CLOUDERA

Building Modern Data Streaming Apps

Tim Spann
Principal Developer Advocate

25-May-2023

BIG DATA FEST
MAY, 24-25 2023

softserve | BIG DATA COMMUNITY

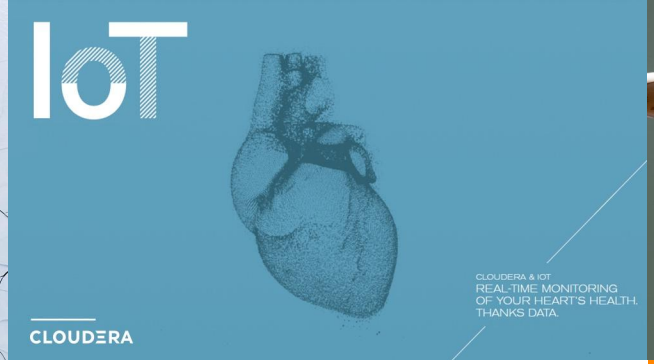
24-25 MAY 2023
ONLINE CONFERENCE

BIG DATA COMMUNITY

softserve

BUILDING MODERN DATA STREAMING APPS

Timothy Spann
Principal Developer Advocate
Cloudera

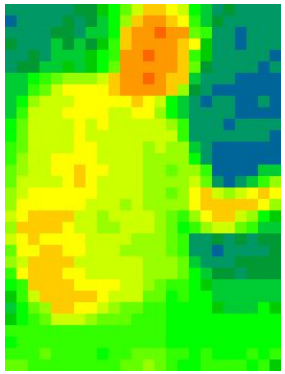


CLOUDERA

EDGE
2AI
CLOUDERA

ENTERPRISE
DATA
CLOUD
CLOUDERA

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://github.com/tspannhw/EverythingApacheNiFi
https://medium.com/@tspann
Apache NiFi x Apache Kafka x Apache Flink x Java

FLaNK Stack Weekly

https://bit.ly/32dAJft

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

Future of Data - Princeton + Virtual

FUTURE OF DATA

AN OPEN SOURCE COMMUNITY

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

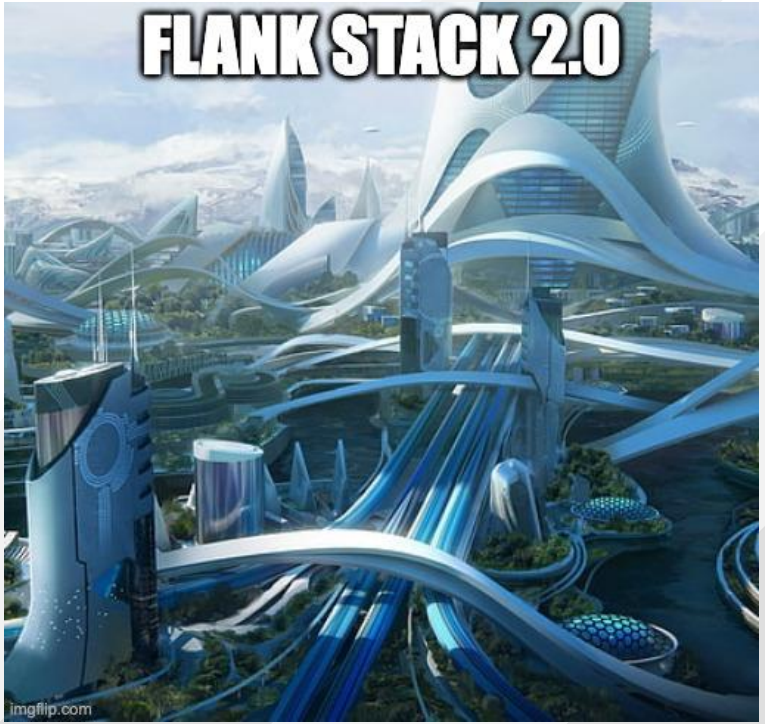
From Big Data to AI to Streaming to Containers to Cloud to Analytics to Cloud Storage to Fast Data to Machine Learning to Microservices to ...

ICEBERG

@PaasDev

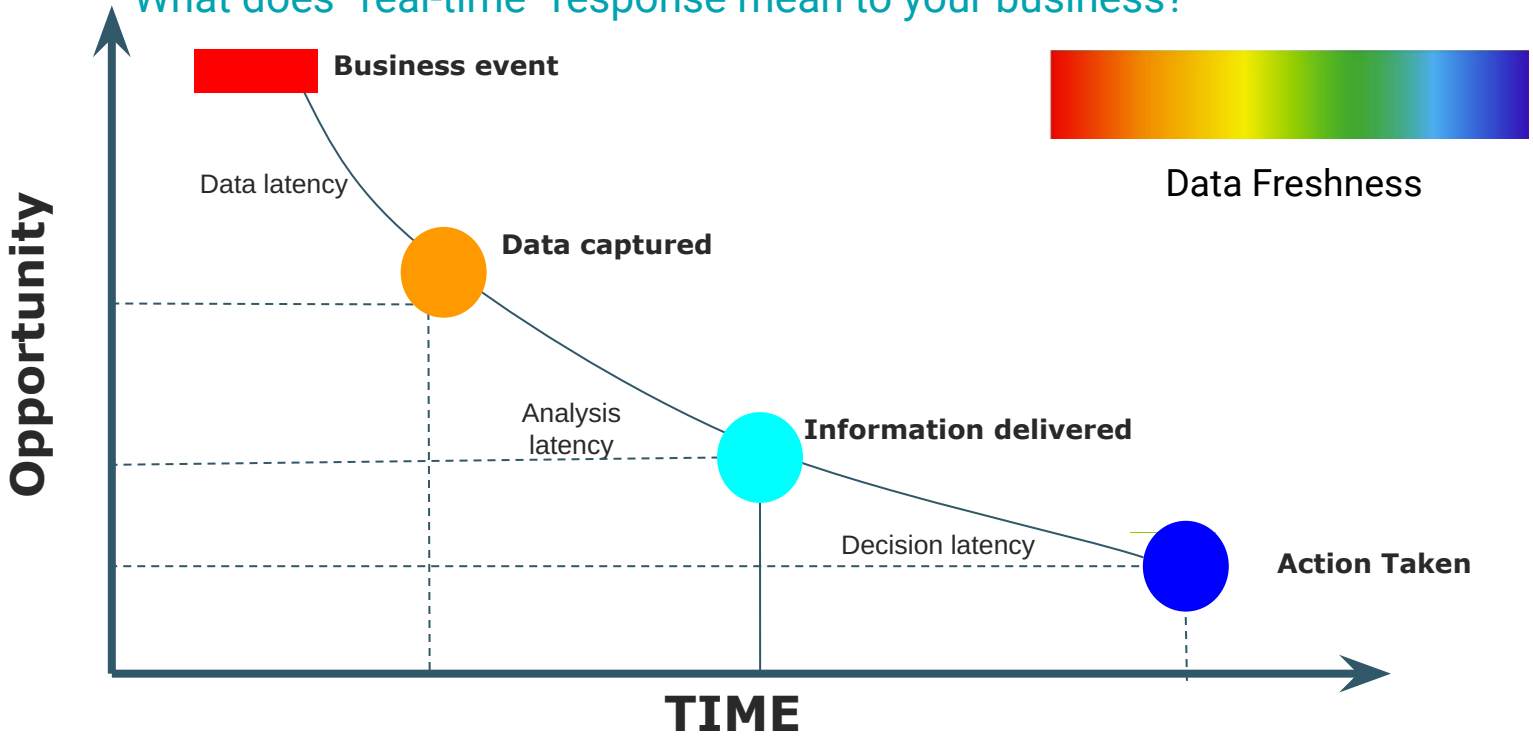
https://openeyes.org.ua/en/donate

STREAMING

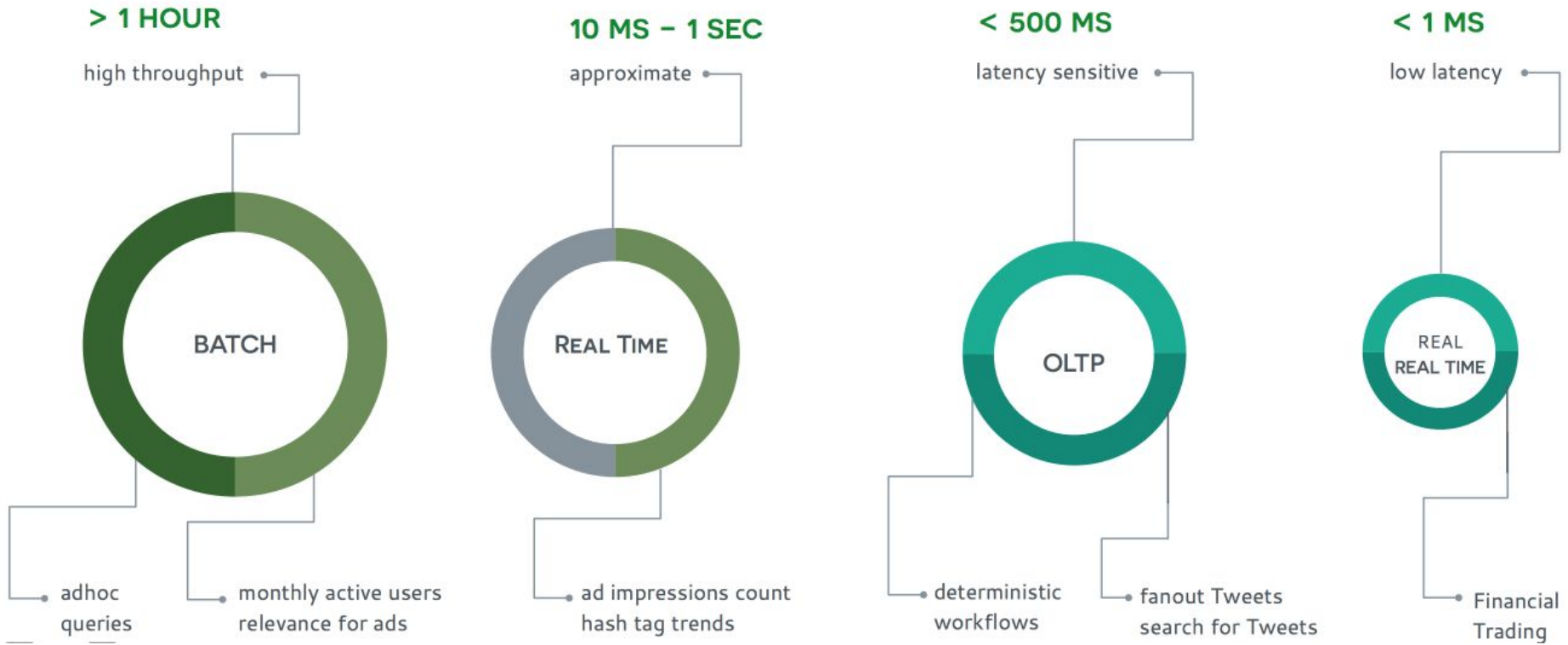


Speed Matters

What does “real-time” response mean to your business?

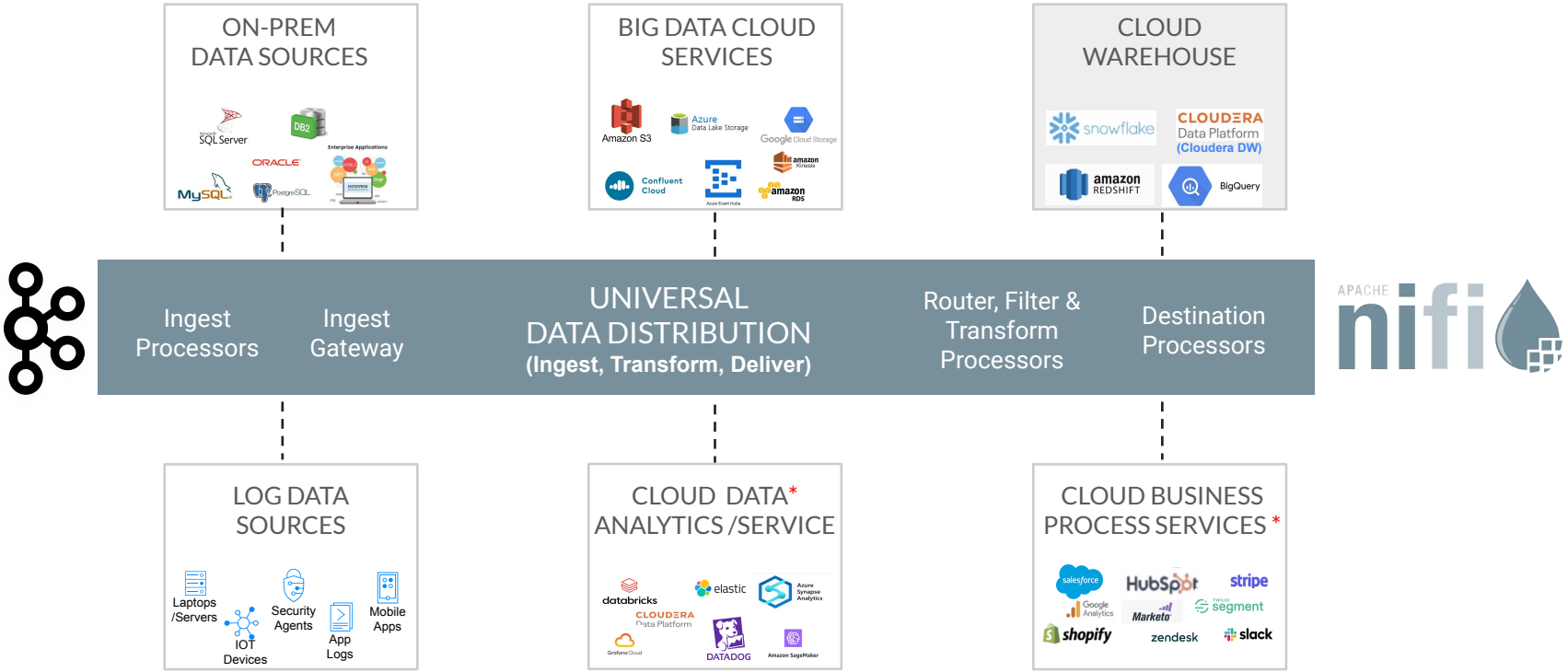


What is Real-Time?



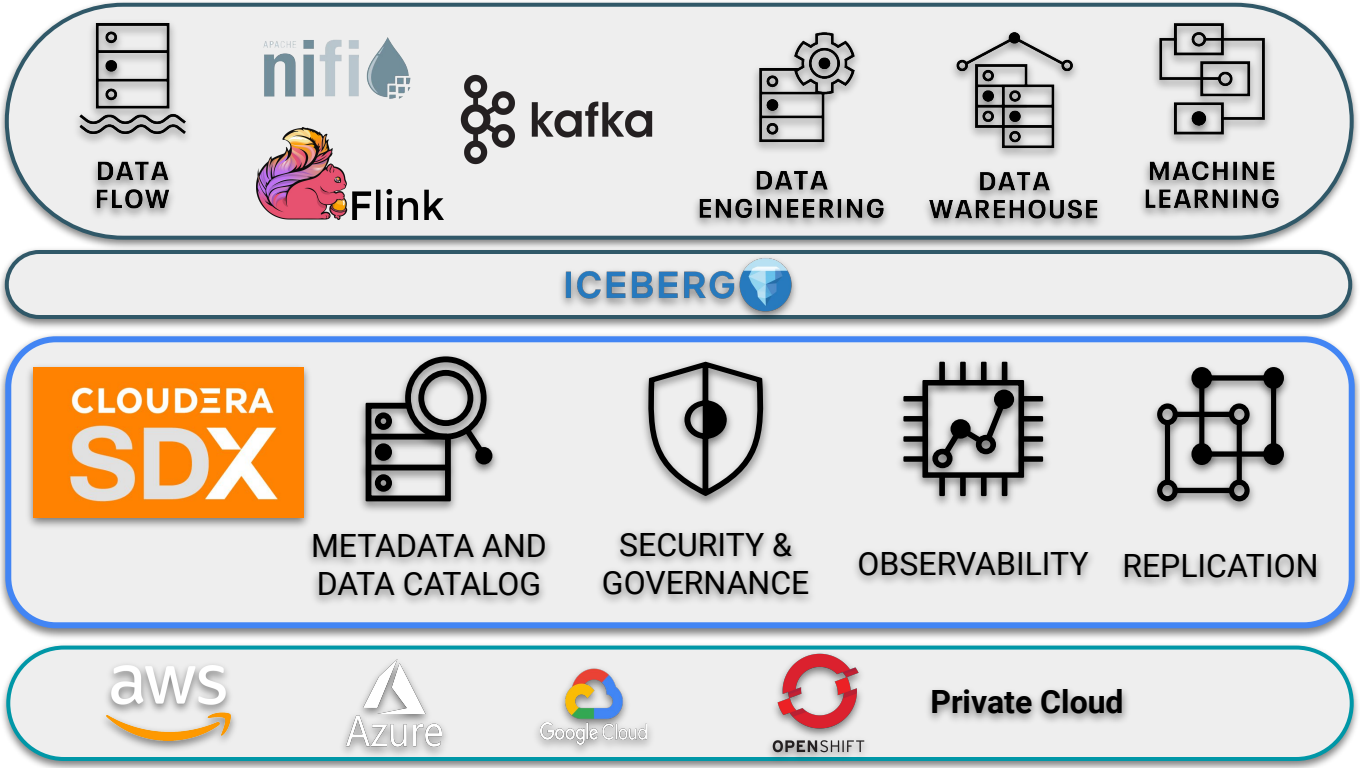
Streaming From … To …

Data distribution as a first class citizen



BUILDING REAL-TIME REQUIRES A TEAM

CDP: AN OPEN DATA LAKEHOUSE



APACHE KAFKA

What is Apache Kafka?

Distributed: horizontally scalable

Partitioned: the data is split-up and distributed across the brokers

Replicated: allows for automatic failover

Unique: Kafka does not track the consumption of messages (the consumers do)

Fast: designed from the ground up with a focus on performance and throughput

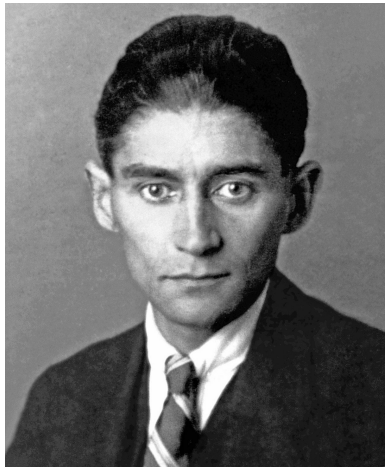
Kafka was built at Linkedin in 2011

Open sourced as an Apache project

Yes, Franz, It’s Kafka

Let’s do a metamorphosis on your data. Don’t fear changing data.

You don’t need to be a brilliant writer to stream data.



Franz Kafka was a German-speaking Bohemian novelist and short-story writer, widely regarded as one of the major figures of 20th-century literature. His work fuses elements of realism and the fantastic.
Wikipedia

What is Can You Do With Apache Kafka?

Web site activity: track page views, searches, etc. in real time

Events & log aggregation: particularly in distributed systems where messages come from multiple sources

Monitoring and metrics: aggregate statistics from distributed applications and build a dashboard application

Stream processing: process raw data, clean it up, and forward it on to another topic or messaging system

Real-time data ingestion: fast processing of a very large volume of messages

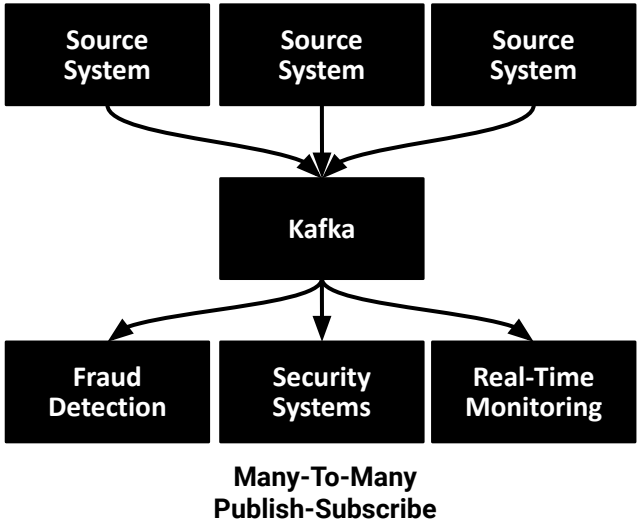
Kafka Terms

- Kafka is a publish/subscribe messaging system comprised of the following components:

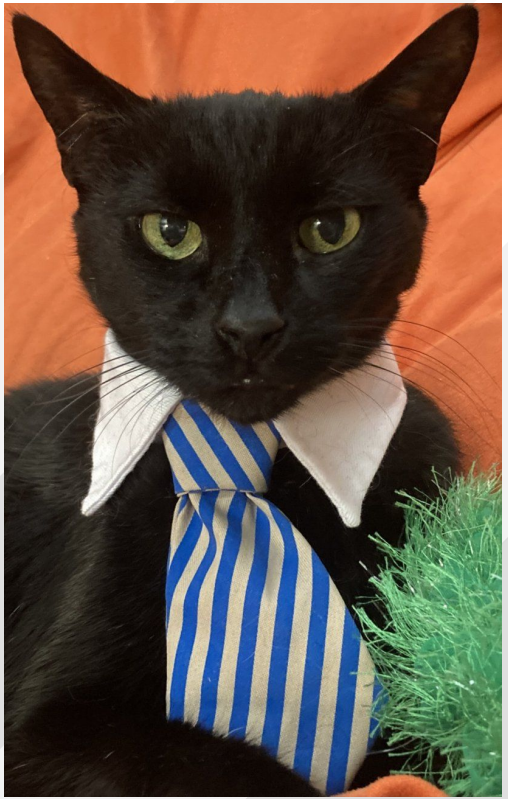
  - **Topic**: a message feed
  - **Producer**: a process that publishes messages to a topic
  - **Consumer**: a process that subscribes to a topic and processes its messages
  - **Broker**: a server in a Kafka cluster

- Highly reliable distributed messaging system
- Decouple applications, enables many-to-many patterns
- Publish-Subscribe semantics
- Horizontal scalability
- Efficient implementation to operate at speed with big data volumes
- Organized by topic to support several use cases

EVENTS

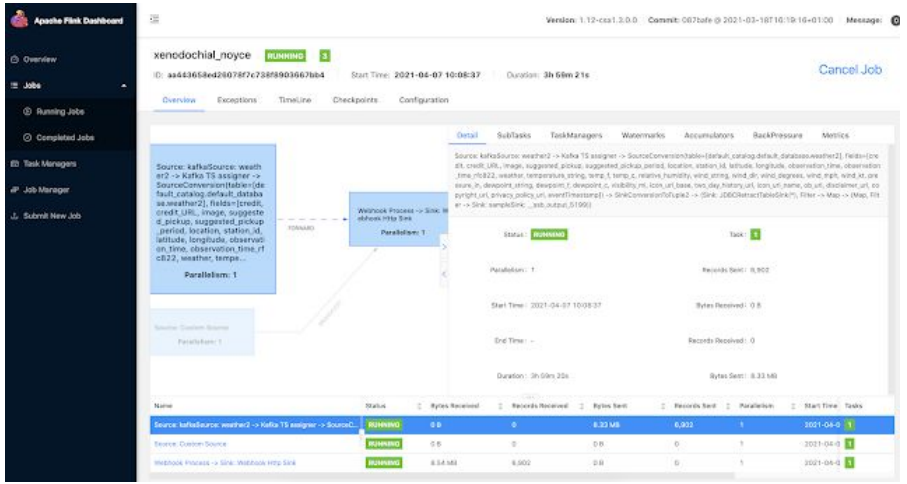


APACHE FLINK



Flink SQL

- Streaming Analytics
- Continuous SQL
- Continuous ETL
- Complex Event Processing
- Standard SQL Powered by Apache Calcite



https://www.datainmotion.dev/2021/04/cloudera-sql-stream-builder-ssb-updated.html

Flink SQL

Key Takeaway: Rich SQL grammar with advanced time and aggregation tools

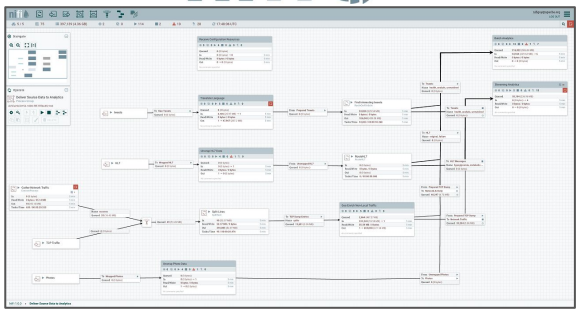
```
-- specify Kafka partition key on output
SELECT foo AS _eventKey FROM sensors

-- use event time timestamp from kafka
-- exactly once compatible
SELECT eventTimestamp FROM sensors

-- nested structures access
SELECT foo.'bar' FROM table; -- must quote nested
column

-- timestamps
SELECT * FROM payments
WHERE eventTimestamp > CURRENT_TIMESTAMP-interval
'10' second;

-- unnest
SELECT b.*, u.*
FROM bgp_avro b,
UNNEST(b.path) AS u(pathitem)
```

```
-- aggregations and windows
SELECT card,
MAX(amount) as theamount,
TUMBLE_END(eventTimestamp, interval '5' minute) as
ts
FROM payments
WHERE lat IS NOT NULL
AND lon IS NOT NULL
GROUP BY card,
TUMBLE(eventTimestamp, interval '5' minute)
HAVING COUNT(*) > 4  -- >4==fraud

-- try to do this ksql!
SELECT us_west.user_score+ap_south.user_score
FROM kafka_in_zone_us_west us_west
FULL OUTER JOIN kafka_in_zone_ap_south ap_south
ON us_west.user_id = ap_south.user_id;
```

DATAFLOW APACHE NIFI

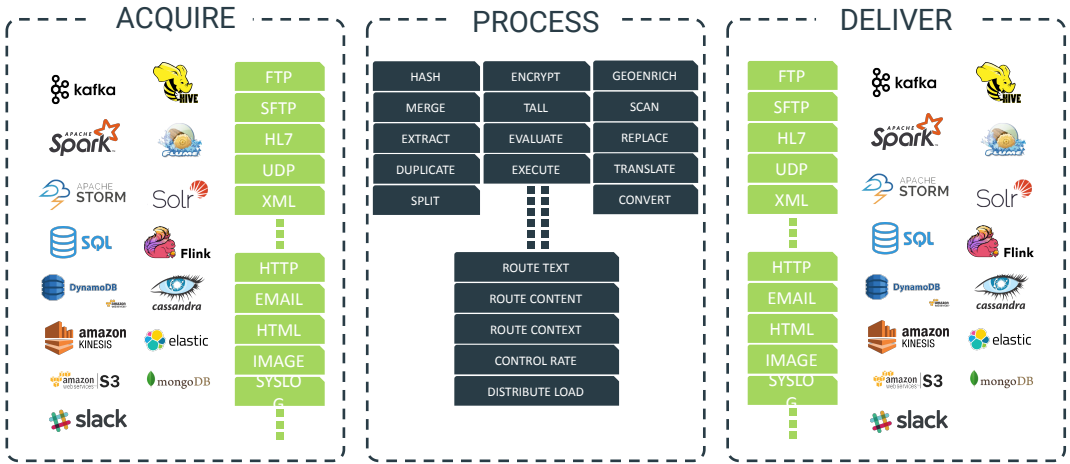
Apache NiFi

Enable easy ingestion, routing, management and delivery of any data anywhere (Edge, cloud, data center) to any downstream system with built in end-to-end security and provenance



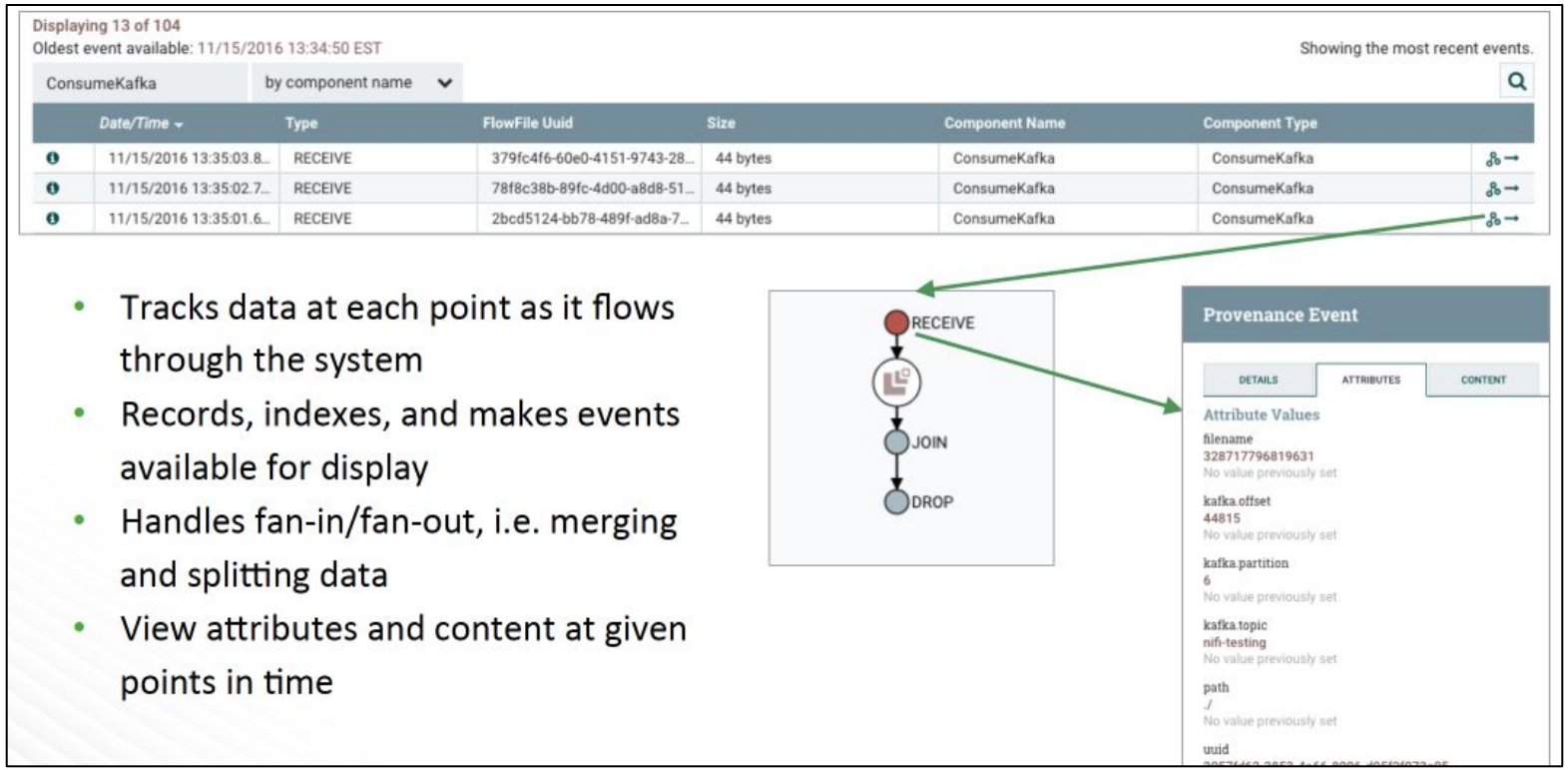
APACHE NIFI registry

Advanced tooling to industrialize flow development (Flow Development Life Cycle)



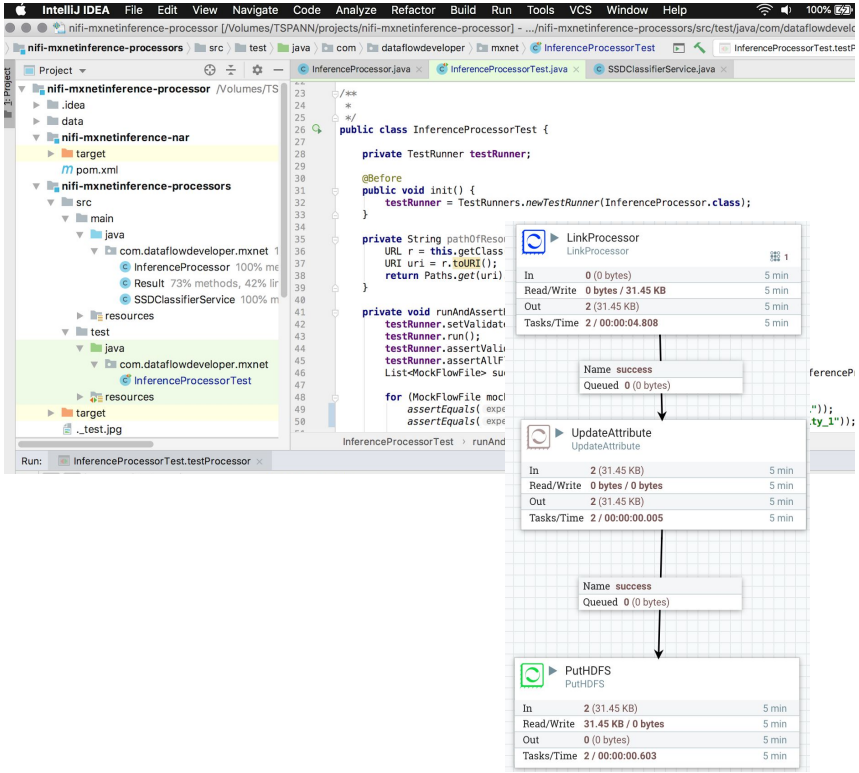
- Over 450 Prebuilt Processors
- Easy to build your own
- Parse, Enrich & Apply Schema
- Filter, Split, Merger & Route
- Throttle & Backpressure
- Guaranteed Delivery
- Full data provenance from acquisition to delivery
- Diverse, Non-Traditional Sources
- Eco-system integration

Provenance



Extensibility

- Built from the ground up with extensions in mind
- Service-loader pattern for…
  - Processors
  - Controller Services
  - Reporting Tasks
  - Prioritizers
- Extensions packaged as NiFi Archives (NARs)
  - Deploy NiFi lib directory and restart
  - Same model as standard components

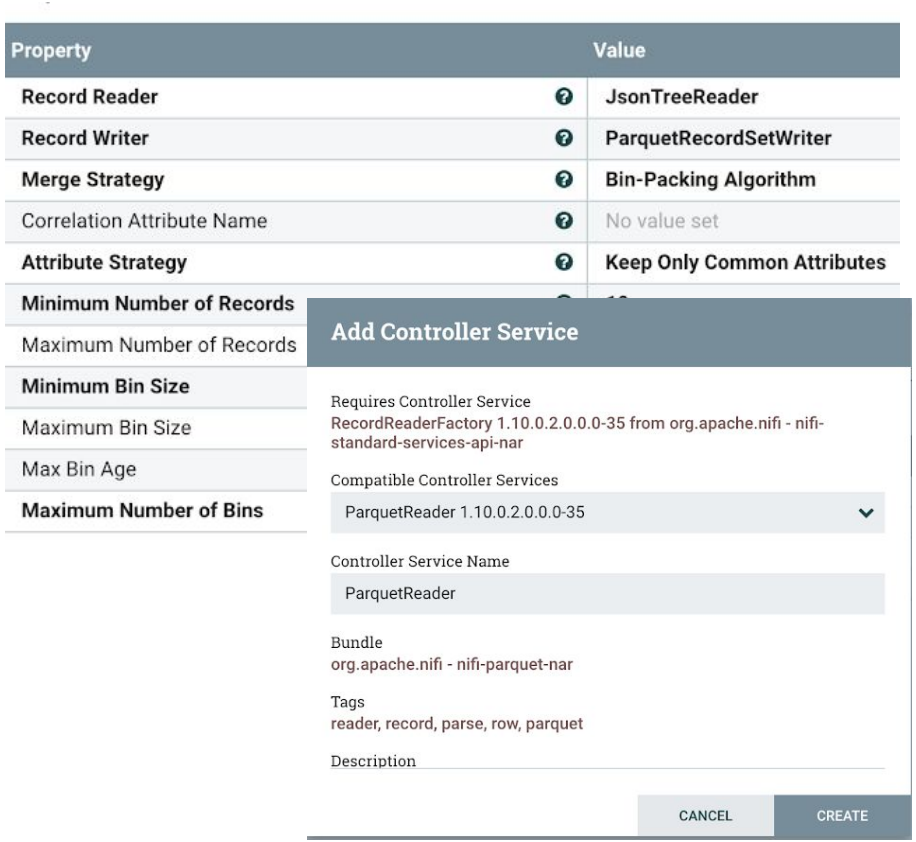


Custom Processors

https://github.com/tspannhw/nifi-extracttext-processor
https://github.com/tspannhw/nifi-tensorflow-processor
https://github.com/tspannhw/nifi-nlp-processor
https://github.com/tspannhw/nifi-convertjsontoddl-processor
https://github.com/tspannhw/nifi-corenlp-processor
https://github.com/tspannhw/nifi-imageextractor-processor
https://github.com/tspannhw/nifi-attributecleaner-processor
https://github.com/tspannhw/linkextractorprocessor
https://github.com/tspannhw/GetWebCamera
https://github.com/tspannhw/nifi-langdetect-processor
https://github.com/tspannhw/nifi-postimage-processor

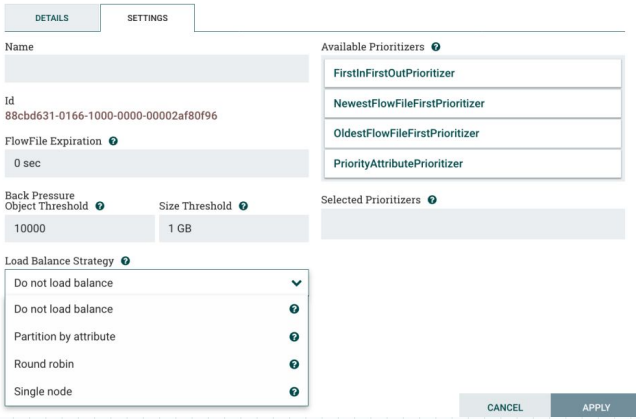
Parquet Reader/ Writers

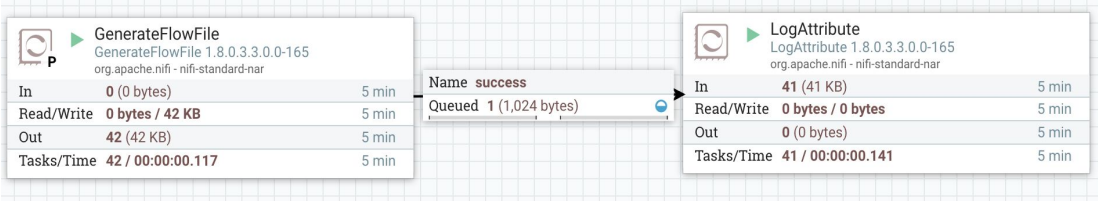
- Native Record Processors for Apache Parquet Files!
- CSV <-> Parquet
- XML <-> Parquet
- AVRO <-> Parquet
- JSON <-> Parquet
- More...



NiFi Load Balancing

- Improve NiFi cluster throughput
- Defined at connection level
- Configurable balancing strategies
- Critical for scale up paradigm in Kubernetes

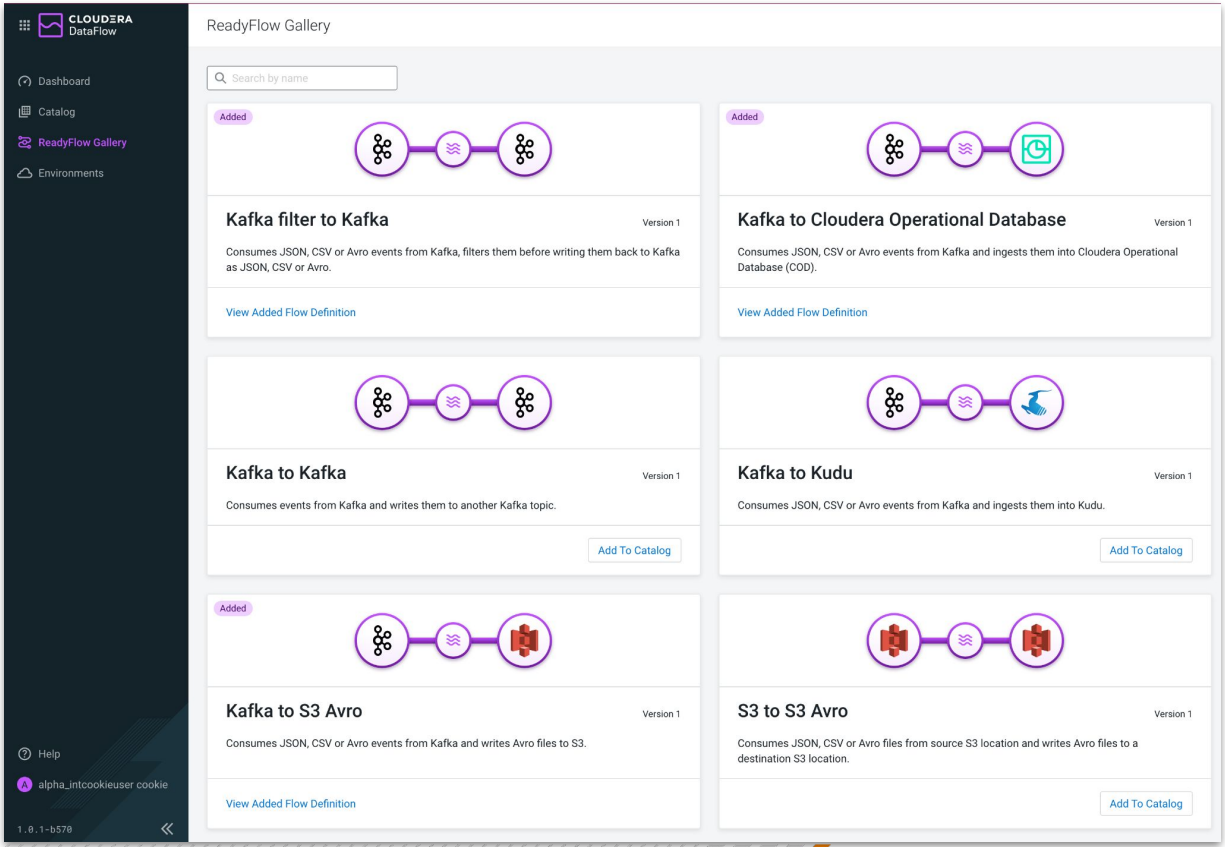




FlowFile Uuid	Size	Component Name	Component Type	Node
0733ad94-3c80-44d7-9fc2-480caaa...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe0.field.hortonworks.com:9...
2bc7b5c1-c164-40fb-9e7e-da57884...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe0.field.hortonworks.com:9...
80dece7a-15c8-4eb7-80ad-176bfe9...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe1.field.hortonworks.com:9...
98d9f9c4-bb47-4fe7-9786-964d027...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe2.field.hortonworks.com:9...
26c165ca-2f6d-4714-8c0c-e1de6e2...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe1.field.hortonworks.com:9...
8bff920b-97a3-4b64-998d-046324a...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe2.field.hortonworks.com:9...
6345a326-4843-442e-b77d-480d20...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe1.field.hortonworks.com:9...
5fc30a5a-641e-4aa0-9c67-3b1d438...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe2.field.hortonworks.com:9...
1e90e7ee-92fe-47b3-9aa4-4094fa8...	1,024 bytes	LogAttribute	LogAttribute	hdf-pm-fe1.field.hortonworks.com:9...

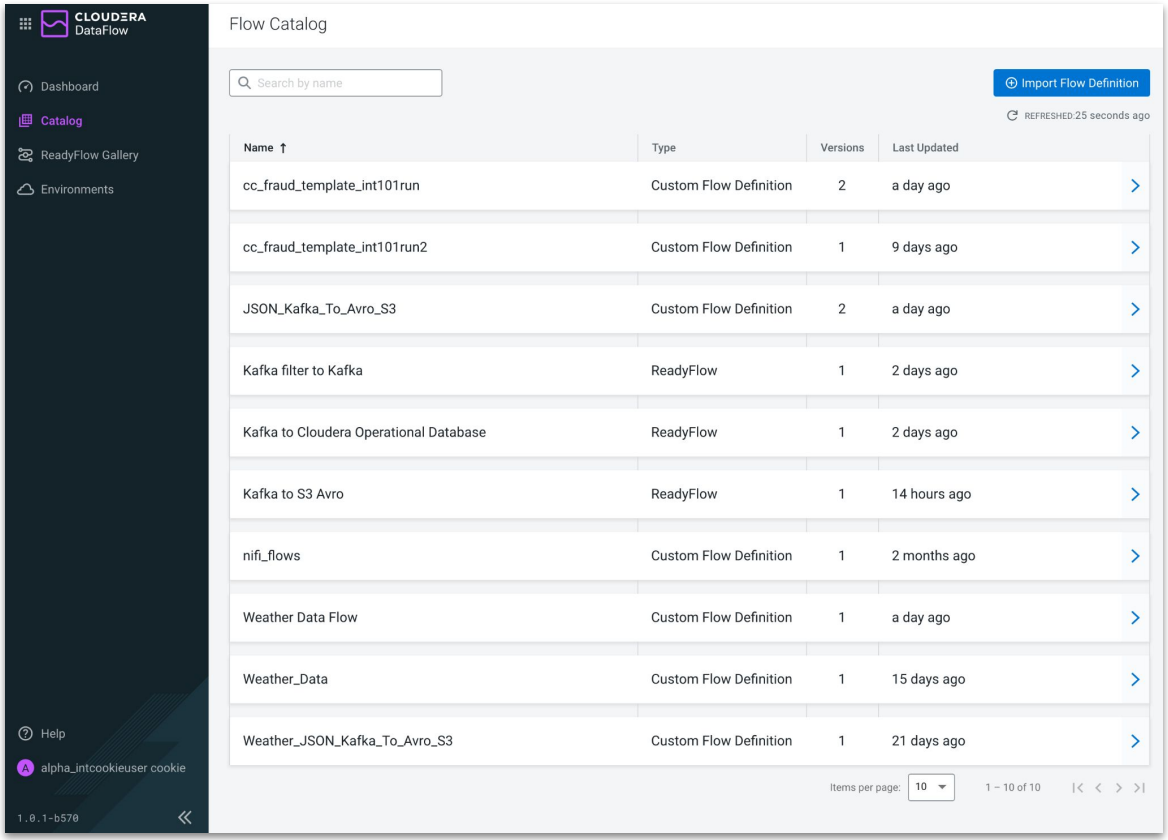
ReadyFlow Gallery

- Cloudera provided flow definitions
- Cover most common data flow use cases
- Optimized to work with CDP sources/destinations
- Can be deployed and adjusted as needed



Flow Catalog

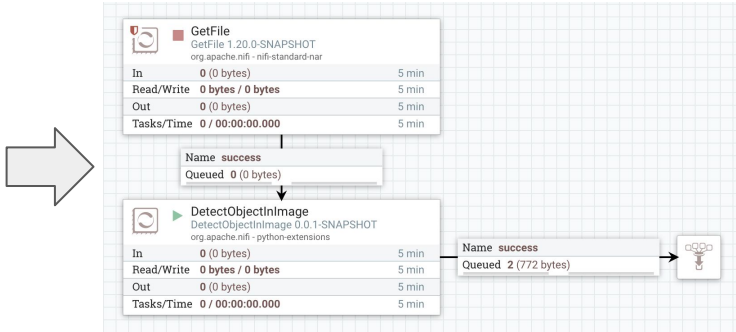
- Central repository for flow definitions
- Import existing NiFi flows
- Manage flow definitions
- Initiate flow deployments



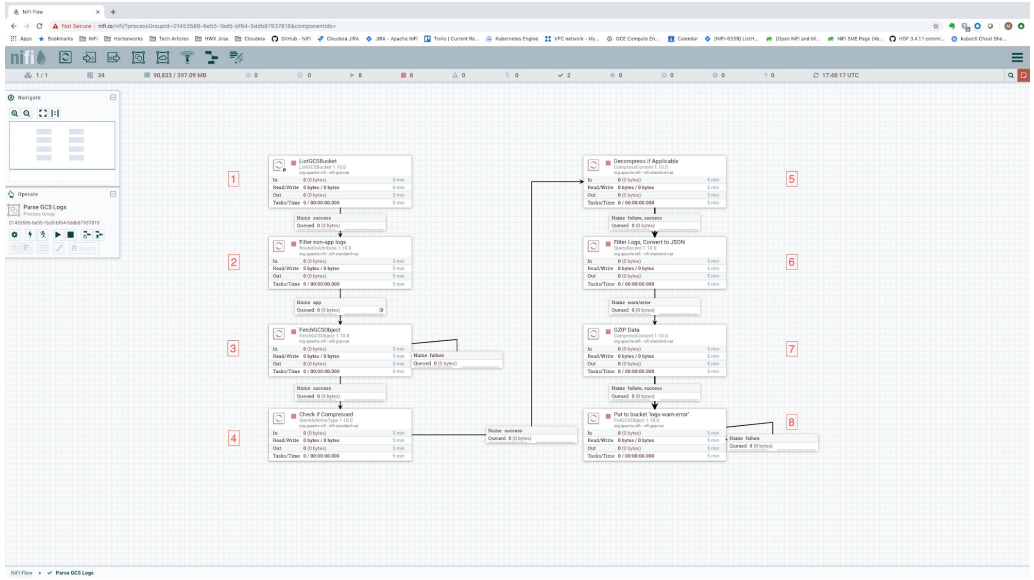
Apache NiFi with Python Custom Processors

Python as a 1st class citizen

```
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:
```



Processing millions of events with NiFi



Nodes	Data rate/sec	Events/sec	Data rate/day	Events/day
1	192.5 MB	946,000	16.6 TB	81.7 Billion
5	881 MB	4.97 Million	76 TB	429.4 Billion
25	5.8 GB	26 Million	501 TB	2.25 Trillion
100	22 GB	90 Million	1.9 PB	7.8 Trillion
150	32.6 GB	141.3 Million	2.75 PB	12.2 Trillion

SOURCES AND SINKS

APACHE ICEBERG

A Flexible, Performant & Scalable Table Format

- Donated by Netflix to the Apache Foundation in 2018
- Flexibility
  - Hidden partitioning
  - Full schema evolution
- Data Warehouse Operations
  - Atomic Consistent Isolated Durable (ACID) Transactions
  - Time travel and rollback
- Supports best in class SQL performance
  - High performance at Petabyte scale

AMQP

AWS Lambda

Airtable

Amazon API Gateway

Amazon CloudWatch

Amazon DynamoDB

Amazon Kinesis Data Firehose

Amazon Kinesis Data Streams

Amazon SQS

Amazon Simple Notification Services (SNS)

Amazon Simple Storage Service (S3)

Apache Accumulo

Apache Cassandra

Apache HBase

Apache Hive

Apache Iceberg

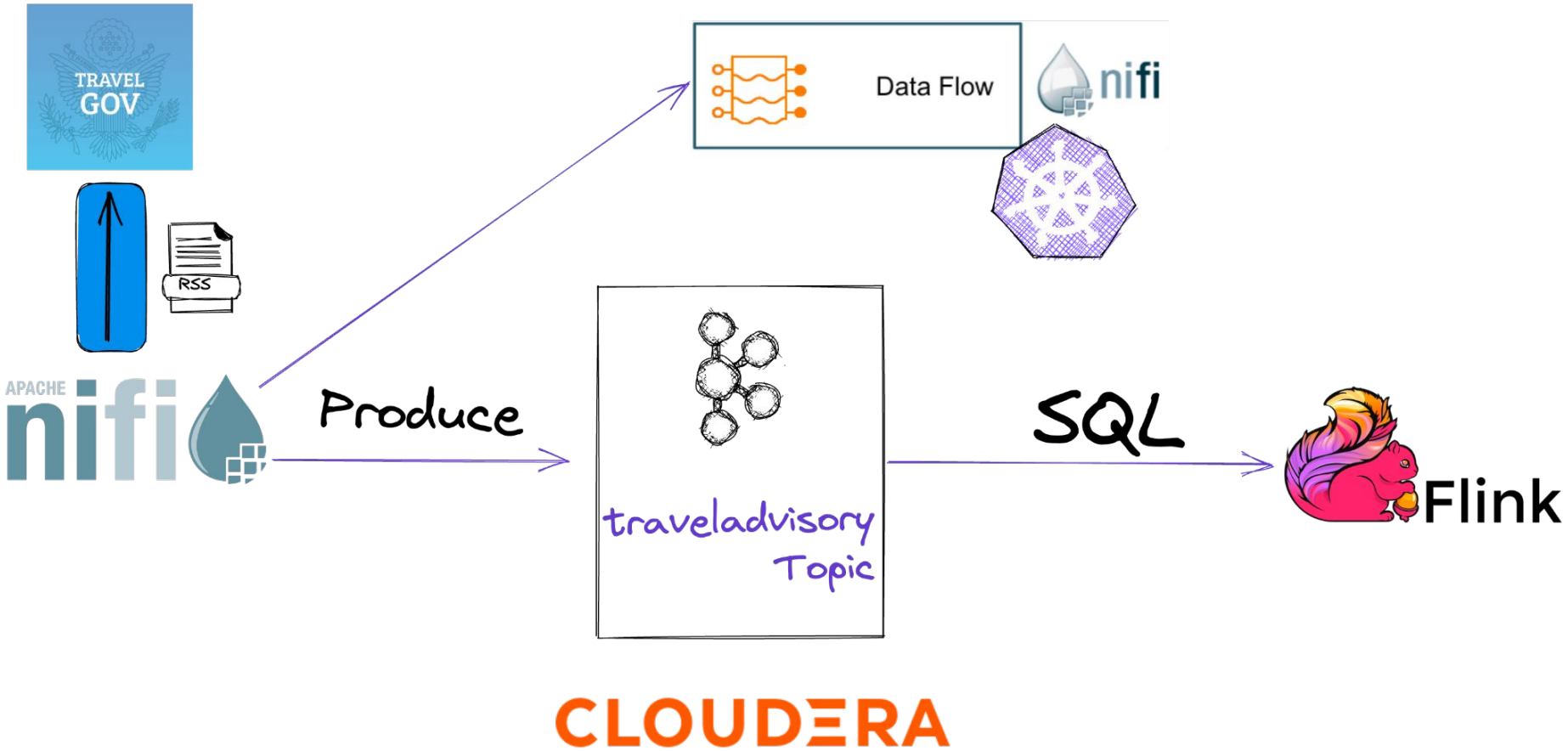
Apache Ignite

Apache Kafka

Apache Kudu

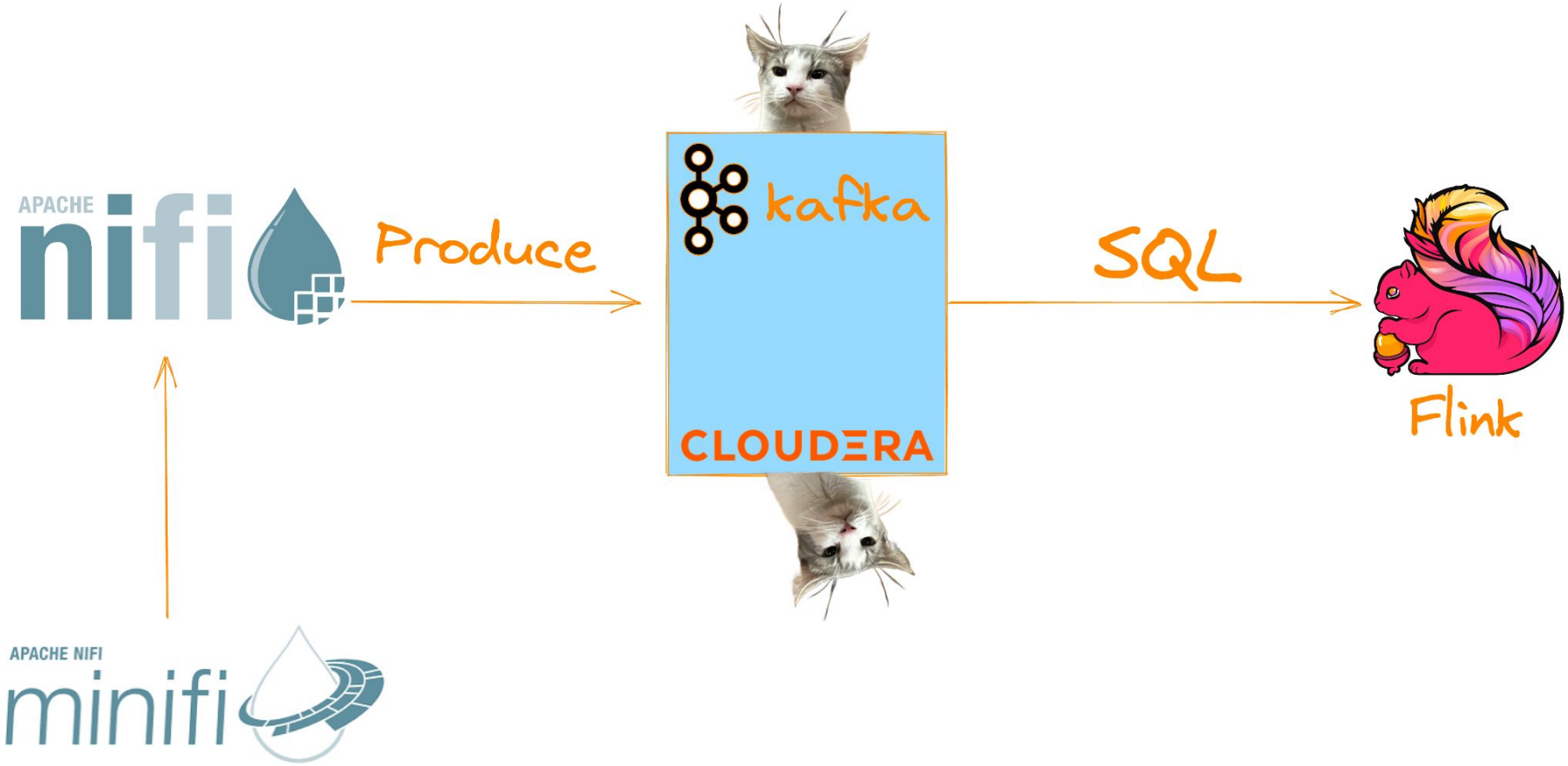
Apache Solr

DEMO AND CODE

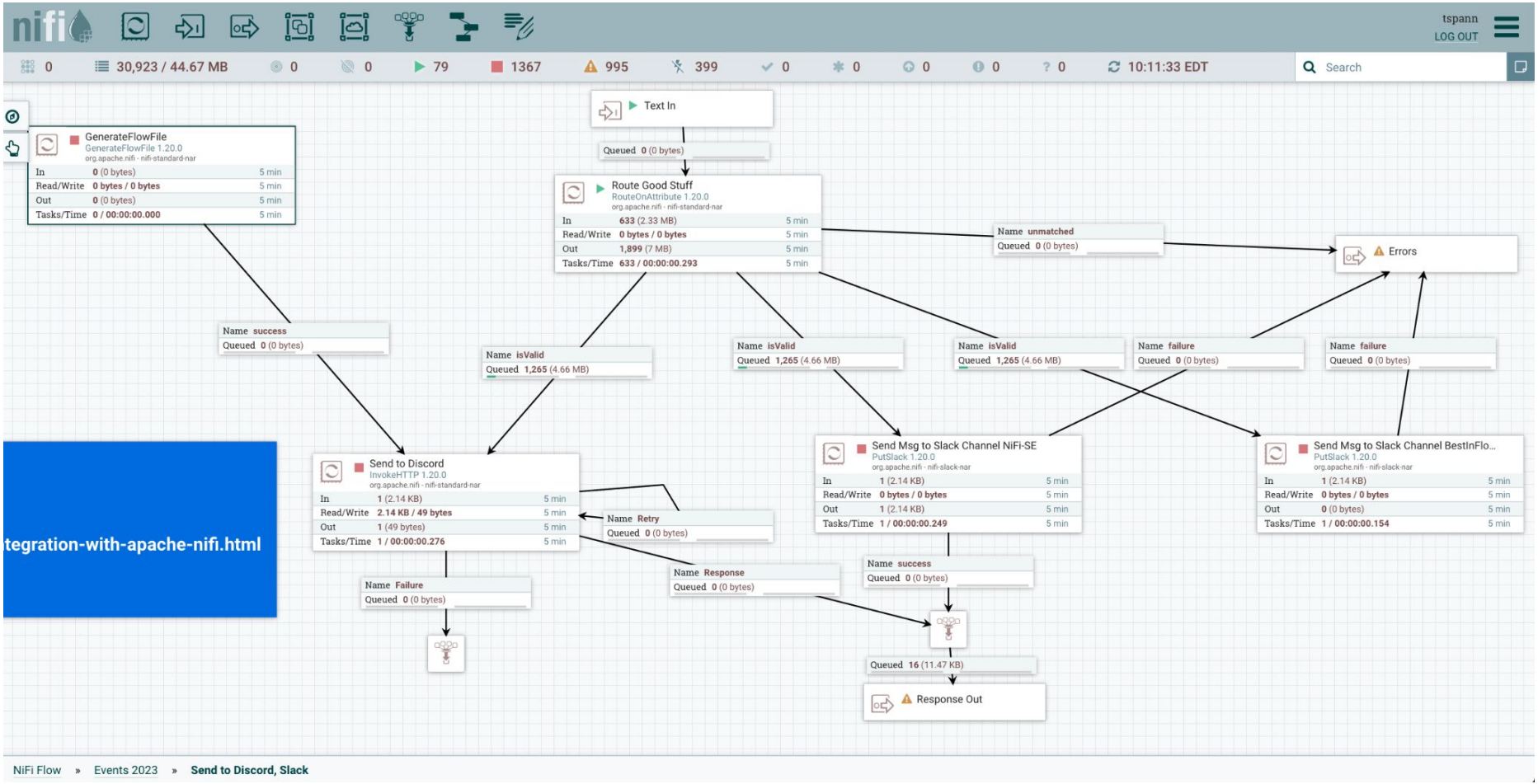


CLOUDERA

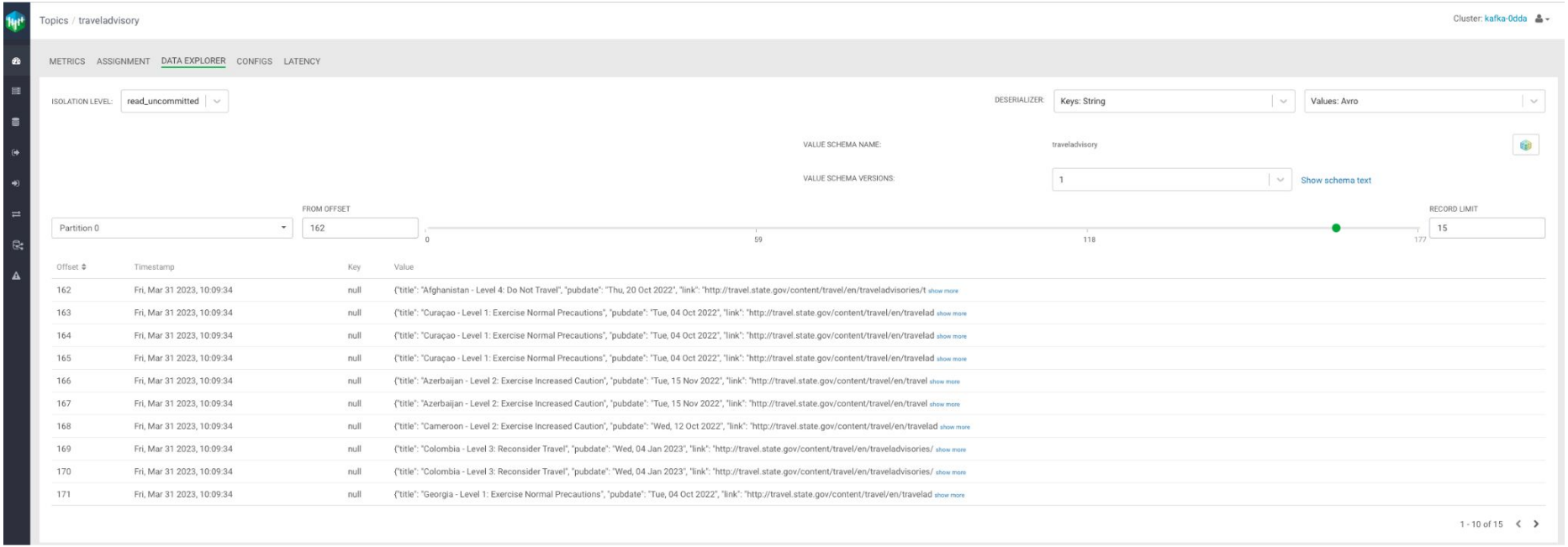
https://github.com/tspannhw/FLaNK-TravelAdvisory

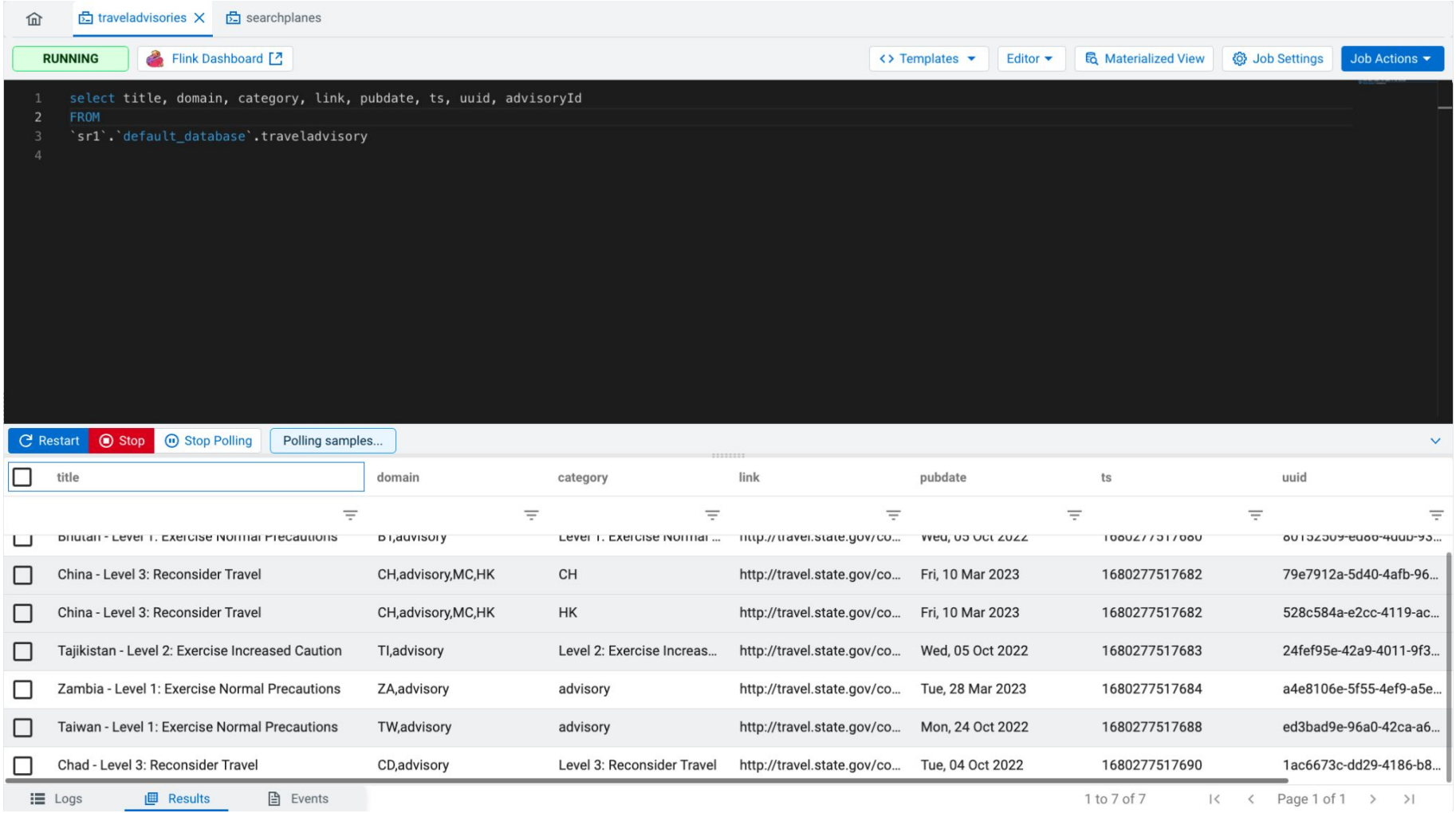


https://github.com/tspannhw/FLaNK-Edge


nifi
tspann
LOG OUT
0
30,923 / 44.67 MB
0
0
79
1367
995
399
0
0
0
0
0
10:11:33 EDT
Search
Text In
Queued 0 (0 bytes)
GenerateFlowFile
GenerateFlowFile 1.20.0
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
Route Good Stuff
RouteOnAttribute 1.20.0
org.apache.nifi - nifi-standard-nar
In 633 (2.33 MB) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1,899 (7 MB) 5 min
Tasks/Time 633 / 00:00:00.293 5 min
Name unmatched
Queued 0 (0 bytes)
Errors
Name success
Queued 0 (0 bytes)
Name isValid
Queued 1,265 (4.66 MB)
Name isValid
Queued 1,265 (4.66 MB)
Name isValid
Queued 1,265 (4.66 MB)
Name failure
Queued 0 (0 bytes)
Name failure
Queued 0 (0 bytes)
tegration-with-apache-nifi.html
Send to Discord
InvokeHTTP 1.20.0
org.apache.nifi - nifi-standard-nar
In 1 (2.14 KB) 5 min
Read/Write 2.14 KB / 49 bytes 5 min
Out 1 (49 bytes) 5 min
Tasks/Time 1 / 00:00:00.276 5 min
Name Retry
Queued 0 (0 bytes)
Send Msg to Slack Channel NiFi-SE
PutSlack 1.20.0
org.apache.nifi - nifi-slack-nar
In 1 (2.14 KB) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 1 (2.14 KB) 5 min
Tasks/Time 1 / 00:00:00.249 5 min
Send Msg to Slack Channel BestInFlo...
PutSlack 1.20.0
org.apache.nifi - nifi-slack-nar
In 1 (2.14 KB) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:00.154 5 min
Name Failure
Queued 0 (0 bytes)
Name Response
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
Queued 16 (11.47 KB)
Response Out
NiFi Flow » Events 2023 » Send to Discord, Slack

```
CREATE TABLE `sr1`.`default_database`.`traveladvisory` (
  `title` VARCHAR(2147483647),
  `pubdate` VARCHAR(2147483647),
  `link` VARCHAR(2147483647),
  `guid` VARCHAR(2147483647),
  `advisoryId` VARCHAR(2147483647),
  `domain` VARCHAR(2147483647),
  `category` VARCHAR(2147483647),
  `description` VARCHAR(2147483647),
  `uuid` VARCHAR(2147483647),
  `ts` BIGINT NOT NULL
) COMMENT 'traveladvisory'
WITH (
  'properties.bootstrap.servers' = 'kafka:9092',
  'avro-cloudera.properties.schema.registry.url' = 'http://schema-registry:7788/api/v1',
  'connector' = 'kafka',
  'avro-cloudera.schema-name' = 'traveladvisory',
  'format' = 'avro-cloudera',
  'topic' = 'traveladvisory',
  'scan.startup.mode' = 'latest-offset'
)
```


Topics / traveladvisory
Cluster: kafka-0dda
METRICS ASSIGNMENT DATA EXPLORER CONFIGS LATENCY
ISOLATION LEVEL: read_uncommitted
DESERIALIZER: Keys: String Values: Avro
VALUE SCHEMA NAME: traveladvisory
VALUE SCHEMA VERSIONS: 1 Show schema text
Partition 0
FROM OFFSET 162
0 59 118 177
RECORD LIMIT 15
Offset Timestamp Key Value
162 Fri, Mar 31 2023, 10:09:34 null {"title": "Afghanistan - Level 4: Do Not Travel", "pubdate": "Thu, 20 Oct 2022", "link": "http://travel.state.gov/content/travel/en/traveladvisories/t show more
163 Fri, Mar 31 2023, 10:09:34 null {"title": "Curaçao - Level 1: Exercise Normal Precautions", "pubdate": "Tue, 04 Oct 2022", "link": "http://travel.state.gov/content/travel/en/travelad show more
164 Fri, Mar 31 2023, 10:09:34 null {"title": "Curaçao - Level 1: Exercise Normal Precautions", "pubdate": "Tue, 04 Oct 2022", "link": "http://travel.state.gov/content/travel/en/travelad show more
165 Fri, Mar 31 2023, 10:09:34 null {"title": "Curaçao - Level 1: Exercise Normal Precautions", "pubdate": "Tue, 04 Oct 2022", "link": "http://travel.state.gov/content/travel/en/travelad show more
166 Fri, Mar 31 2023, 10:09:34 null {"title": "Azerbaijan - Level 2: Exercise Increased Caution", "pubdate": "Tue, 15 Nov 2022", "link": "http://travel.state.gov/content/travel/en/travel show more
167 Fri, Mar 31 2023, 10:09:34 null {"title": "Azerbaijan - Level 2: Exercise Increased Caution", "pubdate": "Tue, 15 Nov 2022", "link": "http://travel.state.gov/content/travel/en/travel show more
168 Fri, Mar 31 2023, 10:09:34 null {"title": "Cameroon - Level 2: Exercise Increased Caution", "pubdate": "Wed, 12 Oct 2022", "link": "http://travel.state.gov/content/travel/en/travelad show more
169 Fri, Mar 31 2023, 10:09:34 null {"title": "Colombia - Level 3: Reconsider Travel", "pubdate": "Wed, 04 Jan 2023", "link": "http://travel.state.gov/content/travel/en/traveladvisories/ show more
170 Fri, Mar 31 2023, 10:09:34 null {"title": "Colombia - Level 3: Reconsider Travel", "pubdate": "Wed, 04 Jan 2023", "link": "http://travel.state.gov/content/travel/en/traveladvisories/ show more
171 Fri, Mar 31 2023, 10:09:34 null {"title": "Georgia - Level 1: Exercise Normal Precautions", "pubdate": "Tue, 04 Oct 2022", "link": "http://travel.state.gov/content/travel/en/travelad show more
1 - 10 of 15


traveladvisories
searchplanes
RUNNING
Flink Dashboard
Templates
Editor
Materialized View
Job Settings
Job Actions
select title, domain, category, link, pubdate, ts, uuid, advisoryId
FROM
`sr1`.`default_database`.traveladvisory
Restart
Stop
Stop Polling
Polling samples...
title
domain
category
link
pubdate
ts
uuid
China - Level 3: Reconsider Travel
CH,advisory,MC,HK
CH
http://travel.state.gov/co...
Fri, 10 Mar 2023
1680277517682
79e7912a-5d40-4afb-96...
China - Level 3: Reconsider Travel
CH,advisory,MC,HK
HK
http://travel.state.gov/co...
Fri, 10 Mar 2023
1680277517682
528c584a-e2cc-4119-ac...
Tajikistan - Level 2: Exercise Increased Caution
TI,advisory
Level 2: Exercise Increas...
http://travel.state.gov/co...
Wed, 05 Oct 2022
1680277517683
24fef95e-42a9-4011-9f3...
Zambia - Level 1: Exercise Normal Precautions
ZA,advisory
advisory
http://travel.state.gov/co...
Tue, 28 Mar 2023
1680277517684
a4e8106e-5f55-4ef9-a5e...
Taiwan - Level 1: Exercise Normal Precautions
TW,advisory
advisory
http://travel.state.gov/co...
Mon, 24 Oct 2022
1680277517688
ed3bad9e-96a0-42ca-a6...
Chad - Level 3: Reconsider Travel
CD,advisory
Level 3: Reconsider Travel
http://travel.state.gov/co...
Tue, 04 Oct 2022
1680277517690
1ac6673c-dd29-4186-b8...
Logs
Results
Events
1 to 7 of 7
Page 1 of 1

Materialized View

Configuration

Primary Key

uuid

Retention (Seconds)

Min Row Retention Count

10000

API Key

traveladvisory1

Enable MV

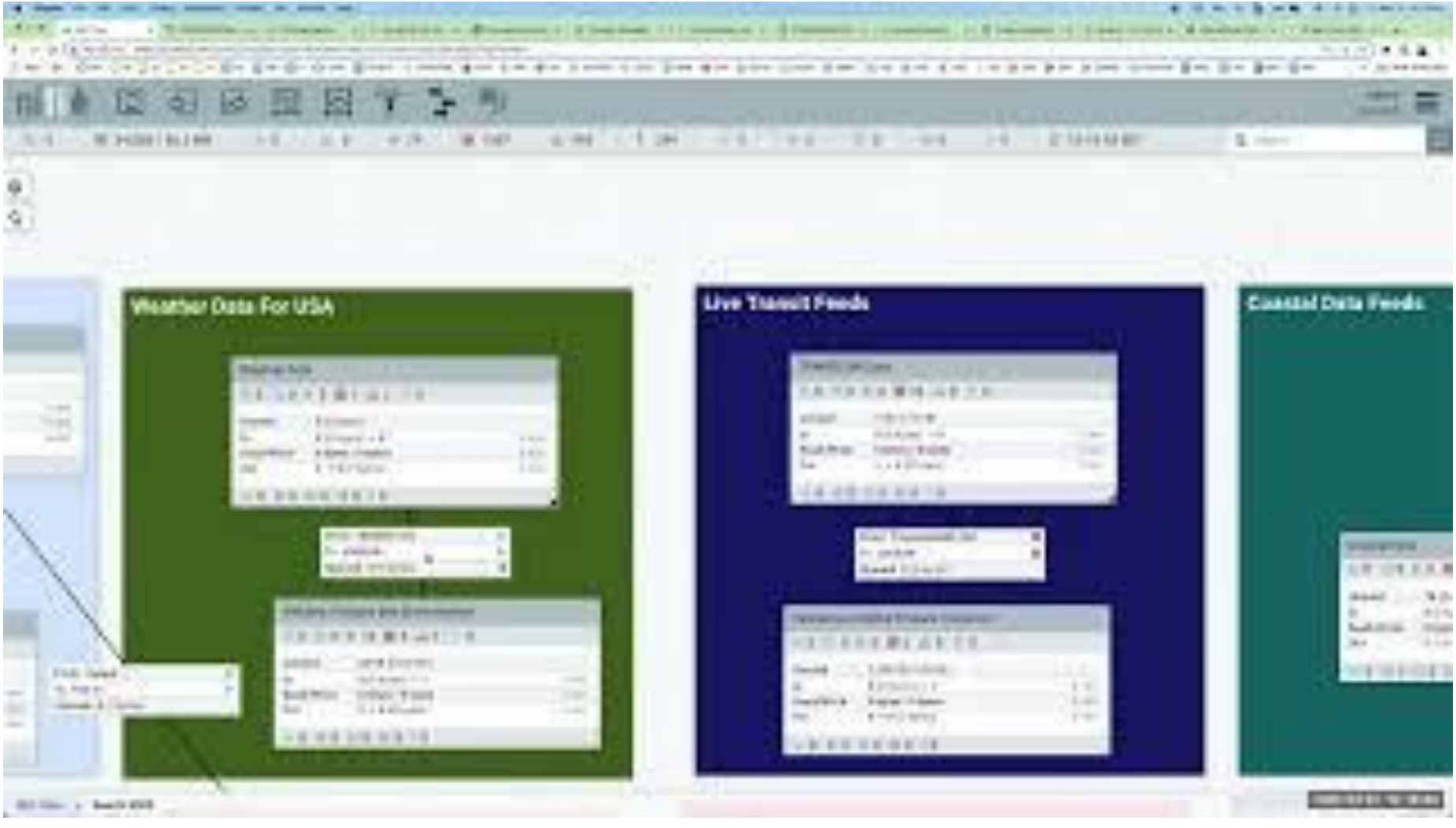
Recreate on Job Start

Ignore NULLs

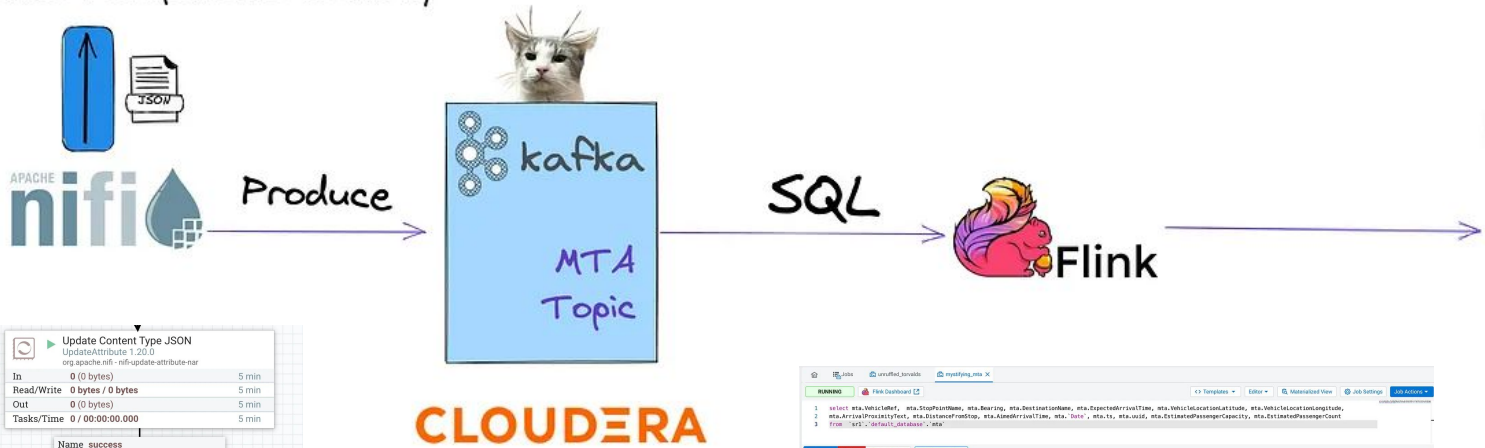
Queries

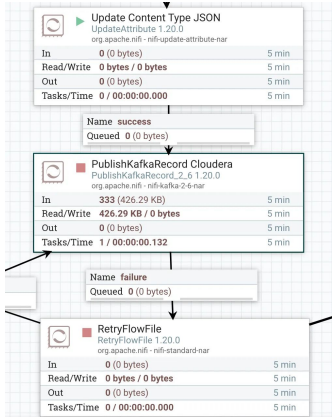
Add New Query

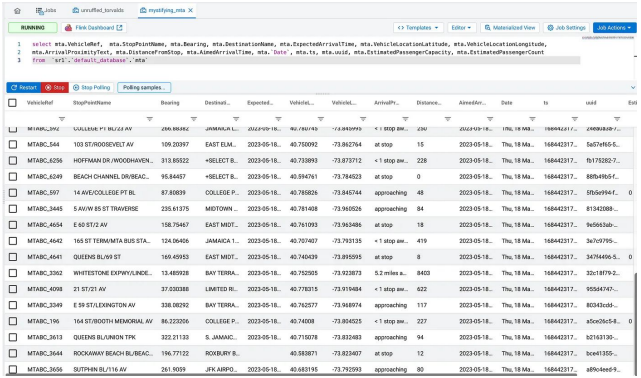
/api/v1/query/5201/travel?key=66ba91a9-507f-422c-bbb4-86250a9f7bb1&limit=100


Weather Data For USA
Live Transit Feeds
Coastal Data Feeds

Metropolitan Transportation Authority



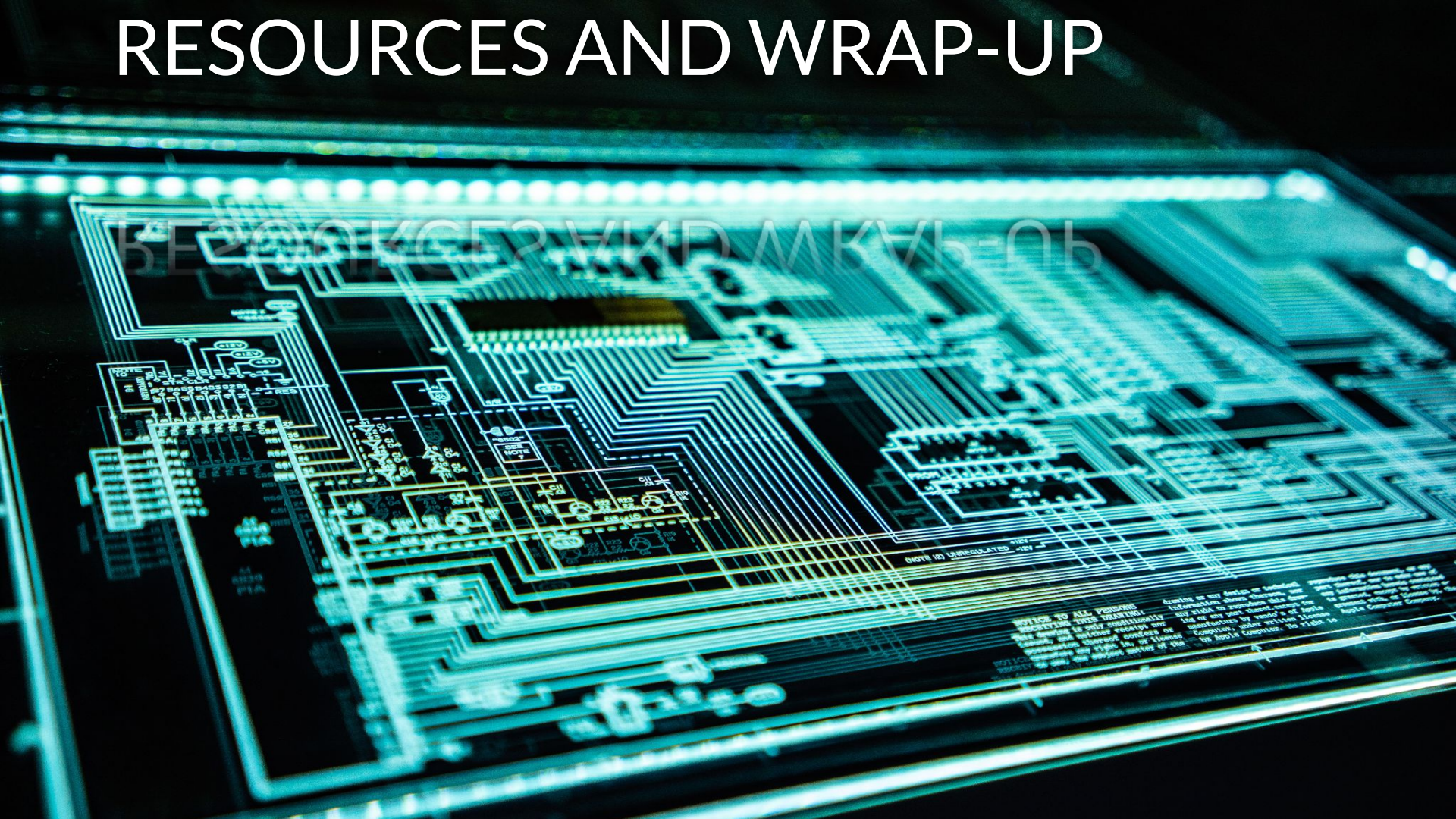




https://medium.com/@tspann/finding-the-best-way-around-7491c76ca4cb

https://github.com/tspannhw/FLaNK-MTA

RESOURCES AND WRAP-UP



Streaming Tech Debt Tips

- Version Control All Assets
- Operationalize with K8
- Use DevOps and APIs
- Latest Java and Python
- Stream Sizing (NiFi, Kafka, Flink)
- Unit and Integration Test
- Backup everything
- Scale in 3s

Streaming Resources

- https://dzone.com/articles/real-time-stream-processing-with-hazelcast-and-streamnative
- https://flipstackweekly.com/
- https://www.datainmotion.dev/
- https://www.flankstack.dev/
- https://github.com/tspannhw
- https://medium.com/@tspann
- https://medium.com/@tspann/predictions-for-streaming-in-2023-ad4d7395d714
- https://www.apachecon.com/acna2022/slides/04_Spann_Tim_Citizen_Streaming_Engineer.pdf

FREE LEARNING ENVIRONMENT

CSP Community Edition

CSP Community Edition

A readily available, dockerized deployment of Apache Kafka and Apache Flink that allows you to test the features and capabilities of Cloudera Stream Processing.

Learn More

- **Gets developers zero to Flink in less than an hour**
  - Experiment with features
  - Develop apps locally
- One docker compose file of CSP which includes:
  - All dependencies required to run
  - Kafka, Kafka Connect and Flink
  - Streams Messaging Manager
  - Schema Registry
  - **SQL Stream Builder Projects**
- Licensed under the Cloudera Community License
- **Unsupported** https://www.cloudera.com/downloads/cdf/csp-community-edition.html
- Community Group Hub (Discussion Forum) for CSP
- Find it on docs.cloudera.com under Applications

Open Source Edition

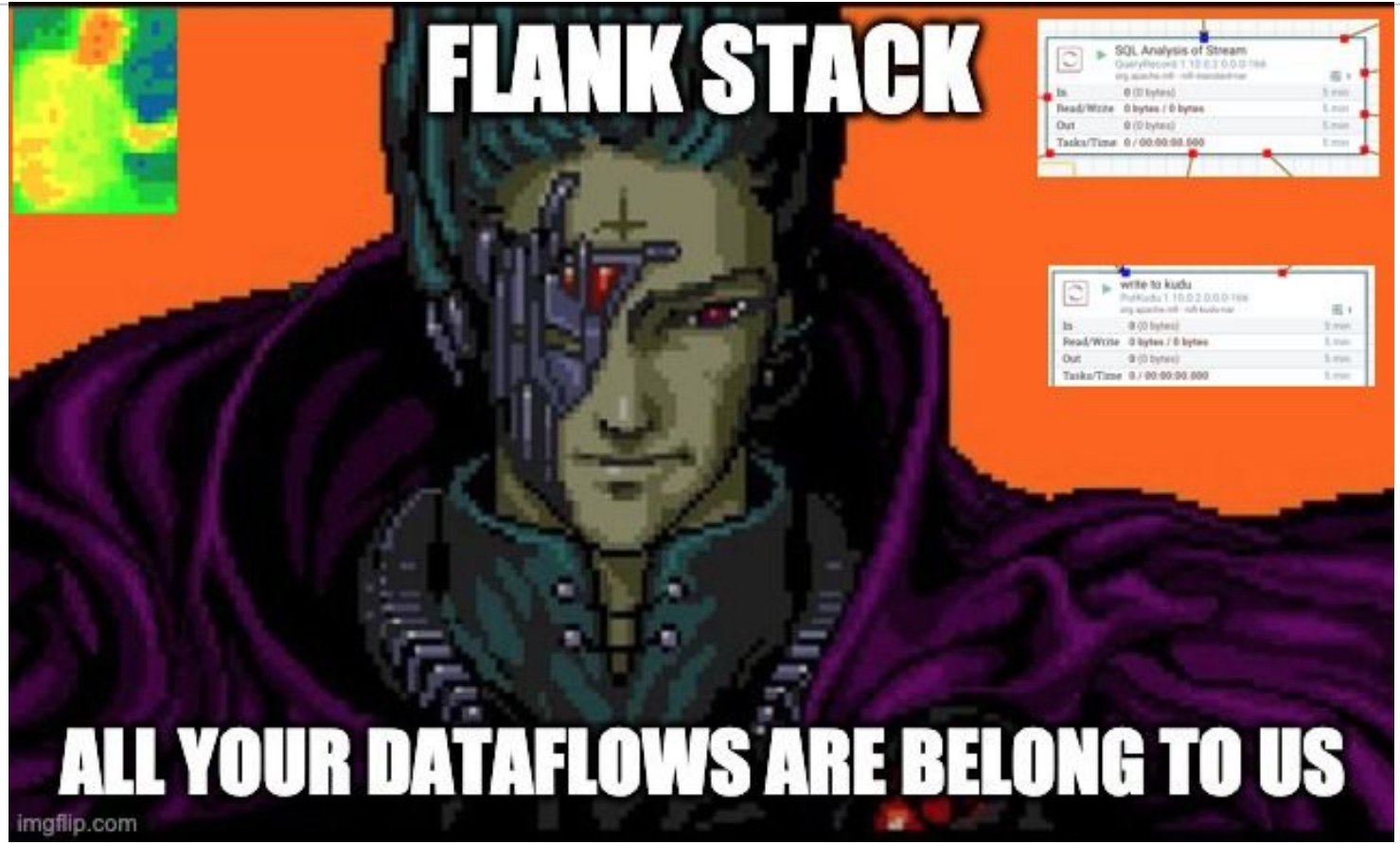
- Apache NiFi in Docker
- Runs in Docker
- Try new features quickly
- Develop applications locally

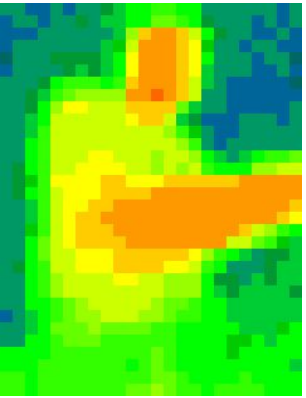
- Docker NiFi
  - `docker run --name nifi -p 8443:8443 -d -e SINGLE_USER_CREDENTIALS_USERNAME=admin -e SINGLE_USER_CREDENTIALS_PASSWORD=ctsBtRBKHRAx69EqUghvvgEvjnaLjFEB apache/nifi:latest`

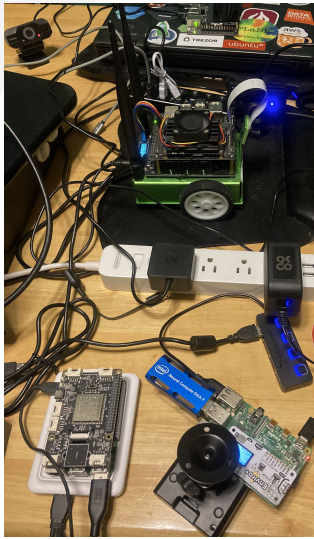
- Licensed under the ASF License
- **Unsupported**

https://hub.docker.com/r/apache/nifi

Resources







THANK YOU