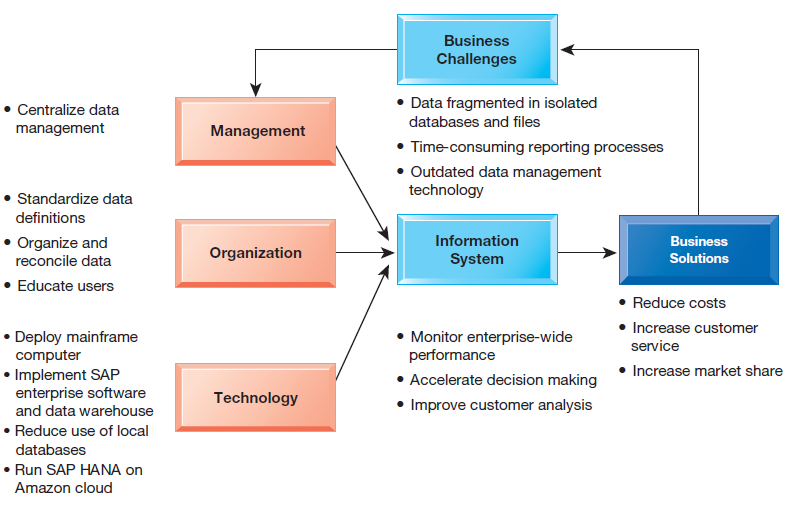
**CH06 Foundations of Business Intelligence: Databases and Information Management**

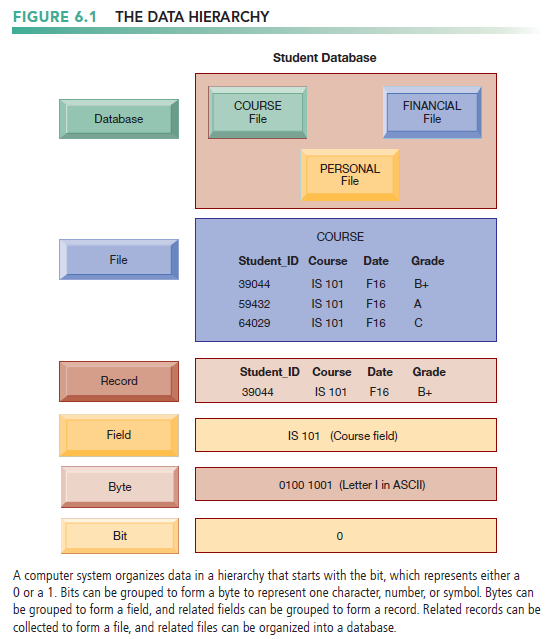
* 在傳統文件環境(traditional file environment)中管理數據資源有哪些問題？
* 資料庫管理系統（DBMS）的主要功能是什麼，為什麼關係式DBMS如此強大？
* 有哪些主要工具和技術可用於從資料庫存取資訊以改善業務績效和決策？
* 為什麼資訊政策，資料管理(data administration)和品質保證(data quality assurance)對於管理公司的數據資源至關重要？
* BAE Systems



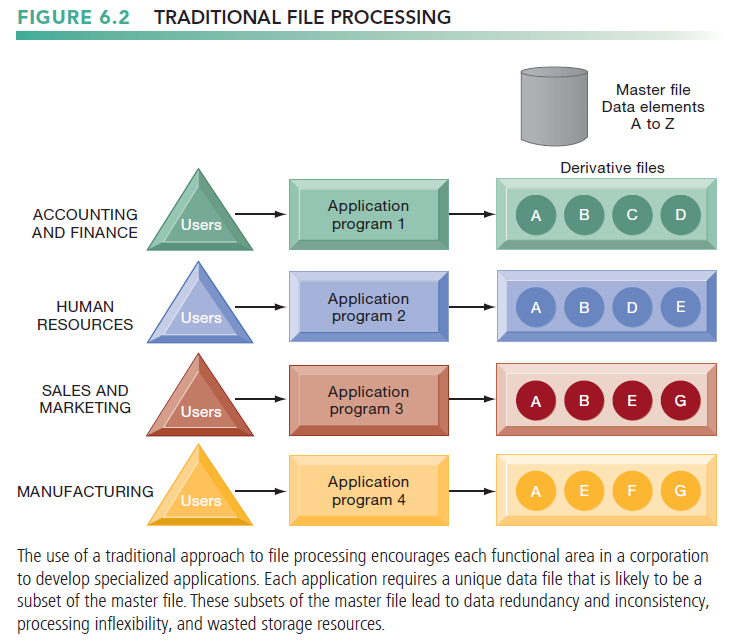
* A state-of-the-art database management system suite of software helps BAE Systems boost efficiency by making it easier to locate and assemble data for management reporting and for processing day-to-day CAD/CAM transactions for final aircraft component assembly

1. What are the problems of managing data resources in a traditional file environment?
   * File Organization Terms and Concepts

* A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases



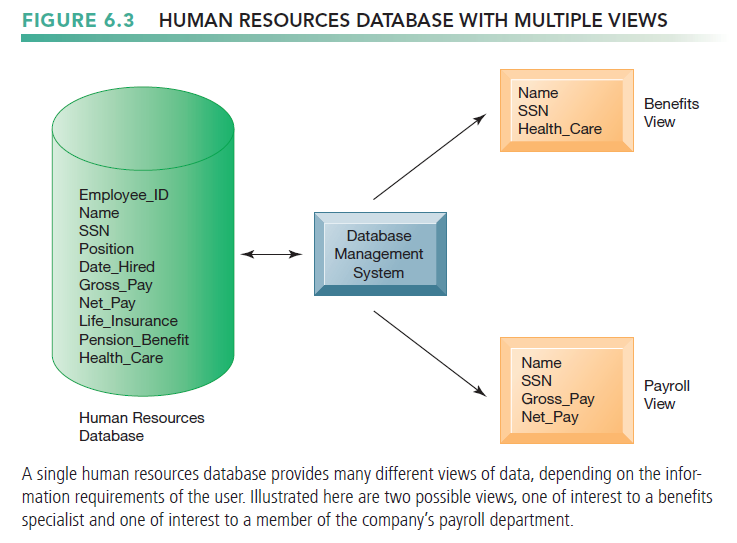
* Bit
* the smallest unit of data a computer can handle
* byte
* A group of bits
* a single character, which can be a letter, a number, or another symbol
* A grouping of characters into a word
* Field
* a group of words, or a complete number (such as a person’s name or age)
* record
* A group of related fields
* File
* a group of records of the same type is called a file
* Database
* A group of related files
* A record describes an entity.
  + An entity is a person, place, thing, or event on which we store and maintain information.
* Attribute
  + Each characteristic, or quality, describing entity
  + Example: Attributes DATE or GRADE belong to entity COURSE
  + Problems with the Traditional File Environment
* Preface
* In most organizations, systems tended to grow independently without a companywide plan.(files maintained separately by different departments)



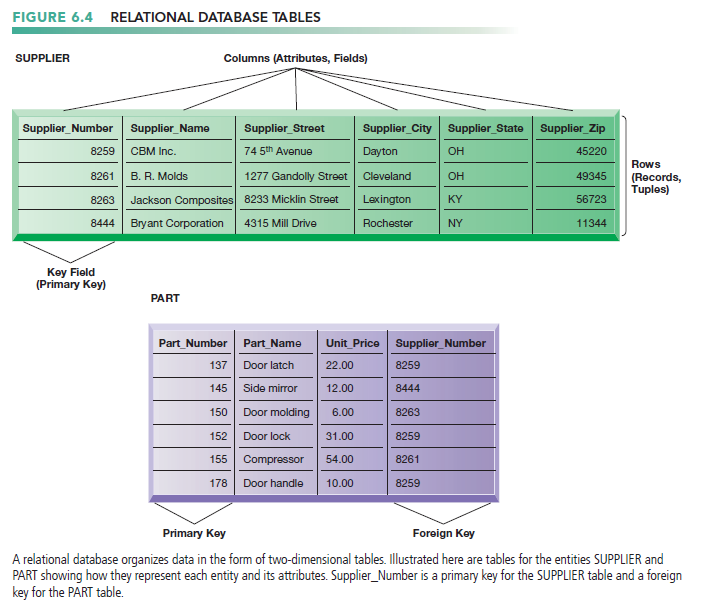
* + The resulting problems are data redundancy and inconsistency, program-data dependence, inflexibility, poor data security, and an inability to share data among applications.
* Data Redundancy and Inconsistency
* Redundancy
  + the presence of duplicate data in multiple data files so that the same data are stored in more than one place or location.
  + occurs when different groups in an organization independently collect the same piece of data and store it independently of each other
  + wastes storage resources
  + leads to data inconsistency
* Inconsistency
  + Same attribute has different values
  + difficult for companies to create customer relationship management, supply chain management, or enterprise systems that integrate data from different sources.
* Program-Data Dependence
* the coupling of data stored in files and the specific programs required to update and maintain those files
  + changes in programs require changes to the data.
* Lack of Flexibility
* deliver routine scheduled reports after extensive programming efforts, but it cannot deliver ad hoc reports or respond to unanticipated information requirements in a timely fashion
* Poor Security
* little control or management of data, access to and dissemination of information may be out of control
* Lack of Data Sharing and Availability
* pieces of information in different files and different parts of the organization cannot be related to one another, it is virtually impossible for information to be shared or accessed in a timely manner.

1. What are the major capabilities of database management systems (DBMS), and why is a relational DBMS so powerful?
   * Preface

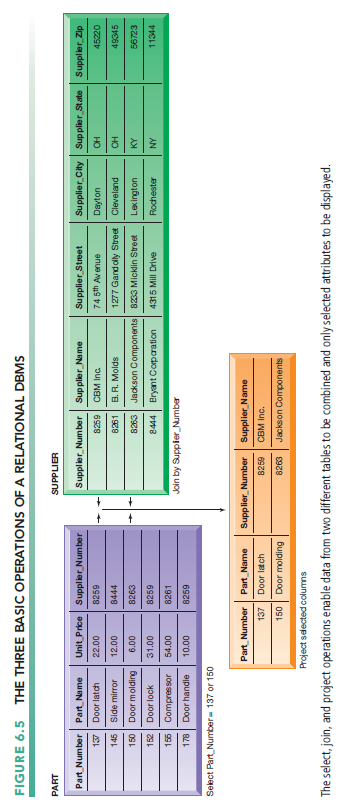
* A more rigorous definition of a database is a collection of data organized to serve many applications efficiently by centralizing the data and controlling redundant data.
  + Database Management Systems
* Preface
* database management system (DBMS)
  + software that permits an organization to centralize data, manage them efficiently, and provide access to the stored data by application programs
  + Interfaces between applications and physical data files
  + Separates logical and physical views of data



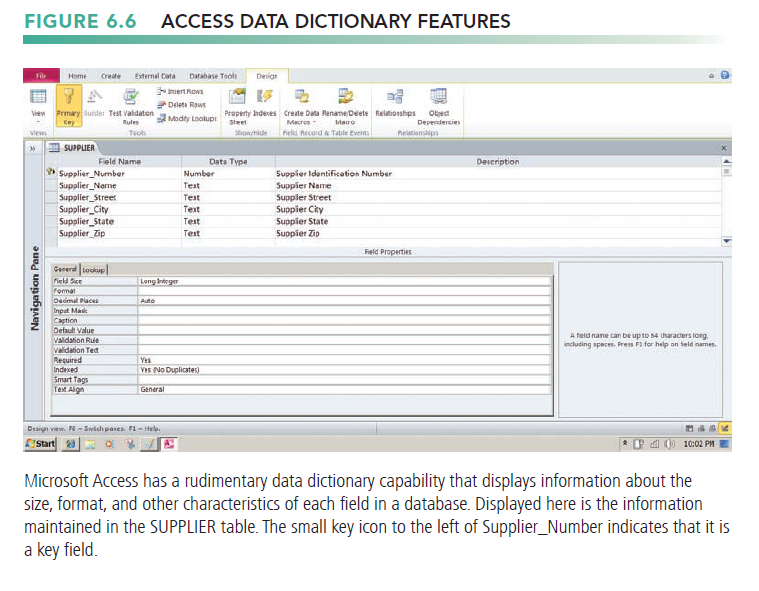
* How a DBMS Solves the Problems of the Traditional File Environment
* reduces data redundancy and inconsistency
  + minimizing isolated files in which the same data are repeated
  + if the organization maintains some redundant data, using a DBMS eliminates data inconsistency
* uncouples programs and data, enabling data to stand on their own.
* Access and availability of information will be increased and program development and maintenance costs reduced
* Enables the organization to centrally manage data, their use, and security.
* Data sharing throughout the organization is easier
* data are presented to users as being in a single location
* Relational DBMS
* Most popular type of DBMS today for PCs as well as for larger computers and mainframes
* Represent data as two-dimensional tables (called relations).
* Tables may be referred to as files.
* Each table contains data on an entity and its 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



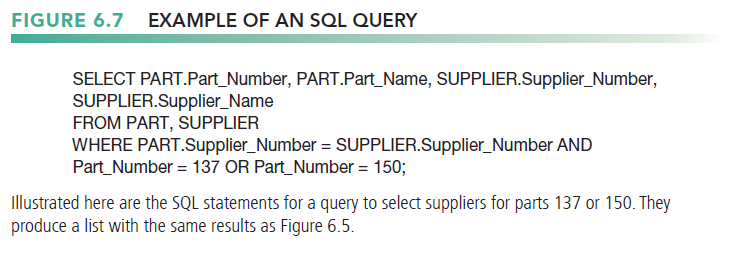
* Operations of a Relational DBMS
* Relational database tables can be combined easily to deliver data required by users, provided that any two tables share a common data element.
* Three basic operations used to develop useful sets of data
* SELECT: Creates subset of data of all records that meet stated 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



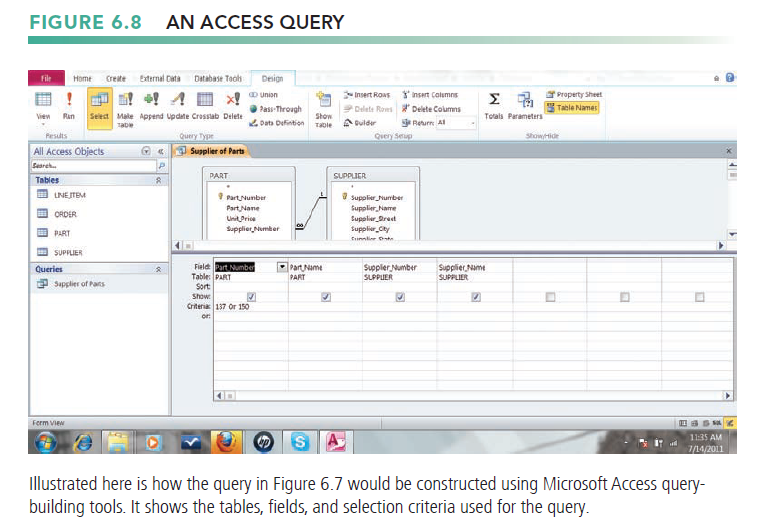
* + Capabilities of Database Management Systems
* Preface
* The most important are its data definition language, data dictionary, and data manipulation language.
* Data definition capability
* Specifies structure of database content, used to create tables and define characteristics of fields
* Data dictionary
* Automated or manual file storing definitions of data elements and their characteristics



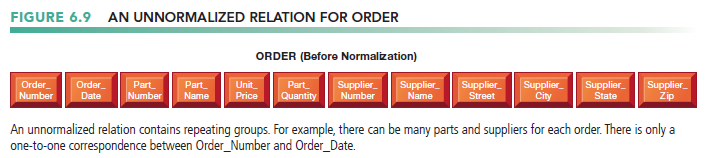
* Querying and Reporting
* Data manipulation language: Used to add, change, delete, retrieve data from database
* Structured Query Language (SQL)
* Microsoft Access user tools for generating SQL

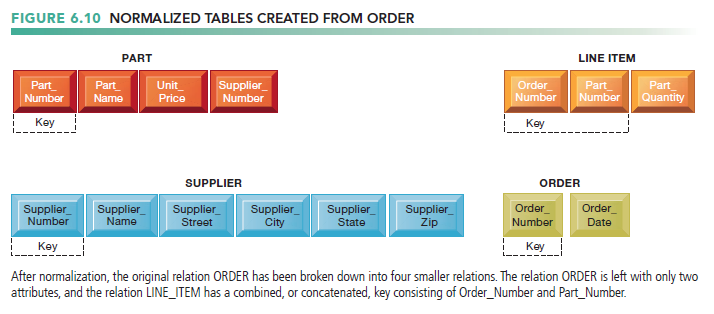


* Many DBMS have report generation capabilities for creating polished reports (Crystal Reports)

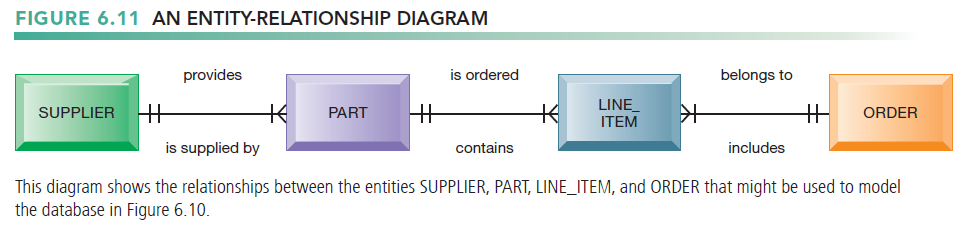


* + Designing Databases
* Normalization and Entity-Relationship Diagrams
* The conceptual (logical) database design
* how the data elements in the database are to be grouped
* Physical design: How database is arranged on direct-access storage devices
* The design processes
* identifies relationships among data elements and the most efficient way of grouping data elements together to meet business information requirements
* identifies redundant data elements and the groupings of data elements required for specific application programs.
* Normalization
* Streamlining complex groupings of data to minimize redundant data elements and awkward many-to-many relationships





* Referential integrity rules
* Used by RDMS to ensure relationships between tables remain consistent
* Entity-relationship diagram
* Used by database designers to document the data model
* Illustrates relationships between entities
* Caution: If a business doesn’t get data model right, system won’t be able to serve business well



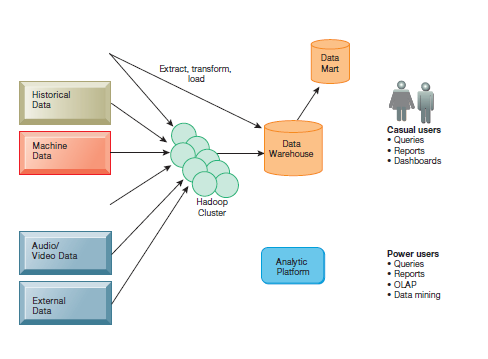
* + Non-relational Databases and Databases in the Cloud
* Non-relational databases: “NoSQL”
* More flexible data model
* Data sets stored across distributed machines
* Easier to scale
* Handle large volumes of unstructured and structured data (Web, social media, graphics)
* Example
* Oracle NoSQL, Amazon’s SimpleDB, MongoDB
* Cloud Databases
* Typically, less functionality than on-premises DBs
* Amazon Relational Database Service, Microsoft SQL Azure
* Private clouds

1. What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?
   * The Challenge of Big Data

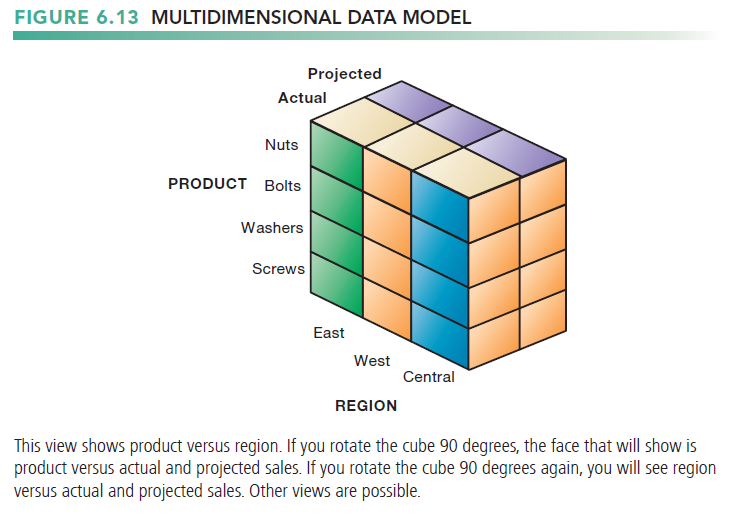
* Big data
* Massive sets of unstructured or semi-structured from Web traffic, social media, sensors and thus not suitable for relational database
* Petabytes, exabytes of data
* Can reveal more patterns and anomalies
  + Business Intelligence Infrastructure
* Preface
* Today includes an array of tools for separate systems, and big data
* Data Warehouses and Data Marts
* Data Warehouses
* Stores current and historical data from many core operational transaction systems
* combined with data from external sources and transformed by correcting inaccurate and incomplete data and restructuring the data for management reporting and analysis
* Consolidates and standardizes information for use across enterprise, but data cannot be altered
* Provides analysis and reporting tools
* Data marts
* subset of a data warehouse
* a summarized or highly focused portion of the organization’s data is placed in a separate database for a specific population of users
* Typically focuses on single subject or line of business
* Hadoop
* an open source software framework managed by the Apache Software Foundation
* enables distributed parallel processing of huge amounts of data across inexpensive computers.
* In-Memory Computing
* breaks a big data problem down into sub-problems, distributes them among up to thousands of inexpensive computer processing nodes, and then combines the result into a smaller data set.
* Key services
* Hadoop Distributed File System (HDFS): data storage
* MapReduce: breaks data into clusters for work
* Hbase: NoSQL database
* can process large quantities of any kind of data, including structured transactional data, loosely structured data, complex data, and unstructured audio and video data.
* Used by Facebook, Yahoo, NextBio
* Analytic Platforms
* Used in big data analysis
* Uses computers main memory (RAM) for data
* Users access data stored in system primary memory, thereby eliminating bottlenecks from retrieving and

reading data in a traditional, disk-based database and dramatically shortening query response times.

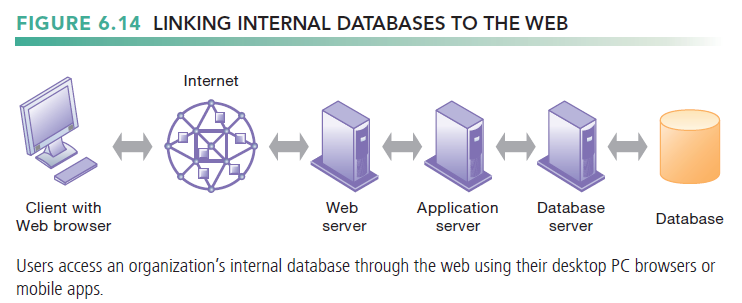
* Can reduce hours/days of processing to seconds
* Requires optimized hardware
* Analytic Platforms
* High-speed platforms using both relational and non-relational tools optimized for large datasets
* Example :
* IBM PureData System for Analytics,



* + Analytical Tools: Relationships, Patterns, Trends
* Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
* 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
* Example: How many washers sold in the East in June compared with other regions?
* OLAP enables rapid, online answers to ad hoc queries



* Data Mining
* more discovery-driven
* insights into corporate data that cannot be obtained with OLAP by finding hidden patterns and relationships in large databases and inferring rules from them to predict future behavior.
* The patterns and rules are used to guide decision making and forecast the effect of those decisions.
* Types of information obtainable from data mining
* Associations are occurrences linked to a single event.
* In sequences , events are linked over time.
* Classification recognizes patterns that describe the group to which an item belongs by examining existing items that have been classified and by inferring a set of rules
* Clustering works in a manner similar to classification when no groups have yet been defined.
* Forecasting uses predictions in a different way. It uses a series of existing values to forecast what other values will be.
* Text Mining
* extract key elements from unstructured big data sets, discover patterns and relationships, and summarize the information.
* Stored e-mails, Call center transcripts, Legal cases, Patent descriptions, Service reports, and so on.
* Sentiment analysis software
* mine text comments in an e-mail message, blog, social media conversation, or survey form to detect favorable and unfavorable opinions about specific subjects.
* Web Mining
* The web is another rich source of unstructured big data for revealing patterns, trends, and insights into customer behavior.
* Understand customer behavior
* Evaluate effectiveness of Web site, and so on
* Web mining looks for patterns in data through content mining, structure mining, and usage mining.
* Web content mining
* the process of extracting knowledge from the content of webpages, which may include text, image, audio, and video data.
* Web structure mining
* Analyzes links to and from Web page
* Web usage mining
* Mines user interaction data recorded by Web server
  + 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
* handles all application operations and data access , between browser-based computers and a company’s back-end business applications or databases. , between browser-based computers and a company’s back-end business applications or databases.
* A CGI script is a compact program using the Common Gateway Interface (CGI) specification for processing data on a web server.
* Database server (hosting DBMS)
* 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



1. Why are information policy, data administration, and data quality assurance essential for managing the firm’s data resources?
   * Establishing an Information Policy

* information policy
* Firm’s rules, procedures, roles for sharing, managing, standardizing, classifying, and inventorying data.
* Data administration
* Establishes policies and procedures to manage data
* Data governance
* the policies and processes for managing the availability, usability, integrity, and security of the data employed in an enterprise with special emphasis on promoting privacy, security, data quality, and compliance with government regulations.
* Database administration
* defining and organizing the structure and content of the database and maintaining the database.
  + Ensuring Data Quality
* More than 25 percent of critical data in Fortune 1000 company databases are inaccurate or incomplete
* Redundant data
* Inconsistent data
* Faulty input
* Before new database in place, need to
* Identify and correct faulty data
* Establish better routines for editing data once database in operation
* Data quality audit
* a structured survey of the accuracy and level of completeness of the data in an information system
* surveying entire data files, surveying samples from data files, or surveying end users for their perceptions of data 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