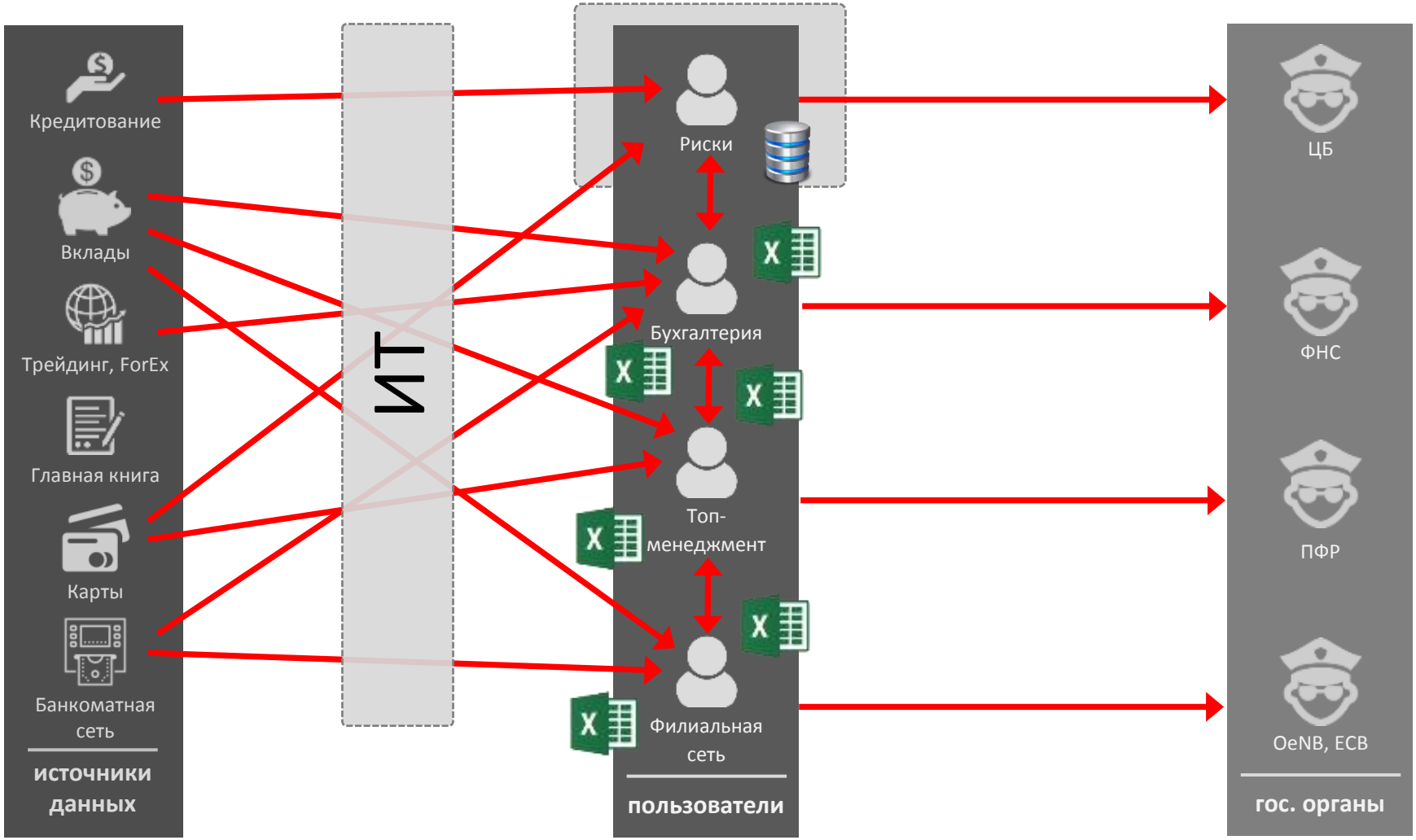
Data Lake в банке

Михаил Сеткин

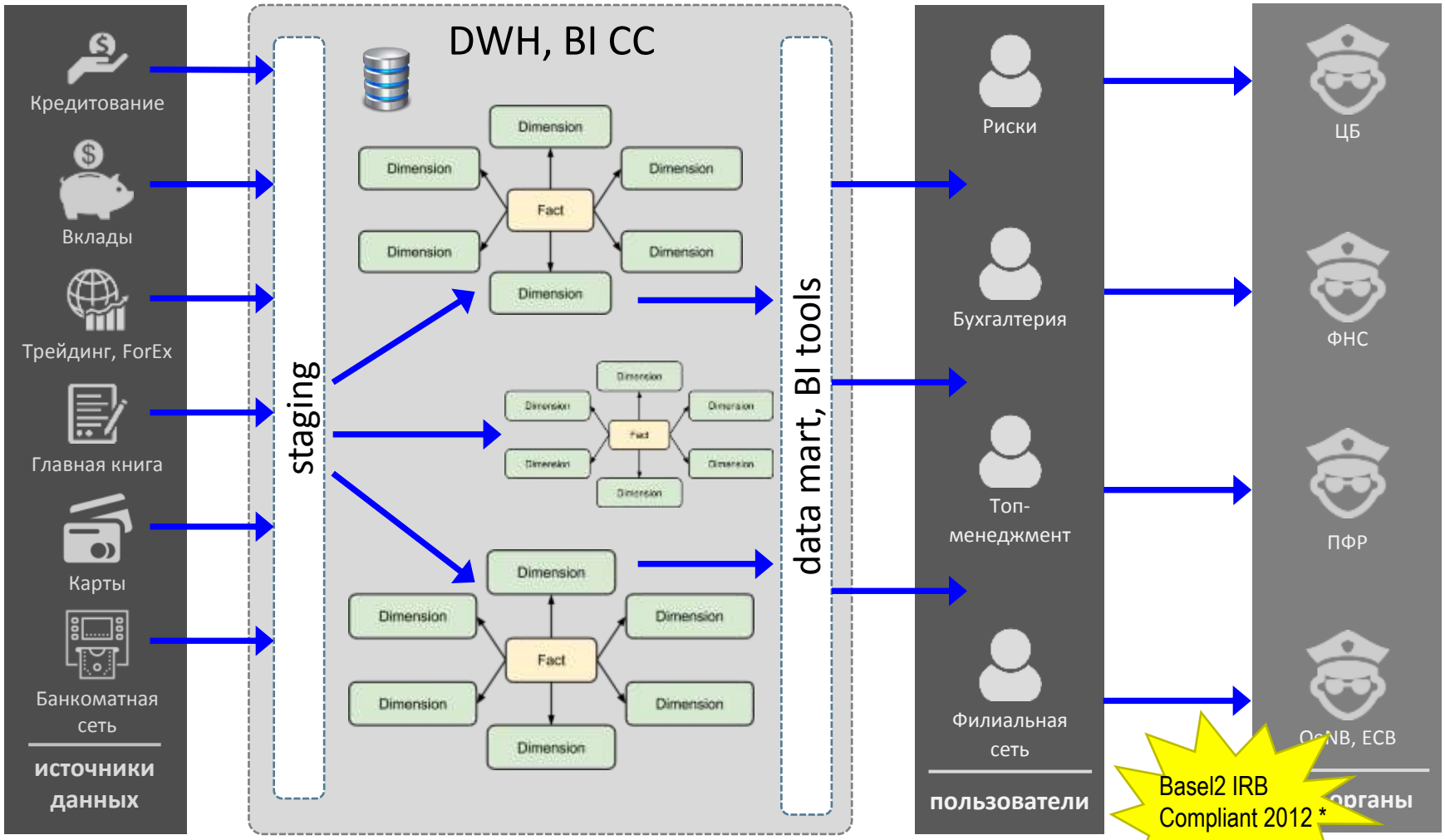
Райффайзен БАНК

Разница
в отношении

Исходное состояние



ИТ-ландшафт: DWH



* https://www.raiffeisen.ru/about/press/releases/?id1630=25150#_ftn11

Характеристики DWH

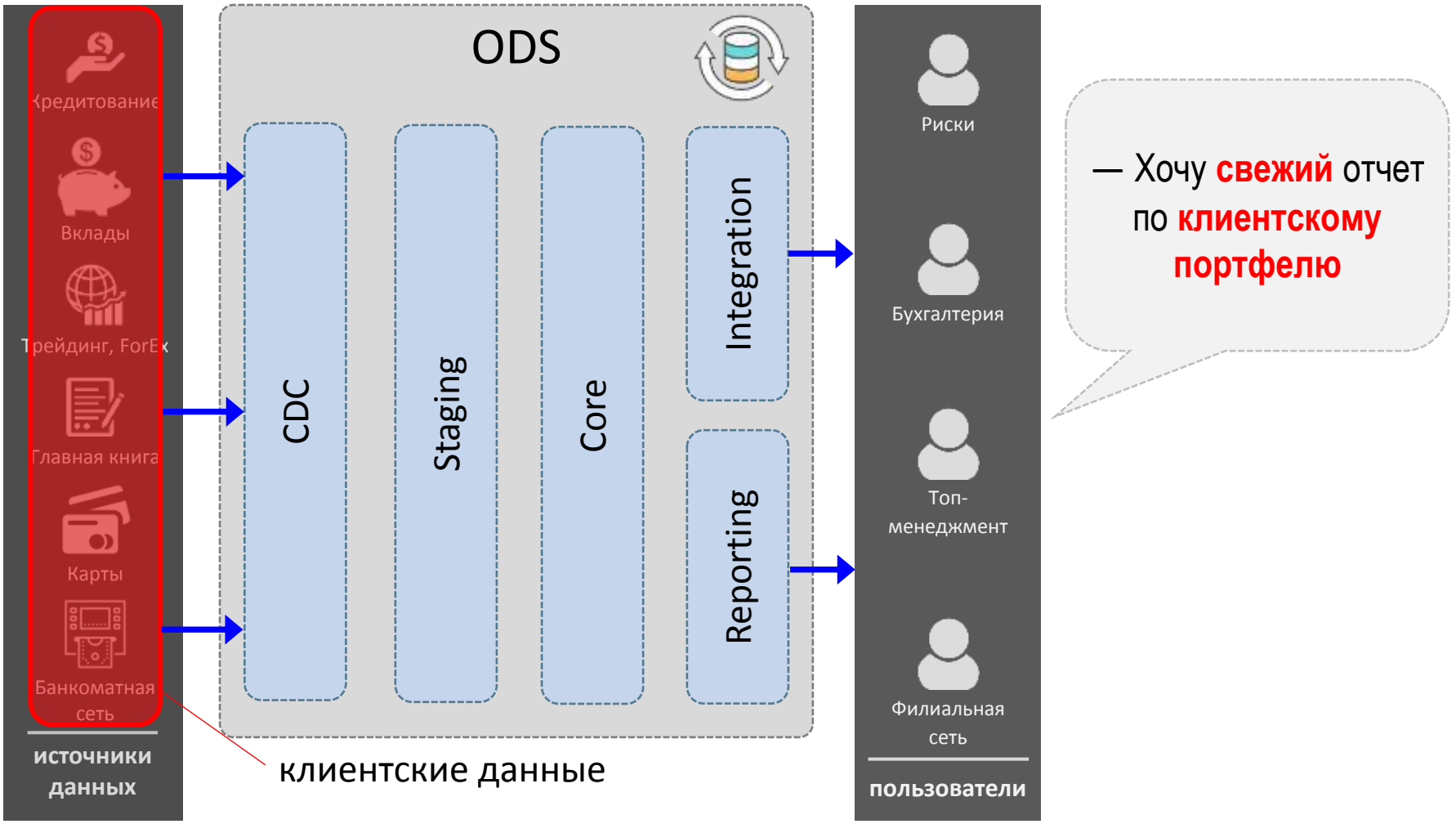
DWH

Хотим быстрее

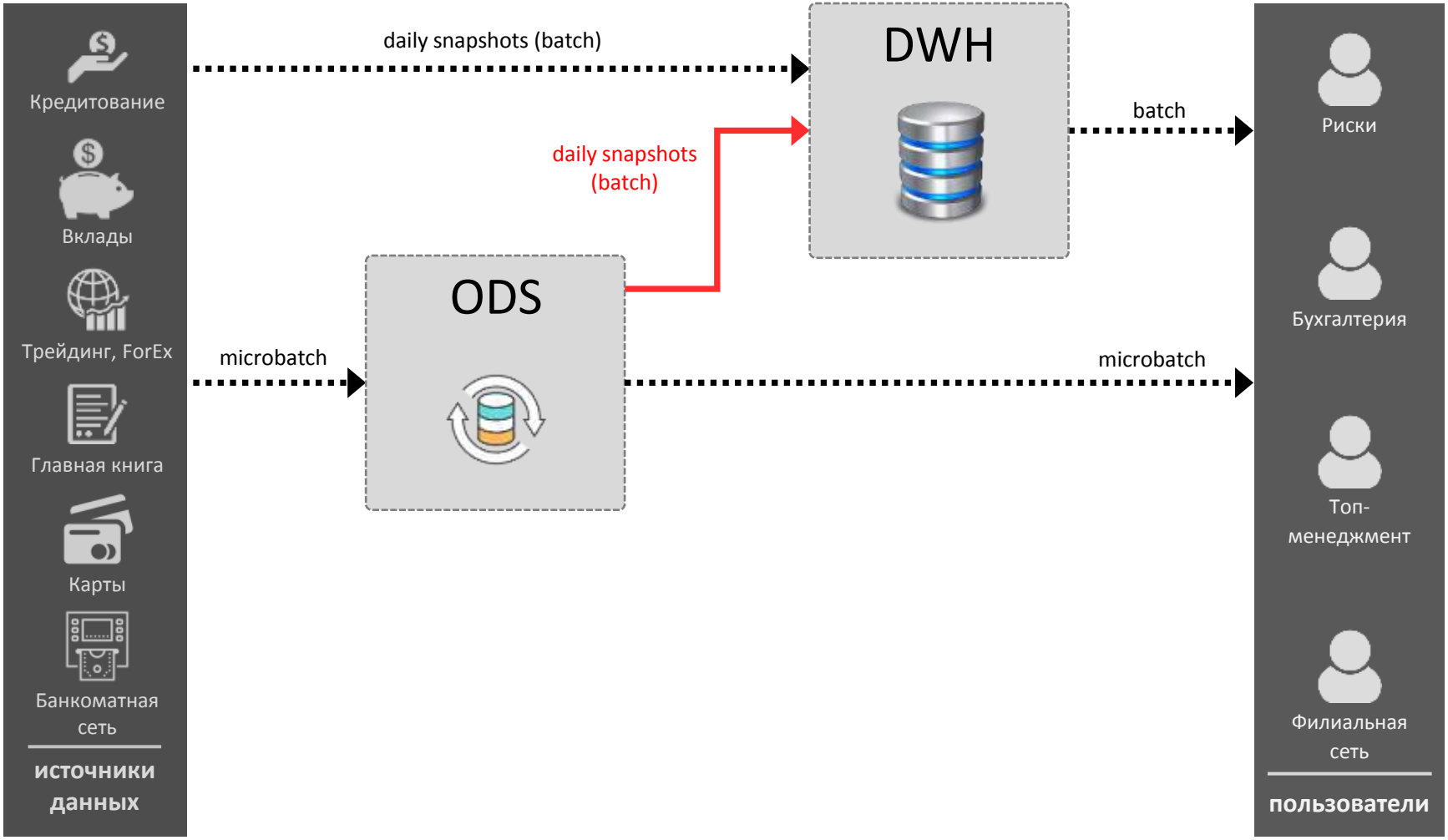
ODS

- Аналитическая отчетность
- Жесткие требования к DQ
- Историчность (SCD type 2)
- Монолитная enterprise-СУБД
- Модель данных – схема «звезда»
- Более 80 систем-источников
- Структурированные данные
- Batch (1 раз в сутки)

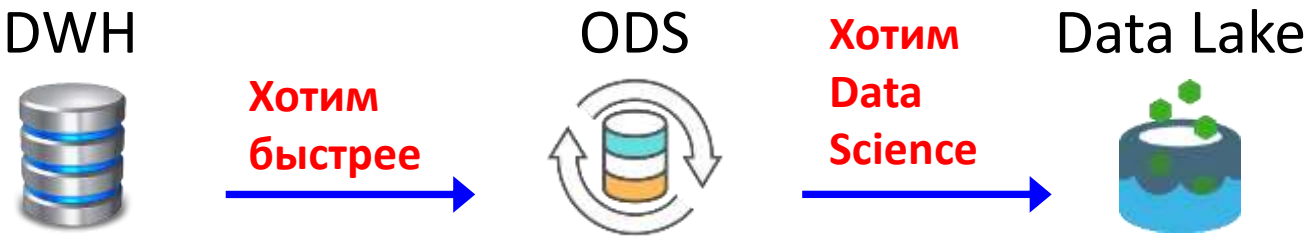
Operational Data Store (ODS)



ИТ-ландшафт: DWH+ODS

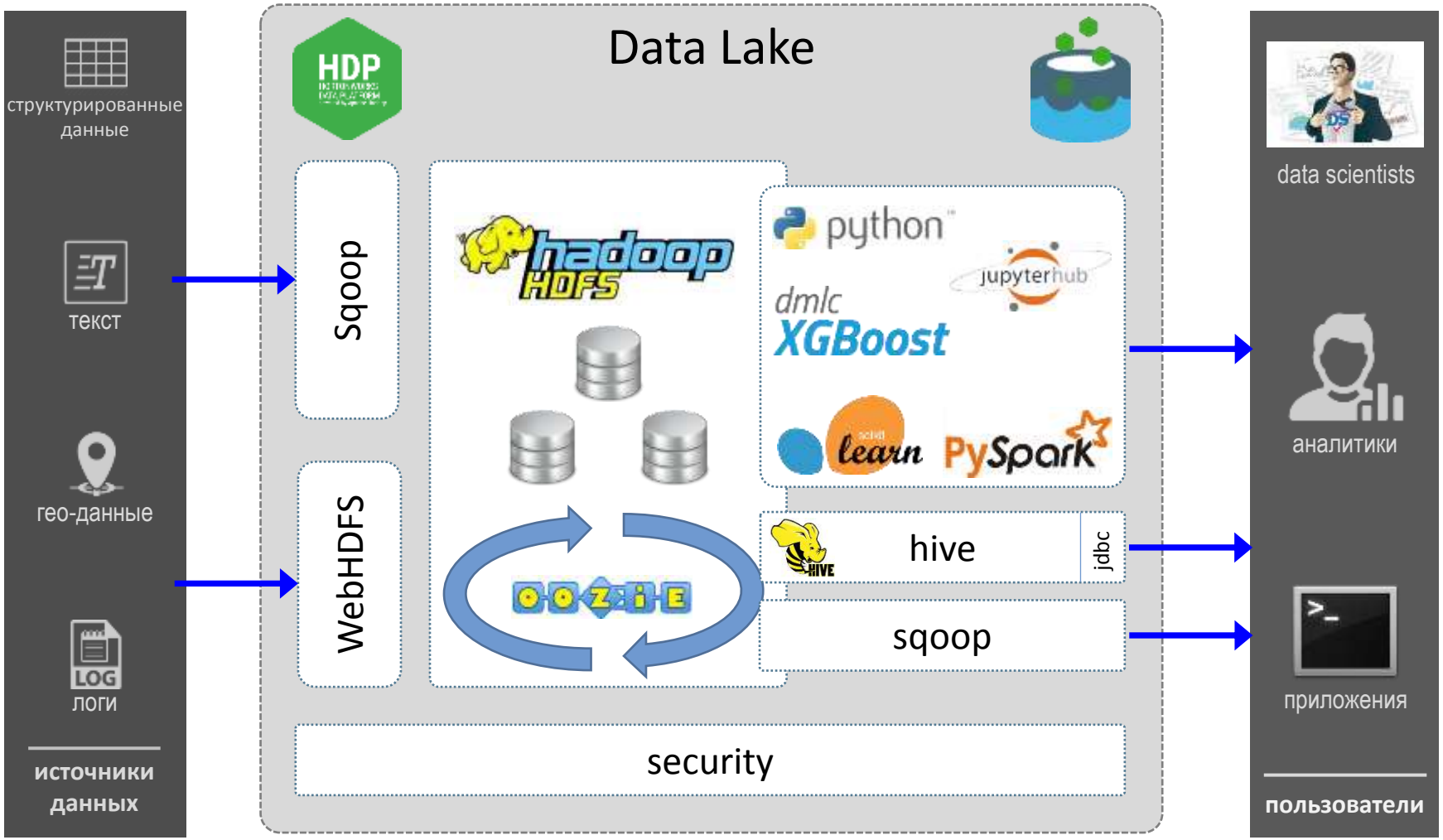


Характеристики ODS



Аналитическая отчетность	**Операционная отчетность и интеграция**
Жесткие требования к DQ	**Жесткие требования к DQ**
Историчность (SCD type 2)	**Отсутствие истории (SCD type 1)**
Монолитная enterprise-СУБД	**Монолитная enterprise-СУБД**
Модель данных – схема «звезда»	**Каноническая модель данных**
Более 80 систем-источников	**Более 80 систем-источников**
Структурированные данные	**Структурированные данные**
Batch (1 раз в сутки)	**Microbatch (1 раз в 1-15 минут)**

Data Lake (batch)

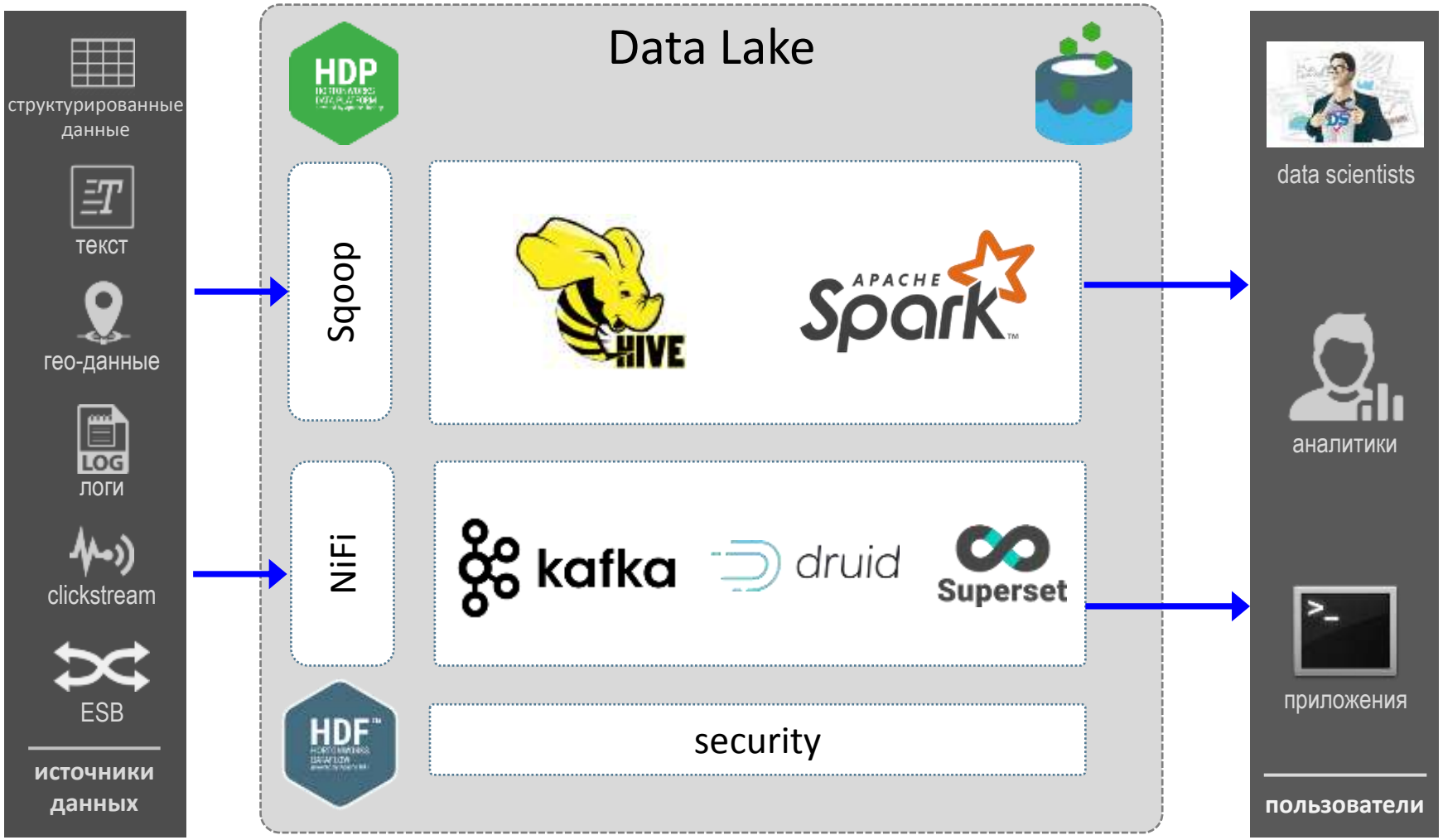


Райффайзен БАНК

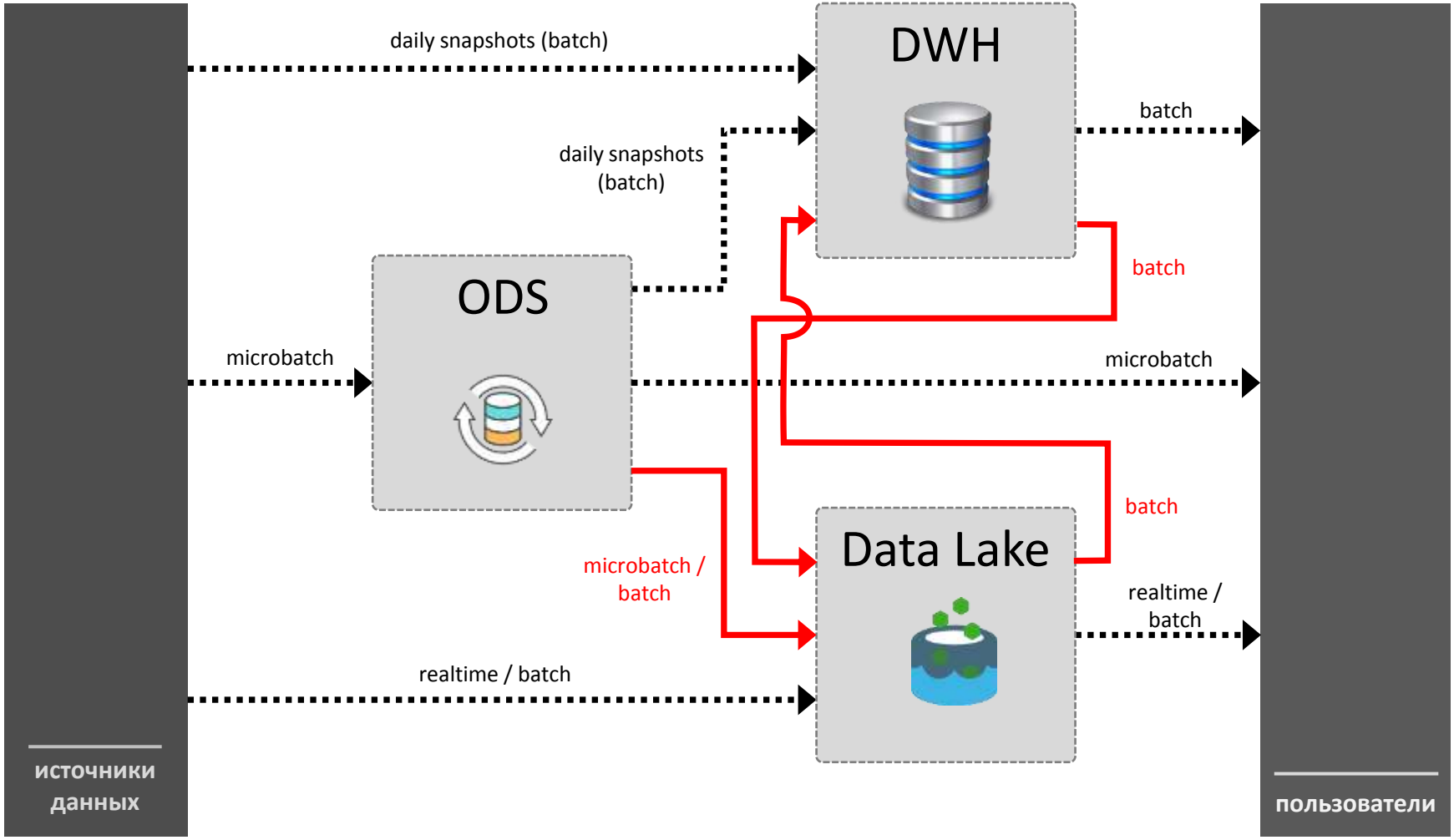
Бонус: Hortonworks Data Flow

HDF™
HORTONWORKS DATAFLOW
powered by Apache NiFi

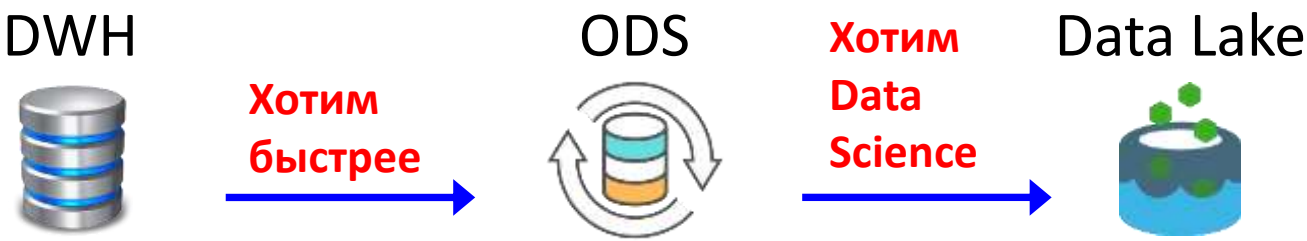
Data Lake (batch+realtime)



ИТ-ландшафт: DWH+ODS+Data Lake



Характеристики Data Lake



Аналитическая отчетность	Операционная отчетность и интеграция	**Data Science**
Жесткие требования к DQ	Жесткие требования к DQ	**Нет требований к DQ**
Историчность (SCD type 2)	Отсутствие истории (SCD type 1)	**Историчность «Full Snapshots»**
Монолитная enterprise-СУБД	Монолитная enterprise-СУБД	**Экосистема Hadoop**
Модель данных – схема «звезда»	Каноническая модель данных	**Структура как в источнике**
Более 80 систем-источников	Более 80 систем-источников	**10 систем-источников**
Структурированные данные	Структурированные данные	**Любые данные**
Batch (1 раз в сутки)	Microbatch (1 раз в 1-15 минут)	**от Batch до Realtime**

Райффайзен БАНК

Решаемые задачи

- Поиск аномалий/statistical outliers
- Кластеризация клиентской базы
- Анализ временных рядов
- Геокодирование
- Предотвращение клиентского оттока
- Анализ логов посетителей сайта/приложения

Райффайзен БАНК

Спасибо за внимание!

- https://habrahabr.ru/users/msetkin/
- https://www.linkedin.com/in/mikhail-setkin-6aa864b1