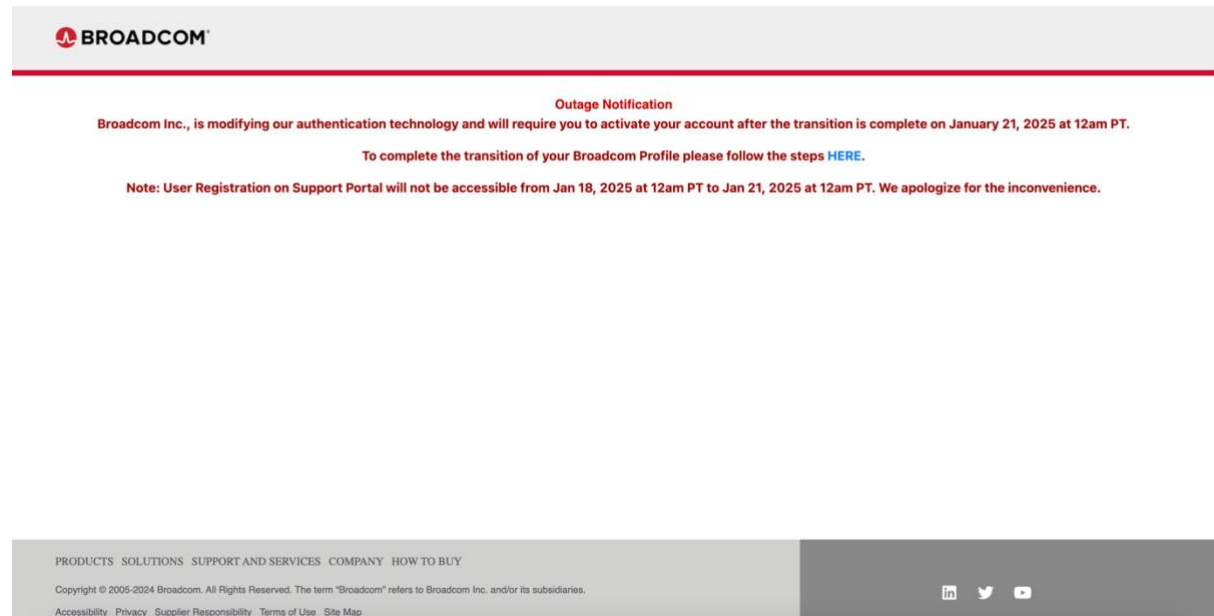


Lab 1 Solution

Due to a temporary block on Broadcom registration until January 21, I used an Amazon EC2 instance for Lab 1 instead. The instance runs the same Linux and "Ubuntu (64-bit)" version as the Broadcom VM. I will switch back to Broadcom's VM once the registration is completed.



The screenshot shows a Broadcom banner with the following text:

BROADCOM

Outage Notification
Broadcom Inc., is modifying our authentication technology and will require you to activate your account after the transition is complete on January 21, 2025 at 12am PT.

To complete the transition of your Broadcom Profile please follow the steps [HERE](#).

Note: User Registration on Support Portal will not be accessible from Jan 18, 2025 at 12am PT to Jan 21, 2025 at 12am PT. We apologize for the inconvenience.

PRODUCTS SOLUTIONS SUPPORT AND SERVICES COMPANY HOW TO BUY

Copyright © 2005-2024 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.
Accessibility Privacy Supplier Responsibility Terms of Use Site Map

LinkedIn Twitter YouTube

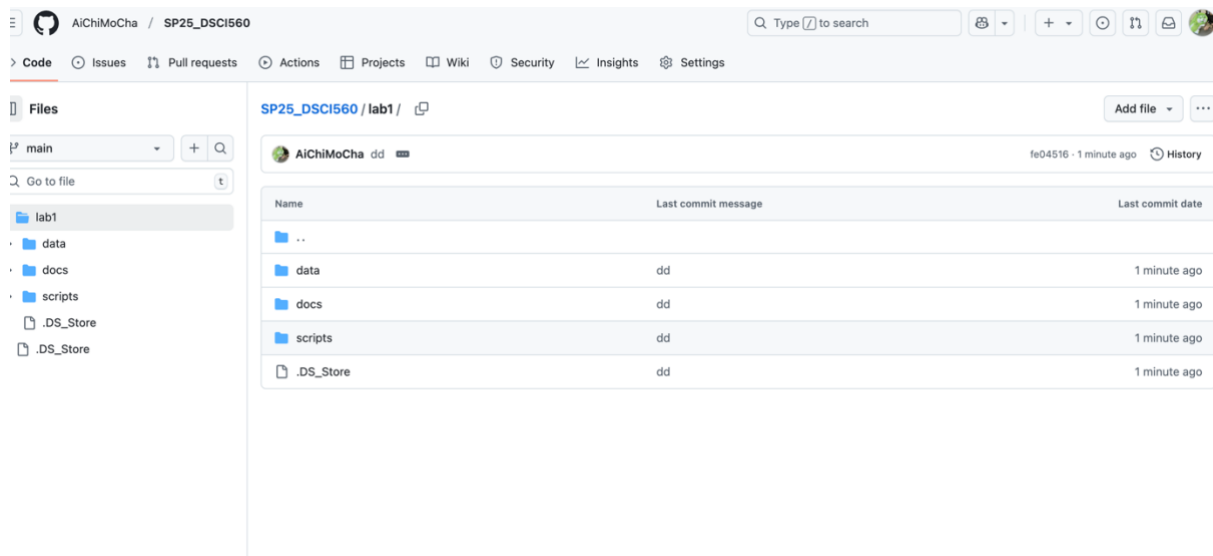
A screenshot of EC2 instance

The image shows two screenshots related to an AWS EC2 instance.

The top screenshot is a terminal window from an Ubuntu 24.04.1 LTS instance. It displays system information as of Sun Jan 19 05:25:05 UTC 2025. The system load is 0.01, with 129 processes and 1 user logged in. The instance's IPv4 address is 172.31.18.40. It also shows Ubuntu Pro security updates and a last login timestamp of Sun Jan 19 05:23:18 2025 from 166.196.78.145.

The bottom screenshot is the AWS Management Console, specifically the 'Instances' page. It shows a table with one instance: 'launch-wizard-1' with key name 'DSCI_560', launched on 2025/01/18 19:10 GMT-8, running Linux/UNIX. Below the table, the monitoring dashboard for instance 'i-039476738090193a5' is visible, showing graphs for Average read latency, Average write latency, Read throughput, Write throughput, Read operations, Write operations, and Average queue length.

The Linux Ubuntu instance is spin up using GitHub for the repository:



2.1. Get Familiar with Linux and Python

Playing around with Linux Terminal

```
ubuntu@ip-172-31-18-40:~$ pip3 --version
pip 24.0 from /usr/lib/python3/dist-packages/pip (python 3.12)
ubuntu@ip-172-31-18-40:~$ mkdir ~/Chenxiao_yu_6024079123
cd ~/Chenxiao_yu_6024079123
mkdir data scripts
touch scripts/task_1.py
ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123$ ls
data  scripts
ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123$ ls scripts/
task_1.py
ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123$ nano scripts/task_1.py
```

2.2. A basic Python Script

Task1 details

```
[ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ ls
task_1.py
ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ cat task_1.py
name = input("Enter your name: ")
print(f"Hello, {name}!")
```

Task1

```
ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123$ python3 scripts/task_1.py
Enter your name: Chenxiao
Hello, Chenxiao!
```

2.3. Web Scraping Task

Task2 details (partial – full in github)

```
[(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ cat web_scraper.py
import os
import logging
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from bs4 import BeautifulSoup

# 设置日志
logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(levelname)s - %(message)s')

# 配置路径
class Config:
    ROOT_DIR = os.path.dirname(os.path.abspath(__file__))
    DATA_DIR = os.path.join(ROOT_DIR, "../data")
    RAW_DATA_DIR = os.path.join(DATA_DIR, "raw_data")
    PROCESSED_DATA_DIR = os.path.join(DATA_DIR, "processed_data")
    BASE_URL = "https://www.cnbc.com/world/?region=world"
```

Task2

+ A snapshot of the webpage from where the data is to be scraped ([CNBC](https://www.cnbc.com/world/?region=world)).

The screenshot shows the CNBC website with a red header bar containing a breaking news alert: "Apple, Google remove TikTok from stores as app halts service in US". Below the header is a navigation bar with the CNBC logo and various market categories. A table of market indices is displayed, showing gains for the DJIA, S&P 500, and NASDAQ, and a slight dip for the VIX. A headline states: "Dow surges more than 300 points, S&P 500 posts best week since period following Trump's election". Below this is a Goldman Sachs banner about 2025 US GDP forecasts. Quick links for various market segments are provided, followed by a "My Portfolio" section and a "LATEST NEWS" section featuring the TikTok news story.

Index	Value	Change
DJIA	43,487.83	+334.70 +0.78%
S&P 500	5,996.66	+59.32 +1.00%
NASDAQ	19,630.20	+291.91 +1.51%
RUSS 2K*	2,275.88	+9.09 +0.40%
VIX	15.97	-0.63 -3.80%

Headline: Dow surges more than 300 points, S&P 500 posts best week since period following Trump's election

Goldman Sachs: According to Goldman Sachs Research, US GDP will grow 2.5% in 2025, outperforming consensus forecasts.

Quick Links: Pro: My Portfolio, Europe market moves, U.S. futures, Mineral 'gold rush', BOE rate cuts, EU cyber rules, BRICS bloc

My Portfolio: PRO, My Portfolio, + LINK YOUR PORTFOLIO

LATEST NEWS:

- 26 MIN AGO: Apple, Google remove TikTok from stores as app halts service in US
- 9 HOURS AGO: Perplexity AI makes a bid to merge with TikTok LLC

```
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ python web_scraper.py
2025-01-19 04:03:39,266 - INFO - Initializing ChromeDriver...
2025-01-19 04:03:41,312 - INFO - Fetching page content...
2025-01-19 04:04:28,821 - INFO - Extracting Market Banner and Latest News...
2025-01-19 04:04:28,829 - INFO - HTML content saved to /home/ubuntu/Chenxiao_yu_6024079123/scripts/../data/raw_data/web_data.html
2025-01-19 04:04:28,831 - INFO - Printing the first ten lines of the saved HTML file:
=== Market Banner ===
<div class="MarketsBanner-marketData" id="market-data-scroll-container">
<a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.DJI">
<div class="MarketCard-row">
<span class="MarketCard-symbol">
DJIA
</span>
<span class="MarketCard-stockPosition">
43,487.83
</span>
```

Task2 output -- under "raw_data" folder named "web_data.html" (partial screenshot)

```
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/data/raw_data$ ls
web_data.html
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/data/raw_data$ cat web_data.html
=== Market Banner ===
<div class="MarketsBanner-marketData" id="market-data-scroll-container">
  <a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.DJI">
    <div class="MarketCard-row">
      <span class="MarketCard-symbol">
        DJIA
      </span>
      <span class="MarketCard-stockPosition">
        43,487.83
      </span>
    </div>
    <div class="MarketCard-row">
      <span aria-hidden="true" class="MarketCard-triangle-up">
      </span>
      <div class="MarketCard-changeData">
        <span class="MarketCard-changesPts">
```

2.4. Data Filtering Task

Task3 code (partial screenshots)

```
[(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ cat data_filter.py
import os
import csv
import logging
from bs4 import BeautifulSoup

# 设置日志
logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(levelname)s - %(message)s')

# 配置路径
class Config:
    ROOT_DIR = os.path.dirname(os.path.abspath(__file__))
    DATA_DIR = os.path.join(ROOT_DIR, "../data")
    RAW_DATA_DIR = os.path.join(DATA_DIR, "raw_data")
    PROCESSED_DATA_DIR = os.path.join(DATA_DIR, "processed_data")
    WEB_DATA_FILE = os.path.join(RAW_DATA_DIR, "web_data.html")
    MARKET_DATA_CSV = os.path.join(PROCESSED_DATA_DIR, "market_data.csv")
    NEWS_DATA_CSV = os.path.join(PROCESSED_DATA_DIR, "news_data.csv")
```

Task3 running

```
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/scripts$ python data_filter.py
2025-01-19 04:38:38,092 - INFO - Reading web_data.html file...
2025-01-19 04:38:38,097 - INFO - Extracting market banner data...
2025-01-19 04:38:38,097 - INFO - Saving data to /home/ubuntu/Chenxiao_yu_6024079123/scripts/../data/processed_data/market_data.csv...
2025-01-19 04:38:38,097 - INFO - CSV file created at /home/ubuntu/Chenxiao_yu_6024079123/scripts/../data/processed_data/market_data.csv
2025-01-19 04:38:38,098 - INFO - Extracting latest news data...
2025-01-19 04:38:38,098 - INFO - Saving data to /home/ubuntu/Chenxiao_yu_6024079123/scripts/../data/processed_data/news_data.csv...
2025-01-19 04:38:38,098 - INFO - CSV file created at /home/ubuntu/Chenxiao_yu_6024079123/scripts/../data/processed_data/news_data.csv
```

Task3 outputs - market data

```
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/data/processed_data$ cat market_data.csv
symbol,stock_position,change_pct
DJIA,"43,487.83",N/A
S&P 500,"5,996.66",N/A
NASDAQ,"19,630.20",N/A
RUSS 2K*, "2,275.88",N/A
VIX,15.97,N/A
```

Task3 outputs - news data

```
(venv) ubuntu@ip-172-31-18-40:~/Chenxiao_yu_6024079123/data/processed_data$ cat news_data.csv
timestamp,title,link
8 Hours Ago,Perplexity AI makes a bid to merge with TikTok U.S.,https://www.cnbc.com/2025/01/18/perplexity-ai-makes-a-bid-to-merge-with-tiktok-us.html
11 Hours Ago,"Solana surges 12% on launch of Trump-themed meme coin, ether falls",https://www.cnbc.com/2025/01/18/crypto-market-today.html
12 Hours Ago,"What to expect from travel prices in 2025, and which spots have the best deals",https://www.cnbc.com/2025/01/18/what-to-expect-from-travel-prices-in-2025.html
13 Hours Ago,Consumer protection agencies at risk in Trump's second term: What it means for you,https://www.cnbc.com/2025/01/18/how-trumps-second-term-could-mean-the-downfall-of-the-fdic-cfpb.html
14 Hours Ago,Why the gold boom is causing a surge in illegal mining,https://www.cnbc.com/2025/01/18/why-the-gold-boom-is-causing-a-surge-in-illegal-mining.html
14 Hours Ago,Google Maps is turning 20 - mapping more countries and adding AI capabilities,https://www.cnbc.com/2025/01/18/google-maps-turns-20-adds-ai-features-new-countries-to-beat-apple.html
```