**Scalable and Data-Intensive Computing in the Cloud**

Instructor: Bill Howe, Phd

Email: billhowe@cs.washington.edu

Phone: (206) 616-5828

Office: Allen 450

Period: Mar 26 - Jun 4, 2012

Class Meeting Time: Mondays, 6:00-9:00 p.m.

Class Meeting Place: ST 359

Web: http://www.cs.washington.edu/homes/billhowe/bigdatacloud

Office Hours: by appointment

**Course Overview:** We will explore the technology landscape at the intersection of “big data” and cloud computing. Articles in the NYTimes, Science, Nature, the Economist, a variety of reports from the federal government and innumerable blog posts have made the case that technology supporting the deep analysis of massive datasets – big data – is now a critical enabling technology for business and research in all fields. Cloud computing has been a catalyst for this trend, democratizing access to the infrastructure required to build big data applications. Scalable data platforms are increasingly deployed in the cloud or offered directly by cloud providers.

The course will be technology driven. We will consider relational databases (specifically in the context of cloud computing), the Hadoop ecosystem and its variants, other NoSQL platforms emphasizing low-latency access, and more. We will work directly with a selected set of these platforms, compare and contrast their relative strengths and weaknesses, and characterize the problems they are designed to solve.

**Learning Objectives:** By the end of this course, students will be able to:

* Develop a basic test application in each of several selected technologies
* Evaluate application requirements and recommend an appropriate data-intensive computing technology platform or platforms to satisfy them
* evaluate an existing data-intensive computing application and identify opportunities for improvement
* Explain the relative strengths and weaknesses of selected major data-intensive computing and data management platforms in use today.
* Explain the major technology trends that are influencing the industry.

**Course Structure:** Each class will consist of a 1-hour lecture, a 1-hour case study and demonstration of a specific system, and 1-hour of discussion and hands-on work. Each week, we will consider a category of scalable data platform through a lecture and consider a representative example from this category in detail through a demonstration. Students will be asked to complete a hands-on homework assignment based on the material presented in class and, in some cases, come prepared to discuss assigned reading. The reading assignments will generally be research papers from relevant computer science conferences.

**Student Assessment:** Assignments: 80%, Participation: 20%. Participation will be a combination of attendance and discussion involvement; in class and online involvement will both contribute. Assignments will typically not be graded in terms of correct/incorrect answers, but students will be expected to demonstrate effort and insight. In this course, discussion between students is not discouraged; the goal is to learn as much as possible in a short time, and discussion is a very efficient way to do this. Some assignments may be completed in groups, depending on students’ experience level. In these cases, a portion of the grade will be based on peer review by one’s group members.

**Textbook:** None. All materials will be on the web. We will use a combination of slides, documentation for the selected systems covered in class, some custom material, and some relevant research papers.

**Prerequisite:** The assignments will involve example-oriented programming assignments in various languages, possibly including Java, C#, and Python. You will NOT be expected to be proficient in any of these languages, but you will be expected to “think computationally.” Specifically, you will work through examples, answer questions about the code, and generally be “brave” with respect to learning new technology – read the documentation, ask questions, try experiments, make some educated guesses, etc.

**Specific Course Topics and Tentative Schedule:**

*Week 1:* Cloud computing review, big data overview, problem types, history and context, methodology

Case study: Amazon EC2, S3

Reading: None

Assignment: <http://www.cs.washington.edu/homes/billhowe/bigdatacloud/lecture1/assignment.html>

1) Setup accounts

2) http://escience.washington.edu/get-help-now/get-started-amazon-web-services

3) <http://www.windowsazure.com/en-us/pricing/free-trial/>

4) Cloud Storage Write Speed Shootout

*Week 2*: Hadoop, Elastic MapReduce (Hadoop vs. RDBMS)

Case study: Elastic MapReduce

Reading:

<http://www.cse.nd.edu/~dthain/courses/cse598z/spring2010/benchmarks-sigmod09.pdf>

<http://www.cs.princeton.edu/courses/archive/spring11/cos448/web/docs/week10_reading2.pdf>

Assignment: Elastic MapReduce Sample

*Week 3:* Iterative Applications. HaLoop, Spark, Daytona

Guest Lecturer: Jaliya

Case study: Daytona

Reading: <http://www.cs.washington.edu/homes/billhowe/pubs/HaLoop.pdf>

Assignment: Run k-means algorithm from SeaFlow data on Daytona.

*Week 4*: Relational Databases in the cloud: MS SQL Azure, RDS, database.com, Google Cloud SQL

Case study: SQL Azure (and SQLShare)

Reading: None

Assignment: Compare SQL Azure and Amazon RDS; SQL exercise

Week 5: key-value stores and NoSQL: Google App Engine DataStore, BigTable, Megastore, PNUTS

Case study: Google App Engine

Reading: http://[www.cidrdb.org/cidr2011/Papers/CIDR11\_Paper32.pdf](http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper32.pdf)

Assignment: Google App Engine

Week 6: key-value stores, cont’d. low-latency, indexing: Hbase, DynamoDB

Case study: DynamoDB

Assignment: DynamoDB speed test

Week 7: Declarative Query, Pig, HIVE, Tenzing, Dremel

Case study: Pig

Reading: <http://infolab.stanford.edu/~usriv/papers/pig-latin.pdf>

Assignment: Elastic MapReduce, Billion Triple Challenge Pig Query

Week 8: Application Studies: Social Networks, Text data, Biological data

Assignment: Design Phase for Mini-Project

1) Social Networks: Provide a solution for graph query over the Billion Triple Challenge Dataset (e.g., triangle query)

OR

2) Text: Provide a solution for keyword search over the entire corpus for the Wikipedia dataset

OR

3) Genomics: Perform one of the analyses described in the talk below using the 1000 genomes dataset

<http://schatzlab.cshl.edu/presentations/2010-03-15.XGen-Scalable%20Solutions.pdf>

Week 9: NewSQL: Column-oriented DBs, CAP theorem revisited

MonetDB, Vertica, VoltDB, JustOne DB, NimbusDB

Case Study: MonetDB

Reading: <http://db.csail.mit.edu/projects/cstore/abadi-sigmod08.pdf>

Assignment: Group case studies

Week 10: Mini-Project Presentations

Exercise: Cloud Data Interview Questions

**The instructor reserves the right to alter the syllabus if circumstances dictate.**