

Data Engineer

Лучшие практики по приготовлению данных. Загрузка, обработка, организация хранения и доступа к данным с использованием современных инструментов

Длительность курса: 98 академических часов

1 Инженер Данных: задачи, инструменты, платформы

1 Инженер Данных. Задачи, навыки, инструменты, потребности на рынке

познакомиться;
обсудить правила работы и общения;
объяснить, кто такой Инженер Данных и какие перед ним стоят задачи.

Домашние задания

1 Домашнее задание: анализ рынка Инженер Данных: РФ, USA, EU

Цель: Изучить рынок Data Engineering и его особенности в РФ, EU, USA
Проанализировать требования к языкам, стеку инструментов и опыту
Поставить цели и акценты на обучение для себя

Источники: Glassdoor, HH, Monster, LinkedIn
Формат: Свободный - Google Doc, Jupyter, просто текст
Вопросы:
- Рынок DE: Количество вакансий, отрасли компаний, размер компаний
- Требования: Популярные языки, технологии, фреймворки
- Цели на обучение, акценты на интересующие темы и технологии

2 Эволюция подходов работы с данными. Базовые принципы и понятия

проследить эволюцию подходов работы с данными;
получить представление о технологиях и инструментах;
рассмотреть принципы построения архитектуры аналитического приложения.

3 Облачные платформы. Дистрибутивы Cloudera и Hortonworks	объяснить, чем отличаются облака и on-premise; выяснить, какие платформы есть для каждого из вариантов; объяснить, чем определяется выбор конкретной платформы; провести обзор инструментов для обработки данных: Cloudera, GCP, AWS. Домашние задания 1 Развернуть дистрибутив Cloudera Цель: Цель этого ДЗ - научиться выполнять базовые операции на кластере Hadoop. В его ходе нужно будет развернуть свой мини-кластер в Google Cloud Platform и создать таблицу в Hive. Инструкции к ДЗ https://docs.google.com/document/d/1iLTiN7D1kM4njOEbF_f_YGT7Jj7OnvT_WjbU3rX6Hzw/edit?usp=sharing
4 Распределенные файловые системы	научиться работать с HDFS; разобраться в архитектуре HDFS.
5 Форматы данных и их особенности	определить факторы выбора формата хранения данных; объяснить разницу между Row-based и Column-based форматами; провести обзор наиболее распространенных форматов.

1	Введение в Scala	изучить основы языка Scala; научиться собирать простейшее Scala-приложение.
2	Apache Spark - 1 часть	познакомиться с Apache Spark; научиться работать с API Spark.
3	Apache Spark - 2 часть	научиться оптимизировать приложения на Spark. Домашние задания 1 Введение в Spark + Гид по безопасному Бостону Цель: В этом задании предлагается собрать статистику по криминогенной обстановке в разных районах Бостона, используя Apache Spark. Подробную инструкцию по выполнению заданий смотрите в документах по ссылкам: 1) Инструкция к занятию "Введение в Scala: https://docs.google.com/document/d/1HPhwYQCQ1AYkIBeQ75RVsYV4MMGBEw_UBeTWp0c0Q678/edit?usp=sharing 2) Гид по безопасному Бостону: https://docs.google.com/document/d/1eIWInbWsLrIDqB4FMMgFTUMNEmiYev9HJUfg9LXxydE/edit?usp=sharing
4	Очереди сообщений, Kafka, Confluent platform	рассказать о нескольких очередях сообщений (Kafka, Rabbit MQ, NATS, ...); описать примеры архитектур с использованием очередей сообщений; рассказать об архитектуре Apache Kafka.
5	Spark Streaming	объяснить различные модели обработки потоков данных; научиться использовать Spark Streaming для обработки потоков данных.
6	Доступ к данным, ноутбуки. Explore and visualize	рассмотреть инструменты интерактивной аналитики.

3 Развертывание ML-моделей

- | | | |
|---|--|--|
| 1 | Production Code на Python.
Организация и Packaging кода | превратить свой код ML модели в готовый вид к работе с продакшн системами;
объяснить организацию и структуру кода, научиться паковать свой код с помощью python подходов. |
| 2 | Docker и REST-архитектура | рассказать о REST архитектуре и ее применении;
пользоваться одним из широко распространенных веб фреймворков Flask и с его помощью превращать свою модель в готовый веб-сервис;
рассказать о docker, для чего он используется и где применяется; использовать docker для packaging своего кода; использовать docker для деплоя на AWS системы. |
| 3 | Деплоймент моделей в Sagemaker | использовать мощный инструмент ML на AWS-Amazon Sagemaker для разработки и тренировки собственных моделей и создания endpoints для работы с веб-сервисами. |

4 Очереди сообщений. Хранилища данных. SQL-доступ.

1 Инструменты выгрузки данных из сторонних систем

научиться классифицировать источники для выбора правильного способа загрузки;
разобраться в форматах передачи данных;
научиться реплицировать оперативные базы данных.

Домашние задания

1 Настроить захват изменений из Postgres (CDC)

Цель: Настраиваем захват изменений из Postgres с помощью Debezium и загрузку этих изменений в Hadoop с помощью StreamSets.

Следуйте инструкциям в описании репозитория:
https://github.com/Gorini4/debezium_cdc/

2 DWH. Хранилища данных - 1 часть

оптимизировать MPP-базы для аналитики на больших объемах данных;
использовать специфичные механизмы MPP-баз для оптимизации запросов.

3 DWH. Хранилища данных - 2 часть

выстраивать работу с СУБД в соответствии с лучшими практиками;
выбрать подход для моделирования витрин в DWH;
использовать фреймворк DBT для построения витрины.

Домашние задания

1 Домашнее задание: проектирование витрины в Vertica (BigQuery)

Цель: В этом ДЗ вы спроектируете схем данных и постройте витрину.

Спроектировать схему данных + Построить витрину
Использовать Vertica (Docker) или BigQuery • Датасет:
Захват данных из divolte (или GCP Public Datasets)
Definition of Done:

- DDL объектов
- DML шагов преобразований
- Опционально: Тестирование на наличие ошибок в данных

4 Хранилища NoSQL. Назначение и особенности

познакомиться с течением NoSQL;
рассмотреть классификацию, причины появления, историческую справку NoSQL СУБД;
понять место NoSQL СУБД в жизни data-инженера;
рассмотреть несколько примеров NoSQL СУБД, как они внутри устроены, как их правильно использовать.

5 **Практика по
NoSQL**

6 **SQL-доступ к
данным. Apache
Hive**

познакомиться с инструментами класса SQL-доступ, их устройством, компонентами;
получить представление об особенностях работы, сценариях использования.

Домашние задания

1 HiveQL

Цель: Практика с Hive на CDH

Развернуть дистрибутив CDH
Самостоятельно проделать манипуляции с Hive с
приложенными скриптами

7 **ElasticSearch**

познакомиться с основными концепциями Elasticsearch;
рассмотреть возможности обработки данных с использованием Logstash;
рассмотреть визуализации в Kibana.

8 **Практика по
ElasticSearch**

- | | | |
|---|---|--|
| 1 | Интеграция, тестирование, развертывание. CI / CD в ML | рассмотреть работу в Jenkins и построение CI/CD пайплайна применительно к DS задачам. |
| 2 | Оркестрация | использовать оркестраторы для автоматизации процессов обработки данных;
написать DAG'и для создания сложных процессов обработки, состоящих из множества этапов. |
| 3 | Мониторинг | рассмотреть инструменты мониторинга - Prometheus, Zabbix, Graphite, Grafana;
объяснить специфику мониторинга процессов обработки данных. |
| 4 | Data Quality. Контроль качества данных, мастер-данные, Troubleshooting | получить представление о том что такое качество данных и как оно влияет на выводы и принимаемые решения;
рассмотреть стратегии управления качеством данных. |

- | | | |
|---|---|---|
| 1 | Выбор темы и организация проектной работы | <p>выбрать и обсудить тему проектной работы;
спланировать работу над проектом;
ознакомиться с регламентом работы над проектом.</p> <p>Домашние задания</p> <p>1 Проектная работа</p> <p>Цель: В качестве курсового проекта необходимо придумать дизайн и архитектуру, а затем - имплементировать data-driven приложение в выбранной доменной области. Работающее приложение должно быть представлено в качестве репозитория в GitHub.</p> <p>Подробнее см. в файле:
https://docs.google.com/document/d/1Onz9oWn_TfdWZmBhRCOIUAyKKLX3gXqcliMLdhzBoow/edit?usp=sharing</p> <hr/> |
| 2 | Консультация по проектам и домашним заданиям | <p>получить ответы на вопросы по проекту, ДЗ и по курсу.</p> <hr/> |
| 3 | Защита проектных работ | <p>защитить проект и получить рекомендации экспертов.</p> |