Telefonica

# How to ingest 16 billion events

## Data ingestion using Apache Nifi

Barcelona, 2019-03-21
Uwe Weber, Telefónica Germany, uwe.weber@telefonica.com

# Mobile freedom in a digital world

>80 percent owned customers:

Premium O2

Non-Premium Blau

Reseller & Ethnic* ALDI TALK

ay yıldız ortel MOBILE Tchibo

Wholesale < 20%

1&1 DRILLISCH mobilcom debitel

| | |
|---|---|
| EUR 7.3 | Billion revenue in 2017 |
| 49.6 | Million customers** |
| 45,4 | Million mobile network customers** |
| No. 1 | No German network operator connects more people |

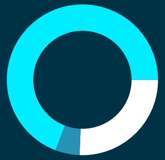
Optimum network

- Outstanding shop network throughout Germany
- fast, reliable mobile network coverage throughout Germany

Network coverage at a glance

- 100% (2G)
- 94% (3G/4G)
- LTE about 85%
- UMTS: ~90%
- ø LTE Speeds: 29.7 Mbit/s
- > 25.000 network sites after integration

Shareholder Structure



69.2% Telefónica Germany Holdings Limited[1]

25.6% Freefloat

5.2% Koninklijke KPN N.V.[2]

[1] Telefónica Germany Holdings Limited is an indirect, wholly owned subsidiary of Telefónica, S.A.

[2] according to press release as of 24.10.2018

*extract

**access numbers according to market practices, as of 30.09.2018

Telefónica

# Auditorium survey

## When did YOU start using mobile telephony?

1992 1995 1997 2000 2005 2010 2014 2019 2021

GSM (2G)

SMS

Prepaid

WAP

UMTS (3G)

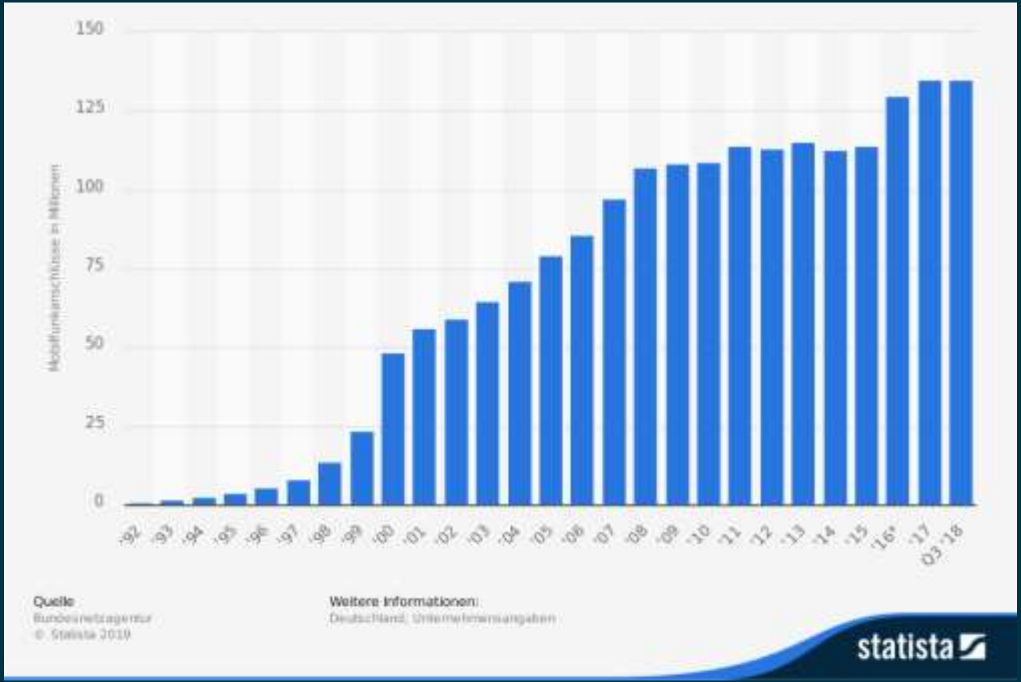
LTE/4G

LTE+

5G

Telefónica

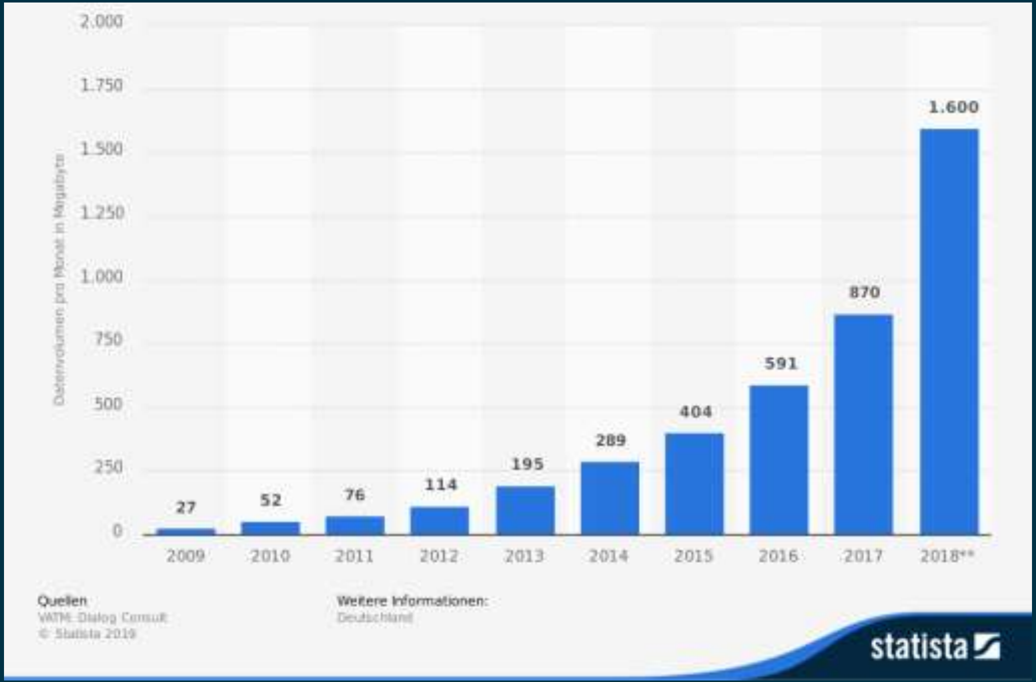
# Mobile phone accounts Germany 1992 - 2018



Quelle
Bundesnetzagentur
© Statista 2019

Weitere Informationen:
Deutschland; Unternehmensangaben

# Monthly data volume Germany

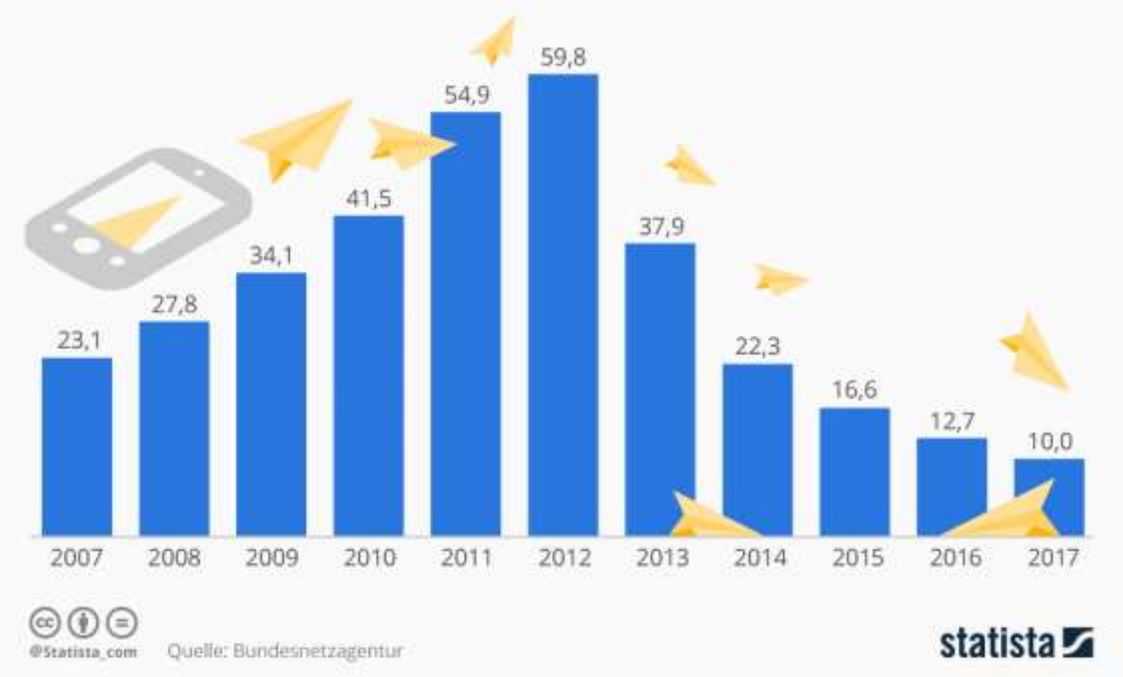


Quellen
VATM; Dialog Consult
© Statista 2019

Weitere Informationen:
Deutschland

Telefónica

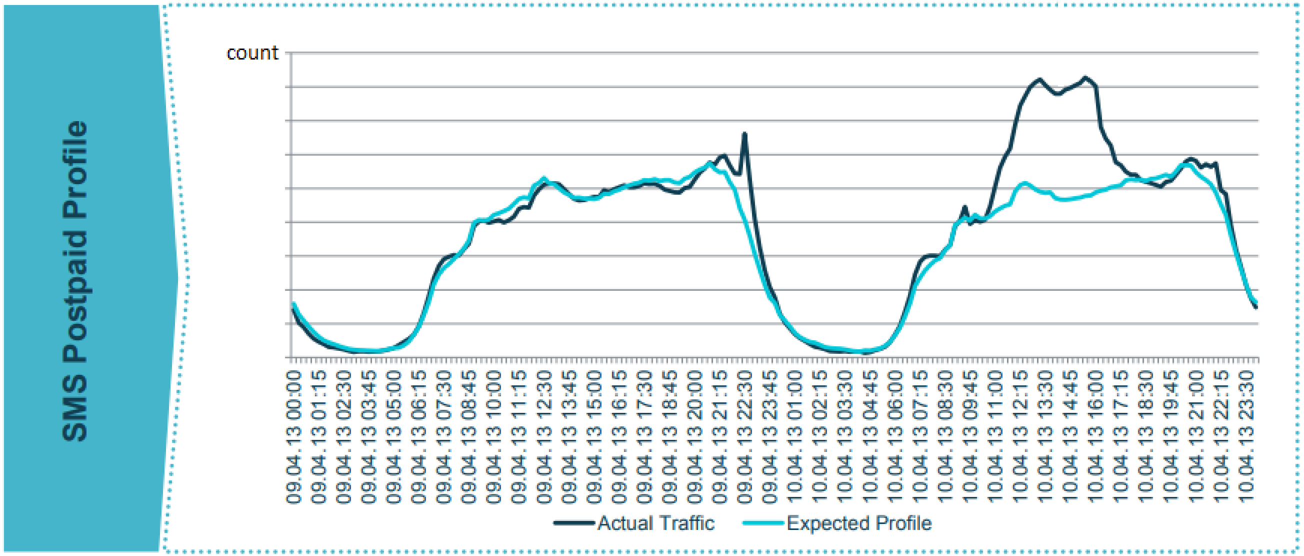
# SMS usage Germany



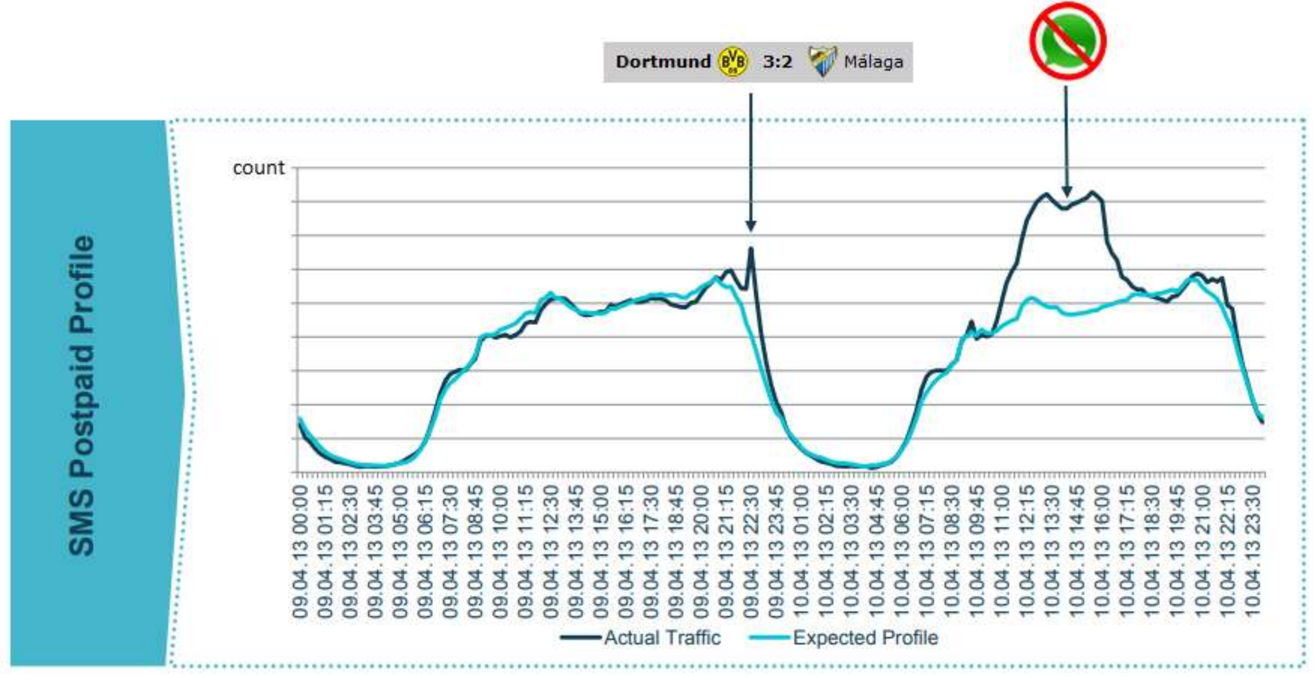
Quelle: Bundesnetzagentur

Telefónica

# But sometimes…



# But sometimes…



Telefónica

# Time has changed

- Telco’s business has changed
  - Tariffs are data usage centric (voice and SMS are flat)
  - High demand for data infrastructure
- Telco’s responsibly has changed
  - Mobile network is essential for some industries
    - Entertainment, streaming services, communication
  - Machine to machine applications
    - Success of companies depends on network stability
    - Defibrillators equipped with SIM cards
  - More to come (5G networks)
    - Autonomous driving

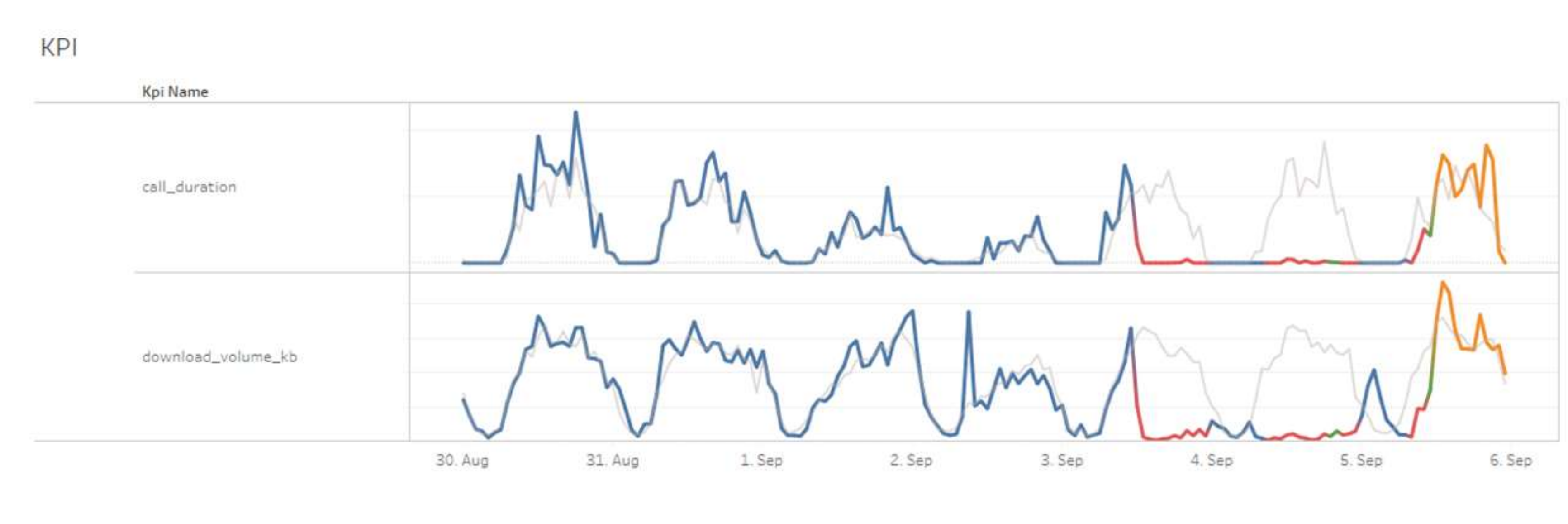
# Early Detection of Mobile Network Outages
# Example: LTE station outage September 2018

03.09 12:00
Station goes (nearly) down.
Seen by Big Data Anomaly Detection

(No automatic alarm from Network Monitoring)

04.09
13:00
Incident

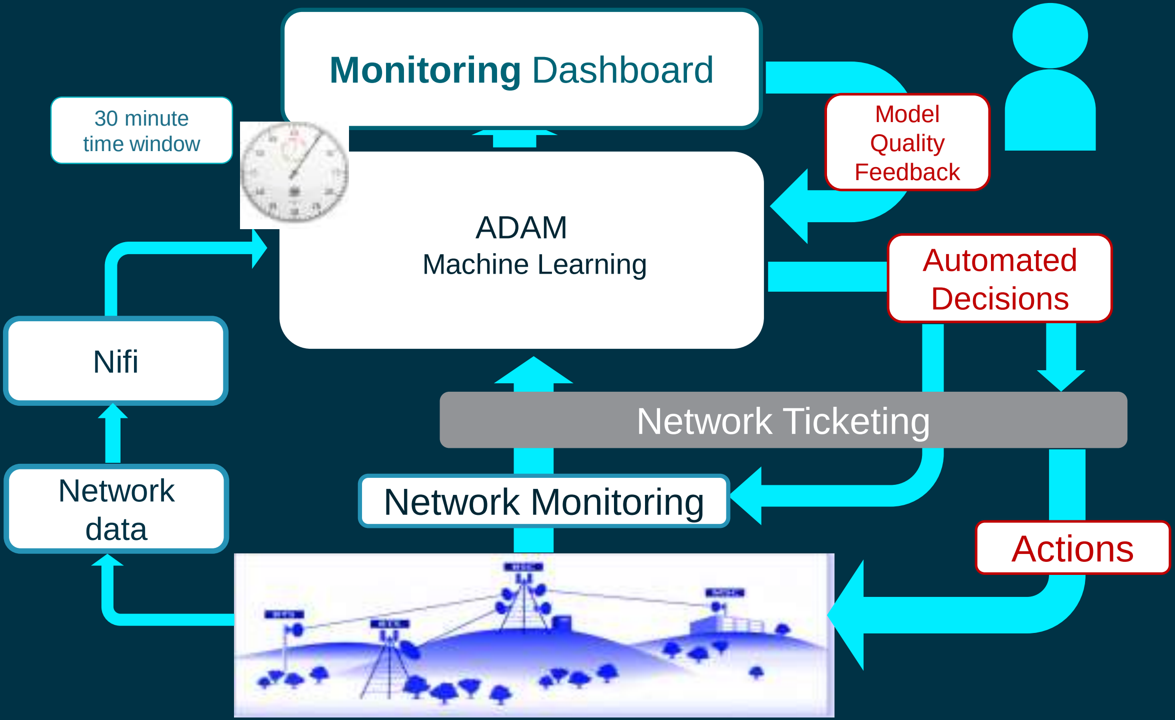
05.09
11:00
Fixed?



Telefónica

# Network Diagnostics: Process Integration

## Automated recommendations/actions integrated into operational process



# Our Big Data / Hadoop environment

- Size does matter
  - Started in 2014: 22 nodes, 352 cores, 768 GB RAM, 300TB disk
  - Today: 75 nodes, 2300 cores, 14TB RAM, 1PB disk
- Use cases: Network planning / analysis
- Redundant management nodes
- Redundant edge nodes for production and analytics
- Secured per Ranger, LDAP and Kerberos
- 10 node Kafka cluster (completely oversized)
- Cluster name is ADAM: Advanced Data Analytic Methods
- Data ingestion per Nifi (HDF), still some Sqoop jobs (going to be replaced by Nifi)

# Nifi evolution

- 2016: Single node installation
  - Stability issues
    - Full file system, corrupted repositories
  - Simple CSV/FTP integrations
- 2017: Two node installation
  - Stability improved
  - Insufficient hardware
  - Splunk real time data integration
- 2018: Three node installation
  - Rollout of high demanding use cases
- Today: Four node installation
  - Moving to Docker
  - Moving from 1.7.1 to 1.8 (HDF 3.3.1)

# Nifi glossary

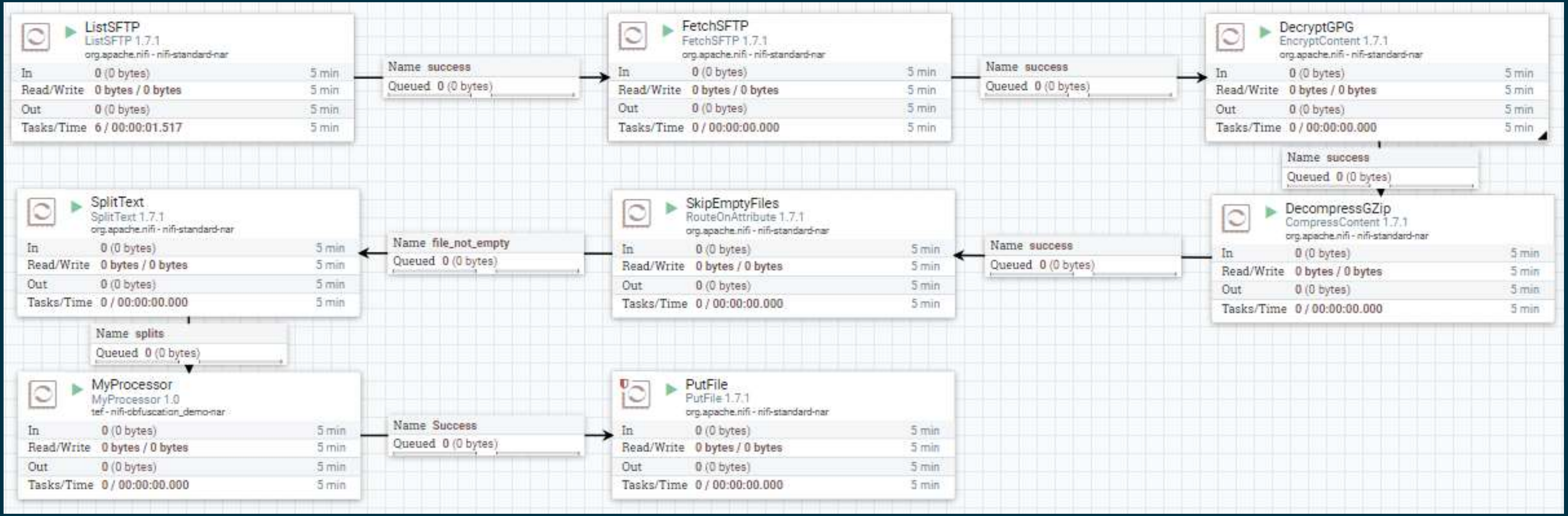
- Processor
  - A working unit
- Flowfile
  - Data
- Queue
  - Waiting data to be processed
- Backpressure
  - Load control/queuing
- Processor Group
  - Logical grouping of Processors which provides a task/procedure
- Cluster
  - Multi instance
- Nifi Registry
  - Version control

# SFTP/CSV file integration

Depending on use cases/data source

- Files may be encrypted/may be not encrypted
- Files are compressed
- Files may contain GDPR relevant data to be filtered
- Target is always a Hive table

# SFTP/CSV file integration



ListSFTP
ListSFTP 1.7.1
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 | 6 / 00:00:01.517 | 5 min |

Name success
Queued 0 (0 bytes)

FetchSFTP
FetchSFTP 1.7.1
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 |

Name success
Queued 0 (0 bytes)

DecryptGPG
EncryptContent 1.7.1
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 |

Name success
Queued 0 (0 bytes)

DecompressGZip
CompressContent 1.7.1
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 |

Name success
Queued 0 (0 bytes)

SkipEmptyFiles
RouteOnAttribute 1.7.1
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 |

Name file_not_empty
Queued 0 (0 bytes)

SplitText
SplitText 1.7.1
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 |

Name splits
Queued 0 (0 bytes)

MyProcessor
MyProcessor 1.0
tef - nifi-obfuscation_demo-nar

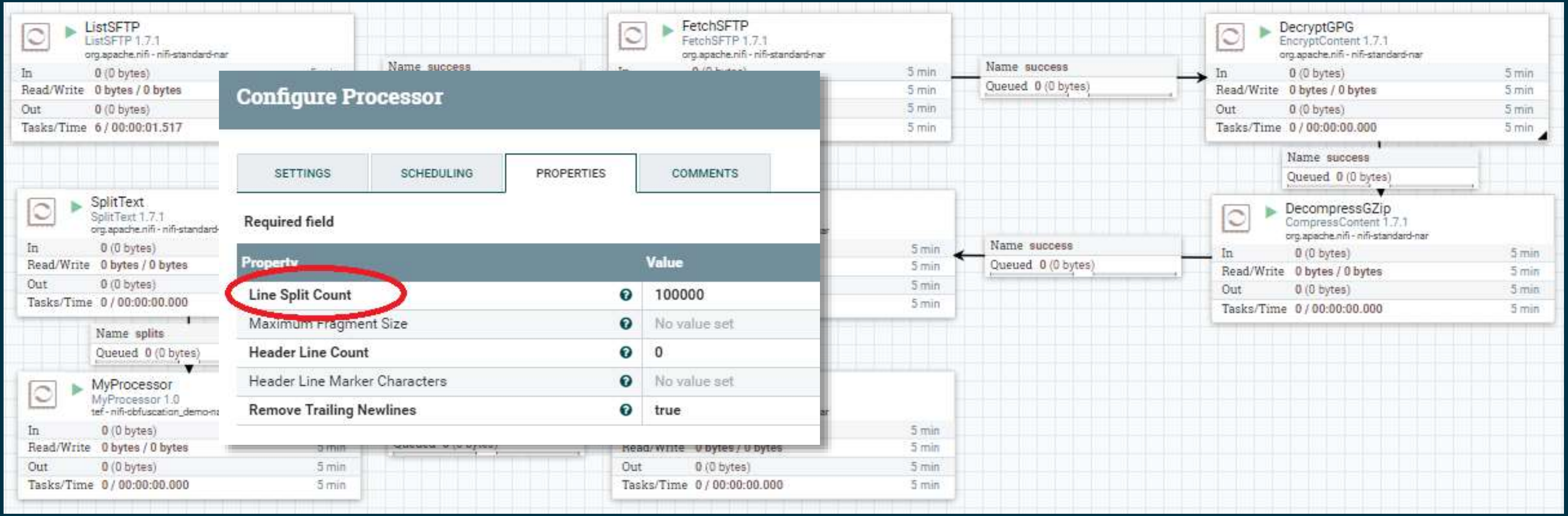
| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

Name Success
Queued 0 (0 bytes)

PutFile
PutFile 1.7.1
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 |

# SFTP/CSV file integration

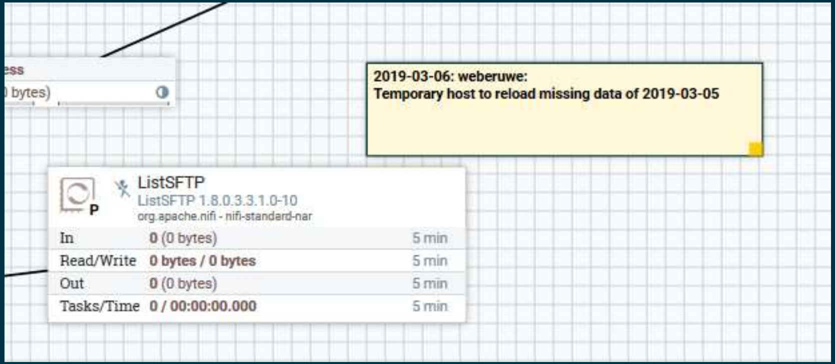


**Configure Processor**

Required field

| Property | Value |
|---|---|
| Line Split Count | 100000 |
| Maximum Fragment Size | No value set |
| Header Line Count | 0 |
| Header Line Marker Characters | No value set |
| Remove Trailing Newlines | true |

# Document your Flows

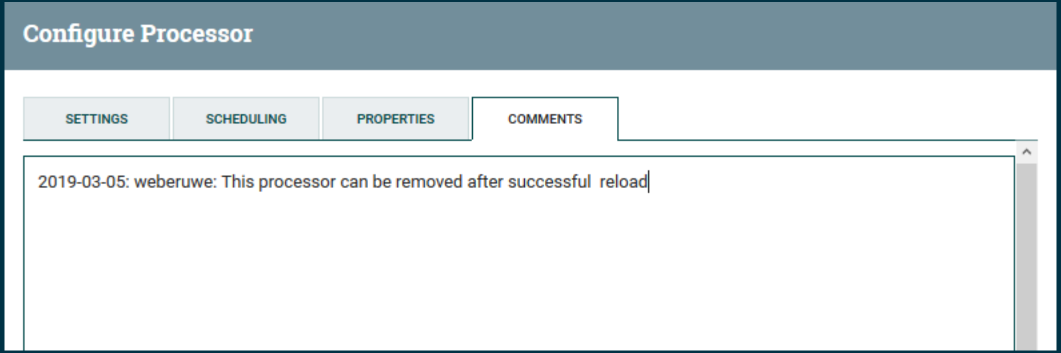


2019-03-06: weberuwe:
Temporary host to reload missing data of 2019-03-05

ListSFTP
ListSFTP 1.8.0.3.3.1.0-10
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
```

- Add text boxes (labels) to describe
  - Contact for source data
  - Contact for target data / customer
  - Error handling
  - Reload strategies
  - … everything which is required for operations
- Processors also contains a “comment” tab



Configure Processor

2019-03-05: weberuwe: This processor can be removed after successful  reload

# Creating a custom Nifi Processor

## Processor input:

```
49111123123123,2019-02-04 12:00:00, 2019-02-04 12:10:00,VOICE,0
49111234234234,2019-02-04 12:00:00, 2019-02-04 12:15:00,LTE,10
49111345345345,2019-02-04 12:05:00, 2019-02-04 12:06:00,UMTS,1
49111567567567,2019-02-04 12:10:00, 2019-02-04 12:15:00,VOICE,0
49111678678678,2019-02-04 12:15:00, 2019-02-04 12:20:00,LTE,20
```

## Processor output:

```
49111--------,2019-02-04 12:00:00, 2019-02-04 12:10:00,VOICE,0
49111--------,2019-02-04 12:00:00, 2019-02-04 12:15:00,LTE,10
49111--------,2019-02-04 12:05:00, 2019-02-04 12:06:00,UMTS,1
49111--------,2019-02-04 12:10:00, 2019-02-04 12:15:00,VOICE,0
49111--------,2019-02-04 12:15:00, 2019-02-04 12:20:00,LTE,20
```

# Creating a custom Nifi Processor

- Create a Java class, inherit from *AbstractProcessor*
- Overwrite method *onTrigger*
- Define processor’s relationships (success, failure)
- Find documentation here:
  - https://community.hortonworks.com/articles/4318/build-custom-nifi-processor.html
  - https://www.javadoc.io/doc/org.apache.nifi/nifi-api/1.7.1

# Creating a custom Nifi Processor code I

```
public void onTrigger(final ProcessContext context, final ProcessSession session)
        throws ProcessException {

    FlowFile flowFile = session.get();
    if ( flowFile == null ) {
        return;
    }

    flowFile = session.write(flowFile, (inputStream, outputStream) -> obfuscate(context, inputStream,
                                                                                outputStream));
    session.transfer(flowFile, SUCCESS);
}
```

Telefónica

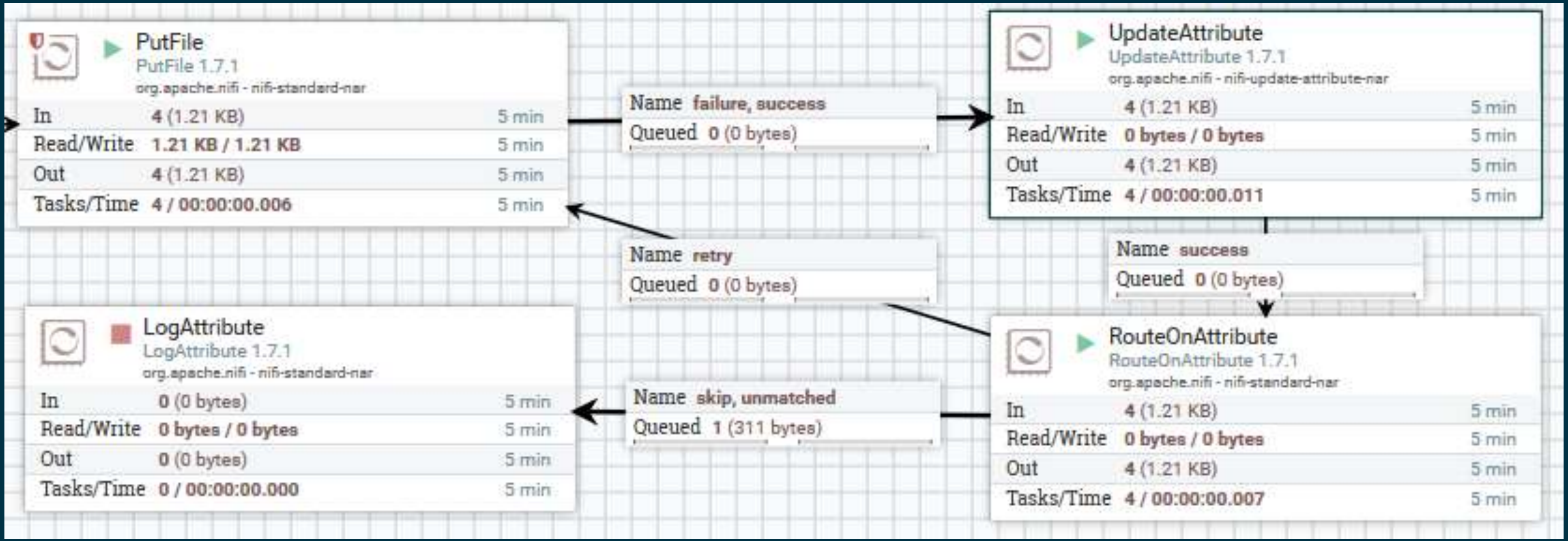
# Creating a custom Nifi Processor code II

```
void obfuscate(ProcessContext context, InputStream inputStream, OutputStream outputStream) {
    LinkedList<String> obfuscatedRecord = new LinkedList<>();

    try (BufferedReader reader = new BufferedReader(new InputStreamReader(inputStream));
         BufferedWriter writer = new BufferedWriter(new OutputStreamWriter(outputStream))) {
        String csvLine;
        while ((csvLine = reader.readLine()) != null) {
            List<String> record = new LinkedList<>(Arrays.asList(csvLine.split(",", -1)));
            String obfuscated = record.get(0).substring(0, 5).concat("--------");
            record.remove(0); record.add(0, obfuscated);
            obfuscatedRecord.add(String.join(",", record));
        }
        writer.write(String.join("\n", obfuscatedRecord));
    } // exception code skipped
```

# Error handling

- Processors can fail due to various reasons
  - Source/target system not available
  - Missing permissions
  - data format mismatch
- Store data in error queue
- Build a retry branch when applicable

# Error handling



**PutFile**
PutFile 1.7.1
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 1.21 KB / 1.21 KB | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.006 | 5 min |

Name failure, success
Queued 0 (0 bytes)

**UpdateAttribute**
UpdateAttribute 1.7.1
org.apache.nifi - nifi-update-attribute-nar

| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.011 | 5 min |

Name retry
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

**LogAttribute**
LogAttribute 1.7.1
org.apache.nifi - nifi-standard-nar

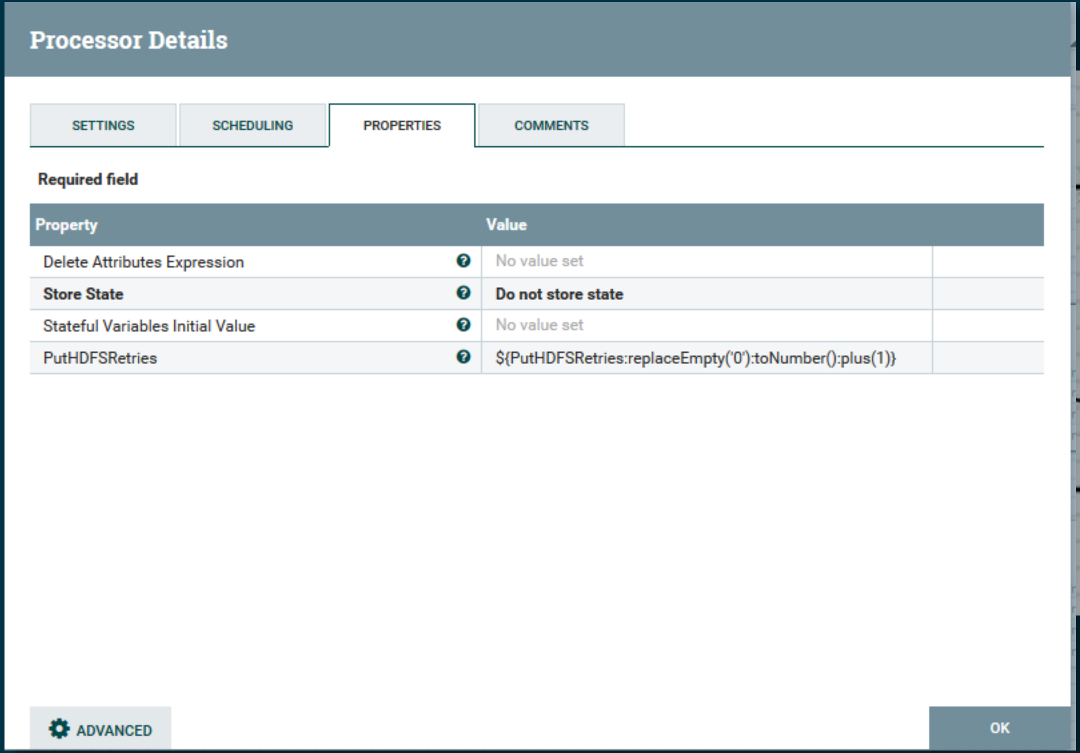
| | | |
|---|---|---|
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

Name skip, unmatched
Queued 1 (311 bytes)

**RouteOnAttribute**
RouteOnAttribute 1.7.1
org.apache.nifi - nifi-standard-nar

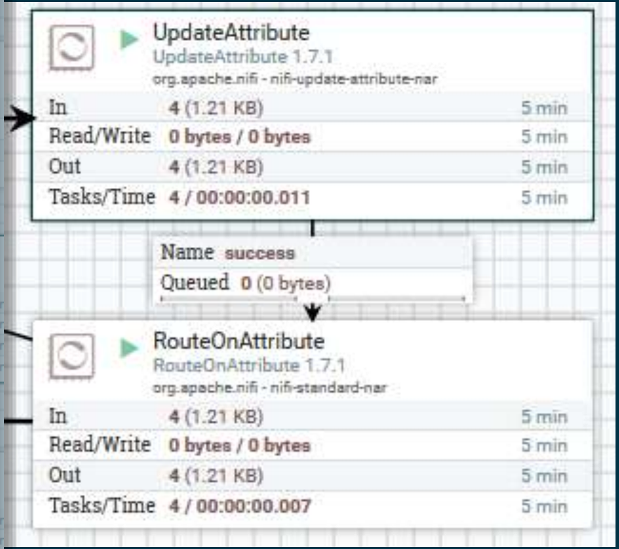
| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.007 | 5 min |

# Error handling



Required field

| Property | Value |
|---|---|
| Delete Attributes Expression | No value set |
| **Store State** | **Do not store state** |
| Stateful Variables Initial Value | No value set |
| PutHDFSRetries | ${PutHDFSRetries:replaceEmpty('0'):toNumber():plus(1)} |



UpdateAttribute
UpdateAttribute 1.7.1
org.apache.nifi - nifi-update-attribute-nar

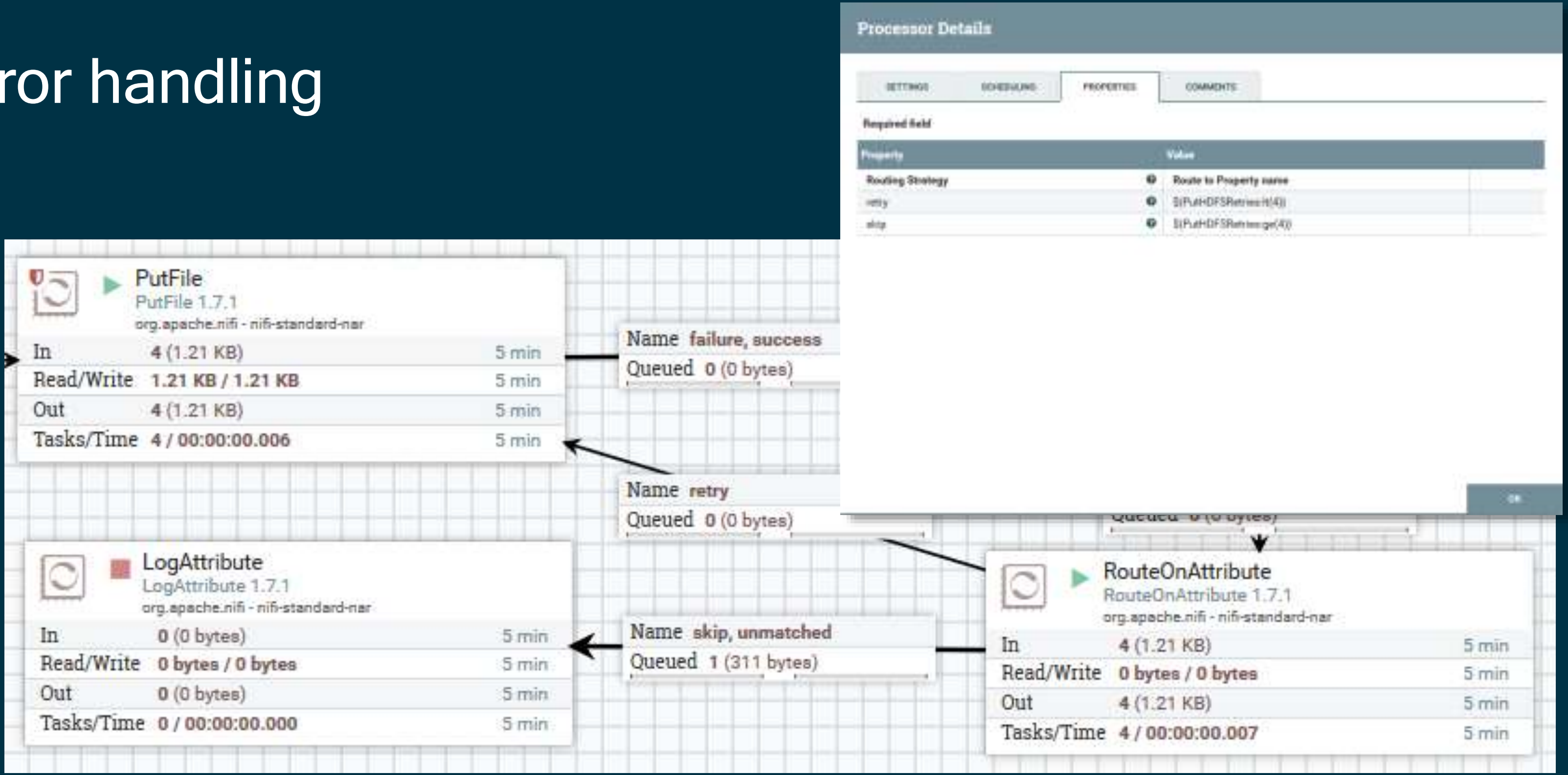
| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.011 | 5 min |

Name success
Queued 0 (0 bytes)

RouteOnAttribute
RouteOnAttribute 1.7.1
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.007 | 5 min |

# Error handling



Processor Details

Required field

| Property | Value |
|---|---|
| Routing Strategy | Route to Property name |
| retry | ${PutHDFSRetries:lt(4)} |
| skip | ${PutHDFSRetries:ge(4)} |

PutFile
PutFile 1.7.1
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 1.21 KB / 1.21 KB | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.006 | 5 min |

Name failure, success
Queued 0 (0 bytes)

Name retry
Queued 0 (0 bytes)

LogAttribute
LogAttribute 1.7.1
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 |

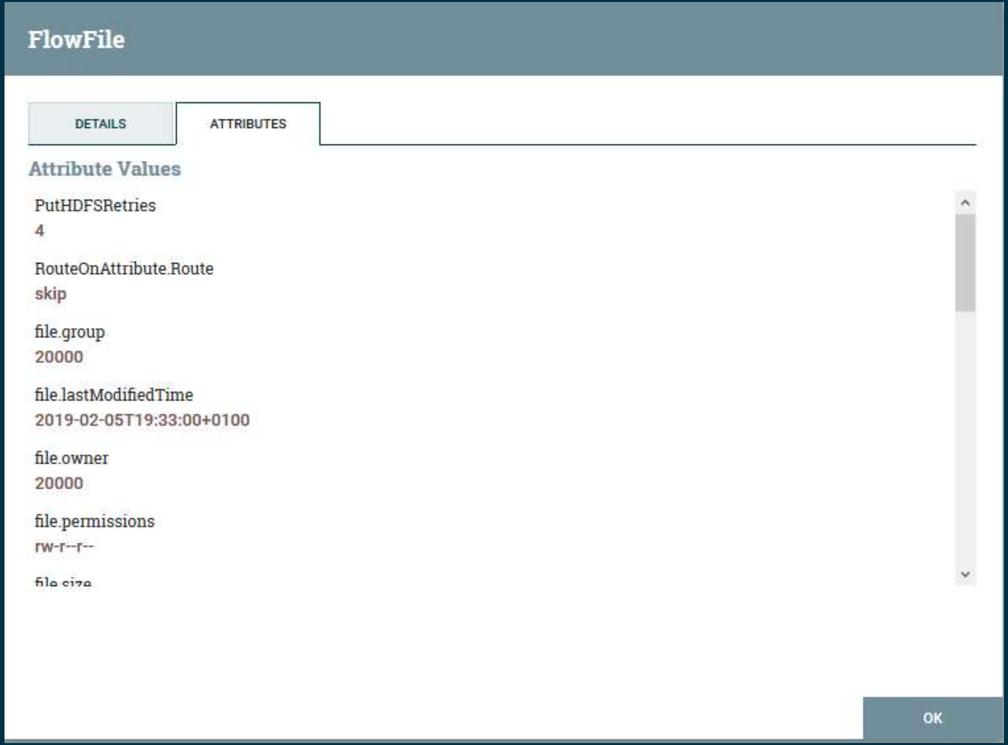
Name skip, unmatched
Queued 1 (311 bytes)

Queued 0 (0 bytes)

RouteOnAttribute
RouteOnAttribute 1.7.1
org.apache.nifi - nifi-standard-nar

| | | |
|---|---|---|
| In | 4 (1.21 KB) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 4 (1.21 KB) | 5 min |
| Tasks/Time | 4 / 00:00:00.007 | 5 min |

# Error handling



PutHDFSRetries
4

RouteOnAttribute.Route
skip

file.group
20000

file.lastModifiedTime
2019-02-05T19:33:00+0100

file.owner
20000

file.permissions
rw-r--r--

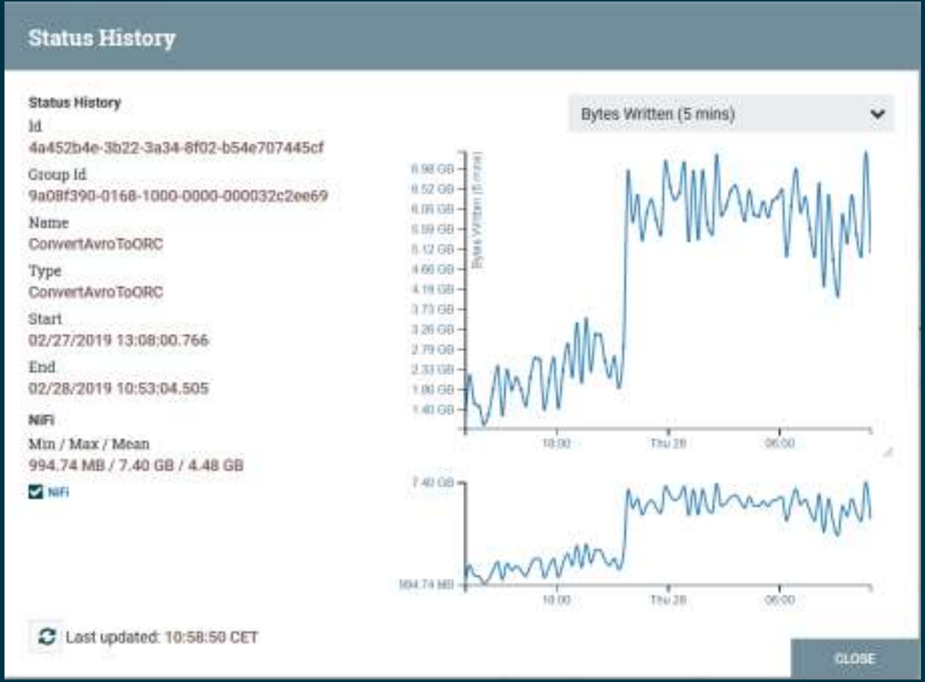
file.size

# Monitoring Nifi Processors

- Build in monitoring available
  - System stats
  - Processor stats
- Limited alarming functionality
- Limited history
- Confusing the more processors you have

➔ Dashboard capabilities would be nice

# Monitoring Nifi Processors



**Status History**

Id
4a452b4e-3b22-3a34-8f02-b54e707445cf

Group Id
9a08f390-0168-1000-0000-000032c2ee69

Name
ConvertAvroToORC

Type
ConvertAvroToORC

Start
02/27/2019 13:08:00.766

End
02/28/2019 10:53:04.505

**NiFi**

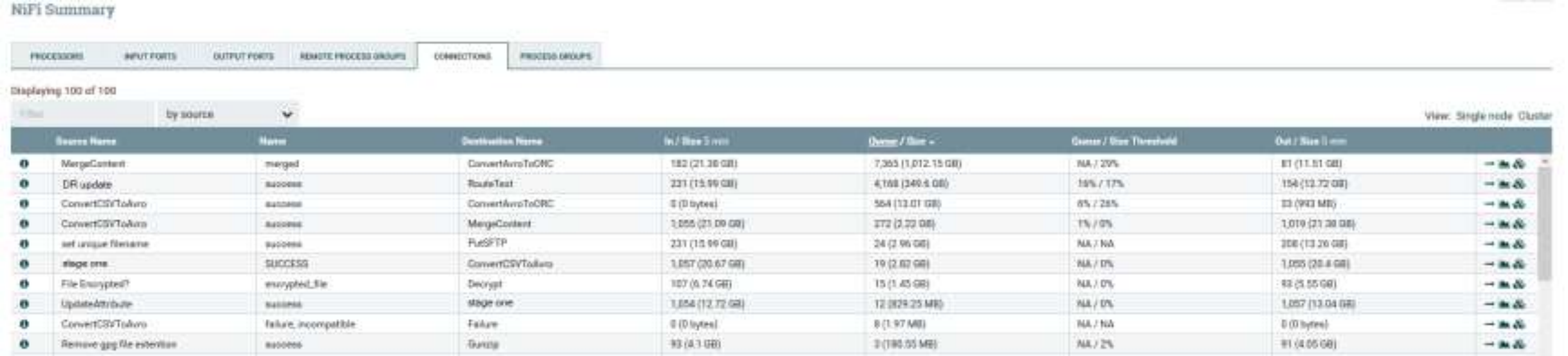
Min / Max / Mean
994.74 MB / 7.40 GB / 4.48 GB

☑ NiFi

Bytes Written (5 mins)

Last updated: 10:58:50 CET

# Monitoring Nifi Processors



NiFi Summary

Displaying 100 of 100

| | Source Name | Name | Destination Name | In / Size 5 min | Queue / Size | Queue / Size Threshold | Out / Size 5 min | |
|---|---|---|---|---|---|---|---|---|
| ❶ | MergeContent | merged | ConvertAvroToORC | 182 (21.38 GB) | 7,365 (1,012.15 GB) | NA / 29% | 81 (11.51 GB) | → |
| ❶ | DR update | success | RouteText | 221 (15.99 GB) | 4,168 (349.6 GB) | 16% / 17% | 156 (12.72 GB) | → |
| ❶ | ConvertCSVToAvro | success | ConvertAvroToORC | 0 (0 bytes) | 564 (13.01 GB) | 6% / 26% | 33 (993 MB) | → |
| ❶ | ConvertCSVToAvro | success | MergeContent | 1,055 (21.09 GB) | 272 (2.22 GB) | 1% / 0% | 1,019 (21.38 GB) | → |
| ❶ | set unique filename | success | PutSFTP | 231 (15.99 GB) | 24 (2.96 GB) | NA / NA | 208 (13.26 GB) | → |
| ❶ | stage one | SUCCESS | ConvertCSVToAvro | 1,057 (20.67 GB) | 19 (2.82 GB) | NA / 0% | 1,055 (20.4 GB) | → |
| ❶ | File Encrypted? | encrypted_file | Decrypt | 107 (6.74 GB) | 15 (1.45 GB) | NA / 0% | 93 (5.55 GB) | → |
| ❶ | UpdateAttribute | success | stage one | 1,054 (12.72 GB) | 12 (829.25 MB) | NA / 0% | 1,057 (13.04 GB) | → |
| ❶ | ConvertCSVToAvro | failure, incompatible | Failure | 0 (0 bytes) | 8 (1.97 MB) | NA / NA | 0 (0 bytes) | → |
| ❶ | Remove gpg file extention | success | Gunzip | 93 (4.1 GB) | 3 (190.55 MB) | NA / 2% | 91 (4.05 GB) | → |

# Monitoring Nifi Processors using REST

```
typeset -r NIFI_URL_BASE="http://<nifi-host>:9090/nifi-api‟

function get_fetch_ftp_byteswritten ()
{
    typeset -r NIFI_PROCESSOR_ID="bd768dc1-8241-45d7-88cd-4e666248776d"
    typeset -r PATTERN_BYTESWRITTEN=".status.aggregateSnapshot.bytesWritten"
    typeset -i bytes_written=0

    bytes_written=$(curl --silent -X GET 
${NIFI_URL_BASE}/processors/${NIFI_PROCESSOR_ID} | jq "${PATTERN_BYTESWRITTEN}")
    echo ${bytes_written}    # return value
}
```

# Monitoring Nifi Processors using REST

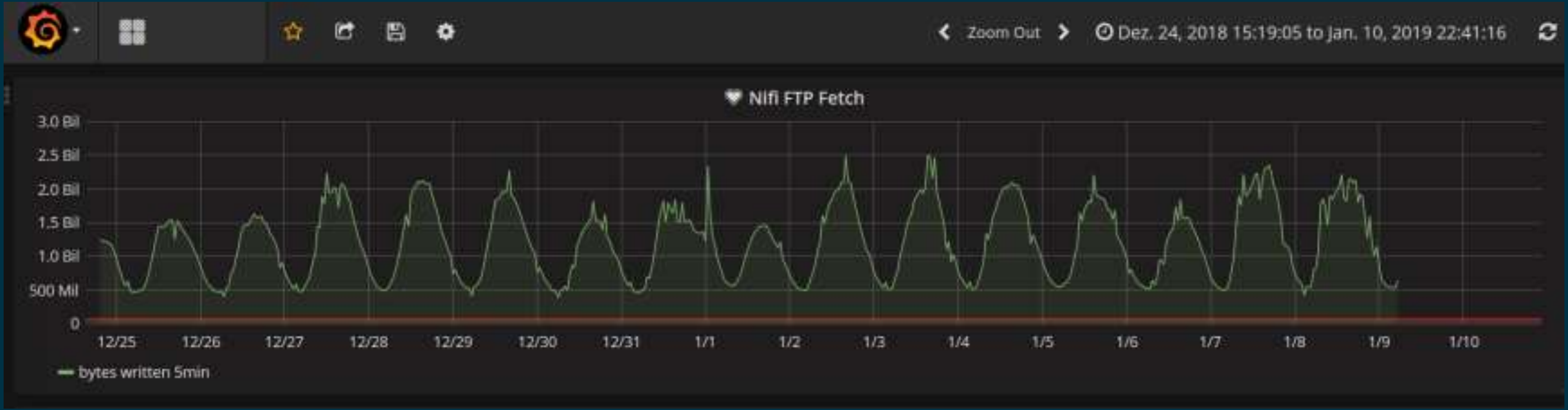
```
"inputRequirement": "INPUT_REQUIRED",
"status": {
  "groupId": "01571011-72aa-14f8-87a2-b692d52d55a6",
  "id": "bd768dc1-8241-45d7-88cd-4e666248776d",
  "name": "Fetch of data",
  "statsLastRefreshed": "17:30:26 CET",
  "aggregateSnapshot": {
    "id": "bd768dc1-8241-45d7-88cd-4e666248776d",
    "groupId": "01571011-72aa-14f8-87a2-b692d52d55a6",
    "name": "Fetch of data",
    "type": "FetchSFTP",
    "runStatus": "Running",
    "bytesRead": 0,
    "bytesWritten": 1676835113,
```

# Monitoring Nifi Processors using REST

## Feed Prometheus

```
cat <<EOF | curl --data-binary @- ${PROMETHEUS_URL_BASE}/metrics/job/data
bytes_written{group="prod",framework="nifi",processor="Fetch of data"} ${bytes_written}
data_file_count{group="prod",framework="hdfs",day="today",type="cs"} ${file_count_cs_today}
EOF
```

# Monitoring Nifi Processors/Grafana



# Monitoring Nifi Processors custom processors

- Implement monitoring capabilities in custom processors
- Full control about what is monitored
  - Technical KPI
  - Business KPI
    - ➔ Real time reporting!
- Freedom of tools
  - We are using Prometheus & Grafana

# Monitoring Nifi Processors with Prometheus

```
io.prometheus.client.exporter.HTTPServer metricsServer;

 if (metricsServer == null) {
            CollectorRegistry registry = new CollectorRegistry();
            registratiosRecords = Counter.build().name("records")
                                         .help("Number of processed records.").register(registry);

            try {
                InetSocketAddress addr = new InetSocketAddress(10888);
                metricsServer = new HTTPServer(addr, registry);
            }
            catch (IOException ex) {
                System.err.println("Failed to start metrics server.");
            }
```

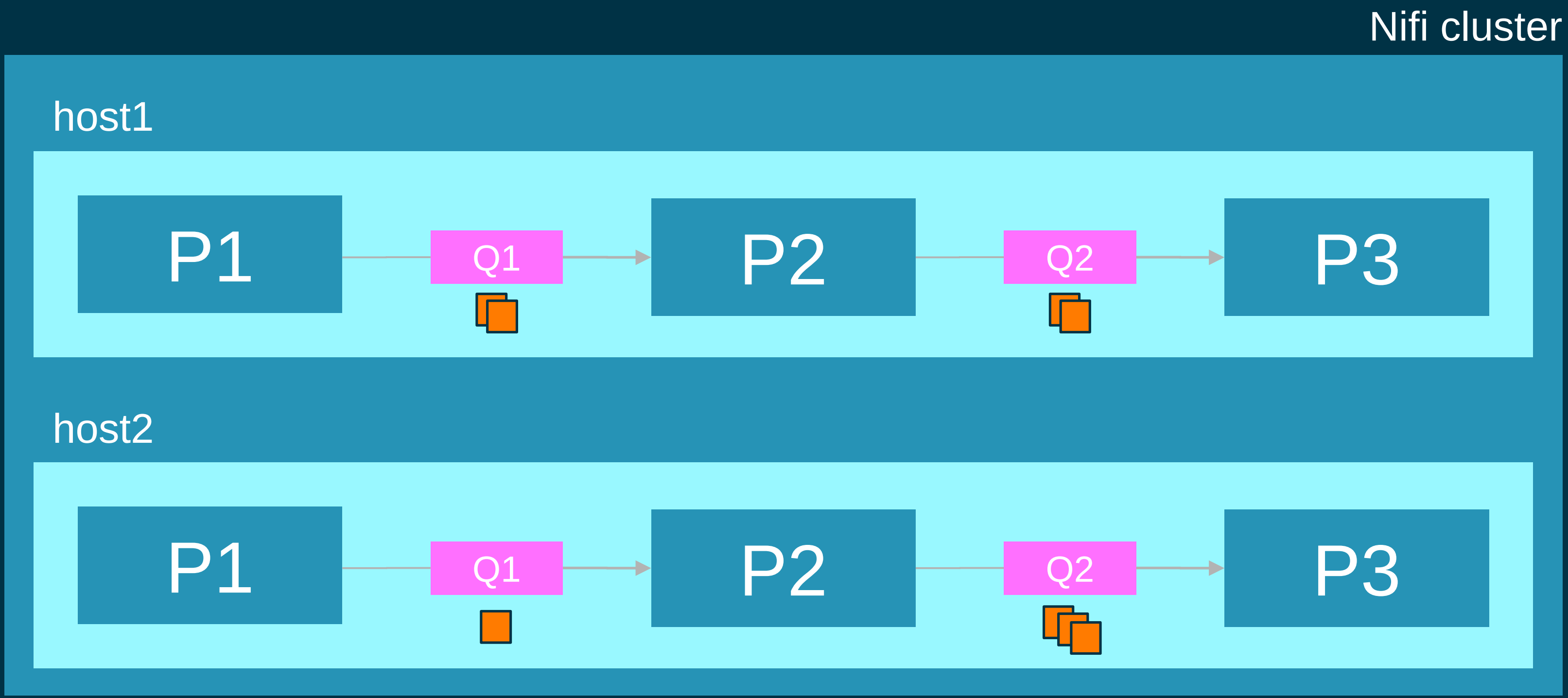
# Monitoring Nifi Processors with Prometheus

```
try (BufferedReader reader = new BufferedReader(new InputStreamReader(inputStream));
            BufferedWriter writer = new BufferedWriter(new OutputStreamWriter(outputStream))) {
           String csvLine;

           while ((csvLine = reader.readLine()) != null) {
               List<String> record = new LinkedList<>(Arrays.asList(csvLine.split(",", -1)));
               String obfuscated = record.get(0).substring(0, 5).concat("--------");
               record.remove(0);
               record.add(0, obfuscated);

               obfuscatedRecord.add(String.join(",", record));
               registratiosRecords.inc();
```

# Why we want a Nifi cluster?

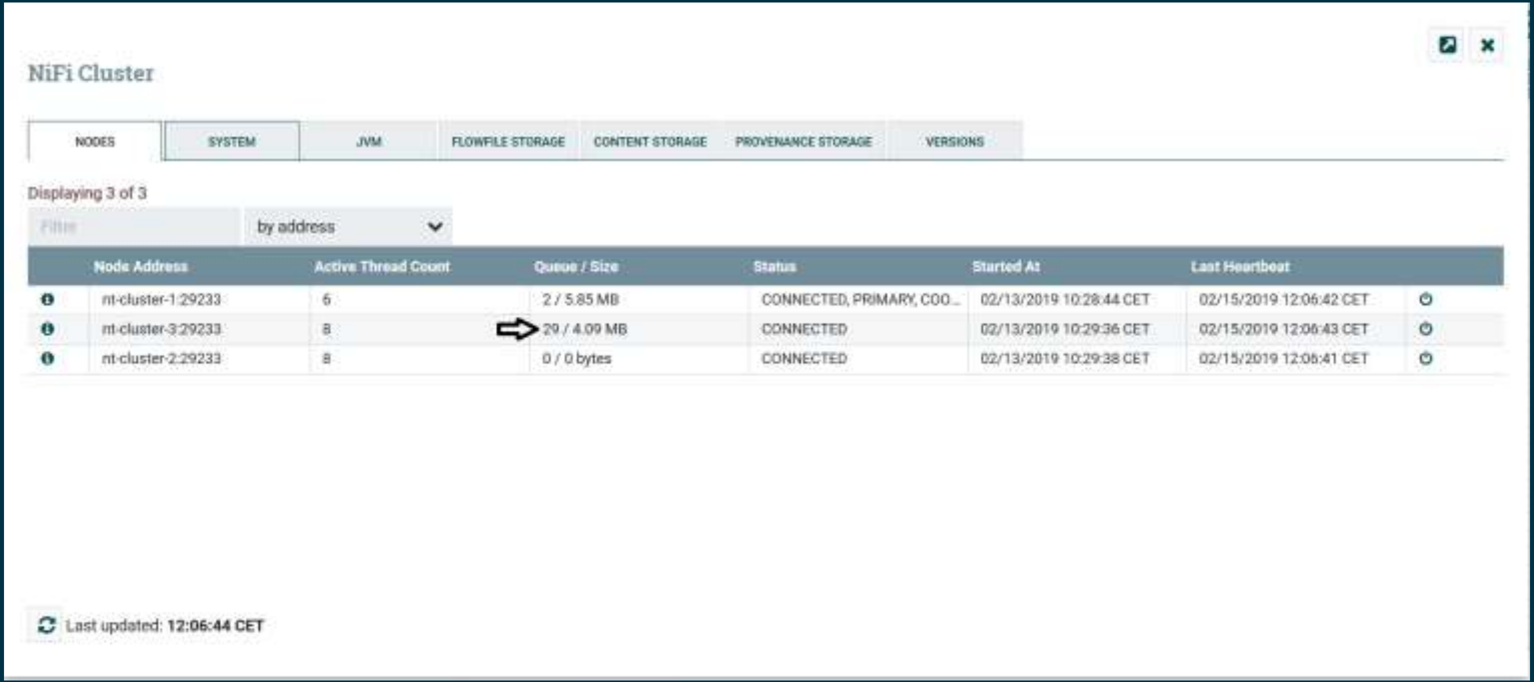
- Parallel Processing
  - Allow multiple processors handle the workload ✓
  - Some processors cannot be executed on multiple nodes ✓
- Workload balancing
  - Better utilization of available resources -
- Availability
  - In case of outage remaining nodes can proceed ✓
  - In case of outage remaining nodes can take over ✗

# Schematic Nifi cluster



Telefónica

# Nifi load balancing



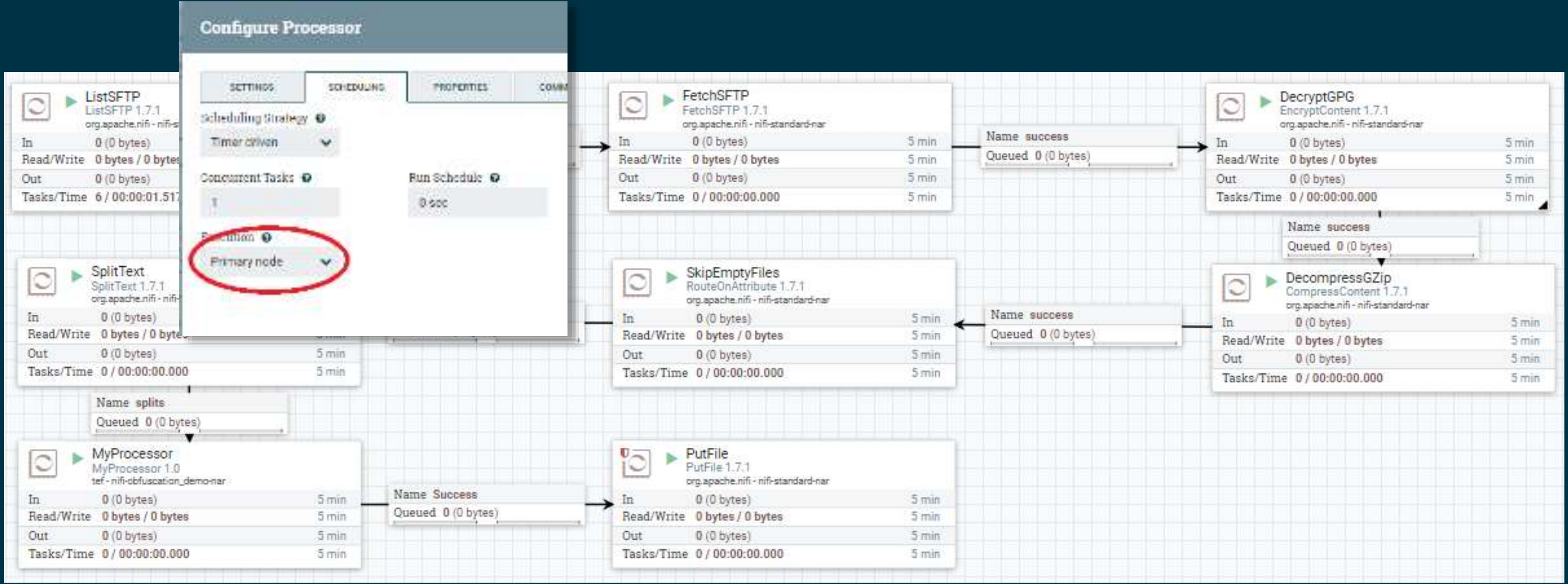
NiFi Cluster

Displaying 3 of 3

| | Node Address | Active Thread Count | Queue / Size | Status | Started At | Last Heartbeat | |
|---|---|---|---|---|---|---|---|
| ℹ | nt-cluster-1:29233 | 6 | 2 / 5.85 MB | CONNECTED, PRIMARY, COO... | 02/13/2019 10:28:44 CET | 02/15/2019 12:06:42 CET | ⏻ |
| ℹ | nt-cluster-3:29233 | 8 | ⇨ 29 / 4.09 MB | CONNECTED | 02/13/2019 10:29:36 CET | 02/15/2019 12:06:43 CET | ⏻ |
| ℹ | nt-cluster-2:29233 | 8 | 0 / 0 bytes | CONNECTED | 02/13/2019 10:29:38 CET | 02/15/2019 12:06:41 CET | ⏻ |

Last updated: **12:06:44 CET**

# Nifi Cluster obstacles



Configure Processor

SETTINGS | SCHEDULING | PROPERTIES | COMM

Scheduling Strategy: Timer driven

Concurrent Tasks: 1

Run Schedule: 0 sec

Execution: Primary node

ListSFTP
ListSFTP 1.7.1
org.apache.nifi - nifi-s
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes)
Tasks/Time 6 / 00:00:01.517

FetchSFTP
FetchSFTP 1.7.1
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

Name success
Queued 0 (0 bytes)

DecryptGPG
EncryptContent 1.7.1
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

Name success
Queued 0 (0 bytes)

DecompressGZip
CompressContent 1.7.1
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

Name success
Queued 0 (0 bytes)

SkipEmptyFiles
RouteOnAttribute 1.7.1
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

SplitText
SplitText 1.7.1
org.apache.nifi - nifi-
In 0 (0 bytes)
Read/Write 0 bytes / 0 bytes
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name splits
Queued 0 (0 bytes)

MyProcessor
MyProcessor 1.0
tef - nifi-obfuscation_demo-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)

PutFile
PutFile 1.7.1
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

# Why we are using Nifi?

Pros

- We were looking for a Swiss Army Knife being able to handle different data sources and formats
- Moving to real time data integration
- Being able to transform data while integration
- Version control
- Extensibility
- Graphical flow control

Cons

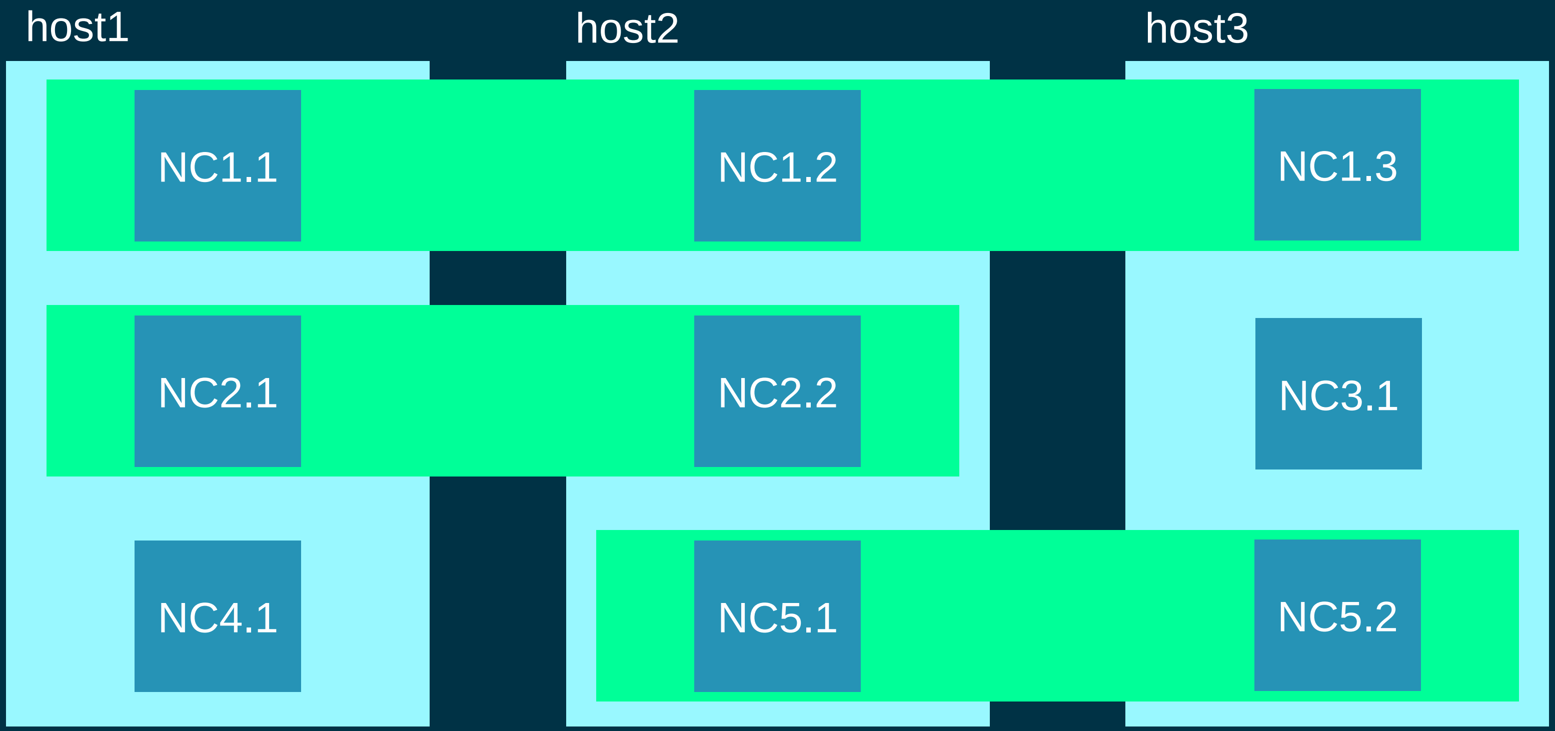
- Monolithic instance
- Super user
- Monitoring / alarming

# Nifi on Docker

Why to Dockerize Nifi flows?

- If one flow is stuck it may not affect other flows
- Using specific Nifi user instead of one mighty user for all Nifi flows
- Monitoring a single instance may become easier
  - Separate log files
  - OS monitoring
- Easy setup for development
- Easy deployment
- Flexibilty
  - But don‘t get a mess!

# Nifi on Docker



NC = Nifi Container

# Nifi for development

```
# doc runnifi weberuwe
[2019-02-05 16:47:23] INFO: max memory set to 1
Using default tag: latest
latest: Pulling from bmidwh/nifi
aeb7866da422: Already exists
4c8dfbaaeab8: Already exists
[…]
Digest: sha256:802700364ad31400af0fed8cdf15f8975dda87c07193247caabfc8c4898ca605
Status: Downloaded newer image for <hostinfo>:5000/bmidwh/nifi:latest
2c0193a4602e9466303b94dd3efc20a35a57c90cb8ebecbb636cacafc3157be1
[2019-02-05 16:49:19] INFO: Nifi container weberuwe_bmidwh_nifi_20190205164723 reachable
under <hostinfo>:9001/nifi
```

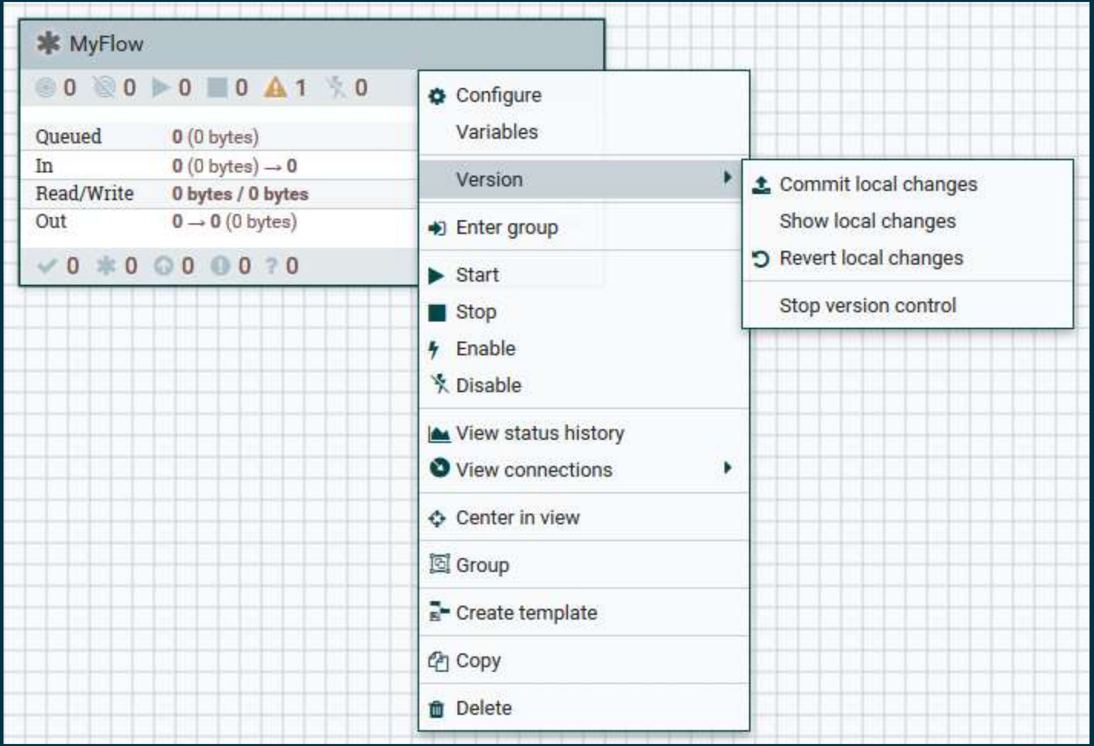
# Setting up a Nifi Dev-Container

Everything is managed by Dockerfile and startup script within container

- Base image is CentOS7
- Copy “empty” Nifi installation
- Copy LDAP, Kerberos config
- Mount local filesystems
  - Hadoop config
  - Nifi repository directories
- Adjust Nifi config (sed!)
- Start Nifi

```
docker run --name "$contname" -h "$contname" --restart=unless-stopped -d -P -e NIFI_RUN_AS_USER=${arg} \
        -e HOST_HOSTNAME=$(hostname) -e NIFI_MAX_MEM_GB=${max_mem} \
        -v /home/${arg}:/home/${arg} \
        -v /etc/profile.d:/etc/profile.d:ro -v /apps/dwh:/apps/dwh:ro \
        -v /etc/hadoop/conf:/etc/hadoop/conf:ro \
        -v /etc/hadoop/conf:/opt/hadoop/etc/hadoop:ro \
        -v /etc/hive/conf:/etc/hive/conf:ro \
        -v /etc/hbase/conf:/etc/hbase/conf:ro \
        -v /etc/kafka/conf:/etc/kafka/conf:ro \
        -v /etc/zookeeper/conf:/etc/zookeeper/conf:ro \
        -v "${NIFI_DOCKER_VOLUME_PATH}/database_repository":/opt/nifi/database_repository \
        -v "${NIFI_DOCKER_VOLUME_PATH}/content_repository":/opt/nifi/content_repository \
        -v "${NIFI_DOCKER_VOLUME_PATH}/provenance_repository":/opt/nifi/provenance_repository \
```

# Nifi Registry Client

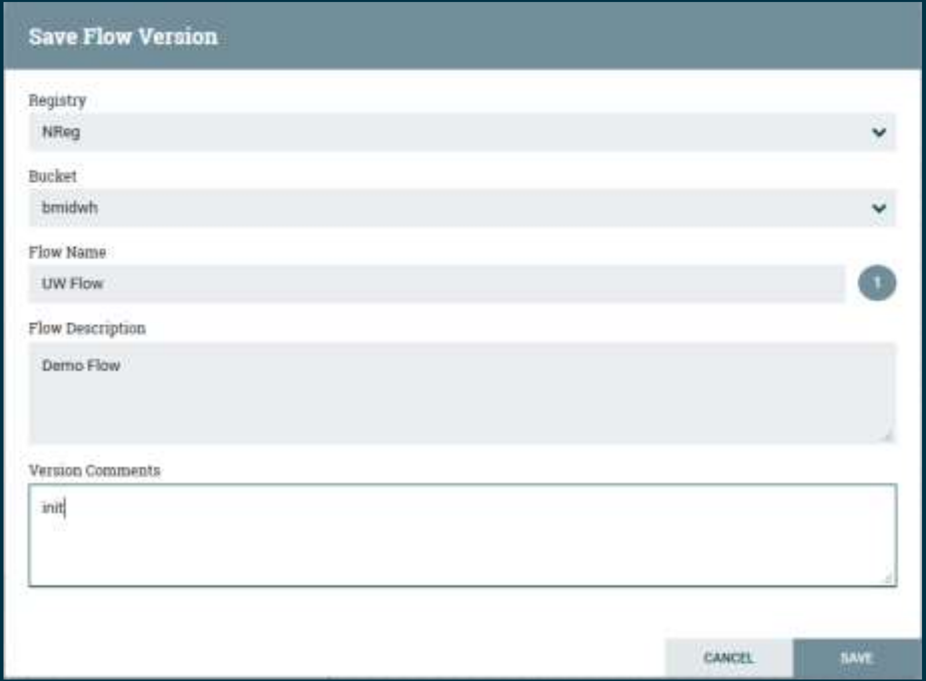


MyFlow

| | |
|---|---|
| Queued | 0 (0 bytes) |
| In | 0 (0 bytes) → 0 |
| Read/Write | 0 bytes / 0 bytes |
| Out | 0 → 0 (0 bytes) |

- Nifi Registry allows version control within Nifi

# Nifi Registry Client



Save Flow Version

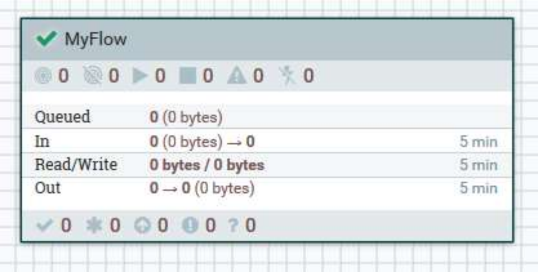
Registry: NReg

Bucket: bmidwh

Flow Name: UW Flow

Flow Description: Demo Flow

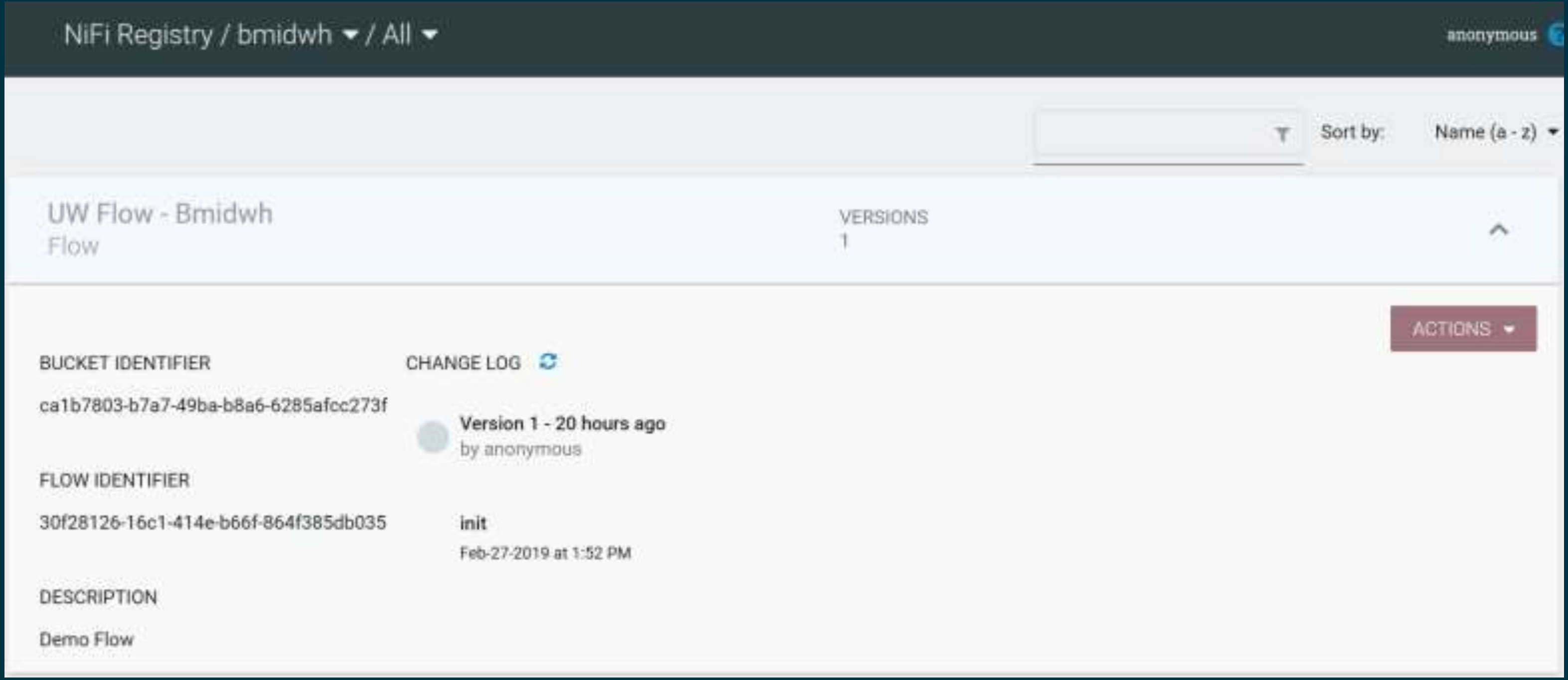
Version Comments: init



MyFlow

| | | |
|---|---|---|
| Queued | 0 (0 bytes) | |
| In | 0 (0 bytes) → 0 | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 → 0 (0 bytes) | 5 min |

# Nifi Registry Backend



UW Flow - Bmidwh
Flow

VERSIONS
1

BUCKET IDENTIFIER
ca1b7803-b7a7-49ba-b8a6-6285afcc273f

FLOW IDENTIFIER
30f28126-16c1-414e-b66f-864f385db035

DESCRIPTION
Demo Flow

CHANGE LOG

**Version 1 - 20 hours ago**
by anonymous

**init**
Feb-27-2019 at 1:52 PM

Telefónica

# Nifi Cluster Deployment

- Developer checks in the flow into Nifi Registry
- Containers are created
  - CI/CD mechanism in Git
    - Invoke custom deploy script
    - Provide information how many “nodes” and where to run
  - Manual checkout from Nifi Registry
  - Manual setting for passwords (!)
- Currently, Docker container are managed by Docker Swarm
  - Moving to Kubernetes in future (?)

# Summary: Nifi Cluster on Docker steps to do

- Have a base OS image
- Install Nifi
- Configure Kerberos, LDAP
- Provide Hadoop client config
- Provide some connectivity libraries you need (e.g. JDBC driver)
- Orchestration is helpful

- Mount local file system into Nifi container
  - Nifi needs to persist flow files
  - Nifi Container cannot change host

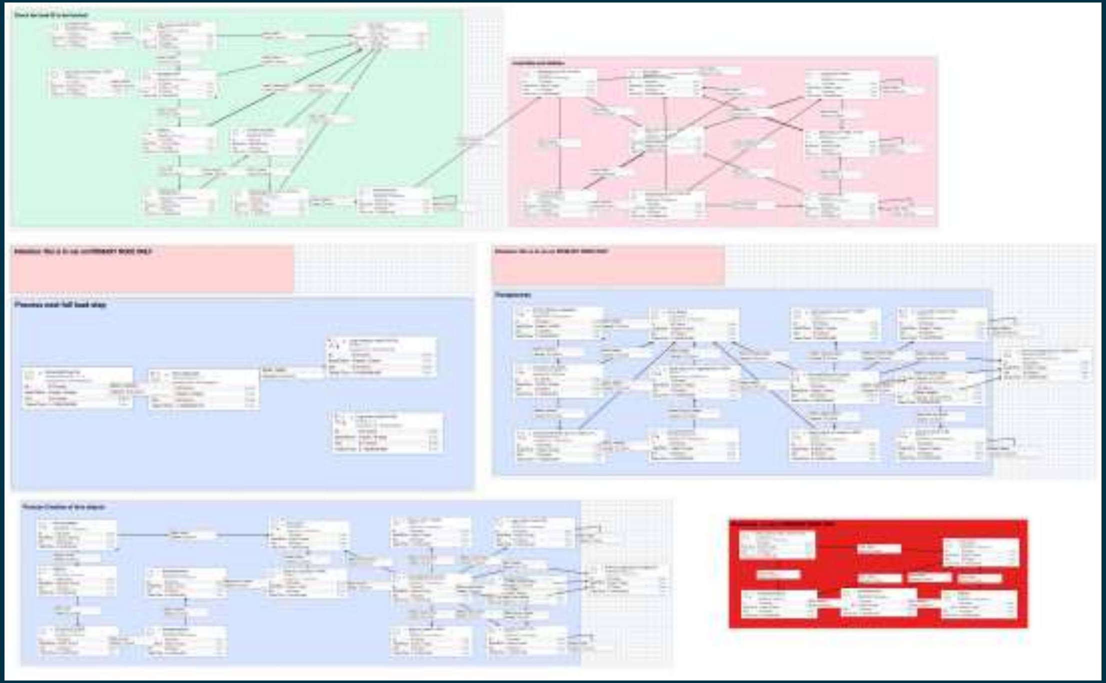
- Create dashboard/portal for your Nifi instances

Telefonica

# Are you doing this?

```
sqoop import ${HADOOP_OPTIONS} \
    --connect ${URL} \
    --username ${USER} \
    --password-file ${PWDFILE} \
    --table ${SOURCE_TABLE} \
    --hive-home ${HIVE_HOME} \
    --hcatalog-home ${HCAT_HOME} \
    --hcatalog-database ${TARGET_SCHEMA} \
    --hcatalog-table ${TARGET_TABLE} \
    --outdir "${SQOOP_CODEGEN_TEMP_DIR}" \
    --columns "${COLUMNS}" \
    --num-mappers 8 \
    --split-by ${SPLIT_COL} \
    --verbose
```

# We do this!



# Database integration with Nifi

- When transferring data from a relational data base to Hive
  - Translation of DDL
  - Datatype mapping
  - Transfer scripts
  - Incremental loading complexity
  - Load control

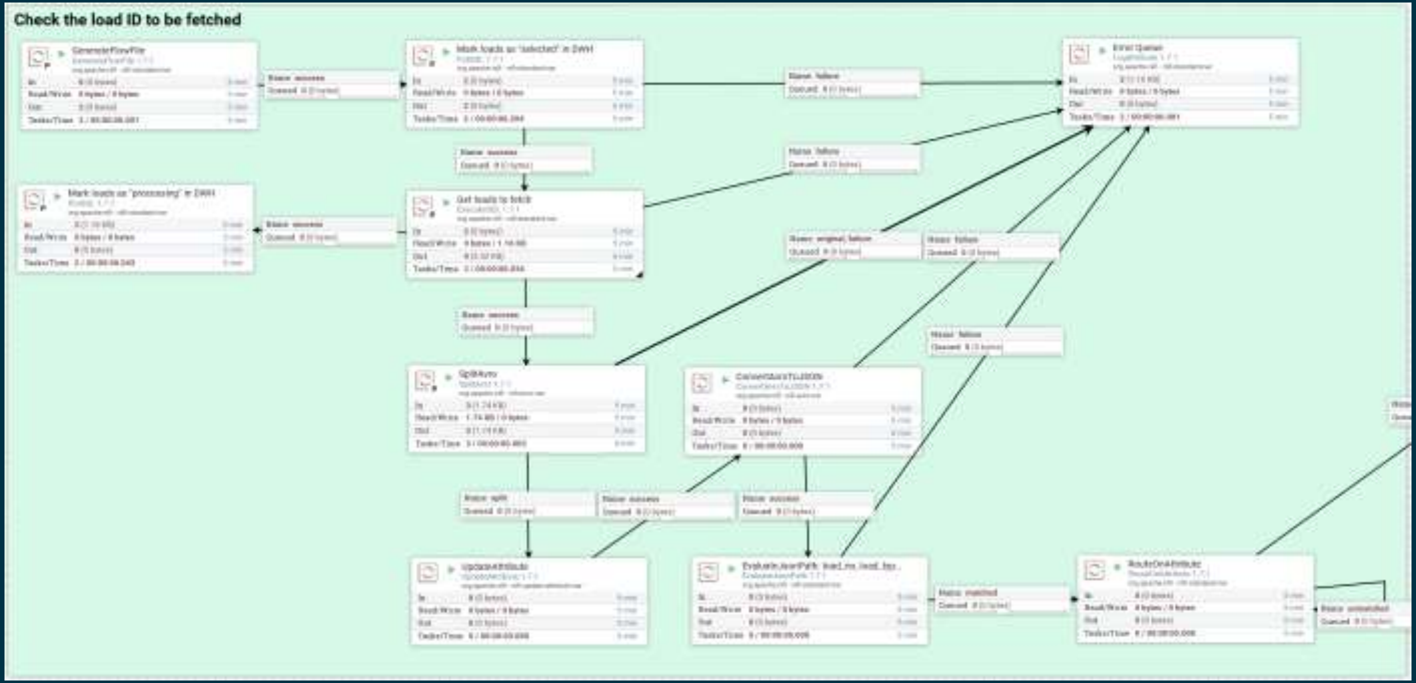
➔ Create a framework that covers all those aspects

# Database integration with Nifi

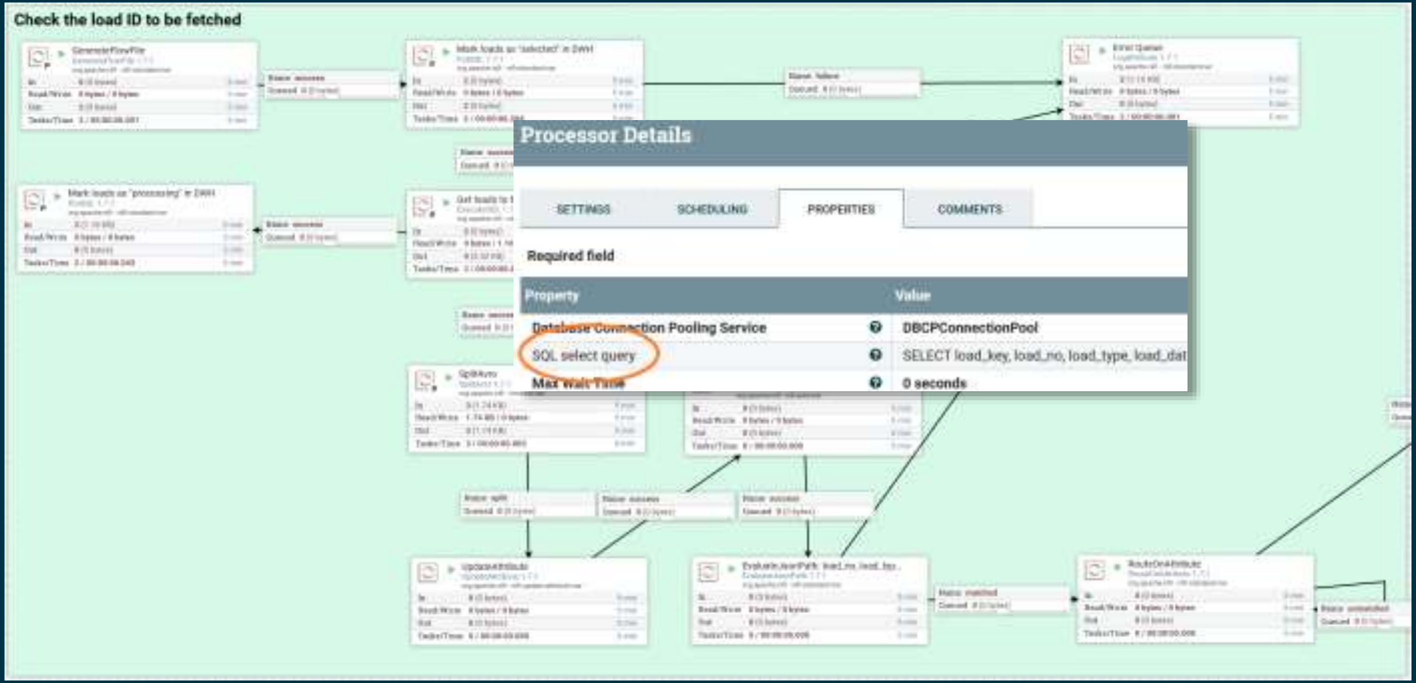
## Have one single place to configure your loads

```
CREATE TABLE T_MTA_LD_CONFIG (
  LOAD_NO                     NUMBER NOT NULL,
  SOURCE_FRAMEWORK            VARCHAR2(32) NOT NULL,
  TARGET_SYSTEM               VARCHAR2(32) NOT NULL,
  IS_ACTIVE                   VARCHAR2(1),
  SRC_OBJECT_ORA_SCHEMA       VARCHAR2(30) NOT NULL,
  SRC_OBJECT_ORA_NAME         VARCHAR2(30) NOT NULL,
  WRK_TABLE_ORA_SCHEMA        VARCHAR2(30) NOT NULL,
  WRK_TABLE_HIVE_SCHEMA       VARCHAR2(64) NOT NULL,
  TGT_TABLE_HIVE_SCHEMA       VARCHAR2(64) NOT NULL,
  NIFI_PROCESSOR_SELECT_EXPR  VARCHAR2(256) NOT NULL,
  NIFI_LOAD_RANGE_COLUMN      VARCHAR2(30),
  NIFI_LOAD_RECORD_COUNT      NUMBER,
  COUNT_WRK_TBL_LOADS         NUMBER,
  ORACLE_PARALLEL_DEGREE      NUMBER,
  LOAD_ID_FRAMEWORK_ASPECT    VARCHAR2(32),
  HDFS_TGT_DIRECTORY_PREFIX   VARCHAR2(256) NOT NULL,
  DESCRIPTION                 VARCHAR2(1024),
  OPT_TGT_TABLE_HIVE_NAME     VARCHAR2(64),
  OPT_TGT_COLUMN_LIST         VARCHAR2(2048)
)
```

# Get meta data: overview



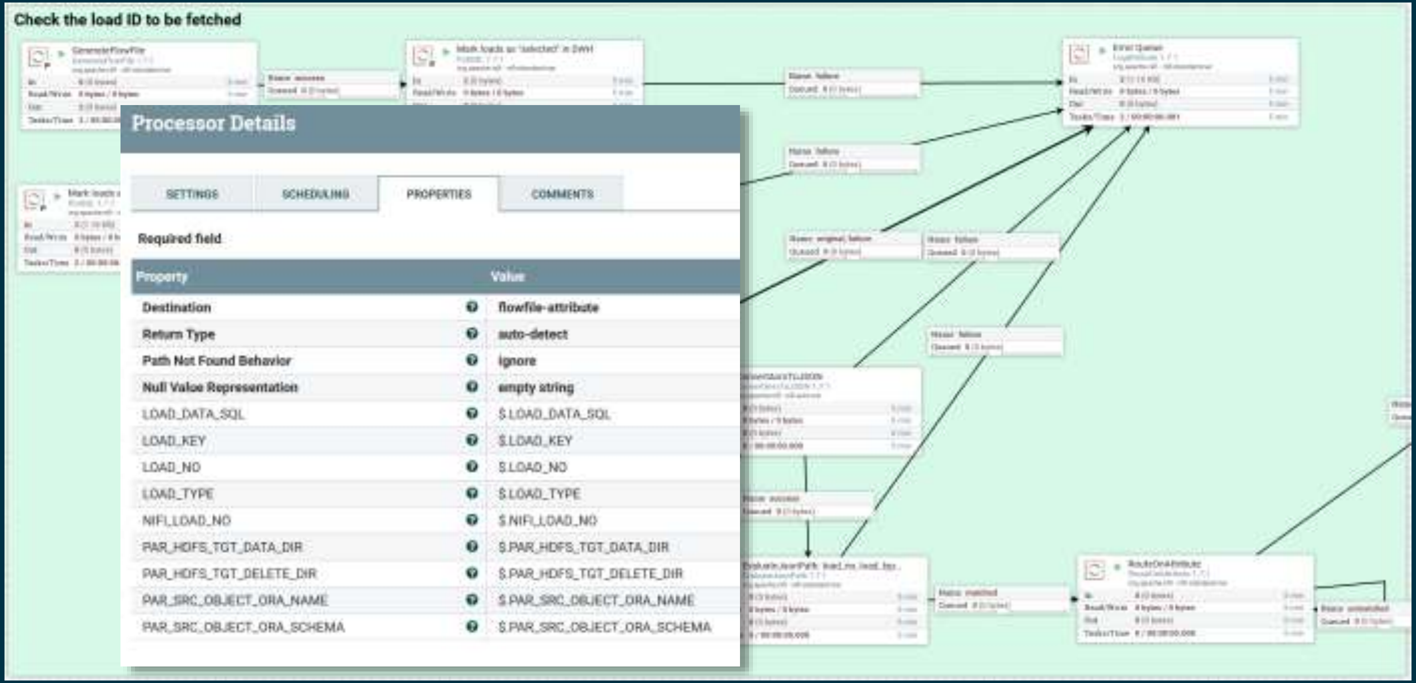
# Get meta data: SQL



Required field

| Property | Value |
|---|---|
| Database Connection Pooling Service | DBCPConnectionPool |
| SQL select query | SELECT load_key, load_no, load_type, load_dat |
| Max Wait Time | 0 seconds |

# Get meta data: set variables



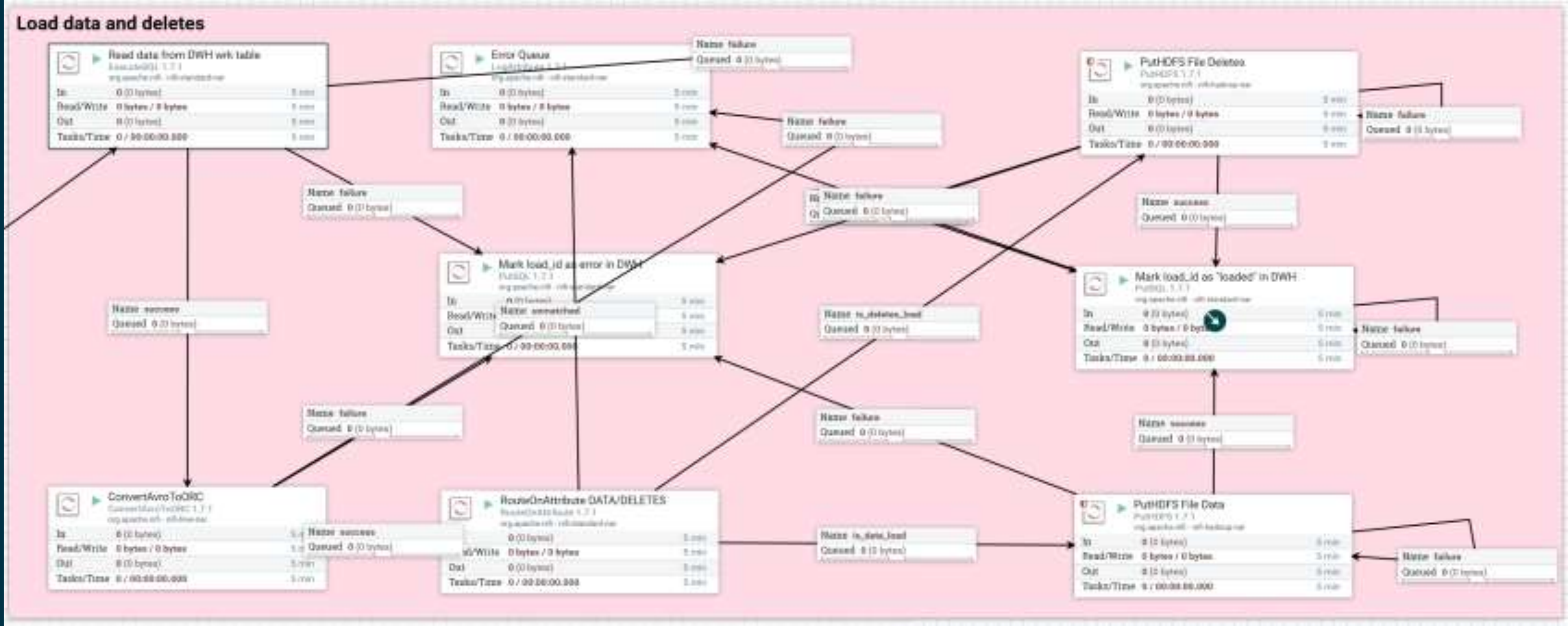
**Processor Details**

SETTINGS | SCHEDULING | PROPERTIES | COMMENTS

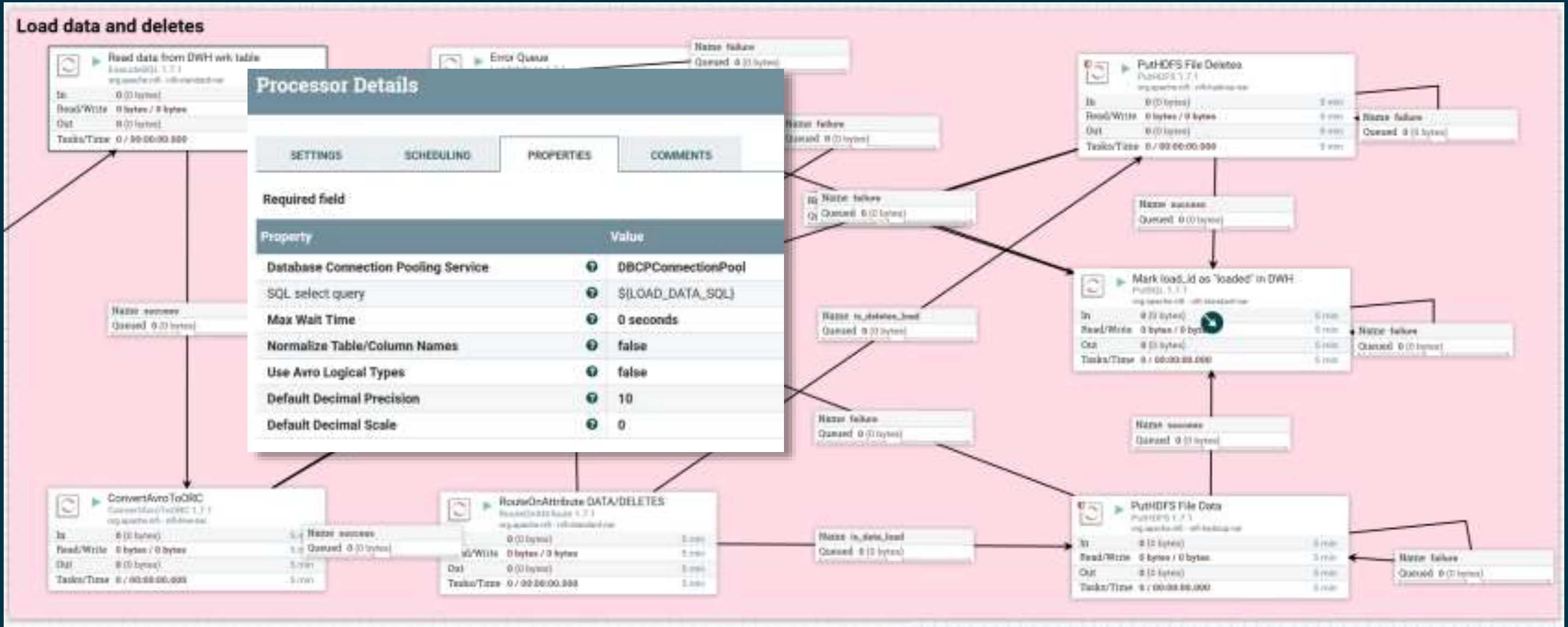
Required field

| Property | Value |
|---|---|
| **Destination** | **flowfile-attribute** |
| **Return Type** | **auto-detect** |
| **Path Not Found Behavior** | **ignore** |
| **Null Value Representation** | **empty string** |
| LOAD_DATA_SQL | $.LOAD_DATA_SQL |
| LOAD_KEY | $.LOAD_KEY |
| LOAD_NO | $.LOAD_NO |
| LOAD_TYPE | $.LOAD_TYPE |
| NIFI_LOAD_NO | $.NIFI_LOAD_NO |
| PAR_HDFS_TGT_DATA_DIR | $.PAR_HDFS_TGT_DATA_DIR |
| PAR_HDFS_TGT_DELETE_DIR | $.PAR_HDFS_TGT_DELETE_DIR |
| PAR_SRC_OBJECT_ORA_NAME | $.PAR_SRC_OBJECT_ORA_NAME |
| PAR_SRC_OBJECT_ORA_SCHEMA | $.PAR_SRC_OBJECT_ORA_SCHEMA |

# Load data: overview


Load data and deletes
Read data from DWH wrk table
ExecuteSQL 1.7.1
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
Error Queue
LogAttribute 1.7.1
org.apache.nifi - nifi-standard-nar
Name failure
Queued 0 (0 bytes)
PutHDFS File Deletes
PutHDFS 1.7.1
org.apache.nifi - nifi-hadoop-nar
Name failure
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
Mark load_id as error in DWH
PutSQL 1.7.1
org.apache.nifi - nifi-standard-nar
Name unmatched
Queued 0 (0 bytes)
Name is_deletes_load
Queued 0 (0 bytes)
Mark load_id as "loaded" in DWH
PutSQL 1.7.1
org.apache.nifi - nifi-standard-nar
ConvertAvroToORC
ConvertAvroToORC 1.7.1
org.apache.nifi - nifi-hive-nar
RouteOnAttribute DATA/DELETES
RouteOnAttribute 1.7.1
org.apache.nifi - nifi-standard-nar
Name is_data_load
Queued 0 (0 bytes)
PutHDFS File Data
PutHDFS 1.7.1
org.apache.nifi - nifi-hadoop-nar

# Load data: SQL



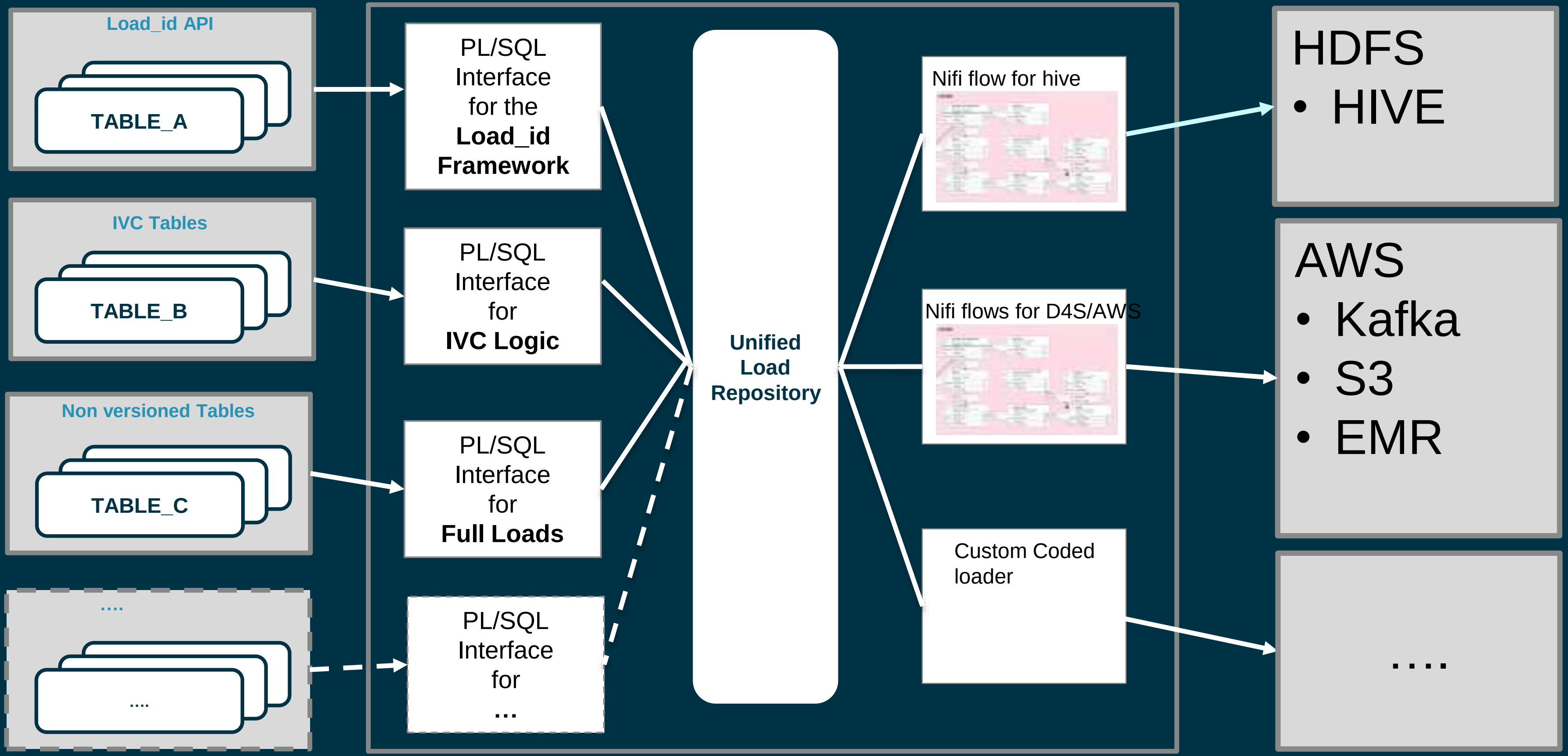
## Processor Details

SETTINGS | SCHEDULING | PROPERTIES | COMMENTS

**Required field**

| Property | | Value |
|---|---|---|
| **Database Connection Pooling Service** | ❓ | DBCPConnectionPool |
| SQL select query | ❓ | ${LOAD_DATA_SQL} |
| **Max Wait Time** | ❓ | 0 seconds |
| **Normalize Table/Column Names** | ❓ | false |
| **Use Avro Logical Types** | ❓ | false |
| **Default Decimal Precision** | ❓ | 10 |
| **Default Decimal Scale** | ❓ | 0 |

# Framework architecture



Load_id API

TABLE_A

IVC Tables

TABLE_B

Non versioned Tables

TABLE_C

....

....

PL/SQL Interface for the **Load_id Framework**

PL/SQL Interface for **IVC Logic**

PL/SQL Interface for **Full Loads**

PL/SQL Interface for **…**

**Unified Load Repository**

Nifi flow for hive

Nifi flows for D4S/AWS

Custom Coded loader

HDFS

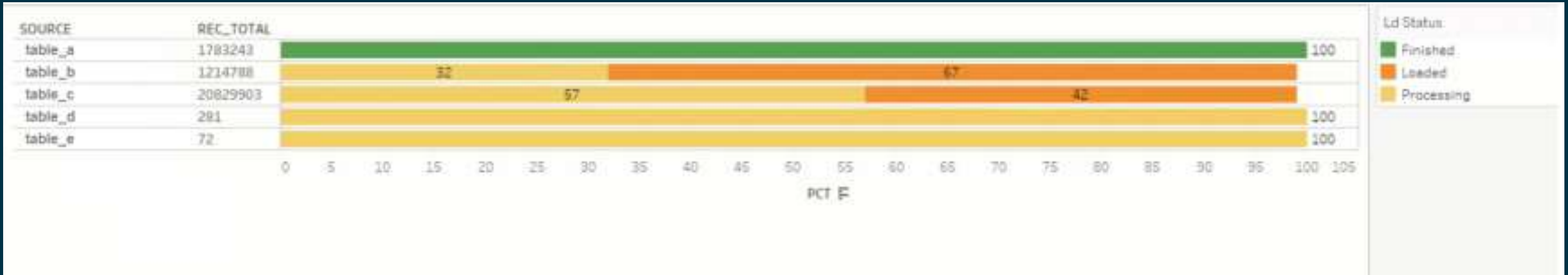
- HIVE

AWS

- Kafka
- S3
- EMR

….

# Tableau Monitor



| SOURCE | REC_TOTAL |
|---|---|
| table_a | 1783243 |
| table_b | 1214788 |
| table_c | 20829903 |
| table_d | 281 |
| table_e | 72 |

# Lessons learned

- Try to have multiple Nifi instances
  - If a single flow is stuck you have to bounce the entire instance including all flows
  - Monitoring and tuning becomes easier
  - Memory
    - Flow files will be kept in memory. This can hurt if flow files are copied instead of moved
- Beware of ETL
  - Single row processing can be expensive and overload your Nifi environment
  - Try to move ETL operations into Hadoop (Spark, Hive, etc.)
- Consider backlog planning
  - What amount of data you expect, what are you disk capacities

# Your questions

Telefónica