Data Science with Python

Data Extraction

Agenda

01 Data Extraction

02 Databases

03 SQL

04 Queries

05 Subqueries

06 Joins

07 HTML

08 DOM

09 Beautiful Soup

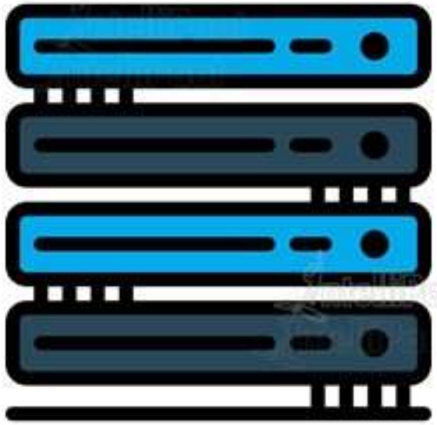
10 HTML Parsers

What is Data Extraction?

What is Data Extraction?

Data extraction is one of the most important steps in Data Science

It is the process of retrieving data from various sources to be used in our Data Science process



Why do we need Data Extraction?

Why do we need Data Extraction?

Data extraction is performed in order to gather data from diverse sources and store it in a data repository

This data can later be cleaned and transformed to be used to derive important insights or to make predictions

Data Extraction Sources

Data Extraction Sources

Data can be extracted from various sources to be used in Data Science for further processing. Some of these sources are:

Databases

Databases store structured data such as tables with relationships, constraints etc. to keep data in a consistent state

Internet

Websites on the Internet contain unstructured data such as text, images, audio, video, etc.

Data Extraction Sources

Data can be extracted from various sources to be used in Data Science for further processing. Some of these sources are:

Databases

Databases store structured data such as tables with relationships, constraints etc. to keep data in a consistent state

Internet

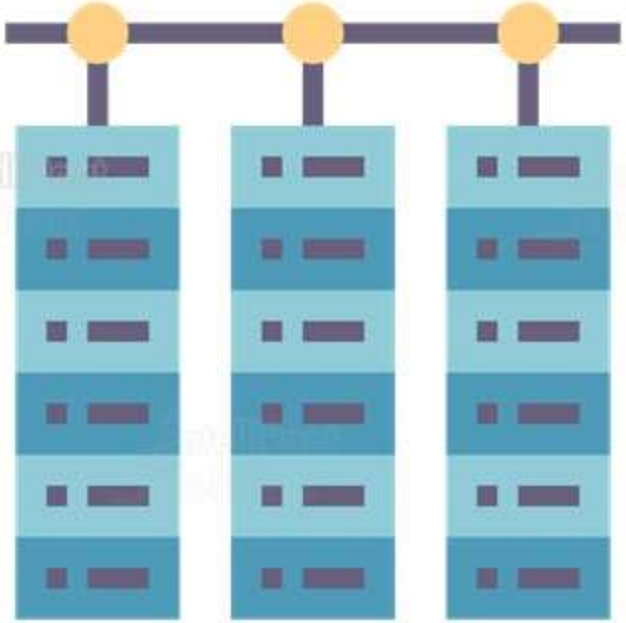
Websites on the Internet contain unstructured data such as text, images, audio, video, etc.

What is a Database?

What is a Database?

A database is an organized and structured collection of interrelated data

Databases are used to store large amounts of data, along with strongly defined constraints



Why do we need Databases?

Why do we need Databases?

We use databases because they provide several benefits over traditional data storage systems

- Security
- Consistency
- Speed
- Ease of Use

Databases provide several security measures so that only authorized users can make necessary changes to the data stored in them

Why do we need Databases?

We use databases because they provide several benefits over traditional data storage systems

Security

Consistency

Speed

Ease of Use

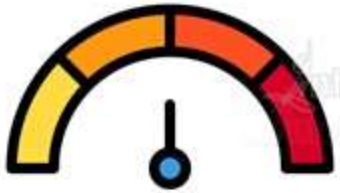
Databases allow us to put constraints on data being stored so that the data always remains in a consistent and accurate state

Why do we need Databases?

We use databases because they provide several benefits over traditional data storage systems

- Security
- Consistency
- Speed
- Ease of Use

Databases allow us to query large amounts of data and extract the desired information quickly and efficiently



Why do we need Databases?

We use databases because they provide several benefits over traditional data storage systems

- Security
- Consistency
- Speed
- **Ease of Use**

Databases make it easy for us to manage and manipulate large amounts of data

What is SQL?

Hands-on: Installing MySQL

What is SQL?

SQL stands for **Structured Query Language**

SQL is a language that is used to interact with relational database management systems

Hands-on: Creating a Database and a Table

SQL Queries

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The SELECT statement is used to retrieve data from a database table in a database management system

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The INSERT statement is used to insert data into a single or multiple database tables

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The UPDATE statement allows us to update a single or multiple rows in a database table

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The DELETE statement allows us to delete a single or multiple rows in a database table

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The WHERE clause is used to put filters or conditions on queries being performed so that these queries only affect the necessary rows

SQL Queries

There are several types of SQL queries that we can perform when interacting with a database

- SELECT
- INSERT
- UPDATE
- DELETE
- WHERE
- HAVING

The HAVING clause is used as the WHERE clause cannot be used with aggregate functions, such as COUNT, SUM, etc.

Hands-on: CRUD Operations

Hands-on: Aggregation and the HAVING Clause

What are Subqueries?

What are Subqueries?

In SQL, a subquery is just a query that is embedded within another query, which is called the main query

A subquery is usually used within a WHERE, HAVING, or FROM clause

Why do we need Subqueries?

Why do we need Subqueries?

As said, a subquery is a query embedded within another query

This is done in order to restrict the amount of data the main query or outer query will work on

Why do we need Subqueries?

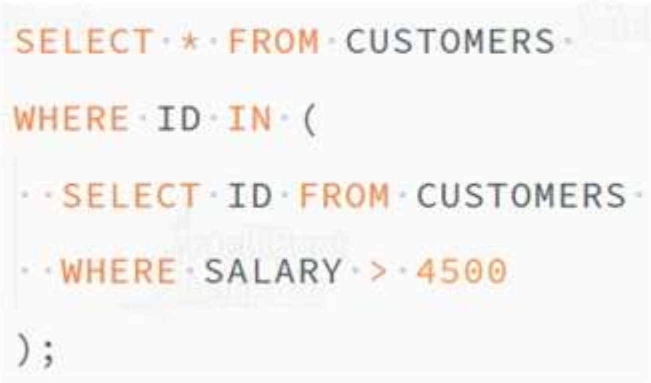
Suppose wish to find the average age of customers from India whose salary is more than 80K in your database

Instead of running two queries, one for getting data and one to find average age, you can use a subquery

Why do we need Subqueries?

This is especially useful when data is stored in an external database system and every read operations has certain cost

By reducing the amount of queries you can save a read operation as well as some bandwidth on the network



```sql
SELECT * FROM CUSTOMERS
WHERE ID IN (
  SELECT ID FROM CUSTOMERS
  WHERE SALARY > 4500
);
```

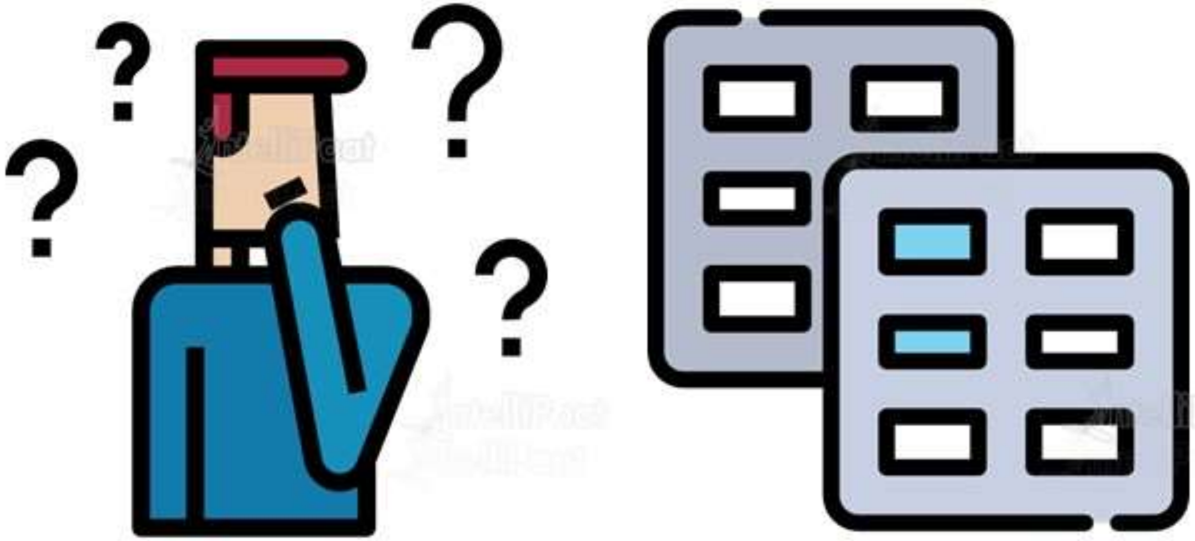
Hands-on: Subqueries

What are Joins?

What are Joins?

A join in SQL is a clause that is used to merge rows from two tables

Joins can be performed based on a common column between the two tables

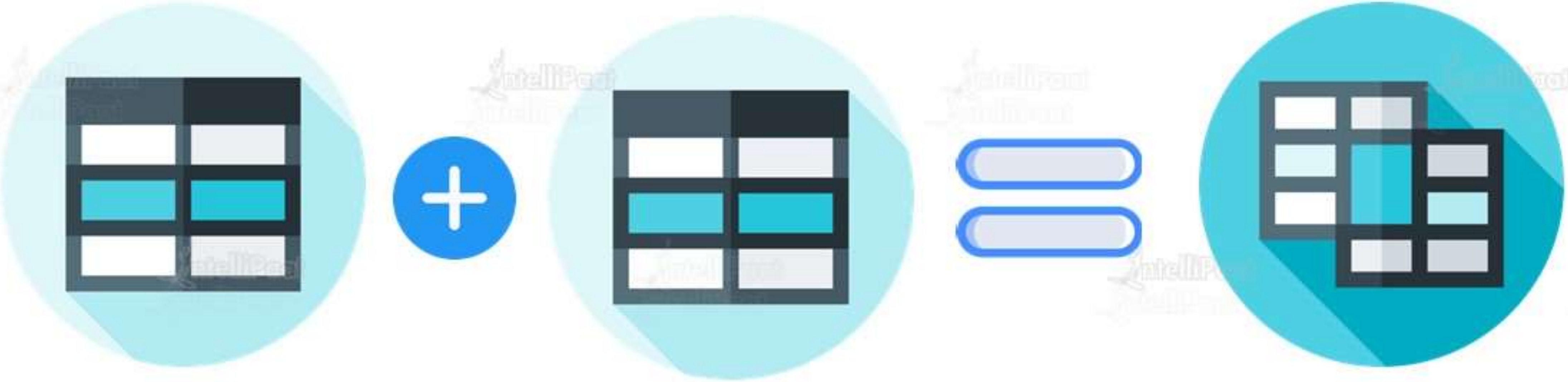


Why do we need Joins?

Why do we need Joins?

Databases are required to store huge volumes of data in an organized way, which can be very difficult sometimes

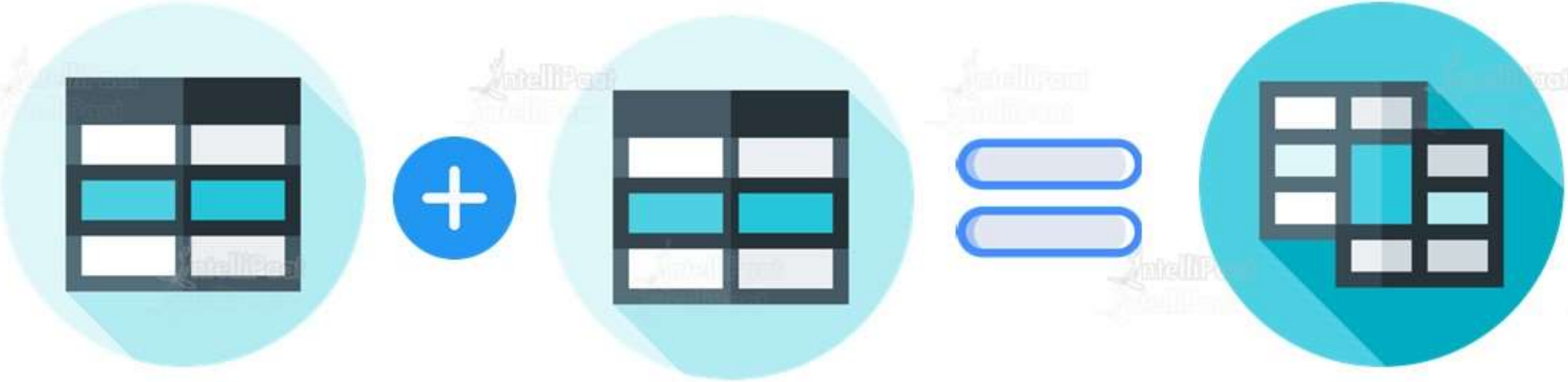
To reduce duplication and maintain consistency of data and relationships a large table is split into multiple small tables



Why do we need Joins?

In large organizations, data is split into multiple strongly interrelated tables

Joins are used to extract data from these interrelated tables in a single query



Types of Joins

Types of Joins

There are four types of joins in SQL

- Inner Join
- Left Join
- Right Join
- Full Join

Inner joins return rows of data that have matching values in the common column from both the left table and the right table

Types of Joins

There are four types of joins in SQL

- Inner Join
- **Left Join**
- Right Join
- Full Join

Left joins return all rows of data from the left table and only the rows that have matching values in the common column from the right table

Types of Joins

There are four types of joins in SQL

- Inner Join
- Left Join
- Right Join
- Full Join

Right joins return all rows of data from the right table and only the rows that have matching values in the common column from the left table

Types of Joins

There are four types of joins in SQL

- Inner Join
- Left Join
- Right Join
- Full Join

Full joins return all rows of data that have matching values in the common column from either the left table or the right table

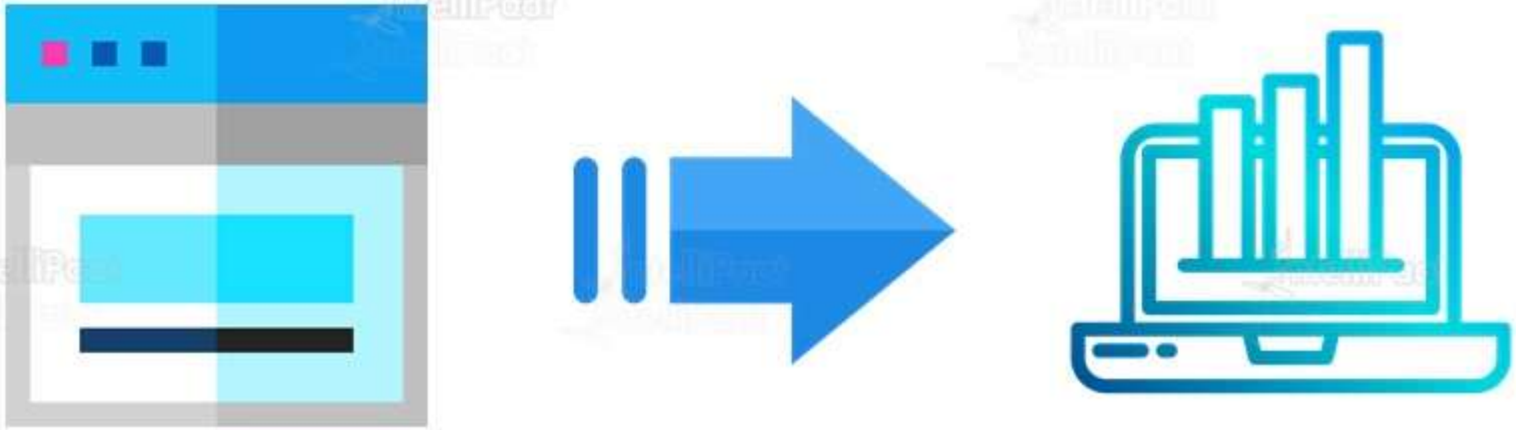
Hands-on: Joins

What is Web Scraping?

What is Web Scraping?

Web scraping is the process by which we extract useful data from a web page on the Internet

Web scraping is done by parsing the HTML content of a website and then navigating to the desired content and extracting it from the web page

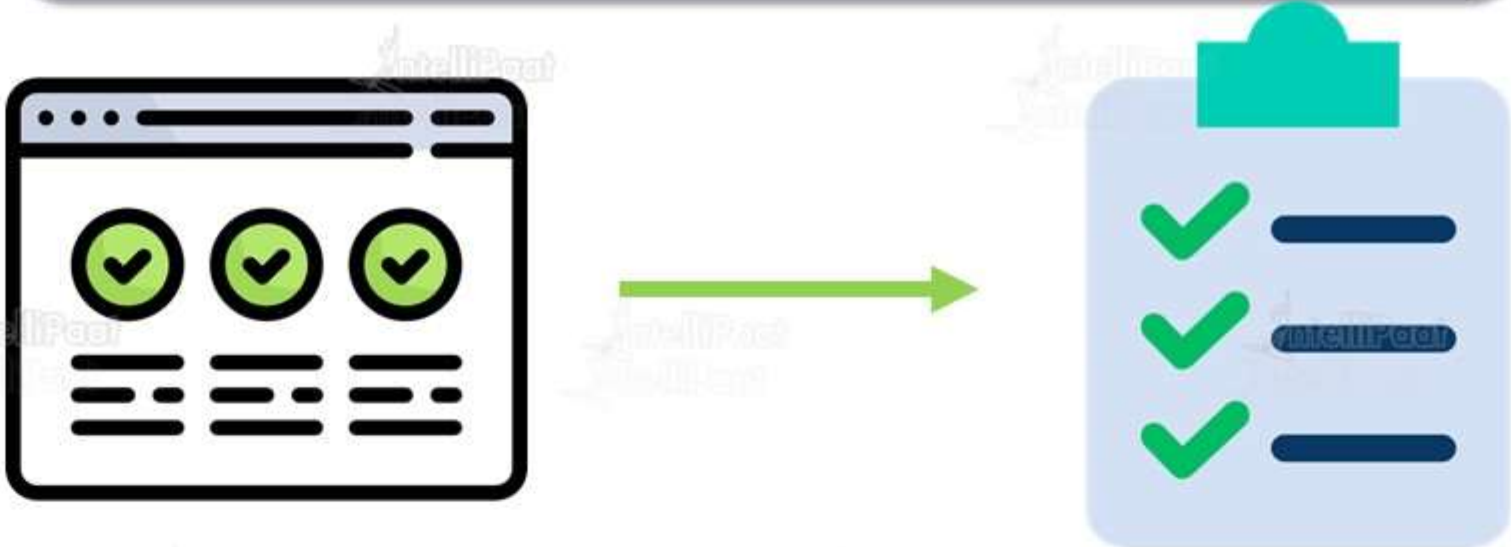


Why use Web Scraping?

Why use Web Scraping?

Sometimes the data we wish to use or stored is on the web

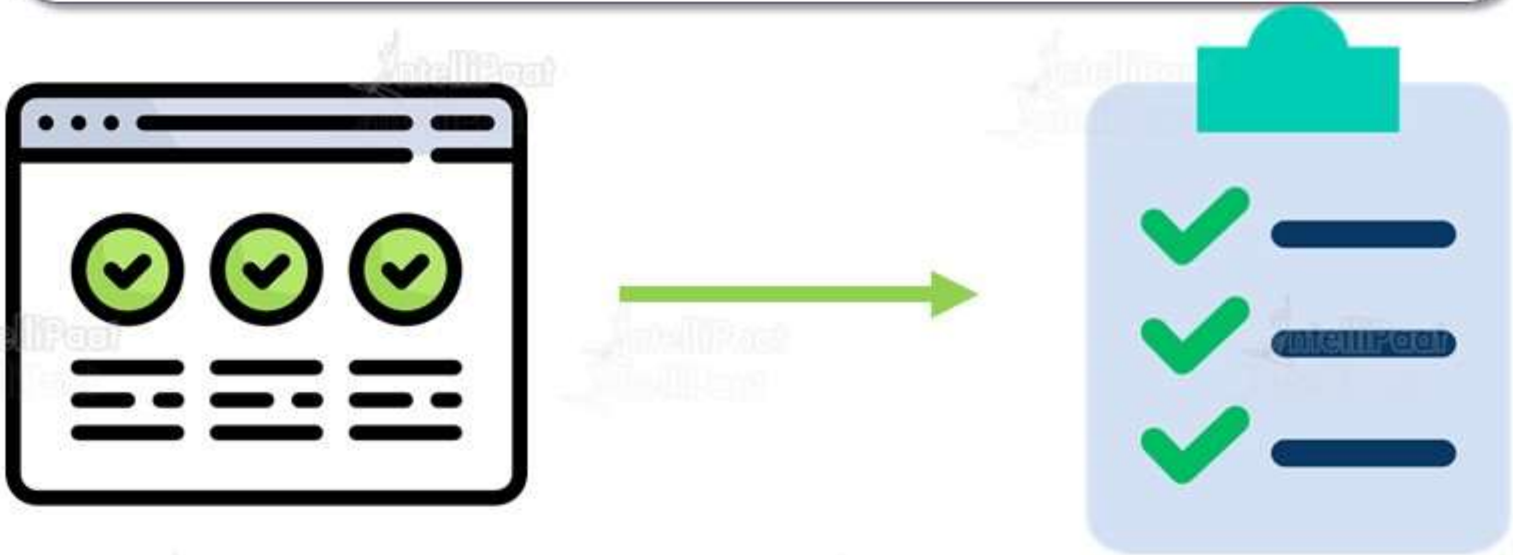
On the web data is stored in a web page which is written in HTML. When we wish to extract data from an HTML page we use web scraping which takes the HTML code of a web page and allows us to define a process to extract useful information out of it



Why use Web Scraping?

You can use Web Scraping in many different fields

In lead generation, market analysis, and in our case collecting and testing data set for Machine Learning. For example, contact details of businesses as well as individuals from yellow pages websites

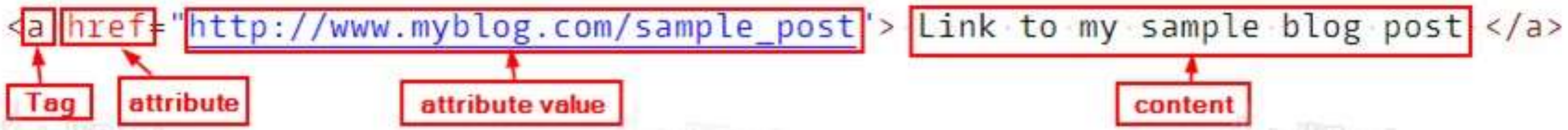


What is HTML?

What is HTML?

HTML (Hypertext Markup Language) is a markup language used to describe the structure and content of a web page

HTML contains data in the from of text encapsulated within tags and attributes, and this information is what we obtain through scraping a web page



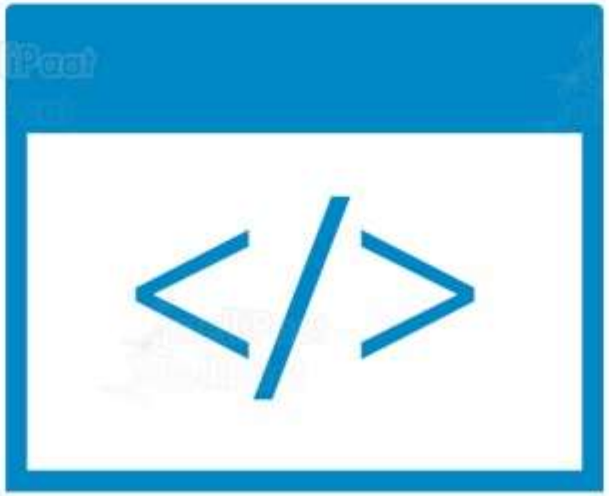
```
<a href="http://www.myblog.com/sample_post"> Link to my sample blog post </a>
```

HTML Tags

HTML Tags

In HTML, tags are the basic building blocks of the structure of a web page

A tag helps a web browser decide what an element on a web page is, e.g., an image, a paragraph, etc.



HTML Attributes

HTML Attributes

HTML attributes are added to HTML tags to provide additional information

Attributes come in key–value pairs, e.g., alt=“sample image"

HTML Selectors

HTML Selectors

In HTML, if we wish to refer to a particular element on the web page, we need to specify some sort of selector on that element

A selector is basically an identifier that is put on the HTML elements so that they can be programmatically accessed later

HTML Selectors

There are two types of selectors available in HTML

ID

Class

An ID is used as a selector when we wish to identify a single element on a web page. An ID can only be assigned to a single element on the page

HTML Selectors

There are two types of selectors available in HTML

ID

Class

A class is used as a selector when we wish to access multiple elements on a web page. A class can be assigned to multiple elements on the page

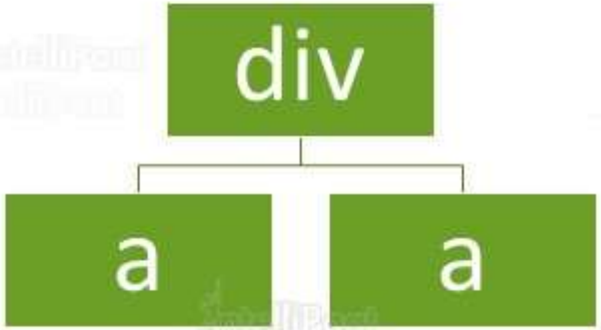
HTML DOM

HTML DOM

When HTML is parsed, it is converted into a hierarchical tree structure

This tree structure is called Document Object Model (DOM). This tree structure is what we use when we access an HTML element on the web page

```
<div>
   <a href="jane">Jane</a>
   <a href="john">John</a>
</div>
```



Beautiful Soup

Beautiful Soup

Beautiful Soup is a Python package that is used to scrape data from web pages

It uses parsing libraries to parse the HTML content of a web page that can then be used to extract contents from that web page

HTML Parsers

HTML Parsers

Beautiful Soup can use either the Python libraries built in HTML parser or a third-party HTML parser such as lxml or html5lib

‘lxml’ is the fastest and the most reliable parser to be used with the Beautiful Soup package

Hands-on: Web Scraping

IntelliPaat

☎ India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

✉ support@intellipaat.com

24/7 Chat with Our Course Advisor