Chapter 3: Data Warehousing

Learning Objectives for Chapter 3

1. Understand the basic definitions and concepts of data warehouses
2. Understand data warehousing architectures
3. Describe the processes used in developing and managing data warehouses
4. Explain data warehousing operations
5. Explain the role of data warehouses in decision support
6. Explain data integration and the extraction, transformation, and load (ETL) processes
7. Describe real-time (active) data warehousing
8. Understand data warehouse administration and security issues

CHAPTER OVERVIEW
The concept of data warehousing has been around since the late 1980s. This chapter provides the foundation for an important type of database, called a data warehouse, which is primarily used for decision support and provides improved analytical capabilities. We discuss data warehousing in the following sections:

CHAPTER OUTLINE

3.1 OPENING VIGNETTE: ISLE OF CAPRI CASINOS IS WINNING WITH ENTERPRISE DATA WAREHOUSE

   Questions for the Opening Vignette

A. WHAT WE CAN LEARN FROM THIS VIGNETTE
### 3.2 DATA WAREHOUSING DEFINITIONS AND CONCEPTS

A. **WHAT IS A DATA WAREHOUSE?**

B. **A HISTORICAL PERSPECTIVE TO DATA WAREHOUSING**

C. **CHARACTERISTICS OF DATA WAREHOUSING**

D. **DATA MARTS**

E. **OPERATIONAL DATA STORES**

F. **ENTERPRISE DATA WAREHOUSES (EDW)**
   - Application Case 3.1: A Better Data Plan: Well-Established TELCOs Leverage Data Warehousing and Analytics to Stay on Top in a Competitive Industry

G. **METADATA**
   - Section 3.2 Review Questions

### 3.3 DATA WAREHOUSING PROCESS OVERVIEW

- Application Case 3.2: Data Warehousing Helps MultiCare Save More Lives
- Section 3.3 Review Questions

### 3.4 DATA WAREHOUSING ARCHITECTURES

A. **ALTERNATIVE DATA WAREHOUSING ARCHITECTURES**

B. **WHICH ARCHITECTURE IS THE BEST?**
   - Section 3.4 Review Questions

### 3.5 DATA INTEGRATION AND THE EXTRACTION, TRANSFORMATION, AND LOAD (ETL) PROCESSES

A. **DATA INTEGRATION**
   - Application Case 3.3: BP Lubricants Achieves BIGS Success

B. **EXTRACTION, TRANSFORMATION, AND LOAD**
   - Section 3.5 Review Questions

### 3.6 DATA WAREHOUSE DEVELOPMENT

- Application Case 3.4: Things Go Better with Coke’s Data Warehouse

A. **DATA WAREHOUSE DEVELOPMENT APPROACHES**
   - The Inmon Model: The EDW Approach
   - The Kimball Model: The Data Mart Approach
   - Which Model Is Best?
     - Application Case 3.5: Starwood Hotels & Resorts Manages Hotel Profitability with Data Warehousing

B. **ADDITIONAL DATA WAREHOUSE DEVELOPMENT CONSIDERATIONS**
   - Technology Insights 3.1: Hosted Data Warehouses

C. **REPRESENTATION OF DATA IN DATA WAREHOUSE**

D. **ANALYSIS OF DATA IN THE DATA WAREHOUSE**

E. **OLAP VERSUS OLTP**

F. **OLAP OPERATIONS**
1. Variations of OLAP
   ♦ Technology Insights 3.2: Hands-On Data Warehousing with MicroStrategy
   ♦ Section 3.6 Review Questions

3.7 DATA WAREHOUSING IMPLEMENTATION ISSUES
   ♦ Application Case 3.6: EDW Helps Connect State Agencies in Michigan
   A. MASSIVE DATA WAREHOUSES AND SCALABILITY
      ♦ Section 3.7 Review Questions

3.8 REAL-TIME DATA WAREHOUSING
   ♦ Application Case 3.7: Egg Plc Fries the Competition in Near Real Time
   ♦ Technology Insights 3.3: The Real-Time Realities of Active Data Warehousing
   ♦ Section 3.8 Review Questions

3.9 DATA WAREHOUSE ADMINISTRATION, SECURITY ISSUES, AND FUTURE TRENDS
   ♦ Technology Insights 3.4: Ambeo Delivers Proven Data-Access Auditing Solution
   A. THE FUTURE OF DATA WAREHOUSING
      ♦ Section 3.9 Review Questions

3.10 RESOURCES, LINKS, AND THE TERADATA UNIVERSITY NETWORK CONNECTION
   A. RESOURCES AND LINKS
   B. CASES
   C. VENDORS, PRODUCTS, AND DEMOS
   D. PERIODICALS
   E. ADDITIONAL REFERENCES
   F. THE TERADATA UNIVERSITY NETWORK (TUN) CONNECTION

Chapter Highlights
Key Terms
Questions for Discussion
Exercises
Teradata University and Other Hands-On Exercises
Team Assignments and Role-Playing Projects
Internet Exercises
  ♦ End of Chapter Application Case: Continental Airlines Flies High with Its Real-Time Data Warehouse
  ♦ Questions for the Case

References
TEACHING TIPS/ADDITIONAL INFORMATION

Expect to spend some time on this chapter, as understanding its content is crucial and many of the concepts are not intuitive to students whose major computer exposure has been at the personal level. You should prepare yourself with some technical examples of how things work, such as screen shots from some of the tools mentioned in the chapter, as most of the technical discussion here is at a more conceptual level. The business discussion is tangible, but the technical part is less so. It will be up to you to connect it to the students.

Section 3.2 defines several fundamental concepts. Students must understand that these are not definitions for the sake of definitions (every textbook has some of those, this one is no exception) but are important to anyone working in the field. The characteristics of a data warehouse in its extended definition, for example, are given not because some researcher thought they’d be theoretically nice, but, more specifically, because they’ve turned out to be important in practice. So, it’s important to really know what each of them is about. (The concept of an operational data store should already be familiar to students, though perhaps not by that name. They should know what metadata are from their use with general databases, too.)

Section 3.3 is brief. It can be covered quickly. It’s important to understand what each step does, but the details of things like ETL will be covered later in this chapter.

Section 3.4, by contrast, is seven pages long—though a lot of that is diagrams. The key point here is that there are a lot of options; the correct one can be chosen only when something is known about the characteristics of the application, and in most cases more than one will work reasonably well.

Section 3.5 discusses ETL, a process that never shows up outside a data warehouse context and that students may therefore never have heard of before this course. The need for ETL arises because databases aren’t all in the same format; chances are that few, if any, of the source databases are in the format of the DW—and even if the formats match, most databases contain dirty data. This is a good place to show screen shots of ETL tools so students will really appreciate how they work.

Section 3.6, another long one, is on the development process. Students will have studied the development and implementation of information systems in general in their introductory MIS course and probably gone deeper into it in Systems Analysis and other courses. Here, it helps to focus on what’s different about data warehouses. For example, while training is needed for both conventional systems and data warehouses, the training needed for a DW is at a higher level: more information literacy than computer literacy. A systems analyst who is used to the correct type of training for ERP or SCM systems, and who attempts to replicate that experience for a DW user community, is courting disaster.

This section covers more than the database development suggested by its title. It also discusses the internal structure of a data warehouse database. It first focuses on the star schema and then describes multidimensional analytical queries. Students who’ve studied database management—likely to be most of the class in a senior-level business school course for MIS concentrators—will want tangible information here. You can provide this with examples showing how a simple set of historical sales data with several dimensions could be stored in an operational database and in a data warehouse organized by the star schema.
If you have provided your students with a good foundation up to this point, Sections 3.7 and 3.8 should go quickly. Section 3.7 raises some issues to prepare for in implementing a DW, including the importance of scalability. In Section 3.8, the business case for real-time DWs is well presented, and the technology introduces no new conceptual difficulties.

Section 3.9 returns to issues that students will be familiar with from a general database management context. The need for “establishing effective corporate and security policies and procedures,” for instance, is not unique to data warehouses. Therefore, you may, again, want to focus on what’s different about them. Answers to Question for Discussion 5 at the end of the chapter can lead to a good class discussion of this.

ANSWERS TO END OF SECTION REVIEW QUESTIONS

Section 3.1 Review Questions

1. Why is it important for Isle to have an EDW?

In the gaming industry, companies distinguish themselves based on their customer relationships. An enterprise data warehouse (EDW) gathers and provides the data needed to tell Isle of Capri what customers respond to, so the casinos can adapt their offerings. The information provided by the EDW lets Isle deepen its understanding of customers, so it can efficiently give them more of the kinds of entertainment they are looking for.

2. What were the business challenges or opportunities that Isle was facing?

Isle of Capri Casinos is one of the largest publicly traded gaming companies in the United States, but the gaming business is competitive. Other casinos offer essentially the same games, so Isle must find ways to make its entertainment and hospitality atmosphere one that exceeds customer expectations. Before Isle implemented the EDW, casino managers had to wait to review monthly data until the second week of the following month. The time lag made it difficult for casinos to identify what actions were appealing to customers in time to respond. Adding to these business challenges, Hurricane Katrina set back initial efforts to set up a data warehouse at the southeastern company.

3. What was the process Isle followed to realize EDW? Comment on the potential challenges Isle might have had going through the process of EDW development.

Isle of Capri brought in two expert suppliers. Teradata provided the core solution; IBM Cognos provided expertise in business intelligence. Isle hired a management team that understood how EDW could support decision making at Isle. That team would be able to help Isle’s managers with the challenge of understanding how they can frame queries and follow-up questions to figure out ways to improve the business. Most likely, the potential of using detailed, up-to-the-minute data would be unfamiliar to many of these managers.

4. What were the benefits of implementing an EDW at Isle? Can you think of other potential benefits that were not listed in the case?
The implementation of EDW brought several benefits related to the timeliness and detail of the data that became available. Instead of five week-old monthly reports, managers can now study a variety of daily, weekly, and monthly reports. The reports segment data by particular properties and customer groups, so managers can zero in on particular problems and successes, easily making comparisons among properties. Managers can submit queries about data sets and receive information within minutes. In this way, managers can find out how particular promotions are affecting customer behavior at particular casinos. The EDW also connects data about casino activity with data about customers’ use of hotels and efforts by Isle’s hosts. This, too, helps the company target promotions and offer customers incentives they value. Even decisions as detailed as where to locate slot machines can be adjusted to boost profits based on data from the EDW.

5. Why do you think large enterprises like Isle in the gaming industry can succeed without having a capable data warehouse/business intelligence infrastructure?

In the past, businesses in the gaming industry could succeed without a capable data warehouse/business intelligence infrastructure because their managers knew as much about customers as their competitors’ management knew. They were all testing ideas for promotion or entertainment and responding to customer behavior at about the same pace. However, when one company such as Isle begins to respond to daily, property-by-property data, choosing the marketing and entertainment options that deliver exactly what most profitably lures customers, competitors without that capability will soon begin to suffer.

Section 3.2 Review Questions

1. What is a data warehouse?

A data warehouse is defined in this section as “a pool of data produced to support decision making.” This focuses on the essentials, leaving out characteristics that may vary from one DW to another but are not essential to the basic concept.

The same paragraph gives another definition: “a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management’s decision-making process.” This definition adds more specifics, but in every case appropriately: it is hard, if not impossible, to conceive of a data warehouse that would not be subject-oriented, integrated, etc.

2. How does a data warehouse differ from a database?

Technically a data warehouse is a database, albeit with certain characteristics to facilitate its role in decision support. Specifically, however, it is (see previous question) an “integrated, time-variant, nonvolatile, subject-oriented repository of detail and summary
data used for decision support and business analytics within an organization.” These characteristics, which are discussed further in the section just after the definition, are not necessarily true of databases in general—though each could apply individually to a given one.

As a practical matter most databases are highly normalized, in part to avoid update anomalies. Data warehouses are highly denormalized for performance reasons. This is acceptable because their content is never updated, just added to. Historical data are static.

3. What is an ODS?

Operational Data Store is the database from which a business operates on an on-going basis.

4. Differentiate among a data mart, an ODS, and an EDW.

An ODS (Operational Data Store) is the database from which a business operates on an ongoing basis.

Both an EDW and a data mart are data warehouses. An EDW (Enterprise Data Warehouse) is an all-encompassing DW that covers all subject areas of interest to the entire organization. A data mart is a smaller DW designed around one problem, organizational function, topic, or other suitable focus area.

5. Explain the importance of metadata.

Metadata, “data about data,” are the means through which applications and users access the content of a data warehouse, through which its security is managed, and through which organizational management manages, in the true sense of the word, its information assets. Most database management systems would be unable to function without at least some metadata. Indeed, the use of metadata, which enable data access through names and logical relationships rather than physical locations, is fundamental to the very concept of a DBMS.

Metadata are essential to any database, not just a data warehouse. (See answer to Review Question 2 of this section above.)

Section 3.3 Review Questions

1. Describe the data warehousing process.

The data warehousing process consists of the following steps:
1. Data are imported from various internal and external sources
2. Data are cleansed and organized consistently with the organization’s needs
3. a. Data are loaded into the enterprise data warehouse, or
2. Describe the major components of a data warehouse.

- **Data sources.** Data are sourced from operational systems and possibly from external data sources.
- **Data extraction and transformation.** Data are extracted and properly transformed using custom-written or commercial software called ETL.
- **Data loading.** Data are loaded into a staging area, where they are transformed and cleansed. The data are then ready to load into the data warehouse.
- **Comprehensive database.** This is the EDW that supports decision analysis by providing relevant summarized and detailed information.
- **Metadata.** Metadata are maintained for access by IT personnel and users. Metadata include rules for organizing data summaries that are easy to index and search.
- **Middleware tools.** Middleware tools enable access to the data warehouse from a variety of front-end applications.

3. Identify and discuss the role of middleware tools.

Middleware tools enable access to the data warehouse. Power users such as analysts may write their own SQL queries. Others may access data through a managed query environment. There are many front-end applications that business users can use to interact with data stored in the data repositories, including data mining, OLAP, reporting tools, and data visualization tools. All these have their own data access requirements. Those may not match with how a given data warehouse must be accessed. Middleware translates between the two.

Section 3.4 Review Questions

1. What are the key similarities and differences between a two-tiered architecture and a three-tiered architecture?

Both provide the same user visibility through a client system that accesses a DSS/BI application remotely. The difference is behind the scenes and is invisible to the user: in a two-tiered architecture, the application and data warehouse reside on the same machine; in a three-tiered architecture, they are on separate machines.

2. How has the Web influenced data warehouse design?

Primarily by making Web-based data warehousing possible.
3. List the alternative data warehousing architectures discussed in this section.

- Independent data marts architecture
- Data mart bus architecture with linked dimensional data marts
- Hub-and-spoke architecture (corporate information factory)
- Centralized data warehouse architecture
- Federated architecture

4. What issues should be considered when deciding which architecture to use in developing a data warehouse? List the 10 most important factors.

1. Information interdependence between organizational units
2. Upper management’s information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

(This list from the text, while clearly intended by the authors as the answer to this review question, does not explicitly say that these are the ten most important factors. Students may choose others.)

5. Which data warehousing architecture is the best? Why?

See Table 3.1 Average Assessment Scores for the Success of the Architectures. What is interesting is the similarity of the averages for the bus, hub-and-spoke, and centralized architectures. The differences are sufficiently small that no claims can be made for a particular architecture’s superiority over the others, at least based on a simple comparison of these success measures.

Section 3.5 Review Questions

1. Describe data integration.

Data integration is an umbrella term that covers three processes that combine to move data from multiple sources into a data warehouse: accessing the data, combining different views of the data, and capturing changes to the data.

2. Describe the three steps of the ETL process.

*Extraction:* selecting data from one or more sources and reading the selected data.
**Transformation**: converting data from their original form to whatever form the DW needs. This step often also includes cleansing of the data to remove as many errors as possible.

**Load**: putting the converted (transformed) data into the DW.

3. **Why is the ETL process so important for data warehousing efforts?**

Since ETL is the process through which data are loaded into a data warehouse, a DW could not exist without it. The ETL process also contributes to the quality of the data in a DW.

**Section 3.6 Review Questions**

1. List the benefits of data warehouses.

   **Direct** benefits include:
   - Allowing end users to perform extensive analysis in numerous ways.
   - A consolidated view of corporate data (i.e., a single version of the truth).
   - Better and more timely information. A data warehouse permits information processing to be offloaded from costly operational systems onto low-cost servers; therefore, end-user information requests can be processed more quickly.
   - Enhanced system performance. A data warehouse frees production processing because some operational system reporting requirements are moved to DSS.
   - Simplification of data access.

   **Indirect** benefits arise when end users take advantage of these direct benefits.

2. List several criteria for selecting a data warehouse vendor, and describe why they are important.

   Six important criteria are: financial strength, ERP linkages, qualified consultants, market share, industry experience, and established partnerships. These are important to indicate that a vendor is likely to be in business for the long term, to have the support capabilities its customers need, and to provide products that interoperate with other products the potential user has or may obtain.

   One could add others, such as product functionality (Does it do what we need?), vendor strategic vision (Does their direction make sense for our future plans and/or is it consistent with industry trends?) and quality of customer references (What do their existing customers think of them?).