Module 2: Data Pipelines

Overview

In Module 2 we address the first task of data pipelines, namely data ingestion, and describe the concept of task orchestration. Finally, we will work on performing batch and streaming ingestion to store data on an S3 bucket using Apache Airflow and Python APIs.

Objectives

Upon completion of this module, students will be able to:

1. Define data ingestion
2. Explain the batch and streaming ingestion
3. Explain the data orchestration
4. Discuss the orchestration mechanism of Apache Airflow
5. Synthesize a solution on performing batch and streaming ingestion to store data on a data lake

Readings (1 hour)

- **A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the Data Cloud Project (https://canvas.vt.edu/courses/176740/files/28995272?wrap=1)**

Watch (1 hour)

- **Lecture 2 Video (https://canvas.vt.edu/media_objects_iframe/m-6dtoP7ZTCW2vVsA2Mqn5Q6uKjiT8jDqC?type=video?type=video)**

- **Lecture 2 Slides (https://canvas.vt.edu/courses/176740/files/28995310?wrap=1)**

Assignment (Lab and Homework) (2 hours)

- **Assignment 1 (https://canvas.vt.edu/courses/176740/assignments/1816310)**

(https://canvas.vt.edu/courses/176740/files/28995187/download?wrap=1) Recitation (1 hour)

Slides (https://canvas.vt.edu/courses/176740/files/29507279?wrap=1)