APPENDIX 13.A: DETAILS OF THE SCHEMA INTEGRATION PROBLEM

This appendix provides details of the schema integration problem not covered in Section 13.4.1. You need to use sample tables (Tables 13A-1 to 13A-4) in the last step of the schema integration process to populate sample tables in the data warehouse design. You also need to use rows in the worksheet for custom product purchases (Table 13A-5).

TABLE 13A-1
Sample Data for the *Supplier* Table

SuppNo	SuppName	SuppEmail	SuppPhone	SuppDisc
S2029929	ColorMeg, Inc.	custrel@colormeg.com	(720) 444-1231	0.10
S3399214	Connex	help@connex.com	(206) 432-1142	0.12
S4290202	Ethlite	ordering@ethlite.com	(303) 213-2234	0.05
S4298800	Intersafe	orderdesk@intersafe.com	(512) 443-2215	0.10
S4420948	UV Components	custserv@uvcomponents.com	(303) 321-0432	0.08
S5095332	Cybercx	orderhelp@cybercx.com	(212) 324-5683	0.00

TABLE 13A-2
Sample Data for the *Product* Table

ProdNo	ProdName	SuppNo	ProdQOH	ProdPrice	ProdNextShipDate
P0036566	17 inch Color Monitor	S2029929	12	$169.00	02/20/2017
P0036577	19 inch Color Monitor	S2029929	10	$319.00	02/20/2017
P1114590	R3000 Color Laser Printer	S3399214	5	$699.00	01/22/2017
P1412138	10 Foot Printer Cable	S4290202	100	$12.00	
P1445671	8-Outlet Surge Protector	S4298800	33	$14.99	
P1556678	CVP Ink Jet Color Printer	S3399214	8	$99.00	01/22/2017
P3455443	Color Ink Jet Cartridge	S3399214	24	$38.00	01/22/2017
P4200344	36-Bit Color Scanner	S4420948	16	$199.99	01/29/2017
P6677900	Black Ink Jet Cartridge	S3399214	44	$25.69	
P9995676	Battery Back-up System	S5095332	12	$89.00	02/01/2017

TABLE 13A-3
Sample Data for the *Purchase* Table

PurchNo	PurchDate	SuppNo	PurchPayMethod	PurchDelDate
P2224040	02/03/2013	S2029929	Credit	02/08/2017
P2345877	02/03/2013	S5095332	PO	02/11/2017
P3249952	02/04/2013	S3399214	PO	02/09/2017
P3854432	02/03/2013	S4290202	PO	02/08/2017
P9855443	02/07/2013	S4420948	PO	02/15/2017

TABLE 13A-4
Sample Data for the *PurchLine* Table

PurchNo	ProdNo	PLQty	PLUnitCost
P2224040	P0036566	10	$100.00
P2224040	P0036577	10	$200.00
P2345877	P9995676	10	$45.00
P3249952	P1114590	15	$450.00
P3249952	P1556678	10	$50.00
P3249952	P3455443	25	$21.95
P3249952	P6677900	25	$12.50
P3854432	P1412138	50	$6.50
P9855443	P4200344	15	$99.00

TABLE 13A-5
Sample Worksheet for Custom Inventory

ProdCode	ProdDesc	Supp	Qty	Unit Price	PurDate	Amount
CPC1	Pencils	Omart	20	$2.00	13-Feb-2017	$40.00
CPC2	Paper	Smart	10	$3.50	14-Feb-2017	$35.00
CPC3	Folders	Pmart	20	$1.50	11-Feb-2017	$30.00

To estimate grain size, you should use these estimates about cardinalities of tables and unique values of some columns.

- Product rows: 1,000
- Supplier rows: 100
- Purchase rows: 100,000 per year
- PurchLine rows: 500,000 per year
- Spreadsheet rows: 1,000 per month; new spreadsheet each month
- Unique products in a spreadsheet for one year: 100
- Unique suppliers in a spreadsheet for one year: 20

APPENDIX 13.B: SOLUTION FOR THE SCHEMA INTEGRATION PROBLEM

1. The dimensions in the problem are reasonably clear. Supplier, calendar, and product are dimensions. Supplier and product come from the ERD and the sample spreadsheet. The calendar dimension is a standard data warehouse dimension. Calendar is a hierarchical dimension. Phone and email can be parsed to be hierarchical as part of the supplier dimension.
   - Supplier
     - SuppNo: ERD only
     - SuppName (Supplier table) | Supp (spreadsheet)
     - SuppPhone: ERD only; hierarchical (country code → area code → prefix → line)
     - SuppEmail: ERD only; hierarchical (top level domain → second level domain → local part)
   - Calendar
     - Date columns in the ERD (ProdNextShipDate, PurchDelDate, and PurchDate) and spreadsheet (PurchDate); hierarchical (year → month → day)
   - Product:
     - ProdNo: ERD only
     - ProdName (ERD) | ProdDesc (spreadsheet)
     - ProdCode: spreadsheet only
2. The measures mostly come from the *PurchLine* table and supply purchases spreadsheet. Measures from related tables are important to associate with the measures from the *PurchLine* table and Supply Purchases spreadsheet.
   - PLQty (PurchLine table) | Qty (spreadsheet); additive measure
   - Amount of purchase: derived additive measure from the spreadsheet
   - PLUnitCost (PuchLine table) | Unit Price (spreadsheet); snapshot measure
   - ProdQOH (Product table) | Stock (Spreadsheet): Semi-additive across products but not useful to add quantity of different products. Usually average across time periods
   - SuppDisc (Supplier table): supplier discount; snapshot measure
   - ProdPrice: product price; snapshot measure indicating the resale price of the product when the purchase occurs
3. The most detailed grain is the combination of individual supplier, individual product, and date.
   - 1,100 products: sum of product rows and unique products in a spreadsheet
   - 120 suppliers: sum of supplier rows and unique suppliers in a spreadsheet
   - Days per year: 365
   - 512,000 purchases of individual products: sum of PurchLine rows and spreadsheet rows (one year)
   - Fact table size is determined from sum of the rows in the PurchLine table and Spreadsheet. Thus, the individual product purchases per year are 512,000.
   - Sparsity estimate:
     - 1 - (fact table size / product of dimensions)
     - (1 – (512,000 / (1,100*120*365)) = 0.98937
     - The data cube has mostly missing cells with slightly more than 1% of cells with non-zero values.

4. The star schema (Figure 13B.1) should support the dimensions and measures specified in parts 1 and 2. There are two relationships between the *Calendar* and *InvFact* tables to record both the purchase and delivery dates. Product type is a new derived column indicating the data source (merchandise for resale or supply for internal usage). *ProdNextShipDate* was dropped in the data warehouse design. The problem did not indicate a clear usage the data warehouse. It could be added as another relationship from *Calendar* to *InvFact* if the date was useful for business intelligence reasoning. The relationship would be incomplete for the spreadsheet data source.
5. The *DelDate* relationship is an incomplete fact-dimension relationship as the delivery date is missing in the supply spreadsheet. It is probably not possible to add to existing data but second data source possibly can be changed in the future so delivery date is collected on the spreadsheet. If delivery date is the same as purchase date for supplies, the same date can be used as a default value.

   There are also missing values for *SuppEmail* and *SuppPhone* for the suppliers from the spreadsheet. Although the Supplies relationship is mandatory, these missing values make the relationship missing for the *SuppEmail* and *SuppPhone* columns. Additional data collection can resolve this incompleteness as no reliable default value exists.
6. The data warehouse tables (Tables 13B-1 to 13B-5) have been derived from the sample date in the source tables and spreadsheet. The delivery date for the supply purchases uses the default value of the purchase date since the values are missing the source data. New primary key values have been generated for data from the spreadsheet data source.

FIGURE 13B.1
ERD for Retail Fitness Database

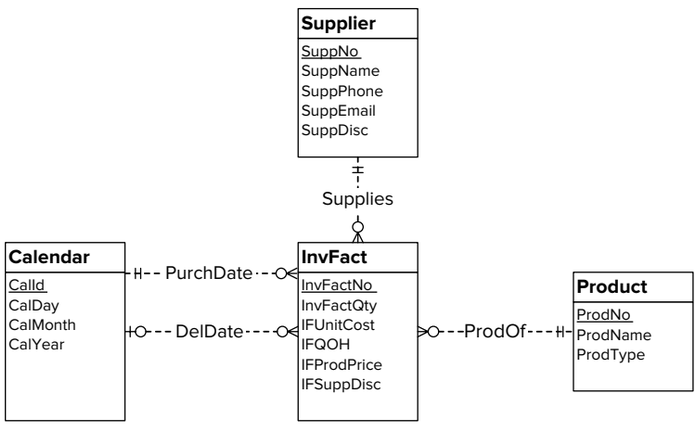


TABLE 13B-1
Sample Data for the *Supplier* Dimension Table

SuppNo	SuppName	SuppEmail	SuppPhone
S2029929	ColorMeg, Inc.	custrel@colormeg.com	(720) 444-1231
S3399214	Connex	help@connex.com	(206) 432-1142
S4290202	Ethlite	ordering@ethlite.com	(303) 213-2234
S4298800	Intersafe	orderdesk@intersafe.com	(512) 443-2215
S4420948	UV Components	custserv@uvcomponents.com	(303) 321-0432
S5095332	Cybercx	orderhelp@cybercx.com	(212) 324-5683
S1111111	Omart		
S1111112	Smart		
S1111113	Pmart		

ProdNo	ProdName	ProdType
P0036566	17 inch Color Monitor	Merch
P0036577	19 inch Color Monitor	Merch
P1114590	R3000 Color Laser Printer	Merch
P1412138	10 Foot Printer Cable	Merch
P1445671	8-Outlet Surge Protector	Merch
P1556678	CVP Ink Jet Color Printer	Merch
P3455443	Color Ink Jet Cartridge	Merch
P4200344	36-Bit Color Scanner	Merch
P6677900	Black Ink Jet Cartridge	Merch
P9995676	Battery Back-up System	Merch
P111111	No 2 pencils	Supp
P111112	Copier paper	Supp
P111113	File folders	Supp

TABLE 13B-2
Sample Data for the *Product* Dimension Table

CalId	CalDay	CalMonth	CalYear
C10000211	1	2	2017
C10000212	2	2	2017
C10000213	3	2	2017
C10000214	4	2	2017
C10000215	5	2	2017
C10000216	6	2	2017
C10000217	7	2	2017
C10000218	8	2	2017
C10000219	9	2	2017
C10000220	10	2	2017
C10000221	11	2	2017
C10000222	12	2	2017
C10000223	13	2	2017
C10000224	14	2	2017
C10000225	15	2	2017
C10000226	16	2	2017
C10000227	17	2	2017

TABLE 13B-3
Sample Data for the Calendar *Dimension* Table

InvFactNo	ProdNo	SuppNo	IFQty	IFUnitCost	IFQOH	IFProdPrice	IFSuppDisc
I2224040	P0036566	S2029929	10	$100.00	12	$169.00	0.10
I2224041	P0036577	S2029929	10	$200.00	10	$319.00	0.10
I2224042	P9995676	S5095332	10	$45.00	12	$89.00	0.00
I2224043	P1114590	S3399214	15	$450.00	5	$699.00	0.12
I2224044	P1556678	S3399214	10	$50.00	8	$99.00	0.12
I2224045	P3455443	S3399214	25	$21.95	24	$38.00	0.12
I2224046	P6677900	S3399214	25	$12.50	44	$25.69	0.12
I2224047	P1412138	S4290202	50	$6.50	100	$12.00	0.05
I2224048	P4200344	S4420948	15	$99.00	16	$199.99	0.08
I2224049	P111111	S1111111	20	$2.00	1		
I2224050	P111112	S1111112	10	$3.50	2		
I2224051	P111113	S1111113	20	$1.50	0		

TABLE 13B-4
Sample Data for the *InvFact* Measure Table (Part 1)

TABLE 13B-5
Sample Data for the *InvFact* Measure Table (Part 2)

InvFactNo	PurchCalNo	DelCalNo
I2224040	C10000213	C10000218
I2224041	C10000213	C10000218
I2224042	C10000213	C10000221
I2224043	C10000214	C10000219
I2224044	C10000214	C10000219
I2224045	C10000214	C10000219
I2224046	C10000214	C10000219
I2224047	C10000213	C10000218
I2224048	C10000217	C10000225
I2224049	C10000223	C10000223
I2224050	C10000224	C10000224
I2224051	C10000221	C10000221