CHAPTER 2

Unit II

Data Warehouse

Syllabus Topics

Data Warehouse, Operational Database Systems and Data Warehouses (OLTP Vs OLAP), A Multidimensional Data Model : Data Cubes, Stars, Snowflakes, and Fact Constellations Schemas; OLAP Operations in the Multidimensional Data Model, Concept Hierarchies, Data Warehouse Architecture, The Process of Data Warehouse Design, A three-tier data warehousing architecture, Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP.

Syllabus Topic : Data Warehouse

2.1 Data Warehouse

Q. Define Data Warehouse. (2 Marks)

- Precisely, a data warehouse system proves to be helpful in providing collective information to all its users. It is mainly created to support different analysis, queries that need extensive searching on a larger scale.
- With the help of Data warehousing technology, every industry right from retail industry to financial institutions, manufacturing enterprises, government department, airline companies people are changing the way they perform business analysis and strategic decision making.

The term Data Warehouse was defined by Bill Inmon in 1990, in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". He defined the terms in the sentence as follows :

Subject Oriented

Data that gives information about a particular subject instead of about a company's ongoing operations.

Integrated

Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

Time-variant

All data in the data warehouse is identified with a particular time period.

Non-volatile

- Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.
- Ralph Kimball provided a much simpler definition of a data warehouse i.e. "data warehouse is a copy of transaction data specifically structured for query and analysis". This is a functional view of a data warehouse. Kimball did not address how the data warehouse is built like Inmondid, rather he focused on the functionality of a data warehouse.

Benefits of Data Warehousing

- **Potential high returns on investment and delivers enhanced business intelligence :** Implementation of data warehouse requires a huge investment in lakhs of rupees. But it helps the organization to take strategic decisions based on past historical data and organization can improve the results of various processes like marketing segmentation, inventory management and sales.
- **Competitive advantage :** As previously unknown and unavailable data is available in data warehouse, decision makers can access that data to take decisions to gain the competitive advantage.

15 16 17

- **Saves Time :** As the data from multiple sources is available in integrated form, business users can access data from one place. There is no need to retrieve the data from multiple sources.
- **Better enterprise intelligence :** It improves the customer service and productivity.
- **High quality data :** Data in data warehouse is cleaned and transferred into desired format. So data quality is high.

2.1.1 Features of Data Warehouse

Characteristics/ Features of a Data Warehouse

A common way of introducing data warehousing is to refer to the characteristics of a data warehouse :

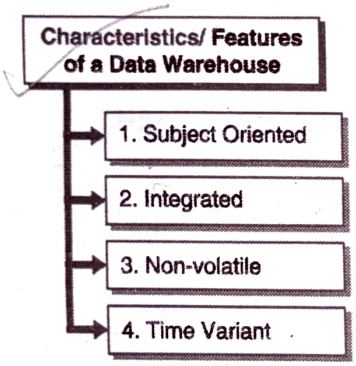


Fig. 2.1.1 : Characteristics/ Features of a Data Warehouse

1. Subject Oriented

- Data warehouses are designed to help analyze data. For example, to learn more about banking data, a warehouse can be built that concentrates on transactions, loans, etc.
- This warehouse can be used to answer questions like "Which customer has taken maximum loan amount for last year?" This ability to define a data warehouse by subject matter, loan in this case, makes the data warehouse subject oriented.

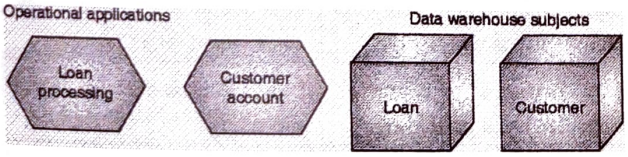


Fig. 2.1.2 : Data Warehouse is subject Oriented

2. Integrated

- A data warehouse is constructed by integrating multiple, heterogeneous data sources like, relational databases, flat files, on-line transaction records.
- The data collected is cleaned and then data integration techniques are applied, which ensures consistency in naming conventions, encoding structures, attribute measures etc. among different data sources.

Example

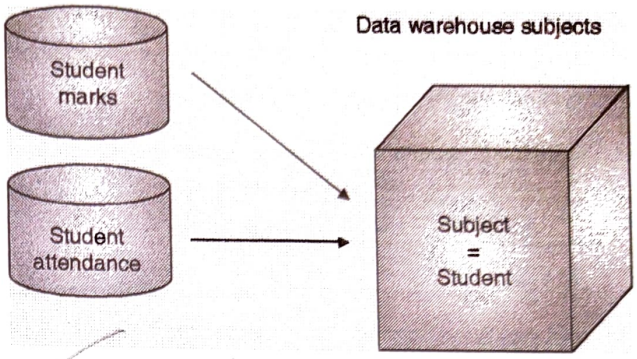


Fig. 2.1.3 : Integrated Data Warehouse

3. Non-volatile

Nonvolatile means that, once data entered into the warehouse, it cannot be removed or changed because the purpose of a warehouse is to analyze the data.

4. Time Variant

A data warehouse maintains historical data. For e.g. A customer record has details of his job, a data warehouse would maintain all his previous jobs (historical information) when compared to a transactional system which only maintains current job due to which its not possible to retrieve older records.

Syllabus Topic : Operational Database Systems and Data Warehouses (OLTP Vs OLAP)

2.2 Operational Database Systems and Data Warehouses (OLTP Vs OLAP)

2.2.1 Why are Operational Systems not Suitable for Providing Strategic Information?

- The fundamental reason for the inability to provide strategic information is that strategic information has been extracted from the existing operational systems.

- These operational systems such as University Record system, inventory management, claims processing, outpatient billing, and so on are not designed in a way to provide strategic information.
- If we need the strategic information, the information must be collected from altogether different types of systems. Only specially designed decision support systems or informational systems can provide strategic information.
- Operational systems are tuned for known transactions and workloads, while workload is not known a priori in a data warehouse.
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)e.g., average amount spent on phone calls between 9AM-5PM in Pune during the month of December.

Operational Database System	Data Warehouse (or DSS -Decision Support System)
Application oriented	Subject oriented
Used to run business	Used to analyze business
Detailed data	Summarized and refined
Current up to date	Snapshot data
Isolated data	Integrated data
Repetitive access	Ad-hoc access
Clerical user	Knowledge user (manager)
Performance sensitive	Performance relaxed
Few records accessed at a time (tens)	Large volumes accessed at a time (millions)
Read/update access	Mostly read (batch update)
No data redundancy	Redundancy present
Database size 100 MB-100 GB	Database size 100GB – few terabytes

2.2.2 OLAP Vs OLTP

- *OLAP* (On Line Analytical Processing) supports the *multidimensional* view of data.
- *OLAP* provides fast, steady, and proficient access to the *various views of information.*
- *The complex queries can be processed.*
- *It's easy to analyze information by processing complex queries on multidimensional views of data.*
- Data warehouse is generally used to analyse the information where huge amount of historical data is stored.
- Information in data warehouse is related to more than one dimension like sales, market trends, buying patterns, supplier, etc.

Definition

Definition given by OLAP council (www.olapcouncil.org) On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

Application Differences

OLTP (On Line Transaction Processing)	OLAP (On-Line Analytical Processing)
Transaction oriented	Subject oriented
High Create/Read/Update/ Delete (CRUD) activity	High Read activity
Many users	Few users
Continuous updates – many sources	Batch updates – single source
Real-time information	Historical information
Tactical decision-making	Strategic planning
Controlled, customized delivery	"Uncontrolled", generalized delivery
RDBMS	RDBMS and/or MDBMS
Operational database	Informational database

Modeling Objectives Differences

OLTP	OLAP
High transaction volumes using few records at a time.	Low transaction volumes using many records at a time.
Balancing needs of online v/s scheduled batch processing.	Design for on-demand online processing.
Highly volatile data.	Non-volatile data.
Data redundancy – *BAD.*	Data redundancy – *GOOD.*
Few levels of granularity.	Multiple levels of granularity.
Complex database designs used by IT personnel.	Simpler database designs with business-friendly constructs.

Model Differences

OLTP	OLAP
Single purpose model – supports Operational System.	Multiple models – support Informational Systems.
Full set of Enterprise data.	Subset of Enterprise data.
Eliminate redundancy.	Plan for redundancy.
Natural or surrogate keys.	Surrogate keys.
Validate Model against business Function Analysis.	Validate Model against reporting requirements.
Technical metadata depends on business requirements.	Technical metadata depends on data mapping results.
This moment in time is important.	Many moments in time are essential elements.

Syllabus Topic : A Multidimensional Data Model

2.3 A Multidimensional Data Model

2.3.1 What is Dimensional Modeling ?

- It is a logical design technique used for data warehouses.
- Dimensional model is the underlying data model used by many of the commercial OLAP products available today in the market.
- Dimensional model uses the relational model with some important restrictions.
- It is one of the most feasible technique for delivering data to the end users in a data warehouse.
- Every dimensional model is composed of at least one table with a multipart key called the *fact table* and a set of other related tables called *dimension tables.*

Syllabus Topic : Data Cubes

2.3.2 Data Cubes

- Multidimensional models is used to inhabit data in multi-dimensional matrices like Data Cubes or Hypercubes. A standard spreadsheet, signifying a conventional database, is a two-dimensional matrix. One example would be a spreadsheet of regional sales by product for a particular time period. Products sold with respect to region can be shown in 2 dimensional matrix but as one more dimension like time is added then it produces 3 dimensional matrix as shown in Fig. 2.3.1.

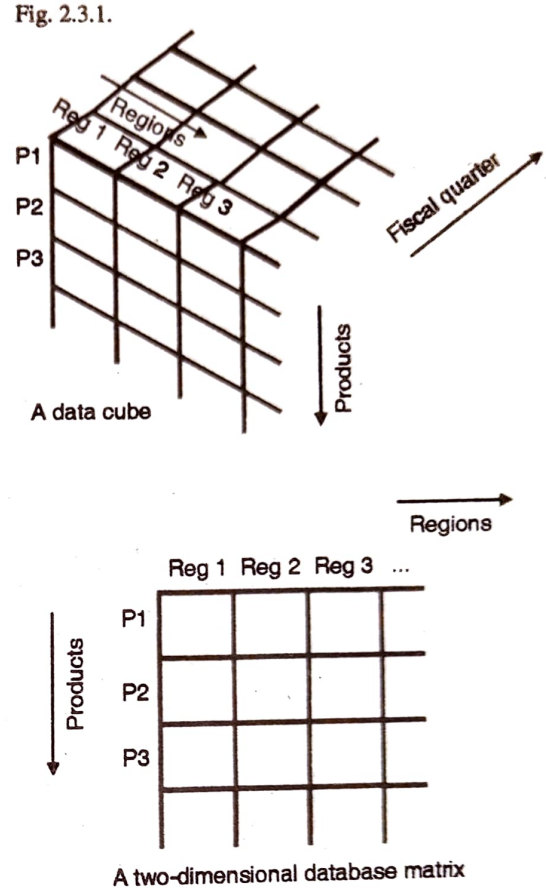


Fig. 2.3.1 : Pictorial view of data cube and 2D database

- A multidimensional model has two types of tables :
  1. **Dimension tables :** contains attributes of dimensions
  2. **Fact tables :** contains facts or measures

Syllabus Topic : Star Schema

2.3.3 Star Schema

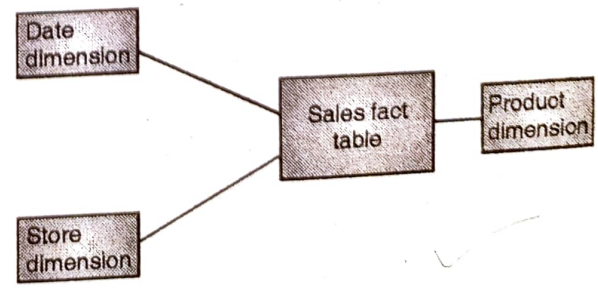


Fig. 2.3.2 : Examples of Star Schema

- Star Schema is the most popular schema design for a data warehouse.
- Dimensions are stored in a Dimension table and every entry has its own unique identifier.
- Every Dimension table is related to one or more fact tables.
- All the unique identifiers (primary keys) from the dimension tables make up for a composite key in the fact table.
- The fact table also contains facts. For example, a combination of store_id, date_key and product_id giving the amount of a certain product sold on a given day at a given store.
- Foreign keys for the dimension tables are contained in a fact table. For eg.(date key, product id and store_id) are all three foreign keys.
- In a dimensional modeling fact tables are normalised, whereas dimension tables are not.
- The size of the fact tables is large as compared to the dimension tables.
- The Facts in the star schema can be classified into three types.

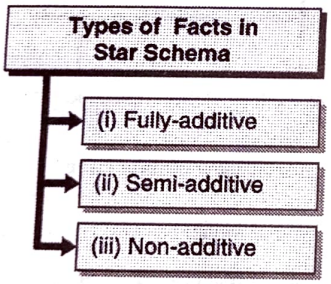


Fig. 2.3.3 : Types of Facts in Star Schema

➜ **(i) Fully-additive**

- Additive facts are facts that can be summed up through all of the dimensions in the fact table.

➜ **(ii) Semi-additive**

- Semi-additive facts are facts that can be summed up for some of the dimensions in the fact table, but not the others.

Example : Bank Balances : You can take a bank account as Semi- Additive since a current balance for the account can't be summed as time period; but if you want see current balance of a bank you can sum all accounts current balance.

➜ **(iii) Non-additive**

- Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table.

E.g. : Ratios, Averages and Variance

☞ **Advantages of Star Schema**

- A star schema describes aspects of a business. It is made up of multiple dimension tables and one fact table. For e.g. if you have a book selling business, some of the dimension tables would be customer, book, catalog and year. The fact table would contain information about the books that are ordered from each catalog by each customer during a particular year.
- Reduced Joins, Faster Query Operation.
- It is fully denormalized schema.
- Simplest DW schema.
- Easy to understand.
- Easy to Navigate between the tables due to less number of joins.
- Most suitable for Query processing.

Syllabus Topic : The Snowflake Schema

2.3.4 The Snowflake Schema

- A snowflake schema is used to remove the low cardinality i.e attributes having low distinct values, textual attributes from a dimension table and placing them in a secondary dimension table.
- For e.g. in Sales Schema, the product category in the product dimension table can be removed and placed in a secondary dimension table by normalizing the product dimension table. This process is carried out on large dimension tables.
- It is a normalization process carried out to manage the size of the dimension tables. But this may affect its performance as joins needs to be performed.
- In a star schema, if all the dimension tables are normalised then this schema is called as snowflake schema, and if only few of the dimensions in a star schema are normalised then it is called as star flake schema.

2.3.5 Star Flake Schema

- It is a hybrid structure (i.e. star schema + snowflake schema).
- Every fact points to one tuple in each of the dimensions and has additional attributes.
- Does not capture hierarchies directly.
- Straightforward means of capturing a multiple dimension data model using relations.

2.3.6 Differentiate between Star Schema and Snowflake Schema

Sr. No.	Star Schema	Snowflake Schema
1.	Star schema contains the dimension tables mapped around one or more fact tables.	A Snowflake schema contains in-depth joins because the tables are split in to many pieces.
2.	It is a de-normalized model.	It is the normalised form of Star schema.
3.	No need to use complicated joins.	Have to use complicated joins, since it has more tables.
4.	Queries results fast.	There will be some delay in processing the Query.
5.	Star Schemas are usually not in BCNF form. All the primary keys of the dimension tables are in the fact table.	In Snowflake schema, dimension tables are in 3NF, so there are more dimension tables which are linked by primary – foreign key relation.

2.3.7 Factless Fact Table

- Factless table means only the key available in the Fact there is no measures available.
- Used only to put relation between the elements of various dimensions.
- Are useful to describe events and coverage, i.e. the tables contain information that something has/has not happened.
- Often used to represent many-to-many relationships.
- The only thing they contain is a concatenated key, they do still however represent a focal event which is identified by the combination of conditions referenced in the dimension tables.
- An Example of Factless fact table can be seen in the Fig. 2.3.4.

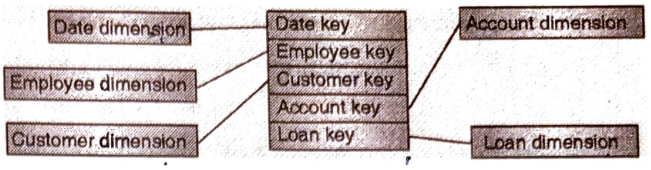


Fig. 2.3.4 : A Factless Fact Table

- There are two main types of factless fact tables :

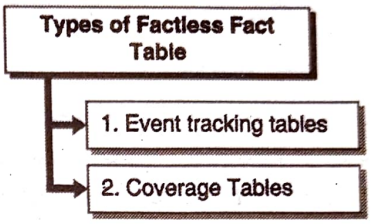


Fig. 2.3.5 : Types of Factless Fact Table

➜ **1. Event tracking tables**

- Use a factless fact table to track events of interest to the organization. For example, attendance at a cultural event can be tracked by creating a fact table containing the following foreign keys (i.e. links to dimension tables): event identifier, speaker/entertainer identifier, participant identifier, event type, date. This table can then be queried to find out information, such as which cultural events or event types are the most popular.
- Following example shows factless fact table which records every time a student attends a course or Which class has the maximum attendance? Or What is the average number of attendance of a given course?
- All the queries are based on the COUNT() with the GROUP BY queries. So we can first count and then apply other aggregate functions such as AVERAGE, MAX, MIN.

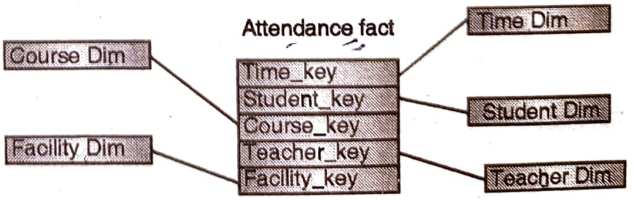


Fig. 2.3.6 : Example of Event Tracking Tables

2. Coverage Tables

The other type of factless fact table is called Coverage table by Ralph. It is used to support negative analysis report. For example a Store that did not sell a product for a given period. To produce such report, you need to have a fact table to capture all the possible combinations. You can then figure out what is missing.

Common examples of factless fact table

- Ex-Visitors to the office.
- List of people for the web click.
- Tracking student attendance or registration events.

Syllabus Topic : Fact Constellation Schema or Families of Star

2.3.8 Fact Constellation Schema or Families of Star

Fact Constellation

- As its name implies, it is shaped like a constellation of stars (i.e., star schemas).
- This schema is more complex than star or snowflake varieties, which is due to the fact that it contains multiple fact tables.
- This allows dimension tables to be shared amongst the fact tables.
- A schema of this type should only be used for applications that need a high level of sophistication.
- For each star schema or snowflake schema it is possible to construct a fact constellation schema.
- That solution is very flexible, however it may be hard to manage and support.
- The main disadvantage of the fact constellation schema is a more complicated design because many variants of aggregation must be considered.
- In a fact constellation schema, different fact tables are explicitly assigned to the dimensions, which are for given facts relevant.
- This may be useful in cases when some facts are associated with a given dimension level and other facts with a deeper dimension level.
- Use of that model should be reasonable when for example, there is a sales fact table (with details down to the exact date and invoice header id) and a fact table with sales forecast which is calculated based on month, client id and product id.
- In that case using two different fact tables on a different level of grouping is realized through a fact constellation model.

Family of stars

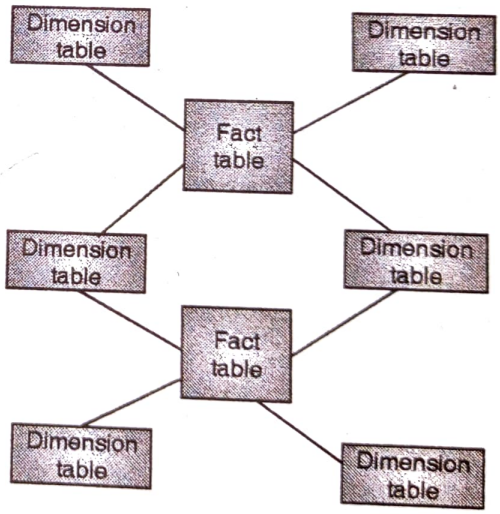


Fig. 2.3.7 : Family of stars

2.3.9 Examples on Star Schema and Snowflake Schema

Ex. 2.3.1 : All electronics company have sales department. Sales consider four dimensions namely time, item, branch and location. The schema contains a central fact table sales with two measures dollars_sold and unit_sold. Design star schema, snowflake schema and fact constellation for same.

Soln. :

(a) Star Schema

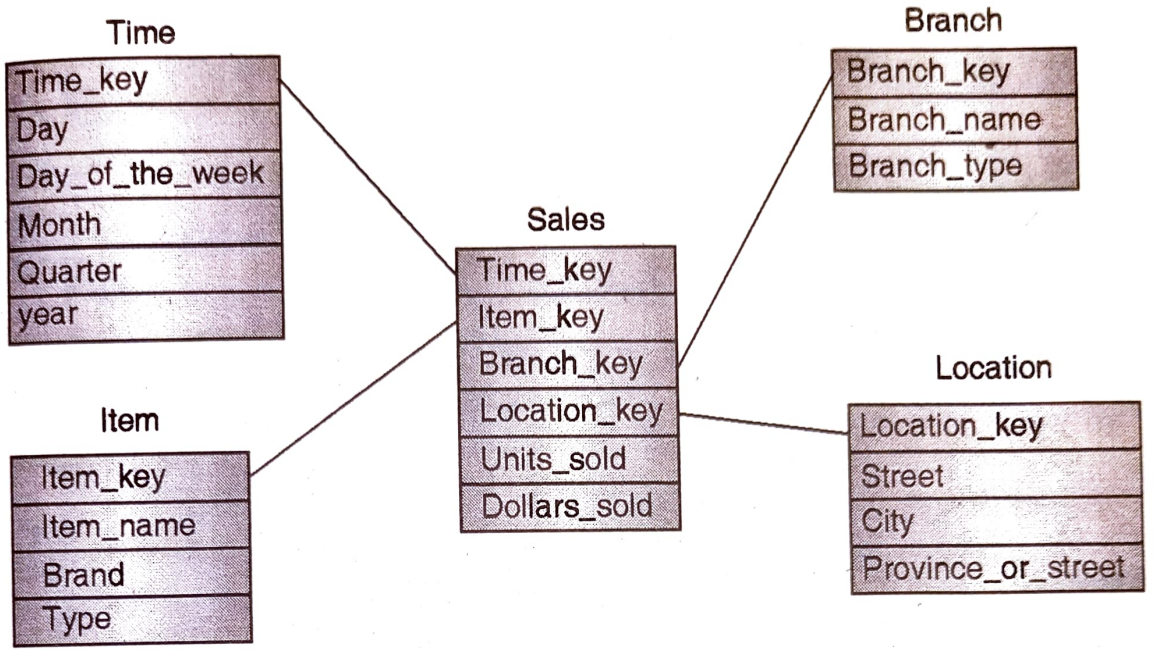


Fig. P. 2.3.1 : Sales Star Schema

(b) Snowflake Schema

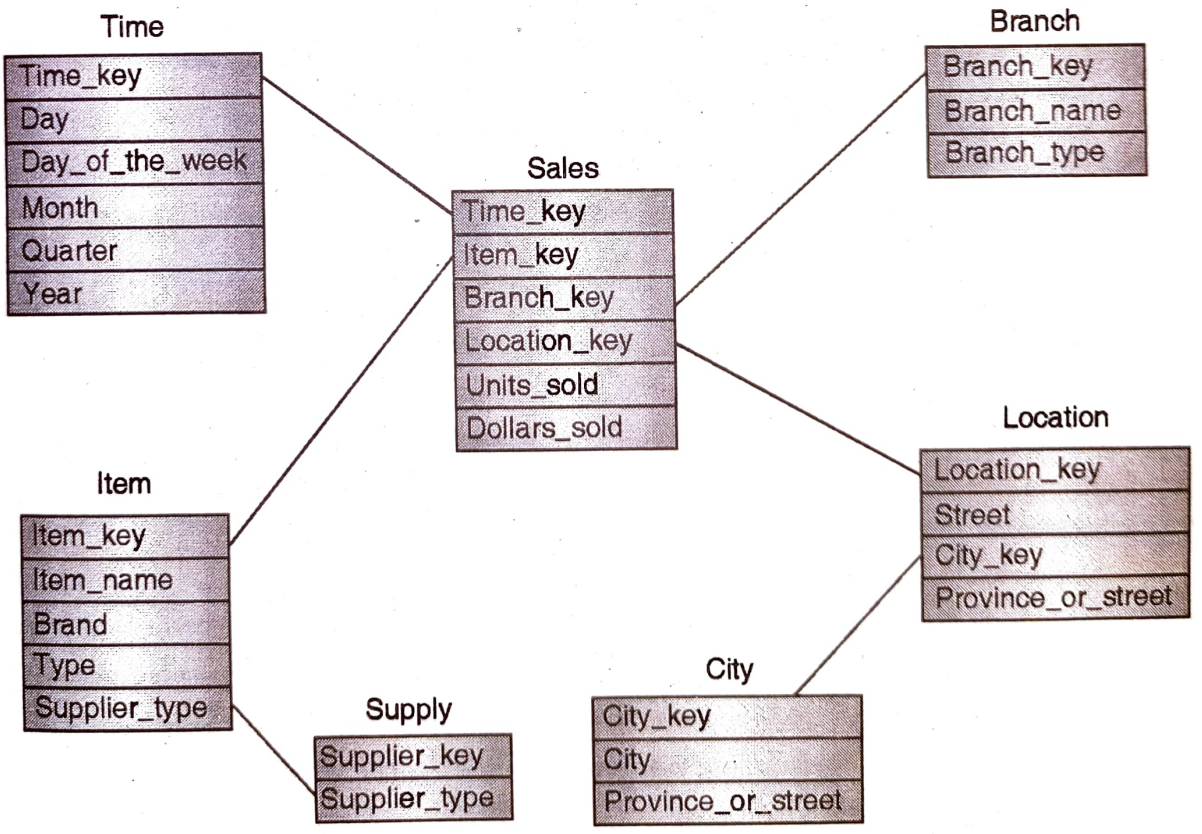


Fig. P. 2.3.1(a) : Sales Snowflake Schema

(c) Fact Constellation

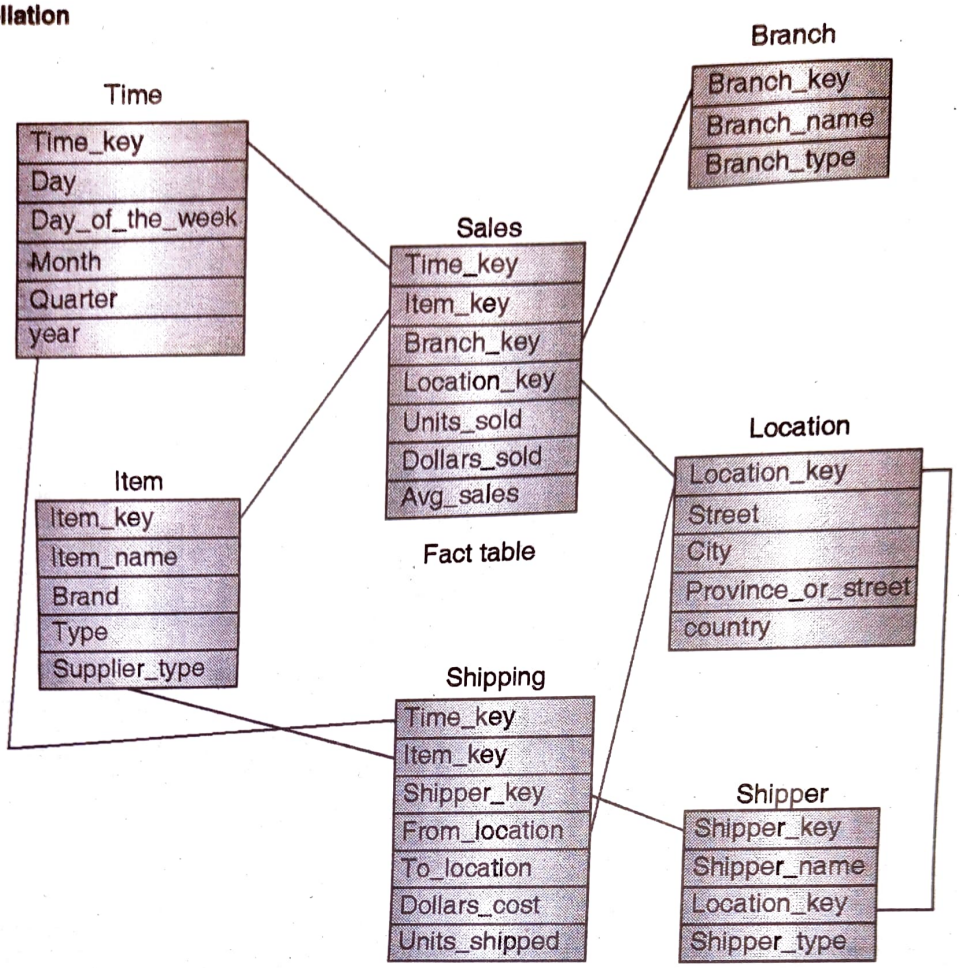


Fig. P. 2.3.1(b) : Fact constellation for sales

Ex. 2.3.2 : The Mumbai university wants you to help design a star schema to record grades for course completed by students. There are four dimensional tables namely course_section, professor, student, period with attributes as follows :

Course_section Attributes : Course_Id, Section_number, Course_name, Units, Room_id, Roomcapacity. During a given semester the college offers an average of 500 course sections

Professor Attributes : Prof_id, Prof_Name, Title, Department_id, department_name

Student Attributes : Student_id, Student_name, Major. Each Course section has an average of 60 students

Period Attributes : Semester_id, Year. The database will contain Data for 30 months periods. The only fact that is to be recorded in the fact table is course Grade

Answer the following Questions

(a) Design the star schema for this problem

(b) Estimate the number of rows in the fact table, using the assumptions stated above and also estimate the total size of the fact table (in bytes) assuming that each field has an average of 5 bytes.

(c) Can you convert this star schema to a snowflake schema ? Justify your answer and design a snowflake schema if it is possible.

Soln. :

(a) Star Schema

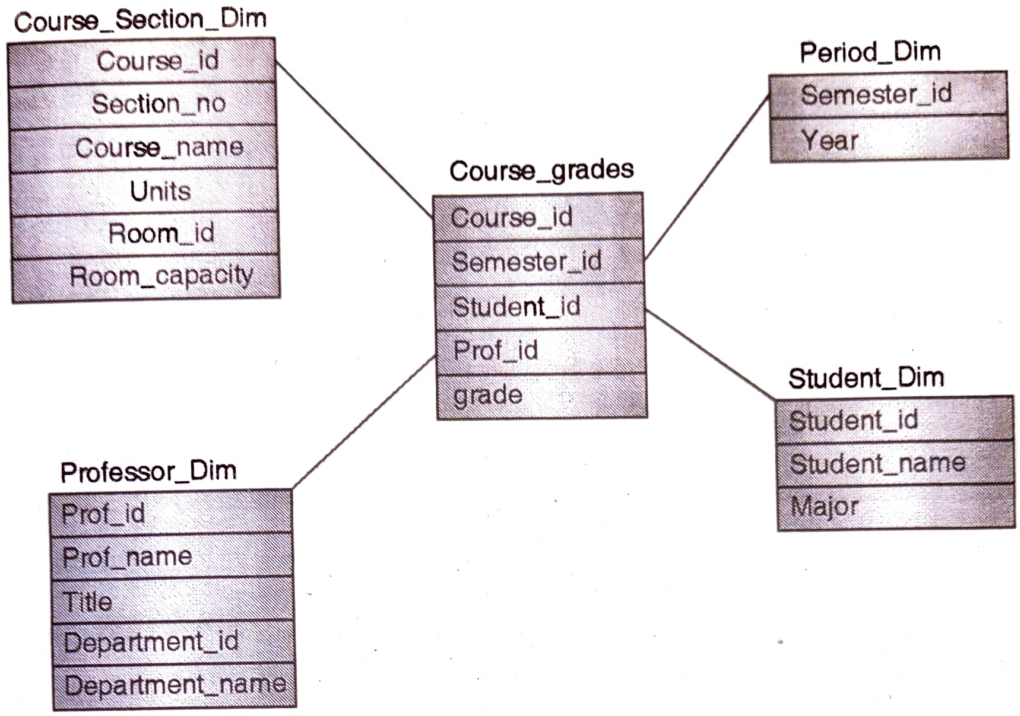


Fig. P. 2.3.2 : University Star Schema

(b) Total Courses Conducted by university = 500

Each Course has average students = 60

University stores data for 30 months

Total Student in University for all courses in 30 months = 500*60 = 30000

Time Dimension = 30 months = 5 Semesters (Assume 1 semester = 6 months)

Now, Number of rows of fact table = 30000*5 = 150000 (one student has 5 grades for 5 semesters)

(c) Snowflake Schema

- Yes, the above star schema can be converted to a snowflake schema, considering the following assumptions
- Courses are conducted in different rooms, so course dimension can be further normalized to rooms dimension as shown in the Fig. P. 2.3.2(a).
- Professor belongs to a department, and department dimension is not added in the star schema, so professor dimension can be further normalized to department dimension.
- Similarly students can have different major subjects, so it can also be normalized as shown in the Fig. P. 2.3.2(a).

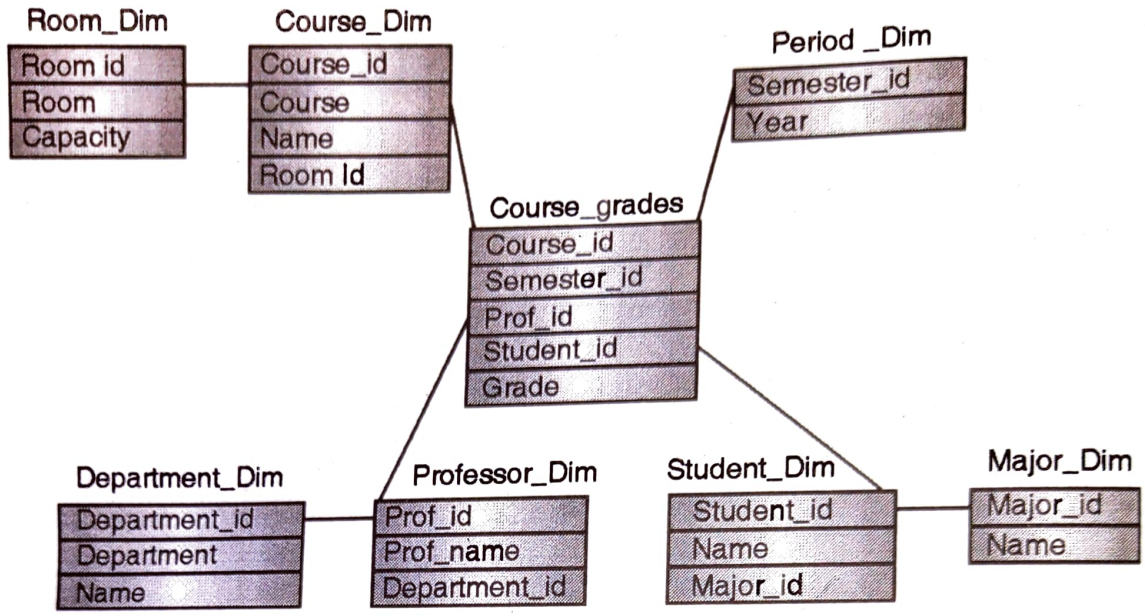


Fig. P. 2.3.2(a) : University Snowflake Schema

Ex. 2.3.3 : Give Information Package for recording information requirements for "Hotel Occupancy" considering dimensions like Time, Hotel etc. Design star schema from the information package.

Soln. :

Information package diagram

- Information package diagram is the approach to determine the requirement of data warehouse.
- It gives the metrics which specifies the business units and business dimensions.
- The information package diagram defines the relationship between the subject or dimension matter and key performance measures (facts).
- The information package diagram shows the details that users want so its effective for communication between the user and technical staff.

Table P. 2.3.3 : Information Package for Hotel Occupancy

Hotel	Room Type	Time	Room Status
Hotel Id	Room id	Time id	Status id
Branch Name	room type	Year	Status Description
Branch Code	room size	Quarter	
Region	number of beds	Month	
Address	type of bed	Date	
city/stat/zip	max occupants	day of week	
construction year	Suite	day of month,	
renovation year		holiday flag	

Facts

(a) Occupied Rooms

(b) Vacant Rooms

(c) Unavailable Rooms

(d) No of occupants

(e) Revenue

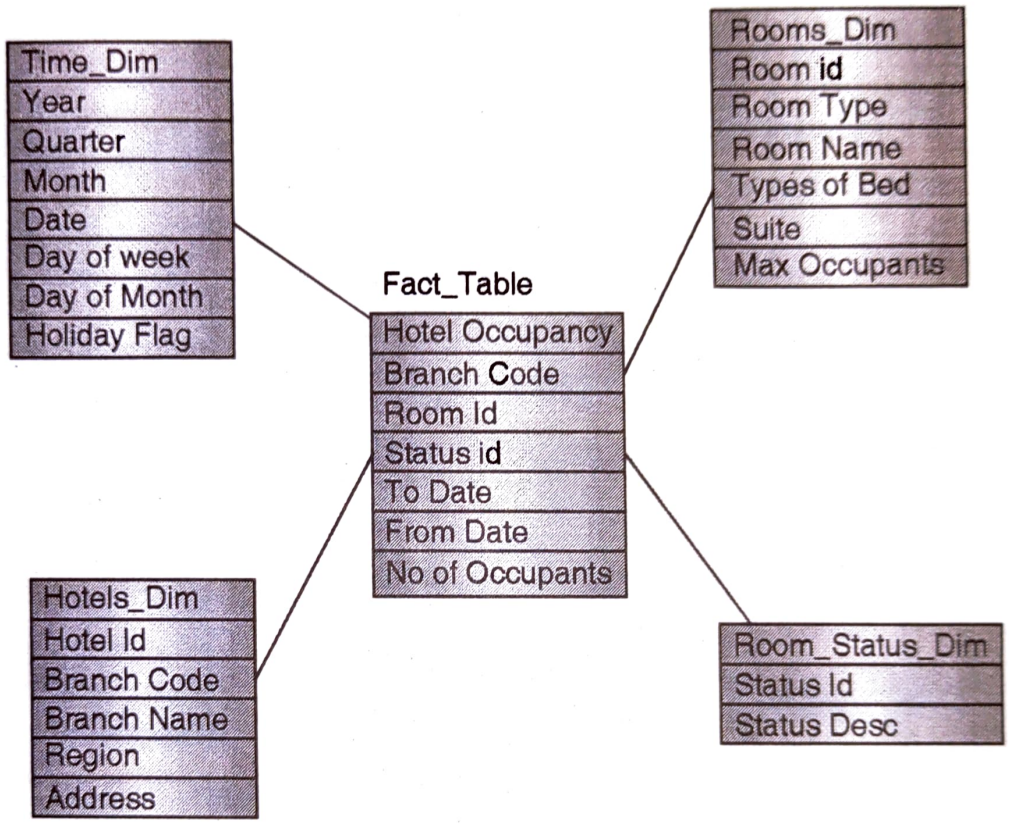


Fig. P. 2.3.3 : Hotel Occupancy Star Schema

Ex. 2.3.4 : For a Supermarket Chain consider the following dimensions, namely Product, store, time , promotion. The schema contains a central fact tables sales facts with three measures unit_sales, dollars_sales and dollar_cost.

Design star schema and calculate the maximum number of base fact table records for the values given below :

Time period : 5 years

Store : 300 stores reporting daily sales

Product : 40,000 products in each store(about 4000 sell in each store daily)

Promotion : a sold item may be in only one promotion in a store on a given day

---

Soln. :

(a) Star schema

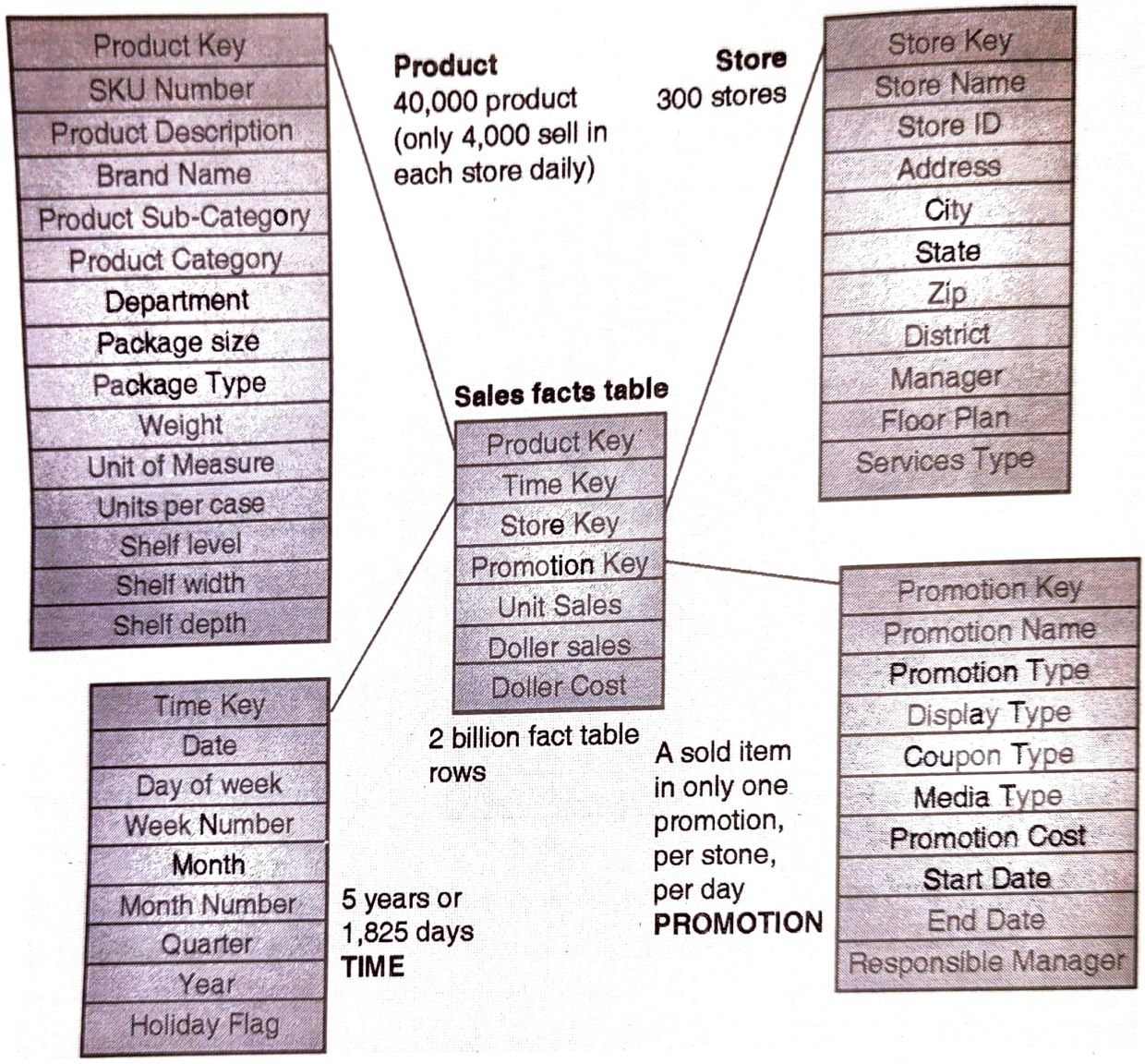


Fig. P. 2.3.4 : Sales Promotion Star Schema

(b) Time period = 5 years × 365 days = 1825

There are 300 stores.

Each stores daily sale = 4000 ; Promotion = 1

Maximum number of fact table records : 1825 × 300 × 4000 × 1= 2 billion

---

Ex. 2.3.5 : Draw a Star Schema for Student academic fact database.

Soln. :

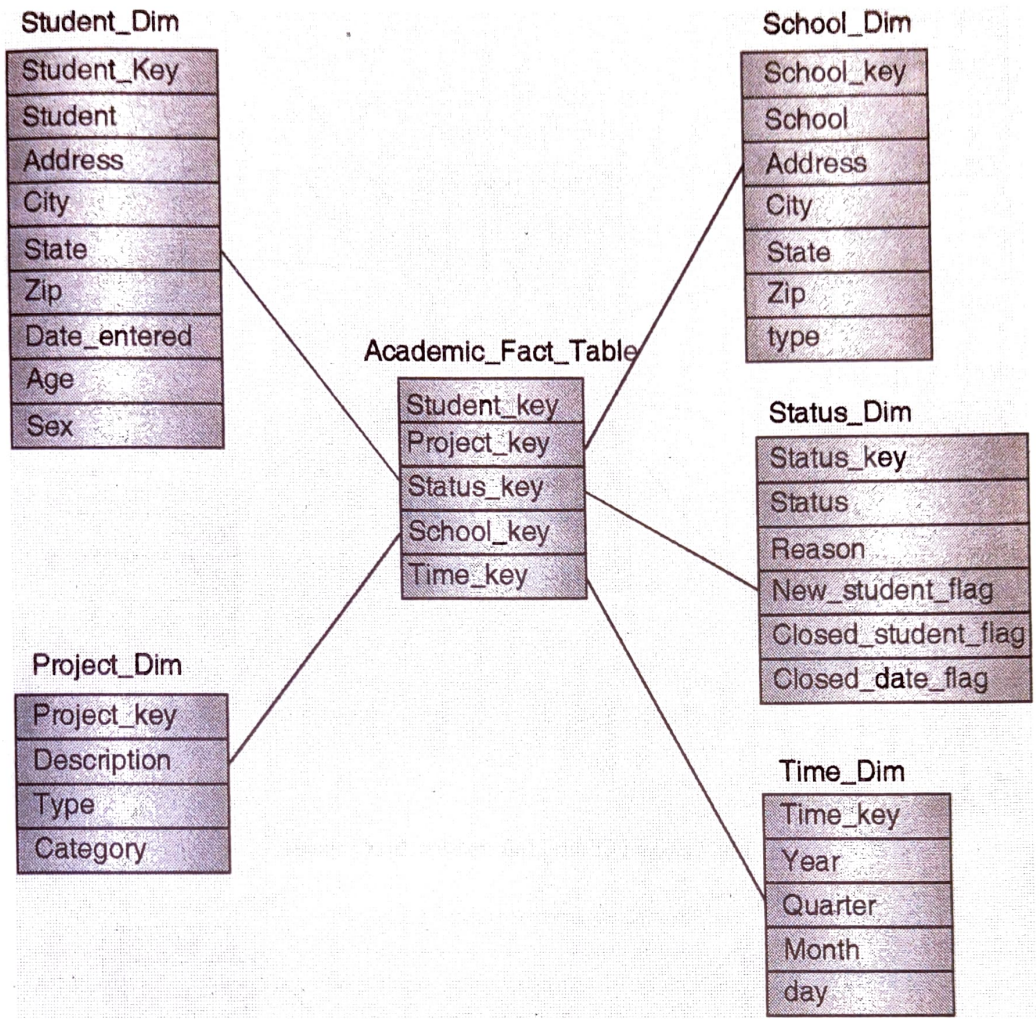


Fig. P. 2.3.5 : Student Academic Star Schema

Ex. 2.3.6 : List the dimensions and facts for the Clinical Information System and Design Star and Snow Flake Schema.

Soln. :

Dimensions

1. Patient
2. Doctor
3. Procedure
4. Diagnose
5. Date of Service
6. Location
7. Provider

Facts

1. Adjustment
2. Charge
3. Age

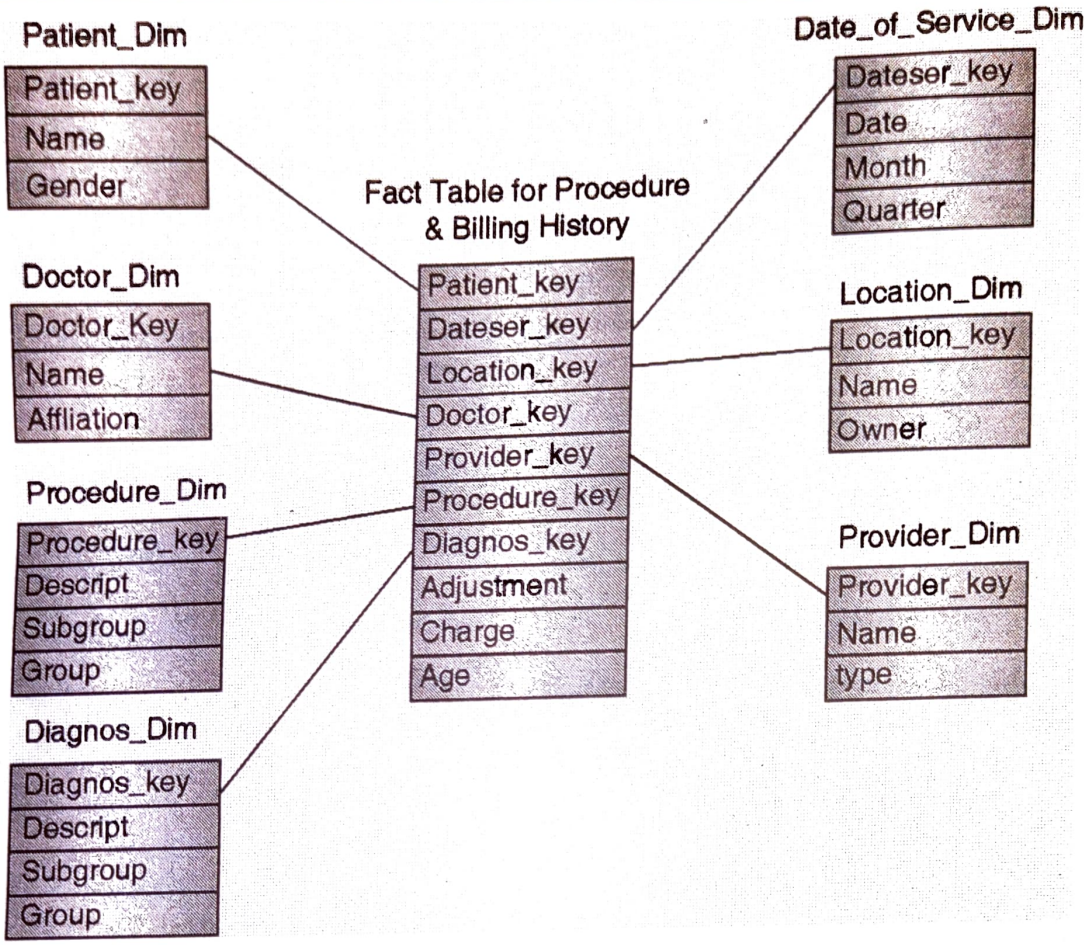


Fig. P. 2.3.6 : Clinical Information Star Schema

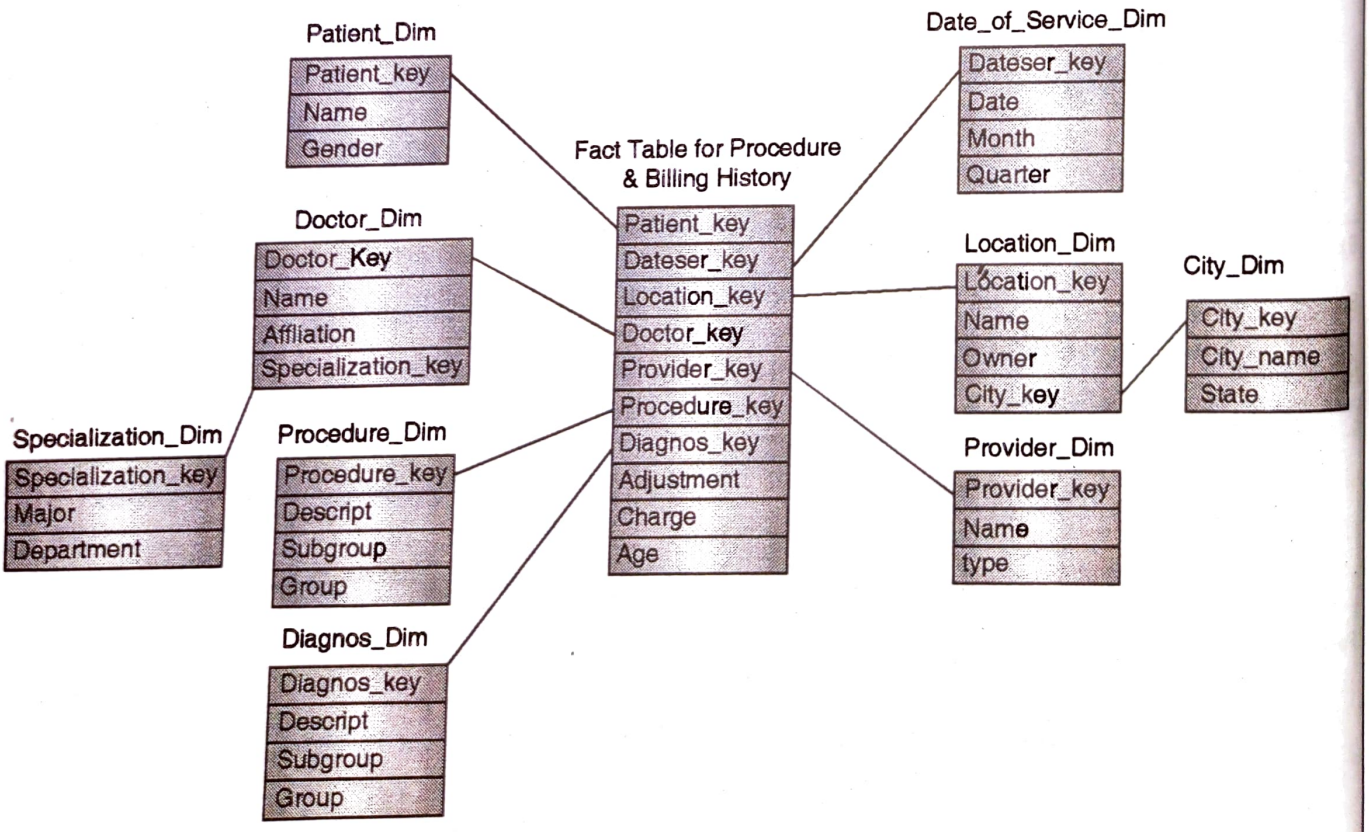


Fig. P. 2.3.6(a) : Clinical Information Snow Flake Schema

Ex. 2.3.7 : Draw a Star Schema for Library Management.

Soln. :

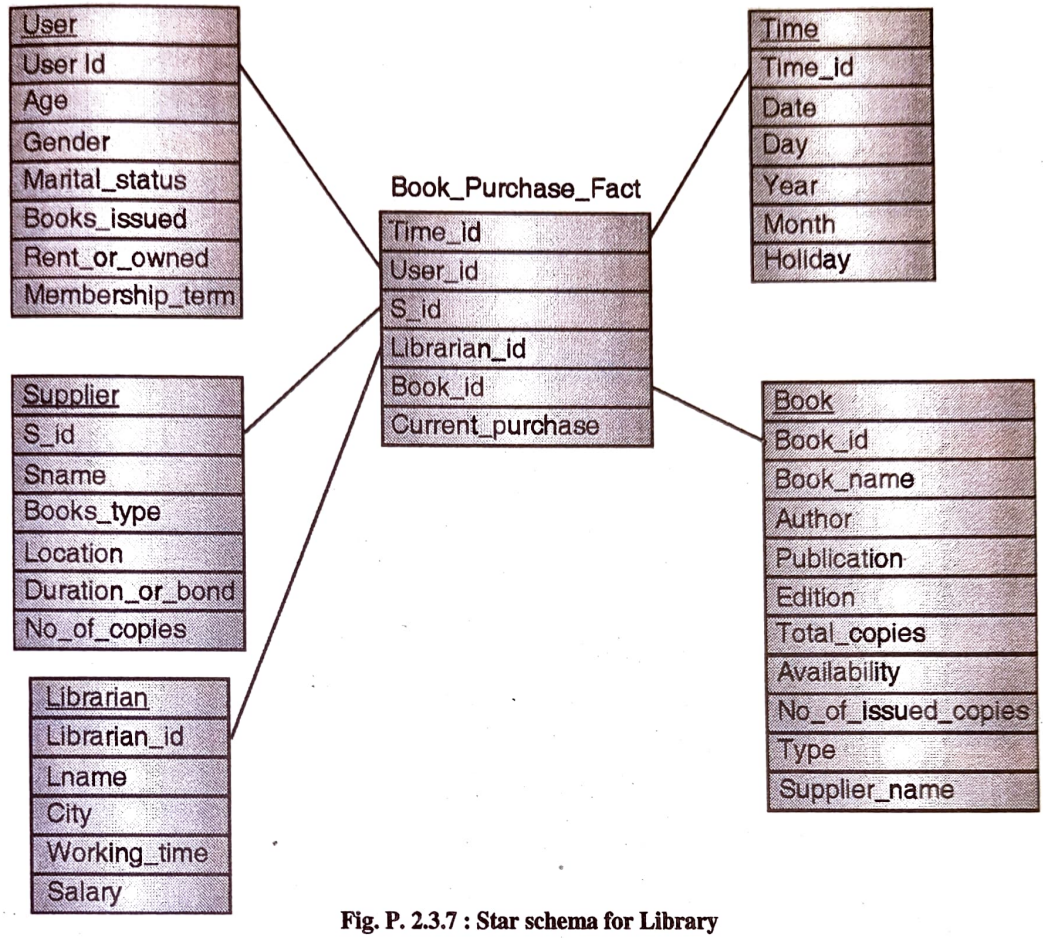


Fig. P. 2.3.7 : Star schema for Library

Ex. 2.3.8 : A manufacturing company has a huge sales network. To control the sales, it is divided in the regions. Each region has multiple zones. Each zone has different cities. Each sales person is allocated different cities. The object is to track sales figure at different granularity levels of region. Also to count no. Of products sold. Create data warehouse schema to take into consideration of above granularity levels for region, sales person and the quarterly, yearly and monthly sales.

Soln. :

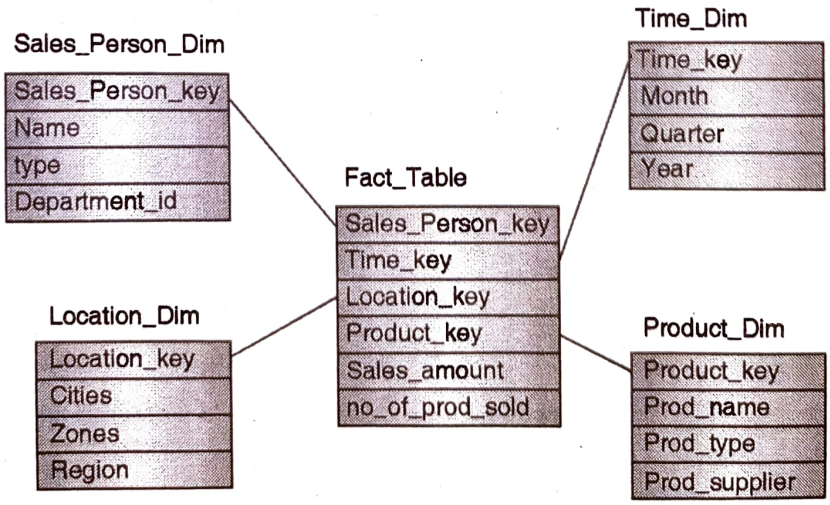


Fig. P. 2.3.8 : Star schema for Sales

Ex. 2.3.9 : A bank wants to develop a data warehouse for effective decision-making about their loan schemes. The bank provides loans to customers for various purposes like House Building loan, car loan, educational loan, personal loan etc. The whole country is categorized into a number of regions, namely, North, South, East, West. Each region consists of a set of states; loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. That data warehouse should record an entry for each disbursement of loan to customer. With respect to the above business scenario.

(i) Design an information package diagram. Clearly explain all aspects of the diagram.

(ii) Draw a star schema for the data warehouse clearly identifying the fact tables, dimension tables, their attributes and measures.

Soln. : (i)

Time	Customer	Branch	Location
Time_key	Customer_key	Branch_key	Location_key
Day	Account_number	Branch_Area	Region
Day_of_week	Account_type	Branch_home	State
Month	Loan_type		City
Quarter			Street
Year			
Holiday_flag			

***(ii)* Star Schema for a Bank**

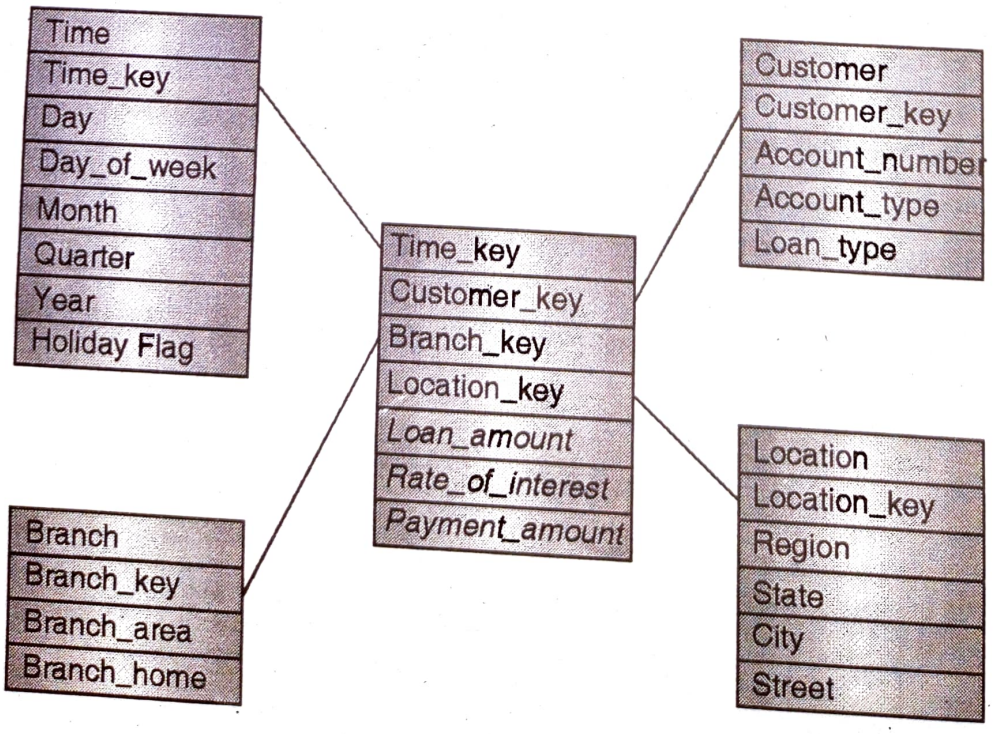
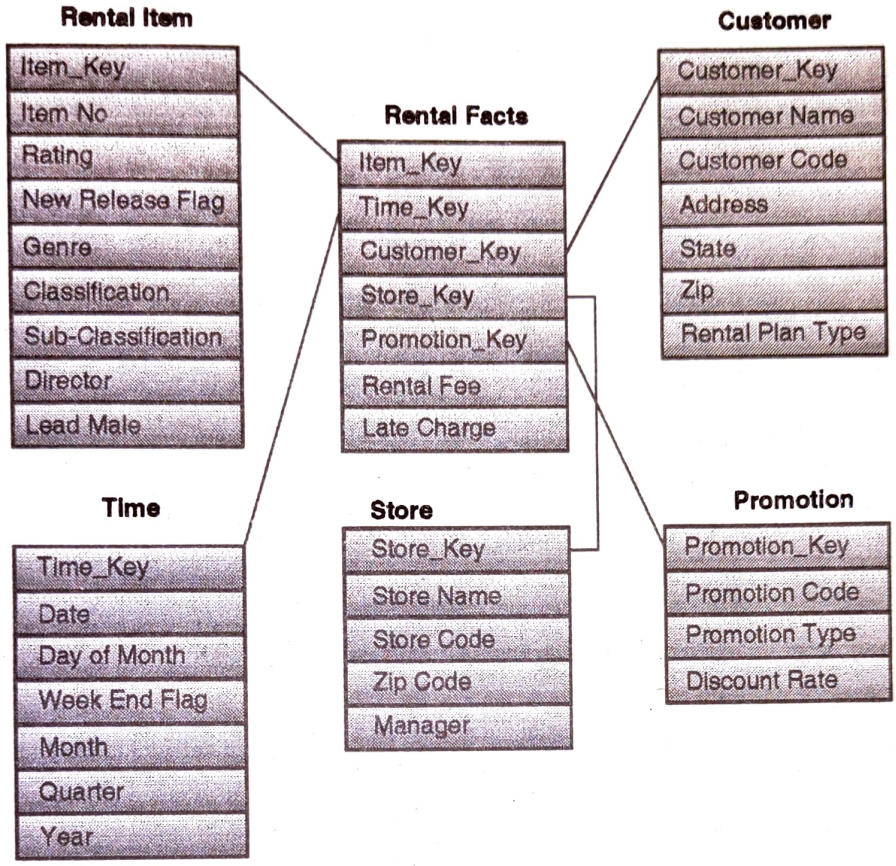


Fig. P. 2.3.9 : Star schema for Bank

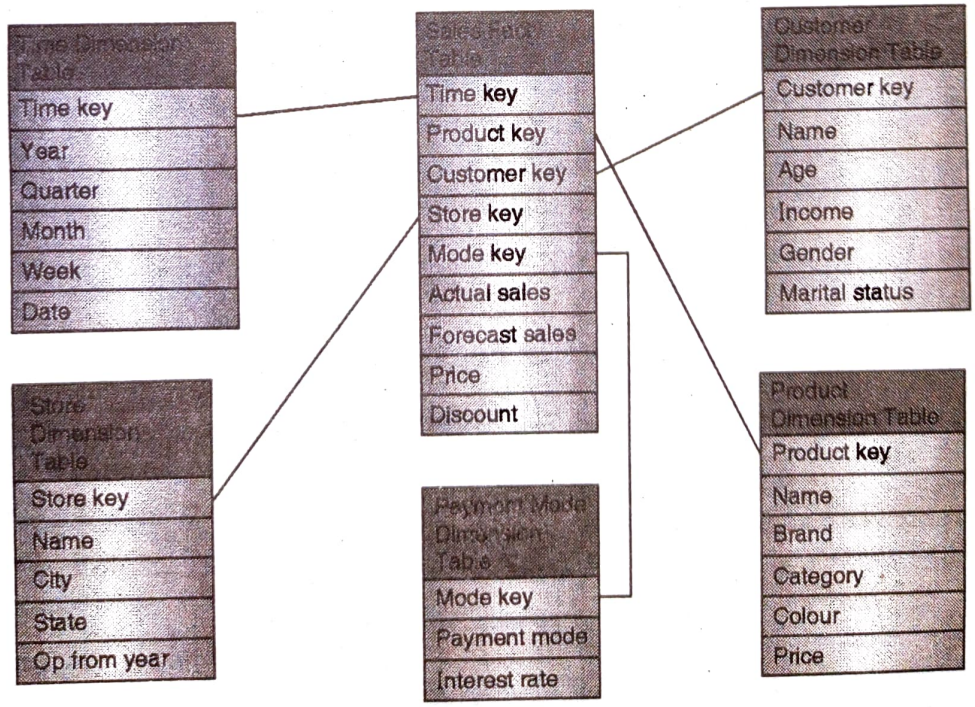
Ex.2.3.10 : Draw star schema for video Rental.

Soln. :



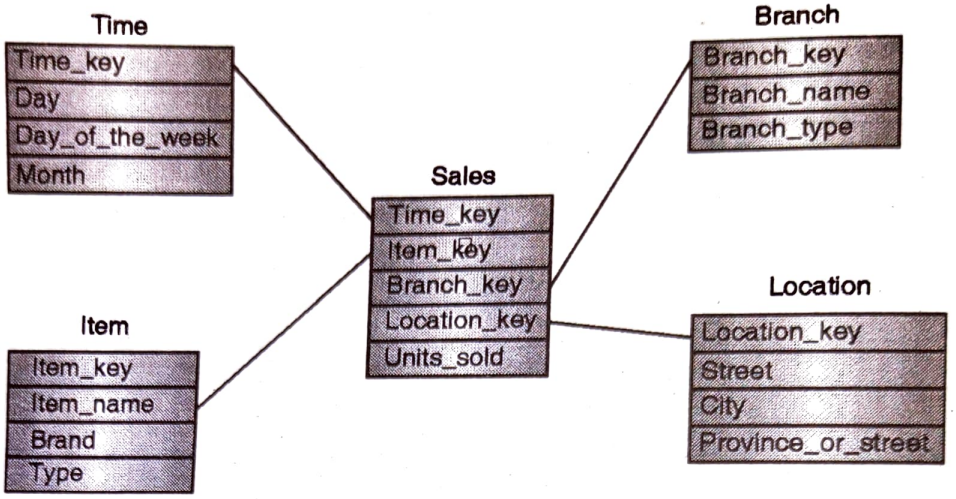
Ex. 2.3.11 : Draw star schema for retail chain.

Soln. :



Ex. 2.3.12 : Design Star schema for autosales analysis of company.

Soln. :



Ex. 2.3.13 : Consider the following database for a chain of bookstores.

BOOKS (Booknum, Primary_Author, Topic, Total_Stock, Price)

BOOKSTORE (Storenum, City, State, Zip, inventory_Value)

STOCK (Storenum, Booknum, QTY)

With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.

(a) Design an information package diagram.

(b) Design a star schema for the data warehouse clearly identifying the fact tables(s), Dimension table(s) their attributes and measures.

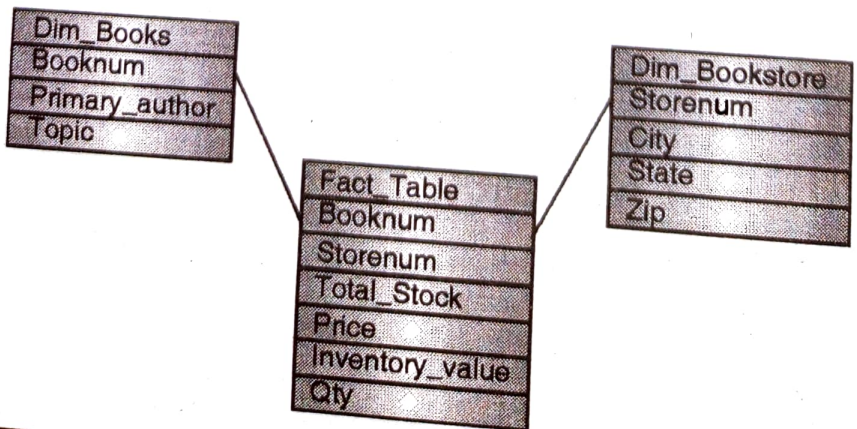
Soln. :

a) Information Package Diagram

BOOKS	BOOKSTORE
Booknum	Storenum
Primary_author	City
Topic	State
	Zip

Facts : Total_Stock , Price, Inventory_value , Qty

b) Star Schema



Ex. 2.3.14 : One of India's large retail departmental chains, with annual revenues touching $2.5 billion mark and having over 3600 employees working at diverse locations, was keenly interested in a business intelligence solution that can bring clear insights on operations and performance of departmental stores across the retail chain. The company needed to support a data warehouse that exceeds daily sales data from Point of Sales (POS) across all locations, with 80 million rows and 71 columns.

(a) List the dimensions and facts for above application.

(b) Design star schema and snow flake schema for the above application.

Soln. :

a) **Dimensions :** Product, Store, Time, Location

Facts : Unit Sales, Dollar Sales, Dollar Cost

b) **Star Schema and snowflake Schema**

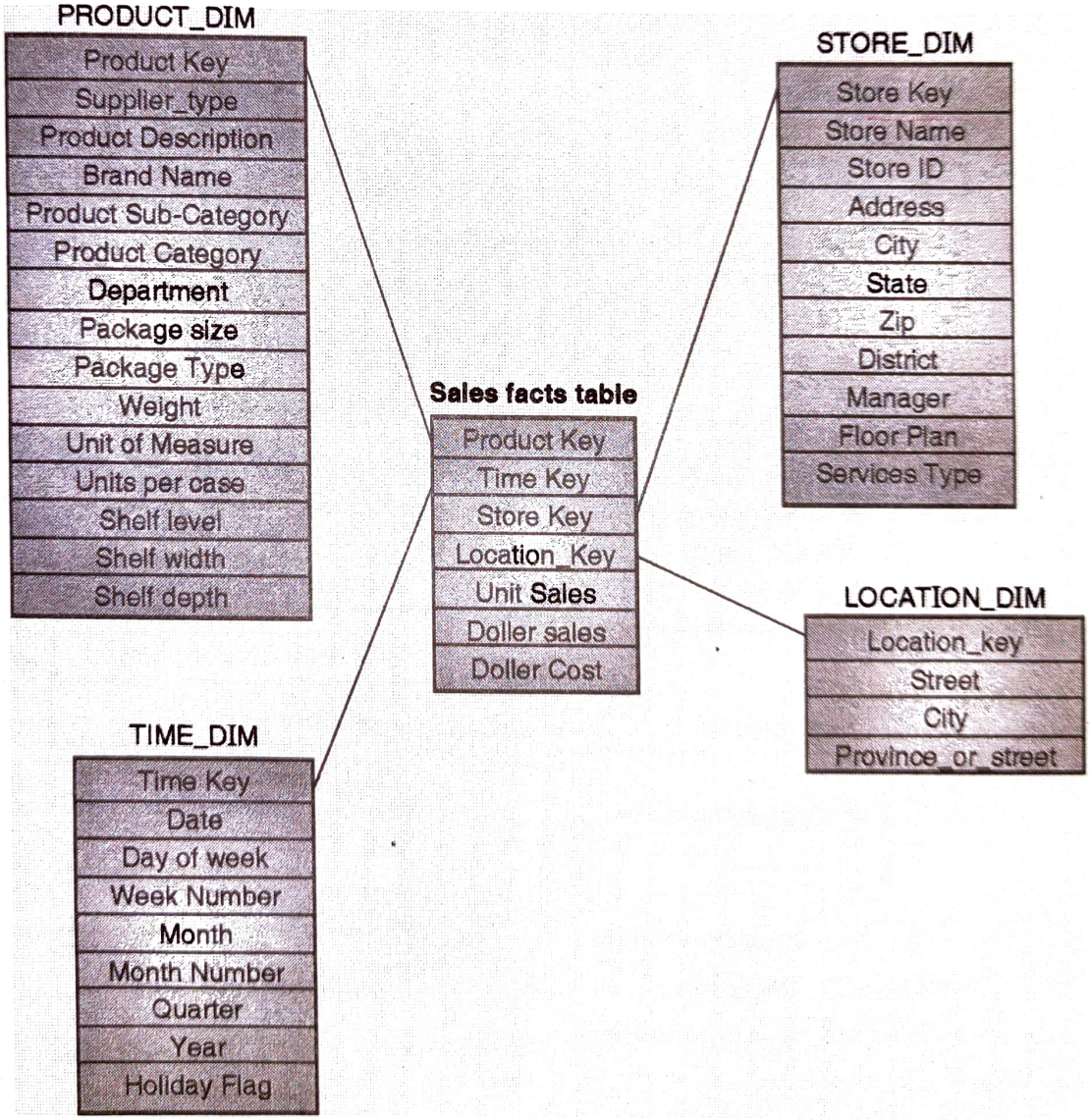


Fig. P. 2.3.14 : Star Schema

Fig. P. 2.3.14(a) : Snowflake Schema

Syllabus Topic : OLAP Operations in the Multidimensional Data Model

2.4 OLAP Operations in the Multidimensional Data Model

The following techniques are used for OLAP implementation

Example

Let us consider a company of Electronic Products. Data cube of company consists of 3 dimensions Location (aggregated with respect to city), Time (is aggregated with respect to quarters) and item (aggregated with respect to item types).

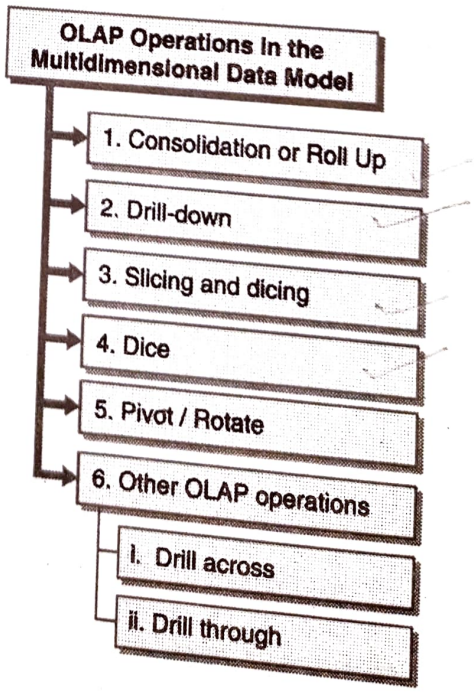


Fig. 2.4.1 : OLAP Operations in the Multidimensional Data Model

1. Consolidation or Roll Up

- Multi-dimensional databases generally have hierarchies with respect to dimensions.
- Consolidation is rolling up or adding data relationship with respect to one or more dimensions.
- For example, adding up all product sales to get total City data.
- For example, Fig. 2.4.2 shows the result of roll up operation performed on the central cube by climbing up the concept hierarchy for location.
- This hierarchy was defined as the total order street <city <province_or_state<country.
- The roll up operation shown aggregates the data by city to the country by location hierarchy.

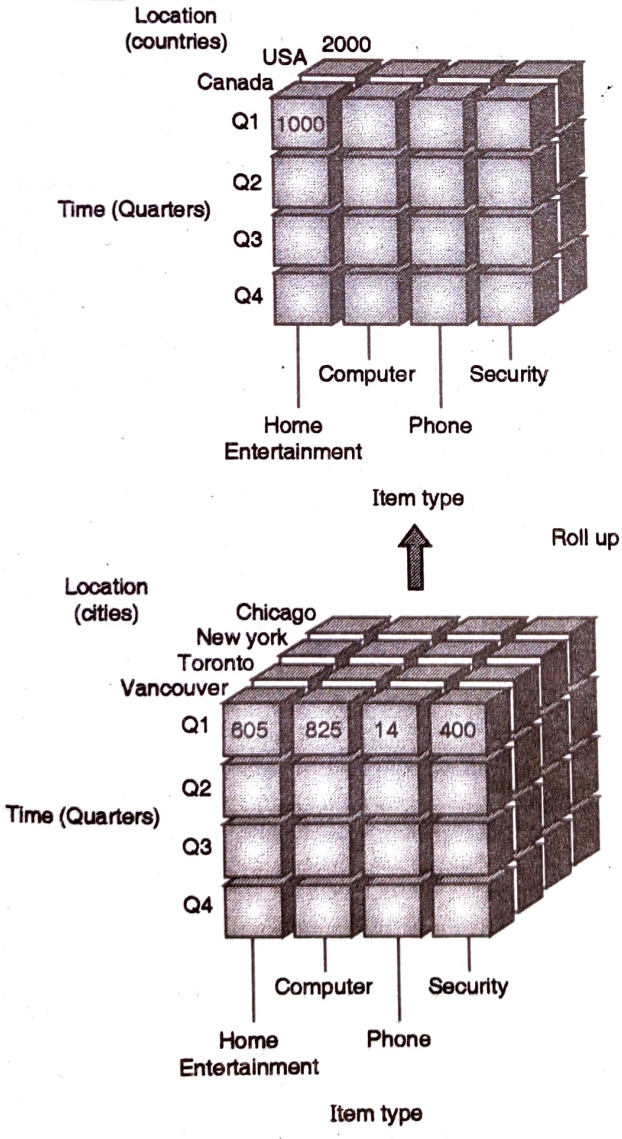


Fig. 2.4.2 : Roll -up or drill up

2. Drill-down

- Drill Down is defined as changing the view of data to a greater level of detail.
- For example, the Fig. 2.4.3 shows the result of drill down operations performed on the upper cube by stepping down a concept hierarchy for time defined as day<month<quarter<year.

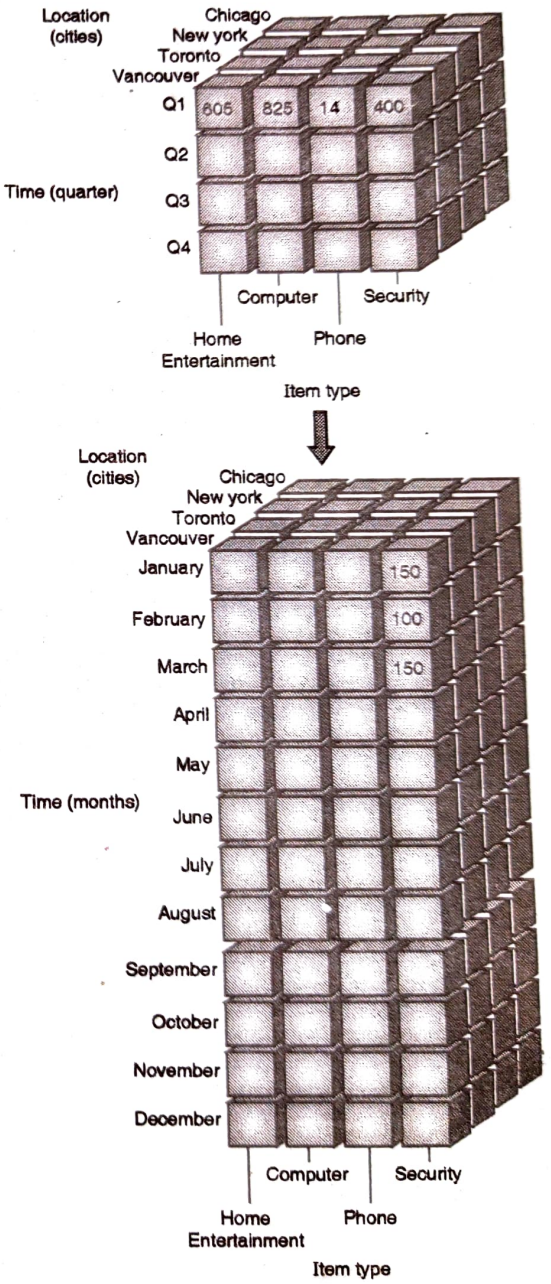


Fig. 2.4.3 : Drill Down

3. Slicing and dicing

- Slicing and dicing refers to the ability to look at a database from various viewpoints.
- Slice operation carry out selection with respect to one dimension of the given cube and produces a sub cube.
- For example, Fig. 2.4.4 shows the slice operation where the sales data are selected from the left cube for the dimension time using the criterion time = "Q1"

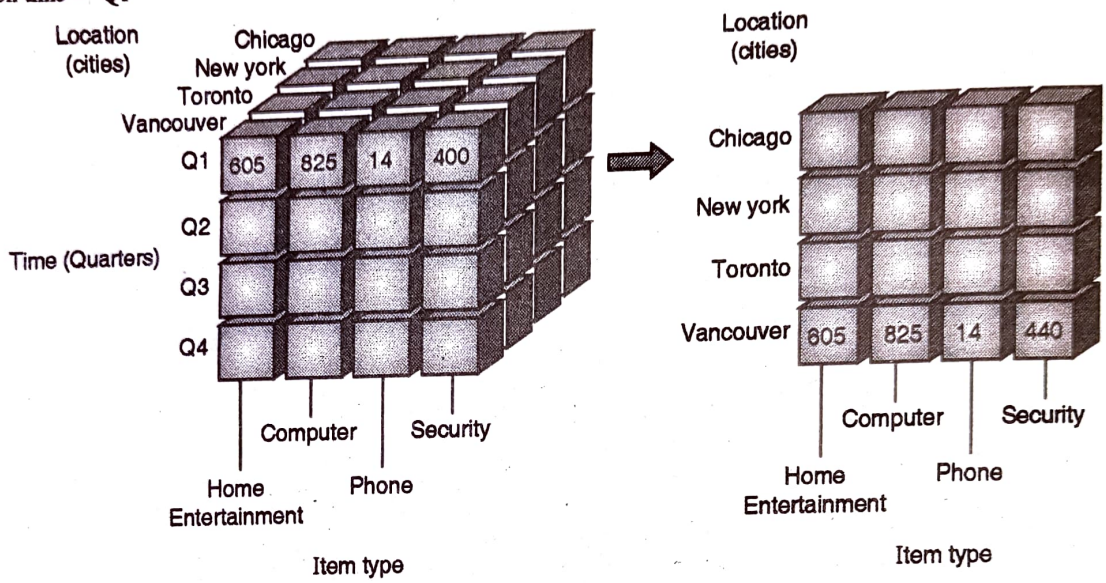


Fig. 2.4.4 : Slice operation

4. Dice

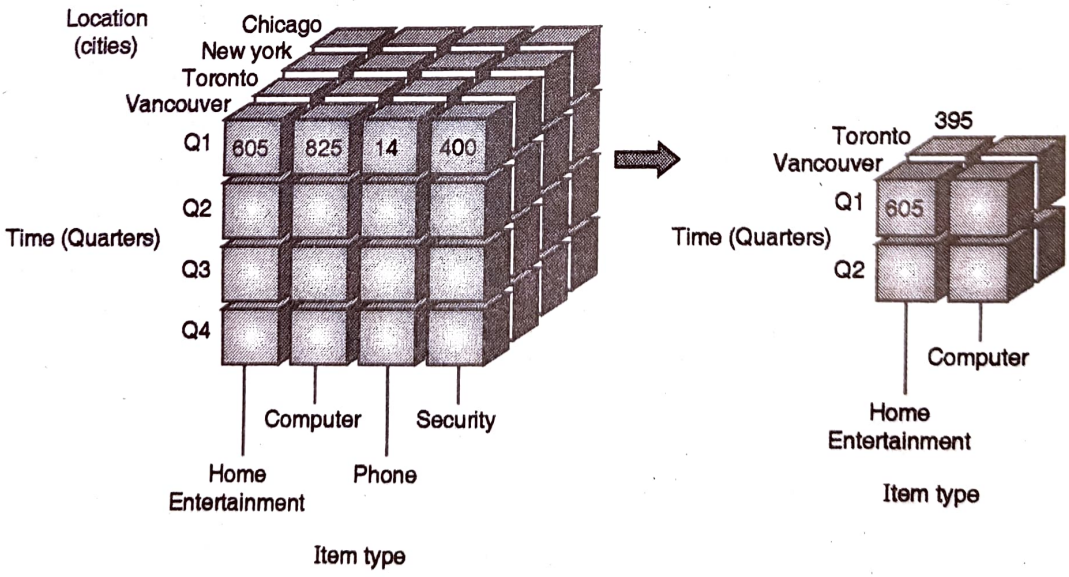


Fig. 2.4.5 : Dice operation

- Dice operation carry out selection with respect to two or more dimensions of the given cube and produces a sub cube.
- For example, the dice operation is performed on the left cube based on three dimension as Location, Time and Item as shown in Fig. 2.4.5 where the criteria is (location= "Torrento" or "Vancouver") and (time = "Q1" or "Q2") and (item= "home entertainment" or "computer").

5. Pivot / Rotate

- Pivot technique is used for visualization of data. This operation rotates the data axis to give another presentation of the data.
- For example Fig. 2.4.6 shows the pivot operation where the item and location axis in a 2-D slice are rotated.

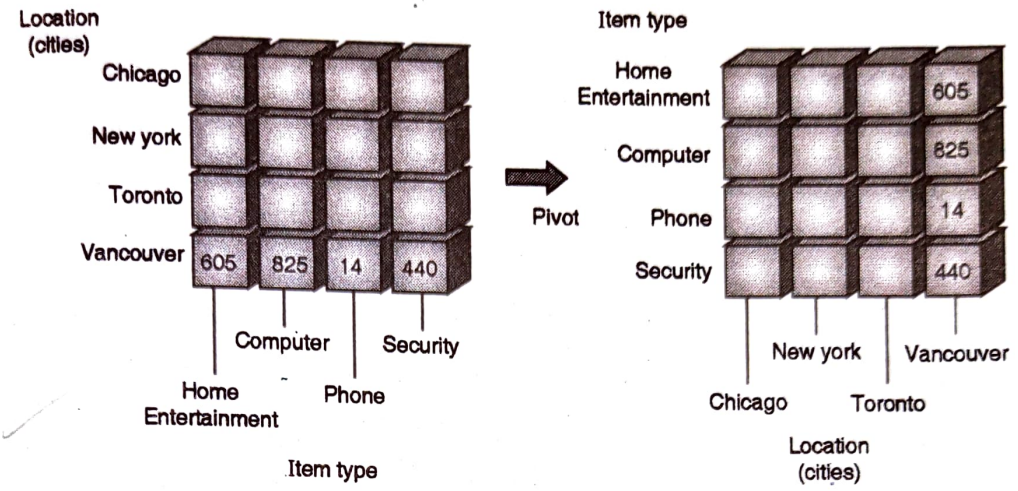


Fig. 2.4.6 : Pivot operation

6. Other OLAP operations

Drill across

This technique is used when there is need to execute a query involving more than one fact table.

Drill through

This technique uses relational SQL facilities to drill through the bottom level of the data cube.

Syllabus Topic : Concept Hierarchies

2.5 Concept Hierarchies

The amount of data may be reduced using concept hierarchies. The low level detailed data (for example numerical values for age) may be represented by higher-level data (e.g. Young, Middle aged or Senior).

Concept hierarchy generation for categorical data

- The users or experts may perform a partial/total ordering of attributes explicitly at schema level :

  E.g. street < city < state < country

- Specification of a hierarchy for a set of values by explicit data grouping :

  E.g. {Acton, Canberra, ACT} < Australia

- Ordering of only a partial set of attributes :

  E.g. only street < city, not others

- By analysing number of distinct values the hierarchies or attribute levels may be generated automatically.

  E.g. for a set of attributes : {street, city, state, country}

  E.g. weekday, month, quarter, year

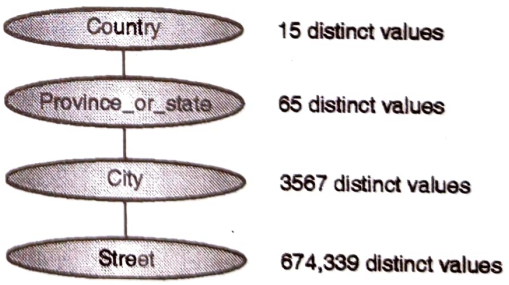


Fig. 2.5.1 : Concept hierarchy example

Syllabus Topic : Data Warehouse Architecture

2.6 Data Warehouse Architecture

- The data in a data warehouse comes from operational systems of the organization as well as from other external sources. These are collectively referred to as ***source systems***.

- The data ***extracted*** from source systems is stored in an area called ***data staging area***, where the data is cleaned, ***transformed***, combined, and duplicated to prepare the data in the data warehouse.
- The data staging area is generally a collection of machines where simple activities like sorting and sequential processing takes place. The data staging area does not provide any query or presentation services.
- As soon as a system provides query or presentation services, it is categorized as a ***presentation server***.
- A presentation server is the target machine on which the data is ***loaded*** from the data staging area organized and stored for direct querying by end users, report writers and other applications.

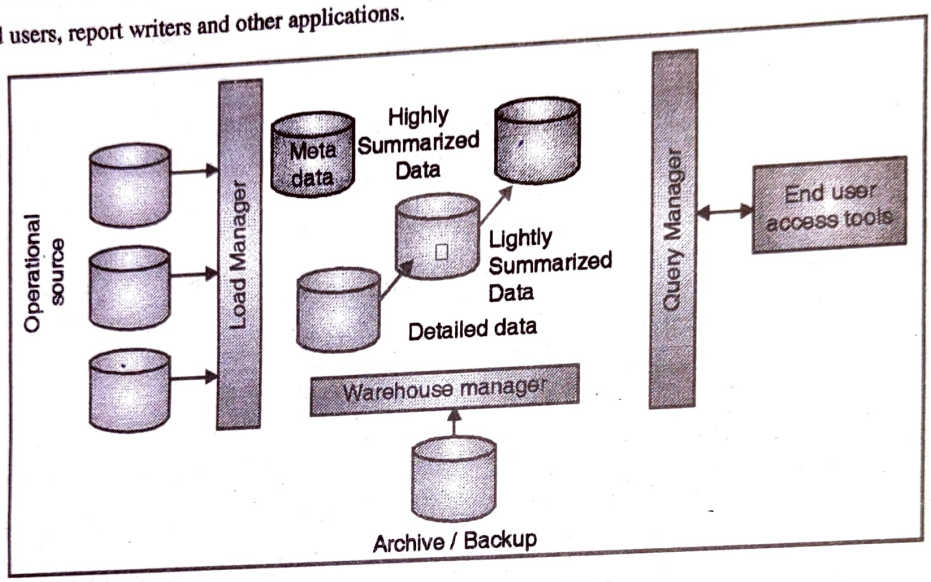


Fig. 2.6.1 : Data Warehouse Architecture

- The three different kinds of systems that are required for a data warehouse are :
  (i) Source Systems (ii) Data Staging Area
  (iii) Presentation servers
- The data travels from source systems to presentation servers via the data staging area. The entire process is popularly known as ETL (extract, transform, and load) or ETT (extract, transform, and transfer). Oracle's ETL tool is called Oracle Warehouse Builder (OWB) and MS SQL Server's ETL tool is called Data Transformation Services (DTS).
- A typical architecture of a data warehouse is shown in Fig. 2.6.1.

Each component and the tasks performed by them are explained below :

1. Operational Source

The sources of data for the data warehouse are supplied from :

- The data from the mainframe systems in the traditional network and hierarchical format.
- Data can also come from the relational DBMS like Oracle, Informix.
- In addition to these internal data, operational data also includes external data obtained from commercial databases and databases associated with supplier and customers.

2. Load Manager

- The load manager performs all the operations associated with extraction and loading data into the data warehouse.
- These operations include simple transformations of the data to prepare the data for entry into the warehouse.
- The size and complexity of this component will vary between data warehouses and may be constructed using a combination of vendor data loading tools and custom built programs.

3. Warehouse Manager

- The warehouse manager performs all the operations associated with the management of data in the warehouse.

- This component is built using vendor data management tools and custom built programs.
- The operations performed by warehouse manager include.
- Analysis of data to ensure consistency.
- Transformation and merging the source data from temporary storage into data warehouse tables.
- Create indexes and views on the base table.

Denormalization

- Generation of aggregation.
- Backing up and archiving of data.
- In certain situations, the warehouse manager also generates query profiles to determine which indexes and aggregations are appropriate.

4. Query Manager

- The query manager performs all operations associated with management of user queries.
- This component is usually constructed using vendor end-user access tools, data warehousing monitoring tools, database facilities and custom built programs.
- The complexity of a query manager is determined by facilities provided by the end-user access tools and database.

5. Detailed Data

- This area of the warehouse stores all the detailed data in the database schema.
- In the majority of cases detailed data is not stored online but aggregated to the next level of details.
- The detailed data is added regularly to the warehouse to supplement the aggregated data.

6. Lightly and Highly Summarized Data

- This stores all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.
- This area of the warehouse is transient as it will be subject to change on an ongoing basis in order to respond to the changing query profiles.
- The main goal of the summarized information is to speed up the query performance.
- As the new data is loaded into the warehouse, the summarized data is updated continuously.

7. Archive and Back up Data

- The detailed and summarized data are stored for the purpose of archiving and back up.
- The data is transferred to storage archives such as magnetic tapes or optical disks.

8. Meta Data

- The data warehouse also stores all the Meta data (data about data) definitions used by all processes in the warehouse.
- It is used for variety of purpose including:
- The extraction and loading process-Meta data is used to map data sources to a common view of information within the warehouse.
- The warehouse management process-Meta data is used to automate the production of summary tables.
- As part of Query Management process - Meta data is used to direct a query to the most appropriate data source.
- The structure of Meta data will differ in each process, because the purpose is different.

9. End-User Access Tools

- The main purpose of a data warehouse is to provide information to the business managers for strategic decision-making.
- These users interact with the warehouse using end user access tools.
- Some of the examples of end user access tools can be:
  - Reporting and Query Tools
  - Application Development Tools
  - Executive Information Systems Tools
  - Online Analytical Processing Tools
  - Data Mining Tools

Syllabus Topic : The Process of Data Warehouse Design

2.7 The Process of Data Warehouse Design

Data Warehouse Design Process

- Choose the grain (atomic level of data) of the business process.
- Choose a business process to model, e.g., orders, invoices, etc.
- Choose the dimensions that will apply to each fact table record.
- Choose the measure that will populate each fact table record

2.8 Data Warehousing Design Strategies or Approaches for Building a Data Warehouse

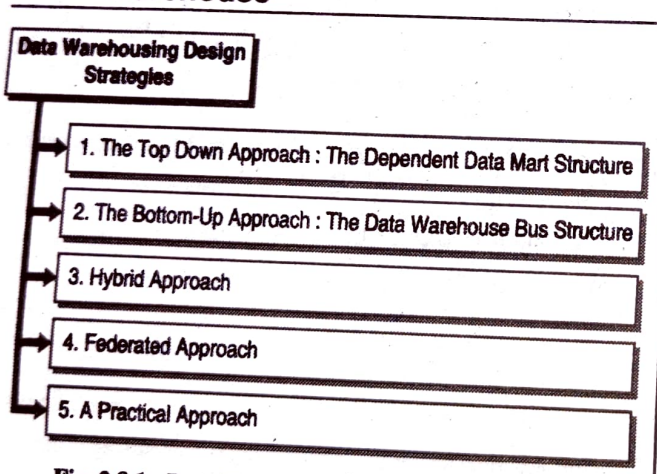


Fig. 2.8.1 : Data Warehousing Design Strategies

2.8.1 The Top Down Approach : The Dependent Data Mart Structure

- The data flow in the top down OLAP environment begins with data extraction from the operational data sources.
- This data is loaded into the staging area and validated and consolidated for ensuring a level of accuracy and then transferred to the Operational Data Store (ODS).
  - The ODS stage is sometimes skipped if it is a replication of the operational databases.

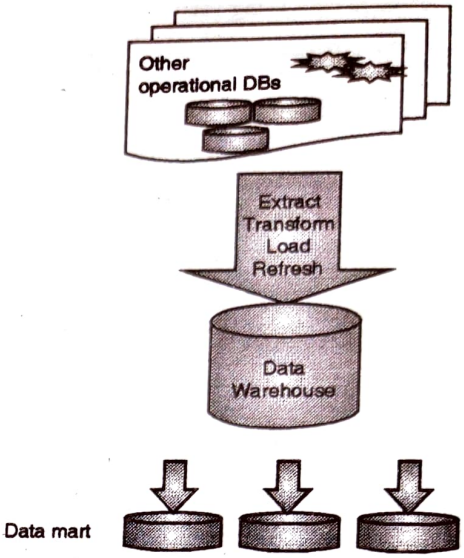


Fig. 2.8.2 : Top Down Approach

- Data is also loaded into the Data warehouse in a parallel process to avoid extracting it from the ODS.
  - Detailed data is regularly extracted from the ODS and temporarily hosted in the staging area for aggregation, summarization and then extracted and loaded into the Data warehouse.
  - The need to have an ODS is determined by the needs of the business.
  - If there is a need for detailed data in the Data warehouse then, the existence of an ODS is considered justified.
  - Else organizations may do away with the ODS altogether.
  - Once the Data warehouse aggregation and summarization processes are complete, the data mart refresh cycles will extract the data from the Data warehouse into the staging area and perform a new set of transformations on them.
  - This will help organize the data in particular structures required by data marts.
- Then the data marts can be loaded with the data and the OLAP environment becomes available to the users.

Advantages of top down approach

- It is not just a union of disparate data marts but it is inherently architected.

- The data about the content is centrally stored and the rules and control are also centralized.
- The results are obtained quickly if it is implemented with iterations.

☞ **Disadvantages of top down approach**

- Times consuming process with an iterative method.
- The failure risk is very high.
- As it is integrated a high level of cross functional skills are required.

2.8.2 The Bottom-Up Approach : The Data Warehouse Bus Structure

- This architecture makes the data warehouse more of a virtual reality than a physical reality. All data marts could be located in one server or could be located on different servers across the enterprise while the data warehouse would be a virtual entity being nothing more than a sum total of all the data marts.
- In this context even the cubes constructed by using OLAP tools could be considered as data marts. In both cases the shared dimensions can be used for the conformed dimensions.
- The bottom-up approach reverses the positions of the Data warehouse and the Data marts.
- Data marts are directly loaded with the data from the operational systems through the staging area.
- The ODS may or may not exist depending on the business requirements. However, this approach increases the complexity of process coordination.
- The data flow in the bottom up approach starts with extraction of data from operational databases into the staging area where it is processed and consolidated and then loaded into the ODS.
- The data in the ODS is appended to or replaced by the fresh data being loaded.
- After the ODS is refreshed the current data is once again extracted into the staging area and processed to fit into the Data mart structure.

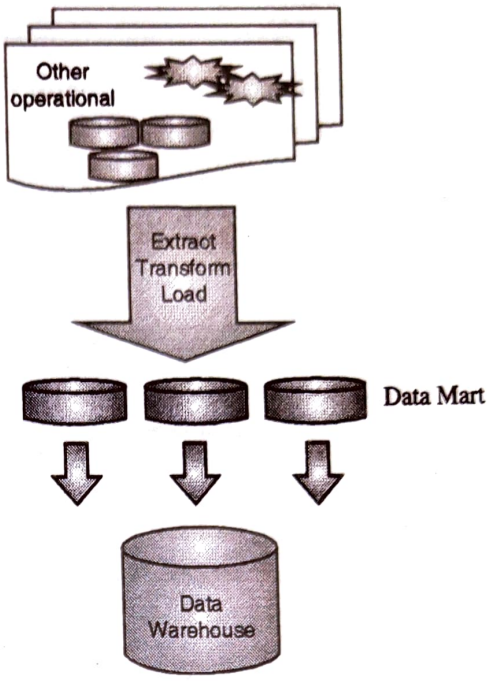


Fig. 2.8.3 : Bottom Up Approach

- The data from the Data mart then is extracted to the staging area aggregated, summarized and so on and loaded into the Data Warehouse and made available to the end user for analysis.

☞ **Advantages of bottom up approach**

- This model strikes a good balance between centralized and localized flexibility.
- Data marts can be delivered more quickly and shared data structures along the bus eliminate the repeated effort expended when building multiple data marts in a non-architected structure.
- The standard procedure where data marts are refreshed from the ODS and not from the operational databases ensures data consolidation and hence it is generally recommended approach.
- Manageable pieces are faster and are easily implemented.
- Risk of failure is low.
- Allows one to create important data mart first.

☞ **Disadvantages of bottom up approach**

- Allows redundancy of data in every data mart.
- Preserves inconsistent and incompatible data.
- Grows unmanageable interfaces.

2.8.3 Hybrid Approach

- The Hybrid approach aims to harness the speed and user orientation of the Bottom up approach to the integration of the top-down approach.
- The Hybrid approach begins with an Entity Relationship diagram of the data marts and a gradual extension of the data marts to extend the enterprise model in a consistent, linear fashion.
- These data marts are developed using the star schema or dimensional models.
- The Extract, Transform and Load (ETL) tool is deployed to extract data from the source into a non persistent staging area and then into dimensional data marts that contain both atomic and summary data.
- The data from the various data marts are then transferred to the data warehouse and query tools are reprogrammed to request summary data from the marts and atomic data from the Data Warehouse.

Advantages of hybrid approach

- Provides rapid development within an enterprise architecture framework.
- Avoids creation of renegade "independent" data marts.
- Instantiates enterprise model and architecture only when needed and once data marts deliver real value.
- Synchronizes meta data and database models between enterprise and local definitions.
- Backfilled DW eliminates redundant extracts.

Disadvantages of hybrid approach

- Requires organizations to enforce standard use of entities and rules.
- Backfilling a DW is disruptive, requiring corporate commitment, funding, and application rewrites.
- Few query tools can dynamically query atomic and summary data in different databases.

2.8.4 Federated Approach

- This is a hub-and-spoke architecture often described as the "architecture of architectures". It recommends an integration of heterogeneous data warehouses, data marts and packaged applications that already exist in the enterprise.
- The goal is to integrate existing analytic structures wherever possible and to define the "highest value" metrics, dimensions and measures and share and reuse them within existing analytic structures.
- This may result in the creation of a common staging area to eliminate redundant data feeds or building of a data warehouse that sources data from multiple data marts, data warehouses or analytic applications.
- Hackney-a vocal proponent of this architecture claims that it is not an elegant architecture but it is an architecture that is in keeping with the political and implementation reality of the enterprise.

Advantages of federated approach

- Provides a rationale for "band aid" approaches that solve real business problems.
- Alleviates the guilt and stress data warehousing managers might experience by not adhering to formalized architectures.
- Provides pragmatic way to share data and resources.

Disadvantages of federated approach

- The approach is not fully articulated.
- With no predefined end-state or architecture in mind, it may give way to unfettered chaos.
- It might encourage rather than dominant in independent development and perpetuate the disintegration of standards and controls.

2.8.5 A Practical Approach

The Steps in the Practical Approach are :

1. The first step is to do Planning and defining requirements at the overall corporate level.

2. An architecture is created for a complete warehouse.
3. The data content is conformed and standardized.
4. Consider the series of supermarts one at a time and implement the data warehouse.
5. In this practical approach, first the organizations needs are determined. The key to this approach is that planning is done first at the enterprise level. The requirements are gathered at the overall level.
6. The architecture is established for the complete warehouse. Then the data content for each supermart is determined. Supermarts are carefully architected data marts. Supermart is implemented one at a time.
7. Before implementation checks the data types, field length etc. from the various supermarts, which helps to avoid spreading of different data across several data marts.
8. Finally a data warehouse is created which is a union of all data marts. Each data mart belongs to a business process in the enterprise, and the collection of all the data marts form an enterprise data warehouse.

Syllabus Topic : A Three-Tier Data Warehousing Architecture

2.9 A Three-Tier Data Warehousing Architecture

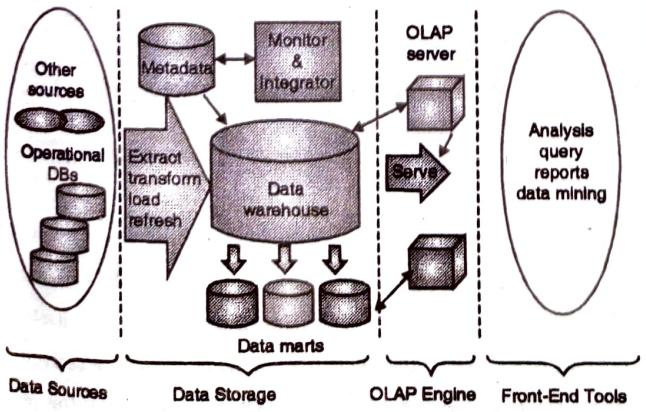


Fig. 2.9.1 : Multi-tier Data warehouse Architecture

1. Bottom Tier (Data Sources and Data Storage)

- It is a warehouse database server, that is generally a RDBMS.
- Using Application Program interfaces (called as gateways), data is extracted from operational and external sources.
- Gateways like, ODBC (Open Database connection), OLE-DB (Open linking and embedding for database), JDBC (Java Database Connection) is supported by underlying DBMS.

2. Middle Tier (OLAP Engine)

OLAP Engine is either implemented using ROLAP (Relational online Analytical Processing) or MOLAP(Multidimensional OLAP).

3. Top Tier (Front End Tools)

- This tier is a client which contains query and reporting tools, Analysis tools, and /or data mining tools.
- From the Architecture Point of view there are three data warehouse Models:

(a) Enterprise Warehouse

- The information of the entire organization is collected related to various subjects in enterprise warehouse.

(b) Data Mart

- A subset of Warehouse that is useful to a specific group of users.
- It can be categorized as Independent vs. dependent data mart.

(c) Virtual warehouse

- A set of views over operational databases.
- Only some of the possible summary views may be materialized.

2.9.1 Data Warehouse and Data Marts

Data Mart defined

- A data mart is oriented to a specific purpose or major data subject that may be distributed to support business needs. It is a subset of the data resource.

- A data mart is a repository of a business organization's data implemented to answer very specific questions for a specific group of data consumers such as organizational divisions of marketing, sales, operations, collections and others.
- A data mart is typically established as one dimensional model or star schema which is composed of a fact table and multi-dimensional table.
- A data mart is a small warehouse which is designed for the department level.
- It is often a way to gain entry and provide an opportunity to learn.
- **Major problem :** If they differ from department to department, they can be difficult to integrate enterprise-wide.

Table 2.9.1 : Differences between Data Warehouse and Data Mart

Sr. No.	Data Warehouse	Data Mart
1.	A data warehouse is application independent.	A data mart is a dependent on specific DSS application.
2.	It is centralized, and enterprise wide.	It is decentralized by user area.
3.	It is well planned.	It is possibly not planned.
4.	The data is historical, detailed and summarized.	The data consists of some history, detailed and summarized.
5.	It consists of multiple subjects.	It consists of a single subject of concern to the user.
6.	It is highly flexible.	It is restrictive.
7.	Implementation takes months to year.	Implementation is done usually in months.
8.	Generally size is from 100 GB to 1 TB.	Generally size is less than 100 GB.

Syllabus Topic : Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP

2.10 Types of OLAP Servers : ROLAP versus MOLAP versus HOLAP

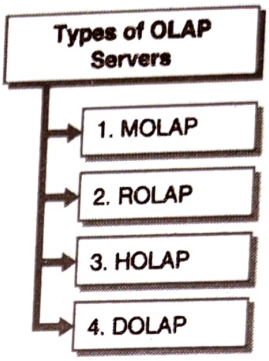


Fig. 2.10.1 : Types of OLAP Servers

Approaches to OLAP Servers

In the OLAP world, there are mainly two different types of OLAP servers: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP). Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.

2.10.1 MOLAP

This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Advantages of MOLAP

1. **Excellent performance :** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
2. **Can perform complex calculations :** All calculations have been pre-generated when the cube is created.

Disadvantages of MOLAP

1. **Limited in the amount of data it can handle :** Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.

2. **Requires additional investment :** Cube technology are often proprietary and do not already exists in the organization. Therefore, to adopt MOLAP technology, chances of additional investments in human and capital resources are needed.

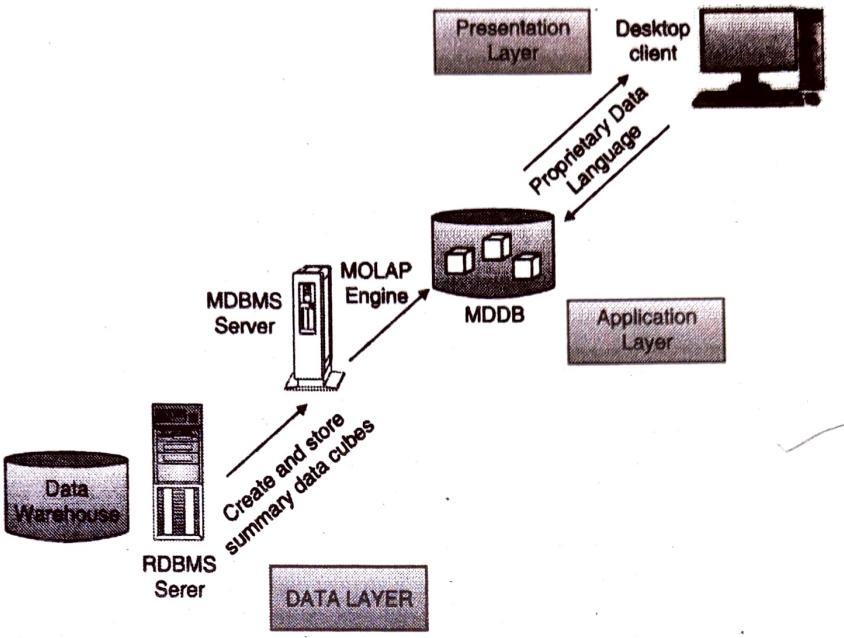


Fig. 2.10.2 : MOLAP Process

2.10.2 ROLAP

This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

Advantages of ROLAP

1. **Can handle large amounts of data**

   The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on amount of data.

2. **Can leverage functionalities inherent in the relational database**

   Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages of ROLAP

1. **Performance can be slow :** Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
2. **Limited by SQL functionalities :** Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL can do.
3. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.

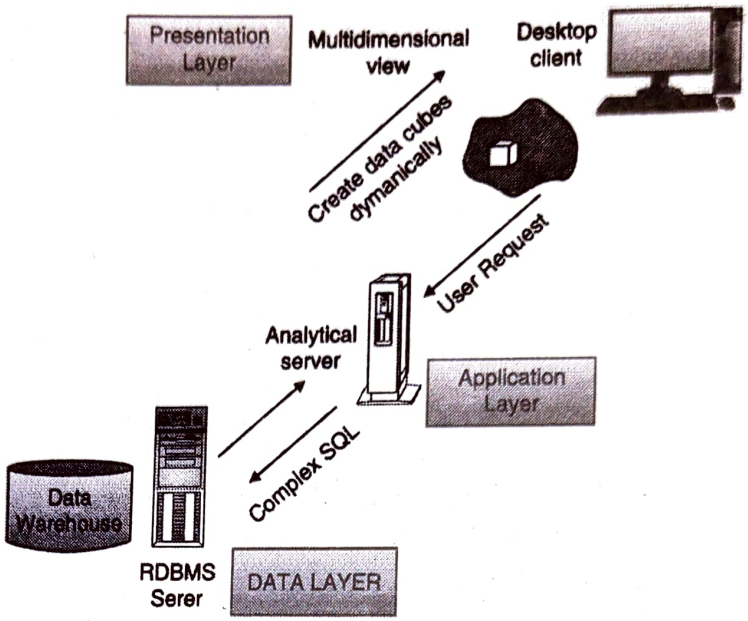


Fig. 2.10.3 : ROLAP Process

2.10.3 HOLAP

- HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information, HOLAP leverages cube technology for faster performance.
- When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data.
- For example, a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store. The Microsoft SQL Server 7.0 OLAP Services supports a hybrid OLAP server.

2.10.4 DOLAP

It is Desktop Online Analytical Processing and variation of ROLAP. It offers portability to users of OLAP. For DOLAP, it needs only DOLAP software to be present on machine. Through this software, multidimensional datasets are formed and transferred to desktop machine.

2.11 Examples of OLAP

Ex. 2.11.1 : Consider a data warehouse for a hospital where there are three dimension (a) Doctor (b) Patient (c) Time and two measures i) count ii) charge where charge is the fee that the doctor charges a patient for a visit.

Using the above example describe the following OLAP operations.

1) Slice 2) Dice
3) Rollup 4) Drill down
5) Pivot

Soln. :

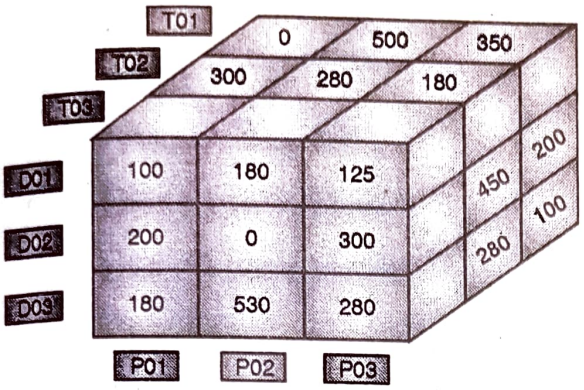
There are four tables, out of 3 dimension tables and 1 fact table.

Dimension tables

1. Doctor (DID, name, phone, location,pin,specialisation)
2. Patient (PID,name, phone , state, city,location, pin)
3. Time (TID,day, month, quarter , year)

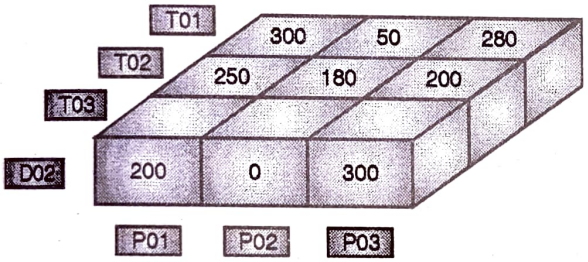
Fact Table

Fact_table (DID,PID,TID, count, charge)

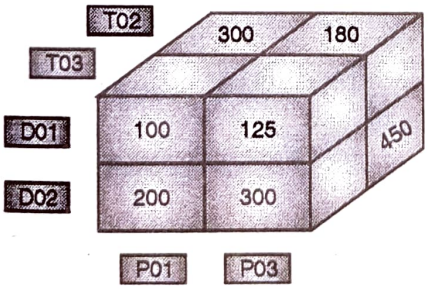


Operations

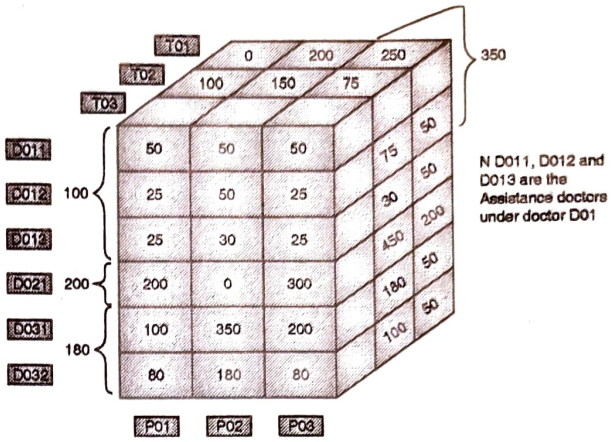
1. **Slice :** Slice on fact table with DID = 2 , this cuts the cube at DID = 2 along the time and patient axis thus it will display a slice of cube, in which time on x and patient on y axis.



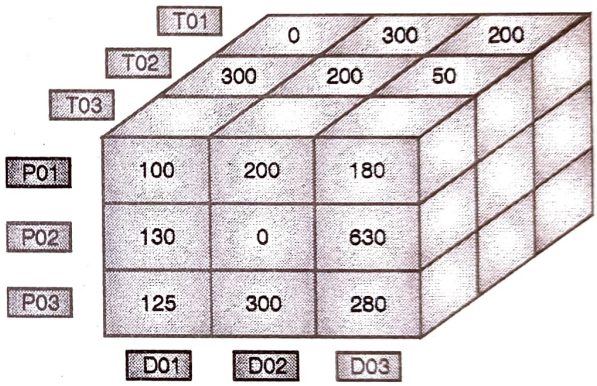
2. **Dice :** It is a sub cube of main cube. Thus it cuts the cube with more than one predicate like dice on cube with DID = 2, and DID = 01 and PID = 01 and PID = 03 and TID = 02, 03



3. **Roll up :** It gives summary based on concept hierarchies. Assuming there exists concept hierarchy in patient table as state->city->location. Then roll up will summarise the charges or count in terms of city or further roll up will give charges for a particular state etc.



4. **Drill down :** It is opposite to roll up that means if currently cube is summarised with respect to city then drill down will also show summarisation with respect to location.



5. **Pivot :** It rotates the cube, sub cube or rolled-up or drilled-down cube, thus changing the view of the cube.

Ex. 2.11.2 : All Electronics Company have sales department consider three dimensions namely

(i) Time

(ii) Product

(iii) store

The Schema Contains a central fact table sales with two measures

(i) dollars-cost and

(ii) units-sold

Using the above example describe the following OLAP operations :

(i) Dice (ii) Slice

(iii) Roll-up (iv) drill Down.

Soln. :

There are four tables, out of these 3 dimension tables and 1 fact table.

For OLAP operations refer Example 2.11.1.

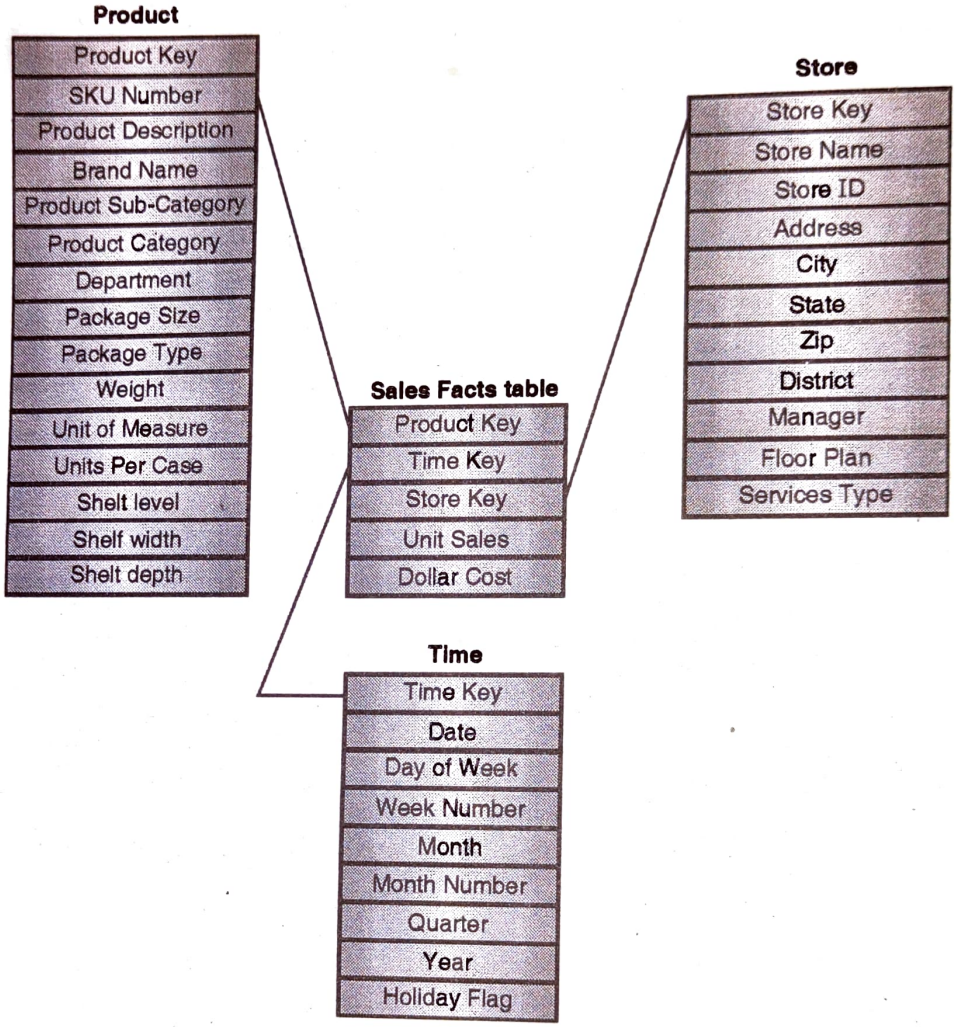


Fig. P. 2.11.2 : Star Schema for Electronics Company sales department