

Программа курса

«GPDE: GreenPlum для инженеров данных и аналитиков данных»

О курсе: 5 дней практического обучения аналитиков, дата-инженеров и администраторов баз данных: анализ Big Data, организация ETL/ELT-процессов, построение корпоративных озер и хранилищ данных с масштабируемой отказоустойчивой open-source СУБД Greenplum Database.

Аудитория:

- инженер данных (Data Engineer);
- аналитик больших данных (Data Analyst);
- администратор больших данных (Big Data Administrator);
- администратор баз данных (Database Administrator).

Практические курсы по Greenplum предназначены для дата-инженеров, администраторов, системных архитекторов, аналитиков и разработчиков, которые хотят получить практические навыки по эффективной организации ETL/ELT-процессов, а также построению корпоративных озер и хранилищ данных на базе MPP-СУБД Greenplum.

Уровень подготовки:

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

Продолжительность курса: 24 академических часа, 5 дней по 8 ак. часов дистанционно

Содержание программы

1. Основы Greenplum: Greenplum как реляционная MPP СУБД

- Введение в Greenplum
- Архитектура Greenplum
- Партиционирование таблиц, индексы
- Транзакции и блокировки
- Оптимизация и планы запросов
- Лабораторная работа

2. Внешние данные и код: Greenplum как агрегатор данных и исполнитель кода

- Механизм внешних таблиц, протоколы, PXF
- Загрузка данных в Greenplum
- Механизм UDF
- Работа с JSON, XML
- Экосистема и дополнительные библиотеки
- Лабораторная работа

3. Greenplum в современных хранилищах данных

- Архитектура озера данных
- ETL, ELT и подготовка данных
- GPSS и интеграция Greenplum с Kafka и Nifi
- Обработка данных из Greenplum с помощью Spark
- Построение многоуровневого хранилища на Greenplum
- Лабораторная работа