Основы работы с аналитическими инструментам кластера Hadoop

Цель тренинга:

Предоставить знания и навыки, необходимые для анализа данных в кластере Hadoop: где найти нужные данные, как управлять данными в кластере, как писать оптимальные запросы; познакомить слушателей с синтаксисом языков, используемых в инструментах кластера; сформировать понимание возможностей и ограничений платформы. Рассматриваются базовые инструменты, входящие в стандартную поставку платформы Horton Data Platform: Hive, Pig, также будет дана справочная информация о перспективных технологиях.

Приобретаемые практические знания:

После окончания курса слушатели смогут:

1. описать принципы работы распределенной файловой системы HDFS, механизм распределённых вычислений MapReduce, экосистему и архитектуру кластера;
2. выполнять выборку и анализ данных в кластере Hadoop.

После окончания тренинга, участники тренинга смогут самостоятельно:

- использовать командную строку и графический интерфейс пользователя для работы с данными в кластере;
- определять, какие данные нужны для решения задачи и где их найти;
- разрабатывать и выполнять запросы на выборку данных в кластере;
- расширять функционал языков своими алгоритмами обработки;

Для кого предназначен тренинг:

Курс рассчитан на системных аналитиков, разработчиков, системных архитекторов или администраторов баз данных.

Метод проведения:

лекции

практика

Тренеры, которые проводят данный тренинг:

Роман Бойко

Продолжительность:

3 дня

Максимальный размер группы:

до 8 человек

Программа тренинга:

День	Время	Темы	Примечание
1	9:30 – 18:00	**Вступление** • Знакомство • Структура курса • Организационные вопросы **Основы работы с кластером** • Как появился Hadoop • Файловая система HDFS • Модель распределённых вычислений MapReduce • Экосистема Hadoop • Источники данных в кластере **Практическая работа:** Знакомство с инструментами Hadoop **Введение в Pig** • Что такое Pig • Особенности Pig • Для каких задач применяется Pig • Как работать с Pig **Основы анализа данных в Pig** • Синтаксис языка • Загрузка данных • Простые типы данных • Ключевые концепции • Вывод и сохранение результатов • Схемы данных • Фильтрация и сортировка • Часто используемые функции **Практическая работа:** Использование Pig для предобработки данных	Лекции, практика

2	9:30 – 18:00	**Обработка сложных типов данных в Pig** • Форматы хранения • Виды сложных типов данных • Группировка данных • Встроенные функции для работы со сложными типами • Операции со сгруппированными данными **Практическая работа:** Применение Pig для решения аналитических задач **Работа с несколькими источниками данных в Pig** • Техники комбинирования источников • Операции JOIN, CROSS, UNION • Способы разбиения источников данных **Практическая работа:** Комбинирование нескольких источников данных в Pig **Расширения Pig** • Параметризация запросов • Макросы и повторное использование кода • Пользовательские функции • Пакеты пользовательских функций • Использование других языков для обработки данных в Pig **Практическая работа:** Работа с пользовательскими функциями **Поиск неисправностей и оптимизация скриптов Pig** • Способы поиска проблем • Логирование исполнения скрипта • Использование GUI Hadoop • Отладка на небольшой выборке данных • План исполнения запроса • Советы по оптимизации **Введение в Hive** • Что такое Hive? • Хранение данных и метаданные • Сравнение с RDBMS • Для каких задач применяется Hive • Интерфейсы для работы с Hive	Лекции, Практика

3	9:30 – 18:00	**Анализ реляционных данных с Hive** • Базы данных и таблицы • Синтаксис HiveQL • Типы данных • Способы объединения таблиц • Часто используемые функции **Практическая работа:** Запуск запросов Hive из командной строки, файла и графического интерфейса **Управление данными в Hive** • Форматы хранения таблиц • Создание баз данных и таблиц • Загрузка данных в таблицы Hive • Изменение таблиц и баз данных • Создание внешних таблиц • Создание View • Сохранение результатов запросов • Контроль доступа к данным **Практическая работа:** Создание таблиц и их наполнение с помощью Hive **Обработка текстовых данных в Hive** • Обзор текстовой обработки • Важные функции текстовой обработки • Регулярные выражения • Анализ тональности текста и использование n-grams **Оптимизация запросов Hive** • Обзор методов оптимизации • Контроль исполнения запроса • Partitioning и bucketing • Индексы **Расширения Hive** • Сериализация и десериализация • Трансформации данных с использованием других языков программирования • Пользовательские функции • Параметризация запросов **Практическая работа:** Обработка данных в Hive с помощью Python и пользовательских функций **Итоги курса**	Лекции, Практика