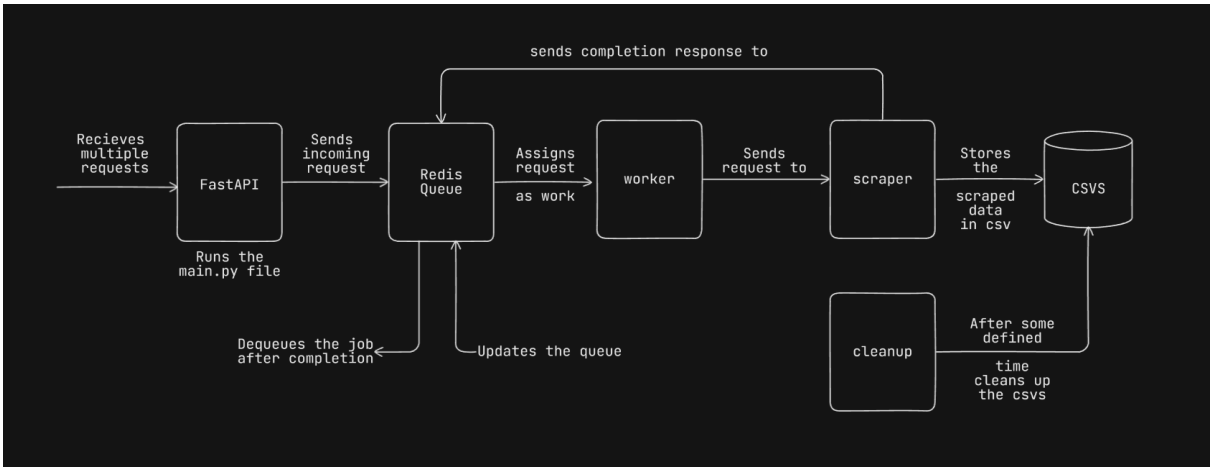
Backend of YT Scraper Project

Pre Requisites:

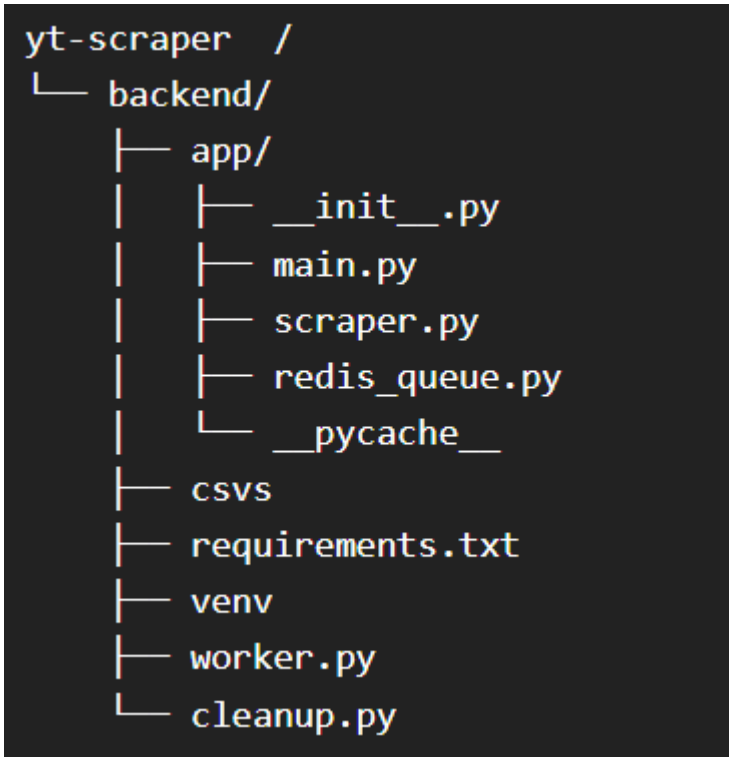
- Must have created an EC2 Instance with Ubuntu AMI
- Assign and allocate Elastic IP (Optional)
- Clone this repository

```
https://github.com/parthjs27/yt-scraper.git
```

Workflow



Folder structure:



Install Dependencies

```
sudo apt update && sudo apt upgrade -y
sudo apt install -y python3 python3-pip python3-venv redis git unzip curl
```

Install Google Chrome and Chromedriver

```
sudo apt update
sudo apt install -y wget unzip
wget https://dl.google.com/linux/direct/google-chrome-stable_current_amd64
sudo apt install -y ./google-chrome-stable_current_amd64.deb
```

Move to backend directory

```
pip install -r requirements.txt
```

Create Virtual Environment

```
python3 -m venv venv
source venv/bin/activate
pip install -r requirements.txt
```

Start Redis Setup

```
sudo systemctl enable redis
sudo systemctl start redis
```

Test if Redis is running

```
redis-cli ping
# Should return: PONG
```

Note: These below 2 commands must be executed on different terminals

Run FastAPI App

```
uvicorn app.main:app --host 0.0.0.0 --port 8000 --reload
```

Run the worker.py file

```
python3 worker.py
```

For Swagger UI

- Go to your browser
- Click on New Tab and paste this URL

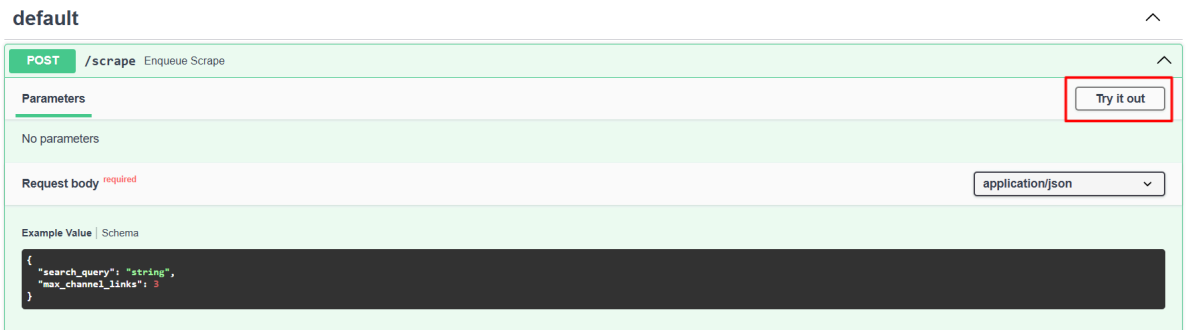
```
http://<EC2 IPv4>:8000/docs
```

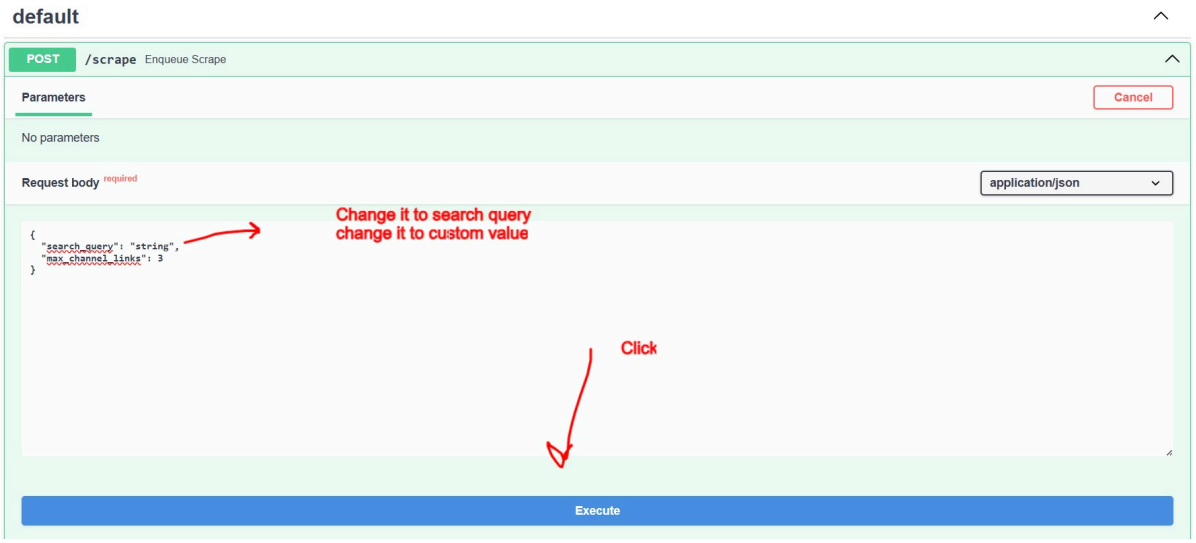
In the /scrape route:

- Click on try out

- replace “string” with your search query
- You can change the max_channel_links to your custom value
- Click Execute

You can see that the scraping process has started in the terminal





```
ubuntu@ip-[redacted]:~$ cd yt-scraper/backend/
ubuntu@ip-[redacted]:~/yt-scraper/backend$ source venv/bin/activate
(venv) ubuntu@ip-[redacted]:~/yt-scraper/backend$ uvicorn app.main:app --host 0.0.0.0 --port 8000 --reload &
[1] 1328
(venv) ubuntu@ip-[redacted]:~/yt-scraper/backend$ INFO:     Will watch for changes in these directories: ['/home/ubuntu/yt-scraper/backend']
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
INFO:     Started reloader process [1328] using StatReload
INFO:     Started server process [1330]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     103.182.159.202:53530 - "GET /docs HTTP/1.1" 200 OK
INFO:     103.182.159.202:53530 - "GET /openapi.json HTTP/1.1" 200 OK
INFO:     103.182.159.202:53531 - "GET /docs HTTP/1.1" 200 OK
INFO:     103.182.159.202:53531 - "GET /openapi.json HTTP/1.1" 200 OK

(venv) ubuntu@ip-[redacted]:~/yt-scraper/backend$ python3 worker.py
INFO:     103.182.159.202:53565 - "POST /scrape HTTP/1.1" 200 OK
2025-06-06 16:54:32,138 - INFO - Processing job_id: 62df13d0-9a40-4507-af1d-82384418a4b8  task: javascript (Max Links: 3)
2025-06-06 16:54:32,138 - INFO - Running scraper for query: javascript
2025-06-06 16:54:32,138 - INFO - Initializing WebDriver for Linux environment...
2025-06-06 16:54:32,138 - INFO - ====== WebDriver manager ======
2025-06-06 16:54:33,195 - INFO - Get LATEST chromedriver version for google-chrome
2025-06-06 16:54:33,210 - INFO - Get LATEST chromedriver version for google-chrome
2025-06-06 16:54:33,225 - INFO - Driver [/home/ubuntu/.wdm/drivers/chromedriver/linux64/137.0.7151.68/chromedriver-linux64/chromedriver] found in cache
2025-06-06 16:54:35,182 - INFO - Handling consent dialogue (if present)...
2025-06-06 16:54:45,342 - WARNING - Consent dialog not found or failed: Message:
Stacktrace:
#0 0x60ba74d32c4a <unknown>
#1 0x60ba747d86e0 <unknown>
#2 0x60ba7482a117 <unknown>
#3 0x60ba7482a311 <unknown>
```

After task completion you will see this message

```
2025-06-06 16:56:15,809 - INFO - Channel details collection complete.
2025-06-06 16:56:15,870 - INFO - Scraping completed for job_id 62df13d0-9a40-4507-af1d-82384418a4b8. Files saved.
2025-06-06 16:56:15,870 - INFO - Closing WebDriver...
2025-06-06 16:56:16,083 - INFO - Task completed successfully for query: job_id: 62df13d0-9a40-4507-af1d-82384418a4b8 javascript
```

You can download the CSV files from the /download/{job_id}/{file_type} route:

- You can get the job Id from the above image
- There are 2 csv generated
  - {job_id}_channel_urls.csv
  - {job_id}_channel_details.csv
- Specify the csv you want to download (channel_details.csv is Recommended)