all the MCQs were new expect 2 of them were from past papers.
Exact wording in not ensured:
1. Benefits of CDC in modern systems? 2
2. List 4 techniques of handling "Multi-Dimensions"? 2
3. In MOLAP, the aggregates are large, but is it possible that some aggregates have null value, give example to justify? 3
4. 5th Orr's law ? 3
Justify your answer as TRUE or FALSE:
1. STDDEV is a distributive aggregate? 2.5
2. The data that is not used is correct? 2.5

Handling Multi-valued Dimensions

1 Drop the dimension.
2 Use a primary value as a single value.
3 Add multiple values in the dimension table.
4 Use "Helper" tables.

My today's paper cs614 11:00 am 25/5/2013
20 mcqs and 14 msqs were from moaaz file
subjective questions.
1)List down any two disadvantages of Molap? (2 marks)
2)List down two step of basic sorted neighborhood? (2 marks)
3)aik statment identify kerni thi k correct hai ya in coorect from Orr's Law ? (3 marks)
4) AIk statment TQM ki identify krni thi correct or incorrect? (3marks)
5)aik fact and dimensions table create kerna tha (5marks)
6)Orr's Law ki aik aur statment 5 marks ki identfy kerni thi k correct hai ya in coorect and ans ko justify b krna tha (5 marks)

Aslamo Alikum bachoo
mera paper aisa tha 6 mcqs moazz bahi ki file me sy thy

1. **List down four basic tasks of data transformation?**
2. **Identify the given statements as correct and incorrect "The approach of TQM refers to the involvement of only 20% employee inthe continuous improvemnt process" and 2nd statement was "orr's law says that data quality is a function of its use not its collection"**
3. **Identify the given statement as correct and incorrect "in Molap the complexity cannot go beyound o(1) in any case" 2nd statement was "Drill down is a cube operation and its basic purpose is to select and project"**
4. **if dirty data in DWH is used by the government for decision making then what would be the effects?explain with exemple**
5. **identify the given statement as correct and incorrect"Transactional fact table always stores the complete records for the event that dont occur?**

---

20 mcqs and 14 msqs were from moaaz file
subjective questions.
1)List down any two disadvantages of Molap? (2 marks)
2)List down two step of basic sorted neighborhood? (2 marks)
3)aik statment identify kerni thi k correct hai ya in coorect from Orr's Law ? (3 marks)
4) AIk statment TQM ki identify krni thi correct or incorrect? (3marks)
5)aik fact and dimensions table create kerna tha (5marks)

6)Orr's Law ki aik aur statment 5 marks ki identfy kerni thi k correct hai ya in coorect and ans ko justify b krna tha (5 marks)

My today's Paper
80% MCQ from Moaaz file and previous sem MCQs...
1)What is the difference between Additive and non-additive? (2 marks)
2)Define and explain one-to-many transformation? (2 marks)
3)If the Government use dirty DWH; what problems could happen ? (3 marks)
4) Dont remember? (3marks)
5)Identify atleast 3 or more facts in the fact table of a company's sales record, company want to know per day aggregate sales in a specific store item wise. (5marks)
6)Clustering is one of the automatic data cleansing technique, list atleast 3 more automatic techniques and list down drawback of clustering if any? (5 marks)

@ Airbone Please question No 3 ,5 or 6 ka Answer bta dain

3)If the Government use dirty DWH; what problems could happen ? (3 marks)

5)Identify atleast 3 or more facts in the fact table of a company's sales record, company want to know per day aggregate sales in a specific store item wise. (5marks)
6)Clustering is one of the automatic data cleansing technique, list atleast 3 more automatic techniques and list down drawback of clustering if any? (5 marks)

---

some mcqs were from past papers some mcqs were difficult
why molap size increses 2 reasons (2 mark)
why data cleasing is important and why it need expert person with domain expertise (5 mark)

---

1)List down any two disadvantages of Molap? (2 marks)
2)List down two step of basic sorted neighborhood? (2 marks)
3)aik statment identify kerni thi k correct hai ya in coorect from Orr's Law ? (3 marks)
4) AIk statment TQM ki identify krni thi correct or incorrect? (3marks)
5)aik fact and dimensions table create kerna tha (5marks)
6)Orr's Law ki aik aur statment 5 marks ki identfy kerni thi k correct hai ya in coorect and ans ko justify b krna tha (5 marks)
paper over all boht easy tha.

Diffrence between additive and nonadditive facts? there was a table and define karna tha.. 5
Fetures of automatic Data Cleansing ? 3 marks
merge/purge in your own words? 5
objective muaaz file realy works and save the time.. some mcqs was from its file and well i had listen all the lectures and then i attempt muaaz file and other data so please listen all lectures for better understanding of this subject.. i really enjoyed the way HE give the lecture really amazing for me.. all questions are concept base so i reffer to listen mustly last 10 lectures for paper and muaaz file for your paper :) hope you enjoyed and best of luck :)

---

dear u can find the answer on page 158
Data cleansing is vitally important to the overall health of your warehouse project and ultimately affects the health of your company. Do not take this statement lightly.
The true scope of a data cleansing project is enormous. Much of production data is dirty, and you don't even want to consider what work cleaning it up would take. By "dirty," I mean that it does not conform to proper domain definitions or "make sense." The age-old adage "garbage in, garbage out" still applies, and you can do nothing about it short of analyzing and correcting the corporate data. What is noise in one domain may be information in another.
Usually the process of data cleansing can not be performed without the involvement of a domain expert because the detection and correction of anomalies requires detailed domain knowledge. Data cleansing is therefore described as semi-automatic but it should be as automatic as possible because of the large amount of data that usually is be processed and the time required for an expert to cleanse it manually

---

page 87
The biggest problem with MOLAP is the requirement of large main memory as the cube size increases. There may be many reasons for the increase in cube size, such as increase in the number of dimensions, or increase in the cardinality of the dimensions, or increase in the amount of detail data or a combination of some or all these aspects. Thus there is an issue of scalability which limits its applications to large data sets.

Exact wording in not ensured:
1. Benefits of CDC in modern systems? 2
2. List 4 techniques of handling "Multi-Dimensions"? 2
3. In MOLAP, the aggregates are large, but is it possible that some aggregates have null value, give example to justify? 3
4. 5th Orr's law ? 3
Justify your answer as TRUE or FALSE:
1. STDDEV is a distributive aggregate? 2.5
2. The data that is not used is correct? 2.5

---

my paper
MCQs are mixed some new too...
2-Marks
4 steps of transformation
how to handle missing records give 4 methods
3-Marks
5th Orrs law dia hua tha ststement main ju kay incorrect tha aap nain explain karna tha ...
table diay huay thay 2 un main values thi drinks soled ki aur % Discounts aap nain batana tha kon si values additive hain aur kon si non-additive
5-marks
Q-1
2-statements di hui thi batana tha correct hain ya incorrect
if duplication in data extracted data it will be syntactically dirty?
main draw back of clustering cleansing technique is that is it lowest computationally complex.
Q-2
same 2-statements di hui thi batana tha correct hain ya incorrect
it is easy to capture the history once the records are overwritten?

STDDIV is descriptive aggregate ?
Remember me in Your Prayers
and Best of luck... Paper is conceptual not theoretical so try to grasp the concepts.

---

1) Describe two ways to simplify ER Model. (2 marks)
2) List down problems associated with Primary Key. (2 marks)
3) Justify your answer (3 marks)
We can dissolve the aggregates to get the original data from which the aggregates were created.
4) Explain benefits of splitting of single field i.e name and address. (3 marks)
5) Justify your answer (5 marks)
a) STDDEV is an example of distributive Aggregate.
b) It is always easy to track down history after overwriting the records.
6) Justify your answer (5 marks)
a) When Attributes or records are missing in extended data then the data is semantically dirty.
b) Pattern based data cleansing technique identify outlier fields and records using the values of Mean and Standard Deviation.

CS-614 My today papper, Subjective
29-05-2013 at 11:00 AM

1. Define one-to-many Transformation with the help of example. 2 Marks
2. Write down four approches which are used to handle Multi-Values Dimensions. 2 Marks
3. Orr's law say's that data quality problem decreases when the age of system increases. 3 Marks
4. ER diagram ahve macro relations among data elements, is this right or wrong. Justify your answer. 3 Marks
5. MOLAP cube, pro-partitioning relation between cordinality and cube size, Justify your answer with logical reason. 5 Marks
6. Last question was related to data duplication. 5 Marks

---

in MOLAP there are many reasons for increase in cube size.list any two of them.....2marks
list down 2 advantages of applying "change data capture" technique in modren system.....2marks
identify given statement correct /incorrect and explain either:"sql cant be used for querying the MolaP cube".....3marks
in your opinion what may be possible reason to use summarization during data transformation.explain with example.....3marks
identify given statement correct /incorrect and explain either:"transformation is process in which we extract data from single/multiple data sources"
"offline extraction is type of logical data extraction".....5marks
last wala yaad nhe

---

my ppr
16 mcqs out of 20 were from past pprs(moaaz file).
subjective questions.
1)List down any two disadvantages of Molap? (2 marks)
2)List down two step of basic sorted neighborhood? (2 marks)
3)aik statment identify kerni thi k correct hai ya in coorect from Orr's Law ? (3 marks)
4) AIk statment TQM ki identify krni thi correct or incorrect? (3marks)
5)aik fact and dimensions table create kerna tha (5marks)

6)Orr's Law ki aik aur statment 5 marks ki identfy kerni thi k correct hai ya in coorect and ans ko justify b krna tha (5 marks)

My cs614 paper
Mcqs:

Grain is the _____ level of data stored in the warehouse.
► **Atomic**
► Summarized
► Aggregated
► Cube

During ETL process of an organization, suppose you have data which can be transformed using any of the transformation method. Which of the following strategy will be your choice for least complexity?
► **One-to-One Scalar Transformation**
► One-to-Many Element Transformation
► Many-to-Many Element Transformation
► Many-to-One Element Transformation

Change Data Capture is one of the challenging technical issues in _____
► **Data Extraction**
► Data Loading
► Data Transformation
► Data Cleansing

_____ is class of Decision Support Environment.
► OLTP
► **OLAP**
► DBMS
► Network

Horizontal splitting breaks a table into multiple tables based upon_____
► Common Row values
► Range of Data.
► Redundant data.
► **Common column values.**

The most common use of range partitioning is on _____.
► **Date**
► Rows
► DSS
► None of these

All data is _____ of something real.
IAn Abstraction
IIA Representation
Which of the following option is true?

► I Only
► II Only
► **Both I & II**
► None of I & II

Pre-computed _____ can solve performance**problems**
►**Aggregates**
►Facts
►Dimensions

Suppose the amount of data recorded in an organization is doubled every year. This increase is _____ . ►Linear
►Quadratic
►**Exponential**
►logarithmic

Experience showed that for a single pass of a magnetic tape that scanned 100% of the records, only _____of the records

- **5%**
- 50%
- 8%
- 60%

To handle dimensions that requires the aggregation of multiple data quality indicators ,the _____can be applied.

- **Min or max operation**
- Complex ratio
- Average weight

In the Information Age, the_____ learning organization is at a distinct disadvantage. The term dysfunctional means "impaired or abnormal functioning."

- Functional
- **Dysfunctional**

ETL is _____steps.

- **Independent and interrelated**
- Independent or interrelated
- Dependent and interrelated

Insurance data warehouses are similar to other data warehouses with a few exceptions: such as **the length of time** that insurance data warehouses exists, in terms of the dates found in the business,

How Much Data is that? 1GB

- **230 or 109 bytes**
- 240 or 1012 bytes
- 220 or 106 bytes

The effects of denormalization on database performance are _____

- **Unpredictable**
- **Predictable**

OLAP is Analytical Processing instead of Transaction Processing. It is also NOT a physical database design or implementation technique, but a **framework**.

The classic statement of _____is “decision making is an iterative process; which must involve the users”.

- **OLAP**
- DWH
- OLTP

ER is a _____design technique that seeks to remove the redundancy in data.

- **Logical**
- Physical

Subject questions:
Write to reason of increase in cube size in MOLAP. 2 marks
Write first two steps of Basic Sorted Neighborhood (BSN) Method. 2 marks
Justify this statement either correct or incorrect
“if defect are found in process of Attribute Domain Validation it is better to fix error in DWH and leave the data source as it is”. 3 marks
Justify the statement valid or invalid with reasons
“Dimension are quantitative and numerical measurements such as sales $".(3 marks)
Identify the given statement as correct and incorrect
1."in Molap the complexity cannot go beyond o(1) in any case"
2."Drill down is a cube operation and its basic purpose is to select and project".5 marks
Identify the given statement as correct and incorrect
1.“Lexical error is a type of coverage anomaly”
2.“Data cleansing process is describe as semi automatic but can be performed without the involvement of a domain expert”. 5 marks

---

Exact wording in not ensured:
1. Benefits of CDC in modern systems? 2
2. List 4 techniques of handling "Multi-Dimensions"? 2
3. In MOLAP, the aggregates are large, but is it possible that some aggregates have null value, give example to justify? 3
4. 5th Orr's law ? 3
Justify your answer as TRUE or FALSE:
1. STDDEV is a distributive aggregate? 2.5
2. The data that is not used is correct? 2.5

---

1. **1. How to simplify an ER data model? 2marks**

Two general methods:
§ De-Normalization
§ Dimensional Modeling (DM)

1. **2. Write basic tasks of Data Transformation? 2marks**

Basic tasks
§ Selection
§ Splitting/Joining
§ Conversion
§ Summarization
§ Enrichment

1. **3. Identify with reason which table is additive and non additive table? 3marks**

- Additive facts are easy to work with
- Summing the fact value gives meaningful results

Additive facts:

§ Quantity sold
§ Total Rs. sales
Non-additive facts:

§ Averages (average sales price, unit price)
§ Percentages (% discount)
§ Ratios (gross margin)
§ Count of distinct products sold

Month	Crates of Bottles Sold
May	14
Jun.	20
Jul.	24
TOTAL	58

Month	% discount
May	10
Jun.	8
Jul.	6
TOTAL	24% ? Incorrect!

1. **4. Write names of splitting table and why it is beneficial in different fields? 3marks**

Splitting Tables
- Horizontal Splitting
- Vertical Splitting

Horizontal Splitting ADVANTAGE
§ Enhance security of data.
§ Organizing tables differently for different queries.
§ Reduced I/O overhead.
§ Graceful degradation of database in case of table damage.
§ Fewer rows result in flatter B-trees and fast data retrieval.

Vertical Splitting

§ Splitting and distributing into separate files with repeating primary key.
§ Infrequently accessed columns become extra "baggage" thus degrading performance.
§ Very useful for rarely accessed large text columns with large headers.
§ Header size is reduced, allowing more rows per block, thus reducing I/O.
§ For an end user, the split appears as a single table through a view.

1. **5. What is significant role of cleansing? 5marks**

Data cleaning (also called data cleansing or scrubbing) is especially required when integrating heterogeneous data sources and should be addressed together with schema-related data transformations. In data warehouses, data cleansing is a major part of the so-called ETL process. Data cleansing deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleansing increases significantly. Most data cleansing is typically performed in a separate data staging area before loading the transformed data into the warehouse. A data cleansing approach should satisfy several requirements. First of all, it should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources. Furthermore, data cleansing should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata. Mapping functions for data cleansing and other data transformations should be specified in a declarative way and be reusable for other data sources as well as for query processing. Especially for data warehouses, a workflow infrastructure should be supported to execute all data transformation steps for multiple sources and large data sets in a reliable and efficient way.

1. **6. What you prefer single MOLAP or cube partition? 5marks**

Partitioned Cubes

§ 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.

MOLAP is a important application on multidimensional data warehouse. We often execute range queries on aggregate cube computed by pre-aggregate technique in MOLAP. For the cube with *d* dimensions, it can generate 2^d cuboids. But in a high-dimensional cube, it might not be practical to build all these cuboids. In this paper, we propose a multi-dimensional hierarchical fragmentation of the fact table based on multiple dimension attributes and their dimension hierarchical encoding. This method partition the high dimensional data cube into shell mini-cubes. The proposed data allocation and processing model also supports parallel I/O and parallel processing as well as load balancing for disks and processors. We have compared the methods of shell mini-cubes with the other existed ones such as partial cube and full cube by experiment. The results show that the algorithms of mini-cubes proposed in this paper are more efficient than the other existed ones.

==========

1)List down any two disadvantages of Molap? (2 marks)

Drawbacks of MOLAP:

§ Long load time (pre -calculating the cube may take days!).
§ Very sparse cube (wastage of space) for high cardinality

2)List down two step of basic sorted neighborhood? (2 marks)

Concatenate data into one sequential list of N records
§ **Steps 1:** Create Keys
§ Compute a key for each record in the list by extracting relevant fields or portions of fields
§ Effectiveness of the this method highly depends on a properly chosen key
§ **Step 2:** Sort Data
§ Sort the records in the data list using the key of step 1
§ **Step 3:** Merge
§ Move a fixed size window through the sequential list of records limiting the comparisons for matching records to those records in the window
§ If the size of the window is *w* records then every new record entering the window is compared with the previous *w-1* records.

3) **Identify the given statements as correct and incorrect "Orr's law says that data quality is a function of its use not its collection"? (3 marks)**
Correct

4) **Identify the given statements as correct and incorrect "The approach of TQM refers to the involvement of only 20% employee in the continuous improvement process"**? **(3marks)**
Correct Statement

The approach of TQM refers to the involvement of all the employee in the continuous improvement process.

==========

1) What is the difference between Additive and non-additive? (2 marks)

1. Additive facts are easy to work with
2. Summing the fact value gives meaningful results

Additive facts:

§ Quantity sold
§ Total Rs. sales
Non-additive facts:

§ Averages (average sales price, unit price)
§ Percentages (% discount)
§ Ratios (gross margin)
§ Count of distinct products sold

2) Define and explain one-to-many transformation? (2 marks)
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, state and zip code by applying a parsing algorithm.

3) If the Government uses dirty DWH; what problems could happen? (3 marks)

§ Decisions taken at government level using wrong data resulting in undesirable results.
§ In direct mail marketing sending letters to wrong addresses loss of money and bad reputation.

5) Identify at least 3 or more facts in the fact table of a company's sales record, company want to know per day aggregate sales in a specific store item wise. (5marks)

6) Clustering is one of the automatic data cleansing technique, list atleast 3 more automatic techniques and list down drawback of clustering if any? (5 marks)

Automatic cleansing techniques
1) Statistical
2) Pattern Based
3) Clustering
4) Association Rules

Drawbacks:
The main drawback of this method is computational time. The clustering algorithms have high computational complexity. For large record spaces and large number of records, the run time of the clustering algorithms is prohibitive.

==========

Why molap size increases 2 reasons (2 marks)
The biggest problem with MOLAP is the requirement of large main memory as the cube size increases. There may be many reasons for the increase in cube size, such as in crease in the number of dimensions, or increase in the cardinality of the dimensions, or increase in the amount of detail data or a combination of some or all these aspects.

Why data cleansing is important and why it need expert person with domain expertise (5 marks)
Data cleansing is vitally important to the overall health of your warehouse project and ultimately affects the health of your company. The original aim of data cleansing was to eliminate duplicates in a data collection, a problem occurring already in single database applications and gets worse when integrating data from different sources.

Usually the process of data cleansing cannot be performed without the involvement of a domain expert because the detection and correction of anomalies requires detailed domain knowledge. Data cleansing is therefore described as semi-automatic but it should be as automatic as possible because of the large amount of data that usually is be processed and the time required for an expert to cleanse it manually.

==========

Features of automatic Data Cleansing? 5 marks

Merge/purge in your own words? 5 marks

==========

2. Write down four approaches which are used to handle Multi-Values Dimensions. 2 Marks
Handling Multi-valued Dimensions
One of the following approaches is adopted:

- Drop the dimension.
- Use a primary value as a single value.
- Add multiple values in the dimension table.
- Use "Helper" tables.

3. Orr's law says that data quality problem decreases when the age of system increases Identify correct or incorrect? 3 Marks

Incorrect "Data quality problems increase with the age of the system!"

4. ER diagram have macro relations among data elements is this right or wrong. Justify your answer. 3 Marks

Correct Statement
ER modeling does not really model a business; rather, it models the micro relationships among data elements. ER modeling does not have "business rules," it has "data rules".

5. MOLAP cube, pro-partitioning relation between cardinality and cube size, Justify your answer with logical reason. 5 Marks

==========

List down 2 advantages of applying "change data capture" technique in modern system……2marks

Identify given statement correct /incorrect and explain either:"SQL can't be used for querying the MOLAP cube".....3marks
Correct Statement
In a MOLAP environment there are no tables, there are no traditional relational structures, hence ANSI SQL can not be used. As a matter of fact, there is no standard query language for querying a MOLAP cube.

In your opinion what may be possible reason to use summarization during data transformation. Explain with example.....3marks

Summarization: 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.

Identify given statement correct /incorrect and explain either:"transformation is a process in which we extract data from single/multiple data sources"

"Offline extraction is type of logical data extraction".....5marks

---

my today paper of CS614
total 26 q
all mcq'z from past paperz
1. additive and non additive facts 2 marks
2. one to many transformation with example 3 marks
3. ek statement d hoi thi about timestamp btana tha true hae yan false with reason
4.star aur snow flock schema ki digrame thi and identify karna tha k kon sa schema hae
5. what is murge/purge when we performed the cleansing in the dataware house 5 marks
6. 5 marks

Product Id	Region ID	Period	Quantity
1	N	Month	25
2	N	Month	50
2	S	Week	30

Find the dimension , primary key & dimension

A.A
#CS614
19/20 MCQs was from Dr Tariq Hanif file...
subjective was easy.
1.Automatic Data Cleansing
2. Handling missing data
3. basic tasks of Data Transformation
one q was regarding Data Extraction ...
3 q mushkil thy yad ni rhy statmnt long thi and table given tha

---

MCQ only 25% from Moaz file

Q. Consider the following facts table having name product sales 5

Product ID	Region ID	Period	quantity
01	N	Monthly	25
02	N	Monthly	50
02	G	Weekly	30

Q. identify the give statement correct or incorrect 5
1. intelligent learning organization nevers shares its information what its employees.
2. Orr's law say that that data which in not used is always correct

Q. identify the give statement correct or incorrect 3
If defects are found in the process of attributes domain validation then it is batter to fix the errors in DHW and leave the data source as it as.

Q. identify the give statement correct or incorrect 3
We can dissolve the aggregation to get the original data from which the aggregates were created.
Q. Mention one factor that lends towards long load time of MOLAO cube. 2
Q. define one to many transformation one example. 2

CS 614 date warehousing paper was easy
Mcqs were mostly from moazz file 5 to 6 mcqs were new .....last 10 lectures say prepare karna 70% paper us may sai tha
Subjectives are:
1) 2 ways to simplifiy ER?
Answer: De normalization and Dimensional modelling
2) 4 data validation techniques?
Answer: See lecture 22
3) identify the following statement as correct or incorrect. Justify your answer in either cases
"The less likely something to happen the less traumatic it will be when it happens"
Answer : incorrect statement....see orr law no.5
4) bhool gaya
5) 2 statements thin or batana tha k correct hain ya nahi
Statement 1: intelligent organizations never shares their info with their employes
Statement 2: data which is not used will always correct
Answer: both are incorrect statement
6) suppose ke sales table hai grain is "total sales by day by store "....now identify 3 facts at least.
Answer soch ki likh Dena 3 facts.....such as total products sold , total sales amount etc some thing like this
Best of luck .pray for me
Thanks

---

12 mcqs were from DrTariqHanif file.. 4 mcqs from files quiz01 and quiz02 and remaining were new..
subjective:
Q1. Write 2 ways to simplify ER. (2 marks)
Q2. Write first two steps of BSN. (2 marks)
Q3. Identify the following statement as correct or incorrect. (3 marks)
"The less likely something to happen the less traumatic it will be when it happens"

Q4. Identify the following statements as correct or incorrect. (3 marks)
"One of the basic function of OLTP is to show the historical background of an organization."
Q.5 Identify the following statements as correct or incorrect. (5 marks)
1. Transformation is the process to extract data from single/multiple sources.
2. Offline extraction is the form of Logical data extraction.
Q.6 Identify the following statements as correct or incorrect. (5 marks)
1: Intelligent organizations never share their info with their employees.
2: Orr's Law says that the data which is not used will always correct.

AOA to ALL..
my cs614 2day paper:
5 marks
Q. Identify the given statements as correct or incorrect and justify your answer in either case.
1. "The Intelligent Learning Organization never shares its information with its employees".
2. "Orr’s Law says that the data which is not used is always correct".
1st statement is correct. And 2nd statement is incorrect.
Stat2: correct is “Orr`s Law says that the data is not used is cannot correct”.
Q. Identify the given statements as correct or incorrect and justify your answer in either case.
1. "Transformation is the process in which we extract the data from single/multiple data sources".
2. "Offline Extraction is a type of Logical Data Extraction".
Both statements are incorrect.

Stat1: correct statement is “Transformation is the process in which we exact the data from multiple data sources”.
Stat2: correct statement is “Offline Extraction is a type of Physical data extraction”.
3 marks
Q. List down any three ways of “handling missing data” during “data cleansing process”.
There are any three ways of “handling missing data” during “data cleansing process”. Handling missing data:
• Dropping records
• Manually filling missing values
• Using the attribute mean or median as filter
• Using a global constants as filter
Data cleansing process as describing as semi automatic but can be performance without the involvement of a domain expert.
Q. Identify the given statement as correct or incorrect and justify your answer in either case. "Standard Query Language can not be used for querying the MOLAP cube".
Above statement is incorrect. Correct statement is; No, standard query language can not be used for querying the MOLAP cube”.
2 marks
Q. The problems associated with the extracted data can correspond to non-primary keys. List down any four problems associated with the non-primary key.
There are any four problems associated with the non-primary key:
Same primary key but different data
Same entity with different keys
Primary key in same information
Source might contain invalid data
Q. Differentiate between Additive Facts and Non-additive Facts.
Additive facts are easy to work with
Summing the fact value gives meaningful results
Additive Facts:
• Quantity sold
• Total Rs. Sales Non-additive facts:
• Averages (average sales price, unit price)
• Percentages (% discount)
• Ratio (gross margin)
• Count of distinct products sold
.....BeSt Of LuCk.....

Until now all shared paper are solved
1. Additive and non-additive facts 2 marks
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. Or Additive facts are easy to work with summing the fact value gives meaningful results.
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.

Examples of non-additive facts are average, discount, ratios etc. page 119
2. One to many transformation with example 3 marks
To store information we need one to many transformations of names. We need to transform name of each student into 3 columns

- First Name

- Last name
- Student Name (middle part of name)

This type of transformation requires scripts. We will write VB Scripts for such transformations.

3. Ek statement d hoi thi about timestamp btana tha true ha ya false with reason

A timestamp in the product table to record the change.

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. If the timestamp information is not available in an operational source system, you will not always be able to modify the system to include timestamps. Such modification would require, first, modifying the operational system's tables to include a new timestamp column and then creating a trigger to update the timestamp column following every operation that modifies a given row.

4. Star aur snow flock schema ki diagram thi and identify karna tha k kon sa schema ha

Star schema: designs usually used to facilitate ROLAP querying.

A star schema is generally considered to be the most efficient design for two reasons. First, a design with de-normalized tables encounters fewer join operations. Second, most optimizers are smart enough to recognize a star schema and generate access plans that use efficient "star join" operations. It has been established that a "standard template" data warehouse query directly maps to a star schema.

Diagram page106, page 110

Snowflake scheme: Sometimes a pure star schema might suffer performance problems. This can occur when a de-normalized dimension table becomes very large and penalizes the star join operation. Conversely, sometimes a small outer-level dimension table does not incur a significant join cost because it can be permanently stored in a memory buffer. Furthermore, because a star structure exists at the center of a snowflake, an efficient star join can be used to satisfy part of a query. Finally, some queries will not access data from outer-level dimension tables. These queries effectively execute against a star schema that contains smaller dimension tables. Therefore, under some circumstances, a snowflake schema is more efficient than a star schema.

6. 5 marks

Product Id	Region ID	Period	Quantity
1	N	Month	25
2	N	Month	50
2	S	Week	30

Find the primary key & dimension

Answer:

Primary key: product Id, region Id

Dimension: Period, quantity

1.Automatic Data Cleansing

1) Statistical
2) Pattern Based
3) Clustering
4) Association Rules

2. List down any three ways of “handling missing data” during “data cleansing process”.

There are any three ways of “handling missing data” during “data cleansing process”. Handling missing data:

Dropping records.

“Manually” filling missing values.

Using a global constant as filler.

Using the attribute mean (or median) as filler.

Using the most probable value as filler.

Data cleansing process as describing as semi-automatic but can be performance without the involvement of a domain expert.
3. basic tasks of Data Transformation
Basic tasks
Selection
Splitting/Joining
Conversion
Summarization
Enrichment

one q was regarding Data Extraction ...
Extraction is the operation of extracting data from a heterogeneous source system for further use in a data warehouse environment. This is the first step of the ETL process. After the extraction, this data can be transformed, cleansed and loaded into the data warehouse.

Identify at least one factor that lead to such relation.
"Normally performance increase when use more of the disk".
Write this answer....
Performance vs space trade-off
Q. identify the give statement correct or incorrect 5
1. Intelligent learning organization never shares its information what its employees.
Answer: Incorrect
Correct is: The intelligent learning organization shares information openly across the enterprise in a way that maximizes the throughput of the entire organization. Page 181
2. Orr's law say that that data which in not used is always correct.
Answer: Incorrect
Correct is: Law 1: "Data that is not used cannot be correct!" page 181

Q. identify the give statement correct or incorrect 3
If defects are found in the process of attributes domain validation then it is batter to fix the errors in DHW and leave the data source as it as.
Answer: **incorrect**
Correct is: if at all possible, fix the problem in the source system. People have the tendency of applying fixes in the DWH. This is a wrong i.e. if you are fixing the problems in the DW; you are not fixing the root cause. Page 190
Q. identify the give statement correct or incorrect 3
We can dissolve the aggregation to get the original data from which the aggregates were created.
Answer: incorrect
Correct is: Aggregation is one-way i.e. you can create aggregates, but cannot dissolve aggregates to get the original data from which the aggregates were created. Page 113
Q. Mention one factor that lends towards long load time of MOLAP cube. 2
Long load time (pre-calculating the cube may take days!). The biggest drawback is the extremely long time taken to pre-calculate the cubes, remember that in a MOLAP all possible aggregates are calculated.

1) 2 ways to simplify ER?
Answer: De normalization and Dimensional modeling
2) 4 data validation techniques?
Answer: Referential Integrity (RI).
Attribute domain
Using Data Quality Rules
Data Histograming
3) Identify the following statement as correct or incorrect. Justify your answer in either case.
"The less likely something to happen the less traumatic it will be when it happens"
Answer: incorrect

Law #5: "The less likely something is to occur, the more traumatic it will be when it happens!" Page 182
6) suppose ke sales table hai grain is "total sales by day by store "....now identify 3 facts at least.
Answer soch ki likh Dena 3 facts.....such as total products sold , total sales amount etc some thing like this
Answer: prepare 2nd assignment solution for this question

lexical error:
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.

Aggregate: refers to a summarization coupled with a calculation across different business elements. An example of aggregation is the addition of bimonthly salary to monthly commission and bonus to arrive at monthly employee compensation values.

Transformation
Molap ki statement di gei thi and pucha gya tha k correct hai ya nahi justify karna tha

What is merge and purge problem.
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.

Aggregate ka question tha

1. Data cleansing and its role. 5 Marks
Data cleansing is the 3rd step in ETL. It is the activity that is used to remove noise from the input data before bringing it into DWH environment. Data cleansing is vitally important to the overall health of your warehouse project and ultimately affects the health of your company. DO not take this statement lightly. The original aim of data cleansing was to eliminate duplicates in a data collection, a problem occurring already in single database applications and gets worse when integrating data from different sources.
Data cleansing is much more than simply updating a record with good data.
2. A table was given and we are required to identify fact, dimensions ,PK from the table. 5 Marks
3. Performance vs more use of disk space. 3 Marks
Performance vs. Space Trade-off
"Maximum performance boost implies using lots of disk space for storing every pre-calculation"
If storage is not an issue, then just pre-compute every cube at every unique combination of dimensions at every level as it does not cost anything. This will result in maximum query performance. But in reality, this implies huge cost in disk space and the time for constructing the pre-aggregates.

Q. BSN method k steps
Basic Sorted Neighborhood (BSN) Method
Concatenate data into one sequential list of N records
Steps 1: create keys
Compute a key for each record in the list by extracting relevant fields or portions of fields
Step 2: sort data
Sort the records in the data list using the key of step 1
Step 3: merge
Move a fixed size window through the sequential list of records limiting the comparisons for matching records to those records in the window

If the size of the window is *w* records then every new record entering the window is compared with the previous *w-1* records

3, 3 number ki aik statement

facts , transformation

5,5 number ki 2, 2 statement

yah transformation, cleansing, logical data extraction etc

LOGICAL DATA EXTRACTION:

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.

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.

Q4. Identify the following statements as correct or incorrect. (3 marks)

"One of the basic function of OLTP is to show the historical background of an organization."

Answer: incorrect

Correct is: OLTP systems don't keep history, cant get balance statement more than a year old

Q5. Identify the given statements as correct or incorrect and justify your answer in either case.

1. "Transformation is the process in which we extract the data from single/multiple data sources".

Answer: INCORRECT

Correct is: "Transformation is the process in which we exact the data from multiple data sources".

2. "Offline Extraction is a type of Logical Data Extraction".

Answer: INCORRECT

Correct is: "Offline Extraction is a type of Physical data extraction".

3 marks

Q. Identify the given statement as correct or incorrect and justify your answer in either case.

"**Standard Query Language can not be used for querying the MOLAP cube".**

Answer: incorrect

Correct statement is: No, standard query language can not be used for querying the MOLAP cube". 2 marks

Q. The problems associated with the extracted data can correspond to non-primary keys. List down any four problems associated with the non-primary key.

There are any four problems associated with the non-primary key:

Same primary key but different data
Same entity with different keys
Primary key in same information
Source might contain invalid data

---

my cs614 paper.... some new mcqs from my paper with solution

1. The telecommunication data warehouse is dominated by the sheer volume of data generated at the call level _____ area.

- **Subject page 35**
- Object
- Aggregate
- Details

1. 4NF has an additional requirements which is
- Data is in 3NF and no null key dependency
- Data is in 2NF and no Multi value dependency
- **Data is in 3NF and no multi value dependency page 48**
- Data is in 3NF and no foreign key table

1. 3NF remove even more data redundancy than 2NF but it is at the cost of
- **Simplicity and performance page 48**
- Complexity
- No of table
- Relations

1. In full extraction, data extracted completely from source. No need to keep track of change to the _____
- **Data source page 133**
- DWH
- Data mart
- Data destination

1. Which is not the characteristics of DWH
- Ad-hoc access
- Complete repository
- Historical data
- **Volatile page 27**

1. Experienced showed that for a single pass of magnetic tape that scanned 100% of the record only_____ of the records.
- **5% page 12**
- 30%
- 50%
- 80%

1. 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 and high performance page 78**

1. _____ are created out of the data warehouse to service the needs of different departments such as marketing, sales etc.
- MIS
- OLAPs
- **Data mart page 31**
- None of the given

Subjective
1) 2 limitation of aggregation. …… 2 marks
2) Realistic data quality…….2marks
Answer: Degree of utility or value of data to business page 180
3) Check whether the given statement is correct or not. Justify in either case.
Rollup is 3cube used to view of table……3marks
4) Three drawback of data redundancy……. 3marks
5) In DWH data is collected from heterogeneous source. Which create redundancy? How it effect decision of an organization…….5marks

---

Cs614_Midterm paper fall 2014 (Held On Dated 12-01-2015)
Total Questions: 26
Objective: 20 marks
2 questions each having 2 marks
2 questions each having 3 marks
2 questions each having 5 marks

Q.21Which cube operation changes the view of data? (2 marks)
Answer (page 80)
Pivot: change the view of data
Q.22
Answer
Q.23 Change data capture (CDC) is most important step in data extraction. Why? (3 marks)
Answer (page 149)
Without Change Data Capture, database extraction is a cumbersome process in which you move the entire contents of tables into flat files, and then load the files into the data warehouse. This ad hoc approach is expensive in a number of ways.
Q.24 Identify the following statements as correct or incorrect. (3 marks)
"One of the basic function of OLTP is to show the historical background of an organization."
Answer (page 122)
OLTP & Slowly Changing Dimensions
• OLTP systems not good at tracking the past. History never changes.
• OLTP systems are not "static" always evolving, data changing by overwriting.
• Inability of OLTP systems to track history, purged after 90 to 180 days.
Actually don't want to keep historical data for OLTP system
Q.25 List down any three ways of "handling missing data" during "data cleansing process". (5 marks)
Answer (page 162)
• Dropping records.
• "Manually" filling missing values.
• Using a global constant as filler.
• Using the attribute mean (or median) as filler.
• Using the most probable value as filler.
Q.26 Suppose there is a table sale. Grain is "sales by day by product by store. Identify at least three facts so that sales table can easily be built. (5 marks)
Answer (page 74)
• Quantity sold
• Amount
• Sales volume
• Total Rs.sales
Objective almost 40% from from moaaz files and subjective totally out of moaaz file all subjective was from last 10 lectures.
Remember me in your prayers!

---

today my paper cs614 at 11 AM
mcqs past m sy ni thy sb new thy muskil b itni itni long statement k sth
long part easy tha
aditive non adtive wala tabale tha 5 marks
data extration m data caputre k about likhna tha 2 marks
molap k disadvantages likhny thy 2 marks
validation techniques k about ak senrio deia tha k kon c technique use hwi btya jay explian b keya jay 5 marks
2 statement thy
1 law jo 5 deye us m sy 4th num wala law corect h ya ni 3 marks
1 agregate wali statment thi btna tha corect h ya ni 3marks

---

mine paper
mcqs from moaaz file except these
pb value

which is rite option which describe market item etc
in subjective
1st quetion is
basic tasks of Data Transformation
2nd is
he problems associated with the extracted data can correspond to non-primary keys. List down any four problems associated with the non-primary key
3rd is what is ranking
4th is which method is used to remove duplication
5th is what is relation between dimensional cardinality and cube in molap
. Performance vs more use of disk space.

ASSLAM O ALIKUM KO ALL
Mcqz past papers main sy thy. subjective main 1.data transformation k 4 step likhny thy.
2.BNS method k 2 steps name likhny thy.
3.merge/purge problem in data cleasing.
ak statement the
we dissolve aggreate into orginal form (incoorect)
we can't dissolve aggreate into orignal form.
ya tha ak 2 chezain yad nai ALLAH ap sbko
dunia aur akhrat ki kamyab dy. AMEEN

Today-Mid-cs614- paper
Assalam o Alaikum!
About half MCQs from previous mid and final files (see only up to page 193 from final MVQs files). Half th4e MCQs were new and I did about 5 correct (sure) , So paper was OK for me.
Here r MCQs which I did wrong, I wrote on paper only the answers , I attempted and now have checked and most of these are wrong. Very tricky MCQs.
One MCQ The data is extracted completely from the source system. Since this extraction reflects all the data currently available in the source system, there's no need to keep track of changes to the data _____ since the last successful extraction. if u do full extraction then no need to record changes in _____. Source, destination, DWH. Page=133, I attempted DWH and it is wrong correct is Source
One MCQ about data in flight (page 152), I had no idea if data can also fly so attempted wrong.
One MCQ was if grains r increased then what increases. In Solved MCQs it was details but in Lecture 14 at 54:05 minutes and in hand out page 118 , it is dimension , so I did dimension.
Question 21- Four basic tasks of data transfer
22- Describe four non-primary key problems when you get data from a source.
23- Describe three costs of data duplication
24- About Dimensions and their hierarchy
25 Which technique will you use for 1 to many and for many to many values? (See 2nd assignment)
26- You have singe MOLAP table. You partition it and use joins. Enquiries require joins among portioned tables. Will you prefer it for performance?

Today paper of cs614

Two drawbacks of MOLAP (2)
What are the syntactical errors? (2)
Aggregates are generated by given data can we do vice versa? Support your answer with example (3)

What are the effects if govt. uses dirty data (3)
Operations of MOLAP with an example (5)
Given statements are correct or incorrect (5)

Steve fox has sent you a message on Virtual University of Pakistan

Subject: CS614 My Today's Paper (4 Ujeeeeee)

Subjective Portion:

Q21:Two Limitations of data aggregation (2 marks)

Answer: Khud check kar lo...

Q22: Correct the statement "Basic purpose of OLTP is to represent the historical picture of company (2marks)

Answer: Correct statement is "Basic purpose of OLTP is to represent the current 60 to 90 days picture of company

Q23: CDC consider most challenging activity in data extraction? why? (3marks)

Answer: Khud check kar lo...

Q24: Mention any two sources of CDC. Also identify its approaches. (3 marks)

Answer: Sources of CDC: (i)Modern system (ii) Legacy system . Approaches khud kar lain.

Q25. If dirty data in DWH is used by Government for decision making then what would be that affects? (5marks)

Answer: Khud check kar lo...

Q26: Find the relationship b/w cube size and cardinality of dimensions in Context of MOLAP cube? (5marks)

Answer: Khud check kar lo...

Objective Portion:

Total MCQS = 20 (20 marks)

Easy way to learn MCQS from Moaaz file and quizzes. Thanks Pray 4 me!! UJEE Check kar lo

---

4 U UJEEEEE

Subjective Portion:
Q21: Two Limitations of data aggregation (**2 marks**)
Answer: Khud check kar lo...

Q22: Correct the statement "Basic purpose of OLTP is to represent the historical picture of company (**2marks**)
Answer: Correct statement is "Basic purpose of OLTP is to represent the current 60 to 90 days picture of company
Q23: CDC consider most challenging activity in data extraction? why? (**3marks**)
Answer: Khud check kar lo...
Q24: Mention any two sources of CDC. Also identify its approaches. (**3 marks**)
Answer: Sources of CDC: (i)Modern system (ii) Legacy system . Approaches khud kar lain.
Q25. If dirty data in DWH is used by Government for decision making then what would be that affects? (**5marks**)
Answer: Khud check kar lo...
Q26: Find the relationship b/w cube size and cardinality of dimensions in Context of MOLAP cube? (**5marks**)
Answer: Khud check kar lo...
Objective Portion:
Total MCQS = 20 (**20 marks**)
Easy way to learn MCQS from Moaaz file and quizzes. Thanks Pray 4 me!! UJEE Check kar lo

---

My Today Paper

Q.NO 1:
purpose of aggregate awarneness.
Solution:
Aggregate awareness allows using pre-built summary tables by some
front-end tools.
the environment is smart enough to develop or compute higher level aggregates using lower level or more detailed aggregates.(ye statement mcq mn b thi)
Q.NO 2:
Two steps of BSN method:
Solution:
Steps 1: Create Keys
Step 2: Sort Data
Step 3: Merge

Today's Paper 24-12-2013
Total Question = 26
Total Mcqs of 1 marks of each = 20
Total 2 Marks Question = 2
Total 3 Marks Question = 2
Total 5 Marks Question = 2
Objective paper MCQ's 90% from past papers
subject questions
1."Change the data capture" is considered as most challenging activity.why? marks 2
2. Which cube operation is used to change the view of data ? 2 marks
3. Identify the correct or incorrect statement?
"Rollup is a cube operation is used to change the view of data" 3 marks
4. Identify the correct or incorrect statement?
"If defects are found in process of attributes Domain validation then it is better to fix the error in DWH ad leave the data source as it is" 3 marks
5. Identify the correct or incorrect Given two statements?
"Orr's Law says that the data whcih is not used is always correct" 5 marks
2nd statement bhool gai ...
6. Product sales table tha us me facts dimentions and primary key ko identify krna tha 5 marks

my cs614 midterm paper 1/01/2104 fall 2013
50% mcqs from past papers others new
(1) what is the purpose of aggregates awareness 2mrks
(2) first 2 steps of BSN method 2mrks
(3) a statment was given we have to identify the correct statement 3mrks
(4) tow diagrams of schema's were given we have to write the name of schema's 3mrks
ans
(1) star schema (2) snowflake schema
(5) a statement was given we have to write correct with reason 5mrks
(1) data is not used always correct
(2) a intelligent organization does not share his information with their employees
(6) explain the drill down and roll up operations with at least one example 5mrks