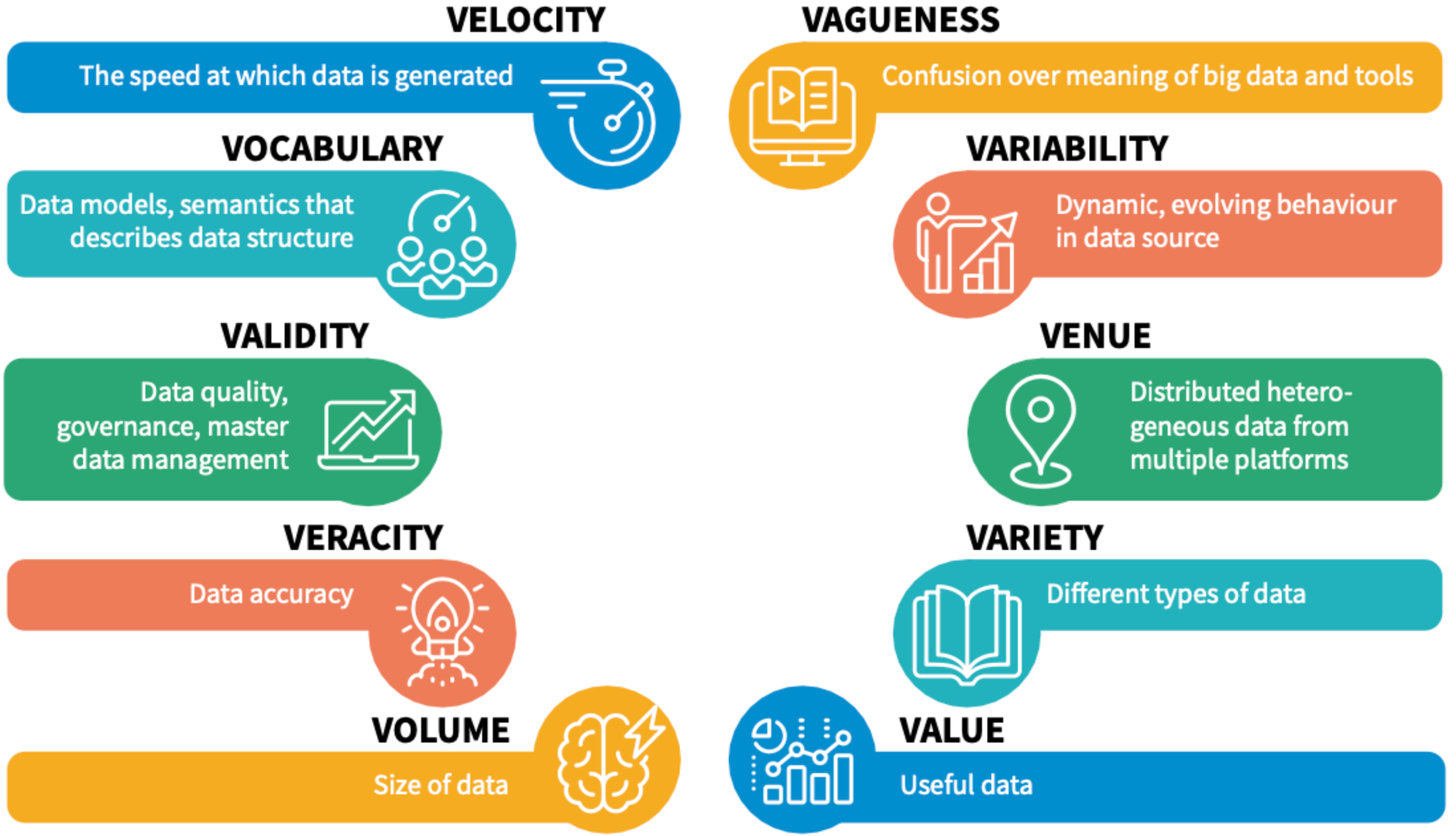
Yandex Cloud

Работа с большими данными в медицине

на платформе Yandex Cloud

Евгений Попов,
руководитель направления
Здравоохранение Yandex Cloud

xV of Big Data



Regulatory considerations on artificial intelligence for health, World Health Organization, 2023

Примеры архитектур хранения данных

Data Warehouse (DWH)

Единое корпоративное хранилище с обработанной и структурированной информацией. Хранилище упрощает анализ полученных данных, но требует структурированности.

Data Vault

Одна из моделей хранилища Data Warehouse с временными отметками размещения данных, которые позволяют проследить изменение хранимой информации во времени.

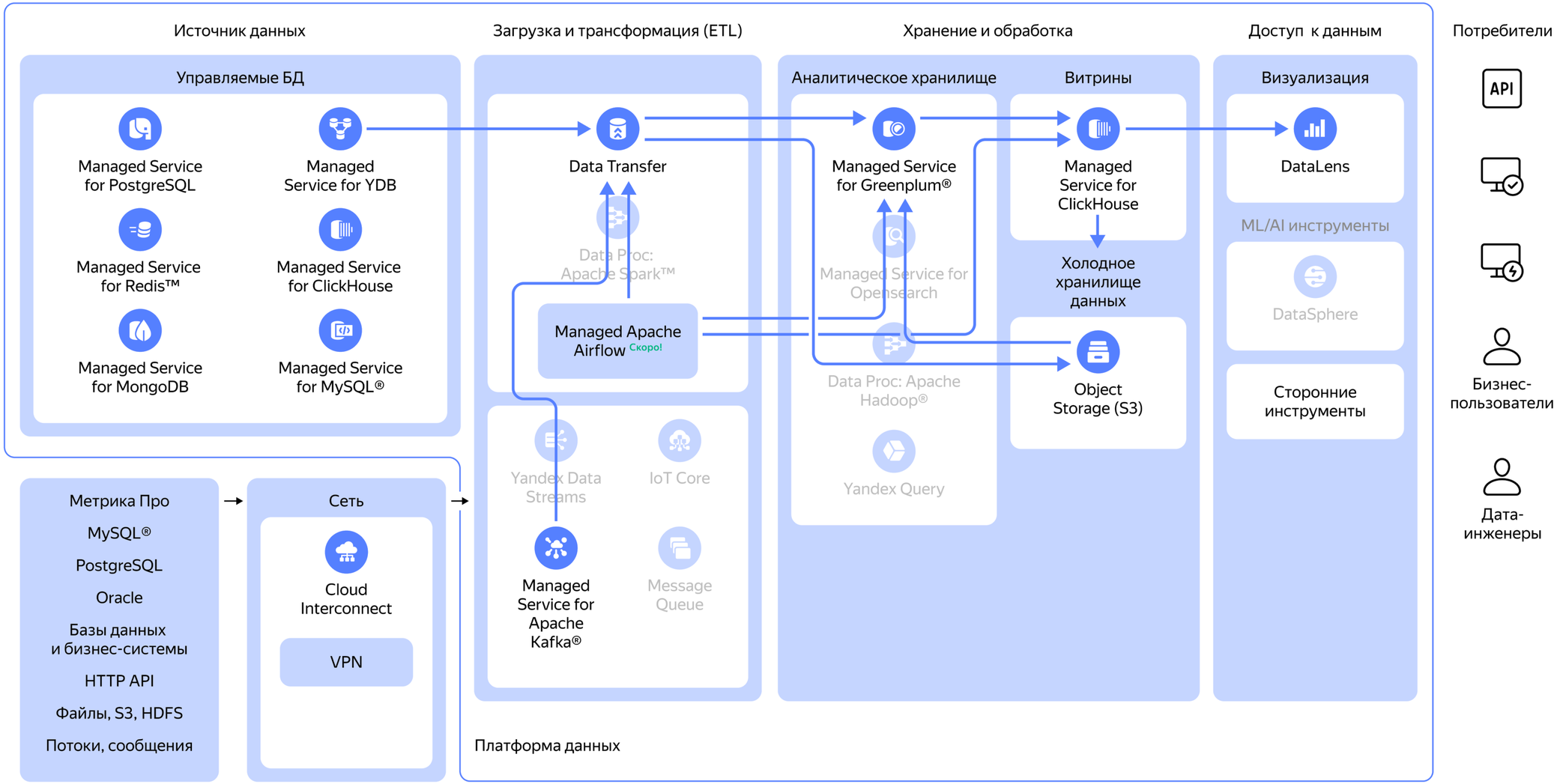
Data Lake

Данные в хранилище поступают непрерывно в неструктурированном или, наоборот, структурированном или слабоструктурированном виде. Используется для сбора данных из разных источников в режиме реального времени.

Data Mart

Хранилище данных, предназначенных для повседневного использования. Поступающую информацию необходимо тщательно обрабатывать, но после этого к ней проще регулярно обращаться

Современное хранилище данных предусматривает смешанные сценарии его использования, рост объёмов данных и их новые источники



С чем сталкиваются компании при внедрении или модернизации хранилища данных?

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.

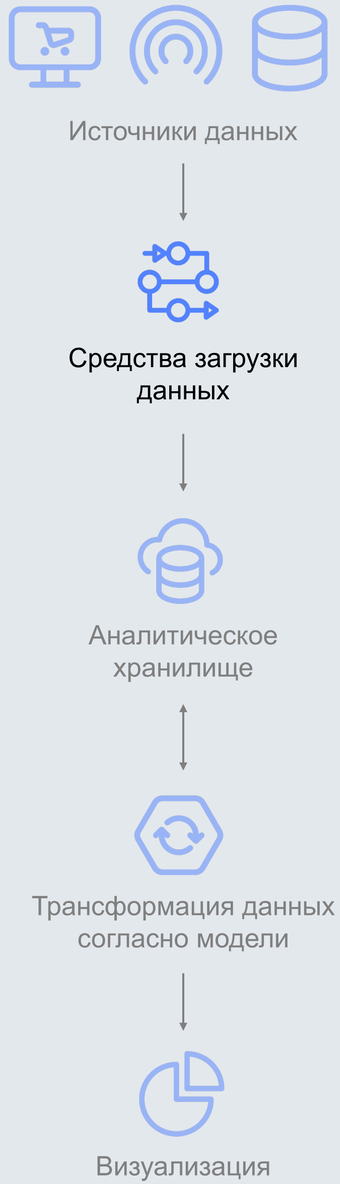
При этом нежелательно или невозможно дополнительно нагружать системы-источники.

Подход CDC* позволяет обеспечить минимальную задержку и нагрузку на источник, но коммерческие реализации такого подхода крайне дороги.

- Oracle Golden Gate
- 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.

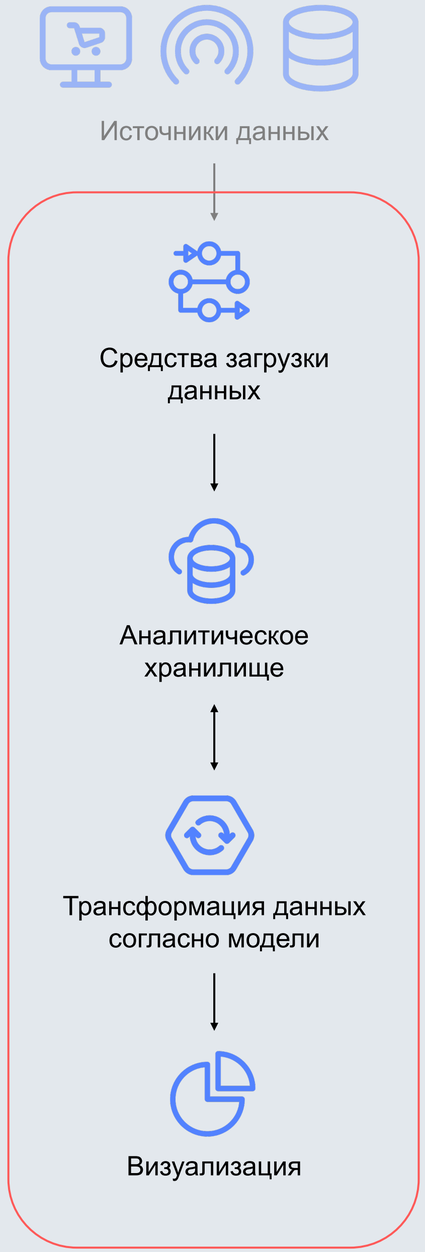
3–8 месяцев

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, обеспечение от поставщика)
- 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

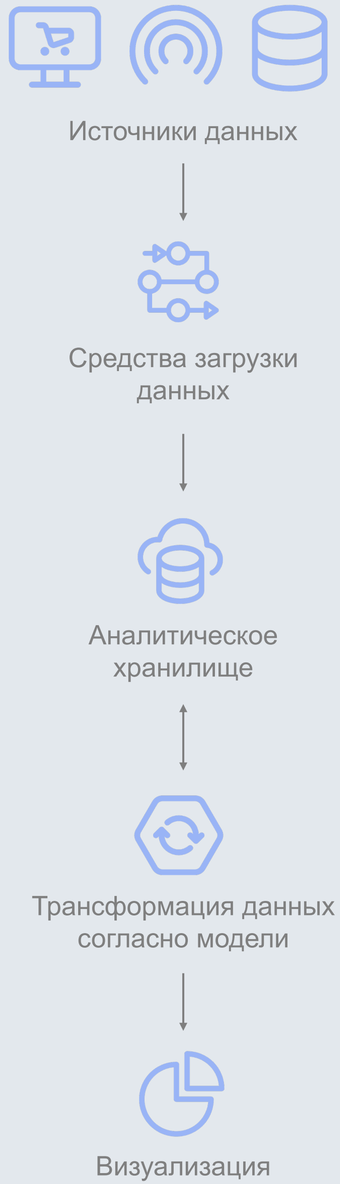


Для эксплуатации инфраструктуры DWH нужны экспертные знания и опыт

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



Компании не могут
быстро разворачивать
и оптимизировать
data-проекты

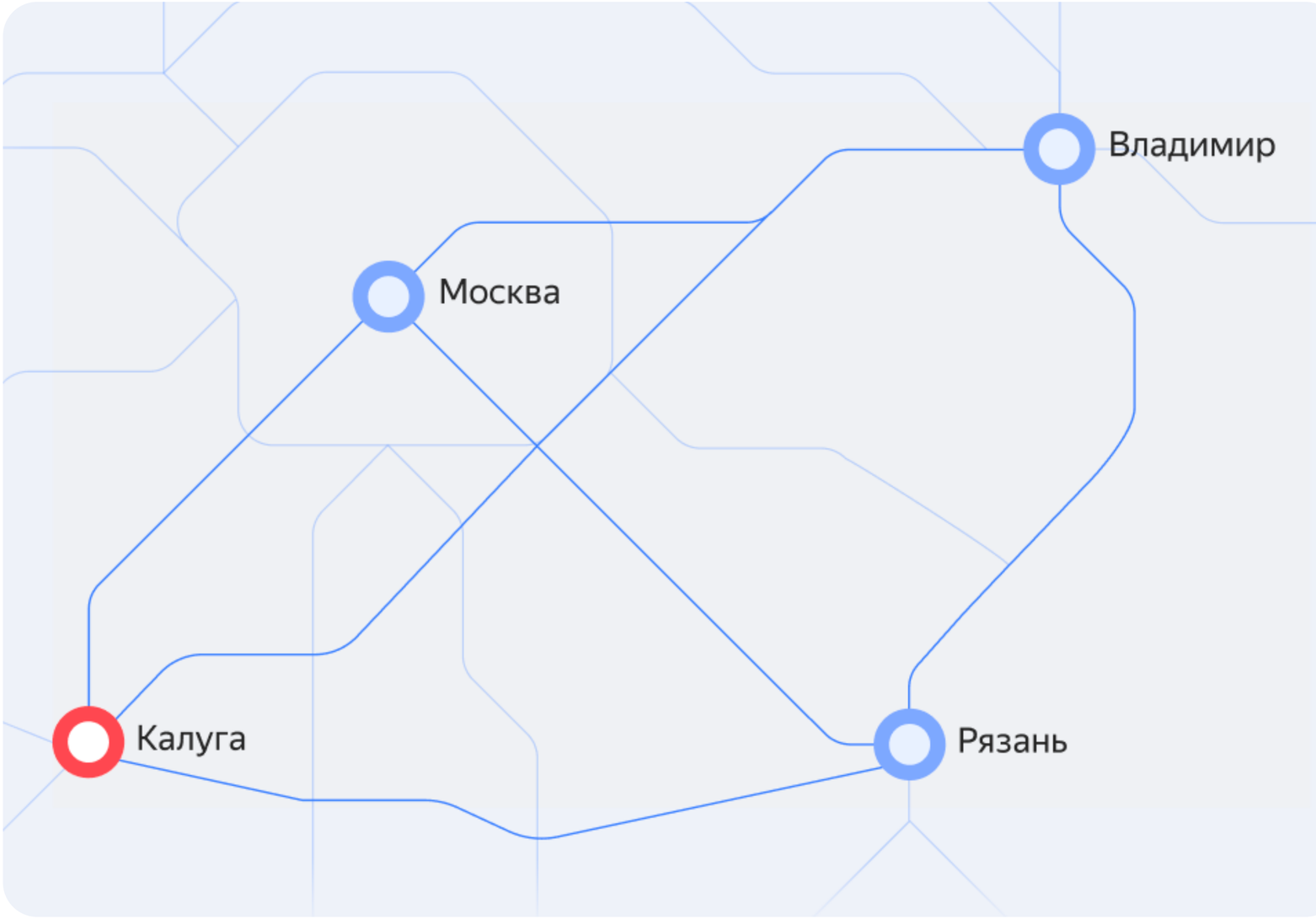
Компании не могут быстро разворачивать и оптимизировать data-проекты

5 сервисов в среднем нужно разрабатывать, поддерживать и интегрировать между собой в архитектуре проекта

Почему хранилище данных в Yandex Cloud?

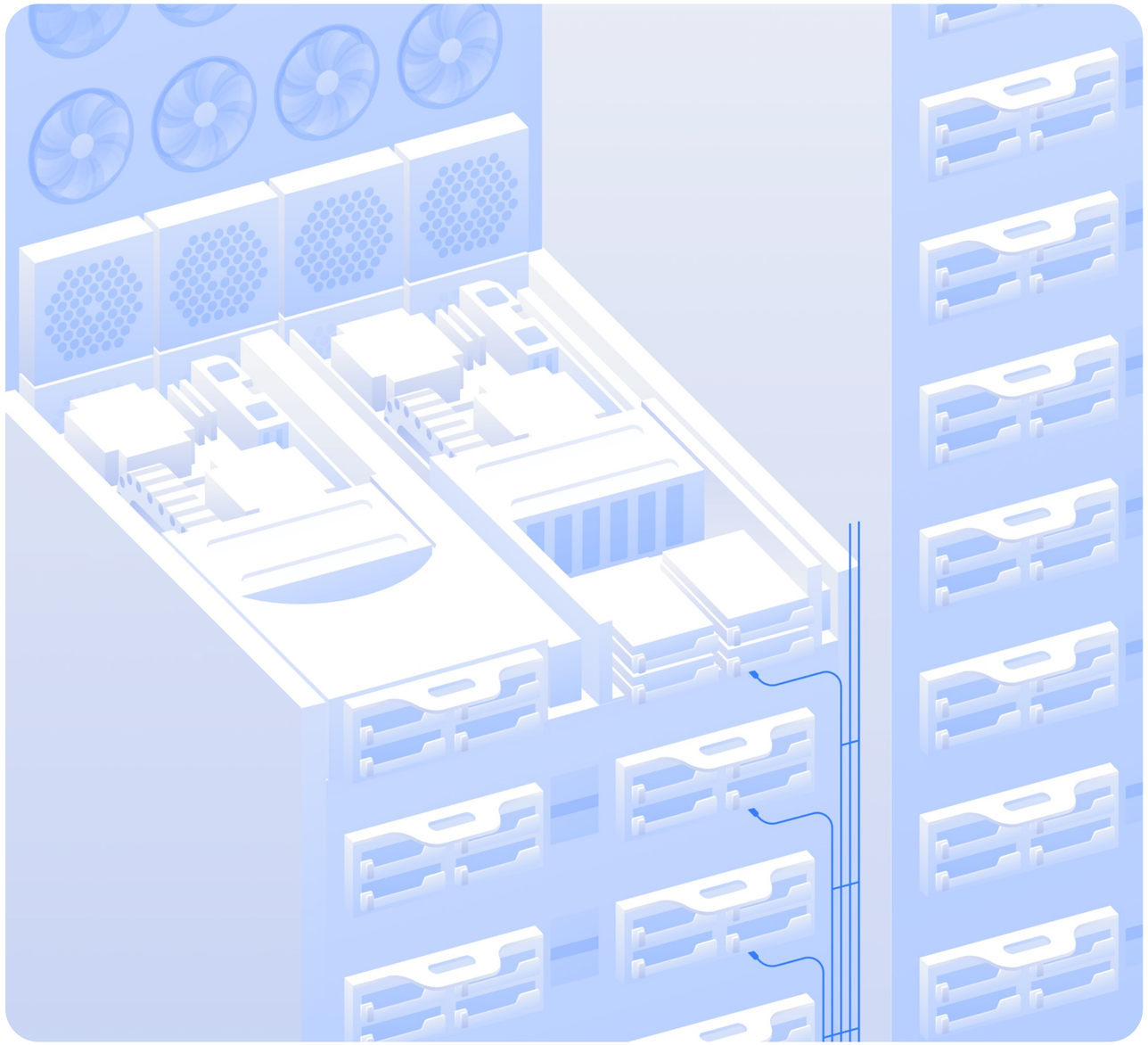
Собственная физическая инфраструктура

- 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

Всестороннее обеспечение безопасности инфраструктуры и защита данных

Cloud Security Alliance

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

ГОСТ Р 57580.1-2017

Безопасность финансовых операций

Реестр программного обеспечения

Запись в реестре
№ 9286 от 20.02.2021

152-ФЗ, УЗ-1

Аттестат соответствия по требованиям приказа ФСТЭК № 21

Стандарты ISO

ISO 27001, ISO 27017, ISO 27018 и ISO 27701 New

Стандарты PCI

PCI DSS для ЦОД и облачных сервисов, PCI PIN и PCI 3DS

GDPR

Общий регламент о защите данных в Европейской зоне

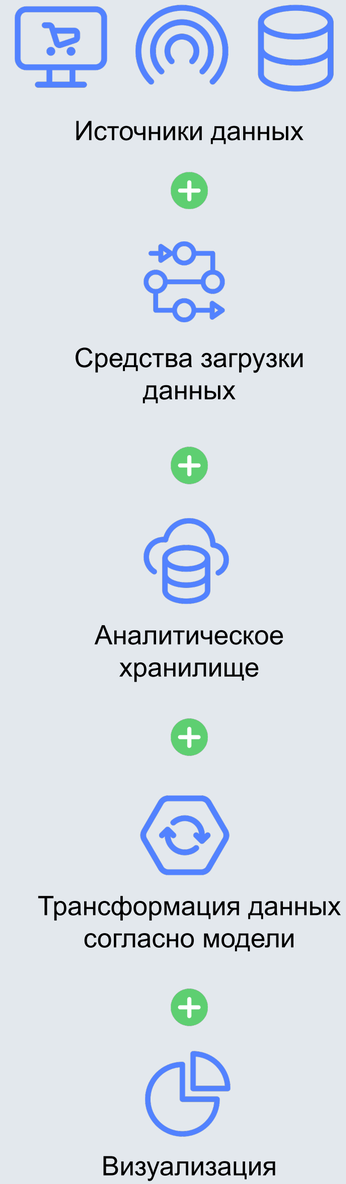
DWH в Yandex Cloud — это экосистема, а не набор отдельных сервисов

Ключевые компоненты хранилища данных интегрируются друг с другом без написания кода:

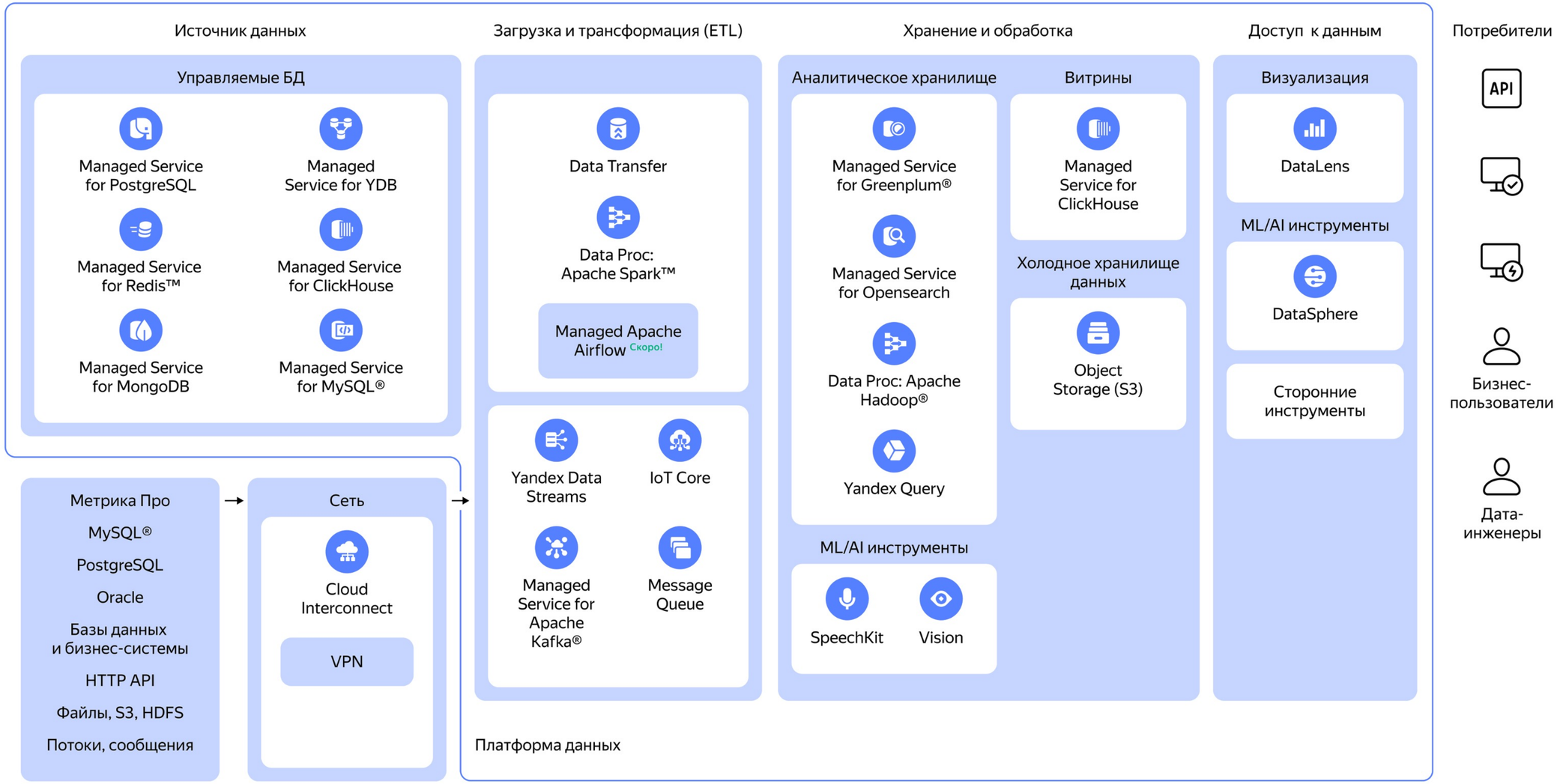
- Источники
  в ваших ЦОДах,
  в облаках
- Аналитические СУБД
- BI
- Machine Learning
- Холодное хранилище

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



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

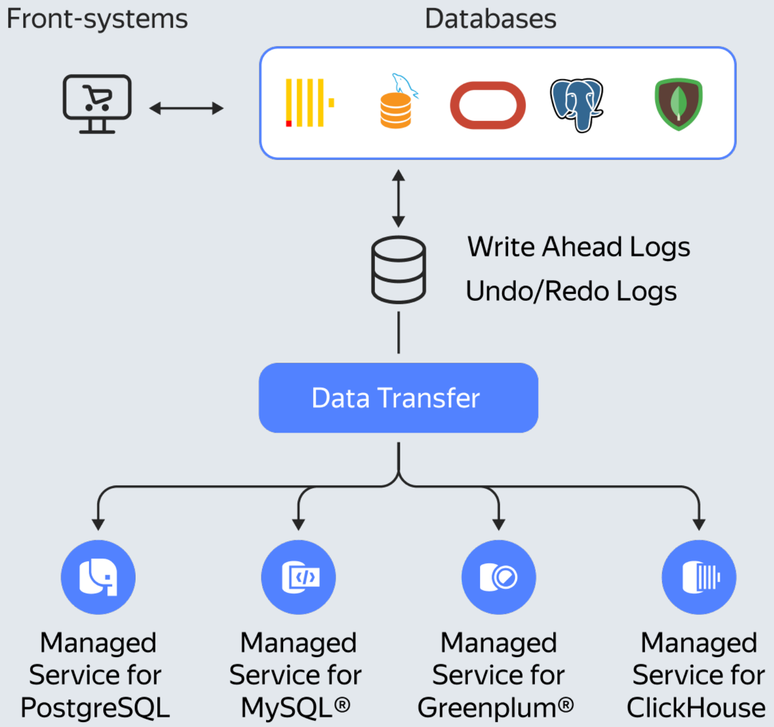
CDC и ETL бесплатно при размещении Data Warehouse в Yandex Cloud

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

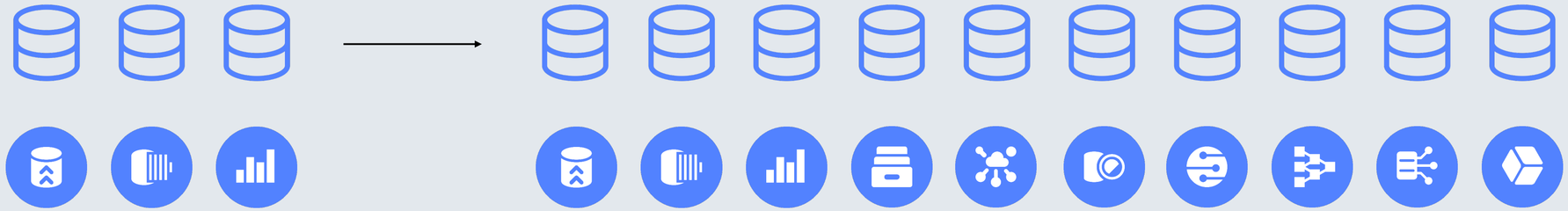
Масштабирование без задержки: требуемые мощности не нужно заказывать заранее

Все компоненты хранилища горизонтально масштабируются через консоль за минуты.

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®



Гибкое управлению стоимостью платформы данных

Гибридное хранение данных в Yandex Cloud

Плата только за потребление

Охлаждение данных в Object Storage

Временные кластеры Data Proc

Гибридное хранилище в Managed ClickHouse

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 Uber, CERN и др. А также в сервисах Яндекса: Такси, Маркет

Широкий выбор экспертов

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


Opensearch

Jupyter


Redis


My SQL®


ClickHouse


Greenplum Database


PostgreSQL


Kafka


MongoDB


Apache Spark


Airflow


YDB

* По России и СНГ

Фокусируйтесь на работе с данными, а не на обслуживании инфраструктуры

Надёжность, производительность и безопасность управляемых сервисов – наш приоритет

Сосредоточьтесь на главном — архитектуре и модели данных

Исключите потребность развивать непрофильные компетенции внутри команды

Мы возьмём на себя:

- информационную безопасность
- мониторинг
- резервирование
- бэкапы
- обновление

Партнёрская экосистема

DWH и аналитика

Glowbyte Consulting
Navicon
Geointellect
Korus Consulting
Bi.Qube
DBI
Softline
Jet Infosystems
Data Stories
Fevlake
Internet Expert
Hilbert Team
Neoflex
DRT (ex-Delloite)
Axenix (ex-Accenture)
GMCS
DataGo
Aero

Визуализация

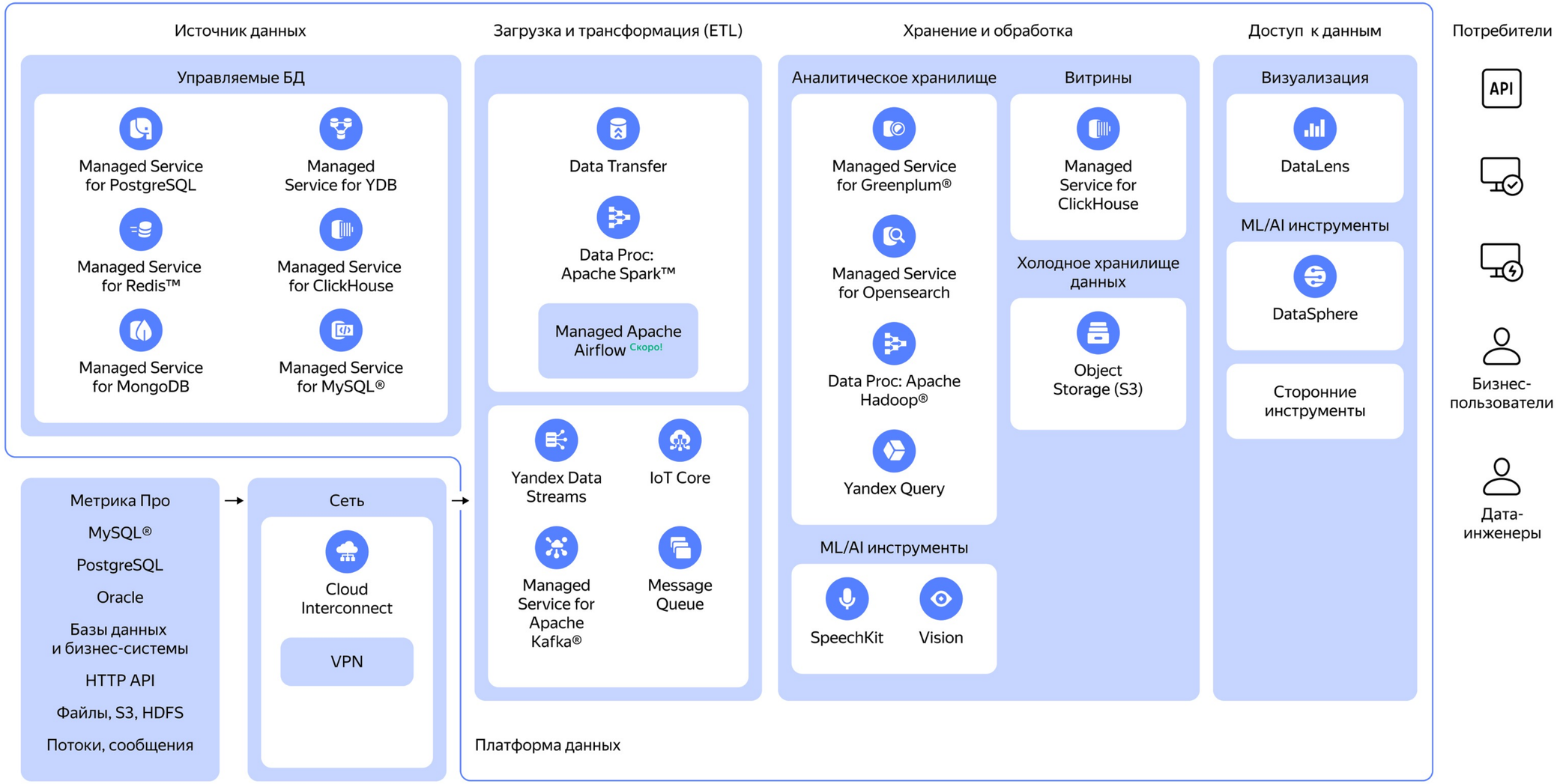
Geointellect
Yolva
Glowbyte Consulting
Navicon
Korus Consulting
Bi.Qube
Aplana
DBI
Softline
DRT (ex-Delloite)
Datanomics (Beltel)

Приложения

Hilbert Team
Express24
Korus Consulting
OpsGuru
Neoflex
DBI
IT Summa
Digital Spirit
Jet Infosystems

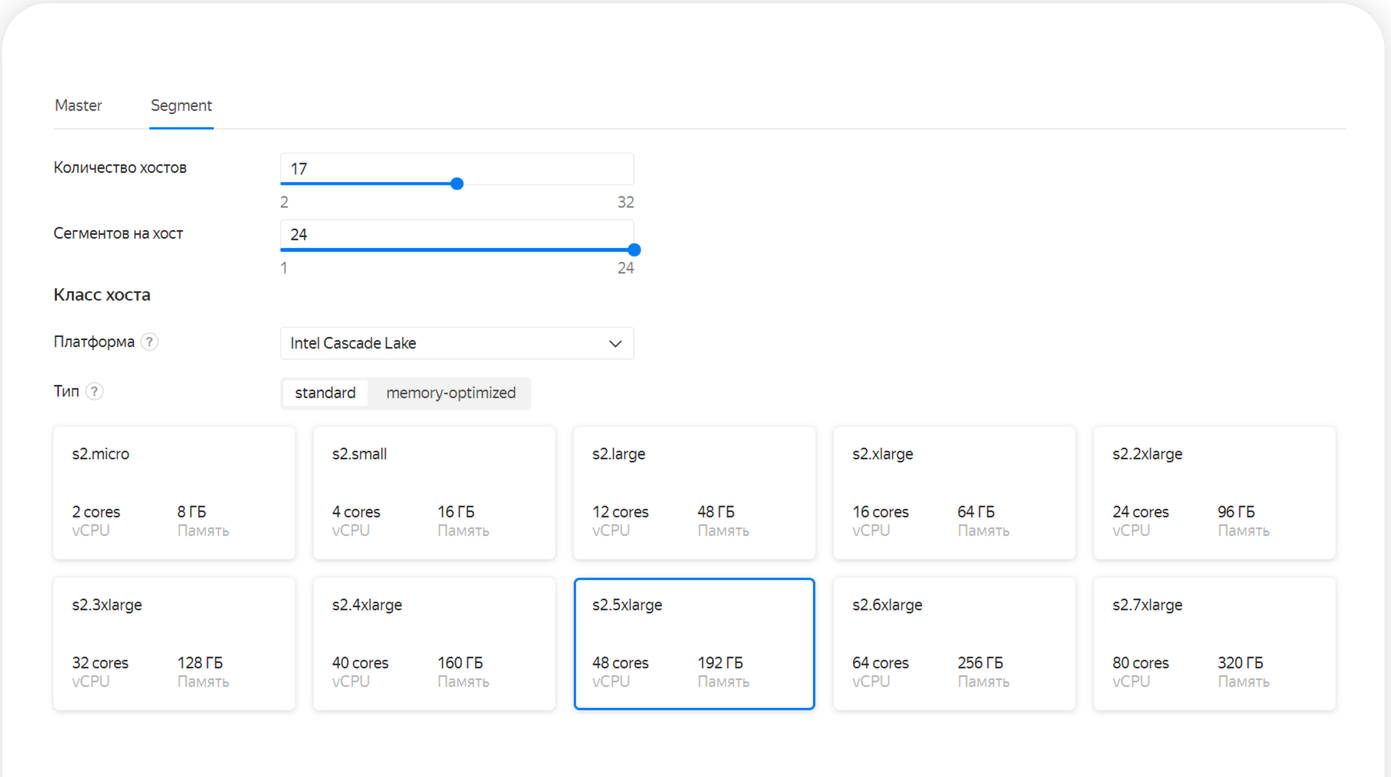
Архитектура
и функционал

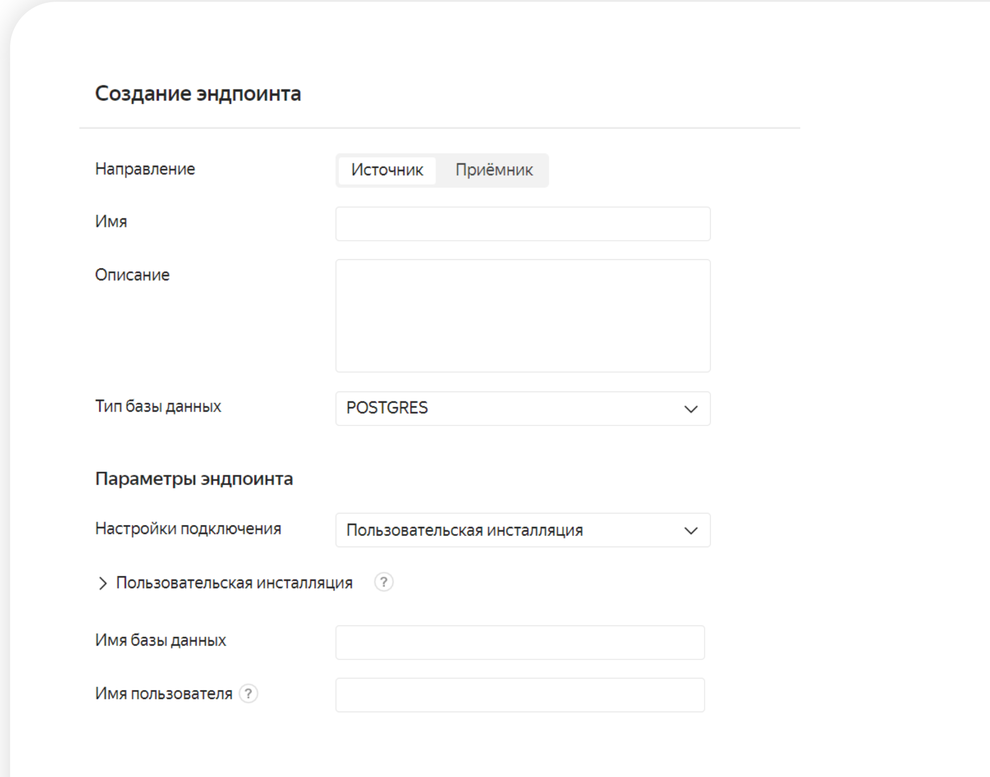
Платформа данных Yandex Cloud



Консоль: функциональный и интуитивно простой интерфейс для архитектора и администратора

Развёртывание, расширение, обновление, интеграция, мониторинг и BI в одном месте





Мониторинг

>340 метрик

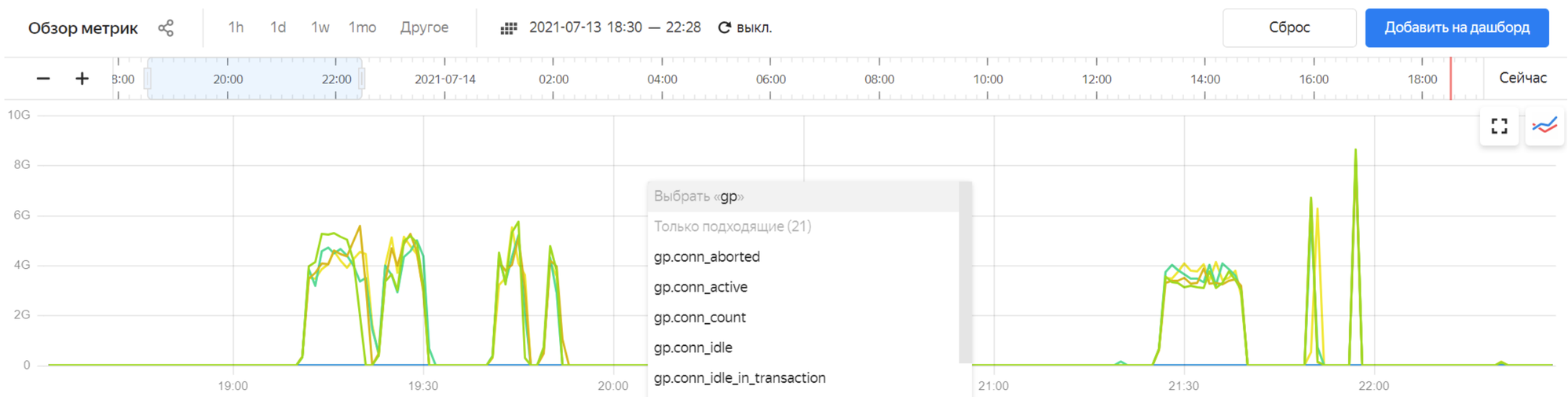
В том числе сервис-специфичных

Фильтрация

Функции, агрегации, интеграции

Алерты

Срабатывают при значимых изменениях

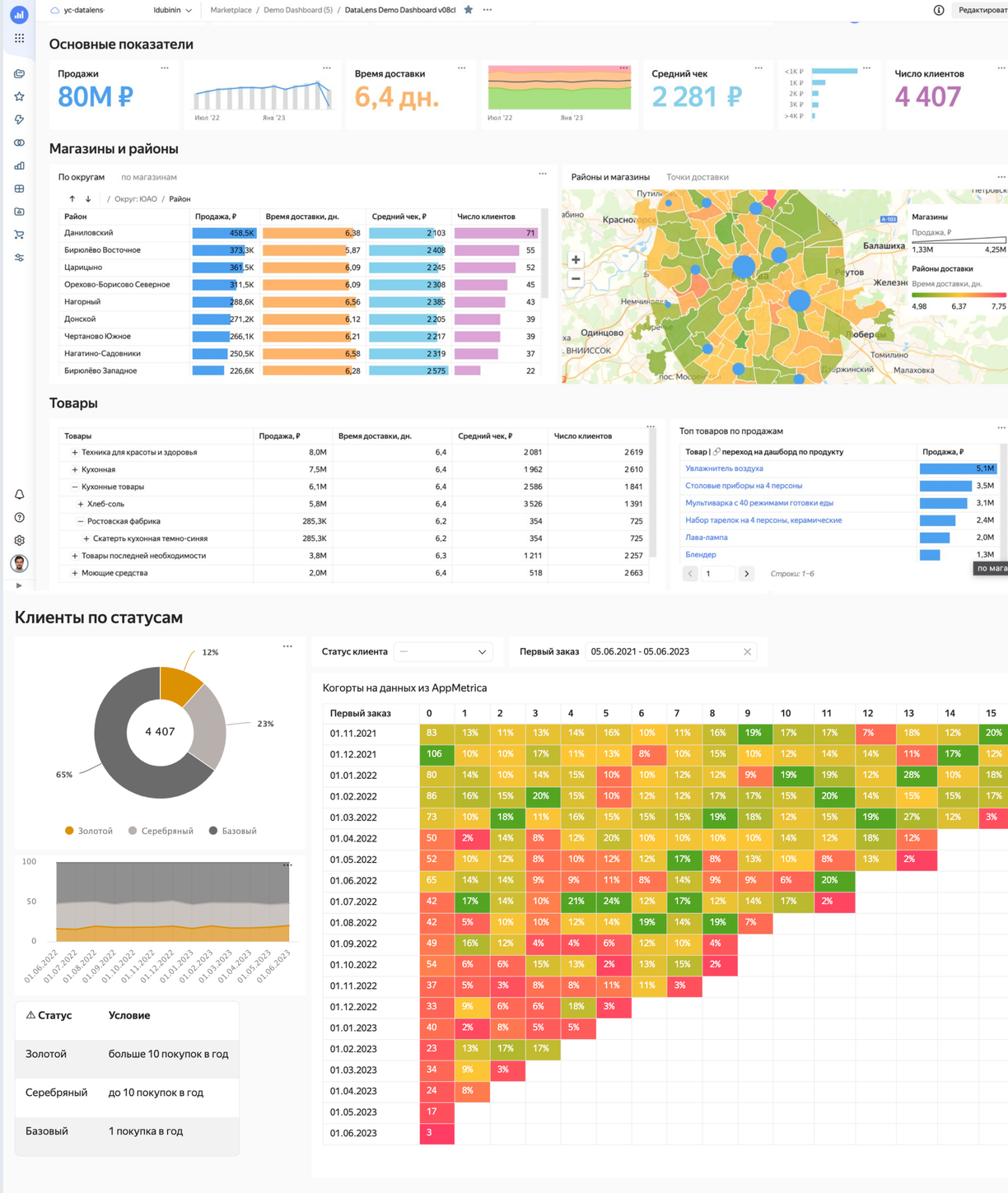


Yandex DataLens

Доступная аналитика любых масштабов

- Корпоративный стандарт BI во всём 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: clck.ru/34K7Jw



Стоимость проекта

Преимущества ценообразования

Отсутствие капитальных затрат

Возможность уменьшить или прекратить потребление в любой момент без потерь

Удобные способы оплаты

Гранты и бесплатное тестирование

Гибкий выбор ресурсов

от 2 до 96 vCPU

от 8 до 576 ГБ RAM

до 8 ТБ ROM на ВМ

до 32 ВМ в кластерах СУБД

Точное прогнозирование

Подробная детализация затрат

Удобная ресурсная модель (затраты на проект)

Простота оценки ресурсов

Открытый калькулятор для расчёта стоимости ресурсов

Доступность для любого типа бизнеса: крупный, средний, малый

Резервы на 1 или 3 года

Compute Cloud

Managed Databases

Почасовая оплата

Сервисы тогда, когда нужно

Платформа решает задачи бизнеса любого размера

Наши сервисы в различных конфигурациях используют как малый и средний бизнес, так и крупный enterprise из топ-5 e-commerce, банков и ритейла.

Компании с хранилищем под production-нагрузки на платформе сервисов Yandex Cloud

М.ВидеоЭльдорадо

KazanExpress

МАГНИТ

Hoff

Мы знаем, какие задачи решают пользователи, и делаем так, чтобы сервисы работали

24/7

Вместе с партнёрами мы готовы помочь вам выбрать оптимальную архитектуру проекта с учётом текущего стека и возможностей интеграции

Поможем построить архитектуру DWH в облаке

Евгений Попов
Руководитель направления
Здравоохранение Yandex Cloud
popov-evgeny@yandex-team.ru

Получите консультацию
Поможем проработать архитектуру
Оптимизируем стоимость