5-2003

Factors in the Design and Development of a Data Warehouse for Academic Data.

Margaret C. Lester
East Tennessee State University

Factors in the Design and Development of a Data Warehouse for Academic Data

---

A thesis presented to
the faculty of the Department of Computer Science
East Tennessee State University

In partial fulfillment
of the requirements for the degree
Master of Science in Information Systems Science

---

by
Margaret C. Lester
May 2003

Dr. Donald Sanderson, Chair
Ms. Lorie Moffitt
Ms. Kellie Price

Keywords: data warehouse, data warehousing, data models, data warehouse design, academic data warehouse

ABSTRACT

Factors in the Design and Development of a Data Warehouse
for Academic Data

by
Margaret C. Lester

Data warehousing is a relatively new field in the realm of information technology, and current research centers primarily around data warehousing in business environments. As new as the field is in these environments, only recently have educational institutions begun to embark on data warehousing projects, and little research has been done regarding the special considerations and characteristics of academic data and the complexity of analyzing such data. Educational institutions measure success very differently from business-oriented organizations, and the analyses that are meaningful in such environments pose unique and intricate problems in data warehousing. This research describes the process of developing a data warehouse for a community college, focusing on issues specific to academic data.

ACKNOWLEDGEMENTS

I would like to thank my thesis committee for their time, interest, and input in this project. I would especially like to thank Dr. Donald Sanderson for his patience, encouragement, and wisdom.

To my family, thank you all so much for your support. I couldn't have done this without you.

Finally, thank you to the faculty and staff at Northeast State who participated in this project. Your interest and input were absolutely critical to the success of this project.

CONTENTS

Page

LIST OF FIGURES

Figure Page

CHAPTER 1
INTRODUCTION

Data warehousing is a relatively new field in the realm of information technology, and as a result, there is not a wealth of analytical documentation of the data warehousing processes. Current research that does exist centers primarily around data warehousing in business environments. As new as the field is in these environments, only recently have educational institutions begun to embark on data warehousing projects, and little research has been done regarding the special considerations and characteristics of academic data, and the complexity of analyzing such data. Educational institutions measure success very differently from business-oriented organizations, and the analyses that are meaningful in such environments pose unique and intricate problems in data warehousing. Therefore, we need a better understanding of how to develop good data warehouses for academic institutions. This research describes the process of developing a data warehouse for a community college, focusing on issues specific to academic data.

1.1 Background

1.1.1 Definition of Data Warehouse

Bill Inmon, who is recognized as the “father of data warehousing”, defines a data warehouse as “a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision-making process” (Inmon 1996). Ralph Kimball, another noted authority on data warehousing, defines it more simply as “a copy of transaction data specifically structured for query and analysis” (Kimball 1996). Both of these definitions stress that a data warehouse is a collection of data separate from the databases that support the day-to-day operations of an organization, and that it is specially designed for the purpose of producing reports. Therefore, the process of designing a data warehouse is somewhat different from the process of developing an on-line transaction

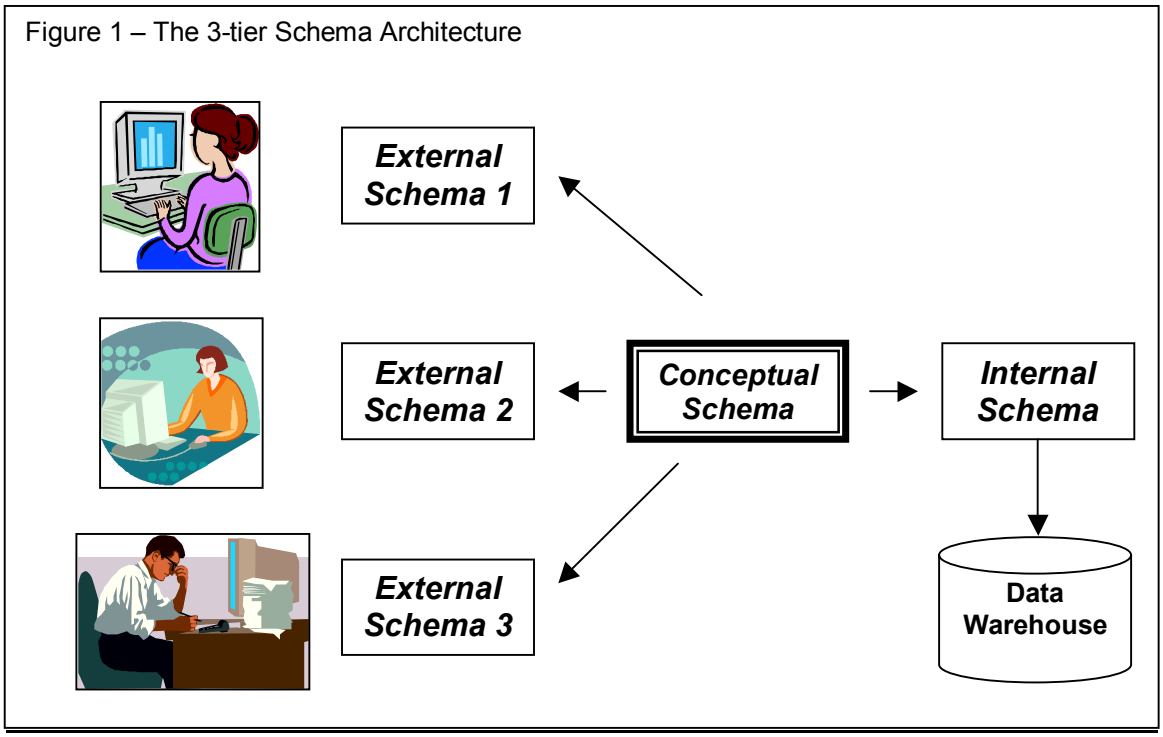
processing (OLTP) database. The next section explains some of these differences in terms of design processes.

1.1.2 Designing a Data Warehouse

Data models have long been used as a communications tool between database developers and the user community. Data models help developers to visualize the structure of a business or organization, but they are descriptive enough for end users to understand so that they can ensure that the model does indeed reflect their business requirements. Figure 1 shows the three-tier architecture of data processing systems, as defined by Hay (Hay 1997). The data models are the primary component of the conceptual level, serving as the translator between the external views of the end users and the internal implementation that is the physical layer.

In OLTP systems, data models describe the business operations in terms of small, predefined transactions. These models tend to be rather complex, and

Figure 1 – The 3-tier Schema Architecture



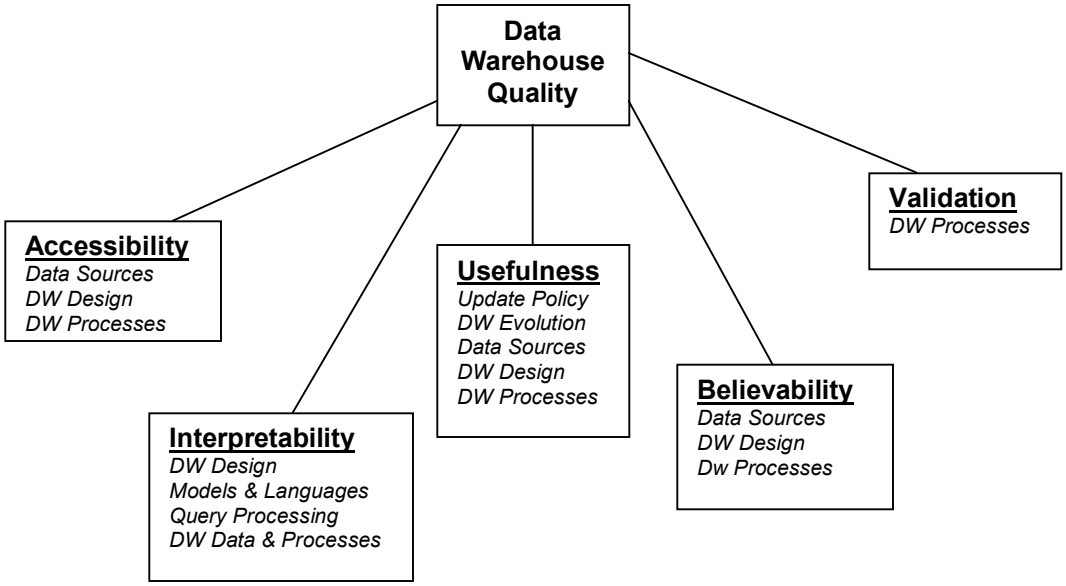
the end users rarely interact directly with them, relying instead on IT personnel to process reports for them. In contrast, data models for data warehouses, also called on-line analytical processing (OLAP) systems, are considerably simpler, because the intent is for end users to interact with the models to produce their own reports. Creating these simplified models requires careful analysis of the data sources in order to discover data that will be most useful for analysis and then organizing that data so that it is easy to access and understand.

1.1.3 Evaluating the Data Warehouse

The purpose of having a good design is to have a product at the end that will be of benefit to the organization. The best judge of whether or not a data warehouse is "good" is the end user. Therefore, the best way to evaluate the quality of the data warehouse is through customer satisfaction surveys. To develop these surveys, we need to understand the factors that measure the quality of a data warehouse, and the components of the warehouse that impact those factors.

The Foundations of Data Warehouse Quality project (DWQ), a 3-year ESPRIT research project (Jarke and Vassiliou 1997), set out to develop a quality framework that is specifically mapped to the data warehouse architecture. Figure 2 shows the five quality factors that were used as the framework for DWQ, and the components of the data warehouse that impact each of them. To evaluate the quality of the data warehouse, we must develop a set of baseline performance requirements, define the metrics to measure the properties of the data warehouse components, and then formulate tests to relate the baseline goals to the metrics.

Figure 2 – The DWQ Quality Framework (adapted from (Jarke and Vassiliou 1997)



1.2 Description of Research

This research will focus on a data warehousing project for Northeast State Technical Community College to support analysis of the Developmental Studies Program. Data will be extracted primarily from the Student Information System (SIS) (SCT Corp. 1997) covering a five-year period. While many factors in the construction of a data warehouse affect the quality of the end product, this research will focus on the data modeling aspect of data warehouse planning and construction. Specifically, it will look at two methods of constructing data warehouse models in an attempt to determine which is most effective for the source data that we have.

The first phase of the research will consist of designing the data warehouse itself. This will be accomplished by analyzing the source data to develop the data models. Then the data models will be used to design the data warehouse tables. Next, data must be extracted and loaded into the data warehouse tables. Finally, the reporting tools will be configured and end users will be trained to produce reports.

The goal of the next phase of research is to develop the baseline measurements to evaluate the quality of the data warehouse. End users will be surveyed to determine their expectations for the warehouse. These measurements will be used in the final phase to develop an evaluation instrument, which will include a customer satisfaction survey to determine if the end product is a quality data warehouse that can provided needed information to the college.

The sections that follow will detail the processes followed and the findings of each of these phases. Section 2 explains the development method that was followed to design and develop the data warehouse, including considerations specific to academic data that proved to be major factors in the design. Section 3 describes the development of the end user survey and the results of that survey, and section 4 shows how the evaluation tool was developed from the survey results, as well as the results of the evaluation. The conclusions of the research and future work are detailed in the final section.

CHAPTER 2
DESIGN AND DEVELOPMENT OF THE DATA WAREHOUSE

2.1 Modeling Multidimensionality

The most common representation of the OLAP data model is a cube because this visualization is the easiest for end users to understand. Each side of the cube represents a dimension of the data being analyzed. The cells of the cube contain the facts relating to the intersection of the dimensions. Table 1 is a simple analysis of grade distributions for courses over a number of semesters. Figure 3 shows how this same information would be represented in a cube. In contrast to the traditional normalized relational model used for databases, the dimensional model is denormalized for reporting speed, as it minimizes the number of table joins required to satisfy queries. Common cube operations are (Kelly 1998):

- Drill-down: zooming in for more detail
- Roll-up: aggregating or summarizing data
- Slice-and-dice: focusing on a specific attribute of a dimension
- Pivot: rotating from one dimension to another

Table 1 – Grade Distribution Table

Course	Semester	A	B	C	D	F	W
MATH0700	00F	53	69	23	14	8	9
	01S	48	58	32	12	11	3
	01U	24	10	12	4	3	1
	01F	67	72	51	10	17	14
MATH0800	00F	30	44	26	12	5	6
	01S	46	24	13	16	7	4
	01U	13	25	27	3	8	3
	01F	38	51	35	14	6	15
ENGL1100	00F	63	38	19	3	4	5
	01S	41	12	28	15	2	7
	01U	24	6	7	4	3	2
	01F	74	22	27	9	6	3

The logical choice for a physical database in which to store this multidimensional model would seem to be the multidimensional database (MDDB). Indeed, MDDB's produce faster query results and provide a logical grouping of data elements and their hierarchies that are easier for end users to understand. However, MDDB's suffer from several significant problems that can make them impractical for an entire data warehouse. First, cells are "reserved" at every intersection of the dimensions, but there may not be data to fill them, resulting in a "sparse" database (Kelly 1998). For example, if we replaced the "Grades" dimension in Figure 3 with a "Student" dimension, and stored student grades in the cells (see Figure 4), very few of the cells would actually contain data, because a single student would only be enrolled in very few of the courses offered by a college during a single semester. In reality, the data would be summarized somewhat, as in Figure 3 before being stored in the MDDB. The disadvantage to this solution is that we have limited some of the options of rearranging the data to expand the types of analysis that can be performed. For example, if we go ahead and aggregate the student data into the "Grades" dimension, we can no longer analyze data based on student attributes such as gender or classification.

Figure 3 – Grade Distribution in a Cube

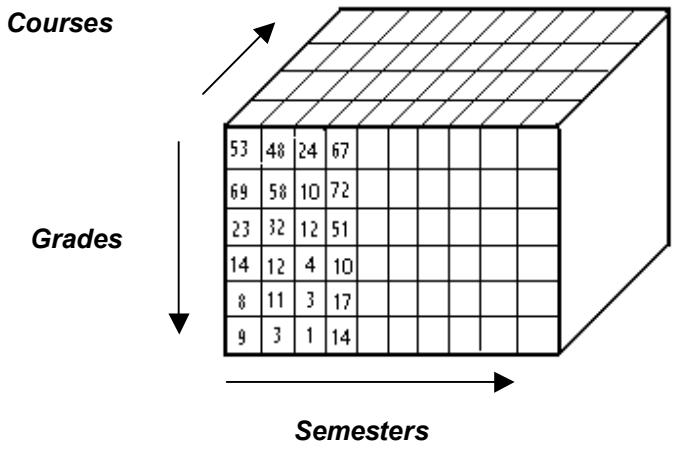
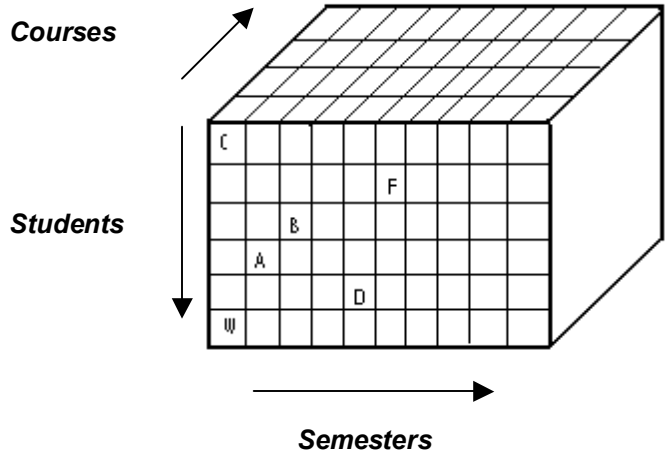


Figure 4 – Multidimensional Cube Without Summarized Dimension



Secondly, most MDDB's have a limit to the number of cells and/or dimensions that can be managed (Kelly 1998). Further, the time dimension, which is always required in a data warehouse, can add orders of magnitude of complexity to the database, particularly if many different measures of periodicity are required. MDDB's can be useful for storing the external views of data marts, where requirements are well-known in advance and users will not be performing frequent drill-down operations. However, the relational database structure provides the flexibility and scalability required for most data warehouses. The problem, then, is how to model the multidimensional views understood by the users in two dimensions.

2.2 Star Schema

The model traditionally used for relational database representation is the entity-relationship (ER) model. However, ER models tend to be complex because they are designed for OLTP systems in which data integrity concerns and transaction speed requirements make data redundancy highly undesirable. In data warehouses, however, query speed is most important and data integrity is

already insured by the data cleansing and transformation processes. Because users will be interacting directly with the model, a design that is easier to understand is necessary. The compromise is the star schema.

The star schema contains many of the components found in ER models: entities, attributes, relationship connections, min-max values, optionality, and primary keys. However, the star schema is a simpler, denormalized structure, and numerical attributes, called “measures”, are generally restricted to the central fact table. Figure 5 shows an ER diagram for a grocery store chain database.

Figure 5 – ER Diagram for Grocery Store Chain Database (adapted from (Kelly 1998))

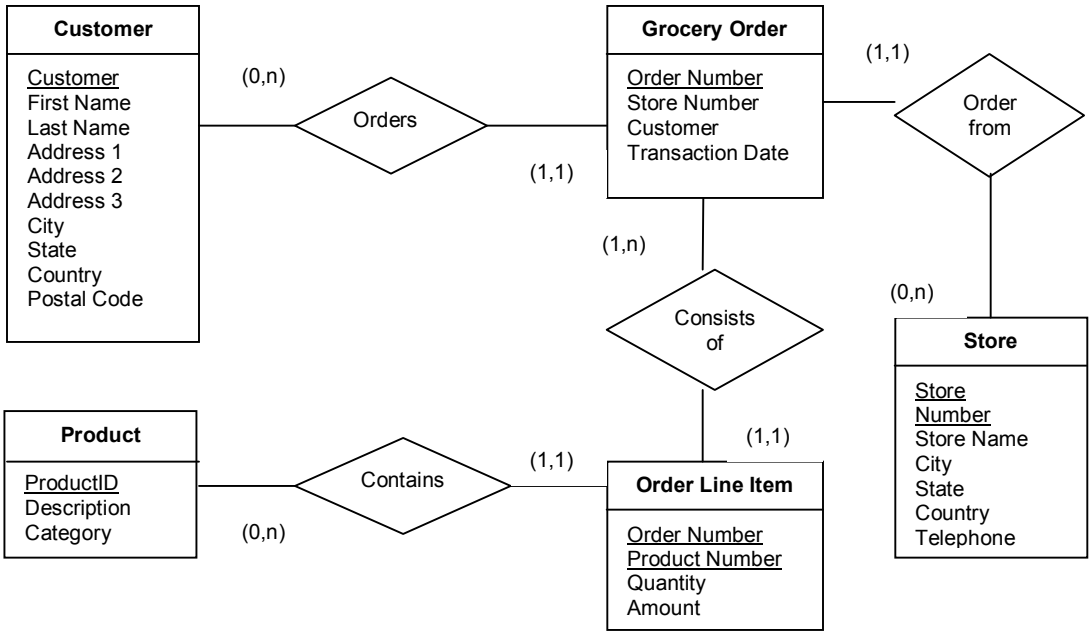
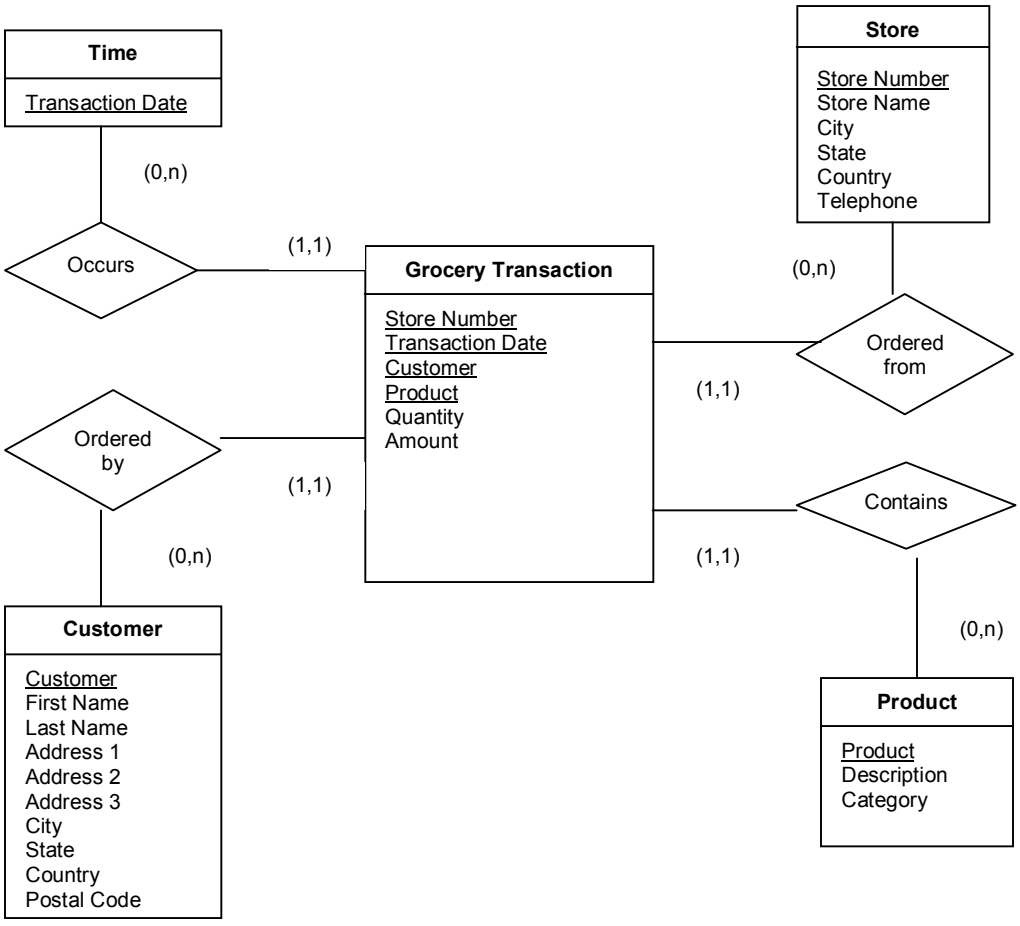


Figure 6 is a star schema that can be constructed from the original schema to support a sales analysis. In some cases, it may be desirable to provide a more normalized format to give users the ability to view even more detailed information. The snowflake schema adds hierarchies to the dimensions and eliminates some of the data redundancy. Snowflake schemas are especially

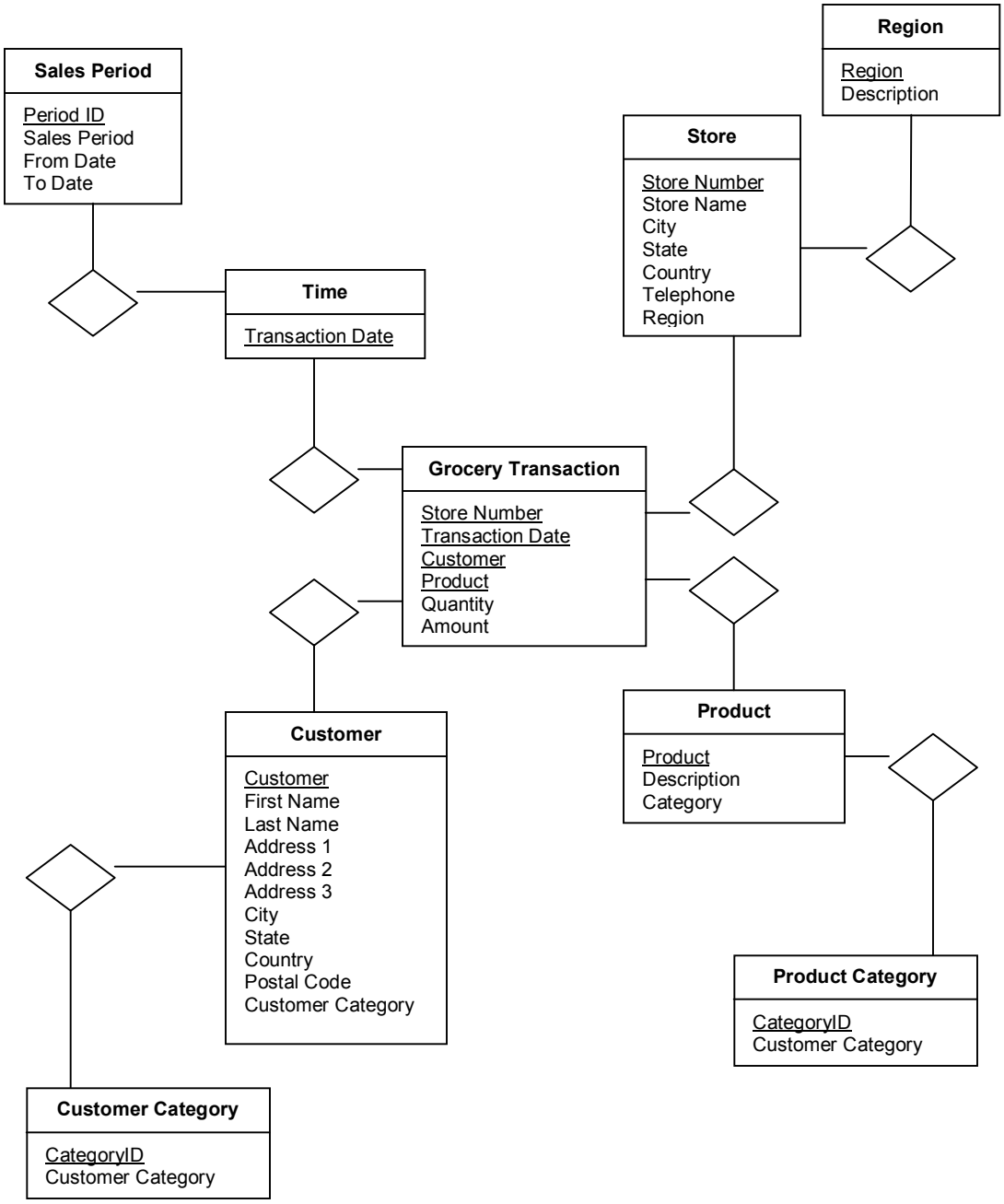
Figure 6 – Star Schema Design from ER Diagram (adapted from (Kelly 1998))



desirable if the redundancy in the original dimension results in an unacceptable storage space burden (Kelly 1998). Figure 7 expands the star schema for the Grocery Store to the more normalized snowflake schema.

While dimensional modeling using the star and snowflake models is generally regarded as the best solution by most data warehouse experts, it is not without its pitfalls. First, denormalization is risky in situations where data will be changing frequently. This is not usually of great concern for data warehouses, because the data being used are static, historical data, but should always be considered. Also, denormalized models tend to be less flexible, which puts even greater emphasis on careful requirements gathering. Secondly, while it is not

Figure 7 – Snowflake Schema Example (adapted from (Kelly 1998))



unusual for the fact table to be very large, the dimension tables should generally be small. In a query operation on the star schema, the dimension tables are joined first to produce the intermediate result table. Then that table is processed against the fact table. If the intermediate table is very large, due to one or more large dimension tables, performance is impacted severely (Kelly 1998).

A third issue that can impact performance, as well as understandability, is "snowflakiness". While the more normalized structure of the snowflake model may be less disconcerting for the traditional database designer, too much of it adds complexity to the model from the user's perspective and reduces the benefits of the star schema by increasing the number of joins required for queries. Finally, data sharing may be an issue in some data warehouse implementations. Assuming that a single star schema represents a data mart, each star schema may be implemented in its own database. If this is the case, then any dimension that is used in other data marts, such as the time dimension, must be duplicated in each of those databases as well, with all of the associated risks (Kelly 1998).

2.3 The Methods

There is disagreement in the data warehouse community about the best method to use for designing the warehouse architecture. The approaches recommended generally fall into one of two categories. The first method involves meticulous analysis of the source data and then constructing models for a comprehensive data warehouse. This method will be referred to as the "bottom-up" method. The second approach, which we will call the "top-down" method, looks at potential data marts that will be needed for the overall data warehouse and then designs data models to support the data marts. The data warehouse is considered complete when all of the data marts have been constructed. The details of each method will be outlined in the sections that follow.

2.3.1 The Bottom-Up Approach

The bottom-up method described in (IBM 1998) begins with a clear understanding of the conceptual view of the business as a whole and then selects the parts of that view that will be useful for business analysis in the data warehouse. Table 2 outlines the six steps of the bottom-up method. The first

Table 2 – The Bottom-Up Method

- **Step 1: Create ER diagram of data source(s)**
- **Step 2: Remove associative entities and subtypes**
- **Step 3: Merge entities in M:N relationships**
- **Step 4: Identify parts of the model to be used in DW**
- **Step 5: Identify the measures and dimensions**
- **Step 6: Create facts**

step is to create the ER diagram of the source data, if one does not already exist. Then, all associative entities and subtype entities are removed from the diagram. Next, the model is simplified by rolling up entities at the ends of many-to-many relationships into single entities. As entities are eliminated, any attributes associated with them that would be useful for analysis should be retained.

Once the model has been simplified, the designer works with business executives and analysts to determine what parts of the model will be valuable in the data warehouse, as well as identifying information that is not currently included in the model. At this point, the dimensional model can be constructed. This is done by identifying the measures that will make up the fact table and the dimensions that are desired and adding a time dimension. The next step is to create the facts. The designer can determine the desired level of detail, or granularity of data, from the requirements. One of the factors to consider in this step is the additivity of the measures. Additivity is the ability of a measure to be summarized. Some measures, such as percentages, are non-additive, as they cannot be summarized across the time dimensions to produce meaningful information. In general, all measures should be fully additive. If they are not, they may need to be broken down further. After all of the facts have been

created, they are merged into a single fact table and linked to the dimension tables, completing the star schema.

2.3.2 The Top-Down Approach

The top-down design method begins by looking at the external user views of the organization by studying query behavior. The potential data marts are then identified by creating materialized views that can satisfy the queries. A materialized view is defined as a preprocessed query that is stored for faster retrieval. A carefully designed materialized view can serve many different queries. There is currently much research in the academic community involving identifying and constructing materialized views, particularly views that are self-maintaining. To build a data mart, materialized views are analyzed to discover the common facts and dimensions being used in the queries (Kimball 1999). The next step is to construct a matrix of the potential data marts (fact tables) and the dimensions they would have in common (Moriarty 1995). Figure 8 shows a sample matrix for proposed academic data marts.

Figure 8 – Matrix of Proposed Facts and Dimensions

Data Marts	***Dimensions***					
	Time	*Course*	*Student*	*Instructor*	*Inventory*	*Payroll*
Grade Distribution	X	X	X	X		
Faculty Work Load	X	X		X		
Cost Study	X	X	X	X	X	X
Room Utilization	X	X			X	

At this point, it may be necessary to have user groups meet to resolve any conflicts in business rules or definitions that might appear in the dimension tables. For example, the Admissions and Records department will view data primarily by academic calendars, while the Business Office will view their data on

a fiscal year basis. In such situations, a compromise must be reached, whether it is a common calendar, or an agreement that both calendars will be represented in the dimension table. Once all conflicts have been resolved, the star schemas are designed using the matrix. Finally, the source data are analyzed to locate the information that will be needed to populate the data marts.

2.3.3 The Better Approach?

Each of these approaches has its advantages and disadvantages, but it is not clear which approach produces the better quality data warehouse. The bottom-up approach is generally viewed as being more flexible and scalable, because the foundation is laid for adding more functionality to the data warehouse by building the entire conceptual schema first. Advocates of the top-down method, however, believe that the quicker implementation time of their method outweighs the benefits of analyzing source data that may never be used (Moriarty 1995). The top-down approach also tends to produce a more streamlined data warehouse, since it does not contain any extra data that end users are not currently accessing in their queries.

2.3.4 Comparison of Methods

The first step in comparing these methods will be to design the data models using the bottom-up method, and then construct a data warehouse based on those models. Then the reports generated from the data warehouse will be compared with current reports from the SIS system. The objective here will be to discover discrepancies between the data warehouse reports and the reports currently being used, because those reports would be the basis of the top-down modeling approach. The next section describes the processes followed to design the data warehouse using the bottom-up method.

2.4 Design and Development Using the Bottom-Up Method

This section describes in detail the processes used to design and develop a data warehouse to support analysis is the Developmental Studies Program (DSP) at Northeast State. This program provides pre-college level training for students who have not demonstrated the necessary skills in the areas of mathematics, reading, and writing to enable them to perform well in college level classes. Readiness is determined by various test scores, such as the ACT and SAT, as well as performance on specialized academic placement tests administered by the college. College administrators are interested in evaluating the success of the program from many angles, such as student and instructor demographics. Success will be measured by student performance in DSP courses, performance in subsequent college-level classes, and graduation rates. Also, DSP student performance will be compared with the performance of students not in the program. The analysis will span a five-year period, from the Spring semester of 1997 through the Fall semester of 2001. The next section describes the process followed to analyze the source data and design the data models.

2.4.1 Analyzing the Source Data

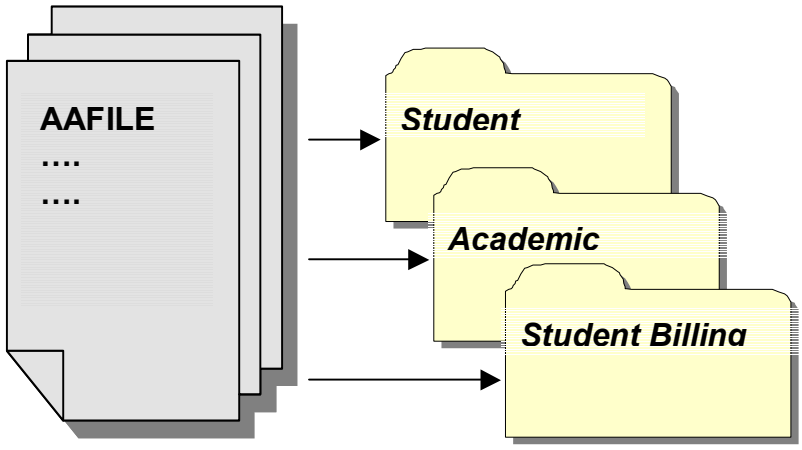
The bottom-up method requires that the source data be modeled in a relational database model – the Entity-Relationship (ER) model. Because such a model does not exist for the source systems, the first step will be to design one. The database system used for all of the administrative operations of the college is a flat file system. It actually consists of three separate systems – the Student Information System (SIS), the Human Resources System (HRS), and the Financial Records System (FRS) – and these systems do not connect with other, except via batch processes. We will only be using the files needed to support the DSP analysis, but our model should be flexible enough to allow for future expansion.

2.4.1.1 High-Level Analysis.

The source data analysis turned out to be an iterative process, looking at the data sources in increasingly finer detail. The first step was to identify the source files that contain the data necessary to support the analysis of the Developmental Studies Program. Initially, three files from SIS were selected, but other files were added as the analysis progressed. The three initial files were the Student Demographics file, the Course Term file, and the Student Term file.

The next step was to look at the internal structure of these files. In general, all of the files in the system have a similar structure: uniquely keyed records with one of more embedded arrays (see Figure 9), referred to as "segments". The segments are usually, but not always, directly related to the subject of the file. For example, the Student Term file contains segments for the student's current academic programs (majors and concentrations) and courses the student is enrolled in for the term, but also contains a segment for financial aid information. Another characteristic of these files is that data elements (the equivalent of a column in a relational database) are frequently duplicated in

Figure 9 – Student Demographic File Format



different files. For example, the Demographics file stores up to 16 instances of academic programs that a student has been enrolled in at the institution. Much of this information, such as the major and department offering the major, is duplicated in the Student Term file, which stores the student's academic program for each term. Also, some of the files have space designated as "user-defined filler" that can be used to store essentially anything the institution wants to store. As a result, data may be stored within a segment that does not necessarily match the "theme" of that segment, but are nevertheless data that is important to the institution and possibly useful for analytical processing.

2.4.1.2 Lower-Level Analysis – The Data Elements.

Some implementations of the bottom-up strategy propose representing every data element present in the source data in the ER schema. However, if there are many data elements in the source files that are not being used, the resulting data warehouse can become "overstuffed" and inefficient. Initial analysis of our source files indicates that many of the fields are currently unused or are inappropriate for analytical processing. Therefore, in order to make a more efficient, understandable, and manageable warehouse, we will stray from a "pure" bottom-up approach. Only data elements that (1) are currently being used and (2) could be useful for analysis will be included in the data warehouse schema. The next task, then, is to look at each data element in each file and decide whether or not to warehouse it. For each of these elements, the following questions were asked:

- Do we use this element?
- Have we been using it for the past 5 years?
- How have we used it (i.e. has it been consistently, correctly entered in the system)?
- If the data element is stored in more than one place, which copy do we consider to be most correct?

These questions establish, in essence, the “business rules”, often cited as a key component of a successful data warehouse.

In order to complete this task, input from the end users was required. The challenge here was to find a common ground on which to communicate regarding the individual elements, because the end users do not recognize the data elements in the context of the files they reside in, but rather by the data entry screens on which they appear. To facilitate this process, a table was designed to record all of the information about each data element:

- The name of the element in the source file definitions
- The element ID in the SIS system
- The data type
- The screen(s) on which the element appears
- Valid values or ranges for the element
- A description of how the element is used in the operational system
- An indicator that the element is suitable for warehousing
- The dimension that the element will be assigned to in the warehouse
- Any other comments regarding special handling of the element for the data warehousing process

The following section describes the most significant findings of the analysis and design decisions that were made as a result of those findings.

2.4.1.2.1 Duplicate Values.

As noted in the previous section, some of the data elements related to a student’s academic program are stored in two separate files: the Student Demographics file and the Student Term file. The major difference between the two is that the Student Demographics file information represents the academic program data at the end of the program (or as of the current date), while the Student Term file represents a snapshot of the data for each term. Because the information in the Student Demographics file can be derived from the information

in the Student Term file, the Student Term file values were chosen for the data warehouse.

In another instance of data duplication, many of the values in the Course Term file are duplicated in the Student Term files, such as the College offering the course, Type Credit, RD Area and Level, and Session Code. In reality, this duplication of the values in the Student Term file is incorrect and can cause problems in calculating a student's grade point average. For example, a course that has been coded as a Developmental Studies course (pre-college level), as indicated by RD Area and Level, is not to be used in calculating a student's grade point average (GPA). The duplication of the data element in the Student Term file permits a data entry operator to code the course differently, for example, as a college-level course, for an individual student. This would cause the course to be included in the GPA calculation. Because these data elements describe a course, rather than an instance of a student taking a course, the Course Term file values will be used.

2.4.1.2.2 Unreliable Data.

One of the major analyses that we want to be able to do is comparison of student performance based on the full-time or part-time status of the course instructor. The SIS system as delivered does not provide a data element to indicate if the course instructor is full-time or part-time. To provide this capability, an unused field in the Course Term file was redefined to store an "A" to indicate a full-time instructor and a "B" to indicate part-time. One problem with this is that the data element is an attribute of the course, rather than the instructor, so that an instructor could be coded as full-time for one course and part-time for another. Although steps have been taken to reconcile these data, this has only occurred during the past two to three years, and data prior to that time are highly unreliable. The solution, then, was to try to find a more reliable source for this information. The Human Resources System (HRS) was the most likely

candidate, because information about an individual's employment status must be reliable for state and federal reporting requirements. This information is stored in the form of a job code that is assigned to each employee for a particular time period of his or her employment contract. The primary drawback to this solution is that the HRS system is based on calendar and fiscal years, rather than semesters, which may complicate the time dimension somewhat. A secondary advantage, however, is that the name field in HRS is much more reliable as well. Two files from HRS will be used to extract this data: the Assignment file and the Employee Data file.

It should be noted that this is a significant finding in favor of the bottom-up approach. The top-down method uses the approach of analyzing current queries and then using those to build the data warehouse views. Currently, we do not have a means to "join" files from SIS with those in HRS, which is what we effectively are doing here. Instead, we are using the SIS full-time/part-time indicator in all queries where the instructor's employment status is of interest. As a result, many of the reports using this value may be inaccurate.

2.4.1.2.3 Preservation of Anonymity.

Federal regulations regarding confidentiality of student records present a somewhat unique challenge to educational institutions when it comes to designing analytical processing systems. While these regulations are moving toward disallowing the use of Social Security numbers for any other type of identification, the SIS and HRS systems currently use the SSN as the Student ID and Employee ID, respectively. Also, the data warehouse will provide users with drill-down capability to view detailed data, and care must be taken to ensure that restrictions on data access that are present in the source databases are preserved in the data warehouse. One measure that will be taken to preserve confidentiality is the use of system-generated ID's in the data warehouse to replace SSN's. These ID's must be unique across the data warehouse because

there will be instances where an employee is also a student. A second measure to be taken is that individual names will not be stored in the data warehouse views, and users with restricted access to the source systems will only have access to appropriate views and not to the underlying tables.

2.4.1.2.4 Storing Calculated Values.

Several data elements stored in the SIS system are actually calculated values, such as the credit hours earned, quality points earned, and grade point average. These values appear primarily as attributes of the Student Term entity. The decision we needed to make was whether or not to recalculate these values before storing them in the data warehouse. Although recalculating may be the more accurate option, in this case we are more interested in representing the "official record". Therefore, the stored calculated values will be extracted directly from the SIS system and loaded into the data warehouse without manipulation.

2.4.1.2.5 Other Findings.

Organizations involved in manufacturing and sales are able to measure their success based on concrete numerical analyses. Educational institutions measure their success by less tangible comparisons. The DSP is successful if a student who enters college unprepared for college-level studies leaves the program with the necessary skills to perform college-level work. Measuring this requires a baseline for each student, which is established by scores on standardized tests. At Northeast State, a student's initial placement into Developmental Studies courses is determined by their performance on ACT and/or SAT standardized tests. If a student did not take either of these, or the scores were no longer valid, he or she may have been required to take another placement test designated by the college. Storing these scores requires the inclusion of two additional source files from the SIS system: the Developmental Studies Placement file, which stores DSP placement data, and the Basis of

Admission file, which stores standardized test data and high school diploma information.

2.4.2 Designing the Entity-Relationship Model

Once the source file analysis was completed, we had a better understanding of how we view the data in our databases and which data elements we consider to be our sources of "truth". We recorded the "warehousability" of each element as part of the analysis and also assigned each element to a prospective entity. The table can now be used to construct the ER diagram, shown in Figure 10.

The entities fall into two groups – those that are static, shaded green in the diagram, and those that change from one semester to the next, shaded pink. Also, many of the attributes are actually codes that are translated by other entities. Because the inclusion of all of these entities adds unnecessary confusion to the diagram, these are noted by an asterisk (*) and a foreign key designation (FK).

In addition to providing a visual representation of our data, the ER diagram will serve as the model for constructing a physical database to serve as the staging area for the data warehouse. Staging areas are commonly used in data warehousing to facilitate the processes of extracting the data from the sources, cleansing the data, and preparing it for loading into the data warehouse structures (commonly referred to as ETL processes). The physical database, which we will refer to as the "Base Warehouse", consists of 41 tables, including the translator tables described in the previous paragraph. After the programs and procedures were written to load these tables, we began the process of designing the data models required to support the specific reports we need.

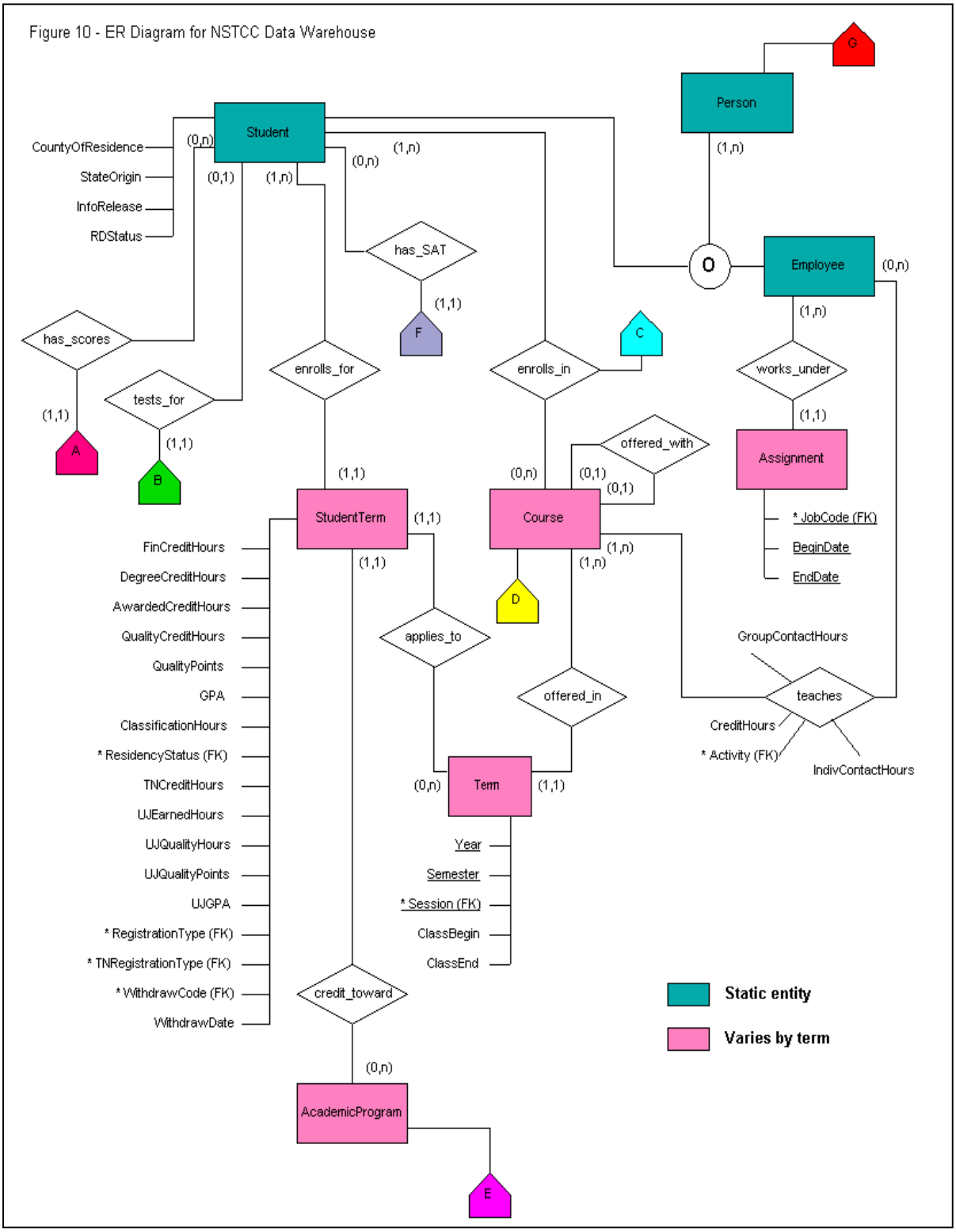


Figure 10 - ER Diagram for NSTCC Data Warehouse
Person
G
Student
CountyOfResidence
StateOrigin
InfoRelease
RDStatus
(0,n)
(0,1)
(1,n)
(0,n)
(1,n)
has_SAT
(1,1)
F
O
Employee
(0,n)
(1,n)
has_scores
(1,1)
A
tests_for
(1,1)
B
enrolls_for
enrolls_in
C
works_under
(1,1)
Assignment
offered_with
(0,1)
(0,1)
(0,n)
(1,1)
StudentTerm
(1,1)
Course
(1,n)
(1,n)
D
* JobCode (FK)
BeginDate
EndDate
FinCreditHours
DegreeCreditHours
AwardedCreditHours
QualityCreditHours
QualityPoints
GPA
ClassificationHours
* ResidencyStatus (FK)
TNCreditHours
UJEarnedHours
UJQualityHours
UJQualityPoints
UJGPA
* RegistrationType (FK)
* TNRegistrationType (FK)
* WithdrawCode (FK)
WithdrawDate
applies_to
offered_in
GroupContactHours
teaches
CreditHours
* Activity (FK)
IndivContactHours
(0,n)
Term
(1,1)
Year
Semester
* Session (FK)
ClassBegin
ClassEnd
credit_toward
(0,n)
AcademicProgram
E
Static entity
Varies by term

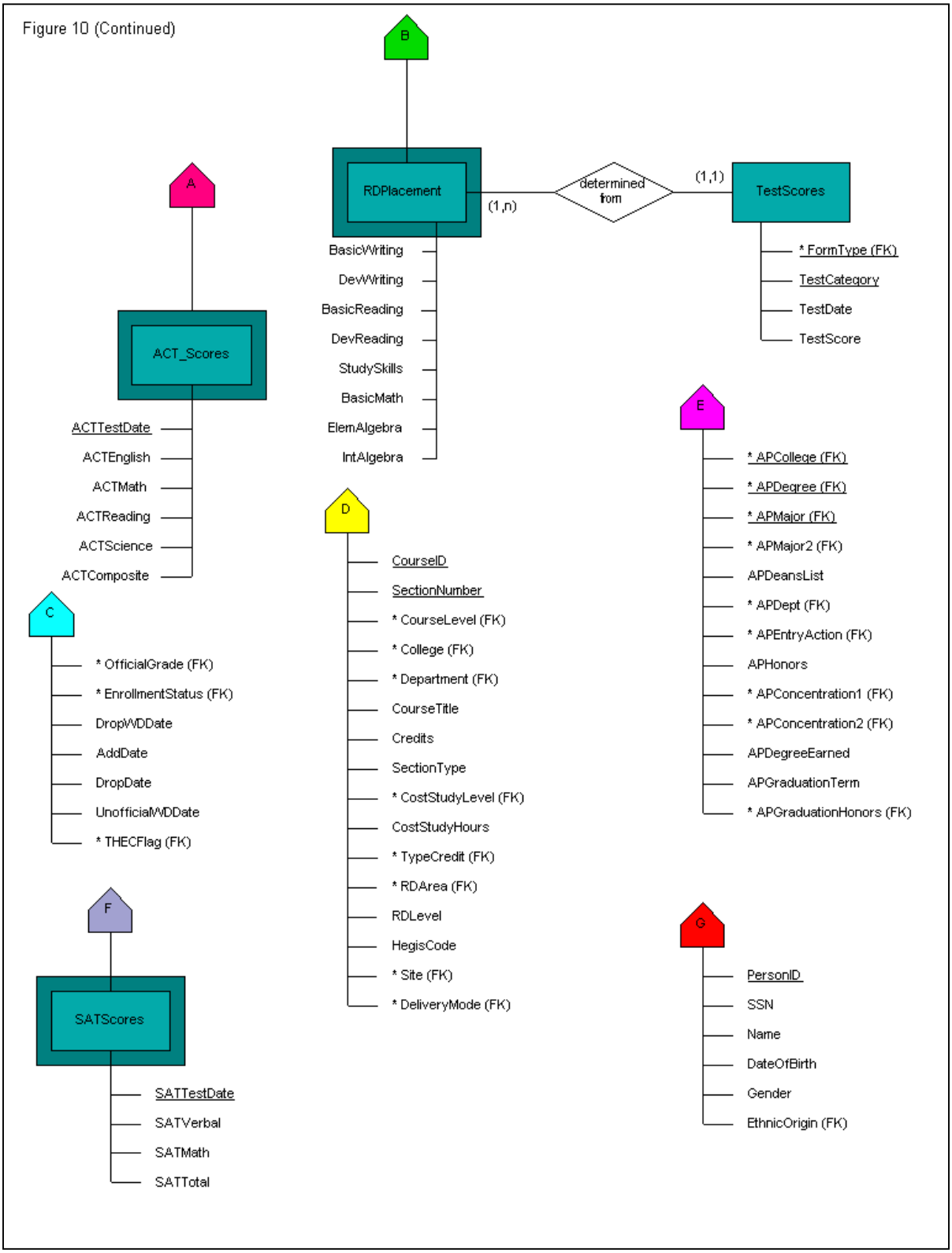


Figure 10 (Continued)
B
RDPlacement
(1,n)
determined from
(1,1)
TestScores
* FormType (FK)
TestCategory
TestDate
TestScore
BasicWriting
DevWriting
BasicReading
DevReading
StudySkills
BasicMath
ElemAlgebra
IntAlgebra
A
ACT_Scores
ACTTestDate
ACTEnglish
ACTMath
ACTReading
ACTScience
ACTComposite
C
* OfficialGrade (FK)
* EnrollmentStatus (FK)
DropWDDate
AddDate
DropDate
UnofficialWDDate
* THECFlag (FK)
D
CourseID
SectionNumber
* CourseLevel (FK)
* College (FK)
* Department (FK)
CourseTitle
Credits
SectionType
* CostStudyLevel (FK)
CostStudyHours
* TypeCredit (FK)
* RDArea (FK)
RDLevel
HegisCode
* Site (FK)
* DeliveryMode (FK)
E
* APCollege (FK)
* APDegree (FK)
* APMajor (FK)
* APMajor2 (FK)
APDeansList
* APDept (FK)
* APEntryAction (FK)
APHonors
* APConcentration1 (FK)
* APConcentration2 (FK)
APDegreeEarned
APGraduationTerm
* APGraduationHonors (FK)
F
SATScores
SATTestDate
SATVerbal
SATMath
SATTotal
G
PersonID
SSN
Name
DateOfBirth
Gender
EthnicOrigin (FK)

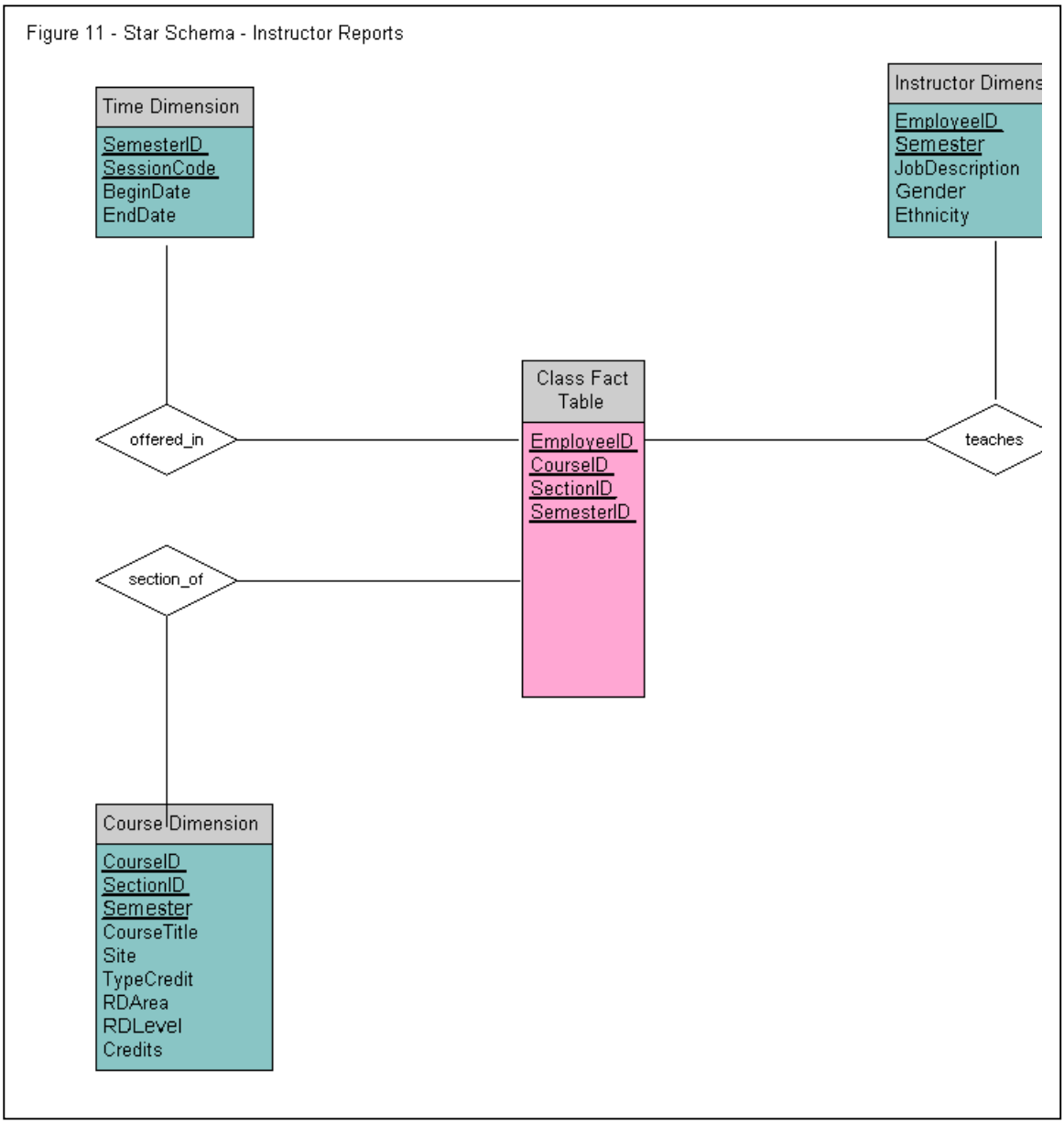
2.4.3 Designing the Star Schema

The first step in the process of designing the star schema for the data warehouse is to identify the measures and dimensions. For the measures, we need to determine the level of detail, or granularity, of data that we need that will provide the most flexibility. The reports that we will be producing actually use a variety of measures, so the objective is to define those measures and determine how they can fit into one or more fact tables.

2.4.3.1 Instructor Analysis Reports.

The first set of reports calculates the counts and percentages of part-time and full-time instructors by various categories. This measure will not actually be stored in the fact table, because we are counting the number of joins, but will be calculated at query time. As far as the dimensions, we need to know if instructors are full-time or part-time, so we will need an Instructor Dimension that has as an attribute the job description. We also want to view the counts by course type (DSP vs. non-DSP) and location, so we need a Course Dimension that stores those attributes. Finally, we will need a time dimension so that we can view the data by individual semesters. The star schema for this set of reports is shown in Figure 11.

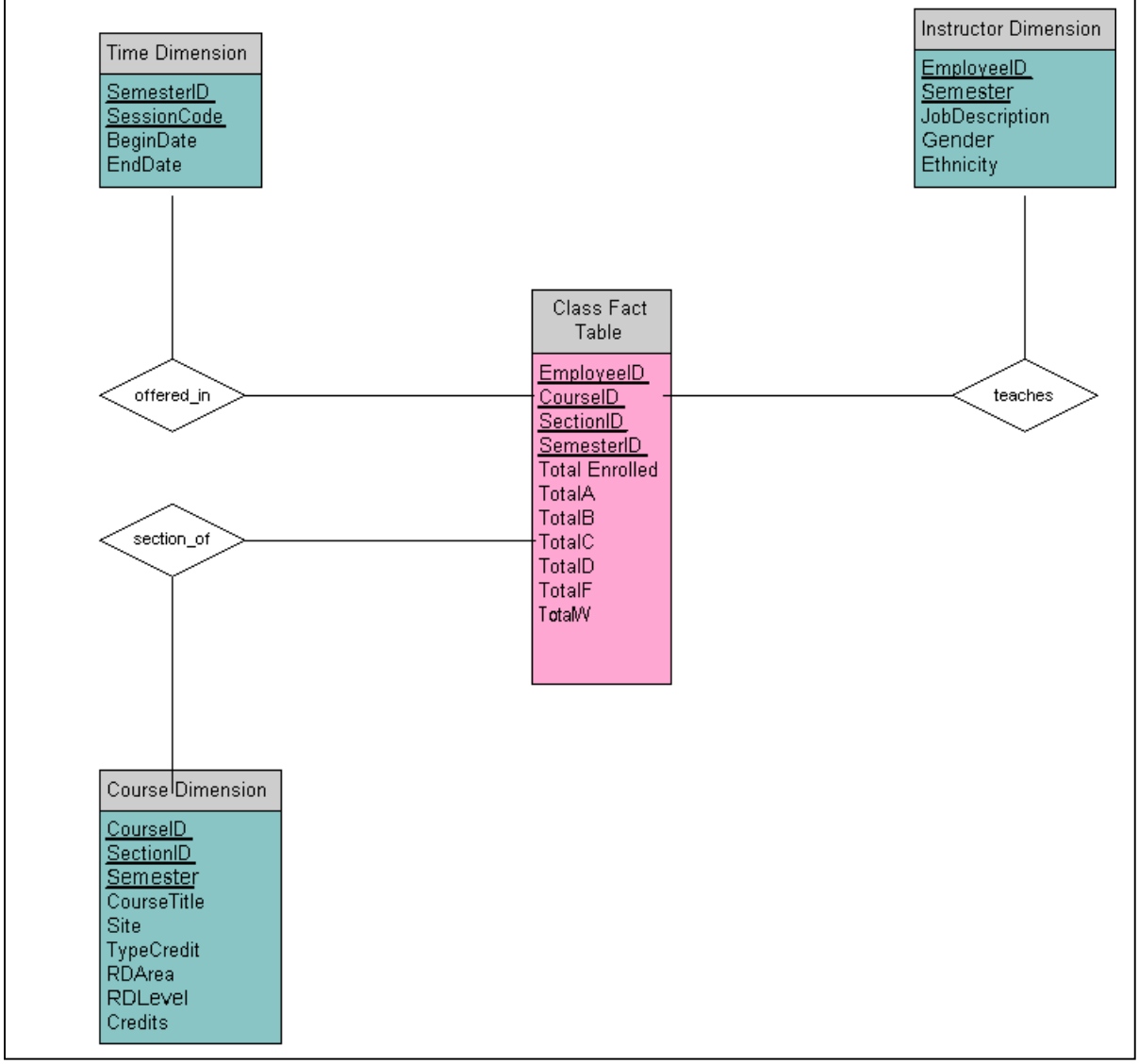
The second set of reports compares grade distributions. The measures in this case will be the total counts for individual courses by the following categories: total enrolled in the class, total receiving a grade of “A”, total receiving a grade of “B”, etc. By adding these measures to the previous fact table, we can use the same star schema to produce both sets of reports. The revised star schema is shown in Figure 12.



2.4.3.2 Student Analysis Reports.

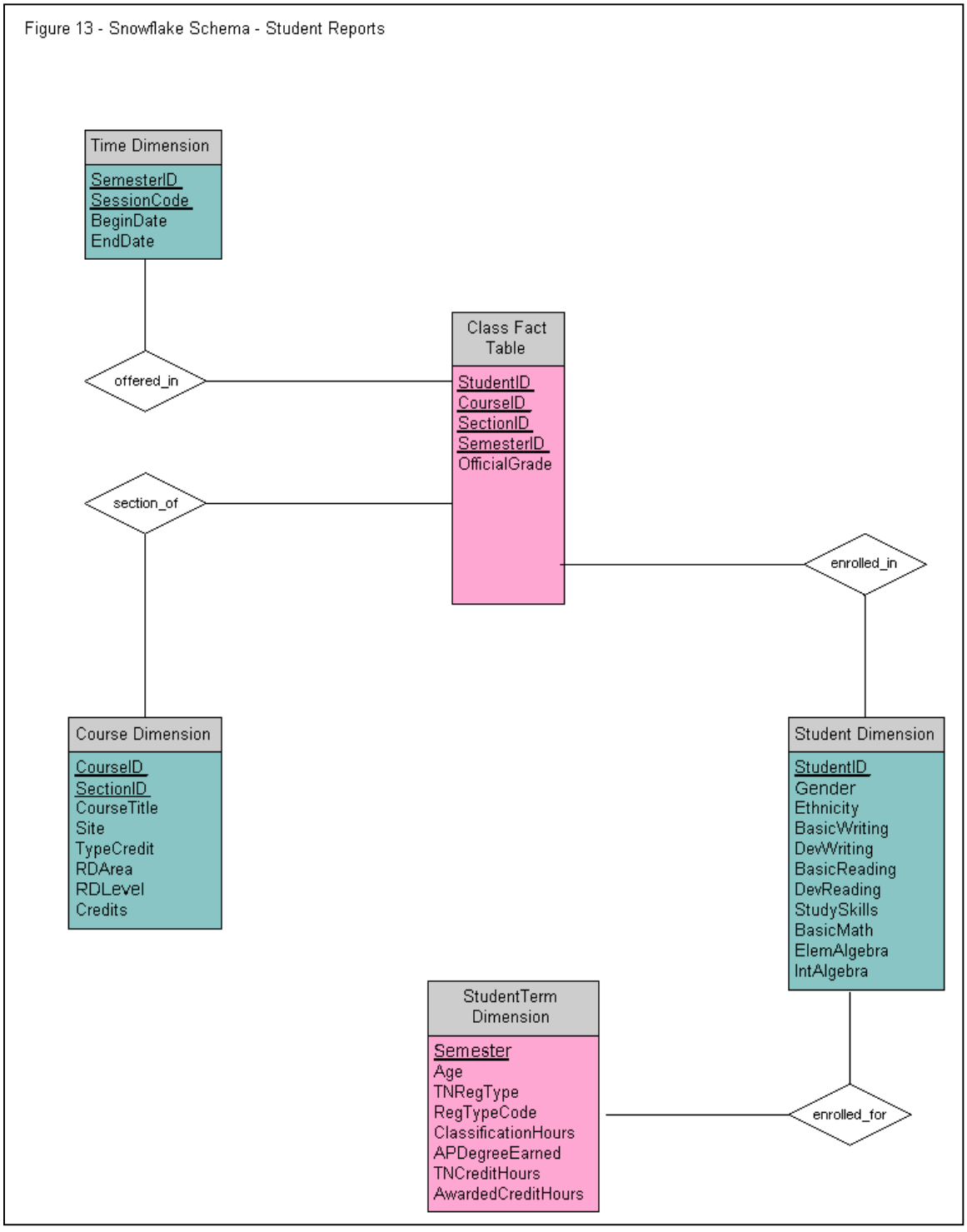
The student analysis reports compare all enrolled students with those designated as DSP students. For one set of reports, we need counts of first-time freshmen and counts of returning students. We want to look at these counts based on certain demographics: ethnic origin, gender, age, and the DSP status. The second set of reports looks at the number of students enrolled in DSP courses by classification (the number of credit hours earned). In both cases, our

Figure 12 - Star Schema - Grade Distribution Reports



measures will not be stored in the fact table, but will be calculated measures. We will need the Term Dimension from the previous set of reports and a Course Dimension, to determine course type. The Student Dimension is bit complicated, because the ethnic origin, gender, and DSP status do not change from one semester to the next, but the age and classification do. The solution is a snowflake schema, as shown in Figure 13. In the next section, these designs and reports will be compared with current reports being produced from the SIS system.

Figure 13 - Snowflake Schema - Student Reports



2.4.4 Comparison of Data Warehouse Reports and SIS Reports

The first step in developing a data warehouse using the top-down approach is to analyze current query behavior and then use the data fields being used in those queries as the attributes in the data warehouse. This method can produce quicker results than the bottom-up method in terms of development time, but it also assumes that the reports currently being generated are producing correct information. To determine how the top-down method compared to the bottom-up method that was actually used to implement the Northeast State data warehouse, programs that are regularly run out of the SIS system to produce reports were analyzed to determine the source elements used in the programs. The report programs were then compared to the documentation of the data warehouse sources to find instances in which the data warehouse is using different data sources for the same information.

Of the 20 reports analyzed, there was only one instance of a data element's being used in the SIS reports that was not included in the data warehouse in any form. This element (the high school from which the student graduated) was not included for two reasons: (1) the data were not requested in the original reports specifications for the warehouse and (2) the element is stored in a file that has not yet been included in the data warehouse because it does not include any other data in the report specifications. There were five cases where data elements used in the SIS reports are not used in the data warehouse, but the information they represent is being calculated or obtained in a different way.

The first case is a field that serves as an indicator of whether or not a student is in the Developmental Studies Program. In the bottom-up analysis, it was determined that the element had been created to make reporting easier, but it is hand loaded and may not be completely reliable. Individuals who regularly use this particular field stated that more accurate information can be derived from

student placement information, so it was implemented that way in the data warehouse.

Another element used in the SIS system that is not used in the data warehouse is the student classification level (i.e. freshman or sophomore). This information can be derived by calculating the number of credit hours earned by the student. Because the requirement for each classification may change over time, it is actually incorrect to determine the classification based on a stored flag, rather than on credit hours. Similarly, the designation of a full-time versus part-time student is calculated from the student's enrolled credit hours, instead of the stored "FT/PT" indicator, which is used in the SIS reports.

Some of the SIS reports are based on the use of fields that are known to be unreliable but are the only means of storing those particular data. Many reports use a field intended to designate how a course is to be listed in the printed class schedule as a filter for class location, rather than using the site code for the class. The site code was generally considered to be the more reliable field, so it was used in the data warehouse. It also provides more flexibility, because end users can group the sites in different ways, based on the particular report they need. Several of the reports requested from the DSP analysis are based on comparisons of full-time instructors with part-time instructors. As pointed out in Section 2.4.1, the field used for this designation in the SIS system is not reliable. For example, in a grade distribution report for the 1998 Fall semester, out of 760 course records, 76 were coded incorrectly:

- 37 records had part time instructors coded as full-time
- 38 records had administrative or support staff coded as full-time
- 1 record had administrative or support staff coded as part-time

The data warehouse uses information from the Human Resources System to determine if the instructor is full-time or part-time. Also, SIS only has two classifications, and administrative and support staff are technically neither.

Because there is a difference of opinion regarding which category these belong to, the data warehouse gives end users the option of including or excluding them as they see fit.

At this point, it appears that the bottom-up method was a good design option for the source data being used. However, an evaluation of the data warehouse is needed to help discover any weaknesses to the approach. The next chapter explains the procedure used to develop the criteria for the evaluation.

CHAPTER 3

DEVELOPING THE CRITERIA FOR THE DATA WAREHOUSE

Much of the current research pertaining to data warehouse quality shows that applying known methods of database design does not necessarily guarantee a quality data warehouse product, mainly because the requirements are very different for a data warehouse. To evaluate the quality of a data warehouse, the Data Warehouse Quality (DWQ) project, referenced in Chapter 1, used the Goal-Question-Metric (GQM) software engineering methodology described by Basili, Caldiera, and Rombach (1994) and expanded it for data warehouse specific measurement. In the GQM method, each part of the data warehouse architecture is defined as an "object", and the user requirements are defined as "goals". The components of the typical data warehouse architecture are shown in Figure 14. Metrics are defined to measure some property of each object in terms of quality. "Quality questions" are then formulated to relate the goals to the metrics. For example, if a high-level goal is "completeness", the data warehouse object might be identified as "sources". The quality questions could then be:

- What is the window of availability of the source(s)?
- How much history is contained in the warehouse?
- How far back is the historical data stored in the warehouse?
- How frequently is data extracted from the source(s)?

Thus, the first step in an evaluation is the development of a set of requirements. The remainder of this chapter describes the process of developing a survey to gather those requirements for the Northeast State data warehouse, and then summarizes the survey results.

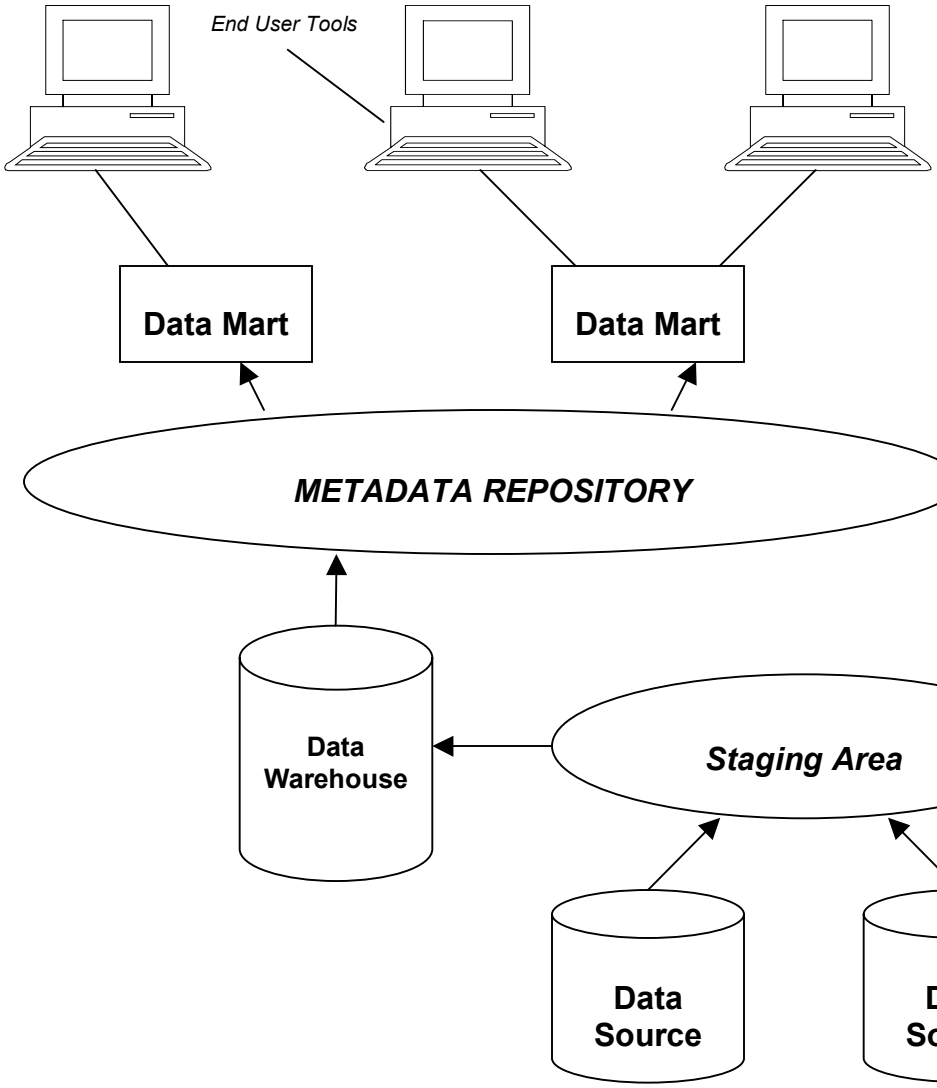
3.1 Developing a Survey

In order to develop a baseline set of end user expectations, a survey was developed to determine the features or characteristics of a data warehouse that would be required for the intended users to consider the data warehouse a successful project. Using the DWQ quality framework (Jarke and Vassiliou 1997), shown in Figure 2, as a guide, first, the quality factors had to be defined, then metrics determined for those qualities, and, finally, questions developed to be asked of the end users in order to establish a range of acceptable values. Of the five quality factors in Figure 2, questions were developed to measure four of them. The definitions for each factor and the types of questions that were asked to establish the requirements for each are in the sections that follow. The complete survey is included in Appendix A.

3.1.1 Accessibility

Accessibility can actually be defined in two ways. First, it is the amount of time that a system is online and available to end users. This factor is fairly easy to measure, because we are primarily concerned with the number of hours per day and the days of the week that the system is available. The second definition is somewhat harder to measure, as it has to do with the end users' perceived ease of accessing information from the data warehouse, as opposed to the current level of access to the source systems. This is one of the more critical success factors because, if the data warehouse makes the data more difficult for the end users to access without providing any extra value in terms of information, the developers may have a difficult time selling the end product to the users (Greenfield 2001).

Figure 14 – The Data Warehouse Architecture



The questions regarding available days and times are fairly straightforward. The users were asked the minimum acceptable hours of availability each day, if they would prefer access during normal working hours over weekend and after-hours access, and how often they might expect to use the system. They were also asked how important the availability of the system is

to them compared to other criteria. Questions regarding the access to data asked about current levels of access to the source system, the importance of viewing detail data, and the necessity of the ability to manipulate data themselves.

3.1.2 Interpretability

The quality factor of interpretability in a data warehouse is primarily concerned with how well users are able to understand the information in the data warehouse reports. Because data warehouse data are often the result of aggregations and summarizations of the source data, it is important that the end users are provided some means to interpret the results they see in reports. Much of this understanding should come from the data models, but end users will also most likely need some form of documentation. The intended users at Northeast State do not have experience working with data models, so there is already an understanding that some training will be necessary to help them understand the models. However, other documentation should be provided in some form to assist users with using the reporting tools and analyzing the results. The only real issue here is the preferred format of documentation. The questions asked attempt to determine the types of documentation (if any) the end users currently use, as well as their stated preference, i.e. hard-copy manuals, online documentation, etc.

3.1.3 Usefulness

There are many factors that determine if the data warehouse is useful, but some can only be measured over a period of time. A good data warehouse is never really “finished”, particularly if it is successful with the users. At Northeast State, for example, the end users did not initially understand that the data warehouse would give them access to reports that were previously unavailable to them, and that they would be able to produce the reports themselves with

minimal IT intervention. Once they realized this, they immediately began asking about possible additions to the warehouse. If the data warehouse is not maintained properly or continually developed and expanded, its usefulness will decline over time. As Larry Greenfield states, “It’s very easy for the users to quickly go sour on a system they were enthusiastic about at roll-out time if the system personnel do not support the maturing of the system” (Greenfield 2001).

This aspect of usefulness cannot be measured until the data warehouse has been operational for a significant period of time, but there are other aspects that can be measured now. Some end users may prefer to have “raw” data available to them, while others may prefer a briefing book type of setup, with several pre-formatted reports readily available to them. We can ask questions regarding the importance of these features. Response time may be an important factor, so we need to determine the maximum amount of time that is acceptable to wait for reports to be run, using various querying scenarios. The reporting tool should be easy to use, else the users will likely fall back into the practice of contacting IT to produce reports for them. One way we can measure ease of use is by determining how many steps an end user is willing to perform to produce a report. Finally, we need to determine the level of detail users want to have available. Users may find that highly summarized reports that do not offer them the ability to drill down to more detail are not particularly useful.

3.1.4 Believability

There are many factors in data warehouse construction that can make believability a critical issue. In the best case, developers will be working with a relational database as a source system, where constraints have been properly applied to ensure that the data are clean and uniform. In most cases, however, developers will be working with systems that may not be very particular about data formatting, or about ensuring integrity by having a data element occurring only once in the database. The result of “dirty” data, where constraints are lax or

nonexistent, may be that calculations in the data warehouse produce results that differ from those in the source system. Also, some summaries may be incomplete because the particular attribute being used in the calculation is not mandatory in the source system. For example, a student record can be entered into the SIS system without a date of birth being specified. This means that summaries and aggregations using calculated ages will not be complete, because the date of birth is not known for some of the records. In cases where a data element is stored more than once in the database, end users may be using different instances of the element for reports, and results may differ based on which copy is used.

In all of these situations, the important issue is ensuring that end users have some way to know how each aggregation and summary was derived. If there are questionable figures in the data warehouse, the end users need to know why the results were not as they expected them to be, or do not match figures from the source system. To determine the best way to communicate this information, the survey questions asked what evidence would be most compelling to the end users to explain the data warehouse results. Methods of conveying the information were suggested as well, such as the ability to drill down and show the detail used to derive summaries, or documentation showing the original source of the data and the definitions of data elements. Users were also asked about how they currently go about reconciling reports that they receive from the SIS system.

3.2 Survey Results

The survey was advertised to the Northeast State campus community via e-mail, with a brief explanation of the data warehouse concept, as well as a description of the DSP analysis the data warehouse was being designed to support. Twenty-one individuals expressed an interest in participating in the project. The survey was distributed to these participants, and 14 of those were

returned. The following sections discuss the results in the context of each of the quality factors described in the previous section.

3.2.1 Accessibility

Most of the users surveyed are currently using the SIS system on a daily basis and have unlimited viewing access to student and course data. It is most important to them to have at least as much access to the data as they have in the SIS system. The days and hours of availability of the data warehouse are not as critical to them. They would like to have access during the week, during normal working hours. Although the answers to a direct question regarding the frequency of data refreshes were not conclusive, all of the respondents agreed that they would use the data warehouse if it was refreshed weekly.

3.2.2 Interpretability

Although the majority of users indicated that they do not currently use hardcopy documentation, most of them indicated that they would prefer a printed (or printable) manual or set of instructions over online help. This may be because the primary documentation available for the SIS system is online, context-sensitive help. The hardcopy documentation is maintained in a rather formidable set of notebooks that is not frequently updated by the vendor. Also, most of those surveyed indicated that they find reports in a tabular format easier to interpret and understand than other formats, such as formatted reports, or charts and graphs.

3.2.3 Usefulness

Most of the individuals surveyed indicated that it was important to them to have data available to them in downloaded text files so that it could be imported into an analysis tool of their choice, but they also indicated that they would like to

have a collection of pre-defined reports from which to choose. Answers to questions regarding response times were not conclusive, perhaps because of the frame of reference of the users. Very few of the individuals surveyed are currently able to produce their own reports from the SIS system, so they did not answer the questions with the idea of submitting queries themselves and waiting for the results. In terms of the ease of use of the reporting tool, nearly all of the responses indicated that the users would not want to perform more than five steps to produce a report. Finally, the end users all agreed that they would like summarized data for all of the dimensions (Student, Instructor, Course, and Semester), but the results were evenly divided regarding the dimensions for which they need detail data.

3.2.4 Believability

The results of this set of survey questions indicate that the users need both access to detail data used to derive summaries and the documentation of the original source of the data in order to reconcile questionable data warehouse reports. The answers to questions regarding the users' current methods of reconciling reports were varied, with several users indicating that they do not question the current reports from the SIS system. Finally, most of the individuals surveyed responded that the most important item of documentation would be the definition of the data elements, followed closely by the date that each element was last refreshed.

The next chapter explains how the results of this survey were used to design an evaluation instrument using the GQM methodology.

CHAPTER 4
EVALUATING THE DATA WAREHOUSE

Now that the requirements for the data warehouse are established, the GQM methodology will be applied to complete the evaluation phase. The next step is to develop a set of quality schemas that define:

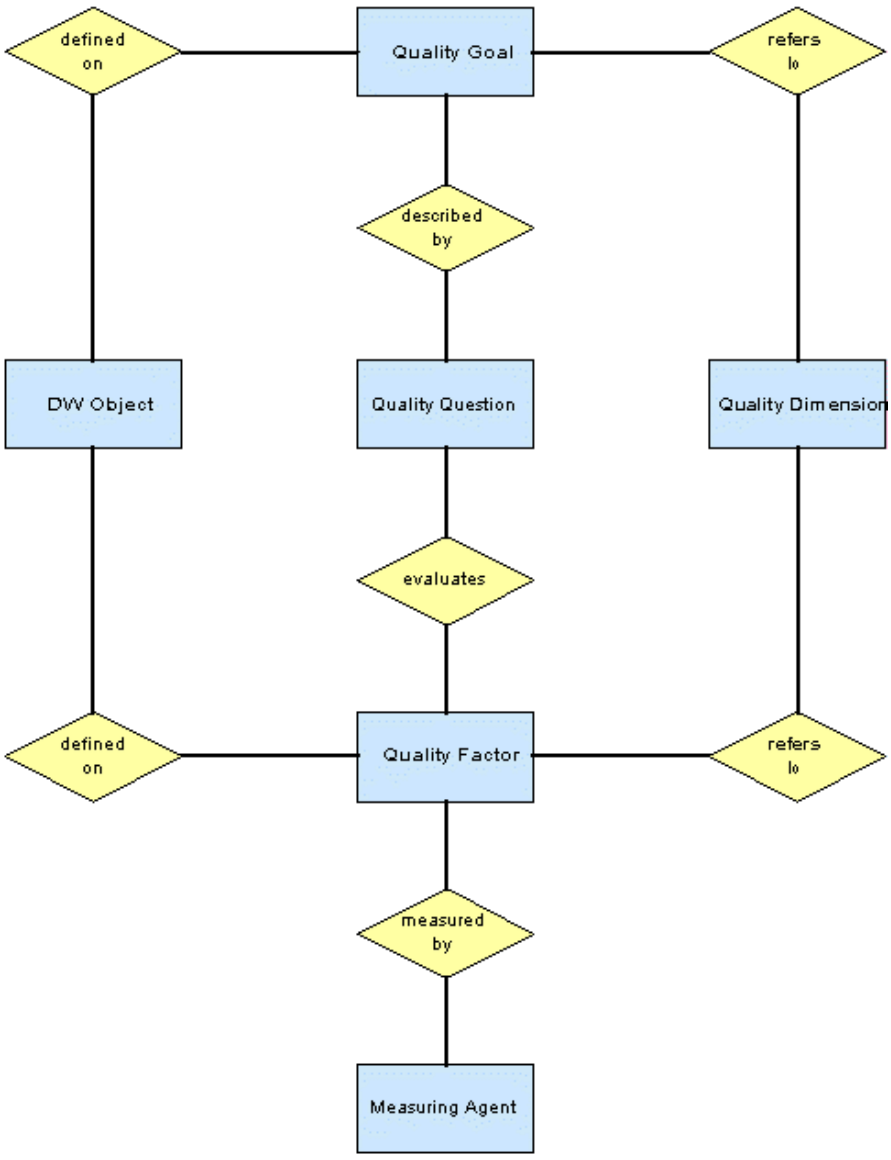
- The object in the data warehouse architecture
- The metrics to be used
- The means to be used to conduct the measurement

Figure 15 shows a template of the quality schema. In the GQM methodology, as adapted by the DWQ project, the schemas were implemented as a database (or as part of the data warehouse meta data), and that database could be queried periodically to measure the state of the data warehouse (Jarke and Vassiliou 1997). For this study, the quality schemas will be used to develop an evaluation tool to be administered to the end users surveyed in Chapter 3. The complete evaluation is included in Appendix B. The sections that follow describe each of the nine schemas that were developed, the evaluation questions or processes that were derived from them, and the results of that evaluation. The final section gives an overall summary of the evaluation, as well as general user perceptions of the warehouse.

4.1 Description of the Quality Schemas

Before the evaluation could begin, the reports required by the DSP analysis request were processed and made available to the users. The data warehouse itself was implemented in a SQL Server 2000 database, so the simplest end user tool to use was the Excel Pivot Table. In some cases, data cubes were defined in Microsoft Analysis Services that were then accessed via Excel, while other reports were generated directly from the SQL Server tables (Microsoft Corp. 2001). Excel was chosen because most of the end users were

Figure 15 - Quality Schema Template



already familiar with it, and it interfaces well with Analysis Services. The evaluation phase was advertised to the same group of individuals who filled out the survey described in Chapter 3, because the evaluation tool was based on their responses to that survey. Fifteen individuals expressed a desire to participate in the evaluation. A training session was held to familiarize the users

with the Excel Pivot table tool, and to explain the processes that were followed to design and construct the data warehouse. Seven individuals attended the training, so the remaining eight were given hardcopy documentation (which was also distributed to those who attended the training) to assist them with using and interpreting the reports. The pivot tables used for the evaluation were made available to the evaluators on the campus network. Of the 15 evaluations distributed, 10 were returned.

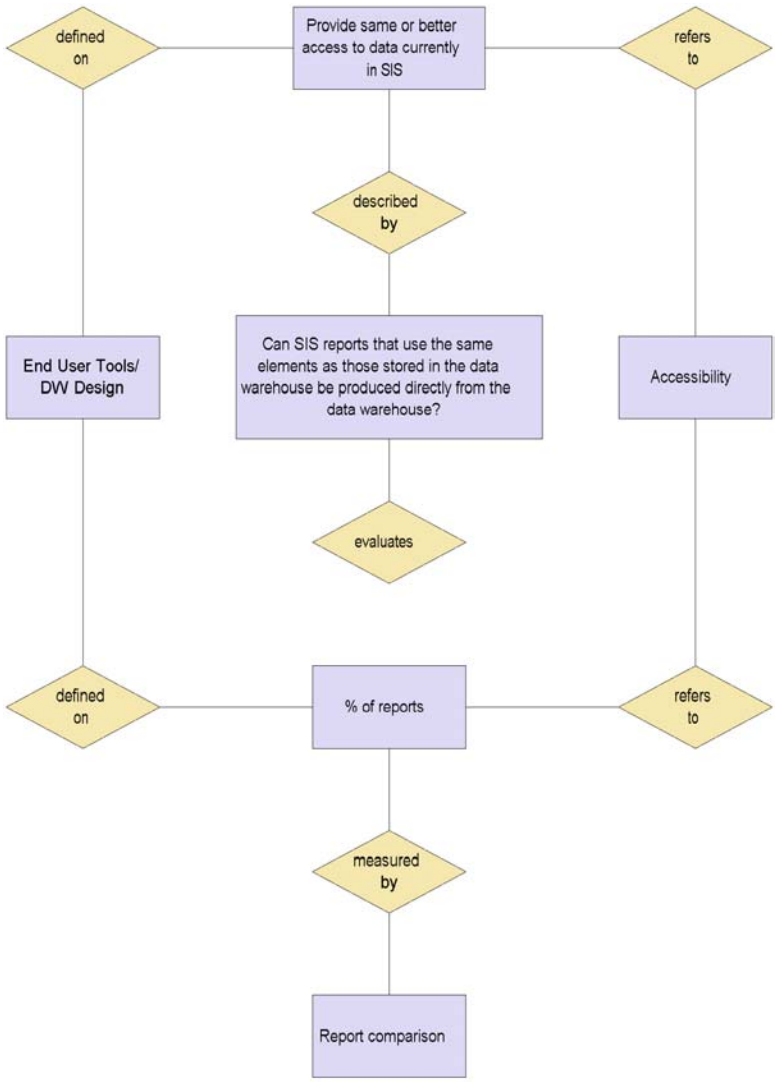
4.1.1 Quality Schema 1

This schema, shown in Figure 16, provides a guideline for measuring accessibility in terms of the ability to obtain at least as much information from the data warehouse as from SIS. The users were provided a set of reports from the data warehouse and were asked the following types of questions:

- Do the data warehouse reports provide as much information as SIS provides?
- Is there information in the data warehouse that has been unavailable to you in SIS?
- Are the data warehouse reports harder to understand, easier, or about the same?

Of the users who regularly use the SIS system and receive reports from the system, all agree or strongly agree that the data warehouse provides at least as much information to them. Eighty percent agree or strongly agree that the data warehouse reports provided information that had not been available to them in SIS, with twenty percent responding with no opinion. The majority of evaluators found the data warehouse reports easier to understand. Overall, the users had very favorable opinions regarding the access to information provided in the data warehouse.

Figure 16 – Quality Schema 1



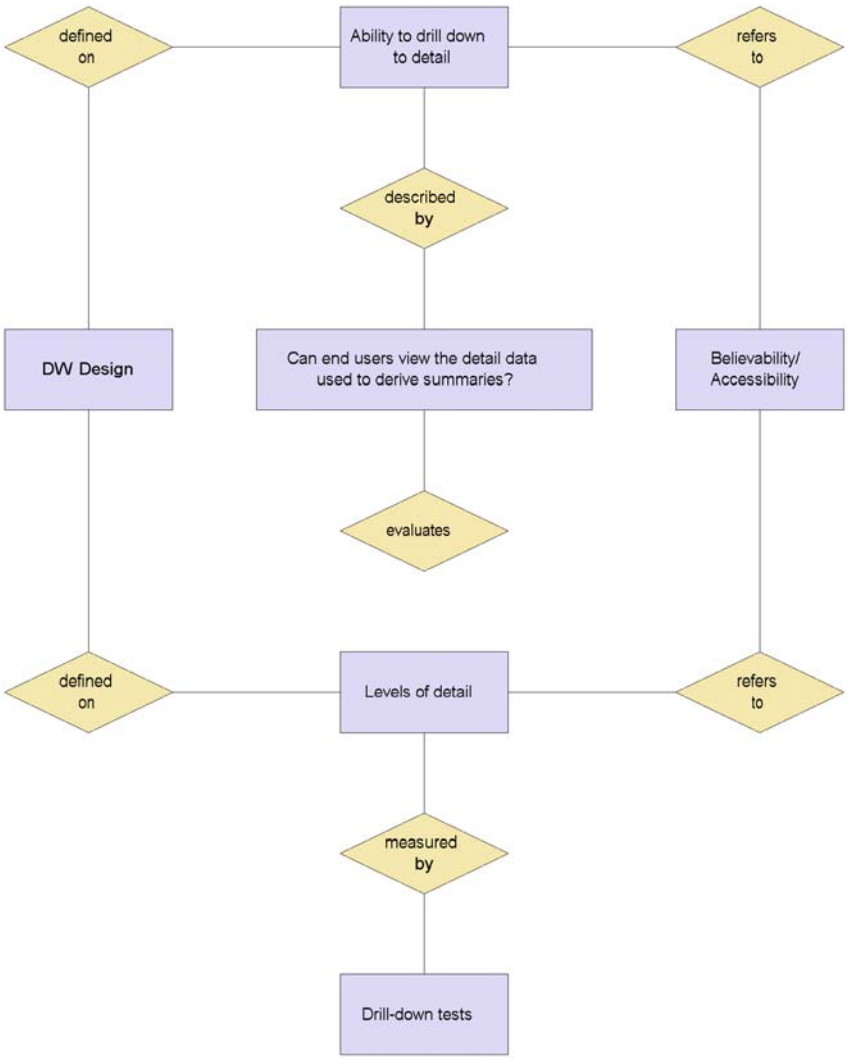
4.1.2 Quality Schema 2

Chapter 3 discussed each of the quality factors individually, but some of the factors overlap in terms of the data warehouse objects and characteristics that affect them. The quality schema in Figure 17 measures believability and accessibility in terms of drill-down capability. Using the Grade Distribution pivot table report, users were asked:

- Was the information in the drill-down reports meaningful to you?

- Is the detail easier to obtain in SIS?
- Are the drill-down reports harder to understand?

Figure 17 – Quality Schema 2



As shown in the quality schema, the data warehouse design is the object that most affects the drill-down capability. This particular part of the design was especially challenging because the designer needs to be able to recognize the hierarchies that exist in the data and then represent them correctly in the star schema designs. The evaluators' impressions in this area were again very

favorable. All of the respondents agree or strongly agree that the drill-down reports provide meaningful information. They also agree that the detail information is easier to obtain than in SIS. Finally, 80% indicated that the drill-down reports were easy to read and understand.

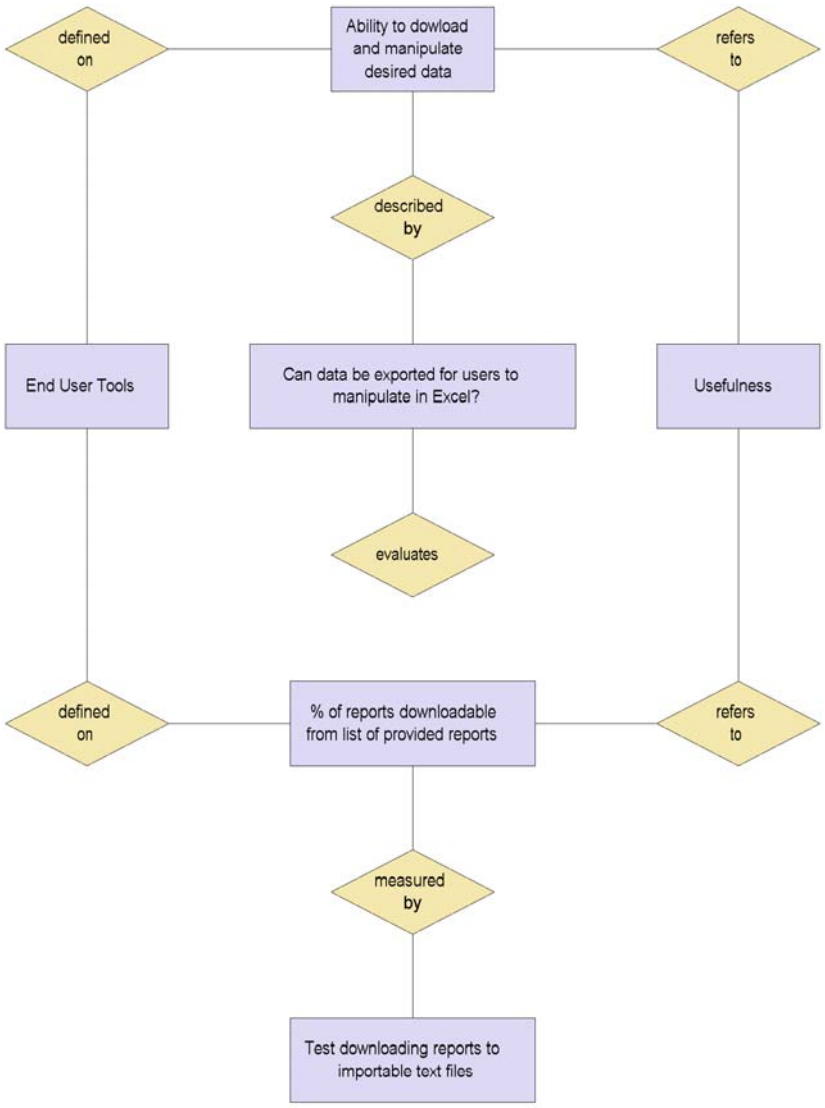
4.1.3 Quality Schema 3

The survey indicated that the majority of end users want to have downloadable data available to them in addition to pre-processed reports. With the tools being used, both of these requirements can be easily satisfied for each report because any table can be exported directly to an ASCII file. Figure 18 shows the quality schema used for the evaluation of usefulness based on the availability of exported data. Of the reports requested for the DSP analysis project, 74% are available as exported text files. About 7% of the requested reports cannot be produced without the addition of census data to the warehouse or are based on data from the source systems that is not reliable.

4.1.4 Quality Schema 4

Although the end users surveyed indicated that the time needed to run reports was not an important issue, this may have been because those surveyed do not have a great deal of experience running their own reports. Processing times may become more of a factor as time goes on. The evaluation represented by the quality schema in Figure 19 assesses the usefulness quality factor based on the time required to run reports. To evaluate this, each of the available reports was processed and timed. The reports were processed via a remote connection using a 56K modem, so these numbers should represent a worst-case scenario. Fifty-nine percent of the reports processed in one second or less. The longest processing time was 28 seconds. The mean processing time was 10.1 and the median was 1 second.

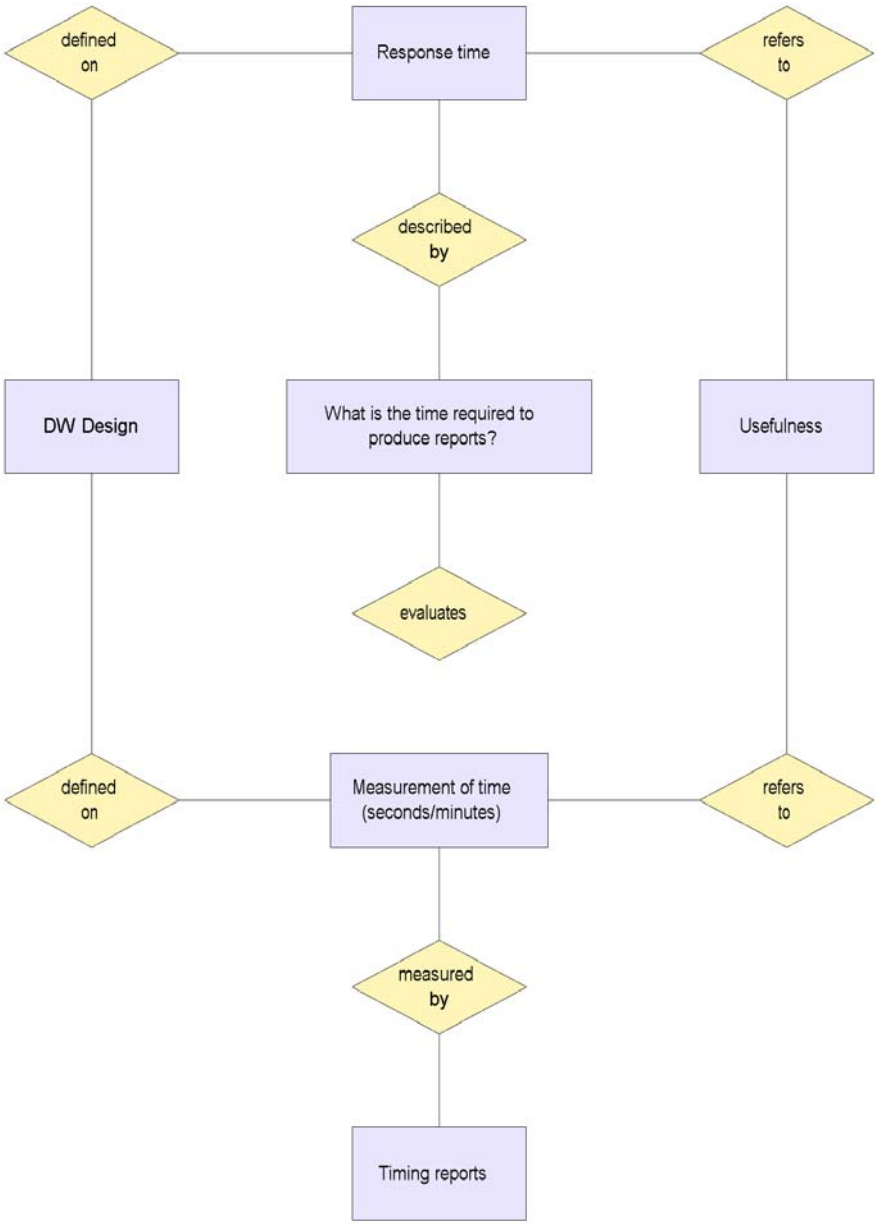
Figure 18 – Quality Schema 3



4.1.5 Quality Schema 5

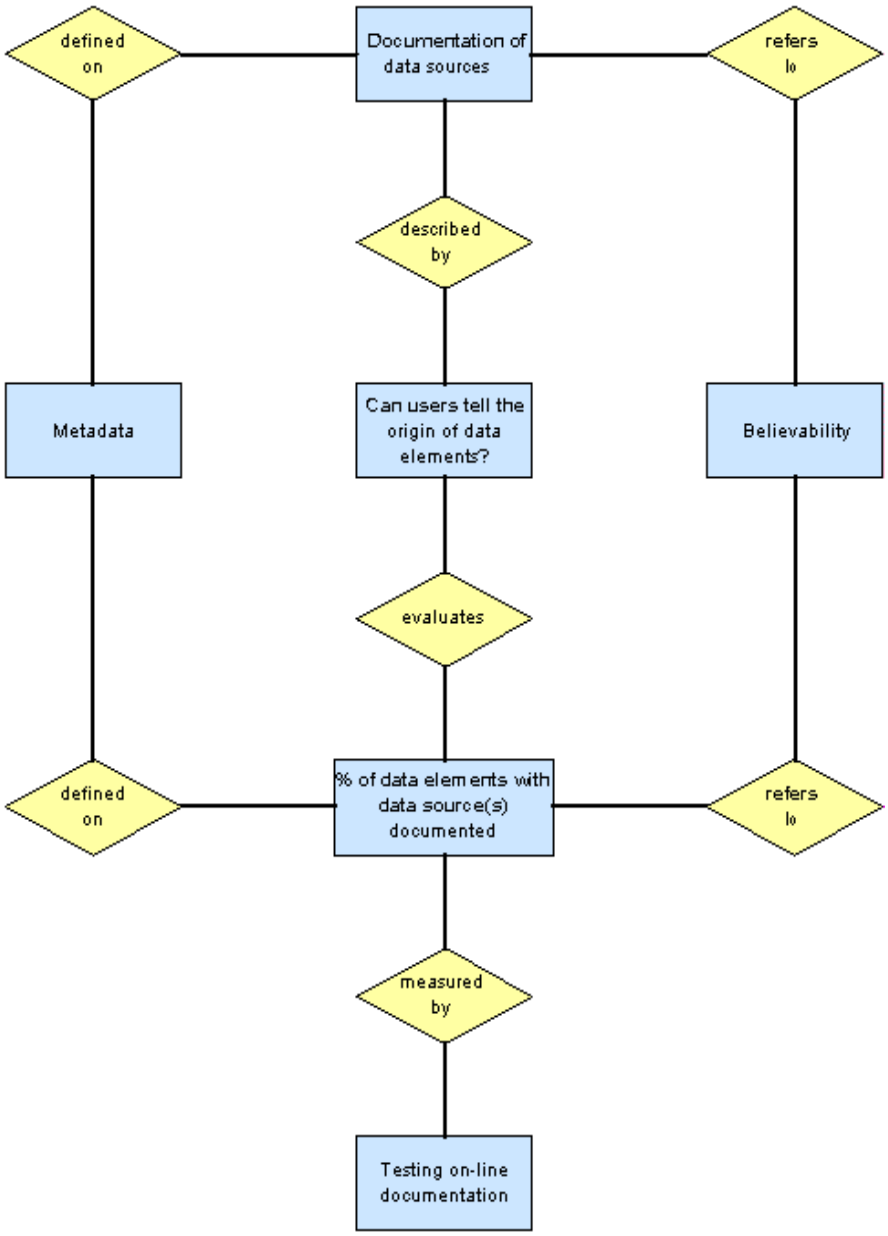
The users who were surveyed indicated that documentation of data sources was an important feature for validating data warehouse reports. The quality schema in Figure 20 shows how documentation of data sources can be

Figure 19 – Quality Schema 4



used to measure the believability factor. Several steps have been taken during the design process to ensure that the data sources are clearly documented and traceable. For example, data elements in the star schema, which the end users may use to help them produce reports, are given meaningful names so that their origin is clear to the users reading it. Also, the source elements from which the data warehouse tables and columns were derived are fully documented in an

Figure 20 - Quality Schema 5



Excel spreadsheet available to the end users. The measurement for this factor is the percentage of data warehouse fields that have their sources documented. One hundred percent of the data elements in the data warehouse are documented.

4.1.6 Quality Schema 6

The evaluation modeled in Figure 21 is similar to the evaluation for reports available as downloads (see section 4.1.3). The end users indicated in the survey they would like to have completed reports available online, as well as downloadable data. Of the requested reports, 74.1% are currently available online as completed reports or Excel Pivot Table reports. As noted in section 4.3.1, 6.9% of the reports cannot be produced from the data currently available in the warehouse.

4.1.7 Quality Schema 7

Figure 22 shows the quality schema that models the evaluation of accessibility in terms of the amount of time the system is online and available. The system was monitored over a period of three weeks, with any downtime being recorded. Other than regular refreshes, which take place on weekends, the system is available 24 hours per day unless a network problem or hardware problem causes the system to be unreachable. Also, any maintenance on the machine is scheduled for after working hours or on weekends. Because the end users' requirement is that the system be available during workings days and hours, this criterion has been met.

4.1.8 Quality Schema 8

The metric chosen to measure the usefulness factor of the data warehouse in terms of the ease of use of the end user tools was the number of steps required to produce a report (see Figure 23). For the survey, the majority of users indicated that they would not want to perform more than five steps to run their own reports. The evaluators were given five reports to process from a given Excel Pivot Table and were asked to record the number of steps (mouse clicks) required to produce the report. Users were instructed to begin counting when

Figure 21 - Quality Schema 6

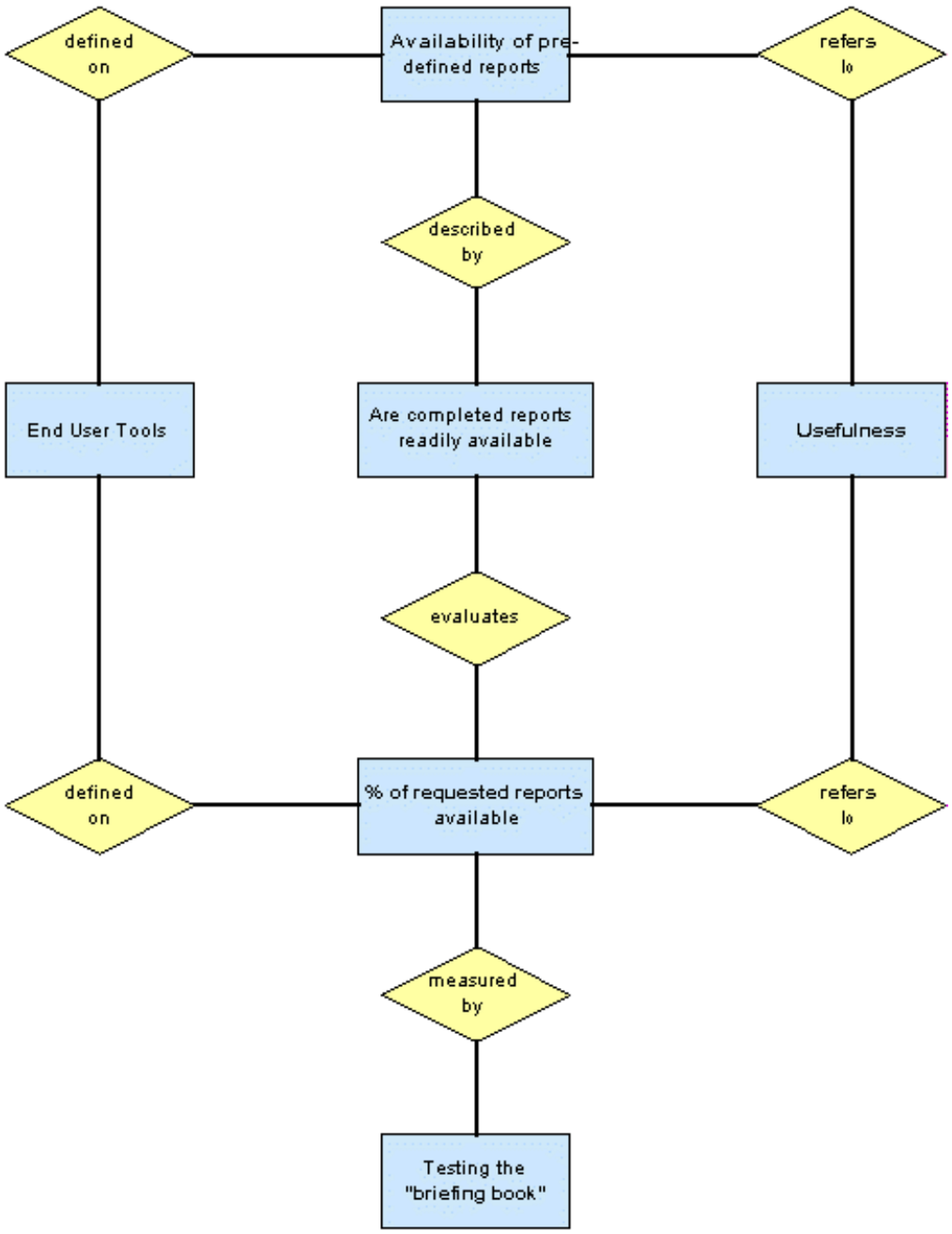
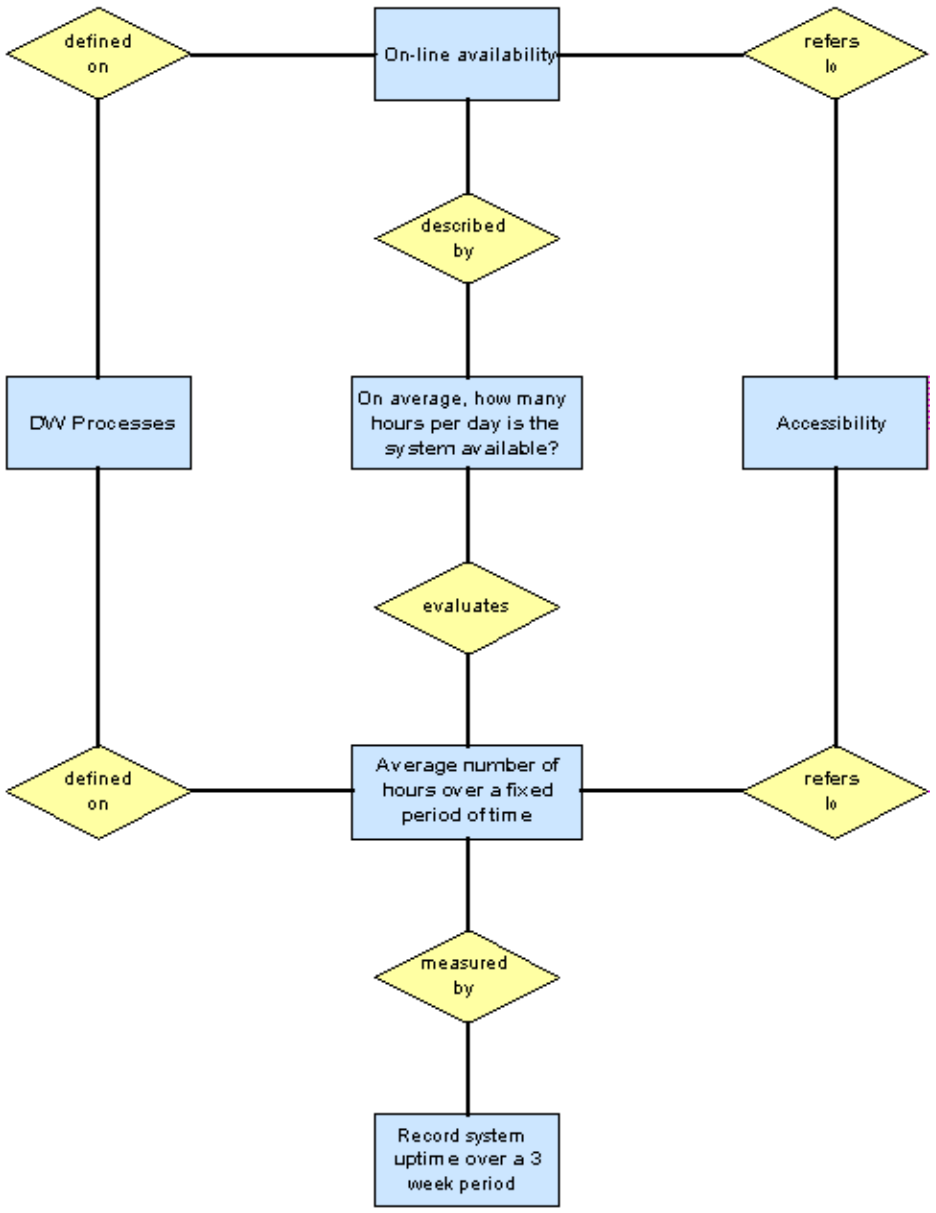
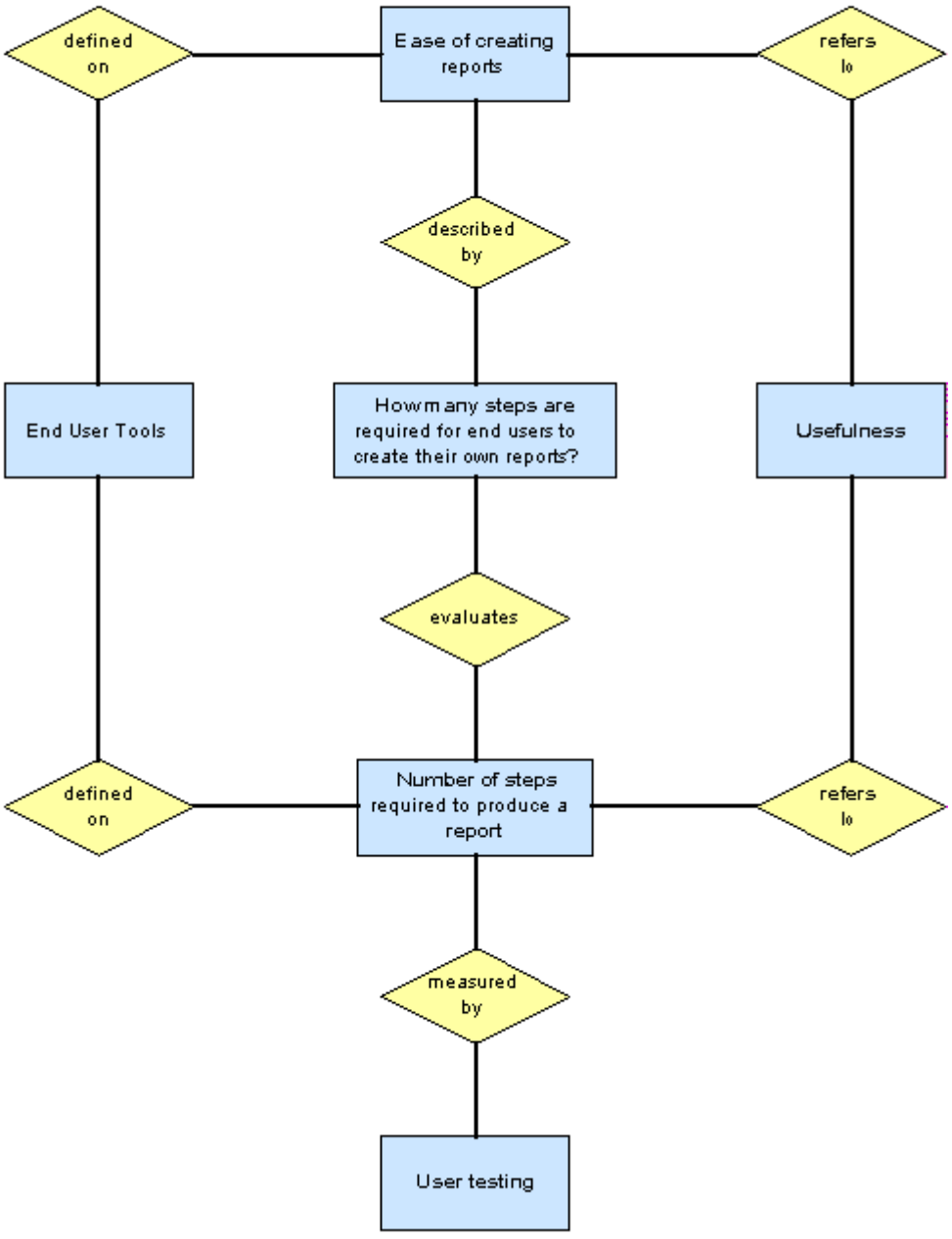


Figure 22 - Quality Schema 7



they had the pivot table visible on the screen. Eighty percent of the individuals completing the survey used the Pivot Table tool. Generally, the evaluators were able to produce each of the reports in five or fewer steps, but even those that found that it required more steps stated that they found the tool easy to use and would like to have more reports available in the same format.

Figure 23 - Quality Schema 8



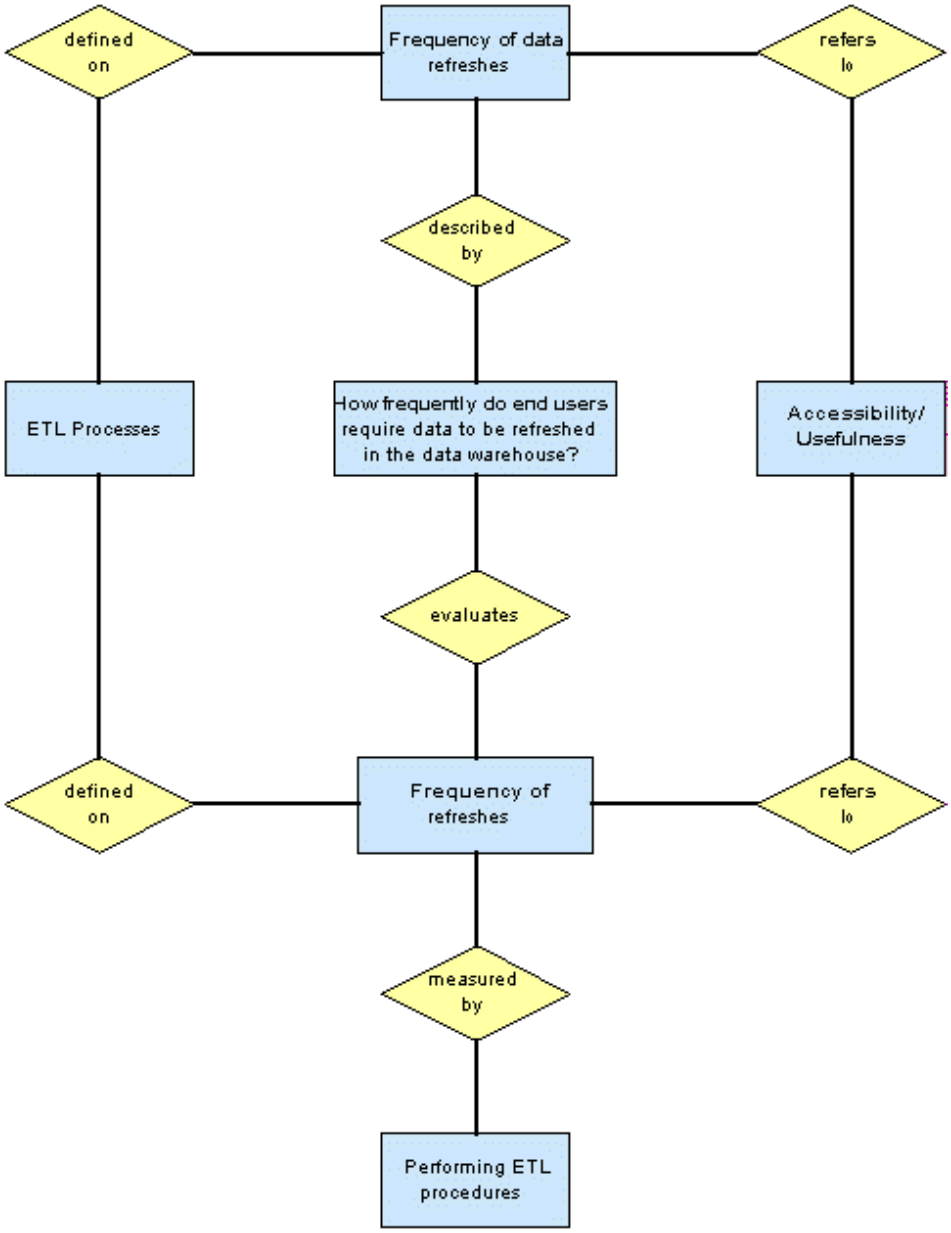
4.1.9 Quality Schema 9

The final quality schema, shown in Figure 24, evaluates the ETL processes being used in the data warehouse in terms of their effect on accessibility and usefulness. The users surveyed indicated that data should be refreshed at least once per week in order for them to be considered useful. Because they also indicated that the system need not be available on weekends, there is a fairly large window of availability for completing the refresh process. Nevertheless, these processes will be competing with normal batch processing somewhat, so they will need to be as efficient as possible. To measure this, the procedure for performing a full refresh of all of the base data warehouse tables and the star schemas was timed from start to finish. The time required to perform a full refresh was approximately 8 hours, which should be easily manageable over a weekend. Also, many of the processes currently being used can be streamlined or simplified to make the procedure run more quickly. In reality, there would be little to be gained by weekly refreshes. Unlike typical sales or production data, which represents on-going, daily data entry processes, academic data tends to be stored in "snapshots". Therefore, the actual date of the refreshes is more significant than the time required to perform them. Eventually, the refreshes will most likely be performed twice per semester – once on the fourteenth day of classes (known as the "census day"), which represents official enrollment numbers, and again at the end of the semester, when all grades have been entered into the system. With this in mind, the refresh times should not be a significant factor.

4.2 Overall Evaluation

The evaluators were asked questions at the end of the evaluation regarding their general perceptions of the data warehouse. Of the users who responded to those questions, the opinions were very favorable. All of the individuals indicated that they would continue to use the data warehouse, rather

Figure 24 - Quality Schema 9



than SIS, for reporting, and that they would like to see the data warehouse expanded to include more types of data. When asked what was the best feature of the data warehouse, the most common response was the ease of use. Others responded that the easy access to comprehensive, in-depth information was the best feature. In answer to the question regarding changes they would like to see

made, the most common response was that they would like to have more reports available and more types of data included.

The weakest areas of the data warehouse currently are in the areas of available reports and documentation for using the Pivot Table tool. As mentioned in section 4.1.2, to make the most effective use of the Pivot Tables, the designer must identify the hierarchies that exist within the data, and these are not always obvious. For example, the grade distribution reports are currently most useful for viewing the grades in the DSP classes, and the design hierarchy only looks at DSP classes versus non-DSP classes. However, more hierarchies exist within the non-DSP classes that could be exploited to make very informative reports for administrators responsible for those classes.

The individuals participating in the data warehouse evaluation represented a varied cross-section of college employees. Not all of these individuals are in decision-making roles, but many expressed the opinion that the information provided by the data warehouse would be very useful to college administrators who are in such roles.

CHAPTER 5
CONCLUSIONS

The purpose of the research described in this paper is to analyze the process of designing and constructing a data warehouse for an academic institution and to document the special characteristics of academic data that affect the data warehousing processes. Some of these characteristics are specific to data sources organized as complex flat file structures. This chapter discusses the conclusions reached as a result of this research.

5.1 Top-Down vs. Bottom-Up Method

The analysis of current reporting behavior, described in section 2.4.4, showed that the top-down method could not meet the requirements for the Developmental Studies Program analysis. The top-down approach makes the assumption that reports currently being produced from the source systems are correct, and the research done in this project has shown that this is not the case. Current reports frequently make use of hand-keyed data elements that represent other "original" elements, thereby introducing an increased possibility for error. In addition, the fact that the source systems are not currently integrated at all means that unreliable data elements must sometimes be used for reporting, even if a more reliable data element exists in another system. Once the unsuitability of the top-down method had been determined, the focus of the research shifted from comparing the two design methods to determining if the reporting requirements could be met using the bottom-up method and documenting problems encountered during the process.

The evaluation of the data warehouse, as detailed in Chapter 4, showed that the bottom-up method was indeed successful in meeting the requirements for the DSP analysis, as well as end user performance requirements. Also, several of the end users expressed an interest in having other types of reports

available from the data warehouse, in addition to those defined for the DSP analysis. Because the bottom-up method was used, much of the data needed for those reports have already been extracted and incorporated into the existing data warehouse.

5.2 Academic-Specific Issues

One of the issues encountered specific to academic data that affects the data warehouse process concerns the time characteristics of the data. In most business-oriented database applications, sales and production data are entered into the system on a continual basis. Time constraints on the data are somewhat artificial, meaning that the information is grouped into time categories for accounting convenience. By contrast, academic data are stored in snapshots, with very definite beginning and ending dates. Primarily, institutions tend to be most interested in the data as they exist at the official census date and at the end of the semester. This means that frequent refreshes of the data are not necessarily important, but the refreshes must be done promptly and completely on those critical dates.

A second characteristic of academic data is that they tend to be more complex and multidimensional than business-oriented data for analytical applications. Research of star schema development for business applications implies that one well-designed star schema can meet most of the reporting needs of a particular area of the organization. In these applications, however, the measures are fairly straightforward, and just a few dimensions exist. In some of the DSP reports, the measure was actually the count of joins of the dimensions, while other measures were evident in the fact table. A single star schema, with all of the dimensions necessary for the reports, would have been too complicated for end users to work with, so the solution was to use more, simpler star schemas to support each different aspect of the analysis (i.e. Instructor analysis, Student analysis, etc.).

5.3 Source Data-Specific Issues

Some of the issues encountered in designing and constructing the data warehouse were specific to the structure of the source data. First is the issue of duplicate data elements, and the necessity of determining a "copy of record". As described in section 2.4, several data elements are duplicated in different files, with no guarantee that all copies are kept in synch. In order to ensure that the data warehouse contains the most reliable data, the developer must engage in analysis to determine which copy of the data end users consider to be most correct. In proper, normalized, relational databases, this is not an issue.

A second characteristic of this particular source data that is not commonly documented in literature is disparate systems. As previously noted, many of the current reports from the Student Information System are based on an unreliable data element, the full-time/part-time status of the instructor, primarily because a more reliable element is not available within that system. While a more reliable source exists in a "sister" application, the Human Resources System, the two applications do not interface with each other at all. Although it is not uncommon to incorporate many different external systems into a data warehouse, this particular situation was complicated by the time characteristics of the data discussed in section 5.2. The SIS data are stored in semester snapshots, but the HRS data are stored on a fiscal year basis. This complicated the task of incorporating that data into the warehouse, as the dates of employment for an instructor had to be matched up to the correct semester beginning and ending dates to determine the instructor's full-time or part-time status for that particular semester.

5.4 Lessons Learned

One of the phases of the research project that proved to be the most cumbersome was the development of extraction routines. Most of these routines are COBOL programs run on the source systems, and the logic for cleansing the data is in those programs. That means that any time the cleansing rules change, the programs have to be modified and recompiled. The preferred method for extracting and cleansing data is to extract the data "as is" from the source systems into a staging area that has an identical structure to the relational database that will contain the clean data. All of the cleansing routines are incorporated into the transformation processes. The benefits of this method are two-fold: (1) the effect of the extraction routines on source system performance is minimized, since the overhead of the cleansing process is removed and, (2) modifications that need to be made to the cleansing or loading routines can be made on the fly.

A second item that proved to be a problem during data loads and refreshes was the definition of foreign keys in the physical database. The foreign keys in the data warehouse that were represented in the logical data models were also implemented in the physical data warehouse database. This resulted in several incomplete table loads, as records with old values for certain fields that are not longer valid (such as major codes or academic program codes) were encountered. Implementing the foreign key constraints does not really make sense for a data warehouse, since data entry control is not an issue. Most of the constraints have subsequently been removed.

Finally, as reports were developed for the end users, the importance of identifying the logical hierarchies that exist in the data became evident. Some of the star schema designs could have been done much differently to take advantage of those hierarchies and provide more flexible and informative reports.

5.5 Future Work

The results of the evaluation indicate that the users who have had the opportunity to use the data warehouse view it as a successful project. Most all of them agreed that it provided them with information they had not had access to previously, in an easy to use and easy to understand format. The key to ensuring that it remains a valuable information resource to the college will be continual modification, expansion, and tuning. As mention in the previous section, the extraction routines need to be rewritten to streamline the ETL processes. In addition, more data sources should be incorporated to expand the data warehouse to meet users' increasing need for information.

BIBLIOGRAPHY

Ballou, Donald P., and Giri Kumar Tayi. 1999. Enhancing Data Quality in Data Warehouse Environments. Communications of the ACM. (January): 73-78.

Basili, Victor R., Gianluigi Caldiera, and H. Dieter Rombach. 1994. The Goal Question Metric Paradigm. Encyclopedia of Software Engineering. New York. John Wiley and Sons, Inc.: 528-532.

Chaudhuri, Surajit, and Umeshwar Dayal, 1997. An Overview of Data Warehousing and OLAP Technology. ACM Sigmod Record. (March).

Gardner, Stephen R. 1998. Building the Data Warehouse. Communications of the ACM. (September): 52-60.

Gibbons Paul, Lauren. 1997. Anatomy of a Failure. CIO Enterprise Magazine. (November). Internet. Available from http://www.cio.com/archive/enterprise/111597_data_content.html; accessed 18 July 2001.

Greenfield, Larry. 2001. The Case Against Data Warehousing. (June). Internet. Available from http://www.dwinfocenter.org; accessed 18 July 2001.

Hadley, Laura. 2000. Developing a Data Warehouse Architecture. Internet. Available from http://www.users.qwest.net/~lauramh/resume/thorn.htm; accessed 25 August 2001.

Hammer, Joachim, Hector Garcia-Molina, Wilburt Labio, Jennifer Widom, and Yuc Zhuge. 1995. The Stanford Data Warehousing Project. IEEE Data Engineering Bulletin. (June): 41-48.

Hay, David C. 1997. From a Relational to a Multi-dimensional Data Base. Internet. Available from http://www.essentialstrategies.com/publications/datawarehouse/relmult.htm; accessed 5 October 2001.

IBM. 1998. Data Modeling Techniques for Data Warehousing. IBM Redbook SG24-2238-00. Armonk. IBM Corp.

Inmon, William H. 1996. Building the Data Warehouse. New York. John Wiley & Sons, Inc.

Inmon, William H. 2001. Little White Lies. Internet. Available from http://www.billinmon.com/library/articles/artlies.asp; accessed 30 December 2001.

Inmon, William H. 1998. Wherefore Warehouse? Byte. (January): 88NA1-10.

Jarke, Matthias, and Yannis Vassiliou. 1997. Data Warehouse Quality: A Review of the DWQ Project, Invited Paper: Proc. 2nd Conference on Information Quality. Cambridge. Massachusetts Institute of Technology.

Kelly, Thomas J. 1998. Dimensional Data Modeling. Internet. Available from http://www.gca.net/solutions/whitepapers/sybase/syb_dim_data_mod.html ; accessed 25 October 2001.

Kimball, Ralph. A Practical Method for Planning a Data Warehouse. Data Webhouse. Internet. Available from http://www.intelligententerprise.com/db_area/archives/1999/990712/webhouse.shtml; accessed 3 September 2001.

Kimball, Ralph. 1996. The Data Warehouse Toolkit. New York. John Wiley & Sons, Inc.

Microsoft Corp. 2001. Microsoft SQL Server 2000 Resource Kit. Redmond. Microsoft Press.

Moriarty, Terry. 1995. Modeling Data Warehouses. Database Programming & Design. (August). Internet. Available from http://www.inastrol.com/Articles/9508.htm; accessed 5 October 2001.

SCT Corp. 1997. SIS Technical Guide. Rochester. SCT Corp.

APPENDIX A

DATA WAREHOUSE SURVEY

1. Rank the following five items in order of importance to you.

_____ Number of hours the system is available.

_____ Days and times the system is available.

_____ Ability to view detail data when needed.

_____ Ability to download and manipulate desired data.

_____ Same or better access to data currently accessed from SIS.

_____ Other (please explain) _____

2. How would you rate the importance to you of the ability to extract "raw" data (i.e. data in text format) from the data warehouse to manipulate at the desktop level, for example, in Excel or Access?

_____ Very important

_____ Somewhat important

_____ Not important

3. For standard reports that you will be processing frequently, what is the maximum acceptable response time?

4. Indicate your current level of use of the SIS system.

_____ I use the system on a daily basis

_____ I use the system several times a week

_____ I occasionally use the system during the semester

_____ I rarely or never use the SIS system

5. In situations where data from the data warehouse conflicts with SIS reports, what would be the most compelling evidence to you that the data warehouse data is correct?

_____ The detail data that was used to derive the summary

_____ Documentation that shows the source(s) of the data

_____ Other (please explain) _____

6. Select the option that best describes your preference for producing reports.

_____ I would like an icon on my desktop that will automatically display my reports

_____ I would like to be able to select from a list of pre-defined reports

_____ I want to be able to produce my own reports using a graphical interface

_____ I want to be able to produce my own reports from a command line interface

_____ Other (please explain) _____

7. What is the minimum number of hours of up time per day that you consider acceptable?

8. Describe your current level of *viewing* access to data in SIS.

_____ All or partial access to student data

_____ All or partial access to instructor data

_____ All or partial access to course data

_____ Unlimited access

_____ Other (please explain) _____

9. For ad hoc reports, what is the maximum acceptable response time?

10. Rank the following in order of your preference regarding documentation for a software application.

_____ Hardcopy user's manual

_____ Online documentation

_____ Context-sensitive help

_____ Diagrams and charts

11. If you are using a graphical interface to create your own reports, what is the maximum number of steps that you want to perform to produce the report?

12. What days of the week would you require the data warehousing system to be available to you?

13. If a particularly complicated ad hoc report required 30 minutes to run, would you use the report again? If it required 1 hour?

14. What best describes your current use of documentation for software applications.

_____ I frequently use hardcopy manuals

_____ I normally use online documentation only

_____ I will call someone for help using the application

_____ I do not use software applications that are not readily usable

15. Considering the reports you currently receive from the SIS system, how do you go about reconciling questionable results from these reports?

16. How do you prefer to view report data, or what format is most meaningful to you? (Rank in order)

_____ Graphical – graphs, pie charts, etc.

_____ Tabular format

_____ Formatted report

_____ Other (Please explain) _____

17. Considering an analytical processing system, if you had to choose between a system that is available to you during normal working hours and one that is available evenings and weekends, which would be most useful to you?

18. How frequently will you require data to be refreshed in the data warehouse?

_____ Daily

_____ Weekly

_____ Monthly

_____ At the end of each semester

_____ Variable throughout semester

_____ Other (please explain)_____

19. Select the level of detail that best describes the level of detail of data that you require. Each data perspective (student, instructor, course, semester) will be referred to as a "dimension". You may select all that apply.

_____ Detail by student dimension

_____ Detail by instructor dimension

_____ Detail by course dimension

_____ Detail by semester dimension

_____ Summaries at all dimensions

20. If the data in the warehouse were refreshed on a weekly basis, would you use the warehouse for producing reports?

21. Rank the following documentation items in order of importance to you.

_____ Origin of data element (i.e. the "system of record")

_____ Data element mapping to original system

_____ Date the element was last refreshed

_____ Modifications to data (summary)

_____ How the data was reconciled

_____ Definition of data element

APPENDIX B

DATA WAREHOUSE EVALUATION

A. Accessibility of Information

For each of the following questions, please put a check beside the answer that best represents your opinion.

1. There is at least as much information in the data warehouse reports as in reports from the SIS system.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

2. The data warehouse reports contain information that has not been available to me in SIS.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

3. How would you compare the data warehouse reports to SIS reports in terms of understandability?

____ a. Data warehouse reports are harder to understand

____ b. Data warehouse reports are easier to understand

____ c. They are about the same

____ d. No opinion

In data warehouse terms, “drilling down” is the ability to view summarized data in more detail. The Grade Distribution report is an example of drilling down. Please answer the following questions about the Grade Distribution report.

4. The information in the drill down reports is meaningful to me.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don’t know

5. The detail information is easier to get to than in SIS.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don’t know

6. The drill down reports are harder to understand.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

B. Using Pivot Tables to Produce Reports

If you are not using the Excel Pivot Table tool, please proceed to section C.

To answer the questions below, record the number of steps required to produce each report. Begin counting steps after you have the Grade Distribution pivot table visible in Excel.

____ 7. List the grade distributions by course for all Development Studies courses.

____ 8. List the grade distributions by on-campus classes vs. off-campus classes.

____ 9. List grade distributions for Basic level courses.

____ 10. List grade distributions by instructor.

____ 11. List total grade distributions by semester

Using the Pivot Table – General Perceptions

Place a check mark next to the answer that best describes your opinion.

12. I found the Pivot Table tool easy to use.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

13. The reports produced by the Pivot Table were easy to understand.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

14. I would like to have more reports available in this format.

____ a. Strongly agree

____ b. Agree

____ c. Disagree

____ d. Strongly disagree

____ e. Don't know

C. General Perceptions of the Data Warehouse

15. Would you continue to use the data warehouse for reporting?

16. Where possible, would you use the data warehouse instead of SIS for reporting?

17. Would you like to see the data warehouse expanded to include more types of data?

18. In your opinion, what was the best feature of the data warehouse reports?

19. If you could make one change to the data warehouse, what would it be?

VITA

MARGARET C. LESTER

Personal Data:	Date of Birth: December 2, 1961
	Place of Birth: Kingsport, Tennessee
	Marital Status: Married
Education:	East Tennessee State University, Johnson City, Tennessee; Music Education, B.M.Ed., 1984
	East Tennessee State University, Johnson City, Tennessee; Information Science, B.S., 1992
	East Tennessee State University, Johnson City, Tennessee; Information System Science, M.S., 2003
Professional Experience:	Director of Computer Services, Northeast State Technical Community College, Blountville, Tennessee, 1997 – Present
	System Manager/Programmer, NSTCC, Blountville, Tennessee, 1993 - 1997
Honors and Awards:	Member of Upsilon Pi Epsilon Computing Science Honor Society
	First Place Poster Presentation, Tennessee Academy of Science Annual Meeting, November 2002