

Automation_DiscountMate Deployment & Scheduling Guide

1. Project Overview

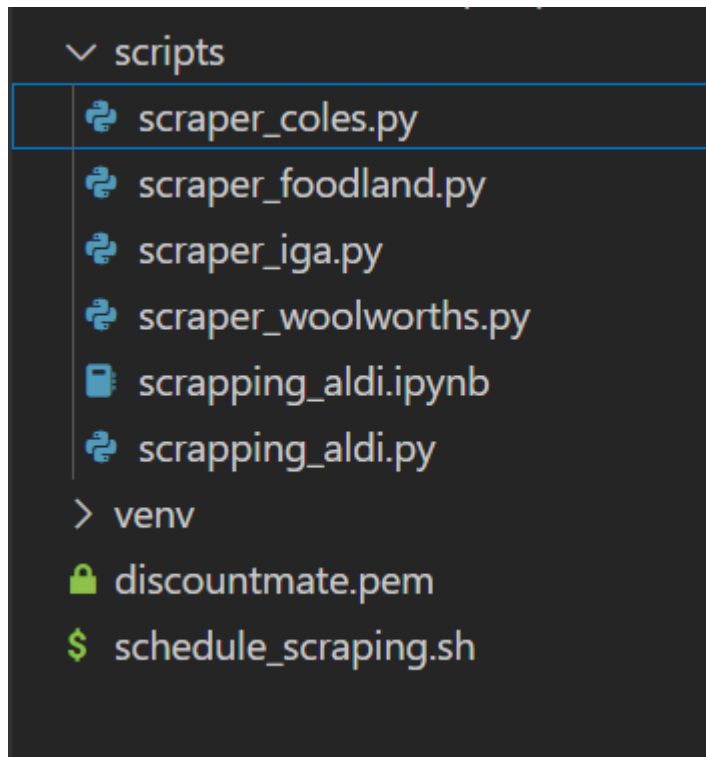
This project automates the scraping of multiple grocery sites (Coles, Foodland, IGA, Woolworths, Aldi) on a schedule. The main driver is the ``schedule_scraping.sh`` script which:

- Activates the Python virtual environment
- Runs all scraper scripts
- Converts and runs the Aldi notebook
- Deactivates the environment

2. Prerequisites

- Local Windows machine with PowerShell or Git Bash
- ``discountmate.pem`` SSH key file
- AWS EC2 instance (Ubuntu or similar)
- Basic shell and SSH familiarity

3. Folder Structure (after transfer to EC2)



4. Step 1: Secure Your SSH Key On Windows, run:

```
chmod 400 discountmate.pem
```

5. Step 2: Copy the Project to EC2, Use SCP from PowerShell or Git Bash:

```
scp -i discountmate.pem -r E:/Automation_DiscountMate \ ubuntu@<EC2-PUBLIC-  
DNS>:/home/ubuntu/
```

6. Step 3: SSH into Your Instance

```
ssh -i discountmate.pem ubuntu@<EC2-PUBLIC-DNS>
```

7. Step 4: Install Python & venv

```
sudo apt update  
sudo apt install python3 python3-venv python3-pip -y
```

8. Step 5: Create & Activate Virtual Environment

```
cd ~/Automation_DiscountMate  
python3 -m venv venv  
source venv/bin/activate  
pip install -r requirements.txt # if available  
pip install jupyter nbconvert # for notebook conversion
```

9. Step 6: Adjust & Test schedule_scraping.sh

- Open and edit the script to use ``/usr/bin/env bash``
- Change ``cd`` to ``/home/ubuntu/Automation_DiscountMate``
- Make executable: ``chmod +x schedule_scraping.sh``
- Test: ``./schedule_scraping.sh``

10. Step 7: Schedule with cron -> "crontab -e"

Add line to run daily at 3 AM:

```
0 3 * * * /home/ubuntu/Automation_DiscountMate/schedule_scraping.sh >>
```

```
/home/ubuntu/scrape.log 2>&
```

```
Asus@Harsh MINGW64 /e/Automation_DiscountMate
$ ssh -i discountmate.pem ubuntu@ec2-3-27-70-143.ap-southeast-2.compute.amazonaws.com
The authenticity of host 'ec2-3-27-70-143.ap-southeast-2.compute.amazonaws.com (3.27.70.143)' can't be established.
ED25519 key fingerprint is SHA256:/z3mcXWN+4THHEr1Jpa50SHk+pBqlojFCYs6g3aQT8s.
This host key is known by the following other names/addresses:
  ~/.ssh/known_hosts:6: ec2-54-252-155-111.ap-southeast-2.compute.amazonaws.com
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-3-27-70-143.ap-southeast-2.compute.amazonaws.com' (ED25519) to the list of known hosts.
Welcome to Ubuntu 24.04.2 LTS (GNU/Linux 6.8.0-1028-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/pro

System information as of Thu May 15 07:35:45 UTC 2025

System load:  0.22          Processes:      108
Usage of /:   45.7% of 6.71GB Users logged in:  0
Memory usage: 21%          IPv4 address for enx0: 172.31.2.233
Swap usage:   0%

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.
   https://ubuntu.com/aws/pro

Expanded Security Maintenance for Applications is not enabled.

39 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

Last login: Thu May  8 08:09:02 2025 from 203.123.64.129
ubuntu@ip-172-31-2-233:~$ ls
Automation_DiscountMate  myscript.sh
ubuntu@ip-172-31-2-233:~$ cd Automation_DiscountMate/
ubuntu@ip-172-31-2-233:~/Automation_DiscountMate$ ls
discountmate.pem  schedule_scraping.sh  scripts  venv
ubuntu@ip-172-31-2-233:~/Automation_DiscountMate$ cd scripts/
ubuntu@ip-172-31-2-233:~/Automation_DiscountMate/scripts$ ls
scraper_coles.py  scraper_foodland.py  scraper_iga.py  scraper_woolworths.py  scrapping_aldi.ipynb  scrapping_aldi.py
ubuntu@ip-172-31-2-233:~/Automation_DiscountMate/scripts$
```

```
ubuntu@ip-172-31-2-233: ~/Automation_DiscountMate/scripts
GNU nano 7.2 /tmp/crontab.J54mk2/crontab
# Edit this file to introduce tasks to be run by cron.
#
# Each task to run has to be defined through a single line
# indicating with different fields when the task will be run
# and what command to run for the task
#
# To define the time you can provide concrete values for
# minute (m), hour (h), day of month (dom), month (mon),
# and day of week (dow) or use '*' in these fields (for 'any').
#
# Notice that tasks will be started based on the cron's system
# daemon's notion of time and timezones.
#
# Output of the crontab jobs (including errors) is sent through
# email to the user the crontab file belongs to (unless redirected).
#
# For example, you can run a backup of all your user accounts
# at 5 a.m every week with:
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/
#
# For more information see the manual pages of crontab(5) and cron(8)
#
# m h dom mon dow  command
0 18 * * * /home/ubuntu/Automation_DiscountMate/schedual_scraping.sh >> /home/ubuntu/Automation_DiscountMate/scrap.log 2>&1
```

11. Troubleshooting Tips

- Permissions: re-check ``chmod`` settings
- Missing modules: activate venv and ``pip install`` errors
- Cron: check ``scrape.log`` for output and errors
- Time-zone: confirm server time with ``date``

12. Appendix: Common Commands

Task	Command
Change to project folder	<code>cd ~/Automation_DiscountMate</code>
Activate venv	<code>source venv/bin/activate</code>
Deactivate venv	<code>deactivate</code>
List crontab entries	<code>crontab -l</code>
View latest scrape log	<code>tail -n 50 ~/scrape.log</code>