

# ТИН

## Как сделать больно NIFI.

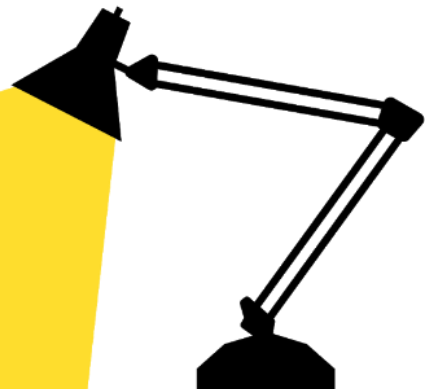
4 задачи, перед которыми стоит основательно подумать, как их делать в NIFI.

Бронислав

09.2023

# КОЕЕ

# 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

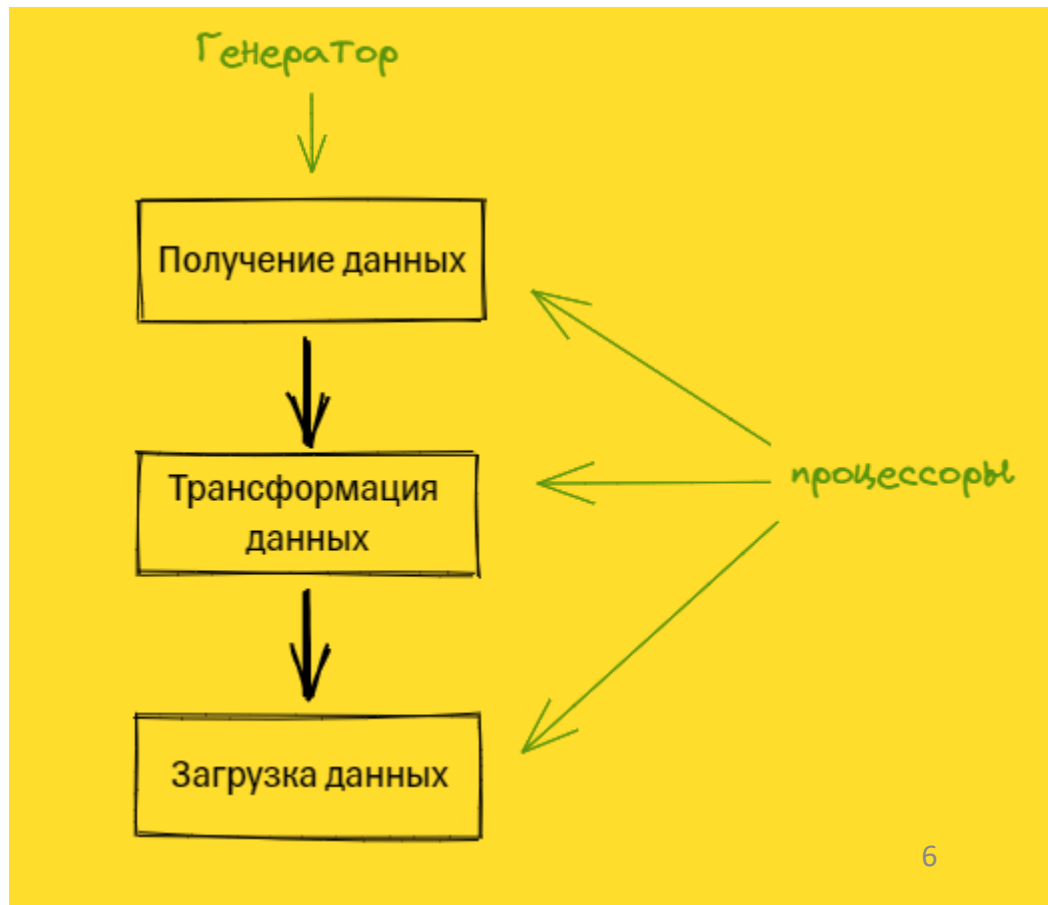
Создали более 400 бизнес-поток

Кастомизируем ядро NiFi

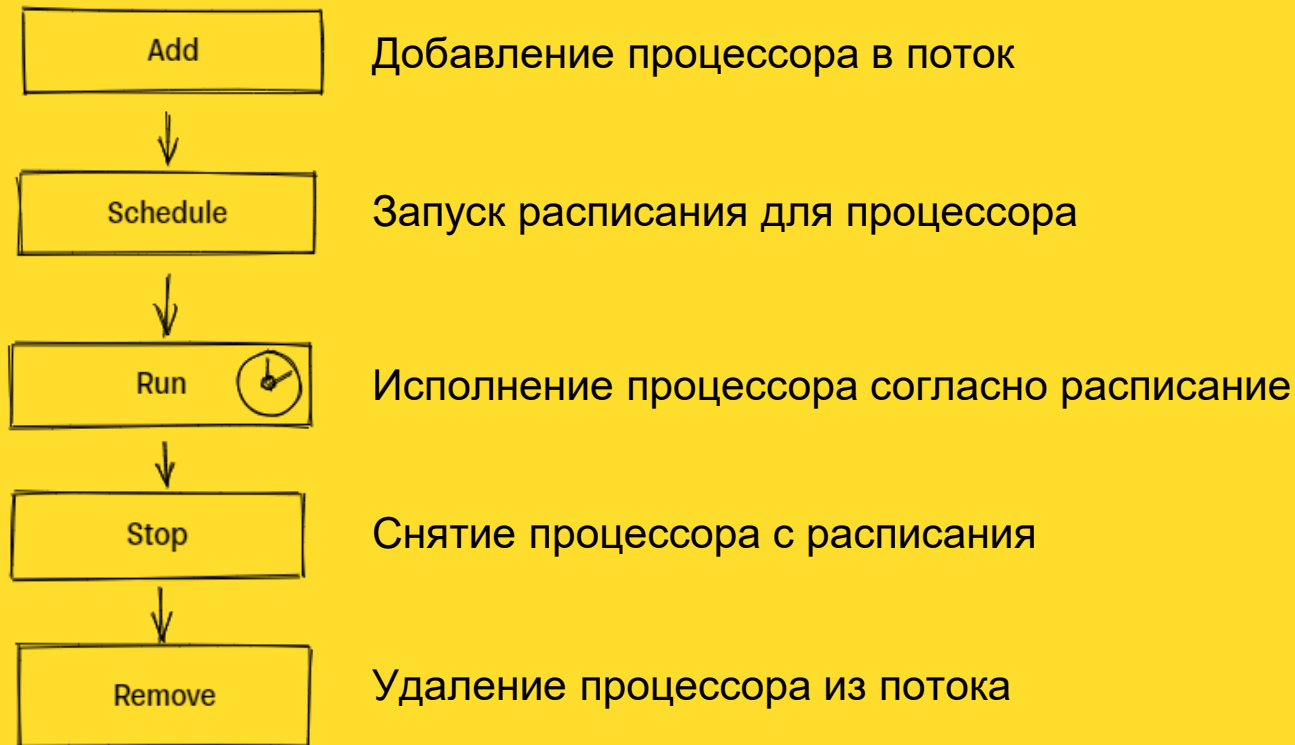
# Процессоры

Процессор – элемент отвечающий за обработку данных

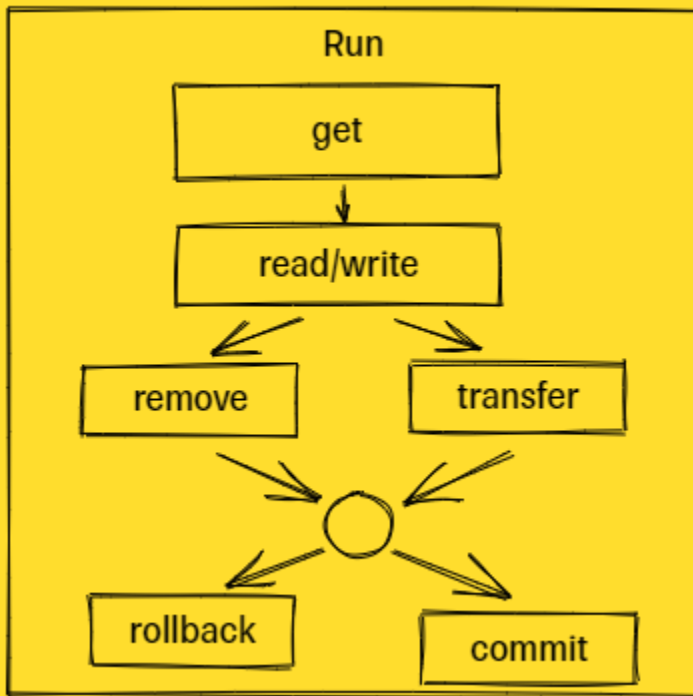
Генератор – процессор не имеющий входящей очереди и создающий новые FlowFile в потоке.



# ЖЦ Процессора



# Исполнение



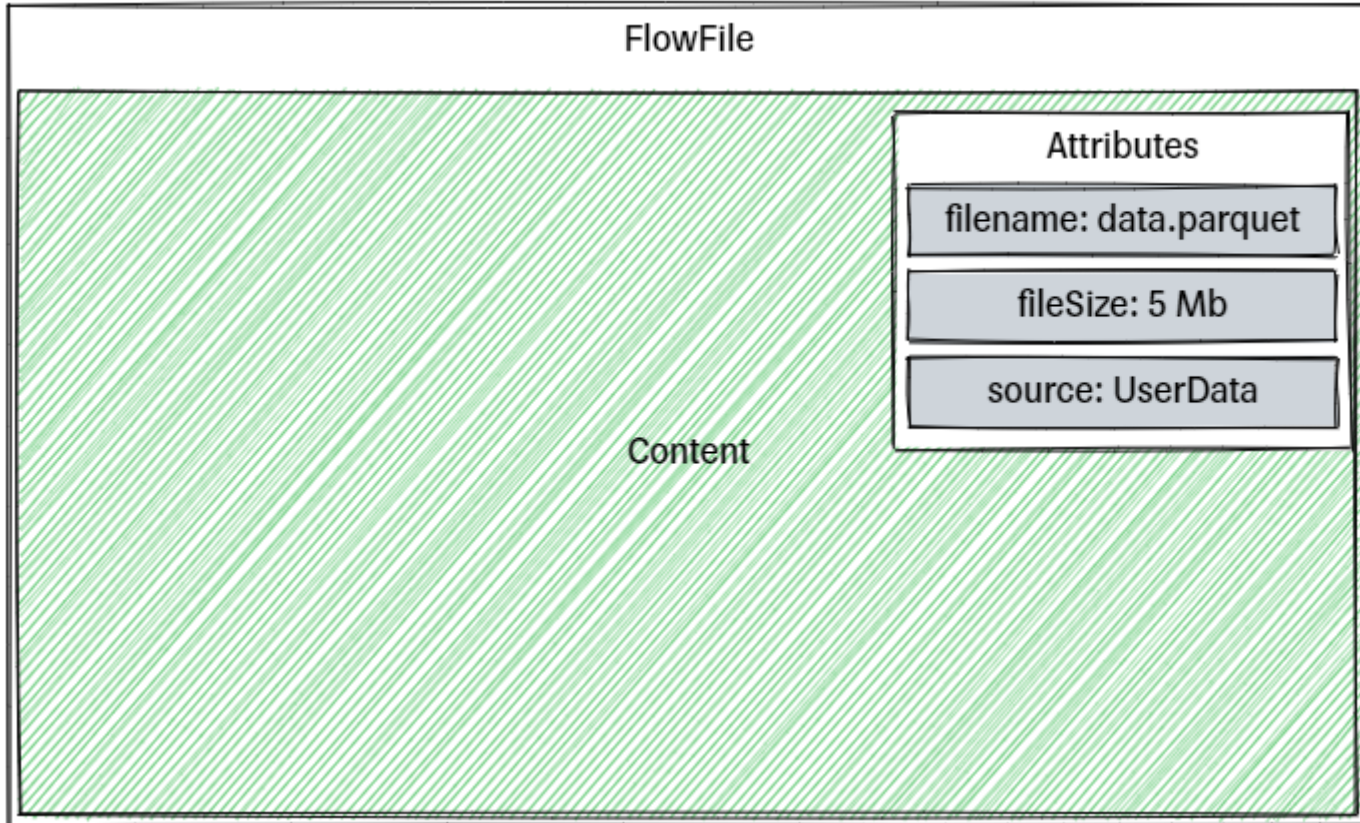
Запись в новый файл

Транзакционный переход

Проверка полной обработки

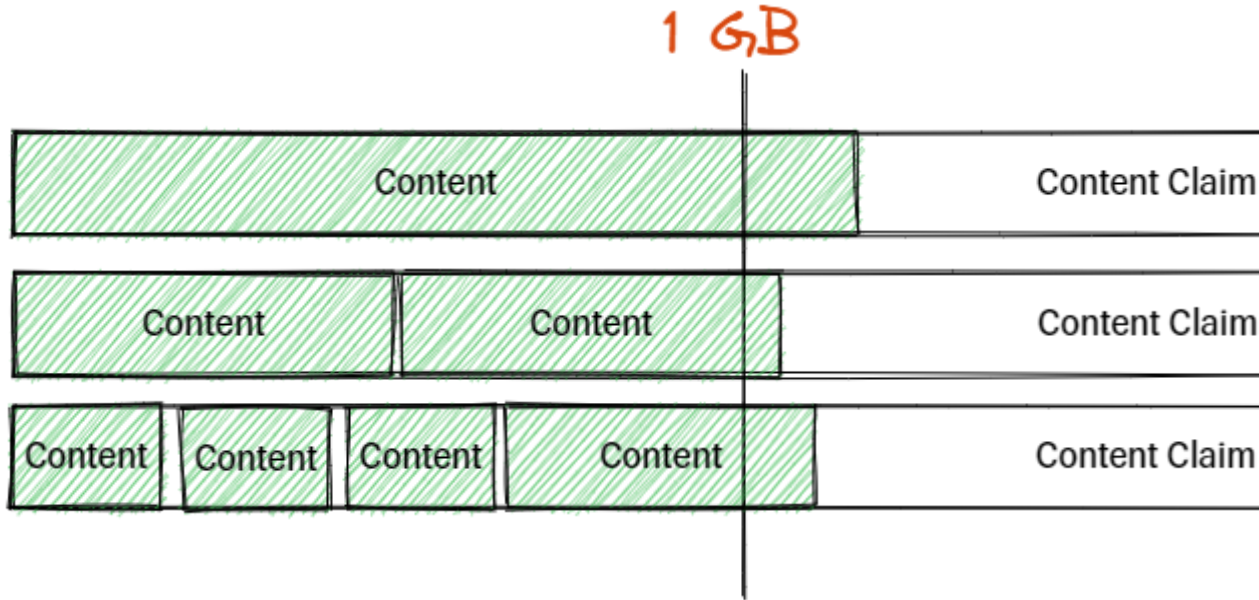


# Προ FlowFile



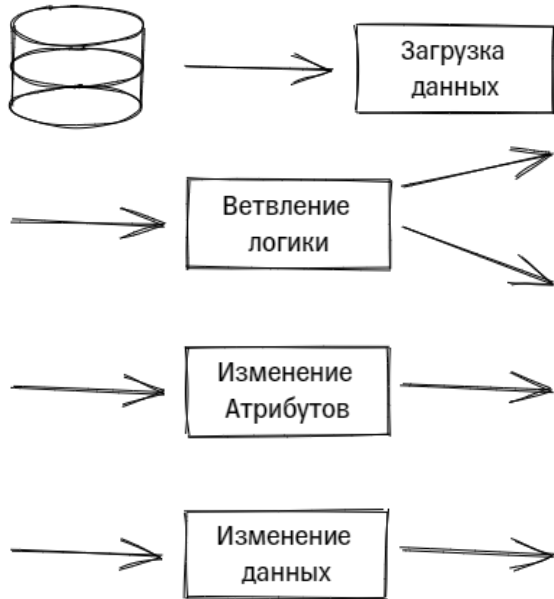
# Структура хранения

Content Repository



# Изменение FlowFile

Операция



Content Repository

Создаем Content

Один Content для нескольких файлов

Content не меняется

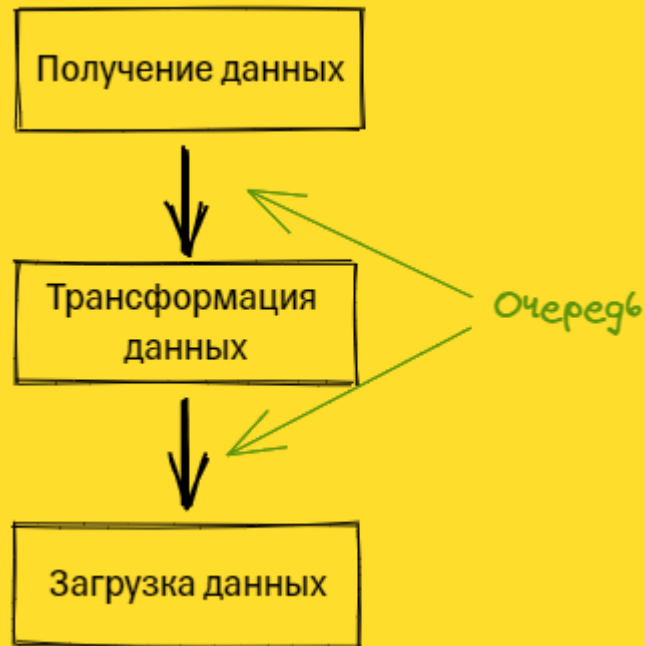
Создается новый Content на основании старого

# 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

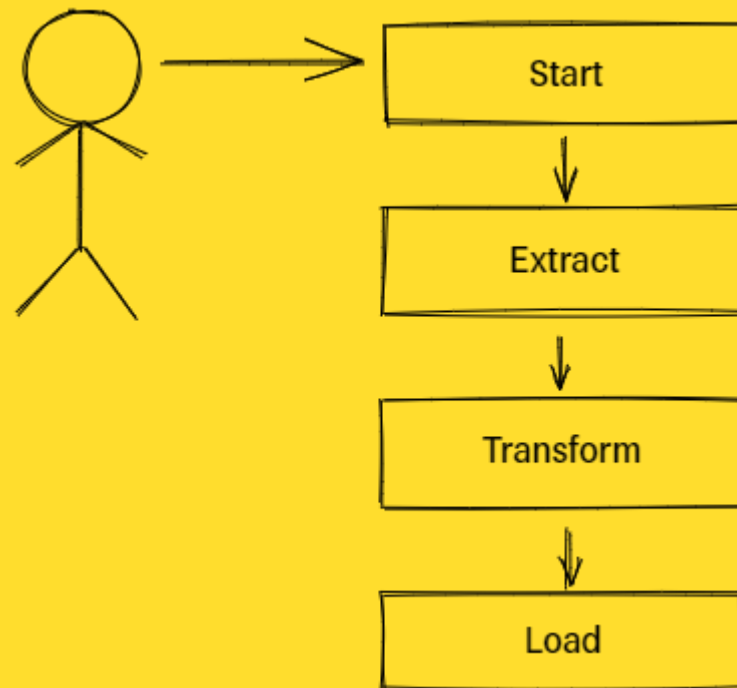


# Ручное управление

Отсутствие события

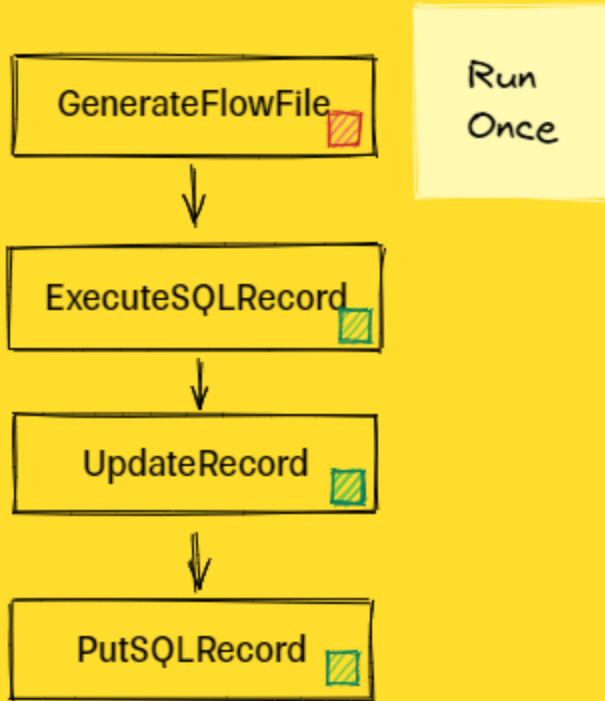
Отсутствие расписания

Запуск по кнопке





# Ручное управление



Выключаем генератор

Доступ к среде

Действие «RunOnce»

# Какие проблемы

Правки

Мониторинг

Контроль

GenerateFlowFile



ExecuteSQLRecord



UpdateRecord



PutSQLRecord

Забыл

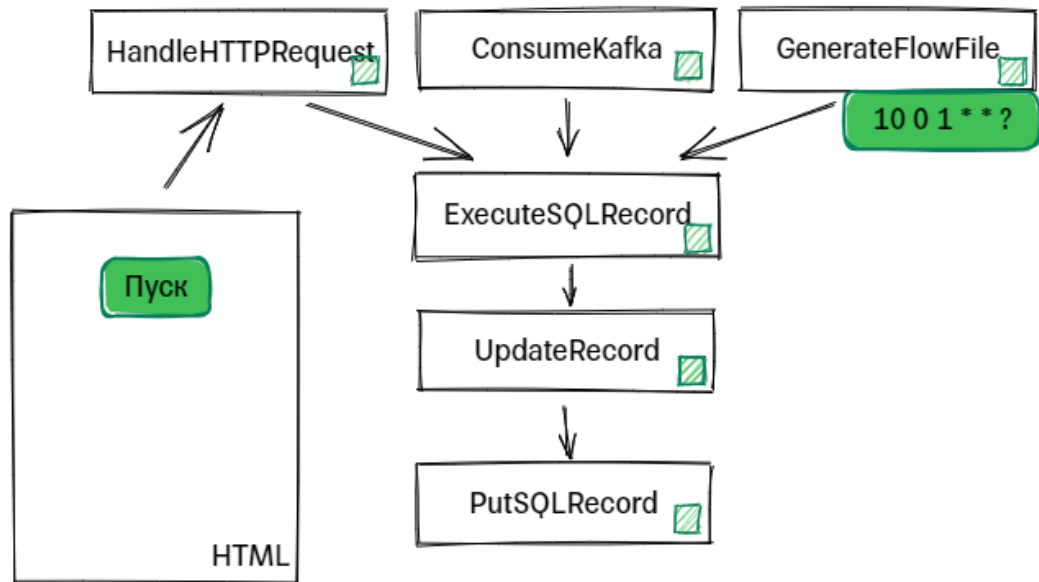
Остановил

# Что делать

Ищем событие

Определяем расписание

Кнопки управления



# Кейс 2

Мы успешно запустили MVP

Данных стало больше

Скорость не достаточна для обработки





# Concurrent Task


**Configure Processor** | ExecuteGroovyScript 1.16.3

Stopped


SETTINGS | SCHEDULING | PROPERTIES | RELATIONSHIPS | COMMENTS

Scheduling Strategy 


Timer driven 


Concurrent Tasks 


1

Run Schedule 

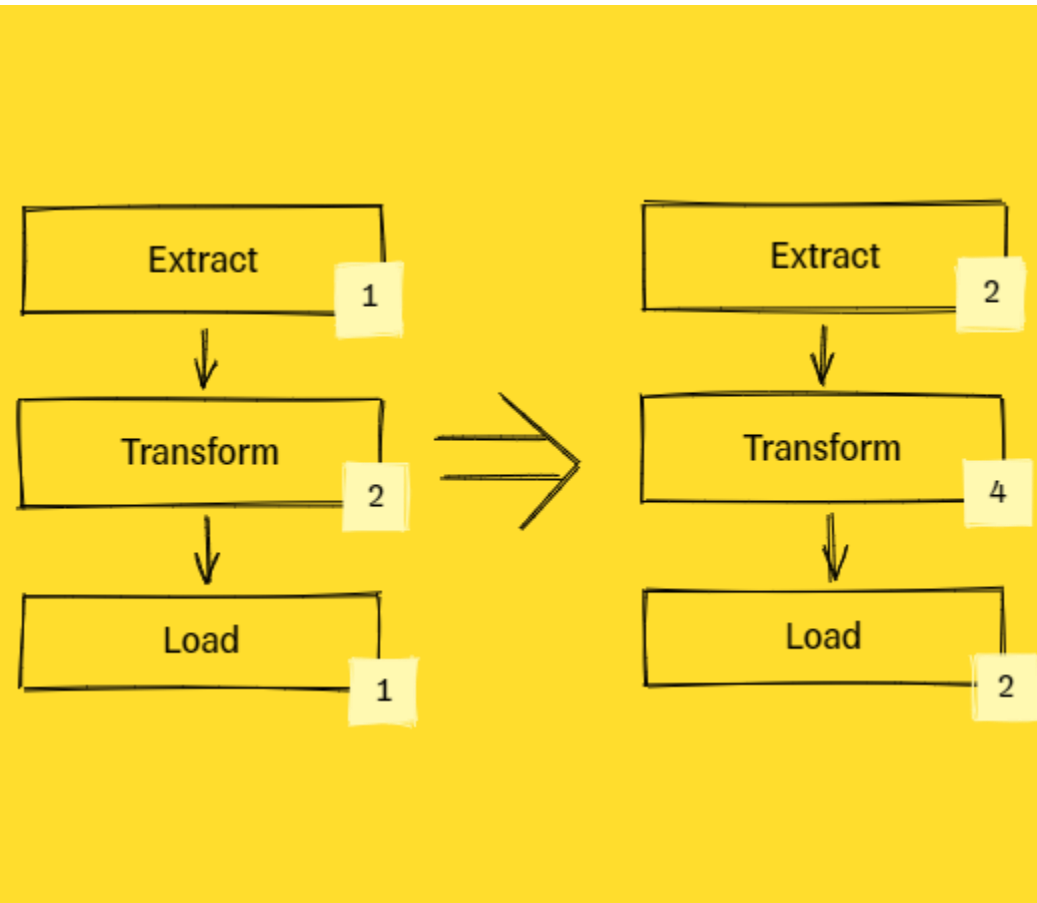
0 sec

Execution 

All nodes 



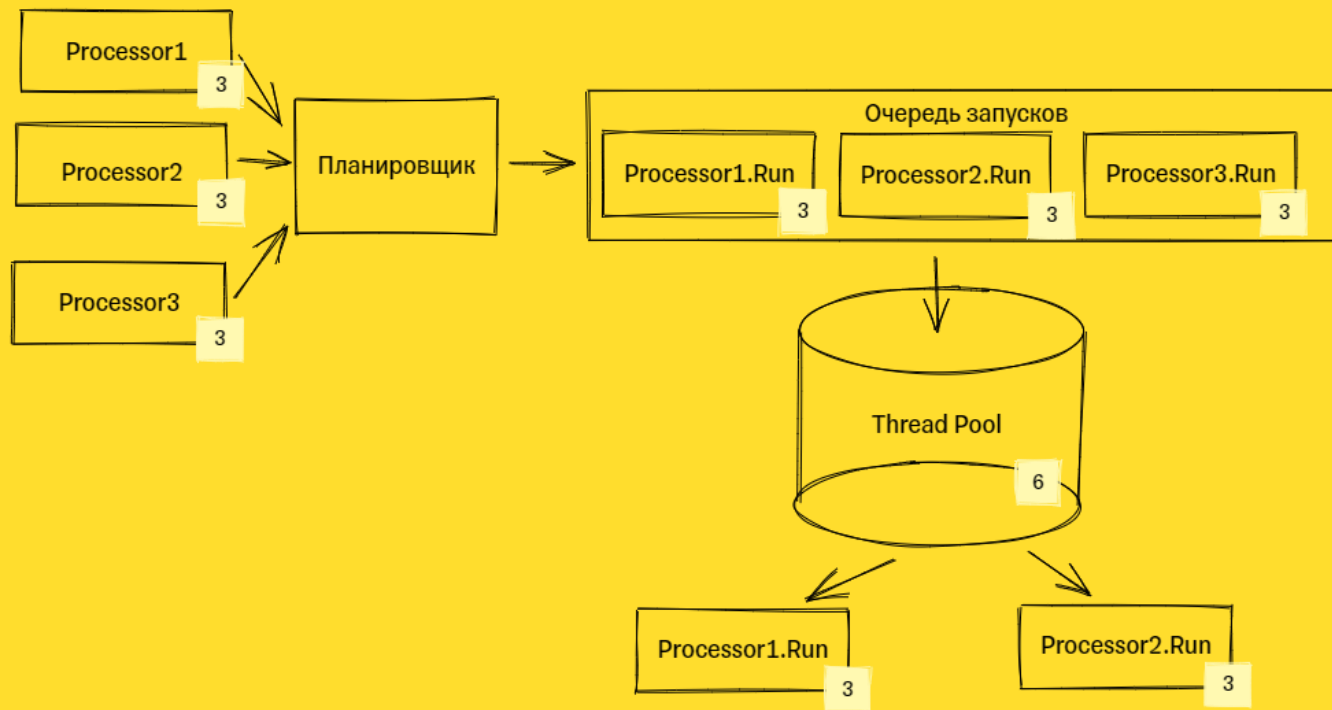
# Ускорение работы



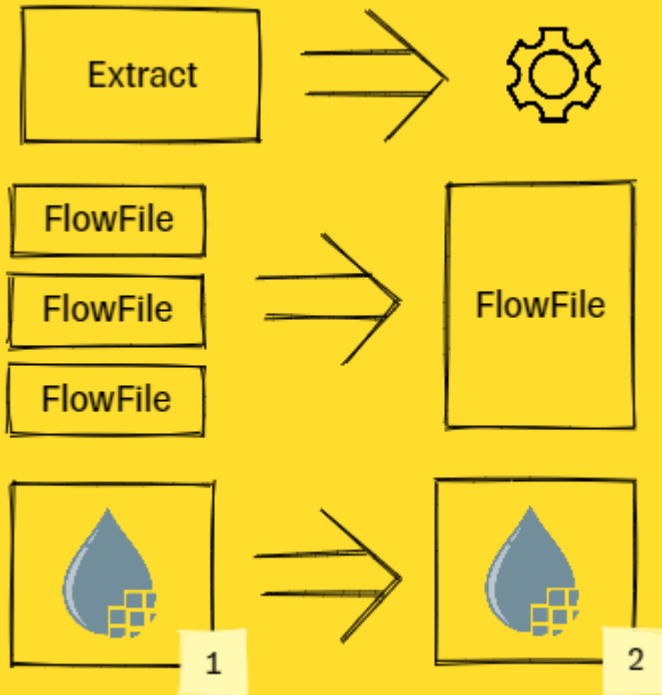
У нас больше данных

Обработка быстрее

# Планировщик



# Как можно по другому



Настройки

Микро-батчингу

Кластер



# Настройки производительности

Configure Processor | UpdateAttribute 1.16.3

Stopped

SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Scheduling Strategy ?

Timer driven



Concurrent Tasks ?

1

Run Schedule ?

0 sec

Run Duration ?

0ms 25ms 50ms 100ms 250ms 500ms 1s 2s



Lower latency

Higher throughput

# Настройки производительности

Configure Processor | ExecuteSQLRecord 1.16.3

Invalid

SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Required field



Property		Value
SQL select query	?	No value set
SQL Post-Query	?	No value set
Max Wait Time	?	0 seconds
Record Writer	?	No value set
Normalize Table/Column Names	?	false
Use Avro Logical Types	?	false
Default Decimal Precision	?	10
Default Decimal Scale	?	0
<u>Max Rows Per Flow File</u>	?	10000
<u>Output Batch Size</u>	?	0
Fetch Size	?	0
Set Auto Commit	?	true

# Балансировка

## Configure Connection

DETAILS

SETTINGS

Name

Id

48bde0de-0188-1000-ffff-ffffaae14198

FlowFile Expiration ?

0 sec

Back Pressure  
Object Threshold ?

10000

Size Threshold ?

1 GB

Load Balance Strategy ?

Round robin

Load Balance Compression ?

Do not compress

Available Prioritizers ?

FirstInFirstOutPrioritizer

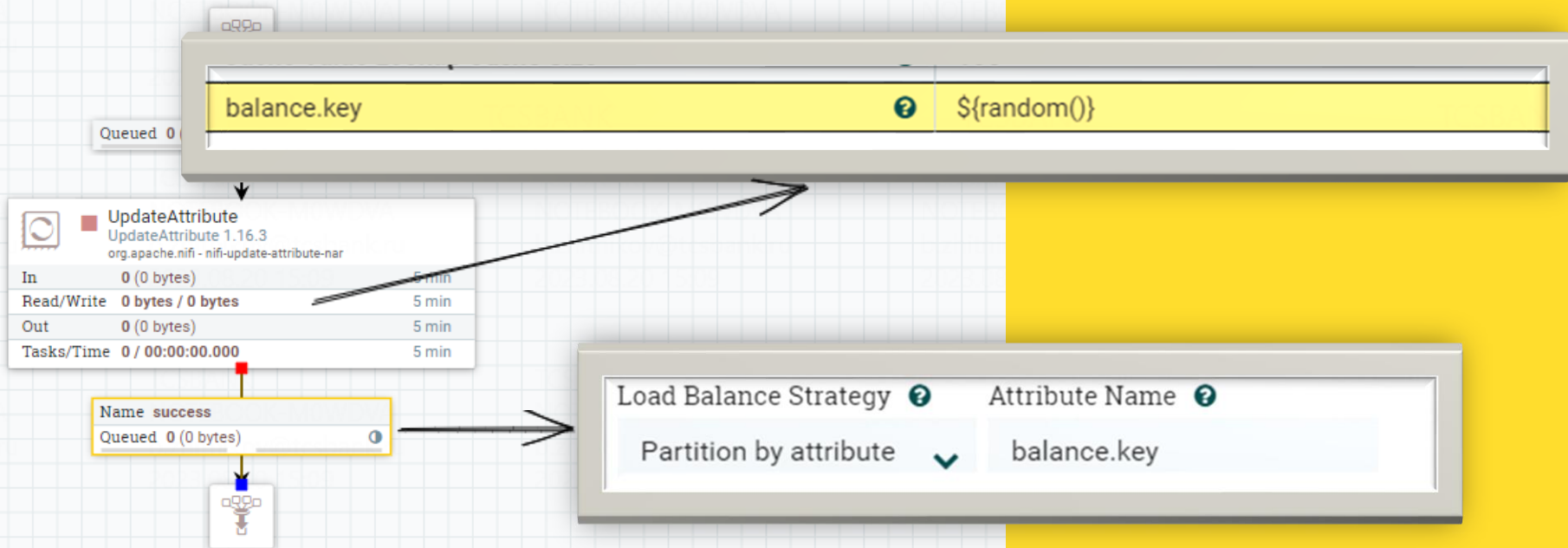
OldestFlowFileFirstPrioritizer

NewestFlowFileFirstPrioritizer

PriorityAttributePrioritizer

Selected Prioritizers ?

# Балансировка



# Кейс 3

Два потока данных с документами

ID документа присутствует в двух потоках

Сохраняем только документы которые прошли по двум потокам



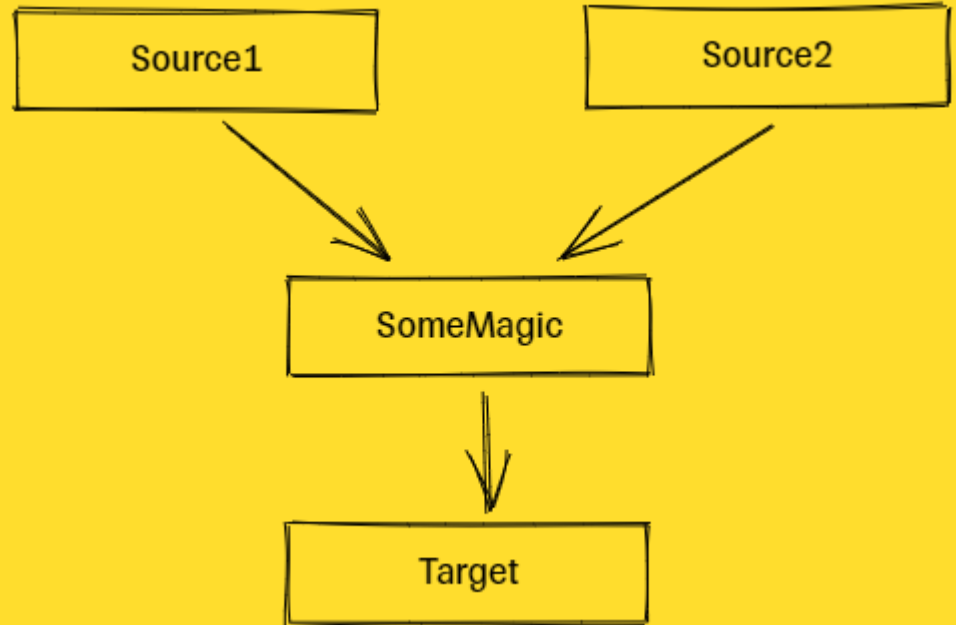
# Join данных

Несколько потоков

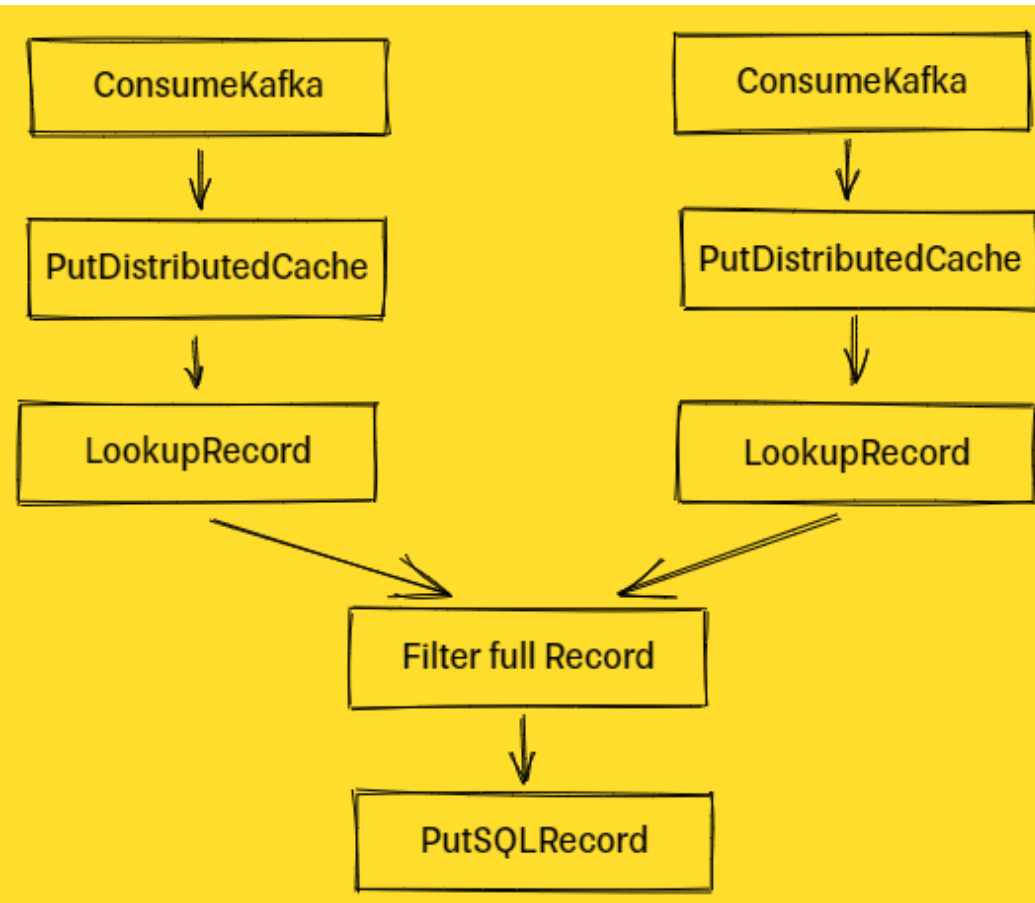
Несколько источников

Связь по ключу

Данные приходят в одно время



# Join данных



Распределенный Cache

LookupRecord

Объединение

# Какие проблемы

Хранение событий

Сложная логика

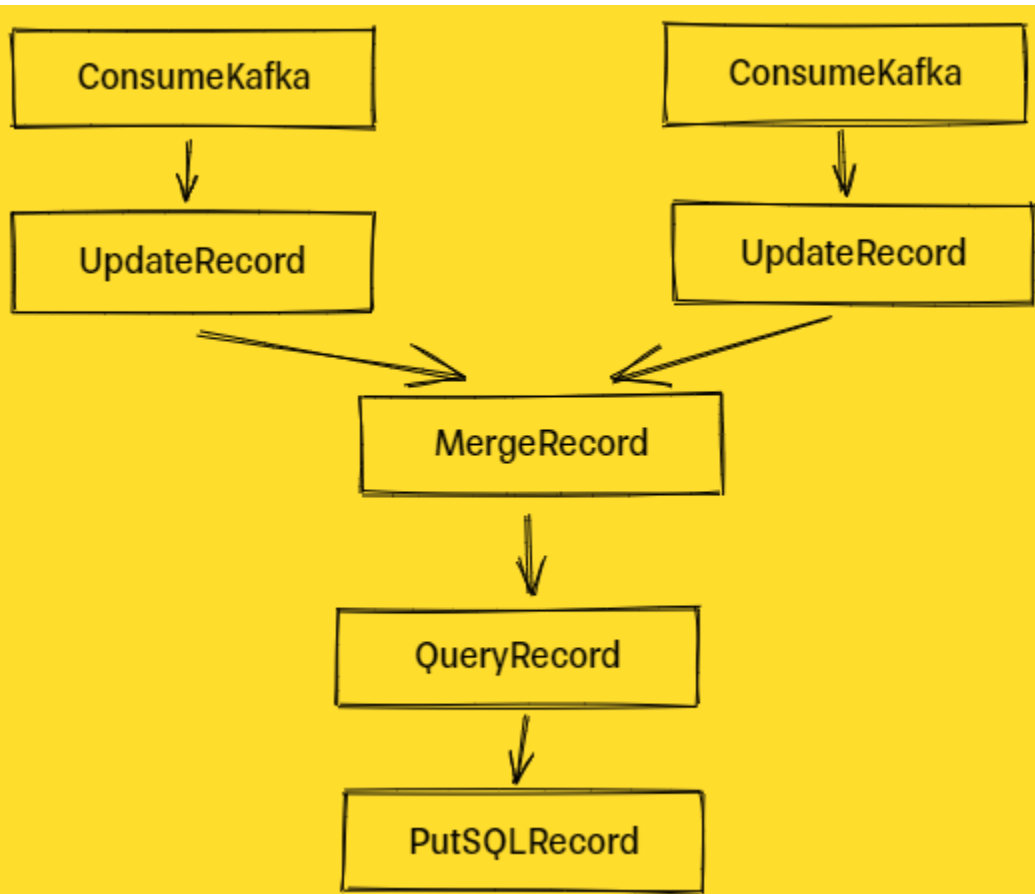
Много  
взаимодействия

сетевого





# Join данных



Добавили поле с Источником

Собрали в один файл

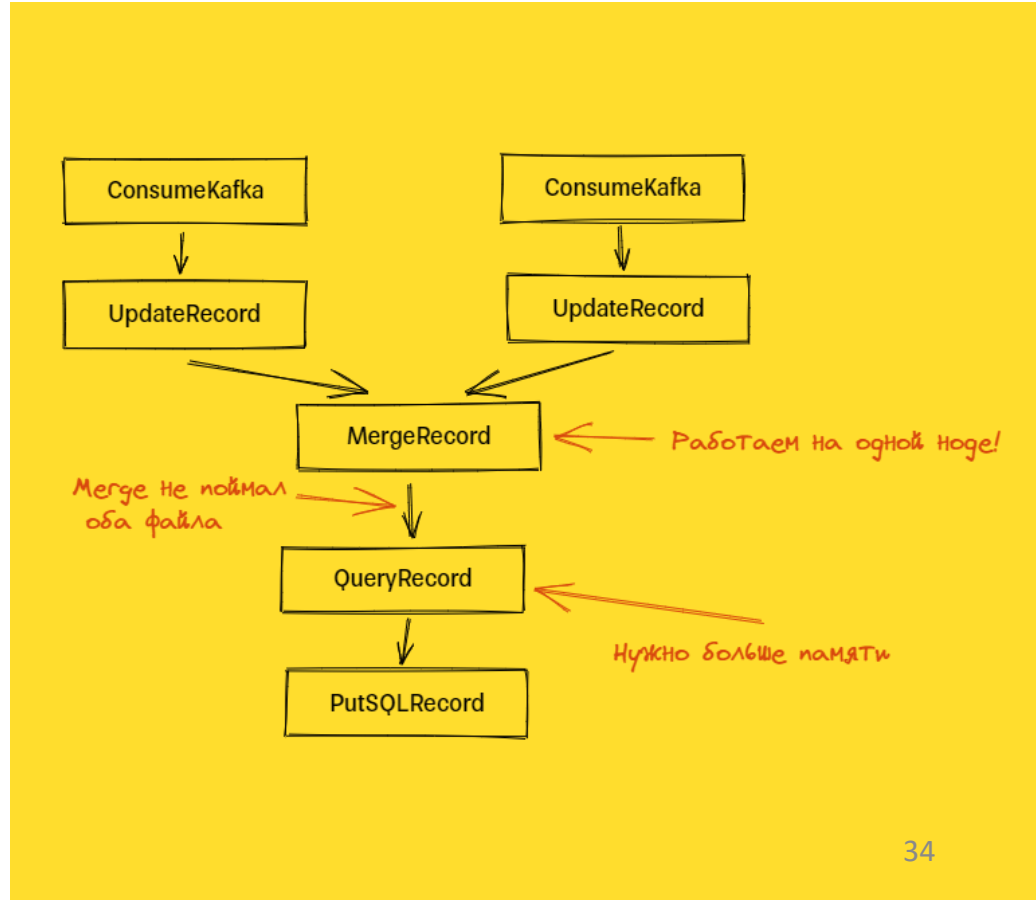
Отфильтровали в QueryRecord

# Какие проблемы

Гарантии Merge

Память в QueryRecord

Отсутствие распределения



# Что делать

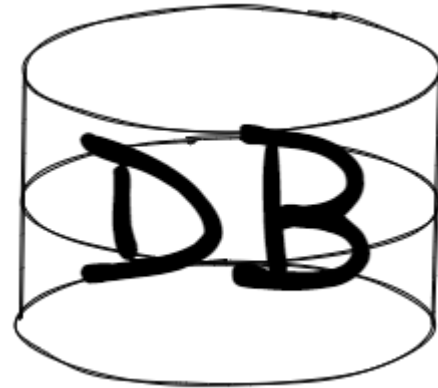
Выполнять Join в Target

Создать промежуточную систему

Использовать  
инструмент

другой

ETL



# Кейс 4

Есть движок расчетов работающий как Batch

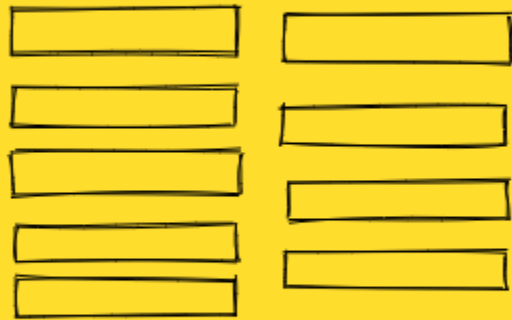
В движке формируют набор данных для выгрузки

Нужно загрузить в хранилище результаты расчета целиком

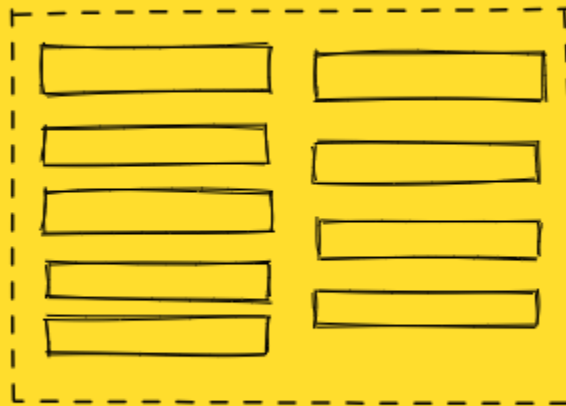


# Консистентный Batch

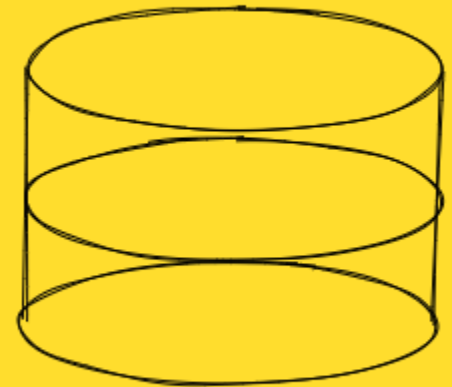
Все или ничего



Поток

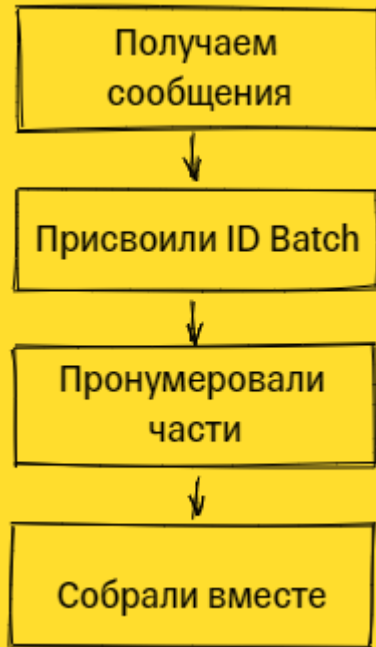


Батч



Транзакция

# Консистентный батч

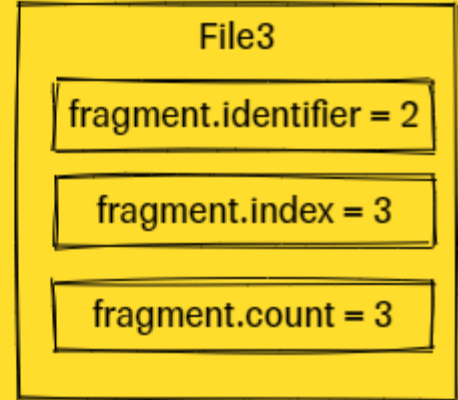
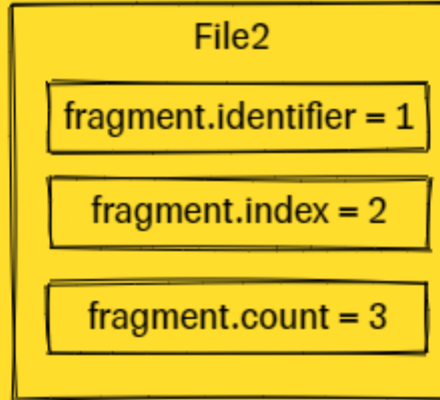
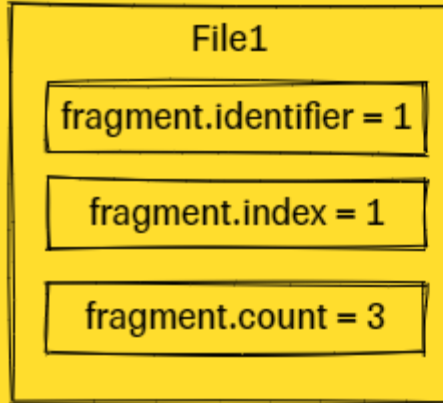


Определение границ

Нумерация частей

Merge процессоры

# Merge Defragment



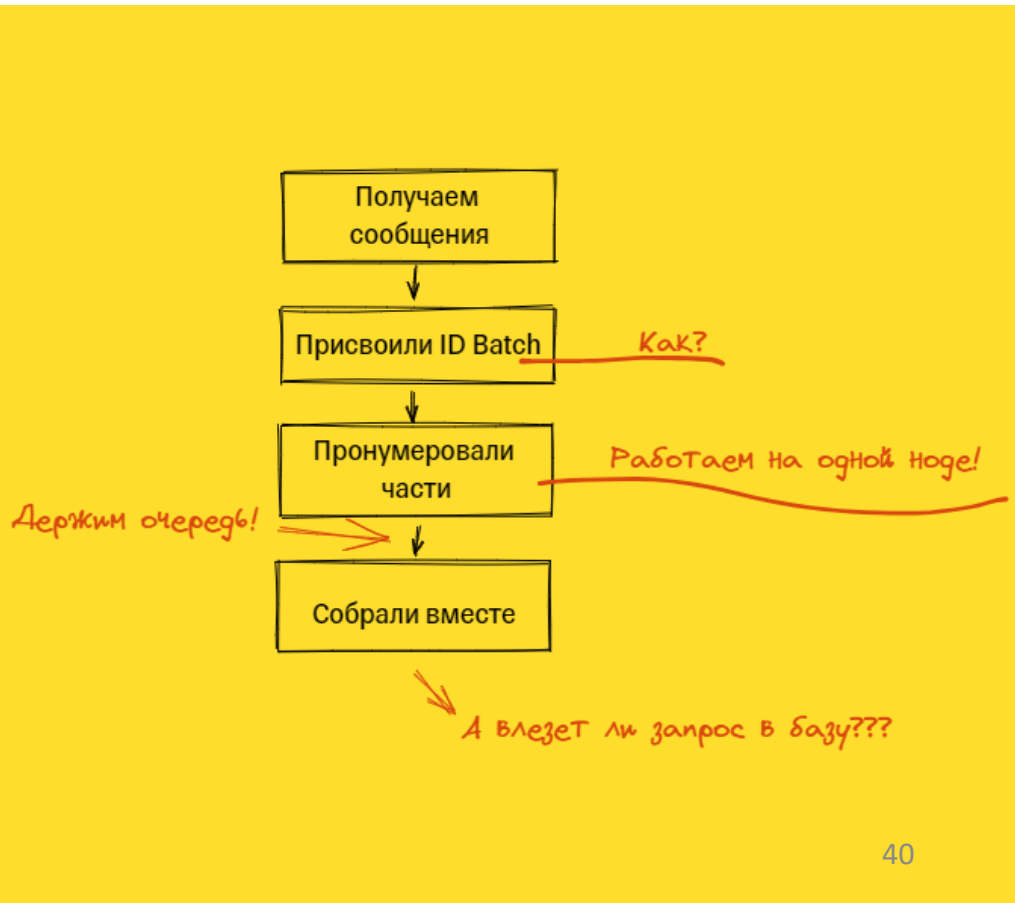
# Какие проблемы

Идентификация Batch

Распределение работы

Накопление очередей

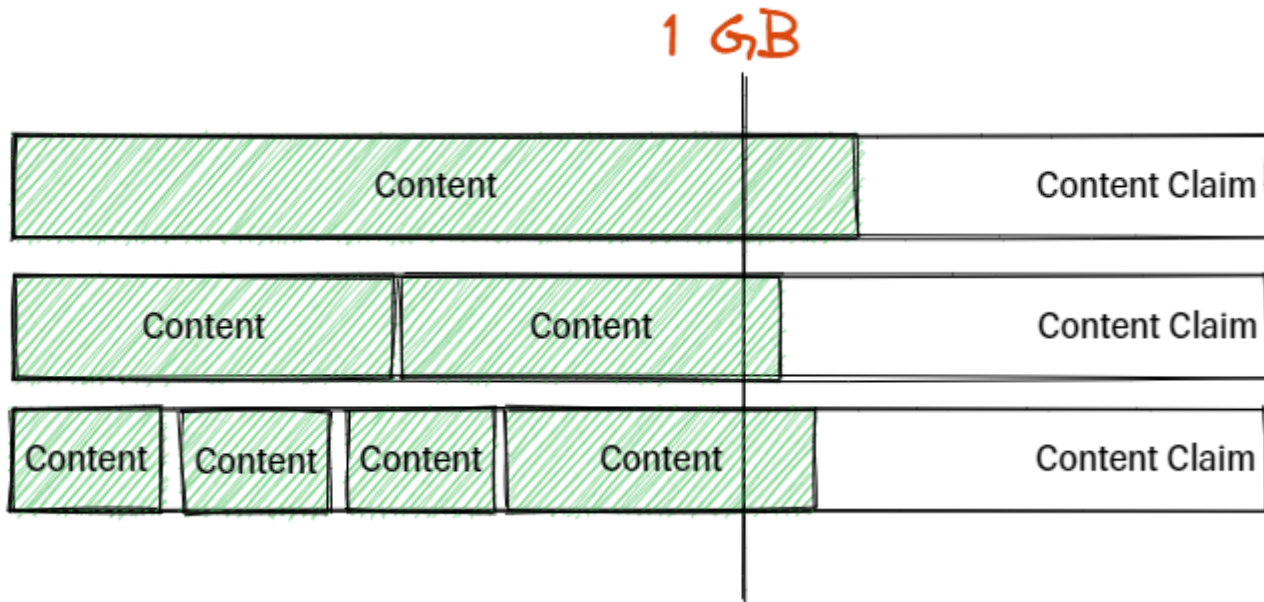
Большие файлы





# Структура хранения

Content Repository



# Что делать

A thick yellow vertical line on the left side of the slide, which branches into two horizontal lines pointing to the right, forming a bracket shape.

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

