ACTIVATE
VIRTUAL SEARCH & AI CONFERENCE

Powered by Lucidworks

Real-Time Cloud Native Open Source Streaming Of Any Data to Apache Solr

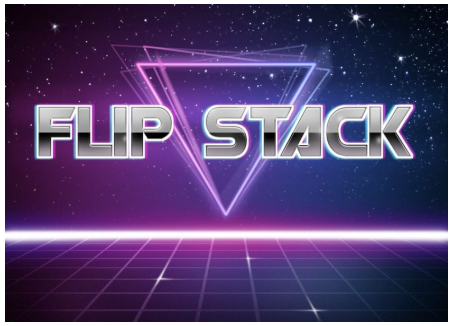
Timothy Spann, Developer Advocate

Speaker Bio

Developer Advocate

DZone Zone Leader and Big Data MVB;
@PaasDev

https://github.com/tspannhw https://www.datainmotion.dev/
https://github.com/tspannhw/SpeakerProfile
https://dev.to/tspannhw
https://sessionize.com/tspann/
https://www.slideshare.net/bunkertor



Stream Native

Agenda

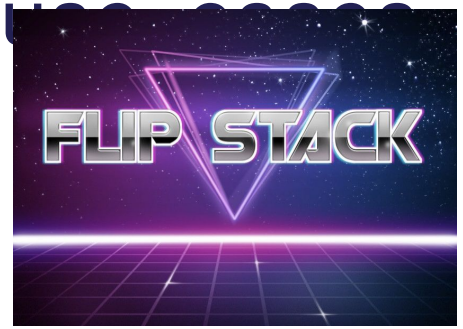
Utilizing Apache Pulsar and Apache NiFi we can parse any document in real-time at scale. We receive a lot of documents via cloud storage, email, social channels and internal document stores. We want to make all the content and metadata to Apache Solr for categorization, full text search, optimization and combination with other datastores. We will not only stream documents, but all REST feeds, logs and IoT data. Once data is produced to Pulsar topics it can instantly be ingested to Solr through Pulsar Solr Sink.

Utilizing a number of open source tools, we have created a real-time scalable any document parsing data flow. We use Apache Tika for Document Processing with real-time language detection, natural language processing with Apache OpenNLP, Sentiment Analysis with Stanford CoreNLP, Spacy and TextBlob. We will walk everyone through creating an open source flow of documents utilizing Apache NiFi as our integration engine. We can convert PDF, Excel and Word to HTML and/or text. We can also extract the text to apply sentiment analysis and NLP categorization to generate additional metadata about our documents. We also will extract and parse images that if they contain text we can extract with TensorFlow and Tesseract.

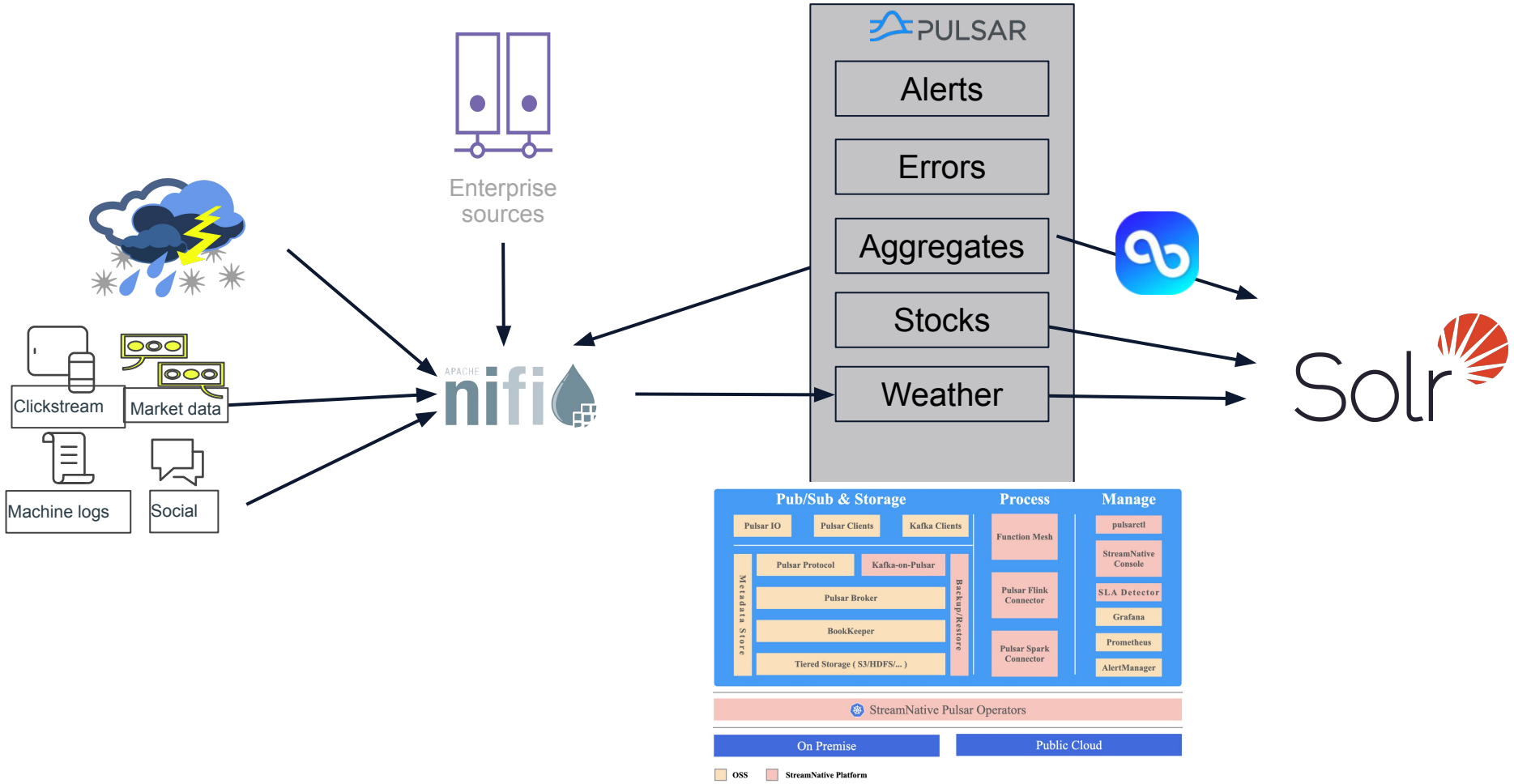
FLiP Stack

- Apache **Fl**ink
- Apache **P**ulsar
- StreamNative's Flink Connector for Pulsar
- Apache +++

Apache projects are the way for all streaming

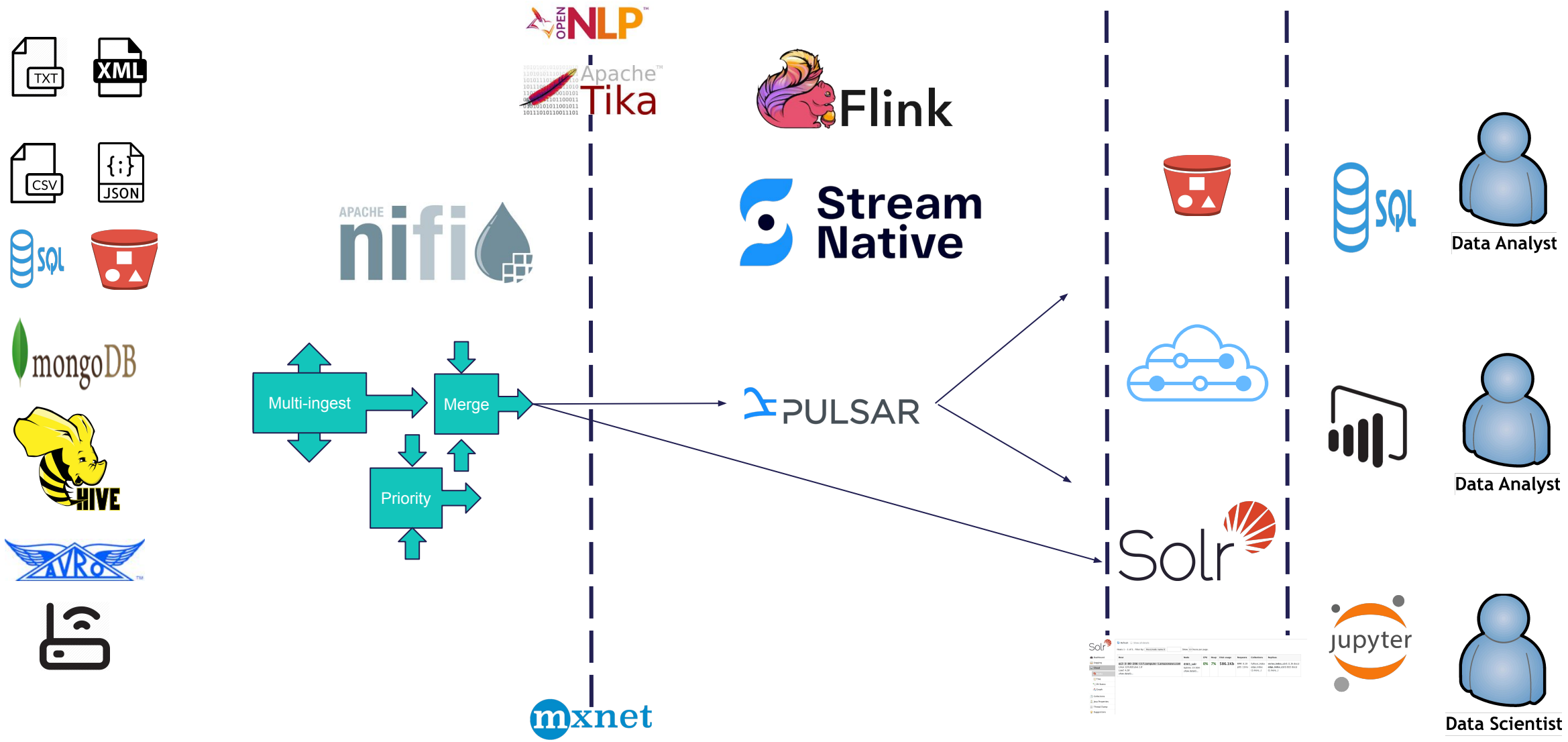


End to End Streaming Demo Pipeline



https://hub.streamnative.io/connectors/solr-sink/2.5.1

All Data - Anytime - Anywhere - Multi-Cloud - Multi-Protocol



ACTIVATE

StreamNative Cloud

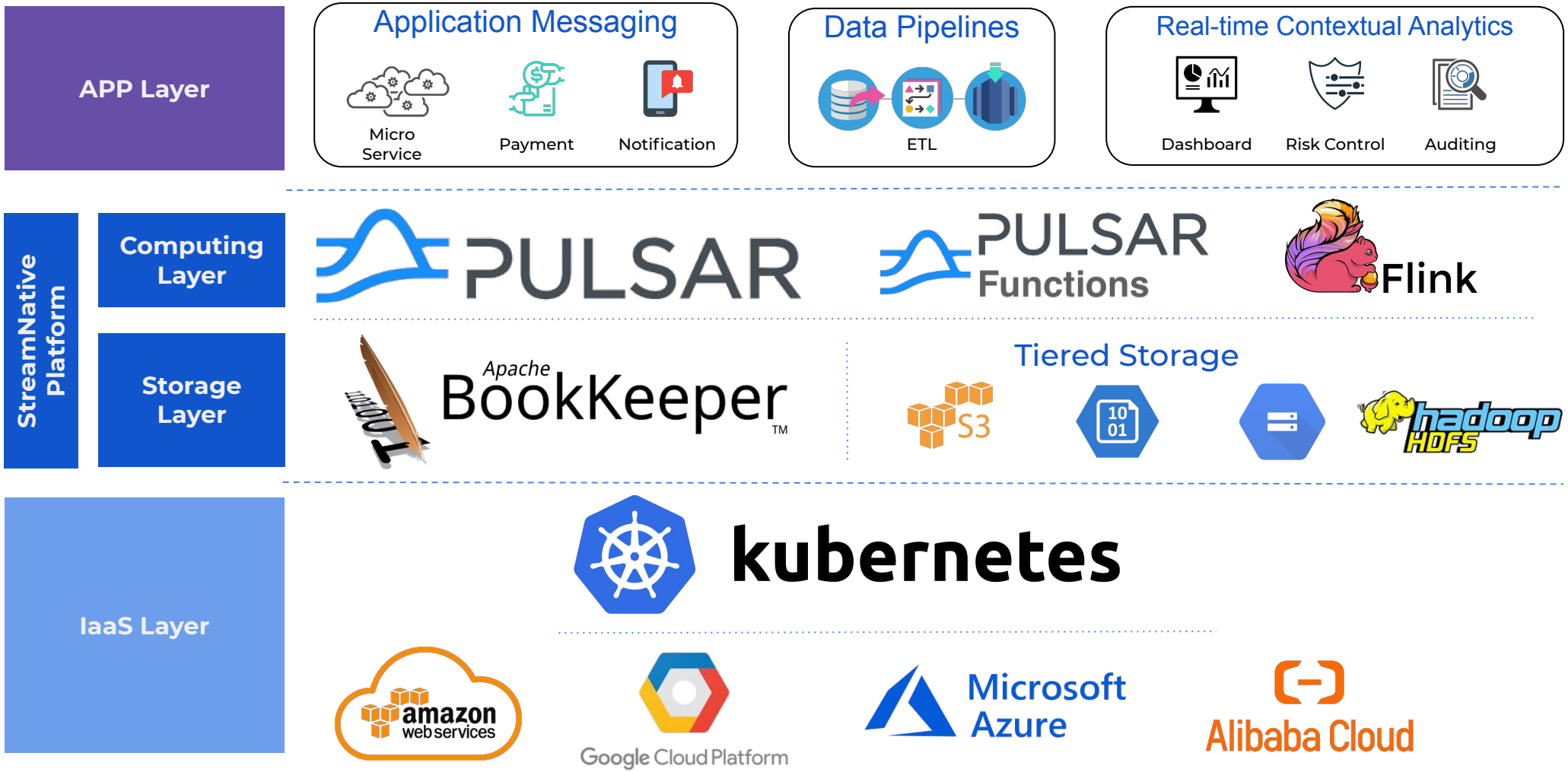
Powered by Apache Pulsar, StreamNative provides a cloud-native, real-time messaging and streaming platform to support multi-cloud and hybrid cloud strategies.

Cloud Native

Built for Containers

Flink SQL

StreamNative Solution



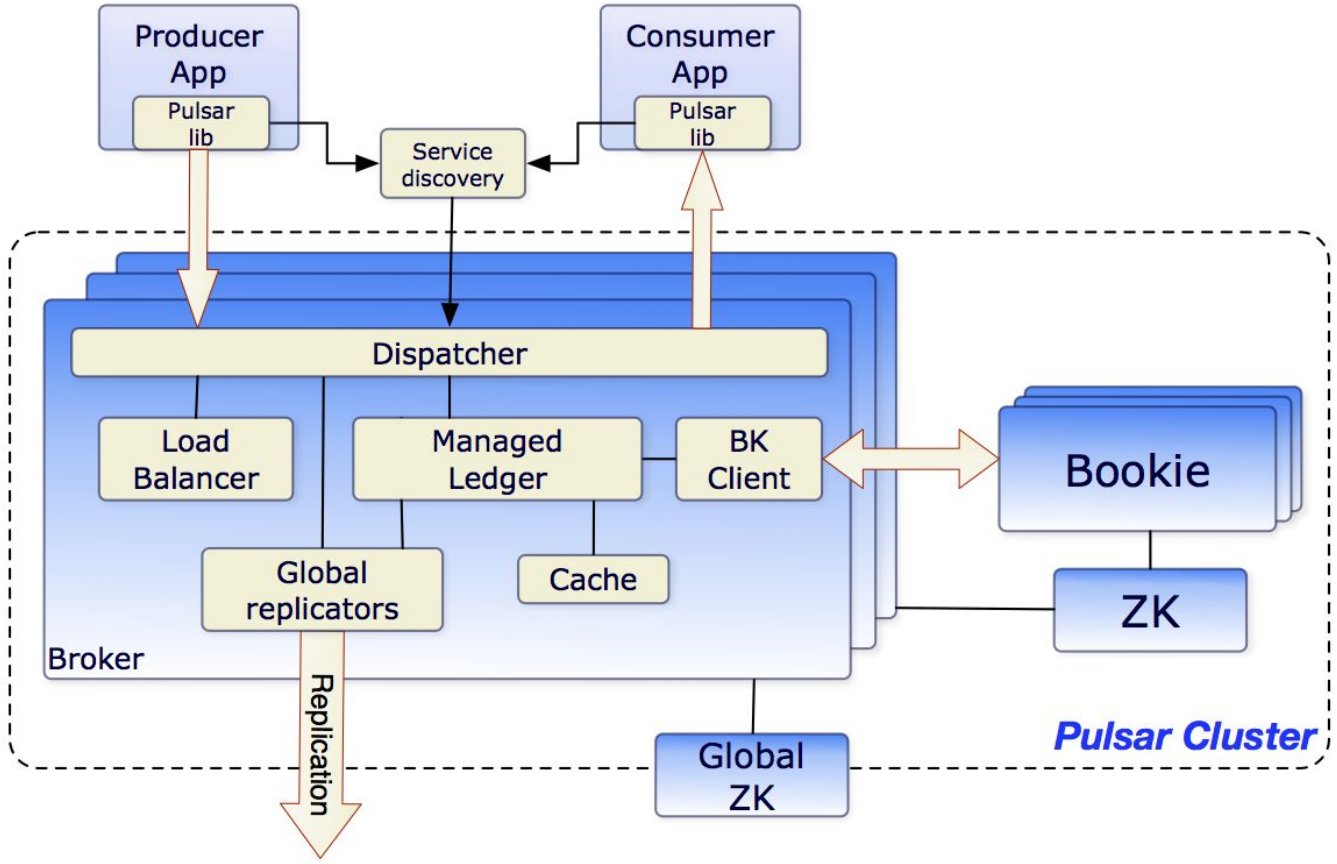
Apache Pulsar

Apache Pulsar is Cloud-Native Messaging and Event-Streaming Platform

Apache Pulsar Overview

Enable Geo-Replicated Messaging

- Pub-Sub
- Geo-Replication
- Pulsar Functions
- Horizontal Scalability
- Multi-tenancy
- Tiered Persistent Storage
- Pulsar Connectors
- REST API
- CLI
- Many clients available
- Four Different Subscription Types
- Multi-Protocol Support
  - MQTT
  - AMQP
  - JMS
  - Kafka
  - ...



What are the Benefits of Pulsar?

Multi-Tenancy

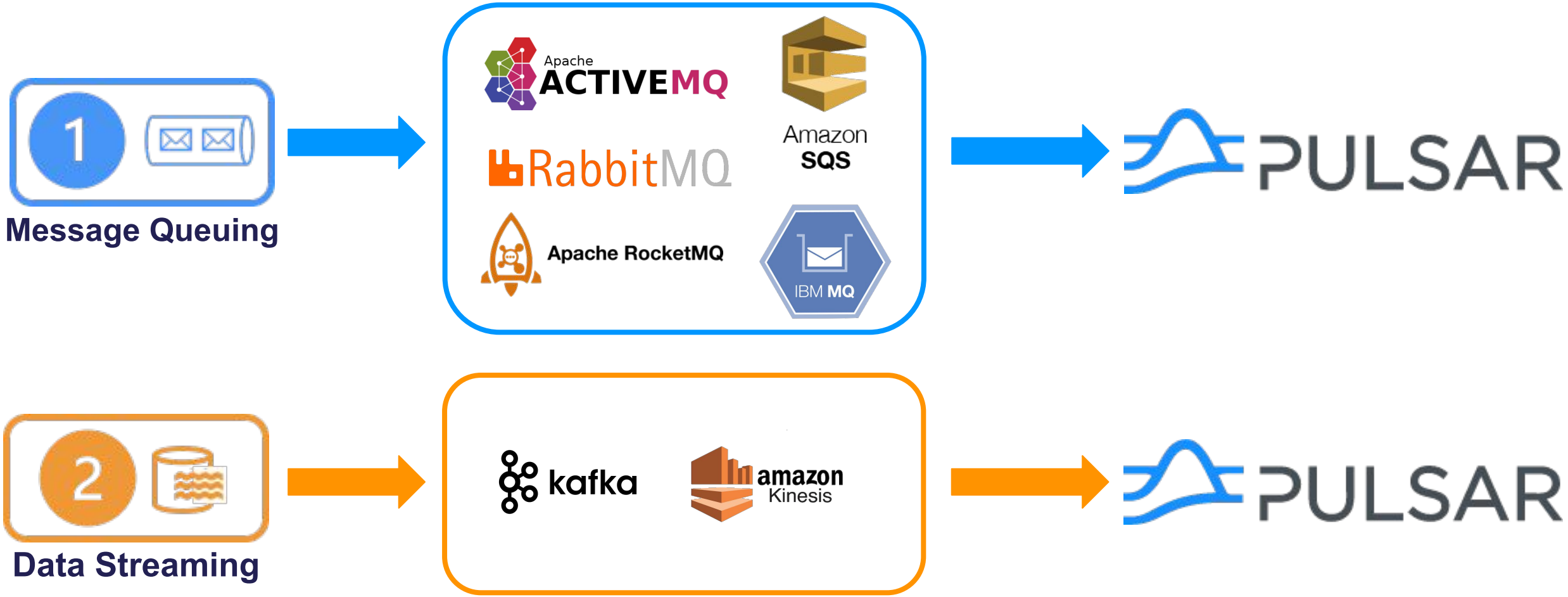
Scalability

Geo-Replication

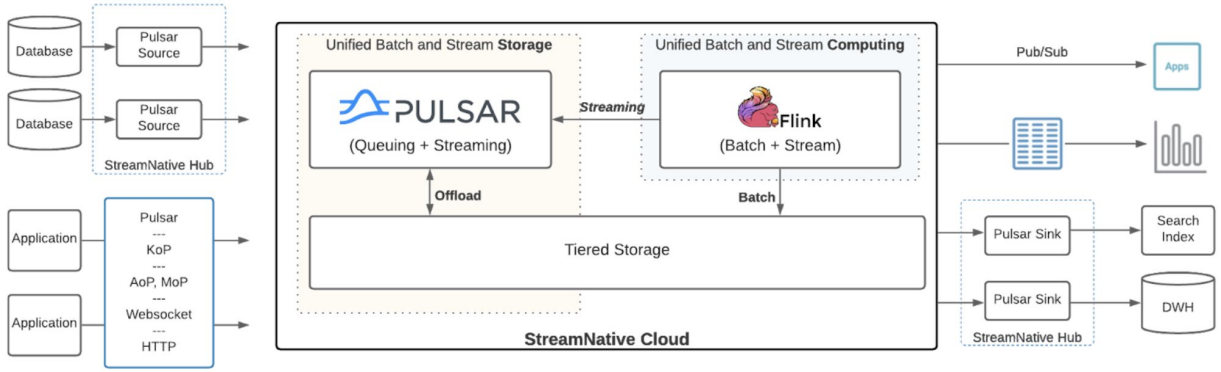
Unified Messaging Model

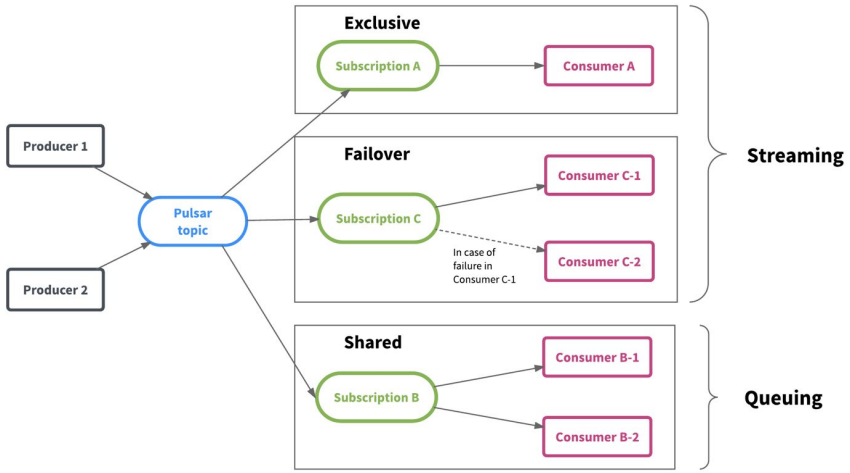
Data Durability

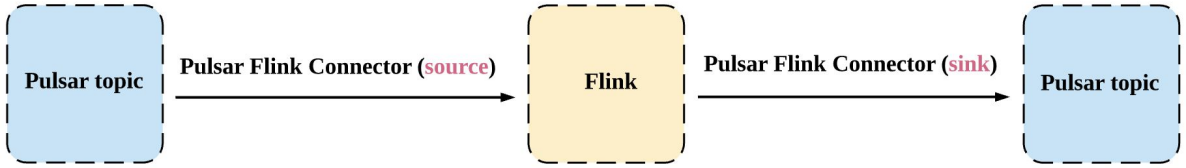
A Unified Messaging Platform



Flink + Pulsar (FLiP)



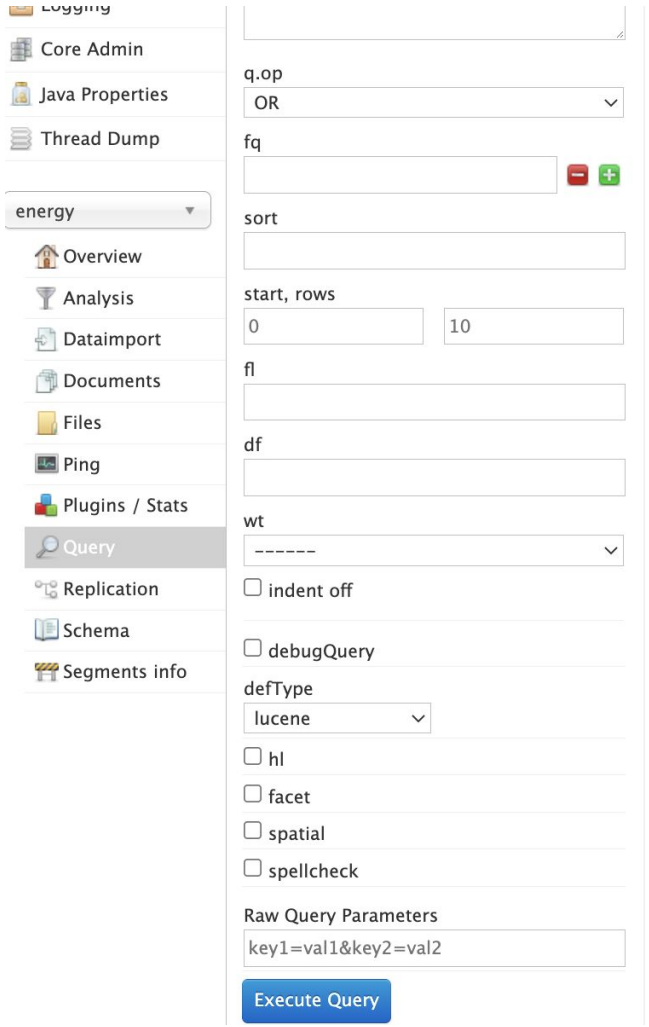




https://flink.apache.org/2019/05/03/pulsar-flink.html
https://github.com/streamnative/pulsar-flink
https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud

Apache Solr

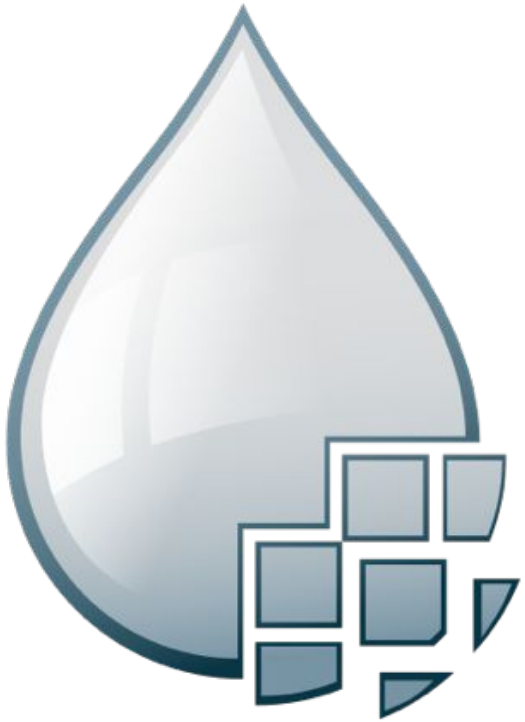
Apache Solr As a Destination




```
"params":{
  "q":"*:*",
  "q.op":"OR",
  "_":"1631799072231"}},
"response":{"numFound":24,"start":0,"numFoundExact":true,"docs":[
  {
    "current":[0.077649],
    "voltage":[117.896479],
    "power":[3.848307],
    "total":[1.251],
    "dev_name":["Wi-Fi Smart Plug With Energy Monitoring"],
    "day1":[0],
    "systemtime":["09/16/2021 09:29:50"],
    "cpu":[3.3],
    "memory":[29.8],
    "diskusage":[32607.4],
    "uuid":["20210916132950_ef6a416c-8675-4332-9f34-401794f08f1e"],
    "macaddress":["1c:bf:ce:1a:7f:a0"],
    "id":"0bdcd768-1868-406d-ae2f-eec27e1ba31f",
    "_version_":1711065259866849280},
  {
    "current":[0.077437],
    "voltage":[117.892421],
    "power":[3.828128],
    "total":[1.252],
    "dev_name":["Wi-Fi Smart Plug With Energy Monitoring"],
    "day1":[0],
    "systemtime":["09/16/2021 09:30:51"],
    "cpu":[3.0],
    "memory":[29.9],
    "diskusage":[32607.3],
    "uuid":["20210916133050_82338f49-7d7f-4c37-b4b0-4ee9042dbc0c"],
    "macaddress":["1c:bf:ce:1a:7f:a0"],
    "id":"b0d0fcb1-e5ed-4ac0-b0c2-499e2b4dab59",
    "_version_":1711065323473469440},
  {
    "current":[0.077039],
```

Apache NiFi

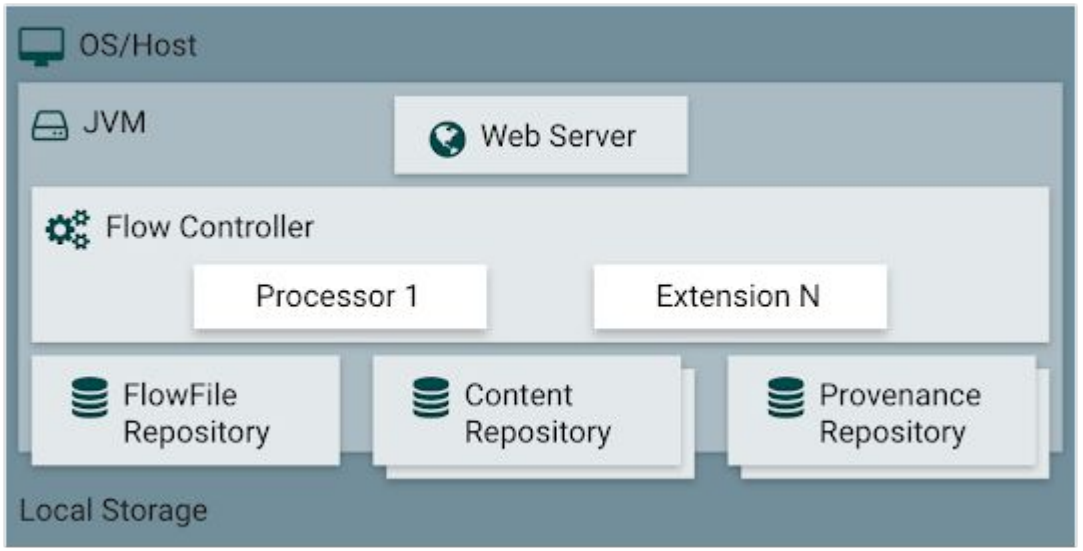
Why Apache NiFi?

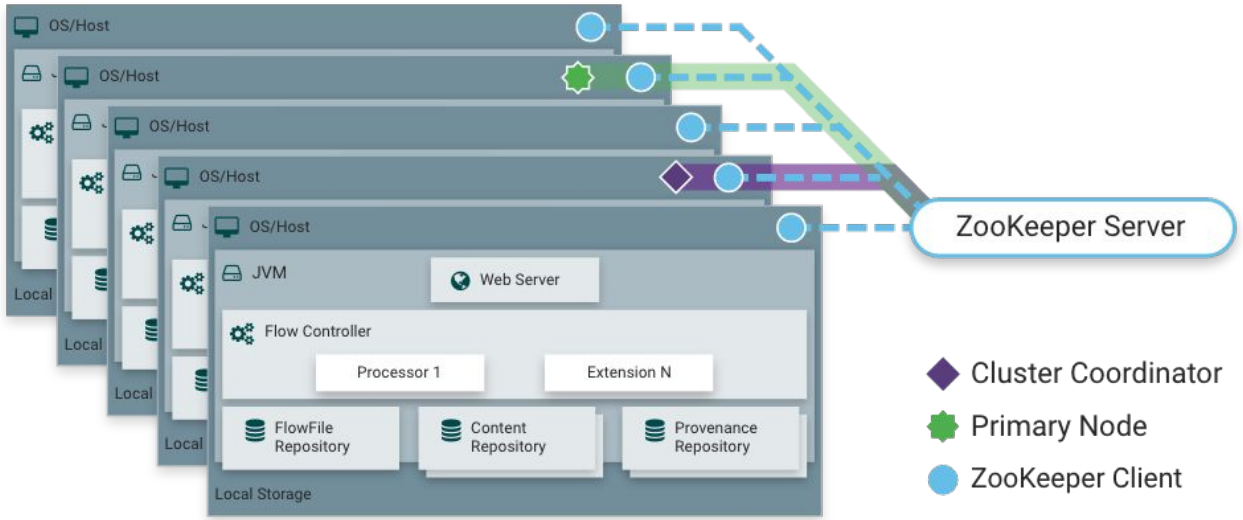


- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models

- Hundreds of processors
- Visual command and control
- Over a sixty sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

Architecture





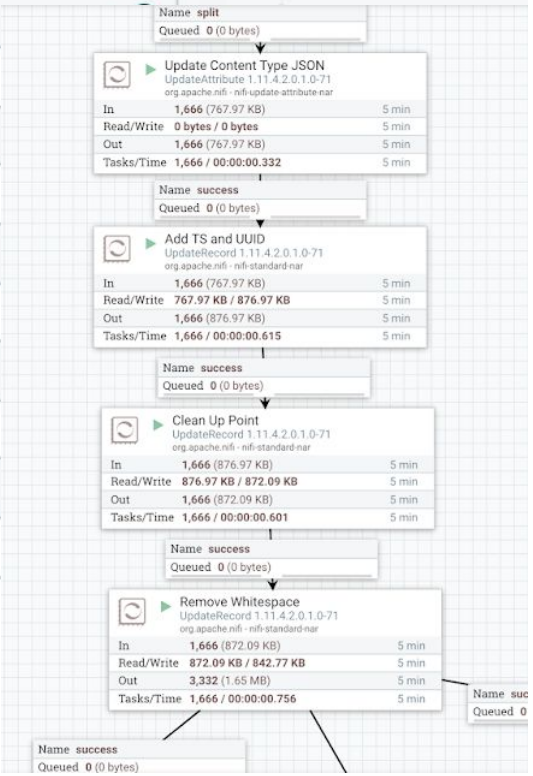
https://nifi.apache.org/docs/nifi-docs/html/overview.html

Record Processors

- XML, CSV, JSON, AVRO and more
- Schemas or Inferred Schemas
- Easily convert between them
- Support SQL with Apache Calcite

Property	Value
Record Reader	**XMLReader**
Record Writer	**JsonRecordSetWriter**
Include Zero Record FlowFiles	**false**
Cache Schema	**true**
query1	SELECT * FROM FLOWFILE

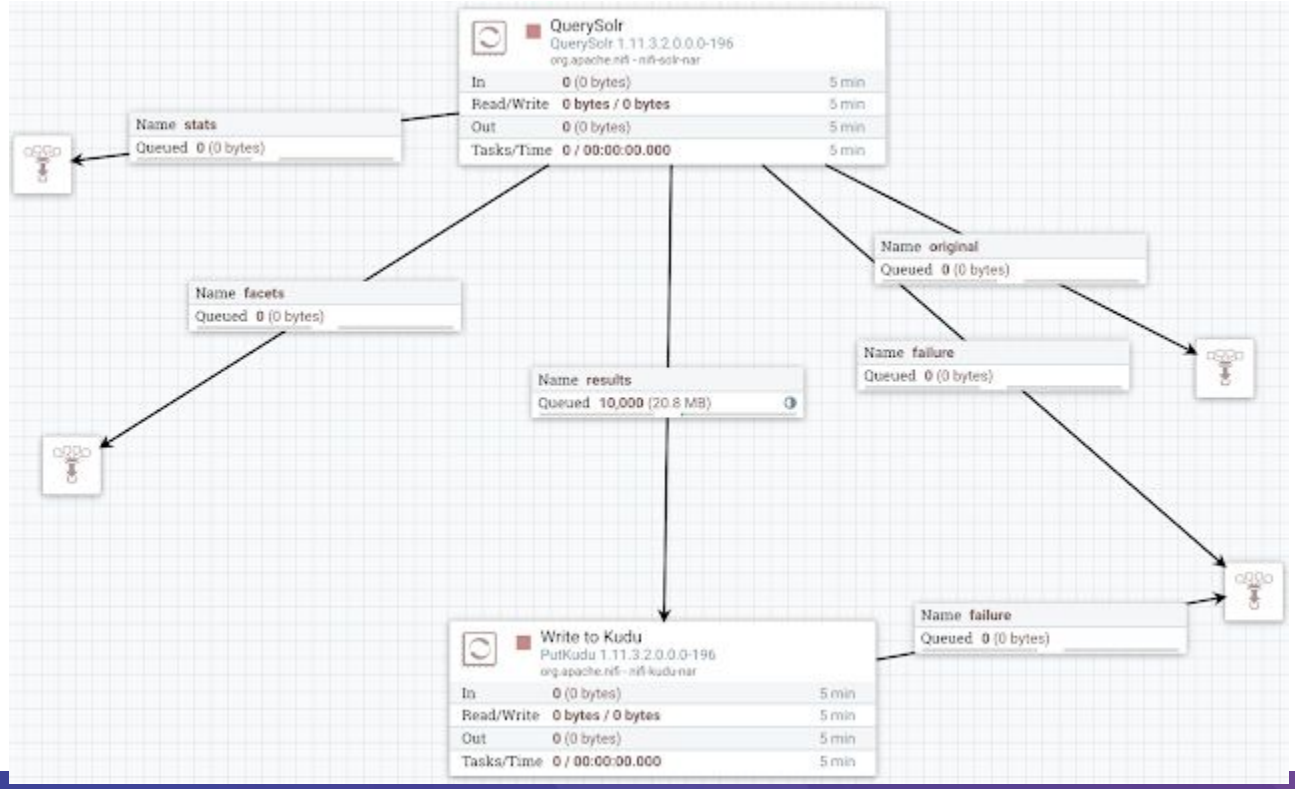
Property	Value
Schema Access Strategy	**Infer Schema**
Schema Registry	AvroSchemaRegistry
Schema Name	${schema.name}
Schema Version	
Schema Branch	
Schema Text	
Schema Inference Cache	
Expect Records as Array	
Attribute Prefix	
Field Name for Content	
Date Format	
Time Format	
Timestamp Format	

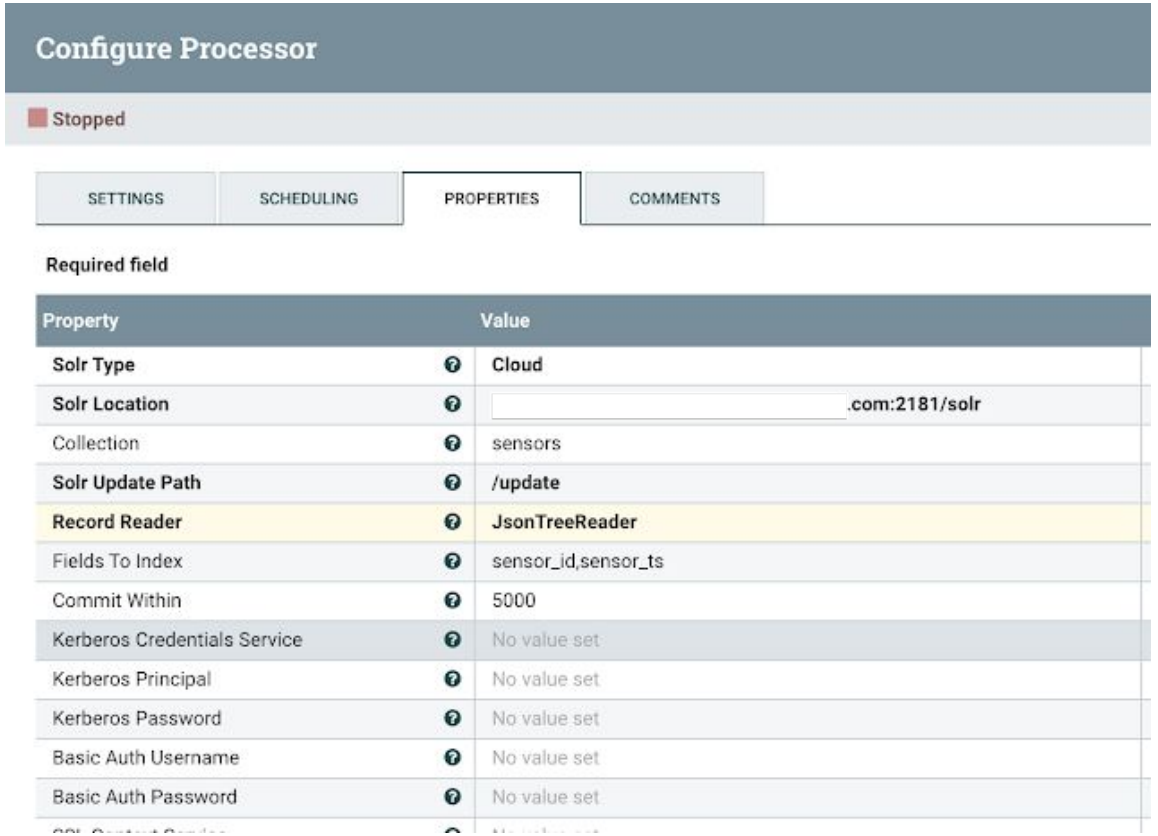


https://www.datainmotion.dev/2019/03/advanced-xml-processing-with-apache.html

SOLR Connectors

- XML, CSV, JSON, AVRO and more
- Schemas or Inferred Schemas
- Use Records or Raw Text
- Support SQL with Apache Calcite



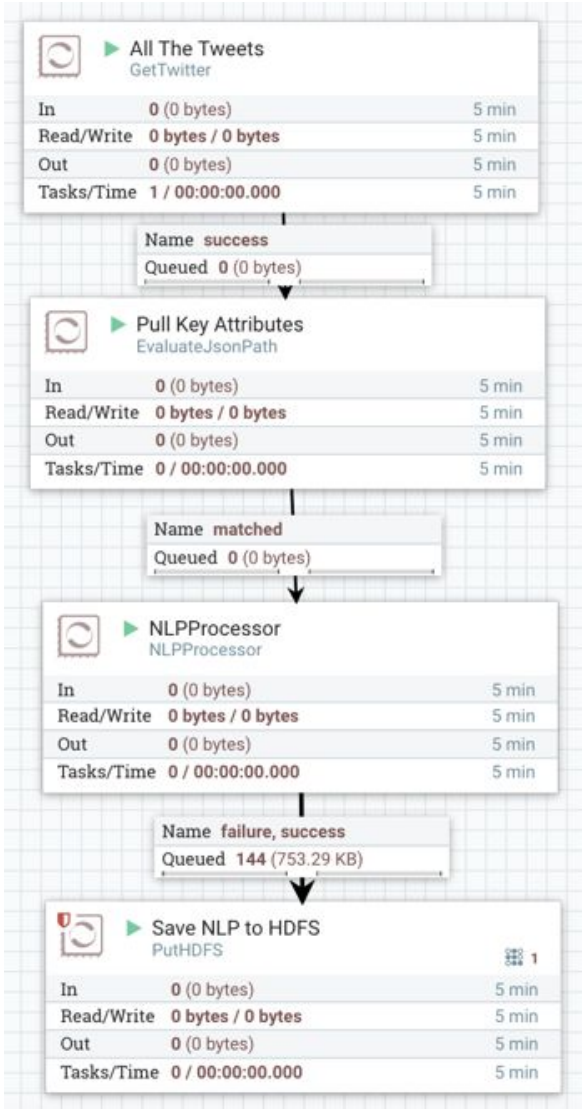


Apache OpenNLP with Apache NiFi

Apache OpenNLP for Entity Resolution Processor
https://github.com/tspannhw/nifi-nlp-processor

Requires installation of NAR and Apache OpenNLP Models (http://opennlp.sourceforge.net/models-1.5/).

This is a non-supported processor that I wrote and put into the community. You can write one too!



https://community.hortonworks.com/articles/80418/open-nlp-example-apache-nifi-processor.html

https://opennlp.apache.org/news/release-190.html

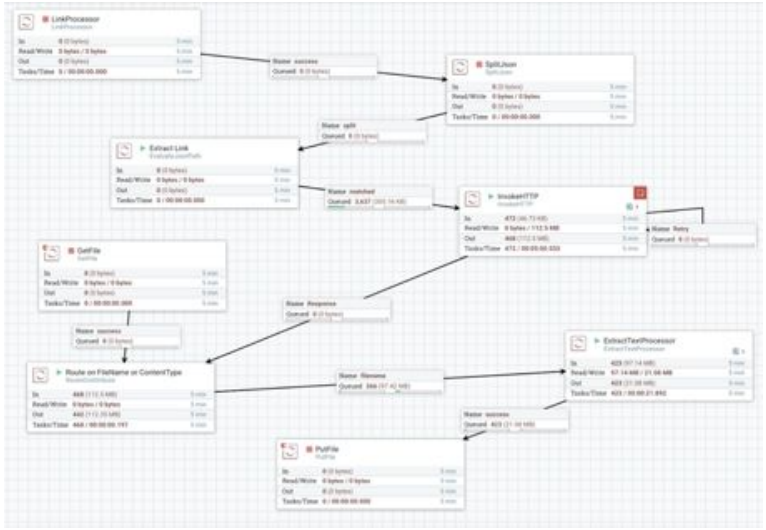
Apache Tika with Apache NiFi

Displaying 1 of 195 extracttextp

Type	Tags
ExtractTextProcessor	extracttextprocessor

Selected Processor:
ExtractTextProcessor

Run Tika Text Extraction from PDF, Word, Excel



View as: original

```
        streaming data. The combination of messaging and processing technologies enables stream
        processing at linear scale.
For example, Apache Storm ships with support for Kafka as a data source using Storm's core
        API or the higher-level, micro-batching Trident API. Storm's Kafka integration also includes
        support for writing data to Kafka, which enables complex data flows between components in a
        Hadoop-based architecture. For more information about Apache Storm, see the Storm User Guide.

Legal notices

    Contents
    Search

    1. Building a High-Throughput Messaging System with Apache Kafka
    2. What's New   1. Apache Kafka
    2. Content Updates


    3. Apache Kafka Concepts
    4. Installing Kafka 1. Prerequisites
    2. Installing Kafka Using Ambari

    5. Configuring Kafka for a Production Environment   1. Customizing Kafka Configuration Settings 1.1. Connection Settings
    1.2. Topic Settings
    1.3. Log Settings
    1.4. Compaction Settings
    1.5. Advanced kafka-env Settings
    1.6. Adding Configuration Properties


    2. Configuring ZooKeeper for Multiple Applications
    3. Enabling Audit to HDFS for a Secure Cluster


    6. Mirroring Data Between Clusters: Using the MirrorMaker Tool  1. Running MirrorMaker
    2. Checking Mirroring Progress
    3. Avoiding Data Loss
    4. Running MirrorMaker on Kerberos-Enabled Clusters


    7. Developing Kafka Producers and Consumers


Search
```

https://github.com/tspannhw/nifi-extracttext-processor

https://community.hortonworks.com/articles/76924/data-processing-pipeline-parsing-pdfs-and-identify.html

https://community.hortonworks.com/articles/81694/extracttext-nifi-custom-processor-powered-by-apach.html

https://community.hortonworks.com/articles/163776/parsing-any-document-with-apache-nifi-15-with-apac.html

https://community.hortonworks.com/content/kbentry/177370/extracting-html-from-pdf-excel-and-word-documents.html

Final Thoughts

Build Your Own Pulsar - SOLR Integration

https://github.com/tspannhw/FLiP-Energy

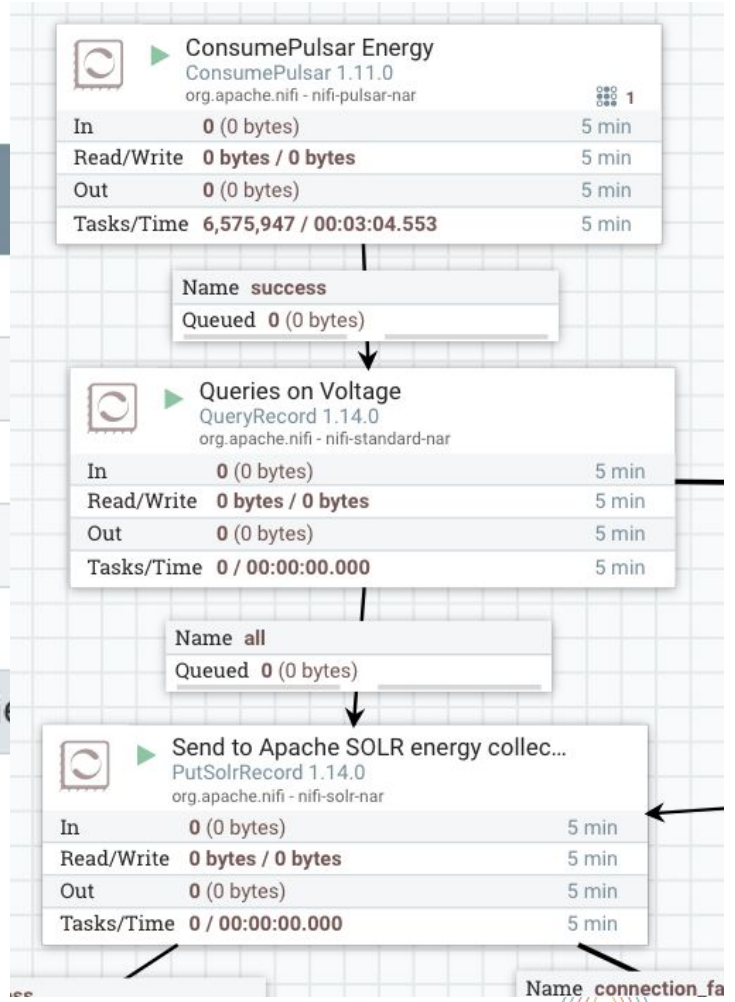

```
bin/pulsar-admin sinks create --tenant public
--namespace default
--name solr-sink-energy
--sink-type solr
--sink-config-file conf/solr-sink-energy.yml
--inputs energy
```

```
}
tspann@Timothys-mbp apache-pulsar-2.8.0 % cat conf/solr-sink.yml
configs:
  solrUrl: "http://localhost:8983/solr"
  solrMode: "Standalone"
  solrCollection: "iot"
```

Build Your Own Pulsar - NiFi Integration

PutSolrRecord

Property	Value
Solr Type	**Standard**
Solr Location	**http://localhost:8983/solr/energy**
Collection	energy
Solr Update Path	**/update**
Record Reader	**CDP Infer JsonTreeReader**
Fields To Index	current,voltage,power,total,diskusage,cpu,systemtime,uuid,m
Commit Within	5000



Connect with the Community & Stay Up-To-Date

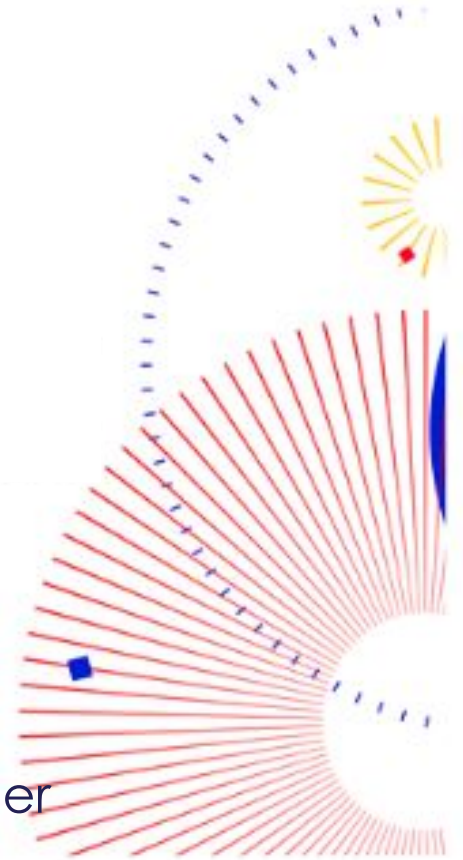
- Join the Pulsar Slack channel - *Apache-Pulsar.slack.com*
- Follow @streamnativeio and @apache_pulsar on Twitter
- Subscribe to Monthly Pulsar Newsletter for major news, events, project updates, and resources in the Pulsar community

Deeper Content

- https://www.datainmotion.dev/2020/04/building-search-indexes-with-apache.html
- https://github.com/tspannhw/nifi-solr-example
- https://github.com/streamnative/pulsar-flink
- https://www.linkedin.com/pulse/2021-schedule-tim-spann/
- https://github.com/tspannhw/SpeakerProfile/blob/main/2021/talks/20210729_HailHydrate!FromStreamtoLake_TimSpann.pdf
- https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud
- https://docs.streamnative.io/cloud/stable/compute/flink-sql
- https://pulsar.apache.org/docs/en/client-libraries-websocket/

 **@PaasDev**  **timothyspann**

https://www.pulsardeveloper.com/



Pulsar Summit Asia

November 20-21, 2021

Contact us at partners@pulsar-summit.org to become a sponsor or partner

Announcing

Flink SQL on StreamNative Cloud

Thank You


ACTIVATE
VIRTUAL SEARCH & AI CONFERENCE
Powered by Lucidworks