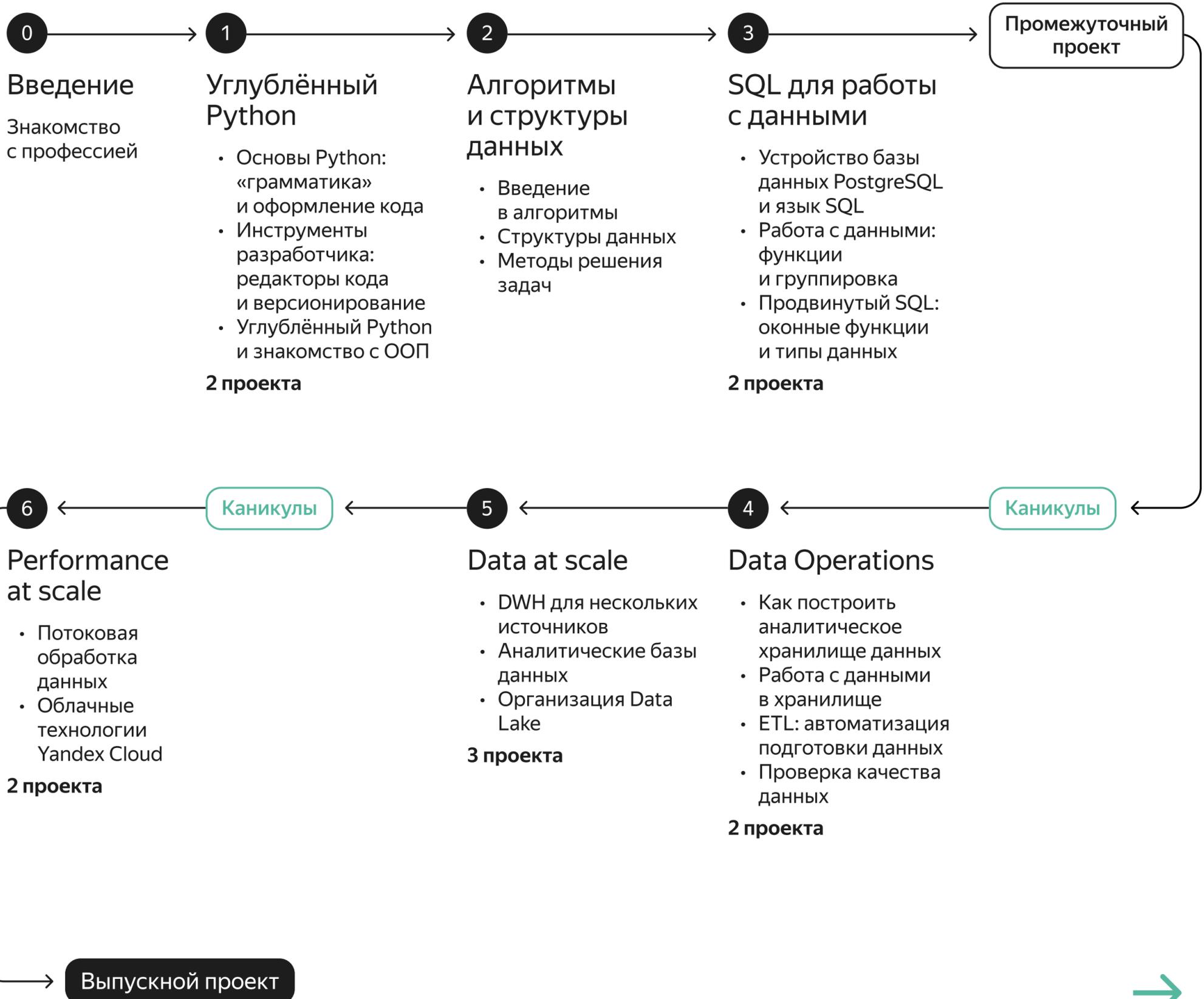


Инженер данных с нуля

Продолжительность курса 11 месяцев

13 проектов в портфолио

12–15 часов учёбы в неделю



Воркшопы

В каждом модуле будут онлайн-занятия, которые проводит наставник.

На этих занятиях вы углубите свои знания, получив практические советы по реализации ваших учебных проектов, и узнаете больше об особенностях работы инженера данных.

00

Введение

~2 часа

Познакомьтесь с профессией инженера данных:

- узнаете о роли и задачах инженера данных в процессе разработки продукта,
- узнаете, что такое базы данных и для чего они нужны,
- напишете свой первый SQL-запрос.

Инструменты
и технологии

SQL

01

Углублённый Python

7 недель
от 84 часов
2 проекта

В этом модуле вы подробно изучите «грамматику» Python, обучитесь правилам оформления кода, работе в виртуальном окружении и основам объектно-ориентированного программирования. Вас ждёт много практики — писать код вы будете практически в каждом уроке. Работать вы будете в онлайн-тренажёре — нашей интерактивной среде.

Языки:

Python 3.9

Инструменты
и технологии

Git

GitHub

редактор кода Visual
Studio Code

Flake8

Pytest

виртуальное
окружение (venv)

Спринт 1

Основы Python

3 недели
36+ часов

Этот спринт станет вашим стартом в мире Python. Начнёте вы, как водится, с основ:

- познакомитесь с переменными, типами данных и функциями,
- научитесь применять встроенные инструменты и модули.

Проект

Разработаете приложение «Холодильник» для регистрации и отслеживания продуктов.



Спринт 2

1 неделя
12+ часов

Инструменты разработчика

В этом спринте вы продолжите изучать Python, а заодно подготовите рабочую среду — установите и настройте необходимые для работы с Python инструменты:

- установите редактор кода и научитесь с ним работать,
- познакомитесь с популярной системой контроля и управления версиями Git,
- изучите требования к оформлению кода и часто встречающиеся ошибки.

Проект

В этом спринте проекта нет. Вместо него — итоговое задание на проверку и закрепление знаний.

Спринт 3

3 недели
36+ часов

Углублённый Python

Новый спринт — новые возможности Python:

- познакомитесь с основами объектно-ориентированного программирования (ООП) — одного из самых популярных подходов программирования,
- узнаете, какие инструменты помогут сделать ваш код более эффективным, безопасным и оптимизированным.

Проект

Разработаете программу для скачивания данных о персонажах и планетах из вселенной «Звёздных войн» и подготовите эти данные для работы в Excel.



02

Алгоритмы и структуры данных

3 недели
от 36 часов

Любую задачу можно решить медленно и неэффективно, а можно — быстро и экономно. Неэффективное программное решение может впустую занять все вычислительные ресурсы сервера и замедлить его работу, а то и вовсе сломать его. Изучение алгоритмов поможет избежать подобных ошибок и ускорить работу программ. Вы научитесь проектировать решения, которые позволят вашим проектам работать эффективнее.

Языки:

[Python 3.9](#)

Инструменты
и технологии

[Яндекс Контест](#)

Спринт 4

Алгоритмы и структуры данных

3 недели
36+ часов

В этом спринте вы изучите основы алгоритмов:

- разберётесь, что такое алгоритмы и зачем они инженеру данных;
- познакомитесь с несколькими популярными алгоритмами и на примерах выясните, что такое «сложность алгоритма»;
- рассмотрите понятие «структура данных» и узнаете, чем она отличается от типа данных;
- познакомитесь с некоторыми распространёнными методами решения алгоритмических задач.

Проект

В этом спринте вам предстоит решить алгоритмическую задачу и помочь марсоходу рассчитать путь доставки.



03

SQL для работы с данными

8 недель
от 96 часов
2 проекта

От базовых команд SQL до продвинутых техник — этот модуль даст вам глубокое понимание работы с базами данных, включая создание, изменение таблиц, манипулирование данными и освоение сложных способов обработки информации с помощью SQL.

Инструменты
и технологии

SQL

PostgreSQL

Спринт 5

Основы SQL и баз данных

4 недели
48+ часов

В этом спринте вы:

- познакомитесь с менеджерами БД: DBeaver, psql;
- научитесь создавать, изменять и удалять таблицы баз данных — CREATE, ALTER, DROP;
- сможете писать запросы на создание, чтение, обновление и удаление данных — INSERT, SELECT, UPDATE, DELETE;
- научитесь выполнять основные манипуляции с данными — объединять информацию из нескольких таблиц, фильтровать данные, группировать и сортировать их;
- поймёте принципы нормализации баз данных и сможете читать и рисовать структуру БД через ER-диаграммы.

Проект

Спроектируете и создадите БД для автосалона «Врум-Бум». Наполните таблицу сырыми данными на основе выгрузки в формате .csv, создадите нормализованные таблицы и напишете несколько запросов для сбора аналитики по этой базе данных.

Спринт 6

Продвинутый SQL для работы с данными

4 недели
48+ часов

В этом спринте вы:

- научитесь использовать подзапросы и общие табличные выражения, в том числе рекурсивные запросы;
- познакомитесь с синтаксисом оконных функций — OVER, PARTITION BY и освоите некоторые аналитические функции ранжирования и смещения;
- научитесь работать с представлениями (Views и Materialized Views), транзакциями и блокировками;
- разберётесь в продвинутых типах данных: uuid, массивах, json и пользовательских типах данных.



Проект

Создадите базу данных для сети ресторанов Gastro Hub. Получите необработанные данные, постройте дополнительные таблицы с продвинутыми типами данных. Создадите представления и напишете несколько аналитических запросов, используя оконные функции и подзапросы.

Промежуточный проект

1 неделя
12 часов

Разработаете практическое решение задачи инженера данных, используя Python и SQL, чтобы увидеть, как эти навыки применяются в реальной работе.

Каникулы —————> 1 неделя



04

Data Operations

10 недель
от 120 часов
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.

В этом модуле вы:

- разберётесь, как в компании строят БД,
- обновите модель данных текущей БД в соответствии с новыми требованиями бизнеса,
- подготовите новые витрины и метрики для аналитиков и менеджеров.

Инструменты
и технологии

SQL

PostgreSQL

Python

Airflow

Спринт 7

3 недели
36+ часов

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. Вместо него — итоговый тест на проверку и закрепление знаний.

Спринт 8

3 недели
36+ часов

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 понятием витрин данных, научитесь строить их и обновлять,
- научитесь работать с инкрементальной загрузкой и транзакциями,
- узнаете, как оптимизировать запросы.

Инструменты
и технологии

SQL

PostgreSQL



Проект

Построите хранилище данных в PostgreSQL.

Спринт 9

3 недели
36+ часов

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.

Спринт 10

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.

В этом спринте вы:

- поймёте, как пользоваться метаинформацией и документацией,
- оцените качество данных.

Проект

В этом спринте проекта нет. Вместо проекта — итоговый тест на проверку и закрепление знаний.

Каникулы —————> 1 неделя



05

Data at scale

8 недель
от 96 часов
3 проекта

Под data at scale (в переводе с англ. — данные в масштабе) понимают комплексный подход к работе с большими объёмами данных. Вы научились обрабатывать данные и теперь готовы к более сложной задаче — данных становится больше. Сначала создадите классический DWH (от англ. Data Warehouse — «хранилище данных»), а затем построите Data Lake для разнообразных данных.

Инструменты
и технологии

PostgreSQL
MongoDB
Airflow
S3
Vertica
Hadoop
MapReduce
HDFS
Apache Spark
(PySpark)

Спринт 11

DWH для нескольких источников

2 недели
24+ часа

Вы продолжите 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 для сети ресторанов многослойный DWH с двумя источниками данных.

Спринт 12

Аналитические базы данных

2 недели
24+ часа

Специфичных неструктурированных данных, которые тоже надо хранить и обрабатывать, становится больше. Вы познакомитесь 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.



Спринт 13

4 недели
48+ часов

Организация Data Lake, ELT

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

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

Каникулы —————> 1 неделя



06

Performance at scale

6 недель
от 72 часов
2 проекта

Performance at scale (в переводе с англ. — производительность в масштабе) подразумевает, что большие объёмы данных необходимо обрабатывать быстро и эффективно. В этом модуле вы научитесь обрабатывать потоковые данные в реальном времени, а также изучите эластичность систем на примере облачных технологий.

Спринт 14

Потоковая обработка данных

3 недели
36+ часов

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.

Спринт 15

Облачные технологии

3 недели
36+ часов

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

Выпускной проект

2 недели
24+ часа
1 проект

Вам предстоит самостоятельно выбрать и реализовать решения для бизнес-задачи. Это поможет вам ещё раз закрепить использование изученных инструментов.



Программа трудоустройства

В дополнение к основной образовательной программе вы можете пройти трек трудоустройства, чтобы быстрее найти новую работу. Программа включает руководства, вебинары, домашние задания. Каждая секция трека длится 2 недели и занимает около 10 часов.

1. Резюме

- изучите подходы к составлению резюме,
- разберётесь в стратегиях поиска работы,
- сформируете резюме совместно со специалистом по трудоустройству.

2. Портфолио и сопроводительное ПИСЬМО

- научитесь создавать портфолио на сервисе GitHub,
- изучите подходы к написанию сопроводительного письма,
- сформируете сопроводительное письмо со специалистом по трудоустройству.

3. Интервью

научитесь презентовать себя и легче проходить собеседования.

4. Нетворкинг и поиск работы

- начнёте погружаться в профессиональное сообщество,
- узнаете, где искать работу.