Программа курса «YARF: AIRFLOW с использованием Yandex Managed Service for Apache Airflow™»

О курсе: 3-хдневный курс обучения по Airflow позволит вам получить и систематизировать знания по использованию этого фреймворка для разработки, планирования и мониторинга рабочих процессов с большими данными. Курс содержит расширенные сведения по установке распределенного кластера Apache Airflow, администрированию и интеграции этой платформы с другими технологиями Big Data в соответствии с лучшими практиками (best practices).

Аудитория: Наши курсы обучения по Airflow ориентированы на системных администраторов, инженеров данных (Data Engineer), архитекторов, DevOps-инженеров, разработчиков Hadoop и прочих Big Data систем.

Уровень подготовки:

- Знание базовых команд Linux (опыт работы с командной строкой, файловой системой, POSIX, текстовыми редакторами vi, nano)
- Базовый опыт программирования Python/bash
- Начальный опыт в экосистеме Apache Hadoop
- Средний опыт программирования SQL

Продолжительность курса: 24 академических часа, 3 дня по 8 ак. часов дистанционно

Содержание программы

1. Введение в Airflow

- Что такое Airflow?
- Почему Airflow?
- История создания
- Аналоги и конкуренты
- Airflow vs Oozie
- “Киты” Airflow
- Настройка образа в YandexCloud

2. Базовый Airflow

- Верхнеуровневая архитектура
- Компоненты: подробнее
- Executors
- LocalExecutor
- Схема учебного стенда
- DAG
- DAG: параметры
- Operators
- Operators: виды
- WEB UI: обзор
- Пайплайн по созданию DAG
- Dag: context
- Operator: основные параметры
- Composition
- EmptyOperator
- BashOperator

- Написание первого DAG
- TaskFlowApi
- PythonOperator

Практика № 1. Создание первого DAG, использование Python и Bash операторов, использование WebUI

- Запуск DAG с ручной конфигурацией

Практика № 2. Написание DAG с ручной передачей параметров

- Переменные и их использование (Variables)

Практика № 3. Применение Variables, default_args

Практика № 4. Применение Variables расширенное

- Connections
- Sensors

Практика № 5. Применение fileSensor

- ExternalTaskSensor

3. Расширенный Airflow

- Trigger Rules

Практика № 6. Использование fileSensor + triggerRules

- Backfill & catchup
- Templates
- Macros
- Yandex Managed Service for PostgreSQL
- Демонстрация ETL процесса на временном DataProc(Spark) кластере в Yandex Cloud

Практика № 7. Использование PostgresOperator, оркестрация ETL процесса

- Hooks

Практика № 8. Применение Hooks

- TaskGroup
- XCOM
- Dynamic Tasks
- XCOM vs Variable

Практика № 9. Финальная практика, включающая в себя все вышеизученное