

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-1

CS535 BIG DATA

PART 0. INTRODUCTION
1. INTRODUCTION TO BIG DATA

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-2

What is Big Data?

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-3

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-4

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-5

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?

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-6

Who is using Big Data?

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-7





Photo Credit:
<https://datafloq.com/read/car-manufacturers-are-using-big-data/1204>

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-8

Connected cars

- Single hybrid plug-in car generates up to 25 gigabytes per hour
- Car manufacturers (including BMW) use big data analytics platforms
 - Eliminates models of vehicles before they go into production
 - Analyze data of prototypes
 - "Error memories" of approximate 15,000 reports
- Connected cars
 - \$130 billion by 2019
 - Traffic problem, re-routing based on the volume of traffic
 - Alerts driver when a road conditions are hazardous by automatically activating anti-lock break
 - This information is shared by the vehicles that are nearby

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-9




8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-10

The Artemis project: Saving "preemies" using Big Data

- The Artemis project
 - Dr. Carolyn McGregor
- Toronto's Hospital for Sick Children, University of Ontario Institute of Technology and IBM
 - Captures and process 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

"Achieving Small Miracles from Big Data", https://www.ibm.com/smarterplanet/global/files/ca_en_us_healthcare_smarter_healthcare_data_baby.pdf

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-11



8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-12

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

KAREN BLUMENTHAL, "Look Who's Peeking at Your Paycheck", The Wall Street Journal, Jan. 13, 2010, <http://www.wsj.com/articles/SB10001424052748703672104574654211904801106>

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-13

CS535 BIG DATA

PART 0. INTRODUCTION
1. INTRODUCTION TO BIG DATA
2. COURSE INTRODUCTION

PART 0. INTRODUCTION 2. COURSE INTRODUCTION

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-14

Goal of this course

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-15

Related courses

- Machine Learning/Statistics
- Distributed Systems
- Database Systems
- Computer Communications
- Computer Security

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-16

About me

- Research area
 - Big Data
 - Storage, retrievals, analytics and visualization
- Projects
 - Synapse
 - A Framework for Ad Hoc Model Construction in Data Streaming Environments
 - Galileo
 - Distributed data storage system for large scale geospatial time-series datasets
 - Columbus
 - Visual Analytics, cloud based workflow engine
 - Mendel
 - Genomic data storage system
- Research group meeting
 - 1:00PM ~ 2:00PM every Friday

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-17

Galileo: Big Picture

- Data is voluminous
 - Outpaces what is available on a single hard drive
- Storage must be over a collection of machines
 - Avoid central coordinators
 - Cope with failures
 - Preserve data locality without introducing storage imbalances
 - And the accompanying query hotspots
 - Support range queries and fast ingest of new data

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-18

Galileo: key features

- Support for large numbers (10^9) of small files
- High throughput storage and retrieval
- Data is multidimensional with multiple types
- Time-series data
- Support for exact match and range queries (with wildcards) along multiple dimensions
- Support for multiple data formats
 - netCDF, BUFR, HDF 4/5, and data from the Defense Meteorological Satellite Program

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-19

Galileo Design Considerations

- **Symmetric** storage nodes
 - No special-function or "controller" nodes
 - Storage and retrievals may go to any node, and will be forwarded to the targeted node(s)
- Incremental scale-up
- Failure-resiliency

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-20

Distributed Hash Tables (DHTs)

- These are decentralized systems
- Each node maintains information about a subset of nodes
- Each node uses its local information to reach any other node in the system
- Typically used as <key, value> stores

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-21

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 Instructor**
 - sangmi@cs
 - Office hour: Tuesday 11:00AM ~ noon and by appointment
 - Office: CSB456
 - URL: <http://www.cs.colostate.edu/~sangmi>

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-22

Course Structure

```

graph LR
    P0[Part 0: Introduction to Big Data] --> P1[Part 1: Batch Computing Models for Big Data Analytics]
    P1 --> P2[Part 2: Scalable Frameworks for Big Data Analytics]
  
```

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-23

Grading Policy

Category	Percentage of Final Grade
Programming Assignments	40 % Assignment 1: 20% Assignment 2: 20%
Term Project	30 % Deliverable 0: 1% Deliverable 1: 5% Deliverable 2: 18% Deliverable 3: 6 %
Quizzes and participation	10 %
Midterm Exam	20%

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-24

Programming Assignments

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-25

What will be the assignments?

- Programming Assignment 1
 - Implementing PageRank algorithm over the Wikipedia pages using MapReduce with Apache Spark
 - Due on 9/28 5:00PM
- Programming Assignment 2
 - Implementing real-time Twitter stream analysis using Apache Storm
 - Due on 11/9 5:00PM

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-26

How do I submit the assignments?

- All of the programming assignments are **individual** 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 student will provide demo of the programming assignment in CSB120
- Each assignment will count 20% of total score of this course

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-27

Term Project

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-28

Objectives of Term Project

- 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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-29

Term Project Grading

- Deliverable 1 (Proposal): 6%
- Deliverable 2 (Final): 18%
- Deliverable 3 (Presentation): 6%
- Late policy for the deliverable submissions
 - Up to a maximum of 2 day past the deadline.
 - 10% penalty per day will be applied

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-30

Deliverables of Term Project

- **Deliverable 0 (Due on August 29 5:00PM)**
 - You are strongly recommended to work on the term project as a team (2 or 3 team members)
 - However, you are allowed to work as an individual
 - Please contact me
 - Topic area (1 or 2 area)
- Please submit it via Canvas by **August 29th 5:00PM**

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-31

Deliverables of Term Project

- **Deliverable 1 (due on 10/7 5:00PM).** Term project proposal
 - Proposal of your project
 - Topic
 - Motivation
 - Literature survey
 - Methodology
 - Timeline
 - Evaluation method
 - Team management
 - Special team interview sessions: 9/27, 9/29, 10/4, 10/6 11:00AM ~ noon
 - Topics and methodology
 - Student presentation (10 minutes per team) will be followed
 - 10/11 and 10/13 in the class

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-32

Deliverables of Term Project

- **Deliverable 2.** Your software and final report
 - Due on 11/30 5:00PM
- **Deliverable 3.** Presentation and Demonstration
 - 12/1, 12/6, 12/8 in the class
 - Demonstration schedule: TBA

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-33

How do I select my topic?

- Step 1. What is your interest? **Deliverable 0**
- Example Area A
 - Real-time Big Data Analytics
 - Are you interested in a streaming data analysis? Are you interested in Twitter message analysis?
- Example Area B
 - Network Data Analytics
 - Are you interested in graph data analysis? Are you interested in social network analysis?
- Example Area C
 - Predictive Data Analytics
 - Are you interested in predicting values? Are you interested in recommendation systems?
- Example Area D
 - Visual Data Analytics
 - Are you interested in visualizing very large dataset?
- You can choose one or combine two of above area

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-34

How do I select my topic?

- Step 2. Is there any other student to work with? **Deliverable 0**
 - Find one or two

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-35

How do I select my topic?

- Step 3. **Deliverable 1**
 - Narrow down your topic
 - Specific problem

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-36

How do I select my topic?

- Step 4. **Deliverable 1**
 - Is it feasible?
 - Do you have a good dataset?
 - Do you have a good tools or environment?
 - Is the workload reasonable?

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-37

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)

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-38

Highlights of the Previous Term Projects

- "Time to Answer" 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?

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-39

Quizzes and Participation

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-40

Quizzes and Participation

- There will be 10+ pop quizzes/BSQs in class
 - 1-5 simple questions
 - Open-book, open-material, not open-Internet
- Minute survey

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-41

Midterm Exam

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-42

Midterm

- 11/17 in class
- 20% of the course score

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-43

Final Letter Grade and Course Policy

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-44

Final Letter Grade

Letter Grade	Total Percentage
A	90.00 % and higher
B	80.00 ~ 89.99 %
C	70.00 ~ 79.99 %
D	60.00 ~ 69.99 %
F	Below 60.00 %

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-45

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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-46

Course Policy

- 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.

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-47

More importantly,

- 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

8/23/2016 CS535 Big Data - Fall 2016 Week 1-A-48

Questions?