Yandex Cloud

Как делать аналитические проекты в облаке

Рассказываем о том, как инструменты и ресурсы Yandex Cloud помогают собирать, хранить, обрабатывать и визуализировать данные в облачной платформе

Если вы руководитель бизнес-команды

Поможем подобрать подходящую архитектуру проекта, рассчитаем его стоимость и сроки реализации, порекомендуем сертифицированного партнёра Yandex Cloud, который будет помогать вам на всех стадиях проекта.
Ждём писем на cloud-sales@yandex-team.ru.

Поделитесь этим документом с вашими IT-специалистами: в нём мы кратко рассказали об облачных инструментах и сервисах для работы с аналитикой

84%

компаний, которые уже внедрили облачные технологии, видят явный экономический эффект*

Если вы технический специалист

Познакомим вас с ресурсами и инструментами для анализа данных, которыми располагает Yandex Cloud. Материалы сгруппировали по этапам работы над аналитическим проектом в облаке. Посмотрите документ полностью или сразу переходите в нужный раздел.

Что внутри?

* Источник: исследование EY для Yandex Cloud. «Облачные платформы в России 2021: как помочь росту бизнеса»

Первые шаги и обзор этапов работы

Работу над проектом по анализу данных условно можно разделить на четыре этапа. Инструменты для каждого из этапов вы найдёте в Yandex Cloud.

1

Сбор и подготовка данных (ETL)

2

Хранение, обработка, и создание витрин

3

Машинное обучение для продвинутой аналитики

4

Визуализация данных

Обзор ресурсов платформы данных

Плейлист видеоматериалов о платформе данных

Корпоративное хранилище данных

Бизнес-аналитика

Рекомендательная система для ритейла и e-commerce

Data Science в облаке

Материалы для новичков в Yandex Cloud

Начало работы в документации

Вебинар о миграции

Сравнение с другими платформами

Этап 1. Сбор и подготовка данных (ETL)

Расскажем о сервисах, которые помогут, которые помогут перенести данные из источника, обработать их и загрузить в хранилище.

Сбор данных

Yandex Data Transfer — сервис для переноса данных. Помогает регулярно поставлять данные в аналитическое хранилище. Данные передаются не только в режиме snapshot, но и в режиме репликации, который поддерживает копию данных в приёмнике в актуальном состоянии.

Видео: вебинар
Документация
Обзор и инструкции

Managed Apache Kafka® — сервис управления потоками данных в связке с Yandex Data Transfer. Если данные поставляются в произвольной форме, а не из БД, можно воспользоваться брокерами сообщений Yandex Cloud, данные из которых Data Transfer сохранит в приёмники.

Видео: вебинар
Документация

Yandex Data Streams — сервис также решает задачу ввода и передачи потоков данных. В отличие от Managed Apache Kafka®, сервис совместим с протоколом AWS Kinesis Data Streams, не требует выделенных виртуальных машин (serverless-модель) и интегрирован с Yandex Query.

Документация

Решение на сайте Yandex Cloud

Видео: доклад на конференции Yandex Scale 2021

Подготовка данных

Yandex Data Proc — сервис трансформации данных в объектное хранилище на временных кластерах Apache Spark, Apache Hadoop для DWH/ML/BI. Вместе с Yandex Object Storage служит для хранения и предварительной обработки большого объёма сырых и слабоструктурированных данных.

Документация

Видео: доклад на DataOps Community Meetup

Apache Airflow — инструмент для создания, мониторинга и оркестрации пайплайнов загрузки и обработки данных как кода. Автоматизировать управление заданиями в Yandex Data Proc, а обогащать данные и готовить их витрины помогут PostgreSQL, Greenplum® и ClickHouse.
Сервис Apache Airflow доступен в Yandex Cloud Marketplace.

Документация

Видео: вебинар

Yandex Query обеспечивает пакетные и потоковые трансформации данных с использованием SQL. Сервис не требует выделенных виртуальных машин (serverless-модель) и интегрирован с Yandex Data Streams для потоковой аналитики.

Документация

Yandex Cloud Functions — сервис для запуска кода в виде функции в безопасном окружении и без создания и использования виртуальных машин. Благодаря Cloud Functions вы можете создавать функции сбора и обработки данных перед поставкой их в хранилище. Часто работает с сообщениями.

Документация

Этап 2. Хранение и обработка данных

В отличие от on-premise решений, управляемые сервисы снимают с пользователя задачи обслуживания и управления инфраструктурой, позволяя уделять больше времени решению задач. Yandex Cloud берёт на себя:

- Развёртывание кластера
- Обновление СУБД
- Резервное копирование
- Безопасность хранилища данных и оборудования
- Реализацию репликации и шардирования*
- Интеграцию с сервисами Yandex Cloud
- Шифрование данных
- Мониторинг

* Шардирование — стратегия горизонтального масштабирования данных, при которой данные размещаются на разных хостах единого кластера. Шардирование доступно для ClickHouse, Greenplum®, YDB, Apache Kafka®, MongoDB® и Redis™

Для хранения и обработки данных используйте управляемые сервисы баз данных Managed PostgreSQL или Managed ClickHouse.
Витрины часто строят на ClickHouse.

Для создания корпоративного хранилища данных многие компании выбирают в качестве ядра Managed Greenplum®, который позволяет строить развитые модели данных с промежуточными слоями на любых объёмах данных.

Курс: Построение корпоративной аналитической платформы

Managed ClickHouse — сервис для управления кластерами быстрой колоночной СУБД для аналитики.

Документация
Видео: вебинар

Managed PostgreSQL — сервис для управления кластерами популярной объектно-реляционной СУБД.

Документация
Видео: вебинар

Managed Greenplum® — сервис для управления кластерами популярной массивно-параллельной СУБД.

Документация
Видео: вебинар
Видео: митап

Yandex Query — сервис, который предоставляет доступ к данным в объектном хранилище различным потребителям, включая пользователей Yandex DataLens.

Документация

Этап 3. Машинное обучение для продвинутой аналитики

Yandex DataSphere — полноценное интегрированное рабочее место специалиста по Data Science. Позволяет сохранять и воспроизводить результаты исследований, работать совместно с командой и расширить возможности стандартных ML-инструментов, например JupiterLab.

Документация
Видео: доклад на конференции Yandex Scale 2022

Этап 4. Визуализация данных

Yandex DataLens — бесплатный BI-сервис для всей команды.

Документация
Видео: вебинар
Видео: инструкция по началу работы

Когда собраны данные и настроены аналитические витрины, всё готово для визуализации и построения дашбордов в Yandex DataLens. Самостоятельный подход даёт полную свободу в создании аналитических отчётов, моделей данных, визуализаций, а также в предоставлении прав доступов даже на уровне строк.

Демо-дашборд – только на чтение
Развернуть демо-дашборд
Курс: Основы работы с DataLens

Благодаря гибкости настроек и разнообразию источников данных вы можете быстро и легко провести анализ сырых данных, отслеживать бизнес-метрики по продуктам в режиме реального времени и расширить аудиторию, заинтересованную в бизнес-аналитике.

М.видео

На 80% сократили время внедрения изменений в дата-продукты при снижении стоимости эксплуатации

Узнать больше

Уральский Банк реконструкции и развития

За 9 месяцев с нуля построили рабочий фреймворк Data Science на Yandex Cloud

Узнать больше

PUBLICIS MEDIA

Перенесли платформу для моделирования рекламных кампаний в Yandex Cloud, сократив время на реализацию высокотехнологичных проектов на 30%

Узнать больше

KazanExpress

В сжатые сроки проанализировали локации для открытия более 90 пунктов выдачи заказов в 25 городах России.

Узнать больше

Остались вопросы?

Обратитесь в службу технической поддержки Yandex Cloud
или свяжитесь с отделом продаж

Если в вашей команде нет специалистов нужного профиля, мы поможем подобрать партнёра для миграции