System Architecture

1. Overview

The **Solution_DataProfiling_Local_Model.py** script extracts validation rules from a PDF document and applies them to a dataset using a locally hosted AI model via GPT4All. The architecture consists of three main components:

1. **PDF Rule Extraction** - Extracts relevant text from a given section of the PDF.
2. **Rule Processing & AI Model Execution** - Uses GPT4All with Llama 3.2 1B Instruct to generate validation rules.
3. **Dataset Validation** - Applies the extracted rules to validate a dataset (CSV file).

2. Components

A. Input Layer

- **PDF File** (e.g., `FR_Y-14Q20220930_i.pdf`): Source document containing validation rules.
- **Dataset CSV** (e.g., `sample_transactions.csv`): Data to be validated.

B. Processing Layer

- **PDF Extraction**: Reads specified sections of the PDF.
- **Local AI Model (GPT4All - Llama 3.2 1B Instruct)**: Generates validation rules based on extracted text.
- **Validation Logic**: Applies generated rules to the dataset.

C. Output Layer

- **Validated Dataset (CSV)**: Output file with validation results.
- **Logs/Errors (if applicable)**: Prints logs and errors for debugging.