LittleBigCode..

AI Solution Creator

Take Home Challenge

Role: Data Engineer
Level: Confirmed

Welcome

Foreword

Before we begin, we would like to **thank you**. We know that these challenges take time and we thank you for your time. We take care to read all the technical responses sent to us and we make sure to give you feedback, regardless of the outcome of your recruitment process.

Purpose

The goal of this challenge is to help us understand your **data engineering abilities**. Challenges, like this take home project, are intended to demonstrate the extent of a candidate's skills on their potential future role. Through this exercise, we want to get an overview of your technological knowledge, your ability to propose solutions or approaches based on the constraints inherent in the business context and the approach you will adopt when faced with a problem.

We do not expect you to spend **more than a 4 hours** completing the exercise. There's no hard time limit so work on it at your convenience. Take your time and submit a solution you feel proud of and ready to discuss with us.

Also, questions are definitely welcome, so ask away!

LittleBigCode..
AI Solution Creator

Description

You are part of a team responsible for the proper provision of a wide range of data for several departments of your company. For the next sprints, you and your team will be in charge of a solution that crawls for articles from a news website, cleanses the response, stores in a SQL or NoSQL database then makes it available to list via an API.

1. Write an application to crawl an online news website, e.g. www.theguardian.com/ or http://www.bbc.com/ using a crawler framework such as Scrapy or other. The application should cleanse the articles to obtain only information relevant to the news story, e.g. article text, author, headline, article url, and remove superfluous content such as advertising and html tags

2. Store the data in the database, for subsequent search and retrieval

3. Write an API that provides access to the content in the database. We should be able to list all articles or get one by its URL

4. Find the top 5 movies by genre

5. In these slides, explain:
   - what are the limitation of your actual solution
   - how you can improve it, using an other technology or an other pattern
   - how you can setup this kind of solution at scale in production with a detailed architecture

How?

Rules

- you can perform the test in the programming language(s) of your choice
- you can use all the technological solutions that seem useful to you
- feel free to use this presentation if you want to add schemas, drawing…
- you will be challenged on the quality of your code and the organization of your solution, but also on the documentation, the architecture and overall solution design, the relevance of the technical choices, the appropriate use of source control...
- you are expected to be able to explain your solution in a face-to-face interview. During the interview, you'll have also to be able to justify the technical choices, your implementation and the advantages and limitations of your solution.

Deliverables

- your application should be able to be run as a stand alone console application, or as a docker container.
- all source code should be able to be built in a Linux environment.
- include database scripts with DDL you deem necessary and any other prequisites
- include any deployment instructions in a readme.md file. These can include manual steps that need to be run first.
- also in a readme.md file, explain:
    - how you can setup this kind of pipeline in production
    - what are the limitation of this solution
    - how you can improve it, using an other technology or an other pattern
    - what are the advantages and disadvantages of using a distributed data processing framework in this situation

What next?

- invite our github user "littlebigcode-codereview" to be a collaborator on the repository
- let us (your interview coordinator) know when you are ready to review
- one of our coaches will contact you to discuss with you about your solution, your choice of solutions, the overall architecture... stay tuned!

Solution limitations

- **Development**
  - Implement unit testing
  - Implement Logging and Monitoring
  - Error Handling
  - Use scheduler to trigger scraping periodically.

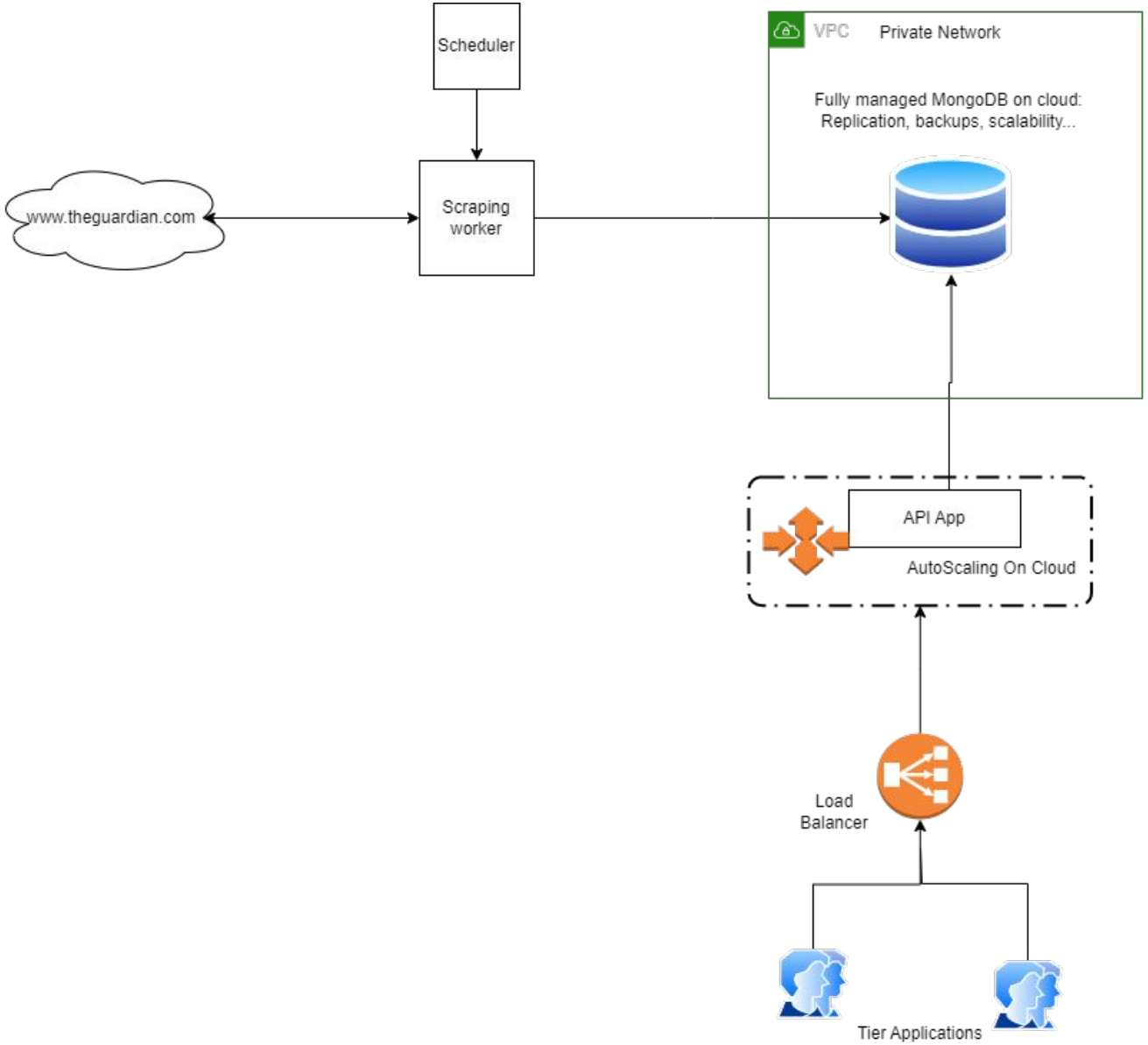
- **Security**
  - Add API Authentication
  - Add API rate limiting
  - Implement Access control and limit access to the database (use a private network, Firewall, security rules, access lists, …)

- **Deployment**
  - Scalability: adjust ressources based on rules and metrics (Auto scaling on cloud)
  - Caching: Use Redis or a caching layer with the database to give most asked data in API

Solution improvements

- Improve the above limitations : Error Handling, logging, caching, security,...
- Build full cloud architecture : low cost and fully managed
- Add more API functionalities to access articles by Headline, Author,...
- Add a semantic search layer in the API : Use NLP models to generate semantic understanding of articles and enable retrieval of most similar articles to a query.

Solution architecture



LittleBigCode..
AI Solution Creator

Thanks

This document includes icons by Flaticon and infographics & images by Freepik