

Инженер данных

01 Кому подойдёт курс

- Специалистам по Data Science и аналитикам:
- освоите новые инструменты, чтобы эффективнее справляться с задачами
 - научитесь строить хранилища и витрины данных, проектировать пайплайны, готовить данные к машинному обучению
- Начинающим инженерам данных:
- систематизируете знания и отработаете их на практике
 - решите проекты, которые войдут в портфолио, выделят вас на фоне других кандидатов и помогут расти в карьере
- Практикующим разработчикам:
- получите навыки и знания инженера данных, чтобы использовать их на текущей должности или сменить работу
 - освоите архитектуры данных, ETL-процессы, Airflow, PySpark и другие инструменты

02 Чему научитесь на курсе

Какие знания и навыки освоите

- Освоите проектирование пайплайнов
- Узнаете, как создавать витрины и хранилища
- Научитесь обрабатывать данные разными инструментами

03 Как проходит курс

- Теория и практика на платформе Практикума
- Практические задания на готовой инфраструктуре
- Воркшопы
- Индивидуальные онлайн-встречи с наставником
- Доступ из любой точки мира в удобное время

Что вас ждёт

Сделаете и добавите в портфолио 8 проектов

Погрузитесь в рабочую среду инженера данных

В качестве итоговой работы сможете выполнить пет-проект

Инженер данных

6,5 месяцев

продолжительность курса

8 проектов

в портфолио

5 ЧАСОВ

00

Вводная часть

- Простая витрина данных

10 НЕДЕЛЬ

01

Data Governance /
Data Operations

- Как построить аналитическое хранилище данных
- Работа с данными в хранилище
- ETL: автоматизация подготовки данных
- Проверка качества данных

8 НЕДЕЛЬ

02

Data at Scale

- DWH для нескольких источников
- Аналитические базы данных
- Организация Data Lake

6 НЕДЕЛЬ

03

Performance at Scale

- Потоковая обработка данных
- Облачные технологии

2 НЕДЕЛИ

04

Итоговый проект

Вводный модуль.

Простая витрина данных

00

5 часов

Устроитесь на работу в IT-компанию как начинающий инженер данных и попробуете выполнить свою первую задачу — получите от лида требования и постройте по ним витрину данных.

Инструменты и технологии

- SQL и Python
- Metabase
- PostgreSQL

Data Governance / Data Operations

01

10 недель

Итоговый тест

2 проекта

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

Содержание модуля

01. Как построить аналитическое хранилище данных Вы поможете молодому и быстрорастущему бизнесу справиться с хаосом в организации данных и спроектируете для него DWH — хранилище данных.	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	Инструменты и технологии - SQL - PostgreSQL Проект В этом спринте проекта нет. Вы продолжите работать с той же задачей в следующем спринте. Вместо проекта — итоговый тест на проверку и закрепление знаний.	3 недели 36+ часов
02. 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.	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	Инструменты и технологии - SQL - PostgreSQL Проект Построите хранилище данных в PostgreSQL.	3 недели 36+ часов

03.	ETL: автоматизация подготовки данных О хранилище данных компании вы теперь знаете почти всё. Пришло время настроить ETL-процессы.	В этом спринте вы: • автоматизируете пайплайн работы с данными • настроите автоматическую выгрузку данных из источников • научитесь регулярно и инкрементально загружать данные в БД	Инструменты и технологии • Python • Airflow • PostgreSQL Проект Построите для e-commerce-проекта пайплайн автоматизированного получения, обработки и загрузки данных (ETL) от источников до витрины.	3 недели 36+ часов
04.	Проверка качества данных Вы хотите быть уверены, что ваши первые пайплайны работают нормально. Качество данных необходимо проверять, а поломки — вовремя отслеживать.	В этом спринте вы: • поймёте, как пользоваться метаинформацией и документацией • оцените качество данных	Инструменты и технологии • Python • Airflow • PostgreSQL	1 неделя 12+ часов

Data at Scale

02

8 недель

3 проекта

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.

Содержание модуля

05.	DWH для нескольких источников 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	Инструменты и технологии • PostgreSQL • MongoDB Проект Спроектируете и реализуете DWH для инхаус-стартапа.	2 недели 24+ часа
06.	Аналитические базы данных 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 Vertica • построите простое хранилище данных в Vertica	Инструменты и технологии • Vertica • PostgreSQL • Airflow • S3 Проект Построите DWH для высоконагруженной системы малоструктурированных данных мессенджера с использованием Vertica	2 недели 24+ часа

07.	Организация Data Lake Классические решения не помогают справиться с объёмом и разнообразием видов данных. Чтобы справиться с новыми вызовами бизнеса, вы постройте и наполните Data Lake.	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	Инструменты и технологии • Hadoop • MapReduce • HDFS • Apache Spark (PySpark) Проект Построите Data Lake, а также автоматизируете загрузку и обработку данных в нём.	4 недели 48+ часов
-----	---	--	---	-----------------------

Performance at Scale

03

6 недель 2 проекта	В этом модуле вы научитесь обрабатывать потоковые данные в реальном времени, а также изучите эластичность систем на примере облачных технологий.
-----------------------	--

Содержание модуля

08.	Потоковая обработка данных 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 • постройте свою стриминговую систему • постройте витрину с использованием real-time данных	Инструменты и технологии • Kafka • Spark Streaming Проект Разработаете систему real-time обработки данных.	3 недели 36+ часов
09.	Облачные технологии 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).	Инструменты и технологии • Яндекс Облако • Kubernetes • kubectl • Redis • PostgreSQL Проект Разработаете инфраструктуры хранения и обработки данных в облаке.	3 недели 36+ часов

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.