GitHub Scraping Exercise for Backend Engineers

Design and implement a scalable system to scrape GitHub data, store it in a database and provide a RESTful interface and CLI tool for querying repository information.

System Components

RESTful Service

- Design and implement endpoints to support the use cases defined below.
- It should follow current best practices in API design

Database

- Choose a suitable database for the problem.
- Design the database schema to store GitHub user and project data.

CLI Tool

- Implement commands for user to interact with RESTful service.
- Use any modern existing libraries like `typer`, `click` in case of Python.

Desired Use Cases

1. Retrieve User Projects

- Accept a GitHub username.
- If the user exists in the database, return the stored projects.
- If the user doesn't exist, scrape the projects from GitHub, store them in the database and return the projects.
- For each project, include the project metadata (name, description, stars, and forks).

2. Get N Most Recent Users

- Return N most recent users saved in the database. Cache for quicker retrieval

3. Retrieve N Most Starred Projects

- Return N most starred projects saved in the database.

Technical Requirements

- Use Python (or your preferred language) with `poetry` for dependency management if using Python.
- Provide technical documentation for your solution.
- Ensure your code is well-structured, readable and maintainable.
- Ensure your code is tested

Evaluation Criteria

- System design and architecture
- Code quality and maintainability
- Technical documentation and testing

Submission Guidelines

- Create a new Git repository for your solution.
- Implement the required features and ensure your code is well-tested.
- Provide technical documentation for your solution.
- Make your repository publicly accessible.
- Be prepared to discuss your design decisions and implementation during the interview.
- If using any code generator tools, mention it in the documentation.
- Feel free to use existing libraries and frameworks to simplify your implementation.