Amazon Lightsail
Containers Hackathon

3141

Aggre-Gator

PROBLEM STATEMENT

Some questions you may want to answer: What is the problem statement? What kind of opportunities does this problem promise?

- ★ Generally, people browse multiple sites to get a complete information/news about a topic. A lot of time is wasted in searching, opening the links, etc. This process is simply boring and we believe that there is a great potential to save time.
- ★ We plan to build a platform that collects news from different sources and deliver them right where the users needs it, i.e., at a single page. On top of that, we plan to develop important features such as grouping relevant news, news summarization.
- ★ This problem badly affected the world during the pandemic, as there were constant changes in policies and health updates. So searching a new policy meant googling three-four results and reading through them.
- ★ This was during the pandemic, when we were actually at homes, in front of our device, a lot of free time. Even if we wanted to keep ourselves updated we couldn’t. Talk about going to normal world and googling three-four results.
- ★ Moreover this platform will serve a great opportunity to individual bloggers who are often left unnoticed by becoming part of this platform they have a chance of reaching to more people meanwhile just maintaining their personal blogs.

DESCRIPTION OF SOLUTION

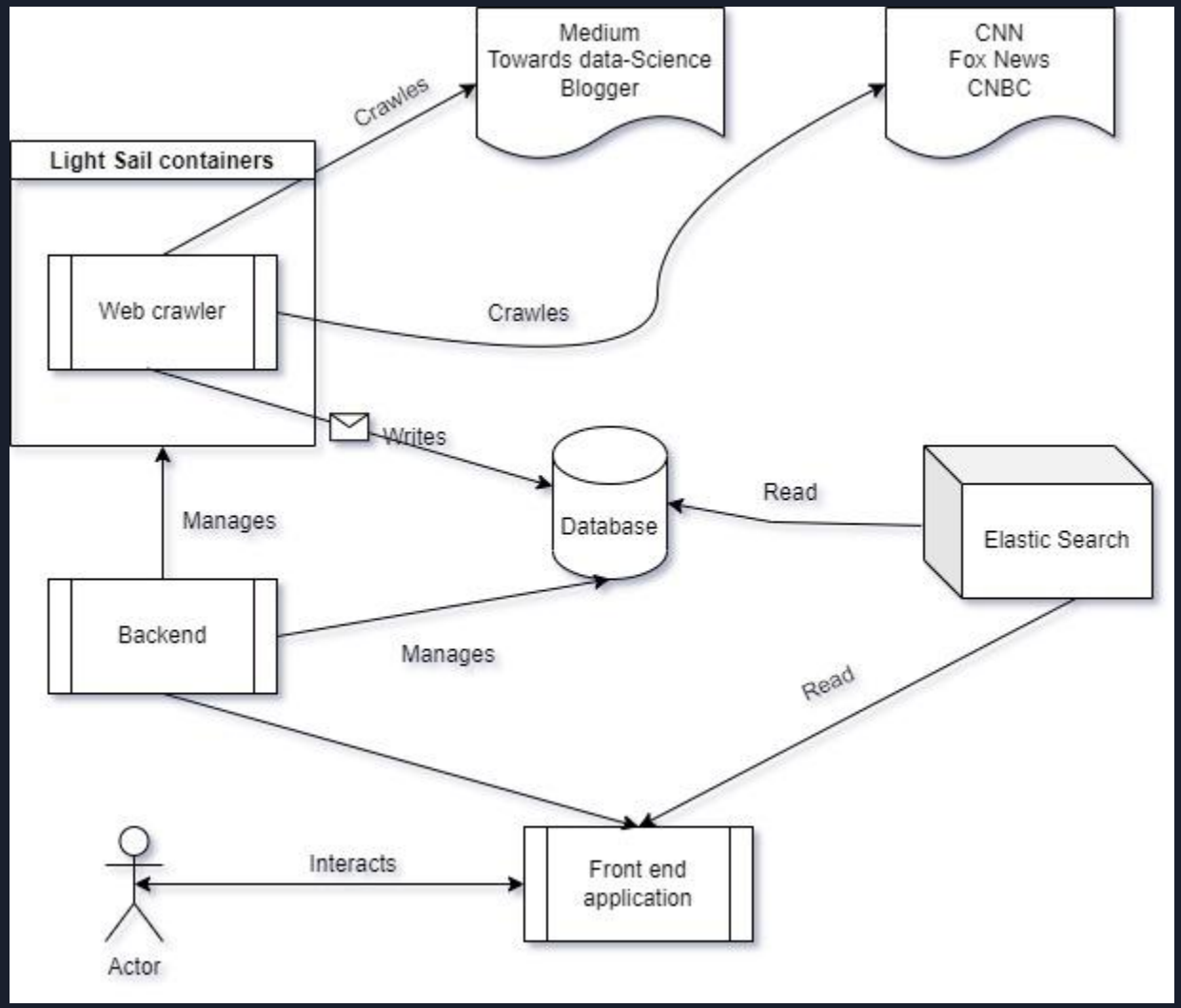
Some questions you may want to answer: How will your project solve the issue? What makes it achievable? If there are other similar solutions out there – what makes yours unique/different?...

- ★ Our project tries to solve the issue of people being in their own ideological/interest bubbles by providing them with a platform which will aggregate news from different sources. This way, they'll be able to access a much larger amount of information from different sources in one convenient place.
- ★ We can make our project easier to achieve by focusing on only two news sites i.e. CNN and Fox News in the beginning, we can thus focus on adding functionality to the site, such as powerful searching, classification, and presentation. By keeping it constrained to only a couple sites at the beginning, we can quickly prototype the solution, while being able to expand the product to cover more websites.
- ★ To differentiate our solution from others which already exist, we will focus on displaying only the latest news and let the user decide which articles to read and where they fall on various spectra rather than showing them our personal curated collection. This way they'll be able to see content beyond what's personalised for them by other content aggregators. Recommendations will only be based on topics of interest.

Some questions you may want to answer: How will you build your solution? What data will you be using? How will you obtain that date? What will be you solution's architecture? What tools & technology will you use? Please make sure to use the Amazon Lightsail Containers to build your solution.

- ★ We have planned to develop a web crawler using Scrapy in python that will search the list of website provided by the user for latest content at a fixed interval.
- ★ The data we need will be gathered from the websites and will be in the form of url to relevant articles.
- ★ As the web crawler will only access publicly available sites the data will be captured from news domains like CNN, Fox News, CNBC, etc.. also various open blogs can also be added.
- ★ The data will be stored in dynamoDB and then cloud search will cache upon it for quick search of relevant results.
- ★ Architecture is presented on next slide.
- ★ Tech stack we will require for our application is :
  - ○ Lightsail for hosting our crawlers, frontend and backend.
  - ○ DynamoDB for storing our crawled data.
  - ○ Cloud Search for quick searches among the post.
  - ○ Front end application will be in react.
  - ○ Backend will use python django.
  - ○ Crawlers will be created using python Scrapy.
  - ○ SNS will be used for notifications.

ARCHITECTURE



TECHNICAL ROADMAP

Some questions you may want to answer: How will you realize your solution? What is your implementation plan and what is the technical roadmap of your solution/MVP?

Implementation Plan:

- We will build a crawler to perform web scraping.
- Deploy the crawler using Amazon Lightsail and let it gather data.
- Store scraped data into Amazon CloudSearch and Amazon DynamoDB.
- Back-end to be deployed to Amazon Lightsail and will connect to CloudSearch and DynamoDB to provide functionality such as searching, perform categorizations, serve homepage and user profiles etc.
- Back-end will also manage
- Front-end to be deployed to Lightsail and will communicate with the backend.
- We plan to use Amazon SNS to send daily/weekly email digests to users.

Technical Roadmap:

- Build crawler
- Setup CloudSearch and DynamoDB
- Build back-end to manage crawler tasks and store data into CloudSearch, DynamoDB.
- Build back-end functionality to serve main page, user profile, provide searching for articles.
- Design front-end to communicate with back-end and show the user the aggregated articles.

TEAM PRESENTATION

Koli
Anuj
25

University of Florida (M.Sc.)

I am lovin it
That's not McDonald's tagline
That's people reaction when they see my code.

Shenoy
Anurag
26

University of Florida (MS)

I'm that guy who keeps saying "Did you know?".

Upadhyay
Ruchika
27

New York University

Sounds Good

PATHAK
AMAN
24

University of Florida (MS)

I'm that ML crazy guy.

Kumar
Naveen
26

Rajasthan Technical University

NA