Chapter 1. Introduction

Chase Wu

Associate Professor
Department of Computer Science
New Jersey Institute of Technology
chase.wu@njit.edu

Collaborative Research Staff
Computer Science and Mathematics Division
Oak Ridge National Laboratory
wuqn@ornl.gov

The 1\textsuperscript{st} Class Attendance Check

- Name
- Program
- Year
- Why do you take this course?
- What is the largest data size you’ve ever dealt with and in what context?
About this course

• Recent Developments and Future Trends on Big Data Computing
• Overview of Big Data Analytics
• Big Data Analytics Platforms, Storage, and Algorithms
• Advanced Topics:
  • Big-Data Movement in High-performance Networks
  • Big-Data Workflow Management
Four V’s of Big Data

**Volume**

- 40 Zettabytes (42 trillion 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**

- 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.

**Variety**

- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes).
- 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.

**Veracity**

- 1 in 3 business leaders don’t trust the information they use to make decisions.
- In one survey, 27% of respondents were unsure of how much of their data was inaccurate.

**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.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
Exascale Computing and Big Data

By Daniel A. Reed and Jack Dongarra

July 2015 Communications of the ACM

https://vimeo.com/129742718
Thanks! ☺️

Questions ?