DataOps Engineer

DataOps Engineer

Длительность курса: 140 академических часов

1 Введение

1	**Вводное занятие**	**Цели занятия:** объяснить в чем специфика области обработки данных и как DevOps соотносится с DataOps. **Краткое содержание:** обсудим практики DevOps, отличия процессов в области обработки данных, практики DataOps, MLOps, Data Governance.
2	**Архитектуры систем обработки данных (1 часть)**	**Цели занятия:** описать архитектуру аналитического решения под конкретную задачу. **Краткое содержание:** эволюция процессов обработки данных, от Excel до DWH/BI.

3 **Архитектуры систем обработки данных (2 часть)**

Цели занятия:

описать архитектуру аналитического решения под конкретную задачу.

Краткое содержание:

эволюция процессов обработки данных, от MapReduce до современных Data Lakes.

Домашние задания

1 Тест по архитектурам систем обработки данных

Цель: В результате выполнения ДЗ вы проверите и закрепите знания по основным паттернам и элементам систем обработки данных.

Необходимо перейти по ссылке "Тест" из материалов к занятию и заполнить форму. После заполнения и отправки формы - отписаться в чат с преподавателем для проверки.

2 Необходимые инструменты

1 Облака и on-premise

Цели занятия:

научиться использовать облачную инфраструктуру и понимать отличия от on-premise

Краткое содержание:

рассмотрим облачную инфраструктуру, подход Infrastructure-as-a-Code, отличия от on-premise, познакомимся с Yandex.Cloud.

---

2 Terraform

Цели занятия:

научиться использовать Terraform для управления инфраструктурой.

Краткое содержание:

рассмотрим основные компоненты Terraform. Разберемся с их использованием.

Домашние задания

1. Создать виртуальную машину в Yandex.Cloud при помощи Terraform

   Цель: В результате у вас должен быть готова конфигурация terraform и созданая при помощи неё виртуальная машина в Yandex.Cloud

   1. Создать проект в Yandex.Cloud
   2. Применить промокод (если нет промокода, напишите в Slack в чат курса)
   3. Создать конфигурацию terraform
   4. Применить её
   5. Убедиться, что результат соответствует ожиданиям
   6. Загрузить код на github или любой другой репозиторий
   7. Скинуть ссылку на код

---

3 Ansible

Цели занятия:

научиться использовать Ansible для развертывания сервисов.

Краткое содержание:

познакомимся с Ansible на практике и напишем первую роль.

Домашние задания

1 Настроить виртуальную машину на Yandex Cloud при помощи Ansible

Цель: В результате домашнего задания, вы научитесь создавать конфигурацию для автоматизации развёртывания при помощи Ansible и установите необходимый софт на свою виртуальную машину в Yandex.Cloud'e

1. Установить себе на рабочий компьютер Ansible
2. Создать playbook для установки и запуска nginx и postgres на виртуальной машине
3. Конфигурацию playbook'a залить на гит и указать ссылку

---

4 **Docker**

Цели занятия:

научиться использовать Docker для запуска контейнеров и создавать свои образы.

Краткое содержание:

познакомимся с Docker, рассмотрим архитектуру и основные команды для управления, напишем первый docker compose скрипт, включая сборку своего Docker-образа.

Домашние задания

1 Создание docker-compose для стандартного веб-приложения

Цель: В результате выполнения ДЗ вы создадите docker-compose для стандартного веб-приложения

Необходимо создать docker-compose конфиг, для веб-приложения.
- Предполагается, что в качестве базы данных используется postgres.
- Само веб-приложение написано на любом выбранном вами фреймворке.
- Предполагается, что Dockerfile приложения находится в той же директории, что compose файл

docker-compose файл необходимо запушить в репозиторий и в ДЗ указать ссылку на него

---

5 **Q&A**

Цели занятия:

получить ответы на вопросы по ДЗ;
получить ответы на вопросы по приложениям.

Краткое содержание:

типичные ошибки при выполнении ДЗ;
наставники ответят на ваши вопросы.

3 ETL

1 Data Storage

Цели занятия:

выбрать хранилище под конкретные задачи обработки данных.

Краткое содержание:

рассмотрим разные типы хранилищ, применяемых в аналитике;
поговорим об Object Storage, MPP-базах, SQL-движках;
обсудим минусы и плюсы каждого из классов решений.

---

2 Дизайн ETL

Цели занятия:

спроектировать типовые ETL-процессы.

Краткое содержание:

рассмотрим различные типы ETL: пакетный, потоковый, лямбда-архитектура, каппа-архитектура.

Домашние задания

1 Проектирование ETL для отслеживания подозрительной активности на сайте бронирования авиабилетов

   Цель: Предлагается спроектировать ETL-процесс, который отвечает на бизнес-задачу и соответствует требованиям. Подобные задачи встречаются на первоначальных этапах создания систем обработки данных.

   В Материалах к занятию открыть документ с заданием `DataOps_ETLDesign_HW`
   1. Ознакомиться с описанием кейса.
   2. Ответить на вопросы в документе, приложить схемы.
   3. Предоставить любую другую информацию о вашем решении по необходимости, прямо в документе.

3	**Data Ingestion**	**Цели занятия:** интегрировать источники данных. **Краткое содержание:** обсудим способы подключения источников, научимся развертывать и использовать NiFi.
4	**Фреймворки для обработки данных**	**Цели занятия:** рассказать, какие существуют фреймворки для обработки больших данных, и работать с приложениями, написанными на них **Краткое содержание:** рассмотрим различные фреймворки: Spark, Flink, Beam, обсудим их архитектуру, основные операции и окружение, в котором они могут исполняться.

5 **CI, мониторинг и логирование для фреймворков обработки данных**

Цели занятия:

настроить обслуживающие системы (CI, мониторинг и логирование) для приложений, написанных на Spark.

Краткое содержание:

рассмотрим нюансы организации мониторинга и логирования для различных типов аналитических приложений;
применим на практике это к готовому Spark-приложению.

Домашние задания

1 Автоматическое развертывание Spark-приложения

Цель: В этом ДЗ вы настроите CI/CD для развертывания Spark-приложения.

Необходимо:
- развернуть GitLab;
- склонировать репозиторий со Spark-приложением;
- добавить конфигурацию .gitlab-ci.yml для сборки и деплоя Spark-приложения на кластер DataProc;
- выложить обновленный репозиторий с конфигурацией на гитхаб.

6 Оркестрация

Цели занятия:

объяснить, что такое оркестраторы, и деплоить наиболее популярный из них - Airflow.

Краткое содержание:

рассмотрим, что такое Airflow, его архитектуру и как его разворачивать.

Домашние задания

1 Развернуть свой Airflow

Цель: На своей виртуальной машине развернуть Airflow при помощи докер-контейнера

1. Написать простой DAG. Hello World будет достаточно
2. Запустить у себя локально Airflow с этим дагом
3. Запустить airflow на своей виртуальной машине
4. Скинуть ссылку на UI

4 Хранилища и платформы

1 **Архитектура аналитических БД**

Цели занятия:

описать архитектуру аналитических базы данных.

Краткое содержание:

рассмотрим общие архитектурные принципы аналитических баз, внутреннее устройство и особенности применения.

2 **SQL-движки Hive, Presto, Impala**

Цели занятия:

описывать архитектуру распределенных SQL-движков

Краткое содержание:

обсудим роль и архитектуру SQL-движков Hive, Presto, Impala;
проанализируем их недостатки и преимущества относительно аналитических БД.

Домашние задания

1 Реализация операции соединения (JOIN) в MapReduce

Цель: В этом домашнем задании пишем алгоритм, по которому выполняется операция соединения данных (SQL JOIN) в распределенных системах (Hadoop и другие). Это помогает лучше понять механику MapReduce и в будущем поможет выявлять узкие места пайплайнов обработки больших данных и грамотно их оптимизировать.

1. В материалах к занятию выложен файл "DataOps_MapReduce_HW". В нем описана задача, которую нужно решить в парадигме MapReduce.
2. Опишите свое решение словами. Перечислите все шаги map и reduce, что они получают на вход, какая функция применяется к входным данным, и что получается на выходе.
3. По желанию, решение можно описать в виде псевдокода или на языке программирования.
4. Решение напишите в файле с заданием и выложите в личный кабинет.

---

3 **Vertica**

Цели занятия:

научиться разворачивать и использовать БД Vertica.

Краткое содержание:

особенности базы Vertica и её поддержки.

---

4 **GreenPlum**

Цели занятия:

научиться разворачивать и использовать БД GreenPlum.

Краткое содержание:

особенности базы GreenPlum и её поддержки.

---

5 **ClickHouse**

Цели занятия:

научиться разворачивать и использовать БД ClickHouse.

Краткое содержание:

особенности базы ClickHouse и её поддержки.

Домашние задания

1 Деплой MPP-базы

Цель: В результате ДЗ вы научитесь деплоить одну из аналитических БД в режиме кластера.

Необходимо:
- Выбрать одну из аналитических баз (ClickHouse, Greenplum, Vertica);
- Для выбранной базы написать скрипты terraform и ansible для развертывания этой базы в режиме кластера из 3 нод.

---

6 **Q&A**

Цели занятия:

получить ответы на вопросы по ДЗ;
получить ответы на вопросы по приложениям.

Краткое содержание:

типичные ошибки при выполнении ДЗ;
наставники ответят на ваши вопросы.

5 Hadoop

1	**Hadoop**	**Цели занятия:** объяснить, что такое Hadoop и какие задачи он может решать. **Краткое содержание:** быстрое введение в Hadoop: для чего он нужен, из каких сервисов состоит.
2	**Развертывание Hadoop**	**Цели занятия:** развернуть Hadoop. **Краткое содержание:** рассмотрим пример развертывания кластера Hadoop.
3	**Мониторинг Hadoop**	**Цели занятия:** настроить мониторинг кластера Hadoop. **Краткое содержание:** организуем мониторинг сервисов Hadoop.

4 **Безопасность Hadoop**

Цели занятия:

научиться обеспечивать контроль доступа для сервисов Hadoop.

Краткое содержание:

настроим контроль доступа с помощью Ranger и Knox.

Домашние задания

1 Настройка политик безопасности в Apache Ranger

Цель: В процессе выполнения вы

развернете локальный набор сервисов, включающих Ranger и HDFS
настроите политики доступа к HDFS

https://github.com/Gorini4/apache-ranger-docker-poc

6 Практики Data Governance

1 **Практики Data Governance**

Цели занятия:

описать весь жизненный цикл данных в компании и выделять основные процессы работы с данными.

Краткое содержание:

обсудим практики Data Governance.

---

2 **Управление метаданными**

Цели занятия:

собирать метаданные в каталог данных.

Краткое содержание:

рассмотрим инструменты Atlas и Amundsen.

---

3 **Контроль качества данных**

Цели занятия:

организовать тесты для контроля качества данных.

Краткое содержание:

фреймворки для контроля качества данных и их интеграция в оркестратор.

---

4 **Организация песочницы**

Цели занятия:

научиться давать пользователям возможность работать с данными.

Краткое содержание:

HUE и Redash для доступа к данным.

Домашние задания

1 Развертывание системы Data Governance

Цель: В ходе ДЗ вы развернете каталог данных и интегрируете его с хранилищем.

Необходимо:
- Написать скрипты для развертывания Amundsen;
- Интегрировать Amundsen с хранилищем.

7 MLOps

1 Практики MLOps

Цели занятия:

описать жизненный цикл моделей машинного обучения.

Краткое содержание:

выясним, что такое классическая модель машинного обучения;
обсудим, какие этапы проходит модель от создания и до работы в проде.

---

2 Инфраструктура для исследований

Цели занятия:

развернуть сервис для исследований на основе JupyterHub.

Краткое содержание:

рассмотрим, как происходит процесс создания модели машинного обучения;
развернем JupyterHub для подобных исследований.

---

3 Контроль качества моделей

Цели занятия:

настроить эксперименты в MLFlow.

Краткое содержание:

разберем процесс CI для моделей;
организуем процесс контроля метрик моделей.

---

4 **Версионирование данных**

Цели занятия:

научиться версионировать наборы данных с помощью DVC.

Краткое содержание:

обсудим хранение фичей, роль тренировочных и тестовых наборов данных;
научимся версионировать данные с помощью DVC.

---

5 **Варианты деплоя моделей - REST**

Цели занятия:

упаковать модели в микросервисы.

Краткое содержание:

изучим различные варианты "упаковки" моделей для развертывания в проде;
научимся деплоить модели в виде микросервисов.

---

6 **Варианты деплоя моделей - Spark**

Цели занятия:

упаковать модели в джобы для выполнения на кластере.

Краткое содержание:

научимся деплоить модели в виде Spark-джоб;
рассмотрим различные ML-библиотеки и способы переноса их моделей из Python в Java.

Домашние задания

1 Деплой модели машинного обучения

Цель: В результате ДЗ вы развернете ML-модель и необходимую инфраструктуру.

Необходимо:
- Написать скрипты для развертывания MLFlow;
- Упаковать модель в контейнер;
- Настроить пайплайн для переобучения и деплоя модели.

8 Проектный модуль

1 Выбор темы и организация проектной работы

Цели занятия:

выбрать и обсудить тему проектной работы;
спланировать работу над проектом;
ознакомиться с регламентом работы над проектом

Краткое содержание:

правила работы над проектом и специфика проведения итоговой защиты;
требования к результату проекта и итоговой документации

Домашние задания

1 Проектная работа

Цель: В этом ДЗ необходимо выбрать тему проектной работы и закрепить её в чате по дз. Написать проект и защить.

1. Выбрать тему и утвердить
2. Сделать проект
3. Защитить проект

---

2 Консультация по проектам и домашним заданиям

Цели занятия:

получить ответы на вопросы по проекту, ДЗ и по курсу

Краткое содержание:

вопросы по улучшению и оптимизации работы над проектом;
затруднения при выполнении ДЗ;
вопросы по программе

---

3 **Защита проектных работ**

Цели занятия:

защитить проект и получить рекомендации экспертов.

Краткое содержание:

презентация проектов перед комиссией;
вопросы и комментарии по проектам.

---

4 **Подведение итогов курса**

Цели занятия:

узнать, как получить сертификат об окончании курса, как взаимодействовать после окончания курса с OTUS и преподавателями, какие вакансии и позиции есть для выпускников (опционально - в России и за рубежом) и на какие компании стоит обратить внимание.

Краткое содержание:

организационные вопросы;
рынок вакансий по направлению;
статистика курса и вопросы по курсу.