Course Overview

CptS 415

What is Big Data?

CptS 415

What is Big Data?

- No official definition!

- A collection of datasets so large that
  - Difficult to process using on-hand database management tools
  - Difficult to process with traditional data processing applications

- The challenge include capture, curation, storage, search, sharing, transfer, analysis and visualization

Big Data - Example

Software application for storing and processing of images and videos:

- *store images / videos in a database*
- *user interface for loading, tagging and searching the images / videos*

Big Data - Example

Software application for storing and processing of all WSU students' images and videos:

- *server computers*
- *application and database running on different servers*

Big Data - Example

Software application for storing and processing of all (global) university students' images and videos:

- *multiple servers across different regions*
- *distributed application and database architecture*

"Data center" by seeweb is licensed under CC BY-SA 2.0

Big Data Applications

Course Details

CptS 415

Instructor Introduction

- B. Tech. - Indian Institute of Technology (India)
- M.S. - Vanderbilt University (Nashville, TN)
- 20+ yrs industry experience
- First Job:
  - Programmer / Data Analyst
- Recent Roles:
  - Software Engineer, SLAC
  - Founder - EpiData, Inc.
- Personal Interests:
  - Snowboarding, Golf, Hiking and Travel

Objective and Scope

- Objective:

  Understand core concepts of Big Data techniques, and develop ability to implement real-world Big Data solution

- Scope:
  - Relational databases and SQL
  - Document data store models
  - NoSQL databases including Graph databases
  - Hadoop and Spark for processing big data

About the Course

- It is not a
  - Programming language course
  - Independent database or data mining course

- This course is
  - Provide design principles for Big Data challenges
  - Overview/survey of state-of-the-art Big Data techniques, tools and principles.
  - Provide pointers to Big Data research projects, papers, tools, and commercial/open-source projects

- This course is unique in
  - A complete overview of major Big Data techniques
  - Algorithm design techniques for Big Data
  - Academic & Industrial practice

Course Format

- Seminar-style course:
  - No Exam!!
  - 6 homework assignments
  - 1 course projects with multiple milestones
- Suggested Textbook:
  - Database System: The Complete Book
  - Big Data Fundamentals: Concepts, Drivers and Techniques
  - Mining of Massive Datasets
- Online Tutorials & Papers
  - Research papers or chapters related to the topics
  - Checkout the resources listed in the syllabus

Grading

Categories	Percent of Overall Grade
Participation	10%
Homework / Assignments	50%
Project	45%
Total	105%

Grading (cont.)

- Participation Activities:
  - Panopto Videos
  - Survey / Poll responses
- Homework / Assignments:
  - Theoretical and design questions
  - Simple implementations
- Project:
  - Big Data project - from design to implementation
  - Presentation and report

Grading (cont.)

Grade	Percent	Grade	Percent
A	93 - 100	C	73 - 76.99
A-	90 - 92.99	C-	70 - 72.99
B+	87 - 89.99	D+	67 - 69.99
B	83 - 86.99	D	63 - 66.99
B-	80 - 82.99	D-	60 - 62.99
C+	77 - 79.99	F	0 - 59.99

Academic Integrity

Academic integrity will be strongly enforced in this course.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students, WAC 504-26-010(3).

Copyright protects intellectual property and work of individuals, including instructors.

Academic Integrity violations examples (but is not limited to)

- Copying/taking a picture of another student's code/work
- Letting another student copy/take a picture of your code/work
- Sending your code/work to another student (i.e. digitally or in print)
- Receiving another's student code/work (i.e. digitally or in print)

Weekly Plan

	Sunday	Monday - Friday	Sunday
Instructor	- Lecture Videos and Slides Released - Assignments Released		
Students		- Watch Lecture Videos - Read Textbook Chapters - Complete Participation Activity	Submit Assignments / Project Milestones

Course Schedule

Dates	Lesson Topic	Participation	Assignment
Week 1 Aug. 21 - Aug. 27	• Course Overview, Big Data Overview	• Week 1 Videos (**due Aug. 25**) • Introductions (**due Aug. 27**)	No Assignment
Week 2 Aug. 28 - Sept. 3	• Relational DBMS	• Week 2 Videos (**due Sept. 1**)	• Project Team Formation (**Sept. 2 - Sept. 3**)
Week 3 Sept. 4 - Sept. 10	• SQL, Relational Algebra	• Week 3 Videos (**due Sept. 8**)	• Assignment 1 (**due Sept. 10**)
Week 4 Sept. 11 - Sept. 17	• XML and JSON	• Week 4 Videos (**due Sept. 15**)	• Project Milestone 1 **(due Sept. 17)** • CATME Peer Eval. Survey (**due Sept. 17**)
Week 5 Sept. 18 - Sept. 24	• Graphs and RDF	• Week 5 Videos (**due Sept. 22**)	• Assignment 2 (**due Sept. 24**)
Week 6 Sept. 25 - Oct. 1	• Distributed Systems and NoSQL Databases	• Week 6 Videos (**due Sept. 29**)	• Project Milestone 2 **(due Oct. 1)** • CATME Peer Eval. Survey (**due Oct. 1**)
Week 7 Oct. 2 - Oct. 8	• Query Processing and Query Optimization	• Week 7 Videos (**due Oct. 6**) • Plus/Delta Survey (**due Oct. 7**)	No Assignment
Week 8 Oct. 9 - Oct. 15	• Graph Query Processing and Approximate Query Processing	• Week 8 Videos (d**ue Oct. 13**)	• Assignment 3 (**due Oct. 15**)
Week 9 Oct. 16 - Oct. 22	• MapReduce	• Week 9 Videos (**due Oct. 20**)	• Project Milestone 3 **due Oct. 22** • CATME Peer Eval. Survey (**due Oct. 22**)
Week 10 Oct. 23 - Oct. 29	• Hadoop	• Week 10 Videos (**due Oct. 27**)	• Assignment 4 (**due Oct. 29**)
Week 11 Oct. 30 - Nov. 5	• Apache Spark	• Week 11 Videos (**due Nov. 3**)	• Project Milestone 4 (**due Nov. 5**) • CATME Peer Eval. Survey (**due Nov. 5**)
Week 12 Nov. 6 - Nov. 12	• Big Data Theory & Practice	• Week 12 Videos (**due Nov. 10**)	• Assignment 5 (**due Nov. 12**)
Week 13 Nov. 13 - Nov. 19	• Data Quality and Data Privacy	• Week 13 Videos (**due Nov. 17**)	• Project Milestone 5 (**due Nov. 19**) • CATME Peer Eval. Survey (**due Nov. 19**)
Thanksgiving Break	Nothing due this week		
Week 14 Nov. 17 - Dec. 3	• NewSQL and In-Memory DBMS	• Week 14 Videos (**due Dec. 1**)	• Assignment 6 (**due Dec. 3**)
Week 15 Dec. 4 - Dec. 10	Project Completion Week •	• Course Evaluation (Course Feedback)	• Project Presentations and Demos (**Dec. 8 - Dec. 9**)
Week 16 Dec. 11 - Dec. 15	• Finals Week		• Project Report **(due Dec. 11)** • CATME Peer Eval. Survey (**due Dec. 11**)

Summary