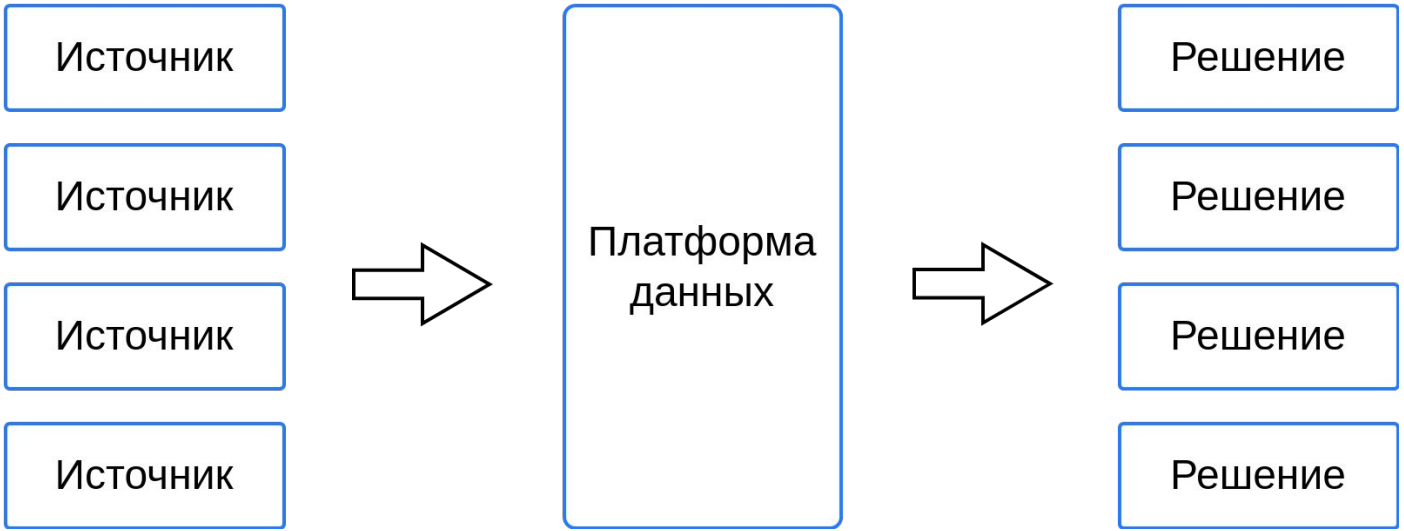
Querify Labs

# CedrusData

## Анализ данных предприятия с помощью технологий Trino и CedrusData

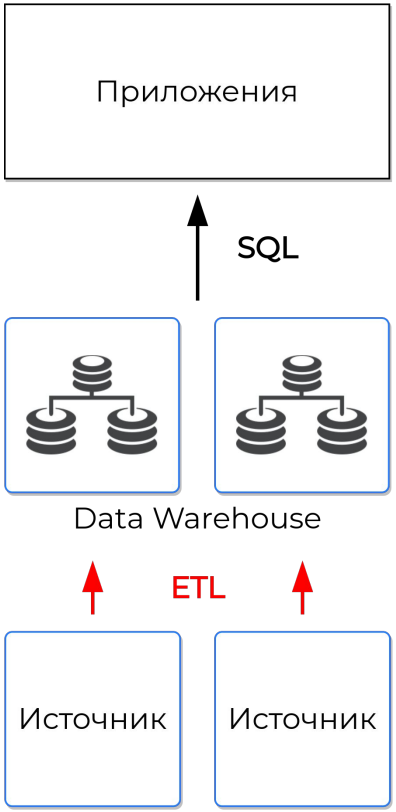
# Принцип




Принятие решений на основе данных:

- Извлечь данные из исторических источников
- Интегрировать
- Проанализировать

# Data Warehouse




Принцип: с помощью ETL перенесем данные из источников в СУБД, оптимизированную под аналитическую обработку.

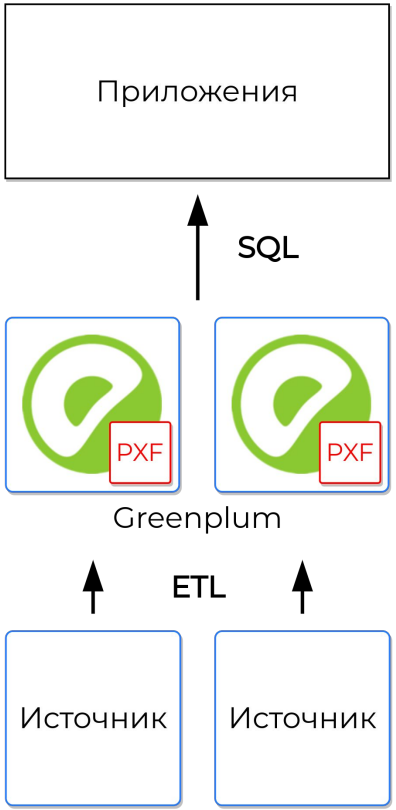
Плюсы:

- Быстро
- Понятно

Минусы:

- Бизнес: долгий цикл внедрения изменений
- Инфраструктура: дорого за счет дублирования данных и потребности в резервировании избыточных мощностей

# Data Warehouse со встроенным ETL




Принцип: все то же самое, только добавим встроенную возможность интеграции данных из внешних источников.

Примеры: Greenplum PXF, Teradata

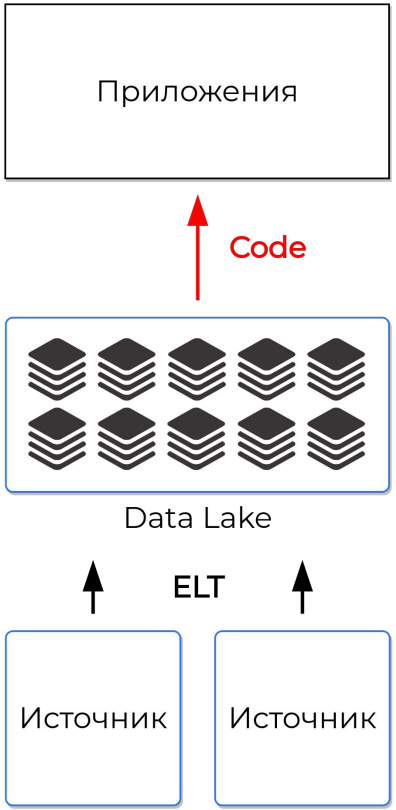
Плюсы:

- Уменьшает потребность в ETL

Минусы:

- Медленно
- Не в приоритете у вендоров
  - Например, PXF развивается с минимальной скоростью

# Data Lake




Принцип: выгрузим сырые данные в дешевое озеро, потом разберемся, что с ними делать.

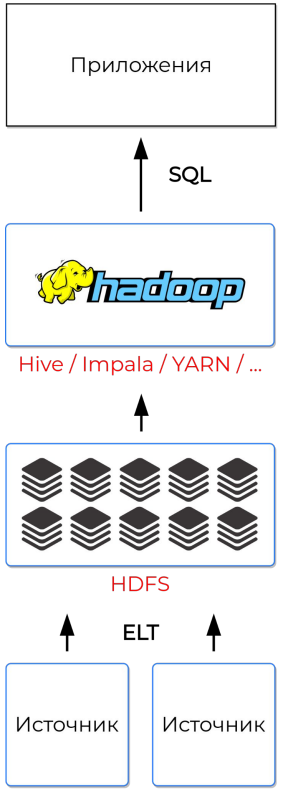
Плюсы:

- Дешевое хранение
- Относительно простой процесс загрузки данных

Минусы:

- Извлечение данных из озера обычно требует написания кода и недоступно рядовым пользователям

# Data Lake + SQL on Hadoop




Принцип: храним данные в HDFS, читаем с помощью Hive / Impala.

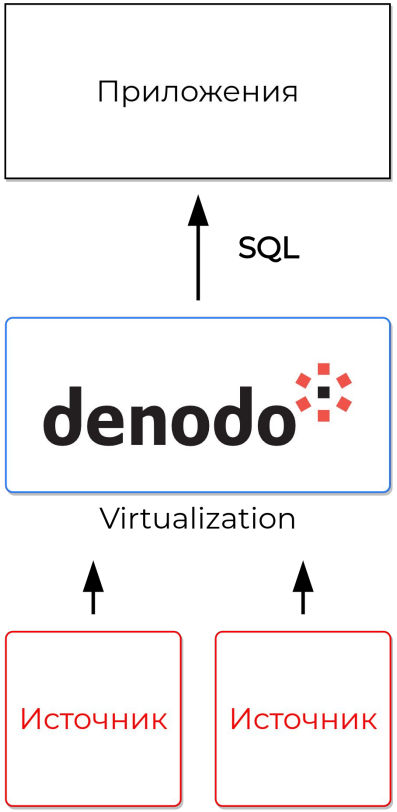
Плюсы:

- SQL

Минусы:

- Legacy
- SQL поверх map-reduce – это медленно (Hive)

# Виртуализация




Принцип: объединяем данные из разных источников «на лету»

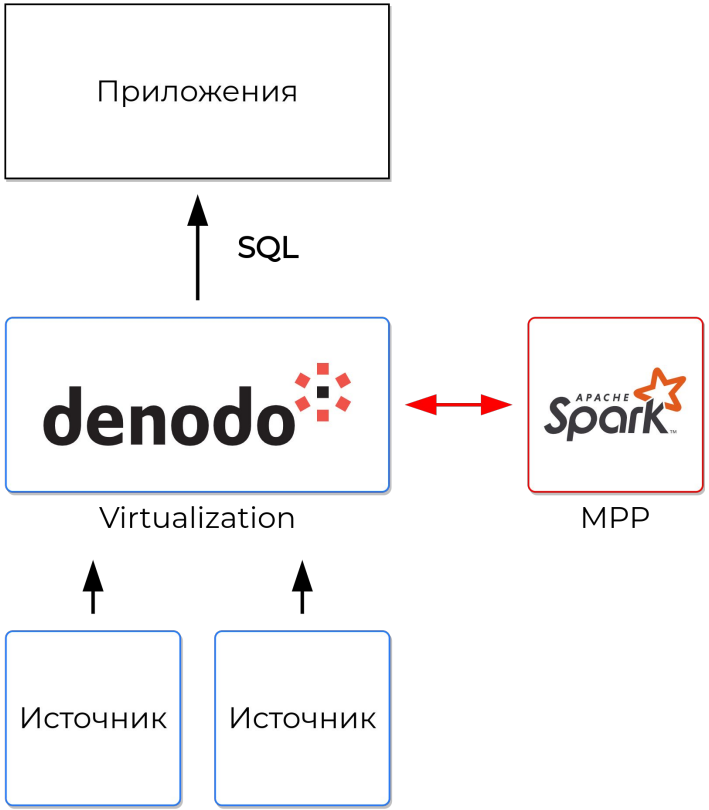
Плюсы:

- Доступ ко всем данным без ETL

Минусы:

- Тяжелые запросы могут дестабилизировать источники
- Нет возможности объединения больших объемов информации из разных источников

# Виртуализация с MPP




Принцип: все то же самое, но делегируем финальную обработку в сторонний движок (напр., Spark, Impala, Presto).

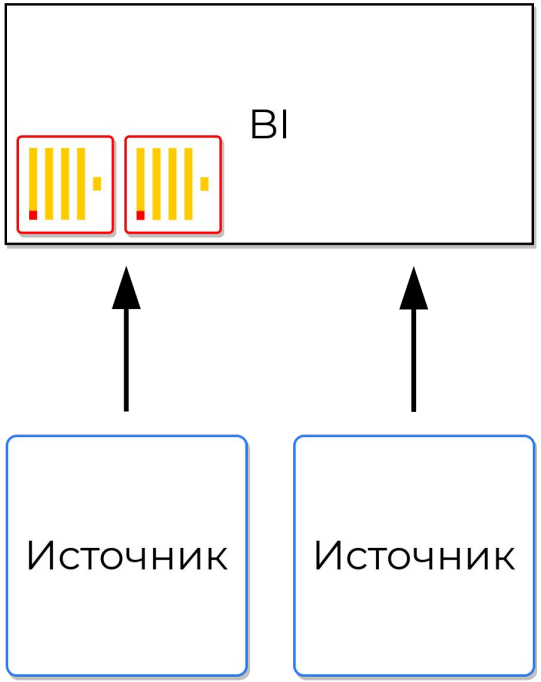
Плюсы:

- Теперь можем обрабатывать данные любого объема

Минусы:

- Относительно медленно за счет потребности в дополнительной передаче данных между системами
- Усложняет инфраструктуру

# Виртуализация




Принцип: загружаем данные во внутренний движок BI инструмента

Плюсы:

- Простая архитектура (на первый взгляд)
- Высокая скорость

Минусы:

- Дублирование данных
- Не поддерживает сложные запросы
- Как интегрировать большие объемы данных с разных систем?

# Trino и CedrusData



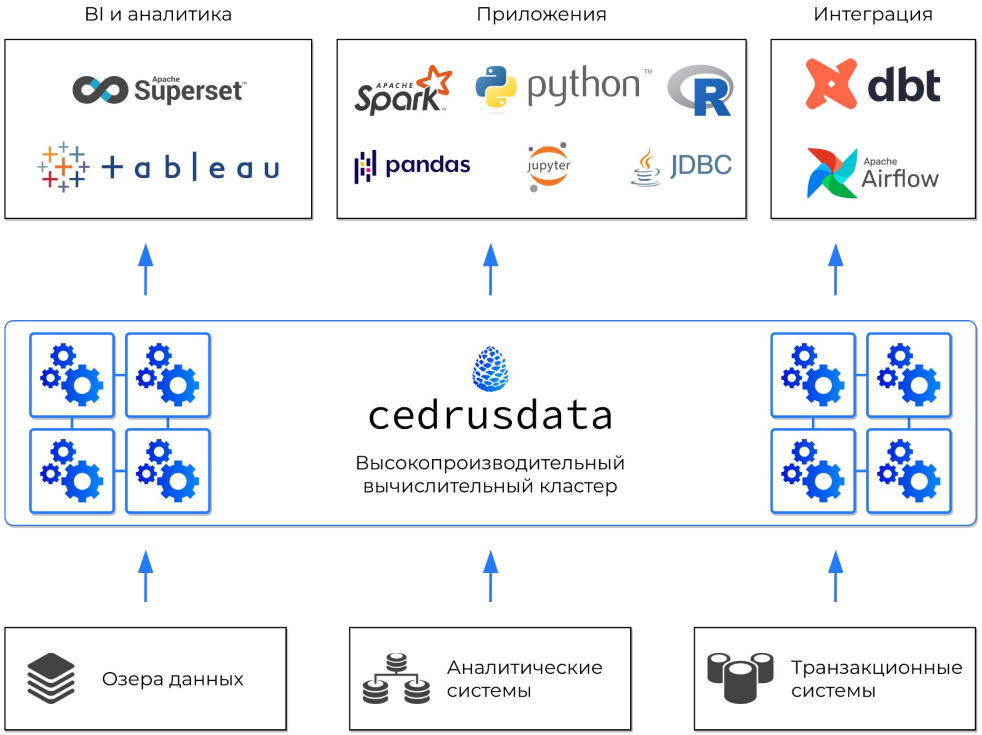

Presto — это open-source технология массивно-параллельной обработки больших данных из разных источников с SQL интерфейсом, разработанная Facebook, и опубликованная в 2013 году. Продукт нацелен на решение задач крупнейших технологических компаний: обработку данных масштаба петабайт на тысячах серверов.

Trino — это форк Presto, развиваемый оригинальными авторами Presto с 2018 года, и нацеленный на решение задач крупного и среднего бизнеса. Имеет активное open-source сообщество и сотни внедрений в компаниях различного размера.

Компании из списка Fortune 500 успешно используют технологии Presto и Trino для обработки больших объемов данных от нескольких терабайт до десятков петабайт.

CedrusData - это коммерческий форк Trino, адаптированный для российского рынка и содержащий дополнительный функционал и улучшения производительности.

# CedrusData




Технология CedrusData использует современные технические решения:

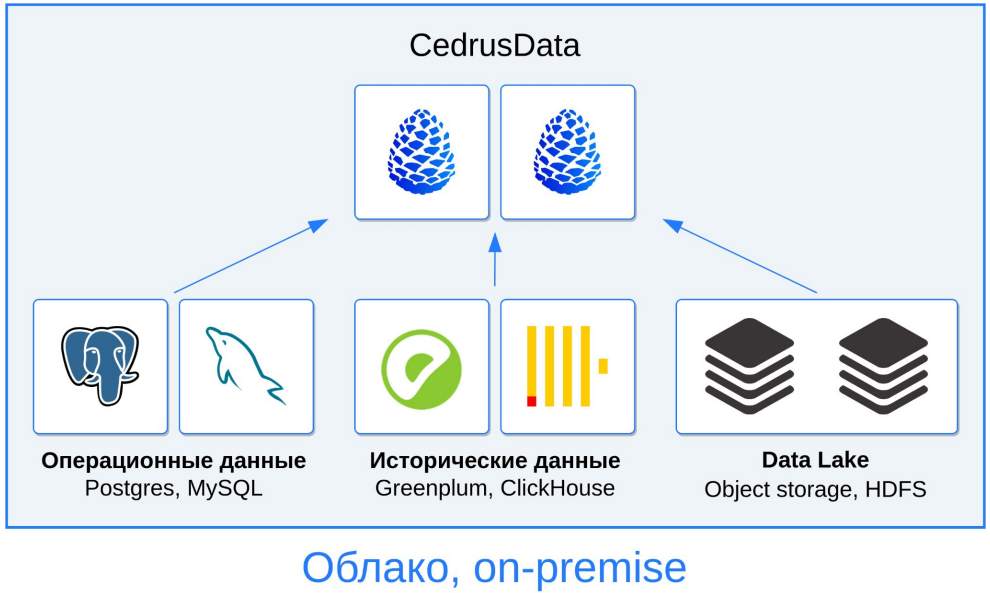
- MPP архитектура с разделением compute и storage
- Федеративное выполнение SQL-запросов
- Локальный дисковый кэш данных

Комбинация этих подходов в одном продукте позволяет компаниям быстро анализировать данные из любых источников с помощью языка SQL.

Преимущества:

- **Скорость для бизнеса**: CedrusData предоставляет доступ ко всем данным организации без необходимости реализации ETL-процессов
- **Экономия для IT**: CedrusData позволяет перенести нагрузку из дорогостоящих корпоративных хранилищ в пользу более дешевого озера данных
- **Гибкость**: CedrusData может быть развернута on-premise или в облаке; как единый центральный кластер организации, или как множество независимых кластеров (data mesh)
- **Итеративное внедрение**: CedrusData обладает эластичной масштабируемостью и не требует обязательной переработки существующей архитектуры платформы данных. Компания может плавно переносить нагрузку в CedrusData по мере необходимости

# Сценарии: исследовательский анализ



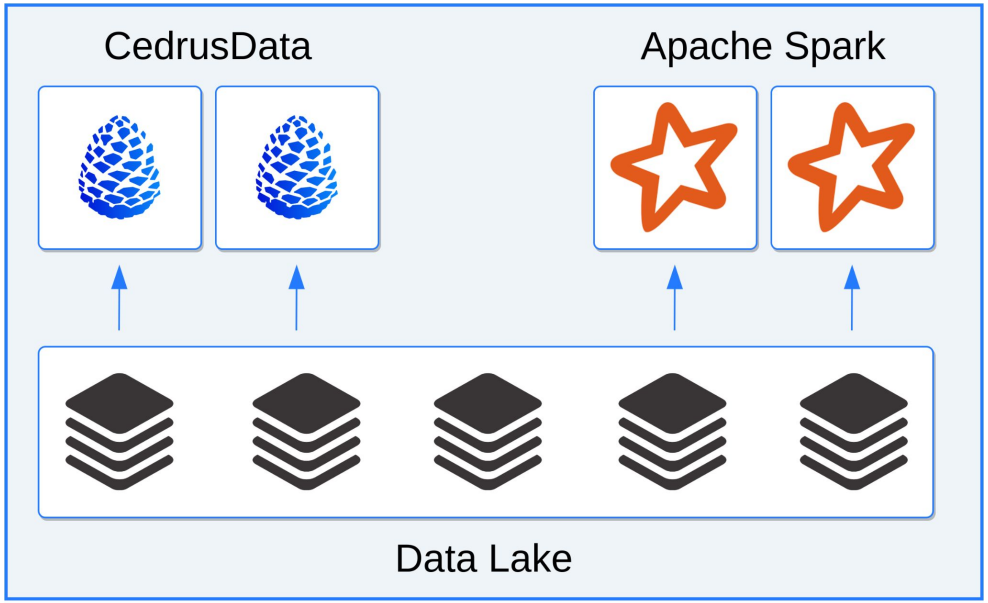

Организации обычно имеют множество источников операционных и исторических данных. Возможности анализа данных из разных источников ограничены необходимостью разработки сложных ETL-процедур.

Технология CedrusData позволяет быстро объединять данные из разных источников с помощью SQL-запросов. Организация может сосредоточиться на реализации новых бизнес-сценариев, вместо разработки инфраструктуры.

Преимущества:

- Широкие возможности исследовательского и интерактивного анализа данных за счет мгновенного доступа ко всем данным организации
- Снижение расходов на корпоративное хранилище данных и ETL

# Сценарии: интерактивная аналитика в озерах данных




Облако, on-premise

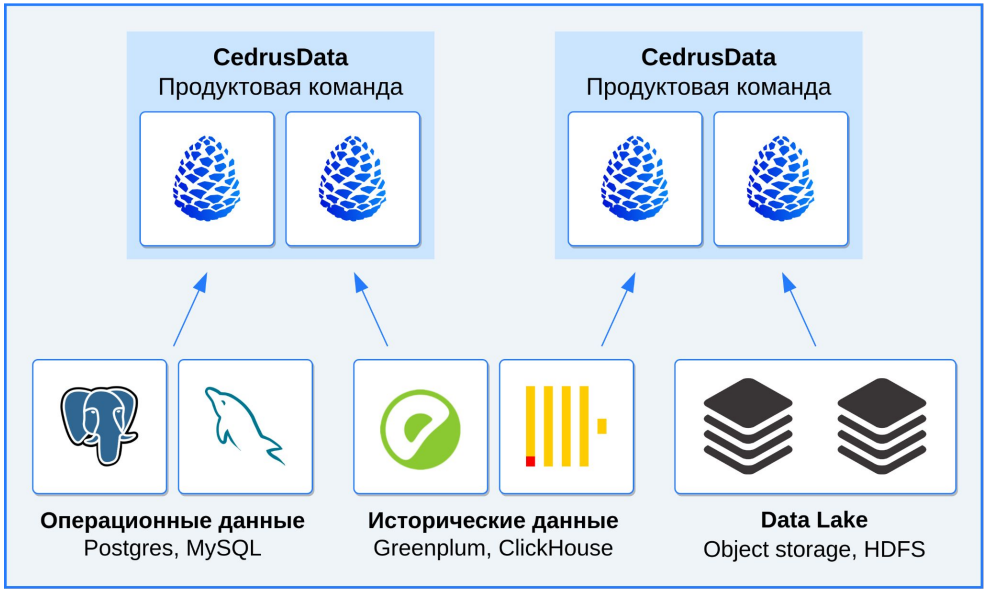
Организации используют озера данных для выполнения задач batch processing (например, построение моделей машинного обучения с помощью Apache Spark), в то время как задачи интерактивной аналитики обычно решаются посредством хранилищ данных.

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

Преимущества:

- Ускорение реализации новых сценариев анализа данных за счет уменьшения потребности в ETL
- Удешевление инфраструктуры за счет уменьшения дублирования данных и переноса нагрузки из корпоративного хранилища в более дешевое озеро данных

# Сценарии: децентрализованная аналитика




Организации обычно используют централизованные платформы данных, в которых все пользователи и приложения делят общие ресурсы кластера, так как архитектура корпоративных хранилищ не позволяет создать несколько кластеров за приемлемые деньги.

Чрезмерная централизация удорожает инфраструктуру и 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.

Преимущества:

- Позволяет продуктовым командам быстро внедрять новые аналитические сценарии
- Позволяет организации гибко управлять расходами на лицензии и инфраструктуру, минимизируя избыточность

# Отличия CedrusData от Trino

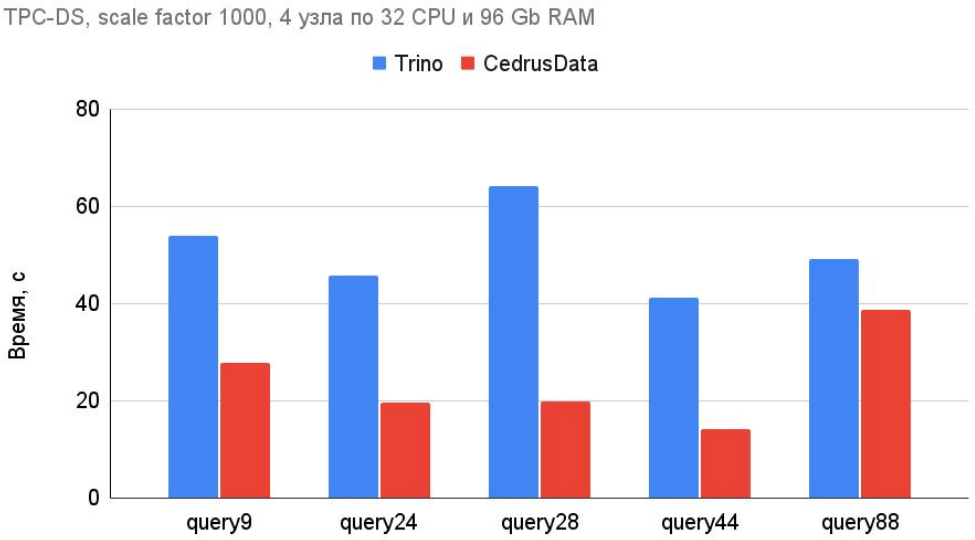
| | Trino | CedrusData |
|---|---|---|
| Базовый функционал Trino | ✓ | ✓ |
| Дополнительные возможности интеграции (Greenplum, Teradata) | | ✓ |
| Улучшения производительности | | ✓ |
| Расширенный мониторинг | | ✓ |
| Оперативное исправление дефектов | | ✓ |
| **Реестр российского ПО** | | ✓ |
| Проверка на отсутствие вредоносного кода | | ✓ |
| Сервисы (поддержка, консалтинг, тренинги) | | ✓ |

# Ориентация на российский рынок

Наша команда создает продукт для российского рынка и внимательно учитывает потребности и запросы отечественных компаний.

- Регистрация в реестре российского ПО (Запись в реестре от 05.12.2022 №15789)
- Проверка совместимости с отечественным системным ПО: Astra Linux, Axiom JDK и др.
- Сервисы для клиентов: помощь в пилотировании, консалтинг, тренинги
- Русскоязычная поддержка
- Публичный roadmap развития продукта на основе потребностей российских заказчиков

# Производительность CedrusData




Команда CedrusData разрабатывает улучшения производительности ядра Trino для более эффективного использования серверных ресурсов. В ряде случаев CedrusData обеспечивает кратное ускорение SQL-запросов по сравне.

Примеры реализованного функционала:

- Локальный дисковый кэш данных [1]
- Кэш метаданных Parquet [2]
- Cost-based оптимизатор запросов [3]

Разрабатываемый функционал:

- Переиспользование повторяющихся операторов [4]
- Материализованные представления (в т.ч. автоматические) [5]
- Выбор оптимального порядка операторов Join [6]

# Контакты

ООО “Кверифай Лабс”
ИНН 7811766769
ОГРН 1217800163790

Конакты:

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