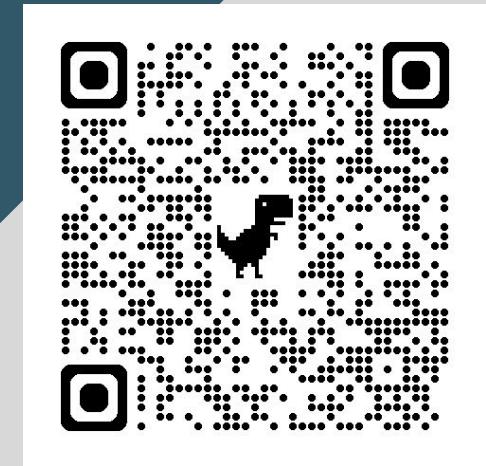
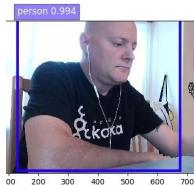




Adding Generative AI to Real-Time Streaming Pipelines

Tim Spann
Principal Developer Advocate

May 1, 2024





TIM SPANN

Twitter: @PaasDev // Blog: datainmotion.dev

Principal Developer Advocate. Field Engineer.
Princeton/NYC Future of Data Meetups.

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

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



 DZone REF CARDS TREND REPORTS EXPERTS

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...



FLaNK Stack Weekly by Tim Spann



<https://bit.ly/32dAJft>

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



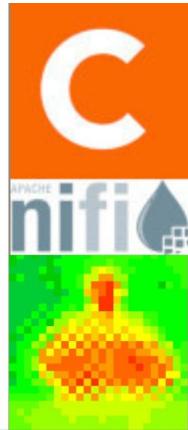
This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java, LLM, GenAI, Vector DB and Open Source friends.

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 ...



CLOUDERA



@PaasDev

RAPID INNOVATION IN THE LLM SPACE

Too much to cover today.. but you should know the common LLMs, Frameworks, Tools

Notable LLMs

Closed Models	Open Models
 OpenAI GPT3.5 GPT4	 Meta AI Llama2 Code Llama
 Claude2	 MISTRAL AI Mistral7B Mixtral8x7B

++ 100s more... check out the HuggingFace LLM Leaderboard (pretrained, domain fine-tuned, chat models, ...)

Popular LLM Frameworks

LangChain	Llamaindex
 LangChain Langchain is a framework for developing apps powered by LLMs <ul style="list-style-type: none">• Python and JavaScript Libraries• Provides modules for LLM Interface, Retrieval, & Agents	 Llamaindex Llamaindex is a framework designed specifically for RAG apps <ul style="list-style-type: none">• Python and JavaScript Libraries• Provides built in optimizations / techniques for advanced RAG

When to use one over the other? Use Langchain if you need a general-purpose framework with flexibility and extensibility. Consider Llamaindex if you're building a RAG only app (retrieval/search)

Open Source Vector DBs



Open Community & Open Models

Hugging Face	HuggingFace
 Hugging Face HuggingFace is an ML community for hosting & collaborating on models, datasets, and ML applications <ul style="list-style-type: none">• Latest open source LLMs are in HuggingFace• + great learning resources / demos https://huggingface.co/	

ENTERPRISE WIDE USE CASES FOR AN LLM



Enterprise Knowledge Base / Chatbot / Q&A

- Customer Support & Troubleshooting
- Enable open ended conversations with user provided prompts



Classification and Clustering

- Categorize and sort large volumes of data into common themes and trends to support more informed decision making.



Content Generation

- Provide detailed and contextually relevant prompts to develop outlines, brainstorm ideas and approaches for content.



Social and emotional sensing

- Gauge emotions and opinions based on a piece of text.
- Understand and deliver a more nuanced message back based on sentiment.



Code assistant:

- Provide relevant snippets of code as a response to a request written in natural language.
- Assist with creating test cases and synthetic test data.
- Reference other relevant data such as a company's documentation to help provide more accurate responses.



Document Summarization

- Distill large amounts of text down to the most relevant points.



Language Translation

- Globalize your content by feeding web pages through LLMs for translation.
- Combine with chatbots to provide multilingual support to your customer base.

WHICH MODEL AND WHEN?

Use the right model for right job: closed or open-source

	Closed Source	Most advanced AI models	Great for a wide range of tasks	Usage can easily scale but so can your costs	Compliance, privacy, and security risks
	Open Source	Rapidly improving AI models	Excel at more specialized tasks	Better cost planning	More control over where & how models are deployed

APPLICATIONS

CLOSED-SOURCE
FOUNDATION MODELS

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



MODEL HUBS
Hugging Face



FINE-TUNED MODELS

Meta (Llama 2)

**OPEN SOURCE
FOUNDATION MODELS**

**MANAGED
VECTOR STORE**

Milvus

**PRIVATE
VECTOR STORE**

CLOUDERA
Open Data Lakehouse



CLOUD INFRASTRUCTURE



SPECIALIZED HARDWARE



NVIDIA



DELL Technologies

CLOUDERA + LLMS

LLM Serving
Serving Framework

LLM Fine Tuning Process
Training Framework

Vector DB

Data Preparation
Data Engineering

Knowledge Repository
Data Storage / Management



milvus



CDP CLOUDERA

minifi

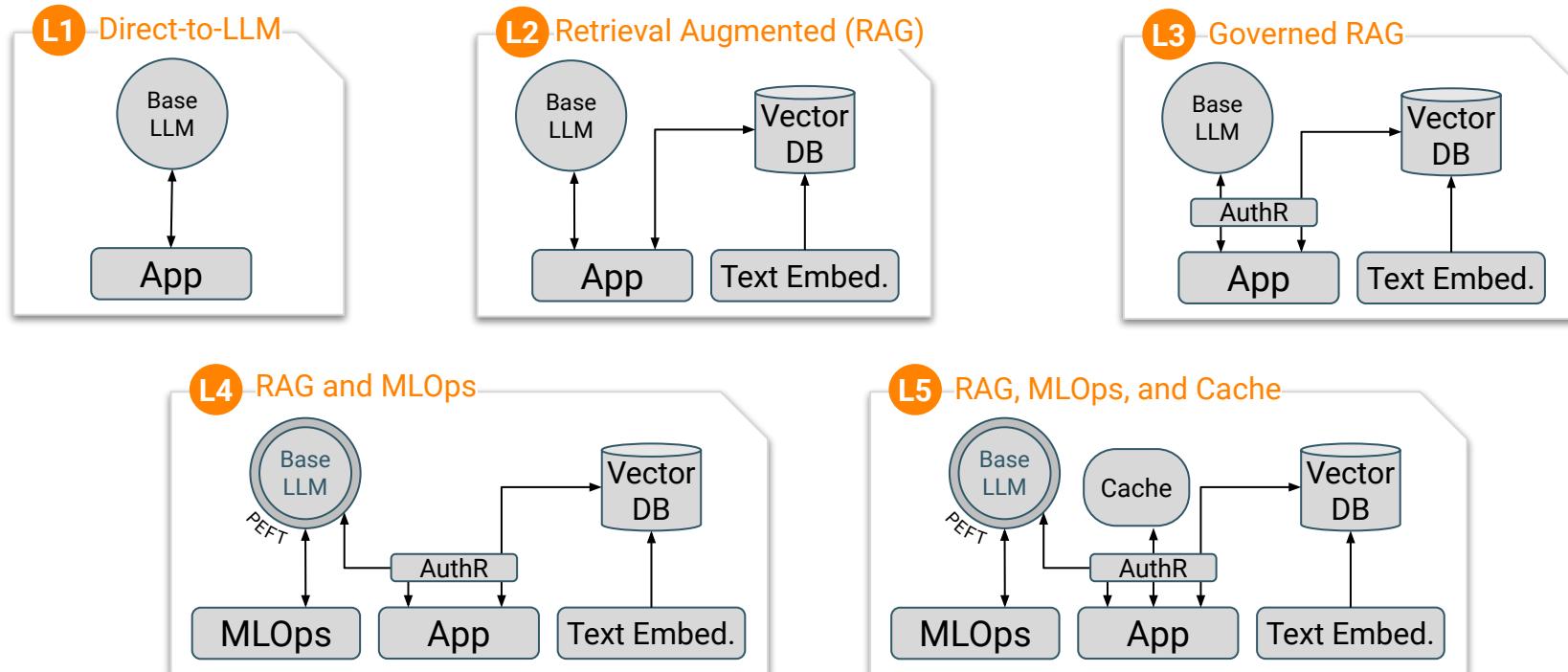


Key:

- GPU Task
- CPU Task

EVOLUTION OF LLM APP ARCHITECTURES

Choose a cost-effective platform that provides all of the necessary components and integrations



ML OPS IN CLOUDERA MACHINE LEARNING

Enabling Production ML At Scale



MODEL DEPLOYMENT & HA SERVING

- One-click deployment of models
- Robust and HA model serving infrastructure



MODEL & PREDICTION MONITORING

- UUID for each prediction
- Analyze metrics granularly to the feature level
- Ground truth to production environments



SHARED DATA EXPERIENCE FOR MODELS

- Automatic model cataloging & lineage
- Governed and secure production workflows



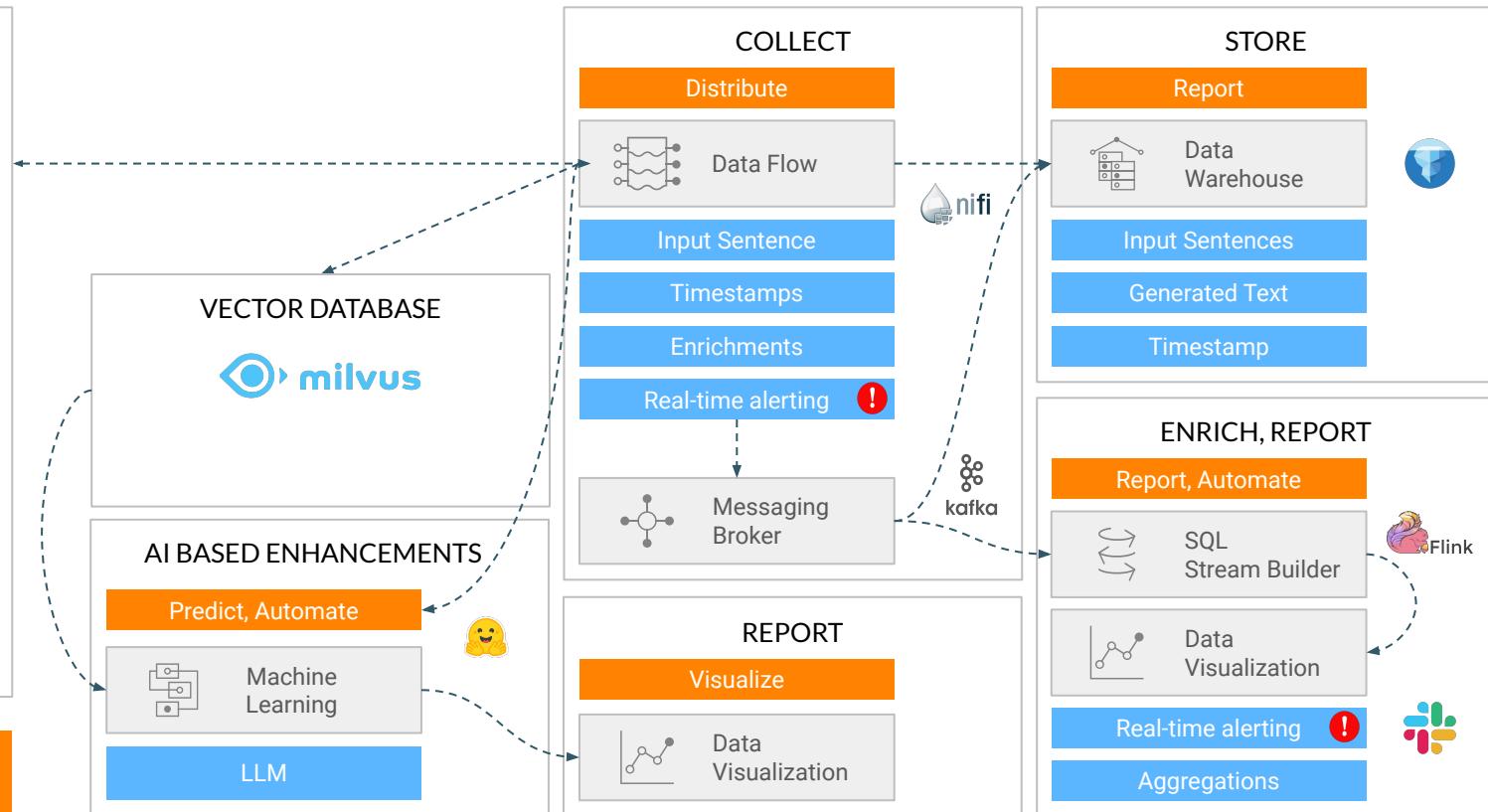
DISTRIBUTED AI COMPUTE

- Cutting Edge frameworks
- Support for Ray, Dask, Modin...

INTERACT

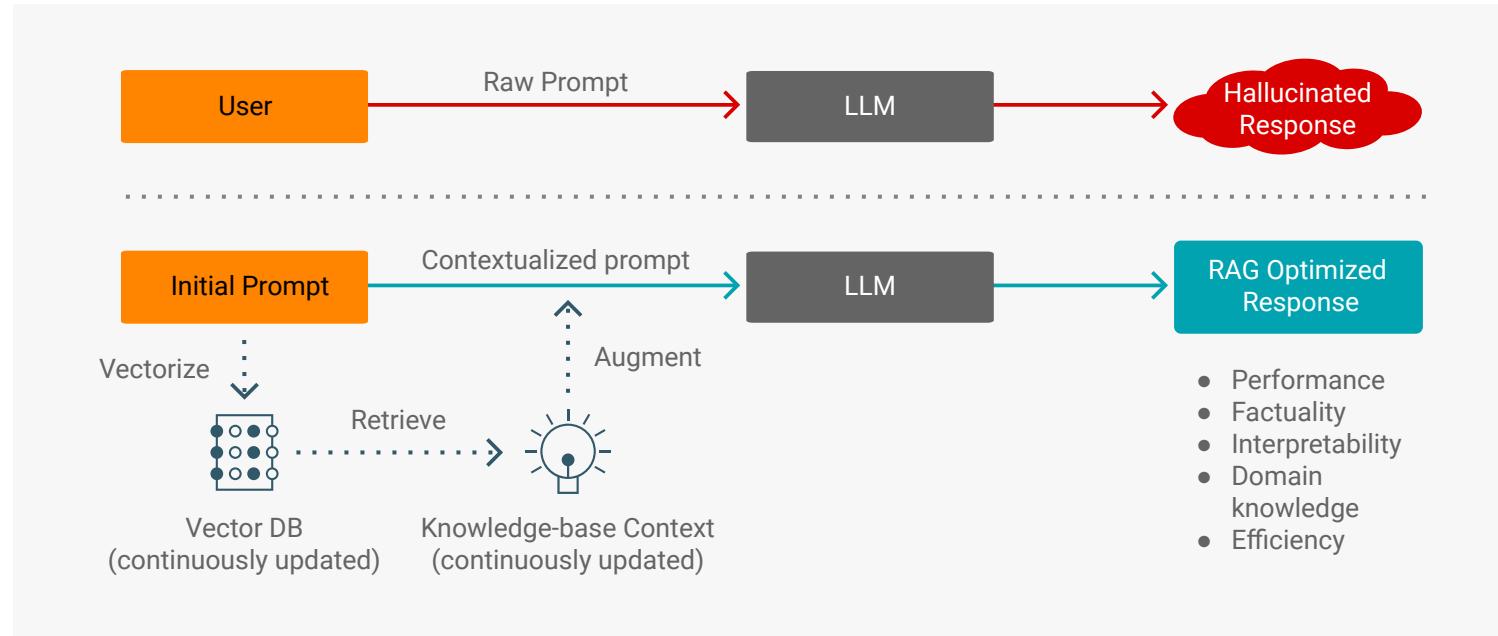
- Live Q&A
 - Travel Advisories
 - Weather Reports
 - Documents
 - Social Media
 - Databases
 - Transactions
 - Public Data Feeds
 - S3 / Files
 - Logs
 - ATM Data
 - Live Chat
 - ...
- Collect

HYBRID CLOUD



REAL-TIME CONTEXT FOR GEN AI

Classic RAG Architecture



Generative AI



NLP / AI / LLM





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

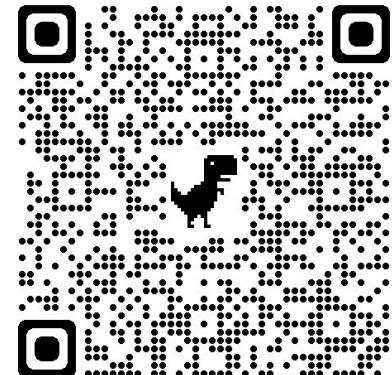
UNSTRUCTURED DATA WITH NIFI

- **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

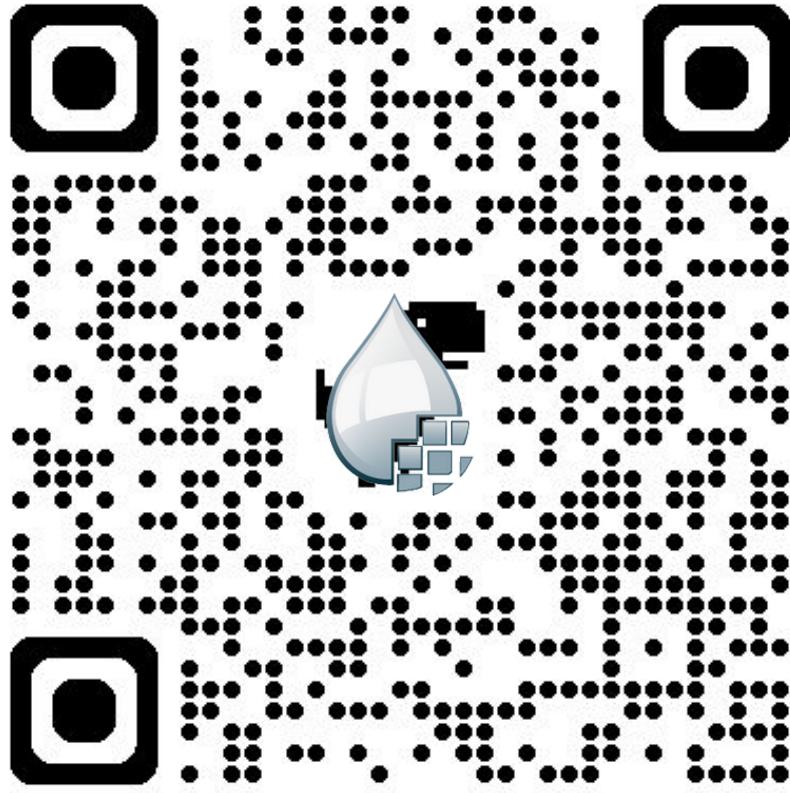


CLOUD ML/DL/AI/Vector Database Services

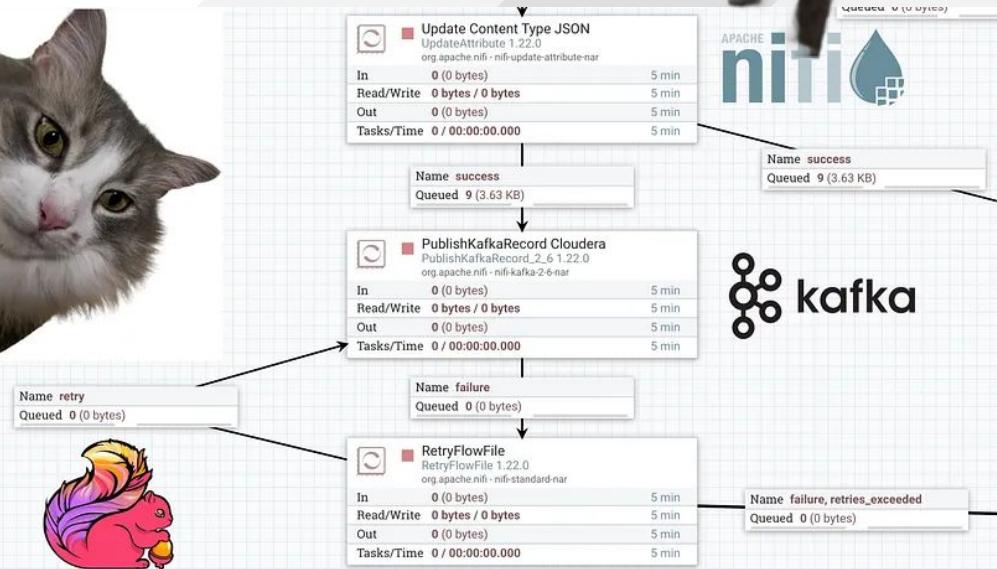
- Cloudera ML
- Amazon Polly, Translate, Textract, Transcribe, Bedrock, ...
- Hugging Face
- IBM Watson X.AI
- Vector Stores Anywhere: Milvus, ...



<https://medium.com/@tspann/building-a-milvus-connector-for-nifi-34372cb3c7fa>

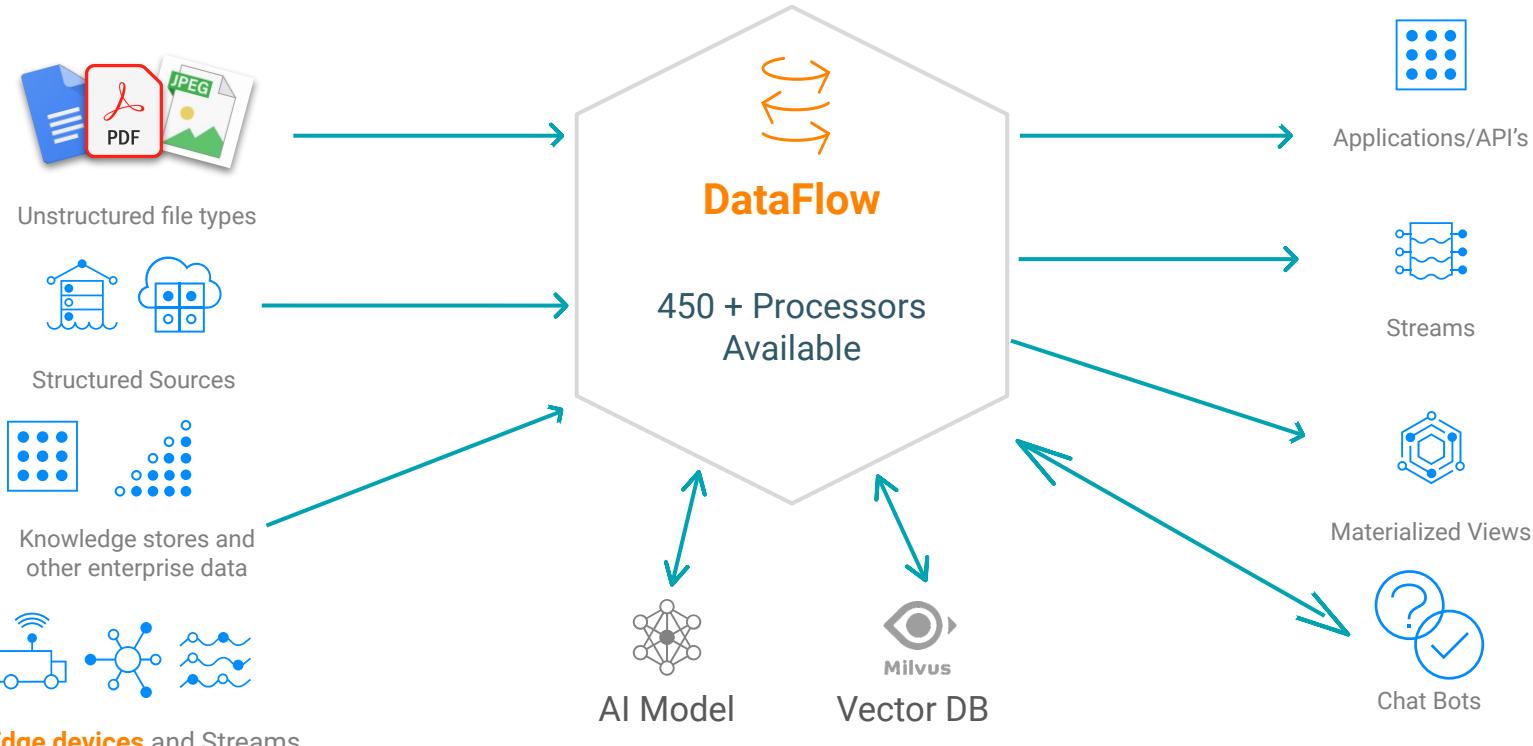


CLOUDERA DATAFLOW



CAPTURE ALL DATA

DataFlow Has a Vast Library of Processors to connect to ANYTHING



CAPTURE ALL DATA

DataFlow is built for MultiModal Data

Traditional ML pipelines:

- Structured
- Batch/Microbatch
- Highly engineered feature sets
- Clearly labeled data
- Metrics, KPI's, etc

Gen AI pipelines:

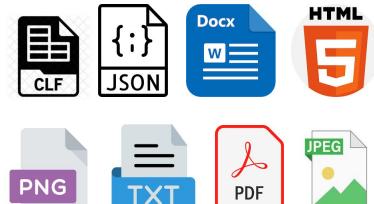
- Multimodal
- Real-time/Streaming
- Parsing, chunking required
- ELT tools poor suited



Structured Data



Multimodal data



PROVENANCE

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

Showing the most recent events.

Component Name	Component Type
ConsumeKafka	ConsumeKafka
ConsumeKafka	ConsumeKafka
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

328717796819631-44815-06f5d077-07



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>

FLINK SQL -> NIFI -> HUGGING FACE GOOGLE GEMINI

f() CALLLLM X CallGenAI

CALLLLM User Defined Function

Properties

Name *

CALLLLM

Description

CALLLLM: call LLM

Output Type *

STRING

Add Input Type

STRING

1 STRING

Test Save

```
function CALLLLM(input) {  
    try {  
        var c = new java.net.URL("http://192.168.1.158:9676/query?calltype=llm&key=" + java.net.URLEncoder.encode(input));  
        c.requestMethod="GET";  
        var reader = new java.io.BufferedReader(new java.io.InputStreamReader(c.getInputStream));  
        var inputLine = new java.lang.String();  
        var out = new java.lang.StringBuilder();  
        if ( reader != null ) {  
            while ((inputLine = reader.readLine()) != null)  
                out.append(inputLine);  
        }  
        reader.close();  
        return out.toString();  
    } catch(err) {  
        return "Unknown: " + err;  
    }  
}  
CALLLLM(sp@)
```



Hugging Face

f() CALLLLM User Defined Function

Properties

Name *

CALLLLM

Description

CALLLLM: call LLM

Output Type *

STRING

Add Input Type

STRING

1 STRING

UDF Tester (CALLLLM)

Parameters

Result

STRING

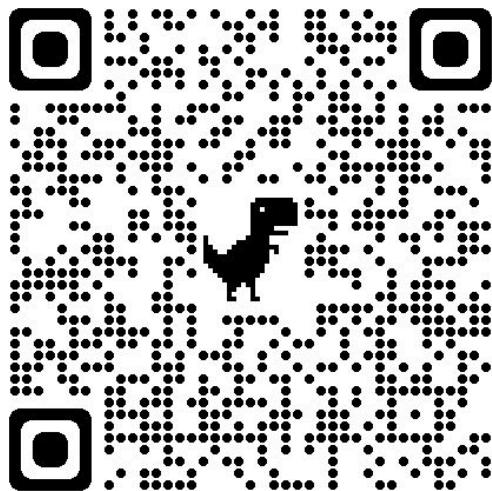
What is apache flink?

Run

Close

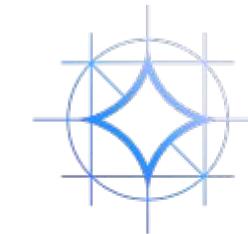
What is apache flink? Flink is an open-source project built for complex distributed data processing on the Hadoop platform. It is mainly developed by a cluster of 43 contributors in Cloudera. In recent years, we have witnessed a technology boom due to Cloud, Big data, Internet of Things, and machine learning. Notably, it emphasizes a few proprietary and open-source components. As a result, technological advancements helped streamline processes within the industry and increased the importance of analytics within this field. F

SSB UDF JS/JAVA + GenAI = Real-Time GenAI SQL



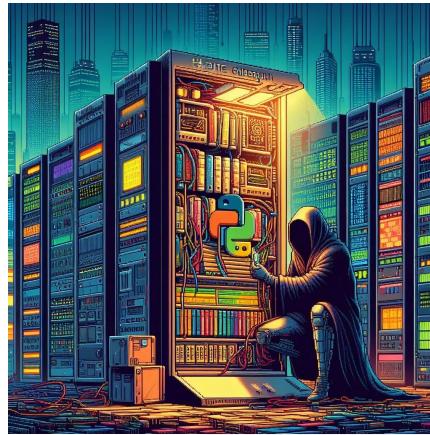
```
SELECT CALLLLM(CAST (messagetext as  
STRING) ) as generatedtext,  
messagerealname, messageusername,  
messagetext,messageusertz,  
messageid, threadts, ts  
FROM flanksslackmessages  
WHERE messagetype = 'message'
```

<https://medium.com/cloudera-inc/adding-generative-ai-results-to-sql-streams-513e1fd2a6af>

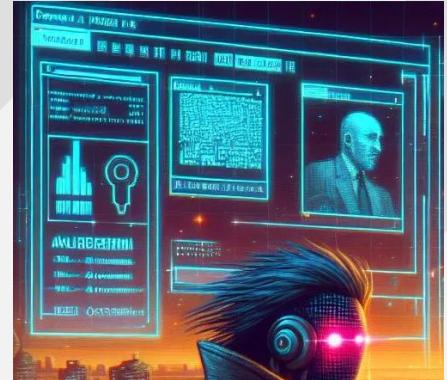


Gemma





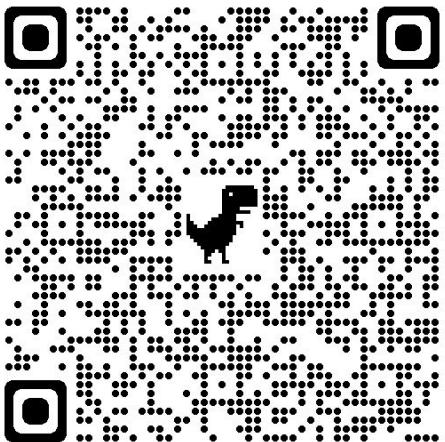
Python Processors





Generate Synthetic Records w/ Faker

- Python 3.10+
- faker
- Choose as many as you want
- Attribute output



Configure Processor | GetFakeRecord 2.0.0-M1

○ Validating

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Include UUID	true
Include CREATED_DT	true
Include EMAIL	true
Include IP_V4	true
Include USER_NAME	true
Include CLUSTER_NAME	true
Include CITY	true
Include COUNTRY	true
Include POSTCODE	true
Include STREET_ADDRESS	true
Include LICENSE_PLATE	true
Include EAN13	true

CANCEL APPLY

flowFile

DETAILS ATTRIBUTES

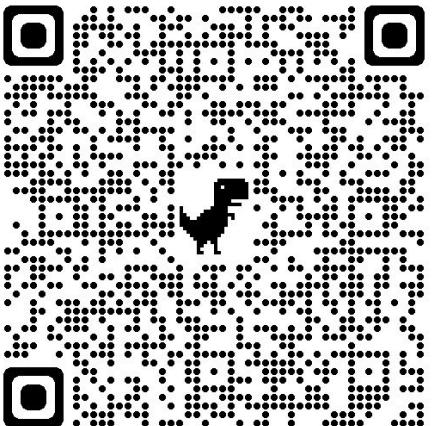
Attribute Values

catchphrase	Ameliorated needs-based matrix
city	West Ashleyshire
clustername	benefit-rate-ask
comment	orchestrate proactive technologies
company	Cruz, Martinez and Edwards
country	Faroe Islands
createddt	2021-01-01



Download a Wiki Page as HTML or WikiFormat (Text)

- Python 3.10+
- Wikipedia-api
- HTML or Text
- Choose your wiki page dynamically



Configure Processor | GetWikiData 2.0.0-M1

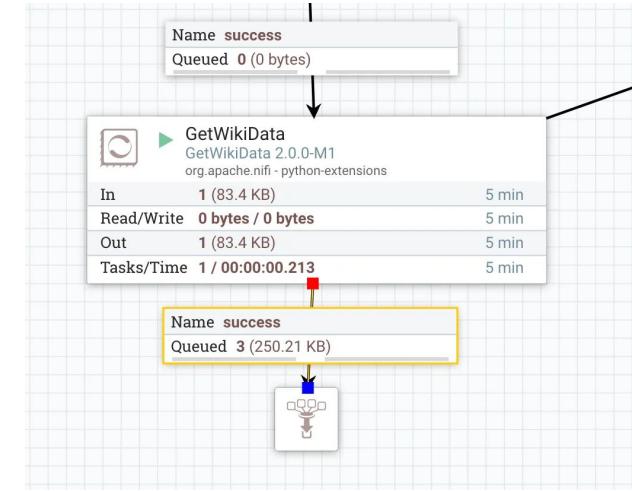
⚠ Invalid

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Plain Text Wiki format or HTML	<input checked="" type="radio"/> text
Wiki Page	<input checked="" type="radio"/> \${company0}

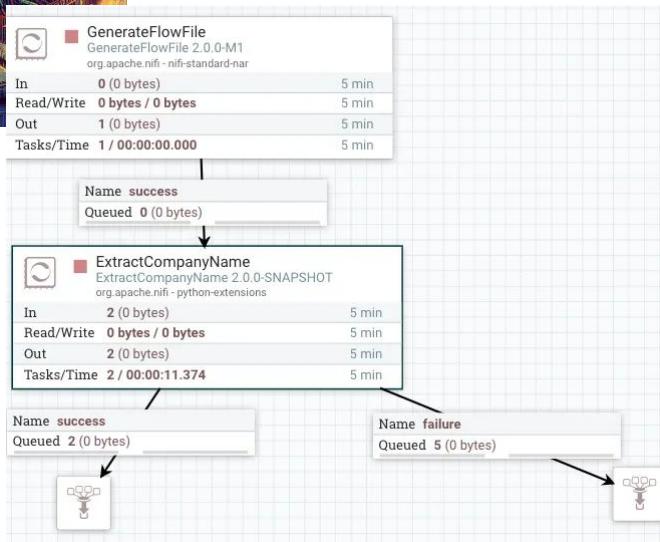
CANCEL APPLY





Extract Company Names

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



Attribute Values

companylist

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

filename

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

parsedcompany

Amazon

path

./

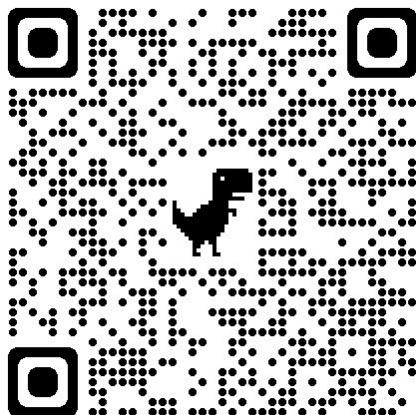
uuid

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



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

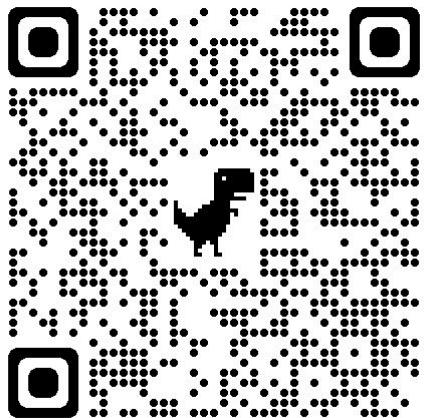


<https://github.com/tspannhw/FLaNK-python-processors>



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

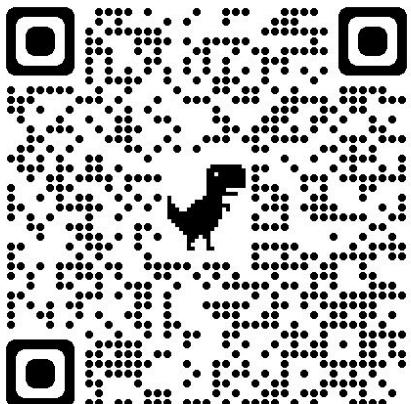


<https://github.com/tspannhw/FLaNK-python-processors>



NSFW Image Detection

- 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

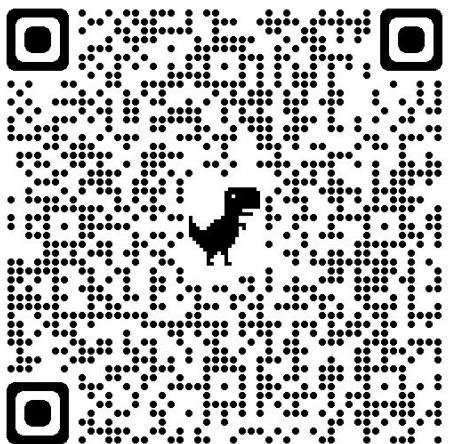


<https://github.com/tspannhw/FLaNK-python-processors>



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>



Extract Entities

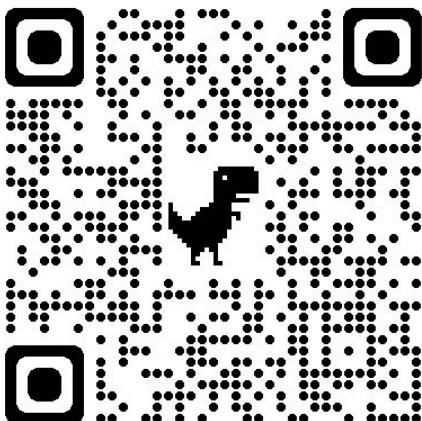
- Python 3.10+
- NLP, SpaCY
- Extract locations
- Extract organizations
- Extract money
- Extract time
- Extract events
- Extract countries
- Extract objects, food, people, quantities

<https://github.com/tspannhw/FLaNK-python-processors/blob/main/ExtractEntities.py>



Parse Addresses

- Python 3.10+
- PYAP Library
- Simple Library if your text includes an address
- Address Parsing
- Address Detecting
- MIT Licensed
- Looking at other libraries, GenAI, DL, ML



<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

DEMO



DEMO #1 - Cloudera Machine Learning - AMPs

CLOUDERA Machine Learning

- Home
- Projects
- Sessions
- Experiments
- Model Deployments
- Model Registry
- Jobs
- Applications
- AMPs
- Runtime Catalog
- Learning Hub
- User Settings
- Site Administration

→ Get Started

Help

2.0.45-b54

Accelerators for ML Projects

All Cloudera Hugging Face Community

Search AMPs Source: Select source Tags: Select tags Deploy External

AMPs(45) AMPs are pre-built, end-to-end ML Projects specifically designed to kickstart your use cases. Explore the featured AMPs below or deploy your own using the Deploy button. [Learn more](#)



Pinecone
Dataflow Machine Learning

Intelligent QA Chatbot with NiFi, Pinecone, and Llama2
ingest data with Cloudera DataFlow from a user-specified website sitemap to create embeddings in a Pinecone vector DB and... deploy a context-aware LLM chatbot app with Cloudera Machine Learning.

Chatbot +10 Deploy Cloudera AMP



Amazon Bedrock
Cloudera Machine Learning

Text Summarization and more with Amazon Bedrock
This AMP demonstrates how to integrate text generation models from the Amazon Bedrock service for text usecases like summarization...

Bedrock +2 Deploy Cloudera AMP



PEFT
Cloudera Machine Learning

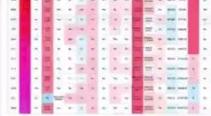
Fine-Tuning a Foundation Model for Multiple Tasks (with QLoRA)
This AMP demonstrates how to improve performance of Large Language Models for specific tasks using distributed fine tuning... techniques like Parameter-Efficient Fine-Tuning(PEFT) and Quantization.

Huggingface +7 Deploy Cloudera AMP



LLM Chatbot Augmented with Enterprise Data
Build a Retrieval Augmented Generation (RAG) Question Answer Large Language Model (LLM) Bot with local documents

Chatbot +9 Deploy Cloudera AMP



Churn Modeling with scikit-learn
Build an scikit-learn model to predict churn using customer telco data.

Churn Prediction +3 Deploy Cloudera AMP



Deep Learning for Image Analysis
Build a semantic search application with deep learning models.

Computer-Vision +2 Deploy Cloudera AMP



Deep Learning for Anomaly Detection
Apply modern, deep learning techniques for anomaly detection to identify network intrusions

Anomaly Detection +3 Deploy Cloudera AMP



Structural Time Series
Applying a structural time series approach to California hourly electricity demand data

Time Series +2 Deploy Cloudera AMP



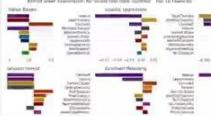
Analyzing News Headlines with SpaCy
Notebook demonstrating entity extraction on headlines with SpaCy

SpaCy +2 Deploy Cloudera AMP



Deep Learning for Question Answering
Explore an emerging NLP capability with WikiQA, an automated question answering system built on top of Wikipedia.

Automated Question A... +3 Deploy Cloudera AMP



Explaining Models with LIME and SHAP
Learn how to explain ML models using LIME and SHAP

Interpretability +2 Deploy Cloudera AMP



Active Learning
Interactive visual workflow of active learning using the MNIST dataset

Active Learning +1 Deploy Cloudera AMP



Strearn
Data for projects, Model training, Model deployment, Monitoring, CI/CD integration, and more

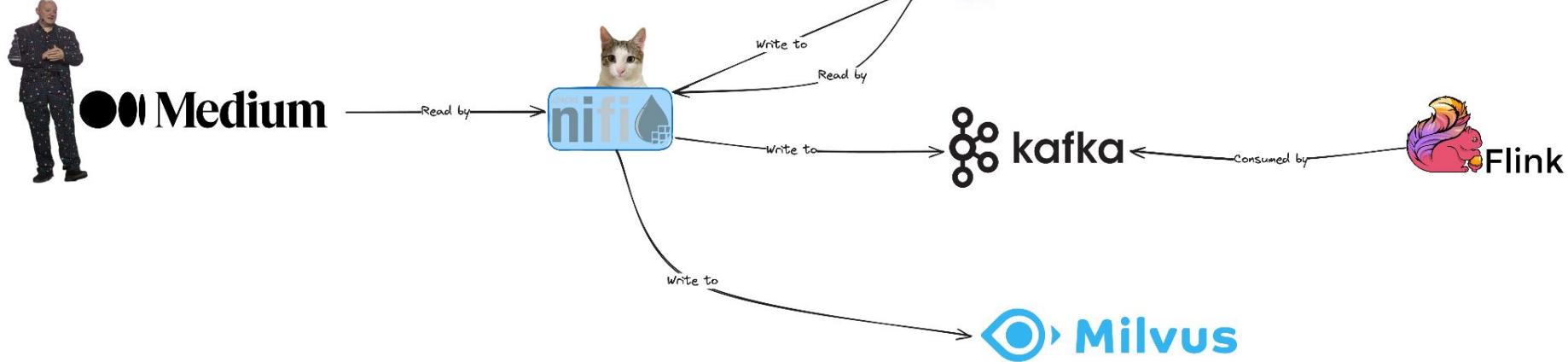
Deploy Cloudera AMP



TPOT + DASH
Data Science Pipeline Optimization Toolkit (TPOT) + Data Analysis and Shiny Dashboards (DASH)

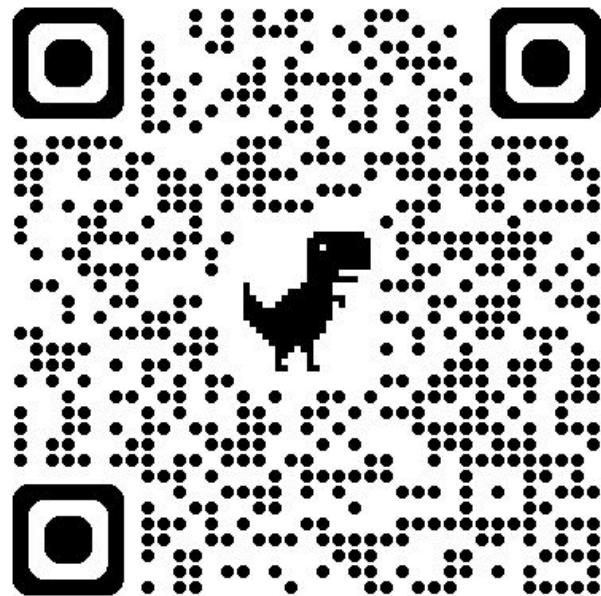
Deploy Cloudera AMP

DEMO #2 - Cloudera DataFlow - Milvus

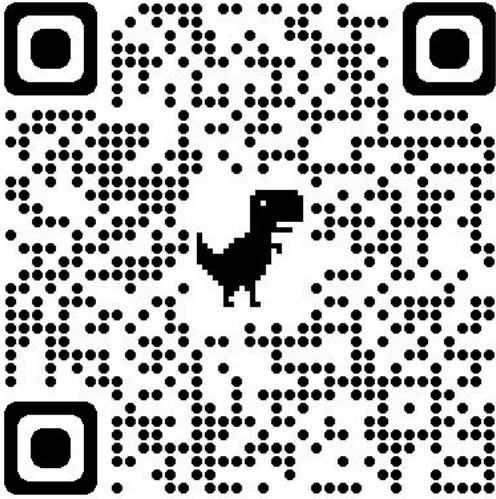


REFERENCES

https://github.com/cloudera/CML_AMP_LLM_Chatbot_Augmented_with_Enterprise_Data



<https://medium.com/@tspann/building-a-milvus-connector-for-nifi-34372cb3c7fa>



CSP Community Edition

A readily available, dockerized deployment of Apache Kafka and Apache Flink that allows you to test the features and capabilities of Cloudera Stream Processing.

[Learn More](#)

CSP Community Edition



- Docker compose file of CSP to run from command line w/o any dependencies, including Flink, SQL Stream Builder, Kafka, Kafka Connect, Streams Messaging Manager and Schema Registry.
 - \$>docker compose up
 - Licensed under the Cloudera Community License
 - **Unsupported Commercially (Community Help - Ask Tim)**
 - Community Group Hub for CSP
 - Find it on docs.cloudera.com (see QR Code)
 - Kafka, Kafka Connect, SMM, SR, Flink, Flink SQL, MV, Postgresql, SSB
 - Develop apps locally

DATA SUMMIT

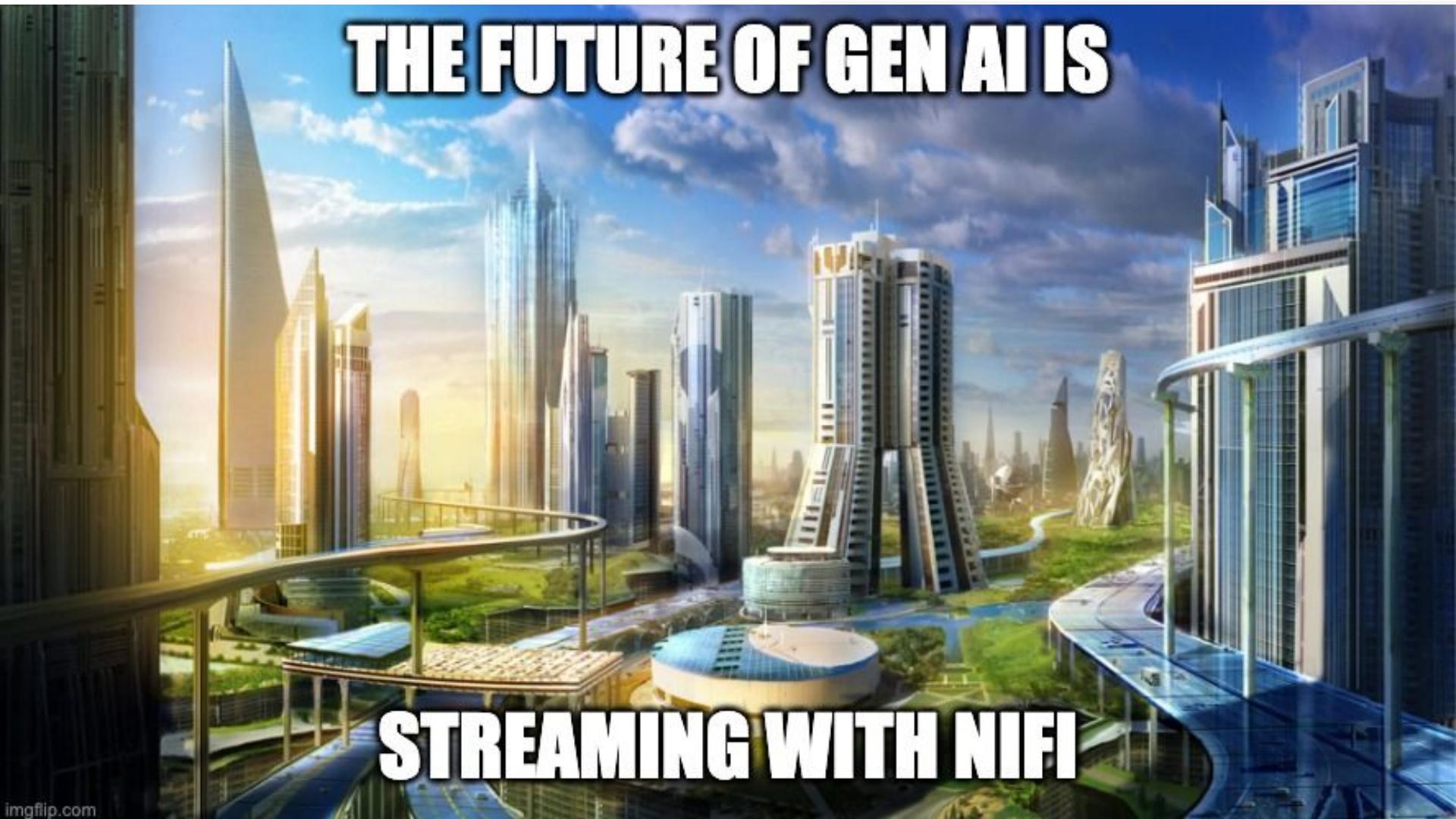
UNLEASH THE POWER OF YOUR DATA

MAY 8–9
BOSTON, MA



TH^AN^O Y^U *



A vibrant, futuristic cityscape featuring numerous skyscrapers of various designs, some with reflective glass facades and others with more organic, rounded shapes. A prominent feature is a large, curved elevated track or bridge system that cuts through the city. The sky is a clear blue with scattered white clouds. In the foreground, there's a circular, low-profile building and some green spaces. The overall atmosphere is one of advanced technology and urban development.

THE FUTURE OF GEN AI IS

STREAMING WITH NIFI