OIDD 3150: Databases for Analytics (Q3 2024) (0.5cu)

Lorin M. Hitt (lhitt@wharton.upenn.edu)

(DRAFT: November, 2023 PRELIMINARY)

Relational databases are the primary way in which business data is stored and processed. This course focuses on the analysis of data in databases and the development of databases to support analytical tasks. Over the course of the semester, students will learn the database language SQL and use this language to perform analytical tasks on existing and self-created databases. In addition, we will cover database scripting languages and extensions.

The course is intended as students with little or no database background and does not presume prior computer science or coding experience. This course is nearly all hands-on coding. Students interested in more conceptual discussions of technology should consider other OIDD offerings.

In the past, OIDD315 was taught combined with OIDD105 as a 1cu class. If you want a similar experience, you can take 315 and 105 simultaneously in Q3 (unfortunately scheduling constraints prevented a Q4 offering).

Course Format:

Class: Class time will be a mix of introducing and discussing the material and in-class exercises where it is easier to learn "hands on". Students will be asked to bring their computers and work individually or in teams to solve problems in class.

Regular Workload: There will be regular activities (approximately once per week) that must be completed at an acceptable level of accuracy for full credit. Students are permitted to omit 1 of these over the course of the quarter. In many cases, time will be allocated in class for doing these exercises.

Quizzes: There will be a mid-quarter checkpoint quiz and final exam covering all of the material in the class.

Project: There will be a small final project where you are asked to identify, import, clean and prepare a database for analysis, then perform an interesting analysis on the data using database techniques. The project should be about the same size as a problem set in terms of content/effort.

Course Materials. The course textbook is:

(Optional but very strongly recommended) Syyverson and Murach (2022). *Murach's SQL Server 2022 for Developers* (ISBN: 1943873062)

This is a trade book available from Amazon and many other outlets. While it is not expensive, it is also not free, so please do not use illegal copies. Prior additions are also acceptable (e.g. 2016, 2019) but page/chapter numbers may not line up with my reading guides.

There will also be online readings and supplementary material describing the datasets that we will be analyzing.

Mandatory Computer Resources

You will need a modern laptop computer (PC or Mac is acceptable) and a fast internet connection. If you have a PC you have the further option to run the database locally but that is not critical or necessary. From time to time, we will also be using the computer labs. As such, if you are not a Wharton student, you need a Wharton domain account which you can obtain from the Wharton Computer Consultants.

Microsoft Azure Data Studio. This is a data management tool that, among other things, can connect to SQL Server. It is a free download for Windows and Macs.

Amazon Web Services – SQL Server. You will be able to create an account and have your own private database server in the cloud. We will be using other services on AWS as well. You may need a credit card to access AWS so plan accordingly. Total expenditure if you manage well is anywhere from $0 to maybe $20-30 at most. Be very careful when you set up your server to pick free/low cost options so you do not incur unnecessary expense.

(Required). If you are using a laptop, get an external mouse (either wired or wireless). This will increase your programming productivity significantly (best $5 you will ever spend!).

(Optional). I also highly recommended that programmers use large screens (24" or better). Studies have shown this increases developer productivity.

<u>Grading and Evaluation</u>.

Weekly Activities (20%). Approximately weekly activities which are graded on a Pass/Fail basis. Any failed items can be remediated (once) to obtain a passing score. Students are permitted to omit 1 of these activities with no penalty (except those labeled "mandatory"). Remediation attempts should be completed within 1 week after receiving the score.

Class Project (20%). There is a class project which involves obtaining, importing, cleaning and analyzing a dataset to answer an interesting question. There are two deliverables: a proposal due around the middle of class and the final product which includes a 5-page writeup of what you did, supporting code/data/analysis files and a short presentation (which probably will not actually be presented). The ideal group size is 3 but you are not required to work in a group. I will consider larger groups on an individual basis for larger/more ambitious projects. It is due at the end of class. On a six-week schedule, this comes up very quickly so the time to start thinking of ideas is *now*.

Exams (50%). There will be an intermediate (10%) and final (40%) exam. In the world of large language models, I can no longer do take home exams so please plan accordingly.

Class Participation (10%). Students are expected to prepare, attend class, actively participate, and make good use of course resources (including the support staff and the instructors out of class time). The class participation grade will reflect our subjective evaluation on these dimensions as well as objective observation of class attendance. You are permitted to miss two classes without penalty. If you decide to not attend class, this will affect your grade.

Grade Distribution. There is no pre-specified grade distribution.

<u>Other Course Policies</u>

Regrades. Regrade requests (other than simple tabulation errors), must be submitted in writing within 1 week after grades for that activity are returned. Answers awarded partial credit are not eligible for regrade consideration unless the original answer was completely correct.

Deadlines. Assignment deadlines are firm because we often review the assignments in class immediately following the deadline. If for some reason you are not able to complete an assignment (e.g., you can't get your code to work…) submit what you have by the deadline. If you have a conflict on a deadline date, skip the assignment or submit it early. You cannot remediate an exercise which you did not submit on time.

Collaboration. You are free to discuss any and all course material with your fellow students and the course staff, including approaches to the assignments. You can also work together on most assignments in small groups. However, you are not allowed to share code or answers on any graded assignments outside your small work team. You are also not permitted to use materials from other courses or to copy code directly from Internet sources. All collaborators or should be identified by name in the submitted documents (distinguishing between your work team and anyone you spoke with in preparation of the assignment). Activities started in class as a group can be completed individually. Group projects assigned or initiated outside of class should have only a group deliverable.

Regardless…I strongly discourage "divide and conquer" strategies on assignments where questions are divided among group members or "you drive, I watch" programming where one student writes all the code and the other watches, gets coffee, etc. You cannot learn these skills without actual personal experience. Programmers write code, and you can't write and test code without touching the computer.

ChatGPT, Language Models and AI. ChatGPT is very good at writing SQL code. However, the purpose of this class is for you to learn how to write this code on your own. This is useful for many reasons including the fact that AI tools often make subtle mistakes (and different mistakes than a human would make) or misinterpret questions. You are welcome to use ChatGPT or other online tools to help with assignments, but you should use them to *assist* not to *complete* your work -- verbatim copying of ChatGPT answers is not permitted.[1] Remember, however, that you will not have access to the Internet on the exams.

[1] How will I know? I teach with some very specific coding conventions that ChatGPT may or may not be able to replicate (especially for more complicated queries). Also, students usually do not preface answers with "As a large language model…". My usual policy is: if an answer is correct and there is no reason to believe it was copied, then you get regular credit. If any answer is wrong and I don't believe it was original work, I give the maximum deduction. This policy is not new or really directed toward AI tools – it applies

Attendance. You are expected to come to class and to be prepared. From time to time, something may happen in class that requires your physical presence. I will also, from time to time, take attendance. You are permitted to miss one of these over the course of the semester before it affects your grade (this is in addition to any University-approved absences such as religious observances). You do not need to tell me why you are missing class or get permission. If you need to miss class due to a religious holiday, I am happy to go over the material by appointment or during office hours or to record a session of the class by request.

Support. There will be office hours by both the instructors as well as undergraduate and graduate teaching assistants. Office hours will be hybrid with both in-person and Zoom simulcast -- for coding classes, Zoom can be better than in-person meetings because we can review code on the screen.

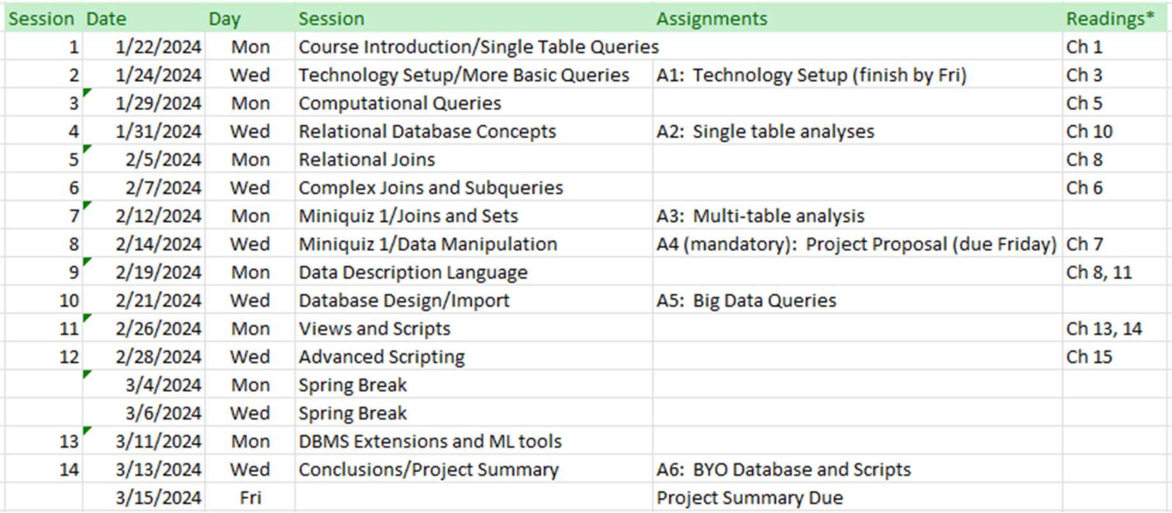
We will be using Piazza, an online discussion tool, for online course questions. A few guidelines about the use of Piazza which will make everyone happier:

- If you have a general question or something about the course material, use Piazza. If you have a personal question, e-mail the instructor.
- Please do not post code to Piazza as an open message. If you need a quick evaluation of your code, post it is a private message to instructors. If you have a more complicated question ("why doesn't this work?") that is probably best done in person or by e-mail.
- Please do not make all your questions private. It defeats the purpose of an open discussion forum (the exception is when you need to post code).
- Please do not spam questions on Piazza. If you have lots of questions, that is best done during office hours (online or in-person).
- You can make your questions anonymous to other students but the instructors and TAs can see your real name… so be nice.
- You too can answer questions on Piazza. This is appreciated by the course staff.

Electronics. We will be using computers in class. However, use of your computer is strictly limited to course work (if you have completed the work for the day or otherwise want to use your computer for other purposes, please leave the room). You are not permitted to make audio or video recordings of class sessions under any circumstances. If you need audio or video of class for some reason, I will arrange it with the school. Cell phones should be turned off or silenced.

just as well to more traditional Internet resources like StackOverflow (which is also a very good resource for finding answers to SQL questions).

Preliminary Schedule (subject to change, more so in the latter part of the semester)



Session	Date	Day	Session	Assignments	Readings*
1	1/22/2024	Mon	Course Introduction/Single Table Queries		Ch 1
2	1/24/2024	Wed	Technology Setup/More Basic Queries	A1: Technology Setup (finish by Fri)	Ch 3
3	1/29/2024	Mon	Computational Queries		Ch 5
4	1/31/2024	Wed	Relational Database Concepts	A2: Single table analyses	Ch 10
5	2/5/2024	Mon	Relational Joins		Ch 8
6	2/7/2024	Wed	Complex Joins and Subqueries		Ch 6
7	2/12/2024	Mon	Miniquiz 1/Joins and Sets	A3: Multi-table analysis	
8	2/14/2024	Wed	Miniquiz 1/Data Manipulation	A4 (mandatory): Project Proposal (due Friday)	Ch 7
9	2/19/2024	Mon	Data Description Language		Ch 8, 11
10	2/21/2024	Wed	Database Design/Import	A5: Big Data Queries	
11	2/26/2024	Mon	Views and Scripts		Ch 13, 14
12	2/28/2024	Wed	Advanced Scripting		Ch 15
	3/4/2024	Mon	Spring Break		
	3/6/2024	Wed	Spring Break		
13	3/11/2024	Mon	DBMS Extensions and ML tools		
14	3/13/2024	Wed	Conclusions/Project Summary	A6: BYO Database and Scripts	
	3/15/2024	Fri		Project Summary Due	