Программа курса
«HDDE: Hadoop для инженеров данных»

О курсе: 5-дневный практический тренинг по batch/streaming обработке больших данных средствами экосистемы Apache Hadoop: Airflow, Spark, Flume, Sqoop, Hive, а также Kafka для организации озера данных (Data Lake) на кластере Hadoop версии 3 и процессов ETL/ELT.
Аудитория: Курс Hadoop для инженеров данных ориентирован на специалистов по работе с большими данными, которые отвечают за настройку и сопровождение ввода данных в Data Lake и хотят получить теоретические знания и практические навыки по подготовке массивов Big Data и специфике процессов ETL/ELT в кластерах Hadoop. Также на нашем курсе Data Engineer освоит тонкости организации pipelines в Hadoop, Batch, stream и real-time процессинга больших данных с использованием компонентов экосистемы Хадуп.
Уровень подготовки:

- Знание базовых команд Linux (опыт работы с командной строкой, файловой системой, POSIX, текстовыми редакторами vi, nano)
- Начальный опыт работы с SQL

Продолжительность курса: 40 академических часов, 5 дней по 8 ак. часов дистанционно

Содержание программы

1. Основные концепции Hadoop и Data Lake

- Основы Hadoop. Основные компоненты, парадигма, история и тенденции развития
- Современные хранилища данных, Data Lake, его архитектура

2. Map Reduce и Yarn

- Ведение в MapReduce. Этапы выполнения задачи в MapReduce и подход к программированию
- Архитектура и задачи YARN. Управление ресурсами и очередями задач, FIFO/Capacity/Fair scheduler

3. Хранение данных в HDFS

- Архитектура HDFS. Операции чтения и записи, блоки HDFS
- Основные команды работы с HDFS
- Дополнительные возможности и особенности HDFS

4. Импорт/экспорт данных в кластер Hadoop — формирование Data Lake

- Импорт и обработка данных в кластере Hadoop
- Интеграция с реляционными базами данных
- Структура хранения данных в таблицах
- Введение в Sqoop: импорт и экспорт данных из реляционных источников

5. Apache Hive

- Введение в Hive и соответствие DDL операций структуре хранения
- Работа с внешними и внутренними таблицами Hive
- Партиционирование данных
- Hive LLAP, Hive on Spark/Tez
- Хранение данных в HDFS: сжатие и форматы файлов (AVRO, ORC, Parquet)

6. Основы Apache Spark

- Архитектура и состав Apache Spark

- Основные абстракции (Dataframe, RDD)
- Spark SQL
- Ввод и вывод данных в Apache Spark

7. Введение в Cloudera Impala

- Введение в Cloudera Impala: особенности архитектура и компоненты
- Взаимодействие Spark, Hive

8. Введение в Apache HBase

- Архитектура и состав Apache HBase
- Основные абстракции и язык запросов

9. Введение в Apache Kafka

- Архитектура и состав Apache Kafka
- Партиции, топики, управление смещением
- Основные API

10. Введение в Apache Airflow

- Архитектура и состав Apache Airflow
- Основные абстракции (DAG, оператор, сенсор)
- Основные операторы (Bash Operator, Python Operator)

Список практических занятий:

- Выполнение и анализ работы Map Reduce приложений
- Особенности запуска задач и использование командной строки YARN
- Работа с HDFS (интерфейс командной строки)
- Импорт/экспорт данных с помощью Apache Sqoop
- Использование Apache Hive для анализа данных
- Обработка данных с использованием Structured API Apache Spark
- Сравнение производительности SQL движков (Hive, Spark, Impala)
- Работа в командной строке с Apache HBase
- Использование Consumer и Producer API в Apache Kafka
- Построение Workflow с использованием Apache Airflow