

CHAPTER TWO

The Data Warehouse

From *Business Intelligence: Making Decisions through Data Analytics*
By Jerzy Surma (A Business Expert Press Book)



© 2011 by Business Expert Press. All rights reserved.

Harvard Business Publishing distributes in digital form the individual chapters from a wide selection of books on business from publishers including Harvard Business Press and numerous other companies. To order copies or request permission to reproduce materials, call 1-800-545-7685 or go to <http://www.hbsp.harvard.edu>. No part of this publication may be reproduced, stored in a retrieval system, used in a spreadsheet, or transmitted in any form or by any means – electronic, mechanical, photocopying, recording, or otherwise – without the permission of Harvard Business Publishing, which is an affiliate of Harvard Business School.

This document is authorized for use only by Ching Yang at University of Wales Trinity Saint David until April 2014.

Copying or posting is an infringement of copyright. Permissions@hbsp.harvard.edu or 617.783.7860.

Do Not Copy or Post

CHAPTER 2

The Data Warehouse

2.1. Introduction

In the first chapter, business intelligence (BI) was defined and a review of its possible applications in selected business areas was presented. The background for those solutions consists of basic BI tools. However, before we look at these in depth, we should look at the technological aspect of BI systems, that is, the structure of data warehouses. It seems obvious that the previous examples of business solutions use various data. However, this chapter answers the question of how data are collected and organized in order to be applied to business analysis.

Figure 2.1 shows the standard structure of a data warehouse. On the left, there are examples of sources of data—for example, logistics data in a transactional system. Within the sources of data, operational data are registered; this is data that is connected with isolated and specific business events. A possible use of a logistics system might be, for instance, to provide information that a given product is in stock. These data are sent to a data warehouse and adequately processed and connected with other data there. The information is read from the data warehouse (see the right part of Figure 2.1) to conduct certain business analyses. For example, information about the delivery and shipment of goods throughout a whole month might be used, within online analytical processing (OLAP), to detect those groups of products that fill the shelves. If these data are connected with information about the storage costs of different products, taken from a controlling system, then a complete analysis of the costs of storage of all goods that fill the shelves is possible.

It is essential to understand why data warehouses are built. There are at least four of them. Thanks to data warehouses you can

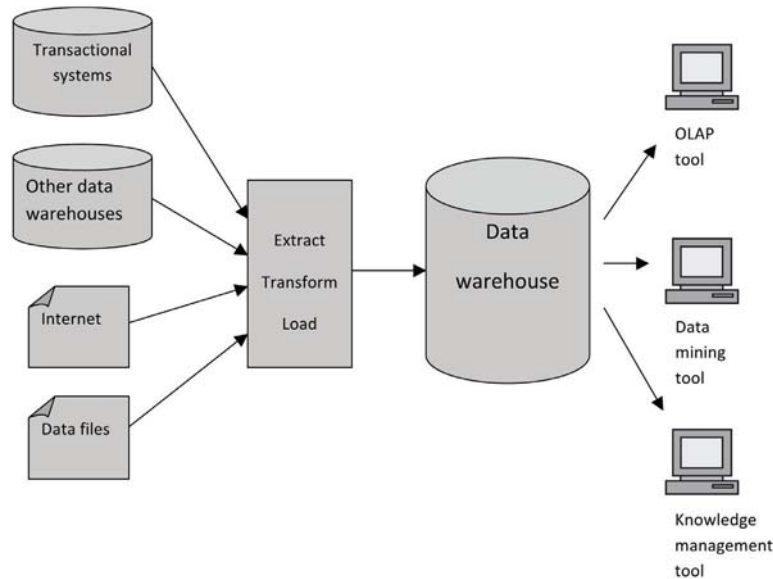


Figure 2.1. An example of the structure of a data warehouse.

Source: author.

- conduct business analyses without any interference in online transaction processing (OLTP).¹ Business analyses usually require complex and time-consuming calculations. However, this should not influence the daily operational work of systems that contain the source data. And thus, for instance, the sales analysis based on the aggregation of data from invoices registered in a transaction system ought not to cause delays in issuing invoices.
- obtain a global picture of a company's data—that is, you can integrate data from different sources, which gives a thorough picture of the business events in a given enterprise. For a telecommunications company, it could be information about the total sales value of a given client on the basis not only of the number of phone calls registered in the billing system but of the value of the accessories purchased by this client in brand stores (these data are registered in the sales system).
- gain access to historical data. Past information can be stored and then accessed later. Thanks to historical data, it is possible to carry out most basic business analyses.

- standardize information. The proper implementation of a data warehouse system enforces the standardization of terms and makes different departments of a company interpret performance indicators (measures) identically and calculate them in a homogenous manner. Consequently, the problem of many versions of the truth depending on the author of a given report is eliminated.

2.2. What a Data Warehouse Is

Bill Inmon is an American computer scientist and is recognized by many as the father of the data warehouse concept. His definition of a data warehouse is widely accepted.² A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decisions.

"Subject oriented" means that the collected data are related to a specific business area—for example, a sales and marketing department. However, in practice, the data warehouse can include different business areas. The data warehouse is restricted to one business area or even to one department, usually called a data mart, which is a local warehouse that is generally a part of a corporate warehouse.

"Integrated" refers to the fact that all data are homogenous in their format, there is a range of acceptable values, and data are provided from most or all of an organization's operational system.

"Time variant" means that all data entered into a data warehouse have a timestamp, which allows companies to track changes and perform analytical research at certain time intervals.

"Nonvolatile" signifies that data, once entered into a warehouse, are remembered and cannot be deleted. After downloading, one can only read the data.

Technologically, data warehouses are operated through database management systems; usually these are databases using a relational model.³ Inmon amended his classic definition of the data warehouse to include a corporate information factory (CIF).⁴ This is a data warehouse that contains all the information resources of an enterprise and the data coming from the external world.

2.3. Logical Model: Star Schema

The way of modeling data in the data warehouse corresponds to the need for reporting of business data. When one makes a report, business measures are analyzed, such as sales quantity, in various contexts (i.e., dimensions), such as products, customers, and time. As Edgar F. Codd, a British computer scientist who invented the relational model for database management, observed, “in principle, there are a great number of different dimensions, which can be used to analyze a certain set of data. This complex perspective that is a multi-faceted, notional picture seems to be the way in which most business people perceive their enterprise.”⁵

Table 2.1 shows a sales report of personal computers (PCs) by a computer hardware distribution company. In this report such measures as sales amount and costs for the product (e.g., purchase costs and logistics costs) are analyzed by looking at the dimension time (May 2008) and product (PC ABC, PC XYZ, and PC QWE).

In order to determine the data that are necessary to generate such a report, three tables must be defined:

1. A fact table (sales and costs; see Table 2.2)⁶
2. A dimension table (product; see Table 2.3)
3. A dimension table (date; see Table 2.4)⁷

A data warehouse model is able to link the separate data on sales, product, and date and then generate the report presented in Table 2.1. It should be pointed out that not all the data from the three tables were

Table 2.1. Demonstration of a Sales Report for a Computer Hardware Distributor, May 2008

Product	Sales amount (\$US)	Costs (\$US)	Profitability (%)
PC ABC	1,000,000	700,000	30
PC XYZ	2,000,000	1,800,000	10
PC QWE	2,000,000	1,500,000	25
Total	5,000,000	4,000,000	20

Note: Profitability (in %) = (qualitative sales – costs total) ÷ qualitative sales.

Source: author.

Table 2.2. A Demonstration Fact Table: Sales

ID_Product	ID_Date	Sales_Quantity	Price	Costs
P1	C5	500	2000	700,000
P2	C5	400	5000	1,800,000
P3	C5	1000	2000	1,500,000
P1	C6	400	2500	8,000,000

Source: author.

Table 2.3. A Demonstration Table for a Dimension: Product

ID_Product	Name	Producer	Processor
P1	PC ABC	Beta	Gamma
P2	PC XYZ	Delta	Gamma
P3	PC QWE	Beta	Theta

Source: author.

Table 2.4. A Demonstration Table for Dimension: Date (Time)

ID_Date	Month	Quarter	Year
C4	April	Q2	2008
C5	May	Q2	2008
C6	June	Q2	2008

Source: author.

used to generate the report and that the calculation of the sales amount requires reference to the sales table (Sales_Quantity and Price). Every line in every table has an unambiguous identifier: the primary key (PK).⁸ In the product table, it is ID_Product; in the date table, it is ID_Date; and in the sales table, the pair ID_Product and ID_Date is the composite key. The sales table is connected with the product table by ID_Product. ID_Product in the sales table is then the so-called foreign key (FK).⁹ The foreign key is a referential constraint between two tables. The foreign key identifies a column or a set of columns in one (referencing) table that refers to a set of columns in another (referenced) table. And analogically the sales table is connected with the date table by ID_Date; ID_Date is also a foreign key in the sales table. It is worth stressing that to one line (row) in the product table, zero, one, or many rows (records) in the sales

table may correspond, while one row in the sales table may correspond exactly to one row in the product table. The sales table and the date table share exactly the same relationship. This type of relationship, described as $1:N$ (one to many) is the basis of modeling data in a star schema (see Figure 2.2). Every rectangle corresponds to one table with data.

The schema presented in Figure 2.2 is called a multidimensional model, or popularly a star schema, because there are tables with dimensions connected by a relationship $1:N$ in the fact table. In formal terms, the schema from Figure 2.2 should be described as follows:

- Sales (**ID_Product**, **ID_Date**, Sales_Quantity, Price, Cost)
- Product (**ID_Product**, Name, Producer, Processor)
- Date (**ID_Date**, Month, Quarter, Year)

Sales, product, and date are called entity sets. A single entity in a set presented in the table is simply a row in this table. Expressions in brackets are descriptions of attributes that are characteristic of a given set of entities. The bold selected attributes represent a key (an identifier) entity in the entity set. Owing to this convention, the schema in Figure 2.2 is called an entity relationship schema, and it is a classical way of modeling databases. In fact, having determined the format of data for specific attributes, one can already generate certain data structures in a selected database management system. The model in the form of a schema of entity sets is called a logical (conceptual) model, and its realization in a specific database is a physical model.

2.4. Sources of Data

Returning to the example of the report from Table 2.1, it is vital to understand how the fact table (i.e., Table 2.2) and other tables were filled



Figure 2.2. A demonstration dimensional model (star model).

Source: author.

in with data. For instance, in order to calculate the sales quantity for a given product, one must

1. check which transactional data system the source data (i.e., the invoices or orders, on the basis of which one can specify the quantity of sales) are stored in.
2. determine the interface between the system with the source data and the data warehouse and enter suitable data. Information from the given day should be added to the previous data to extend the history of sales by another day.
3. aggregate the data from the invoices, which are single transactional business events, to show the results for a whole month; take, for instance, the first row (Sales_Quantity in May 2008 for PC ABC) in Table 2.2. To calculate profitability, you must integrate sales data with the data related to costs, which are downloaded from a financial controlling system.

This simplified scenario shows the key issues connected with designing and using data warehouses. In general, there are the following sources of data:

- *Data files.* These could be both textual and electronic (e.g., Excel files).
- *Transactional systems.*¹⁰ These are systems that, by making repeatable transactions, support the basic business processes of a company. The most popular among them are information management systems such as the following:
 - Enterprise resource planning (ERP), which supports companies in terms of finance, sales and logistics, warehouse management, production, procurement, human resource management, and so on
 - Banking systems for handling customers' accounts
 - Billing systems in telecommunications and energy companies
- *Internet data.* These could be both data taken from Internet pages and, for example, data mined by linking to Internet services that automatically transfer suitable information in the form of a subscription
- *Other data warehouses.*

It should be stressed that the sources of data may be located inside (and those are usually the most important) and also outside companies. The next chapter looks at where this information may come from.

Data from data sources are entered into a warehouse. This seemingly straightforward process usually consists of three phrases—extract-transform-load (ETL)—that are chronologically linked:

1. *Extraction* of data from sources of data. The identification of the sources of data is a complex issue, and you have to specify precisely the information that should be taken to obtain the suitable data (facts and dimensions). You have to refer to unambiguous business terms and connect them with specific data in the system. In some situations, it may turn out that not all the necessary data are registered. As a result, many times you are forced to prepare the source yourself, which is a very costly additional task.
2. *Transformation* of data to the form that allows for appropriate business application. This problem is related to, among other things, the integration of data:¹¹
 - *Integration of formats*—that is, unification of formats of data coming from different systems. For instance, gender can be recorded in a number of different ways: m/f, M/F, male/female, and so on.
 - *Semantic integration*—that is, a uniform interpretation of data. The professional analysis of business requirements allows you to clarify, for instance, whether the sales are calculated on the basis of invoices or orders. If the sales are calculated on the basis of orders, it allows you to clarify whether this means orders received, carried out, and so on. If you are not precise at this stage and you do not work out an unambiguous definition of the term “sales,” the problem of different versions of truth may arise. This applies also to tables with dimensions, particularly if a source system is not integrated and thus does not have standardized master data. Usually, the problem of uniform and unambiguous definitions for all terms used in warehouse is related to the issue of representation of metadata.

Moreover, this also includes any additional activities, including, among others, the verification of the quality of the data and

its preliminary aggregation before its input into the warehouse. The verification of the quality of the data consists of checking if a given date meets previously defined criteria (e.g., you can check whether personal details, including the date of birth of employees, fit in a rational range of values). The preliminary aggregation of data is, for instance, the input not of single transactions from the store's cash registers (i.e., receipts) but of summaries of the sales of each product for the whole day.

3. *Loading* of data into a warehouse. Normally, the loading of data into the warehouse is not a single action but a repeated process, in which data are added to the warehouse at strictly defined time intervals, depending on business needs: once per day, once per month, and so on. Consequently, the warehouse contains the history of business events that enables you to scrupulously analyze their changes over time.

This information demonstrates the importance of properly administering a data warehouse. To sum up, you should pay special attention to the significant difference between source data and data stored into a warehouse. The first type of data concerns operational processing in day-to-day activities in an enterprise, whereas the second one is related to management information, often at a strategic level, which enables you to gain insights into the company's performance from a number of angles.

2.5. Designing a Data Warehouse: Basic Facts

A data warehouse should be precisely designed to satisfy specific business needs. Usually, business requirements are specified by determining what information is necessary to manage a given business area. This allows you to select an appropriate logical model of a warehouse (see the star model) and then the necessary source data. This top-down approach was formulated by Ralph Kimball as a dimensional design model.¹² Kimball is widely recognized as one of the first architects of data warehouse logical models. According to this paradigm, there are four steps to design a data warehouse:

1. *Specify business requirements.* To understand business aims and needs in the context of existing source data, one must define the most important business issues in a given area. See, for instance, Table 2.1: It could be a decision to continue or discontinue selling a product on the basis of its profitability and how this changes over time.
2. *Declare the granularity.* Specify how detailed the data should be in the multidimensional model in the fact table. The most common is the lowest level of granularity, which is not subject to any further division. It definitely gives the greatest analytical flexibility, but it also may require more external memory than others. In the case of the previously mentioned computer hardware distributor, it would mean loading the data into the data warehouse for all the individual items from all invoices and all costs ascribed to the sale of a given product.
3. *Choose the dimensions.* Specifying the granularity determines the fact tables. In the presented example, it is the product dimension (defined as a concrete product; see Table 2.3) and the date dimension (defined as a concrete month; see Table 2.4).
4. *Identify the facts.* Specify precisely which business measures should be included in the fact table. Measures are quantitative, and various mathematical operations can be performed on them (e.g., the highest value, ordering, mean, sum). In the presented example, it is Sales_Quantity, Price, and Costs (see Table 2.2). The calculated measure—that is, Sales amount and Profitability (see the report in Table 2.1)—appears in the report generated for the user by a special tool for reporting as the result of a certain operation on the measures from the fact table.

The result of this four-step process of designing the multidimensional model is a star schema, such as the one in Figure 2.2. On the basis of this schema—that is, to be precise, on the basis of the definitions of sets of entities along with the specifications of data formats for attributes—you can easily create a proper data structure in database languages (e.g., structured query language, or SQL).

The example of modeling a data warehouse will be discussed in Case Study 2.1, and the basics of the business analysis will be presented in the next chapter.

Case Study 2.1.

The implementation of the BI system began with the analysis of business requirements. In the area of sales, it was as follows:

- *Area.* Sales.
- *Provided information or analysis:* Quantity (volume) and amount of sales analysis divided into time period, sales region, product category, and so on

According to guidelines on modeling a multidimensional model, the following activities were conducted:

1. Specify Business Requirements

To analyze business requirements, a sales report was produced providing the amount in percentage terms for selected product groups (in the example, four product groups) in a given period (in the example, January 2007) in stores in specific locations (in the example, stores in Boston). The sales report is shown in Table 2.5 and 2.6.

Table 2.5. Report: sales amount in January 2007

	Store name			
	ALFA_10	ALFA_11	ALFA_12	Total
Category	Sales amount	Sales amount	Sales amount	Sales amount
Juice	1470.05	1471.58	1820.68	4762.31
Dairy products	4432.69	4630.44	5073.86	14136.99
Confectionary	5981.00	6103.04	7558.79	19642.83
Cleaning products	929.52	1276.85	1259.43	3465.80
Total	12813.26	13481.91	15712.76	42007.93

Table 2.6. Report: sales amount (in percentage) in January 2007

	Store name			
	ALFA_10	ALFA_11	ALFA_12	Total
Category	Sales amount (%)	Sales amount (%)	Sales amount (%)	Sales amount (%)
Juice	3.50	3.50	4.33	11.34
Dairy products	10.55	11.02	12.08	33.65
Confectionary	14.24	14.53	17.99	46.76
Cleaning products	2.21	3.04	3.00	8.25
Total	30.50	32.09	37.40	100.00

2. Declare the Grain

The following level of granularity was defined:

- Date: 1 day
- Product: product item (SKU, or stock-keeping unit)
- Store: a single store

3. Choose the Dimensions

The following dimensions were defined:¹³

- Date (see Table 2.7)

Table 2.7. Dimension Date

Date
IDDate (PK)
Date
Full Date Description
Day of Week
Calendar Month
Calendar Quarter
Calendar Year

- Product (see Table 2.8)

Table 2.8. Dimension Product

Product
IDProduct (PK)
Product Name
SKU Number
Brand
Category
Package Type
Weight
Weight Units

In Table 2.9 are presented demonstration data for Product dimension.

Table 2.9. Product: demonstration data

Source: author.

IDProduct	Product name	SKU number	Brand	Category	Package type	Weight	Weight units
1	Chocolate bar Hit	1480	Chocolate bar	Confectionary	Elastic mat.	55	Grams
2	Yogurt Silk	3716	Yogurt	Dairy products	Elastic mat.	400	Grams
3	Buttermilk Joy	1600	Buttermilk	Dairy products	Elastic mat.	500	Grams
4	Chocolate Noir	3161	Chocolate	Confectionary	Paper	100	Grams
5	Crunchy-salty sticks	6008	Salty sticks	Confectionary	Foil	100	Grams
6	Crunchy-salty sticks	0855	Salty sticks	Confectionary	Foil	400	Grams
7	Washing Powder Shine	1025	Washing powder	Cleaning products	Foil	3	Kilos
8	Extra sweet & diet	2265	Beverage	Juice	Can	0.33	Liters
9	Milk chocolate bar	1973	Chocolate bar	Confectionary	Elastic mat.	45	Grams
10	Wafer Andrew	3008	Chocolate bar	Confectionary	Elastic mat.	40	Grams
11	Yogurt Super Cherry	1731	Yogurt	Dairy products	Elastic mat.	500	Grams
12	Kefir Agro	3530	Buttermilk	Dairy products	Elastic mat.	500	Grams
13	Qubus Small	0011	Beverage	Juice	Glass bottle	0.33	Liters
14	Qubus Big	0012	Beverage	Juice	Glass bottle	1	Liters

- Store (see Table 2.10)

Table 2.10. Dimension Store

Store
IDStore (PK)
Store Name
Store Number
Store Address
Store City
Store Manager
Store Region
Selling Square Footage
First Open Date

Table 2.11 presents demonstration data for Store dimension.

Table 2.11. Store: demonstration data

IDStore	Store name	Store number	Store address	Store city	Store manager	Store region	Selling square footage	First open date
1	Alfa_1	1489	Beacon St.	Newton	Jane Davis	MA	137	1-03-2004
2	Alfa_2	3711	Varsov St.	Newton	Trent Tucker	MA	174	1-07-2004
3	Alfa_3	1630	Morton St.	Newton	John Cash	MA	357	1-05-2004
4	Alfa_4	3216	Riverway	Bedford	Steve Lucky	MA	235	1-01-2004
5	Alfa_5	2600	John-Paul St.	Salem	Joe Turner	MA	256	1-09-2004
6	Alfa_6	1085	Tremont St.	Bedford	Horace Grant	MA	136	1-11-2004
7	Alfa_7	1102	Broadway	Burlington	John Novak	VT	185	1-03-2004
8	Alfa_8	2256	Wall St.	Burlington	Larry King	VT	214	1-10-2004
9	Alfa_9	1978	New St.	Newport	Jay Leno	VT	125	1-10-2004
10	Alfa_10	3080	Green St.	Boston	Scott Pippen	MA	115	1-11-2004
11	Alfa_11	1073	Main St.	Boston	John Paxson	MA	200	2-04-2005
12	Alfa_12	1353	Oxford St.	Boston	Michel Jordan	MA	203	1-04-2004

4. Identify the Facts

In Figure 2.3 the retail sales schema (multidimensional model) is presented where the fact table is: Transaction.

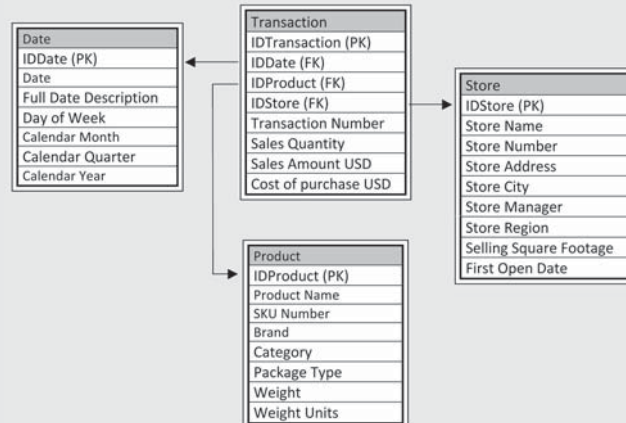


Figure 2.3. Retail sales schema (multidimensional model)

Source: Based on Kimball and Ross (2002)

In Figure 2.4 it is presented how to generate report (shown in Table 2.5) from the multidimensional model shown in Figure 2.3.

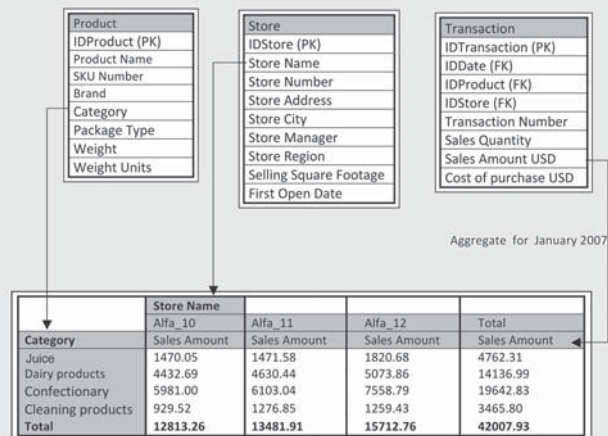


Figure 2.4. Dragging a dimensional attributes and facts into a simple report

Source: author.

In order to generate the required business reports, the following sources of data were specified:

- Cash registers, which register single transactions (receipts)
- ERP system, which includes master data (for dimensions)

Moreover, during the comprehensive analysis of business requirements, the following sources of data were defined:

- ERP system, which includes data about sales, purchases, and logistics costs
- Internet service that makes trade benchmarking data available
- Files
 - Planned sales (once a month)
 - Data concerning location (nearest competition, location, etc.)

Recommended Literature

Data Warehouse

- Inmon, W. H. (2005). *Building the data warehouse* (4th ed.). New York, NY: Wiley.
- Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2010). *Fundamentals of data warehouses* (2nd ed.). Berlin, Germany: Springer Verlag.

Database and Data Warehouse Design

- Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *Unified modeling language user guide* (2nd ed.). Reading, MA: Addison Wesley.
- Imhoff, C., Gallemm, N., & Geiger, J. (2003). *Mastering data warehouse design: Relational and dimensional techniques*. New York, NY: Wiley.
- Kimball, R., & Ross, M. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling* (2nd ed.). New York, NY: Wiley.
- Silverston, L. (2001). *The data model resource book, Vol. 1: A library of universal data models for all enterprises*. New York, NY: Wiley.
- Ullman, J. D., & Widom J. (2007). *First course in database systems* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Case Study (Dimensional Model for a Chain Store)

Kimball, R., & Ross, M. (2002). *Retail sales in the data warehouse toolkit: The complete guide to dimensional modeling* (2nd ed.). New York, NY: Wiley.

Internet Resources

The Data Warehousing Information Center. <http://www.dwinfocenter.org/>
Teradata University Network. <http://www.teradatauniversitynetwork.com/>
Ralph Kimball site. <http://www.ralphkimball.com/>