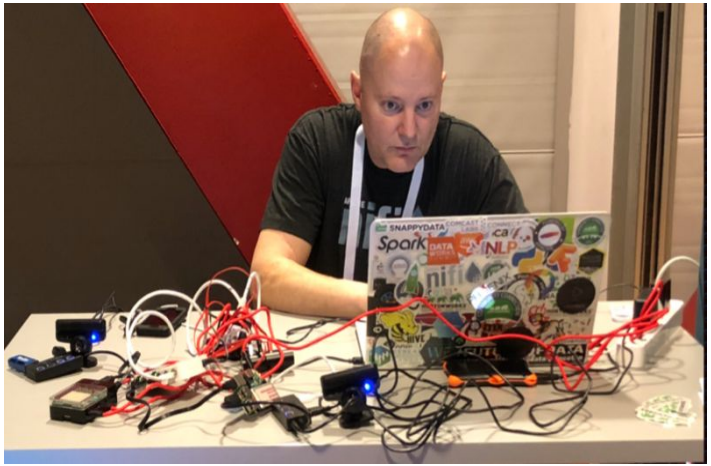
From Air Quality to Aircraft & Automobiles, Unstructured Data Is Everywhere

Tim Spann, Senior Solutions Engineer

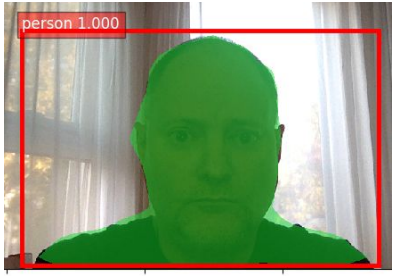
APACHE nifi

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.

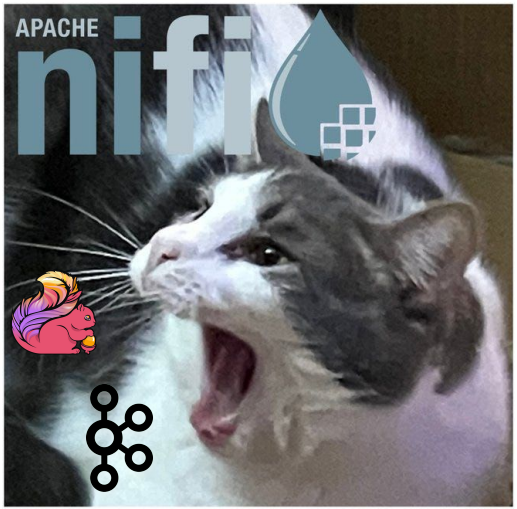
https://medium.com/@tspann
https://github.com/tspannhw



snowflake

Streamlit

AI + Streaming Weekly by Tim Spann



https://bit.ly/32dAJft

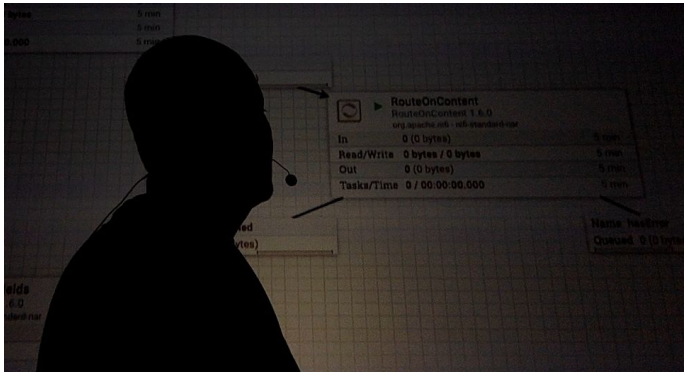
This week in Snowflake, Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Streamlit, Jupyter, Apache Iceberg, Apache Polaris, Python, Java, LLM, GenAI, Vectors and Open Source friends.

Introduction

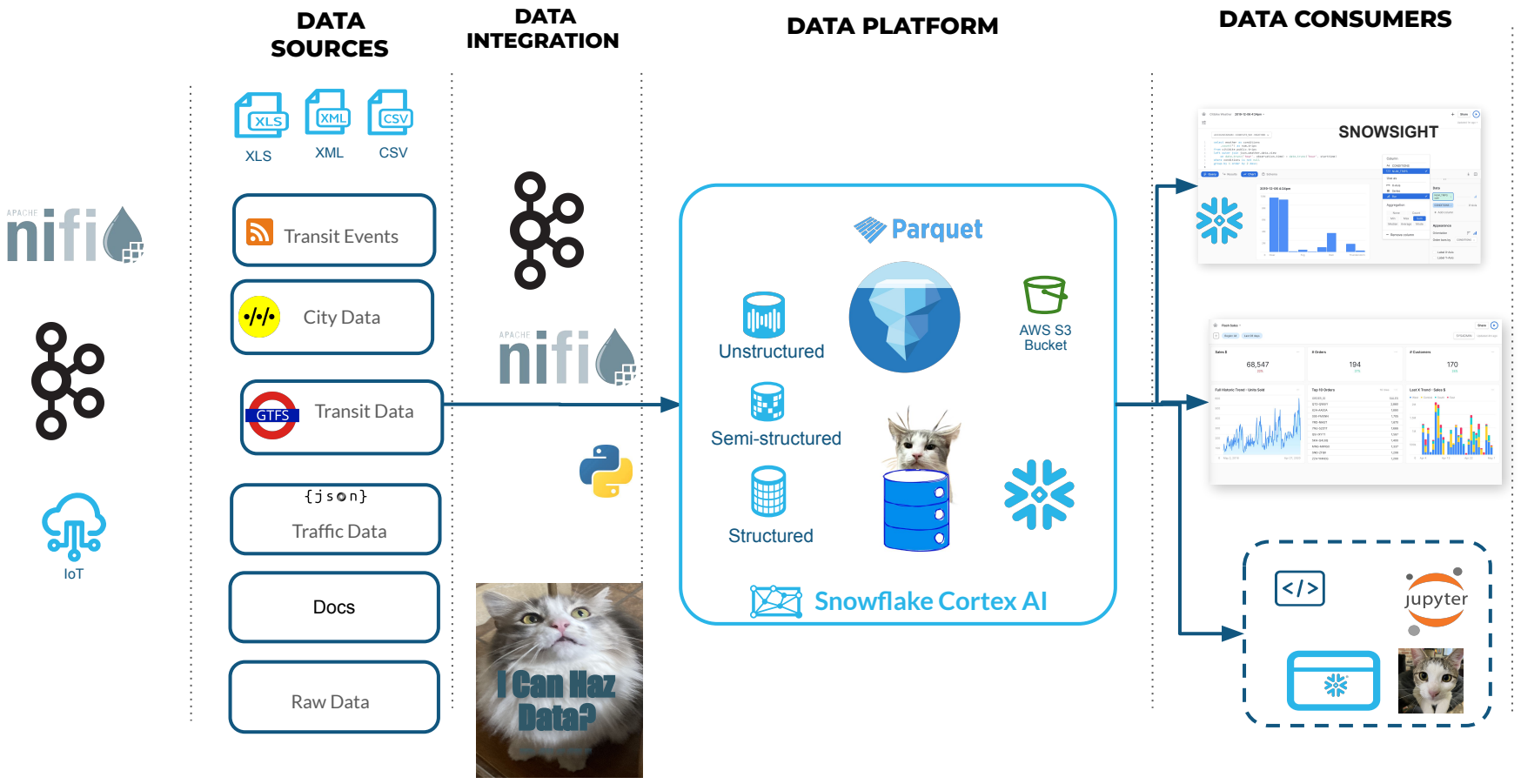
Overview

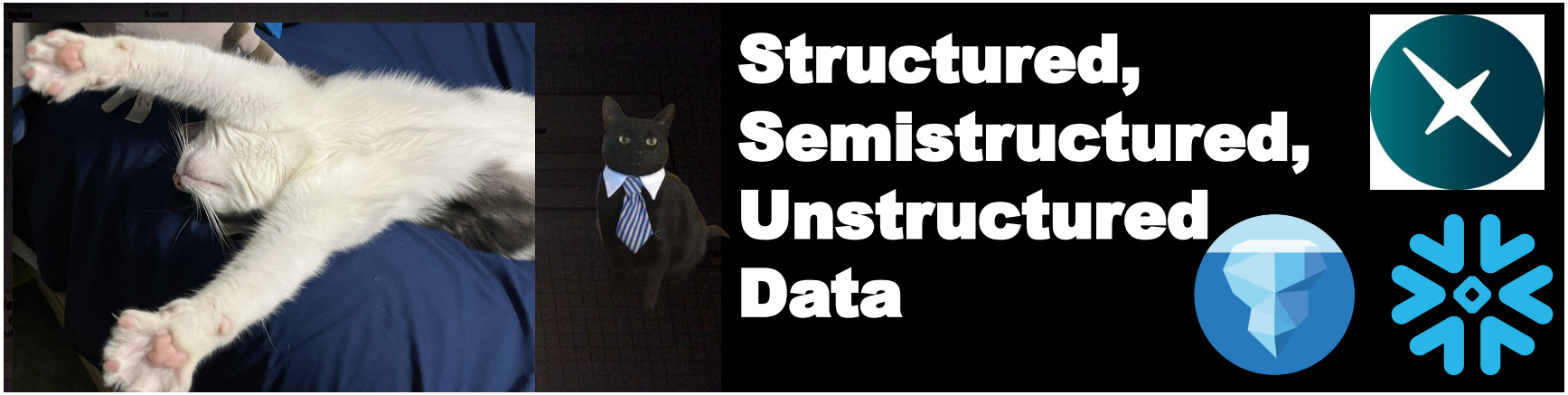
AI

Where, What, Why



Real-Time AI Open Lakehouse





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

Unstructured Data

Unstructured

- Lots of formats
- Text, Documents, PDF
- Images, Videos, Audio
- Email, Slack, Teams
- Logs
- Binary Data Formats
- Zip, Archives
- Variants

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

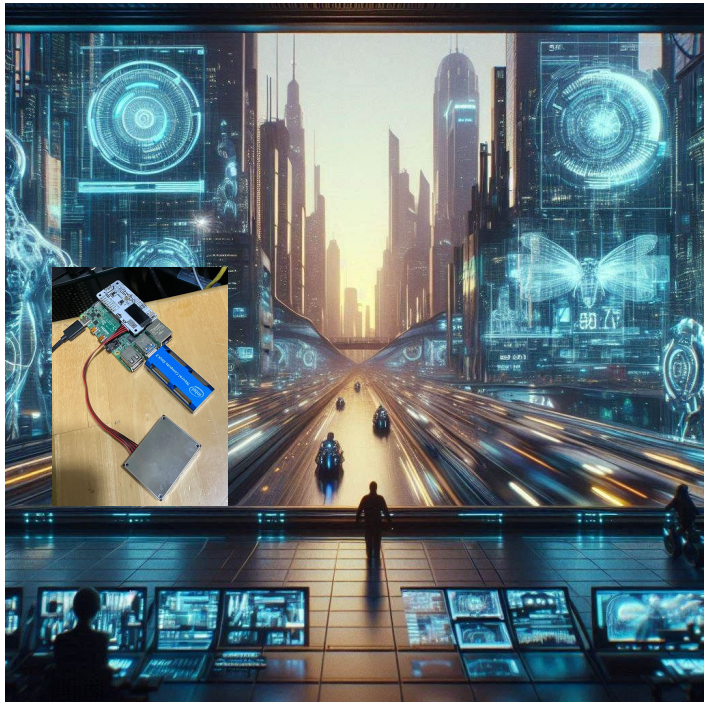
- Snowflake Tables
- Snowflake Hybrid Tables
- Apache Iceberg Tables
- Relational Tables
- Postgresql Tables
- CSV, TSV

Structured

Record-Oriented Data with NiFi

- Readers - Avro, CEF, CSV, Excel, Grok, Protobuf, JSON, Parquet, Scripted, Syslog-5424, Syslog, Windows Event, XML, YAML
- Writers - Avro, CSV, Free From Text, JSON, Parquet, Scripted, XML
- Schema registry integration for retrieving schemas
- Automatic parsing and serialization
- Bundling thousands of records in one microbatch

Table Design



```
create or replace ICEBERG TABLE AQ (
        DATEOBSERVED   STRING,
        HOUROBSERVED    STRING,
        LOCALTIMEZONE   STRING,
        REPORTINGAREA     STRING,
        STATECODE  STRING,
        LATITUDE NUMBER(7,3),
        LONGITUDE NUMBER(7,3),
        PARAMETERNAME  STRING,
        AQI NUMBER(2,0),
        CATEGORYNUMBER  STRING,
        CATEGORYNAME     STRING,
        TS    STRING,
        UUID    STRING,
        AQITMP NUMBER(2,0))
 EXTERNAL_VOLUME = V
CATALOG = 'SNOWFLAKE'
BASE_LOCATION = 'airquality/';
```

https://bit.ly/4jk4QUV

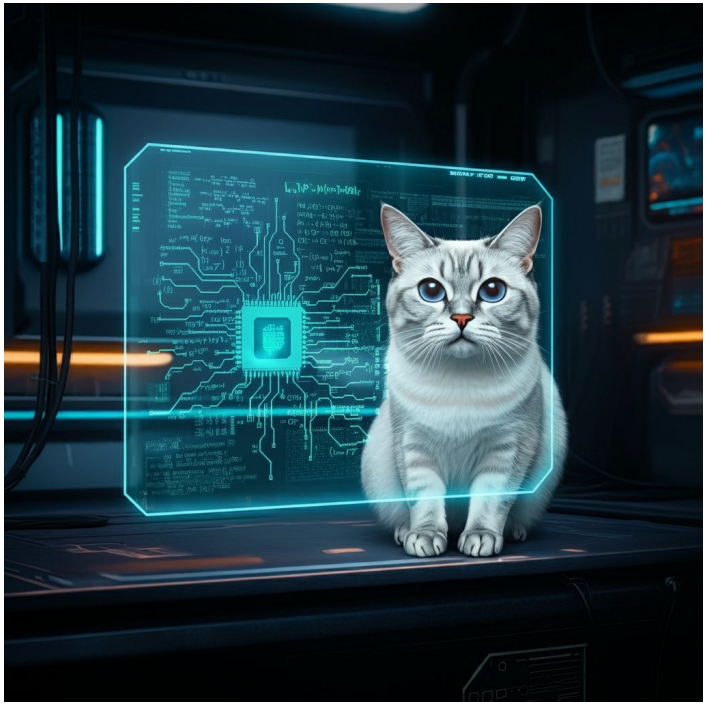
Open LLM Options

- Arctic Instruct
- Arctic-embed-m-v2.0
- Llama-3.3-70b
- Mixtral-8x7b
- Llama3.1-405b
- Mistral-7b
- Deepseek-r1

Continuous Ingest

- REST Feeds to Kafka
- Local Sensors to Kafka
- Apache NiFi Read
- Kafka Connector
- Snowpipe Streaming

Snowflake RAG



Build

Ingest -> Extract ->
Split -> Build Indexes

Serve

Orchestration |
Observability <->
Retrieval <-> Inference

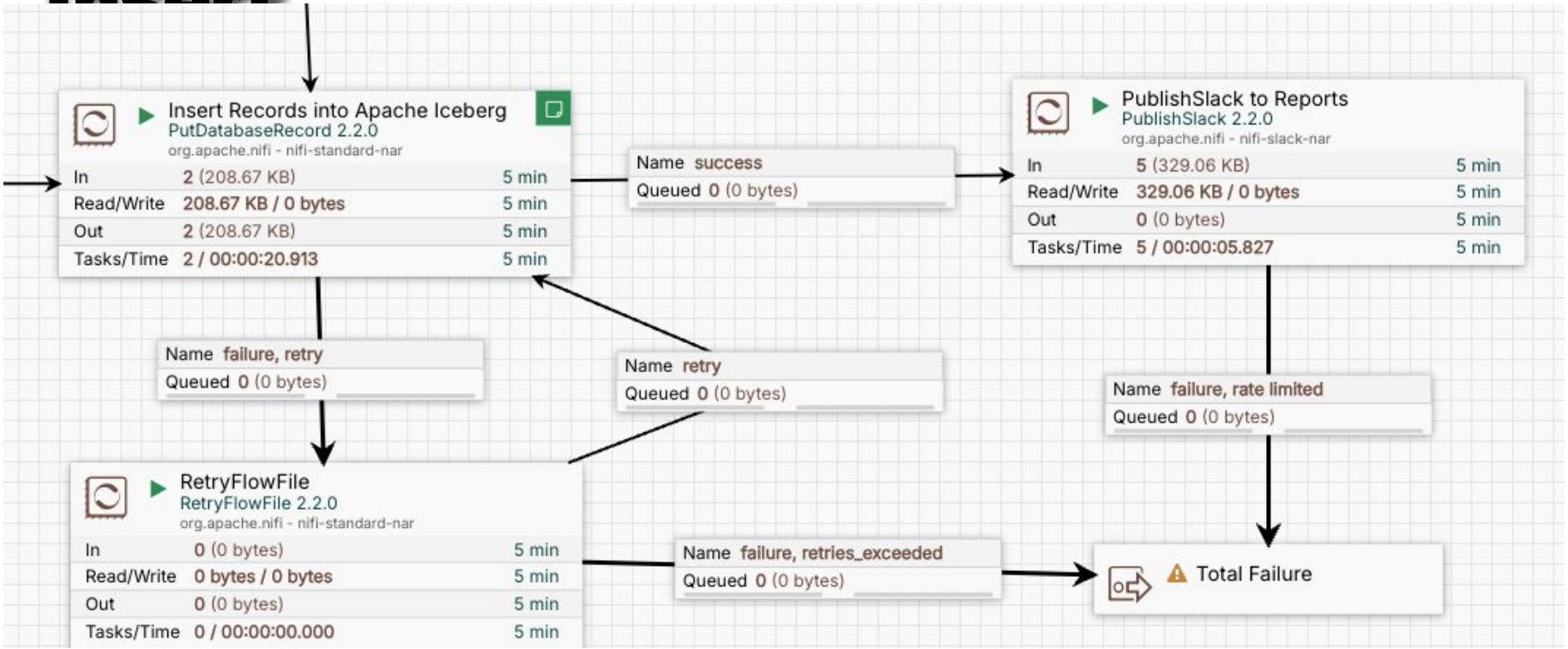
Apache Iceberg™ - Append

- NiFi - PutIcebergTable
- NiFi - PutDatabaseRecord
- Snowpark -

```
df.write.mode("append").
save_as_table("atable_iceberg")
```

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

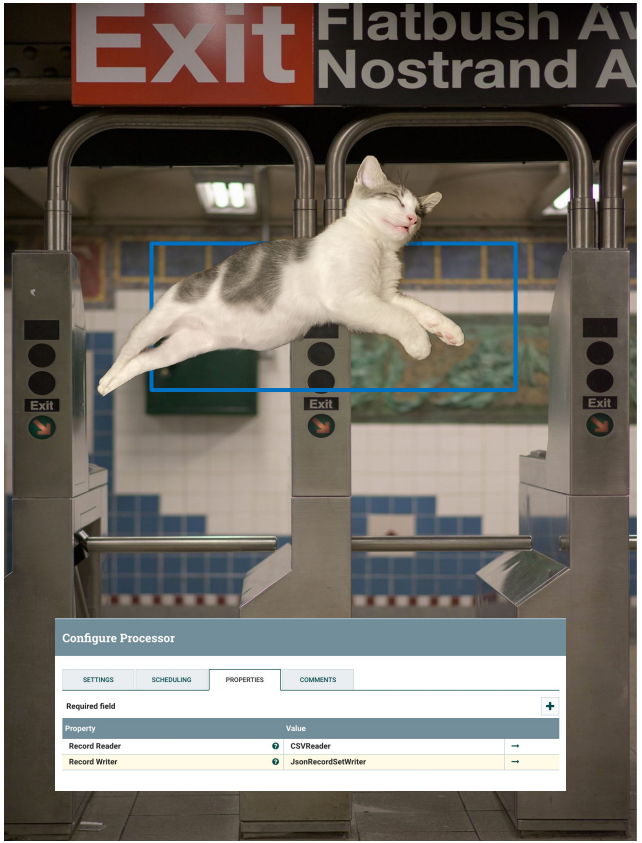
Apache Iceberg™ - JDBC - Insert



Parse Documents

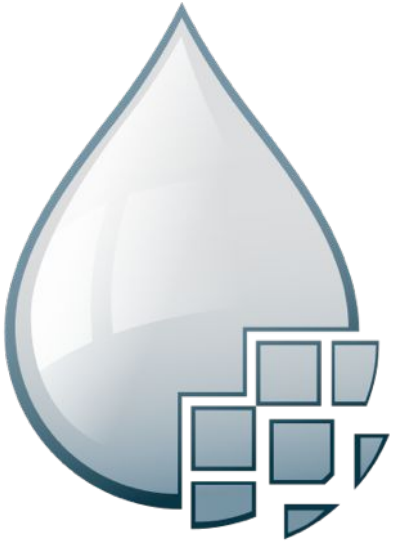
```
SELECT SNOWFLAKE.CORTEX.PARSE_DOCUMENT(
@documents, 'mydoc.pdf', {'mode': 'LAYOUT'});
```

Apache NiFi



- From laptop to 1,000 nodes
- Ingest, Extract, Split
- Enrich, Transform
- Mature, 10 years+
- Any Data, Any Source
- LLM Calls
- Data Provenance
- Back Pressure
- Guaranteed Delivery

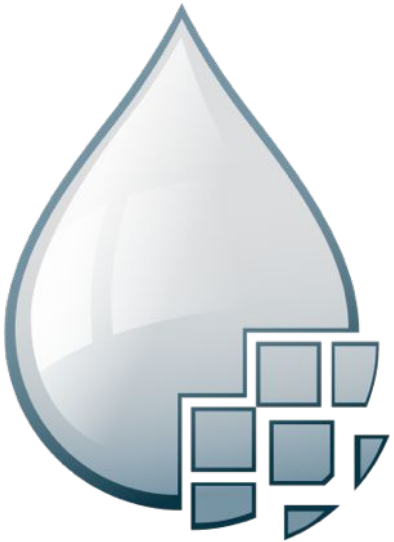
Apache NiFi for Data Ingest, Movement and Routing



- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
- 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
- REST, TCP/IP, UDP
- Enrichment
- Universal Visual Processor
- Simple Event Processor
- Routing
- Feeding data to Central Messaging
- Support for modern protocols
- Kafka Protocol Source/Sink
- Universal REST API

APACHE NIFI 2.4.0 FEATURES

Real-Time Integration and AI

Major Updates:

- Python Integration
- Parameterization
- JDK 21+, Python 3.9
- Provenance / Data Lineage
- Rules Engine for Development Assistance
- Many Cloud Processors
- Integration with Zendesk, Slack, SaaS Tools
- Database Tables as Schemas
- Amazon Glue Schema Registry
- Open Telemetry Support
- Snowflake-jdbc to 3.22.0

PROVENANCE

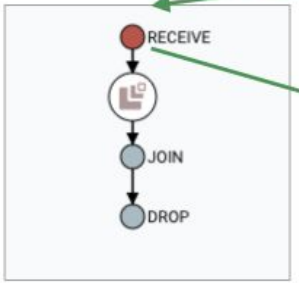
Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST

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

DEMO

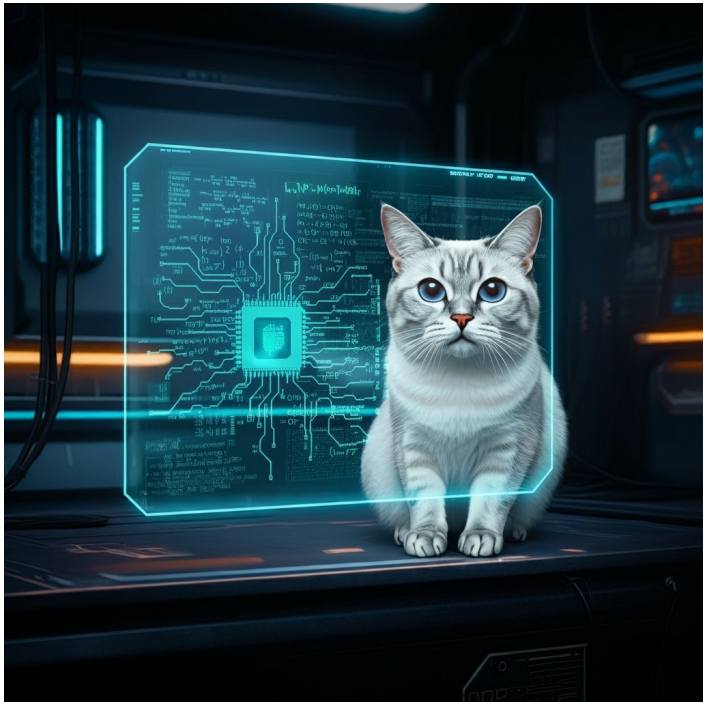
Open Source Edition

- 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

RESOURCES



https://medium.com/@tspann/yes-apache-nifi-can-do-that-c7fcaca5a177


SPIRIT
Business
Cat
ADULT
Size Costume
ONE SIZE FITS MOST


AUTOMATE EVERYTHING
imgflip.com