ШКОЛА БОЛЬШИХ ДАННЫХ

Курс HDDE: Hadoop для инженеров данных

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 "**pipeline**s" - 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 дистрибутивов **Cloudera, HortonWorks, АренаДата.**

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**. Жизненый 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/ Apache Ambari.**
   - ✓ Создание и управление запросами и данными с использованием сервиса **Hue/ Apache Ambari Views**.

3. **Хранение данных в HadoopDFS**
   - ✓ Хранение файлов в **HDFS**: сжатие, sequence файлы. Формат **AVRO**, **CSV**, **ORC**, **Parquet**.
   - ✓ Введение в **Apache Pig**: формат хранения данных, сложные и вложенные типы данных, синтаксис **Pig Latin**, оптимизация операций **Join**.

4. **Импорт/экспорт данных в кластер Hadoop - формирование Data Lakes.**
   - ✓ Импорт и обработка данных в кластере **Hadoop**.
   - ✓ Интеграция с реляционными базами данных.
   - ✓ Структура хранения данных в таблицах.
   - ✓ Сравнительная характеристика решений **Hadoop SQL.**
   - ✓ **Введение в Sqoop:** импорт и экспорт данных **Sqoop**, формат файлов, инкриментальный импорт, **Hive** экспорт.

5. **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.
   - ✓ Hive LLAP, Hive on Spark/Tez

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

7. **Потоковые данные**
   - ✓ **Event Processing System**. Импорт потоковых данных в кластер.
   - ✓ Использование **Kafka** для работы с потоковыми данными.
   - ✓ Использование **Flume** для работы с потоковыми данными.
   - ✓ Интеграция **Flume + Kafka**

Список практических занятий:

- Автоматическая установка **3x**-узлового кластера в облаке **Amazon Web Services** с использованием **Cloudera Manager/Apache Ambari** и поддержка базовых операций с кластером **Hadoop** и **HDFS**. (опционально)
- Управление очередями ресурсов и запуском задач с использованием **YARN**.

- 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**
- 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/ Ambari Views**
- Построение **Dataflow** с использованием **Apache NiFi**

info@bigdataschool.ru
+7 (985) 162-29-63