AM05 Data Mgmt – Lab 4 – Solution

This lab has two parts. The first part asks you create a star schema suitable for a data warehouse or data mart. The second part ask you to parse text data with regular expressions.

Part 1. DataMart Design

Problem 1. You are to construct a *star schema* for Simplified Automobile Insurance Company. The relevant dimensions, dimension attributes, and dimension sizes are as follows

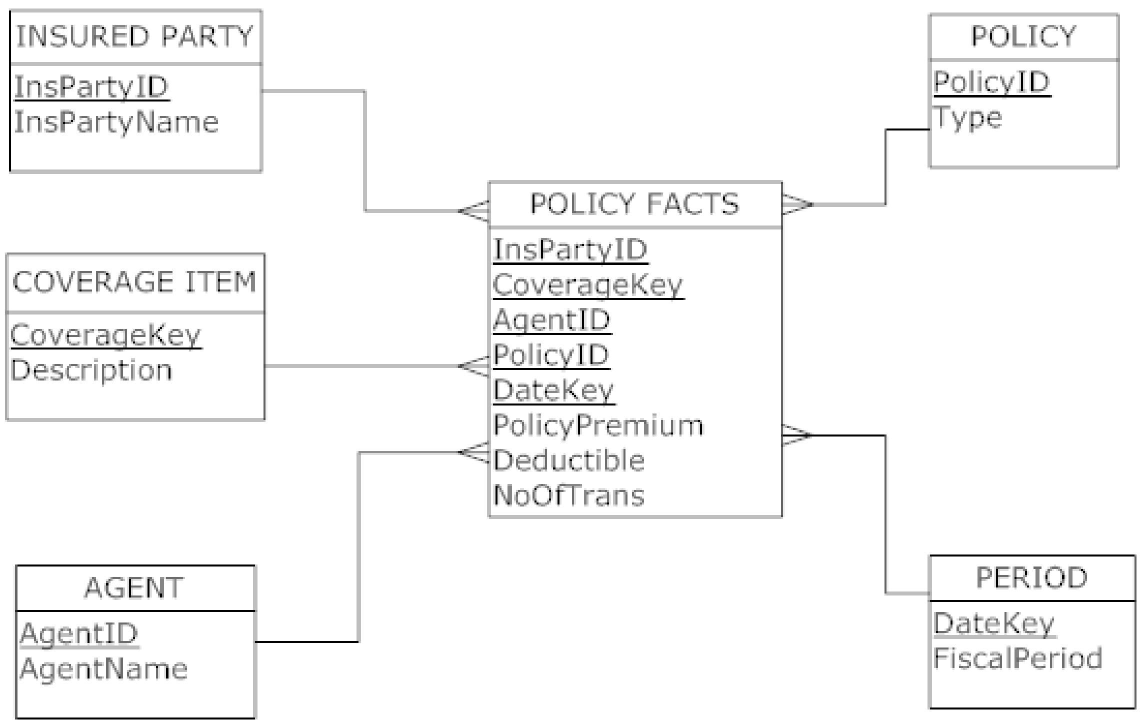
Dimension	Dimension Attributes	Dimension Size comment
InsuredParty	1. InsuredPartyID 2. Name	There is an average of two insured parties for each policy and covered item.
CoverageItem	1. CoverageKey 2. Description	There is an average of 10 covered items per policy.
Agent	1. AgentID 2. AgentName	There is one agent for each policy and covered item.
Policy	1. PolicyID 2. Type	The company has approximately 1 million policies at the present time.
Period	1. DateKey 2. FiscalPeriod	

Facts to be recorded for each combination of these dimensions are PolicyPremium, Deductible, and NumberOfTransactions.

a. Design a *star schema* for this problem.
b. Estimate the number of rows in the fact table, using the assumptions stated previously.
c. Estimate the total size of the fact table (in bytes), assuming that each field has an average of 5 bytes.

Problem 1 solution Simplified Automobile Insurance Company

a. Star schema:



b. To correctly estimate the number of rows, we must make some additional assumptions regarding the fiscal periods and the frequency of changes to policies. We assume the following:

1. The length of a fiscal period is one month
2. The data mart will contain five years of historical data
3. Approximately 5 percent of the policies experience some type of change each month

Therefore, the expected number of rows is:

.05 x 1,000,000 x 10 coverage items x 2 insured parties x 5 years x 12 months per year = 60,000,000 rows

60,000,000 rows x 8 fields x 5 bytes = 2,400,000,000 (2.4 Gigabytes)

9-1. *Simplified Automobile Insurance Company, revised:*

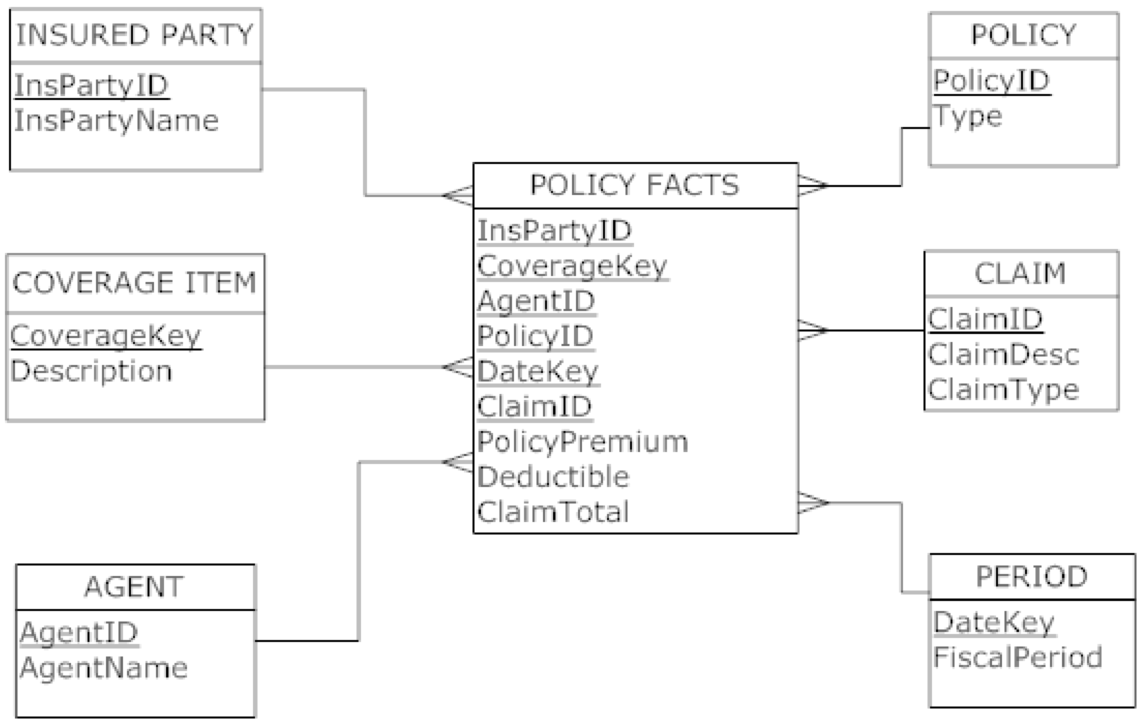
Problem 2. Simplified Automobile Insurance Company would like to add a Claims dimension to its star schema. Attributes of Claim are ClaimID, ClaimDescription, and ClaimType. Attributes of the fact table are now PolicyPremium, Deductible and MonthlyClaimTotal

a. Extend the star schema from the previous problem to include these new data.

b. Calculate the estimated number of rows in the fact table, assuming that the company experiences an average of 2000 claims per month.

Problem 2. Solution

a. An initial version of the extended star schema:



b. Unfortunately, the policy fact table in this schema has a design issue because premium and deductible are not dependent on CLAIM. This would result in two different "grains" or levels of detail in the same table, which is not permitted. A good solution is to create a separate star schema for claim facts (in this case, only ClaimTotal) with all six dimension tables, along with the original star schema. We can then calculate the number of rows in the claims fact table as follows:

60 months x 2,000 claims per month = 120,000 rows in this additional fact table

Part 2. Parsing text data with regular expressions

We started working with text data from a feed of 911 calls from Monroe county New York. I demonstrated in class how to use regular expressions to extract some of the specific data items using regular expressions. Please,

1. Develop regular expressions to extract all the data fields from the raw text data.
2. Use UPDATE queries to place the extracted data in the appropriate columns of the database table specified in the supplied sql file that accompanied the class notes. This will require casting to the data types that match those in the specification of the table.
3. Run some select queries on the data to explore common event types and the proportion of calls responded to by different agencies.

See **SQL.Monroe.911(with.solutions).sql** for possible solution.

Part 2. Parsing Text with Regular Expressions