Utilizing Real-Time Transit Data for Travel Optimization

Tim Spann
Principal Developer Advocate

Sunday October 8, 2023
11:20 AM - 12:00 PM
Room 102
Tim Spann

@PaasDev  www.datainmotion.dev
github.com/tspannhw  medium.com/@tspann
Principal Developer Advocate

Princeton Future of Data Meetup
ex-Pivotal, ex-Hortonworks, ex-StreamNative,
ex-PwC, ex-EY, ex-HPE.

Apache NiFi x Apache Kafka x Apache Flink
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 streaming tables with Flink SQL, but also keep that same contract with Kafka consumers, Iceberg tables and other users of this data. I do not need to many user interfaces to interopt with the system as I want my final decision sent in a Slack message to me and then I’ll get moving. Along the way data will be visible in NiFi lineage, Kafka topic views, Flink SQL output, REST output and Iceberg tables.

Apache NiFi, Apache Kafka, Apache OpenNLP, Apache Tika, Apache Flink, Apache Avro, Apache Parquet, Apache Iceberg.

https://github.com/tspannhw/FLaNK-MTA/tree/main
https://medium.com/@tspann/finding-the-best-way-around-7491c76ca4cb
https://medium.com/@tspann/open-source-streaming-talks-in-progress-3e75af8848b0
https://medium.com/@tspann/watching-airport-traffic-in-real-time-32c522a6e386
AGENDA

- Introduction
- Overview
- Sources & Code
- Apache Kafka
- Apache Flink
- Apache NiFi
- Demos
DATA OVERLOAD
OUTGROWING ONE SERVER
USERS JUST WANTED ONE TACO
Q. What is contain it everything?
A. Wisdom.
# Trains, Planes and Automobiles

<table>
<thead>
<tr>
<th>Information Needed</th>
<th>Data Feed(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local weather conditions</td>
<td>XML, JSON, RSS</td>
</tr>
<tr>
<td>Mass transit status &amp; alerts</td>
<td>XML, JSON, RSS</td>
</tr>
<tr>
<td>Regional highways &amp; tunnels</td>
<td>GeoRSS, XML, ProtoBuf, JSON</td>
</tr>
<tr>
<td>Local social media</td>
<td>JSON</td>
</tr>
<tr>
<td>ADS-B Plane Data</td>
<td>JSON</td>
</tr>
<tr>
<td>Local air quality</td>
<td>JSON</td>
</tr>
</tbody>
</table>
REAL-TIME REQUIRES A PLATFORM
ALL THE TRANSIT DATA

FLANK STACK
INGEST OF ALL TRANSIT DATA

Run collection and streaming on any cloud, server, container or VM
FLaNK-MTA / Urban Transportation

Metropolitan Transportation Authority

APACHE
nifi

kafka
MTA
Topics

SQL
Apache Kafka

**DATA COLLECTION AT THE EDGE**

- **US-West Fleet**
  - C++ agent
  - Minifi
- **US-Central Fleet**
  - C++ agent
  - Minifi
- **US-East Fleet**
  - C++ agent
  - Minifi

**INGEST GATEWAY POWERED BY KAFKA**

- Gateway-West raw-sensors
- Gateway-Central raw-sensors
- Gateway-East raw-sensors

**DATA FLOW APPS POWERED BY NIFI**

**KAFKA CLUSTER GEO 1 DATA SYNDICATE SERVICES**

- Kafka Topic syndicate-transmission
- Kafka Topic syndicate-temp
- Kafka Topic syndicate-geo

**KAFKA CLUSTER GEO 2 DATA SYNDICATE SERVICES**

- Kafka Topic syndicate-transmission
- Kafka Topic syndicate-temp
- Kafka Topic syndicate-geo

**STREAMING ANALYTICS APPS**

- **Micro Batch Analytics**
  - Stream Analytics App
- **Micro Services**
  - Stream Analytics App
- **Complex Low Latent**
  - Stream Analytics App

**Apache Flink**

- Structured Streaming
- Apache Spark
- Apache Kafka
- Apache NiFi
- Apache Flink
I Can Haz Data?
Apache Flink SQL
Democratize access to real-time data with just SQL

```sql
CREATE TABLE `ssb`.`Meetups`.hfbloom (
  `generated_text` VARCHAR(2147483647),
  `ts` VARCHAR(2147483647),
  `x_compute_type` VARCHAR(2147483647),
  `inputs` VARCHAR(2147483647),
  `x_compute_time` VARCHAR(2147483647),
  `x_inference_time` VARCHAR(2147483647),
  `uuid` VARCHAR(2147483647),
  `x_time_per_token` VARCHAR(2147483647),
  `x_compute_characters` VARCHAR(2147483647),
  `eventTimeStmp` TIMESTAMP(3) WITH LOCAL TIME ZONE METADATA FROM 'timestamp',
  WATERMARK FOR `eventTimeStmp` AS `eventTimeStmp` - INTERVAL '3' SECOND
) WITH (  
  'scan.startup.mode' = 'group-offsets',
  'properties.request.timeout.ms' = '120000',
  'properties.auto.offset.reset' = 'earliest',
  'format' = 'json',
  'properties.bootstrap.servers' = 'kafka:9092',
  'connector' = 'kafka',
  'properties.transaction.timeout.ms' = '900000',
  'topic' = 'hfbloom',
  'properties.group.id' = 'llmBloomProps'
)
APACHE NIFI
RECORD-ORIENTED DATA WITH NIFI

- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML

- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML

- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.

- Enable processors that accept any data format without having to worry about the parsing and serialization logic.

- Allows us to keep FlowFiles larger, each consisting of multiple records, which results in far better performance.
PROVENANCE

- Tracks data at each point as it flows through the system
- Records, indexes, and makes events available for display
- Handles fan-in/fan-out, i.e. merging and splitting data
- View attributes and content at given points in time
Python as a First Class Citizen

[Insert code snippet here]

https://github.com/apache/nifi/blob/614947e4ac6798ad80817e82514c39349d5faacb/nifi-docs/src/main/asciidoc/python-developer-guide.adoc
Future of Data - NYC / Princeton + Virtual

https://www.meetup.com/futureofdata-princeton/
https://www.meetup.com/futureofdata-newyork/

From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...
FLaNK Stack Weekly

This week in Apache NiFi, Apache Flink, Apache Kafka, Apache Spark, Apache Iceberg, Python, Java, AI, ML, LLM and Open Source friends.

THANK YOU