SDD - Data Extraction & Visualization of Form-Like Structured Documents

Darshan Satra, Nikhil Sharma, Param Shendekar, Vishal Salgond

15th Oct, 2021

Contents

1 Introduction

1.1 Design Overview

Various enterprises have massive amounts of scanned forms that need to be processed every month and added to the enterprise database.The data is also used for generating insights for the organization. All this needs a lot of time and effort in manual data entry into the computer system and then further analysis. This involves a lot of manual repetitive tasks which take up a lot of time and energy. In enterprises the data that is present in the form is mostly printed text apart from the signature, this gives us a window to utilize the advancements in the area of OCR, Natural Language Processing, etc to deliver a product that will automate the manual, cumbersome, error-prone process and generate insights with the data that we have received.

Using text extraction technique, machine learning and data visualization we intend to provide a solution which automates the manual data entry by extracting information from the uploaded forms and store it in the database of the organization. In addition, we will also be using sentiment analysis on relevant fields and using data analysis to produce meaningful insights for the organization and visualize it using data visualization techniques.

1.2 Requirements Traceability Matrix

	User	**Project**	**Form**	**Dashboard**	**Project List**
Account creation and modification	X	X	X	X	X
Project Creation	X	X			
Analyze Form	X	X	X		
Form Data Visualization and Analytics	X	X	X	X	

2 System Architectural Design

2.1 Chosen System Architecture

In order to create the project, a lot of system architecture were studied and discussed. Keeping in mind the features of the project, the functional requirements, and also the amount of audience that the project might potentially serve to, the Client-Server Architecture was finalized. In this architecture, the users (clients) would be requesting a service from the centralised computer (server), which basically would be a response of the extracted data and its visualization.

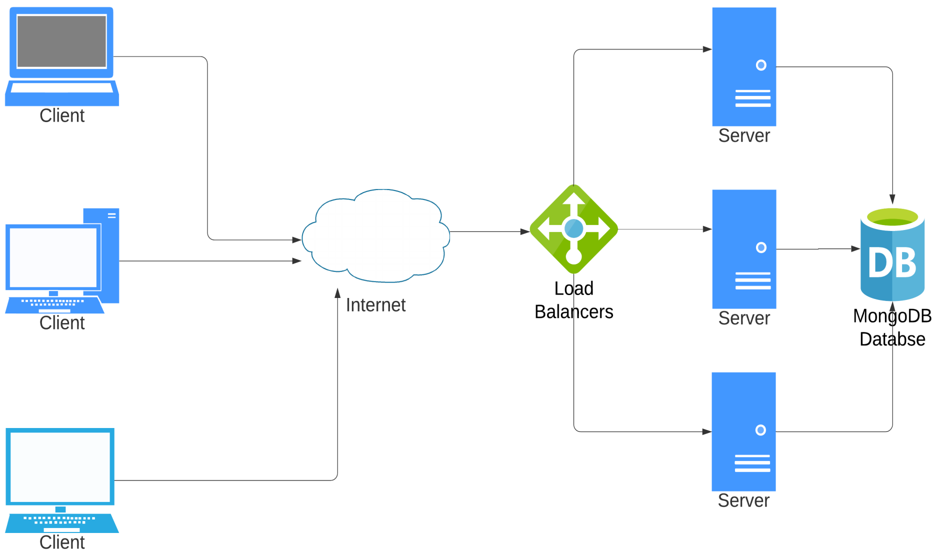


Figure 1: Client-Server Architecture

The server would receive the requests from clients which would involve bulk amount of forms. The required processing would then be done on the server and a response would be sent. The server would host backend scripts pertaining to the machine learning and NLP models.

2.2 Discussion on Alternative Designs

1. Model View Controller

   (a) In this Model-View-Controller (MVC) architecture, the Model is responsible for maintaining the data. The View is used to isolate some amount of data and present the other as per requirement. And lastly, the Controller controls the interaction between Model and View.

   (b) We decided to not follow this architecture as the requirements of processing of forms and subsequent computations were more relevant to the Client-Server Architecture.

2. Data Centric Architecture

   (a) The data centric architecture model would be beneficial in case our project was tending to requirements which help it create or interact with a data center.

   (b) Since that is not the case, it was not used.

2.3 System Interface Description

1. **react-chartjs**: For the visualization purpose of the analytical data, we would need a library for displaying graphs in the project dashboard. We will be using react-chartjs package for visualization purpose, using a library for this purpose would help us minimize the effort of creation of different types of graph.

2. **MongoDB** For the database, we will be using MongoDB, a No-SQL database. We provide a very straightforward interface to Django using PyMongo which helps to interact with MongoDB fairly simple.

3. **AWS S3** Since our system uses forms to extract data, we would want to store these forms somewhere. We will be using AWS S3 for storing the forms temporarily for processing the form.

4. **Operating System** The web application has not specific OS requirement, although the used OS should have a standard browser for accessing our platform.

5. **AWS** Our back end will be hosted in AWS(Amazon Web Services) cloud because it provides an affordable cost and the services provided are much help when trying to build Client-Server Architecture.

3 Detailed Description of Components

3.1 Account creation and modification

The account creation and modification module will be responsible for user management and authentication functionalities like login, registration, updating profile, association projects to users, etc. The constraints for this module is that one email can only be associated with one account.

This module is be composed of Login, Sign-up, Logout, Profile Page and Edit profile functionality. This module is be the base functionality of the entire system. This module will interact with all other functionalities and bind projects to users.

3.2 Project

The project creation module will be responsible for creation and set-up of a project. The constraints for the module will be that a project can only be created and modified by an authenticated user.

The module is composed of creation of project, uploading of blank form and specifying field types to set-up the blue print for the Forms module. This module will depend on the authentication functionality and the blue print and project will be used in the Forms functionality.

3.3 Analyze Form

The Analyze from module will be responsible for text extraction of key and value pairs from the form and storing it in the database. In addition, it will also be responsible for the sentiment analysis on relevant field and storing the sentiment in database. The constraints are that the module will perform sentiment analysis on only text fields like description, review, etc.; Text extraction will only be done on printed text, if there is any handwritten text then the characters should be clearly separated and in uppercase format.

This module will be composed of text extraction model, sentiment analysis model and the API to store results in database. This module will depend on the Project module and the form blue-print setup. This module will interact with the form data visualisation and analytics module by providing the data in appropriate format.

3.4 Form Data Visualization and Analytics

This module will be responsible for performing data analytics on extracted data and display it using data visualization. This module may also be responsible for data pre-processing as data might not be always in required format and structure. The constraints for this module is that there should be sufficient amount of data and appropriate data types should be present to perform meaningful data analytics.

This module will be composed of the backend module which will perform aggregations and provide data in required format, it will include the data analytics model and the data visualization model and an API to show the results. This module will depend on the proper functioning of the Analyze form module because it will be requiring correct and huge amounts of data to derive meaningful inferences.

4 User Interface Design

4.1 Description of User Interface

4.1.1 Screen Images

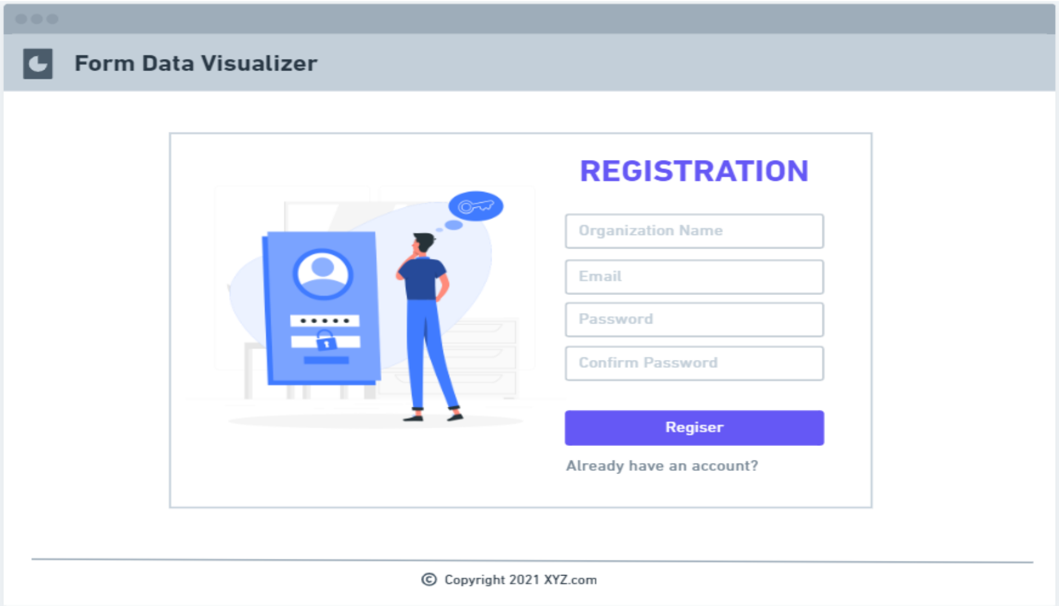


Figure 2: Registration Page

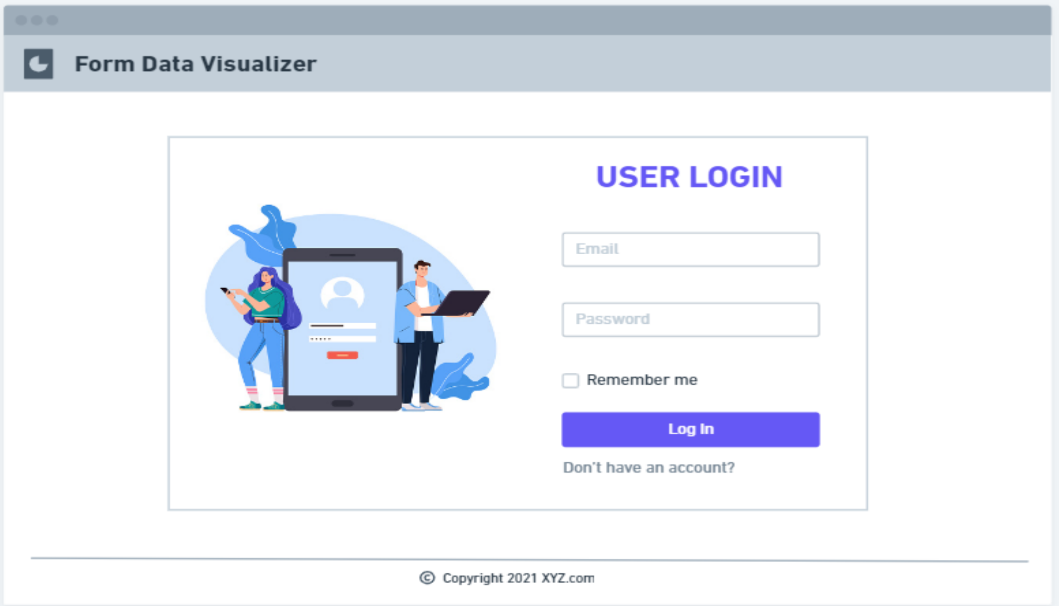


Figure 3: Login Page

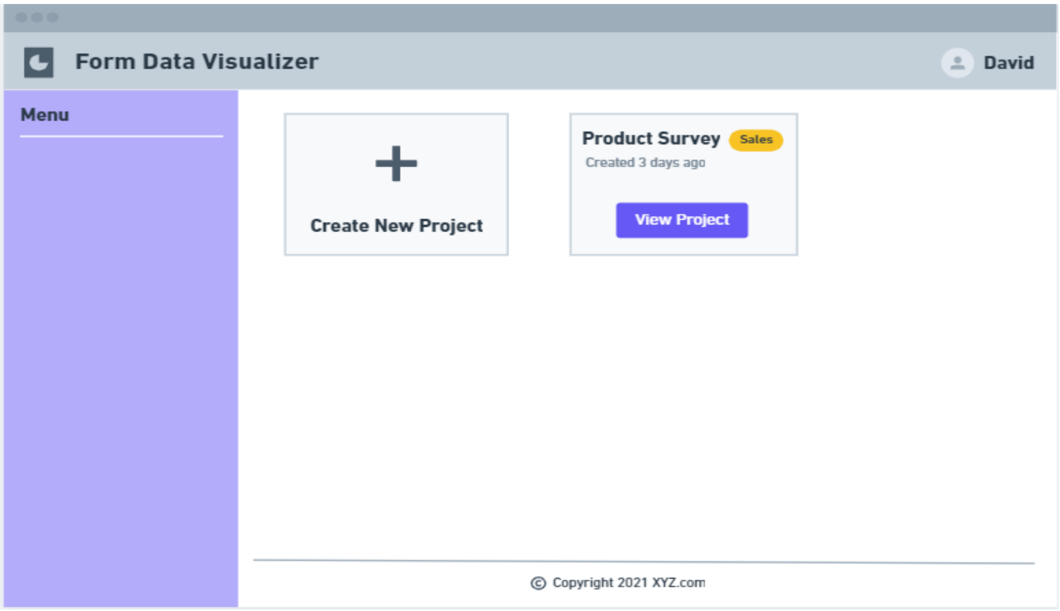


Figure 4: Project List

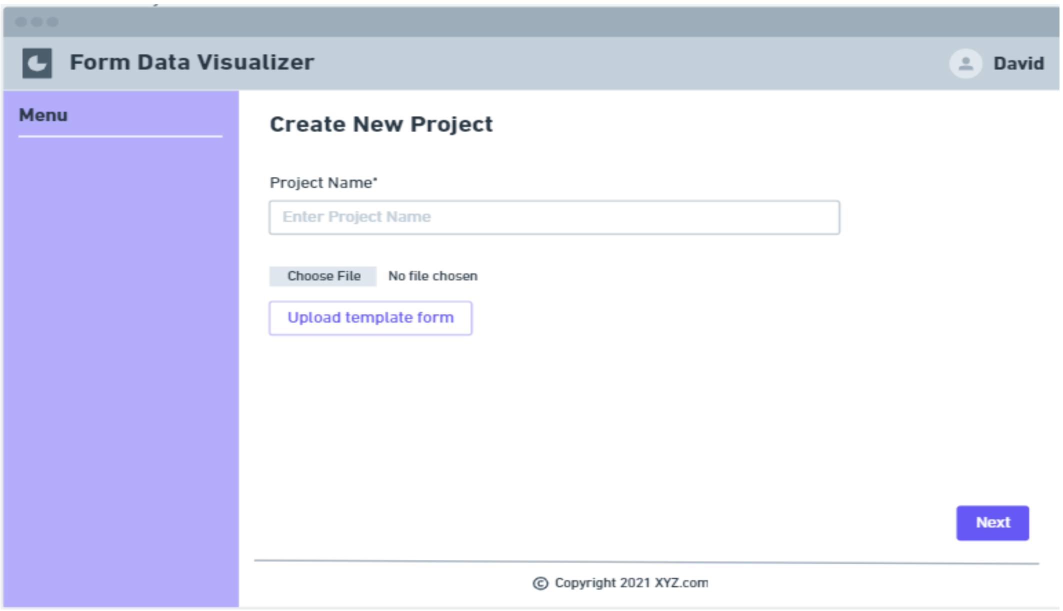


Figure 5: New Project Creation Page

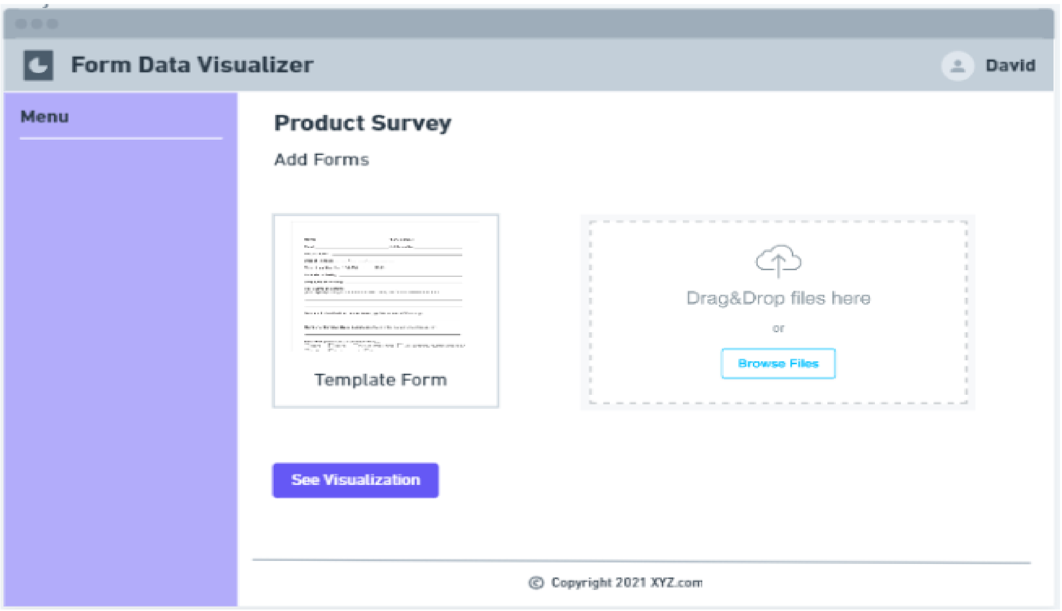
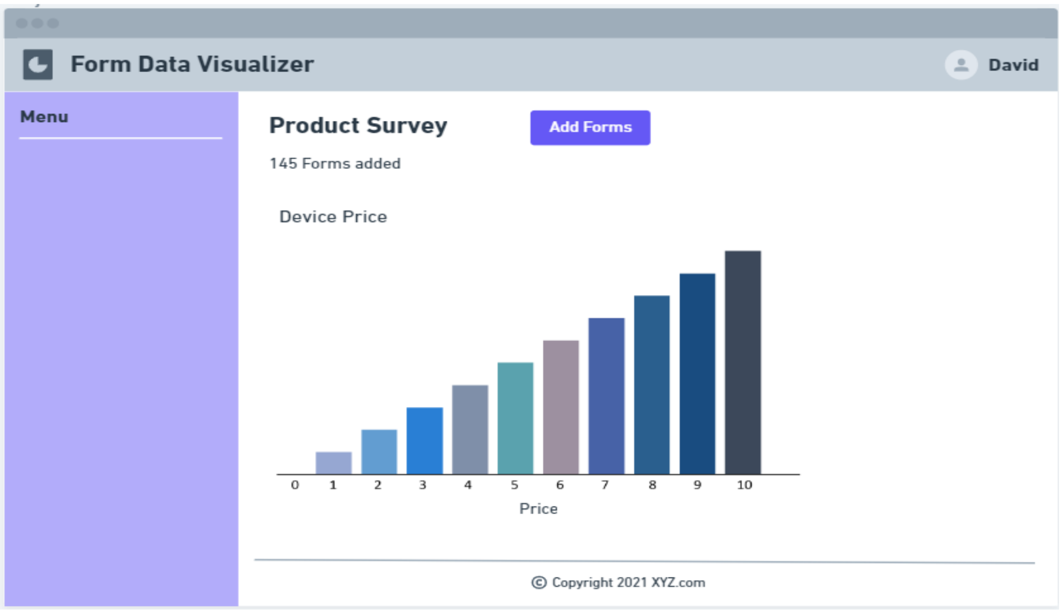


Figure 6: Add Forms Page


Form Data Visualizer

David

Menu

Product Survey

Add Forms

145 Forms added

Device Price

© Copyright 2021 XYZ.com

Figure 7: Insights & Visualization Page

4.1.2 Objects and Actions

Register Page

- **Object:** Textfield
  Action: Taking various details required for registration (Organization Page, Email, Password, Confirm Password) from the user as input.

- **Object:** Register Button
  Action: On pressing this button, the user will get registered in the database provided all the details given by the user fall within the relevant constraints. Once the user is registered, he/she will be able to login.

- **Object:** 'Already Have An Account?' Button
  Action: In case a user is already logged in, he/she will be redirected to the Login page.

Login Page

- **Object:** Textfields
  Action: Taking various details required for authenticating the user to the system (Email, Password) from the user as input.

- **Object:** 'Remember Me' Button
  Action: This will be a radio button, and user will have to press this to instruct their browser to remember their credentials for future ease of access.

- **Object:** Log In Button
  Action: On pressing this button, the user's credentials will then be authenticated against the details existing in system's database, and if match is found, user will be able to log in.

- **Object:** 'Don't Have An Account?' Button
  Action: In case a user doesn't have an account but mistakenly came to the Login page, this button will redirect user to Register page.

Project List Page

- **Object:** Side Navigation
  Action: This will be a side-aligned navigation bar, allowing user for quick access to important pages of the website. Clicking on an item from the menu will redirect user to that page.

- **Object:** 'Create New Project' Button
  Action: A logged in user will have two choices to make, either open an existing project or create a new one. Pressing this button would initiate a new project for the user.

- **Object:** 'View Project' Button
  Action: If user wants to access an older project to reiterate over some insights, then this button will come handy to open that project.

- **Object:** My Profile Button
  Action: This button present on top of the screen will take the user to their profile.

Create New Project Page

- **Object:** Textfield
  Action: This textfield will be for taking input of Project's name. This name will then be referenced whenever a particular project is being accessed or worked upon.

- **Object:** File Upload Button
  Action: This button will be used to upload a blank form which will act as a template for identifying locations of fields and storing related metadata.

- **Object:** 'Next' Button
  Action: This button will be pressed when user is done with uploading of their blank form, and will take them to next step of the process.

Select Fields Page

- **Object:** Radio Buttons
  Action: There will be as many radio buttons as there'd be identified fields which will act as a confirmation from the user whether or not they want a visualization of that field.

- **Object:** Dropdowns
  Action: There will be dropdowns pertaining to each identified fields which will have options of choosing the right datatype of that field.

- **Object:** 'Next' Button
  Action: This button will be pressed when user is done with choosing fields for visualization and assigning their data type, and will take them to next step of the process.

Add Forms Page

- **Object:** Drag & Drop Option
  Action: Using this option, users will be allowed to upload bulk forms which have to be worked upon.

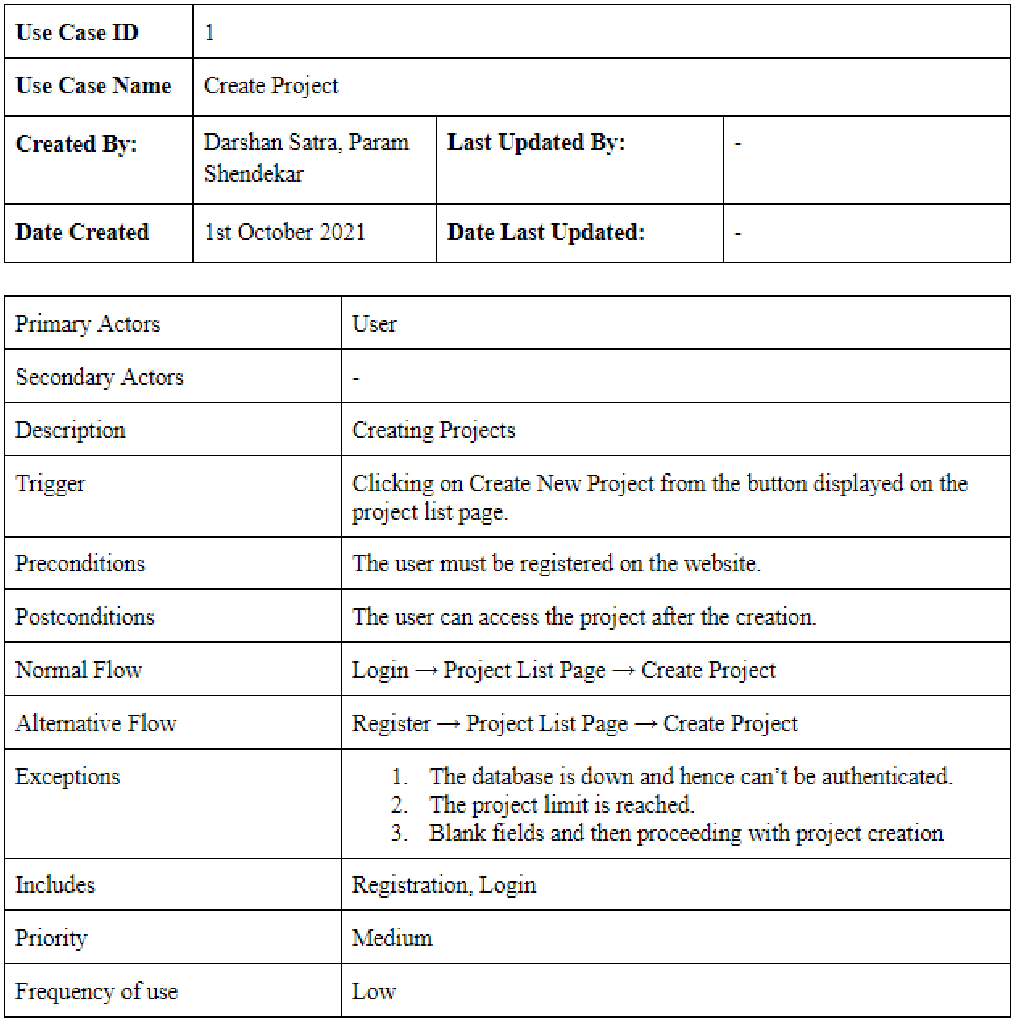
- **Object:** 'See Visualization' Button
  Action: This button will be pressed when user is done with uploading all forms images, and will take them to next step of the process.

Visualizations Page

- **Object:** Add Forms Button
  Action: If after seeing visualizations, user feels to add more forms for processing, then this button will redirect user back to Add Forms page.

5 System Architecture

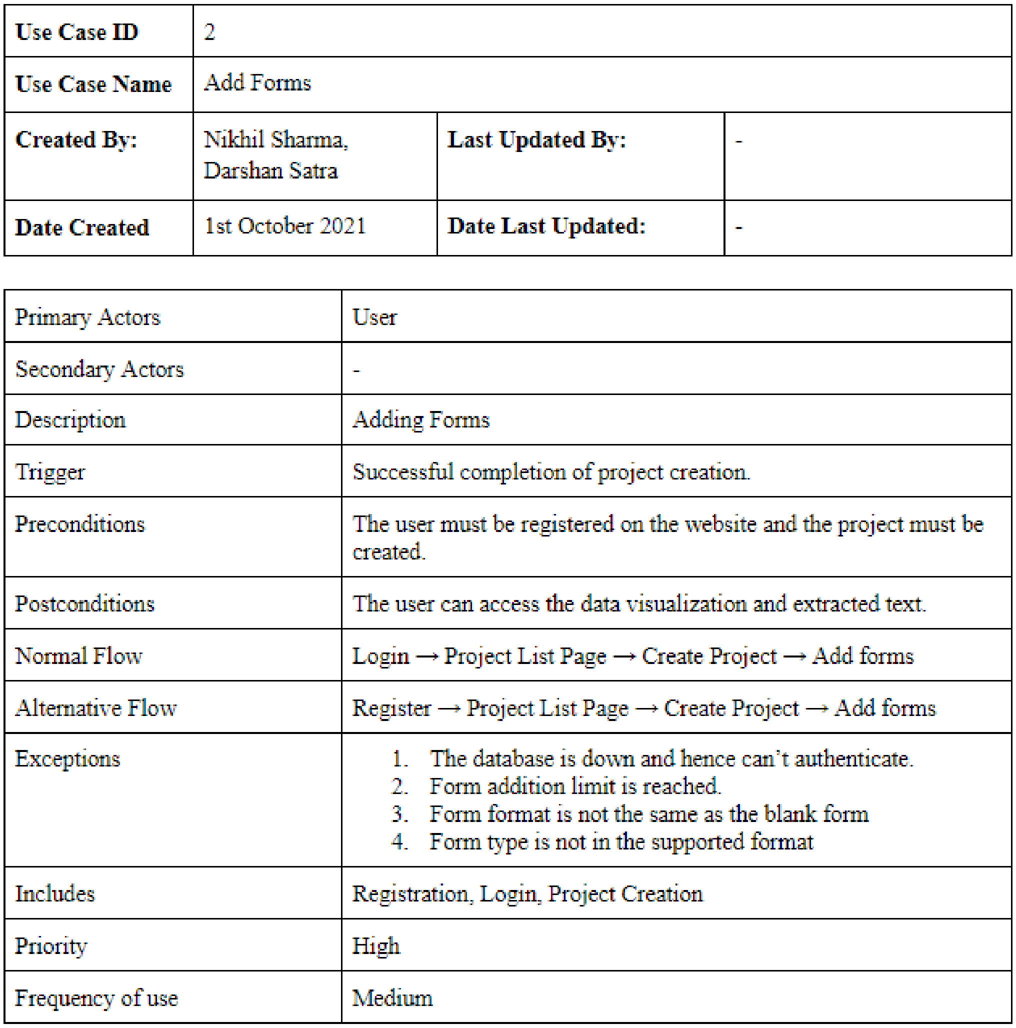
Use Case Specifications



Use Case ID	1		
Use Case Name	Create Project		
Created By:	Darshan Satra, Param Shendekar	**Last Updated By:**	-
Date Created	1st October 2021	**Date Last Updated:**	-

Primary Actors	User
Secondary Actors	-
Description	Creating Projects
Trigger	Clicking on Create New Project from the button displayed on the project list page.
Preconditions	The user must be registered on the website.
Postconditions	The user can access the project after the creation.
Normal Flow	Login → Project List Page → Create Project
Alternative Flow	Register → Project List Page → Create Project
Exceptions	1. The database is down and hence can't be authenticated. 2. The project limit is reached. 3. Blank fields and then proceeding with project creation
Includes	Registration, Login
Priority	Medium
Frequency of use	Low

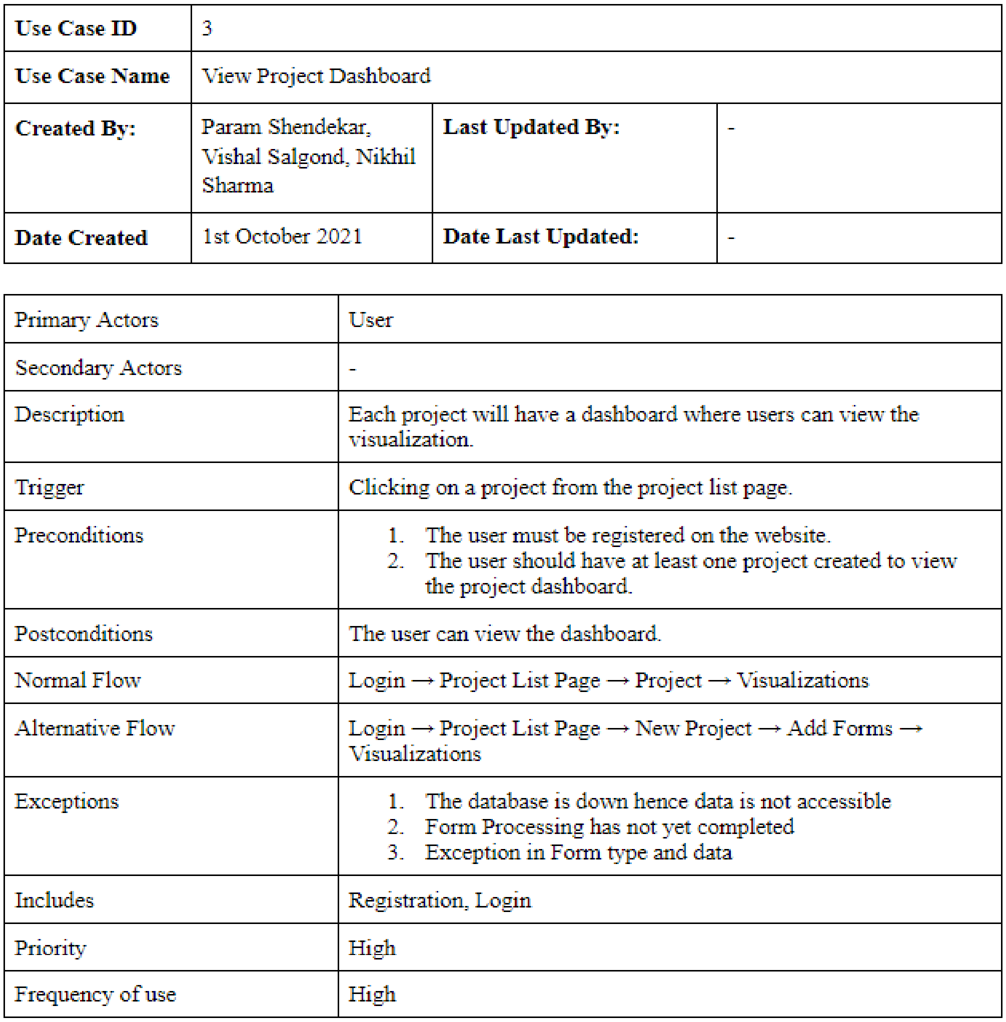
Figure 8: Use Case 1



Use Case ID	2		
Use Case Name	Add Forms		
Created By:	Nikhil Sharma, Darshan Satra	**Last Updated By:**	-
Date Created	1st October 2021	**Date Last Updated:**	-

Primary Actors	User
Secondary Actors	-
Description	Adding Forms
Trigger	Successful completion of project creation.
Preconditions	The user must be registered on the website and the project must be created.
Postconditions	The user can access the data visualization and extracted text.
Normal Flow	Login → Project List Page → Create Project → Add forms
Alternative Flow	Register → Project List Page → Create Project → Add forms
Exceptions	1. The database is down and hence can't authenticate. 2. Form addition limit is reached. 3. Form format is not the same as the blank form 4. Form type is not in the supported format
Includes	Registration, Login, Project Creation
Priority	High
Frequency of use	Medium

Figure 9: Use Case 2



Use Case ID	3		
Use Case Name	View Project Dashboard		
Created By:	Param Shendekar, Vishal Salgond, Nikhil Sharma	**Last Updated By:**	-
Date Created	1st October 2021	**Date Last Updated:**	-

Primary Actors	User
Secondary Actors	-
Description	Each project will have a dashboard where users can view the visualization.
Trigger	Clicking on a project from the project list page.
Preconditions	1. The user must be registered on the website. 2. The user should have at least one project created to view the project dashboard.
Postconditions	The user can view the dashboard.
Normal Flow	Login → Project List Page → Project → Visualizations
Alternative Flow	Login → Project List Page → New Project → Add Forms → Visualizations
Exceptions	1. The database is down hence data is not accessible 2. Form Processing has not yet completed 3. Exception in Form type and data
Includes	Registration, Login
Priority	High
Frequency of use	High

Figure 10: Use Case 3

6 Data Flow Specifications

6.1 Level 0 DFD

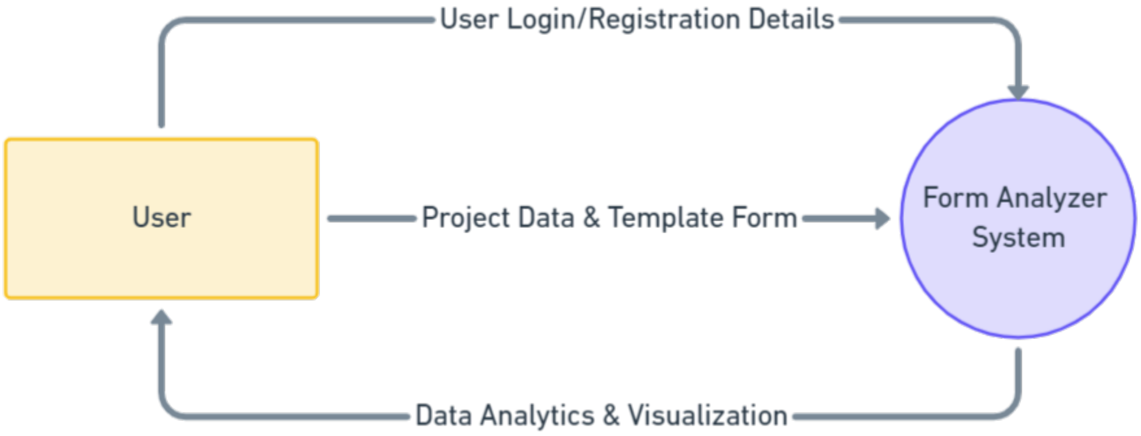

User Login/Registration Details
User
Project Data & Template Form
Form Analyzer System
Data Analytics & Visualization

Figure 11: Level 0 Data Flow Diagram

6.2 Level 1 DFD

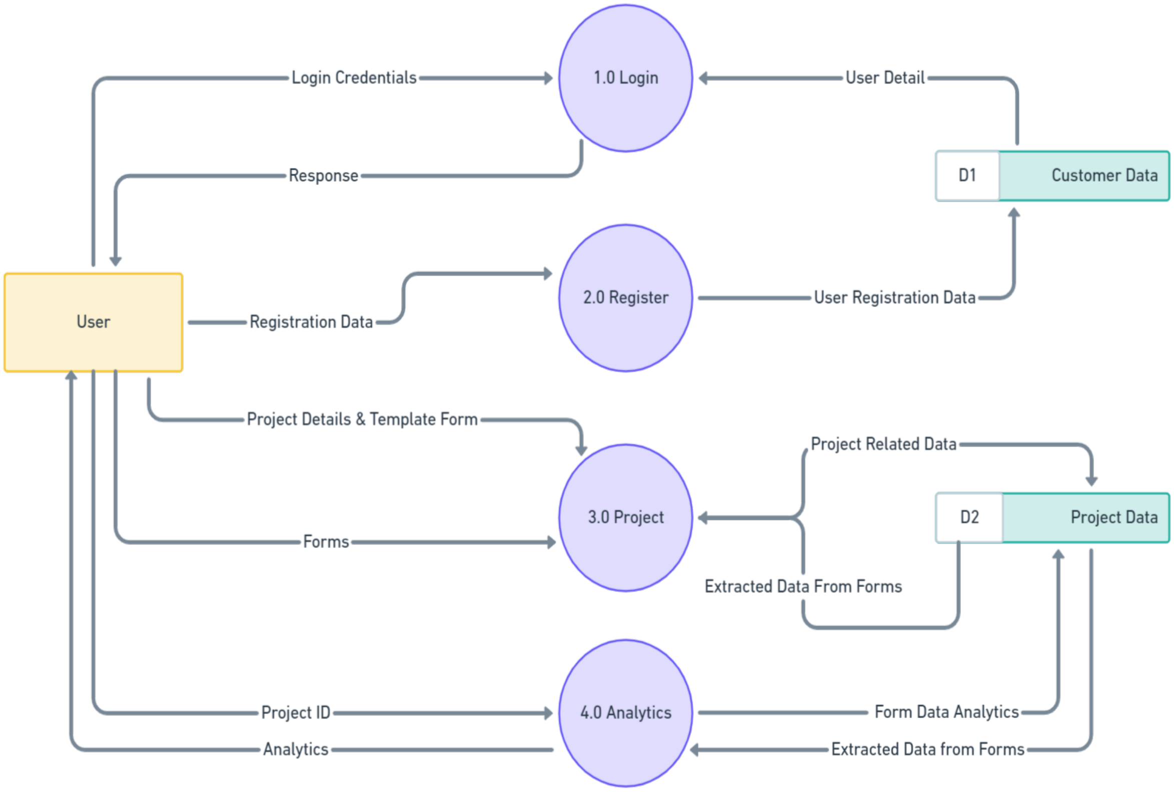


Figure 12: Level 1 Data Flow Diagram