Программа курса
«INTR: Основы Hadoop»

О курсе: 3-дня практического обучения по установке и первоначальной настройке кластера Apache Hadoop — основы Big Data для начинающих и специалистов. Практическое обучение Хадуп для системных администраторов, архитекторов и разработчиков Big Data.
Курс «Основы Hadoop» представляет сокращенную версию курса «Администрирование кластера Hadoop» и проводится параллельно с данным курсом в 3 дня, согласно утвержденной программе, на платформе Arenadata Hadoop Community Edition или Apache Hadoop.
Аудитория: Курс «Основы Hadoop» ориентирован на начинающих и опытных ИТ-специалистов в области больших данных, которые хотят получить теоретические знания и прикладные навыки по установке, настройке и использованию кластера Apache Hadoop версии 3 на базе дистрибутива Arenadata Hadoop Community Edition (Cloudera Data Platform (CDP) Private Cloud для ознакомления).
Уровень подготовки:

- Базовый опыт работы в Linux (обязательно)
- Опыт работы с любым текстовым редактором vi, nano

Продолжительность курса: 24 академических часа, 3 дней по 8 ак. часов дистанционно

Содержание программы

1. Основы Hadoop и Big Data

- Что такое Big Data. Понимание проблемы Big Data
- Эволюция систем распределенных вычислений Hadoop
- Концепция Data Lake и pipelines
- Схемы организации Data Lakes с использованием кластеров Hadoop, NoSQL и платформ потоковой обработки данных

2. Архитектура Apache Hadoop

- Hadoop сервисы и основные компоненты. Name node. DataNode.
- YARN сервис-планировщик
- Демоны HDFS
- Отказоустойчивость и высокая доступность

3. Hadoop Distributed File System

- Архитектура HDFS. Блоки HDFS
- Основные команды работы с HDFS
- Операции чтения и записи, назначения HDFS
- Дисковые квоты. Поддержка компрессии
- Основные форматы хранения данных TXT, AVRO, ORC, Parquet, Sequence файлы
- Импорт (загрузка) данных на HDFS

4. MapReduce

- Ведение в MapReduce. Компоненты MapReduce. Работа программ MapReduce. YARN MapReduce v2/3.
- Ограничения и параметры MapReduce и YARN
- Управление запуском пользовательских задач (jobs) под MapReduce.

5. Дизайн кластера Hadoop

- Сравнение дистрибутивов и версий Hadoop 2/3 (Arenadata Hadoop, Cloudera Data Platform, Apache Hadoop): различия и ограничения
- Требования программного и аппаратного обеспечения
- Планирование кластера
- Масштабирование кластера Hadoop.
- Интеграция с другими решениями: streaming (DataFlow), NoSQL

6. Установка кластера Arenadata Hadoop

- Оптимизация OS для узлов кластера
- Установка Hadoop кластера с использованием ADCM (Arenadata Cluster Manager)
- Выбор начальной конфигурации
- Начальная конфигурация HDFS и MapReduce
- Файлы логов и конфигурации
- Установка Hadoop клиентов
- Установка Hadoop кластера в облаке

7. Операции обслуживания кластера Hadoop

- Дисковая подсистема
- Квоты
- Остановка, запуск, перезапуск (Graceful Shutdown)
- Управление узлами
- Управление обновлениями и создание локального репозитория

8. Оптимизация и управление ресурсами

- Производительность. Файловая система. Data Node и Data layout и партиционирование, bucketing
- Планировщики: FIFO Scheduler. Планировщик емкости (Capacity Scheduler). Гранулярное управление ресурсами (Fair Scheduler). Защита очередей и доминантное управление ресурсами DRF

9. Управление кластером Arenadata Hadoop с использованием ADCM

- Основные операции и задачи ADCM
- Мониторинг кластера.
- Диагностика и разрешение проблем с ADCM

10. Инструментарий Apache Hadoop экосистемы

- Графический интерфейс сервиса HUE/Zeppelin
- Основы Apache Zookeeper
- Введение в Hadoop SQL: Apache Hive, понятие Hive таблицы, установка Hive
- Использование Apache Sqoop — установка и выполнение базовых операций
- Обзор и назначение компонент: Apache Spark, Apache Solr, Apache HBase, Apache Phoenix, Apache Flink, Apache Airflow

Примерный список практических занятий по курсу «Основы Hadoop»:

- Установка кластера и настройка Arenadata Cluster Manager (ADCM)
- Настройка оффлайн репозитория для установки кластера Arenadata Hadoop и RHEL/Centos
- Ручная установка 3х-узлового кластера Hadoop версии 3 с дистрибутива Arenadata Cluster Manager (ADCM) в облаке Amazon Web Services с использованием ADCM
- Базовые операции обслуживания кластера Hadoop и файловые операции HDFS
- Управление ресурсами и запуском задач с использованием YARN и MapReduce
- Знакомство с SQL интерфейсом доступа Apache Hive
- Выполнение базовых операций импорта/экспорта с применением Apache sqoop
- Применение веб-интерфейса HUE/Zeppelin (опционально)

Примечание:

- Доступ к лабораторному стенду на Yandex Cloud предоставляется на время учебных курсов с 8:30 до 18:30 (возможно продление времени по запросу)

- Практические занятия с меткой (опционально) выполняются по желанию и при наличии свободного времени у слушателей