Shree Rahul Education Society's (Regd.)
SHREE L. R. TIWARI
COLLEGE OF ENGINEERING
(Approved by AICTE & DTE, Maharashtra State & Affiliated to University of Mumbai)
NAAC Accredited, NBA Accredited Program, ISO 9001:2015 Certified | DTE Code No. : 3423
Minority Status (Hindi Linguistic)

DEPARTMENT OF COMPUTER ENGINEERING

CSL503 Data Warehousing & Mining Lab

Fifth Semester, 2023-2024 (Odd Semester)

Name of Student :

Roll No. :

Batch :

Venue :

Experiment No. :

Title of Experiment :

Date of Conduction :

Date of Submission :

Particulars	Max. Marks	Marks Obtained
Preparedness and Efforts (PE)	**3**	
Knowledge of Tools (KT)	**3**	
Debugging and Results (DR)	**3**	
Documentation (DN)	**3**	
Punctuality & Lab Ethics (PL)	**3**	
Total	**15**	

Grades – Meet Expectations (3 Marks), Moderate Expectations (2 Marks), Below Expectations (1 Mark)

Checked and verified by

Name of Faculty: Mrs. Harsha Dave

Signature :

Date :

EXPERIMENT NO. 1

TITLE: Case study on building Data warehouse/Data Mart

AIM: One case study on building Data warehouse/Data Mart Write Detailed Problem statement and design dimensional modelling (creation of star and snowflake schema)

THEORY:

Dimensional Modelling:
In order to maximize query performance and enable clear data analysis, dimensional modelling is a data modelling technique used in data warehousing. It arranges information into two primary kinds of tables: dimension tables, which offer descriptive features and context, and fact tables, which include quantitative measurements. Dimensional modelling makes complicated data schemas simpler by breaking up data into these discrete units and creating links between them using foreign keys. This methodology improves the usability and accessibility of data, facilitating effective reporting and analysis that closely matches with business requirements and supports well-informed decision-making processes at different organizational levels.

Elements of Dimensional Data Model:

A dimensional data model typically consists of several key elements:

1. Fact Tables: These tables contain quantitative data or metrics, such as sales revenue, quantity sold, and expenses incurred. Foreign keys to dimension tables are frequently found in fact tables.

2. Dimension Tables: These tables contain descriptive information about the business dimensions connected with the data in the fact tables. Dimensions may encompass time, geography, items, or customers. Dimension tables have attributes that specify the properties of each dimension.

3. Facts: These are numerical or quantitative data points that describe business transactions or events. They are recorded in fact tables and used as quantitative metrics for analysis

4. Attributes: Attributes are descriptive fields found within dimension tables that provide context and information about the dimension. A "time" dimension, for example, may contain properties such as year, quarter, month, and date.

5. Hierarchies: Hierarchies define the logical relationships between attributes within a dimension. They enable drill-down and roll-up capabilities, allowing users to navigate data at different levels of granularity**.**

6. Aggregations: Aggregations are precomputed summaries of data stored in fact tables. They improve query performance by reducing the amount of data that needs to be processed for analytical **queries.**

7. Star Schema or Snowflake Schema: These are common schema designs used in dimensional modeling. The star schema consists of a single fact table surrounded by multiple dimension tables, whereas the snowflake schema further normalizes dimension tables by breaking them into additional sub-tables.

By leveraging these elements, dimensional data modeling creates a structure that optimizes data retrieval and analysis, supports intuitive querying, and enhances the overall usability of data for decision-making and reporting purposes in data warehousing environments.

Steps of Dimensional Modelling

Dimensional modeling involves several key steps to create an effective data model for data warehousing and business intelligence purposes:

1. Identify Business Processes and Requirements: Understand the business processes and identify the key metrics and dimensions that are important for analysis. This requires collaboration with stakeholders to determine the scope and goals of the dimensional model.

2. Identify Dimensions and Hierarchies: Identify the dimensions, such as time, product, and customer, that will provide context to the facts. It is advisable to establish hierarchies within each dimension, such as year, quarter, month, and day for the time dimension, in order to facilitate drill-down and roll-up analysis.

3. Design the Fact Tables: Design fact tables that will store the quantitative data or metrics (facts) of interest. Determine the granularity of each fact table (e.g., daily sales, monthly revenue) based on business requirements.

4. Design the Dimension Tables: Create dimension tables that will contain descriptive attributes related to each dimension. To provide context to the facts stored in the fact tables, it is imperative to establish the attributes and hierarchies within each dimension table.

5. Define Relationships: Establish relationships (foreign keys) between fact tables and dimension tables. This linkage enables users to navigate and analyze data across different dimensions.

6. Optimize for Query Performance: Consider performance optimizations, such as aggregations, indexing, and partitioning, to improve query performance. Aggregations involve precomputing summary values in order to speed up queries, while indexing and partitioning improve data retrieval.

7. Validate and Iterate: Validate the dimensional model with stakeholders to ensure it meets business requirements and supports analytical needs. Iterate on the design based on feedback and changes in business requirements.

8. Implement and Populate: Implement the dimensional model in the data warehouse environment. Populate the fact and dimension tables with data extracted from operational systems or other sources.

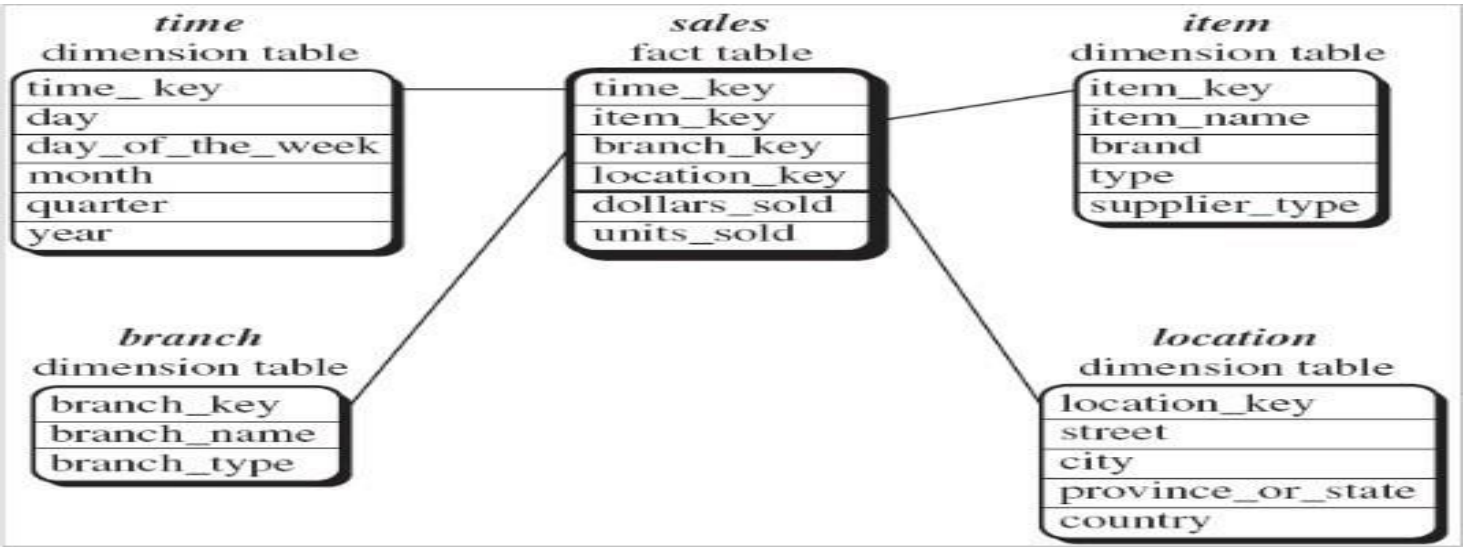
9. Test and Tune: Test the dimensional model to ensure data integrity and accuracy. Tune the model based on performance testing and user feedback to optimize query response times and usability.

10. Document and Maintain: Document the dimensional model, including schemas, relationships, and business rules. Maintain the model over time by updating it to reflect changes in business processes or data sources.

By following these steps, dimensional modeling creates a structured and efficient framework for organizing data in a data warehouse, enabling intuitive analysis and reporting that supports informed decision-making across the organization.

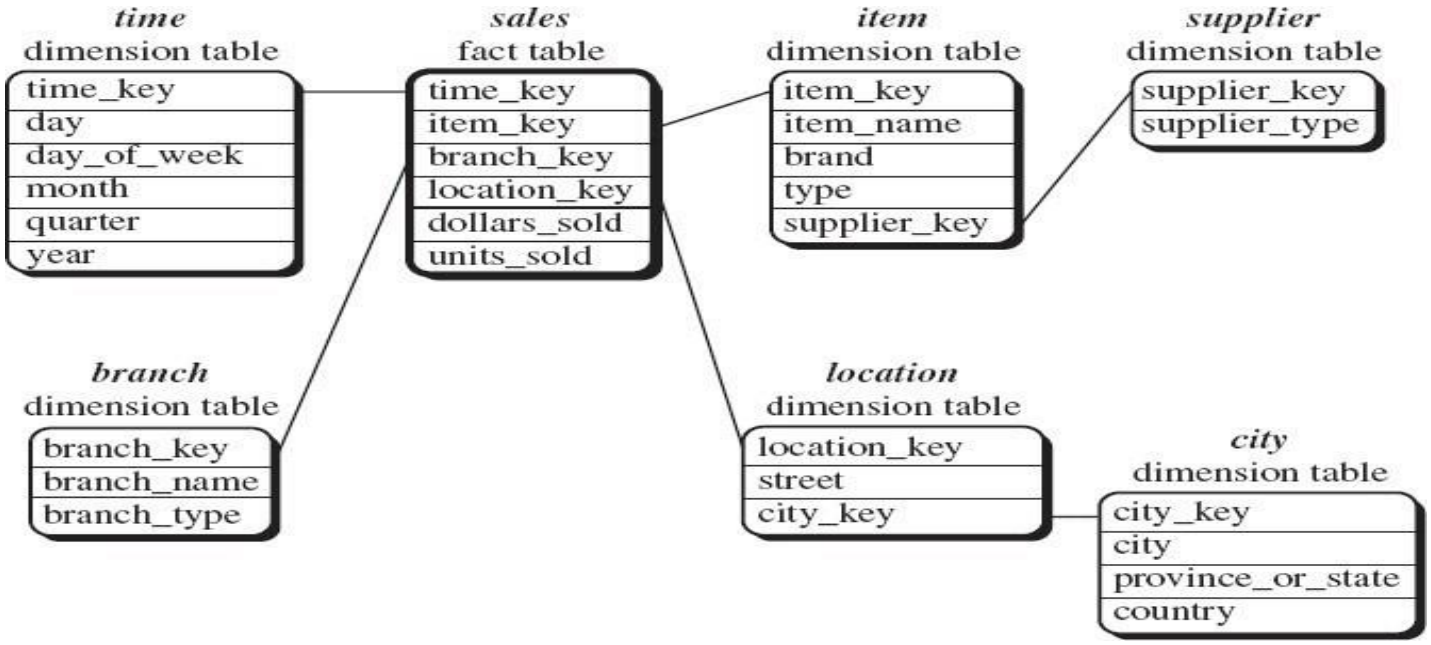
Star Schema:

The star schema is a technique for dimensional modeling used in data warehousing. It consists of a central fact table and dimension tables. The fact table stores quantitative information, such as sales revenue or quantities sold, while the dimension table provides context with descriptive attributes, such as time, product, or customer. Each dimension table is linked to the fact table through foreign keys, creating a star-like structure. This design improves querying and performance by enabling efficient data retrieval and analysis. Star schemas are widely embraced due to their clarity, adaptability to business changes, and efficacy in facilitating business intelligence and analytical applications.



Snowflake schema:

The snowflake schema is a normalized form of dimensional modeling used for data warehousing. It extends the star schema by further normalizing dimension tables into multiple related tables. This process of normalization reduces redundancy by dividing hierarchies into distinct tables. Each dimension table in a snowflake schema can have its attributes stored in additional related tables that are linked through foreign key relationships. While it promotes data integrity and reduces storage space, it can also complicate queries compared to star schemas. Snowflake schemas are suitable for environments where data consistency and space efficiency hold precedence over query performance and simplicity in analytical processing**.**



LAB EXERCISE: Students will select one case study on building Data warehouse/Data Mart. Write DetailedProblem statement and design dimensional modelling (creation of star and snowflake schema)

Problem Statement:

Ex. 1.14.2 : 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

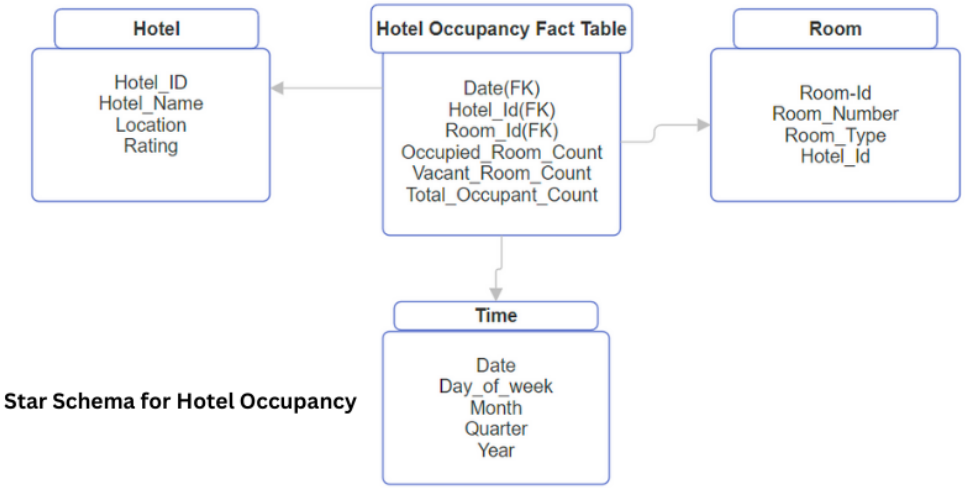
Professor Attributes : Prof_id, Prof_Name, Title, Department_id, department_name

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

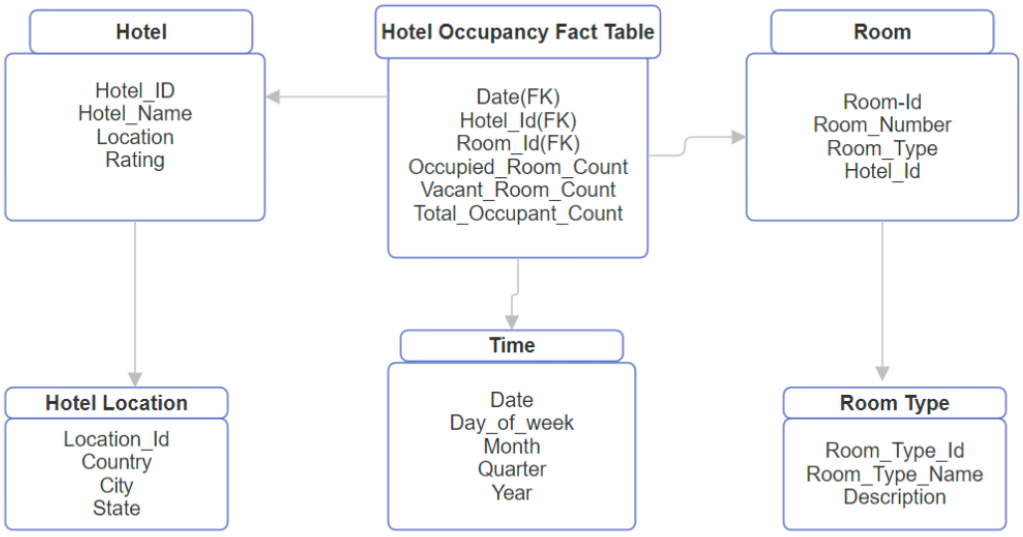
Period Attributes : Semester_id, Year. The database will contain Data for 30 months periods. The only fact that is to be recorded in the fact table is course Grade

OR Information requirements are recorded for "Hotel occupancy" considering dimensions like Hotel, Room and Time. Few Facts recorded are vacant rooms, occupied rooms, number of occupants etc. Answer the following quesitons for this problem : (i) Design the star schema (ii) Can you convert this star schema to a snowflacke schema ? If yes, justify and draw the snowflake schema. **MU - May 18, Dec. 18**

Ex. 1.14.19 : A manufacturing company has a has a huge sales network. To control the sales, it is divided into regions. Each region has multiple zones. Each zone has different cities. Each sales person is allocated different cities. The objective is to track sales figure at different granularity levels for region, sales person and time. Convert the star schema to snowflake schema. **MU - Dec. 17**



Star Schema for Hotel Occupancy



Snowflake Schema for Hotel Occupancy

CONCLUSION: Hence, we have understood the concept of dimensional modelling and we have created thedimensional modelling design using star schema and snowflake schema for the selected case study for buildingData Ware house/Data Mart.