1 Overview of Database

Chapter Objectives

The concept of databases came into being in the second half of 1960s, and since then numerous improvements have been made for more efficient processing of larger amounts of data.
In this chapter, we get an overall picture of databases.

① Grasping the concept of databases by comparing files and databases, and understanding the structures and characteristics of data models to build databases.
② Understanding data normalization and ERD which are the most important things in database design.
③ Understanding the set and relational operations necessary for database manipulations.

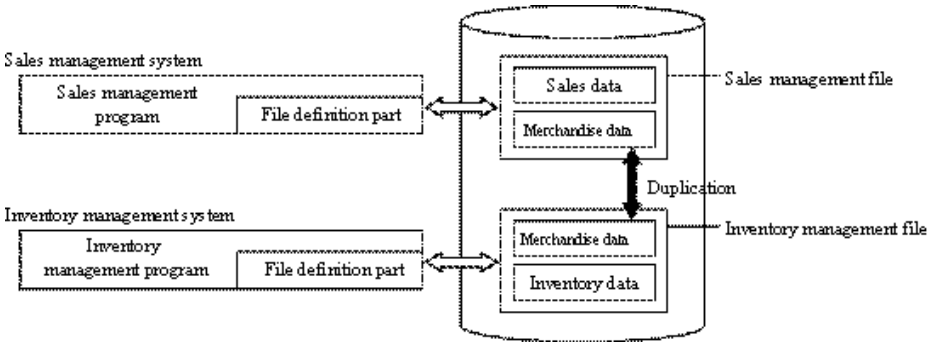
1.1 Purpose of Database

Although we now call a collection of data a database in our daily lives, the word 'database' first appeared in the second half of 1960's.
This section, we'll present the overview and functionalities of the databases which have come to be utilized for efficient processing as the computer application area has expanded.

(1) Problems of file-based systems

In the past file-based systems were created to process large amounts of data efficiently. In such systems, data processing was performed by creating files on magnetic tapes and disks.

Figure 1-1-1 File-Based System



However, as the scale of business and the need to process and operate data for various purposes in various formats increased, some serious problems arised.
The diversification of the purposes and formats of data processing and operation also caused problems.
File-based systems developed for particular uses, for example, have the following problems:

- Because files are created for each application system, a set of same data are recorded in each system, and hardware resources such as magnetic disks are wasted.
- As the data recorded in files is independently changed by the corresponding system, the contents of some data items can be inconsistent with those of the same data items in a different system.
- Because the file definition is included in the program, if file contents and record formats need to be modified, the program also has to be modified.

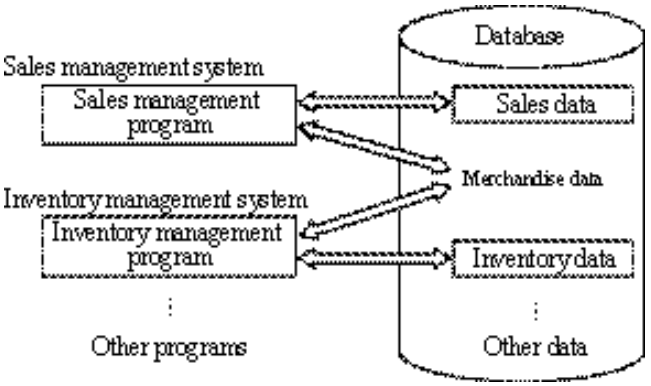
To solve these problems, an idea of database was conceived.

(2) Purposes and functions of database

To solve problems of file-based systems, the following measures are required:

- To eliminate duplication of data items in the related files
- To maintain strict consistency of file contents
- To make programs independent of files

Figure 1-1-2
Concept of Database



More specifically, the following functions and controls are required:

- Data sharing
  By centrally managing files used in an organization data maintenance workload is reduced and data consistency can be maintained.
- Data independent of programs
  By making programs independent of centrally managed databases, program maintenance and modification are become easier.
- Data integrity and failure recovery
  Data integrity must be guaranteed even in the case of supporting a large number of user access, and fast recovery must be made in case of failures.
- Data confidentiality
  Depending on the data contents, access right control is required to allow access only by authorized users.

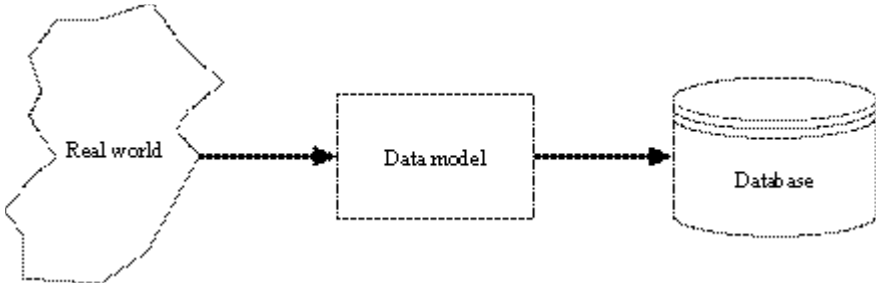
Taking these factors into consideration, databases are built on large-scale direct access storage devices (DASD) such as magnetic disk devices with large storage capacity.

1.2 Database Model

To build a database, a framework which defines the complex real world information and the operations on it is required. This framework is called a "data model." The purposes of data model are as follows:

- To provide conventions for describing data and its structure.
- To define a set of operations for the data represented based on the conventions.
- To provide a framework to describe semantic constraints to correctly represent the information in the real world.

Figure 1-2-1
Data Model



The major roles of a data model can be summarized in the following two items:

- An interface between a database management system (Database Management System software to manage databases: the details explained in Chapter 3) and users. This enables data description and manipulations at the logical level, independent of the physical data storage formats and data retrieval procedures. With this, people can use database without knowing physical-level contents.
- The tool to model the real world
  This provides the framework to represent the data structure and semantics, reflecting the information used in the targeted world as naturally as possible.

1.2.1 Data Modeling

To build a database, the following procedures are carried out to decide its contents:

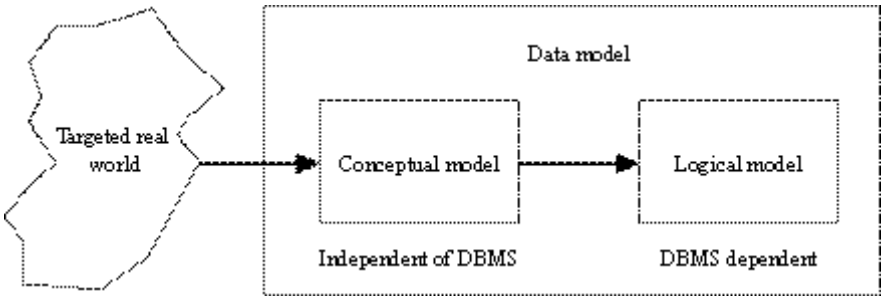
1. Investigate and analyze the complicated information structure, various applications and requirements of the real world.
2. Select information to be arranged into a database.
3. Appropriately structuralize selected data.

These procedures are called "database design." As a result, a mini-world is constructed by modeling and abstracting the targeted world. A series of these processes is generally called "data modeling."
In a database system, data must be described with the manageable data model provided by DBMS. However, describing directly the complex data structure in the real world with the data model provided by DBMS may limit the degree of freedom in representation.

1.2.2 Conceptual Data Model

Even after the completion of a database, natural expressions without constraints imposed by DBMS are necessary to understand the structure and the meaning of data in a database. For this reason, data modeling is generally conducted through at least two steps (Figure 1-2-2).
First, how the target data look like is depicted independently from the data model provided by the DBMS. This is called a "conceptual model." Next, convert this conceptual model into the data model provided by DBMS. This converted model is called a "logical model." This corresponds to the conceptual schema of the three-layer schema mentioned later. A DBMS currently corresponds to either the hierarchical data model, the network data model, or the relational data model.

Figure 1-2-2 Creation Process of Data Model

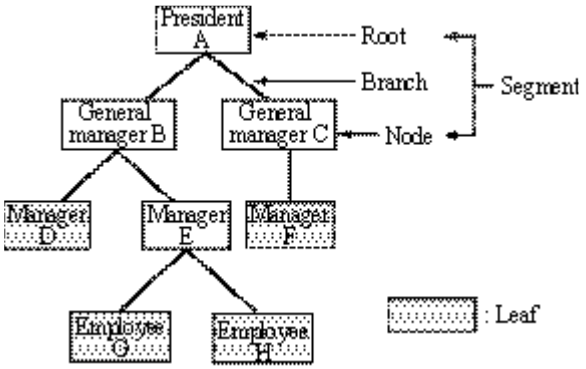


1.2.3 Logical Data Model

(1) Hierarchical data model

The hierarchical data model is a data model employed in IMS (Information Management Systems) which was made public by IBM in 1968. A data set structured based on the hierarchical data model is called the hierarchical database.

Figure 1-2-3
Structure of Hierarchical Data Model



The hierarchical data model consists of the following three kinds of elements:

- Root
  This is the highest-level data, and data retrieval basically begins from the "root."
- Node
  This is the middle-level data. It always has its parent and child (children).
- Leaf
  This is the terminal data, and no data exists below the "leaf" level.

Root and node are sometimes referred to as "segment."
Data are connected by the pointer called branch. The relationship of "root" - "node" and "node" - "leaf" is

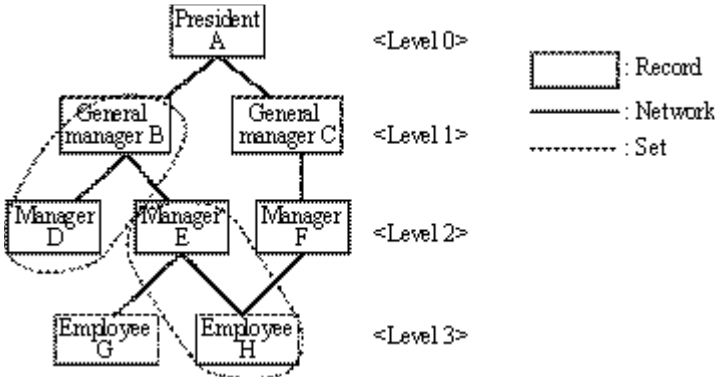
parent and child. A parent can have more than one child, but each child cannot have more than one parent. This is called a parent-child relationship. Therefore, only a single path exists to reach a certain data item.
The Bachman diagram is used to express a hierarchical data model. As shown in Figure 1-2-4, a rectangular box shows a record, and the parent-child relationship is shown by connecting the records with an arrow.

Figure 1-2-4
Bachman Diagram

(2) Network Data Model

A network data model is the one which was employed for IDS (Integrated Data Store) developed by GE in 1963. A data set integrated and based on the network data model is called a network database. Since a network database is designed in accordance with the specifications proposed by CODASYL (Conference on Data Systems Languages), it is also called a CODASYL-type database.
In the network data model, the part corresponding to the segment in the hierarchical data model is called a "record" and records are connected by "network." As records are defined as a parent-child set called "set," a child can have more than one parent. Each hierarchy is called a "level." The levels are defined as level 0, level 1, level 2, ..., and level n, from the highest level towards the lower levels.

Figure 1-2-5 Data Structure of Network Data Model

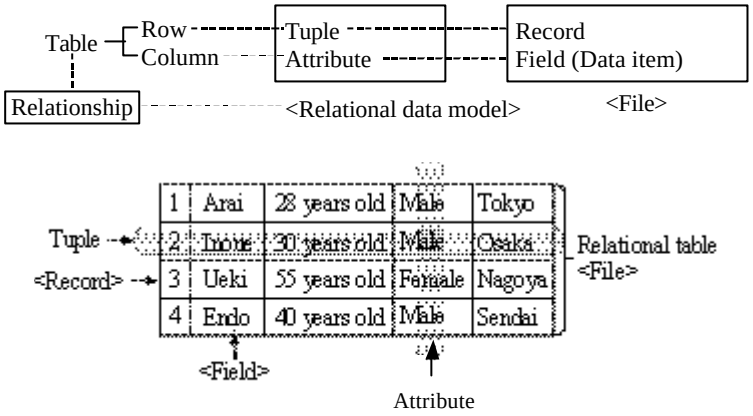


While only one access path to the data exists in the hierarchical data model, multiple access paths can be set in the network data model.

(3) Relational data model

The relational data model is a data model which was proposed by E. F. Codd of IBM in 1970. A data set structured based on the relational data model is called the relational database.
While segments and records are connected by branches and networks in the hierarchical data model and network data model, tables are used in the relational data model. A table consists of rows and columns. A "row" corresponds to a record and a "column" corresponds to a field in a file. In the relational data model, a table is called a "relation," a row a "tuple," and a column an "attribute."

Figure 1-2-6 Structure of Relational Data Model

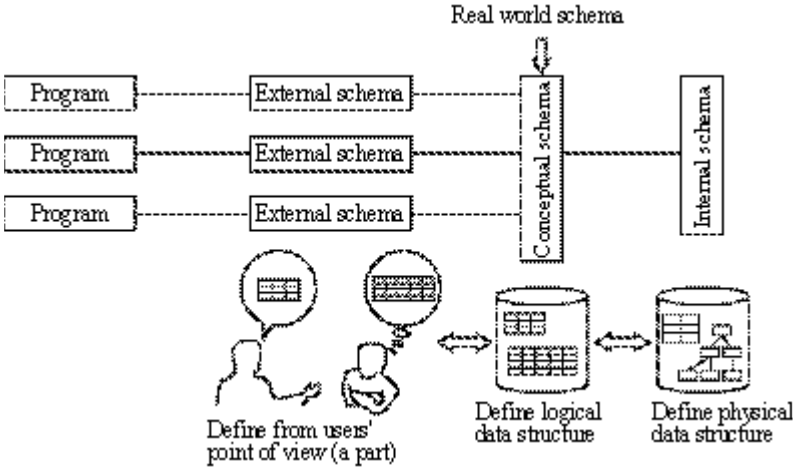


As the structure of the relational data model is simple, data can be freely combined and the operation method is simple enough for end users. The relational data model, therefore, is widely used in various systems ranging from mainframes to personal computers.

1.2.4 3-Tier Schema

As for data modeling, ANSI-SPARC (American National Standard Institute/Systems Planning And Requirements Committee) proposed the 3-tier schema (Figure 1-2-7) in 1978, and it is widely accepted at present.

Figure 1-2-7
3-Tier Schema



In the 3-tier schema, the basic structure of the database system is layered into the following three schemata:

① Conceptual schema

The conceptual schema logically defines the data of the whole real world necessary for the computer system to process. It defines data from its own viewpoint, without taking into consideration the characteristics of computers and programs. One conceptual schema corresponds to one database.

② External schema

The external schema defines the database from the viewpoint of the program using the database. The external schema is considered as part of the data structure defined by the conceptual schema.

③ Internal schema

The internal schema defines how to store physically on storage devices the database defined by the conceptual schema. One internal schema corresponds to one conceptual schema.

The word "schema" as used here means "database description."

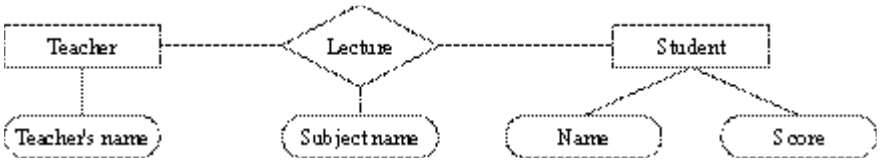
1.3 Data Analysis

1.3.1 ERD

The "Entity-Relationship model (E-R model)" is a diagram expressing the conceptual model, independent of DBMS. The entity-relationship diagram (ERD) is used here. ERD represents the world to be modeled in terms of entities, their relationships and their attributes.
The E-R model consists of the following three elements:

- Entities
  Entities are objects to be managed as depicted by rectangles.
- Relationships
  A relationship indicates a relation between an entity and another entity or a relationship between an entity and a relationship, and is depicted by diamonds.
- Attributes
  Attributes are characteristics of entities and of relationships, and are depicted by ovals.

Figure 1-3-1
E-R Model



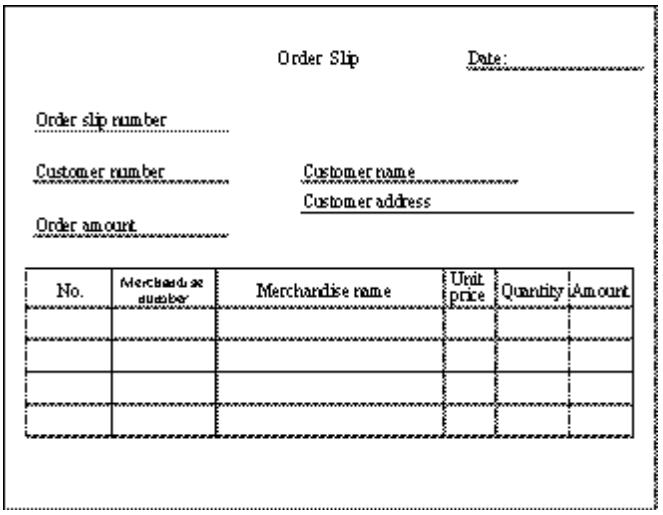
The E-R model in Figure 1-3-1 shows the following:
- "Teacher" and "Student" are connected by "Lecture."
- "Teacher" has "Teacher's name."
- "Student" has "Name" and "Score."
- "Lecture" has "Subject name."

There are three types of relationships: "one-to-one," "one-to-many," and "many-to-many." In Figure 1-3-1, if one teacher gives a lecture to more than one student, and a student receives lectures from more than one teacher, the relationship between "Teacher" and "Student" is "many-to-many."

1.3.2 Normalization

To design a database that fits the users' purposes, the database structure must be thoroughly examined. If not fully examined, users may make demands for other ways to use the database after loading the actual data. Such modifications tend to be very time-consuming and inefficient.
Company A, for example, is a distributor of office automation equipment and uses the order slip shown in Figure 1-3-2.

Figure 1-3-2
Order Slip of Company A



The characteristics of the merchandises, customers, and order-receiving data of Company A are as follows:
- "Customers" are lasting clients and each customer has its own "customer number."
- Each "merchandise" has its "merchandise number" and "unit price."
- "No." is a sequential number for order received for "merchandises."
- "Amount" is calculated by "unit price" × "quantity."
- "Order amount" is the total of "amounts."

Company A plans to design a database of these order slips and related data for efficient order management. For example, when designing a database by the relational data model after deciding the purpose of applications, tables are created by classifying necessary data items to manage. Normalization of data is necessary in this phase. The purpose of normalization is to eliminate the redundancy from data and achieve integrity and consistency of data.
There are five stages for the normalization of a relational database:
- The 1st normalization
- The 2nd normalization
- The 3rd normalization
- The 4th normalization
- The 5th normalization

However, since a relational database requires only the 1st to the 3rd normalization, explanations up to the 3rd normalization are given here.
In the example of Company A, the data items in the order slip can be arranged in a table as shown in Figure 1-3-3.

Figure 1-3-3 Table of Order Slip of Company A (order detail table)

Order slip number	Customer number	Customer name	Customer address	Date	Order amount	No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
						No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
						No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount

The database in this phase is called the unnormalized form (non-1st normal form).
The underlined items here are key items. Key items means the items used to identify records. Thus, if a certain data item is identified, other data items are uniquely determined. This is called "functional dependency (FD)."

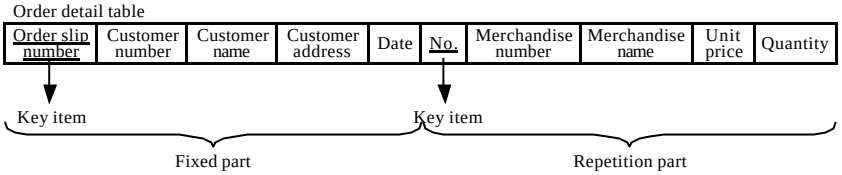
(1) The 1st normalization

There are fixed parts and repetition parts in the unnormalized data as follows:

- Fixed part
  Order slip number, customer number, customer name, customer address, date, and order amount
- Repetition part
  No., merchandise number, merchandise name, unit price, quantity, and amount

In the 1st normalization, data is divided into the fixed part and the repetition part, and the fixed part is overlapped with the repetition part. In this stage, both amount and order amount are excluded because they are decided by calculation of other items, and do not have to be included in the database.
As a result of the 1st normalization, the order slip of Company A is arranged as shown in Figure 1-3-4. This is called the 1st normal form.

Figure 1-3-4
The 1st Normal Form



In the order slip of Company A (unnormalized form), only the slip number was specified as a key item. However, in the 1st normal form, the order slip number and No. are specified as key items because the order slip number cannot specify the repetition items (No., merchandise number, merchandise name, unit price, and quantity). Therefore, combinations of multiple data items such as "slip number + No." are used as concatenated keys.

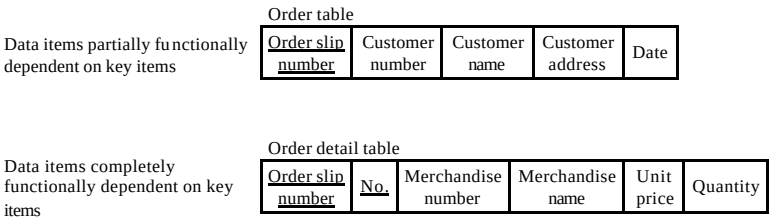
(2) The 2nd normalization

In the 2nd normalization, data items are divided into those data items completely functionally dependent on the key items ("slip number" + "No.") and the data items partially dependent on the key items (functionally dependent on either of the "slip number" or "No.").

- Data items completely functionally dependent on key items
  Merchandise number, merchandise name, unit price, quantity
- Data items partially functionally dependent on key items ("order slip number")
  Customer number, customer name, customer address, date

The result of the 2nd normalization is shown in Figure 1-3-5. This is called the 2nd normal form.

Figure 1-3-5
The 2nd Normal Form



(3) The 3rd normalization

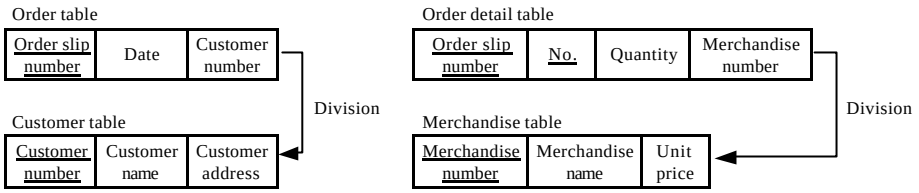
In the 3rd normalization, data items functionally dependent on the data items other than key items, are divided from the data in the 2nd normal form.
The 3rd normalization procedure is as follows:

1. If the customer number is identified, the customer name and the customer address are uniquely determined. So, the order table is divided into the groups of "order slip number and date" and "customer number, customer name, and customer address." "Customer number" is included in the order table to coordinate it to have relationship with the customer table.

2. If the merchandise number is identified, the merchandise name and the unit price are uniquely determined. So, the order table is divided into the groups of "order slip number, No., and quantity" and "merchandise number, merchandise name, and unit price." "Merchandise number" is included in the order table to coordinate it to have relationship with the merchandise table.

The result of the 3rd normalization is shown in Figure 1-3-6. This is called the 3rd normal form.

Figure 1-3-6
The 3rd Normal Form



As the above example, the redundancy of the data can be eliminated by data normalization. Divided tables can be reproduced in the original table in the unnormalized form by means of key items.

Concrete data examples in line with the steps of normalization are shown below. By reference to these examples, we can firmly grasp the image of normalization.

Page 1

November 10, 2000

Order Slip

Order slip number 120131
Customer number 9321 Customer name: Office Ginza Co., Ltd.
Customer address: 1-2-3 Ginza, Chuo-ku
OA Sales Co., Ltd.
Order amount: ¥2,782,000- 138 Soto-kanda, Chiyoda-ku, Tokyo

No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
1	H1010	Notebook-size personal computer	250,000	4	1,000,000
2	H2010	Laser printer	300,000	2	600,000
3	S1040	Integrated software	100,000	1	100,000
4	SP002	A-4 size paper	3,000	2	6,000
5	SP003	B-5 size paper	2,500	4	10,000
6	H0030	Mouse	4,000	4	16,000
7	H1020	Desktop personal computer	180,000	5	900,000
8	S1010	Word processing software	30,000	5	150,000
9		The space below is left blank.			
10					

Page 2

November 18, 2000

Order Slip

Order slip number 120132
Customer number 8109 Customer name: Daiba Sangyo Co., Ltd.
Customer address: 3-2-1 Daiba, Minato-ku
OA Sales Co., Ltd.
Order amount: ¥2,773,000- 138 Soto-kanda, Chiyoda-ku, Tokyo

No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
1	H1010	Notebook-size personal computer	250,000	6	1,500,000
2	H2010	Laser printer	300,000	2	600,000
3	N1030	Terminal adapter	20,000	1	20,000
4	S1040	Integrated software	100,000	4	400,000
5	N0010	LAN cable	1,500	6	9,000
6	N0020	LAN card	5,000	6	30,000
7	S1020	Spreadsheet software	50,000	2	100,000
8	S1010	Word processing software	30,000	2	60,000
9	SP002	A-4 size paper	3,000	10	30,000
10	H0030	Mouse	4,000	6	24,000

Page 3

December 12, 2000

Order Slip

Order slip number 120133
Customer number 9321 Customer name: Office Ginza Co., Ltd.
Customer address: 1-2-3 Ginza, Chuo-ku
OA Sales Co., Ltd.
Order amount: ¥310,500- 138 Soto-kanda, Chiyoda-ku, Tokyo

No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
1	H1020	Desktop personal computer	180,000	1	180,000
2	N1030	Terminal adapter	20,000	1	20,000
3	N0010	LAN cable	1,500	1	1,500
4	N0020	LAN card	5,000	1	5,000
5	S1040	Integrated software	100,000	1	100,000
6	H0030	Mouse	4,000	1	4,000
7		The space below is left blank.			
8					
9					
10					

Page 4

December 12, 2000

Order Slip

Order slip number 120134
Customer number 9321 Customer name: Office Ginza Co., Ltd.
Customer address: 1-2-3 Ginza, Chuo-ku
OA Sales Co., Ltd.
Order amount: ¥1,028,500- 138 Soto-kanda, Chiyoda-ku, Tokyo

No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
1	H1010	Notebook-size personal computer	250,000	2	500,000
2	S1040	Integrated software	100,000	1	100,000
3	H0030	Mouse	4,000	2	8,000
4	SP002	A-4 size paper	3,000	5	15,000
5	SP003	B-5 size paper	2,500	5	12,500
6	N0010	LAN cable	1,500	2	3,000
7	N0020	LAN card	5,000	2	10,000
8	H2010	Laser printer	300,000	1	300,000
9	S1010	Word processing software	30,000	1	30,000
10	S1020	Spreadsheet software	50,000	1	50,000

Order slip/Page 1

Order slip number	Customer number	Customer name	Customer address	Date	Order amount	No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	2,782,000	1	H1010	Notebook-size personal computer	250,000	4	1,000,000
						2	H2010	Laser printer	300,000	2	600,000
						3	S1040	Integrated software	100,000	1	100,000
						4	SP002	A-4 size paper	3,000	2	6,000
						5	SP003	B-5 size paper	2,500	4	10,000
						6	H0030	Mouse	4,000	4	16,000
						7	H1020	Desktop personal computer	180,000	5	900,000
						8	S1010	Word processing software	30,000	5	150,000

Order slip/Page 2

Order slip number	Customer number	Customer name	Customer address	Date	Order amount	No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	2,773,000	1	H1010	Notebook-size personal computer	250,000	6	1,500,000
						2	H2010	Laser printer	300,000	2	600,000
						3	N1030	Terminal adapter	20,000	1	20,000
						4	S1040	Integrated software	100,000	4	400,000
						5	N0010	LAN cable	1,500	6	9,000
						6	N0020	LAN card	5,000	6	30,000
						7	S1020	Spreadsheet software	50,000	2	100,000
						8	S1010	Word processing software	30,000	2	60,000
						9	SP002	A-4 size paper	3,000	10	30,000
						10	H0030	Mouse	4,000	6	24,000

Order slip/Page 3

Order slip number	Customer number	Customer name	Customer address	Date	Order amount	No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	310,500	1	H1020	Desktop personal computer	180,000	1	180,000
						2	N1030	Terminal adapter	20,000	1	20,000
						3	N0010	LAN cable	1,500	1	1,500
						4	N0020	LAN card	5,000	1	5,000
						5	S1040	Integrated software	100,000	1	100,000
						6	H0030	Mouse	4,000	1	4,000

Order slip/Page 4

Order slip number	Customer number	Customer name	Customer address	Date	Order amount	No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	1,028,500	1	H1010	Notebook-size personal computer	250,000	2	500,000
						2	S1040	Integrated software	100,000	1	100,000
						3	H0030	Mouse	4,000	2	8,000
						4	SP002	A-4 size paper	3,000	5	15,000
						5	SP003	B-5 size paper	2,500	5	12,500
						6	N0010	LAN cable	1,500	2	3,000
						7	N0020	LAN card	5,000	2	10,000
						8	H2010	Laser printer	300,000	1	300,000
						9	S1010	Word processing software	30,000	1	30,000
						10	S1020	Spreadsheet software	50,000	1	50,000

The 1st Normal Form

Order detail table

	Order slip number	Customer number	Customer name	Customer address	Date	No.	Merchandise number	Merchandise name	Unit price	Quantity
Page 1	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	1	H1010	Notebook-size personal computer	250,000	4
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	2	H2010	Laser printer	300,000	2
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	3	S1040	Integrated software	100,000	1
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	4	SP002	A-4 size paper	3,000	2
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	5	SP003	B-5 size paper	2,500	4
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	6	H0030	Mouse	4,000	4
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	7	H1020	Desktop personal computer	180,000	5
	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000	8	S1010	Word processing software	30,000	5
Page 2	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	1	H1010	Notebook-size personal computer	250,000	6
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	2	H2010	Laser printer	300,000	2
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	3	N1030	Terminal adapter	20,000	1
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	4	S1040	Integrated software	100,000	4
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	5	N0010	LAN cable	1,500	6
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	6	N0020	LAN card	5,000	6
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	7	S1020	Spreadsheet software	50,000	2
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	8	S1010	Word processing software	30,000	2
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	9	SP002	A-4 size paper	3,000	10
	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000	10	H0030	Mouse	4,000	6
Page 3	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	1	H1020	Desktop personal computer	180,000	1
	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	2	N1030	Terminal adapter	20,000	1
	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	3	N0010	LAN cable	1,500	1
	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	4	N0020	LAN card	5,000	1
	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	5	S1040	Integrated software	100,000	1
	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	6	H0030	Mouse	4,000	1
Page 4	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	1	H1010	Notebook-size personal computer	250,000	2
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	2	S1040	Integrated software	100,000	1
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	3	H0030	Mouse	4,000	2
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	4	SP002	A-4 size paper	3,000	5
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	5	SP003	B-5 size paper	2,500	5
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	6	N0010	LAN cable	1,500	2
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	7	N0020	LAN card	5,000	2
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	8	H2010	Laser printer	300,000	1
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	9	S1010	Word processing software	30,000	1
	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000	10	S1020	Spreadsheet software	50,000	1

The 2nd Normal Form

Order table

	Order slip number	Customer number	Customer name	Customer address	Date
Page 1	120131	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	11/10/2000
Page 2	120132	8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku	11/18/2000
Page 3	120133	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000
Page 4	120134	9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku	12/12/2000

Order detail table

	Order slip number	No.	Merchandise number	Merchandise name	Unit price	Quantity
Page 1	120131	1	H1010	Notebook-size personal computer	250,000	4
	120131	2	H2010	Laser printer	300,000	2
	120131	3	S1040	Integrated software	100,000	1
	120131	4	SP002	A-4 size paper	3,000	2
	120131	5	SP003	B-5 size paper	2,500	4
	120131	6	H0030	Mouse	4,000	4
	120131	7	H1020	Desktop personal computer	180,000	5
	120131	8	S1010	Word processing software	30,000	5
Page 2	120132	1	H1010	Notebook-size personal computer	250,000	6
	120132	2	H2010	Laser printer	300,000	2
	120132	3	N1030	Terminal adapter	20,000	1
	120132	4	S1040	Integrated software	100,000	4
	120132	5	N0010	LAN cable	1,500	6
	120132	6	N0020	LAN card	5,000	6
	120132	7	S1020	Spreadsheet software	50,000	2
	120132	8	S1010	Word processing software	30,000	2
	120132	9	SP002	A-4 size paper	3,000	10
	120132	10	H0030	Mouse	4,000	6
Page 3	120133	1	H1020	Desktop personal computer	180,000	1
	120133	2	N1030	Terminal adapter	20,000	1
	120133	3	N0010	LAN cable	1,500	1
	120133	4	N0020	LAN card	5,000	1
	120133	5	S1040	Integrated software	100,000	1
	120133	6	H0030	Mouse	4,000	1
Page 4	120134	1	H1010	Notebook-size personal computer	250,000	2
	120134	2	S1040	Integrated software	100,000	1
	120134	3	H0030	Mouse	4,000	2
	120134	4	SP002	A-4 size paper	3,000	5
	120134	5	SP003	B-5 size paper	2,500	5
	120134	6	N0010	LAN cable	1,500	2
	120134	7	N0020	LAN card	5,000	2
	120134	8	H2010	Laser printer	300,000	1
	120134	9	S1010	Word processing software	30,000	1
	120134	10	S1020	Spreadsheet software	50,000	1

The 3rd Normal Form

Order table

	Order slip number	Date	Customer number
Page 1	120131	2000/11/10	9321
Page 2	120132	2000/11/18	8109
Page 3	120133	2000/12/12	9321
Page 4	120134	2000/12/12	9321

Customer table

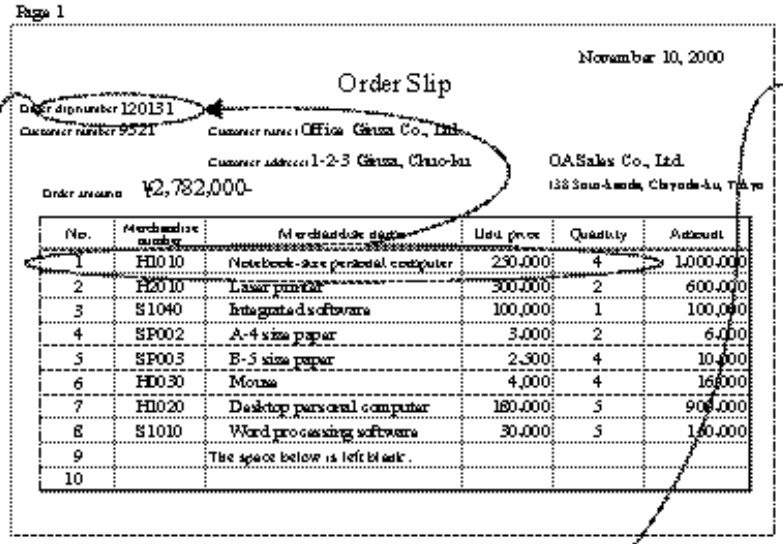
Customer number	Customer name	Customer address
9321	Office Ginza Co., Ltd.	1-2-3 Ginza, Chuo-ku
8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku

Order detail table

	Order slip number	No.	Quantity	Merchandise number
Page 1	120131	1	4	H1010
	120131	2	2	H2010
	120131	3	1	S1040
	120131	4	2	SP002
	120131	5	4	SP003
	120131	6	4	H0030
	120131	7	5	H1020
	120131	8	5	S1010
Page 2	120132	1	6	H1010
	120132	2	2	H2010
	120132	3	1	N1030
	120132	4	4	S1040
	120132	5	6	N0010
	120132	6	6	N0020
	120132	7	2	S1020
	120132	8	2	S1010
	120132	9	10	SP002
	120132	10	6	H0030
Page 3	120133	1	1	H1020
	120133	2	1	N1030
	120133	3	1	N0010
	120133	4	1	N0020
	120133	5	1	S1040
	120133	6	1	H0030
Page 4	120134	1	2	H1010
	120134	2	1	S1040
	120134	3	2	H0030
	120134	4	5	SP002
	120134	5	5	SP003
	120134	6	2	N0010
	120134	7	2	N0020
	120134	8	1	H2010
	120134	9	1	S1010
	120134	10	1	S1020

Merchandise table

Merchandise number	Merchandise name	Unit price
H0030	Mouse	4,000
H1010	Notebook-size personal computer	250,000
H1020	Desktop personal computer	180,000
H2010	Laser printer	300,000
N0010	LAN cable	1,500
N0020	LAN card	5,000
N1030	Terminal adapter	20,000
S1010	Word processing software	30,000
S1020	Spreadsheet software	50,000
S1040	Integrated software	100,000
SP002	A-4 size paper	3,000
SP003	B-5 size paper	2,500


Page 1

November 10, 2000

Order Slip

Order slip number 120131

Customer number 9521

Customer name Office Ginza Co., Ltd.

Customer address 1-2-3 Ginza, Chuo-ku

OA Sales Co., Ltd.

138 Sotokanda, Chiyoda-ku, Tokyo

Order amount ¥2,782,000-

No.	Merchandise number	Merchandise name	Unit price	Quantity	Amount
1	H1010	Notebook-size personal computer	250,000	4	1,000,000
2	H2010	Laser printer	300,000	2	600,000
3	S1040	Integrated software	100,000	1	100,000
4	SP002	A-4 size paper	3,000	2	6,000
5	SP003	B-5 size paper	2,500	4	10,000
6	H0030	Mouse	4,000	4	16,000
7	H1020	Desktop personal computer	180,000	5	900,000
8	S1010	Word processing software	30,000	5	150,000
9		The space below is left blank.			
10					

Order table

Order slip number	Date	Customer number
120131	2000/11/10	9321
120132	2000/11/18	8109
120133	2000/12/12	9321
120134	2000/12/12	9321

Customer table

Customer number	Customer name	Customer address
9321	Office Ginza Co.,Ltd.	1-2-3 Ginza, Chuo-ku
8109	Daiba Sangyo Co., Ltd.	3-2-1 Daiba, Minato-ku

Order detail table

Order slip number	No.	Quantity	Merchandise number
120131	1	4	H1010
120131	2	2	H2010
120131	3	1	S1040
120131	4	2	SP002
120131	5	4	SP003
120131	6	4	H0030
120131	7	5	H1020
120131	8	5	S1010
120132	1	6	H1010
120132	2	2	H2010
120132	3	1	N1030
120132	4	4	S1040
120132	5	6	N0010
120132	6	6	N0020
120132	7	2	S1020
120132	8	2	S1010
120132	9	10	SP002
120132	10	6	H0030
120133	1	1	H1020
120133	2	1	N1030
120133	3	1	N0010
120133	4	1	N0020
120133	5	1	S1040
120133	6	1	H0030
120134	1	2	H1010
120134	2	1	S1040
120134	3	2	H0030
120134	4	5	SP002
120134	5	5	SP003
120134	6	2	N0010
120134	7	2	N0020
120134	8	1	H2010
120134	9	1	S1010
120134	10	1	S1020

Merchandise table

Merchandise number	Merchandise name	Unit price
H0030	Mouse	4,000
H1010	Notebook-size personal computer	250,000
H1020	Desktop personal computer	180,000
H2010	Laser printer	300,000
N0010	LAN cable	1,500
N0020	LAN card	5,000
N1030	Terminal adapter	20,000
S1010	Word processing software	30,000
S1020	Spreadsheet software	50,000
S1040	Integrated software	100,000
SP002	A-4 size paper	3,000
SP003	B-5 size paper	2,500

1.4 Data Manipulation

This chapter explains data manipulation of relational databases by using concrete examples. Data manipulation in information processing consists of four representative set operations (union, difference, intersection, and Cartesian product) and four relational operations (selection, projection, join, and divide) for the relational model.

1.4.1 Set Operation

The following is an explanation of set operations (data manipulation) of union, difference, and intersection using Tables A and B.

Table A: Participants in the Database Course

Employee name	Gender	Extension
Ichiro Higashino	Male	2136
Takako Minamida	Female	2142
Shuhei Nishikawa	Male	2144
Akira Kitayama	Male	2145

Table B: Participants in the Network Course

Employee name	Gender	Extension
Tadanobu Ueno	Male	2134
Ichiro Higashino	Male	2136
Michiko Shimoda	Female	2137
Shuhei Nishikawa	Male	2144
Akira Kitayama	Male	2145
Takao Migita	Male	2146

Of the four set operations, Cartesian product is explained by using Tables C and D on the next page.

(1) Union (A∪B)

Union is also called sum.
For example, union is used for the data manipulation to extract employees who took either of the database courses, or the network course, or both.
When union is used, duplicate tuples (rows) do not exist in the result. Domains of columns corresponding to the two tables must be the same, but column names can be different.

<Operation result>

Employee name	Gender	Extension
Ichiro Higashino	Male	2136
Takako Minamida	Female	2142
Shuhei Nishikawa	Male	2144
Akira Kitayama	Male	2145
Tadanobu Ueno	Male	2134
Michiko Shimoda	Female	2137
Takao Migita	Male	2146

(2) Difference (A–B)

Difference is used to extract employees who did not take the network course, from the participants in the database course.
In the case of difference, as in the case of union, domains of columns corresponding to the two tables must be the same, but column names can be different.

<Operation result>

Employee name	Gender	Extension
Takako Minamida	Female	2142

(3) Intersection (A∩B)

Intersection is also called product.
Intersection is used to extract the employees who took both the database course and the network course.
In the case of intersection, like the above two cases, domains of columns corresponding to the two tables must be the same, but column names can be different.

<Operation result>

Employee name	Gender	Extension
Ichiro Higashino	Male	2136
Shuhei Nishikawa	Male	2144
Akira Kitayama	Male	2145

(4) Cartesian product (C×D)

Cartesian product is used to create a table by combining tuples in the two tables. This operation, however, is transparent to users because it is used for intermediate processing to increase the efficiency of database manipulation.
In Cartesian product, the table name is added before the column name to avoid the duplication of column names, and the number of rows is decided by multiplying the numbers of rows in the two tables.
Table E shows the result of Cartesian product performed on Tables C and D.

Table C: Participant

Employee name	Course code
Masaharu Yamamoto	NE208
Yoko Kawano	DB200

Table D: Course

Course code	Course name
NE208	Network course
DB200	Database course
DB202	SQL course

<Table E: Operation result>

Participant/ Employee name	Participant/ Course code	Course/ Course code	Course/Course name
Masaharu Yamamoto	NE208	NE208	Network course
Masaharu Yamamoto	NE208	DB200	Database course
Masaharu Yamamoto	NE208	DB202	SQL course
Yoko Kawano	DB200	NE208	Network course
Yoko Kawano	DB200	DB200	Database course
Yoko Kawano	DB200	DB202	SQL course

1.4.2 Relational Operation

The following is an explanation of relational operations (data manipulation) of selection, projection, and join using Tables E and F.

Table E: Employee

Employee name	Gender	Extension
Tadanobu Ueno	Male	2134
Ichiro Higashino	Male	2136
Michiko Shimoda	Female	2137
Takako Miyamida	Female	2142
Shuhei Nishikawa	Male	2144
Akira Kitayama	Male	2145
Takao Migita	Male	2146

Table F: Employee Information

Employee name	Native place	Date of employment
Tadanobu Ueno	Tokyo	1993
Ichiro Higashino	Chiba Pref.	1999
Michiko Shimoda	Shizuoka Pref.	1995
Takako Miyamida	Saitama Pref.	1998
Shuhei Nishikawa	Kanagawa Pref.	1995
Akira Kitayama	Fukushima Pref.	1996
Takao Migita	Tochigi Pref.	1994

Of the four relational operations, divide is explained by using Tables G to J on the next page.

(1) Selection

Selection extracts only the rows satisfying the conditions from the specified table.
The following is the result gained by extracting the rows of females from Table E: Employee by selection.

<Operation result>

Employee name	Gender	Extension
Michiko Shimoda	Female	2137
Takako Minamida	Female	2142

(2) Projection

Projection extracts only those columns satisfying conditions from the specified table.
The following is the result gained by extracting the row of gender from Table E: Employee by projection.

<Operation result>

Gender
Male
Female

(3) Join

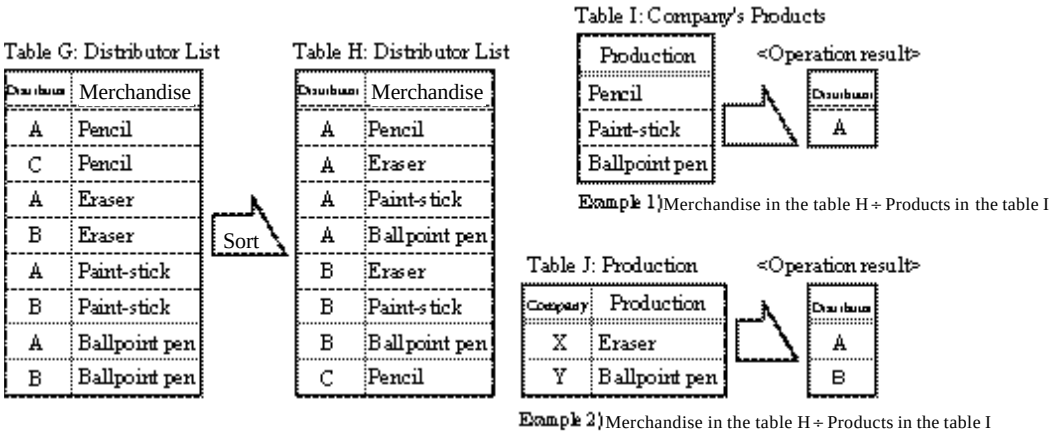
Join is used to create a new table by extracting the necessary columns from the multiple tables.
The table below is an employee list created by extracting all column names from Table E: Employee and Table F: Employee Information by join.

Operation Result: Employee List

Employee name	Gender	Extension	Native place	Date of employment
Tadanobu Ueno	Male	2134	Tokyo	1993
Ichiro Higashino	Male	2136	Chiba Pref.	1999
Michiko Shimoda	Female	2137	Shizuoka Pref.	1995
Takako Miyamida	Female	2142	Saitama Pref.	1998
Shuhei Nishikawa	Male	2144	Kanagawa Pref.	1995
Akira Kitayama	Male	2145	Fukushima Pref.	1996
Takao Migita	Male	2146	Tochigi Pref.	1994

(4) Divide

Divide is used to examine whether the one table completely includes all elements in the other table, by comparing column elements of two tables.
Example 1 below is the divide operation used to extract the distributor that deals in all products in Table I: Company's Products. Example 2 is the divide operation used to extract the distributors that deal in all products in Table J: Production.



Some set and relational operations can be expressed by combining other operations. By combining six operations: union, difference, selection, projection, join, and attribute renaming, all other operations can be expressed. Intersection, for example, can be expressed by using difference as follows:
$A \cap B = A - (A - B)$
In data manipulation of relational databases, at least six operations are necessary.

Exercises

Q1 Choose two effects that can be expected by installing database systems.

a) Reduction of code design works
b) Reduction of duplicate data
c) Increase in the data transfer rate
d) Realization of dynamic access
e) Improvement of independence of programs and data

Q2 Which of the data models shows the relationship between nodes by tree structure?

a) E-R model
b) Hierarchical data model
c) Relational data model
d) Network data model

Q3 Which of the following statements correctly explains relational database?

a) Data are treated as a two-dimensional table from the users' point of view. Relationships between records are defined by the value of fields in each record
b) Relationships between records are expressed by parent-child relationship.
c) Relationships between records are expressed by network structure.
d) Data fields composing a record are stored in the index format by data type. Access to the record is made through the data gathering in these index values.

Q4 Which of the following describes the storage method of databases in storage devices?

a) Conceptual schema
b) External schema
c) Subschema
d) Internal schema

Q5 Which of the following statements correctly explains the 3-tier schema structure of a database?

a) The conceptual schema expresses physical relationships of data.
b) The external schema expresses the data view required by users.
c) The internal schema expresses logical relationships of data.
d) Physical schema expresses physical relationships of data.

Q6 **Which of the following data models is used for the conceptual design of a database, expressing the targeted world by two concepts of entities and relationships between entities?**

a) E-R model
b) Hierarchical data model
c) Relational data model
d) Network data model

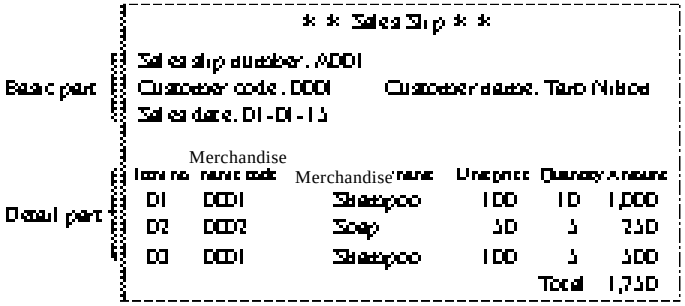
Q7 **In the ERD diagram, the one-to-many relationship, "a company has multiple employees," is expressed as follows:**

Then,

Which of the following statements correctly explains the above diagram?

a) There are multiple companies, and each company has a shareholder.
b) There are multiple companies, and each company has multiple shareholders.
c) One company has one shareholder.
d) One company has multiple shareholders.

Q8 **A database was designed to store the data of the following sales slip. The database is planned to be separated into two tables: the basic part and detail part of the sales slip. The items in the detail part are inputted by reading bar codes on merchandise. Depending on the input method, the same merchandise can appear multiple times in the same sales slip.**
Which of the following combinations is appropriate as key items for the basic part and the detail part? Key values of both parts cannot be duplicated.

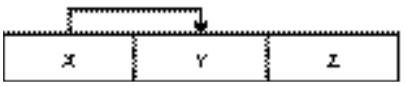


	Basic part	Detail part
a)	Sales slip number	Sales slip number + Item no.
b)	Sales slip number	Sales slip number + Merchandise name code
c)	Customer code	Item no. + Merchandise name code
d)	Customer code	Customer code + Item no.

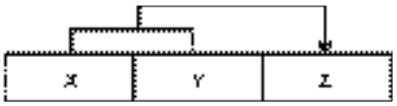
Q9 **Which of the following table structures correctly describes the record consisting of data fields a to e in the 3rd normal form in accordance with the relationships between fields described below?**

[Relationships between fields]

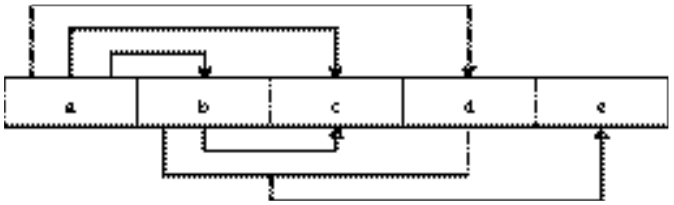
(1) When the value of the field X is given, the value of the field Y can be uniquely identified.



(2) When the values of fields X and Y are given, the value of field Z can be uniquely identified.



[The record to be normalized]



a) a b c d a d e

b) a b c d a d e b c

c) a b c a d e b c d

d) a b d b c b d e

Q10 **A school has recorded information on classes taken by students in the following record format. To create a database from these records, each record must be divided into several parts to avoid the problems of duplicated data. A student takes multiple classes, and multiple students can take one class at the same time. Every student can take a class only once. Which of the following is the most appropriate division pattern?**

Student code	Student name	Class code	Class name	Class finishing year	Score

a)

Student code	Class code

Student name	Class name	Class finishing year	Score

b)

Student code	Student name	Score

Class code	Class name	Class finishing year

c)

Student code	Student name	Class finishing year	Score

Class code	Class name	Student code

d)

Student code	Student name

Class code	Class name

Class finishing year	Score

e)

Student code	Student name

Class code	Class name

Student code	Class code	Class finishing year	Score

Q11 **A culture center examined three types of schemata (data structures) of A to C to manage the customers by using a database. Which of the following statements is correct?**

[Explanation]
A member can take multiple courses.
One course accepts applications from multiple members. Some courses receive no application.
One lecturer takes charge of one course.

Schema A

Member name	Member address	Telephone number	Course name	Lecturer in charge	Lecture fee	Application date

Schema B

Member name	Member address	Telephone number	Course name	Application date

Course name	Lecturer in charge	Lecture fee

Schema C

Member name	Member address	Telephone number

Application date	Member name	Course name

Course name	Member name	Lecture fee

a) In any of the three schemata, when there is any change in the lecturer in charge, you only have to correct the lecturer in charge recorded in the specific row on the database.

b) In any of the three schemata, when you delete the row including the application date to cancel the application for the course, the information on the course related to the cancellation can be removed from the database.

c) In Schemata A and B, when you delete the row including the application date to cancel the application for the course, the information on the member related to the cancellation can be removed from the database.

d) In Schemata B and C, when there is any change in the member address, you only have to correct the member address recorded in the specific row on the database.

e) In Schema C, to delete the information on the member applying for the course, you only have to delete the specific row including the member address.

Q12 **Regarding relational database manipulation, which of the following statements correctly explains projection?**

a) Create a table by combining inquiry results from one table and the ones of the other table.

b) Extract the rows satisfying specific conditions from the table.

c) Extract the specific columns from the table.

d) Create a new table by combining tuples satisfying conditions from tuples in more than two tables.

Q13 Which of the following combinations of manipulations is correct to gain Tables b and c from Table a of the relational database?

Table a

Mountain name	Region
Mt. Fuji	Honshu
Mt. Tarumae	Hokkaido
Yarigatake	Honshu
Yatsugatake	Honshu
Mt. Ishizuchi	Shikoku
Mt. Aso	Kyushu
Nasudake	Honshu
Mt. Kuju	Kyushu
Mt. Daisetsu	Hokkaido

Table b

Mountain name	Region
Mt. Fuji	Honshu
Yarigatake	Honshu
Yatsugatake	Honshu
Nasudake	Honshu

Table c

Region
Honshu
Hokkaido
Shikoku
Kyushu

	Table b	Table c
a)	Projection	Join
b)	Projection	Selection
c)	Selection	Join
d)	Selection	Projection