DATA SCIENCE SUMMIT

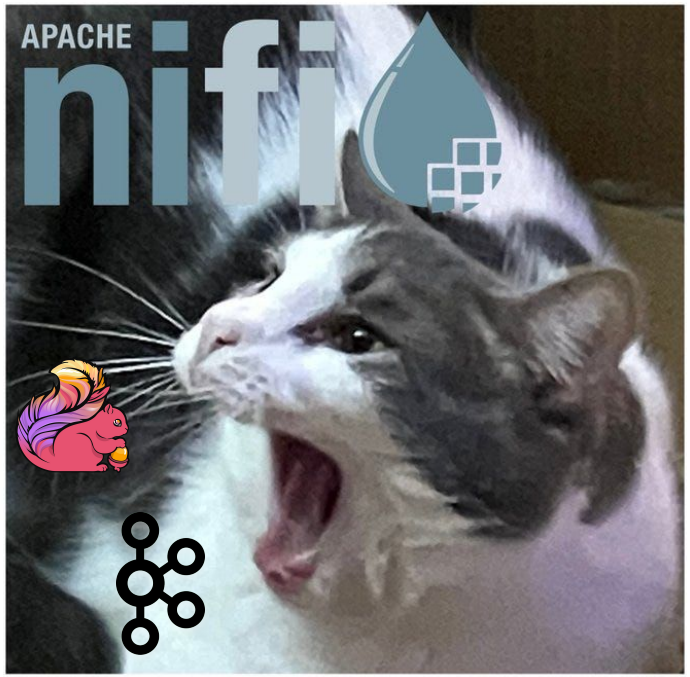
Integrating LLM with Streaming Data Pipelines

Tim Spann
Principal Dev, Cloudera

www.dssconf.pl | 23-24.11.2023 | PGE Narodowy + Online

ORGANIZERS: ACADEMIC PARTNERS Data Science Warsaw Wydział Matematyki i Nauk Informacyjnych POLITECHNIKA WARSZAWSKA

FLaNK Stack Weekly by Tim Spann

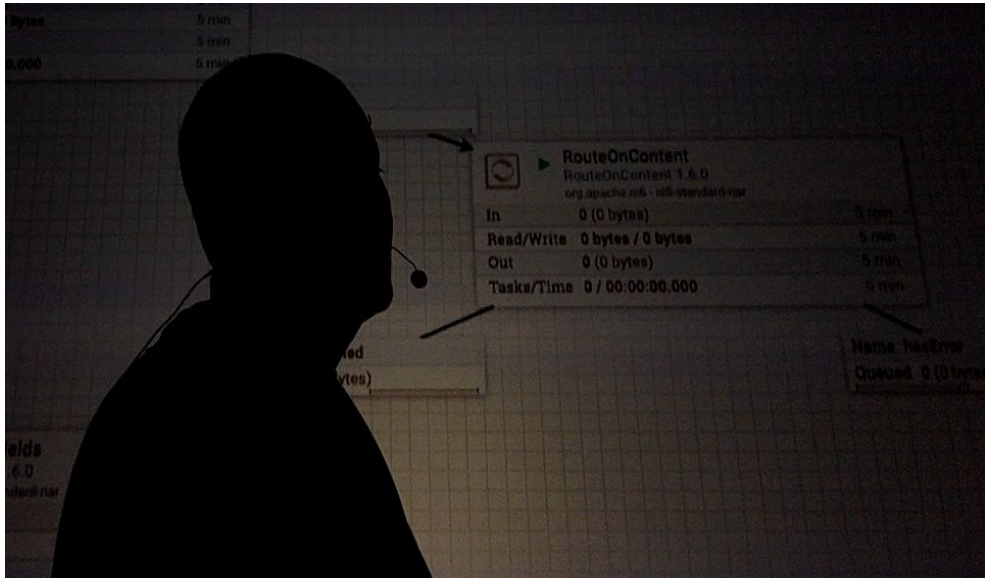


https://bit.ly/32dAJft

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.

Agenda (35 minutes)



Build a Streaming Data Pipeline

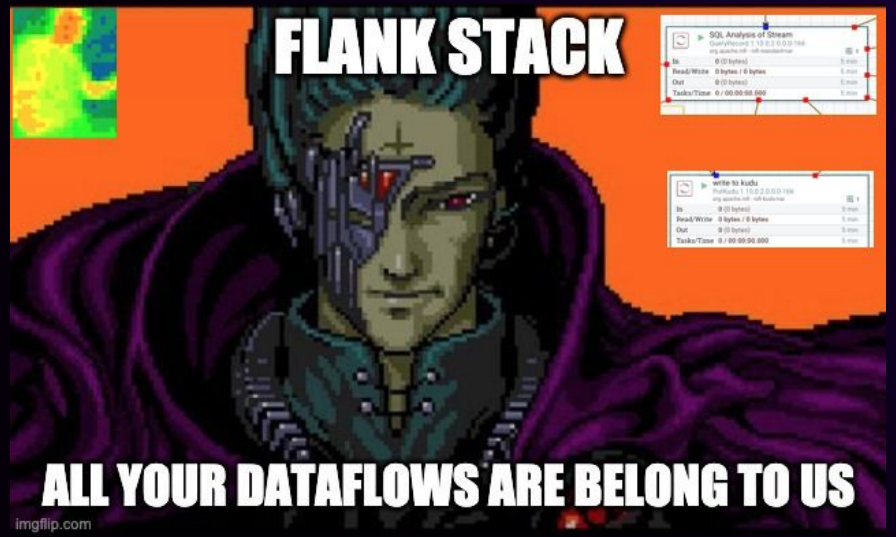
Integrate LLM via HTTPS / REST

Process, Route, Enrich, Transform Results

Store

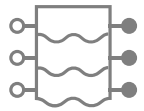
Distribute

Build a Streaming Data Pipeline

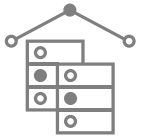


Data Relevance

Any DATA



Streaming data



Data at rest

Change data capture

Any BUSINESS EVENT

Real-time Processing

- Analyze data in motion
- Continuous monitoring
- Trends and anomalies

Any DATA ANALYST

Continuous Results

- No-Code UI
- Author once publish anywhere
- Analytics lifecycle management for dev/ops

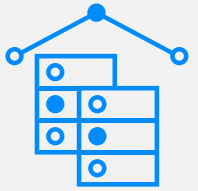
Any DATA CONSUMER

Event-driven apps

AI models

Analytics apps

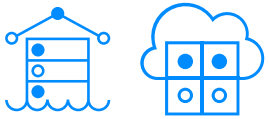
Data products



Data Lakehouse

LLM USE CASE

Unstructured file types



Structured Sources

Other enterprise data

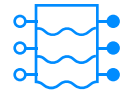
Data in Motion on Cloudera Data Platform (CDP)

Capture, process & distribute any data, anywhere

AI Model

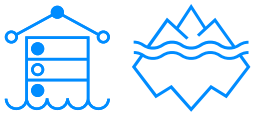
Vector DB

Applications/API's



Streams

Materialized Views



Open Data Lakehouse

Integrating LLM via HTTPS

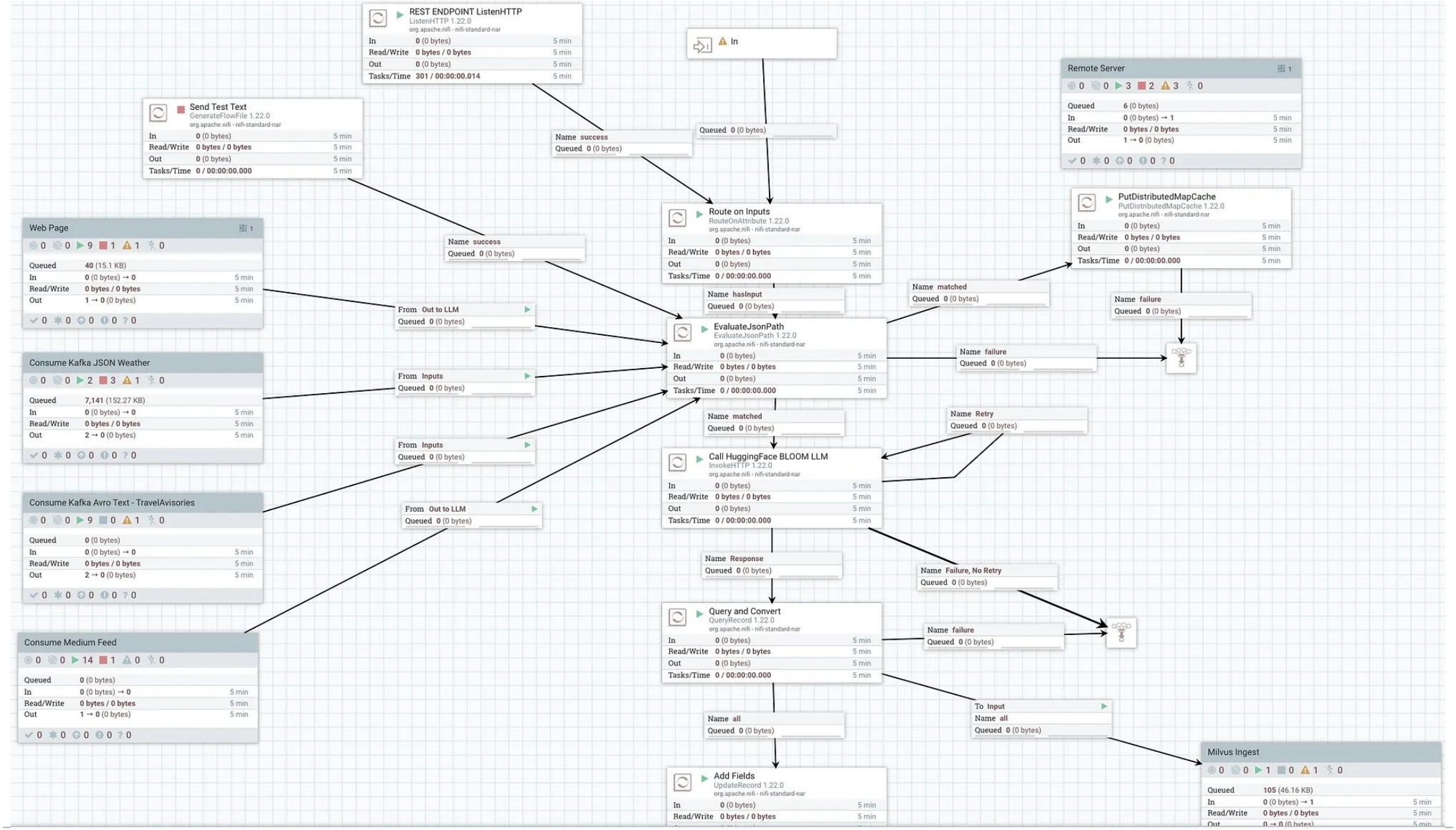
Models Used for Inference API

google/flan-ul2
google/flan-t5-xxl

bigscience/bloom

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

ibm/mpt-7b-instruct2
ibm/granite-13b-instruct-v1
ibm/granite-13b-chat-v1


REST ENDPOINT ListenHTTP
ListenHTTP 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 301 / 00:00:00.014 5 min
In
Send Test Text
GenerateFlowFile 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
Queued 0 (0 bytes)
Remote Server
Queued 6 (0 bytes)
In 0 (0 bytes) → 1 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 → 0 (0 bytes) 5 min
Route on Inputs
RouteOnAttribute 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
PutDistributedMapCache
PutDistributedMapCache 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Web Page
Queued 40 (15.1 KB)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 → 0 (0 bytes) 5 min
Name success
Queued 0 (0 bytes)
Name matched
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
From Out to LLM
Queued 0 (0 bytes)
Name hasInput
Queued 0 (0 bytes)
EvaluateJsonPath
EvaluateJsonPath 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Consume Kafka JSON Weather
Queued 7,141 (152.27 KB)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 2 → 0 (0 bytes) 5 min
From Inputs
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
Name matched
Queued 0 (0 bytes)
Name Retry
Queued 0 (0 bytes)
From Inputs
Queued 0 (0 bytes)
Call HuggingFace BLOOM LLM
InvokeHTTP 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Consume Kafka Avro Text - TravelAvisories
Queued 0 (0 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 2 → 0 (0 bytes) 5 min
From Out to LLM
Queued 0 (0 bytes)
Name Response
Queued 0 (0 bytes)
Name Failure, No Retry
Queued 0 (0 bytes)
Query and Convert
QueryRecord 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name failure
Queued 0 (0 bytes)
Consume Medium Feed
Queued 0 (0 bytes)
In 0 (0 bytes) → 0 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 → 0 (0 bytes) 5 min
To Input
Name all
Queued 0 (0 bytes)
Name all
Queued 0 (0 bytes)
Milvus Ingest
Queued 105 (46.16 KB)
In 0 (0 bytes) → 1 5 min
Read/Write 0 bytes / 0 bytes 5 min
Add Fields
UpdateRecord 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min

HTTPS Calls

For Hugging Face:
https://api-inference.huggingface.co/models/bigscience/bloom

Pass in a simple JSON Prompt

{"inputs":"This is Prompt Text"}

Requires an HTTP Header with Authorization for Bearer YOURTOKEN

YOURTOKEN from https://huggingface.co/settings/tokens

HTTPS Calls

For IBM WatsonX.AI:

https://iam.cloud.ibm.com/identity/token <- First call to get token by sending an API key and grant.

Then send a JSON prompt call model server via https://us-south.ml.cloud.ibm.com/ml/v1-beta/generation/text?version=2023-05-29

Requires an HTTP Header with Authorization for Bearer ${readtoken}

WatsonX.AI JSON Prompt

```
{ "model_id": "meta-llama/llama-2-70b-chat",
  "input": "this is prompt text”,
    "parameters": {
      "decoding_method": "greedy",
      "max_new_tokens": 200,
      "min_new_tokens": 50,
      "stop_sequences": [],
      "repetition_penalty": 1
      },
   "project_id": "0ead8ec4-d137-4f9c-8956-50b0da4a7068" }
```

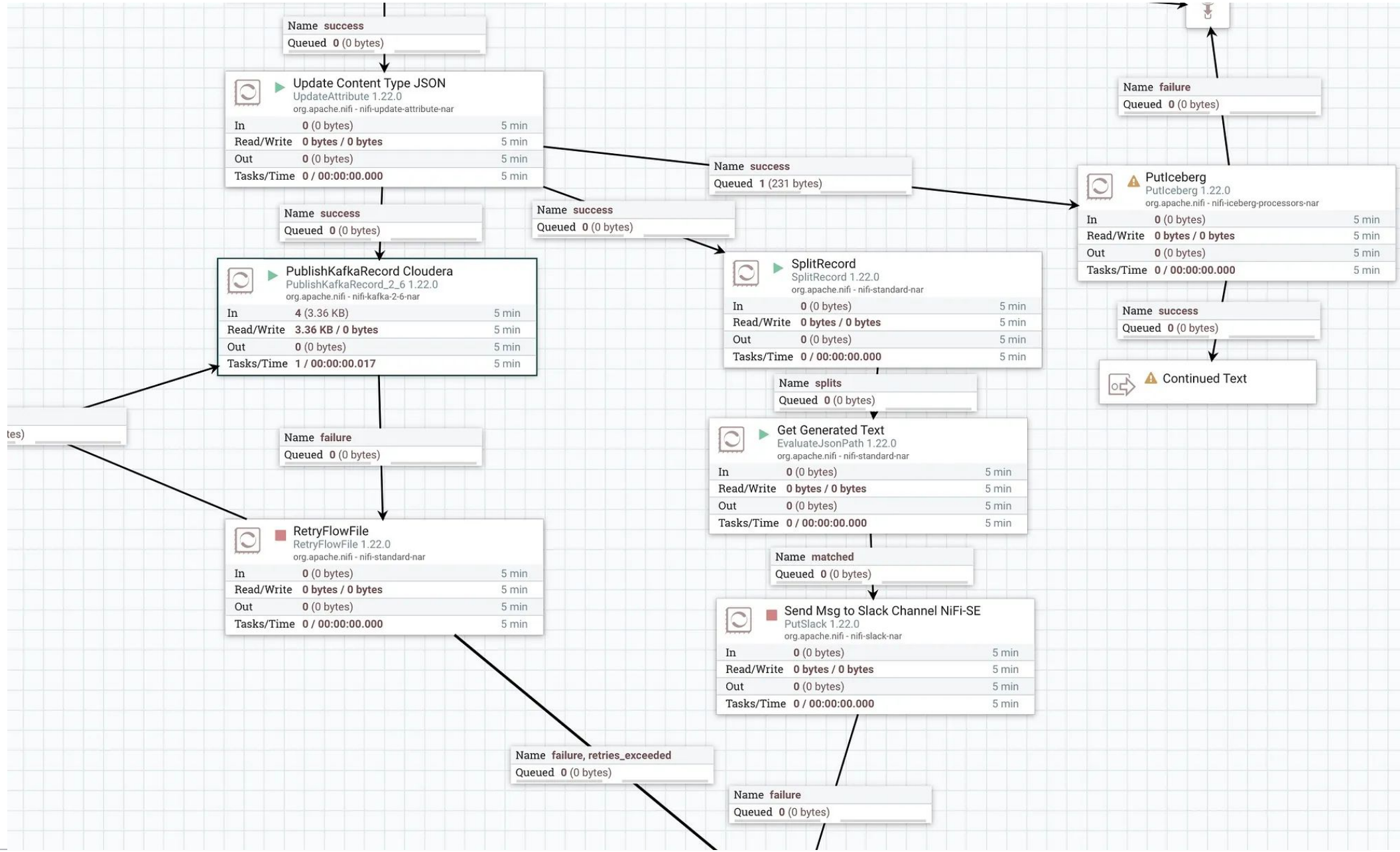
Property	Value
HTTP Method	**GET**
HTTP URL	**https://rss.nytimes.com/services/xml/rss/nyt/US.xml**
HTTP/2 Disabled	**False**
SSL Context Service	No value set
Socket Connect Timeout	**5 secs**
Socket Read Timeout	**15 secs**
Socket Idle Timeout	**5 mins**
Socket Idle Connections	**5**
Proxy Configuration Service	No value set
Proxy Host	No value set
Request OAuth2 Access Token Provider	No value set
Request Username	No value set

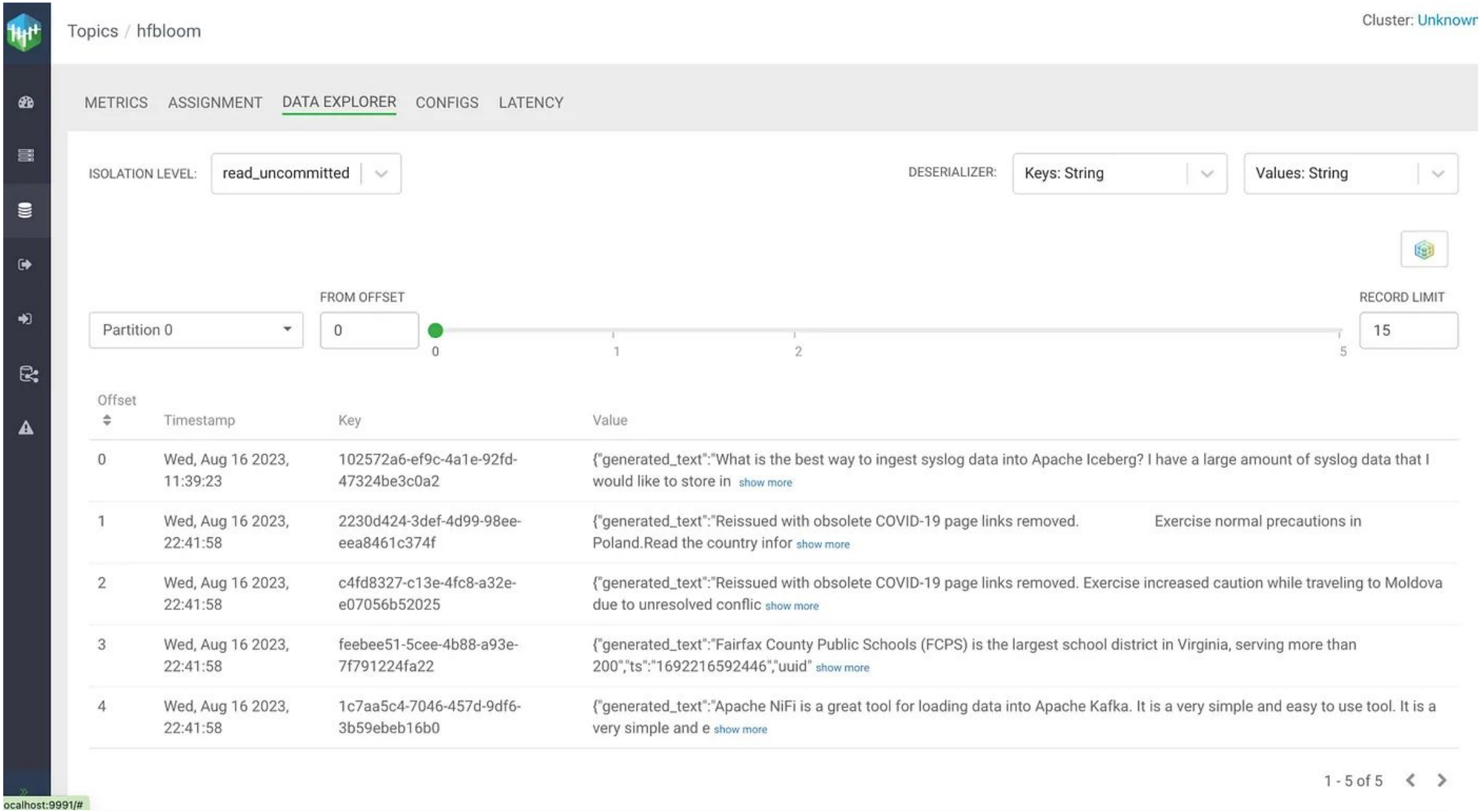
Process, Route, Enrich, Transform Results

Property	Value
Record Reader	**CDP Infer JsonTreeReader**
Record Writer	**BasicJsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**false**
Default Decimal Precision	**10**
Default Decimal Scale	**4**
all	SELECT * FROM FLOWFILE
negative	SELECT lower(label) as sentiment, label, score...
positive	SELECT lower(label) as sentiment, label, score...

Property	Value
Record Reader	**CDP Infer JsonTreeReader**
Record Writer	**JsonRecordSetWriter**
Replacement Value Strategy	**Literal Value**
/textbody	${inputs}
/ts	${now():toNumber()}
/uuid	${uuid}

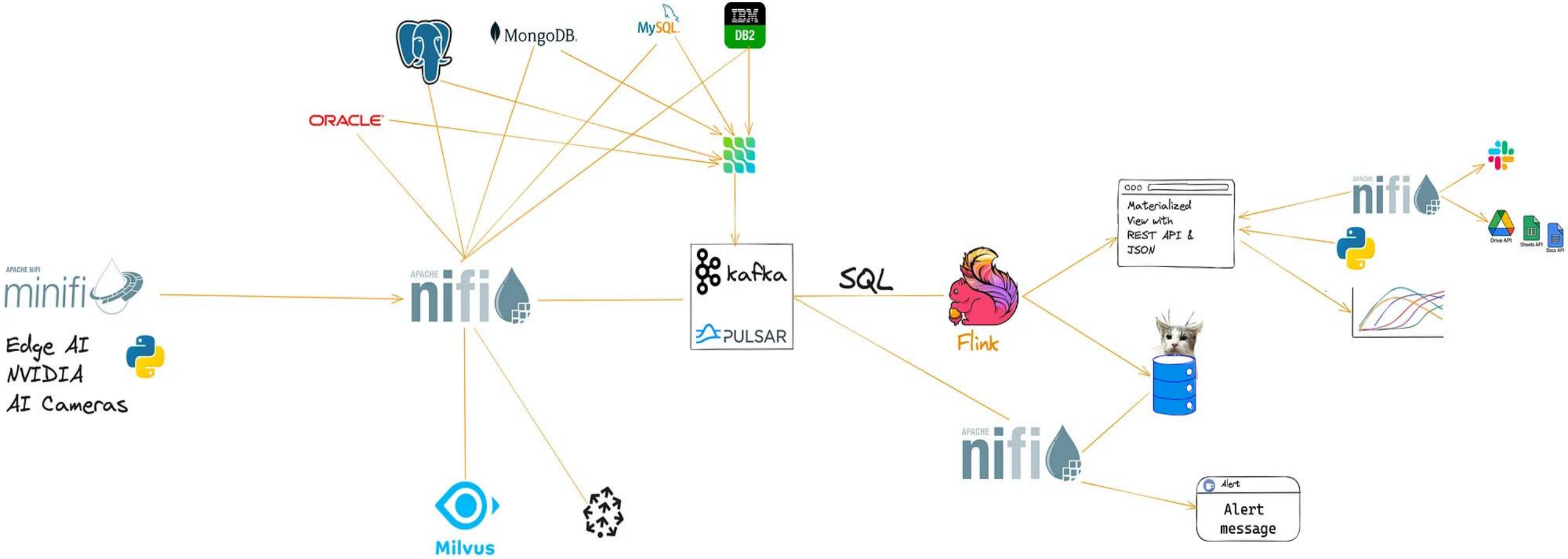
Store


Name success
Queued 0 (0 bytes)
Update Content Type JSON
UpdateAttribute 1.22.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
Name success
Queued 1 (231 bytes)
Name success
Queued 0 (0 bytes)
PublishKafkaRecord Cloudera
PublishKafkaRecord_2_6 1.22.0
org.apache.nifi - nifi-kafka-2-6-nar
In 4 (3.36 KB) 5 min
Read/Write 3.36 KB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:00.017 5 min
SplitRecord
SplitRecord 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name failure
Queued 0 (0 bytes)
PutIceberg
PutIceberg 1.22.0
org.apache.nifi - nifi-iceberg-processors-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name failure
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
Continued Text
Name splits
Queued 0 (0 bytes)
Get Generated Text
EvaluateJsonPath 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
RetryFlowFile
RetryFlowFile 1.22.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name matched
Queued 0 (0 bytes)
Send Msg to Slack Channel NiFi-SE
PutSlack 1.22.0
org.apache.nifi - nifi-slack-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name failure, retries_exceeded
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)


Topics / hfbloom
Cluster: Unknowr
METRICS
ASSIGNMENT
DATA EXPLORER
CONFIGS
LATENCY
ISOLATION LEVEL:
read_uncommitted
DESERIALIZER:
Keys: String
Values: String
FROM OFFSET
RECORD LIMIT
Partition 0
0
15
0
1
2
5
Offset
Timestamp
Key
Value
0
Wed, Aug 16 2023, 11:39:23
102572a6-ef9c-4a1e-92fd-47324be3c0a2
{"generated_text":"What is the best way to ingest syslog data into Apache Iceberg? I have a large amount of syslog data that I would like to store in show more
1
Wed, Aug 16 2023, 22:41:58
2230d424-3def-4d99-98ee-eea8461c374f
{"generated_text":"Reissued with obsolete COVID-19 page links removed. Exercise normal precautions in Poland.Read the country infor show more
2
Wed, Aug 16 2023, 22:41:58
c4fd8327-c13e-4fc8-a32e-e07056b52025
{"generated_text":"Reissued with obsolete COVID-19 page links removed. Exercise increased caution while traveling to Moldova due to unresolved conflic show more
3
Wed, Aug 16 2023, 22:41:58
feebee51-5cee-4b88-a93e-7f791224fa22
{"generated_text":"Fairfax County Public Schools (FCPS) is the largest school district in Virginia, serving more than 200","ts":"1692216592446","uuid" show more
4
Wed, Aug 16 2023, 22:41:58
1c7aa5c4-7046-457d-9df6-3b59ebeb16b0
{"generated_text":"Apache NiFi is a great tool for loading data into Apache Kafka. It is a very simple and easy to use tool. It is a very simple and e show more
1 - 5 of 5
ocalhost:9991/#

Distribute


CUNK ON FLANK
THEY SAY IT'S FOR FAST DATA
IS IT SO FAST THAT IT ARRIVES BEFORE I ASKED FOR IT
imgflip.com


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
Drive API
Sheets API
Docs API
APACHE
nifi
Milvus
Alert
Alert
message

general Company-wide announcements and work-based matters

+ Add a bookmark

Today

Compute Time: 905
Compute Type: gpu+optimized
Inference Time: 905
Queue Time: 0
Request ID: VjKhYJvX664mtgaljgymc
SHA: 053d9cd9fbe814e091294f67fcfedb3397b954bb
Time Per Token: 45
Total Time: 905
Validation Time: 0
R: Fairfax County Public Schools (FCPS) is the largest school district in Virginia, serving more than 200
=====

==== NiFi to Hugging Face BLOOM LLM
AzureML Model Deployment: bloom-deployment
On Date: Thu, 17 Aug 2023 02:05:56 GMT
File Name: 7b8afb65-d5eb-49ee-9a89-d2b870310243
Request Duration: 1339
Request URL: https://api-inference.huggingface.co/models/bigscience/bloom
Compute Characters: 63
Compute Time: 1014
Compute Type: gpu+optimized
Inference Time: 968
Queue Time: 45
Request ID: iPX8lqeK9XwqSZA9e-wCH
SHA: 053d9cd9fbe814e091294f67fcfedb3397b954bb
Time Per Token: 48
Total Time: 1014
Validation Time: 0
R: Apache NiFi is a great tool for loading data into Apache Kafka. It is a very simple and easy to use tool. It is a very simple and easy to use
=====

Resources

DATA
SCIENCE
SUMMIT

Thank you for watching!

Remember to leave your questions and rate the presentation in the section below.

www.dssconf.pl

23-24.11.2023

PGE Narodowy + Online

ORGANIZERS:

ACADEMIC PARTNERS

Data Science Warsaw

Wydział Matematyki i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA