

## Lab 2 – Star Schema

### Overview

TPC is ready to implement its first data mart. In this lab, you will analyze the user requirements for this data mart, design it using dimensional modeling techniques, and implement the schema design in your database.

### After completing this lab exercise, you should be able to:

- Translate user information requirements into a design for a data mart. •
- Identify the data needs and potential data sources for a data mart.
- Implement the design for a data mart in a database schema.

### To do this lab you will need the following:

- 1) Your copies of the *TPC* case study, business rules, and ERD.
- 2) Access to a computer running MySQL and MySQL Workbench.

### Deliverables

Submit the following to MyCourses as a single .zip file with the following name –  
*YourLastName\_Lab2\_201.zip*:

- 1) Answers to the questions in this specification in a separate document (MS Word or .pdf).
- 2) An ER diagram showing your Star Schema, developed in MySQL Workbench and pasted into a MS Word or .pdf document.
- 3) A screenshot of your tables from MySQL Workbench.
- 4) Your MySQL Workbench model *YourLastName\_201.mwb* file
- 5) A *YourLastName\_201.sql* file that contains the dumping your database

## Business Scenario

TPC management has identified financial control and analysis as their top current issue. After talking with the users in the TPC central office in Stratford who are involved with financial control, you find out the following:

- Although each of the three divisions is responsible for financial control (increasing sales and decreasing costs), the Financial Director in Stratford is primarily responsible for overall company performance.
- The three divisions will provide data to the data warehouse in different forms. You will have access to OLTP database for TPC-E. This will provide you with sales data for TPC-E. Since you have access to the developers, they can help you with the data in the OLTP system.
- TPC-W is similar in operation to TPC-E and TPC-W will provide a feed of data for you to use. The data will be similar to that which you have access to for TPC-E. An initial feed will be provided from TPC-W and a monthly feed will be provided with updates each month.
- The data from PEC will be different. There will be a customer and a product feed, but the cost will have to be calculated from manufacturing cost data that will be provided. Formulas for calculation will be provided later.
- Since sales can be made from TPC-E and TPC-W to PEC and from PEC to TPC-E and TPC-W, there will need to be allowances when sales and costs are calculated at the total company level. Sales from one unit to another, although considered as sales for the first unit, are not considered sales for the total company (You can't count sales to yourself). You will need to identify these customer records.
- There may be overlap in customers among the three company units.
- The company's financial performance is measured on an annual, quarterly, monthly and weekly basis. Quarters are based on the normal annual quarters for comparison against other companies. (e.g Quarter 1 is January, February and March ...). The company's fiscal year (financial reporting and tax year), however, goes from April 1 through March 31. The fiscal quarters conform to the fiscal year (e.g. Fiscal quarter 1 is April, May and June ...). As an example, fiscal 2011 will extend from April 1, 2011 through March 31, 2012.
- Invoice numbers are not unique across the three divisions, so it will be necessary to keep track of the division responsible for the sale.
- PEC sometimes requires special shipping for the products they manufacture. The options are "Train", "Truck", "Air", "N/A" (not available or applicable). Sales that have no special requirements are coded as 0 on the invoice. The other divisions do not provide this information.
- PEC also provides data on the sales feed about the ordering method. The options are "Internet", "phone", "email" or "mail". The values are stored as text. This information is not provided by the other divisions.
- Since PEC manufactures equipment, in addition to the sale date, there is also an order date. The time between order and sale can be used to measure the performance of the organization's manufacturing process. The other divisions do not provide this information since they normally ship from stock.
- Payment method is also provided by PEC on the sales feed. The three valid methods are "COD", "charge" or "cash". Again, this is stored as text. It is not provided by the other divisions.
- After the initial load feeds, there will be similar feeds for monthly updates.

The company would like a data mart that would allow them to investigate their financial performance at the gross profit (margin) level historically so as to more effectively manage financial performance. They are interested in having a flexible system that will ultimately allow them to optimize sales (to maximize sales) while keeping costs down. In addition, they want to be able to better manage the relationships with their suppliers. Some of the initial queries and reports they would like are:

- A report that shows the sales, and costs associated with each customer or customer type on an annual, quarterly, monthly or weekly basis.
- A similar report showing top customers.
- A similar report as above at the product level / product type/ business unit.
- The average time in days needed to fulfill an order from PEC.
- The number of orders that are not shipped within 10 days of order from PEC.
- The average number of products and sales per invoice (keep in mind that invoice number is not unique across divisions).
- What is the average number / maximum number of different shipping methods on each invoice?
- The average cost of shipping for a particular product by different methods.
- The percentage of invoices that are COD.
- The most frequent method of ordering a product from PEC.
- What is the average number of products supplied by each supplier?
- Show the total cost of products for each supplier.
- Show sales from one division to another.
- Comparisons should be able to be done from year-to-year, quarter-to-quarter, month-to-month, same month or quarter compared to last year, ... . This should be able to be done on a calendar year basis or a fiscal year basis.
- Sales by type of customer, by state, by product type, by business unit.
- The sales by supplier state to customer state. This would be useful to see if suppliers should ship directly to customers.
- All reports should be able to report sales, costs and gross profit (sales minus costs).

## Part #1. Requirements Gathering – Fill Out an Information Package

NOTE: Record your answers to the questions below in a separate MS Word or .pdf document that will be submitted for grading.

### Step #1-1: Identify the Process

Remember the focus of a data mart is *one* key business process that is important to company success.

*Question:* Which business process will be the focus of this data mart development?

The sales and movement of TPC Management products from TPC-W, TPC-E, and the PEC divisions.

*Define:* Write a statement that defines the scope – i.e. universe of discourse – of this data mart.

This data mart is for the TPC management to analyze and monitor the sales of their products throughout their TPC-W, TPC-E, and PEC divisions.

*Question:* Assuming that the *TPC-E* ERD and other data sources cover various business activities and data systems within the company, what are the source data system(s) that are relevant to this development? Fill out **Table 1** with the details.

*Table 1.* Business Activities & Relevant ERD Tables

Business Activity	Relevant ERD Table(s) or other datas source(s)
TPC-W	Product Sales, customer, product type, fiscal year, calendar year
TPC-E	Product Sales, customer, product type, fiscal year, calendar year
PEC	Product Sales, customer, product type, fiscal year, calendar year, shipping type, sale date, order date, order method, payment method, shipping type

## **Step #1-2: Choose the Grain**

*Question:* What grain options do you see in the scenario?

Product sales by customer, type, date, shipping, order, and/or payment method.

*Question:* What level of detail do you propose for this data mart? Why?

The manufacturing cost and price of a product as it is the lowest level of the users' information needs.

## **Step #1-3: Identify the Dimensions**

*Question:* What business dimensions are relevant to the scenario?

Customer, product type, fiscal year, calendar year, shipping type, sale date, order date, payment method, and order method.

*Question:* Will you have any degenerate dimensions in your model? Explain.

Yes, the invoice number will need to be included to keep track of which division was responsible for a sale.

*Question:* Will you have any role-playing dimensions in your model? Explain.

Yes, there will be a calendar vs fiscal year dimensional tables along with sales date vs order date tables.

*Question:* Will you have any junk dimensions in your model? Explain.

No, there aren't any attributes from any dimensions that seem unimportant or unnecessary to include in their respective dimensional table.

## **Step #1-4: Identify the Facts**

*Question:* What are the key performance metrics needed by the users?

Product sales by customer, type, date, shipping type, order method, and payment.

*Question:* What *type* of fact table schema will this be? (Refer to the Week #4 lecture discussion of schema types.) Explain your reasoning.

Periodic as they will be sending database updates on a monthly basis.

Fill in **Table 2** with the information about the facts that are relevant to this process. Include in your description the reason *why* a given fact is included (i.e. for what will it be used?).

Table 2. Data Mart Fact Group Details

Fact Group: Product Sales

Fact Name	Fact Description	Default Aggregation Rule
Manufacturing Cost	Cost to manufacture the product	Sum
Price	Sell price of the product	Sum

### **Step #1-5: Complete the Process Information Package**

Fill in the Information Package chart in **Appendix A** for this process.

*Question:* Did you identify any hierarchies within the dimensions? If so, list them here.

Fiscal Date -> Year, Quarter, Month, Week, Date

Calendar Date -> Year, Quarter, Month, Week, Date

Sale Date -> Year, Quarter, Month, Week, Date, Time

Order Date -> Year, Quarter, Month, Week, Date, Time

Customer -> Country, State, City, Zip, Street

Product Type -> Level, Product Type, Business Unit

Invoice -> Invoice Number, Division

Draw a dimensional table detail diagram (refer to Week #4/5 Practice Exercise #3) for your dimension(s). Put the attribute for the lowest grain level at the bottom. Surround it with the other time items (attributes) and show relationships with arrows. Clearly delineate any hierarchies.

Fiscal Date -> Year, Quarter, Month, Week, Date

Calendar Date -> Year, Quarter, Month, Week, Date

Sale Date -> Year, Quarter, Month, Week, Date, Time

Order Date -> Year, Quarter, Month, Week, Date, Time

Customer -> Country, State, City, Zip, Street

-> Customer type, Customer Name

Product Type -> Level, Product Type, Business Unit

PEC Order -> Shipping Type

-> Order Method

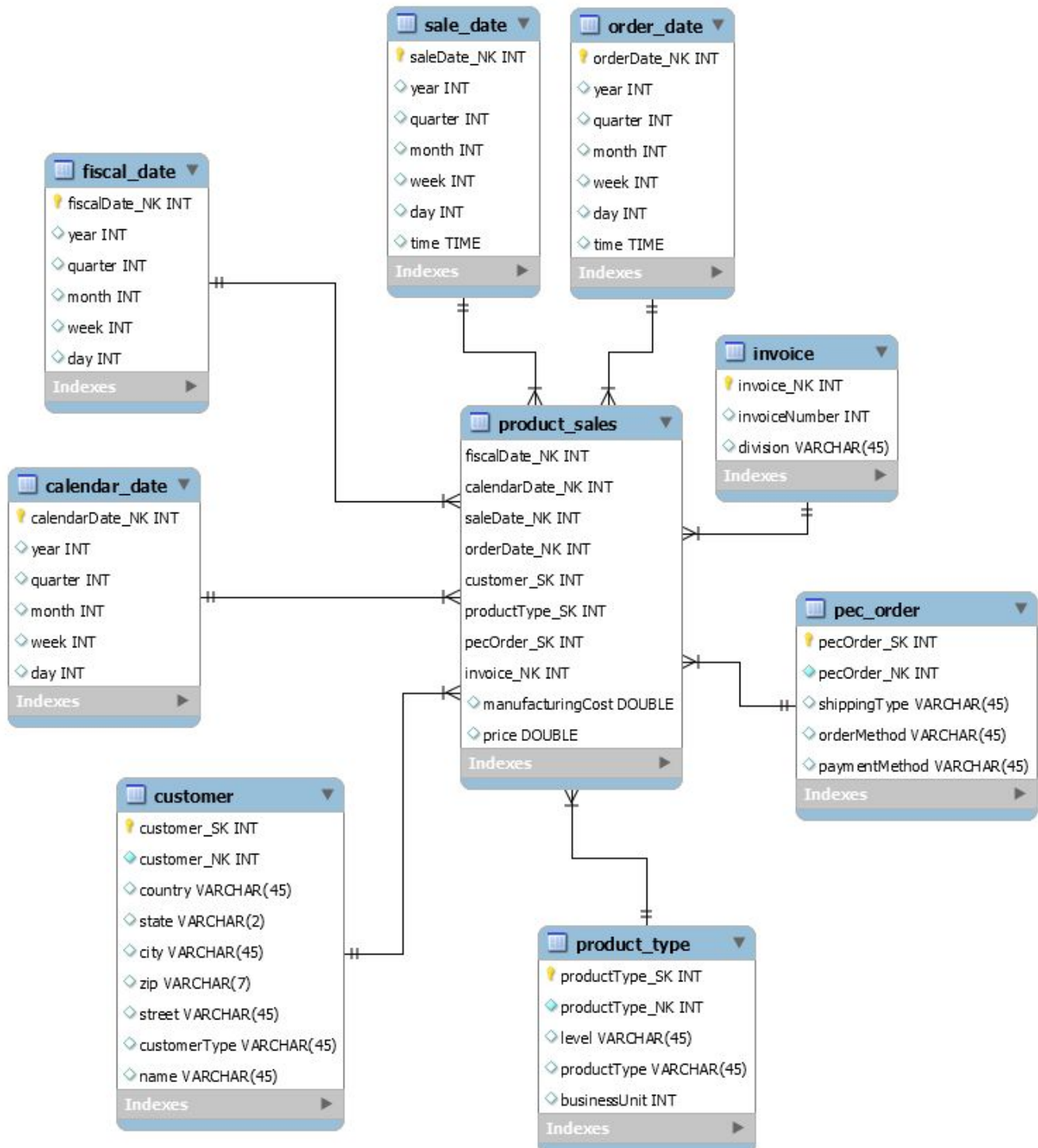
-> Payment Method

Invoice -> Invoice Number, Division

## Part #2. Dimensional Modeling

### Step #2-1: Design Your STAR Schema

Using the information that you have gathered, design a STAR schema for this process. Use MySQL Workbench to formally draw your model. Define tables, attributes, attribute data types, and relationships (with cardinality and participation). Save (paste) your STAR schema in an MS Word or .pdf document and save it to the MyCourse dropbox.





## Step #2-2: Implement the STAR Schema

Create a MySQL database called YourLastName\_FinancialDM that will contain your data mart. This will be similar to what you did in Lab #1 using MySQL Workbench.

Implement your STAR schema in your YourLastName\_FinancialDM data mart. You should save your SQL generated from MySQL Workbench.

You may define your constraints (PK, FK, etc.) and indexes in your model now but you can wait to implement them in your database until after you've loaded the data.

Question: Why would you want to wait?

If there is a large amount of data, it will take a very long time to load it all in if the foreign keys are defined already.

Fill in **Table 3** for the tables that you defined.

*Table 3: Data Mart Tables*

Table Name	Fact or Dimension?
Product Sales	Fact
Fiscal Date	Dimension
Calendar Date	Dimension
Sale Date	Dimension
Order Date	Dimension
Customer	Dimension
Product Type	Dimension
PEC Order	Dimension
Invoice	Degenerate Dimension

Appendix A: Information Package

Process Name: Product Sales

Fiscal Date	Calendar Date	Sale Date	Order Date	Customer	Product Type	PEC Order	Invoice
Year	Year	Year	Year	Country	Level	Shipping Type	Invoice Number
Quarter	Quarter	Quarter	Quarter	State	Product Type	Order Method	Division
Month	Month	Month	Month	City	Business Unit	Payment Method	
Week	Week	Week	Week	Zip			
Day	Day	Day	Day	Street			
		Time	Time	Customer Type			
				Name			

Measured Facts: Manufacturing Cost, Price

Part/Step	Q#	Max Pts.	Pts. Earned	Comments
1-1	1	5		
	2	5		
	Table 1	10		
1-2	1	5		
	2	5		
1-3	1	5		
	2	5		
	3	5		
	4	5		
1-4	1	5		
	2	5		
	Table 2	10		
1-5	1	5		
	2	5		
2-1	1	15		
	2	5		
	3	5		
2-2	1	15		
	2	10		
	3	5		
	Table 3	5		
Appendix	Info Package	10		
	<b>Total</b>	150		