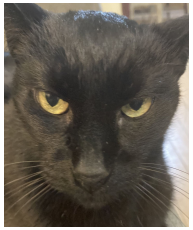
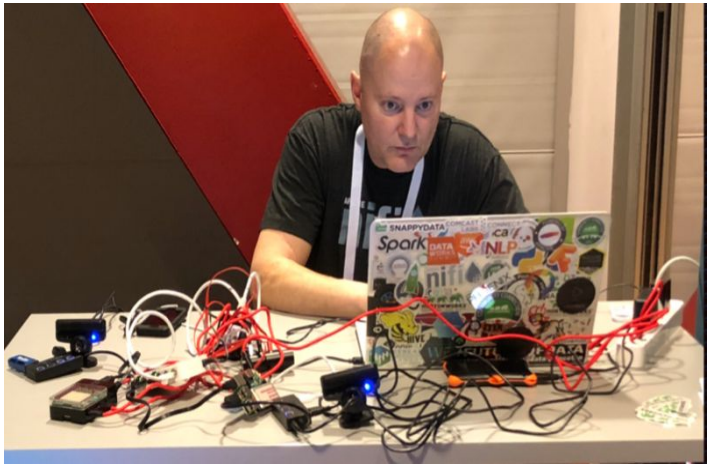
COMMUNITY
THE ASF CONFERENCE
CODE

Utilizing Real-Time Transit Data for Travel Optimization

Tim Spann, Senior Solutions Engineer, Snowflake

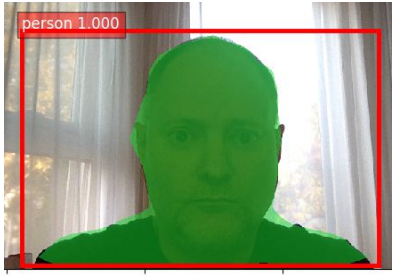


Tim Spann



paasdev.bsky.social
@PaasDev // Blog: **datainmotion.dev**
Senior Solutions Engineer, Snowflake
NY/NJ/Philly - Cloud Data + AI Meetups
ex-Zilliz, ex-Pivotal, ex-Cloudera, ex-HPE, ex-StreamNative, ex-Hortonworks.

https://medium.com/@tspann
https://github.com/tspannhw



snowflake

Streamlit

AI + Streaming Weekly by Tim Spann

https://bit.ly/32dAJft

This week in Snowflake, Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Streamlit, Jupyter, Apache Iceberg, Apache Polaris, Python, Java, LLM, GenAI, Vectors and Open Source friends.

There are a lot of factors involved in determining how you can find our way around and avoid delays, bad weather, dangers and expenses. In this talk I will focus on public transport in the largest transit system in the United States, the MTA, which is focused around New York City. Utilizing public and semi-public data feeds, this can be extended to most city and metropolitan areas around the world. As a personal example, I live in New Jersey and this is an extremely useful use of open source and public data.

Once I am notified that I need to travel to Manhattan, I need to start my data streams flowing. Most of the data sources are REST feeds that are ingested by Apache NiFi to transform, convert, enrich and finalize it for usage in Parquet files stored as Apache Iceberg tables.

https://medium.com/@tspann/populating-an-open-lakehouse-with-codeless-data-streams-04292375ddaf

Nearby mta

MTA ridership
https://medium.com/@tspann/building-transit-ridership-with-cursor-ai-utomatically-d73cbdcba3e9

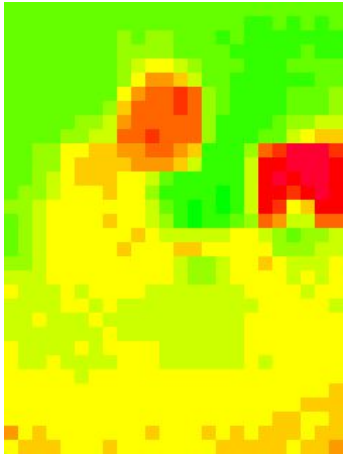
https://medium.com/cloudera-inc/finding-the-best-way-around-7491c76ca4cb

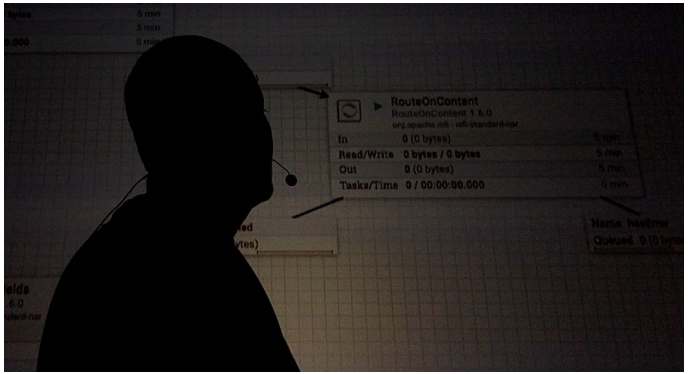
Geospatial
https://medium.com/cloudera-inc/boston-wheres-my-bus-llm-streaming-to-the-rescue-586dfd019237
https://medium.com/cloudera-inc/iteration-1-building-a-system-to-consume-all-the-real-time-transit-data-in-the-world-at-once-4322b160df9d
https://medium.com/cloudera-inc/iteration-2-building-a-system-to-consume-all-the-unsecured-real-time-transit-data-in-the-world-5f365c0cf240
https://medium.com/cloudera-inc/subways-and-transit-updates-in-real-time-30c104c359ef
https://medium.com/cloudera-inc/nyc-traffic-are-you-kidding-me-6d3fa853903b

Introduction

Overview

Get Hacking





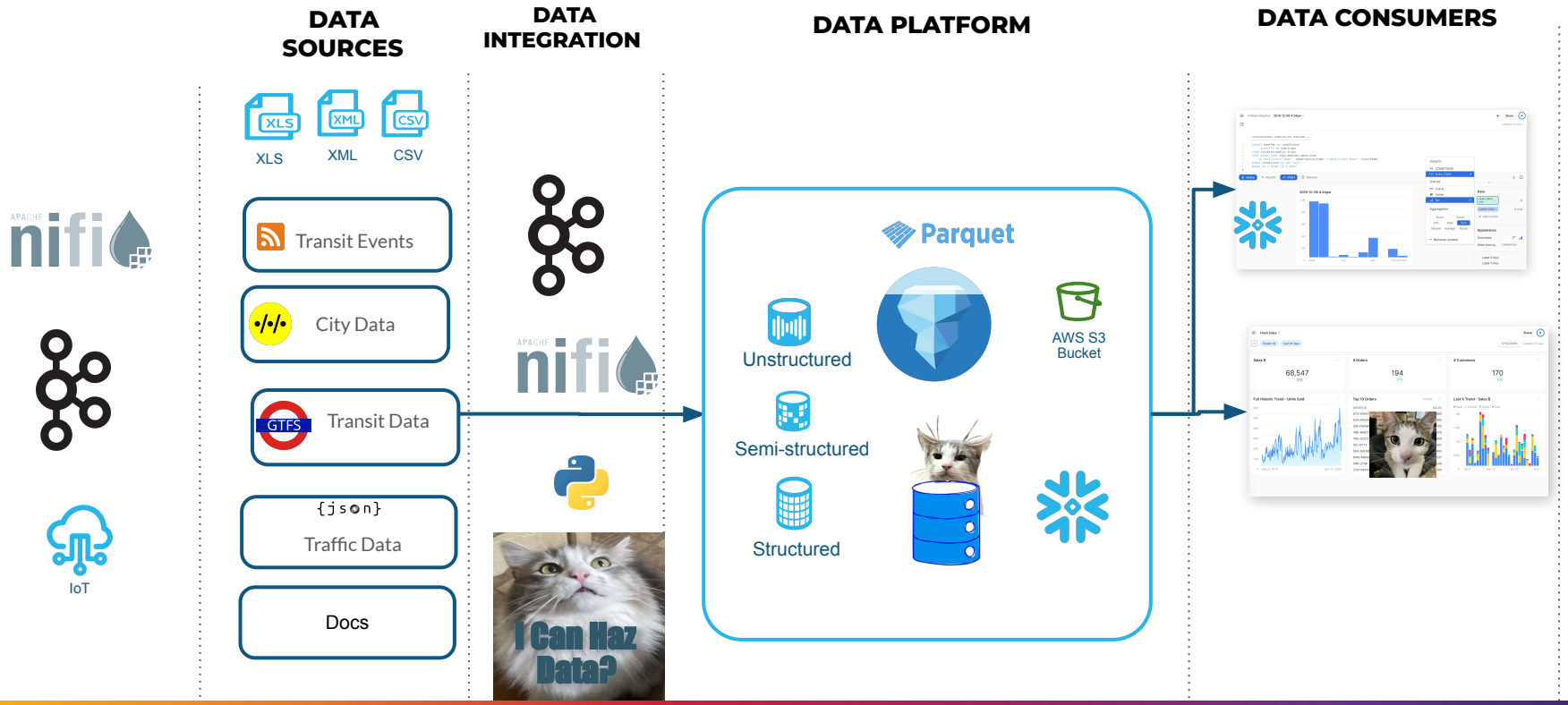
General (Google) Transit Feed Specification

GTFS - Protocol Buffers (binary format)

Organizations:

- Open Transit Software Foundation
- Mobility Data
- Google
- GTFS.org

Real-Time AI Open Lakehouse with Open Data



https://511ny.org/developers/help

https://github.com/tspannhw/hackathon2025

http://web.mta.info/developers/data/nyct/subway/StationComplexes.csv

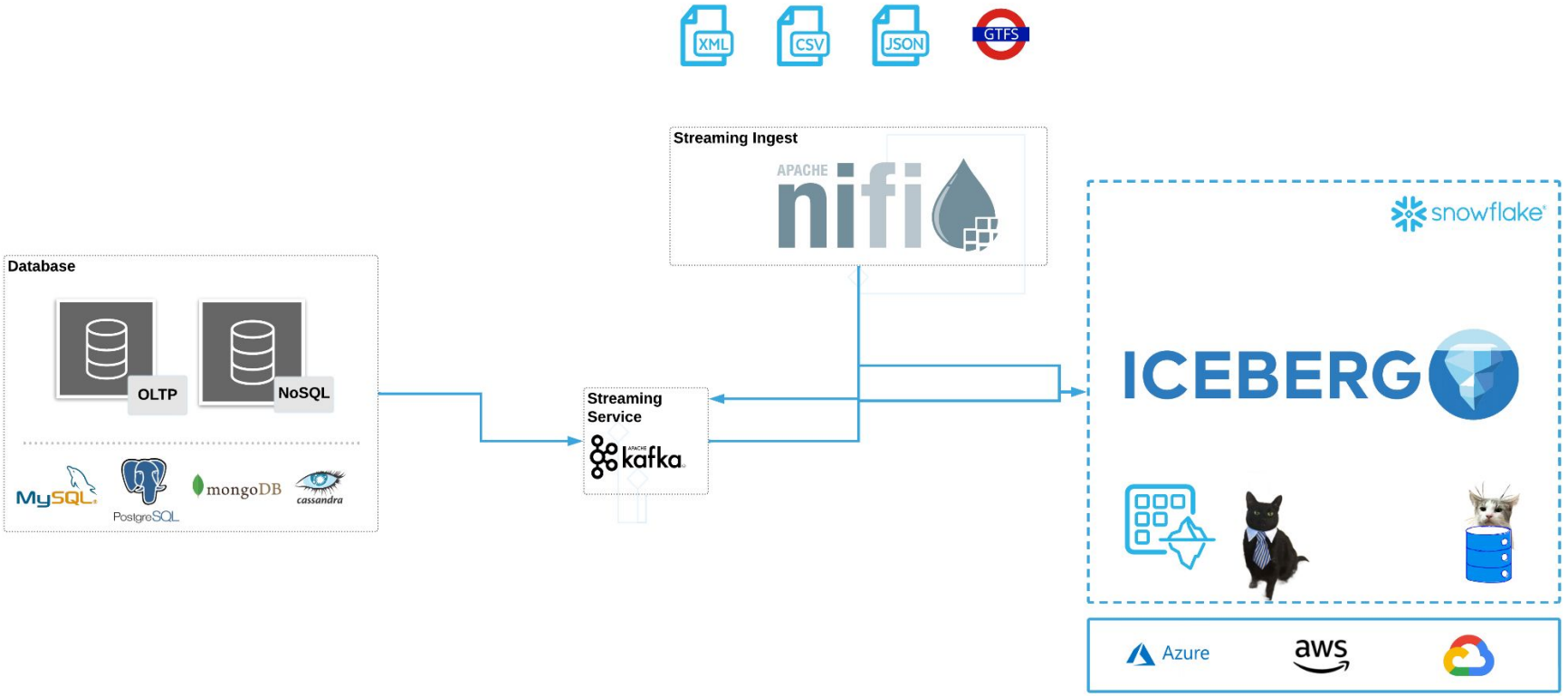
https://data.ny.gov/Transportation/MTA-Daily-Ridership-and-Traffic-Beginning-2020/sayj-mze2/about_data

https://data.ny.gov/Transportation/MTA-Bus-Service-Delivered-2015-2019/tw28-zvtk/about_data

https://data.ny.gov/Transportation/MTA-Congestion-Relief-Zone-Vehicle-Entries-Beginni/t6yz-b64h/about_data


RAGTRTOWN
Map of
NEW YORK
HOUSE NUMBER
AND
TRANSIT GUIDE
RAGTROM COMPANY
HUDSON
RIVER
EAST
RIVER
NORTH
RIVER
NEW YORK
BAY
HOBOKEN
JERSEY CITY


50
Street
50
Street


XML
CSV
JSON
GTFS
Streaming Ingest
APACHE
nifi
Database
OLTP
NoSQL
MySQL
PostgreSQL
mongoDB
cassandra
Streaming
Service
kafka
snowflake
ICEBERG
Azure
aws

Open New York Data Links

Developer Signup
https://511ny.org/developers/resources

Cameras
https://511ny.org/api/getcameras?key=SignUpForFreeKey&format=json

Alerts
https://511ny.org/api/getalerts?key=SignUpForFreeKey&format=json

Events
https://511ny.org/api/getevents?key=SignUpForFreeKey&format=json

Message Signs
https://511ny.org/api/getmessagesigns?key=SignUpForFreeKey&format=json

MTA Subway Hourly Ridership
https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-2025/5wq4-mkjj/about_data

Open New York Data Links

MTA Stations
https://data.ny.gov/resource/i9wp-a4ja.json

MTA Bus
https://bustime-beta.mta.info/developers/siri/#
https://bustime-beta.mta.info/developers/siri/monitored-vehicle-journey

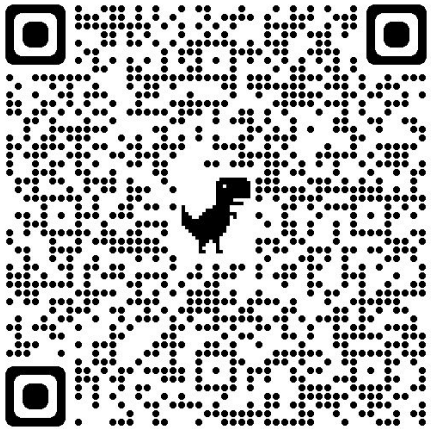
DOT Speed
https://data.cityofnewyork.us/Transportation/DOT-Traffic-Speeds-NBE/i4gi-tjb9/about_data
https://dev.socrata.com/foundry/data.cityofnewyork.us/i4gi-tjb9
https://medium.com/cloudera-inc/nyc-traffic-are-you-kidding-me-6d3fa853903b

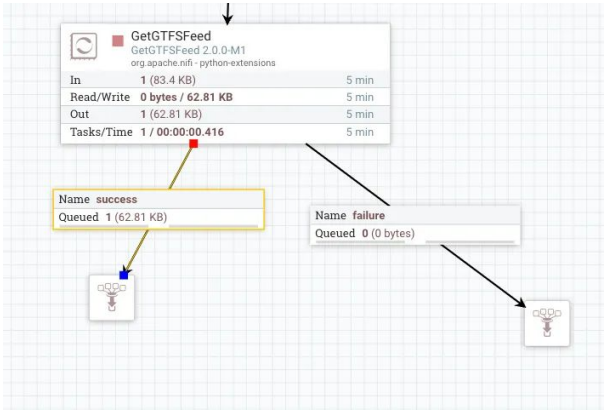
Deeper Resources

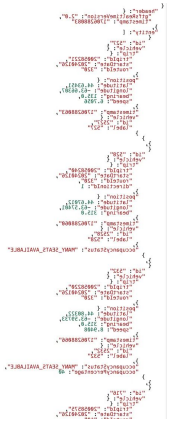
https://github.com/tspannhw/TrafficAI
https://medium.com/@tspann/automating-transit-gtfs-flows-7a5c4d3dafe8
https://medium.com/cloudera-inc/iteration-1-building-a-system-to-consume-all-the-real-time-transit-data-in-the-world-at-once-4322b160df9d
https://medium.com/cloudera-inc/iteration-2-building-a-system-to-consume-all-the-unsecured-real-time-transit-data-in-the-world-5f365c0cf240
https://medium.com/cloudera-inc/subways-and-transit-updates-in-real-time-30c104c359ef
https://medium.com/data-engineering-with-dremio/introducing-apache-iceberg-1-9-0-native-geospatial-support-enhanced-row-lineage-and-more-dead8950d391
https://carto.com/blog/iceberg-geoparquet-carto-reshaping-geospatial
https://forrest.nyc/airflow-ai-iceberg-v3-the-new-stack-for-scalable-geospatial-data/

Get GTFS Data

- Python 3.10+
- GTFS from Transit URL
- Alerts, Trip Updates or Vehicle Positions
- Returns JSON
- google.transit and google.protobuf

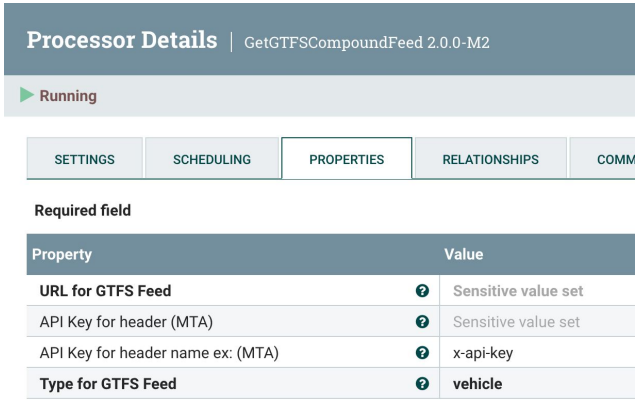




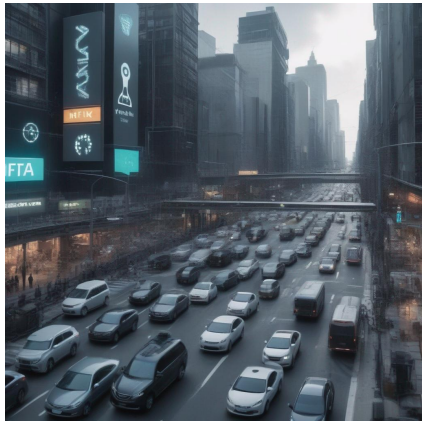


Get Compound GTFS Data

- Python 3.10+
- GTFS to JSON



https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetGTFSCompoundFeed.py



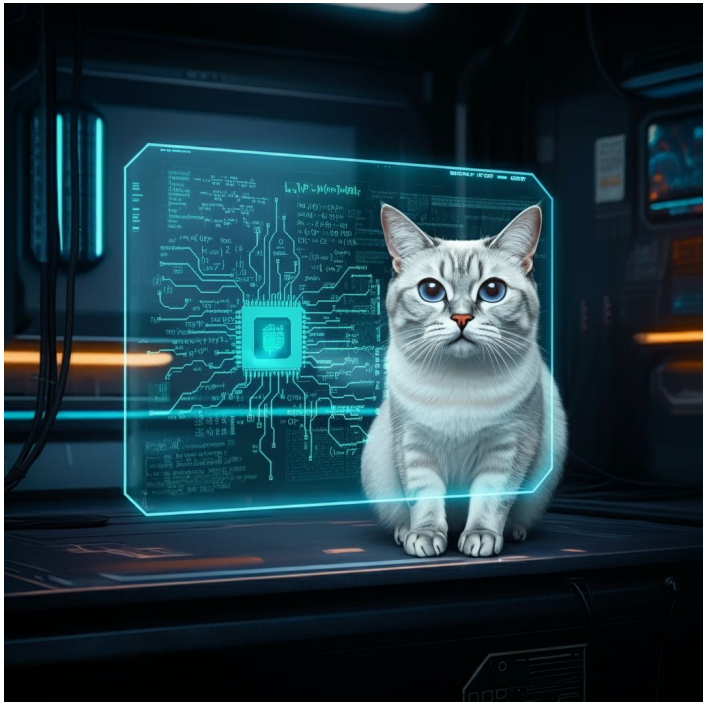
Address To Lat/Long

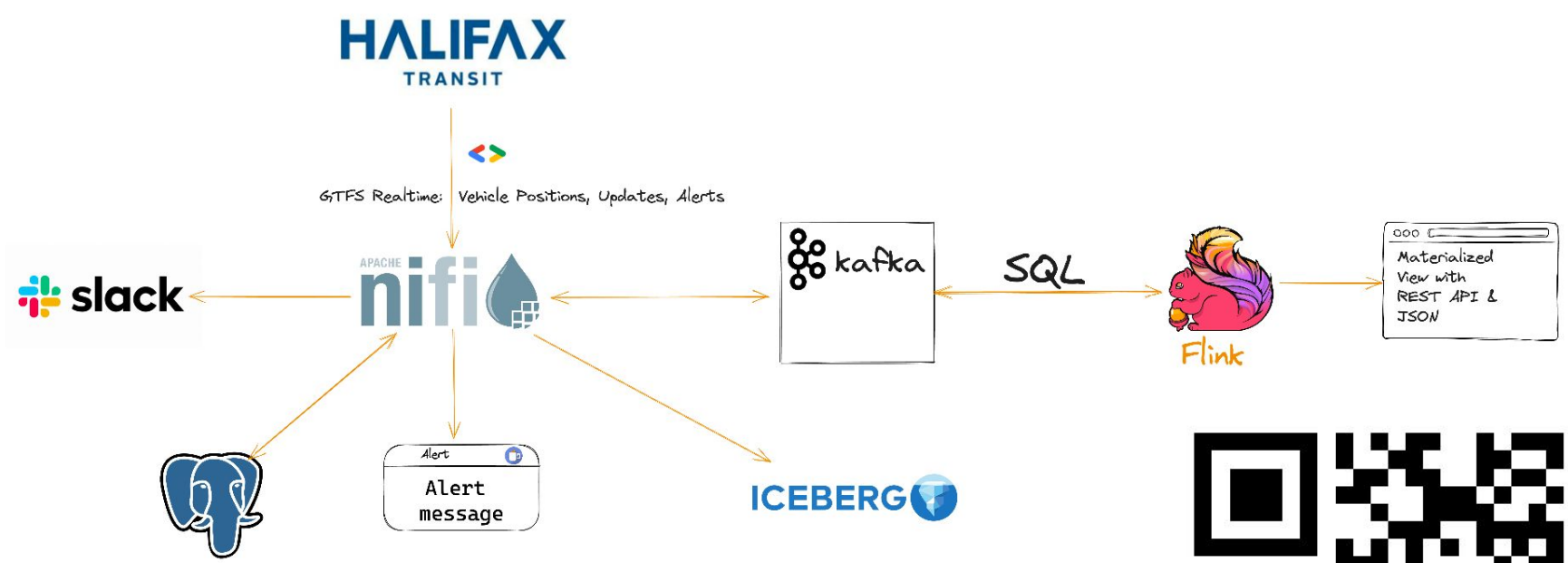
- Python 3.10+
- geopy Library
- Nominatim
- OpenStreetMaps (OSM)
- openstreetmap.org/copyright
- Returns as attributes and JSON file
- Works with partial addresses
- Categorizes location
- Bounding Box

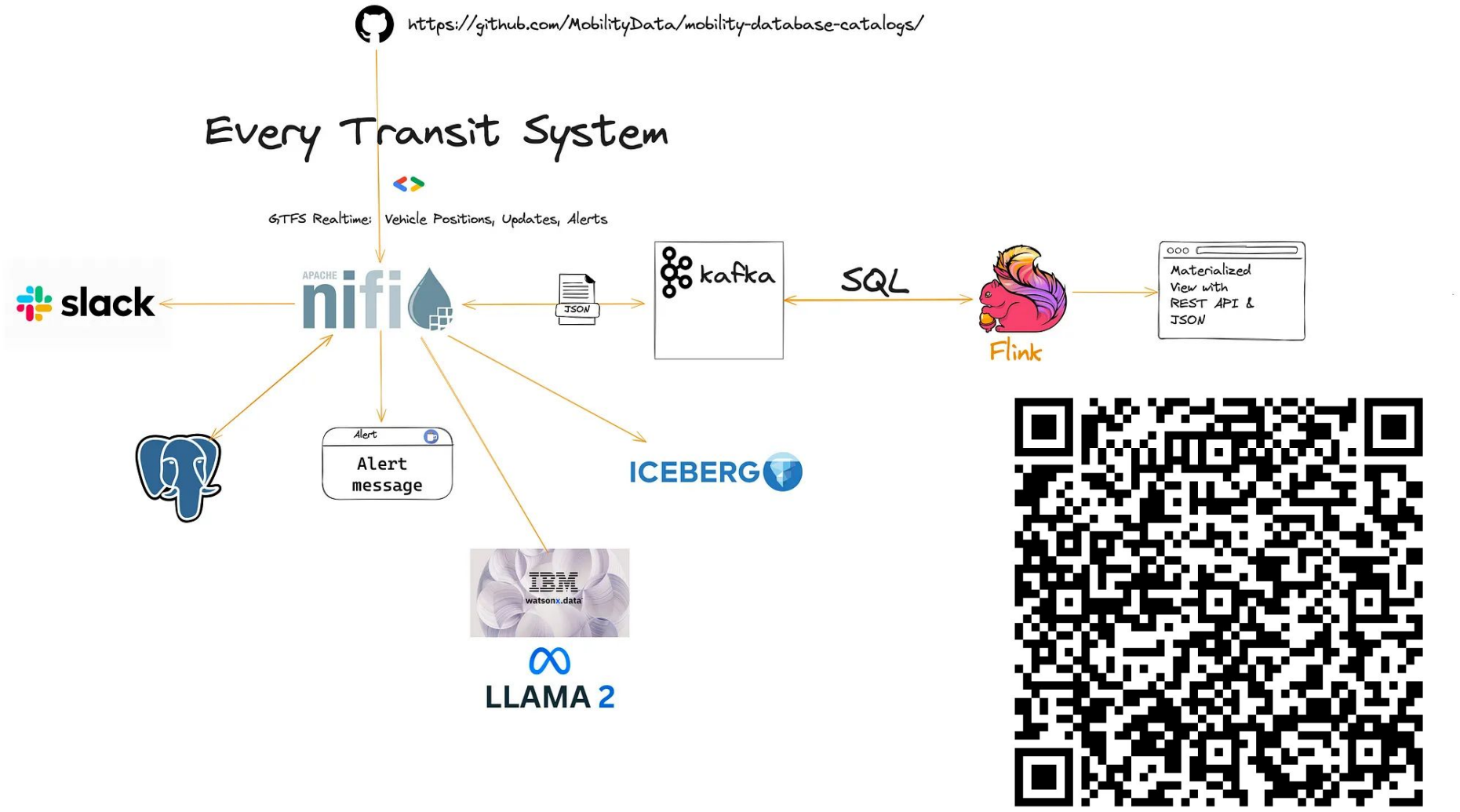
https://github.com/tspannhw/FLaNKAI-Boston

LET'S SEE

RESOURCES




HALIFAX
TRANSIT
GTFS Realtime: Vehicle Positions, Updates, Alerts
slack
APACHE
nifi
kafka
SQL
Flink
Materialized
View with
REST API &
JSON
Alert
Alert
message
ICEBERG


https://github.com/MobilityData/mobility-database-catalogs/
Every Transit System
GTFS Realtime: Vehicle Positions, Updates, Alerts
slack
APACHE nifi
JSON
kafka
SQL
Flink
Materialized View with REST API & JSON
Alert
Alert message
ICEBERG
IBM watsonx.data
LLAMA 2

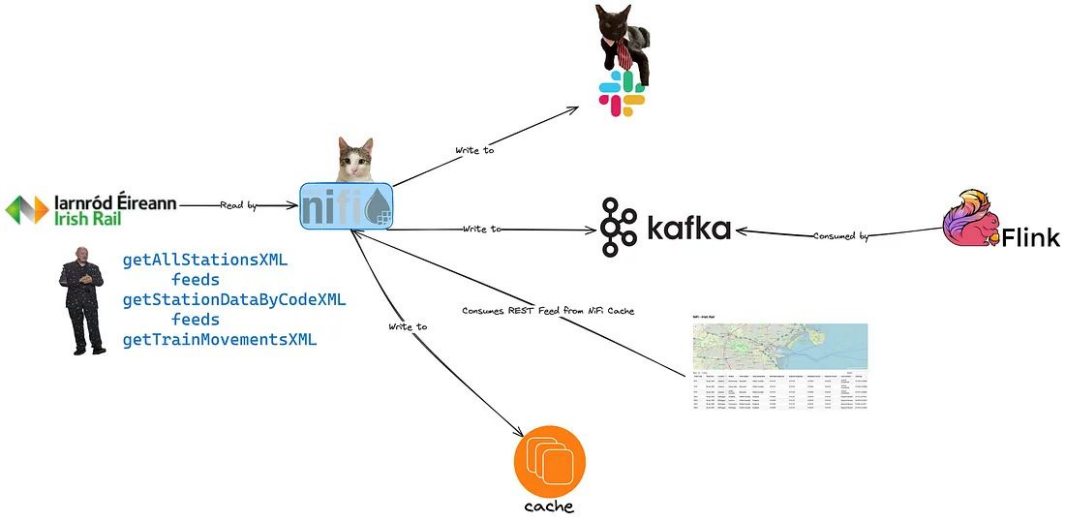
Irish Transit

https://medium.com/@tspann/real-time-irish-transit-analytics-ea76164c9595

Street Cameras

https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-v8-with-python-and-nifi-to-minio-s3-3277e73723ce

Irish Rail Example



https://medium.com/cloudera-inc/events-streams-flows-and-maps-22a8d27c9b4