

BAX421 *Data Management*

Units: 2

Quarter: Fall

Stream: Computing

Mode: Lecture, text/readings, exercises

Instructor:

Julia Fu yanqiufu@gmail.com

Office Hour: Tentatively Wed 5-7 online

Description: Acquiring, preparing and staging data for analytical use. Specific topics include data warehouse, data mart, data lake, data integration and pipeline management (ETL, ELT, and streaming), and data governance. After the class completion, the students will learn key data management concepts, know how to load, clean, and transform raw data for analytical use, and have insight to real-world data operation process.

Book:

The Data Warehouse ETL Toolkit Ralph Kimball ISBN: 978-0-7645-6757-5

Topics:

1. Students will learn core data warehouse and data integration concepts. Topics include: OLAP vs OLTP, a light touch on Star Schema, ETL strategy (incremental vs full replace), and data integration best practice guide. (3 units + 1.5 units exercise)
2. Students will get high-level overview of major analytical data platforms and ETL tools. And they will also get exposure to basic knowledge on how to perform a POC on new tools. (1 unit)
3. Students will learn how big data and modern analytical data platform have re-shaped DW landscape and data integration process. Star Schema to flat table structure; User behavior event schema, ETL to ELT and streaming, etc. (1 unit + 0.5 unit exercise)
4. Students will learn the key components of data governance program. Also understand database cataloging, data dictionary, data quality alert and monitoring, etc. (1 unit)
5. Last but not least, students will learn how various data related roles/functions working together: Data Architect, Data Product manager, BI engineer/analyst, Data Engineer, and various analytics teams. (1 unit)

Teaching Method

1. I'd like to provide solid basics on data warehouse core concepts, and ETL methods. This topic will have a written exam.
2. Then students will form into teams. Members of each team will perform different business roles, and each team will complete one end-to-end data integration project with requirement document, schema design, ETL, and scheduling. Ideally, the data prepared in this class can be used for analytical project or data visualization course.

3. If scheduling permitting, I'll set up one students meet data team session. Needs to pan out details.

Grading Criteria:

1. Assignments: 20%
2. Class Participation: 10%
3. Midterm Exam: 15% (Tentatively on Oct 27th)
4. Final Exam: 15% (Tentatively on Dec 1st)
5. Team Project: 40% (Team project due on mid-night Sunday Dec 3rd)

Exams:

The purpose of the exams is to assess your ability to apply concepts discussed in class. Two exams will be given. Two examinations will be administered in class on the dates indicated in the course schedule. The midterm will cover material from the first half of the course. The non-comprehensive final will cover material from the second half of the course. The exams will include a combination of objective questions, applied problems, and short answer/essays.

Projects:

The goals of the team project are (1) to integrate and apply what you have learned in the class; (2) to give you experience working in team situations; and (3) to enhance your communication and presentation skills. Each group consists of up to 4 students. The working program, a final written report, and an oral presentation in class will be the basis for the project grade.