Frederick McGovern

University of Denver

University College

ICT Program

Cathie Wilson, Instructor

ICT 4410 – Data Warehousing Design

Assignment 1

**Use the material from chapters 1 and 2 of the Marakas textbook, as well as notes, and posted articles and links to answer questions in this assignment.**

SECTION 1 (25 points)

Directions: The answer to each of the following questions is either True or False. Answer questions by placing a T or F in the cell immediately to the right of each question.

Answer ALL questions

|  |  |  |
| --- | --- | --- |
|  |  | T/F |
| 1 | According to Inmon, a data warehouse is a subject-oriented, integrated database designed to support DSS functions where the data is volatile and relevant. | F |
| 2 | The essence of the data warehouse concept is a recognition that the characteristics and usage patterns of operational systems used to automate business processes and those of a DSS are fundamentally similar and symbiotically linked | F |
| 3 | The finer the level of granularity, the greater the chance that unusual and previously unnoticed patterns will go undetected | F |
| 4 | One difference between data warehouses and operational data stores is the presence of *metadata* | T |
| 5 | A data warehouse topology requires a centralized data warehouse that is accessed by one or more decision support tools | F |
| 6 | The most common component of a data warehouse environment is the operational data store | T |
| 7 | Transformation mapping metadata records how data from operational data stores and external sources are transformed on the way into the warehouse | T |
| 8 | Data visualization involves examining data represented by numerical data and occasionally static images | F |
| 9 | One objective of the data warehouse environment is to minimize the impact on operational systems | T |
| 10 | According to Inmon, because data warehouses are highly de-normalized, they are highly redundant. | F |

SECTION 2 (25 points)

Directions: Answer each of the following questions by indicating the letter that corresponds to the correct answer. There is only one right answer for each question. Place answer letter in the cell immediately to the right of each question.

Answer ALL questions

|  |  |  |
| --- | --- | --- |
|  | Question | Answer |
| 1 | Which of the following is not true of a data warehouse?  a. Implicit in its definition is that the data warehouse is physically separated from all other operational systems.  b. The data warehouse replaces the need for all other reporting systems within an organization.  c. The data warehouse holds aggregated data and atomic data for management.  d. None of the above. | B |
| 2 | Which of the following is not true of a data mart?  a. The data mart is often viewed as a way to gain entry into the realm of data warehouses and to make the mistakes on a smaller scale.  b. Vendors of data warehouse applications have found it easier to deal with a small group of isolated users than with the IS department of an entire organization.  c. The data mart is more efficient than a fully-developed data warehouse.  d. None of the above. | C |
| 3 | The essence of the data warehouse environment is that the data contained within the boundaries of the warehouse are \_\_\_\_\_\_\_\_\_\_\_\_\_.  a. integrated  b. consistent  c. streamlined  d. accurate | A |
| 4 | Which of the following is not a characteristic of a data mart:  a. It obtains its data from relatively stable, cleansed and integrated sources  b. It is a set of tables designed for direct access, to analyze data according to predefined parameters  c. It is a reflection of the business rules of the business enterprise  d. It is a set of tables designed for aggregation | C |
| 5 | Which of the following activities would not normally be associated with a data warehouse?  a. Loading  b. Updating  c. Accessing  d. None of the above | B |
| 6 | The *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* represents the source data for the DW. This layer is comprised, primarily, of operational transaction processing systems and external secondary databases.  a. information access layer  b. operational and external layer  c. data access layer  d. process management layer | B |
| 7 | The *\_\_\_\_\_\_\_\_\_\_\_\_\_\_* serves as a sort of interface or middleman between the operational and information access layers and the data warehouse itself. This layer spans the various databases contained within the DW and facilitates common access by the DW users.  a. data access layer  b. application messaging layer  c. information access layer  d. None of the above | A |
| 8 | Which of the following would not be a good example of metadata?  a. The directory of where the data is stored.  b. The rules used for summarization and scrubbing.  c. Where the operational data came from.  d. All of the above are examples of metadata | D |
| 9 | In data mining, exploration for the discovery of important relationships in data may involve which of the following approaches:  a. analysis of summary and “sniffing out” unusual patterns  b. analysis of data samples to discover previously undetected patterns  c. heuristic pattern searching  d. all of the above | D |
| 10 | The process that records how data from operational data stores and external sources are transformed on the way into the warehouse is referred to as:  a. summarization algorithms.  b. transformation mapping.  c. back propagation.  d. extraction history | B |

SECTION C (50 points)

Directions:

Provide complete answers to ALL the following questions based on research and investigation of the topic. Use in-text citations to references and properly cite all references in CMS format.

1. *What are the characteristics of a data warehouse?*

Marakas formally defines four characteristics of a data warehouse.

The first characteristic is that the data warehouse is subject orientation. This refers to the fact that the data warehouse is used for decision making so it needs to be organized around the major subject areas that a company deems important. Subject orientation also means that the data in the warehouse can have many relationships between the tables according to a variety of business rules. (Marakas pg 46)

The second characteristic of a data warehouse is data integration. Data added into a data warehouse can come from many different sources. As such, similar data points may have different naming conventions or measurements. A good example is that one application may indicate gender as M and F while another might use 0 and 1. Data integration ensures that the data in the warehouse has consistency both in naming and units of measurement. (Marakas pg 47)

In a normal operational transactional application, the data in the database is current when the data is accessed. Inmon refers to this as “current-value data” (Inmon pg 33). This is not normally the case in a data warehouse. A third characteristic is that data warehouses are time variant. Instead of a guarantee that the data is currently accurate, a data warehouse would be accurate at some period of time when the data was loaded into the data warehouse. So if one were to look at data loaded 5 years ago, it doesn’t necessarily represent a current look of the data but is instead a look back in time to what the data looked like 5 yrs ago. The different time periods that are loaded into the warehouse are what is used to identify trends in the data. To ensure that the data contains a time stamp of when the data was loaded, it is necessary that the record key is related to a specific time period. (Marakas pg 49)

Marakas mentions one final characteristics. A data warehouse does not normally change or update the data in the warehouse. This means that the database has nonvolatility. Since the designer doesn’t need to be worried about updates, they can focus on optimizing the data warehouse data access. Furthermore, while operational application databases are normalized with no calculated data, the data warehouse design may include aggregate calculations to future improve query optimization. (Marakas pg 50).

Additional characteristics such as a data warehouse utilizes metadata, is comprised of a large amount of data, is not normalized and is summarized into a format that allows decisions to be made, are identified by Marakas in table 2-2 in his book Modern Data Warehousing, Mining, and Visualization. (Marakas pg 46)

1. *List and explain the different layers in the data warehouse architecture.*

The first layer in the data warehouse architecture is the operational and external database layer. This layer if where extraction software is used to incorporate data from a variety of sources into the data warehouse. Marakas refers to this layer as the “source data for the DW”. Its important that performance on the operational application not be impacted by the extraction of data. (Marakas pg 51)

The second layer is the information access layer. This is the layer that contains all tools that would be used by a data analyst to extract and analyze data in the data warehouse. Some of these tools include software and hardware necessary to print reports and develop spreadsheets and charts. (Marakas pg 52)

The third layer is the data access layer. This layer allows for access by the data warehouse users. It allows all users to access data no matter what database the data might reside on in the data warehouse or what tool was used to access the data. It is an interface between the operational and external database layer and the data warehouse. (Marakas pg 52)

The metadata layer is the fourth layer. To enable data access to everyone, a directory of where the data is stored must be maintained. This is an example of metadata. The metadata layer contains all of the metadata information. (Marakas pg 52).

The application messaging layer is known as “middleware”. It is the layer that is used to send information throughout the computer network. Marakas indicates that this layer “can be used to isolate applications, operational or informational, form the extract data format on either end.” (Marakas pg 53).

The physical data warehouse layer is where the data is actually located. Data doesn’t necessarily have to reside in the data warehouse but instead can utilize a view to access the data through the warehouse (Marakas pg 53).

The final layer is the data staging layer. It includes any processes needed to select or combine data from the operational/external databases for loading into the data warehouse. These processes can be quite complex which is why a market has developed to provide tools to reduce this complexity. (Marakas pg 53)

1. *What are metadata? Why are they so important to a data warehouse?*

I was always taught that metadata was ‘data about data”. It provides a description of the lower level data. Examples would include Author, Publisher, maps, and field definitions for all data in the data warehouse. Since I work in insurance, if I see a value such as 9934232901321 the metadata establishes that the meaning of the data ie it’s a customer policy number. Without metadata, I wouldn’t know if it was a policy number or some other data point such as a complex key. Metadata also provides clarity in the data. For example, the date 4/5/2015 would seem to suggest a date in April. However, the data format could be utilizing the European convention and is instead a date in May. The metadata provides a descriptions of what the data actually represents. As Marakas points out, metadata is important to a data warehouse as it allows for recognition of patterns since the meaning of the data is clear to all analysts. (Marakas pg 56).

1. *In data warehousing terms, what kind of user is a “farmer”? How would you characterize an “explorer” user?*

Marakas identifies a farmer as someone who knows what correlation or pattern they are looking for in the data before they start. Thus they are highly predictable submitting queries on a routine basis. The queries are typically small in size and seldom return any substantial amounts of information beyond finding what the analyst was looking for. According to Inmon, a farmer is the foremost user found in a data warehouse. The types of queries submitted by farmers are very similar, varying usually only in the type of data involved in the query (Inmon pg 458)

An explorer on the other hand is very unpredictable, They might submit one query one day, ten the next and zero on the following day. Marakas characterizes these end users as “out of the box” thinker (Marakas pg 16). Queries tend to very long and large and very frequently produce no real insight. Explorers don’t know what they are looking for prior to submitting the query. Inmon refers to explorers as operating in “heuristic mode” (Inmon pg 458). This means that the steps in conducting an analysis depend entirely on the results produced by the current step.

1. *In data mining, what are “false positives” , and why should we be concerned with these?*

Data mining began as statistical analysis of the data and has since evolved to include things such as heuristic reasoning and neural networks. The point of these approaches is to identify a correlation between data points. A perfect correlation would be one in which for every occurrence of data point A there also exists data point B. This would be considered a strong correlation of data but its extremely rare to find a perfect correlation. A normal strong correlation would be for every data point A most of the time there is a data point B.

Sometimes there appears to be a relationship between the data, however, the relationship is purely random. It may in fact be caused by the set of data that the analysis is being conducted upon and disappears when a new data set is analyzed. These are known as spurious correlations (Marakas pg 26)

The problem that can occur is in the area of data classifications. The relationship between two data points, can cause one data point to be classified incorrectly. This is known as a false positive. For example, in the insurance industry there is a statistical relationship between an insured credit score and claims. The higher a person’s credit score, the less likely they are to file a claim so they might get classified as a good driver. If the person is in fact a bad driver then the classification would be a false positive

Developing a predictive model requires the use of a training set. The higher the proportion of false positives in your data set ultimately increase the number of false positives that get produced. This is especially true if utilizing an unsupervised learning paradigm as the pattern for classification is determined by which pattern best matches that of the input data (Inmon pg 135). If the number of false positives equals the number of true positives, the classification is in fact meaningless.

References

Inmon, William H. *Building the Data Warehouse*. 4th ed. Indianapolis, IN: Wiley Pub., 2005.

Marakas, George M. *Modern Data Warehousing, Mining, and Visualization: Core Concepts*. Upper Saddle River, NJ: Prentice Hall, 2003.