- **ZenRows**: This is an API that provides a powerful toolkit to get around all kinds of anti-scraping protections. It supports JavaScript rendering, rotating premium proxies, and headless browsing. It integrates with any programming language and is suitable for advanced web scraping. It is easy to use and has great documentation and support. However, you will need a data parsing library to process the extracted data. It offers 1,000 free API credits after signing up, then plans start from $49/month.
- **Scrapy**: This is an open-source framework for automated web scraping using Python. It supports asynchronous loading, which allows you to scrape many pages at once. It also allows you to export the data in different formats such as JSON, CSV, or XML. However, it cannot handle JavaScript and may have difficulty bypassing some anti-scraping measures. It may also cause memory leaks if not handled properly. It is free to use but requires plug-in proxies bought from third-party providers.
- **ParseHub**: This is a web scraping tool that handles outdated websites and interactive pages. It integrates with Tableau and Google Sheets and offers rotating proxies with a paid plan. It is designed for market researchers and does not require coding skills. However, it limits the number of projects you can create with a free plan to five. After that, plans start from $189/month[3].
- **Apify**: This is a web scraping tool that offers easy scraping of popular sites such as Amazon, Google, or Facebook. It contains more than 200 tools for data extraction and web

automation. It also provides workflow management, process batching, easy access controls, and API integration. However, it may not be suitable for complex web scraping tasks that require custom logic or advanced parsing. It gives $5 in credit for free, then plans start from $49/month.

- **Mozenda**: This is a web scraping tool that offers cloud-based data extraction and management services. It can scrape data from any website and store it in various formats such as CSV, XML, JSON, or SQL. It also provides data quality assurance, scheduling, monitoring, and reporting features. However, it may be expensive and less flexible than other tools. It offers a 30-day free trial after contacting its sales department, then plans start at $99/month.
- **ScraperAPI**: This is a web scraping API that handles proxies, browsers, and CAPTCHAs for you. It can scrape any website with any level of complexity and scale up to millions of requests per day. It also supports geo-targeting, custom headers, and JavaScript rendering. However, it does not provide data parsing or cleaning services. You will need to use another tool or library to process the extracted data. It offers 1,000 API credits for free after signing up, then prices range from $49/month onwards.
- **Octoparse**: This is a web scraping tool that allows you to extract data from any website without coding. It has a visual interface that lets you build your own crawlers by pointing and clicking on the elements you want to scrape. It also supports cloud extraction, IP rotation, pagination, and scheduling features. However, it may not be able to handle complex websites that use dynamic content or anti-scraping

techniques. It offers a 14-day free trial after signing up, then plans start from $89/month.

- **Import.io**: This is a web scraping tool that specializes in e-commerce data extraction and analysis. It can scrape product information, prices, reviews, ratings, images, and more from any online store or marketplace. It also provides data transformation, integration, visualization, and reporting features. However, it may not be able to scrape other types of websites or data sources. It offers a 30-day free trial after contacting its sales department, then plans start from $299/month.

If you are looking for a fast and reliable API that can handle any kind of website and provide you with high-quality data, ZenRows may be the best option for you.

If you are looking for a free and flexible framework that can scrape large volumes of data using Python, Scrapy may be the best option for you.

If you are looking for a user-friendly tool that can scrape outdated and interactive websites without coding, ParseHub may be the best option for you.

If you are looking for a simple tool that can scrape popular sites with predefined templates and workflows, Apify may be the best option for you.

If you are looking for a cloud-based tool that can scrape and manage data for marketing purposes, Mozenda may be the best option for you.

If you are looking for a cheap and scalable API that can handle proxies, browsers, and CAPTCHAs for you, ScraperAPI may be the best option for you.

If you are looking for a no-code tool that can scrape data from any website with a visual interface, Octoparse may be the best option for you.

If you are looking for an e-commerce tool that can scrape and analyze product data from any online store or marketplace, Import.io may be the best option for you.

AWS Data orchestration Tools:

- **AWS Data Pipeline**: This is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals. With AWS Data Pipeline, you can regularly access your data where it's stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR.
- **AWS Step Functions**: This is a service that lets you coordinate multiple AWS services into serverless workflows. You can use AWS Step Functions to create and run a series of steps, such as tasks, choices, loops, and parallel branches, that can invoke other AWS services such as AWS Lambda, Amazon SNS, Amazon SQS, or AWS Batch. You can also use AWS Step

Functions to handle errors, retries, and timeouts in your workflows.

- **AWS Glue**: This is a fully managed service that provides a serverless data integration platform for data preparation, data cataloging, and data loading. You can use AWS Glue to discover your data sources, extract and transform your data using AWS Glue DataBrew or AWS Glue ETL jobs, and load your data into your target data stores such as Amazon Redshift or Amazon Athena. You can also use AWS Glue to create and manage a centralized metadata repository called the AWS Glue Data Catalog.
- **AWS Glue DataBrew**: This is a visual data preparation tool that enables you to explore, clean, and normalize your data without writing code. You can use AWS Glue DataBrew to interactively apply over 250 transformations to your data using point-and-click operations or suggestions from DataBrew. You can also use AWS Glue DataBrew to profile your data quality and generate statistics and visualizations to understand your data better.
- **Azure Data Factory**: This is a cloud-based data integration service that allows you to create data-driven workflows in the cloud or in hybrid environments. You can use Azure Data Factory to orchestrate and automate the movement and transformation of data from various sources such as Azure Blob Storage, Azure SQL Database, Azure Cosmos DB, or on-premises databases. You can also use Azure Data Factory to integrate with other Azure services such as Azure Databricks, Azure Machine Learning, or Azure Synapse Analytics.

What is a DOM:

It's a copy of the html file but in different language(javascript) plus it has methods and objects and properties to make the html page dynamic (when you click on a button something happens)

is a programming interface that allows you to access and manipulate the document as a tree of nodes .

The DOM is a way for computers to understand and change web pages. When you look at a web page, it's like a big tree with lots of branches. Each branch is called an element, and the DOM helps the computer understand how all the elements are connected. It's like a map that shows the computer where everything is on the page. With the DOM, the computer can change things on the page, like colors or pictures, and make it look different. It's like if you had a coloring book and you could change the colors of the pictures in it .

What is smote analysis:

SMOTE is a technique that helps computers learn how to recognize things better. It's like when you learn how to tell the difference between a cat and a dog. Sometimes, the computer has a hard time telling the difference between two things because there are not enough examples of one of them. SMOTE helps the computer make more examples of the thing that it doesn't have enough of, so it can learn better. It's like if you only had one toy car and you wanted to play with it more, you could make more toy cars that look just like it so you can play with them all together [1].

SMOTE stands for Synthetic Minority Oversampling Technique. It is a technique used to address the issue of imbalanced datasets in binary classification problems. In such problems, the class distribution is skewed, and the minority class has very few samples compared to the majority class. SMOTE generates synthetic samples for the minority class by interpolating between positive instances that lie together in

the feature space. This helps to overcome the overfitting problem posed by random oversampling and contributes to more accurate predictions and better model performance [1].

If you want to learn more about SMOTE, you can check out this Analytics Vidhya article that provides a detailed explanation of SMOTE and its application in improving the performance of classifier models.

Web Scraping libraries:

There are several Python libraries and frameworks that can be used for web scraping. Here are some of the most popular ones:

1. **Beautiful Soup**: A Python library that is used to parse HTML and XML documents. It is a great tool for beginners as it is easy to set up and use. Beautiful Soup uses a hierarchical approach to extracting data from an HTML document, allowing you to extract elements using tags, classes, IDs, names, and other HTML attributes .
2. **Requests**: A Python library that is used to handle HTTP requests. It supports multiple HTTP request types, ranging from GET and POST to PATCH and DELETE. You can control almost every aspect of a request, including headers and responses .

3. **Selenium**: A Python library that is used to automate web browser interactions. It can be used for web scraping by automating the process of clicking buttons, filling out forms, and navigating through pages .
4. **PyQuery**: A Python library that is used to parse HTML documents using jQuery syntax .