Яндекс Практикум

Инженер данных

Продолжительность курса 6,5 месяцев

8 проектов в портфолио

00

Вводная часть

Простая витрина данных

5 часов

01

Data Governance / Data Operations

Как построить аналитическое хранилище данных

Работа с данными в хранилище

ETL: автоматизация подготовки данных

Проверка качества данных

10 недель

02

Data at scale

DWH для нескольких источников

Аналитические базы данных

Организация Data Lake

8 недель

03

Performance at scale

Потоковая обработка данных

Облачные технологии

6 недель

04

Итоговый проект

2 недели

Воркшопы

Воркшоп — это онлайн-мероприятие, которое проводит наставник. На воркшопах вы решите новые задачи из практики инженера данных, разберёте и улучшите собственные проекты.

В каждом спринте будет от одного до трёх воркшопов. Они будут проходить в течение всей программы в фиксированное время.

00

Вводная часть

5 часов

Простая витрина данных

Устроитесь на работу в IT-компанию как начинающий инженер данных и попробуете выполнить своё первое задание — получите от лида требования и построите по ним витрину данных.

Инструменты и технологии

SQL и Python

Metabase

PostgreSQL

01

Data Governance / Data Operations

Спринт 1

В этом модуле вы изучите DataOps (от англ. Data Operations — «операции с данными»): начнёте создания хранилища и обработки данных, а закончите работой с метриками и контролем качества данных.

3 недели

36+ часов

1 тест для самопроверки

Как построить аналитическое хранилище данных

Вы поможете молодому и быстрорастущему бизнесу справиться с хаосом в организации данных и спроектируете для него DWH — хранилище данных.

Инструменты и технологии

SQL

PostgreSQL

В этом спринте вы:

— познакомитесь с необходимыми для строительства хранилища данных технологиями

— изучите различные подходы к построению хранилищ

— научитесь работать с требованиями заказчика и выбирать лучший подход для решения поставленной задачи

Проект

В этом спринте проекта нет. Вы продолжите работать с той же задачей в следующем спринте. Вместо проекта — тест на проверку и закрепление знаний.

Спринт 2

3 недели
36+ часов
1 проект

Работа с данными в хранилище

Вы определились с тем, как будете строить хранилище данных, и согласовали требования к нему с заказчиком. Осталось изучить ещё пару вещей, и можно приступать к реализации.

Инструменты и технологии

SQL

PostgreSQL

В этом спринте вы:

— познакомитесь с понятием витрин данных, научитесь строить их и обновлять

— научитесь работать с инкрементальной загрузкой и транзакциями

— узнаете, как оптимизировать запросы

Проект

Постройте хранилище данных в PostgreSQL

Спринт 3

3 недели
36+ часов
1 проект

ETL: автоматизация подготовки данных

О хранилище данных компании вы теперь знаете почти всё. Пришло время настроить ETL-процессы.

Инструменты и технологии

Python

Airflow

PostgreSQL

В этом спринте вы:

— автоматизируете пайплайн работы с данными

— настроите автоматическую выгрузку данных из источников

— научитесь регулярно и инкрементально загружать данные в БД

Проект

Постройте для e-commerce-проекта пайплайн автоматизированного получения, обработки и загрузки данных (ETL) от источников до витрины

Спринт 4

1 неделя
12+ часов

Проверка качества данных

Вы хотите быть уверены, что ваши первые пайплайны работают нормально. Качество данных необходимо проверять, а поломки — вовремя отслеживать.

Инструменты и технологии

Python

Airflow

PostgreSQL

В этом спринте вы:

— поймёте, как пользоваться метаинформацией и документацией

— оцените качество данных

02 Data at scale

Вы научились обрабатывать данные и теперь готовы к более сложной задаче. Сначала создадите классический DWH (от англ. Data Warehouse — «хранилище данных»), а затем построите Data Lake для разнообразных данных.

Спринт 5

2 недели
24+ часа
1 проект

DWH для нескольких источников

Инструменты и технологии

PostgreSQL
MongoDB

Вы продолжаете исследовать DWH, потому что развитие компании и, следовательно, увеличение объёма данных не остановить.

В этом спринте вы:

— построите DWH с нуля на реляционной СУБД
— познакомитесь с MongoDB в качестве источника данных

Проект

Спроектируете и реализуете DWH для инхаус-стартапа

Спринт 6

2 недели
24+ часа
1 проект

Аналитические базы данных

Инструменты и технологии

Vertica
PostgreSQL
Airflow
S3

Специфичных неструктурированных данных, которые тоже надо хранить и обрабатывать, становится больше. Поэтому мы познакомим вас с концепцией аналитических баз данных на примере СУБД Vertica.

В этом спринте вы:

— изучите организацию хранилища в Vertica
— научитесь делать базовые операции с данными в в Vertica
— построите простое хранилище данных в Vertica

Проект

Постройте DWH для высоконагруженной системы малоструктурированных данных мессенджера с использованием Vertica

Спринт 7

4 недели

48+ часов

1 проект

Организация Data Lake

Классические решения не помогают справиться с объёмом и разнообразием видов данных. Чтобы справиться с новыми вызовами бизнеса, вы построите и наполните Data Lake.

Инструменты и технологии

Hadoop

MapReduce

HDFS

Apache Spark (PySpark)

В этом спринте вы:

— рассмотрите архитектуру Data Lake (пер. «озеро данных»)

— научитесь обрабатывать данные в MPP-системе

— наполните Data Lake данными из источников

— потренируетесь в обработке данных с помощью PySpark и Airflow.

Проект

Постройте Data Lake, а также автоматизируете загрузку и обработку данных в нём

03 Performance at scale

В этом модуле вы научитесь обрабатывать потоковые данные в реальном времени, а также изучите эластичность систем на примере облачных технологий.

Спринт 8

3 недели
36+ часов
1 проект

Потоковая обработка данных

Инструменты и технологии

Kafka
Spark Streaming

Трудности с разнообразием данных вы победили, но появилась новая задача — нужно помочь бизнесу быстрее принимать решения. Тут понадобятся знания потоковой обработки данных (англ. streaming).

В этом спринте вы:

— рассмотрите особенности потоковой обработки данных
— построите свою стриминговую систему
— построите витрину с использованием real-time данных

Проект

Разработаете систему real-time обработки данных

Спринт 9

3 недели
36+ часов
1 проект

Облачные технологии

Инструменты и технологии

Яндекс Облако
Kubernetes
kubectl
Redis
PostgreSQL

Теперь вы умеете работать и с большими объёмами данных, и с потоками. Осталось только автоматизировать масштабирование систем с помощью облачных сервисов.

В этом курсе вы познакомитесь с тем, как реализовать уже изученные решения, но в облаке (на примере Яндекс Облака).

Проект

Разработаете инфраструктуры хранения и обработки данных в облаке

04 Итоговый проект

Спринт 10

2 недели
24+ часа
1 проект

Подтвердите, что освоили новые навыки.

Здесь вам будет нужно самостоятельно выбрать и реализовать решения для бизнес-задачи. Это поможет вам ещё раз закрепить использование изученных инструментов, а также самостоятельность.