

Auto Log

Insight Data Engineering Fellowship, Silicon Valley
Gene Der Su

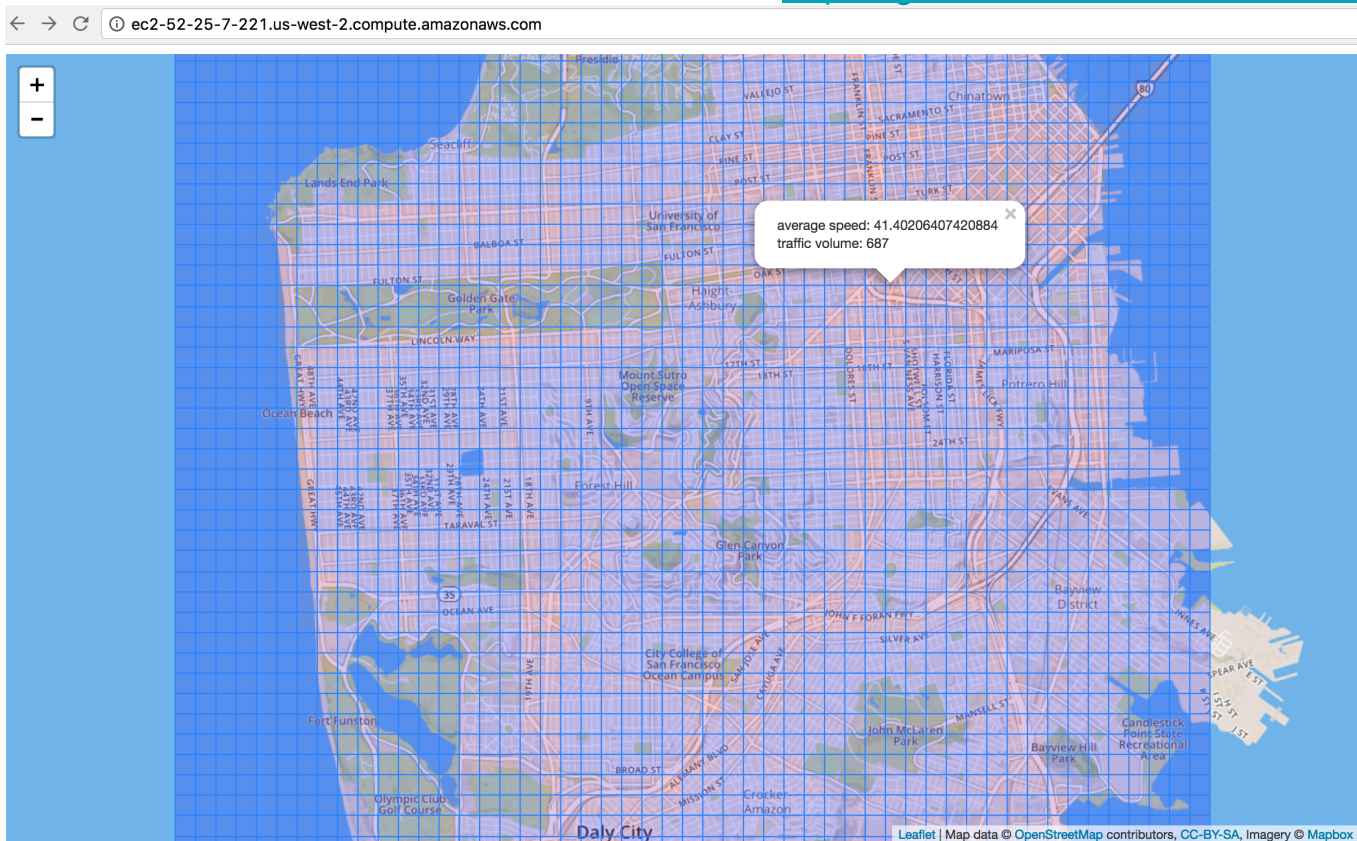
Motivation

- Traffic in the Bay Area is a headache
- This framework can be used in to avoid traffic dense areas and for companies to re-route their customers/ vehicles

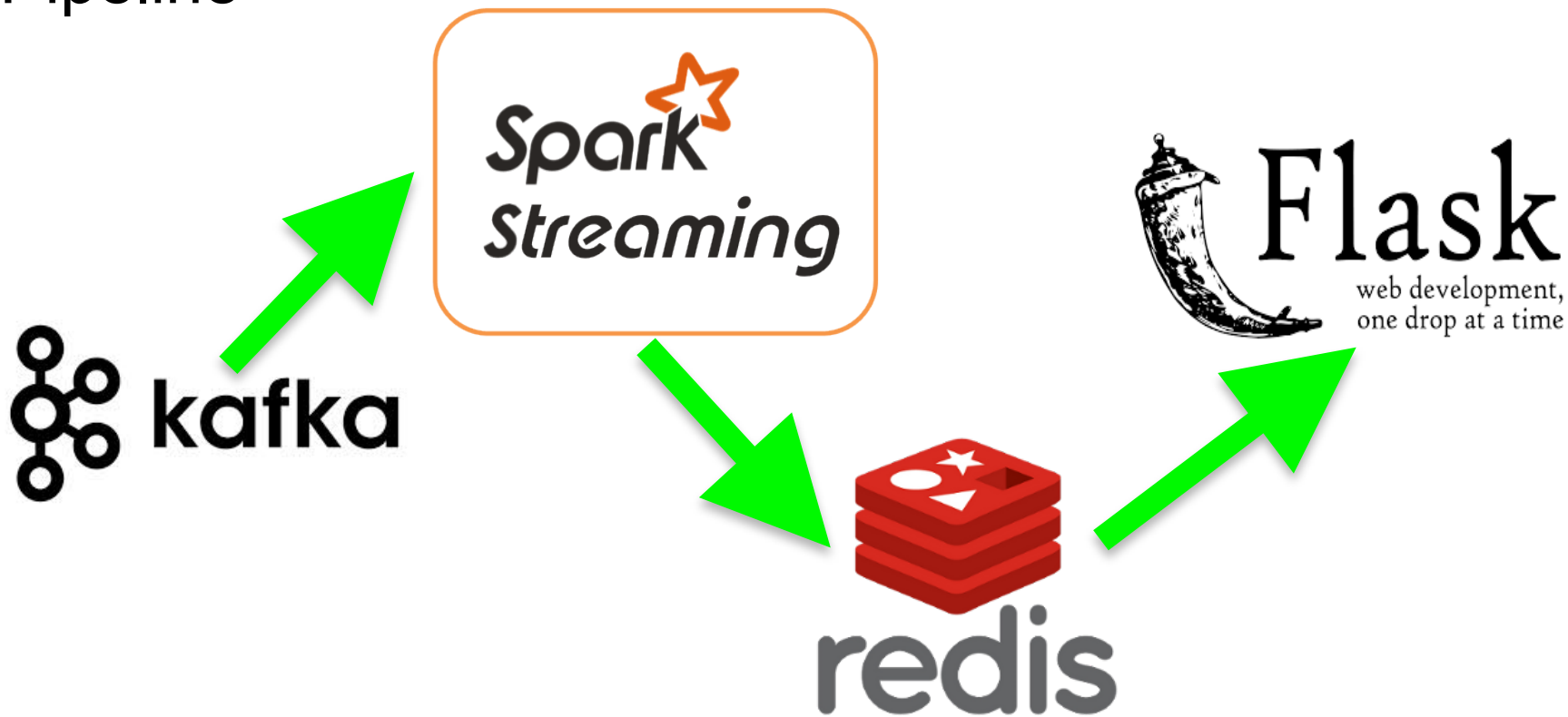
Product

<http://autolog.online/>

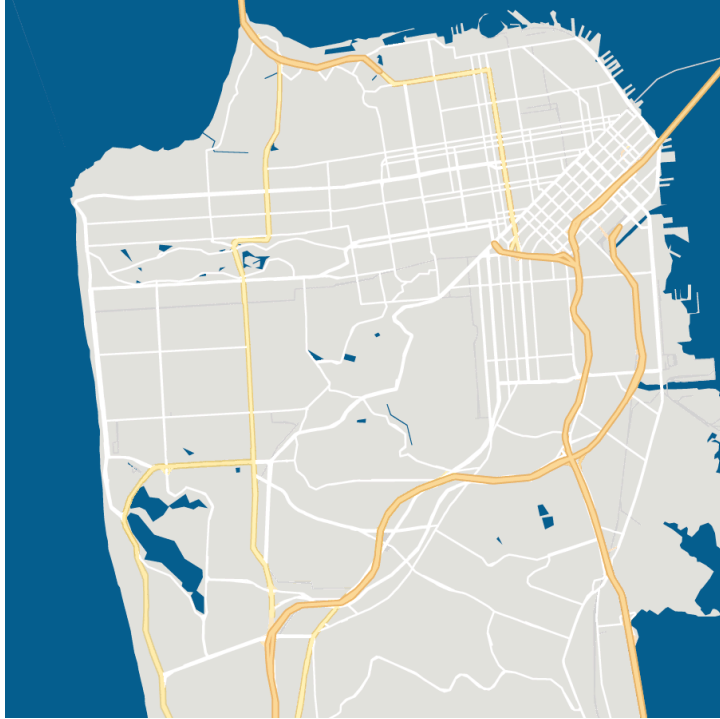
<https://github.com/GeneDer/Auto-Log#product>



Pipeline



Data



```
37.7626081578;-122.510030195;2;3;7.00117542979  
37.6982598733;-122.393299805;3;3;15.3936264874  
37.8096423911;-122.477586013;4;3;6.2898761708  
37.8000992133;-122.460591442;5;3;11.271637845  
37.7710606867;-122.445656818;6;3;15.2550536027
```

- Simulated on the simplified San Francisco map
- (latitude, longitude, car id, simulation iteration, randomly generated speed)

Queries

- The map is divided into 50 by 50 grids
- For each grid, the car density and the average is calculated and updated with the past 20 seconds data
- Since each grid is independent, all of them can be run in parallel on multiple nodes

Challenges

- Set up simulator that can mimic traffics in real world
- Compute and update the data in real time

Other considerations

- Simulated data can behave different than real world
- The simulator can generate **~280MB/minute** on one **t2.micro** with 1 million cars
- The pipeline uses 10 **m4.large**. It costs around \$790.6 per month or **\$26.35 per day** in a 30 day month

About me

Machine Learning Engineer at GoFind.ai

Masters in Computer Science from UC Davis

Bachelor in Applied Math from UC Merced

Love badminton and cycling



Miscellaneous

- The grid is 790 ft. by 790 ft. square (around 2 city blocks covered by each side)
- Maximum processing: 19s @ 7000 messages/sec
- Data can be filtered by timestamp so data transfer delay won't be a big issue
- Mapping of city block to an id can be done for showing specific traffic volume on each street
- Data skew is a good thing (reduce the complexity of computing the graph)