Web Scraping Basics with Python

Clarke Bishop
BIG DATA ENGINEER

@ClarkeBishop www.clarkebishop.com

Overview

Python scraping environment

Download a page to scrape

Extract data to Pandas

Python Scraping Environment

Demo

Jupyter Lab

Pyenv and pipenv

Install required packages

Verify your development environment

Download a Page to Scrape

Demo

Import the needed Python packages

Write Python code to download and save a local copy of the HTML page

Use Jupyter Lab's file pane

Inspect the HTML page with Jupyter

More Chrome Developer Tools tricks

Extract Data to Pandas

Demo

Iterate and clean data

Refine selectors to get the right data

Resolve common scraping problems

List of lists for Pandas

Create a Pandas dataframe

Summary

Python scraping environment

- pyenv
- pipenv

Download a page to scrape

- Requests to download
- Jupyter Lab

Extract and clean data

Load into Pandas