ISTA 322| Data Engineering

Professor Nicholas DiRienzo
Email: ndirienzo@email.arizona.edu
Office Location: NA for Spring 2020
Online office/lab hours: Thursday 11-12:30
Online Zoom office: https://arizona.zoom.us/j/2251199844
SLACK INVITE: https://join.slack.com/t/newworkspace-3wj6593/shared_invite/zt-kpzyvfw1-pQmAEAuzQZklCt3XovLQkg
Prerequisites: ISTA 321 or equivalent machine learning class; ISTA 130 or equivalent python programming class
Final exam: Monday, May 10th from 9am-12pm or 2pm-5pm (you will pick a time that works)

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.

Required Readings:
We'll be using the book Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems by Martin Kleppmann. This will be coupled with other free online resources

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
Weekly assignments	42%	6%	There will be 7 total assignments
Final Project	20%	-	This is a full, independent data engineering project that will be completed in the final weeks of class.
Exams (2)	30%	15%	Two exams, both required

Slack participation	10%	-	Mix of in-class and message board participation. 2% from introduction. 8% from weekly participation. **You must interact/participate an average of once per week for full credit.**

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. This is only a tentative guideline of the weekly topics.

Broad weekly outline

1. Wednesday, January 13 - What is data engineering
    a. Lecture 1 - what is DE
    b. Lecture 2 - Where does data come from and what is it
    c. MAKING ACCOUNTS
2. Wednesday, January 20 - What is and how to work with it
    a. Lecture 3 - Data types and working with data
    b. Notebook 1 - Working with data
    c. Notebook 2 - Functions and loops
    d. HW1 - Python programming and data manipulation
        i. Due Tuesday January 26th
3. Wednesday January 27th - Intro to data manipulation in python
    a. Notebook 2 - Transform flat files - version with fill in parts for lesson
    b. Lecture 4 - Data transforms - slides, video
4. Wednesday, February 3 - Jsons and semi-structured data
    a. Notebook 3 - Flattening Spotify JSON data - completed version
    b. HW2 - Due Tuesday February, 9
    c. HW2 - completed
    d. HW2 - blank
5. Wednesday, February 10 - Intro to SQL
    a. Setting up movies DB
    b. Notebook 4 - Intro to SQL
    c. Lecture - RDBMS - Video here
6. Wednesday, February 17 - More SQL
    a. Ticket sales DB setup
    b. HW3 - Due Tuesday February 23
    c. HW3 - completed
    d. HW3 - blank
    e. Lecture - RDBMS
7. Wednesday, February 24th SQL loads
    a. Expand datatypes part of lesson
    b. Setting up database on AWS - lecture slides
    c. Notebook - test database connection
    d. HW4 - Due Tuesday, March 2nd

i. MAKE - Have have them write several loads

e. Review

8. Midterm - Friday, March 5th

a. Notebook for SQL

9. Monday, March 8th to Friday March 12th - Off week

10. Monday, March 15 - DB normalization & ETL on AWS
    a. Expand conceptual side and examples of normalization and also key specification in HW
    b. Simple lecture on normalization - slides, video
    c. Notebook - Simple ETL (S3 -> colab -> RDB on AWS) - goes with HW4
    d. Video of going through notebook
    e. HW4 - completed - etl - turn into HW5
    f. HW4 - blank - etl - turn into HW5
    g. HW5 - Due Sunday, March 21
11. Monday, March 22 - Tools to deal with lots of data - mapreduce, spark, etc and theory - pyspark
    a. Outline
    b. Why distributed computing and mapreduce - slides, video
    c. Notebook - ipynb file to give intro to databricks - OPEN ON databricks
    d. Link to databricks community edition
    e. Need to expand detail on mapreduce process and find general reading
        i. How it might work on other data
12. Monday, March 29 - More databricks
    a. Notebook - More transform operations on databricks
        i. Note - 'light' lesson as students were really struggling at this point last semester due to covid/elections
    b. HW 5 - blank to be done on databricks - update to HW6!
    c. HW 5 - completed -
        i. HW6 - Due Sunday, April 4 - Move into previous week and then make new shorter one on semi-structured? Current HW is sorta light in that it's just some transforms.
        ii. OR - Connecting to S3 via databricks and brief lesson the week before?
    d. semi-structured data?
13. Monday, April, 5 - Scheduling and Airflow
    a. HW7 - Due Sunday, April 11
        i. Make on scheduling via airflow
14. Monday, April 12 - **Flex/TBD**
15. Monday, April 19 - Final project W/1
    a. Final project goals - slides, video
    b. Notebook - Final project template
    c. Guide to getting data on sheets for import
    d. Points for validating data and question - Due Thursday, April 22
        i. MAKE
16. Monday, April 26 - Final project W/2
    a. Final project due Sunday, May 2
17. Monday, May 3 - Wednesday, May 6 - Final review
18. **Final exam - Monday, May 10**
    a. Study guide
    b. Final exam was short due to breaks I gave them. Could be beefed up.

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).** Indeed, *due dates are firm*, and late work will be accepted only with a verifiable and valid excuse.
2. All holidays or special events observed by organized religions will be honored for those students who show affiliation with that particular religion.
3. Absences pre-approved by the UA Dean of Students (or Dean designee) will be honored.
4. Arriving late and leaving early is extremely disruptive to others in the class. Please avoid this kind of disruption.
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

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: http://deanofstudents.arizona.edu/academic-integrity/students/academic-integrity.
2. It is the University's goal that learning experiences be as accessible as possible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, please let me know immediately so that we can discuss options. You are also welcome to contact Disability Resources (520-621-3268) to establish reasonable accommodations. For additional information on Disability Resources and reasonable accommodations, please visit http://drc.arizona.edu/.
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.
8. 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.