### **Epic 1: Technical User Stories**

User Story 1: Setup Azure Environment

**As an** AI engineer, **I want to** have a secure and scalable environment in Azure, **So that** I can host and manage the data and training models efficiently.

Acceptance Criteria:

**Given** the project requirements, **When** the Azure environment is set up, **Then** it should include necessary services like Azure Blob Storage, Azure Databricks, and Azure SQL Database, with defined access controls and scalable resources.

**User Story 2: Anonymise Data**

**As an** AI engineer, **I want to** receive a database with anonymised sensitive data, **So that** the training data is compliant with privacy regulations and free of PII.

**Acceptance Criteria:**

**Given** the presence of PII in the dataset, **When** the dataset is cleaned, **Then** all sensitive information like names, addresses, and phone numbers should be removed or masked, with a log of actions maintained.

**User Story 3: Data Validation and Cleaning**

**As an** AI engineer, **I want to** receive a clean data set, **So that** I can ensure the quality and accuracy of the data used for training the AI model.

**Acceptance Criteria:**

**Given** the raw dataset, **When** data validation and cleaning scripts are executed, **Then** the resulting dataset should be free of missing, duplicate, or erroneous data, and stored in a validated location.

User Story 4: Data Size

**As an** AI engineer, **I want to** receive a dataset with a large amount of data, **So that** I can accurately train the AI model with enough data to reduce over- and under-fitting

Acceptance Criteria:

**Given** the need to accurately train the AI model, **When** the dataset is being created, **Then** the dataset provided is large enough to support robust model training ensuring that the AI model can learn and generalise effectively for fraud detection.

**User Story 5: Data Type**

**As an** AI Team Member, **I want to** receive a dataset that is significantly diverse

**So that** I can effectively train the AI model to detect fraudulent claims with high accuracy.

**Acceptance Criteria:**

**Given** the data sources are integrated into the system, **When** the data is processed and made available, **Then** the dataset provided is diverse enough to support robust model training, covering various scenarios and examples of claims, ensuring that the AI model can learn and generalise effectively for fraud detection.

#### User Story 6: Data Annotation and Labeling

**As an** AI Team Member **I want to** receive a dataset that is properly annotated and labeled. **So that** the AI model can be trained with clear examples of fraudulent and non-fraudulent claims, enhancing its ability to distinguish between them

Acceptance Criteria

**Given** the data is prepared for training **When** the dataset is delivered **Then** it should include accurate labels for different categories of claims (e.g., fraudulent, non-fraudulent) and be consistently annotated according to predefined guidelines.

#### User Story 7: Data Format and Usability

#### **As an** AI Team Member **I want to** receive the dataset in a format that is easily usable and compatible with AI tools and platforms **So that** I can efficiently integrate the data into the training pipeline without additional data transformation or conversion steps

Acceptance Criteria

**Given** the data is prepared for delivery **When** the dataset is provided **Then** it should be in a standardized format (e.g., CSV, JSON) that is compatible with the AI tools and platforms used, and include necessary metadata and documentation for easy integration.

#### 

#### User Story 8: Data Standardisation **Data Consistency and Integrity**

**As an** AI Team Member **I want to** receive a dataset that maintains consistency and integrity across different data sources **So that** the AI model can be trained on a cohesive and reliable dataset without discrepancies or conflicting information

**Acceptance Criteria**

**Given** the dataset is created from multiple sources **When** the data is reviewed **Then** it should be consistent in terms of format, structure, and content, with integrity checks performed to ensure data accuracy and reliability.

### User Story 9: Data Relevance

**As an** AI Team Member, **I want to** receive a dataset that is highly relevant to current fraud detection scenarios, **So that** the AI model can be trained with data that accurately reflects the latest fraud tactics and trends, enhancing its effectiveness in real-world applications

**Acceptance Criteria**

**Given** the dataset is prepared for training, **When** the data is reviewed for relevance

**Then** the dataset should include recent and relevant examples of fraud claims, accurately representing current patterns and tactics used in fraud detection, with data that reflects up-to-date practices and emerging trends.

User Story 10:

### **Epic 2: Business User Stories**

### **User Story 1: Access to Cleaned Data**

**As an** NRMA agent, **I want to** access the cleaned dataset, **So that** I can review the work done by the data team.

**So that i have confidence that the data will let me do …**

**Acceptance Criteria:**

**Given** the dataset is cleaned, **When** I access it, **Then** I should be able to view the cleaned data with all errors, duplicates, and missing values addressed.

### **User Story 2: Data Anonymisation Verification**

**As an** NRMA agent, **I want to** verify that all PII has been anonymised in the dataset, **So that** I can ensure compliance with privacy regulations before submission.

**Acceptance Criteria: Given** the anonymisation process, **When** I review the dataset,  
**Then** there should be no personally identifiable information (PII) present in any of the data fields.

### **User Story 3: Review of Data Relevance - move to AI team user stories**

**As an** NRMA agent, **I want to** assess the relevance of the dataset to the AI model’s objectives, **So that** I can confirm that the data aligns with the intended use case.

**Acceptance Criteria:**

**Given** the business objectives of the AI model, **When** I review the dataset, **Then** the dataset should contain data points that are directly relevant to the model’s objectives.

### **User Story 4: Data Integrity Check**

**As an** NRMA agent, **I want to** perform an integrity check on the dataset, **So that** I can ensure the data is accurate and reliable.

**Acceptance Criteria:**

**Given** the need for accurate data, **When** I check the dataset, **Then** there should be no errors, inconsistencies, or corrupt data.

### **User Story 5: Documentation Review**

**As an** NRMA agent, **I want to** review the dataset’s accompanying documentation, **So that** I can understand the data sources, preprocessing steps, and any assumptions made.

**Acceptance Criteria:**

**Given** the dataset and its documentation, **When** I review the documentation, **Then** it should clearly explain data sources, preprocessing steps, and any relevant metadata.

### **User Story 6: Verify Data Consistency**

**As an** NRMA agent, **I want to** ensure that the data is consistent across different entries, **So that** I can confirm the reliability of the dataset.

**Acceptance Criteria:**

**Given** the dataset, **When** I examine it, **Then** the data should be consistent in terms of format, structure, and content across all entries.

### **User Story 7: Review of Data Augmentation Techniques**

**As an** NRMA agent (data person in the team), **I want to** review any data augmentation techniques applied to the dataset,  
**So that** I can ensure that the dataset is properly enhanced without introducing bias.

**Acceptance Criteria:**

**Given** the augmented dataset, **When** I review it, **Then** I should be able to confirm that the augmentation techniques are appropriate and do not introduce significant bias.

### **User Story 8: Metadata Inspection**

**As an** NRMA agent, **I want to** inspect the metadata of the dataset, **So that** I can verify the origin, collection methods, and any preprocessing steps applied.

**Acceptance Criteria:**

**Given** the metadata provided, **When** I review it, **Then** it should include comprehensive details about data origin, collection methods, and preprocessing steps.

### **User Story 9: Check for Bias in Data**

**As an** NRMA agent, **I want to** check the dataset for any inherent bias, **So that** I can ensure that the AI model will produce fair and unbiased outcomes.

**Acceptance Criteria:**

**Given** the dataset, **When** I review it, **Then** there should be no significant bias that could lead to unfair outcomes in the AI model.

### **User Story 10: Evaluate Data Coverage**

**As an** NRMA agent, **I want to** evaluate the coverage of different scenarios in the dataset, **So that** I can ensure that the dataset is comprehensive and robust.

**Acceptance Criteria:**

**Given** the need for comprehensive data, **When** I review the dataset, **Then** it should cover a wide range of scenarios relevant to the model’s objectives.

### 

### **User Story 11: Final Approval of Dataset for AI Training**

**As an** NRMA agent, **I want to** give final approval of the dataset, **So that** it can be sent to the AI team for model training.

**Acceptance Criteria:**

**Given** that all reviews and checks are complete, **When** I finalise the dataset, **Then** I should be able to approve it for submission to the AI team, ensuring that it meets all required standards and criteria.

### **User Story 11: Final Approval of Dataset for AI Training**

**As an** NRMA agent (data analyst in the claims team), **I want to** give final approval of the dataset, **So that** it can be sent to the AI team for model training.

As an NRMA agent I want access to an approved and verified data set so that I can have confidence that claims detected as fraudulent are accurate.

**Acceptance Criteria:**

**Given** that all reviews and checks are complete, **When** I finalise the dataset, **Then** I should be able to approve it for submission to the AI team, ensuring that it meets all required standards and criteria.

### **User Story 12: Options Analysis**

**As an** NRMA agent, **I want to** receive an options analysis report, **So that** I can decide on the best way to undertake the project by comparing costs, time and relevance.

**Acceptance Criteria:**

**Given** that the NRMA agent oversees the project, **When** the data team creates the Options Analysis Report, **Then** it should contain various options on how the project will be complete and the recommended solution for comparison so the agent can complete the project in a cost-saving and fast way.

### **User Story 13: High Levels Solutions Design (HLSD)**

**As an** NRMA agent, **As an** NRMA Fraud Analyst, **I want to** receive a High-Level Solution Design (HLSD) from the data team, **So that** I can understand how the data infrastructure and processes align with the overall solution architecture for the fraud detection system.

**Acceptance Criteria:**

**Given** the data team has developed the HLSD, **When** the HLSD is delivered to NRMA  
**Then** it should include, An Overview of the data architecture and flow, Integration Points detailing how data will interface with the AI models and the Data Handling Processes outlining extraction, transformation, and loading (ETL) methods.