Syllabus IS3107 (AY22/23 Sem 1): Data Engineering

This IS3107 module covers the core concepts of data engineering, which span the data engineering lifecycle and principles, data architecture, ETL, data characteristics and the corresponding moving, storage, and processing strategies.

Instructor:

Frank Xing, fxing@comp.nus.edu.sg

Modular credits:

4MCs

Prerequisite:

- BT2102 “Data Management and Visualisation” or CS2102 “Database Systems”
- Some knowledge of database and Python programming.

ILO (Intended Learning Objectives):

- Be able to apply concepts of data engineering to analyze business needs.
- Understand challenges and strategies for corporate data storage and processing.

Assessment:

Participation (10%)

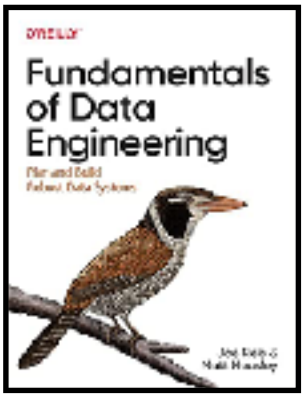
Assignments (4*10% = 40%)

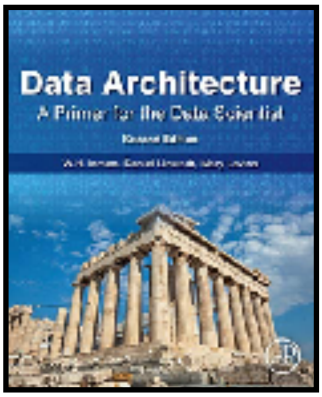
Mini project (30%)

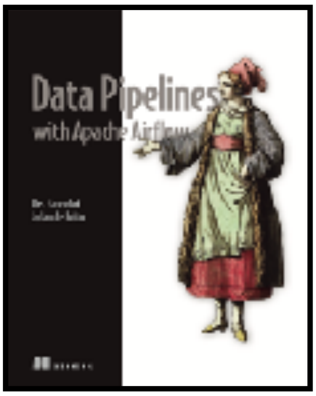
Final Quiz (20%)

Reference materials:

1> FDE: Fundamentals of Data Engineering (2022)
ISBN 978-1-09-810830-4
2> DI: Designing Data-Intensive Applications (2017)
ISBN 978-1-44-937332-0
3> DA: Data Architecture: A Primer for the Data Scientist (2nd edition, 2019)
ISBN 978-0-12-816916-2
4> DP: Data Pipelines with Apache Airflow (2021)
ISBN 978-1-61-729690-1


O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis & Matt Housley





Tentative Lesson Plan:

Week and Date	Lecture Topic	Tutorial & Assignment	Reference Chapters	Dues
Week 1	Introduction to Data Engineering	******	FDE: Ch 1 DA: Ch 4-2	******
Week 2	Data Formats and Encoding	******	FDE: Ch 5,7 DI: Ch 4	******
Week 3	Data Storage and Query	Storing Non-relational Data	FDE: Ch 6,8	******
Week 4	Data Replication	******	DI:Ch 5	Assignment 1
Week 5	Data Partitioning	Regular Data Updates Using Airflow	DI:Ch 6	******
Week 6	Servicing Data from Cloud	******	******	Assignment 2
Recess Week	******	******	******	******
Week 7	Data Architecture	Mini-project Consultation (1hr)	FDE: Ch 3 DA: 1,6,8	******
Week 8	Distributed Data Processing	******	DI:Ch 8 DA:4-3	******
Week 9	MapReducible Algorithms	MapReduce exercises (Python, Hadoop, PySpark)	DI:Ch 10	******
Week 10	Stream Data Processing	******	DI:Ch 11	Assignment 3
Week 11	Counting and Clustering on Streams	Lossy Counting on Twitter Stream	******	******
Week 12	Data Pipeline and Orchestration	******	DP: Ch3	Assignment 4
Week 13	Data Engineering Lifecycle and Final Quiz	******	FDE: Ch2,4,11	Mini-project