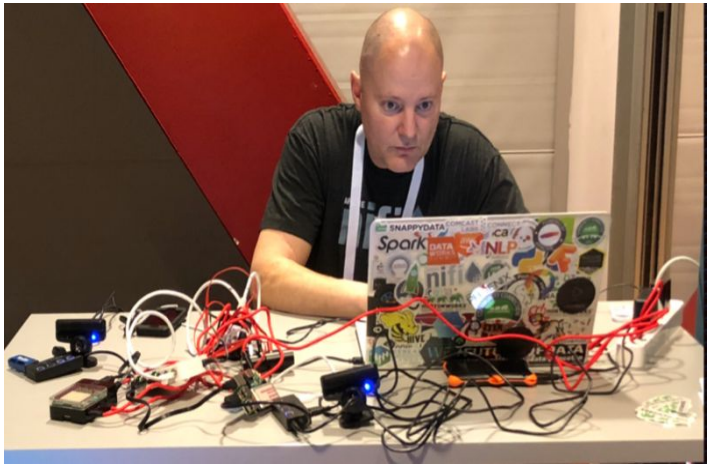
Building Secure RAG Applications with Open Large Language Models

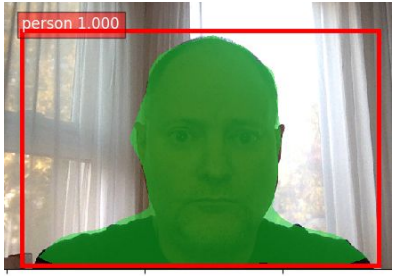
Tim Spann, Senior Solutions Engineer

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-EY, ex-Hortonworks.

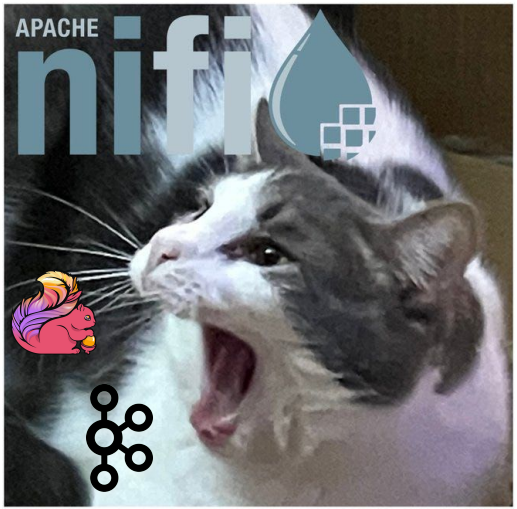
https://medium.com/@tspann
https://github.com/tspannhw



snowflake

Streamlit

AI + Streaming Weekly by Tim Spann



https://bit.ly/32dAJft

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

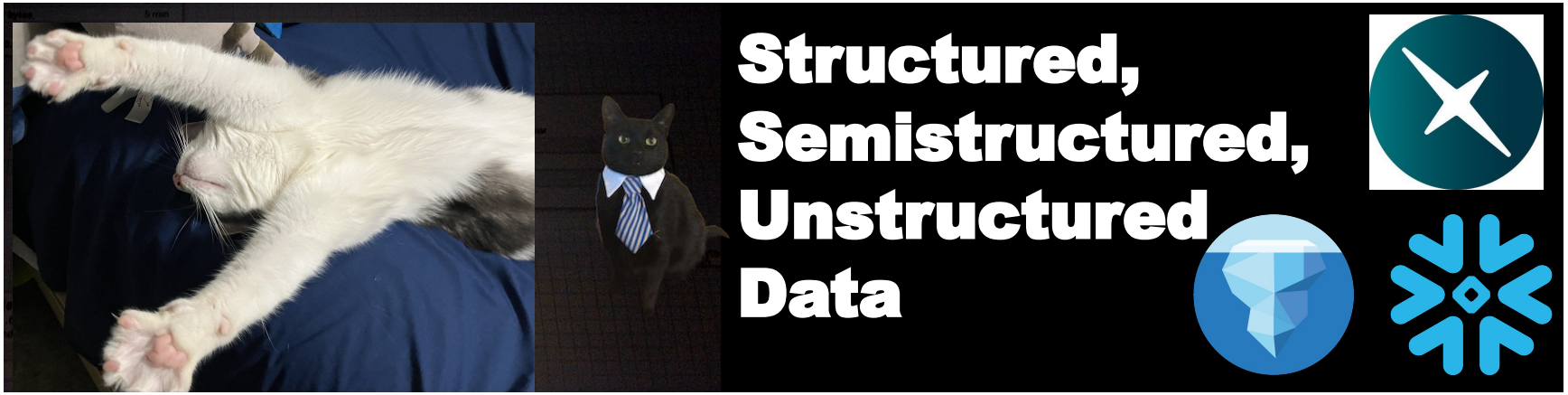
Introduction and Overview

Data

Demo

Resources

Building Secure RAG Apps Requires a Team For Data



When you think of **RAG**, you think of unstructured data like **documents** or giant chunks of **text**.

Unstructured Data

- Lots of formats
- Text, Documents, PDF
- Images, Videos, Audio
- Email
- Variants

Unstructured

Semi-Structured Data

- Open Data like Open AQ - Air Quality Data
- Location, Time,Sensors
- Apache Avro, Parquet, Orc
- JSON and XML
- Hierarchical Data
- Logs
- Key-Value

https://docs.snowflake.com/en/sql-reference/data-types-semistructured

Semi-structured

Structured Data

- Snowflake Tables
- Snowflake Hybrid Tables
- Apache Iceberg Tables
- Relational Tables
- Postgresql Tables
- CSV, TSV

Apache Iceberg™ - Append

- NiFi - PutIcebergTable
- Snowpark -
  df.write.mode("append").
  save_as_table("atable_iceberg")

https://quickstarts.snowflake.com/guide/getting_started_iceberg_tables/

Open Large Language Models

Snowflake Arctic Instruct

https://huggingface.co/Snowflake/snowflake-arctic-instruct

Snowflake's Arctic-embed-m-v2.0

https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v2.0

Llama-3.3-70b, mixtral-8x7b, llama3.1-405b, mistral-7b

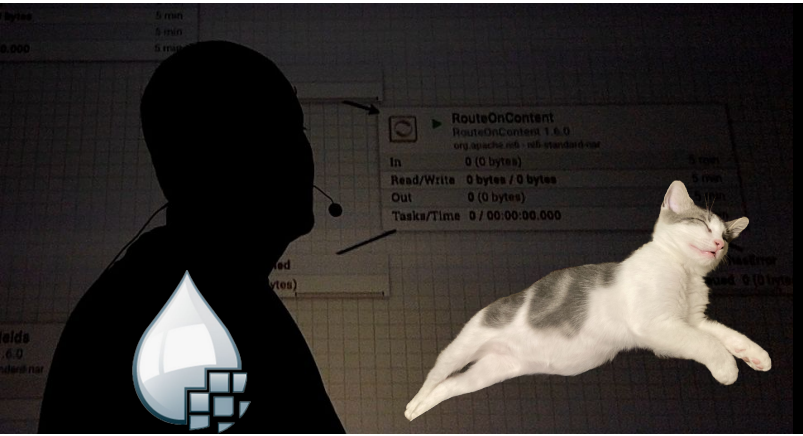
Retrieval Augmented Generation (RAG)

Build

Ingest -> Extract -> Split -> Build Indexes

Serve

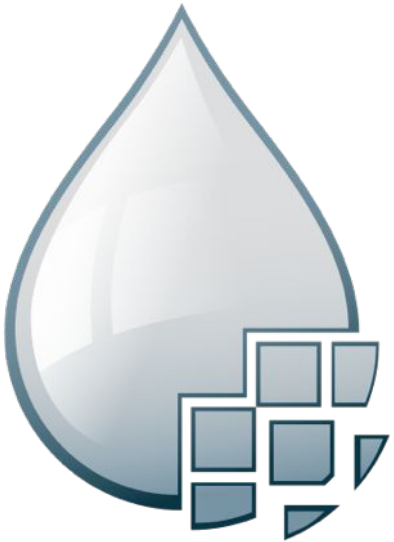
Orchestration | Observability <-> Retrieval <-> Inference



Open Source Option
Apache NiFi

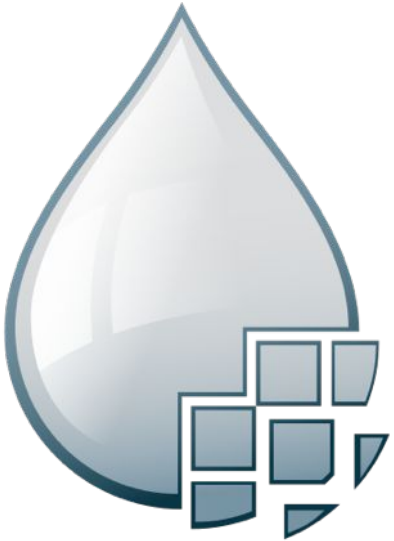
Build
Ingest, Extract, Split, LLM

Apache NiFi for Data Ingest, Movement and Routing



- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models
- Hundreds of processors
- Visual command and control
- Hundreds of sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

The Power of Apache NiFi



- Moving Binary, Unstructured, Image and Tabular Data
- Enrichment
- Universal Visual Processor
- Simple Event Processor
- Routing
- Feeding data to Central Messaging
- Support for modern protocols
- Kafka Protocol Source/Sink
- Pulsar Protocol Source/Sink

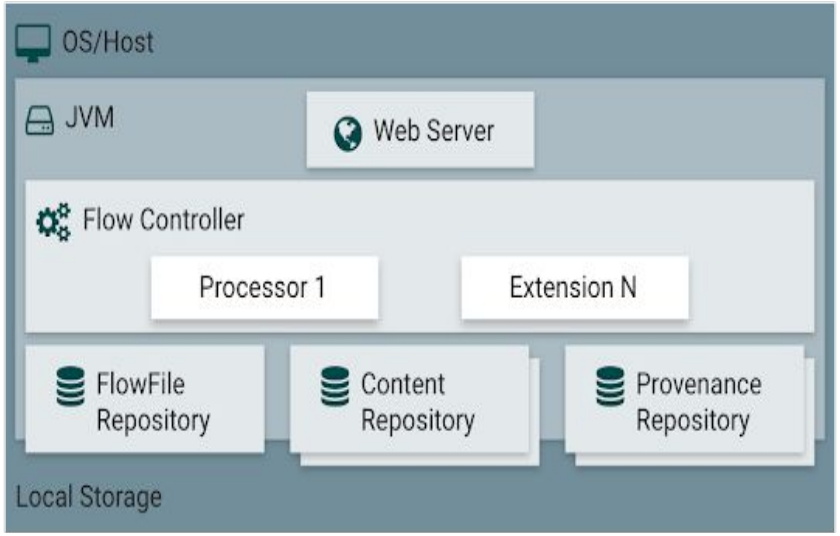
APACHE NIFI 2.0 FEATURES

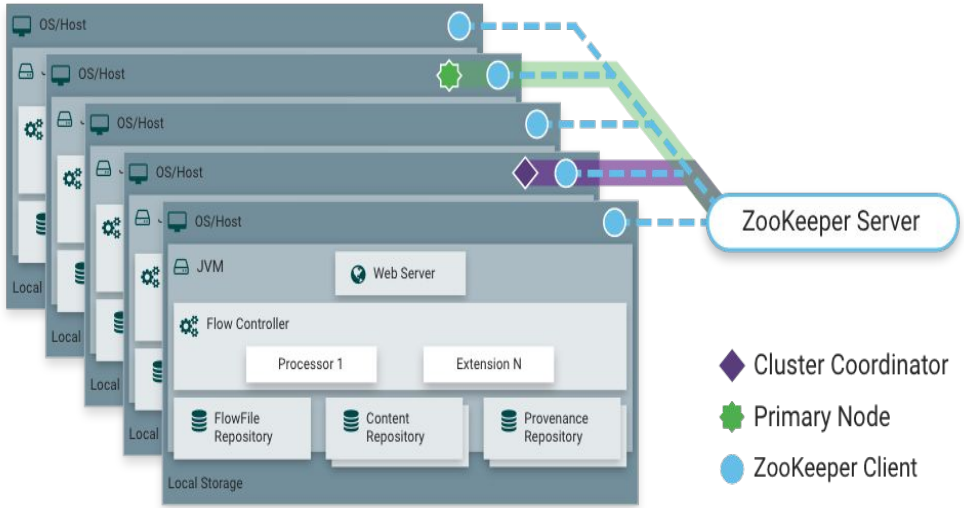
Real-Time Integration and AI

Major Updates:

- Python Integration
- Parameterization
- JDK 21+
- Provenance / Data Lineage
- Rules Engine for Development Assistance
- Additional Azure Processors
- Integration with Zendesk, Slack,
- Database Tables as Schemas
- Amazon Glue Schema Registry
- OpenTelemetry Support

Architecture





https://nifi.apache.org/docs/nifi-docs/html/overview.html

PROVENANCE

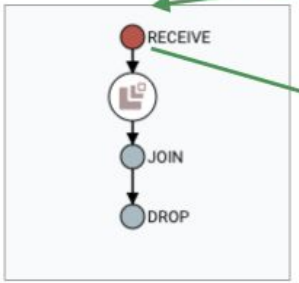
Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST

Showing the most recent events.

ConsumeKafka | by component name

Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type
11/15/2016 13:35:03.8...	RECEIVE	379fc4f6-60e0-4151-9743-28...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:02.7...	RECEIVE	78f8c38b-89fc-4d00-a8d8-51...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:01.6...	RECEIVE	2bcd5124-bb78-489f-ad8a-7...	44 bytes	ConsumeKafka	ConsumeKafka

- 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



Provenance Event

DETAILS | ATTRIBUTES | CONTENT

Attribute Values

filename
328717796819631
No value previously set

kafka.offset
44815
No value previously set

kafka.partition
6
No value previously set

kafka.topic
nifi-testing
No value previously set

path
./
No value previously set

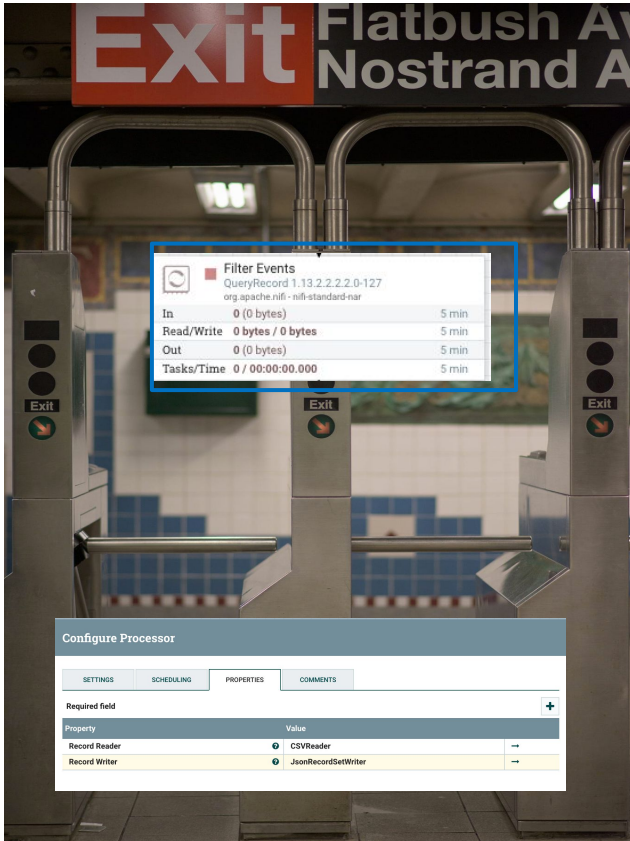
uuid

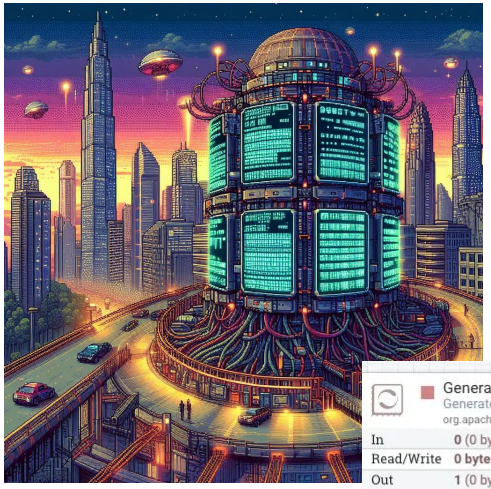
UNSTRUCTURED DATA WITH NIFI

- **Archives** - tar, gzipped, zipped, …
- **Images** - PNG, JPG, GIF, BMP, …
- **Documents** - HTML, Markdown, RSS, PDF, Doc, RTF, Plain Text, …
- **Videos** - MP4, Clips, Mov, Youtube URL…
- **Sound** - MP3, …
- **Social / Chat** - Slack, Discord, Twitter, REST, Email, …
- **Identify Mime Types, Chunk Documents, Store to Vector Database**
- **Parse Documents -** HTML, Markdown, PDF, Word, Excel, Powerpoint

RECORD-ORIENTED DATA WITH NIFI

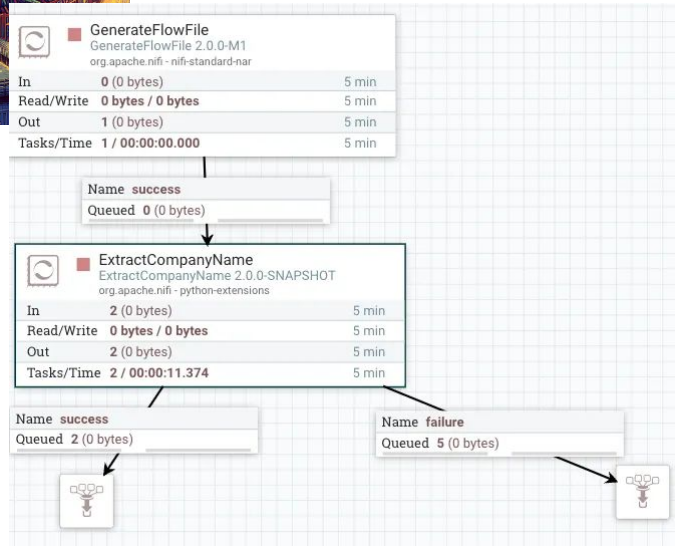
- **Record Readers** - Avro, CSV, Grok, IPFIX, JSAN1, JSON, Parquet, Scripted, Syslog5424, Syslog, WindowsEvent, XML
- **Record Writers** - Avro, CSV, FreeFromText, Json, Parquet, Scripted, XML
- Record Reader and Writer support referencing a schema registry for retrieving schemas when necessary.
- Enable 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.





Extract Company Names

- Python 3.10+
- Hugging Face, NLP, SpaCY, PyTorch

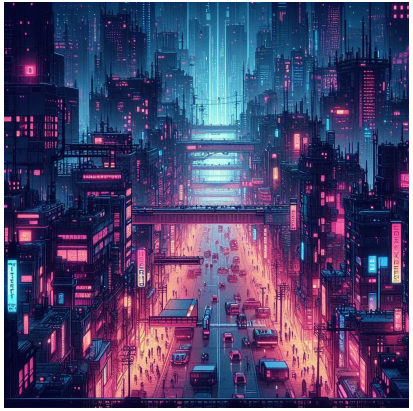


https://github.com/tspannhw/FLaNK-python-ExtractCompanyName-processor

CaptionImage

- Python 3.10+
- Hugging Face
- Salesforce/blip-image-captioning-large
- Generate Captions for Images
- Adds captions to FlowFile Attributes
- Does not require download or copies of your images

https://github.com/tspannhw/FLaNK-python-processors

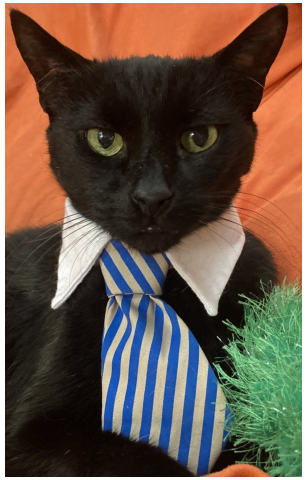


RESNetImageClassification

- Python 3.10+
- Hugging Face
- Transformers
- Pytorch
- Datasets
- microsoft/resnet-50
- Adds classification label to FlowFile Attributes
- Does not require download or copies of your images

https://github.com/tspannhw/FLaNK-python-processors

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

DEMO

RESOURCES AND WRAP-UP

https://www.linkedin.com/in/timothyspann/

Open Source Edition

- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

- Licensed under the ASF License
- Unsupported

Data for Breakfast

Free Data and AI Event

- King of Prussia
- Princeton
- New York
- Virtual

https://www.snowflake.com/events/data-for-breakfast/


SPIRIT
Business
Cat
ADULT
Size Costume
ONE SIZE FITS MOST


AUTOMATE EVERYTHING
imgflip.com