ETL-инструменты на примере NiFi

NiFi для потоков обработки сообщений

Apache NiFi предоставляет возможность управления потоками данных из разнообразных источников в режиме реального времени с использованием графического интерфейса.

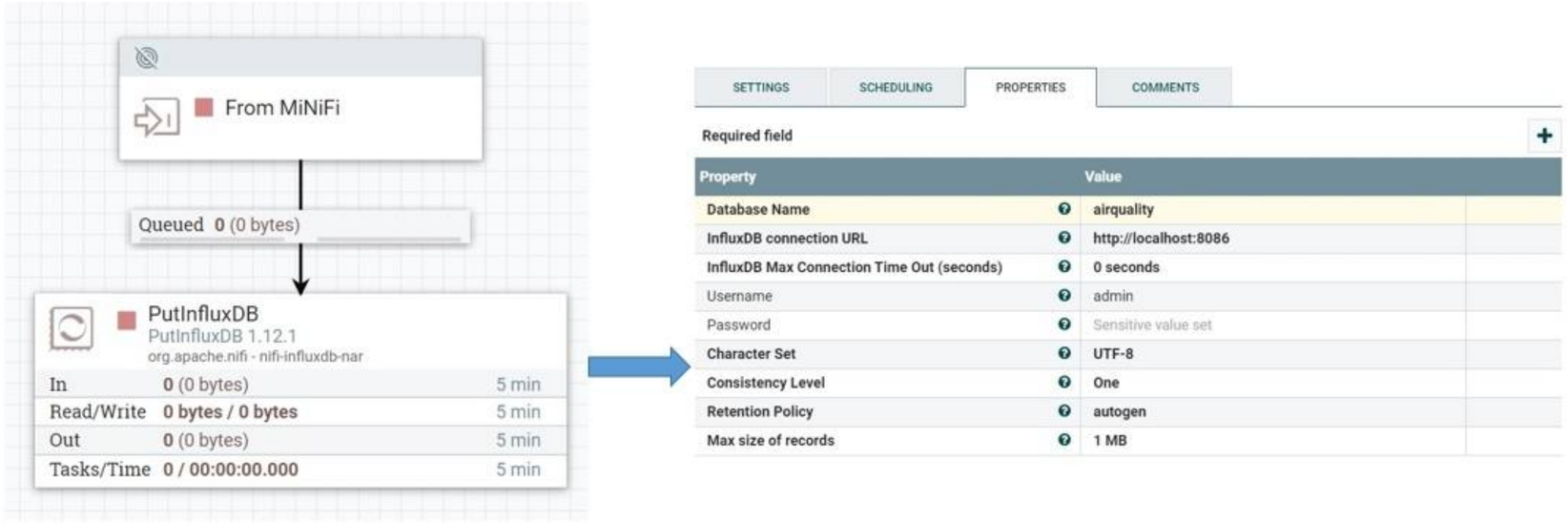
Интересные факты:

- NiFi базируется на технологии NiagaraFiles, ранее разрабатываемой агентством национальной безопасности США
- В ноябре 2014 года исходный код был открыт и передан Apache Software Foundation в рамках программы по передаче технологий NSA Technology Transfer Program.

Компоненты NiFi

NiFi состоит из трех основных компонентов:

- Flow Files – потоки файлов.
- Flow File Processor – своего рода «функции» с входами и выходами параметрами, которые выполняют широкий перечень типовых задач.
- Connections – задают направление движения FlowFiles между процессорами.



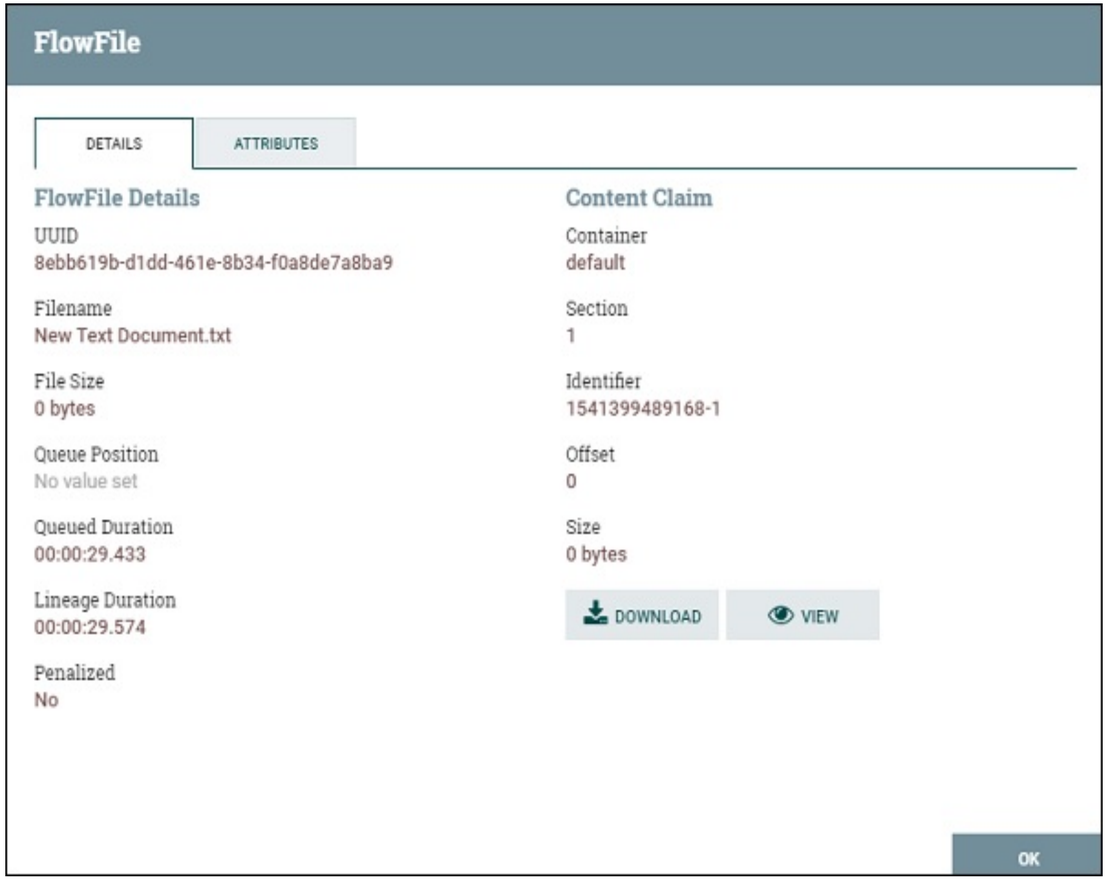
Flow Files – потоки файлов

Поточный файл является основным объектом обработки в Apache NiFi.

Он состоит из двух частей:

- Атрибуты, используемые процессорами NiFi для обработки данных, например:
  - UUID – идентификатор потокового файла
  - FileName – имя потокового файла
  - File Size – размер потокового файлы
  - mime.type - MIME-тип потокового файла

- Контейнер с данными (payload), которые обрабатываются процессорами.

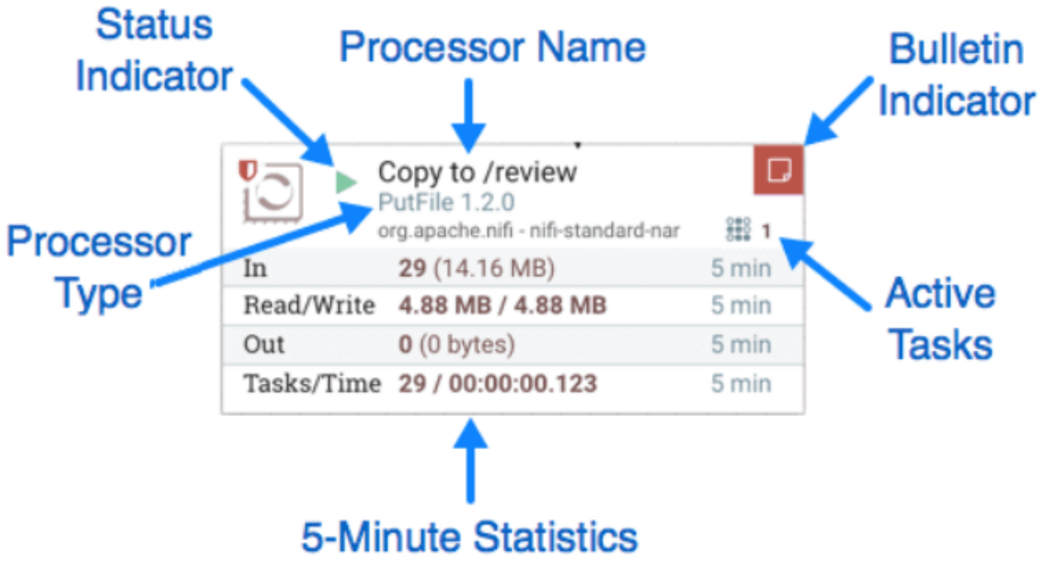


Flow File Processor – «функции»

Процессоры Apache NiFi являются основными блоками создания потока данных. Каждый процессор имеет разные функциональные возможности, что способствует созданию различный выходных потоковых файлов.

Отображаемые в GUI характеристики процессора:

- Processor Name – Пользовательское имя процессора
- Processor Type – Тип используемого процессора
- Active Tasks – Количество активных потоков
- Status Indicator – Статус процессора (включен/ выключен)



Виды Flow File Processor-ов

Название	Примеры
Процессоры загрузки данных	GetFile, GetHTTP, GetFTP, GetKafka, ConsumeKafka
Процессоры маршрутизации и посредничества	RouteOnAttribute, RouteOnContent, ControlRate, RouteText
Процессоры доступа к базам данных	ExecuteSQL, PutSQL, PutDatabaseRecord, ListDatabaseTables
Процессоры извлечения атрибутов	UpdateAttribute, EvaluateJSONPath, ExtractText, AttributesToJSON
Процессоры системного взаимодействия	ExecuteScript, ExecuteProcess, ExecuteGroovyScript, ExecuteStreamCommand
Процессоры преобразования данных	ReplaceText, JoltTransformJSON
Отправка процессоров данных	PutEmail, PutKafka, PutSFTP, PutFile, PutFTP
Процессоры расщепления и агрегации	SplitText, SplitJson, SplitXml, MergeContent, SplitContent
HTTP-процессоры	InvokeHTTP, PostHTTP, ListenHTTP

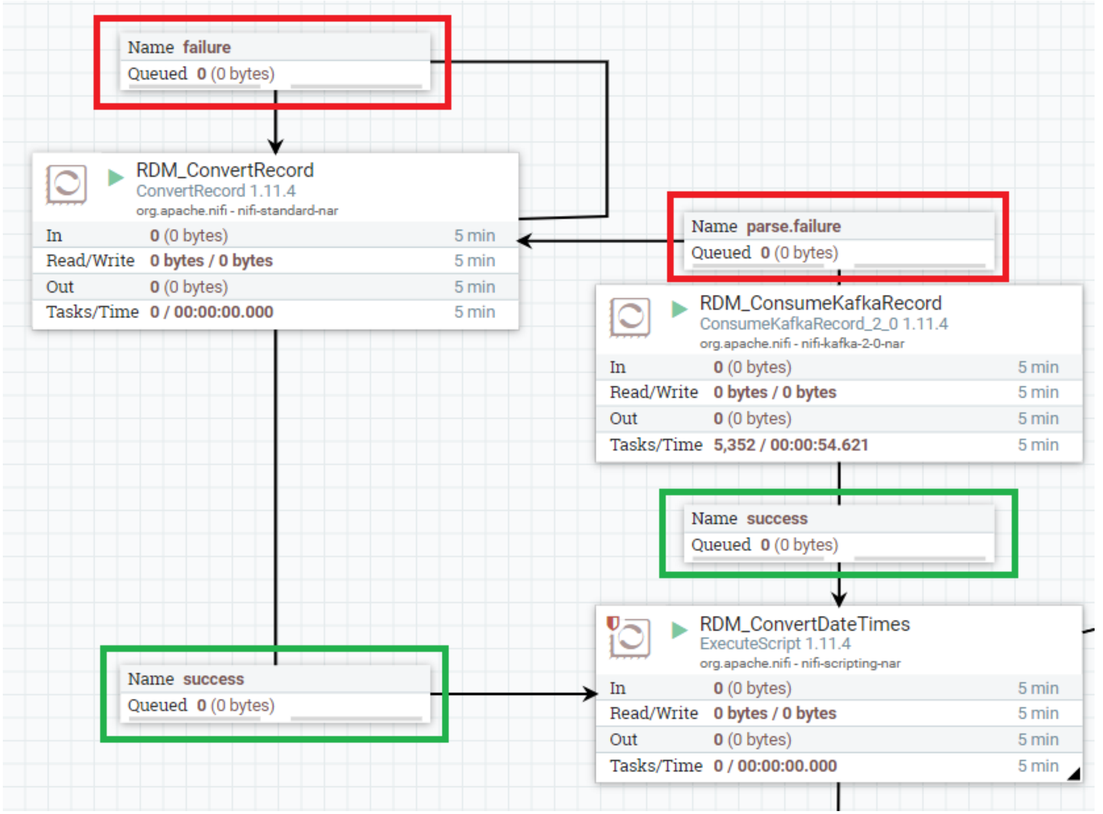
Connections – направление движения FlowFiles

В потоке данных Apache NiFi потоковые файлы перемещаются от одного процессора к другому через соединение.

Каждый раз, когда создается соединение, разработчик выбирает одно или несколько отношений между этими процессорами.

Самые частые соединения:

- Success
- Failure



NiFi Registry

NiFi Registry - система версионирования NiFi Flow. Как гит, только для ETL.

NiFi Registry состоит из двух основных компонентов:

- Flow - поток данных NiFi на уровне группы процессов, который был помещен под контроль версий и сохранен в реестре.
- Bucket - контейнер, который хранит и организует потоки.

Шаблон взаимодействия с NiFi Registry

Типовые действия, которые можно выполнять в NiFi Registry:

- Создать пакет
- Добавить процесс под версионный контроль (работает только над группами)
- Закоммитить NiFi Flow
- Импортировать или экспортировать NiFi Flow