University of Denver

University College

ICT Program

Cathie Wilson, Instructor

ICT 4410 – Data Warehousing Design

Assignment 2

**NAME: \_\_Frederick McGovern\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Use the material from the reading assignments for units 3 and 4 (Marakas ch 3.1-3.3, and 5 and Inmon ch 4-5 and 7-8), posted presentations and outside sources 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 | A sales report showing revenue against forecast targets along with projections for the next two quarters is an example of an EIS output. | T |
| 2 | A well designed EIS will replace the need for MIS applications. | F |
| 3 | An EIS is a special type of DSS designed to facilitate the analysis of information to support strategic decision-making. | T |
| 4 | In the structural perspective of the EIS development framework, the focus is on people and data as they relate to the EIS. | T |
| 5 | Sparcity can significantly increase the storage requirements of a MOLAP hypercube by requiring that space be allocated for all cells rather than just the ones that contain data values | T |
| 6 | The knowledge base in an EIS is the sum of what the executive knows about using the system and all of the support mechanisms designed to assist in its use. | T |
| 7 | The classification approach to data mining searches all details or transactions from operational systems for patterns with a high probability of repetition | F |
| 8 | The clustering approach to data mining is useful when there is a need to create partitions in order to discover patterns in the data | T |
| 9 | The ROLAP structure contains a large number of normalized tables | F |
| 10 | Data mining does not use statistical techniques because the complex patterns in data do not lend themselves to linear regression analysis | 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 an activity performed by managers?  a. Staffing  b. Directing  c. Organizing  d. All of the above. | D |
| 2 | Which of the following is a computer-based system intended to facilitate and support the information and decision-making needs of senior executives by providing easy access to both internal and external information relevant to meeting stated goals of the organization?  a. EIS  b. ES  c. DSS  d. Data Warehouse | A |
| 3 | Which of the following is not true of data mining?  a. Data mining has seen explosive growth in the area of customer relationship management.  b. Data mining is a business solution.  c. Data mining is a technology.  d. None of the above are true | B |
| 4 | Which method for determining executive needs gathers information from a sample of top-level executives concerning the totality of information needs?  a. Null method  b. Normal method  c. By-product method  d. Total study method | D |
| 5 | Which of the following is a reason for the growth in popularity of data mining?  a. Increased volume of data  b. Increased awareness of the inadequacy of the human brain to process multifactorial dependencies or correlations  c. Increased affordability of machine learning  d. All of the above | D |
| 6 | The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ method for determining executive information needs is based on three basic notions: 1) The health of an organization can be determined by comparison to a set of key financial indicators; 2) organizations can be managed based on exception reporting where only those areas operating outside of a preestablished set of norms are of interest; 3) and technology is available to allow for flexible display of key indicator information in graphical form.  a. key indicator  b. critical success factor  c. normal  d. by-product | A |
| 7 | In a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, the data are stored as a multidimensional array where each cell in the array represents the intersection of all of the dimensions. Using this approach, any number of dimensions may be analyzed simultaneously and any number of multidimensional views of the data can be created.  a. hyperion cube  b. hypercube  c. stochastic cube  d. correlation cube | B |
| 8 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ activities fall under the general heading of strategic planning tasks. Executives spent a significant portion of their time focusing on the selection, design, and implementation of improvements and projects intended to improve performance and initiate controlled change within the organization.  a. Entrepreneurial  b. Disturbance management  c. Resource allocation  d. Negotiation | A |
| 9 | Which of the following is not a component of an EIS development framework?  a. Structural perspective  b. Intelligent knowledge base  c. User-system dialog  d. None of the above. | B |
| 10 | There are some cases where it is difficult or impossible to define the parameters of a class of data to be analyzed. In these cases, \_\_\_\_\_\_\_\_\_\_\_\_\_ methods can be used to create partitions so that all members of each set are similar according to some metric or set of metrics thus creating a set of objects grouped together by virtue of their similarity or proximity to each other.  a. association  b. linkage analysis  c. sequencing  d. clustering | D |

**SECTION C (50 points)**

**Directions: Provide complete answers to ALL of 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. List and briefly describe several EIS limitations.***

Marakas on pages 172 through 174 describes numerous EIS limitations as well as pitfalls to avoid. The following is some of the limitations that are discussed:

Cost – Marakas identifies a survey that shows that the mean development costs to implement an EIS were $365,000. As this study was conducted in 1991, I am sure that the costs assessment would be completely different now. However, the cost would still be large and the cost of maintaining the system can easily exceed the initial investment. Marakas points out that “benefits derived from EIS are costly” and an “ongoing commitment of significant organizational resources is critical if the expected benefits are ever to be realized (Marakas pg 172).

Technology Limitations – There is a significant amount of dispersed data of various data formats that must be combined into the EIS. This can be a challenge to EIS designers due to the necessity to work with multiple programming languages, data formats and file structures. As few developers would be versed in every programming layout or data format encountered, it is necessary that the EIS developer seeks out support from other members of the IT department to fill in knowledge gaps (Marakas pg 172).

Organizational Limitations – besides technical limitations, there can exist organizational limitations to developing an EIS. Marakas points out three main categories of potentially negative impacts.

The category is Agenda and Time Biases. This refers to an executive becoming overly reliant of the data that the EIS presents. The problem is that while the EIS certainly provides a measurement of the data it doesn’t provide all the information that an executive would need to understand whether or not the data was both “relevant and necessary” (Marakas pg 173).

The second category is Managerial Syncronization. This refers to the fact that the EIS allows the executive to create Adhoc reporting. Overly reliance on adhoc reporting instead of the established reporting cycles can lead to a disruption to those reporting cycles. This can lead to a loss in management synchronization of working toward a common strategic goal (Marakas pg 173).

The third category is Destabilization. This refers to the fact that an EIS allows executives to make decisions on a much faster manner. However, the quickness in response must be tempered in the change that is actually implemented. Too many changes or changes that are large in nature can destabilize how the company operates. The key is to not overreact to the what is being presented in the EIS. Instead the key is to “affect an average based on small variances in change.” (Marakas pg 173)

***2 Explain the characteristics, role, challenges and benefits of external data in the data warehouse.***

According to Inmon there are two basic types of external data. The first is external data that has been collected by some source ie such as from the Wall Street Journal, industry newsletters, credit reports and competitor sales analysis. This would also include reports internal to the company such as annual report. The second is external data that comes from random articles, reports and other sources.

One problem with the usage of metadata is that external data doesn’t usually have a fixed pattern of appearance. This requires the utilization of monitoring programs to initiate automated notifications or data loads. Another issue with external data is that its undisciplined. That means that the external data requires reformatting in order to transform it so that its compatible with the internal data in the data warehouse. Frequently, the external data is at a granularity that doesn’t match the granularity of the internal data.

It is extremely important to store the metadata associated with the external data. Without such metadata, future users would have no idea of what the context of the data is or whether or not it was from a reliable source. The external data metadata would encompass such things as a document id, date that the data was added into the warehouse, the source of the data and a description of what the data is. Inmon points out that managers typically look at the metadata as opposed to the source document as scanning the metadata helps to quickly identify what documents are relevant and those that are not (Inmon pg 262).

A determination of whether or not the external data should be stored in the data warehouse needs to be made. Inmon points out that it may not be possible or economical to in fact store the data and instead a data entry would be made in the data warehouse identifying where the data can be found. The problem with doing this however, is that it opens up the possibility that other departments might store the same data. By having the external data entirely stored in the data warehouse, it becomes available to all departments that utilize the data warehouse. (Inmon pg 263).

***3 Describe how relational on-line analytical processing (ROLAP) works. (Ch3 - Marakas)***

Relational on-line analytical processing (ROLAP) replaces the multidimensional database server with a very large relational database server. This relational database can have dual granularity containing very detailed data or data that has been summarized. This approach lends itself well to the ability of drilling down from summary data to more detailed data. ROLAP does provide for a large number of administration tools with open SQL interfaces thus increasing portability and scalability. A negative of ROLAP is that the large number of database tables do cause a significant degrade in performance as compared to a MOLAP implementation. One method to reduce this affect is to denormalize the normalized tables into a centralized fact table with surrounding dimensional tables. This is known as a star schema (Marakas pg 76).

**4 *Explain the concept of multidimensional on-line analytical processing (MOLAP).***

Multidimensional on-line analytical processing (MOLAP) is a method of analyzing data across multiple dimensions. This could be as simply as adding a time dimension to the analysis of two other dimensions. However, most MOLAP’s store and analyze data as a n dimensional cube, which Marakas indicates is known as a hypercube (Marakas pg 76). The data in a hypercube is stored as a multidimensional array so that every cell represents the intersection of all dimensions. Since not every data point will have a corresponding dimensional relationship with another data point (ie not all products may have been sold in a given time dimension), there exists the potential for a lot of empty cells. This potential increases with the number of dimensions added to the cube. This is known as sparcity which can result in a need to dramatically increase the needed storage requirements.

**5. *According to Inmon, “the data warehouse requires a simpler set of technological features than its operational predecessors”, at least from a data integrity perspective. Discuss some of the technological requirements that data warehouse designers and developers have to address***

The quote above is from Inmon’s Building the Data Warehouse on pg 159. He further explains his believe by pointing out that data warehouses don’t need online updating which eliminates the need for locking and integrity checks. Despite this, Inmon points out in Chapter 5 that “there are a fair number of technological requirements for the data warehouse”.

Inmon points out that the first and most important technological requirement is for the data warehouse to be able to manage large amounts of data potentially reaching terabytes or petabytes in size. This requires flexibility in accessing the data in the data warehouse. This might be accomplished through indexing, portioning, extensions of data of management of overflow. While its important that large amounts of data be able to be managed, it is also equally important to effectively manage data. As Inmon points out, “the technology used must satisfy the requirements for both volume and efficiency (Inmon pg 161).

A second technological requirement is the ability to manage multiple storage media. This media includes main memory, expanded memory, cache memory. Additional management of media such as magnetic tape, optical disk and fiche may also need to be considered. Each type of media will have a specific speed of access and cost for storage. Due to the volume of data contained in a data warehouse, it is typical to have data reside on more than one type of storage media (Inmon pg 161).

A third technological requirement is the ability to monitor data and index data both freely and easily. The database technology being utilized should allow for both primary and secondary indexing as well as dynamic and temporary indexes. The index should be easy to create with minimal performance degradation to create the index. The monitoring of data helps to answer a number of questions. These questions include whether or not an index is poorly structured, how much data is in overflow and whether or not that amount is excessive, the availability of space and whether or not reorganization needs to be done (Inmon pg 162).

A fourth requirement is whether or not the data warehouse has the ability to interface with other technologies. This would include such things as whether or not the data can pass easily from one database management system to another, whether or not data passage for multidimensional processing can be done, and whether or not the context of the data is lost when the data is moved to a different environment or operating system. The interface needs to be easy to use, efficient and have the ability to operate in batch mode (Inmon pg 163).

In Chapter 5, Inmon discusses sixteen different technological considerations. I have listed in detail the first four. Other considerations include:

* + The ability of the designer to place data physically at the block/page level.
  + The ability to manage data in parallel.
  + To have solid meta data control.
  + To have a rich language interface to the data warehouse.
  + The ability to load the warehouse efficiently.
  + That indexes are efficiently used.
  + That data can be compacted.
  + Provides for support of compound keys.
  + The ability to manage efficiently variable length data.
  + Have the ability to turn off or on the lock manager.
  + Ability to do index only processing
  + Have the ability to restore data from a bulk medium quickly and completely.

(Inmon 164 – 170)

This last point is especially important to ensure disaster recovery,

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.