Storing and Sharing Transit Data at Scale: A Workflow for Big Data Without Big Costs

TransitData 2026 Symposium
University of Toronto

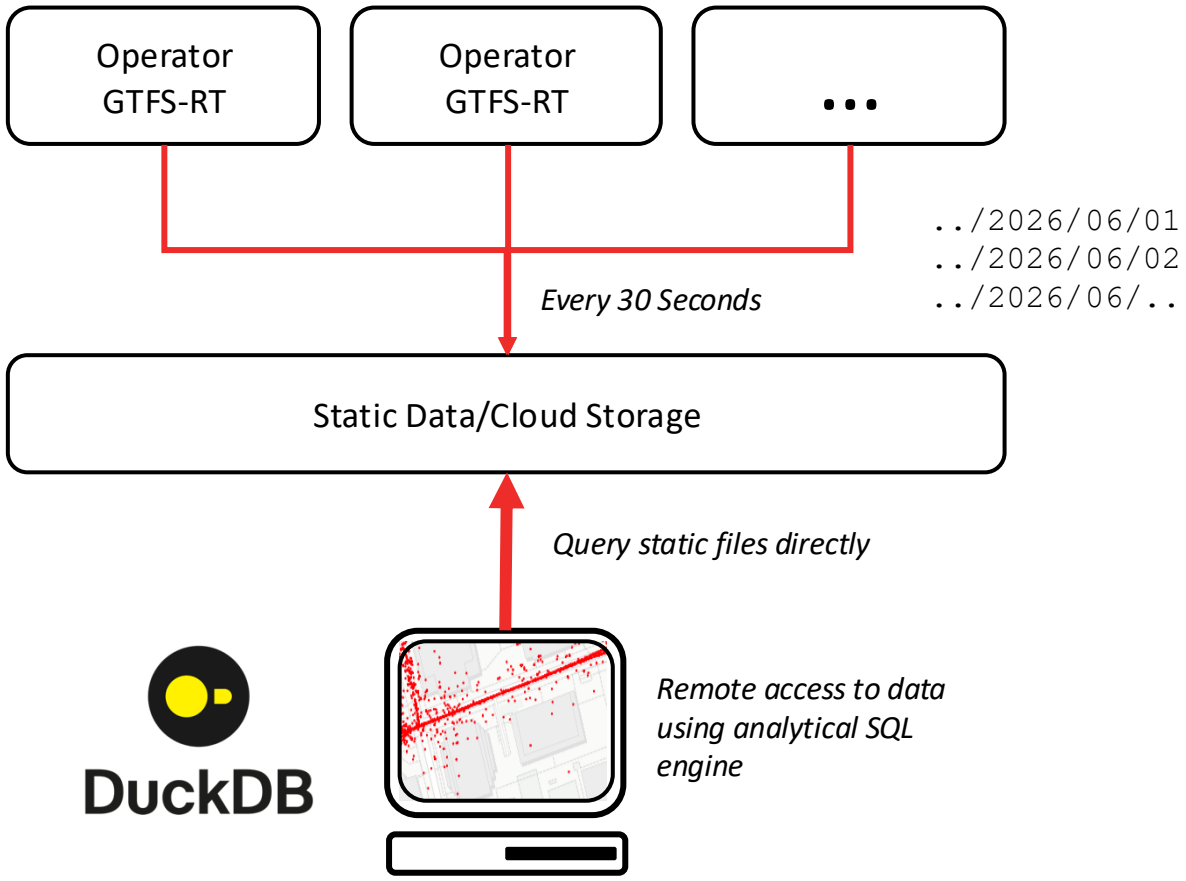
Saadiq Mohiuddin, MSc, P.Eng.
Development Seed, saadiq@developmentseed.org

Saeid Saidi, PhD, P.Eng.
University of Calgary, ssaidi@ucalgary.ca

June 25, 2026

Transit Data is Abundant, But Difficult to Access

- Transit operators collect millions of AVL/APC records every day
- Most timeseries data remains locked in enterprise databases or proprietary systems
- Researchers, students, and planners face barriers to access
- Propose cost effective system to store and distribute historical GTFS-RT Feed for time series and spatial analysis



Databases to Cloud-Native Files

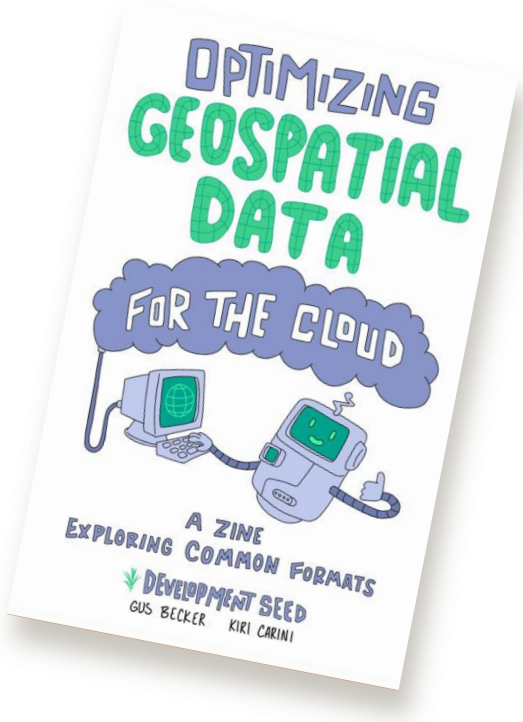
- Instead of downloading, parsing, organizing thousands of large files, we can read part of the data remotely
- Data stored as GeoParquet - columnar format, smaller file size and faster queries
- Works directly with cloud storage (Amazon S3)
- No database credentials or specialized GIS systems required

Row Based

```
trip_id, route_id, latitude
15236, 10-1, 118.2654
15326, 10-1, 118.6548
15689, 10-1, 118.9874
```

Column Based

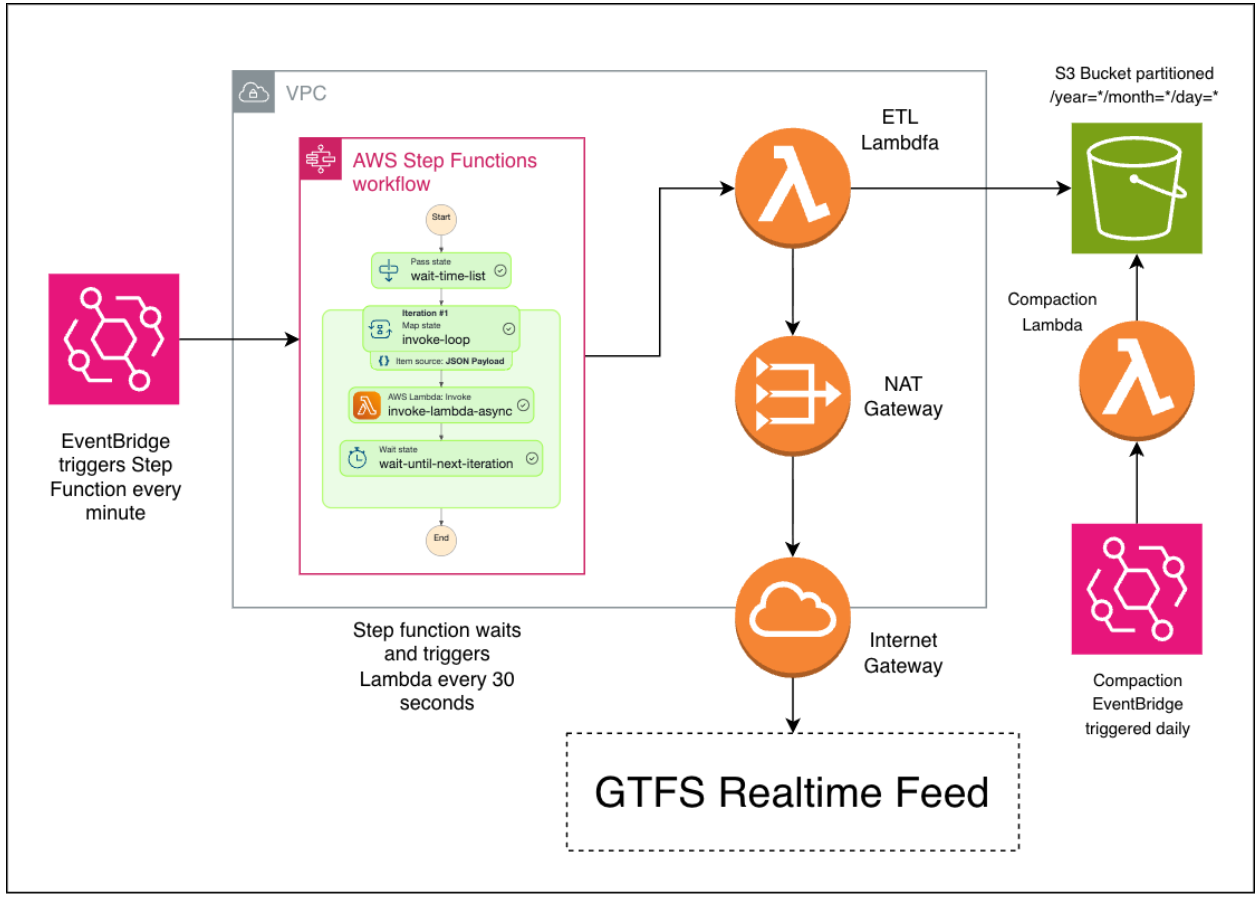
```
trip_id: 15236, 15326, 15326
route_id: 10-1, 10-1, 10-1
latitude: 118.2654,
118.6548,118.9874
```



https://zines.developmentseed.org/zines/cloud-native/

A Low-Cost, Scalable Data Pipeline

- Automatically collects GTFS-Realtime vehicle locations from public GTFS-RT APIs
- Captures data as frequently as every 10 seconds
- Requires minimal maintenance and no dedicated servers



https://github.com/smohiudd/gtfs-realtime-etl

Collecting Data for Six Canadian Transit Systems

- Data collection over six months:
  - Calgary Transit
  - Edmonton Transit Service
  - OC Transpo
  - TransLink
  - Toronto Transit Commission
  - Société de transport de Montréal
- 300 Million vehicle location points collected, frequency of 30 seconds
- **~$100 USD/month** cloud compute costs for data collection
- **$10 USD/month** storage costs

Inspect GTFS-RT endpoint stability

Accessing the Data

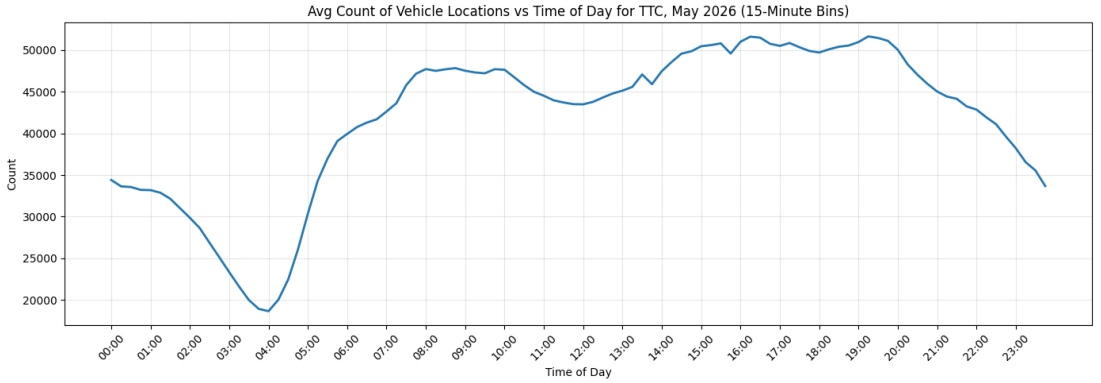
[10]:
```
%%sql
CREATE VIEW IF NOT EXISTS locations AS
SELECT * FROM read_parquet('s3://gtfs-rt-etl-data/{{agency}}/positions/year=2026/month=05/*/*.parquet', hive_partitioning = true);
FROM (DESCRIBE locations);
```

[10]:

	column_name	column_type	null	key	default	extra
0	trip_id	VARCHAR	YES	None	None	None
1	route_id	VARCHAR	YES	None	None	None
2	direction_id	VARCHAR	YES	None	None	None
3	vehicle_id	VARCHAR	YES	None	None	None
4	bearing	DOUBLE	YES	None	None	None
5	speed	DOUBLE	YES	None	None	None
6	timestamp	TIMESTAMP WITH TIME ZONE	YES	None	None	None
7	geohash	VARCHAR	YES	None	None	None
8	geometry	GEOMETRY	YES	None	None	None
9	bbox	STRUCT(xmin DOUBLE, ymin DOUBLE, xmax DOUBLE, ...	YES	None	None	None
10	day	VARCHAR	YES	None	None	None
11	month	VARCHAR	YES	None	None	None
12	year	BIGINT	YES	None	None	None

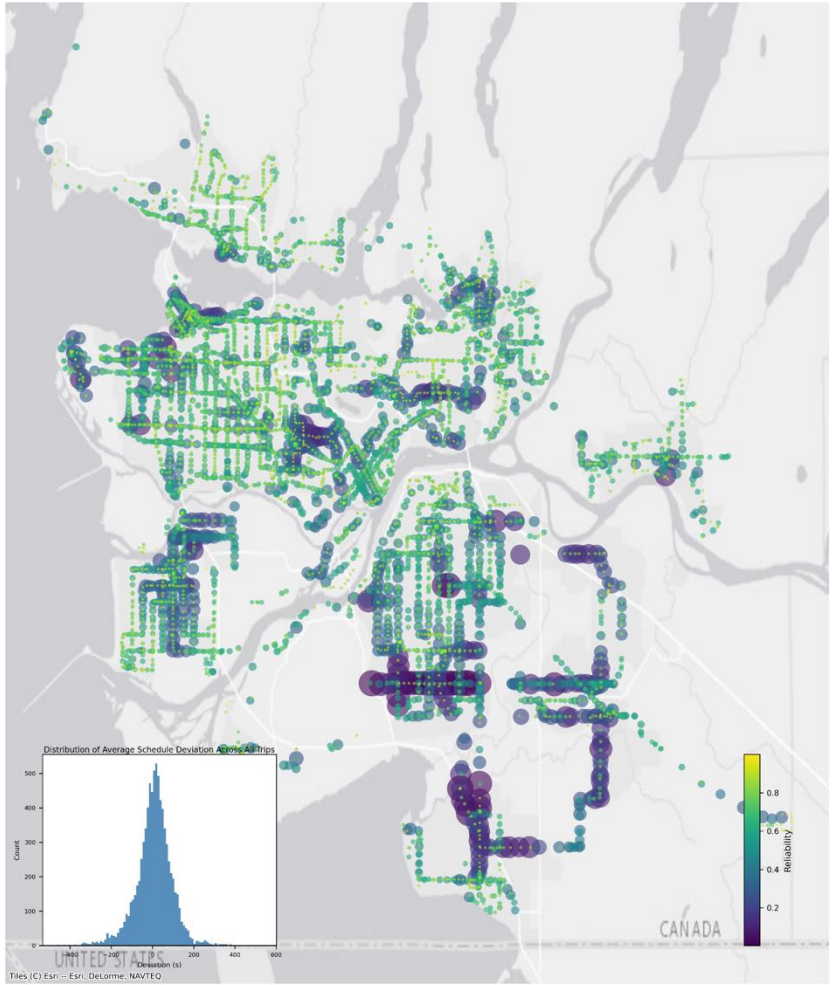
[11]:
```
%%sql df_counts<< WITH daily_bins AS (
    SELECT
        CAST(timestamp AS DATE) AS service_date,
        FLOOR(
            (date_part('hour', timestamp) * 60 +
             date_part('minute', timestamp)) / 15
        ) * 15 AS minute_bin,
        COUNT(*) AS bin_count
    FROM locations
    GROUP BY service_date, minute_bin
)

SELECT
    minute_bin,
    AVG(bin_count) AS avg_count
FROM daily_bins
GROUP BY minute_bin
ORDER BY minute_bin;
```

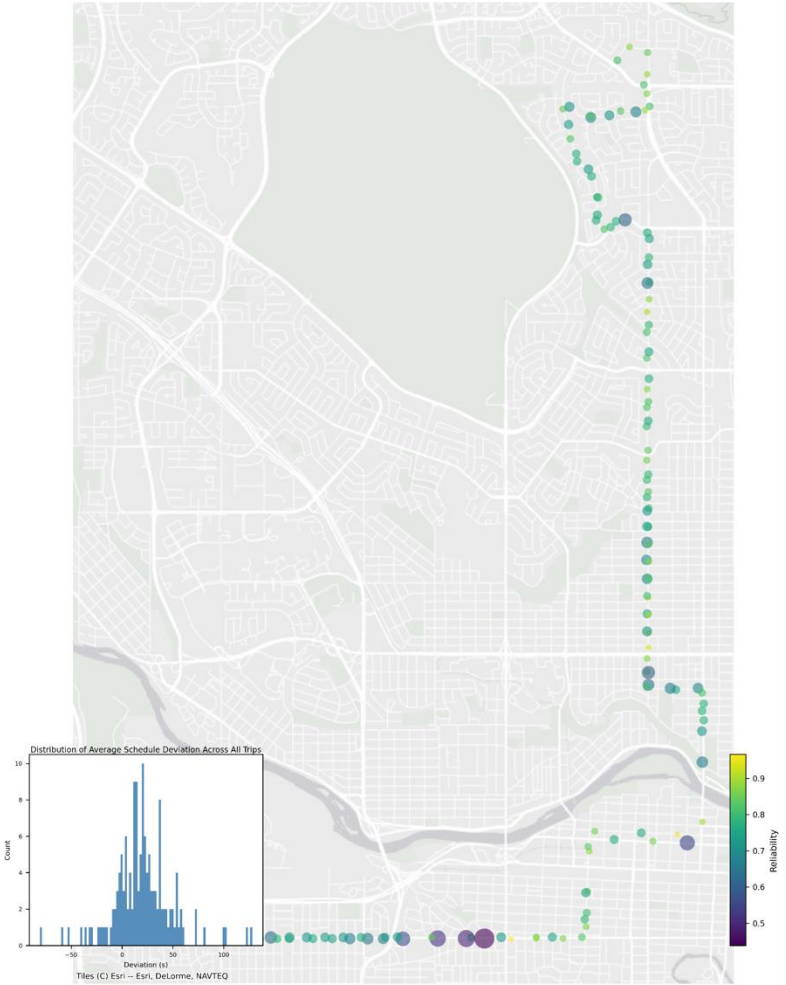
- Remotely query millions of records in a web interface (Jupyter Notebook) without downloading any data



Spatial Queries

- GeoParquet spatial metadata efficiently queries millions points, joining with GTFS (schedule) data
- Open-source, reproducible, and cost effective





Future work

- Use TIDES Specification
- Continue collecting data until end 2026 (12-month duration)
- Explore access, visualization, further analysis

Saadiq Mohiuddin, MSc, P.Eng.
Development Seed, saadiq@developmentseed.org

Saeid Saidi, PhD, P.Eng.
University of Calgary, ssaidi@ucalgary.ca