University of Denver

University College

ICT Program

Cathie Wilson, Instructor

ICT 4410 – Data Warehousing Design

**Final Exam**

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

**Use the material from the textbooks, lecture notes, and outside sources to answer questions in this exam.**

**SECTION 1 (20 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**

|  |  |  |
| --- | --- | --- |
|  | **QUESTION** | **Answer**  **(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.  (True - See a Typical EIS Session Marakas pg. 158) | T |
| 2 | The structural perspective of the EIS development framework builds upon the elements and their relationships contained within the structural model and adds a time dimension to the framework. In this component, the various activities and sequences of events are delineated and actual project management issues relating to time, critical path, and milestones are established.  (False - This would be EIS Development Process Marakas Pg 170) | F |
| 3 | The very essence of the DW environment is that the data contained within the boundaries of the warehouse are integrated. This integration manifests itself through consistency in naming convention and measurement attributes, accuracy, and common aggregation.  (True - See Data Integration section – direct quote Marakas pg. 47) | T |
| 4 | One of the primary advantages of ROLAP versus MOLAP is it results in performance improvements for data access.  (False - The extreme processor overhead results in degraded performance – Marakas pg. 76) | F |
| 5 | The majority of executive time is spent in some form of disturbance management activity.  (True – Figure 5-1 42% Handling Disturbances Marakas Pg 161) | T |
| 6 | According to Inmon, because data warehouses are highly de-normalized, they are highly redundant.  (False – Inmon suggests that data redundancy is a rare occurrence Marakas Pg 51) | F |
| 7 | *Transformation mapping* metadata records how data from operational data stores and external sources are transformed on the way into the warehouse.  (True – First sentence section *Transformation Mapping* Marakas pg. 57) | T |
| 8 | The sequencing approach to data mining relates events in time, based on a series of preceding events.  (True – techniques that use sequencing relate events in time based on prior events Marakas pg. 79) | T |
| 9 | The ability to be proactive in identifying legal issues, thereby foregoing potential financial penalties that might impact a firm would be considered an intangible project benefit.  (True -Table 6-5 Proactive addressing of legal issues Marakas Pg 197) | T |
| 10 | MOLAP is well-suited to handle large numbers of detailed data.  (False – Good at summary data not well suited for large numbers of detailed data Marakas pg. 76) | F |
| 11 | An EIS normally requires unique hardware configurations, making them expensive to build.  (False – EIS Components Pt 2 Reduction in cost of hardware Marakas pg. 166) | F |
| 12 | Organizational culture can have an impact on their readiness to develop a data warehouse because data warehouse development implies cross departmental cooperation and information sharing.  (True – See Analytic Culture of Organization Marakas pg. 188 and Table 6-2 Marakas pg. 190) | T |
| 13 | The GA’s smallest unit is called a chromosome. The chromosome represents the smallest unit of information in the problem domain and can be thought of as the basic building block for a possible solution.  (False – A GA’s smallest unit is called a gene Marakas pg. 140) | F |
| 14 | One of the limitations of a fuzzy logic system is that each rule is dependant upon all of the other rules in the knowledge base.  (False – limitations are cant learn from mistakes and during high complexity unsure which rules are firing Marakas pg. 130) | F |
| 15 | The human language has evolved to allow for the conveyance of meaning through semantic approximation rather than precise content.  (True – See Linguistic Ambiguity 2nd paragraph Marakas pg. 126) | T |
| 16 | Fuzzy logic is a new development in the computing industry that emerged as a result of the development of Internet search engines.  (False – approach to logic that dates back to the days of Plato Marakas pg. 127) | F |
| 17 | The end user in a data warehouse environment deals directly with the application messaging layer.  (False – This would be the Data Access Layer Marakas pg. 52) | F |
| 18 | An EIS normally requires unique software components.  (True – software components are normally highly specialized to meet specific needs Marakas pg. 166) | T |
| 19 | 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.  (True – The third element. Sum of what the what the executive knows about using system Marakas pg. 171) | T |
| 20 | One of the limitations of a fuzzy logic system is that each rule is dependent upon all of the other rules in the knowledge base.  (False - Duplicate of #14) | F |

**SECTION 2 (20 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.**

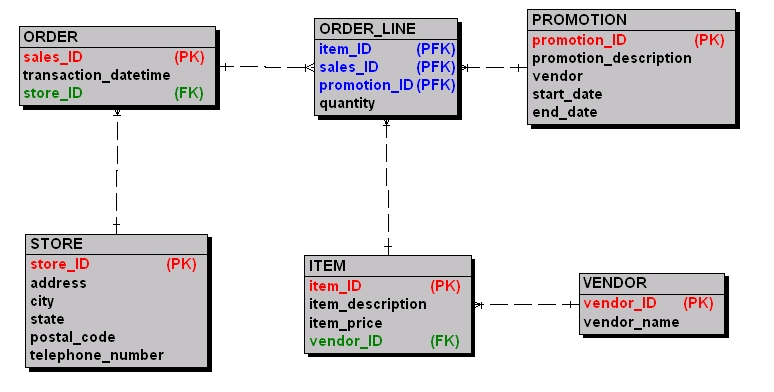
|  |  |  |
| --- | --- | --- |
|  | **QUESTION** | **Answer** |
| 1 | Which data mining technique utilizes linkage analysis to search operational transactions for patterns with a high probability of repetition?  a. Association (See Association section Marakas Pg 79)  b. Cluster  c. Sequence  d. Principal components analysis | A |
| 2 | Which of the following would not be considered a benefit of a client/server environment?  a. Reduction of investment costs in new computer hardware (No - Bullet 2 pg. 166)  b. Establishment of a more flexible and responsive platform (No - Bullet 3 pg. 166)  c. Increased control over locally stored data  d. Multiple views of geographically dispersed data (No - Bullet 1 pg. 166) | C |
| 3 | Using an EIS to explore the underlying data used to compile the current ratio would be an example of which of the following activities?  a. Compilation  b. Aggregation  c. Drill down analysis (See A Typical EIS Session Marakas Pg 158-159)  d. None of the above. | C |
| 4 | The concept of time variant data implies which of the following statements?  a. Data are simply assumed to be accurate as of some moment in time and not necessarily *right now*. (True Marakas pg. 49 Time Variant near end 1st paragraph)  b. Data are assumed to be accurate at the moment they were loaded into the data warehouse. (Also true – next sentence after above)  c. Data are assumed to vary over time.  d. Both a and b | D |
| 5 | The *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* applied to the detail data are of importance to any decision maker analyzing or interpreting the meaning of the summaries. These metadata can also save time by making it easier to decide which level of summarization is most appropriate for a given analysis context.  a. summarization algorithms (See *Algorithms for Summarization* 2nd sentence pg. 58)  b. transformation mapping  c. back propagation  d. extraction history | A |
| 6 | Which of the following is not a method for determining executive information needs?  a. Null method (No – 1st method listed Marakas pg. 164)  b. Normal method  c. By-product method (No - 2nd method listed Marakas pg. 164)  d. Total study method (No – 4th method listed Marakas pg. 165) | B |
| 7 | Which of the following is an advantage of fuzzy systems?  a. Allows for increased association amongst the rules of the system  b. Supports modeling of contradiction (First advantage listed see Marakas Pg 129)  c. Increase system verification capability (No actually cant learn from mistakes)  d. All of the above. | B |
| 8 | Which of the following is not a tangible benefit?  a. Increase in process quality  b. Reduction in employee turnover (See Table 6-5 Marakas Pg 197)  c. Improving time to market  d. Decreased need for travel | B |
| 9 | The ANN layer that relays the final results of the net is referred to as the:  a. input layer.  b. output layer. (Putting a Brain in a Box 2nd paragraph Marakas pg. 132)  c. hidden layer.  d. neural layer. | B |
| 10 | Which layer of the data warehouse architecture does the end user deal directly with?  a. Data access layer  b. Application messaging layer  c. Information access layer (See first sentence in this section Marakas Pg 52)  d. None of the above. | C |
| 11 | \_\_\_\_\_\_\_\_\_\_\_\_ determines the point where the benefits of a project equal its costs.  a. Net present value  b. Breakeven analysis (See Marakas Pg 202)  c. Internal rate of return  d. All of the above. | B |
| 12 | The \_\_\_\_\_\_\_\_\_\_\_\_\_ is not normally connected to the ANN and is assumed to have an input value of 1.0 for the state function. Its purpose is to allow for the individual adjustment of the firing threshold of the neurode to facilitate the final adjustment of the ANN following the learning process.  a. input layer  b. bias input (See *The Bias Input* section Marakas Pg 133)  c. state function  d. output connection | B |
| 13 | Which term is used to refer to a basic database operation that links rows of two or more tables by one or more columns in each table?  a. n-cube analysis  b. table link  c. table join (See footnote 2 Marakas Pg 76)  d. None of the above. | C |
| 14 | Joining a fact table to dimension tables in a star schema is referred to as a:  a. star join. (See *Dimensional Modeling* section 2nd paragraph Marakas Pg 207)  b. data mart.  c. fact analysis.  d. dimensional warehouse. | A |
| 15 | A common example of the use of association methods where a retailer can mine the data generated by a point-of-sale system, such as the price scanner you are familiar with at the grocery store is referred to as:  a. sequencing.  b. linkage analysis.  c. clustering.  d. market basket analysis. (*Association* section 2nd paragraph Marakas pg. 79) | D |
| 16 | Which of the following is not a common characteristic of executives?  a. They are future-oriented and focus on strategic activities.  b. They have narrow spans of control. (All others listed in Table 5-2 Marakas pg. 160)  c. They are responsible for establishing policies.  d. They are concerned with internal and external issues. | B |
| 17 | Which of the following is not a benefit derived from neural computing?  a. Reduced need for experts.  b. Allows for generalization from specific information content.  c. Highly verifiable, especially for complex problems. (A and B listed on table 4-2 pg. 137)  d. All of the above are benefits. | C |
| 18 | Which of the following is not an intangible benefit?  a. Development of new markets (All other listed in Table 6-5 Marakas Pg 196)  b. Increased competitive advantage  c. More timely information  d. Increased workplace safety | A |
| 19 | The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_’s purpose is the consolidation of the weights of the various inputs to the neurode into a single value that can be passed to the transfer function for processing. The value obtained determines the degree of impact the combined inputs will have on the transfer function and, thus, on the final output of the neurode.  a. input layer  b. bias input  c. state function (*The State Function* section 2nd sentence Marakas Pg 133)  d. output connection | C |
| 20 | 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 (See *What Exactly Is An EIS?* Section Marakas Pg 157)  b. ES  c. DSS  d. Data Warehouse | A |

**SECTION C (60 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. Transform the following relational schema for a Sales Database into a Star Schema. At a minimum, you should make use of all the attributes listed in the ERD. Add attributes you think are needed for a meaningful representation of a data warehouse. Identify facts and dimensions. Explain your design reasoning. (10 points)

ER Design for a Sales Database



To create a Star Schema, one needs to create a centralized fact table along with associated dimensional tables. In my experience, textual data is often put into the dimensional tables while numeric data is put into the fact table.

The fact table will be the entity that is most heavily populated. Looking at the crow’s feet notation, it seems to me that all the attributes in the ORDER\_LINE entity should be included in the centralized fact table. I’ll rename this as the Sales fact table. This would mean that both the Promotion and Item tables would need to be transformed into dimensional tables. However, looking at the Item table we see the attribute item price. As this is a numeric data point it should probably (based on context) be moved to the Sales fact table. An argument could be made that we might want to add another column representing the total dollars (i.e. quantity X item\_price) but again without the context of who is using the reports, report performance and what data points they are interested in, I’ve decided to not add a calculated value

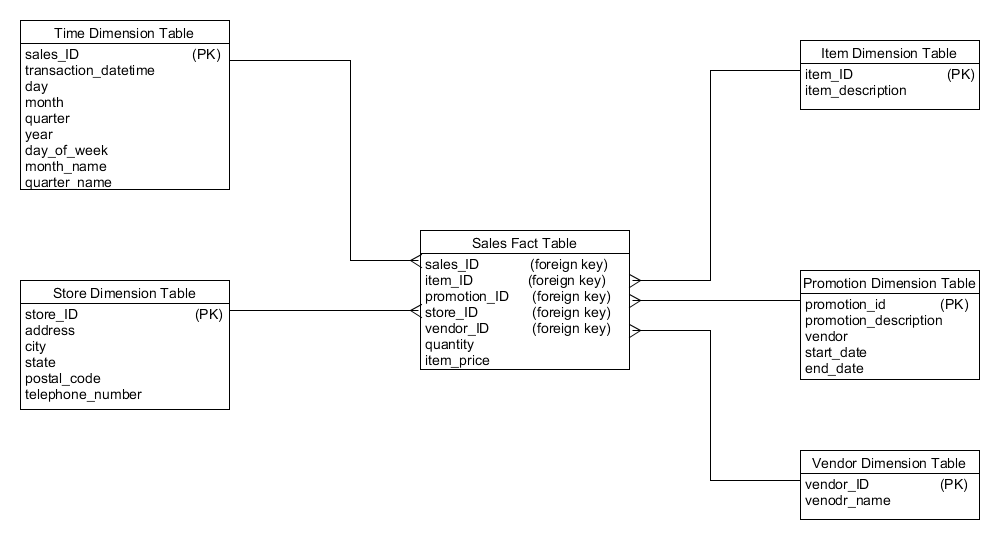
The ORDER table contains the SQL transaction date and since the sales\_ID will be in the fact table, the ORDER entity should be renamed to identify it as the Time Dimension table with the sales\_ID as its primary key.

Making these changes would lead to a snowflake schema as the Store table would be translated to a Store dimensional table but with its join not from the Sales fact table but from the new Time Dimension table. A similar situation exists with the table Vendor. In order to create a Star Schema, we need to take the primary keys of both the Store and Vendor tables and add them to the centralized Sales fact table.

I do have some concerns with the attribute vendor in the promotion dimensional table as well as a separate vendor dimensional table, but since there is no information provided in the diagram above to indicate that the data definitions are identical I will leave it as is. I assume it’s possible that the vendor sponsoring the promotion doesn’t have to be the vendor of the item being purchased.

One other consideration is the newly created time dimension table which currently would only contain the attributes sales\_ID (its primary key) and transaction\_datetime since store\_ID was moved to the Sales fact table. I’ve found from experience that it’s a good idea to have a time dimensional table that contains not only the date but indicators such as what day of the week, what week number, month and quarter the transaction date corresponds to. I’ve also had the need for a Julian date representation of the transaction date but will not add it here. Storing the actual values makes reporting quicker as you wouldn’t need to use SQL functions to transform the date when running the report.

Based on my analysis the final Star Schema would look like this:



1. *Explain the concept of data correlation within the data warehouse framework. What is the significance of data correlation in data mining? Provide various examples and explanations of data correlation.*

A data warehouse is composed of many tables containing various amount of data. By itself, a singular data point doesn’t tell us much as it lacks context but by comparing two or more data points it becomes information. Context is established through the relationship between the data. So in simplest terms data correlation is the mathematical relationship that exists between two or more variables.

In statistics, the Pearson’s correlation coefficient is either represented by the letter p when measuring population or the letter r when dealing with a sample. In most cases, analysts are dealing with a sample of the total data set so r is the most frequently seen variable. Marakas, on pg. 24 of his book, seems to agree with this by defining that correlation is represented by the letter r.

According to George Marakas, there are three types of correlations. The first correlation is known as a perfect correlation. Given a particular data point A, every time A occurs then B also occurs and vice versa. Marakas defines that a perfect correlation would be assigned a value of 1, based on a scale of 0 being no correlation exists to 1 being a perfect correlation exists. However, the Pearson’s correlation coefficient ranges from -1 to 1 with 0 being no relationship, a -1 representing a perfect negative relationship and 1 representing a perfect positive relationship. I believe the difference in definition is that Marakas is simply describing the strength of the relationship in his book as opposed to both strength and direction. I base this reasoning based on his discussion of the spectrum of strengths of correlation found on pg. 26. Despite the difference in definition, Marakas points out that a perfect correlations is an extremely sound basis for correlation. That would be true whether or not it was a perfect positive relationship of a perfect negative relationship. Unfortunately, this type of relationship is extremely rare.

The second correlation defined by Marakas is the strong correlation. In this case, for every A data point occurrence, the data point B usually show up. There are in fact a few times when B doesn’t. Flipping that around, there could be a few times that when you have a B you might not find a matching A data point. Marakas points out that this type of correlation is fairly common and defines that the strength of the correlation variable r would usually be between .25 and .75. Obviously, correlation values above .75 can be found and would be classified as strong correlations, they are just not as common. A strong correlation in terms of the Pearson’s correlation coefficient would also include the negations of this values to indicate the direction of the relations. Regardless, strong correlations represent a relationship that can frequently be exploited to the benefit of a company.

The third correlation discussed by Marakas is the weak correlation. In this case, when A occurs, B might occur a few times. When B occurs, occasional an A will also be present. It could in fact appear that the data points A and B exists independently from one another. While a weak correlation might on the surface appear useless, an analysis of weak correlations over a period of time might indicate a strengthening relationship. A strengthening relationship may indicate a trend that can be exploited by the company. Another analysis that can be done is to identify the correlations of subsets of either A or B. In other words, while a relationship doesn’t appear to exist while looking at the entire population, one may exists when looking at a subset of the population. If a strong correlation then exists on the subset, it allows for exploitation by the company by targeting just that subset of the overall population

Marakas points out that data correlation is at the heart of data mining. It is can be argued that it’s the most important thing one can do with a data set. Data correlation helps the business analysts identify not only why particular patterns exists now but also help to identify future trends. Since data correlation is a mathematical relationship it provides a strong foundation for executives to make business decision. This is one of the reasons why companies invest in data warehouses. Data mining leads to the identification of relationships based on mathematics. These relationships, once identified, can be exploited by companies in an effort to meet established goals such as decreasing costs and increasing revenues.

1. *The internet (world wide web) is a source of a vast amount of information about data warehousing. It is in fact a data warehouse itself. Viewing the web from the perspective of a data warehouse, identify and describe the various warehouse components and provide examples of how an organization might harness the power of the web as a data warehouse.*

It is interesting to think about the evolution of information about you. Long before the introduction of computers, the main method of purchasing an item was to actual visit the business selling the product. If you frequented the establishment enough, the owner got to know you and would start to make recommendations based on purchases that you made. Businesses and technology evolves so that they had to the ability to add direct selling through mail and phone. The advent of computers allowed for the tracking of purchases, allowing for more specific advertising targeting. The age of the internet dramatically increased this potential as more and more information about you was created and stored. Today, primarily due to social networking, the amount of data on the web about you is staggering. In fact in 2013, when looking at the amount of data in the world, experts estimated that about 90% of the data has in fact been generated in just the last two years (SINTEF 2013). Clearly we operate in the era of Big Data and companies have begun to harness all of this data into their own data warehouse.

The advantages of this are immeasurable. For example, if I need to purchase something online, I typically will go to Amazon.com and search for the product. Not only does Amazon have my buying history, so they can tailor ads directly towards me, but I can see a list of products that are frequently purchased in conjunction with the product that I am looking to buy. Combine this with product rankings and online reviews and I feel very comfortable in making a purchase. This information also allows businesses to tailor their advertising on specific sights. For example, I recently was in the market for a shed and did a Google search on the company. Within a matter of days, I started seen advertising for various shed manufacturers when I visited Facebook.

There are currently two methods that data from the web can be incorporated into a company’s data warehouse. The first method is that the customer creates a transaction. The information of what was ordered, who ordered it, where it’s shipping to and how the product is being paid for would all be information captured on the web, then formatted to be stored in a data warehouse. The second method would be through the collection of a web activity into a log file. A web log file contains clickstream data which is a record of every time a user clicks to move to a different location on a web page. So if I customer looks at a product but doesn’t purchase it, the fact that he looked at the product is still captured. Using this data, a business can start to generate an understanding of what the customer is thinking which leads to a better understanding of what products are desirable or are being influenced by advertising and promotions.

The clickstream data, however, is at a level of granularity that just can’t be used by the data warehouse. In order to be useful, the clickstream data is passed through software known as a Granularity Manager (GM). The GM removes bad data and aggregates other data into data that can be used by the data warehouse. Marakas points out that “as a rule of thumb, about 90% of raw clickstream data is discarded or summarized as it passes through the Granularity Manager (Marakas pg. 291).”

Due to the transactional nature of the web, information is not passed directly from the data warehouse back to the web site. A data warehouse response time is too slow to support online support. To handle this issue, data is passed into an operational data from the data warehouse at scheduled intervals. This data is not a copy of the data in the data warehouse but instead its data that is combined to make a profile. For example, this data might include the last activity date, activities they like or types of products that they like. Unlike the data warehouse, which contains historical data, the ODS uses interpretive data. For instance, if the customer has a proclivity to purchase bathing suits and sun glasses, the interpretive data might indicate that the customer likes to go to the beach or might be inclined to travel to Hawaii.

Today an eBusiness can gather data from its customers, both in purchases made and products that they have looked at and store this information into a data warehouse. This information can be combined with other information obtained on the web such as product and vendor information. After analyzing this data, a profile of the customer can be created and stored in the ODS. This allows for rapid tailoring of the website to the consumers likes. Creating consumer specific advertising and promotions along with recommendations based on events based on information in the consumer profile results in more frequent consumer purchases.

1. *What is Online Analytical Processing (OLAP)? What are the two specific approaches used to conduct the analysis? Provide a description of each. Compare and contrast citing advantages and disadvantages of each approach.*

Traditional relational databases have been around for decades and there main focus has been in improving the ability to store transactional data. Primarily known as Online Transaction Processing (OLTP) databases, these databases store very detailed and current data in a normalized schema. Its works extremely well in conjunction with Insert, Update and Delete transactions, Unfortunately, due to the normalization of the data, OLTP databases tend to be slow when queried are issued by the data analyst. As decision support systems and executive information systems where created so as to analyze the data, the complex queries which frequently used aggregation performed extremely poor against the underlying OLTP.

In 1993, the father of relational databases, E.F. Codd coined the term online analytical processing (OLAP) after he concluded that the traditional operational database wasn’t adequate to answer the questions that were being posed by management. He proposed the use of multidimensional databases that allowed for the manipulating of a company’s enterprise data model across many dimensions. In this proposal he set forth twelve rules that should be used in the development and use of an OLAP database. These include such things as providing multidimensional views, consistent and flexible reporting, and utilizing accessible client – server architecture while providing multiuser support.

OLAP databases typically have a low number of transactions. Historical data is typically stored but can be a various levels of granularity.

Marakas points out that the term OLAP has evolved such that it currently represents the “broad category of software technology that enables decision makers to conduct a multidimensional analysis of the consolidated enterprise data (Marakas pg. 74).”

There are two different approaches used in conducting the analysis. These two approaches are the multidimensional online analytical processing (MOLAP) and the relational online analytical processing (ROLAP). Each methods has specific advantages and disadvantages when compared.

The first approach, MOLAP, is characterized by organizing and analyzing data in an n-dimensional cube, known as a hypercube. While difficult to envision, this approach stored data in a multidimensional array, while each cell in the array represent the intersection of all dimensions at that point in the hypercube. This allows the analyst to create multidimensional views of the data each allowing for the analysis of any number of dimensions simultaneously. Marakas does point out that this approach lends itself to the introduction of a lot of empty cells. For example, he points out that while doing a sales system analysis across such things as product, region, time and sales volume that it’s possible that not all products will have been sold in all stores or regions within a specified time period. He defines this as an example of sparcity. As the number of dimensions in a hypercube increase the potential exists that the number of empty cells will also increase. Marakas points out that new techniques such as data compression can reduce the impact of scarcity.

The second approach, and one that I am more familiar with, is ROLAP. In ROLAP, instead of a multidimensional database server we utilize a relational database server that combines both detailed and summarized data. Through the use of these various data views, we can seemingly drill down into the details of the data. Since ROLAP utilizes a relational database there are a vast number of administration tool and SQL interfaces available for use. These means that developers can quickly build tools that are both portable and can scale to the increase in data. Similar to OLTP database, there is a performance hit due to the need to process index and do table joins. An approach has been developed to reduce the processing overhead through the utilization of a star schema. The star schema contains a centralized fact table which contains numerical measurements that exists at the junction of all dimensions. The dimensional table surround the fact table and are used to describe the data that is contained in the fact table. This effectively is denormalizing the data in the database thus decreasing the number of joins that will be needed when querying the data.

The approach of utilizing a star schema in ROLAP still suffers from a performance degradation when compared to the performance of a MOLAP approach. MOLAP, on the other hand suffers from scalability. MOLAP can handle summarized data extremely well but does not do a particular good job when dealing with a high number of detail records. It’s recommended that MOLAP architectures be limited to data warehouses under 50GB in size. ROLAP architecture can handle datasets in excess of the 50GB limit. If access to a combination of summarized and detailed data outweighs the performance gained from utilizing MOLAP, then one should definitely employ a ROLAP architecture.

1. *Two major computer based information processing devices that attempt to mimic the structures and operating principles found in humans and other living organisms are artificial neural networks (ANN) and genetic algorithms (GA). Provide a brief description of each and cite* ***comparative*** *advantages and disadvantages*

Artificial neural networks (ANNs) were an attempt to replicate the cognitive learning process of the human brain. The goal of an ANN is to develop models of a specific problem space through trial and error. For example, an ANN might be presented with astronomical data and asked to classify the data point as either a star or galaxy. The ANN utilizes an algorithm to produce an output and this “guess” is either compared to what the actual result should be or is clustered into sets based on characteristics. If the output matches the actual result no action is performed. If the output doesn’t match then the network analyzes itself to determine which parameters might need to be adjusted to produce the correct result. As more and more data is input into the ANN and then compared the resultant model converges to become fairly accurate.

The ANN mimics the human brain by utilizing an interconnected system of nodes which are referred to as neurodes. Each neurode is associated with various weighted connection. The ANN is composed of numerous layers with the connections going between each layer. There are three basic levels of an ANN. The input layer is the first layer and is the layer that receives the input data. The internal/hidden layer is the second level. It is in this layer that all processing of the input data is done. This layer passes the final ‘guess” to the third layer known as the output layer. This layer just relays the final result.

As mentioned, the ANN converges on an accurate model. To do this it utilizes one of two learning paradigms. The first is unsupervised learning a paradigms. In this paradigm, no comparison is made back to a training set. Instead the ANN continues to process input data and create clusters of records until a discernable pattern emerges. Eventually, the small adjustments to the input weights result in the difference between the input and output patterns being in close approximation to each other. The second learning algorithm is supervised learning. This is the most common paradigm and is also known as back propagation. In this case, the output is compared to the correct answer. These feedbacks result in the adjustment of the weighted connections of the neurode starting from the lower levels and propagating up toward the input layer.

ANN’s have a number of benefits as well as a number of limitations. Marakas points out that a major benefit of an ANN is that it eliminates the need for direct input from domain experts. This can be beneficial in that one may not exists or may be too costly to employ injunction with the analysis. Another advantage is the ability to handle datasets that are noisy or incomplete. Most data sets are far from clean and thus contain some bad data. Marakas points out that “Each of the neurodes looks at only a small portion of the problem and the neurodes and layers are positioned to look at each of these small elements from different angles (Marakas pg. 137).” Thus ANN’s can more easily process noisy data and in some cases can reconstruct a missing data point through inference. Since ANN’s can find subtle relationships in data they are very successful when used to solve complex problems in pattern recognition.

The main limitation of an ANN is that the adjustment of each inputted weight is based on a complex mathematical process. As a result, an ANN cannot explain the reason why it made a decision as to what the final output was. Thus the user must have some faith that output produced is actually correct. In fact the underlying algorithm is so complex that there exists a great disparity in the number of ANN designers compared to those who would like to utilize the ANN for analysis. Additionally, the repetitive nature of comparing the output back to a training set leads itself to a process that is extremely time consuming and can also put excessive demand on existing hardware.

Generic Algorithms (GA), similar to ANN’s, have a basis in biological theory. Instead of neuroscience, GA’s are based on the concept of natural selection, with the small component referred to as a gene. In a specific problem domain, the gene would represent the smallest unit of information used to build a potential solution. Genes can be combined to form a digital bit string which is termed a chromosome and represent a possible solution.

The goal of a GA is to find an optimal solution, based on provided solutions, to a particular problem. The GA is initialized with an initial population of chromosomes that have been a randomly selected from all possible solutions. Each chromosome is first decoded by a decoder function and then evaluated by a fitness function. The fitness function represents a constraint placed on the solution. Marakas provides an example in the Traveling Salesman problem where the typical fitness function might include minimize time or route or fuel consumption. The fitness function is critical in the ranking of solutions and a different ranking can be created from the same solution pool by just changing the function.

One the initial pool of chromosomes has been evaluated, the GA can start to combine/refine the genes within each chromosome to produce newer chromosomes. Those chromosomes that are deemed to the lowest in comparative fitness are weeded out similar in design to natural selection where the weak die and the strong survive. Through crossover and mutation newer chromosomes can be created that are “fitter” than their predecessors.

The solution obtained by a GA greatly depends on the initial chromosome population size. If the value is below 10 it increases the chances that a solution isn’t reached. As the population size grows from 10, the overall processing time increases with no distinct improvement in the provided solution.

A strong advantage of GA over ANN’s is that they tend to find a solution significantly faster and are very predictable in the processing time as it is directly correlated to the number of initial chromosomes. Marakas states that GA’s greatest advantage is that they have the “ability to stretch themselves across a vast solution space in search of the optimal solution (Marakas pg. 144).” Compared to an ANN, the overall algorithm is very straight forward and easy to understand.

On final advantage of a GA is that it can be used to solve problems we don’t know how to solve. Instead of providing possible solutions we can provide the components that would make up a reasonable solution, create a usable fitness function and let the GA derive a solution. Significant domain knowledge isn’t required. The GA just simply needs to recognize if a solution created would be a good one.

1. *Executive information systems (EIS) have been promoted as the solution to every manager’s information processing needs. While true that a properly designed EIS, utilizing a well designed data warehouse, will offer a significant competitive benefits to its users, it is also true that there are a number of limitations and pitfalls that we need to be aware of. Identify and describe some of the limitations of EIS, and offer suggestions on avoiding or overcoming such limitations.*

An Executive Information System (EIS) can bring significant benefits to an organization by producing reports that satisfy executive information needs. This needs may include allocation of resources, determining which projects to implement and getting the information needed to quickly handle the unexpected. While the advantages of an EIS are numerous, there are a number of issues that need to be considered before investing in an EIS.

Of primary concern is the overall cost to implement an EIS. Marakas points out that “a survey of organizations using EISs found the mean development cost to be $365, 000 (Marakas pg. 172).” Marakas then states that the mean operational costs also provided in the survey was in excess of $200,000. So an EIS is expensive with operational costs far exceeding the initial development costs in just a few years. These ongoing operational costs can be a huge influence in the eventual failure of an EIS. Since operational costs consist of things such as training, licensing costs and maintenance a business can make decisions that affect these costs. For example, if a company has practices in place that lower overall turnover, you can eliminate some of the costs associated with training new employees. If a company develops and implements a maintenance schedule these can help to increase the life expectancy of the system. Licensing fees need to be understood up front. Potential does exist to reduce licensing fees and maintenance in the future through the consolidation of databases and cores.

Another concern when implementing an EIS is in regards to the various types of systems and data that currently exist within the organization. The strength of an EIS is the consolidation of dispersed data into the EIS. Due to the vast number of data sources and the media that the data resides in, an EIS developer might need to learn new data formats and programming languages. The technical complexity of working with all this data and integrating it into the EIS and data warehouse may beyond the capability of existing developers or may require the support of numerous individuals. Complexity typically increase with the introduction of legacy systems. It’s imperative that a list of the needed technical experience is created before fully committing to an EIS. Poor understanding may lead to an EIS that performs too slowly or doesn’t have all the desired functionality which ultimately may lead to a decision to continue support for an EIS. If the development time becomes excessive due to inadequate knowledge of a specific technology, a decision to abandon the EIS project becomes a greater concern. This technological assessment ahead of times will also help to reduce costs that might be associated with having to redesign at a later time a poorly performing EIS.

The last pitfall that Marakas discusses is one of Organizational limitations. He subdivides this into three main categories, those being Agendas and Time Biases, Managerial Synchronization and Destabilization. These issues typically are more difficult to address than those limitations already discussed.

Marakas describes Agendas and Biases as issues resulting from the actual use of an EIS. He points out that an EIS should be utilized with other information when making critical decisions of an organization. Over use of an EIS provides a manager information concerning the measurable of the company without the benefit of context. Thus a manager may devote time to an issue identified in measurement but one that might actually be relevant to the overall company strategy. This “information channeling” may make a manager biased to a situation that only appears to be true. Additionally, since the EIS allows for more frequent access to information, managers can sometimes lose the overall enterprise focus and instead start focusing on lower level decisions usually made by lower and midlevel managers. The micromanagement approach is not an effective one and results in the disruption of the daily tasks done by these lower level managers.

Marakas defines Managerial Synchronization as an extension of the above concern. Managers that consistently create Adhoc reporting end up disrupting the dissemination of established reports that are produced on a predetermined schedule. The established reports help managers across the organization conduct a review of key indicators across all levels of management. These reports are designed to ensure that the enterprise will reach its established strategic goals. Over use of Adhoc reporting may cause a manager to lose focus on the overall big picture in favor of one that focuses on short term goals or individual accolades.

The last category is destabilization. It’s important to fully understand the data being provided by the EIS and to take the time to fully understand the consequences of taking action based on the data provided. Quick or impulsive changes to an organization tend to implement destabilization into the company. This can lead to political problems and resistance within management and by employees typically resulting in decreased morale and productivity. It’s better to make small changes in response to data as it’s both easier and faster to implement a small number of changes as well as easier and faster to return to previous functionality if the decision made turns out to be wrong.

1. *Describe the two basic approaches to designing an enterprise model. What are their relative advantages and disadvantages?*

Marakas makes the argument that for a data warehouse project to be successful it needs to have the ability to provide information to as the largest possible audience. To do this, we need to fully understand the information requirements of the enterprise across all levels. This accomplished through the creation of an enterprise model.

The enterprise model is in fact made up of a number of separate models that are integrated to provide an enterprise level overview in a company’s strategies, processes and data. Through the enterprise model we gain an understanding of how the various systems within the enterprise align with the business. Data warehouse designers don’t need to understand every separate model in the enterprise model and instead focus on the corporate process model and process data model. The first describes the processes within an organization and the second the entities that are either required by these processes or generated by these processes.

The construction of an enterprise model requires a number of activities to be done. For example an analysis of the enterprises strengths, weaknesses, threats and opportunities (SWOT analysis) should be undertaken. Additionally, an enterprise should define the entities that exist within an organization and describe the relationship that exists between these entities. Another important task is information needs analysis.

There are two basic approaches to designing a data warehouse – the top down and bottom up approaches.

If the top down approach is taken, then the data warehouse design is based on the constructed enterprise model. As Marakas points out, this “implies a strategic rather than an operational perspective of the data (Marakas pg. 185).” Unfortunately, most organizations don’t take this approach. Of main concern is that there is just too much organizational change at any given time for the model to be truly accurate at a specific point in time. Another concern is that the holistic model might identify organizational deficiencies that could be damaging to senior executives. A final concern, one that I know all too well, is that everyone has established deadlines and thus IS may feel that they don’t have the resources to construct the enterprise model in the necessary time needed to construct the data warehouse.

The bottom up approach focuses on utilizing data in the current existing systems to design the data warehouse. This approach while certainly quicker to produce will result in the data warehouse being restricted to the organizations that it can provide information too. A decision to follow a bottom up approach will certainly result in a decrease in functionality of the data warehouse when compared to the data warehouse created with a top down approach.

Whenever possible it is recommended to utilize a top down approach, as even though it takes longer to build, it will ultimately result in improved functionality. Additionally, building the enterprise model before the data warehouse creates an agreed upon alignment between all information systems and ensures alignments of these systems with the strategic enterprise goals. This ultimately, decreases the organizational change and improves IS productivity.

1. *Data mining, data warehousing, data visualization, knowledge management: How do you see the future of data warehousing? What are some of the challenges facing the designer of a data warehouse, and by extension the knowledge systems built around data warehouses (think technology, www, legal framework, privacy, etc)?*

The first challenge that comes to mind is that data is increasing at an exponential rate and as such we need to develop software that can extract the valuable data from the noise as well as the storage capacity to hold the data. Marakas, in his 2003 textbook, discusses how Large Hadron Collider at CERN generates 20 TB of test data each day. Today we can discuss the Square Kilometer Array which begins construction in 2018 and once completed will have an expected output of 1 petabyte every night (Budavari 2013). A data warehouse typically separates historical data from current data as combining everything together leads to lower performance – after all there is more data to analyze. Users have expectations that queries will return data in a timely fashion. Can our current query algorithms quickly look through petabytes of information?

Additionally, more and more data is in the form of unstructured data coming from social media sites. Will the fact that the amount of unstructured data compared with structured date result in more two tiered data warehouses or will the fundamental logical concept of a data warehouse change so as to primarily support unstructured data first and foremost. I see a trend toward more NoSQL data warehouses or instead of an ODS the development of operational data warehouse that can work with the web in real time. Additional complications arise when you factor in all the additional data available in the cloud. Instead of the singular platform enterprise data warehouse, I believe a move to a multiplatform data warehouse seems like a good idea.

One area that eventually will need to be addressed legally is whom owes that data. Some will argue that it’s the entity that created it while others will argue it’s those who stored it. This will decision will ultimately be made by the courts as more and more data that individuals think should be private are shared. I’ve mentioned on the discussion boards how improvements in voice recognizing is a huge gain for those who use devices to ask questions. While currently regulated in a number of states, companies would love to take all that unstructured data to determine not just what you like but through the listening of ambient sounds what your current environment is and who you are with.

One last thought regarding the future of data warehouses, As Marakas points out, creating an EIS and a data warehouse isn’t cheap and requires a significant investment both in money and resources. Additionally, the development time to build a data warehouse needs to be considered. The introduction of the cloud has led to a number of companies to introduce business solutions hosted on the cloud. First we had Software as a Service, followed by Infrastructure as a Service and Database as a Service. I can envision a time when Data Warehousing as a Service is offered allowing for more small and mid- size companies to have the capability of analytics through a cloud based data warehouse.

**References**

Budavari, Tamas, Laszlo Dobos, and Alexander S. Szalay. "SkyQuery: Federating Astronomy Archives." *Computing in Science & Engineering* 15, no. 2 (2013): 12-20.

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.

SINTEF. (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Retrieved May 29, 2015 from www.sciencedaily.com/releases/2013/05/130522085217.htm