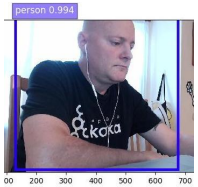
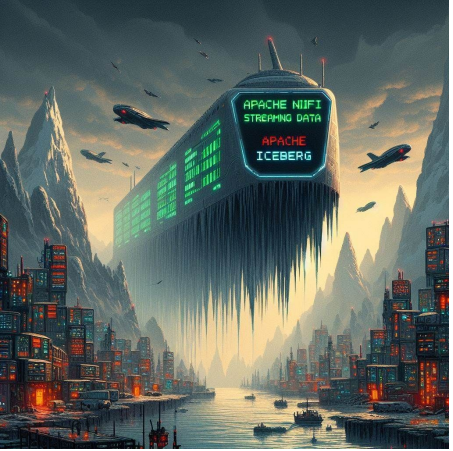
CLOUDERA

Building Real-Time Generative AI Pipelines

Tim Spann
Principal Developer Advocate

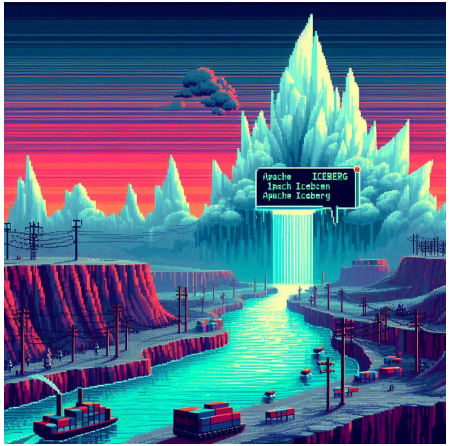
March 15, 2024

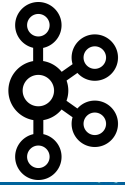



APACHE NIFI
STREAMING DATA
APACHE
ICEBERG


APACHE
ICEBEG
APACHE

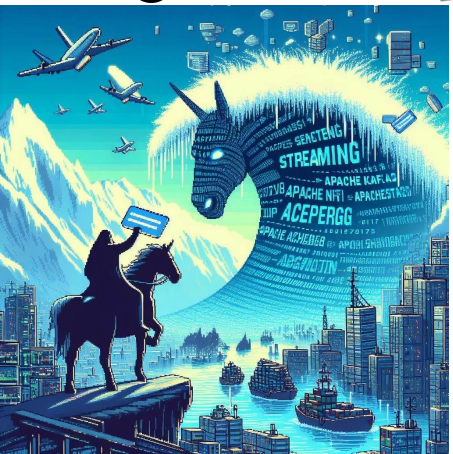

Apache NiFi

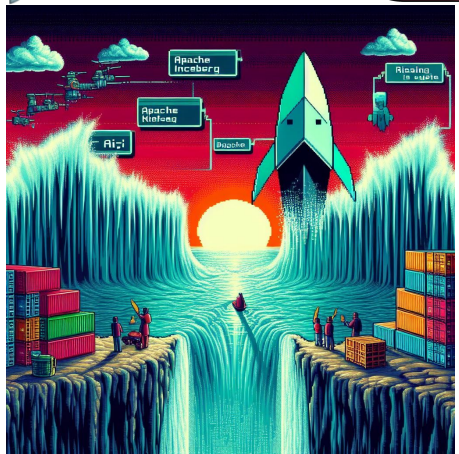

Apache ICEBERG
Ipach Icebcan
Apache Iceberg

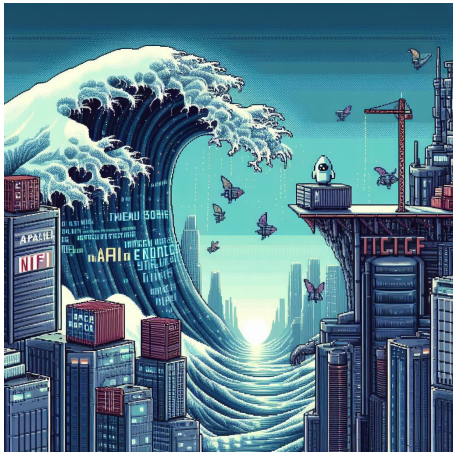



nifi

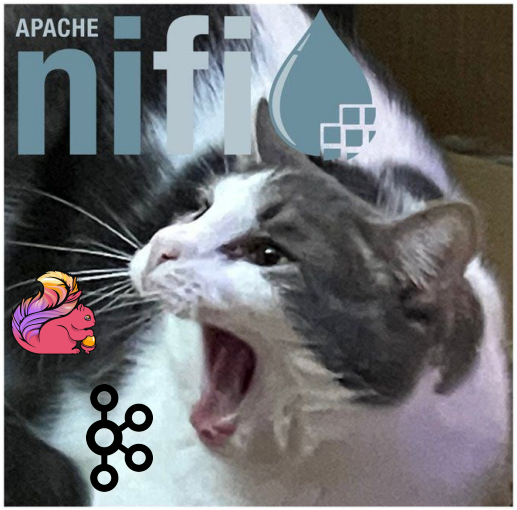

ICEBERG


STREAMING
APACHE KAFKA
APACHE NIFI
ACEPERGG


Apache Inceberg
Apache Ninfeng
Ai:i


APACHE
NIFI

FLaNK Stack Weekly by Tim Spann



https://bit.ly/32dAJft

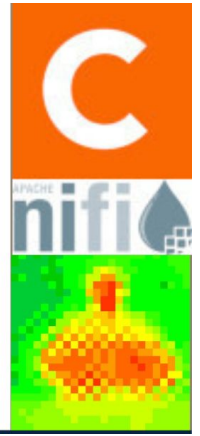
https://www.meetup.com/futureofdata-princeton/

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

Future of Data - NYC + NJ + Philly + Virtual

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY



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

https://linktr.ee/tspannhw

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

Tim Spann

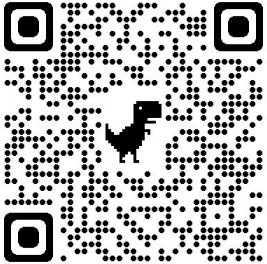
Twitter: @PaasDev Blog: datainmotion.dev

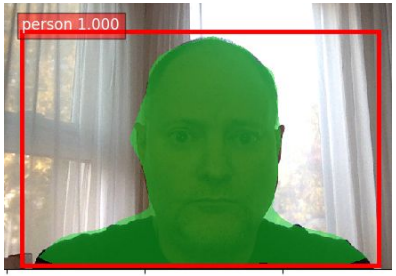
Principal Developer Advocate

Princeton Future of Data Meetup

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

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





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)

Some common Vector DBs

Chroma, Pinecone, Solr, pgvector, Milvus, Weaviate

Open Source vs Self Hosted vs SaaS option

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

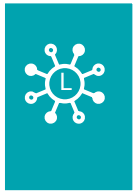
- Gauge emotions and opinions based on a piece of text.
- Understand and deliver a more nuanced message back based on sentiment.

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.

Generative AI

NLP / AI / LLM

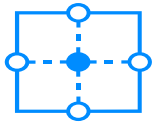
Which Model and When?

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

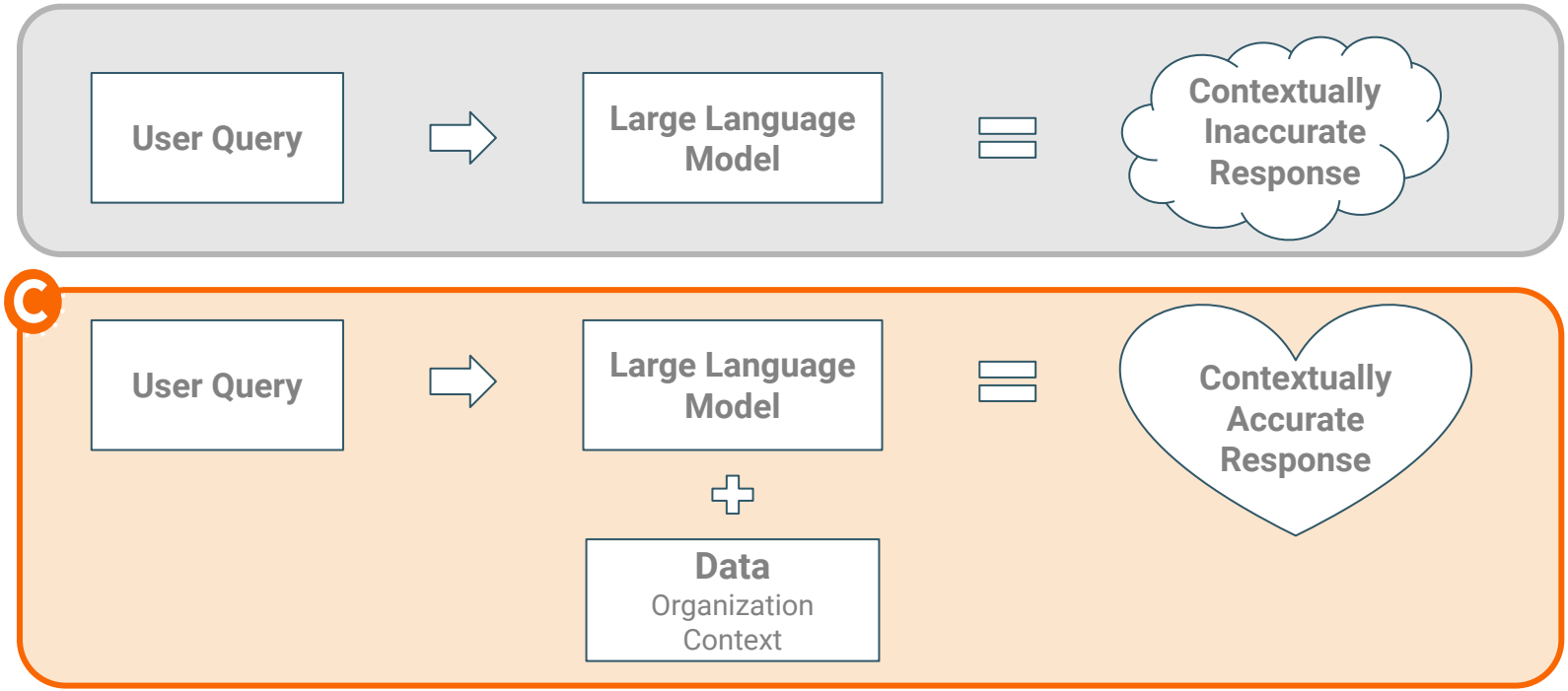
Adoption of Generative AI is a Journey

Identifying AI challenges in the enterprise

Challenges		What's missing
Data integration barriers		• Streamlined access to enterprise data
Rigid model infrastructure		• Modularity • Flexibility • AI Ops
Lack of security and transparency		• Model control • Built-in security • Visibility & governance

Data = Organization Context

Your data enables contextually accurate responses from LLMs



AI 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

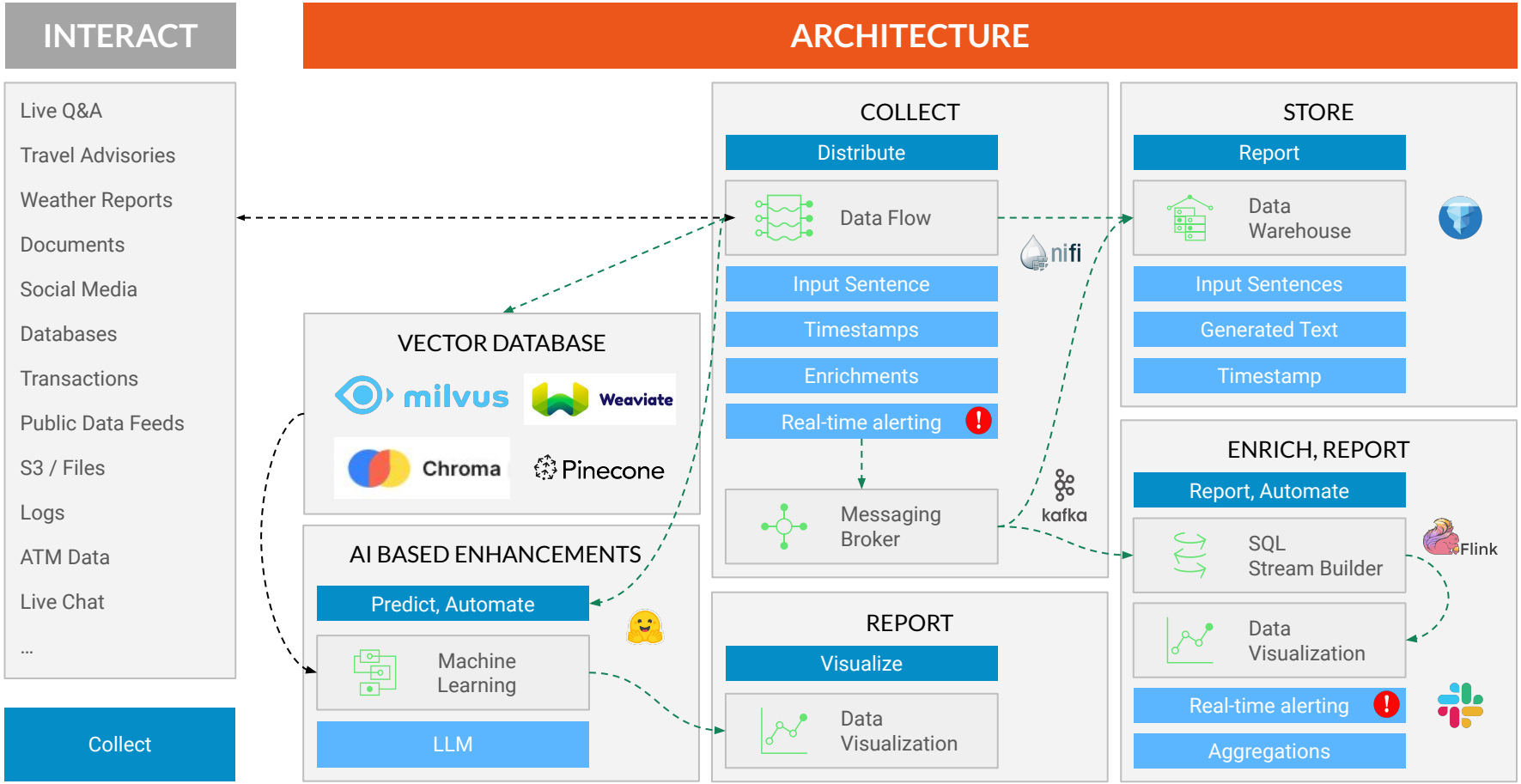
REAL-TIME DATA INGEST & ROUTING ▶ DATA WRANGLING ▶ DATA STORE & VISUALIZATION ▶ AI MODEL TRAINING & SERVING

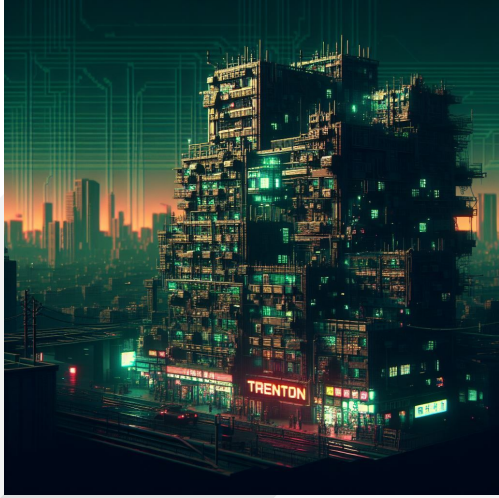
CLOUD INFRASTRUCTURE

aws Azure Google Cloud Red Hat

SPECIALIZED HARDWARE

NVIDIA Google AMD IBM intel DELL Technologies


INTERACT
ARCHITECTURE
Live Q&A
Travel Advisories
Weather Reports
Documents
Social Media
Databases
Transactions
Public Data Feeds
S3 / Files
Logs
ATM Data
Live Chat
...
Collect
VECTOR DATABASE
milvus
Weaviate
Chroma
Pinecone
AI BASED ENHANCEMENTS
Predict, Automate
Machine Learning
LLM
COLLECT
Distribute
Data Flow
nifi
Input Sentence
Timestamps
Enrichments
Real-time alerting
Messaging Broker
kafka
REPORT
Visualize
Data Visualization
STORE
Report
Data Warehouse
Input Sentences
Generated Text
Timestamp
ENRICH, REPORT
Report, Automate
SQL Stream Builder
Flink
Data Visualization
Real-time alerting
Aggregations



DATAFLOW / STREAMING

LLM USE CASE

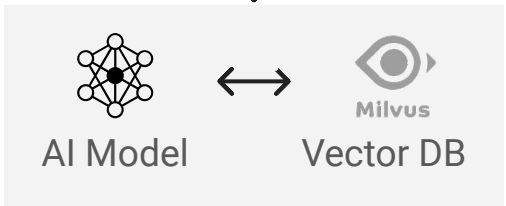
Unstructured file types

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

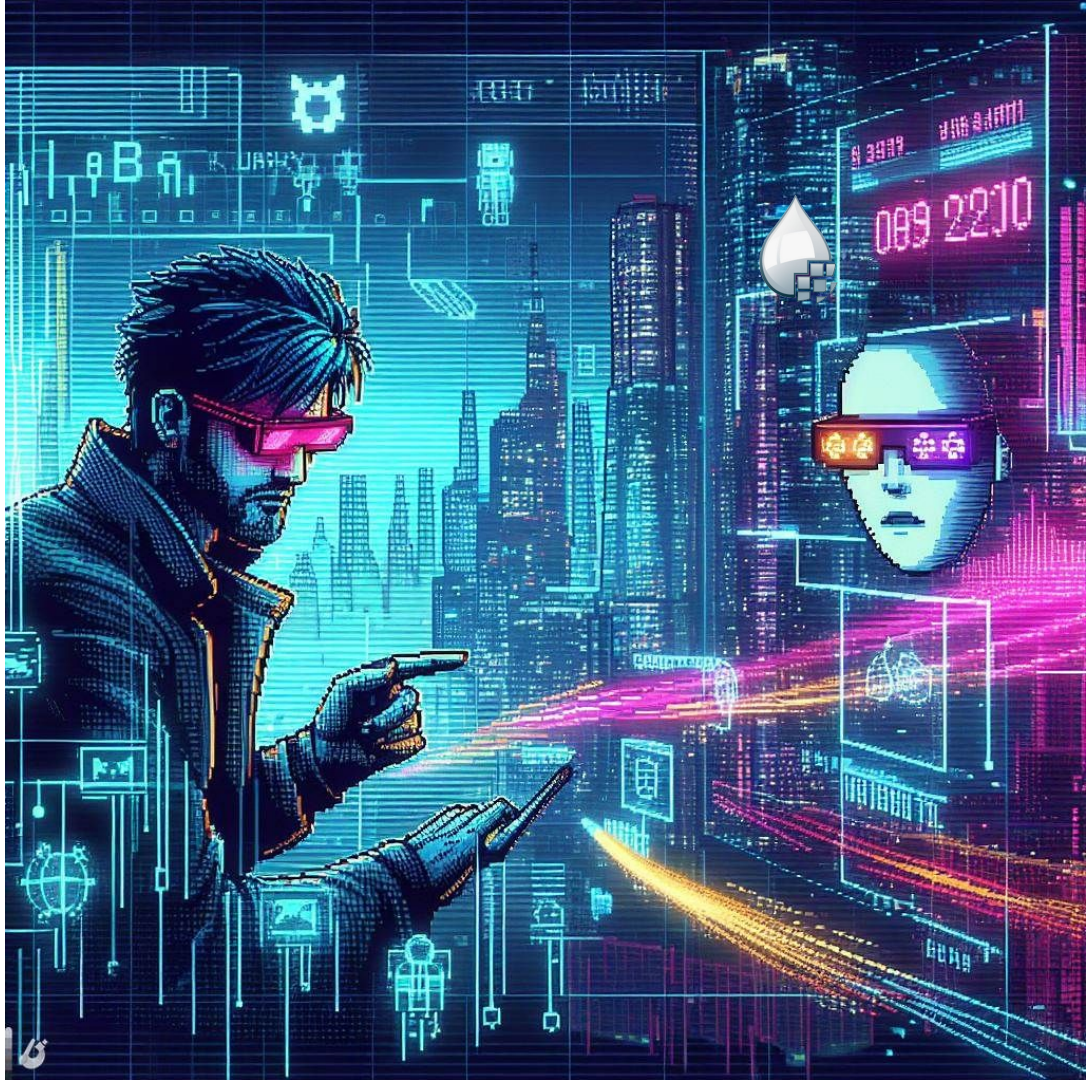
Open Data Lakehouse

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



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

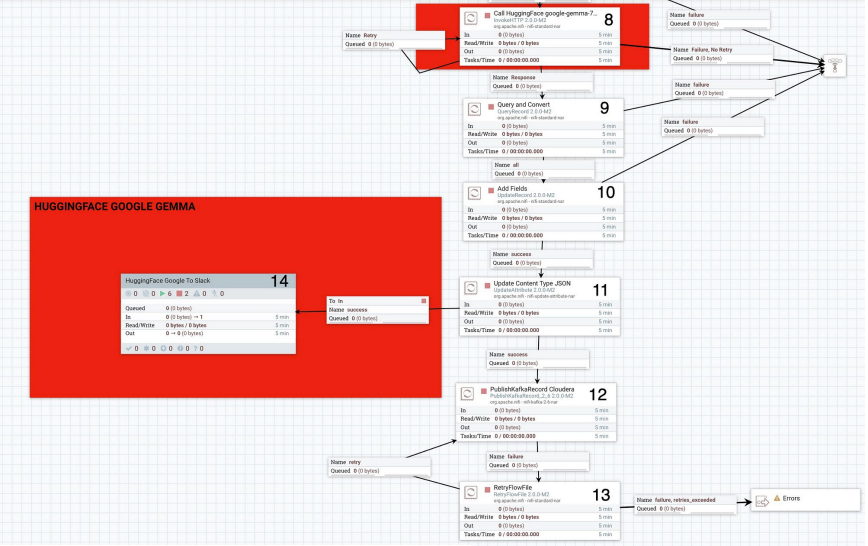
Act as a Discord, REST, Kafka, SQL, Slack bot to roundtrip discussions

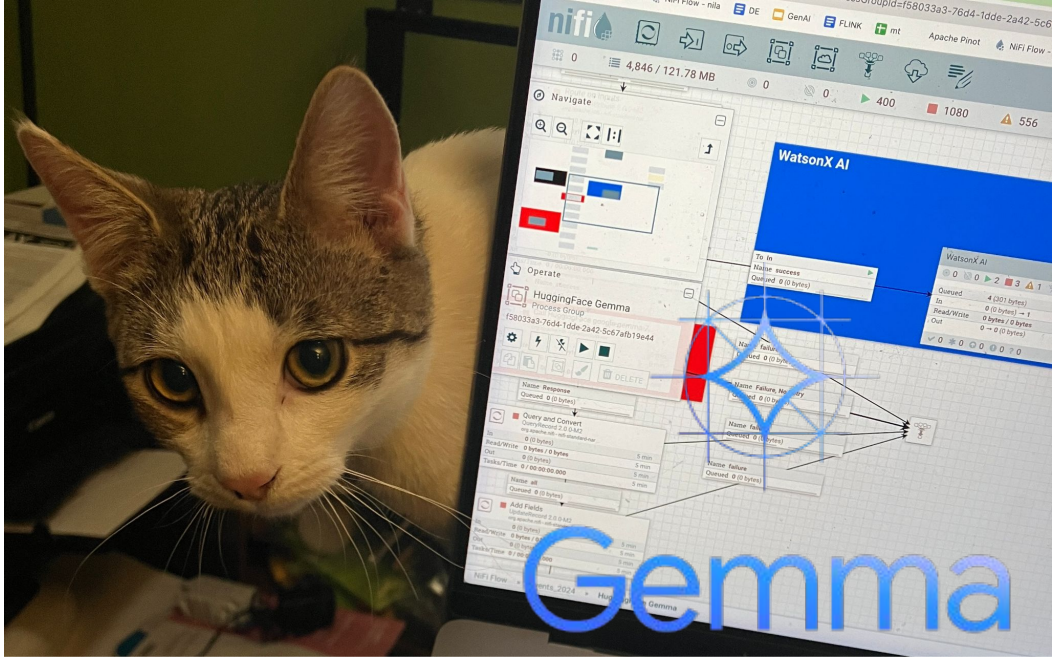
UNSTRUCTURED DATA WITH NIFI

- **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

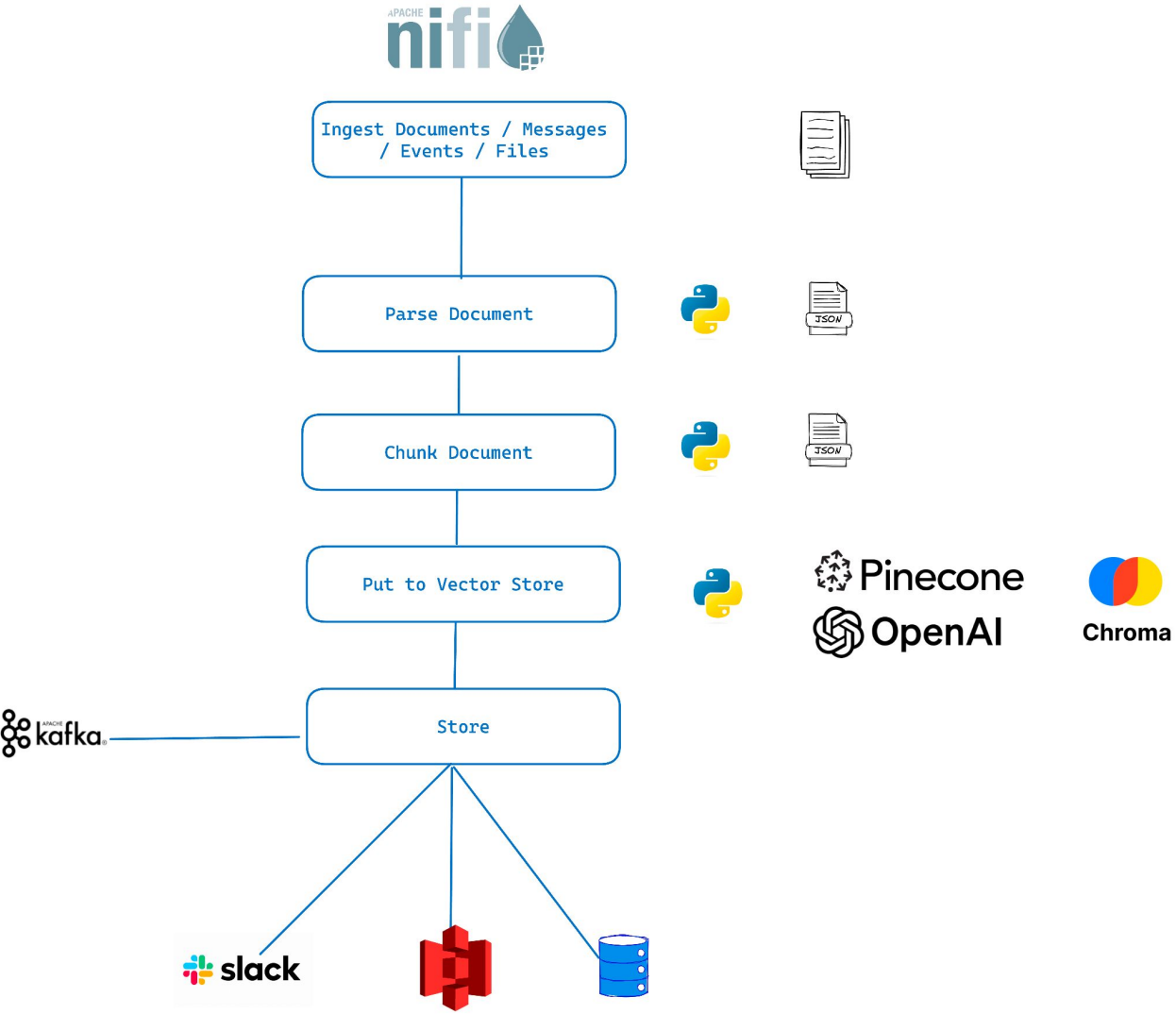
CLOUD ML/DL/AI/Vector Database Services

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

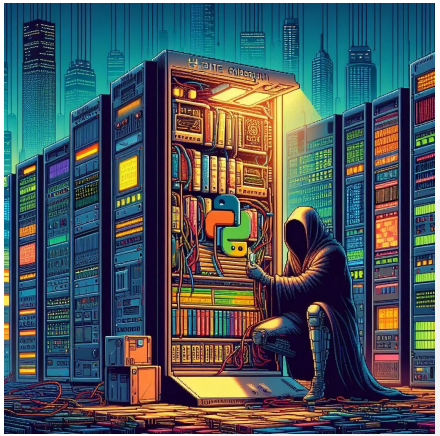




https://medium.com/cloudera-inc/google-gemma-for-real-time-lightweight-open-llm-inference-88efe98e580f



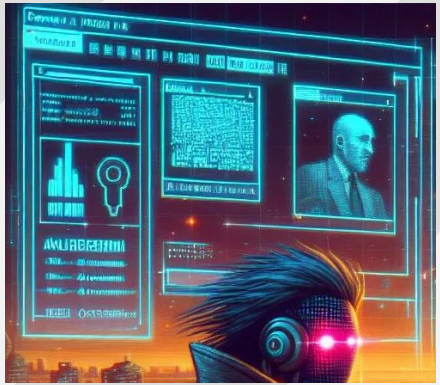


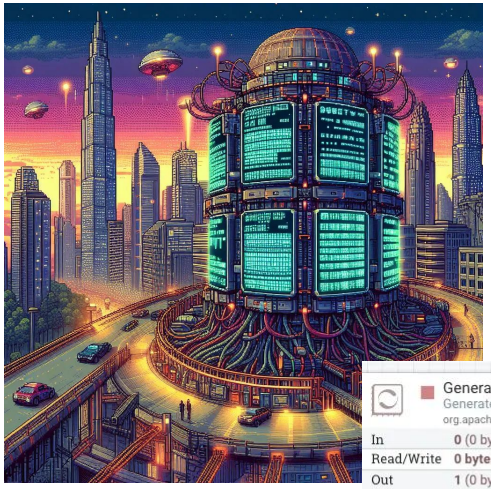



nifi

Python Processors

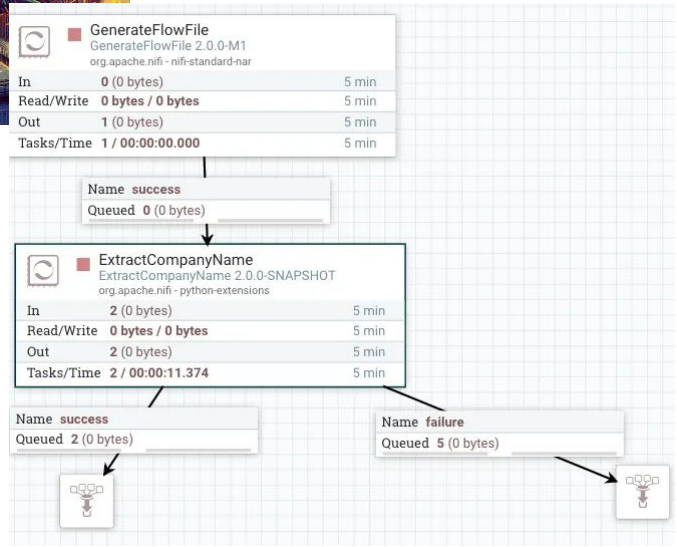






Extract Company Names

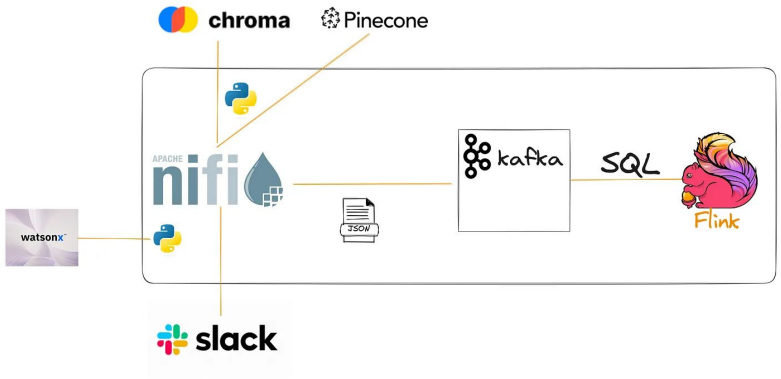
- Python 3.10+
- Hugging Face, NLP, SpaCY, PyTorch



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

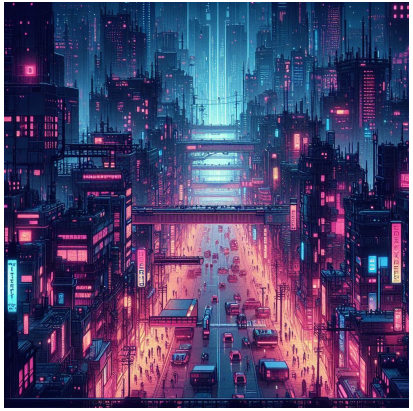


https://github.com/tspannhw/FLaNK-python-watsonx-processor

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

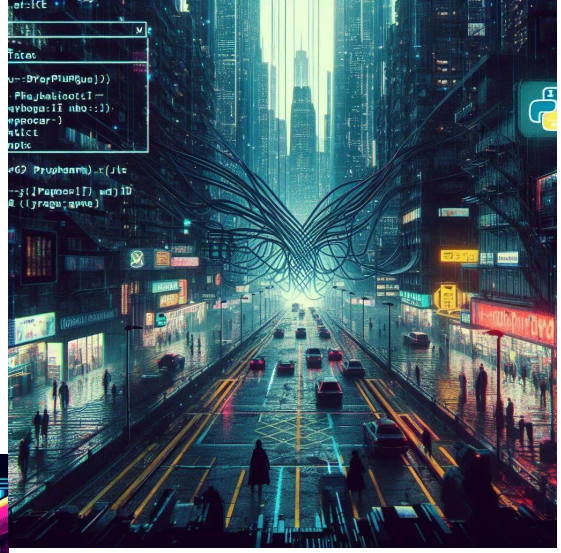
- Python 3.10+
- Hugging Face
- Transformers
- facial_emotions_image_detection
- Image Classification
- Adds labels/scores to FlowFile Attributes
- Does not require download or copies of your images

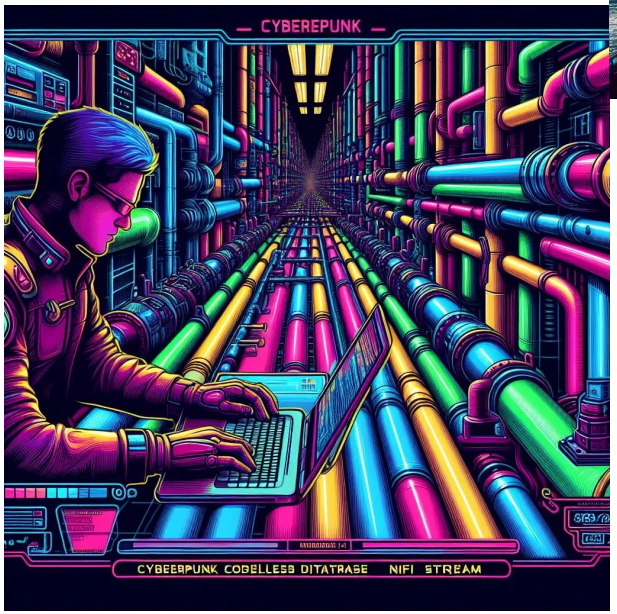
https://github.com/tspannhw/FLaNK-python-processors

Other Python Processors

- Chunk Document, Parse Document
- Prompt Chat GPT
- Put Chroma, Query Chroma
- Put Pinecone, Query Pinecone

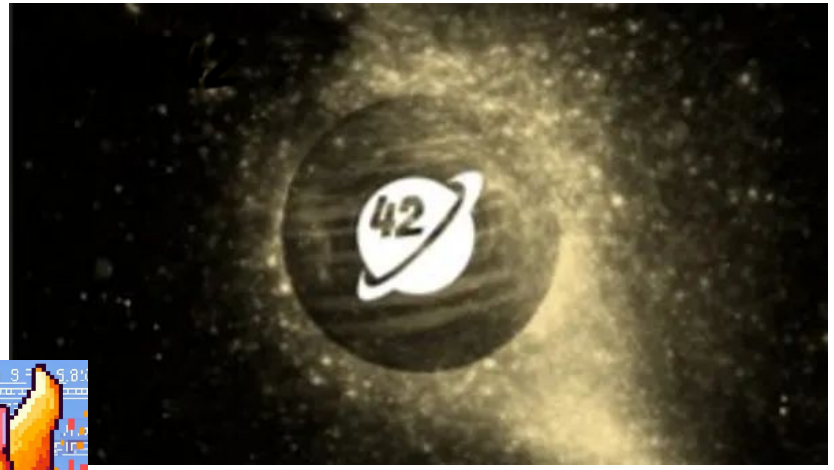
DEMO



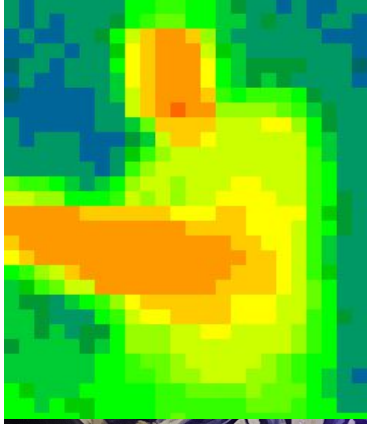


DATA SUMMIT
UNLEASH THE POWER OF YOUR DATA

MAY 8–9
BOSTON, MA



Startup Grind AI Max Summit - April 12 NJ



THANK YOU

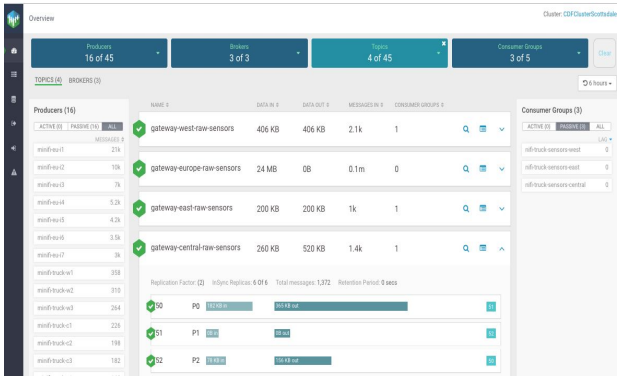
APACHE nifi

CLOUDERA STREAM PROCESSING

Two Major Capabilities: Enterprise Messaging and Powerful Stream Processing

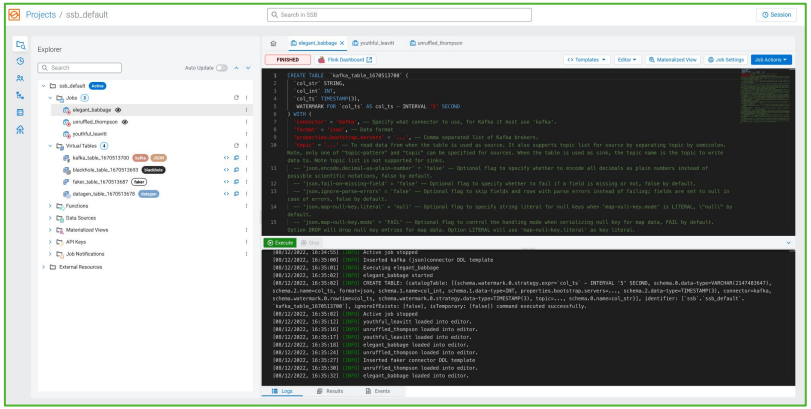
Cloudera Streams Messaging (CSM) Powered by Apache Kafka

Enterprise grade messaging products for *Apache Kafka*. *Streams Messaging Manager* to monitor/operate clusters, *Streams Replication Manager* for HA/DR deployments, *Schema Registry* for centralized schema management, and support for *Kafka Connect* and *Cruise Control*



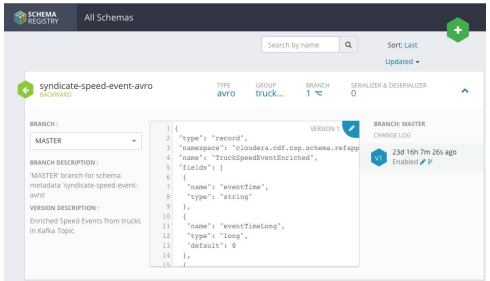
Cloudera Streaming Analytics (CSA) Powered By Apache Flink

Powered by *Apache Flink with SQL StreamBuilder, it provides* low-latency stream processing capabilities with advanced windowing & state management made simple with SQL

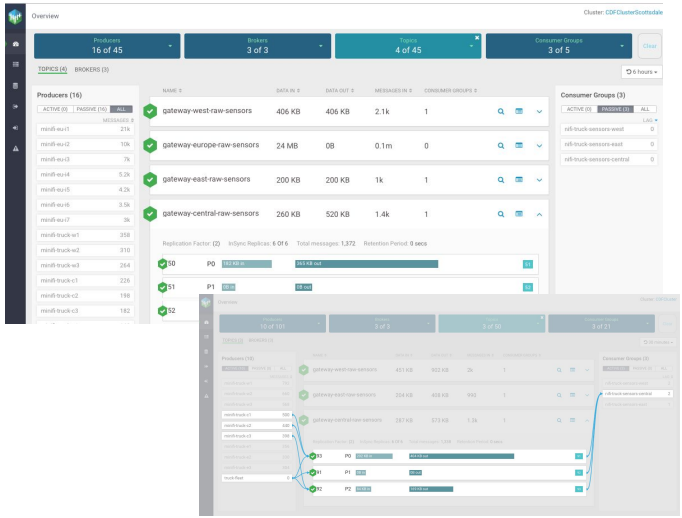


ENTERPRISE MANAGEMENT CAPABILITIES FOR APACHE KAFKA
Extend streams messaging services for Schema Mgmt, Replication & Monitoring

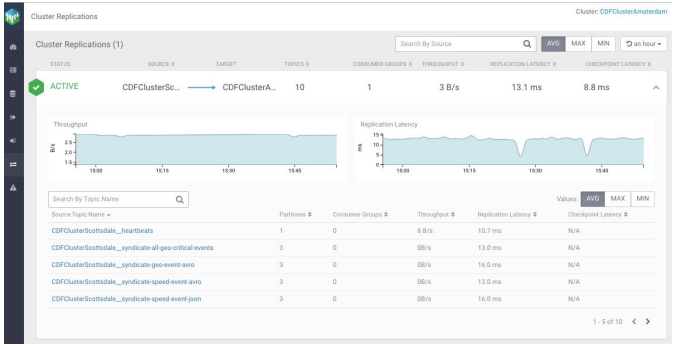
Schema Registry
Kafka Schema Governance



Streams Messaging Manager
Management & Monitoring Service
for all of your Kafka clusters



Streams Replication Manager
Kafka Replication Service
for Disaster Recovery



ENTERPRISE MANAGEMENT CAPABILITIES FOR APACHE KAFKA

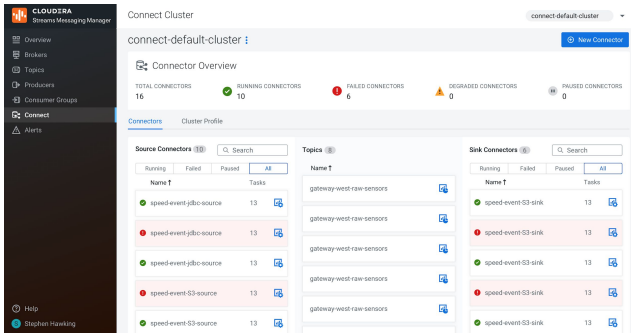
Kafka Data Movement, Operations and Security Made Easier

Kafka Connect Support

Simple Data Movement

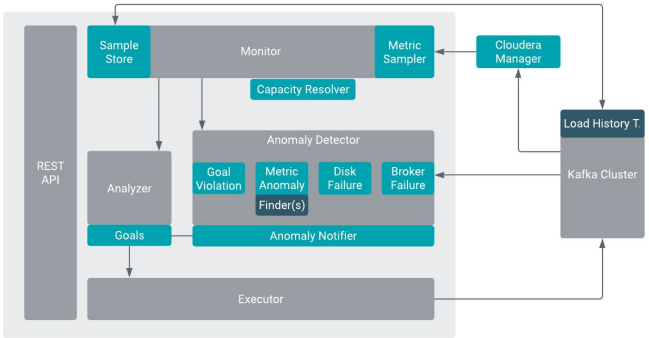
Change Data Capture Connectors

Build Custom Connectors with NiFi



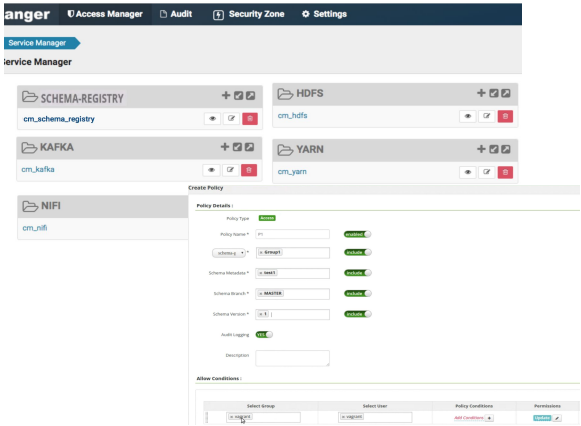
Cruise Control Support

Intelligent Rebalancing & Self-Healing of your Kafka Clusters



Ranger Security

Improved ACL and Audit for Kafka, KConnect and Schema Registry

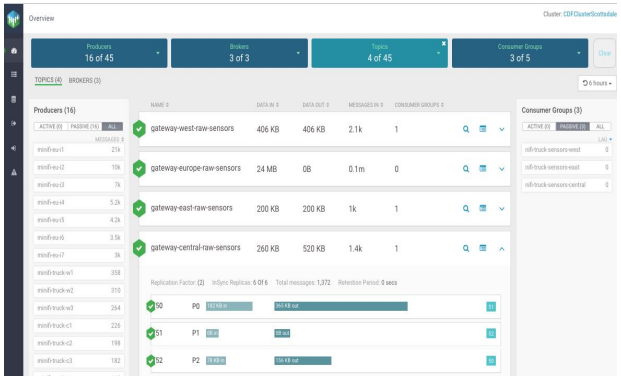


CLOUDERA STREAM PROCESSING

Two Major Capabilities: Enterprise Messaging and Powerful Stream Processing

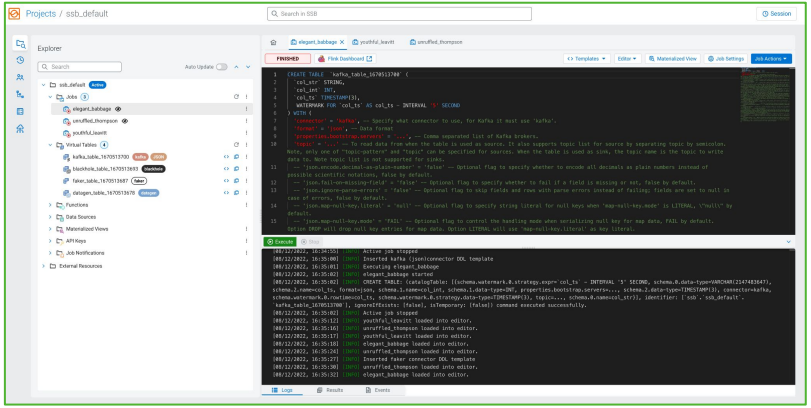
Cloudera Streams Messaging (CSM) Powered by Apache Kafka

Enterprise grade messaging products for *Apache Kafka*. *Streams Messaging Manager* to monitor/operate clusters, *Streams Replication Manager* for HA/DR deployments, *Schema Registry* for centralized schema management, and support for *Kafka Connect* and *Cruise Control*



Cloudera Streaming Analytics (CSA) Powered By Apache Flink

Powered by *Apache Flink with SQL StreamBuilder, it provides* low-latency stream processing capabilities with advanced windowing & state management made simple with SQL



NEXT GENERATION STREAMING ANALYTICS WITH APACHE FLINK

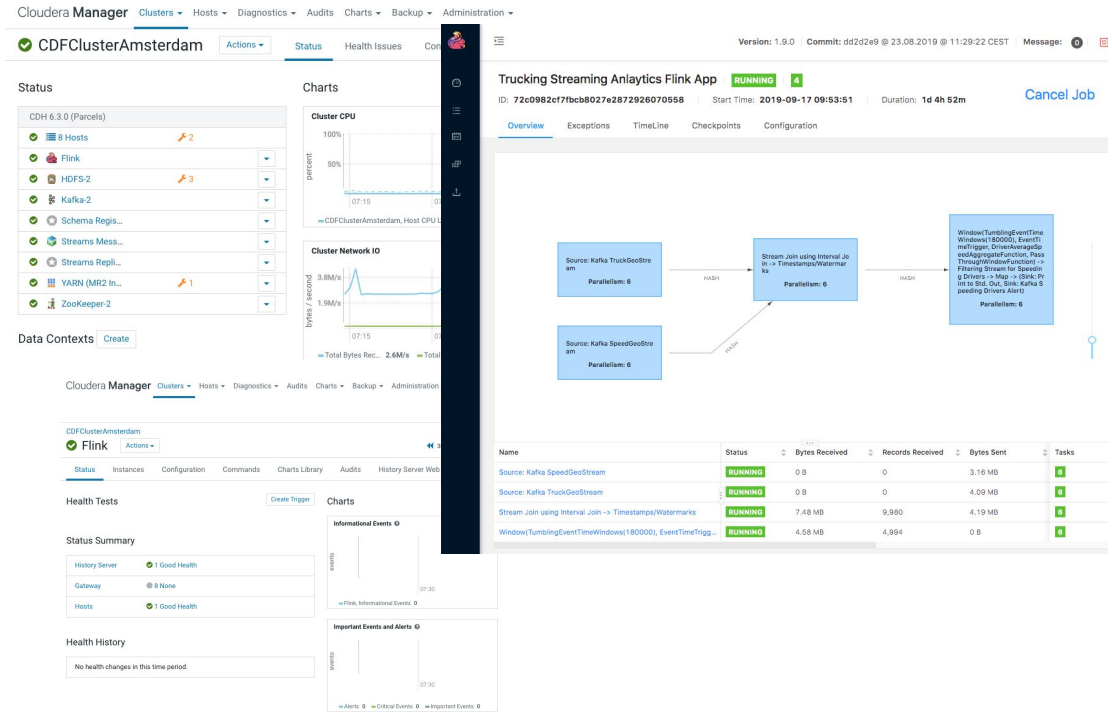
Low latency stateful stream processing

Low Latency

Event Processing

Real-Time Insights

- Flink is a distributed data processing systems ideally suited for real-time, event driven applications.
- Unifies stream and batch processing
- Advanced features - late arriving data, checkpointing, event time processing, Exactly Once Processing



SQL STREAM BUILDER (SSB)

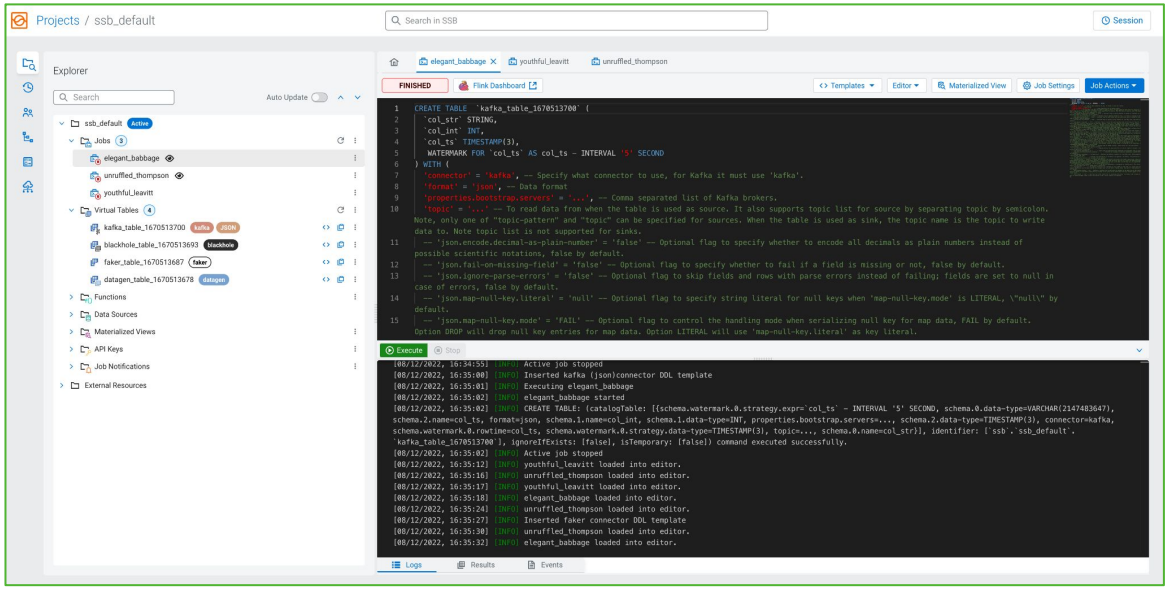
Democratize access to real-time data with just SQL

SQL STREAM BUILDER allows developers, analysts, and data scientists to **write streaming applications** with industry standard **SQL**.

No Java or Scala code development required.

Simplifies access to data in Kafka & Flink. Connectors to batch data in HDFS, Kudu, Hive, S3, JDBC, CDC and more

Enrich streaming data with batch data in a single tool



LLMs ARE FOUNDATION MODELS

Base models that can be adapted for a wide range of use cases

- ➔ Historically, data scientists trained specialized models against narrow datasets to solve specific tasks.
- ➔ LLMs are **Foundation models** that can be adapted to perform a variety of tasks.
  - ◆ It is faster to “adapt” a foundation model than it is to train a specialized model from scratch
  - ◆ Decouples “knowledge” from “intelligence”
  - ◆ Opens up AI use cases to software developers (instead of just specialised data scientists)