Project Documentation:

Requirements:

* Accepts different pdf files (for eg – Aadhar card, driving license, Marksheet) in different sections to collect data.
* Extracts the data from the uploaded pdf file.
* Checks the data and verifies if a correct pdf file has been uploaded.
* Raises an error popup in case a wrong document has been uploaded in the wrong section.
* Extracts required data from the pdf files and uploads the necessary data to the required specific database.

Process:

* Create a website to accept a pdf file from the user, extract the data in the file and display in the website.
* Check the data extracted for certain conditions to enable it to be checked against certain conditions to determine whether it is correct document.
* In case the data does not satisfy conditions, raise a popup stating it is an incorrect document.
* Create a database.
* Extract the required data from the document and upload it to the database for future use.
* Scale the website to include different documents to be uploaded.

Tech Stack -

* Frontend: HTML, CSS, JavaScript
* Deployment: Flask
* Text extraction: Amazon Textract
* Database: MySQL
* Source code language: Python

Front-end:

* Different sections to accept different pdf files.
* Raise a popup in case a wrong file has been uploaded.

Text extraction:

* Create an Amazon S3 Bucket.
* Write a python code to accept file from the frontend, send it to the amazon S3 bucket and extract the text within the file using Amazon Textract.

Database:

* Create multiple databases according to the different types of documents to be accepted.
* Each time a document has been uploaded and verified, extract the required data from the document and send it to the respective database.