

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

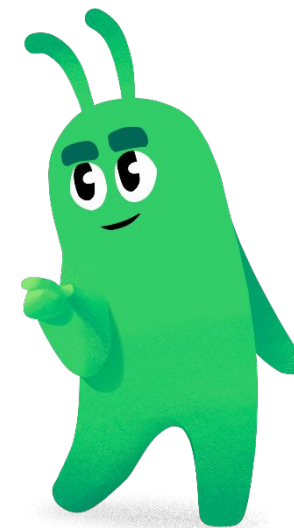


О недостатках

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

1

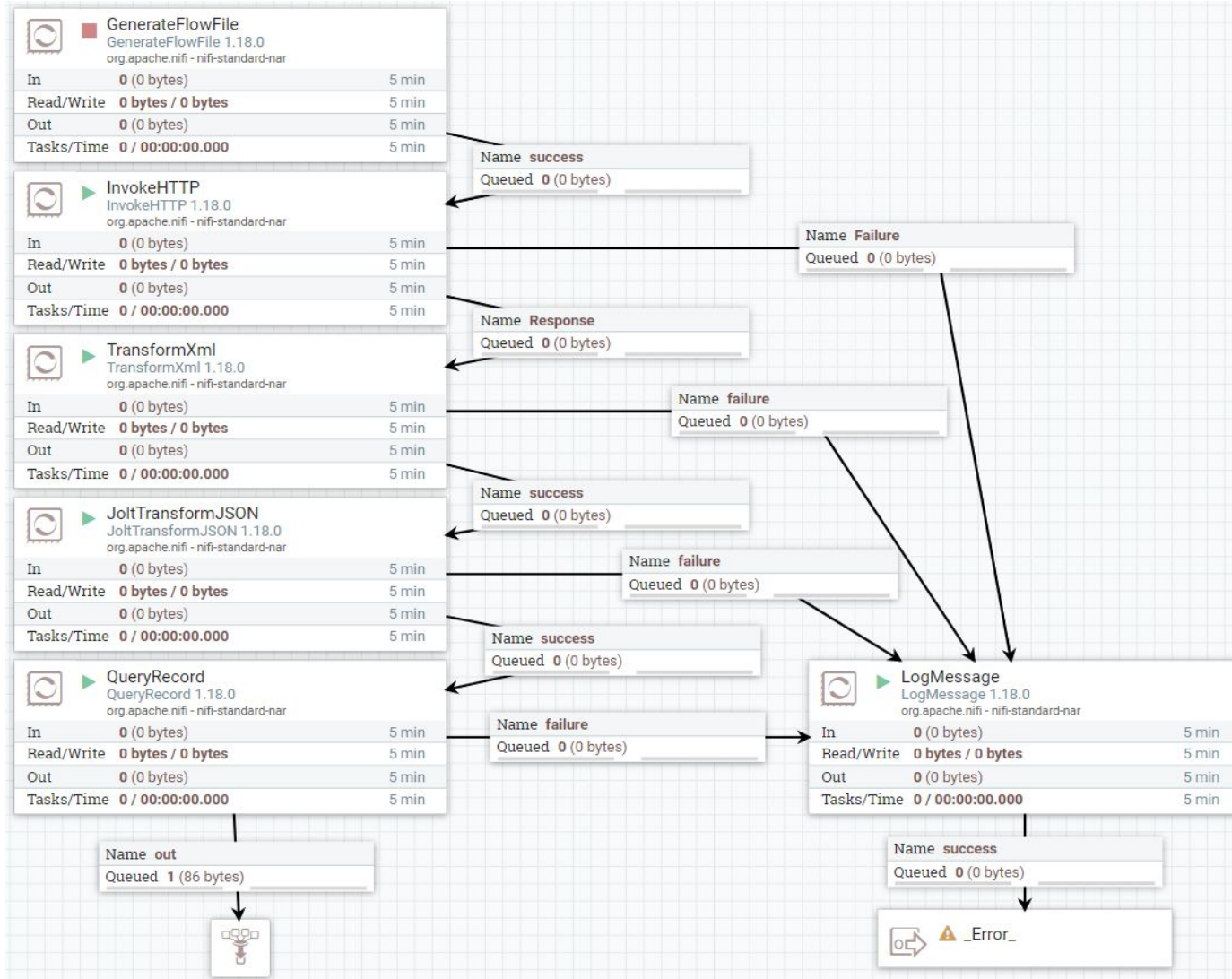
Знакомство с Apache NiFi



Apache NiFi

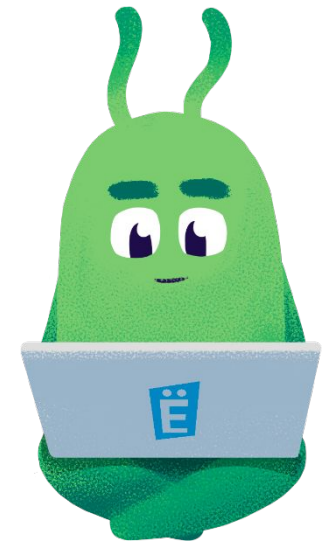


Как выглядит



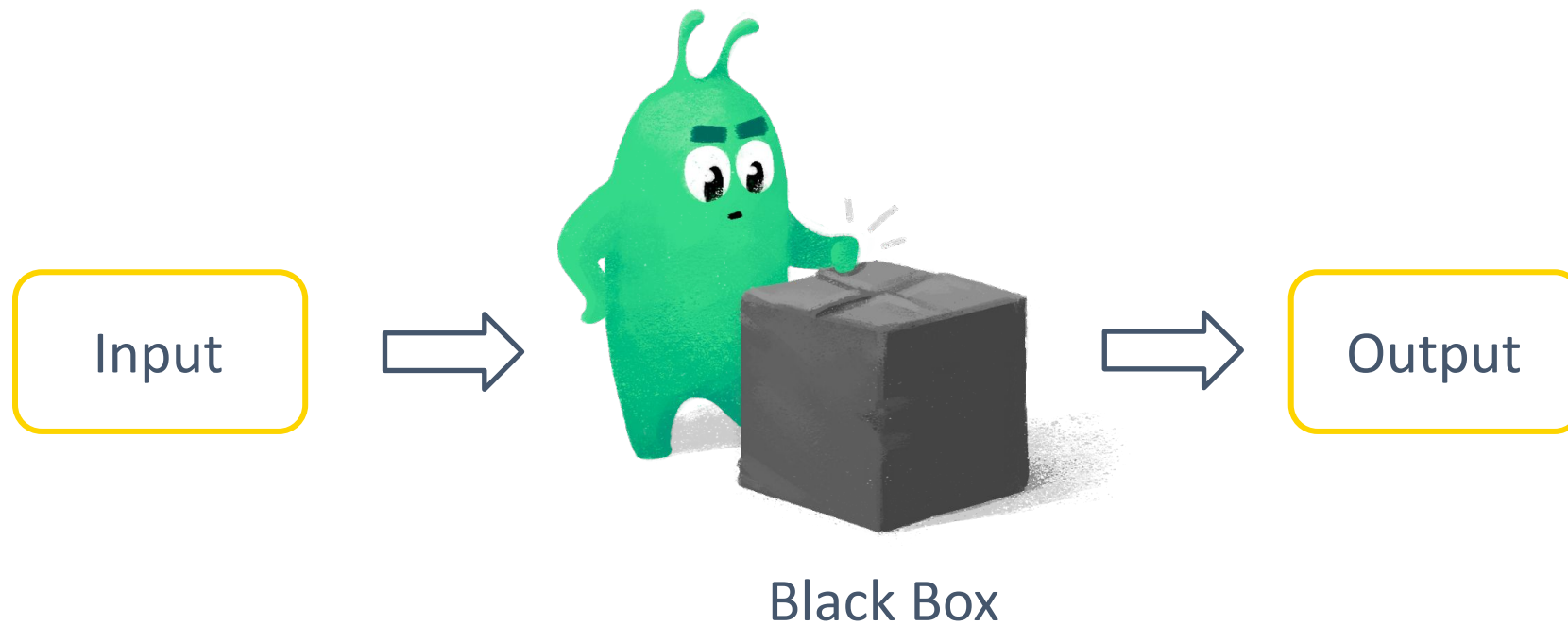
Что такое Apache NiFi

- 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



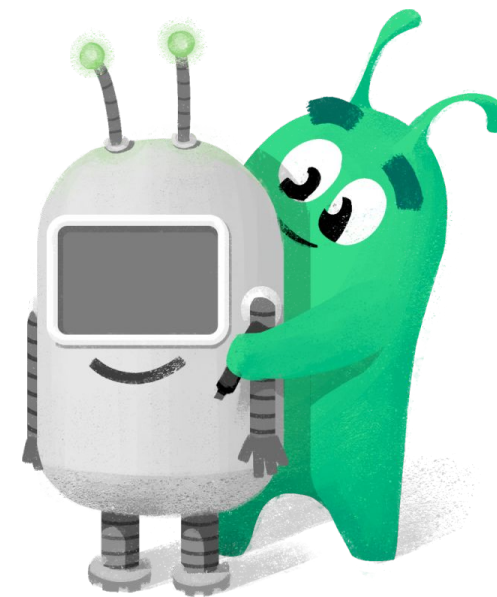
Что такое Apache NiFi

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



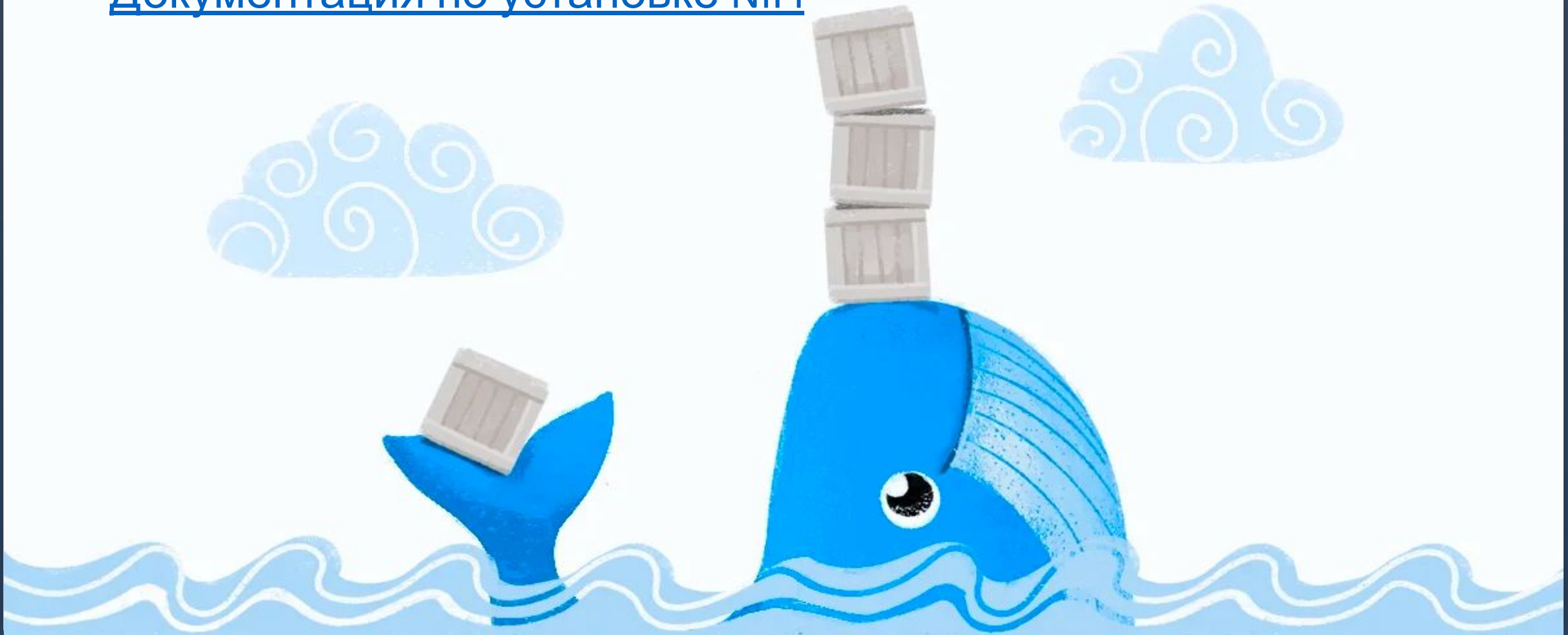
2

Установка



Установка Airflow

[Документация по установке NiFi](#)

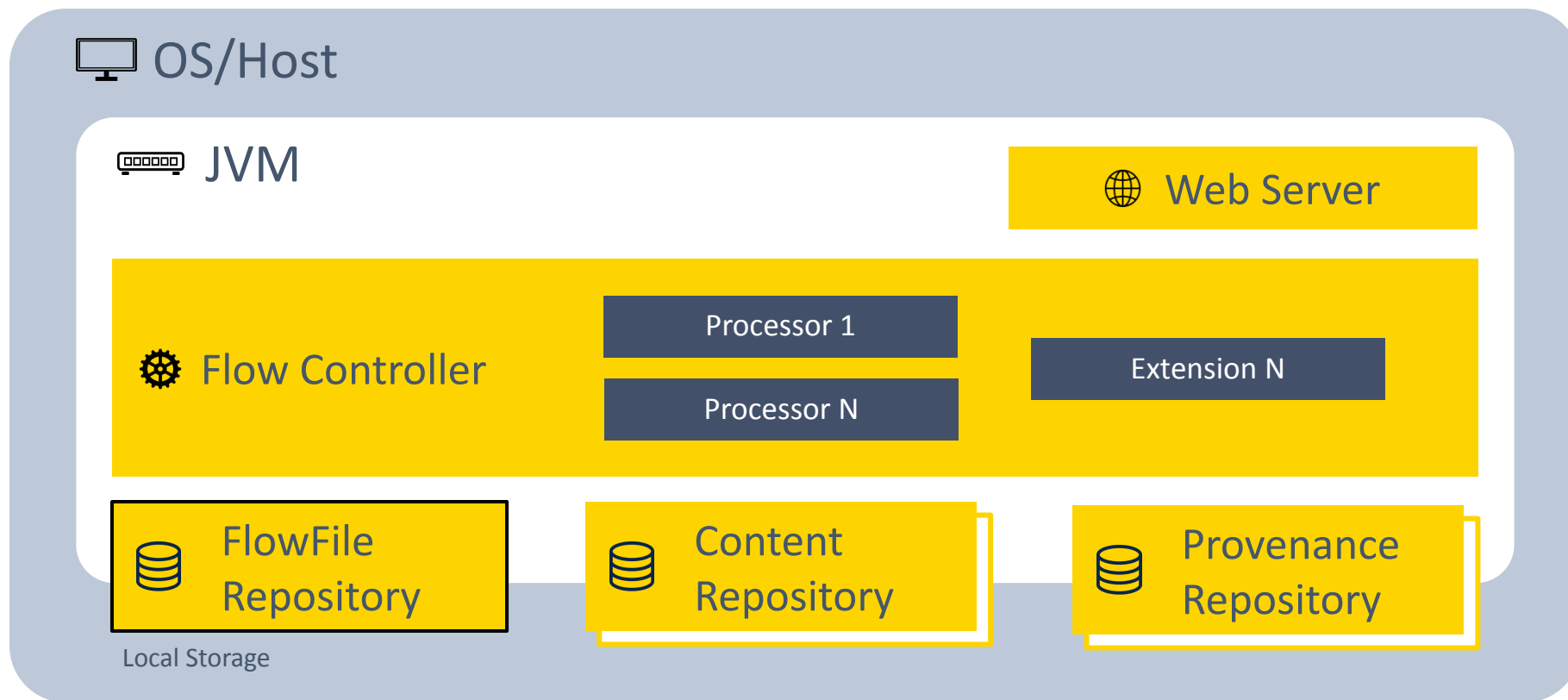


Архитектура и компоненты Apache NiFi



Архитектура Apache NiFi

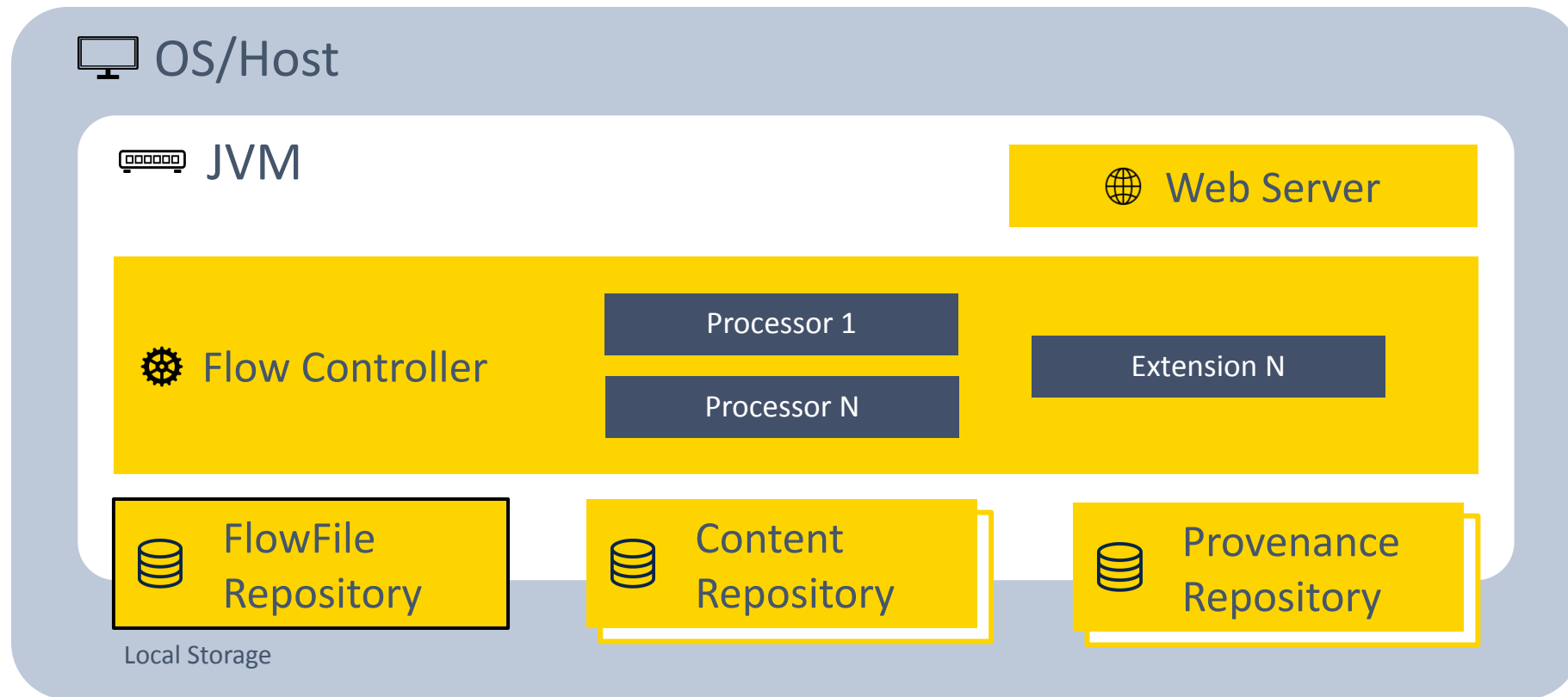
Архитектура состоит из **веб-сервера, контроллера потока и трёх репозиториев**



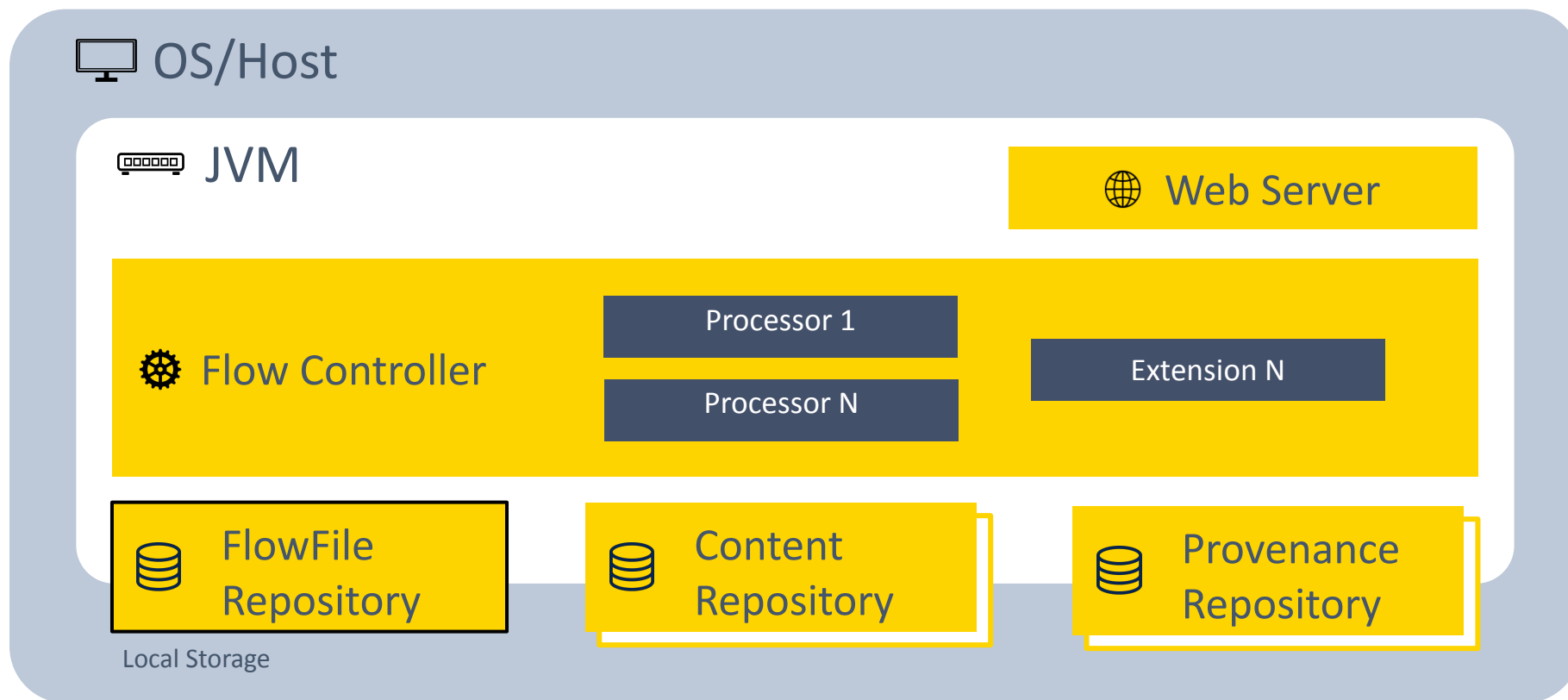
Архитектура Apache NiFi

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**

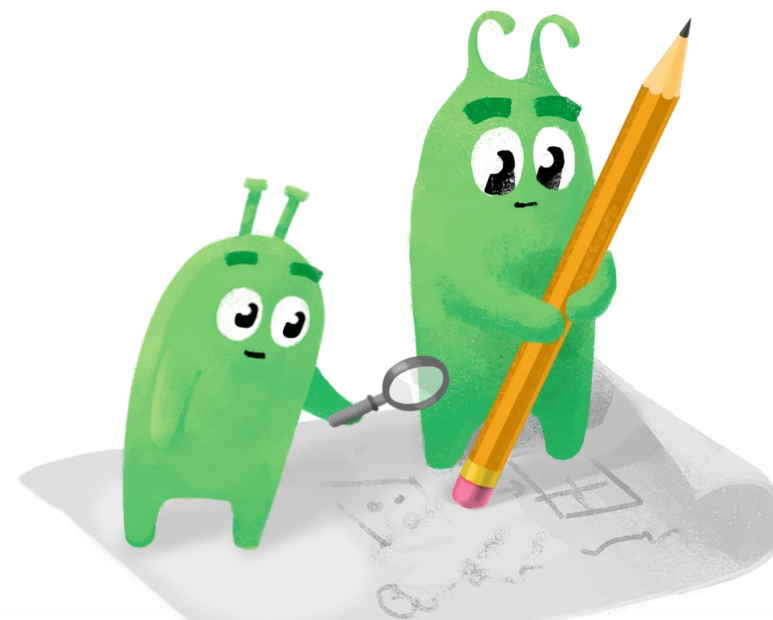
Provenance Repository содержит **историю** о каждом FlowFile

Важно

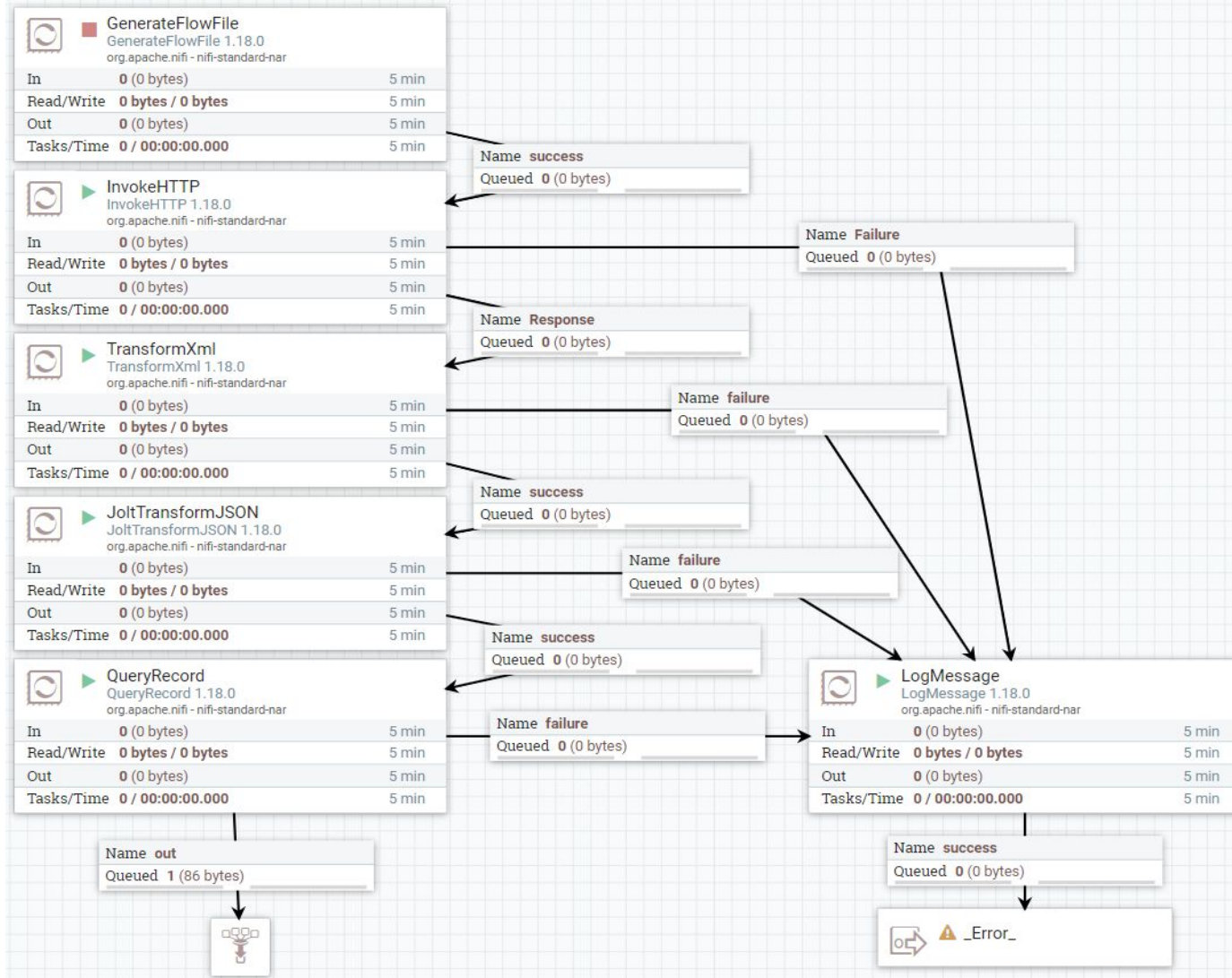
- Выделять больше памяти внутри VM
- FlowFile — особенный, следите за тем, чтобы распределять нагрузку и не использовать большую нагрузку одновременно
- Масштабирование

4

Понятия и компоненты



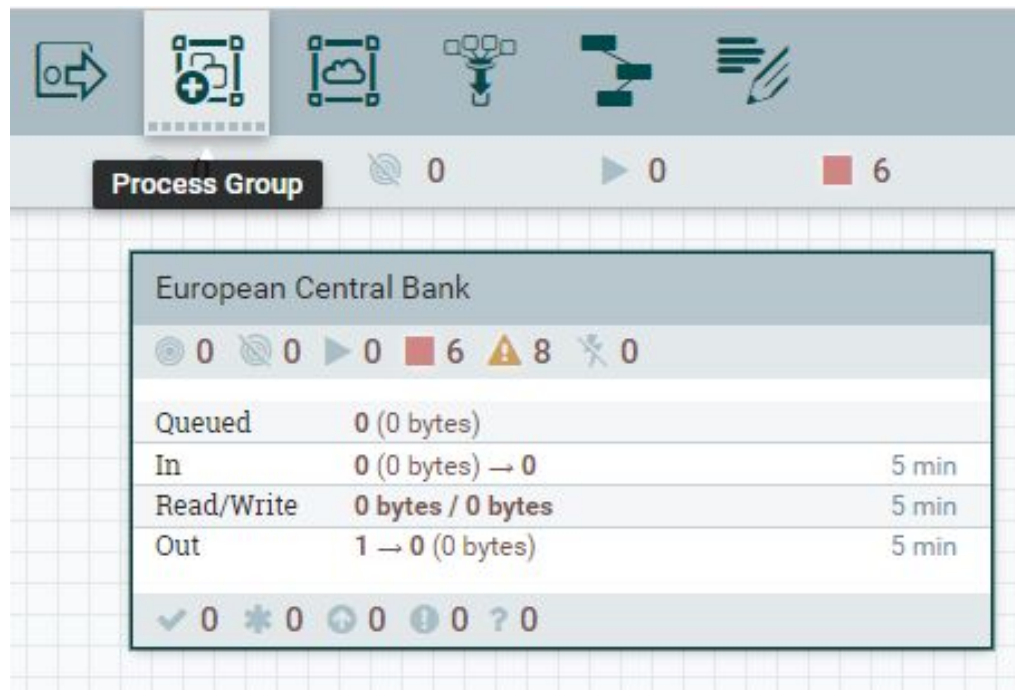
Как выглядит



Группы

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

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



The screenshot displays the Apache NiFi web console interface. At the top, there is a toolbar with icons for navigation and actions. Below the toolbar, a 'Process Group' header shows status indicators: a play button with '0', a red square with '6', and a yellow triangle with '8'. The main content area shows a table for the 'European Central Bank' process group.

European Central Bank		
Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	1 → 0 (0 bytes)	5 min

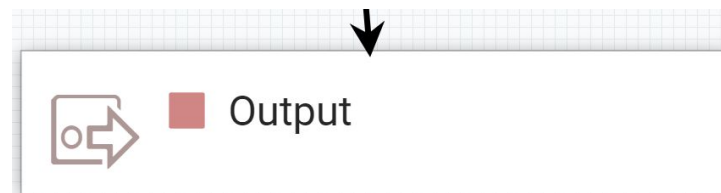
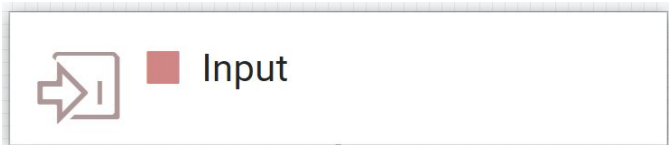
Below the table, there are additional status indicators: a checkmark with '0', a star with '0', a play button with '0', a red square with '0', and a question mark with '0'.

Понятия и компоненты

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

ECB		
🎯 0	🚫 0	▶ 3
■ 6	⚠ 0	✂ 0
Queued	4 (557 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	

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



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

OS/Host

JVM

Web Server

Flow Controller

Processor 1

Processor N

Extension N

FlowFile Repository

Local Storage

Content Repository

Provenance Repository

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

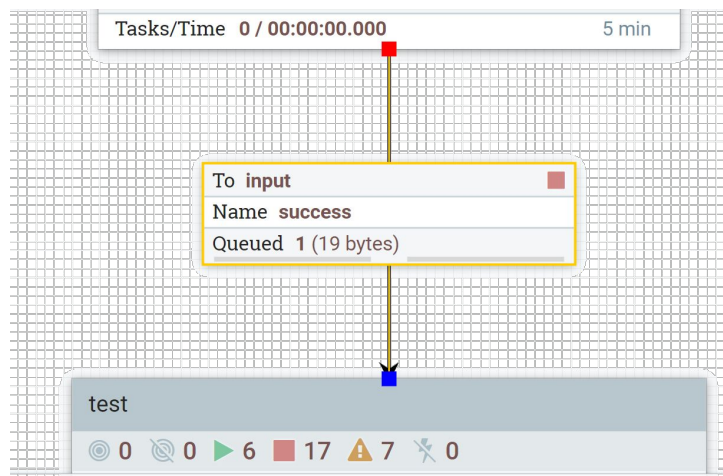
- Например, атрибутом является имя файла, id файла, имя схемы и т.д.
- Система позволяет добавлять собственные атрибуты и присваивать им любые значения
- Значения атрибутов используются в работе процессоров



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**



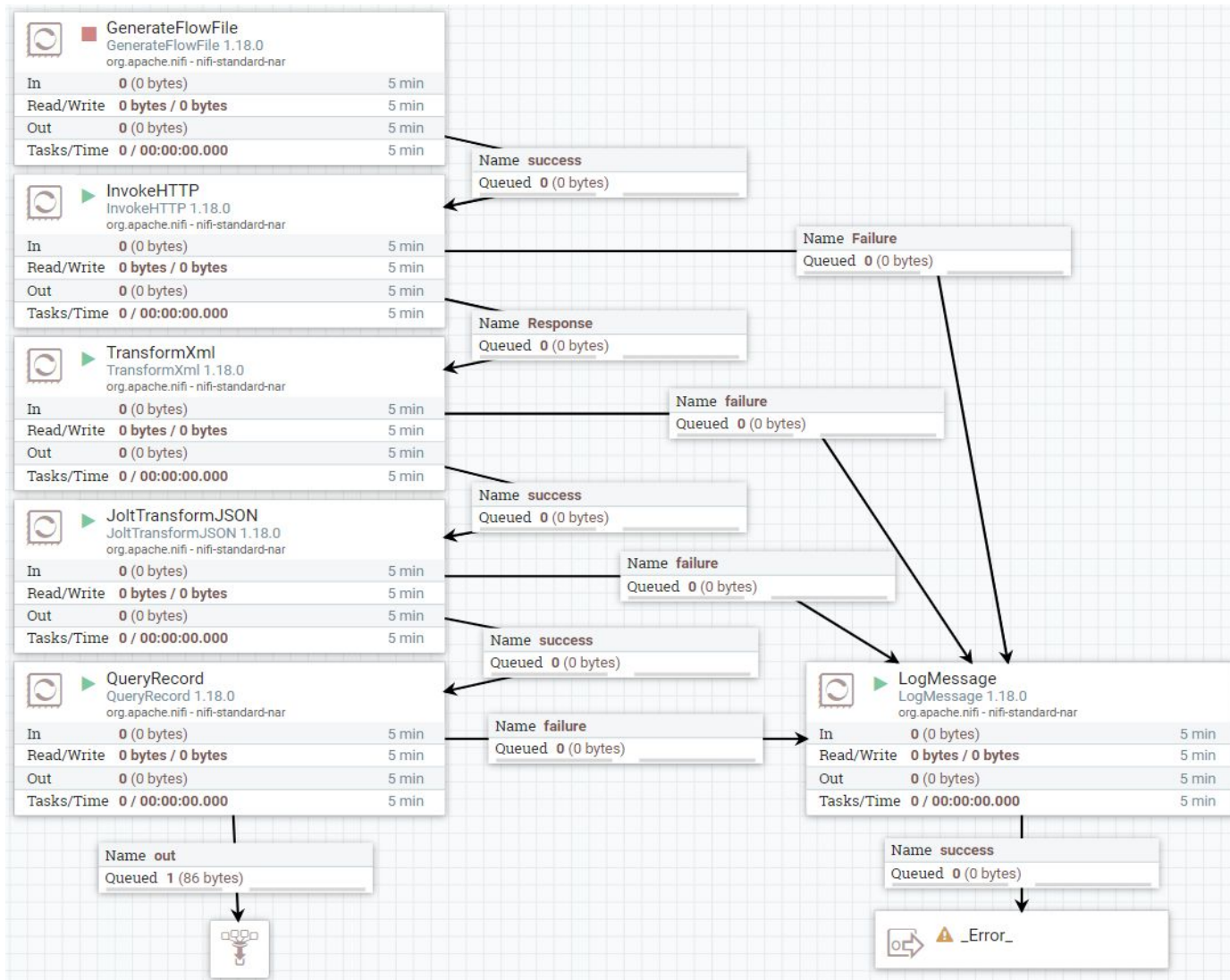
Контроллеры

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

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

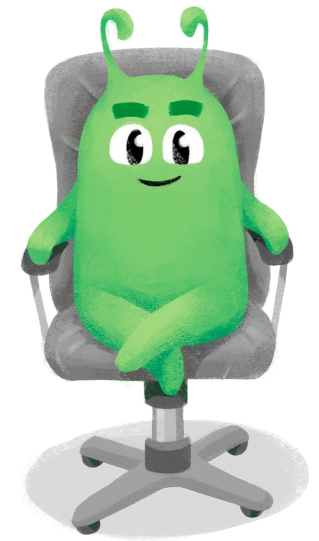
Name ▲	Type	Bundle	State	Scope	
 CSVReader	CSVReader 1.18.0	org.apache.nifi - nifi-record-ser...	 Enabled	test	 
 CSVRecordSetWriter	CSVRecordSetWriter 1.18.0	org.apache.nifi - nifi-record-ser...	 Enabled	test	 
 JsonTreeReader	JsonTreeReader 1.18.0	org.apache.nifi - nifi-record-ser...	 Enabled	test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

Пример потока данных



5

Интерфейс Apache NiFi



Веб-интерфейс NiFi

The screenshot shows the NiFi web interface with several key components highlighted by orange arrows and labels:

- Тулбар** (Toolbar): Located at the top left, containing icons for navigation and flow management.
- Глобальное меню** (Global menu): Located at the top right, containing a search bar and a hamburger menu icon.
- Статус бар** (Status bar): Located below the toolbar, displaying various status indicators and the current time (19:57:07 EET).
- Панель навигации** (Navigation panel): Located on the left side, containing search and navigation icons.
- Панель управления** (Control panel): Located below the navigation panel, displaying the current flow group name and various control icons.
- Контекстное меню** (Context menu): A floating menu in the center-right area, listing actions such as Refresh, Leave group, Configure, Start, Stop, Enable, Disable, and Download flow definition.
- Глобальное меню** (Global menu): A floating menu on the right side, listing actions such as Bulletin Board, Data Provenance, Controller Settings, Flow Configuration History, Node Status History, Templates, Help, and About.

Components Toolbar



Processor



Input Port



Output Port



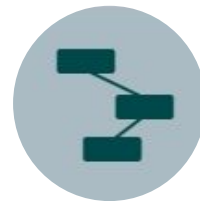
Process Group



Remote Process
Group



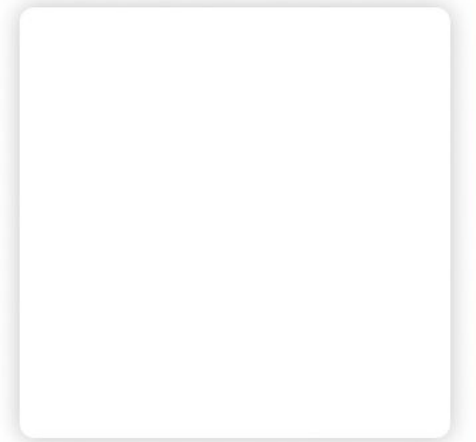
Funnel



Template



Label



Add Processor

Add Processor

Source: all groups ▼ Displaying 32 of 330 Filter

Type ▲	Version	Tags
AttributesToJSON	1.19.1	flowfile, json, attributes
ConsumeKafkaRecord_1_0	1.19.1	PubSub, Consume, 1.0, Inges...
ConsumeKafkaRecord_2_0	1.19.1	PubSub, Consume, Ingest, 2....
ConsumeKafkaRecord_2_6	1.19.1	PubSub, Consume, Ingest, G...
ConsumeTwitter	1.19.1	twitter, json, tweets, social m...
ConvertAvroToJSON	1.19.1	json, convert, avro
ConvertJSONToSQL	1.19.1	database, rdbms, flat, json, i...
ConvertRecord	1.19.1	schema, log, record, csv, free...
EvaluateJsonPath	1.19.1	JSON, JsonPath, evaluate
FlattenJson	1.19.1	flatten, unflatten, json

amazon attributes
aws azure cloud
consume csv
delete fetch get
ingest json
listen logs
message
microsoft pubsub
put query
record restricted
source storage
text update

AttributesToJSON 1.19.1 org.apache.nifi - nifi-standard-nar

Generates a JSON representation of the input FlowFile Attributes. The resulting JSON can be written to either a new Attribute 'JSONAttributes' or written to the FlowFile as content.

CANCEL

ADD



Контекстное меню процессора

The screenshot shows the context menu for a processor named 'GenerateFlowFile' in Apache NiFi. The processor's status is shown in a table on the left, and the context menu is open on the right. The 'Replay last event' option is selected, and its sub-menu is visible.

GenerateFlowFile GenerateFlowFile 1.19.1 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

- Configure
- Disable
- View data provenance
- Replay last event
 - All nodes
 - Primary node
- View status history
- View usage
- View connections
- Center in view
- Change color
- Group
- Create template
- Copy
- Delete

Индикатор состояния процессора

	 GenerateFlowFile GenerateFlowFile 1.20.0 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

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.

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, GetКАFKA и другие



GenerateFlowFile
GenerateFlowFile 1.20.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	3 (0 bytes)	5 min
Tasks/Time	3 / 00:00:00.032	5 min

Name **success**

Queued 3 (0 bytes)

Процессоры, работающие с upstream

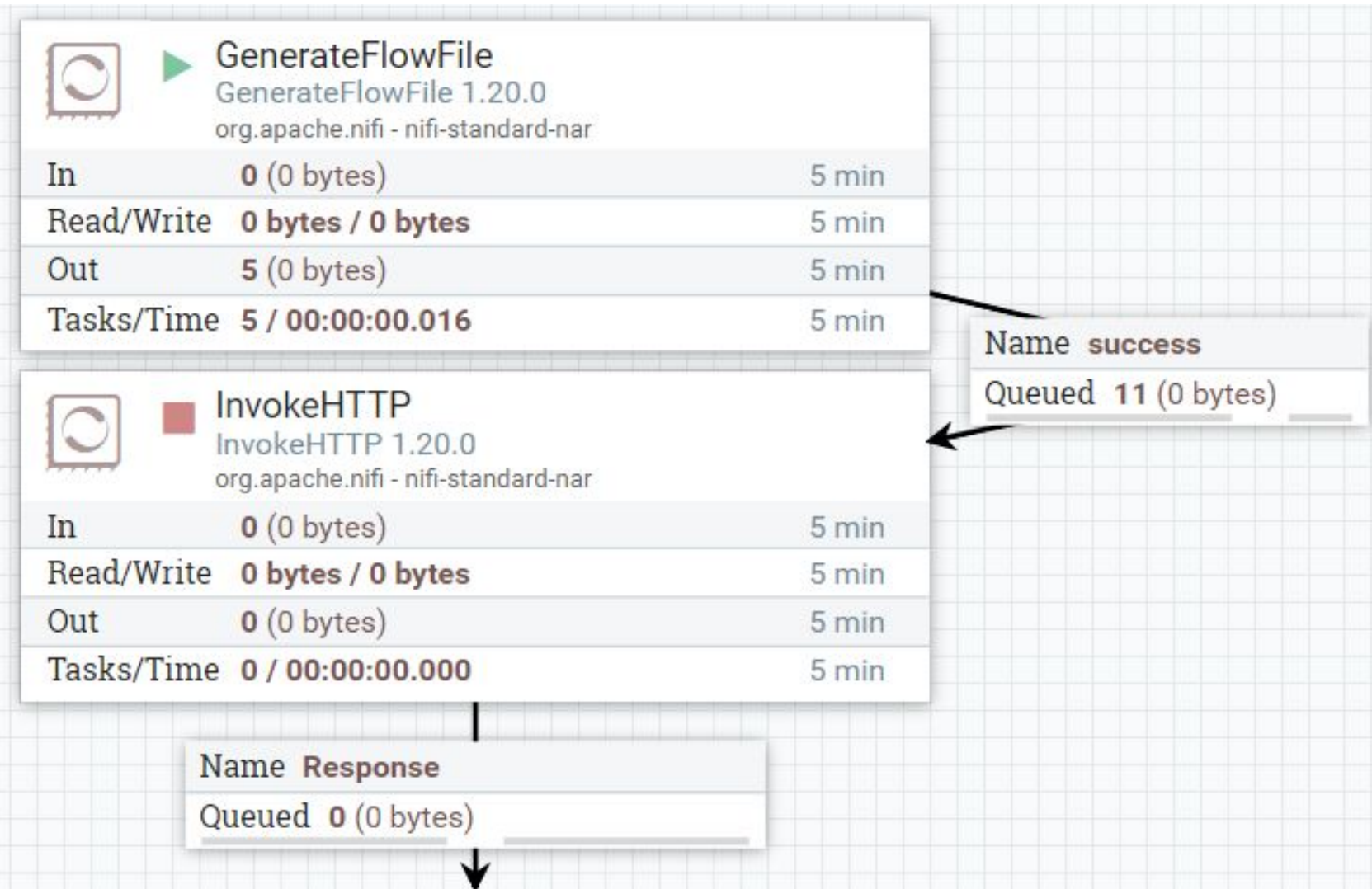
Получают данные из поступающего на **вход** файла, модифицируют **контент** или **атрибуты**

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!**

