

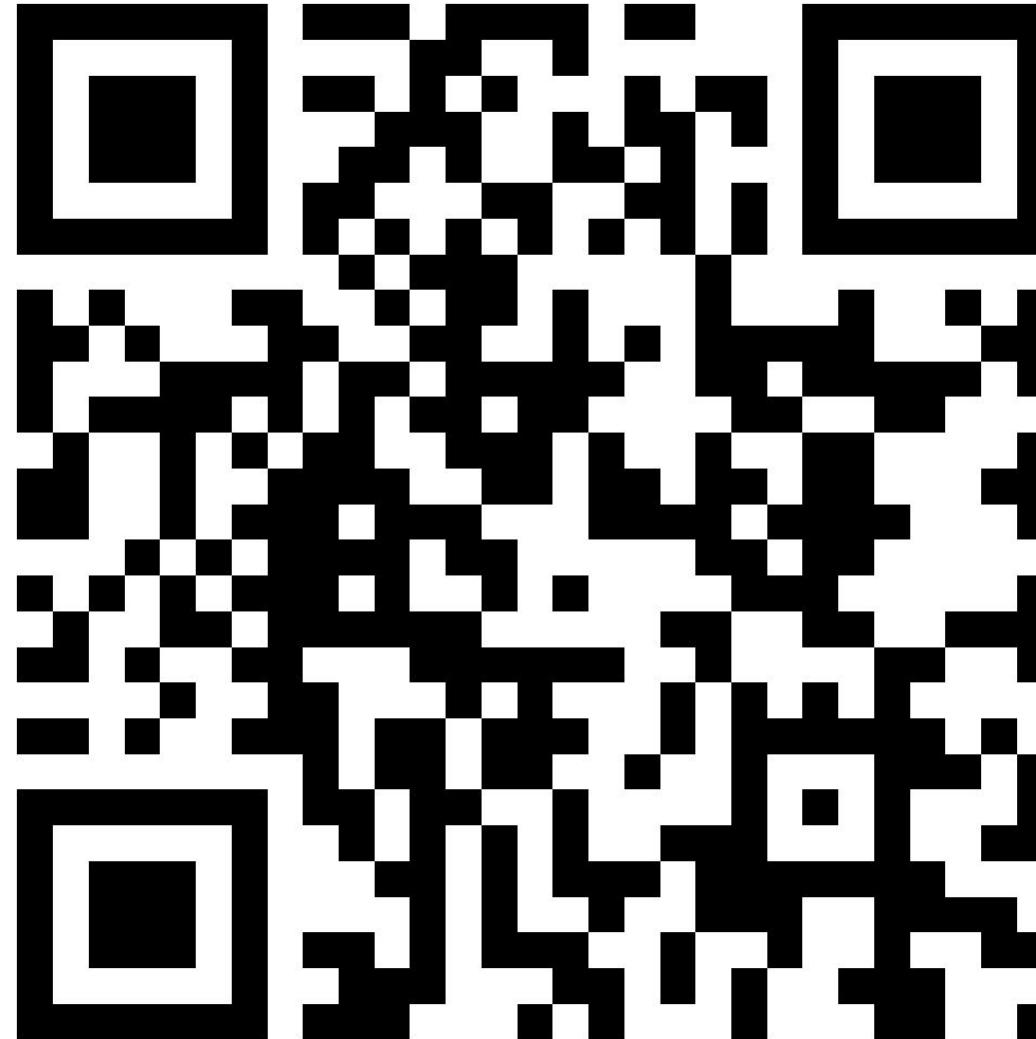
CLOUDERA

Looking at the New Features of Apache NiFi

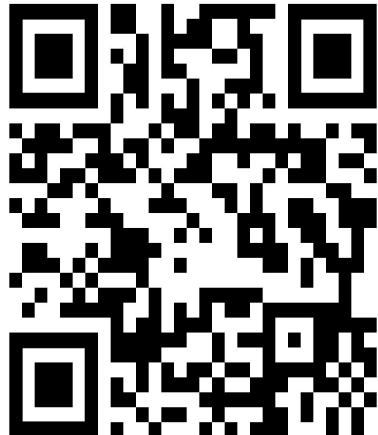
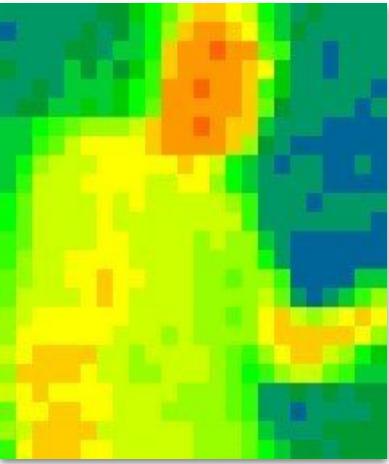
Timothy Spann
Principal Developer Advocate

COMMUNITY
THE ASF CONFERENCE
CODE

Slides, Code, Articles and More...



FLaNK Stack



Tim Spann

@PaasDev // Blog: www.datainmotion.dev

Principal Developer Advocate.

Princeton Future of Data Meetup.

ex-Pivotal, ex-Hortonworks, ex-StreamNative, ex-PwC

<https://medium.com/@tspann>

<https://github.com/tspannhw>

Apache NiFi x Apache Kafka x Apache Flink



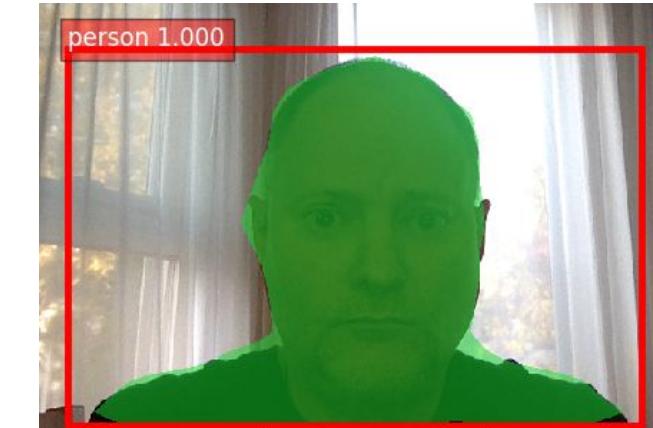
Top IoT Experts



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...



Future of Data - New York + Princeton + Virtual



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

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

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 ...



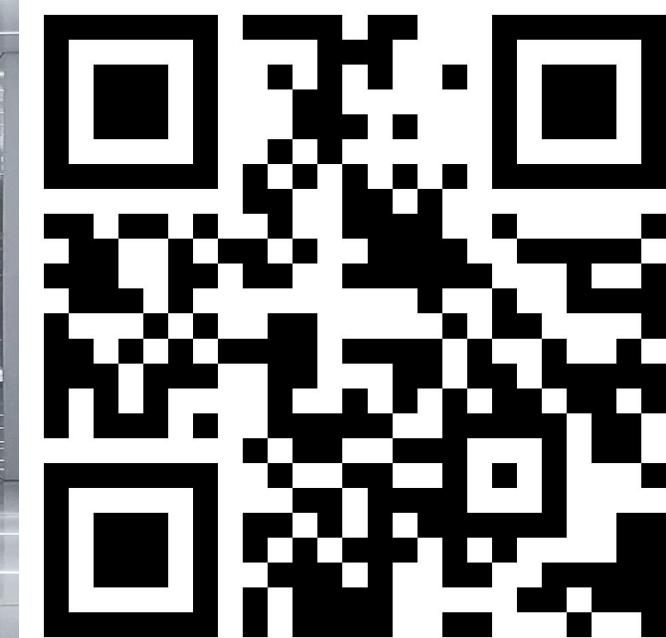
AN OPEN SOURCE COMMUNITY

@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.

My Talk List

Utilizing Real-Time Transit Data for Travel Optimization

Let's Monitor the Conditions at the Conference



Agenda

Apache NiFi has a lot of new features, processors and best practices that have arrived in the last year or so.

I will walk through building flows using the latest tips, techniques and processor.

I will and change a number of data flows utilizing the latest NiFi version and point out gotchas and some never dos. The deck will act as a take-away with notes, tips and guides to what we covered.

==> Any NiFi 1.23+ and 2.0 in progress features people want to see?



0 51,434 / 61.69 MB 0 0 193 914 555 166 0 0 0 0 0 21:02:42 EDT



Records

New ExcelRecord Reader

AmazonGlueSchemaRegistry

<https://issues.apache.org/jira/secure/ReleaseNote.jspa?projectId=12316020&version=12353320>



New to 2023 Processors

GenerateRecord

GetAsanaObject

PutSalesforceObject

QuerySalesforceObject

PutIoTDBRecord

QueryIoTDBRecord

ListGoogleDrive

FetchGoogleDrive

PutGoogleDrive

PutBoxFile

ListBoxFile

FetchBoxFile

PutDropbox

DecryptContent

DecryptContentCompatibility

<https://issues.apache.org/jira/secure/ReleaseNote.jspa?projectId=12316020&version=12353320>



0 51,434 / 61.69 MB 0 0 193 914 555 166 0 0 0 0 0 21:02:42 EDT



New to 2023 Processors

ExtractRecordSchema

RemoveRecordField

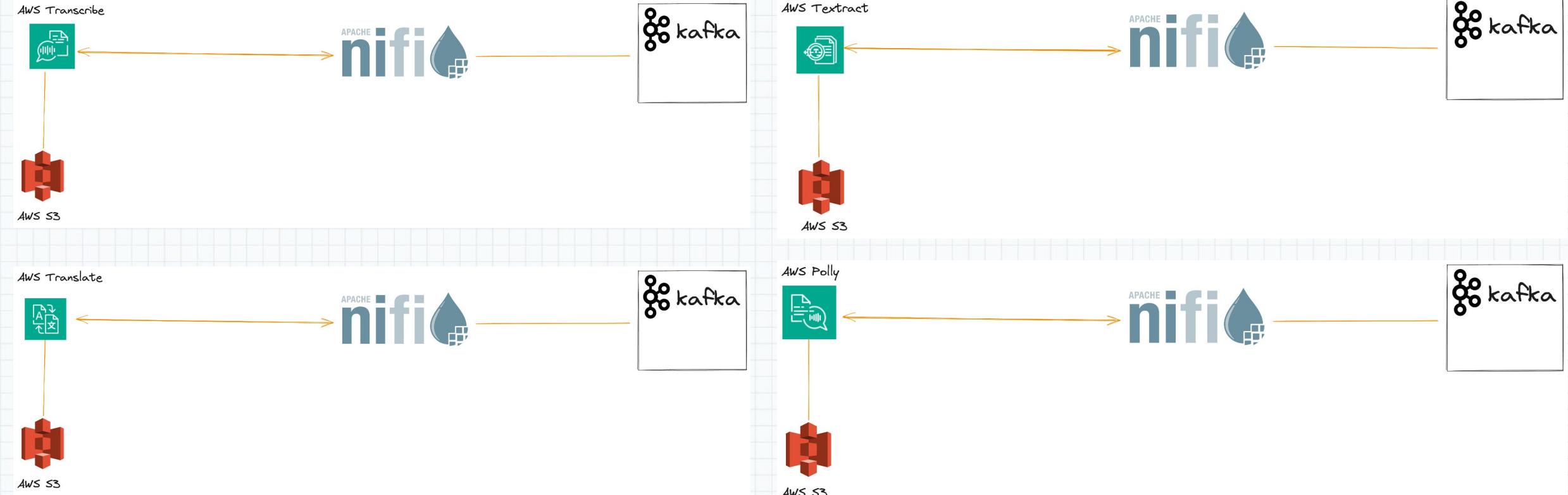
VerifyContentMAC

TriggerHiveMetaStoreEvent

“count” function added to RecordPath

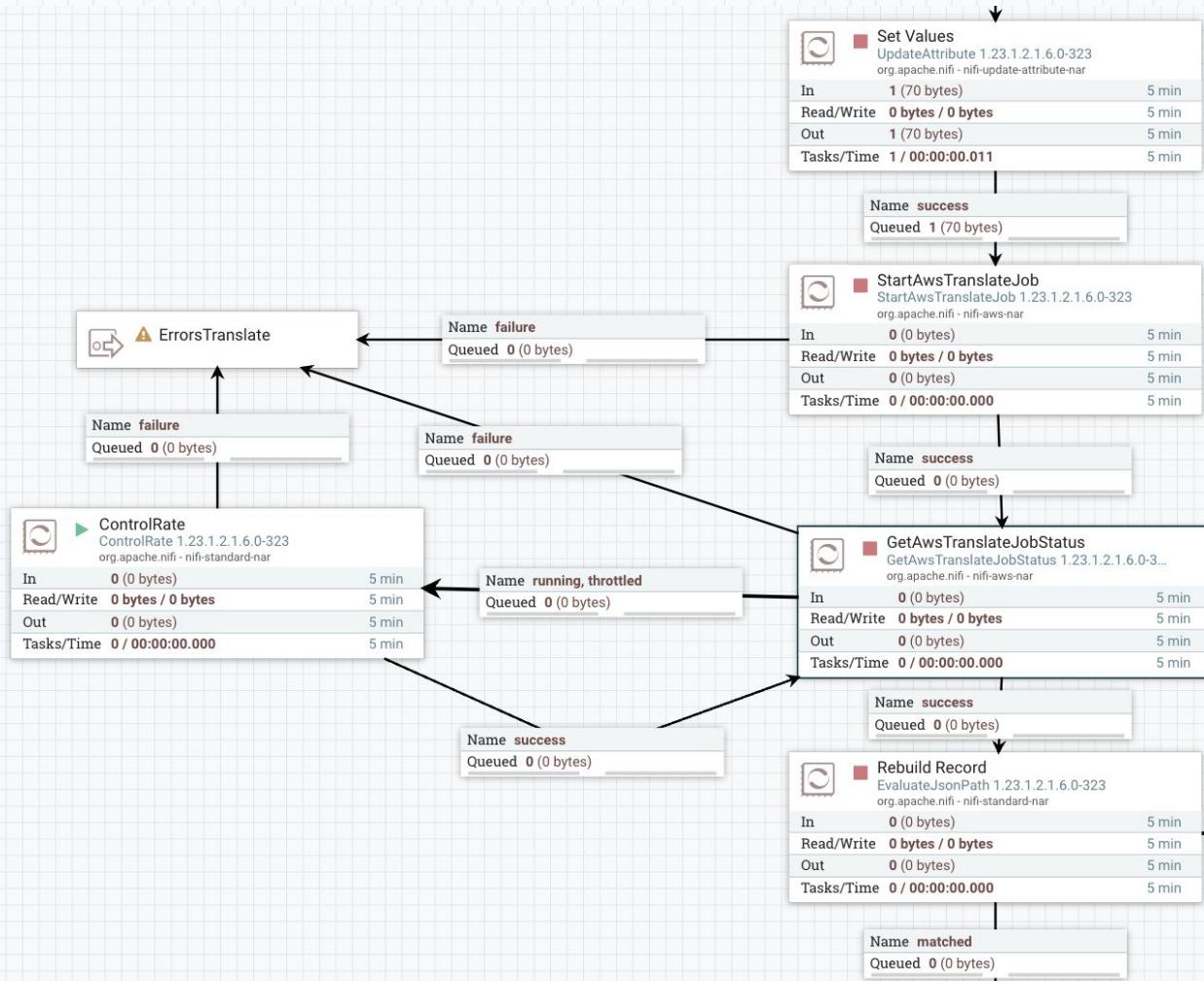


AWS ML Service Processors



<https://github.com/tspannhw/FLaNK-AWSML>

AWS Translate



Deprecating for Removal

Deprecate Lua and Ruby Script Engines

Deprecate ECMAScript Script Engine

Deprecate the Ambari Reporting Task

Deprecate Kafka 1.x components and 2.0 components

XML Templates

Variables

See:

<https://cwiki.apache.org/confluence/display/NIFI/Deprecated+Components+and+Features>

0 51,434 / 61.69 MB 0 0 193 914 555 166 0 0 0 0 0 21:02:42 EDT



Start Using

ExecuteStateless -> run your stateless flows right in a regular NiFi cluster

Parameters

JSON Flow Serialization

Records everywhere







NiFi 2.0 Coming

- Python Integration
- Parameters
- JDK 17, maybe 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>

Thanks to Pierre!



Pierre Villard

Apache NiFi Committer & PMC member |
Working @Cloudera - ex-@Google | Twitter &
Github — @pvillard31 | Blog @
www.pierrevillard.com

Python as First Class (NIFI-11241)

Graphical UI with custom Python based extensions

```
import cv2
import numpy as np
import json
from nifiapi.properties import PropertyDescriptor
from nifiapi.properties import ResourceDefinition
from nifiapi.flowfiletransform import FlowFileTransformResult

SCALE_FACTOR = 0.00392
NMS_THRESHOLD = 0.4      # non-maximum suppression threshold
CONFIDENCE_THRESHOLD = 0.5

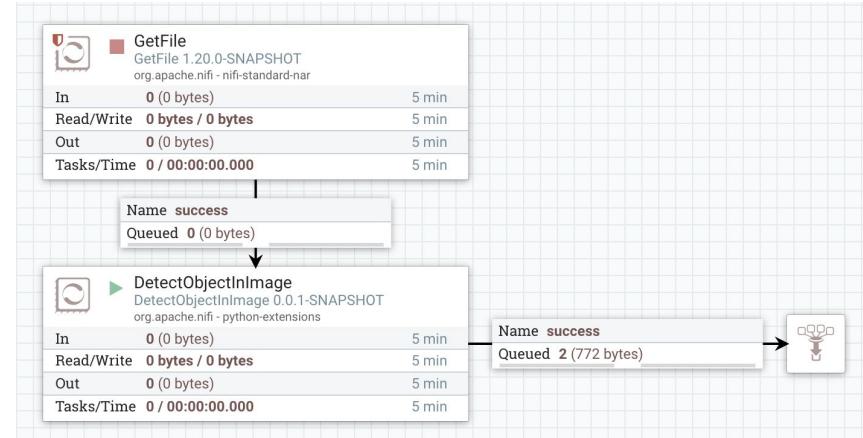
class DetectObjectInImage:
    class Java:
        implements = ['org.apache.nifi.python.processor.FlowFileTransform']
        ProcessorDetails:
            version = '0.0.1-SNAPSHOT'
            dependencies = ['numpy >= 1.23.5', 'opencv-python >= 4.6']

    def __init__(self, jvm=None, **kwargs):
        self.jvm = jvm

        # Build Property Descriptors
        self.model_file = PropertyDescriptor(
            name = 'Model File',
            description = 'The binary file containing the trained Deep Neural Network weights. Supports Caffe (*.caffemodel), TensorFlow (*.pb), Torch (*.t7, *.net), Darknet (*.weights), ' +
                      'DLDT (*.bin), and ONNX (*.onnx)',
            required = True,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.config_file = PropertyDescriptor(
            name = 'Network Config File',
            description = 'The text file containing the Network configuration. Supports Caffe (*.prototxt), TensorFlow (*.pbtxt), Darknet (*.cfg), and DLDT (*.xml)',
            required = False,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.class_name_file = PropertyDescriptor(
            name = 'Class Names File',
            description = 'A text file containing the names of the classes that may be detected by the model. Expected format is one class name per line, new-line terminated.',
            required = True,
            resource_definition = ResourceDefinition(allow_file = True)
        )
        self.descriptors = [self.model_file, self.config_file, self.class_name_file]

    def getPropertyDescriptors(self):
        return self.descriptors

    def onScheduled(self, context):
        # read class names from text file
        class_name_file = context.getProperty(self.class_name_file.name).getValue()
        if class_name_file is None:
```



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)

MiNiFi C++

(small footprint)

MiNiFi Java

(headless version of NiFi)

NiFi Registry

Stateless NiFi

NiFi Deploy Options from Open Source to Managed

Cloudera Edge Flow Manager

(Command & Control of MiNiFi Agents)

Cloudera NiFi for Kafka

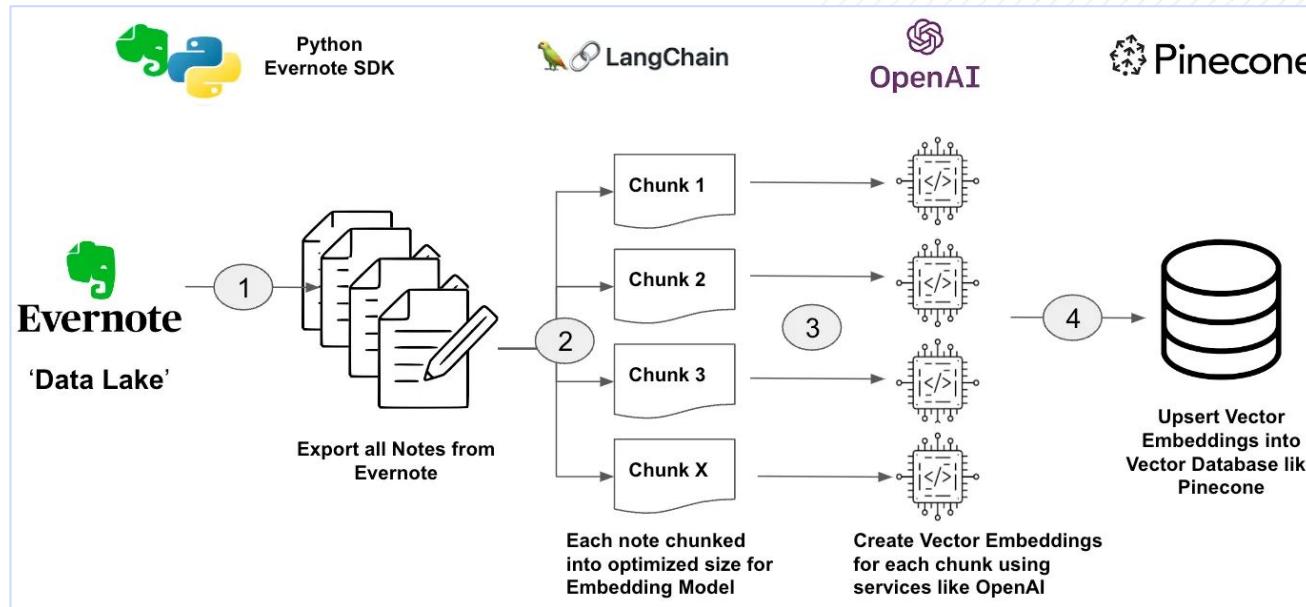
Connect

**NiFi in
Cloudera DataFlow Functions**

Cloudera DataFlow

NiFi 2.0 is coming... <https://medium.com/cloudera-inc/getting-ready-for-apache-nifi-2-0-5a5e6a67f450>

- First-class citizen Python API
- Rules Engine
- NiFi Stateless at Process Group level
- Java 21 (virtual threads, perf improvements, etc)



Closing the gap between data engineers and data scientists...

- Export documentation (Sharepoint, OCR) to build the knowledge base powering your chatbot
- Scrape the internet (Sitemap) to build the knowledge base powering your chatbot
- Real-time streaming ingest of Slack to build the knowledge base powering your chatbot

Evernote AI Chatbot Powered by Apache NiFi using OpenAI, Pinecone & Langchain

When did my right elbow pain occur?

The right elbow pain occurred on January 11, 2023.

Evernote Source: Note: Right Elbow Pain - 01-11-23 (Notebook: Family_Aju_Health),

What doctor did I see for it?

You saw Dr. Patel for your right elbow pain.

Evernote Source: Note: Appointment with Dr. Patel on 6_20 for Right Elbow Pain (Notebook: Family_Aju_Health),

What was the cause for the right elbow pain and what did the doctor suggest?

The cause for the right elbow pain is tennis elbow. The doctor suggested a cortisone shot in the right elbow and taking Diclofenac for prescription pain relief.

Evernote Source: Note: Appointment with Dr. Patel on 6_20 for Right Elbow Pain (Notebook: Family_Aju_Health), Note: Right Elbow Pain - 01-11-23 (Notebook: Family_Aju_Health),

What was the dosage for Diclofenac?

The dosage for Diclofenac is 50 mg, 3 times a week.

Evernote Source: Note: Medications & Prescriptions That I Take (Notebook: Family_Aju_Health),

You:

<https://medium.com/@george.vetticaden/accelerating-ai-data-pipelines-building-an-evernote-chatbot-with-apache-nifi-2-0-and-generative-ai-9d977466ff4c>

DEMO



