СБЕРБАНК
ТЕХНОЛОГИИ

Репликация баз данных в Hadoop

Вадим Сурпин
Руководитель отдела Спецпроектов BigData,
VPSurpin.SBT@sberbank.ru

КАКИЕ ЗАДАЧИ РЕШАЕТ БАНК НА ТЕХНОЛОГИЯХ BIGDATA

ПОДГОТОВКА ДАННЫХ ДЛЯ ПРИНЯТИЯ РЕШЕНИЙ

Маркетинговые акции

Сегментация клиентов

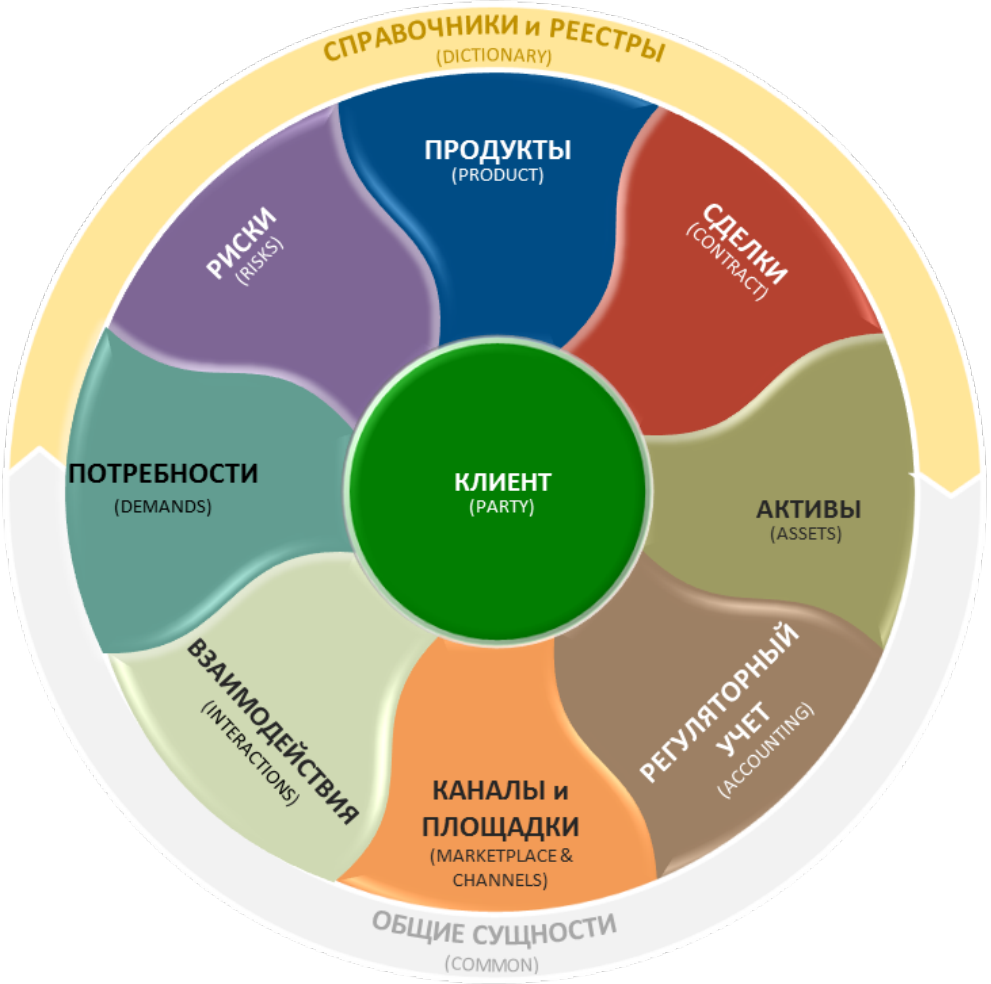
Предсказание оттока клиентов

Расчёт предодобренных кредитов

Предсказание клиентопотока

Предсказание остатков средств в банкоматах

РАЗРАБОТКА «ЕДИНОЙ ВЕРСИИ ПРАВДЫ»



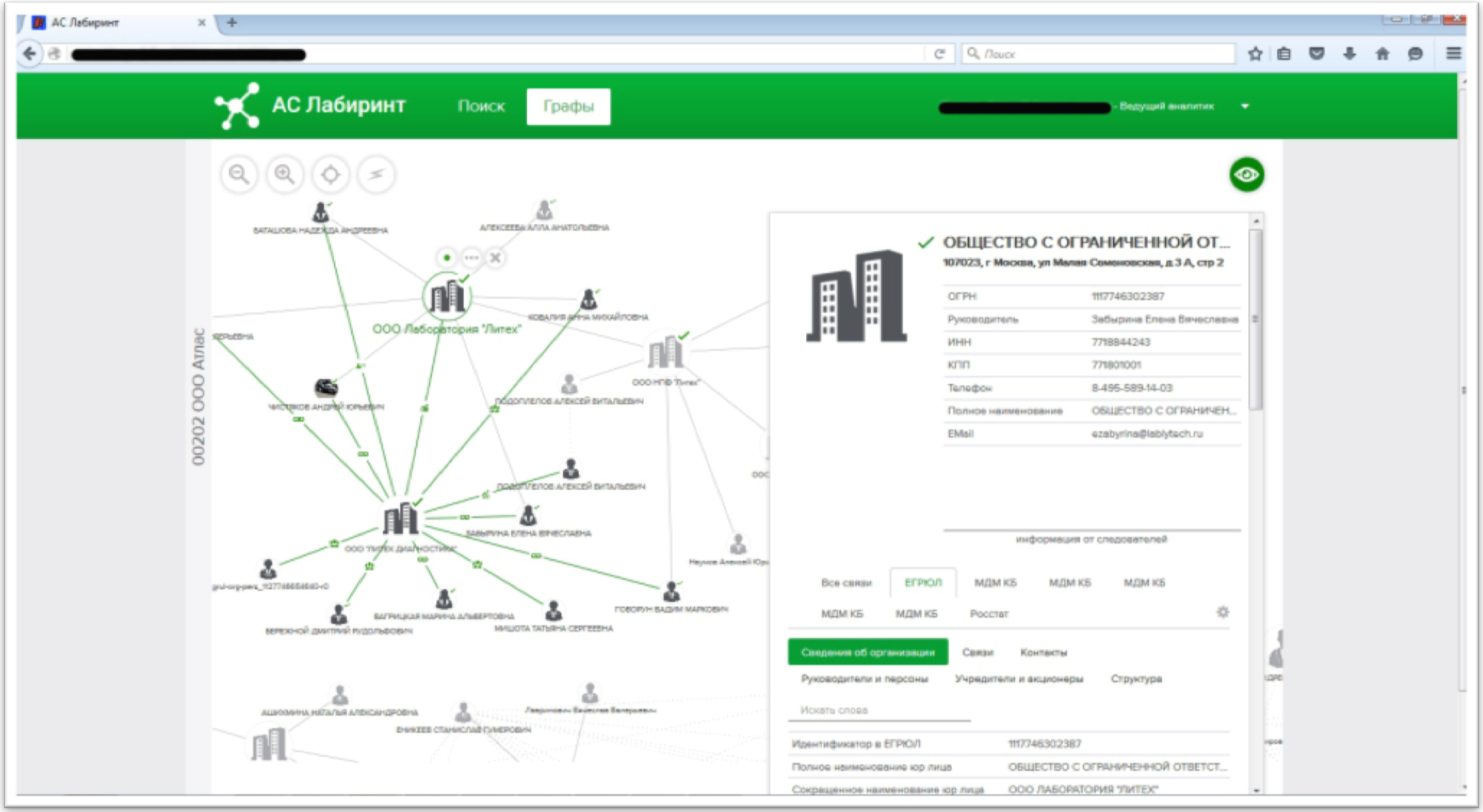
- ❑ Карты, вклады, счета
- ❑ Кредиты и заявки
- ❑ Обращения клиентов
- ❑ GR
- ❑ Терминалы и банкоматы
- ❑ Офисы и электронные очереди
- ❑ Недвижимостью, автоматизация зданий, доступ
- ❑ Кадры и HR
- ❑ ИТ-инфраструктура и сопровождение

В основе -
клиентоцентрическая модель сущностей

ГРАФОВАЯ ЭКОСИСТЕМА

- **Зачем?**
  - ✓ Формирование предложений/триггеров: выявление холдингов, различных индикаторов
  - ✓ Борьба с фродом: поиск массовых руководителей, мошеннических схем, конечных бенефициаров
- **Собственная графовая платформа**
  - ✓ Объем – вся страна
  - ✓ Более 3 млрд. связей
- Абсолютное большинство доступных решений падало или уходило в немыслимые тайм-ауты при загрузке данных
- Фичи:
  - ✓ Хранение разных типов связей
  - ✓ Поиск
  - ✓ Выявление кратчайшего пути
  - ✓ Интерактивная навигация
  - ✓ Поддержка batch алгоритмов

ГРАФОВАЯ ЭКОСИСТЕМА (2)



Факты

СБЕРБАНК ТЕХНОЛОГИИ

ОБЛАКО ДАННЫХ. ФАКТЫ

250

узлов в составе кластера

170

узлов Hadoop data node (HDFS + YARN)

ОБЛАКО ДАННЫХ. ФАКТЫ (2)

6800

CPU Cores

109 Тб

RAM

10 Пб

HDD

СБЕРБАНК ТЕХНОЛОГИИ

ОБЛАКО ДАННЫХ. ФАКТЫ (3)

2.5 ПБ

объем данных, рост до

10 ПБ

в течение 2018г.

15.5 Тб/сутки

объем получаемых изменений

СБЕРБАНК ТЕХНОЛОГИИ

ОБЛАКО ДАННЫХ. ФАКТЫ (4)

2000 – 5000

транзакций в секунду

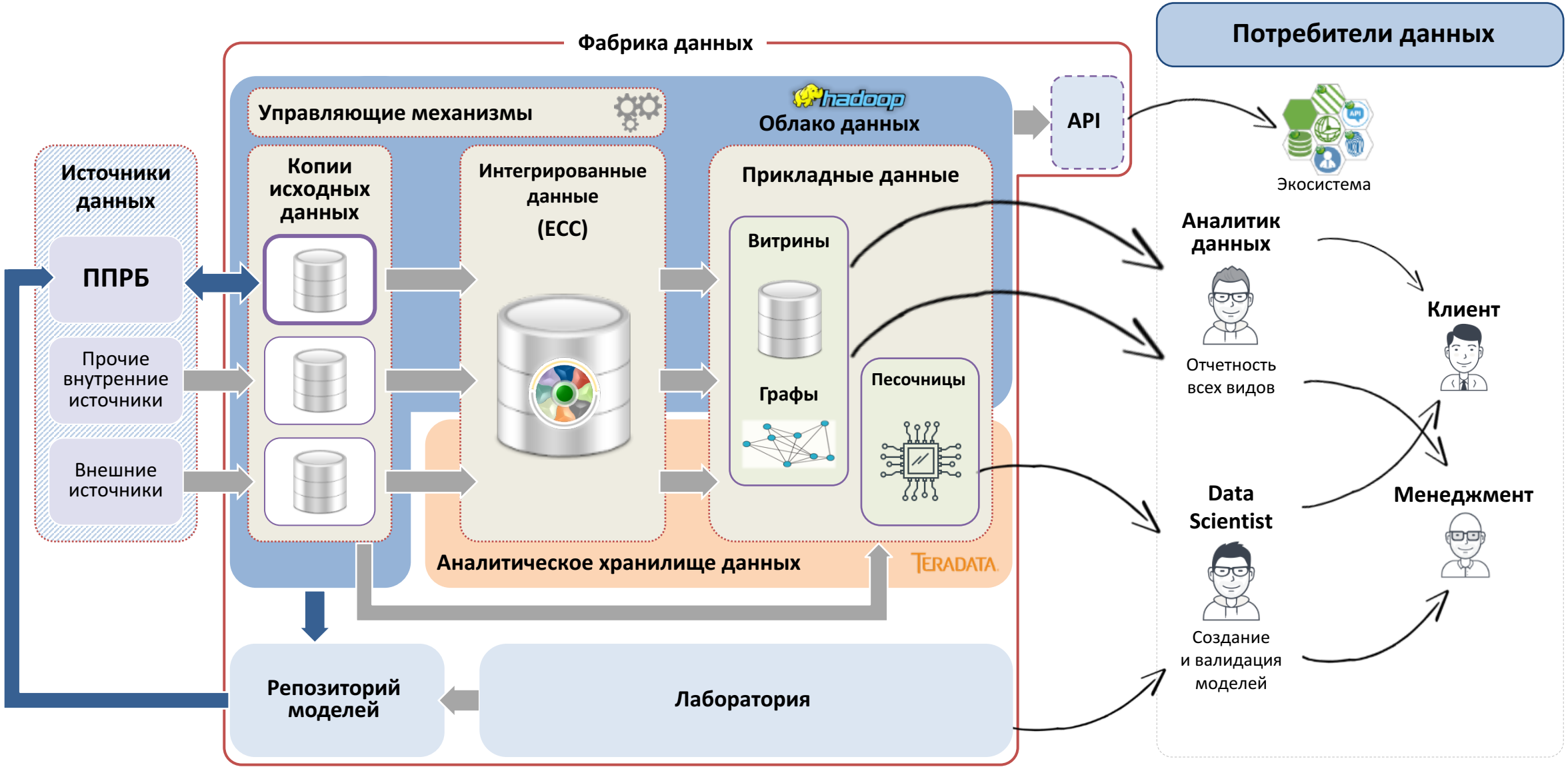
400

параллельных заданий загрузки данных

200 ТБ

объем ежесуточно обновляемой информации в репликах

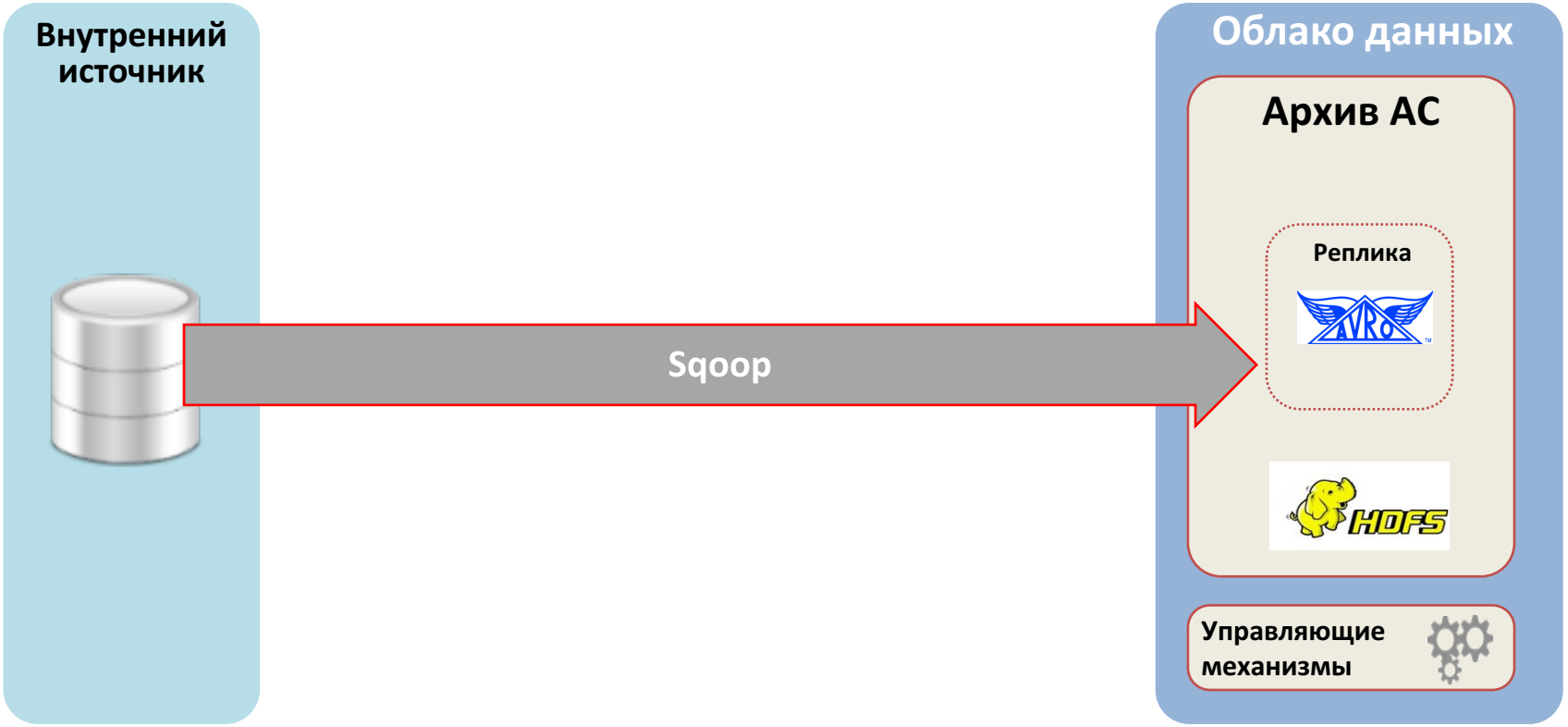
ОСНОВНЫЕ КОМПОНЕНТЫ АРХИТЕКТУРЫ



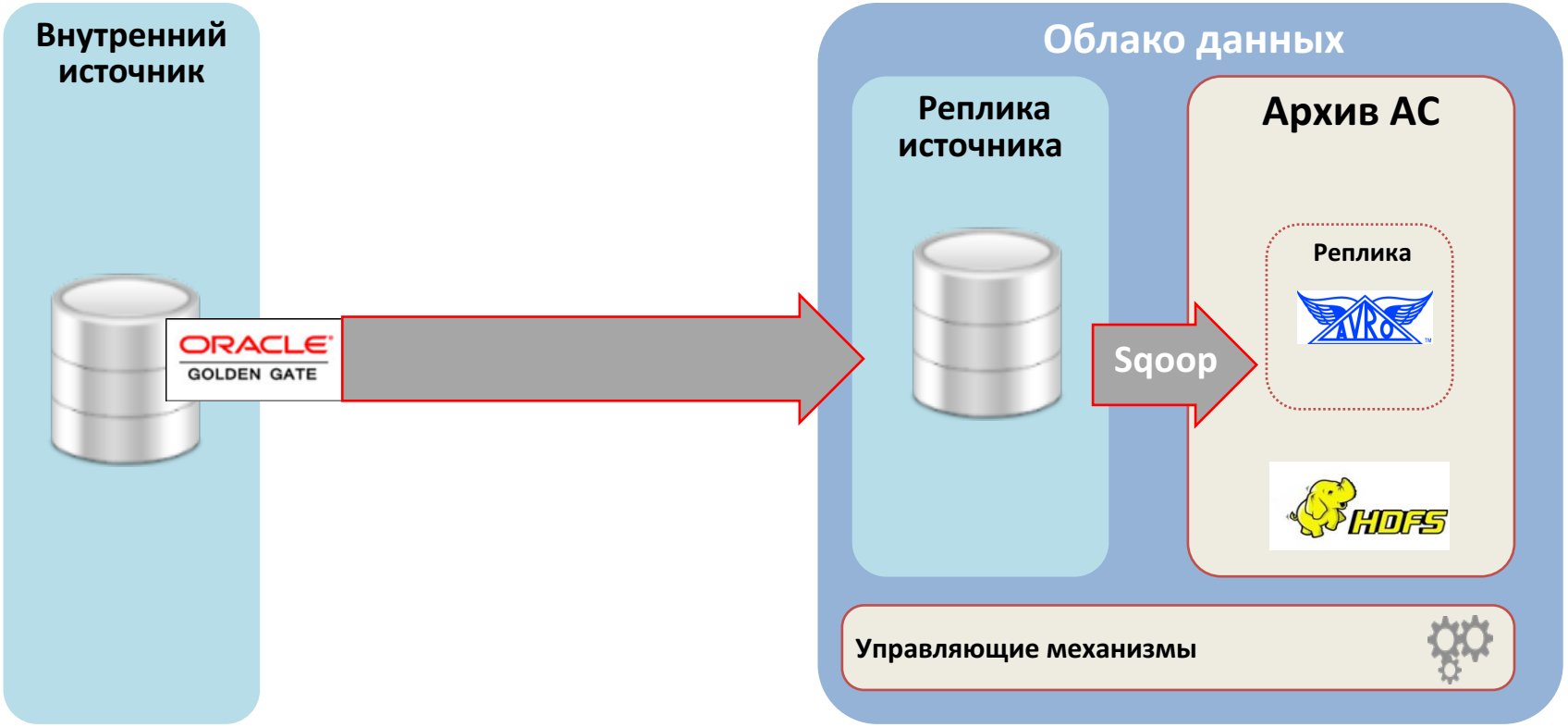
ОСНОВНЫЕ КОМПОНЕНТЫ АРХИТЕКТУРЫ – НАЗНАЧЕНИЕ «ОБЛАКА ДАННЫХ»

- ✓ Единое место сбора, хранения и распространения данных Банка: внутренних, внешних и пользовательских. Интеграция данных в Корпоративную Модель Данных.
- ✓ Функционирование аналитического (исторического) профиля клиентов. Исполнение моделей в пакетном (off-line) режиме.
- ✓ Подготовка специализированных витрин данных для различных потребителей в Банке. Обеспечение потребителей данными в различных режимах (on-line и off-line) и в различных форматах

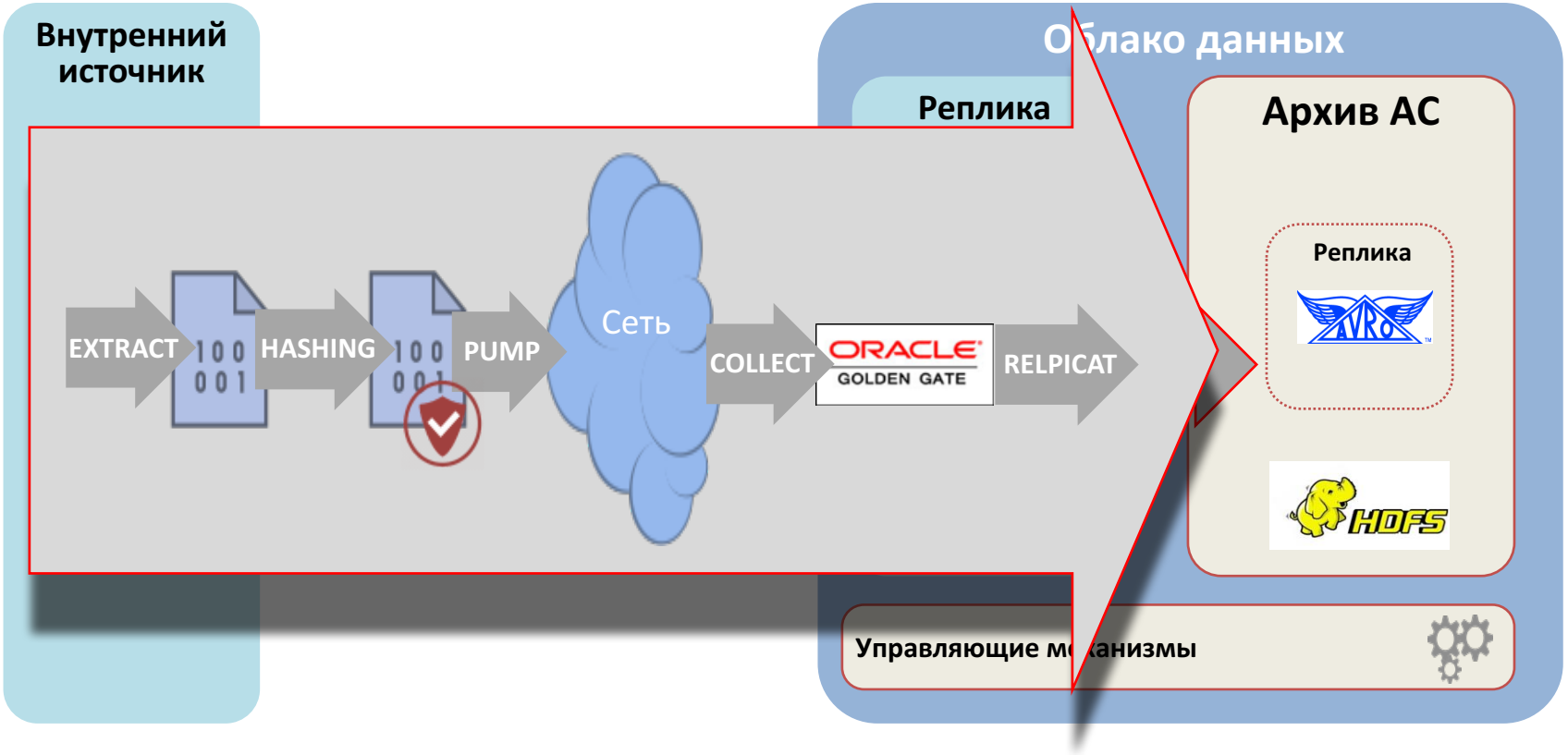
ПЕРВЫЙ ПОДХОД



ВТОРОЙ ПОДХОД



ВТОРОЙ ПОДХОД



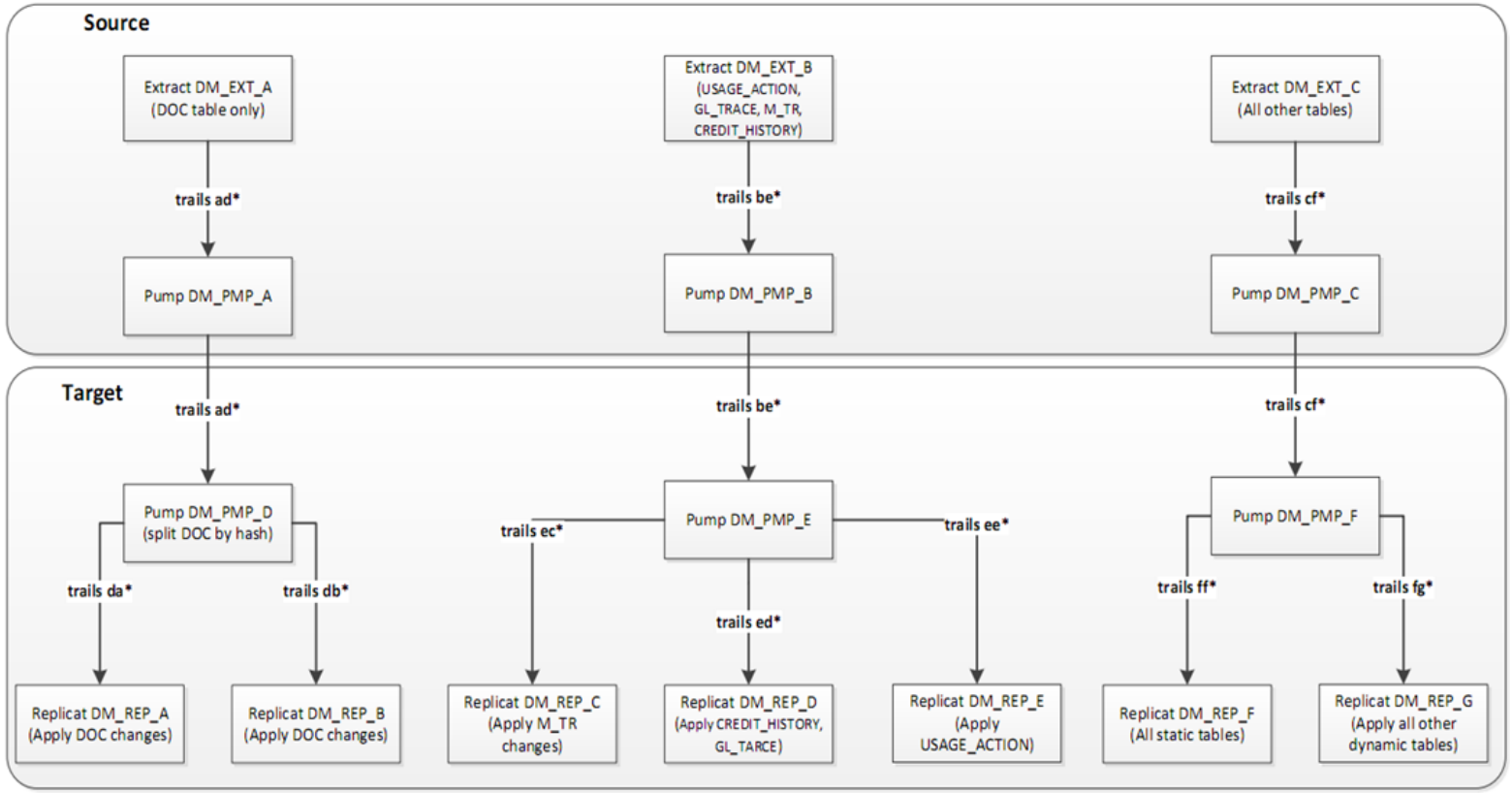
РЕПЛИКАТОР STORK. ПРИМЕРЫ ИСТОЧНИКОВ

Extract – 3 Pump-6 Replicat- 7

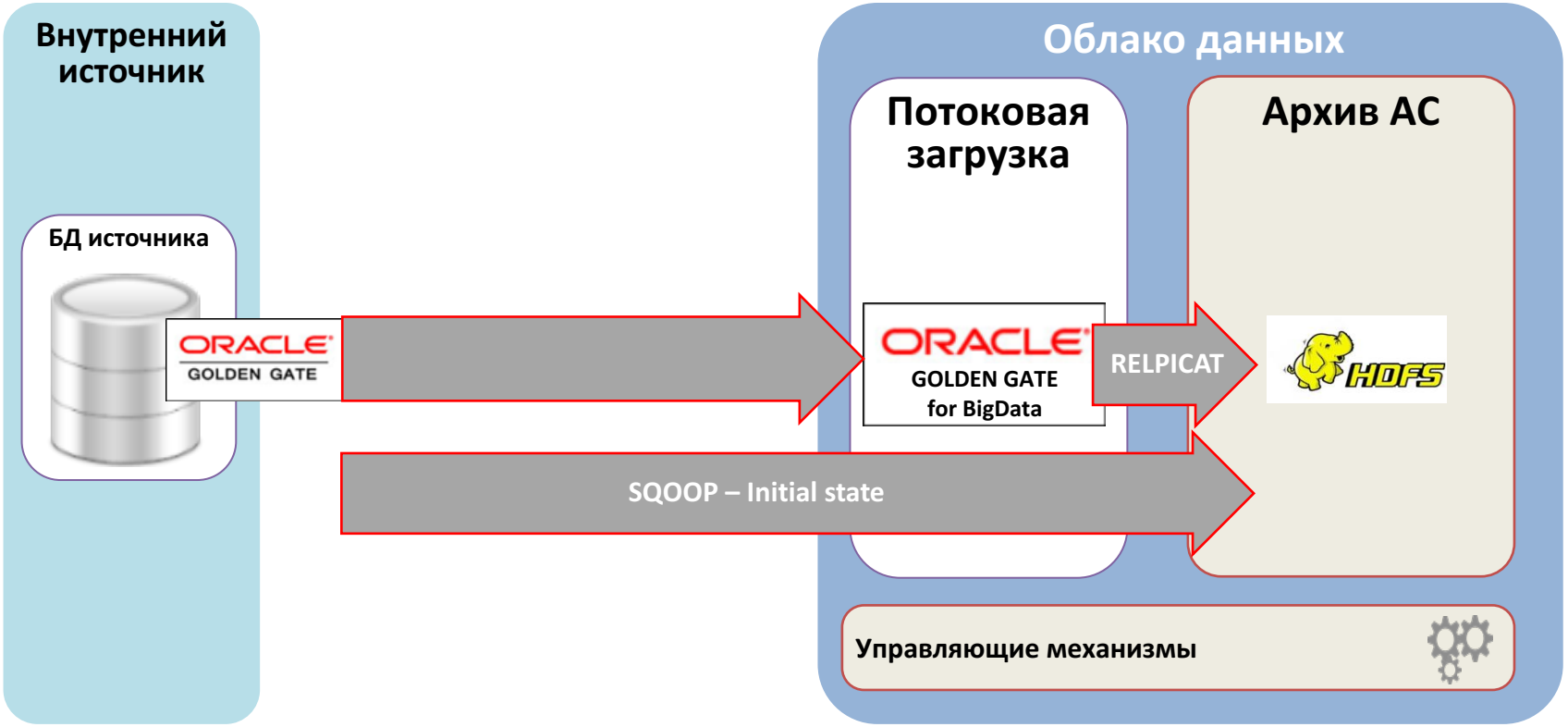
Max: REDO – 1.5 TB per hour
REDO – 22 TB per day

Avg: REDO – 1 TB per hour
REDO – 15 TB per day

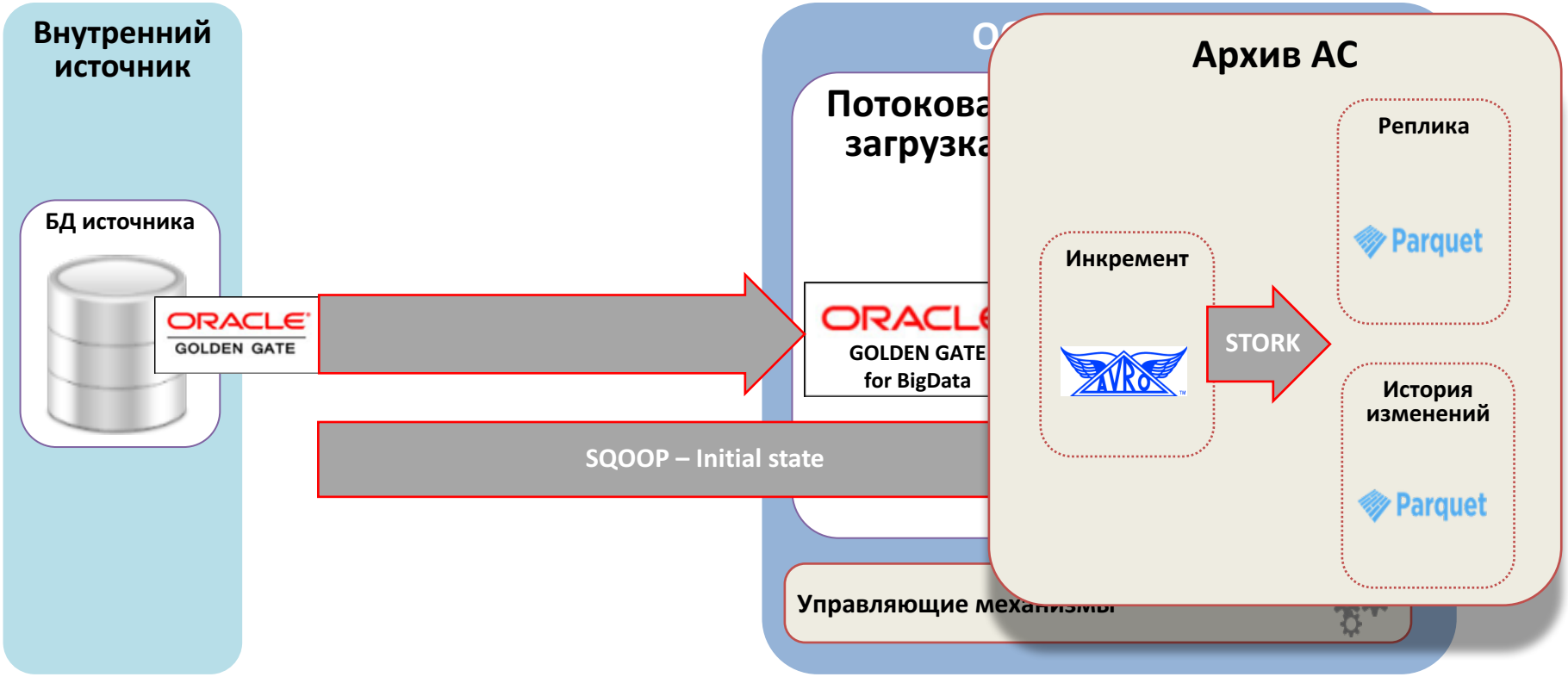
Servers: Oracle RAC 2– M6 (Oracle) , OGG – M7 (Oracle, Veritas)



ТРЕТИЙ ПОДХОД

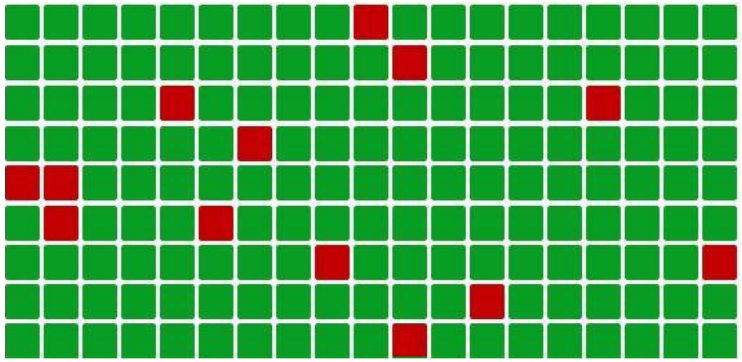


ТРЕТИЙ ПОДХОД



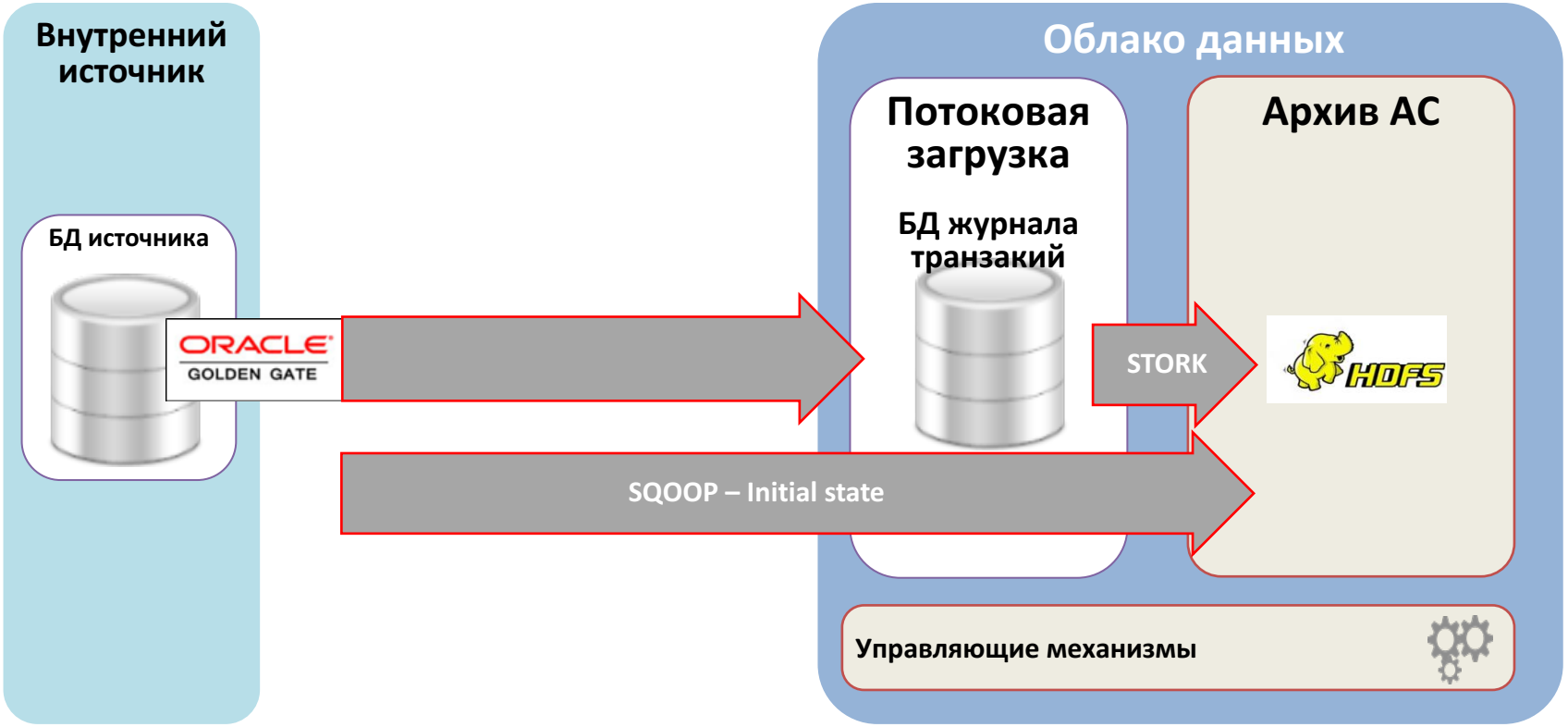
РЕПЛИКАТОР STORK

- «Дырки» в транзакциях и «схлопка»
- Обновление ключа и некорректные операции



- Никто в мире такую задачу не решал
- Объемы
  - ✓ 15 Тб изменений в день
  - ✓ 200 Тб в день нужно переписать
  - ✓ Самая большая таблица – 10 Тб
  - ✓ Наибольшее количество записей в одной таблице – 50 миллиардов

ТРЕТИЙ ПОДХОД: WORKAROUND

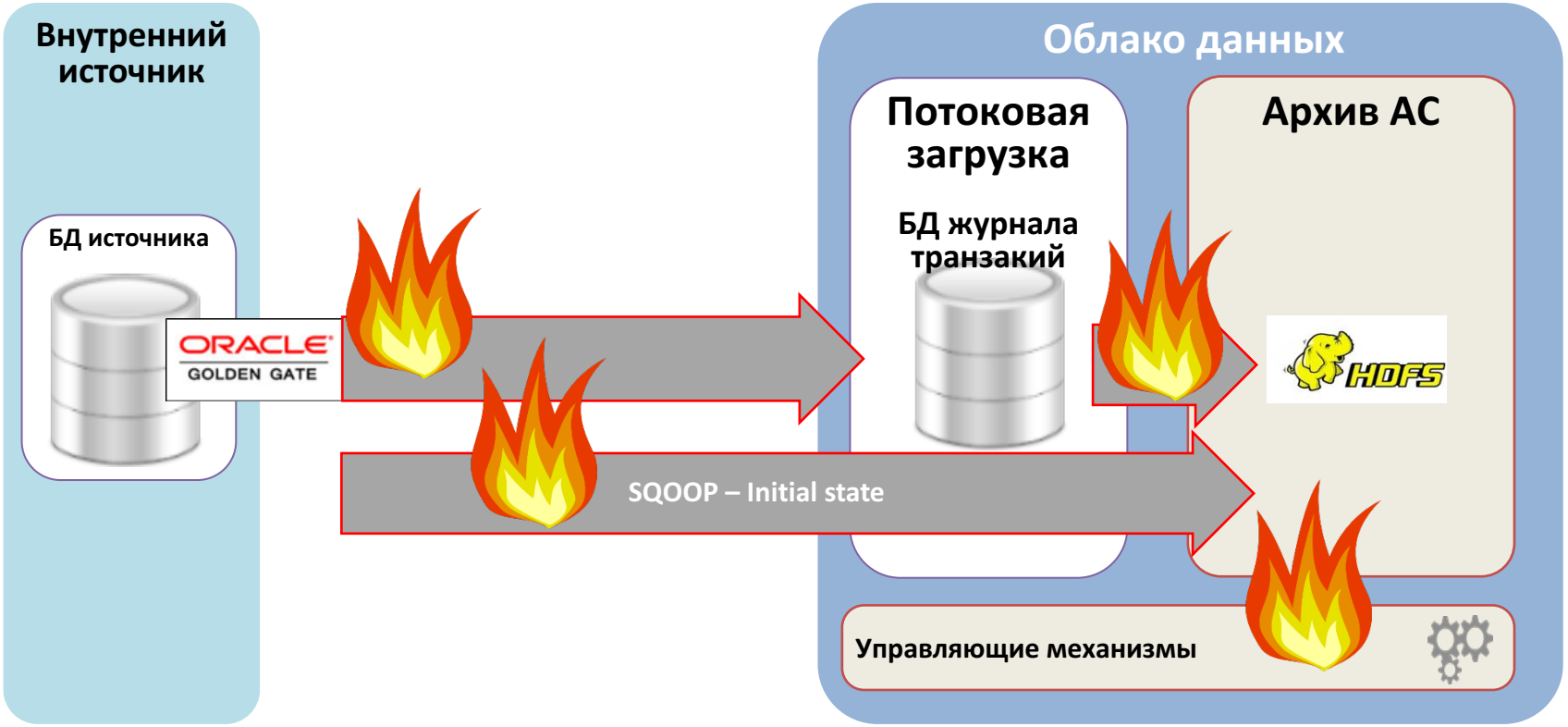


СБЕРБАНК ТЕХНОЛОГИИ

ИНТЕРЕСНЫЕ ЗАДАЧИ, КОТОРЫЕ РЕШАЛИ В ПРОЦЕССЕ РАЗРАБОТКИ

- Инструменты Hadoop
- Интеграция с Базами данных и инструментами репликации
- Особенности информационных систем банка
- Архитектурные и организационные задачи

ТРЕТИЙ ПОДХОД: WORKAROUND

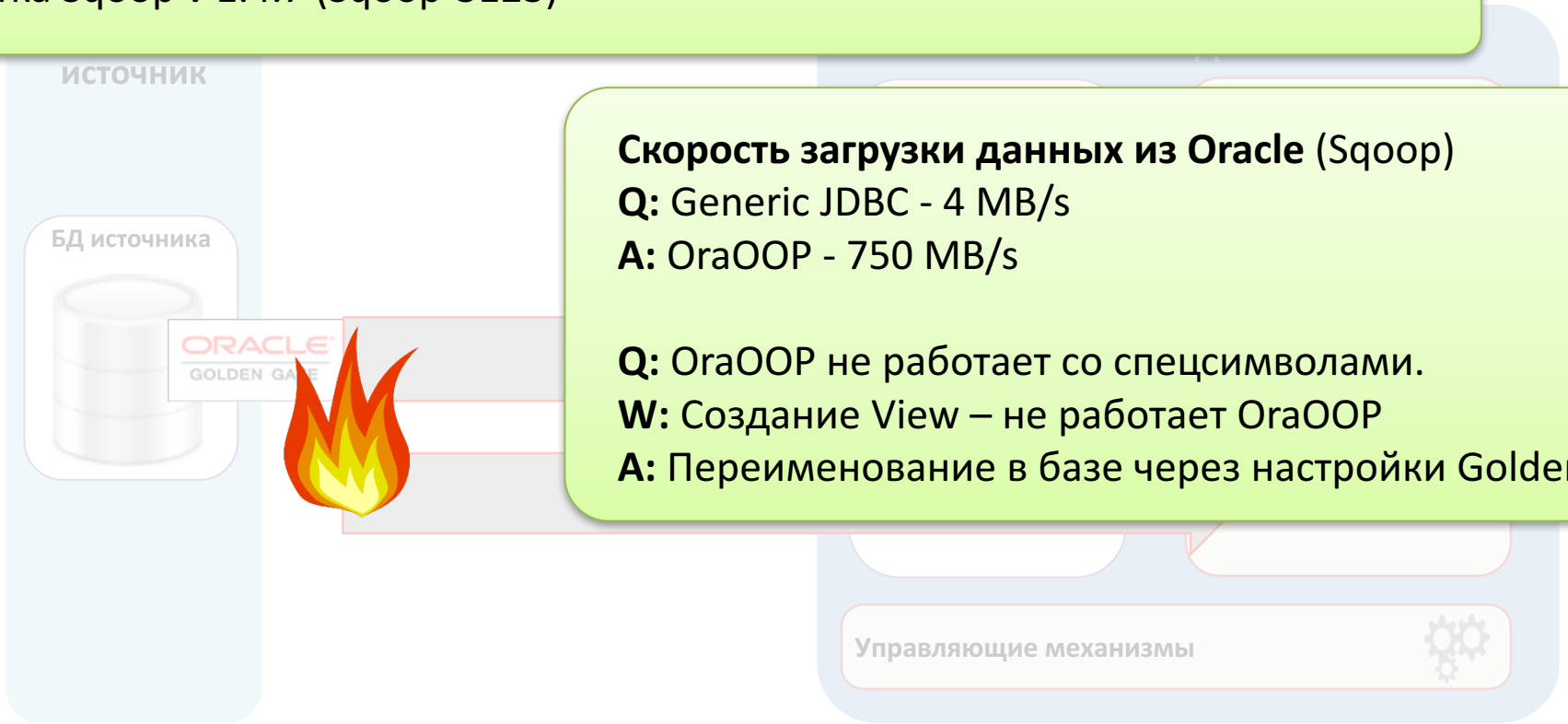


РЕШЕНИЕ ЗАДАЧ С ИНСТРУМЕНТАМИ HADOOP

Спецсимволы в наименовании полей баз данных (Sqoop)
Q: Названия полей в формате AVRO должны быть допустимыми именами Java
A: Доработка Sqoop v 1.4.7 (Sqoop-3123)

РЕШЕНИЕ ЗАДАЧ С ИНСТРУМЕНТАМИ HADOOP (2)

Спецсимволы в наименовании полей баз данных (Sqoop)
Q: Названия полей в формате AVRO должны быть допустимыми именами Java
A: Доработка Sqoop v 1.4.7 (Sqoop-3123)

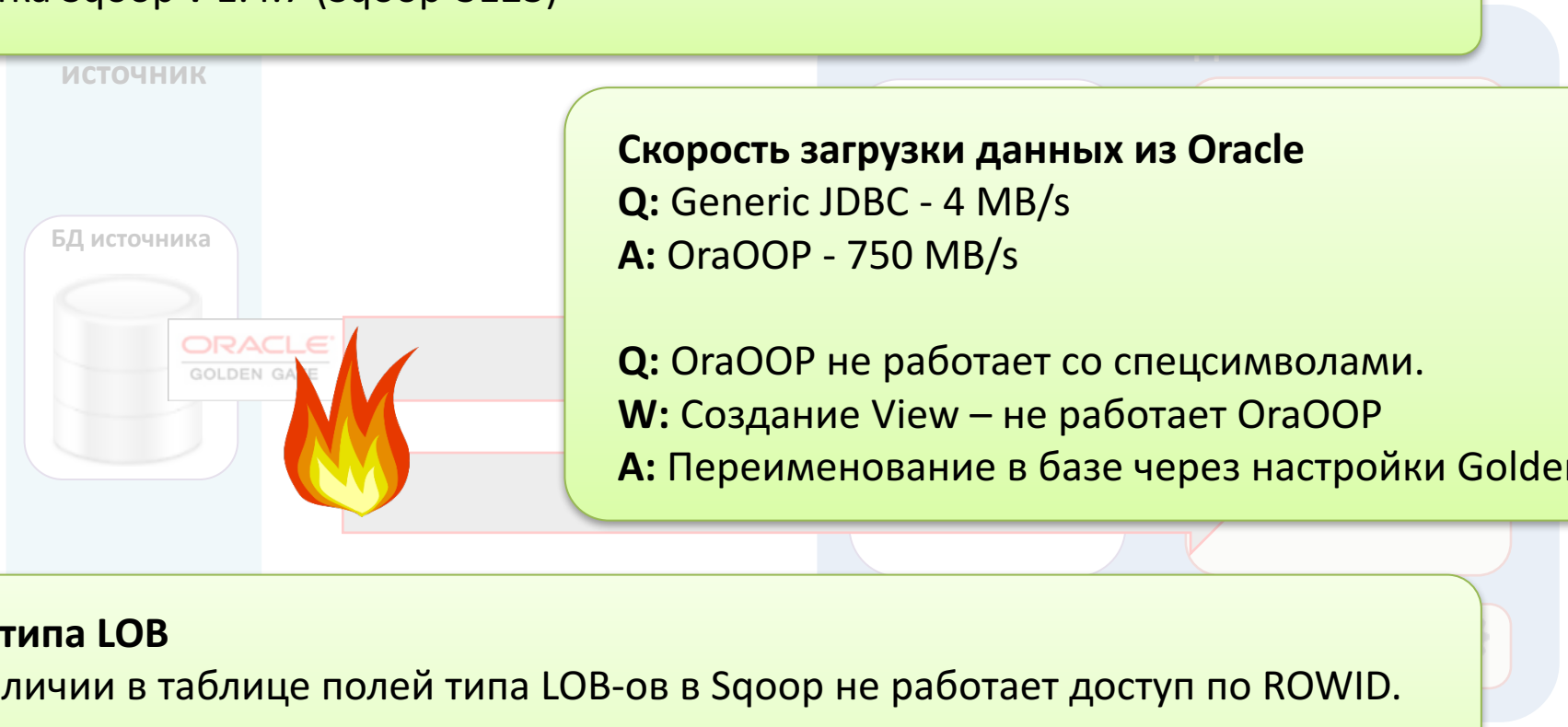


Скорость загрузки данных из Oracle (Sqoop)
Q: Generic JDBC - 4 MB/s
A: OraOOP - 750 MB/s

Q: OraOOP не работает со спецсимволами.
W: Создание View – не работает OraOOP
A: Переименование в базе через настройки GoldenGate

РЕШЕНИЕ ЗАДАЧ С ИНСТРУМЕНТАМИ HADOOP (3)

Спецсимволы в наименовании полей баз данных
Q: Названия полей в формате AVRO должны быть допустимыми именами Java
A: Доработка Sqoop v 1.4.7 (Sqoop-3123)



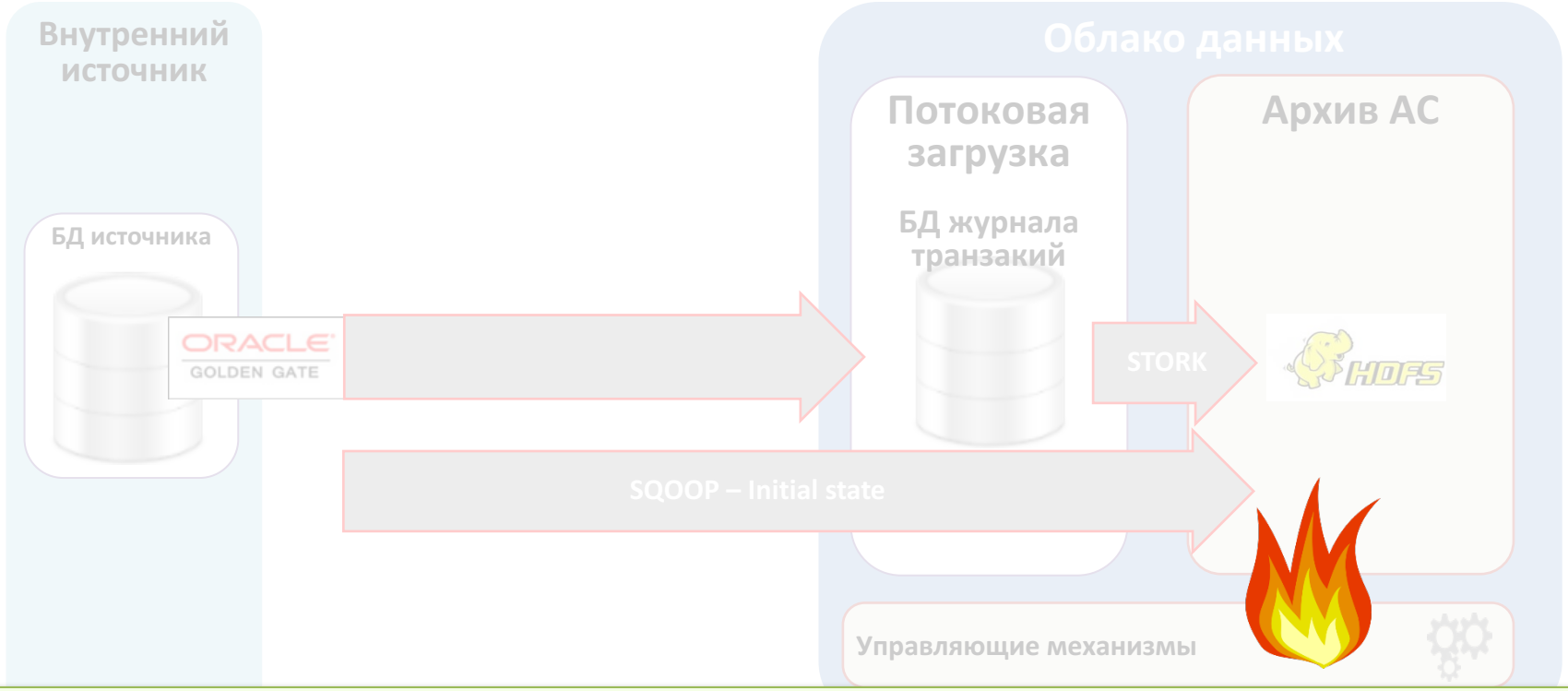
Скорость загрузки данных из Oracle
Q: Generic JDBC - 4 MB/s
A: OraOOP - 750 MB/s

Q: OraOOP не работает со спецсимволами.
W: Создание View – не работает OraOOP
A: Переименование в базе через настройки GoldenGate

Объекты типа LOB
Q: При наличии в таблице полей типа LOB-ов в Sqoop не работает доступ по ROWID.
A: Решение:

- Доработка Sqoop, open-source
- UDF для Hive.

РЕШЕНИЕ ЗАДАЧ С ИНСТРУМЕНТАМИ HADOOP (4)



Стабильность работы Oozie

Q: Проблемы координации потоков – все загрузки стартуют в одно время, реплицируются десятки источников и десятки тысячи таблиц одновременно. Есть вопросы к производительности Oozie.

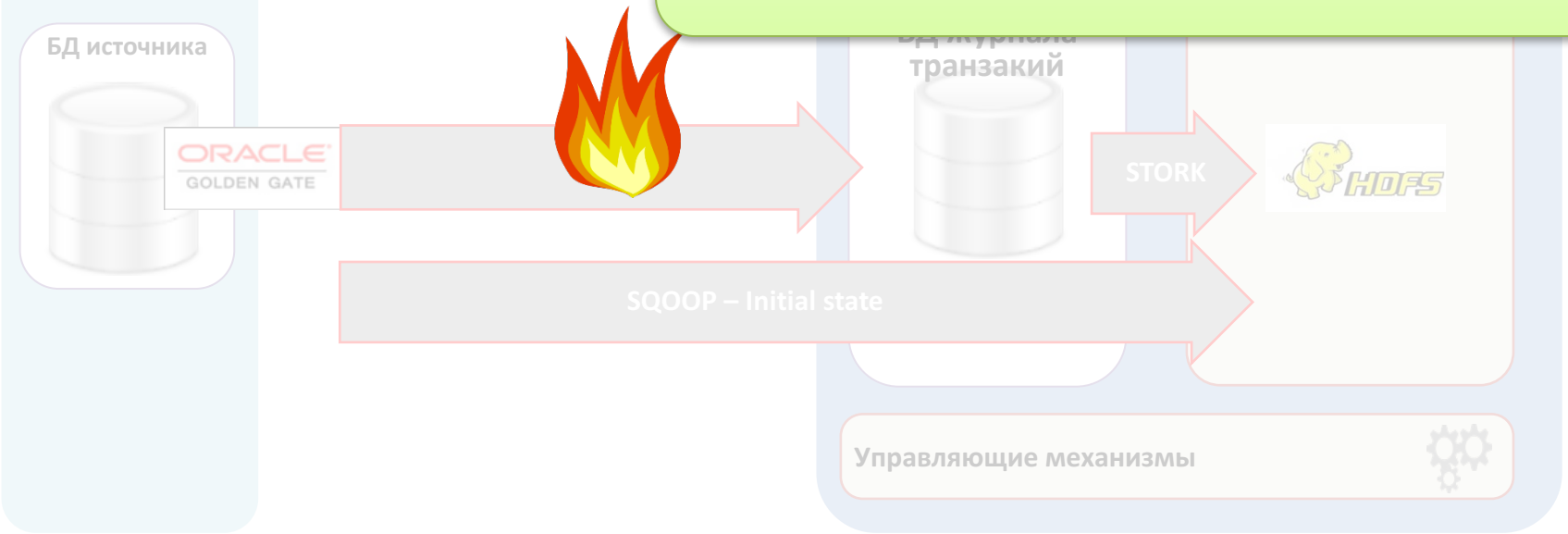
РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (2)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

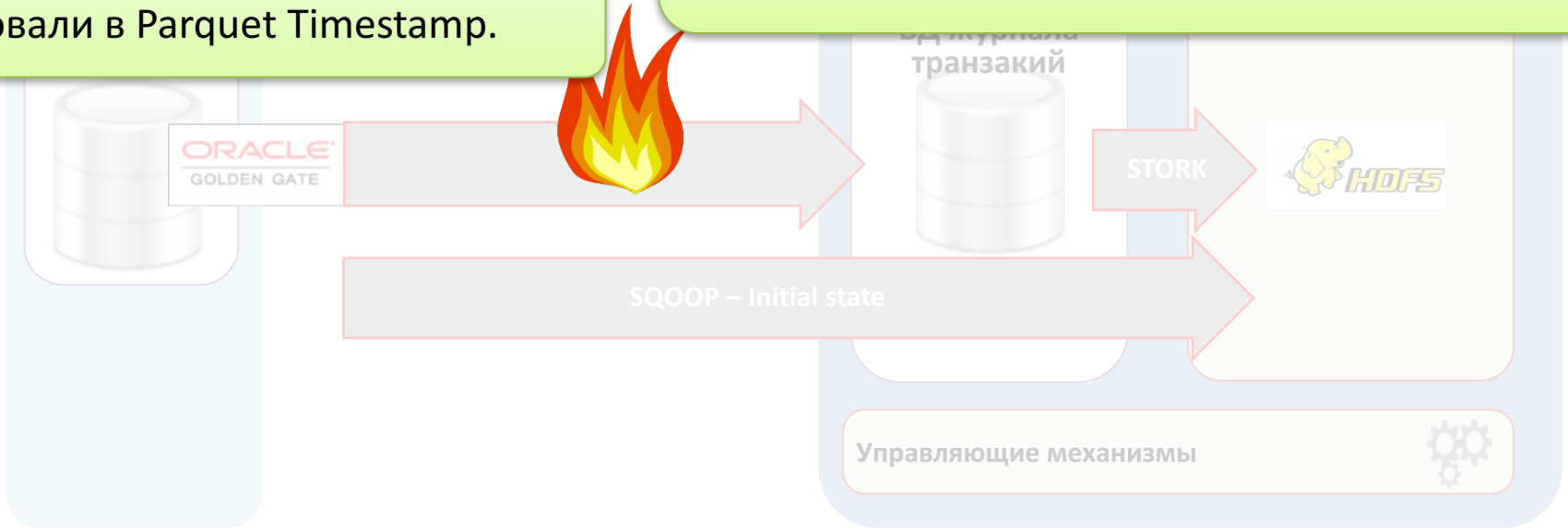


РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (3)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).



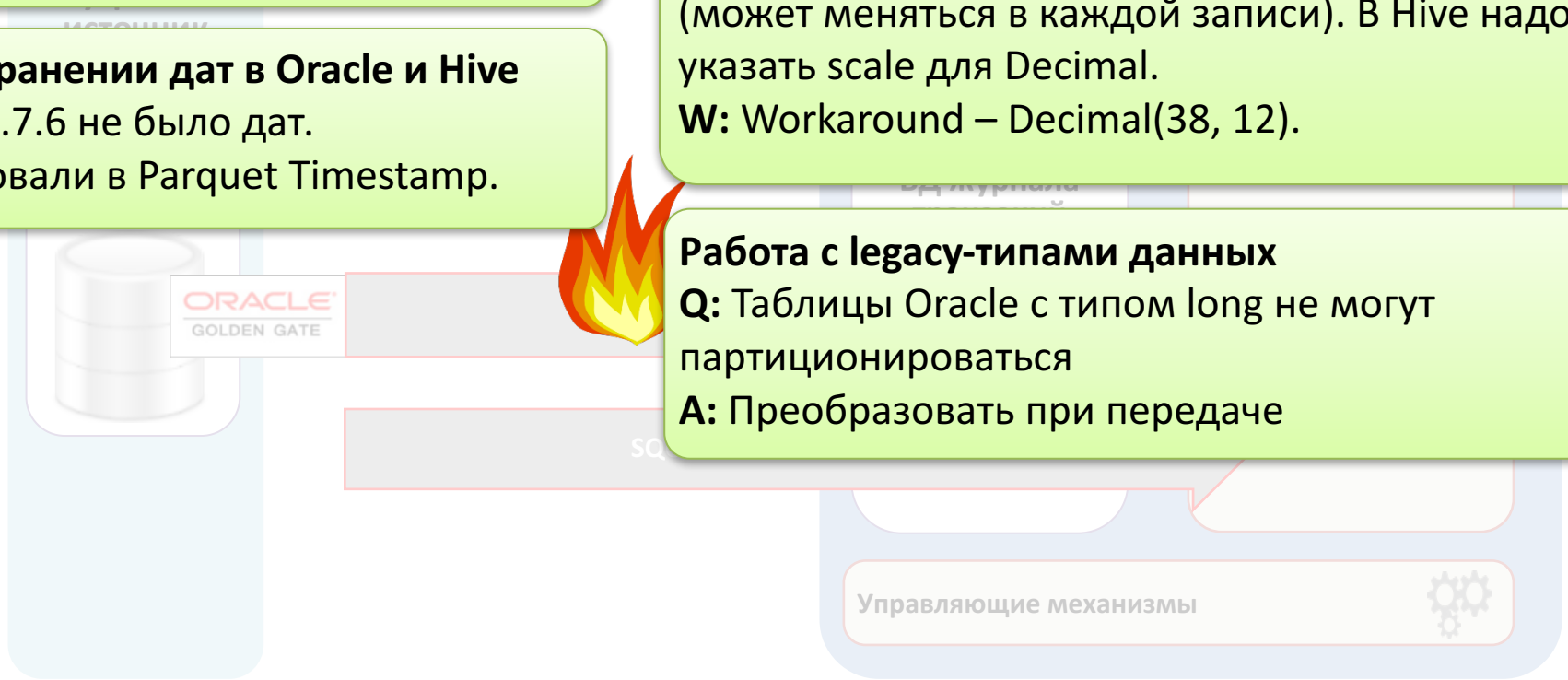
РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (4)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

Работа с legacy-типами данных
Q: Таблицы Oracle с типом long не могут партиционироваться
A: Преобразовать при передаче



РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (5)

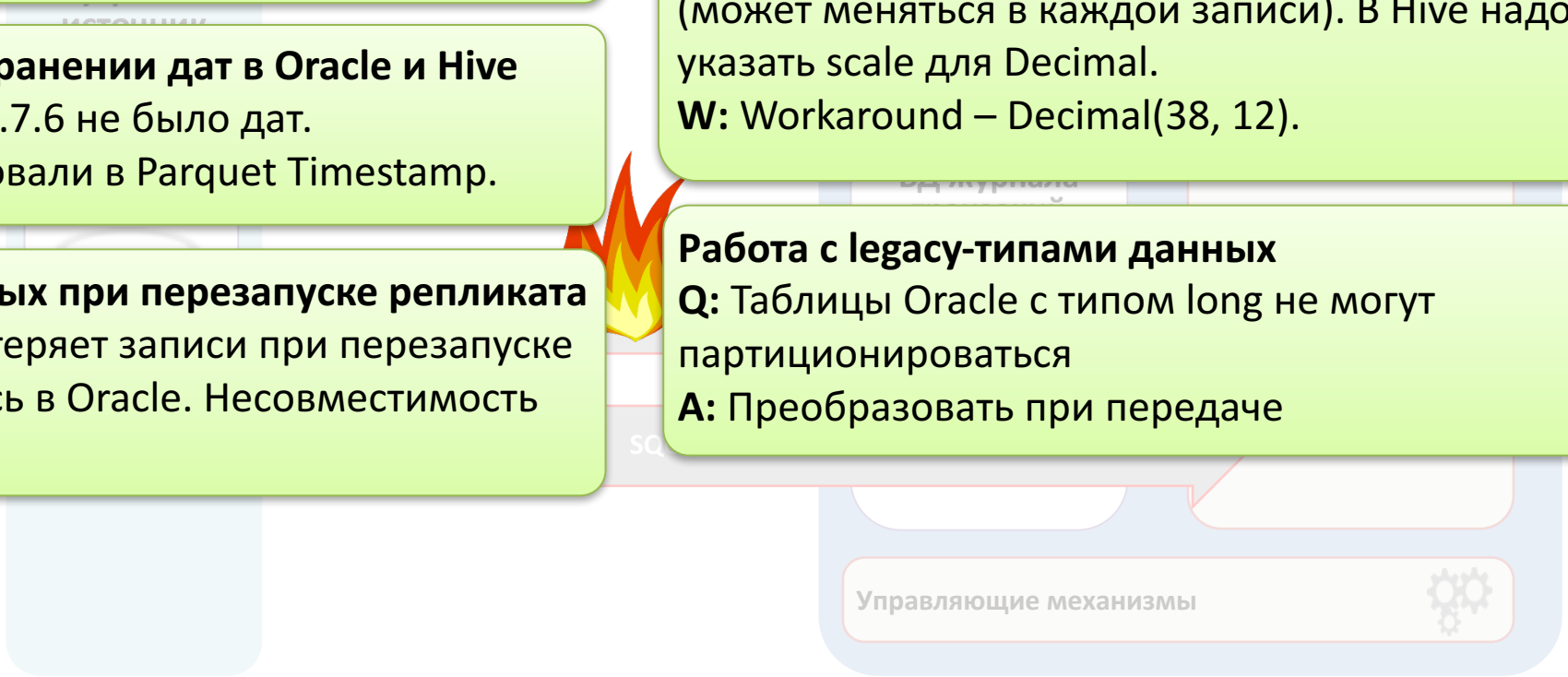
Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Потери данных при перезапуске репликата
Q: Репликат теряет записи при перезапуске
A: Обратились в Oracle. Несовместимость версий?

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

Работа с legacy-типами данных
Q: Таблицы Oracle с типом long не могут партиционироваться
A: Преобразовать при передаче



РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (6)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Потери данных при перезапуске репликата
Q: Репликат теряет записи при перезапуске
A: Обратились в Oracle. Несовместимость версий?

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

Работа с legacy-типами данных
Q: Таблицы Oracle с типом long не могут партиционироваться
A: Преобразовать при передаче

Изменение порядка операций в логе транзакций
Q: NoTCPSourceTimer – опция, которая отключает добавление ко времени операции флюктуации. Настройки OGG по умолчанию добавляют сетевой лаг.
A: Опцию отключить. Сортировать по CSN+SeqNo

РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (7)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Потери данных при перезапуске репликата
Q: Репликат теряет записи при перезапуске
A: Обратились в Oracle. Несовместимость версий?

Партиционирование данных
Q: Переход между границами суток – разные партиции DeltaDB . Получение операции из следующих суток – чтение партиции
A: Решение - разбили по часам.

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

Работа с legacy-типами данных
Q: Таблицы Oracle с типом long не могут партиционироваться
A: Преобразовать при передаче

Изменение порядка операций в логе транзакций
Q: NoTCPSourceTimer – опция, которая отключает добавление ко времени операции флюктуации. Настройки OGG по умолчанию добавляют сетевой лаг.
A: Опцию отключить. Сортировать по CSN+SeqNo

РЕШЕНИЕ ЗАДАЧ ИНТЕГРАЦИИ С ORACLE (8)

Особенности работы журнала транзакций
Q: UPDATE после DELETE
A: Учесть при обработке журнала

Различия в хранении дат в Oracle и Hive
Q: До AVRO 1.7.6 не было дат.
A: Преобразовали в Parquet Timestamp.

Потери данных при перезапуске репликата
Q: Репликат теряет записи при перезапуске
A: Обратились в Oracle. Несовместимость версий?

Партиционирование данных
Q: Переход между границами суток – разные партиции DeltaDB . Получение операции из следующих суток – чтение партиции
A: Решение - разбили по часам.

Различия в числовых типах данных Oracle и Hive.
Q: Потеря точности NUMBER.Oracle тип NUMBER содержит до 38 разрядов (precision), scale можно не указывать (может меняться в каждой записи). В Hive надо заранее указать scale для Decimal.
W: Workaround – Decimal(38, 12).

Работа с legacy-типами данных
Q: Таблицы Oracle с типом long не могут партиционироваться
A: Преобразовать при передаче

Изменение порядка операций в логе транзакций
Q: NoTCPSourceTimer – опция, которая отключает добавление ко времени операции флюктуации. Настройки OGG по умолчанию добавляют сетевой лаг.
A: Опцию отключить. Сортировать по CSN+SeqNo

Операции Truncate, UPDATE первичного ключа, DELETE * и вставка

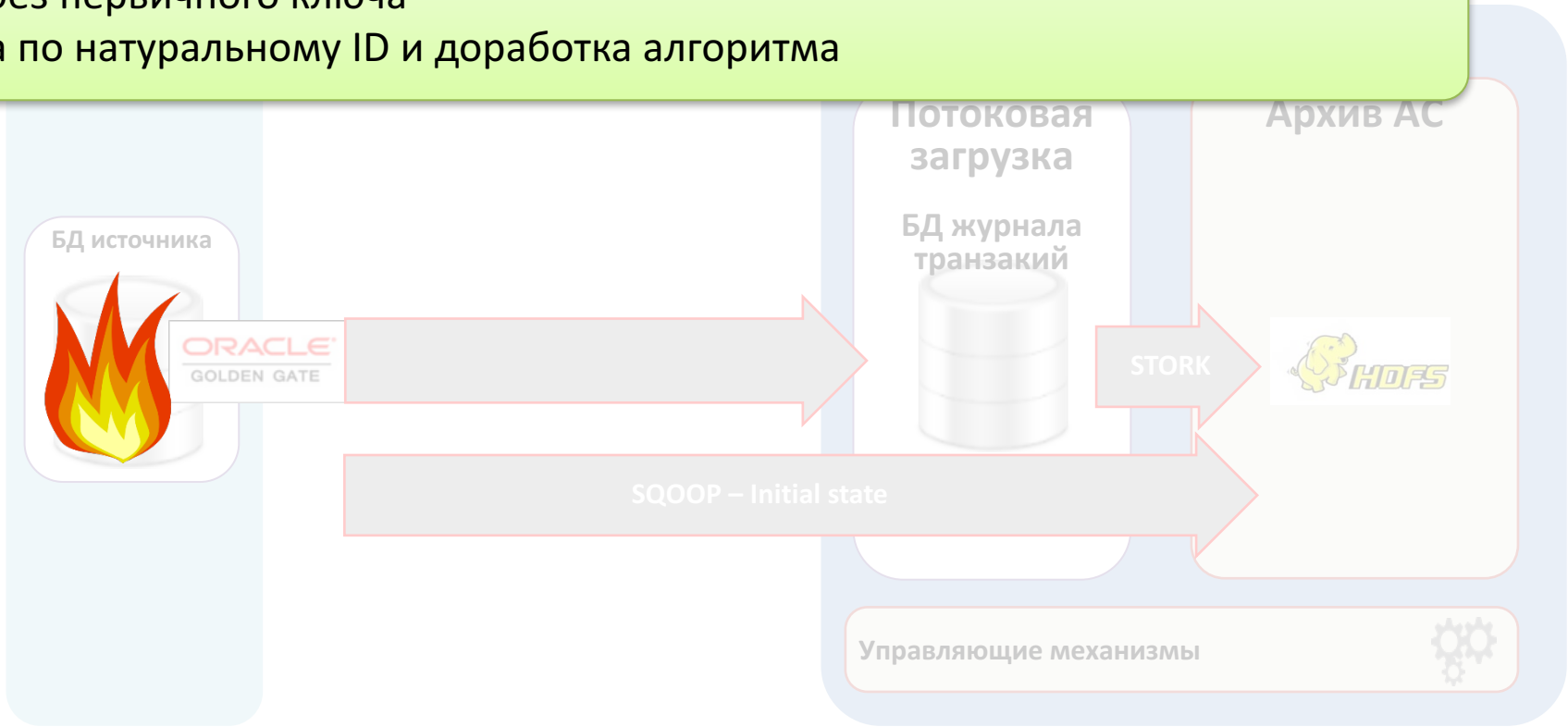
РЕШЕНИЕ ЗАДАЧ С ОСОБЕННОСТЯМИ ИНФОРМАЦИОННЫХ СИСТЕМ БАНКА

Структура данных в legacy-системах

Q: Составной первичный ключ, до 9-ти атрибутов в ключе

Q: Таблицы без первичного ключа

A: Аналитика по натуральному ID и доработка алгоритма



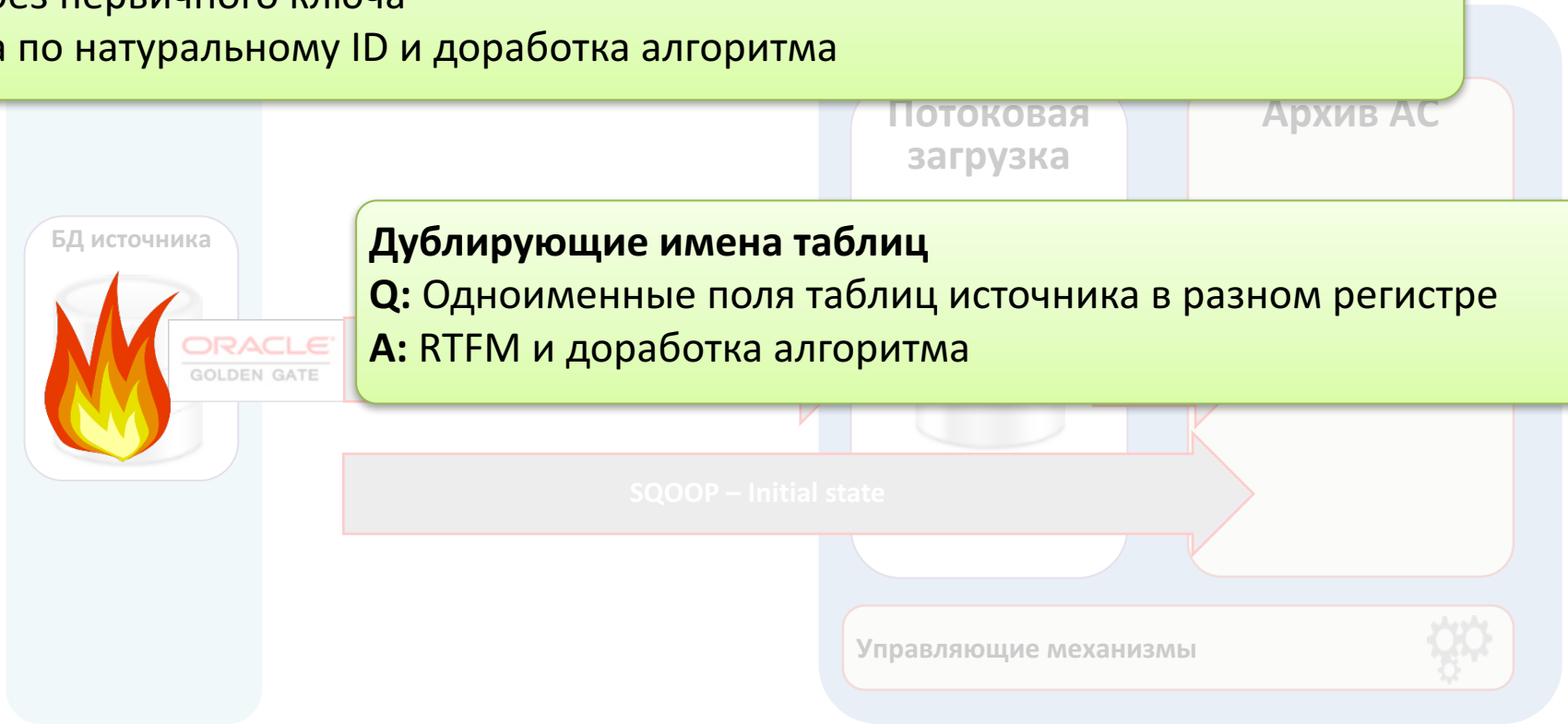
РЕШЕНИЕ ЗАДАЧ С ОСОБЕННОСТЯМИ ИНФОРМАЦИОННЫХ СИСТЕМ БАНКА (2)

Структура данных в legacy-системах

Q: Составной первичный ключ, до 9-ти атрибутов в ключе

Q: Таблицы без первичного ключа

A: Аналитика по натуральному ID и доработка алгоритма



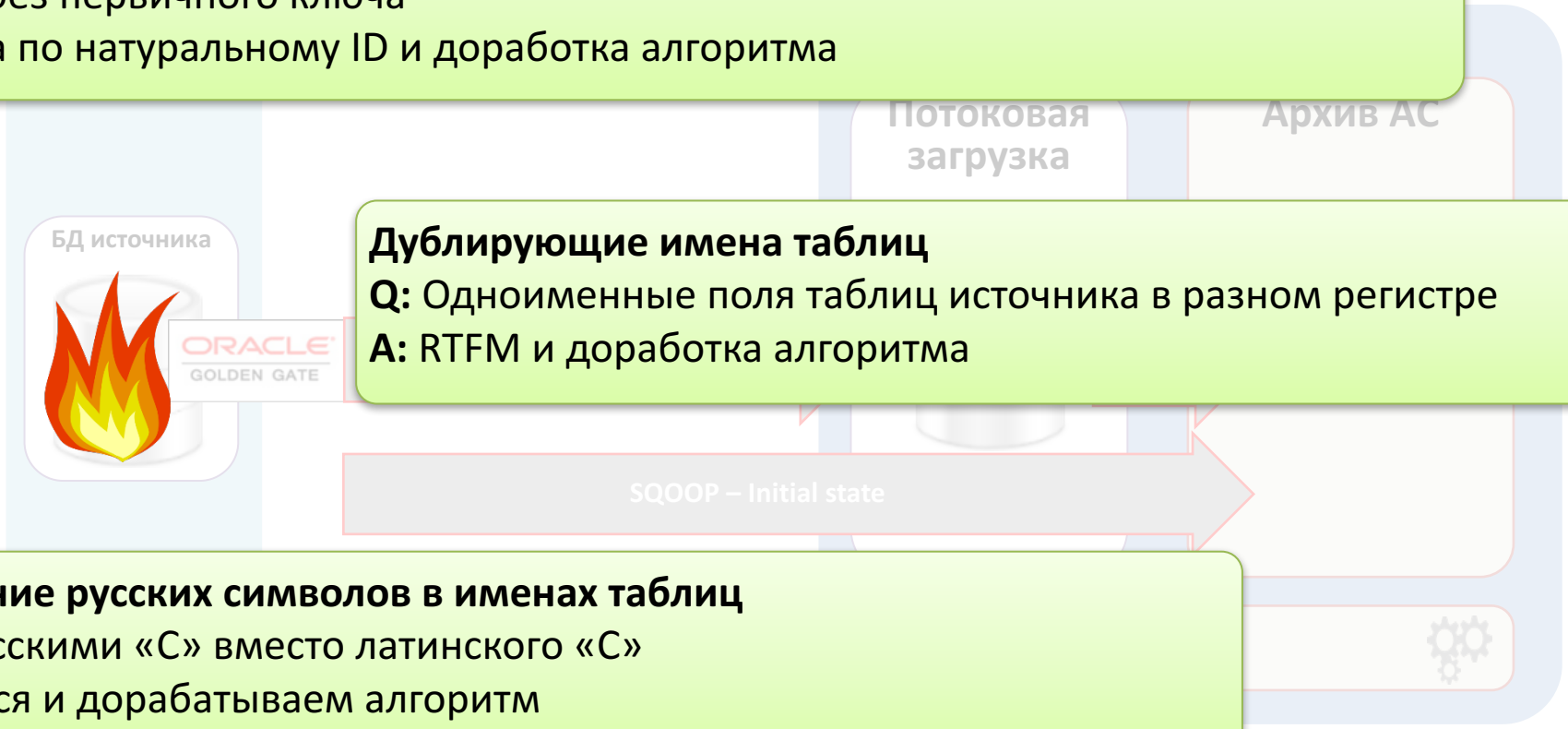
Дублирующие имена таблиц

Q: Одноименные поля таблиц источника в разном регистре

A: RTFM и доработка алгоритма

РЕШЕНИЕ ЗАДАЧ С ОСОБЕННОСТЯМИ ИНФОРМАЦИОННЫХ СИСТЕМ БАНКА (3)

Структура данных в legacy-системах
Q: Составной первичный ключ, до 9-ти атрибутов в ключе
Q: Таблицы без первичного ключа
A: Аналитика по натуральному ID и доработка алгоритма

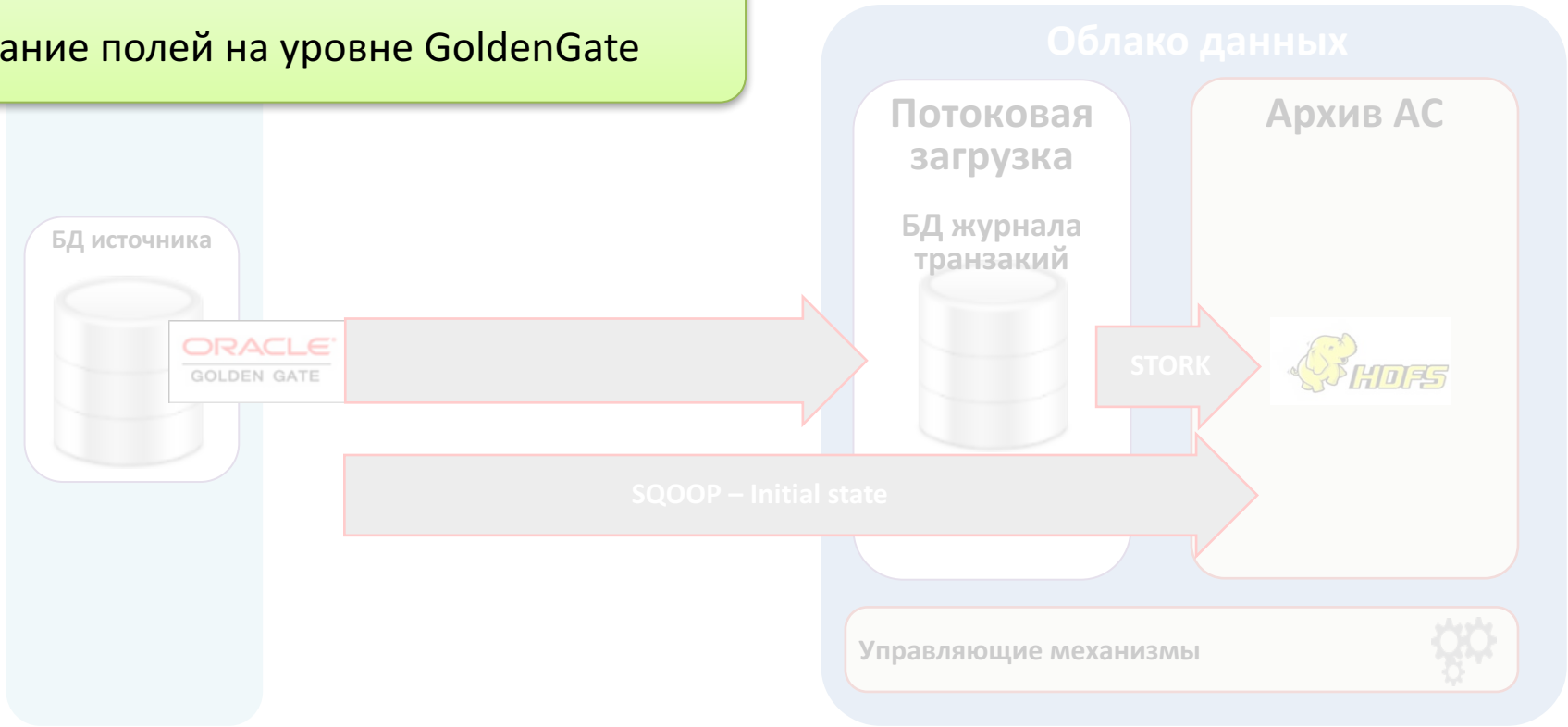


Дублирующие имена таблиц
Q: Одноименные поля таблиц источника в разном регистре
A: RTFM и доработка алгоритма

Использование русских символов в именах таблиц
Q: Поля с русскими «С» вместо латинского «С»
A: Удивляемся и дорабатываем алгоритм

АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate



АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (2)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

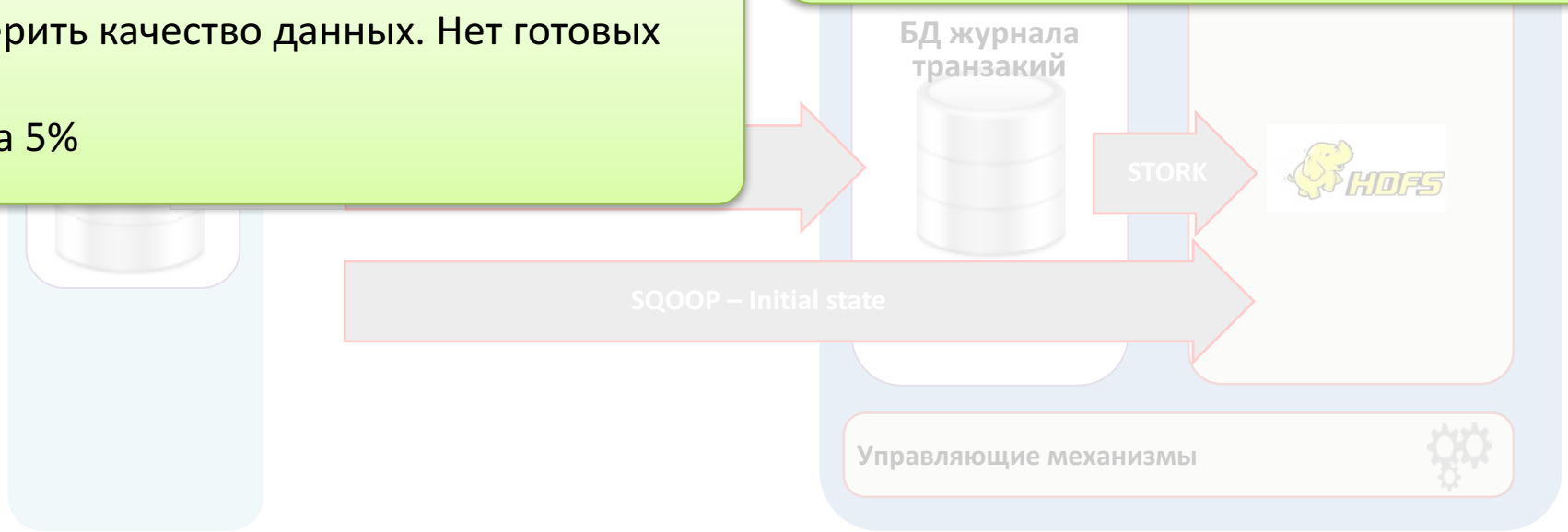
Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (3)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков



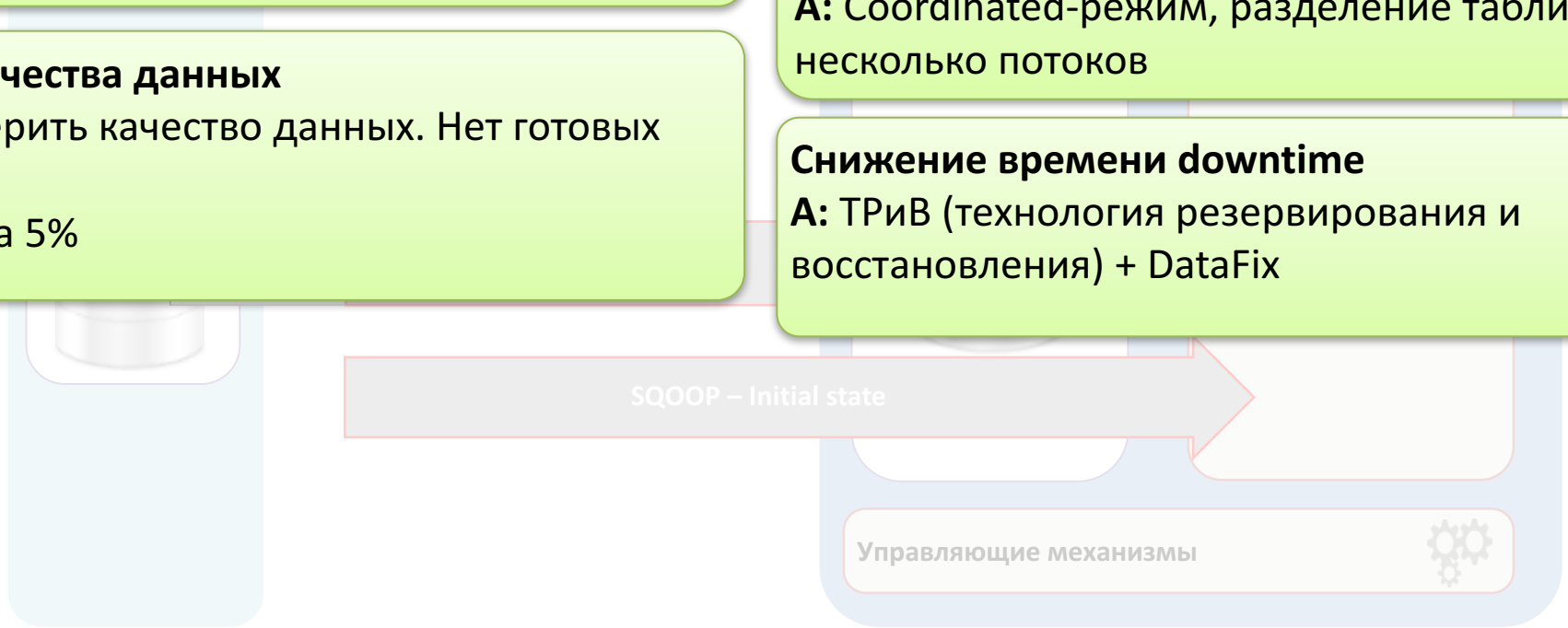
АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (4)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix



АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (5)

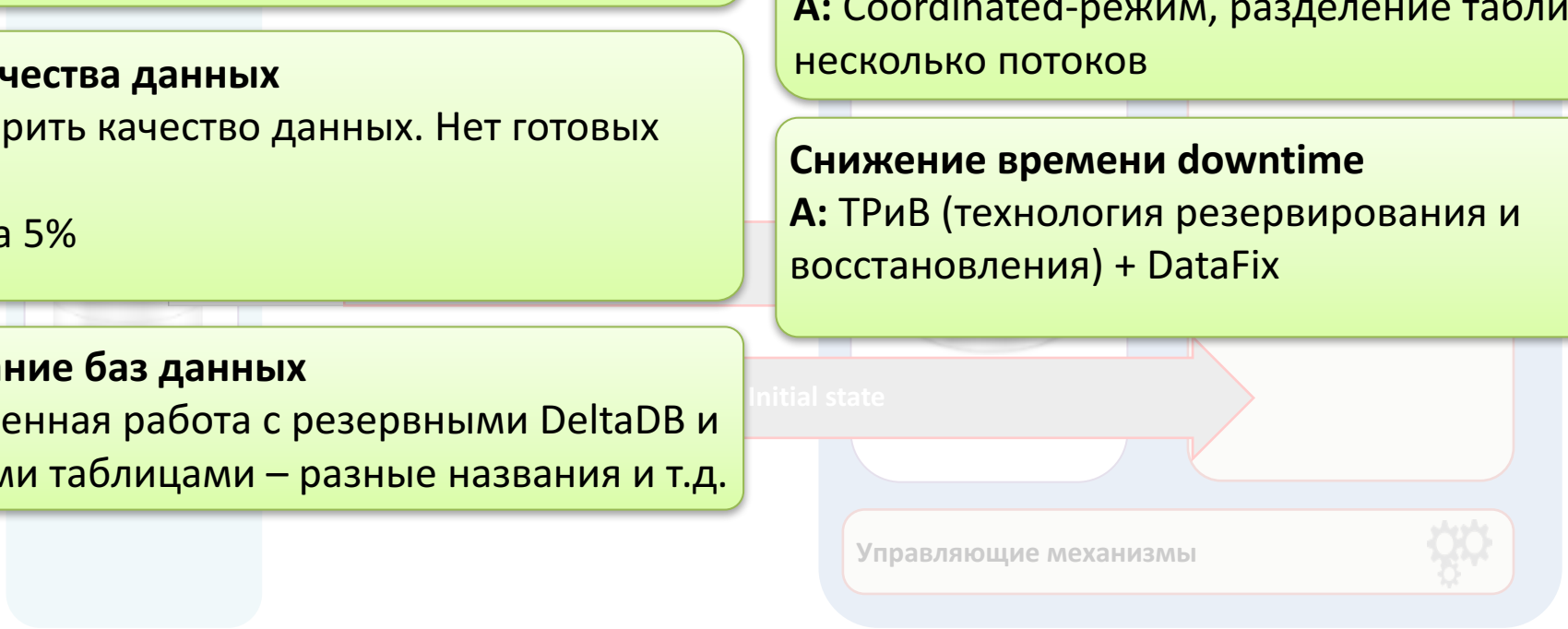
Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Резервирование баз данных
A: Одновременная работа с резервными DeltaDB и разделенными таблицами – разные названия и т.д.

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix



АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (6)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Резервирование баз данных
A: Одновременная работа с резервными DeltaDB и разделенными таблицами – разные названия и т.д.

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix

Обеспечение безопасности
A: Запись, чтение HDFS – Kerberos

АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (7)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Резервирование баз данных
A: Одновременная работа с резервными DeltaDB и разделенными таблицами – разные названия и т.д.

Мониторинг работы репликатора
A: Журналирование, Zabbix, ELK

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix

Обеспечение безопасности
A: Запись, чтение HDFS – Kerberos

АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (8)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Резервирование баз данных
A: Одновременная работа с резервными DeltaDB и разделенными таблицами – разные названия и т.д.

Мониторинг работы репликатора
A: Журналирование, Zabbix, ELK

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix

Обеспечение безопасности
A: Запись, чтение HDFS – Kerberos

Хранение метаданных и тиражирование
A: Генератор конфигураций для init load и apply.

АРХИТЕКТУРНЫЕ И ОРГАНИЗАЦИОННЫЕ ЗАДАЧИ (9)

Обработка конфиденциальной информации
Q: Передача номеров карт между системами запрещена
A: Маскирование полей на уровне GoldenGate

Проверка качества данных
Q: Как проверить качество данных. Нет готовых решений.
A: Статистика 5%

Резервирование баз данных
A: Одновременная работа с резервными DeltaDB и разделенными таблицами – разные названия и т.д.

Мониторинг работы репликатора
A: Журналирование, Zabbix, ELK

Синхронизация потоков при работе в распределённом режиме
Q: Обработка больших (до 2 ТВ/час) таблиц не укладывается в ресурсы
A: Coordinated-режим, разделение таблиц на несколько потоков

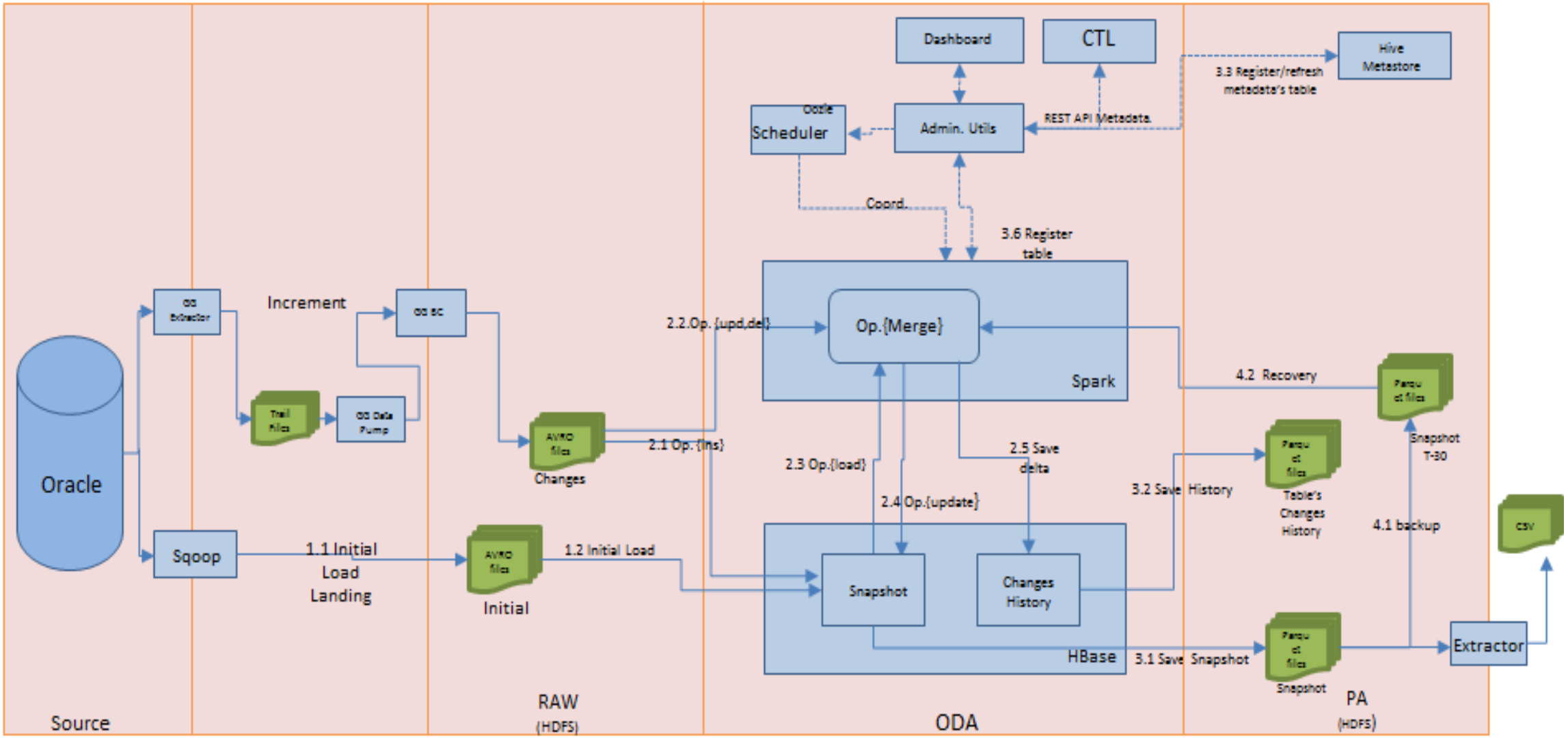
Снижение времени downtime
A: ТРиВ (технология резервирования и восстановления) + DataFix

Обеспечение безопасности
A: Запись, чтение HDFS – Kerberos

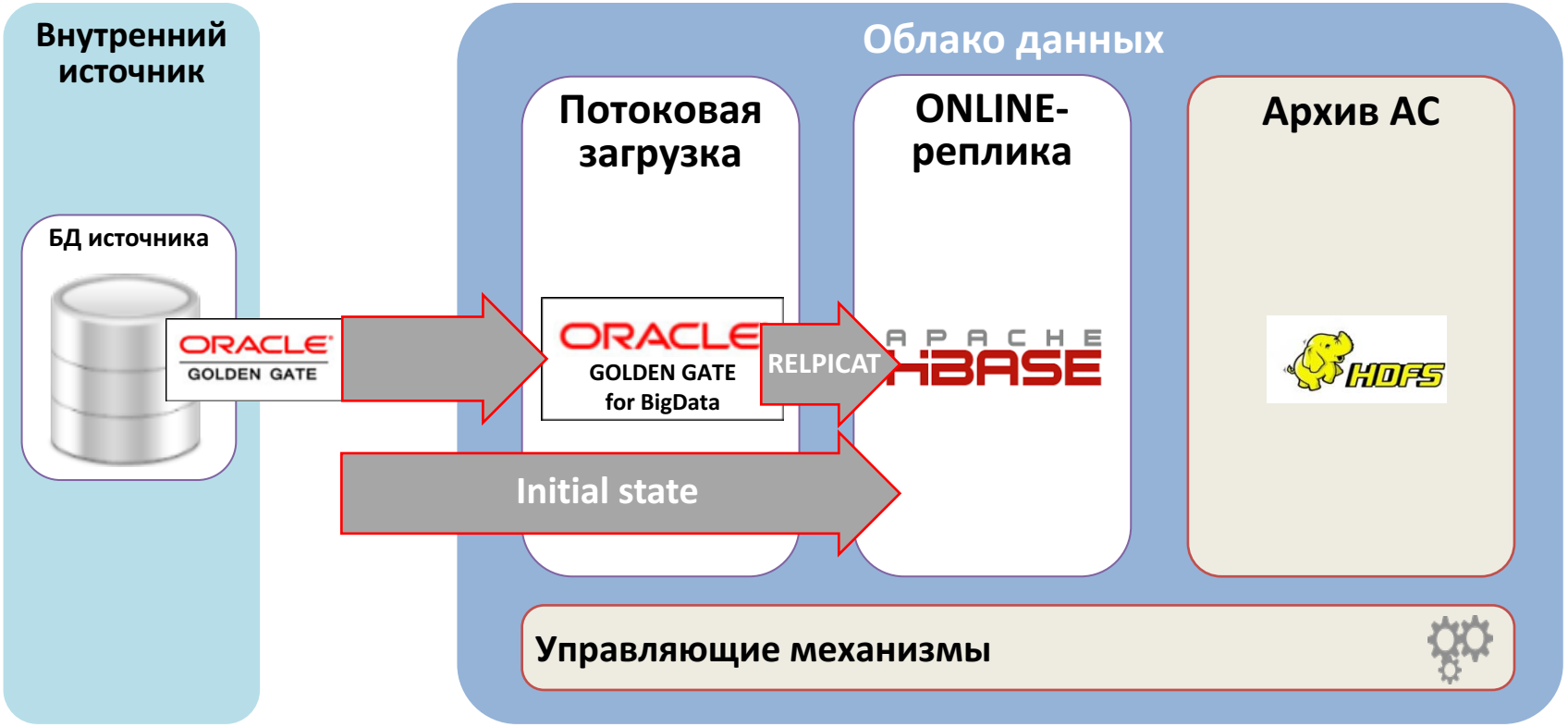
Хранение метаданных и тиражирование
A: Генератор конфигураций для init load и apply.

Поддержка DDL-операций, Секционирование и бакетирование таблиц

АРХИТЕКТУРА РЕШЕНИЯ И «ВСПОМОГАТЕЛЬНЫЕ» ПОДСИСТЕМЫ

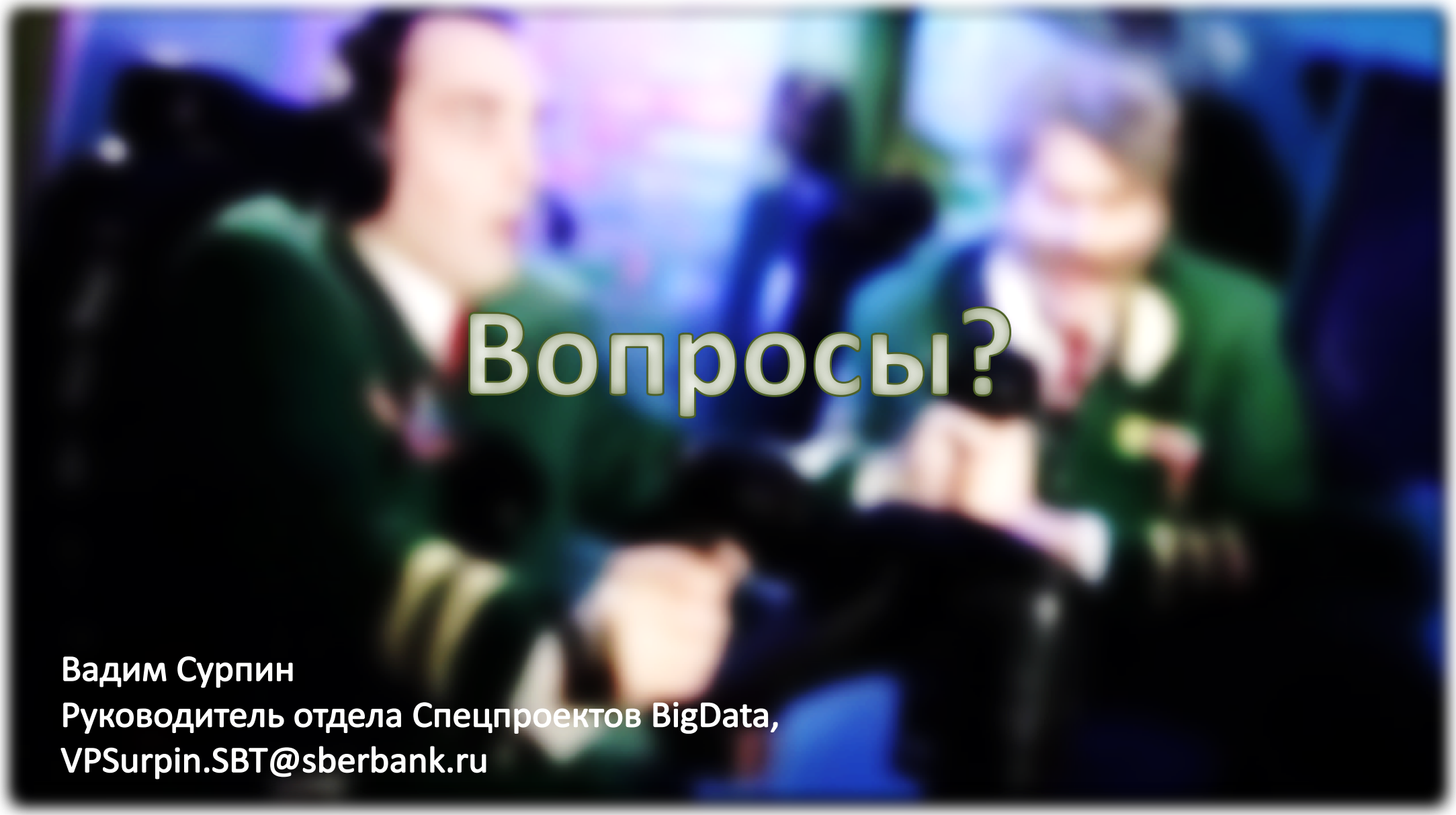


ЧЕТВЁРТЫЙ ПОДХОД



СБЕРБАНК ТЕХНОЛОГИИ

ВОПРОСЫ?



Вадим Сурпин
Руководитель отдела Спецпроектов BigData,
VPSurpin.SBT@sberbank.ru