San José State University
Department of Applied Data Science

DATA 225
Database Systems for Analytics

Section 21
Fall 2023

Course and Contact Information

Instructor:	Ron Mak
Office location:	Clark Hall CL 325 (mostly working from home)
Email:	ron.mak@sjsu.edu
Website:	http://www.cs.sjsu.edu/~mak/
Office hours:	W 6:00 – 7:00 PM, Clark Hall CL 325
Class days/time:	M: 6:00 – 8:45 PM
Classroom:	BBC 102
Prerequisites:	Classified standing or instructor consent.

Course Catalog Description

"Design operational and analytical databases with relational, dimensional, and NoSQL data models; deploy data analytics applications with SQL programming and data access APIs; perform data cleansing, extract, transform, and load (ETL) operations for data warehouses and online analytical processing (OLAP)."

Course Format

This class will meet in person in the classroom. Exams will be given in the classroom.

Faculty Web Page and Canvas

Course materials, syllabus, assignments, grading criteria, exams, and other information will be posted at my faculty website at http://www.cs.sjsu.edu/~mak and on the Canvas Learning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking these websites to learn of any updates. You can find Canvas video tutorials and documentations at http://ges.sjsu.edu/canvas-students

Course Goals

This class emphasizes designing and developing operational and analytical **data management applications.** These applications are client-side Python programs that interact with server-side databases and data warehouses. Analytical topics include performing different types of data analysis such as time series and multiple regression in the database server using SQL, and then graphing the results on the client side using Python. Project teams will use a Python GUI (graphical user interface) library to develop applications where users interact with backend databases and data warehouses through windows, buttons, menus, and dialog boxes.

Course Learning Outcomes (CLO)

Upon successful completion of this course, students will be able to:

CLO 1: Understand and explain the functional aspects of relational and NoSQL databases.
CLO 2: Choose appropriate data models and data management tools for a given application.
CLO 3: Design appropriate relational schemas for operational databases and star schemas for analytical databases.
CLO 4: Implement well-design databases and data warehouses.
CLO 5: Perform data management operations such as normalization and extract-transform-load (ETL).
CLO 6: Program proficiently in SQL, including the use of views and stored procedures.
CLO 7: Write command-line- and GUI-based Python analytical applications that enable users to interactively access and manage data in databases and data warehouses.
CLO 8: Develop a substantial data management project in a small team and provide an oral presentation and a written report about the project.
CLO 9: Confidently discuss issues about database systems for analytics.

Recommended Books

Title:	**Database Systems: Introduction to Databases and Data Warehouses, Edition 2.0**
Authors:	Nenad Jukic, Susan Vrbsky, Svetlozar Nestorov, Abhishek Sharma
Publisher:	Prospect Press, 2021
eTextbook:	978-1-943153-67-1 available from Redshelf.com and VitalSource.com
Paperback:	978-1-943153-68-8 available from Redshelf.com
	Many examples in class will come from this excellent book. We will also use the book's database modeling tools at https://erdplus.com
Title:	**Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data**
Authors:	Wilfried Lemahieu, Seppe vanden Broucke, Bart Baesens
Publisher:	Cambridge University Press, 2018
ISBN:	978-1107186125
	A more advanced text. Well-written with many examples and colorful diagrams.
Title:	**The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition**
Authors:	Ralph Kimball and Margy Ross
Publisher:	Wiley, 2013
ISBN:	978-1118530801
	Ralph Kimball is a pioneer of data warehousing and dimensional data modeling. I have worked closely with Dr. Kimball in the past.

Some Useful Websites

- **SQL Tutorial**
  https://www.w3schools.com/sql/
- **MySQL 8.0 Reference Manual**
  https://dev.mysql.com/doc/refman/8.0/en/
- **MySQL 8.0 Error Message Reference**
  https://dev.mysql.com/doc/mysql-errors/8.0/en/
- **MySQL Connector/Python Developer Guide**
  https://dev.mysql.com/doc/connector-python/en/
- **MySQL Workbench**
  https://dev.mysql.com/doc/workbench/en/
- **PyQt5 Tutorial**
  https://www.tutorialspoint.com/pyqt5/index.htm
- **Tutorials on MySQL Shell and MySQL Workbench**
  http://www.cs.sjsu.edu/~mak/tutorials/index.html

Software to Install

You will need the **MySQL database server** and **Python**.

Install the free MySQL Community Server for your platform (Windows, MacOS, or Linux) from https://www.mysql.com/downloads/. Also install the database management tool MySQL Workbench at https://www.mysql.com/products/workbench/

A good way to install Python is via **Anaconda**: https://www.anaconda.com. This will install the Python interpreter, Jupyter notebook, and several key data science libraries.

After installing Anaconda, you can execute the following commands if you need to update the installed packages to their latest versions:

```
conda update conda
conda update --all
```

You will need to install the **MySQL Connector/Python** which allows a Python program to access a MySQL database:

```
pip3 install mysql-connector-python
```

If you want to create animations with **Matplotlib** inside a Jupyter notebook, you must install the **`ipympl`** tool with these commands:

```
conda install -c conda-forge ipympl
conda install nodejs
jupyter labextension install @jupyter-widgets/jupyterlab-manager
jupyter labextension install jupyuter-matplotlib
```

To create Python programs with a GUI (graphical user interface), install **Qt Designer** which will enable you to design windows with labels, text boxes, menus, buttons, etc.: https://build-system.fman.io/qt-designer-download

Install the **PyQt5 modules** with this command:

```
pip3 install PyQt5
```

Then you will be able to create GUI-based Python applications that enable users to interact with your database via forms.

Project teams

You will form small project teams. *Team membership is mandatory for this class.* The teams will last throughout the semester. Once the teams are formed, you will not be allowed to move from one team to another, so form your teams wisely!

Course Requirements and Assignments

You should have good Python programming skills and be familiar with its development tools.

Weekly individual- and team-based assignments will provide practice with database design techniques and give you experience developing code for database applications. *Each student on a team will receive the same score for each team-based assignment.*

Submit each assignment into Canvas (only one submission per team for a team assignment), where the scoring rubric will be displayed. Late assignments will lose 20 points and an additional 20 points for each 24 hours after the due date.

The university's syllabus policies:

- University Syllabus Policy S16-9 at http://www.sjsu.edu/senate/docs/S16-9.pdf.
- Office of Graduate and Undergraduate Program's Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/

"Success in this course is based on the expectation that students will spend, for each unit of credit, a minimum of 45 hours over the length of the course (normally three hours per unit per week) for instruction, preparation/studying, or course related activities, including but not limited to internships, labs, and clinical practica. Other course structures will have equivalent workload expectations as described in the syllabus."

Exams

The exams will test understanding (not memorization) of the material taught during the semester and now well each of you participated in your team assignments and project. Instant messaging, e-mails, texting, tweeting, file sharing, or any other forms of communication with anyone else during the exams will be strictly forbidden.

Instead of a single midterm examination, there will be three "checkpoint" exams during the semester which will be conducted in Canvas. There will be no make-up examinations unless there is a documented medical emergency.

There will be no final examination — the team data management project takes its place. The project includes an oral presentation at the end of the semester.

Academic Integrity

A data scientist must possess data analytics skills and the integrity to perform the analyses honestly. Therefore, exams in this class test both skills and honesty. The latter means strict adherence to the university's Academic Integrity Policy. **Any violations, including sharing answers, will result in a score of zero for the entire exam.** Repeated violations can result in failing the class and being reported to the Student Conduct Office.

Team Project

In addition to the team-based assignments, each project team will develop a significant data management application during the semester using Python and a database. This project will involve:

- A GUI-based Python application that accesses a well-designed database and/or data warehouse. The application should perform server-side data analytics and display results on the client side.
- Data sources chosen by each team from the internet or other sources, or fake (but realistic) data generated by tools such as **`mockaroo`** (https://www.mockeroo.com).
- Use of data management tools that demonstrates a strong understanding of how to effectively apply the tools in meaningful ways on the chosen data sources.
- A written report (15-20 pp.) that describes the requirements and goals of the project and how they were met. The report should include a description of the project's ETL procedure, entity relationship (ER) diagrams, and screenshots of key application displays.

Each team will give an oral presentation at the end of the semester that includes a demo of its application. The rest of the class (along with the instructor) will score each presentation based on a given set of criteria. *Each student on a team will receive the same score for the project.*

Grading Information

Individual total scores will be computed with these weights:

10%	Checkpoint exam #1*
10%	Checkpoint exam #2*
10%	Checkpoint exam #3*
30%	Assignments**
40%	Database application project***

* *individual scores*
** *individual and team scores*
*** *team scores*

Course grades will be based on a curve. **The median total score will earn a B+.** The database project takes the place of a final exam.

Postmortem Report

At the end of the semester, each student must also turn in a short (under 1 page) individual postmortem report that includes:

- A brief description of what you learned in the course.
- An assessment of your accomplishments for your team assignments and design project.
- An assessment of each of your other project team members.

Only the instructor will see these reports. How your teammates evaluate you will be factored into your course grade.

Technology Requirements

Students are required to have an electronic device (laptop, desktop, or tablet) with a camera and microphone. SJSU has a free equipment loan program available for students: https://www.sjsu.edu/learnanywhere/equipment/index.php

Students are responsible for ensuring that they have access to reliable Wi-Fi during tests. If students are unable to have reliable Wi-Fi, they must inform the instructor, as soon as possible or at the latest one week before the test date to determine an alternative. See Learn Anywhere website for current Wi-Fi options on campus.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Program's Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/.

DATA 225
Database Systems for Analytics

Section 21
Fall 2023
Instructor: Ron Mak

Course Schedule (subject to change with fair notice)

Week	Date	Topics
1	Aug 21	Overview of the course Introduction to databases and database applications Introduction to SQL MySQL and MySQL Workbench *Form project teams*
2	Aug 28	**Operational databases** Python connection to a MySQL database Create databases and load tables Create users and grant privileges Dump and restore a database Data modeling Entity-relationship (ER) diagrams Relationships between entities
	Sept 4	*Labor Day holiday (no classes)*
3	Sept 11	Creating ER diagrams Mapping ER diagrams to relational schemas Mapping relational schemas to physical models Referential integrity constraint EER diagrams
4	Sept 18	Named constraints Table creation and deletion with constraints Update anomalies Functional dependencies and normalization Normal forms
5	Sept 25	***Checkpoint exam #1*** *(held during the first part of the class)* Aggregate functions GROUP BY and HAVING INSERT INTO with SELECT Nested queries
6	Oct 2	Cascading updates and deletes SQL window function Text functions Types of joins Views Indexing

Week	Date	Topics
7	Oct 9	Nested SELECTs and views Database design example GUI-based Python database applications Stored procedures Triggers
8	Oct 16	Object-relational mapping (ORM) ACID properties Concurrency Transaction management Checkpoints and recovery Reliability and RAID
9	Oct 23	Data wrangling Replacing missing values Multiple linear regression Outliers
10	Oct 30	***Checkpoint exam #2*** *(held during the first part of the class)* **Analytical databases** Operational vs. analytical databases Data warehousing and data marts Dimensional modeling and star schemas
11	Nov 6	Dimension tables and fact tables Types of fact tables Fact table granularity Extract-transform-load (ETL) Slowly changing dimensions Online transaction processing (OLTP) Online analytical processing (OLAP)
12	Nov 13	Factless fact tables SQL ROLLUP Time-series analysis Moving averages and exponential smoothing
13	Nov 20	**NoSQL databases** Documents and collections CAP theorem MongoDB queries and joins Python connection to a MongoDB database
14	Nov 27	***Checkpoint exam #3*** *(held during the first part of the class)* Query processing and optimization Security and public key encryption Data virtualization
15	Dec 4	***Oral project presentations***