Database System
and
Big Data

IE 5602 ICT for Industrial Engineering

Introduction

- Database: an organized collection of data
- A database management system (DBMS) is a group of programs that:
  - Manipulate the database
  - Provide an interface between the database and its users and other application programs

Data Fundamentals

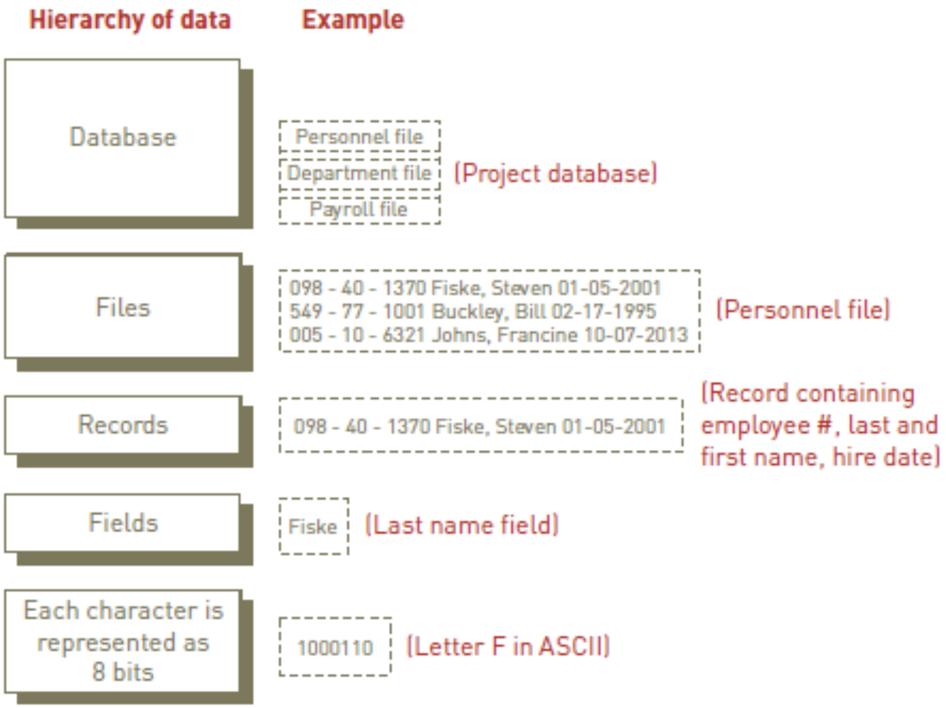
- Without data and the ability to process it:
  - An organization could not successfully complete most business activities
- Data consists of raw facts
- Data must be organized in a meaningful way to transform it into useful information

Hierarchy of Data

- A bit (binary digit) represents a circuit that is either on or off
- A byte is made up of eight bits
  - Each byte represents a character
- Field: a name, number, or combination of characters that describes an aspect of a business object or activity
- Record: a collection of related data fields
- File: a collection of related records

Hierarchy of Data

- Database: a collection of integrated and related files
- Hierarchy of data: bits, characters, fields, records, files, and databases



Hierarchy of data

Example

Database

Personnel file
Department file
Payroll file
(Project database)

Files

098 - 40 - 1370 Fiske, Steven 01-05-2001
549 - 77 - 1001 Buckley, Bill 02-17-1995
005 - 10 - 6321 Johns, Francine 10-07-2013
(Personnel file)

Records

098 - 40 - 1370 Fiske, Steven 01-05-2001
(Record containing employee #, last and first name, hire date)

Fields

Fiske
(Last name field)

Each character is represented as 8 bits

1000110
(Letter F in ASCII)

FIGURE 5.1

Hierarchy of data

Together, bits, characters, fields, records, files, and databases form the hierarchy of data.

Data Entities, Attributes, and Keys

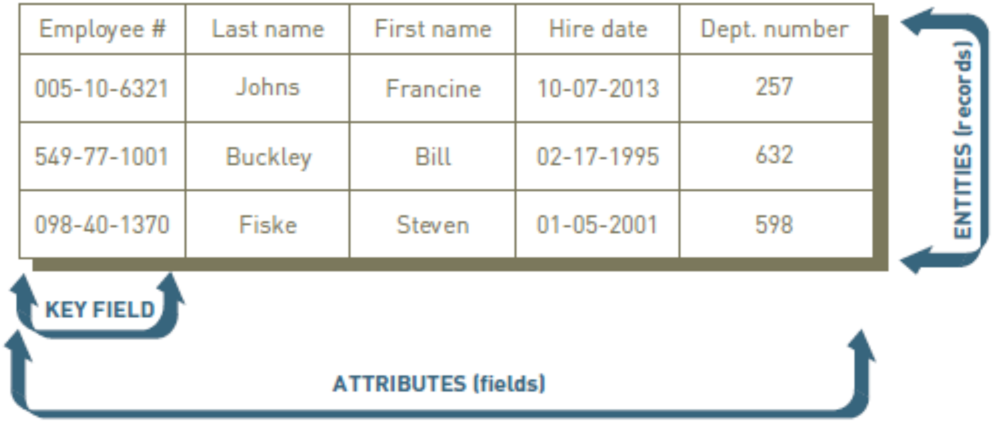
- Entity: a person, place, or thing for which data is collected, stored, and maintained
- Attribute: a characteristic of an entity
- Data item: the specific value of an attribute
- Primary key: a field or set of fields that uniquely identifies the record

Data Entities, Attributes, and Keys

FIGURE 5.2

Keys and attributes

The key field is the employee number. The attributes include last name, first name, hire date, and department number.



Employee #	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-2013	257
549-77-1001	Buckley	Bill	02-17-1995	632
098-40-1370	Fiske	Steven	01-05-2001	598

KEY FIELD
ATTRIBUTES (fields)
ENTITIES (records)

The Database Approach

- Traditional approach to data management
    - Each distinct operational system used data files dedicated to that system
- Database approach to data management
    - Information systems share a pool of related data
    - Offers the ability to share data and information resources
    - A database management system (DBMS) is required

The Database Approach

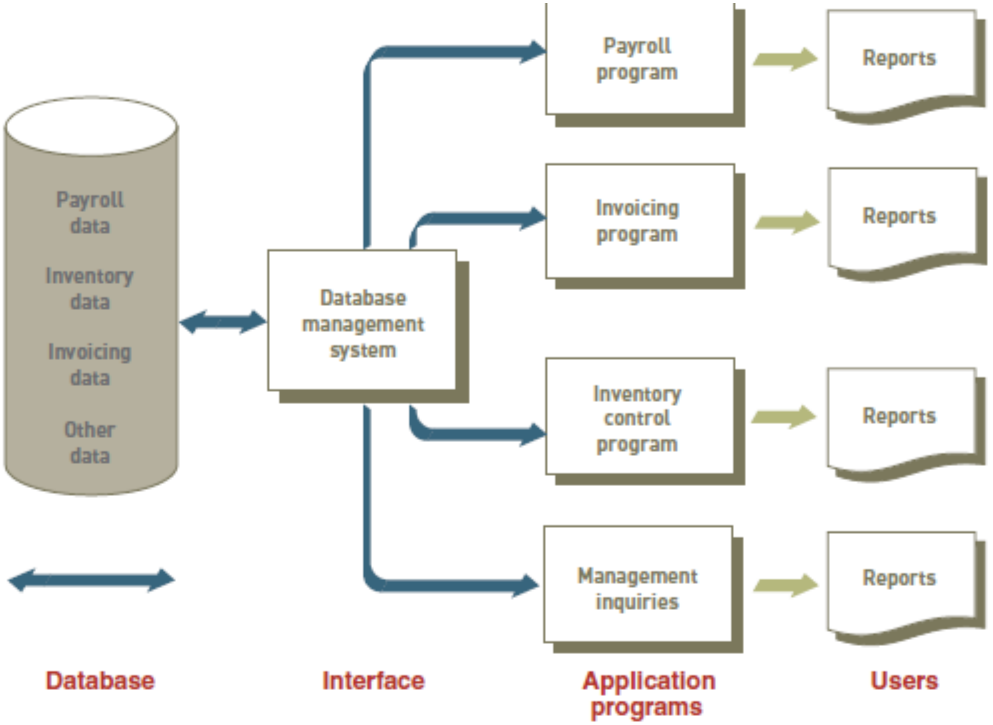


FIGURE 5.4

Database approach to data management

In a database approach to data management, multiple information systems share a pool of related data.

Data Modeling and Database Characteristics

- Considerations when building a database
  - Content: what data should be collected? cost?
  - Access: what data should be provided to which users and when?
  - Logical structure: how should data be arranged so that it makes sense?
  - Physical organization: where should data be physically located?
  - Archiving: how long to store?
  - Security: how can data be protected?

Data Modeling

- Data model: a diagram of data entities and their relationships
- Enterprise data modeling: data modeling done at the level of the entire enterprise
- Entity-relationship (ER) diagrams: data models that use basic graphical symbols to show the organization of and relationships between data

Data Modeling

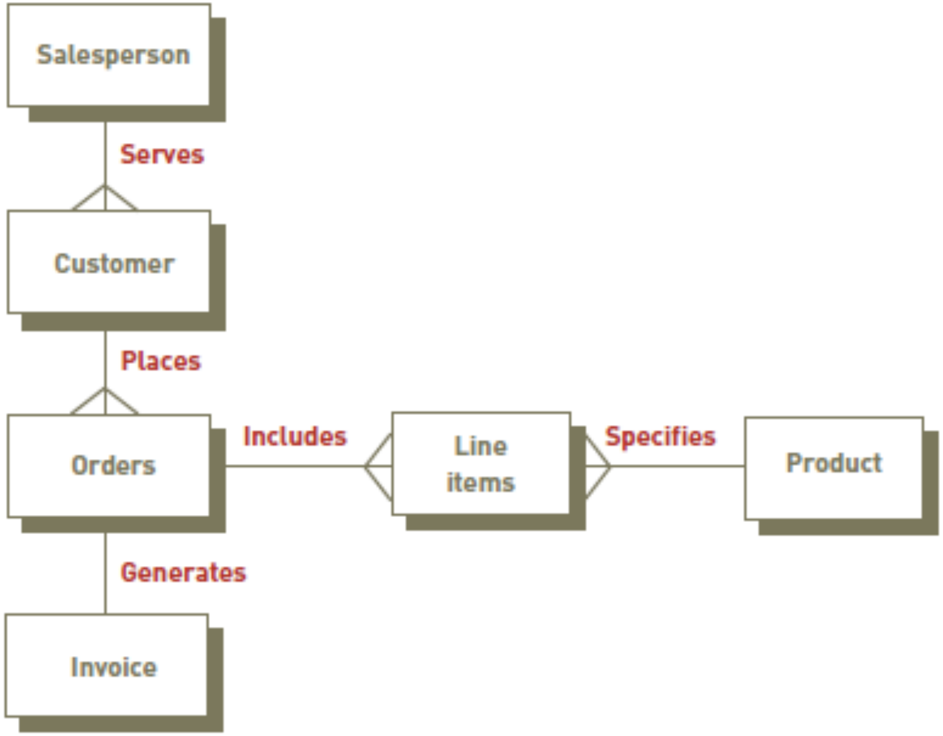


FIGURE 5.6

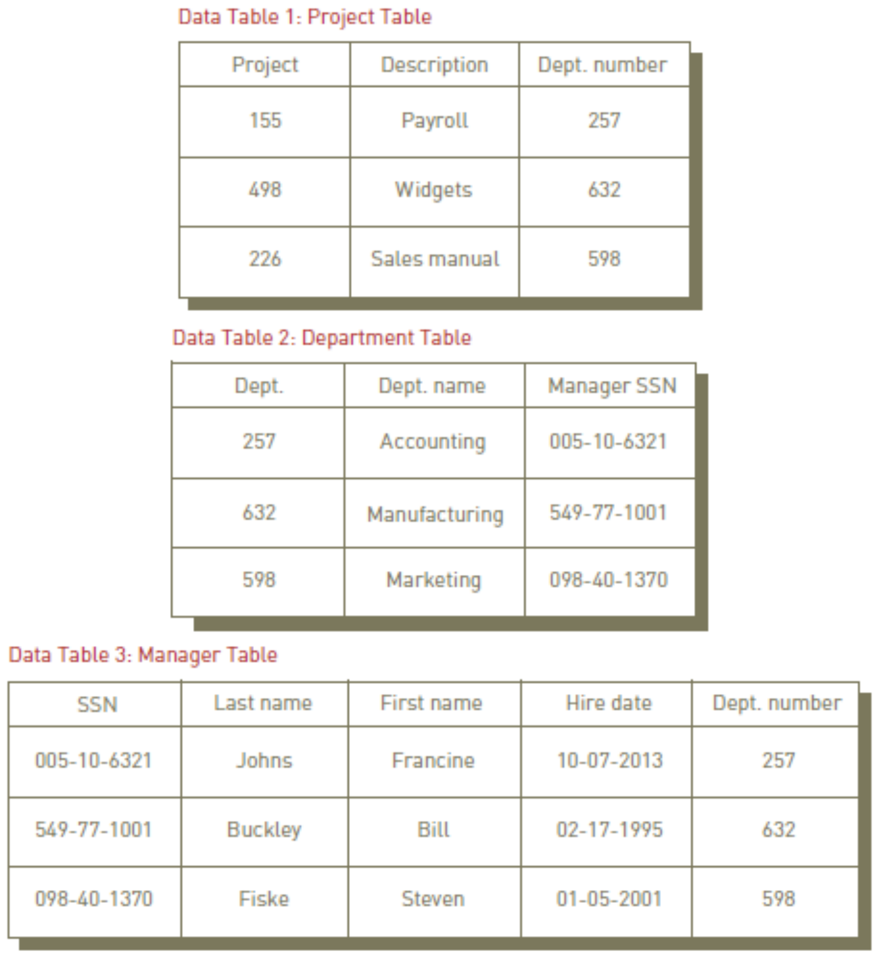
Entity-relationship (ER) diagram for a customer order database

Development of ER diagrams helps ensure that the logical structure of application programs is consistent with the data relationships in the database.

Relational Database Model

- Relational model: a simple but highly useful way to organize data into collections of two-dimensional tables called relations
  - Each row in the table represents an entity
  - Each column represents an attribute of that entity
- Domain: range of allowable values for a data attribute

Relational Database Model



Data Table 1: Project Table

Project	Description	Dept. number
155	Payroll	257
498	Widgets	632
226	Sales manual	598

Data Table 2: Department Table

Dept.	Dept. name	Manager SSN
257	Accounting	005-10-6321
632	Manufacturing	549-77-1001
598	Marketing	098-40-1370

Data Table 3: Manager Table

SSN	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-2013	257
549-77-1001	Buckley	Bill	02-17-1995	632
098-40-1370	Fiske	Steven	01-05-2001	598

FIGURE 5.7

Relational database model

In the relational model, data is placed in two-dimensional tables, or relations. As long as they share at least one common attribute, these relations can be linked to provide output useful information. In this example, all three tables include the Dept. number attribute.

Manipulating Data

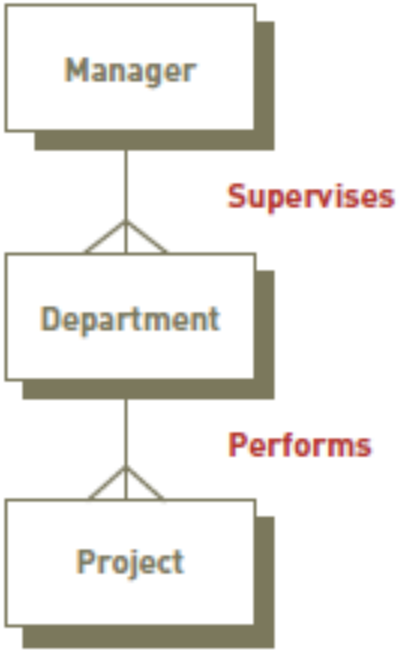
- Selecting: eliminating rows according to certain criteria
- Projecting: eliminating columns in a table
- Joining: combining two or more tables
- Linking: combining two or more tables through common data attributes to form a new table with only the unique data attributes

Manipulating Data

FIGURE 5.8

Simplified ER diagram

This diagram shows the relationship among the Manager, Department, and Project tables.

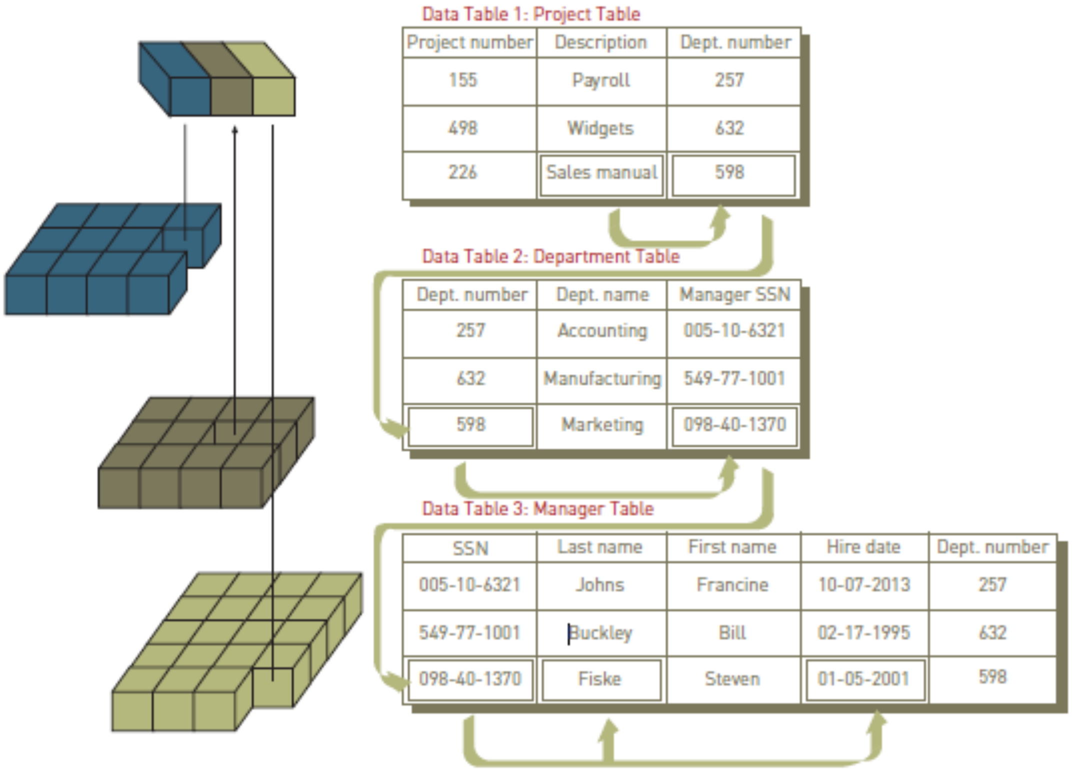


Manipulating Data

FIGURE 5.9

Linking data tables to answer an inquiry

To find the name and hire date of the manager working on the sales manual project, the president needs three tables: Project, Department, and Manager. The project description (Sales manual) leads to the department number (598) in the Project table, which leads to the manager's Social Security number (098-40-1370) in the Department table, which leads to the manager's last name (Fiske) and hire date (01-05-2001) in the Manager table.



Data Table 1: Project Table

Project number	Description	Dept. number
155	Payroll	257
498	Widgets	632
226	Sales manual	598

Data Table 2: Department Table

Dept. number	Dept. name	Manager SSN
257	Accounting	005-10-6321
632	Manufacturing	549-77-1001
598	Marketing	098-40-1370

Data Table 3: Manager Table

SSN	Last name	First name	Hire date	Dept. number
005-10-6321	Johns	Francine	10-07-2013	257
549-77-1001	Buckley	Bill	02-17-1995	632
098-40-1370	Fiske	Steven	01-05-2001	598

Data Cleansing

- Also called data cleaning or data scrubbing
- The process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, irrelevant records that reside in a database
- The cost of performing data cleansing can be quite high
- Different from data validation
  - Which involves the identification of "bad data" and its rejection at the time of data entry

Data Cleansing

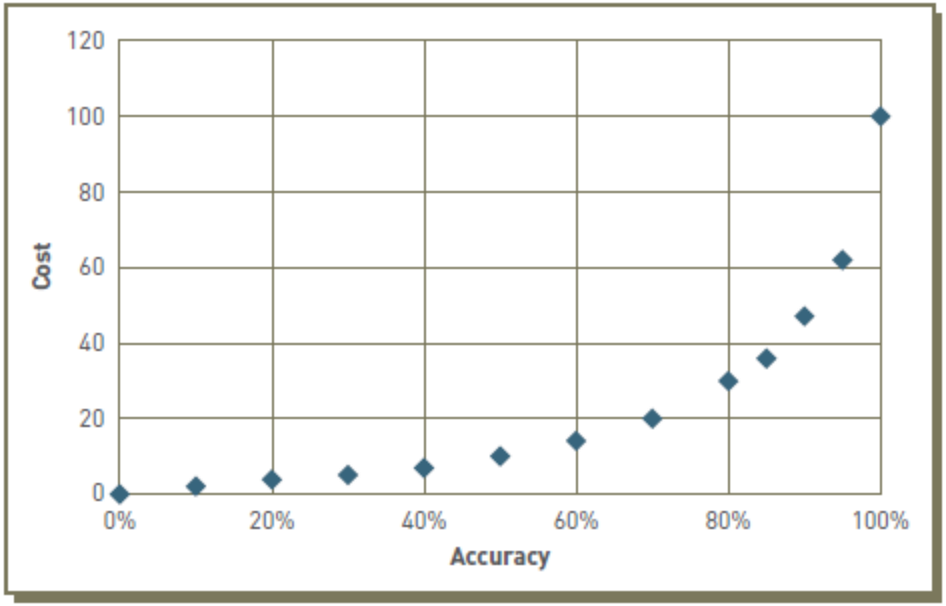


FIGURE 5.11

Tradeoff of cost versus accuracy

The cost of performing data cleansing to achieve 100 percent database accuracy can be prohibitively expensive.

Relational Database Management Systems (DBMSs)

- Creating and implementing the right database system ensures that the database will support both business activities and goals
- Capabilities and types of database systems vary considerably

SQL Databases

- SQL: a special-purpose programming language for accessing and manipulating data stored in a relational database
- SQL databases conform to ACID properties:
  - Atomicity, consistency, isolation, and durability
- 1986: SQL was adopted by ANSI as the standard query language for relational databases

SQL Databases

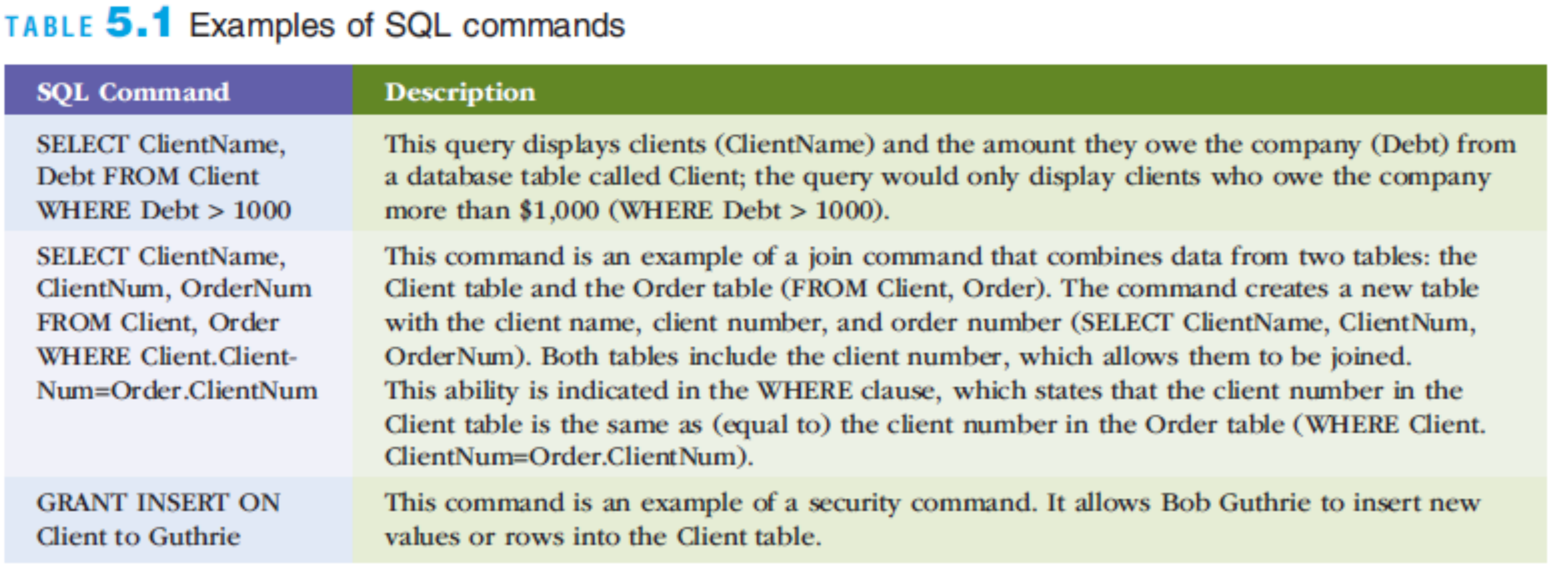
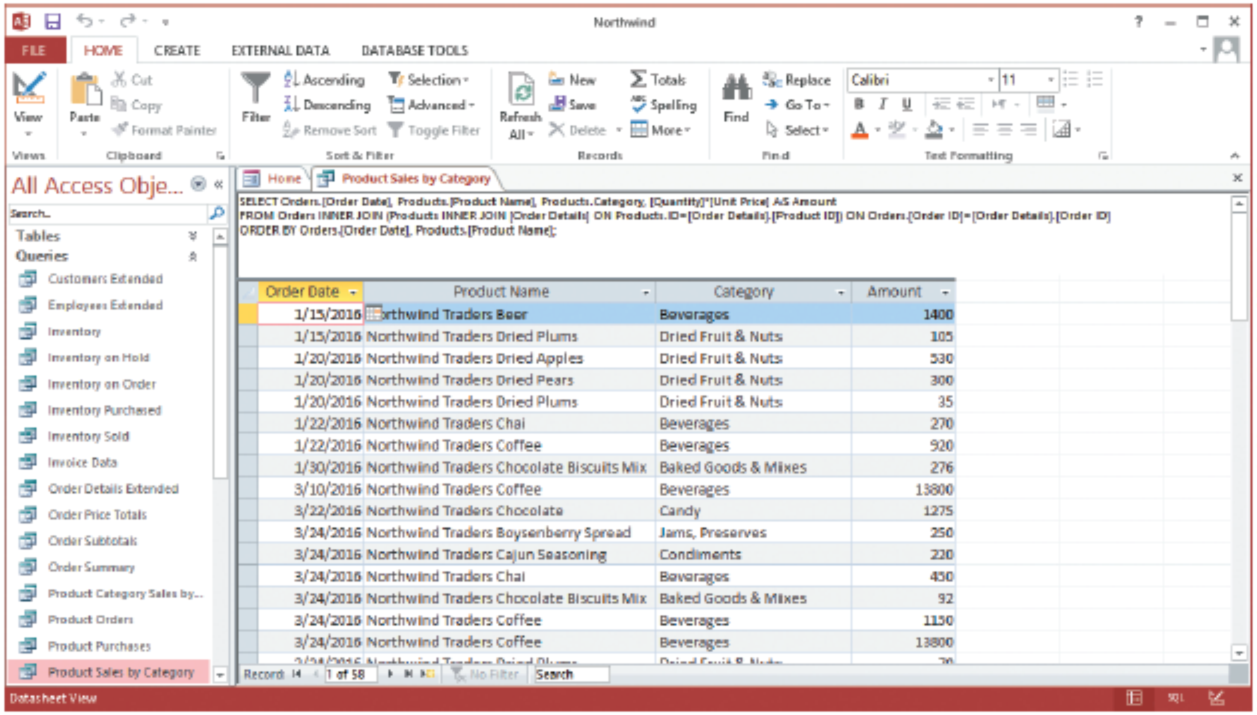


TABLE 5.1 Examples of SQL commands

SQL Command	Description
SELECT ClientName, Debt FROM Client WHERE Debt > 1000	This query displays clients (ClientName) and the amount they owe the company (Debt) from a database table called Client; the query would only display clients who owe the company more than $1,000 (WHERE Debt > 1000).
SELECT ClientName, ClientNum, OrderNum FROM Client, Order WHERE Client.ClientNum=Order.ClientNum	This command is an example of a join command that combines data from two tables: the Client table and the Order table (FROM Client, Order). The command creates a new table with the client name, client number, and order number (SELECT ClientName, ClientNum, OrderNum). Both tables include the client number, which allows them to be joined. This ability is indicated in the WHERE clause, which states that the client number in the Client table is the same as (equal to) the client number in the Order table (WHERE Client.ClientNum=Order.ClientNum).
GRANT INSERT ON Client to Guthrie	This command is an example of a security command. It allows Bob Guthrie to insert new values or rows into the Client table.

SQL Databases



```
SELECT Orders.[Order Date], Products.[Product Name], Products.Category, [Quantity]*[Unit Price] AS Amount
FROM Orders INNER JOIN (Products INNER JOIN [Order Details] ON Products.ID=[Order Details].[Product ID]) ON Orders.[Order ID]=[Order Details].[Order ID]
ORDER BY Orders.[Order Date], Products.[Product Name];
```

Order Date	Product Name	Category	Amount
1/15/2016	rthwind Traders Beer	Beverages	1400
1/15/2016	Northwind Traders Dried Plums	Dried Fruit & Nuts	105
1/20/2016	Northwind Traders Dried Apples	Dried Fruit & Nuts	530
1/20/2016	Northwind Traders Dried Pears	Dried Fruit & Nuts	300
1/20/2016	Northwind Traders Dried Plums	Dried Fruit & Nuts	35
1/22/2016	Northwind Traders Chai	Beverages	270
1/22/2016	Northwind Traders Coffee	Beverages	920
1/30/2016	Northwind Traders Chocolate Biscuits Mix	Baked Goods & Mixes	276
3/10/2016	Northwind Traders Coffee	Beverages	13800
3/22/2016	Northwind Traders Chocolate	Candy	1275
3/24/2016	Northwind Traders Boysenberry Spread	Jams, Preserves	250
3/24/2016	Northwind Traders Cajun Seasoning	Condiments	220
3/24/2016	Northwind Traders Chai	Beverages	450
3/24/2016	Northwind Traders Chocolate Biscuits Mix	Baked Goods & Mixes	92
3/24/2016	Northwind Traders Coffee	Beverages	1150
3/24/2016	Northwind Traders Coffee	Beverages	13800

Microsoft product screenshots used with permission from Microsoft Corporation

FIGURE 5.12

Structured Query Language (SQL)

SQL has become an integral part of most relational databases, as shown by this example from Microsoft Access 2013.

Database Activities

- Providing a user view of the database
- Adding and modifying data
- Storing and retrieving data
- Manipulating the data and generating reports

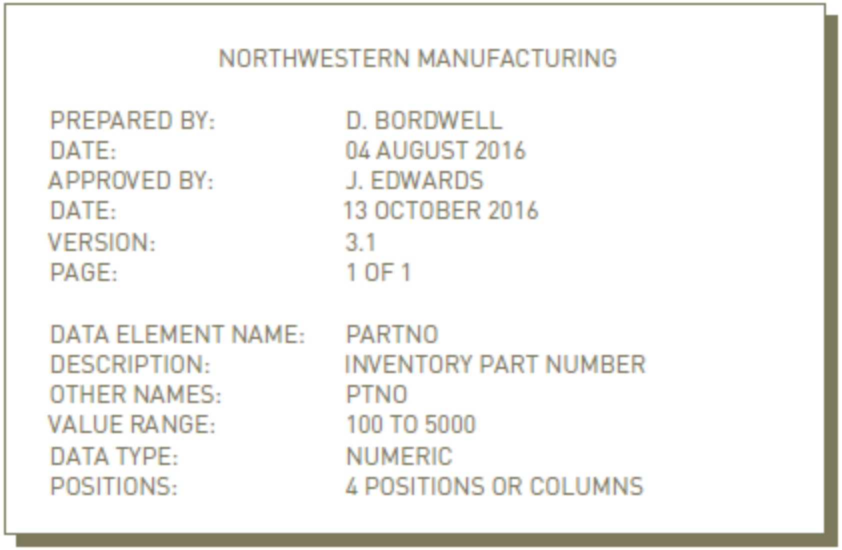
Providing a User View

- Schema: a description of the entire database
- A schema can be part of the database or a separate schema file
- The DBMS can reference a schema to find where to access the requested data in relation to another piece of data

Creating and Modifying the Database

- Data definition language (DDL)
  - A collection of instructions and commands used to define and describe data and relationships in a specific database
  - Allows the database's creator to describe data and relationships that are to be contained in the schema
- Data dictionary: a detailed description of all the data used in the database
  - Can also include a description of data flows, information about the way records are organized, and the data-processing requirements

Creating and Modifying the Database



NORTHWESTERN MANUFACTURING

PREPARED BY:	D. BORDWELL
DATE:	04 AUGUST 2016
APPROVED BY:	J. EDWARDS
DATE:	13 OCTOBER 2016
VERSION:	3.1
PAGE:	1 OF 1
DATA ELEMENT NAME:	PARTNO
DESCRIPTION:	INVENTORY PART NUMBER
OTHER NAMES:	PTNO
VALUE RANGE:	100 TO 5000
DATA TYPE:	NUMERIC
POSITIONS:	4 POSITIONS OR COLUMNS

FIGURE 5.14

Data dictionary entry

A data dictionary provides a detailed description of all data used in the database.

Storing and Retrieving Data

- When an application program needs data, it requests the data through the DBMS
- Concurrency control deals with the situation in which two or more users or applications need to access the same record at the same time

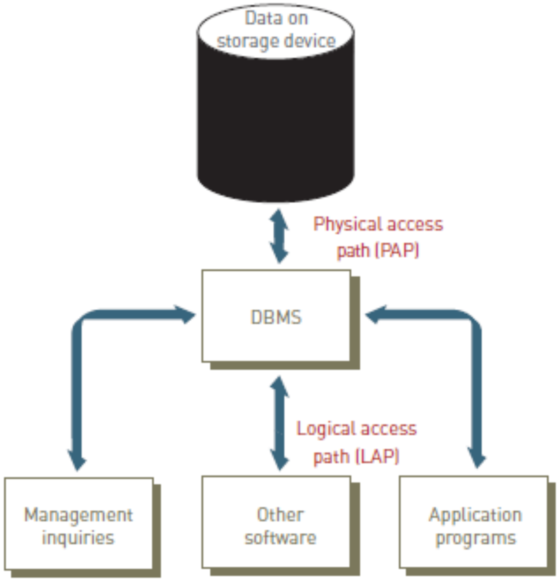


FIGURE 5.15

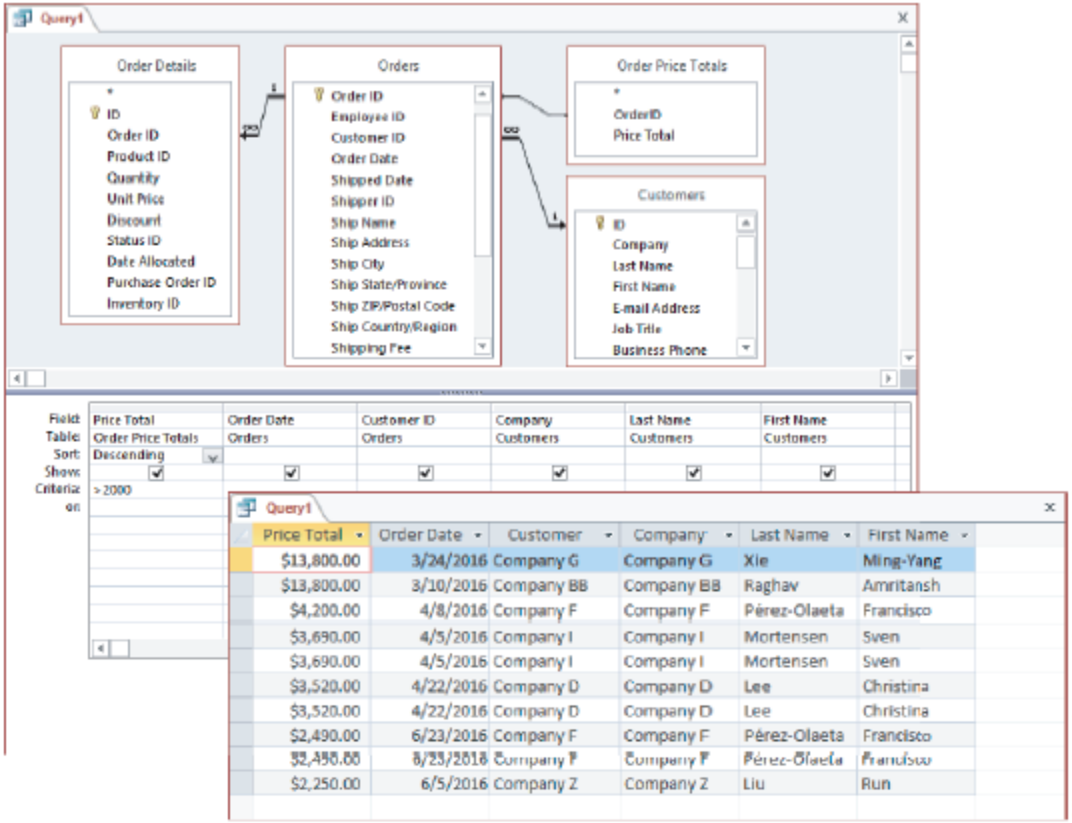
Logical and physical access paths

When an application requests data from the DBMS, it follows a logical access path to the data. When the DBMS retrieves the data, it follows a path to the physical access path to the data.

Manipulating Data and Generating Reports

- Query by Example (QBE) is a visual approach to developing database queries or requests
- Data manipulation language (DML): a specific language, provided with a DBMS
  - Allows users to access and modify the data, to make queries, and to generate reports
- A DBMS can produce a wide variety of documents, reports, and other output that can help organizations achieve their goals

Manipulating Data and Generating Reports



Field:	Price Total	Order Date	Customer ID	Company	Last Name	First Name
Table:	Order Price Totals	Orders	Orders	Customers	Customers	Customers
Sort:	Descending					
Show:	☑	☑	☑	☑	☑	☑
Criteria:	> 2000					
or:						

Query1

Price Total	Order Date	Customer	Company	Last Name	First Name
$13,800.00	3/24/2016	Company G	Company G	Xie	Ming-Yang
$13,800.00	3/10/2016	Company BB	Company BB	Raghav	Amritansh
$4,200.00	4/8/2016	Company F	Company F	Pérez-Olaeta	Francisco
$3,690.00	4/5/2016	Company I	Company I	Mortensen	Sven
$3,690.00	4/5/2016	Company I	Company I	Mortensen	Sven
$3,520.00	4/22/2016	Company D	Company D	Lee	Christina
$3,520.00	4/22/2016	Company D	Company D	Lee	Christina
$2,490.00	6/23/2016	Company F	Company F	Pérez-Olaeta	Francisco
$2,490.00	6/23/2016	Company F	Company F	Pérez-Olaeta	Francisco
$2,250.00	6/5/2016	Company Z	Company Z	Liu	Run

Microsoft product screenshots used with permission from Microsoft Corporation

FIGURE 5.16

Query by Example

Some databases use Query by Example (QBE) to generate reports and information.

Database Administration

- Database administrators (DBAs): skilled and trained IS professionals
  - Works with users to define their data needs
  - Applies database programming languages to craft a set of databases to meet those needs
  - Tests and evaluates databases
  - Implements changes to improve their databases' performance
  - Assures that data is secure from unauthorized access

Database Administration

- Data administrator: a nontechnical position responsible for defining and implementing consistent principles for a variety of data issues
  - Including setting data standards and data definitions that apply across all the databases in an organization
- The data administrator can be a high-level position reporting to top-level managers

Popular Database Management Systems

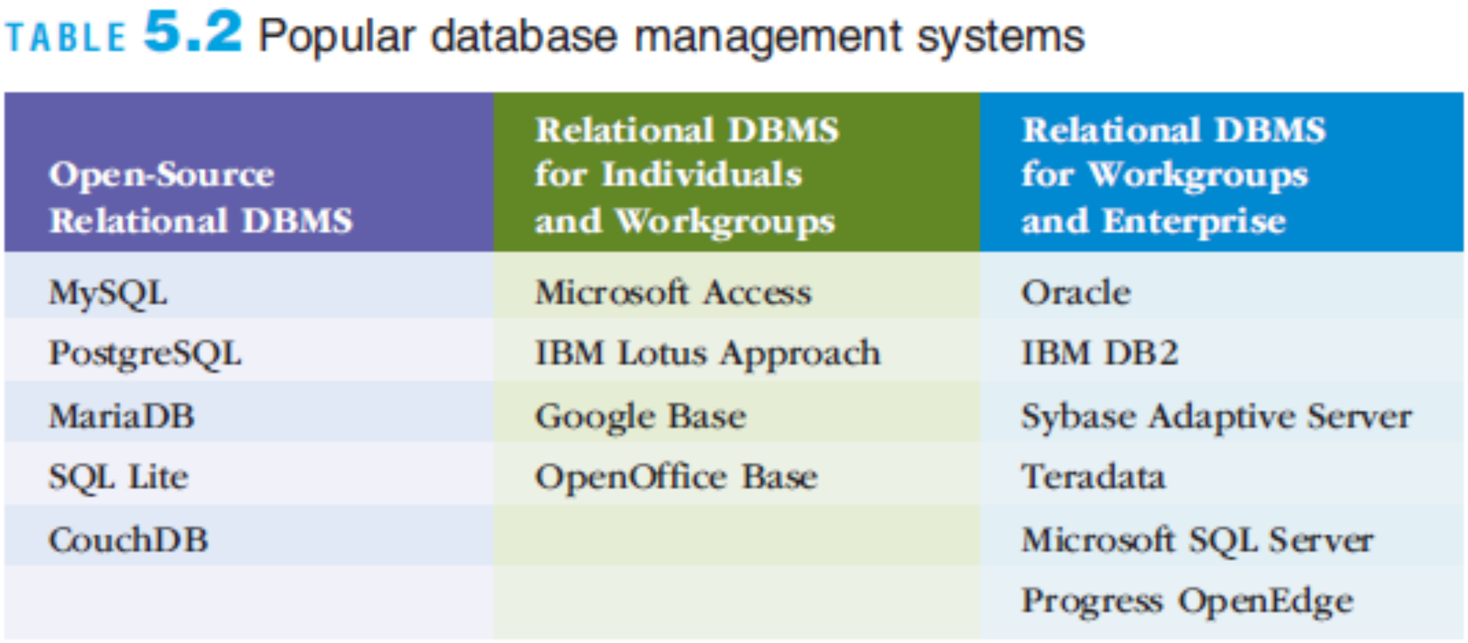


TABLE 5.2 Popular database management systems

Open-Source Relational DBMS	Relational DBMS for Individuals and Workgroups	Relational DBMS for Workgroups and Enterprise
MySQL	Microsoft Access	Oracle
PostgreSQL	IBM Lotus Approach	IBM DB2
MariaDB	Google Base	Sybase Adaptive Server
SQL Lite	OpenOffice Base	Teradata
CouchDB		Microsoft SQL Server
		Progress OpenEdge

Popular Database Management Systems

- Database as a Service (DaaS)
  - The database is stored on a service provider's servers
  - The database is accessed by the client over a network, typically the Internet
  - Database administration is handled by the service provider
- Example of DaaS: Amazon Relational Database Service (Amazon RDS)

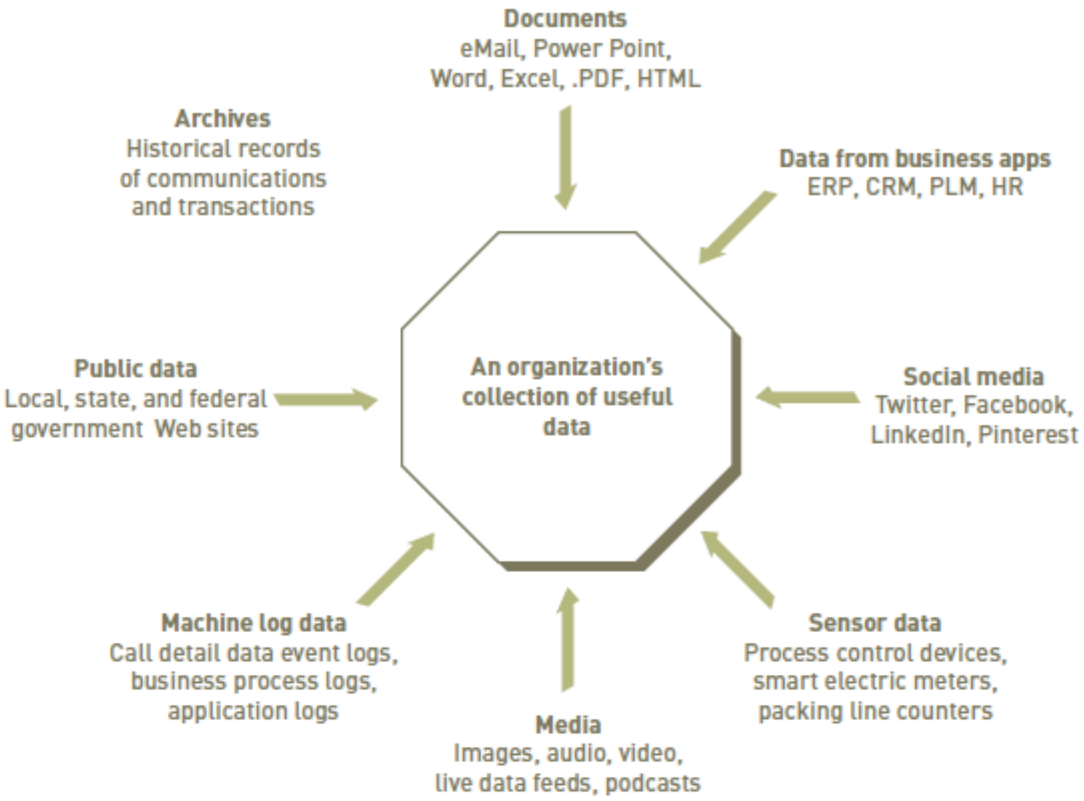
Using Databases with Other Software

- DBMSs can act as front-end or back-end applications
  - Front-end applications interact directly with people
  - Back-end applications interact with other programs or applications
- Example:
  - The Library of Congress (LOC) provides a back-end application that allows Web access to its databases, which include references to books and digital media in the LOC collection

Big Data

- Extremely large and complex data collections
    - Traditional data management software, hardware, and analysis processes are incapable of dealing with them
- Three characteristics of big data (3Vs)
    - Volume
    - Velocity
    - Variety

Sources of Big Data



An organization's collection of useful data

Documents
eMail, Power Point,
Word, Excel, .PDF, HTML

Data from business apps
ERP, CRM, PLM, HR

Social media
Twitter, Facebook,
LinkedIn, Pinterest

Sensor data
Process control devices,
smart electric meters,
packing line counters

Media
Images, audio, video,
live data feeds, podcasts

Machine log data
Call detail data event logs,
business process logs,
application logs

Public data
Local, state, and federal
government Web sites

Archives
Historical records
of communications
and transactions

FIGURE 5.20

Sources of an organization's useful data

An organization has many sources of useful data.

Big Data Uses

- Examples:
  - Retail organizations monitor social networks to engage brand advocates, identify brand adversaries
  - Advertising and marketing agencies track comments on social media
  - Hospitals analyze medical data and patient records
  - Consumer product companies monitor social networks to gain insight into consumer behavior
  - Financial service organizations use data to identify customers who are likely to be attracted to increasingly targeted and sophisticated offers

Challenges of Big Data

- How to choose what subset of the data to store
- Where and how to store the data
- How to find the nuggets of data that are relevant to the decision making at hand
- How to derive value from the relevant data
- How to identify which data needs to be protected from unauthorized access

Data Management

- Data management
  - An integrated set of functions that defines the processes by which data is obtained, certified fit for use, stored, secured, and processed in such a way as to ensure that the accessibility, reliability, and timeliness of the data meet the needs of the data users within an organization
- Data governance
  - Defines the roles, responsibilities, and processes for ensuring that data can be trusted and used by an entire organization

Data Management

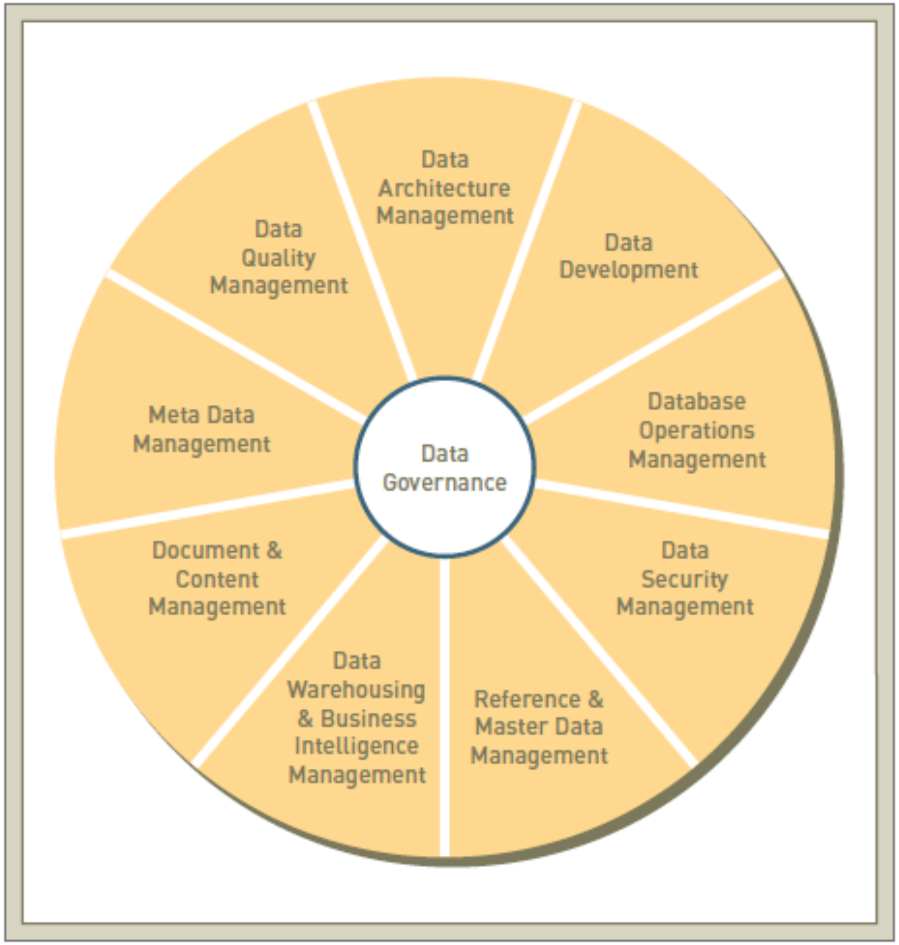

Data Architecture Management
Data Development
Database Operations Management
Data Security Management
Reference & Master Data Management
Data Warehousing & Business Intelligence Management
Document & Content Management
Meta Data Management
Data Quality Management
Data Governance

FIGURE 5.21

Data management

The Data Management Association (DAMA) International has identified 10 basic functions associated with data management.

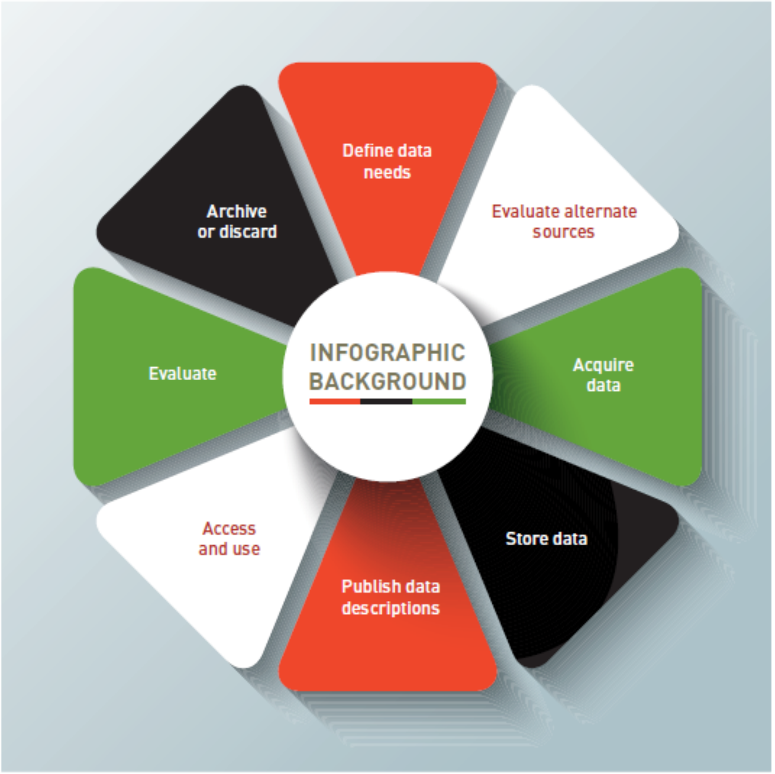
Source: "Body of Knowledge," DAMA International, *https://www.dama.org/content/body-knowledge*. Copyright DAMA International.

Data Management

- Data management is driven by a variety of factors:
  - The need to meet external regulations designed to manage risk associated with financial misstatement
  - The need to avoid the inadvertent release of sensitive data
  - The need to ensure that high data quality is available for key decisions
- Data governance requires business leadership and active participation
  - Use of a cross-functional tea is recommended
  - Team should consist of executives, project managers, line-of-business managers, and data stewards
  - A data steward is an individual responsible for management of critical data elements

Data Management

- Data lifecycle management (DLM)
  - A policy-based approach to managing the flow of an enterprise's data



Abert/Shutterstock.com

FIGURE 5.22

The big data life cycle

A policy-based approach to managing the flow of an enterprise's data, from its initial acquisition or creation and storage to the time when it becomes outdated and is deleted.

Data Warehouses, Data Marts, and Data Lakes

- Data warehouse: a large database that collects business information from many sources in the enterprise in support of management decision making
- ETL process
  - Extract
  - Transform
  - Load

Data Warehouses, Data Marts, and Data Lakes

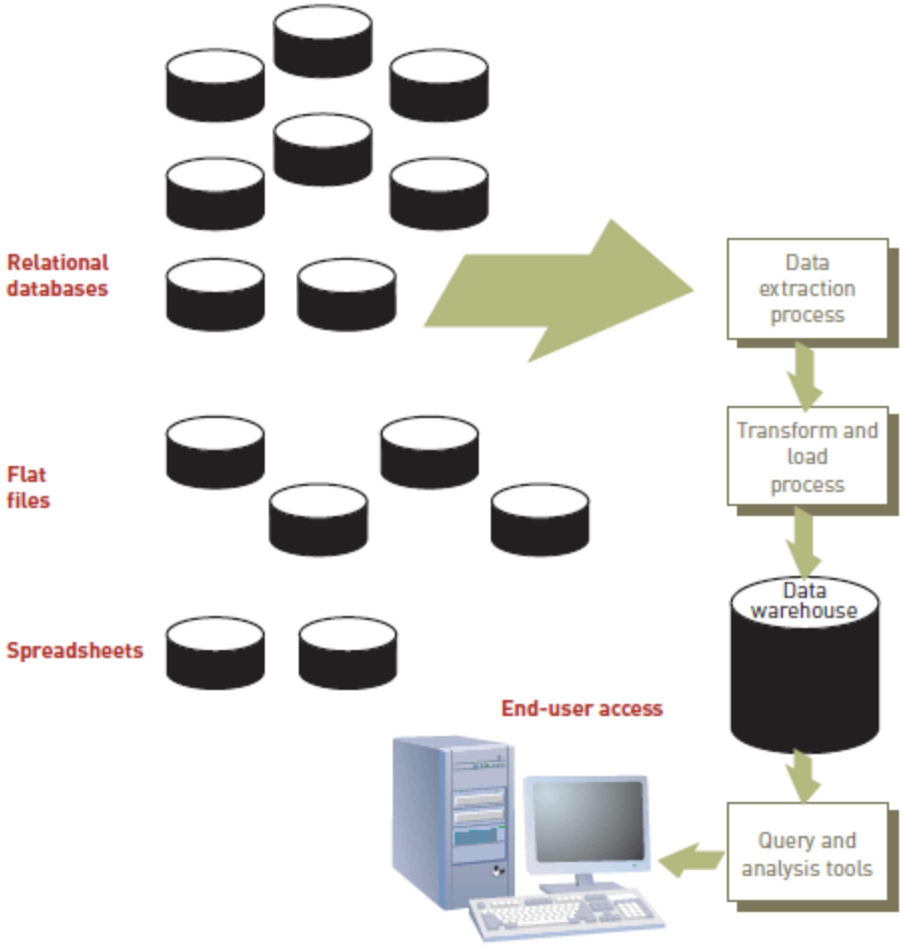


FIGURE 5.23

Elements of a data warehouse

A data warehouse can help managers and executives relate information in innovative ways to make better decisions.

Data Warehouses, Data Marts, and Data Lakes

- Data mart: a subset of a data warehouse that is used by small- and medium-sized businesses and departments within large companies to support decision making
- A specific area in the data mart might contain greater detailed data than the data warehouse
- Data lake: takes a "store everything" approach to big data, saving all the data in its raw and unaltered form
  - Also called an enterprise data hub
  - Raw data is available when users decide just how they want to use the data
  - Only when the data is accessed for a specific analysis is it extracted from the data lake

NoSQL Databases

- NoSQL (Not only SQL) database
  - Provides a means to store and retrieve data that is modeled using some means other than the simple two-dimensional tabular relations used in relational databases
- Advantages:
  - Ability to spread data over multiple servers so that each server contains only a subset of the total data
  - Do not require a predefined schema
  - Data structures are more flexible and can provide improved access speed and redundancy

NoSQL Databases

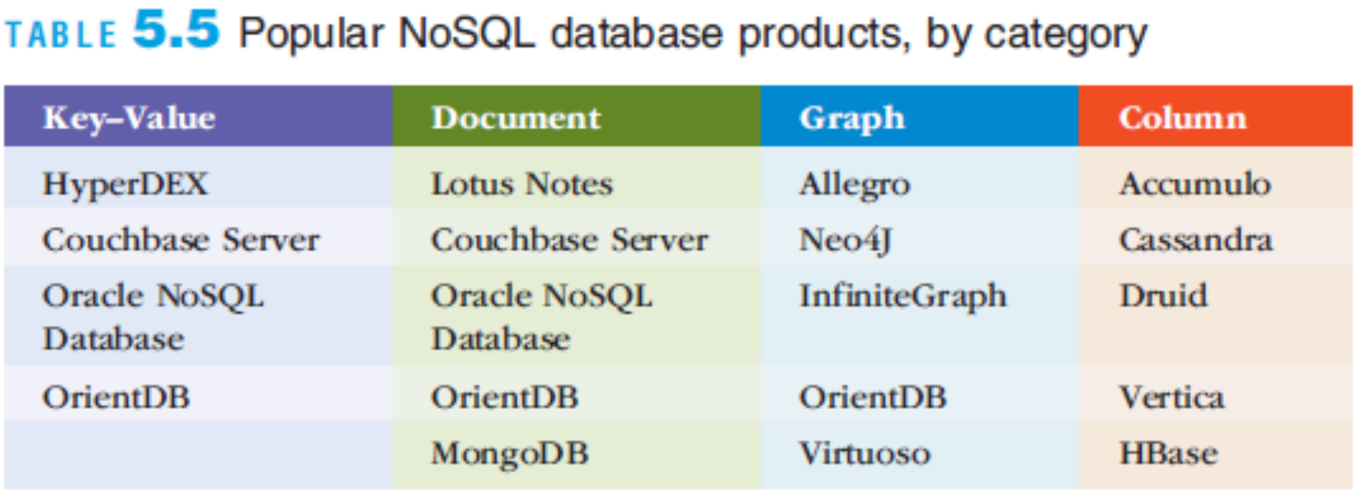


TABLE 5.5 Popular NoSQL database products, by category

Key–Value	Document	Graph	Column
HyperDEX	Lotus Notes	Allegro	Accumulo
Couchbase Server	Couchbase Server	Neo4J	Cassandra
Oracle NoSQL Database	Oracle NoSQL Database	InfiniteGraph	Druid
OrientDB	OrientDB	OrientDB	Vertica
	MongoDB	Virtuoso	HBase

Hadoop

- Hadoop
  - An open-source software framework that includes several software modules that provide a means for storing and processing extremely large data sets
- Has two primary components:
  - A data processing component (MapReduce)
  - A distributed file system (Hadoop Distributed File System, HDFS)

Hadoop

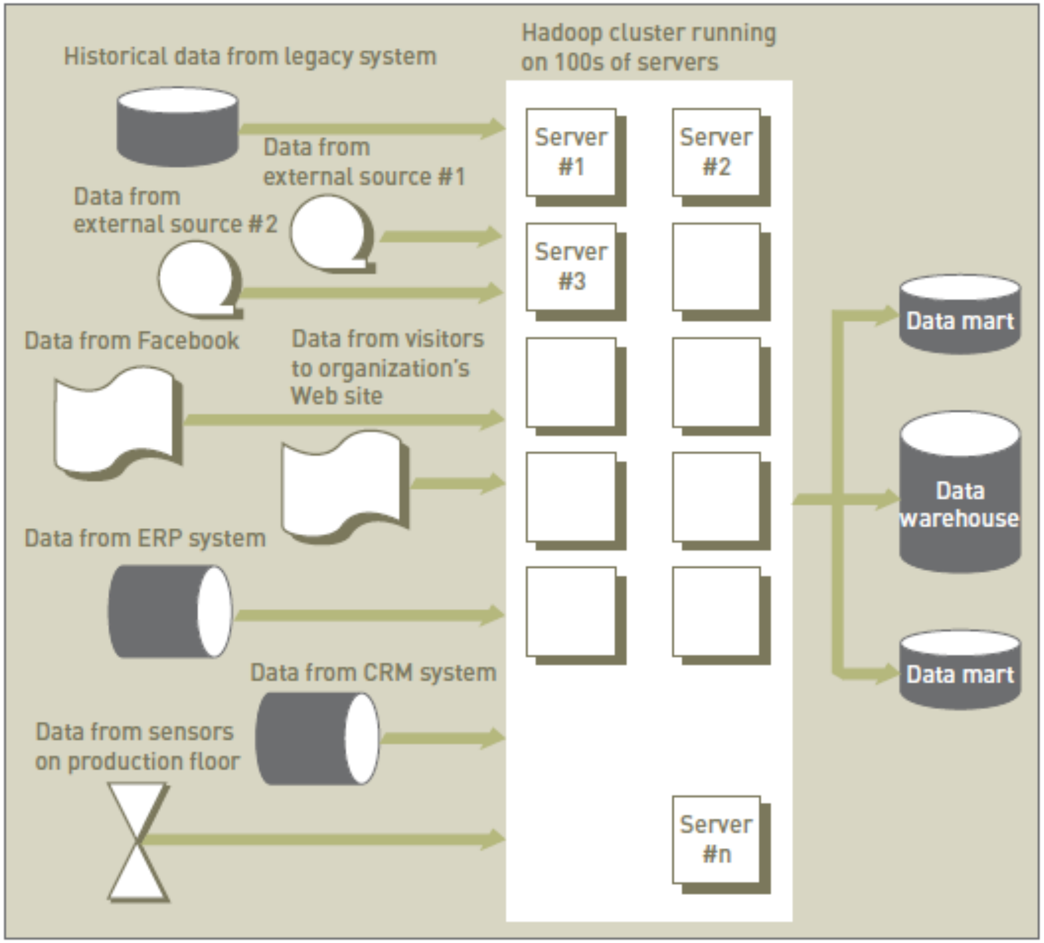

Historical data from legacy system
Hadoop cluster running on 100s of servers
Data from external source #1
Data from external source #2
Server #1
Server #2
Server #3
Data from Facebook
Data from visitors to organization's Web site
Data mart
Data warehouse
Data from ERP system
Data from CRM system
Data mart
Data from sensors on production floor
Server #n

FIGURE 5.24

Hadoop environment

Hadoop can be used as a staging area for data to be loaded into a data warehouse or data mart.

In-Memory Databases

- In-memory database (IMDB)
  - A database management system that stores the entire database in random access memory (RAM)
  - Provides access to data at rates **much faster** than storing data on some form of secondary storage
  - Enables the analysis of big data and other challenging data-processing applications
  - Performs best on multiple multicore CPUs

In-Memory Databases

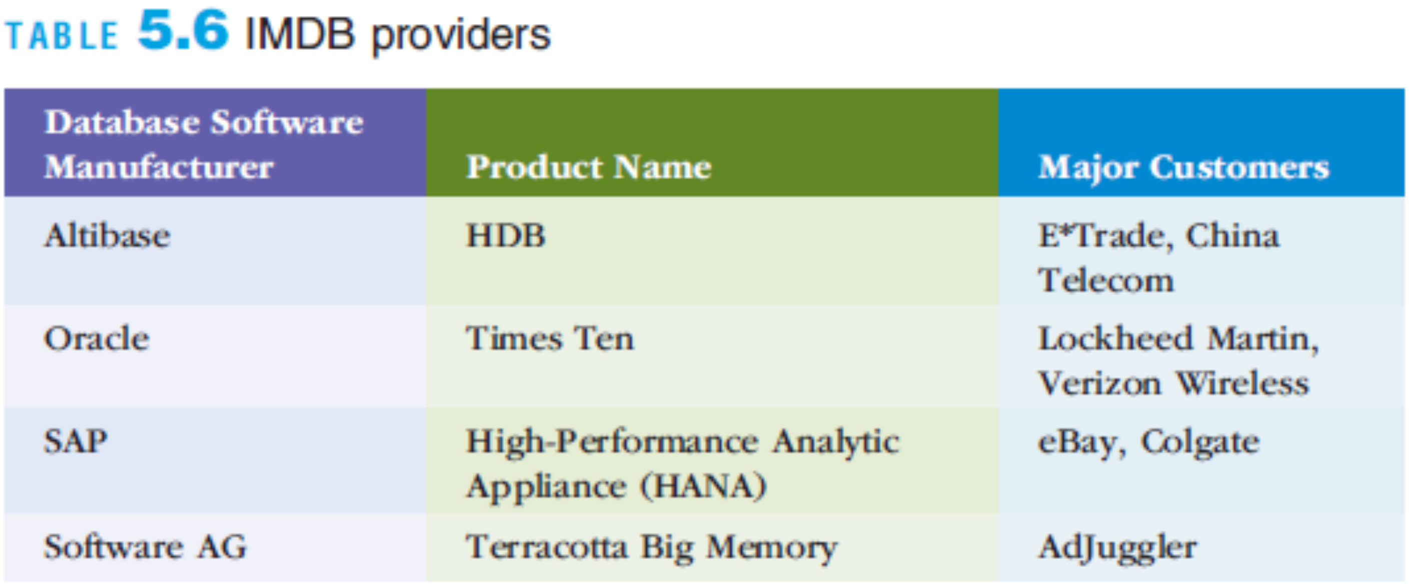


TABLE 5.6 IMDB providers

Database Software Manufacturer	Product Name	Major Customers
Altibase	HDB	E*Trade, China Telecom
Oracle	Times Ten	Lockheed Martin, Verizon Wireless
SAP	High-Performance Analytic Appliance (HANA)	eBay, Colgate
Software AG	Terracotta Big Memory	AdJuggler

Summary

- The database approach to data management has become broadly accepted
- **Data modeling** is a key aspect of organizing data and information
- A well-designed and well-managed database is an extremely valuable tool in supporting decision making
- We have entered an era where organizations are grappling with a tremendous growth in the amount of data available and struggling how to manage and make use of it
- A number of available tools and technologies allow organizations to take advantage of the opportunities offered by big data