**Gen AI-Based Data Profiling System - Design Document**

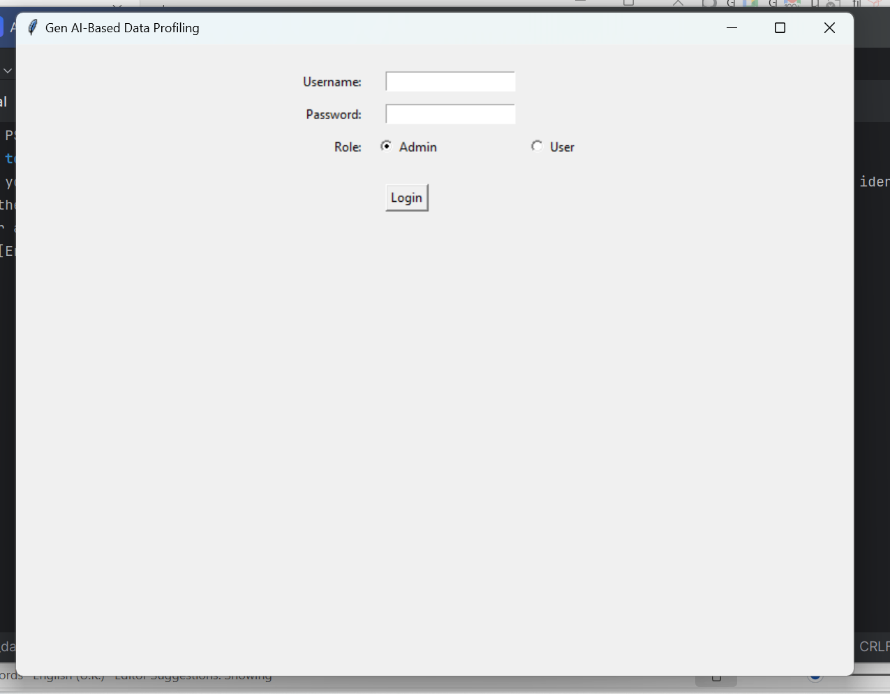
## **1. Introduction**

The **Gen AI-Based Data Profiling System** is a tool that allows users to upload datasets and apply predefined rules using regex-based filtering. The system generates a **data profile report** to analyze the dataset's structure and quality. It is built using Python, **Tkinter for GUI**, **Pandas for data processing**, and **ydata\_profiling for profiling**.

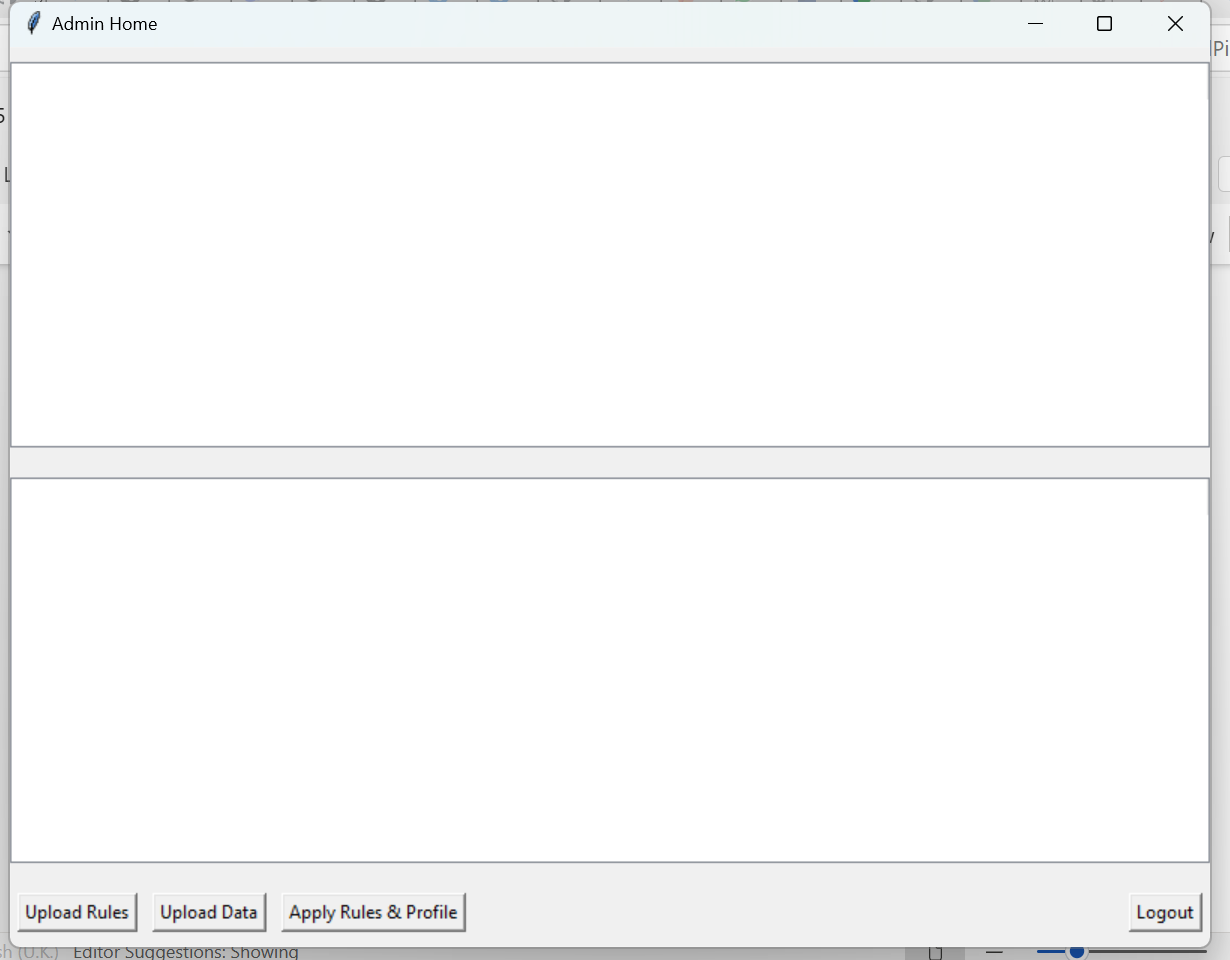
## **2. System Components**

### **2.1 User Interface (UI)**

* **Login System:** Users can log in as Admin or User.



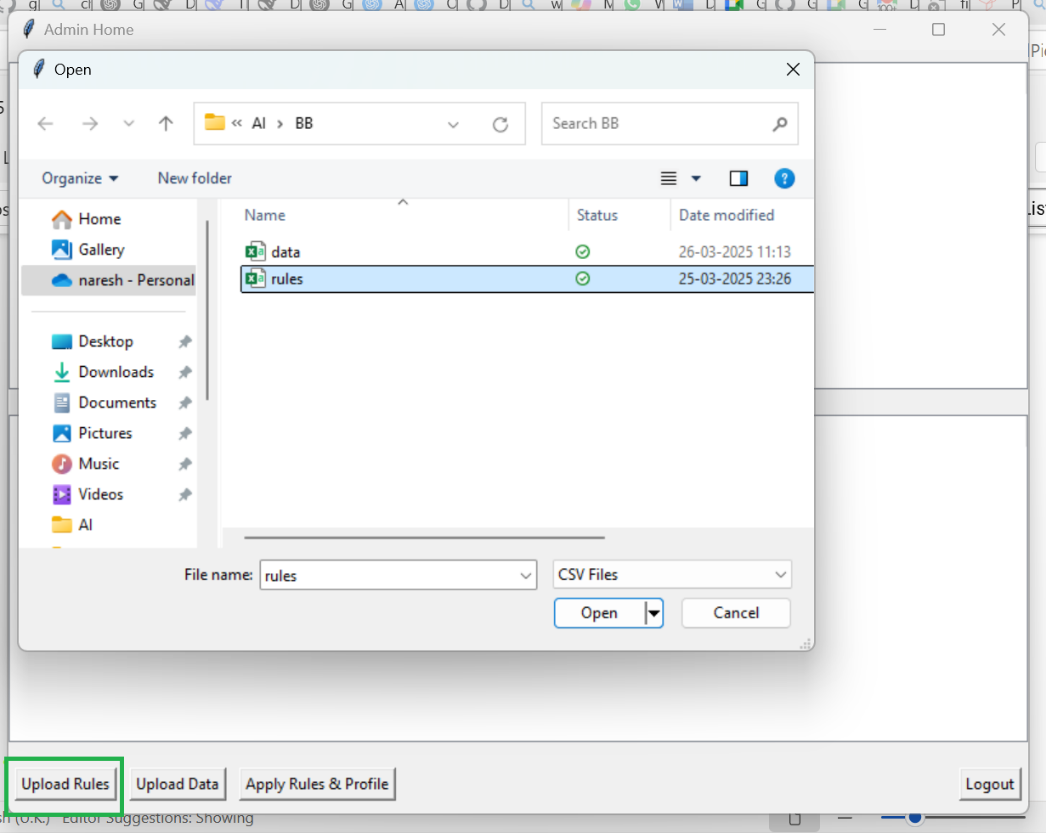
* **Admin Home:** Provides options to upload rules and data files, apply rules, and generate reports.

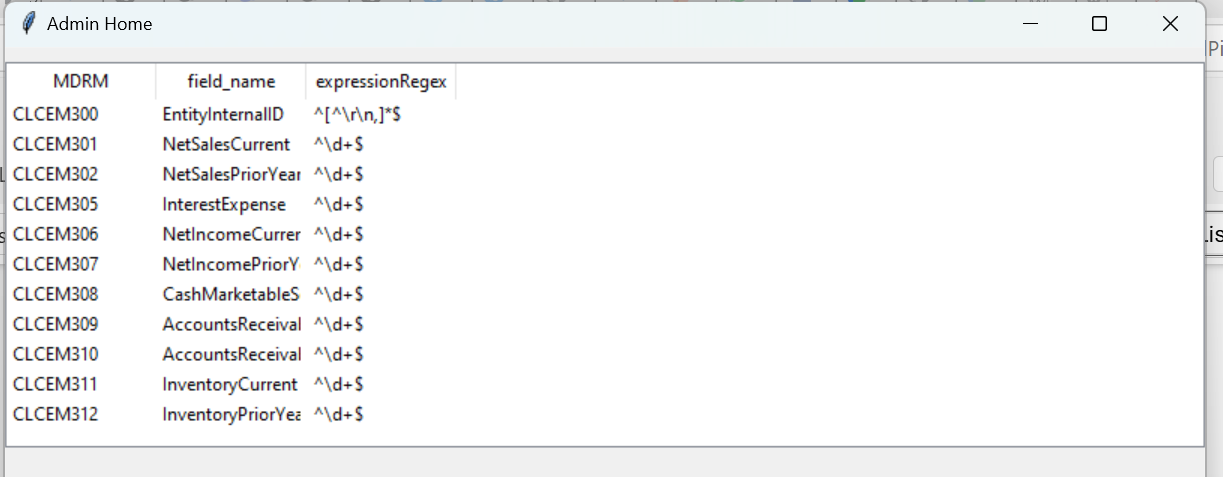


* **Data Visualization:** Displays the uploaded CSV file's content using a Tkinter Treeview widget.

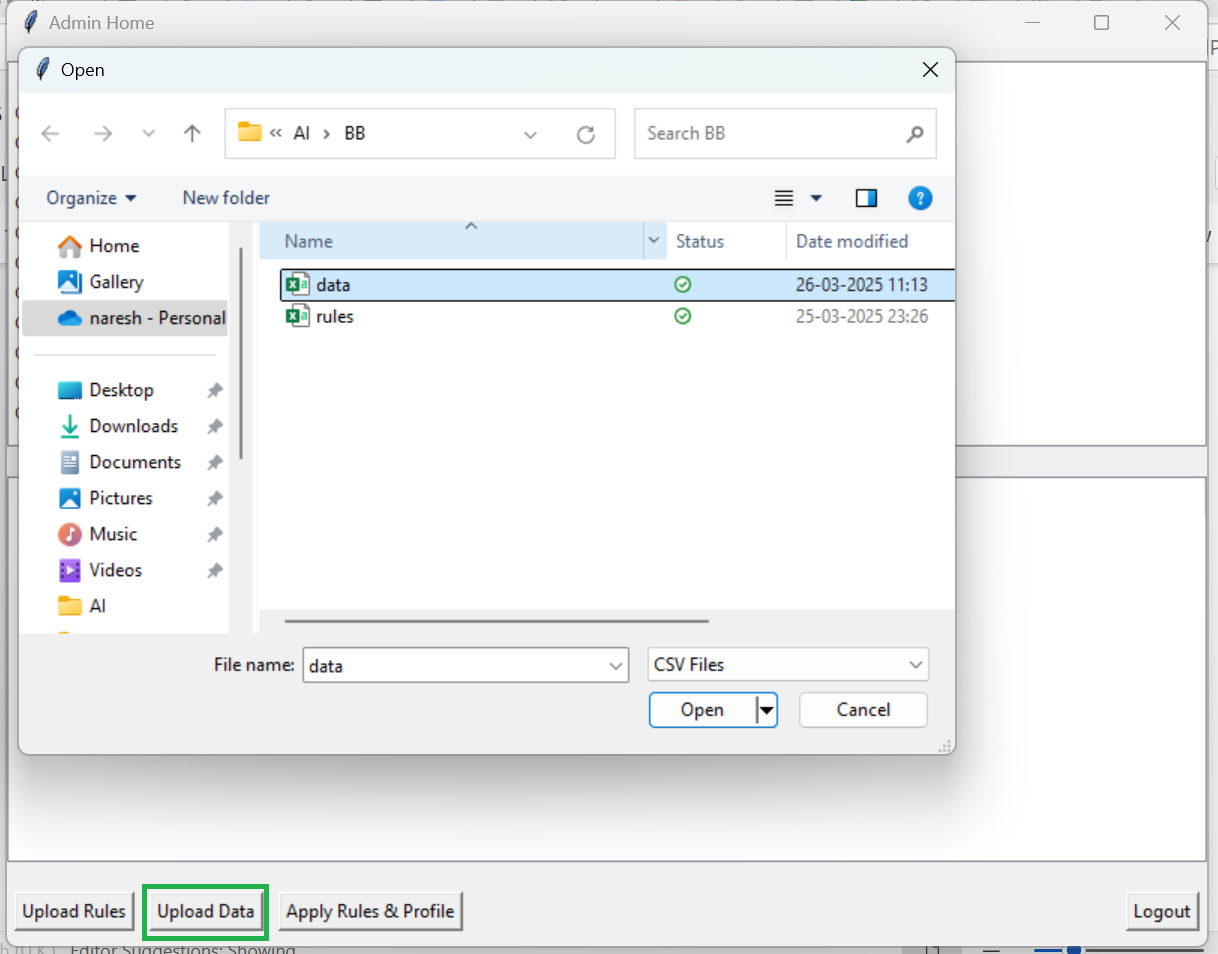
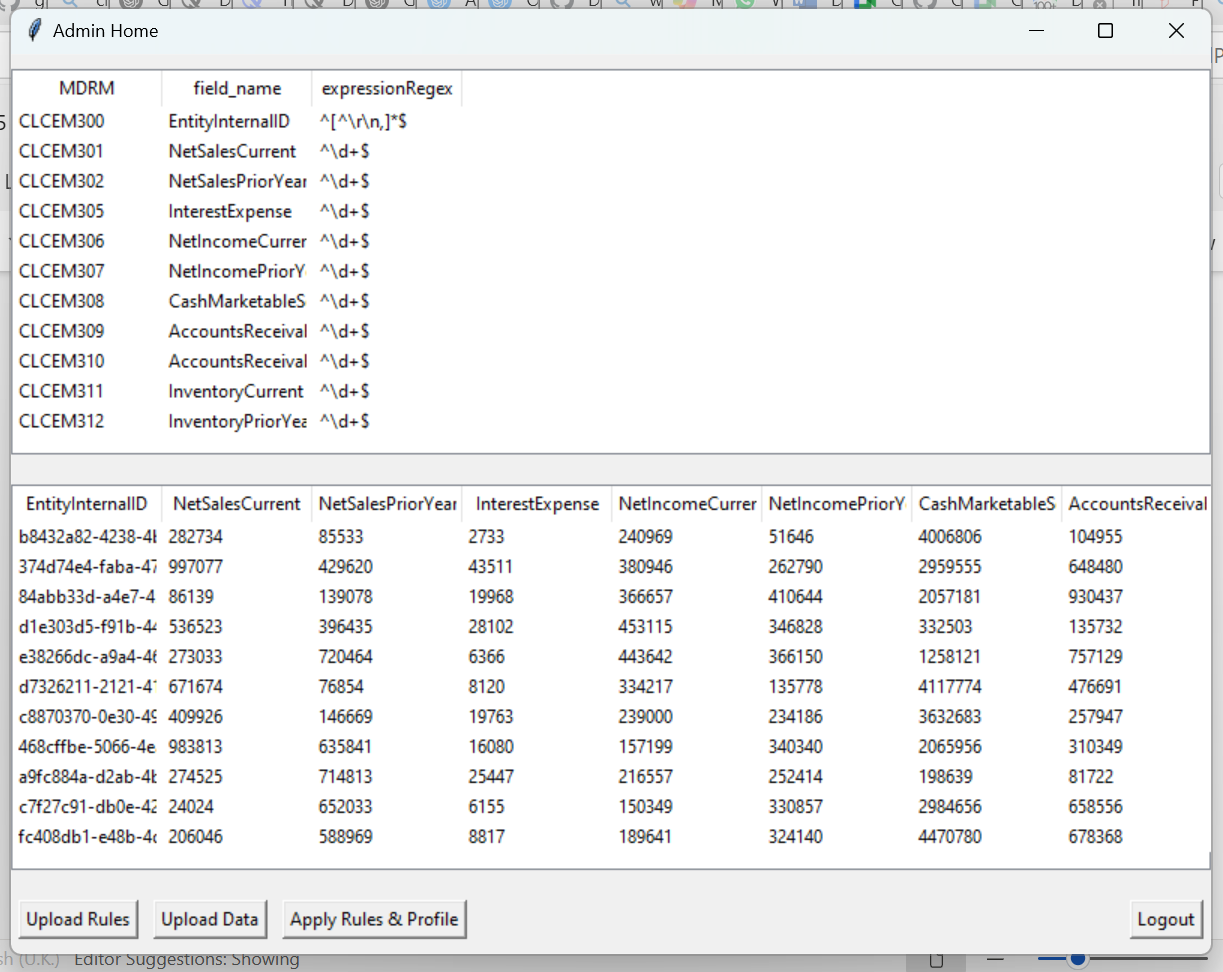
### **2.2 Data Processing**

* **CSV File Upload:** Users can upload datasets and rule files.





* **Rule Application:** Regular expressions (Regex) are applied to filter and clean data based on predefined rules.

* **Data Profiling:** Generates a detailed data profiling report.

## **3. Functional Workflow**

### **3.1 Login & User Authentication**

* The system starts with a **Login Window**, where users enter credentials and select a role (Admin/User).
* If credentials match, Admin is redirected to the **Admin Home Screen**.

### **3.2 Uploading Data & Rules**

* The Admin can upload:
  + **CSV data file** (containing records to be profiled).
  + **Rules file** (containing regex expressions for validation).
* The uploaded files are stored in a predefined directory.

### **3.3 Applying Rules & Profiling Data**

* The **genRegex** module reads the uploaded data and rules.
* It applies regex patterns to the specified fields and filters invalid data.
* The cleaned dataset is saved as **result.csv**.
* A **data profile report** is generated using **ydata\_profiling** and stored as **data\_profile\_report.html**.

## **4. Technologies Used**

* **Python**: Core programming language
* **Tkinter**: GUI framework for user interface
* **Pandas**: Data processing and manipulation
* **ydata\_profiling**: Automated data profiling
* **Regex (re module)**: Data validation and filtering

## **5. Error Handling & Validation**

* Checks if uploaded files are **valid CSV files**.
* Verifies **non-empty datasets** before applying rules.
* Ensures **regex patterns** are valid before filtering.
* Provides appropriate **error messages** using message boxes.

## **6. Output & Reports**

* **Filtered CSV File**: The cleaned dataset after applying rules (**result.csv**).
* **Data Profile Report**: A comprehensive report summarizing dataset properties (**data\_profile\_report.html**).

## **7. Future Enhancements**

* Implement **User Role-Based Access Control**.
* Provide **Graphical Data Insights**.
* Add **AI-driven anomaly detection** to enhance profiling.
* Enable **Web-based UI** instead of a desktop Tkinter application.

This document provides a comprehensive overview of the **Gen AI-Based Data Profiling System**. Let me know if any modifications or additional details are required!