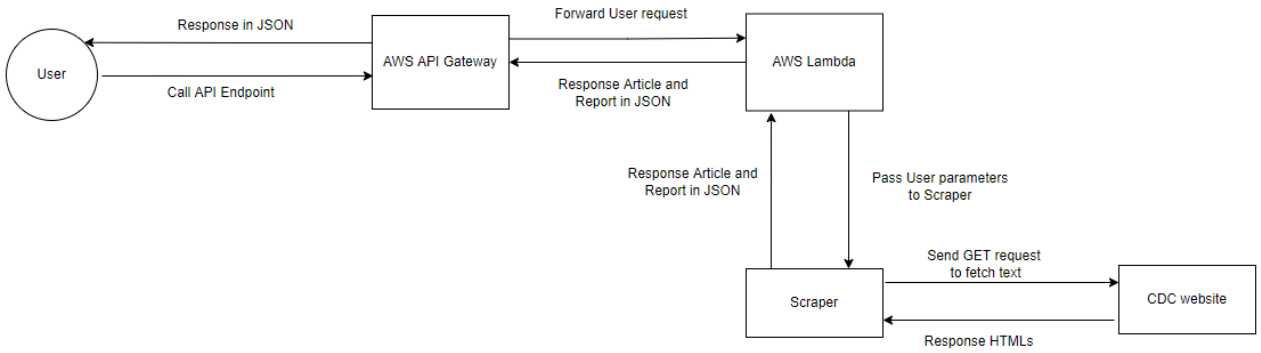
# D2 API design details

Final API architecture:




**To access the programs on AWS and log files, please follow the steps below**.

link to login: https://414306287714.signin.aws.amazon.com/console

Username: Sumeet

password: fVK6z%182hu'EEF

Access key ID: AKIAWA5UUIBRNPCQ6BZO

Secret access key: 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

**IMPORTANT STEP: After login, please change the region in up right corner to** Asia Pacific (Sydney)ap-southeast-2

SWagger UI link:

https://app.swaggerhub.com/apis-docs/GroupName/GroupNameAPI/1.0.0

Implementation Decisions:

1. Python requests is used to send get requests to CDC websites to get the HTML information and extract information which related to user requests from HTML by using beautifulsoup4 package.
2. While parsing HTML pages form CDC website, GeoText package is used to find cities' name from texts.

Challenges and Shortcomings:

1. CDC websites are not well formatted. In most cases, every article and case report has unique structure and format, which means we have to manually find and extract the information which users want, simply parsing HTML is not good enough to address the challenges, therefore, some of user's requests can not be processed properly.
2. Since most of articles and reports are not formatted, key information like event date has to be extracted by using REGEX, locating a information is time-consuming, because there are lots of string comparison take place during scraping the websites. This shortcoming causes users have to wait a real long time to get a response. Some outbreak reports do not have a universal pattern, the scraper can not extract cases, death or hospitalization numbers from texts, therefore '-1' is returned.


3. The API is deployed on AWS Lambda, the performance of the server is not very good, therefore, it causes the scraping process cost more time. (By comparing to a local machine, it takes double amount of time to process the same request)
4. Syndromes from appendix are very challenging to be extracted from texts, because there are many synonyms, which requires NLP to process the texts, which can not be done within two weeks, therefore, our API will specifically tell users that syndromes function is not implemented yet.
5. **IMPORTANT SHORTCOMING**, when the API is deployed on the AWS API Gateway, the time for processing user requests is restricted to 30s, if any request is timed out, a smaller period of interest is required.
   https://docs.aws.amazon.com/apigateway/latest/developerguide/limits.html
   This shortcoming is specified in above note, please search '30' to find the limit.
   This problem only appears in the '**/index**' endpoint, because that endpoint will scrape all the information from CDC when user make a request, the scraper needs more time to process wider period of interest, the reason of shortcoming is that scraping requires lots of regex string matching, string manipulation and the 30s time limit is not configurable. Especially when the key_term=salmonella and period_of_interest interval is longer than a year. In other cases, it only appears sometimes which we believe is due to CDC website's unstable delay in reponse. We recommand to try a few more times when meet 504 : Endpoint request timed out, and don't input period_of_interest interval longer than a year when key_term is salmonella.