IS 631
Enterprise Database Management

Course Syllabus

Fall 2016

Instructor: Songhua Xu

Lecture: Monday 10:00 am – 12:55 pm, CKB 317

Office Hour: Monday 9:00 am - 10:00 am, GITC 5107

E-Mail: songhua.xu@njit.edu

Course textbook:

- Avi Silberschatz, Henry F. Korth, S. Sudarshan, Database System Concept, McGraw-Hill, ISBN 0-07-352332-1, 6th edition.

Supplementary readings:

- Microsoft SQL Server 2012 Step by Step (Step by Step Developer), ISBN-13: 978-0735663862 ISBN-10: 0735663866
- A First Course in Database Systems (3rd edition) by Ullman and Widom.
- Database Systems: The Complete Book (2nd edition) by Garcia-Molina, Ullman, and Widom.
- Database Management Systems (3rd edition) by Ramakrishnan and Gehrke.
- Fundamentals of Database Systems (6th edition) by Elmasri and Navathe.
- P. Rob, C. Coronel, S, Morris DATABASE MANAGEMENT: DESIGN, IMPLEMENTATION, AND MANAGEMENT 10e (Tenth Edition), Thomson/Course Technology – Cengage Learning. ISBN 13: 987-1-111-96960-8. (in case you do not have basic database concepts and knowledge.)

Course Description:

This course introduces the foundations of database systems, focusing on data modeling, query design, and applications. It provides an understanding of the issues in designing and managing database systems as an essential organizational resource. The components of enterprise data management are covered, with a strong emphasis on data modeling as well as the

DBLC (Data Base Life Cycle). Implementing a database using SQL is an art and a science and will be addressed in the course. Data warehousing and data mining issues will also be examined.

Course Requirements:

Before enrolling in this course, students should pass IS331 with a decent performance. All content of IS331 is required for pursing subject matters in this course. More specifically, students should have a good knowledge and working skills with: (1) data modeling (primarily E-R data modeling), (2) relational database design (including database integrity issues), (3) professional and ethical responsibilities of database professionals, (4) query design in SQL, (5) identifying poorly designed databases and their rectification, (6) optimal design of databases invoking conceptual topics in relational decomposition, functional/multivalued dependencies and normalization (through 3NF, BCNF, 4NF and DKNF), (7) Denormalization and read-only/non-updateable databases, including data warehouses, and (8) Structured Query Language (SQL) and popular DBMS products.

Class Communication Space/Learning Management System:

We will be using Moodle, a state-of-the-art, open source, Learning Management System (LMS), and is nationally/internationally the fastest-growing LMS. We will be using this system for both online and face to face sections of the class, where I will be posting additional resources as needed throughout the semester. The powerpoint slides for each lecture will be available for download in Moodle.

Course Goals:

At the end of the course, you should be able to develop a set of business requirements and implement a database that fulfills those requirements.

1. To understand the design and development issues regarding databases and enterprise database management.
2. To convert a set of requirements into an effective database structure.
3. To obtain a strong conceptual foundation of the underpinnings of database design and enterprise database management.
4. To implement a database using some commercial database management systems, such as using SQL within Oracle.
5. To communicate effectively through oral presentations and written documents.

Course Grade Components:

- Class Participation: 10 points
- Homework Assignments: 10 points
- Midterm Exam: 40 points
- Final Exam: 40 points
- Database Project (optional): 10 bonus points

Grading policy:

- Overall course score >=90: A
- Overall course score >=85: B+
- Overall course score >=80: B
- Overall course score >=75: C+
- Overall course score >=70: C
- Overall course score >=60: D
- Overall course score <60: F

Class Project:

As part of your grade for this class, you will be invited to design and implement your own database in the environmental data management domain. This project will be done in pairs. I would like every group to pick a different topic, but there will be little overlap in the final project, so multiple groups can pick the same topic. Your grade will be based on the following criteria:

15% - database specification (consisting of a set of business rules)
25% - Entity Relationship Diagram in Crow's Foot Format
20% - Conversion of ERD to database design
30% - Database implementation and management reports
10% - Presentation (last day of class).
+10% - if you develop a usable system interface that allows the user to input data and run reports. You can use Access, a Windows EXE file, or web pages to carry this out.

Sample Data Specification for a library system (you cannot use this for your project!):

1. The library records information about all documents that are available in its system. Each document is identified by a unique number (*DocumentId*.) It also has a title, a publisher, a publication date, and can be one of three different types: book, journal volume, or conference proceedings.
2. For each document type we need to store additional information. For instance, for books we need to record their ISBN, and for conference proceedings we need to record the date and location of the conference, and the proceedings chair. A journal can have several volumes. The scope and editor(s) of a journal need to be recorded. Every volume of a journal has a number and may have a volume editor who also needs to be recorded.
3. Each document has a single publisher and the publisher address is also recorded.
4. Books have authors. Information about the authors of books is maintained. An author is identified by a number (*AuthorId.)* The name of each author is also recorded.
5. The library system contains several branches, which are identified by a number *(LibId).* We also need to store the name and the location of each branch. Each branch of the library holds a number of copies of a particular document. Each copy of the same document kept by the same library branch is numbered from 1 to n. The total number of copies of each document in the library is needed. We also need to record the position of every copy in a branch. It is encoded by a string of 6 alphanumeric characters (e.g. 001A03 means the third shelve of bookcase A03).

6. The library system keeps track of all readers who are uniquely identified by *ReaderId*. Each reader has a name, an address, a phone number and a type ("student", "senior citizen", "staff" etc.) A reader has to be registered in the database before borrowing a document.
7. Readers have access to the online catalogue of documents and may reserve books by title if they are available. A reserved book has to be picked up before 6 pm; otherwise, the reservation is cancelled. A reader cannot borrow or reserve more than 10 documents. We'll discuss different ways to handle these requiremets.
8. Borrowing is defined as taking out a copy of a document on one date and time (*BDateTime*) and returning it a maximum of 20 days later. *RDateTime* is the date and time on which the copy of the borrowed document is actually returned. (RDateTime is NULL if the document has not yet been returned). Books have to be returned to the branch from which they are borrowed.
9. The same copy of a document can be reserved and/or borrowed by the same reader several times.
10. Documents that are not returned on time are fined at a rate of 20 cents for each day after the due date.
11. A copy of a document cannot be lent to more than one reader at a time, but a reader can borrow multiple copies of documents.

Requirements for Deliverable 1

Deliverable 1 must contain an analysis of the intended database system, and a conceptual (Entity Relationship Diagram – ERD in Crows Foot Format) schema that includes:

- entity types, relationship types (including class/subclass relationship types), and attributes.
- key attributes.
- structural constraints (cardinality ratios and participation constraints).

Requirements for Deliverable 2

Phase 2 Deliverable must contain the goal of this phase of the project, and a logical design of the database (resulting by the mapping of an extended ER (EER) schema to a Relational schema) and the DDL to create the database. It must describe any revisions made to the specification described in Phase 1 Deliverable. You should:

- Translate the EER diagram to a relational schema. This translation should identify primary and foreign keys.
- Include constraints (in words) over and above referential integrity constraints for each table. E.g., "a reader cannot borrow more than 10 documents" is a constraint on the BORROWS table and must be checked before a document is lent to a reader.
- Submit the DDL to create all the appropriate tables with the above keys and constraints

Requirements for Deliverable 3

Phase 3 Deliverable must contain the goal of this phase of the project, and a description of the creation of the database schema and instance and of the application programs. It must also provide any revisions made to the specifications described in Phase 2 Deliverable. It must further describe the problems encountered and justify the solutions.

- The program must run against the data that will be provided after Phase 2 is submitted.
- The database must have ample sample data in its tables so that one can sufficiently perform and illustrate all required tasks.

1. Run SQL command files that populate each table or import the data from a spreadsheet that I will provide. You should be able to run these command files successfully with no errors and no integrity violations. You should also design your DDL so that the script can run multiple times without manual intervention in case you need to recreate your database or modify it due to design flaws.
2. You are to develop the SQL statements to complete the functions described in the requirements
3. Provide a spreadsheet that can access the database you created via an ODBC, pull relevant data for reporting into a sheet in the spreadsheet using the data refresh function, and create pivot table showing relevant management reports.

Requirement for Mid-term Project Deliverable

This mini-project is a database design and redesign task. Each team is asked to find a unique set of environmental monitoring data released by government agencies or nongovernment organizations. The data domain should impact human health conditions. Examples include water and soil pollution monitoring data released by US Environmental Protection Agency (www.epa.gov), chemical and toxical plants across the nation released by US Department of Defense (www.defense.gov) and Department of Energy (www.doe.gov), as well as U.S. Census Bureau (www.census.gov), to name a few. The only environmental dataset that you CANNOT use in your project is the air quality index dataset released by EPA because it has been carefully organized and made available to the public by EPA.

I strongly encourage teams to work on different datasets. In case that multiple teams work on the same dataset, gradings will be assigned in high favor towards the best performing team. The second performing team will suffer significantly in the grade. Therefore, in the interest of maximizing your own grades, I advise you avoid such head-to-head competition to every possibility. Some example datasets are as follows.

1) Airborne chemical emissions: EPA's NEI (National Emission Inventory) database will be used which provide estimates of annual emissions of criteria and hazardous air pollutants from all types of sources.

2) Erometric Information Retrieval System (AIRS) /AIRS Facility Subsystem (AFS): Additional information on air releases will be obtained on AIRS, a computer-based repository for information about air pollution in the United States. This information comes from source reports by various stationary sources of air pollution, such as electric power plants, steel mills, factories, and universities, and provides information about the air pollutants they produce.

3) National Drinking Water Contaminant Occurrence Database (NCOD): This will be used as indicators of drinking water exposure. The NCOD contains occurrence data from both Public Water Systems (PWSs) and other sources (like the U.S. Geological Survey National Water Information System) on physical, chemical, microbial and radiological contaminants for both detections and non-detects.

4) STORET and WQX databases: These are EPA's repository and framework for sharing water monitoring data. The STORET Data Warehouse is a repository for water quality, biological, and physical data.

5, 6) Additional two database Permit Compliance System (PCS) and Safe Drinking Water Information System on water quality: PCS provides information on companies which have been issued permits to discharge waste water into rivers including information on when a permit was issued and expires, how much the company is permitted to discharge, and the actual monitoring data showing what the company has discharged. SDWIS contains information about public water systems and their violations of EPA's drinking water regulations.

7) RadNet, formerly Environmental Radiation Ambient Monitoring System: The RadNet is a national network of monitoring stations that regularly collect air, precipitation, drinking water, and milk samples for analysis of radioactivity. The RadNet network has been used to track environmental releases resulting from nuclear emergencies and to provide baseline data during routine conditions. In addition, EPA's indoor radon level database will be used that is particularly useful to the risk analysis on lung cancer.

8) Superfund sites.

Executing the project involves performing the following four task steps.

Step 1. You are expected to download the raw datasets from one or multiple places of your choice.

Step 2. After examining the data characteristics exhibited by the dataset, you are asked to propose a data model suitable for capturing the data elements and their relationships.

Step 3. Implement your proposed data model through (re-)organizing your downloaded data using a DBMS of your familiarity. NJIT software library offers licenses of multiple DBMS products that you can get for free as a registered student. MS SQL server is recommended but not required.

Step 4. Create a brief project report (no length requirement imposed) addressing the following grading criteria: 1) the size of the datasets you manage to (re-)organize in your project; 2) the clarity and logic soundness of the database model design you propose for the data management problem, including the business rules you choose to model and implement in your database design; 3) how you handle missing and unexpected data elements in the dataset; 4) the data querying performance you can accomplish via using an DBMS of your choice over your established dataset.

You are expected to submit the project report, the raw datasets you download as well as the complete copy of your established database to Moodle by March 28, 2016, 6pm. In the evening of March 28, each team is expected to deliver a 5-10 minutes' brief presentation regarding your project outcomes and achievements. Live demos of your organized database and the query results will be a strong plus. All presentation materials, other than your live demo program(s), need to be uploaded to Moodle before the lecture start time.

The size of a team should be between 1 – 3 people. Grading will be assigned according to the quality of work completed and the number of people working in the team. There will be no

favoritism against or towards smaller or larger teams. If you are not sure about the suitability of your project topic, please double confirm with me through email. To coordinate between the teams so that no particular dataset will be used multiple times between multiple teams, you are encouraged to use Moodle's forum function to chat and self-organize. In case more than one team works on the same dataset, grading will favor the team that delivers the most impressive project outcome. To reduce potential competition between teams, you are strongly recommended not to work on common datasets.

There will be a special project assignment due on 6pm, April 4, 2016. Every student is required to criticize three other projects according to the presented content on the evening of March 28. Comments should be developed according to the grading criteria specified in Step 4 of the project requirement, which are highlighted. Even though there is no length requirement for the submitted comments, detailed comments will be encouraged.

Requirement for Final Project Deliverable

The goal of your project is to generate the cumulative environmental exposure for each person. The expected system output should include: 1) all the environmental monitoring records used in your estimation for each individual person's cumulative exposure, and 2) the cumulated exposure for the person. To derive the cumulative exposure, you need to find for any given year of a person's residence, the closest several pollution monitoring points in your database and then do a distance-based weighted average. (http://en.wikipedia.org/wiki/Inverse_distance_weighting)

Also, in your final project report, you need to manually calculate exposure results for a handful of people as case examples. Please include all steps of your manual calculation so that I can verify your derivation process. Of course, we expect these results should match the automatic output produced by your program. I will announce the human subject ID #s for these people 48 hours before the project due time. Therefore, you are recommended to ensure that your program will run smoothly for all the announced human subjects well before the project due date.

The grading criteria will be based upon:

1) the number of years of environmental monitoring data covered in your database;

2) the number of environmental monitoring elements covered in your database;

3) the complexity and suitableness of the relationships modeled in your database;

3) the correctness of your calculation results;

4) the richness of your cumulative exposure derivation mechanism, mainly the amount of environmental monitoring data points your program would retrieve to derive the cumulative exposure output of each human subject;

5) the computational efficiency of your program in deriving the exposure data.

In your final project report, please discuss your work outcome following the above five aspects. The report does not need to be long. Clearly presented design logics with good performance numbers should be sufficient to receive a score of excellence. A lengthy report will poor performance numbers will not help. I may also randomly run your submitted program to verify the performance numbers you reported. Therefore, please thoroughly discuss in your project report how to configure and run your program by a third-party person.

Our Strict Policy on Collaboration/Cheating:

Every assignment/project is to be regarded as an examination. The NJIT Honor Code will be upheld. A description of the NJIT Honor Code is available for your review at http://www.njit.edu/academics/honorcode.php. Students found cheating or collaborating or plagiarizing will be **immediately** referred to the Dean of Students and the NJIT Committee on Professional Conduct and subject to Disciplinary Probation, a permanent negative marking on their record, **possible dismissal and a definite grade of 'F' in the course. All submitted assignments are carefully checked for similarities, and plagiarism and guilty students will be identified. This also includes use of instructor materials no matter how they were provided to you.**

Policy on Submission of Assignments/Projects : The format of submission will be announced with each assignment/project. Assignments and projects are to be posted in Moodle.

Our Strict Policy on Lateness of Submission: Every assignment/project will have a due date, and all submissions **are expected to be** made by this due date. **Assignments submitted after the due date will not be accepted regardless of any reason you might have.**

Below are the TOPICs covered in the course and the related TEXTBOOK readings. Remember one of the keys to success in IS631 is your own self-discipline - your goal should be to maintain currency each week, and NEVER fall behind! (*Note: this is a very tentative schedule, and I reserve the privilege to modify and edit these topics and textbook readings for the benefit of the course.)*

Week #	TOPIC
0	Please read chapters 1 & 2 prior to the first class meeting
1	Introduction to SQL
2	Intermediate SQL
3	Advanced SQL
4	Formal Relational Query Languages
5	Database Design: The Entity-Relationship Approach
6	Relational Database Design

7	Midterm Exam
8	Storage and File Structure
9	Indexing and Hashing
10	Query Processing
11	Query Optimization
12	Transactions
13	Concurrency Control
14	Recovery System
15	Comprehensive Final Examination

Note: 1. Details of the mid-term and the final exam will be announced later.

2. The syllabus may be changed to be adjusted to provide better educational services. In such a case, the changes will be announced in advance.