

**Apache Airflow for Machine Learning Operations**

**Course Number:** PYTH-226  
**Duration:** 3 days

**Overview**

This Apache Airflow for Machine Learning Operations training course teaches machine learning (ML) engineers how to build and validate training models, upload models to a model registry, and deploy models in a reproducible manner.

Attendees learn machine learning operations and the complexities of creating a reproducible CI/CD pipeline for ML models. Next, students explore options to reduce this gap with Apache Airflow for batch training scenarios (which are the majority). In addition, attendees learn the foundations of Airflow and how it creates reproducible and trustworthy pipelines via DAGs (Directed Acyclic Graphs).

This course focuses on real-world applications of ML using both traditional machine learning algorithms and deep learning algorithms, such as sentiment prediction in a stream of tweets.

Throughout the course, students tackle diverse machine learning problems by creating reproducible pipelines with Airflow.

**Prerequisites**

Students must have basic Python knowledge or object-oriented programming experience. Knowledge of machine learning is helpful but not required.

**Materials**

All Apache Airflow for Machine Learning training attendees receive comprehensive courseware.

**Software Needed on Each Student PC**

This  course is taught using:

* Python 3.5 or later
* Apache Airflow 2.1 or later
* scikit-learn 1.1 or later
* PyTorch 1.8 or later

On request, we can provide either a remote VM environment for the class or directions for configuring this environment on your local PCs.

**Objectives**

* Migrate machine learning training workflows to scalable pipelines in Apache Airflow
* Start with a raw dataset and a model architecture and take the project from beginning to end, culminating in deploying it in the cloud
* Enforce reusability and modularization of pipelines for easy collaboration

**Outline**

* Introduction
* The Scalable Problem of Machine Learning Pipelines
  + What problems arise when trying to create a machine learning model?
  + The components of a machine learning platform
  + Introducing Apache Airflow
  + Airflow architecture
  + How do we represent a machine learning pipeline?
  + Our first DAG
  + Tasks, TaskFlows, and Operators
  + First Pipeline
  + Cresting the datasets for training
* Creating our Machine Learning Pipeline
  + Using custom operators
  + Creating a Train Operator
  + Creating TaskGroups vs subDAGs
  + Sharing data with xCOMs
  + Branching and Triggers
  + Sensors and SmartSensors
  + Adding a sensor to validate enough new data
  + Adding training, validation, and delivery steps to our pipeline
* Mastering Scheduling
  + execution\_date, start\_date, and schedule\_interval
  + Handling non-default schedule\_intervals
  + Playing with time
  + Using Sensors with a correct schedule\_interval
* Enabling Concurrency and Scalability
  + Abandoning SQLite to PostgreSQL
  + Executors: Debug, Local, Celery
  + Concurrency and parallelism
  + Concurrency with Celery
* Hackathon: Sentiment Prediction from Twitter
* Conclusion