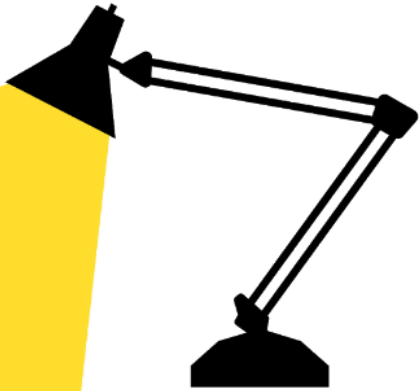
Как сделать больно NIFI.

4 задачи, перед которыми стоит основательно подумать, как их делать в NIFI.

Бронислав

09.2023

О себе



- Работаю в IT с 2004
- В разработке с 2005
- Архитектор в Тинькофф
- Инструмент извлечения сырых данных

О чем поговорим

- О NIFI в Tinkoff
- О том как работает NIFI
- И о 4 кейсах которые заставляют страдать:

#1 Ручной запуск

#2 Увеличение мощности

#3 Join

#4 Консистентный батч

Инструмент извлечения сырых данных

инструмент извлечения сырых данных

О NiFi в Тинькофф

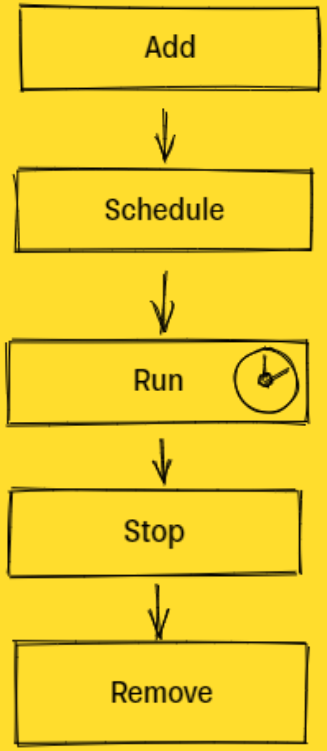
- Запустили NiFi в 2018 году
- Обрабатываем терабайты сырой информации ежедневно
- Создали более 400 бизнес-потоков
- Кастомизируем ядро NiFi

Процессоры

Процессор – элемент отвечающий за обработку данных

Генератор – процессор не имеющий входящией очереди и создающий новые FlowFile в потоке.

ЖЦ Процессора



Добавление процессора в поток

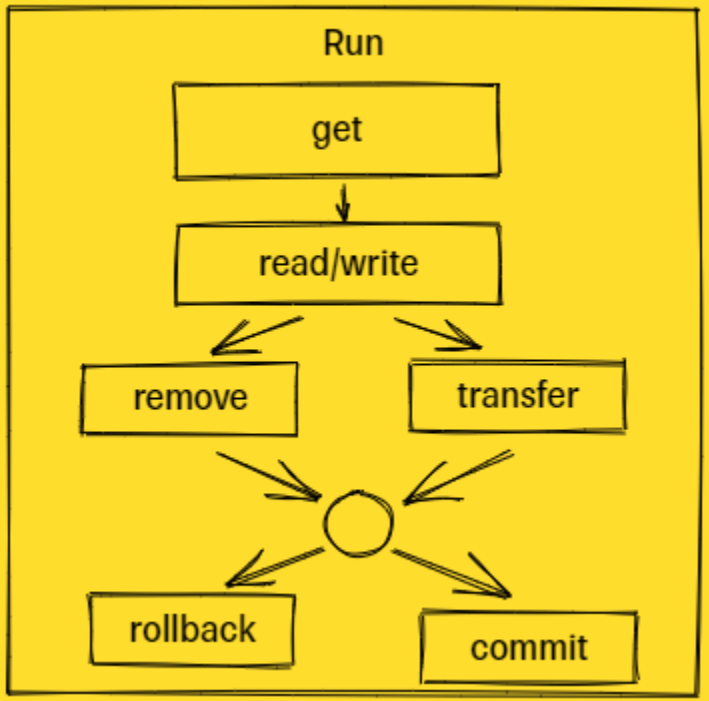
Запуск расписания для процессора

Исполнение процессора согласно расписание

Снятие процессора с расписания

Удаление процессора из потока

Исполнение

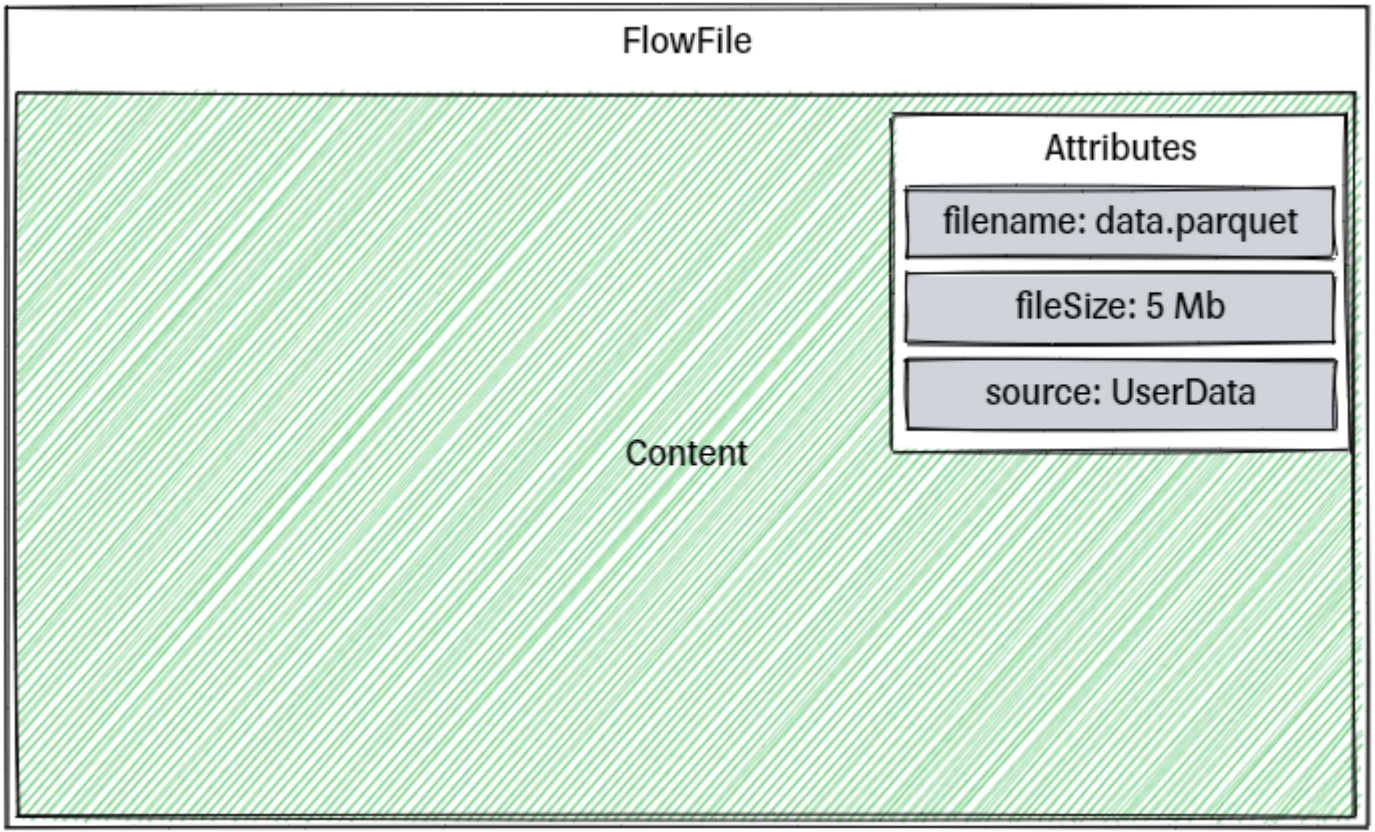


Запись в новый файл

Транзакционный переход

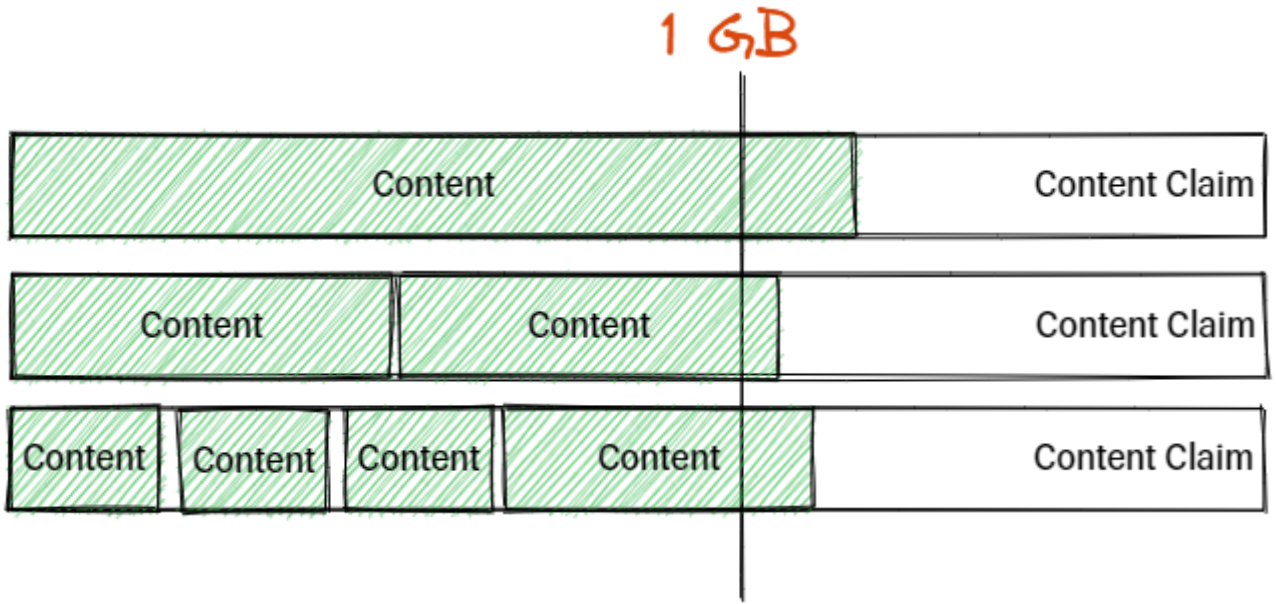
Проверка полной обработки

Про FlowFile



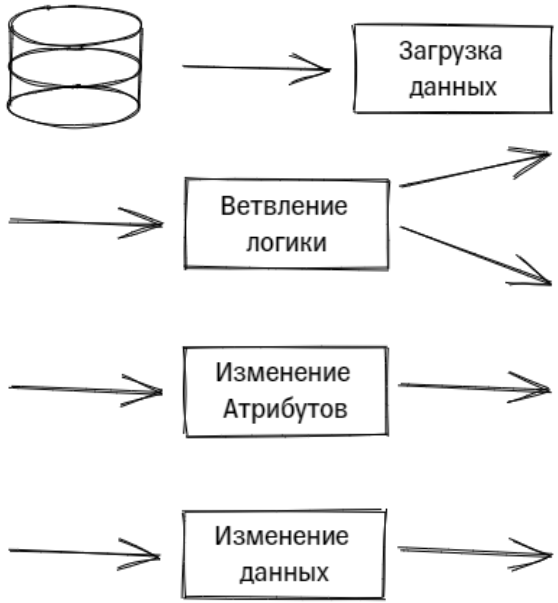
Структура хранения

Content Repository



Изменение FlowFile

Операция



Content Repository

Создаем Content

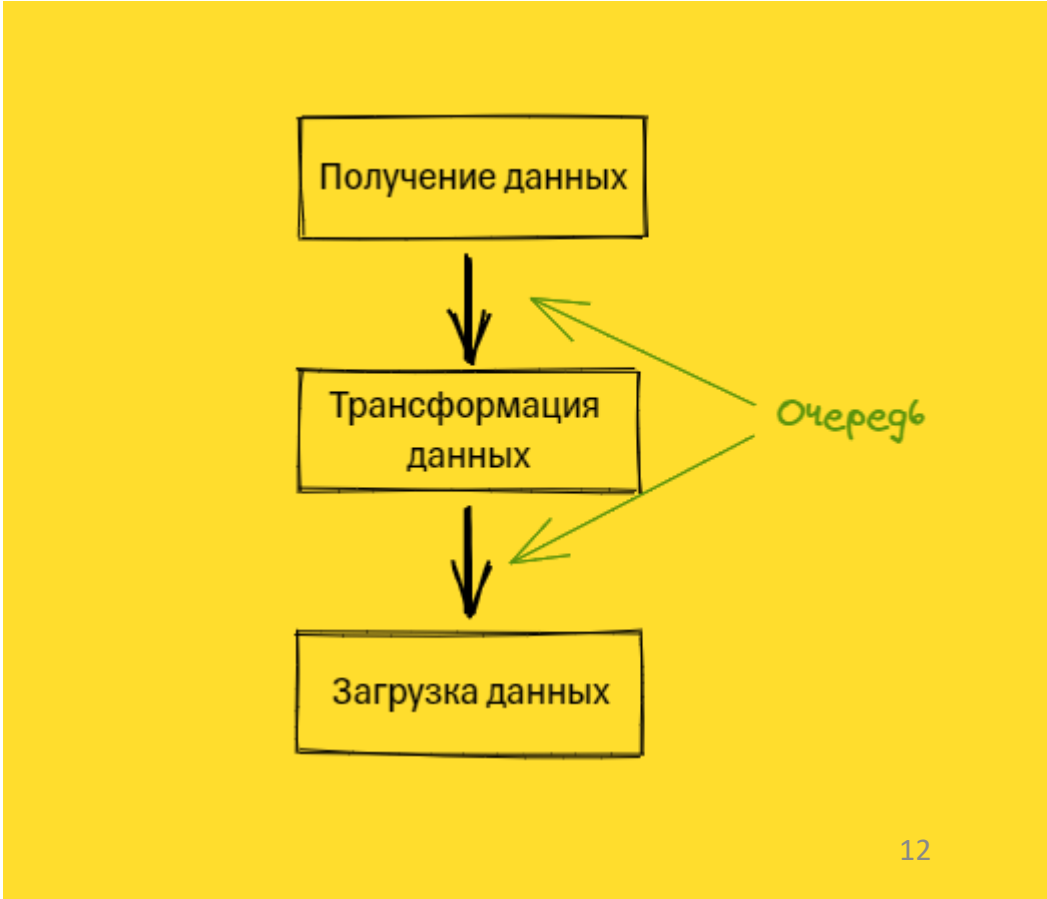
Один Content для нескольких файлов

Content не меняется

Создается новый Content на основании старого

Про очереди

- Хранение FlowFile
- Приоритезация
- Балансировка



Что такое NiFi

- Поток FlowFile
- Файл это контекст

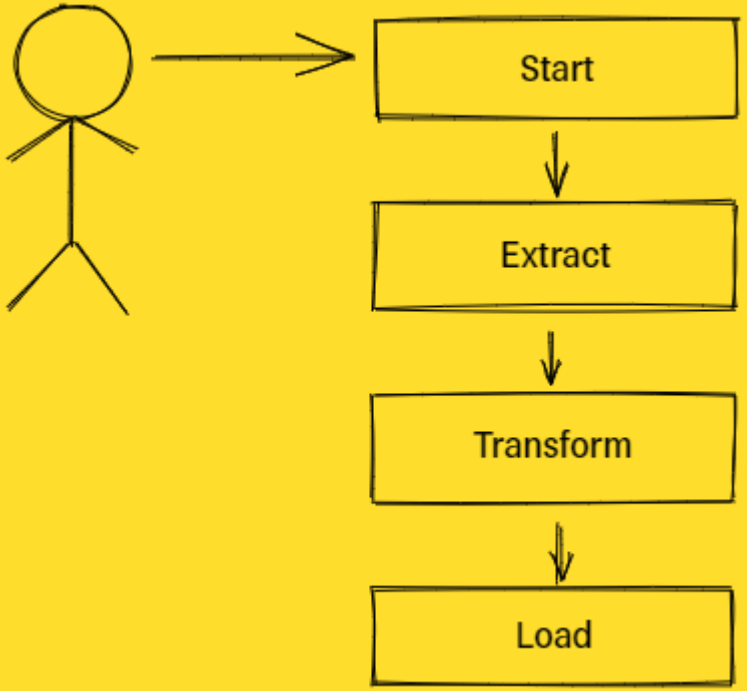
Что такое NiFi

Кейс 1

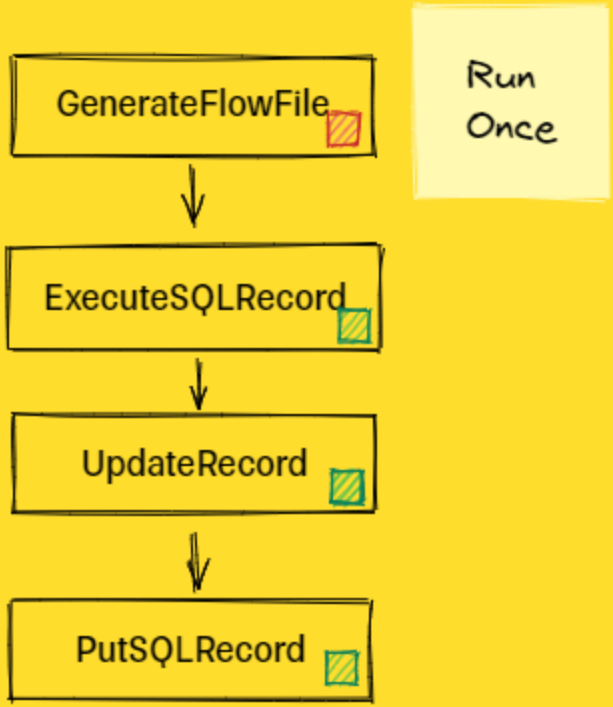
- У нас есть система формирующая данные
- Формирование запускает операционист
- Файл нужно загрузить в хранилище

Ручное управление

- Отсутствие события
- Отсутствие расписания
- Запуск по кнопке



Ручное управление



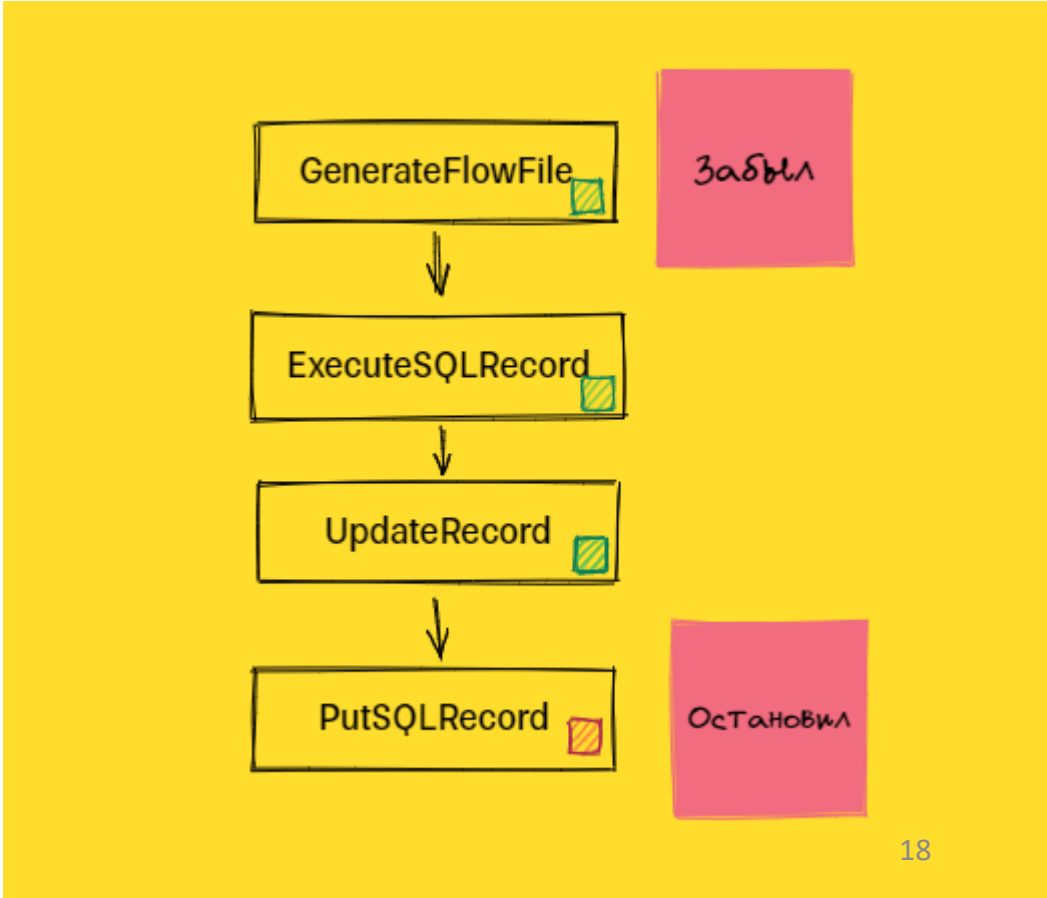
Выключаем генератор

Доступ к среде

Действие «RunOnce»

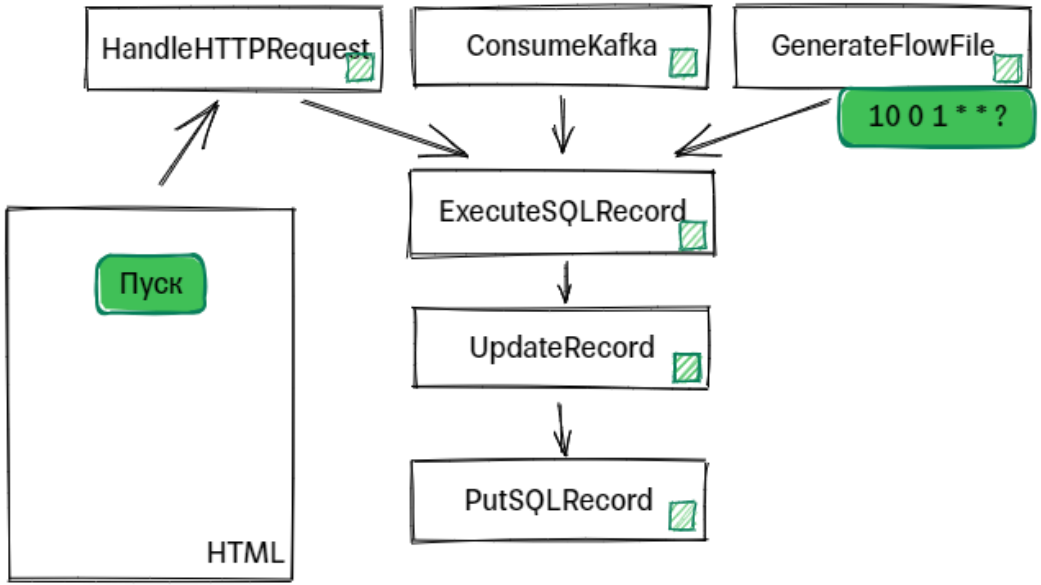
Какие проблемы

- Правки
- Мониторинг
- Контроль



Что делать

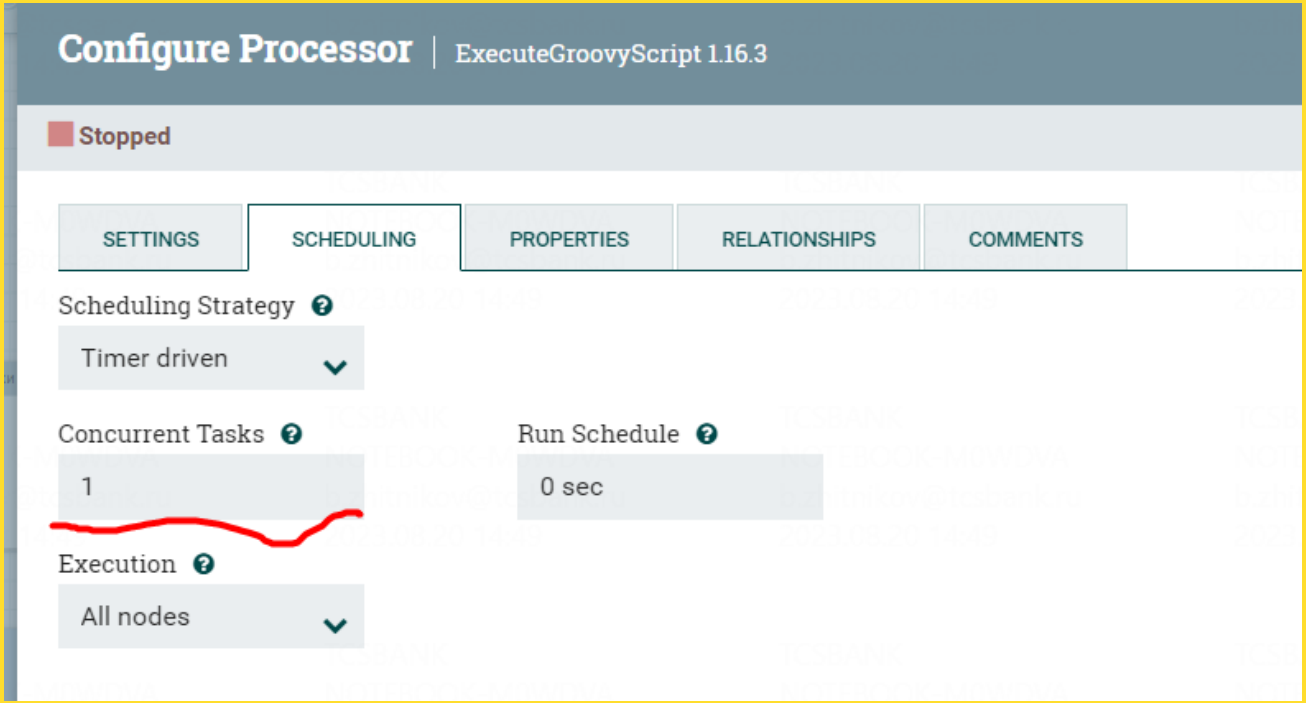
- Ищем событие
- Определяем расписание
- Кнопки управления



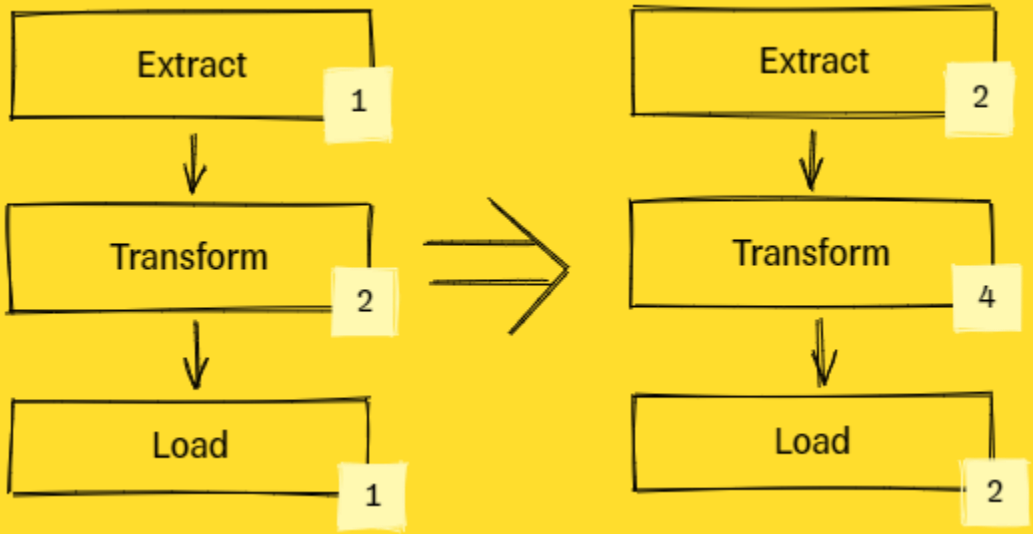
Кейс 2

- Мы успешно запустили MVP
- Данных стало больше
- Скорость не достаточна для обработки

Concurrent Task



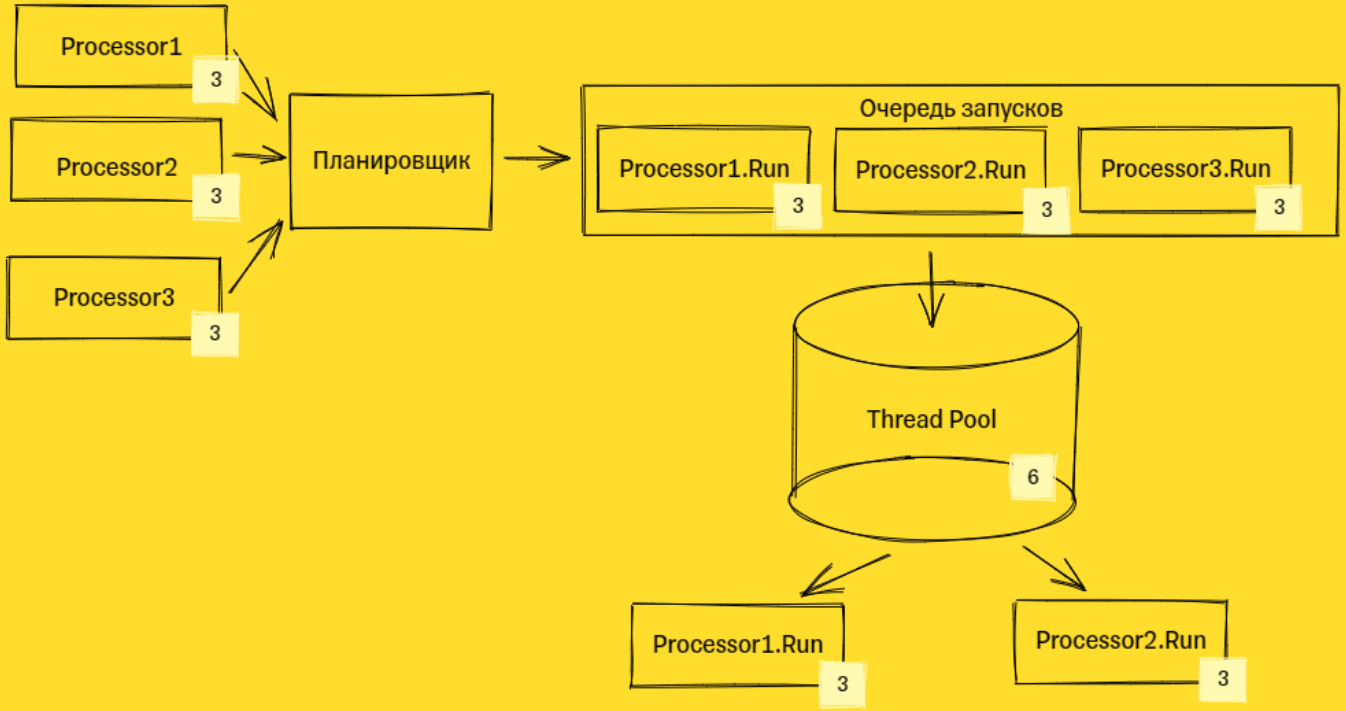
Ускорение работы



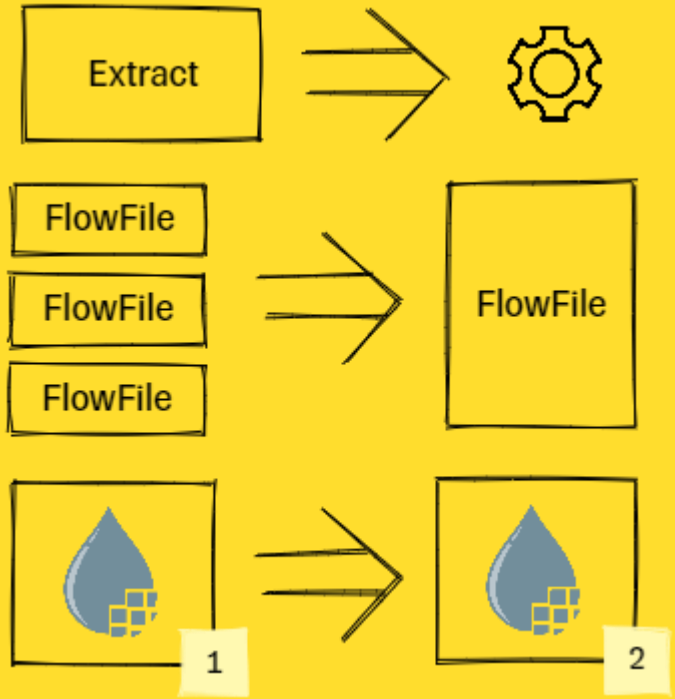
У нас больше данных

Обработка быстрее

Планировщик



Как можно по другому

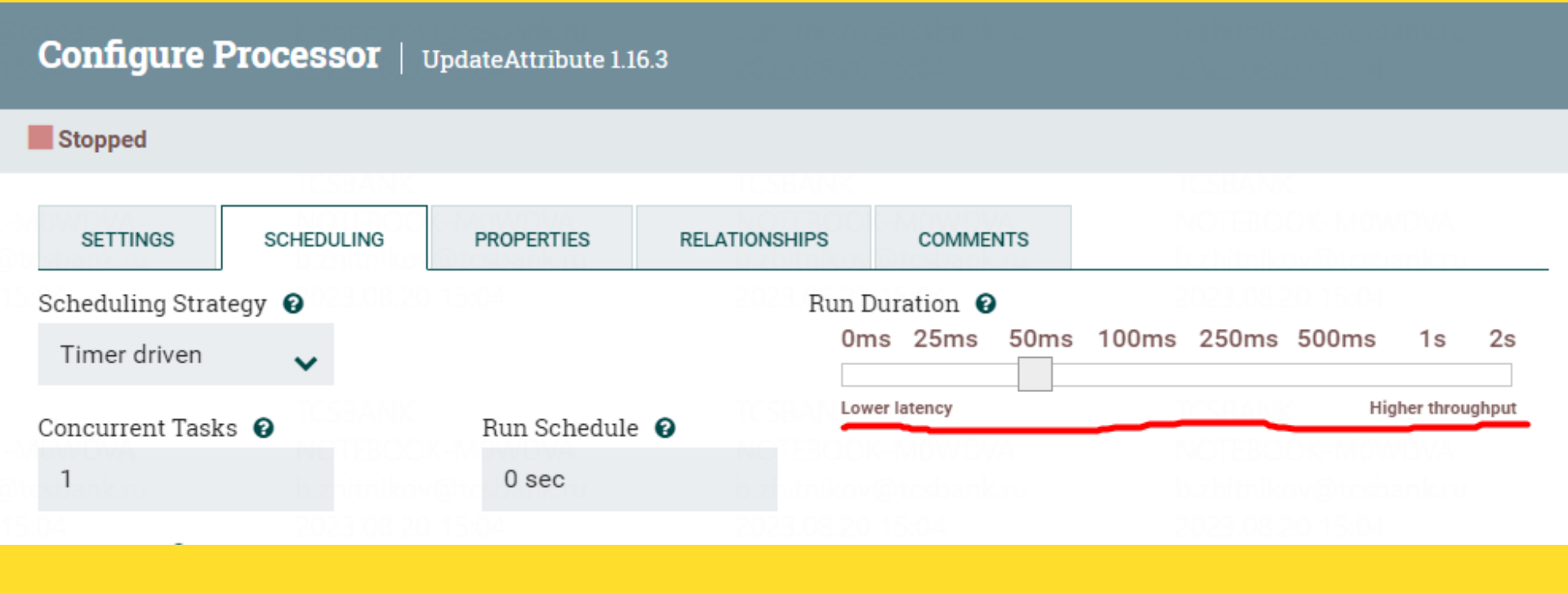


Настройки

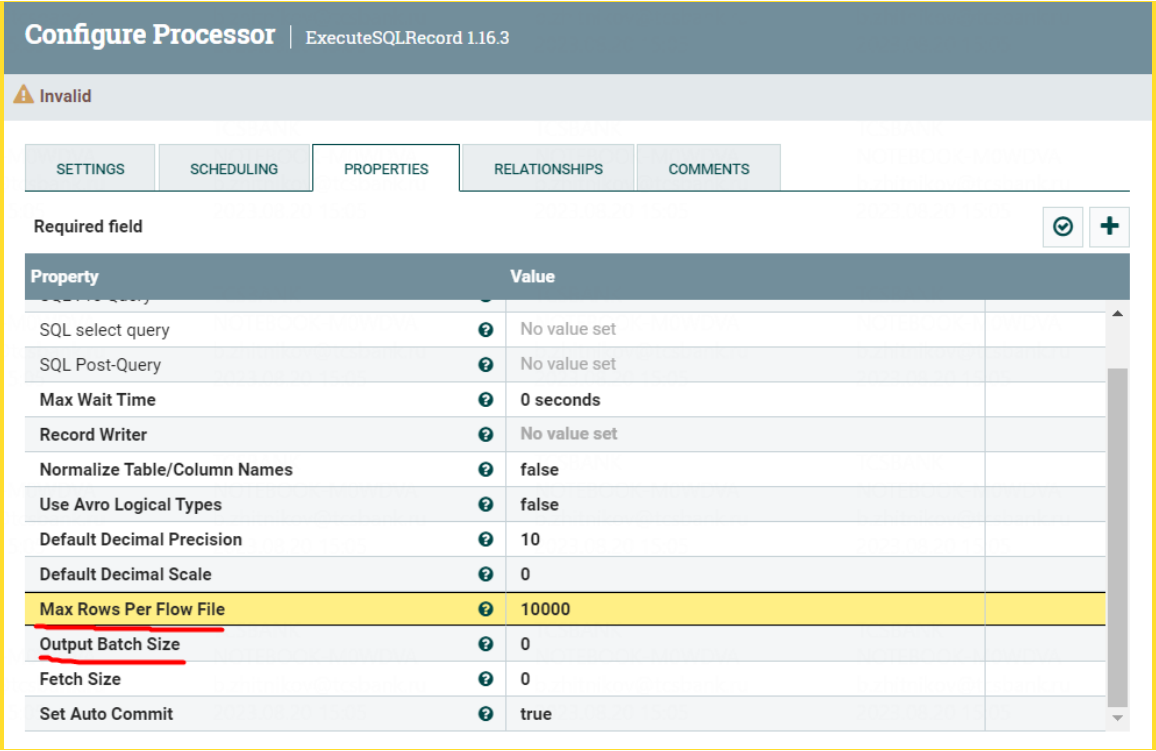
Микро-батчингу

Кластер

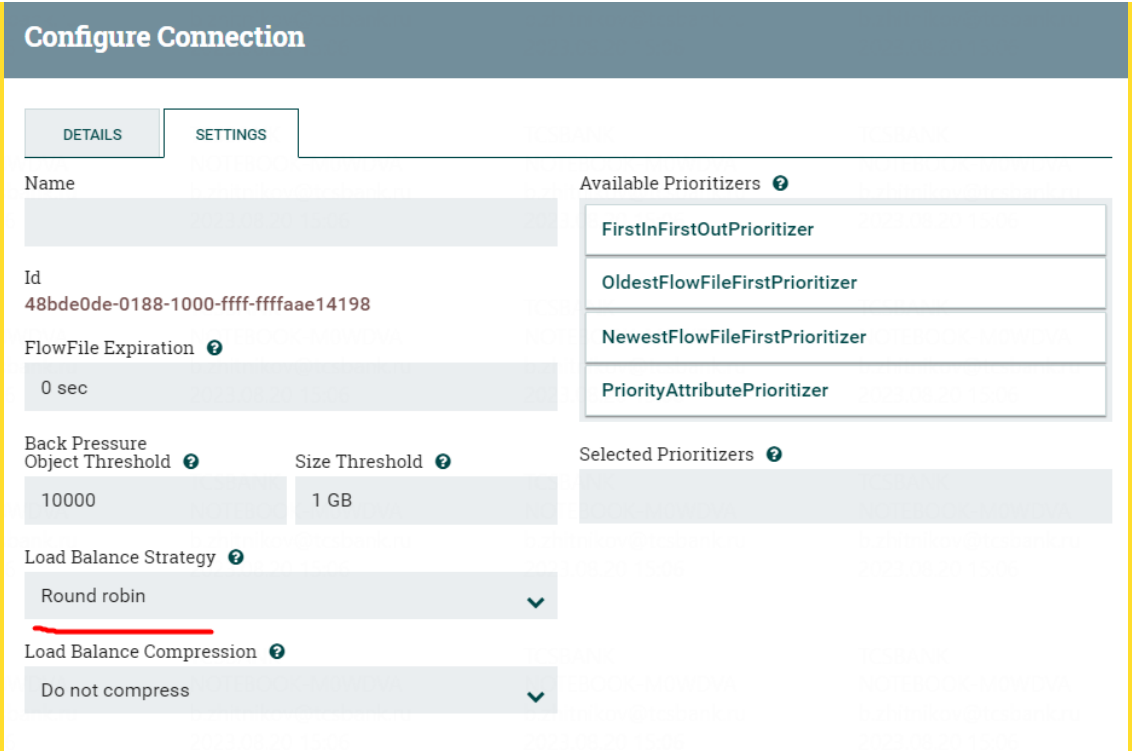
Настройки производительности



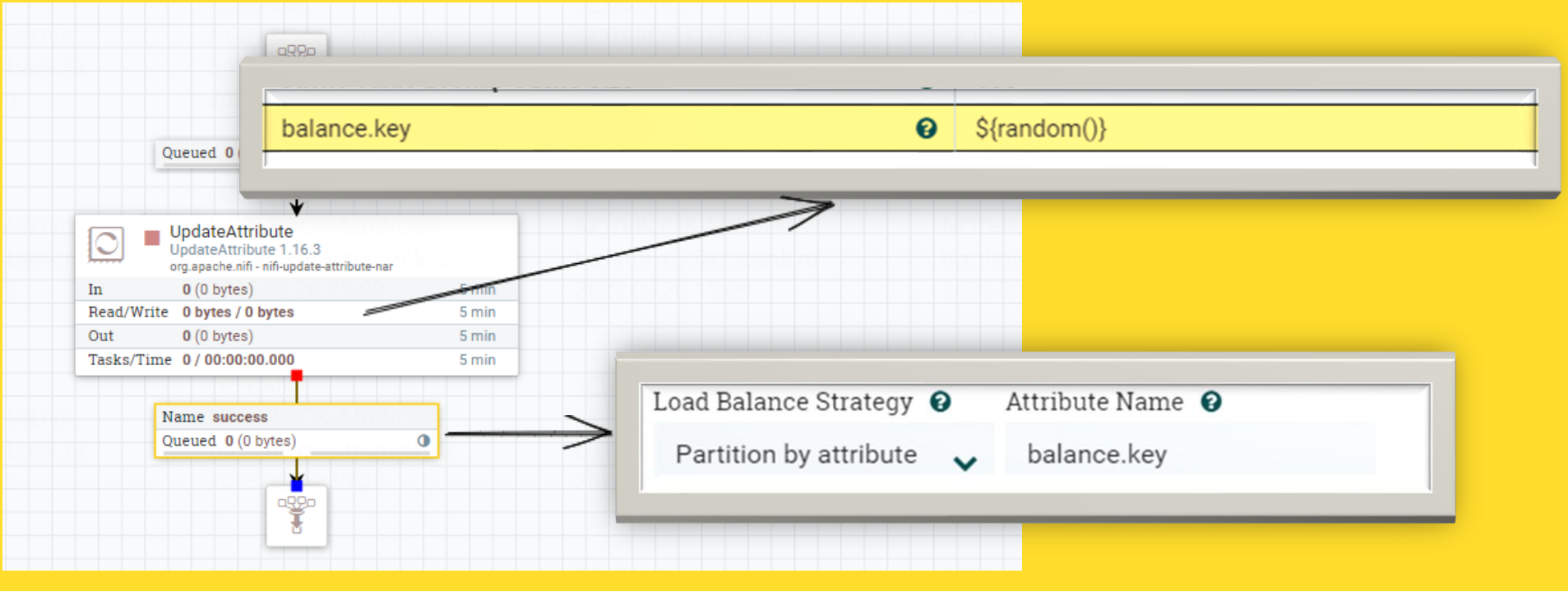
Настройки производительности



Балансировка



Балансировка

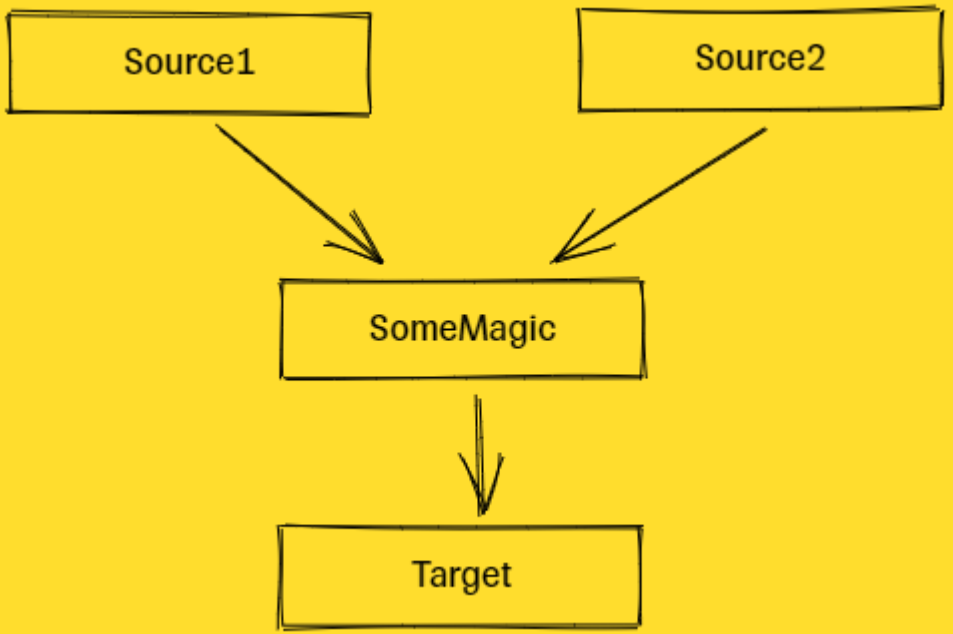


Кейс 3

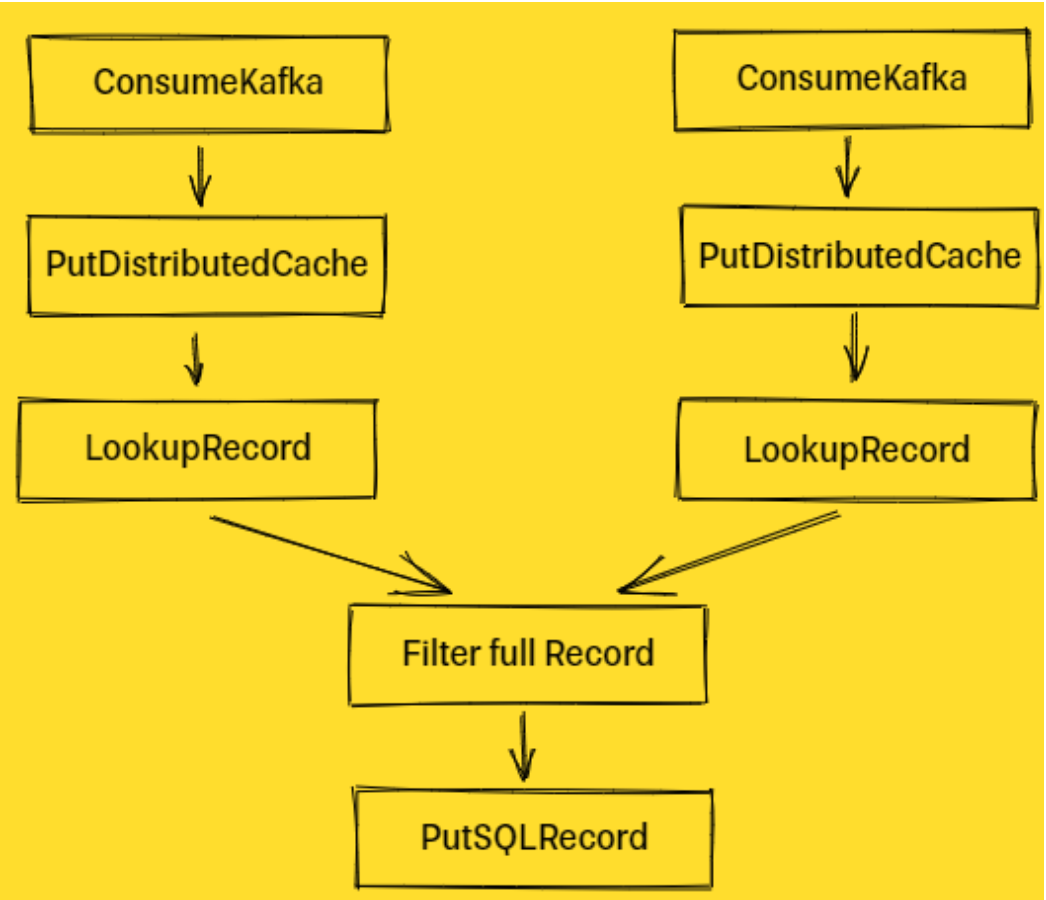
- Два потока данных с документами
- ID документа присутствует в двух потоках
- Сохраняем только документы которые прошли по двум потокам

Join данных

- Несколько потоков
- Несколько источников
- Связь по ключу
- Данные приходят в одно время



Join данных



Распределенный Cache

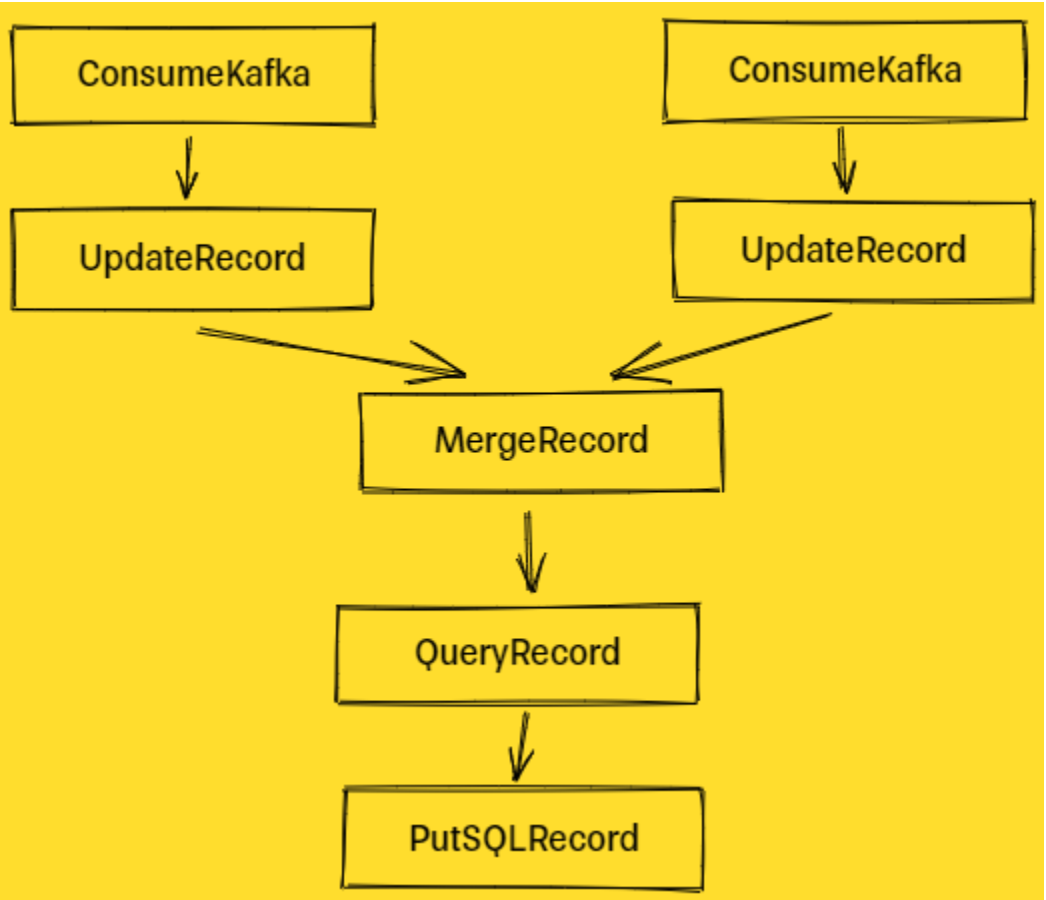
LookupRecord

Объединение

Какие проблемы

- Хранение событий
- Сложная логика
- Много сетевого взаимодействия

Join данных



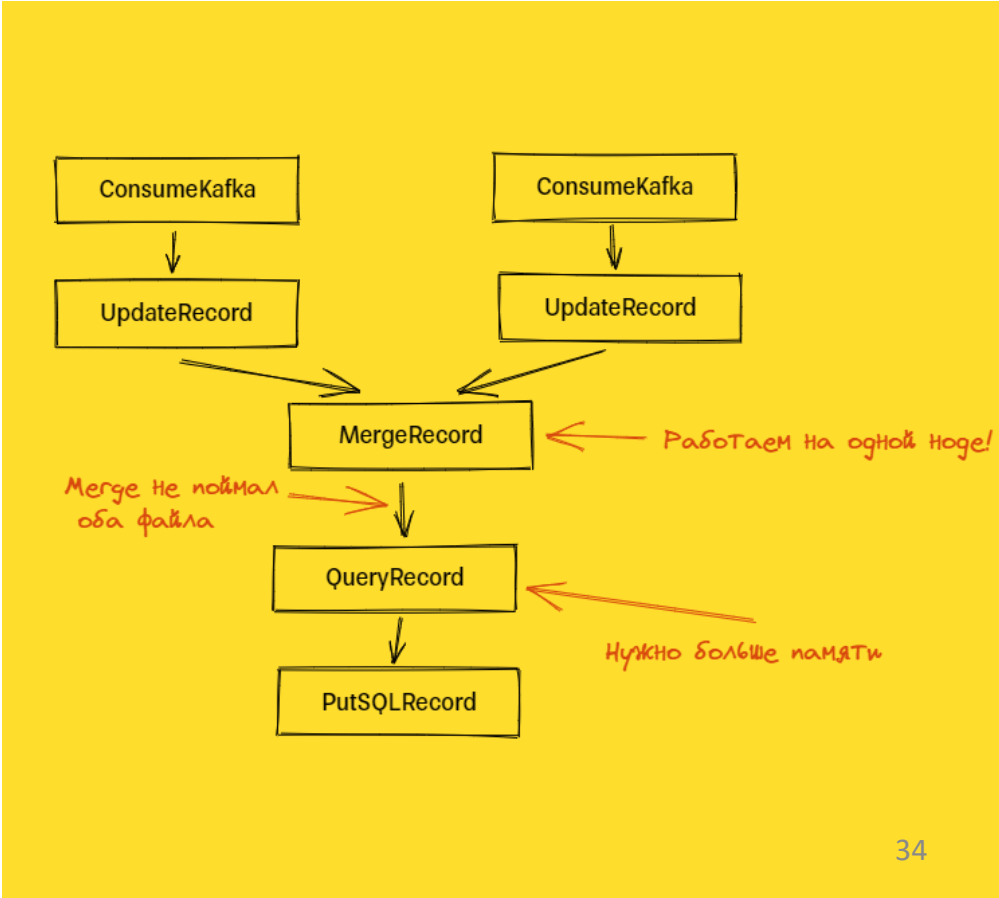
Добавили поле с Источником

Собрали в один файл

Отфильтровали в QueryRecord

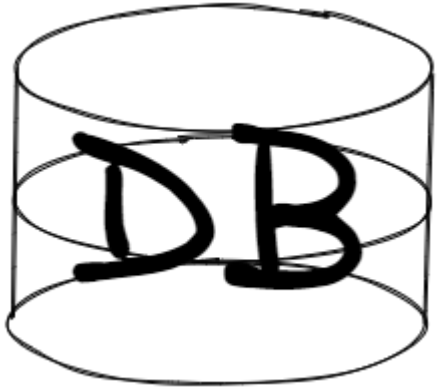
Какие проблемы

- Гарантии Merge
- Память в QueryRecord
- Отсутствие распределения



Что делать

- Выполнять Join в Target
- Создать промежуточную систему
- Использовать другой ETL инструмент

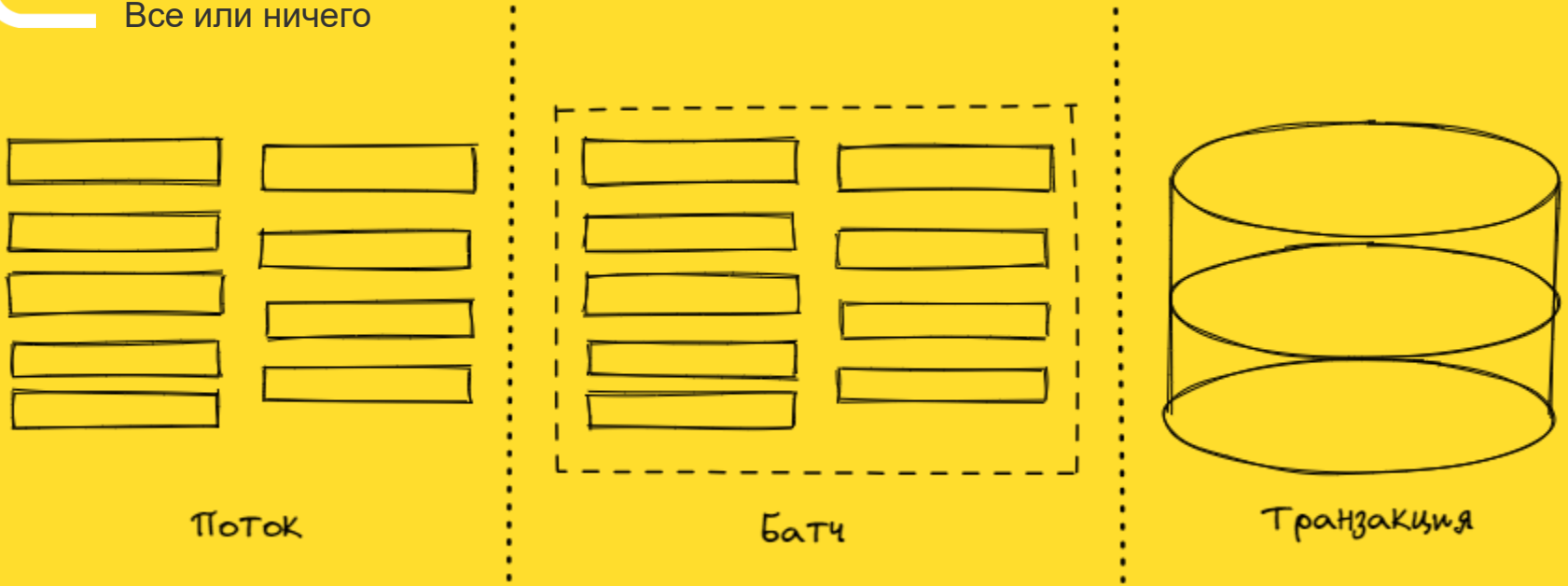


Кейс 4

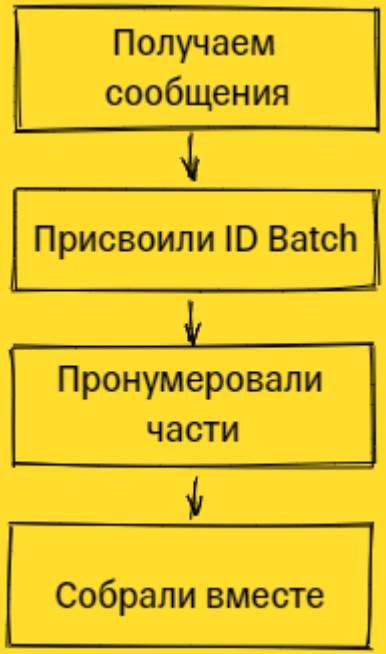
- Есть движок расчетов работающий как Batch
- В движке формируют набор данных для выгрузки
- Нужно загрузить в хранилище результаты расчета целиком

Консистентный Batch

Все или ничего



Консистентный батч

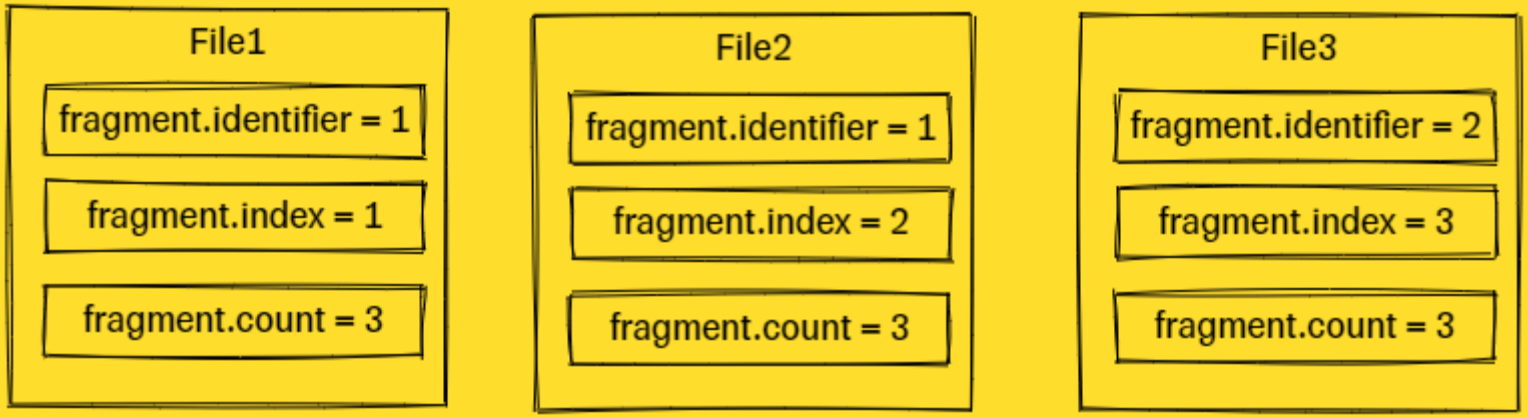


Определение границ

Нумерация частей

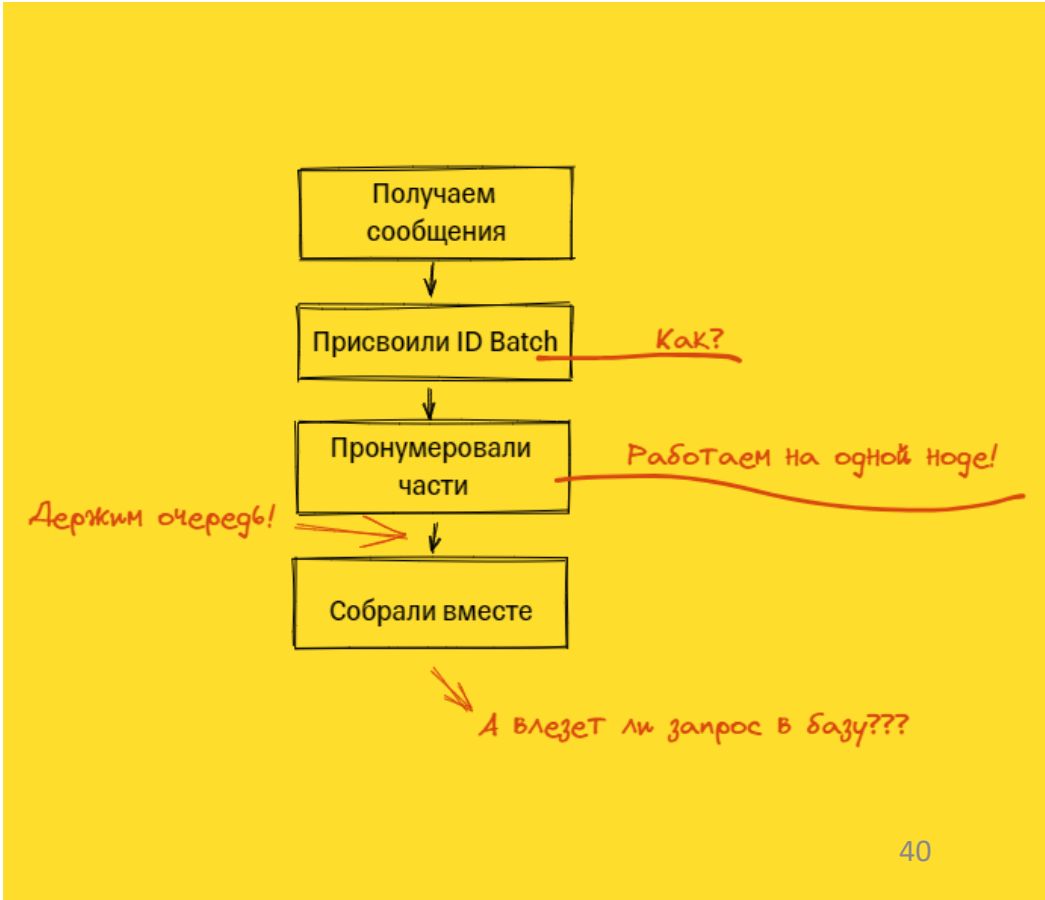
Merge процессоры

Merge Defragment



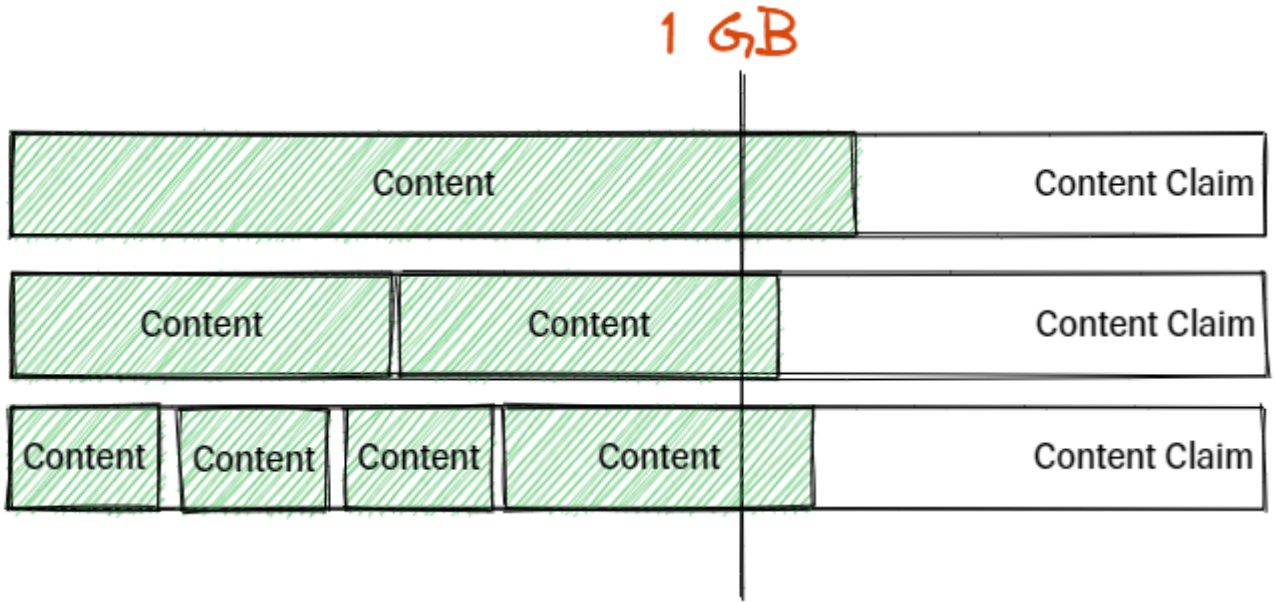
Какие проблемы

- Идентификация Batch
- Распределение работы
- Накопление очередей
- Большие файлы



Структура хранения

Content Repository



Что делать

- Отдельный планировщик
- Batch инструмент

Итоги

NIFI может очень многое

При добавлении фантазии почти все

Но лучше использовать инструмент правильно
И использовать правильный инструмент

Ваши вопросы

Можно задать сейчас

Или позже в группе  nifiusers