Syllabus IS3107 (AY23/24 Sem 1): Data Engineering

This IS3107 module covers the core concepts of data engineering, which span the whole data engineering lifecycle (from ingestion to serving) and intended goals (low latency, high throughput, high reliability, etc). Students will learn about topics including data pipeline and ETL, data formats, data architecture, and data moving, storage, and processing strategies to specific business requirements.

Instructor and tutor:

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

Yuchen WANG, yuchen.wang@u.nus.edu

Course credits and logistics:

4 unit credits

Weekly in LR19, tutorials start from week 3 in LR19 attached seminar room.

L1: Fri 12:00-14:00; T1: Fri 14:00-15:00; T2: Fri 17:00-18:00

Office hour: Tue 16:00-17:00

Prerequisites:

- 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 and fulfil business needs.
- Understand challenges and strategies for corporate data storage and processing.

Assessment:

Participation in discussions and course activities (10%)

Assignments (4*10% = 40%)

Quizzes (10% + 15% = 25%)

Course project (25%)

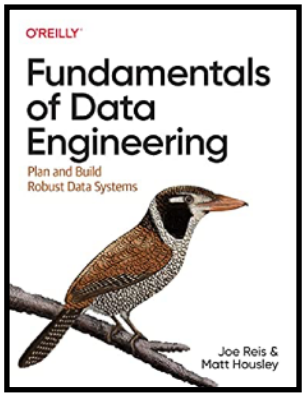
Course reading materials:

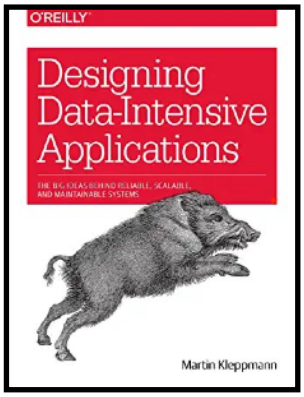
1> FDE: Fundamentals of Data Engineering (2022) ISBN 978-1-09-810830-4

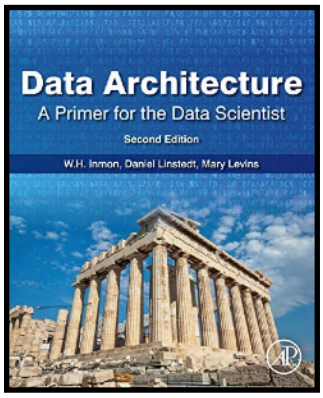
2> DI: Designing Data-Intensive Applications (2017) ISBN 978-1-44-937332-0

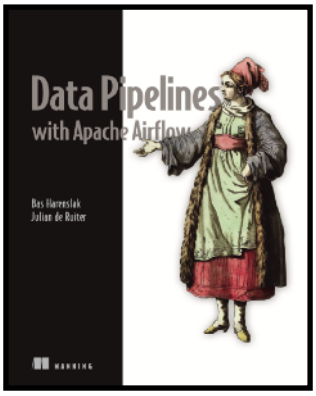
3> DA: Data Architecture: A Primer (2nd edition, 2019) ISBN 978-0-12-816916-2

4> DP: Data Pipelines with Apache Airflow (2021) ISBN 978-1-61-729690-1









Tentative Lesson Plan:

Week and Date	Lecture Topic	Tutorial Topic	Reading	Assessments
Week 1	Introduction to Data Engineering	******	FDE Chapter1	******
Week 2	Data Pipeline and Orchestration	******	DP page 1-85	******
Week 3	Data Storage (Physical and Cloud)	Comparing ETL & ELT performances	FDE Chapter 6	Asnmt. 1
Week 4	Data Organization	Planning storage efficiency	DI page 27-63	******
Week 5	Data Querying	SQL or NoSQL	FDE Chapter 8	Asnmt. 2
Week 6	Data Replication and Partitioning	Query practice	DI Chapter 5,6	******
Recess Week	******	******		******
Week 7	Data Architecture	Replication and partition exercise	DA Chapter 8	Asnmt. 3
Week 8	MapReduce and Hadoop	Designing data warehouse	DI Chapter 10	******
Week 9	MapReducible Algorithms	Map reduce with word count	DI Chapter 10	Asnmt. 4
Week 10	Stream Data Processing	Map reduce with word count ctn'd	DI Chapter 11	******
Week 11	Streaming Algorithms	Sentiment analysis on stream data	DI Chapter 11	******
Week 12	Wellbeing Day	******	******	******
Week 13	Modern Topics in Data Engineering	project consultation	******	mini-project