**1/13/2025**

Self annotation/notes

**DASC 5333 Database Systems for Data Science
CSCI 4333 Design of Database Systems
Spring 2025
General Information and Course Policies**

**1. General Information**

22183 CSCI 4333.1 Design Of Database Systems MW 1:00-2:20 Delta 241
22178 DASC 5333.1 Database Systems for Data Science T 1:00-3:50 Delta 237

This is a first course in database for both BS CS students and MS DS student.

**1.1 Instructor**

Dr. Kwok-Bun Yue, Professor of Computer Science and Computer Information Systems, Chair, Data Science
Delta 163, 281-283-3864, yue at uhcl.edu; URL: <http://dcm.uhcl.edu/yue/>

My regular office hour will be held on 1/14/2025 to 4/28/2025: MW 2:25PM to 3:45PM and Tuesday 3:50-4:10PM. Office hours will be conducted in person (Delta 163) and upon request, via [Zoom meeting: 616 099 762](https://uhcl.zoom.us/j/616099762?pwd=TlhGOXNrQjlCTVhkSTVEeTdBUDBpUT09). (Q: do I need to arrange for it two days ahead?) You can schedule a meeting with me outside my office hours by sending an email to me: yue @ uhcl dot edu. You are encouraged to communicate your questions with me through email. I usually respond quick.

**1.2 Teaching Assistant**

**1.2.1 CSCI 4333.1, CSCI 4333.2 and DASC 5333**

Pavan Kumar Kodavali



For regular correspondence with the TA, send it to UHCL Email Id: kodavali@uhcl.edu. Set up the UHCL spam filter server for your UHCL account to accept this email address as an approved sender. Otherwise, your email may be quarantined by the spam filter server. If you want me to be aware of any particular communications with the TA, you may copy the email to me.

Tentative TA Office hours:

|  |  |
| --- | --- |
| **Day** | **TA Office Hours** |
| Monday | 10:00 AM - 1:00 PM |
| Tuesday | 10:00 AM - 1:00 PM |
| Wednesday | 11:00 AM - 1:00 PM |
| Thursday | 10:00 AM - 1:00 PM & 4:00 PM to 7:00 PM |

Pavan will be stationed in the Delta Lab during his office hours. You may also request Zoom for TA: <https://uhcl.zoom.us/my/pavankodavali>

**1.3 Laboratory Administrations**

You may address account and software problems of the DCM server to the systems administrator, Ms. Krishani Abeysekera. and her assistants. Always copy your email to me.

**1.4 Other Useful Information**

* UHCL emergency hot line (to check weather related closing, for example): 281-283-2221.

**1.5 Textbooks (Recommended, Optional)**

Ricardo., Katherine, & Urban, Susan (2015)*Databases Illuminated*, 3rd Edition, Jones & Bartlett, Mississauga, Ontario, Canada.

**1.6 Course Description**

**CSCI 4333:**

From Catalog: Prerequisite: CSCI 2315. Design of database systems, data description and manipulation languages, data models, entity-relationship model, relational model, SQL, relational algebra, normalization theory, DBMS, Internet, data base design, data flow diagrams and implementation of data base systems. Laboratory instruction.

**DASC 5333:**

From Catalog: Design of database systems, data definition and manipulation languages, data models, entity-relationship model, relational model, SQL, relational algebra, normalization theory, DBMS, Internet, database implementation. Focus on applying DB theory and practice to support data science applications. Laboratory instruction. Prerequisites: DASC 5032 or equivalence

**1.7 Student Learning Outcomes (SLO)**

**CSCI 4333:**

After completing the course, the students are expected to be able to

1. Describe the stages of database design, various database architectures and data models.
2. Explain the concepts of the entity-relationship model, and the relational model.
3. Explain the theoretical background of relational database, including relational algebra and relational calculus.
4. Implement relational database systems using DBMS, SQL, embedded SQL, including both data definition and manipulation languages
5. Explain the importance of normalization of databases, and convert a given relational database into different normal forms.

**DASC 5333:**

After completing the course, the students are expected to be able to

1. Describe the fundamental concepts of database systems.
2. Apply simple data modeling techniques using the UML class diagram and the relational model.
3. Explain the features and theory of the relational model.
4. Implement relational database systems for data science applications using relational DBMS, SQL, and embedded SQL with Python.
5. Explain the importance of normalization of relational databases, and convert a given relational database to appropriate normal forms.
6. Explain the concepts of NoSQL databases and implement simple NoSQL database solutions.

**1.8 Prerequisites**

The following courses or their equivalent are required:

* CSCI 4333: CSCI 2315 Data Structures
* DASC 5333: DASC 4301 Python Programming for Data Science, or equivalent

Languages: The course uses SQL, Python and (MongoDB Query Language (MQL) with Javascript, and/or Cypher (for Neo4j)). No prior SQL, MQL or Cypher language knowledge is assumed. Students are expected to know an object-oriented language, such as Python, Java, C# or C++. Proficiency in Python is important in data science in general, and this course in particular.

**1.9 Course Format**

Traditional lectures, homework and programming assignments.

**2. Course Policies and Guidelines**

Please see: <http://dcm.uhcl.edu/yue/course_policy.html>

**3. Grading Policy**

Grades will be assigned based *solely* on homework and examination scores. *No other factors will be considered.* In particular, students have requested me to reconsider their grades using the following reasons in the past:

1. Expected a higher grade
2. Good course participation
3. Good improvement during the semester; better final grades than mid-term grades
4. Have put in extra efforts
5. Need to avoid probation
6. Financial needs
7. Loss of scholarship
8. Loss of job opportunity
9. Loss of practical training opportunity
10. Avoid probation; avoid suspension
11. Need to graduate
12. Company relocation
13. Immigration status needs
14. Family needs
15. Sickness during the semester
16. and many others.

These requests had all been declined politely but firmly in the past.

There will also be *no* 'special project' that you can work on to improve your grades after the final examination. Anything I offer to one student will be offered to the entire class.

The total score is computed using the following percentages:

Homework: 30%
Mid-term Exam: 30%
Final Exam: 40%

Last Day to Drop/Withdraw: *April 7, 2025 (Monday)*

**Grade Assignment Table**

|  |  |
| --- | --- |
| [92..100] | A |
| [90..92) | A- |
| [87..90)  | B+ |
| [83..87) | B |
| [80..83) | B- |
| [77..80) | C+ |
| [73..77) | C |
| [70..73) | C- |
| [67..70) | D+ |
| [63..67) | D |
| [60..63) | D- |
| [0..60) | F |

o *ask a lot of questions*

*For most assignments, late assignments are accepted with a penalty of 25% deduction per week day after the due date*

* No make-up exam except in *verified emergencies with immediate notification*.

Penalty on cheating will be*extremely severe*

*Install XAMPP:*



MySQL server:





SQL client:



**Introduction to the course
DASC 5333/CSCI 4333**

by K. Yue

**1. Promotion**

* This course is (hopefully) one of the more useful CS/DS courses for students.
* World data is estimated to double every two years.

**2. How to be successful in the course**

General Course Suggestions:

1. Course expectation is demanding.
2. Please consider forming the habit of listening carefully and asking a lot of questions.

General Professionalism:

1. Attitude
2. Be considerate.
3. Be helpful and useful to others.
4. Be a good listener.
5. Be responsive.
6. Hardworking
7. Attention to details.
8. Focus: uni-tasking

Some general tips:

1. Engagement: Participate. Ask questions, a lot of them. Help others. Plan ahead.
2. Preparation: start as early as possible and do not fall behind.
3. Don’t copy and paste. Instead, copy, integrate, and apply.
4. [SEE-I](https://en.everybodywiki.com/SEE-I): State, Elaborate, Exemplify and illustrate.
5. Form good habits.

Some good traits of Computer and Data Scientists:

1. Habits of trying to make sense of stuff.
2. Intellectual curiosity.
3. Tinkering and experimentation.
4. Open-minded, not dogmatic.
5. A large tool set.

**3. Resources**

* Companion materials of our textbook: please consult the course page in Canvas for additional resources related to the textbook.
* Contents of the course will be based mostly on
	1. Lecture notes posted in the course website: <http://dcm.uhcl.edu/yue/courses/joindb/current/index.html>.
	2. Classroom demonstrations.
	3. Assignments.
* Please read the appropriate pages in the textbook and lecture notes in this site *before* coming to the class.
* Document your learning. Bring a notebook to the class. Otherwise, it may be a good idea to print out the notes and bring them to the class so you can make notes during the class.

**4. Introduction**

* *Persistent* data is the backbone of many applications.
* Three main choices of storing persistent data:
	1. Files
	2. Databases: focus of this course.
	3. Cloud-based storage and database.
* Some advantages of DBMS (according to Ricardo, the optional textbook of this class):
	1. Sharing of data
	2. Control of redundancy
	3. Data consistency
	4. Improved data standards
	5. Better data security
	6. Improved data integrity
	7. Balance of conflicting requirements
	8. Faster development of new applications
	9. Better data accessibility
	10. Economy of scale
	11. More control over concurrency
	12. Better backup and recovery procedures
* How do we *make sense* of these 12 different advantages?
	1. Different textbooks may have different collections of the advantages of DBMS because of different classifications.
	2. No need to memorize them.
	3. Better to assimilate them and construct your own list.
	4. Make your own notes. Use [SEE-I](https://en.everybodywiki.com/SEE-I) (In your own words, state, elaborate, and exemplify with examples, and illustrate the concept.)
		+ However, do not overuse metaphors.
	5. Learning through documentation, communications, and teaching.
* What are some disadvantages of DBMS?
	1. Complexity
	2. Cost
	3. Learning curve
	4. Possible single points of failure and bottleneck

**5. A Simple Introduction to the Relational Model**

* Relational databases are the most popular databases: <https://db-engines.com/en/ranking>. It is based on the relational model.
* There are many other data models.
* In layman's term: A *table* (relation) is the basic unit of a relational database.
* A table is composed of many *rows* (tuples).
* Each row has many *column* (attribute) values.
* A primary key is roughly a *minimal* set of columns in a table that*uniquely identify* a row.
* Two tables can be related to each other by *foreign keys*. A foreign key is roughly a column in a table in which its value must be equal to the referenced value of the primary key in another table (called the paren or referenced table).
* Relational DBMS is the most popular DBMS. Examples:
	+ DB-engine ranking: <https://db-engines.com/en/ranking>
	+ Top 10 DBMS in Data Science: <https://towardsdatascience.com/top-10-databases-to-use-in-2021-d7e6a85402ba>
* SQL is the 'glue' in many DB systems.

***Classroom discussion***

Please ask questions about the toy University DB ([toyu](https://dcm.uhcl.edu/yue/courses/joinDB/Spring2025/notes/toyu/toyu.html))