BI Partner

Аналитические Системы для Бизнеса

Data Vault – новая парадигма хранилищ данных для Big Data

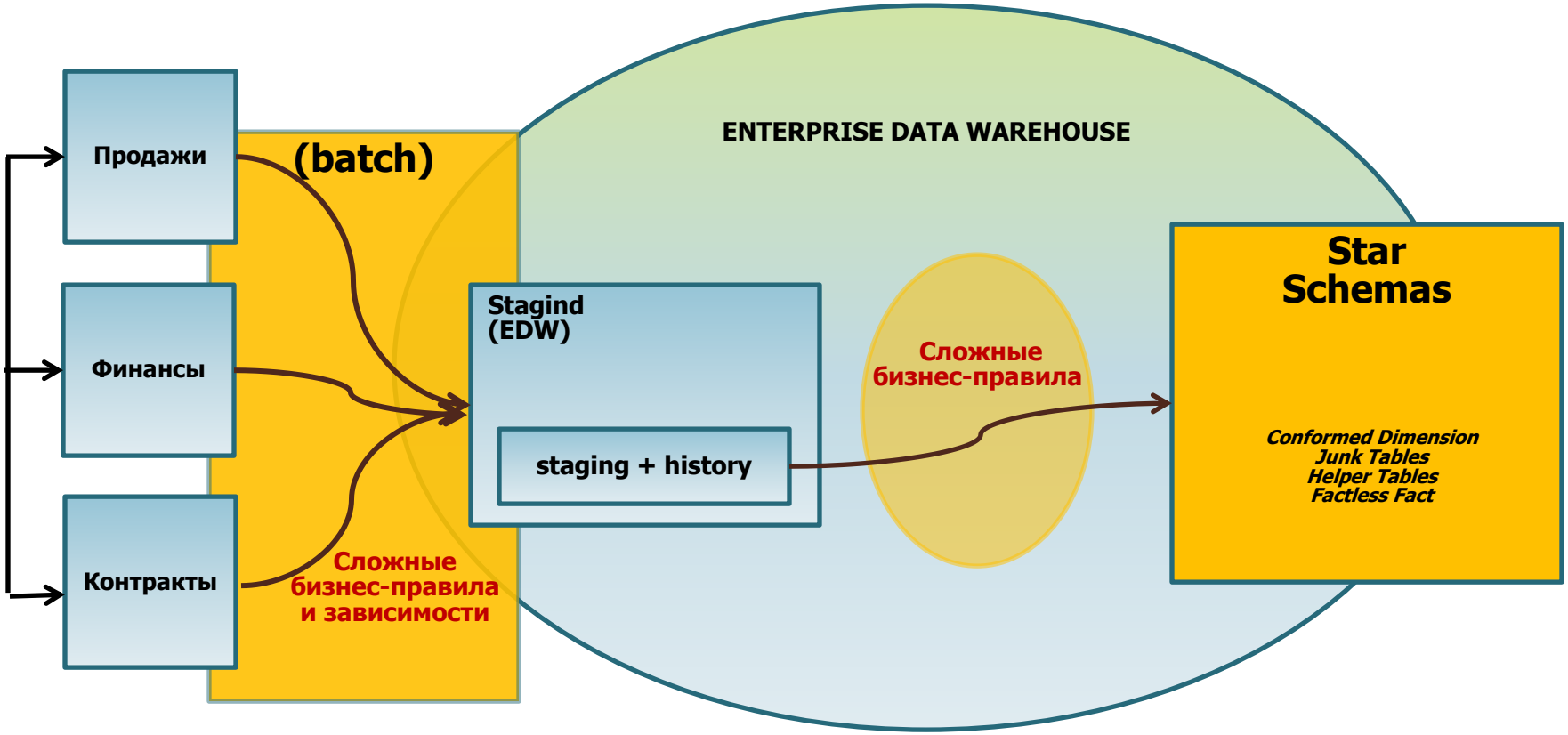
Сергей Сухарев

Руководитель BI-практики

BI Partner

Содержание

Архитектура классических хранилищ данных

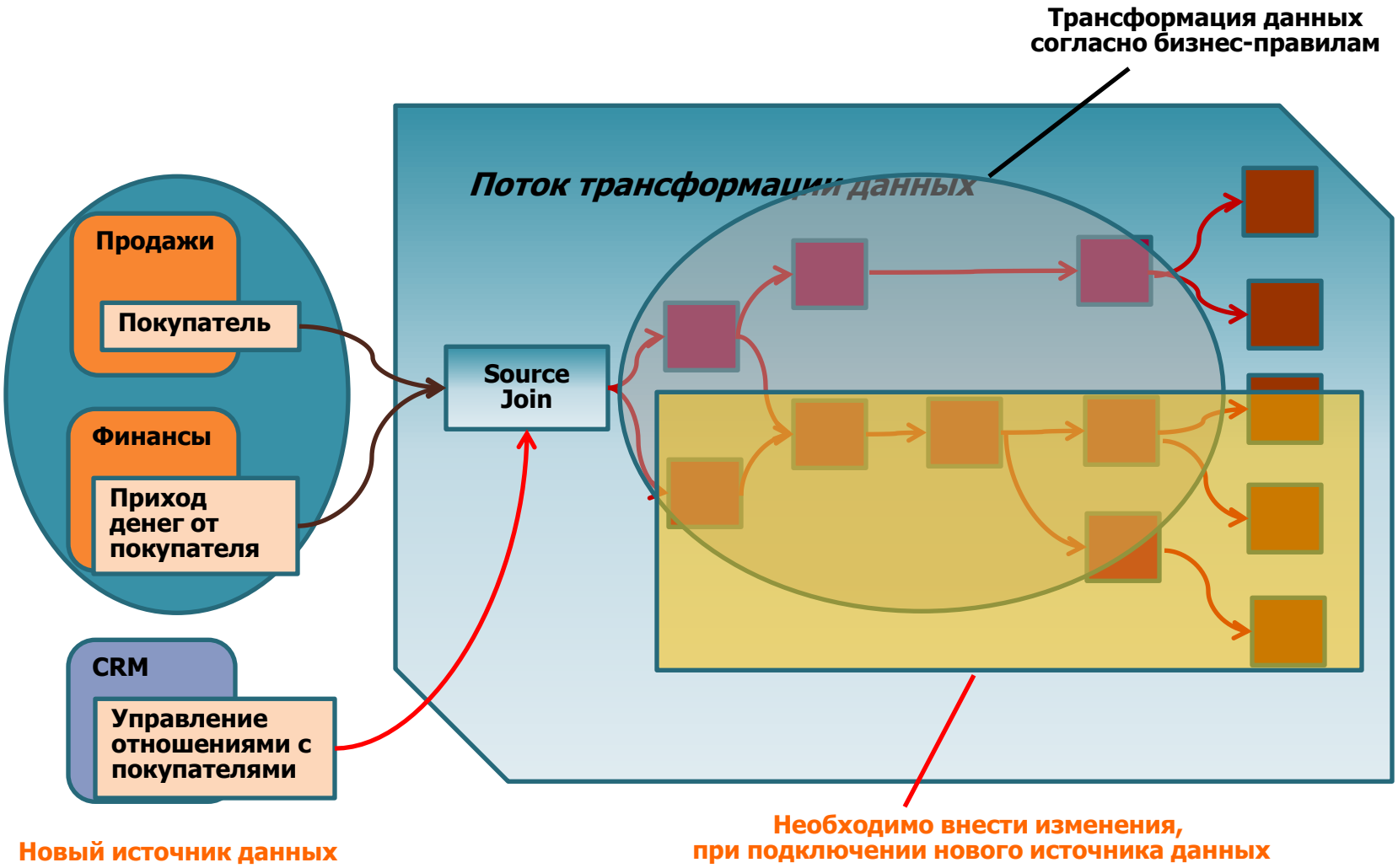


- Зависимость от качества данных
- Не все данные из источника представлены в ХД
- В процессе обработки в данные вносятся изменения

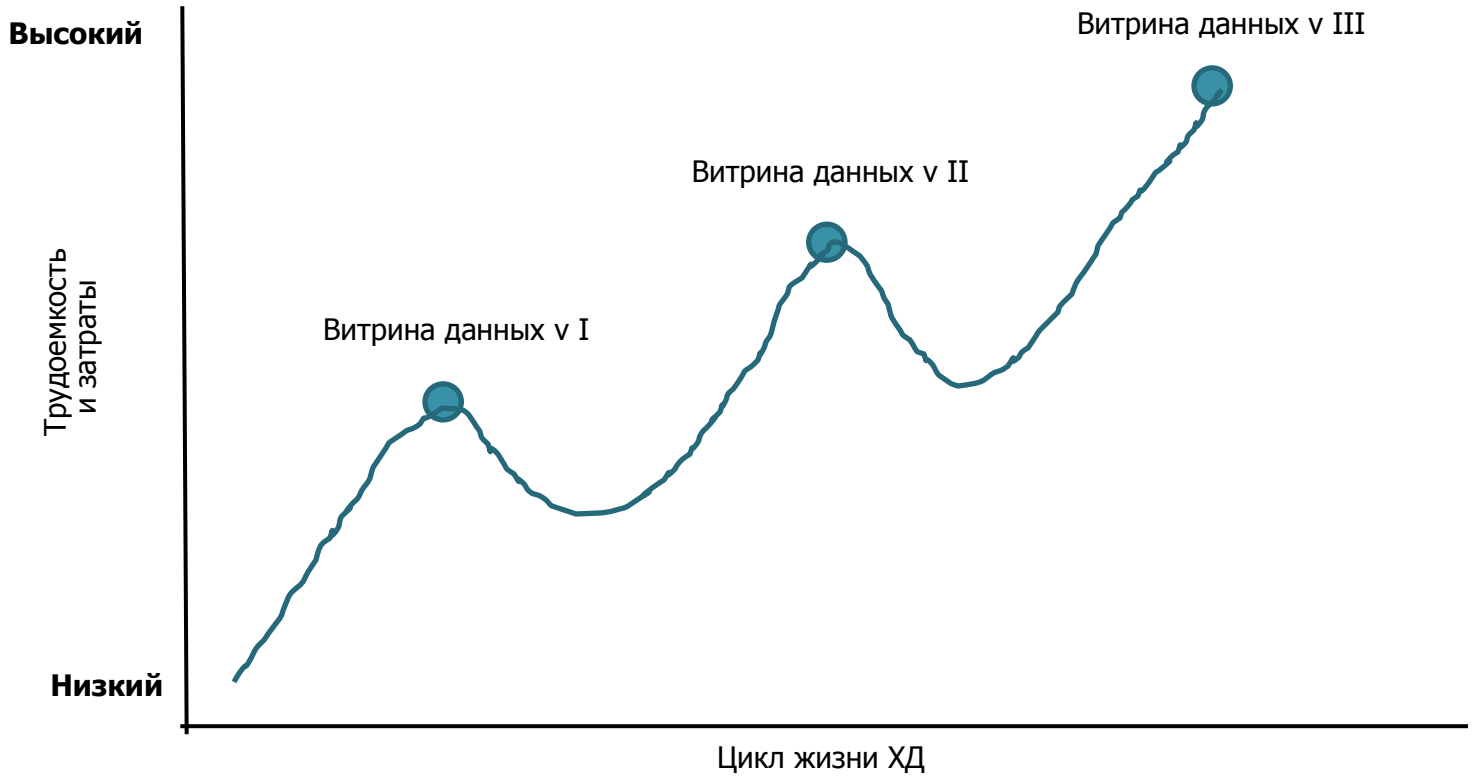
- Высокий риск некорректного расчета показателей
- История изменения первичных данных не может быть восстановлена по данным ХД
- Трудности при проектировании больших систем (не все показатели одинаково трактуются бизнес-пользователями)

BI Partner

Подключение нового источника в классическом ХД



Жизненный цикл классического хранилища данных

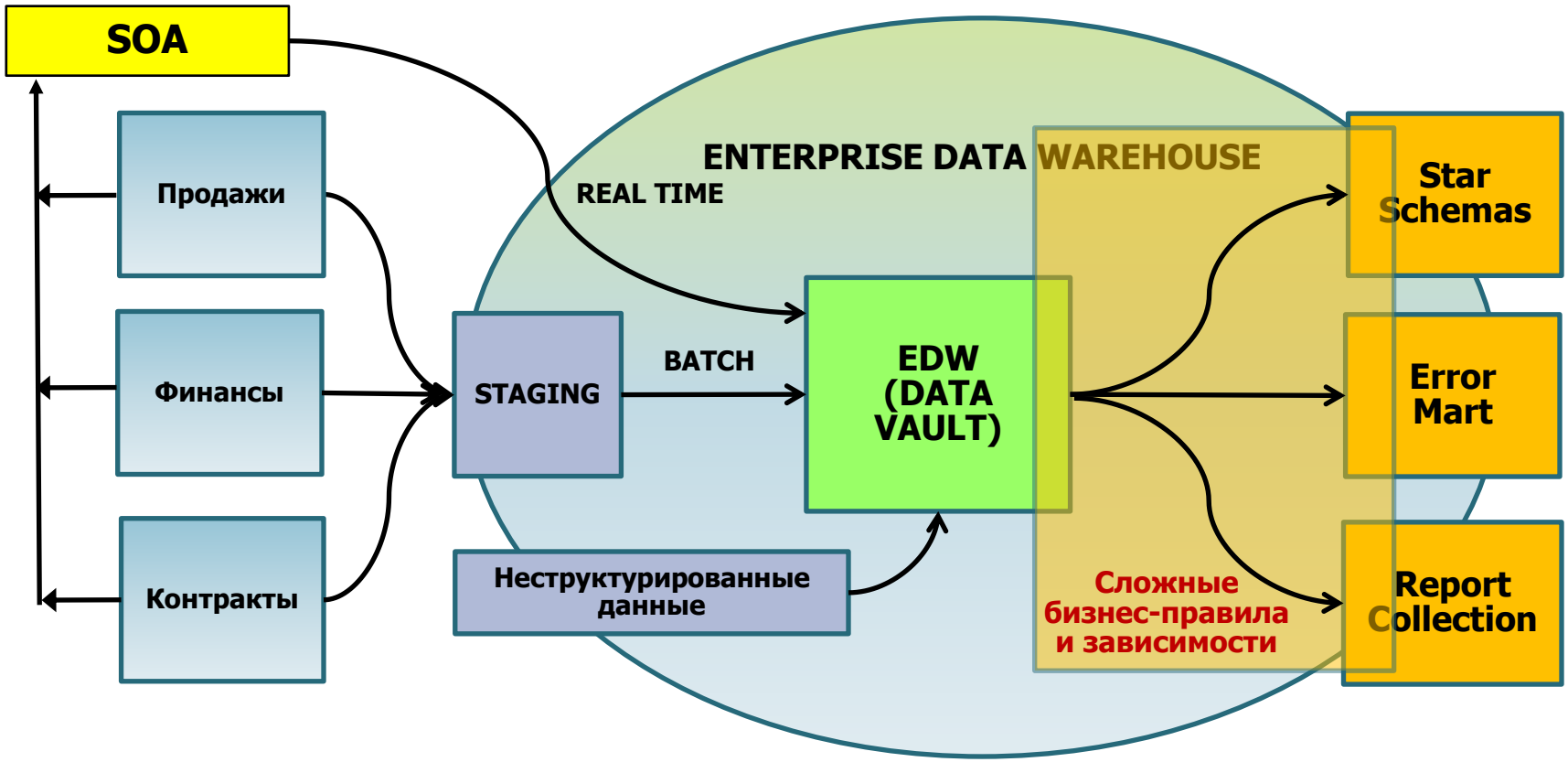


- Изменения и корректировки **согласованных витрин данных** вызывают экспоненциальный рост кривой затрат в течение всего времени жизни хранилища данных
- **Результат: каждое бизнес-подразделение строит собственную витрину данных**

BI Partner

Содержание

Архитектура хранилищ данных 2.0 (DW 2.0)



Основные преимущества

- Гибкость
- Производительность
- Отказоустойчивость
- Масштабируемость
- Легкость в поддержке и развитии
- Легкость аудита данных

Трансформация данных, согласно бизнес-правилам перемещается ближе к бизнесу, это улучшает **время реакции, снижает стоимость и минимизирует последствия** изменений в корпоративном хранилище данных (EDW)

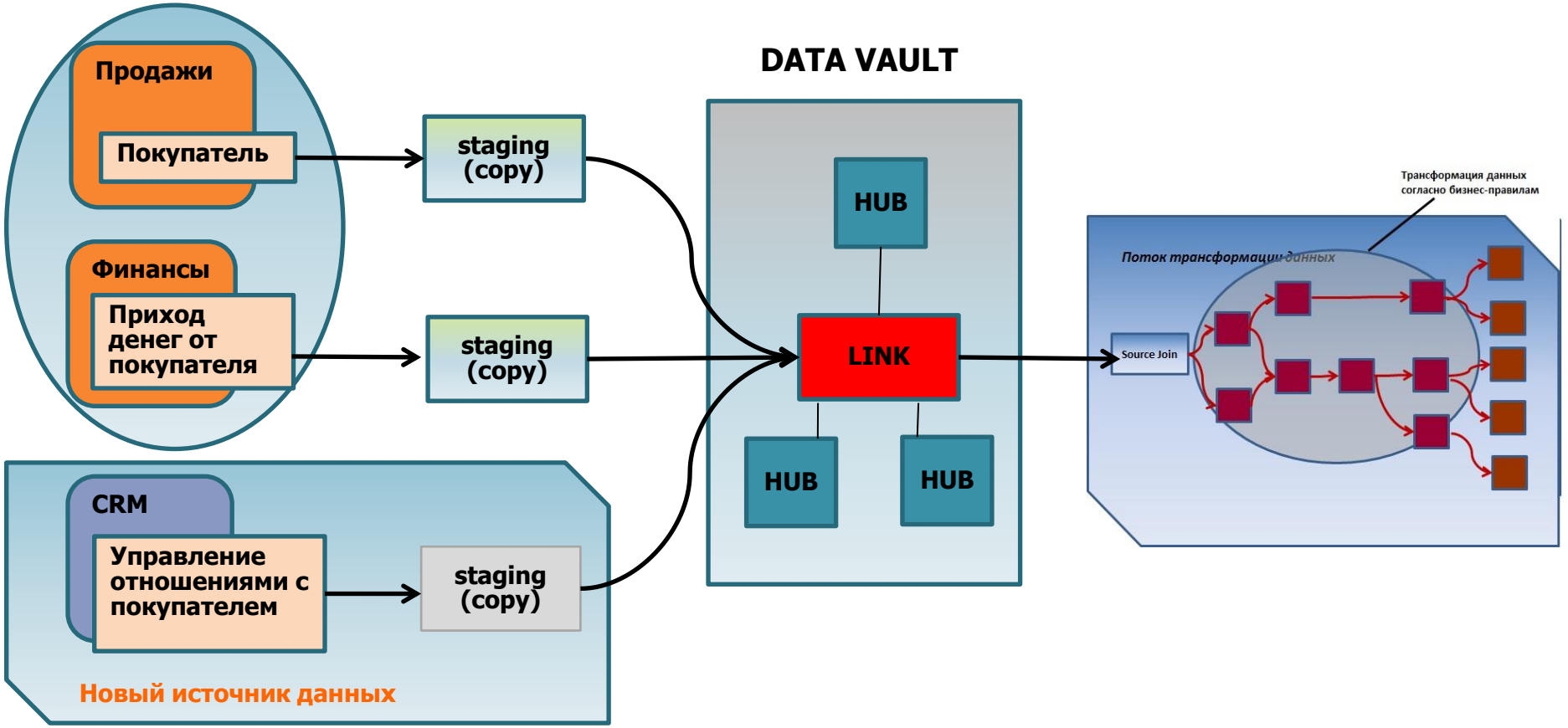
Согласно концепции DW 2.0 хранилища данных разделяются на два слоя:

- Слой хранения данных (собственно хранилище данных)
- Слой представления данных (витрины данных)

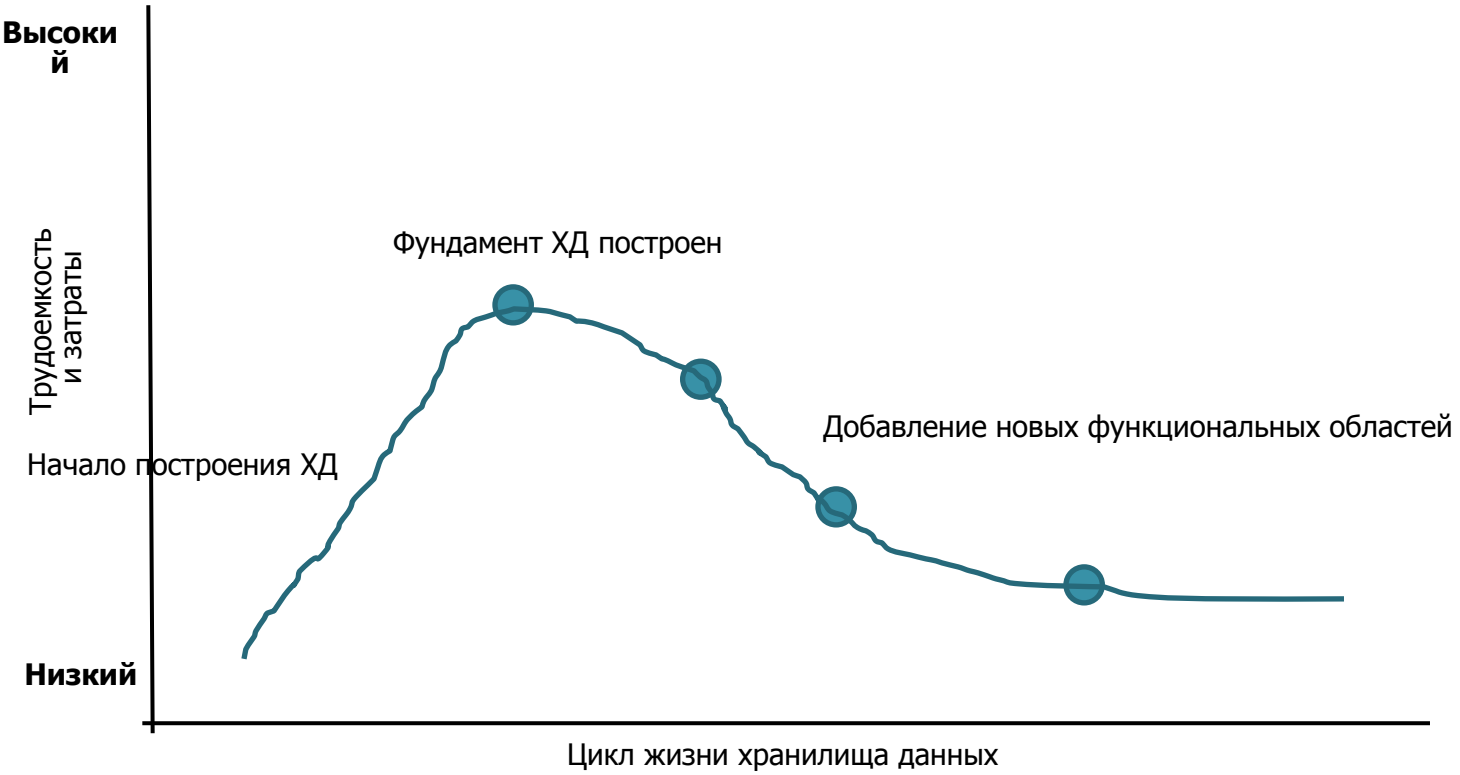
Методология Data Vault соответствует принципам построения DW 2.0

BI Partner

Архитектура Data Vault



Жизненный цикл хранилища данных по Data Vault



В Data Vault **не бывает** реинжиниринга

- Снижает стоимость решения и затраты на обслуживание системы с течением времени
- Упрощает процессы расширения архитектуры системы

BI Partner

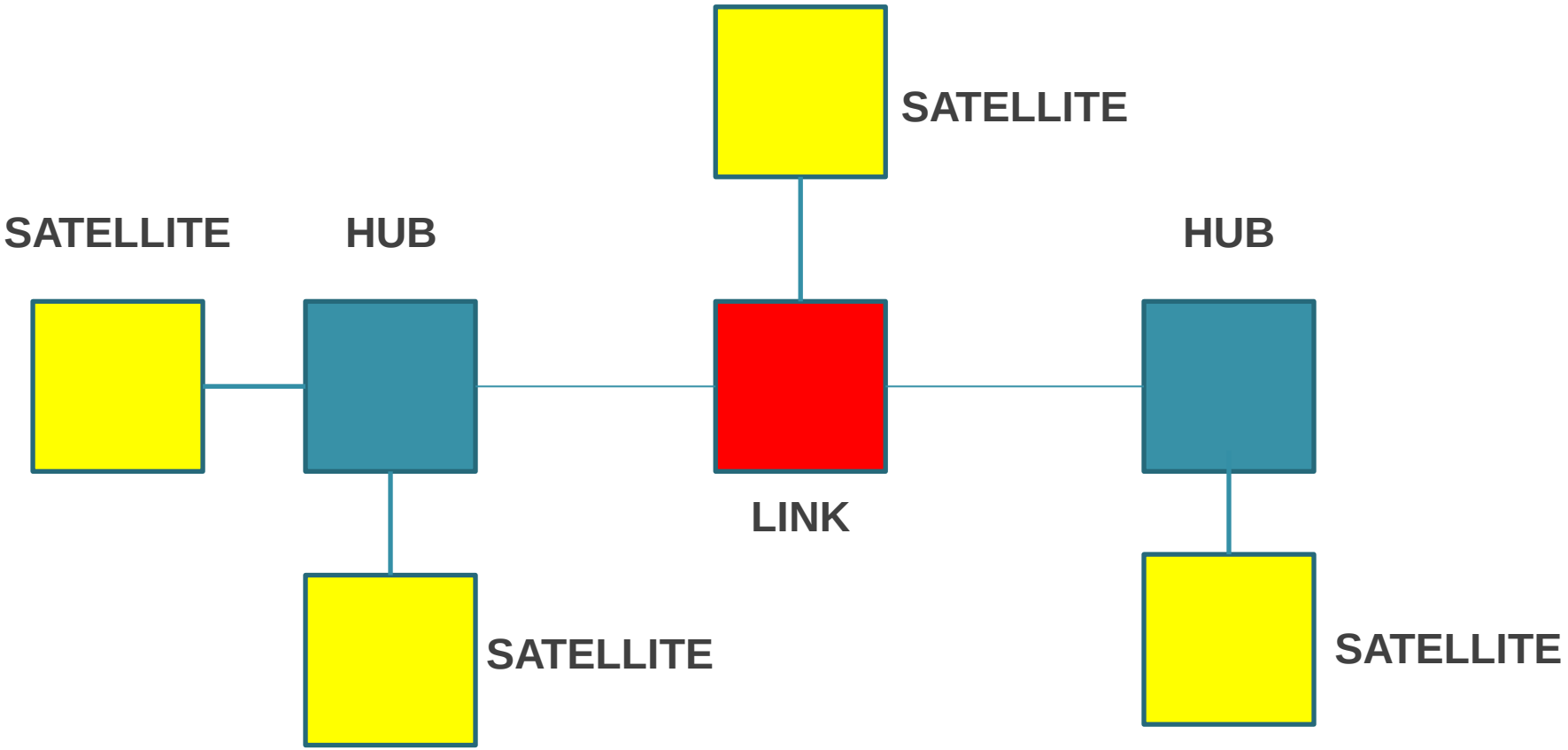
Data Vault. Определение

Data Vault – это детальная, логически связанная, структура, хранящая полную историю изменения данных в источниках. Закрывает одну или несколько бизнес-областей предприятия. Гибридный подход, в котором применяется 3НФ и схема «звезда».

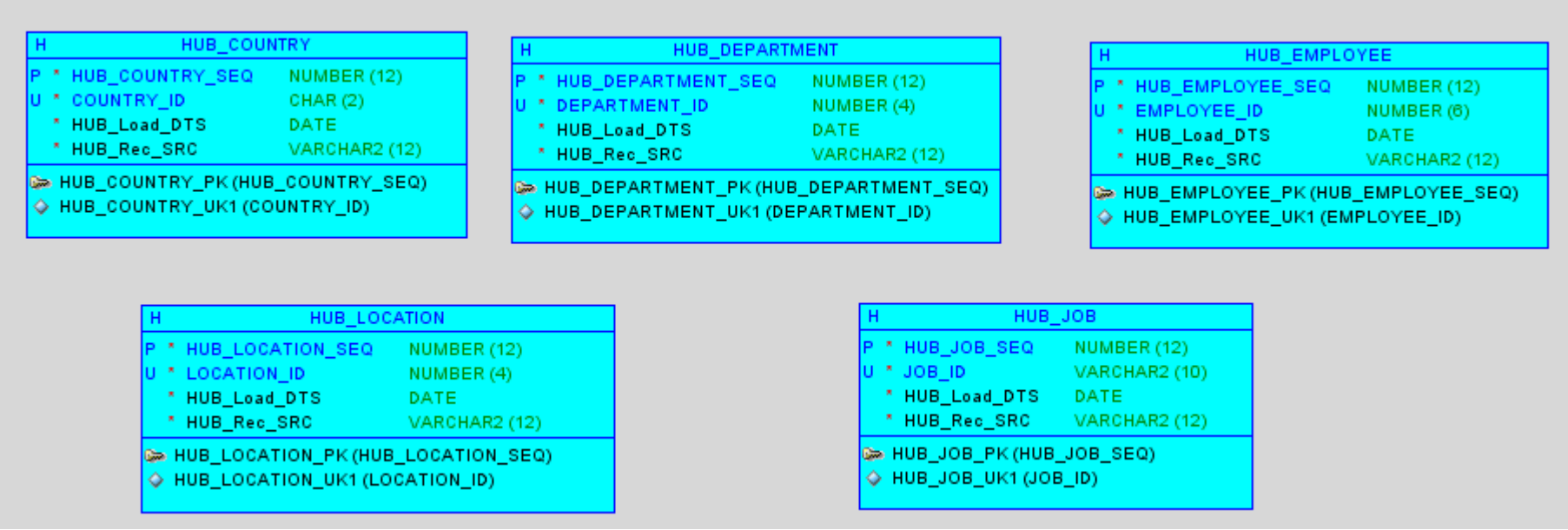
Преимущество при переходе с ХД традиционной структуры (звезда или снежинка) на ХД, организованное по схеме Data Vault:

- **Гибкость.** Быстрое внесение массовых изменений в структуру ХД без переделки предыдущей структуры.
- **Масштабируемость.** Нет ограничений по увеличению размеров и масштабов ХД
- **Производительность.** Позволяет снижать стоимость и сложность процессов загрузки данных

Объекты Data Vault

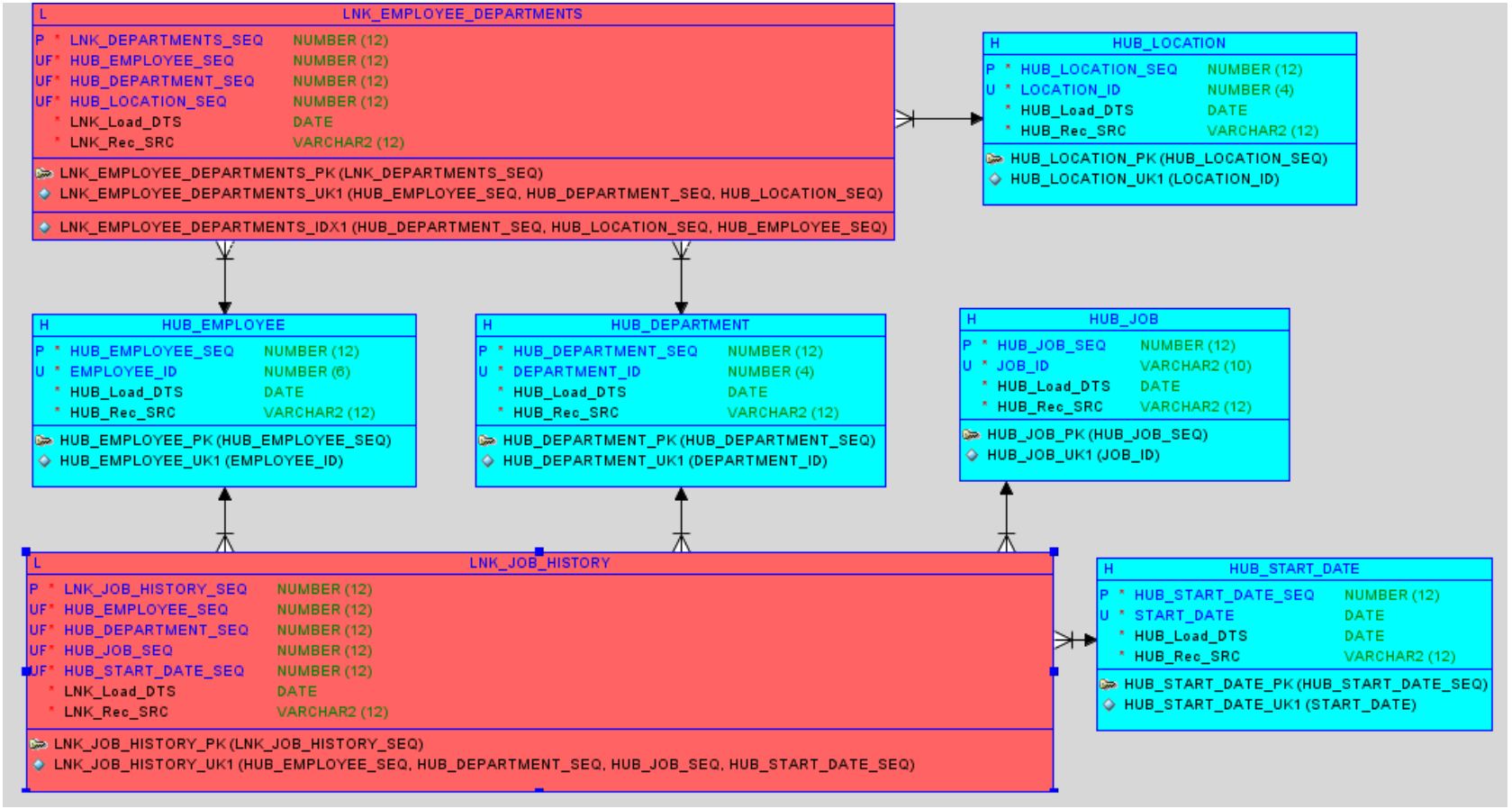


1. Hub



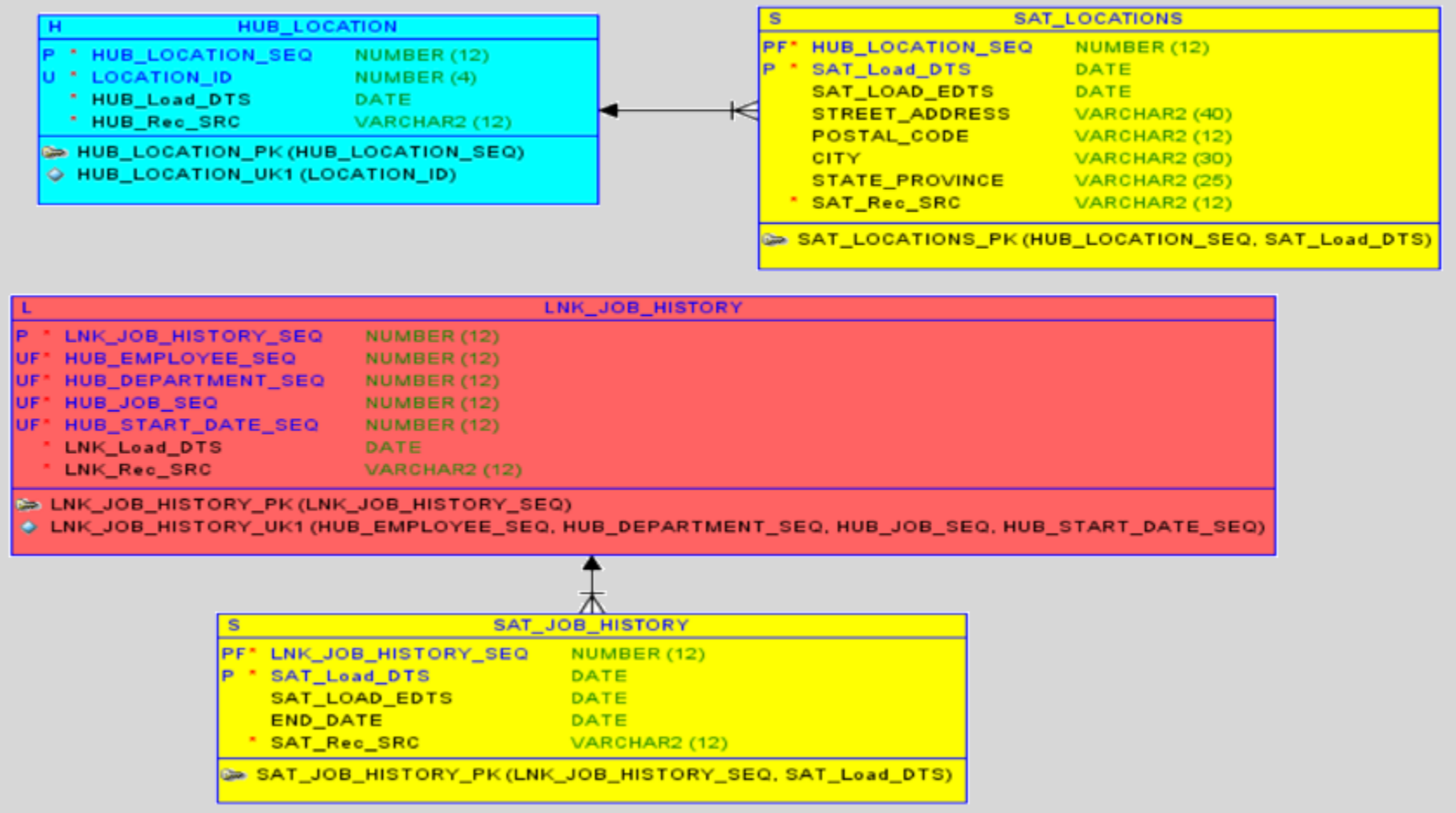
Hub хранит уникальные бизнес-ключи сущностей

2. Link



Link определяет транзакционные, ассоциативные и ролевые связи сущностей

3. Satellite



Satellite содержит:

- Данные.
- Историю изменения данных.
- Историю изменения связей между сущностями.

BI Partner

Содержание

Hadoop

Hadoop вносит изменения в традиционную парадигму хранилищ данных

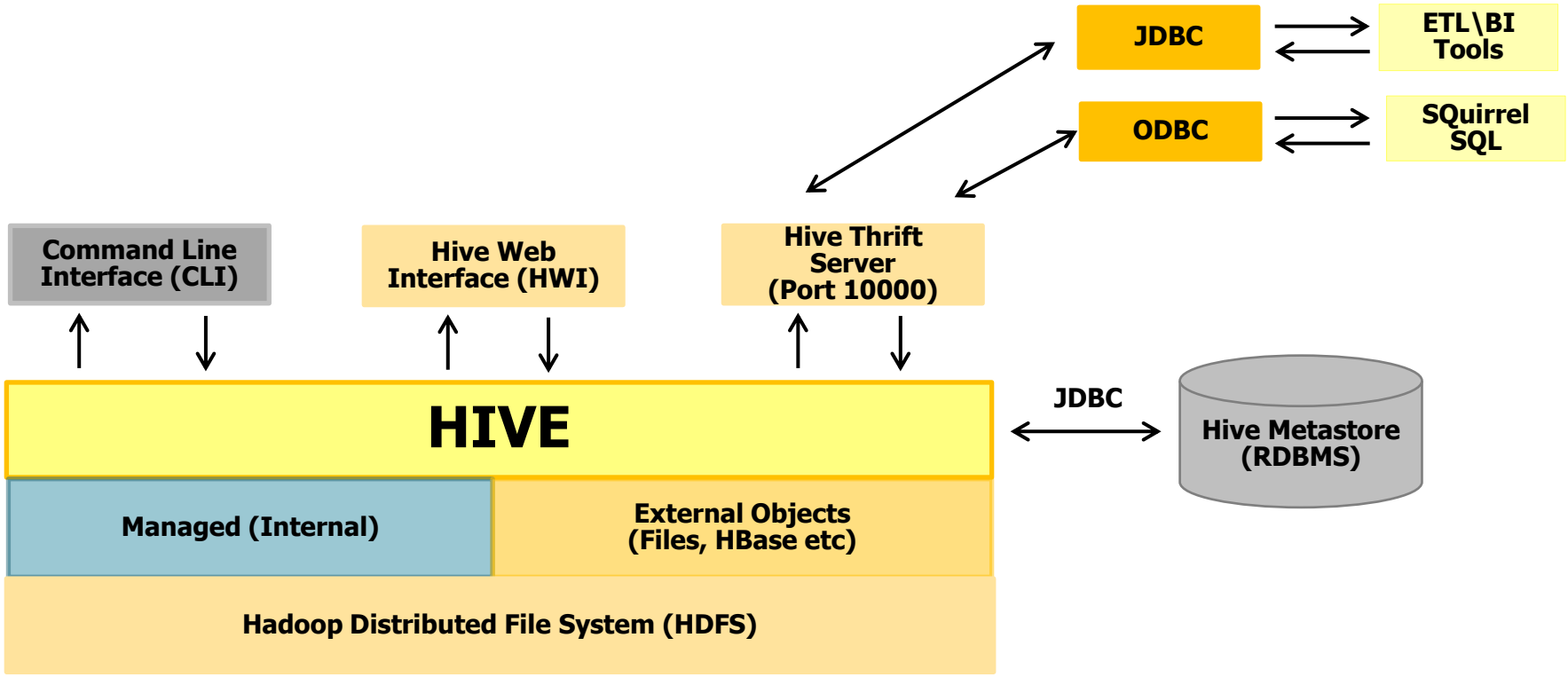
- **Hadoop** не заменяет реляционные базы данных или платформы традиционных хранилищ данных, но его лучше соотношение цена/производительность может помочь организациям снизить затраты, сохраняя при этом существующую инфраструктуру отчетности.

- **Hadoop** может содержать в себе все виды данных: структурированных, частично структурированных и не структурированных. Одно из важных свойств среды – низкая стоимость хранения и легкость масштабирования.

- Одним из эффективных способов использования технологий **Big Data** является применение этой технологии в организации корпоративных хранилищ данных. В этом случае организуется зона первичных данных, содержащая полную историю изменения данных в источниках (идеально подходит для ХД структуры **Data Vault**).

Apache Hive

Apache Hive – хранилище данных для Hadoop



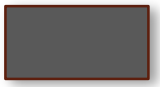
Apache Hive

- **HiveQL** – SQL подобный язык запросов к данным.
- Разработан для работы с данными большого объема (PB).
- Предназначен для анализа данных в т.ч. с использованием ad-hoc запросов.
- Умеет работать с DDL операторами: database \ table \ partition \ bucket.
- Поддерживает стандартные и комплексные типы данных.
- Легко расширяется.
- Не предназначен для небольших баз данных и OLTP систем.

BI Partner

Apache Hive для хранилища данных

- Централизация данных
- MPP и MAP/REDUCE
- Низкая стоимость хранения данных
- Можно использовать сегментирование данных (partitioning)
- Можно использовать сжатие данных
- Можно выбирать формат хранения данных
- Гибкость



- Файловое управление данными
- Данные нельзя обновлять (update) и удалять (delete)
- Для разработки приложений необходимы навыки высококлассного программиста
- Трудно оптимизировать

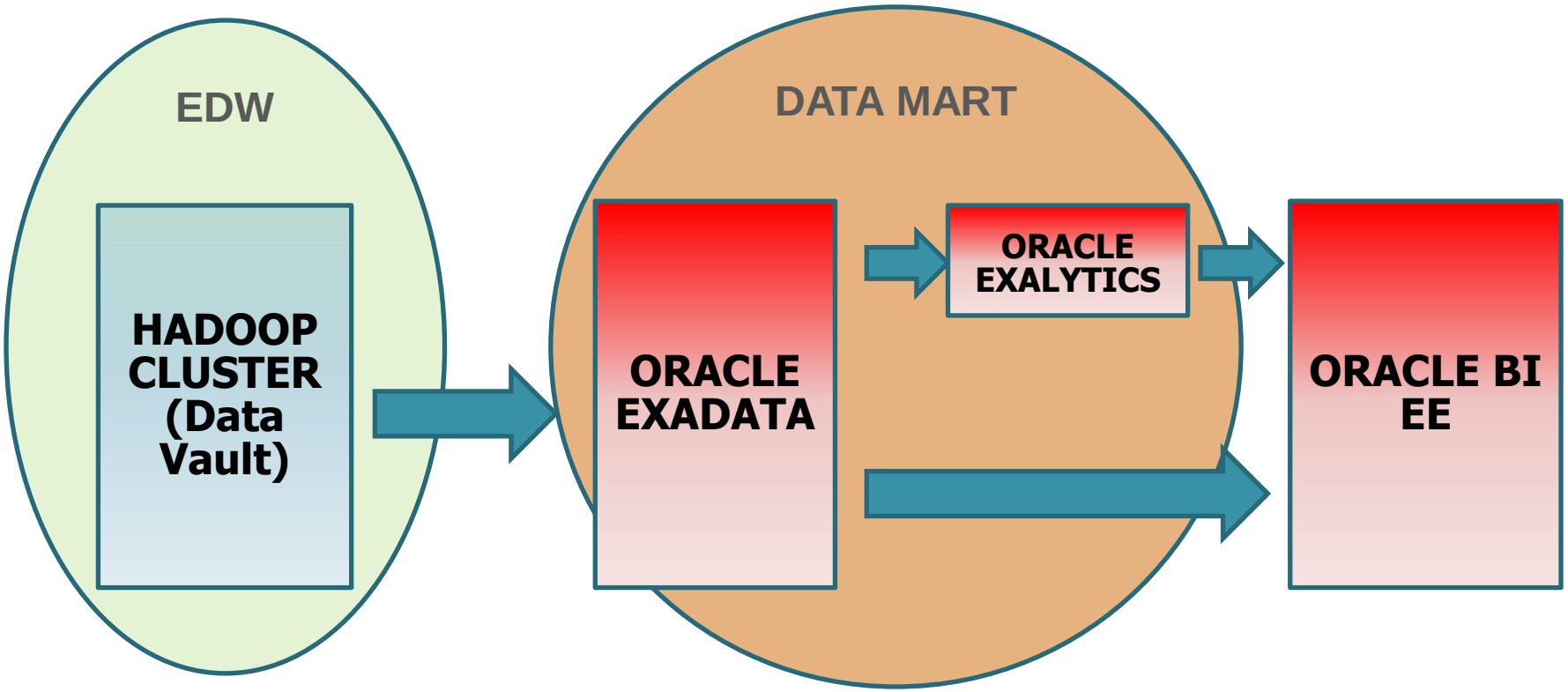
BI Partner

Содержание

- Проблемы классических корпоративных хранилищ данных
- Корпоративные хранилища данных 2.0 (DWH 2.0)
- Хранилища данных в HADOOP
- **Витрины данных на продуктах Oracle и SAP**
- Витрины данных HIVE + HBase (альтернатива)
- BI Partner, предложение к сотрудничеству

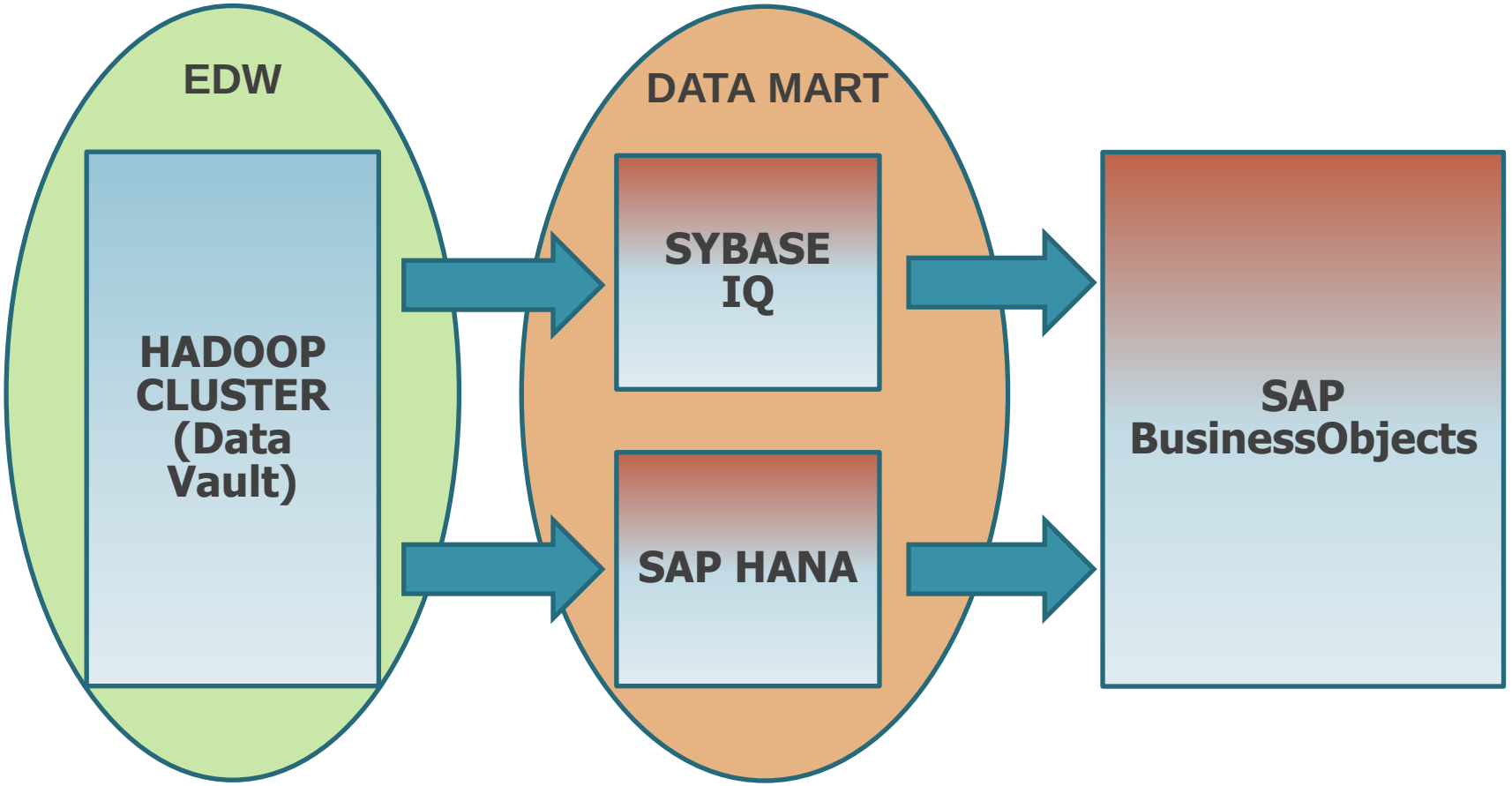
BI Partner

Витрины данных от Oracle



BI Partner

Витрины данных от SAP



BI Partner

Содержание

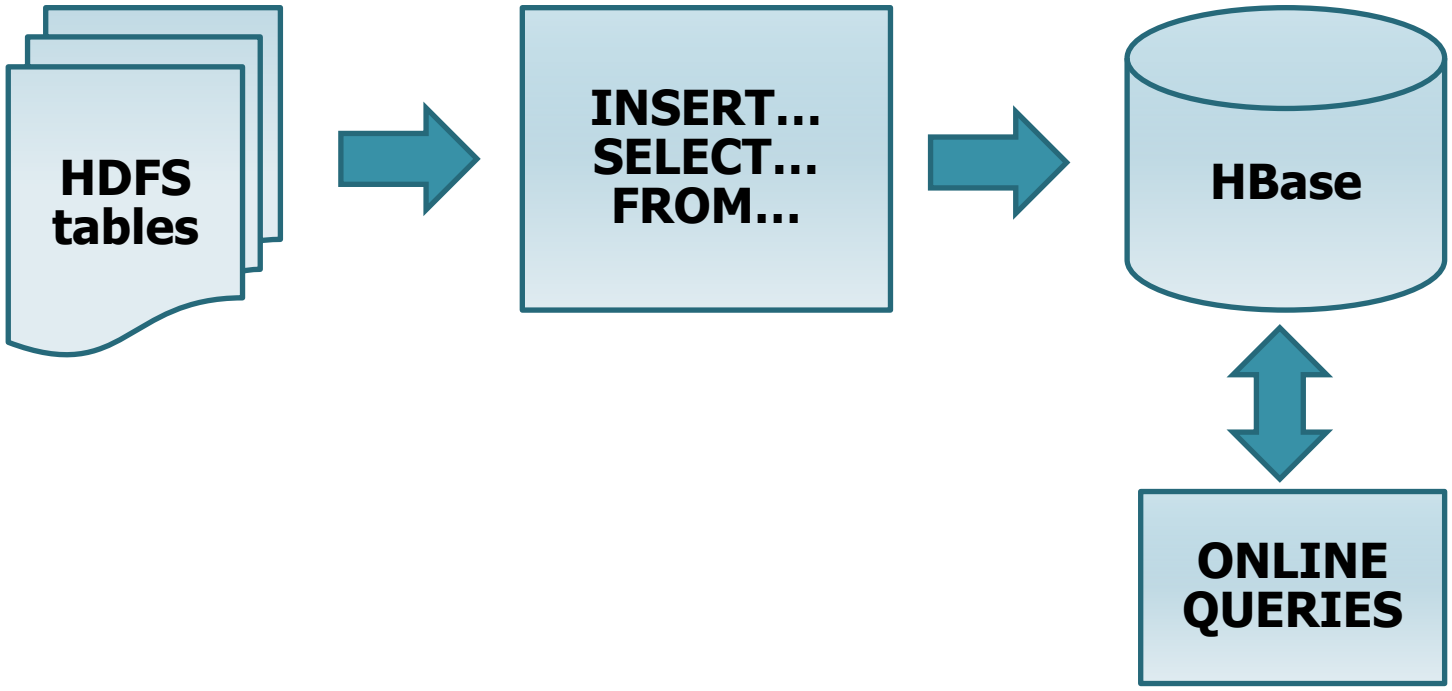
BI Partner

Apache HBase – база данных на Hadoop

- **Apache ™ Hbase** - **не реляционная** (NoSQL) база данных, которая работает поверх Distributed File System (HDFS) в Hadoop.
- Колоночная база данных, которая обеспечивает отказоустойчивость хранения и быстрый доступ к большим объемам разреженных данных.
- Добавляет транзакционные возможности для Hadoop, что позволяет пользователям проводить обновление, вставку и удаление данных.
- Поддерживает автоматическую балансировку нагрузки.
- Поддерживает кэширование в памяти.
- Поддерживает репликацию через центр обработки данных.

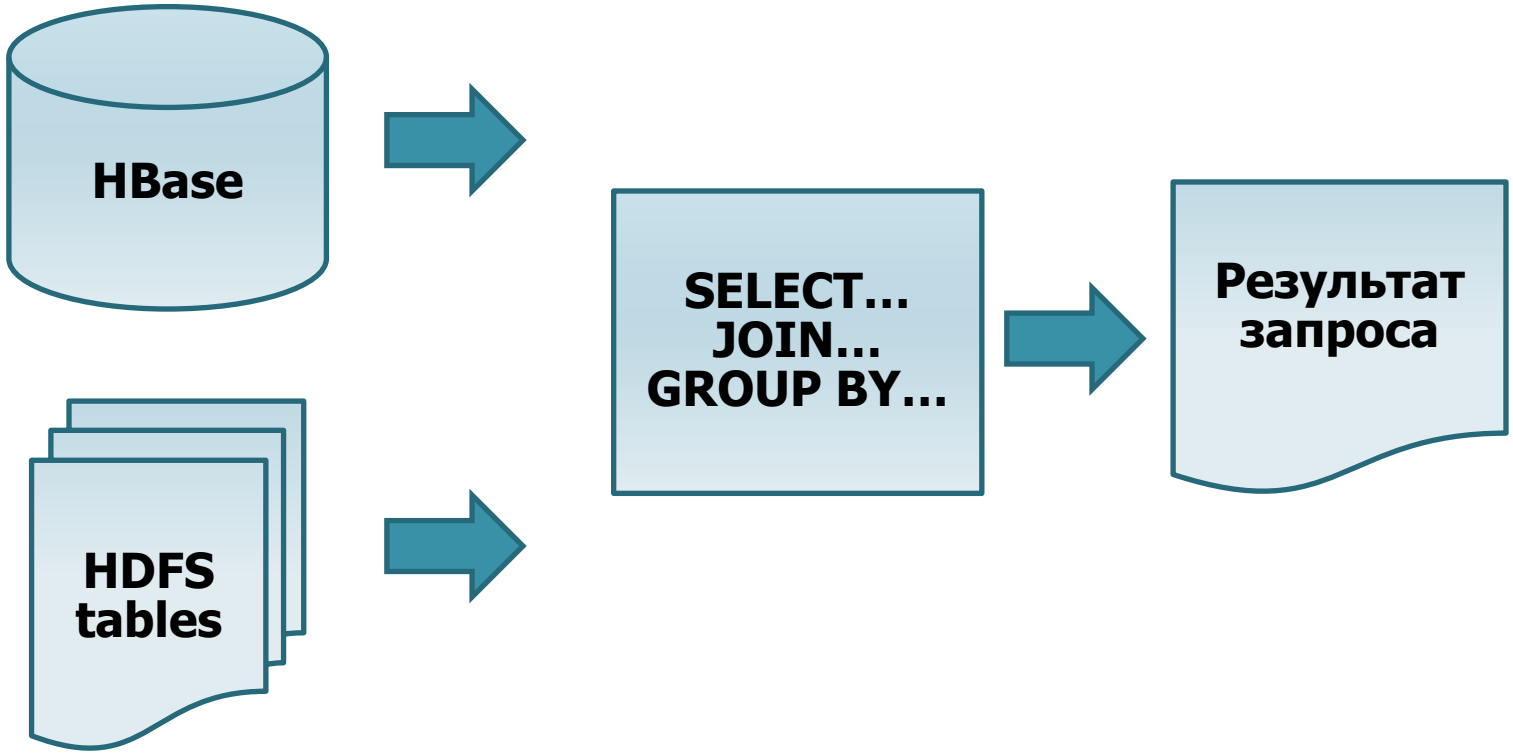
BI Partner

HBase как целевая среда процесса ETL

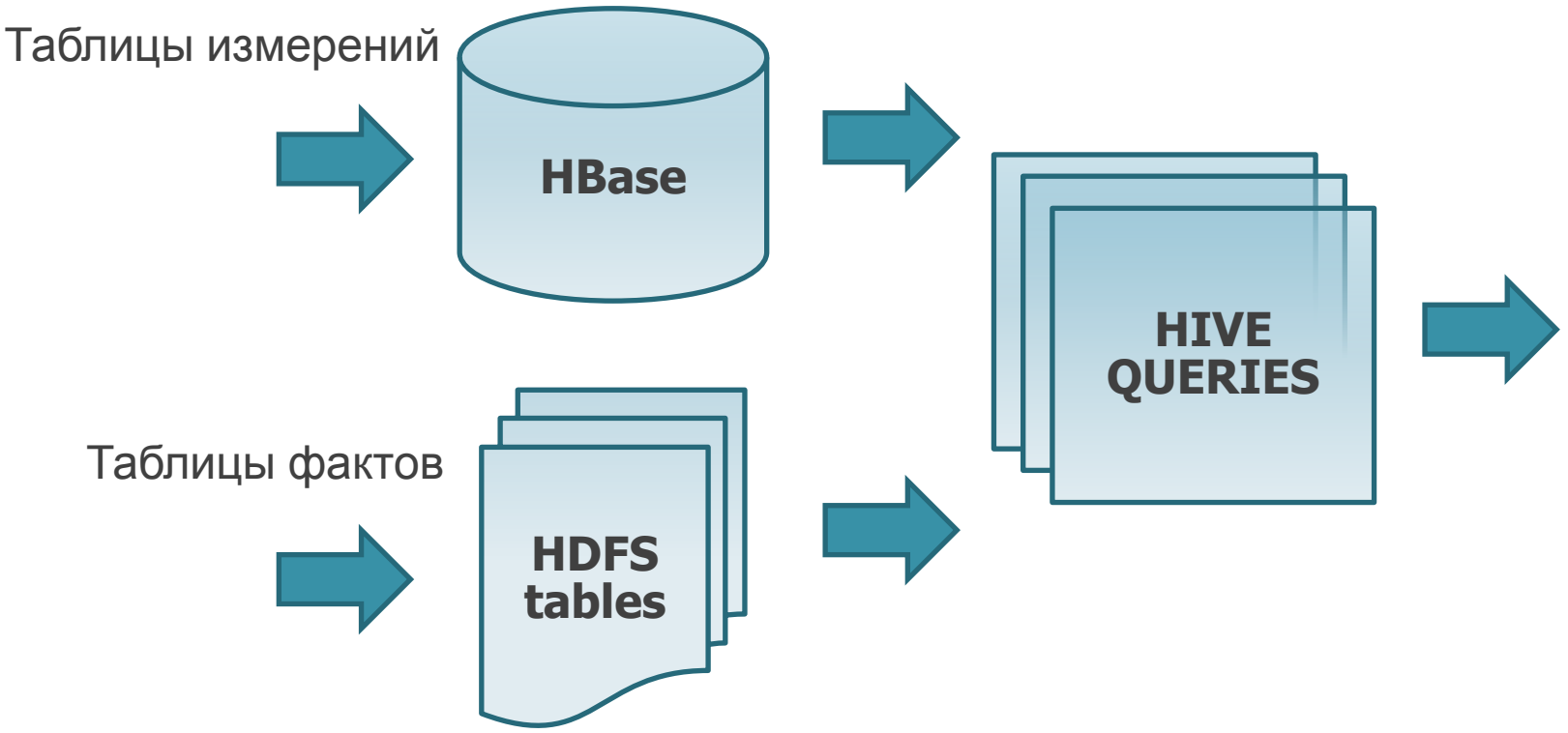


BI Partner

HBase как источник данных



HBase и Hive как высокопроизводительная витрина данных



BI Partner

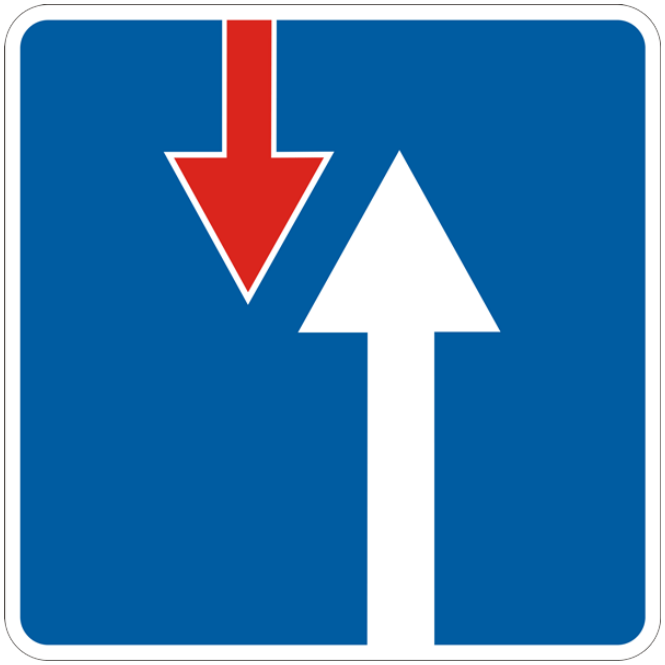
Интеграция HBase и Hive

- Для input\output format, getSplits() и т.д используются базовые классы HBase
- Функции выбора столбца и некоторые фильтры могут быть перенесены вниз к HDFS
- Таблицы HBase могут быть использовань с другими Hadoop-объектами в SQL конструкциях
- DDL операторы Hive преобразуются в DDL операторы HBase

BI Partner

Преимущества Big Data

- **Непомерно высокая стоимость масштабирования традиционных ХД** приводит к экономически неоправданным затратам, и даже если это масштабирование достигнуто, производительность традиционных систем не позволяет производить обработку большого объема данных.
- Применение технологий Big Data **многократно снижает стоимость аппаратных средств** для организации ХД, по сравнению с традиционными СУБД на аналогичных объемах данных.



BI Partner

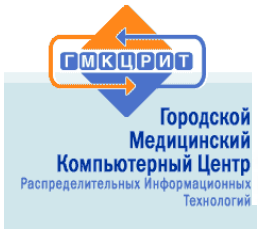
Содержание

О компании BI Partner

- **Год основания: 2002**
- **Первая в России консалтинговая компания, специализирующаяся на бизнес-аналитике**
- **Более 100 проектов в области ХД и BI**
- **Партнер SAP по системам управления эффективностью бизнеса**
  - SAP Reseller
  - SAP PCoE (Partner Centre of Expertise)
  - SAP Preferred Training Partner
  - SAP Subscription-based Hosting Program Provider
- **Партнер Oracle по базовым технологиям и хранилищам данных**
  - Oracle Platinum Partner
  - Специализация **Datawarehousing**
  - Специализация **Oracle Database Performance Tuning**
  - Специализация **Oracle Database**
- **Партнер IBM по бизнес-аналитике и бюджетному управлению**
  - IBM Premium Business Partner

BI Partner

Опыт



BI Partner

Наши конкретные предложения

- Реализовать пилотный проект по построению ХД на технологиях **Big Data**
- Помочь вам развернуть кластер среднего уровня – до 10 узлов на платформе **Hortonworks** или **Cloudera**

www.bipartner.ru

consulting@bipartner.ru

Сергей Сухарев, руководитель BI-практики