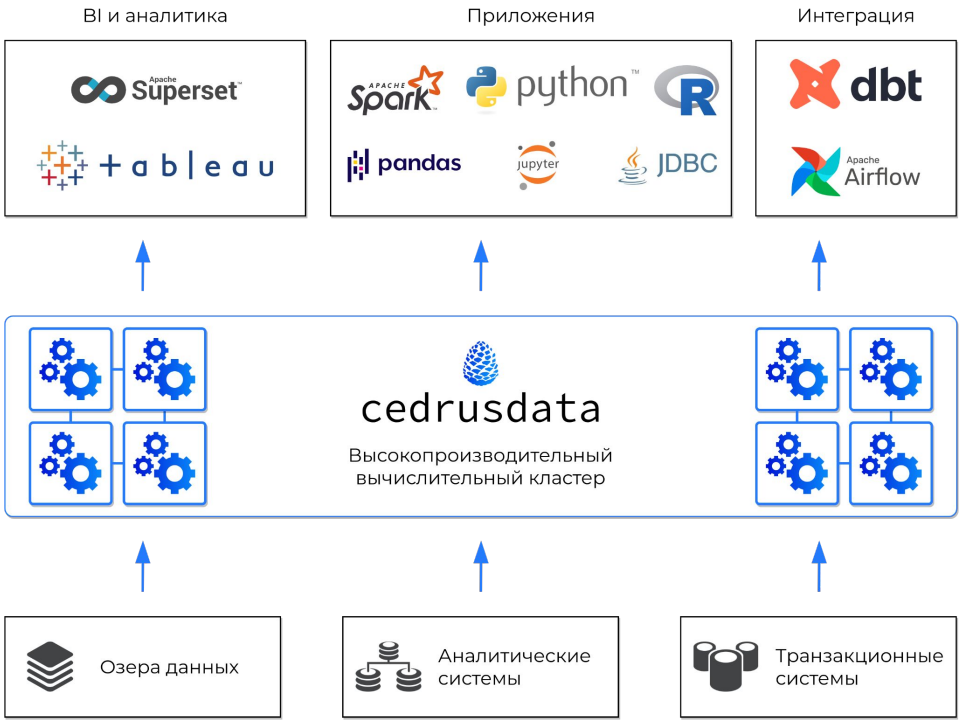
Querify Labs

CedrusData: как обеспечить анализ всех данных предприятия и снизить затраты на инфраструктуру

Обзор

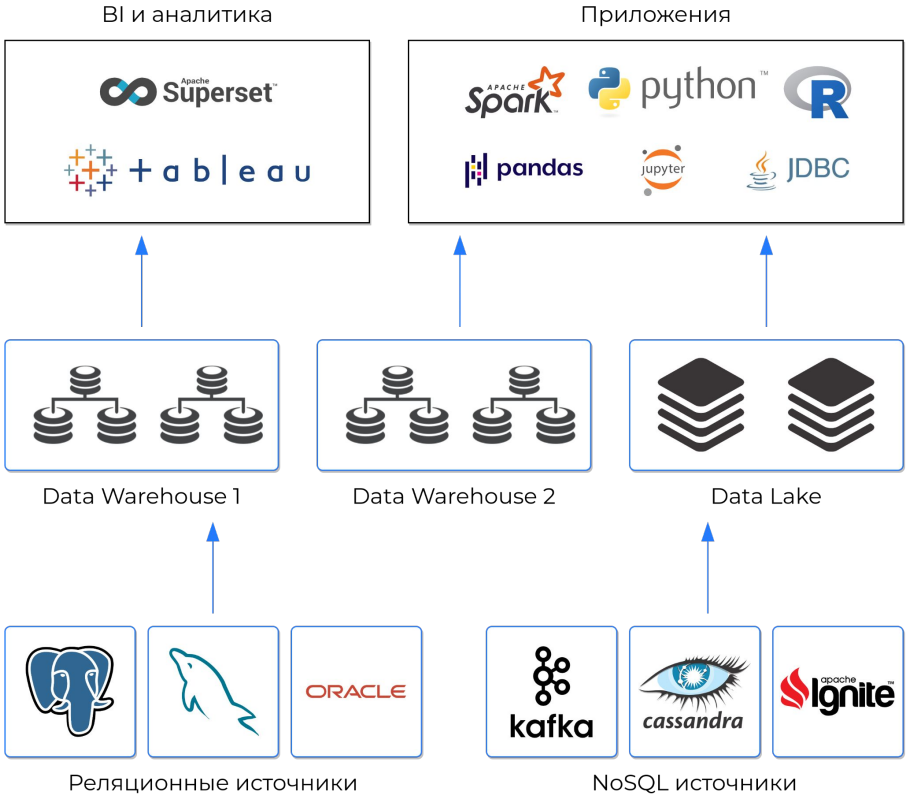


CedrusData — это аналитическая система, которая позволяет компаниям быстро и гибко анализировать все свои данные через единую точку доступа в облаке и on-premise. Основана на популярном open-source проекте Trino.

Особенности:

- Эластичная масштабируемость за счет разделения compute и storage.
- Коннекторы к популярными источникам данных:
  - Озера данных (Hive).
  - Аналитические системы (Greenplum, ClickHouse и др.).
  - Транзакционные системы (Postgres, MySQL, Oracle, SQL Server и др.).
  - NoSQL (Cassandra, MongoDB, Redis, Kafka и др.).
- Легко интегрируется с облачными технологиями (Docker, Kubernetes).

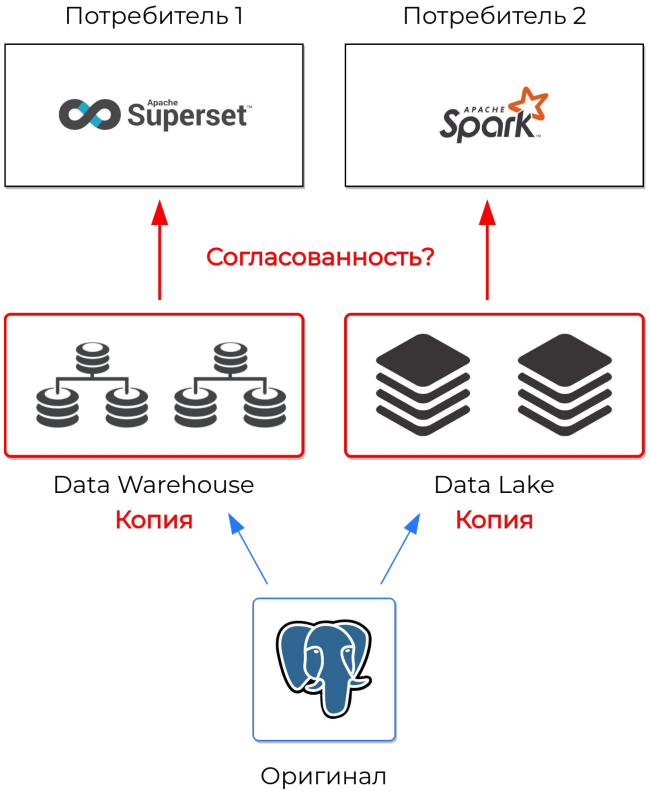
Аналитические платформы



Особенности типичной аналитической платформы предприятия:

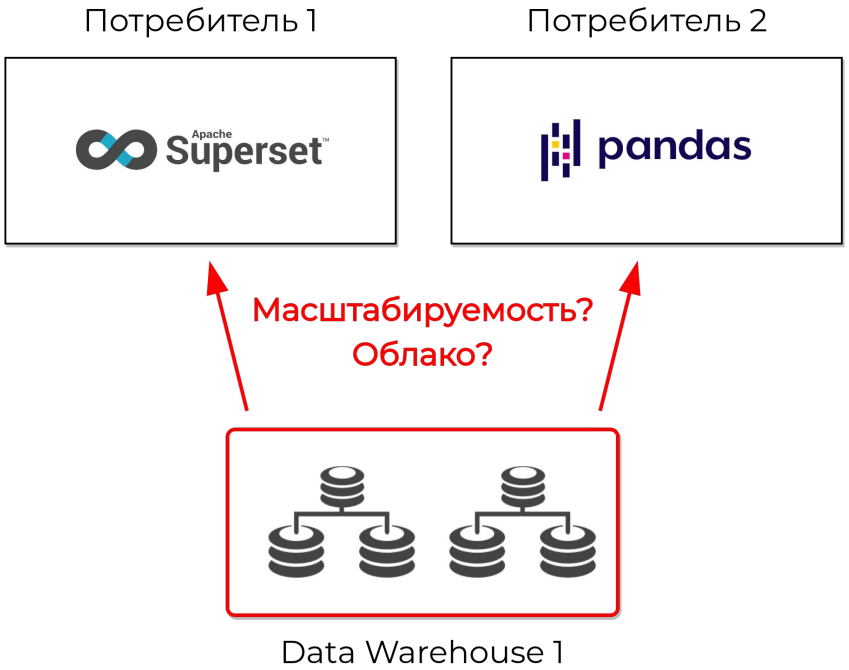
- Множество источников данных.
- Одно или несколько аналитических хранилищ для структурированной информации.
- Озеро данных для работы с сырыми и неструктурированными данными.
- Множество потребителей (пользователи, приложения, продукты) с разнородной нагрузкой.

Издержки: дублирование данных



Загрузка данных в хранилище приводит к созданию дополнительных копий в проприетарных форматах, которые не могут быть напрямую использованы другими приложениями.

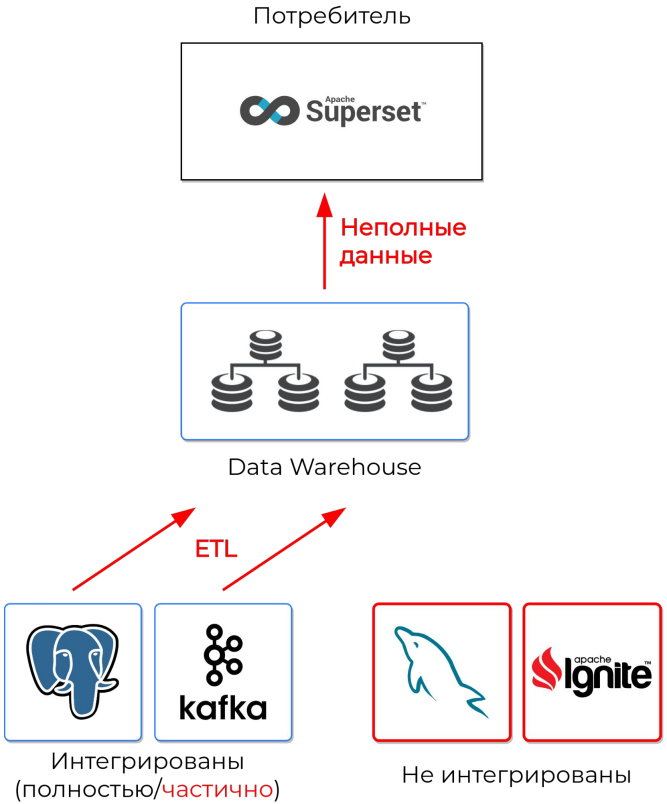
Издержки: проблемы хранилищ



Хранилища данных, построенные на shared-nothing архитектуре (Greenplum, Teradata, Vertica, и т.д.), обладают ограниченной масштабируемостью:

- Сложные аналитические запросы не укладываются в рамки фиксированных схем шардирования.
- Совмещение compute и storage приводит к необходимости совмещения нагрузок от разных приложений в одном кластере и затрудняет эластичное масштабирование в облаке.

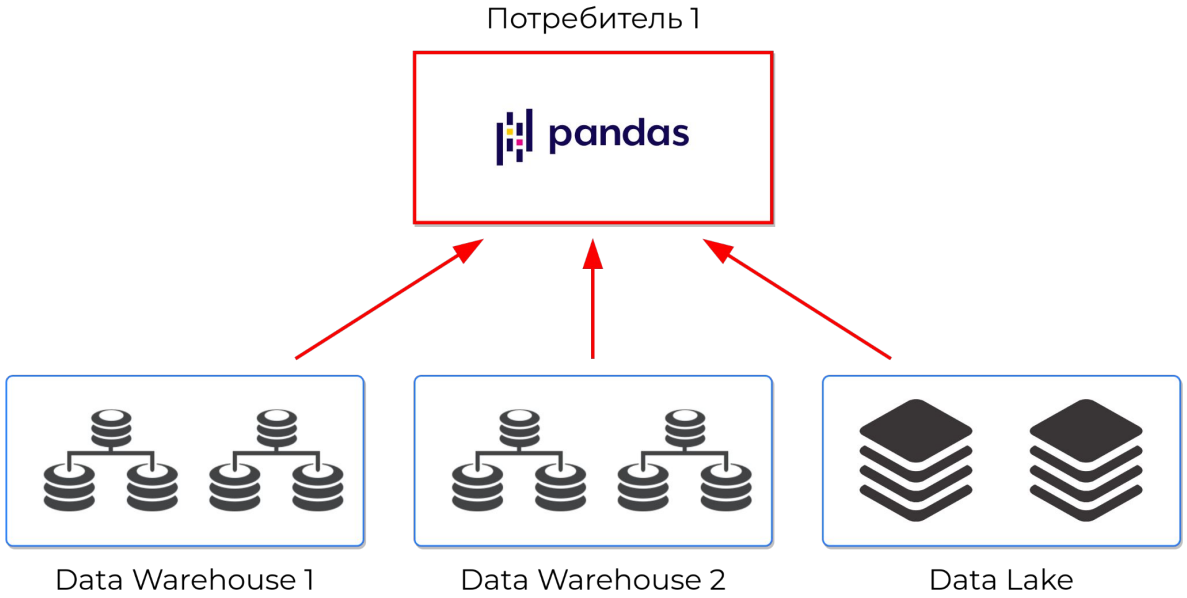
Издержки: ETL



Загрузка данных в корпоративное хранилище происходит с помощью ETL.

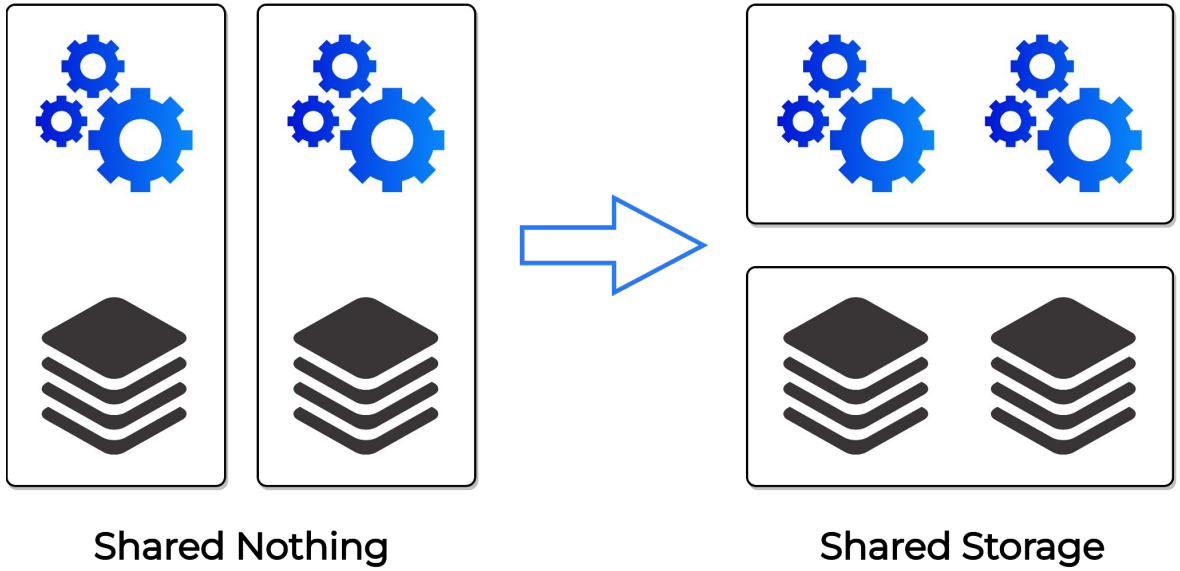
Внесение изменений (новые источники, новые требования, и т.п.) требует вовлечение инженеров платформы данных, что существенно увеличивает сроки реализации новых идей.

Издержки: интеграция данных



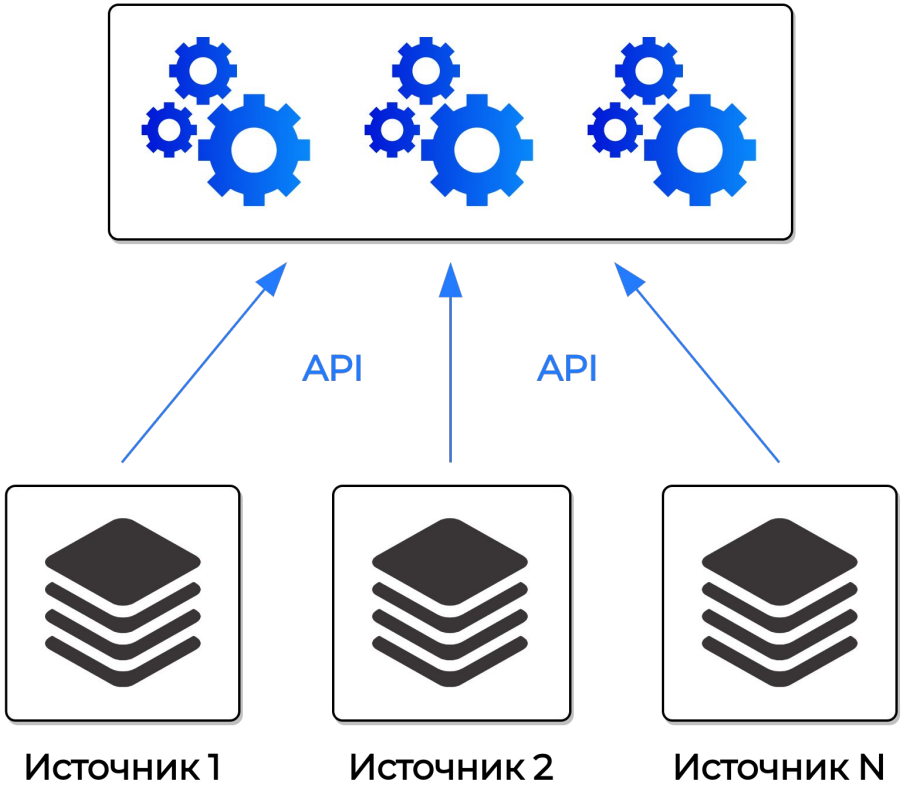
Объединение данных из хранилищ и озера данных затруднено, так как каждая система имеет собственный интерфейс. Это приводит к дальнейшему замедлению скорости реализации новых сценариев, а также увеличивает нагрузку на клиентские приложения.

Shared Storage



- Независимое масштабирование вычислений и хранения.
- Изменение топологии вычислительного кластера не влияет на данные.
- Более гибкое и эффективное использование вычислительных мощностей.
- Отлично подходит для облака.

Виртуализация



Абстрагирование источника данных от вычислений позволяет реализовать сценарий анализа данных из множества источников через единый интерфейс.

Presto, Trino и CedrusData

Presto - технология массивно-параллельной обработки больших данных из разных источников с SQL интерфейсом, разработанная Facebook для внутренних нужд, и опубликованная в 2013 году.

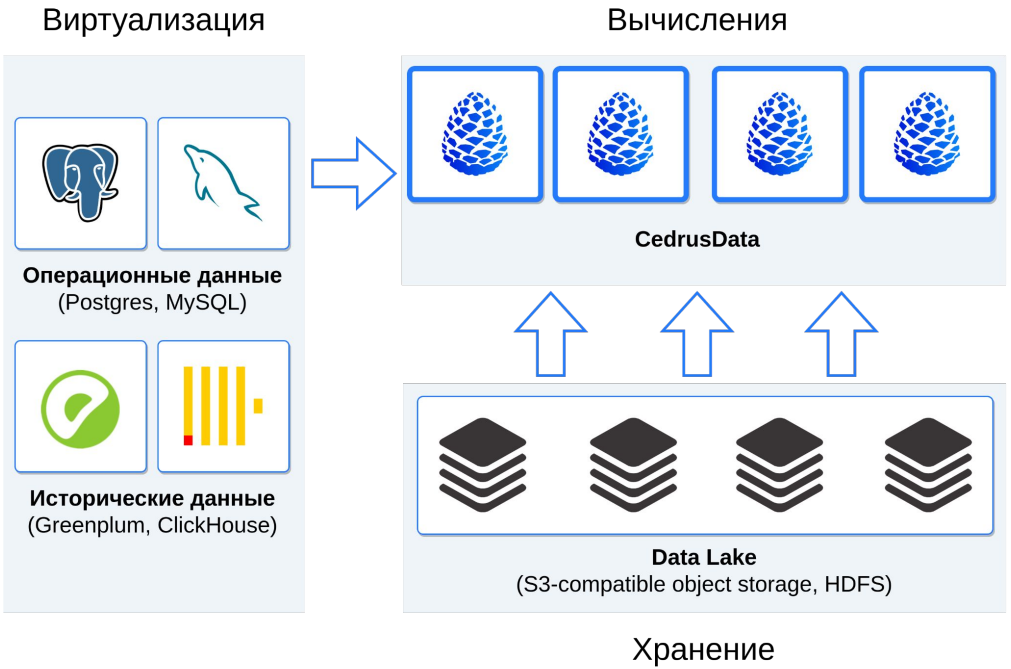
- Нацелен на решение внутренних инфраструктурных задач крупнейших интернет-компаний.
- Поддерживается Meta, Intel, Alibaba, Uber.

Trino - форк Presto, развиваемый оригинальными авторами Presto с 2018 года.

- Нацелен на средний и крупный бизнес.
- Активное сообщество, постоянный поток улучшений.

CedrusData - это коммерческий форк Trino, содержащий дополнительный функционал и улучшения производительности.

Целевая архитектура CedrusData



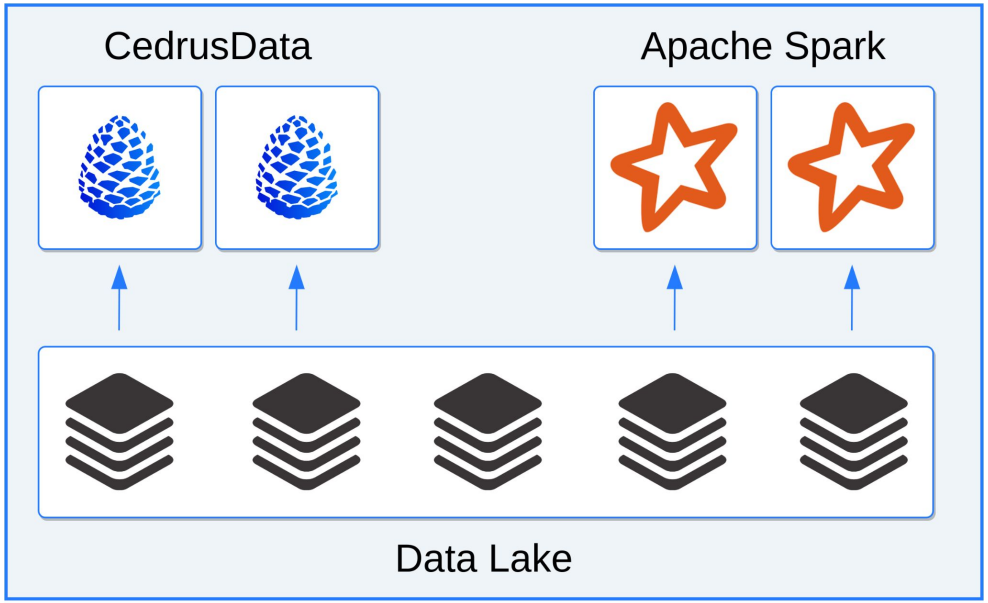
Принципы:

- Отделить данные от вычислений.
- Перенести бОльшую часть аналитических операций в data lake.
- Предоставить возможность интеграции данных между разными системами (виртуализация), в том числе путем отправки запросов к источникам напрямую.

Преимущества:

- Возможность анализа всех данных организации.
- Основной массив данных расположен в дешевом дисковом хранилище (on-premise или в облаке) в открытых форматах, без многократного дублирования.
- Вычисления легко масштабировать в облаке и on-premise.

Сценарий 1: аналитика в озерах данных



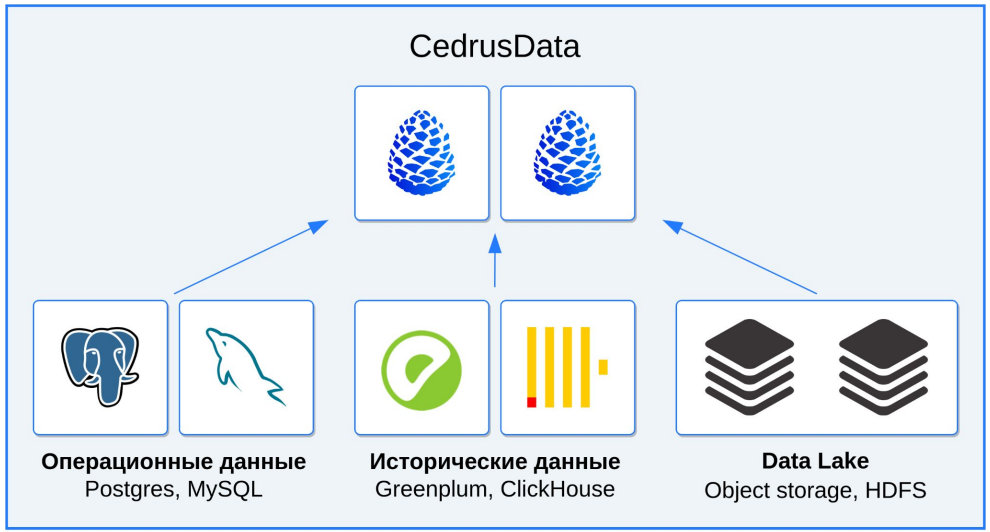
Облако, on-premise

Организации используют озера данных для выполнения задач batch processing (например, ML с помощью Apache Spark), в то время как задачи интерактивной аналитики решаются посредством хранилищ данных.

Технология CedrusData позволяет выполнять задачи интерактивной аналитики путем отправки SQL-запросов к озеру данных.

- Уменьшения дублирования данных.
- Снижение расходов на лицензии и инфраструктуру за счет переноса нагрузки в более дешевое озеро данных.

Сценарий 2: виртуализация данных

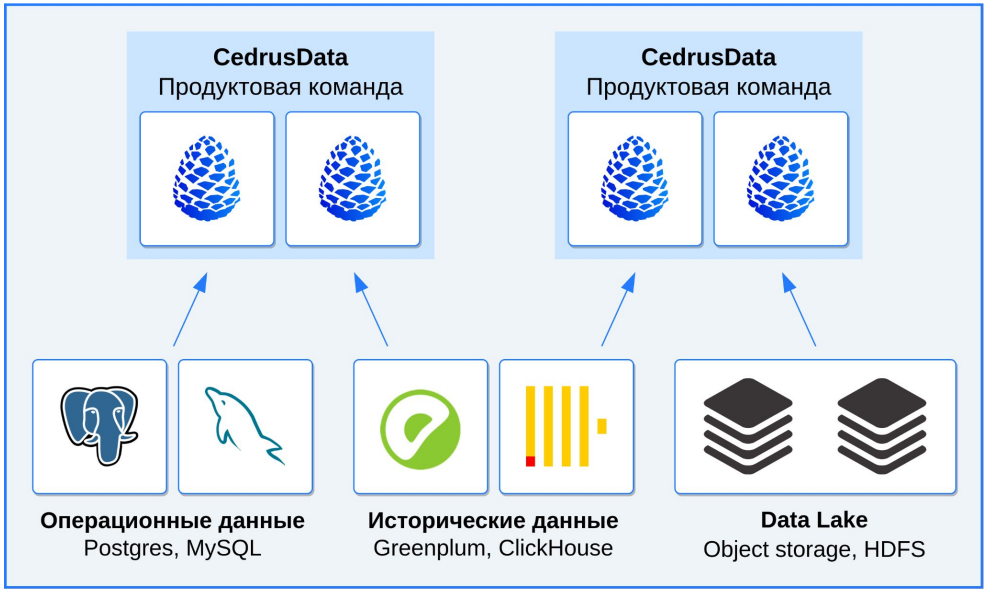


Организации обычно имеют множество источников операционных и исторических данных. Возможности анализа данных из разных источников часто ограничены необходимостью разработки сложных ETL-процедур.

Технология **CedrusData** позволяет быстро объединять данные из разных источников с помощью простых SQL-запросов.

- Позволяет организации сосредоточиться на реализации новых бизнес-сценариев, вместо разработки инфраструктуры.

Сценарий 3: децентрализованная аналитика



Облако, on-premise

Организации часто используют централизованные платформы данных. Чрезмерная централизация замедляет скорость внедрения инноваций, так как внесение изменений требует сложных инженерных и финансовых согласований.

Технология **CedrusData** может быть использована для организации децентрализованных и data mesh архитектур, в которых продуктовые команды работают с выделенными кластерами CedrusData независимо от других команд.

- Позволяет продуктовым командам быстро внедрять новые аналитические сценарии.
- Позволяет организации гибко управлять расходами на лицензии и инфраструктуру без overprovisioning.

Отличия CedrusData от Trino

	Trino	CedrusData
Базовый функционал Trino	✓	✓
Дополнительные коннекторы		✓
Улучшения производительности		✓
Расширенный мониторинг		✓
Оперативное исправление дефектов		✓
Реестр российского ПО		✓
Проверка на отсутствие вредоносного кода		✓
Сервисы (поддержка, консалтинг, тренинги)		✓

Контакты

Контакты:

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