Программа курса
«AIRF: Data Pipeline на Apache Airflow»

О курсе: Наш углубленный 5 дневный курс Apache Airflow поможет освоить самый популярный инструмент для оркестрации, который используют в ведущих IT-компаниях. Вы научитесь превращать хаос в данных в управляемые, автоматизированные data pipelines, став настоящим архитектором данных.
Забудьте о медленных процессах. С Apache Airflow вы научитесь создавать и отслеживать сложные рабочие процессы как код. Этот практический тренинг проведет вас от создания первого DAG до продвинутых техник, включая настройку отказоустойчивых конвейеров и интеграцию с Big Data. Наш Apache Airflow курс — это ваша прямая инвестиция в карьерный рост.

Аудитория:
- Инженер данных
- Аналитик данных
- Разработчик DHW

Уровень подготовки:
- Знание базовых команд Linux.
- Базовый опыт программирования на Python/bash.
- Опыт программирования на SQL и понимание принципов работы реляционных баз данных.

Продолжительность курса: 24 академических часа, 5 дней

Содержание программы

1. Введение в Apache Airflow: концепции, экосистема и первый запуск

Что такое Airflow и зачем он нужен
- Роль Airflow в построении современных дата-пайплайнов
- Основные принципы работы: оркестрация, расписания, мониторинг, устойчивость
- Сравнение с конкурентами
- Когда Airflow — оправданный выбор, а когда — нет

Архитектура и ключевые элементы
- DAG: что это такое, как живет и исполняется
- Operators: виды и применимость
- Понимание жизненного цикла задач

Практическая часть
- Подготовка рабочей среды на виртуальных машинах
- Развёртывание Airflow
- Создание первого DAG и выполнение простых задач с Bash и Python

2. Основы работы с Airflow: параметры, исполнители и управление DAG'ами

Архитектура и компоненты
- Scheduler, Workers, Metadata DB, Web UI — как всё взаимодействует
- Executors: Sequential, Local, Celery и их сценарии использования
- Роль API и возможностей интеграции

Работа с DAG'ами

- Параметры расписаний: schedule_interval, start_date
- Ограничения и параллелизм: concurrency, max_active_runs
- Введение в TaskFlow API

Практическая часть

Создание второго DAG с параметрами
Знакомство с Web UI и базовыми возможностями
Запуск и управление DAG'ами вручную и по расписанию

3. Управление логикой выполнения: композиция, переменные и обмен данными

Композиция задач и зависимости

- Подходы к объявлению связей между задачами
- Композиция через операторы и bit-shift синтаксис
- Организация читаемых пайплайнов

Работа с конфигурацией и состоянием

- Переменные (Variables): хранение параметров и конфигураций
- XCom: обмен данными между задачами, типичные кейсы и антипаттерны
- Практика Backfill & Catchup — как запускать прошлые периоды безопасно

Практическая часть

- Лабораторные упражнения на композицию и использование переменных
- Построение DAG'ов с передачей данных через XCom

4. Интеграции и ожидание внешних событий: Sensors, Connections и Trigger Rules

Интеграции через Connections

- Настройка подключений и безопасное хранение секретов
- Типовые сценарии интеграций

Sensors и работа с событиями

- Принцип работы сенсоров
- ExternalTaskSensor: синхронизация разных DAG'ов
- Trigger Rules: контроль выполнения в зависимости от результатов предыдущих задач
- Условное ветвление и Branch-логика

Практическая часть

- Практика по Sensors и Trigger Rules
- Взаимодействие с PostgreSQL

5. Шаблоны, макросы и расширенные интеграции: создание гибких пайплайнов

Гибкость и параметризация

- Templates & Macros — динамическое формирование задач и путей
- Hooks: использование готовых коннекторов для внешних систем

Работа с хранилищами и БД

- Построение пайплайнов с загрузкой и проверкой данных
- Использование шаблонов в SQL и Python-тасках

Практическая часть

- Задачи, завязанные на S3-файлы
- Использование PostgreSQL, шаблонов и макросов в реальных сценариях

6. Синтез знаний: создание продвинутых end-to-end пайплайнов

Продвинутые конструкции

- TaskGroup — структурирование сложных DAG’ов
- Dynamic Tasks — генерация задач на лету
- Datasets — управление зависимостями на уровне данных
- Лучшие практики и анти-паттерны в продакшене
- Мониторинг, метрики и операционная поддержка
- Способы изоляции окружений в AirFlow

Итоговая практическая работа “Создание связанной системы из нескольких DAG’ов”

- генерация данных и выгрузка их в S3
- чтение, обработка и передача данных через XCom
- запись результата в базу
- использование всех изученных инструментов в одном пайплайне

Версия программы 04.03.2026