C-Tran Data Pipeline

Team Breadbytes

Ebele Esimai, Genevieve Lalonde, Sara Alotaibi, Ahmed Dulaimi

Portland State University
Maseeh College of Engineering
Computer Science
CS410: Data Engineering
Instructor: Bruce Irvin

Ahmed

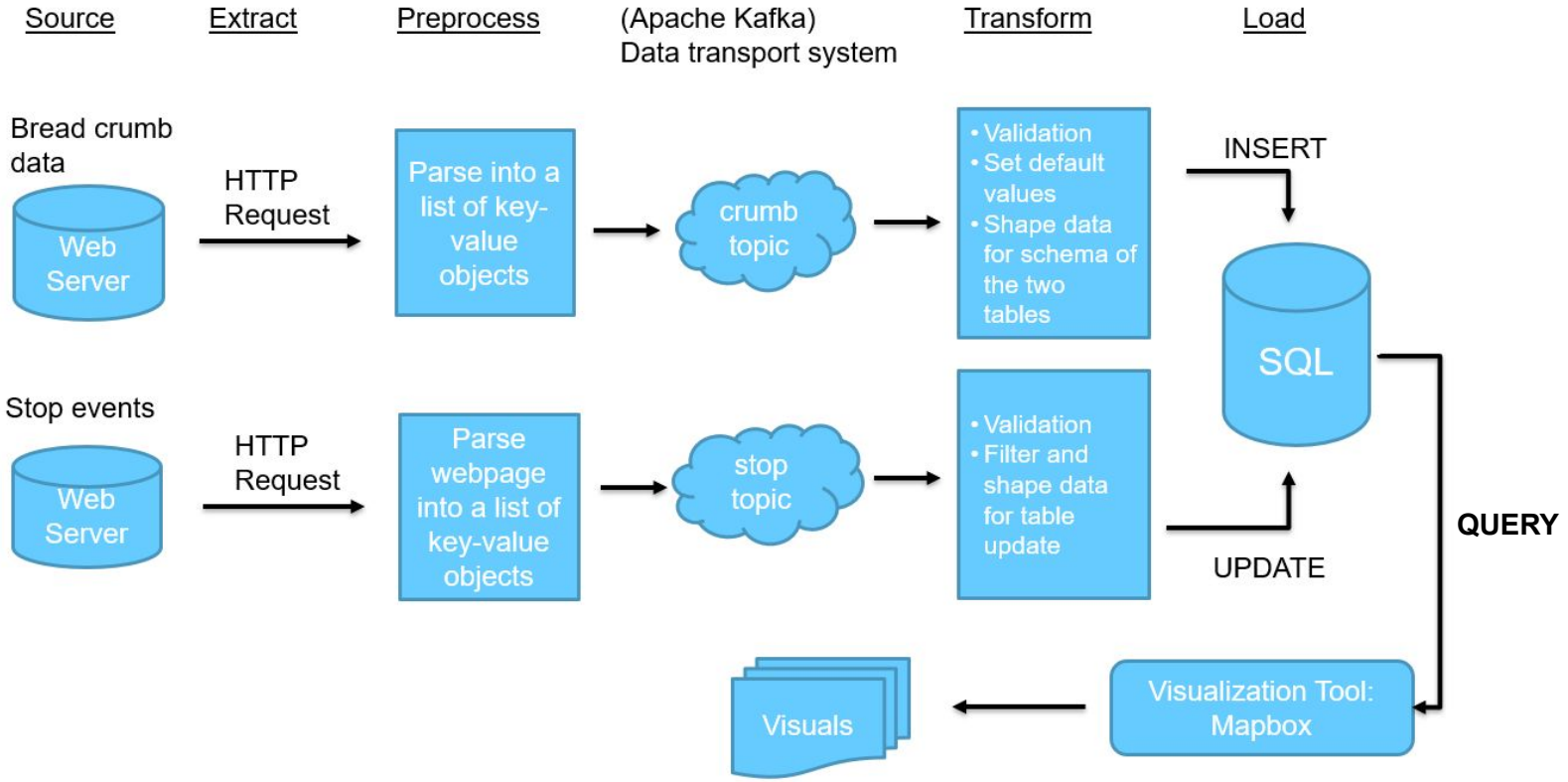
Introduction

C-Tran is the regional transit organization for Clark County in the State of Washington.

C-Tran captures GPS and transit data (bread crumbs) for each bus in its fleet daily.

We built a system that handles validation, transport and storage of the data for further analysis and visualization of congestion and transit patterns

System Architecture



Data Sources

- Data is provided by C-Tran
- Served from 2 simple web servers providing access to a day's worth of C-Tran GPS 5-second interval "breadcrumb" data, as well as stop event data.
- Breadcrumbs contain instantaneous information about a bus trip, including trip index, GPS data, odometer reading, etc.
- Stop event data provides C-Tran vehicles' stop events for a single day of operation


```
{
  "EVENT_NO_TRIP": "167311280",
  "EVENT_NO_STOP": "167311285",
  "OPD_DATE": "07-SEP-20",
  "VEHICLE_ID": "2262",
  "METERS": "269",
  "ACT_TIME": "31702",
  "VELOCITY": "8",
  "DIRECTION": "178",
  "RADIO_QUALITY": "",
  "GPS_LONGITUDE": "-122.604935",
  "GPS_LATITUDE": "45.637342",
  "GPS_SATELLITES": "10",
  "GPS_HDOP": "0.9",
  "SCHEDULE_DEVIATION": ""
},
```


Sample bread crumb

ETL Process

Kafka, an asynchronous event system, is a key component of the system for transporting the data

We split the ETL process into 2 stages in our data pipeline

- **Producer:**
  - Automatically gather the GPS sensor data from the web servers
  - Initially clean and parse it into individual breadcrumb and stop event readings
  - Publish each breadcrumb and stop event to a Kafka topic
- **Consumer:**
  - Automatically consume each breadcrumb and stop event
  - Validate the data
  - Transform the data to the shape needed by the database tables
  - Enhance the data to have absolute timestamps, correct data types and correct default values
  - Load the data into the Postgres database, either inserting new records for each table (bread crumb data) or updating existing records in a table (stop events data)

Pipeline and Data Summaries

Ebele

Pipeline Metrics

Crumb_topic - automated pipeline for daily GPS sensor readings for each bus in C-Tran's fleet.

Stop_topic - automated pipeline for daily stop events for the buses. Augments the bread crumb data

DOW	Average # Kafka Messages
Monday	371,292
Tuesday	368,449
Wednesday	369,417
Thursday	369,199
Friday	371,343
Saturday	175,313
Sunday	134,574

	Week range (06/03/2021 - 12/03/2021)						
	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
# Kafka Messages on crumb_topic	172,896	134,976	365,496	364,554	365,570	373,534	375,773
# Kafka Messages on stop_topic	720	640	1,499	1,702	1,696	1,530	1,523
# rows inserted in breadcrumb table	172,896	134,976	365,496	364,554	365,570	373,534	375,773
# rows inserted in trip table	863	714	1,702	1,702	1,696	1,735	1,733
# rows updated with stop events	720	640	1,499	1,497	1,491	1,530	1,523

Data Summaries

- Data is loaded on PostgreSQL database hosted on a GCP VM.
- Our database has two tables
  - BreadCrumb table: 9,806,199 records (size : 716 MB)
  - Trip table: 44,607 records (size: 2384 kB)
- Some data summaries include
  - Dates of loaded breadcrumb records range from 2020-09-25 to 2020-10-31
  - 104 vehicles in the system with 24 distinct routes
  - South-most location: 45.494323 , -122.683057 (Homestead, Portland, OR)
  - North-most location: 45.866877, -122.408082 (Yacolt, WA 98675)
  - Maximum bus speed : 159, Average bus speed: 10.345

Visualizations

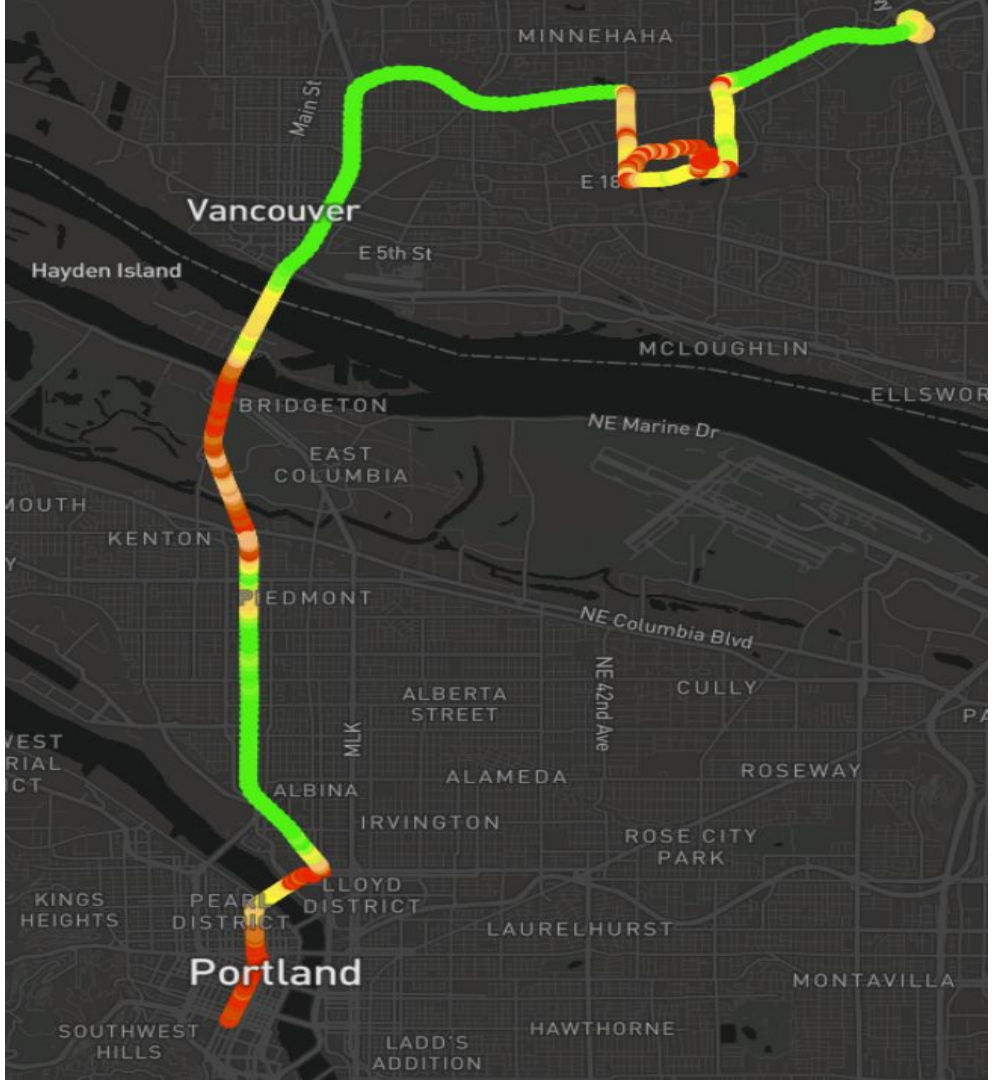
Longest trip

- Trip 169302880 on October 1st, 2020 is 5 hours and 30 minutes long, starting at 4:30 pm and ending at 10:02 pm
- However, the distance is only from the C-Tran facility in vancouver to downtown portland

```
project=# create temp table mytemp(diff interval, trip_id int);
CREATE TABLE
project=# insert into mytemp select age(foo.stp, foo.str) as diff, foo.trip
_id as tid from (select max(tstamp) as stp, min(tstamp) as str, trip_id fro
m breadcrumb group by trip_id) as foo;
INSERT 0 44607
project=# select * from mytemp where diff = (select max(diff) from mytemp);
   diff   |  trip_id
----------+-----------
 05:32:26 | 169302880
(1 row)
```

Longest trip

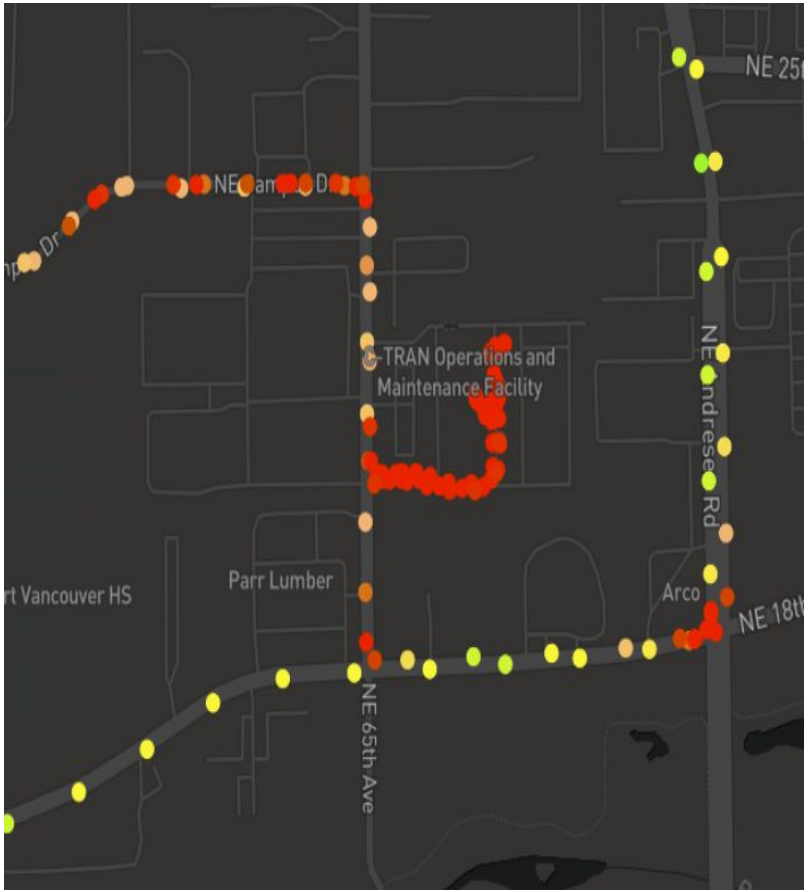
- Trip 169302880 on October 1st, 2020 is 5 hours and 30 minutes long, starting at 4:30 pm and ending at 10:02 pm.
- However, the distance is only from the C-Tran facility in vancouver to downtown portland
- Why is so long?



Ebele

Longest trip

2020-10-01 18:04:38	45.638033	-122.603113	6	4	169302880
2020-10-01 18:04:43	45.638215	-122.603127	357	4	169302880
2020-10-01 18:04:48	45.63836	-122.60313	359	3	169302880
2020-10-01 18:04:53	45.638438	-122.603132	359	1	169302880
2020-10-01 21:39:19	45.638388	-122.60346	258	0	169302880
2020-10-01 21:39:24	45.638378	-122.603445	134	0	169302880
2020-10-01 21:39:29	45.638347	-122.603443	178	0	169302880
2020-10-01 21:39:34	45.638337	-122.603422	123	0	169302880
2020-10-01 21:39:39	45.638278	-122.603283	121	0	169302880
2020-10-01 21:39:44	45.638275	-122.603238	96	0	169302880
2020-10-01 21:39:49	45.638287	-122.603247	333	0	169302880
2020-10-01 21:39:54	45.638287	-122.603235	90	0	169302880
2020-10-01 21:39:59	45.638283	-122.603213	102	0	169302880
2020-10-01 21:40:00	45.638283	-122.603213	0	0	169302880
2020-10-01 21:40:04	45.638333	-122.603233	344	1	169302880
2020-10-01 21:40:09	45.63853	-122.603197	7	4	169302880
2020-10-01 21:40:14	45.638725	-122.603177	4	4	169302880



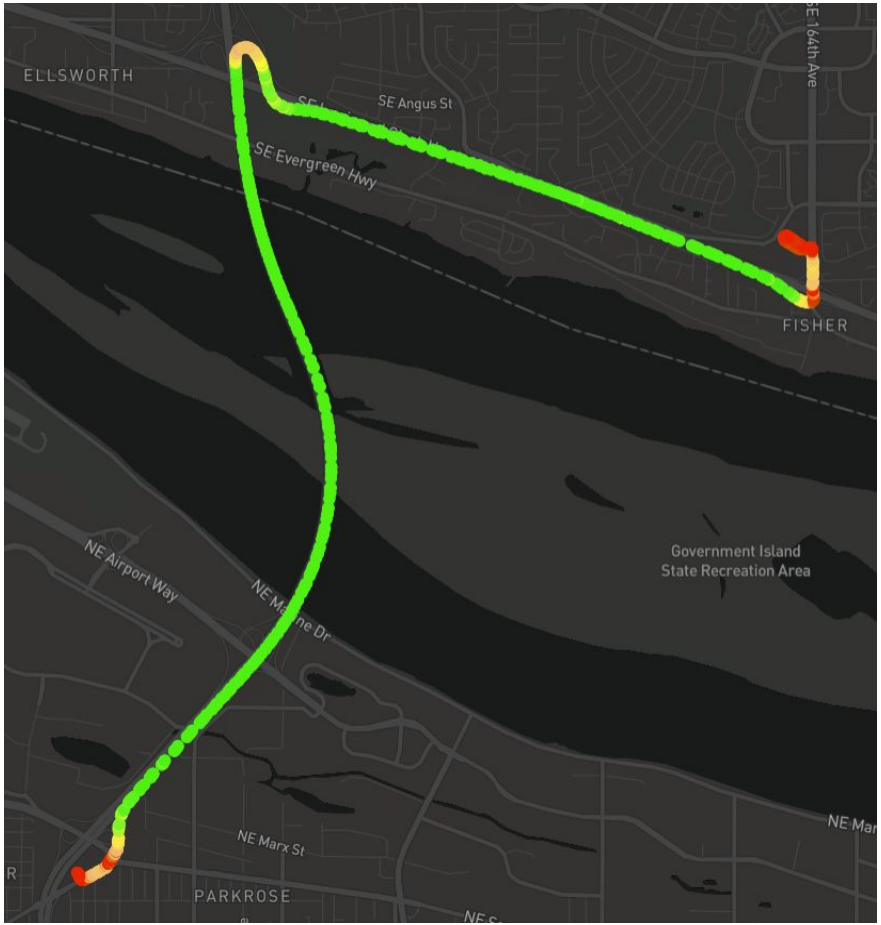
Sara

Morning vs Evening

- Bus 4008 on route 65 on October 18, 2020
- Morning outbound traffic (between 9am and 11 am)
- Evening outbound traffic (between 4pm and 6pm)
- Are there hotspots? If so where are they?

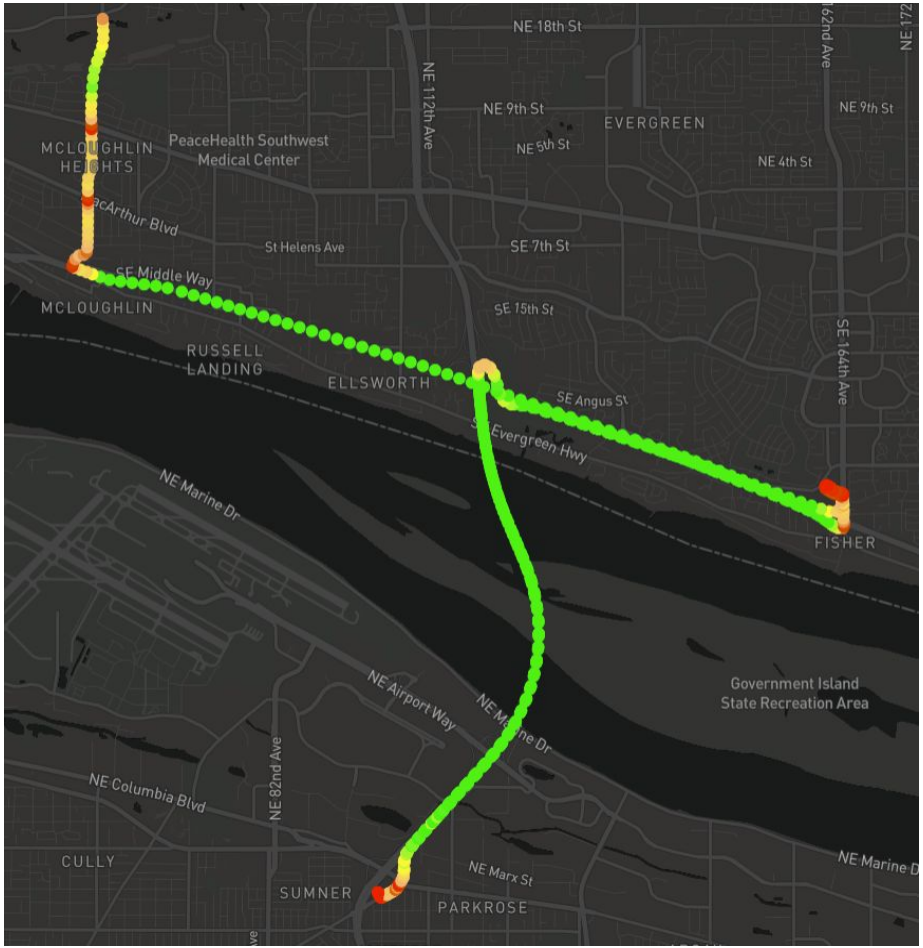
```
project=# select latitude ||' '||longitude, avg(speed) from BreadCrumb b join trip t on b.trip_id = t.trip_id where t.vehicle_i
d = 4008 and t.route_id =65 and date_part('month',b.tstamp) = 10 and date_part('day',b.tstamp) = 18 and date_part('hour',b.tsta
mp) between 16 and 18 group by 1;

project=# select latitude ||' '||longitude, avg(speed) from BreadCrumb b join trip t on b.trip_id = t.trip_id where t.vehicle_i
d = 4008 and t.route_id =65 and date_part('month',b.tstamp) = 10 and date_part('day',b.tstamp) = 18 and date_part('hour',b.tsta
mp) between 9 and 11 group by 1;
project=#
```

Morning



Evening

Sara



Our thoughts...

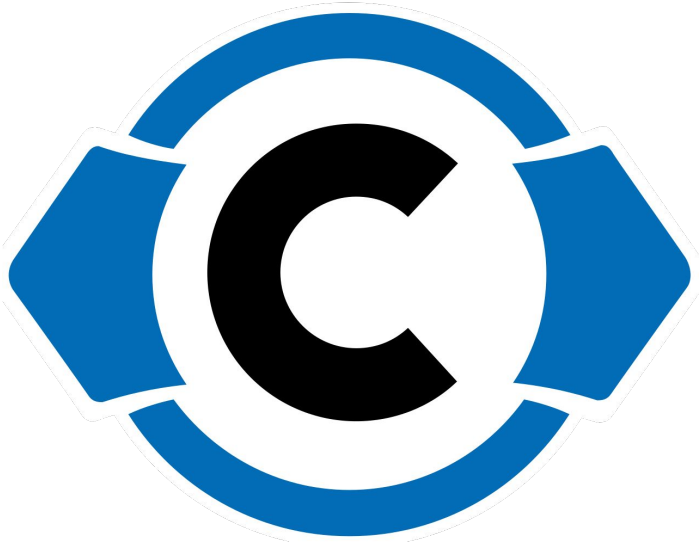
Challenges and actions

- Working with Kafka
  - Fine-tuning the producer code for polling and flushing
  - Dealing with in memory buffer sizes
  - Keeping the pipeline running and automated
- Data validation
  - Understanding the data to have reasonable validation criteria
  - Dealing with missing values and uniqueness
- System setup
  - Memory and disk resizing to accommodate size of data
  - Automation with cron jobs
  - Fault tolerance of pipeline and storage

Sara

Lessons Learned

- Some domain knowledge and understanding of the data are key factors of success
- Extract, Transform,and Load (ETL) process is how most data pipelines are usually designed and structured
- Communication is a key factor in defining requirements and establishing shared understanding of the desired resulting data



Thank you!