Data Analytics

Data Analytics

- Overview
- Data Warehousing
- Online Analytical Processing
- Data Mining

Overview

- **Data analytics**: the processing of data to infer patterns, correlations, or models for prediction
- Primarily used to make business decisions
  - Per individual customer
    - E.g., what product to suggest for purchase (Mining)
  - Across all customers
    - E.g., what products to manufacture/stock, in what quantity (DSS)
- Critical for businesses today

Overview (Cont.)

- Common steps in data analytics
  - Gather data from multiple sources into one location
  - Data warehouses also integrated data into common schema
  - Data often needs to be **extracted** from source formats, **transformed** to common schema, and **loaded** into the data warehouse
    - Can be done as **ETL (extract-transform-load)**, or **ELT (extract-load-transform)**

Overview (Cont.)

- Generate aggregates and reports summarizing data
  - Dashboards showing graphical charts/reports
  - **Online analytical processing (OLAP) systems** allow interactive querying
  - Statistical analysis using tools such as R/SAS/SPSS
    - Including extensions for parallel processing of big data
- Build **predictive models** and use the models for decision making

Overview (Cont.)

- Predictive models are widely used today
  - E.g., use customer profile features (e.g. income, age, gender, education, employment) and past history of a customer to predict likelihood of default on loan
    - and use prediction to make loan decision
  - E.g., use past history of sales (by season) to predict future sales
    - And use it to decide what/how much to produce/stock
    - And to target customers
- Other examples of business decisions:
  - What items to stock?
  - What insurance premium to change?
  - To whom to send advertisements?

Question16-1: Write some objectives and applications of data analytics for a mobile phone company like TeleTalk, Grameen etc.

Overview (Cont.)

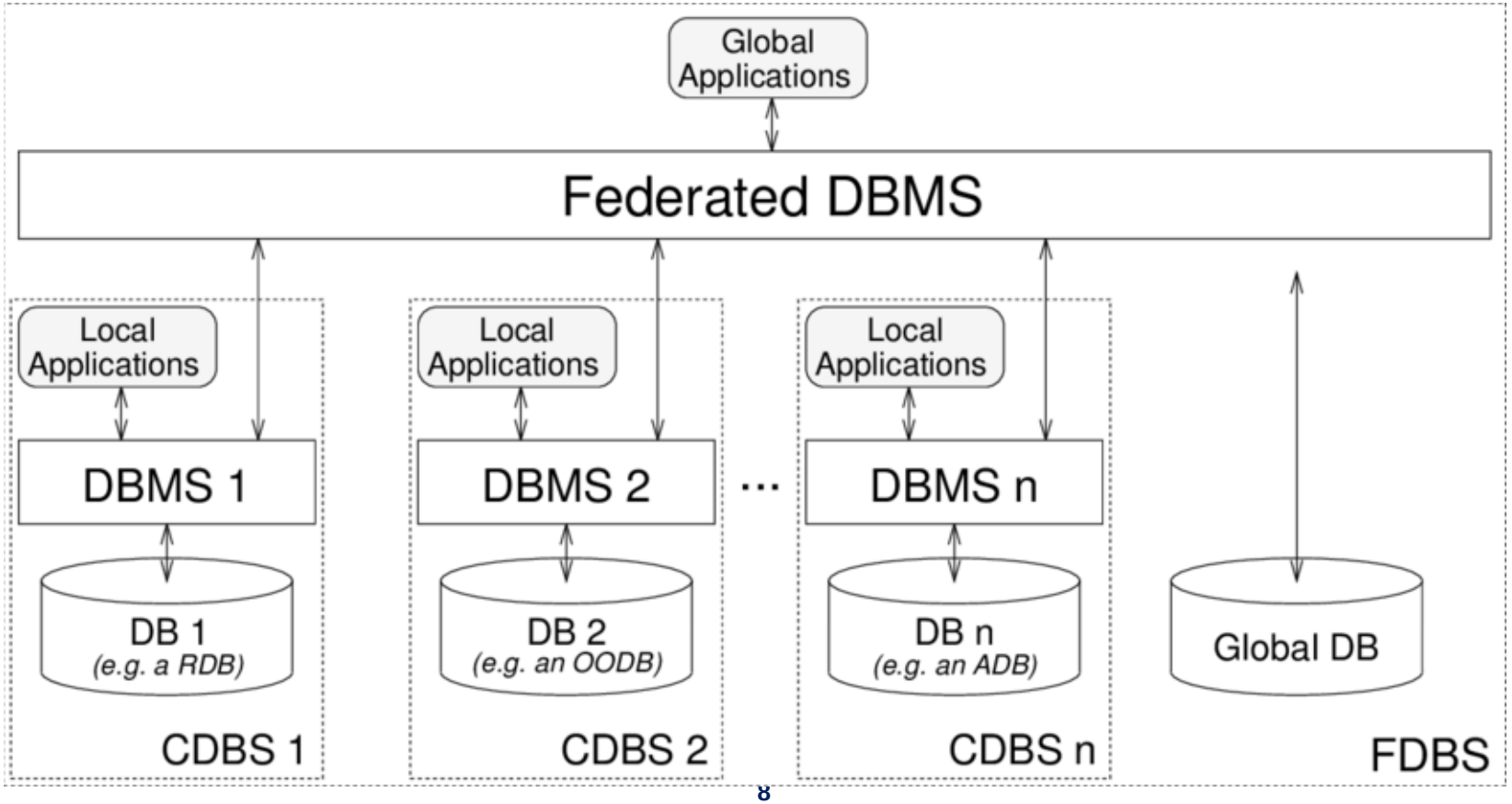
- **Machine learning** techniques are key to finding patterns in data and making predictions
- **Data mining** extends techniques developed by machine-learning communities to run them on very large datasets
- The term **business intelligence (BI)** is synonym for data analytics
- The term **decision support** focuses on reporting and aggregation

Data Integration From Multiple Sources

Many database applications require data from multiple databases

A **federated database system** is a software layer on top of existing database systems, which is designed to manipulate information in heterogeneous databases

Creates an illusion of logical database integration without any physical database integration

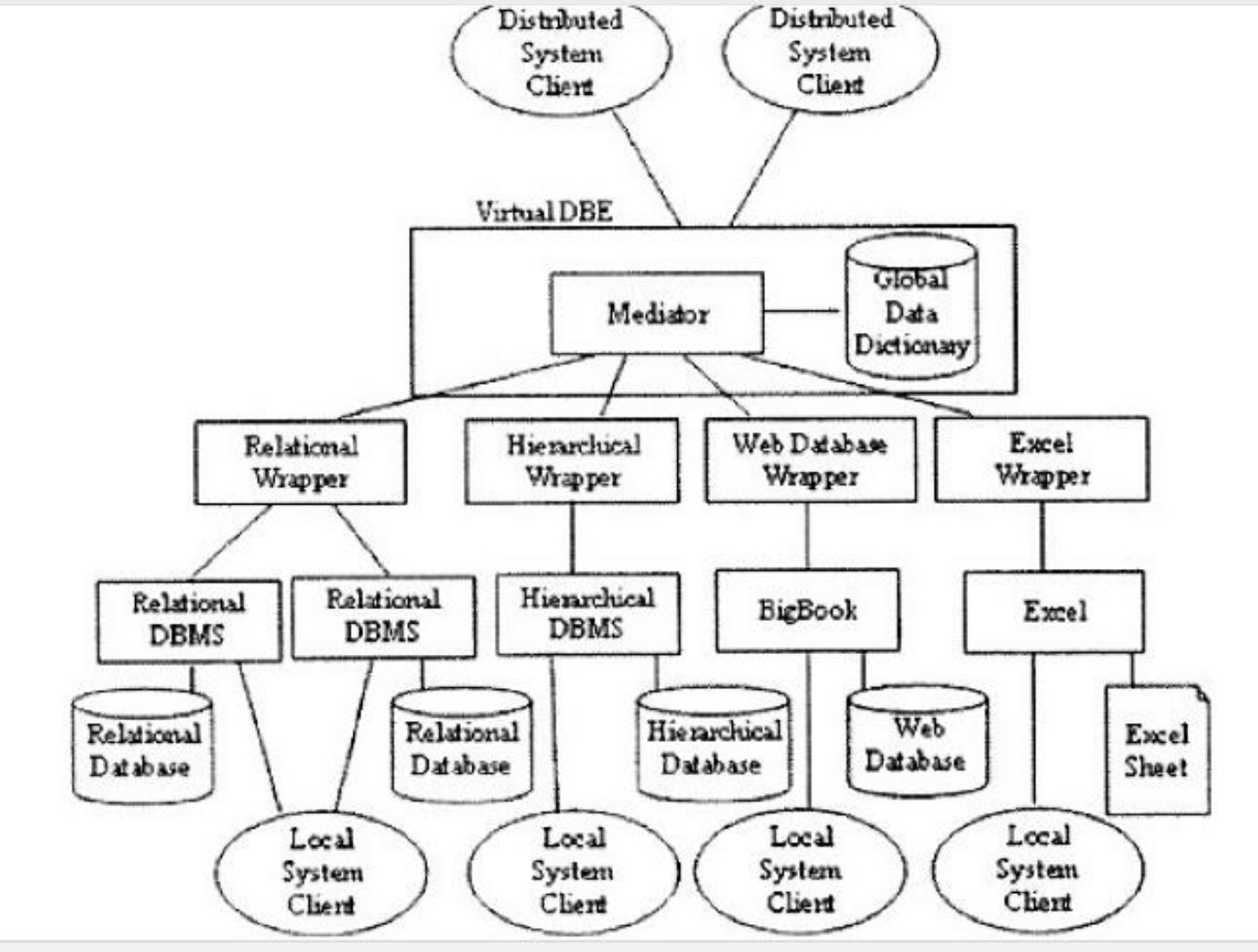


Data Integration From Multiple Sources

- Each database has its **local schema**
- **Global schema** integrates all the local schema
  - **Schema integration**
- Queries can be issued against global schema, and translated to queries on local schemas

Data Integration From Multiple Sources

Wrapper for a data source is a view that translates data from local to a global schema. Wrappers must also translate updates on global schema to updates on local schema



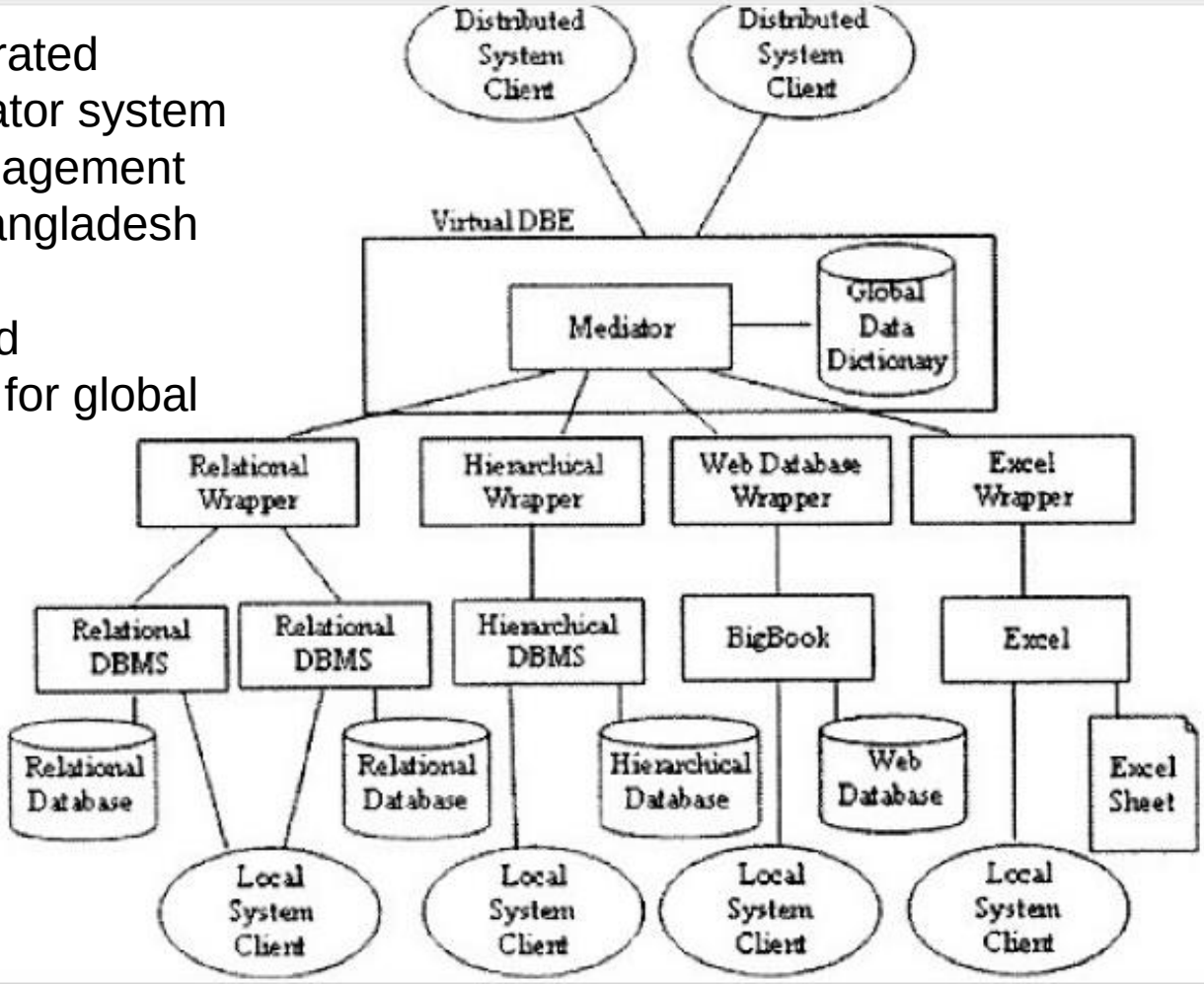
Data Integration From Multiple Sources

- Databases that support common schema and queries, but not updates, are referred to as **mediator** systems

Question 16-2:

a. Explain the use of federated database system / mediator system in the context of the management of higher education in Bangladesh by UGC.

b. Identify some entities and corresponding attributes for global schema for the above.



Data Warehouses Concepts

Data warehouse is an alternative to data integration
Migrates data to a common schema, avoiding run-time overhead
Cost of translating schema/data to a common warehouse schema can be significant

ETL is a process in **Data Warehousing** and it stands for
Extract (E),
Transform (T) and
Load (L).

It is a process in which an **ETL** tool extracts the **data** from various **data** source systems, transforms it in the staging area and then finally, loads it into the **Data Warehouse** system.

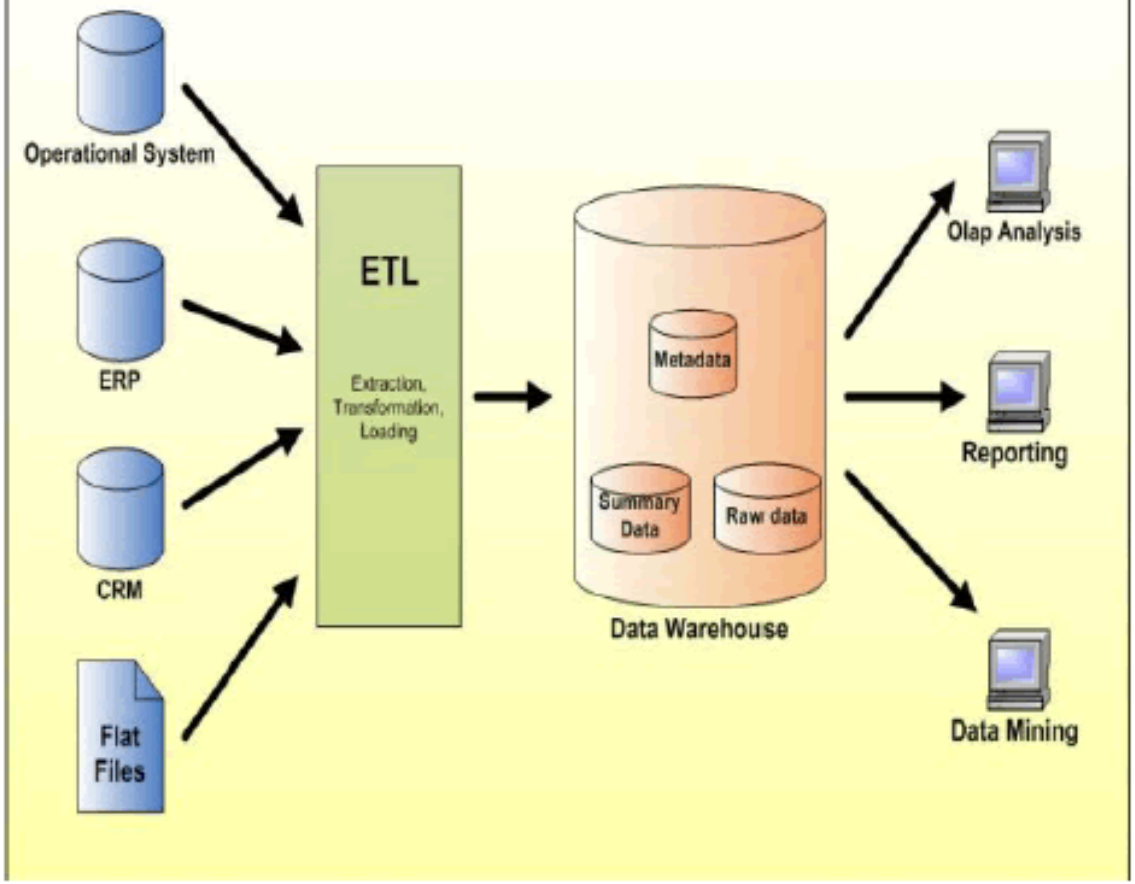


Figure 2: Data Warehouse System.

Data Warehouse Concepts

- A data warehouse is a repository (or archive) of information gathered from multiple sources, stored under a unified schema, at a single site.
- Once gathered, the data are stored for a long time, permitting access to historical data.
- Thus, data warehouses provide the user a single consolidated interface to data, making decision-support queries easier to write.

Data Warehouse Concepts

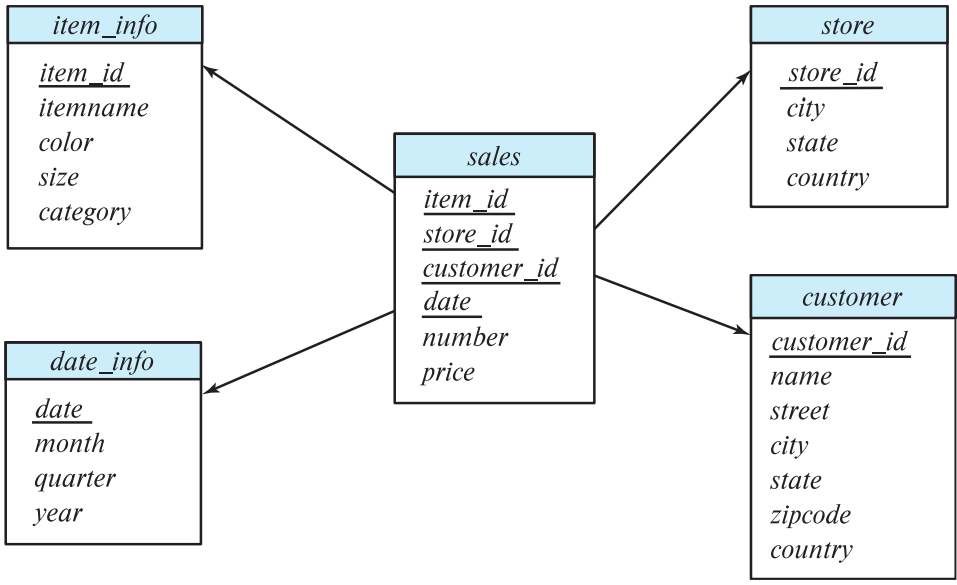
What schema to use. Data sources that have been constructed independently are likely to have different schemas. In fact, they may even use different data models. Part of the task of a warehouse is to perform schema integration and to convert data to the integrated schema before they are stored.

- **Data cleansing**. The task of correcting and preprocessing data is called data cleansing. Data sources often deliver data with numerous minor inconsistencies, which can be corrected.
- **Data transformation:** Transformation of host format to warehouse format

Data Warehouse Concepts

Multidimensional Data and Warehouse Schemas

- Data warehouses typically have schemas that are designed for data analysis, using tools such as OLAP tools.
- The relations in a data warehouse schema can usually be classified as *fact tables* and *dimension tables*.
- Fact tables record information about individual events, such as sales, and are usually very large.
- A table recording sales information for a retail store, with one tuple for each item that is sold, is a typical example of a fact table.



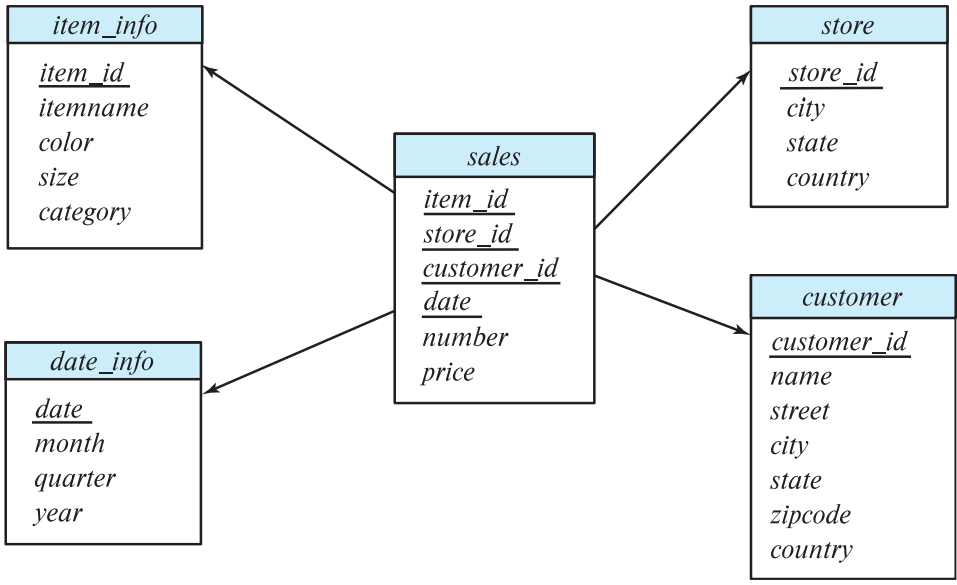
Data Warehouse Concepts

Multidimensional Data and Warehouse Schemas

The attributes in fact table can be classified as either *dimension attributes* or *measure attributes*,

The measure attributes store quantitative information, which can be aggregated upon; the measure attributes of a *sales* table would include the number of items sold and the price of the items.

In contrast, dimension attributes are dimensions upon which measure attributes, and summaries of measure attributes, are grouped and viewed.



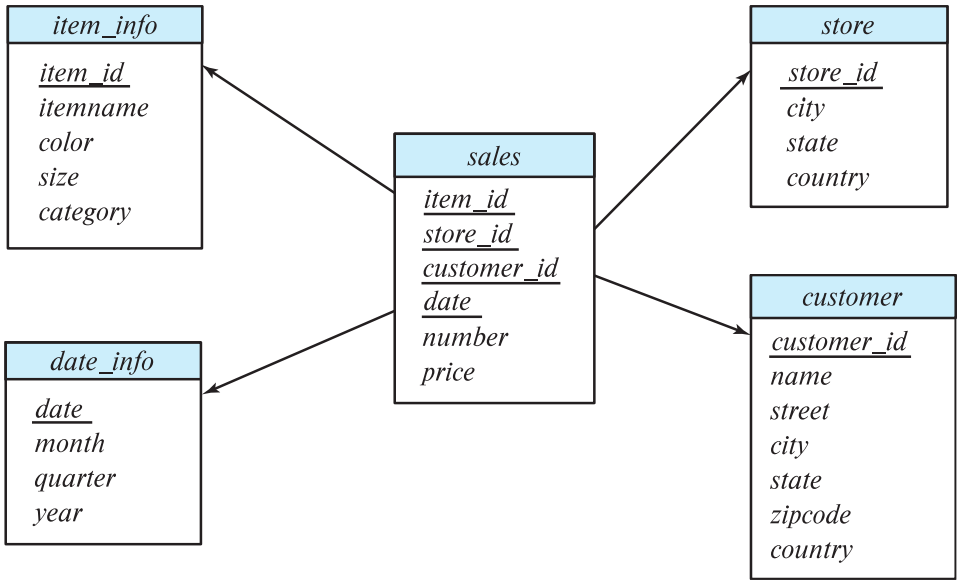
Data Warehouse Concepts

Multidimensional Data and Warehouse Schemas

The dimension attributes of a *sales* table would include an item identifier, the date when the item is sold, which location (store) the item was sold from, the customer who bought the item, and so on.

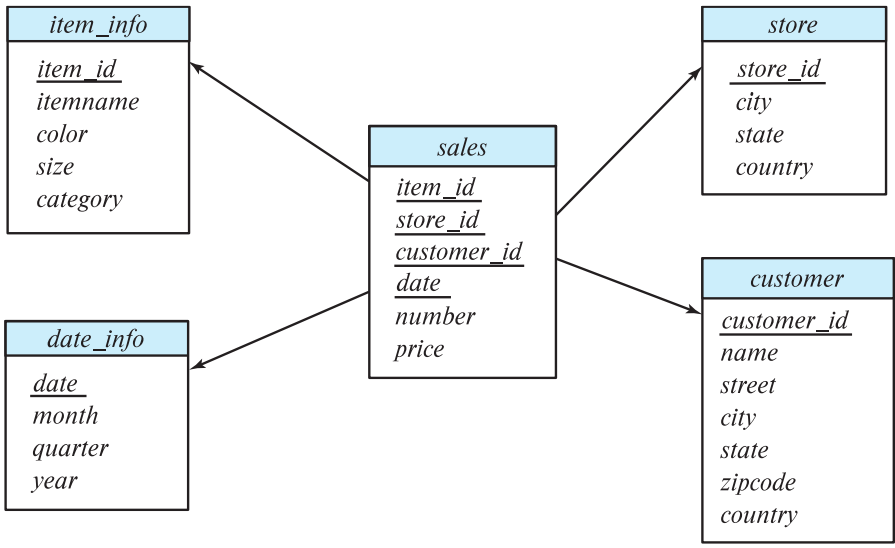
Data that can be modeled using dimension attributes and measure attributes are called multidimensional data.

To minimize storage requirements, dimension attributes are usually short identifiers that are foreign keys into other tables called dimension tables.

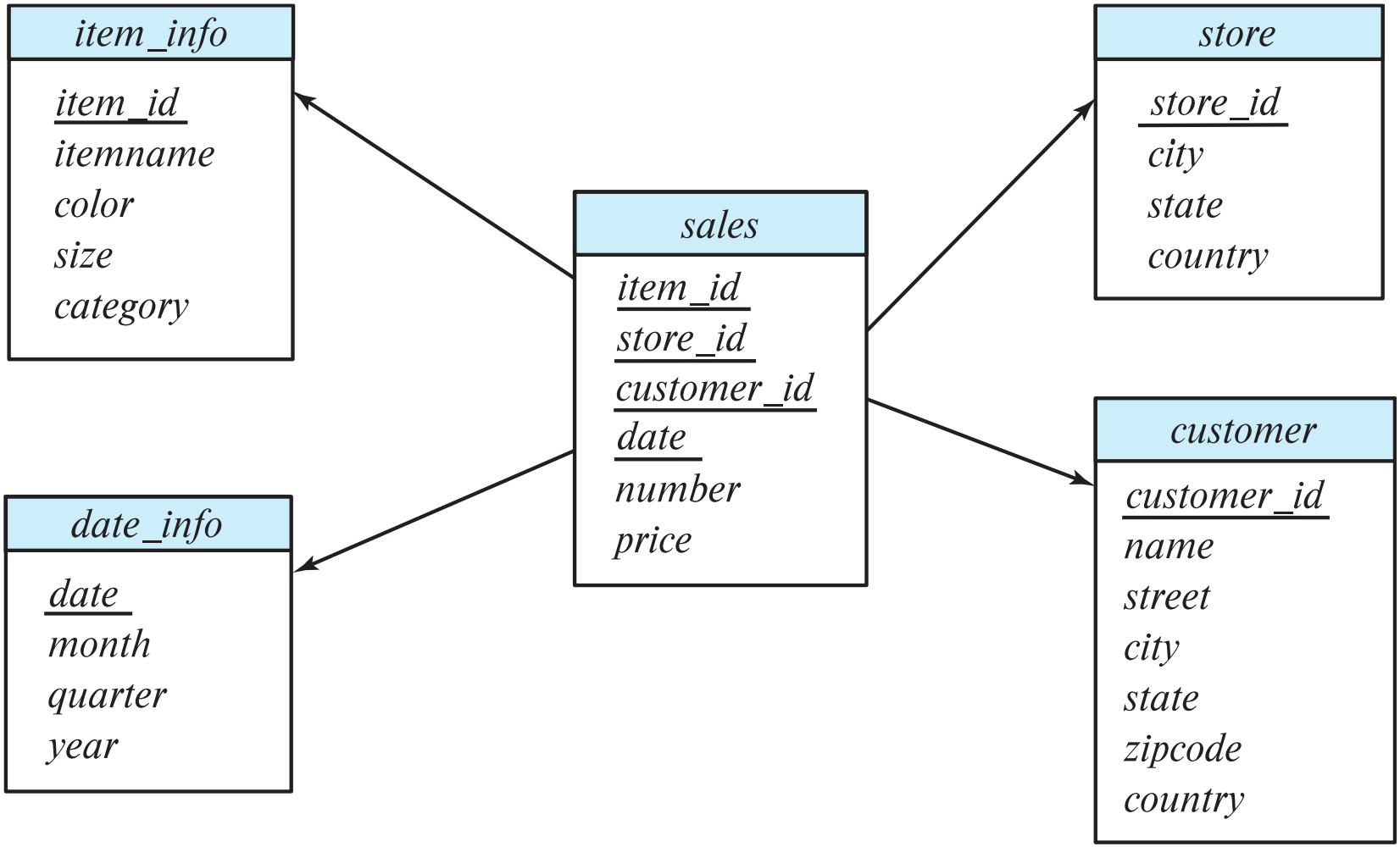


Designing Star Schema

A **fact table sales** would have **dimension attributes item id, store id, customer id, and date, and measure attributes number and price**. The attribute **store id is a foreign key into a dimension table store**, which has other attributes such as store location (city, state, country). The **item id attribute of the sales table would be a foreign key into a dimension table item info**, which would contain information such as the name of the item, the category to which the item belongs, and other item details such as color and size. The **customer id attribute would be a foreign key into a customer table** containing attributes such as name and address of the customer. We can also view the **date attribute as a foreign key into a date info** table giving the month, quarter, and year of each date.



Data Warehouse Schema



Question 17-1:

Design a star schema for the data warehouse of call duration and bill of different types of calls of different users all mobile service phone companies.

Steps 1: Design fact table with measure attributes and dimensional attributes.

Step 2: Design the dimensional tables

Step 3: Design the star schema

Towards Development of Health Data Warehouse: Bangladesh Perspective

- Abstract

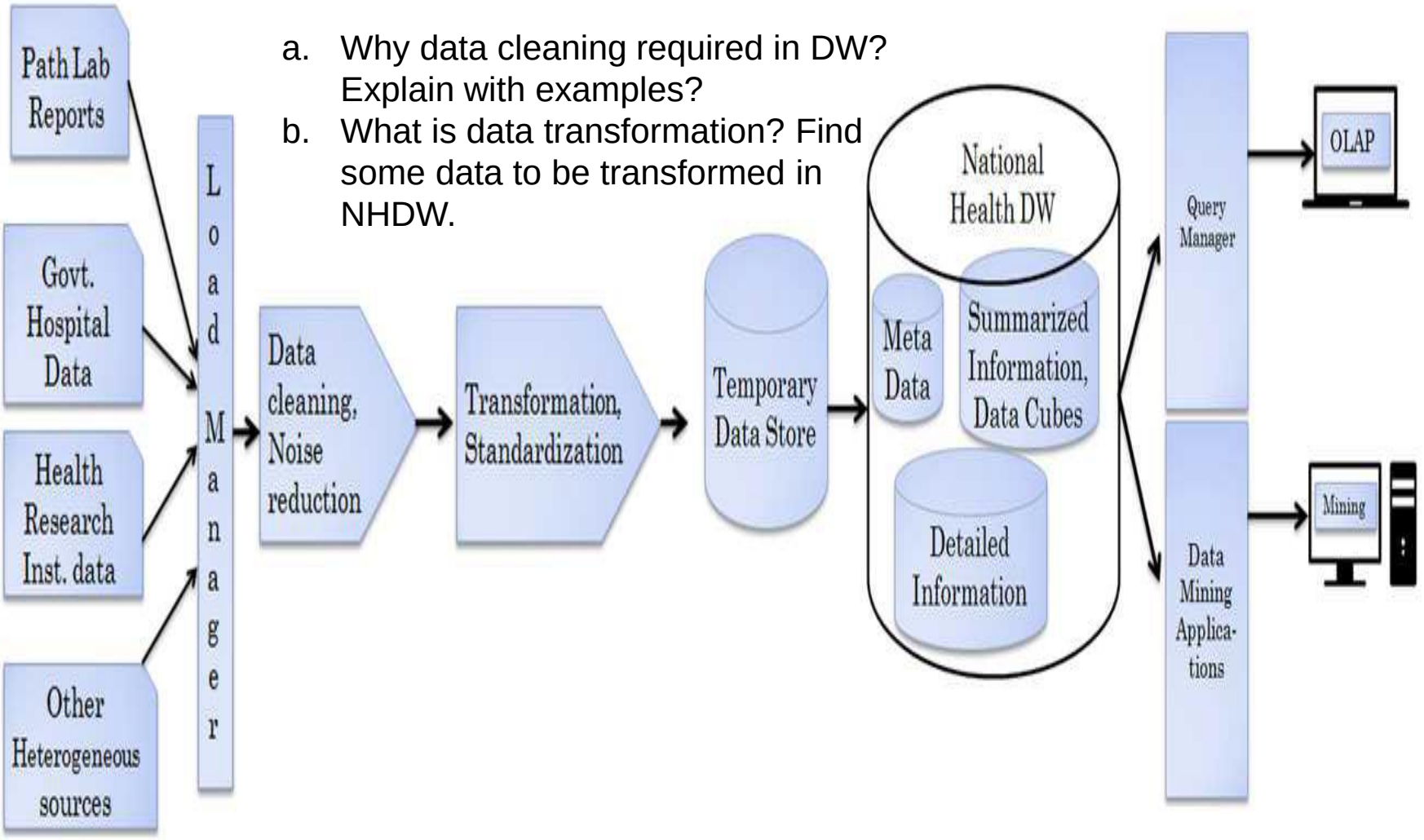
Availability of **timely and accurate data** is essential for **medical decision making (What?)**. Health care organizations face a common problem with large amount of data they have in numerous systems. Such systems are **unstructured and unorganized (How?)**, requires computational time for **data integration (Why?)**. Researchers, medical practitioners, health care providers and patients will not be able to utilize the knowledge stored in different repositories unless synthesize the information from disparate sources. This problem can be solved by Data warehousing.

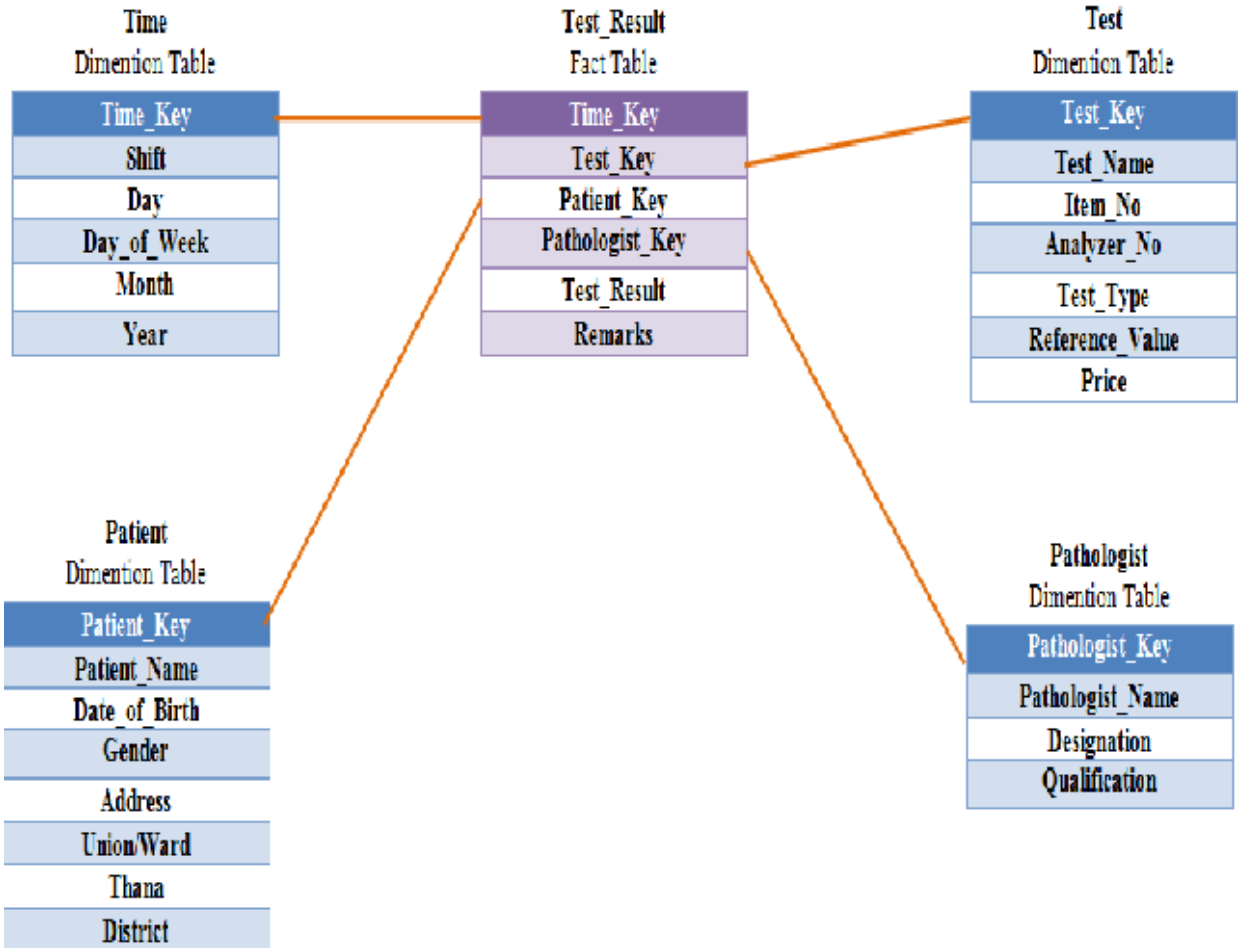
Towards Development of Health Data Warehouse: Bangladesh Perspective

Data warehousing techniques share a common set of tasks, include **requirements analysis, data design, architectural design, implementation and deployment.** Developing Clinical data warehouse is complex and time consuming but is essential to deliver quality patient care. Data integration tasks of medical data store are much challenging when designing clinical data warehouse architecture. This research identifies prospects and complexities of Health data warehousing and Mining in Bangladesh perspective and proposes a data-warehousing model suitable for integrating data from different health care sources.

The Architecture of National DW

a. Why data cleaning required in DW? Explain with examples?
b. What is data transformation? Find some data to be transformed in NHDW.





Question 18-1:

a. Find four useful DSS reports that can be generated from the given star schema from aggregations (average, max, min, sum, count).

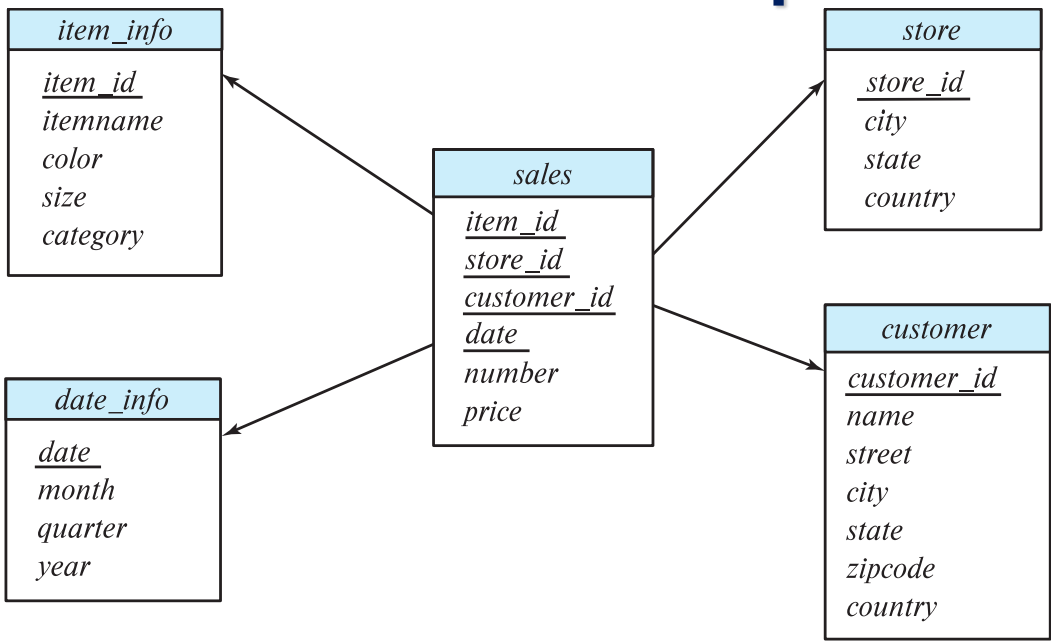
b. Find any missing dimension.

Data Analysis and OLAP

- **Online Analytical Processing (OLAP)**
  - Interactive analysis of data, allowing data to be summarized and viewed in different ways in an online fashion (with negligible delay)
- We use the following relation to illustrate OLAP concepts
  - *sales* (*item_name*, *color*, *clothes_size*, *quantity*)

  This is a simplified version of the *sales* fact table joined with the dimension tables, and many attributes removed (and some renamed)

Example sales relation



How can we get the table
Sales-item (*item_name*, *color*, *clothes_size*, *quantity*)
From star schema?

Answer:
Select *item_name*, *color*, *clothes_size*, number as *quantity*
From item_info a, sales b
Where a.item_id = b.item_id

item_name	*color*	*clothes_size*	*quantity*
dress	dark	small	2
dress	dark	medium	6
dress	dark	large	12
dress	pastel	small	4
dress	pastel	medium	3
dress	pastel	large	3
dress	white	small	2
dress	white	medium	3
dress	white	large	0
pants	dark	small	14
pants	dark	medium	6
pants	dark	large	0
pants	pastel	small	1
pants	pastel	medium	0
pants	pastel	large	1
pants	white	small	3
pants	white	medium	0
pants	white	large	2
shirt	dark	small	2
shirt	dark	medium	6
shirt	dark	large	6
shirt	pastel	small	4
shirt	pastel	medium	1
shirt	pastel	large	2
shirt	white	small	17
shirt	white	medium	1
shirt	white	large	10
skirt	dark	small	2
skirt	dark	medium	5
…	…	…	…
…	…	…	…

Cross Tabulation of *sales* by *item_name* and *color*

clothes_size **all**

color

item_name	dark	pastel	white	total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pants	20	2	5	27
total	62	54	48	164

How can you find the cross-tab of sales?
Write SQL to find the cross-tab.

SQL for cloth_size:
Select sum(quantity) from sales

SQL for total of item_name:
Select item_name, sum(quantity)
From sales
Group by item_name

SQL for total of color:
Select color, sum(quantity)
From sales
Group by item_name

SQL for other cells:
Select item_name, color, sum(quantity)
From sales
Group by item_name, color

- The table above is an example of a **cross-tabulation** (**cross-tab**), also referred to as a **pivot-table**.

Values for one of the dimension attributes form the row headers

Values for another dimension attribute form the column headers

Other dimension attributes are listed on top

Values in individual cells are (aggregates of) the values of the dimension attributes that specify the cell.

Cross Tabulation of *sales* by *item_name* and *color*

clothes_size **all**

color

item_name	dark	pastel	white	total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pants	20	2	5	27
total	62	54	48	164

How can you find the cross-tab of sales?
Write SQL to find the cross-tab.

SQL for cloth_size:
Select sum(quantity) from sales

SQL for total of item_name:
Select item_name, sum(quantity)
From sales
Group by item_name

SQL for total of color:
Select color, sum(quantity)
From sales
Group by item_name

SQL for other cells:
Select item_name, color, sum(quantity)
From sales
Group by item_name, color

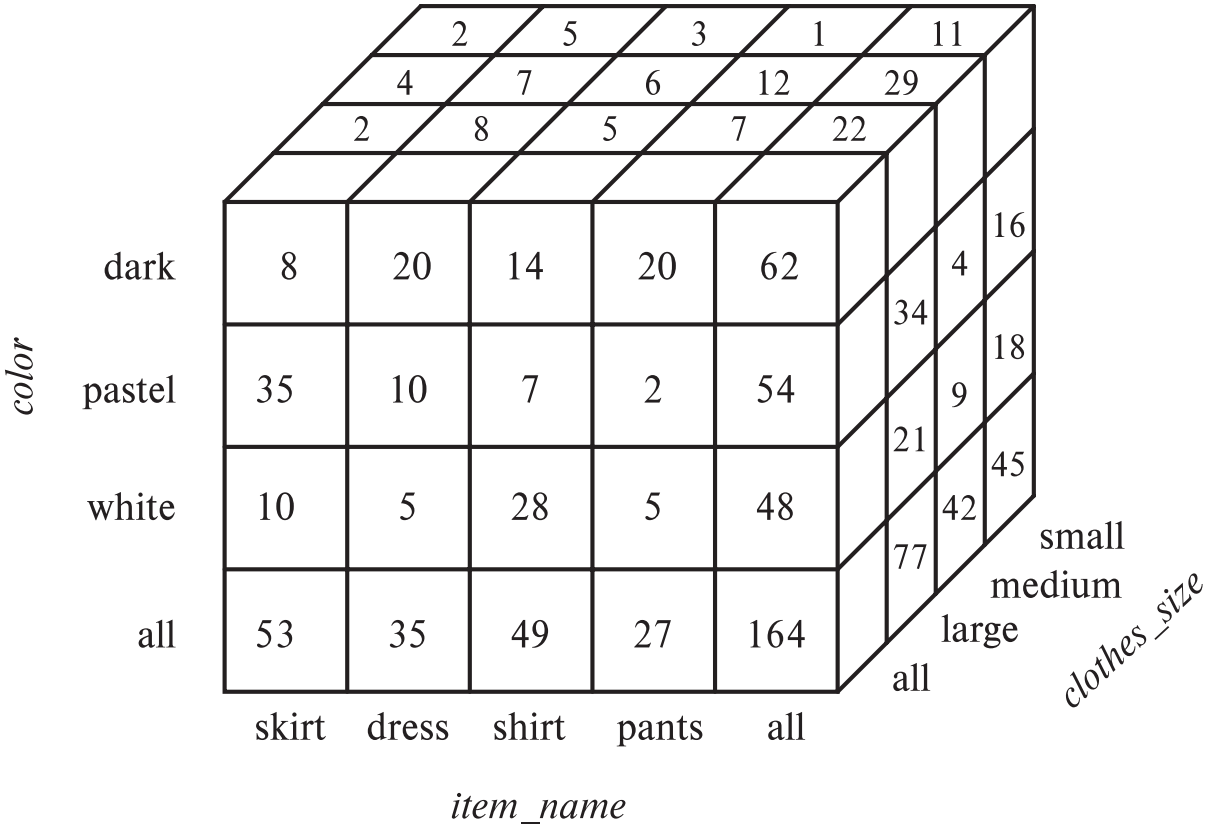
Question 18-2:

a. Write cross tabulation structure of sales by cloth_size and item_name.

b. Write SQL to find the cross-tab

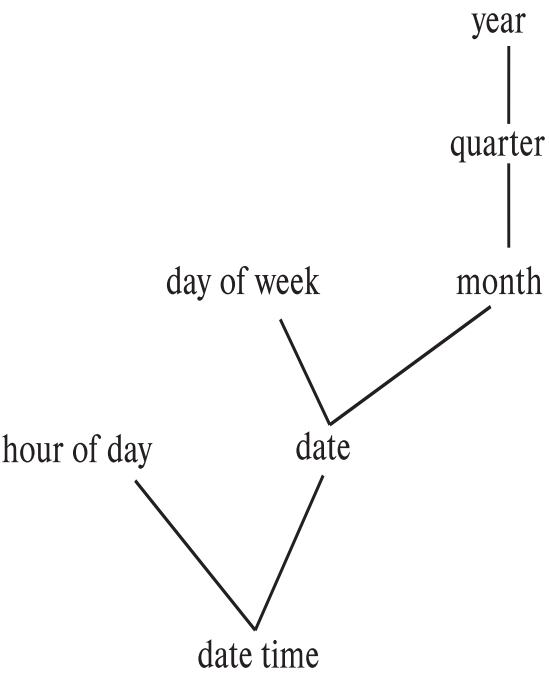
Data Cube

- A **data cube** is a multidimensional generalization of a cross-tab
- Can have n dimensions; we show 3 below
- Cross-tabs can be used as views on a data cube

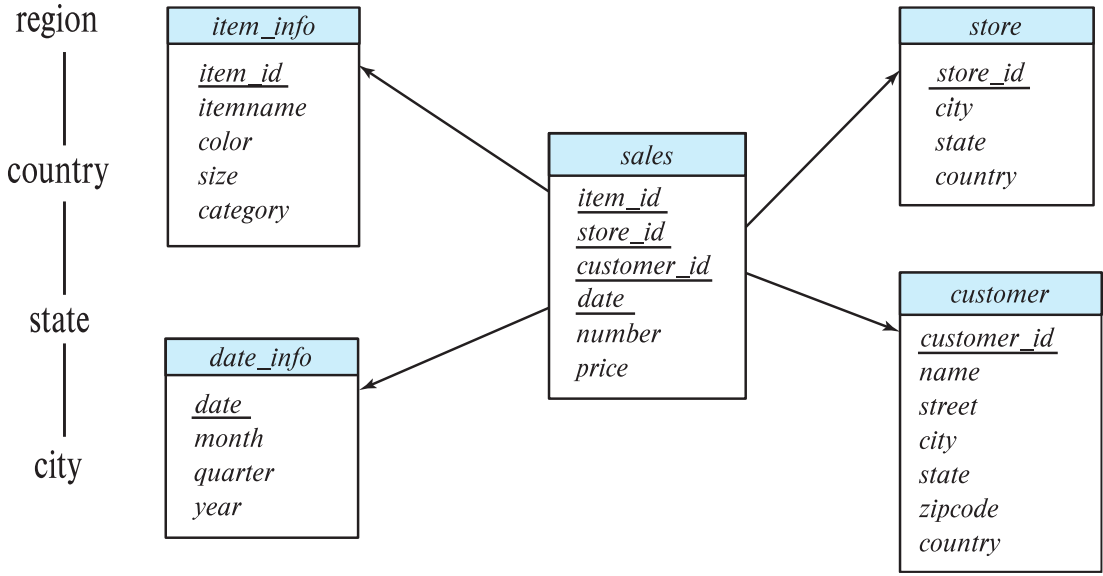


Hierarchies on Dimensions

- **Hierarchy** on dimension attributes: lets dimensions be viewed at different levels of detail
- E.g., the dimension *datetime* can be used to aggregate by hour of day, date, day of week, month, quarter or year



(a) time hierarchy



(b) location hierarchy

Hierarchies on Dimensions

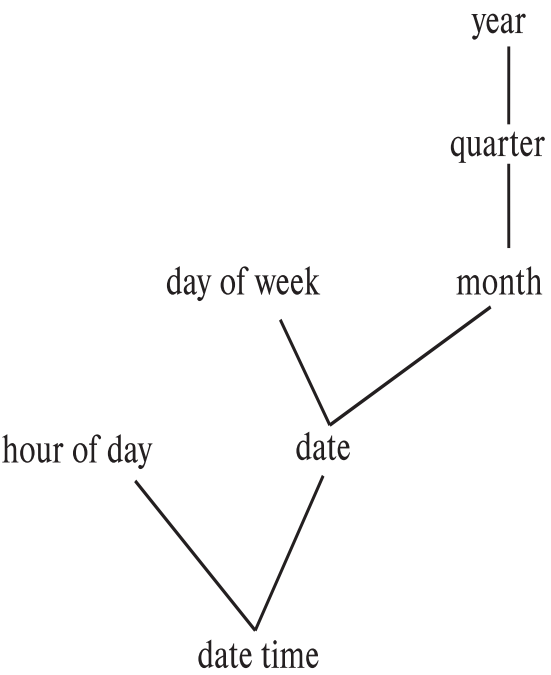
How can you prepare DSS reports based on hierarchy?

Report on date (Year, , quarter, month, date wise report)

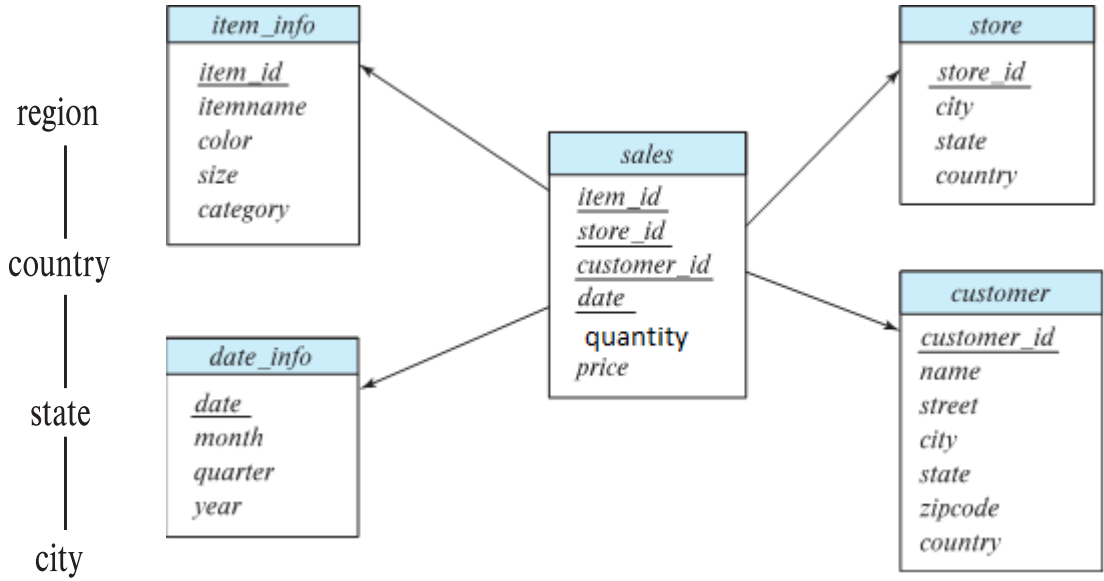
R1 = select year, quarter, month, date, sum(quantity) as tot_d from sales s, date_info d Where s.date = d.date Group by year, quarter, month, date

Report on date (Year, , quarter, month wise report)

R2 = select year, quarter, month sum(tot_d) as tot_m from R1 Where s.date = d.date Group by year, quarter, month, date



(a) time hierarchy



(b) location hierarchy

Hierarchies on Dimensions

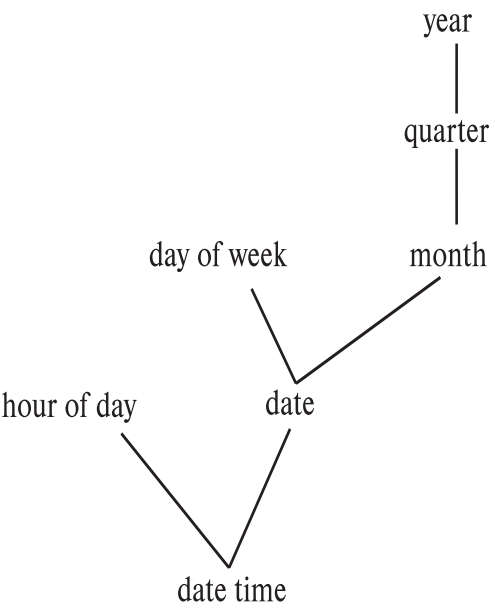
How can you prepare DSS reports based on hierarchy?

Report on date (Year, , quarter, month, date wise report)

R1 = select year, quarter, month, date, sum(quantity) as tot_d from sales s, date_info d Where s.date = d.date Group by year, quarter, month, date

Report on month (Year, , quarter, month wise report)

R2 = select year, quarter, month sum(tot_d) as tot_m from R1 Where s.date = d.date Group by year, quarter, month, date



(a) time hierarchy

(b) location hierarchy

Report on quarter (Year, quarter wise report)

R3 = select year, quarter sum(tot_m) as tot_q from R2
Group by year, quarter

Report on year (Year wise report)

R4 = select year sum(tot_q) as tot_y from R3
Group by year

** R1, R2 R3, and R4 may be stored in the database as materialized view

Question 18-3: Write SQL for all DSS reports on location hierarchy

Cross Tabulation With Hierarchy

- Cross-tabs can be easily extended to deal with hierarchies
- Can drill down or roll up on a hierarchy
- E.g. hierarchy: *item_name* → *category*

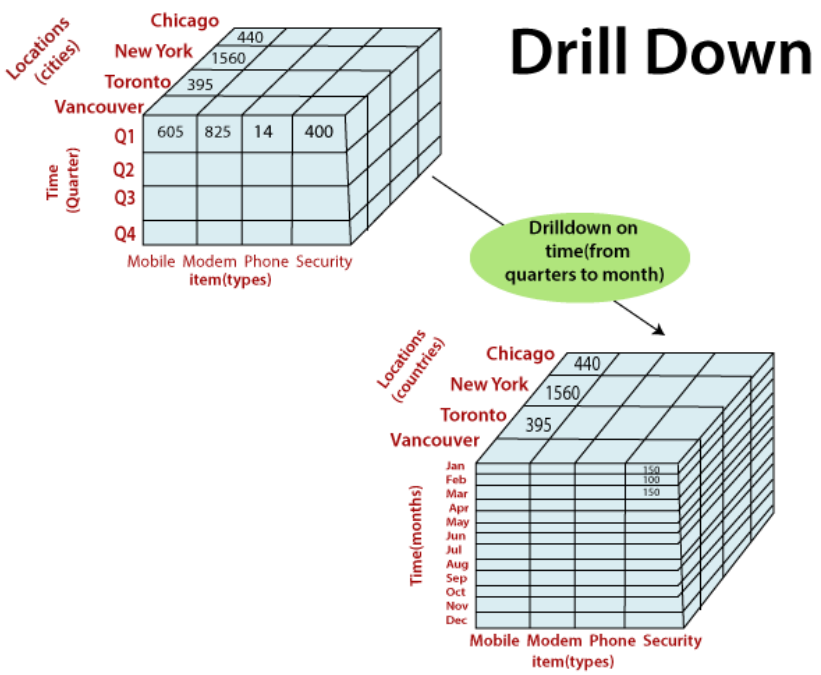
clothes_size: **all**

category	*item_name*	*color*				
		dark	pastel	white	total	
womenswear	skirt	8	8	10	53	
	dress	20	20	5	35	
	subtotal	28	28	15		88
menswear	pants	14	14	28	49	
	shirt	20	20	5	27	
	subtotal	34	34	33		76
total		62	62	48		164

Cross Tabulation With Hierarchy

- Cross-tabs can be easily extended to deal with hierarchies
- Can drill down or roll up on a hierarchy
- E.g. hierarchy: *item_name* → *category*

Drill down is a capability that takes the user from a more general view of the data to a more specific one at the click of a mouse. For example, a report that shows sales revenue by division can allow the user to select a division, click on it and see sales revenue by district within that division.



Cross Tabulation With Hierarchy

- Cross-tabs can be easily extended to deal with hierarchies
- Can drill down or roll up on a hierarchy
- E.g. hierarchy: *item_name* → *category*

The roll-up operation aggregates the data by ascending the location hierarchy from the level of the city to the level of the country. When a roll-up is performed by dimensions reduction, one or more dimensions are removed from the cube. For example, R1 → R2 → R3 → R4.