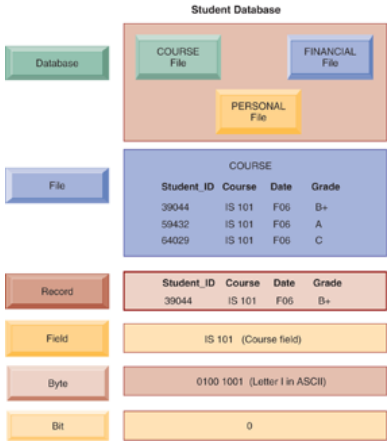
Foundations of Business Intelligence: Databases and Information Management

- **Problem:** HP's numerous systems unable to deliver the information needed for a complete picture of business operations, lack of data consistency
- **Solutions: Build a data warehouse with a single global enterprise-wide database;** replacing 17 database technologies and 14,000 databases in use
- Created consistent data models for all enterprise data and proprietary platform
- Demonstrates importance of database management in creating timely, accurate data and reports
- Illustrates need to standardize how data from disparate sources are stored, organized, and managed

- **File organization concepts**
  - **Computer system organizes data in a hierarchy**
    - **Field:** Group of characters as word(s) or number
    - **Record:** Group of related fields
    - **File:** Group of records of same type
    - **Database:** Group of related files
  - **Record**: Describes an entity
  - **Entity**: Person, place, thing on which we store information
    - Attribute: Each characteristic, or quality, describing entity
      - E.g., Attributes **Date** or **Grade** belong to entity **COURSE**

The Data Hierarchy

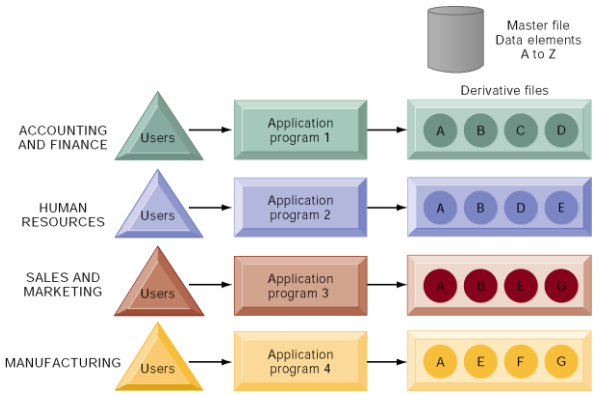


A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can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

Figure 6-1

- **Problems with the traditional file environment (files maintained separately by different departments)**
  - **Data redundancy and inconsistency**
    - **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**

Traditional File Processing

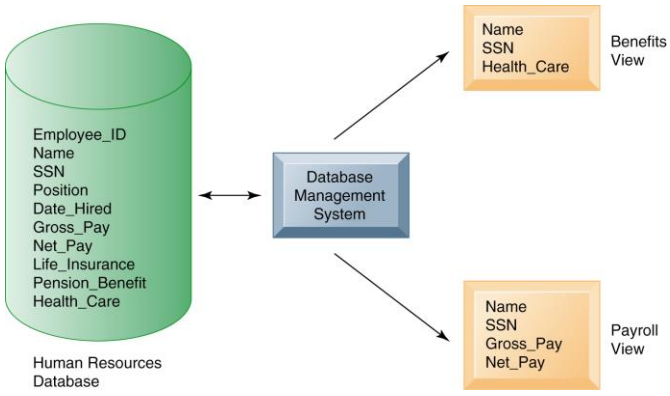


The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications and files. Each application requires a unique data file that is likely to be a subset of the master file. These subsets of the master file lead to data redundancy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

Figure 6-2

- **Database**
  - Collection of data organized to serve many applications by centralizing data and controlling redundant data
- **Database management system**
  - Interfaces between application programs and physical data files
  - Separates logical and physical views of data
  - Solves problems of traditional file environment
    - Controls redundancy
    - Eliminates inconsistency
    - Uncouples programs and data
    - Enables organization to central manage data and data security

Human Resources Database with Multiple Views



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.

Figure 6-3

- **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

Relational Database Tables

SUPPLIER

Columns (Attributes, Fields)

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Rows (Records, Tuples)

Key Field (Primary Key)

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.

Figure 6-4A

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

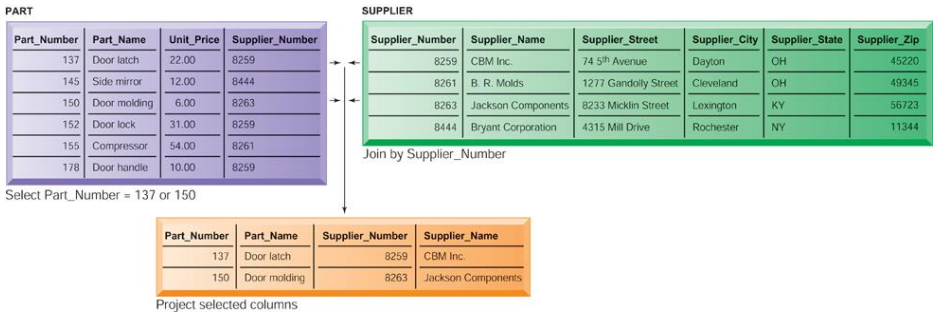
Primary Key

Foreign Key

Figure 6-4B

- **Operations of a Relational DBMS**
- Three basic operations used to develop useful sets of data
  - **SELECT**: Creates subset of data of all records that meet stated criteria
  - **JOIN**: Combines relational tables to provide user with more information than available in individual tables
  - **PROJECT**: Creates subset of columns in table, creating tables with only the information specified

The Three Basic Operations of a Relational DBMS



The select, project, and join operations enable data from two different tables to be combined and only selected attributes to be displayed.

Figure 6-5

- **Object-Oriented DBMS (OODBMS)**
  - Stores data and procedures as objects
  - Capable of managing graphics, multimedia, Java applets
  - Relatively slow compared with relational DBMS for processing large numbers of transactions
  - **Hybrid object-relational DBMS:** Provide capabilities of both OODBMS and relational DBMS

- **Capabilities of Database Management Systems**
  - **Data definition capability:** Specifies structure of database content, used to create tables and define characteristics of fields
  - **Data dictionary:** Automated or manual file storing definitions of data elements and their characteristics
  - **Data manipulation language:** Used to add, change, delete, retrieve data from database
    - Structured Query Language (SQL)
    - Microsoft Access user tools for generation SQL
  - Many DBMS have **report generation capabilities** for creating polished reports (Crystal Reports)

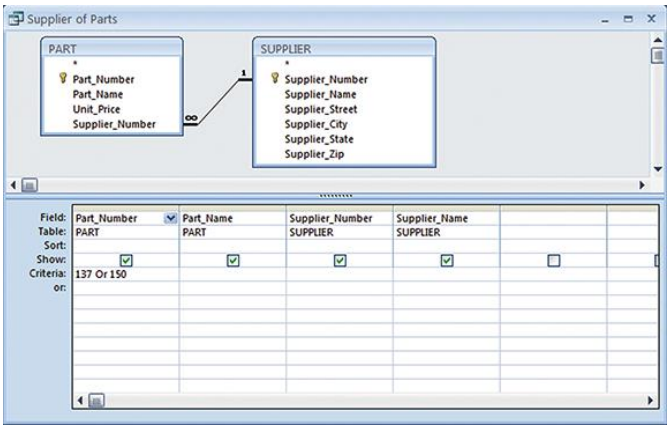
Example of an SQL Query

```
SELECT PART.Part_Number, PART.Part_Name, SUPPLIER.Supplier_Number,
SUPPLIER.Supplier_Name
FROM PART, SUPPLIER
WHERE PART.Supplier_Number = SUPPLIER.Supplier_Number AND
Part_Number = 137 OR Part_Number = 150;
```

Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6-5.

Figure 6-7

An Access Query



Illustrated here is how the query in Figure 6-7 would be constructed using query-building tools in the Access Query Design View. It shows the tables, fields, and selection criteria used for the query.

Figure 6-8

- **Designing Databases**
  - Conceptual (logical) design: abstract model from business perspective
  - Physical design: How database is arranged on direct-access storage devices
- **Design process identifies**
  - Relationships among data elements, redundant database elements
  - Most efficient way to group data elements to meet business requirements, needs of application programs
- **Normalization**
  - Streamlining complex groupings of data to minimize redundant data elements and awkward many-to-many relationships

The Database Approach to Data Management

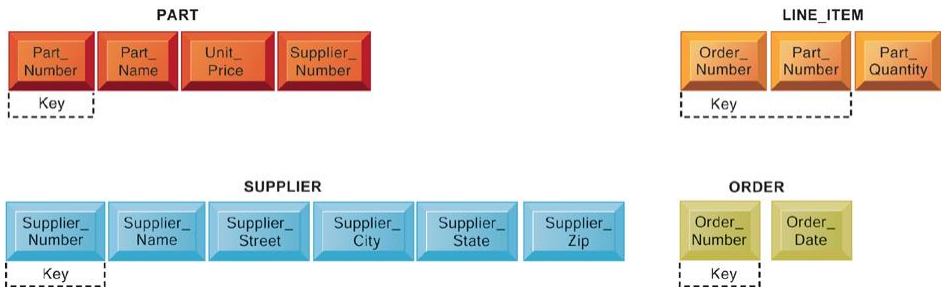
An Unnormalized Relation for Order

An unnormalized relation contains repeating groups. For example, there can be many parts and suppliers for each order. There is only a one-to-one correspondence between Order_Number and Order_Date.

Figure 6-9

The Database Approach to Data Management

Normalized Tables Created from Order

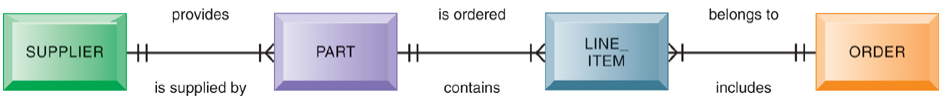


After normalization, the original relation ORDER has been broken down into four smaller relations. The relation ORDER is left with only two attributes and the relation LINE_ITEM has a combined, or concatenated, key consisting of Order_Number and Part_Number.

Figure 6-10

- **Entity-relationship diagram**
  - Used by database designers to document the data model
  - Illustrates relationships between entities
- **Distributing databases**: Storing database in more than one place
  - **Partitioned**: Separate locations store different parts of database
  - **Replicated**: Central database duplicated in entirety at different locations

An Entity-Relationship Diagram



This diagram shows the relationships between the entities ORDER, LINE_ITEM, PART, and SUPPLIER that might be used to model the database in Figure 6-10.

Figure 6-11

The Database Approach to Data Management

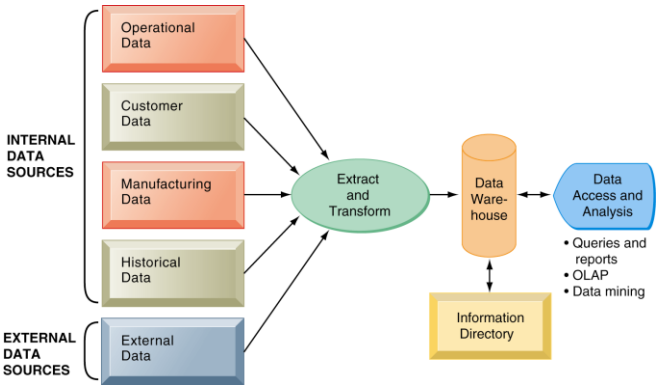
- **Distributing databases**
  - Two main methods of distributing a database
    - **Partitioned**: Separate locations store different parts of database
    - **Replicated**: Central database duplicated in entirety at different locations
  - Advantages
    - Reduced vulnerability
    - Increased responsiveness
  - Drawbacks
    - Departures from using standard definitions
    - Security problems

Using Databases to Improve Business Performance and Decision Making

- Very large databases and systems require special capabilities, tools
  - To analyze large quantities of data
  - To access data from multiple systems
- Three key techniques
  - Data warehousing
  - Data mining
  - Tools for accessing internal databases through the Web

- **Data warehouse:**
  - Stores current and historical data from many core operational transaction systems
  - Consolidates and standardizes information for use across enterprise, but data cannot be altered
  - Data warehouse system will provide query, analysis, and reporting tools
- **Data marts:**
  - Subset of data warehouse
  - Summarized or highly focused portion of firm's data for use by specific population of users
  - Typically focuses on single subject or line of business

Components of a Data Warehouse



The data warehouse extracts current and historical data from multiple operational systems inside the organization. These data are combined with data from external sources and reorganized into a central database designed for management reporting and analysis. The information directory provides users with information about the data available in the warehouse.

Figure 6-13

The IRS Uncovers Tax Fraud with a Data Warehouse

- **Read the Interactive Session: Organizations, and then discuss the following questions:**
  - **Why was it so difficult for the IRS to analyze the taxpayer data it had collected?**
  - **What kind of challenges did the IRS encounter when implementing its CDW? What management, organization, and technology issues had to be addressed?**
  - **How did the CDW improve decision making and operations at the IRS? Are there benefits to taxpayers?**
  - **Do you think data warehouses could be useful in other areas of the federal sector? Which ones? Why or why not?**

- **Business Intelligence:**
  - Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
  - E.g., Harrah's Entertainment analyzes customers to develop gambling profiles and identify most profitable customers
  - Principle tools include:
    - Software for database query and reporting
    - Online analytical processing (OLAP)
    - Data mining

Using Databases to Improve Business Performance and Decision Making

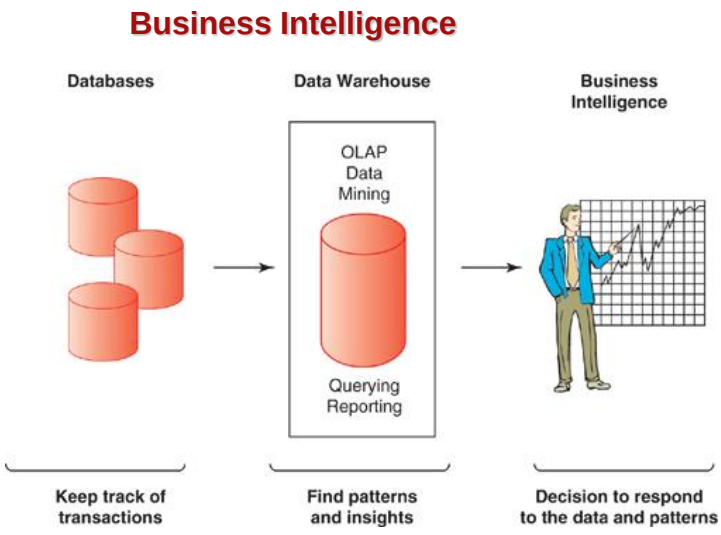


Figure 6-14

A series of analytical tools works with data stored in databases to find patterns and insights for helping managers and employees make better decisions to improve organizational performance.

Using Databases to Improve Business Performance and Decision Making

- **Online analytical processing (OLAP)**
  - Supports multidimensional data analysis
    - Viewing data using multiple dimensions
    - Each aspect of information (product, pricing, cost, region, time period) is different dimension
    - E.g., how many washers sold in East in June compared with other regions?
  - OLAP enables rapid, online answers to ad hoc queries

Using Databases to Improve Business Performance and Decision Making

Multidimensional Data Model

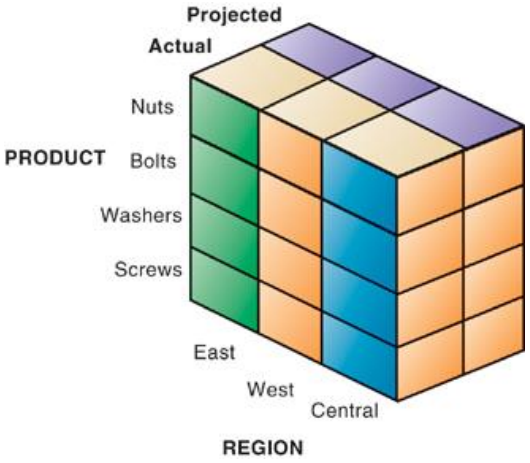


Figure 6-15

The view that is showing is product versus region. If you rotate the cube 90 degrees, the face that will show is product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. Other views are possible.

Using Databases to Improve Business Performance and Decision Making

- **Data mining:**
  - More discovery driven than OLAP
  - Finds hidden patterns, relationships in large databases and infers rules to predict future behavior
  - E.g., Finding patterns in customer data for one-to-one marketing campaigns or to identify profitable customers.
- **Key areas where businesses are leveraging data mining include:**
  - Customer segmentation
  - Marketing and promotion targeting
  - Market basket analysis
  - Collaborative filtering
  - Customer churn
  - Fraud detection
  - Financial modeling
  - Hiring and promotion

- **Data mining:. Types of information obtainable from data mining**
- **Associations**- An association algorithm creates rules that describe how often events have occurred together.
  - Example: When a customer buys a hammer, then 90% of the time they will buy nails.
- **Sequences**- Events linked over time
- **Classification** - Recognizes patterns that describe group to which item belongs-
  - Example: A bank wants to classify its Home Loan Customers into groups according to their response to bank advertisements. The bank might use the classifications "Responds Rarely, Responds Sometimes, Responds Frequently".
- **Clustering -** Similar to classification, but when no groups have been defined; finds groupings within data
  - Example: Insurance company could use clustering to group clients by their age, location and types of insurance purchased.
  - The categories are unspecified and this is referred to as 'unsupervised learning'
- **Forecasting -** Uses series of existing values to forecast what other values will be
  - We'll do this in class with regression analysis
  - Regression deals with the prediction of a value, rather than a class
  - Example: Find out if there is a relationship between smoking patients and cancer related illness.

Data Mining

- A data mining and business analytics team should possesses three critical skills:
  - Information technology
  - Statistics
  - Business knowledge

- **Predictive analysis**
  - Uses data mining techniques, historical data, and assumptions about future conditions to predict outcomes of events
  - E.g., Probability a customer will respond to an offer or purchase a specific product
- **Text mining**
  - Extracts key elements from large unstructured data sets (e.g., stored e-mails)

Artificial Intelligence

- Data Mining has its roots in a branch of computer science known as artificial intelligence (AI)
- The goal of AI is create computer programs that are able to mimic or improve upon functions of the human brain

Artificial Intelligence

- **Neural network**: An AI system that examines data and hunts down and exposes patterns, in order to build models to exploit findings
- **Expert systems**: AI systems that leverage rules or examples to perform a task in a way that mimics applied human expertise
- **Genetic algorithms**: Model building techniques where computers examine many potential solutions to a problem, iteratively modifying various mathematical models, and comparing the mutated models to search for a best alternative

- **Web mining**
  - Discovery and analysis of useful patterns and information from WWW
    - E.g., to understand customer behavior, evaluate effectiveness of Web site, etc.
  - Techniques
    - Web content mining
      - Knowledge extracted from content of Web pages
    - Web structure mining
      - E.g., links to and from Web page
    - Web usage mining
      - User interaction data recorded by Web server

Using Databases to Improve Business Performance and Decision Making

- **Databases and the Web**
  - Many companies use Web to make some internal databases available to customers or partners
  - Typical configuration includes:
    - Web server
    - Application server/middleware/CGI scripts
    - Database server (hosting DBM)
  - Advantages of using Web for database access:
    - Ease of use of browser software
    - Web interface requires few or no changes to database
    - Inexpensive to add Web interface to system

Managing Data Resources

- **Establishing an information policy**
  - Firm's rules, procedures, roles for sharing, managing, standardizing data
    - E.g., What employees are responsible for updating sensitive employee information
  - **Data administration**: Firm function responsible for specific policies and procedures to manage data
  - **Data governance:** Policies and processes for managing availability, usability, integrity, and security of enterprise data, especially as it relates to government regulations
  - **Database administration :** Defining, organizing, implementing, maintaining database; performed by database design and management group

Managing Data Resources

- **Ensuring data quality**
  - More than 25% of critical data in Fortune 1000 company databases are inaccurate or incomplete
  - Most data quality problems stem from faulty input
  - Before new database in place, need to:
    - Identify and correct faulty data
    - Establish better routines for editing data once database in operation

Managing Data Resources

- **Data quality audit**:
  - Structured survey of the accuracy and level of completeness of the data in an information system
    - Survey samples from data files, or
    - Survey end users for perceptions of quality
- **Data cleansing**
  - Software to detect and correct data that are incorrect, incomplete, improperly formatted, or redundant
  - Enforces consistency among different sets of data from separate information systems

Privacy Concerns

- Effective Data Mining requires large sources of data
- To achieve a wide spectrum of data, must link multiple data sources
- Linking sources leads can be problematic for privacy as follows: If the following histories of a customer were linked:
  - Shopping History
  - Credit History
  - Bank History
  - Employment History

- The users' life story can be painted from the collected data
- Hiring, loan, other decision are made by data collected on individuals.
  - What happens if the data is not correct?
- Data aggregators (data brokers) – it's legal to buy and sell personal data.
  - Is this ethical?