CLOUDERA

Introduction to Apache NiFi 1.10

Timothy Spann
Field Engineer, Data in Motion
Cloudera
@**PaasDev**

Welcome to Future of Data - Princeton

FUTURE OF DATA

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 ...

APACHE HBASE

STREAMS MESSAGING MANAGER kafka

@PaasDev

Today's Lead

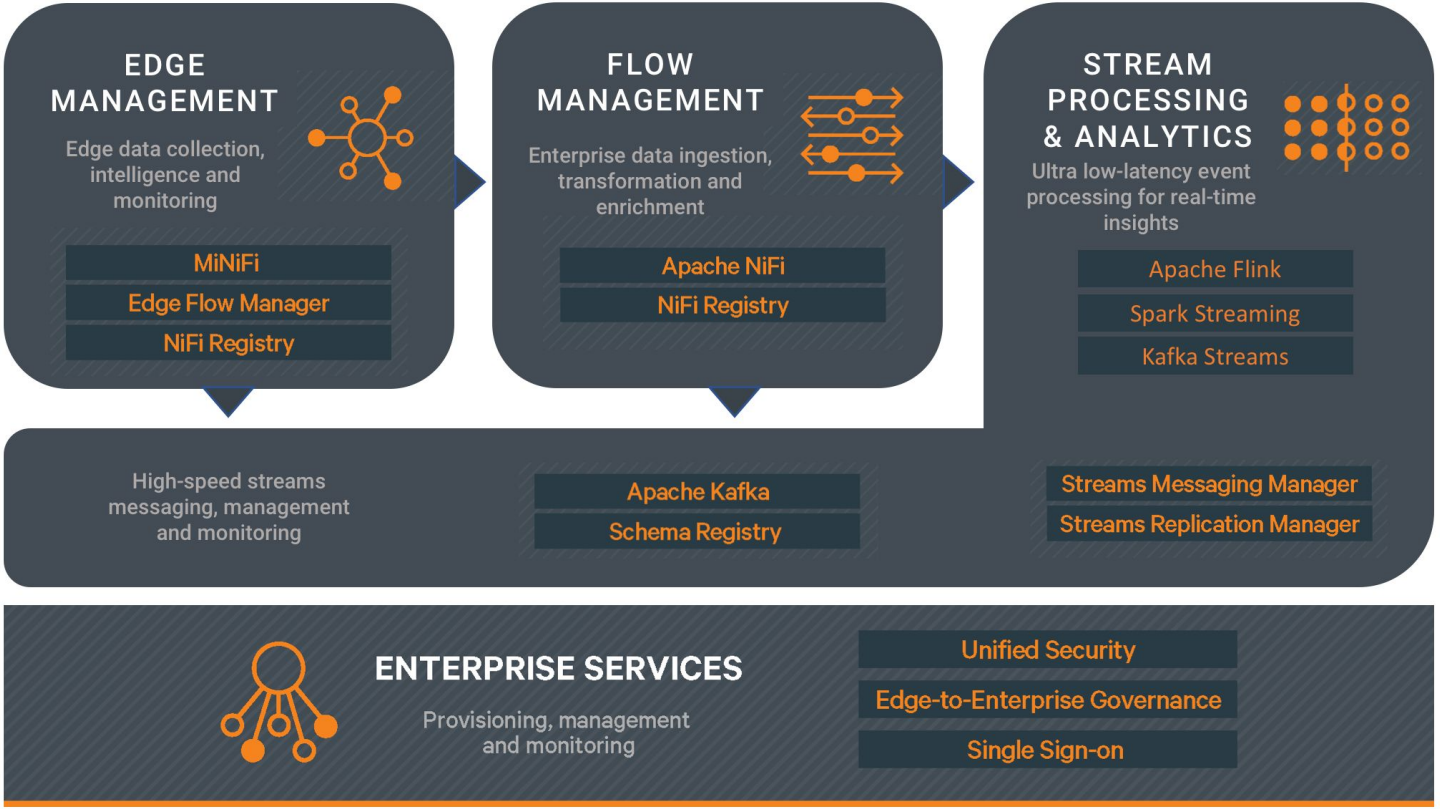
Who am I?

Data in Motion Field Engineer

DZone Zone Leader and Big Data MVB;
Princeton NJ Future of Data Meetup;
ex-Pivotal Field Engineer;
Author of Apache Kafka RefCard
https://github.com/tspannhw
https://www.datainmotion.dev/

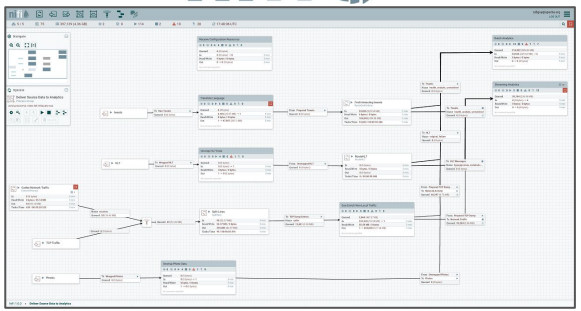
@PaasDev

Cloudera DataFlow (CDF) - Data-in-Motion Platform



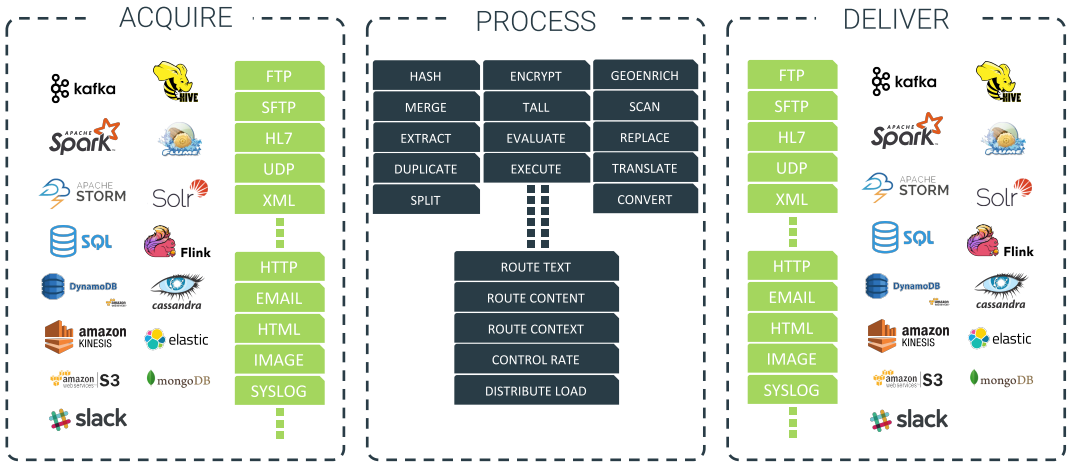
Cloudera Flow Management

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



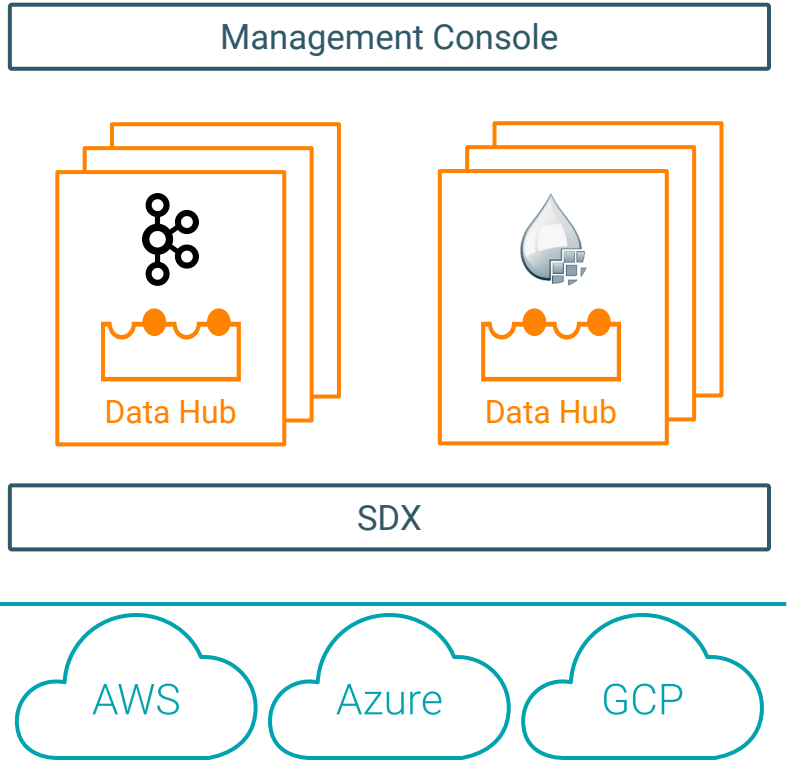
APACHE NIFI registry

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



- Over 300 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

NiFi and Kafka Cluster Definitions in CDP Data Hub



CDP Management Console runs as a web service hosted and managed by Cloudera

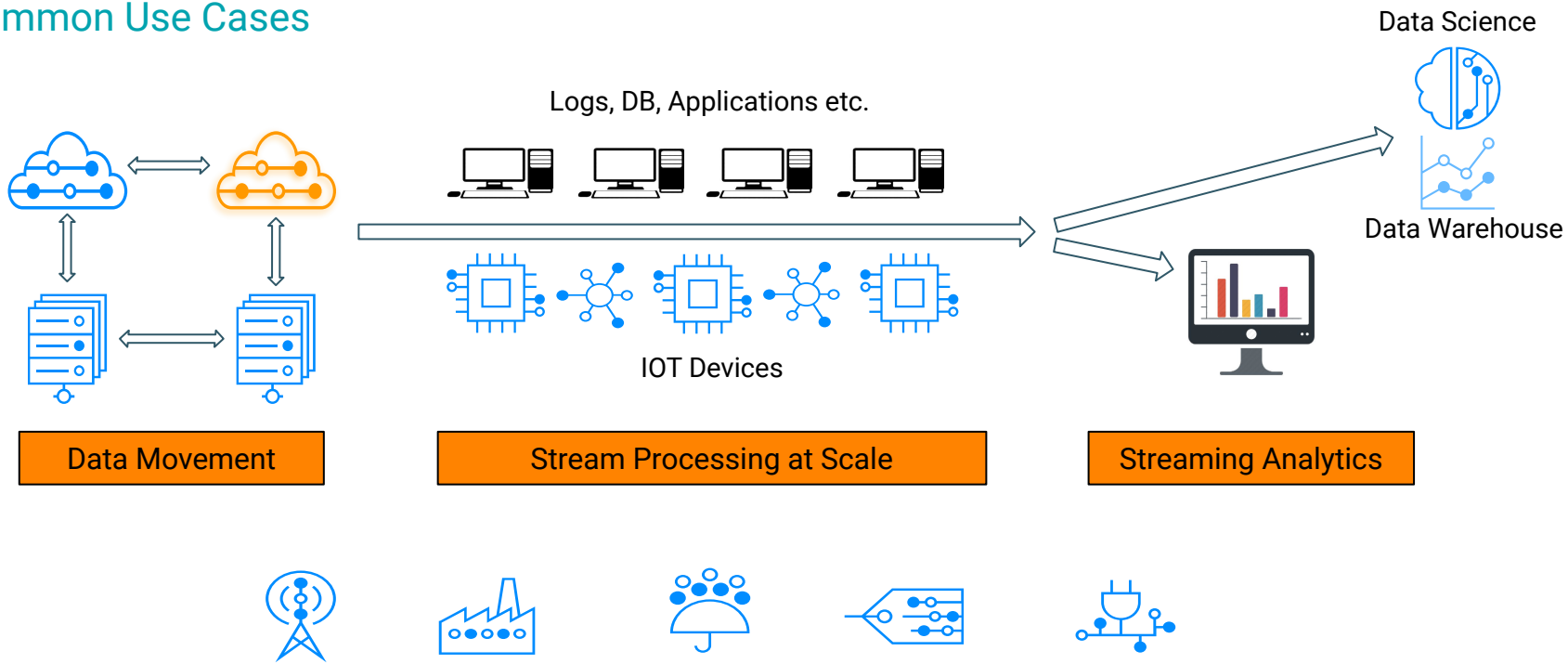
Data Hub clusters running Kafka and NiFi hosted in the customer's cloud environment, but managed by the CDP Management Console

Shared Data Experience (SDX) technologies form a secure and governed data lake backed by object storage (S3, ADLS, GCS)

CDP services are optimized for the elastic compute & 'always-on' storage services provided by the customer's chosen cloud provider

Cloudera Data Flow

Common Use Cases



NIFI 1.10 New Features

Stateless Engine

- Granular containers per flow
- Flows From NiFi Registry

bin/nifi.sh **stateless RunFromRegistry Continuous** --file kafka.json

```
Tspann-MBP15-HW14277:nifi-1.10.0-SNAPSHOT tspann$ bin/nifi.sh stateless
nifi.sh: JAVA_HOME not set; results may vary

Note: Use of this command is considered experimental. The commands and approach used may change from time to time.

Java home (JAVA_HOME):
NiFi home (NIFI_HOME): /Users/tspann/Documents/nifi-1.10.0-SNAPSHOT
Java options (STATELESS_JAVA_OPTS): -Xms1024m -Xmx1024m

14:20:09.696 [main] INFO org.apache.nifi.StatelessNiFi - Unpacking 99 NARs
14:20:37.031 [main] INFO org.apache.nifi.StatelessNiFi - Finished unpacking 99 NARs in 27330 millis
Usage:
       RunFromRegistry [Once|Continuous] --json <JSON>
       RunFromRegistry [Once|Continuous] --file <File Name>

       RunYARNServiceFromRegistry        <YARN RM URL> <Docker Image Name> <Service Name> <# of Containers> --json <JSON>
       RunYARNServiceFromRegistry        <YARN RM URL> <Docker Image Name> <Service Name> <# of Containers> --file <File Name>

    3) RunOpenwhiskActionServer          <Port>

Examples:
    1) RunFromRegistry Once --json "{\"registryUrl\":\"http://172.26.198.107:61080\",\"bucketId\":\"5eec8794-01b3-4cd7-8536-0167c8b4ce8c\",\"flowId\":
\"c5fa1d4f-b453-4bf8-8ff3-362352c418f3\"}"
    2) RunYARNServiceFromRegistry http://127.0.0.1:8088 nifi-stateless:latest kafka-to-solr 3 --file kafka-to-solr.json
    3) RunOpenwhiskActionServer 8080

Notes:
    1) The configuration file must be in JSON format.
    2) When providing configurations via JSON, the following attributes must be provided: registryUrl, bucketId, flowId.
          All other attributes will be passed to the flow using the variable registry interface
```

https://github.com/apache/nifi/blob/ea1becac4fc519c54b8b4d21773e68f8da364755/nifi-nar-bundles/nifi-framework-bundle/nifi-framework/nifi-stateless/README.md

https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Stateless Engine

- See also Parameters
- Docker
- YARN
- Kubernetes (K8)
- Stateful NiFi clusters
- Apache OpenWhisk (FaaS)

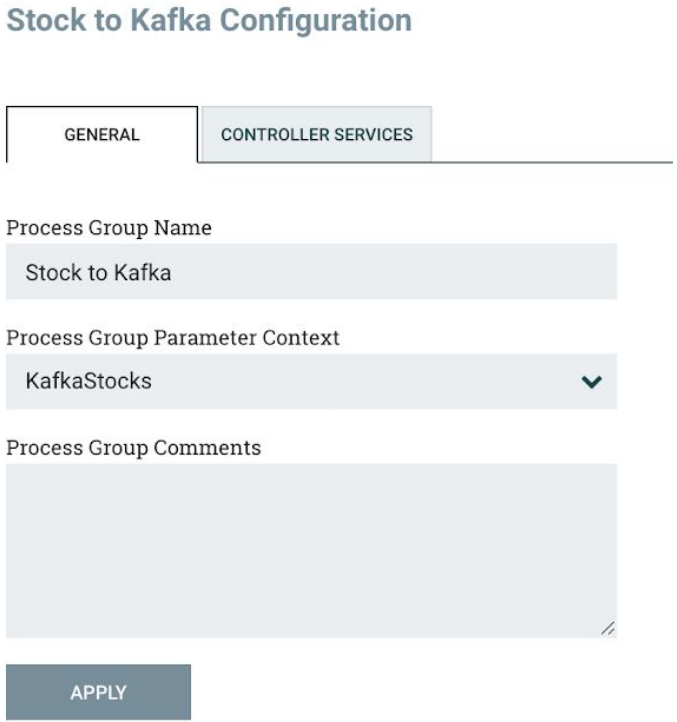

```
{"registryUrl": "http://tspann-mbp15-hw14277:18080",
 "bucketId": "140b30f0-5a47-4747-9021-19d4fde7f993",
 "flowId": "0540e1fd-c7ca-46fb-9296-e37632021945",
 "ssl": {
  "keystoreFile": ""keystorePass": ""keyPass": ""keystoreType": "",
  "truststoreFile":
"/Library/Java/JavaVirtualMachines/amazon-corretto-11.jdk/Contents/Home/lib/security/cacerts",
  "truststorePass": "changeit", "truststoreType": "JKS"
 },
 "parameters": {
  "broker" : "4.317.852.100:9092",
  "topic" : "iot",
  "group_id" : "nifi-stateless-kafka-consumer",
  "DestinationDirectory" : "/tmp/nifistateless/output2/",
  "output_dir": "/Users/tspann/Documents/nifi-1.10.0-SNAPSHOT/logs/output"
 }
}
```

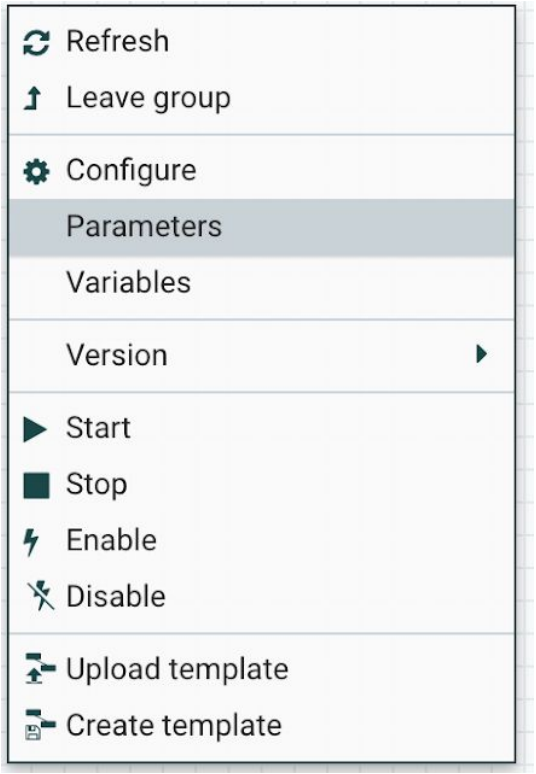

https://github.com/tspannhw/stateless-examples

https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Parameters

- Parameters
- Parameter Context

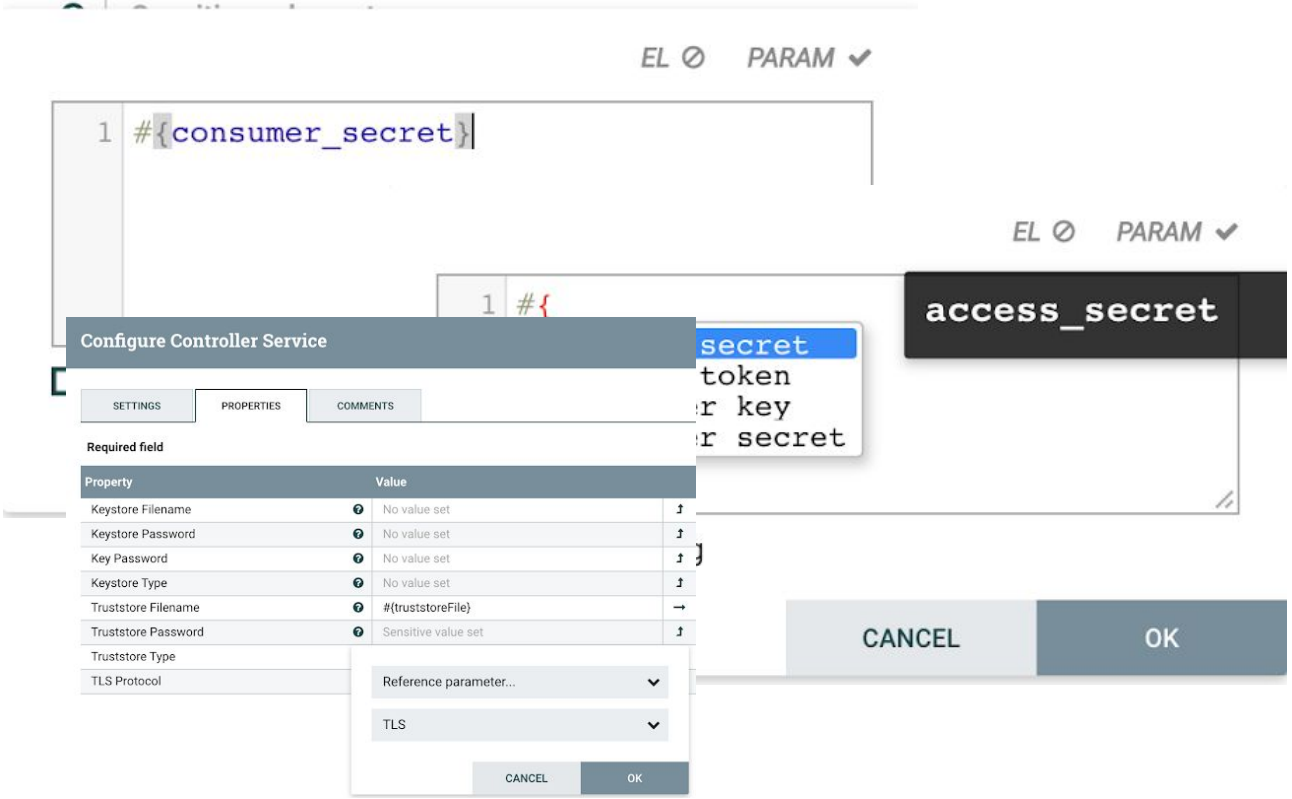




https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Parameters

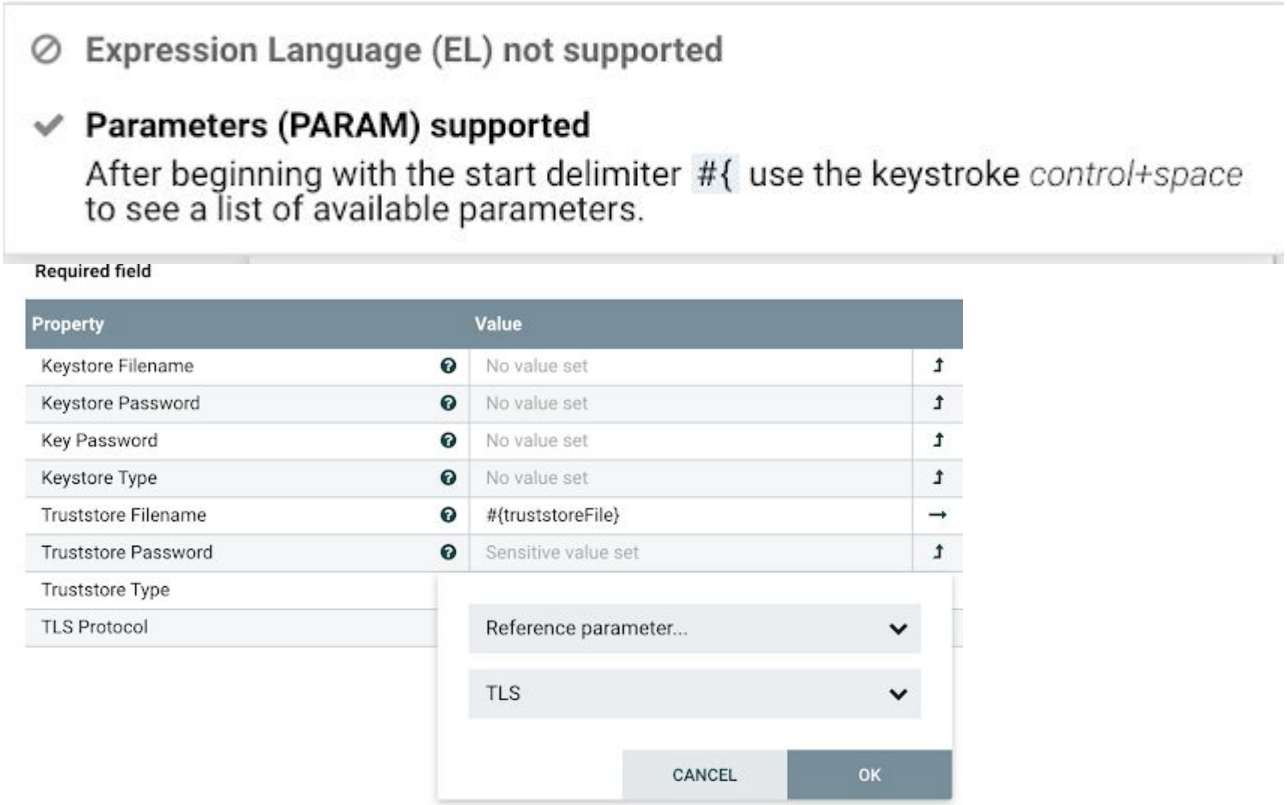
- Advanced Editors
- Easy to Use
- PARAM



https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Parameters

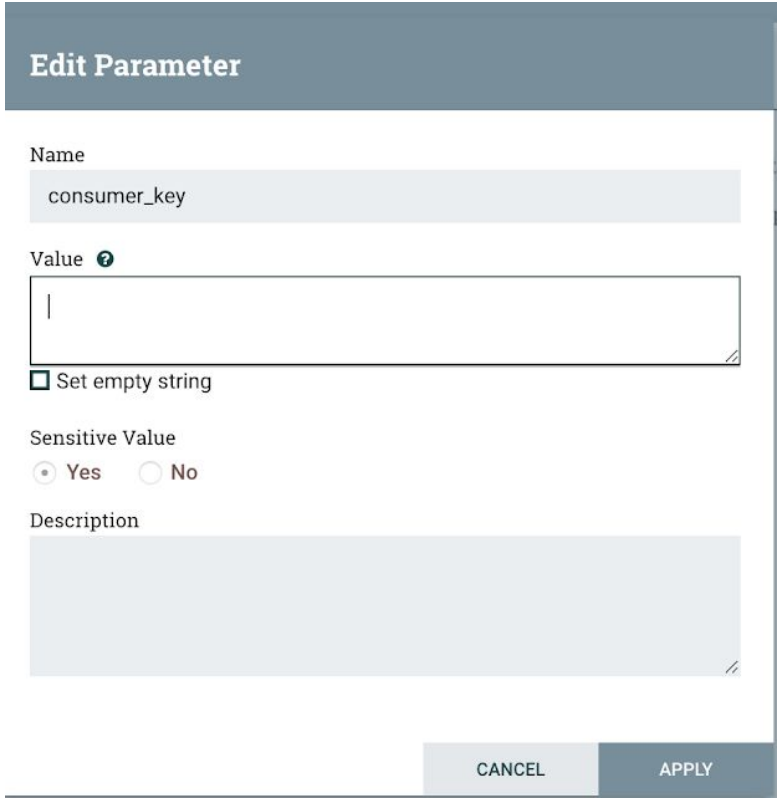
- Configure Externally with JSON Files to Execute Stateless Flows



https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Parameters

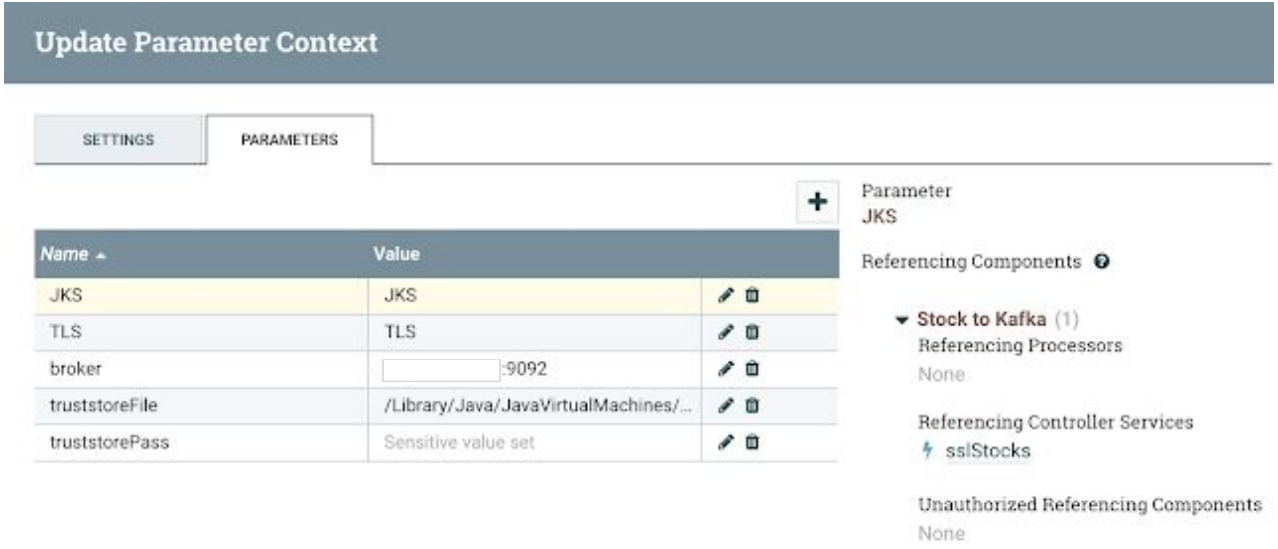
- Create / Edit Parameters from NiFi or in JSON Files



https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Parameter Context

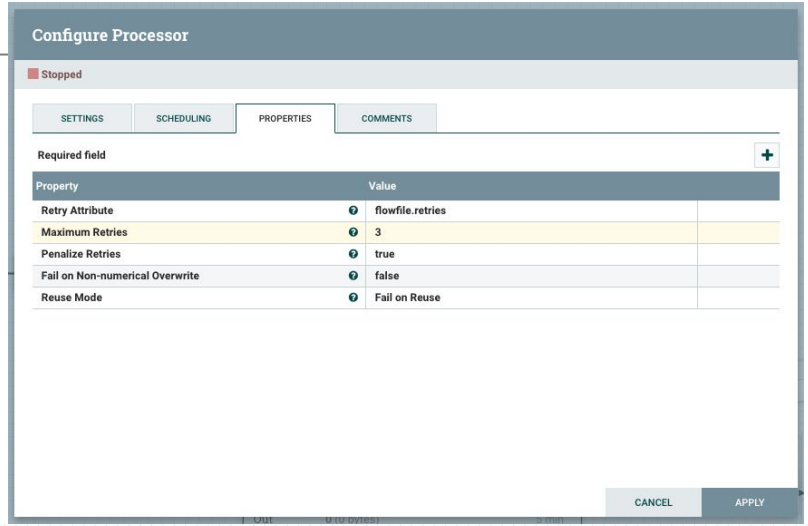
- Sensitive or Normal
- Connect to Multiple Process Groups

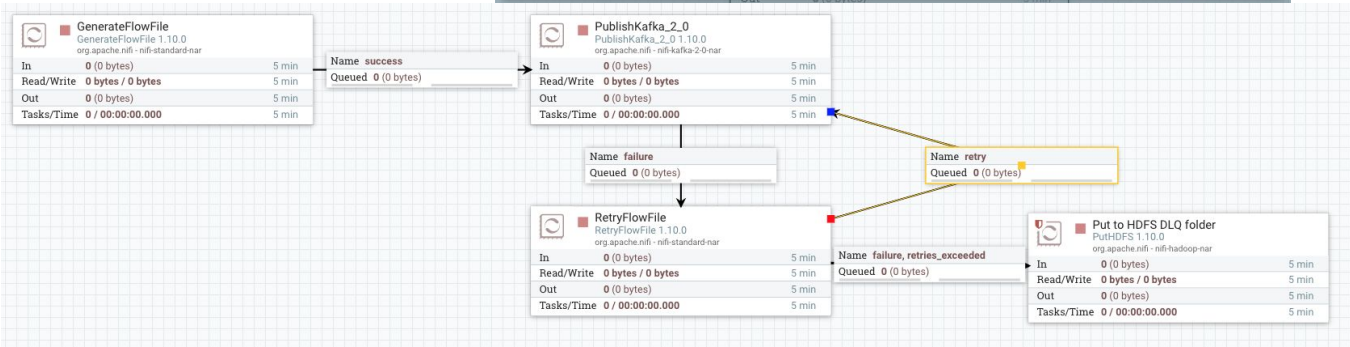


https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

RetryFlowFile

- Configurable Retries
- Maximum #
- Penalties
- When to Fail
- Reuse Mode

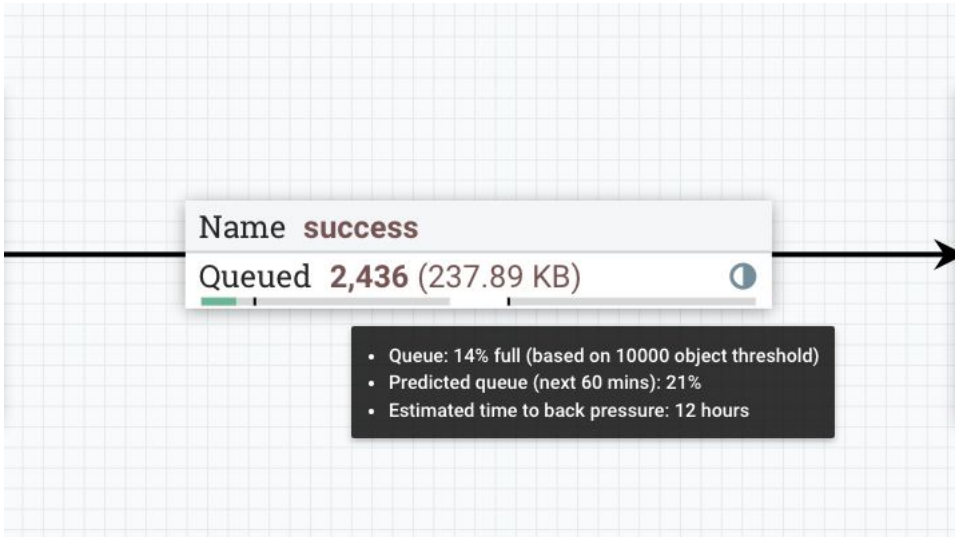




https://medium.com/@abdelkrim.hadjidj/apache-nifi-1-10-series-simplifying-error-handling-7de86f130acd

BackPressure Prediction

- OrdinaryLeastSquares
- SimpleRegression
- Enable analytics feature



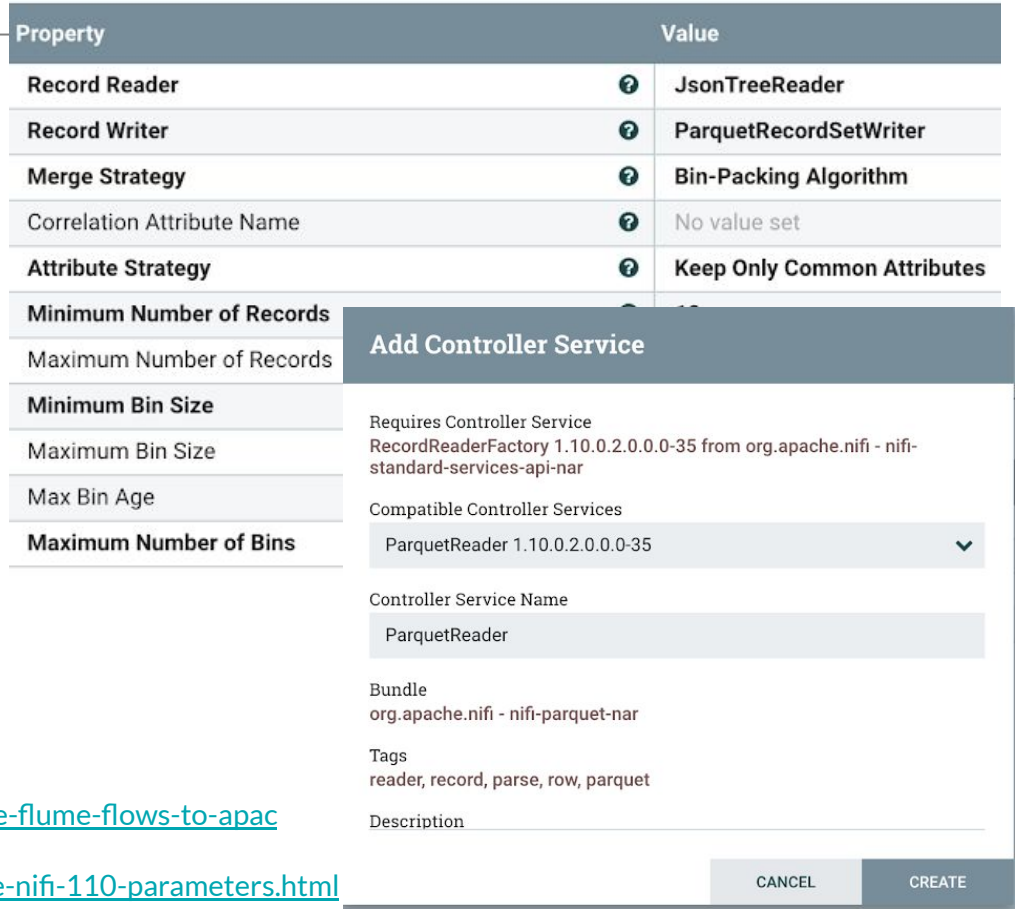
https://youtu.be/Tt8TSlHu7PE

http://lonnifi.blogspot.com/2019/11/back-pressure-prediction-deep-dive.html?es_id=5233333939

ParquetReader / ParquetWriter Records

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

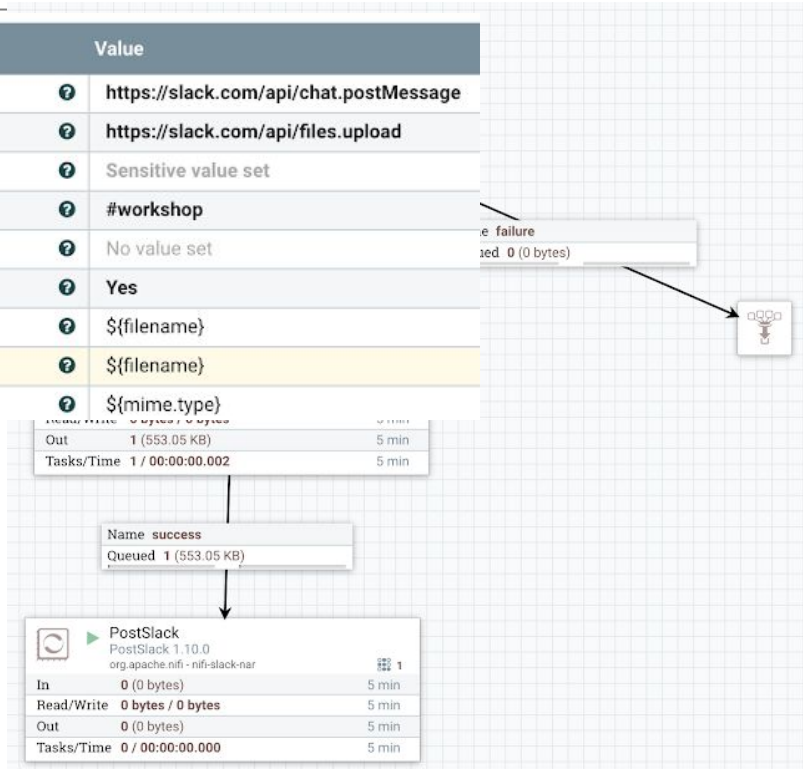
https://www.datainmotion.dev/2019/10/migrating-apache-flume-flows-to-apache_7.html
https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html



PostSlack

- Post Images to Slack

Property	Value
Post Message URL	https://slack.com/api/chat.postMessage
File Upload URL	https://slack.com/api/files.upload
Access Token	Sensitive value set
Channel	#workshop
Text	No value set
Upload FlowFile	Yes
File Title	${filename}
File Name	${filename}
File Mime Type	${mime.type}

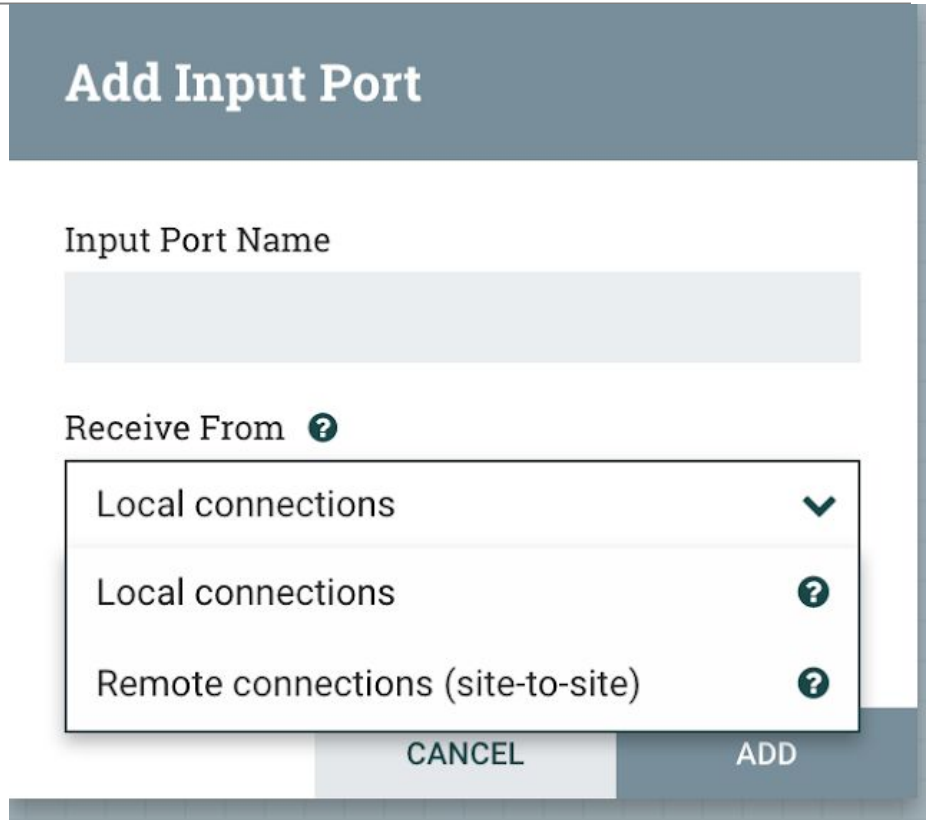


https://www.datainmotion.dev/2019/11/nifi-110-postslack-easy-image-upload.html

https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Remote Input Port in a Process Group

- Put Remote Connections for Site-To-Site (S2S) Anywhere!
- Not only top level
- Drop down simplicity



https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

Many New Features

https://cwiki.apache.org/confluence/display/NIFI/Release+Notes#ReleaseNotes-Version1.10.0

https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html

- Prometheus Reporting Task
- Experimental Encrypted content repository
- PublishKafka Partition Support
- Toolkit module to generate and build Swagger
- GeoEnrichIPRecord Processor
- Command Line Diagnostics
- RocksDB FlowFile Repository
- PutBigQueryStreaming Processor
- Enhanced DevOps and CD/CI

ELT/ETL Lookup Services

- DatabaseRecordLookupService
- KuduLookupService
- HBase_2_ListLookupService

More Resources

- https://www.datainmotion.dev/2019/11/exploring-apache-nifi-110-parameters.html
- https://blog.cloudera.com/adding-nifi-and-kafka-to-cloudera-data-platform/
- https://www.slideshare.net/BryanBende/apache-nifi-sdlc-improvements
- http://lonnifi.blogspot.com/2019/11/back-pressure-prediction-deep-dive.html
- https://github.com/alopresto/slides/tree/master/bdce_2019
- https://github.com/alopresto/slides/tree/master/fodv_2019
- https://github.com/alopresto/slides/tree/master/ioth_2019
- https://medium.com/@abdelkrim.hadjidj/apache-nifi-1-10-series-simplifying-error-handling-7de86f130acd
- https://medium.com/@abdelkrim.hadjidj/hub-and-spoke-architectures-with-nifi-site-to-site-communications-at-any-level-a-nifi-1-10-a8702f77c66e

More Resources

- https://www.datainmotion.dev/2019/11/nifi-110-postslack-easy-image-upload.html
- https://www.datainmotion.dev/2019/11/introducing-mm-flank-apache-flink-stack.html
- https://www.datainmotion.dev/2019/11/nifi-toolkit-cli-for-nifi-110.html
- https://www.datainmotion.dev/2019/09/powering-edge-ai-for-sensor-reading.html
- https://www.slideshare.net/Hadoop_Summit/apache-nifi-crash-course-131483547