Module 5: Data Transformations and Data Provenance

Overview

In this module, we build on our understanding of data transformations while focusing further on data storage and data provenance. We begin by explaining where data transformations are happening. The characteristics of different storage types are described to accommodate the needs of transformed data. Next, we discuss the concepts of data provenance and lineage, identifying their importance in the cloud computing era.

Objectives

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

1. Identify the data pipeline tasks that allow for data transformation
2. Describe the different types of data storage used to accommodate the needs of newly transformed data
3. Define the concepts of data provenance and lineage
4. Discuss the role of data provenance and lineage in the presence of transformed data
5. Identify the importance of provenance and lineage in cloud services and pipelines

Readings (1 hour)

- Establishing Data Provenance for Responsible Artificial Intelligence Systems, K. Werder et. al., ACM Transactions on Management Information Systems, Vol. 13, No. 2, March 2022 (https://canvas.vt.edu/courses/176740/files/28995282?wrap=1)

Watch (30 minutes)

- Module 5 Video (https://canvas.vt.edu/media_objects_iframe/m-6dD3uB8cHX3JV2KC3kcg9XS3A3tv1B5Y?type=video?type=video)

- Module 5 Slides (https://canvas.vt.edu/courses/176740/files/28995308?wrap=1)

Project Proposal, Milestone & Final Project

Work on your Project Milestone

Data Engineering Project (https://canvas.vt.edu/courses/176740/pages/data-engineering-project)

Project Proposal (https://canvas.vt.edu/courses/176740/assignments/1816317)

Project Milestone (https://canvas.vt.edu/courses/176740/assignments/1816316)

Final Project (https://canvas.vt.edu/courses/176740/assignments/1816314)

Recitation (1 hour)

The recording of the optional synchronous ZOOM session for this lecture will be linked here.