Yandex Cloud

Работа с данными и ИИ в Yandex Cloud

Роман Матусевич,
руководитель индустрии
нефтегаза и энергетики в Yandex Cloud

Содержание

Работа с данными: сбор, обработка, хранение и анализ большого объёма данных

Умение работать с данными — необходимое условие для роста любой компании в современном мире

Data-driven-подход
Оперативность и доступность данных
Демократизация данных, ИИ

1. С чем сталкиваются компании при реализации сценариев анализа данных?
2. Почему нужно строить работу с данными в Yandex Cloud?

Хранилище данных — не монолит, а набор связанных сервисов

Сервисы в стеке предстоит тесно интегрировать друг с другом

Помимо аналитического движка хранения и обработки данных, необходимы:

- средства загрузки данных
- средства трансформации данных (ELT)
- инструменты машинного обучения
- средства предоставления доступа к данным (BI и другие)

Интеграция сервисов требует экспертного опыта и знаний в области каждого из них

В процессе интеграции возможны серьёзные проблемы: часто они возникают уже во время эксплуатации продукта

Задача доставки данных из источников сложна и не имеет решения под ключ

Необходимо обеспечить минимальную задержку при доставке данных из источников в хранилище

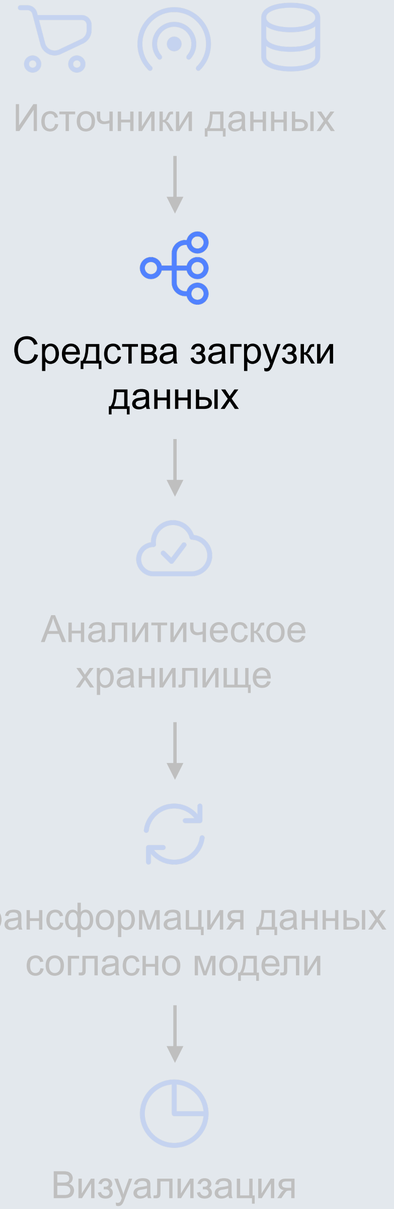
При этом нежелательно или невозможно дополнительно нагружать системы-источники

Подход CDC* позволяет обеспечить минимальную задержку и нагрузку на источник, но коммерческие реализации такого подхода крайне дороги, а доступность и поддержка ограничены

- Oracle GoldenGate
- Informatica CDC
- Qlik Replicate

Open-source-реализации подхода CDC не обладают необходимой стабильностью и требуют колоссальных ресурсов для эксплуатации

- StreamSets
- Debezium



* Change Data Capture

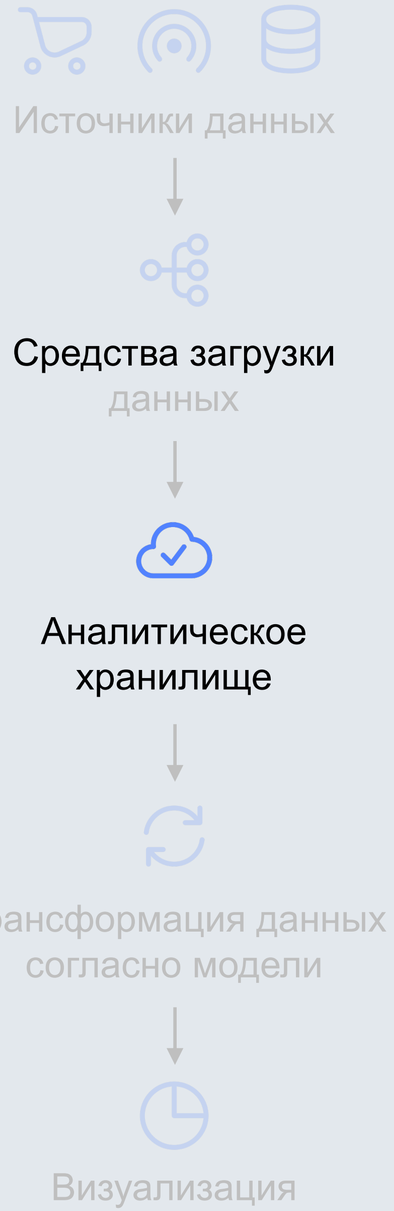
Масштабирование хранилища данных отстаёт от роста бизнеса

Данные могут расти скачкообразно и непредсказуемо: например, при расширении бизнеса, сезонных колебаниях или глобальных изменениях.

Часто масштабирование хранилища не может обеспечить нужную скорость

3–8 месяцев

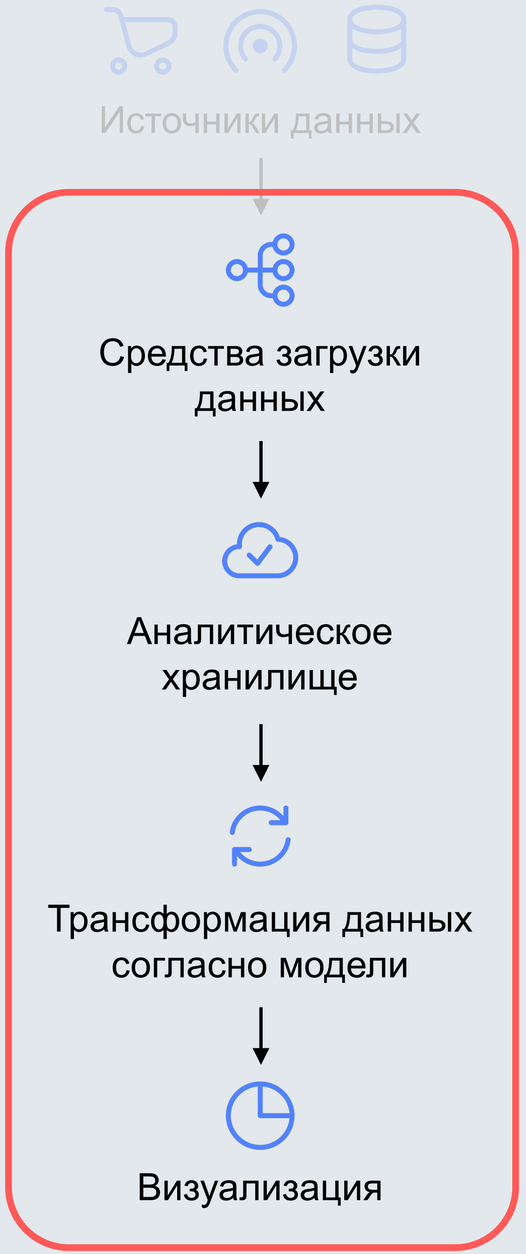
составляет срок поставки серверов, раскатки необходимого ПО и введения их в эксплуатацию



Закрытая экосистема (Vendor Lock)

Часто используются legacy-системы
с закрытым исходным кодом

- Дорогое решение со временем становится ещё дороже из-за изменений внешних условий (курс валют, ограничение доступности)
- Сложно договориться о доработке такого решения под задачи бизнеса
- Найти специалистов для обслуживания систем сложно, а обучение стоит дорого
- Интеграция с решениями часто сложная и затратная
- Legacy-системы могут тормозить развитие смежного ландшафта
- Сложность миграции с таких систем



Для эксплуатации инфраструктуры DWH нужны экспертные знания и опыт

Потребность в экспертах приводит к необходимости строить непрофильный центр компетенции в компании

Квалифицированных кадров не хватает: специалисты редки и стоят дорого

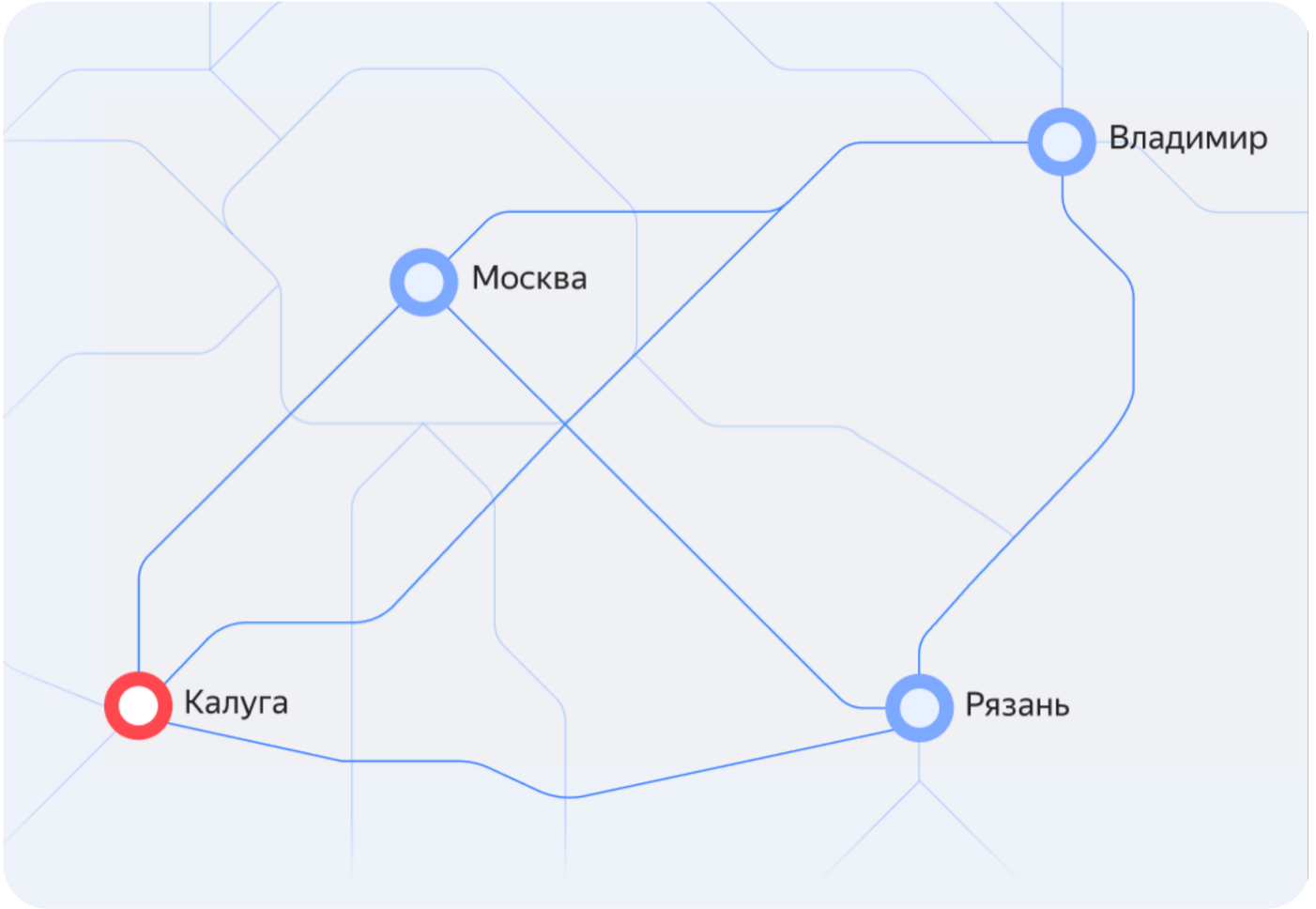
Поиск и обучение — это долго и дорого, а переучивание лишь увеличивает издержки

1. С чем сталкиваются компании при реализации сценариев анализа данных?

2. Почему нужно строить работу с данными в Yandex Cloud?

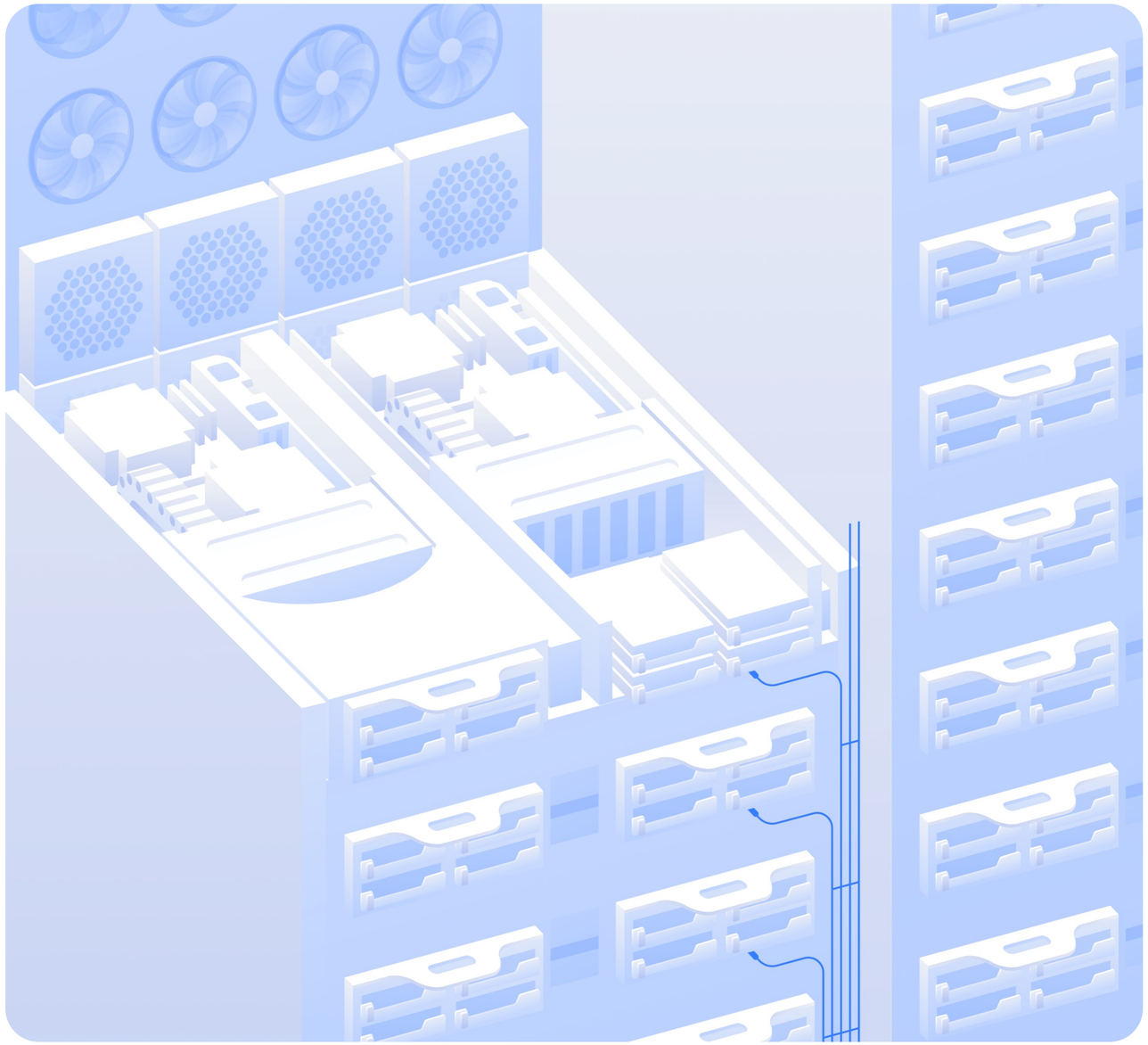
Собственная физическая инфраструктура

- Три зоны доступности
- Дата-центры на расстоянии 300 км друг от друга
- Новый дата-центр в Калуге Скоро!
- Независимое энергоснабжение в каждом дата-центре
- Терабитная полоса пропускания обеспечивается собственной оптоволоконной DWDM-сетью



Серверное оборудование собственной разработки

- Серверные стойки разработаны под дата-центры и наоборот
- Единообразие аппаратного обеспечения: работаем с разными вендорами
- Внутреннее аппаратное обеспечение в нужном интервале температур
- Нагрузка до 500 Вт на сервер
- Режим горячей замены для дисковых накопителей



Платформа Yandex Cloud — единый хаб новых технологий

На слайде указано количество сервисов в каждом направлении

Yandex Cloud — это оптимальная аналитическая платформа

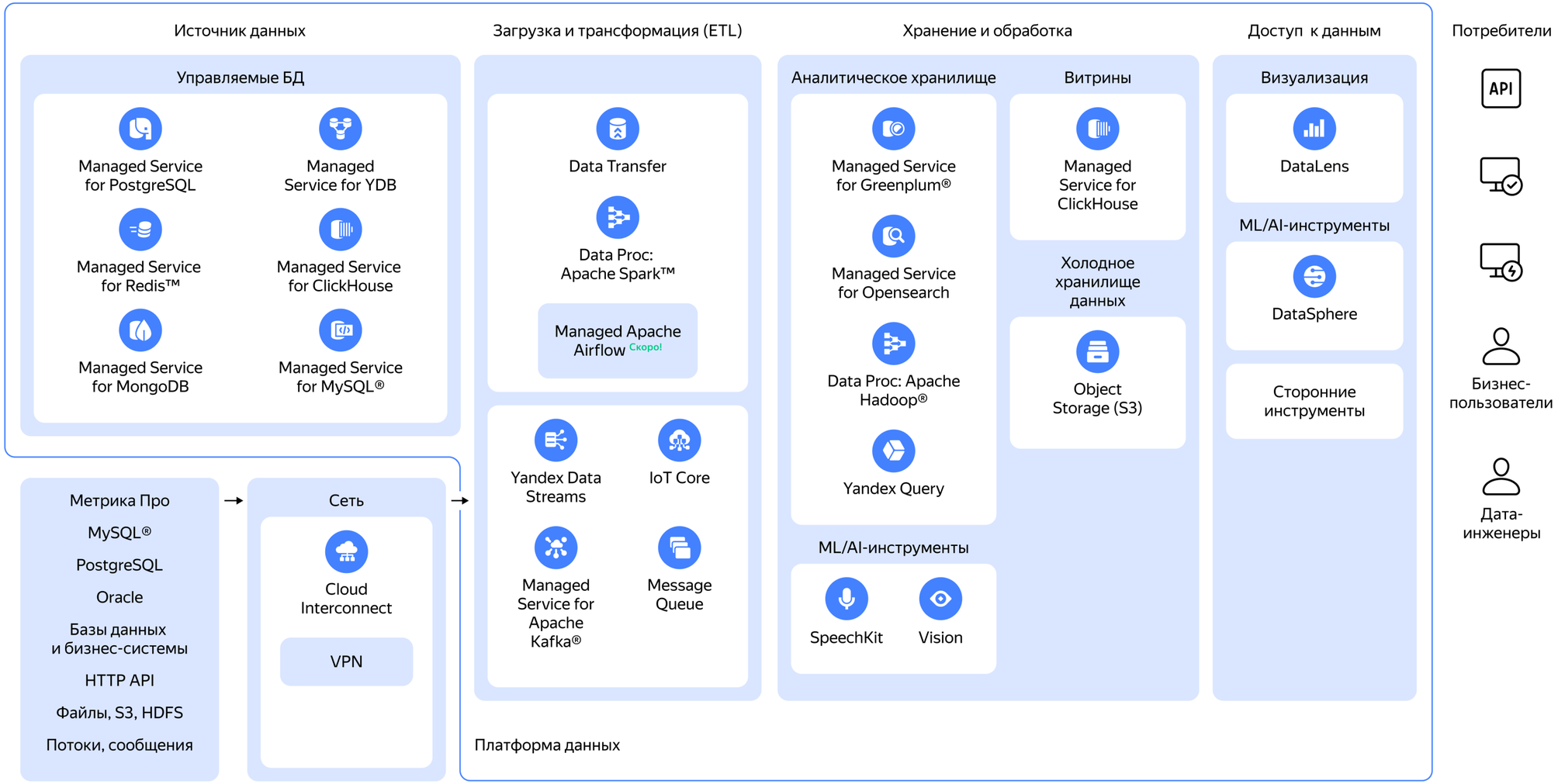
Ключевые компоненты хранилища данных интегрируются друг с другом без написания кода

- Источники
  в ваших ЦОДах,
  в облаках
- Аналитические СУБД
- BI
- Machine Learning
- Холодное хранилище

Техническая поддержка всего ИТ-ландшафта — сервисов и интеграций между ними

Технологии с открытым кодом (open-source): привлекаем партнёров для решения прикладных задач

Платформа данных Yandex Cloud



Отличия Managed Service for Greenplum®
от Data Proc и Managed Service for ClickHouse

Data Proc: храним сырые данные в Object Storage (S3)

Аналитика с ML

Очистка сырых данных

Простая агрегация сырых данных

Хранение истории сырых данных

Подробнее

Managed Service for Greenplum®: строим витрины данных

Преобразование данных согласно модели

ETL/ELT

Подготовка витрин данных

Сложная ad-hoc-аналитика небольшого количества конкурентных пользователей

Подготовка витрин данных

Managed Service for ClickHouse: выполняем запросы пользователей по витринам

Доступ к витринам большого числа пользователей

Простая ad-hoc- и BI-аналитика

Кросс-ДЦ-резервирование витрин

Подготовка витрин данных

CDC- и ETL-движок, доступный как сервис: Data Transfer

CDC и ETL бесплатно

При размещении Data Warehouse в Yandex Cloud

Вариативность источников

Oracle

PostgreSQL

MySQL

MongoDB

ClickHouse

Быстрый старт

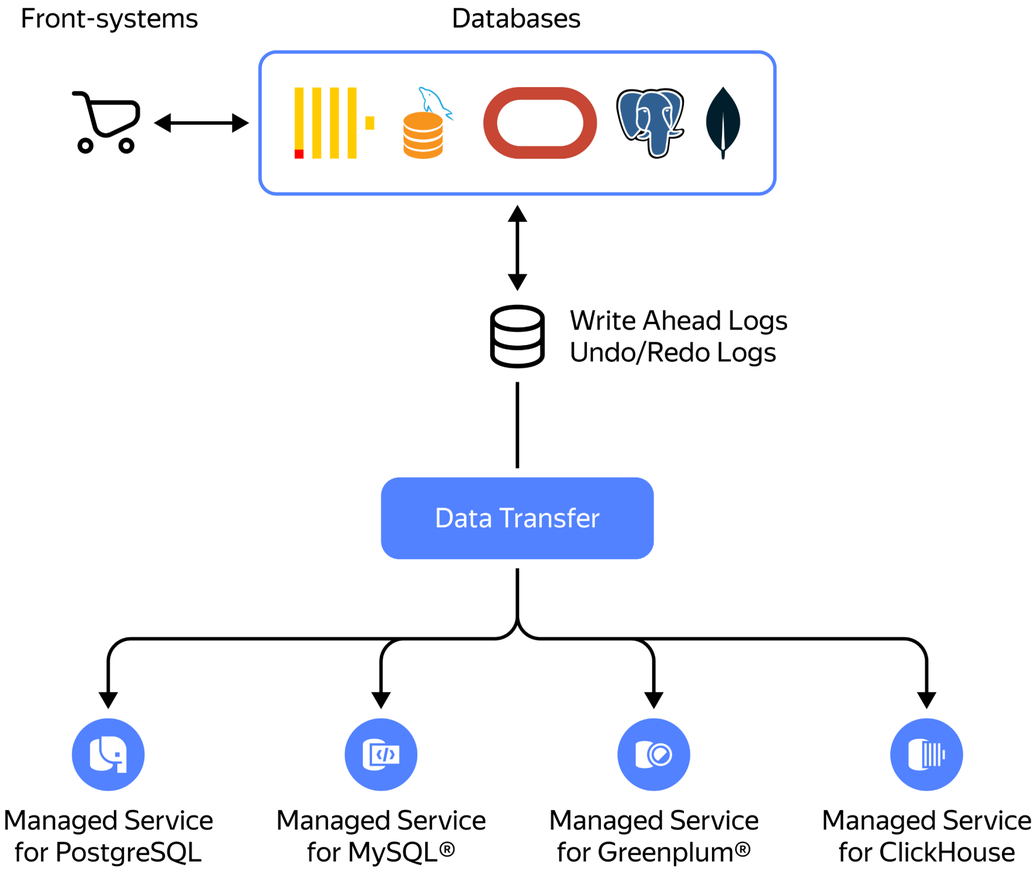
Первичная выгрузка и online-репликация за минуты

Гранулярность

До отдельных таблиц

Гибкий перенос схемы данных

(DDL)



Хранилище растёт вслед за данными

Масштабирование

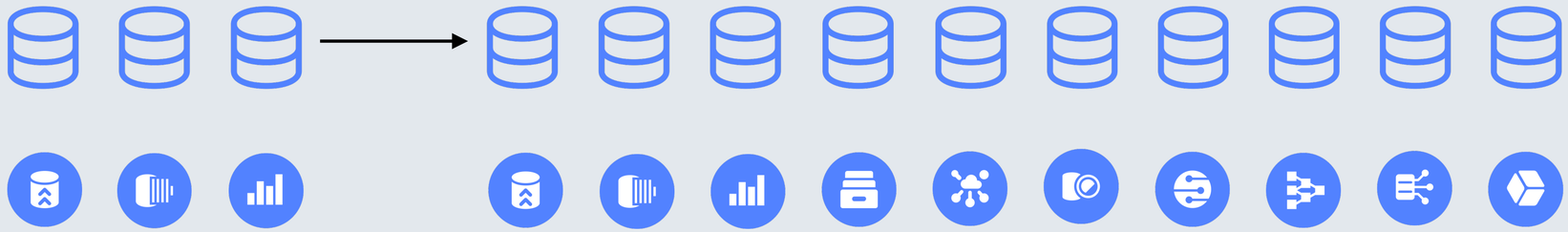
Масштабирование без задержки: требуемые мощности не нужно заказывать заранее

Все компоненты хранилища горизонтально масштабируются через консоль за минуту

CDC, аналитические СУБД, BI, Machine Learning, холодное хранилище

Добавление новых сервисов при изменении характера нагрузки не требует новых экспертных знаний

Например, при внедрении потоковой аналитики на базе Apache Kafka®



Масштабирование

Гибкое управление стоимостью платформы данных

Гибридное хранение данных в Yandex Cloud

Охлаждение данных в Object Storage

Hybrid Storage в Managed ClickHouse (гибридное хранилище)

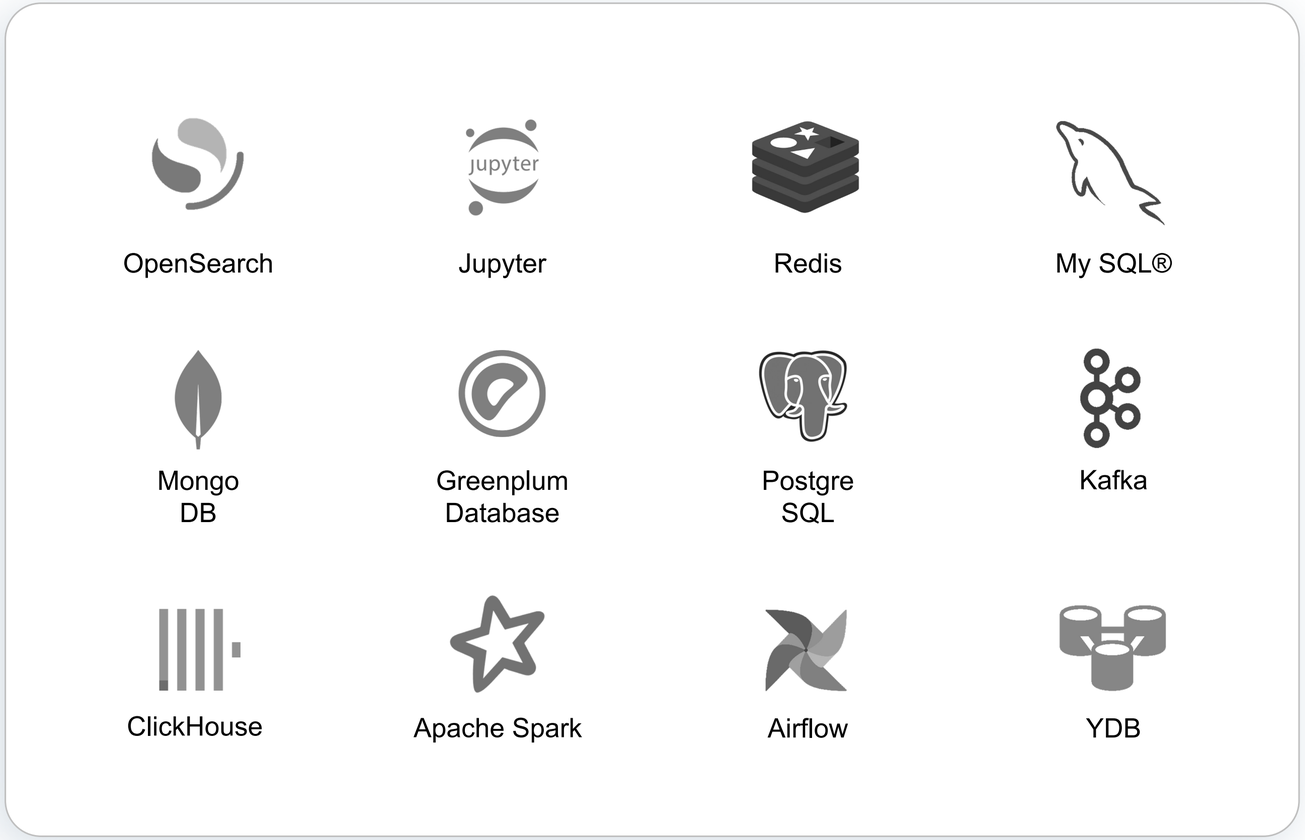
Yezzey в Managed Greenplum

Плата только за потребление

Временные кластеры Data Proc

Serverless

Самый большой портфель сервисов с открытым исходным кодом*



Независимость от вендора

Открытый исходный код у ключевых компонентов хранилища

Проверенные сервисы

Используются в проектах по всему миру, например в Uber, CERN, а также в сервисах Яндекса

Широкий выбор экспертов

В России легко найти специалистов с опытом работы в сервисах, доступных в Yandex Cloud

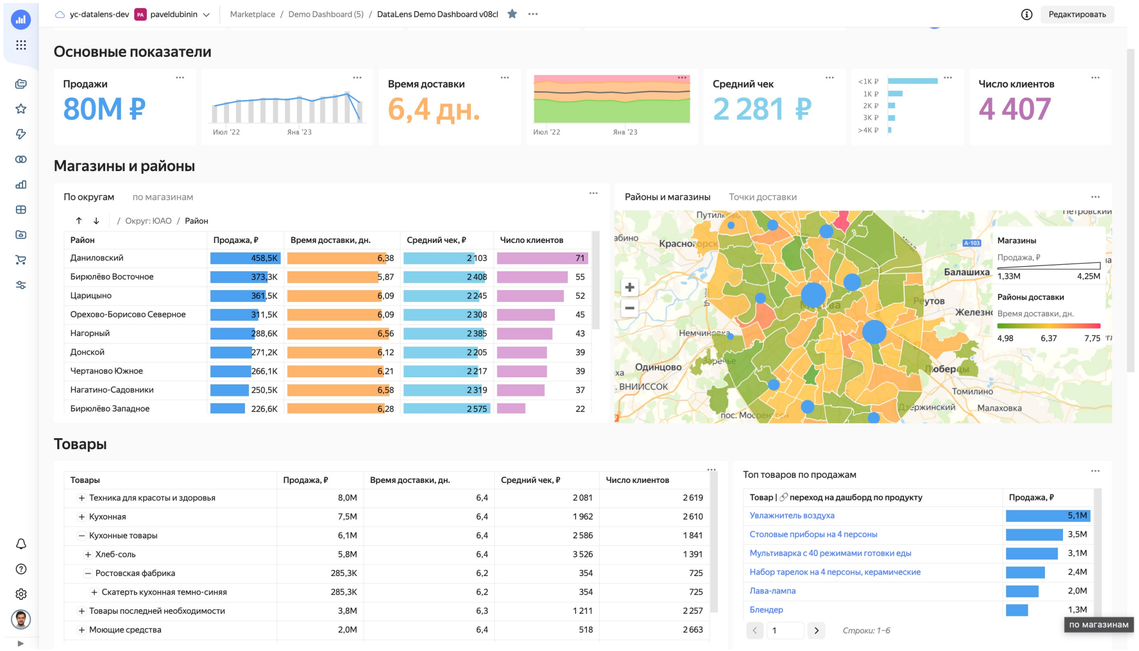
* По России и СНГ

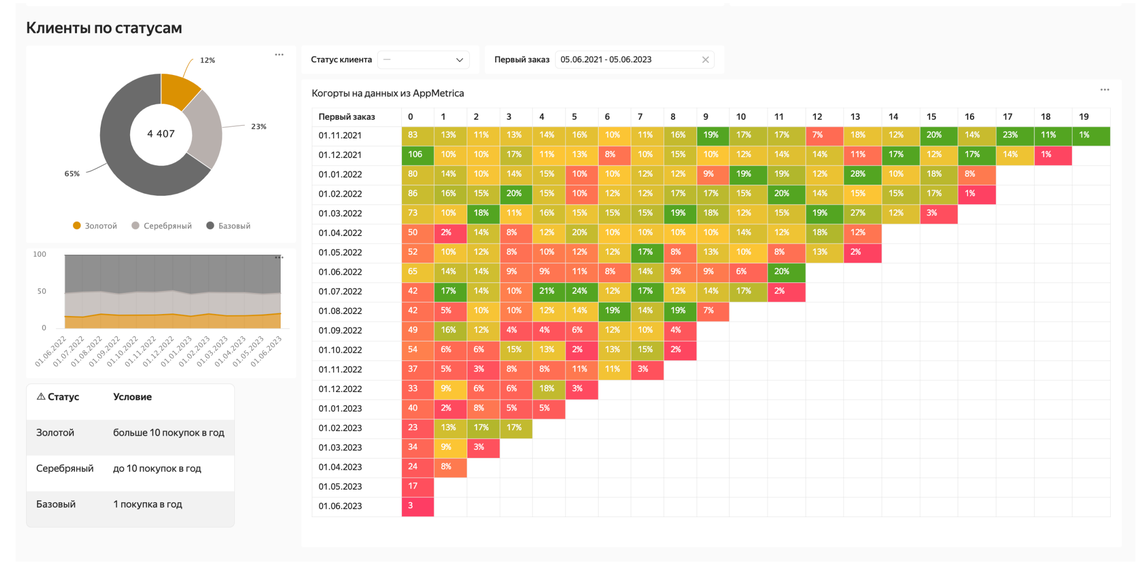
Yandex DataLens

Доступная аналитика любых масштабов!

- Корпоративный стандарт BI во всем Яндексе
- Больше 1000 внешних компаний-пользователей
- Нативный и самый родной инструмент визуализации для ClickHouse
- Быстрое внедрение и меньший TCO по сравнению с классическими BI
- Описание модели данных: джойны, агрегации, трансформации, фильтрации
- Pushdown большинства операций в БД
- Гибкая ролевая модель, RLS, интеграция с LDAP

Демодашборд: datalens.yandex/9fms9uae7ip02





Отвечаем требованиям 152-ФЗ и индустриальных стандартов

152-ФЗ, УЗ-1

Аттестат соответствия
по требованиям
приказа ФСТЭК № 21

PCI DSS

Для ЦОД и облачных сервисов

GDPR

Общий регламент о защите
данных в Европейской зоне

Cloud Security Alliance

Security, Trust, Assurance
and Risk (STAR) по Level 1

Реестр программного обеспечения

Запись в реестре
№ 9286 от 20.02.2021

Стандарты ISO

ISO 27001, ISO 27017
и ISO 27018

bsi. ISO/IEC 27001 Information Security Management | ISO/IEC 27017 Security Controls for Cloud Services | ISO/IEC 27018 Protection of Personally Identifiable Information

ГОСТ Р 57580.1-2017

Безопасность
финансовых операций

Крупнейшие инсталляции DWH на базе Yandex Cloud

Вместе с клиентами мы прошли путь построения и развития DWH, data lake — самых крупных инсталляций на базе Managed Service for Greenplum®

М.ВидеоЭльдорадо

> 200 ТБ

в Managed Service for Greenplum®

Банк из числа пяти крупнейших

100 ТБ

в Managed Service for Greenplum®

E-commerce-компания из числа пяти крупнейших

50 ТБ

в Managed Service for Greenplum®

ВТБ

Создание гибридного озера данных

Задача: начать использовать технологии машинного обучения для более глубокого анализа внутренних и внешних данных

Решение: создали внешнее хранилище «холодных» данных и среду для быстрого прототипирования витрин и отчётов

Результаты: упростился путь применения новых библиотек машинного обучения и тестирования новых технологий. Кратно выросла скорость получения доступа к данным и проверки гипотез дата-сайентистами. Значительно снизилась стоимость владения хранилищем

1 ТБ

новых данных загружается ежедневно

100+

источников обрабатывается озером данных

в несколько раз

выросла скорость получения доступа к данным и проверки гипотез

Managed Service for Kubernetes® Managed Service for ClickHouse Managed Service for PostgreSQL Managed Service for MongoDB

Фокусируйтесь на работе с данными, а не на обслуживании инфраструктуры

Надёжность, производительность и безопасность управляемых сервисов — наш приоритет

1

Сосредоточьтесь на главном — архитектуре и модели данных

2

Исключите потребность развивать непрофильные компетенции внутри команды

3

Мы возьмём на себя:

информационную безопасность

мониторинг

резервирование

бэкапы

обновление

Подайте заявку на получение гранта до 1 млн рублей

clck.ru/34uroH

Спасибо за внимание!
Вопросы?

Роман Матусевич
Руководитель индустрии
нефтегаза и энергетики в Yandex Cloud