

NiFi on a large scale

Архитектура, мониторинг, лучшие практики

Дмитрий Ибрагимов
Data Platform Site Reliability Engineer
2021



Обо мне



- *В IT с 2008*

Дмитрий Ибрагимов

Data Platform Site Reliability Engineer

Dmitry.Ibragimov@leroymerlin.ru

Github/telegram: @diarworld

Обо мне



- *В IT с 2008*
- *В сфере обработки BigData с 2013 года*

Дмитрий Ибрагимов

Data Platform Site Reliability Engineer

Dmitry.Ibragimov@leroymerlin.ru

Github/telegram: @diarworld

Обо мне



- *В IT с 2008*
- *В сфере обработки BigData с 2013 года*
- *Отвечаю за стабильность платформы данных в LM*

Дмитрий Ибрагимов

Data Platform Site Reliability Engineer

Dmitry.Ibragimov@leroymerlin.ru

Github/telegram: @diarworld

Введение

Agenda



1. *0 LMRU DataPlatform*

Agenda

1. *O LMRU DataPlatform*
2. *Наши метрики + архитектура*

Agenda

1. *O LMRU DataPlatform*
2. *Наши метрики + архитектура*
3. *Обзор инструмента – плюсы/минусы*

Agenda

1. *O LMRU DataPlatform*
2. *Наши метрики + архитектура*
3. *Обзор инструмента – плюсы/минусы*
4. *Мониторинг NiFi*

Agenda

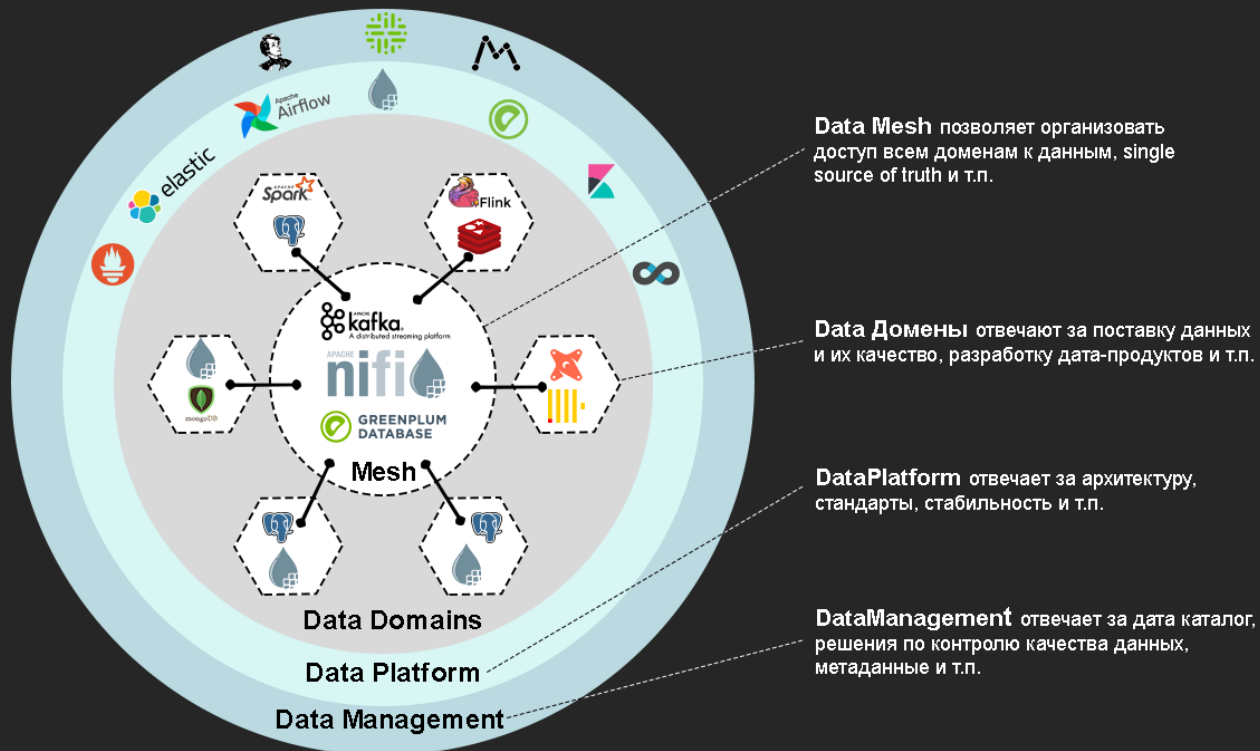
1. *O LMRU DataPlatform*
2. *Наши метрики + архитектура*
3. *Обзор инструмента – плюсы/минусы*
4. *Мониторинг NiFi*
5. *Лучшие практики*

Agenda

1. *O LMRU DataPlatform*
2. *Наши метрики + архитектура*
3. *Обзор инструмента – плюсы/минусы*
4. *Мониторинг NiFi*
5. *Лучшие практики*
6. *Архитектура будущего*

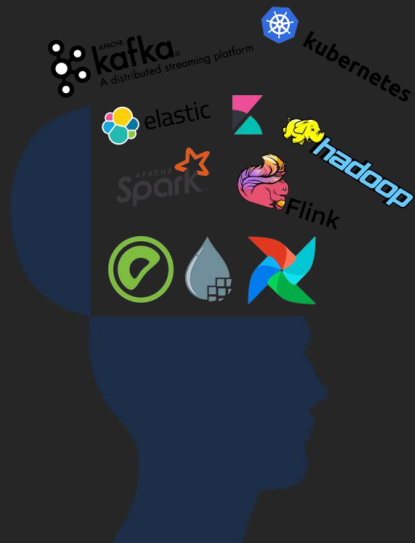
Leroy Merlin experience

DataPlatform LM проект был начат в 2019 году. Цель – создать решение, которое позволит всем сотрудникам компании работать с данными.



Leroy Merlin experience

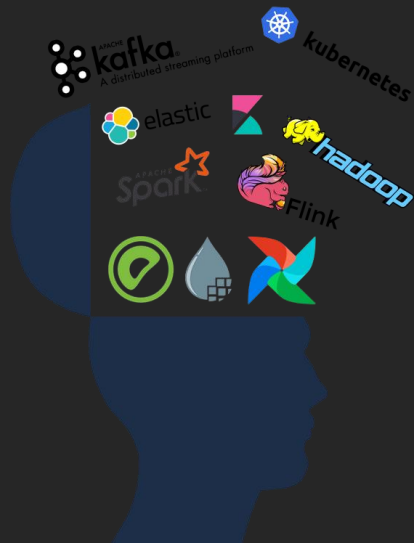
DataPlatform LM проект был начат в 2019 году. Цель – создать решение, которое позволит всем сотрудникам компании работать с данными.



https://habr.com/ru/company/leroy_merlin/blog/561072

Leroy Merlin experience

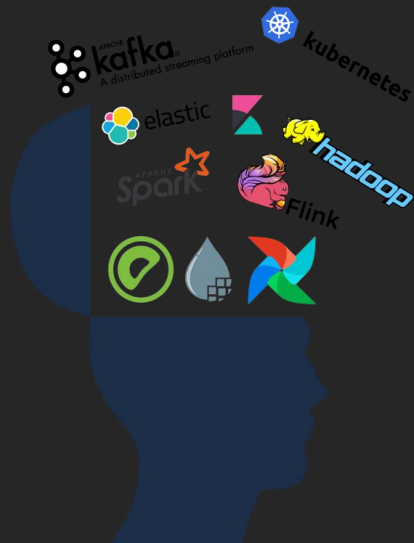
DataPlatform LM проект был начат в 2019 году. Цель – создать решение, которое позволит всем сотрудникам компании работать с данными.



https://habr.com/ru/company/leroy_merlin/blog/561072

Leroy Merlin experience

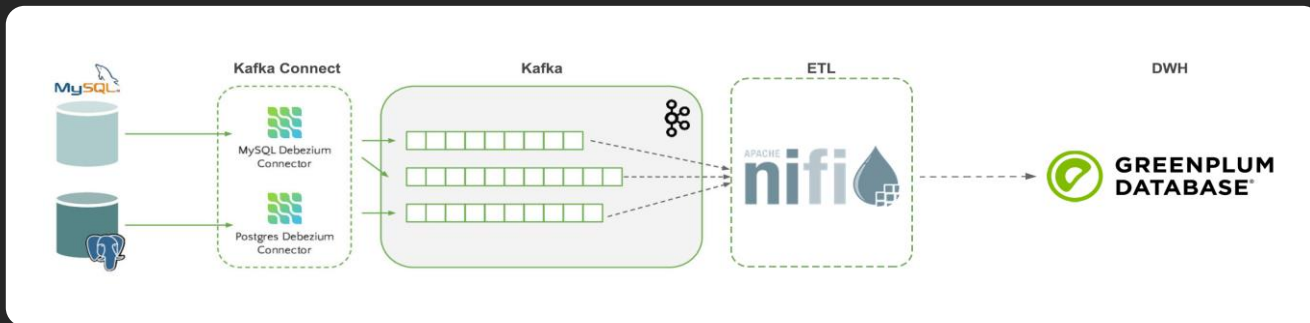
DataPlatform LM проект был начат в 2019 году. Цель – создать решение, которое позволит всем сотрудникам компании работать с данными.



https://habr.com/ru/company/leroy_merlin/blog/561072

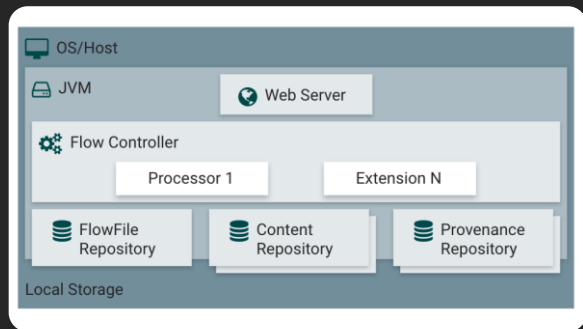
Leroy Merlin experience

DataPlatform LM проект был начат в 2019 году. Цель – создать решение, которое позволит всем сотрудникам компании работать с данными.

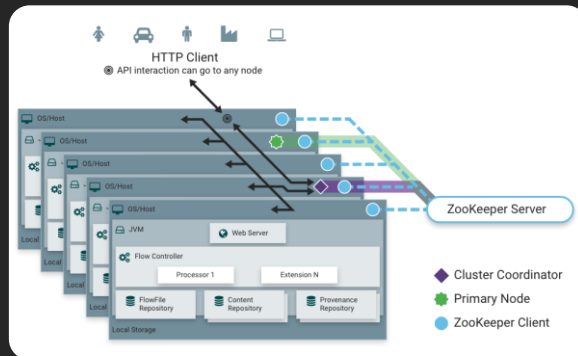


NiFi Architecture

NiFi An easy to use, powerful, and reliable system to process and distribute data.

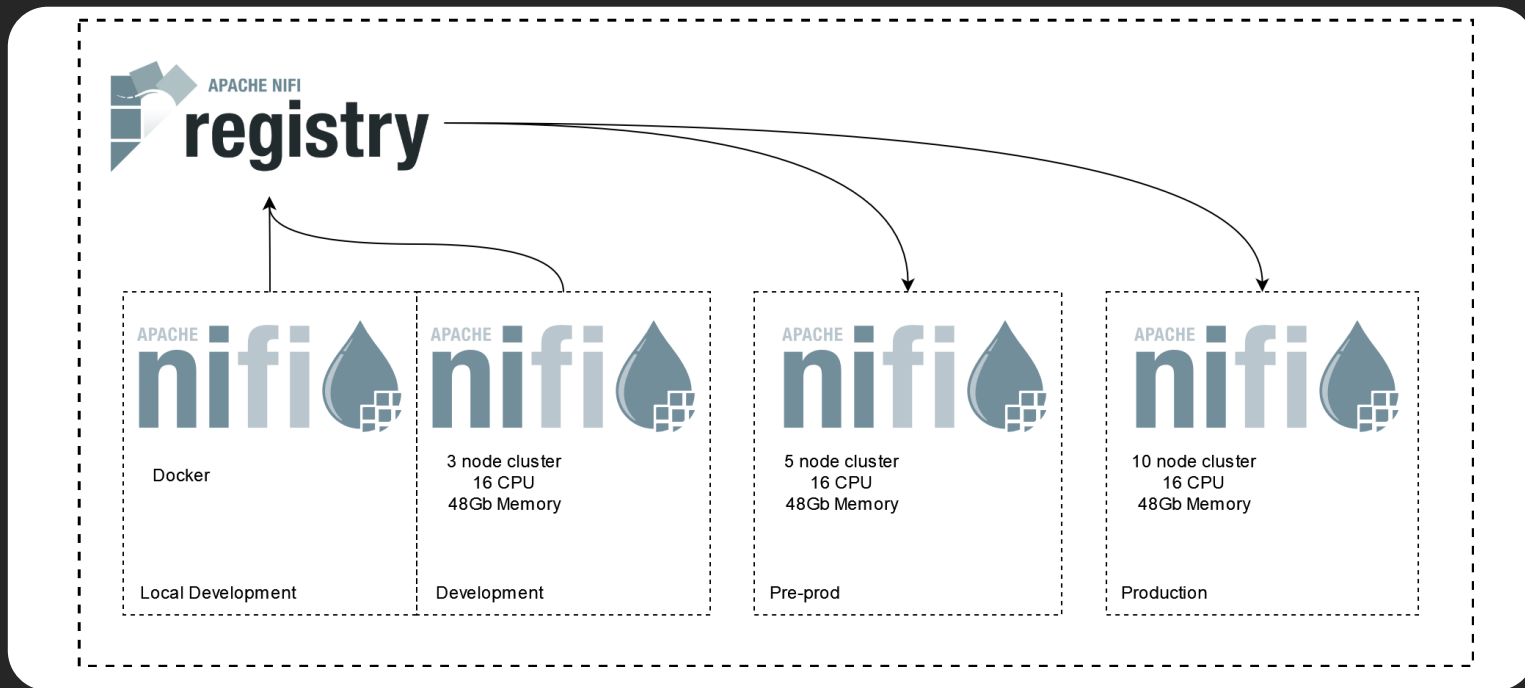


Java приложение с GUI
Может работать в кластерном режиме
Security – LDAP/Kerberos/OpenID/Knox
OpenSource (Apache license)



LM NiFi Architecture As Is

3 кластера:



Leroy Merlin experience

Цифры:

- ~3Tb обрабатываемых входящих сырых данных за день
- Около сотни process group в production от 5 до 20 процессоров в каждой
- 1200 запущенных процессоров
- Текущая версия nifi в проде 1.11.4

Все плюсы и минусы на одном слайде

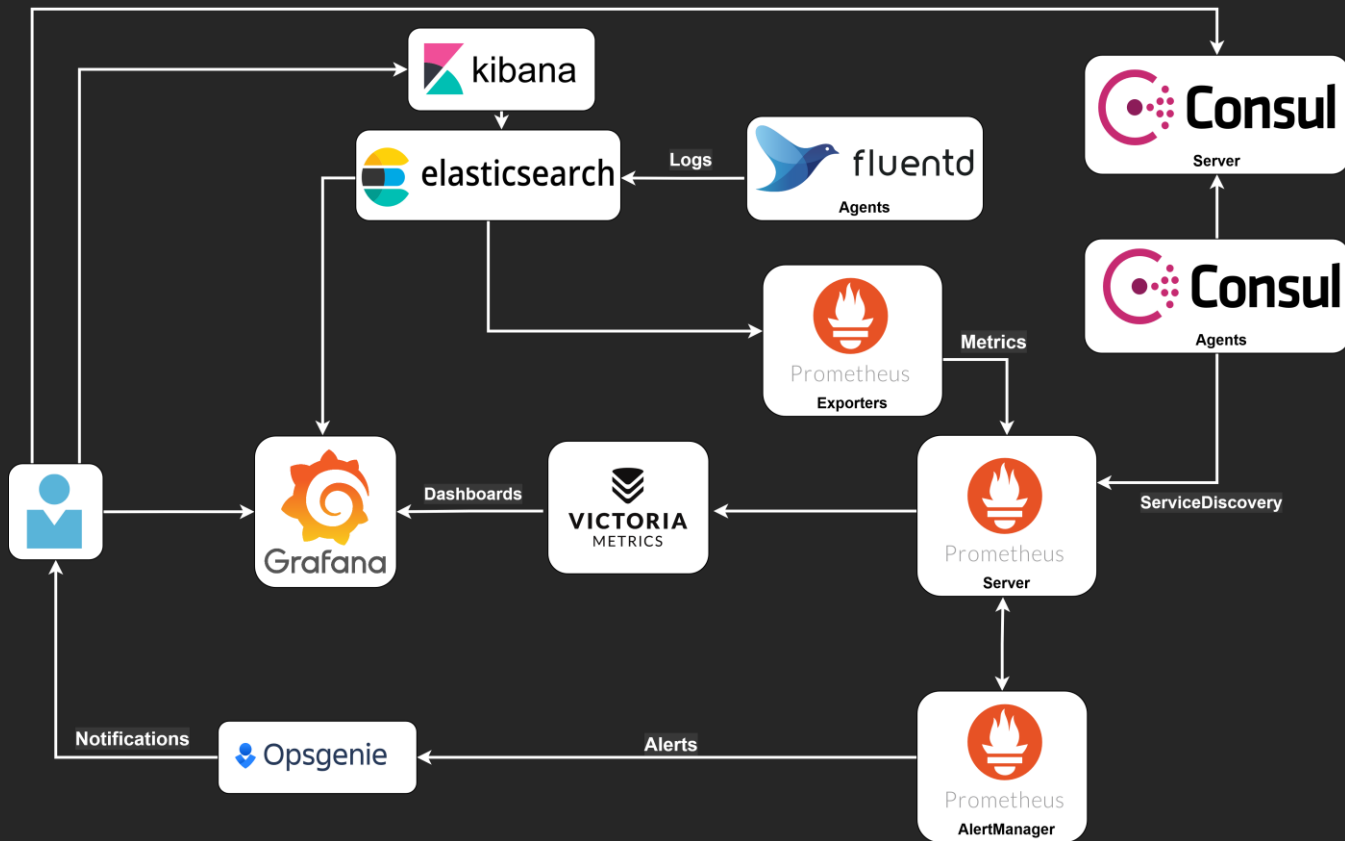
Плюсы:

- Открытость
- Простота (No-Code/Low-Code)
- Гибкость

Минусы:

- Проблемы «коммунальности»
- Выпадение нод из кластера при высокой нагрузке
- Разбалансированность кластера при неравномерном распределении данных

Monitoring Architecture



Базовые метрики

Собираем с помощью стандартного **node-exporter**:

github.com/prometheus/node_exporter

В графанае дашборд **Node Exporter Full**:

grafana.com/grafana/dashboards/1860

По интересующим нас метрикам нужно собрать свой **агрегированный дашборд**

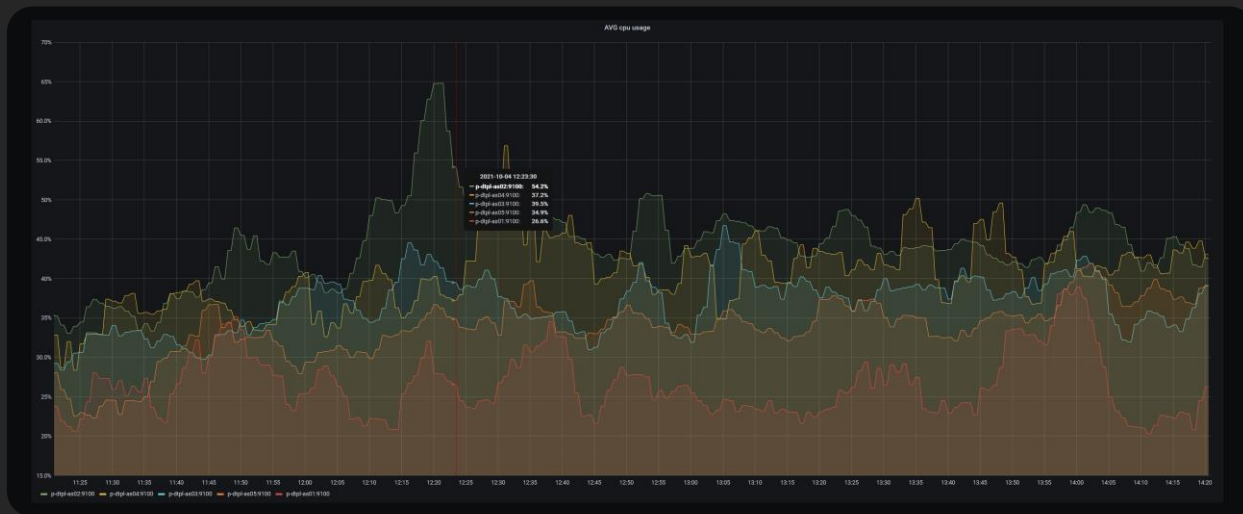
- CPU
- Disks
- IO (read/write)

github.com/diarworld/nifi-exporter-queries/blob/main/prometheus.yml

Базовые метрики

CPU

node_cpu_seconds_total (1 - idle)



Базовые метрики

Cluster unbalanced nodes

1 alert for alertname=nifi-dev-cluster-unbalanced dp_env=dev

[View In AlertManager](#)

[1] Firing

Labels

alertname = nifi-dev-cluster-unbalanced

dp_env = dev

region = dp

service = NiFi

severity = warning

Annotations

dashboard = http://grafana.lmru.tech/d/11HBcV3Zz/nifi?orgId=1&var-instance=dev&var-component=All&var-bulletins_component=All&var-prefix=d

description = NiFi dev unbalanced cluster

summary = CPU load diff between cluster nodes = 74.29314346003201% (>50) for 30min

[Source](#)

Базовые метрики

Disks space alerts

Рекомендуется выносить content, flowfile, provenance storage на отдельные разделы/диски. Самый большой content storage – размер его зависит от количества данных в очередях между процессорами.

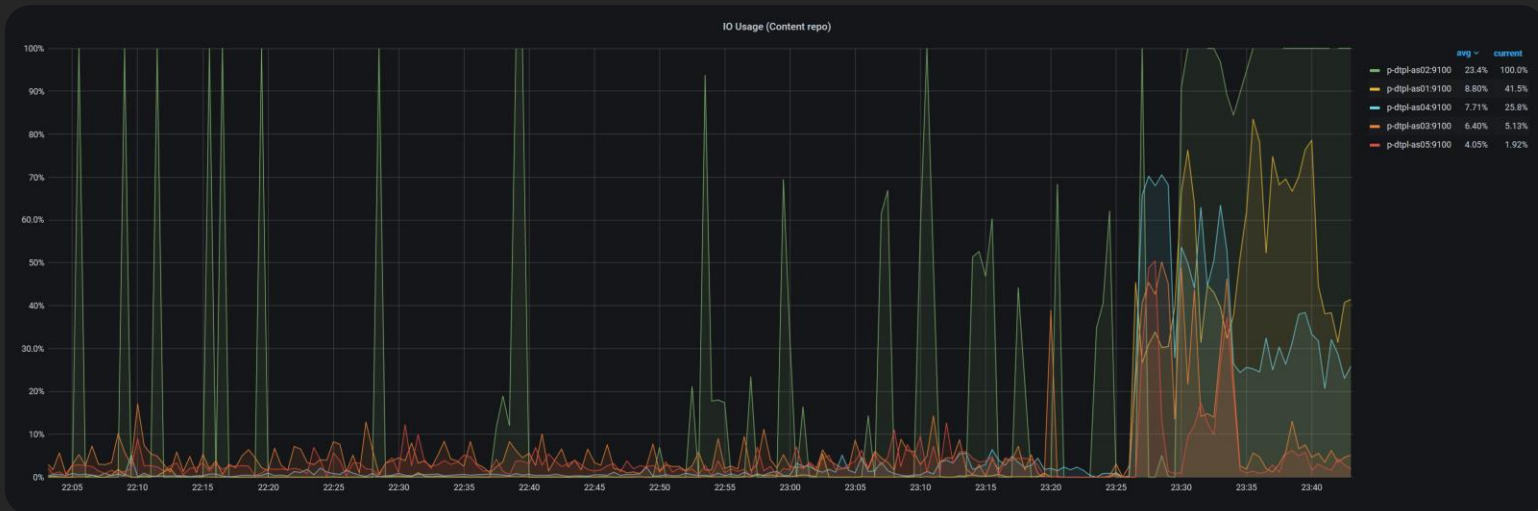
Алертить нужно при >80% использованного места в storages.

github.com/diarworld/nifi-exporter-queries/blob/main/alertmanager-base.yml

Базовые метрики

DiskIO usage

node_disk_io_time_seconds_total



Базовые метрики

DiskIO - Software accelerated network

Вычислительные ресурсы

Платформа ? Intel Ice Lake ▾

vCPU 8
2 96

Гарантированная доля vCPU ? 20% 50% 100%

Для решения любых задач, в том числе для высоконагруженных сервисов.

RAM 16 ГБ
8 ГБ 128 ГБ

Дополнительно ☐ Прерываемая ?

Ускорение сети ☒ Программно-ускоренная сеть ?

Внутренние метрики

Controller settings -> Reporting task -> Add Prometheus reporting task:

NiFi Settings

GENERAL REPORTING TASK CONTROLLER SERVICES REPORTING TASKS REGISTRY CLIENTS

Name	Type	Bundle
BulletinExporterTask	BulletinExporterTask 1.9.2	ru.leroymerlin.data - nifi-reporting-task-nar
PrometheusReportingTask	PrometheusReportingTask 1.11.4	org.apache.nifi - nifi-prometheus-nar

Configure Reporting Task

SETTINGS PROPERTIES COMMENTS

Required field +

Property	Value
Prometheus Metrics Endpoint Port	9092
Instance ID	\$(hostname(true))
Metrics Reporting Strategy	All Components
Send JVM metrics	true
SSL Context Service	No value set
Client Authentication	No Authentication

CANCEL APPLY

Внутренние метрики

Главные метрики:

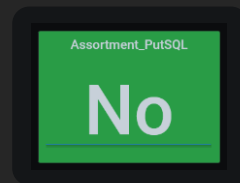
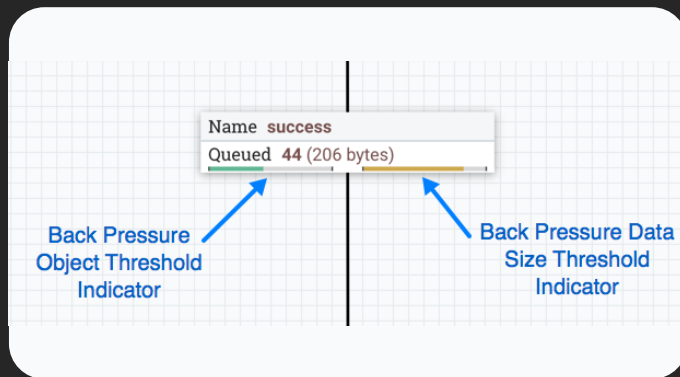
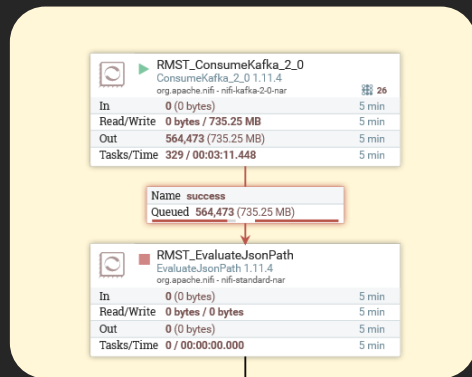
- `nifi_backpressure_enabled`
- `nifi_cluster_node_start_time`
- `nifi_amount_bytes_written`
- `nifi_amount_bytes_read`
- `nifi_size_content_queued_total`
- `nifi_amount_items_queued`

Внутренние метрики

Backpressure:

nifi.apache.org/docs/nifi-docs/html/user-guide.html#Backpressure

nifi_backpressure_enabled



Алертить нужно при наличии backpressure > 30 минут.

Внутренние метрики

Uptime:

nifi_cluster_node_start_time

1.09 day

1.09 day

Uptime
1.09 day

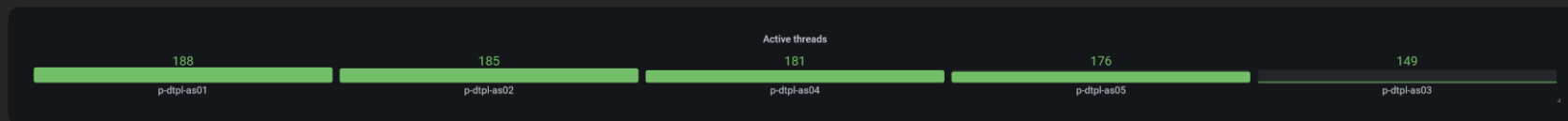
1.09 day

1.09 day

Внутренние метрики

Active threads:

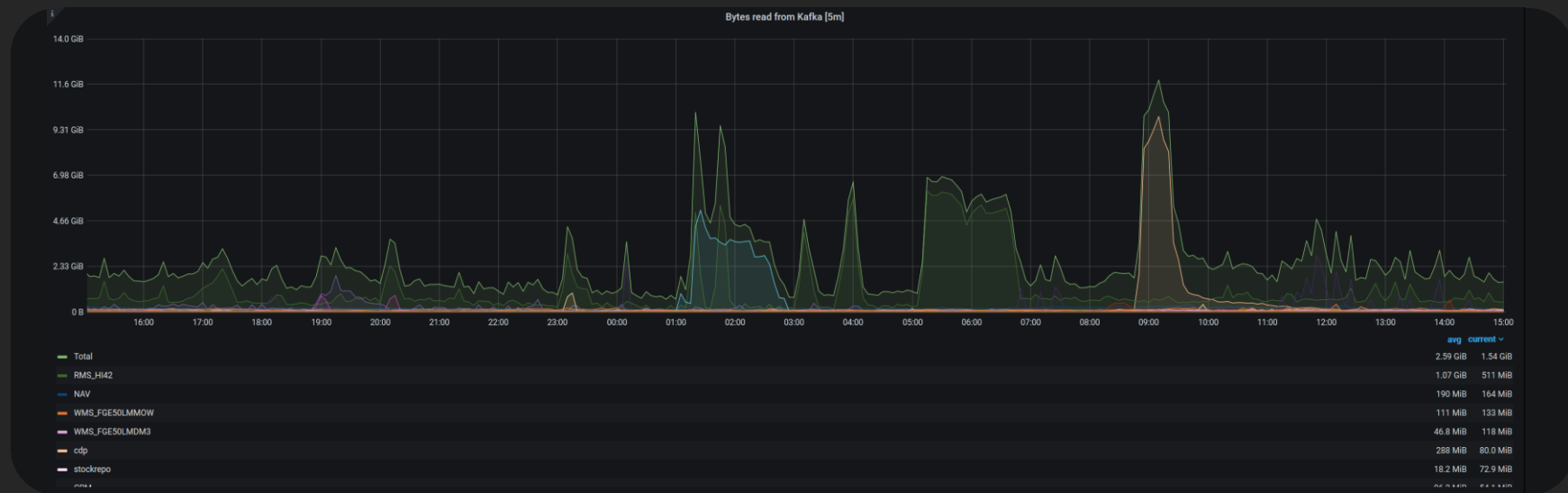
nifi_cluster_status_active_threads



Внутренние метрики

Read/write bytes:

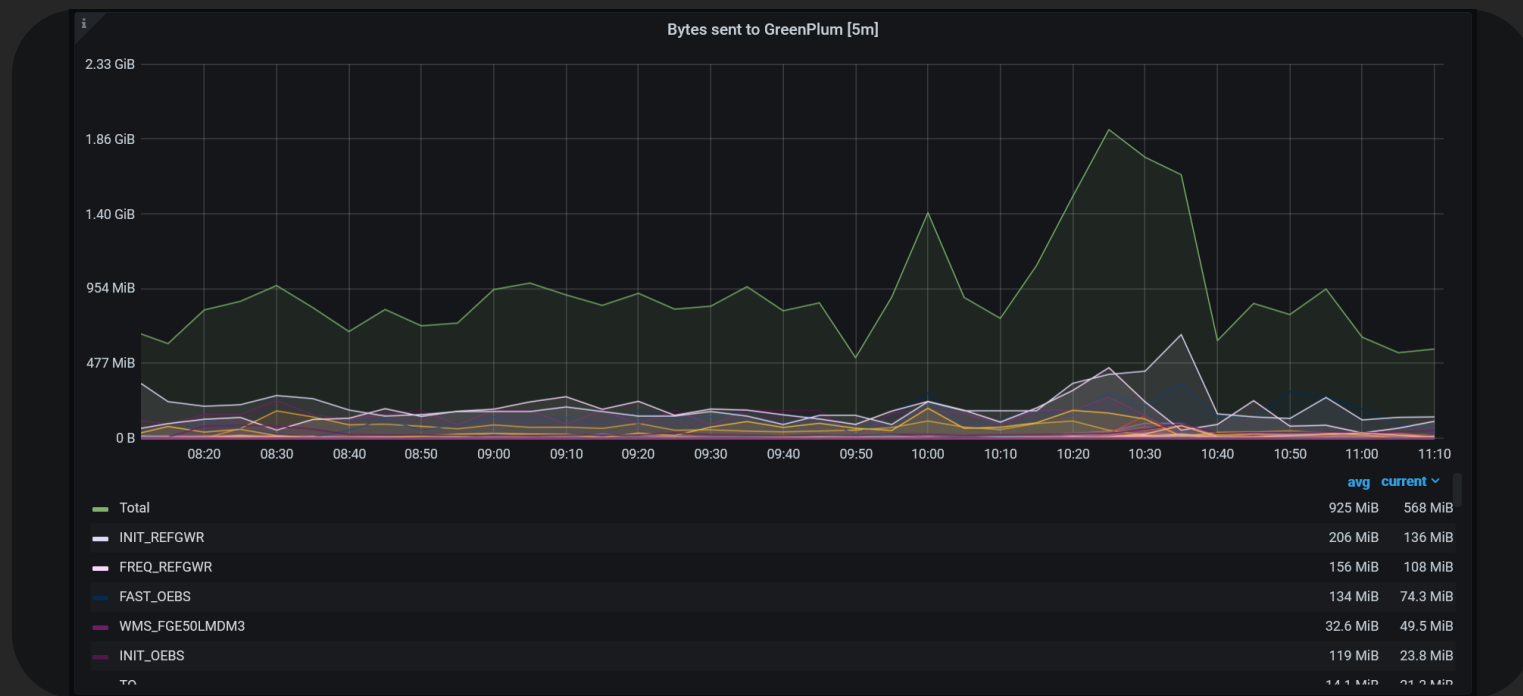
nifi_amount_bytes_written



Внутренние метрики

Read/write bytes:

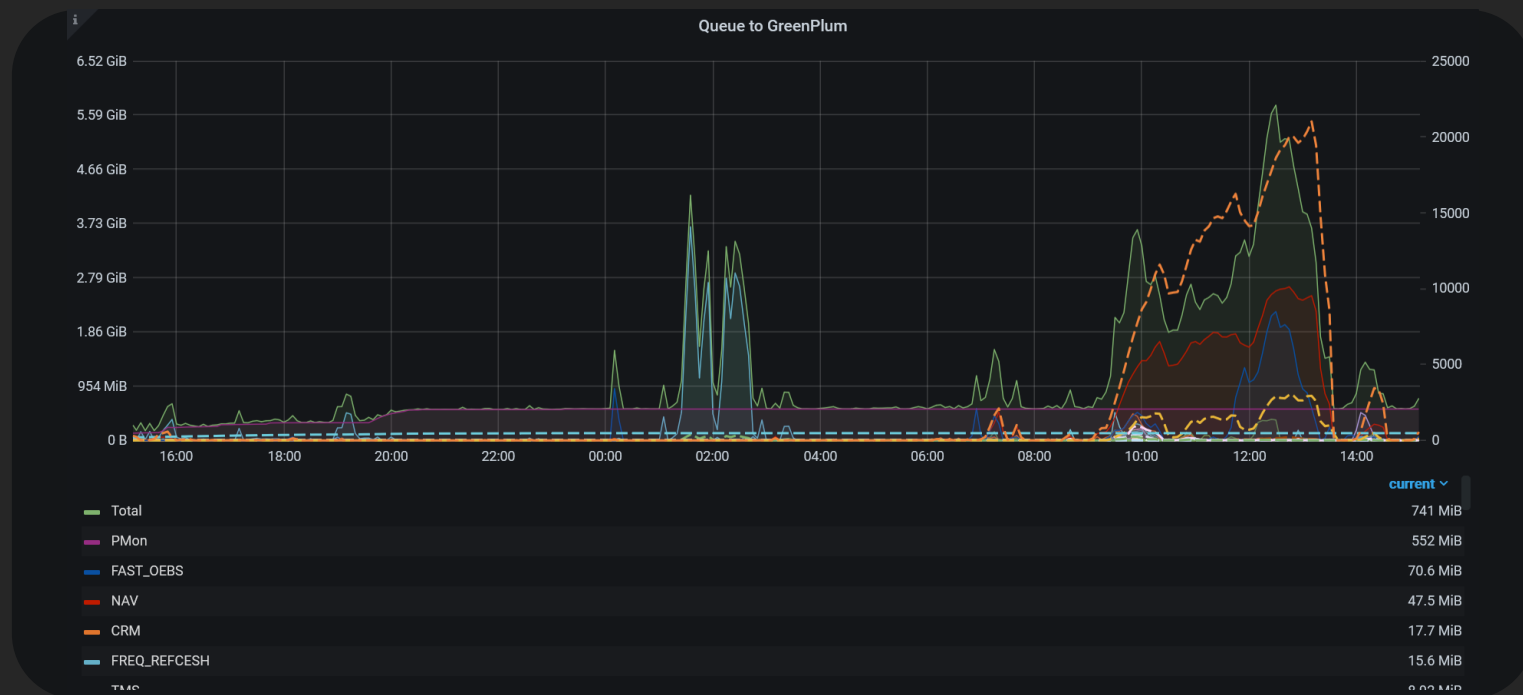
nifi_amount_bytes_read



Внутренние метрики

Queued bytes/items:

$\text{nifi_size_content_queued_total} / \text{nifi_amount_items_queued}$



Внутренние метрики

Алерты:

- Backpressure
- Очереди
- Зависание процессинга
- Uptime
- Активные треды
- nifi_amount_threads_terminated

#50780: [Prometheus]: NiFi prod SV3_NavisionShipments_PutDatabaseRecord processor backpressure

NiFi prod cluster 4 nodes have backpressure enabled for processor SV3_NavisionShipments_PutDatabaseRecord

Operator Help:

<https://portal.support24.online/display/CD/LMDGP++Operator%27s+Help>

Dashboard:

<http://grafana.data.lmru.tech/d/sRH-5RoWk/nifi>

Show more

Priority

P4

Tags

NiFi,

SV3_NavisionShipments_PutDatabaseRecord, dp, info, nifi-putsqldb-backpressure, prod

Routed Teams

Data - Data Platform

Unacknowledge

Close

Other actions...

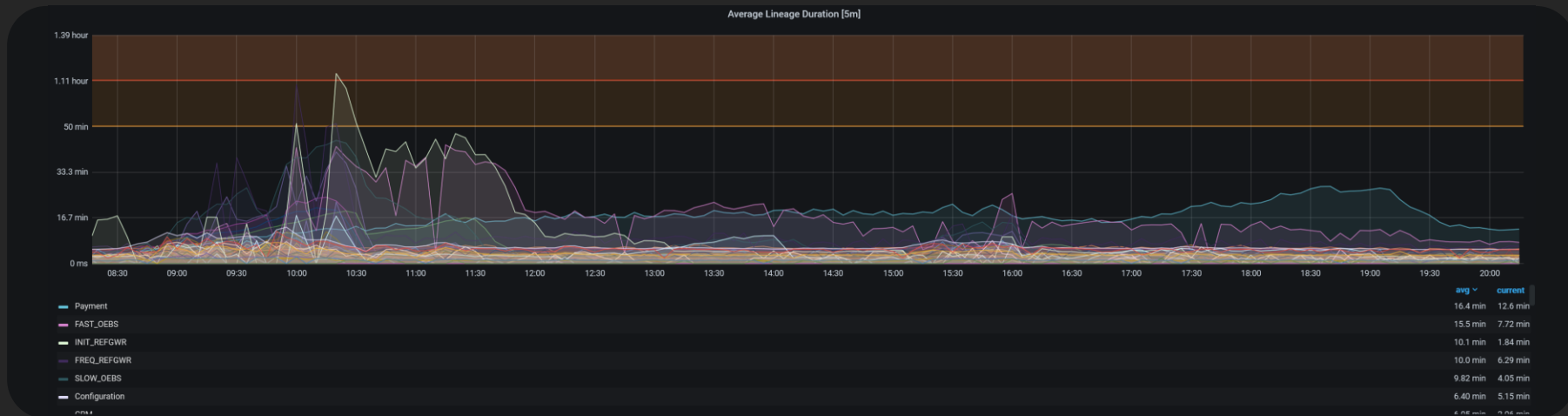


Внутренние метрики

SLO

nifi_average_lineage_duration

The average lineage duration (in milliseconds) for all flow file processed by this component



Blackbox exporter

github.com/prometheus/blackbox_exporter

Мониторим и алертим по недоступности:

- SSH портов
- WEB портов
- Registry порта

#11183: [Prometheus]: GP port is unavailable on prod

Connection to database on 10.220.51.23:5432 is lost

Operator Help:

<https://portal.support24.online/display/CD/LMDGP++Operator%27s+Help>

Dashboard:

<http://grafana.lmru.tech/d/sRH-5RoWk/greenplum>

Show more

Priority

P1

Tags

10.220.51.23:5432, availability,
blackbox_tcp_gp_port, critical, database,
dp, gp-connection-is-lost, greenplum,
greenplum_prod, prod

Routed Teams

Data - Data Platform

#11190: [Prometheus]: PXF has no access at 10.220.51.5:5888

PXF port not accessible at 10.220.51.5:5888

Operator Help:

<https://portal.support24.online/display/CD/LMDGP++Operator%27s+Help>

Dashboard:

<http://grafana.lmru.tech/>

Show more

Priority

P3

Tags

10.220.51.5:5888, PXF, availability,
blackbox_pxf_prod, dp, greenplum_prod,
prod, pxf-port-prod, service, warning

Routed Teams

Data - Data Platform

#726: [Prometheus]: GP node not access via SSH at 10.220.51.24:22

SSH port inaccessible at 10.220.51.24:22

Operator Help:

<https://portal.support24.online/display/CD/LMDGP++Operator%27s+Help>

Dashboard:

<http://grafana.lmru.tech/d/sRH-5RoWk/greenplum>

Show more

Priority

P1

Tags

10.220.51.24:22, Greenplum, availability,
blackbox_tcp_gpprod, critical, dp, gp-node-
not-access-ssh, greenplum_prod, prod,
service

Routed Teams

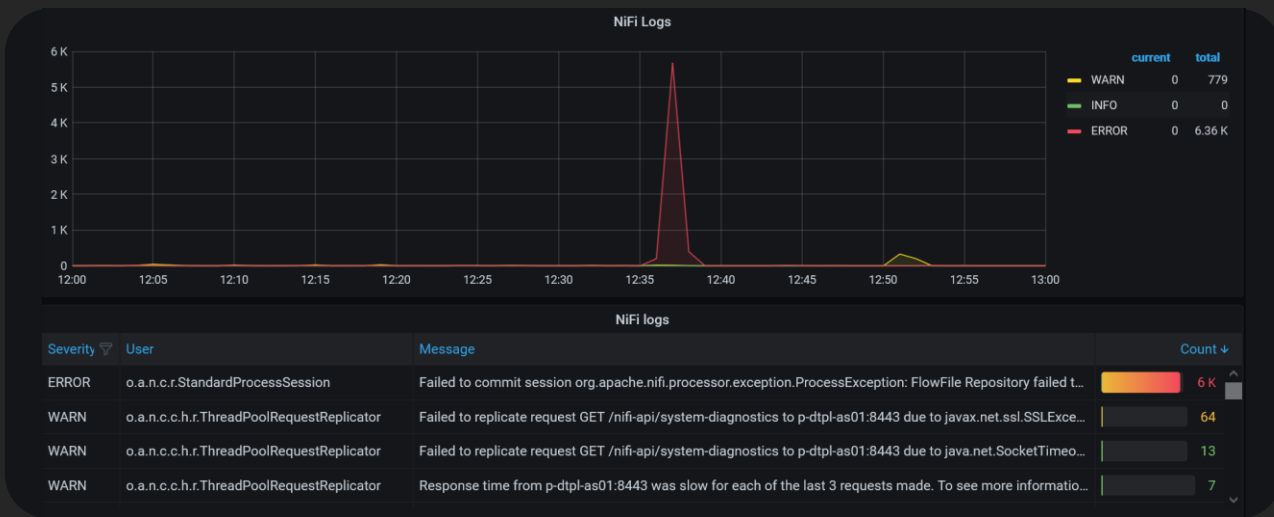
Data - Data Platform

Elasticsearch query exporter

github.com/braedon/prometheus-es-exporter

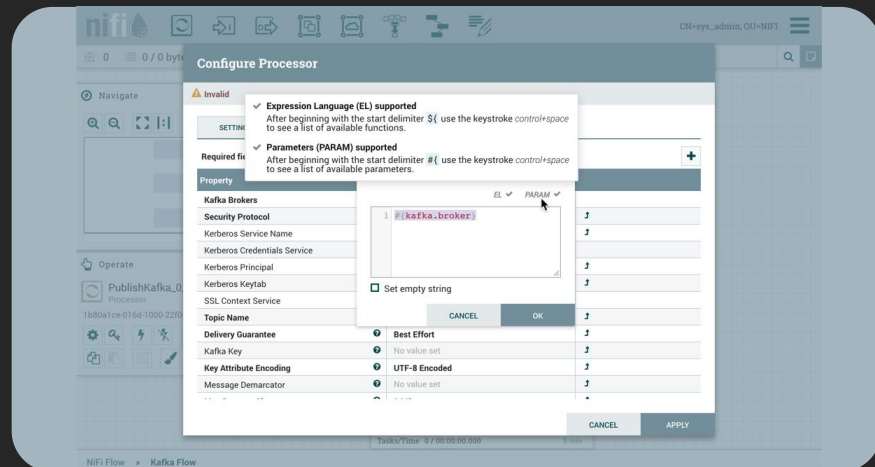
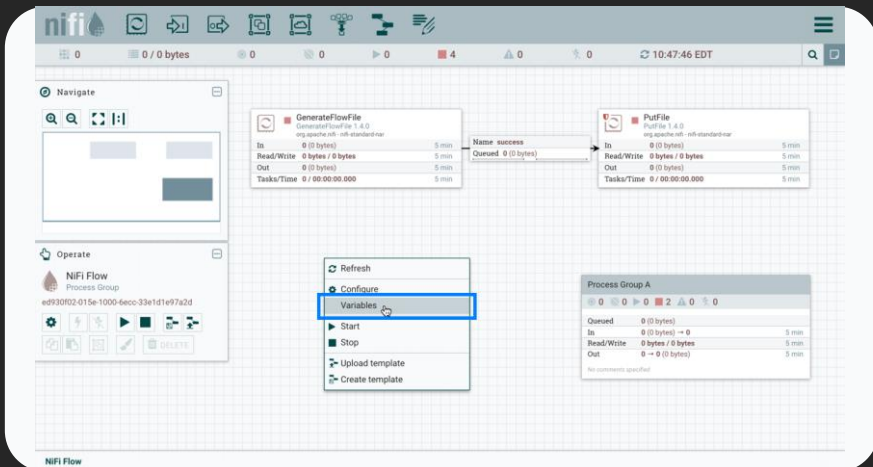
Мониторим и алертим по:

- Out Of Memory в логах
- Резко возросшему числу ERROR в логах



Best practices

Variables vs Parameters Context



Best practices

ControllerServices

Add Controller Service

Source

all groups ▾

Displaying 10 of 28

Serv

Type	Version	Tags
AWSCredentialsProviderControllerService	1.2.0	credentials, provider, aws
CouchbaseClusterService	1.2.0	database, couchbase, connection, ...
DistributedMapCacheClientService	1.2.0	cluster, cache, distributed, state, m...
DistributedMapCacheServer	1.2.0	cluster, server, cache, key/value, dis...
DistributedSetCacheClientService	1.2.0	cluster, cache, set, distributed, state
DistributedSetCacheServer	1.2.0	server, cache, set, distributed, disti...
GCPCredentialsControllerService	1.2.0	gcp, credentials, provider
HBase_1_1_2_ClientService	1.2.0	client, hbase
StandardSSLContextService	1.2.0	p12, jks, certificate, keystore, trusts...
JettyWebSocketServer	1.2.0	server, Jetty, WebSocket

avro cache cluster
comma connection
csv database dbcp
delimited
distributed jdbc
json parse pooling
provider reader
record recordset
result row separated
server set store
writer

StandardSSLContextService 1.2.0 [org.apache.nifi - nifi-ssl-context-service-nar](#)

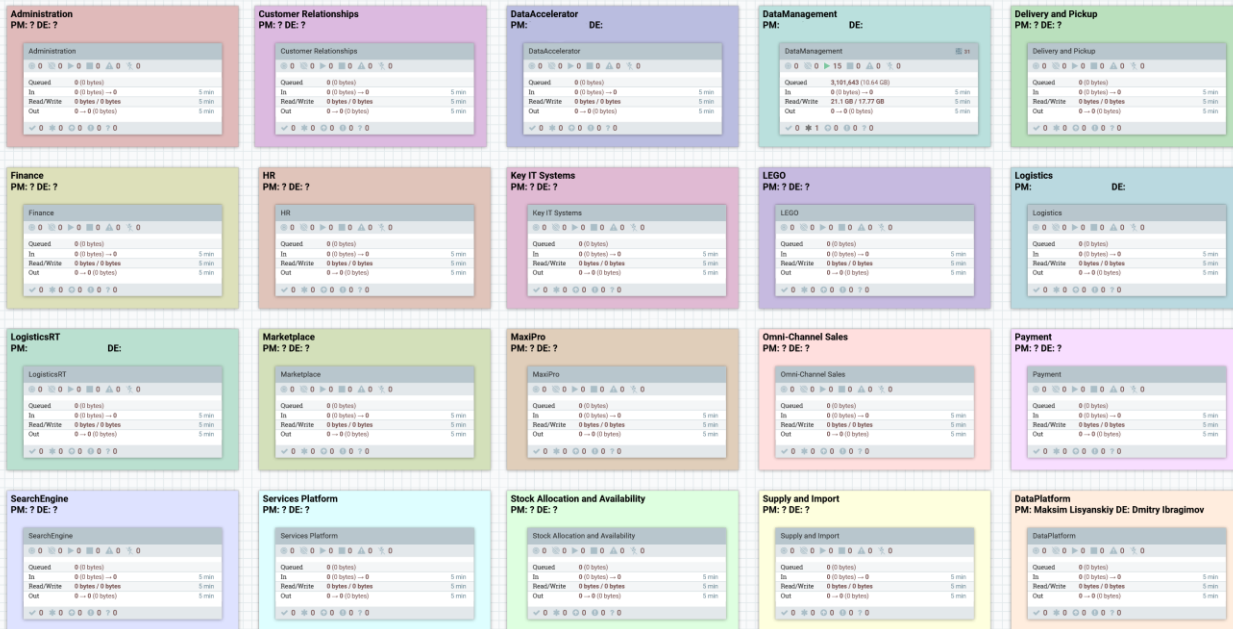
Standard implementation of the SSLContextService. Provides the ability to configure keystore and/or truststore properties once and reuse that configuration throughout the application

CANCEL

ADD

Best practices

Naming conventions



Best practices

CI/CD pipelines for shared libraries/scripts

✓ / nifi / dp--nifi-resources #21 develop Pipeline Changes Tests Artifacts Logout X

Branch: develop 6m 18s No changes
Commit: 3298849 2 months ago Push event to branch develop

Start Checkout Unit & Integration Tests Publish Deploy End

Unit & Integration Tests - 2m 47s [Restart Unit & Integration Tests](#)

✓	> Get contextual object from internal APIs	<1s
✓	> Checks if running on a Unix-like node	<1s
✓	> docker inspect -f . gradle:test - Shell Script	<1s
✓	> Checks if running on a Unix-like node	<1s
✓	> docker pull gradle:test - Shell Script	15s
✓	> Running Tests - Print Message	<1s
✓	> ./gradlew cleanTest test --no-daemon - Shell Script	2m 28s
✓	> **/build/test-results/test/*.xml - Archive JUnit-formatted test results	<1s
✓	> Tell Slack the build went okay - Print Message	<1s

✓ / nifi / dp--nifi-resources #21 develop Pipeline Changes Tests Artifacts Logout X

Branch: develop 6m 18s No changes
Commit: 3298849 2 months ago Push event to branch develop

Start Checkout Unit & Integration Tests Publish Deploy End

Publish - 1m 59s [Restart Publish](#)

✓	> Get contextual object from internal APIs	<1s
✓	> Checks if running on a Unix-like node	<1s
✓	> docker inspect -f . gradle:test - Shell Script	<1s
✓	> ./gradlew artifactoryPublish --no-daemon - Shell Script	1m 55s
✓	> Branch name develop - Print Message	<1s

Best practices

Avoid CtrlC - CtrlV

[EXAMPLE] [WH] TEST

0 0 0 0 0 610

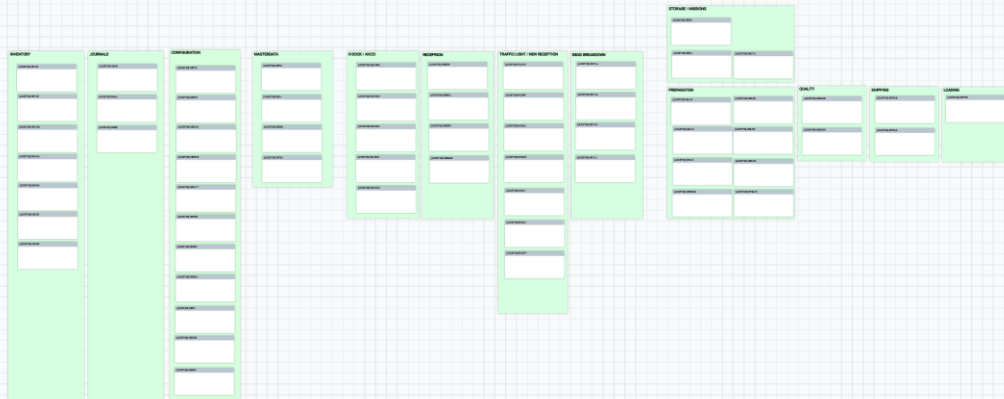
Queued 0 (0 bytes)

In 0 (0 bytes) → 0 5 min

Read/Write 0 bytes / 0 bytes 5 min

Out 0 → 0 (0 bytes) 5 min

0 0 0 0 0 0



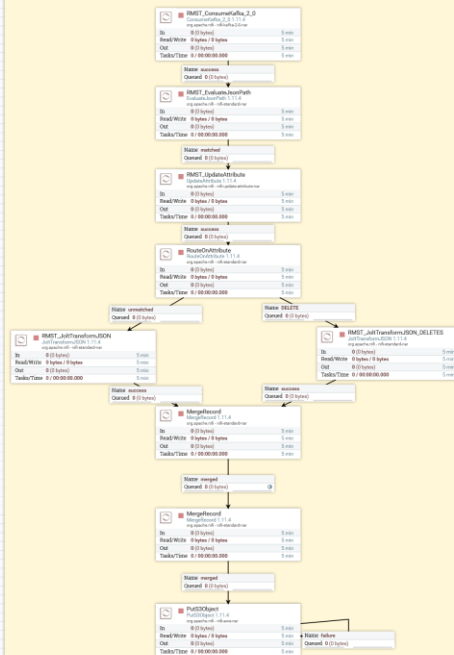
Best practices

Упрощайте!

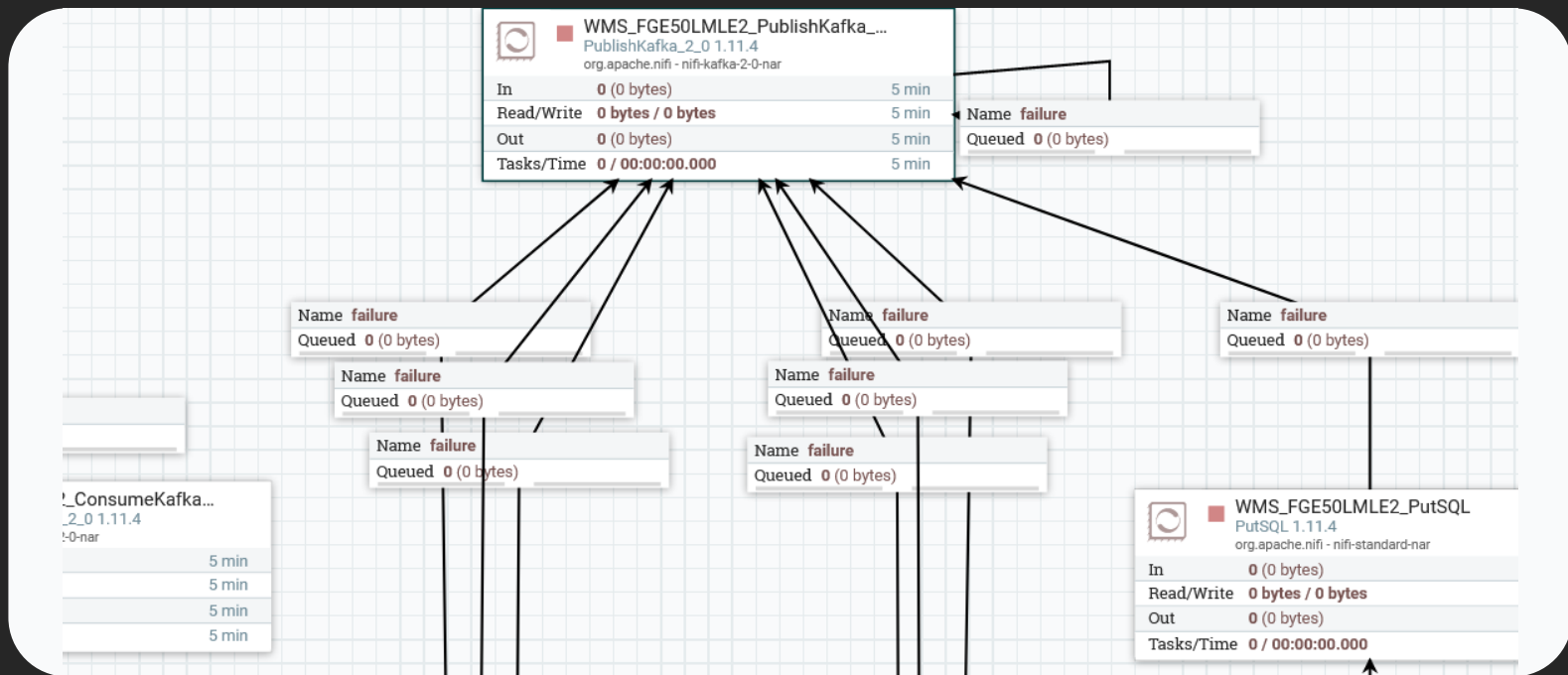
S3 with attribute



S3 with NK



Best practices: Monitoring Driven Development



Best practices

- Избегать ExecuteScript/ExecuteStreamCommand

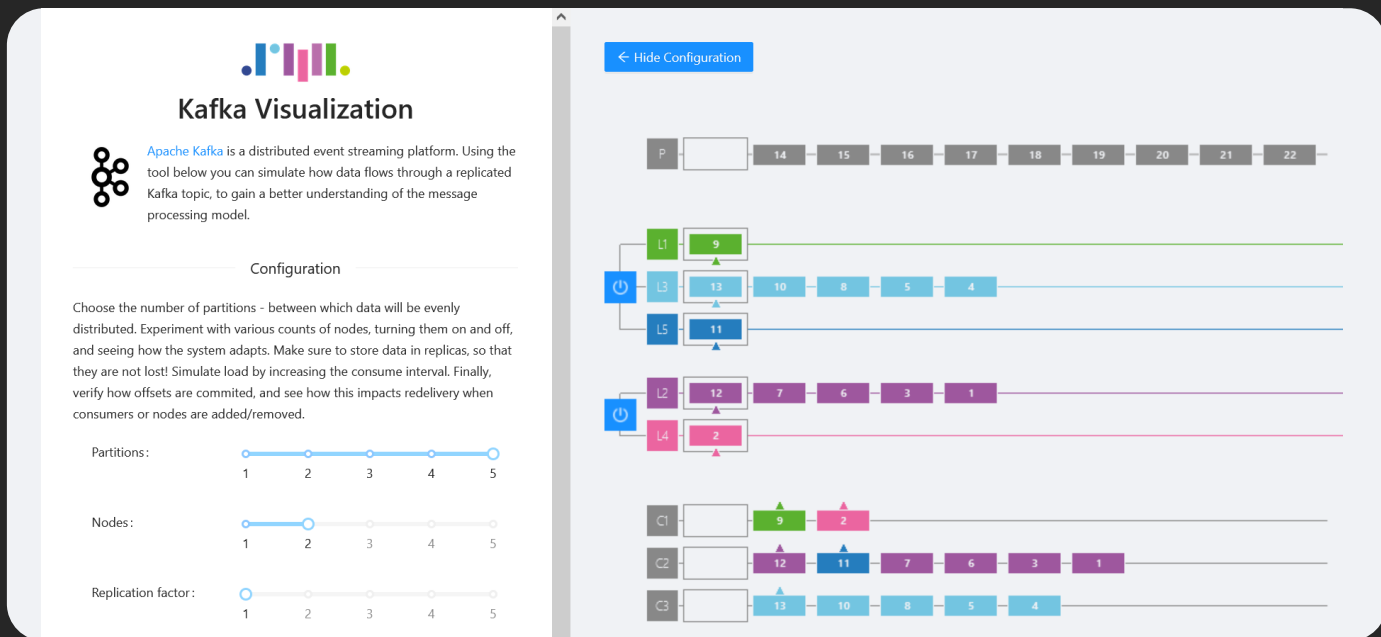
Best practices

- Избегать ExecuteScript/ExecuteStreamCommand
community.cloudera.com/t5/Community-Articles/ExecuteScript-Cookbook-part-1/ta-p/248922

Groovy > Javascript > Jython > JRuby > ExecuteProcess

Best practices

- Контролировать число потоков ConsumeKafka



Best practices

- Контролировать число потоков ConsumeKafka



Kafka Visualization

 **Apache Kafka** is a distributed event streaming platform. Using the tool below you can simulate how data flows through a replicated Kafka topic, to gain a better understanding of the message processing model.

Configuration

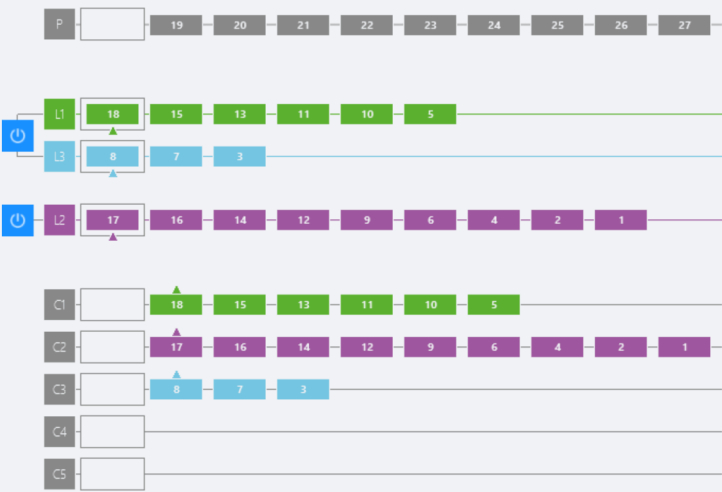
Choose the number of partitions - between which data will be evenly distributed. Experiment with various counts of nodes, turning them on and off, and seeing how the system adapts. Make sure to store data in replicas, so that they are not lost! Simulate load by increasing the consume interval. Finally, verify how offsets are committed, and see how this impacts redelivery when consumers or nodes are added/removed.

Partitions:

Nodes:

Replication factor:

[← Hide Configuration](#)



The diagram illustrates a Kafka topic with 27 partitions (P) distributed across three nodes (L1, L2, L3). Node L1 (green) has 18 partitions (18, 15, 13, 11, 10, 5). Node L2 (purple) has 17 partitions (17, 16, 14, 12, 9, 6, 4, 2, 1). Node L3 (blue) has 8 partitions (8, 7, 3). The diagram also shows five consumers (C1-C5) and their respective partitions: C1 (18, 15, 13, 11, 10, 5), C2 (17, 16, 14, 12, 9, 6, 4, 2, 1), C3 (8, 7, 3), C4 (empty), and C5 (empty).

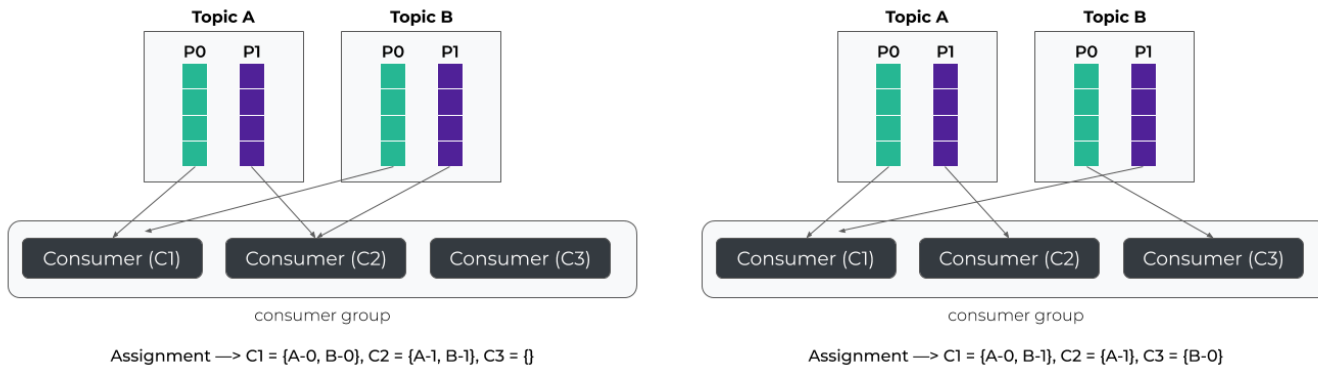
Best practices

- Lifehack: ConsumeKafka change partition.assignment.strategy

Name	Partition	Member	Offset	Lag
data.partition.distribution.example	0	nifi1	806674919	99350015
data.partition.distribution.example	1	nifi1	822090283	87257685
data.partition.distribution.example	2	nifi1	799584443	98361894
data.partition.distribution.example	3	nifi2	37324287	0
data.partition.distribution.example	4	nifi2	37570151	0

Best practices

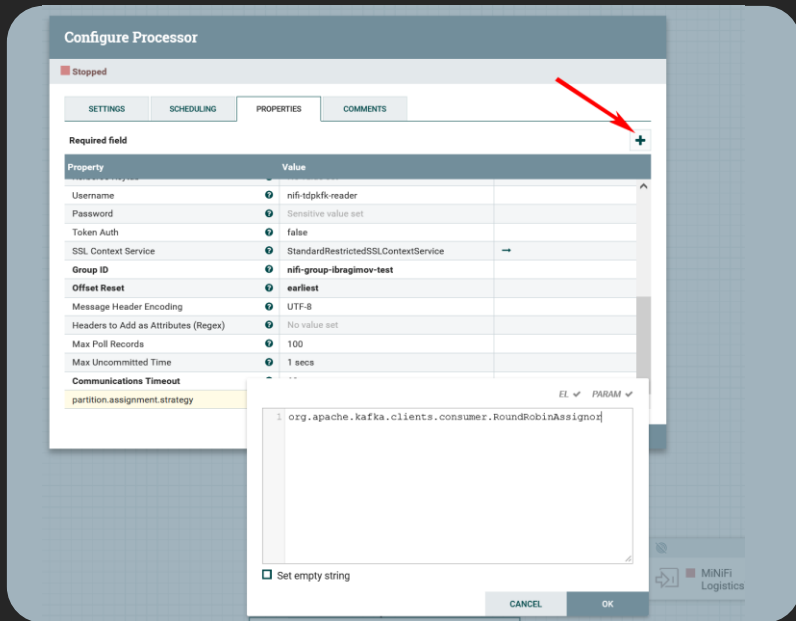
- Lifehack: ConsumeKafka change partition.assignment.strategy



RangeAssignor (default) $\longrightarrow$ RoundRobinAssignor

Best practices

- Lifehack: ConsumeKafka change partition.assignment.strategy
+ partition.assignment.strategy -> org.apache.kafka.clients.consumer.RoundRobinAssignor



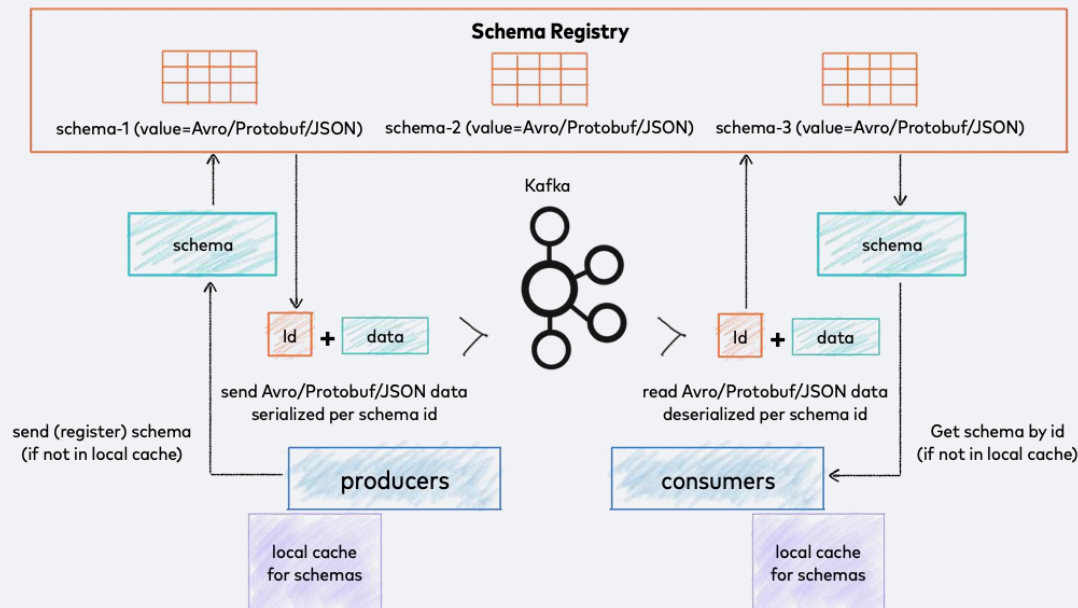
Best practices

- Lifehack: ConsumeKafka change partition.assignment.strategy

Name	Partition	Member	Offset	Lag
data.partition.distribution.example	0	nifi1	806682225	99342709
data.partition.distribution.example	1	nifi2	822095373	87252595
data.partition.distribution.example	2	nifi3	799590047	98356290
data.partition.distribution.example	3	nifi4	37324287	0
data.partition.distribution.example	4	nifi5	37570151	0

Best practices

- Использовать Kafka Schema Registry



Best practices

- Использовать Kafka Schema Registry

issues.apache.org/jira/browse/NIFI-8296

В Nifi < 1.14.0 невозможно получить имя схемы из registry по её id из сообщения

Best practices

- Использовать Kafka Schema Registry

Configure Processor

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Name: LES_ConsumeKafkaRecord ☒ Enabled

Id: 0eb33a12-007d-194f-ae0e-6eb63a86b035

Type: ConsumeKafkaRecord_2_0 1.11.4

Bundle: org.apache.nifi - nifi-kafka-2-0-nar

Penalty Duration: 30 sec Yield Duration: 1 sec

Bulletin Level: WARN

Automatically Terminate Relationships

☒ parse.failure

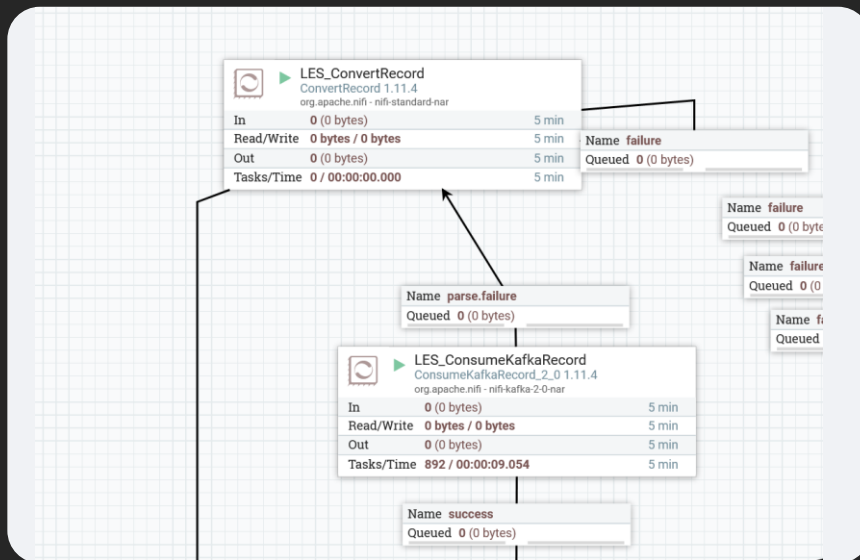
If a message from Kafka cannot be parsed using the configured Record Reader, the contents of the message will be routed to this Relationship as its own individual FlowFile.

☐ success

FlowFiles received from Kafka. Depending on demarcation strategy it is a flow file per message or a bundle of messages grouped by topic and partition.

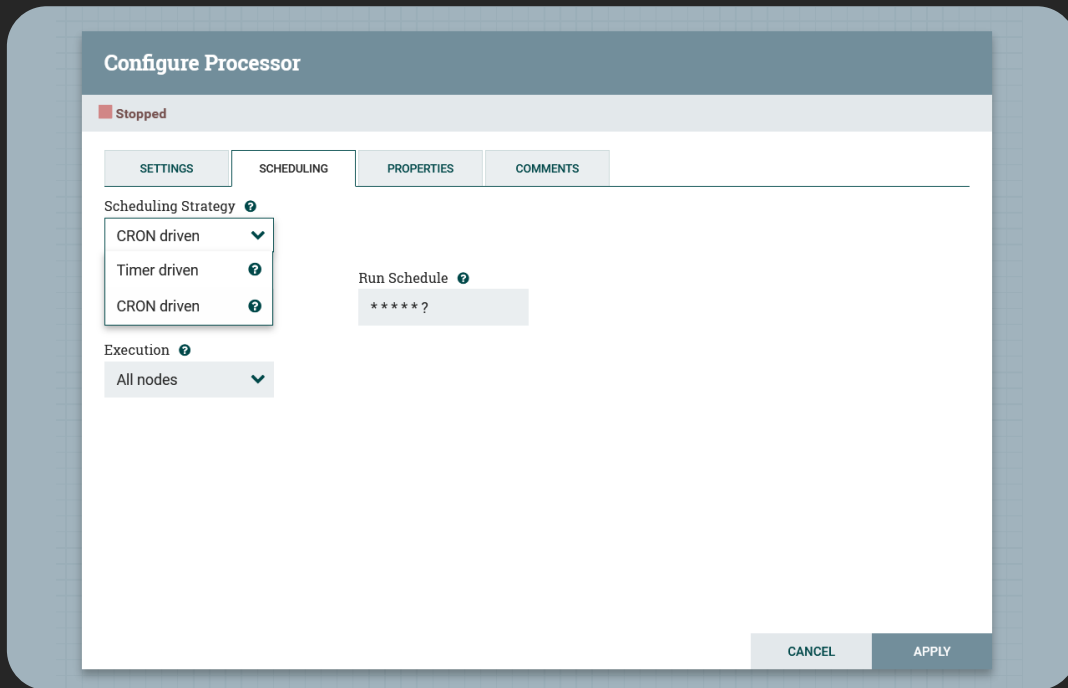
CANCEL APPLY

AVOID!



Best practices

- Не злоупотреблять Cron-driven schedule/Execution Primary Only



The screenshot shows the 'Configure Processor' dialog box in Apache NiFi, specifically the 'SCHEDULING' tab. The dialog has a status bar at the top indicating 'Stopped'. Below this are four tabs: 'SETTINGS', 'SCHEDULING', 'PROPERTIES', and 'COMMENTS'. The 'SCHEDULING' tab is active. It contains a 'Scheduling Strategy' dropdown menu with options 'CRON driven', 'Timer driven', and 'CRON driven'. The 'CRON driven' option is selected. To the right of the dropdown is a 'Run Schedule' field with a cron expression '*****?'. At the bottom of the dialog are 'CANCEL' and 'APPLY' buttons.

Configure Processor

Stopped

SETTINGS SCHEDULING PROPERTIES COMMENTS

Scheduling Strategy ?

CRON driven

Timer driven

CRON driven

Run Schedule ?

*****?

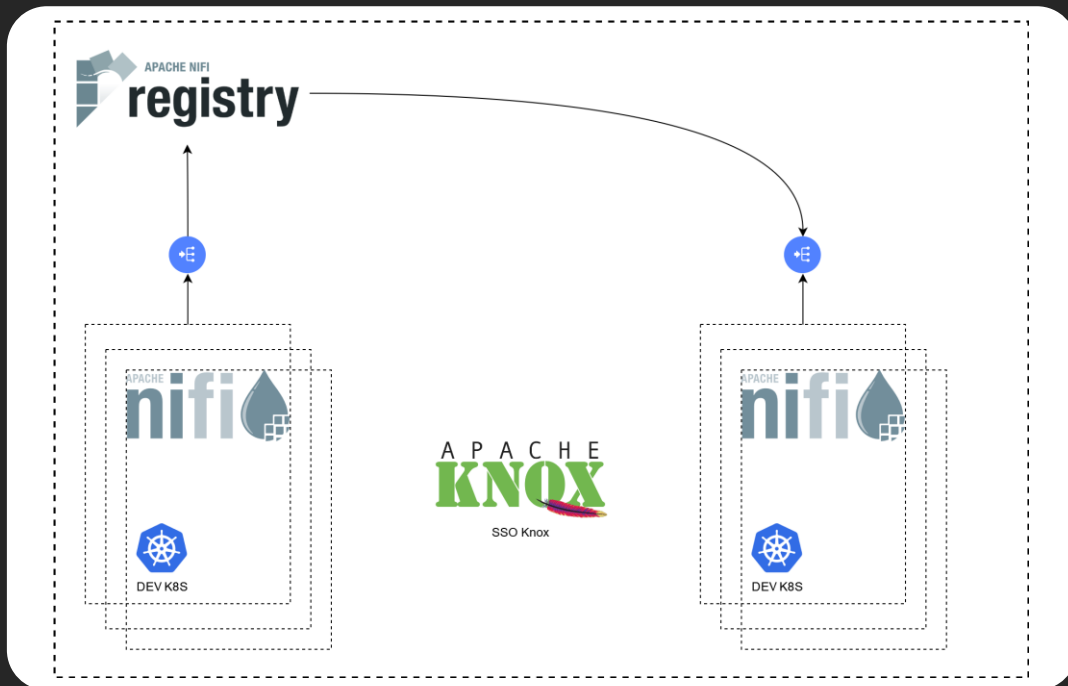
Execution ?

All nodes

CANCEL APPLY

Best NiFi Architecture

«Data Mesh» архитектура



Подводим итоги

Инсайты и выводы

При построении правильной архитектуры и мониторинга NiFi является хорошим решением **потоковой** обработки с **низким порогом** вхождения.

Спасибо за внимание!



Дмитрий Ибрагимов

Data Platform Site Reliability Engineer

Dmitry.Ibragimov@leroymerlin.ru

Github/telegram: @diarworld

IN
DATA
WE
TRUST