CS614- Data Warehousing
SPRING 2012

1) How aggregates awareness helps the users 2 marks

Answer:- (Page 87)

Aggregate awareness allows using pre-built summary tables by some front-end tools. It is smart enough to develop or compute higher level aggregates using lower level or more detailed aggregates.

2) Diff b\w one to one and one to many transformation 2 marks

Answer:- (Page 144)

Simple scalar transformation is a one-to-one mapping from one set of values to another set of values using straightforward rules.

A one-to-many transformation is more complex than scalar transformation. As a data element form the source system results in several columns in the DW.

3) How cube is created in ROLAP 3 marks

Answer:- (Page 88)

Cube is a logical entity containing values of a certain fact at a certain aggregation level at an intersection of a combination of dimensions.

4) What is timestamps 3 marks

Answer:- (Page 150)

The tables in some operational systems have timestamp columns. The timestamp specifies the time and date that a given row was last modified. If the tables in an operational system have columns containing timestamps, then the latest data can easily be identified using the timestamp columns.

دنیا میں سب سے مشکل کام اپنی اصلاح اور سب سے آسان کام دوسروں پر نکتہ چینی کرنا ہے

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

5) factors behind poor data Quality 5 marks
Answer:- (Page 139) Data warehousing fundamentals
Data Pollution Sources
_ System conversions and migrations
_ Heterogeneous systems integration
_ Inadequate database design of source systems
_ Data aging
_ Incomplete information from customers
_ Input errors
_ Internationalization/localization of systems
_ Lack of data management policies/procedures
Types of Data Quality Problems
_ Dummy values in source system fields
_ Absence of data in source system fields
_ Multipurpose fields
_ Cryptic data
_ Contradicting data
_ Improper use of name and address lines
_ Violation of business rules
_ Reused primary keys
_ Nonunique identifiers

6) Which denormalization technique used to reduce foreign keys in a relation 5 marks[/i]
Answer:- (Page 53)
Collapsing Tables technique used to reduce foreign keys in a relation. In general, collapsing tables in One-to-One relationship has fewer drawbacks than others. There are several advantages of this technique, some of the obvious ones being reduced storage space, reduced amount of time for data update, some of the other not so apparent advantages are reduced number of foreign keys on tables, reduced number of indexes (since most indexes are created based on primary/foreign keys).

MIDTERM SPRING 2012
CS614- Data Warehousing

Q#1. Difference between Low granular and high granular (2)
Answer:- (Page 47) Data warehousing fundamentals
Data granularity in a data warehouse refers to the level of detail. The lower the level of detail, the finer the data granularity. Of course, if you want to keep data in the lowest level of detail, you have to store a lot of data in the data warehouse. You will have to decide on the granularity levels based on the data types and the expected system performance for queries.

بہترین تجربہ وہ ہے جس سے نصیحت حاصل ہو

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies
(IEMS) Samundari

Q#2. Difference between one to one and one to many transformation (2)
Answer:- Rep

Q#3. Describe reason to summarization during data transformation (3)
Answer:- (Page 136)
Sometimes you may find that it is not feasible to keep data at the lowest level of detail in your data warehouse. It may be that none of your users ever need data at the lowest granularity for analysis or querying. For example, for a grocery chain, sales data at the lowest level of detail for every transaction at the checkout may not be needed. Storing sales by product by store by day in the data warehouse may be quite adequate. So, in this case, the data transformation function includes summarization of daily sales by product and by store.

Q#4. How clustering and associative rule work (3)
Answer:- (Page 264)
Clustering is the technique of reshuffling, relocating exiting segments in given data which is mostly heterogeneous so that the new segments have more homogeneous data items. This can be very easily understood by a simple example. Suppose some items have been segmented on the basis of color in the given data. Suppose the items are fruits, then the green segment may contain all green fruits like apple, grapes etc. thus a heterogeneous mixture of items. Clustering segregates such items and brings all apples in one segment or cluster although it may contain apples of different colors red, green, yellow etc. thus a more homogeneous cluster than the previous cluster.

Q#5. Splitting of single field transformation is used to store individual components of name and addres in separate field of data warehousing, main reason to doing this (5)
Answer:- (Page 154)
You need to store individual components of names and addresses in separate fields in your data warehouse for two reasons. First, you may improve the operating performance by indexing on individual components. Second, your users may need to perform analysis by using individual components such as road, sector, city etc.

Q#6 explain additive and non additive and examples (5).
Answer:- (Page 119)
There can be two types of facts i.e. additive and non-additive. Additive facts are those facts which give the correct result by an addition operation. Examples of such facts could be number of items sold, sales amount etc. Non-additive facts can also be added, but the addition gives incorrect results. Some examples of non-additive facts are average, discount, ratios etc. Consider three instances of 5, with the sum being 15 and average being 5. Now consider two numbers i.e. 5 and 10, the sum being 15, but the average being 7.5. Now, if the average of 5 and 7.5 is taken this comes to be 6.25, but if the average of the actual numbers is taken, the sum comes to be 30 and the average being 6. Hence averages, if added gives wrong results. Now facts could be averages, such as average sales per week etc, thus they are perfectly legitimate facts.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

MIDTERM SPRING 2012
CS614- Data Warehousing

1. How to gain expressionless?

Answer:- (Page 67)

Expression partitioning is usually deployed when expressions can be used to group data together in such a way that access can be targeted to a small set of partitions for a significant portion of the DW workload.

2. Difference between one to one and one to many transformation (2)

Answer:- Rep

3. In the given scenario What are the additive and non-additive and reason?

(1)

S.no	No. of Item Sold
1	20
2	50
3	40
Total	**110**

(2)

S.no	No. of Item %Discount
1	41
2	52
3	46
Total	**249**

Answer:- (Page)

(1)This is additive because Additive facts are easy to work with Summing the fact value gives meaningful results
Additive facts are:
Quantity sold, Total Rs. sales

(2) This is non-additive because these are easy to work with
Averages (average sales price, unit price)
Percentages (% discount)
Ratios (gross margin)
Count of distinct products sold

جھوٹ انسان اور ایمان دونوں کا دشمن ہے

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

4. What is HOLAP and its features?

Answer:- (Page 96)

The hybrid OLAP (HOLAP) solution is a mix of MOLAP and relational ROLAP architectures that supports queries against summary and transaction data in an integrated fashion. HOLAP environments use MOLAP cubes to support common access paths with reasonably small dimensional cardinality and number of dimensions and relational structures when greater scalability for OLAP queries is required. This coexistence strategy allows exploiting the best of both worlds. Microsoft OLAP Services supports a HOLAP environment, as do tools such as HOLOS. The HOLAP approach enables a user to perform multidimensional analysis on data in the MDDB along with query based probing. However, if the user reaches the bottom of the multidimensional hierarchy and requires further detail data, the smart HOLAP engine automatically generates SQL to retrieve the detail data from the source RDBMS and returns it to the end user. This is done transparently to the user. Several MOLAP vendors, such as Arbor and Oracle, have transitioned to HOLAP architectures that include a ROLAP component. However, these HOLAP architectures are typically more complex to implement and administer than ROLAP or MOLAP architectures individually.

5. You are manager of DWH, how you keep the data Validate in term of secure data of Validation:

6. purposes data profiling

Answer:- (Page 440)

Data profiling is a process of gathering information about columns, It must fulfill the following purposes

• Identify the type and extent to which the transformation is required

• The number of columns which are required to be transformed and which transformation is required, meaning date format or gender convention.

• It should provide us a detailed view about the quality of data. The number of erroneous values and the number of values out of domain.

MIDTERM Fall 2011
CS614- Data Warehousing

Define full extraction?

Answer:- (Page 133)

- The data extracted completely from the source system.
- No need to keep track of changes.
- Source data made available as-is w/o any additional information.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

In round robin distribution is pre_defined.Are you agree are not? Expalin your answer with argument?
Answer:- (Page 66)
In round robin distribution is Not pre-defined. Round-robin spreads data evenly across the partitions, but does not facilitate partition elimination. Round-robin is typically used only for temporary tables where partition elimination is not important and co-location of the table with other tables is not expected to yield performance benefits.

What is merge and purge problem?
Answer:- (Page 168)
Within the data warehousing field, data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in the different data sets or are represented erroneously. Thus, duplicate records will appear in the merged database. The issue is to identify and eliminate these duplicates. The problem is known as the merge/purge problem.

Define quality with real life example? Difference between intrinsic and realistic quality?
Answer:- (Page 180)

- Quality means meeting customer's needs, not necessarily exceeding them.
- Quality means improving things customers *care about*, because that makes their lives easier and more comfortable.

The luxury automobile producer Rolls Royce went bankrupt in the early 1980s. Analysis revealed that, among other things, Rolls Royce was improving components that the luxury automobile customers felt were irrelevant and polishing parts they did not care about. This drove the price beyond what the luxury automobile customer felt was value for their money.
On the other hand, when Lexus decided to make its first major redesign of its highly rated L8 400 luxury automobile, company representatives asked for help from their existing customers. They even visited the homes of a variety of L8 400 owners to observe home furnishings, what kind of leather they had on their brief cases, and other minute details to get an idea of their customer's subconscious expectations.

Intrinsic Data Quality
Intrinsic data quality, simply stated, is data accuracy. Intrinsic data quality is the degree to which data accurately reflects the real-world object that the data represents. If all facts that an organization needs to know about an entity are accurate, that data has intrinsic quality-it is an electronic reproduction of reality. Intrinsic data quality means that data is correct.

Realistic Data Quality
Realistic data quality is the degree of utility and value the data has to support the enterprise processes that enable accomplishing enterprise objectives. Fundamentally, realistic data quality is the degree of customer satisfaction generated by the knowledge workers who use it to do their jobs. Realistic data quality is the degree to which data enables knowledge workers to meet enterprise objectives efficiently and effectively.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Define erreouns data with example?
Answer:- (Page 159)
It related with Non primary key problems. Erroneous data leads to unnecessary costs and probably bad reputation when used to support business processes. Consider a company using a list of consumer addresses and buying habits and preferences to advertise a new product by direct mailing. Invalid addresses cause the letters to be returned as undeliverable. People being duplicated in the mailing list account for multiple letters sent to the same person, leading to unnecessary expenses and frustration. Inaccurate information about consumer buying habits and preferences contaminate and falsify the target group, resulting in advertisement of products that do not correspond to consumer's needs. Companies trading such data face the possibility of an additional loss of reputation in case of erroneous data.

MIDTERM Fall 2011
CS614- Data Warehousing

Why aggregation is one way? Give example
Answer: - (Page 113)
Aggregation is one-way i.e. you can create aggregates, but can not dissolve aggregates to get the original data from which the aggregates were created. For example 3+2+1 = 6 at the same time 2+4 also equals 6, so does 5+1 and if we consider reals, then infinetly many ways of adding numbers to get the same result.

Define one-to-many transformation?
Answer: - (Page 144)
A one-to-many transformation is more complex than scalar transformation. As a data element form the source system results in several columns in the DW. Consider the 6×30 address field (6 lines of 30 characters each), the requirement is to parse it into street address lines 1 and 2, city, sate and zip code by applying a parsing algorithm.

Difference b/w knowledge driven DSS and Model driven DSS?
Answer: - Click here for detail
Knowledge-Driven DSS based on Expert System technologies attempts to reason about an input using its knowledge base and logical rules for problem solving.
Model-Driven DSS has a sequence of predefined instructions (a model) for responding to a change in inputs

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

What is distance function? What are alternative distance functions for typorgaphical mistake?
Answer: - (Page 176)
A number of alternative distance functions for typographical mistakes are possible, including distances based upon (i) edit distance (ii) phonetic distance and (iii) typewriter distance.

Good features of DOLAP that distinguish it from others?
Answer: - (Page 97)
DOLAP typically is the simplified version of MOLAP or ROLAP. DOLAP is inexpensive, it is fast and easy to setup on small data sets comprising of thousands of rows instead of millions of rows. It provides specific cube for the analysis. The DOLAP systems developed are extensions of production system report writers, while the systems developed in the early days of client /server computing aimed to take advantage of the power of the emerging PC desktop machine. DOLAP also provides the mobile operations of OLAP for the people who travel and move extensively, such as sales people. The one obvious disadvantage of DOLAP is that it lacks the ability to manage large data sets. But this is just another technique to suit the business requirement.

Adverse features of MOLAP?
Answer: - Click here for detail
It is limited in the amount of data it can handle. Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.

It requires an additional investment. Cube technology are often proprietary and do not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.

MIDTERM Spring 2011
CS614- Data Warehousing

Difference b/w full and incremental extraction. (5)
Answer: - (Page 133)
Full Extraction

- The data extracted completely from the source system.
- No need to keep track of changes.
- Source data made available as-is w/o any additional information.

خوبصورتی علم وادب سے ہوتی ہے لباس و حسن سے نہیں

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Incremental Extraction

- Data extracted after a well defined point/event in time.
- Mechanism used to reflect/record the temporal changes in data (column or table).
- Sometimes entire tables off-loaded from source system into the DWH.
- Can have significant performance impacts on the data warehouse server.

describe orr's law of data quality (5)

Answer: - (Page 182)

Law #1: "Data that is not used cannot be correct!"
Law #2: "Data quality is a function of its use, not its collection!"
Law #3: "Data will be no better than its most stringent use!"
Law #4: "Data quality problems increase with the age of the system!"
Law #5: "The less likely something is to occur, the more traumatic it will be when it happens!"

horizontal splitting technique: round robin and hash (3)

Answer: - (Page 66,211)

Round-robin spreads data evenly across the partitions, but does not facilitate partition elimination (for the same reasons that hashing does not facilitate partition elimination). Round-robin is typically used only for temporary tables where partition elimination is not important and co-location of the table with other tables is not expected to yield performance benefits (hashing allows for co-location, but round-robin does not).
Hash partitioning is a good and easy-to-use alternative to range partitioning when data is not historical and there is no obvious column or column list where logical range partition pruning can be advantageous.

define one to many transformation (2)

Answer: - Rep

fact table (2)

Answer: - (Page 104)

DM is a logical design technique that seeks to present the data in a standard, instinctive structure that supports high-performance and ease of understanding. It is inherently dimensional in nature, and it does adhere to the relational model, but with some important restrictions. Such as, every dimensional model is composed of one "central" table with a multipart key, **called the fact table**, and a set of smaller tables called dimension tables.

The fact table is a way of *visualizing as* an "un-rolled" cube.

Active DWH?

Answer: - Click here for detail

An active data warehouse (ADW) is a data warehouse implementation that supports near-time or near-real-time decision making.

زندگی میں کامیابی کا یہی راز ہے کہ پریشانیوں سے پریشان مت بنو

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Lexical error?
Answer: - (Page 160)
It is Syntactically Dirty Data. For example, assume the data to be stored in table form with each row representing a tuple and each column an attribute. If we expect the table to have five columns because each tuple has five attributes but some or all of the rows contain only four columns then the actual structure of the data does not conform to the specified format.

MIDTERM Spring 2011
CS614- Data Warehousing

• Define incremental Extraction.[2]
Answer: - Rep

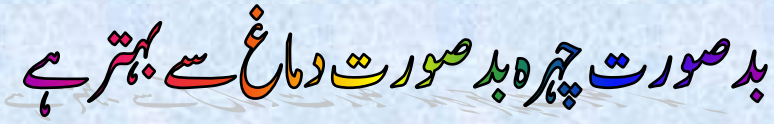
• Define simple one-to-many transformation.[2]
Answer: - Rep

• In Round Robin the distribution.[3]
Answer: - Rep

• Define additive and non-additive of facts.[3]
Answer: - Rep

• What is ELT? Where and why it is used? [3]
Answer: - (Page 147)
ELT: Extract, Load, Transform in which data transformation takes place on the data warehouse server. This is a different kind of approach called ELT which is Extract Load Transform. We extract, we load into the database and then we transform in the parallel database. Then we get all the parallelism for free, because you already have a parallel database. You don't have to buy a separate tool in order to get the parallelization.

• Define Physical Extraction and difference between offline and online extraction.[5]
Answer: - (Page 134)
Depending on the chosen logical extraction method and the capabilities and restrictions on the source side, the extracted data can be physically extracted by two mechanisms. The data can either be extracted online from the source system or from an offline structure.



Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Online Extraction
The data is extracted directly from the source system itself. The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner (for example, snapshot logs or change tables). Note that the intermediate system is not necessarily physically different from the source system. With online extractions, you need to consider whether the distributed transactions are using original source objects or prepared source objects.

Offline Extraction
The data is not extracted directly from the source system but is staged explicitly outside the original source system. The data already has an existing structure (for example, redo logs, archive logs or transportable table-spaces) or was created by an extraction routine. You should consider the following structures:

• Flat files
• Data in a defined, generic format. Dump files
• DBMS-specific format. Redo and archive logs
• Transportable table-spaces

MIDTERM Spring 2011
CS614- Data Warehousing

Q1:- EXPLAIN INCREMENTED EXTRUCTION? MARKS (2)
Answer: - Rep

Q2:- CLANSING CAN BE BREAK DOWN IN HOW MANY STEPS? MARKS (2)
Answer: - (Page 168)
Break down the cleansing into six steps: elementizing, standardizing, verifying, matching, house holding, and documenting.

Q3:- HOW THE FOLLOWING DIMENSION CAN MAKE OF MAX OPERATOR IN THE MIN-MAX OPERATION ? MARKS (3)

Q4:- LOOKING AT THE BENEFITS OF THE CUBE PARTITIONING ANALYZE WHEN IT IS REQUIRED TO PARTITON CUBE ? MARKS (3)
Answer: - (Page 85)

- To overcome the space limitation of MOLAP, the cube is partitioned.
- One logical cube of data can be spread across multiple physical cubes on separate (or same) servers.
- The divide&conquer cube partitioning approach helps alleviate the scalability limitations of MOLAP implementation.
- Ideal cube partitioning is completely invisible to end users.
- Performance degradation does occur in case of a join across partitioned cubes.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Q5:- WRITE BENEFITS OF DOLAP AND HOLAP AND MOLAP AND ROLAP ? MARKS (5)
Answer: - (Page 78)
HOLAP provides a combination of relational database access and "cube" data structures within a single framework. The goal is to get the best of both MOLAP and ROLAP: scalability (via relational structures) and high performance (via pre-built cubes).

MOLAP physically builds "cubes" for direct access - usually in the proprietary file format of a multi-dimensional database (MDD) or a user defined data structure. Therefore ANSI SQL is not supported.

ROLAP or a Relational OLAP provides access to information via a relational database using ANSI standard SQL.

DOLAP allows download of "cube" structures to a desktop platform without the need for shared relational or cube server. This model facilitates a mobile computing paradigm. DOLAP is particularly useful for sales force automation types of applications by supporting extensive slide and dice.

Q6:- IN DATA WAREHOUSE, WE CASE DE-NORMALIZATION WHICH MEANS THERE IS REDUNDANT DATA. BUT IF THERE IS DUPLICATION IN SOURCE SYSTEM WE TRY TO REMOVE THIS DUPLICATION. HOW THE DATA DUPLICATION IN SOURCE SYSTEM HAVE AN EFFECT ON ANALYSIS PROCESS IN DATA WAREHOUSE SYSTEM MARKS (5)
Answer: - (Page 166)
Data duplication can result in costly errors, such as:

- False frequency distributions.
- Incorrect aggregates due to double counting.

Difficulty with catching fabricated identities by credit card companies.
Without accurate identification of duplicated information frequency distributions and various other aggregates will produce false or misleading statistics leading to perhaps untrustworthy new knowledge and bad decisions. Thus this has become an increasingly important and complex problem for many organizations that are in the process of establishing a data warehouse or updating the one already in existence.

Credit card companies routinely assess the financial risk of potential new customers who may purposely hide their true identities and thus their history or manufacture new ones.

The sources of corruption of data are many. To name a few, errors due to data entry mistakes, faulty sensor readings or more malicious activities provide scores of erroneous datasets that propagate errors in each successive generation of data.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

MIDTERM Spring 2011
CS614- Data Warehousing

1. We use de-normalization which means redundant data. IF Duplication is in the source system what are the effects of this duplication in Analysis of Data ware house? 5 marks
Answer: - Rep

2. Briefly describe the features of Dimensional Modeling? 5 marks
Answer: - (Page 104)
DM is a logical design technique that seeks to present the data in a standard, instinctive structure that supports high-performance and ease of understanding. It is inherently dimensional in nature, and it does adhere to the relational model, but with some important restrictions. Such as, every dimensional model is composed of one "central" table with a multipart key, called the fact table, and a set of smaller tables called dimension tables. Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table. This results in a characteristic "starlike" structure or star schema.

The foundation for design in this environment is through use of dimensional modeling techniques which focus on the concepts of "facts" and "dimensions" for organizing data.

Facts are the quantities or numerical measures (e.g., sales $) that we can count and the most useful being those that are additive. The most useful facts in a fact table are numeric and additive. Additive nature of facts is important, because data warehouse applications almost never retrieve a single record form the fact table; instead, they fetch back hundreds, thousands, or even millions of these records at a time, and the only useful thing to do with so many records is to add them up. Example, what is the average salary of customers who's age > 35 and experience more than 5 years?

3. We use CDC in both legacy systems and modern systems. What are the benefits of using CDC in modern systems as compared to legacy systems? 3 Marks
Answer:- Answer: - (Page 151)
1) Immediate.
2) No loss of history
3) Flat files NOT required
4) No incremental on-line I/O required for log tape
5) The log tape captures all update processing
6) Log tape processing can be taken off-line.
7) No haste to make waste.

4. From distributive, algebraic and holistic transformation which is used in statistical analyzer? Give answer with reason. 3 marks

عقل مند کہتا ہے میں کچھ نہیں جانتا جبکہ بے وقوف کہتا ہے کہ میں سب کچھ جانتا ہوں

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

5. Round robin distribution is pre- defined? Do you agreed or not. Give your answer with argument? 2 marks
Answer: - Rep

6. Differentiate between MOLAP and DOLAP in terms of implementation? 2 marks
Answer: - (Page 78)
MOLAP: OLAP implemented with a multi-dimensional data structure.
DOLAP: OLAP implemented for desktop decision support environments.

7. Differentiate b\w offline extractions and online extraction? 2 marks
Answer: - Rep

MIDTERM Spring 2011
CS614- Data Warehousing

Diff between ER and DM? 5marks
Answer: - (Page 78)

ER vs. DM

ER	DM
Constituted to optimize OLTP performance.	Constituted to optimize DSS query performance.
Models the micro relationships among data elements.	Models the macro relationships among data elements with an overall deterministic strategy.
A wild variability of the structure of ER models.	All dimensions serve as equal entry points to the fact table.
Very vulnerable to changes in the user's querying habits, because such schemas are asymmetrical.	Changes in user querying habits can be catered by automatic SQL generators.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Explain Orr's Laws of Data Quality ?5marks
Answer: - Rep

MIDTERM EXAMINATION
Spring 2010
CS614- Data Warehousing (Session - 6)

Question No: 21 (Marks: 2)
Briefly describe snowflake schema.
Answer: - Click here for detail
Snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions.

Question No: 22 (Marks: 2)
Why both aggregation and summarization are required?
Answer: - (Page 155)
- Grain mismatch (don't require, don't have space)
- Data Marts requiring low detail
- Detail losing its utility

Question No: 23 (Marks: 3)
Under what condition smart tools work properly to construct a less detailed aggregate from more detailed aggregate?
Answer: - (Page 91)
Smart tools will allow less detailed aggregates to be constructed from more detailed aggregates (full aggregate awareness) at run-time so that we do not go all the way down to the detail for every aggregation. However, for this to work, the metrics must be additive (e.g., no ratios, averages, etc.). More detailed pre-aggregates are larger, but can also be used to build less detailed aggregates on-the-go.

Question No: 24 (Marks: 3)
What is web scraping? Give some of its uses.
Answer: - (Page 146)
Web scraping means Lot of data in a web page, but is mixed with a lot of "junk".
Problems:
- Limited query interfaces
- Fill in forms
- "Free text" fields
- E.g. addresses
- Inconsistent output
- i.e., html tags which mark interesting fields might be different on different pages.
- Rapid change without notice.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

Question No: 25 (Marks: 5)
After completing the transformation task, data loading activity is started. How many types of data loading strategies are and when each type of strategy is adopted? Explain.

Answer: - (Page 139)
Once we have transformed data, there are three primary loading strategies:

- Full data refresh with BLOCK INSERT or 'block slamming' into empty table.
- Incremental data refresh with BLOCK INSERT or 'block slamming' into existing (populated) tables.
- Trickle/continuous feed with constant data collection and loading using row level insert and update operations.

Question No: 26 (Marks: 5)
What are the drawbacks of MOLAP? Also explain the curse of Dimensionality?
Answer: - (Page 84)
Maintenance issue: Every data item received must be aggregated into *every* cube (assuming "to-date" summaries are maintained). Lot of work.
Storage issue: As dimensions get less detailed (e.g., year vs. day) cubes get much smaller, but storage consequences for building hundreds of cubes can be significant. Lot of space.
Scalability:
Often have difficulty scaling when the size of dimensions becomes large. The breakpoint is typically around 64,000 cardinality of a dimension.

curse of Dimensionality i.e. Scalability
MOLAP implementations with pre-defined cubes as pre-aggregated data perform very well when compared to relational databases, but often have difficulty scaling when the size of dimensions becomes large. The breakpoint is typically around 64,000 cardinality of a dimension. Typically beyond tens (sometimes small hundreds) of thousands of entries in a single dimension will break the MOLAP model because the pre-computed cube model does not work well when the cubes are very sparse in their population of individual cells. Some implementations are also limited to a file size for the cube representation that must be less than 2GB (this is less often an issue than a few years ago). You just can not build cubes big enough, or enough of them to have every thing precomputed. So you get into the problems of scale. As already discussed, it is difficult to scale because of combinatorial explosion in the number and size of cubes when dimensions of significant cardinality are required. There are two possible, but limited solutions addressing the scalability problem i.e. Virtual cubes and partitioned cubes.

جو لوگوں کے سامنے فخر کرتا ہے وہ لوگوں کی نظروں سے گر جاتا ہے

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari

MIDTERM EXAMINATION
Fall 2010

Question No: 21 (Marks: 2)
Briefly describe features of MOLAP.
Answer: - Click here for detail
MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.

They can also perform complex calculations. All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

Question No: 22 (Marks: 2)
Name the steps involved in cleansing of data?
Answer: - Rep

Question No: 23 (Marks: 3)
Briefly describe the features of Star Schema?
Answer: - (Page 107)
Dimensional hierarchies are collapsed into a single table for each dimension. Loss of information
A single fact table created with a single header from the detail records, resulting in:

- A vastly simplified physical data model!
- Fewer tables (thousands of tables in some ERP systems).
- Fewer join resulting in high performance.
- Some requirement of additional space.

Question No: 24 (Marks: 3)
Briefly describe, what is multi-pass BSN Approach?

Multi-pass Approach

- Several independent runs of the BSN method each time with a different key and a relatively small window.
- Each independent run will produce a set of pairs of records which can be merged (takes care of transposition errors)
- Apply the transitive closure property to those pairs of records.
    - If records a and b are found to be similar and at the same time records b and c are also found to be similar the transitive closure step can mark a and c to be similar.
    - The results will be a union of all pairs discovered by all independent runs with no duplicates plus all those pairs that can be inferred by transitivity of equality.

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies
(IEMS) Samundari

Question No: 25 (Marks: 5)
What are the benefits of HOLAP & DOLAP over MOLAP & ROLAP?
Answer: - Rep

Question No: 26 (Marks: 5)
What is the relationship between Data quality and the value of a particular application? When efforts for data quality are logical?
Answer: - Lecture 21-22-23

تم اچھا کرو زمانہ تم کو برا سمجھے یہ اس سے بہتر ہے کہ تم برا کرو اور زمانہ تم کو اچھا سمجھے

Muhammad Moaaz Siddiq – MCS(4th)
Moaaz.pk@gmail.com
Campus:- Institute of E-Learning & Moderen Studies (IEMS) Samundari