CSIE59830 Big Data Systems

Shiow-yang Wu (吳秀陽)
CSIE, NDHU, Taiwan, ROC

What is Big Data?

- The growth of data in volume, velocity, variety and veracity are in such an unprecedented scale that traditional database management systems can no longer handle it properly.
- We need new technologies, new systems and new tools!!

Examples of Big Data

- **Walmart**, the world's biggest retailer with over 20,000 stores in 28 countries, needs to process 2.5 petabytes of data every hour. (1PB=1024TB)
- **Facebook** processes data from more than 2 billion monthly active users worldwide. Every 60 seconds, 136,000 photos are uploaded, 510,000 comments are posted, and 293,000 status updates are posted. That amounts to 1000+ terabytes of data generated per day.

Why Big Data?

From business point of view:

- Big Data can unlock significant value by **making information transparent**.
- Big Data can help organizations collect more **accurate and detailed operational information** to expose variability and boost performance.
- Big Data allows **ever-narrower segmentation of customers** and therefore much more **precisely tailored products or services**.

Why Big Data?

- Sophisticated analytics can substantially improve decision-making, minimize risks, and unearth valuable insights that would otherwise remain hidden.
- Big Data can be used to develop the next generation of products and services.
- Big Data is a "Big Deal"!

What about this course?

- This is an introductory course on big data concepts and processing.
- You will also get hands on experience in using popular open source big data tools such as Hadoop, HBase, Spark, Hama, Storm, etc.
- More on big data processing technologies and systems, less on big data analytics.

Topics 1

- Introduction
  - What is a Big Data?
  - Why Big Data?
  - Examples of Big Data
  - The opportunities and challenges of Big Data
- General purpose big data systems
  - Distributed and cluster computing
  - MapReduce and Apache Hadoop
  - In-memory computation & Apache Spark

Topics 2

- Big data storage
  - Distributed filesystems and big data storage
  - Google GFS
  - Apache HDFS
  - Google BigTable system
- Big structured data processing
  - SQL or NoSQL
  - Apache HBase
  - Cassandra and MongoDB
  - Data Warehousing, Google BigQuery and Apache Hive

Topics 3

- Big graph processing
  - The challanges of big graphs
  - Pregel family of systems
  - GraphLab family of systems
- Big stream processing
  - The challenges of distributed big stream processing
  - Apache Flink
  - Apache Storm
  - Spark Streaming

Topics 4

- Big data analytics, other systems and trends**
  - Google Dremel, Apache Drill and Apache Impala
  - Google Cloud Platform (GCP) vs Amazon Web Services (AWS)
  - Open Data
  - Beyond Hadoop

Administrative Information

- Course Title: Big Data Systems
- Course Number: CSIE59830
- Lecture Time: Tue 09:10~12:00
- Classroom: Engineering Building C309
- Office Hours: Tue 17:00~18:00
- Grading Policy:
  - Assignments 35%
  - Independent Study and Presentation 15%
  - Final exam 25%
  - Term project 25%

Course Related Pages

- Course homepage: http://web.csie.ndhu.edu.tw/showyang/BigDataSys2018s/index.html
- Instructor's homepage: http://web.csie.ndhu.edu.tw/showyang/index.html
- All lecture notes will be available online.

References

- No required textbook.
- Kai Hwang and Min Chen. *Big Data Analytics for Cloud, IoT and Cognitive Computing*. John Wiley & Sons Ltd., 2017.
- Sherif Sakr. *Big Data 2.0 Processing Systems: A Survey*. Springer, 2016.
- Tom White. *Hadoop: The Definitive Guide, 4th Edition*, O'reilly, 2015.
- Jure Leskovec, Anand Rajaraman, Jeff Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2010~2014.
- Mohammed J. Zaki and Wagner Meira JR. *Data Mining and Analysis - Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.
- Donald Miner and Adam Shook. *MapReduce Design Patterns*, O'reilly, 2013.

Assignments

- There will be several programming assignments for you to get hands-on experience in big data processing.
- A lecture topic will start with its origin (mostly Google) and then its open source counterpart (mostly Apache).
- You will learn to use them on the department Hadoop cluster.
- Ask for an account from the TA.

Independent Study

- All students are to form independent study teams of 2~3 members.
- Each team should pick an open souce big data system not discussed in the class as the study target.
- Each team should prepare a presentation and a demonstration of the system.
- Every student must present and demo.

Exam

- No midterm exam. Term project proposal instead.
- There will be a final exam at the final exam week.
- The exam questions will be on the basic concepts, systems and applications of big data processing techniques.
- Only cover the lecture part.
- The final is an open book exam.
- No electronic devices allowed.

Term Project

- There will be a modest scale term project for you to show your creativity.
- You may use any big data systems for your project.
- You should turn in a project proposal by the end of the midterm exam week.
- There will be an end of the semester demo to explain your project to me.
- Turn in the project and report one week after the final exam.

Course Requirements

- Read the assigned readings before the class, participate in the discussion, ask questions!
- Learning by doing. Start early!
- You will need to learn to use UNIX and Python.
- Grading policy revisited:
  - Assignments (35%)
  - Independent sudy and presentation (15%)
  - Final exam (25%)
  - Term project (25%)

Resources

- Wikipedia, Big Data. (http://en.wikipedia.org/wiki/Big_data)
- Wikibook, Data Science: An Introduction. (http://en.wikibooks.org/wiki/Data_Science:_An_Introduction)
- Apache Hadoop (http://hadoop.apache.org/)