# PROFESSIONAL TRAINING REPORT

**at**

**Sathyabama Institute of Science and Technology (Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

## SATYAM PUROHIT

**(Reg. No. 38110512 )**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

**CHENNAI – 600119, TAMILNADU**

**NOVEMBER 2021**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

### (DEEMED TO BE UNIVERSITY)

**Accredited with Grade “A” by NAAC**

**(Established under Section 3 of UGC Act, 1956)**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119**

[**www.sathyabamauniversity.ac.in**](http://www.sathyabamauniversity.ac.in/)

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **SATYAM PUROHIT (38110512)** who carried out the project entitled “**Data Scraping using Blue Prism**” under my supervision from April 2020 to June 2020.

## Internal Guide

## Ms. Yogitha. R M.E.,

**Head of the Department**

**Dr. L. Lakshmanan M.E., Ph.D.,**

**Dr. S. Vigneshwari M.E., Ph.D.,**



## Submitted for Viva voce Examination held on

**Internal Examiner External Examiner**

**DECLARATION**

I SATYAM PUROHIT hereby declare that the Project Report entitled Data Scraping Using Blue Prism done by me under the guidance of Ms. Yogitha. R M.E., (Internal) and Mr. Hemant Gahlot (External) at SMARTBRIDGE HYDERAD is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

SATYAM PUROHIT

## DATE: 11/11/2021

**PLACE: Hyderabad NAME OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing , **Dr. S. Vigneshwari M.E., Ph.D., and Dr. L. Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Ms. Yogitha. R M.E.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

# ABSTRACT

**Data scraping** is a technique where a [computer program](https://en.wikipedia.org/wiki/Computer_program) extracts [data](https://en.wikipedia.org/wiki/Data) from [human-readable](https://en.wikipedia.org/wiki/Human-readable_medium) output coming from another program. Normally, [data transfer](https://en.wikipedia.org/wiki/Data_transmission) between programs is accomplished using [data structures](https://en.wikipedia.org/wiki/Data_structures) suited for [automated](https://en.wikipedia.org/wiki/Automation) processing by [computers](https://en.wikipedia.org/wiki/Computers), not people. Such interchange [formats](https://en.wikipedia.org/wiki/File_format) and [protocols](https://en.wikipedia.org/wiki/Protocol_(computing)) are typically rigidly structured, well-documented, easily [parsed](https://en.wikipedia.org/wiki/Parsing), and minimize ambiguity. Very often, these transmissions are not human-readable at all.

Thus, the key element that distinguishes data scraping from regular [parsing](https://en.wikipedia.org/wiki/Parsing) is that the output being scraped is intended for display to an [end-user](https://en.wikipedia.org/wiki/End-user_(computer_science)), rather than as an input to another program. It is therefore usually neither documented nor structured for convenient parsing. Data scraping often involves ignoring [binary data](https://en.wikipedia.org/wiki/Binary_data) (usually images or multimedia data), [display](https://en.wikipedia.org/wiki/Display_device) formatting, redundant labels, superfluous commentary, and other information which is either irrelevant or hinders automated processing.

Data scraping is most often done either to interface to a [legacy system](https://en.wikipedia.org/wiki/Legacy_system), which has no other mechanism which is compatible with current [hardware](https://en.wikipedia.org/wiki/Computer_hardware), or to interface to a third-party system which does not provide a more convenient [API](https://en.wikipedia.org/wiki/Application_programming_interface). In the second case, the operator of the third-party system will often see [screen scraping](https://en.wikipedia.org/wiki/Screen_scraping) as unwanted, due to reasons such as increased system [load](https://en.wikipedia.org/wiki/Load_(computing)), the loss of [advertisement](https://en.wikipedia.org/wiki/Advertisement) [revenue](https://en.wikipedia.org/wiki/Revenue), or the loss of control of the information content.

|  |  |  |
| --- | --- | --- |
|  | TABLE OF CONTENTS |  |
| **CHAPTER No.** | **TITLE** | **PAGE No** |
|  | ABSTRACT | i |
| **1** | **INTRODUCTION** | **1** |
| **2.** | **AIM AND SCOPE OF THE PRESENT INVESTIGATION** | **2** |

**3. EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED 3**

|  |  |  |
| --- | --- | --- |
| **4.** | **SUMMARY AND CONCLUSIONS** | **5** |
|  | **REFERENCES** | **6** |
|  | **APPENDIX** |  |
|  | SCREENSHOTS | 7 |
|  |  |  |

# INTRODUCTION

Data scraping, also known as [web scraping](https://en.wikipedia.org/wiki/Data_scraping), is the process of importing information from a website into a spreadsheet or local file saved on your computer. It’s one of the most efficient ways to get data from the web, and in some cases to channel that data to another website. Popular uses of data scraping include:

* Research for web content/business intelligence
* Pricing for travel booker sites/price comparison sites
* Finding sales leads/conducting market research by crawling public data sources (e.g. Yell and Twitter)
* Sending product data from an e-commerce site to another online vendor (e.g. Google Shopping)

And that list’s just scratching the surface. Data scraping has a vast number of applications – it’s useful in just about any case where data needs to be moved from one place to another.

# AIM AND SCOPE OF THE PRESENT INVESTIGATION

## 2.1 AIM

There are many ways that businesses use data scraping to their advantage. In almost any case where there is a large body of information, data scraping can function as a way of collecting this data and getting it into useful formats. For instance, in a variant of data scraping called web scraping, a company may take an enormous volume of information from a document or file and format it into an Excel spreadsheet for later use. For example, if there are thousands of real estate listings on the website, a real estate company could scrape that data from the web and get it into a useful format for outreach or classification purposes.

## 2.2 SCOPE

Whether or not you intend to use data scraping in your work, it’s advisable to educate yourself on the subject, as it is likely to become even more important in the next few years.

There are now data scraping [AI on the market that can use machine learning](https://www.targetinternet.com/digital-marketing-insights-5-encounters-with-machine-learning/)to keep on getting better at recognising inputs which only humans have traditionally been able to interpret – like images.

Big improvements in data scraping from images and videos will have far-reaching consequences for digital marketers. As image scraping becomes more in-depth, we’ll be able to know far more about online images before we’ve seen them ourselves – and this, like text-based data scraping, will help us do lots of things better.

Then there’s the biggest data scraper of all – Google. The whole experience of web search is going to be transformed when Google can accurately infer as much from an image as it can from a page of copy – and that goes double from a digital marketing perspective.

If you’re in any doubt over whether this can happen in the near future, try out Google’s image interpretation API, [Cloud Vision](https://cloud.google.com/vision/)

# EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED

## 3.1 Blue Prism

Blue prism is a UK-Based Software Company and is one of the leading robotic process automation tools. It is used to automate mundane tasks such that they could operate without any manual intervention. Blue prism has gained edge over its competitors as it has better security, flexibility, scalability, compliance, and resilience.

## Pre-requirements for Blue Prism

The following are the prerequisites for the Blue Prism. It is the only software which

* Creates and supports a digital workforce of industrial strength and enterprise scale
* Does not require IT skills to implement
* Can be implemented in sprints of 4 to 8 weeks (Start to finish)
* Is very low cost compared to the TCO of alternative solutions
* Provides tremendous payback with self-funding returns and an ROI that has been as high as 80%
* Can be managed within IT infrastructure and processes

## 3.2 MySQL

A [relational database](https://en.wikipedia.org/wiki/Relational_database) organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an [operating system](https://en.wikipedia.org/wiki/Operating_system) to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

MySQL is [free and open-source software](https://en.wikipedia.org/wiki/Free_and_open-source_software) under the terms of the [GNU General Public License](https://en.wikipedia.org/wiki/GNU_General_Public_License), and is also available under a variety of [proprietary](https://en.wikipedia.org/wiki/Proprietary_software) licenses.

# SUMMARY AND CONCLUSIONS

We have successfully extracted the data from the flipkart by auto launching the website. This project performed the following 

* Auto Launch the website
* Extract the Data
* Store the Data Extracted

# REFERENCES

* <https://smartinternz.com/Student/guided_project_info/6021#>
* <https://youtu.be/fpt7c0QgXDA>
* <https://youtu.be/mnDXBmtRQ8E>
* <https://youtu.be/iVYtNdCYgFo>

# APPENDIX

## SCREENSHOTS



