Полная программа

Spark Developer

Spark Developer

Длительность курса: 106 часов

Модуль 1. Введение

Тема 1
Что такое Spark

Цель занятия
- познакомиться с API для работы с приложениями Spark: spark-shell, spark-submit, Idea, jupiter;
- познакомиться с режимами работы Spark-приложения: local, client, cluster;
- познакомиться с менеджерами ресурсов для Spark;
- познакомиться с основными компонентами архитектура Spark приложения;
- познакомиться с возможностями управления.

Краткое содержание
- Spark;
- spark-shell;
- Zeppelin;
- Jupyter;
- spark-submit.

Домашние задания

Развёртывание среды работы со Spark
Цель
Научиться разворачивать и запускать Spark для последующих разработки и запуска приложений.

Тема 2
Первые шаги в Scala

Цель занятия
- познакомимся со Scala;
- изучить особенности Scala;
- изучить как начать работать со Scala;
- изучить базовые понятия Scala.

Краткое содержание
- Scala;
- sbt;
- IDEA.

Тема 3
Дальнейшие шаги в Scala

Цель занятия
- углубить понимание Scala.

Краткое содержание
- Scala;
- ФП;
- ООП.

Тема 4
Практика работы со Scala

Цель занятия
- попрактиковаться в работе со Scala;
- создать простое приложение.

Краткое содержание
- Scala;
- sbt или maven;
- IDEA.

Домашние задания

Коллекции данных
Цель
Научиться практическому использованию наиболее употребляемых методов работы с данными и немного вспомнить математику (теорию вероятности), в виду того, что скалу очень часто применяют для анализа данных, то это будет полезно.

Модуль 2. Большие данные

Тема 1
Hadoop, HDFS

Цель занятия
- познакомиться с Hadoop;
- изучить HDFS.

Краткое содержание
- Hadoop;
- HDFS.

Тема 2
Обзор Hive

Цель занятия
- познакомиться с Hive;
- изучить архитектуру Hive;
- рассмотреть предназначение Hive и аналогов.

Краткое содержание
- Hive;
- Impala;
- Presto.

Тема 3
HiveQL

Цель занятия
- научиться писать запросы на HiveQL;
- научиться создавать таблицы в Hive.

Краткое содержание
- форматы файлов;
- типы данных;
- DDL;
- DML;
- запросы к данным.

Домашние задания

Построение аналитической витрины в Hive
Цель
В это ДЗ вы напишите код для построения аналитической витрины в Hive.

Тема 4
Spark в Hadoop, YARN

Цель занятия
- изучить как запускать Spark в Hadoop.

Краткое содержание
- YARN.

Тема 5
Spark в Kubernetes

Цель занятия
- изучить как запускать Spark в Kubernetes.

Краткое содержание
- Kubernetes.

Модуль 3. API

Тема 1
DataFrame

Цель занятия
- изучить DataFrame API.

Краткое содержание
- DataFrame.

Домашние задания

Аналитическая витрина на основе сырых данных, используя Spark
Цель
Выполнив домашнее задание вы получите опыт работы с RDD API, DataFrame API, Dataset API. Научитесь строить аналитическую витрину на основе сырых данных, используя Spark и различные API.

Тема 2
Dataset, SparkSQL

Цель занятия
- познакомиться с Dataset API;
- изучить отличия от DataFrame;
- запустить SQL в Spark.

Краткое содержание
- Dataset;
- SparkSQL.

Тема 3
RDD

Цель занятия
- изучить RDD API.

Краткое содержание
- RDD.

Тема 4
UDF и UDAF

Цель занятия
- изучить UDF и UDAF;
- изучить особенности.

Краткое содержание
- UDF;
- UDAF.

Тема 5
Apache Arrow в PySpark

Цель занятия
- изучить Apache Arrow и его использование в Spark.

Краткое содержание
- Arrow;
- PySpark.

Тема 6
Pandas API

Цель занятия
- изучить Pandas API.

Краткое содержание
- Pandas;
- PySpark.

Модуль 4. Источники данных

Тема 1
Файлы и их форматы

Цель занятия
- изучить как в Spark работать с файлами;
- изучить особенности форматов файлов.

Краткое содержание
- Data Sources;
- Parquet;
- ORC;
- CSV;
- JSON;
- Avro;
- Protobuf.

Тема 2
Базы данных, Hive

Цель занятия
- изучить как приложению Spark подключиться к СУБД по JDBC и к Hive.

Краткое содержание
- JDBC;
- Hive.

Тема 3
Собственный источник данных

Цель занятия
- изучить как создать собственный источник данных (коннектор).

Краткое содержание
- DataSource API.

Домашние задания

Разработка собственного коннектора на Spark
Цель
В данном ДЗ вы поработаете с DataSource API V2, научитесь писать свои собственные коннекторы для Spark.
Задача - доработать data source для Postgres для партиционированного чтения.

Тема 4
(Бонус) Kafka

Цель занятия
- научиться описывать архитектуру Apache Kafka;
- научиться использовать консольные утилиты и клиенты для работы с Kafka;
- научиться оптимизировать запись и чтение топиков Kafka.

Краткое содержание
- Kafka.

Тема 5
Structured Streaming

Цель занятия
- изучить структурный стриминг;
- изучить интеграцию с Kafka.

Краткое содержание
- Structured Streaming;
- Kafka.

Домашние задания

Читаем и пишем в Kafka, используя Structured Streaming
Цель
Выполнив это ДЗ вы научитесь используя Spark Structured Streaming читать из Kafka и писать в Kafka

Модуль 5. Дополнительные возможности

Тема 1
Spark ML

Цель занятия
- рассмотреть использование Spark для ML;
- изучить разработку и внедрение моделей ML в Spark.

Краткое содержание
- SparkML.

Домашние задания

Разработка модели в Spark ML
Цель
Выполнив это ДЗ вы научитесь обучать и сохранять модели Spark ML.

Тема 2
Работа с графами

Цель занятия
- познакомиться с графами;
- узнать, как работать с графами в Spark.

Краткое содержание
- GraphX;
- GraphFrames.

Тема 3
Тестирование приложений Spark

Цель занятия
- рассмотреть, как тестировать Spark приложения.

Краткое содержание
- Test Containers;
- Scalacheck.

Домашние задания

ДЗ на тему урока

Тема 4
Консультация по домашним заданиям

Цель занятия
- получить ответы на вопросы по ДЗ.

Краткое содержание
- типичные ошибки при выполнении ДЗ;
- ответы на ваши вопросы.

Модуль 6. Промышленное использование

Тема 1
Оркестрация процессов обработки данных

Цель занятия
- узнать зачем нужны оркестраторы и как их использовать;
- познакомиться с Oozie и Airflow.

Краткое содержание
- Oozie;
- Airflow.

Тема 2
Мониторинг Spark приложений

Цель занятия
- познакомиться с мониторингом и его использованием в Spark.

Краткое содержание
- SparkUI;
- Grafana.

Тема 3
Методы оптимизации приложений Spark

Цель занятия
- рассмотреть типичные проблемы Spark приложений;
- изучить способы оптимизации Spark приложений.

Краткое содержание
- Out of Memory Exception;
- Shuffle;
- Adaptive Query Execution;
- Dynamic Partition Pruning.

Модуль 7. Проектная работа

Тема 1
Выбор темы и организация проектной работы

Цель 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

Проектная работа
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.

Тема 2
Консультация по проектам и домашним заданиям

Цель 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.

Тема 3
Защита проектных работ

Цель занятия
- защитить проект и получить рекомендации экспертов;
- 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
- презентация проектов перед комиссией;
- вопросы и комментарии по проектам;
- организационные вопросы;
- рынок вакансий по направлению;
- проведение собеседований;
- статистика курса и вопросы по курсу.