Hands-on Lab : Web Scraping

Estimated time needed: **30 to 45** minutes

Objectives

In this lab you will perform the following:

- Extract information from a given web site
- Write the scraped data into a csv file.

Extract information from the given web site

You will extract the data from the below web site:

In [1]:
```
#this url contains the data you need to scrape
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA032
```

The data you need to scrape is the **name of the programming language** and **average annual salary**.
It is a good idea to open the url in your web broswer and study the contents of the web page before you start to scrape.

Import the required libraries

In [2]:
```
# Your code here
from bs4 import BeautifulSoup # this module helps in web scrapping.
import requests  # this module helps us to download a webpage
```

Download the webpage at the url

In [3]:
```
#your code goes here
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA032
data  = requests.get(url).text
```

Create a soup object

In [4]:
```
#your code goes here
soup = BeautifulSoup(data,"html.parser")  # create a soup object using the variable
```

In [10]:
```
table = soup.find('table')
for i, row in enumerate(table.find_all('tr')[1:], start=1):  # Skip header
    cols = row.find_all('td')
    print(f"Row {i} has {len(cols)} columns")
    for j, col in enumerate(cols, start=1):
        print(f"  Column {j}: {col.text.strip()}")
```

```
Row 1 has 5 columns
  Column 1: 1
  Column 2: Python
  Column 3: Guido van Rossum
  Column 4: $114,383
  Column 5: Easy
Row 2 has 5 columns
  Column 1: 2
  Column 2: Java
  Column 3: James Gosling
  Column 4: $101,013
  Column 5: Easy
Row 3 has 5 columns
  Column 1: 3
  Column 2: R
  Column 3: Robert Gentleman, Ross Ihaka
  Column 4: $92,037
  Column 5: Hard
Row 4 has 5 columns
  Column 1: 4
  Column 2: Javascript
  Column 3: Netscape
  Column 4: $110,981
  Column 5: Easy
Row 5 has 5 columns
  Column 1: 5
  Column 2: Swift
  Column 3: Apple
  Column 4: $130,801
  Column 5: Easy
Row 6 has 5 columns
  Column 1: 6
  Column 2: C++
  Column 3: Bjarne Stroustrup
  Column 4: $113,865
  Column 5: Hard
Row 7 has 5 columns
  Column 1: 7
  Column 2: C#
  Column 3: Microsoft
  Column 4: $88,726
  Column 5: Hard
Row 8 has 5 columns
  Column 1: 8
  Column 2: PHP
  Column 3: Rasmus Lerdorf
  Column 4: $84,727
  Column 5: Easy
Row 9 has 5 columns
  Column 1: 9
  Column 2: SQL
  Column 3: Donald D. Chamberlin, Raymond F. Boyce.
  Column 4: $84,793
  Column 5: Easy
Row 10 has 5 columns
  Column 1: 10
```

```
Column 2: Go
Column 3: Robert Griesemer, Ken Thompson, Rob Pike.
Column 4: $94,082
Column 5: Difficult
```

Scrape the `Language name` and `annual average salary`.

In [11]:
```python
# Find the table
table = soup.find('table')  # or use soup.find('table', {'class': 'your-table-class
table = soup.find('table')
language_salary = []

for row in table.find_all('tr')[1:]:  # Skip header
    cols = row.find_all('td')
    if len(cols) >= 4:
        language = cols[1].text.strip()  # Column 2: Language
        salary = cols[3].text.strip()    # Column 4: Salary
        language_salary.append((language, salary))

# Display results
for lang, sal in language_salary:
    print(f"{lang}: {sal}")
```

```
Python: $114,383
Java: $101,013
R: $92,037
Javascript: $110,981
Swift: $130,801
C++: $113,865
C#: $88,726
PHP: $84,727
SQL: $84,793
Go: $94,082
```

Save the scrapped data into a file named *popular-languages.csv*

In [12]:
```python
# your code goes here
import pandas as pd

# Assuming you have a list of tuples like:
# [('Python', '$114,383'), ('Java', '$101,013'), ...]
language_salary = [
    ('Python', '$114,383'),
    ('Java', '$101,013'),
    ('R', '$92,037'),
    ('Javascript', '$110,981'),
    ('Swift', '$130,801'),
    ('C++', '$113,865'),
    ('C#', '$88,726'),
    ('PHP', '$84,727'),
    ('SQL', '$84,793'),
    ('Go', '$94,082')
]

# Convert to DataFrame
df = pd.DataFrame(language_salary, columns=["Language", "Average Salary"])
```

```
# Save to CSV
df.to_csv("popular-languages.csv", index=False)

print("✅ Data saved to 'popular-languages.csv'")
```

```
✅ Data saved to 'popular-languages.csv'
```

Authors

Ramesh Sannareddy

Other Contributors

Rav Ahuja

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-17	0.1	Ramesh Sannareddy	Created initial version of the lab