Программа курса
«SPARK: Анализ данных с помощью современного Apache Spark»

О курсе: 6-дневный курс обучения по использованию распределенной платформы Apache Spark для работы с большими массивами данных, в том числе — неструктурированных и потоковой обработки информации.
Вы пройдете путь от основ архитектуры Spark до работы с продвинутыми компонентами, такими как GraphX, ML, Structured Streaming и Delta Lake. Программа охватывает как классические подходы (RDD, DataFrames, Spark SQL), так и актуальные тренды: интеграцию с Kubernetes, pandas API в Spark и управление данными через Delta Lake.

Аудитория:

- Data Engineers и аналитики, работающие с большими данными.
- Разработчики, желающие создавать масштабируемые ETL-процессы и ML-модели.
- Архитекторы, планирующие внедрение Spark в облачные среды (Kubernetes)

Уровень подготовки:

- Знание базовых команд Linux (опыт работы с командной строкой, файловой системой, POSIX)
- Начальный опыт программирования (Python)
- Начальный опыт в экосистеме Hadoop

Продолжительность курса: 32 академических часа, 6 дней

Содержание программы

1. **Обзор Apache Spark**
   - Архитектура Spark. Обзор компонентов Spark и их назначения
2. **Основные абстракции Apache Spark**
   - Трансформации и действия, Lazy Evaluation
3. **Знакомство с Dataframes**
   - Structured API и основная абстракция Spark – Dataframe
4. **Знакомство со Spark RDD**
   - Low Level API, использование Resilient Distributed Dataset
5. **Apache Spark SQL**
   - Получение данных из SQL-источников и обработка данных с помощью Spark SQL
   - Отправка данных в SQL СУБД и работа с Hive QL
   - Spark SQL и Hadoop
6. **Работа с источниками данных**
   - Ввод и вывод в Apache Spark
   - Работа с файлами и базами данных
7. **Производительность и параллелизм в Apache Spark**
   - Планы выполнения запроса: логические и физические
8. **Конфигурирование Apache Spark**

- Принципы конфигурирования и основные настройки

9. Spark Streaming и Structured Streaming

- Виды потоковой обработки в Apache Spark
- Особенности исполнения streaming кода
- Checkpoint в Spark Streaming

10. GraphX и ML

- Место и особенности графовых моделей в программировании
- Задачи машинного обучения и проблематика больших данных
- Основные возможности Spark ML

11. Обработка слабоструктурированных данных

- Работа с JSON и XML файлами, особенности и возможности

12. Современный Spark

- pandas API в spark
- Spark Connect: долгоживущие сессии
- Spark on Kubernetes (будущее в настоящем)
- Delta Lake - технологическая основа LakeHouse