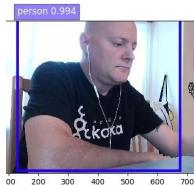




Building Real-Time Generative AI Pipelines

Tim Spann
Principal Developer Advocate

March 15, 2024





APACHE
NIFI



APACHE
ICEBERG



FLaNK Stack Weekly by Tim Spann



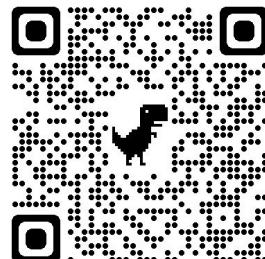
<https://bit.ly/32dAJft>

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



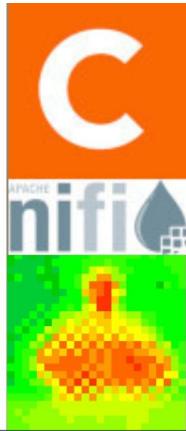
This week in Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java and Open Source friends.

Future of Data - NYC + NJ + Philly + Virtual



FUTURE OF DATA

AN OPEN SOURCE COMMUNITY



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

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 ...



@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>

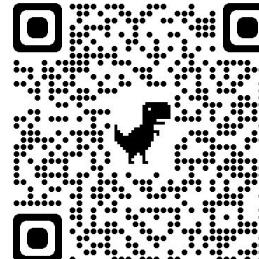


DZone REFCARDS TREND REPORTS E

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...



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)

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
 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.

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

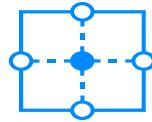
Identifying AI challenges in the enterprise

Challenges

Data integration barriers



Rigid model infrastructure



Lack of security and transparency



What's missing

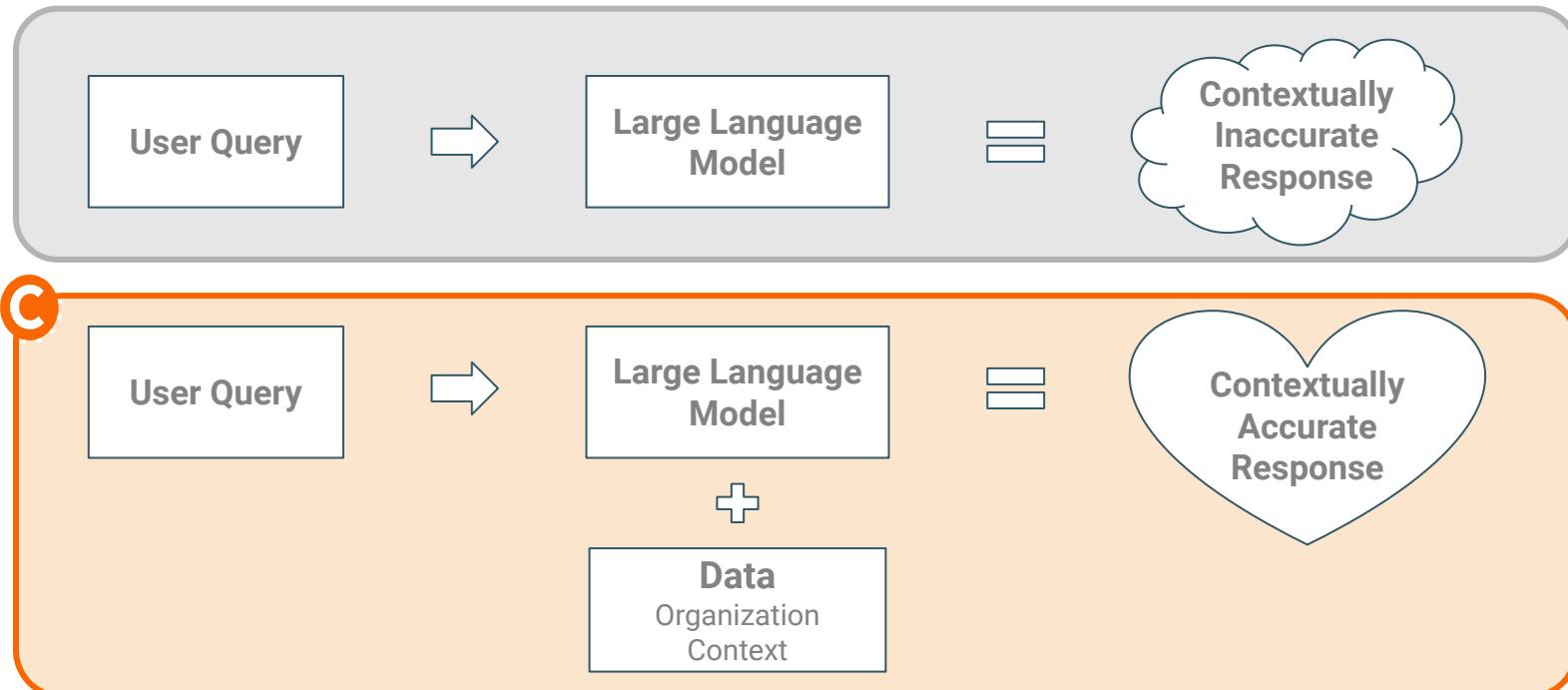
- Streamlined access to enterprise data

- Modularity
- Flexibility
- AI Ops

- 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...



MODEL HUBS
Hugging Face



FINE-TUNED MODELS

Meta (Llama 2)

OPEN SOURCE FOUNDATION MODELS

MANAGED VECTOR STORE
Pinecone



PRIVATE VECTOR STORE
Milvus, Solr*

CLOUDERA
Open Data Lakehouse



CLOUD INFRASTRUCTURE



Azure

Google Cloud

Red Hat

SPECIALIZED HARDWARE



NVIDIA

Google

AMD

IBM

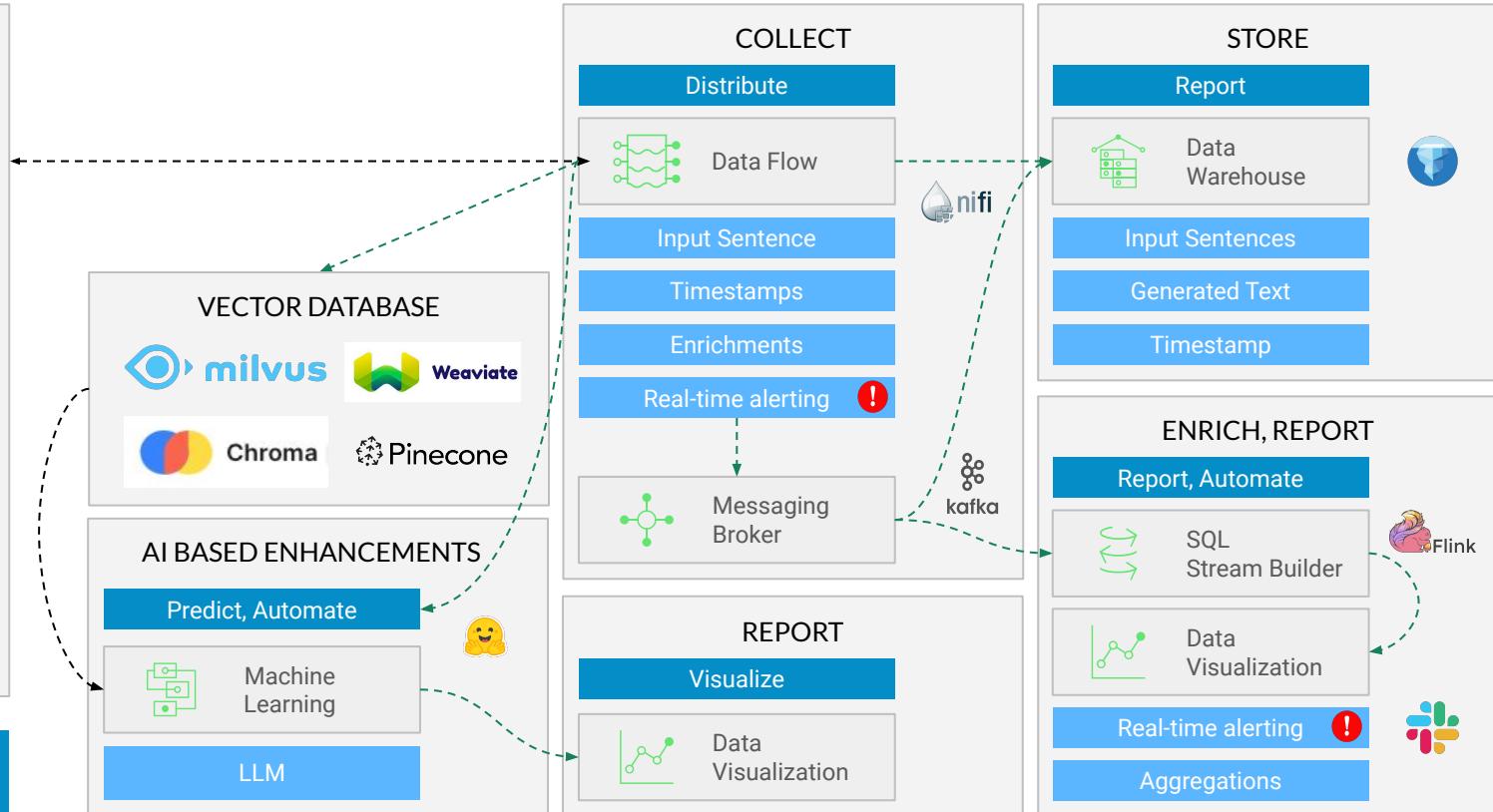
intel

DELL Technologies

INTERACT

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

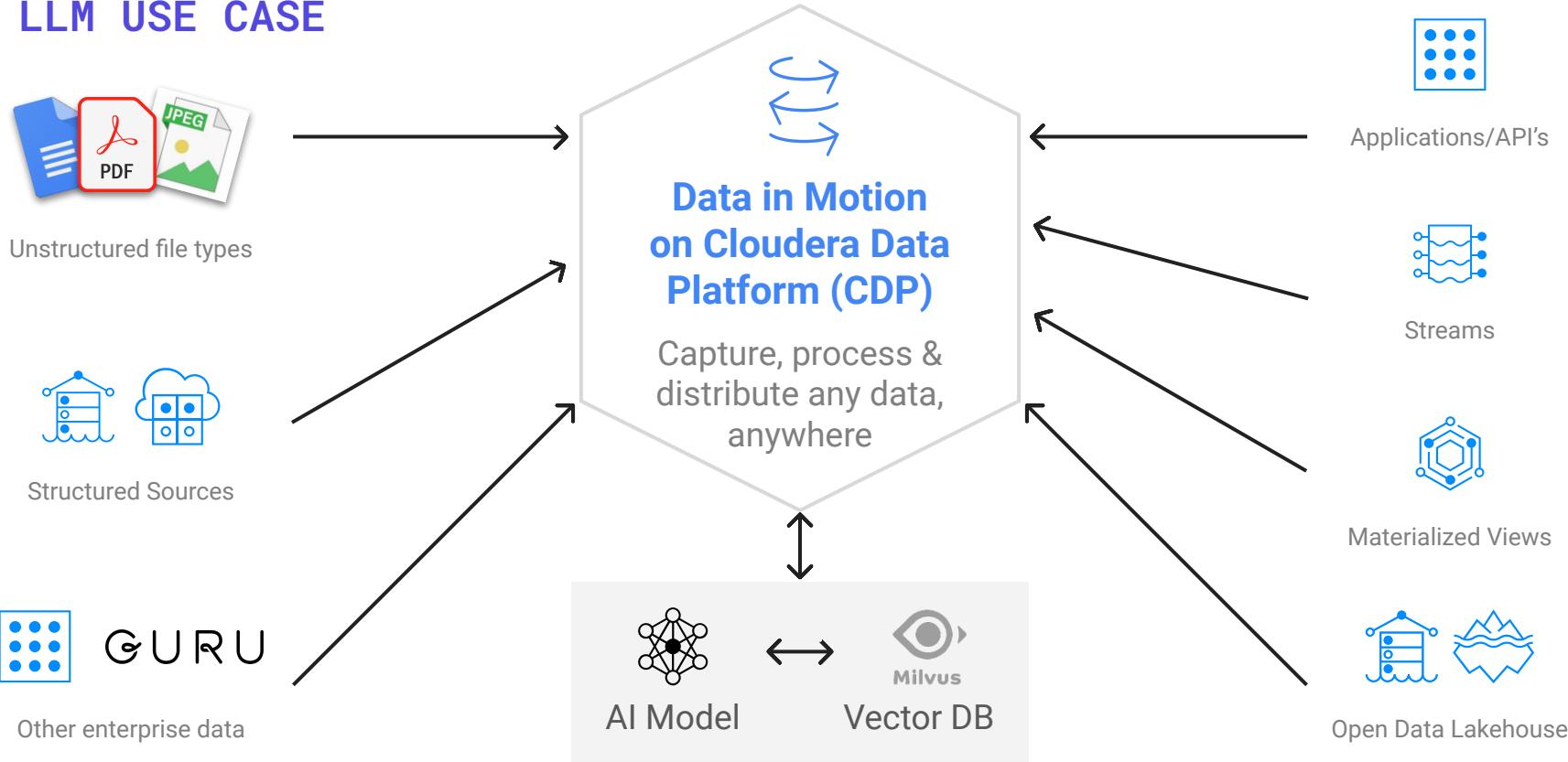
ARCHITECTURE



DATAFLOW / STREAMING



LLM USE CASE





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

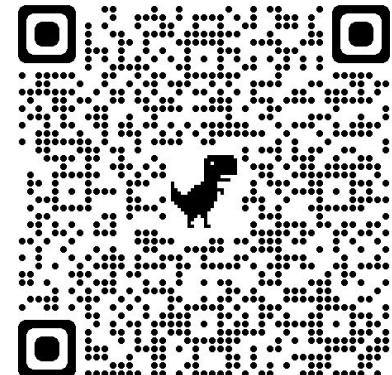
UNSTRUCTURED DATA WITH NIFI

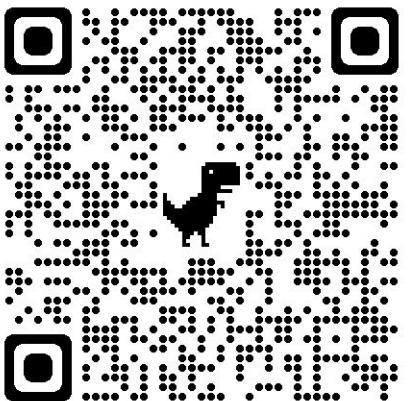
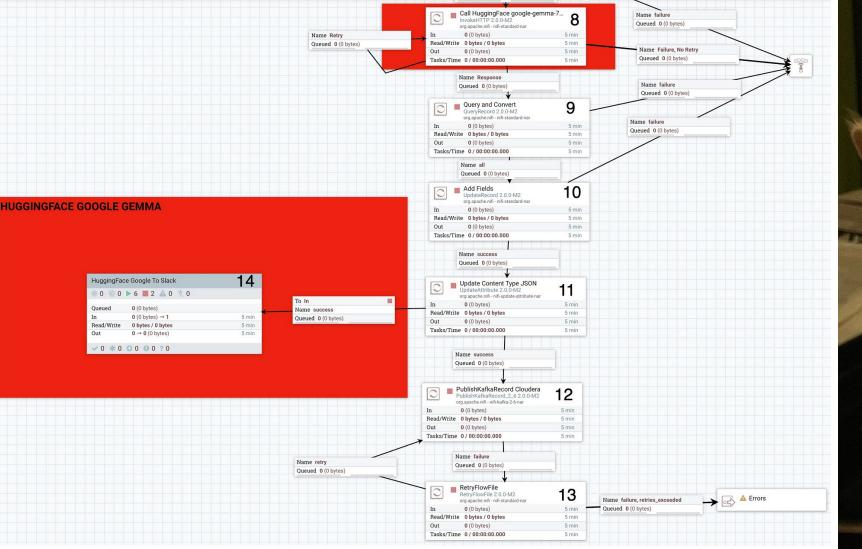
- **Archives** - tar, gzipped, zipped, ...
- **Images** - PNG, JPG, GIF, BMP, ...
- **Documents** - HTML, Markdown, RSS, PDF, Doc, RTF, Plain Text, ...
- **Videos** - MP4, Clips, Mov, Youtube URL...
- **Sound** - MP3, ...
- **Social / Chat** - Slack, Discord, Twitter, REST, Email, ...
- **Identify Mime Types, Chunk Documents, Store to Vector Database**
- **Parse Documents** - HTML, Markdown, PDF, Word, Excel, Powerpoint



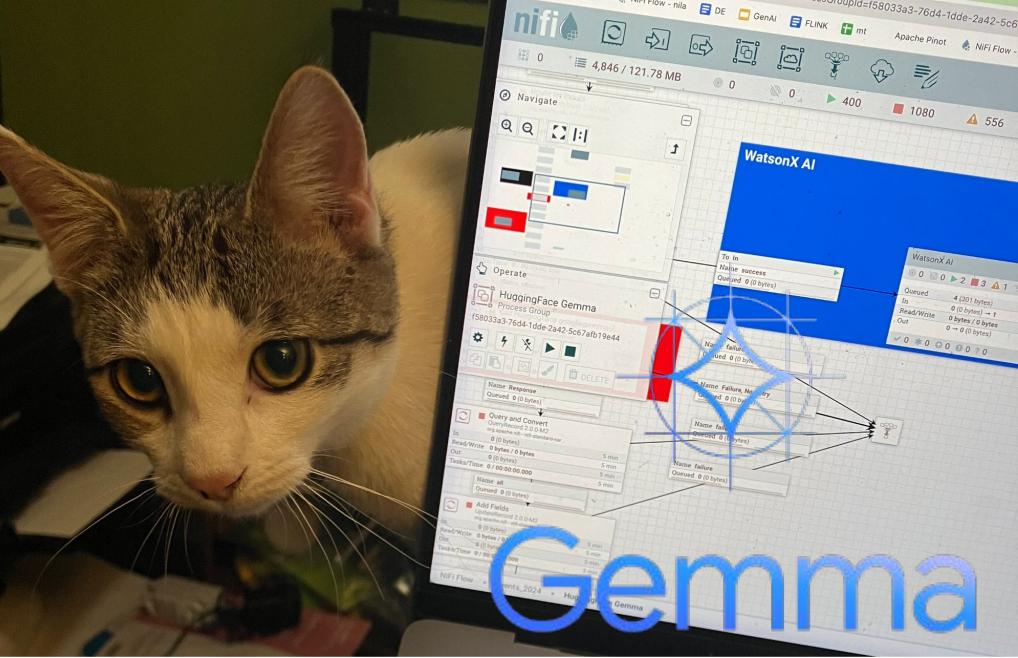
CLOUD ML/DL/AI/Vector Database Services

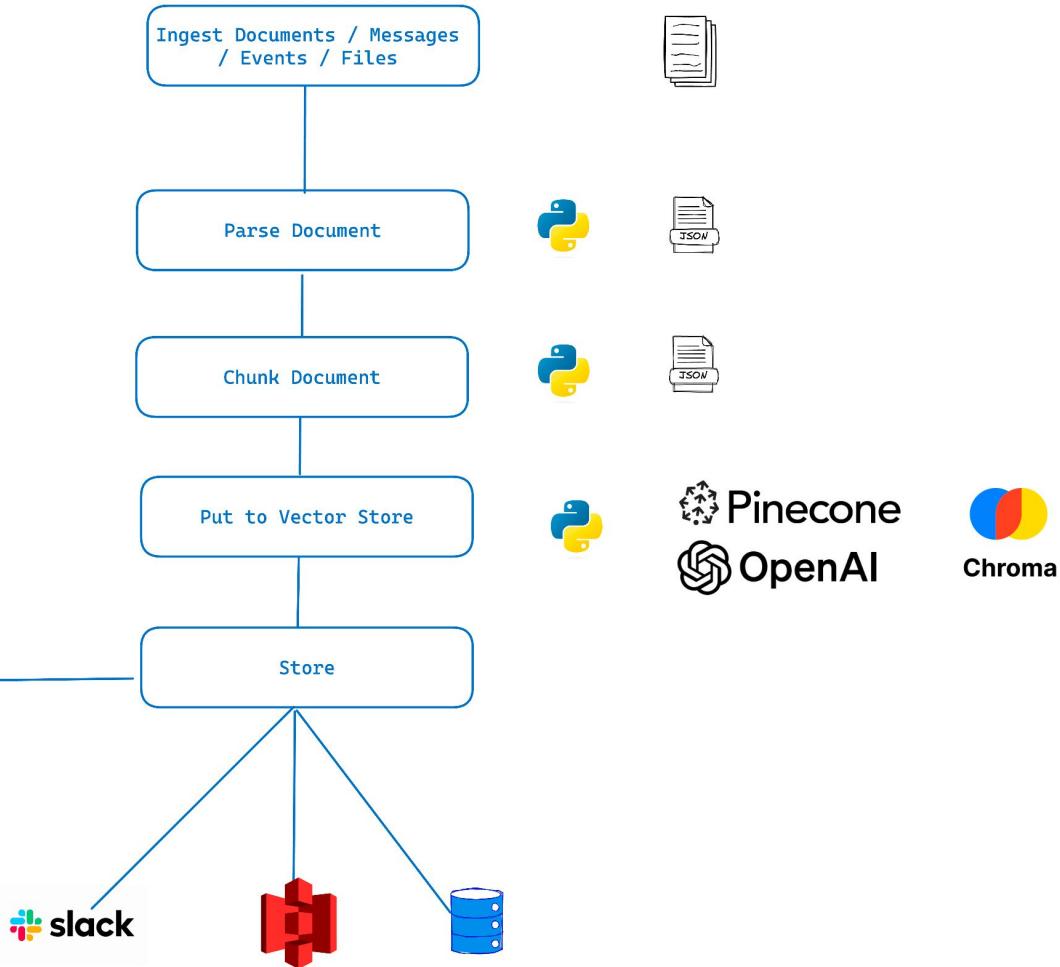
- Cloudera ML
- Amazon Polly, Translate, Textract, Transcribe, Bedrock, ...
- Hugging Face
- IBM Watson X.AI
- Vector Stores Anywhere: Weaviate, Pinecone, Milvus, Chroma DB, SOLR, ...

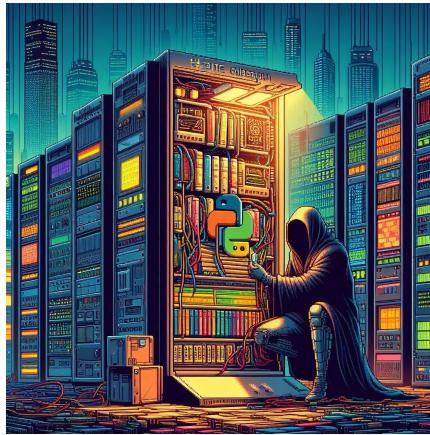




<https://medium.com/cloudera-inc/google-gemma-for-real-time-lightweight-open-llm-inference-88efe98e580f>







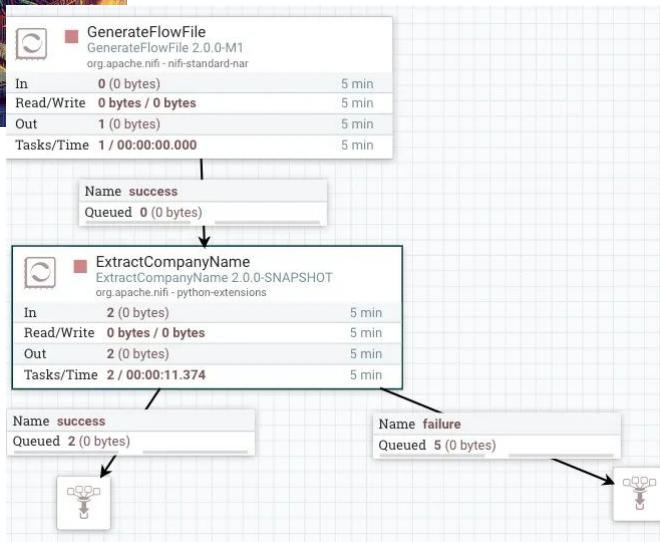
Python Processors





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

./

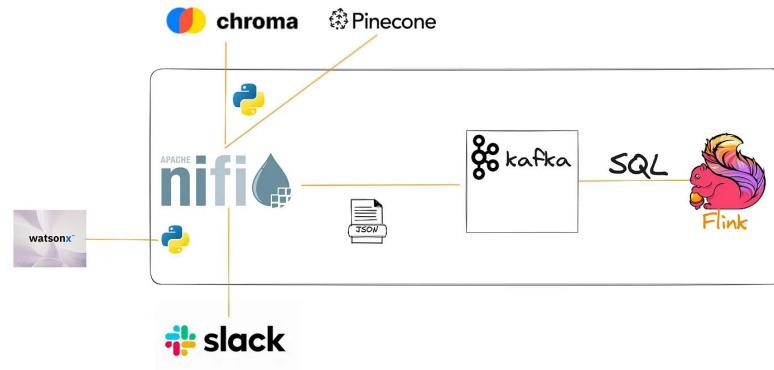
uuid

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



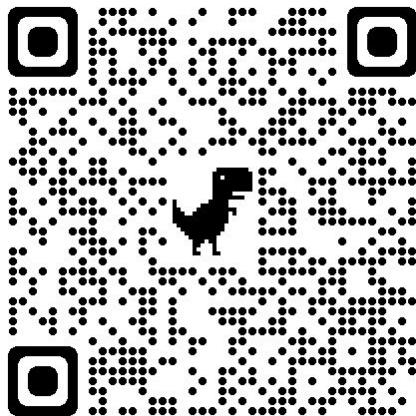
WatsonX SDK To Foundation

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



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

© 2024 Cloudera, Inc. All rights reserved.



CaptionImage

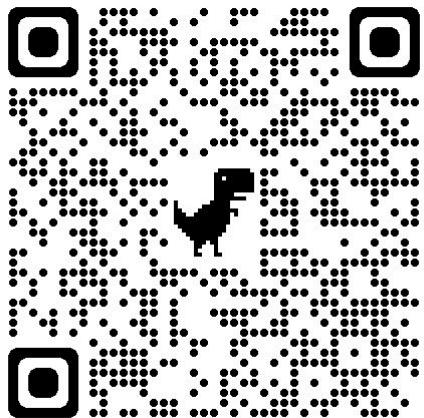
- Python 3.10+
- Hugging Face
- Salesforce/blip-image-captioning-large
- Generate Captions for Images
- Adds captions to FlowFile Attributes
- Does not require download or copies of your images

<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

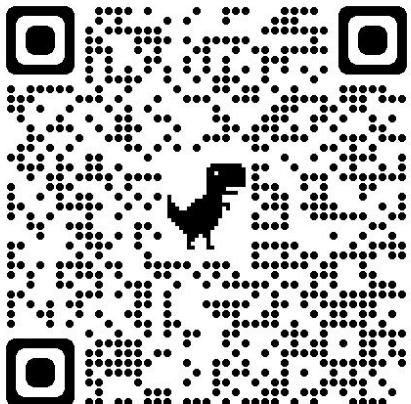


<https://github.com/tspannhw/FLaNK-python-processors>



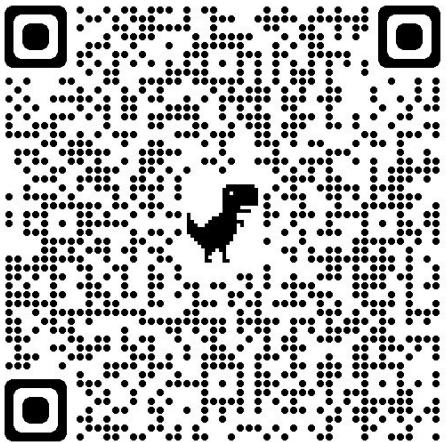
NSFW Image Detection

- 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





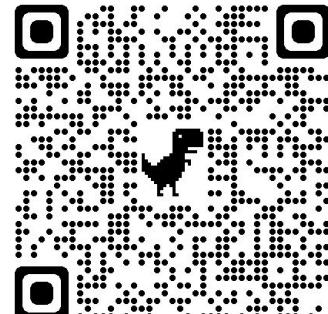
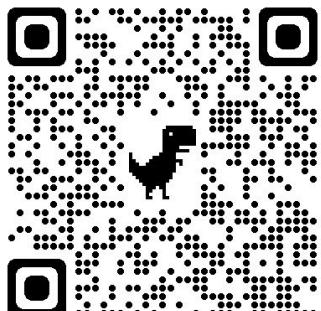
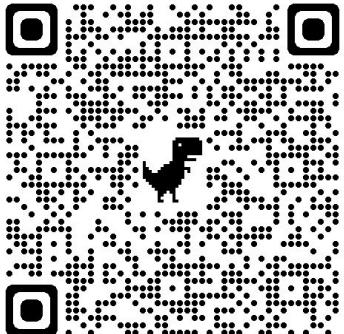
FacialEmotionsImageDetection



- 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

Other Python Processors

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



© 202

DEMO





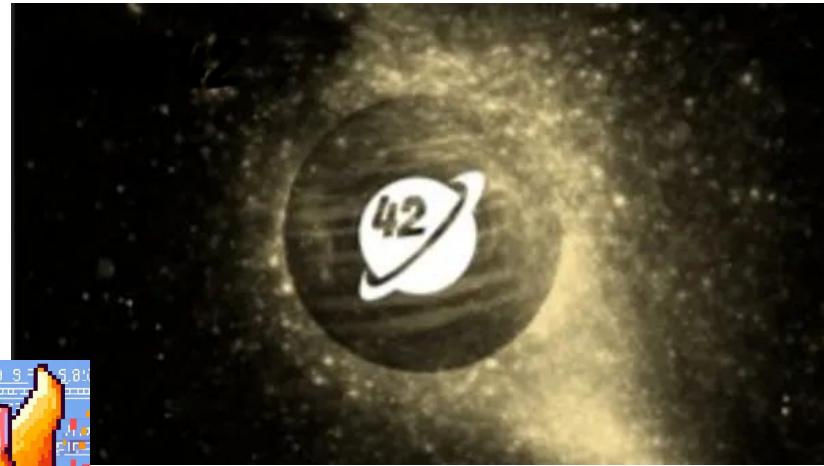
MAY 8–9
BOSTON, MA

star*tree | CLOUDERA

IN-PERSON MEETUP

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

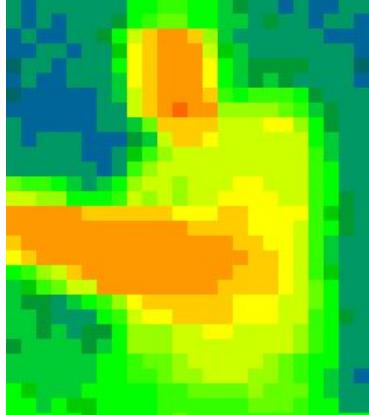
March 28 | 05:30 PM EST | NYC



Startup Grind
AI Max
Summit -
April 12 NJ



TH^NO^YU[★]



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

The screenshot shows the CSM interface with the following data:

- Producers:** 16 of 45
- Brokers:** 3 of 3
- Topics:** 4 of 45
- Consumer Groups:** 3 of 5

Under "Producers (16)", there are 16 listed, including gateway-west-raw-sensors, gateway-europe-raw-sensors, gateway-east-raw-sensors, and gateway-central-raw-sensors. Each producer has metrics like DATA-IN, DATA-OUT, MESSAGES-IN, and CONSUMER GROUPS.

Under "Consumer Groups (3)", there are three groups: rfid-track-sensors-west, rfid-track-sensors-east, and rfid-track-sensors-central, each with its own metrics.

At the bottom, it shows a replication factor of 1, 16 replicas, 1,372 total messages, and a retention period of 0 seconds.



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

The screenshot shows the CSA StreamBuilder interface with the following details:

- Explorer:** Shows a tree view of the project structure, including topics, tables, and external resources.
- Pipeline:** Shows the StreamBuilder code for creating a pipeline. It includes code for reading from various sources (Kafka, MySQL, JDBC, File, etc.) and writing to sinks (Kafka, MySQL, JDBC, File). It also includes logic for joins, windows, and state management.
- Logs:** Shows the command-line logs for the pipeline execution, including successful table creation and state registration.

ENTERPRISE MANAGEMENT CAPABILITIES FOR APACHE KAFKA

Extend streams messaging services for Schema Mgmt, Replication & Monitoring

Schema Registry

Kafka Schema Governance

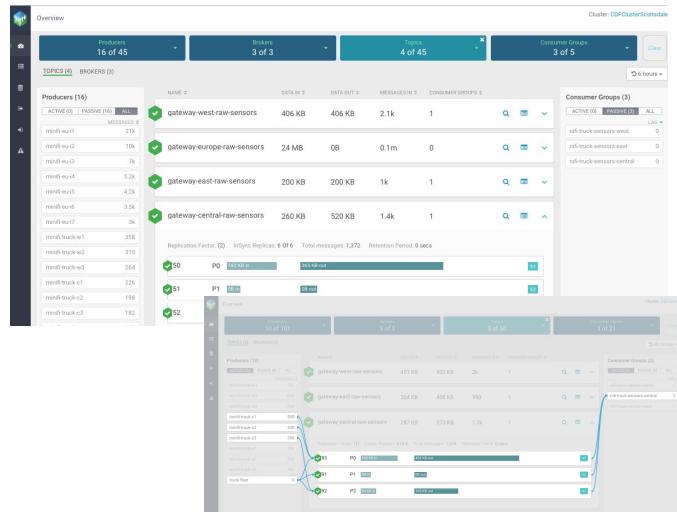
The screenshot shows the Schema Registry interface with the following details:

- Branch:** BACKWARD COMPATIBLE
- Type:** avro
- Group:** truck...
- Branch:** 1
- SERIALIZER & DESERIALIZER:** 0
- VERSION DESCRIPTION:** metadata branch for schema metadata syndicate-speed-event-avro
- VERSION DESCRIPTION:** Enriched Speed Events from trucks in Kafka Topic
- SCHEMA:**

```
1 { "type": "record", "name": "TruckEvent", "fields": [ { "name": "eventTime", "type": "string" }, { "name": "truckId", "type": "long", "default": 0 } ] }
```

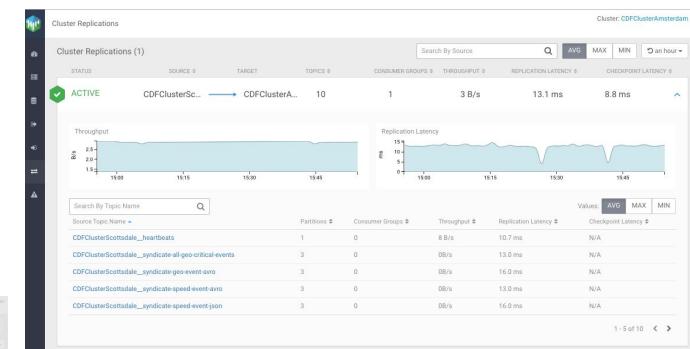
Streams Messaging Manager

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

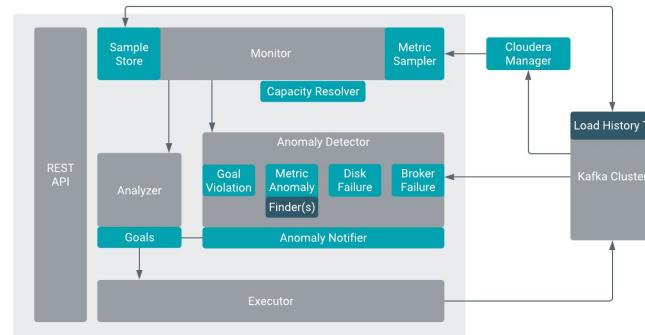
Kafka Connect Support

Simple Data Movement
Change Data Capture Connectors
Build Custom Connectors with NiFi

The screenshot shows the Cloudera Manager interface for Kafka Connect. It displays a summary of connectors: 16 total, 10 running, 6 failed, 0 degraded, and 0 paused. Below this, there are two tabs: 'Connectors' and 'Cluster Profile'. The 'Connectors' tab is active, showing a list of source and sink connectors. Under 'Source Connectors', there are 13 entries, mostly 'speed-event-jdbc-source' with various task counts (e.g., 13, 10, 13). Under 'Sink Connectors', there are 13 entries, mostly 'speed-event-S3-sink' with task counts (e.g., 13, 13, 13). The 'Cluster Profile' tab is visible at the bottom.

Cruise Control Support

Intelligent Rebalancing & Self-Healing of your Kafka Clusters



Ranger Security

Improved ACL and Audit for Kafka, KConnect and Schema Registry

The screenshot shows the Ranger interface for managing security policies. It includes sections for 'SCHEMA-REGISTRY', 'KAFKA', 'NIFI', and 'HDFS'. In the 'SCHEMA-REGISTRY' section, a policy for 'cm_schema_registry' is being edited, setting 'Policy Type' to 'PFI' and 'Schema Version' to 'V1'. Other tabs like 'Access Manager', 'Audit', and 'Security Zone' are also visible. The 'Audit' tab shows audit logs for the 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

The screenshot shows the Apache Flink UI interface. At the top, there's a search bar for 'Search in DB' and a 'Log' button. The main area is titled 'youthful_harvest' and shows the job status as 'FINISHED'. Below the status, it says 'clogat' as the 'Data Handler'. The UI displays various metrics and task details for the completed job.

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

The screenshot shows the Cloudera Manager interface with the following sections:

- Flink Application Overview:** Shows the application is **RUNNING** (ID: 72c0982cf7fbcb8027e2872926070558), started at 2019-09-17 09:53:51, and has been running for 1d 4h 52m. It includes tabs for Overview, Exceptions, Timeline, Checkpoints, and Configuration.
- Health Status:** Displays the status of various components:
 - CDH 6.3.0 (Parcels):** 8 Hosts (2 healthy, 3 warning, 3 critical).
 - Flink:** 1 healthy, 1 warning.
 - HDFS-2:** 2 healthy, 1 warning.
 - Kafka-2:** 1 healthy, 1 warning.
 - Schema Registry:** 1 healthy.
 - Streams Metrics:** 1 healthy.
 - Streams Replicator:** 1 healthy.
 - YARN (MR2 In):** 1 warning.
 - ZooKeeper-2:** 1 healthy.
- Charts:** Shows Cluster CPU usage (~100% for CDFClusterAmsterdam, Host CPU Usage) and Cluster Network IO (~3.8M/s for Kafka SpeedGeoStream, 1.9M/s for Kafka TruckGeoStream).
- Data Contexts:** A button to "Create" new data contexts.
- Flink Application Details:** Shows the Flink application's configuration and execution details, including a complex Stream Join using Interval Join with Timestamps/Watermarks and ThroughputInducedFunction + TimestampInducedFunction.
- Health Tests:** Summary of system health across History Server, Gateway, and Hosts.
- Health History:** No health changes in this time period.
- Important Events and Alerts:** No alerts or critical events.

SQL STREAM BUILDER (SSB)

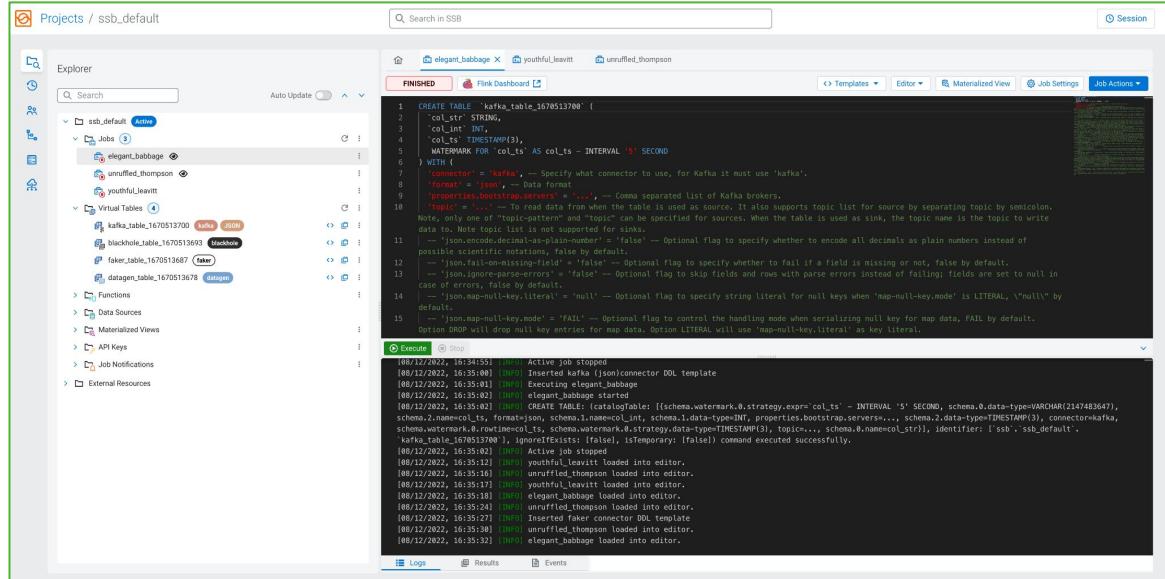
Democratize access to real-time data with just SQL

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)

