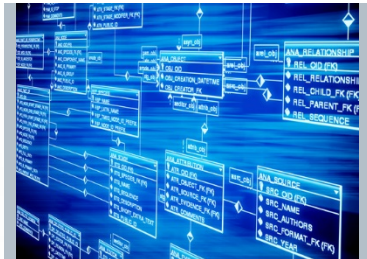
Introduction to Computer Science

William Hsu
Department of Computer Science and Engineering
National Taiwan Ocean University

Chapter 9: Database Systems

You can have data without information, but you cannot have information without data.
Daniel Keys Moran

Data is a precious thing and will last longer than the systems themselves.
Tim Berners-Lee

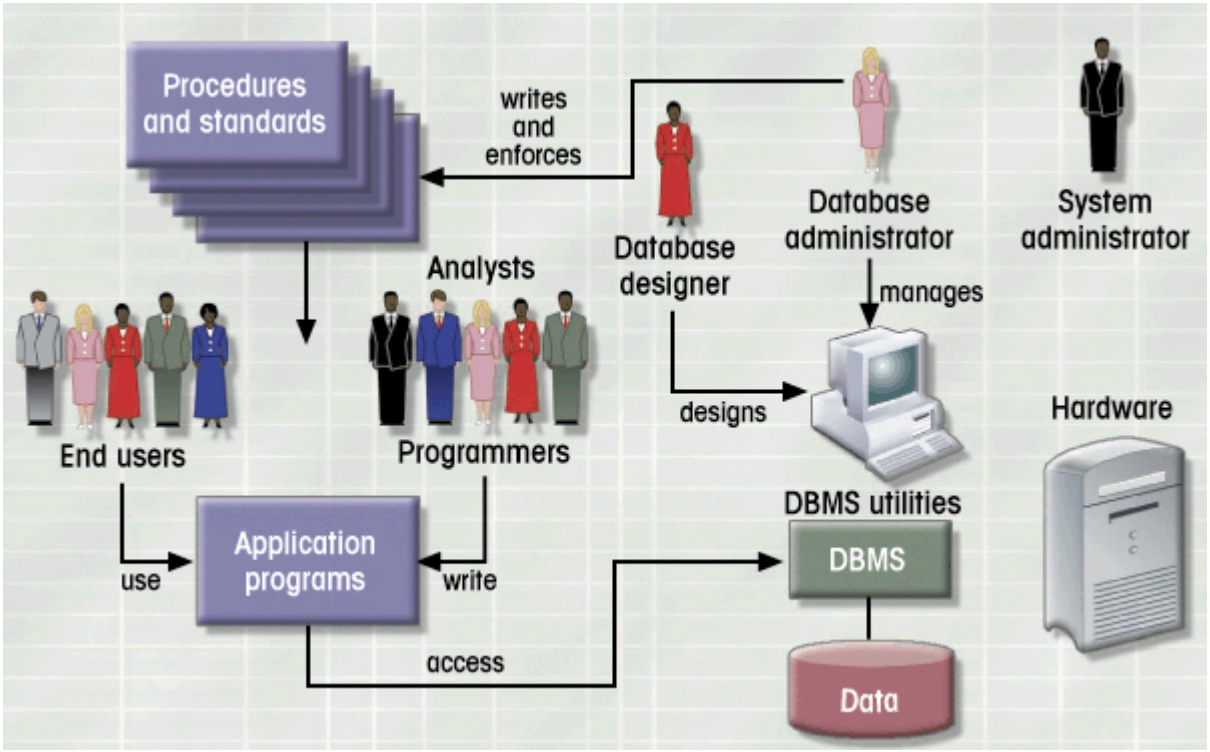



- 9.1 Database Fundamentals
- 9.2 The Relational Model
- 9.3 Object-Oriented Databases
- 9.4 Maintaining Database Integrity
- 9.5 Traditional File Structures
- 9.6 Data Mining
- 9.7 Social Impact of Database Technology

Database

- A collection of data that is multidimensional in the sense that internal links between its entries make the information accessible from a variety of perspectives.

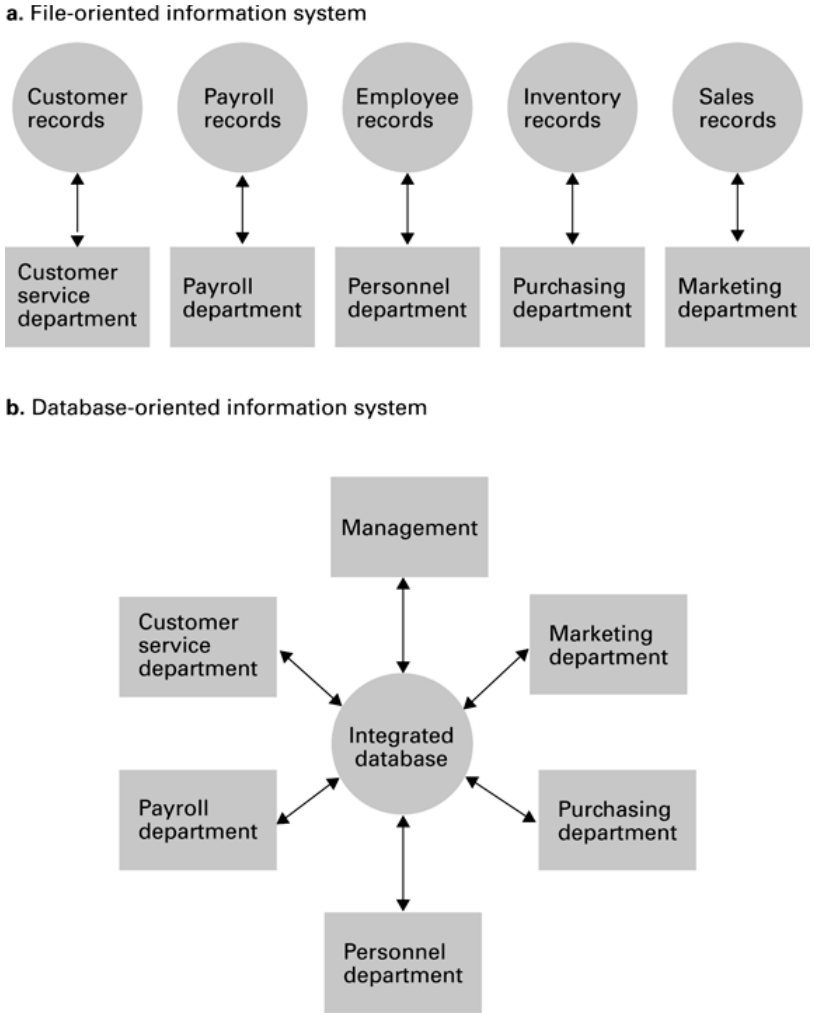
Database system environment



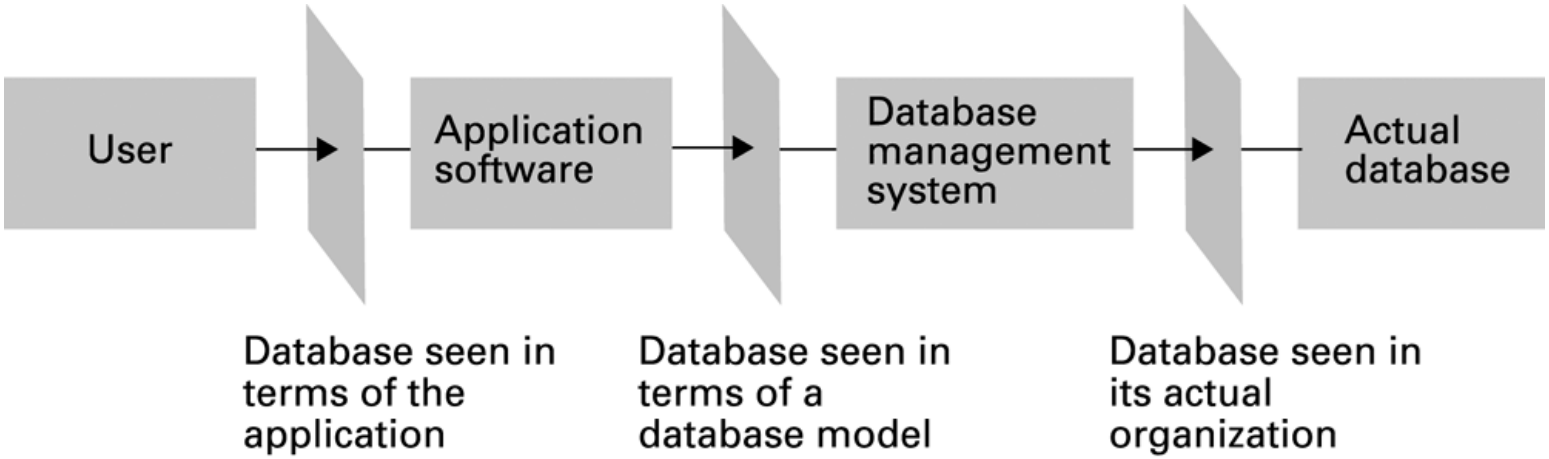
- **Hardware**
- **Software**
  - OS
  - DBMS
  - Applications
- **People**
- **Procedures**
- **Data**

Database Systems: Design, Implementation, & Management: Rob & Coronel

A File versus a Database Organization

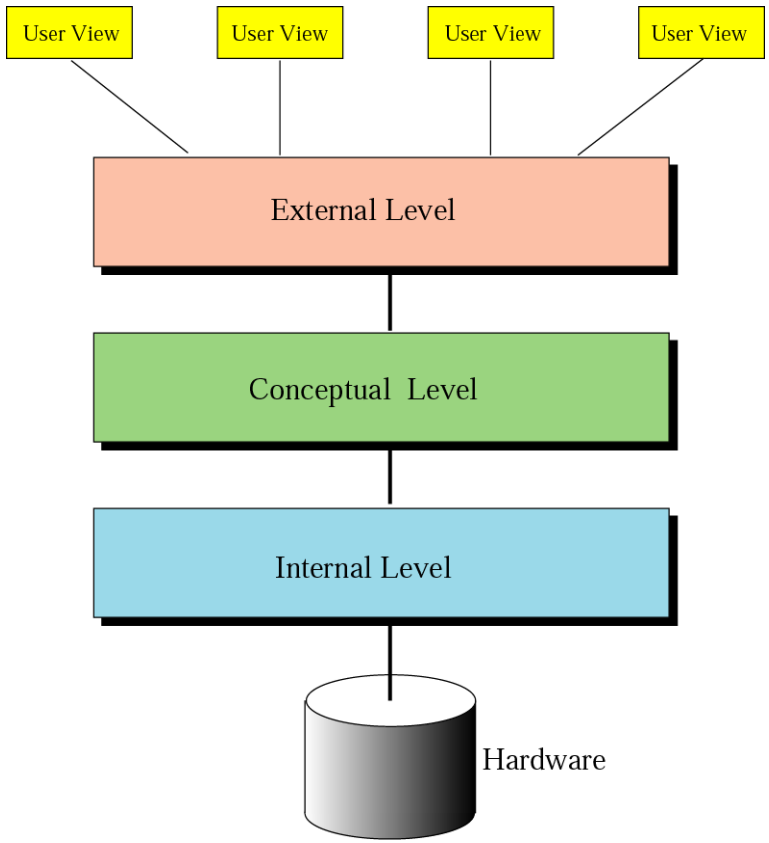


The Conceptual Layers of a Database Implementation



Database architecture

- Internal Level: Interact directly with the hardware
- Conceptual Level: (1) Define the logical view of the data. (2) Define the data model. (3) Contain the main functions of the DBMS (4) Intermediary level that free users from dealing with internal level
- External Level: (1) Interact directly with users (2) Display data in familiar format



Schemas

- **Schema:** A description of the structure of an entire database, used by database software to maintain the database.
- **Subschema:** A description of only that portion of the database pertinent to a particular user's needs, used to prevent sensitive data from being accessed by unauthorized personnel.

Database Management Systems

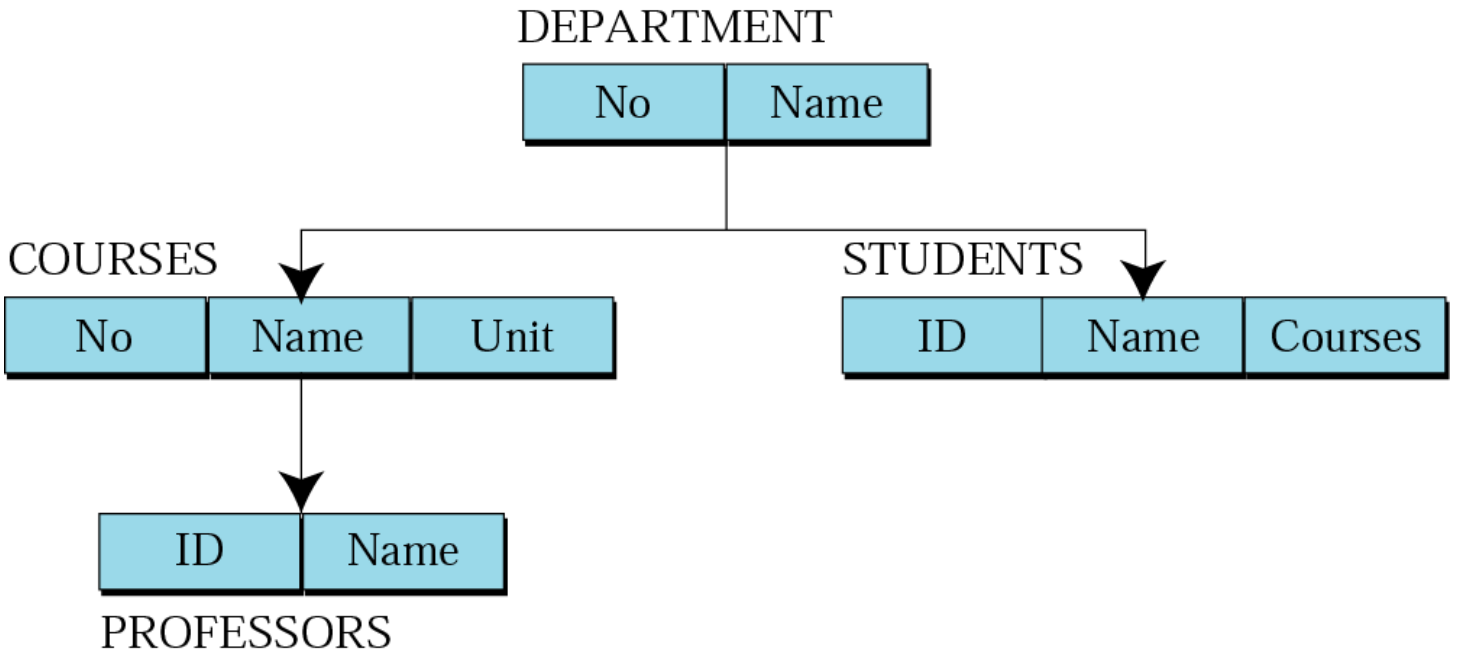
- **Database Management System** (DBMS): A software layer that manipulates a database in response to requests from applications.
- **Distributed Database:** A database stored on multiple machines.
  - DBMS will mask this organizational detail from its users.
- **Data independence:** The ability to change the organization of a database without changing the application software that uses it.

Database Models

- **Database model:** A conceptual view of a database.
  - Hierarchical Model
  - Network Model
  - Relational database model.
  - Object-oriented database model.
  - Column database model.

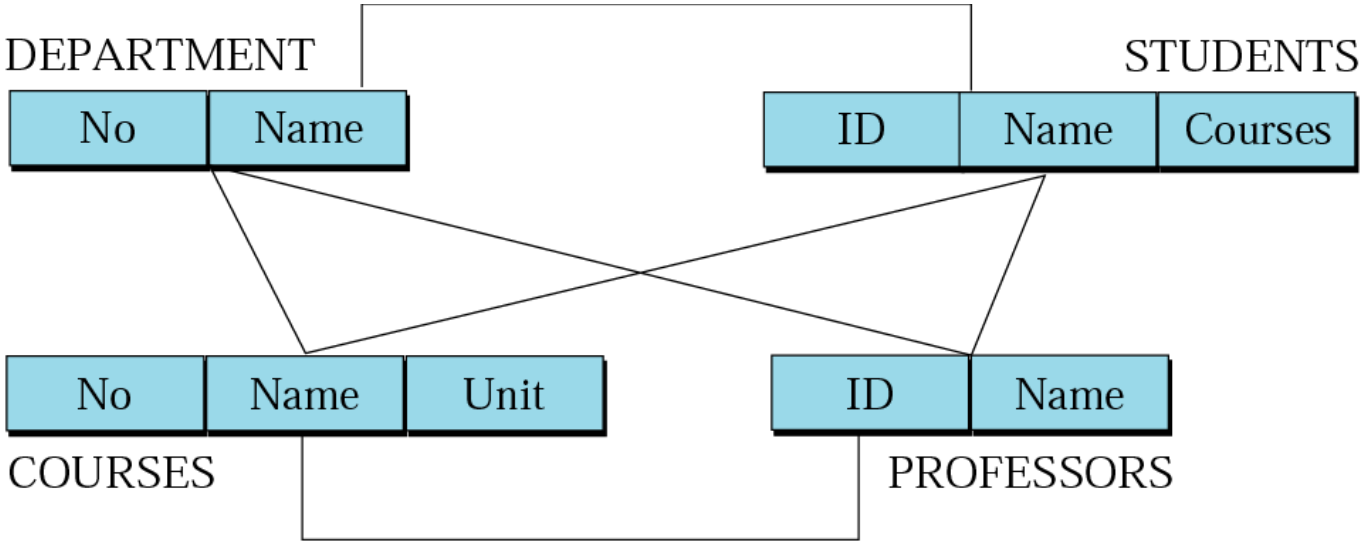
Hierarchical model

- Data are organized in an upside down tree
- Each entity has one parent and many children
- Old and not used now



Network model

- Entities are organized in a graph
- Entities can be accessed through several paths
- Old and not used

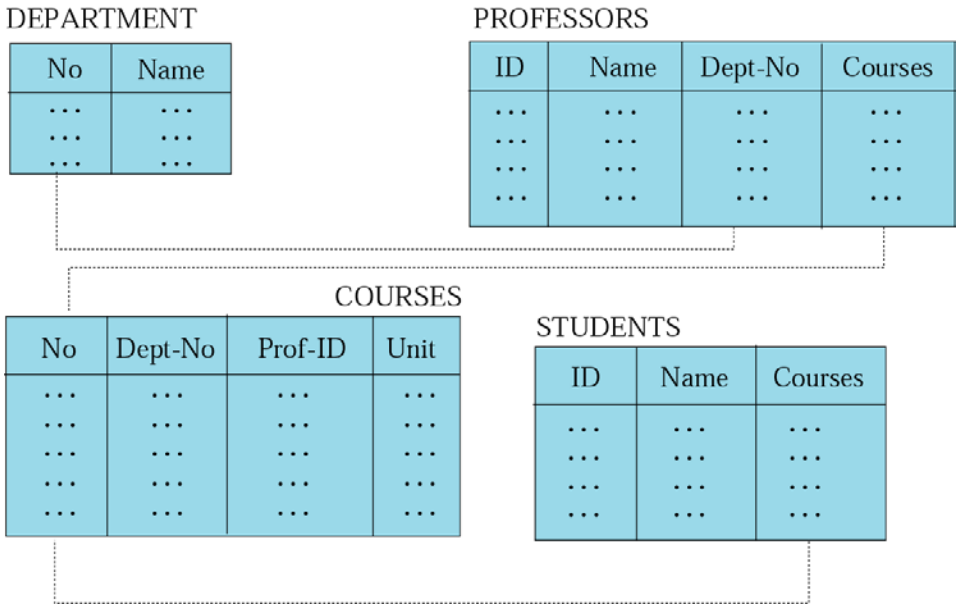


Relational Database Model

- **Relation:** A rectangular table.
  - **Attribute:** A column in the table.
  - **Tuple:** A row in the table.

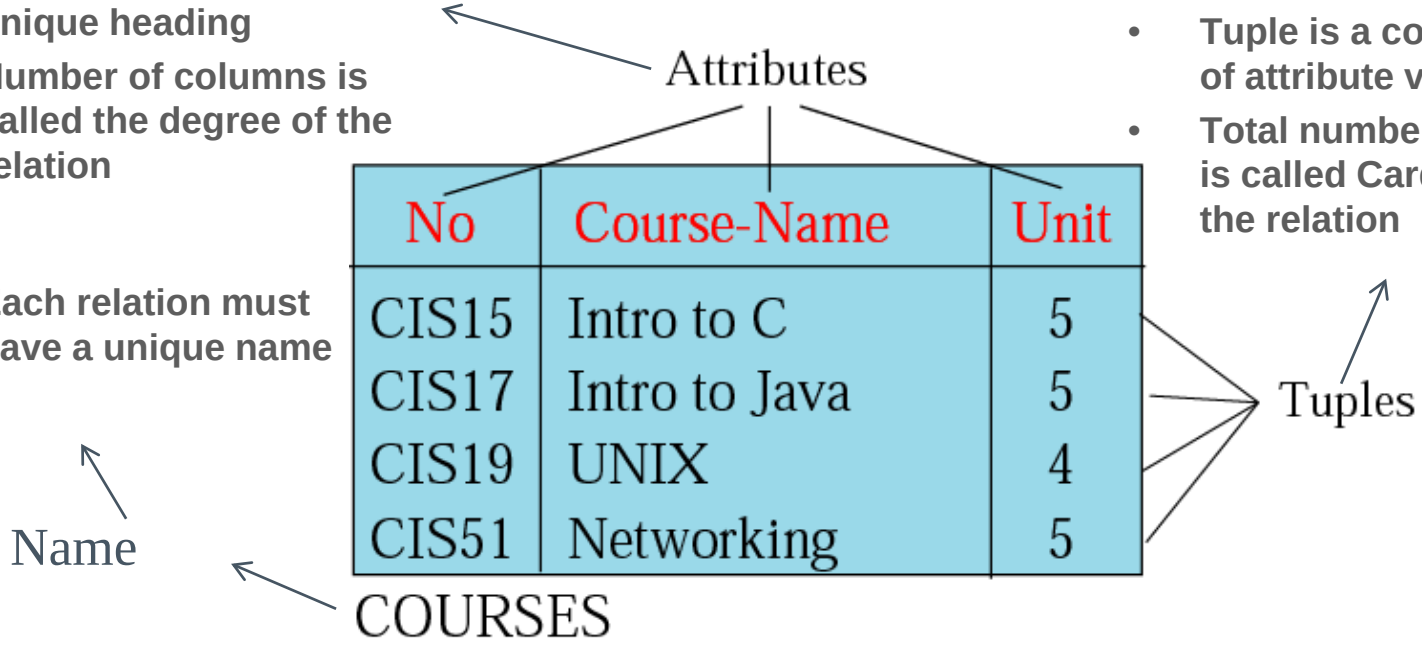
Relational Database Model

- Data are organized in two dimensional tables (relations)
- Tables re related to each other
- Relational Database Management System (RDBMS) are more common model used today



Relation (Name, Attributes, Tuples)

- Attributes are the column heading
- Each column must have a unique heading
- Number of columns is called the degree of the relation
- Each relation must have a unique name

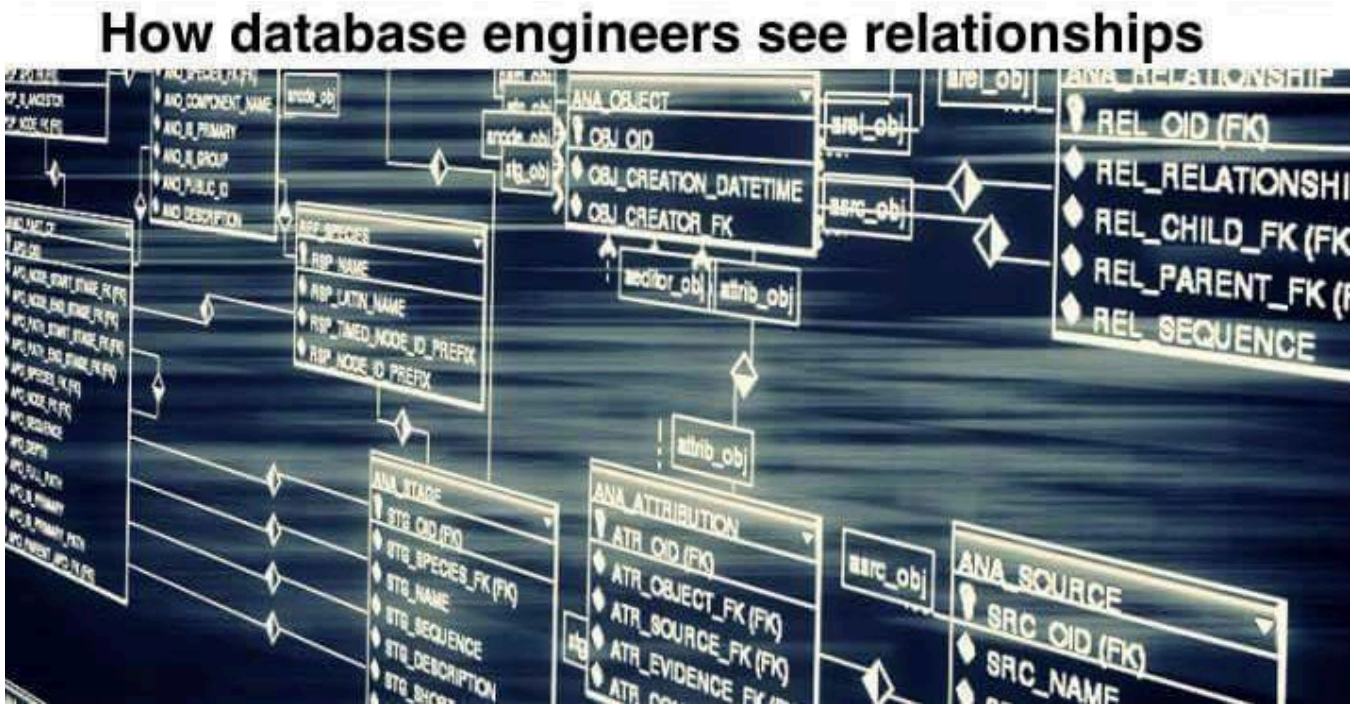


- Tuple is a collection of attribute value
- Total number of rows is called Cardinality of the relation

› Relation appears in 2 dimensional table

› That doesn't mean data stored as table; the physical storage of data is independent of the logical organization of data

Relationships



Relationships

A Relation Containing Employee Information

Empl Id	Name	Address	SSN
25X15	Joe E. Baker	33 Nowhere St.	111223333
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555
•	•	•	•
•	•	•	•
•	•	•	•

Relational Design

- Avoid multiple concepts within one relation.
  - Can lead to redundant data.
  - Deleting a tuple could also delete necessary but unrelated information.

Improving a Relational Design

- **Decomposition:** Dividing the columns of a relation into two or more relations, duplicating those columns necessary to maintain relationships.
  - **Lossless** or **nonloss** decomposition: A “correct” decomposition that does not lose any information.

A Relation Containing Redundancy

Empl Id	Name	Address	SSN	Job Id	Job Title	Skill Code	Dept	Start Date	Term Date
25X15	Joe E. Baker	33 Nowhere St.	111223333	F5	Floor manager	FM3	Sales	9-1-2007	9-30-2008
25X15	Joe E. Baker	33 Nowhere St.	111223333	D7	Dept. head	K2	Sales	10-1-2008	*
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999	F5	Floor manager	FM3	Sales	10-1-2007	*
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555	S25X	Secretary	T5	Personnel	3-1-1999	4-30-2006
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555	S26Z	Secretary	T6	Accounting	5-1-2006	*
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

Data fields

An Employee Database Consisting of Three Relations

EMPLOYEE relation

Empl Id	Name	Address	SSN
25X15	Joe E. Baker	33 Nowhere St.	111223333
34Y70	Cheryl H. Clark	563 Downtown Ave.	999009999
23Y34	G. Jerry Smith	1555 Circle Dr.	111005555

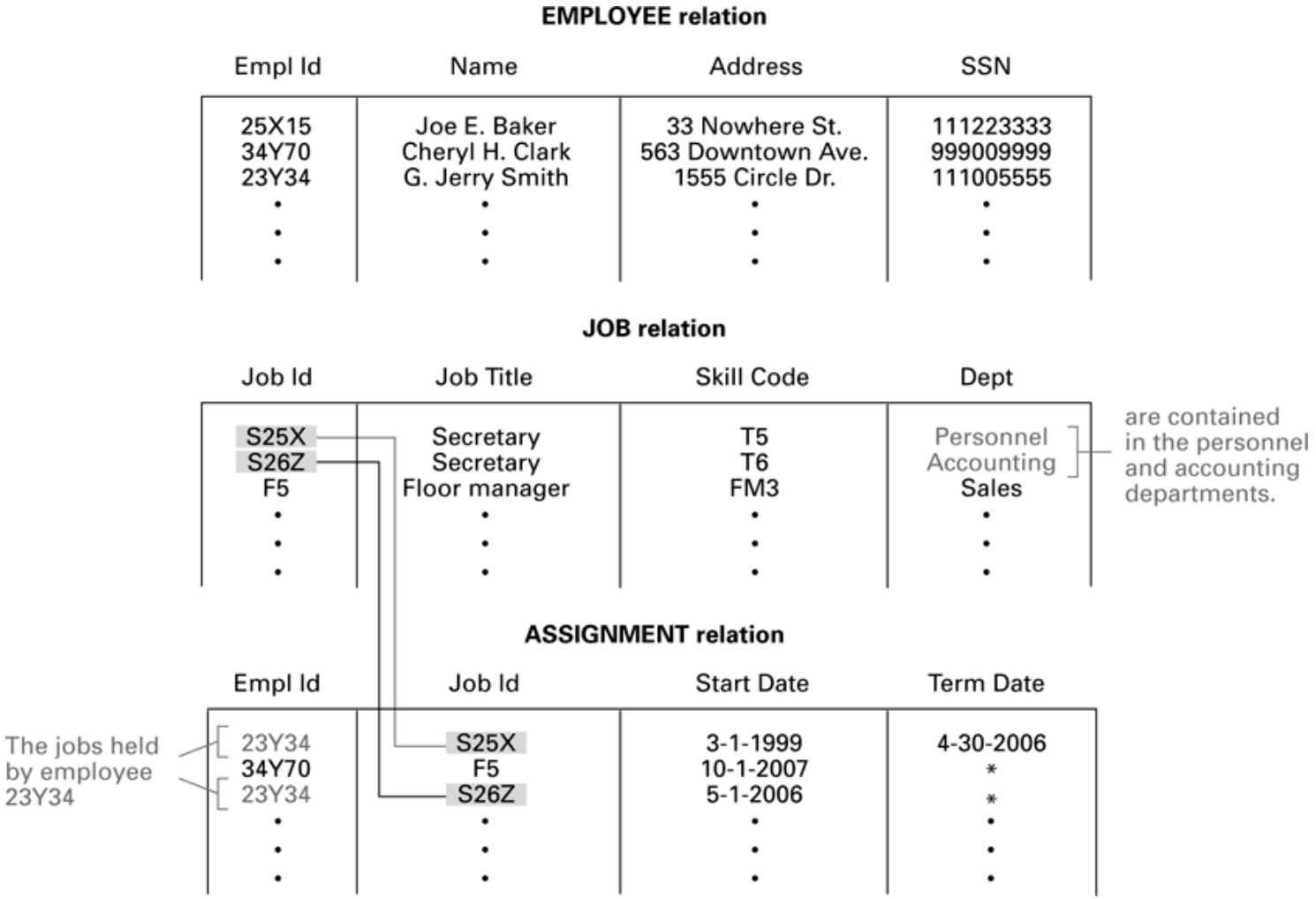
JOB relation

Job Id	JobTitle	Skill Code	Dept
S25X	Secretary	T5	Personnel
S26Z	Secretary	T6	Accounting
F5	Floor manager	FM3	Sales
•	•	•	•
•	•	•	•
•	•	•	•

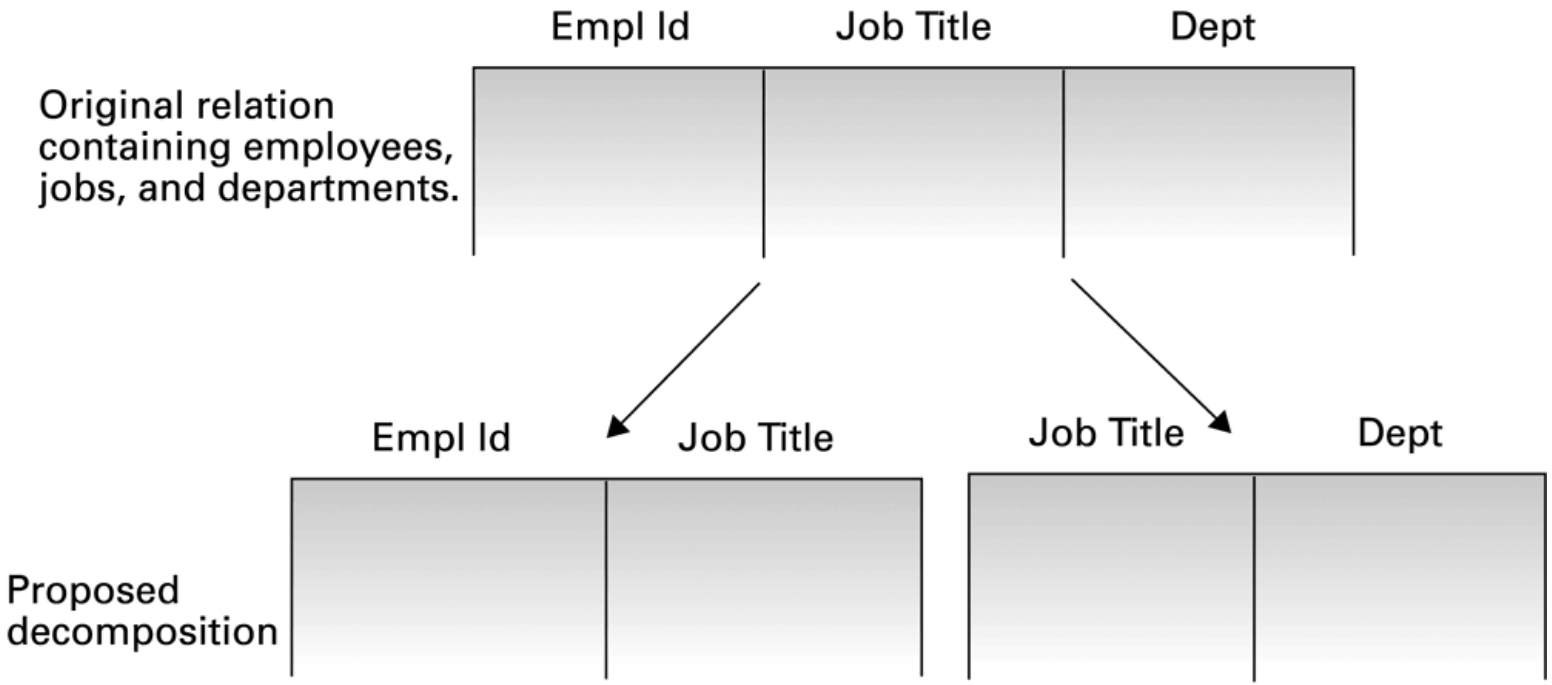
ASSIGNMENT relation

Empl Id	Job Id	Start Date	Term Date
23Y34	S25X	3-1-1999	4-30-2006
34Y70	F5	10-1-2007	*
23Y34	S26Z	5-1-2006	*
•	•	•	•
•	•	•	•
•	•	•	•

Finding the Departments in which Employee 23Y34 has Worked



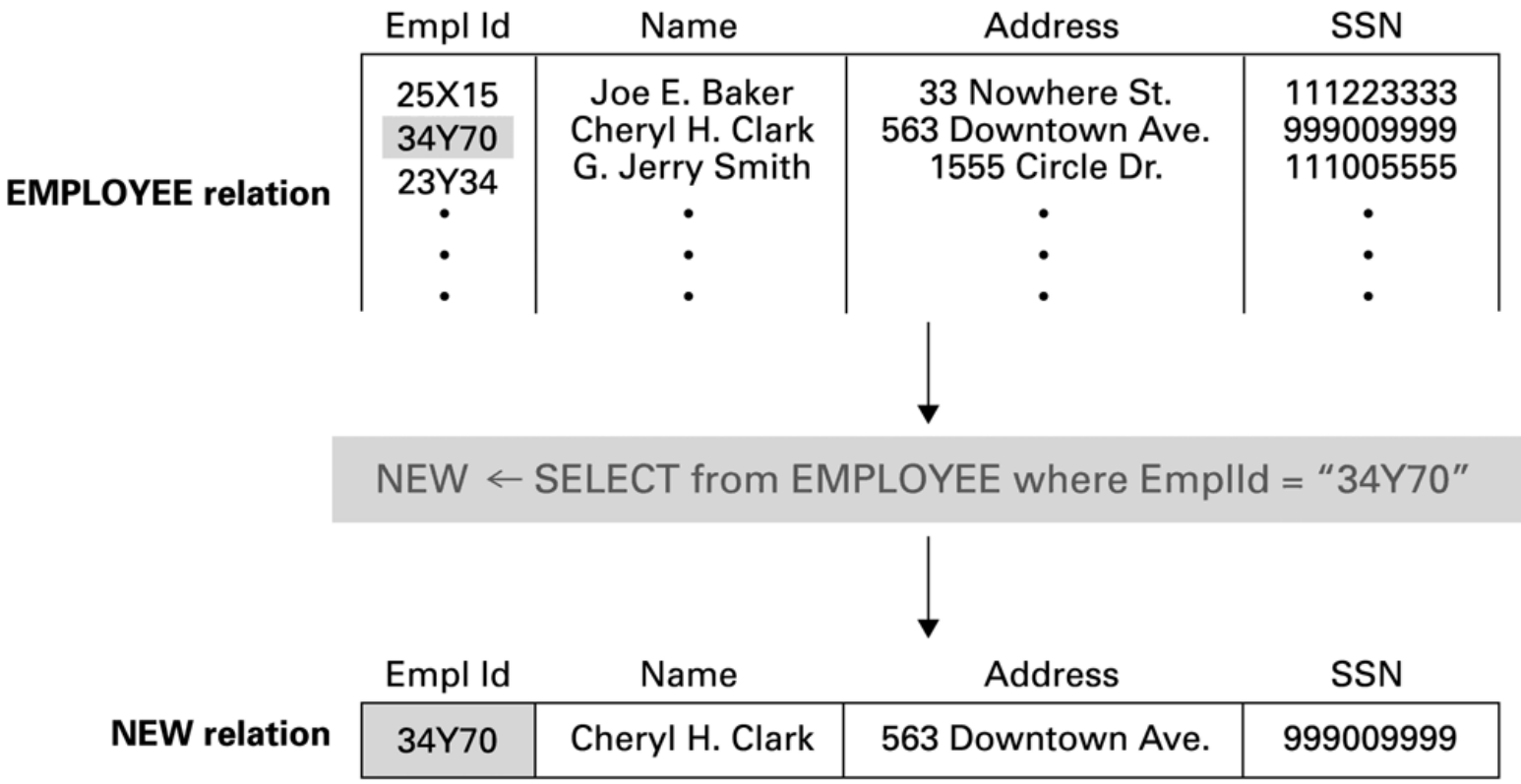
A Relation and a Proposed Decomposition



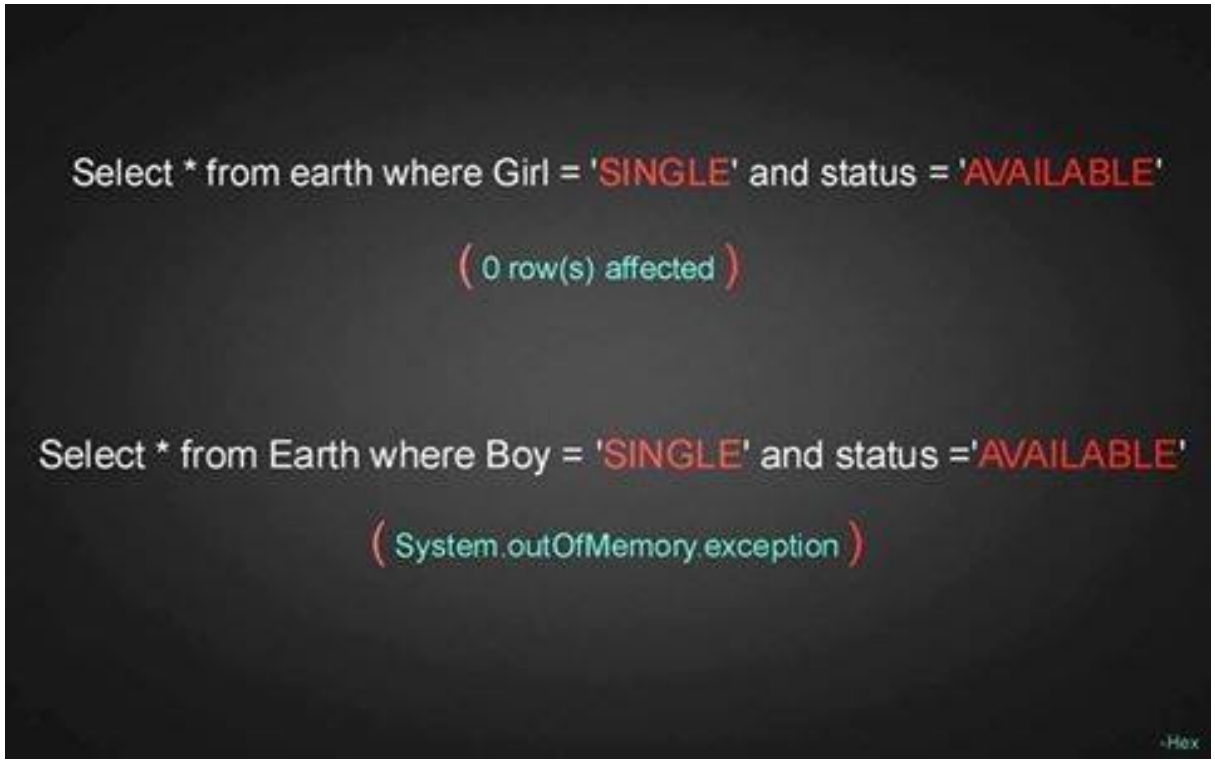
Relational Operations

- **Select:** Choose rows.
- **Project:** Choose columns.
- **Join:** Assemble information from two or more relations.

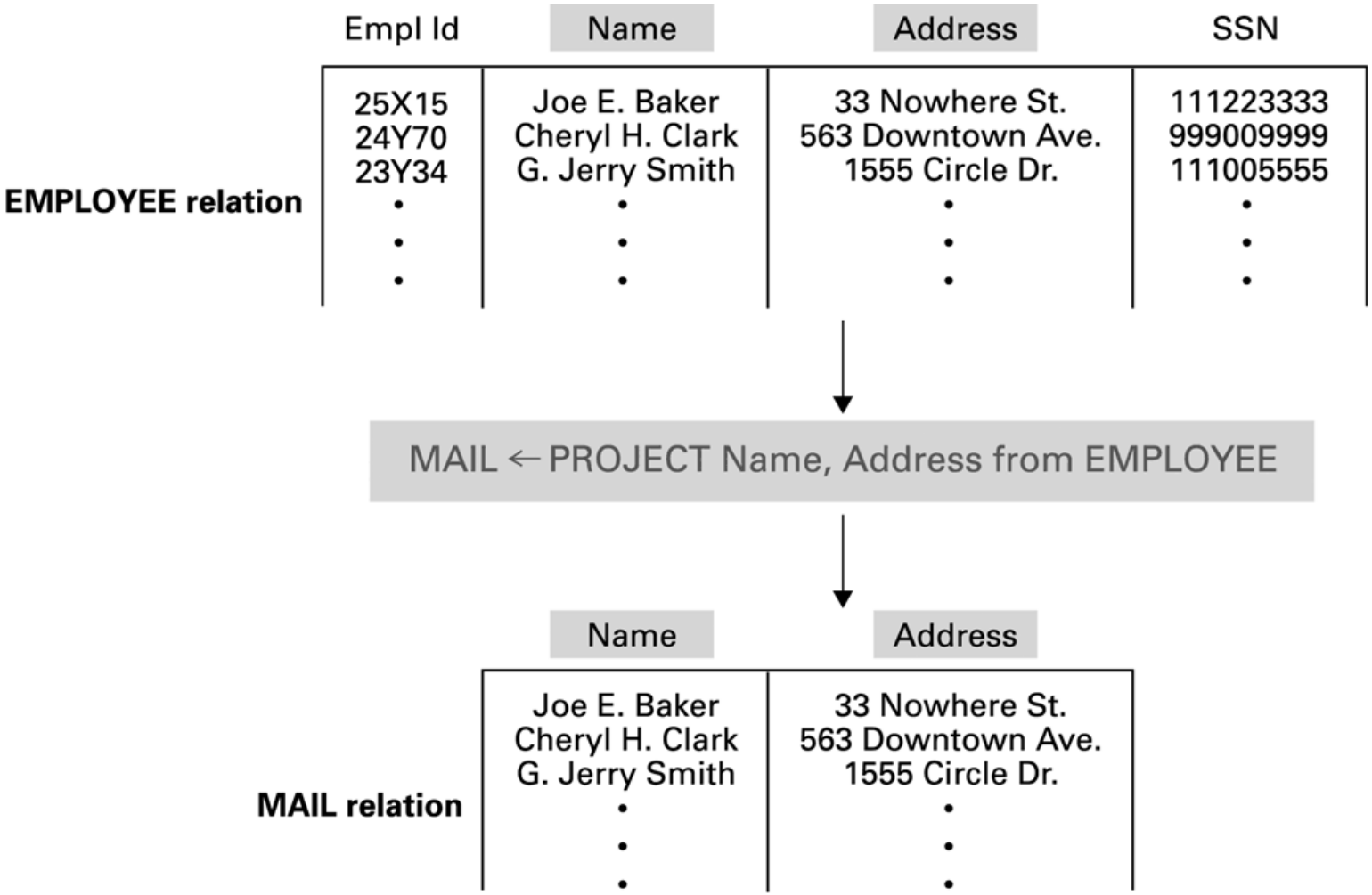
The SELECT Operation



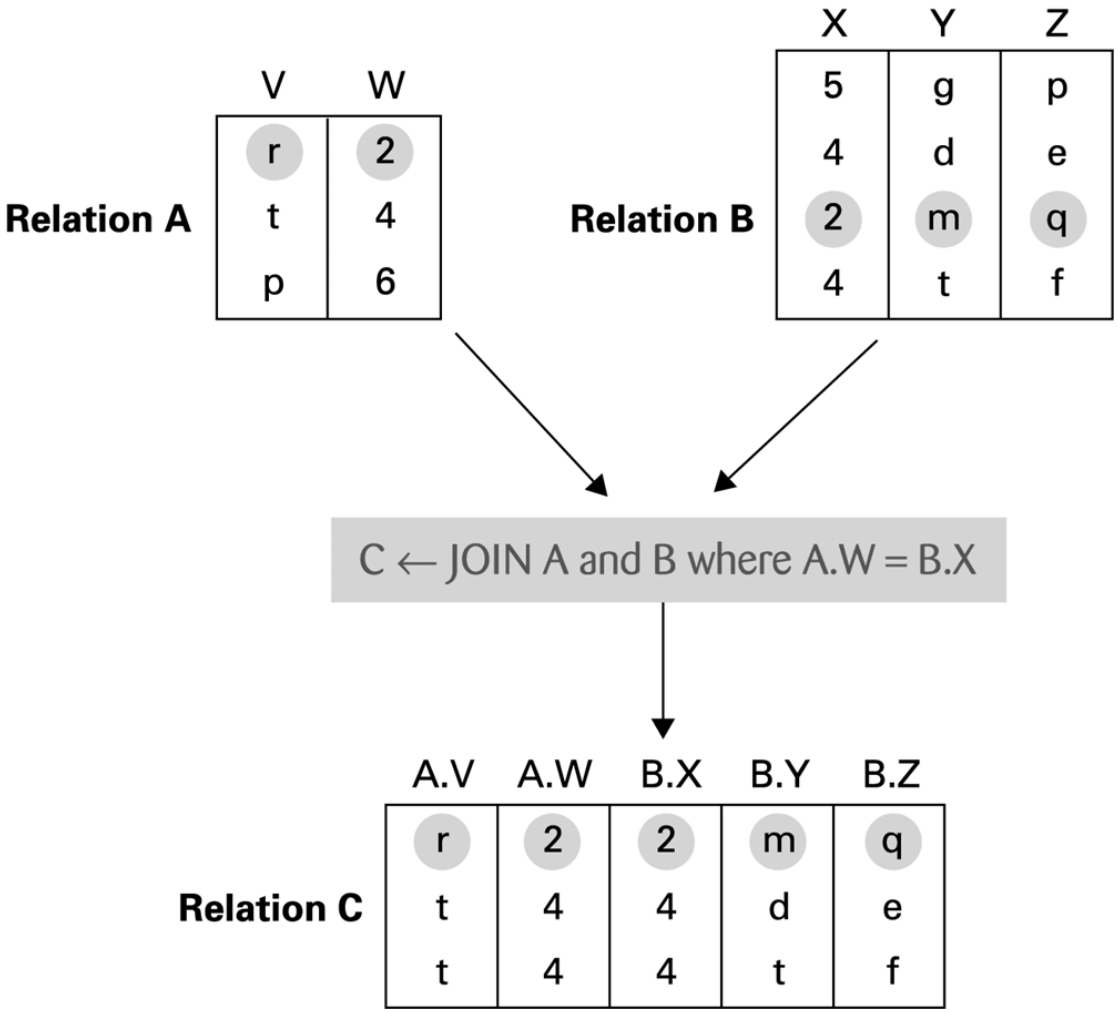
SELECT operation



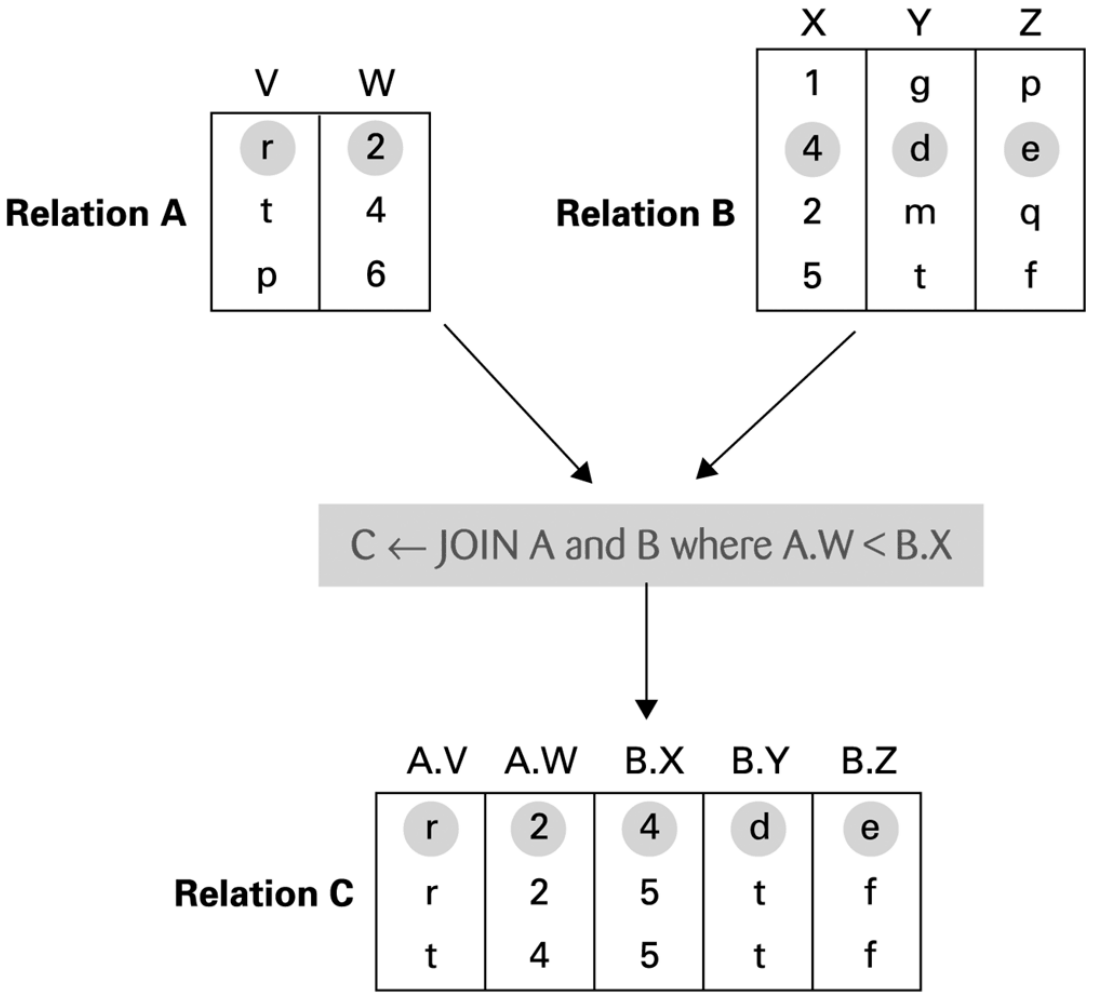
The PROJECT Operation



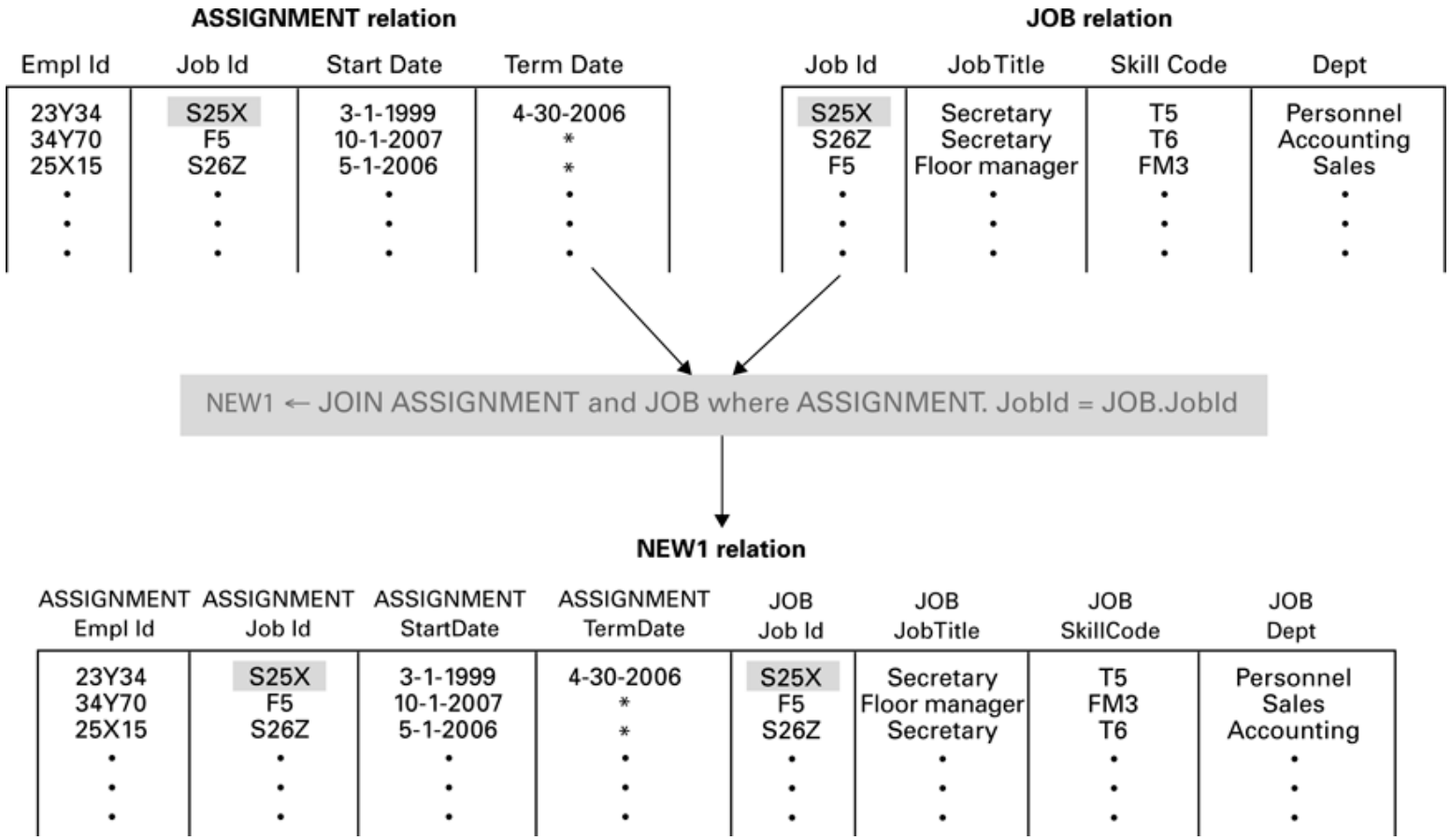
The `JOIN` Operation



Another Example of the `JOIN` Operation



An Application of the `JOIN` Operation



Structured Query Language (SQL)

- Operations to manipulate tuples.
  - `insert`
  - `update`
  - `delete`
- Operations to manipulate tables.
  - `truncate`
  - `drop`

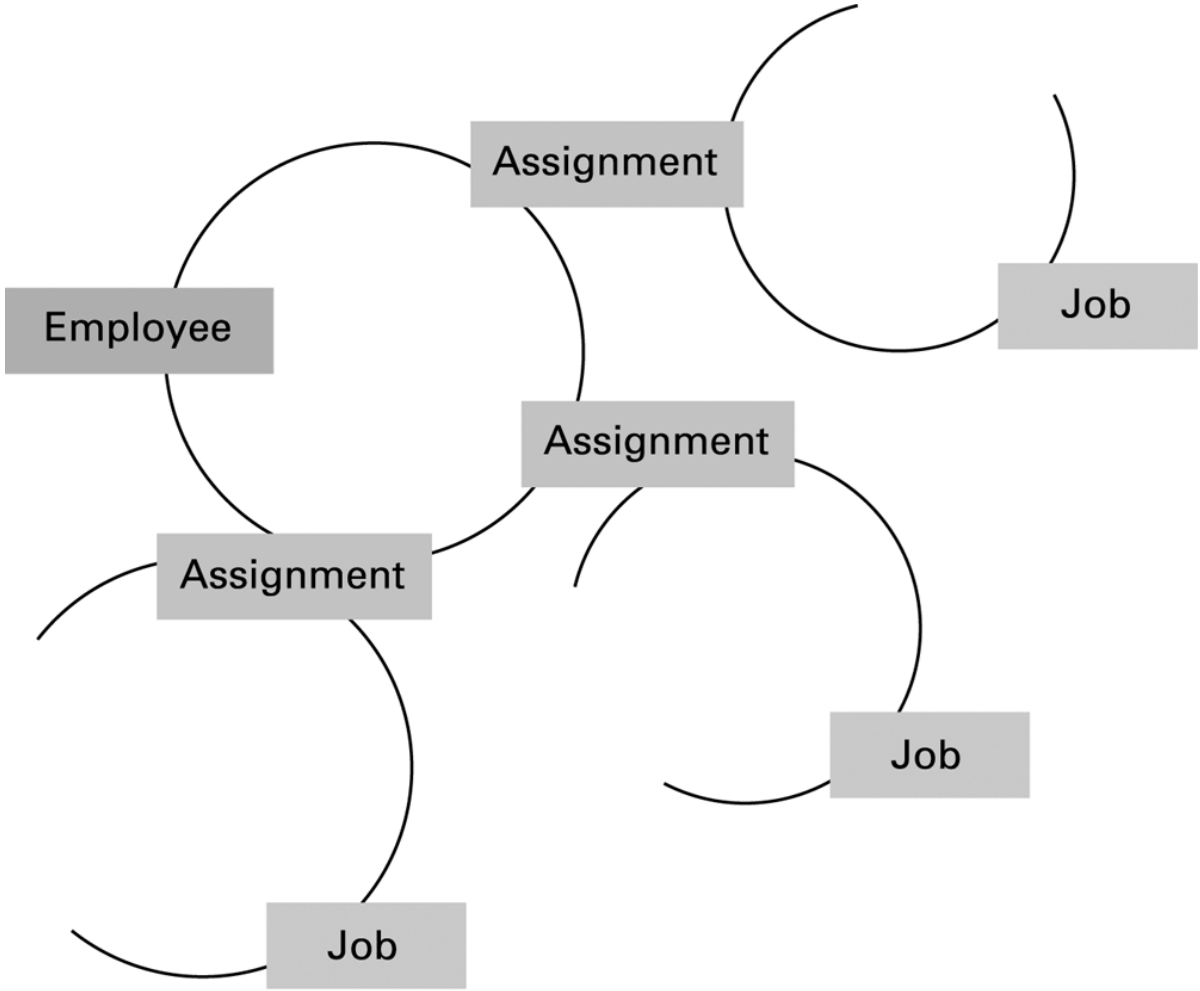
SQL Examples

```
SELECT EmplId, Dept
FROM Assignment, Job
WHERE Assignment.JobId = Job.JobId
      AND Assignment.TermData = '*';


INSERT INTO Employee
VALUES ('43212', 'Sue A. Burt',

        '33 Fair St.', '444661111');

DELETE FROM Employee
       WHERE Name = 'G. Jerry Smith';


UPDATE Employee
       SET Address = '1812 Napoleon Ave.'
       WHERE Name = 'Joe E. Baker';
```

Object-oriented Databases

- **Object-oriented Database:** A database constructed by applying the object-oriented paradigm.
  - Each entity stored as a persistent object.
  - Relationships indicated by links between objects.
  - DBMS maintains inter-object links.

The Associations Between Objects in an Object-oriented Database



Advantages of Object-oriented Databases

- Matches design paradigm of object-oriented applications.
- Intelligence can be built into attribute handlers.
- Can handle exotic data types:
  - Example: multimedia

Maintaining Database Integrity

- **Transaction:** A sequence of operations that must all happen together.
  - Example: transferring money between bank accounts.
- **Transaction log:** A non-volatile record of each transaction's activities, built before the transaction is allowed to execute.
  - **Commit point:** The point at which a transaction has been recorded in the log.
  - **Roll-back:** The process of undoing a transaction.

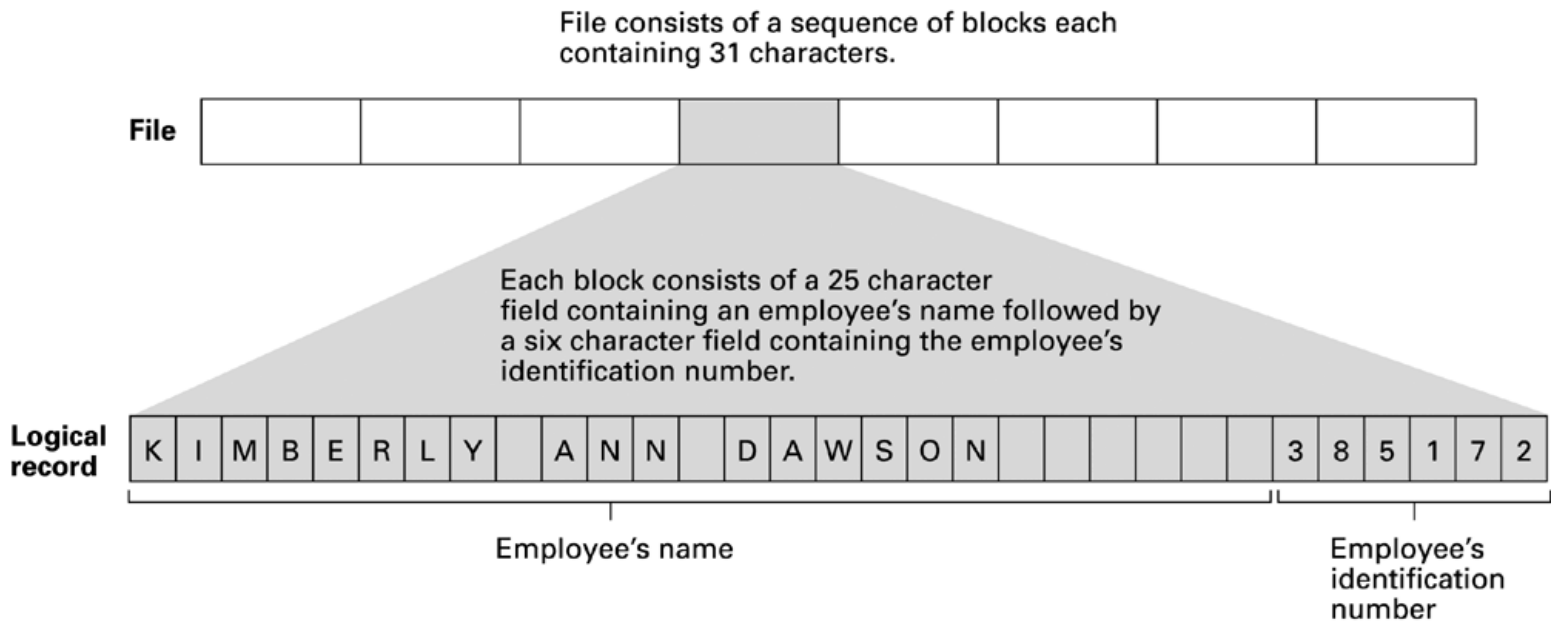
Maintaining Database Integrity (continued)

- Simultaneous access problems.
  - Incorrect summary problem.
  - Lost update problem.
- **Locking** = preventing others from accessing data being used by a transaction.
  - **Shared** lock: used when reading data.
  - **Exclusive** lock: used when altering data.

Sequential Files

- **Sequential file:** A file whose contents can only be read in order.
  - Reader must be able to detect end-of-file (EOF).
  - Data can be stored in logical records, sorted by a key field.
    - Greatly increases the speed of batch updates.

The Structure of a Simple Employee File Implemented as a Text File



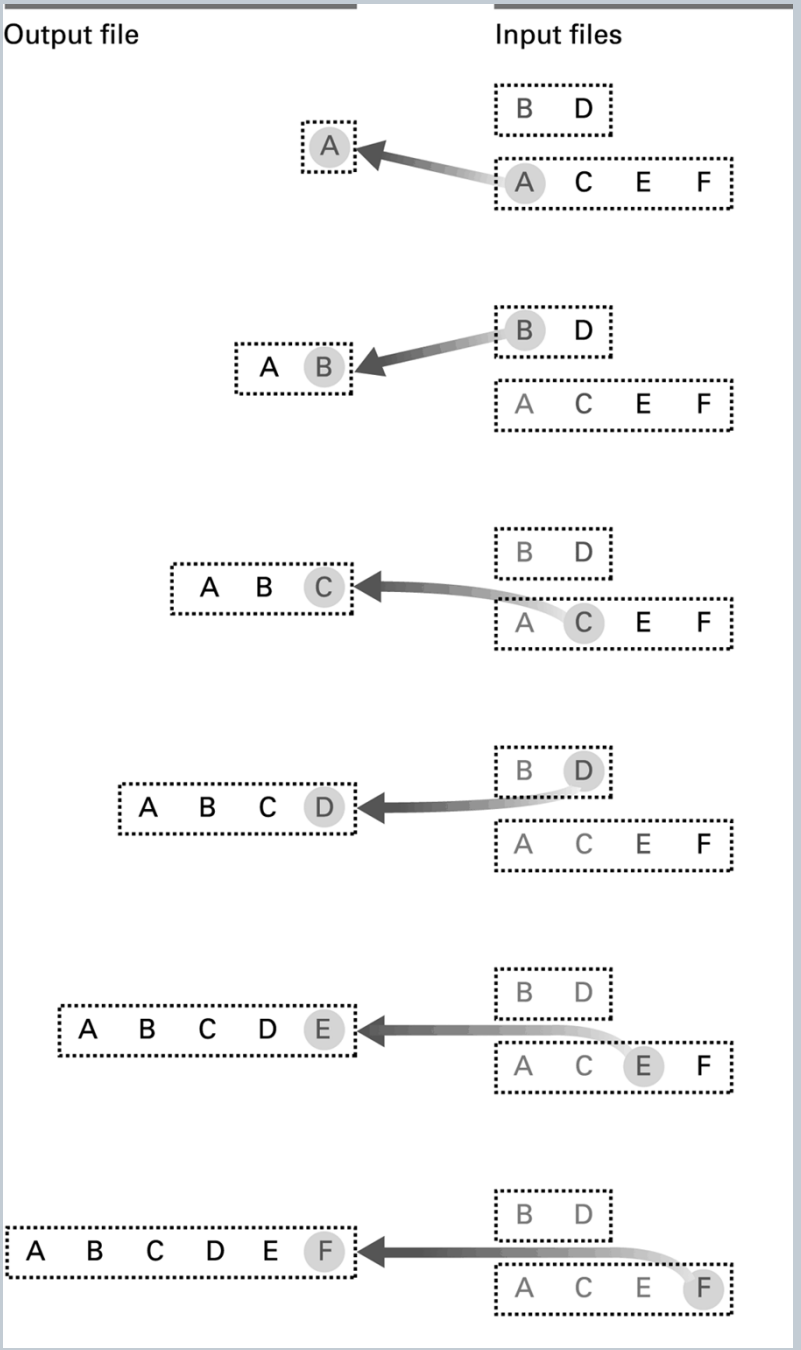
A Function for Merging Two Sequential Files

```
def MergeFiles (InputFileA, InputFileB, OutputFile):

    if (both input files at EOF):

        Stop, with OutputFile empty
    if (InputFileA not at EOF):

        Declare its first record to be its current record
    if (InputFileB not at EOF):

        Declare its first record to be its current record
    while (neither input file at EOF):

        Put the current record with the “smaller” key field value in OutputFile

        if (that current record is the last record in its corresponding input file):

            Declare that input file to be at EOF

        else:

            Declare the next record in that input file to be the file’s current record
            Starting with the current record in the input file that is not at EOF,
            copy the remaining records to OutputFile
```

Applying the Merge Algorithm

Letters are used to represent entire records.

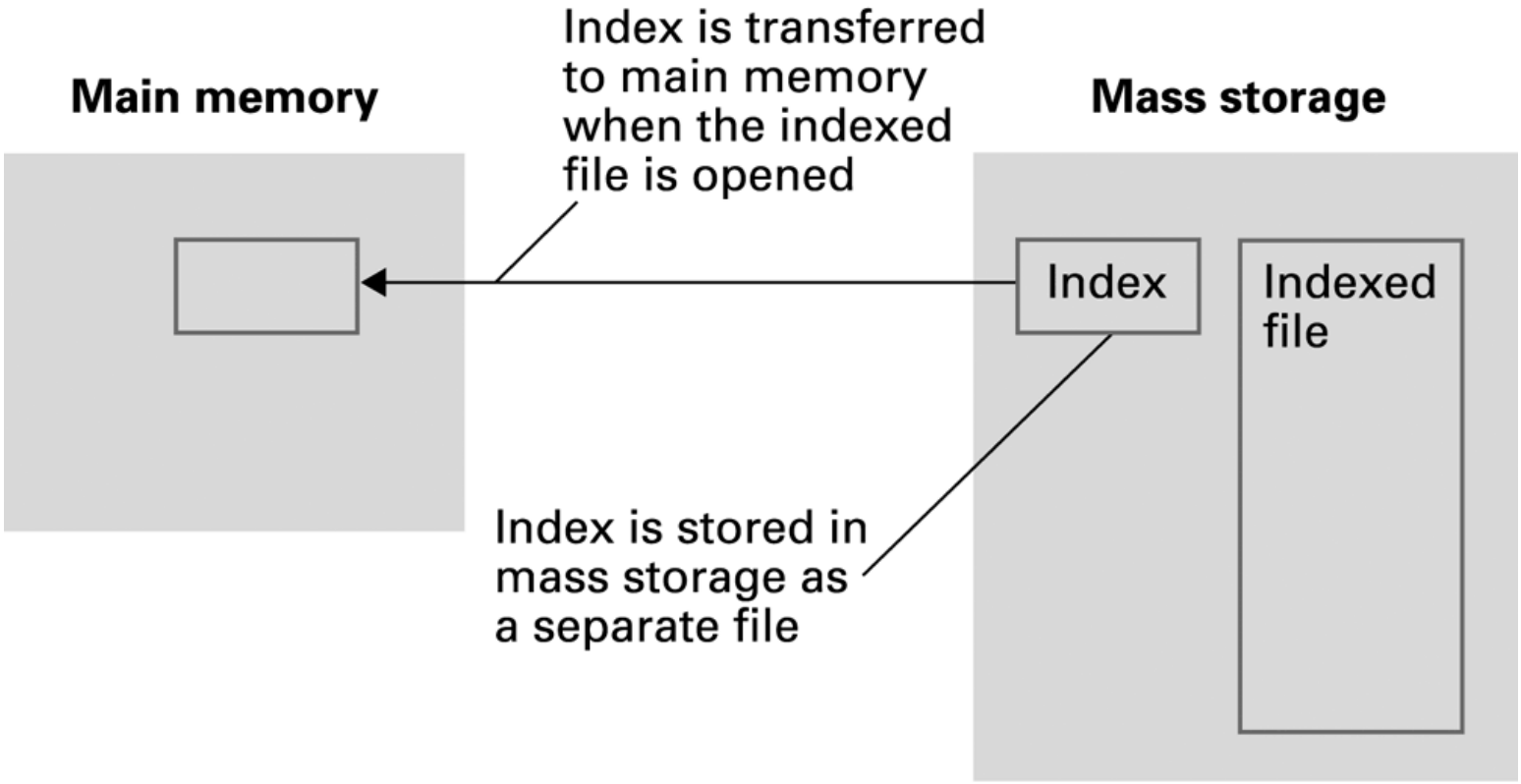
The particular letter indicates the value of the record's key field.



Indexed Files

- **Index:** A list of key values and the location of their associated records.

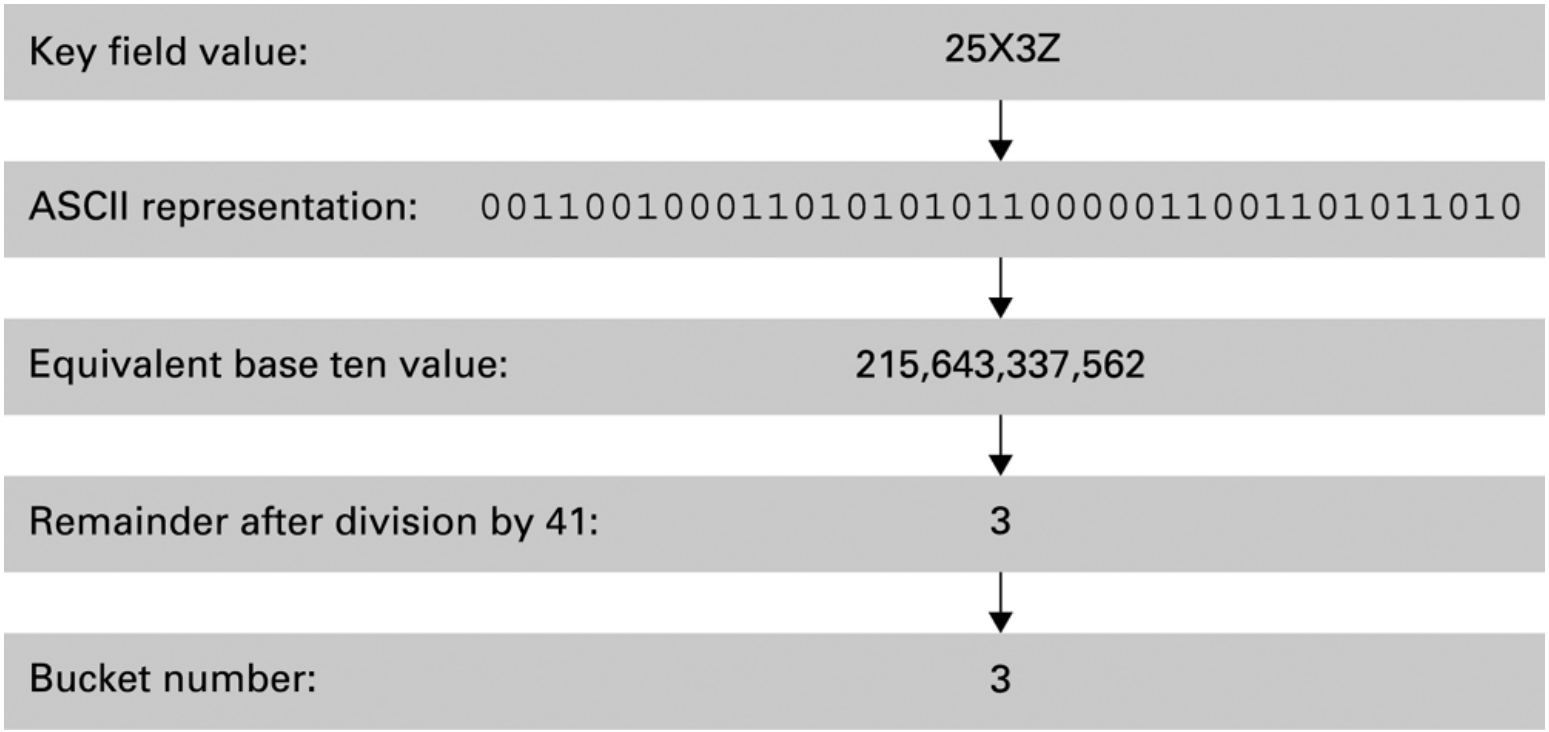
Opening an Indexed File



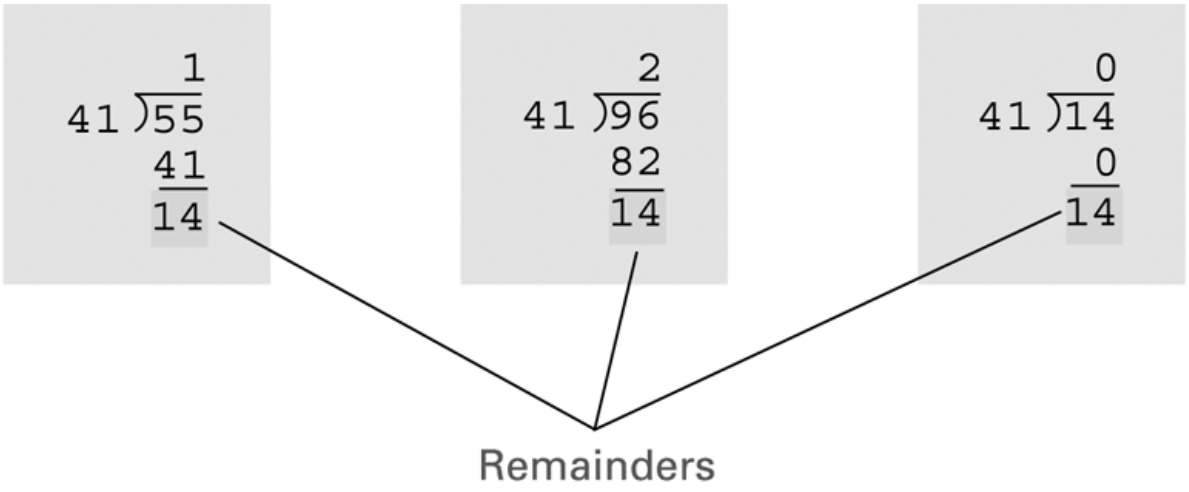
Hashing

- Each record has a key field.
- The storage space is divided into **buckets.**
- A **hash function** computes a bucket number for each key value.
- Each record is stored in the bucket corresponding to the hash of its key.

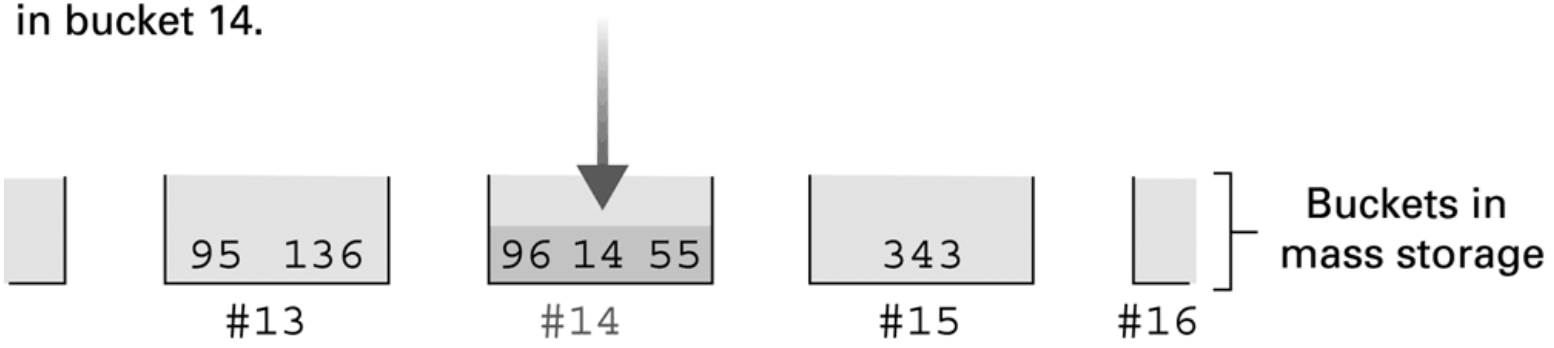
Hashing the Key Field Value 25X3Z to one of 41 Buckets



The Rudiments of a Hashing System



When divided by 41, the key field values of 14, 55, and 96 each produce a remainder of 14. Thus these records are stored in bucket 14.



Collisions in Hashing

- **Collision:** The case of two keys hashing to the same bucket.
  - Major problem when table is over 75% full.
  - Solution: increase number of buckets and rehash all data.

Data Mining

- **Data Mining:** The area of computer science that deals with discovering patterns in collections of data.
- **Data warehouse:** A static data collection to be mined
  - **Data cube:** Data presented from many perspectives to enable mining.

Data Mining Strategies

- Class description
- Class discrimination
- Cluster analysis
- Association analysis
- Outlier analysis
- Sequential pattern analysis

Social Impact of Database Technology

- Problems
  - Massive amounts of personal data are being collected.
    - Often without knowledge or meaningful consent of affected people.
  - Data merging produces new, more invasive information.
  - Errors are widely disseminated and hard to correct.
- Remedies
  - Existing legal remedies often difficult to apply.
  - Negative publicity may be more effective.

Migrations