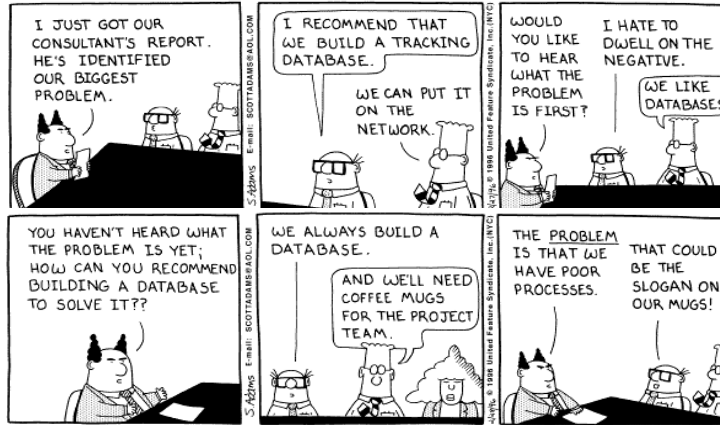


CS5208: Foundations on Database Systems

(<http://www.comp.nus.edu.sg/~cs5208>)



Copyright © 1996 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited

“Knowledge is of two kinds: we know a subject ourselves,
or we know where we can find information upon it.”

-- Samuel Johnson (1709-1784)

CS5208

1

Lecture 1

Introduction & Data Design and Modeling

CS5208

2

What: Database Systems Today

Ask Jeeves®
Search the web

match.com®

Members [sign in here](#)

[How it works](#) | [Success stories](#) | [Subscribe Now](#)

FREE FOR A LIMITED TIME
HOW TO FIND THE RIGHT PERSON IN 90 DAYS
A NEW STEP BY STEP GUIDE

A NEW STEP BY STEP GUIDE
Get your **FREE GUIDE** Today! [GO](#)

Love is complicated. match is simple.

Start browsing now. It's free.

I am a seeking between and
located in (city/zip code) [Browse](#)

Watch our new TV ads >>>

What: Database Systems Today

Application - SAP NetWeaver Portal - Microsoft Internet Explorer

Address: https://www.nus.edu.sg/portal/

NUS **MyNUS**

Home Staff Academic Research International Customer

Exec, Prof & Non-Acad Faculty & Research

Application

Detailed Navigation

- Exec, Prof & Non-Acad
- Update Personal Data
- Salary
- Leave
- Application
 - Overview / Cancellation
 - Dept Calendar
 - Report
 - Help
- Conflict of Interest Annual C
- Performance Management
- Performance Management
- Medical & Healthcare
- Travel/External Training and
- Housing Benefits Annual D
- Request for ...
- Others

Portal Favorites

Name:

Absence type:

Absent from (dd mm yyyy):

To (inclusive) (dd mm yyyy):

Duration of leave: Days

Approver:

Notes:

[Post](#)

[Replace Approver](#)

[Check](#) [Submit](#)

Note: Please provide the contact information below during your leave if it is not the same as the HR record at [MyNUS Portal > Staff Matters > Update Personal Data > Address & Contact](#).

Overseas Addr:

Email Addr:

Fax No:

Phone No:

Existing entitlements	Deductible to	Entitlement	Used	Planned & approved*	Leave Balance	Unit
Pro-rated Vacation Leave	31.12.2009	28.00	0.00	1.00	88.00	Days

* Cases pending approval are excluded.

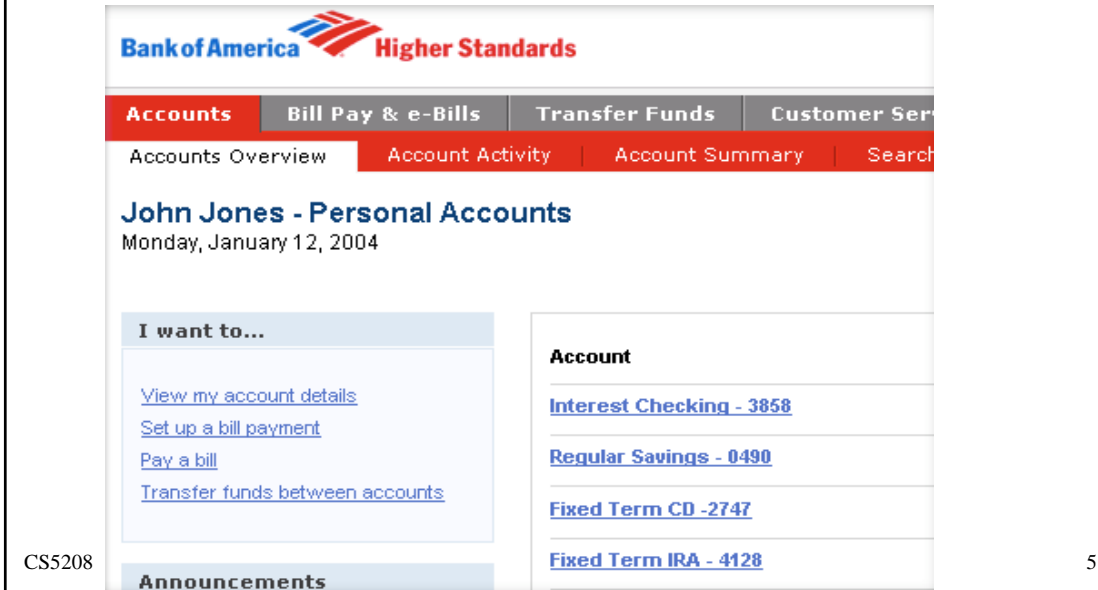
Note: Leave is calculated on a calendar year basis and will have to be pro-rated, where applicable, based on the staff member's period of resident service with the University.

If you encounter problems using this facility, please contact the NUS IT Care at tel: 68742000 or email: itc@nus.edu.sg

CS52

4

What: Database Systems Today



The screenshot shows the Bank of America website interface. At the top is the Bank of America logo with the tagline "Higher Standards". Below the logo is a navigation bar with tabs: "Accounts", "Bill Pay & e-Bills", "Transfer Funds", and "Customer Service". Under the "Accounts" tab, there are links for "Accounts Overview", "Account Activity", "Account Summary", and "Search". The main heading is "John Jones - Personal Accounts" with the date "Monday, January 12, 2004".

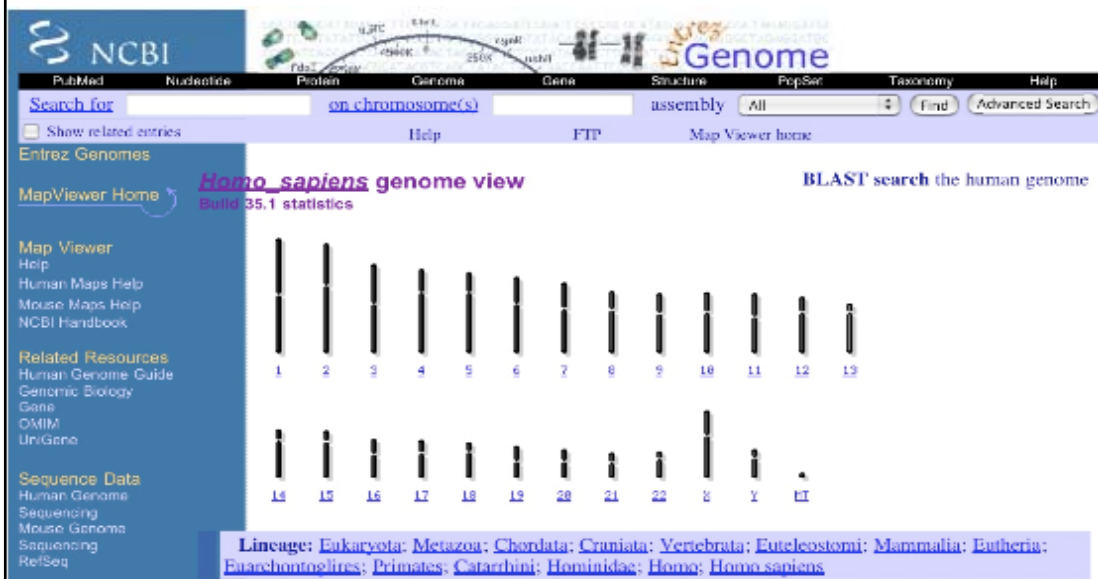
On the left, under "I want to...", there are links: "View my account details", "Set up a bill payment", "Pay a bill", and "Transfer funds between accounts".

On the right, under "Account", there is a list of accounts: "Interest Checking - 3858", "Regular Savings - 0490", "Fixed Term CD -2747", and "Fixed Term IRA - 4128".

At the bottom left, there is a section for "Announcements".

CS5208 5

What: Database Systems Today



The screenshot shows the NCBI Entrez Genomes website. The top navigation bar includes links for "PubMed", "Nucleotide", "Protein", "Genome", "Gene", "Structure", "PopSet", "Taxonomy", and "Help". The "Genome" link is highlighted.

Below the navigation bar, there is a search bar with the text "Search for" and a dropdown menu showing "on chromosome(s)". There are also buttons for "assembly", "All", "Find", and "Advanced Search".

On the left, there is a sidebar with links for "Entrez Genomes", "MapViewer Home", "Map Viewer", "Help", "Human Maps Help", "Mouse Maps Help", "NCBI Handbook", "Related Resources", "Human Genome Guide", "Genomic Biology", "Gene", "OMIM", "UniGene", "Sequence Data", "Human Genome", "Sequencing", "Mouse Genome", "Sequencing", and "RefSeq".

The main content area displays the "Homo sapiens genome view" with a "Build 35.1 statistics" link. It shows a map of the human genome with chromosomes 1 through 22, X, and Y. Each chromosome is represented by a vertical bar with a scale. Below the map, there is a line of text: "Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo; Homo sapiens".

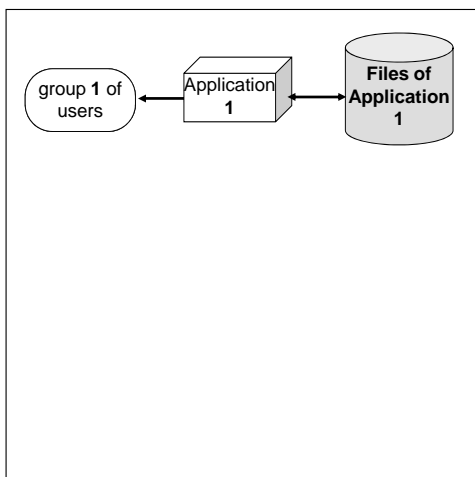
What Is a Database System?



- Database:
a very large, integrated collection of data.
- Models a real-world enterprise
 - Entities (e.g., course, instructor)
 - Relationships
(e.g., Tan *teaches* Database Technology)
- A Database Management System (DBMS) is a software system designed to **store, manage, and facilitate access to** databases.

CS5208

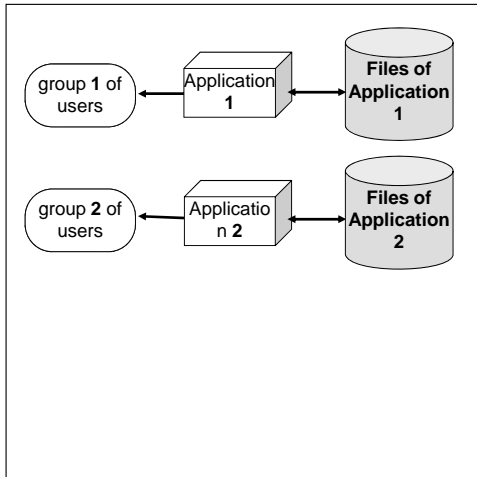
File-based vs database approach



CS5208

8

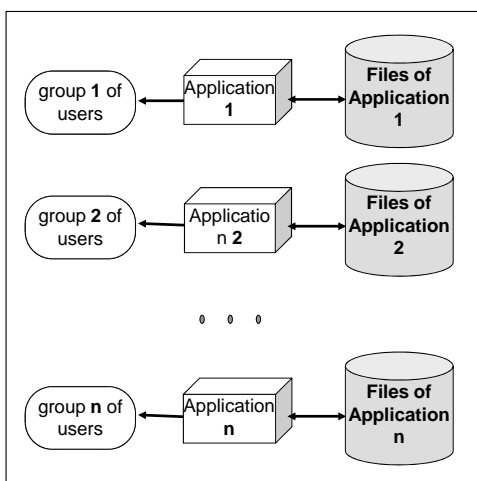
File-based vs database approach



CS5208

9

File-based vs database approach



CS5208

10



Is a File System a DBMS?

- Thought Experiment 1:

- You and your project partner are editing the same file.
- You both save it at the same time.
- Whose changes survive?

A) Yours B) Partner's C) Both D) Neither E) ???

- Thought Experiment 2:

- You're updating a file.
- The power goes out.
- Which of your changes survive?

Q: How do you write programs over a subsystem when it promises you only "???" ?

A: Very, very carefully!!

A) All B) None C) All Since last save D) ???

CS5208

11

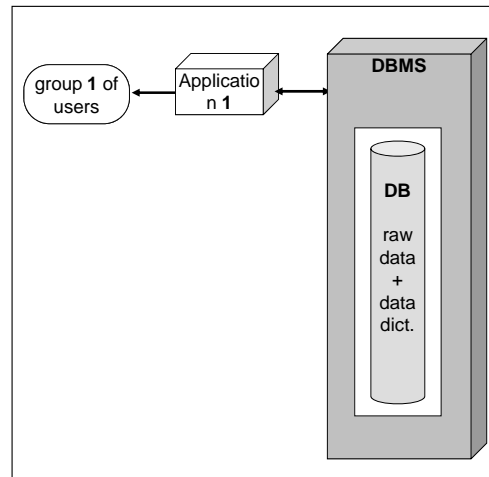
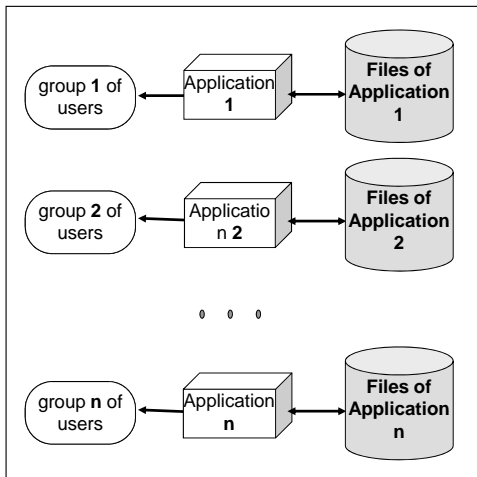
OS Support for Data Management

- Data can be stored in RAM
 - this is what every programming language offers!
 - RAM is fast, and random access
 - Isn't this heaven?
- Every OS includes a File System
 - manages *files* on a magnetic disk
 - allows *open, read, seek, close* on a file
 - allows protections to be set on a file
 - drawbacks relative to RAM?

CS5208

12

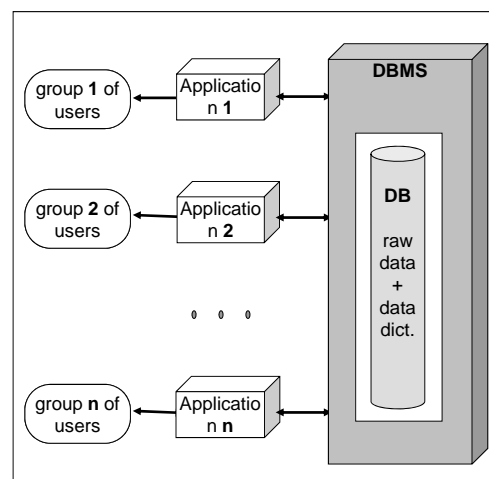
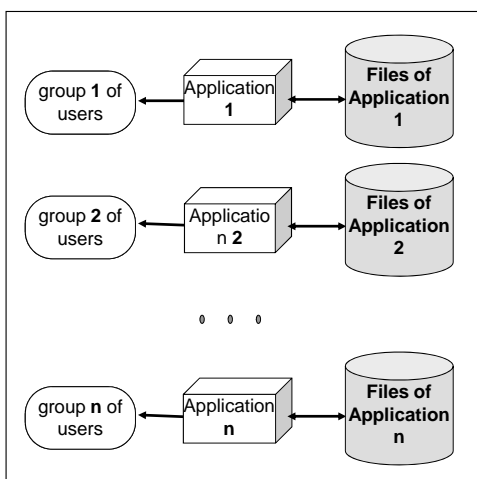
File-based vs database approach



CS5208

13

File-based vs database approach



CS5208

14

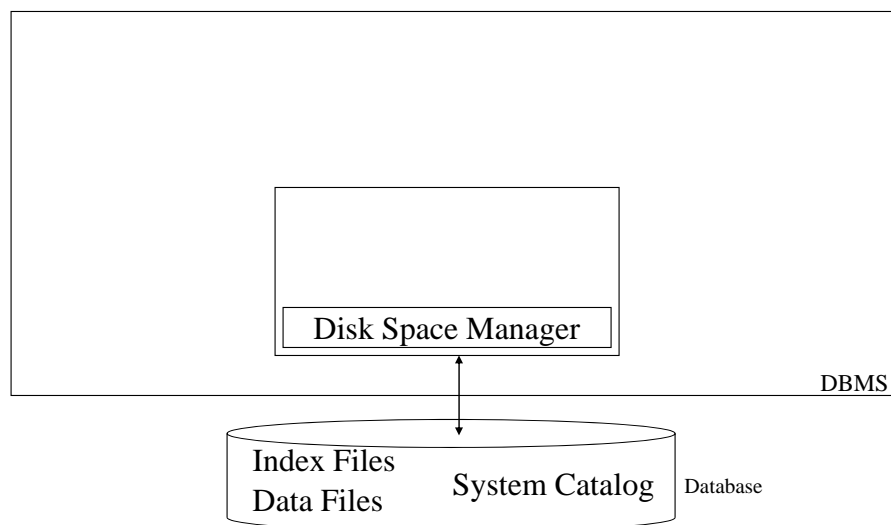
Capabilities of a Modern DBMS

- **Persistence** - permanent storage of data
- **Efficiency** - manage *large* volumes of data and *ad-hoc* queries efficiently
- **High-level access** - data model & language for defining database structures, retrieval and manipulation
- **Transaction management** - provide correct, concurrent access to the database by many users at once
- **Access control** - limit access by unauthorized users
- **Integrity management** - assure compliance to known constraints imposed by application semantics
- **Resiliency** - ability to recover from system failures without losing data

CS5208

15

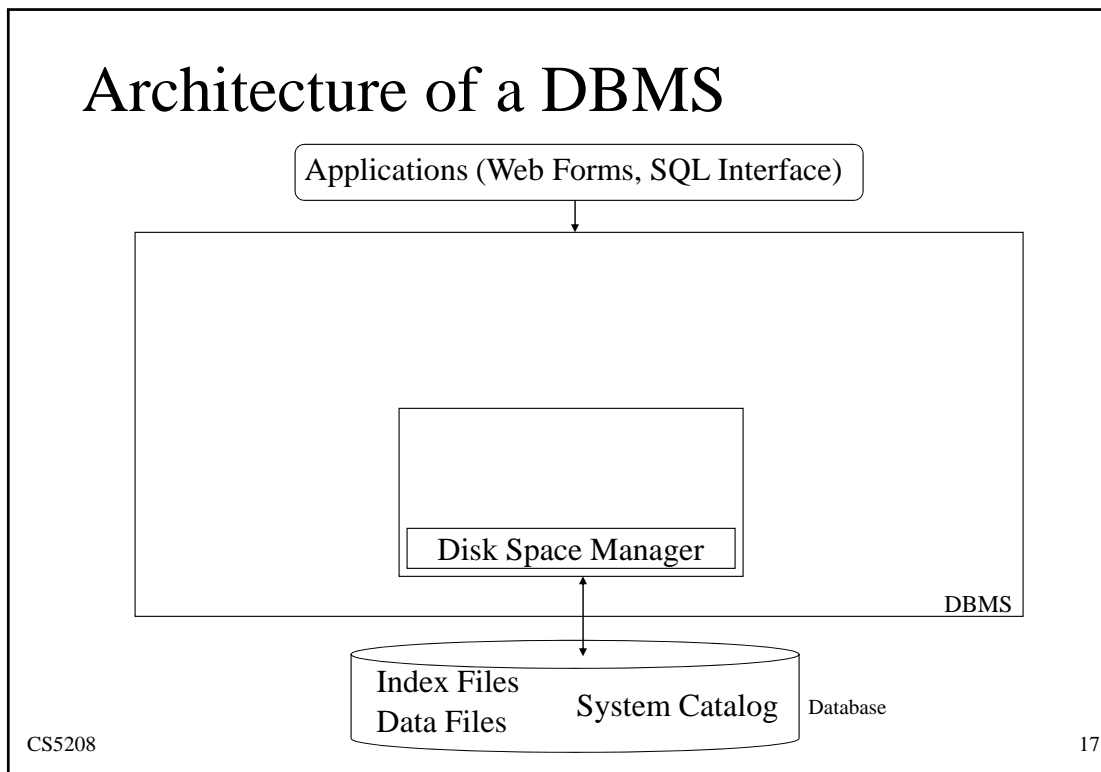
Architecture of a DBMS



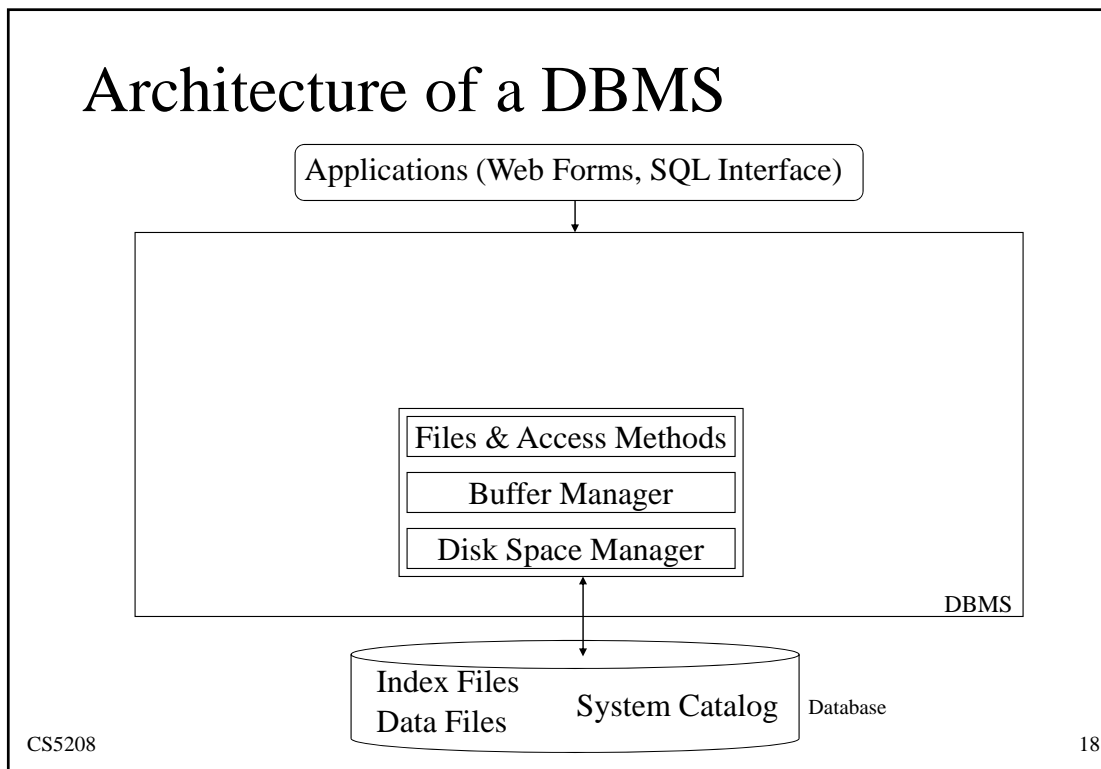
CS5208

16

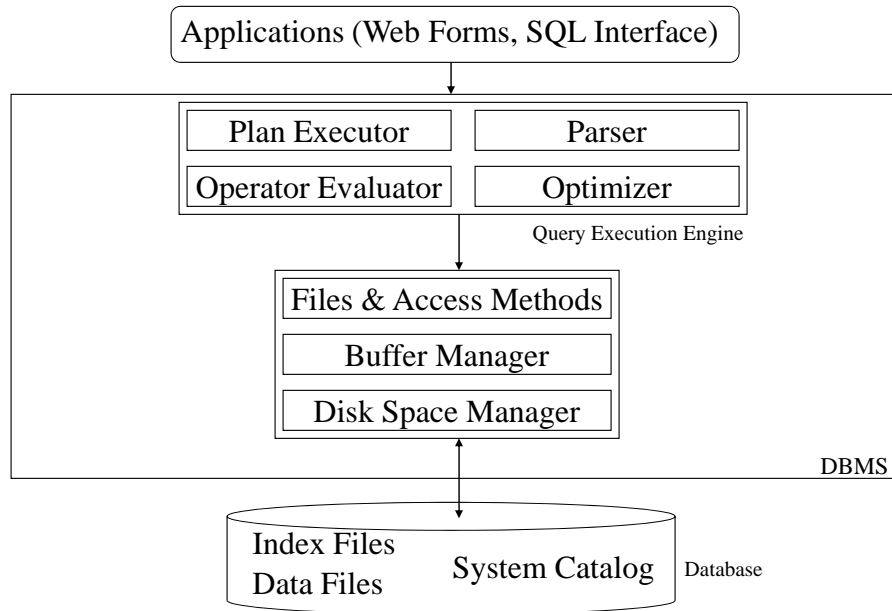
Architecture of a DBMS



Architecture of a DBMS



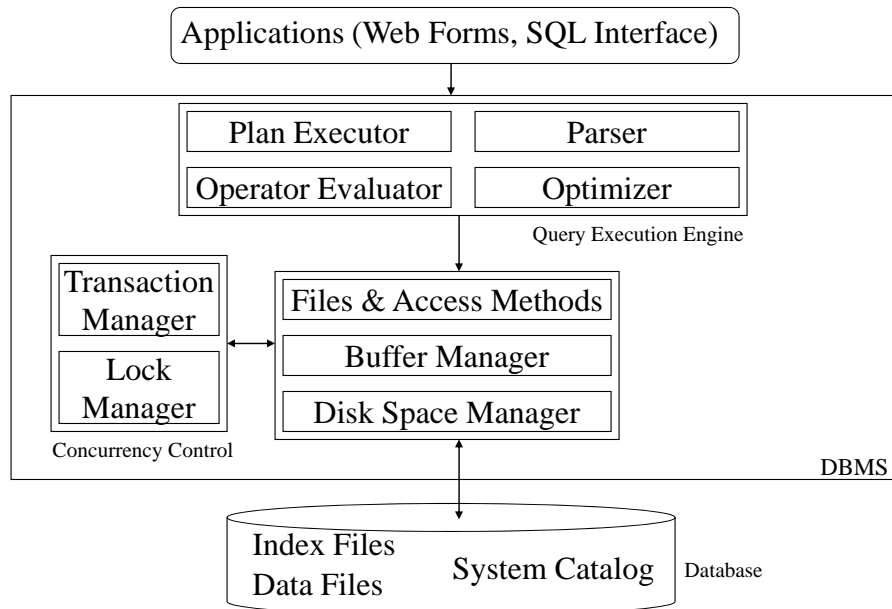
Architecture of a DBMS



CS5208

19

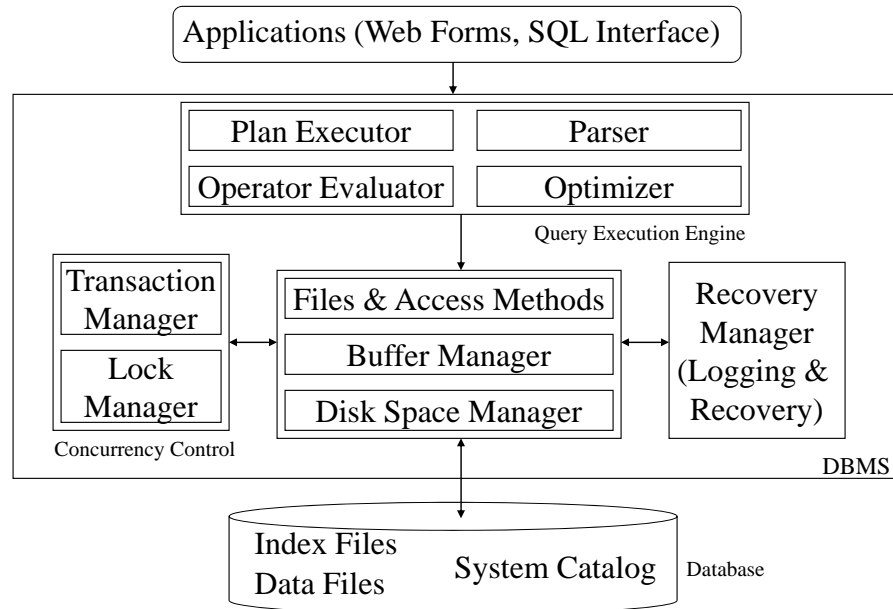
Architecture of a DBMS



CS5208

20

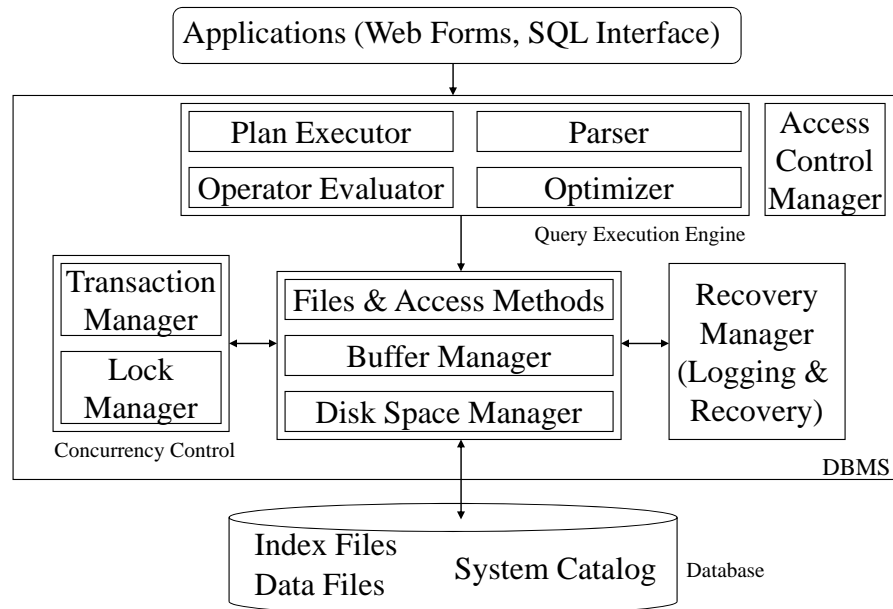
Architecture of a DBMS



CS5208

21

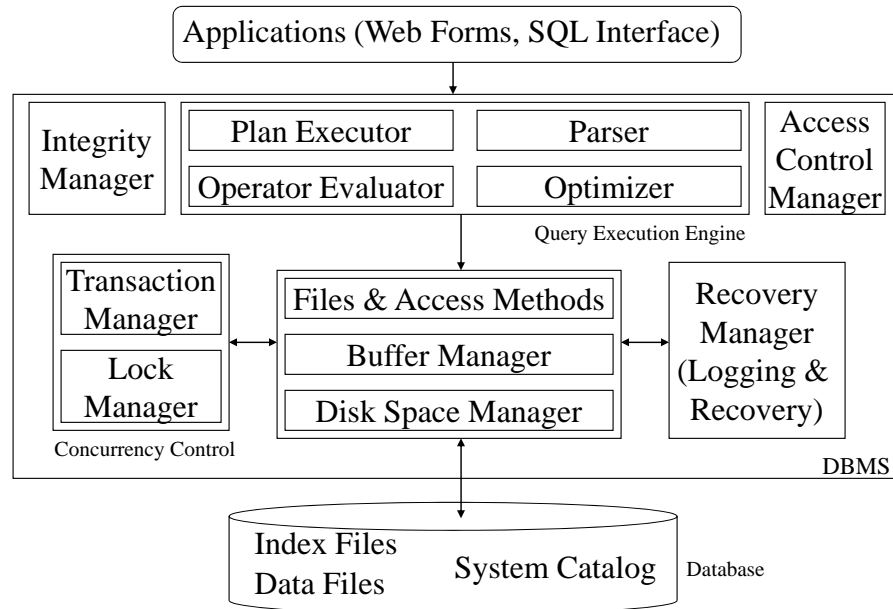
Architecture of a DBMS



CS5208

22

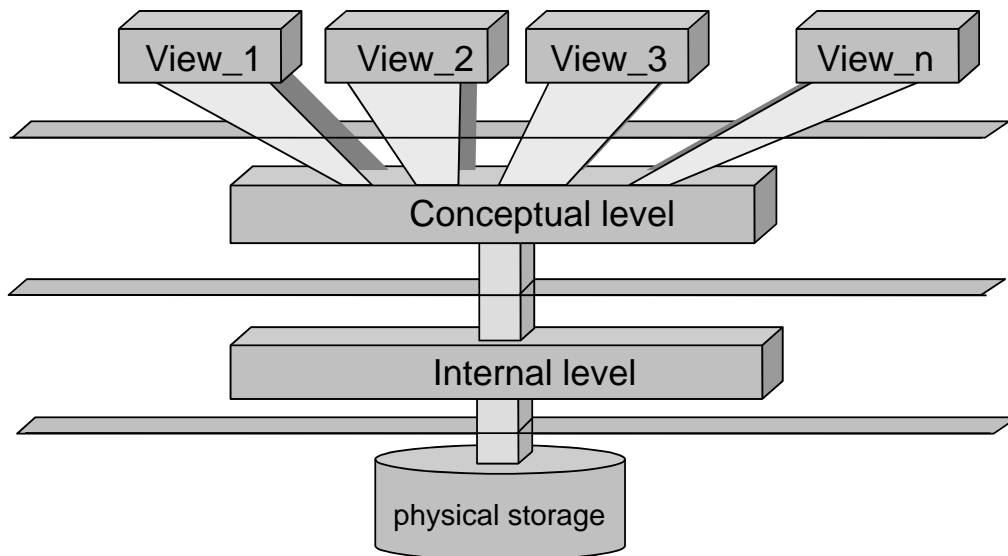
Architecture of a DBMS



CS5208

23

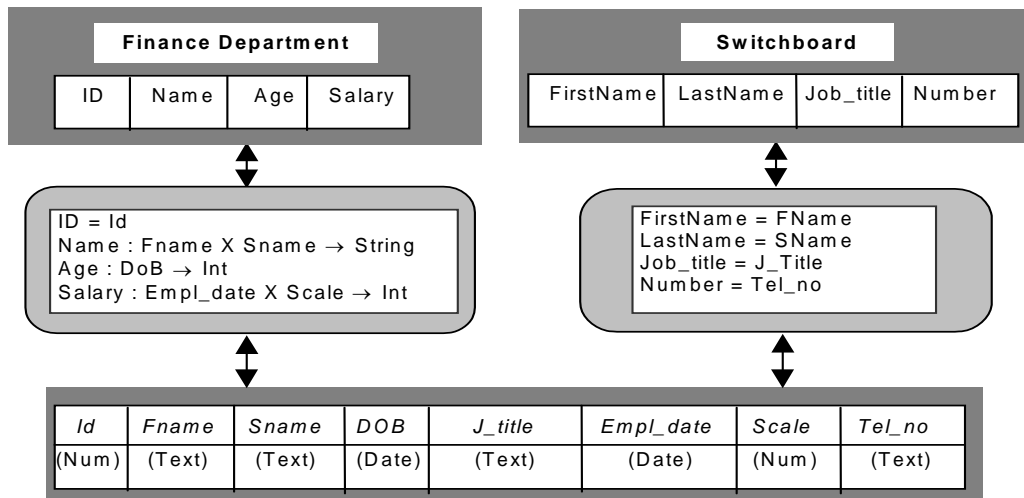
3-level Abstraction



CS5208

24

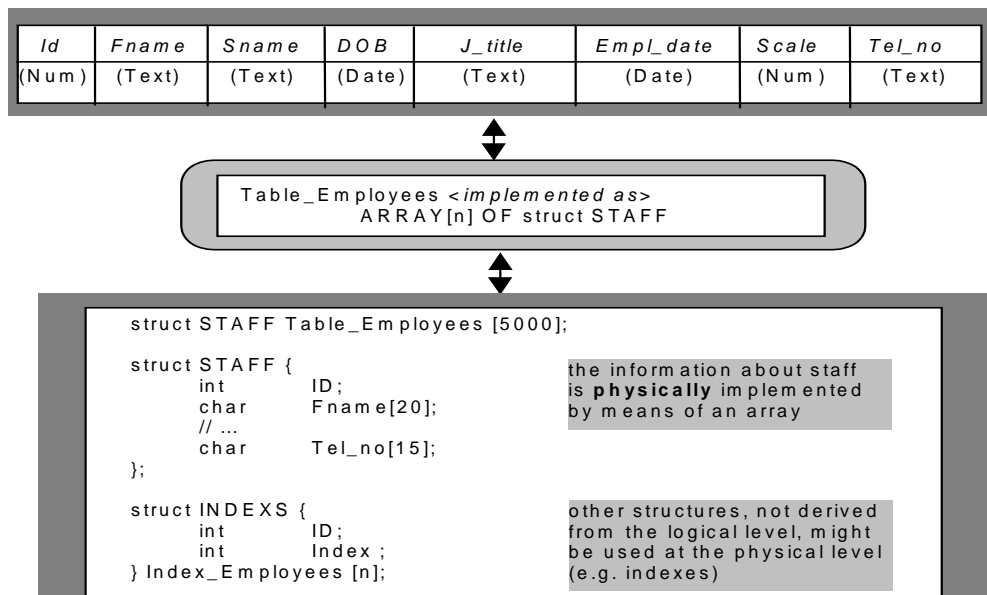
External / conceptual example



CS5208

25

Conceptual / internal - example



CS5208

26

Data Independence

- Applications insulated from how data is structured and stored.
- Ability to modify a schema definition in one level without affecting a schema definition in the next higher level.
- The interfaces between the various levels and components should be well defined so that changes in some parts do not seriously influence others.
- *Logical* and *physical* data independence

CS5208

27

Current Commercial Outlook

- A major part of the software industry:
 - Oracle, IBM, Microsoft
 - also Sybase, Informix (now IBM), Teradata
 - smaller players: java-based dbms, devices, OO, ...
- Well-known benchmarks (esp. TPC)
- Lots of related industries
 - data warehouse, document management, storage, backup, reporting, business intelligence, ERP, CRM, app integration
- Relational products dominant and evolving
 - adapting for extensibility (user-defined types), adding native XML support.
 - Microsoft merging file system/DB for "longhorn" (abandoned?)
- Open Source coming on strong
 - MySQL, PostgreSQL, BerkeleyDB

CS5208

28

Why Study Databases??



- Shift from computation to information
 - “Big-Data” phenomenon
 - always true for corporate computing
 - Web made this point for personal computing
 - more and more true for scientific computing
- Need for DBMS has exploded in the last years
 - Corporate: retail swipe/clickstreams, “customer relationship mgmt”, “supply chain mgmt”, “data warehouses”, etc.
 - Scientific: digital libraries, Human Genome project, NASA Mission to Planet Earth, physical sensors, grid physics network
- DBMS encompasses much of CS in a practical discipline
 - OS, languages, theory, AI, multimedia, logic
 - Yet traditional focus on real-world apps

CS5208

Intellectual Outlook: Research Trends

- Heavy weight DBMS
 - Extend existing DBMS capabilities for advanced applications
- Light weight DBMS
 - Component-based DBMS
 - Build and use what are necessary
- Autonomic & Self tuning DBMS
 - Making the DBMS “intelligent” to minimize maintenance cost

CS5208

30

Databases make these folks happy ...

- DBMS vendors, programmers
 - Oracle, IBM, MS, Sybase, NCR, ...
- End users in many fields
 - Business, education, science, ...
- DB application programmers
 - Build enterprise applications on top of DBMSs
 - Build web services that run off DBMSs
- Database administrators (DBAs)
 - Design logical/physical schemas
 - Handle security and authorization
 - Data availability, crash recovery
 - Database tuning as needs evolve



...must understand how a DBMS works

CS5208

Database Design

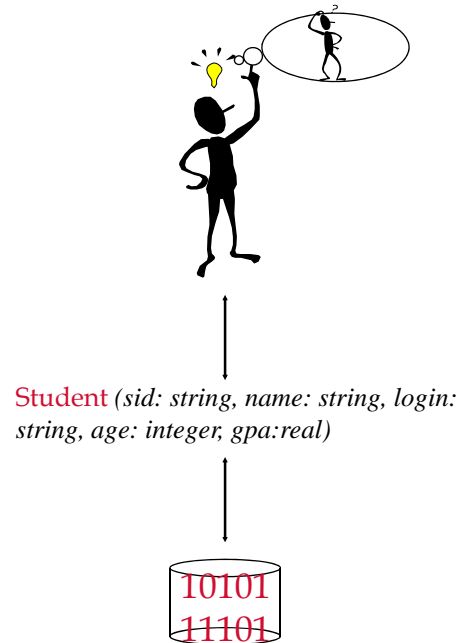
- Why do we need it?
 - Agree on structure of the database before deciding on a particular implementation.
- Issues:
 - What to model?
 - How are things related?
 - What constraints exist?
 - How to achieve *good* designs?

CS5208

32

Data Models

- DBMS models real world
- *Data Model* is link between user's view of the world and bits stored in computer
 - A tool for describing data, data relationships, data semantics and data constraints
- Many models exist
 - Relational Model
 - Entity-Relationship Model
 - Object-oriented Model



CS5208

33

Entity Relationship Model

- Diagrams to represent designs

Entity = object of interest

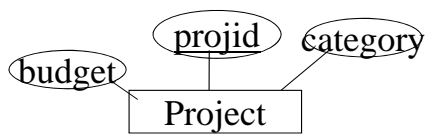
Entity set = set of similar entities.

Attribute = property of entities in an entity set

Relationship = association among entity sets

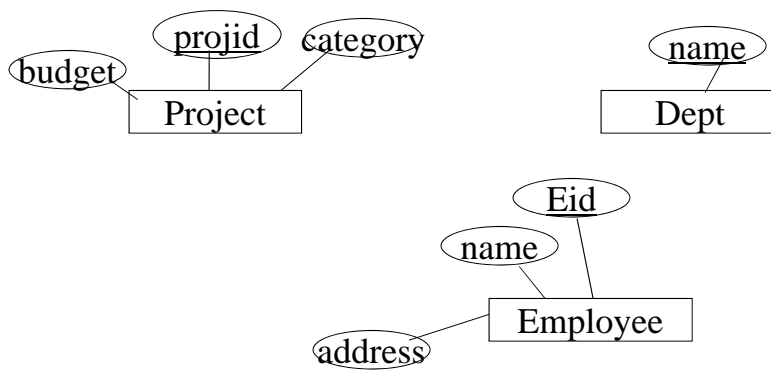
CS5208

34



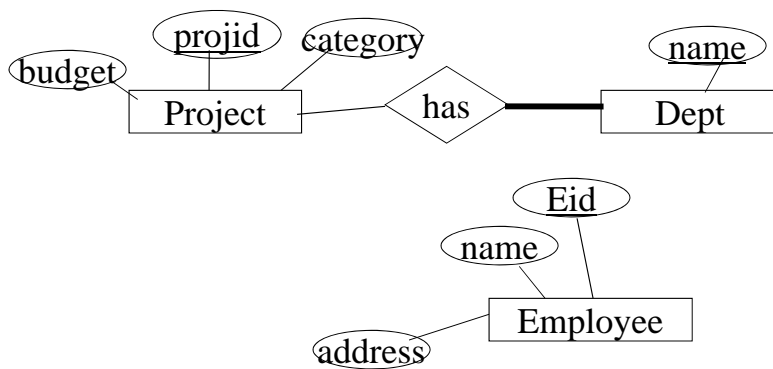
CS5208

35



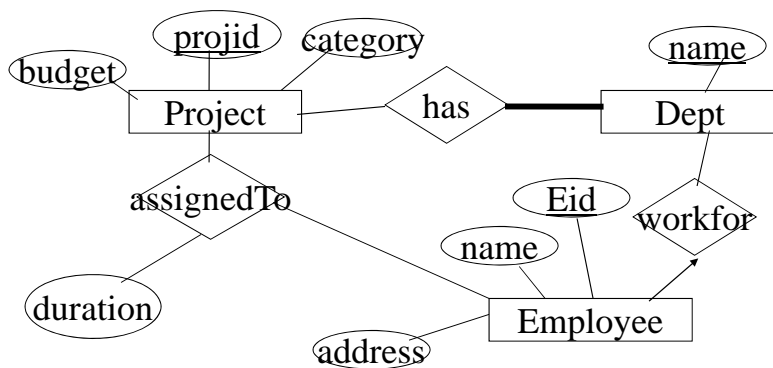
CS5208

36



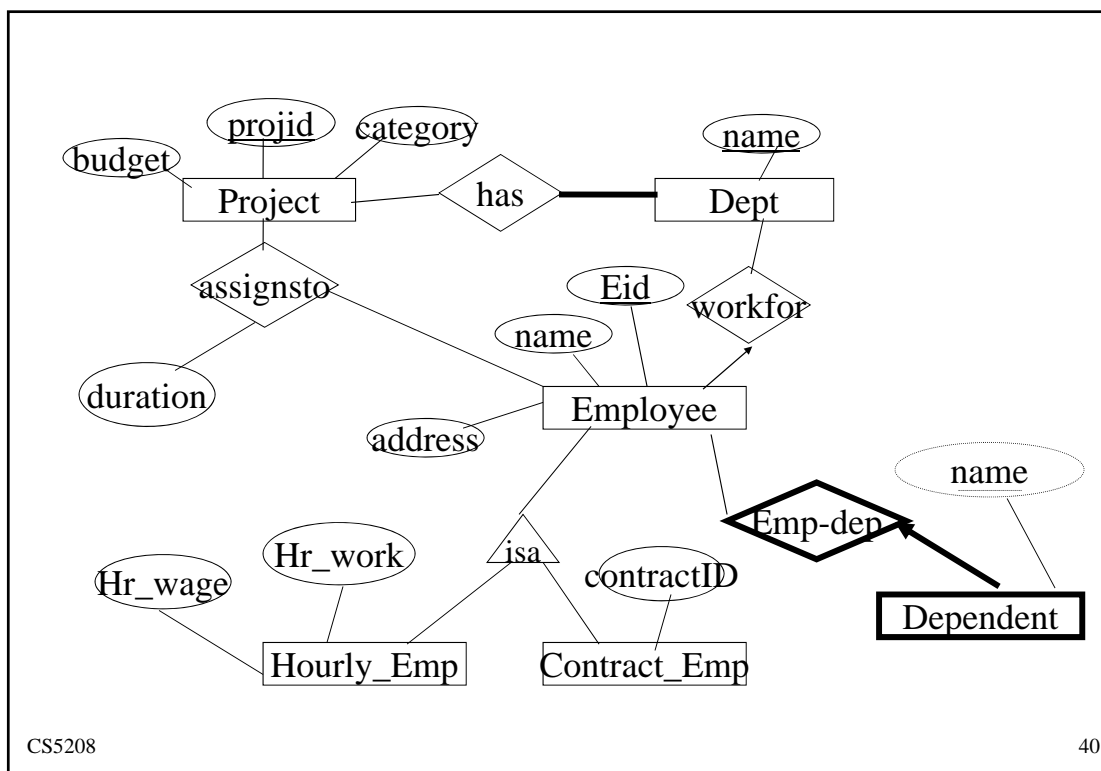
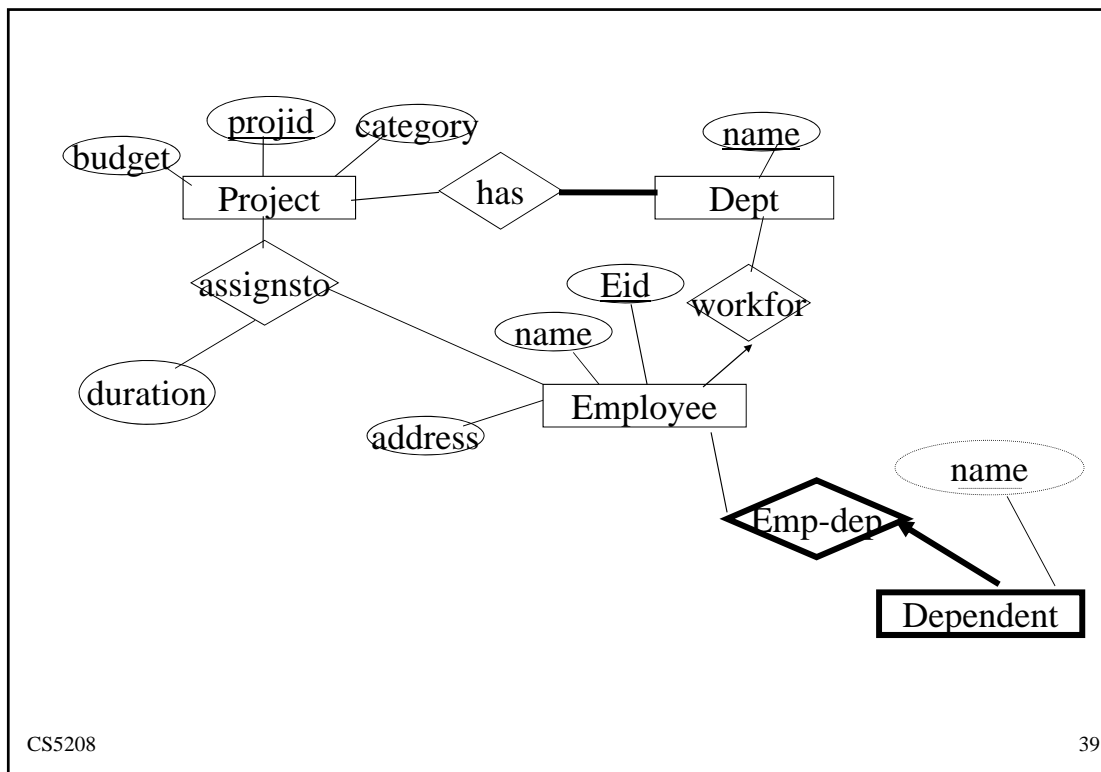
CS5208

37



CS5208

38



Relational Data Model

- a data model in which all data is modelled as relations (tables)
 - a way of looking at data
- a prescription for a way of
 - representing data
 - manipulating data (relational algebra)
 - representing integrity constraints

CS5208

41

A Relation A relation contains a SET of tuples

PARTS(Name: String; Price: Real; Category: String; Manufacturer: String)

Attribute names

Name	Price (\$)	Category	Manufacturer
Gizmo	19.99	gadgets	GizmoWorks
Power gizmo	29.99	gadgets	GizmoWorks
SingleTouch	149.99	photography	Canon
MultiTouch	203.99	household	Hitachi

Tuples (record) Each attribute has an atomic type

CS5208

42

Integrity

- restrictions on data defined by users
 - on individual tables
 - $\text{age} > 18; \text{salary} < 100\text{k}$
 - on more than one table
 - *if* budget $< 10\text{M}$ *then* salary $< 50\text{k}$
- implicit in the data model

CS5208

43

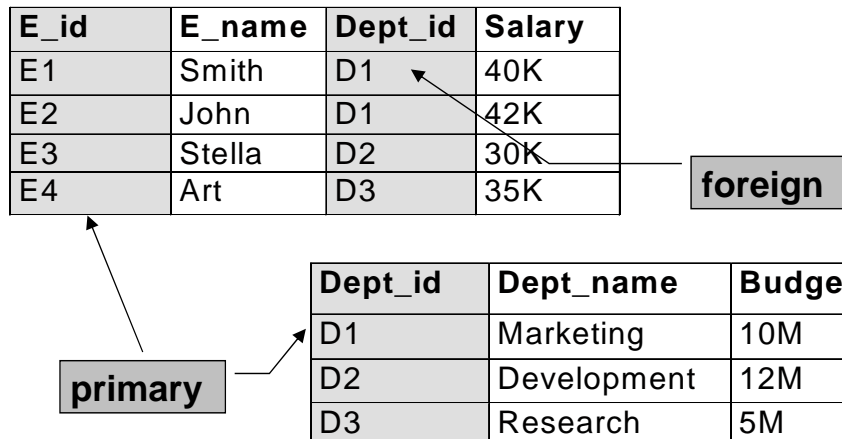
Integrity Constraints (ICs)

- IC: condition that must be true for *any* instance of the database
 - e.g., *domain constraints*
 - Each attribute has values taken from a *domain*. ICs are specified when schema is defined.

CS5208

44

Primary and foreign keys



CS5208

45

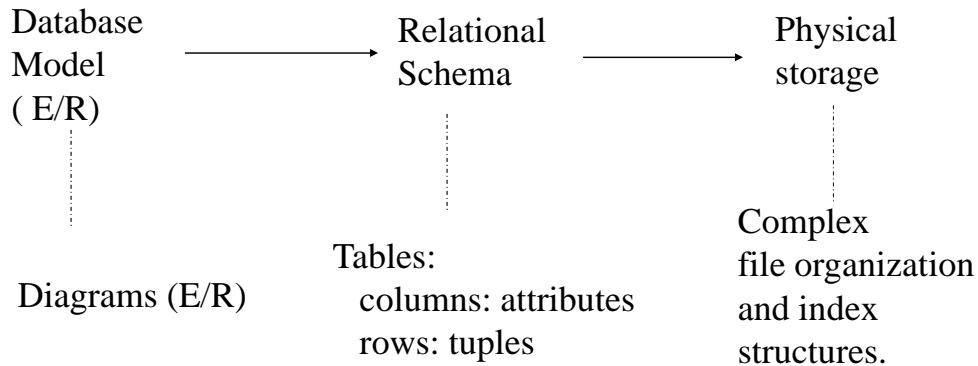
More Integrity Constraints

- *Key constraints*: each tuple must be distinct. A key is a subset of fields that uniquely identifies a tuple (*superkey*), and for which no subset of the key has this property.
- *Referential integrity constraints*: a field in one relation may refer to a tuple in another relation by including its key (*foreign key*). The referenced tuple must exist in the other relation for the database instance to be valid.
- Typically, a relation may have several *candidate* keys one of which is chosen as the *primary* key.

CS5208

46

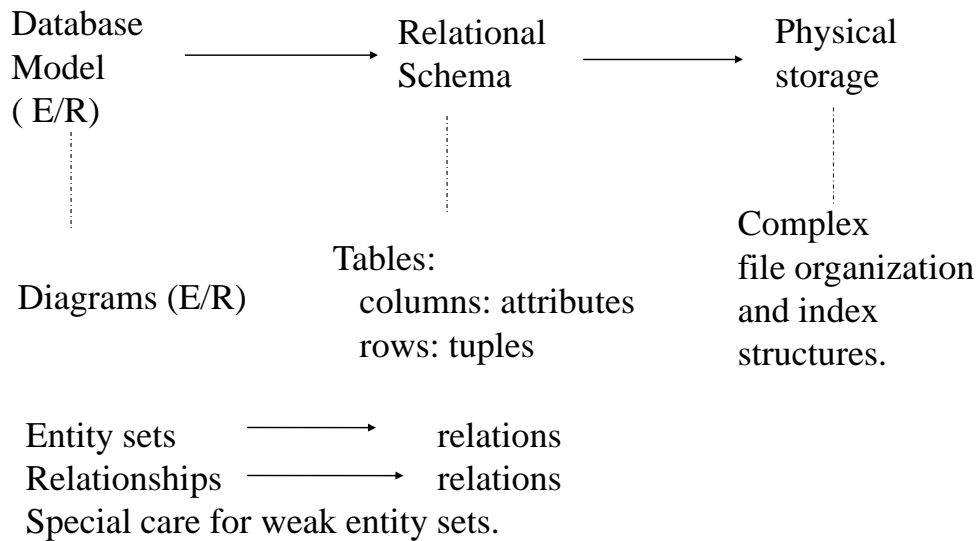
From E/R Diagrams to Relational Schema



CS5208

47

From E/R Diagrams to Relational Schema



CS5208

48

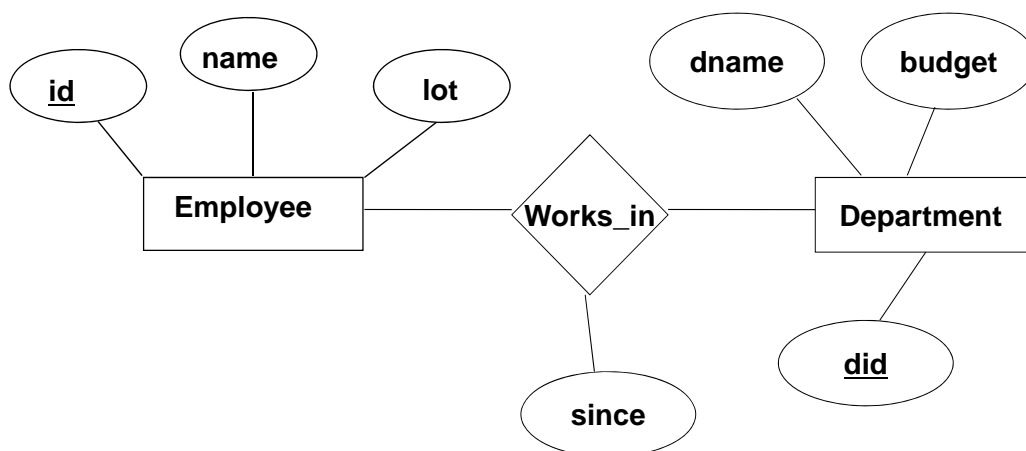
Relationships to Relations

- In translating a relationship set to a relation, attributes of the relation must include:
 - Keys for each participating entity set (as foreign keys).
 - This set of attributes forms a *superkey* for the relation.
 - All descriptive attributes.

CS5208

49

Example



CS5208

50

Example continued

Relation Employee

id : CHAR(9),
name : CHAR(20),
lot : INTEGER
PRIMARY KEY id

CS5208

51

Example continued

Relation Employee

id : CHAR(9),
name : CHAR(20),
lot : INTEGER
PRIMARY KEY id

Relation Department

did : INTEGER,
dname : CHAR(20),
budget : REAL
PRIMARY KEY did

Relation Works_In

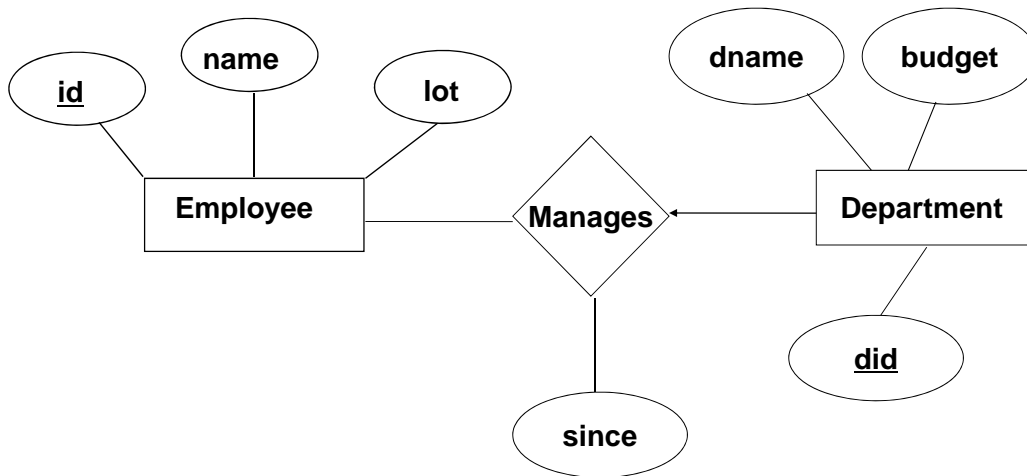
id : CHAR(9),
did : INTEGER,
since : DATE,
PRIMARY KEY (id, did),
FOREIGN KEY (id) REFERENCES Employee,
FOREIGN KEY (did) REFERENCES Department

CS5208

52

Key constraints

Each dept has at most one manager, key constraint on Manages.



CS5208

53

Key Constraints

- Map relationship to a table:
 - Note that ***did*** is the key now!

Relation Manages

id : CHAR(9),

did : INTEGER,

since : DATE,

PRIMARY KEY did,

FOREIGN KEY id REFERENCES Employee,

FOREIGN KEY did REFERENCES Department

CS5208

54

Key Constraints

Since each department has a unique manager, we could instead combine Manages and Departments.

CS5208

55

Key Constraints

Since each department has a unique manager, we could instead combine Manages and Departments.

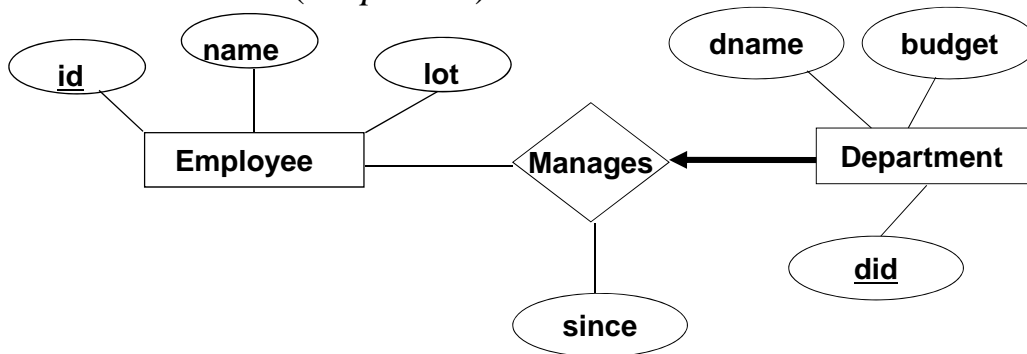
```
Relation Dept_Mgr
  did : INTEGER,
  dname : CHAR(20),
  budget : REAL,
  id : CHAR(11),
  since : DATE,
  PRIMARY KEY did,
  FOREIGN KEY id REFERENCES Employee
```

CS5208

56

Participation Constraints

- Does every department have a manager?
 - If so, this is a *participation constraint*: the participation of Departments in Manages is said to be *total* (vs. *partial*).



CS5208

57

Participation Constraints

- Every *did* value in Department table must appear in a row of the Manages table (with a non-null id value!)

CS5208

58

Participation Constraints

- Every *did* value in Department table must appear in a row of the Manages table (with a non-null id value!)

```
Relation Dept_Mgr
  did : INTEGER,
  dname : CHAR(20),
  budget : REAL,
  id : CHAR(9) NOT NULL,
  since : DATE,
  PRIMARY KEY did,
  FOREIGN KEY (id) REFERENCES Employee,
  ON DELETE NO ACTION
```

CS5208

59

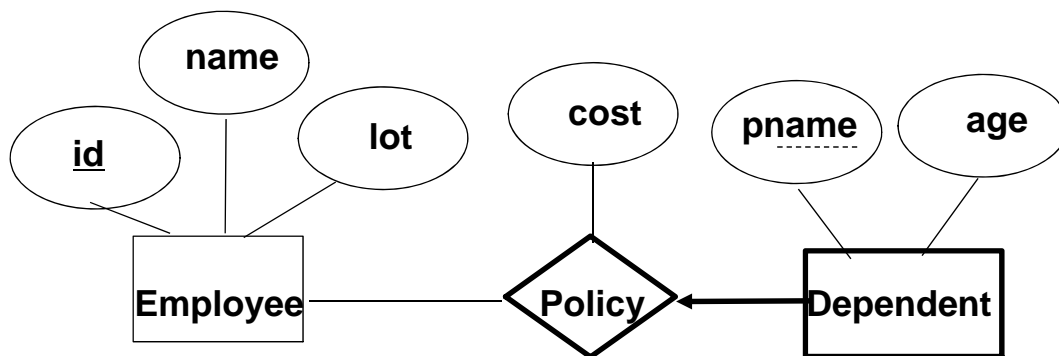
Weak Entities to Relations

- A *weak entity* can be identified uniquely only by considering the primary key of another (*owner*) entity.
 - a one-to-many relationship set (1 owner, many weak entities).
 - Weak entity set must have total participation in this *identifying* relationship set.

CS5208

60

Weak Entities to Relations



CS5208

61

Translating Weak Entity Sets

- Weak entity set and identifying relationship set are translated into a single table.
 - When the owner entity is deleted, all owned weak entities must also be deleted.

```
Relation Dep_Policy
pname CHAR(20),
age INTEGER,
cost REAL,
id CHAR(9) NOT NULL,
PRIMARY KEY (pname, id),
FOREIGN KEY (id) REFERENCES Employee,
ON DELETE CASCADE
```

CS5208

62

Translating ISA Hierarchies

- 3 relations: Employees, Hourly_Emps and Contract_Emps.
- *Hourly_Emps*: Every employee is recorded in Employees.
 - extra info recorded in Hourly_Emps (*hourly_wages*, *hours_worked*, *id*);
 - must delete Hourly_Emps tuple if referenced Employees tuple is deleted.

You are now a *trained* database designer!