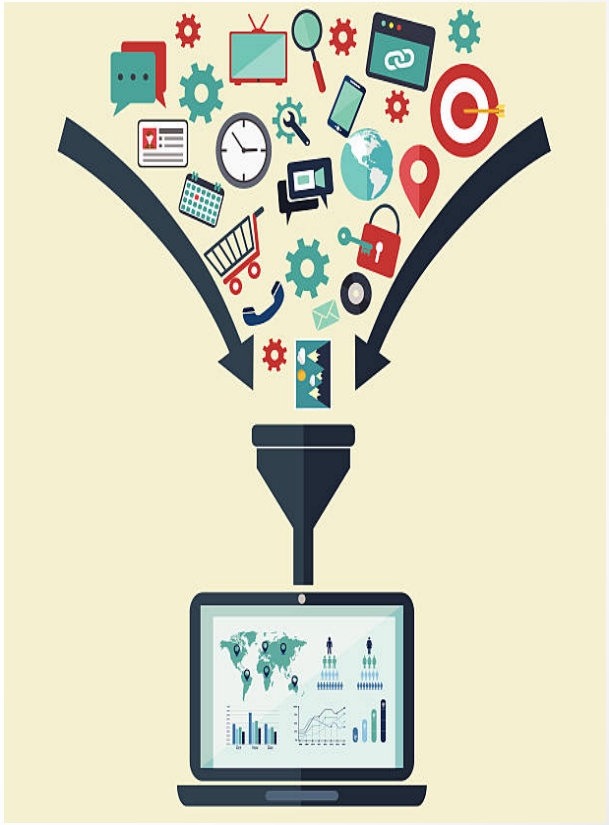
Data Collection

Table Contents

Definition

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection.



Introduction

- During data collection, the researchers must identify the data types, the sources of data, and what methods are being used. We will soon see that there are many different data collection methods.
- There is heavy reliance on data collection in research, commercial, and government fields.
- Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.


Data Collection
Methods
01 Forms and Questionnaires
02 Interview
03 Observation
04 Documents and Records
05 Focus Groups
06 Oral Histories
07 Combination Research
08 Online Tracking
09 Online Marketing Analytics
10 Social Media Monitoring
GlobalPatron

Why Do We Need Data Collection?

- Before a judge makes a ruling in a court case or a general creates a plan of attack, they must have as many relevant facts as possible. The best courses of action come from informed decisions, and information and data are synonymous.
- The concept of data collection isn't a new one, as we'll see later, but the world has changed.

Why Do We Need Data Collection?

- The data collection process has had to change and grow with the times, keeping pace with technology.
- Whether you're in the world of academia, trying to conduct research, or part of the commercial sector, thinking of how to promote a new product, you need data collection to help you make better choices.

Why Do We Need Data Collection?

- While the phrase “data collection” may sound all high-tech and digital, it doesn’t necessarily entail things like computers, big data, and the internet.

- Data collection could mean a telephone survey, a mail-in comment card, or even some guy with a clipboard asking passersby some questions.

Methods of Data Collection

The following are seven primary methods of collecting data in business analytics.

- Surveys
- Transactional Tracking
- Interviews and Focus Groups
- Observation
- Online Tracking
- Forms
- Social Media Monitoring

Methods of Data Collection

- Data collection breaks down into two methods. As a side note, many terms, such as techniques, methods, and types, are interchangeable and depending on who uses them.

- One source may call data collection techniques "methods," for instance.

Data Collection Methods

Primary

Survey/
Questionnaire

Interview

Observation

Experiment

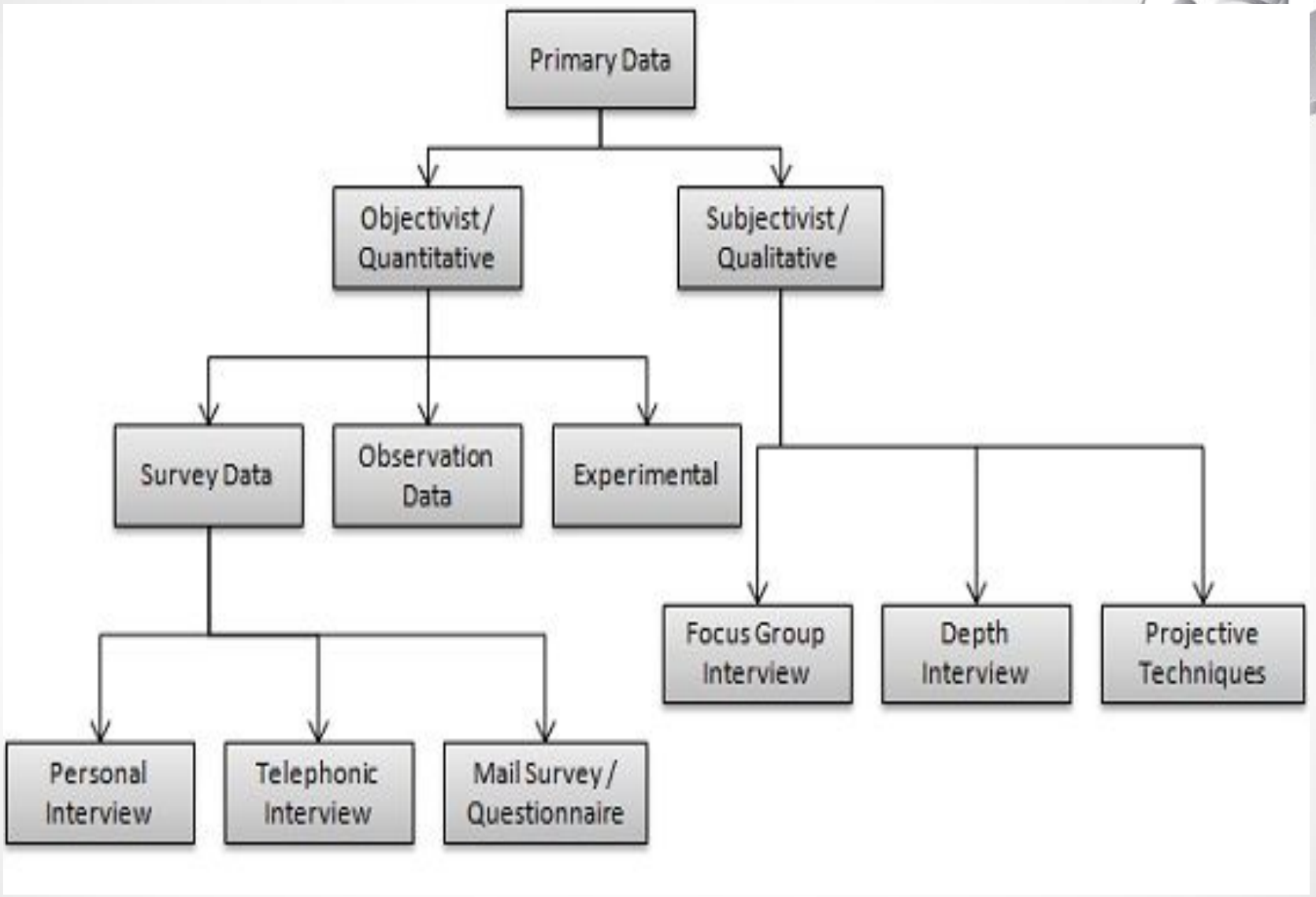
Secondary

Literature
Review

Government
Database

Commercial
Database

Web
Scraping


Primary Data
Objectivist / Quantitative
Subjectivist / Qualitative
Survey Data
Observation Data
Experimental
Focus Group Interview
Depth Interview
Projective Techniques
Personal Interview
Telephonic Interview
Mail Survey / Questionnaire

Methods of Data Collection

Primary

- As the name implies, this is original, first-hand data collected by the data researchers.
- Primary data results are highly accurate provided the researcher collects the information.

Methods of Data Collection

Secondary

- Secondary data is second-hand data collected by other parties and already having undergone statistical analysis.
- This data is either information that the researcher has tasked other people to collect or information the researcher has looked up. Simply put, it's second-hand information.

Primary Data Collection Techniques

Surveys and Questionnaires: Researchers design structured questionnaires or surveys to collect data from individuals or groups. These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms.

Interviews: Interviews involve direct interaction between the researcher and the respondent. They can be conducted in person, over the phone, or through video conferencing. Interviews can be structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational).

Observations: Researchers observe and record behaviors, actions, or events in their natural setting. This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention.

Experiments: Experimental studies involve the manipulation of variables to observe their impact on the outcome. Researchers control the conditions and collect data to draw conclusions about cause-and-effect relationships.

Data Collection Techniques

Focus Groups

- Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.

Secondary Data Collection:

Published Sources: Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data.

Online Databases: Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.

Government and Institutional Records: Government agencies, research institutions, and organizations often maintain databases or records that can be used for research purposes.

Publicly Available Data: Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research.

Past Research Studies: Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyze the data to gain insights or build upon existing knowledge.

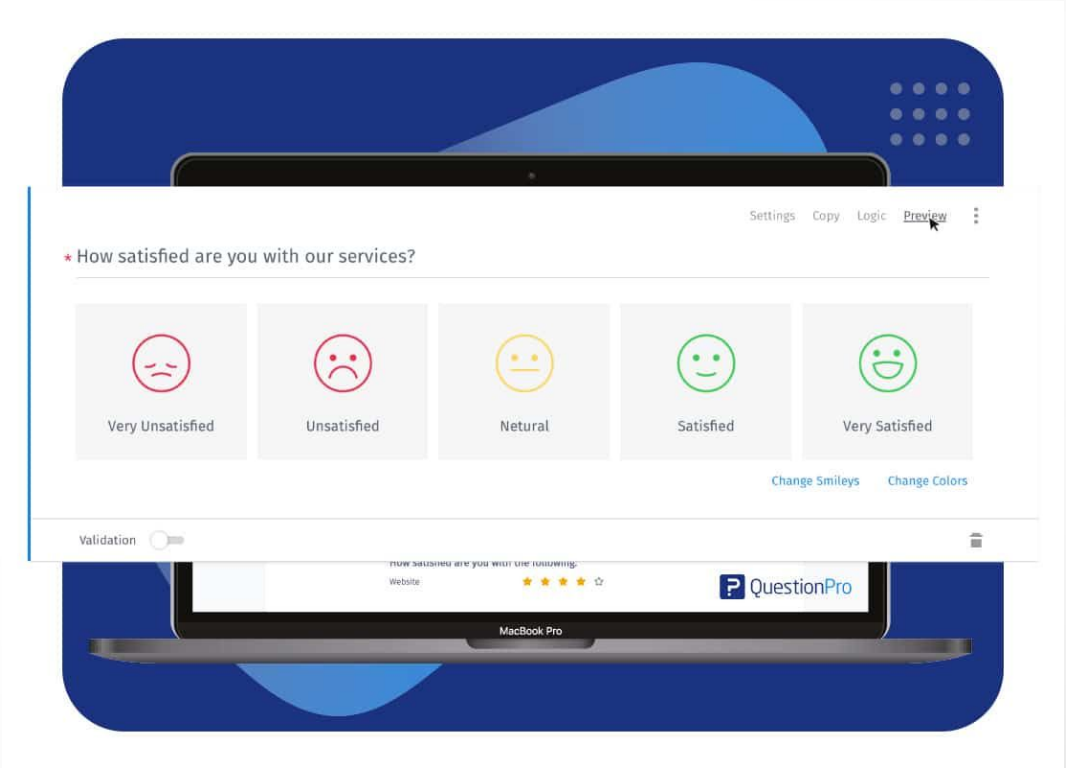
Data Collection Techniques

- Financial Statements
- Sales Reports
- Retailer/Distributor/Deal Feedback
- Customer Personal Information (e.g., name, address, age, contact info)
- Business Journals
- Government Records (e.g., census, tax records, Social Security info)
- Trade/Business Magazines
- The internet

Best Data Collection Tools

QuestionPro

QuestionPro is a survey platform and one of the best data collection tools. They have easy-to-use software with tools for creating, distributing, and analyzing online surveys, polls, forms, and quizzes. With these tools, users can easily gather and analyze data.



Best Data Collection Tools

Magpi

Magpi is an application that provides mobile forms for field data collection. This information can then be put into online dashboards and reports that update in real time or can be added to online systems. Magpi is often used by businesses that want to do maintenance surveys, equipment inspections, site inspections, and progress reports.

Zonka Feedback

It is one of the most widely used data collection tools and feedback collection tools. Zonka is also widely used to analyze customer interviews or surveys.

Zoho

You may develop forms to collect data, share them online, and receive fast alerts with Zoho. There are more than 30 different types of fields, themes that can be changed, templates for different situations, and a simple user interface. It lets you make beautiful and useful forms for all your needs.

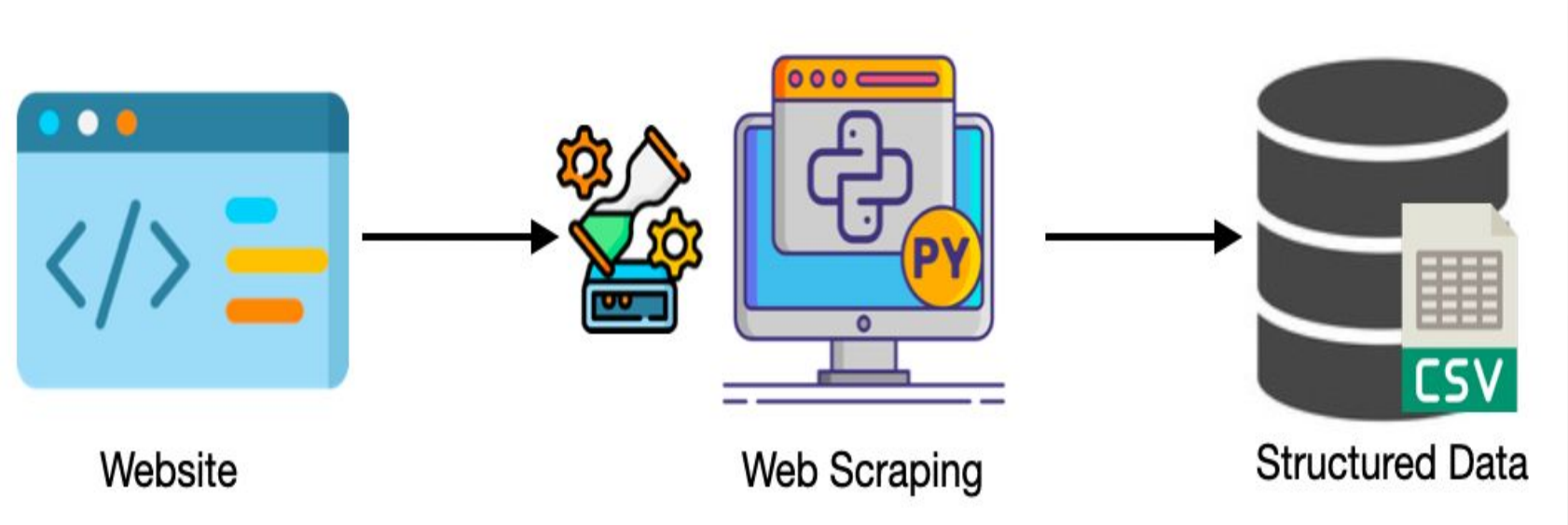
Best Data Collection Tools

Paperform
It is a dependable data collection tool that allows anyone to easily develop forms or product pages. With Paperform, you may gather more than 20 different sorts of data online, including files, eSignatures, emails, text, addresses, and photographs.

Device Magic
Device Magic is one of the best tools for collecting data. It lets you manage your teams at different job sites or survey locations.

web scraping

Web scraping is a crucial technique in data science that involves extracting data from websites. It allows data scientists to gather a large amount of data from the web, which can be used for analysis, modeling, and decision-making



Why Web Scraping is Important?

Data Collection: Web scraping is used to collect data from websites that do not provide an API or other straightforward means of data access.

Real-Time Data: It allows for the collection of up-to-date data that can be used for real-time analytics.

Diverse Data Sources: Scraping enables the aggregation of data from multiple sources to provide comprehensive datasets.

Tools and Libraries for web scraping

Python Libraries: Popular libraries like BeautifulSoup, Scrapy, and Selenium are often used for web scraping in Python.

BeautifulSoup: Ideal for simple scraping tasks and parsing HTML and XML documents.

Scrapy: A more robust framework for building web crawlers and scraping large datasets.

(A web crawler, or spider, is a type of bot that is typically operated by search engines like Google and Bing)

Selenium: Useful for scraping dynamic websites that require interaction, as it automates web browsers.

R Libraries: Packages such as rvest and RSelenium can be used for web scraping in R.

Steps in Web Scraping

Identify the Data: Determine what data you need and where it is located on the website.

Inspect the Website: Use browser developer tools to inspect the HTML structure and identify patterns for the data you want to extract.

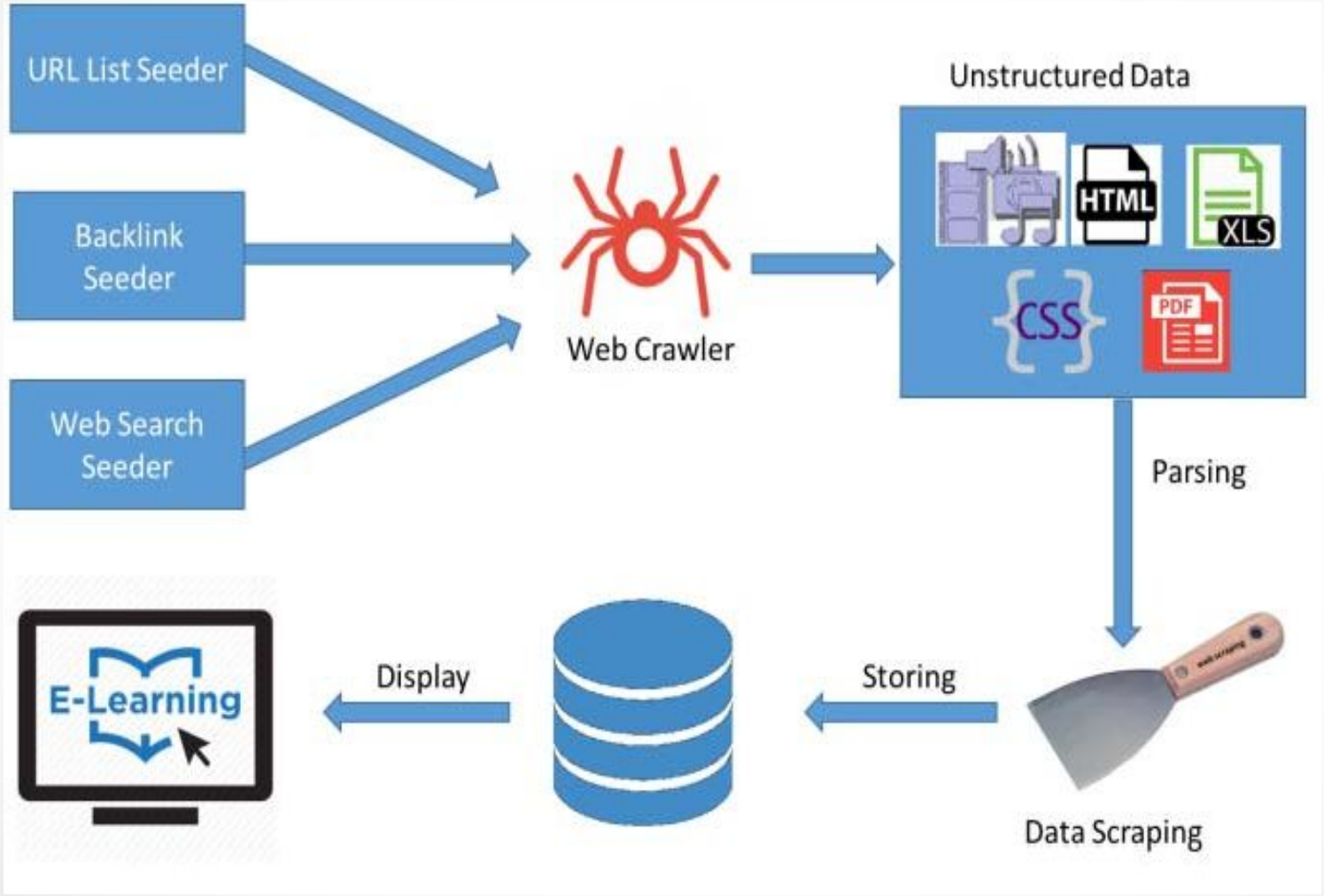
Write the Scraper: Use a web scraping library to write a script that navigates the website and extracts the data.

Store the Data: Save the scraped data in a suitable format (e.g., CSV, JSON, database) for analysis.

Data Cleaning and Preprocessing: Clean the scraped data to handle missing values, duplicates, and inconsistencies.

Analysis: Use the cleaned data for various analyses, such as statistical analysis, machine learning, or visualization.

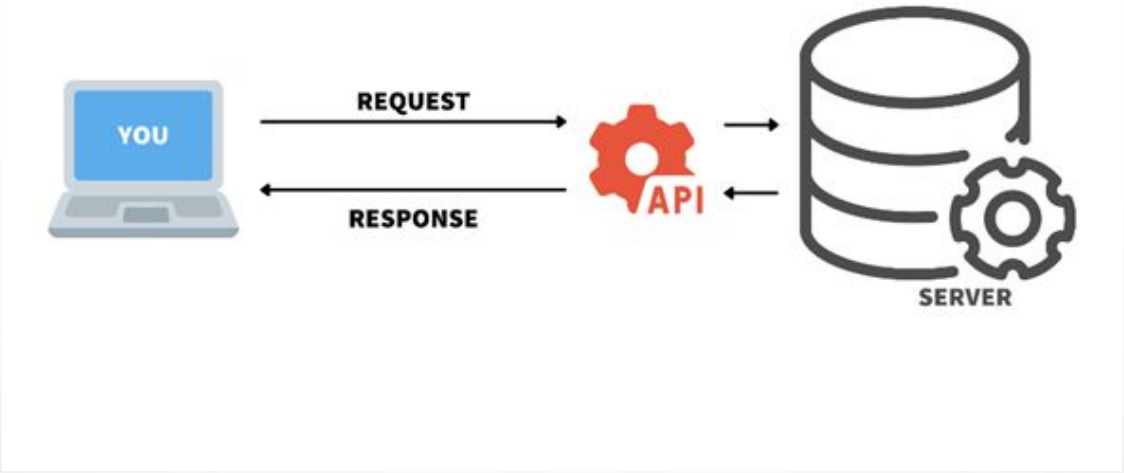
Steps in Web Scraping



What is an API?

An API is a set of rules and protocols that allows different software applications to communicate with each other. It defines the methods and data formats that applications can use to request and exchange information.

APIs (Application Programming Interfaces) play a significant role in data science by providing a standardized way to access data and services. They enable data scientists to interact with external systems, gather data, and integrate various functionalities into their applications.



Examples

Facebook API

Facebook API provides an interface to a large amount of data generated everyday. The innumerable post, comments and shares in various groups & pages produces massive data. And this massive public data provides a large number of opportunities for analyzing the crowd.

It is also incredibly convenient to use Facebook Graph API with both R and python to extract data.

Google Map API

Google Map API is one of the commonly used API. Its applications vary from integration in a cab service application to the popular Pokemon Go.

You can retrieve all the information like location coordinates, distances between locations, routes etc. The fun part is that you can also use this API for creating the distance feature in your datasets as well.

Twitter API

Just like Facebook Graph API, Twitter data can be accessed using the Twitter API as well. You can access all the data like tweets made by any user, the tweets containing a particular term or even a combination of terms, tweets done on the topic in a particular date range, etc.

Twitter data is a great resource for performing the tasks like opinion mining, sentiment analysis.

Types of APIs

Web APIs: These are the most common in data science, allowing access to web services over HTTP/HTTPS. Examples include REST and SOAP APIs.

Library/Framework APIs: Interfaces provided by software libraries or frameworks (e.g., TensorFlow API for building machine learning models).

Operating System APIs: Allow applications to interact with the operating system (e.g., Windows API).

Why APIs are Important in Data Science

Data Access: APIs provide access to large datasets from various sources, such as social media platforms, financial markets, and public databases.

Real-Time Data: They enable the integration of real-time data into applications, essential for dynamic analyses and dashboards.

Automated Workflows: APIs facilitate automation by allowing scripts and applications to interact with other software systems.

Scalability: Using APIs, data scientists can scale their solutions by leveraging cloud services and distributed computing resources.

How to Use APIs in Data Science

Find the Right API: Identify the API that provides the data or functionality you need. Check for documentation and any access restrictions.

Access Credentials: Obtain necessary credentials like API keys or tokens, usually required for authentication.

Read the Documentation: Understand the API endpoints, request methods (GET, POST, PUT, DELETE), and data formats (usually JSON or XML).

Make API Requests: Use HTTP clients like requests in Python or httr in R to send requests to the API and handle responses.

Handle Rate Limits: Be aware of rate limits to avoid exceeding the allowed number of requests in a given period.

Process the Data: Parse and clean the data received from the API to make it suitable for analysis.

Tools and Libraries for APIs

Python: requests, http.client, aiohttp for asynchronous requests.

R: httr, curl for making API calls.

Postman: A tool for testing and developing APIs, providing an interactive interface to send requests and analyze responses.

Example: Using an API in Python

Here's a simple example of using the requests library to fetch data from a public API:

```
import requests

# Define the API endpoint and parameters
url = 'https://api.openweathermap.org/data/2.5/weather'
params = {
    'q': 'London',
    'appid': 'your_api_key',
    'units': 'metric'
}

# Send a GET request to the API
response = requests.get(url, params=params)

# Check if the request was successful
if response.status_code == 200:
    data = response.json()
    print(f"Temperature in London: {data['main']['temp']}°C")
else:
    print('Failed to retrieve data', response.status_code)
```

Data collection done incorrectly, include the following

- Erroneous conclusions that squander resources
- Decisions that compromise public policy
- Incapacity to correctly respond to research inquiries
- Bringing harm to participants who are humans or animals
- Deceiving other researchers into pursuing futile research avenues
- The study's inability to be replicated and validated

Limitations of Data Collection

Quality assurance and quality control are two strategies that help protect data integrity and guarantee the scientific validity of study results.

Each strategy is used at various stages of the research timeline:

- Quality control - tasks that are performed both after and during data collecting
- Quality assurance - events that happen before data gathering starts

10 benefits of collecting customer data

1. Know your customer.
2. Identify high-value audiences.
3. Turn one-time buyers into repeat customers.
4. Improve and personalize the in-store experience.
5. Prevent customer churn.
6. Improve digital advertising (without violating privacy regulations).
7. Improve cross-functional collaboration.
8. Keep track of your marketing performance.
9. Save costs.
10. Offer seamless customer experiences.

Conclusion

- ✔ Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Thanks