



**VIT<sup>®</sup>**  
**Vellore Institute of Technology**  
(Deemed to be University under section 3 of UGC Act, 1956)

**School of Computer Science and Engineering (SCOPE)**  
**B.Tech – Computer Science and Engineering with Specialization in**  
**Artificial Intelligence and Robotics**  
**Fall Semester 2022-23**

**November, 2022**

*A project report on*

**Large Scale Data Migration Using UIPath**

*Submitted in partial fulfillment for the J Component project of*

**CSE3119 – Robotic Process Automation**

*By*

**Ashutosh Ardu (20BRS1262)**

Signature of the Candidate

Sakthivel V

Signature of the Faculty

## Large Scale Data Migration Using UIPATH

**Abstract.** With so many organizations moving away from purchased software and towards cloud-based business process applications, there's a need to transfer data into those new cloud applications. Data Migration is a multi-step process that begins with an analysis of the legacy data and culminates in the loading and reconciliation of data into new applications. With the rapid growth of data, organizations are in constant need of data migration. Data migration can be a complex process where testing must be conducted to ensure the quality of the data. Migration can be very expensive if the best practices are not followed and the hidden costs are not identified at the early stage. In the past, a large-scale transfer like that would require someone to write a complex database code or to task people with the mind-numbing project of manually copying and pasting information. When a content migration depends on a single hand-coded program, there's the risk that the complexity of the code will lead to unforeseen mistakes, or that the incredibly difficult trick of formatting everything correctly will fail. The other option to move large amounts of data was to have humans copy and paste everything. The drawbacks are: paying all that extra staff time, a large amount of human error possible from inattention, a painfully slow upgrade, and the most boring job in history. The goal is to achieve smooth, error-free and easy content migration, system migration, data migration – there are so many ways to refer to the same thing: moving data from one system to a new one and aim is to set up a workflow to retrieve multiple fields and input them into another database with quick speed and accuracy.

**Keywords:** Data Migration, Legacy Data, Database.

# 1 Introduction

Data migration is the process of moving data from one location to another, one format to another, or one application to another. These days, data migrations are often started as firms move from on-premises infrastructure and applications to cloud-based storage and applications to optimize or transform their company. Our project helps in data migration in Customer Information recording application, where all the data of potential customers is stored in different formats in a single location, but the data itself is unordered. Our project extract data from this unordered data by applying different methods depending upon the type of file it is dealing with and stores it in a excel sheet. Now that data organisation is achieved meaning data from different file format is extract and is organised in a single file, the application loads the data onto a cloud portal with the of automation tool called UI Path.

## 2 Motivation

Our data is scattered. Locked in different systems, separated in silos, data is all over the place, and it's accumulating at a staggering rate. But data can be so much more valuable when we can get it all in the same place. When we have all of our data together, we can gain a 360-degree view of our organizations. We can then use information about our customers, products, and services to make faster, better, more informed business decisions. Data migration is an essential component of data management and is the key to gaining this 360-degree view. By implementing our project, we can gain a consolidated view of any organization and unlock the full value of your data. In today's world, data migrations for commercial reasons have become common. With so many organizations moving away from purchased software and towards cloud-based business process applications, there's a need to **transfer data into those new cloud applications**. And we intend that our project achieves this task with easy, accuracy and ease.

## 3 Background

In this section, we will introduce various techniques through which data is extracted from a given file of a particular type.

### 3.1 OCR Engine

This technique is used to extract information from image files. **Optical character recognition** is the electronic or conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image. An **OCR Engine** is the part of OCR software that does the actual character recognition, analysing the pixels on an image and figuring out what characters they represent. This raw result can be further processed using Grooper's OCR Synthesis capabilities, producing the final OCR result used by Data Extractors to match text in a document and return the result. OCR Engines themselves have four phases: Pre-Processing, Segmenting, Character Recognition and Post-Processing.

### 3.2 REGEX

This technique is used to extract information from a text file. A **regular expression** is a sequence of characters that specifies a search pattern in text. Usually, such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. Regular expressions are used in search engines, in search and replace dialogs of word processors and text editors, in text processing utilities.

### 3.3 Excel Extraction

This technique is used to extract data from an excel file. Microsoft Excel is a software used for storing, organising and manipulating data. These processes can be made more efficient with the help of UI Path activities such as Excel Application Scope and Workbook activities. Using them, we can perform a wide range of operations including reading a cell, writing in a cell, executing macro, filtering DataTable and many more.

## **4 Related Problem**

### **4.1 Image Processing**

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Nowadays, image processing is among rapidly growing technologies. It forms core research area within engineering and computer science disciplines too. Image processing basically includes the following three steps- importing the image via image acquisition tools, analysing and manipulating the image and Output in which result can be altered image or report that is based on image analysis. There are two types of methods used for image processing namely, analogue and digital image processing. Analogue image processing can be used for the hard copies like printouts and photographs. Image analysts use various fundamentals of interpretation while using these visual techniques. Digital image processing techniques help in manipulation of the digital images by using computers. The three general phases that all types of data have to undergo while using digital technique are pre-processing, enhancement, and display, information extraction.

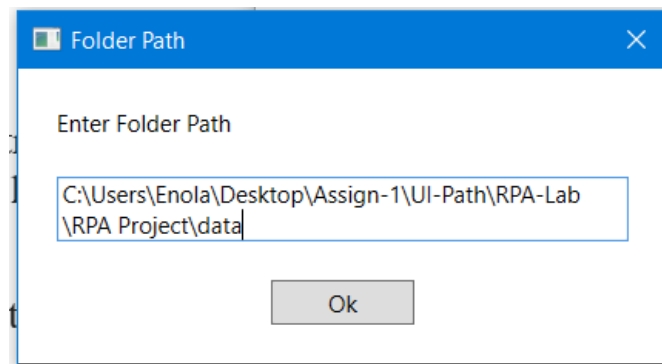
### **4.2 Computer Vision**

Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs and take actions or make recommendations based on that information. Computer vision enables computers to see, observe and understand. Computer vision works much the same as human vision, except humans have a head start. Human sight has the advantage of lifetimes of context to train how to tell objects apart, how far away they are, whether they are moving and whether there is something wrong in an image. Computer vision trains machines to perform these functions, but it has to do it in much less time with cameras, data and algorithms rather than retinas, optic nerves and a visual cortex. Because a system trained to inspect products or watch a production asset can analyse thousands of products or processes a minute, noticing imperceptible defects or issues, it can quickly surpass human capabilities.

## 5 Modules

### 5.1 File Module

This module prompts the user to select a particular location within the system, a location from which the user wants to extract the data from the files present in that particular directory.

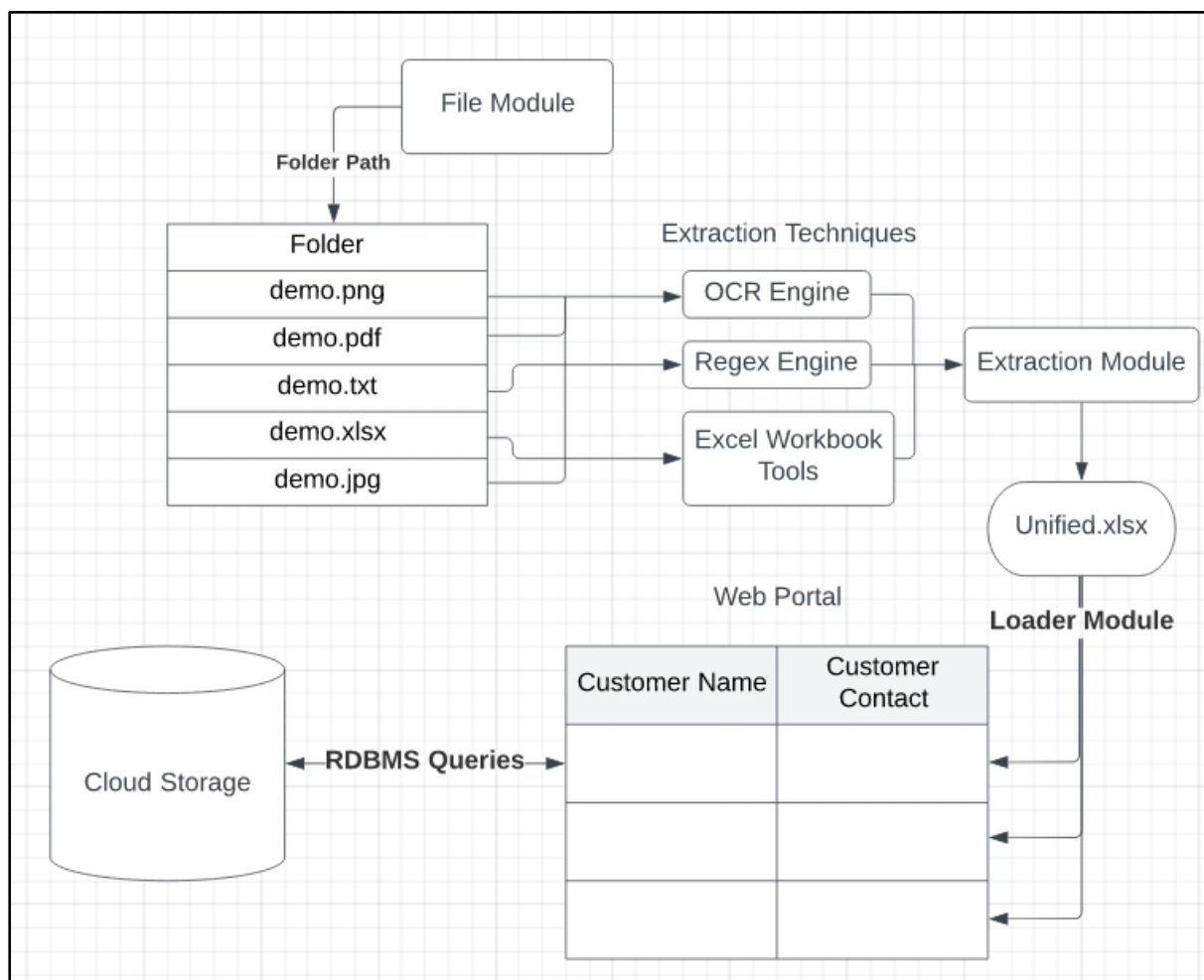


### 5.2 Extraction Module

This module is responsible for extracting data from the files of different types from the location mentioned by the user through the File Module. The module checks the type of each file using a for each method. After confirming the file type, the module applies the specified technique for data extraction as mentioned in its workflow. For e.g., OCR Engine method for extracting data from PDF and Image files, REGEX for extracting data from text file or xml files or json files and using Excel methods (present in UI Path) to extract data from excel files. After extracting the data from the file using the file specific technique for data extraction the workflow/module stores all the data extracted in a unified excel workbook for further data loading.

**5.3 Loader Module.** This module is responsible for migration data from the unified excel workbook to the cloud storage via a web portal interface which is hosted online/locally. Using Excel workbook read activities (UI Path) the module reads the data row wise and stores it inside a data table variable. Using Web Recording activities (UI Path) the module loads the data present in the data table variable under the specified tag into the web portal which uses MySQL database as its backend. The web portal uses HTML5, CSS3 and ExpressJS for its frontend and backend consists of MySQL database applied using NodeJS. The module first opens the web portal in the browser application and first identifies the label and enters the appropriate data from the data table variable in the input form under each label and once all input forms are filled the submit button clicked. On clicking the button, the inputted data is converted into an RDBMS insert query by ExpressJS and is sent to the database for insertion, similar steps are followed for record searching process. Hence the module iterates through all the excel data ranges and transfers the data to cloud storage via the web portal with the help of Web Recording functionality of UI Path.

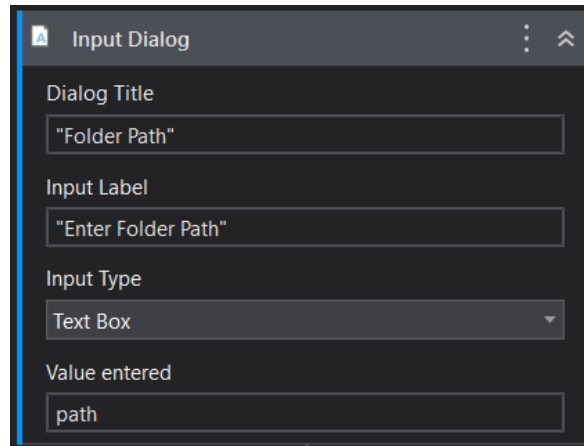
### Proposed Model



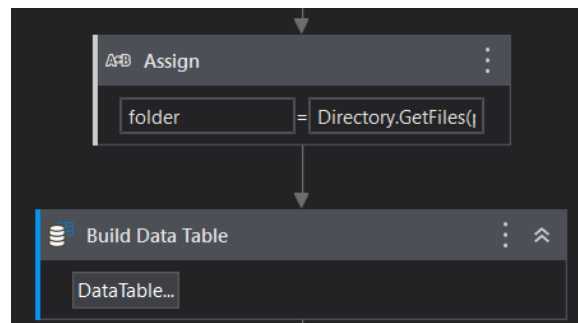
## 6 Result

### Workflow

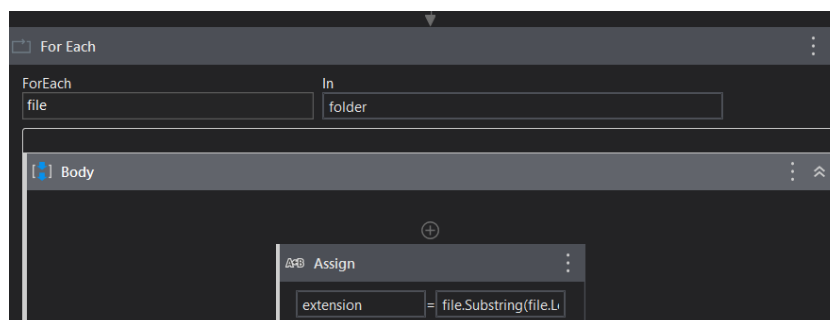
Taking the file path where all the customer data in different file format is stored



Extracting the list of files present in that directories and building a datatable to store all the customer data present in different file formats.

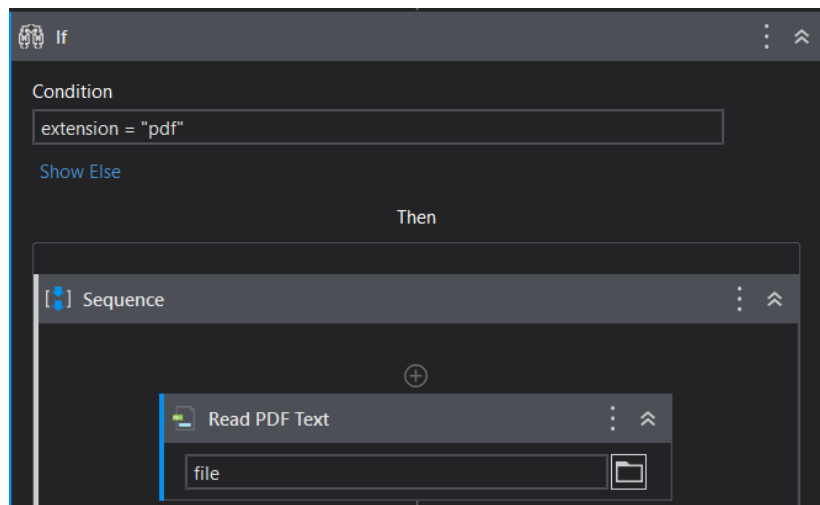


Looping through all the files present in the directory and extracting their file extension to identify the file formats

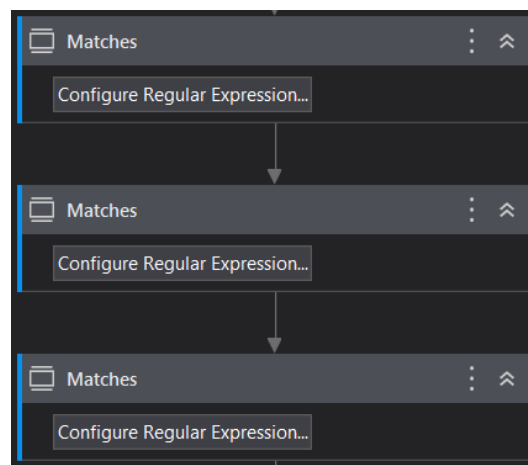




If current file type is of PDF format, extracting the text present in the pdf and applying regex modules to extract customer name, email and phone numbers



## Regex Module



## Regex Module for extracting name

Value	Quantifiers
<code>(([a-z]+) : ([a-z]+(\[a-z]+\))?[a-z]+)</code>	At least one (1 or more) ▾

## Regex Module for extracting email

RegEx	Value	Quantifiers
Email	<code>((?&gt;[a-zA-Z\d!#\$%&amp;'+\-\./=?^_`{ }~\x20"]((?=[\x01-\x7f]) ^\\ \\</code>	At least one (1 or more)

## Regex Module for extracting phone number

Value	Quantifiers
<input type="text" value="(91)?([0-9]{3}[- ]?[0-9]{3}[- ]?[0-9]{4})"/>	At least one (1 or more) ▾

Data Row Append Module for adding the extracted customer data using regex

For Each

ForEach: currentItem, In: pdfName

Body

Add Data Row

ArrayRow \*: {pdfName(pdfInd).Groups(2).ToString, pdfEmail(p...

DataRow \*: The DataRow object to be added to the DataTable. I...

DataTable \*: dt

## Array Object loaded into the Datatable

```
ArrayRow (Object[])
1 {pdfName(pdfInd).Groups(2).ToString, pdfEmail(pdfInd).ToString, pdfNumber(pdfInd).ToString}
```

If the file is of type Text (.txt) then we extract the data in text format and then apply the regex module followed by Data row append module to add it to the common datatable

If

Condition: extension = ".txt"

Show Else

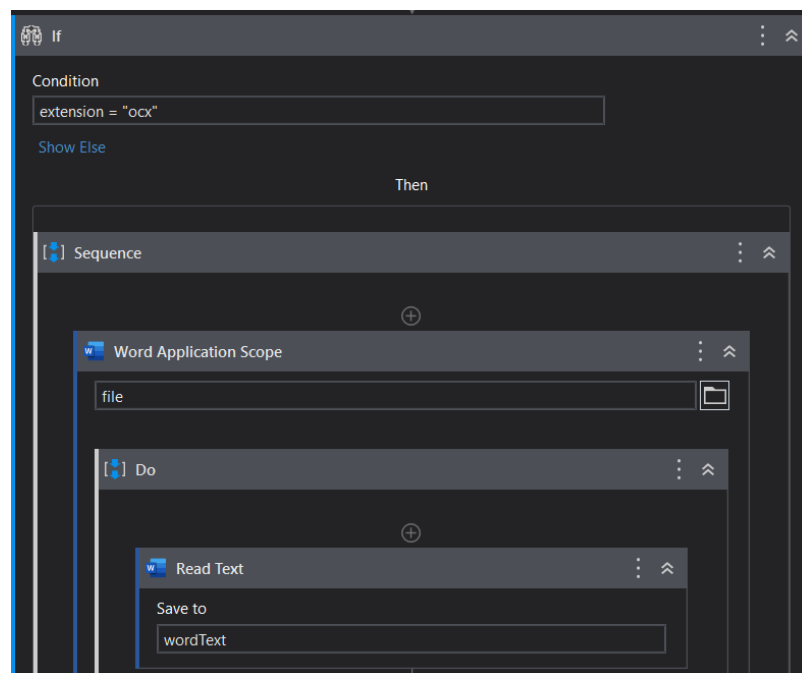
Then

Sequence

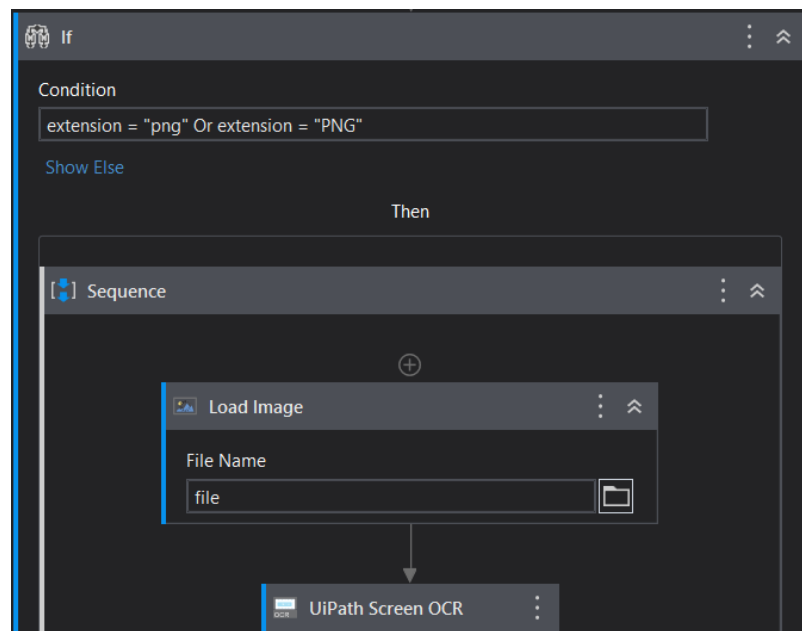
Read Text File

Filename: file

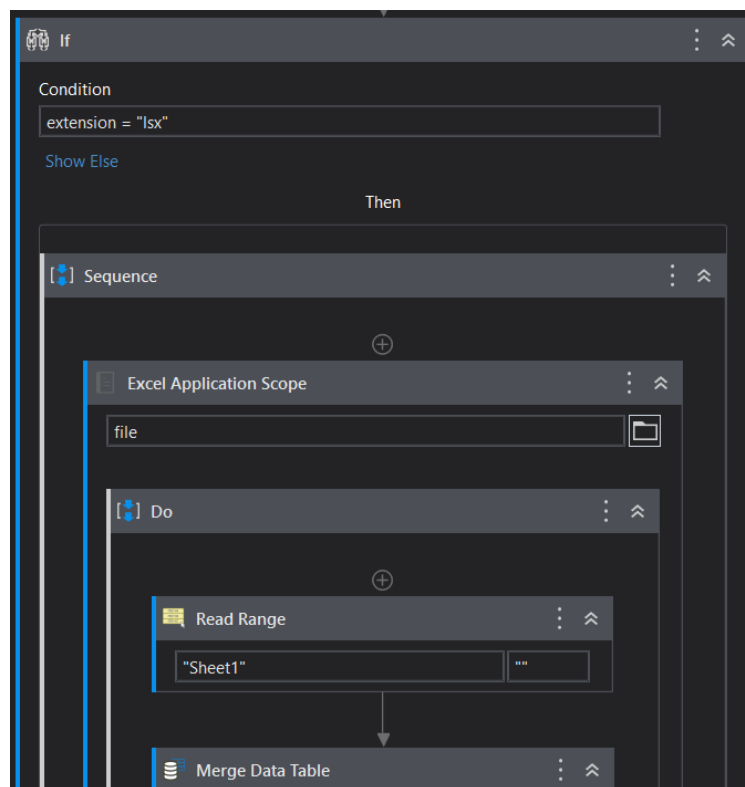
If file is a docx type file then we extract text using the following activities and later we apply the regex module to extract the customer data followed by data row append module to add the data to the common datatable



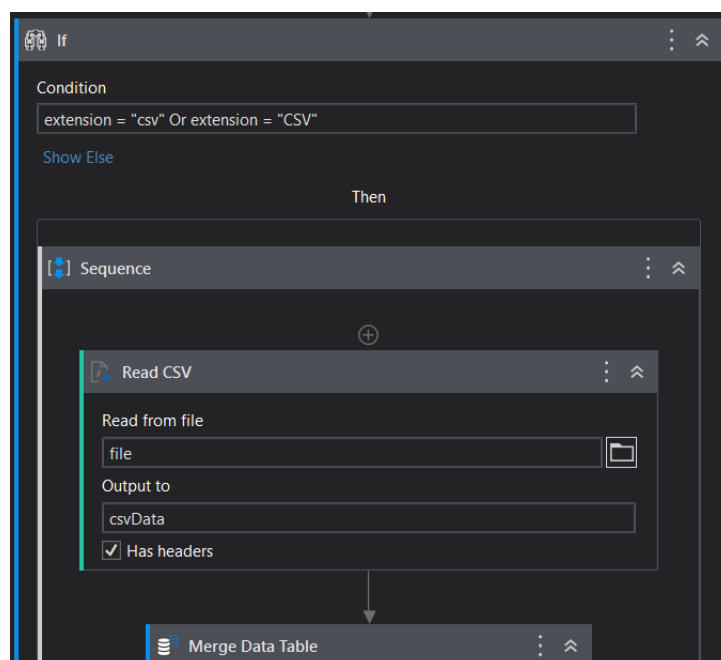
If file is an image file, then we apply OCR activities to extract the data from the image file and later apply regex module and data row append module for further data processing.



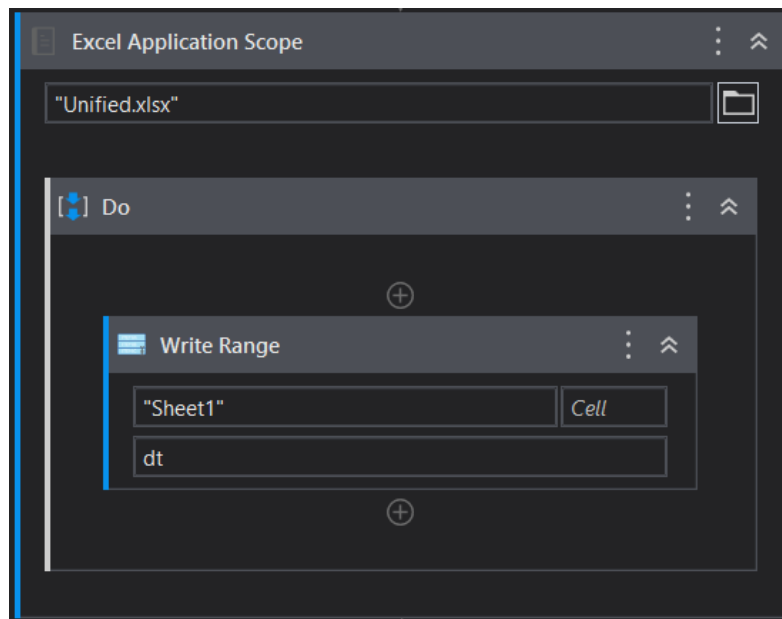
If file is of excel format, then we read the data using read range activity and later we simply merge the data table produced by read range to parent data table where we are storing the unified customer data from all file formats



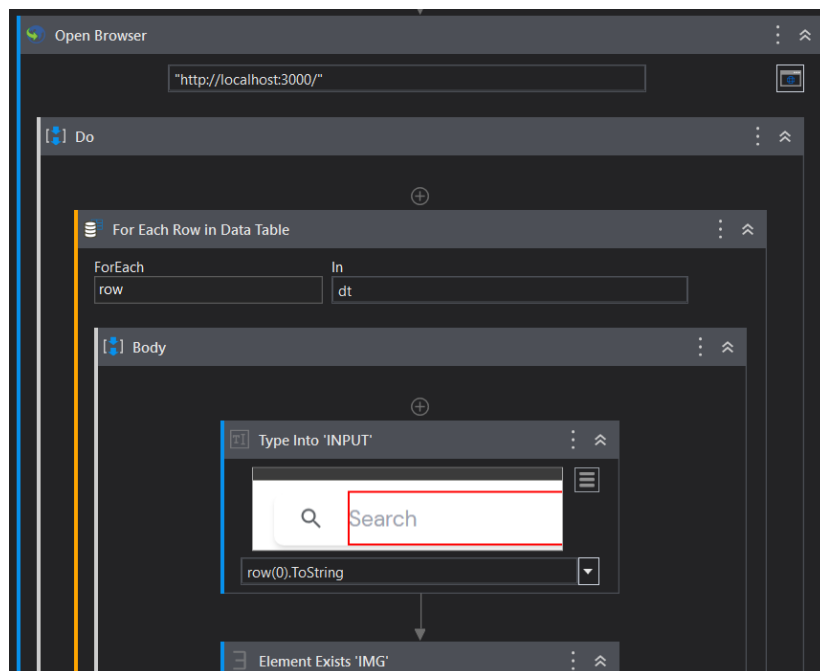
If file of CSV format then we extract the data into a datatable using read CSV activity and merge this extracted data table to the parent unified datatable



After data is extracted from all the files present in the path provided, the datatable holding all the customer information is converted to an excel file

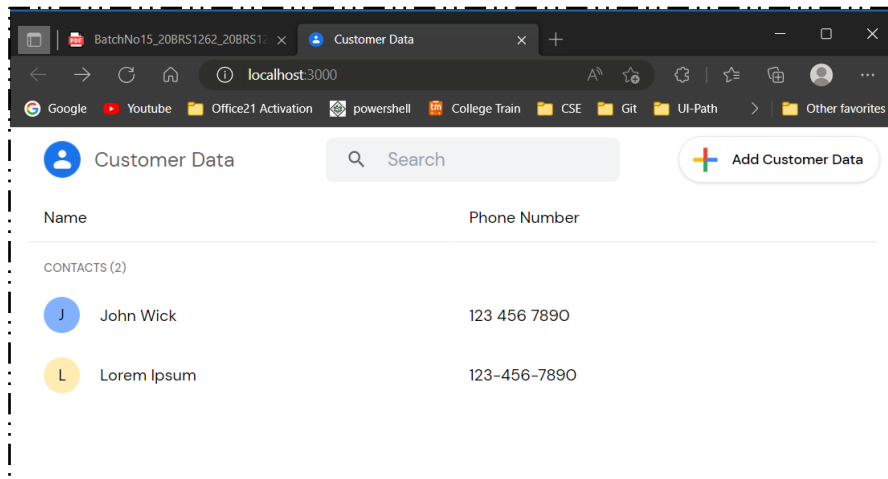


We open a browser and append the data present in the unified excel file to the company's website. Firstly, customer data is extracted row-wise and appending before appending the activities checks whether the customer is already registered in the website by searching in the name in the website search bar and identifying if it has found any match using Find Element activity, if not found then it creates a new customer contact.

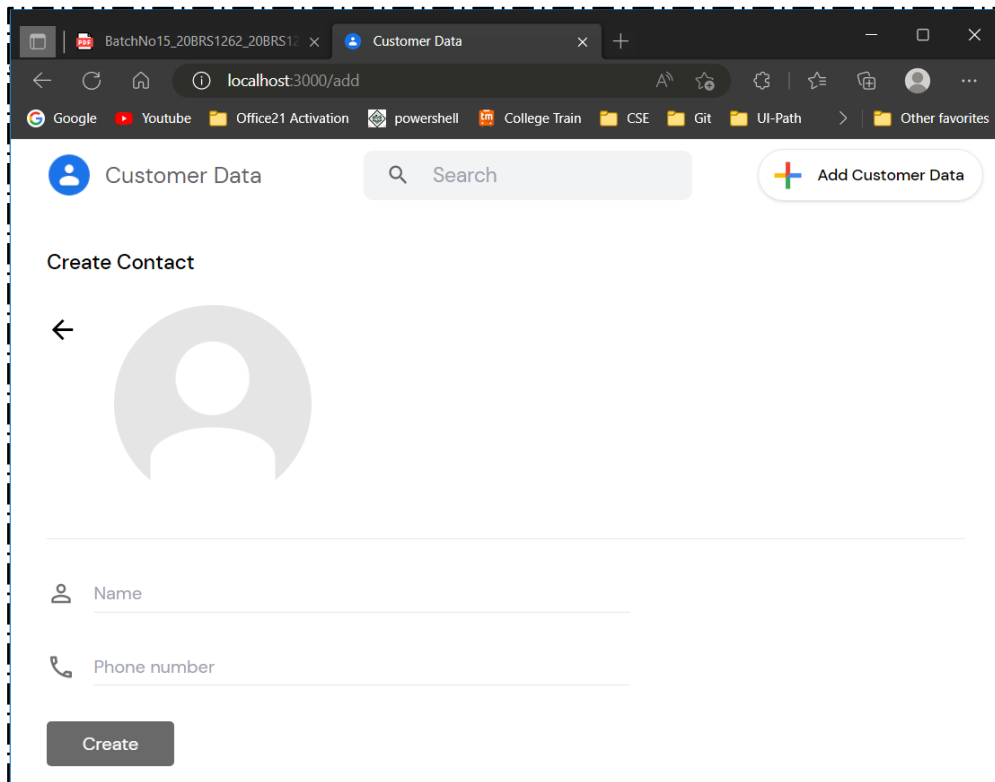


## Website

### Home Page (Initial State)

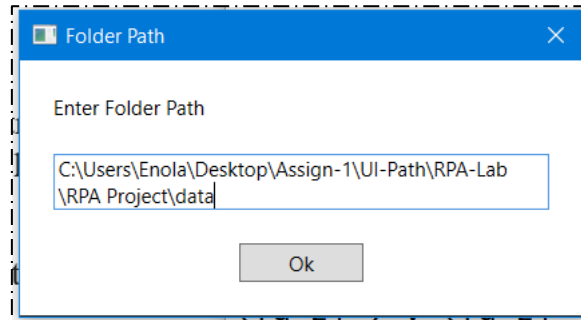


### Create Customer Page



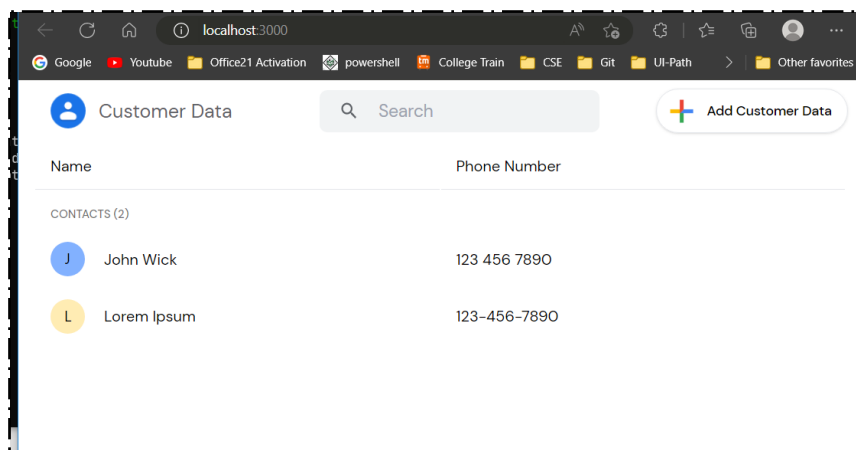
## Workflow in action

Step 1: Taking the folder path where the customer data is stored

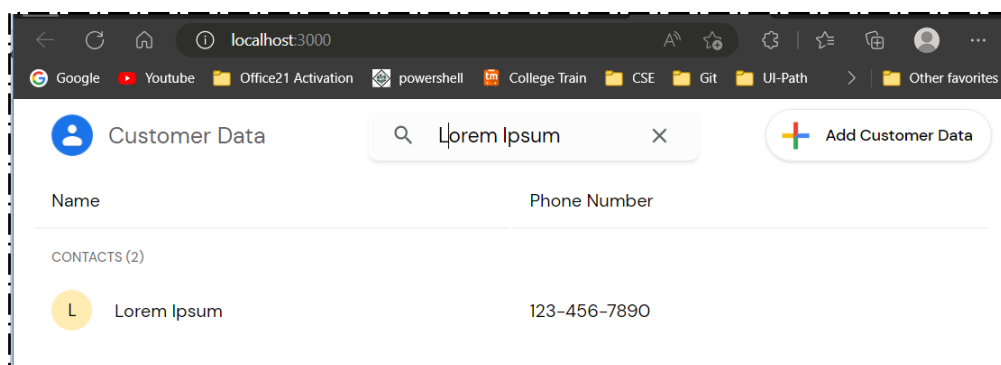


Step 2: The workflow loops through all the files present in the path and extracts data from all the files and stores it a common unified excel file

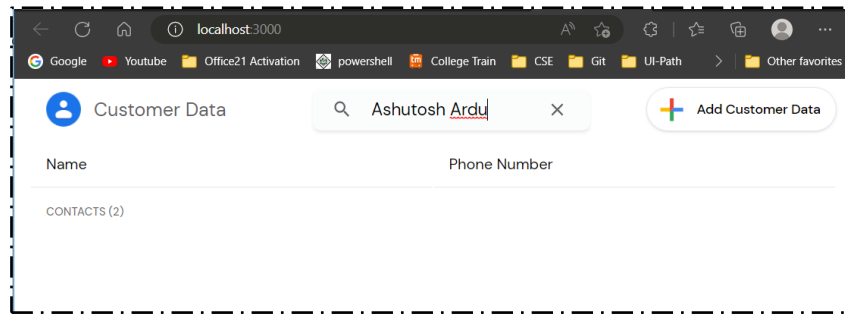
Step 3: The workflow now opens the web portal and loads the customer data onto the server/database.



If the customer entry is already present it continues on to the next data row in the unified excel file



If customer data is not yet entered then it creates a new customer contact



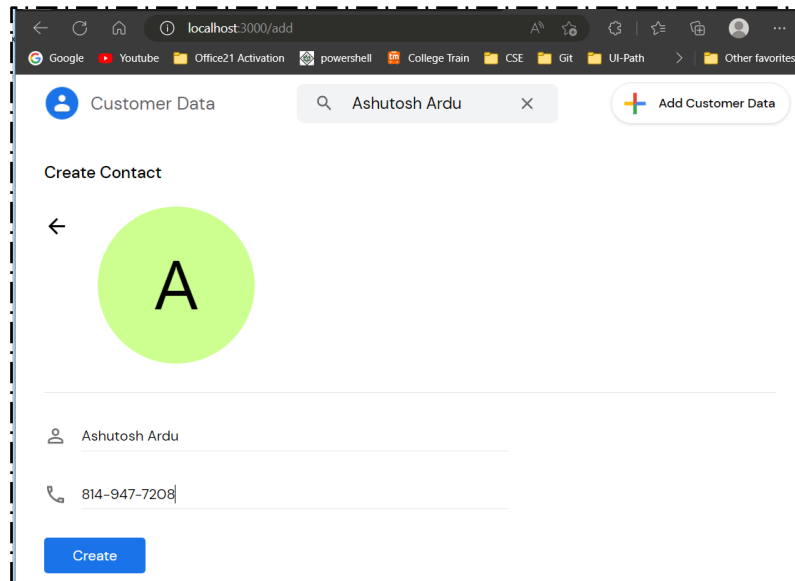
Customer Data

Search: Ashutosh Ardu

Add Customer Data

Name Phone Number

CONTACTS (2)



Customer Data

Search: Ashutosh Ardu

Add Customer Data

Create Contact

←

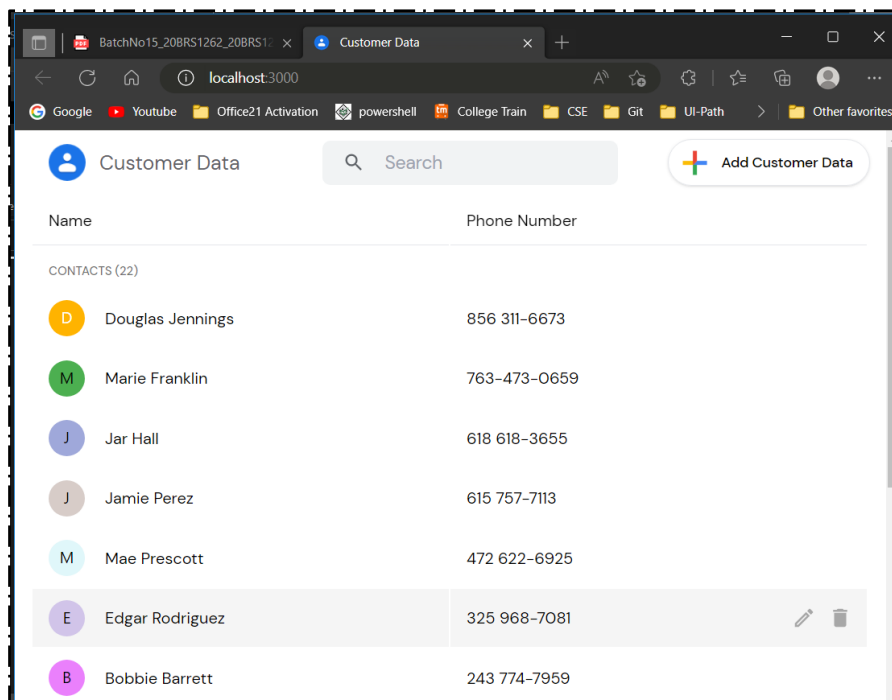
A

Ashutosh Ardu

814-947-7208

Create

Website (Final State after running the workflow)



Customer Data

Search

Add Customer Data

Name	Phone Number
CONTACTS (22)	
D Douglas Jennings	856 311-6673
M Marie Franklin	763-473-0659
J Jar Hall	618 618-3655
J Jamie Perez	615 757-7113
M Mae Prescott	472 622-6925
E Edgar Rodriguez	325 968-7081
B Bobbie Barrett	243 774-7959



## 7 Conclusion

Data migrations have been present ever since IT solutions were introduced and extensive research has been performed in order to further optimize the migration approach. Nevertheless, the data migration phase remains an activity with many challenges which, to be avoided, require appropriate attention. By means of this project we have attempted to provide insight into the main areas of attention that need to be managed by the project in order to avoid a failing migration. The journey to the successful Data Migration is not an easy one, but if done well, it can have many benefits. However, without the correct precautions and visibility, you can easily end up with suboptimal performance or infrastructure spend. Data migration is a necessary process, but not an impossible one to conquer. The key is to prepare for it very early on, and continue to monitor data migration throughout the life of the project. Project timelines tend to become more rigid as time passes, so it really makes sense to meet migration head on. A devoted team with a clearly defined project plan from the inception of the project, armed with automated tools where applicable, is indeed the formula to success.

## 8 References

1. [OCR Engine a Detailed Wiki by Wiki Grooper - Documentation](#)
2. [Tesseract OCR Documentation by UIPath - Documentation](#)
3. [Solutions to Data Migration Problems by Katie Behrens - A UIPath Blogger](#)  
[A deeper dive into Data Migration by Simanta Shekhar Sarmah - Research Paper](#)
4. [Data migration: A theoretical perspective by Qing Wang - A Research Paper](#)
5. [Extraction data from JSON reference by Deepak Rai - A Video](#)
6. [Image text extraction Reference by Marcelo Cruz - A Video](#)
7. [Website using Reactjs reference by Abhisheik G - A Video](#)
8. [Data Migration in Cloud by Mehreen Ansara, Muhammad Ashraf and Mubeen Fatima - A Research Paper](#)
9. [Future Scope of Data Migration by Arif Iqbal and Ricardo Colomo-Palacios - A Research Paper](#)
10. [Image Processing Using UIPath by UIPath - A Documentation](#)
11. [Extracting Data from PDF by UIPath - A Documentation](#)
12. [Backend for React JS using Express - A Video](#)
13. [React Js Docs by React Js - Documentation](#)