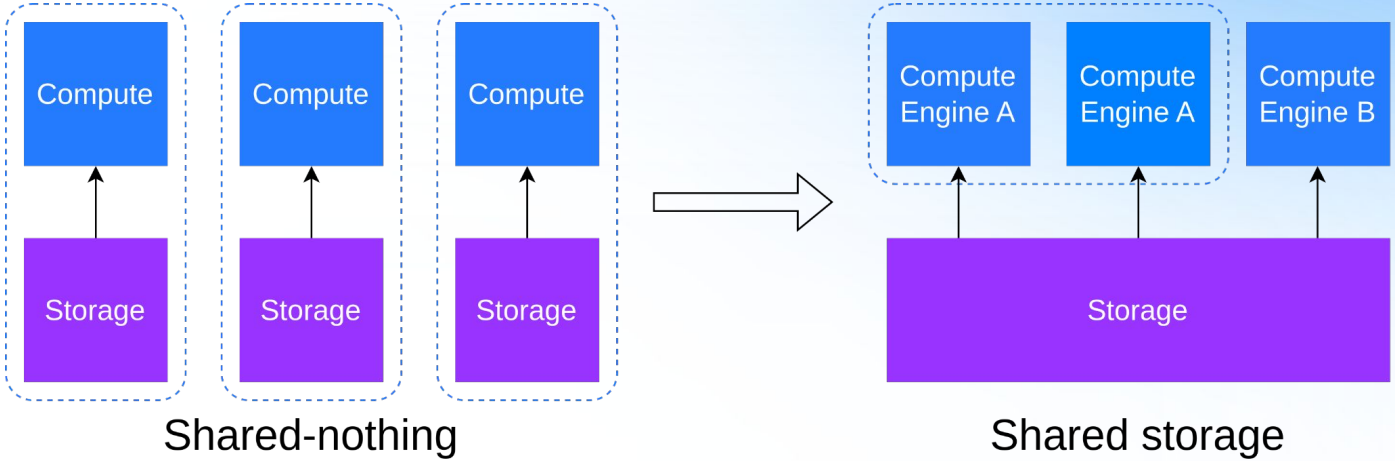
Querify Labs

Современные технологические тренды в аналитике больших и малых данных

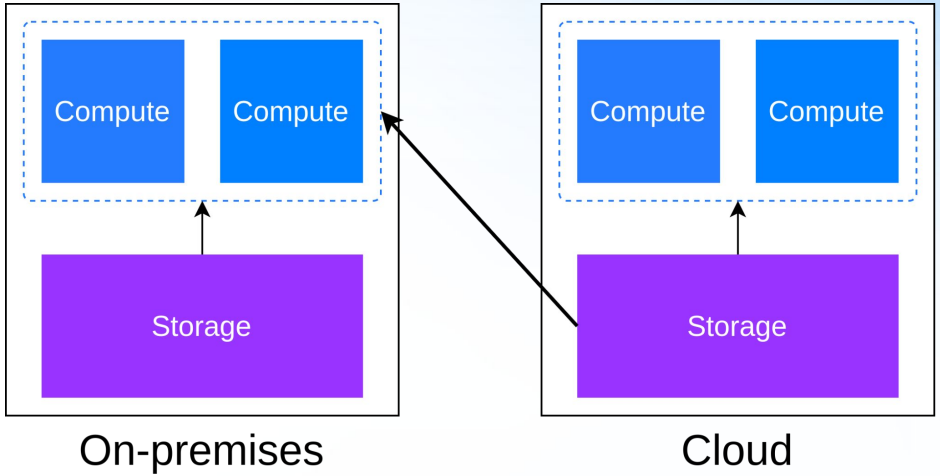
Владимир Озеров
Кверифай Лабс / CedrusData

Disaggregated storage



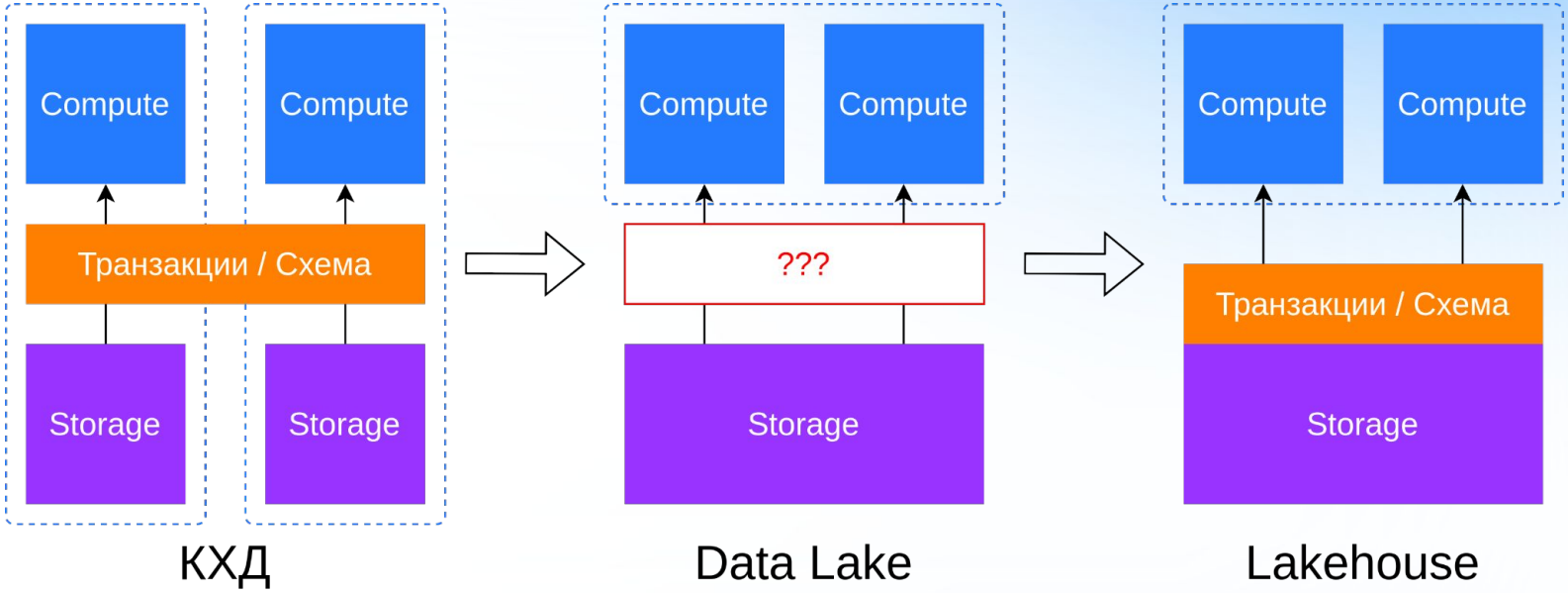
- Shared-nothing системы не отвечают современным требованиям масштабируемости и производительности:
  - Нужны разные движки под разные нагрузки
  - Нужно независимое масштабирование вычислительных ресурсов без перемещения данных
  - Решение: отделение compute от storage
- Драйверы: более быстрое железо; рост количества пользователей; увеличение вариативности нагрузок
- Продукты: Snowflake/BigQuery, Apache Spark, **CedrusData/Trino**

Cloud



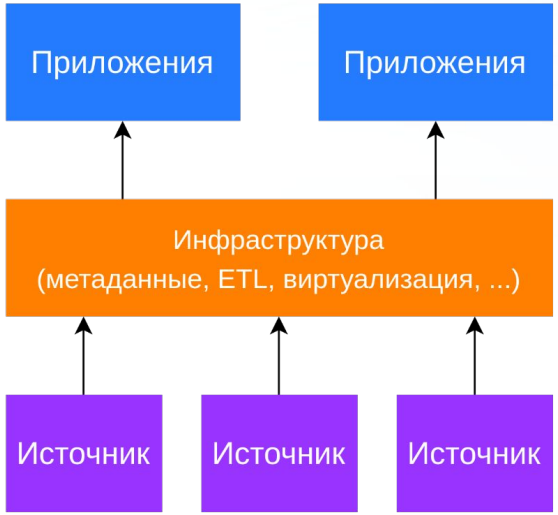
- Преимущества облака:
  - Гранулярная аренда ресурсов
  - Снижение расходов на поддержку инфраструктуры
- Отделение compute от storage позволяет организовать гибридную аналитическую инфраструктуру
- Ключевые продукты: S3-совместимые облачные объектные хранилища (Yandex, VK, ...)

Lakehouse

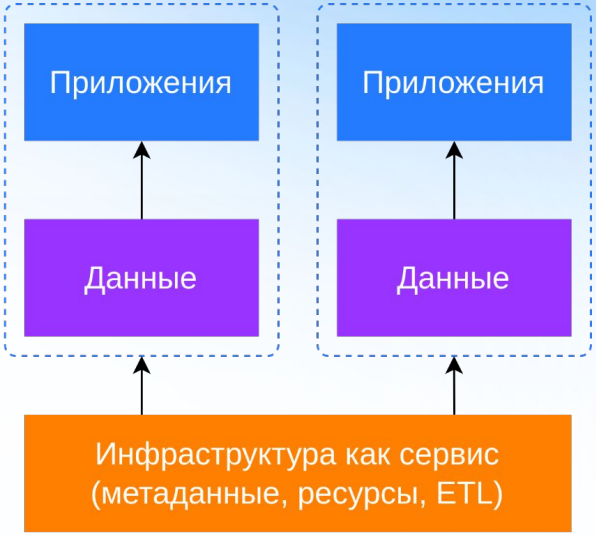


- Поддержка транзакций и эволюции схемы на уровне data lake помогает переносить
- Ключевые технологии: Apache Iceberg, Apache Hudi, Delta Lake

Data fabric и data mesh



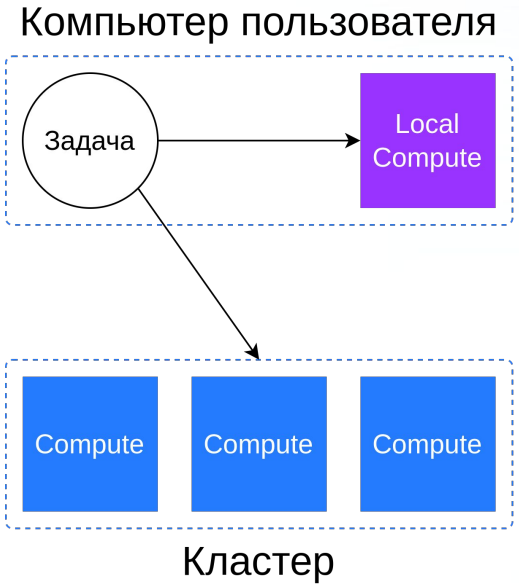
Data Fabric



Data Mesh

- Общие задачи: ускорение time-to-insight
- Технологические драйверы: отделение compute от storage; объектные хранилища; открытые форматы данных; табличные форматы; виртуализация

Small data



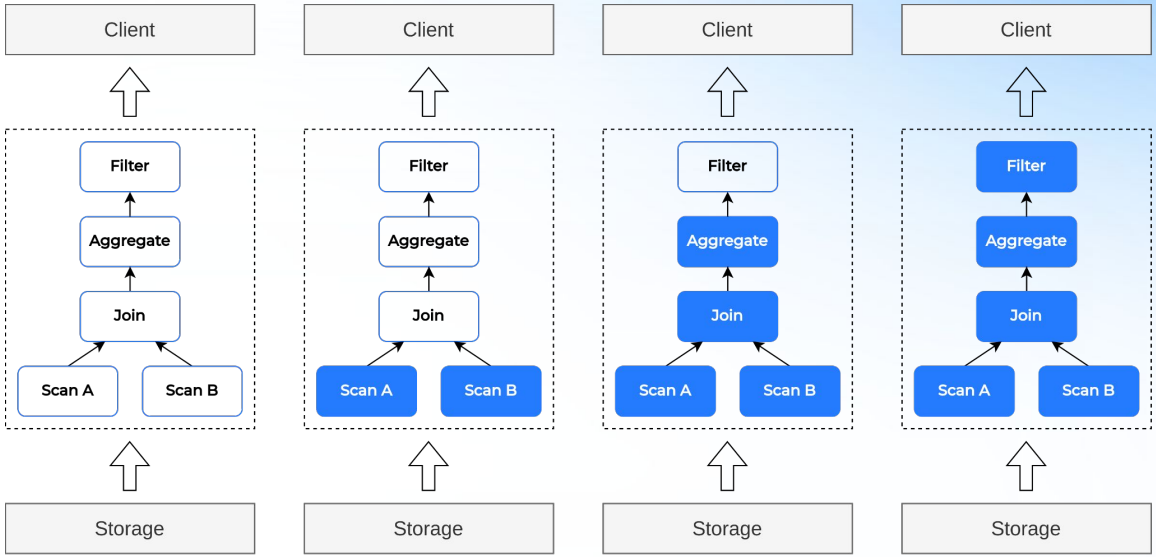
DuckDB

Ibis

Polars

- Наблюдение: значительное количество пользовательских задач могут быть выполнены **локально**!
- Продукты:
  - DuckDB, (+ скоро ClickHouse) – встраиваемые аналитические СУБД
  - Ibis, Polars – библиотеки для эффективной работы с data frame

Оптимизации на примере кэширования



- Наблюдение: значительное количество аналитических запросов выполняют одни и те же операции над медленно изменяющимися данными:
  - Дисковый кэш: снимает нагрузку с data lake
  - Материализованные представления: убирает повторяющиеся подзапросы
  - Кэш результатов: убирает повторяющиеся запросы

Динамика российского рынка

- Последовательная миграция с устаревшего стека:
  - Hadoop >>> Object Storage
    - Накопление опыта с on-premises решениями (напр., CEPH)
    - Активная работа над облачными решениями
    - ПАКи
  - КХД >>> Lakehouse

Динамика российского рынка

- Последовательная миграция с устаревшего стека:
  - Hadoop >>> Object Storage
    - Накопление опыта с on-premises решениями (напр., CEPH)
    - Активная работа над облачными решениями
    - ПАКи
  - КХД >>> Lakehouse
- Низкий уровень доверия бизнеса и вендоров
  - Причины:
    - разрыв отношений с западными вендорами
    - технологическая отсталость и злоупотребления российских вендоров
  - Тренды:
    - “форкнем open-source”
    - “начнем продавать свои внутренние наработки”
  - Если мы продолжим действовать таким образом:
    - Большинство наработок окажутся невостребованными
    - Технологическое отставание будет нарастать

CedrusData и Trino/DuckDB

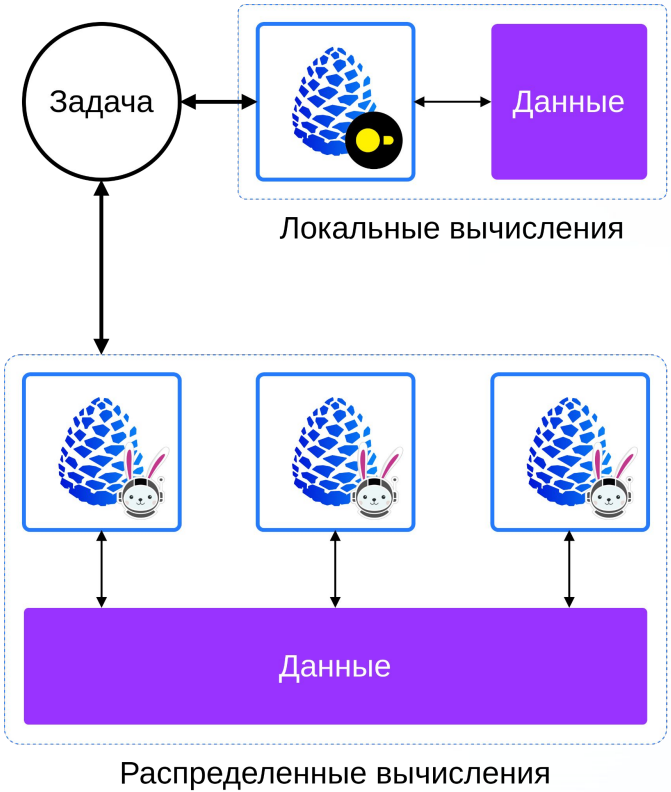
cedrusdata

trino

DuckDB

- **CedrusData** – это набор технологий для современных аналитических платформ на основе open-source продуктов:
  - **Trino** – масштабируемый распределенный compute для больших данных
  - **DuckDB (private beta)** – высокопроизводительные локальные вычисления
- Обеспечивает эффективную обработку сложных аналитических запросов в любых окружениях и любых архитектурах
- Принципы:
  - SQL как основной интерфейс доступа
  - Оптимизация под аналитические нагрузки
  - Отделение compute от storage
  - Виртуализация и интеграция данных из различных источников
  - Многослойное кэширование для минимизации нагрузки на платформу данных

CedrusData и Trino/DuckDB



Поддержка современных архитектурных подходов:

- Lakehouse (SQL к data lake, Iceberg/Hudi/Delta Lake)
- Data fabric (виртуализация, ETL, управление доступом)
- Data mesh (отделение compute от storage, ETL)

Динамика:

- Системная работа над снижением пользовательской нагрузки на платформу данных:
- Развитие интеграций с новыми приложениями и источниками данных
- Высокая расширяемость

Контакты

ООО «Кверифай Лабс»
ИНН 7811766769
ОГРН 1217800163790

Конакты:

- Телеграм: https://t.me/cedrusdata
- Сайт: https://cedrusdata.ru
- Email: info@cedrusdata.ru
- Телефон: +7(812)9839840