Разработка **Real Time Decision Making** решения на базе **Hortonworks Data Platform** & **Data Flow**

Алексей Кузнецов

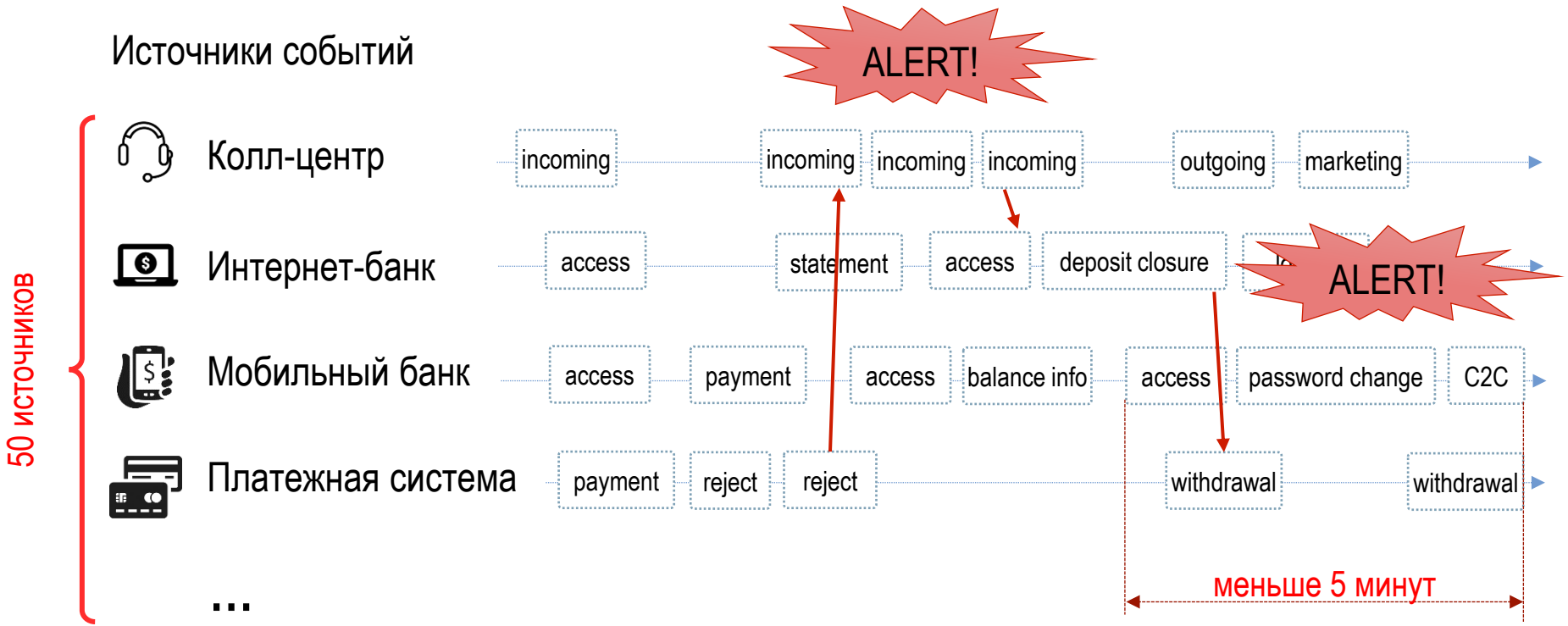
Data Engineer, Software Architect

Антон Ярмолюк

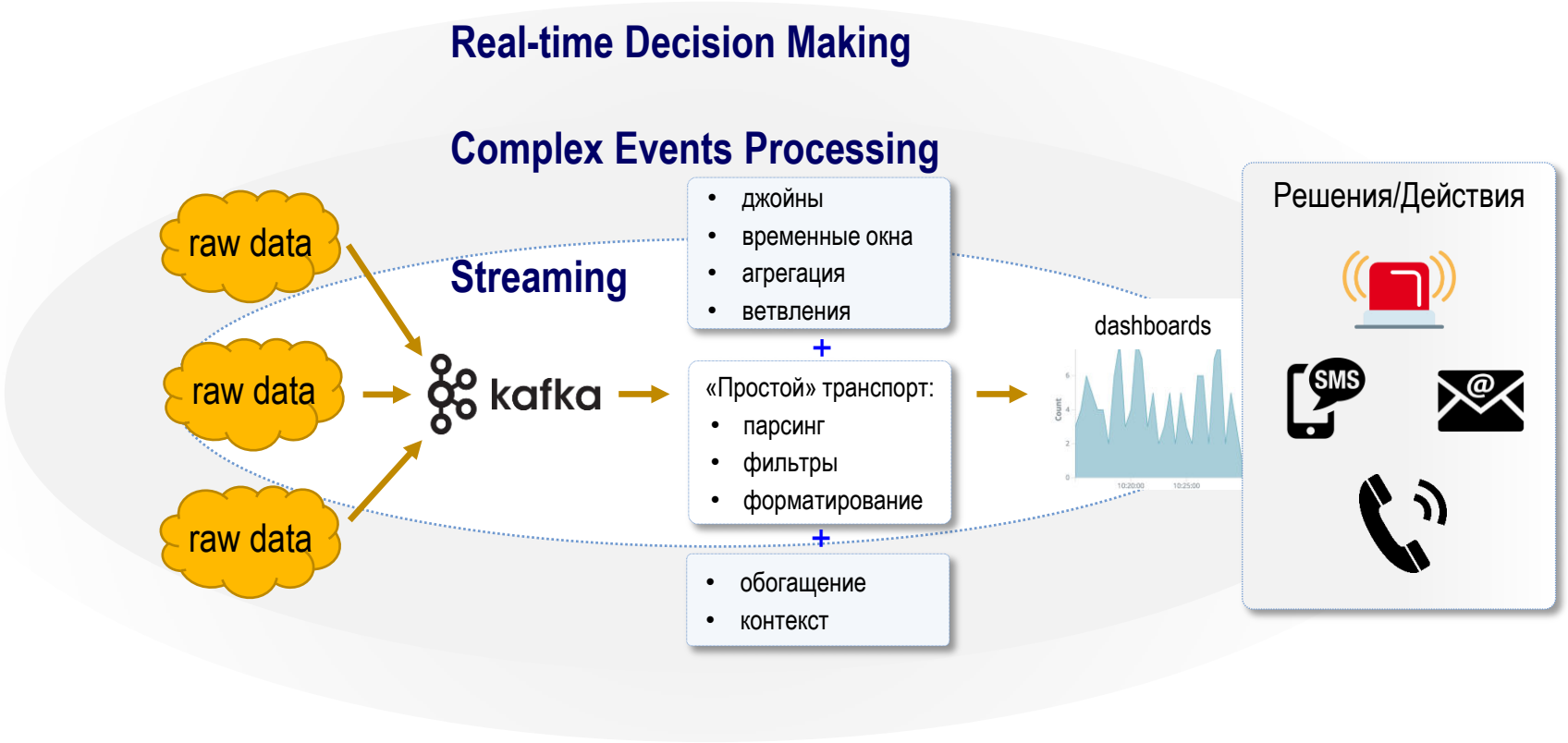
Data Engineer, Chapter Lead

Райффайзен БАНК

Задача

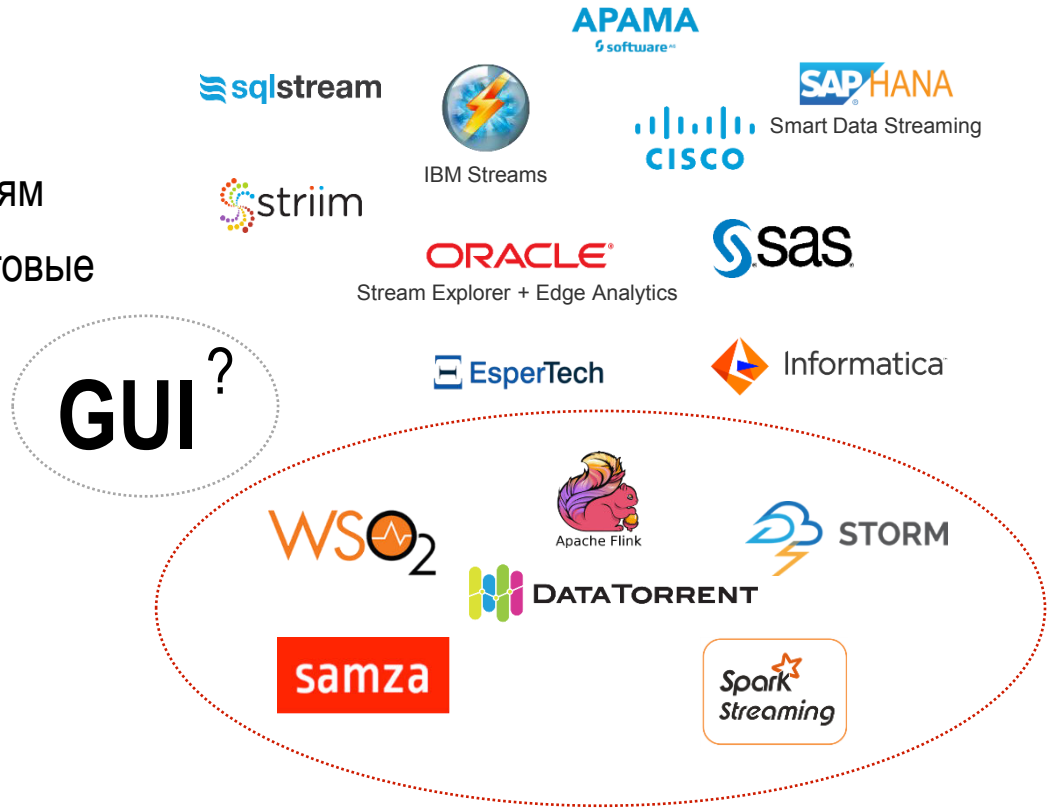

Источники событий
50 источников
Колл-центр
Интернет-банк
Мобильный банк
Платежная система
...
ALERT!
incoming
incoming
incoming
incoming
outgoing
marketing
access
statement
access
deposit closure
ALERT!
access
payment
access
balance info
access
password change
C2C
payment
reject
reject
withdrawal
withdrawal
меньше 5 минут

Аналитика Real-time

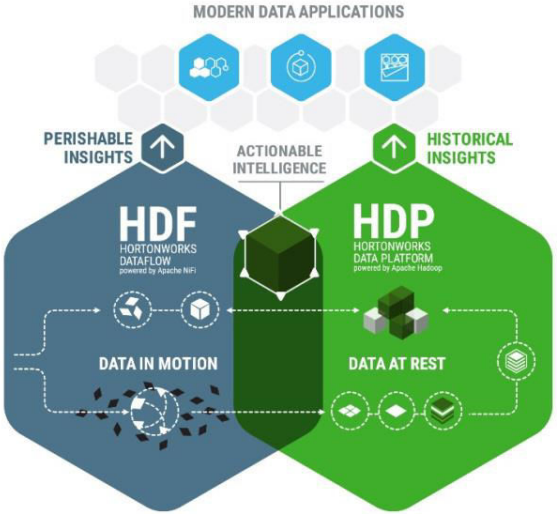


Наш кейс

Дать возможность бизнес-пользователям **самостоятельно** проводить маркетинговые кампании на платформе RTDM



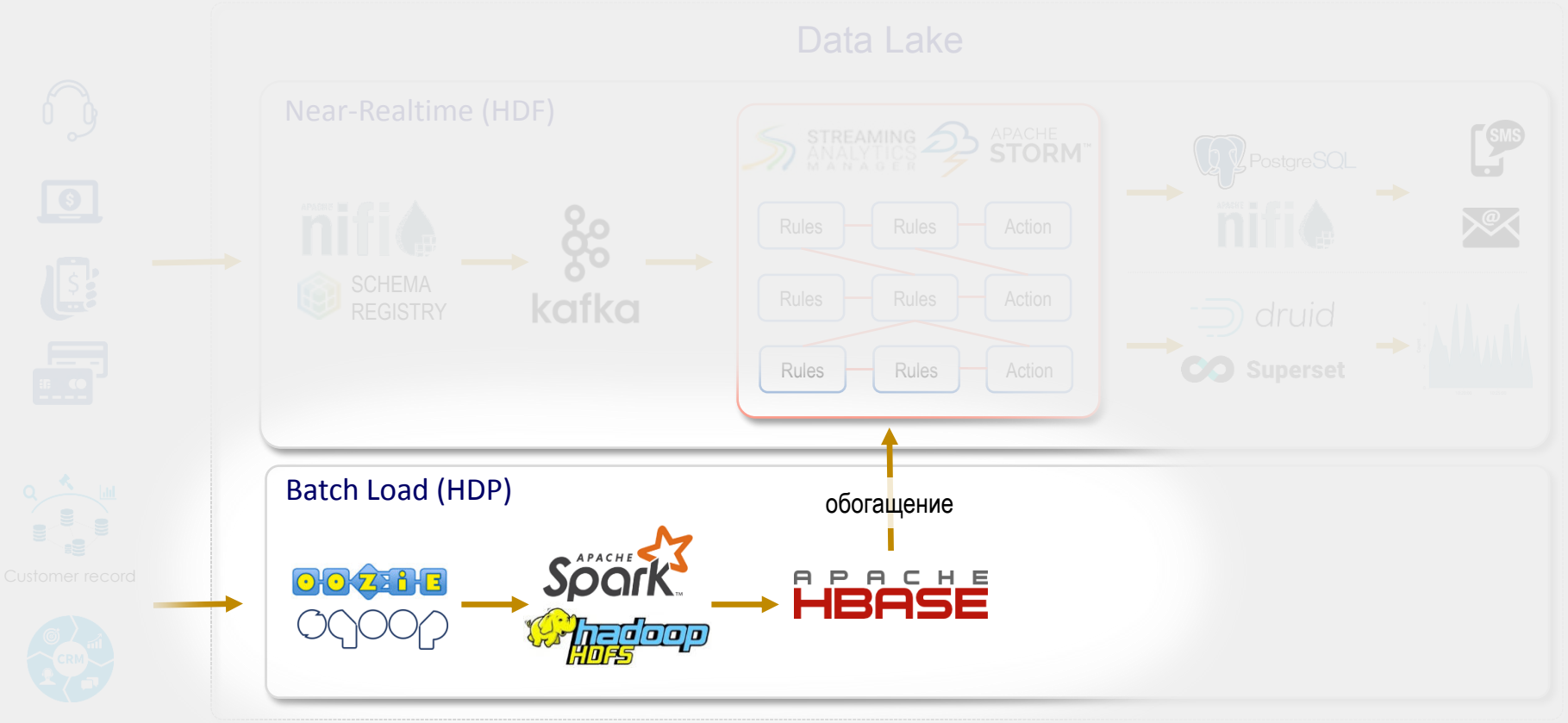
Hortonworks Data Flow 3.0



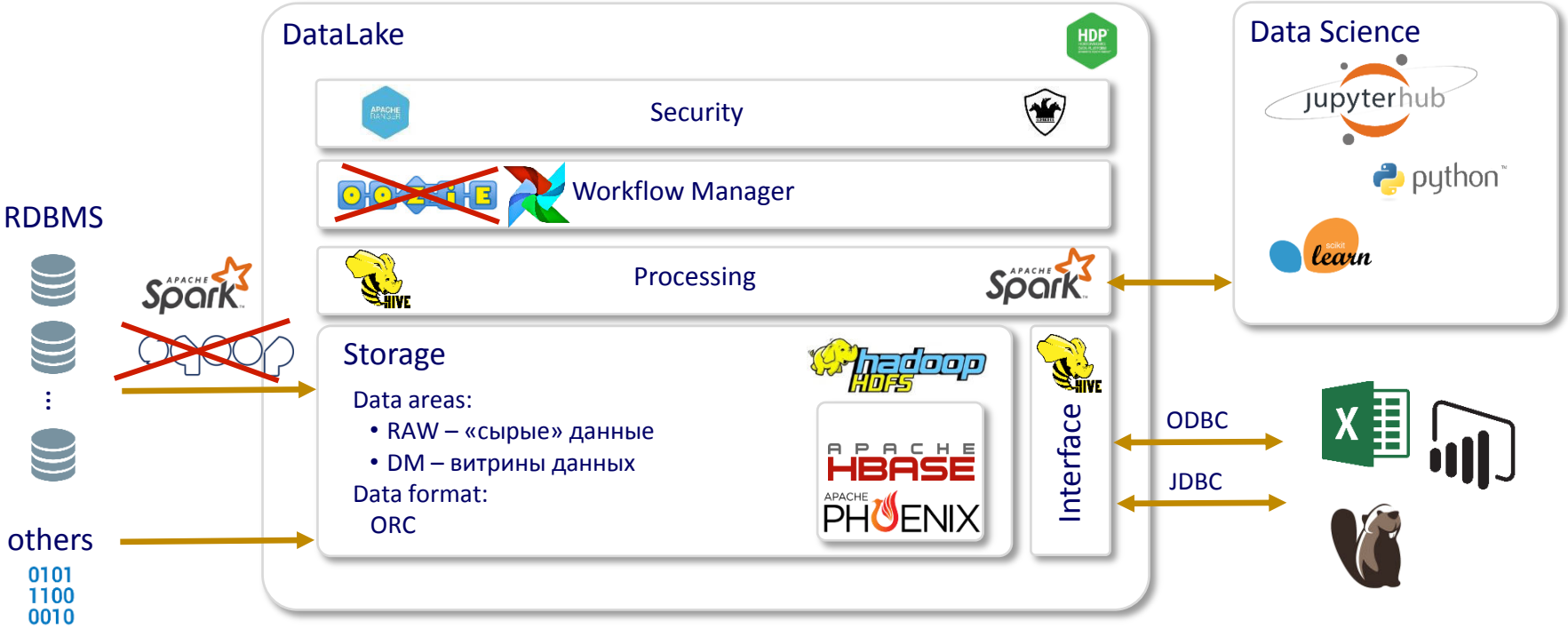
GUI

Pics source: https://hortonworks.com

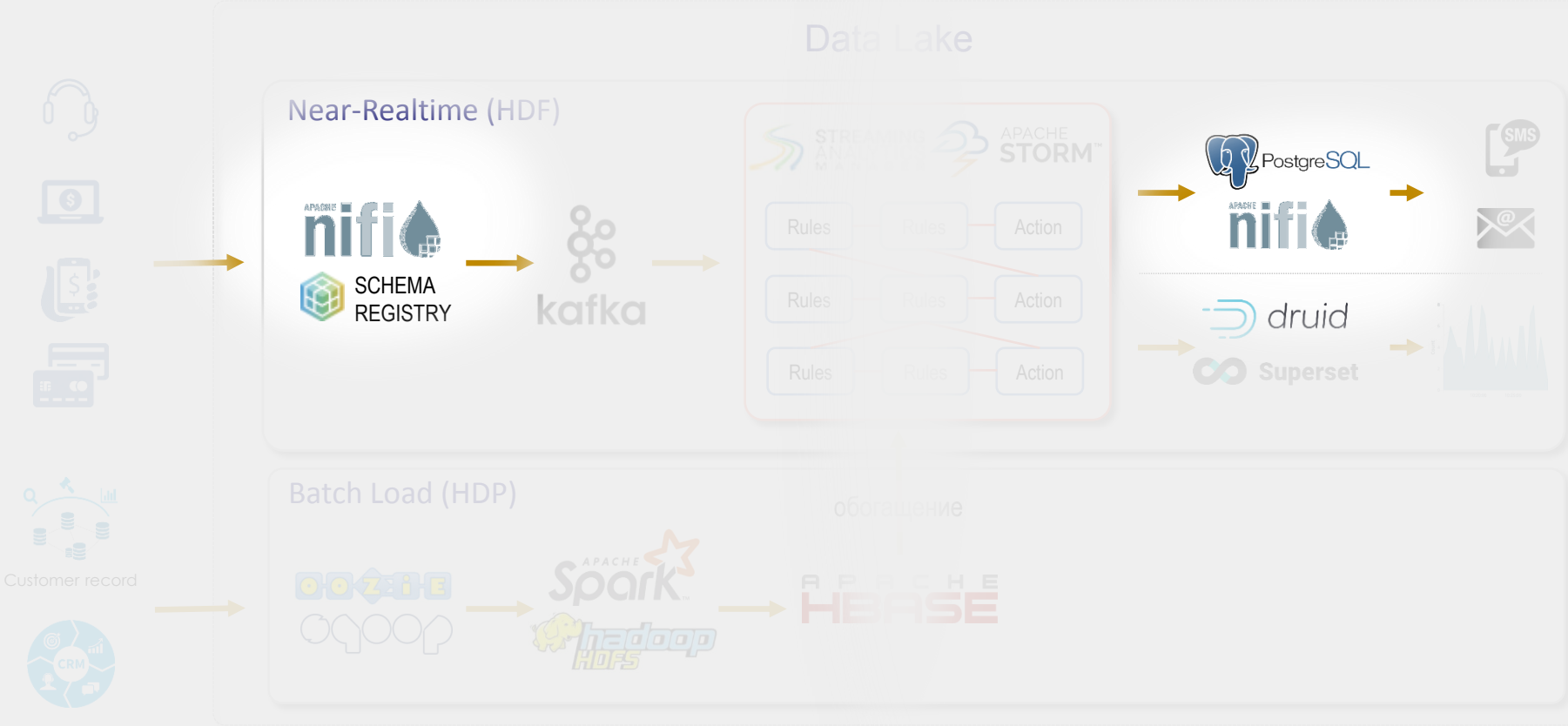
Архитектура RTDM-решения



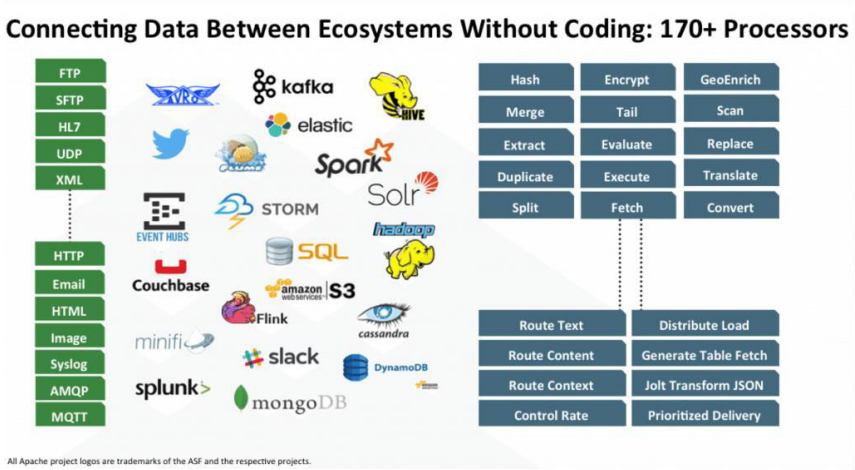
HDP - Batch Load



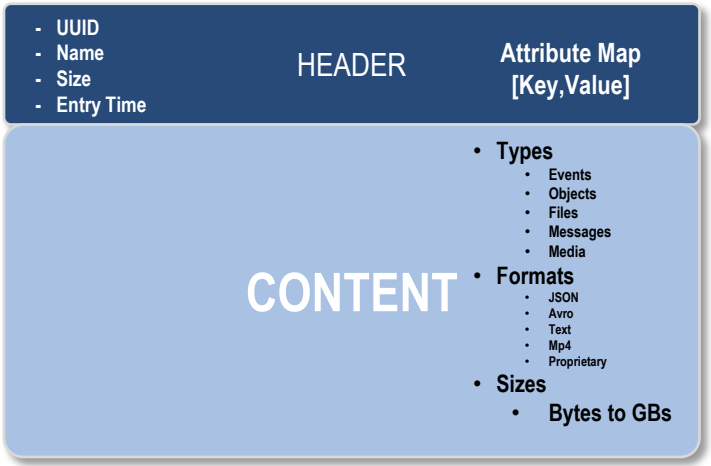
Архитектура RTDM-решения



HDF – Apache NiFi

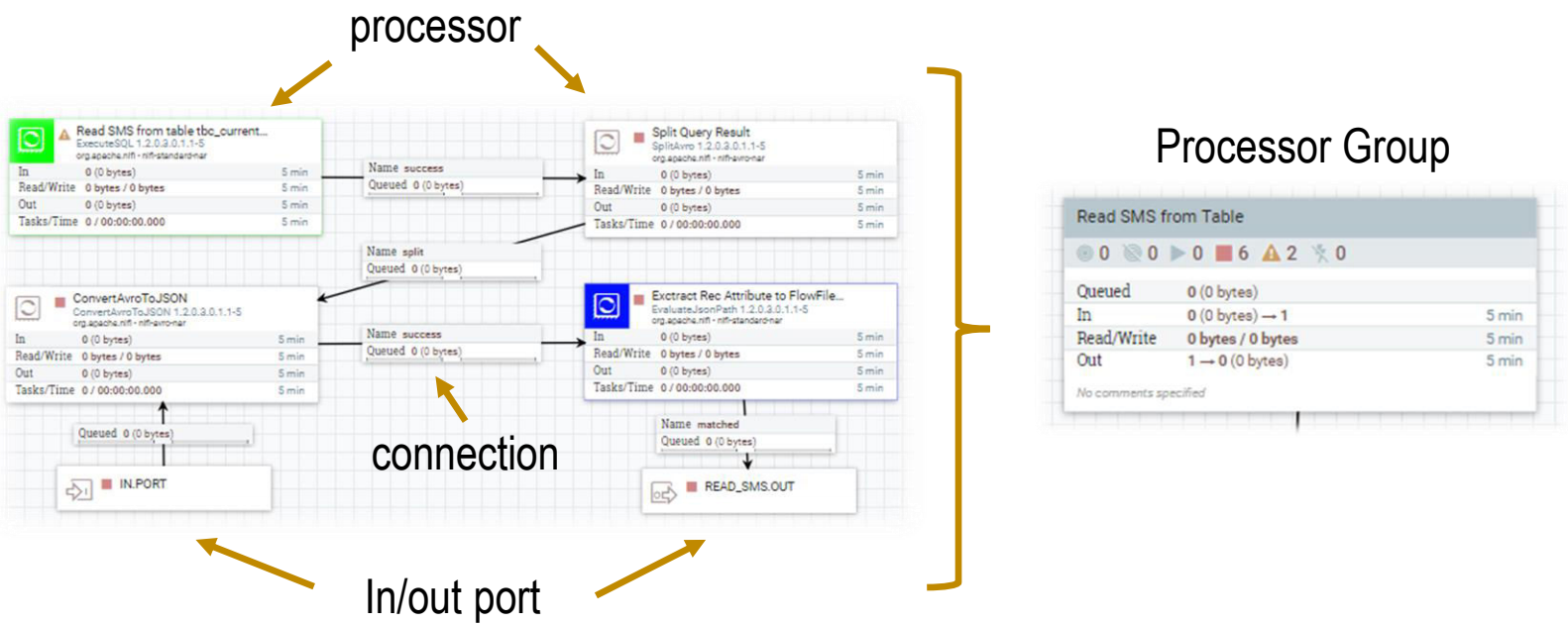


- Разработка АНБ, передан в Open Source в 2014 году
- Реализует концепцию Flow Based Programming (FBP)
- Визуальный интерфейс разработки

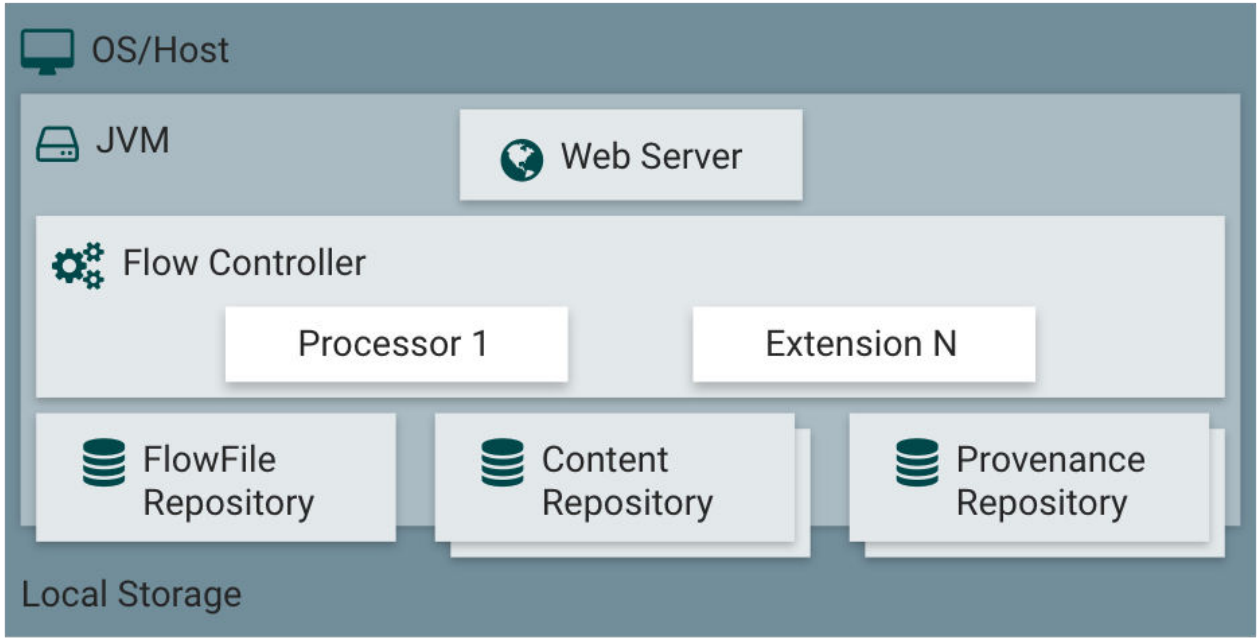


Основная единица передаваемых данных - FlowFile

HDF – Apache NiFi

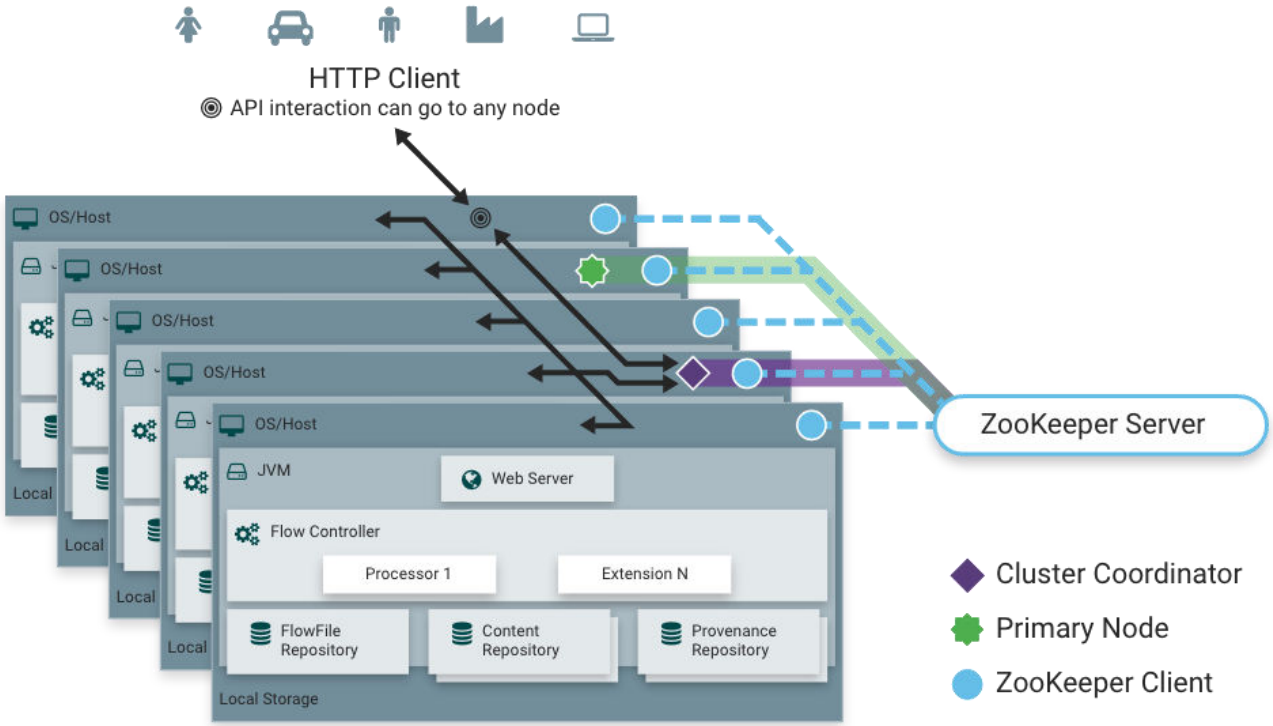


HDF – Apache NiFi



https://nifi.apache.org/docs/nifi-docs/html/nifi-in-depth.html

HDF – Apache NiFi



https://nifi.apache.org/docs/nifi-docs/html/nifi-in-depth.html

HDF – Apache NiFi

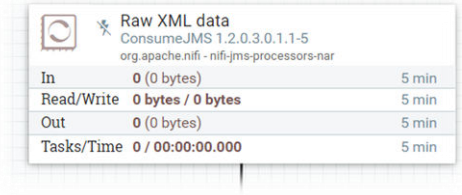
MQ Topic

Очередь руками (Production)

Durable подписка

Non-durable подписка

JMS

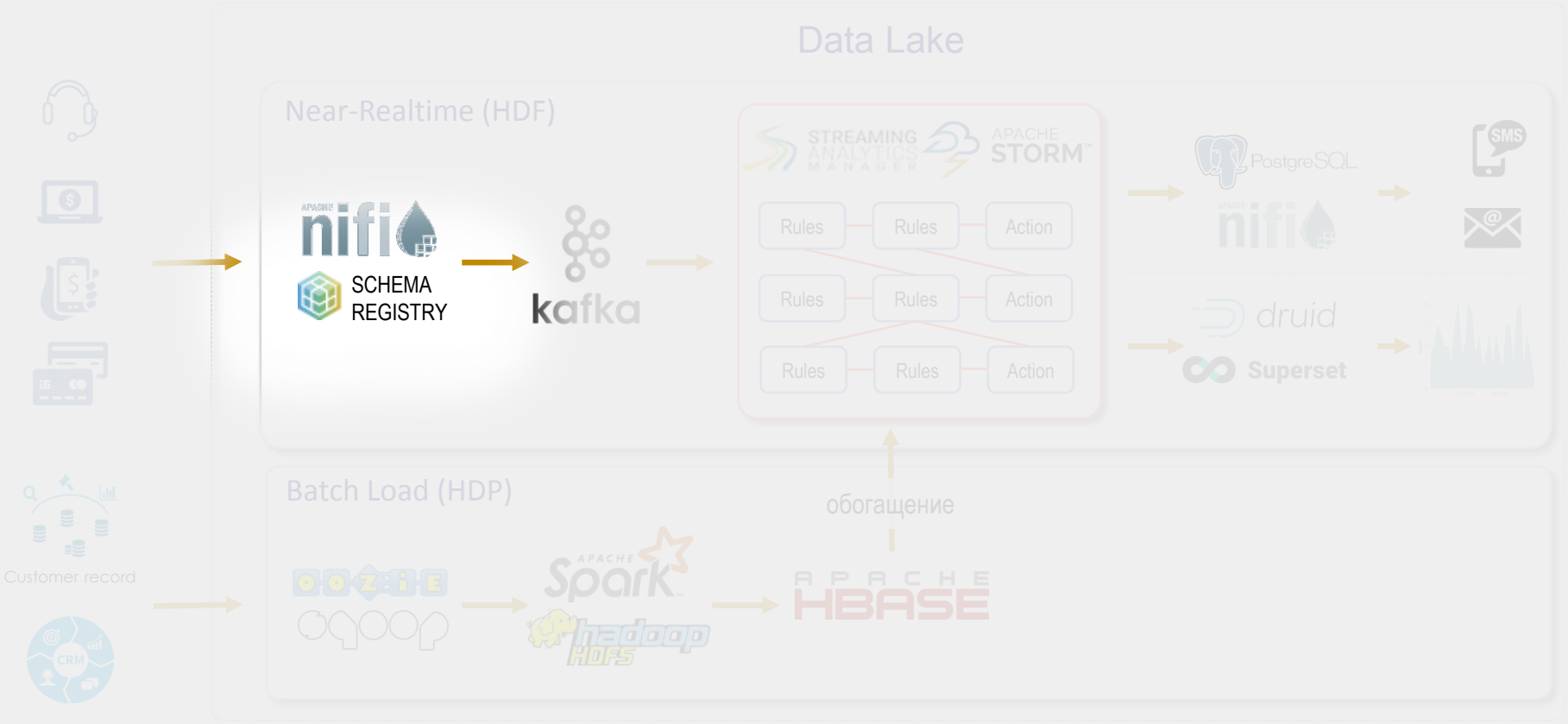


Автоматически создается очередь под каждого подписчика
Имя подписчика генерируется автоматически на основание hostname

queue_a
queue_b

Включили Only Primary Node

Архитектура RTDM-решения



Schema Registry

Райффайзен БАНК

kafka

Topic

message #1: 011110001100 → message #2: 011110001100 →

SCHEMA REGISTRY

Schema

```
{
 "type": "record",
 "namespace": "ru.raiffeisen.bigdata.sam",
 "name": "SOURCE.RUB_TRANSFER",
 "fields": [
  {
   "name": "clientID",
   "type": "string"
  },
  {
   "name": "amount",
   "type": "double"
  }
 ]
}
```

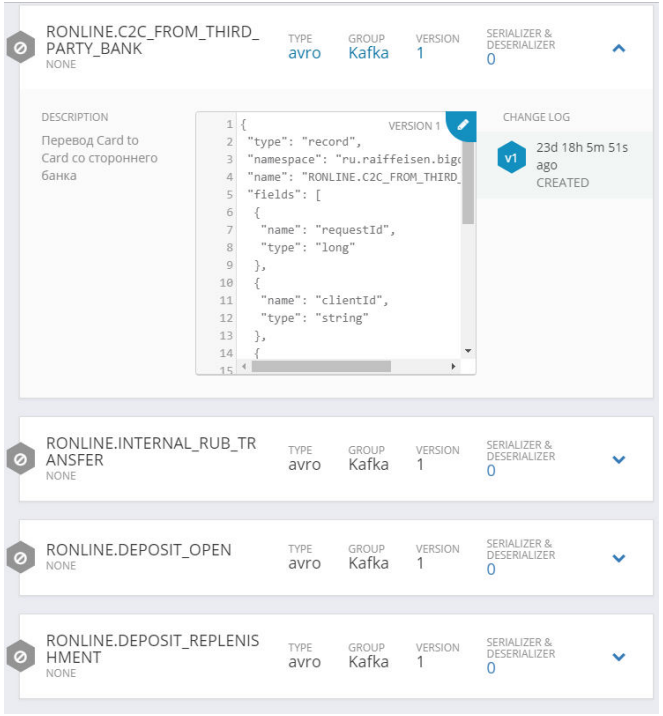
STREAMING ANALYTICS MANAGER

APACHE STORM™

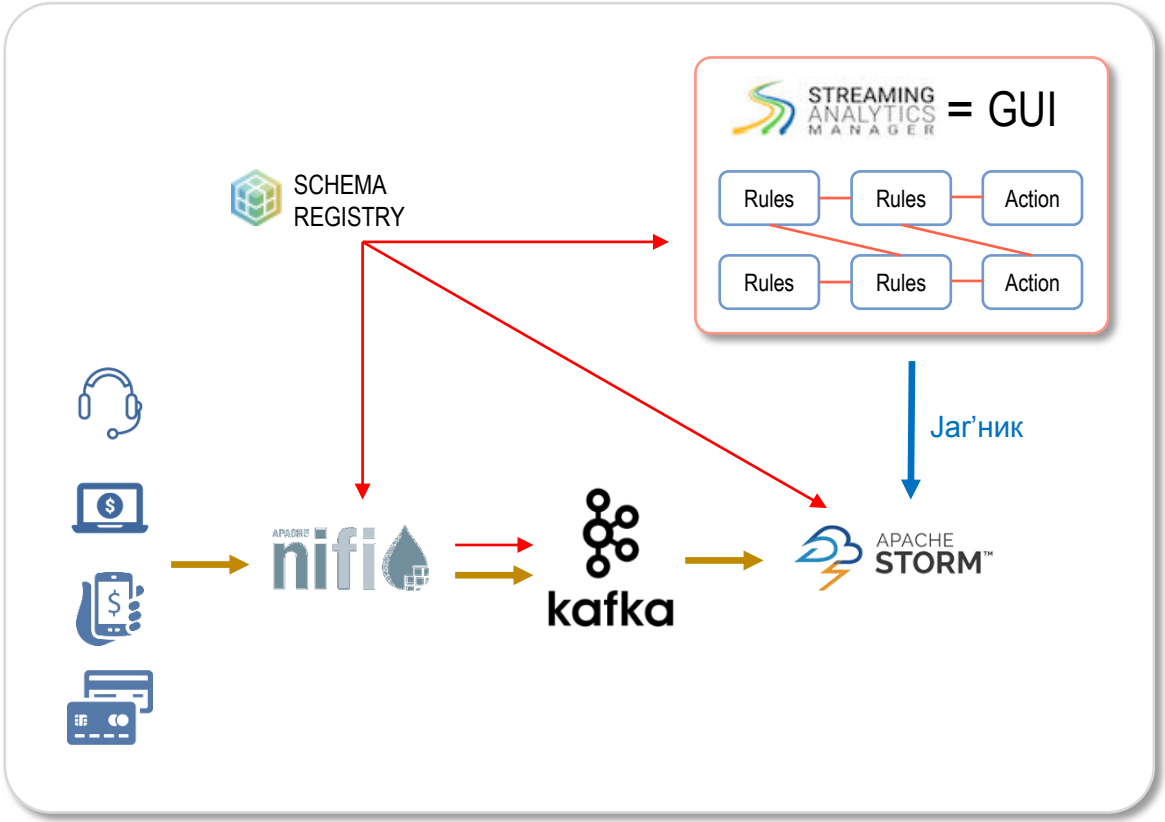
Stream

clientID	amount	clientID	amount
client #1	500 p	client #2	1000 p

SCHEMA REGISTRY GUI

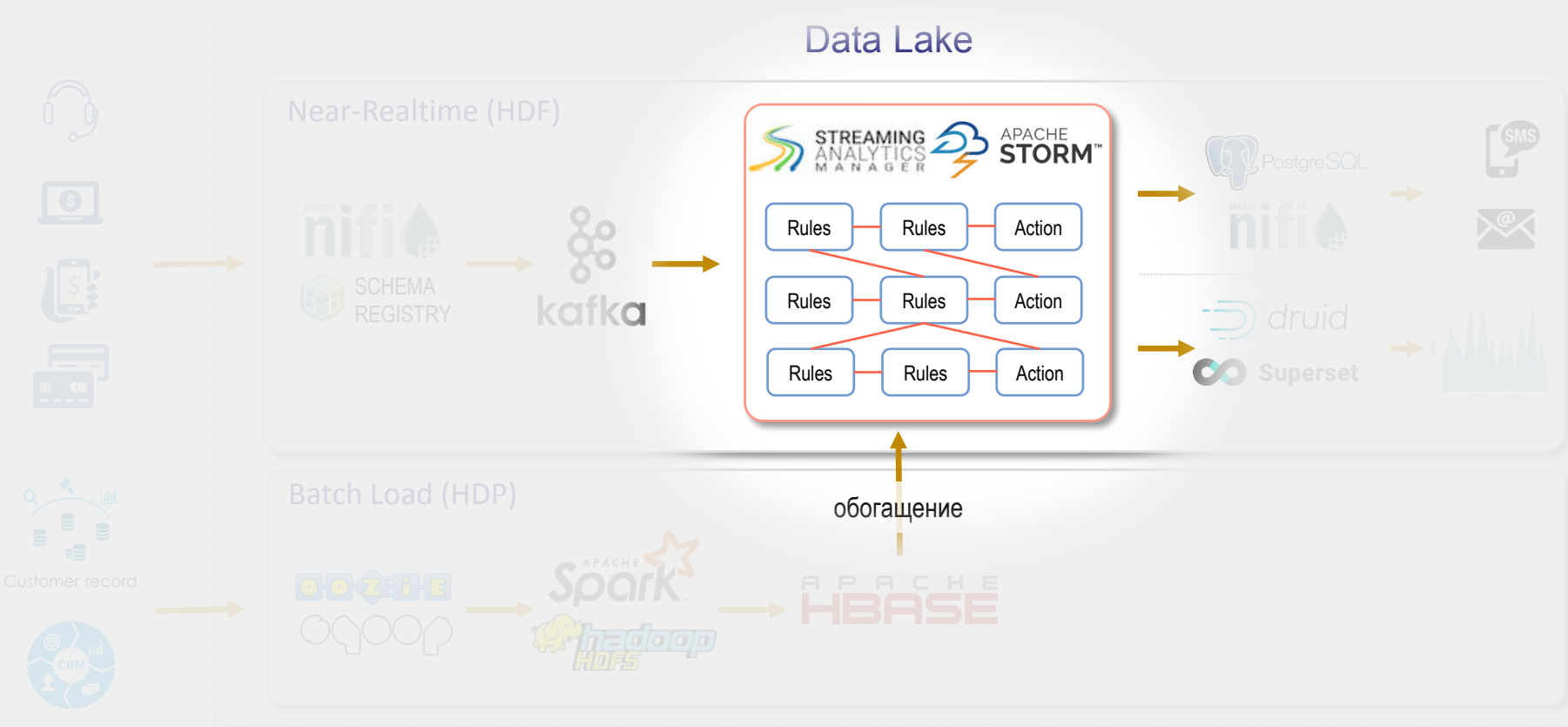


Schema Registry

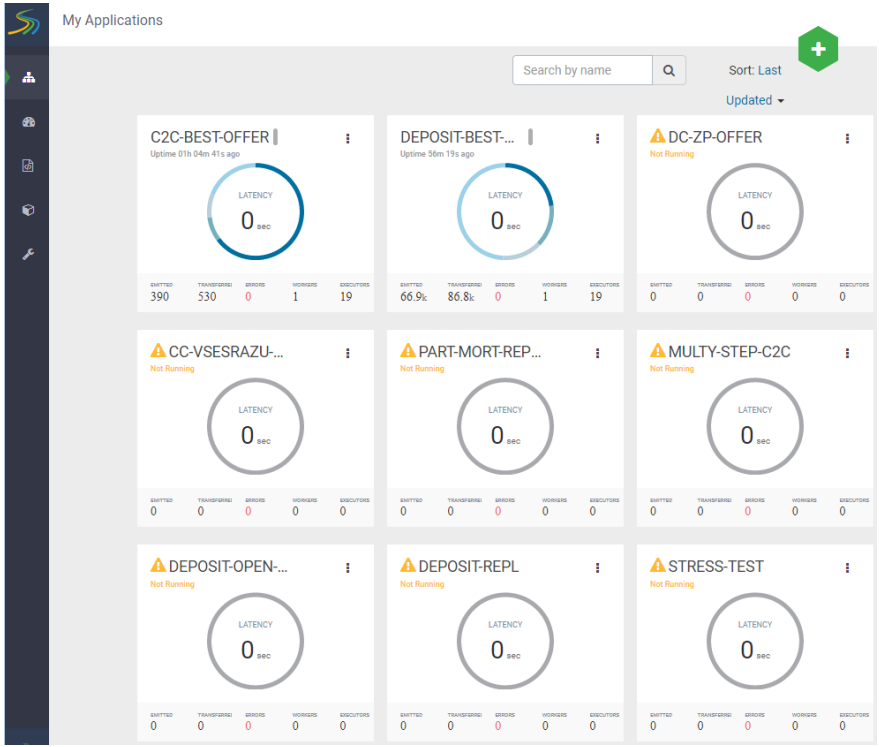


- **Schema Registry**
  - создаем и описываем схему “стрима”
- **Nifi**
  - читаем схему “стрима” (метаданные)
  - сериализуем и отправляем в Kafka
- **Kafka**
  - не знает о том, что в ней лежит
- **SAM**
  - читаем схему “стрима” (метаданные)
- **Storm**
  - де-сериализуем “стрим” в “данные”

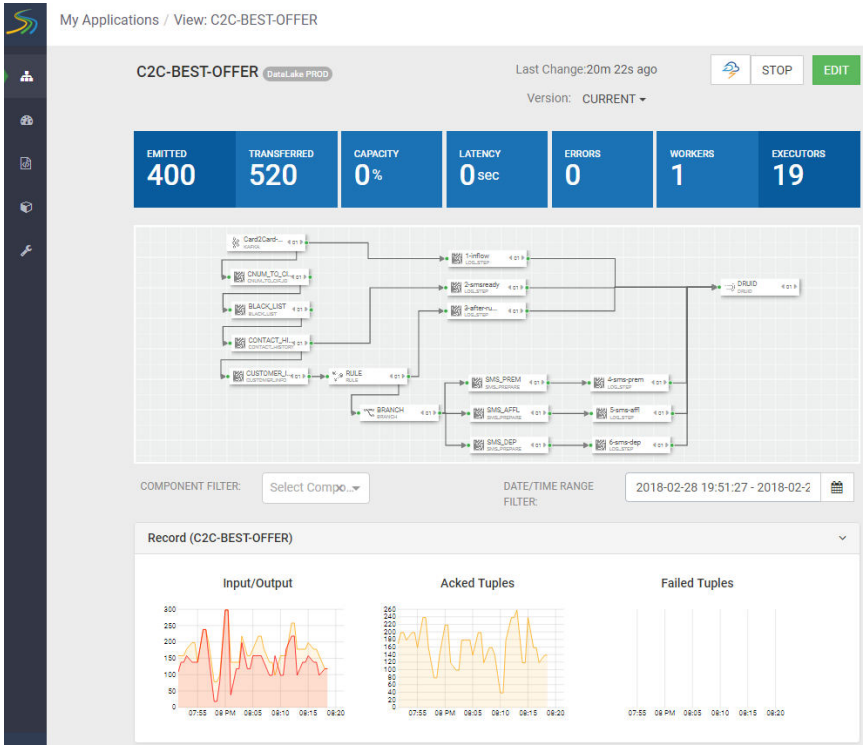
Архитектура RTDM-решения



Streaming Analytics Manager (SAM)

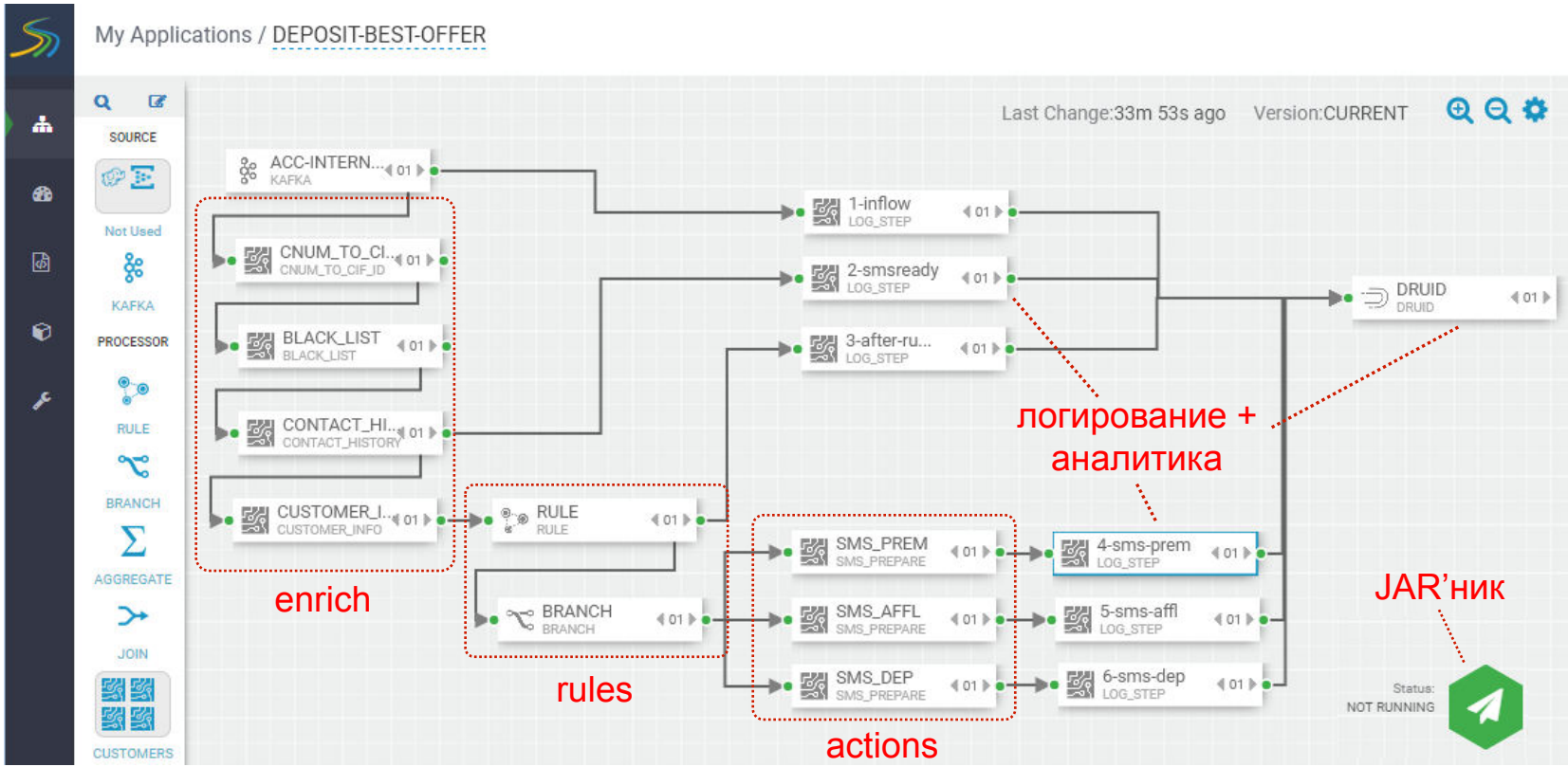


Список кампаний

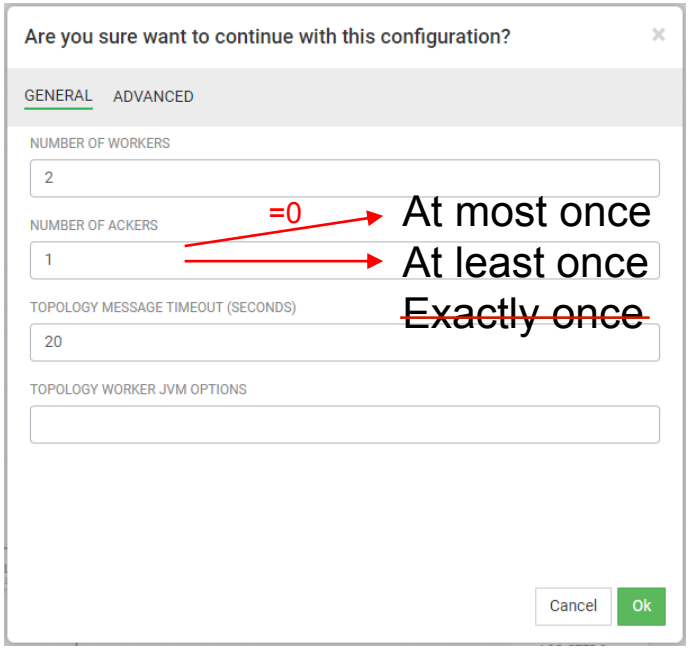


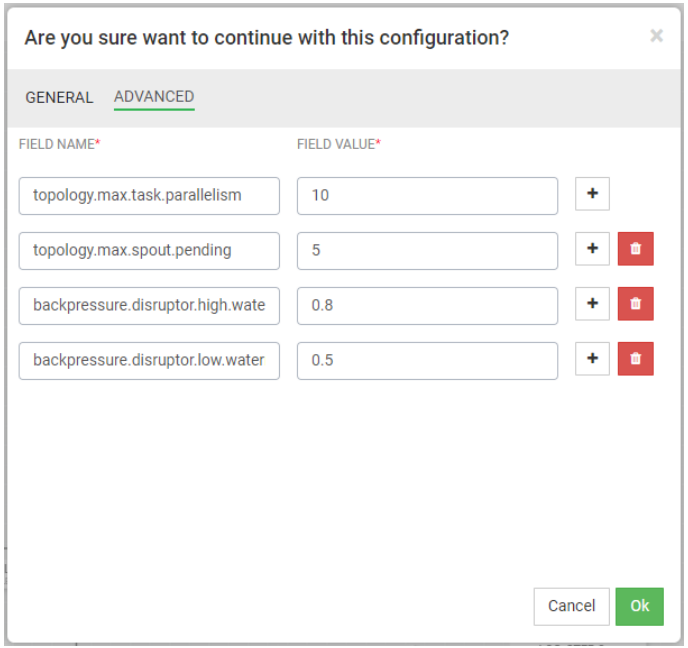
Сводная информация по кампании

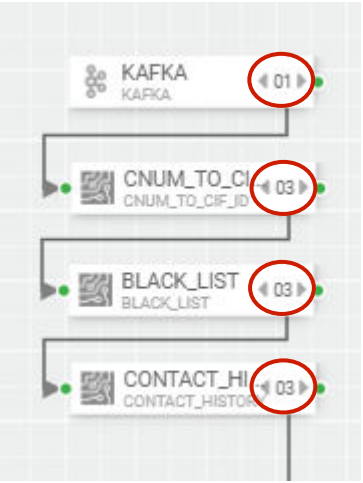
Streaming Analytics Manager (SAM)



Storm (via SAM)





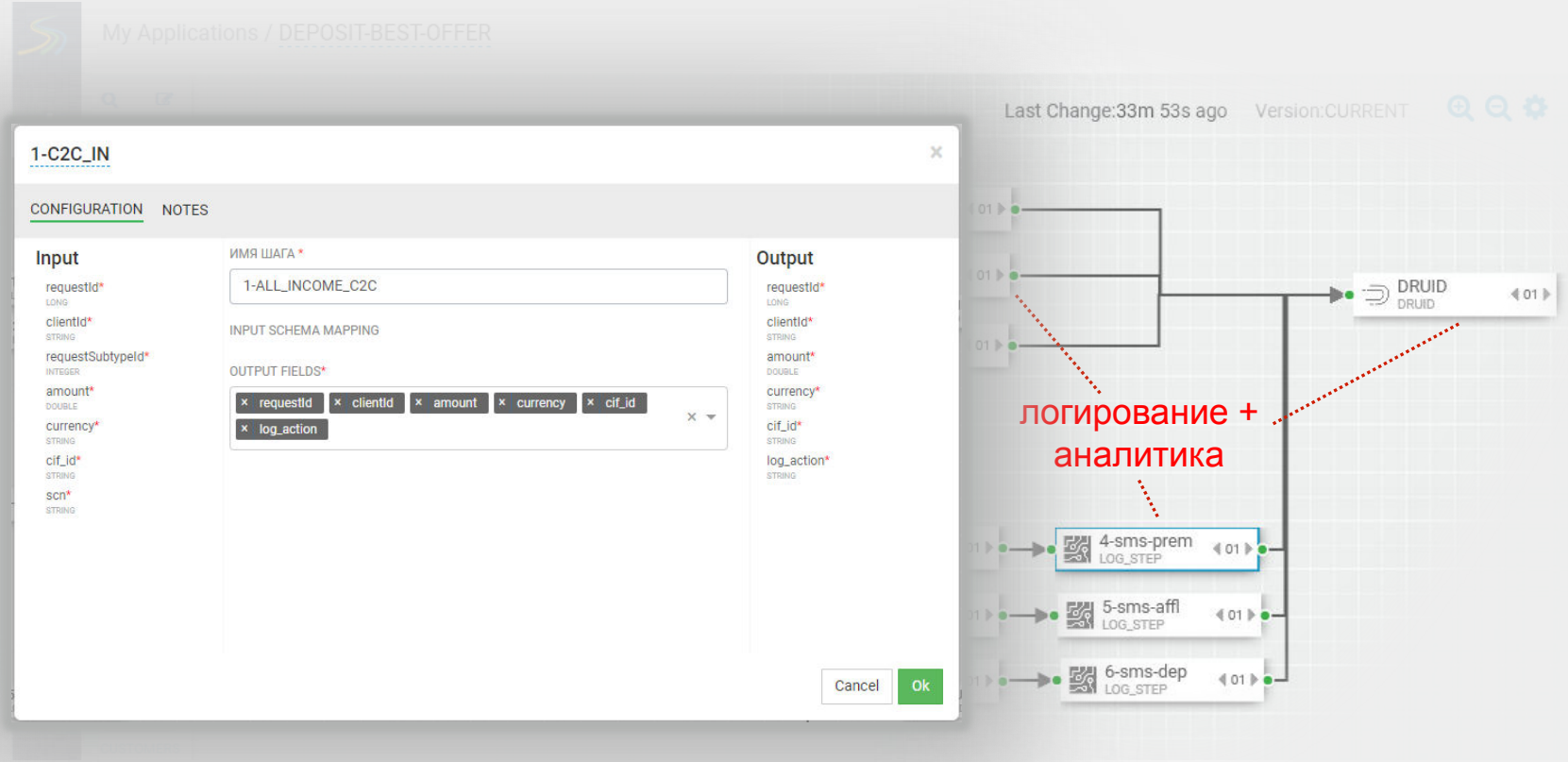


- кол-во Executors

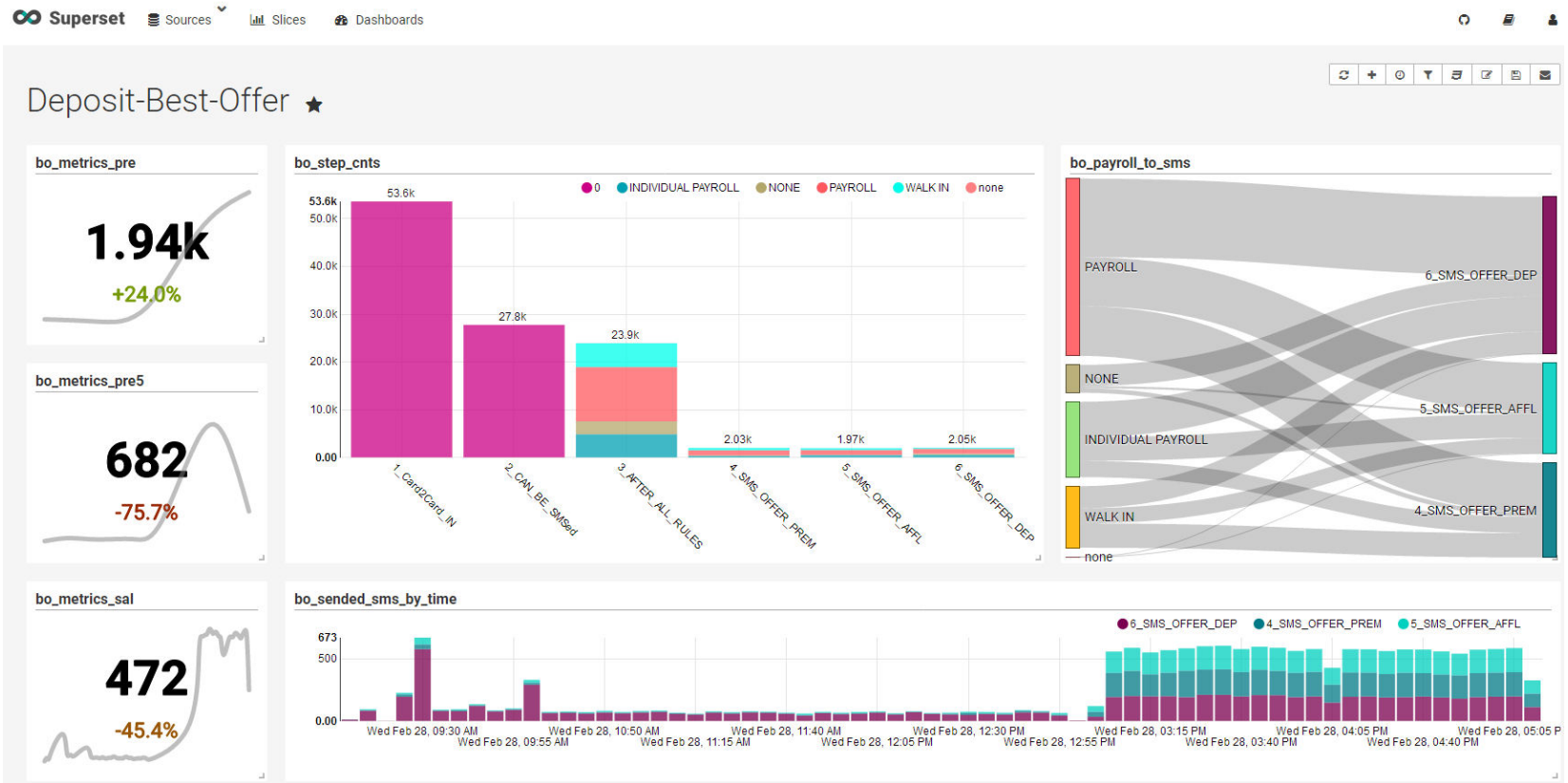
- Готовность к дублям
- Идемпотентность операций

- Крайне мало преднастроенных процессоров
- Разработка кастомных процессоров

SAM + DRUID


My Applications / DEPOSIT-BEST-OFFER
Last Change:33m 53s ago
Version:CURRENT
1-C2C_IN
CONFIGURATION
NOTES
Input
requestId*
LONG
clientId*
STRING
requestSubtypeId*
INTEGER
amount*
DOUBLE
currency*
STRING
cif_id*
STRING
scn*
STRING
ИМЯ ШАГА *
1-ALL_INCOME_C2C
INPUT SCHEMA MAPPING
OUTPUT FIELDS*
requestId
clientId
amount
currency
cif_id
log_action
Output
requestId*
LONG
clientId*
STRING
amount*
DOUBLE
currency*
STRING
cif_id*
STRING
log_action*
STRING
Cancel
Ok
DRUID
DRUID
01
логирование +
аналитика
4-sms-prem
LOG_STEP
5-sms-affl
LOG_STEP
6-sms-dep
LOG_STEP

Druid + Superset: Аналитика и Дашборды



Druid (by Hortonworks?)

- Druid Historical
- Druid Broker
- Druid Coordinator
- Druid Router
- Druid Overlord
- Druid MiddleManage

HDF

CPU = 4 RAM = 16 HDF Master	CPU = 4 RAM = 24 Storm Nifi
CPU = 4 RAM = 16 Storm Druid Hist	CPU = 4 RAM = 24 Druid MM Overlord

- ✓ Иногда может зависнуть без видимых причин, загружая поток со скоростью 1-2 события в секунду
- ✓ Есть подозрение, что это связано с тем что не хватает ресурсов на индексацию
- ✓ Открывается слишком много “временных окон” для индексации? – ~~не~~ факт

Развертывание в Enterprise

SAM -> Maven

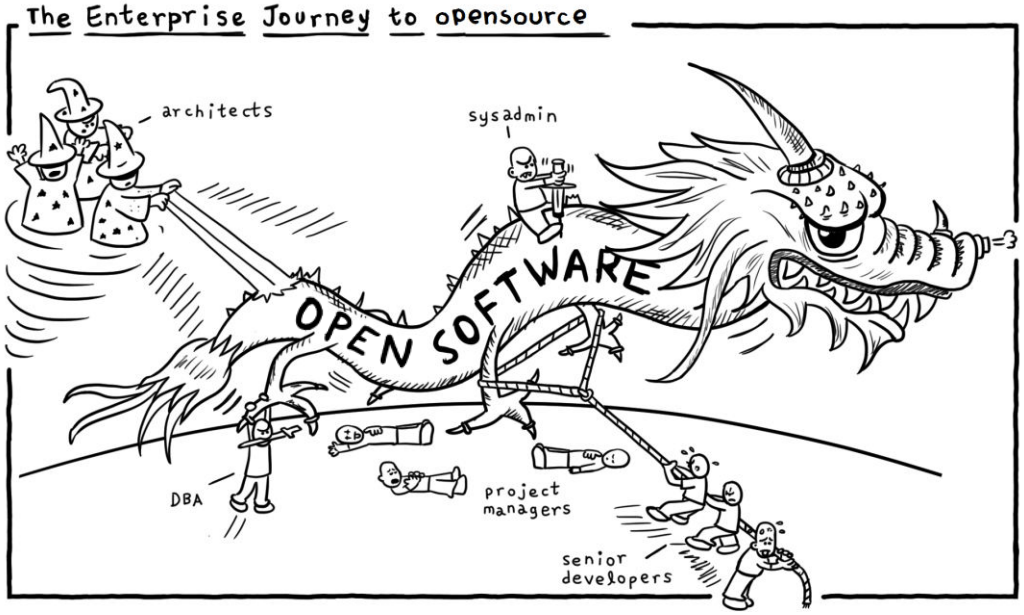
Собирает JAR-файл для Storm'а

при **<u>каждом</u>** запуске SAM-кампании

→ доступ в интернет с Prod-серверов

решение:

1. выкачать библиотеки и сложить в локальную папку
2. патч SAM (переписать класс)



Pic source: https://builttoadapt.io/enterprise-journey-to-the-cloud-aa9e9024a976

Hortonworks (HDP+HDF)

Каждое обновление платформы "как в первый раз":

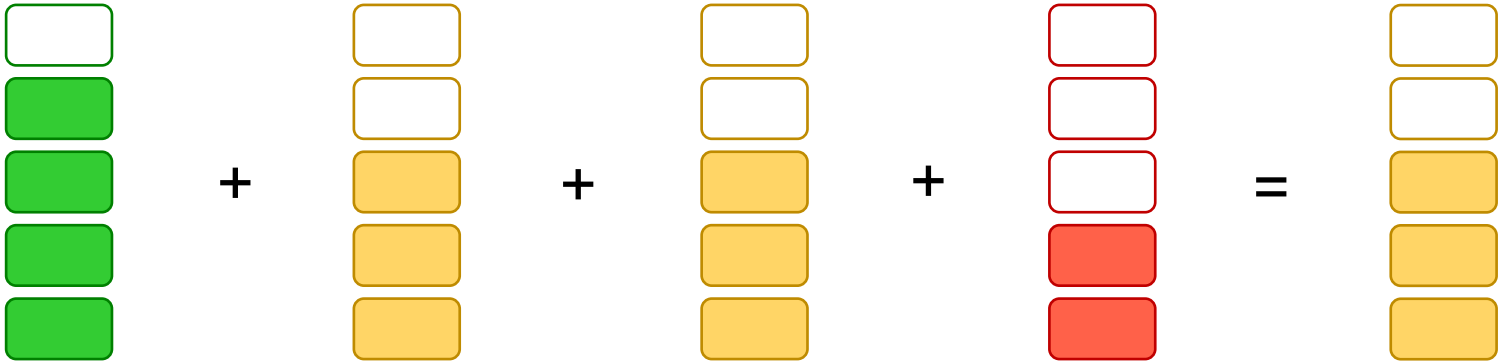
- Необходимых для работы конфигов/JAR-файлов может не быть в поставке
- Регулярные конфликты старых и новых версий библиотек (вычищаем руками)

Райффайзен БАНК

Опыт HDF после 2-х недель пилота

(и 3-х месяцев развертывания)

GUI + разработка + управление кампаниями + Druid = готовность к Production



- сырость текущей версии
- скудные возможности кастомизации
- жесткое задание схемы данных (in/out)
- нельзя изменять работающую кампанию
- интерфейс SAM – отдельно, Storm – отдельно
- потребляет много RAM
- показал низкую производительность

Выводы

ОНО реально РАБОТАЕТ !

НО, Вам придется:

- запастись терпением
- успокоительными
- ввести запрет на мат на рабочем месте

ЧТО, можно сделать лучше:

- DRUID –> ClickHouse
- Storm –> Spark Structured Streaming
- Автоматическое логирование

Спасибо!

aleksey.kuznetsov@raiffeisen.ru

anton.yarmolyuk@raiffeisen.ru