CS 598 YP Spring 2024: ML and Data Systems

2:00 PM - 3:15 PM (Tue/Thu)
1310 Digital Computer Laboratory

- CRN: 48260, Section YP (https://courses.illinois.edu/schedule/2024/spring/CS/598)
- Instructor: Yongjoo Park (https://yongjoopark.com):
    a. Email: yongjoo@g.illinois.edu
    b. Office Hours: 3:20pm-4:00pm Tuesday, Thusday at Rm 2114 Siebel
- TA: Billy Li
    a. Email: zl20@illinois.edu
    b. Office Hours: TBD
- TA: Hanxi Fang
    a. Email: hanxif2@illinois.edu
    b. Office Hours: TBD

Prerequisites

1. **Background:** "CS400 (AI) + CS411 (DB)" OR equivalent knowledge.
2. **Programming:** Proficiency in one of the following programming languages: C, Java, etc. We don't provide programming practices; you should be able to write code independently.

Course Components (or Summary of Workload)

1. **Paper Reading/Reviews:** We will cover research papers that have transformed how we manage data. You are required to read and understand them thoroughly. We will have quizzes every class.
2. **Projects:** We will solidify our understanding of ML+data systems by extending state-of-the-art systems such as Apache Arrows, PyTorch, etc.
    a. Project 1: Online Aggregation
    b. Project 2: DataFrame
    c. Project 3: Tensors
    d. Project 4: Independent research
3. **(Optional) Paper Presenting:** You are encouraged (although not required) to present one paper (i.e., one of the papers we cover in this course in addition to your final project presentations).

Online Discussion: TBD. As Campuswire retires, we will need to find a new platform.

Course Overview

Topic Overview

- **ML for Data Systems**
  - Approximate query processing: Sampling, Sketching, Online Aggregation (Project 1)
  - Database for unstructured data: NoScope, VStore, Lucene, ElasticSearch
  - System optimization: Cloud Computing, Serverless, Indexes
- **Data Systems for ML (and Data Science)**
  - DataFrame (of Apache Arrow for new Pandas Dataframe) - Project 2
    - Columnar layout, Zero-copy
  - Tensor (of PyTorch) - Project 3: Storage/memory layout
  - Interactive Data Science: ElasticNotebook

Lecture Structure

Each lecture will include one student presentation (if a volunteer exists for extra credits) and the instructor's recap. Then, the class will discuss the paper. Finally, a short quiz will be given to review materials, which will also serve as attendance checks.

Overview	5 mins
(If applicable) Student Presentation	25 mins
Instructor Presentation	30 mins
Summary/Discussion	10 mins
Quiz	10 min

[WIP: weekly schedule and lecture slides]

Why don't I have access to the above link? It's because your current web browser is associated with a Gmail account different from the university-provided email. To resolve the issue, ensure your browser is logged in with the netid@illinois.edu email address. Find more information on this page regarding the Illinois G account: https://answers.uillinois.edu/illinois/page.php?id=55049

Evaluation

There are five required components (E1-E5) and one extra-credit component (E6):

	Individual Weight
E1. Daily reviews	15%
E2. Quizzes	15%
E3. Exam	20%
E4. Projects	40%
E5. Participation	10%
E6. Optional Presentation	5%
sum	105%

The sum is intentionally 105%. Scores are not curved relative to others; that is, someone else's scoring more with extra credit items will not lower your score.

E1. Daily Reviews (15%)

We will cover one paper per class. Please submit your review one day before each class by 11:59 PM. For example, if we have a class on January 25, 2024 (Thursday), please submit your review on Canvas by 11:59 pm Central Time on January 24, 2024. If you pass this time, you simply cannot upload your reviews and will be deducted points (if you miss more than 80%). To consider personal circumstances, you can miss up to 20% of reviews, without any penalty. That is, your final score will be calculated as follows.

(your review score) = min(raw score, 80) / 80 * 100

Please do not use public AI (e.g., ChatGPT, Google Bard) for writing reviews. ***However, if you are running an open-source language model (e.g., Meta LLaMA) on your laptop, that is OK.***

E2. Quizzes (15%)

To briefly review the paper reading, we will have a quiz in each class. Each quiz will be a short multi-choice question. Please do not use public AI (e.g., ChatGPT, Google Bard) for writing reviews.

However, if you are running an open-source language model (e.g., Meta LLaMA) on your laptop, that is OK.

E3. Exam (20%)

We will have one in-person midterm. The midterm will cover the papers we study throughout the semester, focusing on their core insights.

This is an open-book exam. Moreover, you are allowed to use any AI services including ChatGPT and others. There is no guarantee that ChatGPT will give you correct answers though.

E4. Projects (40%)

The programming projects will be a major part of this course. The projects will be individual; no group projects. There will be four projects, related to the following topics:

- Online Aggregation
- DataFrame
- Tensor
- Research (you will have to choose one). The goal of this research component is to improve any of the existing systems and algorithms. For the existing systems, you can choose the ones we cover in an earlier project (e.g., DataFrame), or you can choose a different one. However, you may not simply use an existing system (e.g., writing SQL queries). You must change source code in a meaningful way.

Late submission policy: 20% reduction for each day. Even if you submit one second after a due time, your submission is regarded as one-day late. The deadline will always be 11:59 PM Central.

E5. Participation (10%)

Class Participation (5%)

You are expected to participate in class discussions in any forms such as asking questions, answering someone else's questions, etc. Chatting with your classmates do not count. Your participation will be checked using the Google Forms submitted by you.

Attendance (5%)

You will get a full 5% if you come to class without disrupting the class.

Final Letter Grade

A: [90, ∞)
B: [80, 90)
C: [70, 80)
D: [60, 70)
F: [-∞, 60)

where "[" means an inclusive lower bound, and ")" means an exclusive upper bound.

Other Resources

Academic Integrity Policy and Procedure

Anti Racism and Inclusivity Statement

Accessibility and Accommodations

Religious Observations

Sexual Misconduct Reporting Obligations

Statement on CS CARES and CS Values and Code of Conduct

All members of the Illinois Computer Science department - faculty, staff, and students - are expected to adhere to the CS Values and Code of Conduct. The CS CARES Committee is available to serve as a resource to help people who are concerned about or experience a potential violation of the Code. If you experience such issues, please contact the CS CARES Committee. The instructors of this course are also available for issues related to this class.