Problem Statement

Write a scalable microservice in Scala, which exposes an API endpoint that takes a list of URLs, crawls the data concurrently, and returns the crawled data.

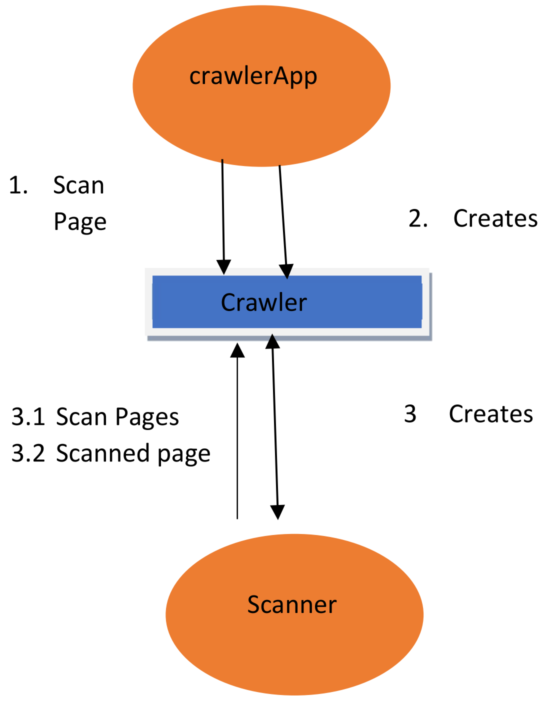
Solution:

I have attempted to create a web crawler as shared in the github.

Below is the outline of the solution drafted:

- **CrawlerApp** raise http request for the websites: google.com and github.com
  - It will check if there are any exceptions while raising the request and throw error if necessary
  - It will initiate Crawler object
- After **Crawler** object gets initated
  - Crawler is now ready to scan the pages, so it will try to establish connection with website.
  - It will wait for sometime for the website to respond for the request. If response time is more then the wait time, then the process terminates.
  - If it responds within the specified time, then it scanner object.
- After **scanner** object gets initiated
  - It will start parsing the URI, the parsed URI, resolved against the given base Uri, if it shows to a web page, which will probably contain more URIs.
  - Scans the page at the URI, which has the given link depth. When the page will be successfully scanned, a Page Scanned command will be sent to the Crawler. If in the page it encounters href-s to URIs, corresponding ScanPage commands will be sent to the Crawler in order to scan them, too.

More detailed explanation, I have added in the code, where ever it is required.


crawlerApp
1. Scan Page
2. Creates
Crawler
3.1 Scan Pages
3.2 Scanned page
3 Creates
Scanner