Unlocking Your Data's Potential with Python

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-Hortonworks.

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

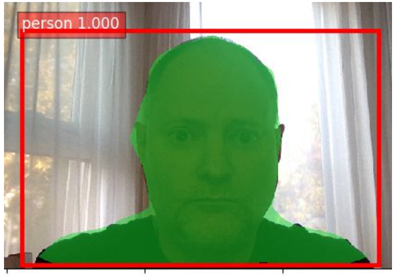
DZone. REFCARDS TREND REPORTS E

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...



snowflake

Streamlit

AGENDA

Introduction

Overview

Superfriends

Where, What, Why

Unlocking Data Requires a Team

Structured, Semistructured, Unstructured Data

Semi-structured

Semi-Structured Data

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

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

Unstructured Data

Unstructured

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

Structured Data

Structured

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

Python and Apache Iceberg™

Apache Iceberg™ - Overview

“Apache Iceberg is a high performance open-source format for large analytic tables.” - Wiki

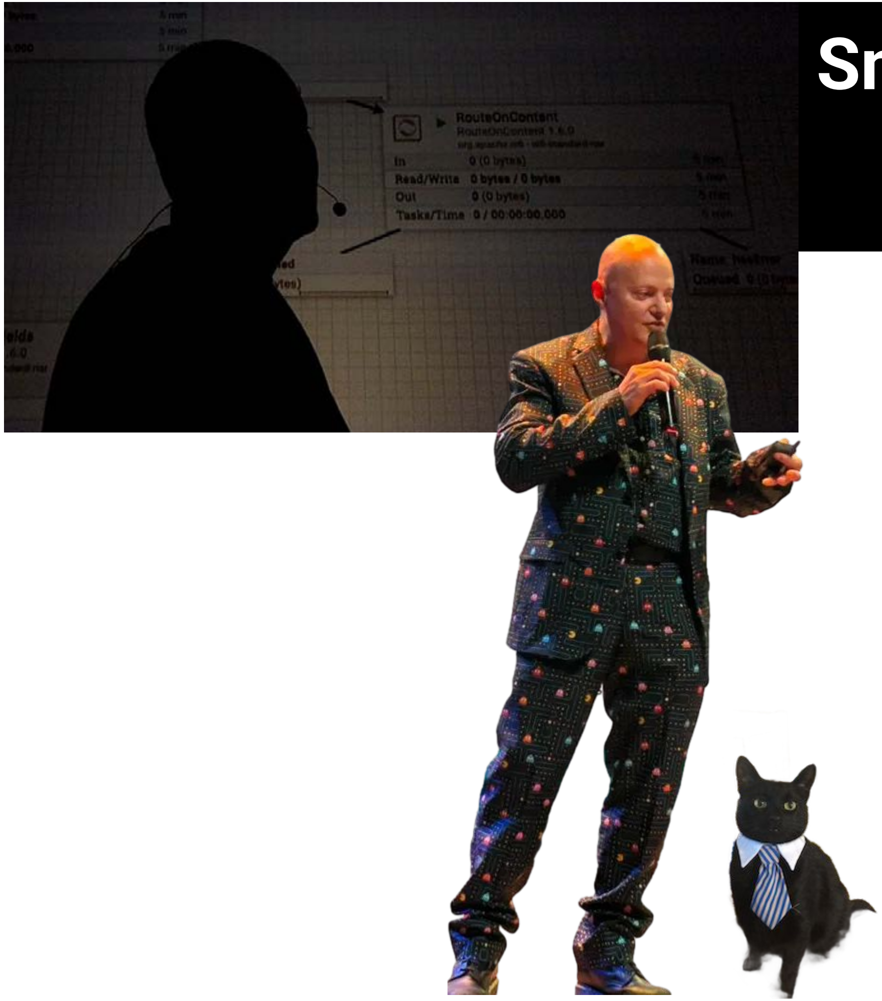
- Strong SQL Support
- Support for many engines (Snowflake, Trino, Flink, Presto, Hive)
- Support for many data catalogs like Polaris & Nessie
- Full Schema Evolution, Time Travel, Hidden Partitioning
- Rollback, Data compaction

Apache Iceberg™ - Append

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

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

Snowpark


RouteOnContent
RouteOnContent 1.6.0
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes)
Tasks/Time 0 / 00:00:00.000

Snowpark

LIBRARIES

Data Engineering, Machine Learning, Streamlit

FAST DATA ACCESS

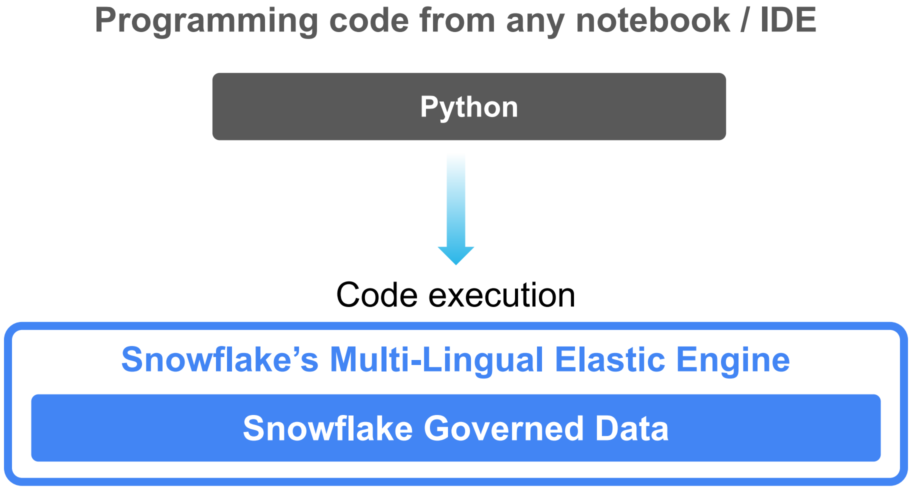
Secure data sessions, Dataframes, pandas

WHY USE IT

Build scalable data pipelines, ML models, apps, and other data processing tasks using any language in Snowflake without any governance or security trade-offs

HOW TO USE IT

Write custom code from any notebook or IDE and automatically push down processing into Snowflake's elastic compute engine


Programming code from any notebook / IDE
Python
Code execution
Snowflake's Multi-Lingual Elastic Engine
Snowflake Governed Data

Snowpark with Python

Libraries & code execution environments

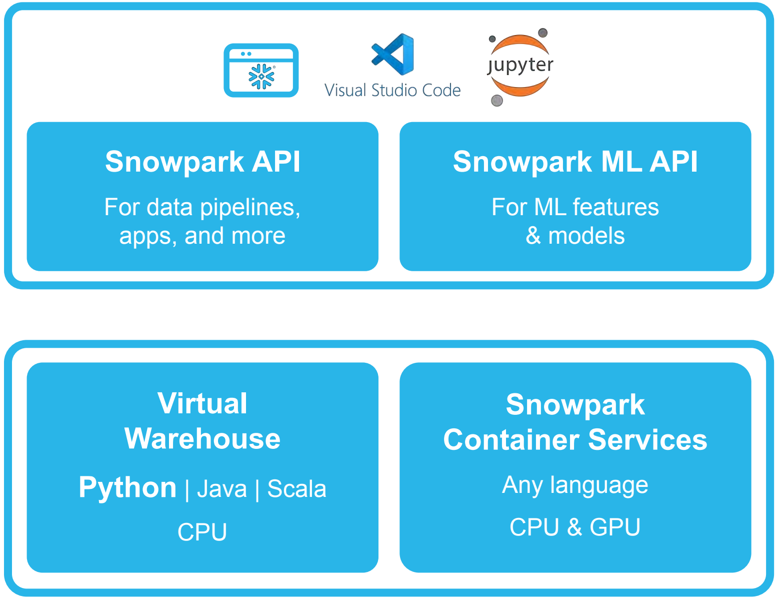
CODE DEVELOPMENT & DEPLOYMENT IN CLIENT-SIDE LIBRARIES

Use libraries with popular **Python** frameworks pre-installed in Snowflake Notebooks or downloaded into your IDE of choice and push down processing to Snowflake

CODE EXECUTION ENVIRONMENTS IN SNOWFLAKE'S ENGINE

Run **Python** and other programming code next to your data in Snowflake. Automatically push down processing in multi-lingual runtimes built right into Snowflake's elastic compute engine

From Snowflake Notebooks or any IDE



Snowpark API

For data pipelines, apps, and more

Snowpark ML API

For ML features & models

Virtual Warehouse

Python | Java | Scala

CPU

Snowpark Container Services

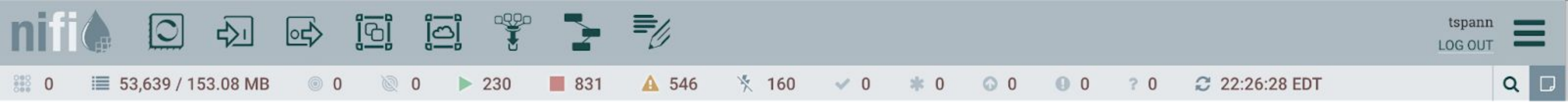
Any language

CPU & GPU

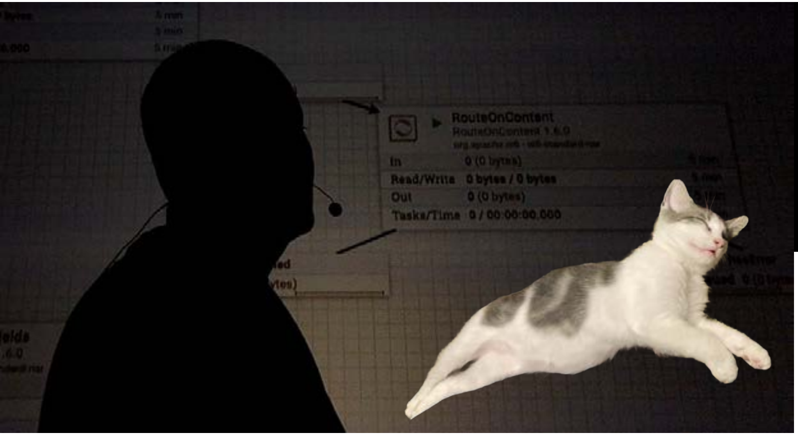
Storing Data

Ingest into Tables

- RedisVL
- Milvus
- Snowflake
- Apache Pinot
- PgVector



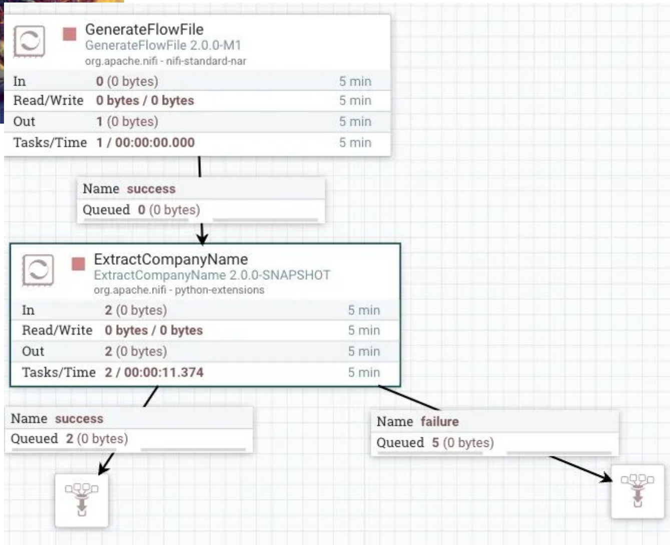
0 53,639 / 153.08 MB 0 0 230 831 546 160 0 0 0 0 0 22:26:28 EDT



Apache NiFi - Python

Extract Company Names

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



GenerateFlowFile
GenerateFlowFile 2.0.0-M1
org.apache.nifi - nifi-standard-nar

```
In          0 (0 bytes)             5 min
Read/Write  0 bytes / 0 bytes       5 min
Out         1 (0 bytes)             5 min
Tasks/Time  1 / 00:00:00.000        5 min
```

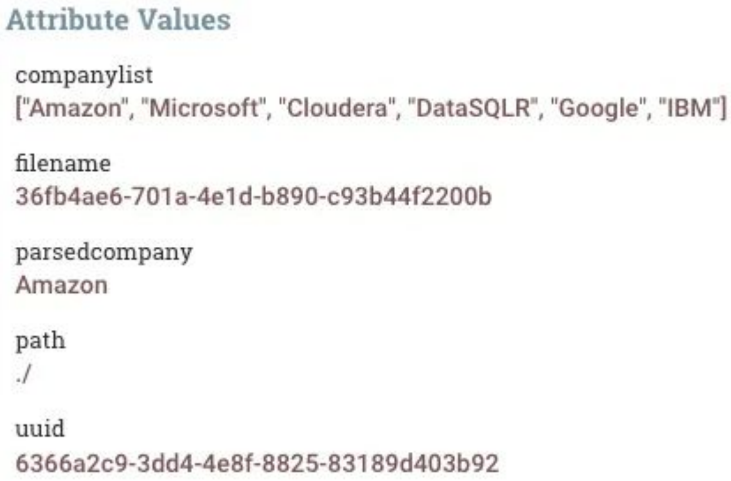
```
Name    success
Queued  0 (0 bytes)
```

ExtractCompanyName
ExtractCompanyName 2.0.0-SNAPSHOT
org.apache.nifi - python-extensions

```
In          2 (0 bytes)             5 min
Read/Write  0 bytes / 0 bytes       5 min
Out         2 (0 bytes)             5 min
Tasks/Time  2 / 00:00:11.374        5 min
```

```
Name    success
Queued  2 (0 bytes)
```

```
Name    failure
Queued  5 (0 bytes)
```



Attribute Values

companylist
["Amazon", "Microsoft", "Cloudera", "DataSQLR", "Google", "IBM"]

filename
36fb4ae6-701a-4e1d-b890-c93b44f2200b

parsedcompany
Amazon

path
./

uuid
6366a2c9-3dd4-4e8f-8825-83189d403b92

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

NSFWImageDetection

- Python 3.10+
- Hugging Face
- Transformers
- Falconsai/nsfw_image_detection
- Adds normal and nsfw to FlowFile Attributes
- Gives score on safety of image
- Does not require download or copies of your images

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

FacialEmotionsImageDetection

- Python 3.10+
- Hugging Face
- Transformers
- facial_emotions_image_detection
- Image Classification
- Adds labels/scores 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

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

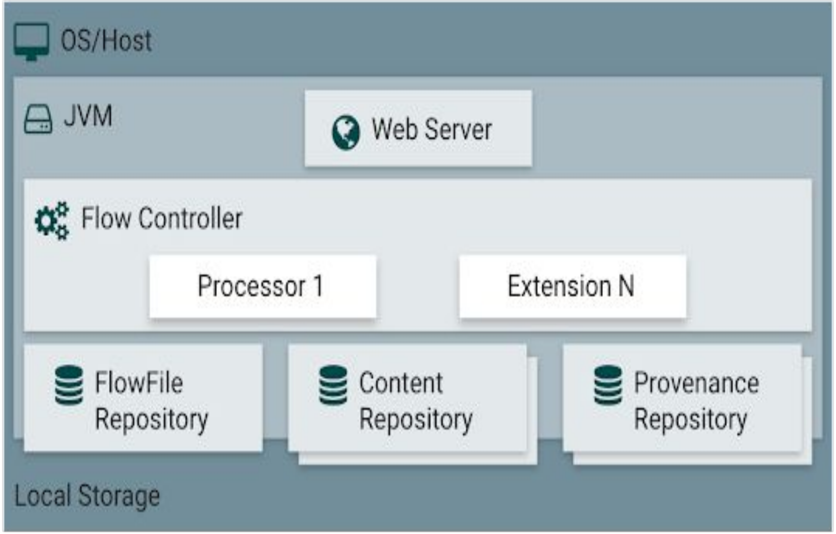
APACHE NIFI 2.0 FEATURES

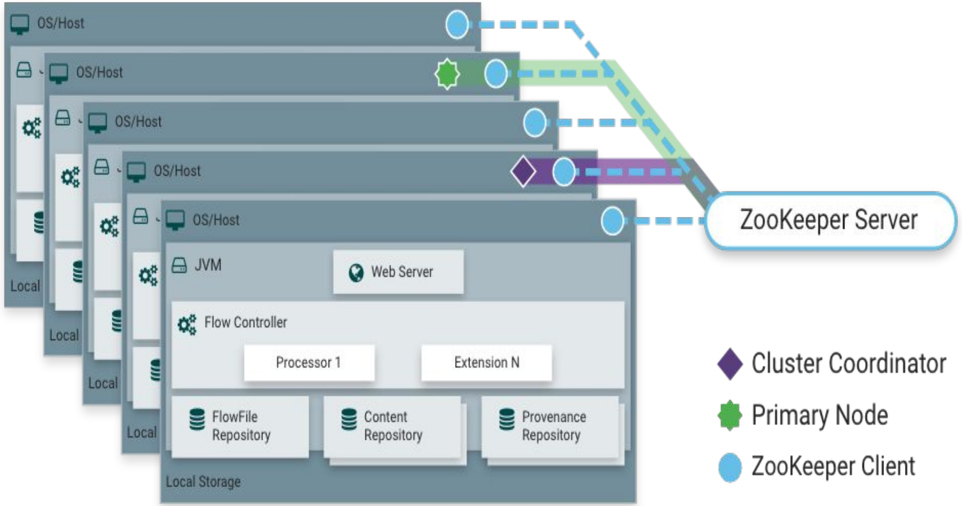
DataFlow is built for Real-Time Integration and AI

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

DEMO

RESOURCES AND WRAP-UP

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

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.

SPIRIT
Business
Cat
ADULT
Size Costume
ONE SIZE FITS MOST

AUTOMATE EVERYTHING
imgflip.com