ISTA 322| Data Engineering

Professor Dan Charbonneau
Email: dcharbonneau@arizona.edu
Online office/lab hours: Thursdays 2pm-3pm
Online Zoom office: https://arizona.zoom.us/j/87815157914
SLACK INVITE: https://join.slack.com/t/ista322fa247w-flo4590/shared_invite/zt-2sk34kvys-6323E36K3ybTuXvZw2P7hg
Prerequisites: ISTA 130 or CSC 110 or ECE 175 or professor consent

Course Description: This course will be inviting for a wide variety of students from across disciplines, and they will learn how to use industry standard tools and practices to make large data sets usable for scientists and other decision makers. From data collection and preparation, to the creation of big data stores, databases, or systems to make data flow, this course will focus on the practical work needed to prepare big data for analyses across contexts. Students will be introduced to a variety of technical tools for data management, storage, use, and manipulation.

Course Objectives: The objective of this course is to train students in the tools and theory behind data engineering so that they can be deployed in the real world. This will revolve around extracting data from static and streaming sources, learning how to transform it in a way to be used for later analytics/machine learning, and data science applications, storing and querying that data in a database. Additional emphasis will be placed on learning some cloud computing methods.

Expected Learning Outcomes
1. Students will be able to extract data from static and streaming data sources
2. Students will develop and apply skills for data cleaning, munging, and transformation so that they are usable for later analysis.
3. Students will be able to generate and query SQL databases and access said databases from a Python interface
4. Students will understand major cloud and distributed computing technologies and be able to leverage them for data engineering application.

Course Competencies:
This class addresses the following ISTA BS competencies:

F1.2 – Students will demonstrate facility using basic research methods, for example: research design; statistics and analysis; organization, identification, and location of data and information including open- and closed-access sources; and/or presentation of findings in oral, written and multi-media form, including proper use of and citation of sources.

DAISBS2.2 – Students will establish the ability to exercise the four key techniques of computational thinking (decomposition, pattern recognition, abstraction, and algorithms) in solving information and data challenges.

DAISBS2.3 – Students will acquire the skills of collecting, manipulating, and analyzing different types of data at different scales, and interpreting the results properly.

Required Readings:
We'll mainly be using free online resources. For additional content, I'd recommend Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems by Martin Kleppmann. This isn't required for the course, but rather a good source to dig a little deeper

Required Materials:
All students will need a functional laptop that can be brought to campus for each class.

Complete List of Assignments with Grade Breakdown:

Activity	Total Percent	Unit Percent	Activity & Notes
NB assignments (D2L quiz)	17.5%	2.5%	Total of 7 notebooks Each notebook will have an associated quiz (unlimited attempts possible) asking questions about the work you're doing in the NB. This is meant to help guide you and ensure that you understand the material in the NB.
HW assignments	30.5%	5% (FP@5.5%)	Total of 6 assignments (5 HW & 1 FP) Graded on correctness. I will be testing your code, so make sure that everything runs correctly *in situ* before uploading
Final Project	30%	-	This is a full, independent data engineering project that will be completed in the 2nd half of the class. You will present your project at the end of the semester (10 pts for presentation, 20 pts for project/code)
Lectures (playposit)	22%	2%	11 lectures (10 playposit and lecture 7 submitting screenshots to D2L)

Assignment submission
HW assignments will be google colab notebooks and should be submitted as follows:

1) First create a copy of the notebook in your drive and rename it to "ISTA_322_F23_assignmentname_firstname_lastname". (e.g. my copy of homework 1, which I'm submitting in Fall 2023 (F23) would be ISTA_322_F23_HW1_dan_charbonneau)

2) When you are ready to submit. Prepare three files: the python file (File->Download->Download .py), the notebook file (File->Download->Download .ipynb), and **PDF** version of your notebook (**after running all cells**).
Note: you can take a screenshot and create the pdf out of them.

3) Create the directory name firstname_lastname_assignmentname (e.g my directory would be dan_charbonneau_hw1) put all three files in it, ZIP, and submit in D2L.

Incorrect filenames or submission formats will result in a loss of 50% of your grade

Grade Distribution:
90-100% = A "exemplary, far beyond reqs/expectations"
80-89% = B "exceeds requirements/expectations"
70-79% = C "meets requirements/expectations"
60-69% = D "falls short of requirements/expectations"
< 60% = E "repeat of course needed"

Course Schedule:

Course schedule is subject to change.

Week	Date	Topic	Lectures	Due	Notebooks	HW	Due
1	17-Oct	What is data engineering	1 & 2	27-Oct	NB1		27-Oct
		What is data and how to work with it	3	27-Oct		HW1	27-Oct
2	28-Oct	Intro to data manipulation in python	4	3-Nov	NB2		3-Nov
		Jsons and semi-structured data			NB3	HW2	3-Nov
3	4-Nov	Intro to SQL	5	10-Nov	NB4		10-Nov
		More SQL	6	10-Nov	NB5	HW3	10-Nov
4	11-Nov	SQL loads	7	17-Nov			17-Nov
		DB normalization	8	17-Nov	NB6	HW4	17-Nov
5	18-Nov	ETL with real data	9	24-Nov		HW5	24-Nov
		Final project details				FP1	24-Nov
6	25-Nov	ETL with real data	10	1-Dec	NB7	FP2	1-Dec
		Free time to work on projects					
7	2-Dec	*Free time to work on projects*					
		Final project presentations (notebook and video)	11	11-Dec	Final project		11-Dec

Requirements for the Course:

The material in this course will rapidly build on itself. Thus, missing even one day may potentially set you behind dramatically. Thus, all students are expected to attend every class period. Students will be programming daily and thus need to bring their laptop with a functional python environment every day.

Attendance, Due Dates, and Missing Work:

1. **Missed class assignments or exams cannot be made up without a well-documented, verifiable, excuse (for example, a physician's medical excuse).**
2. All HW assignments (except HW5) can be resubmitted up to one week after the original due date for up to 90% of the grade.
3. All holidays or special events observed by organized religions will be honored for those students who show affiliation with that particular religion.
4. Absences pre-approved by the UA Dean of Students (or Dean designee) will be honored.
5. The UA's policy concerning Class Attendance and Administrative Drops is available at: https://catalog.arizona.edu/policy/class-attendance-participation-and-administrative-drop

Specific course plagiarism/cheating policies
Cheating and plagiarism is absolutely not tolerated in this class (including copy/pasting part or whole code from online sources or other students). See below for UA and iSchool codes of conduct for details.

In this class, any student found cheating will be reported to the dean of students and the following minimal sanctions will be recommended (possibly additional sanctions based on context):

- Any student found to be cheating will receive an automatic 0 for the entire assignment or exam (yes, a single block of plagiarized code an assignment gets you a 0 for that assignment).
- Multiple infractions (regardless of whether students were notified of the first infraction; ie if multiple infractions are identified at the same time) will result in an automatic failing grade for this class (with more severe sanction recommendations submitted to the dean's office, possibly including a note on your transcript)

Working with other students and troubleshooting together is perfectly fine, and in fact highly encouraged. But... ***all of the code you submit must be your own (ie working on an assignment together and copy/pasting the code you came up with together into both assignments is not acceptable)***

Course Conduct and Campus Policies (be familiar with all campus policies):

1. Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of independent effort unless otherwise instructed. Students are expected to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See: https://deanofstudents.arizona.edu/policies/code-academic-integrity.
2. Accessibility and Accommodations: At the University of Arizona, we strive to make learning experiences as accessible as possible. If you anticipate or experience barriers based on disability or pregnancy, please contact the Disability Resource Center (520-621-3268, https://drc.arizona.edu) to establish reasonable accommodations.
3. The Arizona Board of Regents' Student Code of Conduct, ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to one's self. See: http://policy.arizona.edu/threatening-behavior-students.
4. All student records will be managed and held confidentially. http://www.registrar.arizona.edu/ferpa/default.htm
5. Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.
6. UA Non-discrimination and Anti-harassment policy: http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy.
7. Confidentiality of Student Records: http://www.registrar.arizona.edu/ferpa/default.htm.

Academic Integrity in this course
Information contained in this syllabus, other than the grade and absence policy, may be subject to change without advance notice as deemed appropriate by the instructor.

This policy agreed upon by faculty in the UArizona iSchool applies in addition to the Dean of Students' Code of Academic Integrity.

Students in courses at the UArizona iSchool are expected to maintain rigor in their academic performance with intent to learn, practice, and overcome challenges toward personal growth and enrichment. As

future professionals in digital environments, iSchool students are also expected to exercise transparency and integrity in collaborations and in the use of tools and resources that may aid completion in assignments for our courses.

Consider the following PROHIBITED practices in this course, unless the instructor has specifically written instructions or permission to do otherwise:

- Posting a question on an online site such as Chegg.com, and copying and pasting some or all of the response into an assessment
- Posting an assessment from the course on online sharing sites such as Course Hero. Aiding other students in violation of academic integrity is also a violation, and is potential copyright infringement.
- Generating and submitting, in whole or in part, text or code through Artificial Intelligence such as ChatGPT, QuillBot, and text summarizers
- Using, in whole or in part, computer code not written by the student (for example, from another student, a book, or the internet) in an assignment or project. This includes using such code in modified or unmodified form.
- Searching for *solutions* to projects or assignments on the internet or through other tools, when your instructor intended for you to learn the solution through exercises (e.g. Googling for the solution to a question on an assignment).
- Simultaneously submitting the same assignment as another student enrolled into the course without prior permission from the instructor