**PART A: BIG DATA TECHNOLOGY**

1. INTRODUCTION TO BIG DATA

Sangmi Lee Pallickara  
Computer Science, Colorado State University  
http://www.cs.colostate.edu/~cs535

---

**What is Big Data?**

- Things one can do at a large scale that cannot be done at a smaller one  
  • To extract new insights  
  • Create new forms of values

- Big Data is about analytics of huge quantities of data in order to infer probabilities  
  • Big Data is NOT about trying to “teach” a computer to “think” like humans  
  • Providing a quantitative dimension it never had before

---

**The three(or four) Vs in Big Data**

- **Volume**  
  • Voluminous  
  • It does not have to be certain number of petabytes or quantity.

- **Velocity**  
  • How fast the data is coming in?  
  • How fast you need to be able to analyze and utilize it

- **Variety**  
  • Number of sources or incoming vectors

- **Veracity**  
  • Can you trust the data itself, source of the data, or the process?  
  • User entry errors, redundancy, corruption of the values  
  • Data cleaning

---

**Related research areas**

- **Storage systems**  
  • How can we efficiently resolve queries on massive amounts of input data?  
  • The input dataset may be presented in the form of a distributed data stream

- **Machine learning**  
  • How can we efficiently solve large-scale machine learning problems?  
  • The input data may be massive, stored in a distributed cluster of machines

- **Distributed computing**  
  • How can we efficiently solve large-scale optimization problems in distributed computing environments?  
  • For example, how can we efficiently solve large-scale combinatorial problems, e.g. processing of large scale graphs?
Connected cars

- Single hybrid plug-in car generates up to 25 gigabytes per hour

- Traffic problem, re-routing based on the volume of traffic
- Alerts driver when road conditions are hazardous by automatically activating anti-lock break
- The information is shared by the vehicles that are nearby

The Artemis project:
Saving "preemies" using Big Data

- The Artemis project
- Dr. Carolyn McGregor

- Captures and processes the patients’ data in real time
- 16 different data streams
- Heart rate, respiration rate, temperature, blood pressure and blood oxygen level
- Around 1,260 data points per second
- System detects subtle changes that may signal the onset of infection 24 hours before overt symptoms appear

Photo Credit: https://datafloq.com/read/car-manufacturers-are-using-big-data/1204
Look Who’s Peeking at Your Paycheck

• Experian’s Income Insight
• Estimates people’s income level
• Based on their credit history
• Trains the estimation model using selected credit history and tax information from IRS


My Big Data Lab at Colorado State University

- Algorithmic and systems design
  - Scalable analytics over voluminous datasets on complex distributed architectures
- Research has been deployed in the following domains
  - Precision agriculture, atmosphere science, environmental biology, ecology, civil engineering, bioinformatics, and public health
- Awards
  - Cochran Family Professorship 2018–2021
  - IEEE TCSC Award for Excellence in Scalable Computing (Mid-Career Researcher) 2018
  - National Science Foundation CAREER Award 2016
- Funded by
  - The National Science Foundation
  - The Advanced Research Projects Agency Energy (Department of Energy)
  - Department of Homeland Security
  - The Environmental Defense Fund
  - Google, Amazon, and Hewlett Packard

Communications

- Course Website
  - http://www.cs.colostate.edu/~cs535
- Announcements: Check the course website at least twice a week.
- Schedule (course materials, readings, assignments)
- Policies
- Canvas
  - Assignment submission
  - Discussion board
  - Grades
- Contact Me
  - sangmi@colostate.edu
  - Office hour: Friday 10:00AM ~ 11:00AM and by appointment
- Office: CSB456
- URL: http://www.cs.colostate.edu/~sangmi

http://www.cs.colostate.edu/~cs535
Goal of this course

- Understanding fundamental concepts in Big Data Analytics
- Learn about existing technologies and how to apply them

Computing systems
Algorithms and models

Course Structure

Part A: Big Data Technology

- Purpose
  - Understand concepts of Big Data computing environment
  - Hands-on experience

- Topics
  - Introduction to Big Data
  - Lambda Model
  - Distributed file system
  - Quick view of MapReduce
  - Introduction to Apache Spark
  - Analytics with Apache Storm

Part B: Kernel I ~ IV

- Purpose
  - Understanding different aspects of Big Data research with lectures and workshops

- Kernels
  - Kernel I: Big Data Advanced Analytics Case Study
  - Kernel II: Scalable Analytics Algorithms
  - Kernel III: Large scale Graph Analysis
  - Kernel IV: Big Data Storage with Analytics

- Duration: 2 weeks
  - 3 (x 75 minutes) Lectures
  - 1 (x 75 minutes) workshop

Course Component | Programming Assignments

- Programming Assignment 1
  - Implementing link analysis algorithm over the Wikipedia pages using Apache Spark
  - Due on 2/26 5:00PM
  - The description of PA1 is available now.

- Programming Assignment 2
  - Implementing real-time Twitter stream analysis using Apache Storm
  - Due on 3/26 5:00PM
Course Component | Programming Assignments
- All of the programming assignments are group submission
- Submission should be via canvas
- Late policy for the assignment submissions
  - Up to a maximum of 2 day past the deadline.
  - 10% penalty per day will be applied
- Each group will provide demo of the programming assignment in CSB120
- Each assignment will count 10% of total score of this course

Course Component | Quizzes
- 7-8 Quizzes
- Two lowest scores will be eliminated
- Quizzes will count 20% of total score of this course

Course Component | Term Project
- Objectives
  - Students identify their topics for the term project
  - Students provide methodology to solve their problem
  - Students implement software solution
  - Students provide evaluation scheme for their software

Course Component | Term Project
- Term project grading (40% of this course)
  - Proposal document: 4%
  - Proposal presentation (peer review): 2%
  - Final paper: 24%
  - Final demonstration: 2%
  - Final presentation (peer review): 3%
  - Participation: 4%

- Late policy for the deliverable submissions
  - Up to a maximum of 2 day past the deadline.
  - 10% penalty per day will be applied

Course Component | Term Project
- Highlights of the Previous Term Projects
  - Supporting Emergency Response During Natural Disasters with Twitter Data
  - Winning Words in the Supreme Court
  - Mendel: A Distributed Storage System for Efficient Sequence Alignment and Similarity Searching (published in IEEE IPDPS 2016)
  - Processing Smart Grid Data In Real Time (DEBS grand challenge 2014)

- Highlights of the Previous Term Projects
  - Time to Answert for Questions on stackoverflow.com using Map Reduce
  - Analysis of words for spell checking in search queries using digitized books and articles
  - Efficient Boolean Symmetric Searchable Encryption
  - Big Data I/O Performance Improvement Using Buffered B-Tree Algorithm
  - Node and Metadata Visualization in a Distributed Hash Table
  - Who is Building Wikipedia?

http://www.cs.colostate.edu/~cs535
Grading | Overview

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignments</td>
<td>20% (1% participation included)</td>
</tr>
<tr>
<td>Assignment 1: 10%</td>
<td></td>
</tr>
<tr>
<td>Assignment 2: 10%</td>
<td></td>
</tr>
<tr>
<td>Term Project</td>
<td>40% (1% participation included)</td>
</tr>
<tr>
<td>D0: Term project planning 1%</td>
<td></td>
</tr>
<tr>
<td>D1: Proposal document 4%</td>
<td></td>
</tr>
<tr>
<td>D2: Final paper 24%</td>
<td></td>
</tr>
<tr>
<td>D3: Final presentation 3%</td>
<td></td>
</tr>
<tr>
<td>Participation: 4%</td>
<td></td>
</tr>
<tr>
<td>Quizzes</td>
<td>20% (2% participation included)</td>
</tr>
<tr>
<td>Workshop</td>
<td>20% (2% participation included)</td>
</tr>
</tbody>
</table>

Grading | Participation Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignments</td>
<td>20% (1% participation included)</td>
</tr>
<tr>
<td>Assignment 1: 10%</td>
<td></td>
</tr>
<tr>
<td>Assignment 2: 10%</td>
<td></td>
</tr>
<tr>
<td>Term Project</td>
<td>40% (1% participation included)</td>
</tr>
<tr>
<td>D0: Term project planning 1%</td>
<td></td>
</tr>
<tr>
<td>D1: Proposal document 4%</td>
<td></td>
</tr>
<tr>
<td>D2: Final paper 24%</td>
<td></td>
</tr>
<tr>
<td>D3: Final presentation 3%</td>
<td></td>
</tr>
<tr>
<td>Participation: 4%</td>
<td></td>
</tr>
<tr>
<td>Quizzes</td>
<td>20% (2% participation included)</td>
</tr>
<tr>
<td>Workshop</td>
<td>20% (2% participation included)</td>
</tr>
</tbody>
</table>

Grading | Final Letter Grade

<table>
<thead>
<tr>
<th>Letter Grade</th>
<th>Total Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90.00 % and higher</td>
</tr>
<tr>
<td>B</td>
<td>80.00 ~ 89.99 %</td>
</tr>
<tr>
<td>C</td>
<td>70.00 ~ 79.99 %</td>
</tr>
<tr>
<td>D</td>
<td>60.00 ~ 69.99 %</td>
</tr>
<tr>
<td>F</td>
<td>Below 60.00 %</td>
</tr>
</tbody>
</table>

Course Component | Course Policy

- No make-up for missed quizzes and exams
  - Except for the case where student provided an advance written notice to the instructor based on an emergency
  - Supporting paper works will be requested
  - Two lowest quiz scores will be eliminated at the end of semester

- No Cell-phones in the class.
- No Laptops in the class.
  - If you need to use a laptop during lectures, please sit in the back row.
  - I will ask you to turn off your laptop if needed.

- Attend the class, ask questions, and discuss.
- Check the course web page and canvas regularly
- Try new technologies and apply them
- Share your experiences with other students in class

http://www.cs.colostate.edu/~cs535

Spring 2019 Colorado State University
Questions?