Data Engineer | Programming Task

Overview:

This task requires you to collect structured data from a website. You will be writing a web crawler that will visit a website. Fetch html pages, parse them in the desired format and save them in an output file.
You should use python language(Version 3.xx) for this programming task. Feel free to use any framework/library(requests, selenium, scrapy etc.).

Submission:
You are required to submit this task using GitHub. Please create a profile on GitHub; if you don't already have one. Upload all the code and results to a GitHub repository and share the link to the repository in the email.

You will have **one week** to submit this assignment.

Scoring:
Primary scoring will be done on the correctness of your code and output. However, we will be considering several other factors for evaluation e.g.
- Code readability
- Performance
- Documentation
- Error Handling
- Test cases

You are encouraged to improvise on the above list, as this is only for the hint.

Tasks:

The website you need to crawl is the company listing on below webpage. The website link is: https://www.adapt.io/directory/industry/telecommunications/A-1
This site contains company information having multiple pages from A-Z.

1) You need to collect the basic raw data available in the link. Bonus points will be given if you collect more. The desired output format should be

[
{'company_name':'A & L Personnel Services'**,** 'source_url': '/company/a---l-personnel-services'},
{'company_name**':**'A+ Conferencing', 'source_url': '/company/a--conferencing},
…
]

Save this output to file named **company_index.json**.

2) In the second part you need to fetch a detailed company profile based on the links you crawled in part 1. The detailed company profile should include: (Any additional raw datapoints can also be crawled)

Company_name	Company_location
Company_website	Company_webdomain
Company_industry	Company_employee_size
Company_revenue	Contact_name
Contact_jobtitle	Contact_email_domain

The output format should be a list of dictionaries as follows.

[
{“company name”: “A & L Personnel Services”, “company_location”: “Gregory, Michigan” "company_website": http://www.cac.net”, “company_webdomain”: “cac.net”, “company_industry”: “Telecommunications”, “company_employee_size”:None, “company_revenue” :None, “contact_details”: [{“contact_name”: “Doug Waite”, “contact_jobtitle”: “owner”, “contact_email_domain”: “cac.net”, “contact_department”: “Finance and Administration”},{“contact_name”: “Jim Mason”, “contact_jobtitle”: “Club Directior”, “contact_email_domain”: “cac.net”, “contact_department”: “Other”}]},
…
]

Save the output in **company_profiles.json**.

3) Load the company_index and company_profile in the database of your choice MySQL, PostgreSQL, MongoDb, SQLite etc. Preferably two tables/collections.

4) Build Test cases and Data point validators

5) Please provide answers to the two question below

- Briefly describe the architecture of your application
- Which database engine you choose and why?

All the Best !