

Инженер данных

Пояснения к словам иностранного происхождения

Здесь собраны пояснения к словам иностранного происхождения, которые используются в программе курса, но пока не включены в словарную норму русского языка.

01

Кому
подойдёт
курс

Начинающим инженерам данных:

- Систематизируете знания и отработаете их на практике
- Выполните проекты для портфолио, которые помогут вам выделиться среди других кандидатов и развиваться в профессии

Специалистам по Data Science и аналитикам:

- Освойте новые инструменты, чтобы эффективнее справляться с задачами
- Научитесь строить хранилища и витрины данных, проектировать пайплайны, готовить данные к машинному обучению

Практикующим разработчикам:

- получите знания и навыки инженера данных, чтобы применять их на текущей должности или перейти в новую профессиональную область
- освоите принципы построения архитектуры данных, ETL-процессы, Airflow, Greenplum, ClickHouse, PySpark и другие инструменты

02

Чему научитесь
на курсе

Какие знания и навыки
освоите

- Научитесь проектировать пайплайны
- Узнаете, как создавать витрины и хранилища
- Научитесь обрабатывать данные с помощью разных инструментов

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 9 проектов

Погрузитесь в рабочую
среду инженера данных

В качестве итоговой работы
выполните пет-проект

Инженер данных

6,5 месяца

продолжительность курса

9 проектов

в портфолио

2 ЧАСА

00

Вводная часть

- О курсе
- Знакомство
- Введение в платформу данных: от разрозненных инструментов к целостной системе
- Платформа данных: по шагам пайплайна загрузки данных
- Платформа данных: добавление поля
- Входное тестирование

9 НЕДЕЛЬ

01

Data Governance / Data Operations

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

11 НЕДЕЛЬ

02

Data at Scale

- DWH для нескольких источников
- MPP СУБД — Greenplum
- Clickhouse
- Организация Data Lake

4 НЕДЕЛИ

03

Performance at Scale

- Поточковая обработка данных
- Облачные технологии

2 НЕДЕЛИ

04

Итоговый проект

2 часа	В этом спринте познакомитесь с программой курса, разберёте пайплайн данных на практике, доработаете его и пройдёте тест по SQL и Python, чтобы оценить свою готовность к обучению. Инструменты и технологии • SQL • Python • PostgreSQL
--------	--

Data Governance / Data Operations

01

9 недель Итоговый тест 3 проекта	В этом спринте вы изучите DataOps (от англ. Data Operations — «операции с данными»): начнёте с создания хранилища и обработки данных, а завершите спринт настройкой метрик и проверок качества данных.
--	---

Содержание модуля

01. Как построить аналитическое хранилище данных 2 недели, 24+ часов	В этом спринте вы: • познакомитесь с необходимыми для строительства хранилища данных технологиями • изучите подходы к построению хранилищ и моделированию данных по схемам «звезда» и «снежинка» • научитесь работать с требованиями заказчика и выбирать оптимальное решение	Инструменты и технологии • SQL • PostgreSQL • DBeaver 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.
02. Работа с данными в хранилище 2 недели, 24+ часов	В этом спринте вы: • узнаете, что такое витрина данных и как её создавать • освоите инкрементальную загрузку, транзакции и работу с представлениями • изучите методы оптимизации запросов для повышения производительности	Инструменты и технологии • SQL • PostgreSQL • DBeaver Проект Построите хранилище данных в PostgreSQL.

03.	ETL: автоматизация подготовки данных 3 недели, 36+ часов	В этом спринте вы: - развернёте Airflow с помощью Docker и настроите его на своём устройстве - познакомитесь с интерфейсом, архитектурой и компонентами Airflow - напишете несколько пайплайнов обработки данных для разных сценариев - освоите дополнительные функции Airflow для повышения эффективности работы	Инструменты и технологии - Python - SQL - Airflow - PostgreSQL - Docker Проект Построите пайплайн загрузки нового объекта из API в DWH со сборкой витрины и тестами.
04.	Проверка качества данных 2 недели, 24+ часов	В этом спринте вы: - изучите характеристики качества данных и методы их оценки - освоите поиск аномалий, дата-контракты и тестирование данных на Python - научитесь реализовывать проверки качества на больших объёмах данных в PostgreSQL	Инструменты и технологии - Python - SQL - Airflow - PostgreSQL Проект Проверите качество данных уже существующего хранилища.

Data at Scale

02

11 недель 4 проекта	В этом спринте вы создадите классический DWH (от англ. Data Warehouse — «хранилище данных»), познакомитесь с MPP-системами Greenplum и ClickHouse для обработки больших объёмов информации, а затем построите Data Lake для разнообразных данных.
------------------------	---

Содержание модуля

05.	DWH для нескольких источников 2 недели, 24+ часов	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.
06.	Greenplum 3 недели, 36+ часов	В этом спринте вы: - изучите архитектуру Greenplum и научитесь проектировать таблицы и схемы хранения для аналитических задач - освоите диагностику и оптимизацию производительности: анализ планов запросов, поиск «узких мест» и балансировку сегментов - научитесь диагностировать сбои и предотвращать проблемы с перегрузкой узлов	Инструменты и технологии - Greenplum - SQL Проект Построите хранилище данных в Greenplum.

07. ClickHouse	В этом спринте вы: - определите, для каких задач подходит ClickHouse - развернёте ClickHouse и получите доступ к нему через различные интерфейсы - спроектируете базы данных: научитесь выбирать подходящие движки, обрабатывать потоки данных и выполнять SQL-запросы - освоите ClickHouse как инструмент хранения и обработки больших объёмов данных	Инструменты и технологии - Clickhouse - SQL - OLAP Проект Построите хранилище данных в ClickHouse, учитывая специфику БД.
----------------	---	--

08. Организация 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.
---------------------------	---	---

Performance at Scale

03

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

Содержание модуля

09. Потоковая обработка данных	В этом спринте вы: - рассмотрите особенности потоковой обработки данных построите свою стриминговую систему - построите витрину с использованием real-time данных	Инструменты и технологии - Kafka - Spark Streaming Проект Разработаете систему real-time обработки данных.
--------------------------------	--	---

10. Облачные технологии	В этом спринте вы: 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 облачную инфраструктуру для хранения и обработки данных в облаке.
-------------------------	--	--

2 недели,
24+ часа

- Подтвердите, что освоили новые навыки.
- Самостоятельно выберете и реализуете решения для бизнес-задачи. Это поможет вам ещё раз закрепить использование изученных инструментов.