URL Explorer System - Requirements

Objective:

Develop a scalable backend system that collects unique URLs and their corresponding HTML content.

Requirements:

URL Fetching and Parsing:

- Develop a system that can fetch the raw HTML content of a given URL.
- Parse the fetched HTML content to extract all the URL links present on the page.

Recursive Link Extraction:

- Implement a mechanism to traverse the extracted URL links and find their corresponding sub-URL links.
- Continuously explore each discovered link to identify additional sub-URL links.

Data Storage:

- Design a database schema to store the unique URLs and their corresponding raw HTML content.
- Store the unique URLs and their raw HTML content in the database.

API Endpoints:

- Create API endpoints to accept incoming URLs and trigger the fetching, parsing, and storage process.
- Implement endpoints to retrieve stored URLs and their associated raw HTML content.

Scalability and Performance:

- Optimize the system to handle a large volume of URLs efficiently.
- *Bonus*: Implement caching mechanisms to minimize redundant fetch requests for already visited URLs and improve performance.

Error Handling and Logging:

- Implement error-handling mechanisms to handle cases of invalid URLs, network errors, and other exceptions during the fetching and parsing process.
- Include logging functionality to capture relevant events and errors for debugging and monitoring purposes.

Testing and Documentation:

- Write comprehensive unit tests to validate the fetching, parsing, and storage functionalities.
- Document the system's architecture, algorithms, database schema, and API endpoints in a README.md file.

Notes:

- The requirements focus on fetching, parsing, and storing unique URLs along with their raw HTML content.
- The emphasis is on capturing and storing the unique URLs and their corresponding raw HTML content rather than explicitly tracking the relationship between URLs and sub-URLs.
- Make reasonable assumptions and consider industry best practices when developing this System.
- The focus should be on demonstrating your backend engineering skills, data handling, and problem-solving abilities.

Delivery

Time Estimation:

- Approximately 4 - 6 hours, please capture your time

- We value quality over speed so please take the time you need to deliver a well-structured and thought-out solution.

Instructions

1. Upload the completed project on your GitHub account and send us the link.
2. Document your assumptions potential areas for further improvement, and instructions on how to start the project.
3. Please use online diagramming tools like draw.io or excalidraw.com for the system design.