



Utilizing Real-Time Transit Data for Travel Optimization

Tim Spann

SR SE, Snowflake



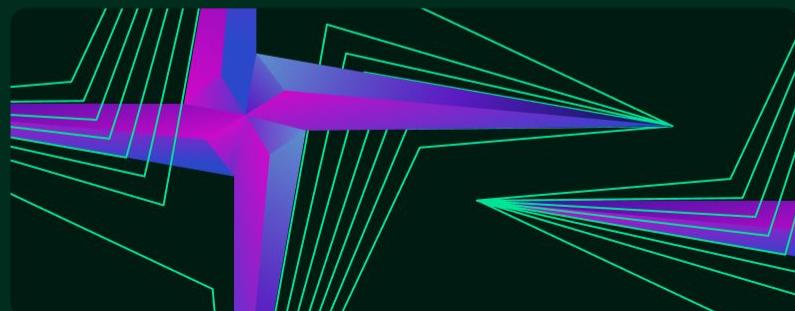
DATA
SCIENCE
SUMMIT

A lecture selected by a Program Council
consisting of recognized leaders
in the Data and AI field.

Warsaw,
20.11.2025 - 21.11.2025



OFFICIAL LECTURE OF THE DATA SCIENCE SUMMIT



Tim Spann

paasdev.bsky.social

@PaasDev // Blog: datainmotion.dev

Senior Solutions Engineer, Snowflake

NY/NJ/Philly - Cloud Data + AI Meetups

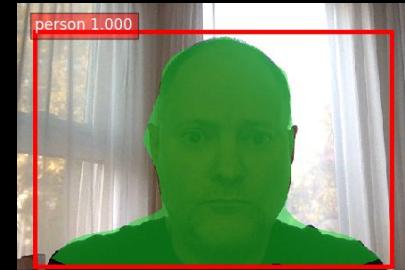
ex-Zilliz, ex-Pivotal, ex-Cloudera, ex-HPE,
ex-StreamNative, ex-Hortonworks.

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



DZone REFCARDS TREND REPORTS EXPERTS
Top IoT Experts

Tim Spann
Principal Developer Advocate, Cloudera
<https://github.com/tspannhw/SpeakerProfile/>
Tim Spann is a Principal Developer Advocate in Data in Motion for Cloudera. He works with Apache NiFi, Apache Pulsar, Apache...



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.**

The MTA Data Challenge

- ✓ The largest transit system in the US.
- ✓ Generates a 'staggering' amount of **data** every minute.
- ✓ Data is siloed: GTFS-RT for subways, SIRI for buses.
- ✓ Key data (like service status) can be in outdated, hard-to-parse XML formats.
- ✓ Internal tools often run reports for the *next day*, not in real-time.



stockcake.com

Our Objective: From "Next Day" to "Right Now"

1

Single, Real-Time View

The Goal

To build a predictive system that processes thousands of real-time data points, avoids delays, and provides the *truly* optimal route, right now.

We want to move from historical analysis to real-time, actionable decisions.



Fueling the System: Data Sources



MTA GTFS-RT

Real-time subway vehicle positions, trip updates, and service alerts.



MTA SIRI

Real-time bus locations, estimated times of arrival, and service status.



External Data

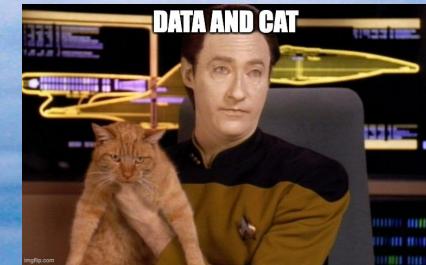
Real-time weather APIs, NYC traffic cameras, news, search, public safety alerts, and city event schedules.

General (Google) Transit Feed Specification

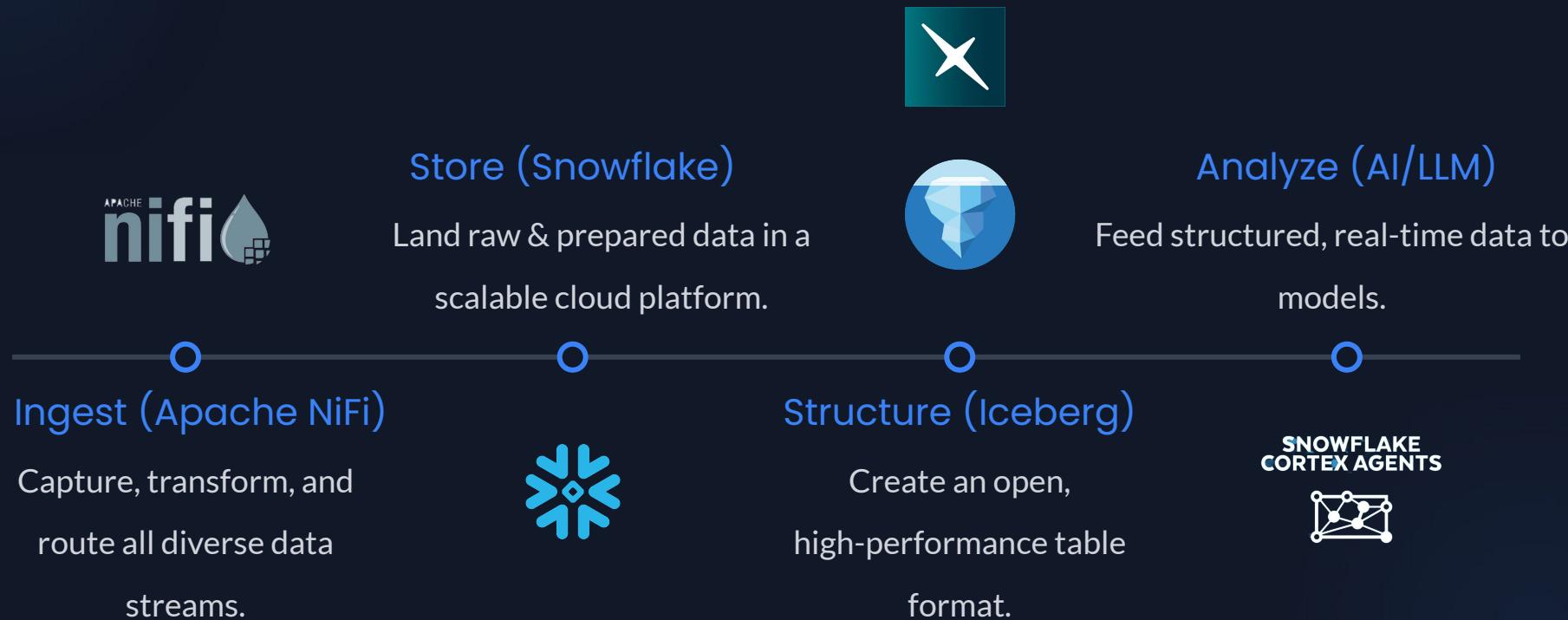
GTFS - Protocol Buffers (binary format)

Organizations:

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



Transit Real-Time Data Pipeline



Real-Time AI Open Lakehouse with Open Data



Step 1: Ingest with NiFi

The Data Flow "Swiss Army Knife"

Apache NiFi is the key to taming this chaos.

- Visually connects to any source (REST, Kafka, Protobuf, XML...).
- Perfect for parsing messy, outdated formats like RSS and XML
- Guarantees data delivery & full data provenance.



Step 2: Store & Structure



Snowflake

The cloud **data** platform. It decouples storage & compute, handling massive scale & concurrency for our **AI** queries.

Apache Iceberg

The open table format. **Apache Iceberg** enables high-performance analytics on the data lake and provides time-travel & schema evolution for our models.



Step 3: The Brain (AI & GenAI)



Predictive AI: Models (like XGBoost) forecast delays based on current crowding, weather, and known signal issues.

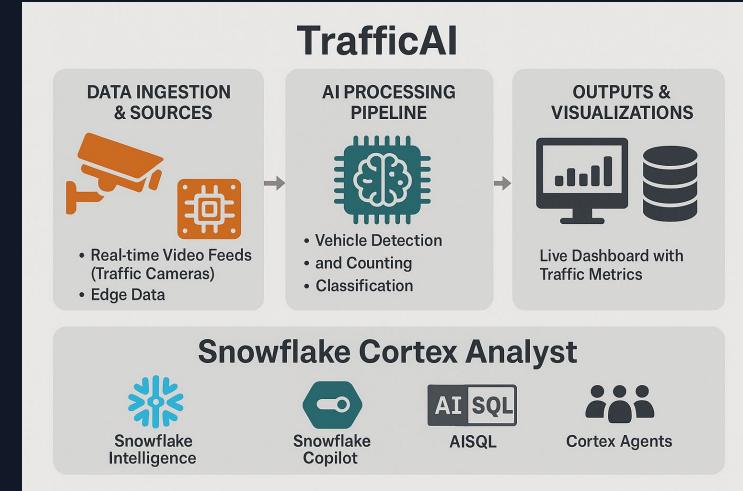


Generative AI (LLMs): The LLM translates complex model outputs ("Delay_Prob_A=0.85") into simple, natural language advice.



Real-Time Decisions: The entire system runs on a continuous flow of **data**, allowing the **GenAI** to update its recommendations at any moment.

Cortex AI SQL



```
SELECT SNOWFLAKE.CORTEX.COMPLETE("claude-4-5-sonnet",
    "Analyze this traffic image and describe what you see.
    Respond in compact JSON format.",
    TO_FILE("@TRAFFIC", "cars.jpg")) INTO :result;
```

Easy, Connected, Trusted



SNOWFLAKE CORTEX AI

AGENTIC BUSINESS INSIGHT

Agent Apps
Snowflake Intelligence PU Soon

Agent APIs
Cortex Agents GA Soon

STATE OF THE ART RETRIEVAL

Unstructured Data Retrieval
Cortex Search

Structured Data Retrieval
Cortex Analyst

SCALABLE AI PROCESSING

Multimodal AI-powered SQL
Cortex AISQL PU

Document Processing
Document AI, Parse, Embed

MODELS

OpenAI

ANTHROPIC

Meta

MISTRAL AI_

deepseek

snowflake®

GOVERNANCE

RBAC

Guardrails

Evaluations

Monitoring

AI Gateway



DATA



DATA INGESTION

Fully managed extraction,
load, preprocess, and
activation of multimodal data



© 2023



Putting It All Together

Avoid the 2 train. Based on current signal problems at 72nd St and platform overcrowding at Times Square, I recommend taking the Q train to 34th St. This will save you an estimated 15 minutes.

– Your Real-Time Travel Assistant

Streamlit Apps New York City Map

Packages ▾

```

1 # Import python packages
2 import streamlit as st
3 from snowflake.snowpark.context import get_active_session
4 from snowflake.snowpark.functions import *
5 from snowflake.snowpark.types import *
6 import json
7 import pandas as pd
8 import numpy as np
9 import pydeck as pdk
10 from snowflake.snowpark.files import SnowflakeFile
11 from snowflake.snowpark.functions import col
12 from streamlit.components.v1 import html
13
14 st.set_page_config(layout="wide")
15
16 st.title(f"New York City")
17
18 session = get_active_session()
19
20 traffic_latlon = session.table('DEMO.DEMO.VWNYCTRAFFIC')
21
22 st.markdown('#### Traffic Station Locations')
23
24 # Convert to pandas DataFrame
25 df = traffic_latlon.to_pandas()
26
27 # Create PyDeck layer
28 layer = pdk.Layer(
29     "ScatterplotLayer",
30     df,
31     get_position=["LONGITUDE", "LATITUDE"],
32     get_color=[255, 0, 0, 160],
33     get_radius=100,
34     pickable=True
35 )
36
37 # Create view state centered on NYC
38 view_state = pdk.ViewState(
39     latitude=40.7128,
40     longitude=-74.0060,
41     zoom=10,
42     pitch=0
43 )
44
45 # Create deck with view state
46 deck = pdk.Deck(
47     layers=[layer],
48     initial_view_state=view_state,
49     map_style="mapbox://styles/mapbox/light-v9",

```



▼ {

- "confidence_level" : "high"
- "total_count" : 12
- ▼ "vehicle_types" : [
 - 0 : "cars"
 - 1 : "SUVs"
 - 2 : "pickup trucks"
]

}

I-87 NB 15.5 GMCB



▼ {

- "confidence_level" : "high"
- "total_count" : 8
- ▼ "vehicle_types" : [
 - 0 : "passenger cars"
 - 1 : "SUV"
 - 2 : "pickup truck"
]

}

I-87 NB 14.9a GMCB



Model Context Protocol

The image displays three screenshots of a software interface, likely a browser-based application, demonstrating the Model Context Protocol (MCP). The background features a blurred view of a server room with multiple server racks.

Screenshot 1: MCP Tools Overview

This screenshot shows the MCP Tools section of the interface. It includes:

- A sidebar with a user profile (bunkertor@gmail.com, Pro Trial) and navigation links: General, oo Chat, oo Tab, Models, Rules, Indexing & Docs, Network, Beta, and Docs.
- A "Tools & Integrations" section with "Integrations" for GitHub (Connect) and Slack (Connect).
- An "MCP Tools" section showing "mcp-server-snowflake" (4 tools enabled) and a "New MCP Server" button.

Screenshot 2: MCP API Requests and Traffic Data

This screenshot shows the MCP API requests and traffic data. It includes:

- A "Retrieve camera data from 5th Avenue" request with details:
 - General: Multi-lane city avenue, commercial buildings, retail stores, sidewalks with pedestrians, clear/sunny, moderate congestion, cars, taxis, commercial vehicles.
 - Notable: Bike lanes, crosswalks, digital billboards (Times Square), street trees, typical NYC urban canyon effect, yellow taxis, and moderate to light traffic.
- A note: "If you want full details for a specific intersection or a particular time, let me know!"
- A "mcp.json" request: "Get the details from 9 ave".
- Callouts for "Called contexttraffic" and "Called TRAFFICIMAGESEARCH".
- A summary: "Here are the latest details from traffic cameras on 9th Avenue in New York:"
- Details for 1. 207 Street @ 9 Avenue:
 - Timestamp: 2025-06-23 12:40:08 PM
 - Direction: Facing East
 - Traffic: Heavy congestion, multiple lanes with backed up traffic, slow moving, cars and trucks
 - Weather: Clear, good visibility
 - Image: !207 St @ 9 Ave
- Details for 2. 9 Avenue @ 23 Street:
 - Timestamp: 2025-06-23 11:45:55 AM
 - Direction: Facing East
 - Environment: Daylight, street parking both sides, tree-lined, clear weather
 - Road: 2 lanes, clear and dry, yellow divider, asphalt
 - Traffic: Light, large truck and several parked cars
 - Image: !9 Ave @ 23 St
- Details for 3. 9 Avenue @ 57 Street:
 - Timestamp: 2025-06-23 11:44:58 AM
 - Direction: Facing North
 - Environment: Mix of red brick and modern buildings, tall residential/commercial, wide sidewalks, some trees, clear/sunny
 - Traffic: Light to moderate, multiple vehicles, yellow taxi, wide zebra crossing
 - Image: !9 Ave @ 57 St
- Details for 4. 9 Avenue @ 49 Street:
 - Timestamp: 2025-06-23 11:44:58 AM
 - Direction: Facing North
 - Environment: Mix of red brick and modern buildings, tall residential/commercial, wide sidewalks, some trees, clear/sunny
 - Traffic: Light to moderate, multiple vehicles, yellow taxi, wide zebra crossing
 - Image: !9 Ave @ 49 St

Screenshot 3: MCP Agent Interface

This screenshot shows the MCP Agent interface, featuring:

- A "mcp.json" entry: "Plan, search, build anything".
- A "Agent ID" dropdown set to "Auto".

Key Takeaways



Apache NiFi is Key

Apache NiFi is essential for taming diverse, messy, and real-time sources.

Open Foundations

Snowflake & Apache Iceberg provide a scalable, open foundation for AI.



AI Makes Data Actionable

AI/LLMs turn raw numbers into simple, human-readable advice.



Questions?

<https://github.com/tspannhw/TrafficAI>

DEMO

IS THIS ENOUGH DATA?



RESOURCES





<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-Beginning/t6yz-b64h/about_data

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://busitime-beta.mta.info/developers/siri/#>

<https://busitime-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/>

<https://medium.com/@tspann/automating-transit-gtfs-flows-7a5c4d3dafe8>

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

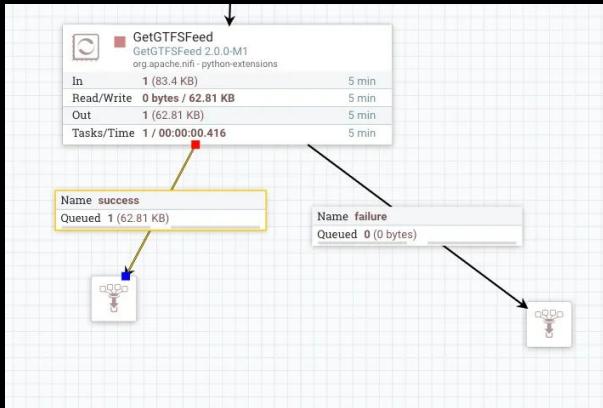
<https://medium.com/cloudera-inc/transit-in-sao-paulo-brasil-flank-style-eaec6753cc63>

<https://medium.com/cloudera-inc/flank-for-halifax-canada-transit-2d4dc5b8ad64>



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

trip_update

trip_update

vehicle

vehicle

alert

Reference parameter...

Processor Details | GetGTFSCompoundFeed 2.0.0-M2

Running

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMM

Required field

Property	Value
URL for GTFS Feed	Sensitive value set
API Key for header (MTA)	Sensitive value set
API Key for header name ex: (MTA)	x-api-key
Type for GTFS Feed	vehicle

<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



Thank you for watching!

Remember to leave your questions and rate the presentation in the section below.



DATA
SCIENCE
SUMMIT

A lecture selected by a Program Council consisting of recognized leaders in the Data and AI field.

Warsaw,
20.11.2025 - 21.11.2025



OFFICIAL LECTURE OF THE DATA SCIENCE SUMMIT

ACADEMIC
PARTNERS