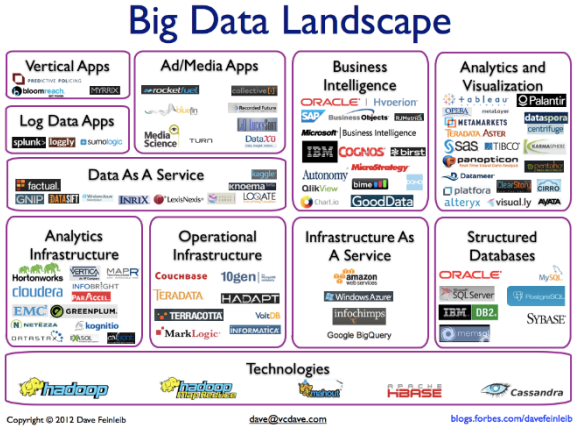


Seminar on big data management

Lecturer: Jiaheng Lu

Spring 2017

We are in the era of big data

- Lots of data is being collected
  - Web data, e-commerce
  - Bank/Credit Card transactions
  - Social Network
  - Scientific data

Data sizes

- Byte (B)
- Kilobyte (KB)
- Megabyte (MB)
- Gigabyte (GB)
- Terabyte (TB)
- Petabyte (PB)
- Exabyte (EB)
- Zettabyte (ZB)
- Yottabyte (YB)

How much data?

- Google processes 20 PB a day (2008)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN atomic facility generates 40TB per second.

- In 2009, total data is about 1ZB, in 2020, it is estimated to be 35ZB.

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

Four V's


BIG DATA
VOLUME
DATA SIZE
VELOCITY
SPEED OF CHANGE
VARIETY
DIFFERENT FORMS OF DATA SOURCES
VERACITY
UNCERTAINTY OF DATA

- Watch videos about big data

Outline

- About the seminar
- Practical information and requirement
- Seminar topics
- Our schedule

The seminar is about

- Big data management

- Data querying, exploration, sampling, sharing, cleansing, cloud data management, big data benchmark and applications.

At the end of the seminar

- You should be able to tell what these terms stand for! And more…

Hadoop Mapreduce

Cassadra

NOSQL

Spark RDD

Pig, Hive, Mahout

HBase, Flume

Volume, Velocity, Variety

RDF, KB

Jaccard, Cosine, Edit distance

Cloud data management

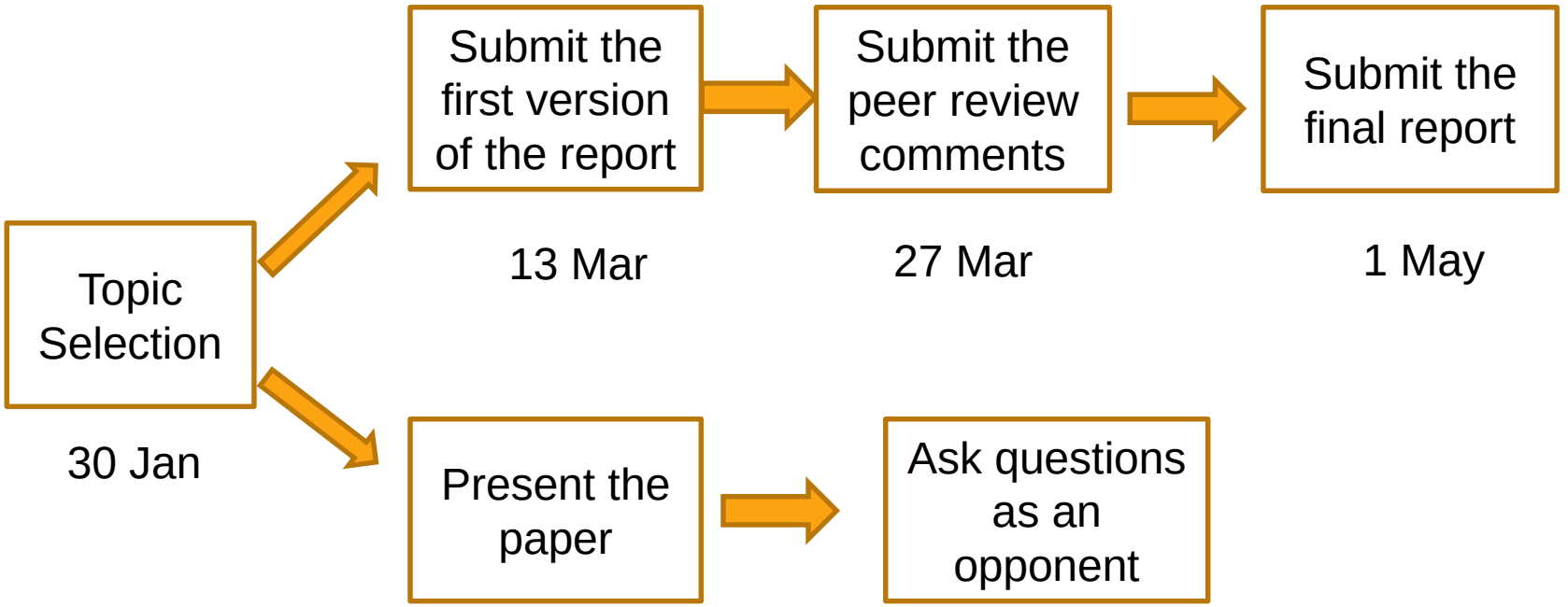
After this seminar

- Students are expected to
  - Have a decent understanding of big data challenge
  - Conduct research on one of topics related to big data management
  - Know how to read/write/review a technical paper
  - Know how to present a paper

More formally

- **Pick** a topic from the offered topics
- **Read** papers on that topic
- **Present** the paper
- **Write** a report on the topic
- **Review** two other reported written by your classmates
- **Ask** questions as an opponent for the presentation by your classmates
- **Attend** the lectures (at least 80%)

Deadlines for each task



Topic assignment

- **Submit your list- the preferred 3 topics**
- **If you have something in mind which is not listed in, please send an email to the teacher**

- **Same topics will be assigned to more than one person.**

Start researching your topics immediately after topic assignment

Topics of this seminar

- Big data survey
- Hadoop and Spark platforms
- Cloud data management
- Graph data management
- Data sampling
- Data exploration
-

Topics of this seminar

- Approximate data processing
- Data cleansing
- Knowledge base
- Big data benchmark
- Big data applications

Hadoop and Spark platforms

- Two open-sources platforms for big data processing

Spark
Lightning-Fast Cluster Computing

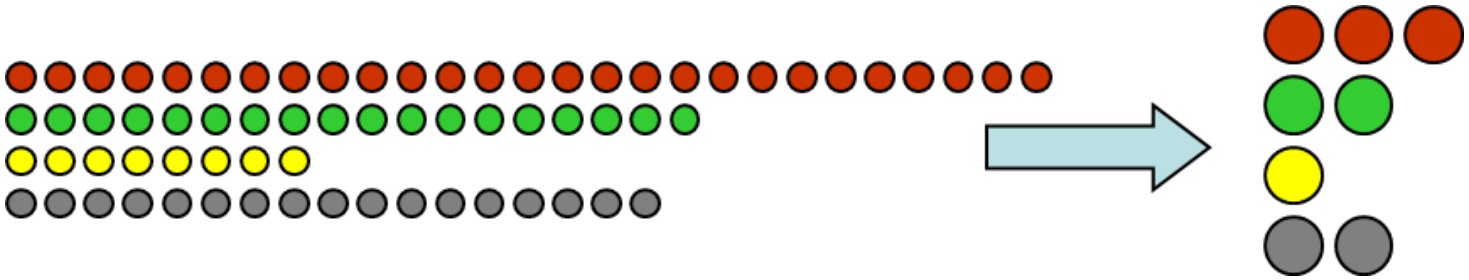
Cloud data management

- Cloud data management is to deploy database systems in the cloud.

- New challenges:
- Data is stored at an untrusted host
- Data is replicated across large geographic distances
- Compute power is elastic

Data sketches and sampling

- It is not always possible to store the big data in full
  - Many applications (telecoms, ISPs, search engines) can’t keep everything
- It is inconvenient to work with data in full
- It is faster to work with a compact summary
  - Better to explore data on a laptop than a cluster



Graph data management

- Graph data management has long been a topic of interest for database researchers.

- New application domains for big data including social networks and the Web of data.

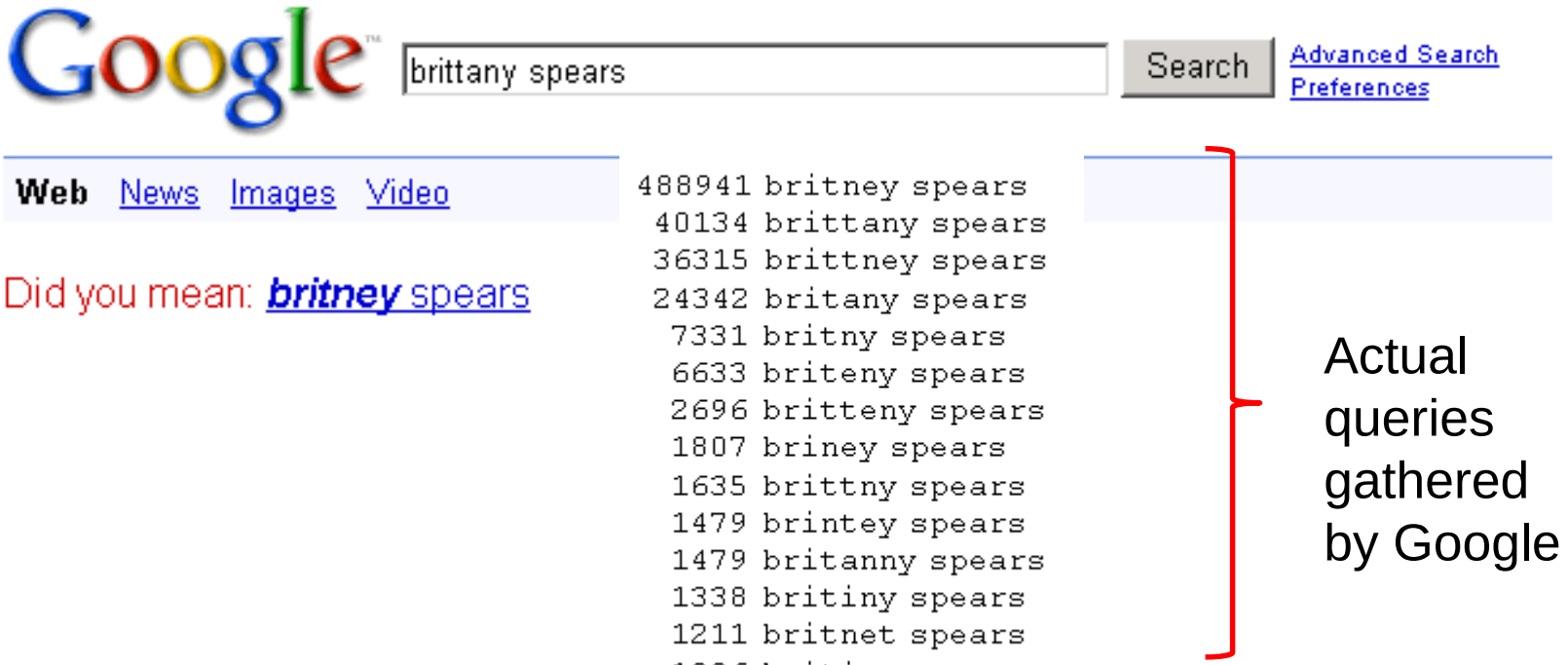
Data exploration

- Data exploration is about efficiently extracting knowledge from big data even if we do not know exactly what we are looking for.

- Topics:
  - Query Result Visualization
  - Query by example
  - Approximation query processing
  - Interactive interface

Approximate string processing

- String data is ubiquitous. Approximate string processing tolerates the error with string matching.



Data cleansing

- Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Example:

Gender	Frequency
2	1
F	12
M	13
X	1
f	2

Knowledge base

- A knowledge base (KB) contains a set of concepts, instances, and relationships.

- Applications:
  - query understanding
  - Deep Web search
  - In-context advertisement
  - Event monitoring in social media
  - Product search, and social mining.

Big data benchmark

- Create a standard benchmark to assist in the evaluation of different big data systems.

- Performance
- Scale-up
- Elastic speedup
- Availability

Big data applications

- Big data will have many applications in different areas:
  - Science and research
  - Public health
  - Customer relation management
  - Machine and Device Performance
  - Security and Law Enforcement
  - Optimizing Cities and Countries

Task for Next week

- Perform 1st pass on the papers in the seed papers list – All papers available on the course home-page

- Select interesting papers for your presentation and report

Conclusion

- Big data seminar focuses on the presentation and paper writing skill training on the topic of big data
- Select the topic of your interests and start to work as soon as possible
- Please enter your topic preferences here:
- http://goo.gl/forms/wa4aGImDfE
- Please enter your available time for presentation and opponent here:
- http://goo.gl/forms/SsR56FTuf6