СЛЁРМ Обучение для инженеров и технических лидеров

slurm.io

Дата-инженер

Архитектура, термины, интерфейс и базовый функционал

Николай Акимов

Николай Акимов

О спикере

- Работаю в компании TaskData (ООО «ТаскДата») инженером по внедрению
- Профессионально в IT уже больше 20 лет
- Начинал как любитель с паяльником в 1989 году со сборки своего компьютера
- В корпоративной IT среде начал карьеру системным администратором в Коринтия Невский Палас Отель. Далее Гранд Отель Европа. В итоге проработал почти 15 лет в крупной международной компании HRS Hospitality & Retail Systems (платиновый партнёр Oracle)
- Большой опыт работы с БД Oracle вырастает из гостиничного продукта Oracle Opera PMS и опыт проведения тренингов по обучению персонала
- Активно участвую в опенсорс проектах на github. C++, Groovy
  Моя страница https://github.com/vomikan
- Активно участвую в группе по поддержке сообщества NiFi в Телеграм
  https://t.me/nifiusers

План занятия

1. Знакомство с Apache NiFi
2. Установка Apache NiFi
3. Архитектура Apache NiFi
4. Понятия и компоненты Apache NiFi
5. Интерфейс Apache NiFi
6. Классификация процессоров

Представьте себе…

- Легкая интеграция со множеством сторонних приемников и источников данных
- Графический веб-интерфейс
- 2 режима работы с данными
- Сервис версионирования Registry,
- Мощность и масштабируемость
- Поддержка SQL
- Активное развитие и поддержка сообщества

Преимущества NiFi

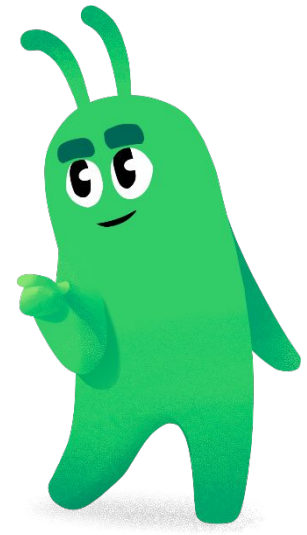
- Легкая интеграция со множеством сторонних приемников и источников данных
- Графический веб-интерфейс
- 2 режима работы с данными
- Сервис версионирования Registry,
- Активное развитие и поддержка сообщества
- Мощность и масштабируемость
- Поддержка SQL

О недостатках

- Неоднозначность гарантированной доставки сообщений
- Чувствительность к отключению узла от кластера
- Проблема с сохранением состояния в случае переключения основного узла

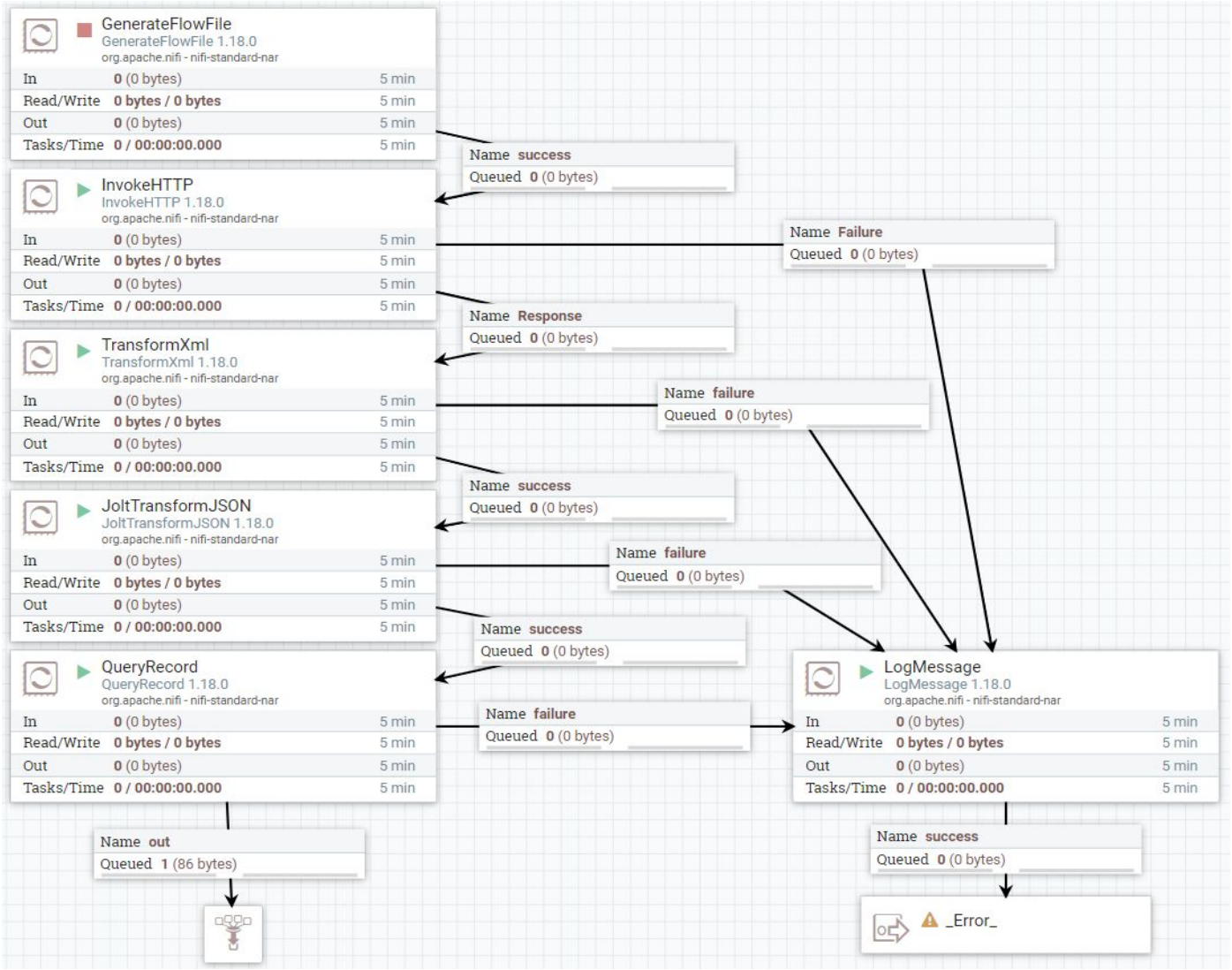
1

Знакомство с Apache NiFi



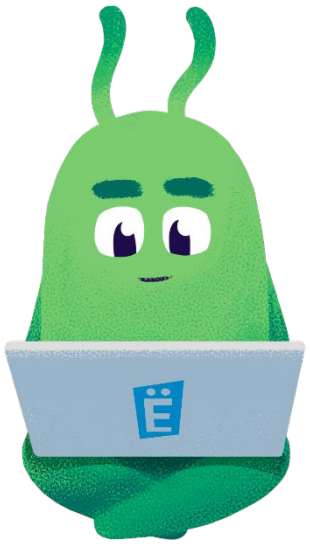
Apache NiFi

Как выглядит



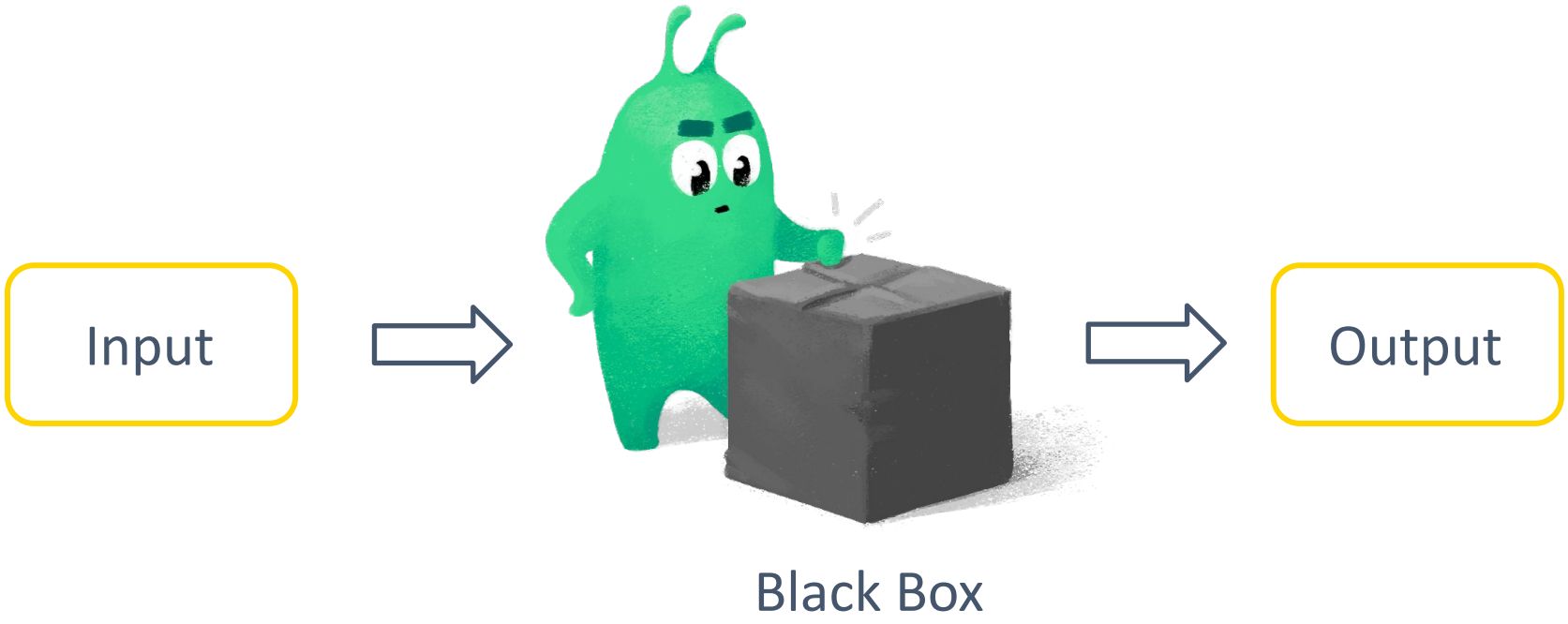
Что такое Apache NiFi

- NiFi – Niagara Files
- NiFi — open source ETL инструмент, умеет работать со множеством систем, включая системы класса Big Data и Data Warehouse
- Hive, HBase, Solr, Cassandra, MongoDB, ElasticSearch, Kafka, RabbitMQ, HDFS, HTTPS, SFTP и другие
- Особенность — кроссплатформенность: удобно устанавливается и работает как в Windows, так и Linux-системах



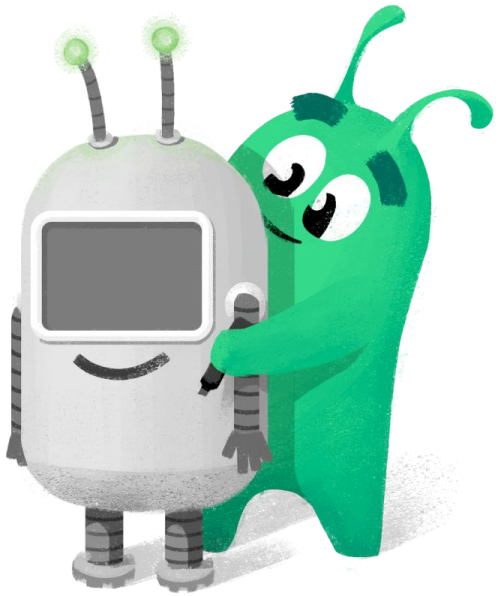
Что такое Apache NiFi

NiFi опирается на концепцию «Flow Based Programming» (FBP)



2

Установка



Установка Airflow

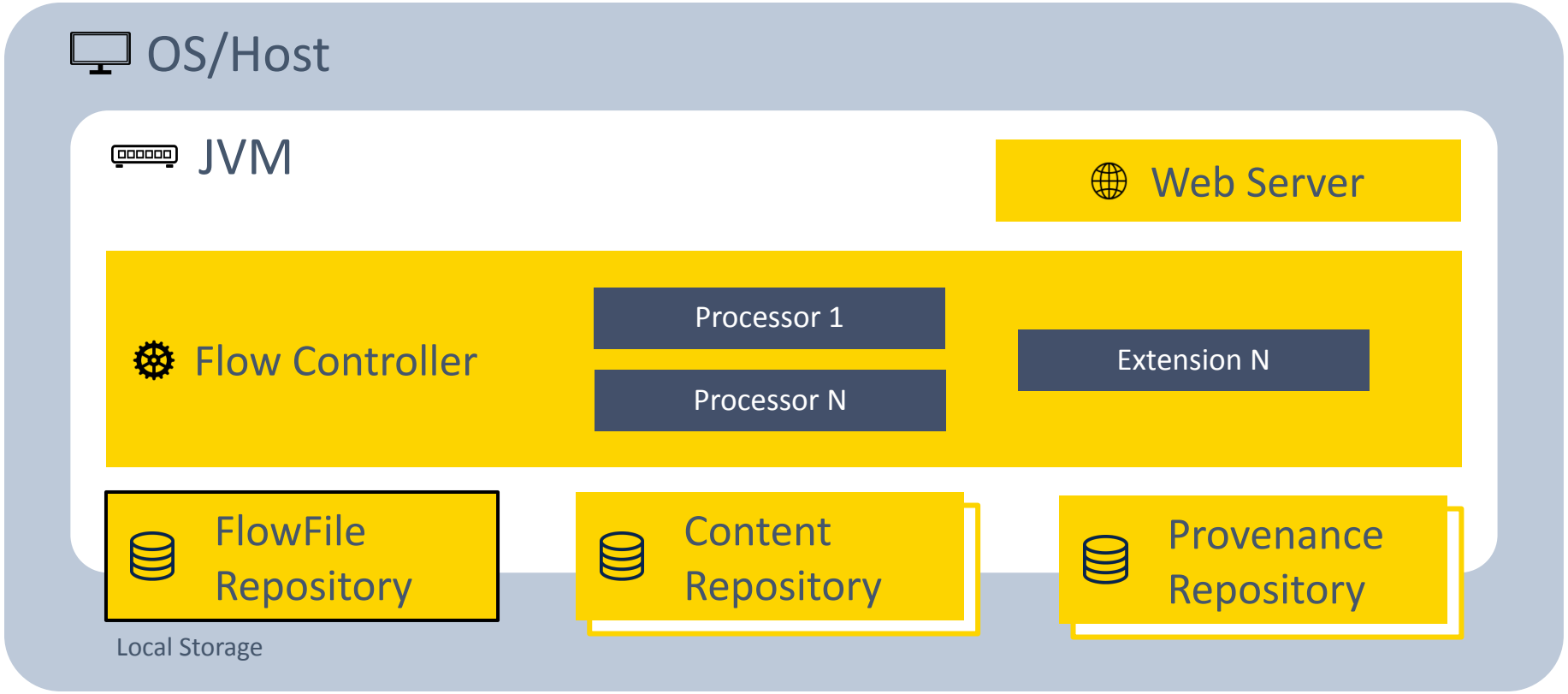
[Документация по установке NiFi]

3

Архитектура и компоненты Apache NiFi

Архитектура Apache NiFi

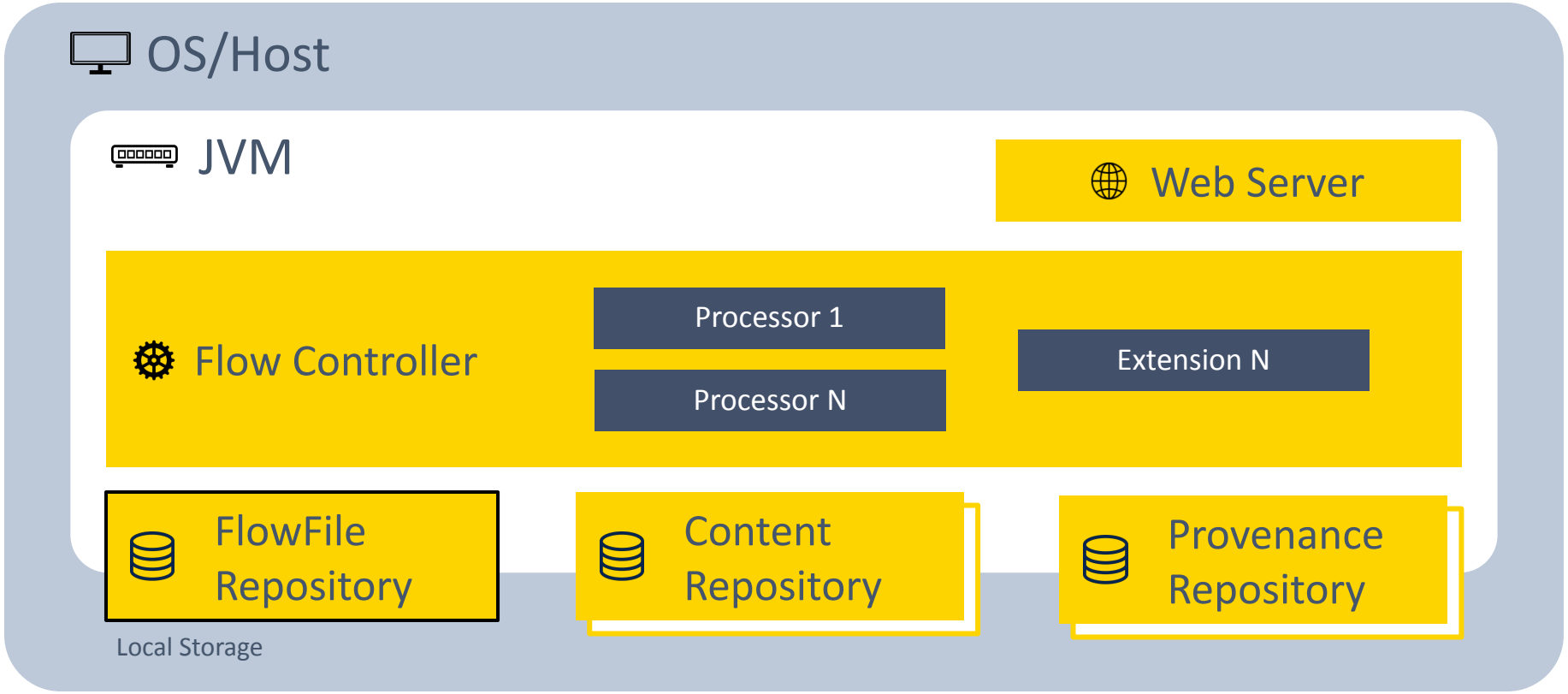
Архитектура состоит из **веб-сервера**, **контроллера потока** и **трёх репозиториев**



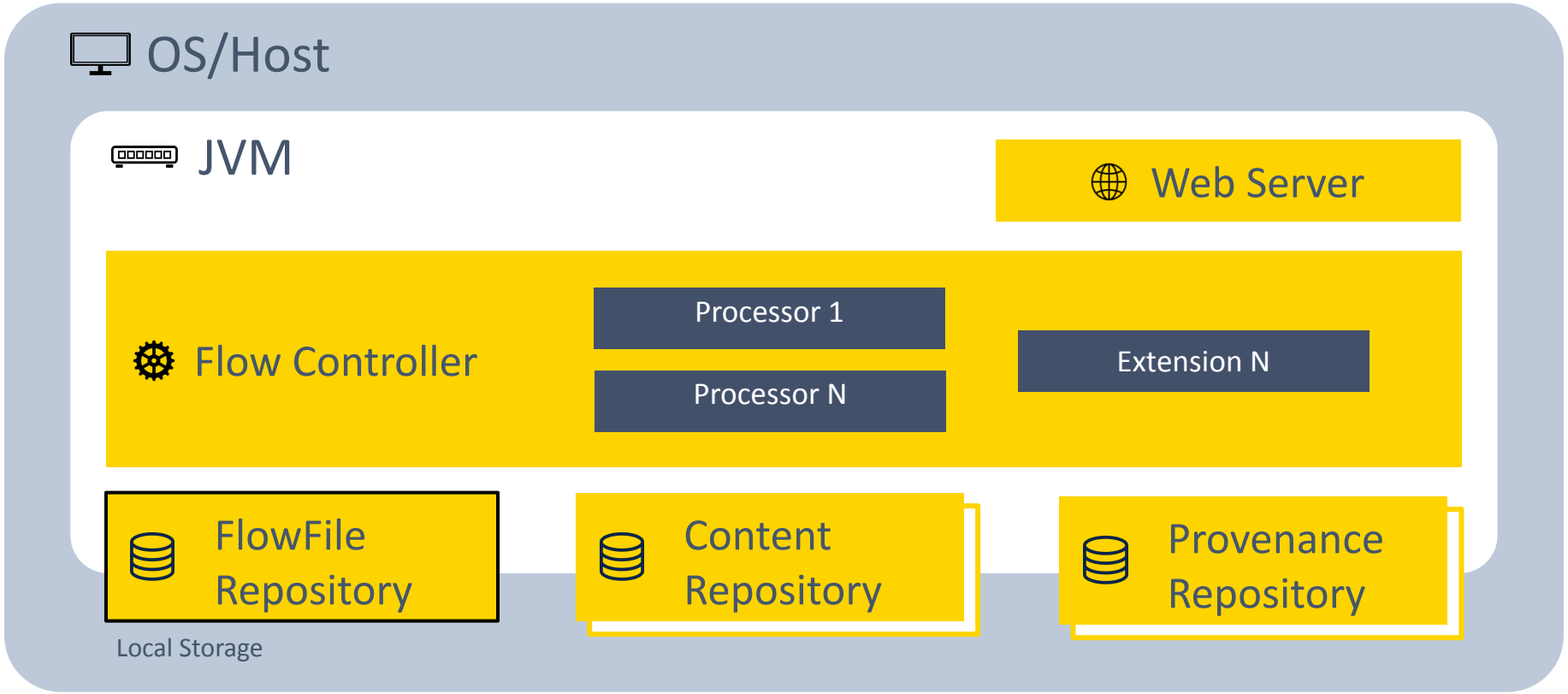
Архитектура Apache NiFi

Web Server — предоставляет веб-интерфейс и REST API

Flow Controller — Контроллер потока сохраняет информацию о том, как процессоры соединяются и управляет потоками, которые используют процессоры



Архитектура Apache NiFi



FlowFile Repository — хранилище, в котором NiFi содержит известную ему информацию **о каждом** существующем в **данный момент** FlowFile в системе

Content Repository, в нём находится содержимое всех FlowFile, то есть сами **передаваемые данные**

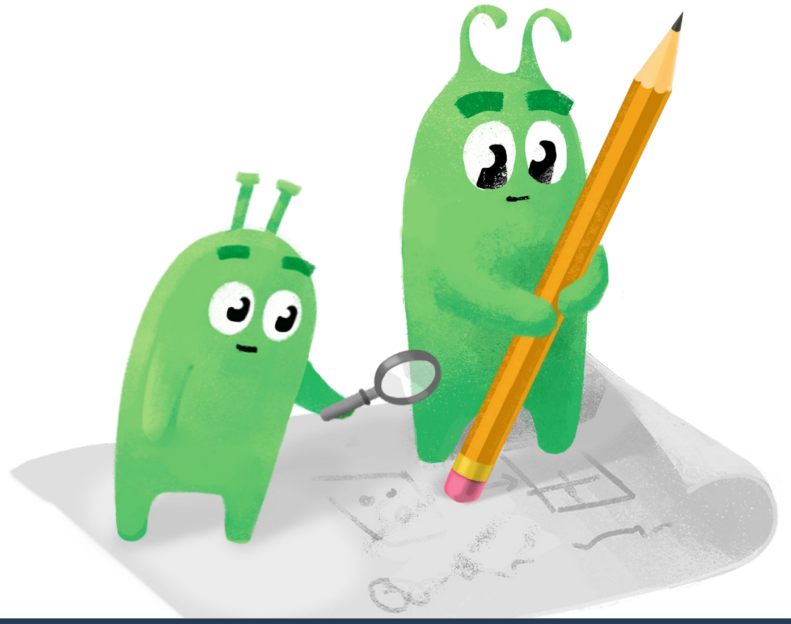
Provenance Repository содержит **историю** о каждом FlowFile

Важно

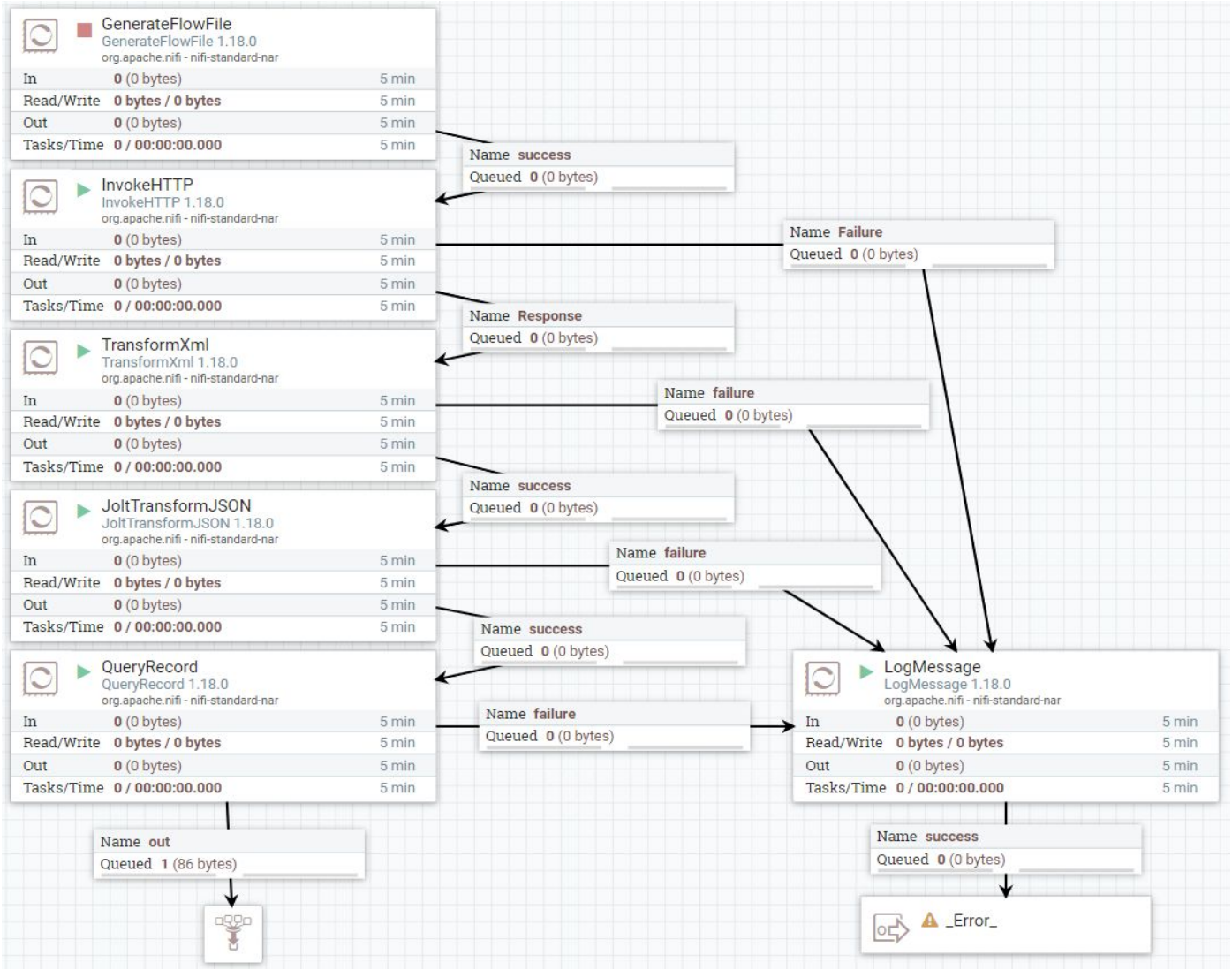
- Выделять больше памяти внутри ВМ
- FlowFile — особенный, следите за тем, чтобы распределять нагрузку и не использовать большую нагрузку одномоментно
- Масштабирование

4

Понятия и компоненты



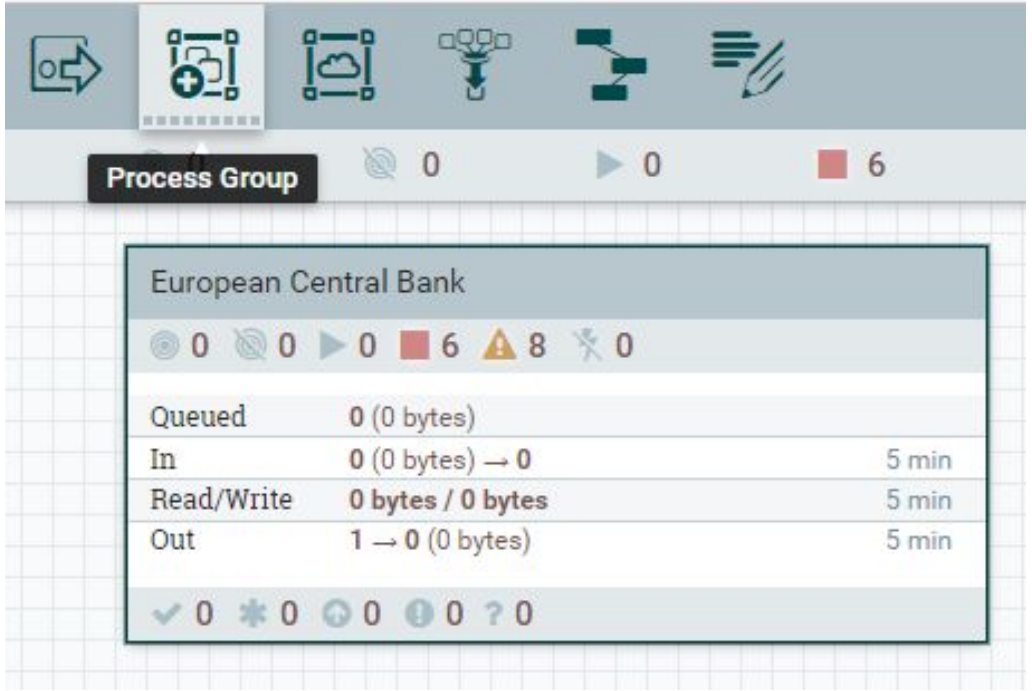
Как выглядит



Группы

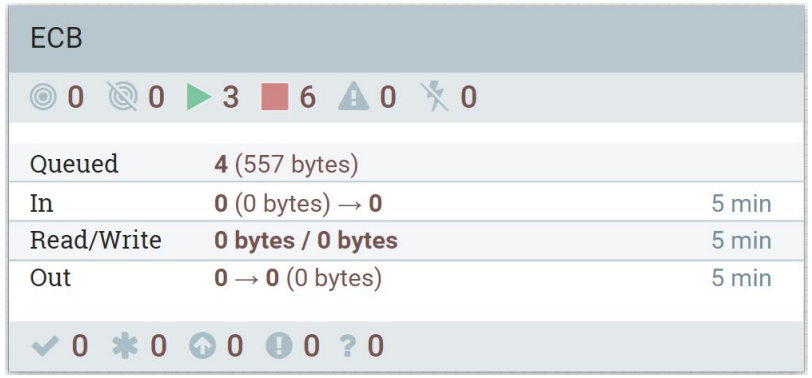
В Apache NiFi можно разместить **потоки данных в разных группах процессов**.

Они могут содержать разные проекты, задачи или подгруппы внутри группы.

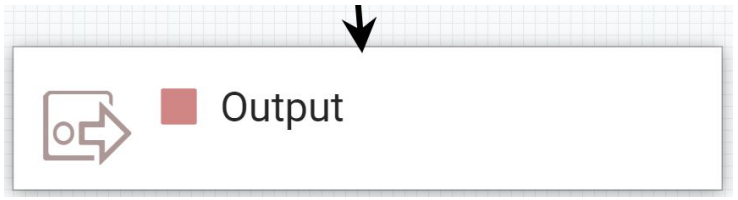


Понятия и компоненты

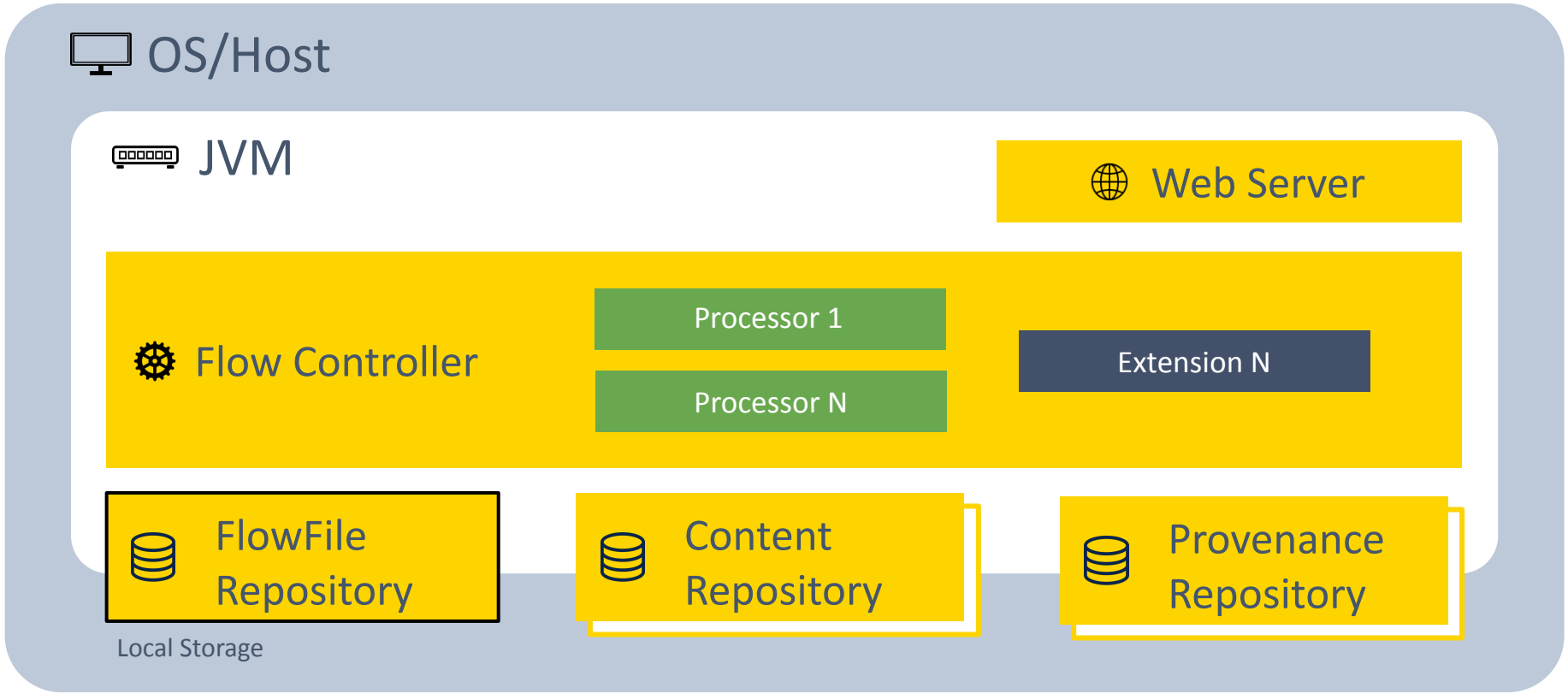
Process Group — набор процессоров, их подключений и прочих элементов DataFlow



Для получения и отправки данных из Process Groups используются **Input/Output Ports**



FlowFile Processor — алгоритм (black box), который выполняет необходимую работу в NiFi



Процессор

Процессор — может выполнять следующие задачи:

- Создание данных
- Извлечение данных
- Преобразование данных
- Сохранение данных в целевой системе
- Добавление/удаление/изменение атрибутов
- Маршрутизация
- Разделение
- Слияние

Потоковый файл

FlowFile (Потоковый файл) — это данные в системе NiFi

- Файл состоит **из контента и метаданных**
- Файл может создаваться **по расписанию или событию**
- Файл может быть **бинарным или текстовым**

Метаданные

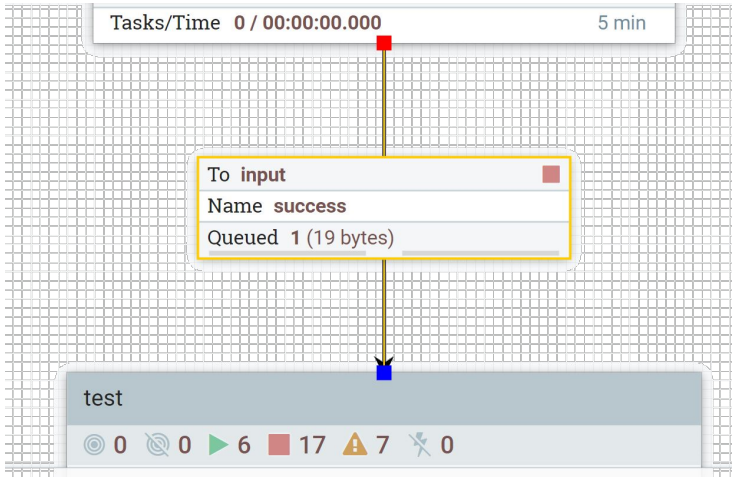
Метаданные — это атрибуты

- Например, атрибутом является имя файла, id файла, имя схемы и т.д.
- Система позволяет добавлять собственные атрибуты и присваивать им любые значения
- Значения атрибутов используются в работе процессоров

Соединение

Соединение (Connection) — обеспечивает подключение и передачу FlowFile между различными процессорами и некоторыми другими сущностями NiFi

Connection помещает FlowFile **в очередь**, после чего **передает** его далее **по цепочке**



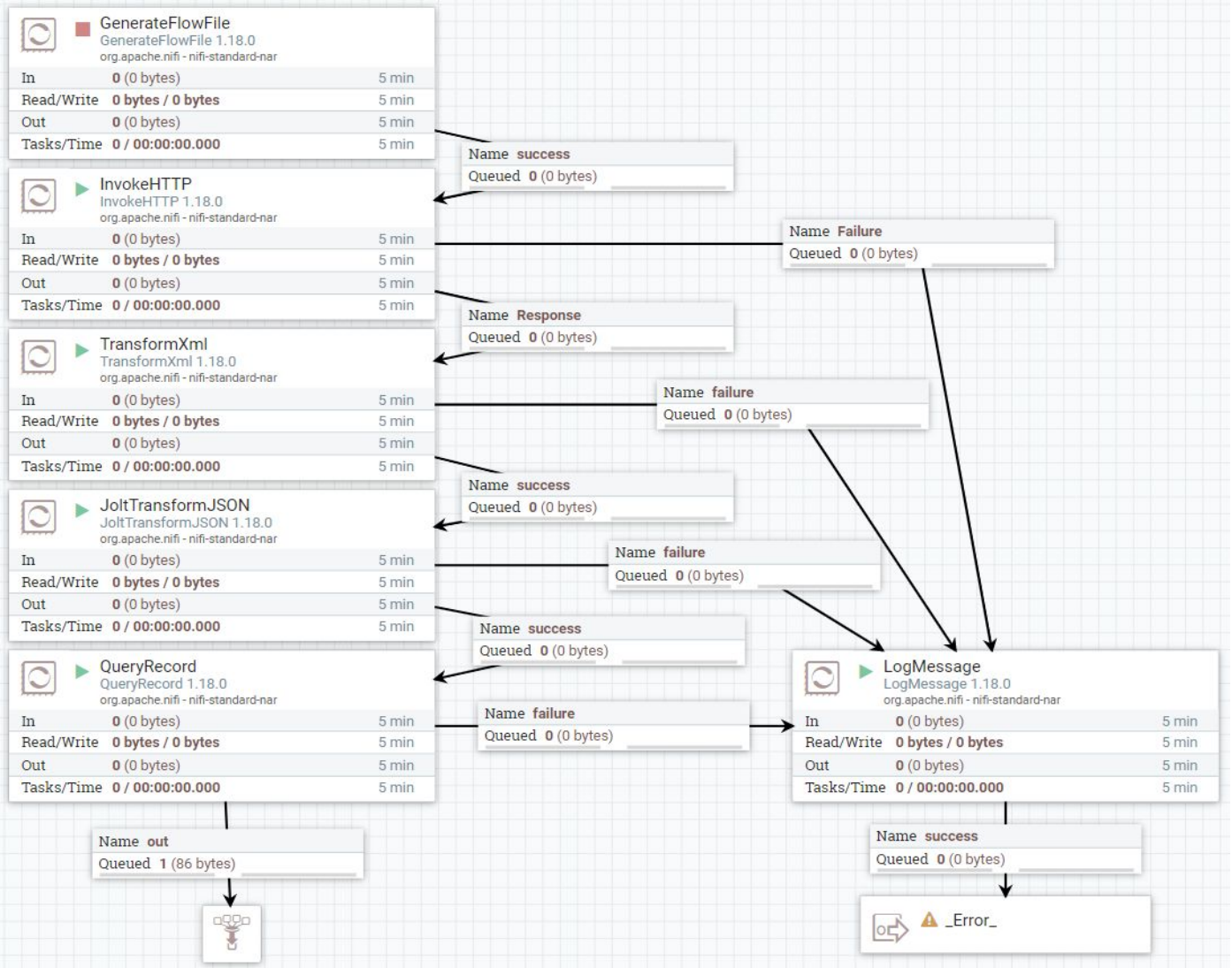
Контроллеры

Это объекты, которые знают, как сделать какую-то работу

- **Reader –** контроллер, который знает, как **читать** данные
- **RecordSetWriter –** контроллер, который **форматирует** данные на выходе процессора

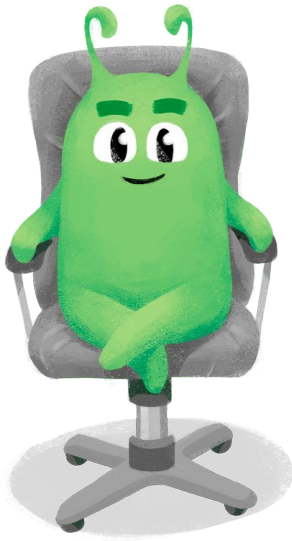
	Name	Type	Bundle	State	Scope	
	CSVReader	CSVReader 1.18.0	org.apache.nifi - nifi-record-ser...	Enabled	test	
	CSVRecordSetWriter	CSVRecordSetWriter 1.18.0	org.apache.nifi - nifi-record-ser...	Enabled	test	
	JsonTreeReader	JsonTreeReader 1.18.0	org.apache.nifi - nifi-record-ser...	Enabled	test	
	ParquetRecordSetWriter	ParquetRecordSetWriter 1.18.0	org.apache.nifi - nifi-parquet-nar	Enabled	test	

Пример потока данных

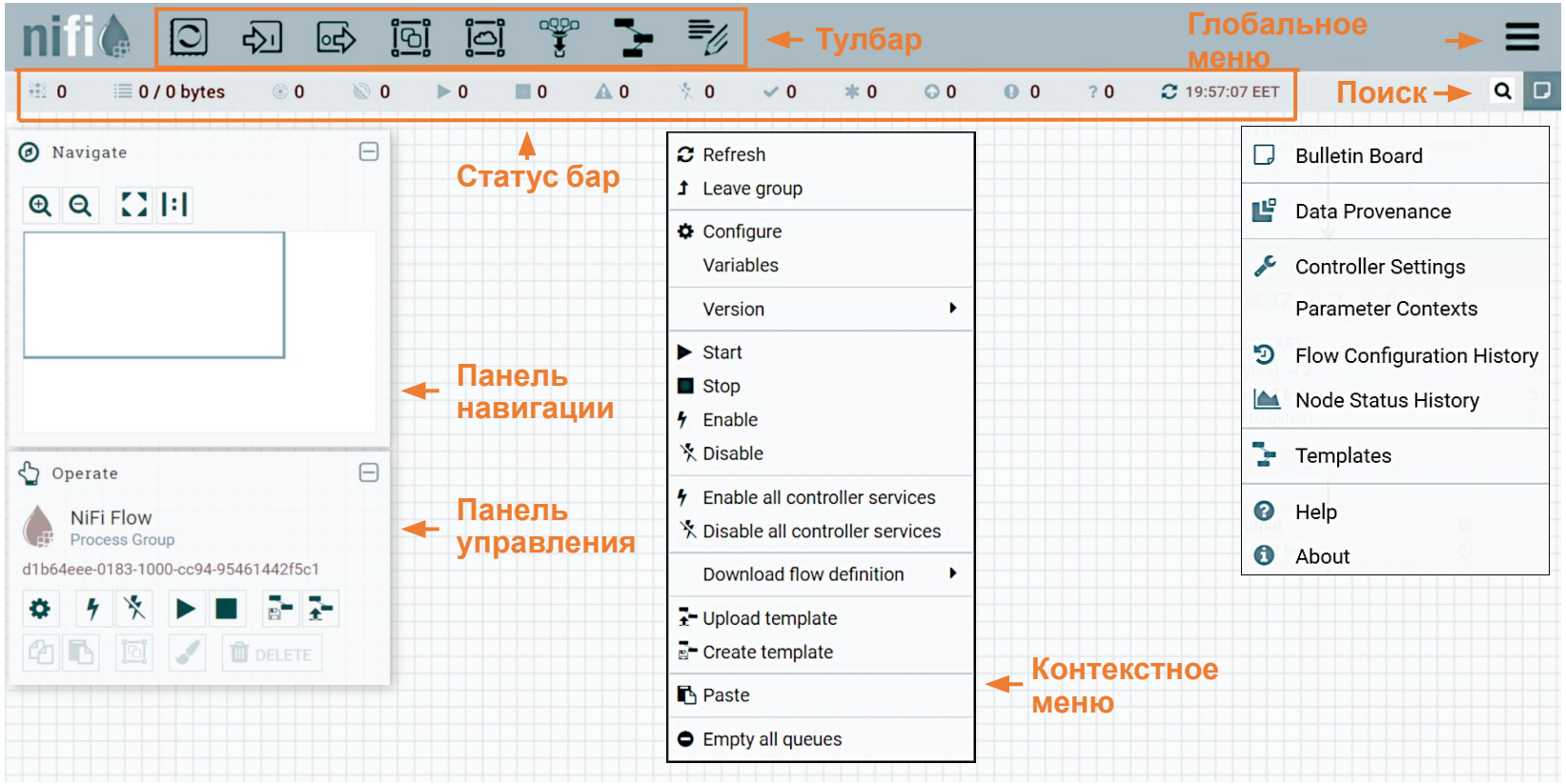


5

Интерфейс Apache NiFi



Веб-интерфейс NiFi



Components Toolbar

Processor

Input Port

Output Port

Process Group

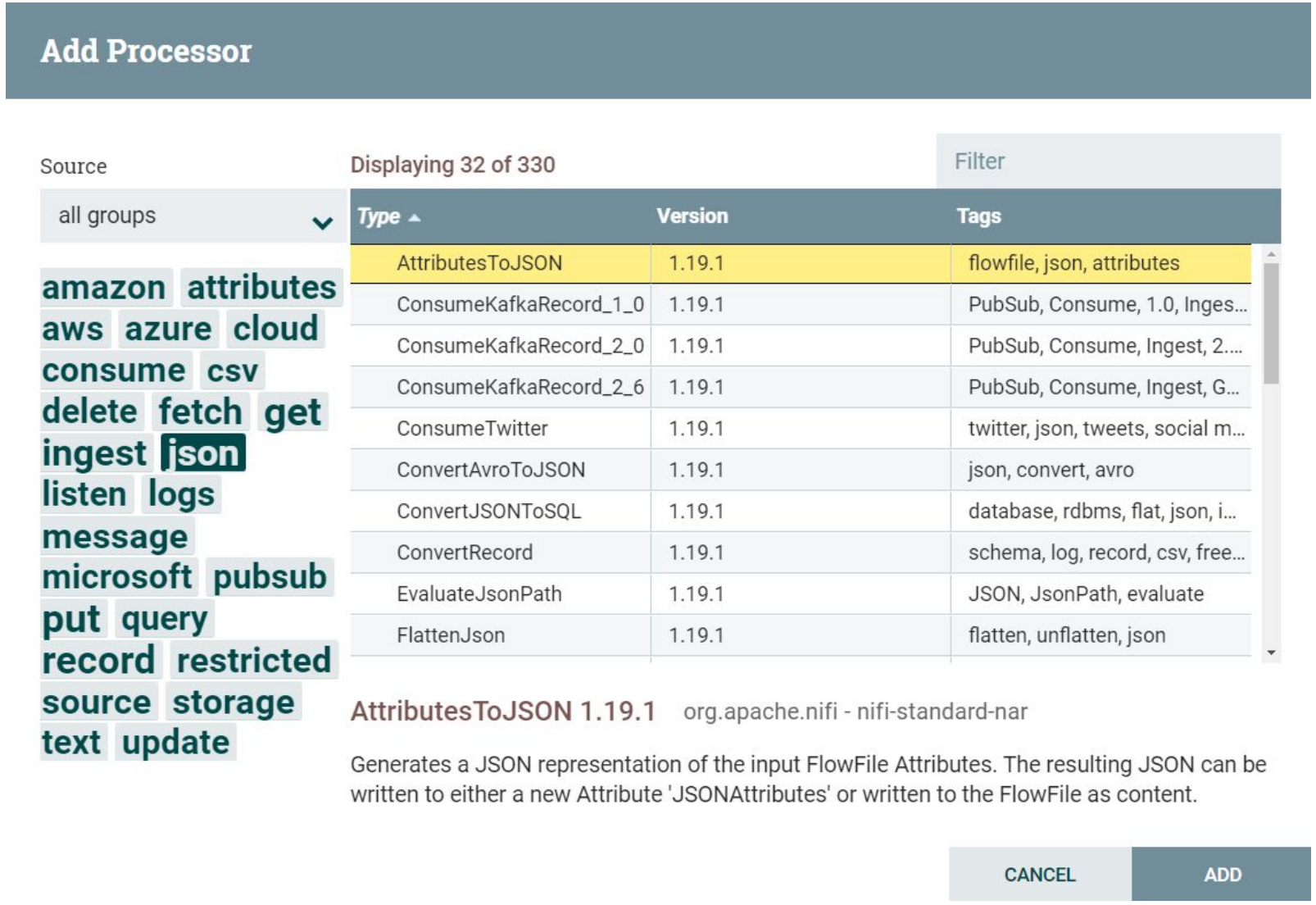
Remote Process Group

Funnel

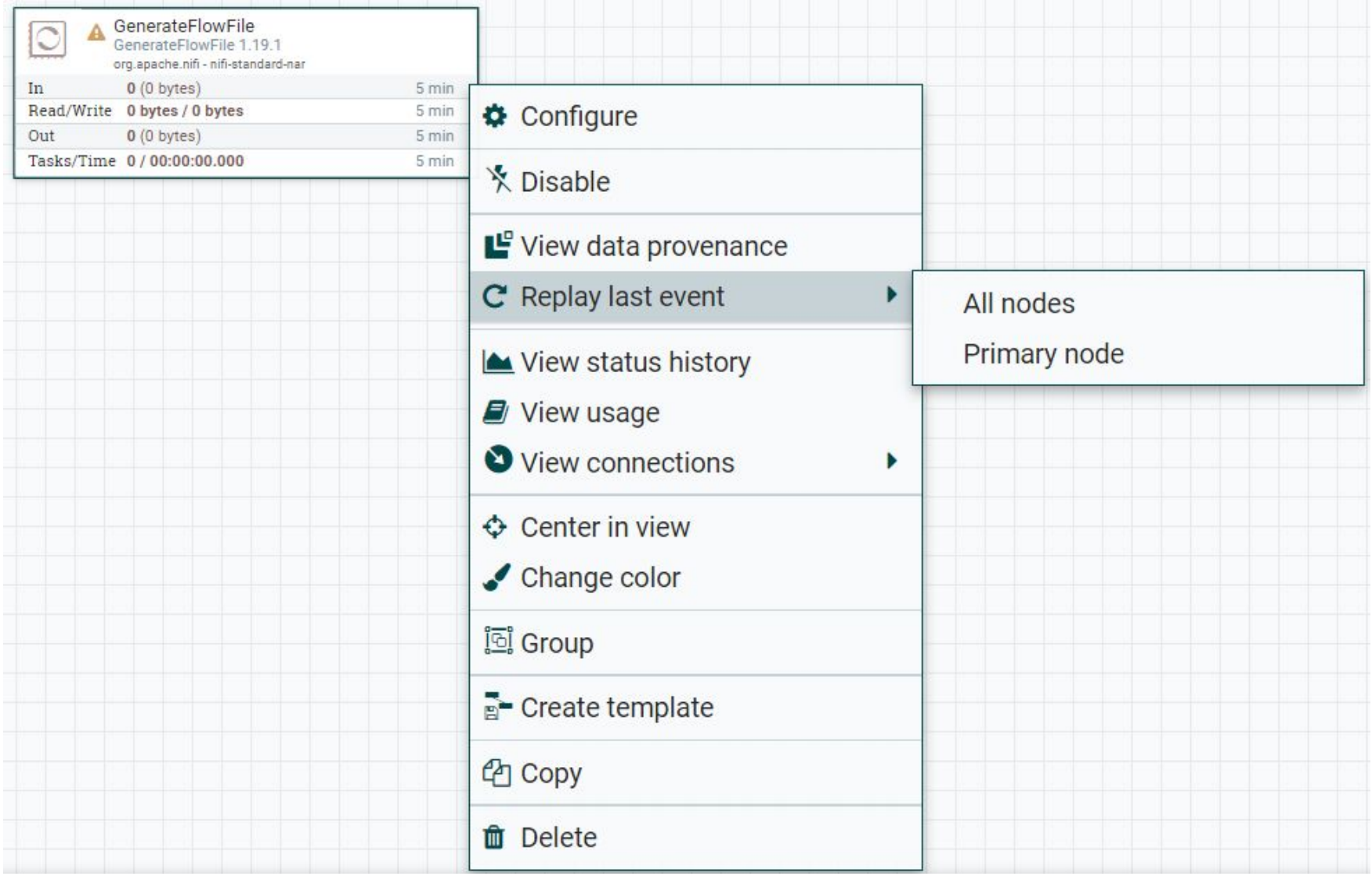
Template

Label

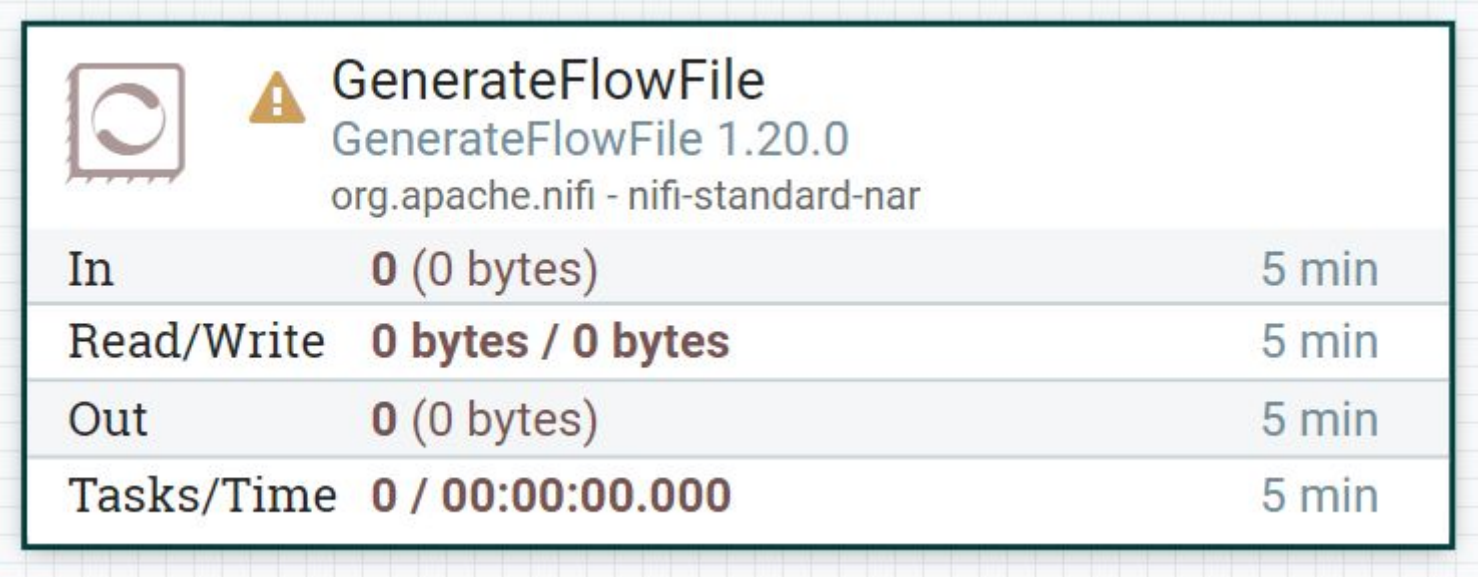
Add Processor



Контекстное меню процессора



Индикатор состояния процессора



Индикатор состояния: показывает текущее состояние процессора.

- **Running (работает)**: процессор в данный момент работает;
- **Stopped (остановлен)**: процессор остановлен;
- **Invalid (невалидный)**: процессор не сконфигурирован должным образом;
- **Disabled (отключен)**: процессор не работает и не может быть запущен, пока он не будет включен. Этот статус не указывает, валидный ли процессор.

6

Классификация процессоров

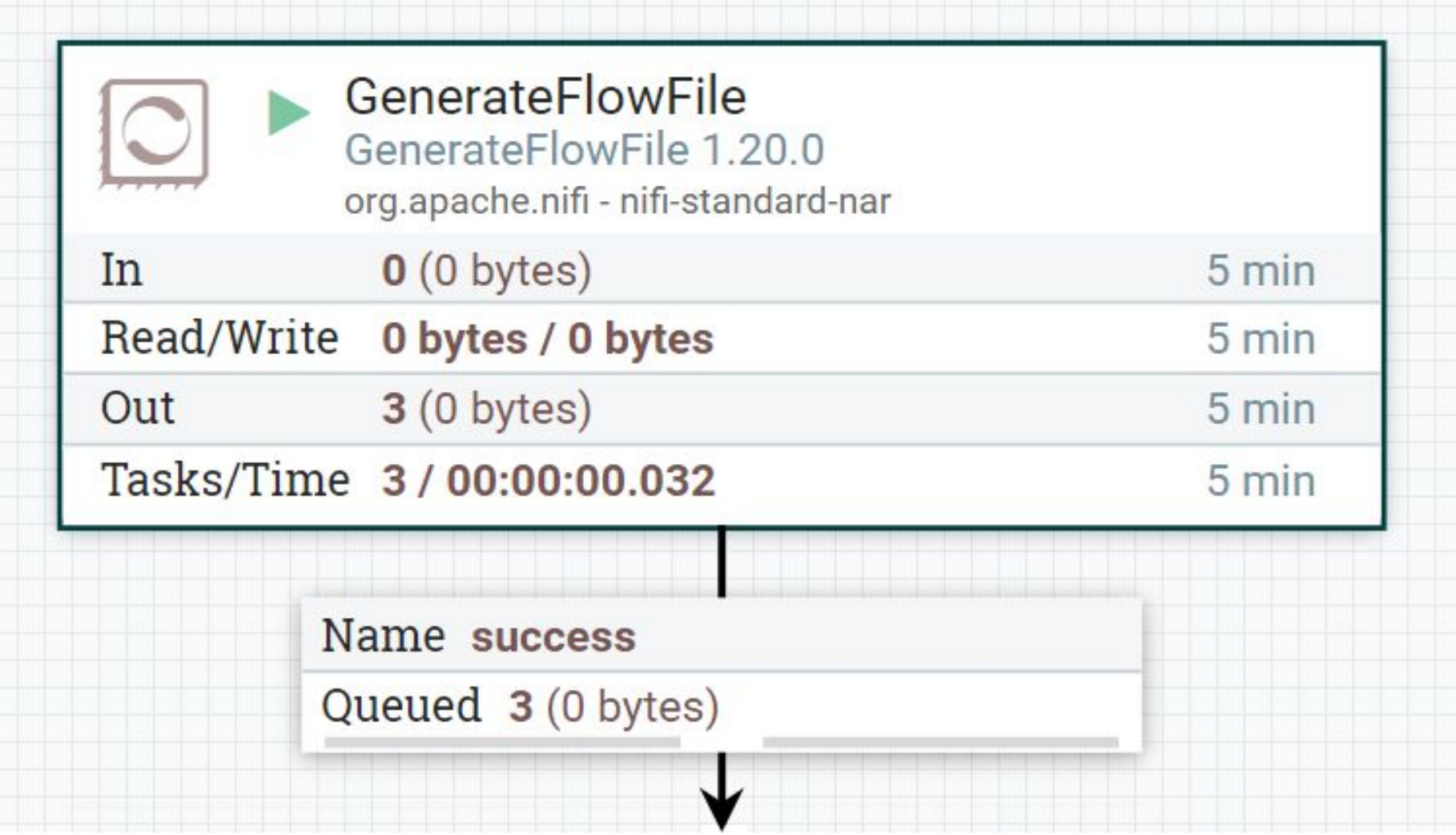
Процессоры, работающие без (upstream)

Условно — генераторы данных

Могут не только создавать контент, но и получать его извне

Для них критично определение периода запуска:
если поставить значение **по умолчанию** 0 секунд,
они будут выполнять постоянные опросы источника.

GenerateFlowFile, GetFile, GetHTTP, GetFTP, GetKAFKA и другие



Процессоры, работающие с upstream

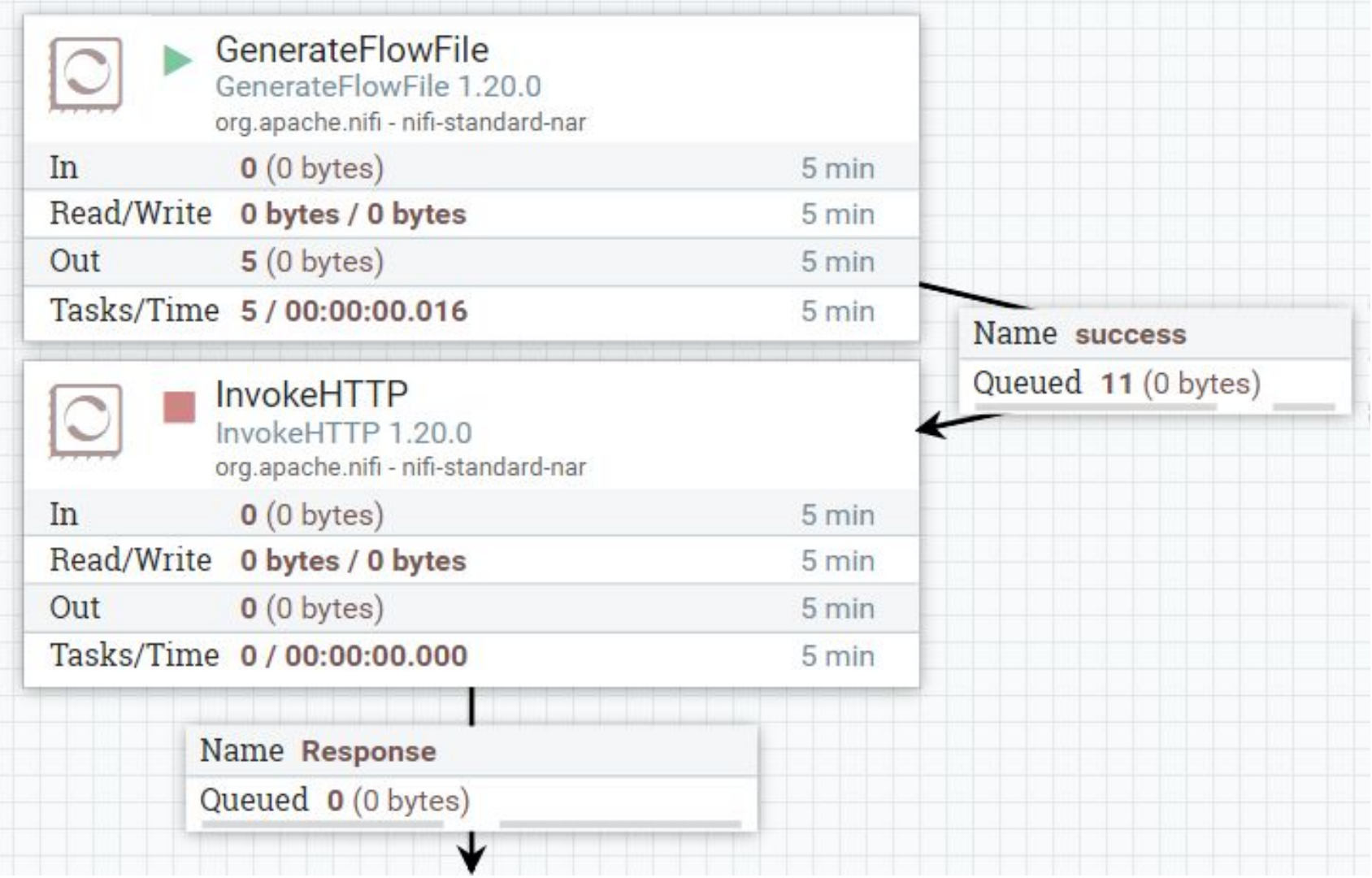
Получают данные из поступающего на **вход** файла,
модифицируют **контент** или **атрибуты**

Они выполняют работу только при наличии
файлов во входящей очереди

Расписание регулирует максимальную частоту запусков
при наличии файлов,для них нормально выставлять 0 секунд

Для более тонкой настройки потока можно выравнивать движение
по потоку за счет периода запуска

Процессоры работающие с upstream



Дополнительные материалы

Перечень книг, полезных в изучении NiFi

Итоги. О чём поговорили:

1. Знакомство с Apache NiFi
2. Установка Apache NiFi
3. Архитектура Apache NiFi
4. Понятия и компоненты Apache NiFi
5. Интерфейс Apache NiFi
6. Классификация процессоров

Спасибо
за внимание!