Due: 09/06/20 Name: Ellie Parobek

**Rochester Institute of Technology Golisano College of Computing and Information Sciences School of Information**

# Practice Exercise #2 (PE02) Information Packages

## Overview

In this exercise, you will have an opportunity to investigate the information package approach to requirements gathering that is proposed by your textbook author for designing data warehouses and data marts.

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

- Show how the information-package approach can be used to determine the schema for a data mart.
- Design and explain the components of a STAR schema, justifying component placement.
- Discuss the importance of careful, schema design for data warehouse/mart performance.

## Business Scenario

The XYZ Indemnity Corporation is a large insurance enterprise that sells insurance policies, annuities, and other financial vehicles. They would like to design a data mart that will allow them to monitor the sale of insurance policies through their sales and marketing organizations.

In talking with them, you find out the following:

- Insurance policies are sold by company agents to clients nationwide.
- For a policy, it is important to know the policy category (individual, group, industry), type (home owners, auto, life, casualty, disability, health, etc), the face value (coverage level) of the policy, the sale and effective dates of the policy, whether or not the policy is currently in force, and the deductible level. Each policy is identified by a 10-digit policy identification number.
- For this portion of the business, it is important to know the names and hire dates of agents. Agents are identified by the standard company employee identifier, their social security number.
- Each agent sells in a given sales territory. Each territory has an identifier and relates to a given zip code. They also keep track of the date that the territory was last redistricted (the zip code was added or deleted). Territories are grouped into sales districts.
- The company keeps extensive information about their clients. However, the information most relevant to policy sales is the policy holder's name and address (street, city, state, and zip code). Customers are identified by a twelve-character, alphanumeric customer ID.

Sales and marketing managers are interested in viewing sales data by territory (or district), effective date, policy category, and face value. They also want to be able to get individual reports on their agents that cover policy sale amounts. Information on sales territories is also important here. Some of the queries and reports they would like are:

- Total number of policy sales per month by territory (district) by type (category) of policy

- Total number of policy sales per quarter by territory (district) by type (category) of policy
- Total number of policy sales per month by agent by type (or category) of policy
- ...
- Total face value of all policies by month of effective date
- Total face value of all policies by year of effective date
- ...
- Total number of policies in force/not in force by agent
- Total face value of all policies sold by an individual agent
- ...
- Total number of policy sales per month by agent territory (or district) and client zip code
- Top selling[1] agent by territory (or district) by month
- ...

In addition to their internal company data, the company purchased a master zip code table from the US Postal Service and tables of various other national actuarial and risk assessment statistics.

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

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:* What are the user department(s) that are relevant to this development?

Sales and marketing.

*Choose:* Business department(s) [ are ] [ are not ] the same as the process under investigation.

*Question:* What would be the source data system(s) that are relevant to this development?

Policies, agents, sales territory/districts, clients, policy types/categories.

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

This data mart is for the XYZ Indemnity Corporation to monitor the sales of their insurance policies throughout their sales and marketing organizations.

Step #1-2: *Choose the Grain*

*Question:* What information is grain selection based upon?

The users' information needs; a single row in the fact table.

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

The number of policy sales by territory/district, by type/category, by time frame, by agent, or by client.

*Question:* What "best practice" applies to the determination of grain? Why?

Keeping the fact table at the lowest grain because if it is not at the lowest, there will need to be a new dimensional table added.

*Question:* What level of detail – i.e. grain – do you propose for this data mart?

Number of policies sold per agent to which client under what category by date.

1 Top-selling means? Number of policies sold? Highest total face value? Assume: # policies.

Step #1-3: *Identify the Dimensions*

*Fill in:* Dimensions are single-valued entities that describe the important business measurements that the user wants to track.

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

Agent, client, type and date.

*Question:* Do any of the dimensions include hierarchies? If so, list hierarchies in each dimension.

Agent contains territory which contains district and type contains category.

Step #1-4: *Identify the Facts*

*Fill in:* Facts are the important numeric, additive values that the user wants to monitor to determine business success or failure.

*Explain:* "Facts are typically additive" means that …

They are measured values that business users can easily analyze to determine the outcome of the business process.

*Questions:* a) What are the key performance metrics needed by the users?

The number of policies sold per agent, per client, or per type.

b) What attributes do these facts imply will be needed in the fact table?

Type foreign key, agent foreign key, and client foreign key.

Fill in the Information Package below.

**Subject:** Insurance Policy Sales

| Policy | Agent | Type | Client | Date |
|---|---|---|---|---|
| Category | Agent Name | Type Name | Client Name | Day |
| Type | Hire Date | Type Description | Street | Month |
| Face Value | Territory Name | Type ID | City | Year |
| Sale date | Territory ID | Category Name | State | Quarter |
| Effective Date | Territory Zip Code | Category Description | Zip Code | |
| In Force | Territory Redistricted Date | Category ID | Customer ID | |
| Deductible Level | District Name | | | |
| Policy ID | District ID | | | |

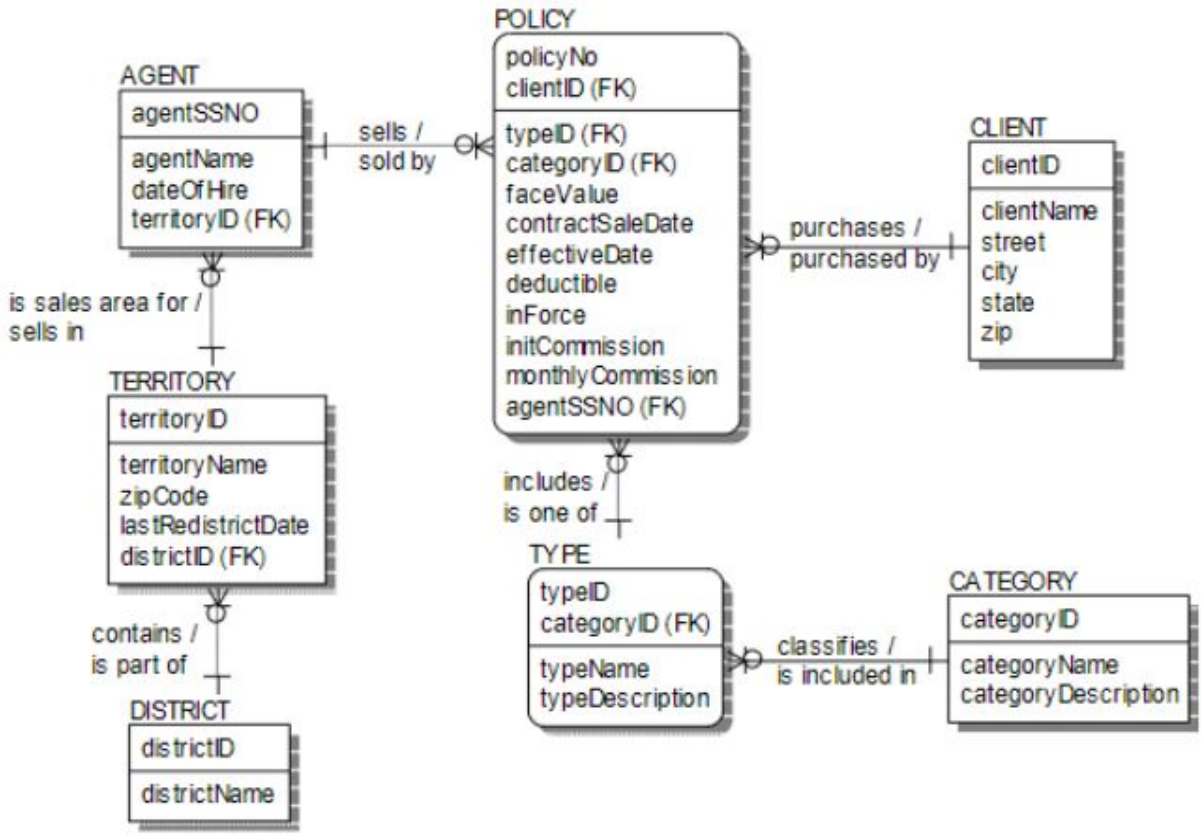
**Measured Facts:** Policy sales, agent sales, category/type sales, client purchases, sales per timeframe

## Part #2. Dimensional Modeling – Create the STAR Schema

Step #2-1: *Model the STAR Schema*

Use the ERD in **Figure 1** below as the data definition source for modeling your STAR schema.

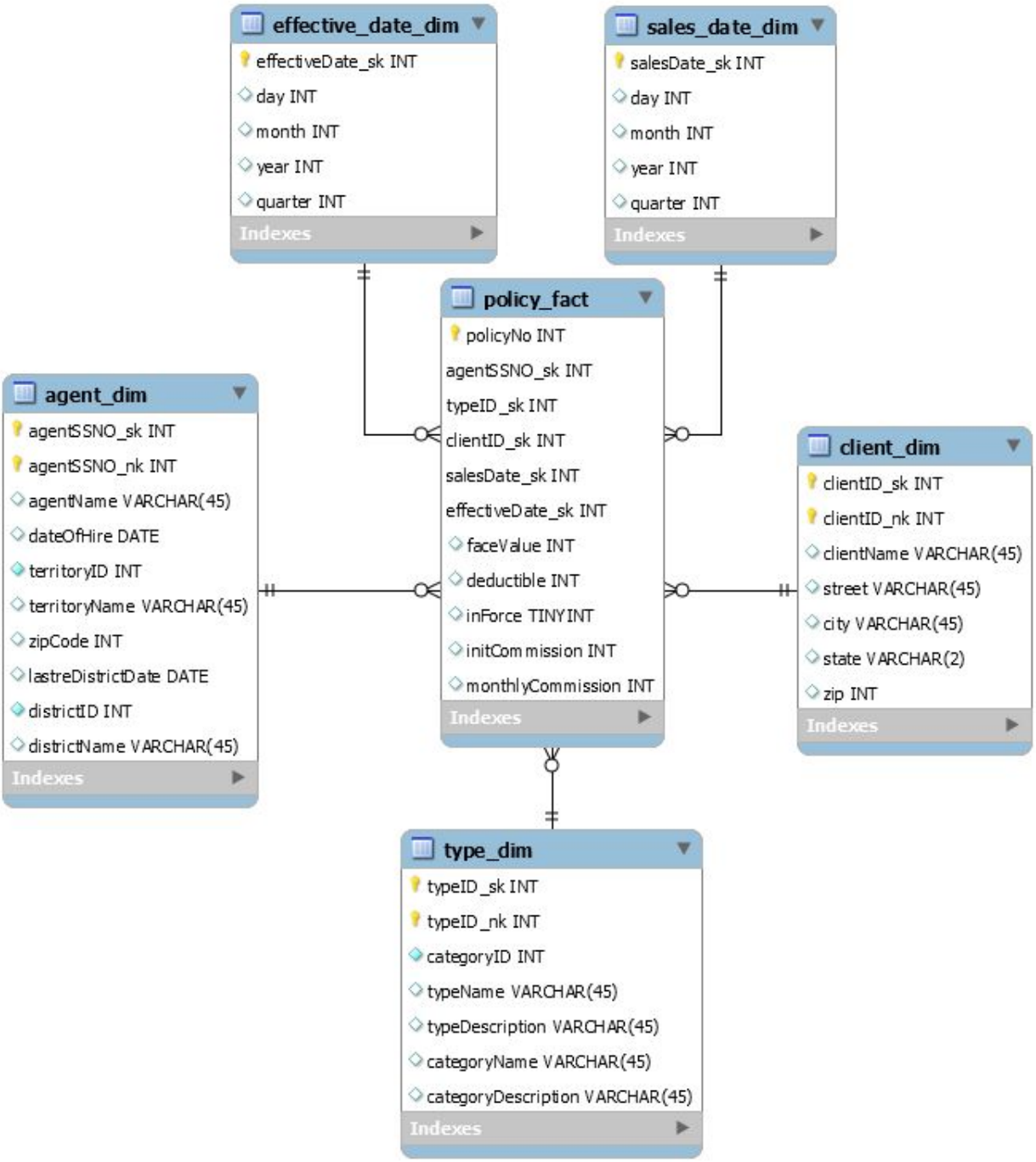
**Figure 1.** *Insurance Policy Sales ERD*




*Question:* Hierarchies are [ modeled separately from ] [ included as dimension attributes in ] the key dimensions in a dimensional model.

*Question:* To be the most meaningful for the user, dimensional attributes should be defined as [ text values ] [ integer values ] .

Step #2-2: *Create your Dimensional Model using MySQL Workbench, save it as a pdf file, and submit it to MyCourses PE02 Drop Box. And bring a hard copy of the pdf file to next class.*



## DW-PE02 Gradesheet

**Name:** _____

| Requirements | Grade | Grade Earned |
|---|---|---|
| **Part #1 – Step #1-1: Identify the Process** | | |
| - Question: User department(s) | 2 | |
| - Choose: | 2 | |
| - Question: Source data system(s) | 2 | |
| - Define: Universe of discourse | 2 | |
| **Part #1 – Step #1-2: Choose the grain** | | |
| - Question: Grain selection | 2 | |
| - Question: Grain options | 2 | |
| - Question: Best practice & Why? | 10 | |
| - Question: level of detail | 2 | |
| **Part #1 – Step #1-3: Identify dimensions** | | |
| - Fill in | 2 | |
| - Question: Business dimensions | 2 | |
| - Question: Hierarchies | 2 | |
| **Part #1 – Step #1-4: Identify dimensions** | | |
| - Fill in | 2 | |
| - Explain: Facts | 2 | |
| - Question: Key performance metrics | 2 | |
| - Question: Attributes | 2 | |
| **Information Package** | | |
| - Has subject | 2 | |
| - At least five dimensions (must have two date dimensions) | 12 | |
| - Measured Facts | 4 | |
| **Part #2 – Step #2-1: Model the star schema** | | |
| - Question | 2 | |
| - Question | 2 | |
| **Part #2 – Step #2-2: Draw dimensional model** | | |
| - Has all dimensions in the Information Package | 15 | |
| - Facts mentioned in the Information Package | 5 | |
| - Role playing dimensions | 10 | |
| - Relationships between fact and dimensions including cardinalities | 10 | |
| **No Hard Copy** | **-10** | |
| **Total Grade:** | **100** | |
| **Graded By:** | | |