CS165 Syllabus - Fall 2015

Professor: Stratos Idreos
URL: http://stratos.seas.harvard.edu
Office: MD139
e-mail: stratos@seas.harvard.edu

TFs
Manos Athanassoulis [manos@seas.harvard.edu]
Michael S. Kester [kester@seas.harvard.edu]
Lukas Maas [maas@seas.harvard.edu]
Abdul Wasay [awasay@seas.harvard.edu]
Alex Liu [zezhouliu@college.harvard.edu]
TF’s Office: MD 136

Class website: http://daslab.seas.harvard.edu/classes/cs165

What is this class about?
We are in the big data era and data systems sit in the critical path of everything we do, i.e., in businesses, in sciences, as well as in everyday life. This course will be a comprehensive introduction to modern data systems. The primary focus of the course will be on modern trends that are shaping the data management industry right now such as column-store and hybrid systems, shared nothing architectures, cache conscious algorithms, hardware/software co-design, main-memory systems, adaptive indexing, stream processing, scientific data management, and key-value stores. We will also study the history of data systems, traditional and seminal concepts and ideas such as the relational model, row-store database systems, optimization, indexing, concurrency control, recovery and SQL; In this way, we will discuss both how data systems evolved over the years and why, as well as how these concepts apply today and how data systems might evolve in the future.

What is this class not about?
This class is not a traditional introduction on how we use a database system and how to write SQL. Instead, this is a systems class about data system design. You will learn how data systems work at their core and how to design new systems for emerging applications and hardware. By the way, if you know how systems work, you also become better at using them!
Why take this class?
Data is everywhere. Every year we create even more data. As it stands, every two days we create as much data as much we created from the dawn of humanity up to 2003 [Eric Schmidt, Google]. Sciences, businesses and everyday life are severely affected. Data systems are in the middle of all this. Data systems is how we store and access data, i.e., they are the backbone for any data-driven application. It is a $100B industry, growing 10% every year [Economist, “Data, data everywhere”].

At the same time data systems research and the whole industry are going through a major and continuous transition; given that new data-driven scenarios and applications continuously pop up, there is a continuous need to redefine what is a good data system design in such dynamic environments.

What is the expected learning outcome?
   a) To become familiar with the history and evolution of data systems design over the past 4-5 decades.
   b) To understand the basic tradeoffs in designing and implementing modern data systems through a step-by-step hands-on experience.
   c) To be able to design a new data system given a data-driven scenario and eventually build a functional prototype.
   d) To be able to understand which data system is a good fit given the needs of an application.
   e) To deepen C programming, debugging, and performance profiling skills.

Class Philosophy
CS165 has unlimited office hours, unlimited late days for deliverables, relies on the latest research papers instead of a standard text book, lectures are based on interaction and discussion instead of just “lecturing”, many of the quizzes and problem sets are actually open research problems and most of all it is fun!

The instructor and the TFs are here to help you all days and at all times through out the semester. You may request as many meetings as you like and as much help as you want.

The class is also geared towards engaging creative thinking and problem solving to give students a feeling of how computer science research takes place. Many of our students in the past have successfully engaged in research projects with DASLab and published research papers.

From your side you should be aware that this is a heavy class that combines knowledge about system design, algorithm design, data structures and has a hefty systems project. You are going to learn state-of-the-art techniques that are being applied in the real world
right now. Following the material of the class and performing a successful project requires serious weekly commitment throughout the semester.

**Who can take this class?**
Prior knowledge of *C programming and systems programming*, as well as a good understanding of computer architecture and in particular the *memory hierarchy (cache memories)* is very important for this class. Courses providing systems background (like CS50 and in particular CS61 or equivalent) are essential. Good hacking, algorithm designing, and data structures skills are also required.

A self-evaluation guide will be posted on the class website to help you understand if you qualify for the course and how much material you might need to cover. The course (lectures, sections and office hours) is designed so you can acquire the necessary background even if you miss some essential knowledge. So we have you covered. However, you should be aware that if you did not breeze through the self-evaluation test you will have to put more hours to successfully go through the course.

Talk to the instructor if you have not taken CS61 or if you do not feel completely comfortable with the self-test but you still think you are ready for CS165.

**Interaction in Every Class**
In every class there will be a 20-30 minute session where students will work on problems in groups of 3-4 students. The instructor and the TFs will be walking around in the classroom to participate in the discussions and brainstorm with the students. The problems will be based on material that has been presented in class and these discussions will be used to either solve open problems or to introduce new ideas. The topics in our midterms will resemble the topics and expectations during those interactive sessions and we will also use those sessions to brainstorm about the milestones of the semester project.

**Lectures**
The class meets twice a week: Mondays and Wednesdays 1:00-2:30pm, room *MD G115*. Class starts at 1:10pm. Classes are designed to be discussion-based and slides will be used mainly to drive discussions as opposed to delivering the material. For some of the classes you will be required to read part of the reading material upfront as homework and we will use the class time to discuss design choices and solve problems.

**Office hours**
We have office hours every day of the week. Starting Week 1, Prof. Stratos Idreos will hold office hours every week day 2:30-3:30pm at his office in MD139.
Sections
Sections are offered 4 times per week by the TFs: Sections will take place Sunday to Wednesday 6:30-8pm in room MD136.

The purpose of sections is twofold. First, sections are used as a slot for students to ask questions about the material of the class and the project. Second, sections are used to deliver material about the class, i.e., to go more deeply into some of the concepts discussed in class, to do additional quizzes, or to deliver background material that is needed to follow next week’s class or material needed for the project. Unless mentioned otherwise, every week all slots will cover the same material so you may go to any of the slots or go to all of them if you have questions.

Attendance
Based on the philosophy of the course, attendance in both section and lectures is optional. The best way to learn, though, is through discussion and interaction with the instructor and the TFs. Our classes and lectures are not about “lecturing” – they are about interaction. We hope to see you there!

Brainstorming sessions
It is a tradition in CS165 and CS265 to schedule several brainstorming sessions throughout the semester. Typically we bring food and drinks and have a relaxed time discussing projects, open research topics, careers in industry and academia, grad school and anything else you may have in mind.

Research Tuesdays
Discuss with the instructor and DASlab PhDs and postdocs about research! First, DASlab researchers will present their recent work on data systems research and connect it with the material you are learning in class. Then, you will get the chance to talk with them about their research, open problems and be exposed to research opportunities. Snacks and drinks will be provided.

Office hours and sections for Extension School students
If the existing slots do not work (e.g., due to time differences), we will include additional slots for office hours and sections that will work for those that cannot make the existing slots.

Feedback
We welcome feedback and ideas about the course at any point during the semester. Just come and chat with us during office hours!
Guest lectures
Every semester we are arranging 1-2 guest lectures from leaders in data system design from industry and academia. Past guest lecturers in our 2014/2015 classes include: Guy Lohman from IBM Research, Erietta Liarou from EPFL Lausanne, Alkis Simitsis and Georgia Koutrika from HP Labs and Nikita Shamgunov from MemSQL.

Required textbook
The class is about state-of-the-art data system design. There is no textbook for that. Thus, we use recent research papers and surveys which will be posted on the course website and you will have access to them through the Harvard network. We also use the following textbook: Database Management Systems, by Raghu Ramakrishnan and Johannes Gehrke. This textbook is a great source for all the seminal and traditional topics that we will cover.

Slides/Notes
The slides used during the course will be available online before each class. If there is material that we want to communicate to you only after class, this will be available shortly after each class.

SLIDES ARE NOT NOTES! You should not expect the slides to cover the material in detail. The class is based on discussion and problem solving; the slides are tailored to drive the discussion as opposed to serving the material.

In each class one or more students will be assigned to take notes, which we will then make available to everyone. Afterwards, any student will be able to jump in and enrich the notes further. Collaborative note taking and editing will be part of your class participation grade and a great way to recite the material and also see how your fellow students perceive it.

Link to the notes: http://tinyurl.com/cs165-notes

Class Project: Building a Main-memory optimized Column-store
The class has a semester-long running project. The project is about designing and implementing a prototype of a modern main-memory optimized column-store database system. By the end of the project you will have designed, implemented and evaluated several key elements of a modern data system and you will have experienced several design tradeoffs in the same way they are experienced in all major data* industry labs.

This is a heavy but fun project! We will also point to several open research problems throughout the semester that may be studied on top of the class project and that you may decide to take on as a research project if you like.
The project will have a total of 5 milestones with specific expected software deliverables, which will be accompanied with a design document. The deliverables will be tested using predefined automated unit tests for functionality and, as extra credit, for performance.

We will give you starting code that implements the basic client-server functionality, i.e., communication, so you can focus on building the server side code, i.e., the essential core data processing algorithms and data structures of a database system. In addition, whenever applicable we will let you know if there are existing libraries that is OK to use.

There is a dedicated project website:  
http://daslab.seas.harvard.edu/classes/cs165/project.html

**Evaluation**: Individual deliverables should pass the provided tests. However, you will not be judged only on how well your system works; it should be clear that you have designed and implemented the whole system, i.e., you should be able to perform changes on-the-fly, explain design details, etc.

At the end of the semester each student will have a 1-hour session with the instructor and another 1-hour session with the TFs where they will demonstrate the system and answer design questions about the current design and about supporting alternative functionality. [Tip: From past experience we found that frequent participation in office hours, brainstorming sessions and sections implies that the instructor and the TFs are very well aware of your system and your progress which makes the final evaluation a mere formality for these cases.]

**Collaboration policy**: The project is an individual project: the final deliverable should be personal, you must write from scratch all the code of your system and all documentation and reports. Discussing the design and implementation problems with other students is allowed and encouraged! We will do so in the class as well and during office hours, sections and brainstorming sessions.

**Late days policy**: We allow for 1000 late days or until Harvard requires us to upload your grade! The more input you give us the more we can help you learn. The schedule of deliverables is a reasonable schedule that we believe will help so that you distribute the project load properly through the semester. This is a heavy project that requires commitment through the whole semester and cannot be done in 2-3 weeks at the end. Not submitting on time will have no side-effects on your grade but at the same time we will not be able to provide you with any feedback on your progress until we have your design documents and your code. Submitting on time means you will get feedback within a week.
Extra points for bonus tasks: We will regularly assign extra tasks or you can come up with your own extra tasks for the various components of the project. With these extra tasks you gain extra points.

Leaderboard: We will have a running competition and an anonymous leaderboard infrastructure so you can continuously test your system against the rest of the class.

Best projects: The best 3 overall projects will gain additional extra points. "Best" is defined in terms of elegant system design, code quality, system efficiency and documentation.

What is a successful project? A successful project passes all the predefined tests we provide and the students successfully pass the final face-to-face evaluation.

A successful final evaluation is one where the student is able:

1. to fully explain every detail of the design, and
2. to propose efficient designs for new functionality on the spot.

A project will get extra points for going the extra mile to provide solutions that are efficient and are elegantly implemented and designed. Efficiency is judged by comparing against other systems. Elegance is judged by the course staff. Participating in office hours and extra sessions guarantees that you get feedback about your design often.

A project also gets extra points for doing some of the bonus tasks we provide or bonus tasks that the student comes up with.

Quizzes & Midterms
We will do several quizzes during class. Books and notes may be open during quizzes.

Similarly, we will have 2 midterms. Books and notes may be open.

Both quizzes and midterms are not designed to test how much you know. Instead, they stress your ability to come up with new solutions. Feedback on midterms and quizzes will be provided within a week.

Online discussions
We will use Piazza for online discussions.
Piazza home page: http://piazza.com/harvard/fall2015/cs165/home
Sign up: http://piazza.com/harvard/fall2015/cs165
Assessment and grading
- Class project 60%
- Quizzes and class participation: 10%
- Midterms (2): 30%
- Bonus points: Extra tasks for the project: 10%
- Bonus points: Best projects: 5%
(Each of the 5 milestones of the project account for 20% of the total project grade)

Extension School
Lectures will be broadcasted live Mondays/Wednesdays 1-2:30pm. Lectures will also be available for on-demand broadcast within 24 hours after each class. Students will be able to watch the live or recorded broadcast through their browser using the Matterhorn player. The link to the broadcasts for CS165 is here: http://matterhorn.dce.harvard.edu/engage/ui/index.html#/2016/01/14861.

You should also monitor the canvas website where extension school is going to post a virtual classroom link as of September 2nd, available at: https://canvas.harvard.edu/courses/4284.

Extension school students will be able to participate to sections and office hours via web-conference tools: We will use Blackboard. The course stuff will also be online with Backboard during class time for extension school students who can participate during class time. We will do that as of Week 3. For remote students who cannot follow the live schedule we will organize broadcasts where the course staff will also be online to participate during interactive sessions (e.g., a quiz).

Grading: For extension school students the 10% “class participation” portion of the grade will be distributed equally in project (65%) and midterms (35%).

Midterms: Extension School will contact students directly regarding administrative preparations and options for midterms. Midterms are proctored and we also allow the new option to take the midterm directly through Canvas with a camera. Local extension school students may come to campus and take the midterm together with college students on midterm day.

Piazza: To participate in piazza you need a Harvard email address. If you do not have one you can create one here: http://www.extension.harvard.edu/resources-policies/resources/computer-e-mail-services.

Extension school shows the class starting date to be September 3. This is when the first recorded video is available. However, extension school students will still be able to stream the September 2 class as normal.
Plagiarism
You are responsible for understanding Harvard and Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting "the wrong draft", or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two, free, online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Accessibility
Harvard and the Extension School are committed to providing an accessible academic community. The Disability Services Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.