Introduction to Information Systems

Data Science Education Program

Chapter #4

Databases and data warehouses

Chapter #4 Overview

LEARNING OBJECTIVES

1. Explain the nature of information resources in terms of structure and quality, and show how metadata can be used to describe these resources.
2. Compare file processing systems to the database, explaining the database's advantages.
3. Describe how a relational database is planned, accessed, and managed, and how the normalization process works.
4. Explain why multiple databases emerge, and how master data management helps address the challenge of integration.
5. Describe how a data warehouse is created, and explain the challenges and value of big data.
6. Explain how the human element and ownership issues affect information management.

Key Terms and Concepts

KEY TERMS AND CONCEPTS

structured information
unstructured information
semi-structured information
metadata
table
record
field
data definition
batch processing
database
database management software (DBMS)
relational database
data model
primary key
autonumbering
normalization
functionally dependent
foreign keys
Structured Query Language (SQL)
interactive voice response (IVR)
scalability
referential integrity
database schema
data dictionary
shadow system
master data management
data steward
data warehouse
extract, transform, and load (ETL)
big data
data mining

Online Simulation

An online, interactive decision-making simulation that reinforces chapter contents and uses key terms in context can be found in MyMISLab™.

Chapter #4 Topics

- The nature of information resources
  - Structured, unstructured, and semi-structured information
  - Metadata
  - The quality of information
- Managing information from filing cabinets to database
  - Tables, records, and fields
  - The rise and fall of file processing systems
  - Databases and database management software
- Developing and managing a relational database
  - Planning the data model
  - Accessing the database and retrieving information

Chapter #4 Topics

- The ethical factor: (*ethical issues in database design: the case for ethnic identification*)
- Multiple databases and the challenge of integration
  - Shadow systems
  - Integration strategies and master data management
- Data warehouses and big data
  - Ownership issues
  - The challenge of big data
  - Strategic planning, business intelligence, and data mining
- The challenge of information management: the human element
  - Ownership issues
  - Databases without boundaries
  - Balancing stakeholders: information needs

Introduction

Introduction

- There is a global tidal wave of data and information engulfing *individuals*, *organisations* (of all types), and *governments*
- The huge volume of data and information is driven by the technological advances which provide the capability to:
  - Capture, store, and process data into information (hopefully useful to users!)
  - Analyse and investigate the available data and information (e.g., *big data analytics*)
  - *Information is power! But a strategic approach to it is needed*
- In this chapter we consider:
  - Databases and data warehousing
  - Apply the concepts introduced to the *Volunteer Now! Online Simulation*

The nature of information resources

The modern workspace and information

- Figure 4.1.:
  - Shows a typical modern
- Figure 4.2.:
  - Shows the nature and types of information

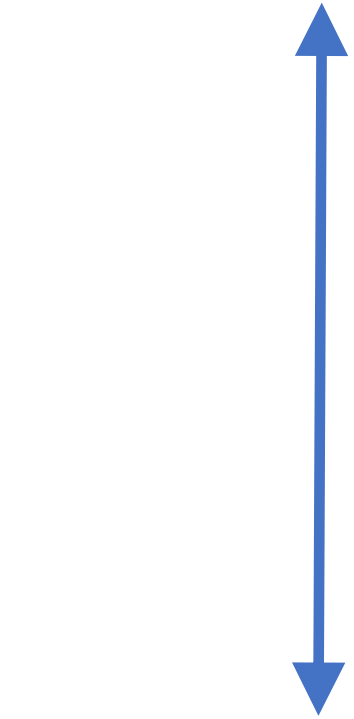
FIGURE 4-1
The modern workspace: An information storehouse.

FIGURE 4-2
Types of information resources.

Type of Information Resource	Example
Structured information	A sales transaction with clearly defined fields for date, customer number, item number, and amount
Unstructured information	Manila folder containing assorted items about a lawsuit, such as photos, handwritten notes, newspaper articles, or affidavits
Semi-structured information	A web page with a title, subtitle, content, and a few images

The nature of information and decisions

Unstructured Information



Unstructured Information

Metadata

What is metadata?

- Figure 4.3.:
  - Provides an illustration of *metadata*

PRODUCTIVITY TIP
Adding metadata to the properties of your documents, photos, and videos makes them easier to search and locate later. Right-clicking on the filename usually brings up a menu that includes Properties. You can also remove information from a file's properties so others will not see it.

structured information
Facts and data that are reasonably ordered, or that can be broken down into component parts and organized into hierarchies.

unstructured information
Information that has no inherent structure or order, and the parts can't be easily linked together.

semi-structured information
Information category that falls between structured and unstructured information. It includes facts and data that show at least some structure, such as web pages and documents, which bear creation dates, titles, and authors.

metadata
Data about data that clarifies the nature of the information.

Photo Metadata	Description
Photo title	Ocean beach scene
Date taken	12/15/2011
License type	Royalty free
Photographer	Felipe DiMarco
Key words	Ocean, waves, outdoors, sunshine, beach, vacation, swimming, swimmers, fishing, surf

FIGURE 4-3
Metadata for a beach scene photo.

The quality of information

Information quality metrics

- Not all information has high quality - there are important characteristics that affect quality:
  - *Accuracy* (errors reduce the quality of the information
  - *Precision* (rounding may not be acceptable)
  - *Completeness* (Omitting required information can be an issue)
  - *Consistency* (inconsistency in how data is represented is an issue)
  - *Timeliness* (this can be an issue – also, there is no clear definition of the term)
  - *Bias* (unbiased information is required as biased information lacks objectivity, and that reduces its value and quality
  - *Duplication* (Information can be redundant, resulting in misleading and exaggerated summaries)

Surveys, data, and information

- Data collected by online surveys illustrate many of the problems surrounding information quality
- Sample populations and respondents can be biased - for example:
  - Virtual Surveys Ltd (a company that specializes in web-based research):
    - Discovered that one person completed an online survey 750 times because a raffle ticket was offered as an incentive
- To avoid relying on poor quality data:
  - Managers must identify suitable unbiased research populations
  - The population must reflect the socio-demographic profile
  - The survey must define what constitutes quality for the information they need

Managing information: from filing cabinets to the database

Introduction

- Managing information – from filling cabinets to the database:
  - Information management predates the digital age
  - Before the invention of the lateral filing cabinet in 1898 businesses:
    - Organised documents by putting them in envelopes (or) in rows of small pigeonholes that lined entire walls from top to bottom
    - The change to vertical manila folders, neatly arranged in cabinet drawers, was quite an improvement for record keeping, and much appreciated by file clerks (see Figure 4-4)
- The information age and the digital revolution dating from the 1960's:
  - Resulted in the introduction of Computerised systems
  - Digital systems and the database technologies are reliant on the concept of the record

Early Information Management Methods

- Figure 4.4.
  Shows a filling cabinet and a pigeon hole approach to the management of documents and information

FIGURE 4-4
Early information management approaches.

Tables, records, and fields

Database Fundamentals

- Recall that in INFO 151 we introduced the basics of database technologies and *Relational Database Management Systems* (RDMS) where we introduced MySQL and:
  - *Tables*, *records*, and *fields*
  - Each field will have a *data definition* which specifies the properties such as:
    - *Alphabetic*, *numeric*, and the *cardinality*

table
A group of records for the same entity, such as employees. Each row is one record, and the fields of each record are arranged in the table's columns.

record
A means to represent an entity, which might be a person, a product, a purchase order, an event, a building, a vendor, a book, a video, or some other "thing" that has meaning to people. The record is made up of attributes of that thing.

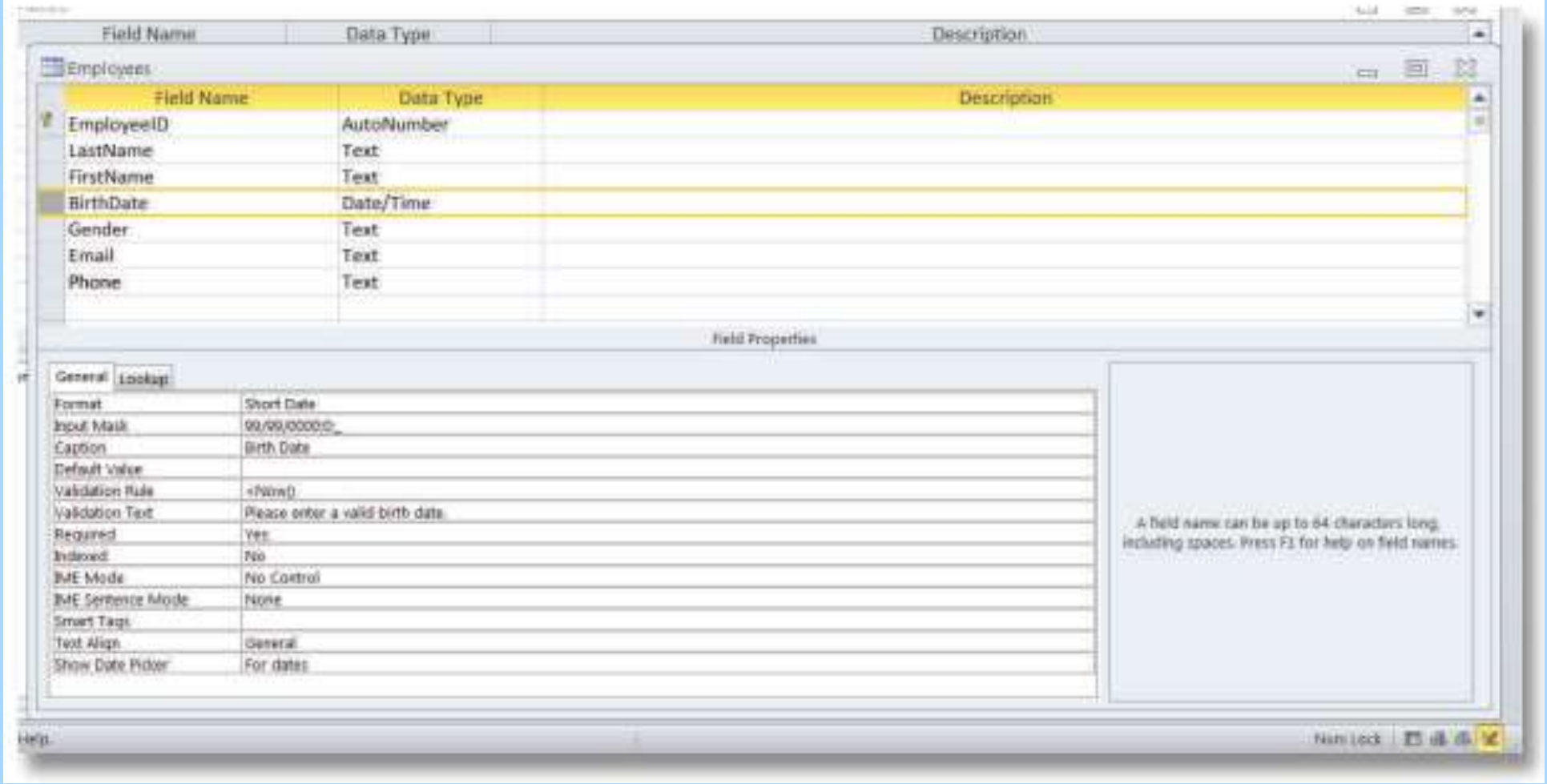
field
An attribute of an entity. A field can contain numeric data or text, or a combination of the two.

data definition
Specifies the characteristics of a field, such as the type of data it will hold or the maximum number of characters it can contain.

A data definition from Office 365 Access

FIGURE 4-5

Data definition for the field "birthdate" in MS Access.



The rise and fall of file processing systems

File Processing Systems

- Manual records
- Electronic records
  - Office systems
  - Batch processing
  - Desktop processing
- Potential issues:
  - Data redundancy and inconsistency
  - Lack of integration
  - Inconsistent data definitions
  - Data dependence

Batch Processing (1)

- Batch processing:
  - Electronic records were created and stored as computer file
  - Computer programs were created to add, delete, or edit the records
  - Each department maintained its own records with its own computer files, each containing information that was required for operations
- For example:
  - The payroll office maintained personnel records and had its own computer programs to maintain and manage its set of file
  - To generate payroll checks:
    - The payroll system's computer programs would read each record in the file and print out checks and payroll stubs for each person using the information contained in the files for that department

Batch Processing (2)

- Batch processing:
  - That kind of activity is called batch processing
  - The program is sequentially conducting operations on each record in a large batch
  - Accounts payable and receivable, personnel, payroll, and inventory were the first beneficiaries of the digital age.
  - Compared to the manual method of generating a payroll (where deductions and taxes were computed by hand and each check was individually typed) the monthly batch processing of computer-generated checks was revolutionary
  - However:
    - It didn't take long for problems to surface as other offices began to develop their own file processing systems
    - Understanding problems is critical in understanding why databases offer benefits

Data Redundancy and Inconsistency

- Figure 4.6. shows:
  - The inconsistencies created by computer programs operating on it's own records
  - The result is redundant and inconsistent information

FIGURE 4-7
Information in separate file processing systems is difficult to integrate. For example, a report listing hourly rates by gender would need extra programming effort in this business.

Lack of Data Integration (1)

- Figure 4.7. models:
  - The challenge is integrating data in separate systems (files)
  - Listing hourly rates by gender requires additional programming

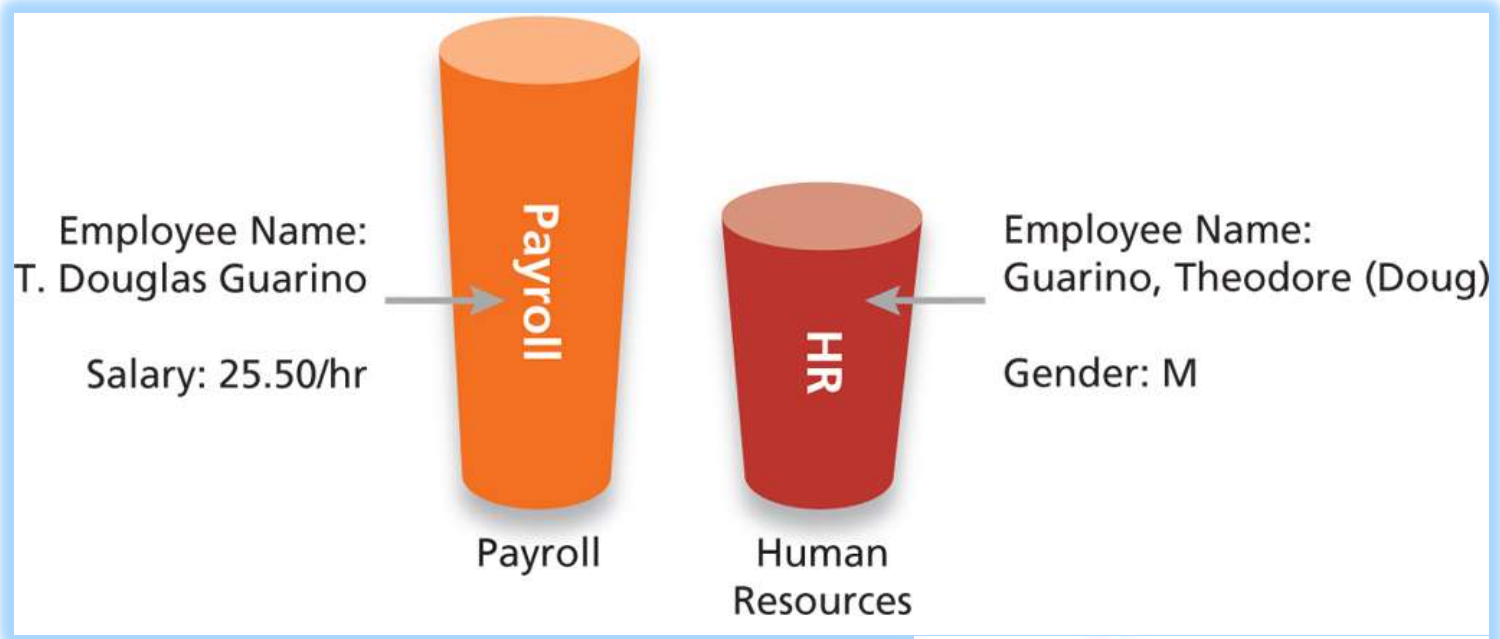


FIGURE 4-7
Information in separate file processing systems is difficult to integrate. For example, a report listing hourly rates by gender would need extra programming effort in this business.

Lack of Data Integration (2)

- Figure 4.8. shows:
  - How data integration issues affect customers
  - The issue illustrated here is that customers cannot resolve inconsistencies in their accounts

FIGURE 4-8
Separate file processing systems lead to a fragmented customer interface, frustrating customers who have to contact several offices to straighten out inconsistencies.

Inconsistent Data Definitions (1)

- When programmers write code to handle files:
  - Differences in format are created
  - Phone numbers may include the dashes and be formatted as a text field in one system but be treated as numbers in another
- A problem is how people use the system
  - Data definitions may seem similar across systems (but they are used differently and summaries become misleading)
  - For example
    - The personnel department may categorize software purchases as *computers*
    - The sales lump *software* with pencils, staplers, and clocks as *supplies* (because less paperwork is needed to justify the purchase)

Inconsistent Data Definitions (2)

- The result can be issues in knowing (for example) how much is spent on technology because of the human element in information systems
  - Figure 4.9.: demonstrates the inconsistencies

FIGURE 4-9
When data definitions are inconsistent, the meaning of different fields will vary across departments and summaries will be misleading. Note how the three departments use categories in different ways.

Department	Object Code	Amount	Category	Description
Sales	4211	1888.25	Computers	Desktop Computers
Sales	4300	249.95	Computer supplies	Image editing software
Sales	4100	29.99	Office supplies	Flash drive
Personnel	4211	59.00	Computers	Stastical software
Personnel	4300	14.95	Computer supplies	Flash drive
Personnel	4211	2500.21	Computers	Laptop Computers
Warehouse	4211	59500.00	Computers	Web server
Warehouse	4211	2500.00	Computers	Printer/copier/scanner/fax

Data Dependence

- Early systems became maintenance nightmares because the programs and their files were so interconnected and dependent on one another
- The programs:
  - Defined the fields and their formats, and business rules were all hard-coded or embedded in the programs
  - Even a minor change to accommodate a new business strategy took a lot of work
  - IT staff were constantly busy (but kept falling behind anyway)
- The disadvantages to the file processing approach:
  - Resulted in a better way of organizing structured data using a database

Databases and relational database management

Database Topics in Chapter #4 **NOT** Covered

- Info 151 introduced:
  - Database basics with the basics of relational databases including building a simple database using MySQL with web-based access using PHP
  - Furthermore: there will be a course later in 2021 which focuses on database
- Refer to Info 151 for our overview of database basics
  - We will **NOT** cover the following topics:
    - *Databases and Database Management Software* (pages 134-137)
    - *Developing and Managing a Relational Database* (pages 137-145)
  - For Chapter #4 we will cover the following sections:
    - Sections #4, #5, #6 (pages 145-152) will be addressed in this course

Natural Language Interfaces

- Are designed to correctly interpret:
  - Queries entered in natural language
  - Spoken queries in natural language
- Current systems include:
  - Microsoft *Cortana*: a virtual assistant for many Microsoft operating systems and applications
  - Amazon: smart speaker with *Siri*:
    - An intelligent assistant that claims to offer a faster, easier way to get things done on your Apple devices
  - Such systems are however at an embryonic stage when considered from a NLP perspective

Natural Language Processing

NATURAL LANGUAGE INTERFACES To many, the holy grail of query languages is the capability to understand and correctly reply to natural language queries, either spoken or typed. Although vendors have attempted to make end-user queries easier to do, the ability to correctly interpret a person's question is still limited, though many promising applications are underway.[6] Apple's Siri, for instance, can interpret a range of spoken questions and search its databases. "What is the best pizza parlor near here?" is something Siri could answer, partly because it knows your location through GPS, and it can query Yelp's database of restaurant reviews (www.yelp.com). But it can't easily answer highly unstructured questions, or questions that rely on databases Siri cannot access.

For business queries, the natural language query systems work well when the questions use a limited vocabulary. For example, "Which employees make more than $100,000 per year?" could be translated into SQL with reasonable accuracy. However, problems arise when the vocabulary is vague, the attribute names can be confused, or the question itself is not clear. Even the question about the high-earning employees could be interpreted more than one way. For example, did the user intend to include benefits and stock options? Should "employees" include part-time people? Natural language query systems are improving very rapidly, however, as Siri and IBM's Watson demonstrate.

Did You Know?

IBM's "Watson," named after the company's founder, is the supercomputer that trounced the top two human players in the TV game show *Jeopardy*. Its ability to understand complex human language queries is astounding. To help Watson interpret more informal speech, researchers fed it slang repositories such as the Urban Dictionary, but they forgot to teach good manners. When Watson starting swearing, they had to wipe those memories clean.[7]

Multiple databases and the challenge of integration

Multiple Databases

- Multiple databases and the challenges in integration:
  - Integration strategies and master data
- Shadow systems
- Integration strategies and master data management:
  - Master data management
    - Efforts to achieve common uniform definitions for entities and attributes
  - Data stewards
    - Assigned as a ‘watchdog’ and ‘bridge-builders’ to manage data definitions

shadow system
Smaller databases developed by individuals outside of the IT department that focus on their creator's specific information requirements.

master data management
An approach that addresses the underlying inconsistencies in the way employees use data by attempting to achieve consistent and uniform definitions for entities and their attributes across all business units.

data steward
A combination of watchdog and bridge builder, a person who ensures that people adhere to the definitions for the master data in their organizational units.

data warehouse
A central data repository containing information drawn from multiple sources that can be used for analysis, intelligence gathering, and strategic planning.

Multiple Databases and Information Systems

- Single databases have proved to be very useful
- However:
  - As organisations develop and change multiple databases may be created in an ad-hoc manner
  - When organisations merge each will have a database
  - In both cases the existence of multiple databases can be a significant problem for the reasons discussed in earlier sections of this chapter
  - Cloud-based services are adding to the problem as databases can be created almost ‘on-the-fly’
- These factors pose important organisational issues for data integtity and information systems

Shadow systems

Shadow Systems (1)

- Although the integrated enterprise database is a critical resource:
  - Changes to support new features can be very slow
  - People want to get their jobs done as efficiently as possible, and sometimes the quick solution is to create a shadow system
  - These are smaller databases developed by individuals or departments that focus on their creator's specific information requirements
  - They are not managed by central IT staff (indeed: they may not even know they exist
- Shadow systems:
  - Are easy to create with tools like Access, Excel, and XML
  - However: the data and information they hold may not be consistent with what is in the corporate database

Shadow Systems (2)

- Another hazard is:
  - Departments may have problems when the creator leaves because no one else knows quite what the shadow system does
  - These problems lead to serious headaches for managers who need enterprise-wide summaries to make decisions.
  - Managers may receive many *versions of the truth* from different sources since the information is *inconsistent*
- Companies should:
  - Try to reduce shadow systems and integrate systems as much as possible
  - This promotes data and information consistency in reports used for planning or compliance

Integration strategies and master data management

Integration Strategies and Interfaces

- An integration strategy can involve:
  - The creation of interfaces (or bridges) between databases to link common fields
- Using this approach:
  - A field that is updated in one database (such as an email address):
    - Is copied to the same fields in other databases that maintain that information
    - In the *downstream* databases the email address would be in read-only format, so that end users could not update it there

Integration Strategies and Master Data Management

- A broader strategy to address underlying inconsistencies in the way people use data is *master data management*:
  - This approach attempts to achieve uniform definitions for entities and their attributes across all business units, and it is especially important for mergers
  - There must be agreement on how everyone will define terms such as *employee*, *sale*, or *student*.
  - For example: should *employees* include temporary contractors or student workers?
  - The most successful efforts at master data management focus mainly on a key area (such as *customers*), or on a limited number of entities that are most important
  - Teams from across the company meet to identify the differences and find ways to resolve them. Data stewards may then be assigned as watchdogs and bridge builders to remind everyone about how data should be defined

Master Data Management and Data Warehousing

- Master data management:
  - Has less to do with technology than with *people*, *processes*, and *governance*.
- For example:
  - Nationwide Insurance launched a master data management initiative to resolve its fragmented environment with 14 different general ledger platforms
  - The results were slow in coming but eventually were dramatic:
    - Initially it once took the company 30 days to close the books
    - Within a year the time to close the books was reduced around 15 days
- Another integration strategy is the *data warehouse*:
  - This approach is even more effective when master data management efforts reconcile data inconsistencies and improve data quality from multiple sources
  - In the following slides we introduce data warehousing and consider big data

Data warehouses and big data

The Data Warehouse

- A *data warehouse* is:
  - A central repository containing information drawn from multiple sources
  - The data (in the data warehouse) is useful in analysis and investigation, intelligence gathering, and strategic planning
  - Figure 4.22. shows some typical examples of data sources that may contribute to a data warehouse
  - A critical *internal* source of data is *operational* data drawn from an organisations own systems such as:
    - *Customer records*, *transactions*, *inventory*, *assets* and *liabilities*, *human resources information*, and other *relevant company data and information*
    - Such data cab range from current data to historical data going bac over many years

The Data Warehouse (Figure 4.22.)

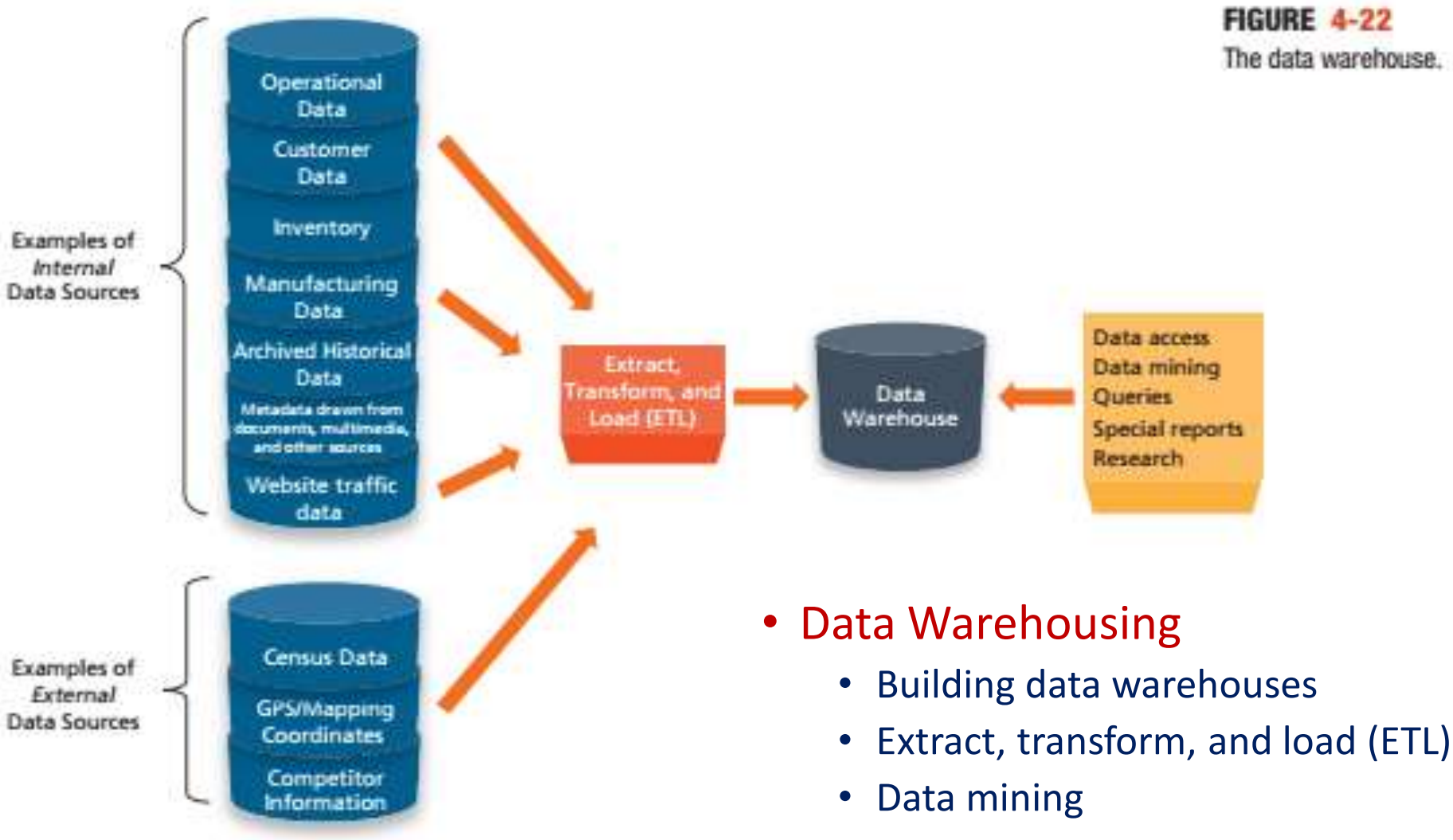


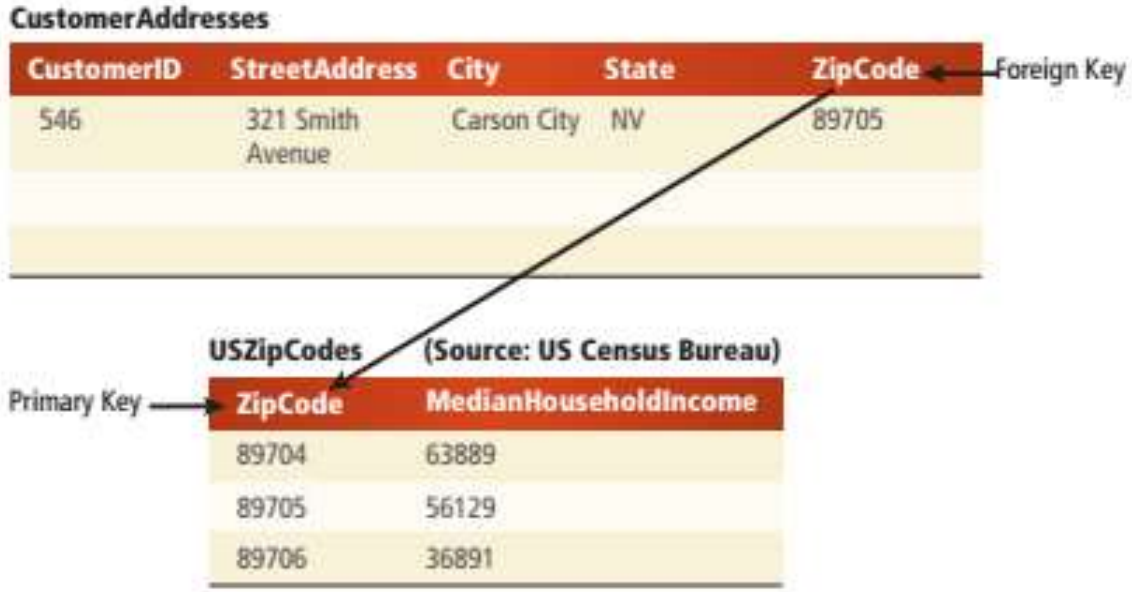
FIGURE 4-22
The data warehouse.

- Data Warehousing
  - Building data warehouses
  - Extract, transform, and load (ETL)
  - Data mining

Figure 4.23.

- Figure 4.23. shows:
  - How the link between two tables can be made
  - Adding an *external resource* can inform preferences and behaviour
  - *Customers shopping habits* can be analysed and used for predicting future activity

FIGURE 4-23
External sources of data can be added to the warehouse to increase its value. Here, a table from the U.S. Census Bureau containing the median household incomes for each zip code can be linked to the customers by means of the zip code field.



Data Warehousing

- The ability to draw on high-quality information from an organization's information systems and external sources enables:
  - Trends can be identified
  - Historical patterns can be identified
  - Reports can be generated for compliance purposes
  - Research into the organisations activity can be created
  - Strategy can be investigated, planned, and monitored
- While databases that support the day-to-day business contain much information that goes into a warehouse:
  - The data are typically not in a format that works well for broad analyses

Data Warehousing

- . As we saw earlier:
  - Organizations often have more than one database
  - A second reason the operational database is not a good candidate for high-level management reporting is that the DBA has to optimize its performance for operations
  - Fast customer response and data entry come first:
    - Not complex queries that answer the bigger, strategic questions
- Strategic planning queries may typically span many years of data:
  - Computing such queries carries high levels of computational overhead
  - Therefore: it makes sense to run them on a separate data warehouse (not the operational database)

Building the data warehouse

Creating a Data Warehouse

- How do we create a data warehouse?
  - A common strategy for drawing information from multiple sources is extract, transform, and load (ETL) (Figure 4-22).
  - The first step is to extract data from its home database, and then transform and cleanse it so that it adheres to common data definitions.
  - As we discussed, this is not a minor challenge, and computer programs rarely can handle it alone. Data drawn from multiple sources across organizations, or even within the same organization, might be defined or formatted differently.
  - If the organization has already made progress with master data management this transformation process can be smoother
  - Indeed, attempts to build a data warehouse often expose a lot of “dirty data” (e.g., inconsistent name spellings)

Creating a Data Warehouse

- To address inconsistent name spellings there is an increase in interest in master data management:
- For example: consider the *Rensselaer Polytechnic Institute* (RPI)
  - The need for improved data quality for strategic planning was recognised due to conflicting reports
  - The institute's data warehouse effort uncovered many problems that needed cross-functional teams to resolve before data could be loaded into the warehouse
  - Once they agreed on definitions:
    - Data stewards were appointed to watch over how the fields were used
  - The transformation process applies to external resources that will enrich the value of the data warehouse for intelligence gathering and marketing

A Data Warehouse

- The *U.S. Census Bureau* records need preparation prior to loading into a data warehouse:
  - The motivation is to ensure the zip code fields are the same data type when they are linked
- Some reformatting might be necessary whenever the company adds external data to its warehouse.
- Following transformation the data is loaded into the data warehouse (typically another database)
- At frequent intervals, the load process repeats to keep it up to date

A Data Warehouse

- The DBA optimizes the warehouse for complex reporting and queries, without having to worry about slowing down customers and staff
- Many data warehouses take advantage of standard relational database architectures, and most DBMS products can be optimized for use as a warehouse
- Some include tools to help with the extractions, transformations, and loading, as well
- Organizations also use alternative data warehouse architectures:
  - Figure 4.24. shows typical data warehouse architectures
  - The examples relate to large data sets (often incorrectly termed *big data*) and Internet based systems (which typically use NoSQL)

Figure 4.24.

Data Warehouse Architectures	
Relational database	Companies often use the same relational DBMS for their data warehouse as they use for their operational database, but loaded onto a separate server and tuned for fast retrieval and reporting.
Data cubes	This architecture creates multidimensional cubes that accommodate complex, grouped data arranged in hierarchies. Retrieval is very fast because data are already grouped in logical dimensions, such as sales by product, city, region, and country.
Virtual federated warehouse	This approach relies on a cooperating collection of existing databases; software extracts and transforms the data in real time rather than taking snapshots at periodic intervals.
Data warehouse appliance	The appliance is a prepackaged data warehouse solution offered by vendors that includes the hardware and software, maintenance, and support.
NoSQL	Database management systems suited for storing and analyzing big data. NoSQL stands for "not only SQL."
In-memory database	Relies on main memory to store the database, rather than secondary storage devices, which vastly increases access speeds.

FIGURE 4-24
Data warehouse architectures.

A typical web-based systems database architecture

The challenge of big data

The Challenge of Big Data

- Building a data warehouse from operational databases and adding some external sources are manageable (and very useful) for most organizations
  - However: consider all the other sources of data (including the Internet which is arguably a primary source of data)
- Consider a company's website
  - Even a medium sized company might have thousands of hits per day, and each visitor might click dozens of times
  - Consider also how much semi-structured and unstructured information flows through Twitter, YouTube, Facebook, and Instagram, some of which could give the company a competitive advantage if analyzed quickly

Figure 4.25.

- The figure shows the Tweeting statistics for a clothing retailer (H&M)
- Given a system with sufficient speed:
  - *Trends can be spotted early and reactions initiated*
  - *This can address sales trends and problems*

FIGURE 4-25
Twitter posts can be part of the big data a company can analyze.

Big Data

- A barrage of tweets (see Figure 4.25.) would get the attention of managers at H&M
- The amount of data available is also exploding because so much is gathered automatically using: *sensors*, *cameras*, *RFID readers*, and *mobile devices*
- In consumer electronics, for instance, devices designed to monitor your personal health readings can transmit information to a smartphone app so you can see real-time displays
- This *Internet of Things* (IoT)," in which so many devices are collecting and transmitting data to one another means that data is accumulating at a breathtaking rate (ar faster than even Moore's Law would predict)

What is big data?

What is Big Data?

- Big data refers to collections of data that are so enormous in size, so varied in content, and so fast to accumulate that they are difficult to store and analyze using traditional approaches
- The three "Vs" are the defining features for big data (see Figure 4-26):
  - *Volume*: data collections can take up petabytes of storage, and are continually growing
  - *Velocity*: many data sources change and grow at very fast speeds. The nightly ETL process often used for data warehouses is not adequate for many real-time demands
  - *Variety*: relational databases are very efficient for structured information stored in tables, but businesses can benefit from analyzing semi-structured and unstructured data as well

Big data technologies

Big Data Technologies

- Relational databases may be part of any effort to analyze big data:
  - But a number of new technologies are under development to better handle the *three Vs*
- For example:
  - Database platforms that don't rely on relational structures are emerging, called NoSQL (Not Only SQL) databases which may be created using *horizontal* and *vertical scaling* (as discussed in Info 151)
  - These don't require fixed schemas with clear data definitions for each attribute
  - They also don't generally enforce strict rules the way a relational database does
  - Also: there are in-memory databases (e.g, HP) where the database itself is stored in main memory rather than on a separate hard drive
- This emerging technology:
  - Is extensively used in Internet-based systems
  - It can increases access speeds and is gaining popularity for applications that need very fast response times

Big Data and Hadoop

- Another useful technology for big data is Hadoop:
  - An open source software that supports distributed processing of big data sets across many computers
  - The software manages file storage and local processing, and can be scaled up to thousands of computers in the cloud.
- Internet radio service Pandora uses Hadoop to analyze some 20 billion *thumbs up* and *down* ratings that 150 million users click when each song plays.
  - The company can accurately predict customer preferences and create tailored playlists
  - A case study in Chapter 6 describes Pandora's technology in more detail

Big Data and Hadoop

- Gaming site *King.com* also uses Hadoop:
  - To explore the company's big data
  - The data includes the activities of more than 60 million registered users who play billions of games every month
- The games are free:
  - The company monetises the operations generating revenue by selling in-game products to players, such as extra lives
- While their MySQL relational database worked well for a time:
  - As discussed in Info 151 MySQL has scaling issues
  - The increased volume (for *King.com*) was just too much to handle

Hadoop Functionality

- With Hadoop and related big data technologies:
  - Organisations can identify and investigate trends and patterns that uncover behavior patterns that would be easy to miss in smaller samples (but that can help improve the games)
- For the *King.com* example:
  - Are players getting stuck too much at certain levels?
  - Do they abandon games with certain features?
  - The director of data warehousing says:
    - *We need to know everything we can. Without that . . . we would be blind*

Hadoop Functionality

- The U.S. National Security Agency:
  - Employs big data technologies to track terrorist activity using Internet and phone records
  - The debate about the agency's work heated up in 2013
  - The public came to understand how very powerful big data can be
- There are significant and far reaching ethical and moral challenges for the big data paradigm
  - The *Ethical Factor* in Chapter #3 explores the ethical implications surrounding big data, especially for privacy rights

Strategic planning, business intelligence, and data mining

Data Mining

- Strategic Planning, Business Intelligence, and Data Mining Data warehouses and big data efforts should make important contributions to strategic planning
- Along with the tools and approaches described in Chapter 7:
  - Access to all of this data opens up a wealth of opportunities for managers seeking insights about their markets, customers, industry, and more
- They become:
  - The main source of business intelligence that managers tap to understand their customers and markets, as well as to make strategic plans
  - The findings are especially valuable when they can accurately predict future events.

Data Mining

- For example, data mining is a type of intelligence gathering that uses statistical techniques to:
  - Explore large data sets
  - Search for hidden patterns and relationships that are undetectable in routine reports:
    - In fact: such searches are used to identify patterns that were unknown and were not being looked for
- In financial markets:
  - The volume and speed of transactions could mask unusual patterns in individual stocks
  - Data mining analytics can be used to uncover efforts to manipulate the prices

The Data Scientist

- The data scientist is an increasingly relevant role in many organisations and:
  - There is a high demand for the role
  - Companies need people who have the skills to identify the most promising data sources, build data collections, and then make meaningful discoveries
  - They need to know the difference between data mining, which leads to important findings, and “data drudging,” which sniffs out relationships that might just occur by accident and that have little value
  - They also must be able to make a compelling case when they do find trends that could add competitive advantage
- A successful data scientist:
  - Is a combination of hacker, analyst, communicator, trusted advisor, and (most of all) a curious individual

The challenges of information management: the human element

The Human Element

- As with all technology-related activities:
  - Managing information resources is not just about managing technology, databases, and big data
  - People and processes form a critical component in information systems and the management of information
- An understanding of a number of important factors:
  - How people interact with information system
  - How people *view*, *guard*, and *share* the information resources they need is a critical ingredient for any successful strategy
  - Such considerations form a central factor as people may not share information:
    - This is important in an organisational context where research groups will (in theory) share their research findings

Ownership issues

Ownership of Information?

- In the workplace:
  - Information resources are found almost everywhere, from file cabinets and desk drawers to the electronic files on portable media and computer hard drives
- While companies may set the policy that all information resources are company-owned:
  - In practice, people often view these resources more protectively, even when compliance and security don't demand tight access controls
  - Norms about how records are used emerge over time, and though many are unwritten, they can certainly affect employees' behavior
  - Salespeople may want to protect access to their own sales leads, or whole departments might want to control who has access to records that they maintain
  - They may prefer that employees outside the department only have *read-only* privileges
- Customers may raise ownership issues:
  - For example: a customer with no last name may request that the DBA change the *surname* field to optional rather than *required*

Ownership of Information?

- Information ownership issues have to be negotiated among many stakeholders
- An additional challenge is:
  - The time required to implement changes to an integrated enterprise database where many people are affected and will want input into the design and development process
  - The design and development process can be time consuming:
    - This applies to the IT staff and analysts (to analyse the impact) and also stakeholders
- Changes to existing file processing systems:
  - Were time-consuming for the IT staff because of the way the code was written (e.g., using Cobol – see the next slide)
  - Changes to the integrated database take less time from IT, but more from the end users

Cobol

- *COBOL* (common business-oriented language) is a compiled English-like computer programming language designed for business use
  - It is imperative, procedural and (since 2002) object-oriented
  - COBOL is primarily used in business, finance, and administrative systems for companies and governments
- COBOL:
  - Is still widely used in applications deployed on mainframe computers, such as large-scale batch and transaction processing jobs
- However:
  - Due to its declining popularity and the retirement of experienced COBOL programmers programs are being migrated to new platforms, rewritten in modern languages or replaced with software packages
  - Most programming in COBOL is now purely to maintain existing applications:
  - However, many large financial institutions were still developing new systems in COBOL as late as 2006 due to the mainframe processing speed

Databases without boundaries

Databases Accessed from Outside an Organisation

- Another example of how the human element interacts with information management involves databases without boundaries:
  - This relates to use-cases where people outside the enterprise enter and manage most of the records
  - These contributors feel strong ownership over their records
- Instagram (for example) registered protests when the company changed its terms of service so that it could sell the photos people upload without their permission and without compensation
  - The Facebook-owned photo sharing site quickly backtracked and changed the policy back, especially after competitors began touting how they would never sell private photos

Databases Accessed from Outside an Organisation

- *Craigslist.com* illustrates other ways in which the human element affects information management
  - Founder Craig Newmark initially sought to help people in San Francisco find apartments and jobs
  - The site soon became the world's largest database of classified ads
  - The side-effect was that this major revenue source for print newspapers dried up
  - Newmark's concerns:
    - Are less about the database technology than about the health of the community and the relentless threats from spammers and fraudsters who can destroy trust in the site.
- Databases without boundaries are also part of emergency disaster relief where:
  - Online databases can help victims find missing family members, organize volunteers, or link people who can provide shelter to those who need it
  - For example: Google launched a *person finder* database after bombs exploded at the Boston Marathon
    - A valuable lesson from the efforts to build databases without boundaries is simply the need to plan for *scalability*
- The growing capabilities of relational databases, along with big data technologies and cloud computing, are essential to support these worldwide repositories

Balancing stakeholders information needs

Balancing Stakeholders' Information Needs

- How should managers balance the information needs of so many stakeholders?
  - Top-level management needs strategic information and insights from big data, along with accurate, enterprise-wide reports
  - Operating units must have reports on transactions that match their operations, and they need information systems that are easily changed to support fast-moving business requirements
  - Customers want simpler user interfaces (usability) that work quickly and reliably, and don't want error messages notifying issues related to failed merging of ed and our computer systems
  - Government agencies: want companies to submit compliance reports using government's definitions, because they have to compile their own summaries
- Meeting all these needs:
  - Represents a balancing act that requires leadership, compromise, negotiation, and well-designed databases
  - As a shared information resource, the database fulfills its role exceptionally well to provide a solid backbone for the whole organization and all its stakeholders

Summary

Chapter #4 Review

- In this lecture we have considered:
- The nature of information resources including:
  - Structured, unstructured, and semi-structured information, metadata, and the quality of information
- Managing information from filing cabinets to database
  - Tables, records, and fields, and the rise and fall of file processing systems
- Multiple databases and the challenge of integration
  - Shadow systems and the integration strategies and master data management
- Data warehouses and big data
  - Ownership issues, the challenge of big data, and strategic planning, business intelligence, and data mining
- The challenge of information management: the human element
  - Ownership issues, databases without boundaries, and balancing stakeholders: information needs

Database Topics in Chapter #4 **NOT** Covered

- Info 151 introduced:
  - Database basics with the basics of relational databases including building a simple database using MySQL with web-based access using PHP
  - Furthermore: there will be a course later in 2021 which focuses on database
- Refer to Info 151 for our overview of database basics
  - We will **NOT** cover the following topics:
    - *Databases and Database Management Software* (pages 134-137)
    - *Developing and Managing a Relational Database* (pages 137-145)
  - For Chapter #4 we will cover the following sections:
    - Sections #4, #5, #6 (pages 145-152) will be addressed in this course

Reading and coursework

Reading and Coursework

- Read and understand the key terms and concepts on page 154
- Work through the:
  - Chapter review questions (page 154)
  - Projects and discussion questions (page 154-155)
  - Application exercises (page 155-156)
- Apply the concepts introduced to the *Volunteer Now! Online Simulation* exercise
- Read and consider the (two) case studies (pages 157-158) and answer the related discussion questions
- Review the (two) *e-projects* on pages 159-160

KEY TERMS AND CONCEPTS

structured information
unstructured information
semi-structured information
metadata
table
record
field
data definition
batch processing
database
database management software (DBMS)
relational database
data model
primary key
autonumbering
normalization
functionally dependent
foreign keys
Structured Query Language (SQL)
interactive voice response (IVR)
scalability
referential integrity
database schema
data dictionary
shadow system
master data management
data steward
data warehouse
extract, transform, and load (ETL)
big data
data mining

CHAPTER REVIEW QUESTIONS

4-1. What are three categories that describe the nature of information resources? Give an example of each. How do you characterize the relationships within each category of information?

4-2. What is metadata? What does metadata describe for structured information? For unstructured information? Give an example of each type of metadata.

4-3. How will you ensure that the data collected in a survey is valuable?

4-4. Discuss how you could manage your information (like notes), based on the approaches on information management discussed in the chapter.

4-5. What are the major disadvantages of file processing systems? What are four specific problems associated with file processing systems?

4-6. Following the file processing model of data management, what three architectures emerged for integrated databases? What are the advantages of each? Are there disadvantages?

4-7. What are the steps in planning a relational data model? Are there benefits to the planning stage?

4-8. How can a relational database model overcome problems arising from traditional file processing systems, for example, in a library?

4-9. What is the typical strategy to access a database? How do users access an Access database? Are there other strategies to access database systems?

4-10. What is the role of the database administrator in managing the database? What is the career outlook for this job?

4-11. What is SQL? How is it used to query a database?

4-12. What is IVR? How is it used to query a database?

4-13. What is a shadow system? Why are shadow systems sometimes used in organizations? How are they managed? What are the advantages of shadow systems? What are the disadvantages?

4-14. What is master data management? What is a data steward? What is the role of master data management in an organization's integration strategy?

4-15. What is a data warehouse? What are the three steps in building a data warehouse?

4-16. Identify the external and internal data sources for a data warehouse to be built for a library.

4-17. What are four examples of data warehouse architectures? Which approach is suitable to meet today's growing demand for real-time information?

4-18. What is big data? What are the defining features of big data?

4-19. What is data mining? What is the difference between data mining and data dredging? What is the goal of data mining?

4-20. What are examples of databases without boundaries?

4-21. How do ownership issues affect information management? How do information management needs differ among stakeholder groups?

PROJECTS AND DISCUSSION QUESTIONS

4-22. Why is metadata becoming increasingly important in this age of digital information? What types of metadata would you expect to see attached to these information resources?

a. Book

b. Digital photograph

c. MP3 file

d. Zappos.com web page for men's athletic shoes

4-23. The concept of relationships is fundamental to relational database design. Briefly describe three relationships that explain how records in a database might be logically related to one another. What are examples of each type of relationship? At your university, what is the relationship between students and courses? What is the relationship between advisors and students?

4-24. Target marketing uses databases and data warehouses to identify potential customers that a business wants to reach based on factors that describe a specific group of people. For example, target markets may be identified by geographic area, by age group, by gender, or by all three factors at one time. One of the leading providers of business and consumer information is infoUSA.com. Visit their website at www.infousa.com to learn how they compile data from multiple sources. (Online at //www.infousa.com/faq/.) How does their process compare to extract, transform, and load (ETL)? Prepare a brief summary of your findings that describes the infoUSA five-step process of building a quality database.

4-25. Visit YouTube.com and search for "R. Edward Freeman Stakeholder Theory" to learn more about stakeholder groups. Are you a stakeholder at any of the following organizations? List several stakeholders at each of these organizations and describe the kind of information each stakeholder needs.

a. A university

b. A regional bank

c. Toyota Motor Corporation

4-26. The idea of data warehousing dates back to the 1980s. Today, data warehousing is a global market worth billions of dollars. What is the relationship between operational databases and data warehouses? Why are data warehouses created, and how do organizations use them? What types of decisions do data warehouses support? Have you ever searched a data warehouse? Visit FedStats. gov and search "MapStats" to see what facts are available for your home state. Prepare a list of five interesting facts about your home state to share with your classmates.

4-27. Lisa Noriega has a problem with unstructured data. As her catering business grows, Lisa wants to analyze contracts to learn if over-budget projects result from using inexperienced project managers. Lisa wants to set up a database and she wants you to identify the records she will need. Work in a small group with classmates to identify the three entities that have meaning for her catering business. What are the attributes of these entities? What are probable data definitions of the attributes? What is the relationship between records and tables? What is the relationship between fields and attributes? Prepare a 5-minute presentation of your findings.

4-28. The Drexel Theatre is a small, family-owned cinema that screens independent and classic films. The lobby is decorated with vintage movie memorabilia including an original poster of Arnold Schwarzenegger, the Terminator, and his famous quote, "I'll be back." The theatre has a collection of 5,000 movies on DVD. It hires part-time workers for ticket and concession sales, as well as janitorial and projection services. It shows one of its movies every evening at 7:00 p.m. The owner of the Drexel plans to implement a relational database to handle operations. He has asked you to develop the data model for managing the film inventory. He wants to track movies, genres (categories), actors, and languages. He wants a description of each entity's attributes, and he wants an explanation of how to use primary keys and foreign keys to link the entities together. Work in a small group with classmates to plan the data model. Prepare a 5-minute presentation that includes an explanation of primary keys and foreign keys.

APPLICATION EXERCISES

4-29. EXCEL APPLICATION: Managing Catering Supplies

Lisa Noriega developed the spreadsheet shown in Figure 4-27 so that she can better manage her inventory of disposable catering supplies. Download the spreadsheet named Ch04Ex01 so you can help her with the inventory analysis.

Lisa listed her inventory items in "Case" quantities, but she now wants to analyze items according to "Pack" quantities and create a price list to show to her customers. For example, a case of Heavy Duty Deluxe Disposable Plastic Knives has 12 packs of 24 knives each. She wants to calculate a "Sales Price per Pack" based on her cost plus a 25% markup.

Lisa asks that you complete the following operations and answer the following questions.

- Create columns that list Case Pack, Packs on Hand, and Cost per Case Pack for each item. Use a formula to calculate the Cost per Case Pack.
- Create a column that lists Sales Price per Pack. Use a formula to calculate a 25% markup. Set up an assumption cell to input the percentage markup rather than include the markup value in the formula.
- Format the spreadsheet to make it easy to read and visually appealing.

1. What is Lisa's total investment in disposable catering supplies?
2. What is the total sales value of her inventory?
3. How much profit will she make if she sells all of her inventory at a 25% markup?
4. How much profit will she make if she uses a 35% markup instead?

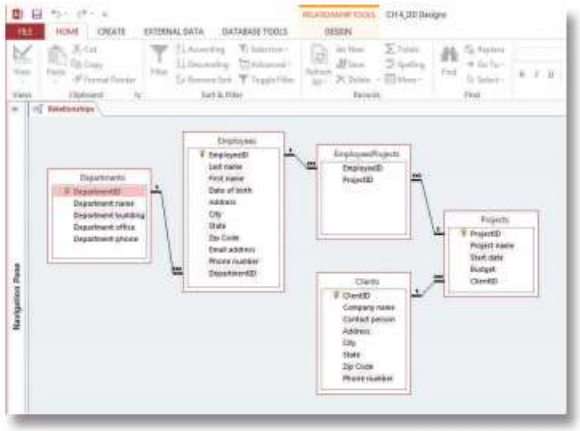
FIGURE 4-27
Catering supplies spreadsheet.

	A	B	C	D	E	F
1	Item No.	Description	Color	Unit	Unit Cost	On Hand
2	630K-C	Deluxe Heavy Duty Disposable Plastic Knives	Clear	Case (12 pk/24 each)	13.61	3
3	630F-C	Deluxe Heavy Duty Disposable Plastic Forks	Clear	Case (12 pk/24 each)	13.61	3
4	630S-C	Deluxe Heavy Duty Disposable Plastic Spoons	Clear	Case (12 pk/24 each)	13.61	3
5	630K-B	Deluxe Heavy Duty Disposable Plastic Knives	Black	Case (12 pk/24 each)	13.61	2
6	630F-B	Deluxe Heavy Duty Disposable Plastic Forks	Black	Case (12 pk/24 each)	13.61	2
7	630S-B	Deluxe Heavy Duty Disposable Plastic Spoons	Black	Case (12 pk/24 each)	13.61	2
8	5454W	54" × 54" Plastic Table Cover	White	Case (24)	19.15	1
9	5454B	54" × 54" Plastic Table Cover	Beige	Case (24)	19.15	2
10	5454BL	54" × 54" Plastic Table Cover	Black	Case (24)	19.15	1
11	549W	54" × 108" Plastic Table Cover	White	Case (24)	30.24	4
12	549B	54" × 108" Plastic Table Cover	Beige	Case (24)	30.24	2
13	549BL	54" × 108" Plastic Table Cover	Black	Case (24)	30.24	1
14	72W	72" Round Plastic Table Cover	White	Case (24)	52.08	3
15	72B	72" Round Plastic Table Cover	Beige	Case (24)	52.08	3
16	72BL	72" Round Plastic Table Cover	Black	Case (24)	52.08	1
17	537	13" × 17" Linen-Like Napkins	White	Case (6 pk/50 each)	45.90	6
18	28500	10" × 10" Linen-Like Napkins	White	Case (8 pk/50 each)	42.00	3
19	1010W	10" × 10" 2-ply Beverage Napkins	White	Case (12 pk/50 each)	18.14	6
20	1010B	10" × 10" 2-ply Beverage Napkins	Black	Case (12 pk/50 each)	18.14	1
21	1010R	10" × 10" 2-ply Beverage Napkins	Red	Case (12 pk/50 each)	18.14	1
22	1313W	13" × 13" 2-ply Luncheon Napkins	White	Case (12 pk/50 each)	25.20	3
23	1313B	13" × 13" 2-ply Luncheon Napkins	Black	Case (12 pk/50 each)	25.20	2
24	1010WED	10" × 10" Wedding Bells Print Napkins	White	Case (12 pk/50 each)	26.46	3
25	1313WED	13" × 13" Wedding Bells Print Napkins	White	Case (12 pk/50 each)	37.80	4
26	CC12	12 oz. Classic Crystal Tall Plastic Glasses	Clear	Case (12 pk/20 each)	78.33	1
27	CC16	16 oz. Classic Crystal Tall Plastic Glasses	Clear	Case (12 pk/20 each)	107.73	1
28	CCR9	9 oz. Classic Crystal Plastic Rocks Glasses	Clear	Case (12 pk/20 each)	53.90	3

4-30. ACCESS APPLICATION: DD-Designs

Devon Degosta set up an Access database to manage her web design business. She has asked you to create a report that summarizes and identifies projects that are assigned to more than one employee. Recreate the Access database with the table names, attributes, and relationships as illustrated in Figure 4-28. Download and use the information in the spreadsheet Ch04Ex02 to populate the tables. Create a report that lists each project by name and the names of the employees assigned to it. Devon wants the report to include the client name and the project budget. What other reports would Devon find useful?

FIGURE 4-28
DD_Designs database schema.



CASE STUDY #1

U.K. Police Track Suspicious Vehicles in Real Time with Cameras and the License Plate Database

Almost every city street in London is under constant video surveillance, partly as a reaction to terrorist attacks. These closed-circuit cameras initially created tapes that could be viewed later, but the technology now is far more capable. The cameras are equipped with automatic license plate recognition capabilities, which use optical character recognition to decipher the license plate numbers and letters in near real-time (Figure 4-29).

The camera's data is sent to the national ANPR Data Centre in north London, which also houses the Police National Computer. Cameras are widespread throughout the city, and many are mounted on police vehicles. Each camera can perform 100 million license plate reads per day. Each vehicle's plate number is combined with the camera's GPS location and a timestamp, so the Oracle database at the Data Center contains detailed information about the whereabouts of almost every vehicle.

Since the database is linked to the Police National Computer, police on the beat can query it to see whether a nearby vehicle is flagged for some reason. Cross-checking the license plate information against the crime database can turn up vehicles involved in crimes or registered to wanted criminals. In one case, a police constable was killed during a robbery, and police were able to track the getaway car because its license plate was read by the cameras. For cameras mounted on vehicles, the officer does not even need to send a query. An audio alert goes off when the camera's image matches a flagged license plate number, prompting the police to investigate.

Beyond criminal activity, the police database contains extensive information linked to the license plate data. For instance, a car might show that it is registered to someone who owes parking fines, or who is uninsured. The data might also show that the license plate is attached to the wrong vehicle, pointing to stolen plates.

The data are maintained for 5 years, creating a rich repository for data mining. One study found that certain cars triggered no flags, but seemed to be making impossibly quick journeys from one end of town to the other. Police discovered that car thieves were trying to outwit ANPR by "car cloning," in which the perpetrators duplicate a real license plate and attach it to a stolen car of the same make and model.

Law enforcement agencies see the license plate database, the cameras that feed it, and its integration with police data as a revolutionary advance, even though there are still gaps in coverage and the technology itself is not perfect. For example, rain, fog, and snow can interfere, and the plate itself might be blurred by mud. The plates themselves vary quite a bit, with different colors, fonts, and background images. Despite the drawbacks, police departments in the United States and other countries are rapidly adopting the system, buying camera-equipped cars, and developing smartphone access to databases.

Privacy advocates, however, are concerned about the mounting power of integrated databases and surveillance technologies to scrutinize human behavior. One judge remarked, "A person who knows all of another's travels can deduce whether he is a weekly churchgoer, a heavy drinker, a regular at the gym, an unfaithful husband, an outpatient receiving medical treatment, or an associate of a particular individual or political group." The United Kingdom is tightening regulations to provide better protections for citizens in an attempt to balance privacy concerns against the enormous value these databases offer to law enforcement.

FIGURE 4-29
Capturing license plate numbers for law enforcement.

Source: Ann Cantelow/Shutterstock.

Discussion Questions

4-31. Describe the manner in which data elements are linked across databases.

4-32. What technical and physical challenges does this information system face?

4-33. What human capital capabilities for law enforcement are necessary to make the database more effective?

4-34. What are the relevant considerations to balance the police's ability to investigate versus the citizens' need for privacy?

Sources: Crump, C. (March 19, 2013). ACLU in court today arguing that GPS tracking requires a warrant. ACLU.org, http://www.aclu.org/blog/technology-and-liberty/aclu-court-today-arguing-gps-tracking-requires-warrant, accessed March 24, 2013.
Du, S., Ibrahim, M., Shehata, M., & Badawy, W. (2013). Automatic license plate recognition (APLR): A state of the art review. *IEEE Transactions on Circuits & Systems for Video Technology* (Feb. 2013), *23*(2), 311–325. Retrieved from Business Source Premier, April 4, 2013.
Mathieson, S. A., & Evans, R. (August 27, 2012). Roadside cameras suffer from large gaps in coverage, police admin. *The Guardian*, http://www.guardian.co.uk/uk/2012/aug/27/police-number-plate-cameras-network-patchy, accessed March 24, 2013.
National Vehicle Tracking Database, http://wiki.openrightsgroup.org/wiki/National_Vehicle_Tracking_Database, accessed March 24, 2013.
Police in Jackson, MS, use Genetec license plate recognition technology. (March 14, 2013). Government Security News, http://www.gsnmagazine.com/node/28730?c=law_enforcement_first_responders, accessed March 24, 2013.

CASE STUDY #2

Colgate-Palmolive Draws on Its Global Database to Evaluate Marketing Strategies

With more than $17 billion in annual sales, Colgate-Palmolive's global operations span dozens of countries. The consumer products giant makes iconic brands such as Colgate toothpaste, Irish Spring soap, Palmolive dish detergent, and Softsoap shower gel and sells them around the world. Besides taking a "bite out of grime" with its soaps and personal hygiene products, the company also makes Science Diet pet foods.

Founded by William Colgate in 1806, the Manhattan-based company specialized in soap, candles, and starch. The "Palmolive" brand, featuring fragrant soaps made from palm and olive oils rather than foul-smelling animal fats, was added in mid-century through a merger. The company began expanding abroad by purchasing local soap and toothpaste companies in the 1930s, first in Europe, and then later in emerging economies in Asia and Latin America. In Latin America, for example, Colgate-Palmolive captured 79% of the market for oral care products after it acquired companies in Brazil and Argentina. More than 80% of its net sales now come from other countries, many in Latin America.

Managing this sprawling global empire requires a dedication to consistency, not just in the products themselves, but in the data that tracks every aspect of the company's operations and performance. Colgate's integrated back-end database and enterprise software, supplied by SAP, supports a consistent approach to master data management. CIO Tom Greene says, "With SAP, the product masters and the customer groupings are all driven by the same master data." With everyone using the same integrated system, Greene avoids the problem of redundant and inconsistent data entered into separate systems. Disputes about which is the correct "version of the truth" disappear.

Greene relies on this consistent back-end database for the Colgate Business Planning (CBP) initiative, which guides Colgate's investment decisions around the world. Marketing managers for consumer products confront a bewildering array of choices to promote products, from advertising campaigns and TV spots to discount coupons, rebates, and in-store displays. Most companies judge the success of such investments by measuring "uplift"—the difference between actual sales with the promotion and a projection of what sales might have been without the promotion. But CBP, combined with the integrated master database, allows Colgate management to dig far deeper, measuring actual profit, loss, and return on investment.

The detailed metrics can be broken down for individual products, regions, and retailers, providing a very clear window into how much any investment contributed to the company's profit. Corporate headquarters taps these finely tuned results to plan new investments. It is not a cookie cutter approach, however. Guided by their knowledge of local markets, subsidiary managers can tweak the plans to better fit local conditions. Since the results are all tallied consistently, drawing on the database, managers know what works and what doesn't.

Margins are critical in consumer products, so this deeper insight pays off. Thanks to CBP, Colgate reinvested $100 million in promotions found to be more profitable, and its long-term goal is $300 million—a sum that could be reinvested in promotions, or added to the company's bottom line. As Greene puts it, "You have to understand the technology, but the most important thing . . . is to understand the business so you can marry the two together."

Discussion Questions

4-35. What type of data does Colgate-Palmolive use, and what types of decisions does Colgate-Palmolive make based on the data?

4-36. Why is it important to Colgate-Palmolive for the data to be integrated across systems?

4-37. What business benefits does Colgate-Palmolive achieve through use of this data?

4-38. What will be more suitable for the architecture of this data warehouse, a relational database or data cubes?

Sources: Colgate-Palmolive Company. (March 24, 2013). Hoover's Company Records. Retrieved March 24, 2013 from Hoover's Online.
Colgate World of Care Website, www.colgatepalmolive.com, accessed August 26, 2013.
Henschen, D. (September 13, 2010). Data drives Colgate investment decisions. *Information Week*, *1278*, 38–39.
Maxfield, J. (February 27, 2013). What makes Colgate-Palmolive one of America's best companies. Motley Fool, http://www.fool.com/investing/general/2013/02/27/what-makes-colgate-palmolive-one-of-americas-best.aspx, accessed March 24, 2013.

E-PROJECT 1 Identifying Suspects with a License Plate Database: Constructing Queries with Access

An Access database from a hypothetical small island nation contains simulated license plate information and violation records, and it will illustrate how police are identifying cars involved in crimes or traffic offenses. Download the Access file called Ch04_Police to answer the following questions.

4-39. What are the three tables in the database? For simplicity, the LicensePlates table in this e-project uses LicensePlateNumber for its primary key. Why might that work for a small island nation, but not for the United States?

4-40. Why is PlateImagesID the primary key for the PlateImages table, rather than LicensePlateNumber?

4-41. A police officer spots a car illegally parked on a dark street, with license plate LCN5339. Query the database and list any crimes or other violations that are linked to this license plate.

4-42. A citizen reports a robbery to the police, but she can only remember the first three letters of the car's license plate (JKR). She thinks it was a black or dark blue Toyota. Which car is the best candidate, and who is the owner?

4-43. Letters such as G and C are often confused by eyewitnesses. Some witnesses to a hit-and-run accident reported that the license plate started with LGR, but they said they weren't sure. Construct a query to retrieve records that might match either LGR or LCR, and list the candidates.

4-44. The homicide division learned that a vehicle with a license plate number DYV4437 was observed near a murder scene, and they would like to speak to the owner, who might be able to shed light on the case. If the cameras have picked up the license plate at some time, it should be in the PlateImages table. Construct a query to retrieve the latitude and longitude of the car's most recent location.

E-PROJECT 2 Building a Database for Customer Records

In this e-project, you will construct a database of customer purchases for a small concession stand near "Four Corners," the point in the United States at which the Utah, Colorado, Arizona, and New Mexico state lines meet. Much of the data will be imported from Excel files.

4-45. Open Access and create a new database called FruitStand.

4-46 Create a table called Products with the following fields:

ProductID (The first field defaults to the name ID, as the table's primary key. Change the name to ProductID. Leave it as autonumber and as the primary key.)

ProductName (Text data type, field size 25 characters)

Price (Currency data type)

4-47 Enter the records in the following table. Note that you do not enter the ProductID; it is an autonumbering field that generates the next value. Save your work.

ProductID	ProductName	Price
1	apple	$0.45
2	pear	$0.70
3	watermelon	$2.75
4	grapefruit	$1.50
5	avocado	$1.25

4-48 Download the Excel file Ch04_FruitStand, and import the two worksheets, labeled Customers and Purchases. Identify the CustomerID as the primary key for Customers, and PurchaseNumber fields as the primary key for Purchases, rather than letting Access create its own primary keys.

a. What fields are contained in the Customers table? Generate a list of all your customers, sorted by CustomerID.
b. What fields are contained in the Purchases table? What are the foreign key(s) in the Purchases table, and which table(s) do they reference?

4-49 Use Access to Create Query (Query Design), join Customers to Purchases (on CustomerID) and Purchases to Products (on ProductID), and answer the following questions:

a. Create a query that returns all the purchases from customers from Nevada (NV). Which fruit do people from that state seem to prefer?
b. How many pears have been sold? (Click on Totals in the Design Ribbon to bring up options to report grouped totals. Your query should Group By ProductName. Include the Quantity field, and in the Total row, select Sum for Quantity.)
c. How many watermelons have been sold?
d. List all the states your customers come from, and the number of customers from each one. (Use COUNT under the CustomerID field from Customers.) From which state do most of your customers come?

4-50 List the countries your customers come from, sorting the data by CountryName. What problem do you encounter? What would you do to the database to improve your ability to analyze the data by country?