SafeGraph or OpenStreetMap? This is the question.

MUSA 509 FINAL PROJECT

BY Eugene Chong, Bingchu Chen

01 WHY IS IT A QUESTION?

What is the Use Case?

COST EFFICIENCY

SafeGraph data is expensive and relatively complete. OSM data is free but much less complete.

Chances are we don't need 100% complete dataset.

Maybe, OSM's POI data is enough.

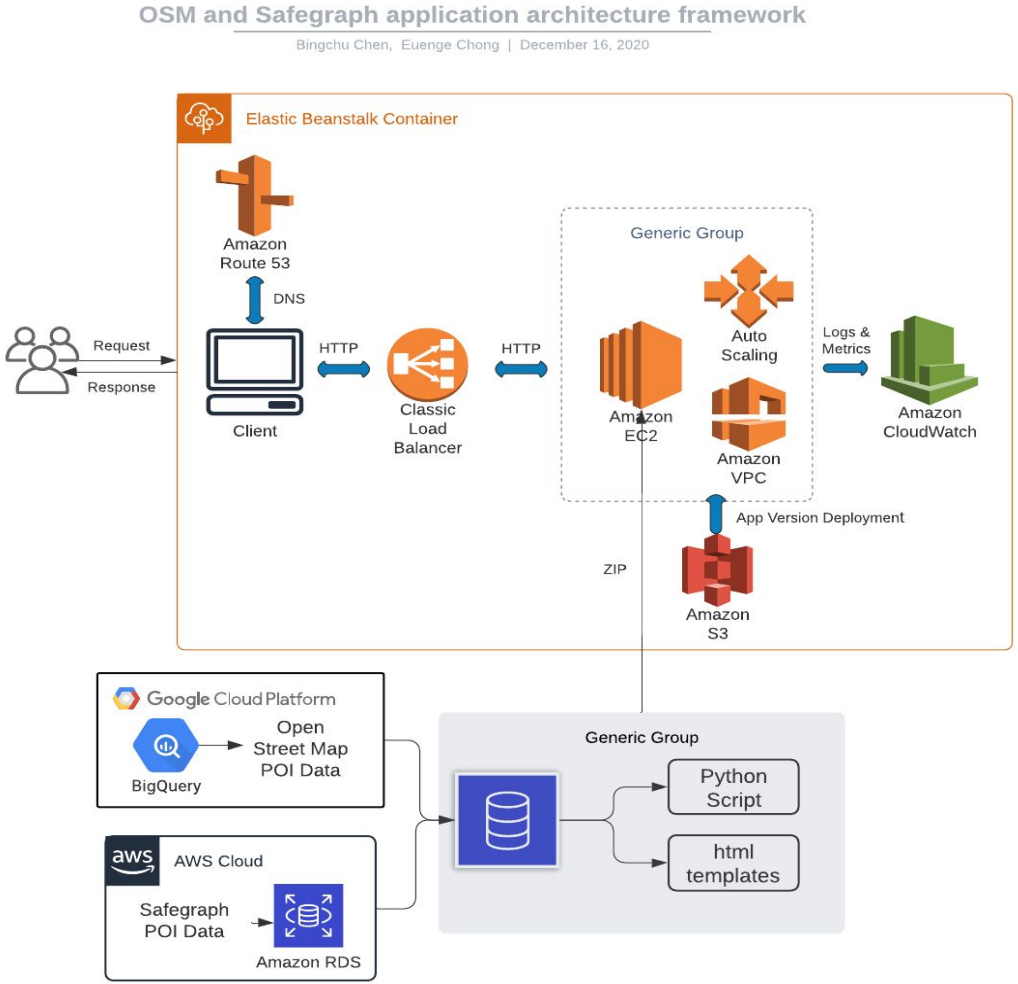
CONTRIBUTE MAP DATA

Help developers/editors find where to contribute on OSM

02 WHAT IS OUR SOLUTION?

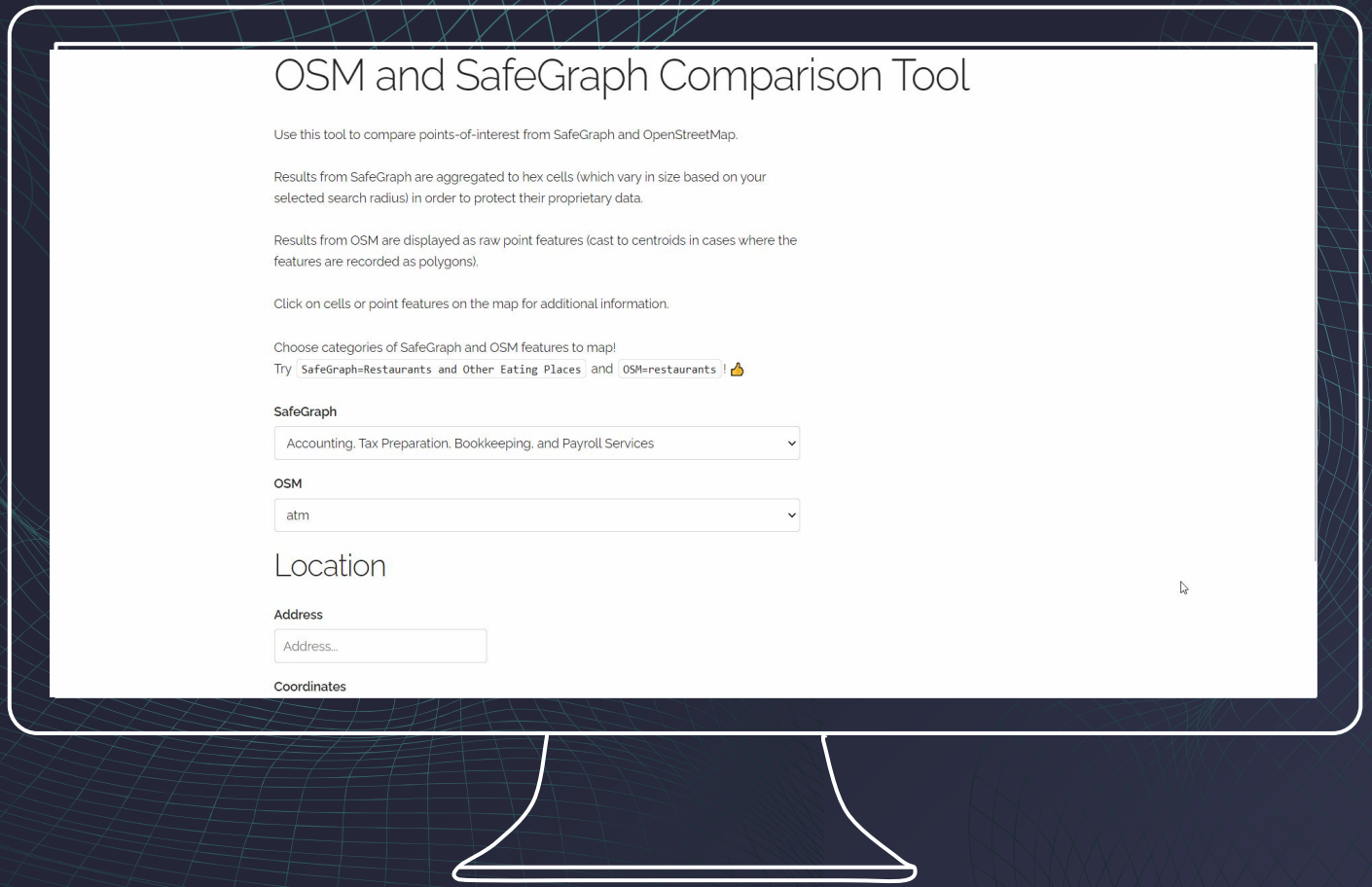
ARCHITECTURE

Our website is host on Elastic Beanstalk.



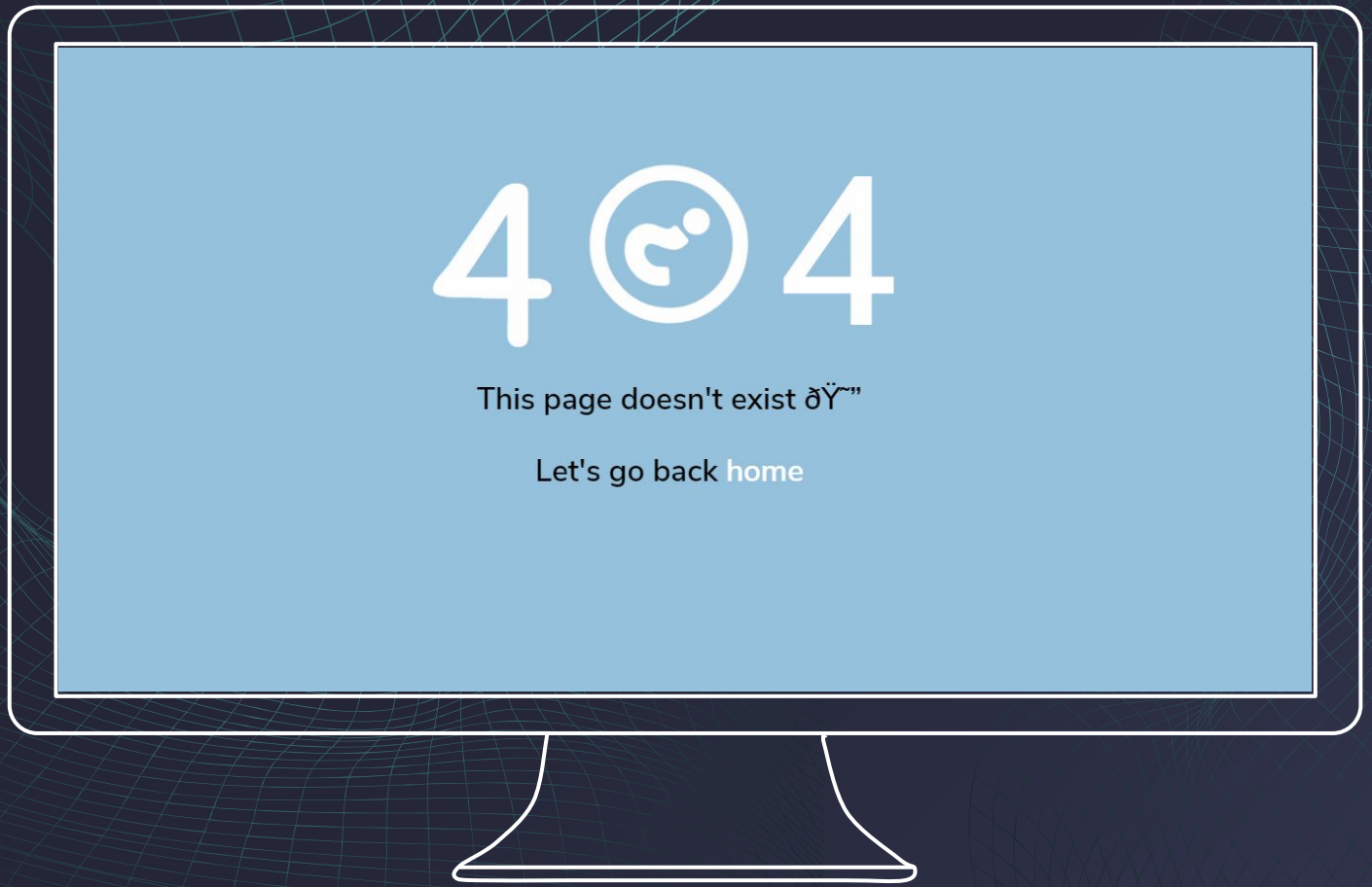
DESKTOP SOFTWARE

http://osmsg-env.eba-ac7vpbmy.us-east-1.elasticbeanstalk.com/



DESKTOP SOFTWARE

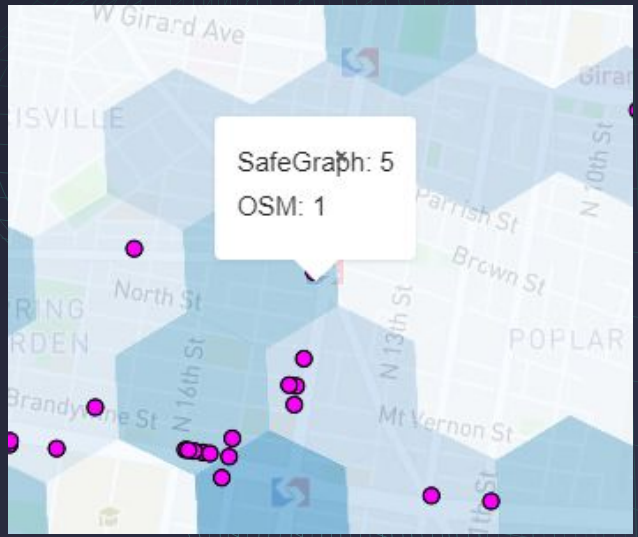
http://osmsg-env.eba-ac7vpbmy.us-east-1.elasticbeanstalk.com/



03 WHAT IS THE STRUCTURE?

How we made it?

Data



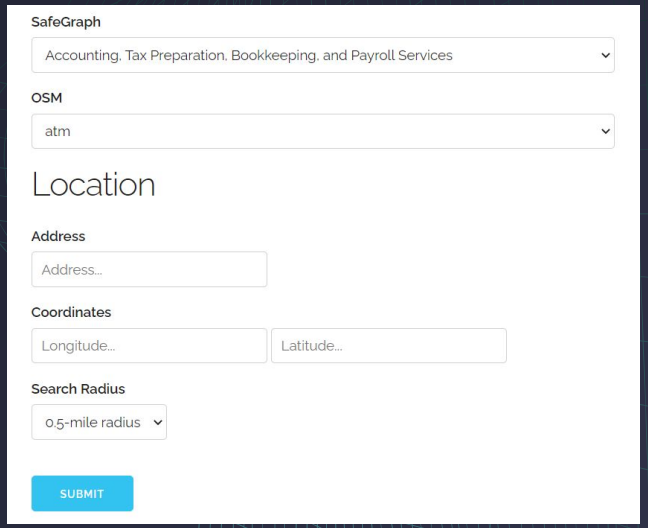
The data of this project comes from two sources:

- SafeGraph POI data stored in the postGIS
- OpenStreetMap data from the Google Bigquery.

We limit the total data set to those inside the boundary of United States and Canada, and to those that belong to categories with more than 10k instances.

Because we wanted to be careful not to re-post SafeGraph's POI point data (and possibly violate their TOS), we aggregated the original point data to H3 hex grids..

Frontend



- In the user interface page, we designed a couple of questions to help the users find the location and the OSM category/SafeGraph category code they want.
- Through the "submit" button, the app will navigate the users to our main interface where data are displayed on a map.
- All the pages are linkable.
- Users can always return to the landing page through the two buttons in the lower-left corner and change the location, category of interest, and search radius.

Backend

```
Flask>=1.1.2,<2.0.0
pandas>=1.1.3,<2.0.0
sqlalchemy>=1.3.20,<2.0.0
geopandas>=0.8.1,<1.0.0
bokeh>=2.2.3,<3.0.0
matplotlib==3.3.2
h3==3.7.0
GeoAlchemy2>=0.8.4,<1.0.0
google-api-core==1.23.0
google-auth==1.24.0
google-cloud-bigquery==2.6.1
google-cloud-bigquery-storage==2.1.0
google-cloud-core==1.4.4
google-crc32c==1.0.0
google-resumable-media==1.1.0
googleapis-common-protos==1.52.0
psycopg2-binary==2.8.6
requests>=2.25.0,<3.0.0
rtree==0.9.4
shapely>=1.7.1,<3.0.0
numpy>=1.19.2,<2.0.0
pyarrow
plotly
```

A Python/Flask backend is provided with a PostgreSQL host on AWS RDS.

- A Python/Flask backend is provided with a PostgreSQL server on AWS RDS.
- The requirements.txt file for the core dependencies of the project is provided here.

04 WHAT ARE THE CHALLENGES?

The challenges we met ?

- Slow queries. The SafeGraph data was ~7mm rows, and filtering and performing spatial aggregations means each query takes at least 20s, even with indexing.
- Adjusting H3 grid size. Hard-coded with search radius, but ideally this would be more dynamic and changed with user's zoom level.
- Hosting GeoPandas with spatial indexing operations on Elastic Beanstalk.

“This is harder than it seems. Definitely.”

— SOMEONE FAMOUS (i.e. us)

05 WHAT FUTURE IMPROVEMENTS?

Next-step for the web development

- Matching OSM categories with SafeGraph NAICS categories so that users can compare the POIs from both sources with one drop-down menu.
- Enable users to upload their own spatial geometry file.
- Come out a great website name.
- ...

THANKS!

Does anyone have any questions?