.

Creating a Data Management system 1.0

PURDUE University, Fort Wayne
Computer Science Department - Software Engineering
Prof. Dr. Venkata Inukollu

Team 6: Deepika Rajashree Penuballi, Sai Swetha Reddy, Likitha Yellinedi
Client: Venkateshwar Madasu

Table Of Content

Introduction

In the current day job market, there is a high need for data analysis due to the large scope of the roles available at a firm, technologies used, experience, client requirements, etc. This leads to the need for Data analysis as it could be tedious work to filter out a profile for a given role – as each organization can use a similar but slightly different naming convention for its titles. Our project works on different hiring platforms from the perspective of a User trying to find a job in the IT industry and a hiring official who wants to research the current job market. It performs web scraping on hiring websites – it extracts the data based on multiple factors like geographical constraints, technologies, experience, Domain, if it permits salary margin, etc.

This extracted data is then transformed(filtered) upon our requirement, we try to categorize the roles based on certain factors as mentioned above and provide a report with our analysis which would be reliable to the user – this would make the task of finding a job based upon multiple factors instead of searching for a job role with a title the user is aware of, user can now know the more options available at the job market for the same languages used, technologies, requirements, day to day tasks at their desired location and other factors. The report is generated based on User input – the user is given a series of options to select from on which factors the transformations are performed. Users would be able to select the website they want to perform the job search, experience, skill sets, etc.

To be precise, this project aims to create an ETL tool from a student/job finder perspective to make it easy to find the options available for them in the given market. We also try to provide an analysis of the overall job market depending on the same factors and provide a report.

Objective

The objective of this project is building a real time ETL Project from scratch – Web scraping data, performing data transformations, and loading into source file which later user to perform analysis and finally creating a dashboard for Data Visualization.

E - Extract the data using Web scraping
T - Transform the data using NLP techniques.
L - Load/Save final data as CSV for data analysis

- Job Seekers/Users can have more options to apply from – as similar Job roles among different organizations are grouped under a single title.

- Performing Data Analytics using Tableau – Dashboard/ Data reports are created from the transformed data. Using Tableau for data visualization, can also connect to the data sources, dashboards, and visualizations, and share them with users.

- HR officials can perform analysis using these reports on the current day job market and various other factors.

Project Constraints

• **Time**: Given the course time frame, this would be a major constraint of the project. There is certainly a limit on what one can achieve considering the requirements, and scope of the project.

• **Scope**: The scope of the project determines the detail that would be needed in the project deliverables. We expect certain limitations as the scope increases, we have limited our project to one website due to constraints surrounding web scraping.

• **Risk**: Web scraping should be done carefully as it deals with extracting real-time data. Web Scraping can be limited due to website privacy/security policies. We need to consider the legal limitations and analyze them to find more techniques and know our limitations while scraping websites.

• **Resources**: The technologies indeed have licenses or sometimes even limitation with respect to the OS. For example, Power automated desktop is limited to Windows OS. And this tool constitutes to 50% of the project. Similarly, Tableau Desktop License is available for Purdue students with a year subscription. Later must be subscribed with certain amount.

• **Data:** Data loss or data inconsistency and delay of regular tasks.

Requirements Glossary

Source Data: For our Project, we choose to perform web scraping instead of using the existing Datasets. Our project works on different hiring platforms from the perspective of a User trying to find a job in an IT industry and a hiring official who wants to research the current job market. It performs web scraping on hiring websites – it extracts the data based on multiple factors like geographical constraints, technologies, experience, Domain, if it permits salary margin etc. Certain websites do have constraints and are not considered legal to perform web scraping. So, we tried to choose the websites that were not causing a problem in extracting ethical data.

Web scraping: The process of scraping and importing information obtained from a website to a spreadsheet is known as web scraping. Data scraping assists in obtaining data from the web and transferring that data into human-readable output.

Extracting the Data: By using Power Automate Desktop, web scraping is done easy by its built-in HTTP connector. It connects to the website and retrieves the web page. It then performs data parsing actions to extract the required data from the web page. We installed a plugin that supports power automation to record all our actions on the web page, showing relevant data to be extracted on a page. By selecting the HTML content of the page, Power Automate automates the process of retrieving similar data across the website on multiple pages.

Transforming data: ETL techniques are used in data transformation. These database approaches are integrated into a single medium for retrieving data from one data store and storing it in another. The extract is a method of collecting data from a database. During extraction, data is acquired from so many other sources.

CSV file: This format is commonly used to import and export data from spreadsheets and databases.

Data Modeling*:* Describes how data flows via an application or group.

Data Reporting*:* The process of gathering and submitting data that leads to correct evaluations of the facts is known as data reporting; faulty data reporting can lead to grossly ignorant decision-making based on misleading information.

Handling data*:* Data Handling is used to manage and portray data in a systematic manner, which is especially crucial when the data presented is huge and complicated. Processing data in a way that makes it easier for people to interpret and comprehend the information provided. As a result, data management refers to the act of gathering, recording, and expressing data in the form of a diagram or graph for convenience.

Tableau: Tableau's approach to data management varies from existing approaches in that it displays information and integrates management operations into the Tableau analytic platform, where people are actively analyzing.

Technical Requirements

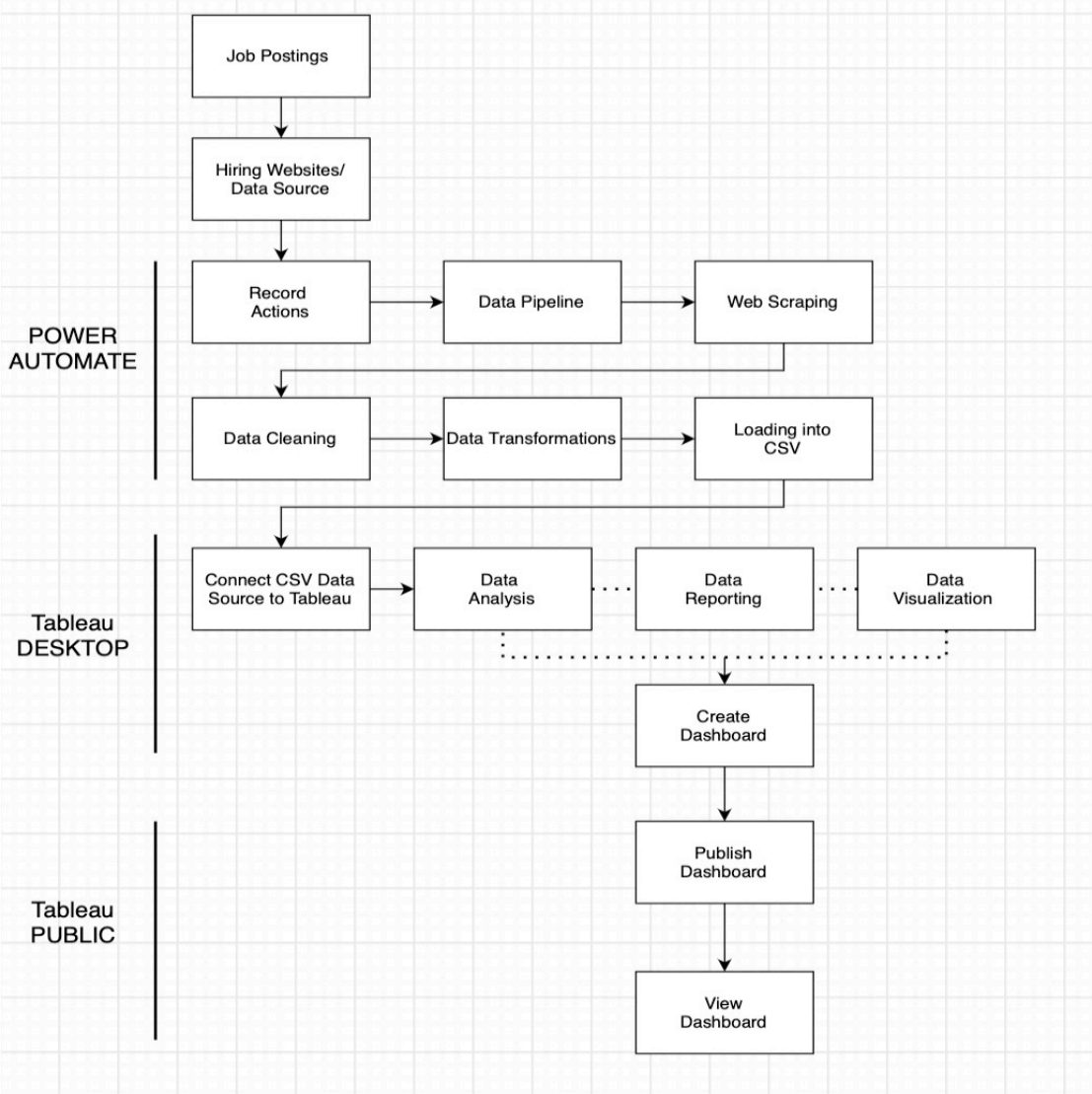
Power Automate Desktop: This tool lets us perform all the ETL actions along with performing web scraping while extracting data from web sources.

Microsoft Office: Data is stored and handled in Excel.

Tableau: This tool helps to understand and visualize data. Creating reports and dashboards made it easy with no need of coding.

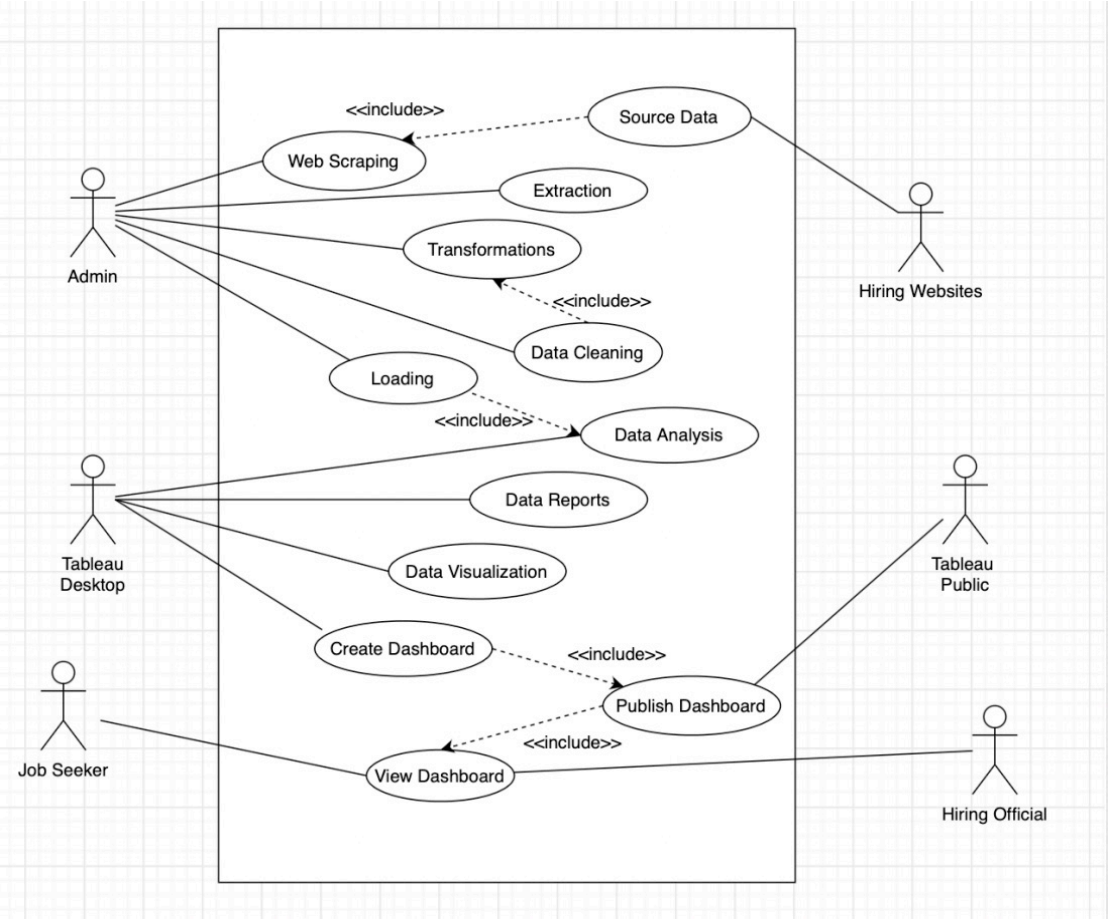
GitHub: Software development hosting service that uses GIT. It will be used to upload project files and collaborate with team members across different project versions.

High Architecture diagram



As per the client requirement, we have not published the dashboard on tableau public.

Use Case Diagram



Actors/Users:

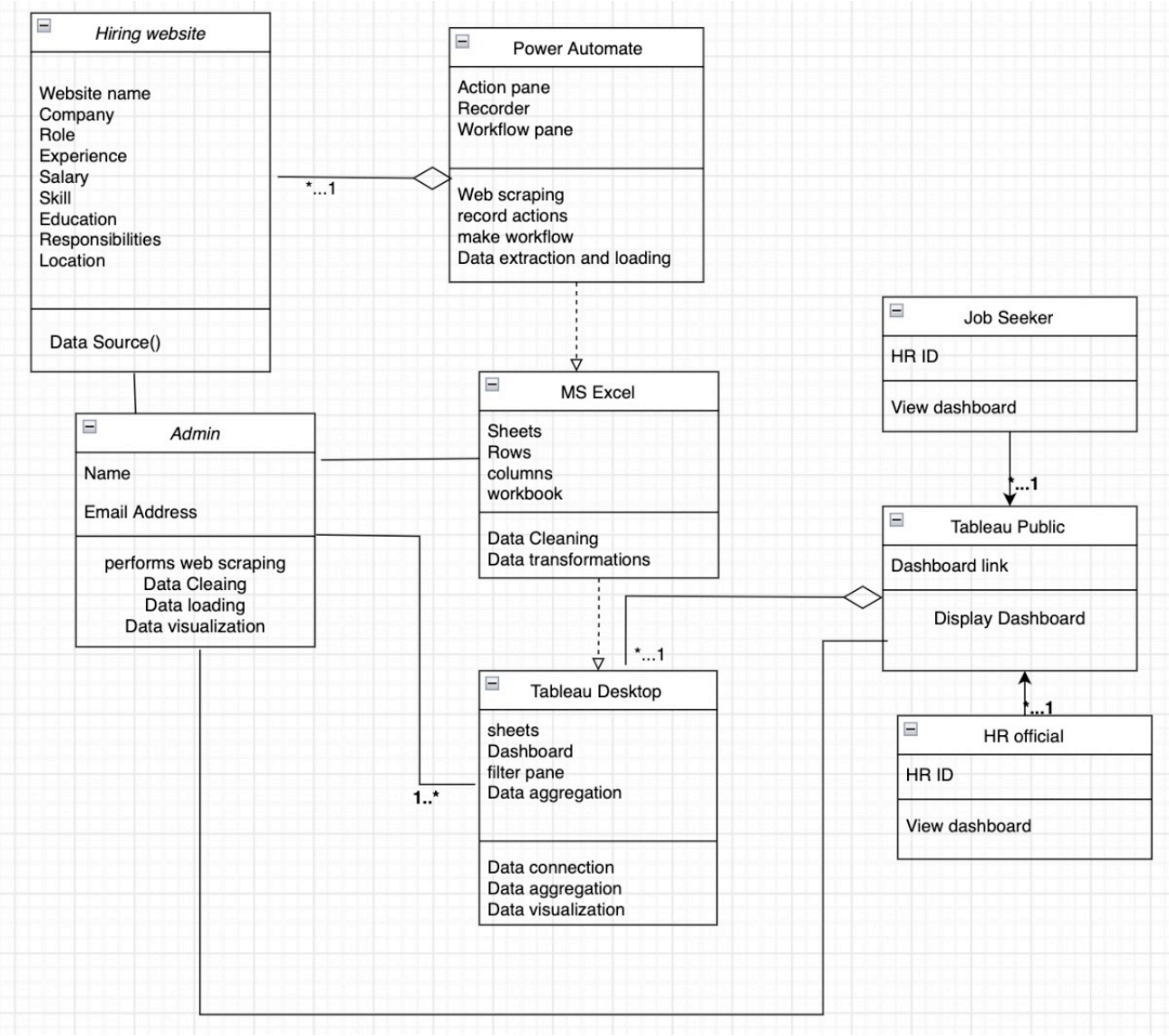
- *User* – Student/Job finder: Any prospective job finder who needs to know all the options available under certain factors – that match the job description but has a different job title among different organizations.

- *Hiring official* – Agent/HR: To give a report based on the same factors on the current day Job market.

- *Hiring Websites* – LinkedIn/Indeed: Any job hiring website that allows us to perform web scraping at a legal level.

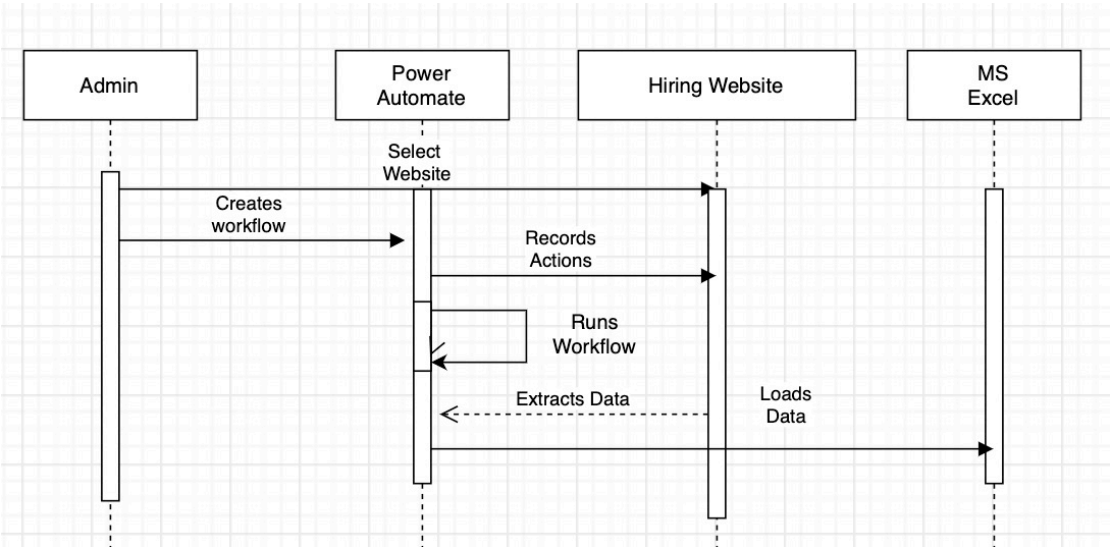
- *Admin*: Uses Power Automate to perform Web Scraping and ETL process. Generates reports/dashboard using Tableau.

Class Diagram

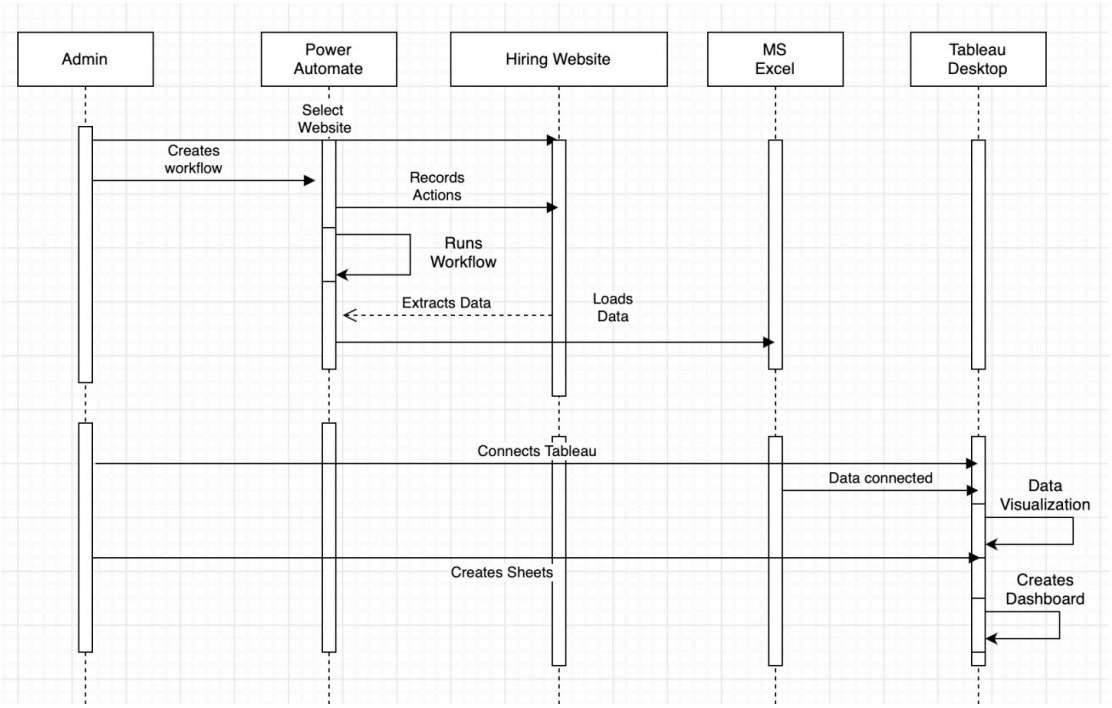


Sequence Diagram

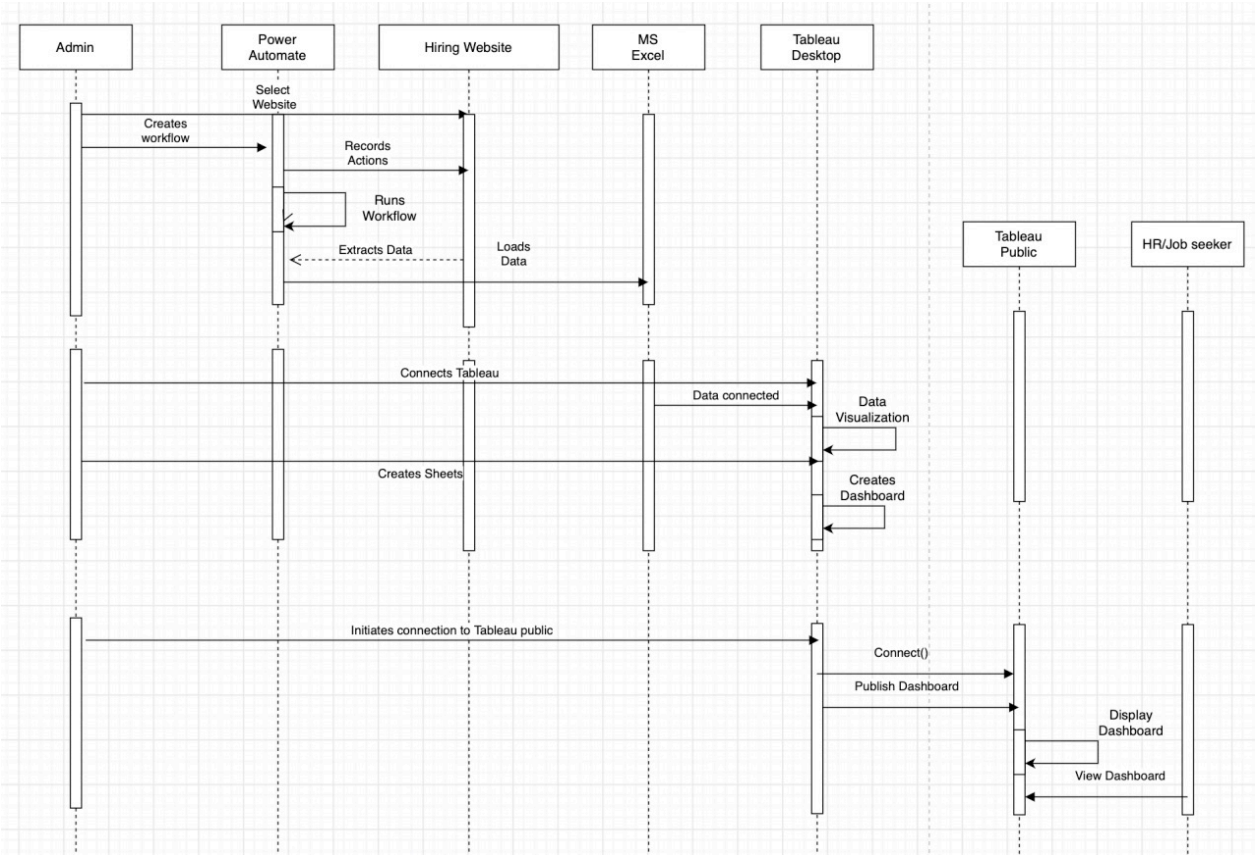
Use Case -1: Web Scraping data from Source website and loading data into excel.



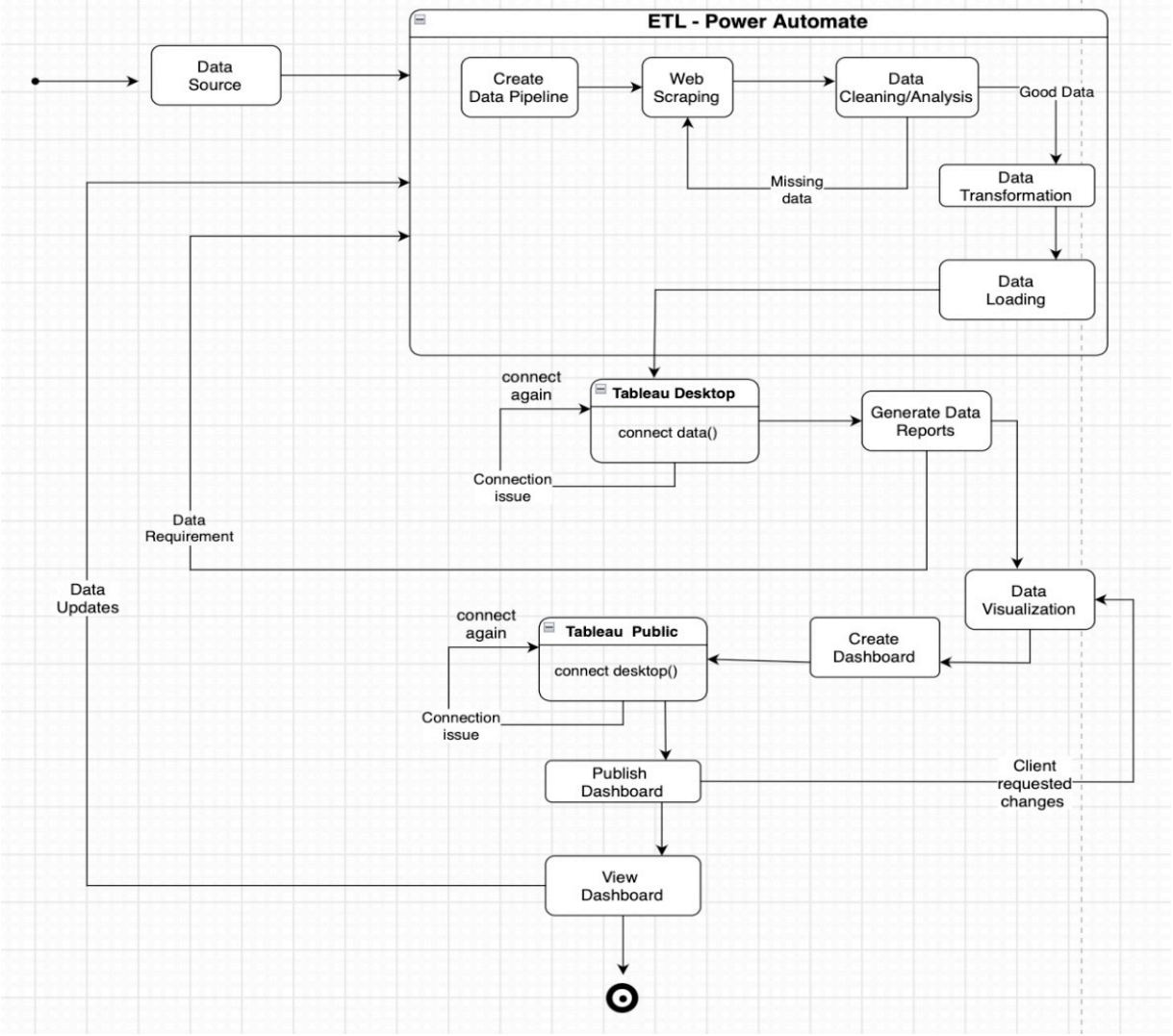
Use Case -2: Data visualization using Tableau and creating Dashboard.



Use Case – 3: Job seeker/HR Gets access to Dashboard.



State Chart Diagram

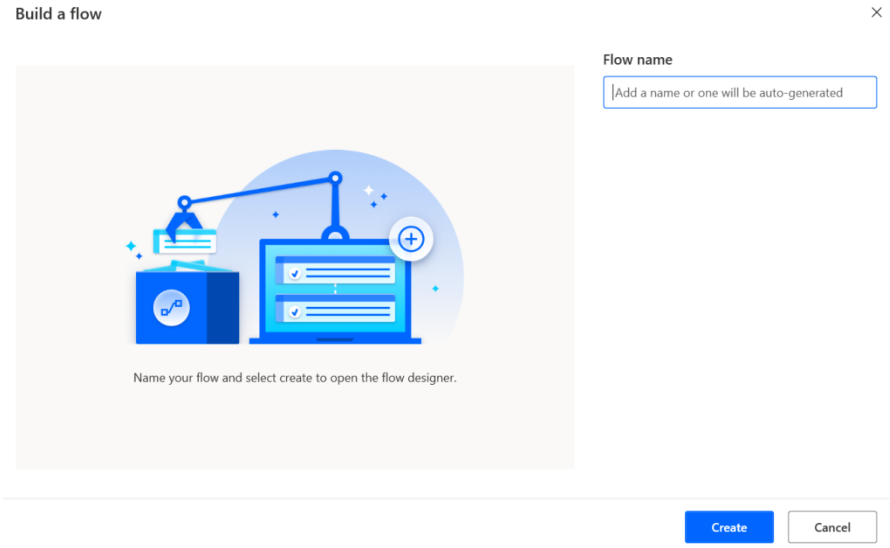


Power Automate

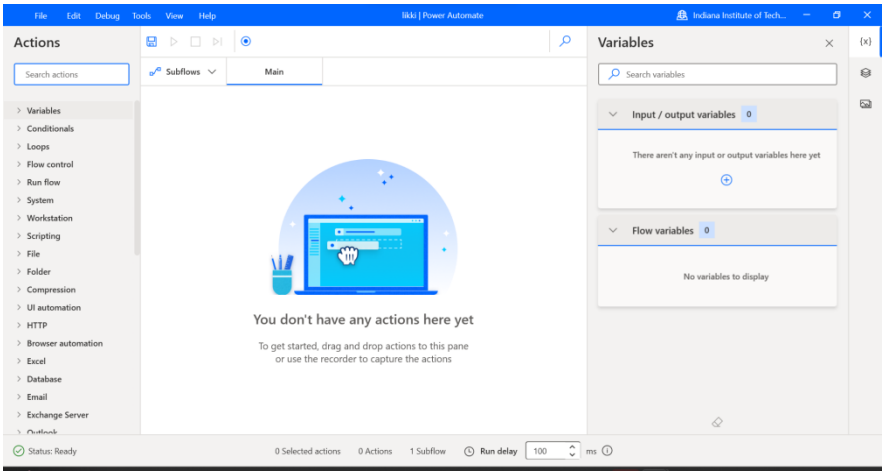
Power Automate working:

https://drive.google.com/file/d/1FvR3UmETiStNEysrlSZBy5eF3Oh5i143/view?usp=sharing

The first time that you're opening it up, you're going to see something like figure below. Click on New Flow, this is going to create your first automation.



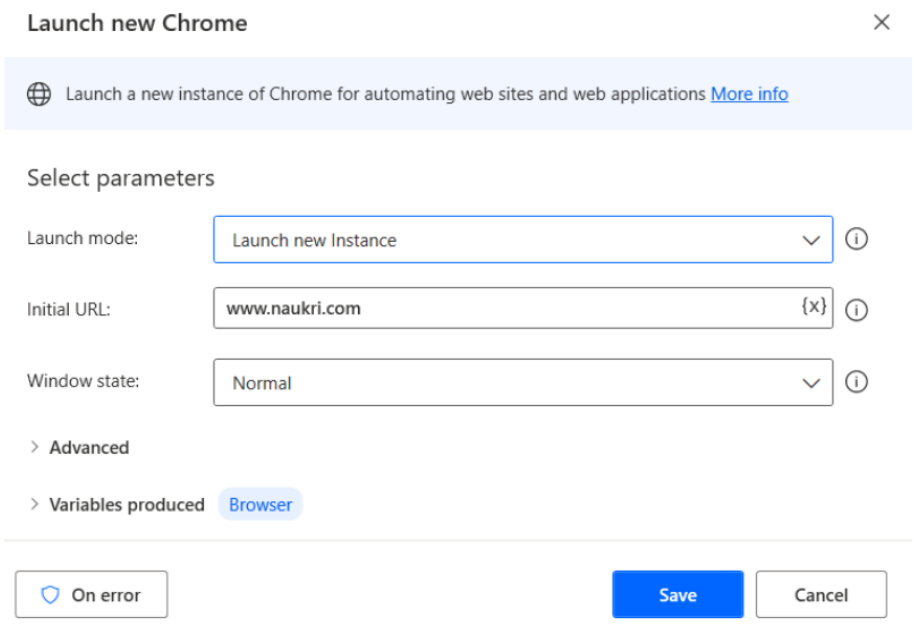
Power Automate is going to bring up the designer, and it's going to take a few seconds for it to get things ready for you. Once it does that, we are going to see this view on the left side. We have the Actions panel. Actions are the primary building blocks of any of your desktop flows.



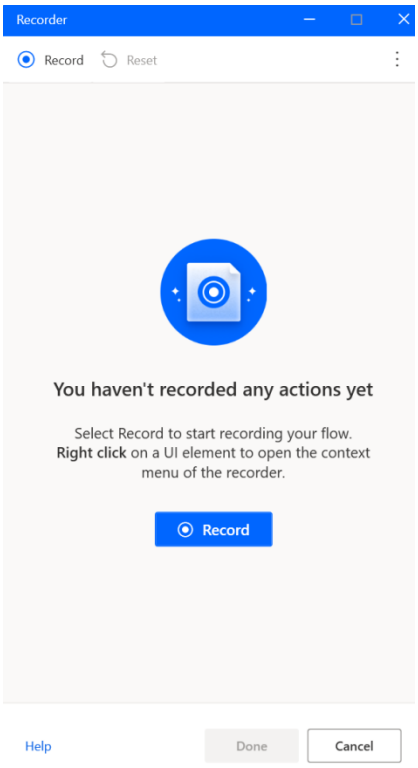
Browser Automation:

Here we can launch a new browser, so depending on which browser you use, make your selection. We are using Google chrome, so we are going to select launch new chrome. We want to navigate to the "Naukri" website. Then, we can select the Window state if it needs to be normal, maximized, or minimized.

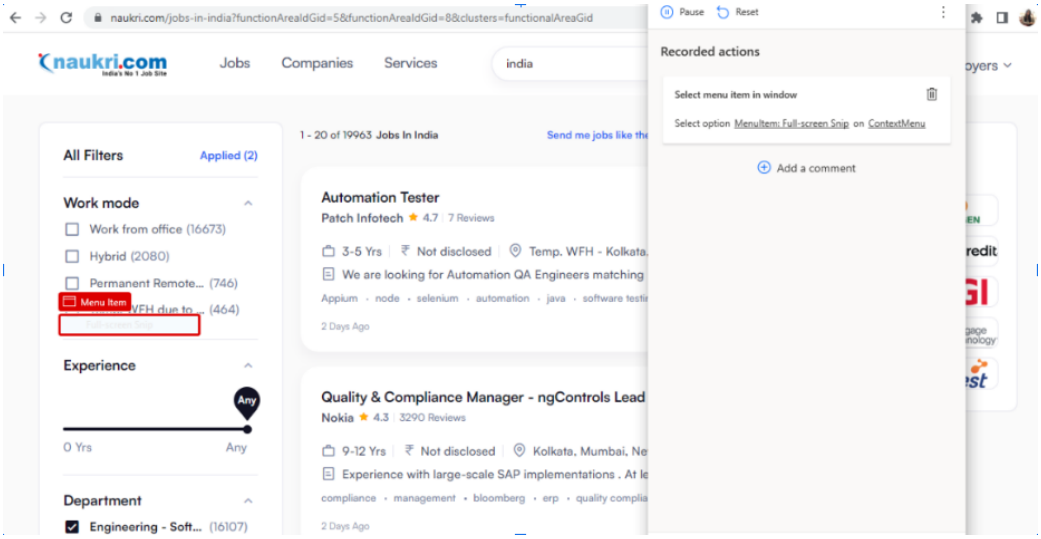
We're going to stick with normal. Notice here on the bottom, it tells us the variable that was produced called Browser. So, this is an automatic name that it picked up. You can rename this if you want for naming convention. But this way, you can always refer to this action.



Use Recorder to Create Flow of Actions. This part gets highlighted in red. Just right-mouse-click, you will get a dialog box on the side, and you will have an ability to extract the element value. If you want to extract the text, select text.

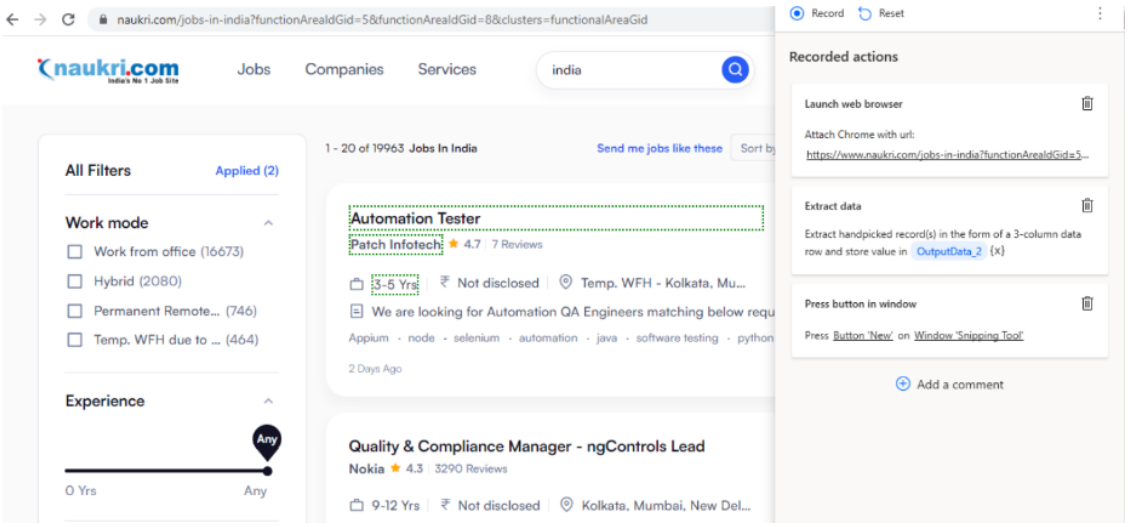


Snippet of Recorder

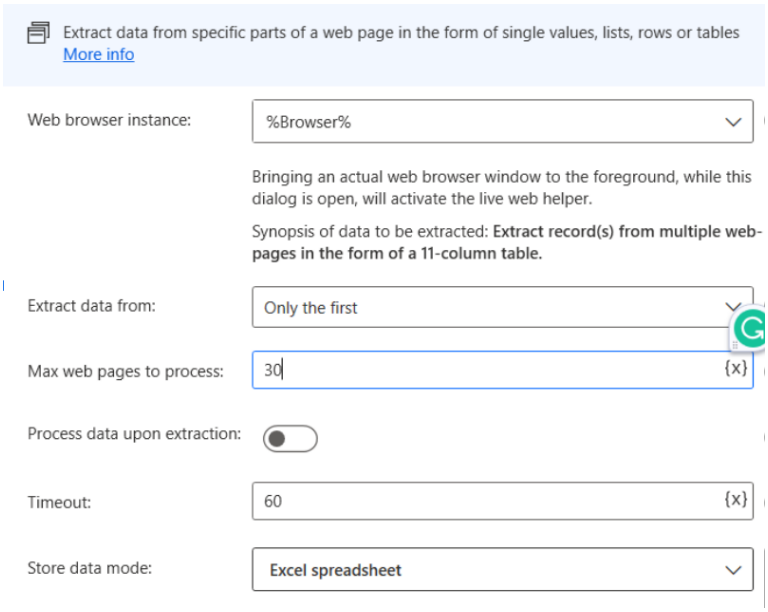


Snippet of Highlighters on the website while recording

Now notice, the moment you did it for the second element, all the other elements on the screen are highlighted in green as well. Power Automate picked up a pattern. This is like Excel's Flash Fill functionality. It picks up a pattern and it applies it to all the elements on this page.



Store Extracted Data in Excel Spreadsheet. I am extracting the first 30 web pages to process.



Data

Transforming the Data: In an ETL Process, transforming the data per requirement is done through data manipulations. For our projects, we needed to optimize the results for the user such that – similar roles with different role names across different organizations must be grouped under one single title - this would increase the scope of finding a job as users now have more options instead of a job role with a title the user is aware of. We also provide analysis on the overall job market depending on certain factors and provide a report.

Data Cleaning: We perform some basic data cleaning operations on the data extracted – trimming, pad the text, handle any null values, parsing and replacing text on the columns using the data cleaning options available on Microsoft Excel.

Columns Fetched	Description of the Column
Job Title	Role name
Role-specific	Multiple job titles under a specific Role
Company Name	Company Name
Minimum Experience	Gives minimum range of experience
Maximum Experience	Gives maximum range of experience
Minimum Salary	Gives minimum range of Salary
Maximum Salary	Gives maximum range of Salary
Description	Role Description
Primary Skill	Top skills required for the role
Secondary Skill - 1	Secondary skills categorized into two columns
Secondary Skill - 2	Secondary skills categorized into two columns
City	Job Location
State	Job Location
Mode of work	Hybrid/Remote

Tableau Desktop

Once we have the excel – One of our major user requirements is fulfilled. Later, the same file is exported to Tableau as source for Data Visualization. Connect the Data Source – Excel that is extracted using Power Automate.

We created 19 sheets of graphs for the Data Analysis and the creation of 3 Dashboards. Each of them as explained below:

Sheets:

- *Role Count:* Total Number of Job postings

- *D1 – Distinct Company/Role Count:* Total Number of Companies that have job postings on the website and number of distinct roles.

- *D1 – Min/Max Salary:* Salary Range we are performing the data analysis.

- *D1 – Min/Max Experience:* Experience Range we are performing the data analysis.

- *D1 – Company Overview:* Overview of Positions/Salary and experience a company is offering.

- *D1 – Role Overview:* Overview of Roles/Salary and experience a company is offering.

- *D1 – Location Overview*: Number of positions available at each Location.

- *D1 – Mode of work:* Ration of Hybrid/ Work from Home.

- *Overview Button:* Button used in dashboard to launch Overview dashboard 1.

- *Job Overview Button:* Button used in dashboard to launch Job overview dashboard 2.

- *Company Overview Button:* Button used in dashboard to launch Company Overview dashboard 3.

- *D2 – Role/Position openings:* Overview of Positions/Salary and experience a company is offering.

- *D2 – Role/Primary Skills Analysis:* Analysis of Primary skills for a role.

- *D2 – Role/Secondary skill 1 Analysis:* Analysis of Secondary skill 1 for a role.

- *D2 – Role/Secondary skill 2 Analysis:* Analysis of Secondary skill 2 for a role.

- *D3 – Jobs/location:* Overview of Positions/location a company is offering.

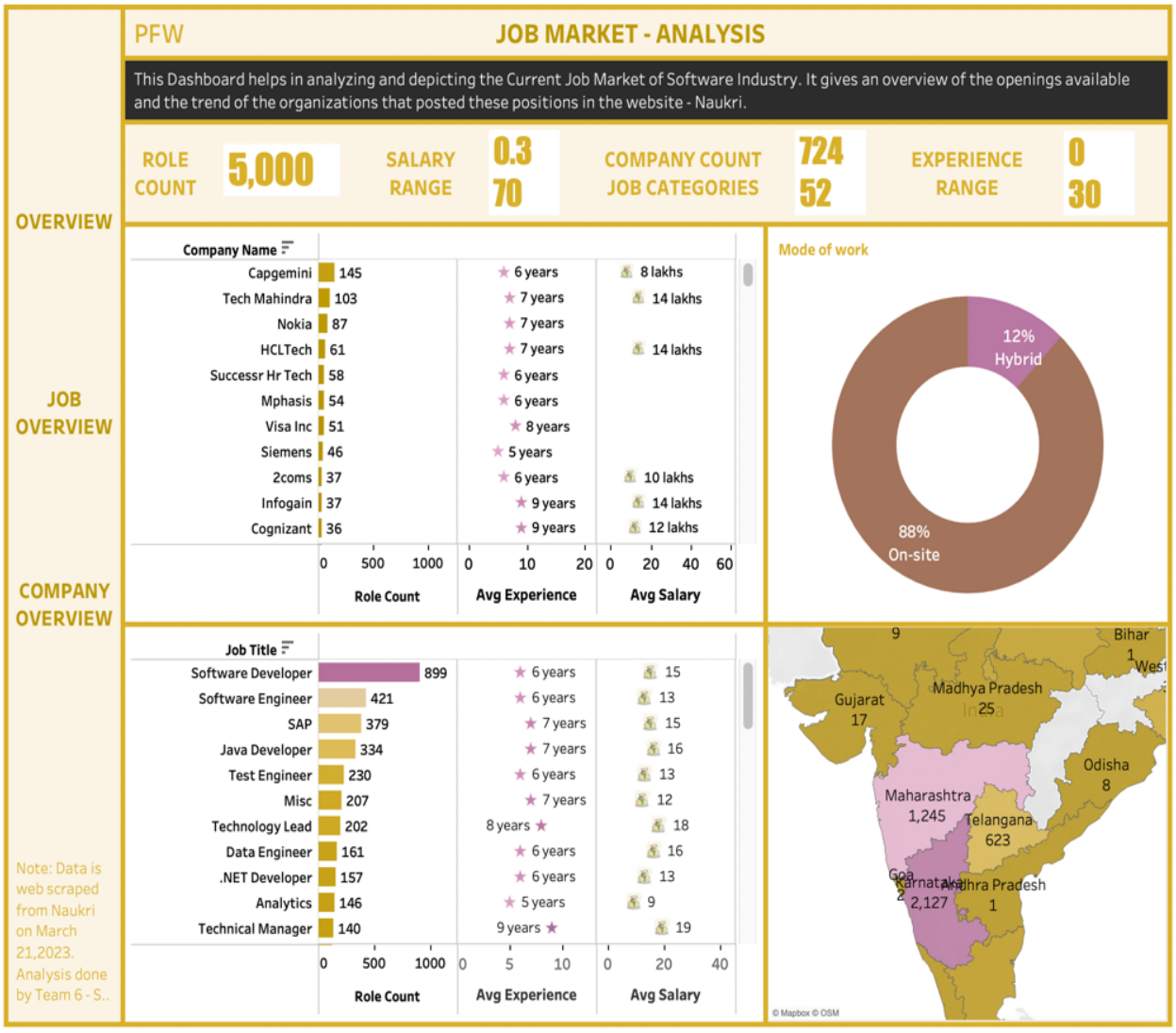
- *D3 – Company/salary disclosed* – Analysis on Salary disclosure of a particular company,

- *D3 – Company/Skill*- Top Skills required in a particular company.

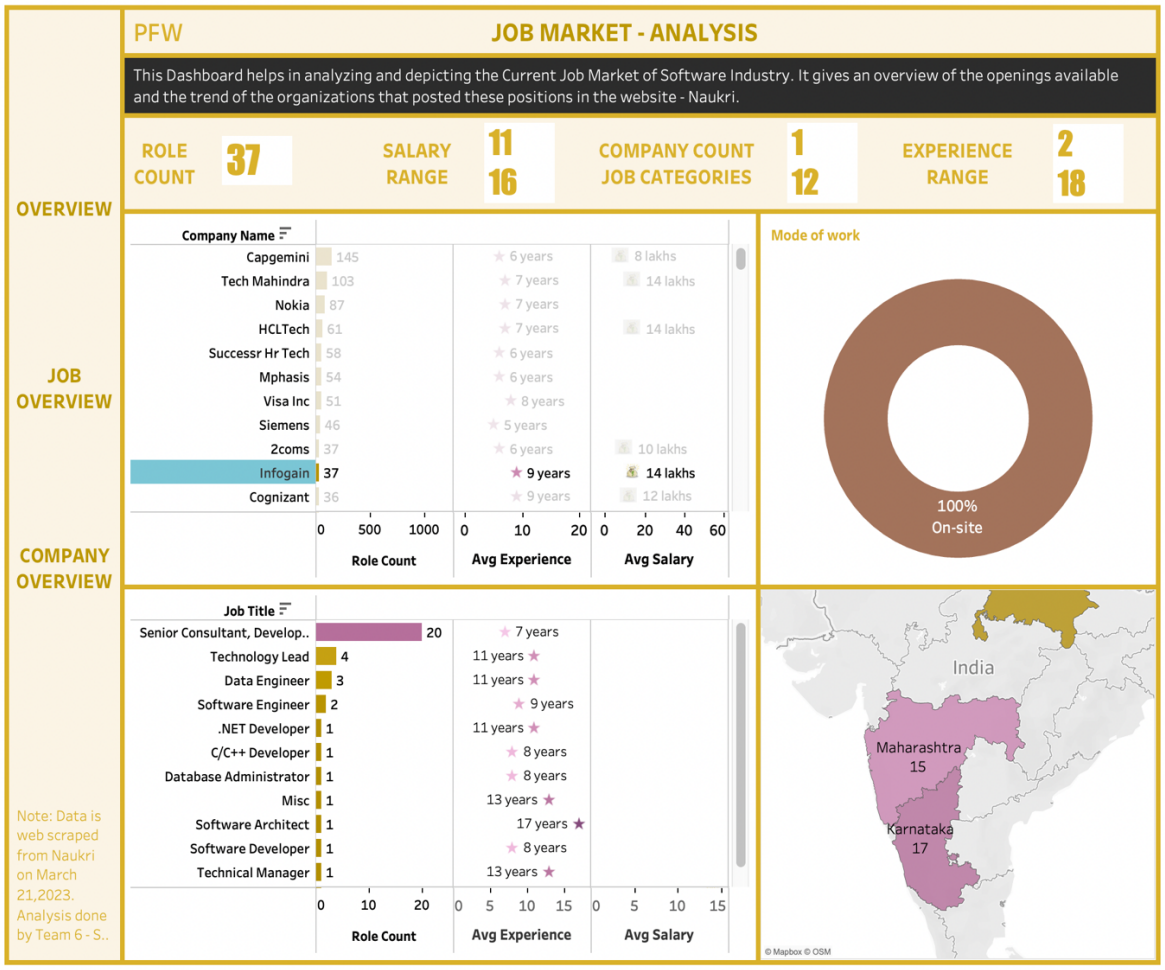
- *D3 – Avg Experience/salary* – Average experience/ salary a company is offering for a given role.

Demo

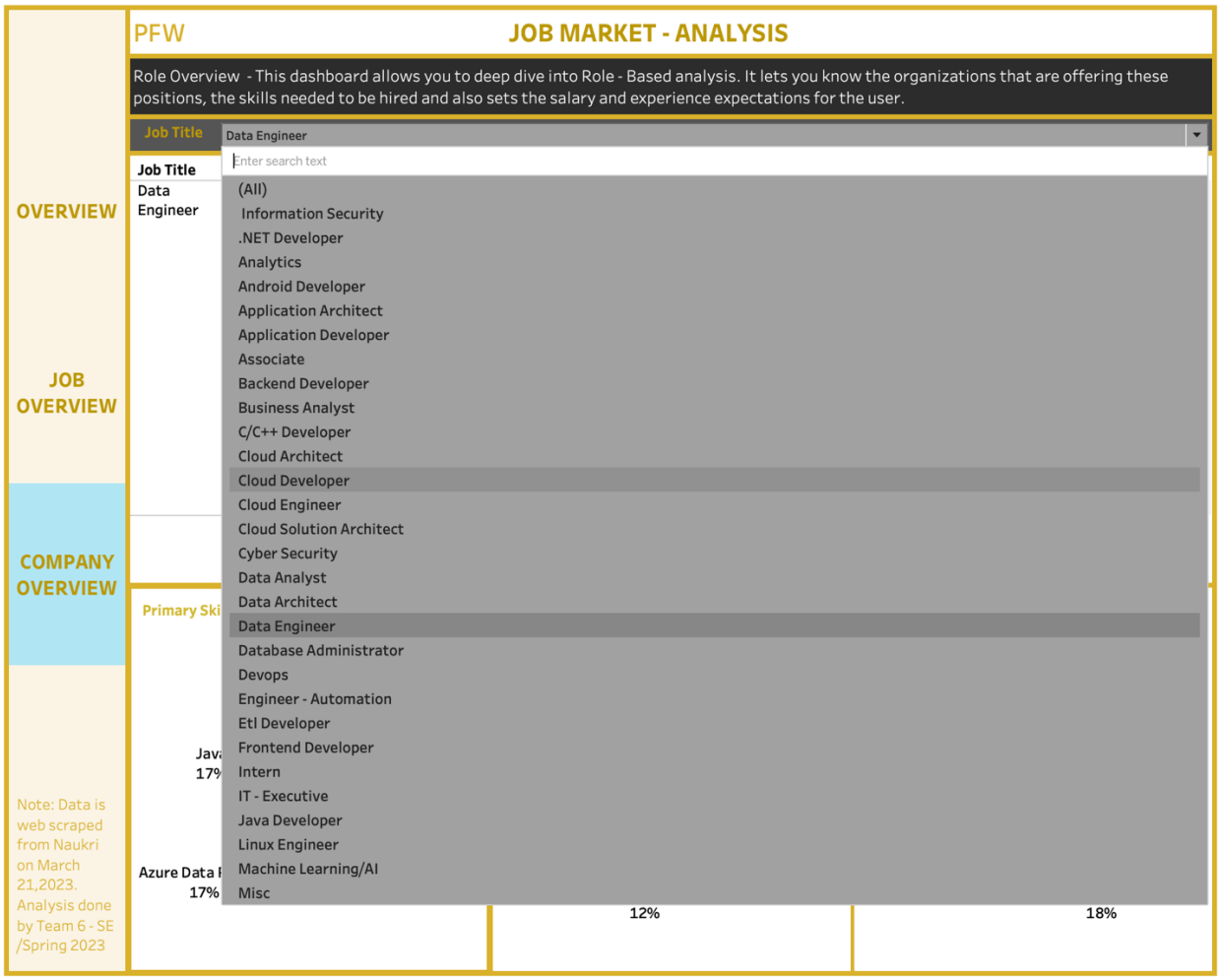
Dashboard 1 – Overview of the Job Market



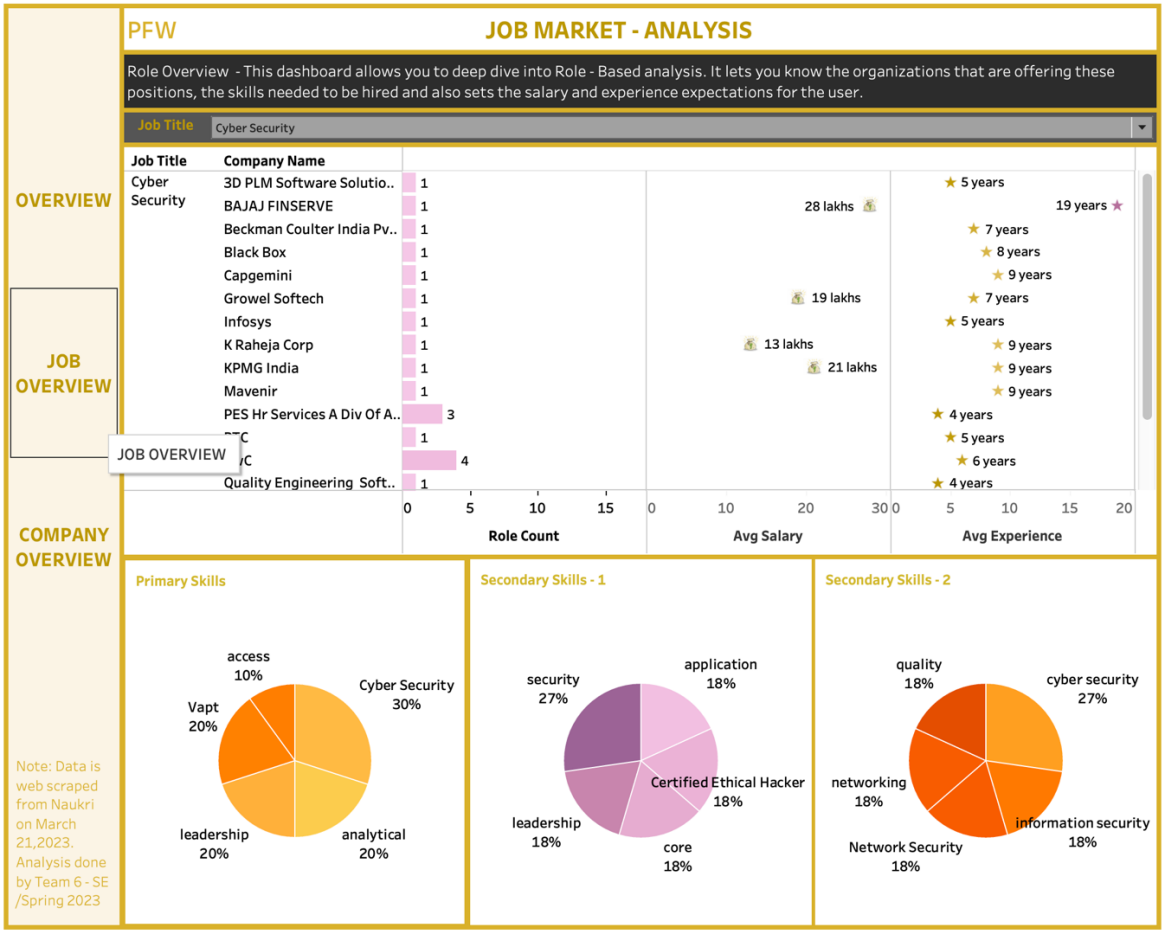
Dashboard 1 – Overview of the Job Market with respect to the Company.



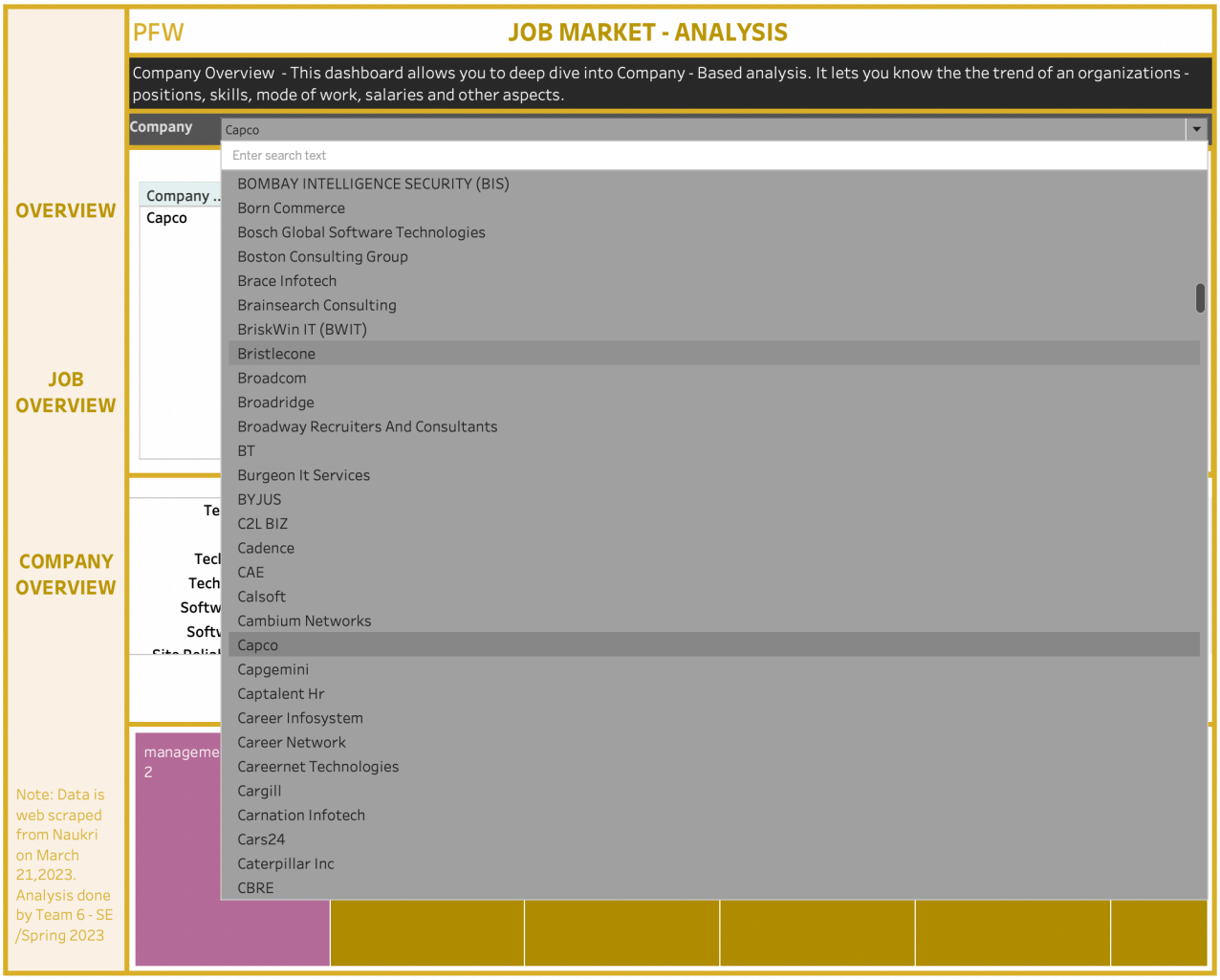
Dashboard 2 – Deep dive of the Job Market with respect to the Role – We have a drop down where you can choose the Job title.



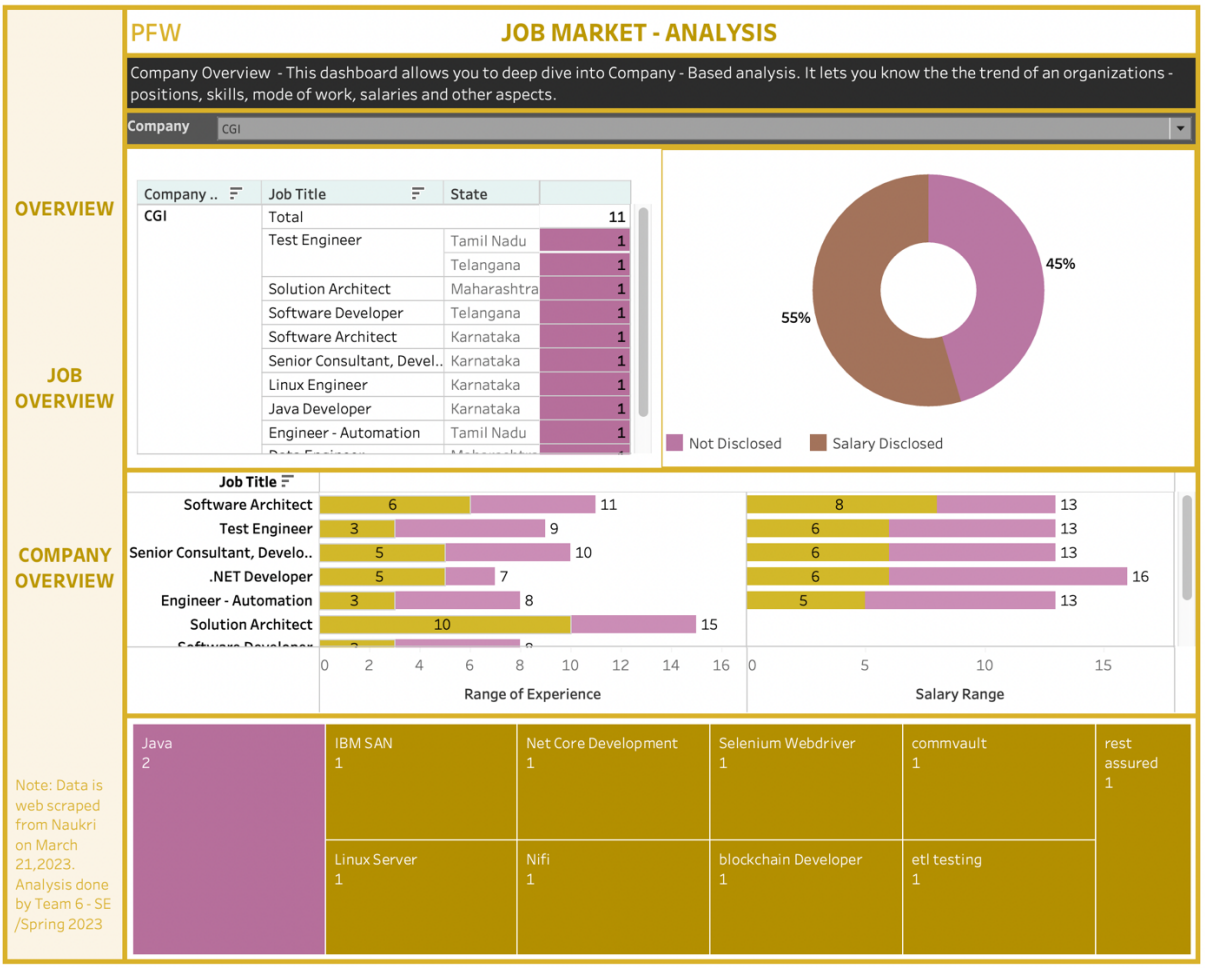
Dashboard 2 – Deep dive of the Job Market with respect to the Role.



Dashboard 3 – Deep dive of the Job Market with respect to the Company – We have a drop down where you can choose the Company.



Dashboard 3 – Deep dive of the Job Market with respect to the Company.



Testing

Test Case ID	Test Case	Pre - Condition	Test Steps	Test Data	Expected Result	Actual Resul	Test Status	Action Required
PA_1	Launch new Chrome website	Website html link	1. Create new Flow 2. Create a Launch New Chrome step and give the website.	https://www.naukri.com/software-jobs?k=software	Launch the website in less than 3 seconds	Website launched in less than 3 seconds	Pass	NA
PA_2	Record Actions on website	Data needed has to be stated.	1. Plug in for recorder installed.	Website layout	All key elements highlighted automatically.	Required elements in naukri hughlighte	Pass	NA
PA_3	Extracting Data on multiple web pages	Recorded actions.	1. Need to have stepd of workflow. 2. Website. 3. Run the workflow.	Recorded steps - columns to be fetched. - given 11 columns	Fetch Data from first 100 web pages.	Failed while extracting	Fail	Web Scraping issue - caused while launching the same website multiple times.
PA_3	Extracting Data on multiple web pages	Recorded actions.	1. Need to have stepd of workflow. 2. Changed Website. 3. Run the workflow.	Recorded steps - columns to be fetched. - given 11 columns	Fetch Data from first 100 web pages.	Extracted data at the first run	Pass	Note - Keep changing the links to avoid errors while extracting data.
PA_4	Save data into excel	Data	1. Create a save excel step in the workflow. 2. Run Workflow.	workflow	Save excel with data	excel saved with data	Pass	NA
TA_1	Connect Data Source	Need Formatted data in Excel	1. Click on connect data source 2. Select Excel 3. Click "Open"	Job Market Excel	Connected and shows columns	Connected and showed columns	Pass	NA
TA_2	Space Requirement	Takes minimum 2 GB	Open workbook and refresh sheets	19 sheets and 3 dashboards created	Running without crashing	Tableau shut down unexpectedly	Fail	Yes - Closed other tasks running on system
TA_2	Space Requirement	Takes minimum 2 GB	Open workbook and refresh sheets	19 sheets and 3 dashboards	Running without crashing	Ruuning as expected	Pass	NA

Conclusion

Creating a Data Management system V 1.0 started with performing web scraping of websites that had job listings. Instead of coding, we used a power tool provided by Microsoft. One of the major advantages using this tool would be that the user does not need to have coding knowledge. One must perform analysis on what data they are exactly looking for and they can proceed further. Tableau has a very keen role in the current day Job Market – It lets Analysis along with visualization. It provides various services starting from workflows to publishing dashboards. For this project – we have used Tableau Dashboard. However, SQL acts as a base in the working of Tableau and thus the user needs to be proficient in SQL while handling large volume of data to perform data visualization.

Overall, this project can be concluded as a Data management system for Job market analysis which is helpful for both the job seekers and HR individuals who needs to have information about the on goings in the market during this period.

Future Work

“Each of us can make a difference”

As it says, everything can be done better or done differently. Below are some of the aspects can be worked on with more time and scope.

- Project can be enhanced with respect to Data; that is, we can work on scraping data from more websites and combining them into single excel.

- Certain Data transformations were done on Excel- which can be automated.

- Automation can be used to periodically web scrape given website with no manual effort.

- More columns can be added to the sheet for better analysis.

References

- https://powerautomate.microsoft.com/en-us/
- https://www.tableau.com
- https://www.youtube.com/watch?v=DgBZiBIgh3w&t=649s
- https://www.naukri.com
- https://www.linkedin.com/
- https://careers.google.com
- https://www.indeed.com
- https://www.edureka.co/blog/web-scraping-with-python/
- https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl
- https://mydataprovider.com
- https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset
- https://arxiv.org/pdf/2005.02780v1.pdf
- https://learn.microsoft.com/en-us/power-
- automate/guidance/planning/transforming-formatting-data
- https://learn.microsoft.com/en-us/training/modules/pad-text-manipulation/2text-actions