DEVELOPING A SEARCH ENGINE IN E-COMMERCE WITH THE AID OF REAL TIME SCRAPING

BY TEAM-12

1. SIRISHA G (19104147)
2. SAKTHIVEL M (19104141)
3. SUSANTH S (19104163)

III CSE - C

GUIDED BY

T.K.P RAJAGOPAL (ASST.PROFESSOR)
Department Of Computer Science Engineering

WEB DEVELOPMENT

-> Tim Berners-Lee, a British Scientist,invented the world wide web in 1989.

-> Web development refers to the building, creating, and maintaining of websites.

-> It includes aspects such as web design, web publishing, web programming, and database management.

Benefits

Cost
Cheaper way compared to other app developments

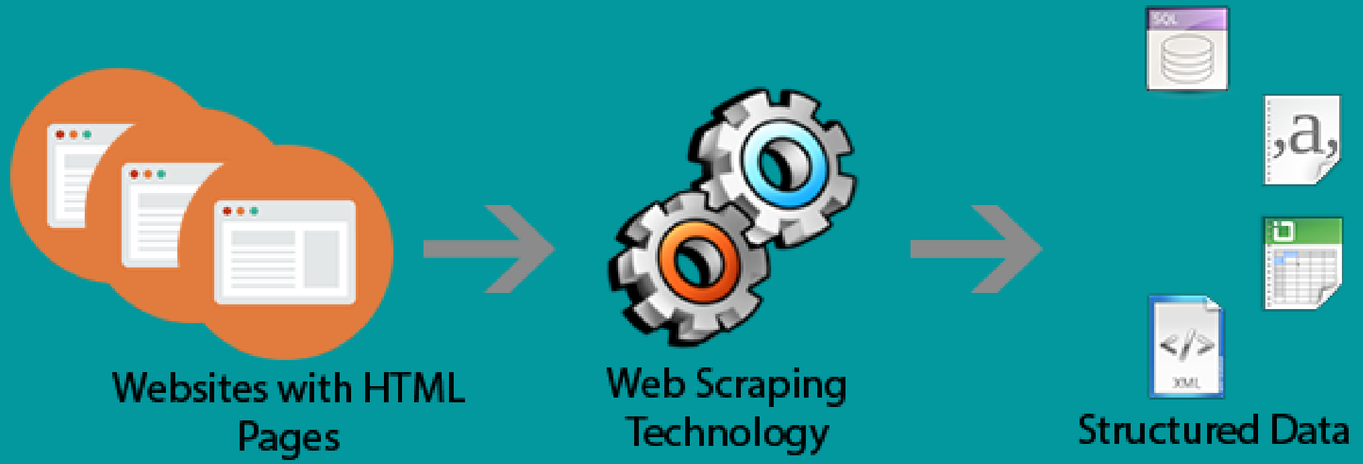
Updates
gets updated directly to most recent version

Customization
Finds easy to customize web apps

Platform
easily adopt to Windows , MacOS, etc.

Abstract

The main aim of the project is to extract data from websites and these data can be saved as Mysql database which is open source language by using web scraping technology.



Web scraping is acknowledged as an efficient and powerful technique for collecting data and these techniques have been customized from smaller ad hoc, that are able to convert entire websites into well-organized dataset.

Introduction

->Web scraping is also known as web extraction , is a technique to extract data from the World Wide Web and save it to file system or database.

->Web data is scrapped by Hyper-text Transfer Protocol and accomplished by a user or automatically by a bot.

Literature survey

Renita Crystal Pereira provided web scraping summary and techniques and tools that face several complexities as data extraction isn't that simple. The measurement level of web scraper will vary with the measurement units of the original source file, making it very difficult to interpret the data.

Federico Polidoro concentrated on the outcomes of web scraping evaluation strategies with particular orientation to user electronics services and goods throughout the sector of commodity price studies.

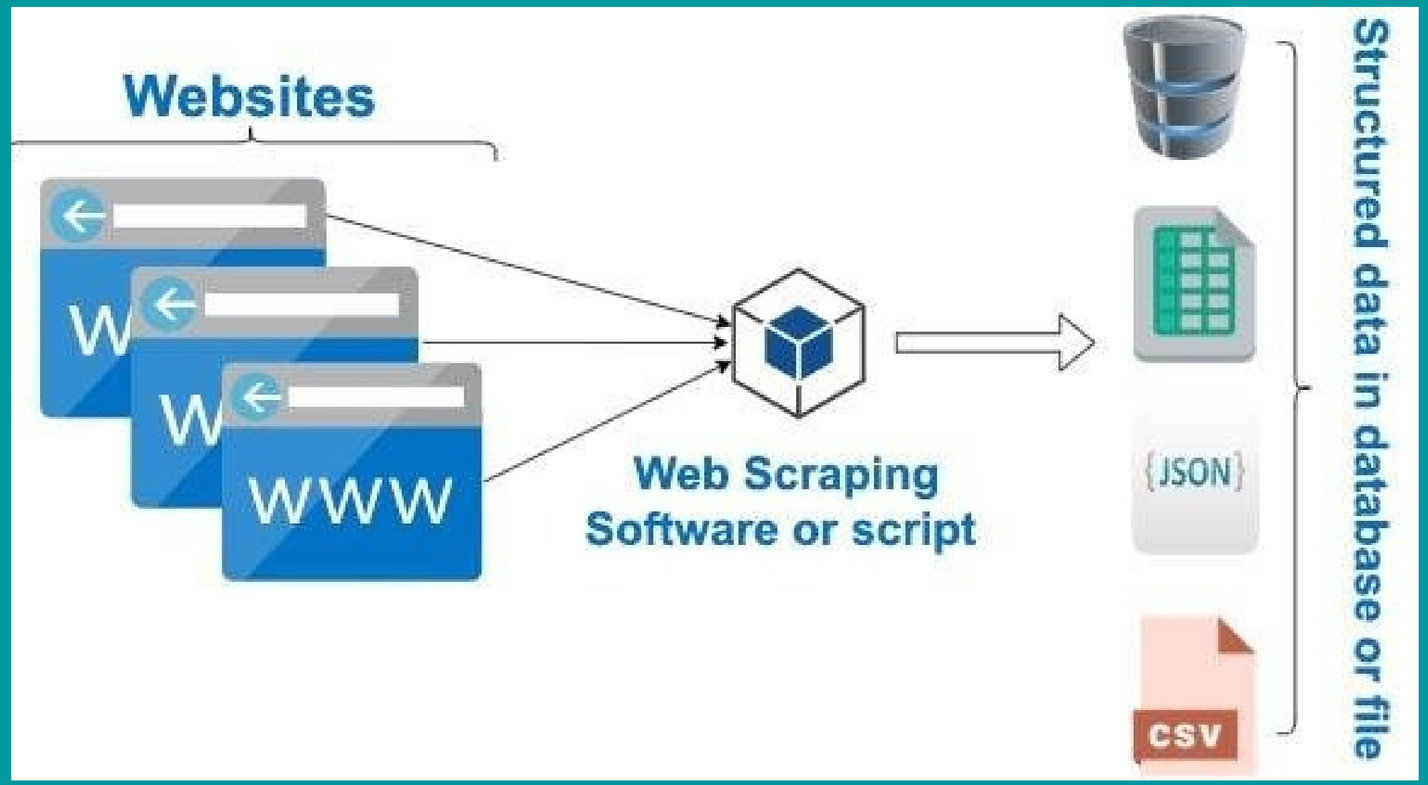
Kaushal Parikh proposed a web scraping detection with the help of machine learning and is valuable for research dependent companies.

Sameer Padghan projected an approach where data extraction is done from web pages in assistance with web scraping easily.

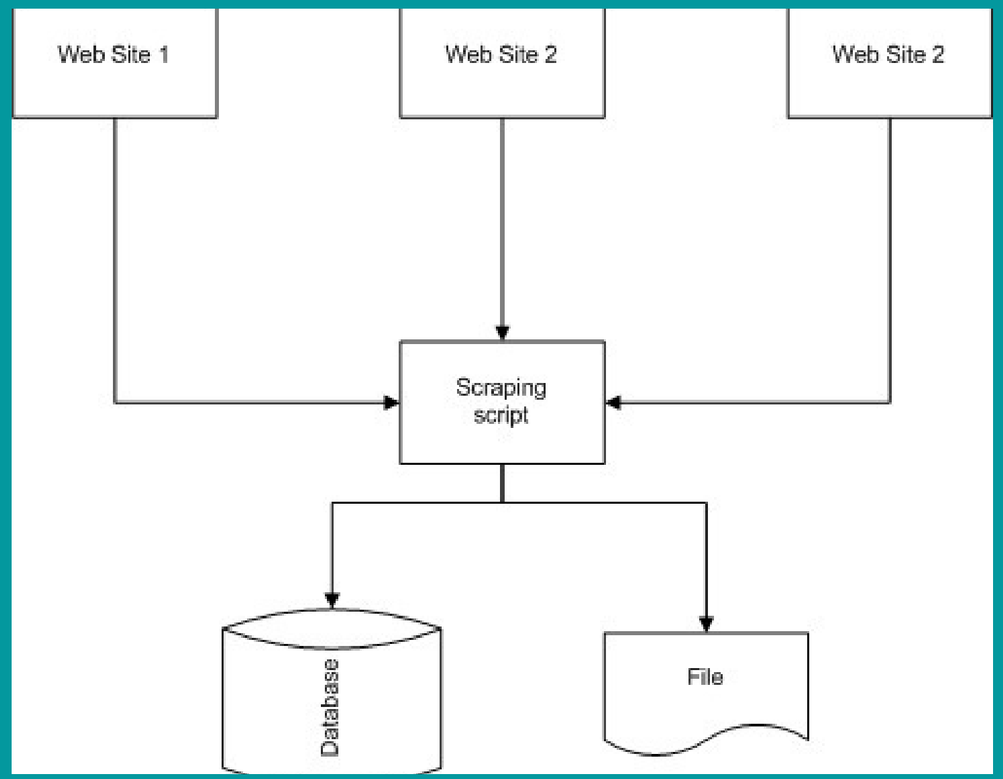
Anand Saurkar discovered a latest technique named web scraping and i s used to produce structured data based on unstructured data available on internet.

Proposed system

In the first proposed system,the most important step is to check robots.txt file to ensure that we have the permission to access the web page without violating any terms or conditions.



The web scraper will be given one or more URLs to load before scraping. The scraper then loads entire HTML code. More advanced scrapers will render the entire website including CSS and Javascript elements.

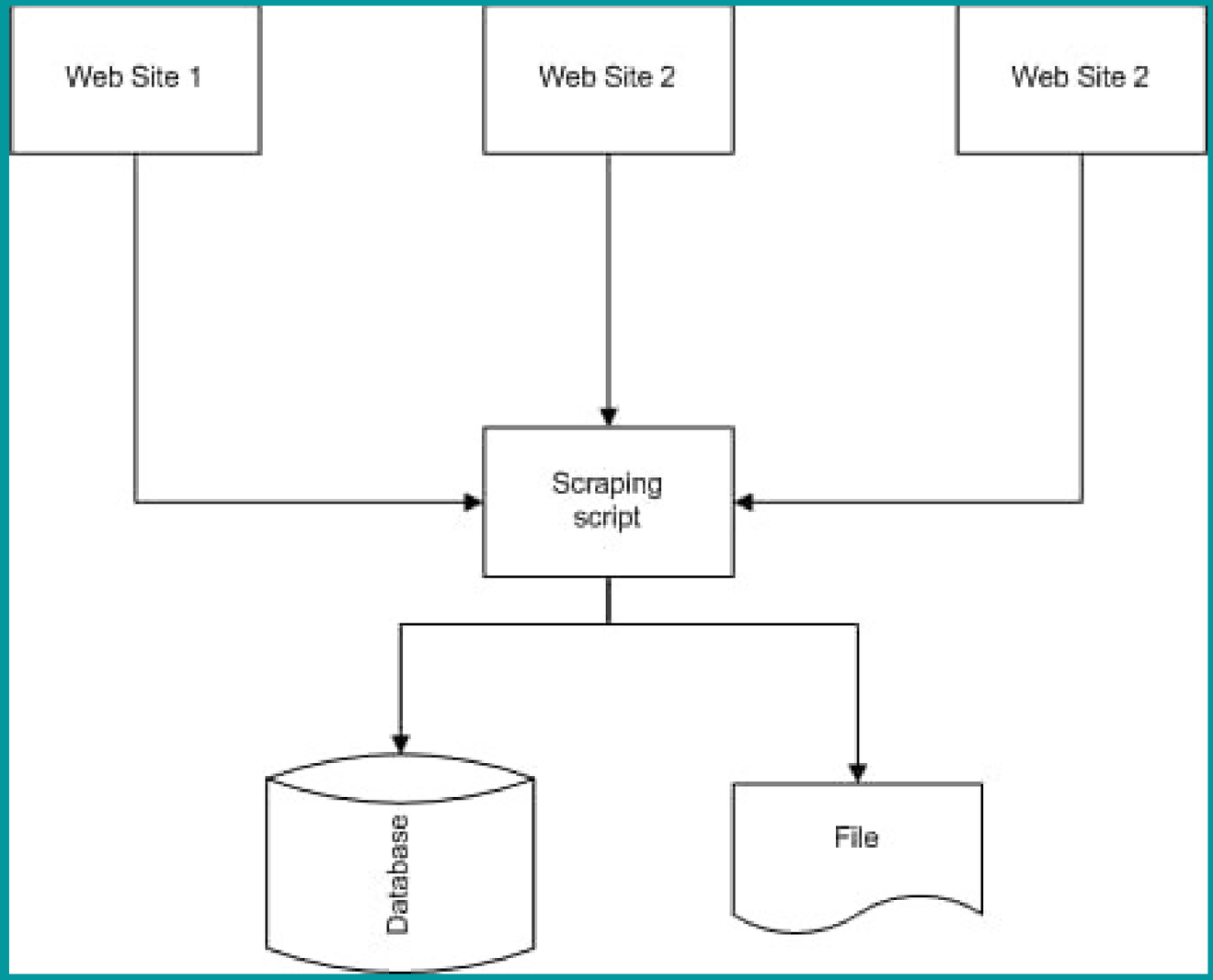


Web Scraper	one (or) more URLs	HTML code

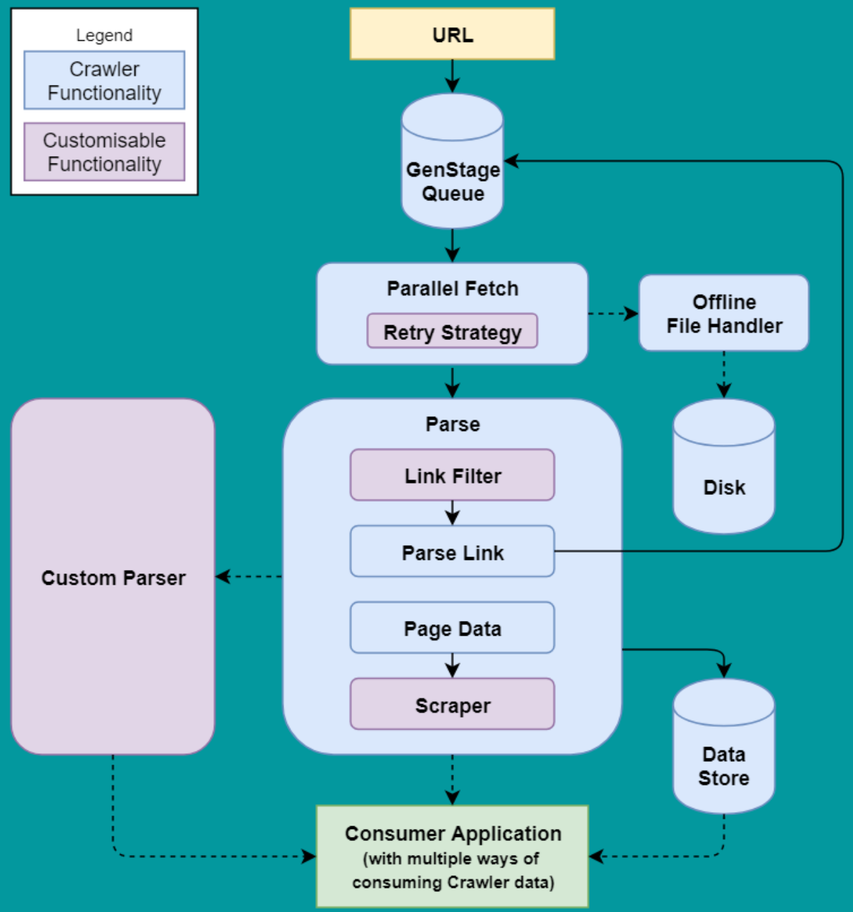
The user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.

Diagrammatic Representation:

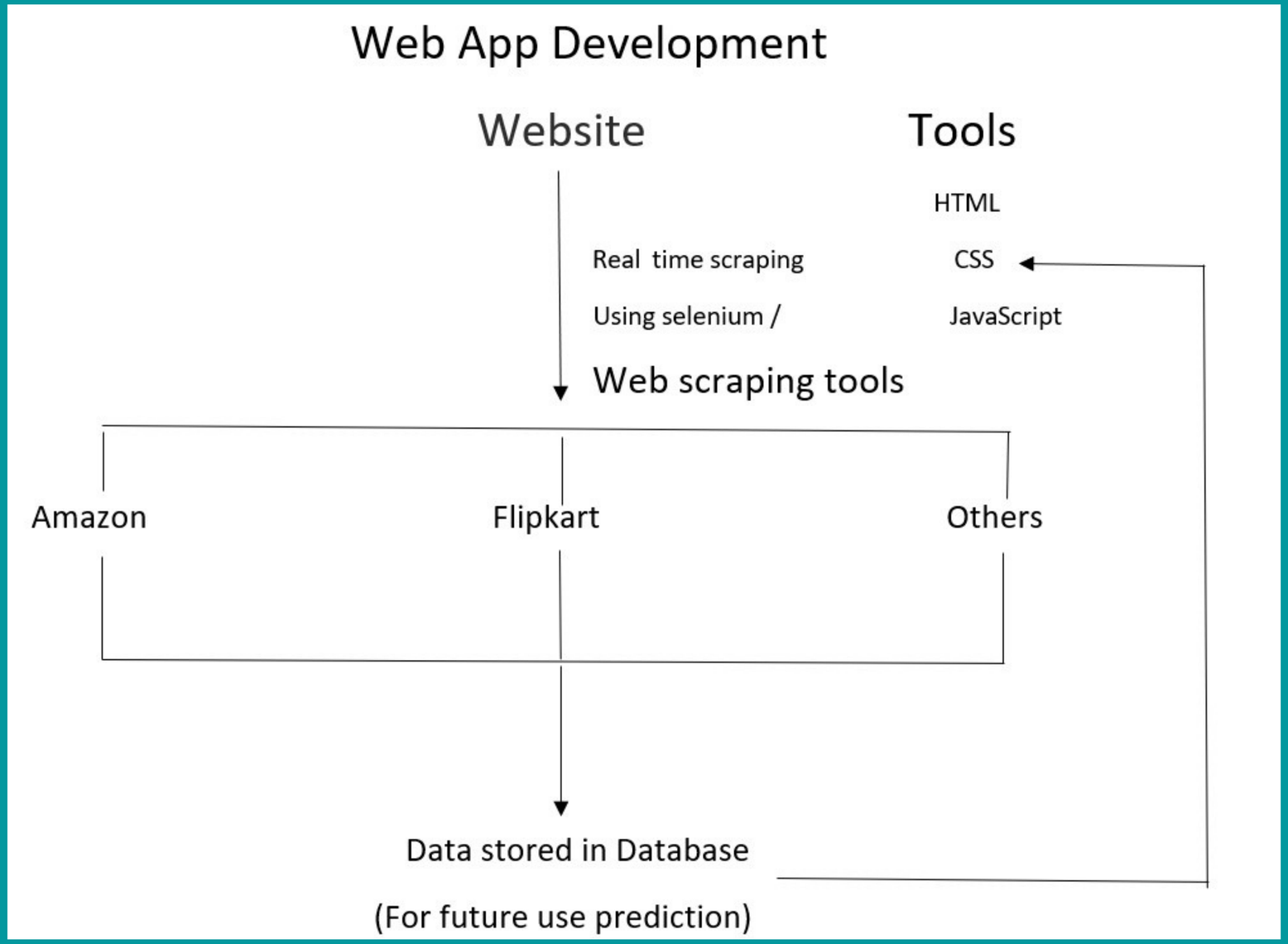
overview of web scraping



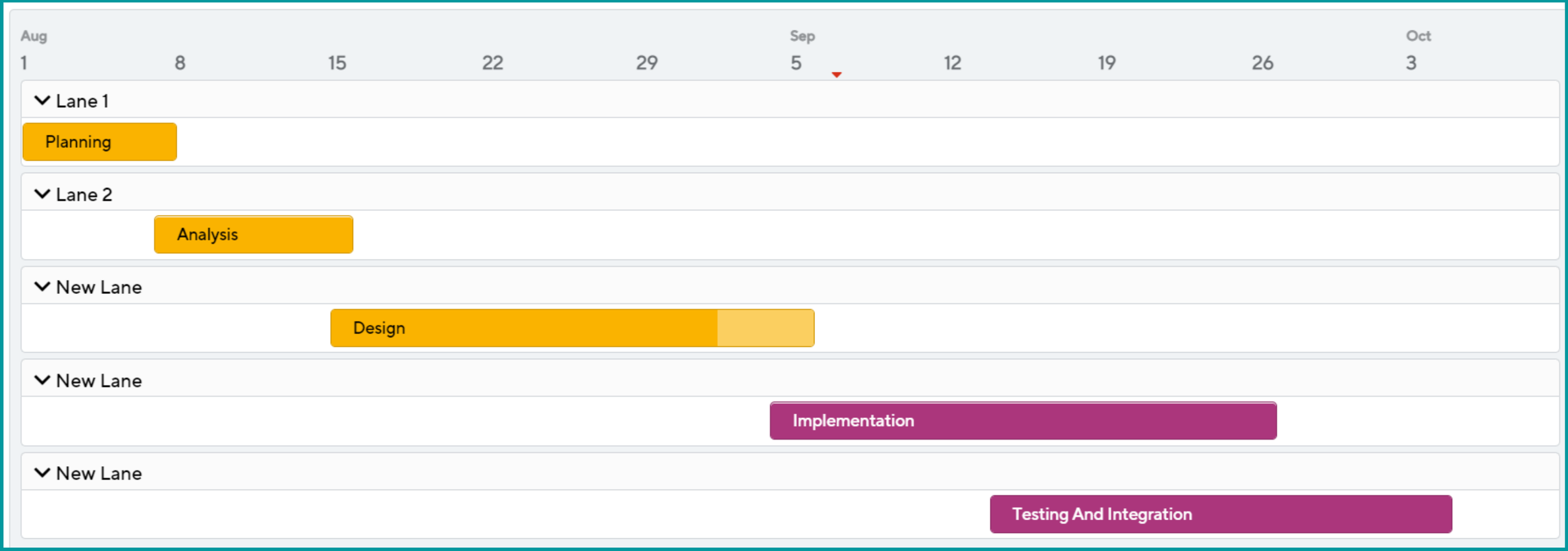
Architecture diagram



Modules split up



Gantt chart


Aug
Sep
Oct
1
8
15
22
29
5
12
19
26
3
Lane 1
Planning
Lane 2
Analysis
New Lane
Design
New Lane
Implementation
New Lane
Testing And Integration

Backend Source code python

```
@app.route('/about')
def about_script():
return render_template("about.html")

@app.route('/login')
def login_script():
return render_template("login.html")

@app.route('/mobiles')
def mobiles_script():
return render_template("mobiles.html")
```

```
@app.route('/laptops')
def laptops_script():
return render_template("laptops.html")

@app.route('/tv')
def tv_script():
return render_template("tv.html")
```

```
@app.route('/tvcompany/<variable7>')
def tvlist_script(variable7):

    if variable7== "samsung":

        url8="https://www.mysmartprice.com/electronics/pricelist/samsung-tv-price-list-in-india.html"
    elif variable7== "sony":
        url8="https://www.mysmartprice.com/electronics/pricelist/sony-tv-price-list-in-india.html"
    elif variable7== "lg":
        url8="https://www.mysmartprice.com/electronics/pricelist/lg-tv-price-list-in-india.html
```

```
else:
url8="https://www.mysmartprice.com/electronics/pricelist/toshiba-tv-price-list-in-india.html"

product_images_tv=[]
product_images_tv2=[]
product_names_tv=[]
tvspecs=[]
product_link_tv=[]

page=requests.get(url8)
content = page.content
soup=BeautifulSoup(content,"html.parser")
```

```
name=soup.find('h1',attrs={'class':'list-info__ttl js-list-ttl'})
company_name_tv=name.text

for images in soup.find_all('img', attrs={'class':'prdct-item__img'}):
product_images_tv.append(images.get('src'))

for images in soup.find_all('img', attrs={'class':'prdct-item__img'}):
product_images_tv2.append(images.get('data-lazy-src'))

for product in soup.findAll('a', attrs={'class':'prdct-item__name'}):
product_names_tv.append(product.text)
```

```
for productspecs in soup.findAll('div', attrs={'class':'prdct-item__spcftn-wrpr'}):
    tvspecs.append(productspecs.text)

for info in soup.findAll('div', attrs={'class':'list-info__dscrptn'}):
    productinfo_tv=info.text

tags = soup.find_all('a',attrs={'class':'prdct-item__name'})
for tag in tags:
    tag=tag.get('href')
    tag=tag.split("w.")[1]
    tag=tag.split("/")[2]
    product_link_tv.append(tag)
```

```
return render_template("tvcompany.html",
tvspecs=tvspecs,product_link_tv=product_link_tv,variable7=variable7,
company_name_tv=company_name_tv,
productinfo_tv=productinfo_tv,
product_images_tv=product_images_tv,
product_images_tv2=product_images_tv2,
product_names_tv=product_names_tv)
```

Frontend source code HTML

```
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8">
<title>login</title>
<link rel="stylesheet" type="text/css" href="{{url_for('static',filename='style.css')}}">

</head>
<link rel="stylesheet" type="text/css" href="
{{url_for('static',filename='loginstyle.css')}}">
<body>
```

```
<div class="login-box">

<h1>Login</h1>

<div class="text-box">
<i class="fa fa-user" aria-hidden="true"></i>
<input type="text" placeholder="Username" name="" value="">
</div>

<div class="text-box">
<i class="fa fa-lock" aria-hidden="true"></i>
<input type="password" placeholder="Password" name="" value="">
```

<input class="btn" type="button" name="" value=" sign in">

</body>
</html>
<section class="about">
<h1 >ABOUT US</h1>

Our main objective is to help people compare, understand, analyze and utilize the price data in online shopping. We've planted this Pricemania to change the way people shop online and also help businesses to maximize their profit.


```
</section>

<section id="newsletter">
<div class="container">
<h1>Subscribe for latest updates</h1>
<form>
<input type="email" name="enter email">
<button type="submit" class="button1">Subscribe</button>
</form>
</div>
</section>
```

<section class="pbox">

</section>
<section class="brands">

<h1>MOBILES</h1>

<h1>MOBILES</h1>
<img src="static/logos/mi.png">

<img src="static/logos/asus.png">

<img src="static/logos/oppo.png">

<img
src="static/logos/vivologo.png">
<img src="static/logos/apple.jpg">

<img
src="static/logos/nokia.png">
<img

src="static/logos/huawei.png">
<img src="static/logos/lenovo.jpg">

<img src="static/logos/moto.png">

<img
src="static/logos/samsung.png">
<img
src="static/logos/allmobiles.png">

```
</section>

<section id="newsletter">
<div class="container">
<h1>Subscribe for latest updates</h1>
<form>
<input type="email" name="enter email">
<button type="submit" class="button1">Subscribe</button>
</form>
</div>
</section>
```

output

Apple iPhone 12

- Apple A14 Bionic
- 4 GB RAM
- 64 GB internal storage
- 2815 mAh battery

- Dual (12 12) MP Rear, 12 MP Front Camera rear
- 6.1 inches (15.49 cm) Screen
- Dual, Nano-eSIM SIM
- iOS v14

Apple iPhone 11

- Apple A13 Bionic
- 4 GB RAM
- 128 GB internal storage
- 3110 mAh battery

- Dual (12 12) MP Rear, 12 MP Front Camera rear
- 6.1 inches (15.49 cm) Screen
- Dual, Nano-eSIM SIM
- iOS v13.0

Apple iPhone 13 Pro Max

- Apple A15 Bionic
- 6 GB RAM
- 64 GB internal storage
- 4352 mAh battery

- Triple (12 12 12) MP Rear, 12 MP Front Camera rear
- 6.7 inches (17.02 cm) Screen
- Dual, Nano-eSIM SIM
- iOS v15

Apple iPhone 11

Websites	Prices
amazon.in	₹43,999
Flipkart	₹49,900
TATA CLiQ	₹60,900

HP Pavilion Gaming 15-ec0101AX (167W1PA) Laptop (15.6 Inch | AMD Quad Core Ryzen 5 | 8 GB | Windows 10 | 1 TB HDD)

- 15.6 inch ScreenAMD Quad Core Processor8 GB Memory1 TB HDD

HP 15s-GR0011AU (35K34PA) Laptop (15.6 Inch | AMD Dual Core Ryzen 3 | 8 GB | Windows 10 | 1 TB HDD)

- 15.6 inch ScreenAMD Dual Core Processor8 GB Memory1 TB HDD

HP Chromebook 14a-na0030nr (9LL05UA) Laptop (14 Inch | Celeron Dual Core | 4 GB | Google Chrome | 32 GB SSD)

HP 15s-GR0011AU (35K34PA) Laptop (15.6 Inch | AMD Dual Core Ryzen 3 | 8 GB | Windows 10 | 1 TB HDD)

Websites	Prices
TATA CLiQ	₹39,288

Subscribe for latest updates

Subscribe

Conclusion

"If programming is magic, then web scraping is wizardy" (Mitchell,2015) said Ryan Mitchell. The presence of internet led to increasing source of information that can be accessed so that information seeking activities become the most common activities performed and become one of the activities that took quite a bit (Adam, 2012). The internet will be remembered as the first place we can collect huge amount of data without spending a lot of energy or money. Whether in ecommerce or e-marketing, the use of the technique of web scraping will be the key to sucess as it will provide insight into the targeting market and help decision makers.

This project focused on how to create a web scraping application using the simple html dom library function ,and the tools that the framework provided, as well as the use of PHP tools that were available. The main intention was to create a web application using the web scraping and retrieved useful information from a website. This goal was met, and a robust application was created.

Limitations

Web-scraping can be also challenging if you don't have the proper tools. Largely, you're completely at the mercy of the target website, and that website can change at anytime - without notice. Or, it may contain faulty JavaScript that causes it to crash and exhibit surprising behavior. The server that hosts the website may crash, or the website may undergo maintenance. Many potential problems can occur during a lengthy web-scraping session, and you have very little influence on any of them.

Future works

This would allow me to use a more precise method of parsing that would ensure that every piece of data I wanted was collected. The ideal tool would allow me to create a unique program to extract the information that I needed. I hoped that
this body of work was able to expose academicians, students and practitioners to the con cept and necessity of web scraping and demonstrate a tool that the average computer user could utilize on their own.

References

- R. Mitchell, Web Scrapping with Python., O’Reilly Media, 2015
- Đ. Petrović and I. Stanišević, "Web scrapping and storing data in a database a case study of the used cars market", 2017 25th Telecommunication Forum (TELFOR), pp. 1-4, 2017
- D. S. Sirisuriya, A comparative study on web scraping, 2015
- Osmar Castrillo-Fernández, "Web Scraping: Applications and Tools", European Public Sector Information Platform Topic Report, no. 2015, December 2015
- Carlos, Iglesias Mercedes Garijo Jose Ignacio Fernandez-Villamor and Jacobo Blasco-Garcia, "A Semantic Scraping Model for Web Resources", Applying Linked Data to Web Page Screen Scraping.
- E. Vargiu dan and M. Urru, Exploiting Web Scraping in a Collaborative Filtering Based Approach to Web Advertising, Italy, 2012


thank you..