MODULE- 1

Data Mining:- Concepts and Applications, Data Mining Stages, Data Mining Models, Data Warehousing (DWH) and On-Line Analytical Processing (OLAP), Need for Data Warehousing, Challenges, Application of Data Mining Principles, OLTP Vs DWH, Applications of DWH.

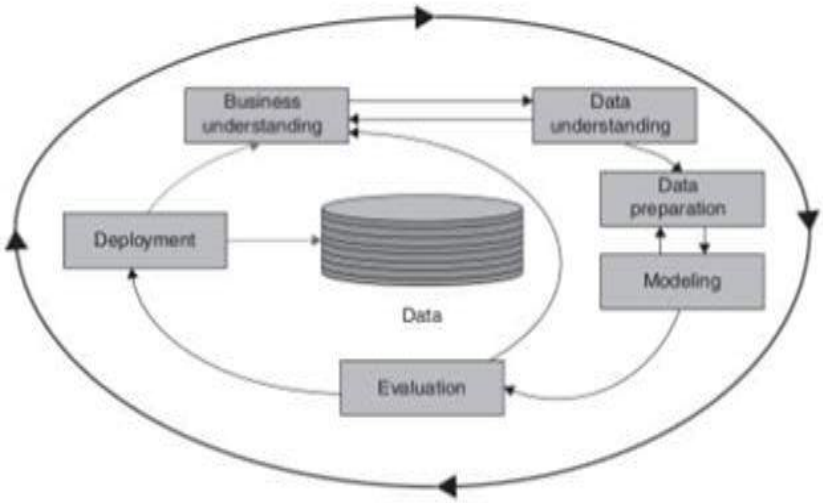
Data mining

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. "We are living in the information age" is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business society, science and engineering, medicine, and almost every other aspect of daily life

What Is Data Mining?

Data mining refers to extracting or mining knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Data mining stages



Data mining stages

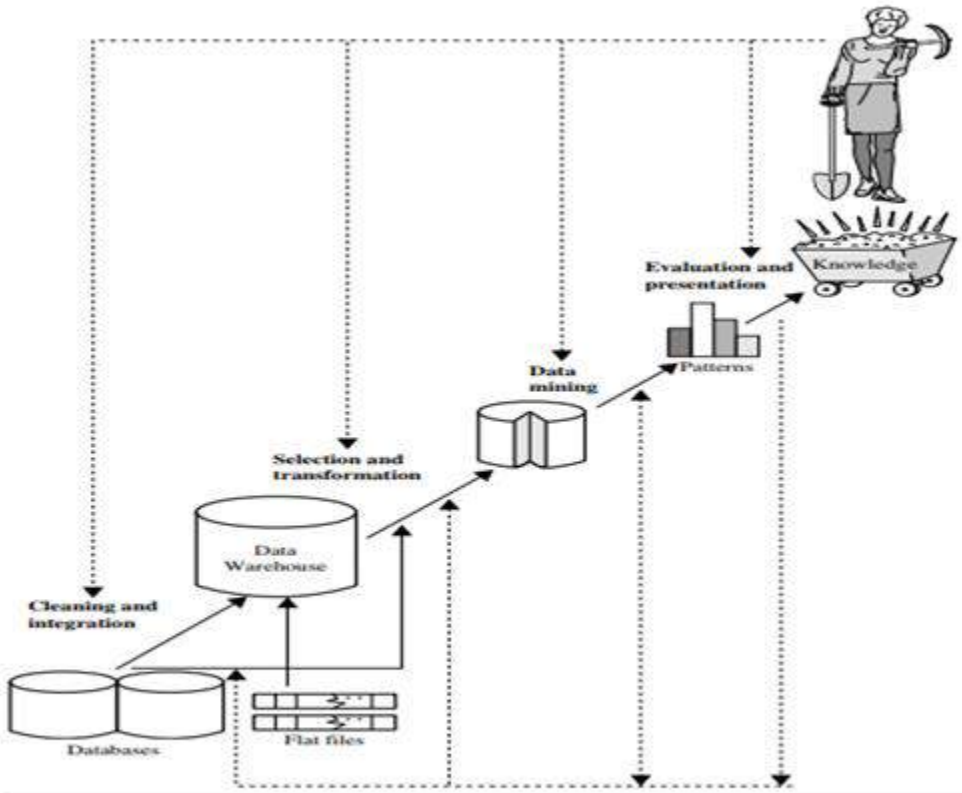
Business understanding- involves understanding the domain for which data mining has to be performed. Domain can be financial, educational data domain etc:
Once the domain is understood properly, the domain data has to be understood next. Here the relevant data in the needed format will be collected and understood.
Data preparation (data preprocessing) involves data cleaning, data integration, data selection and data transformation). It is the important step in the sense that the data is to be made suitable for further processing and mining.

After data preprocessing, data will undergo mining (modeling) and then passes through the stage pattern evaluation and finally it should be deployed. (Knowledge presentation).

KDD steps – STEPS IN KNOWLEDGE DISCOVERY FROM DATA

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The terms knowledge discovery in databases (KDD) and data mining are often used interchangeably.
Over the last few years KDD has been used to refer to a process consisting of many steps, while data mining is only one of these steps.

Knowledge discovery in databases (KDD) is the process of finding useful information and patterns in data. Data mining is the use of algorithms to extract the information and patterns derived by the KDD process.



Data mining as a step in the process of knowledge discovery.

Knowledge discovery as a process is depicted in Figure 1.4 and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

Architecture of a data mining system/ components

The architecture of a typical data mining system may have the following major components: Database, data warehouse, World Wide Web, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

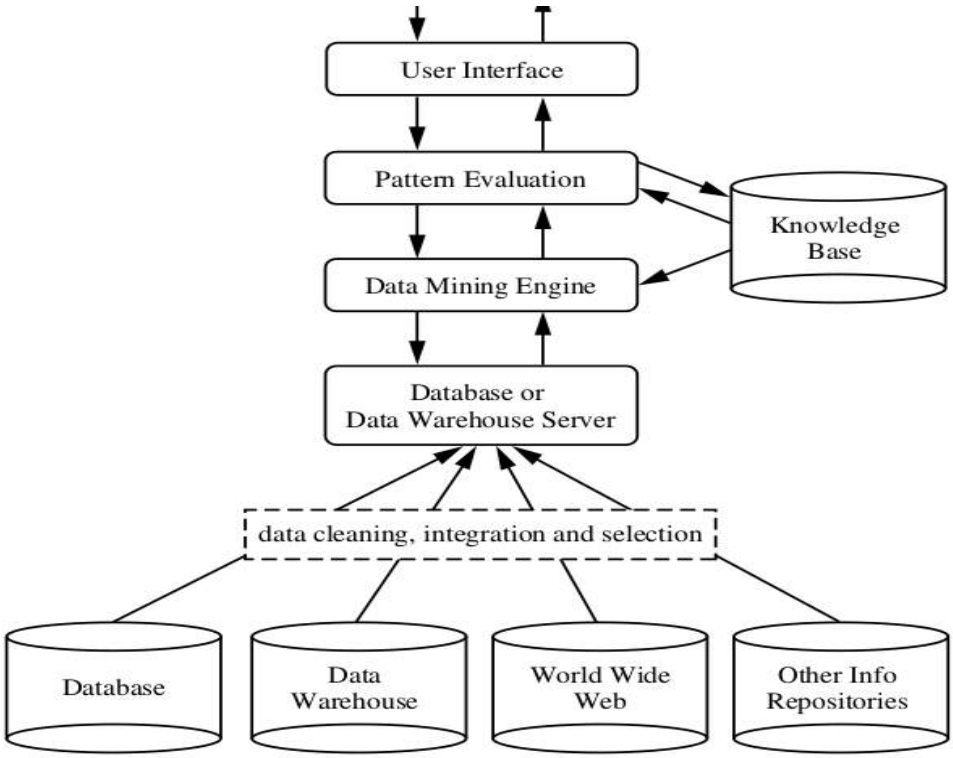
Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include con- cept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures (and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.



Data mining tasks/functionalities/ applications:

1. Classification: The goal is to classify a new data record into one of the many possible classes, which are already known.
   Eg: In loan database, to classify an applicant as a prospective or defaulter, given his various personal and demographic features along with previous purchase characteristics.
2. Estimation: Predict the attribute of a data instance. Eg: estimate the percentage of marks of a student, whose previous marks are already known.
3. Prediction: Predictive model predicts a future outcome rather than the current behavior. Eg: Predict next week's closing price for the Google share price per unit.
4. Market basket analysis(association rule mining)
   Analyses hidden rules called association rule in a large transactional database.
   {pen, pencil-> book} – Whenever pen and pencil are purchased together, book is also purchased.
5. Clustering Classification into different classes based on some similarities but the target classes are unknown.
6. Business intelligence
   Business intelligence technologies provide historical, current and predictive views of business operations. Without data mining, many businesses may not be able to perform effective market analysis, compare customer feedback on similar products, discover strength and weaknesses of their competitors, retail highly valuable customers and make

smart business decisions. Data mining is the core of business intelligence. Classification and prediction techniques are the core of predictive analytics in business intelligence. Clustering plays a central role in customer relationship management, which groups customers based on their similarities.

7. Web search engines
   A web search engine is a specialized computer server that searches for information on web. Various data mining techniques are used in all aspects of search engines ranging from crawling (e.g.: deciding which pages should be crawled and the crawling frequency), indexing (e.g.: selecting pages to be indexed and deciding to which extend the index should be constructed) and searching (e.g.: deciding how pages should be ranked, which advertisements should be added etc. :)
8. Business data analytics
9. Bioinformatics
10. Web mining
11. Text mining
12. Social network data analysis

Data mining models

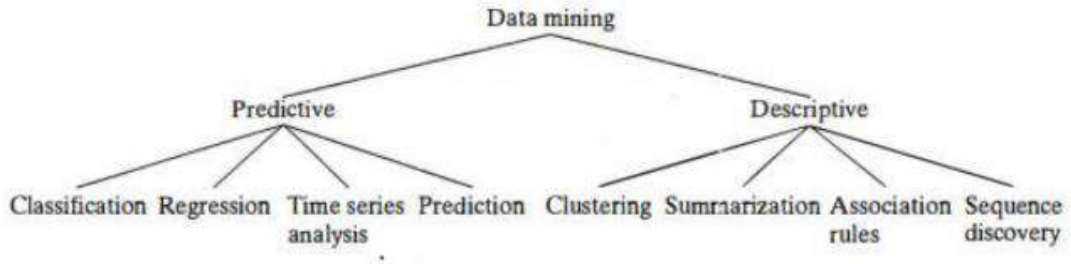
Data mining involves many different algorithms to accomplish different tasks. All these algorithms attempt to fit a model to the data.

The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined.

Data mining algorithms can be characterized as consisting of three parts

- Model- The purpose of the algorithm is to fit a model to the data.
- Preference- Some criteria must be used to fit one model over another.
- Search- All algorithms require some technique to search the data.

Data mining models can be either predictive model or descriptive model.



Data mining models and tasks.

A predictive model makes a prediction about values of data mining known results found from different data. Predictive modeling may be made based on the use of other historical data. Predictive modeling data mining tasks include classification, regression, time series analysis, prediction.

A descriptive model identifies patterns or relationships in data. Unlike predictive model, a descriptive model serves as a way to explore the properties of the data being examined, not to predict new properties. Clustering, summarization, association rules and sequence discovery.

DATA WAREHOUSE

Data Warehouse: Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse refers to a database that is maintained separately from an organization's operational databases. A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process"

1. **Subject-oriented**: A data warehouse is organized around major subjects, such as customer, supplier, product, and sales. A data warehouse focuses on the modelling and analysis of data for decision makers (not on day to day transaction). Provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
2. **Integrated**: data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
3. **Time-variant**: Data are stored to provide information from a historical perspective. Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
4. **Non-volatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: *initial loading of data and access of data.*

Data warehousing is the process of constructing and using data warehouses.

- The construction of a data warehouse requires data cleaning, data integration, and data consolidation.
- The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge workers" (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse.

Data warehousing is very useful from the point of view of heterogeneous database integration. The traditional database approach to heterogeneous database integration was a 'query- driven' approach data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.

Difference between Operational Database systems and Data Warehouse

- Operational Database systems
  - Main task is to perform on-line transaction and query processing. These systems are called **on-line transaction processing (OLTP)** systems.

- They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

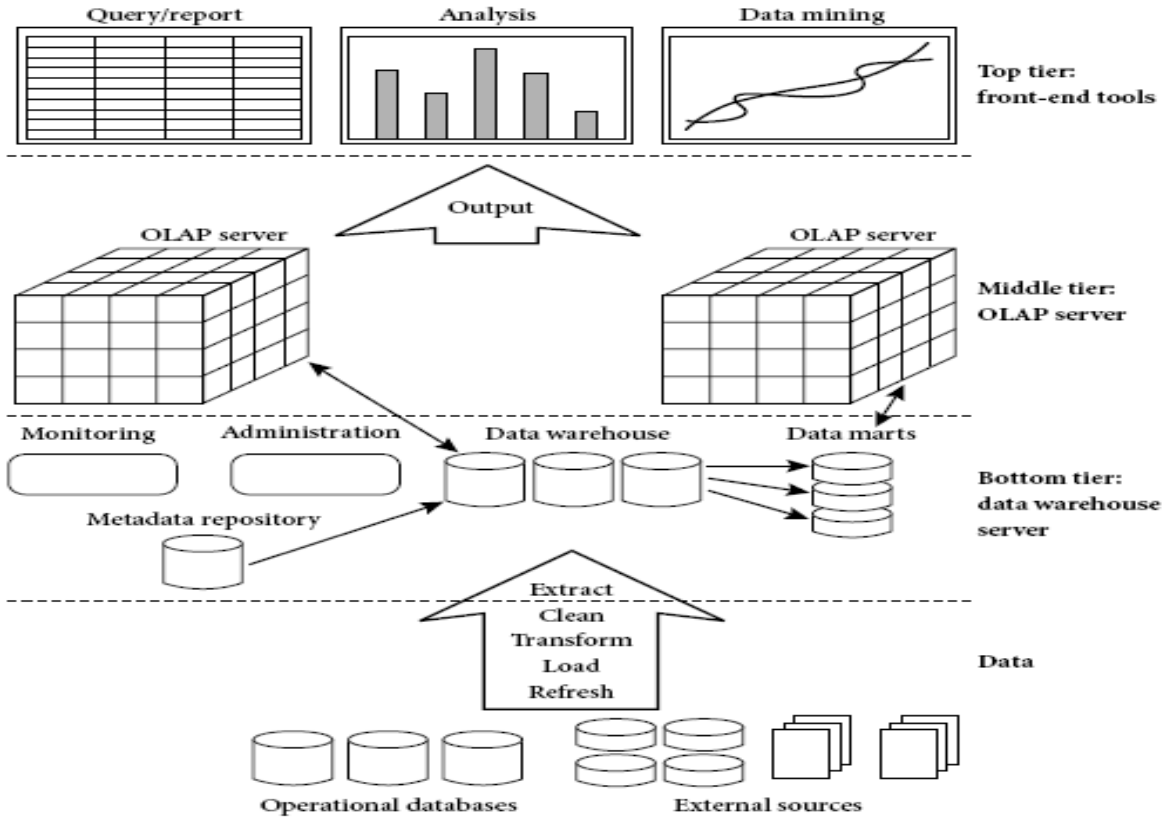
- Data Warehouse
  - Serve users or knowledge workers in the role of data analysis and decision making.
  - Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as **on-line analytical processing (OLAP)** systems.

Difference between OLTP and OLAP

- **Users and system orientation**:
  - OLTP system is *customer-oriented and is used for* transaction and query processing by clerks, clients, and information technology professionals.
  - OLAP system is *market-oriented and is used for data analysis by knowledge* workers, including managers, executives, and analysts.
- **Data contents**:
  - OLTP system manages current data
  - OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.
- **Database design**:
  - An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
  - An OLAP system typically adopts either a *star or snowflake model and a subject oriented* database design.
- **View**:
  - An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.
  - An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
  - OLAP systems also deal with information that originates from different organizations.
  - OLAP data are stored on multiple storage media.
- **Access patterns**:
  - The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.
  - Accesses to OLAP systems are mostly read-only operations although many could be complex queries.

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., 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	historical; 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
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	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data warehouse multi-tier/ three-tier architecture



Data warehouse adopt a three tier architecture, these are:-

1. Bottom Tier (Data warehouse server)
2. Middle Tier (OLAP server)
3. Top Tier (Front end tools)

Bottom tier

- Is a warehouse database server (almost a relational dbase system)
- Data is fed using Back end tools and utilities.
- It also contains Meta data repository.

  Data Warehouse Back-End Tools and Utilities

  - **Data extraction**, which typically gathers data from multiple, heterogeneous, and external sources
  - **Data cleaning**, which detects errors in the data and rectifies them when possible
  - **Data transformation**, which converts data from legacy or host format to warehouse format
  - **Load**, which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions
  - **Refresh,** which propagates the updates from the data sources to the warehouse

  Metadata Repository

  - Data about data.
  - In a data warehouse, metadata are the data that define warehouse object

  Data mart

  Data mart is a department subset of data warehouse that focuses on selected subjects, and thus its scope is department wide. Star or snowflake schema is commonly used for data mart.

Middle Tier

The middle tier is an OLAP server that is typically implemented using either

(1) A relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations;

(2) A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

(3) HOLAP- hybrid OLAP

Top tier

- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.

Data warehouse modeling

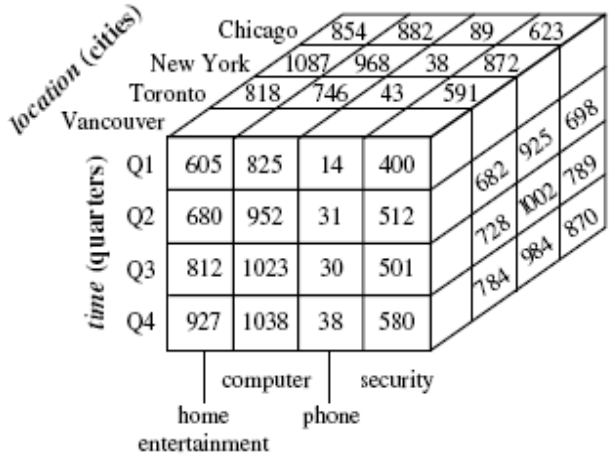
- Data warehouses and OLAP tools are based on multidimensional data models.
- A data cube allows data to be modeled and viewed in multiple dimensions defined by dimensions and facts.
- Dimensions are the perspectives or entities with respect to which an organization wants to keep records.
- Dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold.

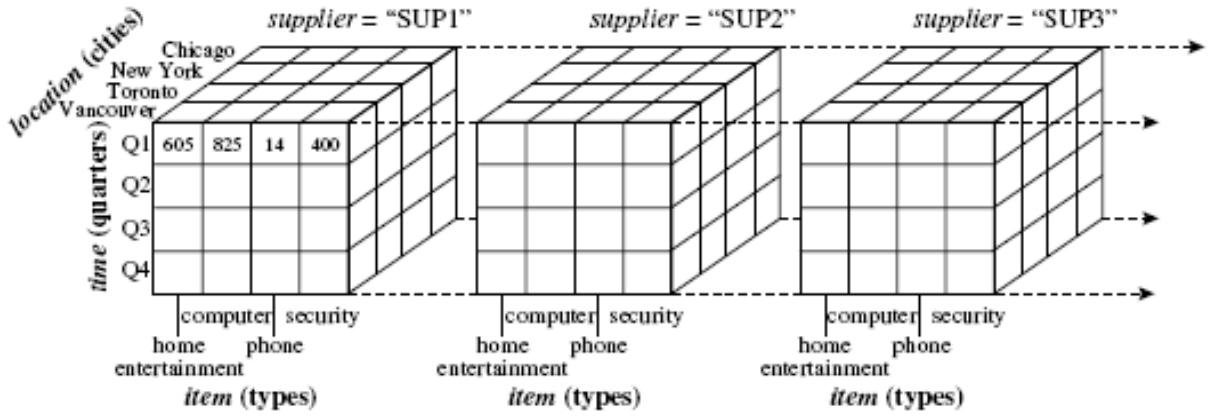
- Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
- Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
- Facts are numerical measures→ the quantities by which we want to analyze relationships between dimensions.
    - Fact table contains name of facts (such as dollars_sold) and keys to each of the related dimension tables.

A 2-D view of sales data for *AllElectronics* according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

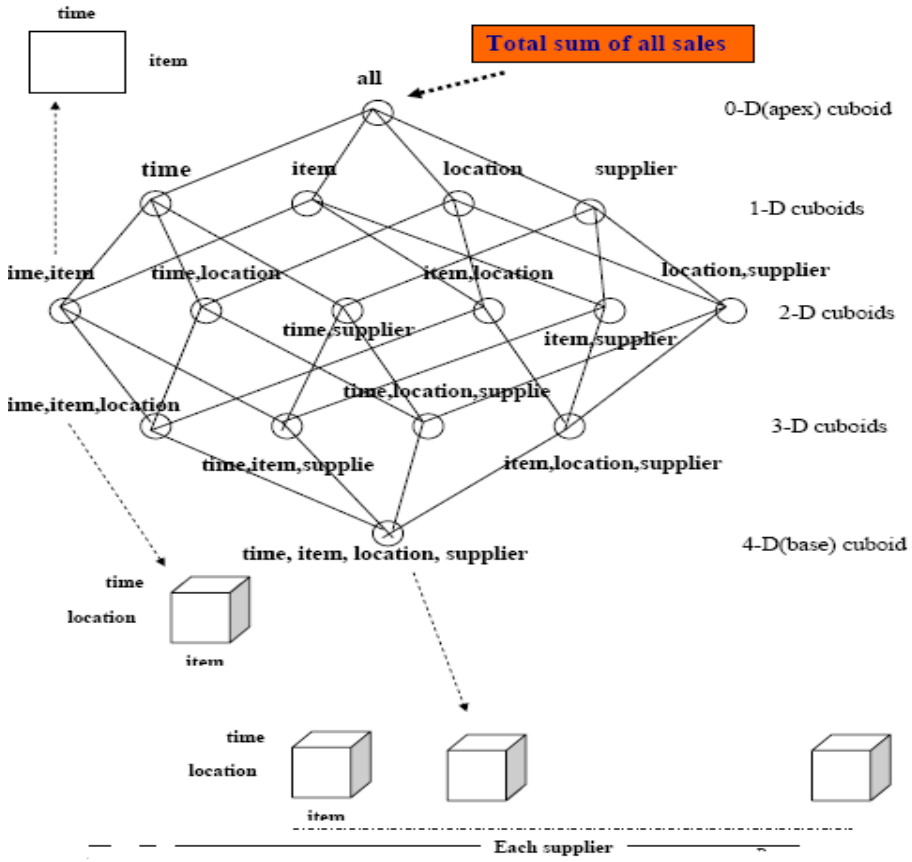
	location = "Vancouver"			
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

	location = "Chicago"				*location* = "New York"				*location* = "Toronto"				*location* = "Vancouver"			
	item				*item*				*item*				*item*			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580





- In data warehousing literature, an n-D base cube is called a base cuboid (returns the total sales for any combination of dimensions). The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid.
- The lattice of cuboids forms a data cube Can generate cuboid for each of the possible subsets of the given.



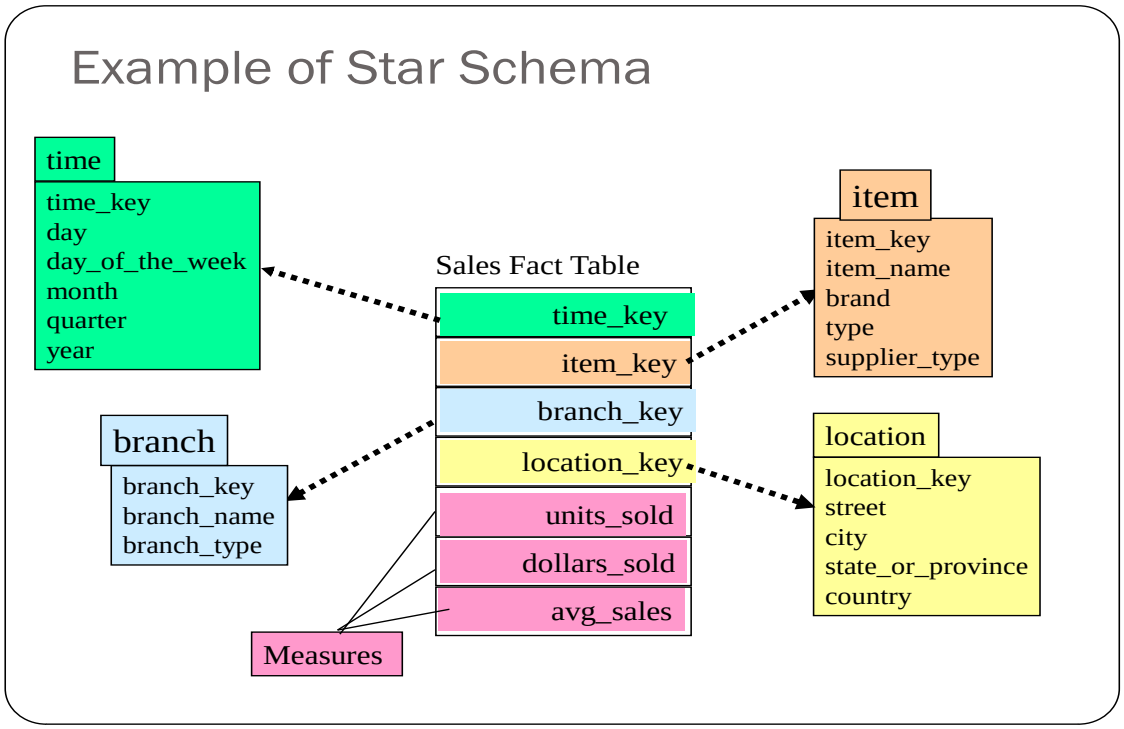
Schema for multidimensional data model

The most popular data model for a data warehouse is a multidimensional model which can exist in the form of a star schema, snow flake schema or a fact constellation schema.

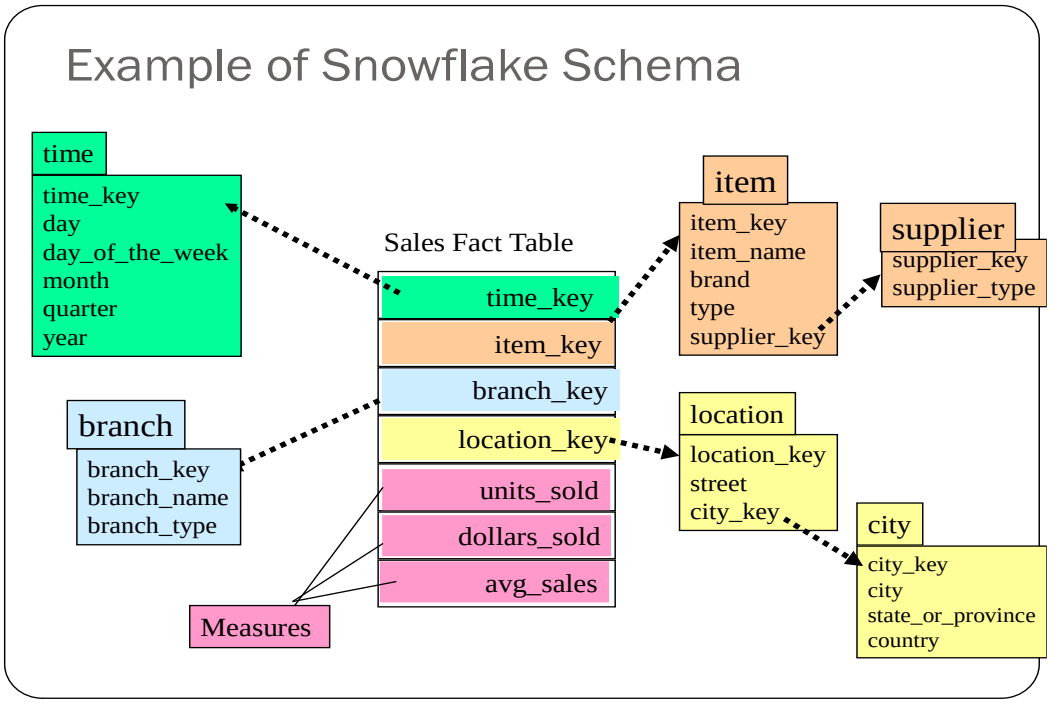
- Star schema:
  In this schema a data warehouse contains
  - A large central fact table containing the bulk of data with no redundancy and
  - A set of smaller attendant tables called dimension tables for each dimension.

The schema graph resembles a star burst with the dimension tables displayed in radial pattern around the central fact table.

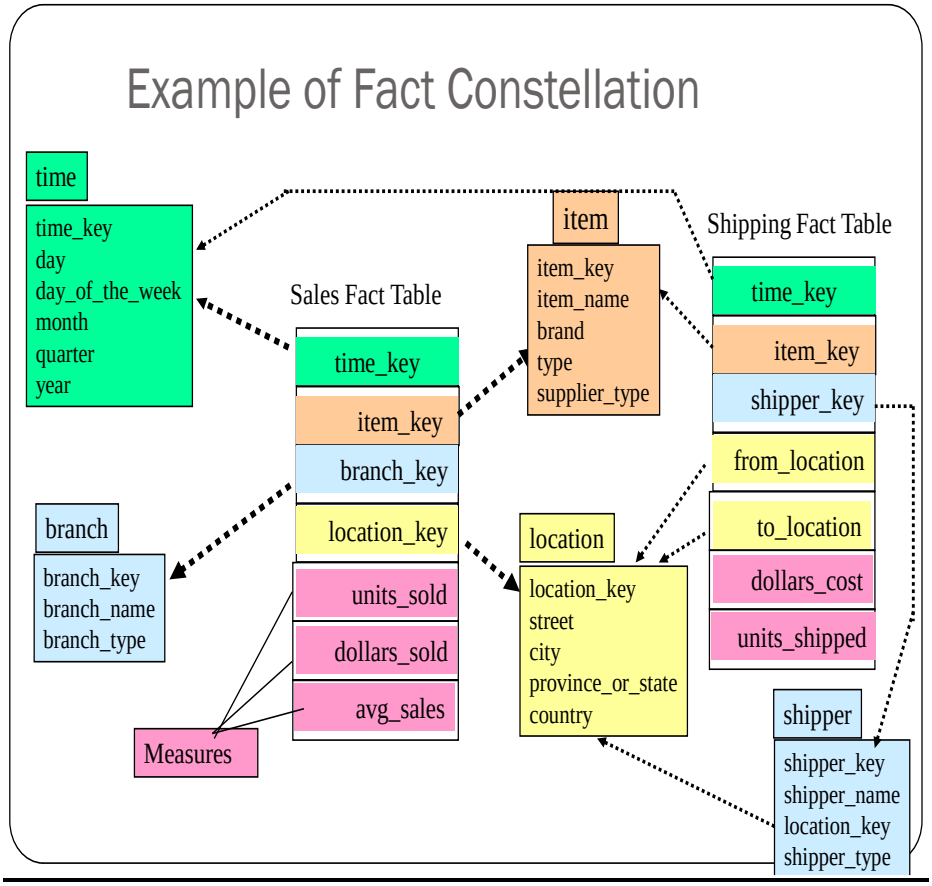
Each dimension is represented by only one table and each table contains a set of attributes. This constraint may introduce some redundancy. For example if we have a branch in two cities of the same country, it will create redundancy among the country attribute in location dimension table.



- Snowflake schema:
  A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake. The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space. However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure 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, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

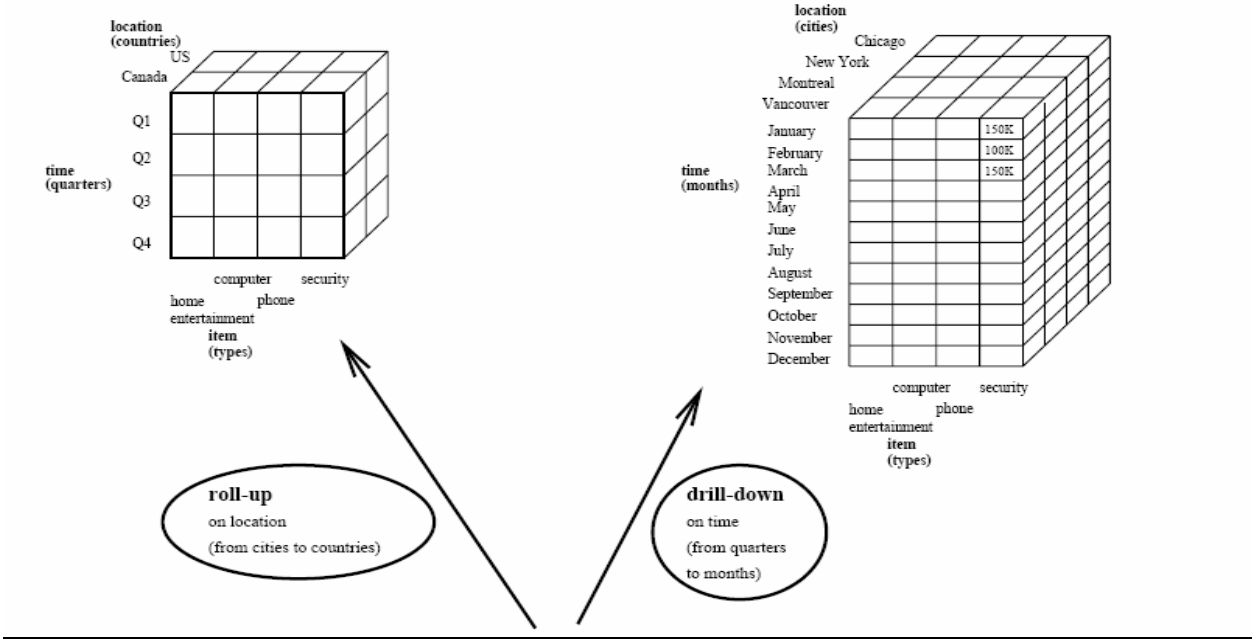


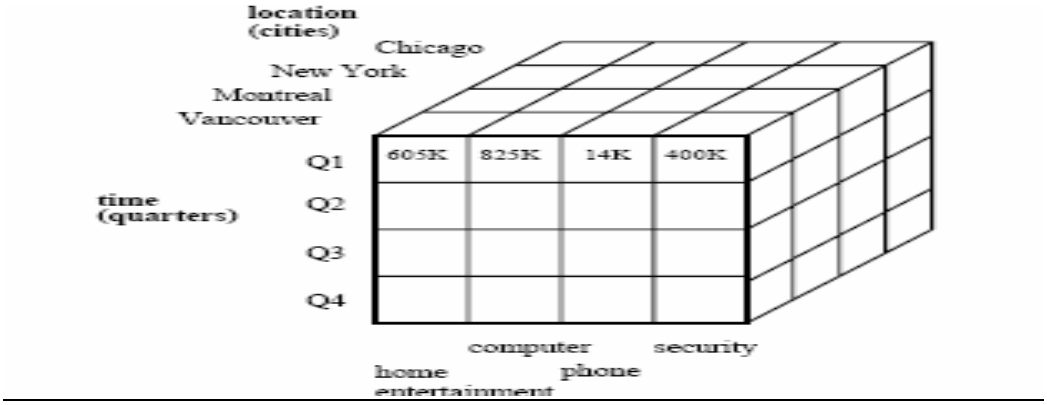
- Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

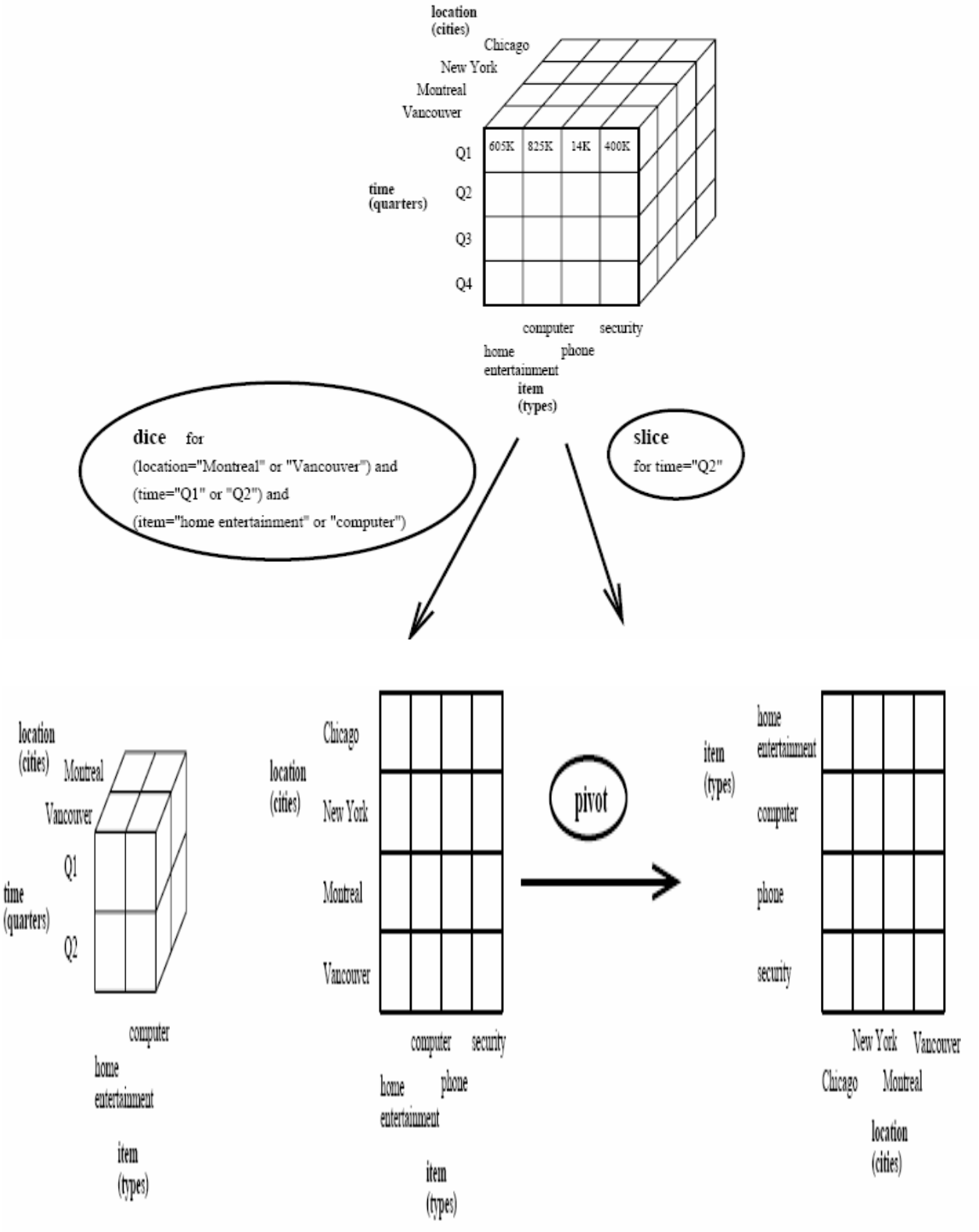


OLAP operations/ Data cube operations

1. Roll-up/ drill up
   Performs aggregation on a data cube by climbing up hierarchy or by dimension reduction
2. Drill down (roll down): reverse of roll-up. Stepping down a concept hierarchy from higher level summary to lower level summary or detailed data, or introducing new dimensions.
3. Slice and dice: *select* & *project*
   Slice → selection on one dimension of the cube.
   Dice → selection on 2 or more dimensions.
4. Pivot
   It is also known as rotate. It rotates the data axis to view the data from different perspectives. Eg: Item and location axes in a 2 D slice are rotated.






location
(cities)
Chicago
New York
Montreal
Vancouver
605K
825K
14K
400K
Q1
Q2
Q3
Q4
time
(quarters)
computer
security
home
entertainment
phone
item
(types)
dice for
(location="Montreal" or "Vancouver") and
(time="Q1" or "Q2") and
(item="home entertainment" or "computer")
slice
for time="Q2"
location
(cities)
Montreal
Vancouver
Q1
Q2
time
(quarters)
computer
home
entertainment
item
(types)
Chicago
New York
Montreal
Vancouver
location
(cities)
computer
security
home
entertainment
phone
item
(types)
pivot
item
(types)
home
entertainment
computer
phone
security
New York
Vancouver
Chicago
Montreal
location
(cities)

Data warehouse models

Enterprise warehouse

- An enterprise warehouse collects all of the information about subjects spanning entire organization.
- It typically maintains detailed data as well as summarized data and can range in size from a few gigabytes to hundreds of gigabytes or beyond.
- Implemented on traditional mainframes or parallel computer architecture platforms

Data mart

- Contains a subset of corporate wide data that is value to a specific group of users. Scope is confined to specific selected subjects.
  Eg: A marketing data mart may confine its subject to customer item and sales.
- The data contained in a data mart tend to be summarized.
- It is implemented on low cost departmental servers.

Virtual warehouse

- A virtual warehouse is a set of views over operational database.

Need for data warehousing

1. The data ware house market supports such diverse industries as manufacturing, retail, telecommunications, and health care. Think of a personnel database for a company that is continually modified as personnel are added and deleted... If management wishes determine if there is a problem with too many employees quitting. To analyze this problem, they would need to know which employees have left, when they left, why they left, and other information about their employment. For management to make these types of high-level business analyses, more historical data not just the current snapshot are required.
   A data warehouse is a data repository used to support decision support systems
2. The basic motivation is to increase business profitability. Traditional data processing applications support the day-to-day clerical and administrative decisions, while data warehousing supports long-term strategic decisions.

3. For increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending)

4. For repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine tune production strategies; analyzing operations and looking for sources of profit.

5. For managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

6. The below figure shows a simple view of a data warehouse. The basic components of a data warehousing system include data migration, the warehouse, and access tools. The data

are extracted from operational systems, but must be reformatted, cleansed, integrated, and summarized before being placed in the warehouse.

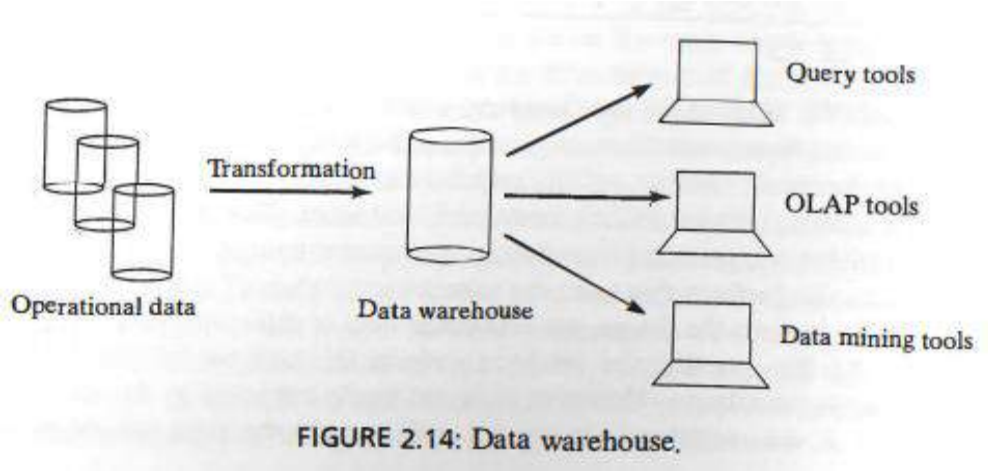


FIGURE 2.14: Data warehouse.

Challenges for Data Warehousing

1. Unwanted data must be removed.
2. Converting heterogeneous sources into one common schema. This problem is the same as that found when accessing data from multiple heterogeneous sources. Each operational database may contain the same data with different attribute names. For example, one system may use "Employee ID," while another uses "EID" for the same attribute. In addition, there may be multiple data types for the same attribute.
3. As the operational data is probably a snapshot of the data, multiple snapshots may need to be merged to create the historical view.
4. Summarizing data is performed to provide a higher level view of the data. This summarization may be done at multiple granularities and for different dimensions.
5. New derived data (e.g., using age rather than birth date) may be added to better facilitate decision support functions.
6. Handling missing and erroneous data must be performed. This could entail replacing them with predicted or default values or simply removing these entries. The portion of the transformation that deals with ensuring valid and consistent data is sometimes referred to as data scrubbing or data staging.
7. Data warehouse queries are often complex. They involve the computation of large groups of data at summarized levels, and may require the use of special data organization, access, and implementation methods based on multidimensional views.
8. **Data Quality** – In a data warehouse, data is coming from many disparate sources from all facets of an organization. When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.

9. **Understanding Analytics** – When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed.
10. **Quality Assurance** – The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader could make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate.
11. **Performance** – Building a data warehouse is similar to building a car. A car must be carefully designed from the beginning to meet the purposes for which it is intended. Yet, there are options each buyer must consider to make the vehicle truly meet individual performance needs. A data warehouse must also be carefully designed to meet overall performance requirements. While the final product can be customized to fit the performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start.
12. **Designing the Data Warehouse** – People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. Usually, there is a high level perception of what they want out of a data warehouse. However, they don't fully understand all the implications of these perceptions and, therefore, have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians building the data warehouse. The typical end result is a data warehouse which does not deliver the results expected by the user. Since the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery.
13. **User Acceptance** – People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization.
14. **Cost** – A frequent misconception among credit unions is that they can build data warehouse in-house to save money.. The harsh reality is an effective do-it-yourself effort is very costly.

Applications of DWH

There are three kinds of data warehouse applications: information processing, analytical processing, and data mining.

1) Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.

2) Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized

and detailed forms. The major strength of online analytical processing over information processing is the multidimensional data analysis of data warehouse data.

3) Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Different areas are:

- **Banking Industry**
  - In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.
  - Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.
  - Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity.
- **Finance Industry**
  - Revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.
- **Consumer Goods Industry**
  - They are used for prediction of consumer trends, inventory management, market and advertising research.
  - In-depth analysis of sales and production is also carried out.
- **Government and Education**
  - The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.
  - The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers.
  - Criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.
  - Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management.
- **Healthcare**
  - All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.
- **Hospitality Industry**
  - A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services.

- They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

- **Insurance**
  - The warehouses are primarily used to analyse data patterns and customer trends, apart from maintaining records of already existing participants.
  - The design of tailor-made customer offers and promotions is also possible through warehouses.
- **Manufacturing and Distribution Industry**
  - A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyse current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.
  - They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyse previous data and customer feedback to evaluate the weaker product lines and eliminate them.
  - For the distributions, the supply chain management of products operates through data warehouses.

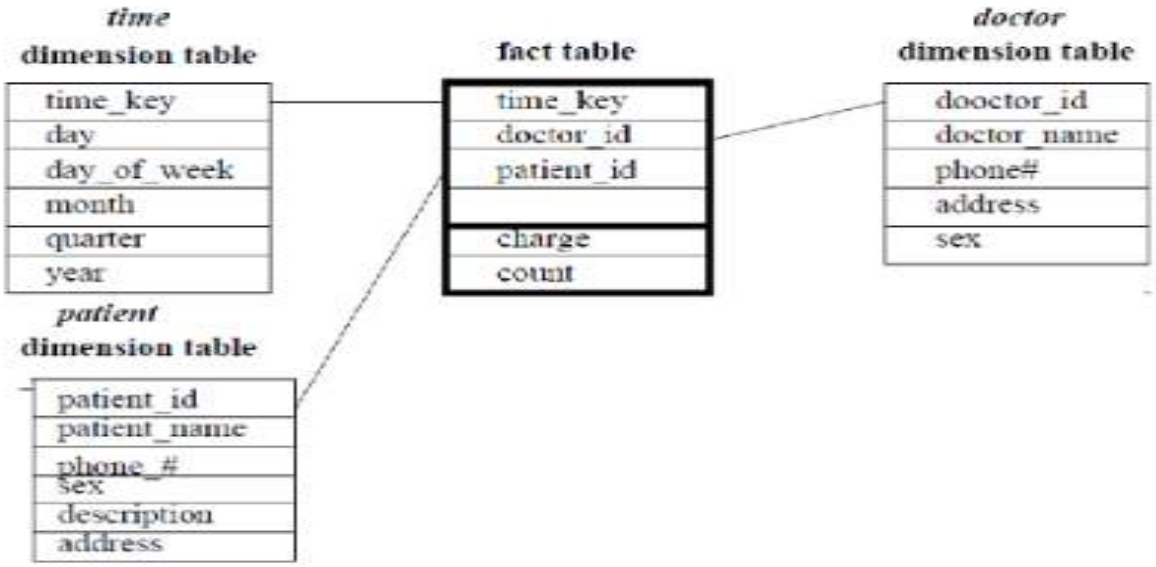
- **The Retailers**
  - Retailers serve as middlemen between producers and consumers.
  - They use warehouses to track items, their advertising promotions, and the consumers buying trends.
  - They also analyse sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.
- **Services Sector**

  Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources
- **Telephone Industry**
  - The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.
  - Analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.
- **Transportation Industry**
  - In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.
  - To analyse customer feedback, performance, manage crews on board as well as analyse customer financial reports for pricing strategies.

Problems on data warehouse

1. Suppose that a data warehouse consists of three dimensions: time, doctor and patient and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.
   a) Draw a star schema for the above data warehouse.
   b) Starting with the base cuboid [day; doctor; patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004.

   Solution (a)

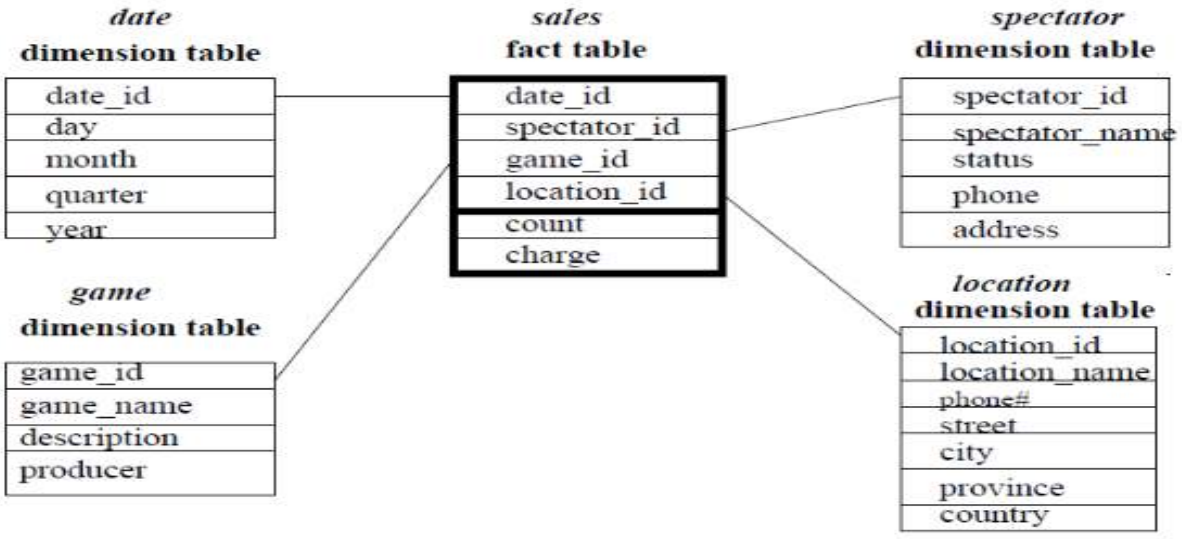


   Solution (b)

   - Roll up on time from day to year
   - Slice for time= 2004
   - Roll up on patient from individual patient (patient_id) to all.

2. Suppose that a data warehouse for University consists of four dimensions date, spectator, location and game and two measures count and charge, where charge is the fare that a spectator pays when watching a game on the given date. Spectator may be students, adults or seniors, with each category having its own charge rate. (May 2019)
   a) Draw a star schema diagram for the data warehouse
   b) Starting with the basic cuboid [date,spectator,location,game] ,what specific OLAP operation should be performed in order to list the total charge paid by student spectators at GM_PLACE in 2010.

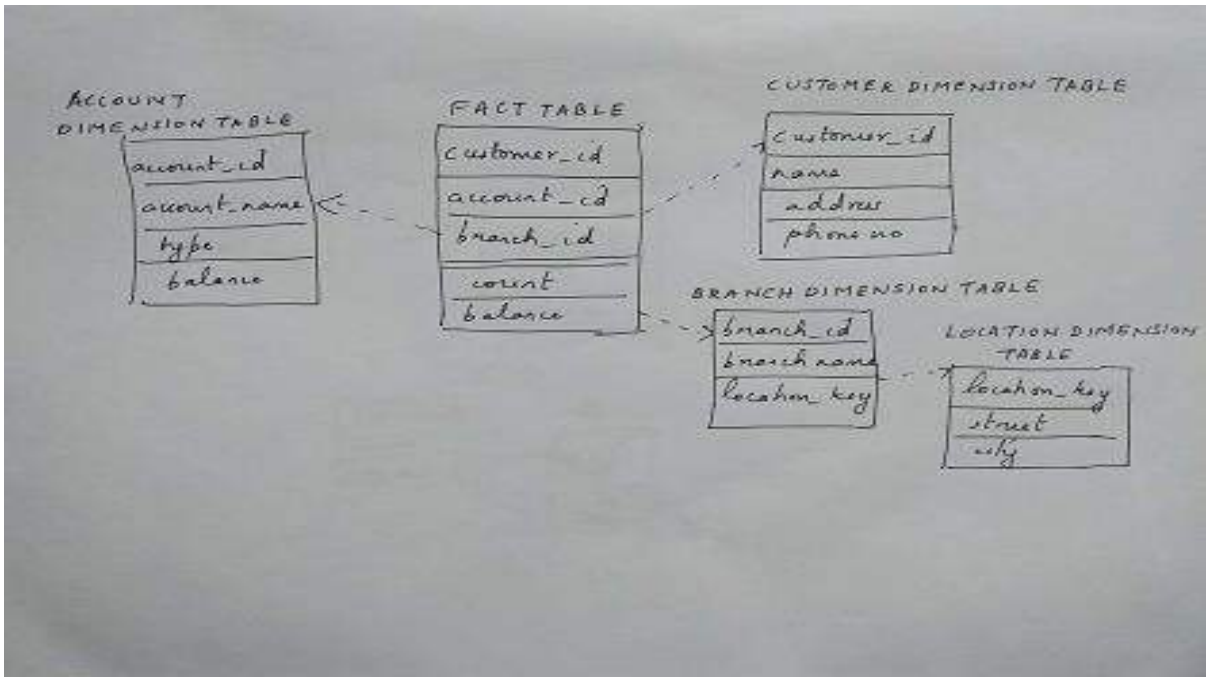
Solution (a)



Solution (b)

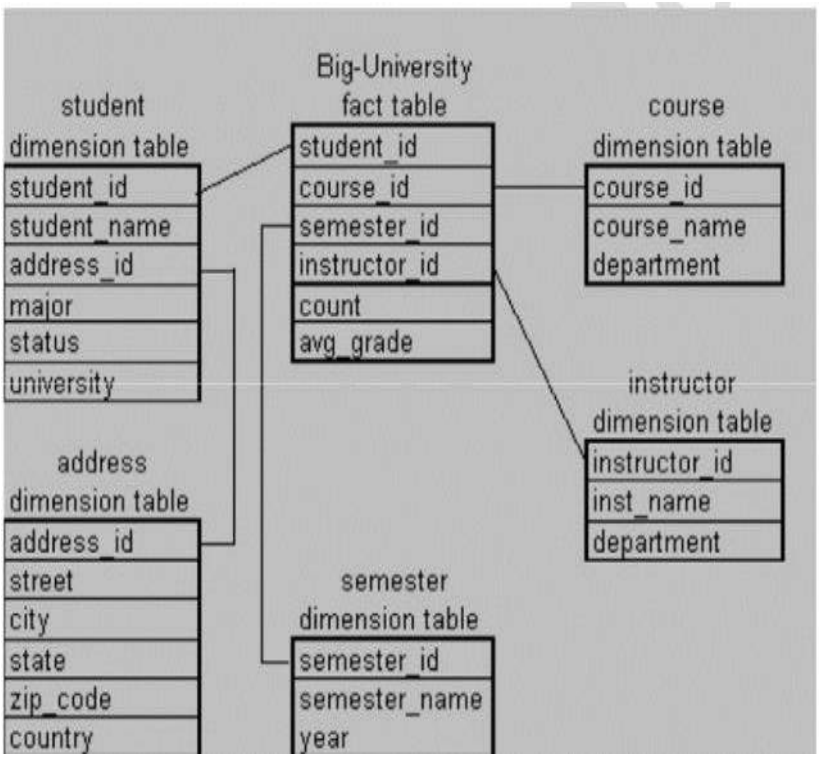
- Roll up on date from date id to year.
- Roll up on spectator from spectator id to status.
- Roll up on location from location id to location name.
- Roll up on game from game id to all.
- Dice with status=‘students’, location name= ‘GM_PLACE’ and year= ‘2010’.

3. Suppose a data warehouse consists of three dimensions customer, account and branch and two measures count (number of customers in the branch) and balance. Draw the schema diagram using snowflake schema. (October 2019)



4. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. (September 2020)
   a) Draw a snowflake schema diagram for the data warehouse.
   b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.

Solution (a)



Solution (b)

1. Roll-up on course from course_id to department.
2. Roll-up on student from student_id to university.
3. Dice on course, student with department ="CS" and university = "biguniversity"
4. Drill-down on student from university to student_name

MODULE 1- UNIVERSITY QUESTIONS

MAY 2019

1. How is data mining related to business intelligence? (4 marks).

2. Differentiate between OLTP and OLAP. (4 marks).

3. Explain various stages in knowledge discovery process with neat diagram. (5 marks).

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

a) Draw a star schema diagram for the data warehouse. (6 marks)

b) Starting with the basic cuboid [date,spectator,location,game] ,what specific OLAP operation should be performed in order to list the total charge paid by student spectators at GM_PLACE in 2010. (3 marks).

OCTOBER 2019

1. How is data warehouse different from a database? How are they similar? (4 marks).

2. Compare star and snowflake dimension table. (4 marks).

3. Suppose a data warehouse consists of three dimensions customer, account and branch and two measures count (number of customers in the branch) and balance. Draw the schema diagram using snowflake schema. (3 marks).

SEPTEMBER 2020

1. List out the four major features of data warehouse as defined by William H. Inmon, the father of da warehousing. (4 marks)

2. Draw a suitable figure that shows data mining as a process of knowledge discovery. (2 marks)

3. Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. (5 marks)

   a) Draw a snowflake schema diagram for the data warehouse.

   b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student.

4. Explain different OLAP operations on multi-dimensional data with suitable examples. (6 marks).

5. A data warehouse can be modeled by either a star schema or a snowflake schema. Describe the similarities and the differences of the two models. (3 marks)

Prepared by

Dr. Hema Krishnan

Assistant Professor Senior Grade, CSE, FISAT