Web Scraper

Usage instructions:

Whether you are a data scientist, market researcher, or software developer, this AMI provides the tools you need to implement efficient and effective web scraping solutions, all while maintaining scalability and ease of use.

General Knowledge of Python & web scraping extraction is required.

Launch the product via 1-click. **Please wait until** the instance passes **all** status checks and is running. You can connect using your Amazon private key and '**ubuntu**' login via your SSH client.

To update software, use: **sudo apt-get update**

Option 1: Using Beautiful Soup Python: Data saved to S3 Bucket

To set up a Python-based web scraping system and save the scraped data to a CSV file saved directly in an S3 bucket, you can follow these steps:

Requirements: You will need to create and know the following info about your AWS account.

- Create a S3 bucket (ex: webscrapfolder)
- Have permissions set to Public
- Create a Bucket policy (Sample/Example below)

```
{
   "Version": "2012-10-17",
   "Statement": [
      {
         "Effect": "Allow",
         "Principal": {
            "AWS": "arn:aws:iam::0XXXXXXXXX:user/XXXX"
         },
         "Action": [
            "s3:GetObject",
            "s3:PutObject",
            "s3:ListBucket"
         ],
         "Resource": [
            "arn:aws:s3:::webscrapefolder",
            "arn:aws:s3:::webscrapefolder/*"
         ]
      }
   ]
}
```

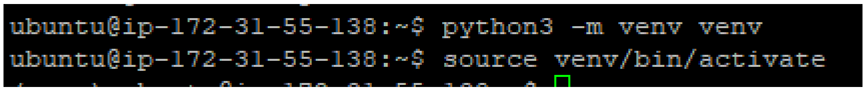
- IAM User (Create a user in Identity & Access Management Console)
- The ARN Name (Copy the ARN Name)
- Access Key (Create and Copy access key)

AWS Help: https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html

- **Now prepare Your Python Environment. Use the following command:**

  python3 -m venv venv

  source venv/bin/activate

  ```
  ubuntu@ip-172-31-55-138:~$ python3 -m venv venv
  ubuntu@ip-172-31-55-138:~$ source venv/bin/activate
  ```

- **Configure AWS CLI: To use the AWS CLI, you need to configure it with your credentials, region, and output format:**

  aws configure

  You will be prompted to enter your credentials in the “Access management” tab.

  AWS Access Key ID:
  AWS Secret Access Key
  Default region name:
  Default output format

  Your server should now be connected to your AWS S3 bucket
  https://docs.aws.amazon.com/cli/latest/

- **Still Inside the environment: (venv)**

  To view list of directories, use the :

  ls -la

- **We have created some sample script in python to get you started:**

  sudo nano scrape_quotes.py

  sudo nano scrape_cnbc.py

- **To run the Script use:**

  python scrape_quotes.py

  python your_script_name.py

  python generate_document.py

  For Beautiful Soup Documentation:
  See: https://www.crummy.com/software/BeautifulSoup/bs4/doc/#

Option 2: Using the Scrapy Framework: Data saved to MySQL Database

- Create the Scrappy environment

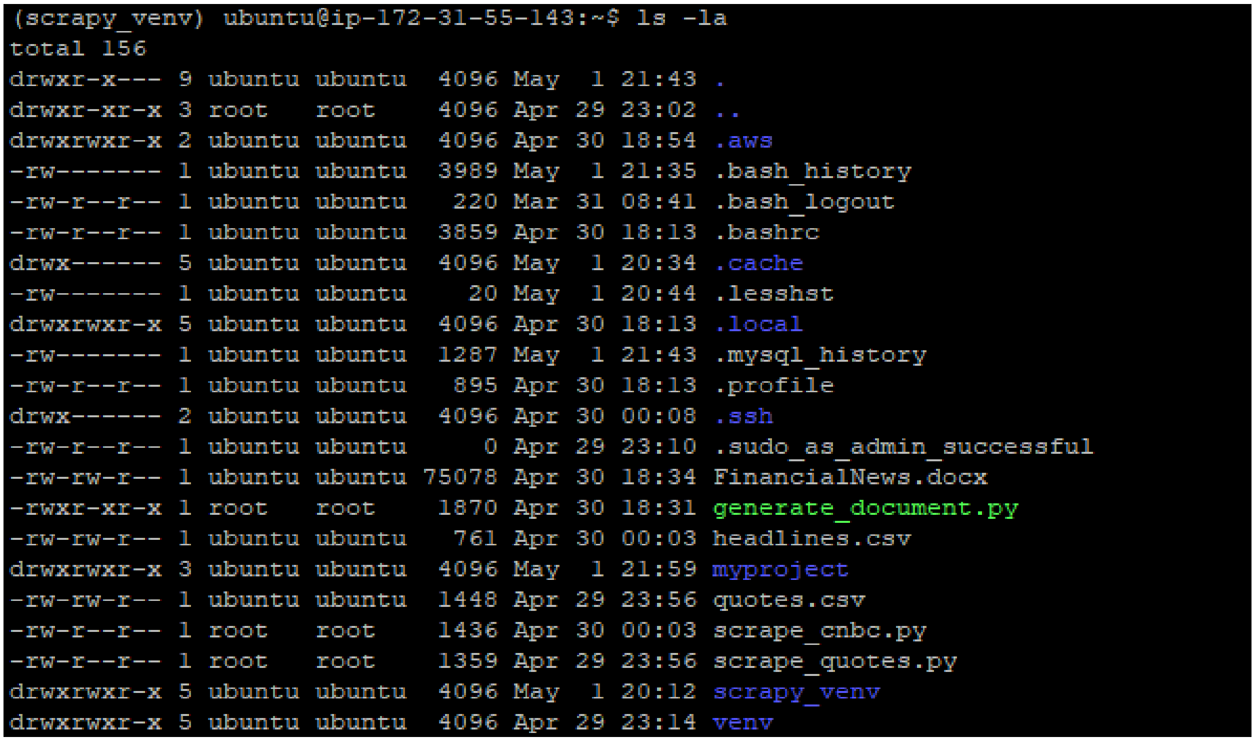
  ```
  python3 -m venv scrapy_venv
  ```

- Activate the virtual environment

  ```
  source scrapy_venv/bin/activate
  ```

- To view all the scrapy directories:

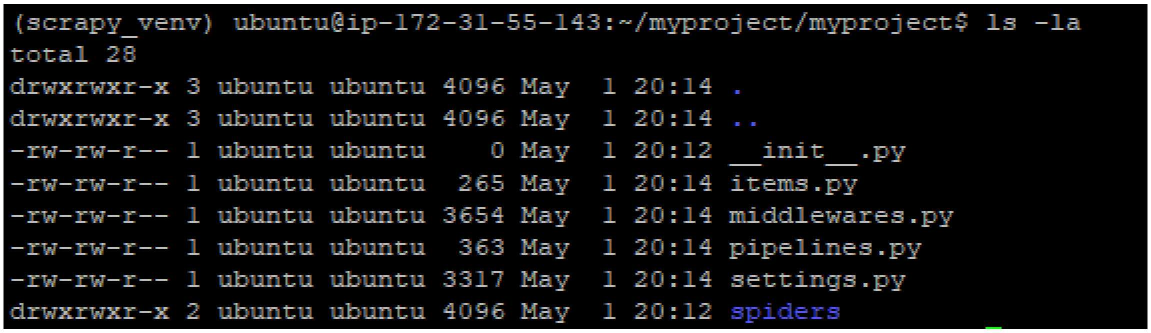
  ```
  ls -la
  ```



```
(scrapy_venv) ubuntu@ip-172-31-55-143:~$ ls -la
total 156
drwxr-x--- 9 ubuntu ubuntu  4096 May  1 21:43 .
drwxr-xr-x 3 root   root    4096 Apr 29 23:02 ..
drwxrwxr-x 2 ubuntu ubuntu  4096 Apr 30 18:54 .aws
-rw------- 1 ubuntu ubuntu  3989 May  1 21:35 .bash_history
-rw-r--r-- 1 ubuntu ubuntu   220 Mar 31 08:41 .bash_logout
-rw-r--r-- 1 ubuntu ubuntu  3859 Apr 30 18:13 .bashrc
drwx------ 5 ubuntu ubuntu  4096 May  1 20:34 .cache
-rw------- 1 ubuntu ubuntu    20 May  1 20:44 .lesshst
drwxrwxr-x 5 ubuntu ubuntu  4096 Apr 30 18:13 .local
-rw------- 1 ubuntu ubuntu  1287 May  1 21:43 .mysql_history
-rw-r--r-- 1 ubuntu ubuntu   895 Apr 30 18:13 .profile
drwx------ 2 ubuntu ubuntu  4096 Apr 30 00:08 .ssh
-rw-r--r-- 1 ubuntu ubuntu     0 Apr 29 23:10 .sudo_as_admin_successful
-rw-rw-r-- 1 ubuntu ubuntu 75078 Apr 30 18:34 FinancialNews.docx
-rwxr-xr-x 1 root   root    1870 Apr 30 18:31 generate_document.py
-rw-rw-r-- 1 ubuntu ubuntu   761 Apr 30 00:03 headlines.csv
drwxrwxr-x 3 ubuntu ubuntu  4096 May  1 21:59 myproject
-rw-rw-r-- 1 ubuntu ubuntu  1448 Apr 29 23:56 quotes.csv
-rw-r--r-- 1 root   root    1436 Apr 30 00:03 scrape_cnbc.py
-rw-r--r-- 1 root   root    1359 Apr 29 23:56 scrape_quotes.py
drwxrwxr-x 5 ubuntu ubuntu  4096 May  1 20:12 scrapy_venv
drwxrwxr-x 5 ubuntu ubuntu  4096 Apr 29 23:14 venv
```

- Change directories into the “myproject/myproject” folder:

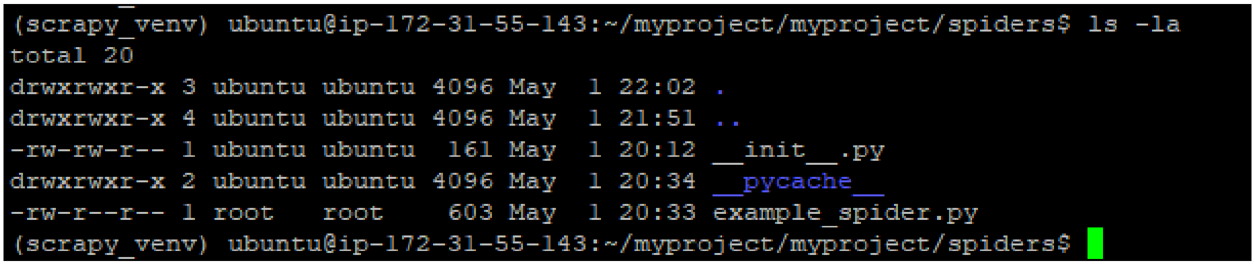
  ```
  cd myproject/myproject
  ls -la
  ```



```
(scrapy_venv) ubuntu@ip-172-31-55-143:~/myproject/myproject$ ls -la
total 28
drwxrwxr-x 3 ubuntu ubuntu 4096 May  1 20:14 .
drwxrwxr-x 3 ubuntu ubuntu 4096 May  1 20:14 ..
-rw-rw-r-- 1 ubuntu ubuntu    0 May  1 20:12 __init__.py
-rw-rw-r-- 1 ubuntu ubuntu  265 May  1 20:14 items.py
-rw-rw-r-- 1 ubuntu ubuntu 3654 May  1 20:14 middlewares.py
-rw-rw-r-- 1 ubuntu ubuntu  363 May  1 20:14 pipelines.py
-rw-rw-r-- 1 ubuntu ubuntu 3317 May  1 20:14 settings.py
drwxrwxr-x 2 ubuntu ubuntu 4096 May  1 20:12 spiders
```

- **A sample script has been provided to get you started in the “spiders” folder.**

  cd spiders

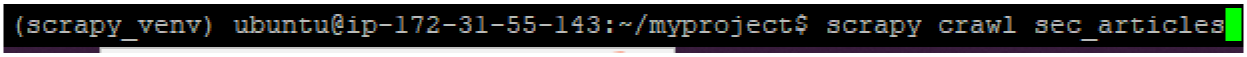


```
(scrapy_venv) ubuntu@ip-172-31-55-143:~/myproject/myproject/spiders$ ls -la
total 20
drwxrwxr-x 3 ubuntu ubuntu 4096 May  1 22:02 .
drwxrwxr-x 4 ubuntu ubuntu 4096 May  1 21:51 ..
-rw-rw-r-- 1 ubuntu ubuntu  161 May  1 20:12 __init__.py
drwxrwxr-x 2 ubuntu ubuntu 4096 May  1 20:34 __pycache__
-rw-r--r-- 1 root   root    603 May  1 20:33 example_spider.py
(scrapy_venv) ubuntu@ip-172-31-55-143:~/myproject/myproject/spiders$
```

  sudo nano sec_articles_spider.py

- **To run the script, go back to the “myproject” directory and run:**

  scrapy crawl sec_articles



```
(scrapy_venv) ubuntu@ip-172-31-55-143:~/myproject$ scrapy crawl sec_articles
```

- **You will find the results in MySQL database. MySQL database Credentials:**

  mysql -u scrappy-user -p

  Pass: **CCscrappy!!!**

  USE scrappyDB;

  SELECT * FROM articles;

- **To exit environment, use the following command:**

  deactivate

<u>Additional Help:</u>

https://scrapfly.io/blog/web-scraping-with-scrapy/#start-scrapy-project

https://docs.scrapy.org/en/latest/