Syllabus IS3107 (AY21/22 Sem 2):

Data Engineering

Lecturer:

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

Teaching Assistants:

Gao Yuting gao.yuting@u.nus.edu.sg

Joel Quek joelq@comp.nus.edu.sg

Time and mode of instruction: online, no recording

L1: Fri 1830-2030 (Frank)

T1: Fri 1000-1100 (Frank) T2: Fri 1200-1300 (Joel)

T3: Fri 1200-1300 (Yuting) T4: Fri 1300-1400 (Yuting)

T5: Fri 1400-1500 (Joel) T6: Fri 1600-1700 (Yuting)

T7: Fri 2030-2130 (Joel) Tutorial cap size: 30pax

Modular credits: 4MCs

Enrolment size: 195. To ensure teaching quality, enrolment approval requests (e.g. co-taking w/ ATAP etc.) after 2021 Dec 7th 23:59PM, unfortunately, will NOT be entertained. However, students who bid and secured a slot due to logistic confusion can proceed this time.

This IS3107 module covers the core concepts of data engineering, which include ETL, data pipeline design, data moving and processing, big data solutions, data on cloud, and the business value of data. A topical emphasis is on financial data and applications.

Prerequisite:

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

ILO (Intended Learning Objectives):

- Be familiar with concepts and skills required for a data engineer.
- Be able to design data pipelines according to task-specific needs.
- Understand the value of data engineering in business practices.

Structure:

This module has a 10 hrs per week workload:

2-hr lecture + 1-hr tutorial + 3-hr preparation + 4-hr project.

This module has 12 lectures and 9 tutorials in total.

No class for Week 13 as it falls on a Singapore gazetted public holiday.

Assessment:

Two quizzes (15%+20% = 35%)

One course project (55%) [39 teams of 5pax]

- 35% on report + 10% on presentation + 10% on peer evaluation.

Involvement - attendance or active participation (10%)

- Marked from tutorials

No final exam.

Other details:

Week 1 Introduction to Data Engineering

In depth reading: Sysco's big data lake

Week 2 Data Formats and Processing

In depth reading: XML and JSON Are Like Cardboard

Week 3 Basics of Data Pipeline

In depth reading: The RADStack Architecture

Tutorial 1 Data Formats and Processing Exercise

Week 4 Data Pipeline Design

Tutorial 2 Your Minimum ETL Example in Python

Week 5 Cloud Computing and Cloud DB

In depth reading: Good cloud services

Tutorial 3 Getting Familiar with Apache Airflow

Week 6 Distributed Data Processing

In depth reading: Text Processing with MapReduce

Tutorial 4 Your own DAG file

****

Recess week + release of course project.

****

Week 7 NoSQL (Quiz 1 on Mar 4)
Tutorial 5 ETL in Airflow

Week 8 Stream Data Processing
In depth reading: Twitter Heron
Tutorial 6 Accessing MongoDB in Airflow

Week 9 Scalable Machine Learning
In depth reading: Communication-Efficient Learning
Tutorial 7 Advanced features in Airflow

Week 10 Data Pricing and Valuation
In depth reading: Datasheets for Datasets
Tutorial 8 Q&A Session for Course Project

Week 11 Data in the Financial Industry
In depth reading: Discovering Business Models of Data Marketplaces
Tutorial 9 Exercise for data pricing and valuation

Week 12 Financial Data Engineering Case Studies (Quiz 2 on Apr 8)

Week 13 No class but deadlines for-
Teaching feedback & final project: the Friday before reading week: Apr 15, 2022