Data Warehousing / OLAP for Analytics

Fall 2015
DNSC / ISTM 6211, Section 11
Tuesdays, 7:10 - 9:40pm, SMPA 309

Daniel Chudnov, daniel.chudnov@gmail.com
Office (9-5): Gelman 606, (202) 994-0684
Office hours: Decision Sciences Department, Thursdays 4-7pm, (202) 556-3282
Teaching Assistant: Bohan Zhang, bohanzhang@gwmail.gwu.edu

Description

This course provides a practical grounding in relational databases with a focus on data warehousing and dimensional modeling, along with hands-on experience in these tools and other traditional and contemporary methods for managing and analyzing data at scale, such as the Unix command line and Apache Spark. We will focus on using these tools for the middle phases of data analysis: wrangling, exploring, and modeling, with an emphasis on delivering reproducible data analyses. This course is complementary to other foundational courses in the Business Analytics program; as such, topics and techniques from Statistics, Programming, Data Mining, and Optimization may be present as use cases,
but will not be a focal point for grading.

Learning Objectives

- Develop practical experience implementing and working with relational databases, data modeling, and dimensional modeling for data warehouses
- Gain skill in wrangling and exploring data with a variety of tools inside and outside of databases
- Understand and be able to develop and deliver reproducible data analyses

Readings

Required: Adamson, Christopher. Star Schema: The Complete Reference. New York: McGraw Hill, 2010.

Highly recommended: Janssens, Jeroen. Data Science at the Command Line. Sebastopol: O'Reilly, 2015.

Optional: Silverschatz, Korth, Sudarshan. Database System Concepts, 5th or 6th edition. Boston: McGraw Hill, 2006 or 2010. (Slides available: http://codex.cs.yale.edu/avi/db-book/)

Optional: Beaulieu, Alan. Learning SQL, 2nd edition. Sebastopol: O'Reilly, 2005.

Please plan to complete 50-100 pages of required readings assigned each week. In addition to these titles, other readings will be assigned, primarily using resources available for free over the web. All readings will be marked as Required, Recommended, or Optional. You are expected to complete all Required readings prior to the week for which they are assigned; lectures, discussion material, and assignments will draw heavily from these and from the Recommended readings.

Software

Our computing environment is Unix, specifically Ubuntu GNU/Linux 14.04. A virtual image based on the Data Science Toolbox (datasciencetoolbox.org) and including support for Jupyter Notebooks, Python, R, MySQL, PostgreSQL, and Apache Spark is provided through instructions available at (tinyurl.com/dbplus-vm). This is **the same VM** to be used in Professor Kanungo's Programming course; you only need to install it once.

If you have an OS X or Linux machine of your own, you may be able to install, configure, and use some or all of these tools for yourself. If you have a Windows machine, several of these pieces will likely not work in the native environment.

Using the provided virtual machine is highly recommended. It is the only option supported by your instructor.

Lectures

This is an on-campus, in-person course. Students are expected to attend all lectures.

Each class session will include a discussion of assigned readings, a demonstration / workshop of student work, a lecture component, and often some additional time set aside to begin work on assignments.

A number of guest lecturers are invited to visit and discuss their own data tools and workflows; the schedule for these visits is not yet set.

Grading

- 10% - Participation (discussion, demos, acknowledged assistance, etc.)
- 50% - Assignments (8-10 total)
- 10% - Midterm exam (early November, in class)
- 10% - Book review (due October, November)
- 20% - Final project (groups of 1-3 people)

Participation: you are expected to engage in class discussion, share thoughts on readings and assignments, demonstrate your work at times, and offer constructive feedback about your peers' work.

Assignments: problem sets assigned weekly provide an opportunity to practice new skills; most will require the development of a Jupyter notebook demonstrating solutions to problems. For each assignment, document the steps you took in your work clearly, identifying tool dependencies and assumptions along the way, so that the work may be reproducible. In this way you will gain expertise in documenting your technical work while also developing a narrative voice appropriate to data analysis. Each assignment should center around one executed notebook with inline output, packaged together with all ancillary scripts developed to support the notebook, in a single zip file with both the PDF export and .ipynb source of the notebook. All weekly assignments are due and must be turned in by 7pm on the following Tuesday, prior to the start of class, unless otherwise noted. Late assignments will be subject to a 10% penalty per day.

You must submit your own original work; see **Assistance** below for details.

If you have questions or require clarification from the instructor or TA about specific assignments, please post your questions to the discussion board for that assignment in Blackboard. This allows other students who might have similar questions to see and review what questions have been asked already. If we receive such questions via private email, we will refer you to the discussion boards instead.

Midterm exam: a one-hour, closed book/laptop essay exam will review your understanding of key concepts learned to date.

Book review: you are expected to read and review at least one book to supplement class assignments and dig deeper into a subject of your choice, such as using the command line, relational database or data warehouse design and theory, or distributed or noSQL database tools. A list of suggested titles will be provided in early September; a 300-500 word written review and five-minute presentation will be due between late September and late November.

Final project: select a substantial (at least 200,000 records) transactional dataset, scrub it, model it with a relational design, transform it into a form suitable for analysis, and prepare a notebook describing your process and exploring the transformed data, providing several descriptive statistics and basic visualizations. Use your tools of choice from among what we have studied together. A five-minute presentation and 15-25 page notebook writeup from each group will be due in December.

Absences

This is an on-campus, in-person class; you are expected to be on time and present for each class session. If you must miss any part of any class session due to religious observance, illness, or other extenuating circumstances, let me know in advance. Any absence not arranged ahead of time will count against participation score.

Assistance

Except for the book review and midterm exam, which you must complete by yourselves, you are encouraged to seek assistance from and to offer assistance to your peers. This may come in the form of reviewing each other's work, study or review sessions, pair programming, debugging, or discussion and documentation of tips and tricks on the class discussion board or on Github. We will do some of these reviews in class together. Even so, you are each required to turn in your own original work. Given the nature of the work for this course, duplication should be easy to spot. Whenever you receive assistance, **acknowledge it explicitly** in your writeup, naming those who provided you with assistance and the manner of assistance they provided. This is both good professional practice and good professional courtesy. The contributions of those named in acknowledgements will count toward their respective participation scores.

Ground Rules

It is our mutual obligation to ensure our classroom is a welcoming place for everyone participating in this class.

Silence all your devices before class begins. If you must use them, be discreet, be brief, and do not distract or annoy your classmates.

Take responsibility for the quality of discussions.

Listen to each other attentively; do not interrupt, and do not monopolize discussion.

Ask for clarification if you are confused.

Be especially thoughtful when guests join us; they are offering their time, so please close your laptops, put down your phones, and give them your full attention.