CLOUDERA

Using Apache NiFi, Apache Kafka, RisingWave, and Apache Iceberg with Stock Data and LLM

Karin Wolok
Developer Relations, Dev Marketing, and Community Programming @ Project Elevate

Tim Spann
Principal Developer Advocate, Cloudera

29-February-2024

Tim Spann

Twitter: @**PaasDev** // Blog: **datainmotion.dev**

Principal Developer Advocate.

Princeton Future of Data Meetup.

ex-Pivotal, ex-Hortonworks, ex-StreamNative, ex-PwC, ex-HPE

https://medium.com/@tspann

https://github.com/tspannhw

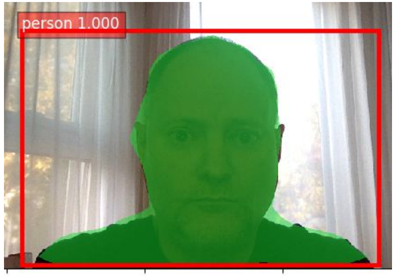
DZone. REFCARDS TREND REPORTS E

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…



CLOUDERA

Future of Data - NYC + NJ + Philly + Virtual

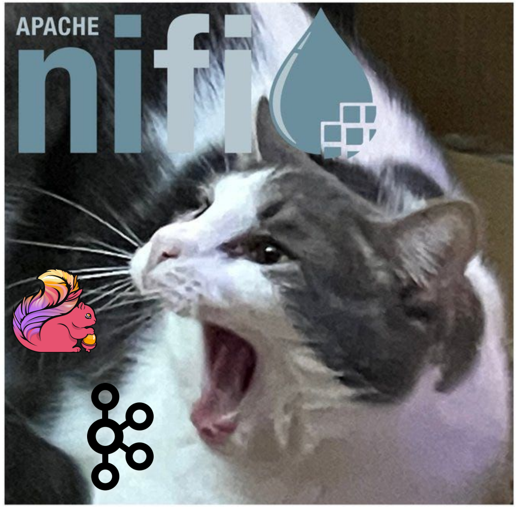
https://www.meetup.com/futureofdata-princeton/
https://www.meetup.com/futureofdata-newyork/

From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

FLaNK Stack Weekly by Tim Spann

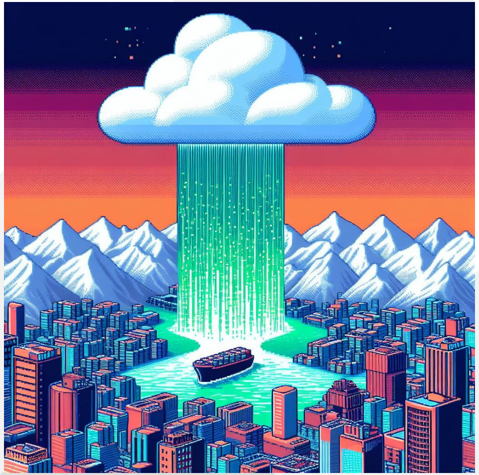


https://bit.ly/32dAJft

https://www.meetup.com/futureofdata-princeton/

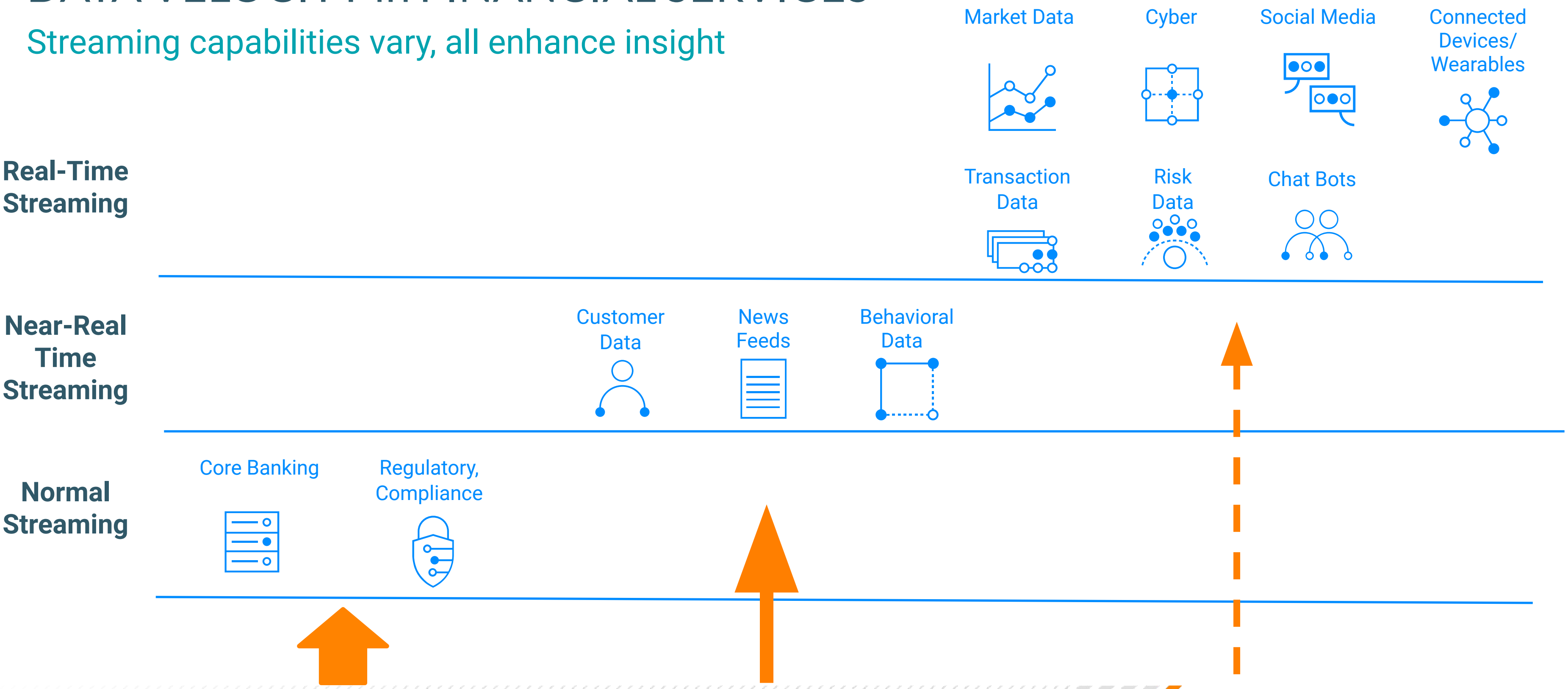
This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java, LLM, GenAI, Vector DB and Open Source friends.

OVERVIEW



DATA VELOCITY in FINANCIAL SERVICES

Streaming capabilities vary, all enhance insight



Real-Time Streaming

Market Data

Cyber

Social Media

Connected Devices/ Wearables

Transaction Data

Risk Data

Chat Bots

Near-Real Time Streaming

Customer Data

News Feeds

Behavioral Data

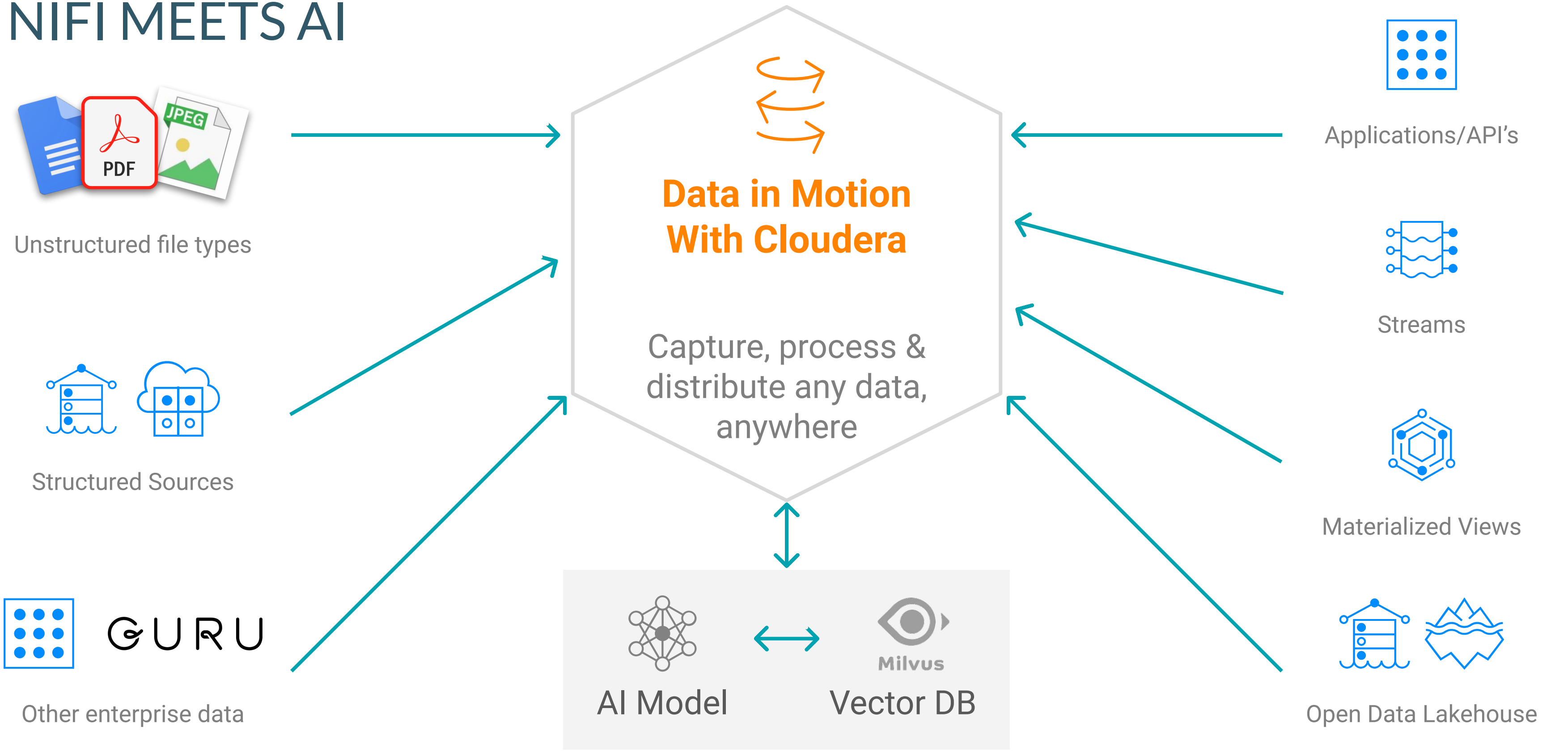
Normal Streaming

Core Banking

Regulatory, Compliance

CLOUDERA

NIFI MEETS AI



Unstructured file types

Structured Sources

Other enterprise data

Data in Motion With Cloudera

Capture, process & distribute any data, anywhere

AI Model

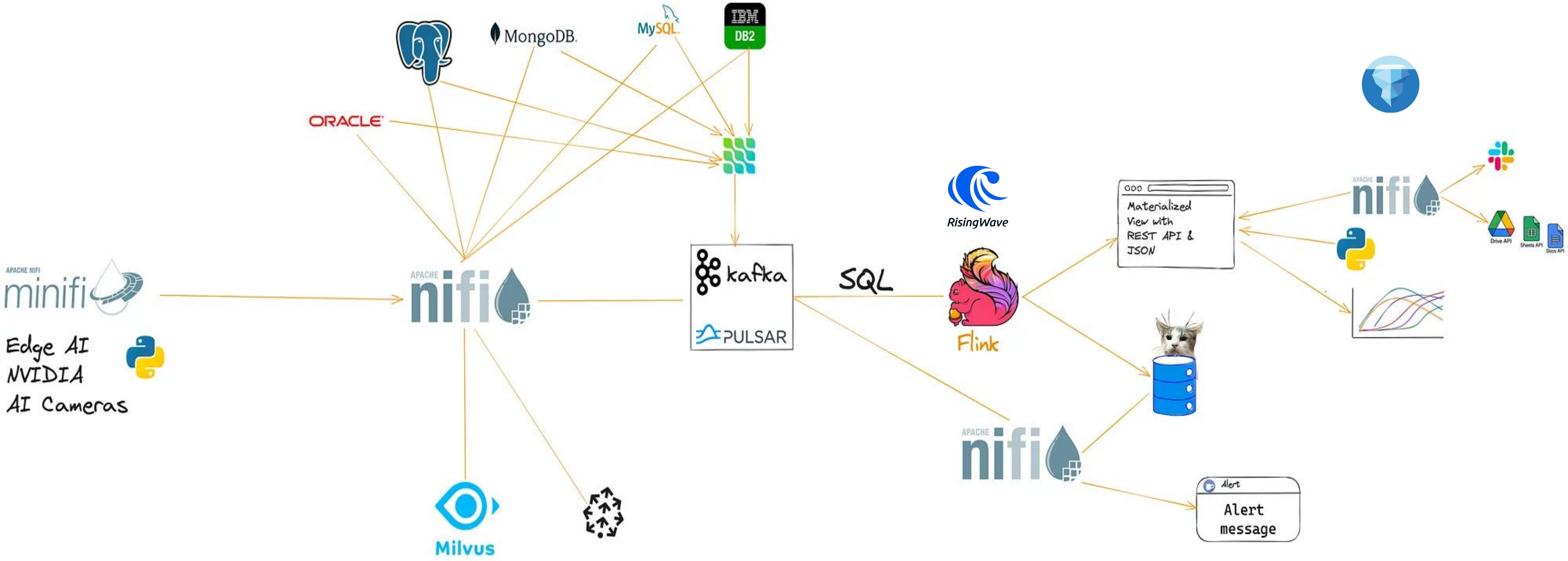
Vector DB

Applications/API's

Streams

Materialized Views

Open Data Lakehouse



Edge AI
NVIDIA
AI Cameras

Materialized
View with
REST API &
JSON

Alert
message

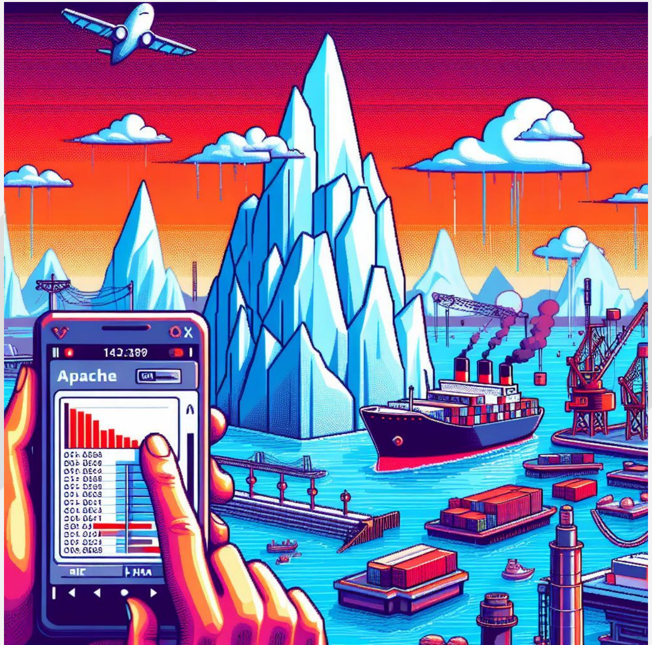
FLaNK for Halifax Canada Transit — NiFi, Kafka, Flink, SQL, GTFS-RT | by Tim Spann | Cloudera | Dec, 2023 | Medium

Never Get Lost in the Stream. NiFi-Kafka-Flink for getting to work… | by Tim Spann | Cloudera | Dec, 2023 | Medium

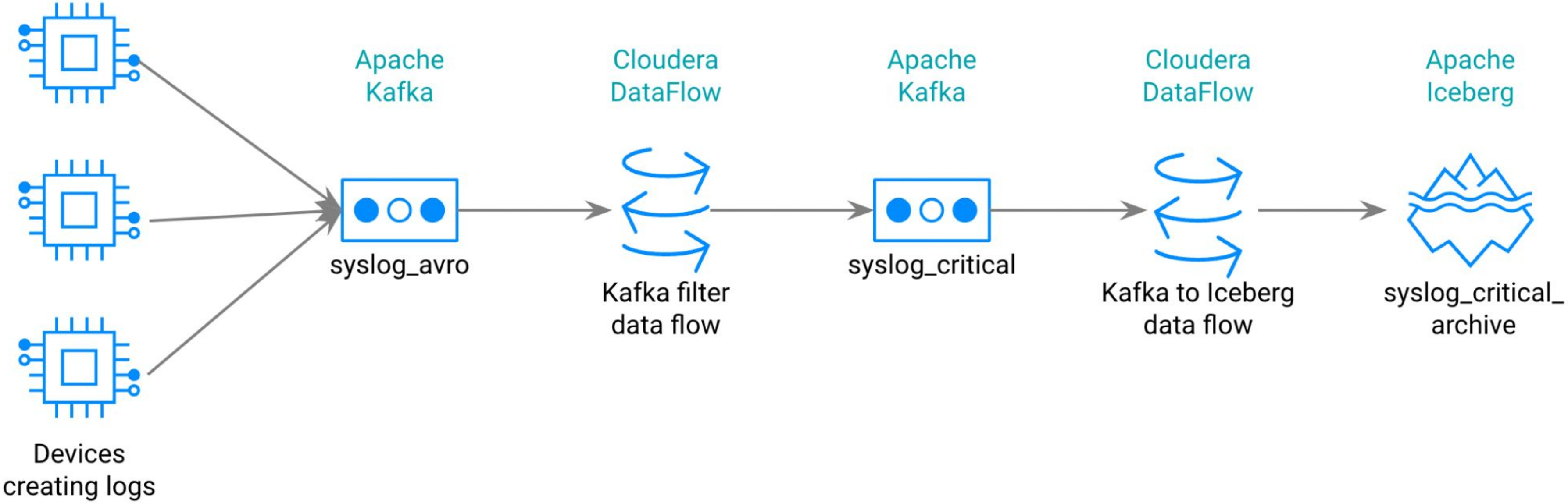
Iteration 1: Building a System to Consume All the Real-Time Transit Data in the World At Once | by Tim Spann | Cloudera | Medium

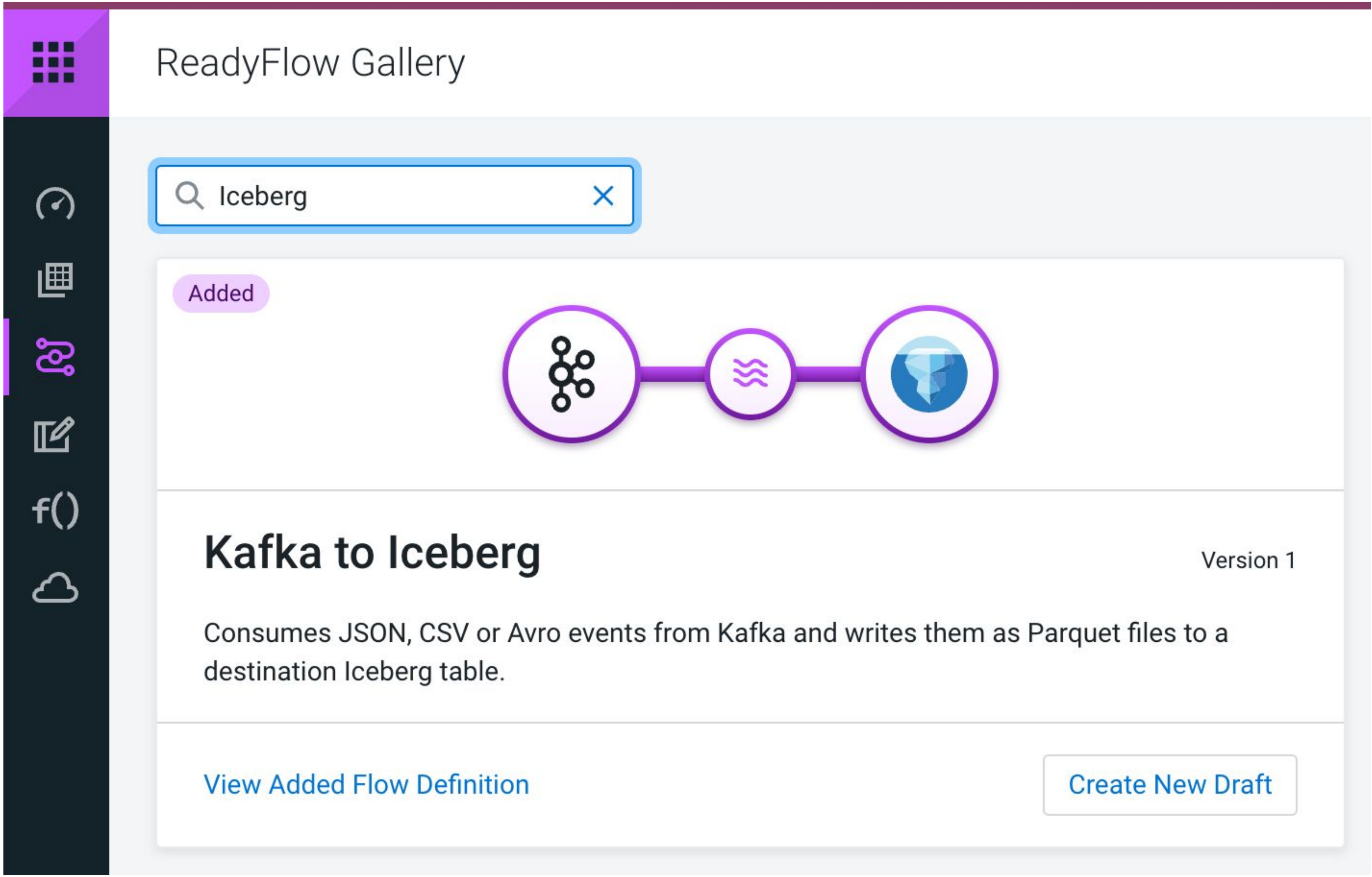
Watching Airport Traffic in Real-Time | by Tim Spann | Cloudera | Medium

APACHE ICEBERG



ICEBERG

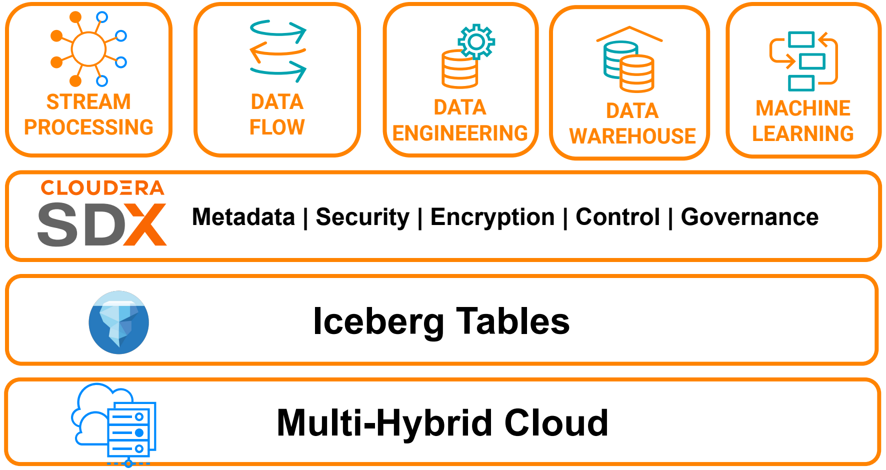

Apache
Kafka
Cloudera
DataFlow
Apache
Kafka
Cloudera
DataFlow
Apache
Iceberg
syslog_avro
Kafka filter
data flow
syslog_critical
Kafka to Iceberg
data flow
syslog_critical_
archive
Devices
creating logs



Kafka to Iceberg

Consumes JSON, CSV or Avro events from Kafka and writes them as Parquet files to a destination Iceberg table.

Cloudera’s Open Data Lakehouse



STREAM PROCESSING | DATA FLOW | DATA ENGINEERING | DATA WAREHOUSE | MACHINE LEARNING

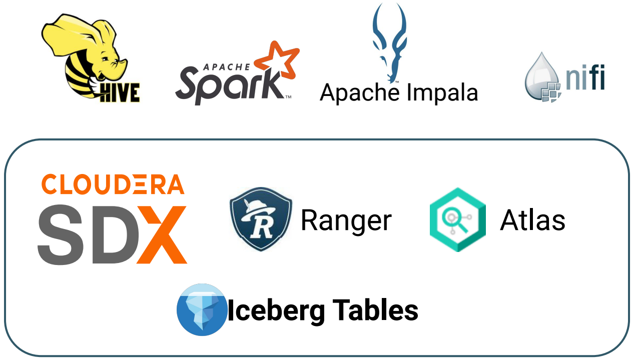
CLOUDERA SDX: Metadata | Security | Encryption | Control | Governance

Iceberg Tables

Multi-Hybrid Cloud

- □ Multi-function analytics for Streaming, Data Engineering, Data Warehouse and AI/ML with integrated data services
- □ Common security and governance policies and data lineage with SDX integration
- □ Common dataset with all CDP analytics engines without data duplication and movement
- □ Deployment freedom with Multi-Hybrid Cloud

Compute Engine Interoperability & SDX Integration



- **Snapshot isolation** ensures **consistent** data access and processing with various compute engines including **Hive**, **Spark**, **Impala** and **Nifi**

- **Security & Governance** support (e.g. FGAC) through **Ranger** integration

- Data **lineage** support through **Atlas** integration

tspann
LOG OUT

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

DATAFLOW
APACHE NIFI

Apache NiFi - developed 17 years ago by the NSA

2006
NiagaraFiles (NiFi) was first incepted
at the National Security Agency (NSA)

November 2014
NiFi is donated to the Apache Software Foundation (ASF) through NSA's Technology Transfer Program and enters ASF's incubator.

July 2015
NiFi reaches ASF top-level project status

Apache NiFi in a few numbers

A very active project with a dynamic community & comparison with ACEU 2019

2800+ members on the Slack channel (535+ - 4 years ago)

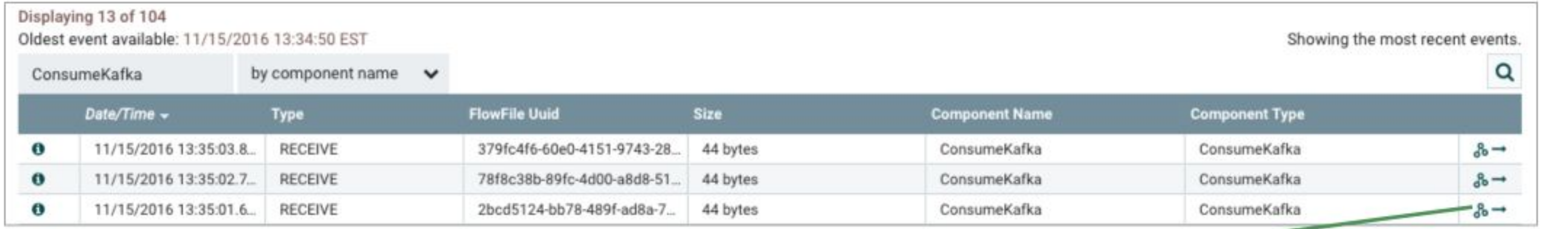
475+ contributors on Github across the repositories (260+ - 4 years ago)

65 committers in the Apache NiFi community (45 - 4 years ago)

Apache NiFi 1.25.0 is the latest release, NiFi 2.0.0-M2 is in alpha.

14M+ docker pulls of the Apache NiFi image (1M+ - 4 years ago)

PROVENANCE



Displaying 13 of 104

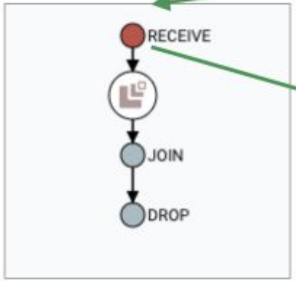
Oldest event available: 11/15/2016 13:34:50 EST

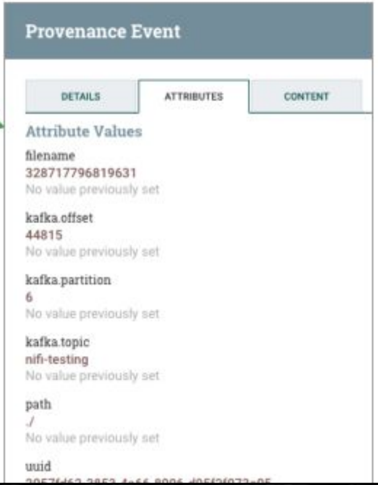
Showing the most recent events.

ConsumeKafka | by component name

	Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type	
ℹ	11/15/2016 13:35:03.8...	RECEIVE	379fc4f6-60e0-4151-9743-28...	44 bytes	ConsumeKafka	ConsumeKafka	→
ℹ	11/15/2016 13:35:02.7...	RECEIVE	78f8c38b-89fc-4d00-a8d8-51...	44 bytes	ConsumeKafka	ConsumeKafka	→
ℹ	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

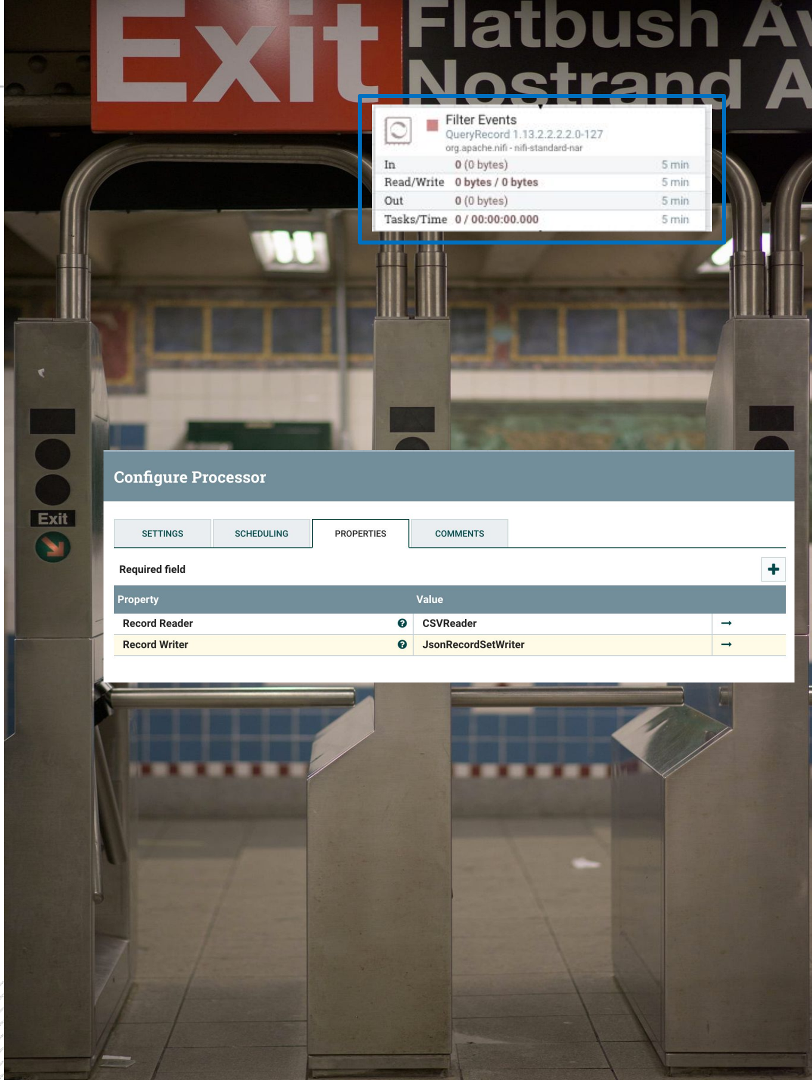
kafka.topic
nifi-testing
No value previously set

path
./
No value previously set

uuid

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.



Filter Events
QueryRecord 1.13.2.2.2.2.0-127
org.apache.nifi - nifi-standard-nar

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

Configure Processor

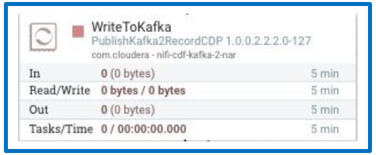
Required field

Property	Value
Record Reader	CSVReader
Record Writer	JsonRecordSetWriter

CLOUDERA

APACHE KAFKA

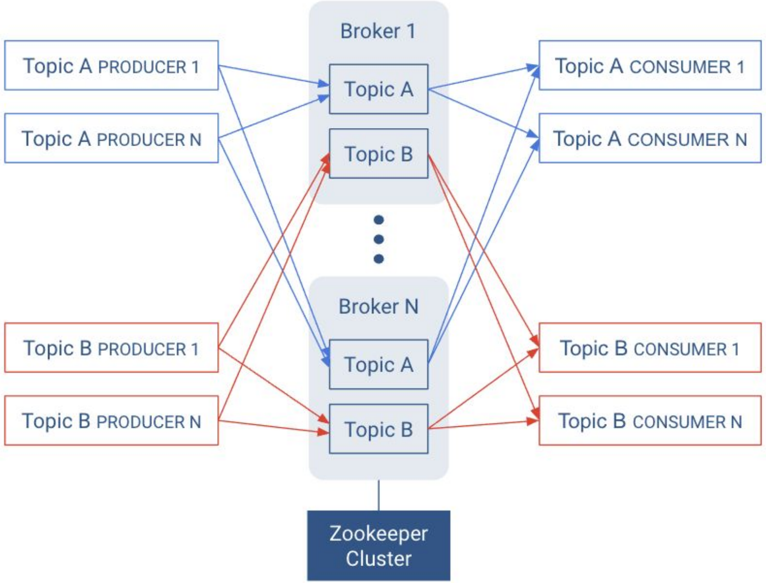
STREAMS MESSAGING WITH KAFKA

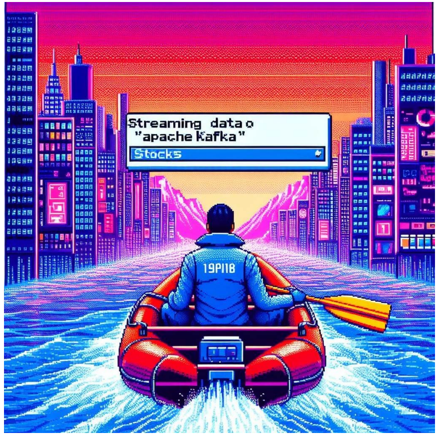


WriteToKafka
PublishKafka2RecordCDP 1.0.0.2.2.2.0-127
com.cloudera - nifi-cdf-kafka-2-nar

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

- Highly reliable distributed messaging system.
- Decouple applications, enables many-to-many patterns.
- Publish-Subscribe semantics.
- Horizontal scalability.
- Efficient implementation to operate at speed with big data volumes.
- Organized by topic to support several use cases.



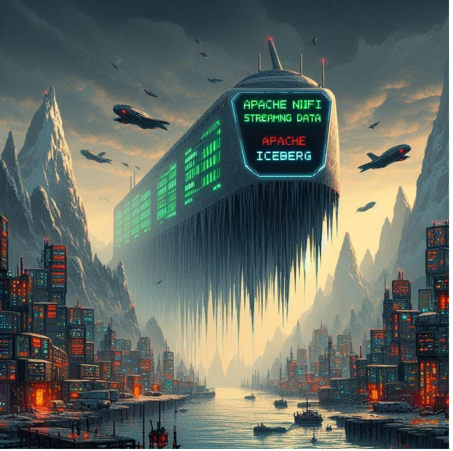

Streaming data o
"apache Kafka"
Stocks
19PII8

DEMO

I Can Haz Data?

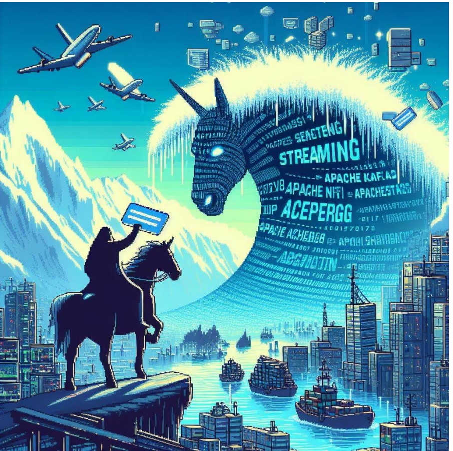
APACHE KAFKA

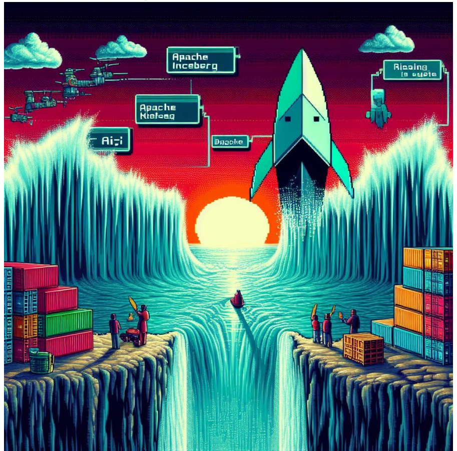
https://medium.com/@tspann/cdc-not-cat-data-capture-e43713879c03

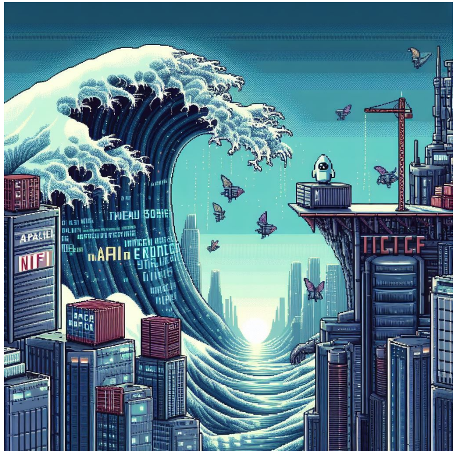


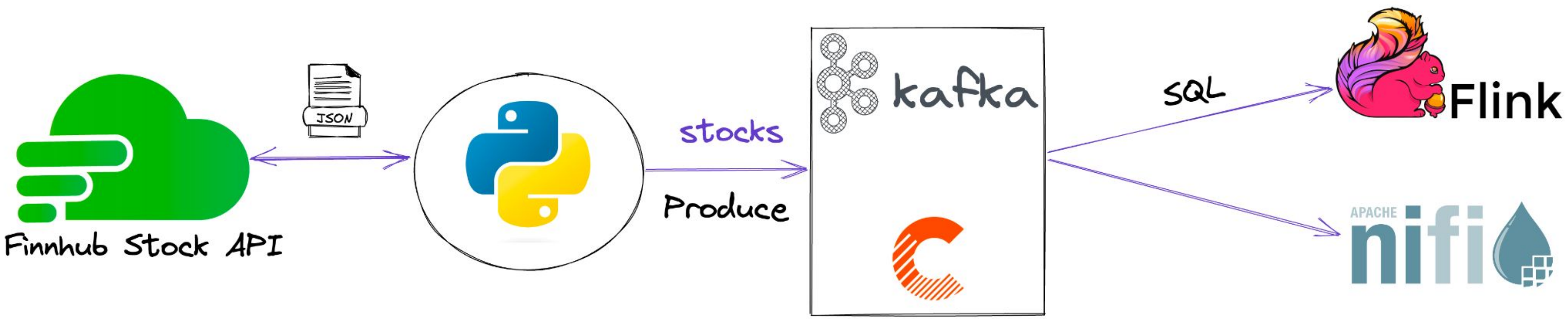
nifi

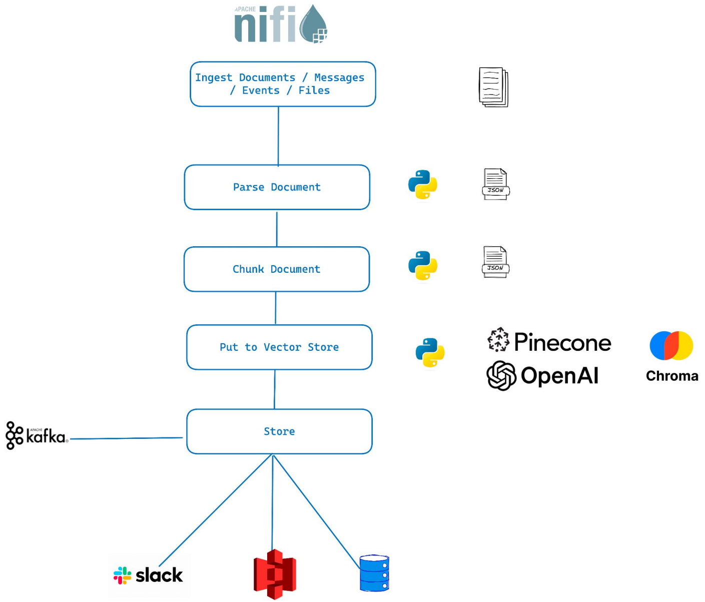
ICEBERG









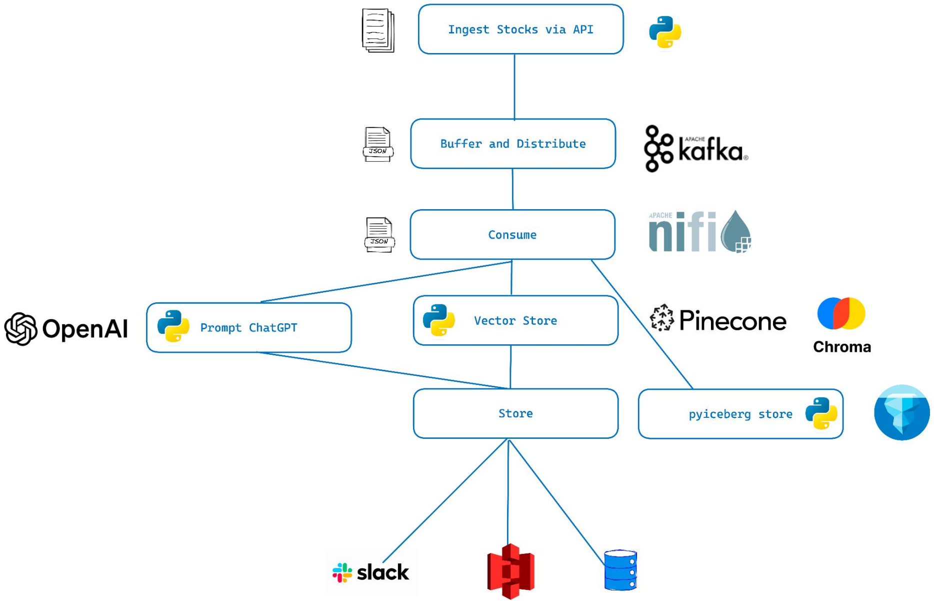

Ingest Documents / Messages / Events / Files

Parse Document

Chunk Document

Put to Vector Store

Store


Ingest Stocks via API

Buffer and Distribute

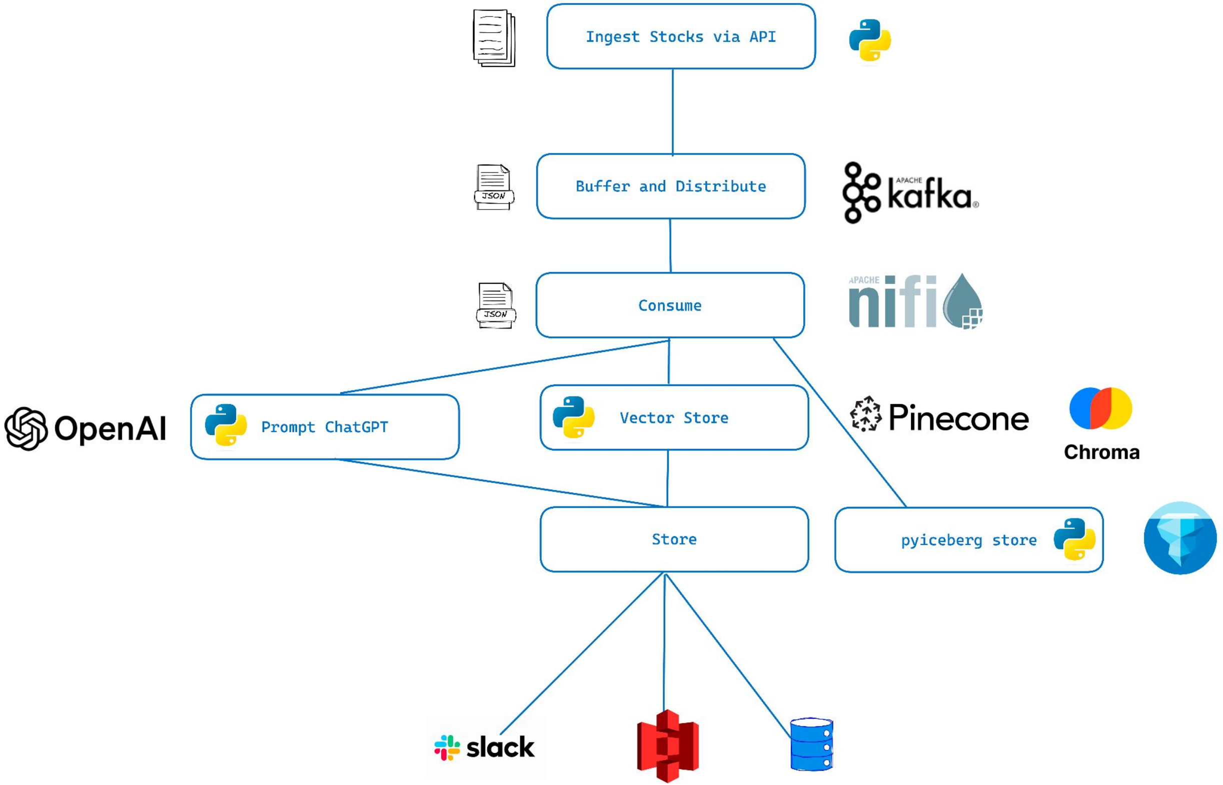
Consume

Prompt ChatGPT

Vector Store

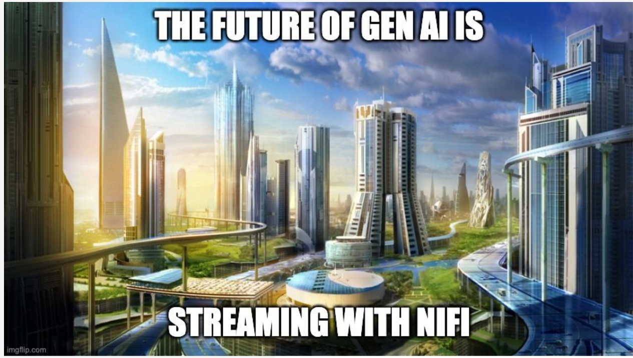
Store

pyiceberg store



https://github.com/tspannhw/PaK-Stocks
https://github.com/tspannhw/FLaNK-Py-Stocks
https://medium.com/cloudera-inc/let-nifi-worry-about-those-stocks-for-you-57d5f16b5e6b

APACHE
APACHE
APACHE


THE FUTURE OF GEN AI IS
STREAMING WITH NIFI
imgflip.com

THANK YOU

GEN AI NEEDS STREAMING
imgflip.com

APACHE NIFI
APACHE NIFI