Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division

M.S. Systems Engineering at Wipro Info Tech (WIMS)

First Semester 2014 - 2015 (October 2014 to March 2015)

Comprehensive Exam (Regular) – Answer Key

Course Number	:	**SEWI ZG514**
Course Title	:	Data Warehousing
Type of Exam	:	Open Book
Weightage	:	60 %
Duration	:	180 Minutes
Date of Exam	:	1st March 2015

No. of Pages : 5
No. of Questions : 9

Session : AN

Note:
1. Please read and follow all the instructions given on the cover page of the answer script.
2. Start each answer from a fresh page. All parts of a question should be answered consecutively.

---

1. For following statements, indicate True or False with proper justification: (5)

   A. It is a good practice to drop the indexes before the initial load.

   True. Index entry creations during mass loads can be too time-consuming. So drop the indexes prior to the loads to make the loads go quicker. You may rebuild or regenerate the indexes when the loads are complete

   B. The choice of index type depends on cardinality.

   True. Bit-map index can be used only for low cardinality data

   C. The importance of metadata is the same for data warehouse and an operational system.

   False. In an operational system, users get information thru predefined screens and reports. In DW, users seek information thru ad-hoc queries.

   D. Backing up the data warehouse is not necessary because you can recover data from the source systems.

   False. Information in DW is accumulated over long periods and elaborately preprocessed

   E. MPP is a shared-memory parallel hardware configuration.

   False. MPP is a share-nothing hardware architecture.

2. Make the selection of best option among the given multiple choices. (1 mark each)

i. The reason(s) for partitioning the fact table is/are
I. To increase the performance. II. To implement access control. III. To assist backup/recovery.
(a) Both (I) and (II) above
(b) Both (II) and (III) above
(c) ***Both (I) and (III) above***
(d) All (I), (II) and (III) above.

ii. Fact table contains:
a) Measures and multiple Foreign keys
b) One primary key and textual data.
c) Multiple Foreign keys and textual data
d) None of these

iii. An operational system is which of the following:- (mid-ranjita)
a) A system that is used to run the business in real time and is based on historical data.
b) A system that is used to run the business in real time and is based on current data.
c) A system that is used to support decision making and is based on current data.
d) A system that is used to support decision making and is based on historical data.

iv. Which of the following is/are produced in the technical blueprint stage of data warehouse delivery process?
I. Detailed design of database. II. Essential components of database design.
III. Server and data mart architecture. IV. Backup and recovery strategy.
a) Only (III) above
b) Both (I) and (IV) above
c) (I), (III) and (IV) above
d) (II), (III) and (IV) above

3. Consider a data warehouse, where the fact data is calculated to be 36GB of data per year, and 4 years' worth of data are to be kept online. The data is to be partitioned by month and four concurrent queries are to be allowed.
   Compute the partition size, Temporary Space and Space Required for this scenario. (6 marks)

   Partition size P = 36GB per year / 12 = 3 GB
   T = (2n +1)P = [(2 x 4) + 1]3 = 27 GB
   F = 36GB X 4 years = 144 GB
   Space Required = 3.5F + T = 3.5 X 144 + 27 = 531 GB

4. While designing fact table it is being advised to use non-intelligent keys; but while designing summary table it is advised to use intelligent key. Explain.

   Unlike fact tables, summary tables are re-created on a regular basis, possibly every time new data is loaded. The purpose of the summary is to speed up queries, and reduce data joins with dimension data takes significant amounts of time

5. Discuss the merits and demerits of using views from the perspective of security of data warehouse.

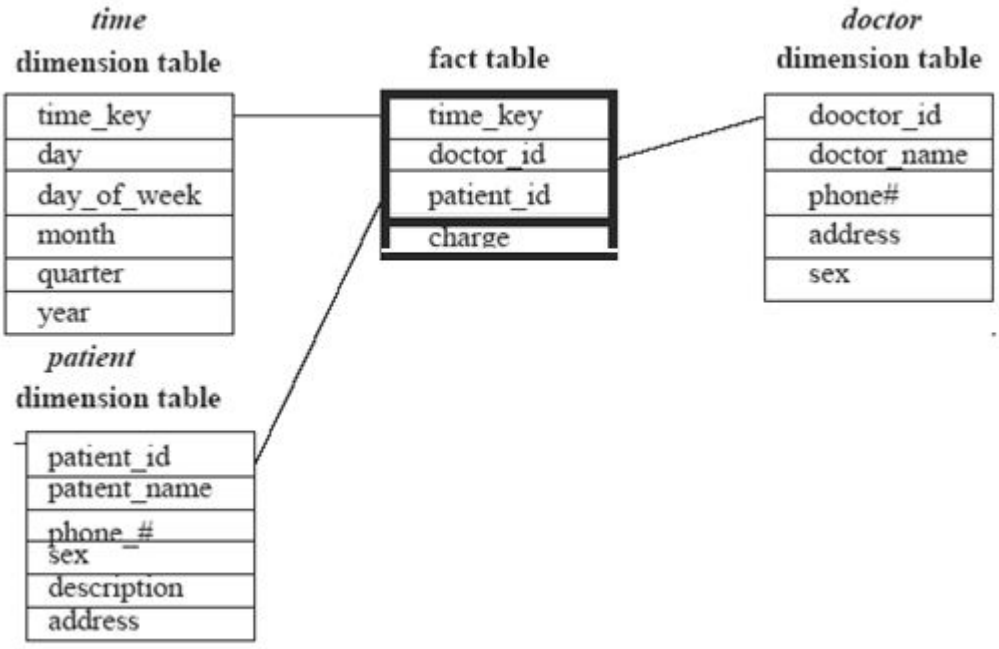
   Views are easier option to define security initially. Later it will cause challenges.
   Some of the common restrictions that may apply to the handling of views are:
   - restricted data manipulation language (DML) operations,
   - lost query optimization paths,
   - restrictions on parallel processing of view projections.

   The use of views to enforce security will impose a maintenance overhead. In particular, if views are used to enforce restricted access to data tables and aggregations, as these changes, the views may also change.

6. Give the merits and demerits of the usage of mini-dimensions over type2 SCD (slowly changing dimensions)

   The Type 2 method tracks historical data by creating multiple records for a given natural key in the dimensional tables with separate surrogate keys and/or different version numbers. With Type 2, we have unlimited history preservation as a new record is inserted each time a change is made. In some DW schemas, e.g. bank accounts, there are a wide variety of attributes to describe accounts, customers, and households, including monthly credit bureau attributes, external demographic data, and calculated scores to identify their behavior, retention, profitability, and delinquency characteristics. It is unreasonable to rely on the type 2 SCD technique to track changes in the account dimension given the dimension row count and attribute volatility, such as the monthly update of credit bureau attributes. . Instead, we break off the browseable and changeable attributes into multiple minidimensions, such as credit bureau and demographics minidimensions, whose keys are included in the fact table. One of the compromises associated with minidimensions is the need to band attribute values in order to maintain reasonable minidimension row counts. Rather than storing extremely discrete income amounts, such as $31,257.98, we store income ranges, such as $30,000-$34,999 in the minidimension.

7. Suppose that a hospital data warehouse consists of the three dimensions time, doctor, and patient, and a measure for fees charged for a visit.
   (a) Draw a schema diagram for the above data warehouse
   (b) Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?



(a) Starting with the base cuboid [**day, doctor, patient**], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

1. roll up from day to month to year
2. slice for year = "2004"
3. roll up on patient from individual patient to all

8. An electronics stores chain is keen on obtaining the following information
    I. Sale of products by make, model, location, price range
    II. Profitability of various products. Here consider design options for a stable cost price, and volatile cost price scenarios.
    a. Design a dimensional model that can help management obtain the answers.
    b. Management have noticed that sale of products drop when there are promotional offers from competitors. Incorporate necessary elements in the design so that impact of competitor promotion can be identified. (6+4)

    a)The dimensional model can have Sales fact table with location, product, and time dimensions. Analysis by price range can be achieved by having unit price as a fact table measure. Profitability can be obtained by including cost price in fact table. If cost price is stable, it can be made an attribute in product dimension.
    b) The promotional offers from competitors can be incorporated in time dimension to measure their impact on store sales.

9. You work for an IT consulting organization. Your client mentioned about performance concerns with their data warehouse. Send a questionnaire (with at least 10 questions) to understand their situation. (10)

    Questions should check on
    a) Schema and sizes of fact and dimension tables
    b) Types of Indexes : B-tree, Hash, Clustered, Bit-map
    c) Data Partitioning
    d) Hardware (processor, memory, network) configuration
    e) Aggregates
    f) Referential Integrity Checks
    g) MOLAP tools
    h) ETL Tools
    i) Most frequent queries
    j) Concurrent Users