DATABASES

(601.315/415/615)

Prof. David Yarowsky

Department of Computer Science
Johns Hopkins University
yarowsky@jhu.edu

August 31, 2021

601.315/415/615 - DATABASES

Instructor: Prof. David Yarowsky
Hackerman 324G
410-516-5372
yarowsky@jhu.edu

Head CA: Conner Delahanty
cdelaha1@jhu.edu

Meeting Time: Tu,Th: 3:00-4:15 PM (Baltimore Time)

Classroom: Online (details distributed to registered students)

Office Hours: TBA and by appointment. zoom details will be provided
CAs - TBA and by appointment. zoom details will be provided.

Textbooks

Required Textbook:

- A. Silberschatz, H. Korth and S. Sudarshan, *Database System Concepts*, 7th Edition, McGraw Hill, 2019, ISBN: 978-0-07-802215-9 or 6th Edition, McGraw Hill, 2010, ISBN: 978-0-07-352332-3.

Other Potentially Useful Textbooks:

- A reference guide of your choice for stored procedures and advanced mysql.
- Electronic references to the above are available on the class website.

Course Requirements

Class Participation:	5%
Homeworks (4):	24%
Midterm:	15%
Final Exam:	28%
Final Project:	28%

- Homeworks will include paper-and-pencil exercises and MySQL implementation exercises
- The midterm will cover material roughly through 10/19/21.
- The final exam will be cumulative, with approximately 1/3 of the content based on pre-midterm material.

Lateness Policy

- One homework assignment may be handed in up to 5 days late without penalty.
- No other late homeworks will be accepted.
- Final projects handed in late will receive a penalty of 10% for every day late.

Computer Science Academic Integrity Code

Academic honesty is required in all work you submit to be graded. **You must solve all homework and programming assignments entirely on your own (Homeworks 1/2/4), unless group work is specified in writing (Homework 3, Project)**. This means you must not show your program code, problem solutions, or work to other students. However, you may discuss assignment specifications with others in the class to be sure you understand what is required by the assignment. If you use fragments of source code from sources other than your text (such as on-line resources), you must put a reference to that effect in your homework submission. **Falsifying program output or results is prohibited.** Please see your professor if there are any questions about what is permissible. Students who cheat will suffer a serious course grade penalty in addition to being reported to university officials. You must abide by JHU's Ethics Code, available at http://jhunix.hcf.jhu.edu/~ethicsbd.

Solutions to Previous Exams and Homeworks

A copy of the previous year's midterm (and one other midterm) and their solutions will be explicitly distributed to students for practice and guidance regarding expectations, and students are encouraged to study using them. Likewise , Homework 4 is composed of questions given on previous final exams, and is intended as preparation for the final exam, with sample solutions given **after** HW4 is submitted but before the final exam.

With the above exceptions, students are explicitly forbidden from looking at or using other 601.315/415/615 exams, homeworks and/or sample solutions.

601.315 vs. 601.415/615

- 601.315/415/615 will be share common lectures.
- They will differ primarily in terms of assignments and grading.
- Homeworks in 601.415/615 will include 1 or more additional problems and the final project will include additional component(s) not required for 601.315.
- Exams will differ somewhat and will be graded on a different scale.
- Nevertheless, 601.415/615 should be manageable by advanced undergraduates and upperclass students are encouraged to enroll.

601.315/415/615 vs. 601.316/416/616:

- *Databases* (315/415/615, Fall) and *Database Systems* (316/416/616, Spring) are complementary courses and make a natural course sequence (see below).
- 315/415/615 focuses on:
  - how to design and use a database;
  - formal database models, theory and foundations;
  - database programming languages, especially SQL and PL/SQL;
  - object-oriented and XML-based data models and future directions (including data mining and natural language interfaces).
  - The final project will be application-focused (e.g. how to design an implement a database for a novel task) including practical execution of the concepts studied in the class.
- In contrast, 316/416/616 will focus on:
  - database internals and systems, including query and join processing, indexing, file organization, estimation and optimization

- database architectures, streaming and partitioning.
- The course project(s) will focus on database system internals and their development.

Can I take 316/416/616 as a stand-alone course without 315/415/615?

- Yes, 316/416/616 does not have 315/415/616 as a formal prerequisite.
- You should have some database experience before taking 316/416/616, however, either through prior employment or via a prior course.
- Graduate students who have prior database employment experience or have taken a prior course in database systems are normally expected to begin directly with 416.
- Anyone with a research focus in the databases area should certainly begin directly with 416/616.

Can I take 315/316 or 415/416 or 616/616 as a 2-course sequence?

- Yes.
- There will be modest overlap of material (10%) but taught via different perspectives and emphasis, and will serve as a good refresher.
- If you have not taken a prior course in databases and are interested in both the theory/applications and systems sides of the field, then this sequence makes a lot of sense and is encouraged.
- The instructors will work to make this a natural 2-course sequence.

Can I take 315/416 as a sequence?

- Yes, 416 does not require 415 as a prerequisite, but you should have done well in 315 and be prepared to do some background catchup to meet the expectations of the 416 instructor.

Can I take 415/316 as a sequence?

- Yes, if you are an undergraduate and would like to continue focusing on database systems and database systems internals but a less difficult level, then this sequence could make sense.

Final Projects

- Students will be able to select final projects of interest to them from a fairly diverse set of options.
- Details will be provided in class.
- Students may work in teams of 1 or 2 people.
- A project proposal will be due in early November, including a detailed system specification and design.
- The final project submission, including a full database implementation in MySQL, will be due shortly after the end of classes in December.
- For most projects, students will be required to populate and test their implemented database design with substantial quantities of *real world data* extracted from the world wide web or other online sources.

Sample Final Project Domains (previous years)

- Used car information (by model and year, from Edmunds)
- World geography and population data (from CIA world fact book)
- Movie industry data (directors, producers, actors, films, etc.)
- Olympic sports data
- JHU Fencing club and Anime film club
- Connecticut volunteer emergency rescue organization
- Fantasy hockey league
- Representations of acoustic data for speech recognition
- Astronomical and pharmaceutical databases for research support
- Bibliographic database for medical robotics
- Human genome databases
- Internet proxy server database
- Stock market news and price correlations (data mining)

Sample Final Project Domains (continued)

- Natural language interfaces to an earthquake database
  - Which country had the greatest number of earthquakes in 2020?
  - What was the magnitude of the most powerful earthquake in China?
  - What was the average magnitude of 2020 earthquakes in Asia?
  - List the years in which there are at least two earthquakes of magnitude greater than 7 on the same continent.
  - Which country had the most powerful earthquake in 2020?

```
SELECT Countryname
FROM   Quake
WHERE  magnitude IN
    ( SELECT MAX (magnitude)
      FROM Quake
      WHERE Year = 2017 )
```

SEGMENT 1 - Survey of Data Models

- Network and Hierarchal models (of historical interest)
- Entity-Relationship model (formal conceptual framework)
- Relational model
  - Formal representations: relational algebra and calculus
  - Relational query languages: SQL, QBE (Query-by-Example)
- Object-Oriented models

SEGMENT 2 - Database Design and Implementation

- Formal Analysis:
  - Integrity constraints
  - Domain constraints
  - Triggers
  - Functional dependencies
  - Normalization
- Practical Database Implementation:
  - MySQL (a detailed exploration)
  - Embedded SQL (in a host language like C or Perl)
  - PL/SQL and stored procedures

SEGMENT 3 - Database System Internals

- Query processing
- Query optimization
- Transaction processing
- Recovery systems
- Database security
- Database system architectures
- Parallel databases
- Distributed databases

SEGMENT 4 - Emerging Technologies and Applications

- Decision support systems
- Data mining
- Data warehousing
- Natural language interfaces
- Spatial, geometric and geographic databases
- nosql, datalog and xml-based data models
- DNA and Human Genome databases
- Multimedia Databases (image, sound, video, etc.)
- Very large text databases and information retrieval
- The impact of the WWW on database technology (and vv.)

$\Rightarrow$ 601.466 - Information Retrieval and Web Agents