INFO 6210
Data Management and Database Design
Spring 2018 Course Syllabus

Course Information

Professor: Nik Bear Brown
Email: nikbearbrown@gmail.com
Office: 328 Dana Hall
Office hours:
Monday 10-11AM
Tuesday 10-11AM
Wednesday 10-11AM
Thursday 10-11AM
Friday 10-11AM

Course website: Blackboard (for raw scores, uploading assignments, getting materials, & forums)
Piazza: https://piazza.com/northeastern/spring2018/info6210

Course Prerequisites

Engineering students only.

Course Description

Studies design of information systems from a data perspective for engineering and business applications; data modeling, including entity-relationship (E-R) and object approaches; user-centric information requirements and data sharing; fundamental concepts of database management systems (DBMS) and their applications; alternative data models, with emphasis on relational design; SQL; data normalization; data-driven application design for personal computer, server-based, enterprisewide, and Internet databases; and distributed data applications.

Communication

Communication between instructor and students is through
- Email via the Blackboard distribution list
- Announcements posted on Blackboard
- Notes posted on the Blackboard discussion board
- Private email exchanges

Course Structure

- Regularly test students on paper/algorithmic exercises
- Evaluate students' implementation competency, using assignments that require coding on given datasets

Copyright © 2017 Northeastern University. All Rights Reserved

- Evaluate ability to setup data, code, and execute using python language
- Student will be required to do “data digging”: run analysis scripts and failure analysis
- Final project is typically asking and answering a “real world” question of interest using machine learning techniques

Learning Objectives

By the end of this course, students should be able to do the following:

Theory

Understand the Entity Relationship Model (ERM)
Understand Relational Algebra
Understand Relational Calculus

Conceptual knowledge

• Understand Relational Databases
• Understand SQL (SQL, NULL, integrity constraints, views, SQL Integrity constraints, outer join, SQL functions, user-defined aggregates, triggers)
• Understand SQL transactions
• Understand Storage and Indexing
• Understand Query Optimization
• Understand Normal Forms
• Understand NoSQL
• Understand Graph Databases
• Understand Map-Reduce
• Understand Apache Spark Dataframes

Practical experience

• Collect clean and munge real-world data
• Store, search and display collected data.

Course GitHub

The course GitHub (for all lectures, assignments and projects):

https://github.com/nikbearbrown/NEU_COE

nikbearbrown YouTube channel

Over the course of the semester I’ll be making and putting additional data science and machine learning related video’s on my YouTube channel.

https://www.youtube.com/user/nikbearbrown

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Copyright © 2017 Northeastern University. All Rights Reserved

Schedule

Week	Topic	Assignments
1) Week 1	Entity Relationship Model (ERM) AWS Google Cloud	Readings; Assignment 1
2) Week 2	Database Project Data gathering, cleaning and munging	Readings
3) Week 3	Relational Algebra Relational Calculus SQL	Readings; Assignment 2
4) Week 4	SQL Normal Forms Normal Forms; Recovery Keys Indexing Views	Readings; Project proposal
5) Week 5	SQL Triggers Transactions Query Optimization	Readings
6) Week 6	Accessing a database from code Backing up data	Exam 1
7) Week 7	NoSQL MongoDB	Readings; Assignment 3
8) Week 8	MapReduce & Hadoop Introduction to the Hadoop Stack Introduction to Hadoop Distributed File System (HDFS)	Readings; Project progress report.
9) Week 9	Introduction to Map/Reduce	Readings; Assignment 4
10) Week 10	Apache Spark Apache Spark SQL and Dataframes	Readings; One programming assignment. (Your choice of lessons 2 or 3)
11) Week 11	Apache Spark MLib Apache Spark Graphframes	Readings; Assignment 5
12) Week 12	Break	Spring recess
13) Week 13	Graph Databases Neo4j	Readings; Assignment 6
14) Week 14	Research Project Presentations	Readings; A draft of the final project for feedback

Copyright © 2017 Northeastern University. All Rights Reserved

Finals	Exam 2	

Teaching assistants

The Teaching assistants for Spring 2018 are:

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

- Exam
- Completion of assignments.
- Completion of a database project asking and answering a "real world" questions.

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown nikbearbrown@gmail.com. In an online course, it's important that a student reaches out early should he/she run into any issues.

Grading Policies

Students are evaluated based on their performance on assignments, performance on exams, and both the execution and presentation of a final project. If a particular grade is required in this class to satisfy any external criteria—including, but not limited to, employment opportunities, visa maintenance, scholarships, and financial aid—it is the student's responsibility to earn that grade by working consistently throughout the semester. Grades will not be changed based on student need, nor will extra credit opportunities be provided to an individual student without being made available to the entire class.

Grading Rubric:

The following breakdown will be used for determining the final course grade:

Assignment	Percent of Total Grade
Exam	10%
Assignments	50%
Database Project One*	15%
Database Project Two*	15%
Portfolio	10%

Copyright © 2017 Northeastern University. All Rights Reserved

* Note that the assignments and drafts related to the research project rather than the programming assignments. I expect to use the following grading scale at the end of the semester. You should not expect a curve to be applied; but I reserve the right to use one.

Score	Grade
93 – 100	A
90 – 92	A-
88 – 89	B+
83 – 87	B
80 – 82	B-
78 – 79	C+
73 – 77	C
70 – 72	C-
60 – 69	D
<60	F

Scores in-between grades. For example, 82.5 or 92.3 will be decided based on the exams.

Blackboard:

You will submit your assignments via Blackboard. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard represents only the raw scores. Not normalized or curved grades. An .Rmd file ALONG with either a .DOC or .PDF rendering of that .Rmd file must be submitted with each assignment.

Due dates

Due dates for assignments are usually every other Monday at midnight.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Only ONE extension will be granted per semester.

Course Materials

Required text (All free online)

Some textbooks are all available for free to NEU students via SpringerLink (http://link.Springer.com/). You must access SpringerLink from an NEU IP address to have full access and/or download these books.

Copyright © 2017 Northeastern University. All Rights Reserved

If you are off-campus, in order to access resources provided through the Northeastern library outside the network, you should use their bookmarklet to load any page through the proxy:
http://library.northeastern.edu/bookmarklet

Required Texts

The *required* textbooks we will be using in this class are:

Database Systems A Pragmatic Approach (2014) (This will be the primary book)
Authors: Elvis C. Foster, Shripad V. Godbole
ISBN: 978-1-4842-0878-6 (Print) 978-1-4842-0877-9 (Online)
http://link.springer.com/book/10.1007/978-1-4842-0877-9

Principles of Distributed Database Systems, Third Edition (2011)
Authors: M. Tamer Özsu, Patrick Valduriez
ISBN: 978-1-4419-8833-1 (Print) 978-1-4419-8834-8 (Online)
http://link.springer.com/book/10.1007/978-1-4419-8834-8

Beginning Database Design From Novice to Professional (2012)
Authors: Clare Churcher
ISBN: 978-1-4302-4209-3 (Print) 978-1-4302-4210-9 (Online)
http://link.springer.com/book/10.1007/978-1-4302-4210-9

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Pro Hadoop Data Analytics
Designing and Building Big Data Systems using the Hadoop Ecosystem
Authors: Kerry Koitzsch 2017
ISBN: 978-1-4842-1909-6 (Print) 978-1-4842-1910-2
https://link.springer.com/book/10.1007/978-1-4842-1910-2

Pro Apache Hadoop
Authors: Sameer Wadkar, Madhu Siddalingaiah 2014
ISBN: 978-1-4302-4863-7 (Print) 978-1-4302-4864-4
https://link.springer.com/book/10.1007/978-1-4302-4864-4

Pro Spark Streaming
The Zen of Real-Time Analytics Using Apache Spark
Authors: Zubair Nabi 2016
ISBN: 978-1-4842-1480-0 (Print) 978-1-4842-1479-4
https://link.springer.com/book/10.1007/978-1-4842-1479-4

Copyright © 2017 Northeastern University. All Rights Reserved

Beginning Neo4j (2015)
Authors: Chris Kemper
ISBN: 978-1-4842-1228-8 (Print) 978-1-4842-1227-1 (Online)
http://link.springer.com/book/10.1007/978-1-4842-122

Recommended Texts

Big Data Made Easy
A Working Guide to the Complete Hadoop Toolset
Authors: Michael Frampton 2015
ISBN: 978-1-4842-0095-7 (Print) 978-1-4842-0094-0
https://link.springer.com/book/10.1007/978-1-4842-0094-0

The Definitive Guide to SQLite (2010)
Authors: Grant Allen, Mike Owens
ISBN: 978-1-4302-3225-4 (Print) 978-1-4302-3226-1 (Online)
http://link.springer.com/book/10.1007/978-1-4302-3226-1

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Beginning CouchDB (2009)
Authors: Joe Lennon
ISBN: 978-1-4302-7237-3 (Print) 978-1-4302-7236-6 (Online)
http://link.springer.com/book/10.1007/978-1-4302-7236-6

Beginning Neo4j (2015)
Authors: Chris Kemper
ISBN: 978-1-4842-1228-8 (Print) 978-1-4842-1227-1 (Online)
http://link.springer.com/book/10.1007/978-1-4842-122

Software

PostgreSQL (http://www.postgresql.org/)
Windows installers - Mac OS X - Linux downloads (http://www.postgresql.org/download/)

sqldf and SQLite (Relational databases)

- sqldf https://cran.r-project.org/web/packages/sqldf/sqldf.pdf
- SQLite https://www.sqlite.org/download.html

Copyright © 2017 Northeastern University. All Rights Reserved

Riak, Redis, and HBase (NoSQl databases)

- Riak http://docs.basho.com/riak/kv/2.1.4/downloads/
- Redis http://redis.io/download
- HBase https://hbase.apache.org/

MongoDB and CouchDB (NoSQl document databases)

- MongoDB https://www.mongodb.com
- CouchDB http://couchdb.apache.org/

Neo4J (graph database)

- Neo4J https://neo4j.com/download/

python Anaconda

- https://www.continuum.io/anaconda-overview

R (Statisical programming language)

- R project https://www.r-project.org/

RStudio (IDE)

- RStudio https://www.rstudio.com/products/rstudio/download3/

Python Tutorials

Dive into Python http://diveintopython.org

Python 101 – Beginning Python http://www.rexx.com/~dkuhlman/python_101/python_101.html

The Official Python Tutorial http://www.python.org/doc/current/tut/tut.html

The Python Quick Reference http://rgruet.free.fr/PQR2.3.html

Python Fundamentals Training – Classes http://www.youtube.com/watch?v=rKzZEtxIX14

Python 2.7 Tutorial Derek Banas· http://www.youtube.com/watch?v=UQi-L-_chcc

Python Programming Tutorial - thenewboston http://www.youtube.com/watch?v=4Mf0h3HphEA

Google Python Class http://www.youtube.com/watch?v=tKTZoB2Vjuk

Nice free CS/python book https://www.cs.hmc.edu/csforall/index.html

datacamp.com https://www.datacamp.com/tracks/python-developer

Copyright © 2017 Northeastern University. All Rights Reserved

R Tutorials

LearnR
https://youtu.be/p3i7Kz6C_-4?list=PLFAYD0dt5xCwDNFdrqeNoU9t-nhAWkbKe

Try python @codeschool: http://tryr.codeschool.com

Datacamp python Tutorials
https://www.datacamp.com/

rstudio online learning
https://www.rstudio.com/online-learning/

Participation Policy

Participation in discussions is an important aspect on the class. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the midterms. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source is not allowed. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Copyright © 2017 Northeastern University. All Rights Reserved

Assignments are due by 11:59pm on the due date marked on the schedule. Late assignments will receive a 5% deduction per day that they are late, including weekend days. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Only ONE extension will be granted per semester.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Northeastern University Copyright Statement: This course material is copyrighted and all rights are reserved by Northeastern University. No part of this course material may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual, or otherwise, without the express prior written permission of the University.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline,

Copyright © 2017 Northeastern University. All Rights Reserved

become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.

Copyright © 2017 Northeastern University. All Rights Reserved