Design a Request based Deep web-crawling application

Create web crawler that takes URL and depth as parameter and create a tree of child pages linked to the URL. application is expected to provide deep-crawling solution, that it goes through multiple levels in link hierarchy. Create a simple API endpoint that would take URL as a parameter and return JSON representing the below format.

total\_links – total Links crawled.

total\_images – total images crawled.

details – information about each link crawled.

Response should be in json format:

{

                "total\_links": "",

                "total\_images": "",

                "details": [{

                                                "page\_title": "",

                                                "page\_link": "",

                                                "image\_count": ""

                                },

                                {

                                                "page\_title": "",

                                                "page\_link": "",

                                                "image\_count": ""

                                }

                ]

}

Requirement:

- Develop Rest Endpoint to Submit the Crawl Request and return the acknowledgement token. It should take two input Parameters - Base url and depth.

- Application should pick all the unprocessed request and initiate the deep crawling as per the request.

- Post processing request should be marked as processed and results should be persisted.

- Develop endpoint to fetch the results.

- Develop endpoint to fetch the status of the request ( Submitted, In-Process, Processed, Failed )

- Use concurrency to optimize the processing.