NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NTU Academy for Professional and Continuing Education

(SCTP) Advanced Professional Certificate

Data Science and AI

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2.8 Out of Core Computation

Module Overview

2.1 Introduction to Big Data and Data Engineering

2.2 Data Architecture

2.3 Data Encoding and Data Flow

2.4 Data Extraction and Web Scraping

2.5 Data Warehouse

2.6 Data Pipelines and Orchestration

2.7 Data Orchestration and Testing

2.8 Out of Core/Memory Processing

2.9 Big Data Ecosystem and Batch Processing

2.10 Event Streaming and Stream Processing

Lesson Objectives

- Use alternative dataframe libraries to Pandas for more performant and scalable data processing.
- Use DuckDB to query data that is too large to fit in memory.

Limitations to Pandas

Alternatives to Pandas

- Polars
  - Built in Rust (very fast)
  - Uses Apache Arrow backend (efficient memory format)
  - Optimized for performance (multi-threading, query optimization)
  - Comparison with Pandas [link]
- Dask
- Spark

🐼 Pandas vs 🐍 Dask vs ⚡ Polars vs 🔥 Spark

Feature	Pandas	Polars	Dask	Spark
Parallelism	single-core	multi-core(local)	multi-core(local) or distributed cluster	Distributed (cluster-first, but can run locally)
Cluster support	No	No	Yes (optional)	Yes (built for it)
Dataset size	Must fit in memory	Efficient in-memory, usually must fit in memory	Larger than memory, out-of-core, distributed	Massive (petabyte scale, big data pipelines)
API familiarity	Standard	Different API	Pandas-like	SQL/Dataframe and RDD APIs

Lesson Plan

1. Follow the Jupyter notebook