Practice Mini Case for Data Warehouse Design

This mini case study provides practice and preparation for Assignment 2. You will get to apply most of the concepts from module 3.

This mini case study contains two data sources with sample data along with a statement of business needs. Using the data sources and business needs, you will specify a dimensional model with dimensions, measures, and grain, create a schema design for the data warehouse that integrates the data sources, identify summarizability problems in the design, and populate data warehouse tables from sample rows in the data sources.

Data Sources

The case study involves two data sources for a retail firm. The Purchase database supports purchase transactions to replenish retail inventory. A purchase consists of a heading with the purchase number, date, payment method, delivery date, and supplier. A purchase contains a collection of products with the quantity and unit cost recorded on a purchase line along with links to the product and purchase heading. Each product has one preferred supplier. However, a purchase can use a non-preferred supplier if necessary.

Individual stores of the retail firm also maintain an inventory of custom products ordered from local suppliers. These products are ordered through the purchase spreadsheets for custom products. Inventory practices for custom products are informal. New products are typically purchased when the manager senses new demand for local items.

The ERD in Figure 1 supports the purchase database. Tables 1 to 4 show sample data for the tables in the purchases database. The supply purchase spreadsheet (Table 5) contains a sample of purchases of custom products from local suppliers. The Stock column in the spreadsheet indicates the quantity in stock at the time of purchase.

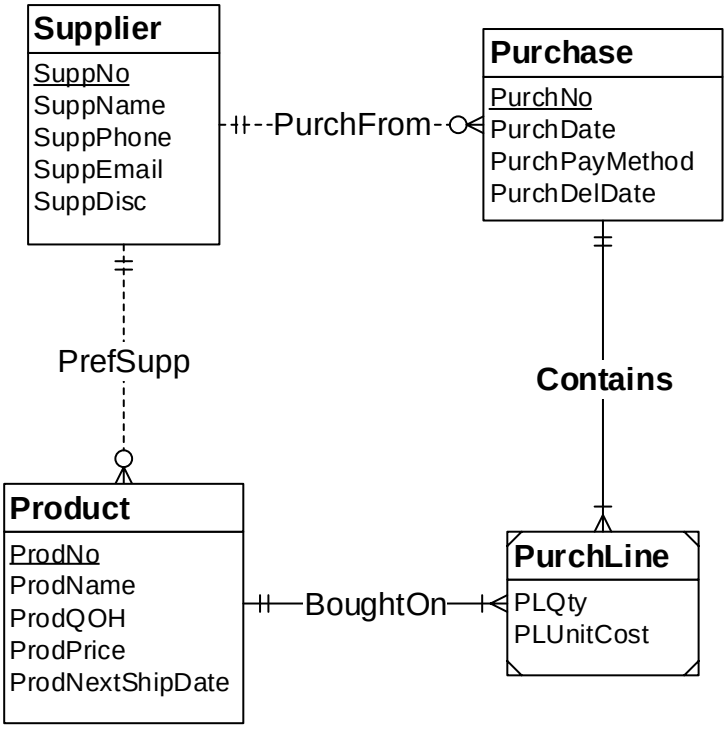

Figure 1: ERD for Retail Purchase Operations

Table 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 2: Sample Data for the *Product* Table

ProdNo	ProdName	SuppNo	ProdQOH	ProdPrice	ProdNextShipDate
P0036566	17 inch Color Monitor	S2029929	12	$169.00	02/20/2021
P0036577	19 inch Color Monitor	S2029929	10	$319.00	02/20/2021
P1114590	R3000 Color Laser Printer	S3399214	5	$699.00	01/22/2021
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	01/22/2021
P3455443	Color Ink Jet Cartridge	S3399214	24	$38.00	01/22/2021
P4200344	36-Bit Color Scanner	S4420948	16	$199.99	01/29/2021
P6677900	Black Ink Jet Cartridge	S3399214	44	$25.69	
P9995676	Battery Back-up System	S5095332	12	$89.00	02/01/2021

Table 3: Sample Data for the *Purchase* Table

PurchNo	PurchDate	SuppNo	PurchPayMethod	PurchDelDate
P2224040	02/03/2013	S2029929	Credit	02/08/2021
P2345877	02/03/2013	S5095332	PO	02/11/2021
P3249952	02/04/2013	S3399214	PO	02/09/2021
P3854432	02/03/2013	S4290202	PO	02/08/2021
P9855443	02/07/2013	S4420948	PO	02/15/2021

Table 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 5: Sample Spreadsheet Data for Custom Product Purchases

ProdCode	ProdDesc	Supp	Qty	Stock	Unit Price	PurchDate	Amount
CPC1	Souvenir 1	Omart	20	1	$2.00	13-Feb-2021	$40.00
CPC2	Souvenir 2	Smart	10	2	$3.50	14-Feb-2021	$35.00
CPC3	Souvenir 3	Pmart	20	0	$1.50	11-Feb-2021	$30.00

Data source size statistics

To compute 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 worksheet each month
- Unique products in a worksheet for one year: 100

- Unique suppliers in a spreadsheet for one year: 20

Business Needs

The main purpose of the data warehouse is to track inventory balances over time. Inventory balances are a type of snapshot. Snapshots are typical in applications in which balances are involved, such as account balances in financial services, enrollment in courses, reservations in hospitality and travel, and head count in personnel management. Snapshots cannot be aggregated over time correctly. Summing quantities and values over time is not meaningful.

The basic values for inventory tracking are quantity on hand and inventory value. Inventory valuation can be complex as many accounting methods exist to value inventory. For this case, the purchase price or unit cost of the inventory can be used for valuation. The data warehouse should support detailed tracking of inventory to the individual product, purchased by date, and supplier.

Here are typical computations for analyzing and tracking inventory balances using the quantity on hand and inventory value:

- The average quantities and stock values in each period
- The opening and closing balances for each period
- The change in inventory levels between consecutive periods and parallel periods
- The minimum and maximum inventory levels in a period
- The relative contribution of the stocked item to the overall stock value

Problems

1. You should identify dimensions, map dimensions to data sources, and specify dimension hierarchies. For each dimension, you should identify its data sources and attributes in each

data source. For hierarchical dimensions, you should indicate the levels from broad to narrow.

2. You should specify measures, related data sources, and measure aggregation properties.
3. Identify the grain in your dimensional design using the business needs as a guideline. You should then indicate relative storage requirements for the grain using the statistics for the data sources. Using the cardinality estimates provided, you should determine either the fact table size or sparsity and then compute the unknown grain size variable. For example, you should compute sparsity if the fact table size is given.
4. Extend your analysis to design a star schema (or variation) to support inventory analysis. Indicate usage of design transformations (flatten and merge) in your schema design. For each table, you should define the table name, primary key, and columns. You do not need to write complete CREATE TABLE statements.
5. Identify summarizability problems in your star schema and indicate preferred resolutions of the summarizability problems. For incomplete dimension-fact relationships, you should also indicate if columns in a dimension table allow null values.
6. You should populate your data warehouse tables based on the data in the sample tables and spreadsheet. You do not need to write SQL INSERT statements or insert the data into your tables. You can just show table listings in your solution. You should indicate mappings from data sources into tables. For example, a mapping may involve generating new primary key values for a data warehouse table or using a default value for a missing value.