

Workshop: Администрирование Hadoop кластера под управлением Apache Ambari (HortonWorks Data Platform/ArenaData Hadoop)

Примечание: с 1 января 2019 года данный курс проводится в объединенном формате по дистрибутивам Hadoop версии 2 компаний **Cloudera/HortonWorks/ArenaData** на выбор для пользователей. Для корпоративного формата обучения возможна выделенная программа по одной версии дистрибутива Hadoop (по запросу).

Длительность: 40 ак. часов

О курсе

Apache Hadoop является наиболее популярной открытой платформой для распределенного хранения больших данных и параллельных вычислений. В рамках данного курса вы получите теоретические знания и практические опыт по планированию и развертыванию распределенных вычислительных кластеров на базе **Hadoop** на базе дистрибутива **HortonWorks Data Platform/ArenaData Hadoop**, мониторингу и оптимизации производительности системы, резервному копированию и аварийному восстановлению узлов кластера и отдельных компонент, настройки безопасности системы **Kerberos** на базе **Hadoop**. Курс построен на сквозных практических примерах развертывания и администрирования **Hadoop** кластера, в том числе в облачной инфраструктуре; использования компонент **Hadoop** для запуска задач распределенных вычислений с тестовыми данными. Практические занятия выполняются в кластерной среде **Amazon Web Services** с использованием дистрибутивов **HortonWorks Data Platform/ArenaData Hadoop** программного обеспечения **Apache Ambari**.

Аудитория

Системные администраторы, системные архитекторы, разработчики Hadoop желающие получить практические навыки по установке, конфигурированию, обслуживанию и управлению кластером **Hadoop** с использованием дистрибутива **HortonWorks Data Platform/ArenaData Hadoop**.

Соотношение теории к практике 40/60

Необходимая предварительная подготовка

- Начальный опыт работы в **Unix**, опыт работы с текстовым редактором **vi** (желателен)

Программа

<mailto:sales@bigdataschool.ru>

+7 (985) 162-29-63

1. **Введение в Big Data**

Что такое **Big Data**. Понимание проблемы **Big Data**. Эволюция систем распределенных вычислений **Hadoop**. Принципы формирования **pipelines** и **Data Lake**.

2. **Архитектура Apache Hadoop**

Hadoop сервисы и основные компоненты. **Name node**. **Data Node**. **YARN** сервис. Планировщик. **HDFS**. Отказоустойчивость и высокая доступность.

3. **Hadoop Distributed File System**

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

4. **MapReduce**

Введение в **MapReduce**. Компоненты **MapReduce**. Работа программ **MapReduce**. **YARN MapReduce v2**. Ограничения и параметры **MapReduce** и **YARN**. Управление запуском пользовательских задач (jobs) под **MapReduce**.

5. **Дизайн кластера Hadoop**

Сравнение дистрибутивов и версий **Hadoop 2/3 (HortonWorks Data Platform, Cloudera Distributed Hadoop, MapR, ArenaData Hadoop)**: различия и ограничения.

Требования программного и аппаратного обеспечения. Планирование кластера. Масштабирование кластера **Hadoop**. Отказоустойчивость **Hadoop**. **Federated NameNode**. **Hadoop** в облаке.

Сравнение **Cloud** решений для **Hadoop**. **Amazon EMR**.

Интеграция с другими решениями: **streaming (DataFlow)**, **NoSQL**.

6. **Установка кластера**

Установка **Hadoop** кластера. Выбор начальной конфигурации. Оптимизация уровня ядра для узлов. Начальная конфигурация **HDFS** и **MapReduce**. Файлы логов и конфигураций. Установка **Hadoop** клиентов. Установка **Hadoop** кластера в облаке.

Установка и настройка кластера **Hadoop** в изолированном окружении (offline).

7. **Операции обслуживания кластера Hadoop**

Дисковая подсистема. Квоты. Остановка, запуск, перезапуск. Управление узлами. Сетевая топология. Управление обновлениями и создание локального репозитория.

8. **Оптимизация и управление ресурсами**

Поиск узких мест. Производительность. Файловая система. **Data Node**. Сетевая производительность. **FIFO scheduler**. Планировщик емкости (**Capacity scheduler**). Гранулярное управление ресурсами (**Fair scheduler**). Защита очередей и доминантное управление ресурсами **DRF**.

9. **Управление кластером Hadoop с использованием Apache Ambari**

Установка **Apache Ambari**. Интерфейс управления **Apache Ambari**. Базовые операции обслуживания и управление задачами с использованием **Apache Ambari**. Диагностика и **troubleshooting** с **Apache Ambari**.

10. Безопасность Hadoop

Безопасность по умолчанию. Встроенные компоненты безопасности дистрибутива **HortonWorks/ArenaData Hadoop: Apache Ranger, Apache Atlas, Apache Knox.**

Многопользовательский режим. Аутентификация и авторизация. **Kerberos, keytabs, principals.** Установка и конфигурирование **Kerberos** в **Hadoop**. Аудит доступа.

Резервное копирование и аварийное восстановление. Репликация данных и **snapshotting**.

Конфигурирование высокой доступности **Name node (HA)**.

Best practices HortonWorks/ArenaData Hadoop .

11. Мониторинг

Apache Zookeeper. Встроенные средства мониторинга **Apache Ambari Metrics**. Логи сервисов и компонент. Внешние системы мониторинга: **Zabbix, JMX.**

12. Troubleshooting

Name Node. Восстановление **Name node. Data Node.**

13. Инструментарий Hadoop экосистемы дистрибутива HortonWorks/ArenaData Hadoop

Графический интерфейс сервиса **Zeppelin/HUE.**

Введение **Apache Pig.**

Введение в **Apache Hive/Tez**, понятие **Hive** таблицы, установка **Hive/Tez.**

Введение в **Apache sqoop** - установка и выполнение базовых операций.

Введение в **Apache Flume** - установка и выполнение базовых операций.

Введение в **Apache Spark** - установка и выполнение базовых операций.

Обзор и назначение компонент: **Apache Kafka, Apache HBase, Apache NiFi, Apache Flink, Apache Zookeeper.**

Примерный список практических занятий:

- Ручная установка кластера **Hadoop** с дистрибутива **HortonWorks Data Platform/ArenaData Hadoop** на локальной системе **3x-узловый кластер**
- Установка **3x-узлового кластера** в облаке **Amazon Web Services** с использованием **Apache Ambari**
- Базовые операции с кластером **Hadoop** и файловые операции **HDFS.**
- Управление ресурсами и запуском задач с использованием **YARN MapReduce.**
- Управление кластером с использованием **Apache Ambari** (развертывание сервисов, репликация, мониторинг, alerting и т.д.)
- Настройка аутентификации **Kerberos** для кластера **Hadoop** под управление **Apache Ambari**
- Установка и выполнение базовых операций в **Apache Hive, Apache sqoop, Apache Flume, Apache Spark.**
- Выполнение задач в веб-интерфейсе **Zeppelin(опционально).**
- Настройка высокой доступности **Name Node.**
- Настройка мониторинга кластера **Hadoop** с использованием **Zabbix (опционально)**

Примечание:

- Доступ к лабораторному стенду на **Amazon Web Services** 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