Initial Requirements

Problems

We are manually classifying our data and it’s way too slow and causing big blockers in the company. Therefore we are building a project as part of our data team to crawl and classify domains correctly to give the desired output which will be saved into our database,.

This can be built in python or PHP and will be deployed in an AWS serverless environment where it will process jobs generated from PHP.

**Crawl a website domain then classify:**

1. Introduction
2. Industry
   1. Linkedin Industry(Business Domain)
   2. Sub Industry
   3. Keywords
3. Find domain social channels
   1. Search for and save social channels on page.
4. Find domain HQ
   1. Parent / subsidiary legal company
   2. Address
5. Phone number
6. Emails
7. Email domain name
8. Product or Service?
9. Is Ecommerce?

**Tasks**

Research crawling a domain:

1. What data points are available by crawling a domain? Research these and list below as part of the first step in this project.

-

**Outcome**

To reduce our manual categorization to 95% - currently as we crawl domains for our company intelligence we have to classify these manually for our platform - this is a huge burden especially with new data. Hence why we are starting this project as it will have direct impact on our productivity and free up team members