# DataGeN - AI Powered system to generate safe, secure, compliant and usable datasets

Empowering organizations to generate secure, realistic synthetic datasets for safe and efficient software development and testing.

| Team Name | Hack to the Future |
|---|---|
| Team Leader | Gauri Karkhile |
| Member 1 | Vaibhav Rahalkar |
| Member 2 | Shruti Mone |
| Member 3 | Gaurav Daund |

## Introduction

- Traditional data generation approaches are inefficient for modern, fast-paced development environments.
- Need for robust solutions that can create secure, realistic, and high-volume data without compromising feature distribution or integrity.
- Generates synthetic data that mimics real-world datasets, providing realistic scenarios for software development and testing.

# AI-Driven Data Synthesis

## Input Data

DataGeN utilizes the organization's database as input, along with domain knowledge

## AI Models

We used CTGAN to generate realistic, scalable, and synthetic data by learning patterns and relationships from the original dataset.

## Output Data

The output dataset mimics real data, with PII, PHI, and other confidential information removed or anonymized for safe use.

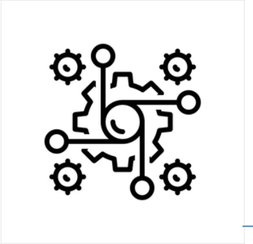

Offline Delivery


User Input


Database Connection


Data Preprocessing


Anonymization


Synthetic Data Generation


Output Generation


User Access

# Approach

- **Offline Secure Delivery:** Product delivered offline to ensure maximum data security.
- **User Input:** User provides database credentials (username, password, database name) for secure access.
- **Automated Data Fetching:** System directly connects to the organization's database and retrieves data.
- **Data Processing:** Presidio is used for synthesizing and anonymizing PII and PHI from the original dataset.
- CTGAN model then synthesizes realistic data based on the patterns learned from the original data.
- **Data Generation:** A synthetic, privacy-compliant dataset is generated for safe use in testing or development.
- **Output:** The synthetic dataset is ready for immediate use, ensuring privacy and data integrity.

# Conclusion

The DataGeN project effectively provides a secure and compliant solution for generating synthetic data. Our approach ensures that organizations can generate safe datasets without compromising privacy. With an easy offline delivery and straightforward user input, DataGeN streamlines the process, enhancing data privacy compliance while supporting efficient software development and testing.