Birla Institute of Technology and Science, Pilani

Work Integrated Learning Programmes Division

BITS-Wipro Infotech Collaboration: M. Tech. Systems Engineering

II SEMESTER 2015-2016

COMPREHENSIVE SEMESTER EXAMINATION (Regular)

Course Title	**: Data Warehousing**	**Date**	**:**
Course No.	**: SEWP ZG514**	**Nature**	**: Open book**
Weightage	**: 60 marks**	**Duration**	**: 3 Hrs.**

Note: Attempt all the questions

Q1. (a) List five types of data warehouse architectures. Explain any two of them with the help of suitable diagrams
[5].

Answer:

List:

- Independent Data Marts
- Data Mart Bus Architecture
- Hub-and-Spoke
- Centralized Data Warehouse
- Federated Architecture

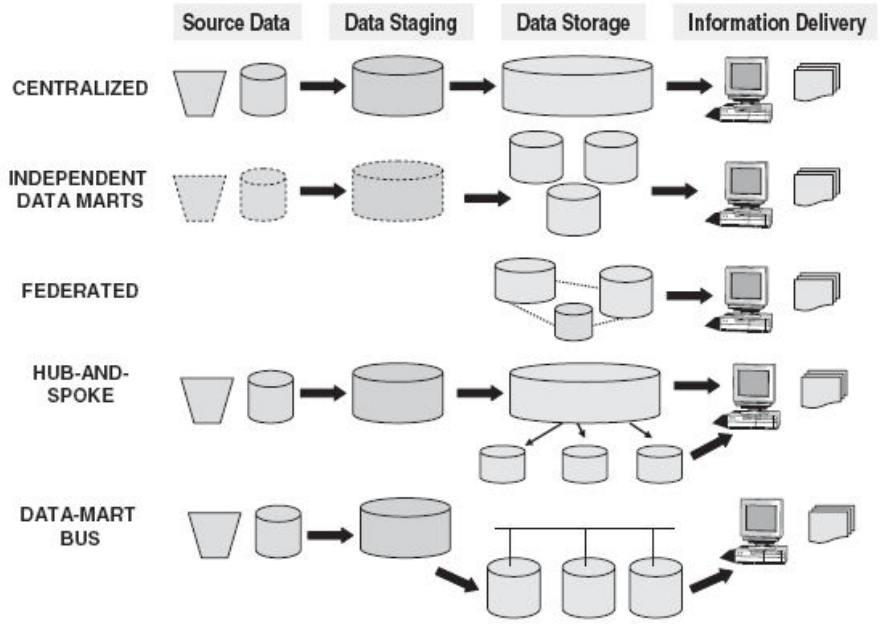


Figure 2-6 Data warehouse architectural types.

Centralized Data Warehouse

This architectural type takes into account the enterprise-level information requirements. An overall infrastructure is established. Atomic level normalized data at the lowest level of granularity is stored in the third normal form. Occasionally, some summarized data is included. Queries and applications access the normalized data in the central data warehouse. There are no separate data marts.

Independent Data Marts

This architectural type evolves in companies where the organizational units develop their own data marts for their own specific purposes. Although each data mart serves the particular organizational unit, these separate data marts do not provide "a single version of the truth." The data marts are independent of one another. As a result, these different data marts are likely to have inconsistent data definitions and standards. Such variances hinder analysis of data across data marts. For example, if there are two independent data marts, one for sales and the other for shipments, although sales and shipments are related subjects, the independent data marts would make it difficult to analyze sales and shipments data together.

Federated

Some companies get into data warehousing with an existing legacy of an assortment of decision-support structures in the form of operational systems, extracted datasets, primitive data marts, and so on. For such companies, it may not be prudent to discard all that huge investment and start from scratch. The practical solution is a federated architectural type where data may be physically or logically integrated through shared key fields, overall global metadata, distributed queries, and such other methods. In this architectural type, there is no one overall data warehouse.

Hub-and-Spoke

This is the Inmon Corporate Information Factory approach. Similar to the centralized data warehouse architecture, here too is an overall enterprise-wide data warehouse. Atomic data in the third normal form is stored in the centralized data warehouse. The major and useful difference is the presence of dependent data marts in this architectural type. Dependent data marts obtain data from the centralized data warehouse. The centralized data warehouse forms the hub to feed data to the data marts on the spokes. The dependent data marts may be developed for a variety of purposes: departmental analytical needs, specialized queries, data mining, and so on. Each dependent dart mart may have normalized, denormalized, summarized, or dimensional data structures based on individual requirements. Most queries are directed to the dependent data marts although the centralized data warehouse may itself be used for querying. This architectural type results from adopting a top-down approach to data warehouse development.

Data-Mart Bus

This is the Kimbal conformed supermarts approach. You begin with analyzing requirements for a specific business subject such as orders, shipments, billings, insurance claims, car rentals, and so on. You build the first data mart (supermart) using business dimensions and metrics. These business dimensions will be shared in the future data marts. The principal notion is that by conforming dimensions among the various data marts, the result would be logically integrated supermarts that will provide an enterprise view of the data. The data marts contain atomic data organized as a dimensional data model. This architectural type results from adopting an enhanced bottom-up approach to data warehouse development.

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

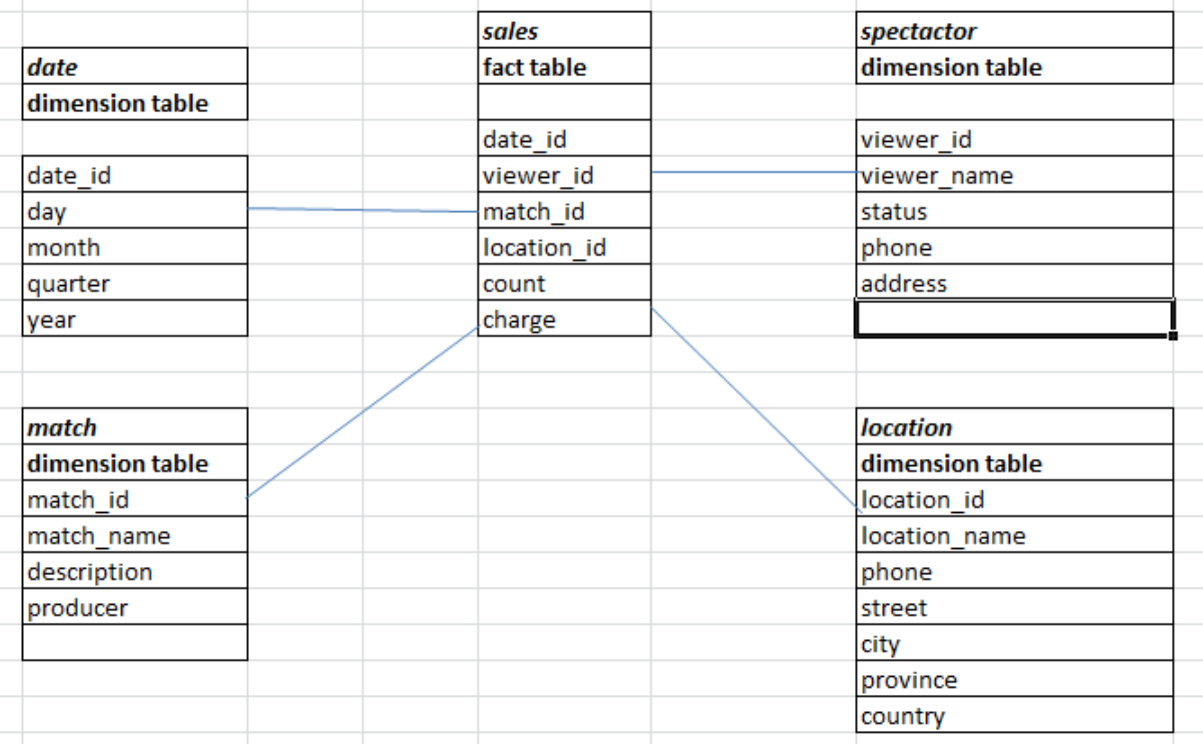
Ans. Slicing and Dicing is a feature that helps us in seeing the more detailed information about a particular thing.

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.

Q2. 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. [4]
(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? 5 Marks

Answer :

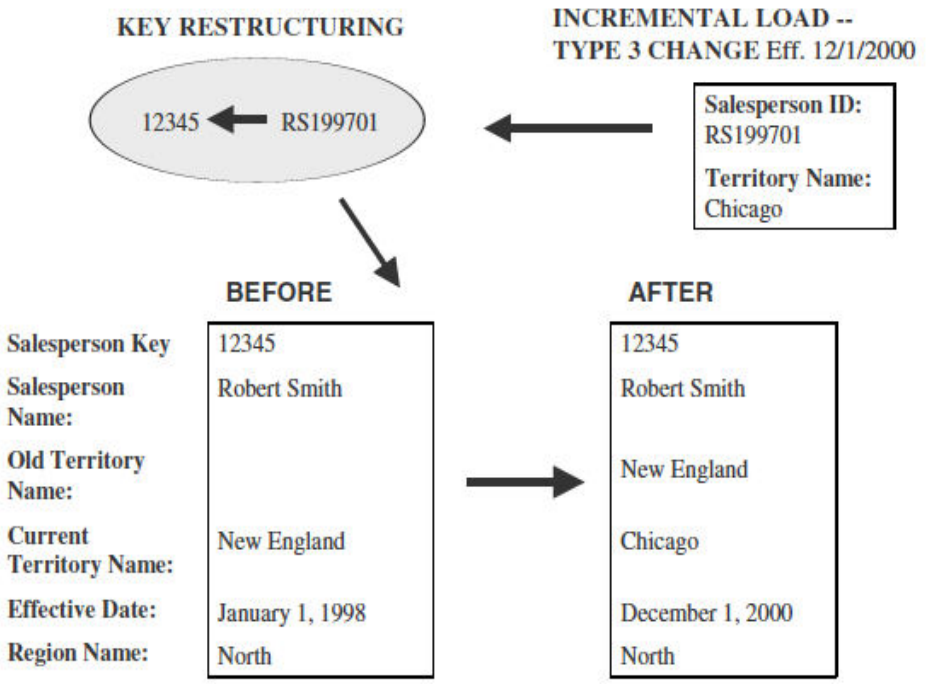
(a) Star schema diagram for the data warehouse.



Q3. (a) With the suitable example, explain the TYPE III change for the slowly changing dimension [5].

Ans:

- Type 3 changes: When you want to compare old and new values of attributes for a given period
- Please note that in Type 2 changes the old values and new values were not comparable before or after the cut-off date (when the address was changed)
- Solution: Add a new column of attribute



(b) 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. 7 Marks

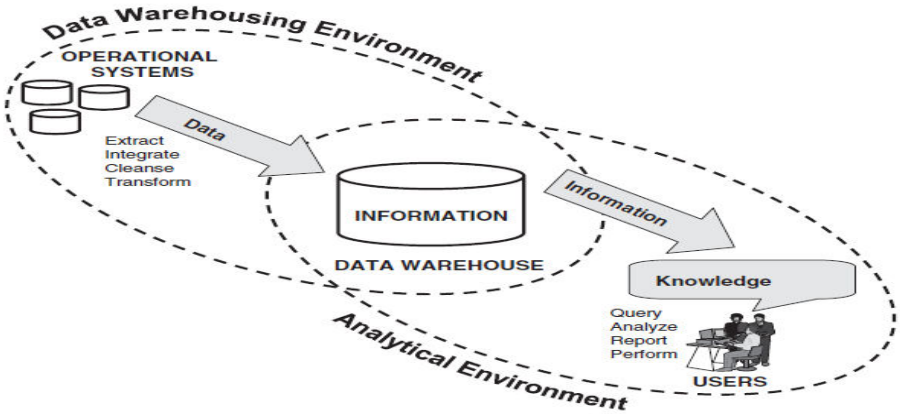
Answer :

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

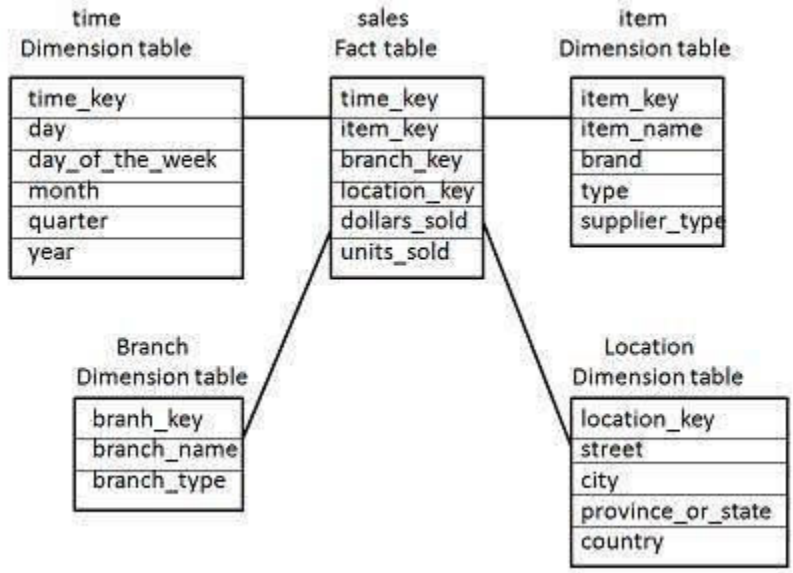
Business intelligence for an enterprise as composed of two environments:
Data to Information. In this environment data from multiple operational systems are extracted, integrated, cleansed, transformed and stored as information in specially designed repositories.

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.



Q4. Draw the start schema for sales data of a company with respect to the four dimensions, namely time, item, branch, and location. [7]



Q5. (a) What is dimensional modeling? [2]

Dimensional model consists of dimension and fact tables. Fact tables store different transactional measurements and the foreign keys from dimension tables that qualifies the data. The goal of Dimensional model is **not** to achieve high degree of normalization but to facilitate easy and faster data retrieval.

Ralph Kimball is one of the strongest proponents of this very popular data modeling technique which is often used in many enterprise level data warehouses.

If you want to read a quick and simple guide on dimensional modeling, please check our Guide to dimensional modeling.

(b) . What is Check Integrity?[3]

Checks the validity of the active universe including its structure, joins, cardinalities, objects, contexts, and conditions. It can also detect whether there are any loops. You can check the entire universe or only certain of its components.

(c). What is a surrogate key? Why are we going for surrogate keys? [2]

Surrogate keys are the keys that are maintained within the data warehouse instead of keys taken from source data systems.

1. Data tables in various source systems may use different keys for the same entity.
2. Keys may change or be used in the data source systems.
3. Changes in organizational structures may move the keys in the hierarchy.

(d). What are OLAP, MOLAP, ROLAP, DOLAP, and HOLAP? Examples? [6]

OLAP - On-Line Analytical Processing: Designates a category of applications and technologies that allow the collection, storage, manipulation and reproduction of multidimensional data, with the goal of analysis.

MOLAP - Multidimensional OLAP: This term designates a Cartesian data structure more specifically. In effect, MOLAP contrasts with ROLAP. In the former, joins between tables are already suitable, which enhances performances. In the latter, joins are computed during the request. Targeted at groups of users because it's a shared environment. Data is stored in an exclusive server-based format. It performs more complex analysis of data.

DOLAP - Desktop OLAP**: Small OLAP products for local multi dimensional analysis Desktop OLAP. There can be a mini multi dimensional database (using Personal Express), or extraction of a data cube (using Business Objects). Designed for low-end, single, departmental user. Data is stored in cubes on the desktop. It's like having your own spreadsheet. Since the data is local, end users don't have to worry about performance hits against the server.**

ROLAP - Relational OLAP. Designates one or several star schemas stored in relational databases. This technology permits multidimensional analysis with data stored in relational databases. Used for large departments or groups because it supports large amounts of data and users.

HOLAP: Hybridization of OLAP, which can include any of the above.

(e) Write Advantage of different types of portioning in data warehousing? [4]

Q7. List the dimensions and facts for Clinical information system and design star and snowflake schema. [8]

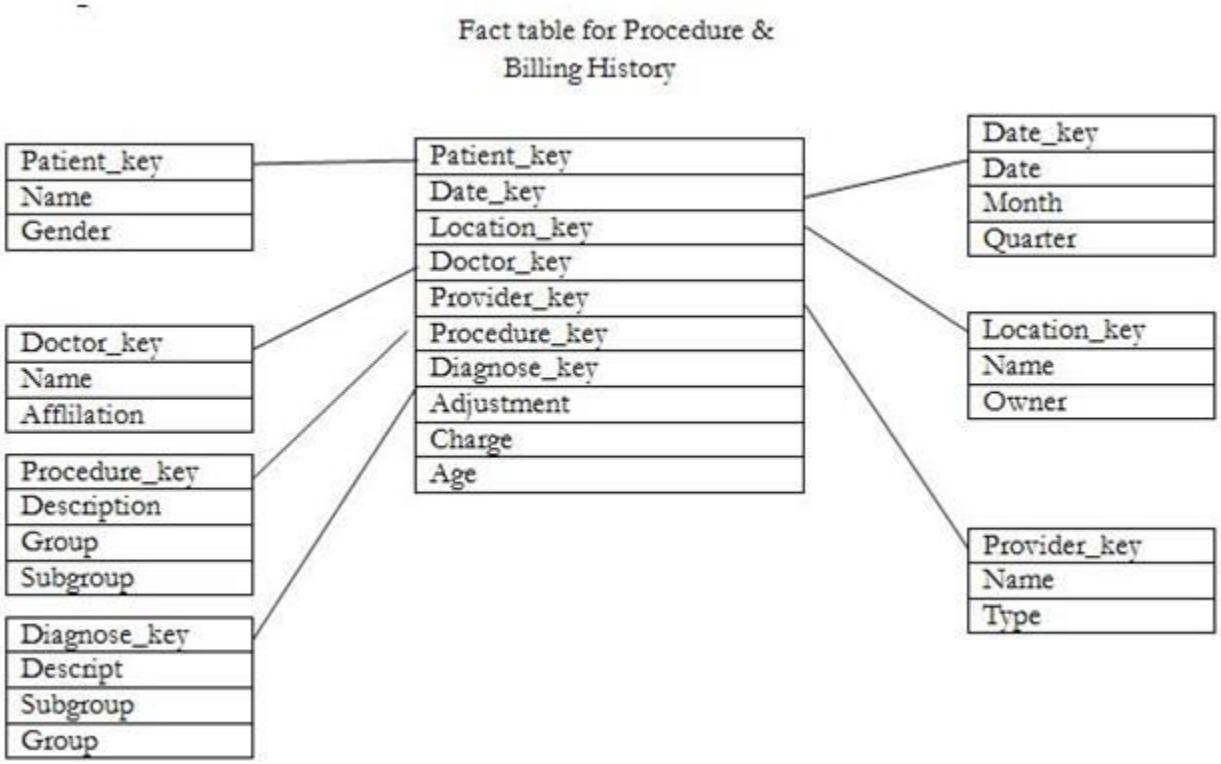
Dimensions:

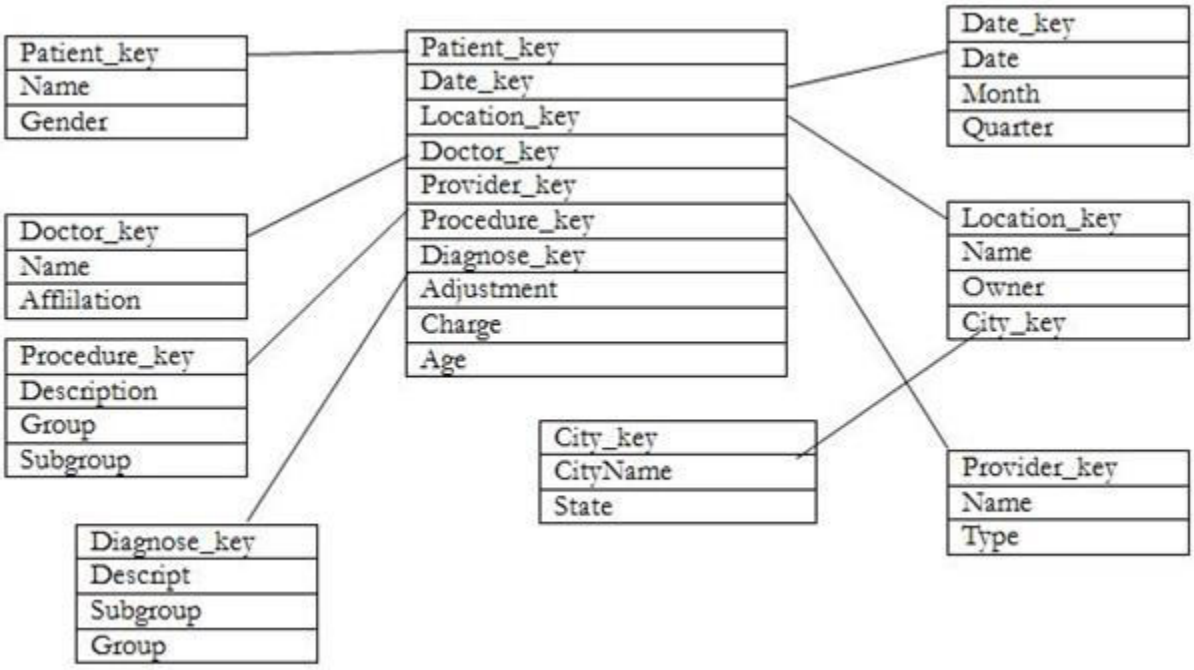
1. Patient
2. Doctor
3. Procedure
4. Diagnose

5. Date of Service
6. Location
7. Provider

Fact:

1. Adjustment
2. Charge
3. Age




Patient_key
Name
Gender
Doctor_key
Name
Affilation
Procedure_key
Description
Group
Subgroup
Diagnose_key
Descript
Subgroup
Group
Patient_key
Date_key
Location_key
Doctor_key
Provider_key
Procedure_key
Diagnose_key
Adjustment
Charge
Age
City_key
CityName
State
Date_key
Date
Month
Quarter
Location_key
Name
Owner
City_key
Provider_key
Name
Type