documentation, and align teams before making changes, |
|  | Failure to meet regulatory compliance | Inability to comply with industry specific regulations or data privacy laws, leading to legal issues. | Low | Legal penalties, reputation damage, or project cancellation. | High | Research on particular laws and document where we would need to ensure we meet compliance and standards. (Eg. Removal of personal identifiable information) when cleaning data. |
|  | Inadequate Resource allocation | Insufficient allocation of human or financial resources leads to slowdowns or compromised project quality | High | Deadlines missed, team burnout, reduced overall output. | High | Conduct thorough resource planning and regularly review resource usage to ensure adequate staffing and budget. |
| 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 | Low | 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 * Inadequate software integration, causing reduced efficiency * Scalability issues due to increased data volume or resource constraints * High frequency of system outages or downtime * Data quality issues leading to inaccurate results or system errors * Inconsistent version control leading to code conflicts or bugs * Security breaches or unauthorised access to sensitive data. |
| 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 / Business Risks | * Weak leadership or lack of clear direction from the project manager * Insufficient support from mentors * Delays in decision making or approvals from stakeholders * Shifting organisational priorities leading to deprioritization of the project. * Budget constraints impacting project resources and scope. * Market changes affecting the relevance or direction of the project. * Inadequate resource allocation leading to bottlenecks or delays. * Failure to manage changes in project scope or requirements effectively. * Policy changes disrupting the flow of the project. |
| 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** | P1: Skills Gaps  R1: Ambiguous Requirements  R2: Changing requirements | P1: Skills Gap | R1:Ambiguous Requirements  R2:Changing Requirements  O7: Inadequate Resource Allocation |
| **Medium** | T4: Scalability Issues  O4: Market changes | P2:Team collaboration  P3: Resource Availability  T1: System downtime  T6: Algorithm Complexity  T9:Version Control Conflicts  O5: Inadequate change management | O1: Budget constraints  E2: Underestimation of Complexity  E3:Timeline Overruns  T5: Data Quality Issues  T7: Model overfitting  T10:Insufficient Data for modelling  T11: Bias in Data |
| **Low** | E1: Resource Overcommitment | T3:Inadequate Software integration  E4: Missed Milestones  O3: Shifting priorities | O2: Policy Changes  T2: Data Security Breach  T8:Data Privacy Compliance  O6: Failure to meet regulatory compliance |

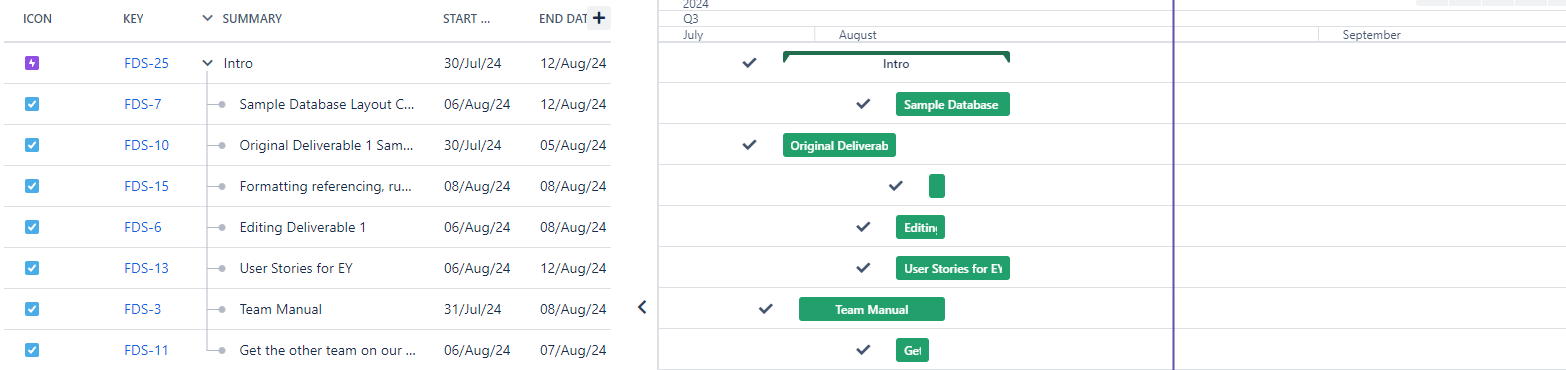
## **Appendix 4: Project Phases and Key Milestones**

| **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 | * Find additional data sources * Clean and enrich dataset 2 * Clean and enrich dataset 3 * Test various imputation methods * Create a MVP Proposition document * Submit deliverable 3 draft to client * Find and test methods for creating synthetic data * Create synthetic data | * Create merge dataset * Send 20,000 rows to the AI team * Creation of MVP | Project:   * N/A   Product:   * MVP1 presentation * Test dataset |
| Sprint 5  25/09/- 8/10  2024 | * Complete deliverable 3 tasks * Refine data pipeline * Resolve Azure permission issues * Move project onto Azure | * Feedback from client on Deliverable 3 * Set up the Azure infrastructure | Project:   * Updated Deliverable 2 * Design Document * Test Cases Document * Prototype/MVP   Product:   * MVP2 |
| 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: |

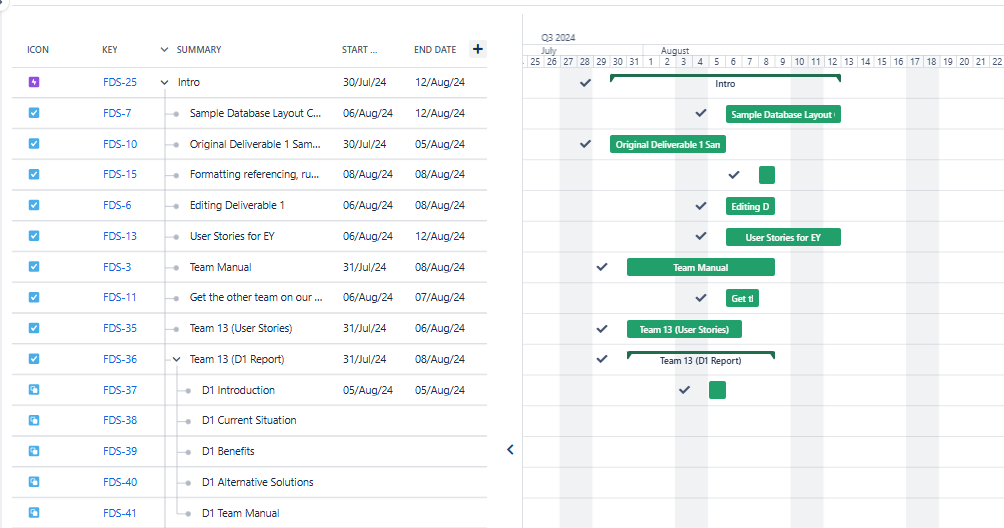
## 

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

Gantt Chart for Team 14



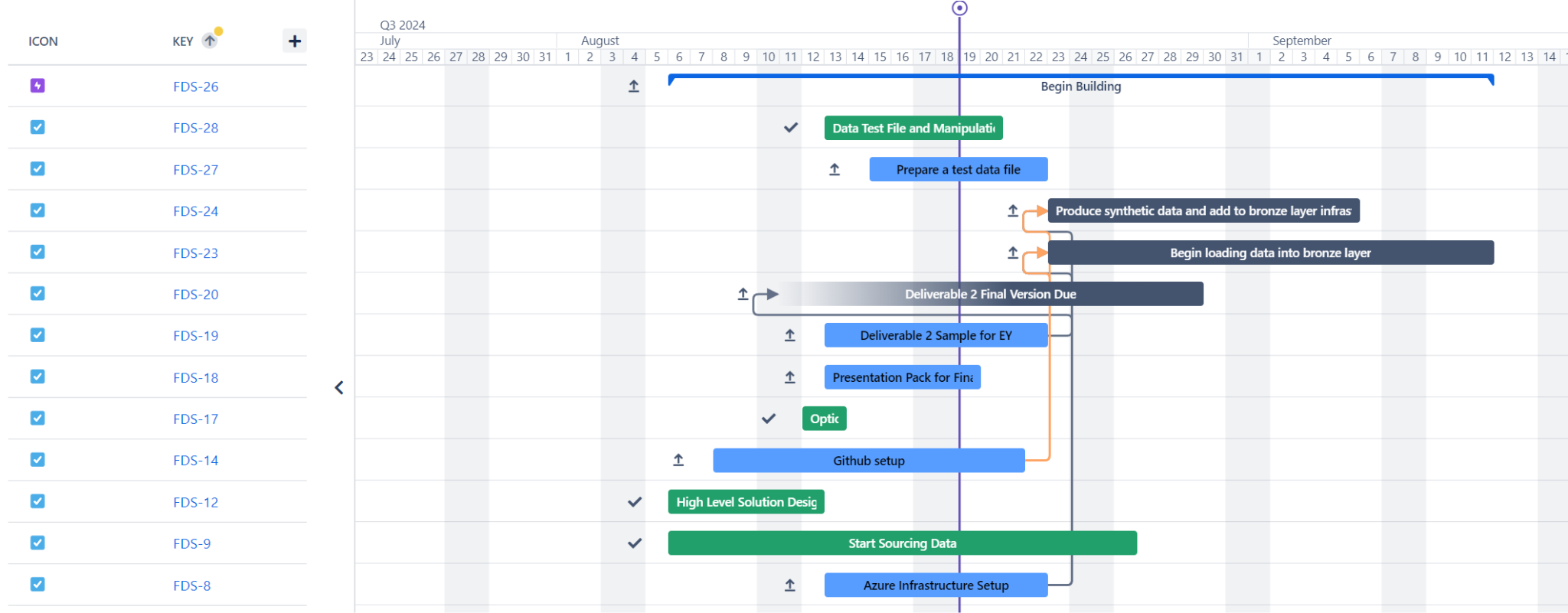
Gantt Chart combined with Team 13



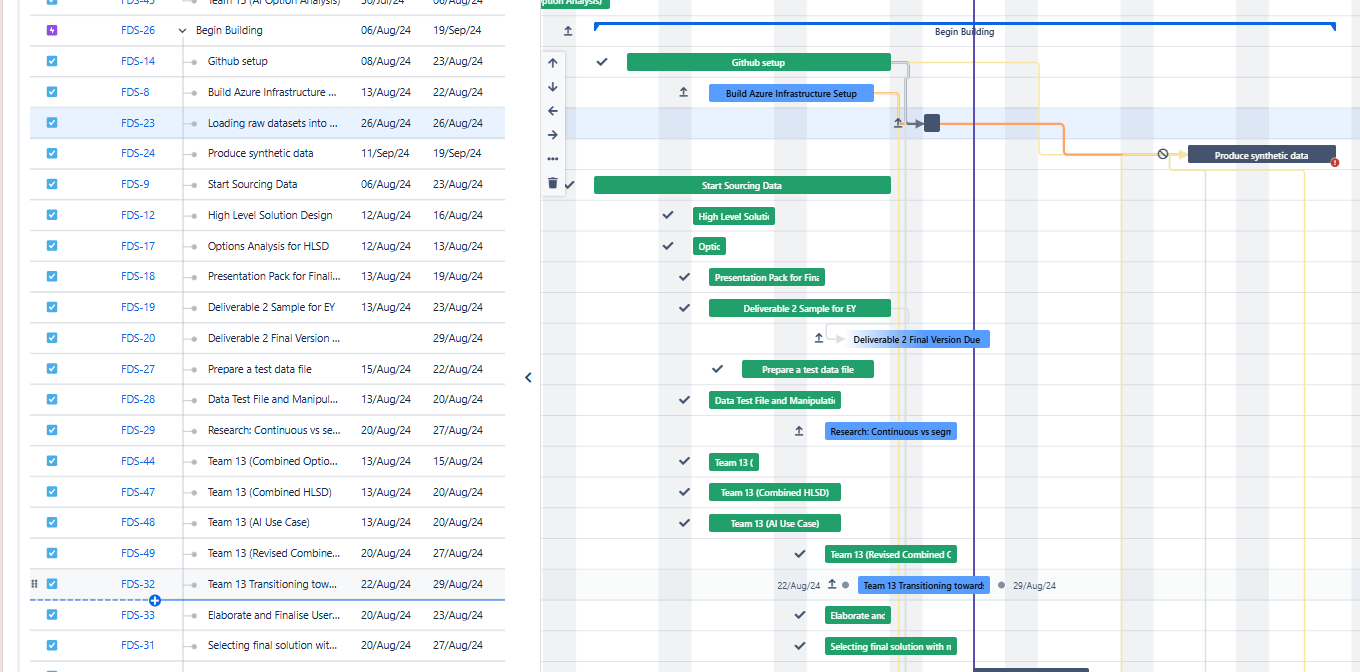
## 

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

Gantt Chart for Team 14



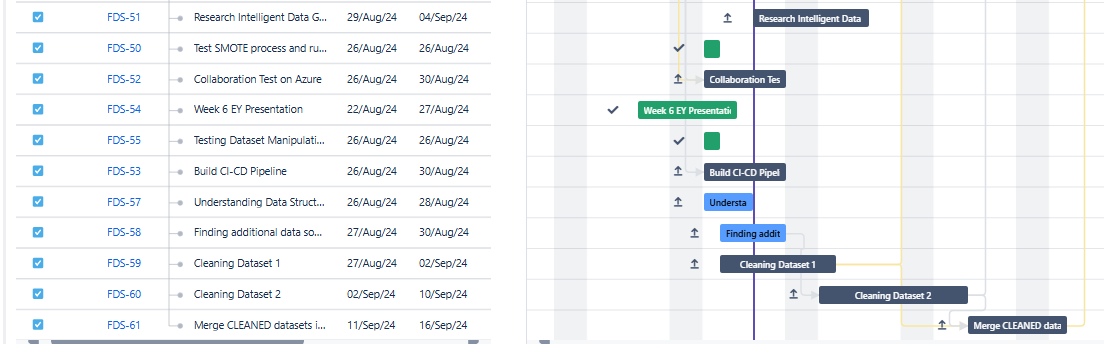
Gantt Chart combined with Team 13

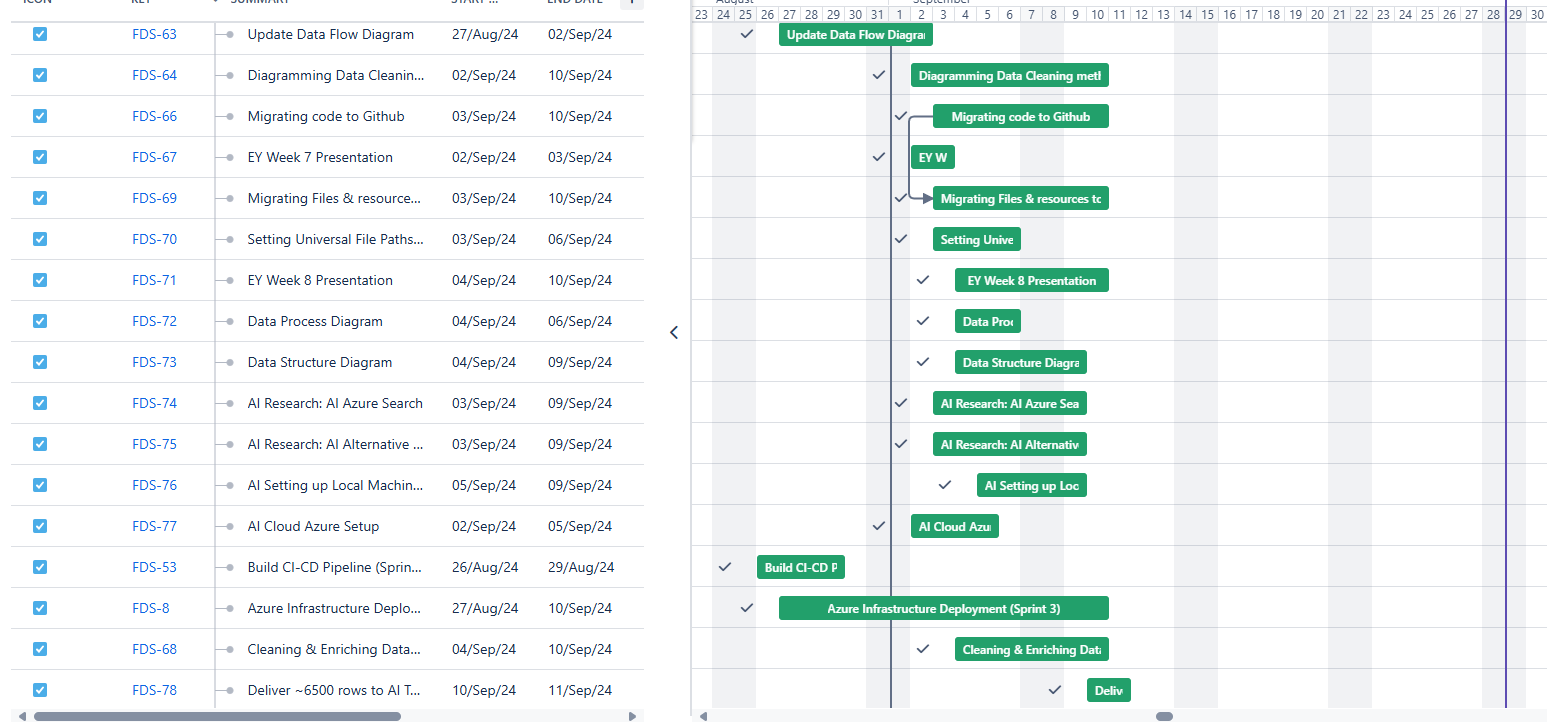


## 

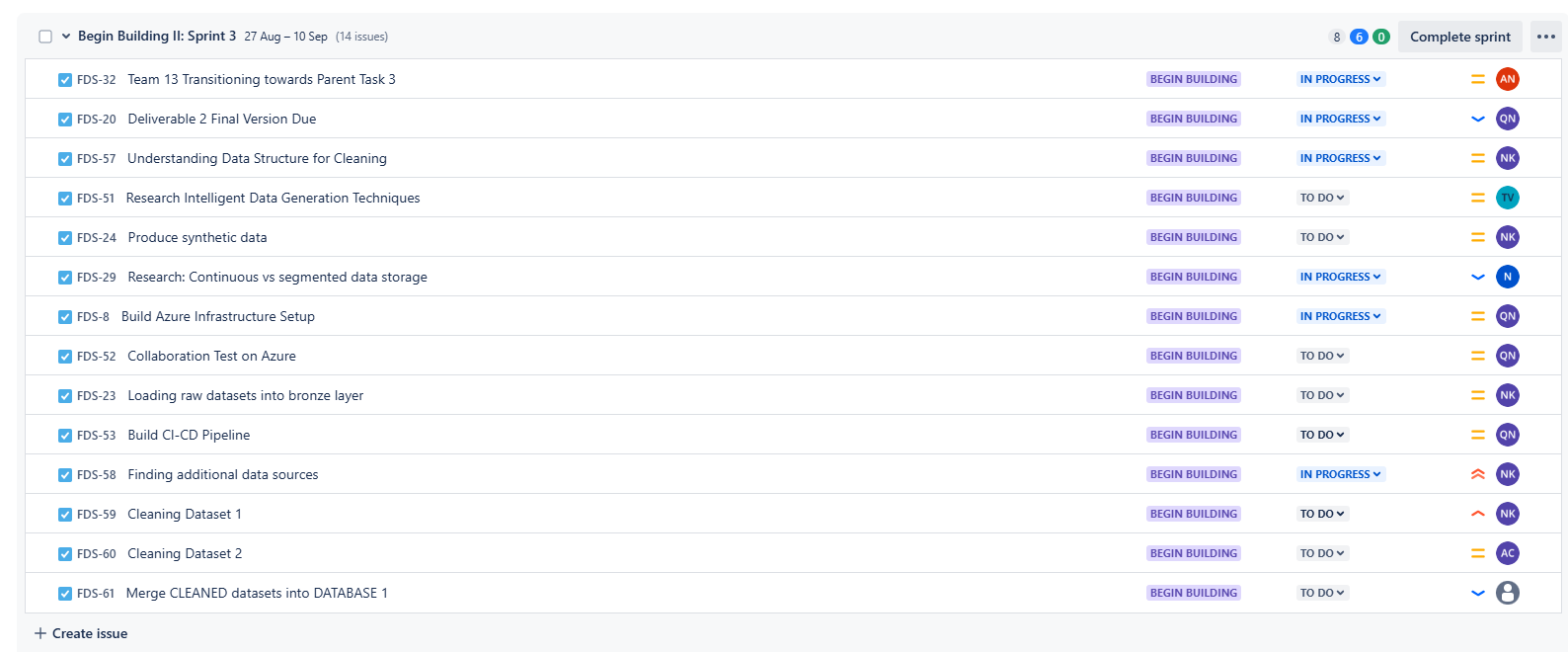
## **Appendix 7: Sprint 3 Gantt Chart**

Gantt Chart combined with Team 13



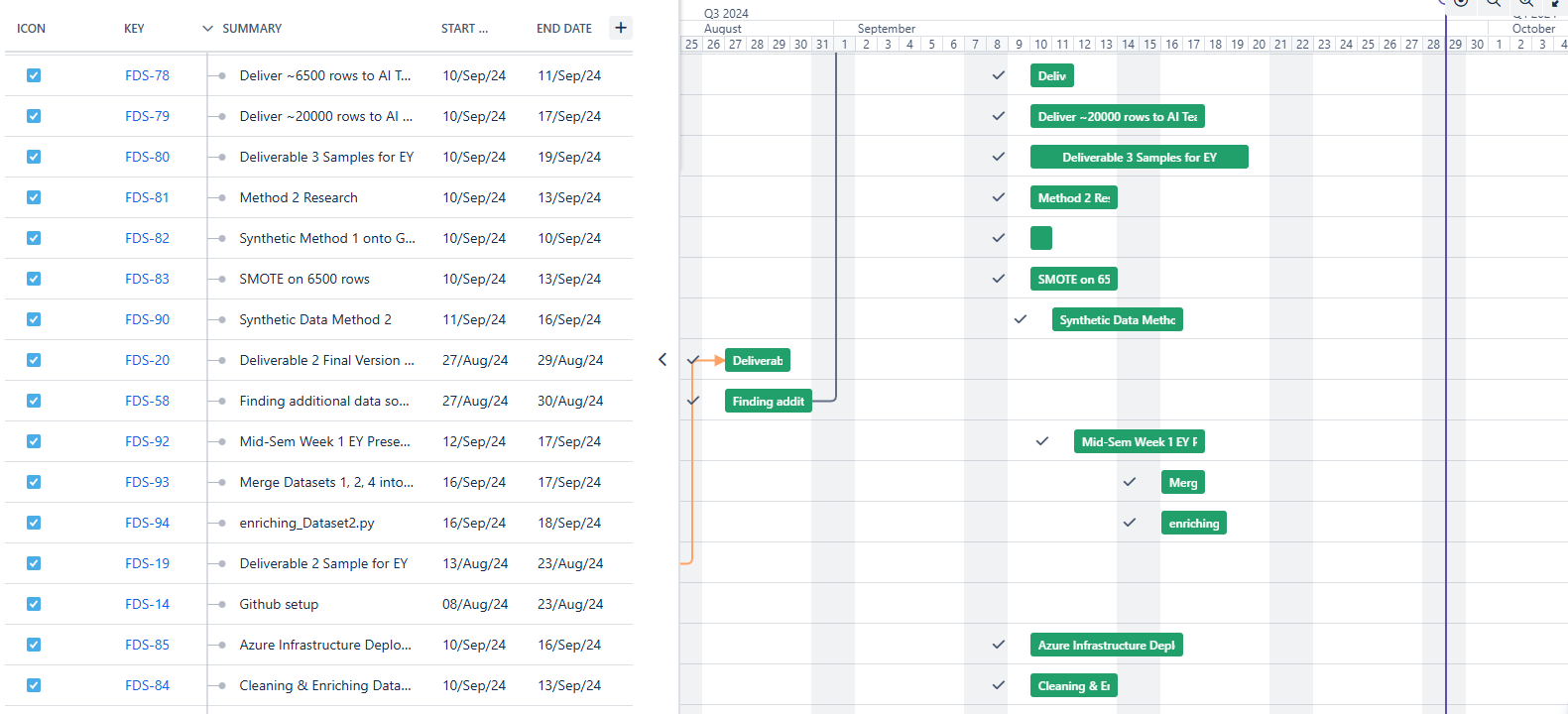


Sprint 3 Backlog



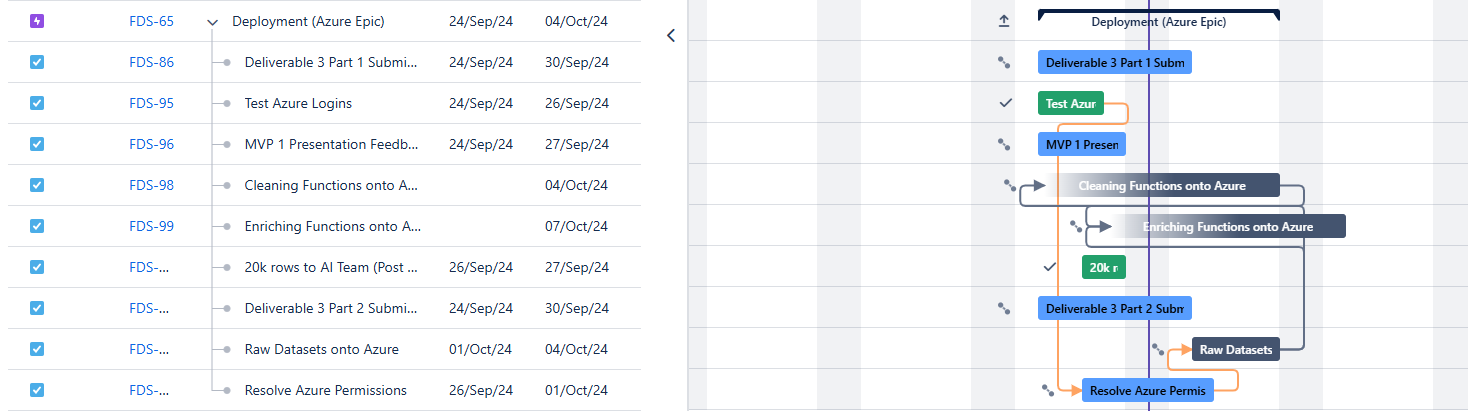
## **Appendix 8: Sprint 4 Gantt Chart**

Gantt Chart combined with Team 13



## **Appendix 9: Sprint 5 Gantt Chart**

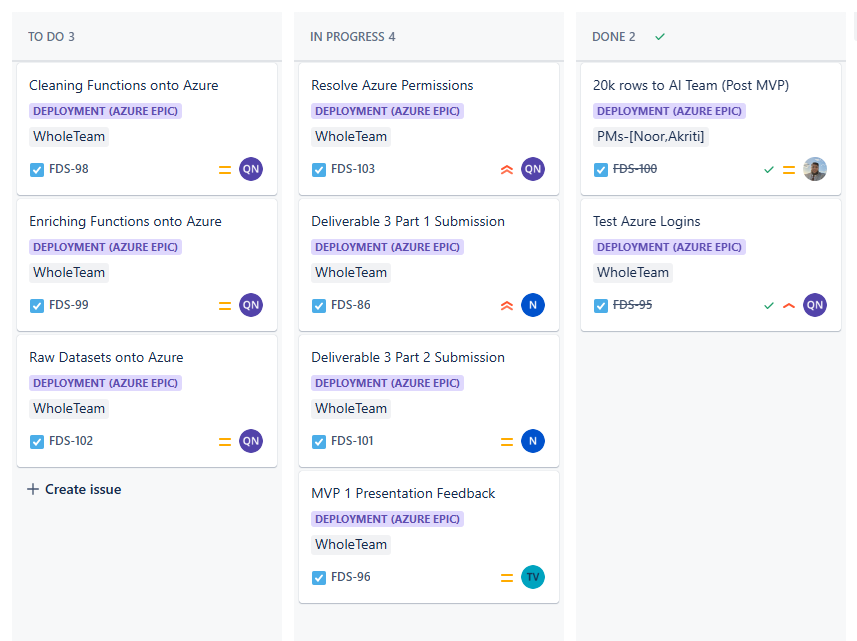
Gantt Chart combined with Team 13



Sprint 5 Backlog



Sprint 5 board



## **Appendix 10: 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, Noorullah, Ninuri and Tash | Mapping dataset structure, pseudocode for cleaning, designing development pipeline |
| AI Team | Alan and Noorullah | Join with the AI team to help integrate the dataset with the AI model and assist in any technical issues that arise with the dataset and the AI model. |
| Whole Team | Ninuri, Tashiya, Alan, Noorullah, and Aasnayem(Adam) | Project-related tasks such as the Deliverables handing to the University. |

**Updated Requirements Document/ Scoping Document**

### **Updated Requirements / Scoping Document (0.25 marks Del 3 and 0.1 marks for Del 4)**

Add a revision history table at the start of your document to clarify all changes made from Deliverable 2. No new sections are required. Please review feedback to Deliverable 2 submission and consider any project changes or sponsor feedback and update the document accordingly.

# **Revision History Table**

| **Version** | **Last Update** | **Author(s)** | **Changes** |
| --- | --- | --- | --- |
| V2\_S1  Data Understanding | 28/09/24 | Noorullah | * For each column mentioned data type and column description * Mentioned the size of raw data file and how the different fields are split in the raw data * Incorporated mentor feedback by editing the way column descriptions are presented to be clearer to read |
| V2\_S2  Data Preparations | 28/09/24 | Adam | * Included the steps taken to handle outliers * Incorporated the various data transformation techniques * Discussed how the data was integrated with examples |
| V2\_S3  Modelling | 28/09/24 | Ninuri and Tashiya |  |
| V2\_S4  Evaluation | 28/09/24 | Adam |  |
| V2\_S5  Deployment | 28/09/24 | Noorullah |  |

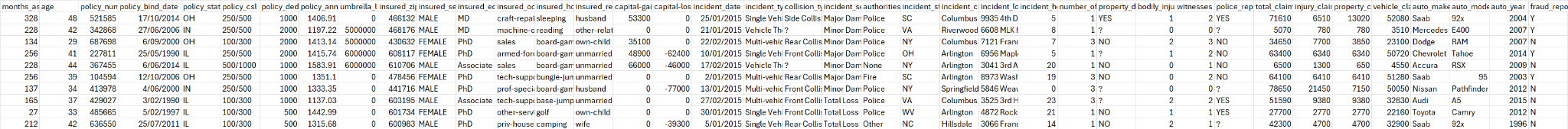
# **1. Data Understanding**

## **1.1 Initial Data Sources**

* Primary Dataset:
  + **Data Source:** 3 Kaggle Datasets. Dataset 1 with 1000 rows, Dataset 2 with 10000 rows (only 1575 relevant to automobile insurance), Dataset 3 with 10303 rows.
  + Content: Datasets include 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.

Detailed column Descriptions of each dataset:

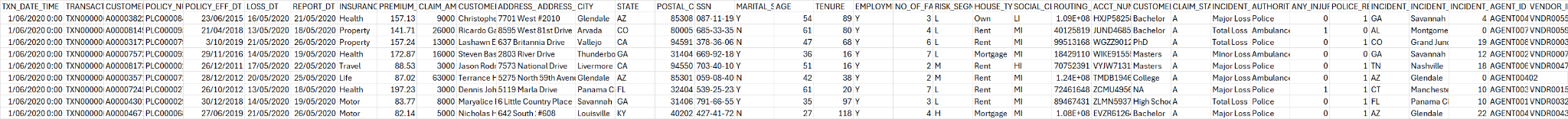
### **Raw Dataset 1: 1000 rows ~ 261 KB**



| Column Name | Data Type | Description |
| --- | --- | --- |
| months\_as\_customer | Numerical | Number of months the individual has been a customer |
| age | Numerical | Age of the customer |
| policy\_number | Categorical | Unique identifier for the insurance policy |
| policy\_bind\_date | Date | The date when the policy was created/bound |
| policy\_state | Categorical | The state where the policy was issued |
| policy\_csl | Categorical | Combined Single Limit for the policy |
| policy\_deductable | Numerical | The amount of deductible in the policy |
| policy\_annual\_premium | Numerical | The annual premium paid for the insurance policy |
| umbrella\_limit | Numerical | The umbrella limit in the policy |
| insured\_zip | Categorical | Zip code of the insured person |
| insured\_sex | Categorical | Gender of the insured person |
| insured\_education\_level | Categorical | Education level of the insured person |
| insured\_occupation | Categorical | Occupation of the insured person |
| insured\_hobbies | Categorical | Hobbies of the insured person |
| insured\_relationship | Categorical | Relationship status of the insured person |
| capital-gains | Numerical | Capital gains of the insured person |
| capital-loss | Numerical | Capital loss of the insured person |
| incident\_date | Date | Date of the incident |
| incident\_type | Categorical | Type of incident |
| collision\_type | Categorical | Type of collision |
| incident\_severity | Categorical | Severity of the incident |
| authorities\_contacted | Categorical | Whether authorities were contacted |
| incident\_state | Categorical | State where the incident occurred |
| incident\_city | Categorical | City where the incident occurred |
| incident\_location | Categorical | Specific location of the incident |
| incident\_hour\_of\_the\_day | Numerical | Hour of the day when the incident occurred |
| number\_of\_vehicles\_involved | Numerical | Number of vehicles involved in the incident |
| property\_damage | Categorical | Whether property damage occurred |
| bodily\_injuries | Numerical | Number of bodily injuries |
| witnesses | Numerical | Number of witnesses to the incident |
| police\_report\_available | Boolean | Whether a police report is available |
| total\_claim\_amount | Numerical | The total amount claimed |
| injury\_claim | Numerical | Claim amount for injuries |
| property\_claim | Numerical | Claim amount for property damage |
| vehicle\_claim | Numerical | Claim amount for vehicle damage |
| auto\_make | Categorical | Make of the vehicle involved |
| auto\_model | Categorical | Model of the vehicle involved |
| auto\_year | Numerical | Year of the vehicle involved |
| fraud\_reported | Boolean | Whether fraud was reported |

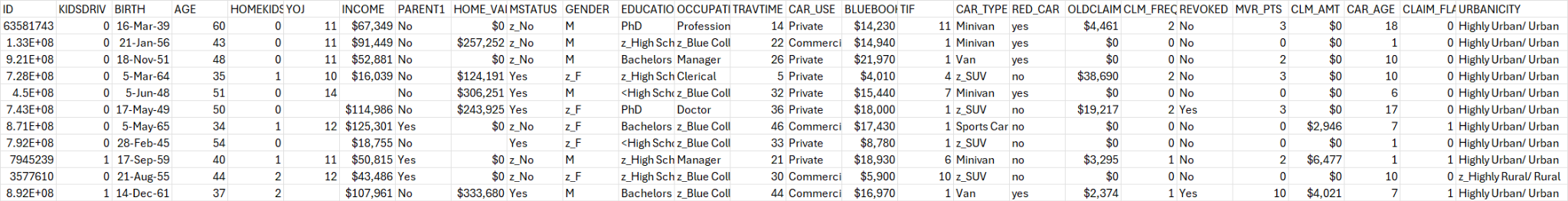
### 

### **Raw Dataset 2: 10000 rows ~ 2870 KB**



| Column Name | Data Type | Description |
| --- | --- | --- |
| TXN\_DATE\_TIME | Date/Time | The date and time of the transaction |
| TRANSACTION\_ID | Categorical | Unique identifier for the transaction |
| CUSTOMER\_ID | Categorical | Unique identifier for the customer |
| POLICY\_NUMBER | Categorical | Unique identifier for the policy |
| POLICY\_EFF\_DT | Date | The policy effective date |
| LOSS\_DT | Date | Date when the loss occurred |
| REPORT\_DT | Date | Date when the incident was reported |
| INSURANCE\_TYPE | Categorical | Type of insurance |
| PREMIUM\_AMOUNT | Numerical | Premium amount for the insurance policy |
| CLAIM\_AMOUNT | Numerical | Amount claimed |
| CUSTOMER\_NAME | Categorical | Name of the customer |
| ADDRESS\_LINE1 | Categorical | Address line 1 of the customer |
| ADDRESS\_LINE2 | Categorical | Address line 2 of the customer |
| CITY | Categorical | City of the customer |
| STATE | Categorical | State of the customer |
| POSTAL\_CODE | Categorical | Postal code of the customer |
| SSN | Categorical | Social Security Number |
| MARITAL\_STATUS | Categorical | Marital status of the customer |
| AGE | Numerical | Age of the customer |
| TENURE | Numerical | Number of years the customer has been with the insurance company |
| EMPLOYMENT\_STATUS | Categorical | Employment status of the customer |
| NO\_OF\_FAMILY\_MEMBERS | Numerical | Number of family members |
| RISK\_SEGMENTATION | Categorical | Risk category assigned to the customer |
| HOUSE\_TYPE | Categorical | Type of house the customer owns or rents |
| SOCIAL\_CLASS | Categorical | Social class of the customer |
| ROUTING\_NUMBER | Categorical | Customer’s bank routing number |
| ACCT\_NUMBER | Categorical | Customer’s bank account number |
| CUSTOMER\_EDUCATION\_LEVEL | Categorical | Education level of the customer |
| CLAIM\_STATUS | Categorical | Status of the claim |
| INCIDENT\_SEVERITY | Categorical | Severity of the incident |
| AUTHORITY\_CONTACTED | Boolean | Whether the authorities were contacted |
| ANY\_INJURY | Boolean | Whether any injury was reported |
| POLICE\_REPORT\_AVAILABLE | Boolean | Whether a police report is available |
| INCIDENT\_STATE | Categorical | State where the incident occurred |
| INCIDENT\_CITY | Categorical | City where the incident occurred |
| INCIDENT\_HOUR\_OF\_THE\_DAY | Numerical | Hour of the day when the incident occurred |
| AGENT\_ID | Categorical | Unique identifier for the insurance agent |
| VENDOR\_ID | Categorical | Unique identifier for the vendor |

### **Raw Dataset 3: 10000 rows ~ 1544 KB**



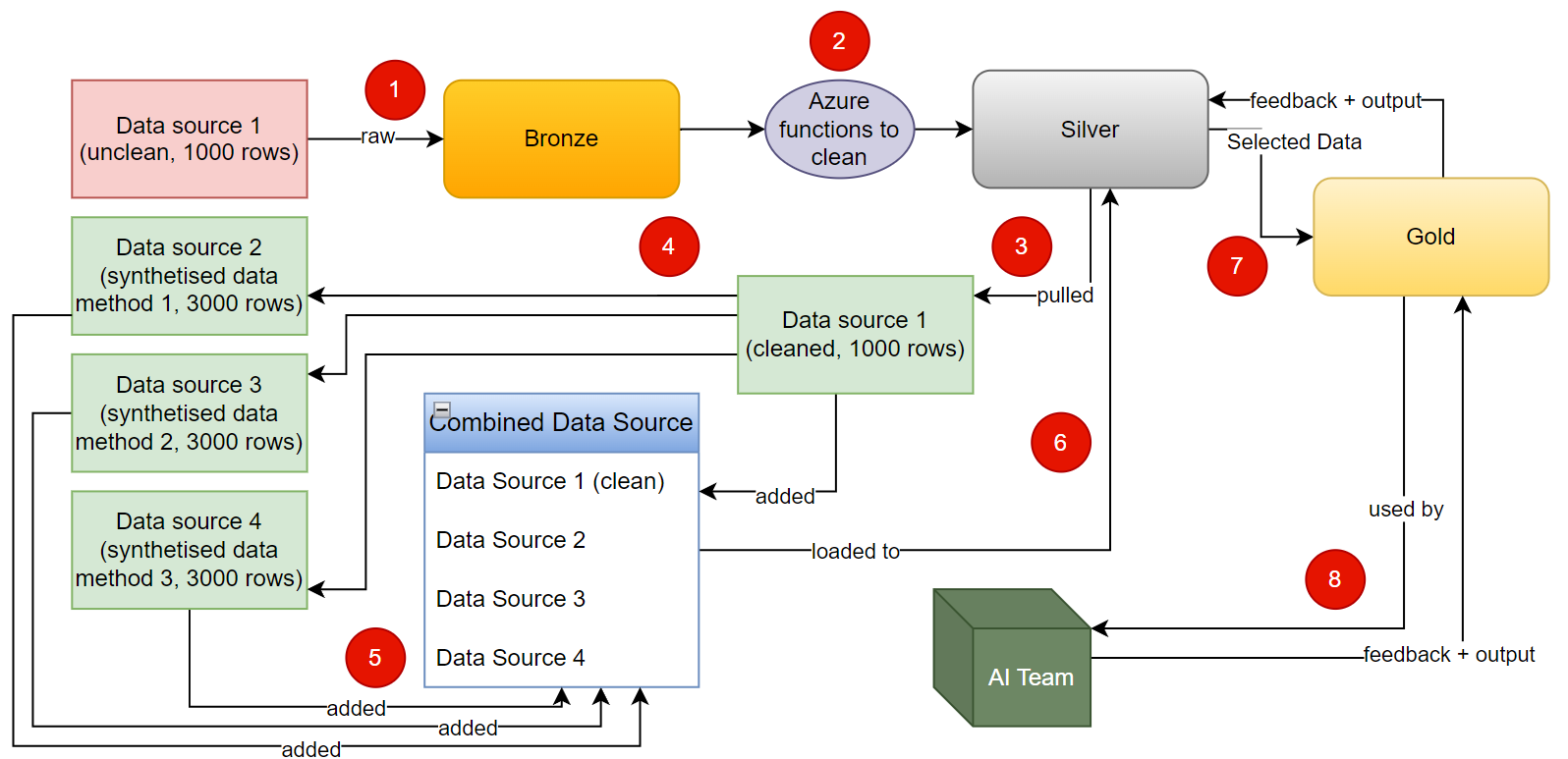
| Column Name | Data Type | Description |
| --- | --- | --- |
| ID | Categorical | Unique identifier for the customer |
| KIDSDRIV | Numerical | Number of kids driving in the household |
| BIRTH | Date | Birth date of the customer |
| AGE | Numerical | Age of the customer |
| HOMEKIDS | Numerical | Number of kids living in the household |
| YOJ | Numerical | Years on the job |
| INCOME | Numerical (Currency) | Income of the customer |
| PARENT1 | Categorical | Whether the customer is a single parent |
| HOME\_VAL | Numerical (Currency) | Value of the home |
| MSTATUS | Categorical | Marital status |
| GENDER | Categorical | Gender of the customer |
| EDUCATION | Categorical | Education level of the customer |
| OCCUPATION | Categorical | Occupation of the customer |
| TRAVTIME | Numerical | Travel time to work |
| CAR\_USE | Categorical | Whether the car is used for private or commercial purposes |
| BLUEBOOK | Numerical (Currency) | Value of the car |
| TIF | Numerical | Time in force of the insurance policy |
| CAR\_TYPE | Categorical | Type of car |
| RED\_CAR | Boolean | Whether the car is red |
| OLDCLAIM | Numerical (Currency) | Amount claimed in prior incidents |
| CLM\_FREQ | Numerical | Frequency of claims |
| REVOKED | Boolean | Whether the driver's licence has been revoked |
| MVR\_PTS | Numerical | Motor Vehicle Record points |
| CLM\_AMT | Numerical (Currency) | Amount claimed in the current incident |
| CAR\_AGE | Numerical | Age of the car |
| CLAIM\_FLAG | Boolean | Whether a claim was filed |
| URBANICITY | Categorical | Urban/rural classification |

* + **Widespread Availability:** It has been discovered that dataset 1 is identical across multiple sources, websites, and forums, suggesting that it is widely used and well-vetted (explained more in 1.3).
  + Despite dataset 1 being a good starting point and it seeming reliable due to multiple independent sources pointing to it, it alone is not enough for the AI team’s model. More data is needed (other datasets). Dataset 1 also needs to still 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, 8,000 additional rows will be generated using two different data synthesis methods (Remapping + Mathematical Sampling and CTGAN (Generative Adversarial Networks, a Neural Network based method). This will then be combined to give a dataset with a total of about 20,000 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**

* Original Dataset:
  + **Quality:** The original datasets, sourced from Kaggle and other widely referenced platforms, are well-vetted and frequently cited in various studies, reports, and articles. These datasets share many similarities across different platforms, often pointing back to the same original Kaggle source. For instance, the first dataset contains 1000 records with 39 diverse columns covering multiple categories of insurance-related data. Dataset 2 and Dataset 3 further add to the breadth and depth of data, with additional rows and varying fields. The large number of records and diverse data types help support data quality and robustness.
  + **Structure:** These datasets have a generally standardised structure, which simplifies their integration. However, care must be taken to address the nuances and differences between the datasets to ensure that all critical elements, such as those necessary for fraud detection, are accurately captured. For example, Dataset 3, being messier and less standardised, required more rigorous cleaning.
  + **Completeness:** While the original datasets are comprehensive in capturing their intended data points, they may not account for all specific scenarios or edge cases relevant to our project. The synthetic data generation methods applied further enhance the completeness of the datasets, providing additional rows to help cover such edge cases and better balance the data.
* Synthetic Data:
  + **Quality:** The synthetic data, generated from cleaned and enriched source datasets, is expected to be of high quality, as it draws directly from the original data's patterns and relationships. The data has been validated to ensure it introduces meaningful variance while maintaining consistency with key attributes, particularly the fraud column, which remains reliable across both synthetic methods.
  + **Structure:** The synthetic data mirrors the structure of the original cleaned and enriched datasets, ensuring that it integrates smoothly with Dataset 1, Dataset 2, and Dataset 3. This consistent structure ensures that the synthetic rows are compatible with the original datasets, helping to create a comprehensive dataset for AI model training.
  + **Fraud Distribution:** A key part of the evaluation process involved ensuring that the synthetic data accurately reflects the distribution of fraud cases. Both synthetic data generation methods produced a similar number of fraudulent records (around 400 in each method), which aligns with the distribution in the original data. This consistency reinforces the legitimacy of the synthetic data and ensures that fraudulent cases are well represented, providing a robust base for model training.
  + **Imbalances and Biases:** Since fraudulent claims are naturally underrepresented in the data, maintaining a balanced representation in the synthetic data was critical. While no oversampling methods like SMOTE were used, the synthetic data successfully preserved the fraud ratios, ensuring that machine learning models trained on this data will learn to detect fraudulent cases without skewed bias.

## **1.4 Mechanisms for Evaluation**

* Data Quality Assessment:
  + **Initial Dataset:** Perform a detailed exploratory data analysis on the initial datasets 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 datasets 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. After further discussion with the mentors and the clients, we decided to go ahead with 3 datasets instead of the sole dataset. Dataset 1 contains 1000 rows, Dataset 2 contains 1500 rows and Dataset 3 contains around 10800 rows.
* **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. Outliers can make the model perform poorly. Outliers can be detected by performing statistical tests like the z-score test. Using the results from the test, we can either remove the outliers which are extreme or transform them if they are not too extreme.
* **Data Transformation:** We used various data transformation techniques while normalising our dataset. We standardised our data so that our data is consistently scaled. An example of this is the gender column in dataset 3. The column had ‘M’ and ‘z\_F’ as the two possible outputs. So, we converted ‘z\_F’ to ‘F’ to match with the other datasets.
* **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. After discussing with the mentors and the AI team, we decided to go ahead with 2 different synthesisation methods. Each method will produce 4000 rows of synthetic data. The first method is called Remapping and Mathematical Sampling where data is remapped and sampled to generate rows. The second method is called Conditional Tabular Generative Adversarial Network (CTGAN) where relationships and patterns are learnt to produce high quality, synthetic data.
* **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. We have actually integrated 3 datasets and 2 synthesisation methods to combine it to one complete dataset with 20,900 rows approximately. The 3 datasets that are cleaned and enriched are merged first. Then, the synthesised methods are combined with the 3 datasets. This is done using the ‘pd.concat’ function in the pandas library. We also make sure that we have an index which is continuous starting from 1 to the end.
* **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 two different methods. Each method will synthesise 4000 rows of data. The methods are Remapping and Random Sampling, CTGAN(Conditional Tabular Generative Adversarial Networks). 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. 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. Thus, the AI team will be using machine learning models like neural networks to classify the data. Therefore, while AI modelling does not fall within the scope of our project with EY, below we will be discussing some aspects of machine learning modelling that will be taking place in relation to cleaning the dataset.

## **3.1. Modelling Techniques**

Filling missing values is an essential part of preparing the dataset for subsequent analysis and model training. Various machine learning techniques including Random Forest Classifier, Random Forest Regressor, Support Vector Regression(SVR) and K-Nearest Neighbours(KNN), Can be utilised to estimate and replace these missing values by analysing the patterns present in the available data. These approaches tend to yield more precise imputations than conventional mean,median mode imputation as they take into account the intricate relationships among different features.

### **3.1.1. Random Forest Classifier:**

* **Reason it was Chosen:** it is good for imputing missing values in categorical variables through predicting classes based on patterns in the dataset.
* **Intended Outcome:** Impute categorical values
* **Training procedure:** the dataset will be split into rows with filled categorical values and rows with missing categorical values. The model will then be trained with the dataset with full categorical values and then predict the rows with missing values with the most likely option based on patterns observed in the dataset.
* **Imputation Process:** The most frequent class will be imputed into missing categorical rows

Benefits of Random Forest Classifier:

* Robust for categorical data imputation
* Can handle large datasets with missing categorical data
* Can handle high-dimensional data
* Can manage complex, non-linear relationships

Limitations of Random Forest Classifier:

* Requires sufficient training data with complete values
* Computationally expensive for large datasets

### **3.1.2. Random Forest Regressor:**

* **Reason it was Chosen:** To help estimate missing values in numerical variables. It builds multiple decision trees and combines their outputs to generate predictions, similar to the classification method.
* **Intended Outcome:** Impute numerical values
* **Training procedure:** This regressor is trained using rows with complete numerical data, allowing it to learn the relationships between the target numerical variable and other features. It leverages ensemble learning techniques to forecast missing values.
* **Imputation Process:** Missing numerical values are replaced with the predicted outputs from the regressor, which takes into account the effects of all other features in the dataset.

Benefits of Random Forest Regressor:

* Handles continuous and discrete variables
* Effective in detecting complex data patterns
* Non parametric
* Does not assume in particular distribution of the data

Limitations of Random Forest Regressor:

* It is resource intensive as it requires large computational power and memory.

### **3.1.3. Support Vector Regression (SVR):**

* **Reason it was Chosen:** Highly effective in imputing missing numerical values by maximising the margin of error around the predicted values to identify the best fit.
* **Intended Outcome:** Impute numerical values
* **Training procedure:** SVR is trained on the available data, using methods like the radial basis function(RBF) or polynomial kernels to account for non-linear relationships. The model minimises the difference between actual and predicted values within a set margin, ensuring a good fit.
* **Imputation Process:** When dealing with missing continuous values, the SVR model predicts the most likely value for each gap by analysing the relationships between the existing data points and features.

Benefits of Neural Networks

* SVR works well in situations with non linear interactions and high dimensional data which works great with simpler methods struggle,
* Flexibility allows it to handle more complex data structures.

Limitations of Neural Network

* SVR can be sensitive to extreme values and may require substantial computing power for larger datasets.
* Fine tuning parameters like the type of kernel and regularisation is necessary for optimal performance.

### **3.1.4. K-Nearest Neighbours (KNN):**

* **Reason it was Chosen:** KNN is a simple yet effective method for filling in missing data, suitable for both numerical and categorical values. It works by finding by finding
* **Intended Outcome:** Impute both categorical and numerical values.
* **Training procedure:** KNN does not require formal training. Instead, it calculates the similarity between data points with gaps and those with complete information. For numerical data, it averages the nearest neighbours, and for categorical data, it chooses the most frequent category.
* **Imputation Process:** For each Missing entry, KNN identifies the closest complete data points and fills in the gaps by considering the values from those neighbouring records.

Benefits of K-Nearest Neighbours(KNN):

* Straightforward to apply
* Works for various types of data without needing assumptions about distribution
* Adaptable to different datasets and types of variables

Limitations of K-Nearest Neighbours(KNN):

* Time consuming when it comes to large datasets, as the algorithm compares each point to others
* Deciding the right number of neighbours(K) and similarity measure is crucial for getting accurate results.

### **3.1.5. 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:

### **Random Forest Classifier:** Able to handle large datasets and high dimensional data. Is also able to handle complex, non linear relationships.

### **Random Forest Regressor:** For imputing discrete and continuous variables by detecting complex data patterns.

### **Support Vector Regression (SVR):** Is able to work great with nonlinear interactions and high dimensional data. Can handle complex data structures.

### **K-Nearest Neighbours (KNN):** Straightforward approach that can adapt to different datasets and different types of variables which offers flexibility.

## **3.2. Evaluation Metrics**

To gauge the performance of the imputation techniques, the following metrics will be applied:

* **Mean Squared Error(MSE):** This is used for numerical data, MSE calculates the average of the squared differences between the actual values and the imputed ones, which helps to assess the accuracy of the imputations,
* **Mean Absolute Error (MAE):** This is used for numerical imputations, MAE provides a more intuitive measure of error by averaging the absolute differences between the actual and imputed values.
* **Accuracy:**This metric evaluates how often the imputation model correctly predicts missing categorical values, offering a simple way to measure success for categorical data.
* **F1-Score:** Particularly useful for categorical variables, the F1- score balances precision and recall, ensuring the imputation method works well even with imbalance data.
* **Cross-Validation:** To test the consistency of the imputation models, techniques like K-fold cross validation will be applied. This helps ensure the models performance holds steady when applied to different parts of the dataset.

## **3.3. Hyperparameter-tuning**

To enhance model performance and boost the accuracy of imputations, fine tuning of hyperparameters will be conducted in the following areas:

* **Random Forest Classifier/ Regressor:** Critical parameters to adjust include the number of trees, tree depth, the minimum number of samples required to split a node and the number of features used for splits.
* **SVR:** Key aspects to optimise include selecting the appropriate kernel(e.g. RBF or polynomial) , setting the regularisation strength (C) and determining the epsilon value, which controls the error margin.
* **KNN:** The primary factors to refine are the numbers of neighbours (K) and the type of distance metric to ensure better imputations.

To find the optimal hyperparameter settings for each model, Grid search or random search methods will be utilised.

* **GridsearchCV -** creates a grid of all the possible hyperparameter combinations and selects the combination with the nest performance by evaluating each combination through cross-validation.
* **RandomSearchCV** - randomly selects a predefined number of combinations from the distribution and the best combination is selected through cross validation

## **3.4. Bias and Fairness**

When handling missing values, it’s important to be mindful of potential biases that may arise, especially if the missing data is not randomly distributed (eg. certain demographic groups may be more missing entries). To address this:

* **Bias Detection:** Regularly check the imputation results across various groups to ensure that no group is unfairly impacted.
* **Bias Mitigation:** Adjust the imputation models or apply reweighting techniques to prevent skewed outcomes that could disproportionately affect specific groups and lead to biassed conclusions.

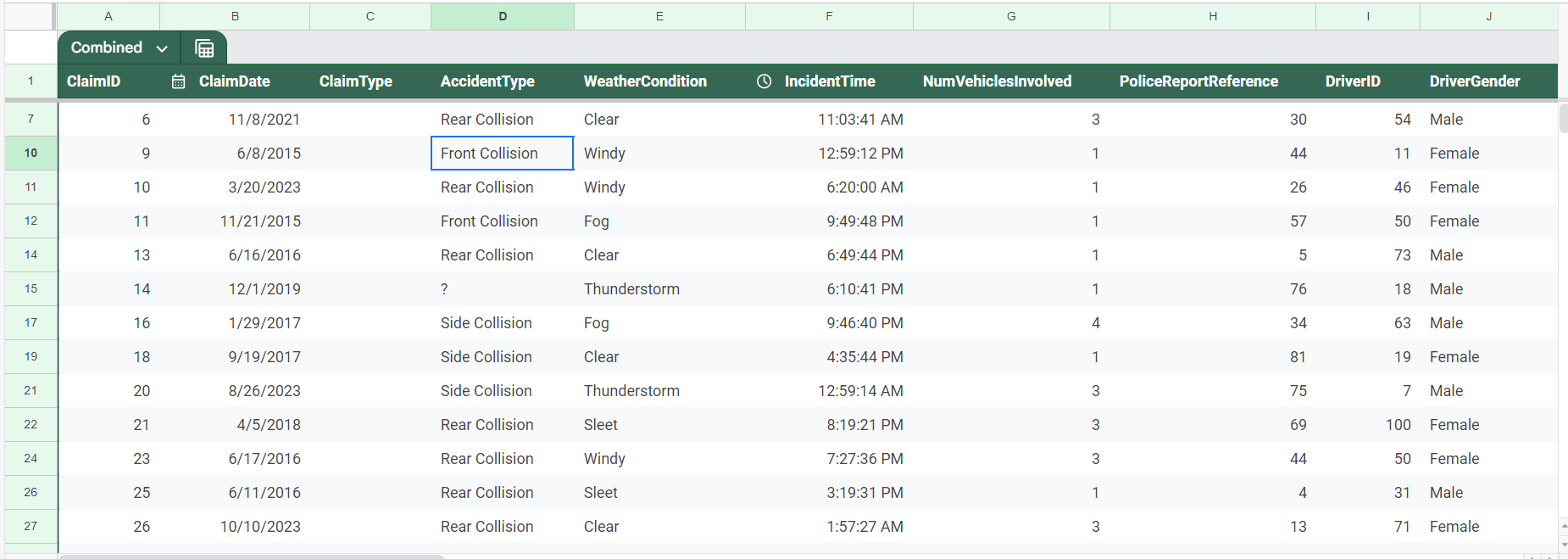
## **3.5. Test Dataset**

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,

However, with further discussion and testing with the AI team have found that there are a few more requirements:

* **Missing values:** there must be no missing values as the model does not train properly if there are any.
* **Column names:** the names of columns have to be uniform across all datasets including names and upper/lower casing. If not, the code has to be changed to reflect each different column name.
* **Numerical values:** the given data should have no categorical values and must be transformed to numerical values.



***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. Evaluation Techniques**

We can add additional evaluation techniques to ensure that the performance of the model is good. We can do this by using cross validation techniques which will split the training and the testing data. One specific example of this can be the K-Fold Cross Validation which is a common technique used to split the data. In this process, the dataset is divided into K equal parts where K is the number of equal parts or folds that the dataset is divided into. K can be chosen by us and it determines the number of times the model is trained and tested. After all the training and testing is done, the overall result is averaged to give a more accurate result for the model. Another cross validation technique that can be used is the Stratified K-Fold Cross Validation technique where the concern of imbalanced datasets can be addressed. For example, if the amount of fraudulent cases is insufficient in our dataset, we can use Stratified K-Fold Cross Validation to ensure that each fold has a similar distribution and balance the ‘Fraud’ column. These cross validation methods ensure that the model is reliable and the model is not dependent on a singular train/test split. Multiple splits will help improve the accuracy of the model.

## **4.3. 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.

## **4.4. 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.

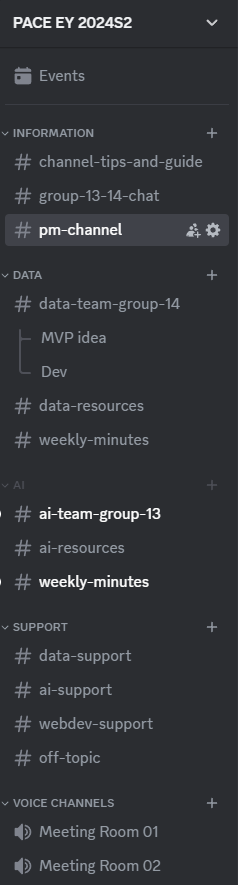
## **2.4. Respect**

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**]

## **4.4 Documentation of Resolutions and Agreed Changes:**

In order to maintain transparency and foster team unity in the run, it's crucial to record any solutions reached when dealing with conflicts and agreed upon adjustments between parties involved. This approach can anticipate disagreements in the future, align everyone's understanding, and establish clear directions for making decisions later on.

4.4.1 **Recent Resolutions and Agreed-Upon Changes**

1. **Resolution: Timeline Misalignment for Deliverables**
   1. **Conflict:** A miscommunication occurred regarding the deadlines for submitting the deliverable draft to clients. Some team members thought they had more time than others, leading to delays.
   2. **Resolution:** The team agreed to implement a new, centralised timeline in Jira, ensuring that all deliverables and deadlines are visible to every member in real-time.
   3. **Agreed Change**: A calendar reminder will be sent to all team members at the start of each week, detailing upcoming tasks and deadlines to avoid further misalignments.
2. **Resolution: Communication Delays During Remote Collaboration (Working around everyone's busy schedule)**
   1. **Conflict:** We were finding it difficult to find times for meetings to practise MVP or team discussions due to everyone's different schedules.
   2. **Resolution**: The team members each provided a schedule of their working hours, so others know when they will be non responsive. We also have allocated two specific times each week for a compulsory meeting where all team members have to attend which will be Saturday 11am and Tuesday 9am, where all team members will be free for the continuation of the project. This allowed us to know we could have our discussions, practises, and collaboration during this time.
   3. Agreed Change: Change to have set times and days for meetings where everyone is available. Also Team members will now set their availability status on Discord, ensuring everyone is aware of their availability for urgent matters. The weekly meeting also is used to identify and address any potential communication bottlenecks.

# **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 Mahagoda | * 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 the 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. |
| V3\_S4 | 28.09.24 | Tashiya Vilathgamuwa | * Added Documentation of resolutions and agreed changes to change management section |

# 

# **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