Advanced Data Management Technologies

Unit 1 — Introduction

J. Gamper

Free University of Bozen-Bolzano
Faculty of Computer Science
IDSE

Acknowledgements: I am indebted to Michael Böhlen and Stefano Rizzi for providing me their slides, upon which these lecture notes are based.

Outline

Outline

Course Organization

- Course page
  - http://www.inf.unibz.it/dis/teaching/ADMT or https://ole.unibz.it
  - Here you can find the schedule, lecture notes, office hours, etc.

- Organization
  - The course consists of lectures and a project
  - Lectures are organized as frontal teaching classes
  - The lab is organized as a project
  - Lab hours are used to discuss with you the progress in the project
  - We also plan an excursion to a company

Course Content

- The course introduces advanced data management technologies:
  - Data Warehousing and Business Intelligence
  - OLAP
  - Extract, Transform, Load
  - NoSQL
  - Main-memory Databases
  - Column-oriented Databases
  - Distributed Databases
  - Distributed Data Structures

- The course is research-oriented
  - Many concepts we discuss are not available in commercial (DBMS) systems
  - There is no single course book; much of the material is based on research papers

Exam

- The assessment of the course consists of two parts:
  - theory (60%): assessed with a written exam at the end;
  - project (40%): assessed through a presentation, demo and final report about the project.
- Both parts must be positive to pass the exam.
- A positive project is required for attending the theory part.
- The final grade is the weighted average between the two parts.

Outline

Literature and Resources

- Journal Publications
  - ACM Transaction on Database System (TODS)
  - IEEE Transactions on Knowledge and Data Engineering (TKDE)
  - The VLDB Journal
  - Information Systems
- Conference Publications
  - ACM SIGMOD International Conference on Management of Data (SIGMOD)
  - International Conference on Very Large Databases (VLDB)
  - International Conference on Extending Database Technology (EDBT)
  - IEEE International Conference on Data Engineering (ICDE)
- DB & LP Bibliography (maintained by Michael Ley, Uni Trier, Germany)
  - http://www.informatik.uni-trier.de/~ley/db/

(Commercial) Products

- Oracle
- DB2 (IBM)
- Microsoft SQL Server
- Teradata
- Sybase
- Ingres
- Informix
- PostgreSQL
- PC "DBMSs": Paradox, Access, ...
- ...

DB Research and Practice has Many Aspects

- Design of languages
- Development of algorithms
- Data modeling
- User interface design
- Design of migration strategies
- Distributed computing
- High data volumes and efficiency
- New data models and systems
  - XML/semi-structured databases
  - Temporal, spatial, moving object databases
  - Stream data processing
- ...

The Relational Data Model/1

- Data are stored in relations/tables

employee

Name	**Dept**	**Salary**
Tom	SE	23K
Lena	DB	33K

department

Dname	**Manager**	**Address**
SE	Tom	Boston
DB	Lena	Tucson

project

PId	**Dept**	**From**	**To**
14	SE	01.01.2005	31.12.2005
173	SE	15.04.2005	30.10.2006
201	DB	15.04.2005	31.03.2006

- SQL as query (and data definition) language
  - Intergalactic dataspeak [Stonebreaker]

The Relational Data Model/2

- A domain D is a set of atomic data values.
  - phone numbers, names, grades, birthdates, departments
  - each domain includes the special value `null` for unknown or missing value
- With each domain a data type or format is specified.
  - 5 digit integers, yyyy-mm-dd, characters
- An attribute A_i describes the role of a domain in a relation schema.
  - PhoneNr, Age, DeptName
- A relation schema $R(A_1, ..., A_n)$ is made up of a relation name R and a list of attributes.
  - $employee(Name, Dept, Salary)$, $department(DName, Manager, Address)$
- A tuple t is an ordered list of values $t = (v_1, ..., v_n)$ with $v_i \in dom(A_i)$.
  - $t = (Tom, SE, 23K)$
- A relation r of the relation schema $R(A_1, ..., A_n)$ is a set of n-ary tuples.
  - $r = \{(Tom, SE, 23K), (Lene, DB, 33K)\}$
- A database DB is a set of relations.
  - $DB = \{r, s\}$
  - $r = \{(Tom, SE, 23K), (Lene, DB, 33K)\}$
  - $s = \{(SE, Tom, Boston), (DB, Lena, Tucson)\}$

Properties of Relations

- A relation is a set of tuples, i.e.,
  - no ordering between tuples and
  - no duplicates (identical tuples) exist.
- Attributes within tuples are ordered.
  - At the logical level it is possible to have unordered tuples if the correspondence between values and attributes is maintained
  - e.g., $\{Salary/23K, Name/Tom, Dept/SE\}$

DB Interfaces

- The success of DBs also depends on the ease of data access.
- Various interfaces to DBs exist, e.g.,
  - Terminal interface (sqlplus, etc.)
  - OCI (Oracle Call Interface)
  - X/Open SQL CLI (Call Level Interface)
  - ODBC (Open Data Base Connection), iODBC for Unix
  - JDBC (Java Database Connectivity)
  - DBI (Perl DB Interface)
  - Embedded SQL

Outline

New Trends/1

- In the light of new trends, the relational model is not sufficient anymore!
- At least three interrelated megatrends in the last few years
  - Big Data
  - Big Users
  - Cloud Computing

New Trends/2

- **Big Data**
  - Database volumes have grown continuously since the earliest days of computing, but that growth has intensified dramatically over the past decade
  - e.g., social networks, Facebook, Google, geo location data, sensor-generated data, scientific data, Internet of Things, Industry 4.0, etc.
  - Huge data repositories, e.g., in astronomy, finance, Web, ...

- **Big Users**
  - Not long ago, 1,000 daily users was a lot and 10,000 was an extreme case.
  - Today, millions of users a day is not uncommon, and users have very different needs.
  - As a consequence, developers need more flexibility to store/access the data.
- **Cloud Computing**
  - Has placed new challenges on the database.
  - Provide computing resources on demand with a "pay-as-you-go" model.
  - Traditional RDBMSs were unable to provide these types of elastic services.

The Need for Advanced Data Management Technologies/1

- With the increase in data and users, applications have changed dramatically over the last 15 years, and so have the data management needs of those apps.
- Relational databases are schema-based, hence rather rigid; new more flexible and scalable data models are needed.
- ACID properties are not always needed; scalability is more important!
- Data is distributed, thus database solutions are needed that are distributed on large numbers of hosts across a network.

$\Longrightarrow$ We study new data management technologies in the second part of the course.

The Need for Advanced Data Management Technologies/2

- On the other hand, the immense value of data has been recognized by businesses.
- Thus, analysis and mining of data has become an important tool in decision making for most businesses.
- An exponential increase in operational data has made computers the only tools suitable for providing data for decision-making performed by business managers.
- The massive use of techniques for analyzing enterprise data made information systems a key factor to achieve business goals.

$\Longrightarrow$ We study Business Intelligence and Data Warehousing in the first part of the course.

Summary

- New trends in the last few years: big data, big users, cloud computing
- With the increase in data and users
  - new data management technologies are needed: more flexible, scalable, relaxed ACID
  - businesses recognized the immense value of data for decision making