BİL401/BİL501 Distributed Data Processing and Analysis «BigData»

Spring 2014

TOBB University of Economics and Technology

Department of Computer Engineering

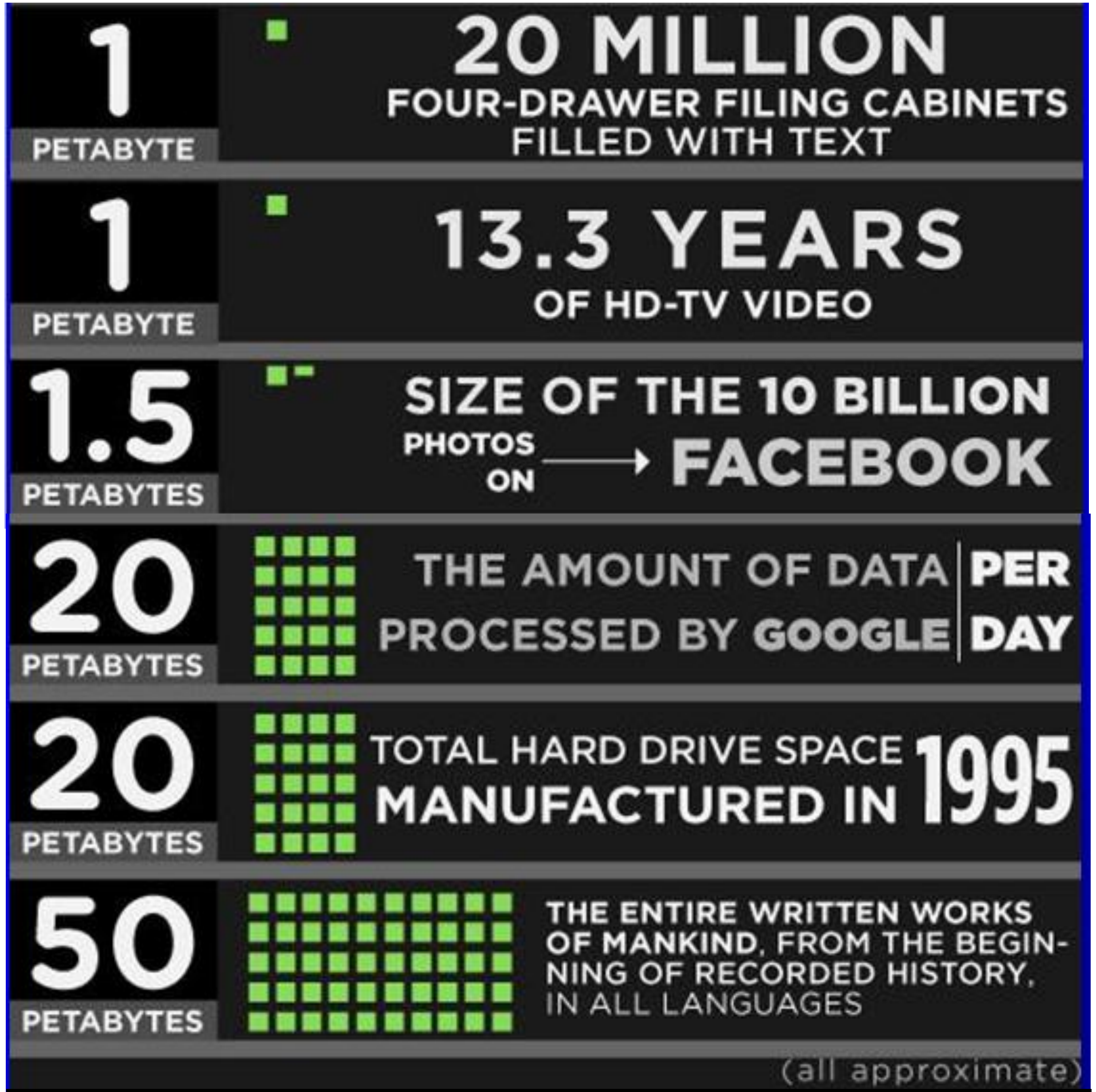
IBM TR

Outline

- Motivation for the course
- Course logistics
- Lecturers
- Schedule
- Evaluation

How big is Big Data?

- *1 byte = 8 bit*
- *1 MB = 10^6 B = 1 million byte*
- *1 GB = 10^9 B = 1 billion byte*
- *1 TB = 10^{12} B*
- *1 PB = 10^{15} B = 250.000 DVD*



1 PETABYTE — 20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT

1 PETABYTE — 13.3 YEARS OF HD-TV VIDEO

1.5 PETABYTES — SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK

20 PETABYTES — THE AMOUNT OF DATA PROCESSED BY GOOGLE | PER DAY

20 PETABYTES — TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995

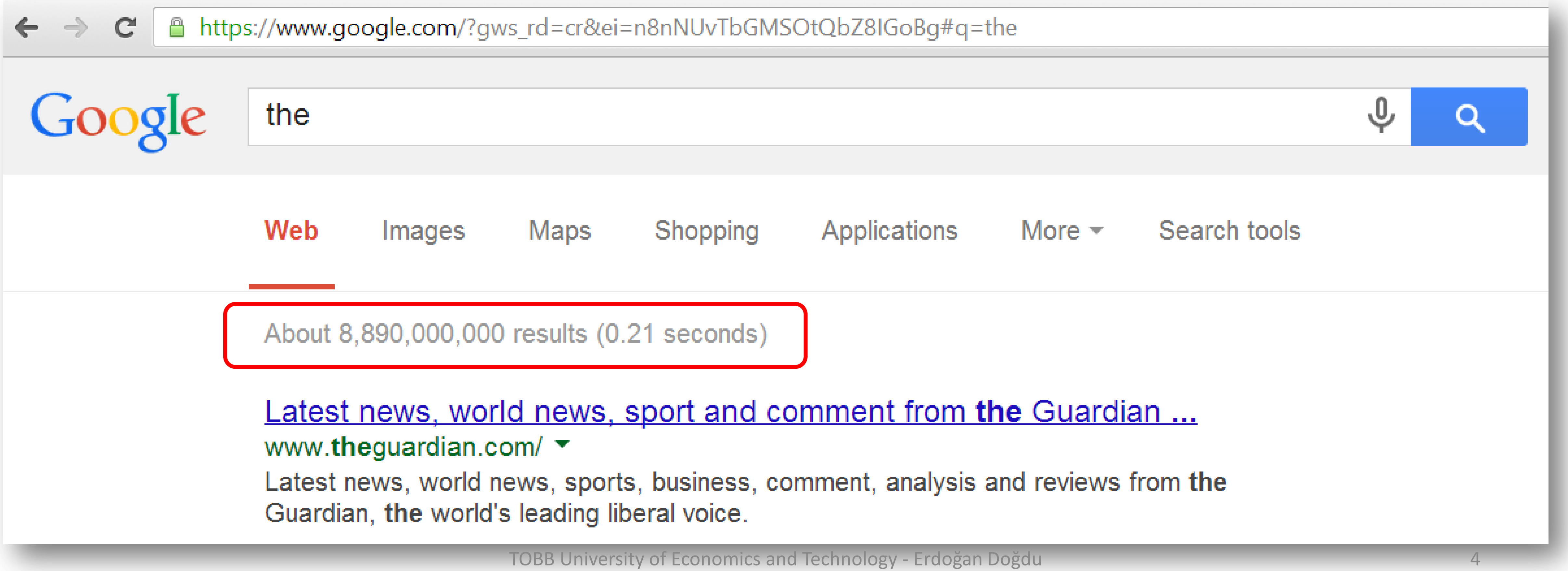
50 PETABYTES — THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

(all approximate)

Facebook currently stores more than 100 petabytes of data.

Where is Big Data?

- Web pages. How many?



https://www.google.com/?gws_rd=cr&ei=n8nNUvTbGMSOtQbZ8IGoBg#q=the

the

About 8,890,000,000 results (0.21 seconds)

Latest news, world news, sport and comment from **the** Guardian ...

www.**the**guardian.com/

Latest news, world news, sports, business, comment, analysis and reviews from **the** Guardian, **the** world's leading liberal voice.

Web in numbers

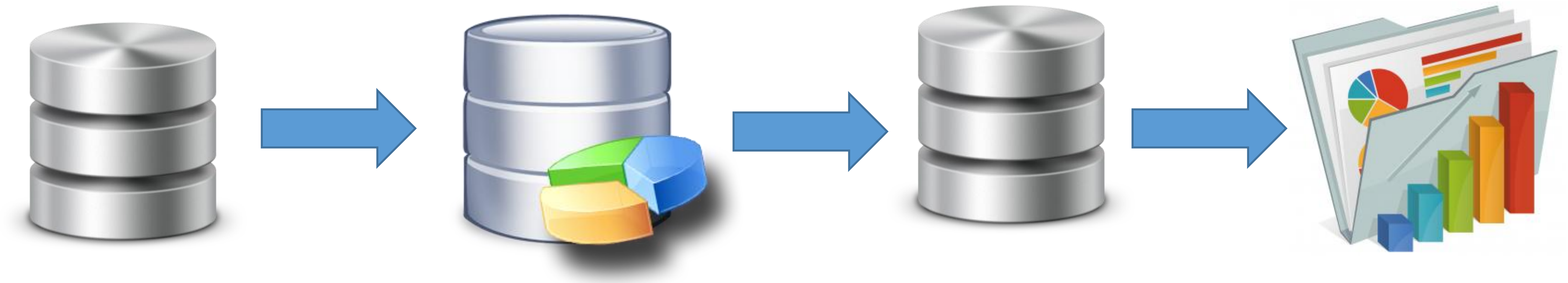
- Facebook, 1 billion users (Sep 2013)
- Twitter, 200 million users, 400 million tweets daily (60% from mobile devices) (Sep 2013)
- Google, 100 billion queries a month (May 2013)

In constrast, typical large enterprises:

- 5.000-50.000 servers
- Terabytes of data, millions of tx/day

Big data technology

- Traditional «**business intelligence**» using databases



DB

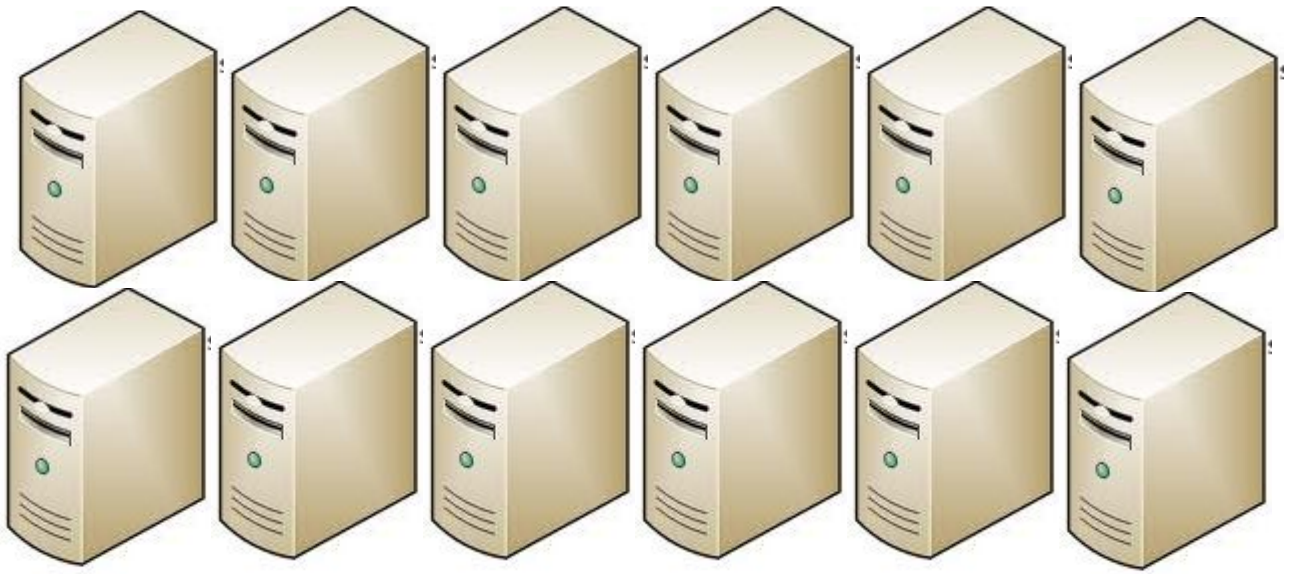
Datawarehouse

More DB

Reports

Big data technology

- Facebook, Twitter, LinkedIn, eBay, Amazon did not use «traditional databases» for big data
  - Massive **parallelism**
  - **Map-Reduce** paradigm



Web intelligence using big data

- Online advertisement – predicting interest
- Consumer sentiment – predicting behavior
- Detecting events – predicting impact
- Intelligent question answering – Watson, Google knowledge graph
- Categorizing, recognizing people, faces, people
- Intelligent public services – smart grids, water distribution, etc.
- Analysing **all** email and watching Web activity – predicting terrorists
- …

Data analytics

- Data → Information
- Finding patterns
- Classification
- Predicting
- Data mining
- Business intelligence
- Data analytics on big data
    - Applying known methods in parallel on distributed data

Big Data Jobs

- **10 hot job titles** that did not exist 5 years ago
- ***LinkedIn** study on **259 million members** (November 2013)*

1. iOS Developer
2. Android Developer
3. Zumba Instructor
4. Social Media Intern
5. **Data Scientist**
6. UI/UX Designer
7. **Big Data Architect**
8. Beachbody Coach
9. Cloud Services Specialist
10. Digital Marketing Specialist

- http://talent.linkedin.com/blog/index.php/2014/01/top-10-job-titles-that-didnt-exist-5-years-ago-infographic

Course

- Thursday 10:30 / two hour lecture
- Friday 08:30 / lecture or **lab**

- **Lecture**
  - IBM experts, Erdoğan Dogdu, Murat Özbayoğlu
- **Lab**: TM107
  - IBM Tools: IBM BigInsights

Course objectives

- Understand big data **concepts**
- Learn **distributed data processing algorithms**, techniques and methods on big data.
- Learn **data analysis methods** on big data.

Learning outcomes

- Write **map/reduce** methods to process big data
- Use advanced **distributed data processing techniques** and tools on big data
- **Develop map/reduce based applications** for processing big data
- Understand big data **analysis methods** and techniques
- Choose appropriate big data analysis methods for specific big data problems and apply

Textbook and Resources

- *Harness the Power of BigData*, McGraw-Hill, 2013 http://public.dhe.ibm.com/common/ssi/ecm/en/imm14100usen/IMM14100USEN.PDF
- *Understanding the BigData*, McGraw-Hill, 2012 http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF
- *Hadoop for Dummies*, Robert D. Schneider, Wiley, 2012 http://public.dhe.ibm.com/common/ssi/ecm/en/dcm03002usen/DCM03002USEN.PDF
- *Hadoop Documentation,* https://hadoop.apache.org/docs/r1.2.1/index.html
- *Big Data University*, http://bigdatauniversity.com

Topics

- Map-Reduce
- Hadoop
- Storage, Indexing
- BigData in-motion, Real Time Analytics

Grading

Work	%
Assignments	20%
Exam (midterm)	25%
Participation	3%
Attendance	2%
Project/Research	50%