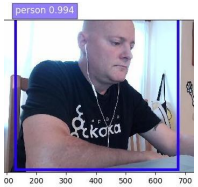
CLOUDERA

Codeless Generative AI Pipelines

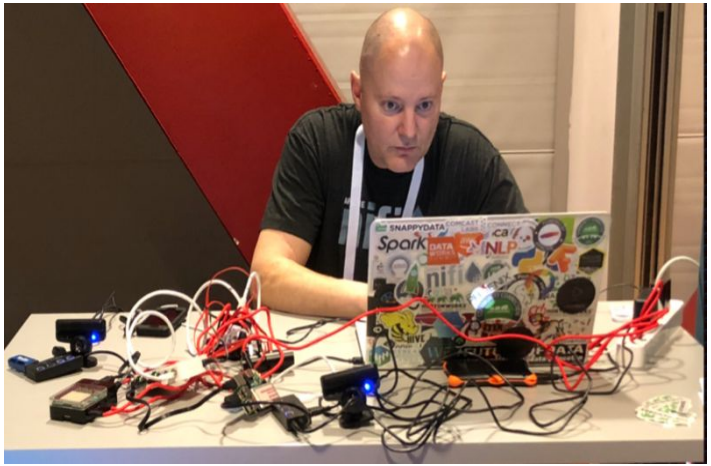
Tim Spann
Principal Developer Advocate

22 Feb 2024



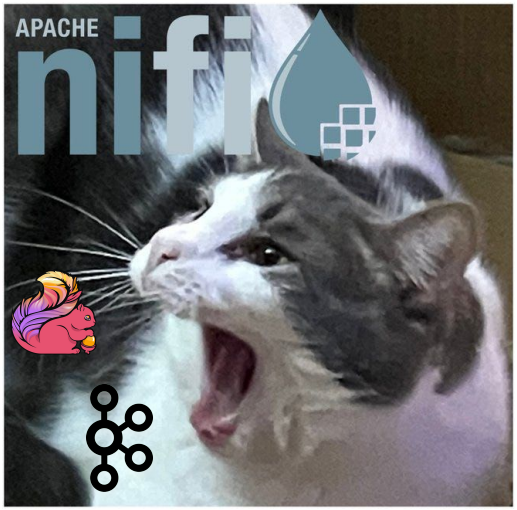
Microsoft NYC - Times Square

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

FLaNK Stack Weekly by Tim Spann

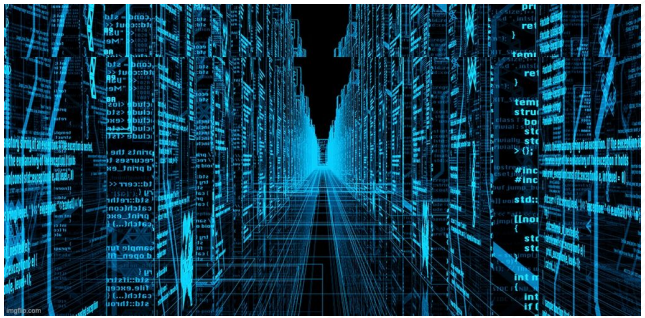


https://bit.ly/32dAJft

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.

Future of Data - NYC + NJ + Philly + Virtual



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

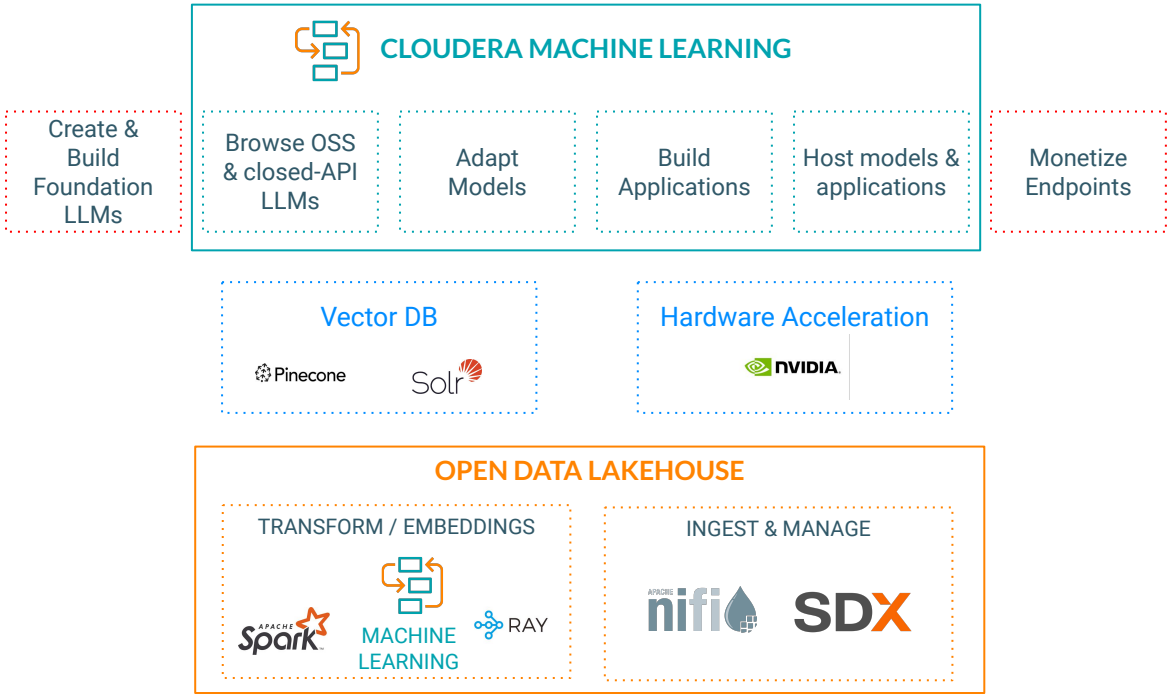
From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

HOW | BUILD AI WITH CDP

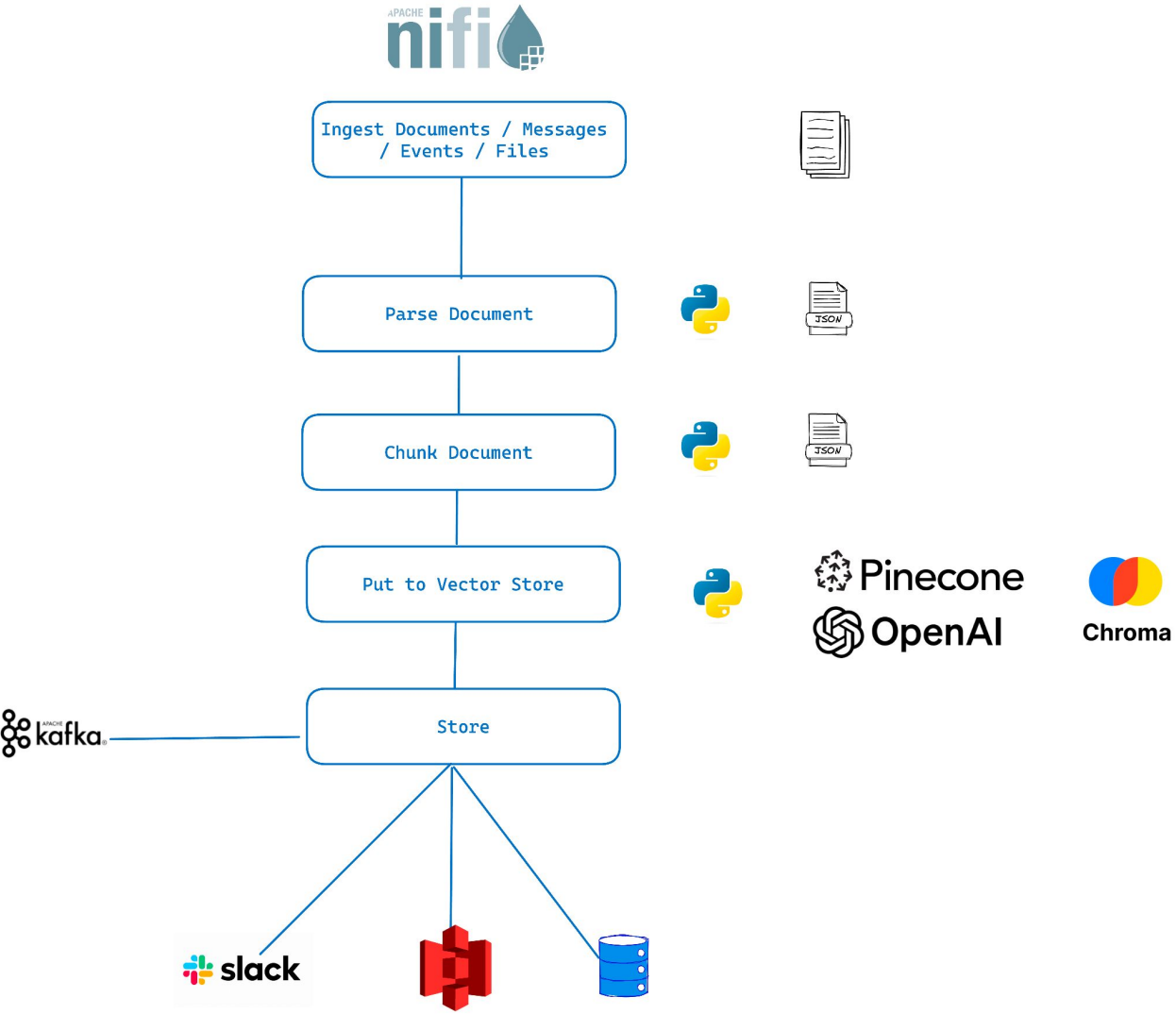
Enable the full Enterprise AI Application Lifecycle



*CML **enables more** developers to create **AI applications** using the latest models from Model Hubs*

*Leverage a broad spectrum of technology partners to **store embeddings** and **train models***

*CDP can manage the full data lifecycle on its **open data lakehouse** while SDX manages a secure **Data Fabric***



ECOSYSTEM PARTNERSHIPS

Best of breed capabilities for best in class Enterprise AI

FOUNDATION

- Widest range of Foundation Models
- Serverless integration with CDP for fast time to value

SEARCH

- Cloud-based semantic search made easy and at scale
- Store and manage AI representations of data in the public cloud

PERFORMANCE

- Optimized GPU performance & accelerated data science pipelines

TOOLING

- Access to open source innovation through CML AMPs
- Embedded into CML (Model Registry & Serving)

COMPUTE

- Tune, manage, scale AI models and applications
- Integrated into CML Sessions

Cloudera Generative AI Stack

Applied Machine Learning Prototypes (AMPs)

APPLICATIONS

CLOSED-SOURCE FOUNDATION MODELS

APIs: OpenAI (GPT-4 Turbo)
Amazon Bedrock: Anthropic (Claude 2), Cohere…

Hugging Face **MODEL HUBS**

FINE-TUNED MODELS

Meta (Llama 2) **OPEN SOURCE FOUNDATION MODELS**

Pinecone **MANAGED VECTOR STORE**

Milvus, Solr* **PRIVATE VECTOR STORE**

CLOUDERA Open Data Lakehouse

REAL-TIME DATA INGEST & ROUTING ▶ **DATA WRANGLING** ▶ **DATA STORE & VISUALIZATION** ▶ **AI MODEL TRAINING & SERVING**

CLOUD INFRASTRUCTURE

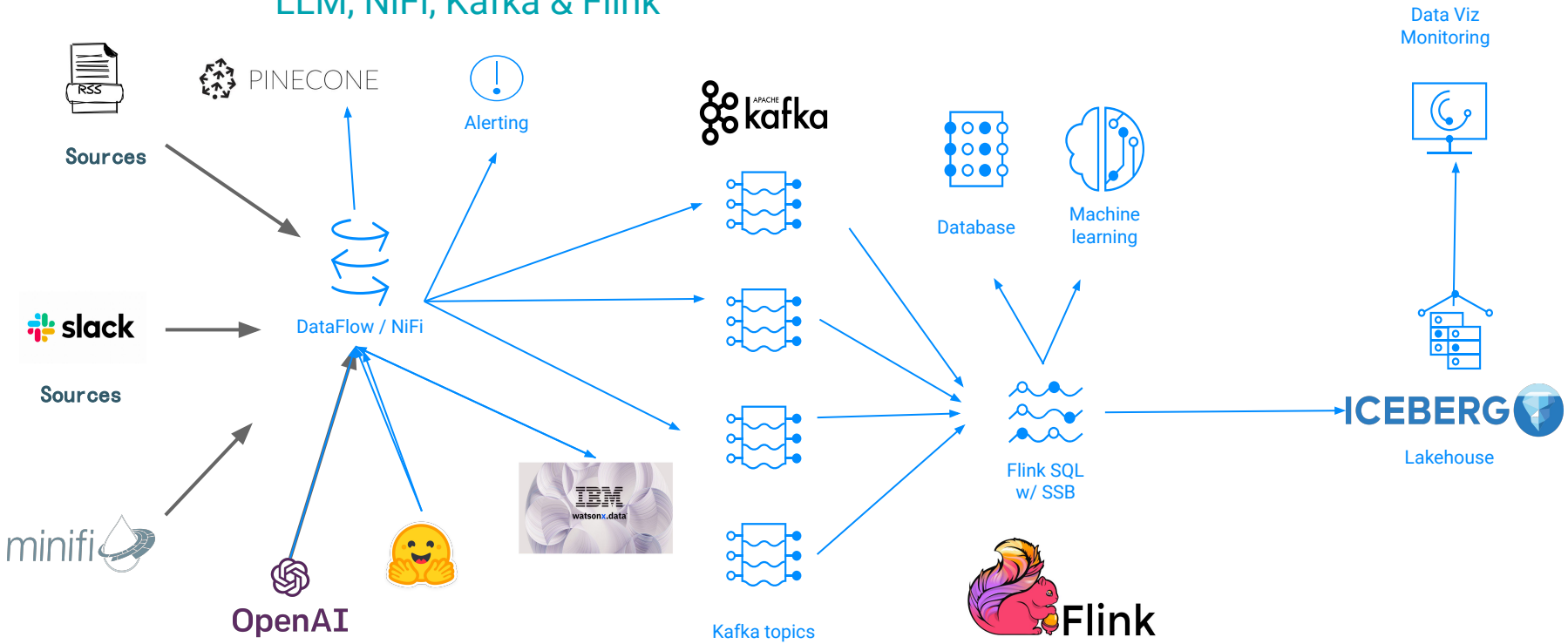
aws Azure Google Cloud Red Hat

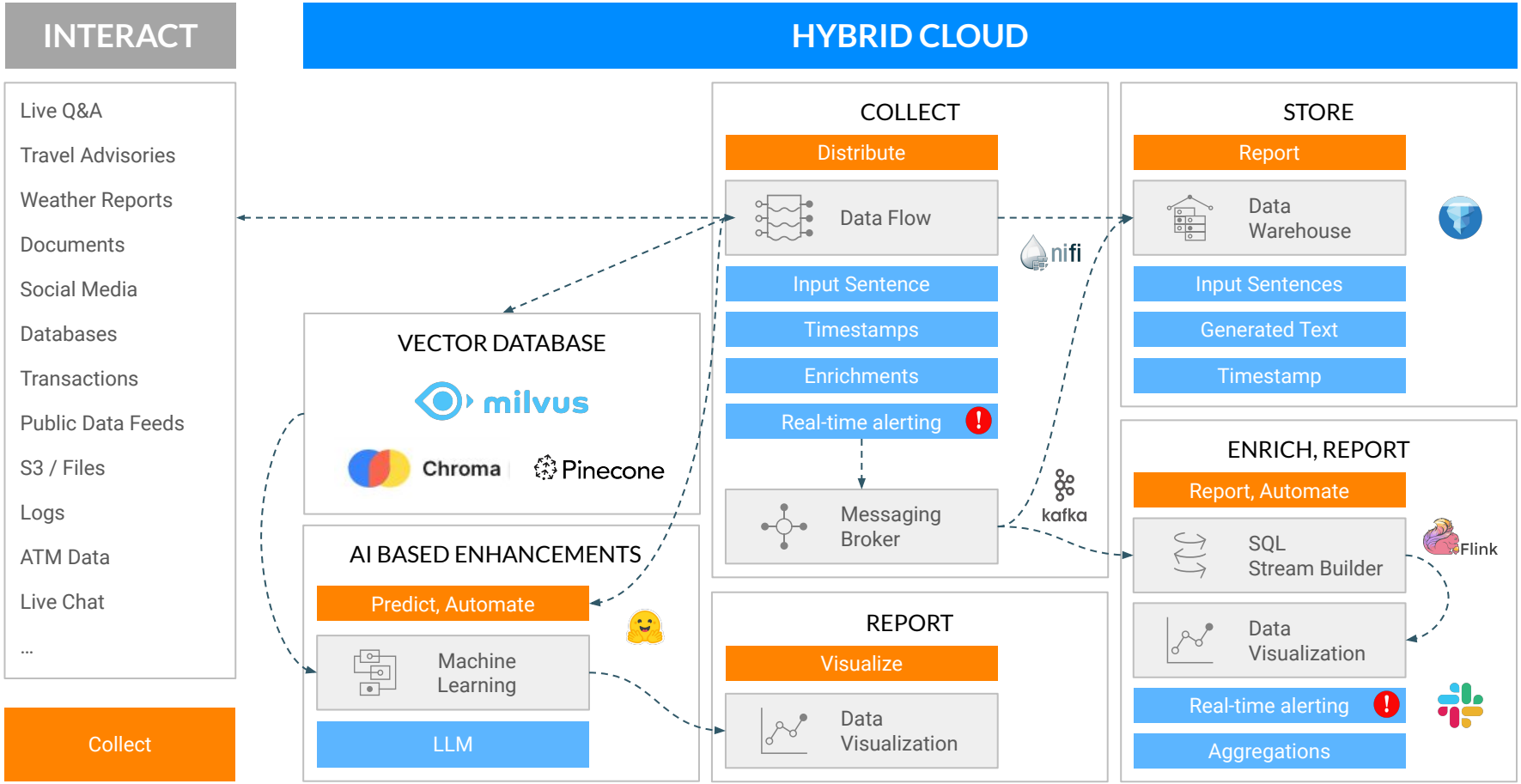
SPECIALIZED HARDWARE

NVIDIA Google AMD IBM intel DELL Technologies

Architecture in the context of Codeless GenAI Pipelines

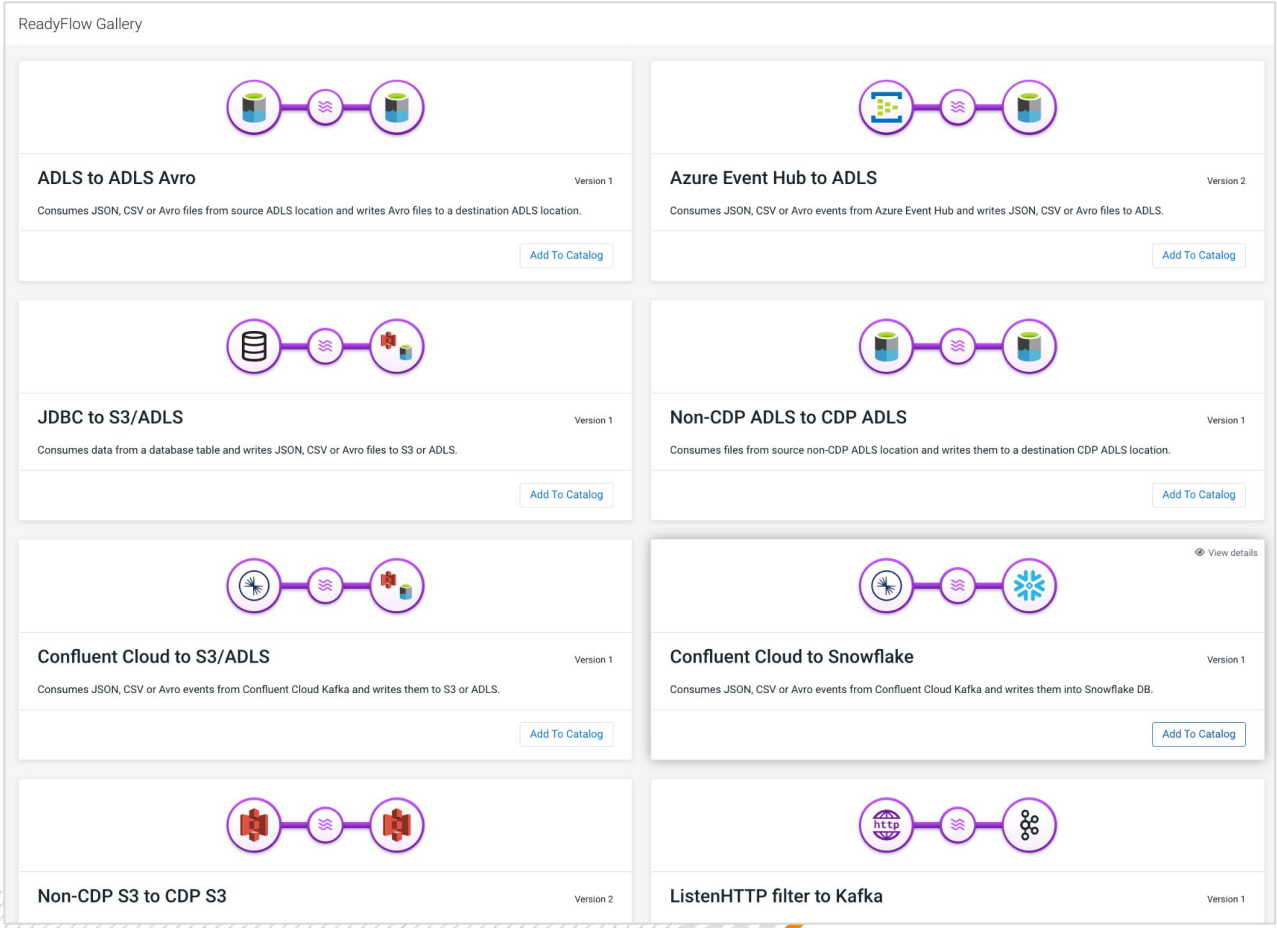
LLM, NiFi, Kafka & Flink





ReadyFlow Gallery

Leverage pre-built flow templates to quickly customize and deploy new data flows



Cloudera + LLMs

LLM Serving
Serving Framework

LLM Fine Tuning Process
Training Framework

Vector DB

Data Preparation
Data Engineering

Knowledge Repository
Data Storage / Management

CML RAY NVIDIA LangChain Streamlit python

CDF nifi

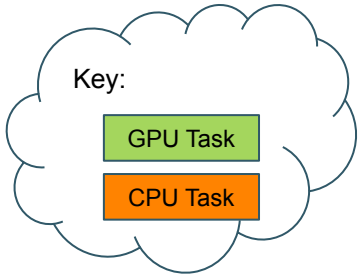
Streaming Classification
Real-Time Model Deployment

milvus Pinecone Chroma

CDE Spark python DASK RAY

minifi

CDP CLOUDERA



LLM USE CASE

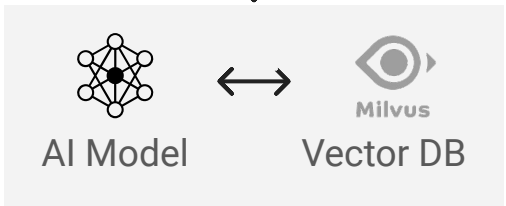
Unstructured file types

Structured Sources

Other enterprise data

Data in Motion on Cloudera Data Platform (CDP)

Capture, process & distribute any data, anywhere



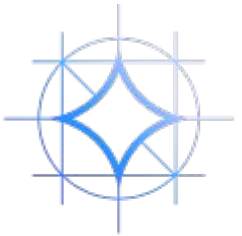
AI Model Vector DB

Applications/API's

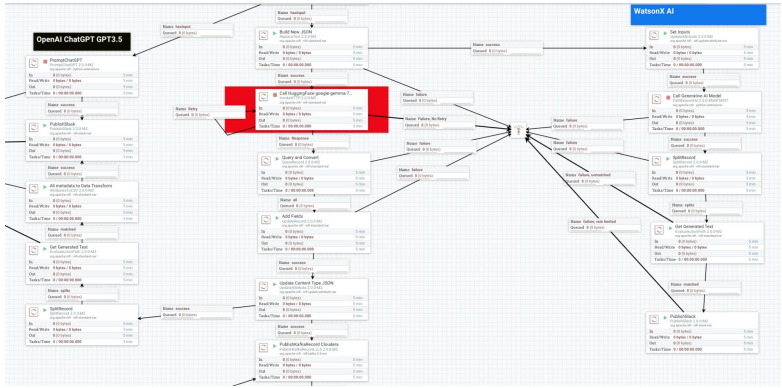
Streams

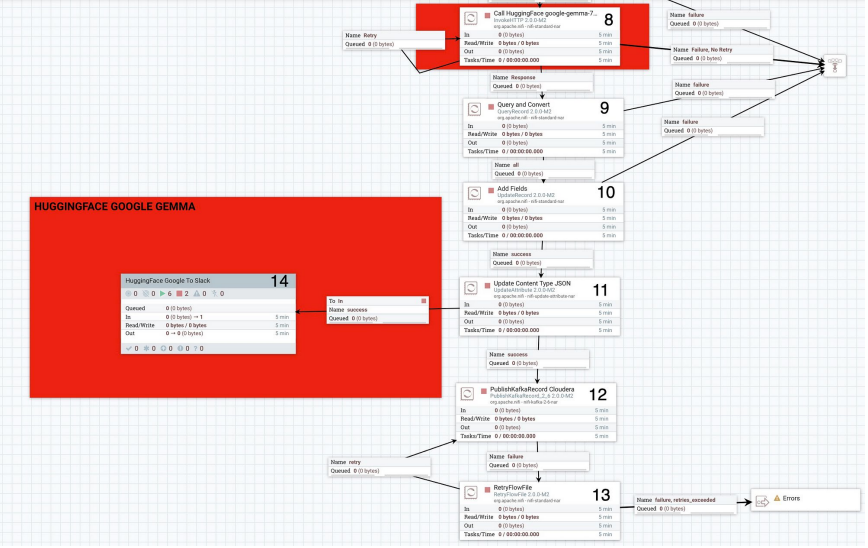
Materialized Views

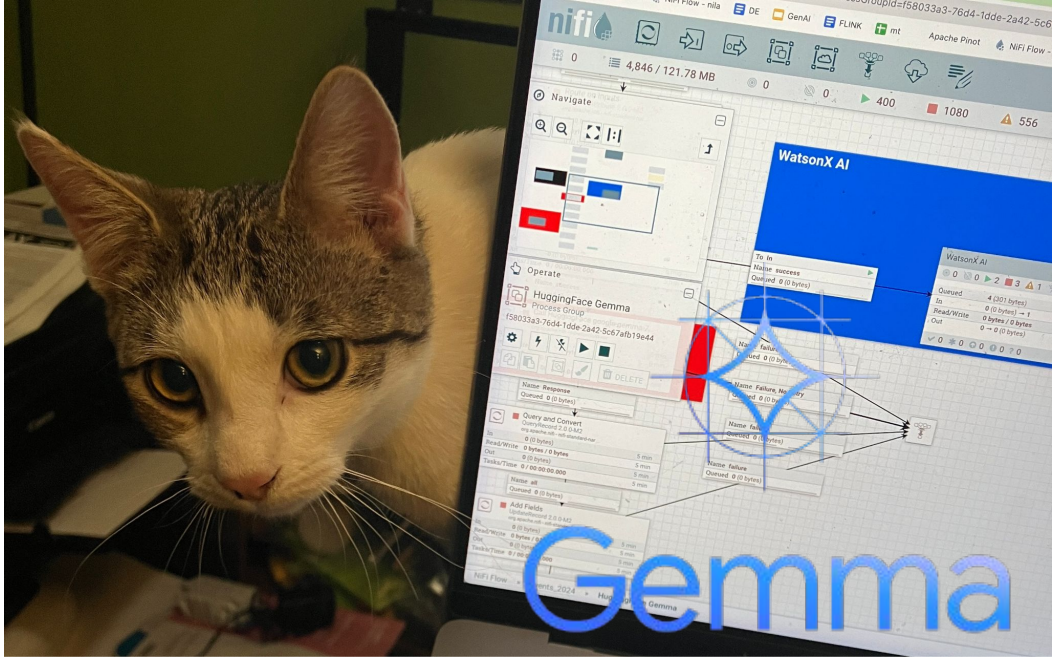
Open Data Lakehouse

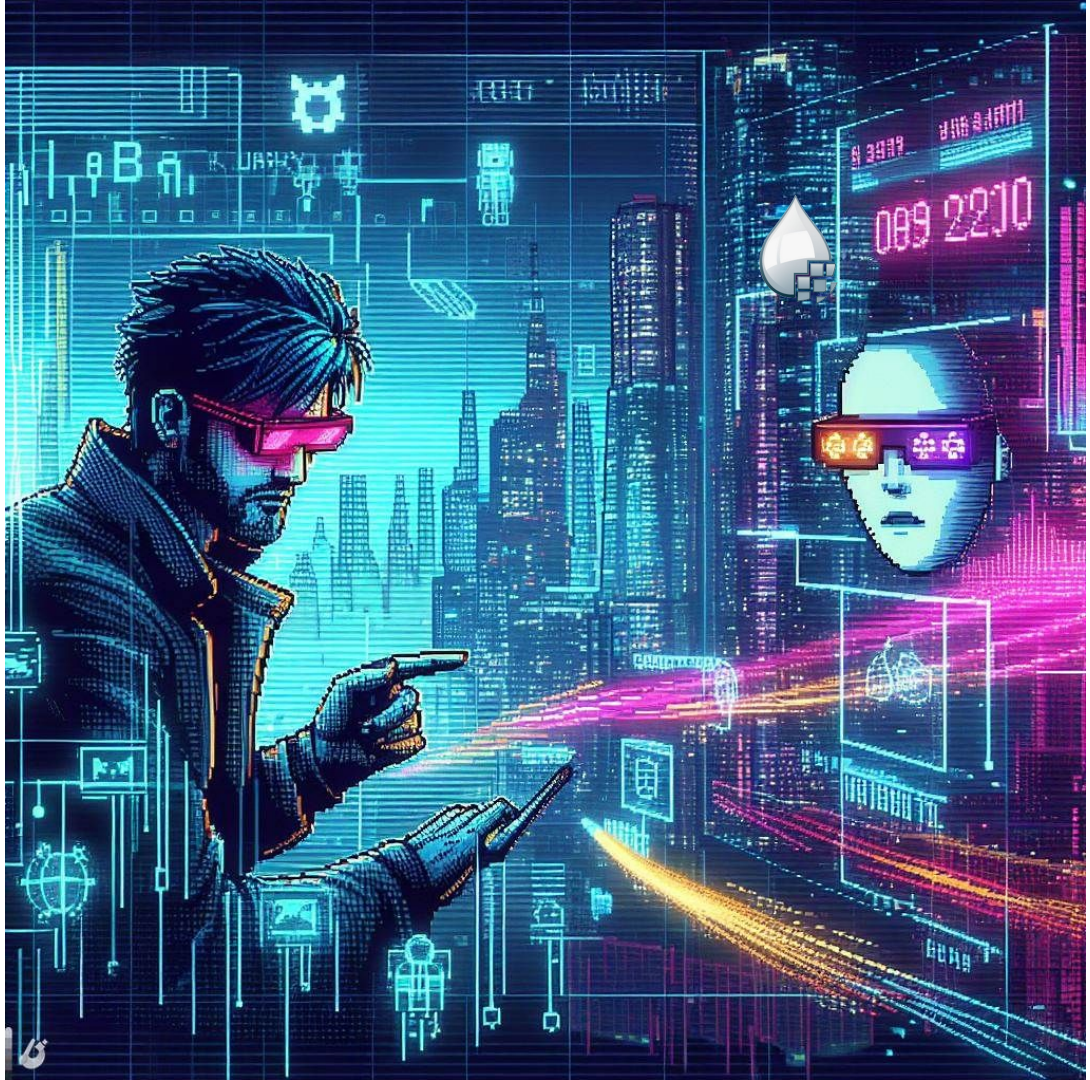


Gemma




HUGGINGFACE GOOGLE GEMMA
HuggingFace Google To Slack
14
Call HuggingFace google-gemma-7...
8
Query and Convert
9
Add Fields
10
Update Content Type JSON
11
PublishKafkaRecord Cloudera
12
RetryFlowFile
13
Errors


nifi
WatsonX AI
Gemma



DataFlow Pipelines Can Help

External Context Ingest

Ingesting, routing, clean, enrich, transforming, parsing, chunking and vectorizing structured, unstructured, semistructured, binary data and documents

Prompt engineering

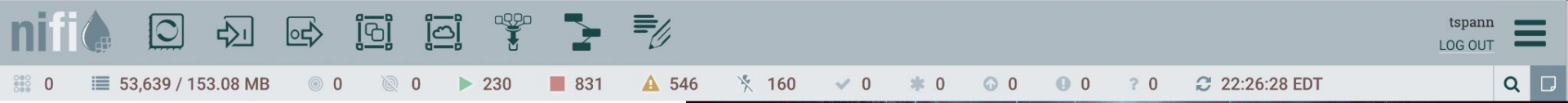
Crafting and structuring queries to optimize LLM responses

Context Retrieval

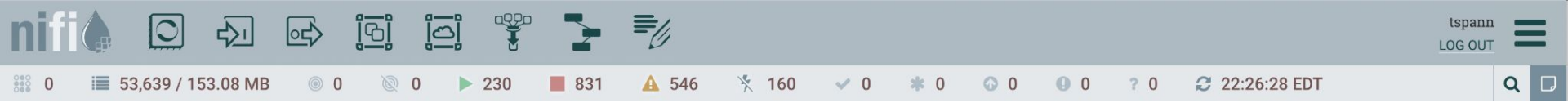
Enhancing LLM with external context such as Retrieval Augmented Generation (RAG)

Roundtrip Interface

Act as a Discord, REST, Kafka, SQL, Slack bot to roundtrip discussions



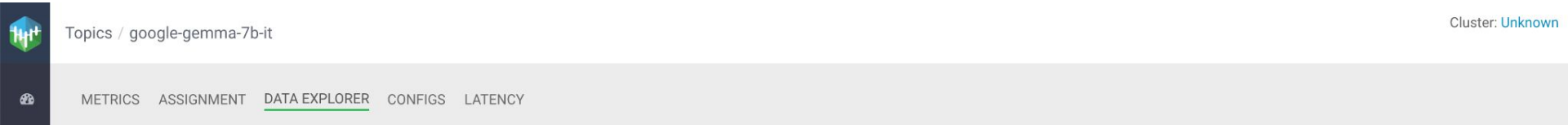
DATAFLOW APACHE NIFI



Apache NiFi in a few numbers

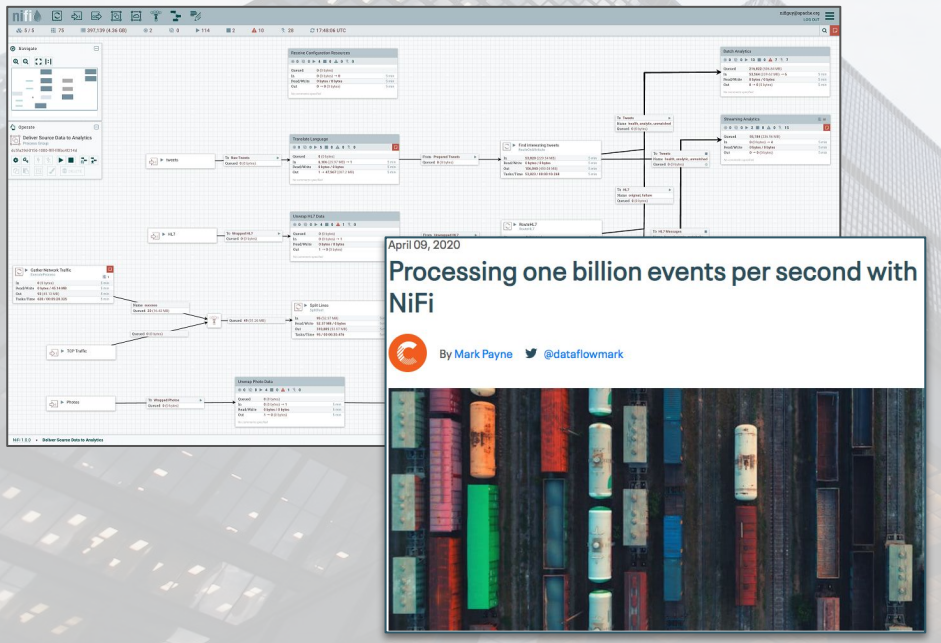
A very active project with a dynamic community & comparison with ACEU 2019

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)



CLOUDERA DATAFLOW - POWERED BY APACHE NiFi

Ingest and manage data from edge-to-cloud using a no-code interface



- #1 data ingestion/movement engine
- Strong community
- Product maturity over 11 years
- Deploy on-premises or in the cloud
- Over 400+ pre-built processors
- Built-in data provenance
- Guaranteed delivery
- Throttling and Back pressure

PROVENANCE

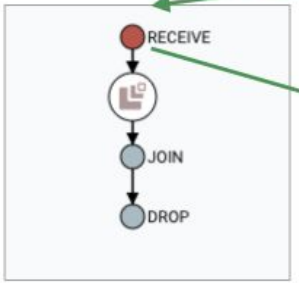
Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST

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

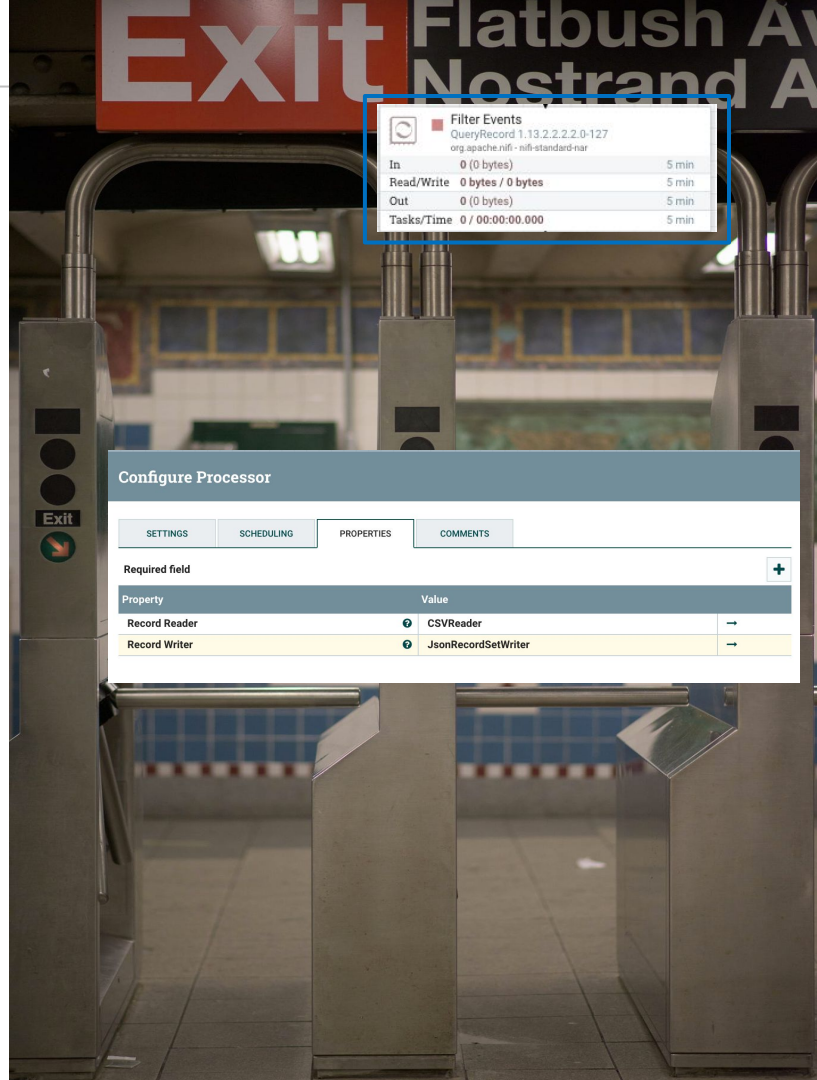
kafka.topic
nifi-testing
No value previously set

path
./
No value previously set

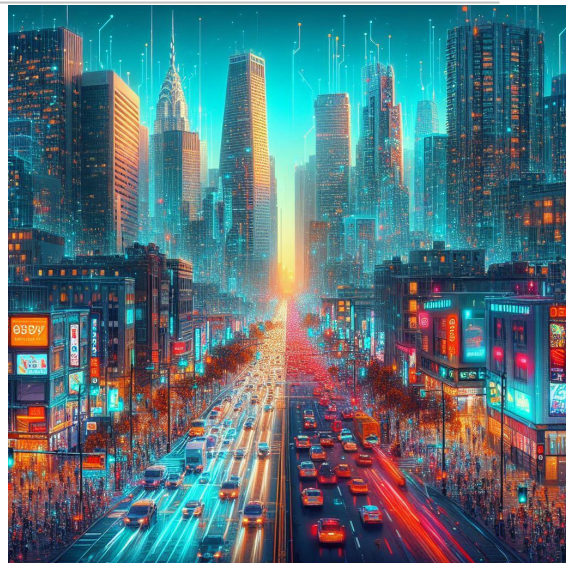
uuid

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.



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

CLOUD ML/DL/AI/Vector Database Services

- **Cloudera ML**
- **Amazon Polly, Translate, Textract, Transcribe, Bedrock, …**
- **Hugging Face**
- **IBM Watson X.AI**
- **Vector Stores Anywhere:** Pinecone, Milvus, ChromaDB, SOLR, …

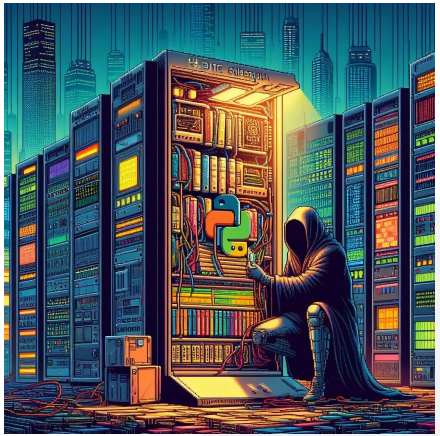
NiFi 2.0.0 Features

- Python Integration
- Parameters
- JDK 21+
- JSON Flow Serialization
- Rules Engine for Development Assistance
- Run Process Group as Stateless
- flow.json.gz

https://cwiki.apache.org/confluence/display/NIFI/NiFi+2.0+Release+Goals

https://medium.com/cloudera-inc/getting-ready-for-apache-nifi-2-0-5a5e6a67f450

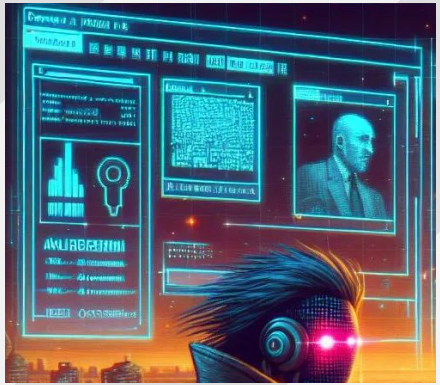


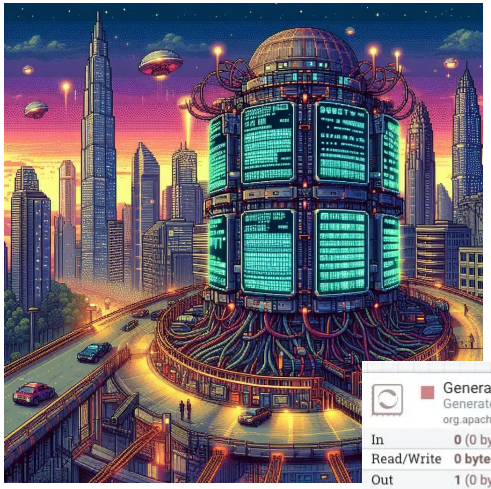



nifi

Python Processors

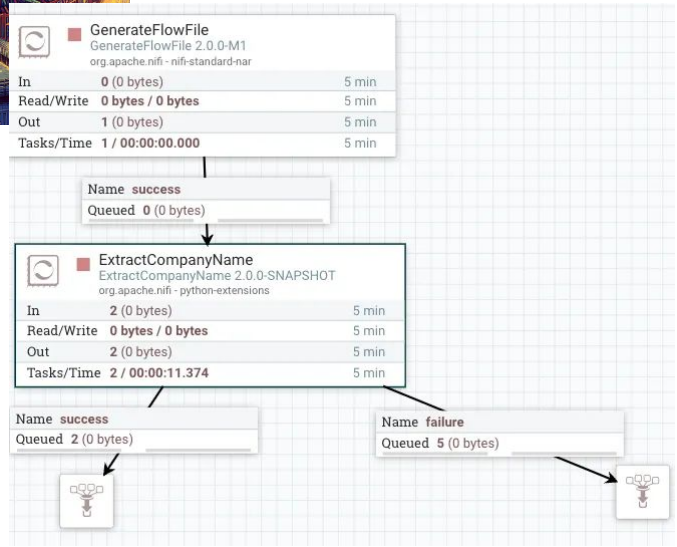






Extract Company Names

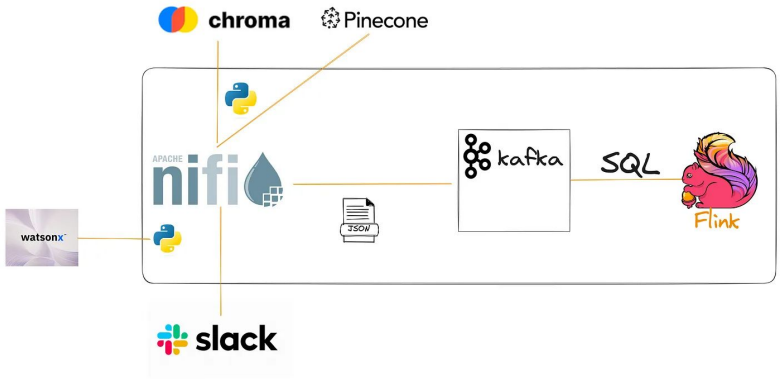
- Python 3.10+
- Hugging Face, NLP, SpaCY, PyTorch



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

WatsonX SDK To Foundation

- Python 3.10+
- LLM
- WatsonX.AI Foundation Models
- Inference
- Secure
- Official SDK from IBM



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

Other Python Processors

- Chunk Document, Parse Document
- Prompt Chat GPT
- PutChroma, QueryChroma
- PutPinecone, QueryPinecone

DEMO

[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]

CONF42 PYTHON 2024

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

BUILDING REALTIME AI APPLICATIONS WITH APACHE FLINK

February 28, 2024

5:30–7:30 PM EST

Timothy Spann
Principal Developer Advocate
Cloudera

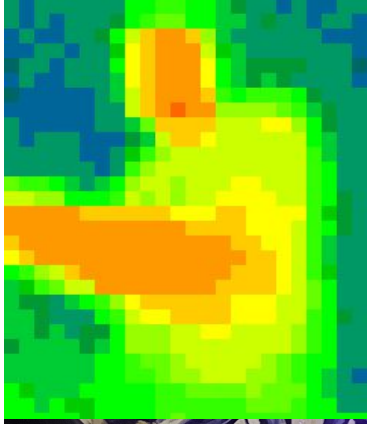
Matthias Broecheler
Founder
DataSQRL

starTree | CLOUDERA

IN-PERSON MEETUP

Discover Data Delights: A Slice of Real-Time Analytics and GenAI!

March 28 | 05:30 PM EST | NYC



THANK YOU

APACHE nifi