DATA WAREHOUSING & MINING

Unit-2: Data Warehousing

Parameshwar R Hegde, PhD

Basic Concepts

- **A subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process**
  - ➢ Provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions
  - ➢ That support information processing by providing a solid platform of consolidated historic data for analysis

- The four keywords—subject-oriented, integrated, time-variant, and nonvolatile —distinguish data warehouses from other data repository systems

Data Warehousing Keywords

- **Subject-oriented**
  - ➢ Organized around major subjects

    Eg., customer, supplier, product and sales
  - ➢ Instead of day-to-day operations and transaction processing a data warehouse focuses on the modeling and analysis of data for decision makers
  - ➢ Hence, data warehouses typically provide a simple and concise view of particular subject issues
  - ➢ Excludes data that are not useful in the decision support process

Data Warehousing Keywords

- **Integrated**
  - ➢ Integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records
  - ➢ Data cleaning techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on

- **Time-variant**
  - ➢ Provide information from an historic perspective (Eg., past 5–10 years)
  - ➢ Data warehouse contains, either implicitly or explicitly, a time element

Data Warehousing Keywords

- **Nonvolatile**
  - Physically separate store of data transformed from the application data found in the operational environment
  - A data warehouse does not require transaction processing, recovery and concurrency control mechanisms
  - Requires only two operations in data accessing: initial loading of data and access of data

Data Ware House Applications for an Organization

- **Increasing customer focus, which includes the analysis of customer buying patterns**
  - Eg., Buying preference, buying time, budget cycles and appetites for spending
- **Repositioning products and managing product portfolios by comparing the performance**
  - Eg., Sales by quarter, by year and by geographic regions in order to fine-tune production strategies
- **Analyzing operations and looking for sources of profit**
- **Managing customer relationships, making environmental corrections, and managing the cost of corporate assets**

Differences between Database and Data Warehouse

- **Online transaction processing (OLTP) systems -** Online operational database systems is to perform online transaction and query processing
  - ➢ Cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting

- **Online analytical processing (OLAP) systems -** Users or knowledge workers in the role of data analysis and decision making
  - ➢ Can organize and present data in various formats in order to accommodate the diverse needs of different users

Differences between Database and Data Warehouse

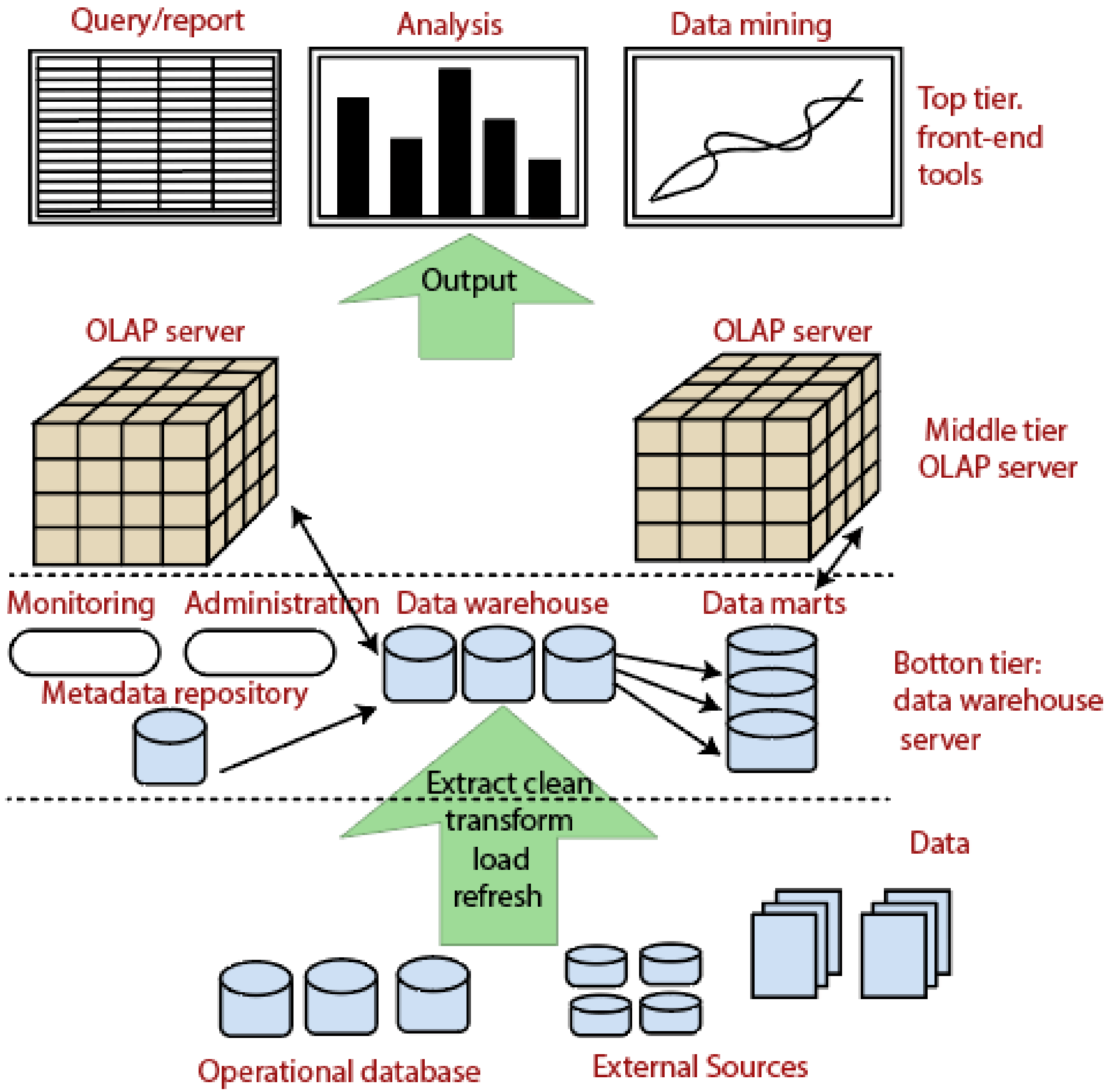
Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (Eg., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/ write	mostly read

Differences between Database and Data Warehouse

Feature	*OLTP*	*OLAP*
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data Warehousing: A Multitiered Architecture

Three-tier data warehousing architecture



https://www.javatpoint.com/three-tier-data-warehouse-architecture

Data Warehousing: A Multitiered Architecture

- **Top-tier:** Contains front-end tools for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP-generated data

- **Middle-tier:** Consists of an OLAP server for fast querying of the data warehouse
  1. Relational OLAP (ROLAP) model, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations
  2. Multidimensional OLAP (MOLAP) model, i.e., a particular purpose server that directly implements multidimensional information and operations

- **Bottom-tier**: Consists of the Data Warehouse server, which is almost always an RDBMS. It may include several specialized data marts and a metadata repository

Data Warehouse Modeling

- **Data Cube**
  - ➢ Data to be modeled and viewed in multiple dimensions
  - ➢ Dimensions are the perspectives or entities with respect to which an organization wants to keep records

    Eg., Sales data warehouse; time, item, branch, and location are dimensions
  - ➢ Each dimension may have a table associated with it, called a dimension table

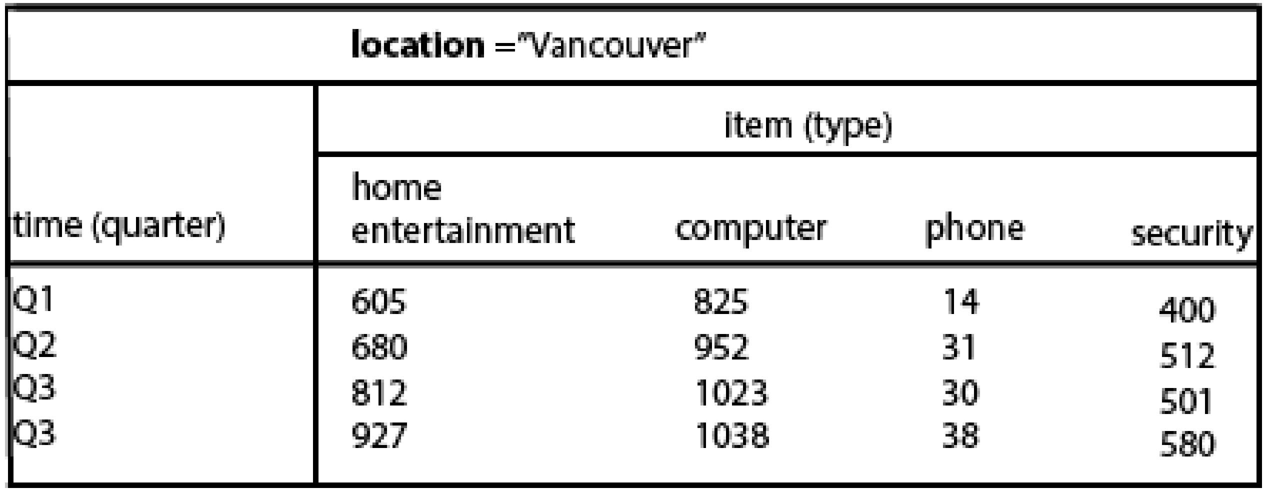
    Eg., Dimension table for item may contain the attributes item_name, brand and type

- Although we usually think of cubes as 3-D geometric structures, in data warehousing the data cube is *n*-dimensional

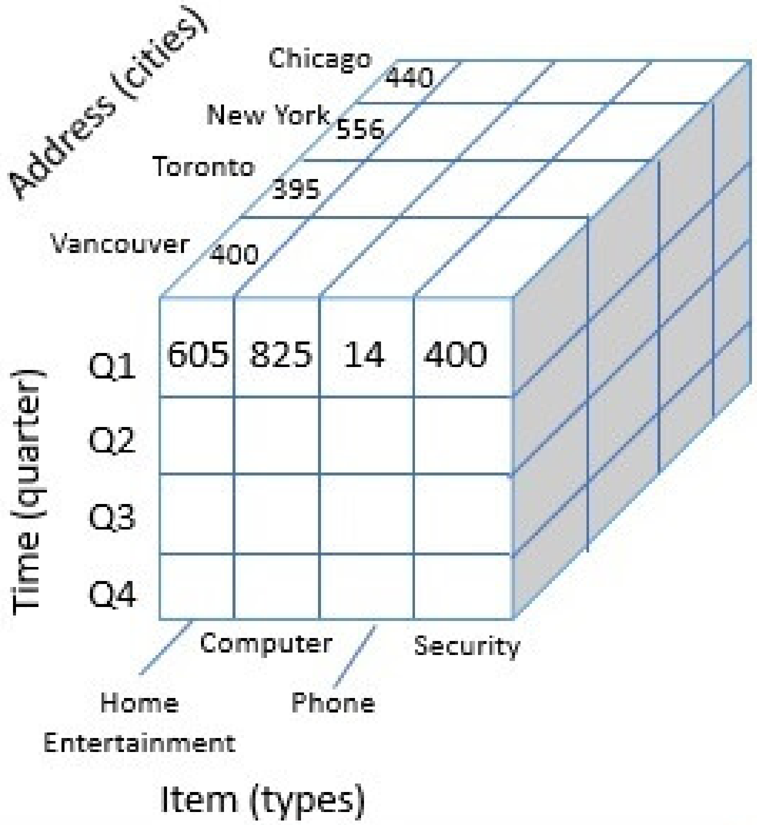
Data Warehouse Modeling

- **Data Cube – 2D and 3D view**

2-D view of Sales Data



location ="Vancouver"				
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q3	927	1038	38	580



Data Cube *AllElectronics* Sales

Data Warehouse Modeling

- **Data Cube – 2 types**

  1. **Multidimensional data cube:**
     - ➢ Helps in storing large amounts of data by making use of a multi-dimensional array
     - ➢ Increases its efficiency by keeping an index of each dimension. Thus, dimensional is able to retrieve data fast

  2. **Relational data cube:**
     - ➢ Helps in storing large amounts of data by making use of relational tables
     - ➢ Each relational table displays the dimensions of the data cube. It is slower compared to a multidimensional data cube

Data Warehouse Modeling

1. **Data Mart**
   - Contains a subset of organization/ wide data
   - Contains data specific to particular group
   - Small in size
   - Very flexible
   - Why?
     - Access control strategy
     - Different hardware platform
     - Speed up execution
   - 3 types of data marts
     - Dependent
     - Independent
     - Hybrid

Data Warehouse Modeling

2. **Enterprise**
   - Collects all the data information and subject spanning of an entire organization
   - Provides wide data integration
   - Integrated from operational system and external information providers
   - Attributes
     - Has a single vision of truth
     - Multiple subject area
     - Normalized design
     - Mission critical environment i.e., can handle any situation
     - Scalable

Data Warehouse Modeling

3. **Virtual**
   - Virtual view of database
   - Have a logical description of the databases and statements
   - Creates single database from all the data sources
   - Allow to access distributed data through single query
   - Why?
     - Fast access
     - Abstraction – Hiding information of source (location)
     - Virtualizing data access
     - Transformation – Source data for consumer
     - Combine result sets from across multiple source system

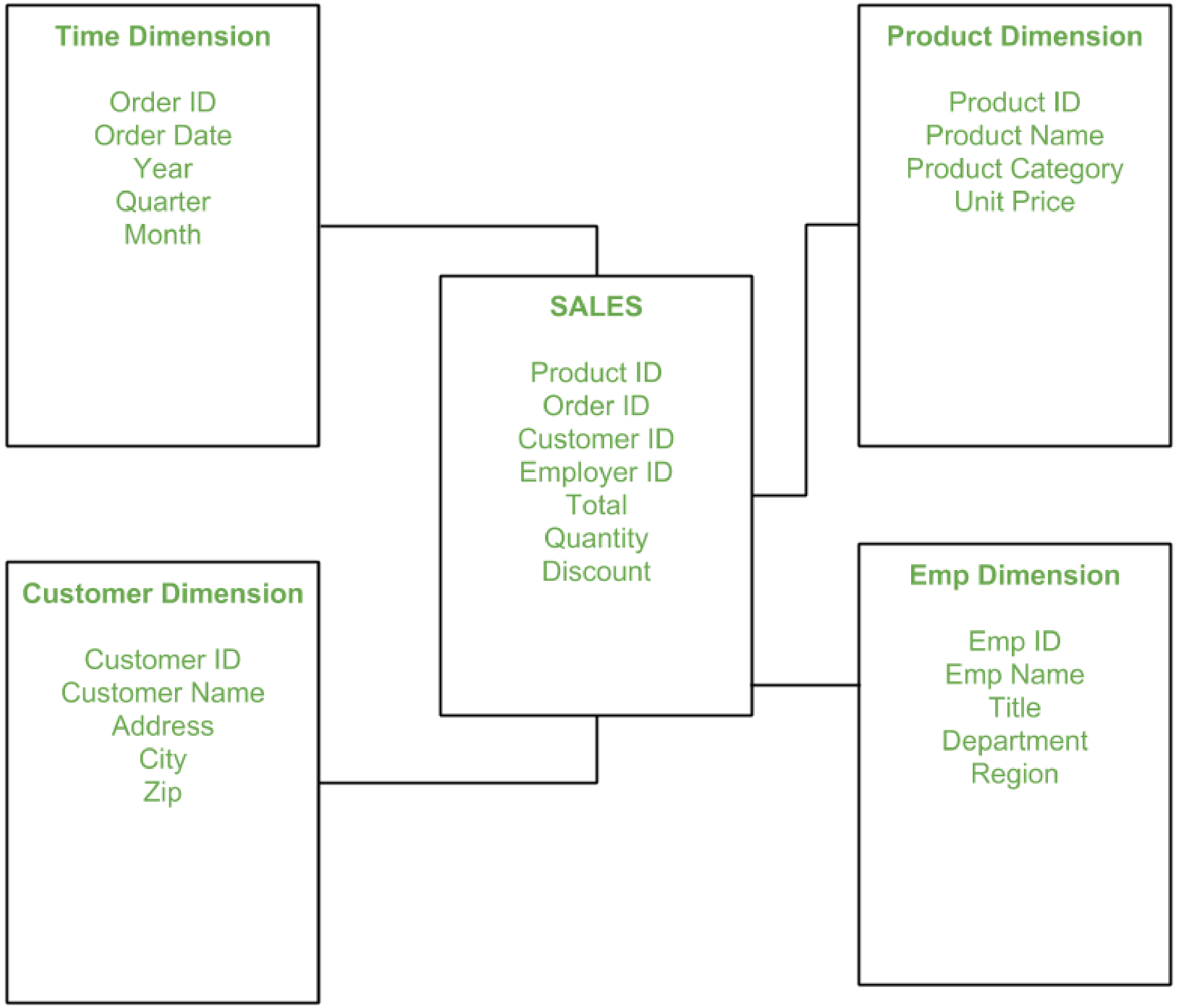
Data Warehouse Modeling

- **Schemas –**
  1. **Star:**
     - Most common modeling paradigm is the star schema, in which the data warehouse contains
       I. Large central table (fact table) containing the bulk of the data, with no redundancy
       II. Set of smaller attendant tables (dimension tables), one for each dimension
       III. Schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table

Data Warehouse Modeling

- **Schemas –**

  1. Star:



Time Dimension

Order ID
Order Date
Year
Quarter
Month

Product Dimension

Product ID
Product Name
Product Category
Unit Price

SALES

Product ID
Order ID
Customer ID
Employer ID
Total
Quantity
Discount

Customer Dimension

Customer ID
Customer Name
Address
City
Zip

Emp Dimension

Emp ID
Emp Name
Title
Department
Region

https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/

Data Warehouse Modeling

- **Schemas –**

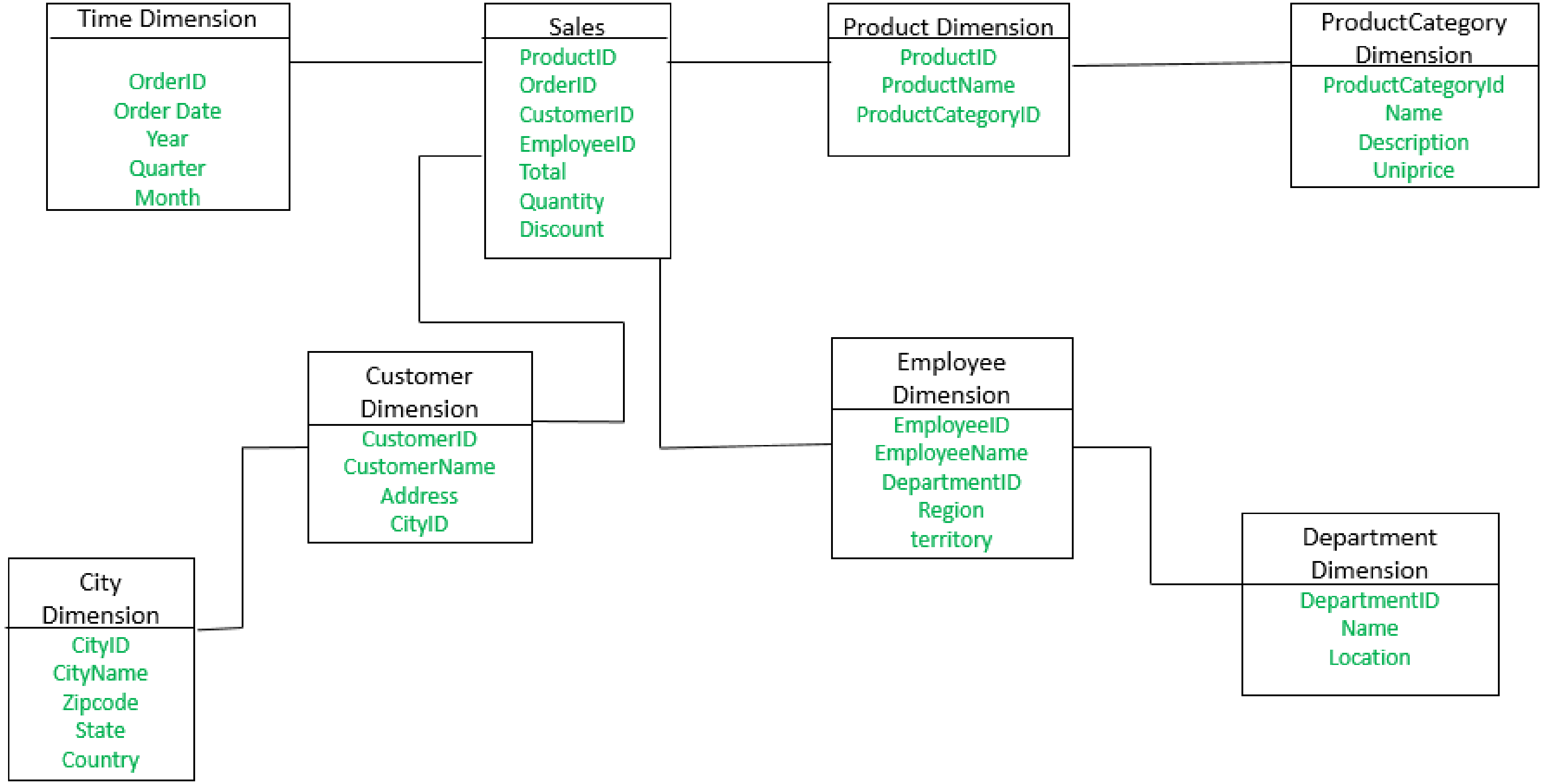
  2. Snowflake:

  - Variant of the star schema model - materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship
  - Some dimension tables are normalized, thereby further splitting the data into additional tables
  - Easy to maintain and saves storage space
  - Can reduce the effectiveness of browsing, since more joins will be needed to execute a query
  - Consequently, the system performance may be adversely impacted; hence, not popular

Data Warehouse Modeling

- **Schemas –**

  2. Snowflake:



https://www.geeksforgeeks.org/snowflake-schema-in-data-warehouse-model/

Data Warehouse Modeling

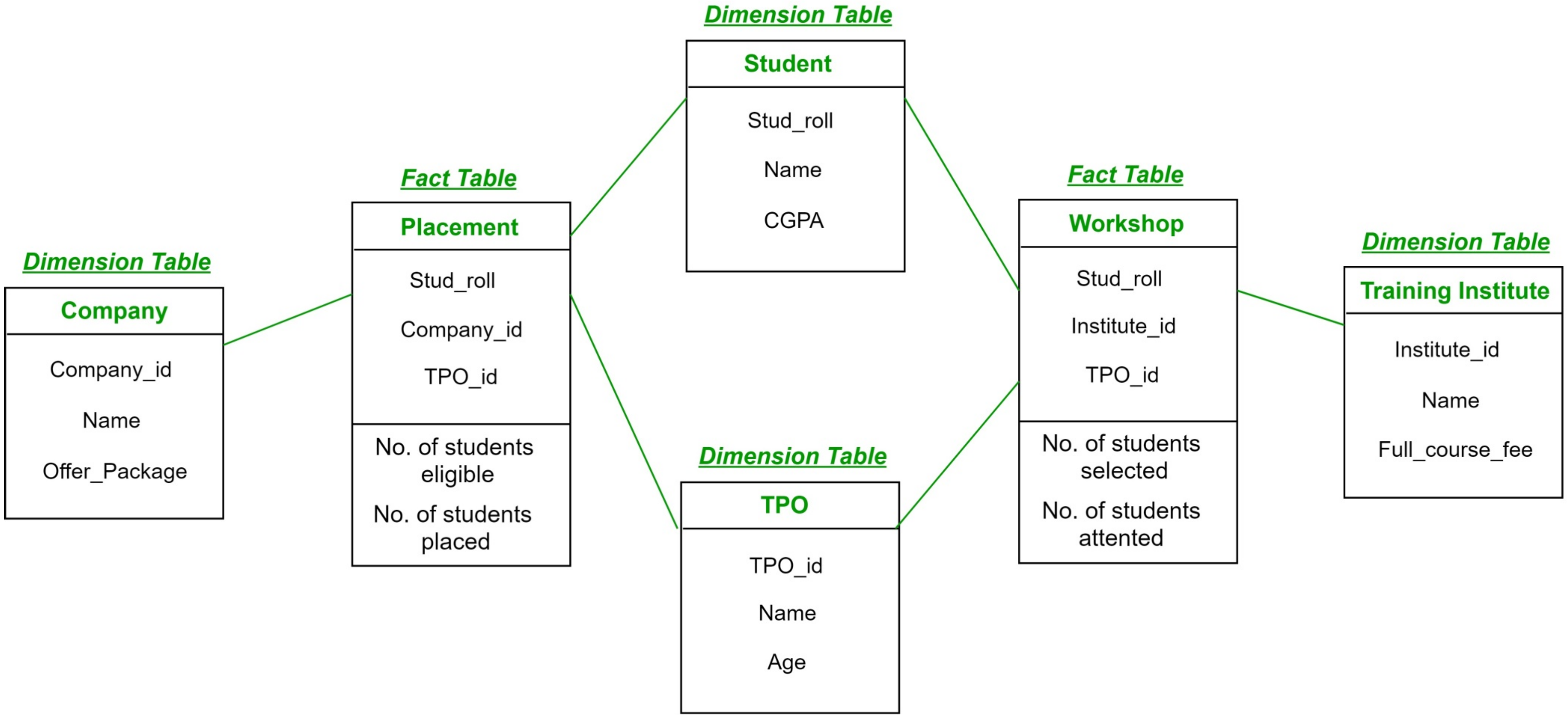
- **Schemas –**

  3. **Fact constellation:**
     - Can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation
     - Commonly used, since it can model multiple, interrelated subjects
     - The widely used schema for Data warehouse designing and it is much more complex than star and snowflake schema
     - Provides a flexible schema
     - More complex and hence, hard to implement and maintain

Data Warehouse Modeling

- **Schemas –**

  3. Fact constellation:



https://www.geeksforgeeks.org/fact-constellation-in-data-warehouse-modelling/

Data Warehouse Modeling

- **Typical OLAP Operations**
  1. **Roll-up:** Aggregate certain similar data attributes having the same dimension together
     - ➢ Eg., Customer daily income can use a roll-up operation to find the monthly income of his salary

  2. **Drill-down:** Allows us to take particular information and then subdivide it further for coarser granularity analysis
     - ➢ It zooms into more detail
     - ➢ Eg., In Country column; splits India into states, districts, towns, cities, villages

Data Warehouse Modeling

- **Typical OLAP Operations**

3. **Slicing:** Filters the unnecessary portions
    - User doesn’t need everything for analysis, rather a particular attribute
    - Eg., country=“India”, will display only about India

4. **Dicing :** Does a multidimensional cutting
    - Not only cuts only one dimension but also can go to another dimension and cut a certain range of it
    - Eg., Annual salary of Karnataka state employees

Data Warehouse Modeling

- **Typical OLAP Operations**
  5. **Pivot:** Basically transforms the data cube in terms of view
     - ➢ Doesn’t change the data present in the data cube
     - ➢ Eg., During, comparing year versus branch; user can change the viewpoint and compare branch versus item type

- **Advantages of data cubes:**
  - ➢ Data cubes store large data in a simple way
  - ➢ Helps in giving a summarized view of data
  - ➢ Provides quick and better analysis
  - ➢ Improve performance of data

Data Warehouse Design and Implementation

- **Data warehouse design – 4 views**
  1. **Top-down view** allows the selection of the relevant information necessary for the data warehouse
     - ➢ This information matches current and future business needs
     - ➢ Used by the company CEOs/ people who knows every aspects of business
  2. **Data source view** exposes the information being captured, stored and managed by operational systems
     - ➢ Then decides how to select and understand each piece of information
     - ➢ Data sources are often modeled by traditional data modeling techniques, such as the entity-relationship model or CASE (computer-aided software engineering) tools

Data Warehouse Design and Implementation

- **Data warehouse design – 4 views**

  3. **Data warehouse view** includes fact tables and dimension tables
     - ➢ Represents the information that is stored inside the data warehouse
     - ➢ Depending on the size and the complexity of data decides what data schemas can be used to organize the data
  4. **Business query view** is the data perspective in the data warehouse from the end-user's viewpoint
     - ➢ Supports variety of queries that user need to use in an effective and efficient way to represent the data

Data Warehouse Design and Implementation

- **Warehouse design process steps**

  Eg., All electronics

  - ➢ Planning: Specific, achievable and measurable
    - ▪ Without a proper planning data warehouse can take years to build
  - ➢ Requirement study
  - ➢ Problem analysis
  - ➢ Warehouse design
  - ➢ Data integration and testing
  - ➢ Development of data warehouse

Data Warehouse Design and Implementation

- **Warehouse design process**
  1. Top-down/ bottom-up approach
     - **Top-down:** Starts with overall design and planning (for mature enterprise)
     - **Bottom-up:** Starts with experiments and prototypes (rapid)
  2. Software engineering point of view
     - **Waterfall:** Structure and systematic analysis at each step before proceeding to the next
     - **Spiral:** Rapid generation of increasingly functional systems, short turn around time (iteration)
     - **Agile:** Do not assume that specifications are complete or even stable

Data Warehouse Design and Implementation

- **Warehouse design process**
  3. Typical warehouse design process
     - Choose a business process to model; Eg., orders, invoices etc.
     - Choose the grain (atomic level of data) of the business process; Eg., individual transactions/ daily snapshots
     - Choose the dimensions that will apply to each fact table record; i.e, time, item, customer, supplier and transaction type
     - Choose the measure that will populate each fact table record; Typical measures are numeric additive quantities like dollars_sold and units_sold

Data Warehouse Design and Implementation

- **Data warehouse usage**
  - **Information processing**
    - Supports querying, basic statistical analysis and reporting using tables, charts and graphs
  - **Analytical processing**
    - Multidimensional analysis of data warehouse data
    - Supports basic OLAP operations
  - **Data mining**
    - Knowledge discovery form hidden patterns
    - Supports associations, Constructing analytical models, performing classifications and predictions and presenting using visualization tools

Data Warehouse Design and Implementation

- **Data warehouse implementation**
  - Represented by data cubes
  - Things to consider during implementation
    1. Efficient cube computation technique
    2. Access methods
    3. Query processing techniques

Data Warehouse Design and Implementation

- **Data warehouse implementation**
  1. Efficient cube computation technique
     - Pre computation of all part of the data cube inorder to reduce the response time and improve the performance
     - Compute cube operator computes aggregates of overall subsets of the dimensions specified in the operation
     - Curse of dimensionality is a problem occurs when many of the dimensions have associated concept hierarchies
     - Total number of cuboids for n dimensional data cube is 2^n
     - As the number of dimensions, number of conceptual hierarchies increase the storage space required will exceeds the actual data
     - To overcome the curse of dimensionality materialization of cuboid is important

Data Warehouse Design and Implementation

- **Data warehouse implementation**
  1. Efficient cube computation technique
     - Cube materialization
       - I. No materialization – Full cuboid is used and no computational techniques
       - II. Full materialization – Full cuboid is used and certain computational techniques will be used
       - III. Partial materialization – Select only the relevant cuboids for query process
     - Cuboid selection
       - I. Frequently referred cuboids by previous query process
       - II. Iceberg method – Set a threshold value to the cuboid and which all sets that select them for query process
       - III. Updated the materialized cuboid

Data Warehouse Design and Implementation

- **Data warehouse implementation**
  2. Access method – 2 types of structured methods
     1. Bitmap index
        - Value in the dimension will be represented as one bit otherwise zero
        - Whenever we search a value in the cuboid wherever we have bit 1 we can access that value
        - Reduce the response time
     2. Join index
        - If we have more than one cuboid access then use groupby option
        - Join the dimensions to execute the query

Data Warehouse Design and Implementation

- **Data warehouse implementation**
  3. Query processing techniques
     - Use the materialized cuboid and perform the OLAP operations to speedup the execution
       - Selection of relevant materialized cuboid
       - Selection of proper operation to be performed on that cuboid

Video Tutorials

- Understanding Data Mart
- Big Data vs Data Science vs Data Analytics