# 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)

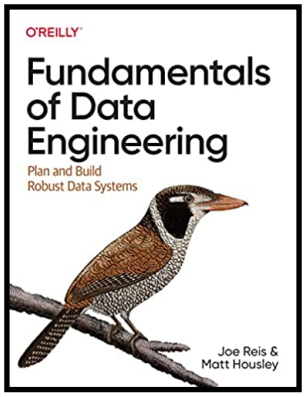
ISBN 978-1-44-937332-0

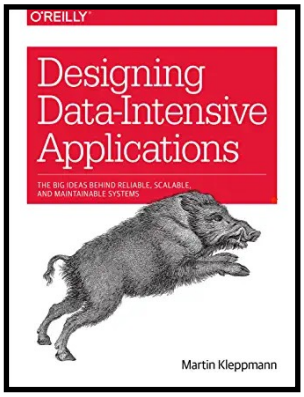
3> DA: Data Architecture: A Primer for the Data Scientist (2nd edition, 2019)

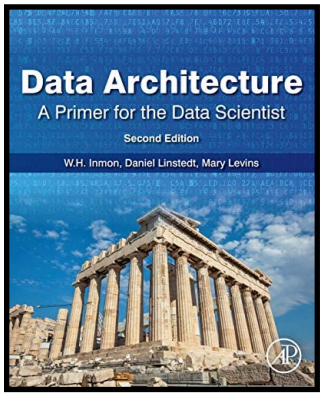
ISBN 978-0-12-816916-2

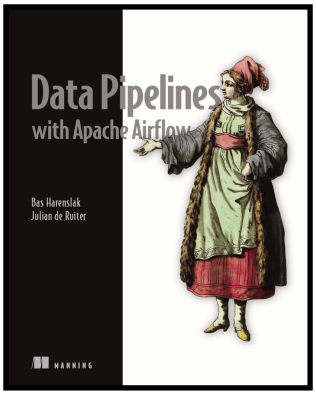
4> DP: Data Pipelines with Apache Airflow (2021)

ISBN 978-1-61-729690-1




## Tentative Lesson Plan:

| Week and Date | Lecture Topic | Tutorial Topic | Reference Chapters | Dues |
|---|---|---|---|---|
| Week 1 | Introduction to Data Engineering | ****** | FDE: Ch 1<br>DA: Ch 4-2 | ****** |
| Week 2 | Data Formats and Encoding | ****** | FDE: Ch 5,7<br>DI: Ch 4 | ****** |
| Week 3 | Data Storage | Combining data from different formats | FDE: Ch 6 | ****** |
| Week 4 | Data Querying | Storing Data with NoSQL | FDE: Ch 8 | ****** |
| Week 5 | Data Replication and Partitioning | Querying Data in NoSQL | DI:Ch 5,6 | Assignment 1 |
| Week 6 | Servicing Data from Cloud | Data Partitioning with Cassandra | ****** | Assignment 2 |
| Recess Week | ****** | ****** | ****** | ****** |
| Week 7 | Data Architecture | Cassandra DB on cloud | FDE: Ch 3<br>DA: 1,6,8 | Assignment 3 |
| Week 8 | Distributed Data Processing | Business case study (DA) | DI:Ch 8<br>DA:4-3 | ****** |
| Week 9 | MapReducible Algorithms | Mini-project Consultation | DI:Ch 10 | ****** |
| Week 10 | Data Pipeline and Orchestration | MapReduce exercises (Python, Hadoop, PySpark) | DP: Ch3 | ****** |
| Week 11 | Stream Data Processing | Update database with Airflow | DI:Ch 11 | Assignment 4 |
| Week 12 | Counting and Clustering on Streams | Word Count with Kafka | ****** | ****** |
| Week 13 | Data Engineering Lifecycle and Final Quiz | ****** | FDE: Ch2,4,11 | Mini-project |