

Курс: Hadoop для Data Developer

Длительность: 40 ак. часов

О курсе

Данный курс направлен на формирование практических и теоретических навыков планирования, формирования и сопровождения **Data Lake** (озеро данных). Рассматриваются примеры интеграции, настройки и обслуживания "**pipelines**" - традиционных источников поступления данных (корпоративные базы данных, web логи, файловые системы, интернет данные, транзакции) для хранения и последующего анализа больших данных. Практические занятия выполняются в **AWS** и локальной кластерной системе с использованием дистрибутива **Cloudera Hadoop**.

Аудитория

Специалисты по работе с большими данными ответственные за настройку и сопровождение ввода данных в **Data Lake**, а также желающие получить теоретические знания и практические навыки по подготовке больших данных, специфики использования процессов **ETL** в кластерах **Hadoop**, и организации **Batch**, **stream** и **real-time** процессинга больших данных с использованием компонентов экосистемы **Hadoop**.

Соотношение теории к практике 40/60

Предварительная подготовка

- Начальный опыт работы в **Unix/SQL**, текстовый редактор **vi**
- Начальный опыт работы в **Hadoop** (желателен)

Программа курса

1. Основные концепции Hadoop

Основы Hadoop. Жизненный цикл аналитики больших данных. Хранение, накопление, подготовка и процессинг больших данных. Тенденции развития **Hadoop**.

Архитектура HDFS. Операции чтения и записи, назначения **HDFS**. Блоки **HDFS**. Основные команды работы с **HDFS**.

Ведение в MapReduce. Компоненты **MapReduce**. Работа программы **MapReduce**. Архитектура **YARN**. Способы обработки распределенных данных с использованием **Apache Spark**, **YARN** и **MapReduce v2/v3**.

Управление ресурсами и очередями задач. FIFO/Capacity/Fair scheduler.

2. Инструменты управления кластером

Выполнение базовых операций с **Cloudera Manager**. Настройка компонент Apache **ZooKeeper**, **Oozie**. Создание и управление запросами и данными с использованием сервиса **Hue**.

3. Хранение данных в HadoopDFS

Хранение файлов в **HDFS**: сжатие, sequence файлы. Формат **AVRO, RCfile, ORC, Parquet**.

Введение в **Apache Pig**: формат хранения данных, сложные и вложенные типы данных, синтаксис **Pig Latin**, оптимизация операций **Join**.

4. Apache Spark

Архитектура **Apache Spark**. Введение в **Spark: RDD & Datasets**, доступ к внешним данным из **Spark**, интеграция с **Hadoop**, запуск приложений, **Spark streaming, Spark SQL, Datasets, Dataframes**.

5. Импорт/экспорт данных в кластер Hadoop.

Импорт и обработка данных в кластере **Hadoop**. Интеграция с реляционными базами данных. Структура хранения данных в таблицах. Технологии **NoSQL**. Сравнительная характеристика решений **Hadoop SQL**.

Введение в Sqoop: импорт и экспорт данных **Sqoop**, формат файлов, инкрементальный импорт, **Hive** экспорт.

6. Apache Hive

Введение в **Hive**: структура **Hive** таблиц, синтаксис **HiveQL**, формат хранения файлов, работа с внешними и внутренними таблицами **Hive**, оптимизация **Join** операций. Операции импорта и экспорта данных и взаимодействия с внешними источниками. Настройка производительности.

7. Cloudera Impala

Введение в **Cloudera Impala**: архитектура и компоненты, **Impala** синтаксис, типы данных, написание запросов, загрузка данных, взаимодействие **Spark, Hive**. Оптимизация **Impala** запросов.

8. Поточковые данные

Event Processing System. Импорт потоковых данных в кластер. Использование **Kafka** для работы с потоковыми данными.

Использование **Flume** для работы с потоковыми данными. Визуализация потоковых данных.

Список практических занятий:

- Автоматическая установка 3х-узлового кластера в облаке **Amazon Web Services** с использованием **Cloudera Manager** и поддержка базовых операций с кластером **Hadoop** и **HDFS**.
- Управление ресурсами и запуском задач с использованием **YARN MapReduce**.
- Использование **Apache Pig** для подготовки данных, операции **JOIN**
- Использование **Apache Hive** для анализа данных
- Оптимизация запросов **JOIN** в **Apache Hive**
- Настройка **partition** и **bucket** в **Apache Hive**
- Инкрементальный импорт/экспорт данных с помощью **Apache sqoop**
- **SQL** аналитика данных с помощью **Cloudera Impala**
- **Batch** процессинг данных с использованием **Apache Spark**
- **Потоковая** обработка данных с использованием **Apache Spark**
- Импорт данных с помощью **Apache Flume**
- Построение **Event Processing System** с использованием **Apache Flume** и **Kafka**
- Создание и управление запросами **sqoop, MapReduce, Hive, Impala** с использованием веб-интерфейса **HUE**