

# DataEng S22: Project Assignment 3

## Data Integration

**Due date:** May 22, 2022 @10pm PT

Submit at: <https://forms.gle/9UKRF6aCvXGESjYt7>

Congratulations! By now you have a working, end-to-end data pipeline. Unfortunately, it does not have enough data to properly implement our Data Scientist's visualization. To fill out information such as "route ID" you need to access another source of data and build a new pipeline to integrate it with your initial pipeline. Here are your steps:

- A. access the stop event data
- B. build a new pipeline for the stop event data
- C. integrate the stop event data with the bread crumb data
- D. testing

### A. Stop Event Data

Access C-Tran "Stop Event" data at this URL: <http://www.psudataeng.com:8000/getStopEvents>

As with the previous data source, this data set gives all C-Tran vehicle stop events for a single day of operation.

### B. New Pipeline

Your job is to build a new pipeline that operates just like the previous one, including use of Kafka, automation, validation and loading.

### C. Integrate Stop Events with Bread Crumbs

The two pipelines (BreadCrumb pipeline and StopEvent pipeline) must update the values in the Trip table such that all of the columns of both tables are filled correctly.

[5/20/2022] Alternatively, it would be OK to load the StopEvent data into a separate table and then use SQL views to integrate the two datasets. S

### D. Visualization

[MapboxGL](#) is a data visualization tool that allows you to view your bread crumb data and display it on a map. Your job is to integrate this tool with your database tables so that you can query the

breadcrumb and trip data in your database server, transform to geoJSON format and display the resulting map visualization. To get started, [see this guide](#).

[5/20/2022] Alternatively, you may use an alternative visualization tool (such as folium) to create the required visualizations. We do not provide any guides for doing it, but you are free to do so if you prefer. The submitted visualizations must be equivalent or superior to the visualizations produced by the provided MapBoxGL based visualization tool.

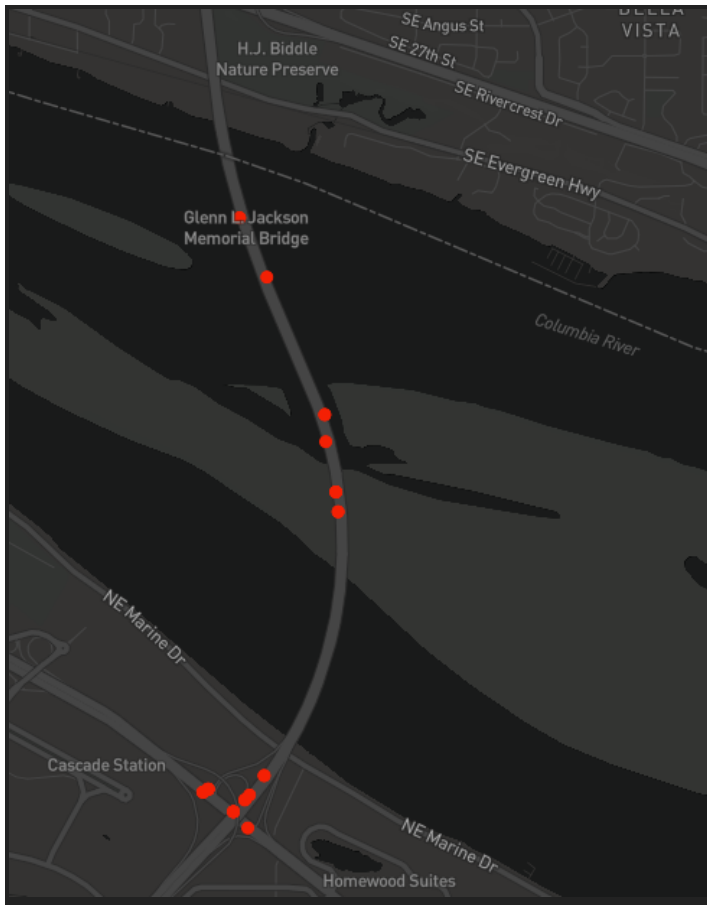
## Submission

Make a copy of this document and update it to include the following visualizations. For each visualization extract from your database a list of {latitude, longitude, speed} tuples and then use the provided visualization code (see Section D above) to display bus speeds at all of the corresponding geographic coordinates. So, for example, if you are asked to visualize a “trip”, then you must query your database to find all of the {latitude, longitude, speed} tuples for that trip, and then display a map showing the recorded/calculated bus speed at each {latitude,longitude} location.

No need to produce software that neatly displays trips, routes, dates, times, etc. onto the visualization itself. Instead, just paste a screen capture of the map-based speed visualization into your submission document and then include a text description of the contents of the visualization. For example, text like this: “Bus Speeds for all outbound trips of route 65 between 9am and 11am on Sunday October 32, 2020.”

Visualization 1. A visualization of speeds for a single trip for any bus route that crosses the Glenn Jackson I-205 bridge. You choose the day, time and route for your selected trip. To find a trip that traverses this bridge, consider finding a trip that includes breadcrumb sensor points within this bounding box: [45.592404, -122.550711, 45.586158, -122.541270]. Any bus trip that includes breadcrumb points within that box either crosses the bridge or goes swimming in the Columbia river!

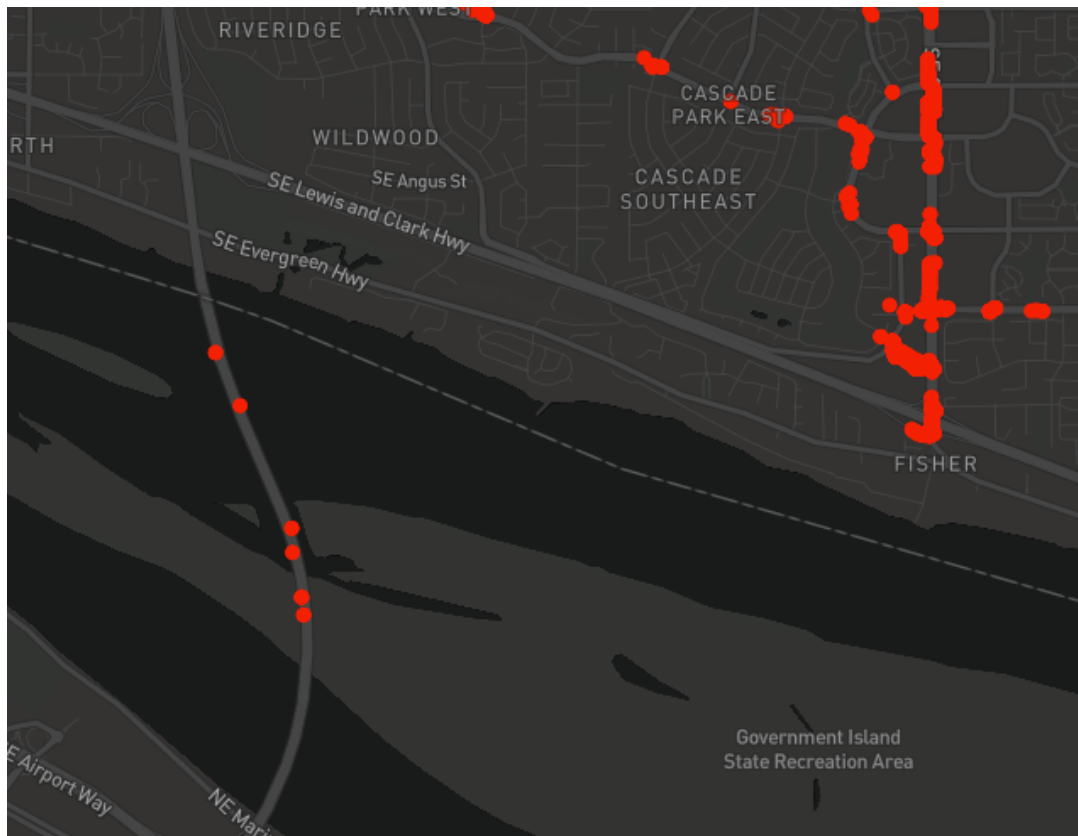
```
SELECT count(trip_id) FROM breadcrumb WHERE latitude between  
'45.586158' and '45.592404' and longitude between '-122.541270' and  
'-122.550711'
```



Visualization 2. All outbound trips that occurred on [route 65](#) on any Friday (you choose which Friday) between the hours of 4pm and 6pm.

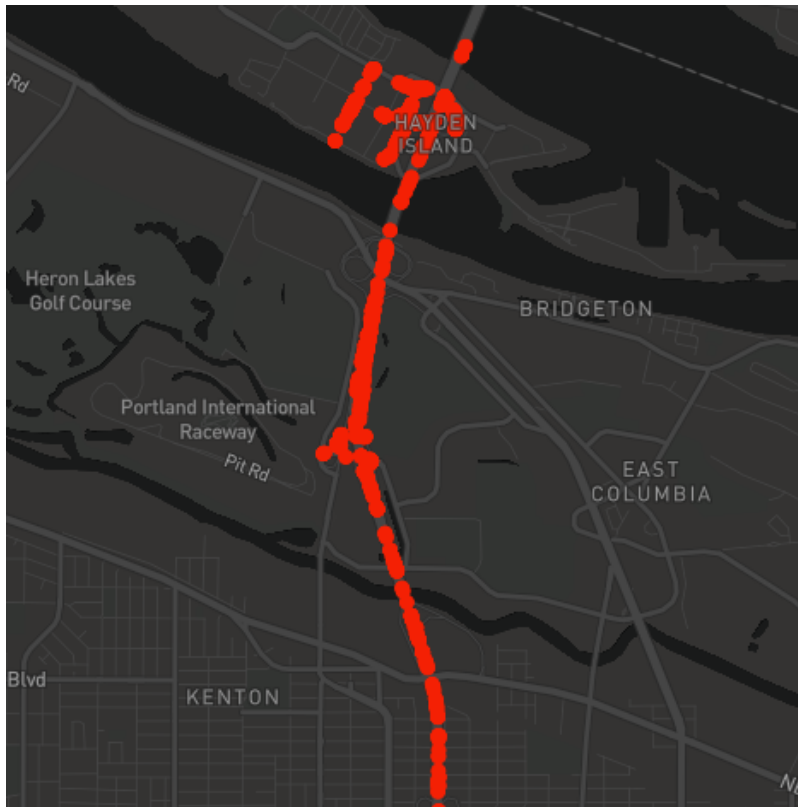
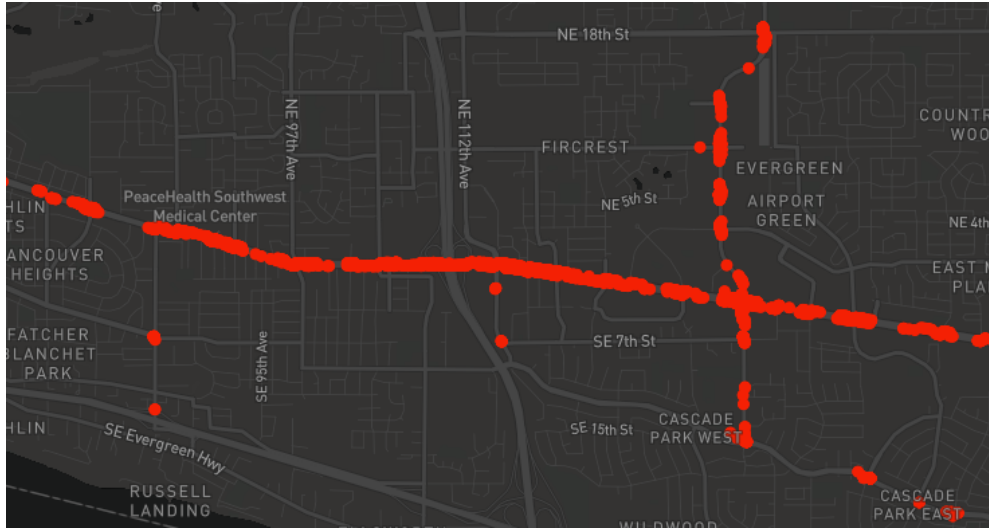


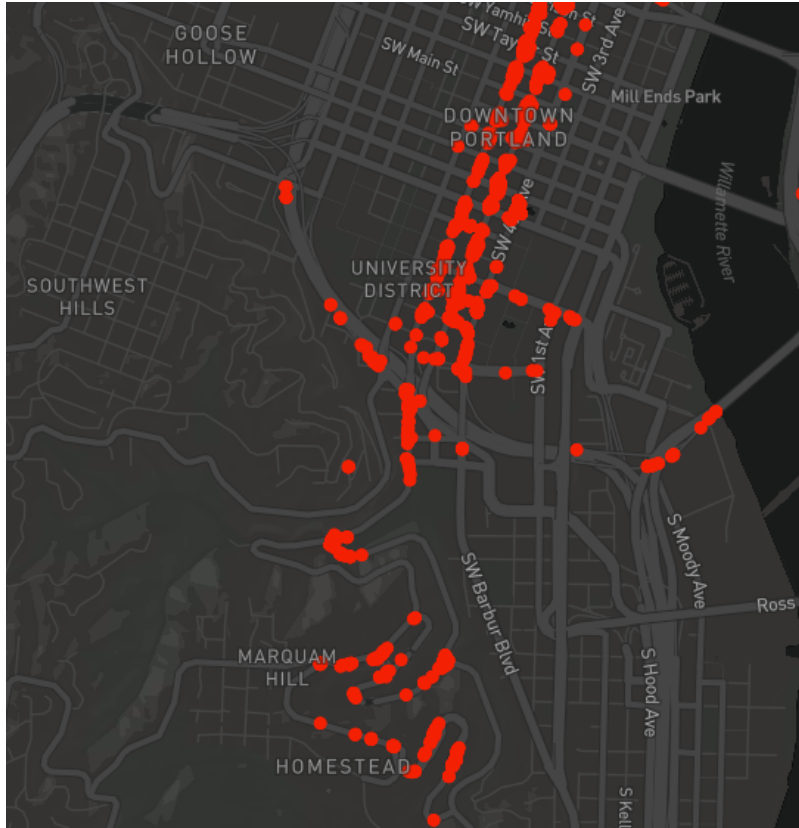
Visualization 3. All outbound trips for route 65 on any Sunday morning (you choose which Sunday) between 9am and 11am.



Visualization 4. The longest (as measured by time) trip in your entire data set. Indicate the date, route #, and trip ID of the trip along with a visualization showing the entire trip.

Visualization 5a, 5b, 5c, .... Three or more additional visualizations of your choice. Indicate why you chose each particular visualization.





## Your Code

Provide a reference to the repository where you store your code. If you are keeping it private then share it with Bruce and Genevieve.