“Обработка больших данных”

Big Data Processing

by

Петр Ермаков https://www.hse.ru/staff/ermakov

Аннотация

Обработка больших объемов данных является одной из ключевых компетенций в современных IT-профессиях. В курсе будут рассмотрены современные подходы к обработке больших данных. Рассмотрим экосистему Hadoop, познакомимся с MapReduce подходом и научимся создавать код под Hadoop-кластер на (java и python). Узнаем как работает HDFS, HIVE. Узнаем когда нужен Spark и в чем его главные плюсы и минусы. Научимся обрабатывать данные в режиме реального времени с использованием Spark Streaming. Познакомимся и научимся применять алгоритмы машинного обучения на Spark.

Пререквизиты

- Базы данных
- Программирование на Python

Система оценки:

Текущий контроль: 7 задач - по 12 каждая

Экзамен: 40 баллов

8-10 баллов выставляется при сумме больше 100 баллов;

6-7 баллов – при сумме >90 баллов;

5-6 баллов – при сумме >80 баллов.

Экзамен засчитывается автоматом, если качественно выполнены все задания.

Состав учебного курса

1. Методология MapReduce, введение в Hadoop
2. Hadoop Streaming, MapReduce на java и python
3. HDFS, Parquet
4. Hive
5. Spark
6. Стриминговая обработка (в том числе Spark Streaming)
7. Spark ML (4 занятия)
8. Работа с GIS данными в Hadoop
9. Работа с графами
10. Машинное обучение на кластере с использованием H2O (2 занятия)

Список литературы

1. Hadoop: The Definitive Guide, 4th Edition, Storage and Analysis at Internet Scale, Tom White, O'Reilly Media
2. Learning Spark, Lightning-Fast Big Data Analysis, Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell, O'Reilly Media
3. Advanced Analytics with Spark, Patterns for Learning from Data at Scale, Sandy Ryza, Uri Laserson, Josh Wills, Sean Owen, O'Reilly Media