Efficient Data Management for Intelligent Urban Mobility Systems

DI-CPS, CPS-IoT Week 2021

Michael Wilbur
Vanderbilt University

Philip Pugliese
Chattanooga Area
Regional Transportation Authority
(CARTA)

Aron Laszka
University of Houston

Abhishek Dubey
Vanderbilt University

Motivation

- Modern AI-driven urban mobility applications require working with large-scale, multivariate, spatiotemporal data streams.
- Current solutions typically involve an ad-hoc combination of open-source and proprietary technologies.

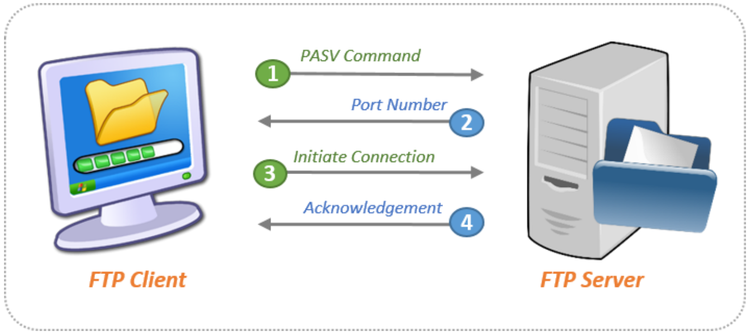
Overview

Integrated data management and processing framework for intelligent urban mobility systems currently in use by the Chattanooga Area Regional Transportation Agency (CARTA).

Motivation

Managing Datasets

- CSVs, JSON, GeoJSON
- Google Drive (Box ect.)
- On-premises SQL server
- FTP server

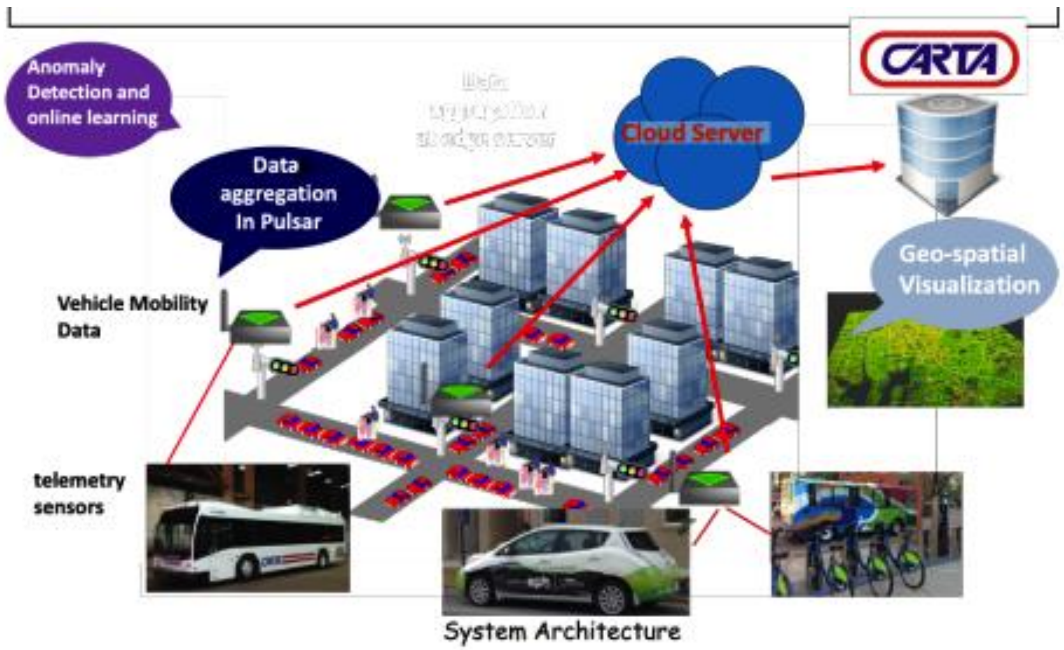


Managing APIs

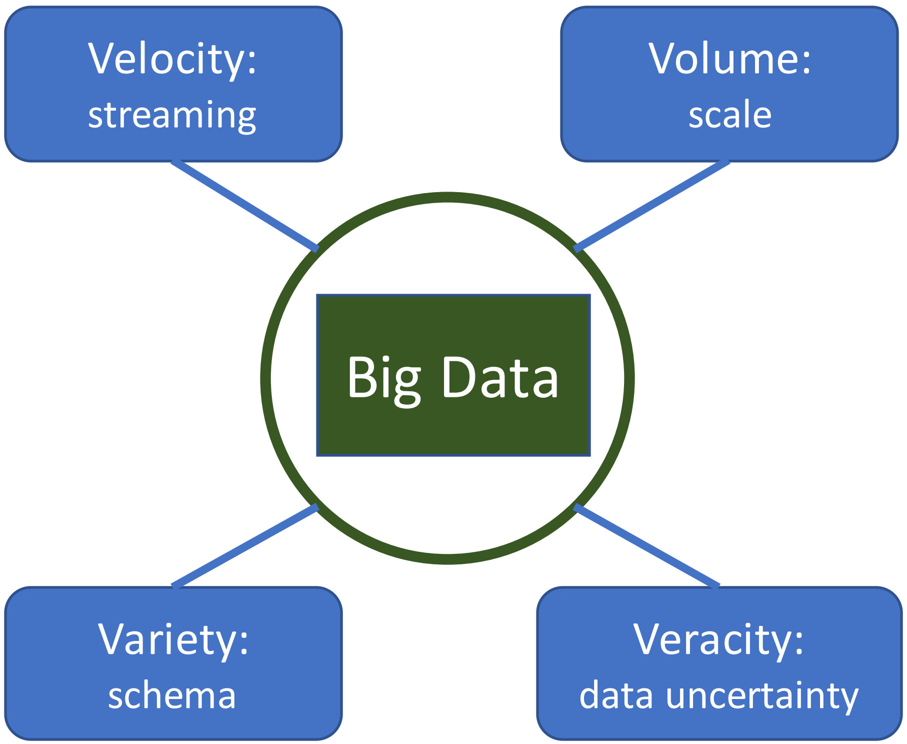
- REST, websockets
- Batch downloads
- Load into database

City-wide Data Collection

- 200 GB per month and growing
- A variety of APIs: REST, websockets
- Streaming, batch downloads and static data
- Multiple agencies: academic research groups, national labs, cities

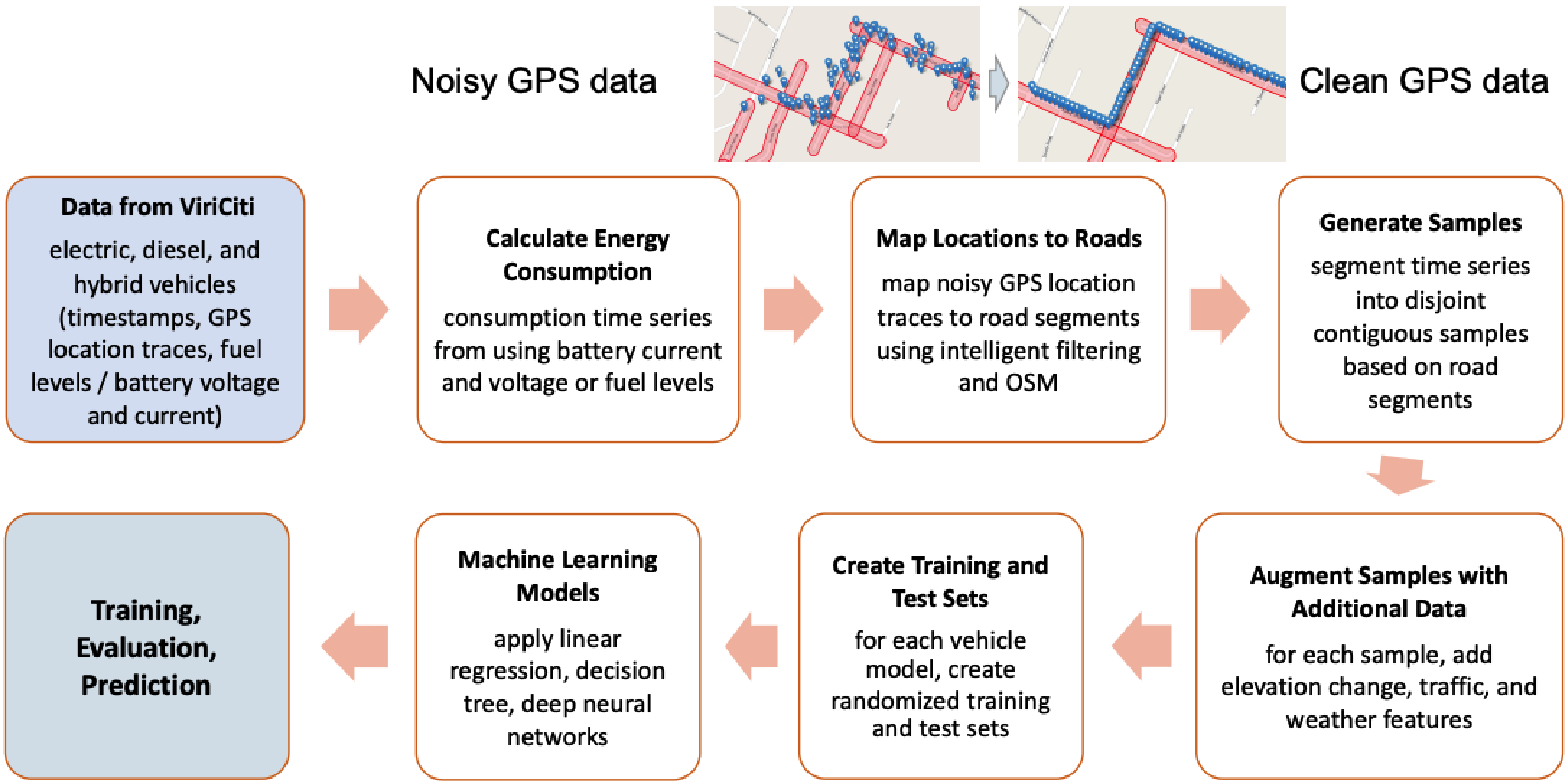


Data Challenges



Spatiotemporal Smart City Applications

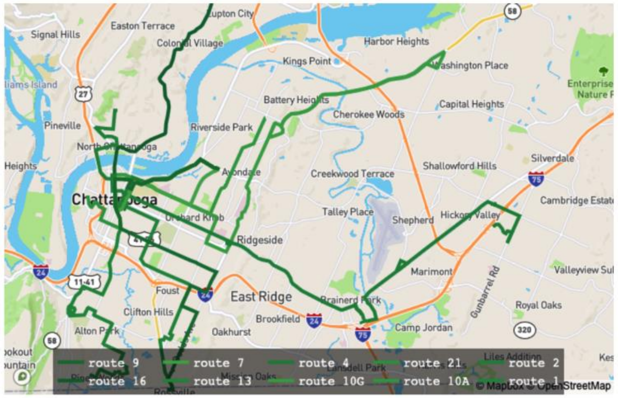
Motivating Example: Energy Prediction



1. Ayman, Afiya, et al. "Data-Driven Prediction of Route-Level Energy Use for Mixed-Vehicle Transit Fleets." *SmartComp* (2020).

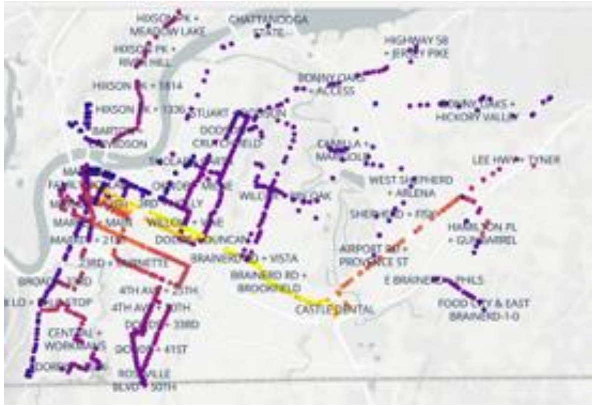
2. Sivagnanam, A. Ayman, M. Wilbur, P. Pugliese, A. Dubey, and A. Laszka, Minimizing Energy Use of Mixed-Fleet Public Transit for Fixed-Route Service, in *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

Applications



Vehicle Assignment: using the energy prediction models, assign EV, HV and ICEVs to bus transit trips to minimize cost/energy consumed

Paratransit dispatch: automated dispatching algorithms for paratransit



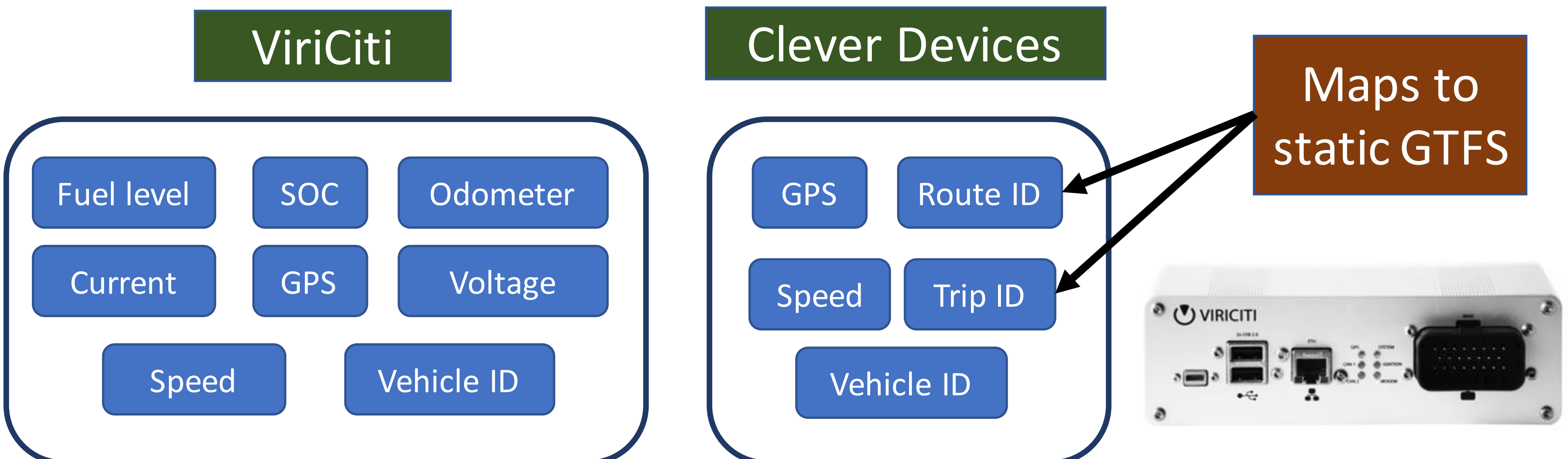
Flexible transit using predictive demand models

Problem Overview

- Data modeling and management
- Data synthesis and stream processing
- Efficient data retrieval
- Monitoring
- Presentation

Vehicle Telemetry and Service

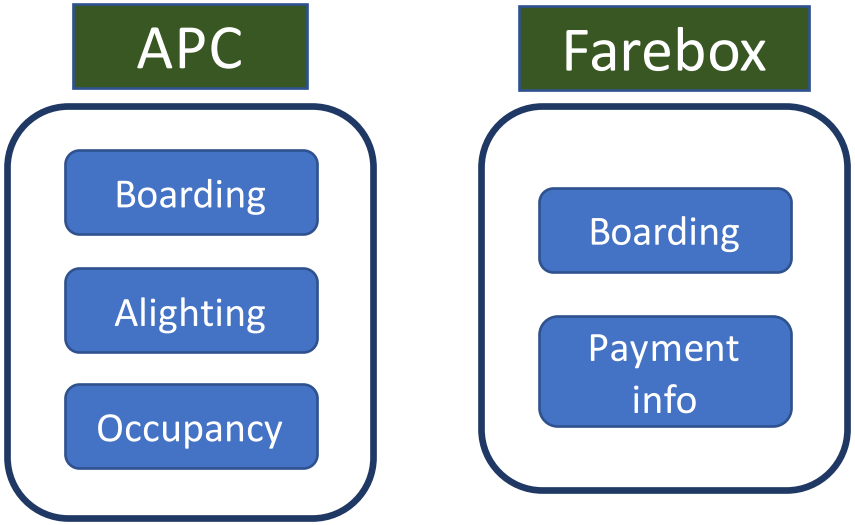
- CARTA manages a mixed fleet of 50 diesel, 3 electric and 7 hybrid vehicles.
- Vehicles are equipped with telemetry kits from our partners ViriCiti and Clever Devices.
- Data is available through websocket APIs at ~1 Hz.



ViriCiti DataHub Kit

Ridership Feeds

- Automated Passenger Counter (APC): infrared lights at doors track boardings + alightings.
- Farebox: from payments.
- Video feeds.



Farebox only includes boardings, however is much more accurate.

External Sources

- Weather: DarkSky
- Traffic: HERE and INRIX
- Road network: OpenStreetMap
- Elevation: Tennessee TNGIC
- Scheduling: static GTFS

State of the art

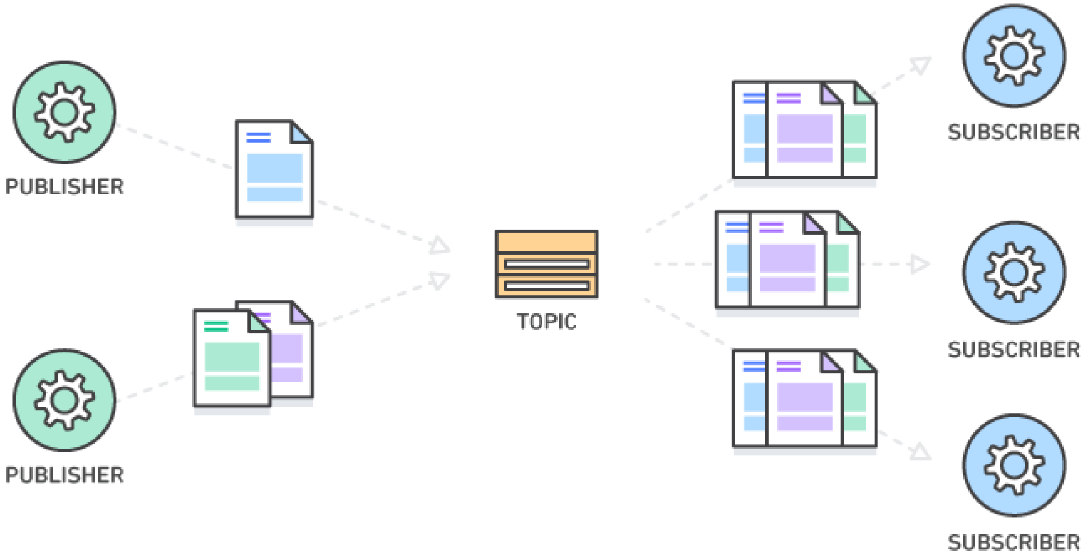
Data Store Solutions

- Pub-sub: Kafka, MQTT, Pulsar
- Stream processing: Apache Spark & Storm
- SQL
- NoSQL: MongoDB, Redis, Elasticsearch
- Hadoop/Spark
- GIS specific data stores: PostGIS, ESRI

AWS alone offers over 200+ services

Handling Real-time Data: Pub-Sub

- Publishers: typically python scripts that pull data from various APIs and write to a topic
- Subscribers: listen on topic.
  - Write to external databases
  - Stream processing



NoSQL: MongoDB

- Data is stored as JSON records
- Highly scalable
- General purpose: great for aggregation
- Easy to use and learn
- Native geospatial support

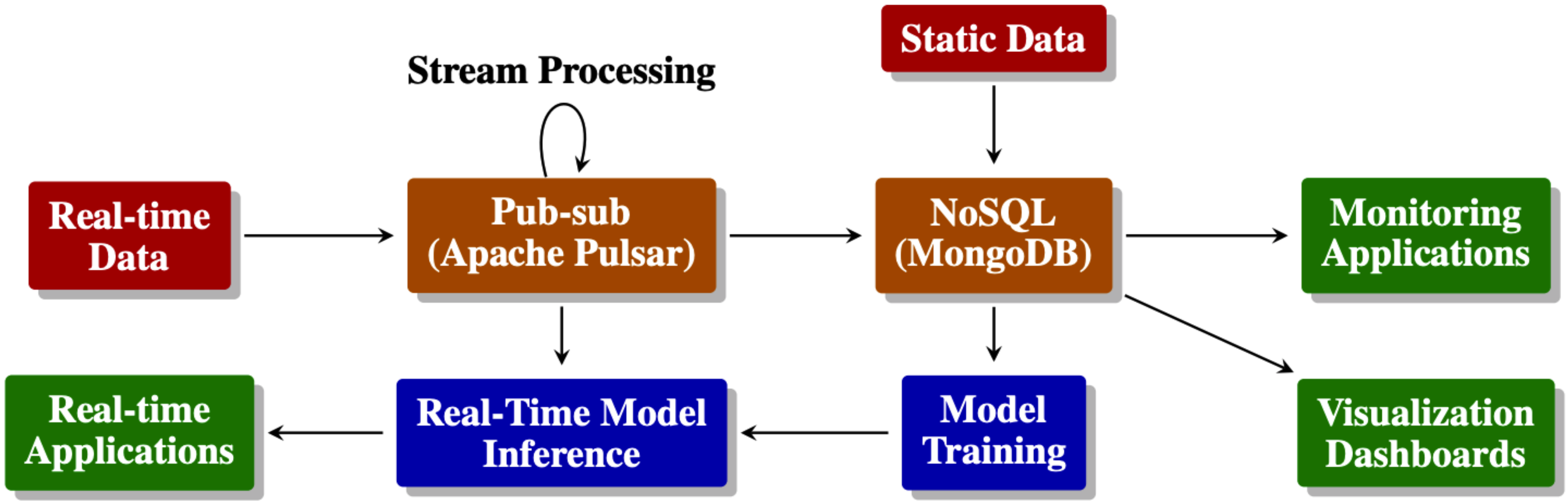

```
_id: ObjectId("60a3ea36b920e654e72f4964")
vid: "726"
tmstmp: "20210518 12:24"
lat: "35.043392874977805"
lon: "-85.3093490600586"
hdg: "168"
pid: "247"
rt: "33"
des: "SHUTTLE PARK SOUTH"
pdist: "5873"
oid: "160725"
rid: "1"
or: "false"
blk: "5002"
tripid: "179044020"
tripdyn: "0"
srvtmstmp: "20210518 12:24"
dly: "false"
spd: "11"
tablockid: "3301DTS"
tatripid: "179044"
origtatripno: "179044"
zone: null
mode: "0"
psgld: "HALF_EMPTY"
timestamp: 1621355040
⌄ geometry: Object
    type: "Point"
  ⌄ coordinates: Array
      0: "-85.3093490600586"
      1: "35.043392874977805"
```

S3/Athena

- S3: AWS cloud object storage
- Store static datasets
- Offload data from MongoDB
- Athena: SQL interface for querying data in S3

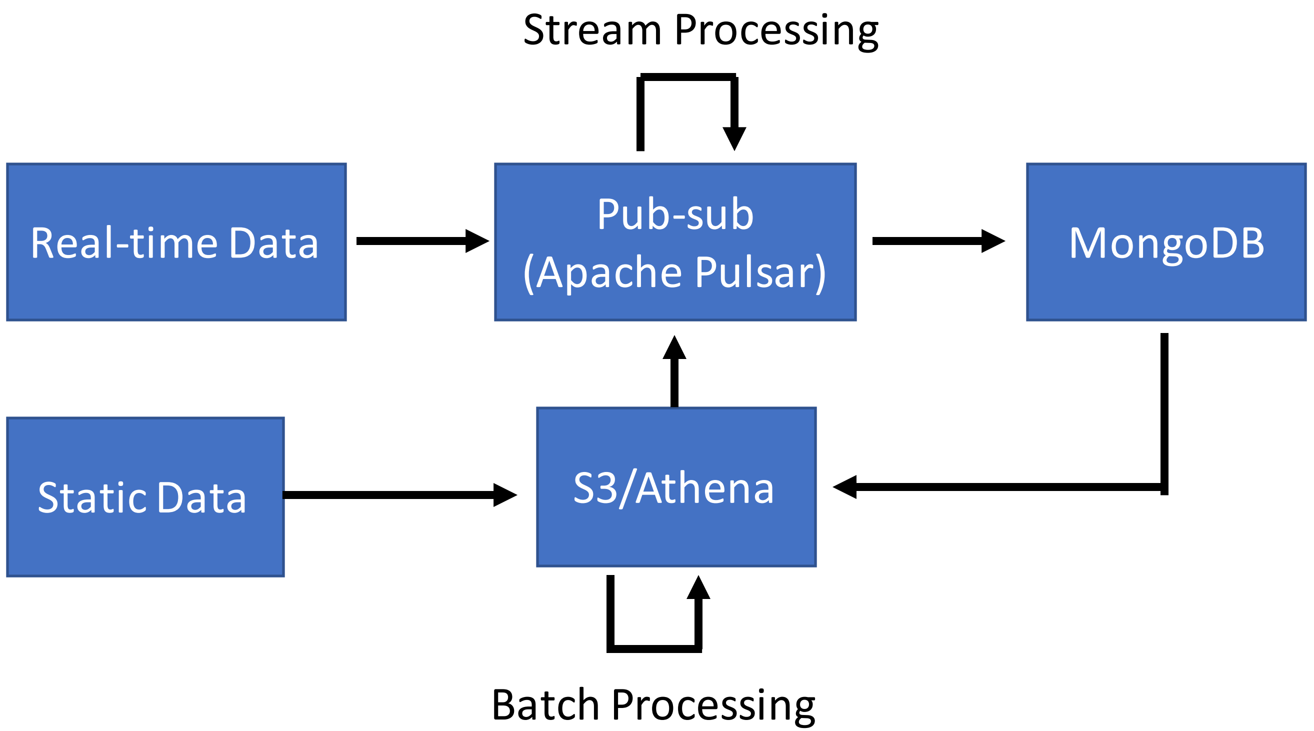
Data Architecture Overview

- Apache Pulsar – distributed topic-based pub-sub
- Topic naming: tenant/class/source
- MongoDB – document based NoSQL data store



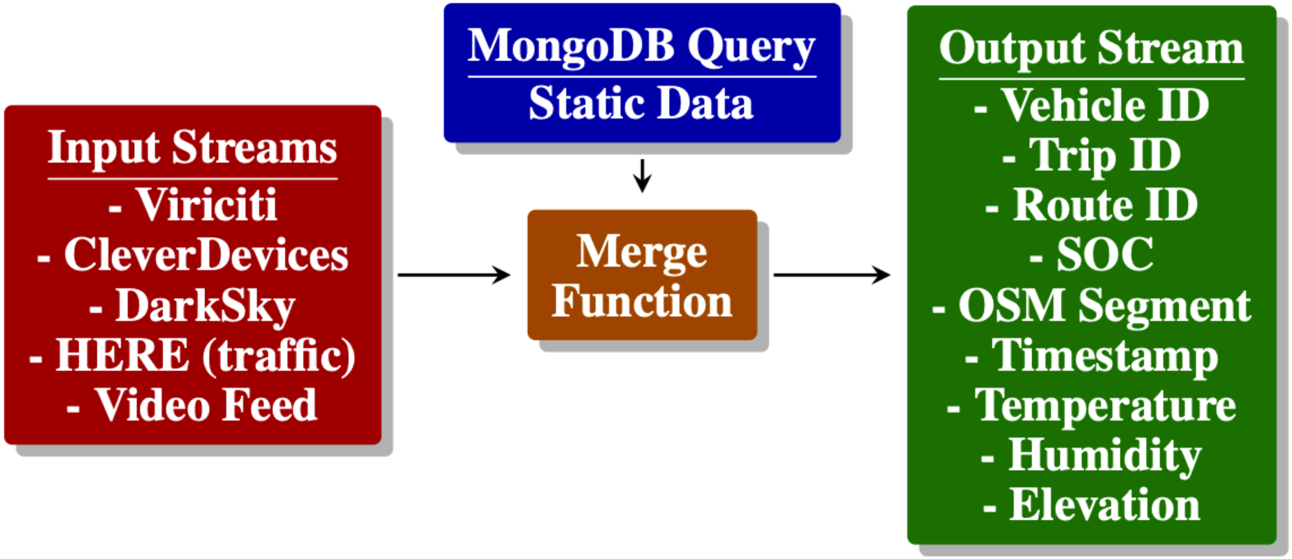
Data Architecture: updated

- S3: storage of static datasets
- Athena: SQL interface for querying data in S3



Data Synthesis and Stream Processing

- Stream processing implemented as Pulsar Functions.
- Real-time streams are merged with external sources such as geospatial features, GTFS schedules, OSM
- Example: real-time bus telemetry data streams are merged with weather, traffic and GTFS.



Data Synthesis – Map Matching

Mapping vehicles to OpenStreetMaps (1)

- OSRM map matching
- Google map matching API

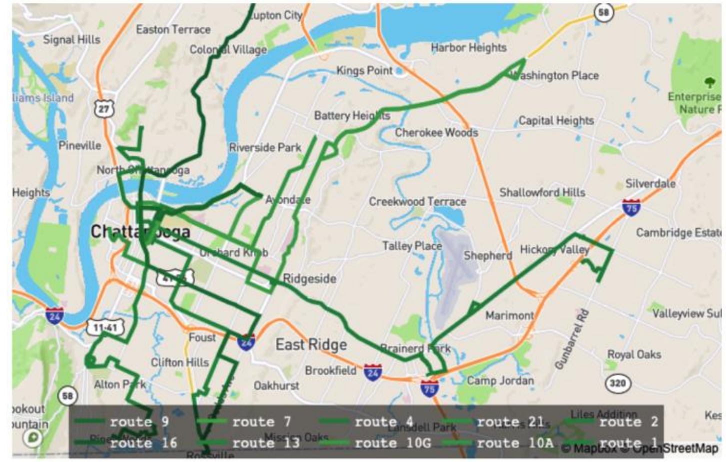
1. Ayman, Afiya, et al. "Data-Driven Prediction of Route-Level Energy Use for Mixed-Vehicle Transit Fleets." *SmartComp* (2020).

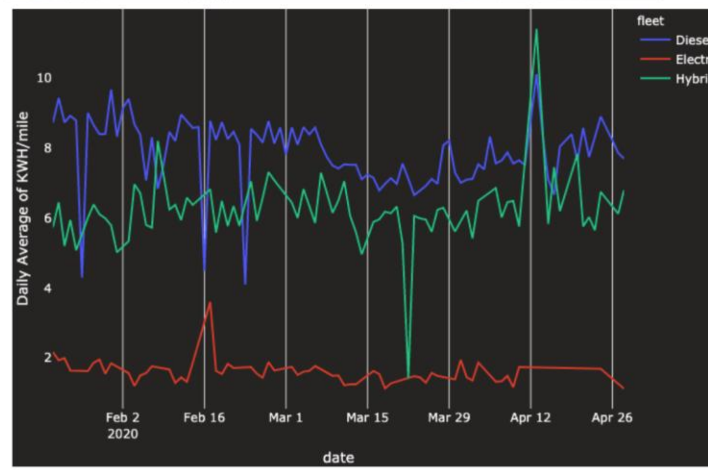
Visualization Dashboards

- Two dashboards: energy consumption and occupancy.
- Tools: Python, Plotly and Dash.
- Runs on Google Cloud App Engine.

Current dashboards provide access to historical performance

We would like to eventually include operational guidance tools





Example of energy consumption dashboard

Visualization Dashboards - Occupancy

Wilbur, Michael, et al. "Impact of COVID-19 on Public Transit Accessibility and Ridership." *arXiv preprint arXiv:2008.02413* (2020).

Email Monitoring System

- Runs nightly.
- Summarize state of the system.
- Compare number of new messages on each topic compared to historical daily averages.
- Flag anomalies.

Future Work

Machine Learning

- **Data analytics** of transit agencies
  - estimating ridership patterns
  - predicting ridership using machine learning models
- **Passenger guidance** application
  - suggestions on how to avoid crowded vehicles

Energy and Ridership Optimization

- **Proactive optimization** of fixed-route transit services
  - Maximize transit accessibility while minimizing crowding
  - Minimize energy use
- **On-demand prioritization and dispatch** for microtransit and paratransit services
  - Assign the calls to on-demand transit in anticipation of the fixed line schedule

Resources

- Overview of this project, the dashboards and published papers are available at: https://smarttransit.ai/