Data Management for Analytics

Fall 2016
ISTM 6212, Section 10
Tuesdays, 7:10 - 9:40pm, SMPA 309

Instructor: Daniel Chudnov, daniel.chudnov@gmail.com
Office hours: GWSB Decision Sciences Dept., Tuesdays 4-7pm, (202) 556-3282
Email availability: Sunday mornings through Friday afternoons
GitHub: https://github.com/gwsb-istm-6212-fall-2016

Description

This course provides a practical grounding in managing data for analytics work. We emphasize traditional and contemporary tools for data wrangling, databases, and data warehousing, with a focus on schema design and dimensional modeling, and with substantial hands-on experience using these tools and other contemporary methods for managing and analyzing data at scale. 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 theoretical foundation and practical experience working with a variety of traditional and contemporary data management tools, enabling students to work productively with any product or toolkit they might encounter
- Gain skill in wrangling and exploring data with a variety of tools inside and outside of databases
- Understand and be able to develop, deliver, and review reproducible data analyses

Readings

Required: Adamson, Christopher. Star Schema: The Complete Reference. New York: McGraw Hill, 2010.
Recommended: Janssens, Jeroen. Data Science at the Command Line. Sebastopol: O'Reilly, 2015.
Recommended: Beaulieu, Alan. Learning SQL, 2nd edition. Sebastopol: O'Reilly, 2009. Available online at GWU: http://findit.library.gwu.edu/item/11839919

Please plan to complete 25-50 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 class session the following

week; lectures, discussion material, and projects will draw heavily from these and from the Recommended readings.

Software

Our primary computing environment is Jupyter (http://jupyter.org/), a web-based notebook system ideal for data wrangling and analysis. A server configured to support Jupyter notebooks for class lectures, exercises, and projects is available at http://datanotebook.org/. There are several important things to know about this server:

- It hosts notebooks temporarily. After a certain amount of idle time, the server will reset and your notebooks will be removed. Save your work often by downloading it to your local machine.
- It has limited capacity. It is there for you, for use in this class. Please do not share the URL outside of class - if too many people use it, it might become unavailable to you.
- Please contact your instructor if you are unable to use the server, and please explain the problem you are seeing in detail.

You can install and run Jupyter notebooks on your own machine, most easily using Anaconda (https://www.continuum.io/downloads), although some extra software installed at datanotebook.org might be hard to replicate on your own. Some assignments will require this extra software. **Using datanotebook.org is highly recommended. It is the only option supported by your instructor**.

An additional environment we will use is GitHub (https://github.com/), where lectures and assignments will be posted. You will be expected to upload your assignments to GitHub and to review your peers' work using GitHub. It will also serve as a version control system for you and a remote backup of your work.

We will likely use additional tools, mostly hosted and available through the web.

Lectures

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

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

Grading

- 30% - Exercises (6-8 total)
- 50% - Projects (4-6 total)
- 10% - Project reviews (two per project)
- 10% - Participation (discussion, demos, acknowledged assistance, etc.)

Exercises: short sets of exercises assigned regularly provide an opportunity to review and practice new concepts learned in lectures and in readings. Time will be provided during class to begin exercises, and each will be due on Friday mornings. If you stay on top of readings and follow lectures closely, the exercises should not require more than 1-2 hours to finish. Late exercises will be subject to a 10% penalty per day.

Projects: problem sets provide an opportunity to practice and integrate new skills. For each project, describe and document the steps you took in your work clearly, explaining your assumptions and identifying tool dependencies 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 project should center around one executed notebook with inline output, packaged together with any ancillary scripts developed to support the notebook, in a single folder including the .ipynb source of the notebook. All projects are due and must be turned in by 12pm on the following Tuesday, prior to the start of class, unless otherwise noted. Late projects will be subject to a 10% penalty per day.

Project reviews: each student will be asked to review other students' work on projects, reproduce their code, and offer direct, constructive feedback. Be supportive of each other -- this is an opportunity to get used to having your data work reviewed by peers, and to be both the reviewer and the reviewee. Reviews will be assigned following project deadlines and will be due before the start of the following Tuesday's lecture.

Final project: working in groups of up to four students, select a substantial (at least 250,000 records) 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.

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

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

If you have questions or require clarification from the instructor 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. Most questions about assignments sent via private email will be referred to the discussion boards instead.

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 your instructor know in advance. Any absence not arranged ahead of time will count against participation score.

Assistance

Although you must complete assignments 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 for all assignments unless directed otherwise. 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 should guests join us; they are offering their time, so please close your laptops, put down your phones, and give them your full attention.

It is easy to distract ourselves through apps, chat, news, sports, etc. Sometimes a quick response to a text is necessary or unavoidable; please be brief. Remember why you are in class, and work to keep your focus during our time together.

On Writing

One mark of a professional is the ability to communicate through clear, concise writing, especially with technical topics like those we cover in our course. The quality of your writing will be a factor in grading all assignments. Always use complete sentences, full punctuation, and formatting appropriate to your text.

If English is not your native or strongest language, consider this an opportunity to improve. Your instructor is looking for you to make progress, not for you to be perfect.

Plagiarism is never acceptable. Always write for yourself, in your own voice. Any submitted assignment with verified plagiarism may be marked down substantially or rejected; any repeat offenses will be reported.

Please familiarize yourself with the Code of Academic Integrity:
https://studentconduct.gwu.edu/code-academic-integrity

Recommended Reading

These titles are related to our course material and should give you plenty to dig into on any of the topics you want to learn more thoroughly on your own.

Unix / Linux

Gancarz, The UNIX Philosophy; Shotts, The Linux Command Line; Hyde, Write Great Code, Volume I: Understanding the Machine (online through GW Libraries at http://findit.library.gwu.edu/item/5966168), chapters 2-5

Databases

Silberschatz et al., Database System Concepts; Celko, SQL for Smarties; Date, Database in Depth; Hernandez, Database Design for Mere Mortals; Lukaszewski, MySQL for Python

Warehouses

Corr / Stagnitto, Agile Data Warehouse Design; Hughes, Agile Data Warehousing Project Management; Inmon, Building the Data Warehouse; Kimball, The Data Warehouse Toolkit

noSQL

Karau et al., Learning Spark: Lightning-Fast Big Data Analysis; Redmond, Seven Databases in Seven Weeks; Ryza et al., Advanced Analytics with Spark: Patterns for Learning from Data at Scale; Sadalage, Fowler, NoSQL Distilled; Sankar, Karau, Fast Data Processing with Spark

Writing

Dupre, BUGS in Writing; Greene, Writing Science in Plain English; Oliver, A Poetry Handbook; Zinsser, On Writing Well