GenAI Hackathon - Topic

[Document Control 1](#_Toc985947859)

[1.1 Document History 1](#_Toc1348527147)

[1.2 Review Panel 1](#_Toc1581728769)

[1.3 Approvals 1](#_Toc1635553781)

[1.4 Supporting Documents 2](#_Toc304618703)

[Project Details 2](#_Toc411113244)

[Introduction 2](#_Toc649809661)

[Description of the Problem 3](#_Toc1613160636)

[Solution Requirements 3](#_Toc1850811790)

[Solution Overview 3](#_Toc675475206)

[Components 3](#_Toc771044311)

[Capabilities 4](#_Toc31724556)

[Architecture 4](#_Toc464110285)

[Deliverables 4](#_Toc107793120)

[Conclusion 4](#_Toc1673891417)

## Document Control

### 1.1 Document History

|  |  |  |  |
| --- | --- | --- | --- |
| Modified By | Modified On | Comments/Notes | Version |
| Lavanya, Shyam | 07/02/2025 | LLM (Mistral 7b) | 1.0 |
| Shyam, Lavanya | 14/02/2025, 15/02/2025 | Deployed LLM on EC2, | 1.1 |
| Charan | 14/02/2025, 15/02/2025 | Front End (React JS) | 1.2 |
| Lavanya, Shyam, Charan | 17/02/2025 | End to end Setup | 1.3 |

### 1.2 Review Panel

|  |  |
| --- | --- |
| Name | Role |
| GenAI Organizing Team | Coordinators |

### 1.3 Approvals

|  |  |  |
| --- | --- | --- |
| Version | Approval Date | Approver Details |
| 1 |  |  |

### 1.4 Supporting Documents

|  |  |  |
| --- | --- | --- |
| Document | Location | Owner |
|  |  |  |
|  |  |  |

## Project Details

|  |  |
| --- | --- |
| Project Title | **Synthetic Data Generator** |
| Project Sponsor | InfoServices |
| Approval Body |  |
| Approved Budget |  |
| Project Lead |  |
| Team | DataPalette |

## Introduction

Synthetic data generation is the process of artificially creating realistic datasets that mimic real-world data while ensuring privacy and compliance. This technique is widely used for training machine learning models, testing applications, and conducting data-driven analysis without exposing sensitive information.

## Description of the Problem

Organizations often struggle with:

* **Data Privacy Concerns** – Using real-world data may violate compliance regulations like GDPR.
* **Data Scarcity** – Limited availability of diverse datasets can hinder model training.
* **Testing Challenges** – Developers require varied datasets to test applications under different scenarios.

## Solution Requirements

An effective synthetic data generator should:

* Allow user-defined inputs via a **React-based frontend**.
* Accept **structured JSON input** defining schema & record count.
* Utilize **Mistral-7B-Instruct v1** for intelligent data generation.
* Output data in **structured CSV format**.
* Ensure scalability and efficiency via **AWS infrastructure**.

## Solution Overview

The proposed solution leverages an **LLM-based synthetic data generator** hosted on AWS. Users provide structured input via a **React-based UI**, which is then processed by **Mistral-7B-Instruct v1** running on an **AWS g5.2xlarge** instance to generate high-quality synthetic datasets in CSV format.

## Components

|  |  |  |
| --- | --- | --- |
| **Component** | **Technology** | **Function** |
| **Data Input** | JSON (from React app) | Specifies schema & record count |
| **Model** | Mistral-7B-Instruct v1 | Generates synthetic data |
| **Compute** | AWS EC2 (g5.2xlarge) | Hosts Mistral 7B for automation |
| **Storage** | Amazon S3 | Stores input/output data |

## Workflow

1. **User Interaction (React App)**

* The user inputs schema details via a **React-based frontend**.

1. **Backend Processing**

* JSON input is sent to the **Mistral-7B-Instruct v1 model** hosted on **AWS g5.2xlarge**.
* The model generates structured synthetic data following the provided schema.

1. **Data Output (CSV Format)**

* The model outputs structured **CSV data**.

1. **Storage & Retrieval**

* The generated CSV is stored in **Amazon S3**.
* The user can **download the dataset** from the React app.

## Capabilities

### **Scalable** – Handles large data volumes efficiently.

### **Customizable** – Allows user-defined schema configurations.

### **Secure** – Ensures compliance with privacy regulations.

## Deliverables

* **Fully functional synthetic data generator.**
* **React-based UI for input configuration.**
* **API to process and generate synthetic datasets.**
* **Secure AWS S3 storage for generated datasets.**

## Conclusion

This solution enables enterprises to generate realistic datasets while preserving privacy. It accelerates testing, enhances ML training, and eliminates the need for sensitive real-world data.