CLOUDERA

Looking at the New Features of Apache NiFi

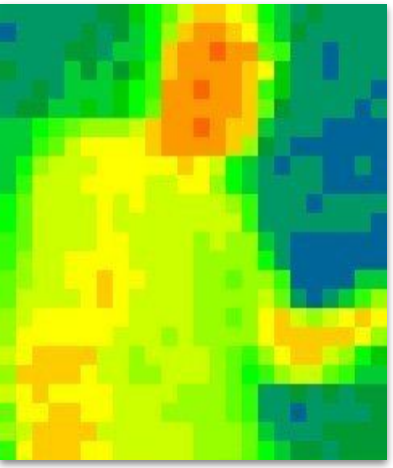
Timothy Spann

Principal Developer Advocate

Sunday October 8, 2023
4:10PM - 4:50 PM
Room 102

Slides, Code, Articles and More…

FLaNK Stack



Tim Spann

@PaasDev // Blog: **www.datainmotion.dev**

Principal Developer Advocate.

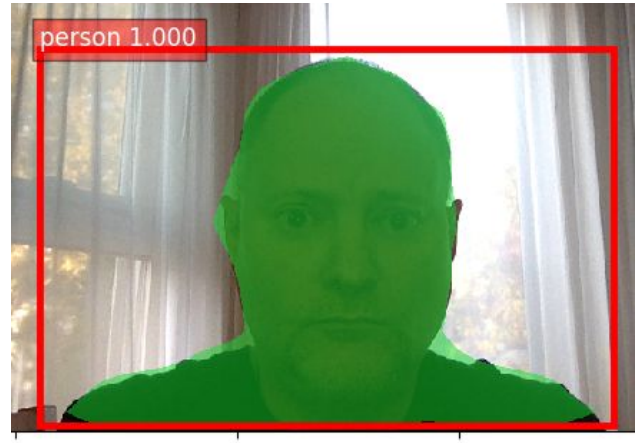
Princeton Future of Data Meetup.

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

https://medium.com/@tspann

https://github.com/tspannhw

Apache NiFi x Apache Kafka x Apache Flink



Future of Data - New York + Princeton + Virtual

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

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 ...

@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?

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

New to 2023 Processors

ExtractRecordSchema

RemoveRecordField

VerifyContentMAC

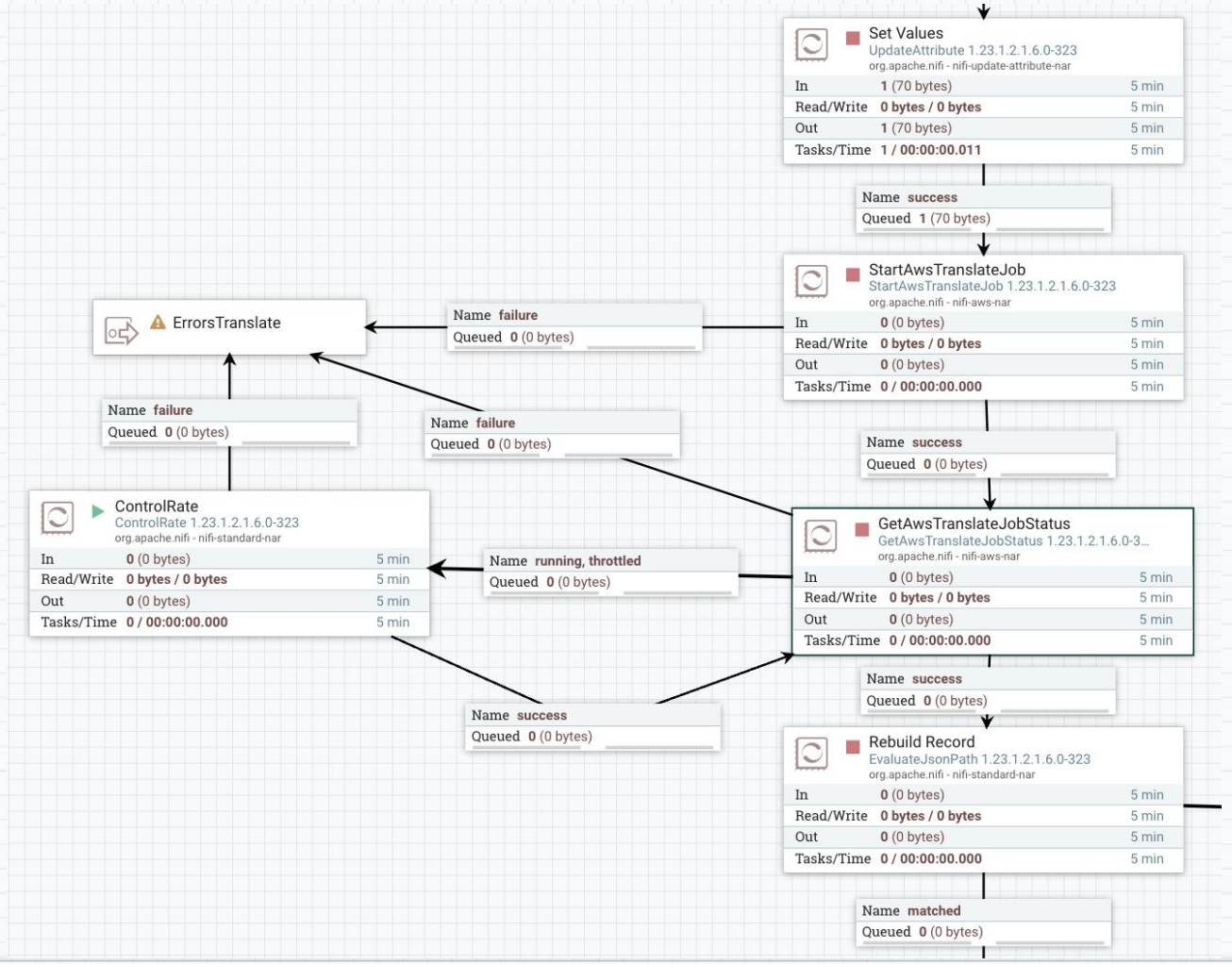
TriggerHiveMetaStoreEvent

“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

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

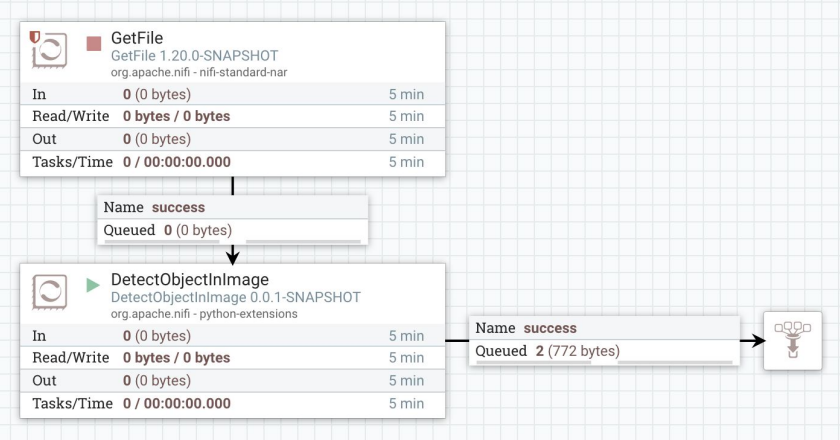
Apache NiFi Committer & PMC member | Working @Cloudera - ex-@Google | Twitter & Github — @pvillard31 | Blog @ www.pierrevillard.com

Python as First Class (NIFI-11241)

NEW in NiFi 2.0

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']
    class 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)

NiFi Deploy Options from Open Source to Managed

MiNiFi C++
(small footprint)

MiNiFi Java
(headless version of NiFi)

NiFi Registry

Stateless NiFi

Cloudera Edge Flow Manager
(Command & Control of MiNiFi Agents)

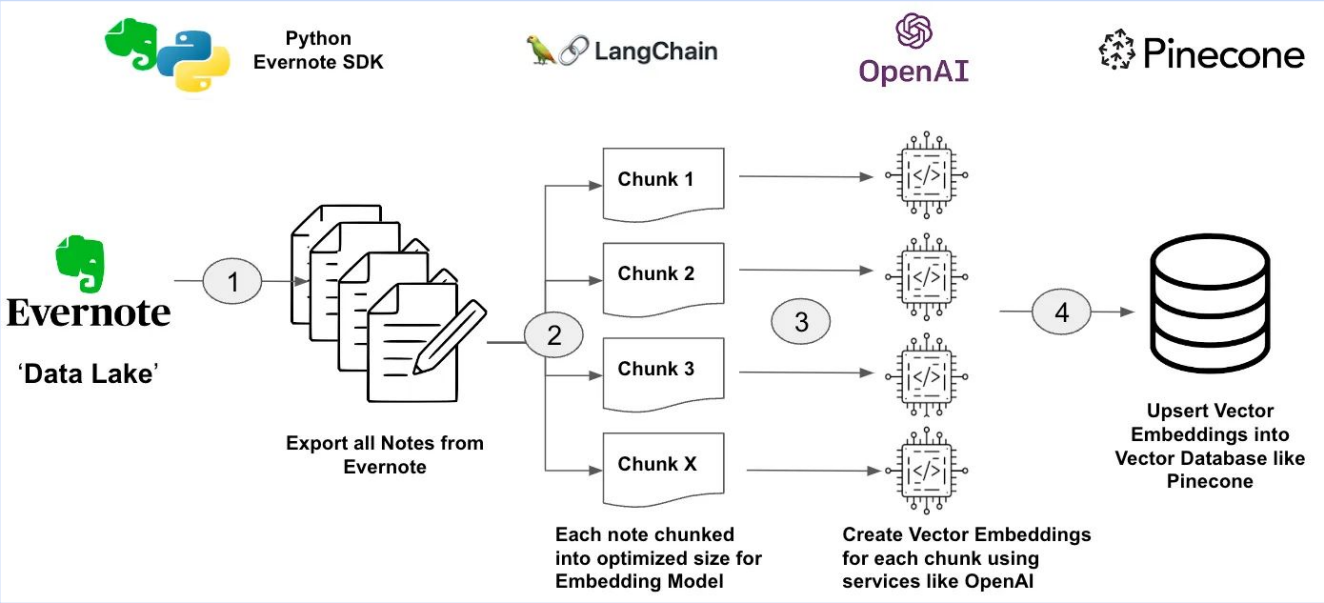
Cloudera NiFi for Kafka Connect

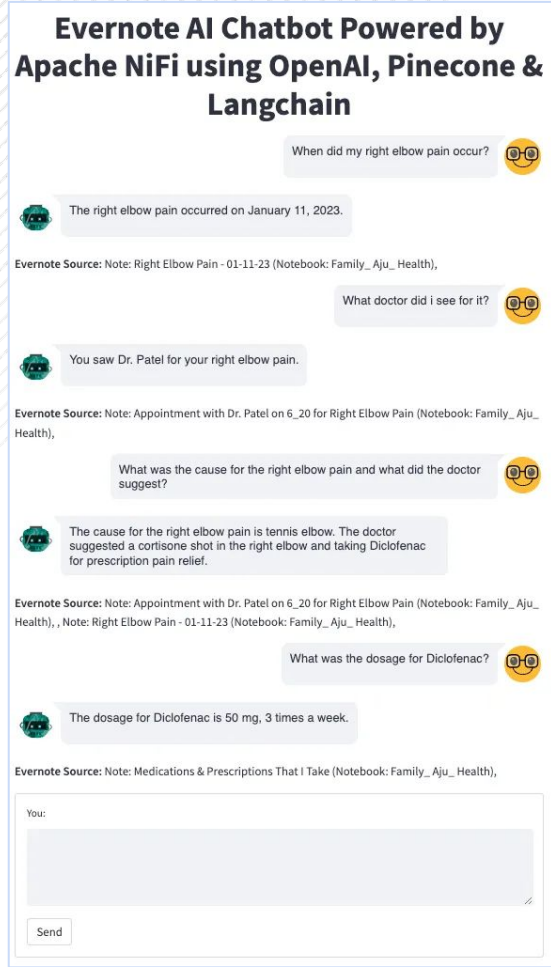
NiFi in Cloudera DataFlow Functions

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)**



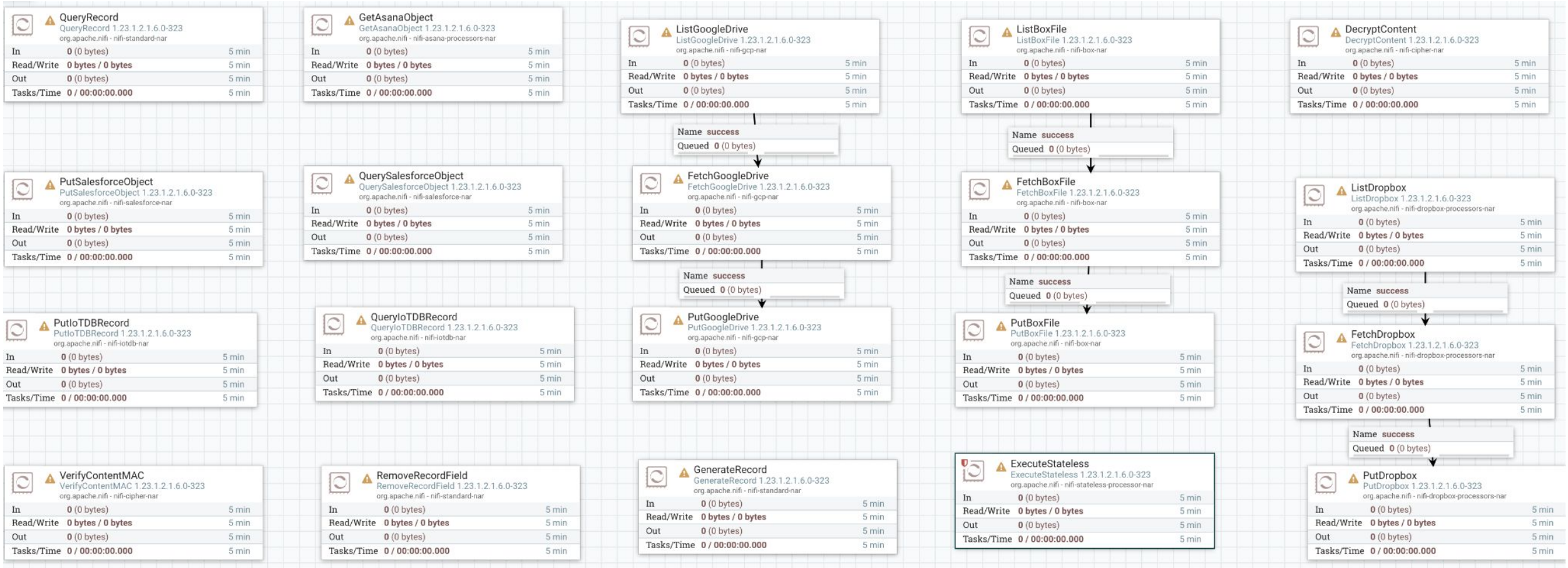


Closing the gap between data engineers and data scientists…

- 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

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

DEMO


QueryRecord
QueryRecord 1.23.1.2.1.6.0-323
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
GetAsanaObject
GetAsanaObject 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-asana-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
ListGoogleDrive
ListGoogleDrive 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-gcp-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)
ListBoxFile
ListBoxFile 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-box-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)
DecryptContent
DecryptContent 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-cipher-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
PutSalesforceObject
PutSalesforceObject 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-salesforce-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
QuerySalesforceObject
QuerySalesforceObject 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-salesforce-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
FetchGoogleDrive
FetchGoogleDrive 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-gcp-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)
FetchBoxFile
FetchBoxFile 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-box-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)
ListDropbox
ListDropbox 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-dropbox-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 success
Queued 0 (0 bytes)
PutIoTDBRecord
PutIoTDBRecord 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-iotdb-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
QueryIoTDBRecord
QueryIoTDBRecord 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-iotdb-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
PutGoogleDrive
PutGoogleDrive 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-gcp-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
PutBoxFile
PutBoxFile 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-box-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
FetchDropbox
FetchDropbox 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-dropbox-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 success
Queued 0 (0 bytes)
VerifyContentMAC
VerifyContentMAC 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-cipher-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
RemoveRecordField
RemoveRecordField 1.23.1.2.1.6.0-323
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
GenerateRecord
GenerateRecord 1.23.1.2.1.6.0-323
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
ExecuteStateless
ExecuteStateless 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-stateless-processor-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
PutDropbox
PutDropbox 1.23.1.2.1.6.0-323
org.apache.nifi - nifi-dropbox-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

THANK YOU