

Advanced Data Engineering

Haruo Yokota

Office: West 8(E)-705 (ex. 3505)

Email: yokota@cs.titech.ac.jp

What is Data Engineering?

- An academic/research area managing “Data”
 - *How to store, search for, retrieve and process data efficiently*
- The word “Data Engineering” has been used since 1980’s
 - As society’s name (Japan) / conference’s name (US)
- Tightly related to “Databases”
 - The concept of Data Engineering is much broader
- Other expressions:
 - Management of Data (also name of a group in society)
 - Data Intensive Processing (vs. Computation Intensive)

2019/6/13

Advance Data Engineering (©H. Yokota)

2

Targets of Data Engineering (DE)

- One of the most essential technologies for the information era
 - We need data (information)
 - Information is the source of power
 - Alvin Toffler, “The Third Wave”, 1980
 - We need to keep the data for a long time
 - We need to retrieve the data efficiently
- The amount of information increase rapidly
 - So called “*Information Explosion*” or “*Big Data*”

2019/6/13

Advance Data Engineering (©H. Yokota)

3

Big Data is 'Buzzword du Jour'; CS Academics 'Have the Best Job'

- Communications of the ACM
 - Michael Stonebraker (MIT)
- Big data can mean:
 - Big volumes of data, but "small analytics"
 - Big analytics on big volumes of data
 - Big velocity
 - Big variety
- We can do many things based on database technologies

<http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext>

2019/6/13

Advance Data Engineering (©H.Yokota)

4

Activities in Data Engineering

- This research area is very active!
- Want to know activities of some research area
 - Keep your eyes on *conferences* or other *events* of related societies
 - Both international and domestic
 - They reflect activities of the research area
 - They are also good indicators to find recent trends of research

2019/6/13

Advance Data Engineering (©H.Yokota)

5

Top 3 Major Int'l Conferences

- **SIGMOD**: International Conference on Management Data
 - Annual conference from 1975 sponsored by ACM
- **VLDB**: International Conference on Very Large Data Bases
 - Annual conference from 1975 sponsored by VLDB Endowment
- **ICDE**: International Conference on Data Engineering
 - Annual conference from 1984 Sponsored by IEEE Computer Society

2019/6/13

Advance Data Engineering (©H.Yokota)

6

Other Int'l Conferences

- **PODS**: Principles of Database Systems
- **EDBT**: International Conference on Extending Database Technology
- **DASFAA**: Database Systems for Advanced Applications
- **DEXA**: Database and Expert Systems Applications
- **ER**: International Conference on Conceptual Modeling
 - Formerly: International Conference on the Entity Relationship Approach
- **ICDT**: International Conference on Database Theory
- **ICDM**: International Conference on Data Mining
- ...

There are many conference related to data engineering

2019/6/13

Advance Data Engineering (©H.Yokota)

7

Japanese Related Organizations

- Related Organizations in Japan
 - [Special Interest Group on Database Systems](#) (1973-)
 - In IPSJ (The Information Processing Society of Japan)
 - [Technical Group on Data Engineering](#) (1986-)
 - In IEICE (The Institute of Electronics, Information and Communication Engineers)
 - [ACM SIGMOD Japan Chapter](#) (1994-)
 - [DBSJ: The Database Society of Japan](#) (2002-)

2019/6/13

Advance Data Engineering (©H.Yokota)

8

Related Domestic Events

- **DEIM Forum** (formerly DEWS from 1990)
 - Data Engineering and Information Management
 - DEIM2018: in Awara, Fukui, 3/4-3/6, 2018
 - DEIM2019: in Huistenbosch, Nagasaki, 3/4-3/6, 2019
- **WebDB Forum** (formerly DBWeb from 2000)
 - WebDB2018: at **Tokyo Tech**, 9/12-14, 2018
 - WebDB2019: at Kogakuin Univ., 9/8-9/9, 2019
- **SoC** (Social Computing Symposium)
 - SoC2018: at LIFULL (Kojimachi) 6/22-23, 2018
 - SoC2019: at Ochanomizu W. U. 6/22, 2019

2019/6/13

Advance Data Engineering (©H.Yokota)

9

Dynamism of Research Area

- Disappeared (Several Years Ago)
 - E-commerce, Materialized Views, XML
- Current Trend / New Area
 - Graph Database
 - Social Media and Crowd
- Still Remain
 - Data Mining, Data on the Cloud
- Continuously appeared
 - Traditional Database Technologies
 - Indexing, Query Processing, Storage, Privacy, Security

2019/6/13

Advance Data Engineering (©H.Yokota)

10

In This Course

- Both Traditional DB & New DE Topics
 - Data Warehouse & OLAP (Data Mining)
 - Storage & Advanced Indexing
 - Cost Estimation of Data Engineering Operations
 - Parallel Database Processing
 - Skew Handling
 - Distributed Database Operations
 - Databases in the Cloud
 - NoSQL & Key Value Storage
 - XML Database, RDF Database

2019/6/13

Advance Data Engineering (©H.Yokota)

11

Background Requirements

- Assuming basic knowledge of
 - Relational Databases
 - Relational Algebra, SQL, etc.
 - Transaction Processing Concept
 - ACID Property, Logging, etc.
 - Computer Architecture
 - Parallel & Distributed Computer Architecture
- Prepare Short Course of Databases

2019/6/13

Advance Data Engineering (©H.Yokota)

12

Reference Books

- Since these topics are up-to-date, there is no text book describing them. Just for reference.
- "Transaction Processing: Concept and Techniques,"
 - Jim Gray and Andreas Reuter,
 - Morgan Kaufmann Publishers, 1993
 - ISBN: 1-55860-190-2
- "An Introduction to Database Systems",
 - C. J. Date,
 - Addison-Wesley, 7th edition, 2000
 - ISBN: 0-201-38590-2
- "A First Course In Database Systems",
 - J. D. Ullman, J. Widom,
 - Prentice Hall, 1997,
 - ISBN: 0-13-861337-0

2019/6/13

Advance Data Engineering (©H.Yokota)

13

Additional Information

- Evaluation
 - Assignments (given in each class) and Final Report
- OCW
 - <http://www.ocw.titech.ac.jp/>

2019/6/13

Advance Data Engineering (©H.Yokota)

14
