CLOUDERA

Adding Generative AI to Real-Time Streaming Pipelines

Tim Spann
Principal Developer Advocate

26-October-2023

EVOLVE NYC

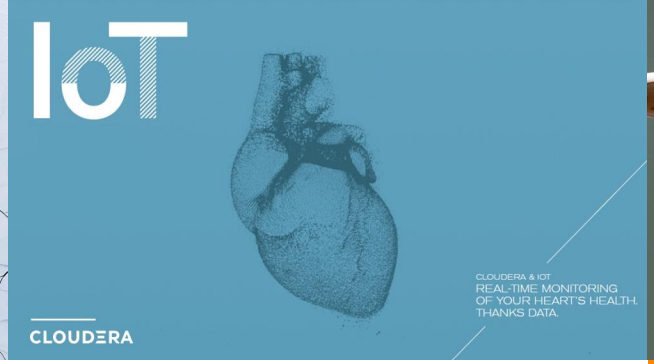
NOV 2, 2023
THE GLASSHOUSE
660 12TH AVE, NEW YORK, NY 10019

presented by

CLOUDERA IBM intel aws

Want to elevate your business with data and AI? Join us at Evolve, a free global live event! Connect with innovators for a day of insights, and leave with actionable solutions that inspire and energize.

https://events.dzone.com/dzone/Data-Pipelines-Investigating-the-Modern-Day-Stack

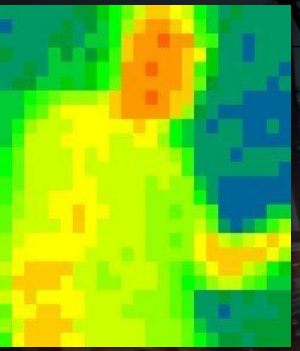


CLOUDERA

EDGE
2AI
CLOUDERA

ENTERPRISE
DATA
CLOUD
CLOUDERA

Tim Spann

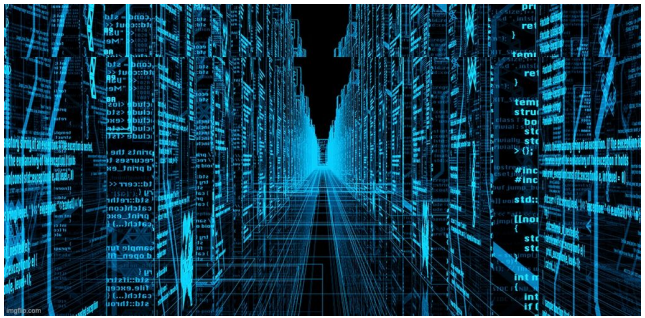


@PaasDev www.datainmotion.dev
github.com/tspannhw medium.com/@tspann
Principal Developer Advocate

Princeton/NYC/Philly Future of Data Meetup
ex-Pivotal, ex-Hortonworks, ex-StreamNative,
ex-PwC, ex-EY, ex-HPE.

Apache NiFi x Apache Kafka x Apache Flink x AI

Future of Data - NYC + NJ + Philly + Virtual



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

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

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

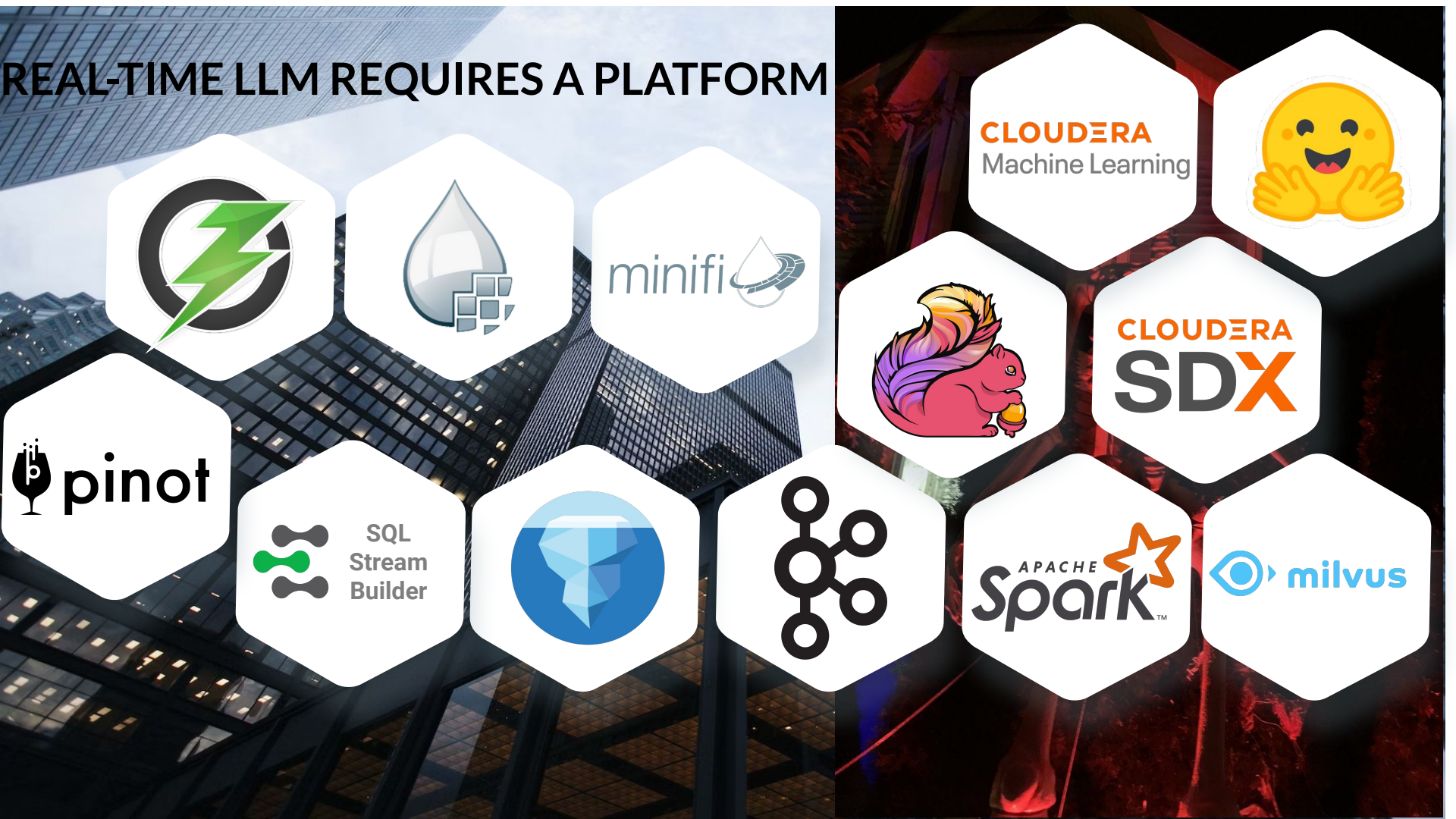
@PaasDev

FLaNK Stack Weekly

https://bit.ly/32dAJft

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

REAL-TIME LLM REQUIRES A PLATFORM



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

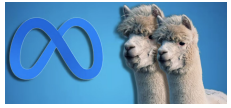
Applications/API's

Streams

Materialized Views

Open Data Lakehouse

meta-llama/llama-2-70b-chat



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

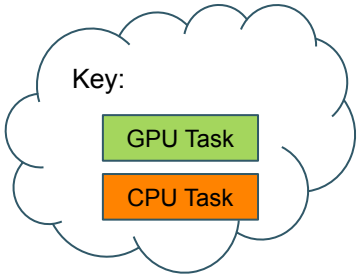
Streaming Classification
Real-Time Model Deployment

milvus Pinecone Chroma

CDE Spark python DASK RAY

minifi

CDP CLOUDERA



Streaming LLM Going Forward

- FLARE pattern
- RAG (retrieval augmented)
- Kafka Integration with Vector Stores
- NiFi Integration with HuggingFace Transformers
- Current Research

LangStream

CLOUDERA
Machine Learning

https://github.com/mit-han-lab/streaming-llm

Streaming LLM With Java
(NiFi, Kafka,Flink, Spark)

https://github.com/deepjavalibrary/djl-demo/tree/master/huggingface/nlp/src/main/java/com/examples
https://pub.towardsai.net/deploy-huggingface-nlp-models-in-java-with-deep-java-library-e36c635b2053

https://milvus.io/docs/create_collection.md

https://github.com/langchain4j/langchain4j

https://github.com/mariofusco/droolsGPT

DJL

Milvus

LangStream

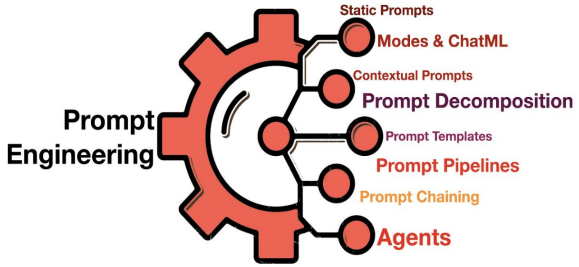
LLAMA 2


SPIRIT
Streaming LLM
MAGIC!
ADULT
Size Costume
ONE SIZE FITS MOST

Greg Phillips - GETTING THE BEST RESULTS FROM LLMs

Prompt Engineering ($)

getting better responses through better questions & context



Retrieval Augmented Generation ($$)

adding search results as context to the prompt

Parameter-Efficient Fine Tuning ($$$)

training a smaller “adapter” to combine with a LLM

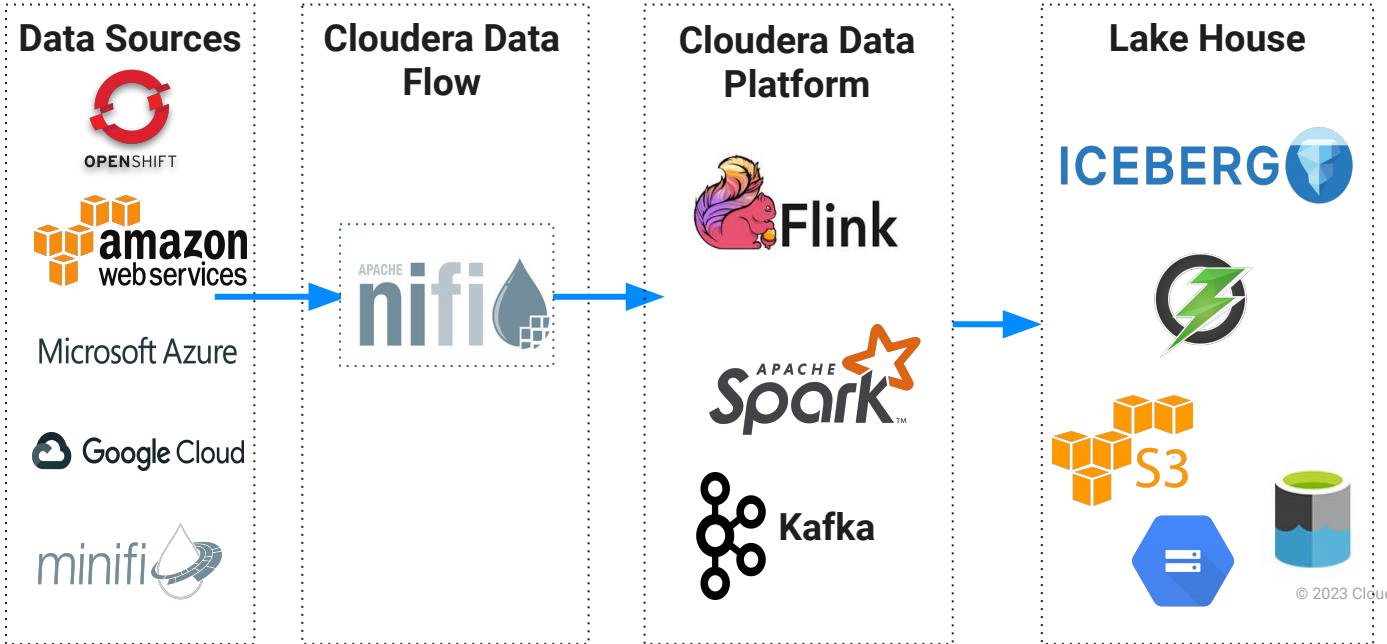
Re-Training a Model ($$$$$)

creating a new model from a base LLM & enterprise-specific data

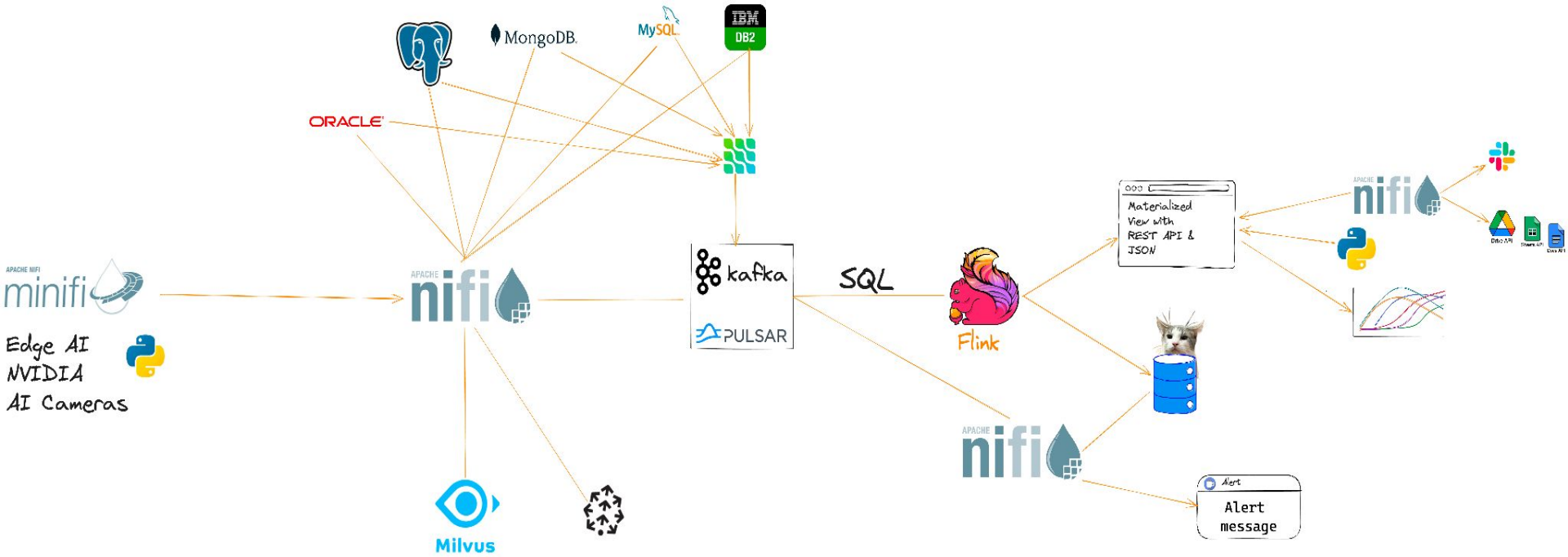
Or use a combination of multiple approaches…

INGEST

Run collection and streaming on any cloud, server, container, bare metal, device or VM

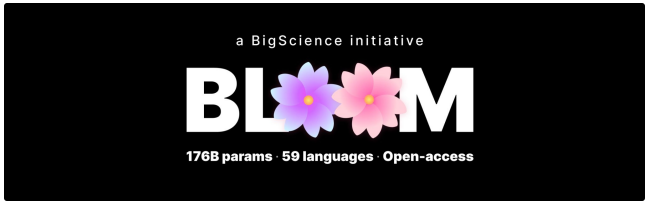


ENRICH


MongoDB.
MySQL
IBM
DB2
ORACLE
APACHE NIFI
minifi
Edge AI
NVIDIA
AI Cameras
APACHE
nifi
kafka
PULSAR
SQL
Flink
Materialized
View with
REST API &
JSON
APACHE
nifi
APACHE
nifi
Milvus
Alert
Alert
message

FUNNEL

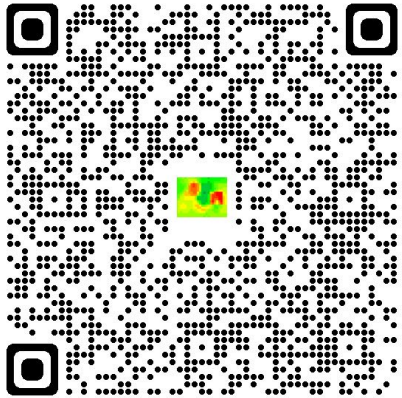
https://github.com/tspannhw/FLaNK-HuggingFace-DistilBert-SentimentAnalysis



watsonx.ai

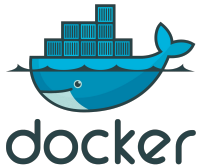
CLOUDERA
Machine Learning

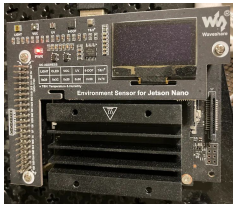
https://github.com/tspannhw/FLaNK-LLM



DISTRIBUTE

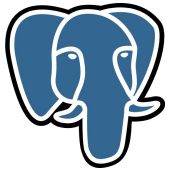
DEPLOY





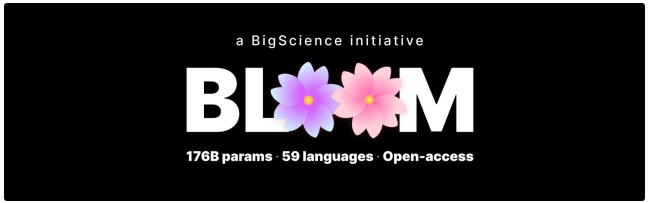
https://github.com/tspannhw/FLaNK-Edge-Models

STORE



Generative AI

https://github.com/tspannhw/FLaNK-HuggingFace-DistilBert-SentimentAnalysis



watsonx.ai

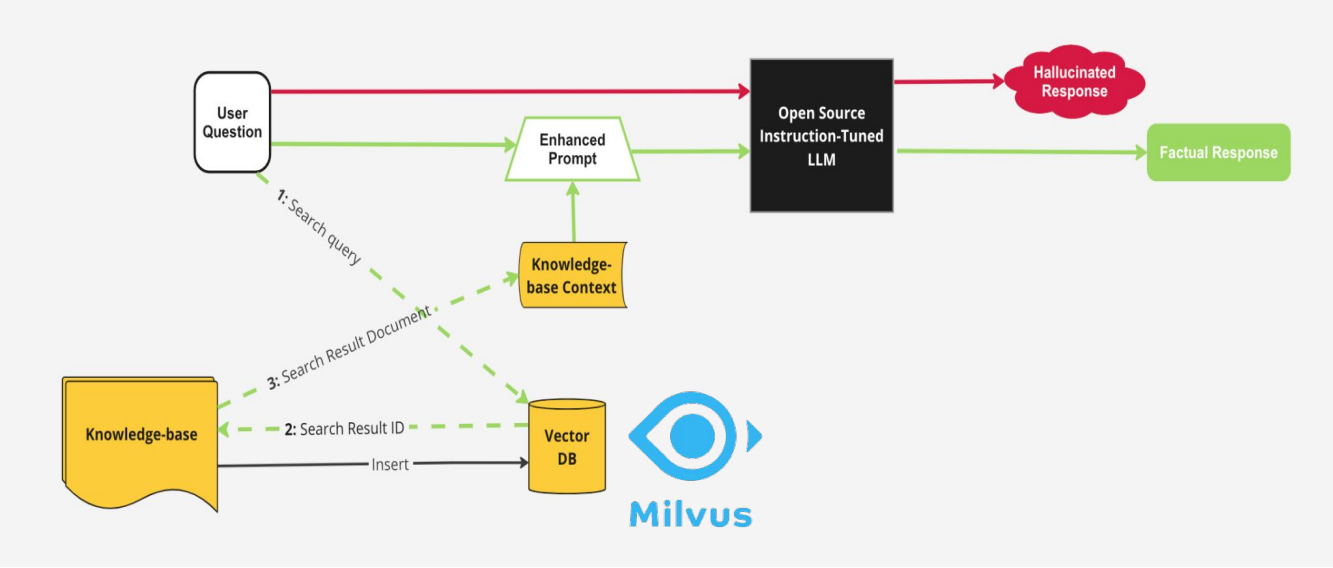
CLOUDERA
Machine Learning

https://github.com/tspannhw/FLaNK-LLM

NEW APPLIED ML PROTOTYPE RELEASE

LLM Chatbot Augmented with Enterprise Data

- Hosts an LLM from HuggingFace
- Creates and populates **Milvus** vector database with Cloudera Docs
- Demonstration of Retrieval Augmented Generation (RAG) Architecture
- Deploys ChatBot-like web application to interact with base and RAG LLMs



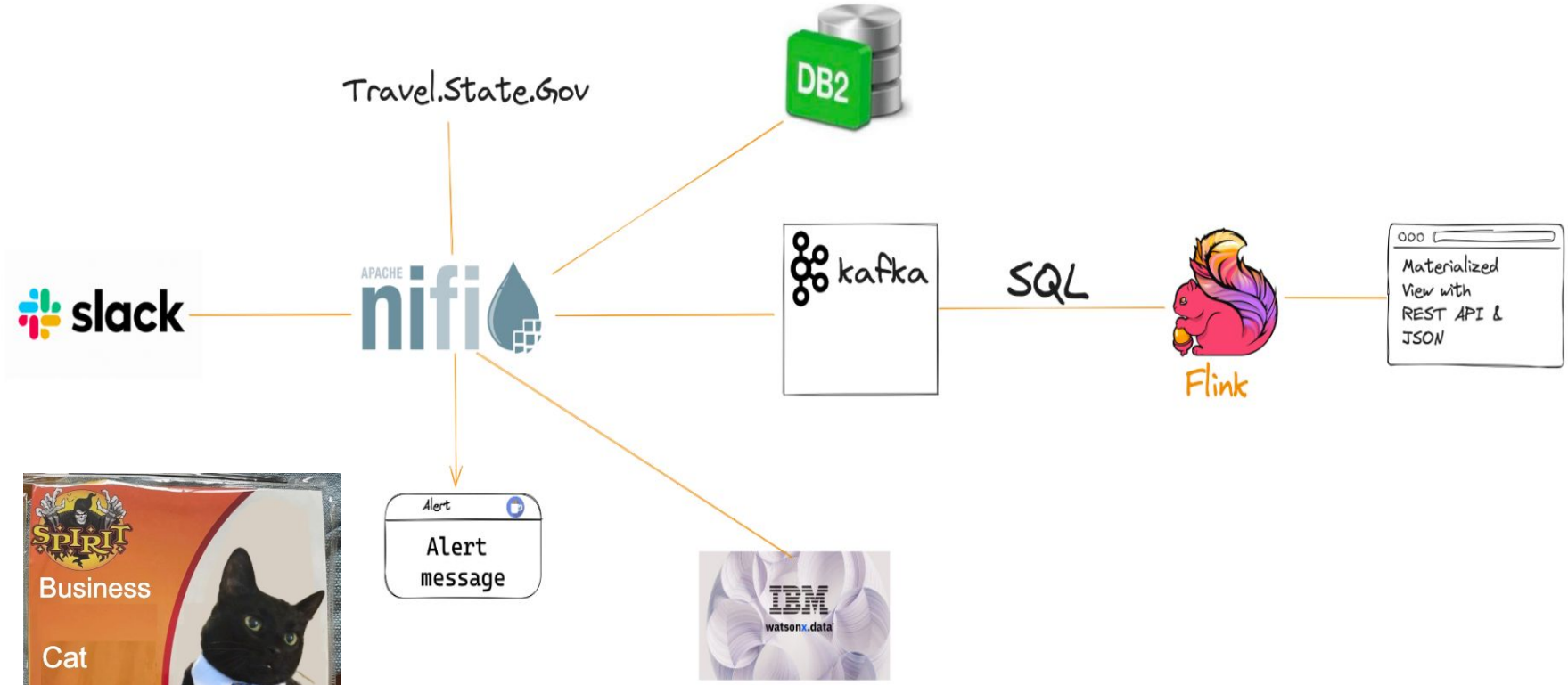
https://github.com/cloudera/CML_AMP_LLM_Chatbot_Augmented_with_Enterprise_Data

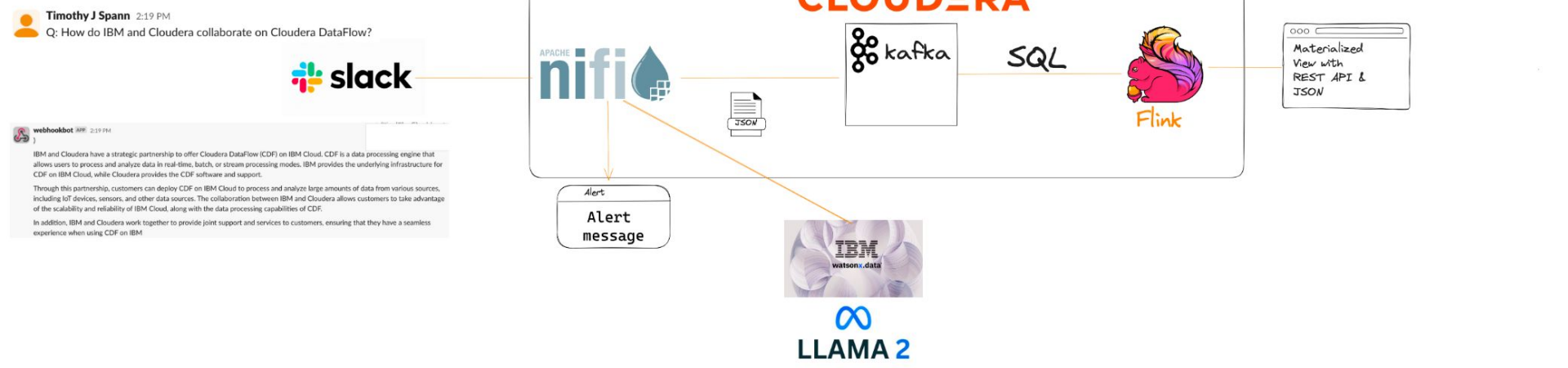
IN THE CURRENT CML PRODUCT, YOU CAN …

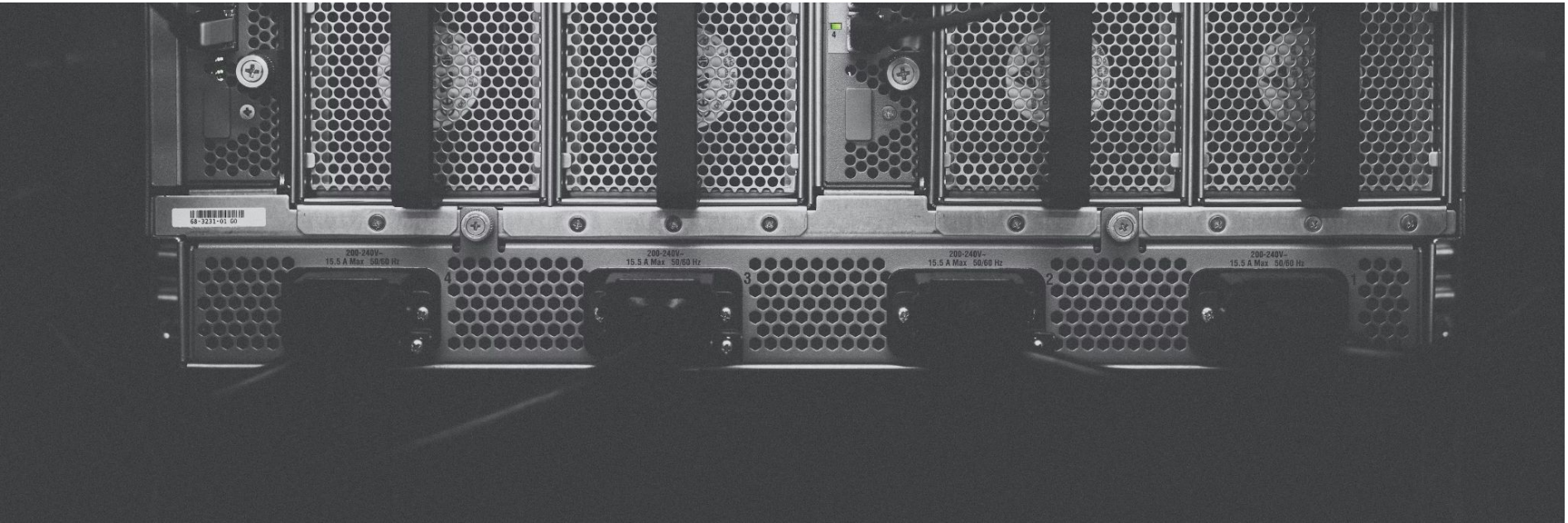
- Host and serve an open source LLM
- Create and host enterprise ready applications as front ends to these LLMs
- Instantiate a vector database to do semantic search on your enterprise knowledge base
- Provide enterprise specific context to an LLM to generate factual responses

All this without making any external calls to openAPI or any other SAAS AI service

DEMO AND Q&A




Timothy J Spann 2:19 PM
Q: How do IBM and Cloudera collaborate on Cloudera DataFlow?
slack
webhookbot APP 2:19 PM
)
IBM and Cloudera have a strategic partnership to offer Cloudera DataFlow (CDF) on IBM Cloud. CDF is a data processing engine that allows users to process and analyze data in real-time, batch, or stream processing modes. IBM provides the underlying infrastructure for CDF on IBM Cloud, while Cloudera provides the CDF software and support.
Through this partnership, customers can deploy CDF on IBM Cloud to process and analyze large amounts of data from various sources, including IoT devices, sensors, and other data sources. The collaboration between IBM and Cloudera allows customers to take advantage of the scalability and reliability of IBM Cloud, along with the data processing capabilities of CDF.
In addition, IBM and Cloudera work together to provide joint support and services to customers, ensuring that they have a seamless experience when using CDF on IBM
CLOUDERA
APACHE nifi
JSON
kafka
SQL
Flink
Materialized View with REST API & JSON
Alert
Alert message
IBM watsonx.data
LLAMA 2


68-3231-01 G0
200-240V~
15.5 A Max 50/60 Hz
4
3
2
1

FREE LEARNING ENVIRONMENT

Cloudera Streams Processing - Community Edition

- Kafka, KConnect, SMM, SR, Flink, and SSB in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- 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**
- Community Group Hub for CSP
- Find it on docs.cloudera.com under Applications

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

Open Source Edition

- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

- Licensed under the ASF License
- **Unsupported**

https://hub.docker.com/r/apache/nifi

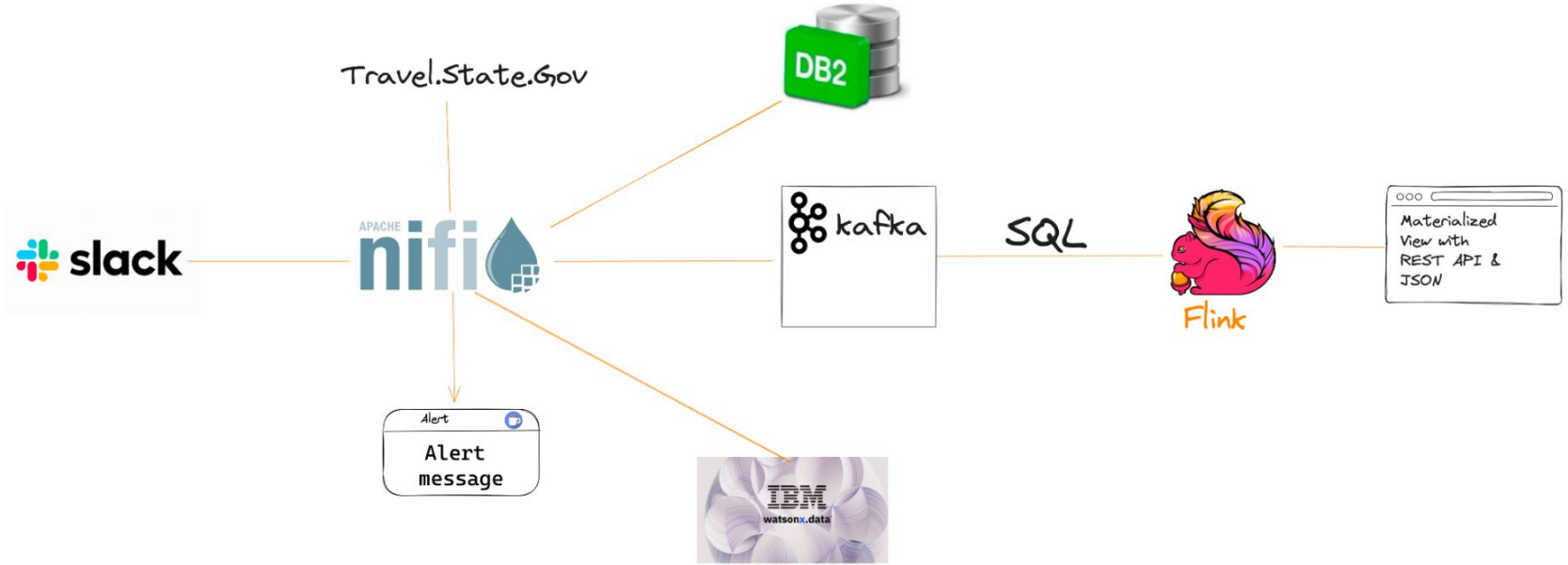
RESOURCES AND WRAP-UP

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 1.23.2 is the latest release, NiFi 2.0 coming soon (NiFi 1.10 - 4 years ago)
14M+ docker pulls of the Apache NiFi image (1M+ - 4 years ago)

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



https://www.youtube.com/watch?v=Y1JeOrJIoKI
https://www.youtube.com/watch?v=he7bcQDbAIQ&t=6s
https://www.youtube.com/watch?v=RPz7Xm4fLF4

Streaming Resources

- https://dzone.com/articles/real-time-stream-processing-with-hazelcast-and-streamnative
- https://flipstackweekly.com/
- https://www.datainmotion.dev/
- https://www.flankstack.dev/
- https://github.com/tspannhw
- https://medium.com/@tspann
- https://medium.com/@tspann/predictions-for-streaming-in-2023-ad4d7395d714
- https://www.apachecon.com/acna2022/slides/04_Spann_Tim_Citizen_Streaming_Engineer.pdf

Java - LLM Resources

- https://github.com/TheoKanning/openai-java
- https://github.com/Knowly-ai/langtorch
- https://onnxruntime.ai/docs/get-started/with-java.html
- https://github.com/tjake/Jlama
- https://github.com/HamaWhiteGG/langchain-java
- https://nightlies.apache.org/flink/flink-ml-docs-master/docs/operators/feature/sqltransformer/
- https://research.ibm.com/blog/retrieval-augmented-generation-RAG