BAX 421 – 001: Data Management

TERM: Fall 2024

LECTURES: Fridays: 5:00 p.m. – 7:00 p.m.

INSTRUCTOR: Mehul Rangwala mrangwala@ucdavis.edu

OFFICE HOURS: My hours for Q&A will be the same as for the BAX-441. I will take questions for both the classes.

TA DISCUSSION SECTIONS: We have two TAs for this course. Each will have an hour of discussion session via Zoom every week. Additional details will be shared on Canvas.

COURSE DESCRIPTION: Introduction to the extraction, assembly, storage and organization of data in IT systems. The course covers the concepts of data modeling, entity relationship models, and SQL to help businesses convert data into insights needed to drive business strategies. Use of MySQL and Microsoft SQL Server for writing SQL queries. Students will design and deploy a database solution using MySQL. Students will also learn how to connect databases to the RStudio computing environment and to the data visualization tool Tableau and visualize query results within Tableau.

PREREQUISITES: None.

REFERENCE TEXTBOOKS:

1. SQL: The Complete Reference, 3rd Edition
ISBN-13: 978-0071592550
ISBN-10: 0071592555

2. Database Systems, 14th edition
by Carlos Coronel and Steven Morris
ISBN-13: 9780357673102

3. Database Systems: Introduction to Databases and Data Warehouses, 2nd edition
by Nenad Jukić, Susan Vrbsky, Svetlozar Nestorov, Abhishek Sharma
ISBN: 978-1-943153-67-1

4. Fundamentals of Relational Database Management Systems by S. Sumathi and S. Esakkirajan
ISBN-13: 978-3-642-08012-8
You can get this downloadable ebook from our library. No need to purchase this.
https://link.springer.com/book/10.1007%2F978-3-540-48399-1

5. Data Modeling and Database Design
by Dr. Narayan S. Umanath, Richard Scamell
ebook ISBN-10: 1305473035 | ISBN-13: 9781305473034

NOTES AND HANDOUTS: They will be available on Canvas.

TOPICS TO BE COVERED: A detailed schedule is available at the end of the syllabus.

COMPUTER PACKAGES:

1. **MySQL**. You can either use the MySQL command line console or install MySQL Workbench.

2. **Microsoft SQL Server 2019** with **Microsoft SQL Server Management Studio.**

GRADING:

Attendance	10%
Homework (Individual)	30%
Project (Individual)	40%
Final Exam	20%

Project Breakdown:

1. Dataset selection	5%
2. Entity Relationship (ER) diagram	10%
3. Database creation, table population, business questions	10%
4. SQL queries + visualizations	10%
5. Final video presentation	5%

GRADING RULES: Clerical scoring errors will be corrected without hassle, but for other re-grades you must hand back the work and send an email; the entire assignment will be subject to re-grading. You must submit any re-grading requests via email message within 5 calendar days from when the assignment is returned. In your message, you should clearly explain why you are requesting a re-grade. While I will consider the specific concerns cited in your message, I will re-grade the entire assignment. Your new score might be higher, lower, or the same as a result. Please remember that small changes in your grade on a single assignment might not affect your overall course grade.

LATE SUBMISSION: Assignments should be submitted on the date and time that they are due (as stated on each assignment.) Late assignments will be accepted with a 10% penalty per day and a score of zero will be awarded if submitted later than 4 days after the due date.

CLASS ATTENDANCE: Attending all the classes is **mandatory** in the sections that you are assigned to. Switching sections due to schedule conflicts will not be allowed under any circumstances. Attendance will be taken at the beginning of every class. I'll take attendance by calling people's names randomly (rather than going down the list in alphabetical order.) If your name is called and you arrive later, then you will be marked absent for that class session. **No negotiation.** Also, to earn attendance points, you need to attend the **entire** class session. Leaving midway or arriving late (after the attendance is taken) will count as not attended.

FINAL EXAM: The final exam will be scantron-based, closed-book, closed-notes, closed-computer, closed-internet. The format of the exam will vary. They will be multiple-choice questions testing your conceptual understanding of SQL and data modeling concepts. The questions may entail analyzing SQL query results, SQL queries, data models, or anything else that we have talked about in the class. **No practice questions will be given.** The homeworks and notes serve as key resources for preparation of the exam. Please note that the purpose of the exams is to assess your understanding of the concepts covered in the class. Working on homeworks is not mutually exclusive from preparing for the exams. If you work on the homeworks, understand the notes, know how to write SQL, and know how to build/read data models, then it doesn't matter how questions are framed, you should be able to answer them.

Learning Objectives:

1. Understand the fundamentals of relational database modeling and database normalization.
2. Learn the concepts of Structured Query Language (SQL) and evaluate how it can be used to retrieve and transform data from relational databases.
3. Retrieve data from the database using SQL joins, grouping, subqueries, aggregate functions, and window functions.
4. Learn how to develop a data architecture solution from scratch.

Academic Honor Code:

All students are expected to adhere to the University of California, Davis' Code of Conduct as noted here: http://sja.ucdavis.edu/files/cac.pdf.

Schedule on the Next Page

Schedule (Tentative): This is a tentative schedule. Contents and sequence are subject to change.

Date	Assignments Due	Topics Covered
09/27/2024		Data modeling and database design – Part 1
10/04/2024		Data modeling and database design – Part 2
10/11/2024		Data modeling and database design – Part 3
10/18/2024		Data modeling and database design – Part 4[1]
10/22/2024 NO CLASS	**Project Dataset selection**	**NO CLASS**
10/24/2024 NO CLASS	**Homework 1**	**Data modeling homework**
10/25/2024		SQL – Part 1 (Joins and UNION)
11/01/2024		SQL – Part 2 (Aggregate functions and Grouped queries)
11/05/2024 NO CLASS	**Project ERD**	**NO CLASS**
11/07/2024 NO CLASS	**Homework 2**	**SQL homework – Joins and grouped queries**
11/08/2024		SQL – Part 3 (Subqueries)
11/15/2024		SQL – Part 3 (Subqueries continued) SQL – Part 4 (Data manipulation) SQL – Part 5 (Window functions)
11/21/2024 NO CLASS	**Project database creation, table population, business questions**	**NO CLASS**
11/22/2024		SQL – Part 5 (Window functions)
12/06/2024		SQL – Part 6 (Recursive queries in SQL Server)
12/07/2024 NO CLASS	**Homework 3**	**SQL homework – Subqueries, Window functions, Recursive queries in SQL Server**
12/12/2024 NO CLASS	**Project – SQL queries + visualizations in Tableau + Video presentation**	**NO CLASS**
12/13/2024	Final Exam (in-class) – 5:00 PM – 7:00 PM	

[1] Guest speaker

Final Project Guidelines

For the project, you should select a comprehensive data set, design, and develop a database for it, formulate business questions, write SQL queries to answer the business questions, and visualize your results to show insights. The project is divided into the following phases:

Phase 1: Dataset selection	Select a comprehensive dataset for your database. The dataset that you choose should be mapped into at least 6 tables in the database. Please note, you do not need to create these tables at this time; however, some forethought should be applied to avoid any surprises when you get to the subsequent phases of the project. Write a one-page summary containing the following: 1. Source of your data. If downloaded from the internet, then please share the link. 2. How many data files? What are the relationships among various data files? 3. How many tables do you anticipate in your database? Your response at this stage can be approximate.
Phase 2: Entity Relationship (ER) diagram	Using the principles of normalization, data modeling, and data lake architecture we will discuss in the class, create a full ER diagram. Advocate a data warehousing framework for raw data storage (staging layer), cleansing (transformation layer) and structured data (warehouse layer). You can use Lucidchart or any similar tool for diagramming the entity relationship model.
Phase 3: Database creation, table population, and business questions.	In this phase, you will create the database, database tables, and the corresponding relationships in MySQL. You will populate the tables in MySQL. Finally, you will write at least 8 business questions which can be answered by writing SQL queries. You do not have to write these queries in this phase, but your questions should be such that they can be answered using SQL queries. Also, your questions and queries should be of varying complexities. Your deliverable will be a three to four-page report containing the following information: 1. Discussion of how you converted the dataset into tables. 2. Challenges faced during importing of your data and how did you overcome these data importation challenges. 3. A complete data dictionary for every table in your database. 4. The list of business questions.
Phase 4: SQL + data visualizations	In this phase, you will write SQL queries for the business questions that you created in Phase 3. Your SQL queries should be of varying complexity. You will be graded based on the complexity of queries and the insights you obtain from the queries. Please note that not all queries need to have visualizations. Your queries should be submitted as a **separate (*.sql) file**. Your visualizations can be submitted as a separate PDF file. Do not submit Tableau workbooks or files. You can prepare a dashboard, if you like.
Phase 5: Final video presentation (no in-class presentation)	A video presentation of no more than 15 minutes. In this video presentation, you will show your PowerPoint containing the business questions and the results/visualizations of the SQL queries. Each business question should be on its own slide. Please do not show your SQL queries on the slides; instead, you should present business questions and the insights from the business questions

	that were derived using the SQL queries. You are encouraged to show visualizations for any query results if you have.