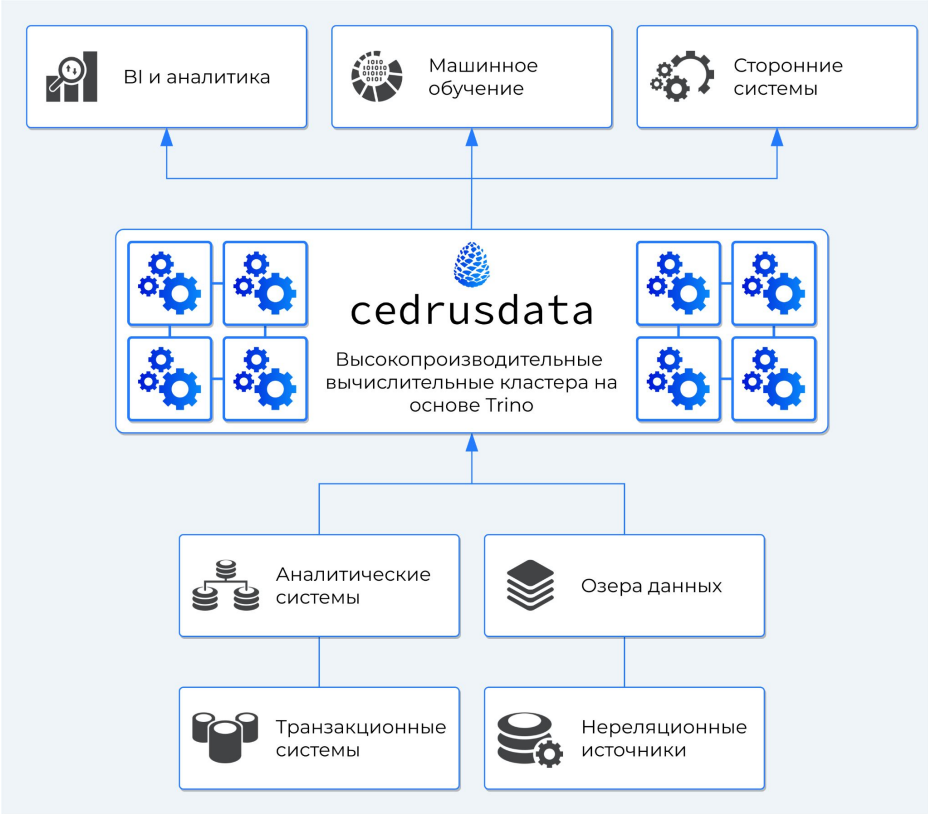
Querify Labs

CedrusData

Высокопроизводительная аналитическая платформа для крупного и среднего бизнеса

Обзор



CedrusData это вычислительная платформа, которая позволяет компаниям быстро и гибко анализировать все свои данные через единую точку доступа.

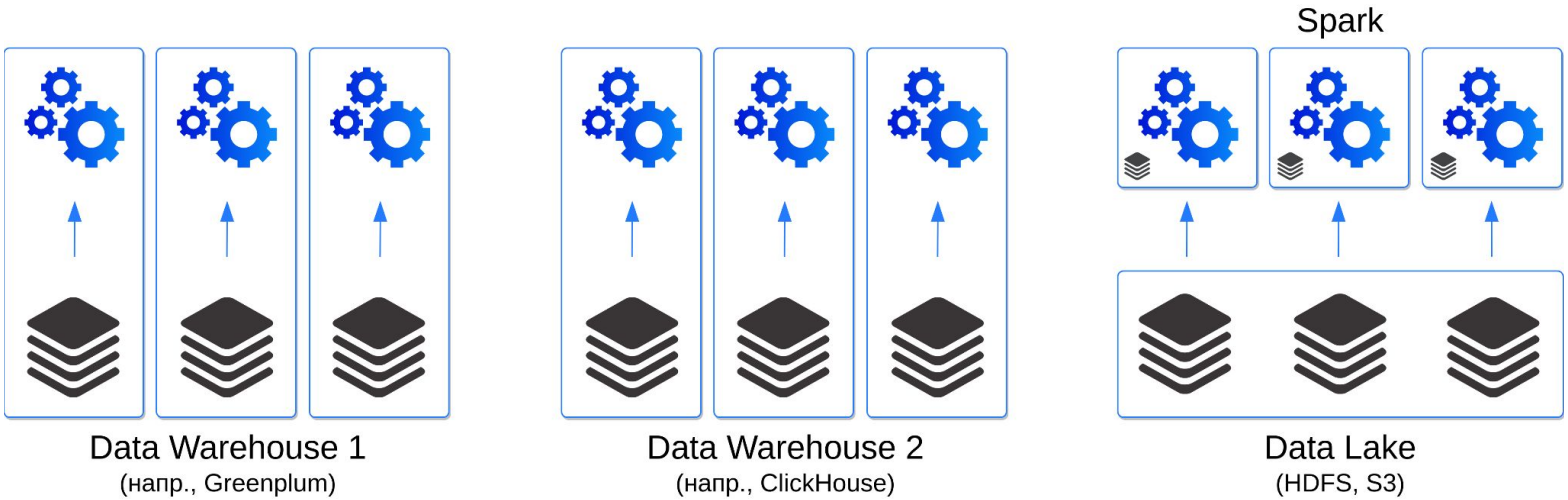
Оптимизирована под работу с облачными инфраструктурами и озерами данных. Проста в развертывании и эксплуатации. Основана на популярном open-source проекте Trino.

Разработкой CedrusData занимаются инженеры компании Querify Labs, которые ранее работали над проектами Apache Ignite, ClickHouse и Yandex Database.

Состояние современной аналитики

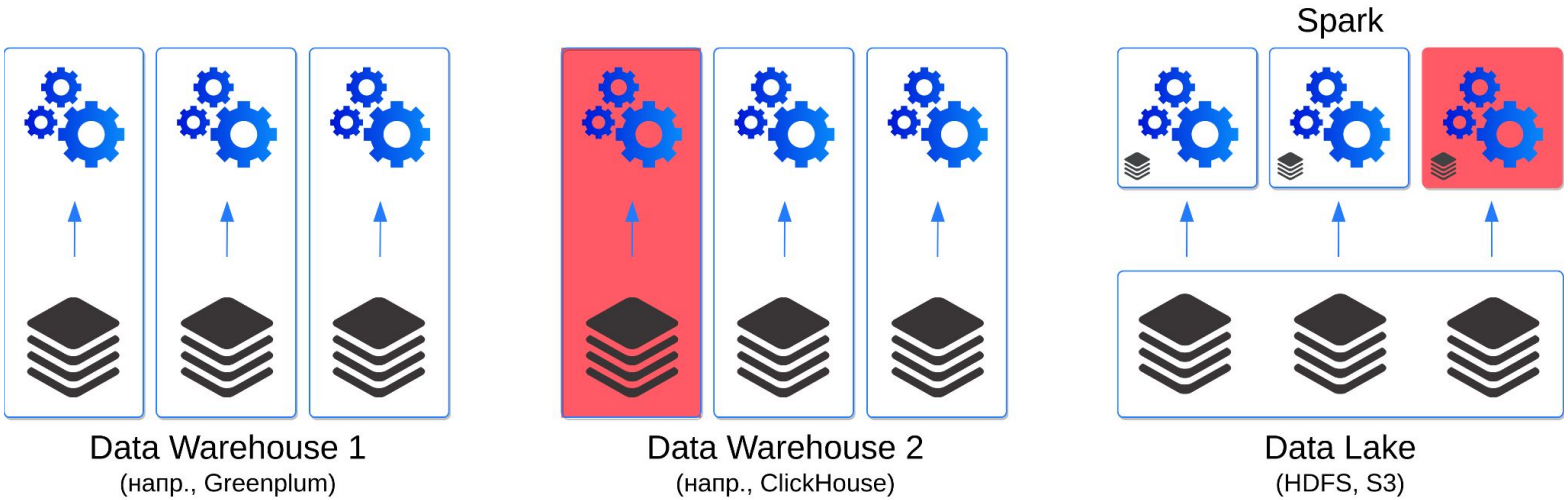
- Продолжающийся рост количества данных (~20% в год). При этом, потребности в storage опережают потребности в compute.
- Увеличение количества источников данных.
- Активный переход в облако.
- Ускорение time-to-market и привлечение большего количества нетехнических пользователей к работе с данными.

Классическая архитектура



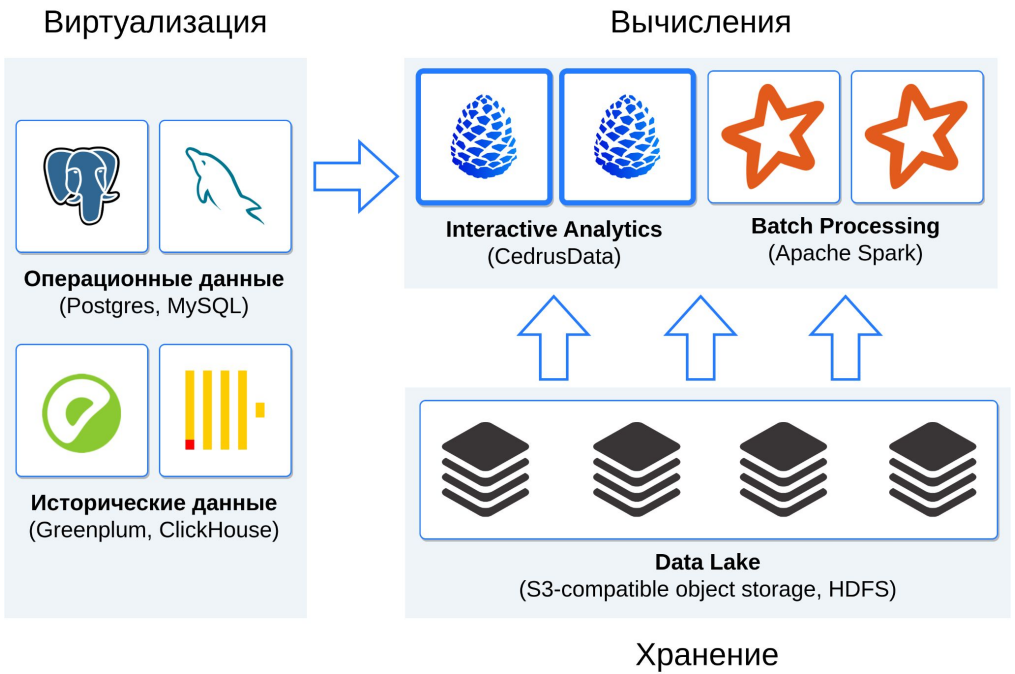
- Аналитическая платформа предприятия обычно включает одно или несколько хранилищ данных (**data warehouse**) и озеро данных (**data lake**).
- Data warehouse хранит данные в подготовленном виде. Подходит для интерактивной аналитики (interactive analytics).
  - Популярные технологии: Greenplum, ClickHouse.
- Data lake хранит данные в сыром виде в дешевом файловом хранилище. Подходит для batch processing и ML.
  - Популярные технологии: Spark.

Классическая архитектура: проблемы



- Затруднен сквозной анализ данных организации.
- Многократное дублирование данных, совмещенных с вычислениями, приводит к неэффективному использованию вычислительных ресурсов (диски, CPU).
- Shared-nothing кластера обладают низкой отказоустойчивостью, что обусловливает необходимость создания сложных инфраструктурных платформенных решений (загрузка данных, разделение ресурсов между пользователями, мониторинг, и т.п.). Это увеличивает TCO и приводит к излишней **централизации** и уменьшению скорости внедрения изменений.

Современный взгляд



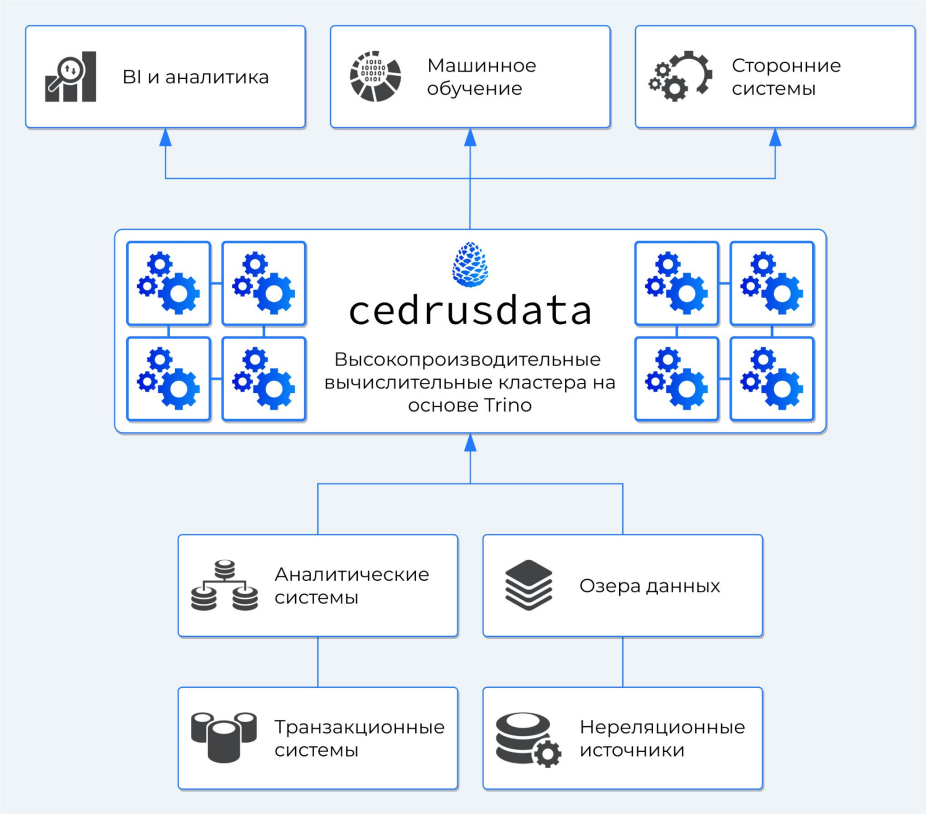
Принципы:

- 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 сквозного 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.

Технология CedrusData



CedrusData это массивно-параллельная распределенная система с SQL-интерфейсом на основе open-source проекта Trino.

Кластер CedrusData состоит из вычислительных узлов, которые обрабатывают SQL-запросы, но не хранят данные. Данная архитектура обеспечивает эластичное масштабирование в облаке и on-premise, и применяется в таких популярных продуктах, как Snowflake и Databricks Lakehouse.

Богатый набор коннекторов позволяет извлекать данные из популярных источников:

- Озера данных под управлением Hive Metastore или Apache Iceberg.
- Хранилища: Greenplum, ClickHouse, Druid, Pinot.
- Реляционные СУБД: Postgres, MySQL/MariaDB, Oracle, SQL Server.
- Нереляционные источники: Cassandra, MongoDB, Redis, Kafka, Elasticsearch, Prometheus.

SQL-интерфейс позволяет подключаться к кластеру CedrusData из популярных BI и ML инструментов, а также любых приложений, поддерживающих интерфейс JDBC.

Технология Trino

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

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

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

- Нацелен на малый, средний и крупный бизнес.
- Развивается преимущественно компанией Starburst.
- Активное сообщество, постоянный поток улучшений.

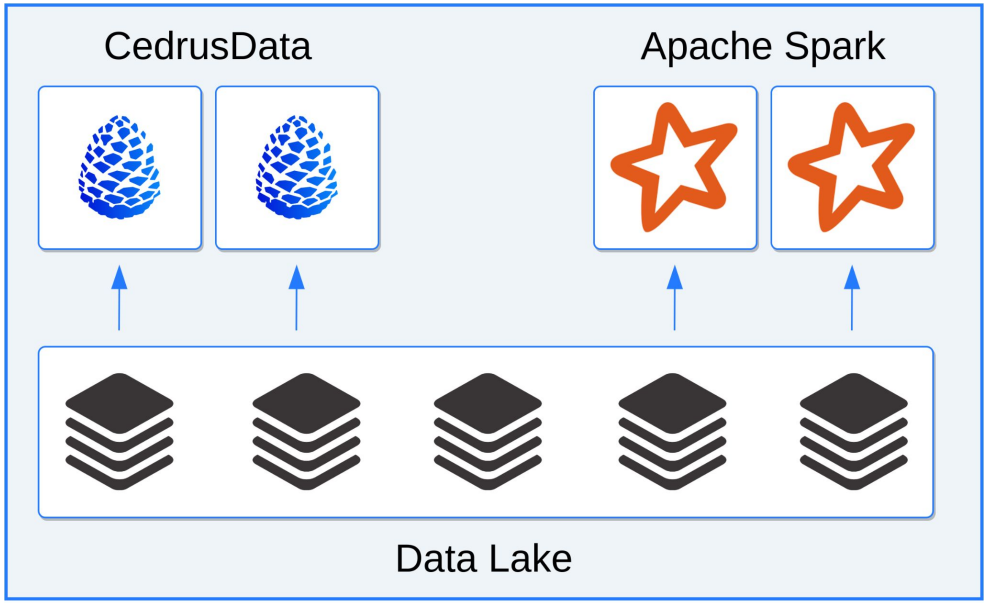
CedrusData основан на Trino.

Преимущества CedrusData перед Trino

	Trino	CedrusData
Базовый функционал Trino	✓	✓
Исправления ошибок и проблем с безопасностью		✓
Улучшения производительности [1]		✓
Расширенный мониторинг		✓
Совместимость с российском ПО Linux и Java [2]	✓	✓
Реестр российского ПО		✓
Проверка на отсутствие вредоносного кода		✓
Поддержка 24x7 и professional services		✓

1. Официальные данные о совместимости Trino с российским ПО не представлены.

Сценарии использования CedrusData

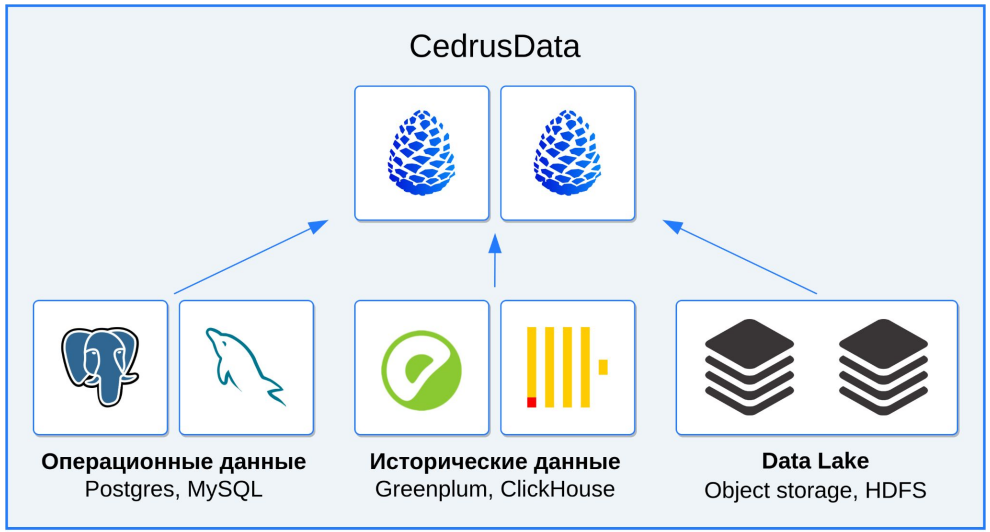


Облако, on-premise

Сценарий 1: интерактивная аналитика поверх озер данных.

- Организация периодически выгружает операционные данные в озеро данных (например, Yandex Object Storage, MinIO).
- Пользователь выполняет задачи интерактивной аналитики с помощью CedrusData, и задачи batch processing с помощью Apache Spark.

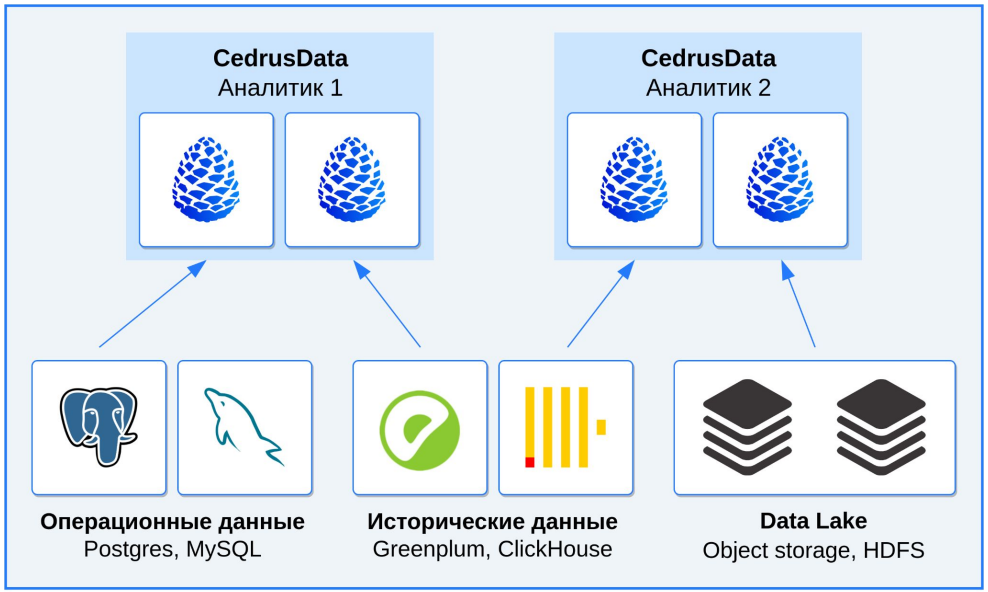
Сценарии использования CedrusData



Сценарий 2: сквозная аналитика всех данных организации (виртуализация).

- Организация использует несколько систем для решения своих аналитических задач. Необходимо объединить данные из разных систем для построения сводных отчетов.
- Пользователь использует CedrusData для написания SQL-запросов, которые объединяют данные из разных систем.

Сценарии использования CedrusData



Облако, on-premise

Сценарий 3: децентрализованная ad-hoc аналитика.

- Аналитик данных использует мощную рабочую станцию или кластер в облаке с большим количеством памяти и CPU.
- Аналитик загружает часть исторических данных из озера данных, подключает другие источники.
- Аналитик использует CedrusData для сквозного ad-hoc анализа данных.

Контакты

ООО “Кверифай Лабс”
ИНН 7811766769
ОГРН 1217800163790

Конакты:

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