UpsellX Collector

Background:

Consider a digital advertisement agency UpsellX, which works with different small or medium scale business (SMBs), provides them with the best advertisement solution across different platforms (Google Ads, Facebook Ads etc). It needs to know as much information possible about the clients to understand their business well and recommend them the best campaign plans. But they don't like to take all the information directly from the client as its time consuming and boring, sometimes it's not possible. So they plan to scrap public information available on the web to provide kickass experience to clients during the onboarding process by giving the prefilled profile. The client will only provide his company website, and your job will be to collect as much information possible from this seed.

Problem:

You will be required to design a **collector platform**, which will have a rest **API** that takes the website URL and collects public data from website and multiple sources; returns combined data back in JSON format. Data will certainly be **stored** for future use, as we don't like to scrap the same information twice.

Deliverables:

You need to submit a GitHub link containing:

1. Necessary design documents which should describe the challenges of this system and how you have addressed it.
2. System diagram that you would like to suggest.
3. An implementation of the service, we prefer it to be implemented using serverless architecture in AWS.
4. Describe different sources from where we can get data for business.
5. Define Schema to store such data and test data if needed.
6. README file to run and test your implementation.

<u>Bonus Point:</u>

1. Dockerize the application so that we can run and deploy it in a local environment as well.