Big Data Management and Analytics

Lecture Notes
Winter semester 2016 / 2017
Ludwig-Maximilians-University Munich

© Prof. Dr. Matthias Renz 2015

Based on lectures by
Donald Kossmann (ETH Zürich), as well as
Jure Leskovec, Anand Rajaraman, and Jeff Ullman (Stanford University)
Course Logistics

• Course website:
  • http://www.dbs.ifi.lmu.de/cms/Big_Data_Management_and_Analytics
  • Registration for this lecture is now open via Uniworx
  • Registration required to attend the exams!!!

• Organization:
  • Load: 3+2 hours weekly
  • Required: Lecture "Database Systems I" or equivalent
  • Beneficial: Lecture "Knowledge Discovery in Databases I" or equivalent

• Lecture: Prof. Dr. Matthias Schubert

• Assisting: Daniyal Kazempour
Why this course?

• Big Data is big
  • $ and science: choose your poison
We are drowning in data … but starving for information

- Exponential grows in data

  $600$ to buy a disk drive that can store all of the world’s music

  5 billion mobile phones in use in 2010

  30 billion pieces of content shared on Facebook every month

  40% projected growth in global data generated per year vs.

  $5$ million vs. $400$ Price of the fastest supercomputer in 1975’ and an iPhone 4 with equal performance

  235 terabytes data collected by the US Library of Congress by April 2011


- Data contains value and knowledge

We are drowning in data … but starving for information

- Exponential grows in data

$600 to buy a disk drive that can store all of the world's music

5 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs.

$5 million vs. $400 Price of the fastest supercomputer in 1975 and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011


- Data contains value and knowledge

We are drowning in data ... but starving for information

- Exponential grows in data
  - $600 to buy a disk drive that can store all of the world’s music
  - 5 billion pieces of content shared on Facebook every month
  - 40% projected growth in global data generated per year vs.
  - $5 million vs. $400 growth in global
  - Data contains value and knowledge

Why this course?

• **Big Data is big**
  • $ and science: choose your poison
  • Big Data approaches required for Data Science “move data from raw to relevant”
Data Science (~eScience/Industry 4.0)

- **The Fourth Paradigm:**
  Age of data driven exploration
  → **Data Science** (eScience / Industry 4.0)

- **Science Paradigms**

  - **Empirical** — describing natural phenomena
  - **Theoretical** — using models, generalizations
  - **Computational** — simulating complex phenomena
  - **Data driven** — Data Science unify theory, experiment, and simulation

- Time
  - Thousand years ago
  - Last few hundred years
  - Last few decades
  - Today

**Formula:**
\[
\left(\frac{a}{a}\right)^2 = \frac{4\pi G p}{3} - K \frac{c^2}{a^2}
\]
Data Science (~eScience/Industry 4.0)

• The Fourth Paradigm:
  Age of data driven exploration
  → Data Science (eScience / Industry 4.0)

• Data Science
  • Data captured by instruments or generated by simulator
  • Processed by software
  • Information/knowledge stored in computer
  • Scientist/Analyst analyzes database / files using data management and statistics

[Informatik Pionier Jim Gray]
[Hey, Tansley, Tolle: Fourth Paradigm, 2009]
The Fourth Paradigm: Age of data driven exploration
→ Data Science (eScience / Industry 4.0)

Data Science

- Data captured by instruments or generated by simulator
- Processed by software
- Information/knowledge stored in computer
- Scientist/Analyst analyzes database / files using data management and statistics

“Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data.”

[Hey, Tansley, Tolle: Fourth Paradigm, 2009]
Why this course?

• **Big Data is big**
  • $ and science: choose your poison
  • Big Data approaches required for Data Science “move data from raw to relevant”

• **Big Data is exciting**
  • gives a new twist to almost everything
  • allows you to reinvent the wheel
Why this course?

- **Big Data is big**
  - $ and science: choose your poison
  - Big Data approaches required for Data Science “move data from raw to relevant”

- **Big Data is exciting**
  - gives a new twist to almost everything
  - allows you to reinvent the wheel

- **Big data is old**
  - opportunity to teach you some fundamental technology
Outline of this course

• Introduction (Motivation and Overview)
• Introduction to Big Data — the four V's
• NoSQL
• Hadoop / HDFS / MapReduce & Applications
• Spark
• Data Stream Processing & Applications & Algorithms
• High-Dimensional Data
• Graph Data Processing
  (Link Analysis, Page Rank, Community Detection)
This course is mainly based on a mixture of existing external lectures, Surveys, Papers and Reports on Big Data.

There is NO, or better, I’m not aware of a single book or script that is equivalent to this course (and addresses all issues discussed in this course).

Since Big Data is a quite new and hot topic, standards and basic concepts are quite dynamic => The Web is a very appropriate source of relevant information.

External lectures basically used for this course:
- Big Data: Donald Kossmann & Nesime Tatbul, Systems Group ETH Zurich - http://www.systems.ethz.ch/node/217

Further material will appear at our web page (check for updates during the course / open to further suggestions!)