# Real Time AI Pipeline Architectures with Flink SQL, NiFi, Kafka and Iceberg

**Timothy Spann**
Senior Solution Engineer, Snowflake

snowflake®

# Tim Spann

paasdev.bsky.social
@PaasDev // Blog: datainmotion.dev

Senior Solutions Engineer, Snowflake
NY/NJ/Philly - Cloud Data + AI Meetups

PAST EXPERIENCE:

ex-Zilliz | ex-Pivotal | ex-Cloudera | ex-HPE | ex-StreamNative | ex-EY | ex-Hortonworks

https://medium.com/@tspann

https://github.com/tspannhw


snowflake

# AI + Streaming Weekly by Tim Spann


snowflake

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.

# Streaming Enterprise Lakehouse

Connect your data to securely build or scale your high-performing lakehouse

## Secure your data estate

Protect your data against lock-in, unwanted access, and disasters

## Conquer data complexity

Connect any data to streamline pipelines with fully managed infrastructure

## Empower your teams

Deliver trusted AI, Analytics and more to any data customer

# REAL-TIME AI PIPELINE CORE TECHNOLOGIES

Leveraging best-in-class open-source for scalable, low-latency data processing.

## APACHE FLINK: STREAM PROCESSING

Stateful stream processing.
Flink SQL for unified batch and stream analytics.

## APACHE NIFI: DATA FLOW

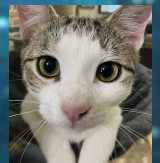
Automated data flow management.
Easy ingestion, routing, and transformation.

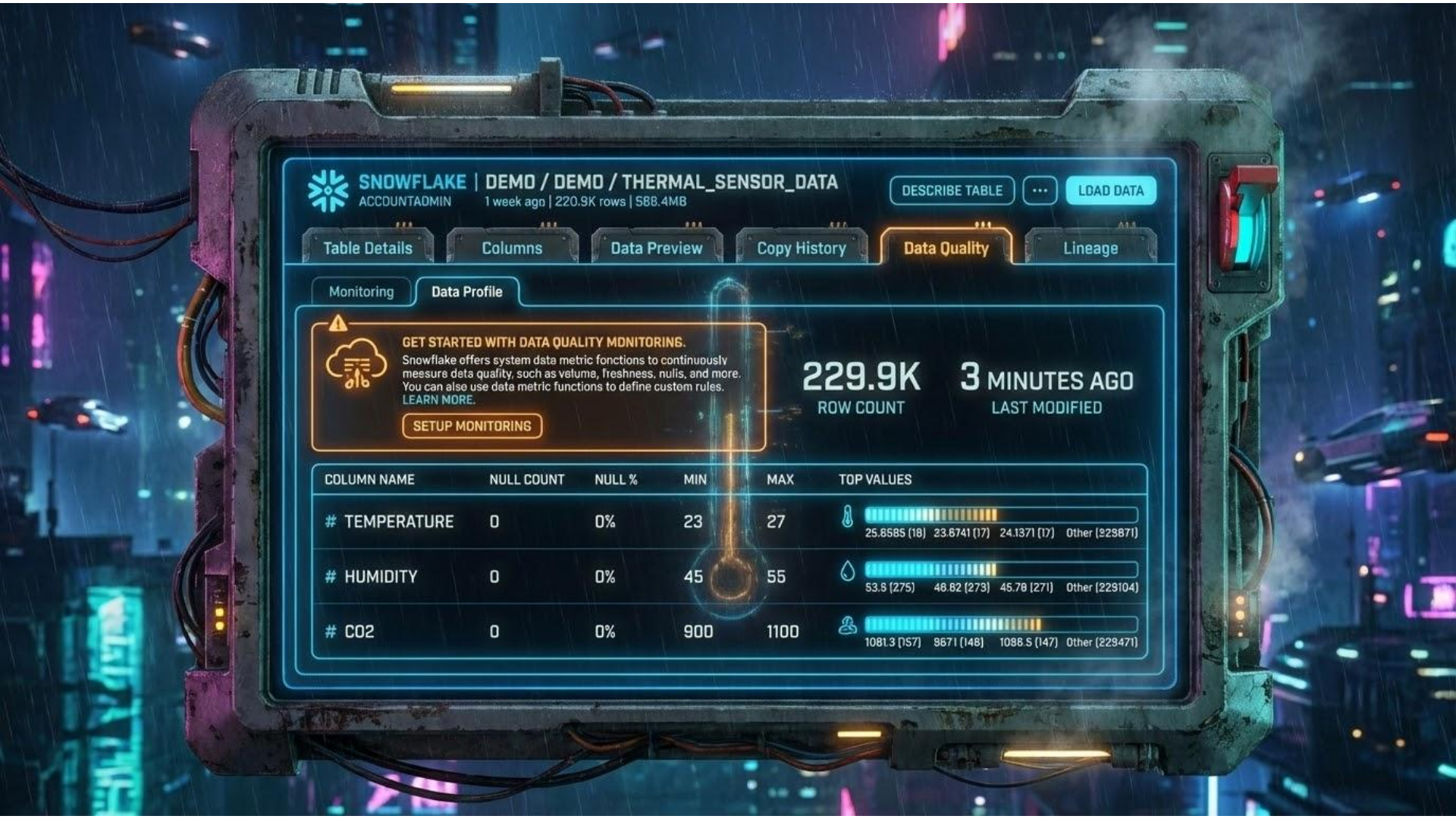
## APACHE KAFKA: EVENT STREAMING

High-throughput, distributed event streaming platform.
The central nervous system.

## APACHE ICEBERG: TABLE FORMAT

Open table format for huge analytic datasets.
ACID transactions on the data lake.




SNOWFLAKE | DEMO / DEMO / THERMAL_SENSOR_DATA
ACCOUNTADMIN
1 week ago | 220.9K rows | 588.4MB
DESCRIBE TABLE
LOAD DATA
Table Details
Columns
Data Preview
Copy History
Data Quality
Lineage
Monitoring
Data Profile
GET STARTED WITH DATA QUALITY MONITORING.
Snowflake offers system data metric functions to continuously measure data quality, such as volume, freshness, nulls, and more. You can also use data metric functions to define custom rules.
LEARN MORE.
SETUP MONITORING
229.9K
ROW COUNT
3 MINUTES AGO
LAST MODIFIED
COLUMN NAME
NULL COUNT
NULL %
MIN
MAX
TOP VALUES
# TEMPERATURE
0
0%
23
27
25.8585 (18) 23.8741 (17) 24.1371 (17) Other (223871)
# HUMIDITY
0
0%
45
55
53.8 (275) 48.82 (273) 45.78 (271) Other (223104)
# CO2
0
0%
900
1100
1081.3 (157) 967.1 (148) 1088.5 (147) Other (223471)

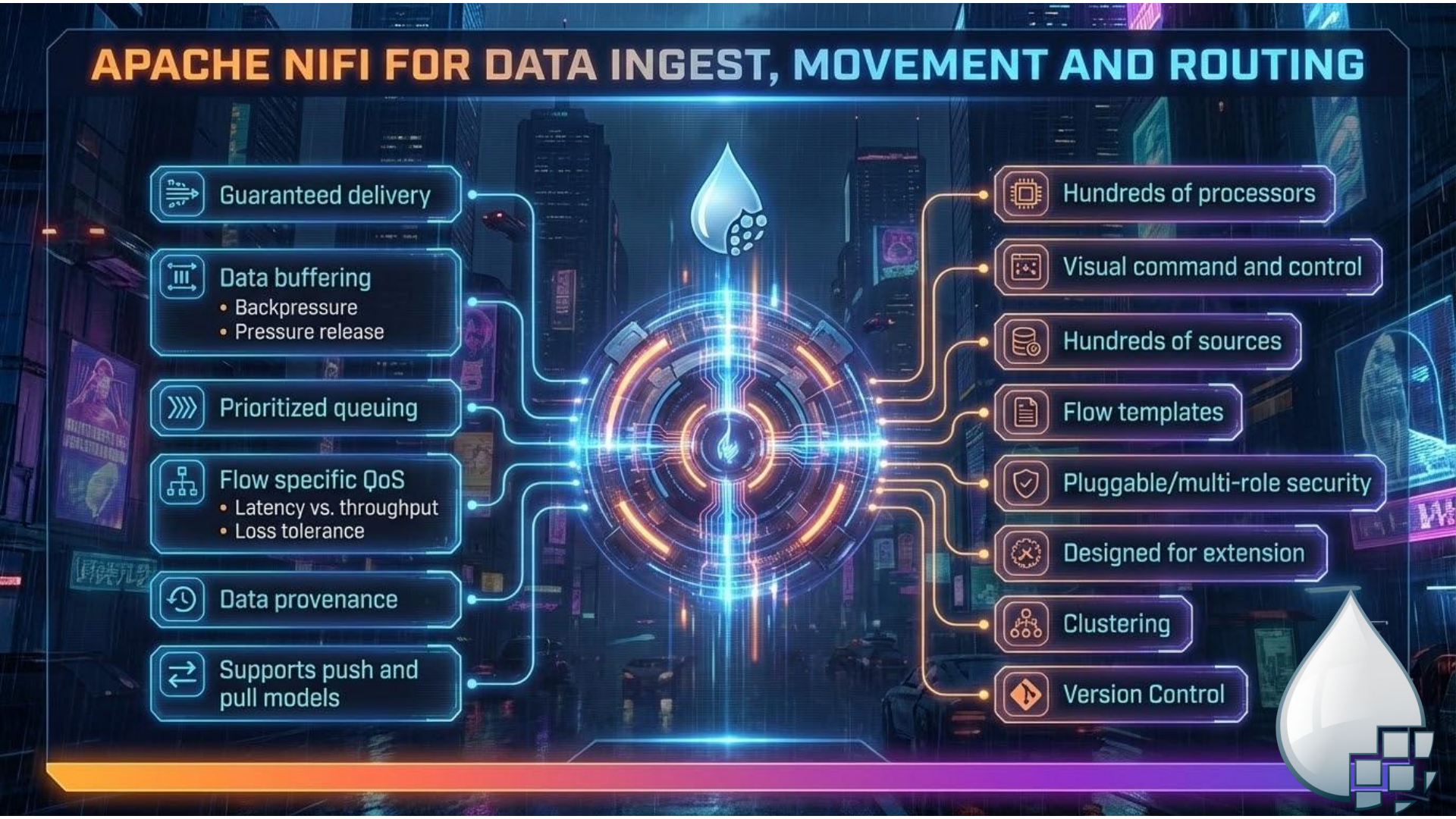
# SEMI-STRUCTURED DATA

Semi-structured


- 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

# APACHE NIFI FOR DATA INGEST, MOVEMENT AND ROUTING



# APACHE FLINK FOR STATEFUL STREAM PROCESSING

## Low-latency processing

- Sub-second results
- Real-time insights

## True streaming & batch unification

- Single API
- Consistent semantics

## Stateful computations

- Managed state
- Exactly-once guarantees

## Fault tolerance with savepoints

- Automatic recovery
- Versioned state

## Event-time semantics

- Handling late data
- Watermarks

## Scalable to thousands of nodes

- Horizontal scaling
- High throughput

# WHAT IS ICEBERG?


Apache Iceberg is *an* ***open table format for huge analytic datasets.***

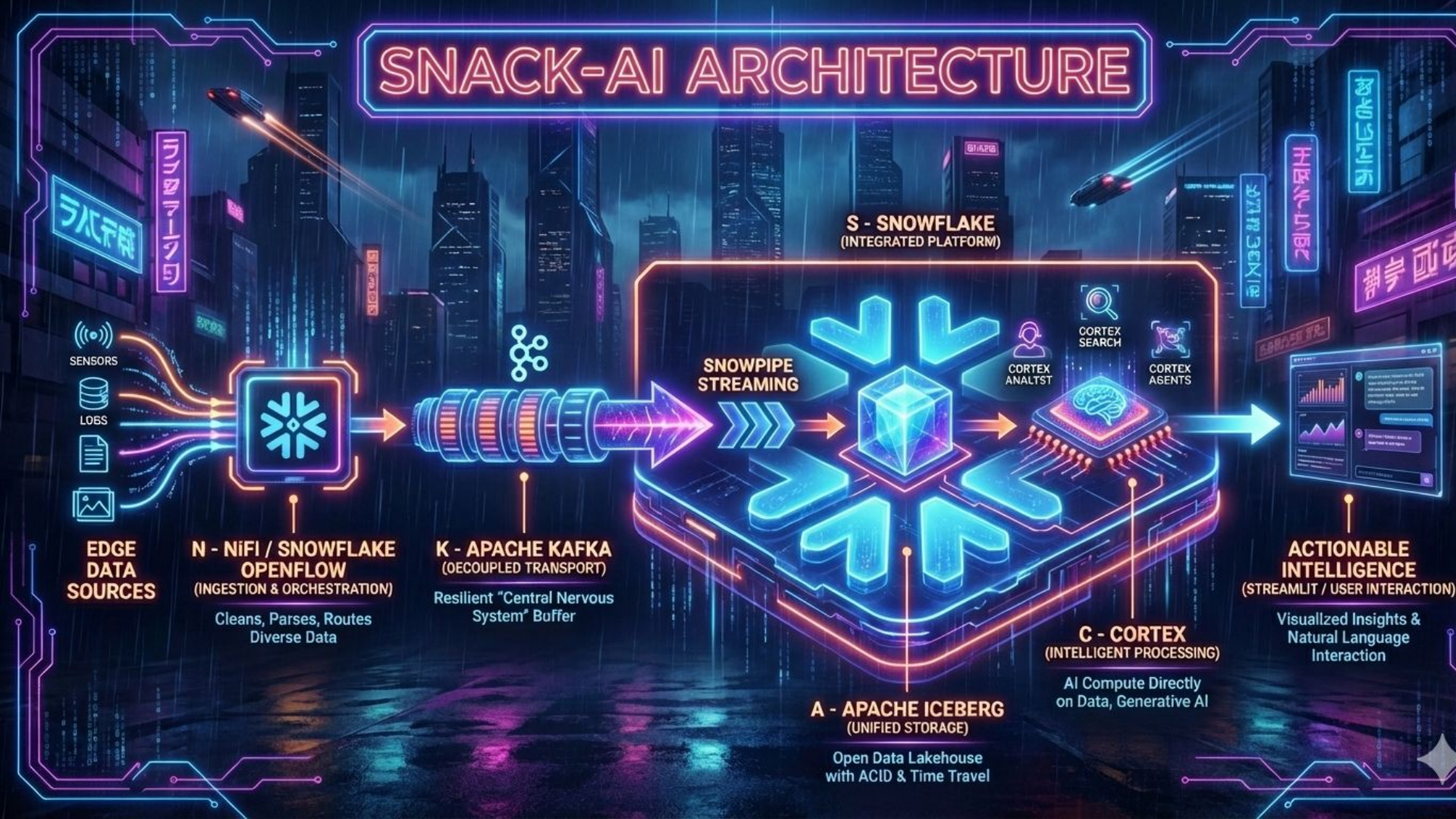
Evolution of Hive; addresses long-standing gaps, consistency, & performance issues.

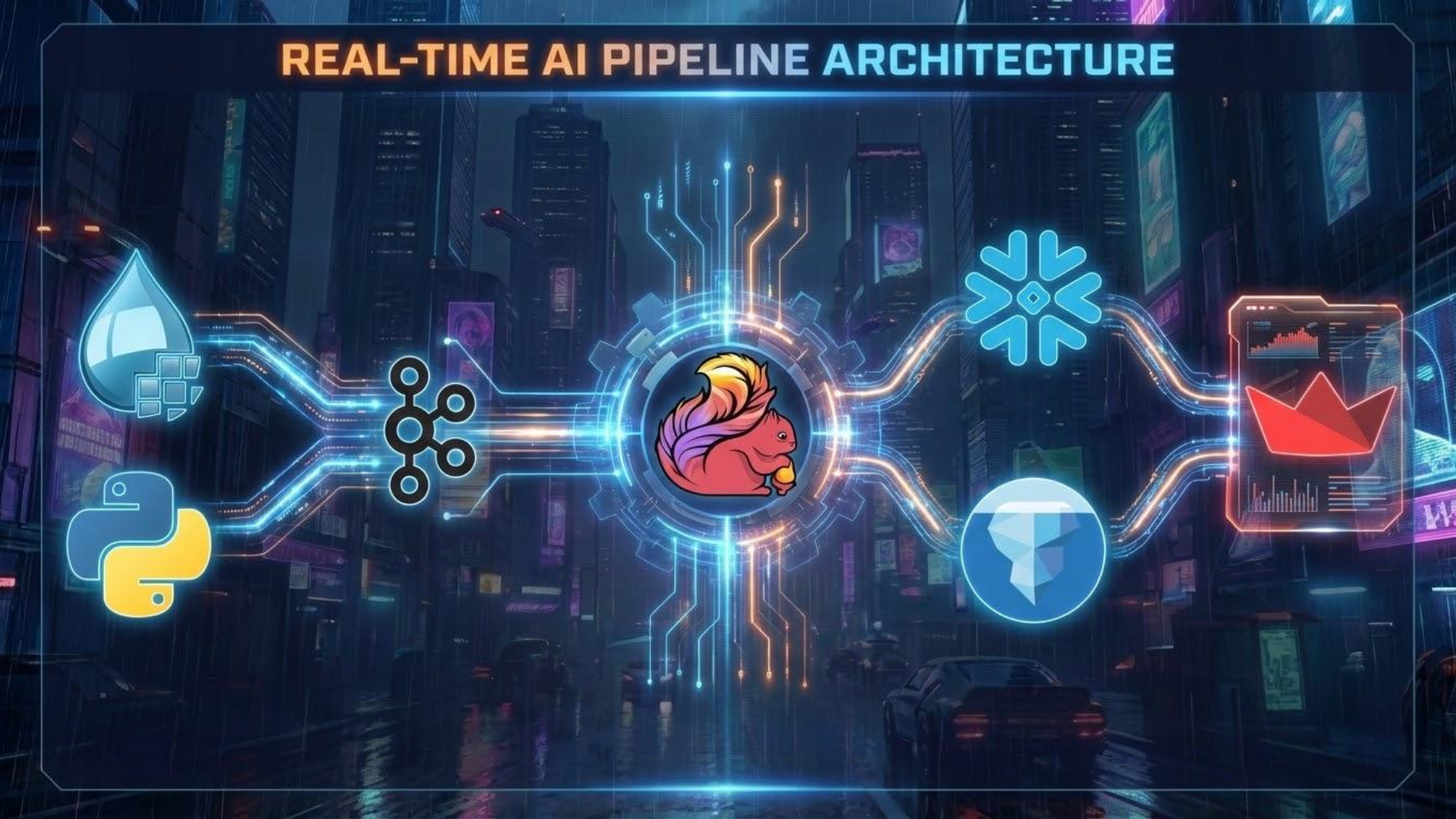
Agnostic to processing engines, file formats, and frameworks

Provides several features customers want from data lake tables, including:

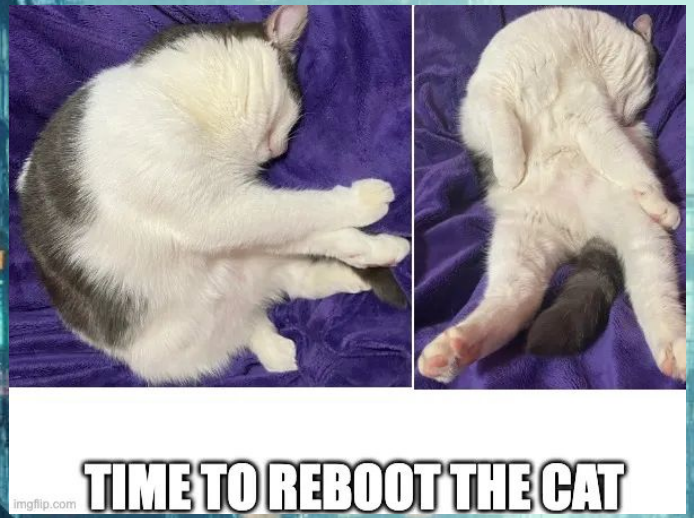
- *Schema evolution*
- *Hidden partitioning*
- *Time travel*
- *Increased performance (fast filtering, scanning)*
- *Safe concurrency*

Netflix donated Iceberg to the Apache Software Foundation, where it is developed by many today


SNACK-AI ARCHITECTURE
S - SNOWFLAKE
(INTEGRATED PLATFORM)
SENSORS
LOGS
SNOWPIPE
STREAMING
CORTEX
ANALTST
CORTEX
SEARCH
CORTEX
AGENTS
EDGE
DATA
SOURCES
N - NiFI / SNOWFLAKE
OPENFLOW
(INGESTION & ORCHESTRATION)
Cleans, Parses, Routes
Diverse Data
K - APACHE KAFKA
(DECOUPLED TRANSPORT)
Resilient "Central Nervous
System" Buffer
A - APACHE ICEBERG
(UNIFIED STORAGE)
Open Data Lakehouse
with ACID & Time Travel
C - CORTEX
(INTELLIGENT PROCESSING)
AI Compute Directly
on Data, Generative AI
ACTIONABLE
INTELLIGENCE
(STREAMLIT / USER INTERACTION)
Visuallzed Insights &
Natural Language
Interaction


REAL-TIME AI PIPELINE ARCHITECTURE


REAL-TIME IOT DATA PIPELINE WITH MQTT, NIFI, FLINK, ICEBERG & JUPYTER
MQTT
Data
jupyter




# RESOURCES AND WRAP-UP

# REFERENCES

## General Guides & Github

https://github.com/tspannhw/FLaNK-python-processors
https://nifi.apache.org/nifi-docs/python-developer-guide.html

## Transit & GTFS Real-Time flows

https://medium.com/@tspann/automating-transit-gtfs-flows-7a5c4d3dafe8
https://medium.com/@tspann/real-time-irish-transit-analytics-ea76164c9595
https://medium.com/cloudera-inc/boston-wheres-my-bus-llm-streaming-to-the-rescue-586dfd019237
https://medium.com/cloudera-inc/real-time-in-boston-part-1
https://medium.com/cloudera-inc/real-time-in-boston-part-1-0f92d7da3496

## Air Quality & Environmental Data

https://medium.com/@tspann/real-time-enrichment-of-air-quality-data-3ce670e4fc5b
https://medium.com/@tspann/real
https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-...v8-with-python-and-nifi-to-minio-s3-3277e73723ce

## Real-Time Streaming & AI/ML Integration

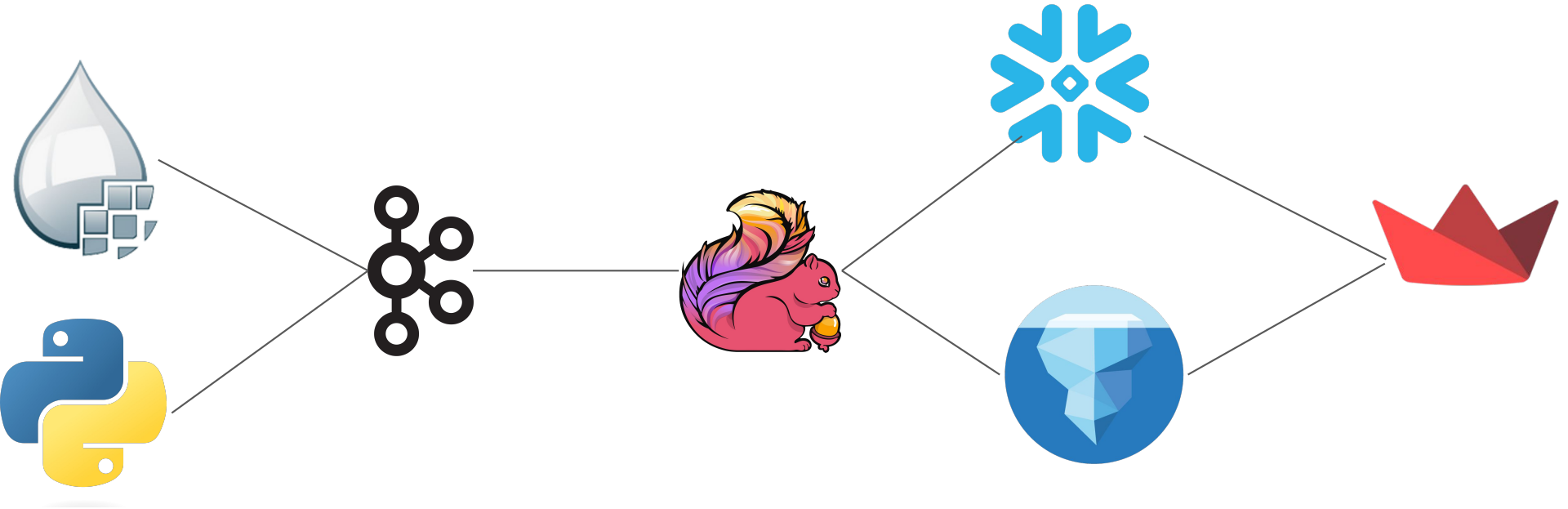
https://medium.com/@tspann/populating-an-open-lakehouse-with-codeless-data-streams-04292375ddaf
https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-v8-with-python-and-nifi-to-minio-s3-3277e73723ce

## Miscellaneous Cases & Articles

https://medium.com/@tspann/yes-apache-nifi-can-do-that-c7fcaca5a177
https://medium.com/cloudera-inc/nyc-traffic-are-you-kidding-me-6d3fa853903b

# REFERENCES


https://medium.com/cloudera-inc/searching-slack-from-apache-nifi-9ed562aa2397
https://medium.com/@tspann/yet-another-python-processor-45aaae6fe406
https://medium.com/cloudera-inc/building-a-library-of-python-processors-6b5517404a58
https://github.com/tspannhw/FLaNK-python-processors
https://github.com/tspannhw/CortexAISearchForAirQuality

https://jasonhhughes.medium.com/apache-flink-reading-apache-iceberg-tables-via-snowflakes-horizon-iceberg-rest-catalog-apis-7f64ecd201be

https://docs.snowflake.com/user-guide/tables-iceberg-query-using-external-query-engine-snowflake-horizon

https://iceberg.apache.org/docs/1.6.0/flink/?h=flink#preparation-when-using-flink-sql-client


https://bit.ly/4qNEiQe

bitly

snowghostbreakers.com