Big Data Systems

CS 5965/6965 – FALL 2014

Today …

- General course overview
- Q&A
- Introduction to Big Data
- Data Collection
- Assignment #1

General Course Information

- Course Web Page
  - http://www.cs.utah.edu/~hari/teaching/fall2014.html
  - also on canvas
- Text & References
  - No official textbook
  - will use online resources and papers
- Apt/CHPC accounts ← request asap
- TA - Anusha Buchireddygari
  - Email – anusha.buchi@gmail.com
  - Office hours

Class Interaction …

- I strongly encourage discussions & interactions
- Extra credits for strong participation

Assignments & Projects

- Assignment 1 – Due in 1 week – 10% credit
- Assignment 2 – 15% credit – 3 weeks
- Assignment 3 – 15% credit – 2 weeks
- Assignment 4 – 15% credit – 2 weeks
- Assignment 5 – 15% credit – 2 weeks
- Final Project – 30% credit – 6 weeks
- Presentations during exam week

3 parts
- Easy
- Medium
- Real world problem

Groups of 2

Things you should know

- No formal prerequisite
- Good Programming Skills (any language)
  - C/C++/Java preferred for assignments (except Assignment 1)
  - Make a case if you would like to use any other language
- Be prepared to learn new programming tools & techniques
- Git – create account on github
- Sequential algorithms, complexity analysis

This is a systems course

- Big data is a broad concept that covers several aspects of computer science
- We shall focus on the computer systems aspect – e.g.
  - How various parts of a big data computer system(hardware, software and applications) are put together?
  - What are the appropriate approaches to realize high-performance, scalability, reliability, and security in practical big data systems?
- Not the right course if you are expecting to learn about algorithmic design and theoretical foundations for machine learning & data mining.

What will we cover ?

- Data Collection
- Parallel Algorithms
- Hadoop & mapreduce
- Analytics beyond Hadoop – Spark
- Graph Algorithms - GraphLib
- Real-time analytics - Storm
- MPI
- Other solutions
- Data Storage
- Data Security, Privacy
- Data visualization
- Big Data Applications

Things that you need to do now…

- Get a github account and join **UtahCS6965**
- Request a chpc account – send email to TA with uid

- Use git → commit often

Where to look for help

- Course website / Canvas
- Email
- TA office hours → MEB 3XXX → Mon,Thu 3-5pm
- My office hours → MEB 3454 → Tue 12-2pm

Questions?

What is Big Data ?


DO WE HAVE ANY ACTIONABLE ANALYTICS FROM OUR BIG DATA IN THE CLOUD?
YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.
MAYBE IN-MEMORY COMPUTING WILL ACCEL-ERATE YOUR APPLICA-TIONS.
PLUNGE, PLUNGE, PLUNGE.
Dilbert.com DilbertCartoonist@gmail.com
1-9-13 ©2013 Scott Adams, Inc./Dist. by Universal Uclick

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

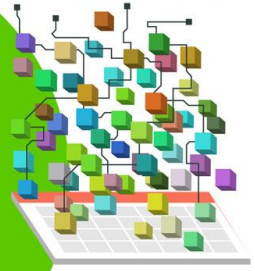
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

The recognition that data is at the center of our digital world and that there are big challenges in collecting, storing, processing, analyzing, and making use of such data.

What is Big depends on the application domain

Kinds of Data

- Web data & web data accesses
- Emails, chats, tweet
- Telephone data
- Public databases – Gene banks, census records, …
- Private databases – medical records, credit card transactions, …
- Sensor data – camera surveillance, wearable sensors, seismograms
- Byproduct of computer systems operations – power signal, CPU events, …
- … … …

Data is valuable

- Google & facebook make money by mining user data
  - for revenue from advertisements
- Financial firms analyze financial records, real-time transactions and current events for profitable trades
- Medical records can be processed for better – and potentially cheaper – health care
- Roadways are monitored for traffic analysis and control
- Face detection in airports

Collection of Big Data

- The amount of data available is increasing exponentially
- But, it is still challenging to collect it
  - Difficult to get access
  - Redundancy
  - Noise
- Example – Web data collection
  - Assignment 1

Processing & Analysis of Big Data

- Large datasets do not fit in memory
- Processing large datasets is time consuming
  - Parallel processing is necessary → challenging
- Message Passing Interface (MPI) – (1991)
  - Low(er) level C/Fortran API for communication
  - Powerful, hard(er) to code, !fault tolerant
- Mapreduce – Google (2004)
  - Originated from web data processing
  - ease of programming, fault tolerant
  - limited semantics
- Spark, GraphLab, Storm, ……

Storage & I/O

- Storage and I/O are critical for big data performance and reliability
- Hardware: disks, Flash, SSD, nonvolatile memory, 3D memory
- Parallelism: RAID, parallel data storage, DFS
- Data durability and consistency

Data Centers

- Racks of machines & Storage
- Data is growing faster than storage
- Energy consumption & cooling are big concerns
- Being built in colder areas and closer to cheap(er) energy sources

Energy

- Energy efficiency in data centers
  - Huge financial and environmental issue
- Data center construction from low-power computers
  - Think of a stack of tablets
  - Final project
- Data centers on renewable energy
  - Hydro, wind, solar
  - Stampede cooling

Data privacy & protection

- Misuse of big data is a big concern
  - A person's online activities can reveal all aspects of the person's life
- Systems need to provide clear guidelines on data privacy and protection
  - Sensitive clinical information
- Understand how the big data world operates
  - as an user
  - as a developer

Data Collection

Challenges of Big Data Collection

- Challenging to acquire a lot of data
  - Resources – memory, bandwidth, processing power
  - Parallelism
- What is useful?
  - Filter while acquiring
  - Identify redundancy
  - Collect topic-specific
- Challenging to collect from distributed sources

Web Crawling

- Collect published web content
  - Start with representative page
  - Parse content and follow hyperlinks
  - Recurse
- Why ?
  - Search engines
  - Advertisements
  - Extract additional information beyond what is provided by search engines

Goals

- Collect good-quality pages (content)
- Filter on the fly (low-content, redundancy)
- Parallel
  - Crawl efficiently (low client resources)
  - Crawl quietly (low server resources)
- Avoid parsing the same page more than once
- Predict page quality before retrieval

Redundancy removal

- Avoid parsing & following the same page more than once
  - Record all URLs that have been parsed & followed
  - Same page might have different URLs
- URL normalization
  - Host portion is case insensitive
  - Resolve percent encoding
- URLs that look completely different may also be the same page
  - Compute and match page content checksum

Other considerations

- Assume crawling is limited by some resource
  - Total time
  - Bandwidth / local storage
- Depth-first search or Breadth-first search
- Hyperlink with a high probability of pointing to a high-quality page
  - High in-links
  - Pointed from high-quality pages (BFS)
  - Spam identification
- Topic-specific crawling
  - Infer topic from text surrounding the link, keywords

Variations …

- Save only images
  - Students interested in image processing
  - Save all images on a topic
- Save pdf files
  - Collect academic papers on a given topic
- Audio, video, etc …

Scalable web crawling

- Resources
  - CPU → network operations, parsing the page content
  - I/O → write to disk
  - Network bandwidth
  - Memory?
- Parallel Web crawling
  - Use multiple CPUs
  - Use multiple storage devices
- Challenges
  - Synchronization
  - Network bandwidth → bottleneck

Crawler etiquette

- Crawling is not always welcome
- Robot exclusion standard (robots.txt)
- Have a polite mode
  - Limit crawler bandwidth
  - Limit per site

Questions?

Classroom & Permission Codes