CS226
Big-Data Management
Instructor: Ahmed Eldawy
Welcome (back) to UCR!
Class information

- Classes: Monday, Wednesday, Friday 2:10 – 3:00 PM at WCH 142
- Instructor: Ahmed Eldawy
- Office hours: Monday & Wednesday 4:00-5:00 PM @357 WCH. Conflicts?
- Website:
  - iLearn (Any UCRX students?)
- Email: eldawy@ucr.edu
  - Subject: “[CS226] …”
Course work

- Active participation in the class (5%)
- Reading and review tasks (10%)
- Class presentation (15%)
- Assignments (20%)
- Project (50%)
Project

- Groups of 3-4 students
- Group Selection
- Project proposal
- Literature survey
- Report outline
- Final report
- Project presentation
Course goals

» What are your goals?

» Understand what big data means

» Identify the internal components of big data platforms

» Recognize the differences between different big data platforms

» Explain how a distributed query runs on big data
Super Hero
Big-data Expert

- Understand how the big-data platforms really work
- Control those thousands of processors efficiently to carry out your task
Syllabus

- Overview of big data
- Big-data storage
- Big-data processing
- Big-data indexing
- Big-SQL processing
- Programming packages

09/28/2018
Introduction
All of the information

Information you need!
Jan 2012: World Economic Forum Report

Big Data, Big Impact: New Possibilities for International Development

The amount of data in the world is exploding - large portion of this comes from the interactions over mobile devices being used by people in the developing world - people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realize the potential for channeling these torrents of data into actionable information that can be used to identify needs & provide services for the benefit of low-income populations. This discussion note is a Call-to-action for stakeholders for concerted action to ensure that this data helps the individuals and communities who create it.
Interest in Big Data in the US

- **March 2012:** Obama administration unveils BIG DATA initiative: $200 Million in R&D investment

- **June 2013:** Washington Post is calling Obama “The Big Data President”

FOR IMMEDIATE RELEASE
March 29, 2012

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a “Big Data Research and Development Initiative.” By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation’s most pressing challenges.
Interest in Big Data in Europe

- March 2014: David Cameron and Angela Merkel talking about Big Data in a Computer Expo in Hannover, Germany
The creation and consumption of data continues to grow by leaps and bounds and with it the investment in big data technology, hardware, software and services to manage it.
Four Three V’s of Big Data

**Volume**
- Scale of Data
- 40 Zettabytes (45 billion gigabytes) of data will be created by 2020, an increase of 300 times from 2005
- 6 billion people have cell phones
- World population: 7 billion

**Velocity**
- Analysis of streaming data
- The New York Stock Exchange captures 1 TB of trade information during each trading session
- By 2016, it is projected there will be 18.9 billion network connections – almost 2.5 connections per person on earth

**Variety**
- Different forms of data
- It's estimated that 2.5 quintillion bytes (2.3 trillion gigabytes) of data are created each day
- Most companies in the U.S. have at least 100 terabytes (100 billion gigabytes) of data stored
- Dependent 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

**Veracity**
- Uncertainty of data
- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes)

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

**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

**Poor data quality costs the US economy around**
- $3.1 trillion a year

**There was unsure of how much of their data was accurate**
- 27% of respondents
Big Data Vs Big Computation

- Full scans (e.g., log processing)
- Range scans
- Point lookups
- Iterations
- Joins (self, binary, or multiway)
- Proximity queries
- Closures and graph traversals
Big Data Applications

- Web search
- Marketing and advertising
- Data cleaning
- Knowledge base
- Information retrieval
- Internet of Things (IoT)
- Visualization
- Behavioral studies
Publicly Available Datasets

- Data.gov
- Data.gov.uk
- Twitter Streaming API
- Yahoo! Webscope [http://webscope.sandbox.yahoo.com/]
- GDELT [http://www.gdeltproject.org/]
- Instagram API
Big Data Landscape 2012

© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

http://mattturck.com/2012/06/29/a-chart-of-the-big-data-ecosystem
Big Data Landscape 2014
