

## Программа курса

# «PRUS: Практическое применение Big Data аналитики для решения бизнес-задач»

**О курсе:** Решите реальный бизнес-кейс на аналитической платформе, используя современные инструменты стека Big Data. Принимайте управленческие решения на основе ценных инсайтов, построив собственный аналитический ETL-конвейер для работы с корпоративными хранилищами и озерами данных.

**Аудитория:**

**ИТ-архитектор**

Технический директор (СТО)

Руководитель IT-проектов

Аналитик данных

**Уровень подготовки:** опыт написания SQL запросов, работа с командной строкой linux

**Продолжительность курса:** 32 академических часа, 8 дней по 4 ак. часов дистанционно

## Содержание программы

### 1. Постановка бизнес-задач и знакомство с BigData архитектурой

- Определяем бизнес-потребность. Формируем ключевые требования к решению, которое позволит удовлетворить бизнес-потребность и принести ценность
- Теория по архитектуре аналитического приложения BigData
  - **Практическое задание** — построение архитектуры анализа данных в зависимости от бизнес-кейсов.

### 2. Инструменты для анализа больших данных

- Глубокое погружение в экосистему инструментов для анализа больших данных: Spark, Hive, Dask
- Сравнительный анализ их возможностей и областей применения
  - **Практическое задание** — практическая работа по анализу данных с использованием Apache Spark/Hive.

### 3. DWH хранилища данных, слои накопления данных (stg, ods, dds, dim, mart)

- Предназначение концептуальных слоев накопления данных
- Методы очистки и трансформации данных
- Разбор подходов к пакетной (batch) и потоковой (stream) загрузке данных
  - **Практическое задание** — реализация Batch/Streaming загрузки данных различными способами.

### 4. Способы организации хранения сырых данных (stg слой)

- Теоретическое описание слоя хранения данных и его принципов
- Обзор Big Data репозиторий, HDFS и форматов хранения данных (Parquet, Avro, ORC)
- Концепция EDW, ее преимущества и недостатки
- NoSQL-подход к хранению данных.
  - **Практическое задание** — демонстрация работы с файлами в HDFS и S3.

### 5. Инструменты для управления процессами обработки и перемещения данных.

- Введение в Apache Airflow
- Архитектура и основные компоненты
- Принципы построения DAG-ов
- Углубленное изучение Airflow.
- Создание сложных пайплайнов перемещения и обработки данных

- **Практическое задание** — настройка и демонстрация работы Airflow на примере простых задач. Реализации пайплайна обработки и перемещения данных с помощью Airflow.

## 6. BI инструменты для аналитики данных

- OLAP/OLTP базы данных
- BI инструменты (SuperSet, PowerBI, и др.) Архитектура и область применения
- Лучшие практики визуализации данных
  - **Практическое задание** — построение дашбордов в BI-инструменте на основе данных, подготовленных в предыдущих модулях.

## 7. Введение в ML

- Обзор кейсов использования ML в бизнесе
- Разбор ML модели кейса
  - **Практическое задание** — встраивание модели ML в pipeline обработки данных.