International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

xx

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 15th Dec 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-13&issue=Issue4

10.48047/IJIEMR/V13/ISSUE 04/24

TITLE: Web Scraping for E-Commerce Website

Volume 13, ISSUE 04, Pages: 216-224

Paper Authors **1Vaishnavi Deshmane,2Prof. Jitendra Musale,3Prof. Shweta Joshi,4Vedant Chinta,5Kaif Gokak,6Isha Dalbhanjan**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Web Scraping for E-Commerce Website

[1]Vaishnavi Deshmane, [2]Prof. Jitendra Musale, [3]Prof. Shweta Joshi, [4]Vedant Chinta, [5]Kaif Gokak, [6]Isha Dalbhanjan

[1, 2, 3, 4, 5, 6]Computer Engineering
Anantrao Pawar College of Engineering and Research, Pune, India
vaishnavideshmane2003@gmail.com, jitendra.musale@abmspcoerpune.org, shweta.joshi@abmspcoerpune.org, vedantchinta223@gmail.com, kaifgokak@gmail.com, dalbhanjanisha523@gmail.com

Abstract
Web scraping has become an invaluable tool for E-Commerce companies seeking a competitive advantage in the dynamic online marketplace. The summary provides an overview of this practice and highlights its importance, methods, and benefits for E-Commerce businesses. E-Commerce websites, with their extensive product catalogs and ever-changing pricing strategies, are data-rich environments. Web analytics is an computerized process of retrieving data from these websites, allowing companies to collect important information about products, prices, reviews and competition. The retrieved data can be used for various purposes including market analysis, price optimization and customer sentiment analysis. This roundup covers the main methods used in E-Commerce web scraping, with a focus on using Python-based libraries such as BeautifulSoup and Scrapy, as well as popular headless browsers such as Selenium. These tools allow you to extract structured data from websites, provide companies with valuable information and improve their decision-making ability. The benefits of web scraping for E-Commerce websites are numerous. Companies can gain a competitive advantage by tracking competitors' prices, tracking product availability, and identifying trends in customer reviews. This data can support pricing strategies, inventory management, and marketing efforts, ultimately leading to increased sales and profitability. However, it is important that companies familiarize themselves with the legal and ethical issues surrounding web scraping. Many websites' terms of service prohibit scraping, which can lead to legal consequences if you're not careful. Therefore, the summary also raises the question of the importance of following ethical principles and adhering to website guidelines.

Keywords — Web Scraping, E-Commerce, Python, BeautifulSoup, Scrapy, Selenium, Pandas

Introduction
In the dynamic and highly competitive world of E-Commerce, gaining a strategic advantage has never been more important. E-Commerce companies operate in an environment where prices are changing, product catalogues are growing rapidly, and customer sentiment is constantly changing. To succeed in this dynamic landscape, many companies are turning to a revolutionary

technique called web scraping. Web scraping is the art and science of automatically extracting data from websites and has revolutionized the way E-Commerce companies collect and use information. In this introduction, we embark on a journey to explore the world of web scraping for E-Commerce websites and shed light on its meaning, methods, and the transformative impact it can have on digital businesses. E-Commerce websites are data treasures full of product details, customer reviews, pricing information, and more. While this data is easily accessible to users, extracting and organizing it at scale is a Herculean task. This is where web scraping comes into play, offering a solution that can save companies countless hours of manual data collection and analysis. Web scraping methods and tools have evolved over the years, making it more accessible and effective. From Python-based libraries like BeautifulSoup and Scrapy to headless browsers like Selenium, companies can use a variety of techniques to extract structured data from E-Commerce websites. Using these tools, they can get product descriptions, pricing data, inventory information, customer reviews, and even competitor data. The potential benefits of web scraping in the E-Commerce industry are enormous. E-Commerce companies can use web scraping to monitor competition, optimize pricing strategies, monitor product availability, and gain valuable insights into customer sentiment. By leveraging the data gained from web browsing, companies can make informed decisions that increase sales, improve customer experience, and ultimately increase profitability. However, caution is advised when web scraping. Many websites' terms of service specifically prohibit scraping, which can raise legal and ethical issues if not done knowingly. It is important for companies to find a balance between gaining a competitive advantage and respecting the rules and rights of website owners. In the following pages, we delve into the world of web scraping for E-Commerce websites, exploring methods, ethical considerations, and a wide range of applications. As we explore this landscape, we will see how web scraping has become an essential tool for E-Commerce companies that want to succeed in the digital marketplace, where information has power and every data point counts. The main objective of this paper is to make the above mentioned process user friendly such that they should be able to get the best E-Commerce site from which they can buy the product.

Ease of Use

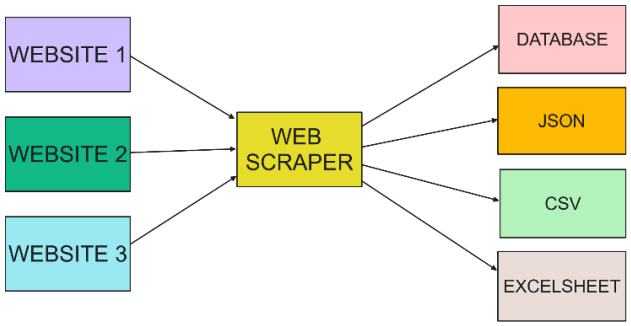

Fig. Working of Web Scraper

A. Traditional Copy-Paste:

Simple form web scraping involves manually copying data from a web page and pasting it into a text file or table. While this method is straightforward, it can become error-prone and tedious, especially when dealing with large datasets.

B. HTTP Parsing:

This technique is utilized to extract data from both static and dynamic web pages. Data retrieval is achieved by sending HTTP requests to a remote web server.

C. Web Scraping Software:

Various software tools are available for customizing web scraping solutions. These tools may include features such as automatic

recognition of page structures, logging interfaces to eliminate the need for manual coding, scripting functions for content extraction and transformation, and database interfaces for storing retrieved data in local databases.

Importance

Web scrapers are important for E-Commerce websites for several reasons. First, it can be used to check competitors' prices and ensure that prices are competitive. This is important because consumers are becoming more price sensitive and often look for the best price before buying. Web scrapers can be used to monitor customer reviews and identify potential problems with products or services. This will improve your products and services and help you avoid negative customer reviews. Web scrapers can be used to identify new product opportunities and conduct research on consumer trends. This information can be used to develop requirements for new products and services. Web scrapers can be used to collect information about consumer behaviour and demographics. This data can be used to improve marketing campaigns and better target ads. Overall, a web scraper is a powerful tool that can help E-Commerce websites improve their performance in various ways. By monitoring competitive pricing, customer research, identifying innovation opportunities, and collecting data on customer behaviour, E-Commerce websites can use web scrapers to gain a competitive advantage and increase sales.

Literature Survey

Prof. P.S. Gaikwad, Kaushal Parmar, Rohit Yadav, and Datta Supekar [1] proposed a scheme aimed at extracting product descriptions from multiple e-commerce websites and displaying them on a single website for comparison. Various tools such as BeautifulSoup, Selenium, and Scrapy are used to extract data. Once extracted, the data is stored in a MySQL database and displayed in a similar format on their web app.

Niranjan Krishna, Anvith Nayak, Sana Bad, and Dr. Sandhya N [2] used various filtering methods to select the most appropriate pages associated with a given web page in their study on web scraping.

Vlad Krotov, Leigh Johnson, and Leiser Silva [3] discussed legal and ethical developments in web scraping projects. They also mentioned the development of an R package to independently search and retrieve data from Dice.com, aimed at stimulating further research and providing value to the wider community.

SCM de S Sirisuriya [4] conducted a comparative study on web scraping techniques and well-known web scraping software, focusing on data extraction from educational websites.

Shakra Mehak, Rabia Zafar, Sharaz Aslam, and Sohail Masood Bhatti [5] utilized web mining and scraping techniques to identify top products from e-commerce websites. They developed a framework based on HTML, CSS, and PHP, which dynamically retrieves and displays results without storing data in a local database, aiming to improve security and usability.

Gandhe Vineeth Kumar, Hema M S, Aishwarya R, and K R Mamatha [6] described their proposed strategy for solving e-commerce pricing problems using web scraping and machine learning concepts, allowing users to track and compare prices, receive email alerts, and forecast future price increases.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Proposed System

The provided figure offers a detailed description of the proposed system. The backend of the system utilizes web scraping techniques, including various libraries such as BeautifulSoup, Selenium, and Scrapy, to extract information about products from different E-Commerce websites. The dashboard provides a graphical user interface (GUI) in the form of a website/web app, enabling users to interact with the system. Information about the products is then presented on the website.When a client requests information about a product, a query is executed in the local database, and the retrieved data is converted into a readable format before being stored in the database. The homepage of the website displays information about products, allowing users to view details and compare prices to choose the product that best fits their needs.

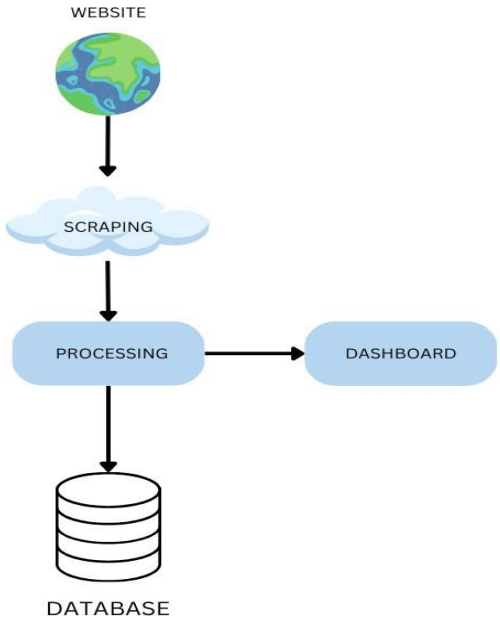

Fig. Proposed System

Website Scraping:

This is how data is extracted from networks. It may involve loading web pages, parsing HTML content, and extracting relevant information. Common tools for this task are web scraping libraries in Python and other programming languages (e.g., BeautifulSoup, Scrapy).

Data Processing:

Once data is extracted, it generally needs to be cleaned, transformed, and organized for further analysis. This step ensures that the data is in a usable format for the next stage of the process.

Database Storage:

The processed data is then stored in a database. Databases are collections of data that can be accessed, maintained, and updated. Common types include relational databases (e.g., MySQL, PostgreSQL) or NoSQL databases (e.g., MongoDB), depending on the nature of your data.

Dashboard Usage:

This may involve creating a user interface (dashboard) to visualize and interact with the processed data. Visualization tools such as Tableau, Power BI, or custom dashboards can be used in conjunction with libraries such as D3.js or Plotly and web frameworks such as Flask or Django for this purpose.

Existing System

Web scraping is utilized for retrieving and tracking online transactions, web crawling, data mining, monitoring online price changes and comparisons, gathering product reviews, collecting real estate inventory, monitoring sky data, detecting website changes, and conducting presence and reputation analysis, as well as link building and network integration. Websites are developed using text languages such as HTML and XHTML, often containing a wealth of useful data and information. However, most web applications are designed for end-users and are not intended for automation. Hence, various web content exploration tools have been developed. Web Scraper serves as an application programming interface (API) for scraping data from internet companies such

as Amazon AWS and Google, offering end-users free access to web scraping tools, services, and public data.

Research Methodology

1. Introduction: The purpose of this research is to investigate and analyse the use of web scrapers in the context of E-Commerce websites. Web scrapers are tools or programs used to extract data from websites automatically. In the context of E-Commerce, they can be employed for various purposes such as price monitoring, product catalog maintenance, competitive analysis, and more. This research aims to understand the current state of web scrapers in E-Commerce, their effectiveness, challenges, and potential improvements.

2. Research Objectives: The primary objectives of this research are as follows: a. To assess the current use of web scrapers in E-Commerce. b. To identify the key benefits and challenges associated with web scraping in E-Commerce. c. To explore the ethical and legal implications of web scraping in E-Commerce. d. To propose best practices and recommendations for using web scrapers in E-Commerce.

3. Research Design: This research will employ a mixed-methods approach, combining both qualitative and quantitative research methods to achieve the objectives mentioned above. The research design is as follows:

4. Ethical and Legal Analysis:

- Terms of Services and Website Policies: Always review and abide by the website's Terms of Service and Privacy Policy. These documents may explicitly prohibit or restrict web scraping activities. Violating these terms can result in legal consequences.

- Copyright and Intellectual Property Be cautious when scraping content that is protected by copyright, trademarks, or other forms of intellectual property. Ensure that your scraping activities do not infringe on these rights.

- Use of Scraped Data Ensure that the scraped data is used for lawful and ethical purposes. Do not engage in activities that harm individuals or businesses, such as price manipulation, misinformation, or fraudulent schemes.

- Responsible and Ethical Scraping: Follow ethical web scraping practices. Use reasonable and responsible scraping techniques, avoid causing harm to the target website, and aim to be transparent about your scraping activities.

5. Best Practices and Recommendations:

- Use an API if available: Many E-Commerce websites offer APIs that allow you to access their data in a structured and legal manner. If an API is available, use it as your primary data source instead of scraping the website.

- Limit the frequency of requests: Avoid sending too many requests to the website in a short time frame. Excessive scraping can put a strain on the target server and may lead to IP blocking or other countermeasures.

6. Data Collection and Sampling: For the online surveys, a convenience sampling method will be used, targeting E-Commerce professionals and website operators. The selection will include a diverse range of businesses in terms of size and industry. For in-depth interviews, purposive sampling will be employed to include experts and experienced practitioners in the field.

7. Data Analysis: Data collected from surveys and interviews will be analyzed using

qualitative and quantitative data analysis techniques. Thematic analysis will be used to identify common themes and patterns in the responses, and statistical analysis will be applied to survey data when appropriate.

8. Ethical Considerations: Ensure that all research activities comply with ethical guidelines and data protection regulations. Obtain informed consent from survey participants and interviewees. Anonymize data and protect sensitive information.

9. Expected Outcomes: The research aims to provide a comprehensive understanding of web scrapers in E-Commerce, their benefits, challenges, and ethical implications. The outcomes will include practical recommendations for E-Commerce businesses and insights into the future of web scraping in this context.

10. Conclusion: The research methodology outlined here will enable a systematic investigation of web scrapers for E-Commerce websites, providing valuable insights for businesses and practitioners in the field. The research findings are expected to contribute to a better understanding of the role of web scrapers in E-Commerce and inform best practices in their use.

Advantages

1. Web scrapers provide a crucial service at a competitive yet reasonable cost. Data is collected regularly and analyzed to serve the client. Web scrapers offer these features in a budget-friendly manner.

2. Individual research and conservation are particularly important, especially for other websites/channels where the collected data may be useful for understanding audience behavior and crafting targeted information for the public. Online user behavior influences website strategies, enabling companies to understand their audience better and offer them what they truly prefer.

3. Web scraping provides an efficient and accurate forecast of prospects. By analyzing user habits, desires, and preferences, advanced predictive analytics can be performed. Understanding customer expectations allows companies to better prepare for the future.

4. One of the most common advantages of web scraping for e-commerce websites is price comparison. Every company benefits from monitoring the prices of competing websites, from eBay to Amazon sellers, using web scraping. This enables companies to compare prices for the same products or services across different platforms, ensuring transparency for customers.

5. New products and images are constantly emerging on the internet, but they can be challenging to access. Web scraping audits all websites for these updates and more. By tracking competitors through this method, companies can streamline their processes without the need for additional procedures.

System Architecture

The web scraping process is typically divided into three stages:

Fetching Stage: This stage involves accessing the desired website and retrieving relevant information using the HTTP protocol. This is achieved by sending an HTTP GET request to the target URL and obtaining the HTML page as a response. Libraries like BeautifulSoup can be utilized for this purpose.

Extraction Stage: In this stage, information is collected from websites using either regular expressions or HTML parsing libraries. Regular expressions, or regexes, are powerful patterns for matching and extracting specific

data (e.g., email addresses or phone numbers) from HTML content. Alternatively, HTML parsing libraries like BeautifulSoup in Python or Cheerio in JavaScript provide a structured way to navigate and extract data from HTML documents based on tags, attributes, and hierarchy.

Transformation Stage: The extracted data is prepared for analysis in this stage. Unstructured text data is converted into structured formats such as JSON, CSV, or databases using tools like Pandas in Python. Data cleaning is essential to eliminate duplicates, correct errors, and standardize formats for reliable analysis. Data cleansing tools and scripts are typically employed for this purpose. Additionally, enrichment may occur at this stage by combining data with external datasets or performing calculations to obtain additional information.

Presentation Stage: In this stage, structured data becomes accessible and understandable to users. Data visualization tools such as Tableau, Power BI, or libraries like Matplotlib in Python and D3.js in JavaScript are used to create interactive graphs, charts, and dashboards for visually representing data. Reports can be generated automatically or on-demand in human-readable formats such as PDF, Word, or HTML reports. Business intelligence tools provide detailed analysis, data mining, ad hoc queries, and interactive dashboards, enabling users to make informed decisions based on structured data.

Tools required are used to scrape the web:

1.BeautifulSoup:

BeautifulSoup stands as another indispensable Python library widely employed for parsing data from XML and HTML documents. It facilitates the organization of parsed content into navigable and searchable data structures, significantly easing the process of traversing through extensive datasets. Many data analysts rely on BeautifulSoup as their primary tool for extracting and processing web data.

2.Scrapy:

Scrapy emerges as a robust Python-based application framework tailored for web crawling and structured data extraction. Renowned for its versatility, Scrapy finds applications in data mining, information processing, and archival of historical content. Beyond its primary purpose of web scraping, it serves as a comprehensive web crawling solution, capable of extracting data through APIs as well.

3.Pandas:

Pandas, another versatile Python library, finds extensive use in data manipulation and indexing tasks. When combined with BeautifulSoup, it offers a potent toolkit for web scraping applications. One of the key advantages of utilizing pandas is its seamless integration within the Python ecosystem, enabling analysts to conduct the entire data analytics process within a single language framework. This eliminates the need for switching between different programming languages, such as R, streamlining the workflow for data professionals.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

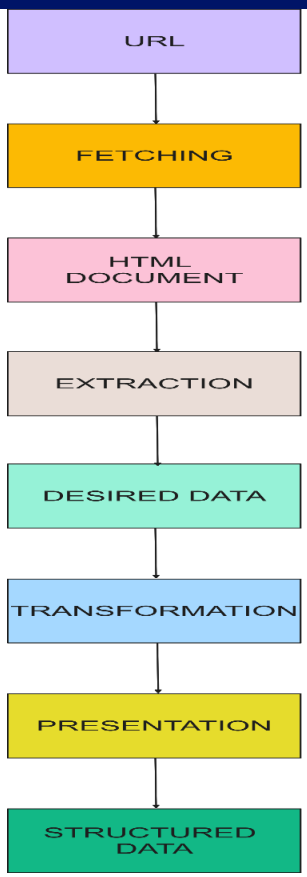

Fig. System Architecture of Web Scraper

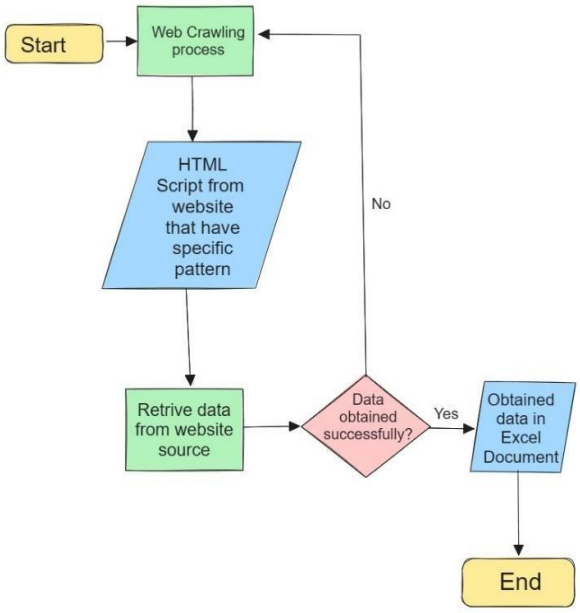

Fig. Flowchart

Algorithm

Sr. No.	Steps
1.	Identify the target website and the data to be scraped.
2.	Request the HTML code of the target website.
3.	Parse the HTML code to extract the desired data.
4.	Store the scraped data in a database or other data structure.
5.	Repeat steps 2-4 until all of the desired data has been scraped.

Algorithm Description

The product name is taken as user input and then passed to various e-commerce sites from where the data is downloaded/extracted. Since data cannot be downloaded directly from a website, we need several web scraping tools/libraries for this purpose.

Conclusion

Web scrapers are invaluable tools for e-commerce websites, offering valuable insights and a competitive edge over competitors. However, ethical and legal considerations must be prioritized to ensure responsible and sustainable usage. This paper aims to enhance the effectiveness of our web scraping process. We have found that most web scrapers are usually designed to perform common tasks decently but look quite generic. The future of web scraping in e-commerce holds great promise and challenges, calling for ongoing innovation and ethical practice. Web scraping can speed up consumer research by reducing the cost and time of data collection.

Future Scope

Web scraping is becoming increasingly important as more data is added to the online world. Many companies now offer their customers customized tools that gather

information from across the Internet and organize it into useful and easy-to-understand information Valuable human resources that manually navigate through each web page and data collection decreases. Web Scrapers that have been developed have code for each and every individual website and crawlers do great scraping. If the web page has a complex structure, it requires more coding to decompress the data compared to a simple one. The future of web scraping is indeed exciting and over time it will become increasingly important for any business.

References

[1] Prof. P.S.Gaikwad, Kaushal Parmar, Rohit Yadav (2021), Datta Supekar,"IMPLEMENTATION OF WEB SCRAPING FOR E-COMMERCE WEBSITE".
[2] Niranjan Krishna,Anvith Nayak,Sana Bad,Dr.Sandhya N (2022),"A Study on Web Scraping".
[3] Vlad Krotov, Leigh Johnson, Leiser Silva," (2020) Tutorial: Legality and Ethics of Web Scraping".
[4] SCM de S Sirisuriya,(2015)"A Comparative Study on Web Scraping".
[5] Shakra Mehak, Rabia Zafar, Sharaz Aslam, Sohail Masood Bhatti, (2019) "Exploiting Filtering approach with Web Scrapping for Smart Online Shopping - Penny Wise: A wise Tool for Online Shopping"
[6] Gandhe Vineeth Kumar,Hema M S,Aishwarya R,K R Mamatha, (2022) "Web Scraping for E-Commerce Websites".
[7] Rabiyatou DIOUF, Edouard Ngor SARR,Ousmane SALL,Babiga BIRREGAH,Mamadou BOUSSO,Sény Ndiaye MBAYE,"Web Scraping: State-of-the-Art and Areas of Application " -2019 IEEE International Conference on Big Data (Big Data)
[8] Zhao, B. (2017). Web scraping. Encyclopedia of big data, 1-3.
[9] https://research.aimultiple.com/ai-web-scraping/
[10] https://monashdatafluency.github.io/python-scraping/
[11] Moaiad Ahmad Khder,(2021)"Web Scraping or WebCrawling: State of Art, Techniques, Approaches and Application"
[12] https://it-s.com/the-future-of-web-scraping-services/
[13] https://limeproxies.netlify.app/blog/future-of-web-scraping
[14] L. a. L. B. Zhang, "Aspect and entity extraction for opinion mining," in Zhang, Lei and Liu, Bing, Berlin, Heidelberg, Springer, 2014, pp. 1--40.
[15] Antonakis, John, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive (2010), "On Making Causal Claims: A Review and Recommendations," The Leadership Quarterly, 21 (6), 1086-120.
[16] Datta, Hannes, George Knox, and Bart J. Bronnenberg (2018), "Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery," Marketing Science, 37 (1), 5-21.
[17] S. Upadhyay, V. Pant, and S. Bhasin, (2017) "Articulating the Construction of a Web Scraper for Massive Data Extraction."
[18] Ohidujjaman, Hasan, M. & Huda, M.N. (2013). E-Commerce Challenges, Solutions and Effectiveness Perspective Bangladesh.