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: 35pax

Modular credits: 4MCs

Enrolment size: 169 students, 34 project groups

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.

Assessment:

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

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

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

Involvement - attendance or active participation (10%)

- Marked from tutorials

No final exam.

Reference materials:

Data Architecture: A Primer for the Data Scientist (2nd edition, 2019)
ISBN 978-0-12-816916-2

Data Pipelines with Apache Airflow (2021)
ISBN 978-1-61-729690-1

Week 1 Introduction to Data Engineering
*5Vs of big data *role of data engineer *basic concepts in data engineering
In depth reading: Sysco's big data lake

Week 2 Data Formats and Processing
*CSV, JSON, XML *cmd & NLTK *performance & implementation
In depth reading: XML and JSON Are Like Cardboard

Week 3 Basics of Data Pipeline
*DAG, ETL, OLT(A)P *lambda/kappa architecture
In depth reading: The RADStack Architecture
Tutorial 1 Data Formats and Processing Exercise

Week 4 Data Pipeline Design
*SQL query optimization *workflow transition *state space search
In depth reading: Data-centric workflow optimization
Tutorial 2 Your Minimum ETL Example in Python

Week 5 Cloud Computing and Cloud DB
*Cloud servicing *cloud db architecture, Redshift and Snowflake *AWS EC2 and S3
In depth reading: Google cloud services
Tutorial 3 Getting Familiar with Apache Airflow

Week 6 Distributed Data Processing

*MapReduce *PageRank and K-means on MapReduce *Hadoop

In depth reading: Text Processing with MapReduce

Tutorial 4 Your own DAG file

****

Recess week + release of course project.

****

Week 7 NoSQL-part1 (Quiz 1 on Mar 4)

*Key-value store *chord

Tutorial 5 ETL in Airflow

Week 8 NoSQL-part2

*Wide-column *HBase

Tutorial 6 Counselling Session for Course Project

Week 9 [Guest Lecture by Edwin Law] Data Engineering Activities at Grab

*Management, governance, security and other enterprise-level concerns

Tutorial 7 Accessing MongoDB in Airflow

Week 10 Stream Data Processing

*DSMS *counting/clustering on streams *Storm and Spark

In depth reading: Twitter Heron

Tutorial 8 Advanced features in Airflow

Week 11 Data Pricing and Valuation

*Strategies *Shapley value *marketplace

In depth reading: Datasheets for Datasets

Tutorial 9 Exercise for data pricing and valuation

Week 12 Data Engineering in the Financial Industry (Quiz 2 on Apr 8)

*Financial data types *case studies

In depth reading: Discovering Business Models of Data Marketplaces

Week 13 No class as it falls on a Singapore gazetted public holiday, but deadlines for-
1) teaching feedback & 2) course project: Apr 15, 23:59 PM