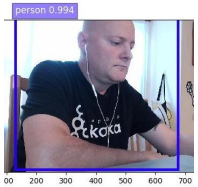
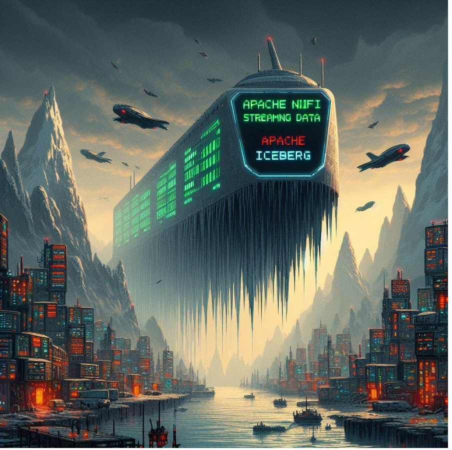
CLOUDERA

Adding Generative AI to Real-Time Streaming Pipelines

Tim Spann
Principal Developer Advocate

May 1, 2024



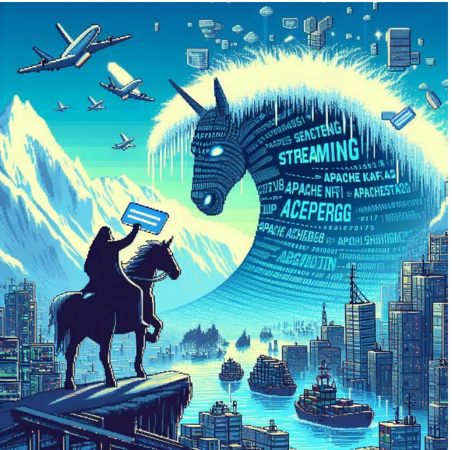

APACHE NIFI
STREAMING DATA
APACHE
ICEBERG

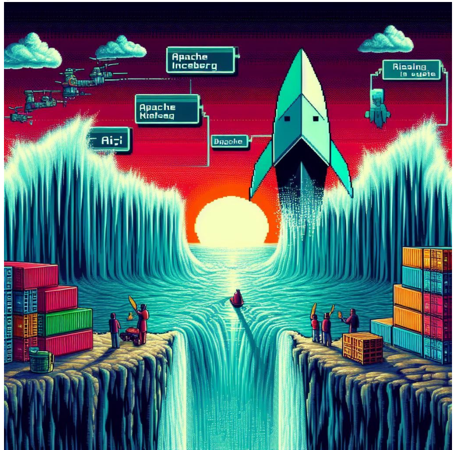

APACHE
APACHE

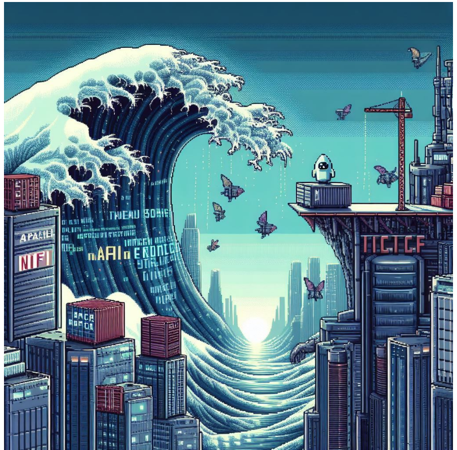

Apache NiFi

nifi

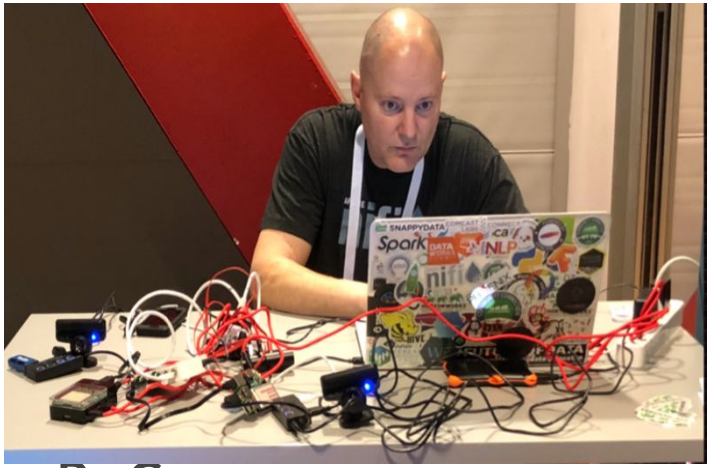
Milvus


STREAMING


Apache
Iceberg


NIFI

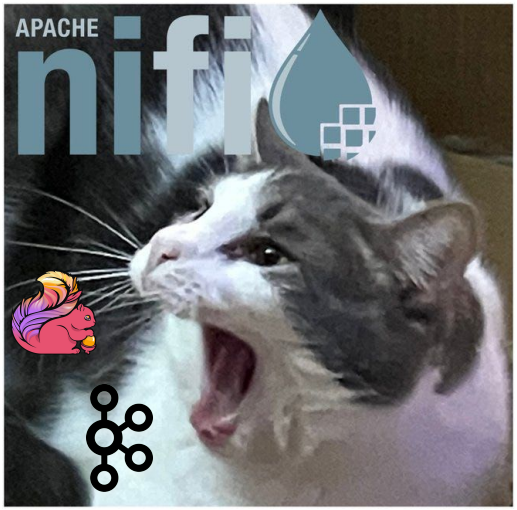
TIM SPANN



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

FLaNK Stack Weekly by Tim Spann

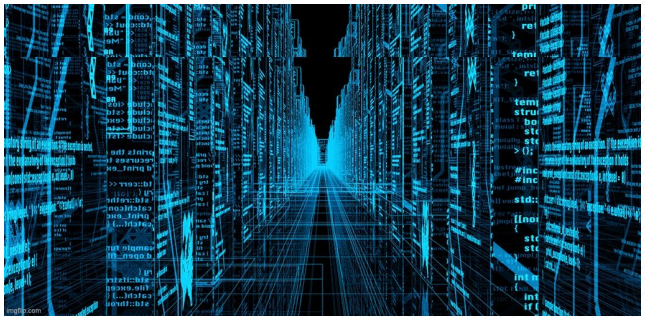


https://bit.ly/32dAJft

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

RAPID INNOVATION IN THE LLM SPACE

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

Notable LLMs

Closed Models

- OpenAI — GPT3.5, GPT4
- Claude2

Open Models

- Meta AI — Llama2, Code Llama
- MISTRAL AI_ — Mistral7B, Mixtral8x7B

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

Popular LLM Frameworks

LangChain

Langchain is a framework for developing apps powered by LLMs

- Python and JavaScript Libraries
- Provides modules for **LLM Interface**, **Retrieval**, & **Agents**

LlamaIndex

LLamaIndex is a framework designed specifically for RAG apps

- 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

Milvus

Open Community & Open Models

Hugging Face

HuggingFace is an ML community for hosting & collaborating on **models**, **datasets**, and **ML applications**

- 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

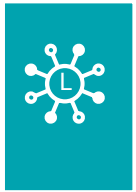
- Gauge emotions and opinions based on a piece of text.
- Understand and deliver a more nuanced message back based on sentiment.

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.

Adoption dependent upon an Enterprise's risk tolerance, restrictions, decision rights and disclosure obligations.

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

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

MANAGED VECTOR STORE

Milvus

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

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

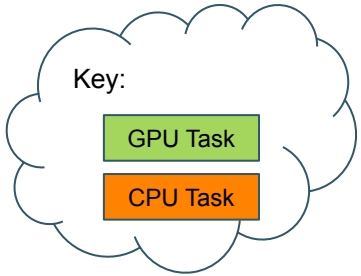
Streaming Classification
Real-Time Model Deployment

milvus

CDE Spark python DASK RAY

CDP CLOUDERA

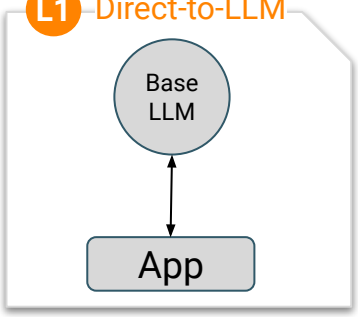
minifi



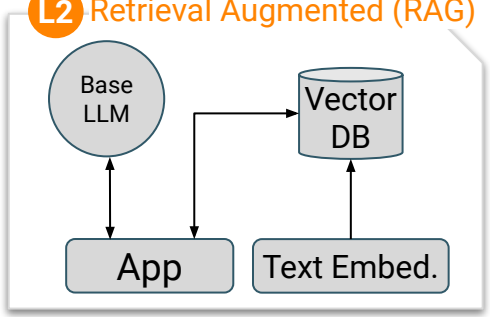
EVOLUTION OF LLM APP ARCHITECTURES

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

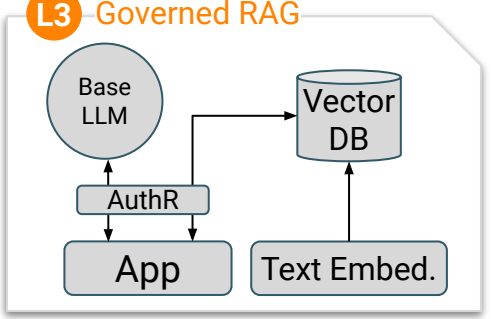
L1 Direct-to-LLM



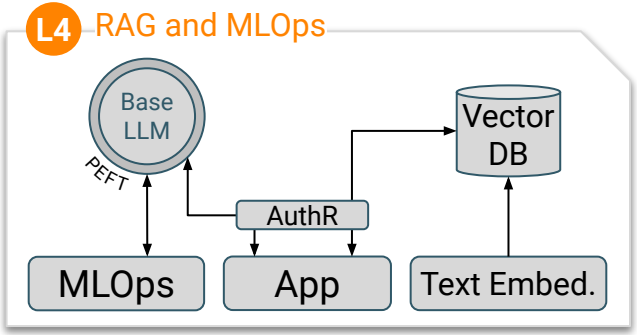
L2 Retrieval Augmented (RAG)



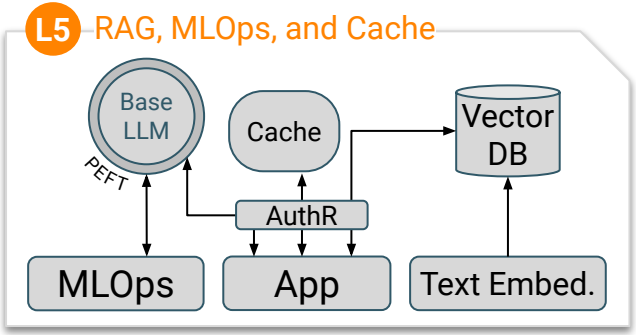
L3 Governed RAG



L4 RAG and MLOps



L5 RAG, MLOps, and Cache



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

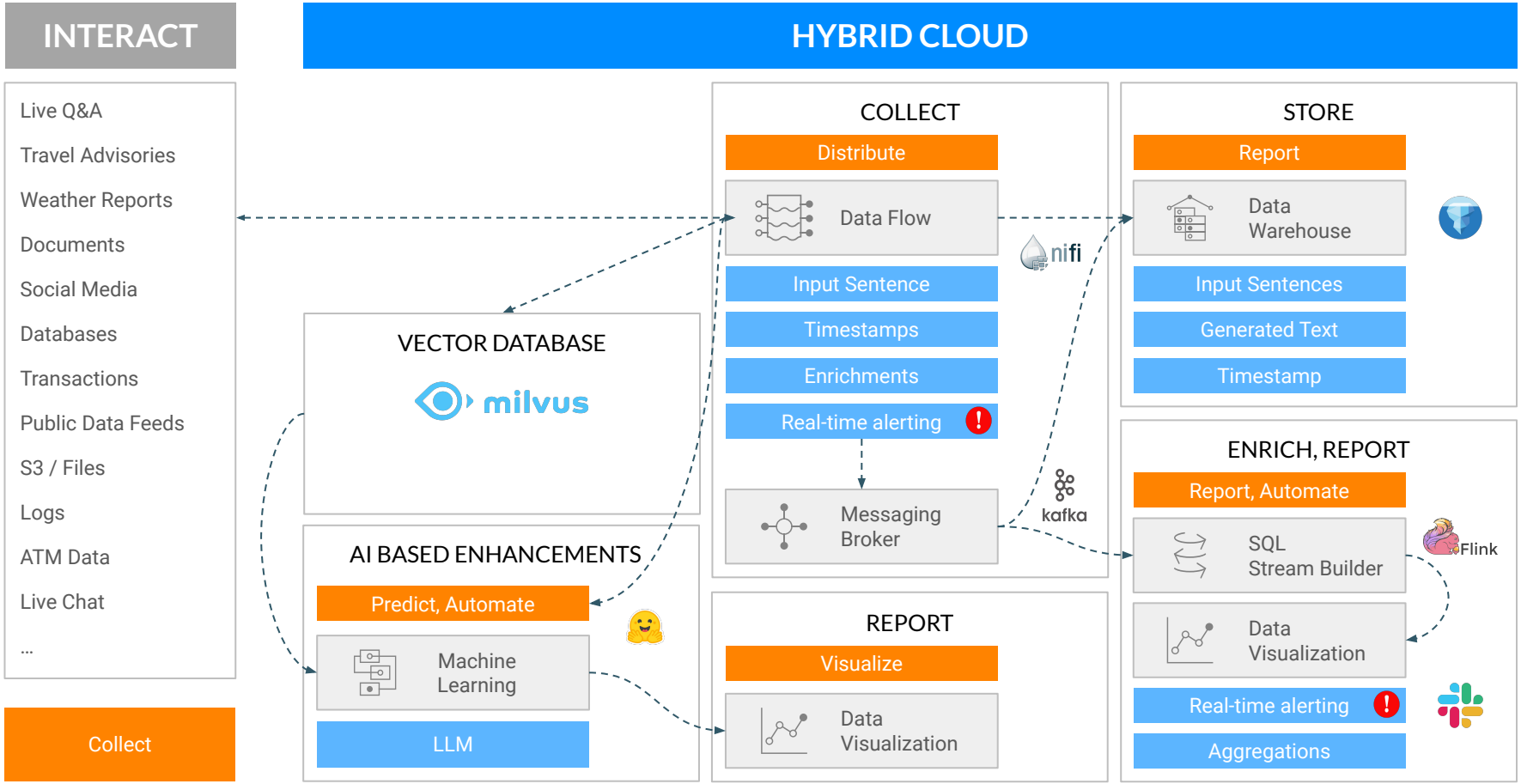
CLOUDERA SDX

SHARED DATA EXPERIENCE FOR MODELS

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

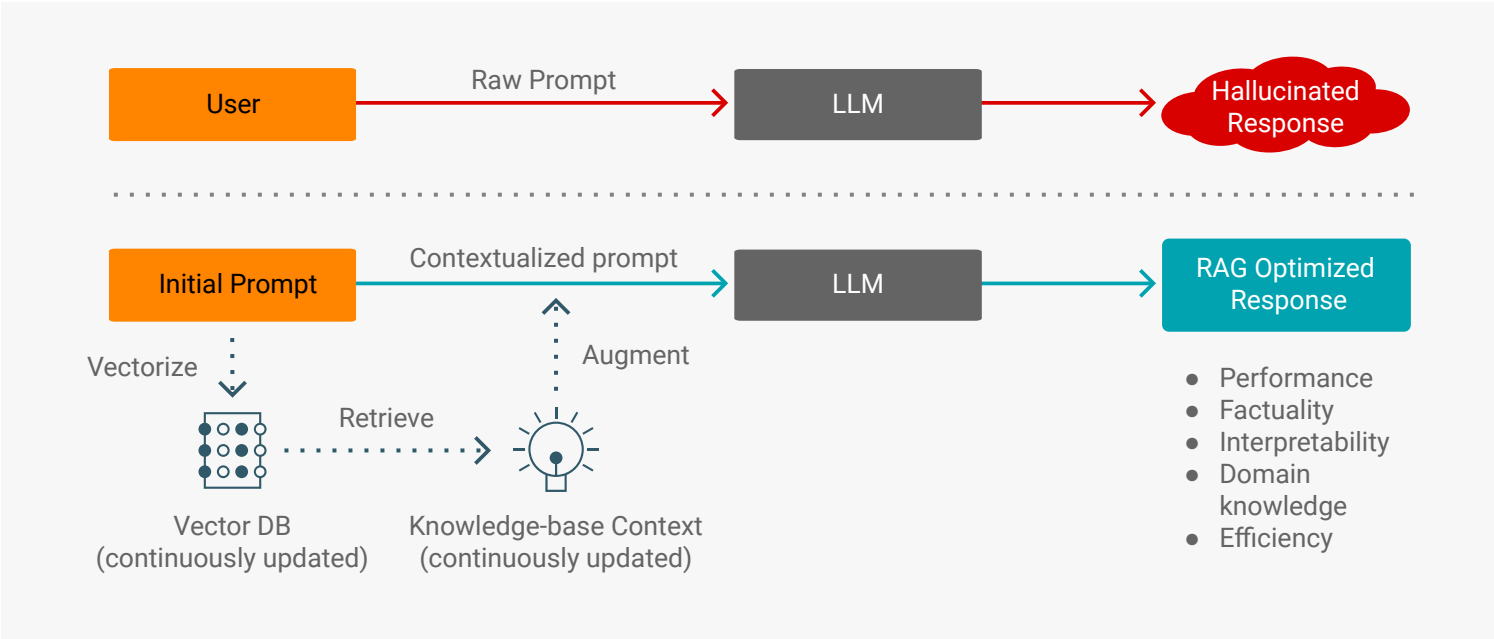
DISTRIBUTED AI COMPUTE

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



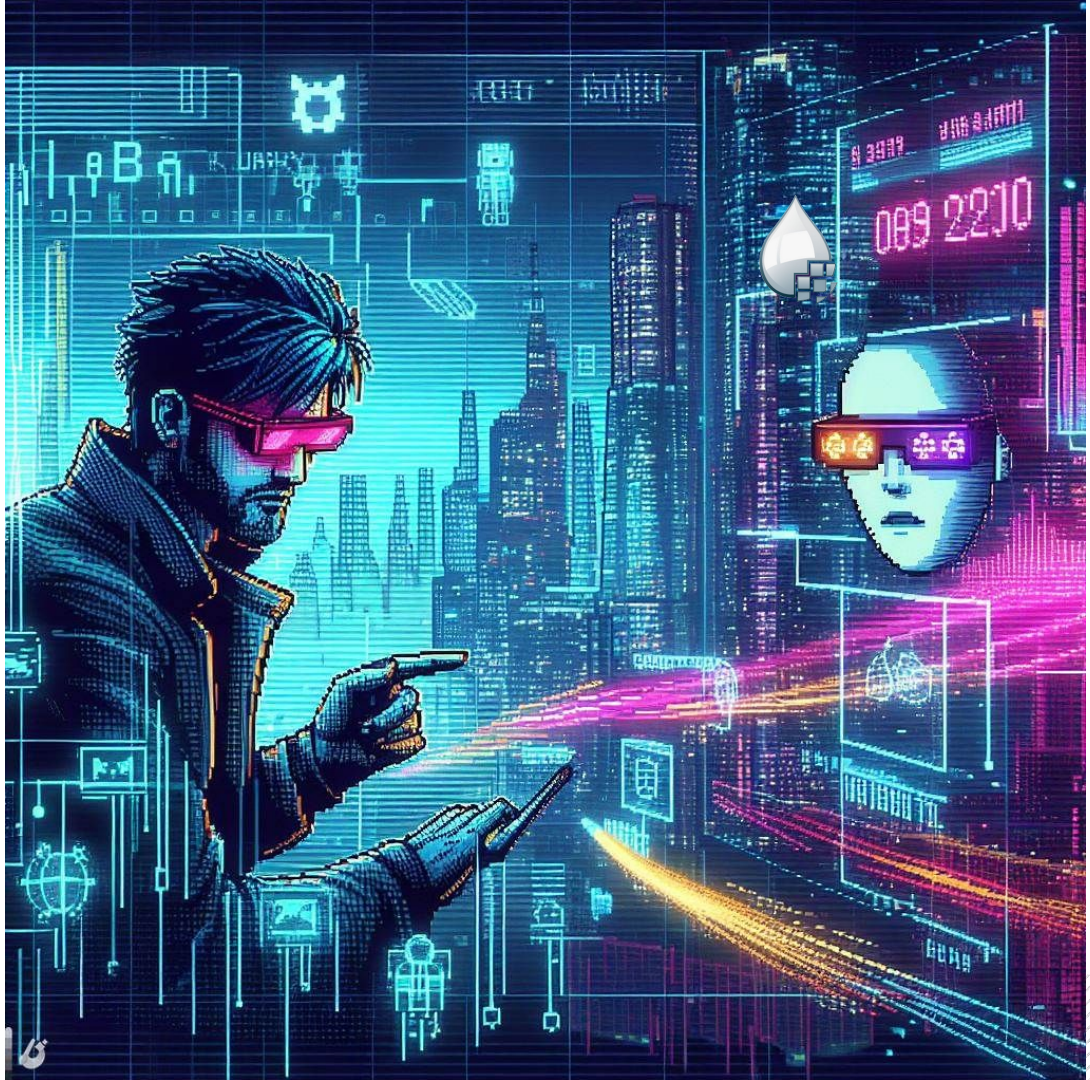
REAL-TIME CONTEXT FOR GEN AI

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

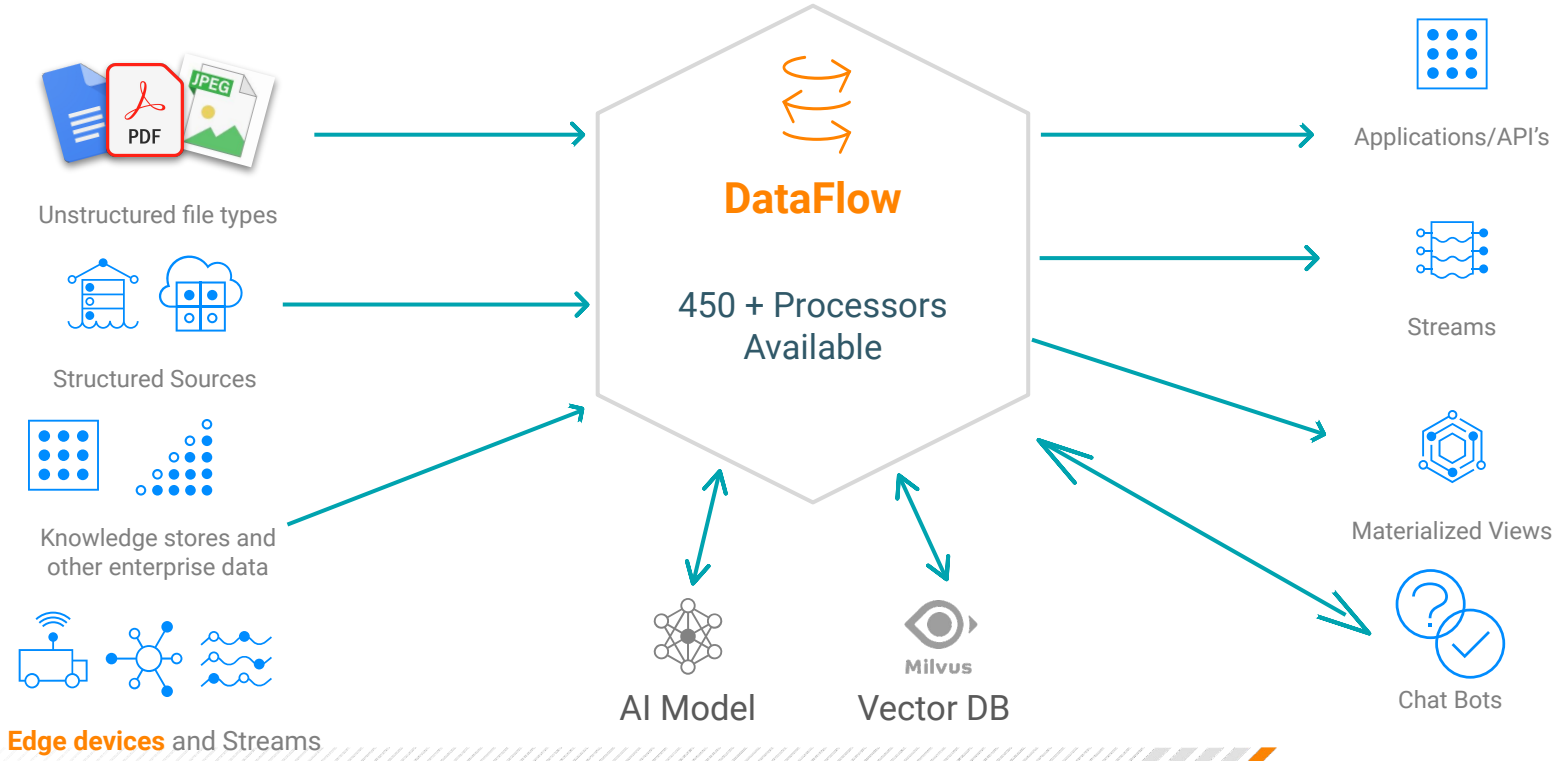
- Cloudera ML
- Amazon Polly, Translate, Textract, Transcribe, Bedrock, …
- Hugging Face
- IBM Watson X.AI
- Vector Stores Anywhere: Milvus, …

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

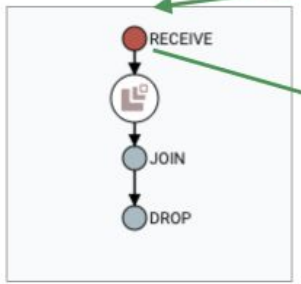
PROVENANCE

Displaying 13 of 104
Oldest event available: 11/15/2016 13:34:50 EST
Showing the most recent events.

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

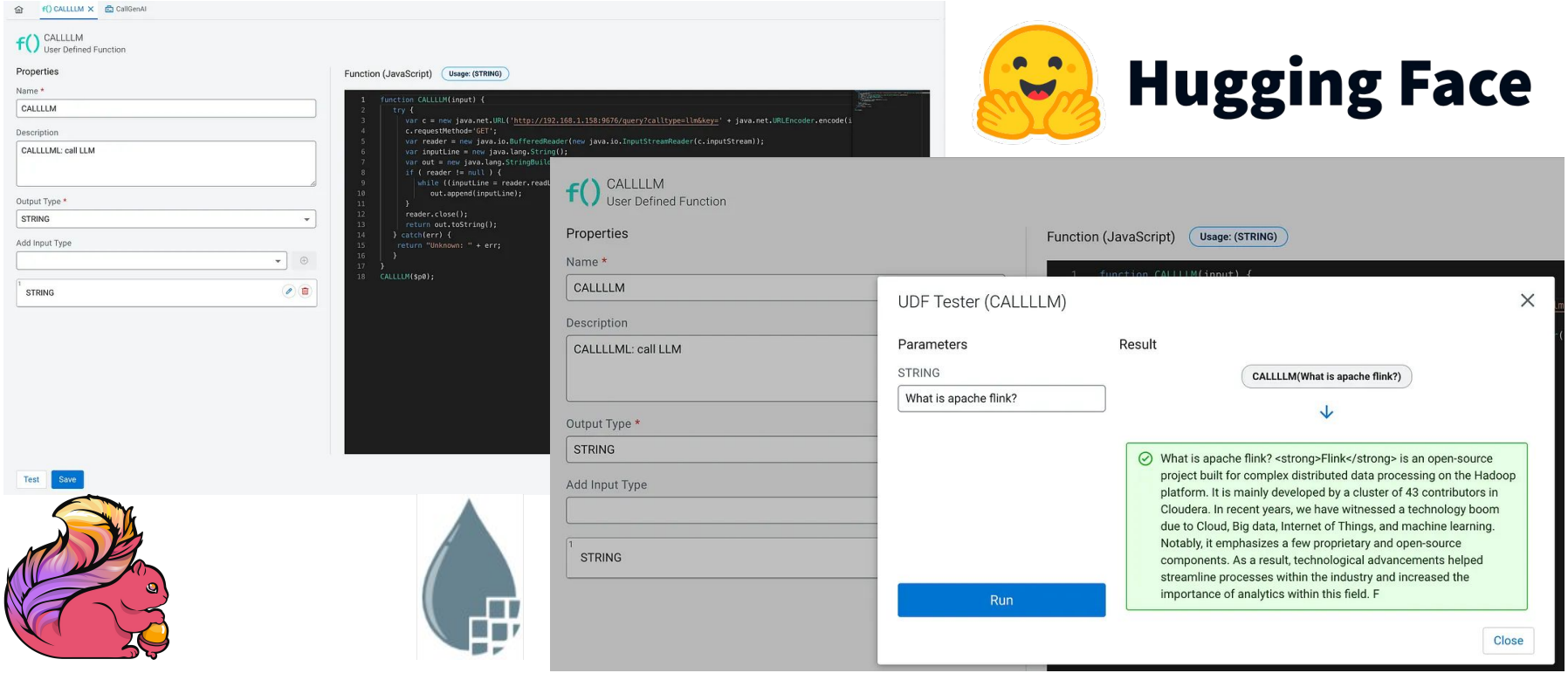
NiFi 2.0.0 Features

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

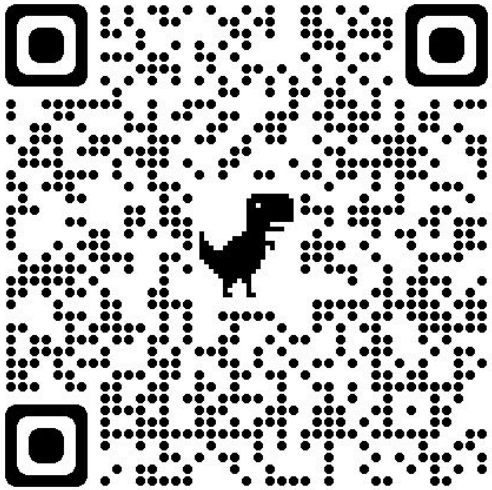
https://cwiki.apache.org/confluence/display/NIFI/NiFi+2.0+Release+Goals

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

FLINK SQL -> NIFI -> HUGGING FACE GOOGLE GEMINI



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

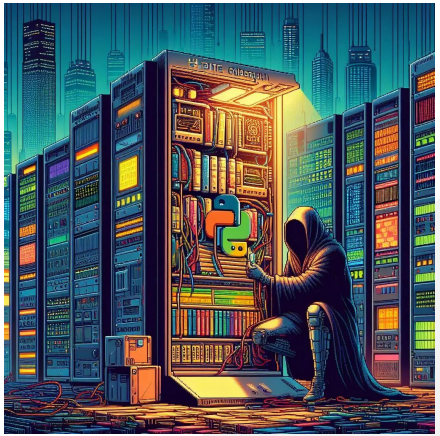


```
SELECT CALLLLM(CAST(messagetext as
STRING)) as generatedtext,
messagerealname, messageusername,
messagetext,messageusertz,
messageid, threadts, ts
FROM flankslackmessages
WHERE messagetype = 'message'
```

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

Gemma

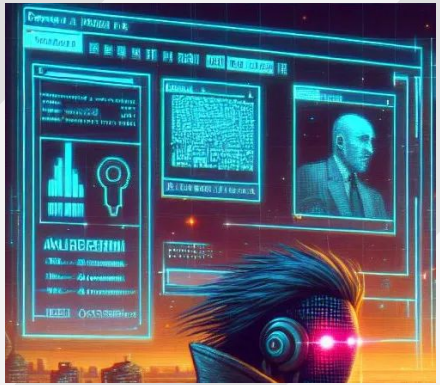





nifi

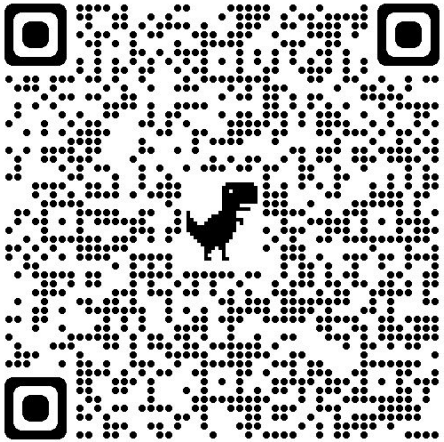
Python Processors

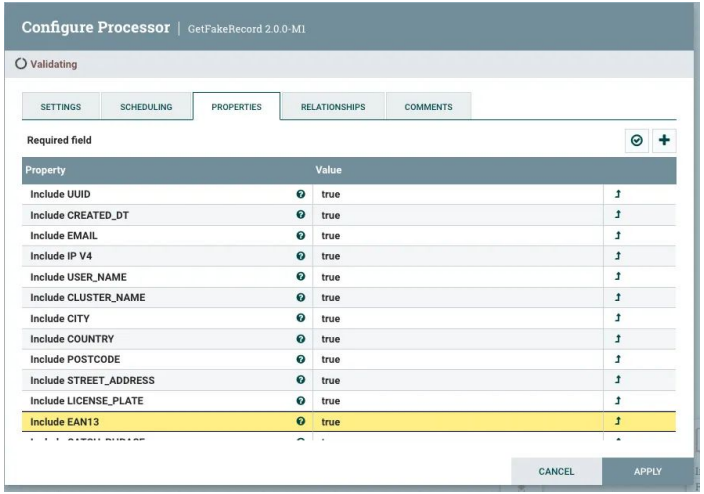


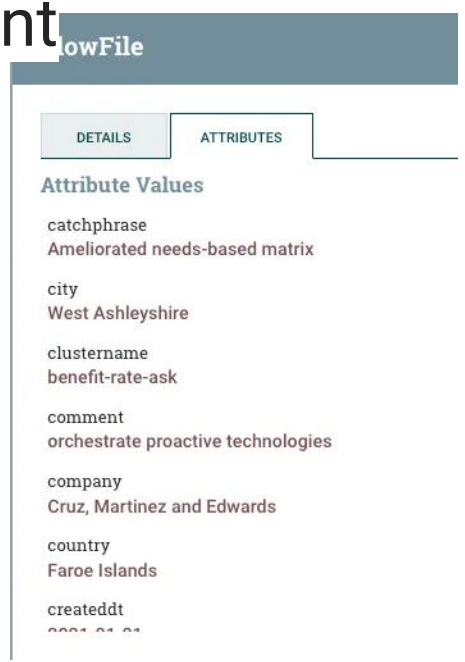


Generate Synthetic Records w/ Faker

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

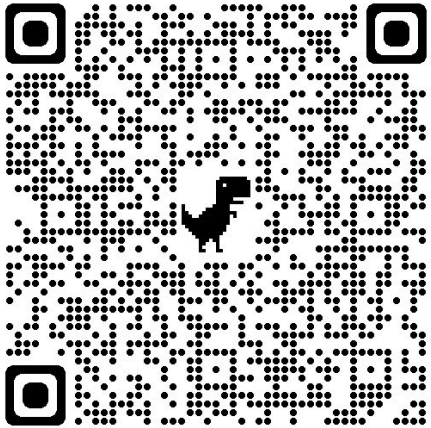


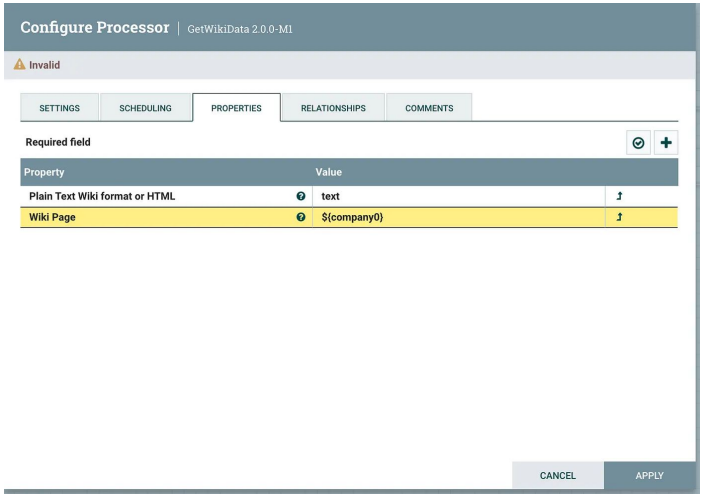


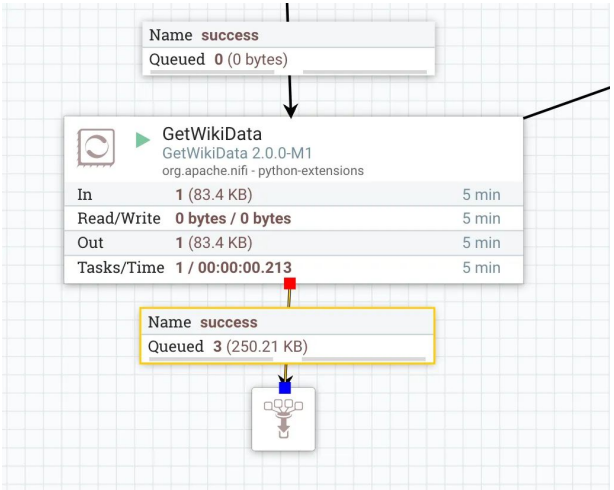


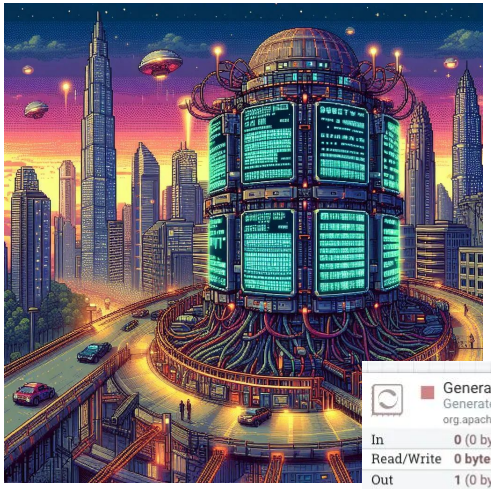
Download a Wiki Page as HTML or WikiFormat (Text)

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



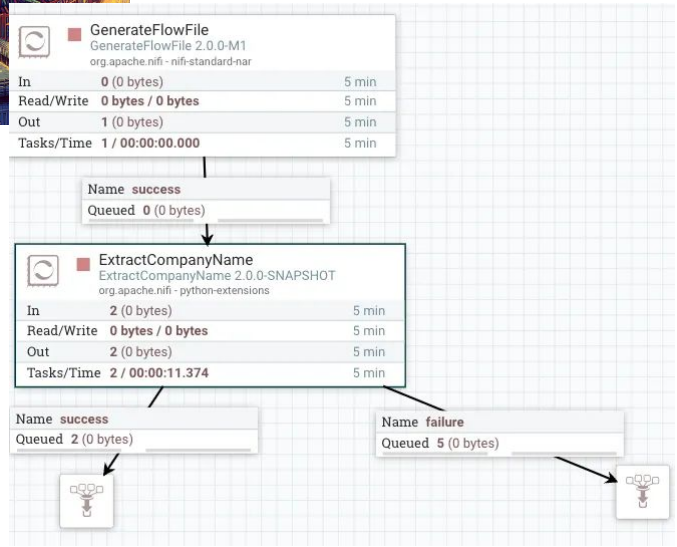






Extract Company Names

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

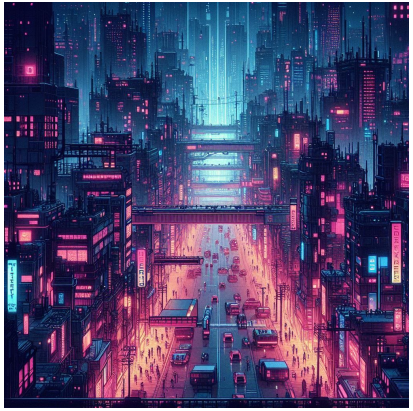


https://github.com/tspannhw/FLaNK-python-ExtractCompanyName-processor

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

Extract Entities

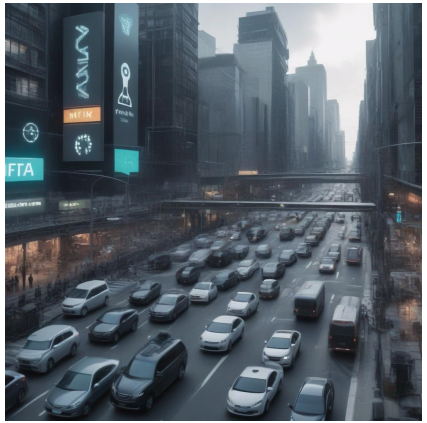
- Python 3.10+
- NLP, SpaCY
- Extract locations
- Extract organizations
- Extract money
- Extract time
- Extract events
- Extract countries
- Extract objects, food, people, quantities

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

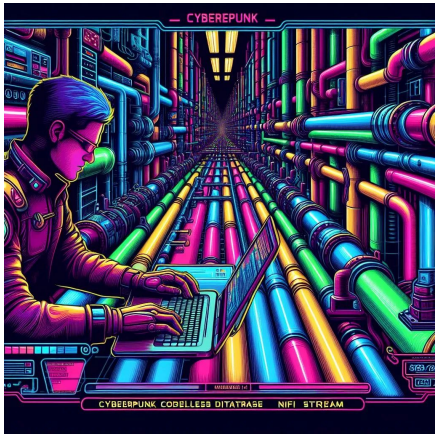


Address To Lat/Long

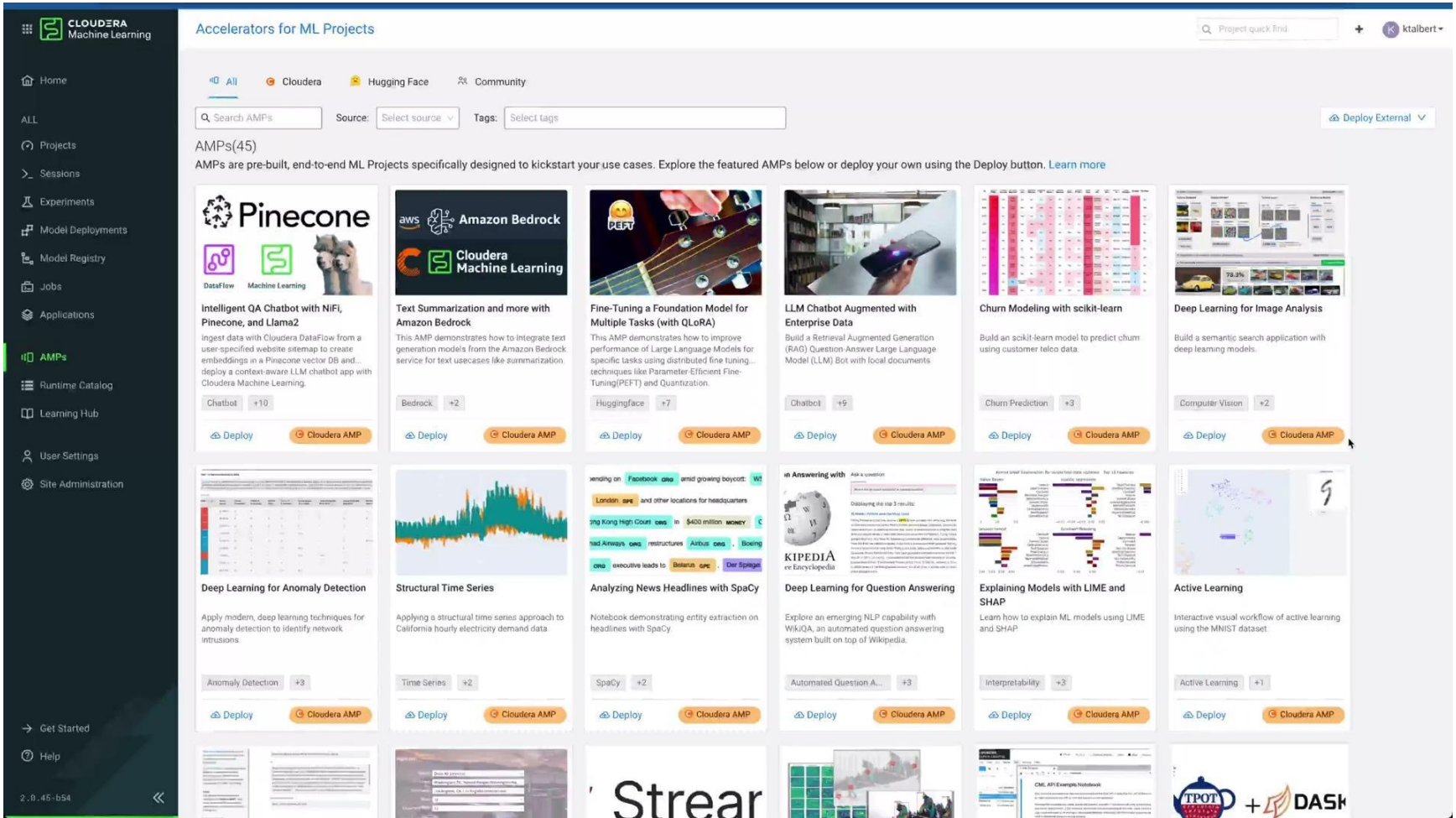
- Python 3.10+
- geopy Library
- Nominatim
- OpenStreetMaps (OSM)
- openstreetmap.org/copyright
- Returns as attributes and JSON file
- Works with partial addresses
- Categorizes location
- Bounding Box

https://github.com/tspannhw/FLaNKAI-Boston

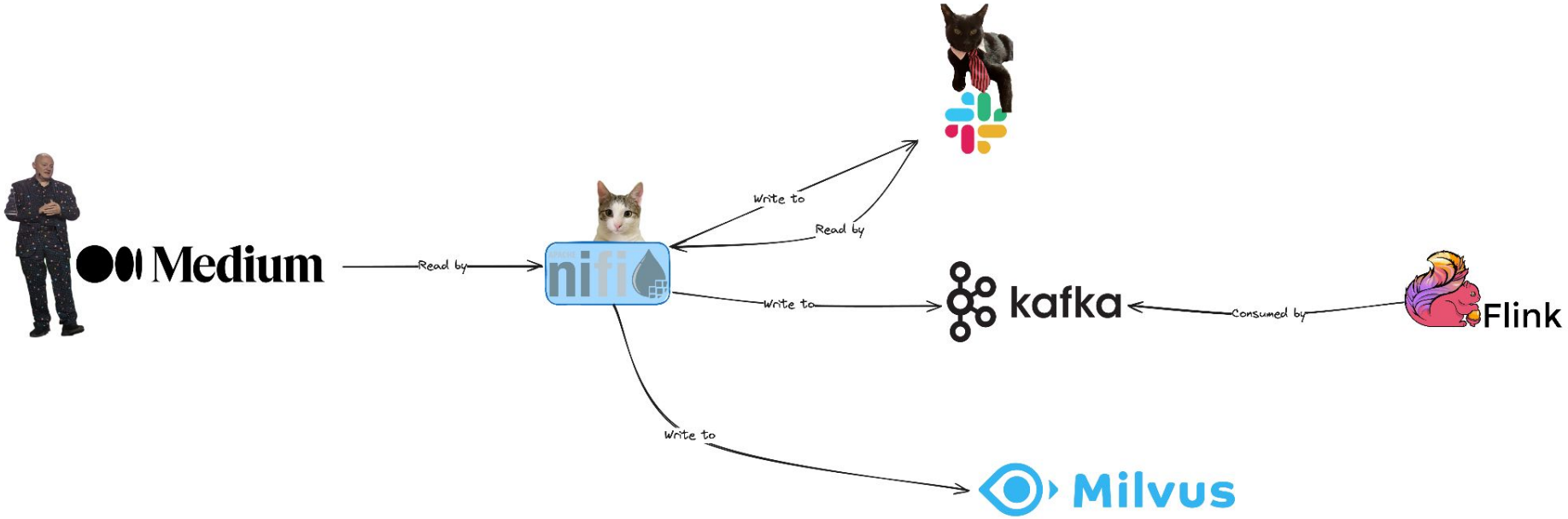
DEMO



DEMO #1 - Cloudera Machine Learning - AMPs

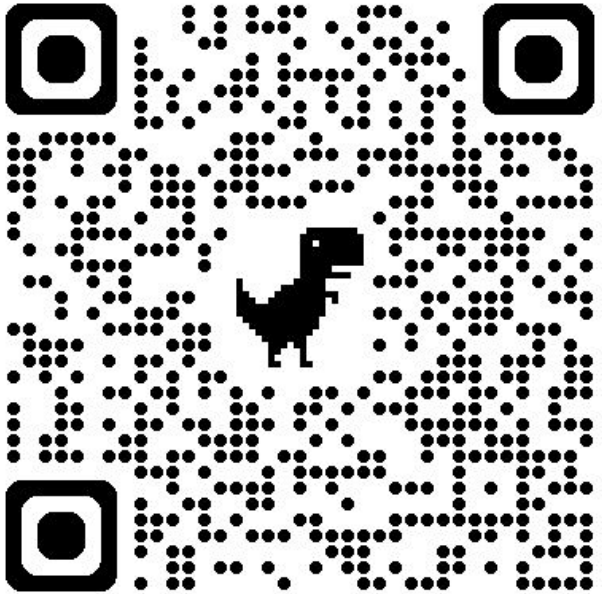


DEMO #2 - Cloudera DataFlow - Milvus



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

- 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

CSP Community Edition

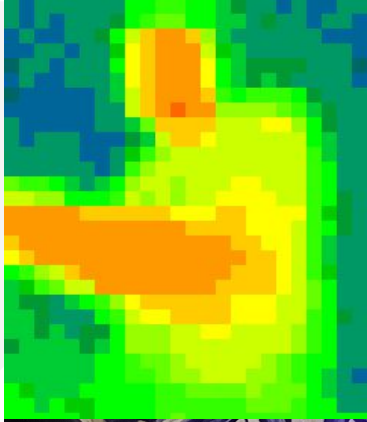
A readily available, dockerized deployment of Apache Kafka and Apache Flink that allows you to test the features and capabilities of Cloudera Stream Processing.

Learn More

MAY 8–9
BOSTON, MA

THANK YOU

APACHE nifi




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