# **DATA TEAM MVP PROPOSITION**

### **1. Introduction**

The MVP we plan to suggest is a comprehensive presentation that outlines the key components of our data processing pipeline, from infrastructure setup to the final dataset. This aims to provide better insights on our approach to managing and processing data to produce a quality dataset. This presentation will demonstrate our codebase encompassing our infrastructure design, code, and the data pipeline. Additionally, the presentation will also include the results of our data pipeline which is a prototype dataset with roughly around 30 features that have been cleaned, synthesised and enriched.

### **2. Problem Statement**

The main challenge faced by the client is that the current fraud detection system for insurance claims is not efficient and accurate. Thus, this project aims to improve EY’s existing fraud detection system by leveraging GenAI. Thus, for the 1st MVP, the goal is to design and build a data pipeline of Azure to produce a high quality dataset to train an AI model.

### **3. Solution Overview**

**MVP Concept:** The MVP will be a detailed presentation covering two main aspects:

* + **Codebase:** An explanation on various aspects of our codebase on Github which make up the data pipeline.
  + **Final Dataset:** An initial version of the dataset, including key columns that provide essential insights into incidents, driver demographics, and insurance claims.

**Key Features:**

* + **Infrastructure as Code:** Terraform code that defines the infrastructure.
  + **Automated Deployment:** GitHub Action scripts, that automatically deploy the infrastructure as well as, deploy Azure Functions scripts to the cloud.
  + **Data Loading:** PowerShell scripts and ADLS GEN 2 REST API for loading raw data into Azure.
  + **Data Processing Pipeline:** Scripts for cleaning, synthesising, and enriching the data.
  + **Final Dataset:** Columns such as AccidentType, IncidentTime, NumVehiclesInvolved, PoliceReportBool, DriverGender, DriverAge, DriverExperience, LicenceType, EducationLevel, VehicleAge, InsurancePremium, InsuranceAccess, TimeAsCustomer, Fraud, IncidentSeverity, NumBodilyInjuries, AuthoritiesInvolved, TotalClaimAmount

**Scope & Limitations:** The MVP will focus on showcasing the infrastructure setup and initial data pipeline. At this stage we will not be looking at data warehousing options and we will not be looking at the creation of a feedback loop that takes outputs from the AI model to further enhance the dataset.

### 

### **4. Testing & Validation**

**Metrics for Success:**

* + Successful automated deployment of the infrastructure.
  + Accurate and efficient loading of raw data into Azure.
  + Generation of a clean, enriched dataset ready for analysis.

**User Testing:** Feedback from the client on the presentation will be collected to identify any gaps in understanding or technical details.

**Validation Strategy:** Validate the infrastructure setup by deploying to a test environment, and validate the data pipeline by comparing raw and processed data outputs.

### 

### **5. Cost Estimate**

**Monthly Budget Overview:** Azure Data Lake storage ($0.45), Azure Functions ($0), and Application Insights ($0), Github Actions ($0)

### 

### **6. Risks & Mitigation**

**Potential Risks:**

* + **Deployment Failures:** Risks associated with automated deployment scripts failing.
  + **Data Inconsistencies:** Potential issues with data quality during the loading or processing stages.
  + **Data Synthesisation:** Issues with data quality and size when sythesising data

**Mitigation Strategies:**

* + **Testing and Monitoring:** Regular testing of deployment scripts and real-time monitoring of data quality.
  + **Fallback Procedures:** Backup plans in case of deployment failures, including manual deployment options.

### 

### **7. Future Roadmap**

After presenting the MVP, the next steps will include expanding the data pipeline to include data warehousing options as well as a feedback loop with the AI model to take its outputs and improve the dataset.

### **8. Conclusion**

The MVP presentation is expected to provide a clear understanding of the infrastructure and data pipeline setup, along with an initial dataset that can be used for preliminary analysis and evaluation