DATA MANAGEMENT

Data base concepts

DATA MANAGEMENT

- Data management ..most difficult challenges facing today's organizations
- Helps improve productivity..people can find what they need without having to conduct long & difficult search.
- Goal of data management is to provide infrastructure & tools to transform raw data into usable corporate information of highest quality.
- Data ..organizations informational assets

- How you manage financial assets similarly manage data assets.
- WHY????????? To maximize earnings companies invest in data management technologies that increase:
- Opportunity to earn revenues(CRM)
- Ability to cut expenses(inventory management)
- Managers need rapid access to correct , comprehensive & consistent data across enterprise to improve business processes & performance.

- Make decisions , serve customers based on data available to them
- DATABASE or DATAWAREHOUSE
- Databases store enterprise data that their business applications create or generate such as sales , accounting & employee data.
- Data entering databases from POS terminals, online sales other sources stored in organized format so that they can be managed & retrieved.

- Viability of business decisions depends on access to high quality data & quality data depends on effective approaches to data management.
- Information workers…constrained by data that cannot be trusted…incomplete , out of context,outdated,inaccurate,inaccesible , may require weeks to analyze.
- Decision maker facing uncertainty to make intelligent decisions.

- Data errors & inconsistencies..lead to mistakes & lost opportunities..failed deliveries, invoicing blunders ,problems synchronizing data from multiple locations.
- In addition data analysis errors…resulted from use of inaccurate formulas or untested models.
- In retail sector cost of errors due to unreliable data..$40 billion annually
- Healthcare industry…cost of errors billion dollars & thousands of lives.

SO DATA MANAGEMENT IS ALL ABOUT

- DESIGN OF INFRASTUCTURE TO PROVIDE EMPLOYEES WITH COMPLETE TIMELY ACCURATE ACCESSIBLE UNDERSTANDABLE & RELEVANT DATA.
- Is a structured approach for capturing , storing , processing ,integrating ,distributing ,securing & archiving data effectively throughout their life cycle.
- Life cycle…identifies the way data travel through an organization from their capture or creation to their use in supporting data driven solutions such as SCM,CRM,EC….enterprise applications that require current & readily accessible data to function properly.

BASIC DATABASE TERMS

- 0 or 1.....bit
- Collection of 8 bits..... A byte/character
- E.g.: A....
- Collection of bytes/characters.... Field/column
- E.g.: CODE, NAME, ADDRESS,DEPT NAME,CITY are examples of field

- Record/Row: A record is a collection of multiple fields that can be related as a unit.
- In database terminology each row is a record.
- Table/ File : collection of logically related multiple records.. E.g.: A collection of all the employee records of a company would be an employee file.

Figure 18.2 *Table*

- Every record in a file has same set of fields
- Database is a collection of multiple related files (tables).{logical group of related files}
- Records of all non-commercial customers who have a mortgage loan at a financial institution would constitute a data file.
- All customer loan files, such as mortgages & auto ,personal & home equity loans could be grouped to create a non-commercial loan database.

Organizing Data in a Traditional File Environment

THE DATA HIERARCHY

A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.

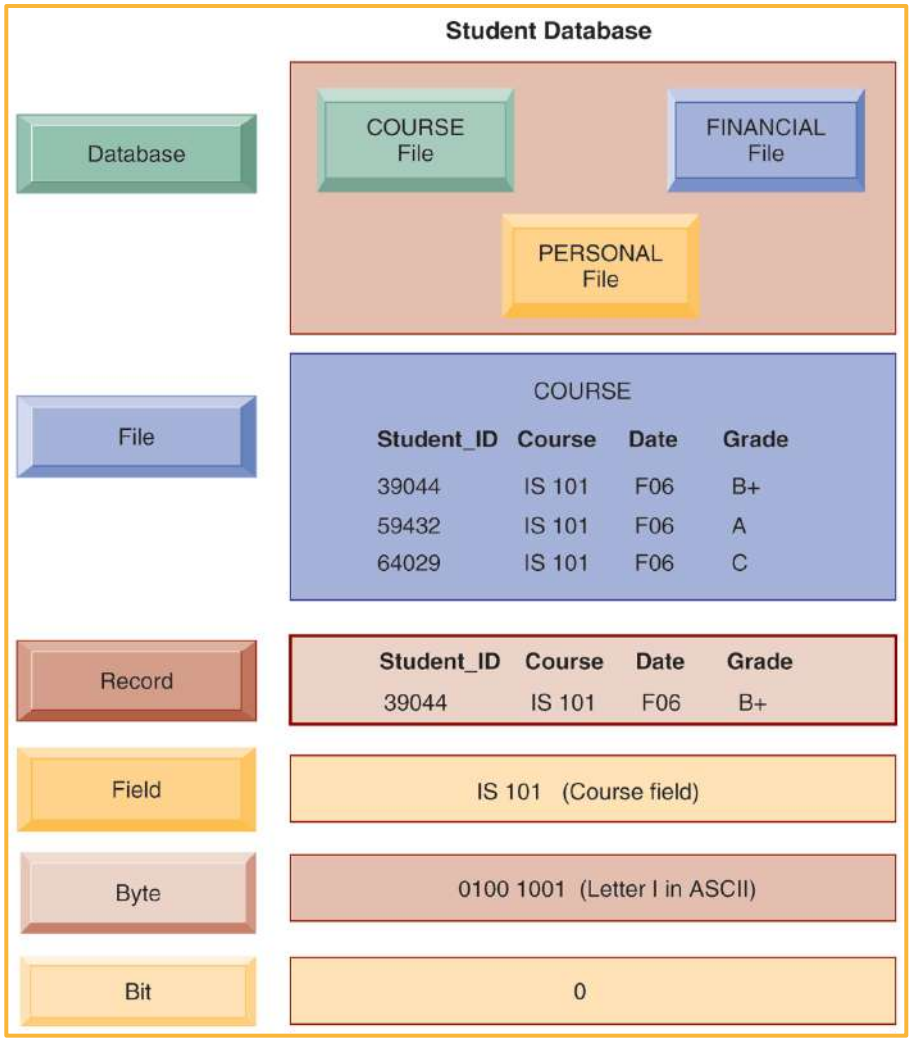


FIGURE 6-1

LOGICAL DATA CONCEPTS: REFERS TO THE MANNER IN WHICH DATA IS VIEWED BY THE PROGRAMMER OR END USER / WAY USER DESCRIBES REALITY

- Entity: is an object that has its existence in the real world.(record describes an entity)
- Includes all those things about which data is collected.
- Entity: tangible....Student/non tangible...job title
- Customer buys goods

Student_ID	Course	Date	Grade
39044	IS 101	F06	B+

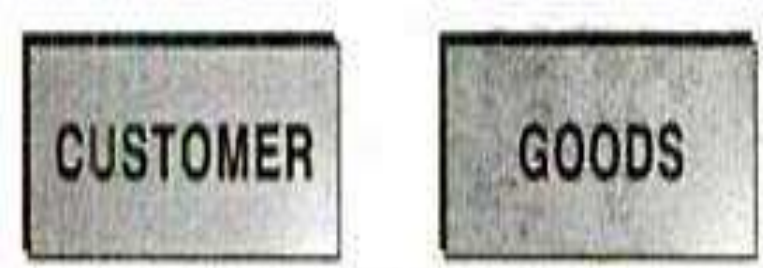
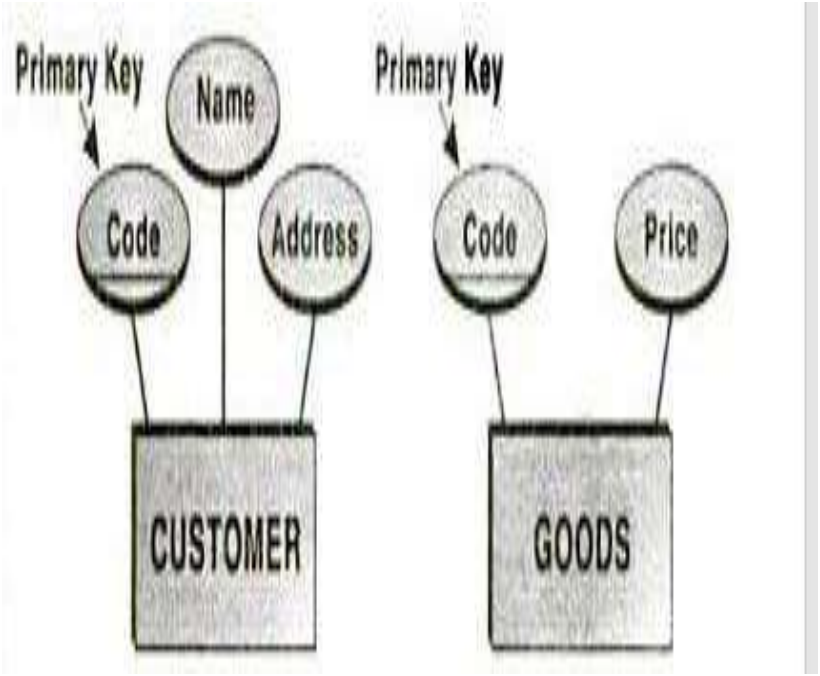
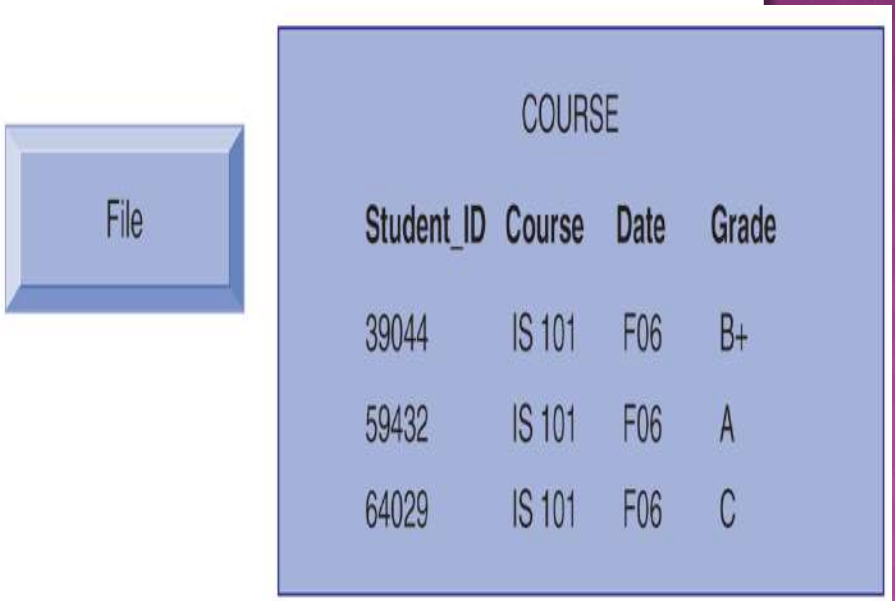


Figure 18.4 *Entities*

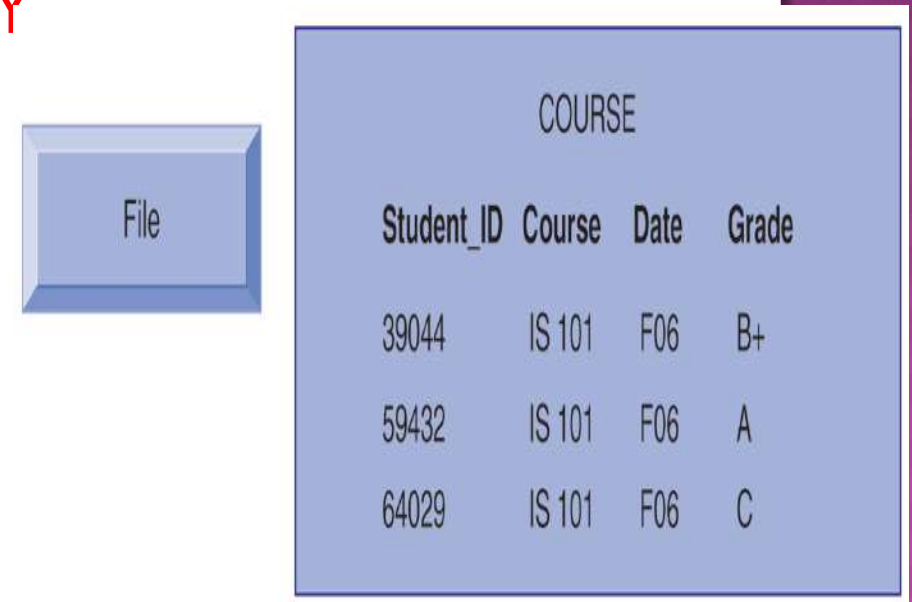
ATTRIBUTE

- Each characteristic describing entity is attribute.
- In a database..entities represented by tables..attributes by columns.





- Each record in a database needs an attribute(field) to uniquely identify it so that record can be retrieved ,updated & sorted……PRIMARY KEY
- Student_ID…PRIMARY KEY
- It is usually numeric



- SECONDARY KEYS: nonunique fields that have some identifying information
- FOREIGN KEYS: keys whose purpose is to link two or more tables together

RELATIONSHIP

- Is an association , dependency or link between two or more entities & is represented by diamond symbol.

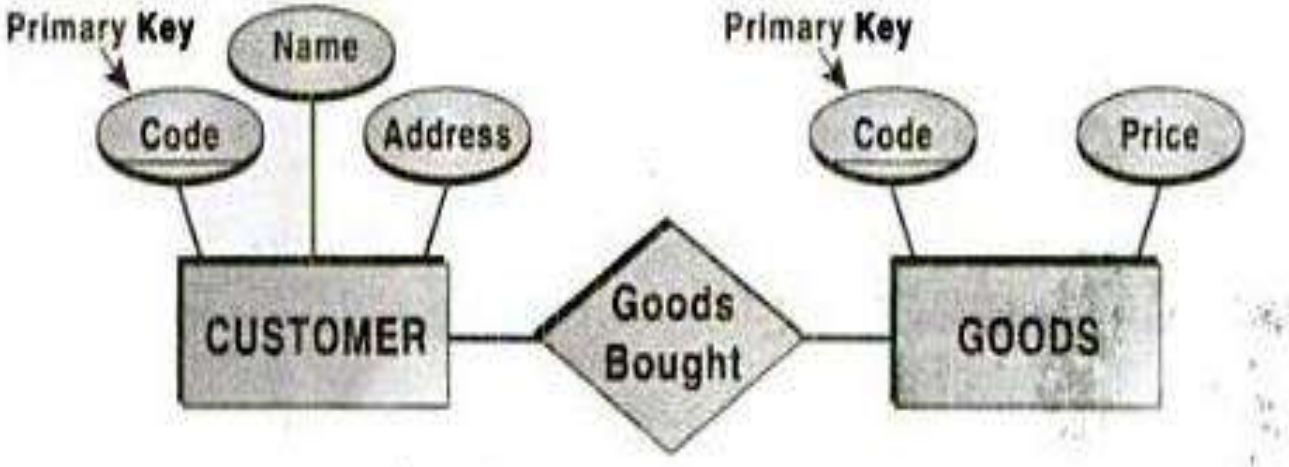


Figure 18.6 *Entities, Attributes, and Relationship*

BINARY RELATIONSHIP TYPES: ONE TO ONE RELATIONSHIP

- One record in a table is related to only one record in another table.
- A department cannot be headed by more than one departmental head

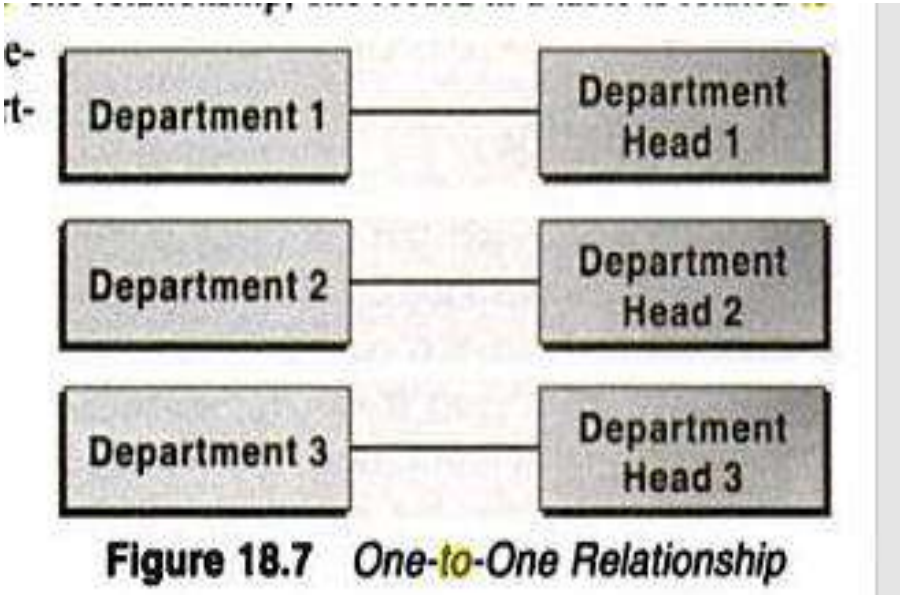


Figure 18.7 *One-to-One Relationship*

ONE TO MANY RELATIONSHIP

- One record in a table (parent table) can be related to many records in another table(child table)

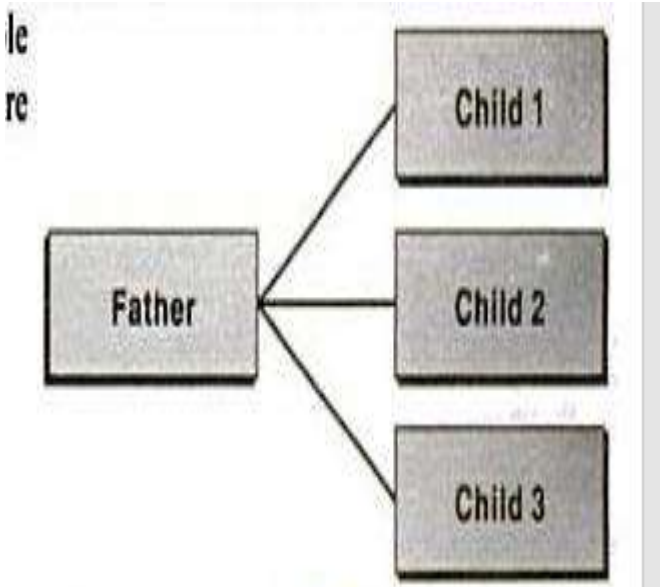


Figure 18.8 *One-to-Many Relationship*

MANY TO MANY RELATIONSHIP

- One record in table can be related to one or ore records in another table, & one or more records in the second table can be related to one or more records in the first table.

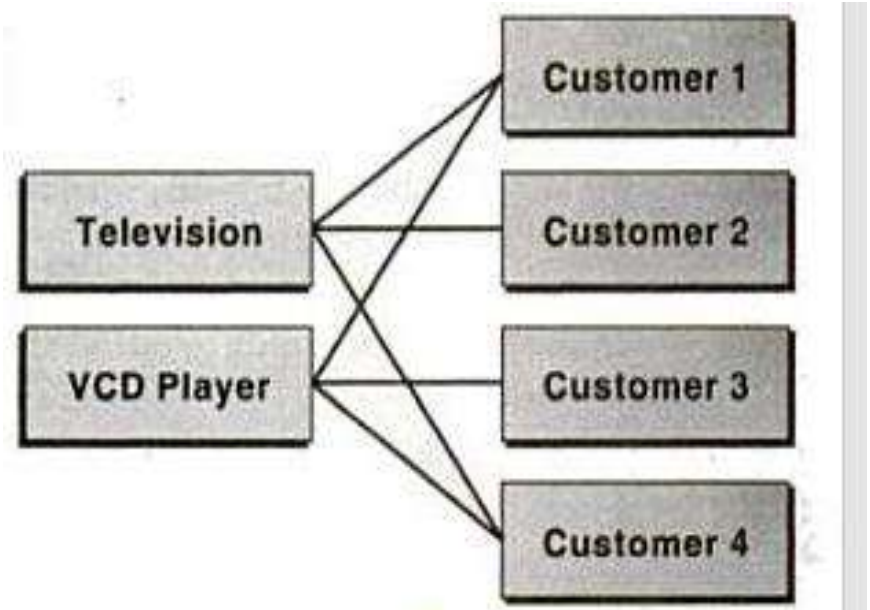


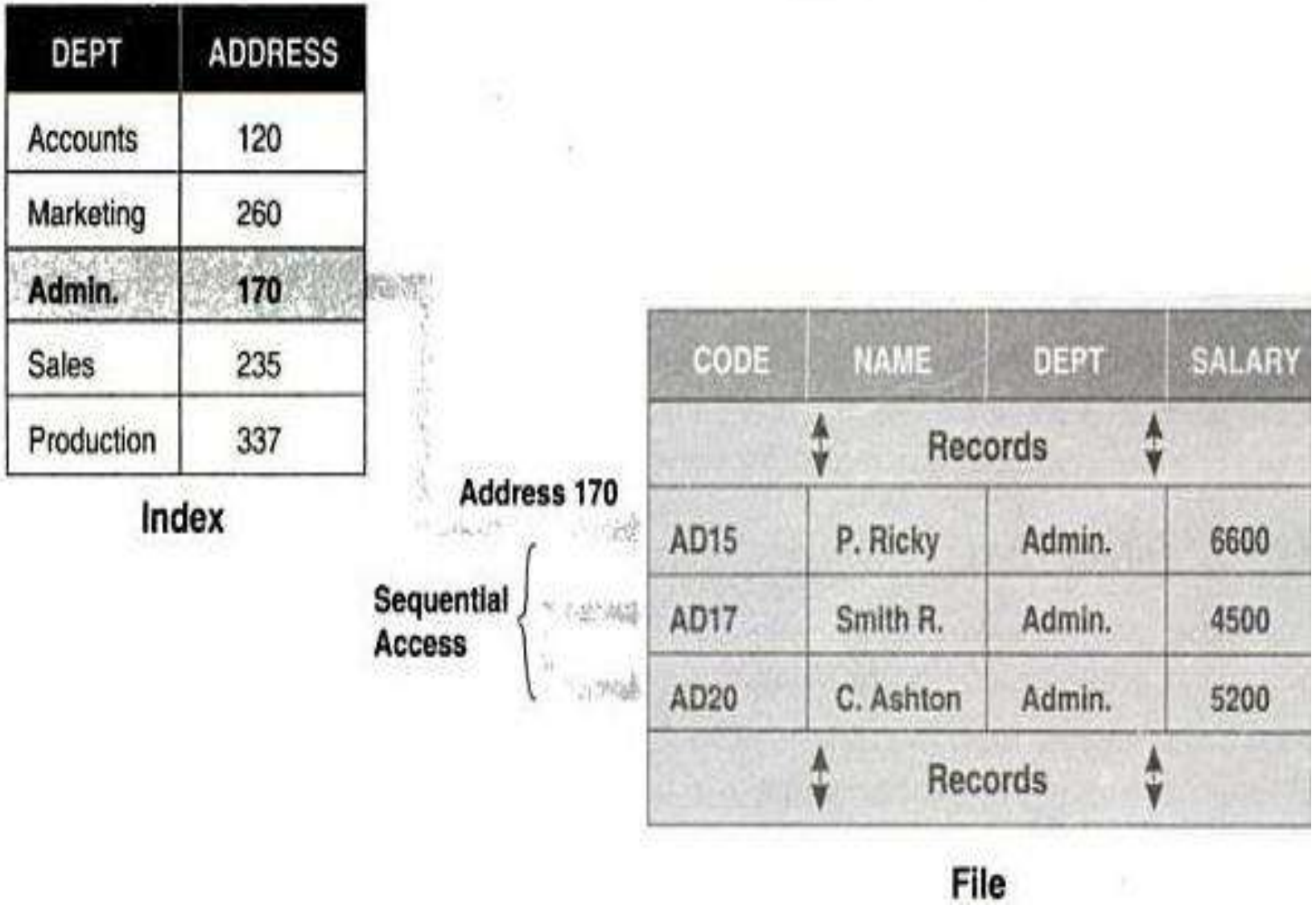
Figure 18.9 *Many-to-Many Relationship*

Physical data concepts: refer to the manner in which data is physically stored on the hardware(accessing records from computer files)

- Sequential file ORGANIZATION: Arrangement determines how individual records can be accessed & how long it takes to access them. (operation like tape recorder)
- Direct file organization or random file organization: access to any record directly without having to traverse the sequence of records (DVD Drive)
- Magnetic tape uses sequential file organization
- Magnetic disks use direct file organization

INDEXED SEQUENTIAL ACCESS METHOD

- Uses an index of key fields to locate individual records.
- An index to a file lists the key field of each record & where the record is physically located on storage media.
- Records are stored on disks in their key sequence.
- To locate a specific record system looks at the index(track index) to locate general location(identified by cylinder & track numbers) containing the record.
- Then points to beginning of that track & reads the records sequentially until it finds correct record.


DEPT
ADDRESS
Accounts
120
Marketing
260
Admin.
170
Sales
235
Production
337
Index
Address 170
Sequential Access
CODE
NAME
DEPT
SALARY
Records
AD15
P. Ricky
Admin.
6600
AD17
Smith R.
Admin.
4500
AD20
C. Ashton
Admin.
5200
Records
File

FILE MANAGEMENT SYSTEMS

- Computer system essentially organizes data into hierarchy that begins with bits & proceeds to bytes,fields,records,files & databases.
- Bit..smallest unit for representing data..0 or 1
- 8bits…byte(10101101).represents single character
- Characters combined…group of words…field
- Key characteristic of a field..All of entries are related in some way

- Cust_name contains names of customers not their number or address
- Related fields..vendor name,address,account data constitute a record
- Collection of related records.. File or Data file(records of all customers who have a mortgage loan..data file)
- Logical group of related files..DATABASE..(All customer loan files such as mortgages & auto, personal & home equity loans..grouped to create noncommercial loan database)

LIMITATIONS OF DATA FILE ENVIRONMENT

- Organizations when began to automate processes they started with one application at a time usually accounting, billing or payroll
- Each application designed to be standalone system worked independently of other applications
- Eg: For each pay period Payroll application would use its own employee & wage data to calculate & process payroll
- No application would use those data without manual intervention.
- This data file approach led to redundancy.

Organizing Data in a Traditional File Environment

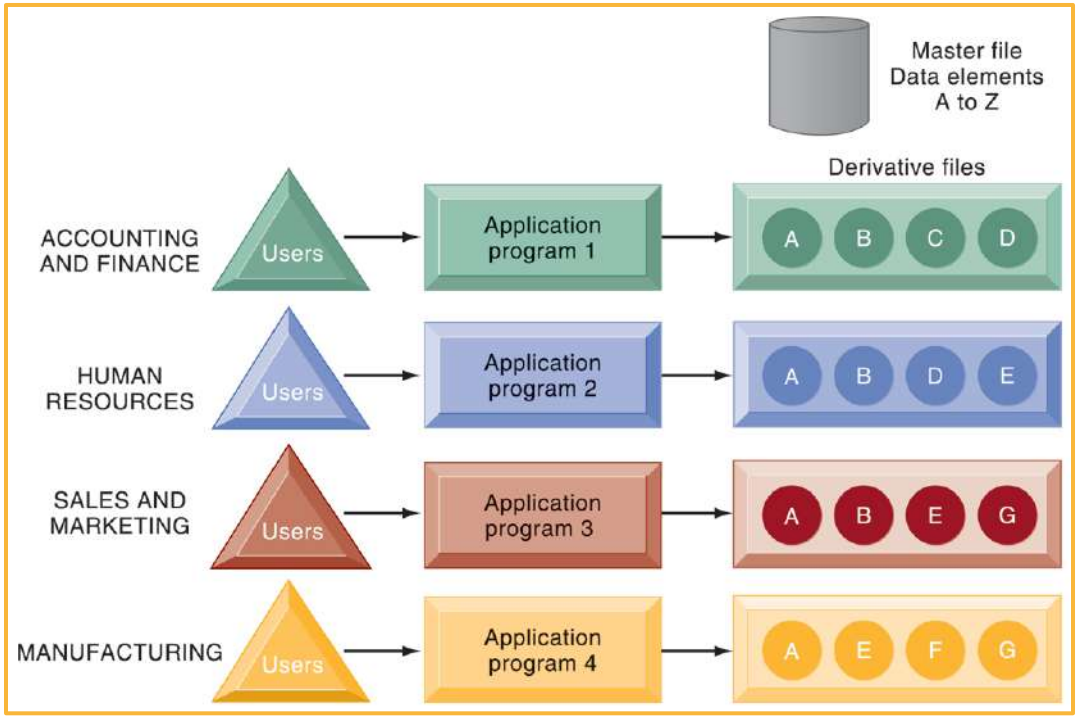


FIGURE 6-2

The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications. Each application requires a unique data file that is likely to be a subset of the master file. These subsets of the master file lead to data redundancy and inconsistency, processing inflexibility, and wasted storage resources.

LIMITATIONS OF DATA FILE ENVIRONMENT

- Data redundancy: Data duplicated in several files..loan example each data file contains records about customers loans . Many of these customers will be represented in other data files.
- Wastes physical storage media
- Difficult to obtain comprehensive view of customers
- Increases costs of entering & maintaining data

LIMITATIONS OF DATA FILE ENVIRONMENT

- Data inconsistency: Actual data values are not synchronized across various copies of the data.
- Financial institution having customers with several loans..each loan there is a file containing customer fields(e.g name,address,email,telephone number)
- Change in customer address in only one file creates inconsistencies with address field in other files.

LIMITATIONS OF DATA FILE ENVIRONMENT

- Data isolation: File organization creates silos of data making it extremely difficult to access data from different applications
- Eg: A manager who wants to know which products customers are buying & which customers owe more than $1000 not be able from a data file system.
- To get results he have to filter & integrate data manually from multiple files.

LIMITATIONS OF DATA FILE ENVIRONMENT

- Data security:
- Lack of data integrity
- Data concurrency..one application updating a record another accessing same record simultaneously. To prevent concurrency applications & data need to be independent of one another.
- File environment applications & data dependent
- To tackle with these problems..development of databases & DBMS

Organizing Data in a Traditional File Environment

- Problems with the traditional file environment (files maintained separately by different departments)
  - Data redundancy:
    - Presence of duplicate data in multiple files
  - Data inconsistency:
    - Same attribute has different values
  - Program-data dependence:
    - When changes in program requires changes to data accessed by program
  - Lack of flexibility
  - Poor security
  - Lack of data sharing and availability

DATABASE MANAGEMENT SYSTEMS

- A program that provides access to databases
- DBMS permits an organization to centralize data ,manage them efficiently & provide access to the stored data by application programs.
- Range in size & capabilities from simple Microsoft Access to full featured Oracle & DB2 solutions.

- DBMS acts as interface between application programs & physical data files.
- Provides users with tools to add ,delete,maintain,display,print,search,select,sort & update data.
- These tools range from easy to use natural language interfaces to complex programming languages used for developing sophisticated database applications.

ADVANTAGES & CAPABILITIES OF DBMS

- Permanence
- Querying
- Concurrency (transactions & locking)
- Backup & replication
- Rule enforcement
- Security
- Computation
- Change & access logging
- Automated optimization

- Companies use DBMSs in a broad range of IS.
- Some DBMS like access can be loaded onto single users computer & accessed in adhoc manner to support individual decision making.
- Others like IBMs DB2 are located on multiple interconnected mainframe computers to support large scale TPSs such as order entry & inventory control systems.
- DBMSs like Oracle 11g are interconnected throughout an organizations LAN(private networks managed & owned by organization) giving departments access to corporate data.

- A DBMS enables different users to share data & process resources
- HOW?? Single unified database meet different requirements of so many users??
- How single database be structured so that sales personnel can view customer , inventory & production maintenance data while HR department maintains restricted access to private personnel data?

The Database Approach to Data Management

HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS

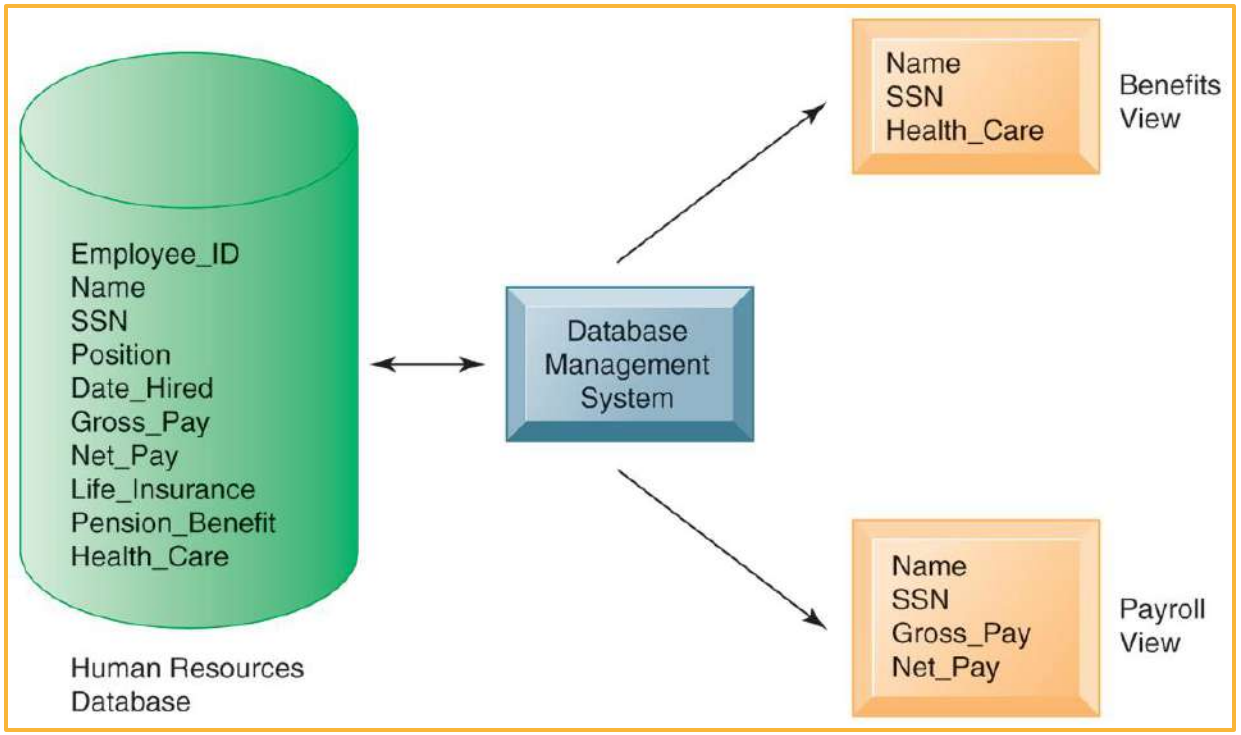


FIGURE 6-3

A single human resources database provides many different views of data, depending on the information requirements of the user. Illustrated here are two possible views, one of interest to a benefits specialist and one of interest to a member of the company's payroll department.

- DBMS provides 2 views of data: PHYSICAL VIEW & A LOGICAL VIEW
- PHYSICAL VIEW..deals with actual physical arrangement & location of data in the DASDs(Direct access storage devices).
- Database specialist use physical view to configure storage & processing resources.

- Business user is interested in using the information not in how it is stored.
- LOGICAL or USERs view of data is meaningful to the user.
- DBMS provides endless logical views of the data.
- This feature allows users to see data from business-related perspective rather than from technical viewpoint.
- Way in which you see the data(logical view) can vary but storage of data(physical view) is fixed.

The Database Approach to Data Management

- Relational DBMS
  - Represent data as two-dimensional tables called relations or files
  - Each table contains data on entity and attributes
- Table: grid of columns and rows
  - Rows (tuples): Records for different entities
  - Fields (columns): Represents attribute for entity
  - Key field: Field used to uniquely identify each record
  - Primary key: Field in table used for key fields
  - Foreign key: Primary key used in second table as look-up field to identify records from original table

The Database Approach to Data Management

RELATIONAL DATABASE TABLES

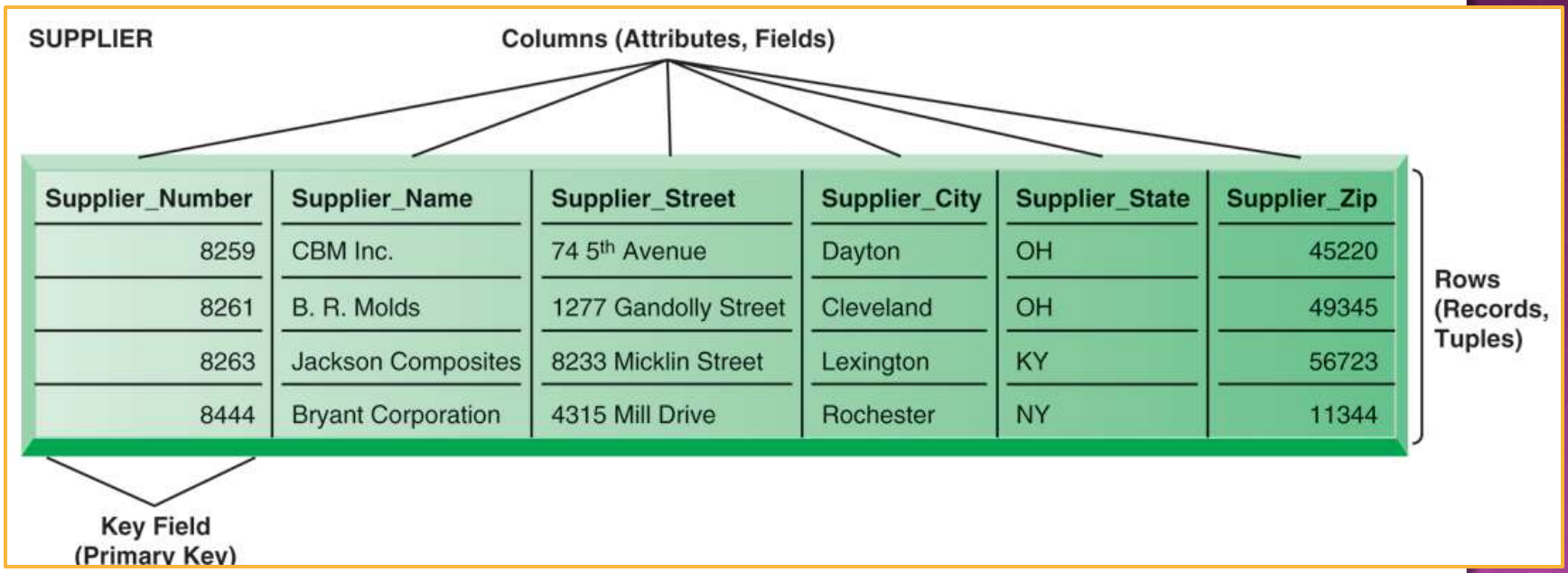
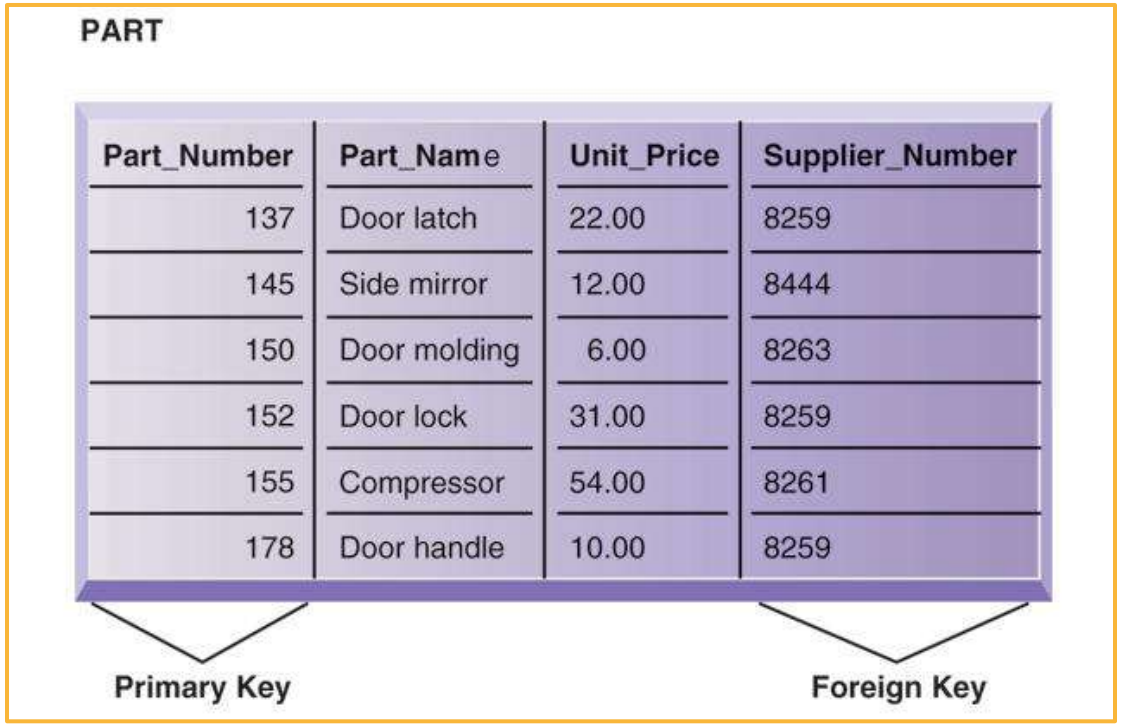


FIGURE 6-4

A relational database organizes data in the form of two-dimensional tables. Illustrated here are tables for the entities SUPPLIER and PART showing how they represent each entity and its attributes. Supplier Number is a primary key for the SUPPLIER table and a foreign key for the PART table.

The Database Approach to Data Management

RELATIONAL DATABASE TABLES (cont.)



PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Primary Key

Foreign Key

FIGURE 6-4
(cont.)

A relational database organizes data in the form of two-dimensional tables. Illustrated here are tables for the entities SUPPLIER and PART showing how they represent each entity and its attributes. Supplier Number is a primary key for the SUPPLIER table and a foreign key for the PART table.

TYPES OF DATABASES

- Data flow into companies from many sources..clickstream data from web,e-commerce applications,POS terminals,filtered data from CRM,Supply chain,ERP applications
- Two basic types of databases:
- Centralized databases
- Distributed databases

CENTRALIZED DATABASES

- Stores all related files in one physical location.
- Decades main database platform consisted centralized database files on large, mainframe because of enormous capital & operating costs associated with alternative systems.
- Application processing is shared between numerous clients & a server...DB/2 from IBM & Oracle server from Oracle

CENTRALIZED DATABASE

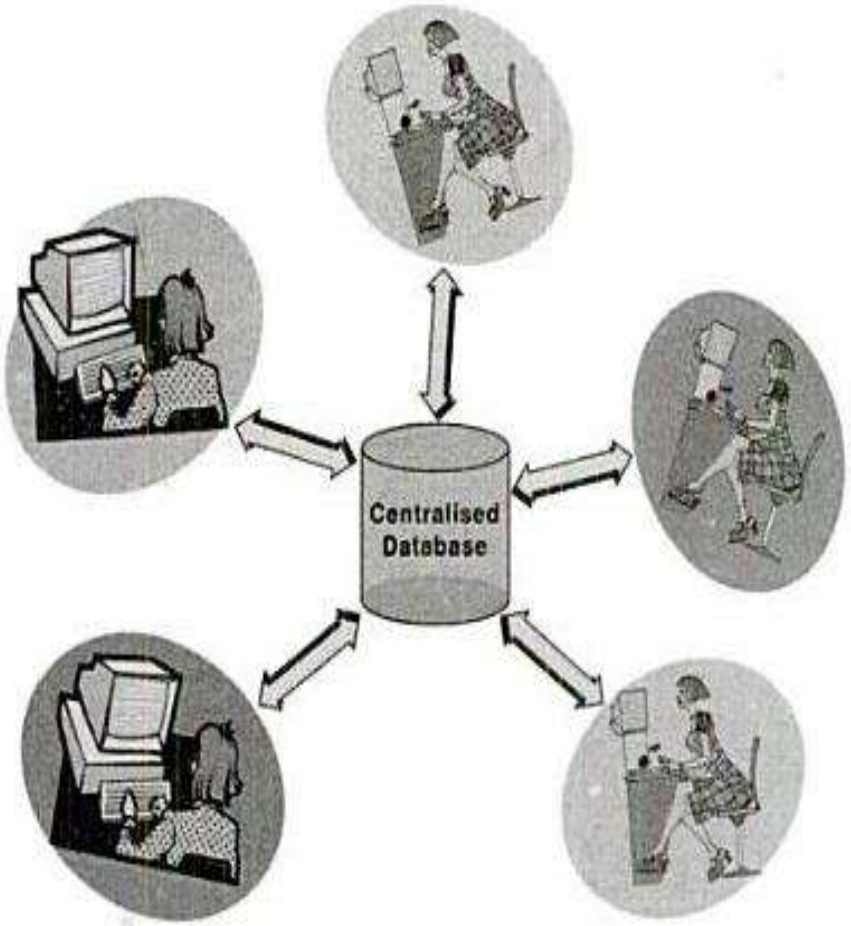


Figure 18.21 *Centralised Database*

CENTRALIZED DATABASES

- Advantages: Multiple processors are applied to overall task , back end & front end are being done in parallel. Thus response time & throughput are improved.
- Different clients able to access same server machine

- Files made more consistent with one another..physically kept at one location..file changes made in supervised & orderly fashion.
- Files are not accessible except via centralized host computer where they can be protected more easily from unauthorized access or modification.
- Vulnerable to single point failure.

CENTRALISED DATABASE

- Disadvantages: Since all data is stored at one place any discrepancy may result in database corruption
- Server must be able to grow in power & capacity to accommodate more clients otherwise it will become bottleneck.
- Centralised computing is more complex because proper processing requires close communication between clients & server hence specialised & expensive tools are necessary.

DISTRIBUTED DATABASE

- A distributed database has complete copies of a database or portions of a database.

Two types:

- Replicated
- partitioned

PARTITIONED DATABASE

- Partitioned: Separate locations store different parts of database(part that meets users local needs)
- Partitioned databases provide response speed of localized files without need to replicate all changes in multiple locations.
- Data in files can be entered more quickly & kept more accurate by users immediately responsible for the data.

DISADVANTAGE OF PARTITIONED DATABASE

- Widespread access to potentially sensitive company data can significantly increase security problems.

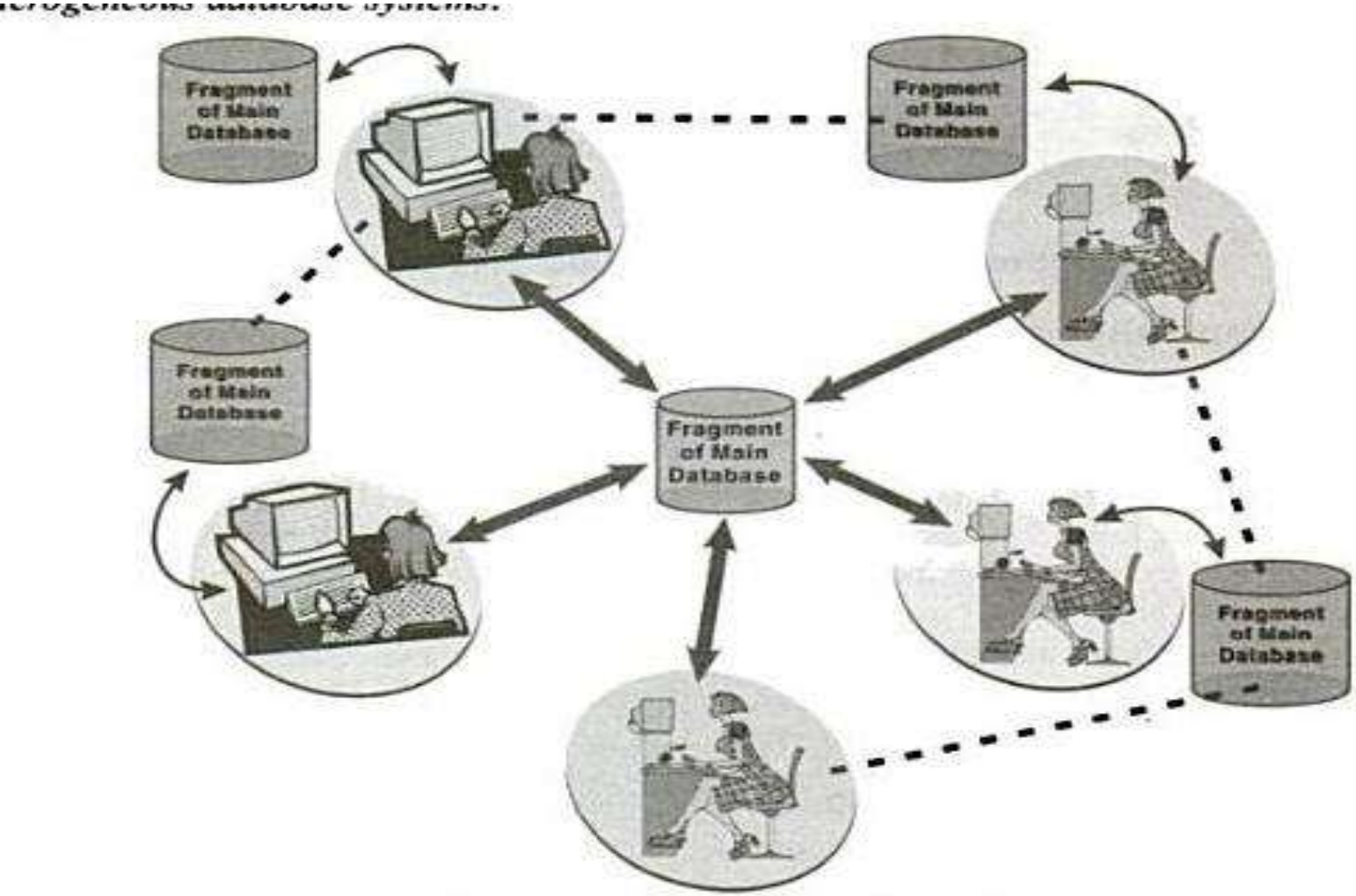
REPLICATED DATABASE

- Stores complete copies of entire database in multiple locations.
- Arrangement provides backup in case of a failure or problem
- Improves response time as it is closer to users
- Expensive to set up & maintain..replica must be updated as records are added to, modified in & deleted from any of the databases.
- Updates may be done at end of day or other schedule as determined by business needs
- If not done various databases will have conflicting data.

DISTRIBUTED DATABASE

- Distributing databases: Storing database in more than one place
- Advantages: continue to function at some reduced level even when a component fails
- Speeds up query processing
- Reduce communication costs
- Easier & more economical to add a local computer

DISTRIBUTED DATABASE



DISTRIBUTED DATABASE

- Disadvantages: Complex software required
- Various sites must exchange messages & perform additional calculations to ensure proper coordination among the sites

NORMALISATION

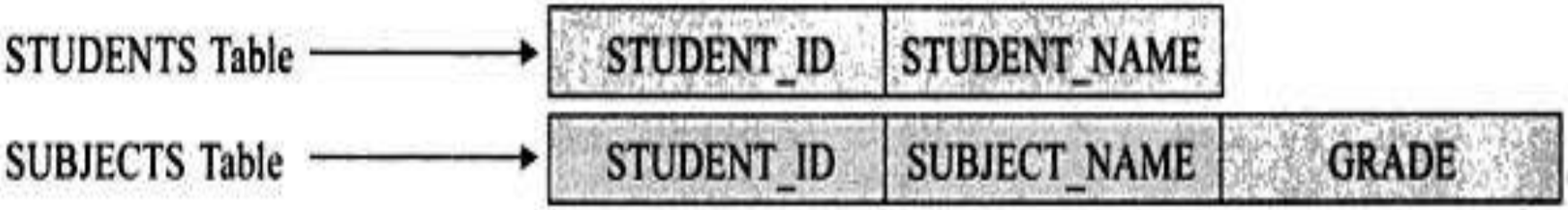
- **Database normalization** is the process of organizing the fields and tables of a relational database to minimize redundancy and dependency.
- Normalization usually involves dividing large tables into smaller (and less redundant) tables and defining relationships between them.
- The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database using the defined relationships.

- Rules of normalisation are referred to as normal forms:
- 1NF
- 2NF
- 3NF

FIRST NORMAL FORM(1NF)

- A table is in 1NF if it contains no repeated groups i.e no two fields stores same kind of information in a single table.

STUDENT_ID	STUDENT_NAME	SUBJECT_1	SUBJECT_2



Example

item	colors	price	tax
T-shirt	red, blue	12.00	0.60
polo	red, yellow	12.00	0.60
T-shirt	red, blue	12.00	0.60
sweatshirt	blue, black	25.00	1.25

Table is not in first normal form because:

- Multiple items in color field
- Duplicate records / no primary key

Example

item	color	price	tax
T-shirt	red	12.00	0.60
T-shirt	blue	12.00	0.60
polo	red	12.00	0.60
polo	yellow	12.00	0.60
sweatshirt	blue	25.00	1.25
sweatshirt	black	25.00	1.25

Table is now in first normal form.

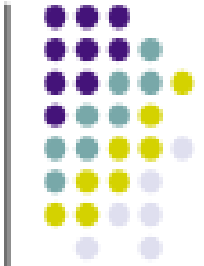
SECOND NORMAL FORM

- Depends on the concepts of primary key & functional dependency
- A database is in 2NF if it is in 1NF & every attribute is fully functionally dependent on the primary key.
- Thus the relation is in 1NF with no repeating groups & all non key attributes must depend on the whole key not just some part of it.

SECOND NORMAL FORM

- Functionally dependent means given a primary key value of any attribute can be retrieved.
- If the student_ID is given , the value of student Name can be obtained
- Hence student name is functionally dependent on student_ID

Example



item	color	price	tax
T-shirt	red	12.00	0.60
T-shirt	blue	12.00	0.60
polo	red	12.00	0.60
polo	yellow	12.00	0.60
sweatshirt	blue	25.00	1.25
sweatshirt	black	25.00	1.25

Table is not in second normal form because:

- **price** and **tax** depend on **item**, but not **color**

Example

item	color
T-shirt	red
T-shirt	blue
polo	red
polo	yellow
sweatshirt	blue
sweatshirt	black

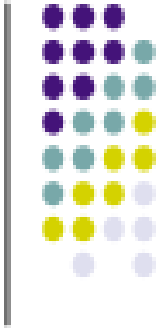
item	price	tax
T-shirt	12.00	0.60
polo	12.00	0.60
sweatshirt	25.00	1.25

Tables are now in second normal form.

THIRD NORMAL FORM

- A table is said to be in 3NF if all the non-key fields are independent i.e. no two non-key fields of the table are dependent on each other.
- Removes redundant data by removing fields that are not wholly dependent on the primary key.
- A table is said to be in 3NF if it is in 2NF & every non-key field is non-transitively dependent on the primary key.

Example



item	color
T-shirt	red
T-shirt	blue
polo	red
polo	yellow
sweatshirt	blue
sweatshirt	black

item	price	tax
T-shirt	12.00	0.60
polo	12.00	0.60
sweatshirt	25.00	1.25

Tables are not in third normal form because:

- **tax** depends on **price**, not **item**

Example

item	color
T-shirt	red
T-shirt	blue
polo	red
polo	yellow
sweatshirt	blue
sweatshirt	black

item	price
T-shirt	12.00
polo	12.00
sweatshirt	25.00

price	tax
12.00	0.60
25.00	1.25

Tables are now in third normal form.