Mini Case for the Assignment in Module 3

The assignment in module 3 gives you the opportunity to apply most of the concepts from module 3 and some of the concepts from module 2. Since the assignment is closely based on the practice problems in module 3, you should attempt the mini case study for the practice problems and then review the solution details.

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

Fitness Unlimited is a leading provider of exercise centers with a variety of fitness programs and membership options. Fitness Unlimited maintains a retail database to track sales of services and merchandise. In the ERD for the retail database (Figure 1), a sale contains a heading (*Sale*) with sales date and a collection of merchandise recorded in the M-N relationship *Contains*. Service purchases are recorded in the *ServPurchase* entity type with 1-M relationships from ServiceCategory and Member. Typical services are lessons, premium equipment usage, and social events. The *MemTypeOf* relationship is optional for members because guest members can use a fitness center and purchase merchandise and services on a short term basis without having a paid membership. Tables with sample rows are shown after Figure 1.
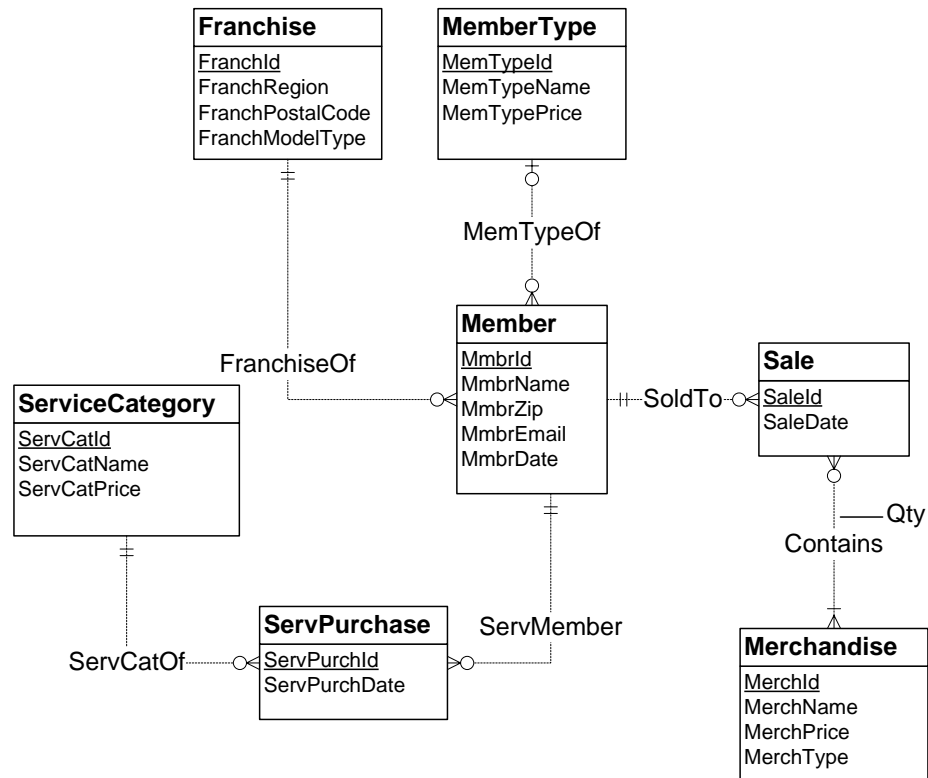
Figure 1: ERD for Retail Fitness Database

Franchise			
FranchId	*FranchRegion*	*FranchPostalCode*	*FranchModelType*
F1	Northwest	98011	Full
F2	Mountain	80111	Medium
F3	Central	45236	Limited

MemberType		
MemTypeId	*MemTypeName*	*MemTypePrice*
M1	Platinum	$1,000
M2	Gold	$800
M3	Value	$300

ServiceCategory		
ServCatId	*ServCatName*	*ServCatPrice*
SC1	Ball machine	$15
SC2	Private lesson	$75
SC3	Adult class	$150
SC4	Child class	$125

Merchandise

MerchId	MerchName	MerchPrice	MerchType
MC1	Wilson balls	$3	Balls
MC2	Wilson racket	$200	Racket
MC3	Adidas shoes	$100	Shoes
MC4	Racket stringing	$40	Racket

Member

MmbrId	MmbrName	MmbrZip	MemTypeId	MmbrDate	FranchId	MmbrEmail
1111	Joe	98011	M1	1-Feb-2009	F1	joe@serv1.com
2222	Mary	80112	M2	1-Jan-2010	F2	mary@serv2.com
3333	Sue	45327	M3	3-Mar-2011	F3	sue@serv3.com
4444	George	45236			F3	george@serv4.com

Sale

SaleId	SaleDate	MmbrId
1111	10-Feb-2013	1111
2222	13-Feb-2013	2222
3333	13-Feb-2013	2222
4444	14-Feb-2013	3333

Contains

MerchId	SaleId	Qty
MC1	1111	2
MC2	1111	1
MC4	2222	1
MC3	3333	1
MC4	4444	1

ServicePurchase

ServPurchId	ServPurchDate	MmbrId	PassCatId
1111	13-Feb-2013	1111	SC1
2222	14-Feb-2013	2222	SC2
4444	15-Feb-2013	4444	SC3

Franchises also sell special events to corporate and other organizations. Since special event promotions and sales are not standard among franchises, spreadsheets are typically used to track special events. The franchise sales database was never extended to accommodate special event sales. The Special Events Worksheet shows a typical format for tracking special event sales by a franchise. Most franchises use a similar spreadsheet.

Special Events Worksheet					
Corporate Customer Id	*Corporate Customer Name And Location*	*Event Type Code*	*Event Name*	*Event Date*	*Amount*
CC1	First Data, Greenwood Village	L-A	Adult Social	13-Feb-2013	$1,000
CC2	DU Tennis, Denver	L-B	Pioneer Social	14-Feb-2013	$500
CC3	Creek High School, Greenwood Village	L-C	Team Practice	21-Feb-2013	$200

Data Source Size Estimates

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

- Franchise rows: 350

- Franchise postal codes: 200

- MemberType rows: 10

- Merchandise rows: 500

- MerchType values: 30

- ServCategory rows: 20

- Member rows: 50,000

- Member zip codes: 500

- Sale rows: 150,000 per year

- Contains rows: 450,000 per year

- ServicePurchase rows: 100,000 rows per year

- SpecialEvents Worksheet rows: 300 per year per franchise with 200 franchises using this

  spreadsheet

- 150 unique customers per special event worksheet

Business Needs

The data warehouse should support analysis of merchandise sales and service purchases by franchise, merchandise or service type, and customer over time. For merchandise, sales

amount is computed as quantity times selling price. For services purchases, each unit sale is

recorded separately so only the service price at the time of purchase is recorded.

For customer, merchandise sales should be tracked by

- zip code,

- membership date, and

- member type.

For franchise, merchandise sales should be tracked by

- franchise region,

- postal code, and

- model type.

The corporate sales office wants a high level of flexibility for sales analysis. For data

mining analysis, they need detail by individual customer, product or service, and franchise, and

date. For typical reporting applications,

they need detail by

- customer location,

- franchise location, and

- product or service type, and

- week.

Problems

You should design a star schema (or variation) to support revenue analysis. You should pay close attention to the grain of the fact table, the major part of the star schema diagram. As part of the design, you should identify all relevant dimensions with hierarchies specified. In your documentation, you should identify summarizability problems in your star schema and indicate mapping from data sources into tables.

You should populate your data warehouse tables based on the data in the operational tables and spreadsheet. You do not need to insert the data into your tables. You can just show table listings in your solution document. Your sample rows should include all revenue events in the range February 10, 2013 to February 21, 2013.

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. 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 potential 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 data into database 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.

Solution Quality

Quality is rather subjective in data warehouse designs, but some elements are less subjective. I suggest that you address these quality items in the appropriate part of your solution.

- Schema pattern: You should use a recognized schema pattern: star, constellation, or snowflake schema.

- Fact table selection: You should study fact table selection in the solution for the practice mini case for inspiration. Typically, the fact table combines a two level solution in a source schema into a single fact table. For example, an order heading and order detail are usually combined into a fact table recording the order details with dimension relationships to capture the order heading.

- Missing data in populated tables: You should ensure that your populated tables include all revenue events shown in both data sources. The best check on your schema design is to map sample rows from the data sources to the data warehouses.

- Simplicity: Typically, a data warehouse schema design simplifies the schemas of the underlying data sources. Simplification can involve combining some elements of data sources in decisions about dimensions and fact tables.

Grading

The assessment method for this assignment is peer review. Each problem has an equal grade.

Submission

You need to submit 6 documents to the peer review for module 3 assignment. Each document contains a full solution for the problem. You should neatly format your documents so that it can be easily graded. Please write the problem number at the top of the page.