# Data Collection Plan

# For

# Goodness and Mercy School, Kaduna, Nigeria

## 1. Data Generation

To generate synthetic data for a senior secondary school, we initially utilized the **Faker** library. However, recognizing that the generated data lacked the cultural and contextual specificity required for the Nigerian context, we enhanced the dataset by scraping authentic Nigerian male and female names from various sources. We also incorporated Nigerian locations for better geographical relevance and modeled challenges that a typical West African child might face, such as health issues or socio-economic factors. These steps ensured that the dataset was more reflective of real-world situations in Nigerian secondary schools.

Additionally, data quality checks were applied to mimic real-world scenarios. This data was then pushed to two locations:

- A Blob Storage container on Azure for backup and historical data storage.
- A Postgres database on Aiven, where the data was used for analysis and model-building.

## 2. Data Collection and Real-Time Management

The figure below shows the process flow of data collection, pipelining, warehousing, automation and reporting needs.

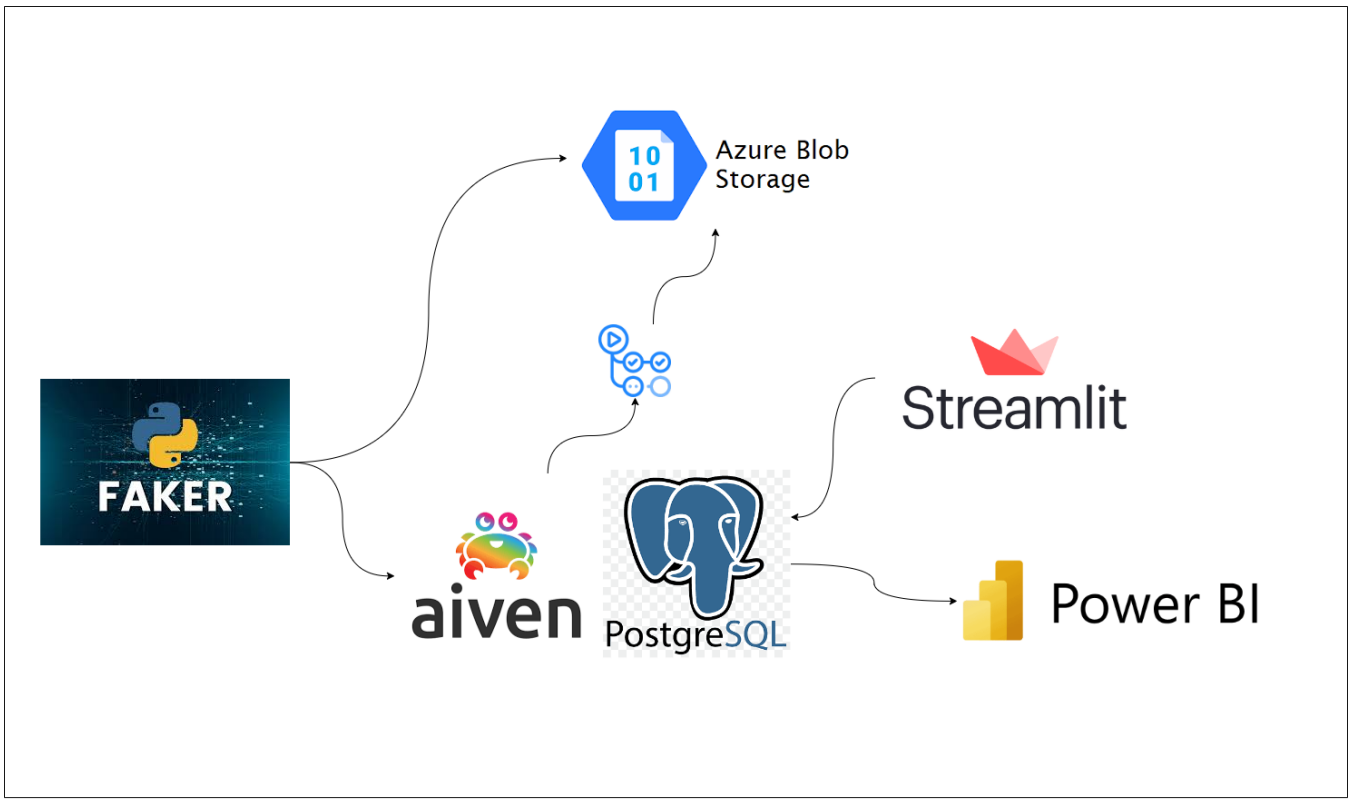


**Figure 1: Data Pipeline Architecture**

To facilitate real-time data collection and updates, we developed a custom **web application** using Streamlit. This application enables school administrators to easily insert new data and update existing records in the database. The app as seen below allows for seamless data entry, ensuring that the school's database remains current.

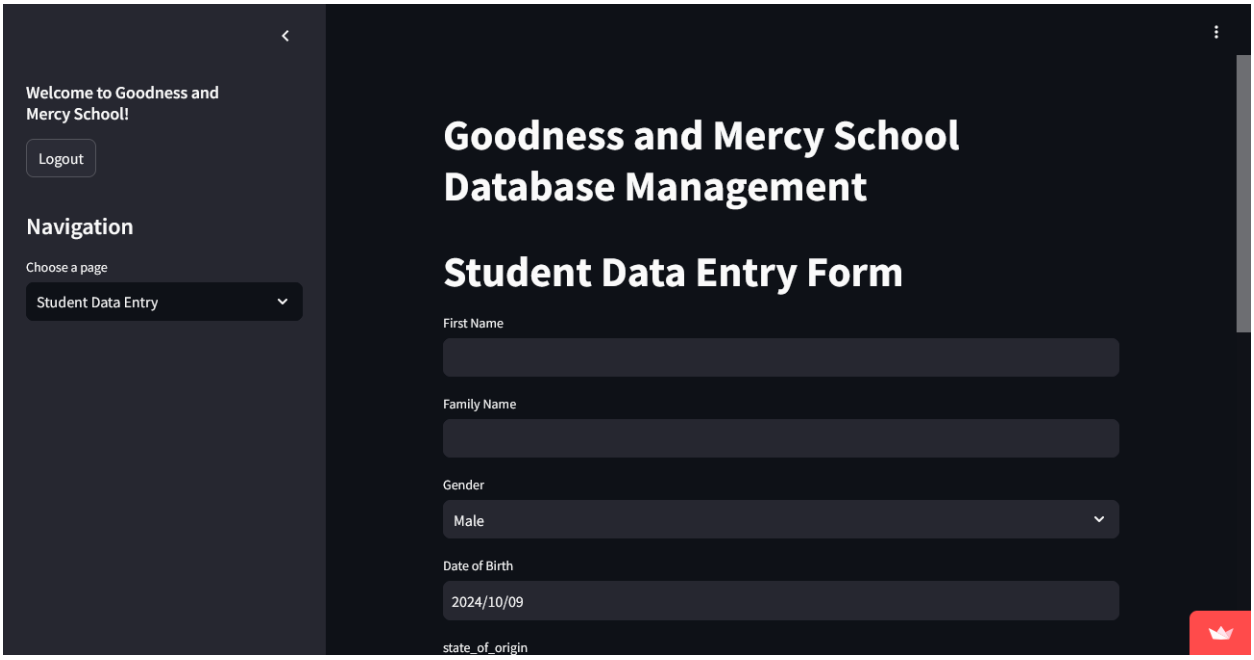


**Figure 2**: Snapshot of the Streamlit app for data entry/ collection

Just like in the pipeline architecture above, new data collected through this web interface is pushed to the postgressql database where it is further connected to Microsoft Power BI where it handles the data reporting needs of the school as in the image below.

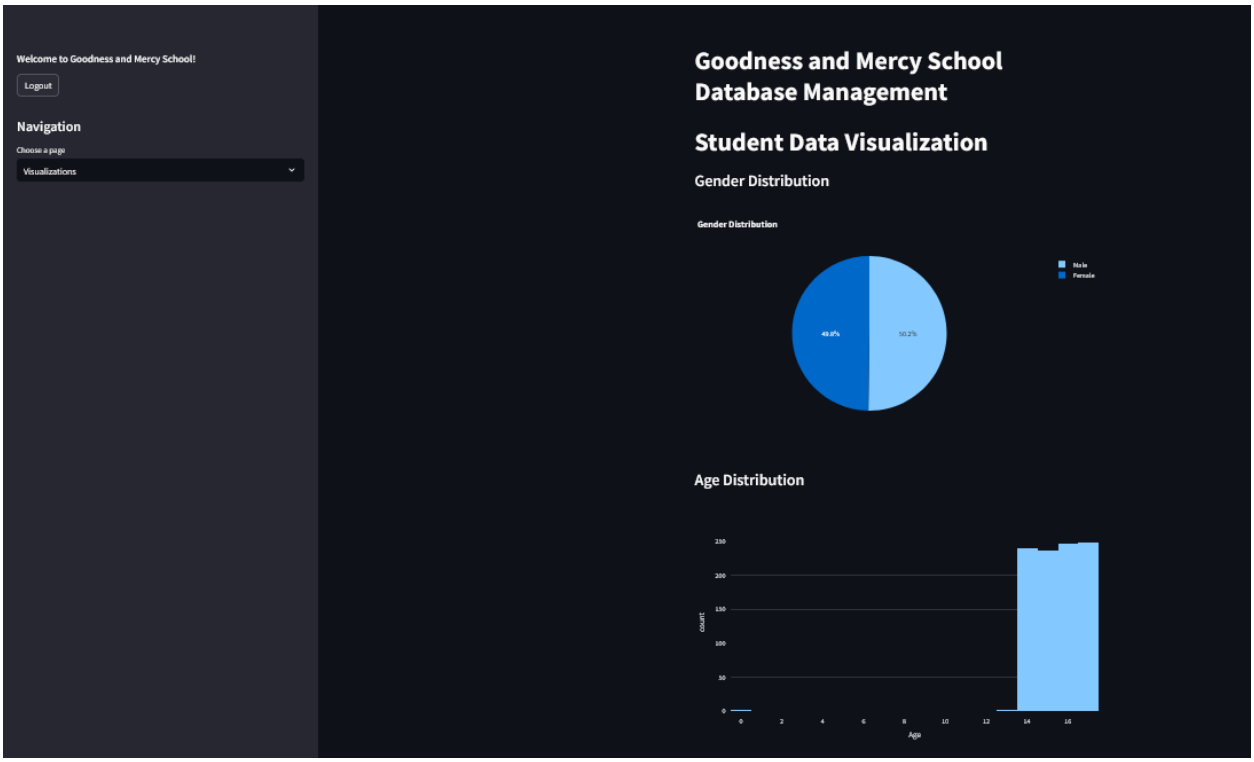


**Figure 3**: Snapshot of the Streamlit app for the reporting needs of the students

Please note that in order to access this interface, the administrators would need to login using the following details:

**Username: datafest_school**

**Password: datafest_school**

Furthermore, after every academic session (approximately three months), the database is queried, and all data within that time frame is moved to the Azure Blob Storage as a PARQUET file. This process has been automated using GitHub Actions to ensure that the data pipeline is both reliable and efficient.

**Tools Overview:**

- **Faker**: For initial data generation.
- **BeautifulSoup:** for web scraping to enhance the dataset with realistic Nigerian names and locations.

- **Streamlit**: For building a data entry web app to facilitate real-time data management.
- **Azure Blob Storage**: For data backup and historical storage.
- **Postgres on Aiven**: For storing and querying the active dataset.
- **GitHub Actions**: For automating the data transfer process from the database to the Blob Storage.
- **Power BI**: visualizations and reporting needs.

These methods and tools combined enabled us to design an enterprise data solution for Goodness and Mercy School data collection, pipelining, warehousing, automation and reporting needs.