R·I·T

**Department of Information Sciences and Technologies**

# 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*.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*.mwb file
5) A *YourLastName*.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 in 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 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 are 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?

**Solution: Financial Control and Analysis will be the business process to focus of this data mart Development.**

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

**Solution: Scope of Data mart: To create Data mart to investigate financial performance at the gross profit (margin) level to manage financial performance more effectively.**

*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 data source(s) |
|---|---|
| **Sales** | **Customer, Products, Customer Type, Product Type, Sales Date, Business Unit, Order Date.** |
| **Purchasing** | **Customer, Products, Customer Type, Product Type, Sales Date, Business Unit, Order Date.** |
| | |
| | |
| | |

## Step #1-2: Choose the Grain

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

**Solution: Grain options for this scenario are sales by daily basis, sales by weekly basis, sales by type of customer, sales by type of product, sales by state, sales by business unit, product sales on weekly basis.**

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

**Solution: Determination of grain at lowest level i.e. Row Level for measurement in fact table is best practice because of Addition of Attributes and Dimensions at Lowest – Level i.e. it is easy to slice, dice, roll-up, drill-down basically OLAP operations so we don't need to go deep into transaction system. It even becomes resilient to dynamic changes.**

## Step #1-3: Identify the Dimensions

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

**Solution: Customer, Product, Payment Method, Sales_Date, Order_Date, Supplier_Details**.

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

**Solution: Yes, there will be degenerate dimensions in my model. Invoice Number details i.e. Invoice ID it is not unique across divisions as it doesn't incorporate within any dimensional tables. A degenerate dimension is a dimension key in the fact table that does not have its own dimension table, because all the interesting attributes have been placed in analytic dimensions.**

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

**Solution: Yes, Date dimensions: Eg: Sales_Date and Order_Date they are role playing dimensions. Role playing dimensions are dimensions playing multiple roles. For example, sometimes product is processed in short amount of time depending upon how much long time would it take to ship. Therefore, date dimensions play multiple roles.**

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

**Solution: Yes, there are junk dimensions shipping method, ordering method, payment method they are low cardinality attributes. These dimensions can be combined to one dimensional table rather than separate this will make the model less complex increase efficiency and user-friendly model.**

## Step #1-4: Identify the Facts

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

**Solution: Key Performance Indicators are ways to measure the process/performance of business. They are Product cost, Quantity of Product, Sales Amount.**

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

**Solution: Accumulating Snapshot is the type of fact table Schema because it has two role playing dimensions describing entire life-cycle of process that keep record and track of orders i.e Sales_Date and Order_Date.**

_

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: Financial_Analysis_Fact | | |
|---|---|---|
| **Fact Name** | **Fact Description** | **Default Aggregation Rule** |
| **Sales Amount** | **The total amount i.e the total cost purchased by the customer.** | **Sum-Fully Additive** |
| **Number of days** | **The number of days taken to process and deliver the order.** | **Semi-Additive** |
| **Quantity** | **The total count of quantity sold.** | **Sum-Fully Additive** |
| **Product Cost** | **The cost of the product.** | **Sum-Fully Additive** |

## 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.
**Solution: Yes, there are hierarchies within the dimensions.**

**Sales_Date: (Calender and Fiscal) Year -> Quarter -> Month -> Week -> Type of Day -> Day**

**Order-Date: (Calender and Fiscal) Year -> Quarter -> Month -> Week -> Type of Day -> Day**

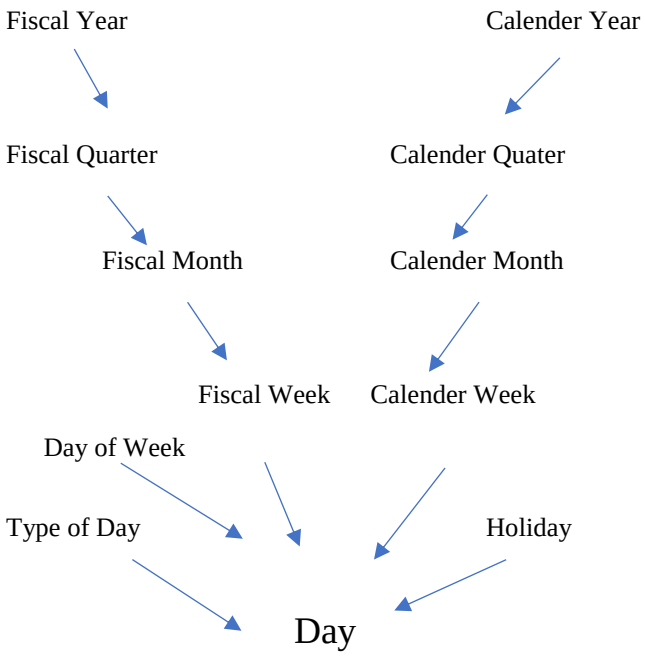
**Customer: Country-> State-> City -> Street Address -> Zip -> Customer Name**

**Supplier: Country-> State-> City -> Street Address -> Zip -> supplier Name**

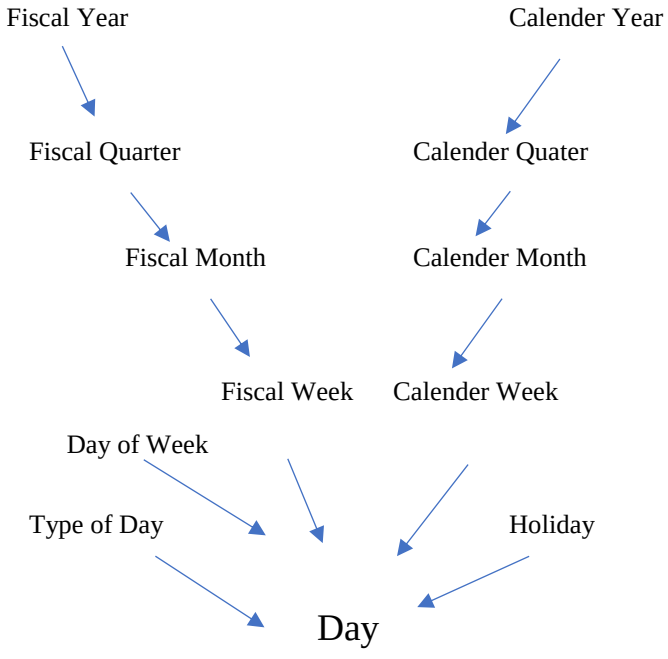
**Product: Product Type -> Product category -> Product Description -> Product Name -> Product Price**

Draw a dimensional table detail diagram (refer to Week #4/5 Practice Exercise #3) for your **time** 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.

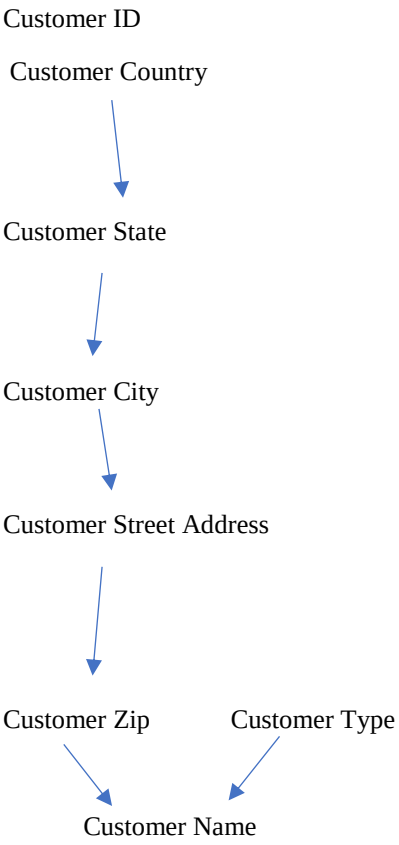
## Order_Date_Dimension



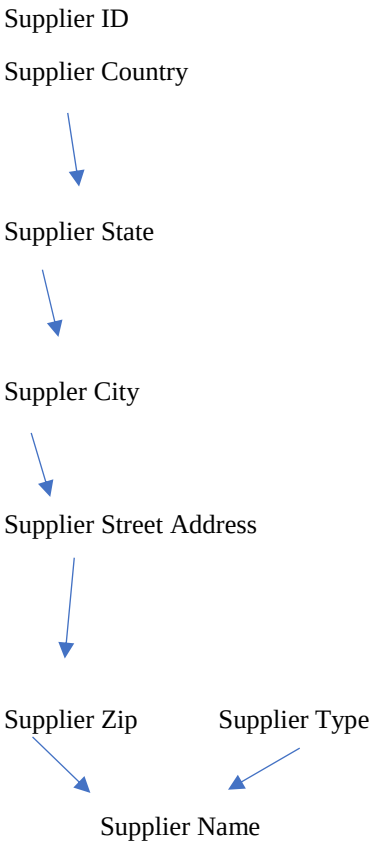

## Sales_Date_Dimension



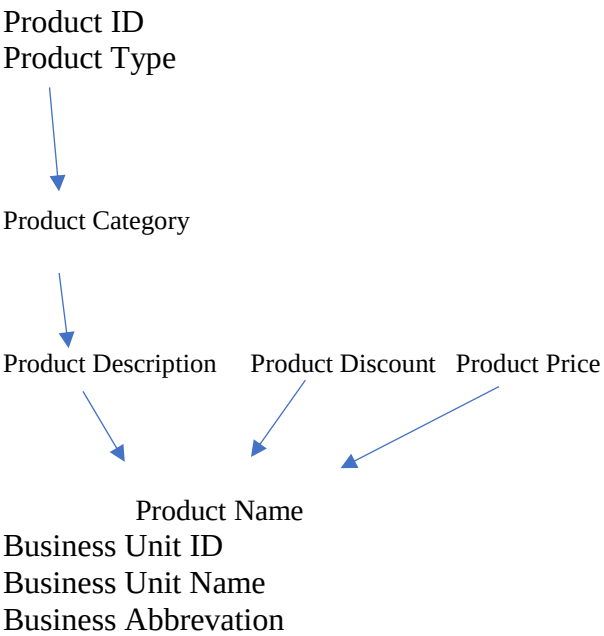
## Customer_Dimension




## Supplier_Dimension




## Product_Dimension




## Payment_Method

Payment Method
Ordering Method
Shipping Method

# 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. Print your STAR schema and submit it to your instructor.

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

Create a MySQL database called TPC_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 TPC_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?

Solution: Waiting to add the foreign key, primary keys, etc constraints until the data is loaded will help data to load faster because when constraints are available time taken increases as data is checked for every constraint when inserted in the table.

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

*Table 3:* Data Mart Tables

| Table Name | Fact or Dimension? |
|---|---|
| **Customer_Dimension** | **Dimension** |
| **Product_Dimension** | **Dimension** |
| **Sales_Date_Dimension** | **Dimension** |
| **Order_Date_Dimension** | **Dimension** |
| **Financial_Anlysis_Fact** | **Fact** |
| **Supplier_Dimension** | **Dimension** |
| **Payment _Method** | **Dimension** |

# Appendix A: Information Package

Process Name: Financial_Analysis

| Sales_Date_ Dimension | Order_Date Dimension | Customer_Dimen Dimension | Product_Dimen Dimension | Supplier_Dim-en Dimension | Payment_Method D | | |
|---|---|---|---|---|---|---|---|
| Sales_Date_Sk | Order_Date_Sk | Customer_SK | Product_SK | Supplier_Sk | Payment_SK | | |
| Calender Year | Calender Year | Customer ID | Product Type | Supplier ID | Payment Method | | |
| Calender Quater | Calender Quater | Customer Name | Product Name | Supplier Name | Shipping Method | | |
| Calender Month | Calender Month | CustomerAddress 1 | Product Category | Supplier Address 1 | Ordering Method | | |
| Calender Week | Calender Week | CustomerAddress 2 | Product Description | Supplier Address 2 | | | |
| Fiscal Year | Fiscal Year | Customer Country | Product ID | Supplier Country | | | |
| Fiscal Quater | Fiscal Quater | Customer State | Business Unit ID | Supplier State | | | |
| Fiscal Month | Fiscal Month | Customer City | Business Unit Name | Supplier City | | | |
| Fiscal Week<br>Day of Week<br>Type of Day | Fiscal Week<br>Day of Week<br>Type of Day | Customer Street Address | Business Unit Abbrevation<br>Product Discount | Supplier Street Address | | | |
| Day | Day | Customer Zip<br>Customer Type | Product Price 1<br>Product Price 2 | Supplier zip | | | |
| **Measured Facts:** _Sales_Amount, Product_Cost, Number_of_Days, Quantity. | | | | | | | |

ISTE-DW Lab #2 Grade Sheet Name: Jyotsna Namdeo Nakte

| 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 | | |
| | **Total** | 150 | | |