Course Outline

Tabular Data Processing with Dask Dataframe

Course Summary

Description

This class module focuses on Dask Dataframe, a simple model for working with tabular data that may be too large to fit in memory or to process on a single machine.

The Dask Dataframe class is recommended for engineers or data scientists who typically work with tabular (row/column) data and related tools, like SQL databases.

Objectives

At the end of this course, students will understand:

- How Dask Dataframe extends Pandas to larger datasets
- How to select, filter, transform, and join data
- Understand performance with partitioning and indexes

Topics

- Introduction
- Core Dataframe API and Operations
- Data Access
- Best Practices

Audience

The Dask Dataframe class is recommended for engineers or data scientists who typically work with tabular (row/column) data and related tools, like SQL databases.

Prerequisites

Students should have experience in Python and Dask programming, both at a basic level.

Duration

One day

Course Outline

Tabular Data Processing with Dask Dataframe

Course Outline

I. Introduction
- A. Python and Pandas for tabular data
- B. Limitations of Pandas
- C. Dask Dataframe model
- D. Key similarities/differences compared to Pandas

II. Core Dataframe API and Operations
- A. Reading data
- B. Selecting records and columns
- C. Using indexing to select records
- D. Filtering datasets
- E. Combining datasets (joins, unions)
- F. Custom functions
- G. Aggregations and sorting (groupby, sort)
- H. Custom aggregation
- I. Window (rolling) operations

III. Data Access
- A. Read CSV and Parquet data and best practices for performant reading
- B. Read JSON and text data with Dask bag
- C. Read custom formats with Dask delayed
- D. Writing data efficiently for future access

IV. Best Practices
- A. General query improvement patterns
- B. Minimizing expensive data transfer
- C. Launching work and preserving results with “persist”
- D. Partition sizing