Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division
M.S/M. Tech (Software Engineering) at Wipro Technologies (WASE)
II Semester 2015 - 2016
Comprehensive Examination (Regular)_ANSWER KEY

Course Number : SEWP ZG514
Course Title : DATA WAREHOUSING
Type of Exam : Closed Book
Weightage : 60 Marks
Duration : 3 Hours
Date of Exam : 24 July,2016 Session : FN(9 to 12 Noon) Note:

No. of Pages : 11
No. of Questions : 06

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.

1a. Why do we need a separate Data Warehouse? What data is stored in a warehouse? Discuss with example how do we represent this data? **8 Marks**

Answer :

Main reasons:

1. OLTP systems require high concurrency, reliability, locking which provide good performance for short and simple OLTP queries. An OLAP query is very complex and does not require these properties. Use of OLAP query on OLTP system degrades its performance.

2.An OLAP query reads HUGE amount of data and generates the required result. The query is very complex too. Thus special primitiveshave to provided to support this kind of data access.

3.OLAP systems access historical data and not current volatile data while OLTP systems access current up-to-date data and do not need historical data.

----- 3 Marks

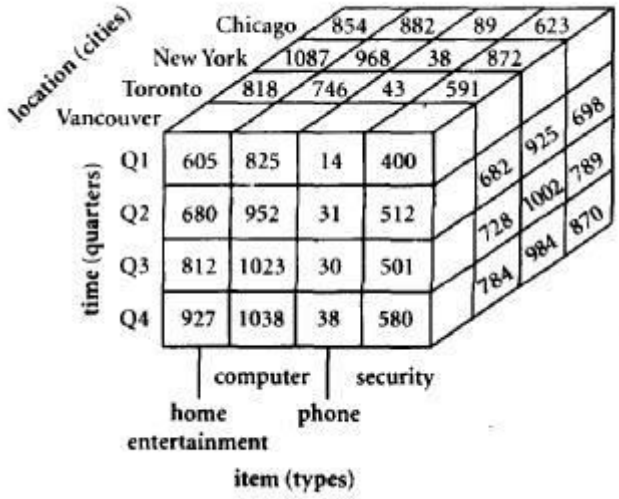
Thus, Solution is to have a separate database system which supports primitives and structures suitable to store, access and process OLAP specific data, in short have a data warehouse.

In simple words: Subject(s) per Dimension
Example: If our subject/measure is ‘quantity sold and if the dimensions are : Item Type, Location and Period then, Data warehouse stores the items sold per type, per geographical location during the particular period.

----- 2 Marks

Data Cube

This multi-dimensional data can be represented using a data cube as shown below.



This figure shows a 3-Dimensional dataModel.
X –Dimension : Item type
Y –Dimension : Time/Period
Z –Dimension : Location
Each cell represents the items sold of type ‘x’, in location ‘z’during the quarter‘y’. This is easily visualized as Dimensions are 3.

----- 3 Marks

1b. What is slicing and dicing? Explain with real time usage and business reasons of it's use.
2 Marks.

Answer :

Slicing and Dicing is a feature that helps us in seeing the more detailed information about a particular thing.

For eg: You have a report which shows the quarterly based performance of a particular product. But you want to see it in a monthly wise. So you can use slicing and dicing technique to drill down to monthly level.

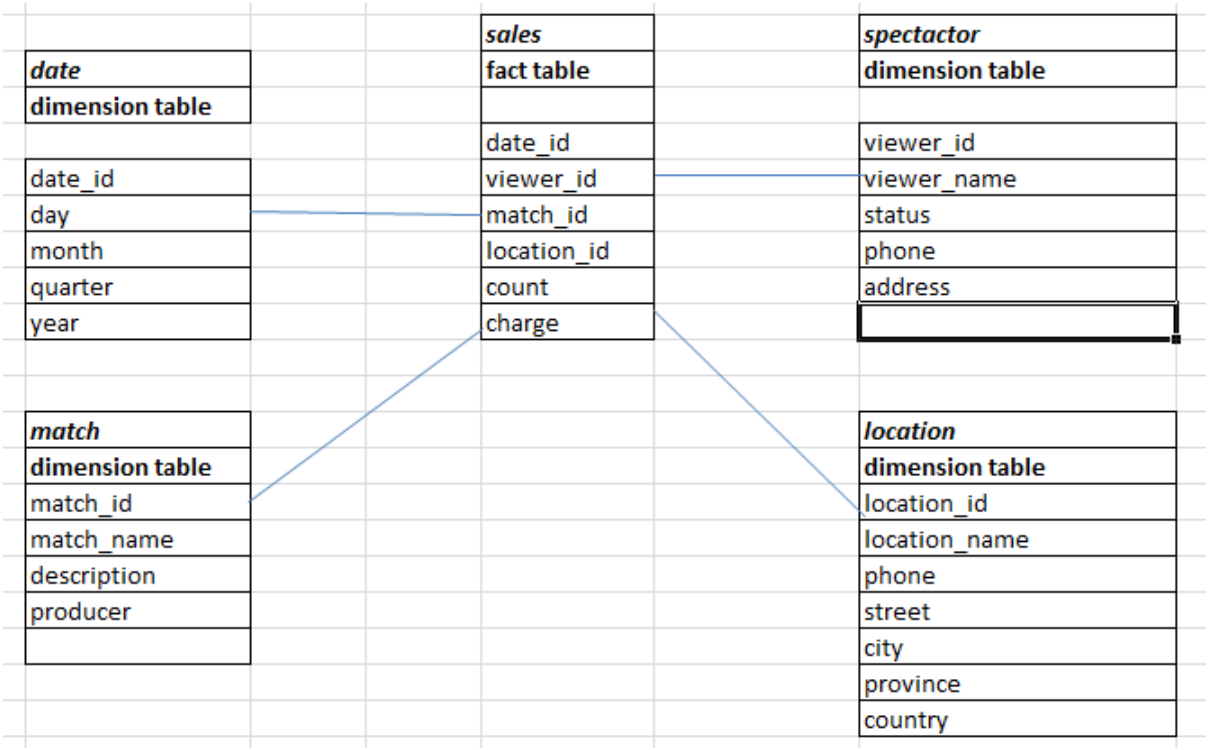
2a. Suppose that a data warehouse consists of four dimensions, date, viewer, location, and match, and the two measures, count and charge, where charge is the fare that a viewer pays when watching a match on a given date. Viewers may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboid [date, viewer, location, match], what specific OLAP operations should one perform in order to list the total charge paid by seniors viewer at National Stadium in 2015? **7 Marks**

Answer :

(a) Star schema diagram for the data warehouse.



Students should specify all suitable attributes and required keys for both fact and dimension tables.

----- 4 Marks

(b) The specific OLAP operations to be performed are:

•Roll-up on date from date id to year.

•Roll-up on game from game id to all.

•Roll-up on location from location id to location name.

•Roll-up on spectator from spectator id to status.

•Dice with status="students", location name="National Stadium", and year=2015.

----- 3 Marks

2b. A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer.

3 Marks

Answer :

Star schema and snowflake schema are similar in the sense that they all have a fact table, as well as some dimensional tables. The major difference is that some dimension tables in the snowflake schema are normalized, there by further splitting the data into additional tables. The advantage for star schema is its simplicity, which will enable efficiency, but it requires more space. For snowflake schema, it reduce some redundancy by sharing common tables: The tables are easy to maintain and save some space.
However, it is less efficient, and the saving of space is negligible in comparison with the typical magnitude of the fact table. Therefore, empirically, star schema is better simply because of efficiency has higher priority over space, if it is not too huge.

----- 2 Marks

Sometimes in industry, to speed up processing, people "denormalize data from a snowflake schema into a star schema" .
Another option here is that "some practitioners use a snowflake schema to maintain dimensions, and then present users with the same data collapsed into a star".

----- 1 Marks

3a. Radhika is one of the most popular Project Managers in the DWH department of a company. Users are very happy with the services provided by her team, especially the efficiency with which the response was provided and the quality of the response. One day, she provided a complex report to the CEO, the CEO called her and asked the following questions with respect to the report:

- How may previous years' data has been used to create this report?
- There were some major changes in the data formats done last year. How have these been considered in the report?
- The company has added new locations last quarter to its operations. Have these been included?
- There have been extra operations planned during the last two quarters during weekends and holidays. Have these been considered in the report?
- What logic has been used to compute the total revenue and profitability?

How will Radhika answer the questions to his satisfaction and keep her reputation? What features of the Data Warehouse will she use to defend her position?

6 Marks

Answer :

Radhika would be used 1 or 5 years historical data, which will be depend on strategies of the company.
Radhika has to used different Data Marts, if data marts has been created based on time variant.
Radhika has to use date dimension effectively for considering weekends and holidays transactional data.
Radhika will make use of the DWH Metadata to answer the questions. She will provide the processing logic and methodology documented in the metadata to explain the same to the users, including the CEO.
The ETL metadata indicates which data from the operational data is extracted into the DWH. It will contain details of any special data included in the reports.
The query processing metadata will indicate what formulae given by the users is being used for doing the computations in the reports etc.
The metadata will also indicate how data in different formats over the past years will be integrated to answer queries spanning large time frames.

3b. (i) Suppose there are on average 10,000 sale transactions in a grocery store daily. If the transactions contain an average of 4 items (from 250,000 items in store), how many rows would be added to the store's data mart's base fact table every week? Assume the store is open 7 days a week.

(ii) Suppose the grocery store's data mart also has a 2-way aggregate fact table that summarizes weekly sales of item categories. If there are on average 100 items per category and items from all categories are sold each week, how many additional rows would be added to this fact table per week? **4 Marks**

Answer :

(i) Each of the 4 items of a transaction will form 1 row in the fact table (since the grain for a grocery store is the sale of one item).
Each day there are 10,000 sales transactions, which results in 40,000 (=10,000x4) rows added to the fact table.
Hence every week there will be 280,000 (=40,000x7) rows added to the fact table.
----- 2 Marks

(ii) There are 100 items per category on an average.
No of categories = 250,000/100 = 2,500
Weekly sales aggregate of item categories will have one fact table row per item category. Since at least one item is sold per week in a category, there will be one row for each of the 2500 categories in the fact table. (if more than one item in a category is sold in a week, the details will be aggregated into the aggregate same row for that category).
Hence 2500 aggregate rows will be added to the fact table every week.
----- 2 Marks

4a. What is meant by the selectivity for a column in a physical table? What type of indexing technique is suitable for low-selectivity data? Explain with example. **3 Marks**

Answer :

The selectivity is a measure of how much variety is in the values of a given table column in relation to the total number of rows in a given table.

---- 1.5 Marks

Bitmapped indexes are ideally suitable for low-selectivity data because it take significantly less space than B-Tree indexes for low-selectivity columns.

E.g. Example may vary from student to student.

---- 1.5 Marks

4b. Why dimension tables are wide and the fact table is deep? Explain with example.

3 Marks

Answer :

A dimension table contains higher granular information so have lesser number of records. As it needs to have all the necessary details means more columns related to the grain of the table, means wider in nature. A fact table has the lowest granularity of a subject area. Lower grain causes more number of rows in the Fact table, means deeper the table.

-----2 Marks

Example may vary from student to student.

-----1 Marks

4c. Consider the following relation Cars: **4 Marks**

Brand	Type	Color	Risk
Opel	Corsa	Grey	Low
Opel	Corsa	Red	Medium
Peugeot	206	Black	Medium
BMW	A	Black	Low

(i) Construct a bitmap index for the attributes Brand and Color for this table.

(ii) Indicate how these two bitmap-indices can be used to answer the query: Give the total number of red Opel cars with a medium risk score.

Answer :

i)Bitmap indices for Brand and Color:

BRAND

Opel	Peugeot	BMW
1	0	0
1	0	0
0	1	0

0	0	1

COLOR

Grey	Red	Black
1	0	0
0	1	0
0	0	1
0	0	1

-----2 Marks

ii) First we intersect the bitmap for Opel with the bitmap index for Red by taking the logical AND of the bitmaps, resulting in the bitmap: (0 1 0 0). Then, for all 1- entries, we directly access the corresponding tuple in the relation, check the condition Risk=medium, and if this condition is satisfied, we count the tuple. Hence, in this case, we access the second tuple only. As the tuple satisfies the condition, it is counted.
The final result 1 is returned.
Often the following error was made: as a first step, a bitmap index for
Risk is constructed, and then the intersection of three bitmaps is made.
Constructing the bitmap index for Risk, however, requires a full scan of the database first and hence results in an evaluation strategy that is less efficient than not using any bitmap index at all.

-----2 Marks

5. An insurance company, with branches all over the country, wants to develop a data warehouse for effective decision-making about their insurance policies. There are a number of different types of insurance like Auto insurance, Home insurance, Industrial insurance, etc. The entire country is categorized into four regions, namely, North, South, East and West. Each region consists of a set of states. There may be different types of customers like individuals, institution, industry, etc. The data warehouse should record an entry for each policy issued to each customer along with the premium paid.
With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.

i. Follow four-Step Dimensional Design Process

ii. Design a star schema for insurance claim by clearly identifying the fact table, dimensional table(s), their attributes and measures along with the primary key and foreign key relationships.

(Marks 10)

Answer :

The four key decisions made during the design of a dimensional model include:
1. Select the business process.
2. Declare the grain.

3. Identify the dimensions.

4. Identify the facts

Business processes are the operational activities performed by your organization, such as regestering an insurance policy, processing an insurance claim, registering employees, or snapshotting every account each month.

---- 1 Mark

Declaring the grain is the pivotal step in a dimensional design. The grain establishes exactly what a single fact table row represents. **---- 1 Mark**

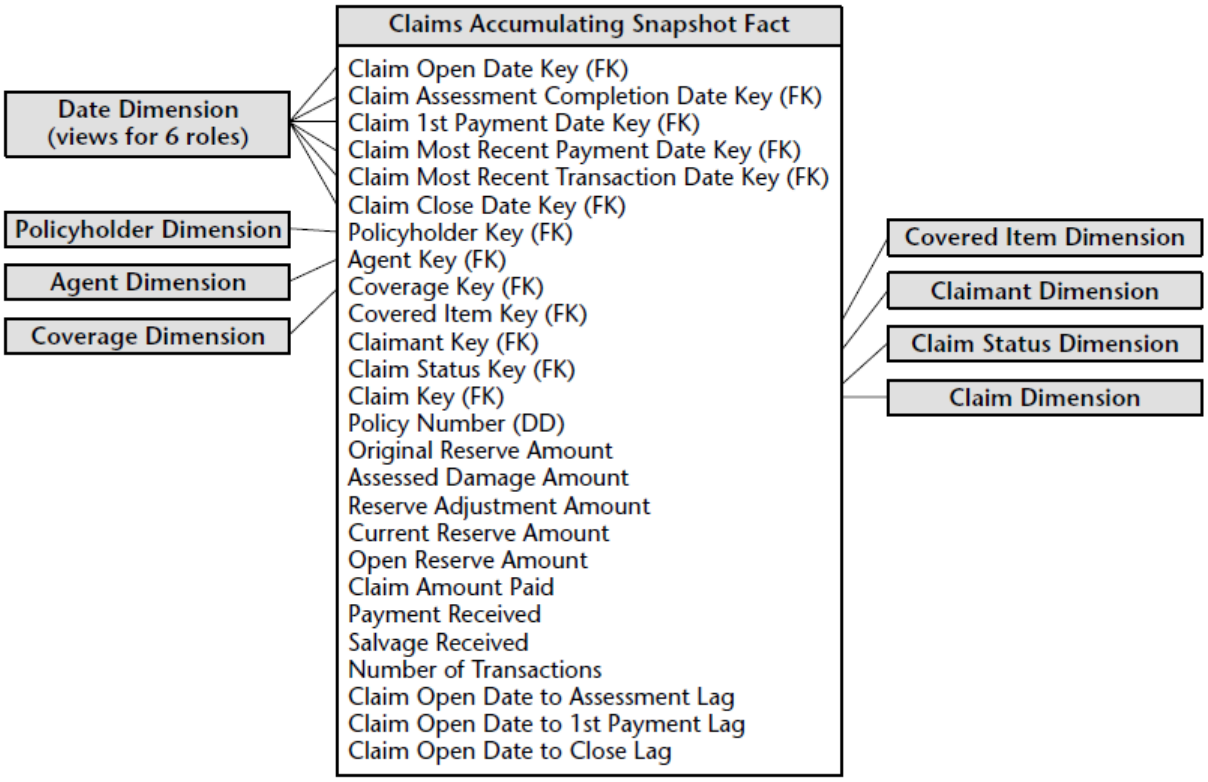
Dimensions are as follows: -----**3 marks**

- Date
- Policy /Insurance
- Claim
- Coverage
- Insured party
- Third Party
- Transaction(Premium)
- Employee
- Region

Fact Details are as follows : ----**2 Marks**

- Transaction Date Key
- Effective Date Key
- Insured party Key
- Employee Key
- Coverage Key
- Policy Key
- Claim Key
- Third Party Key
- Transaction Key
- Amount

Identification of attributes for above mentioned dimension and design of star schema along with primary key and foreign key ----**3 Marks**



Sample schema diagram.

6a. What are the advantages and disadvantages of having finest granularity data in the data warehouse and data marts? 5 marks

Answer :

Most important design issue. Controls the volume of data in the DW.

Advantages:

- Reusability & Flexibility – DW provides an invaluable foundation for many different types of DSS processing. Organizations may build a DW for one purpose & then discover that it can be used for many other kinds of DSS processing. Although infrastructure for the DW is expensive & difficult to build, it has to be built only once. Granular data is the key to reusability. Different departments are able to look at the data as they wish to see it. A related benefit is the ability to reconcile data, if needed. Once there is a single foundation (granular data) on which everybody relies, if there is a need to explain a discrepancy in analyses between two or more departments, then reconciliation is relatively simple. Another related benefit is flexibility. Any department can alter how it looks at data.
- Allows us to issue more versatile queries. The lower the level of granularity, the more versatile the query can be issued.
- Future unknown requirements can be accommodated easily on the foundation of granular data

- Simplified ETL process as no summarization is required.
- Granular data is an ideal source of data for the downstream Data Mining applications
- Dimension of any granularity can be added to an existing star schema.
- In the bottom-up approach, the integration of data marts becomes easy if the data marts are having the most granular data (super marts). In the top-down approach, each data mart that sources its data from the DW may have its own unique requirements. Any such requirement can be handled by the most granular data

Disadvantages:

- Data volume could become unmanageable. For example, the web log data generated by the web based eBusiness environment (often called clickstream data) is at a very low level of granularity. It must be edited, filtered, and summarized before its granularity is fit for the DW environment.

6b. You are a senior analyst in the IT department of a company manufacturing automobile parts. The marketing VP is complaining about the poor response by IT in providing strategic information. Draft a proposal to him introducing the concept of business intelligence and how data warehousing and analytics as part of business intelligence for your company would be the optimal solution. **5 Marks**

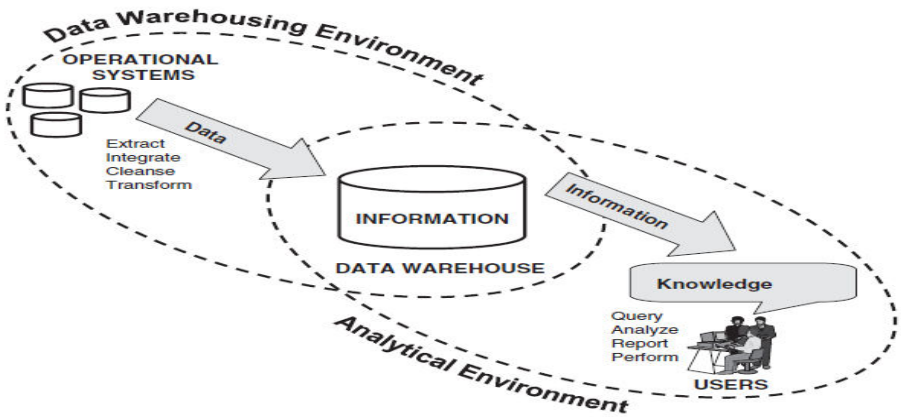
Answer :

The initial challenges following the adoption of early data warehousing systems forced companies
to take a second look at providing decision support

Business intelligence for an enterprise as composed of two environments:
Data to Information. In this environment data from multiple operational systems are extracted, integrated, cleansed, transformed and stored as information in specially designed repositories.

Information to Knowledge. In this environment analytical tools are made available to users to access and analyze the information content in the specially designed repositories and turn information into knowledge.

In today's businesses, extraction, consolidation, transformation, and storing of data as strategic information is a formidable task. Again, using this information with sophisticated tools for proper decision making is equally challenging.


Data Warehousing Environment
OPERATIONAL SYSTEMS
Data
Extract
Integrate
Cleanse
Transform
INFORMATION
DATA WAREHOUSE
Information
Knowledge
Query
Analyze
Report
Perform
USERS
Analytical Environment