On-Line Analytical Processing

(OLAP)

OLAP

- OLAP (online analytical processing) is a computing method that enables users to easily and selectively extract and query data in order to analyze it from different points of view.

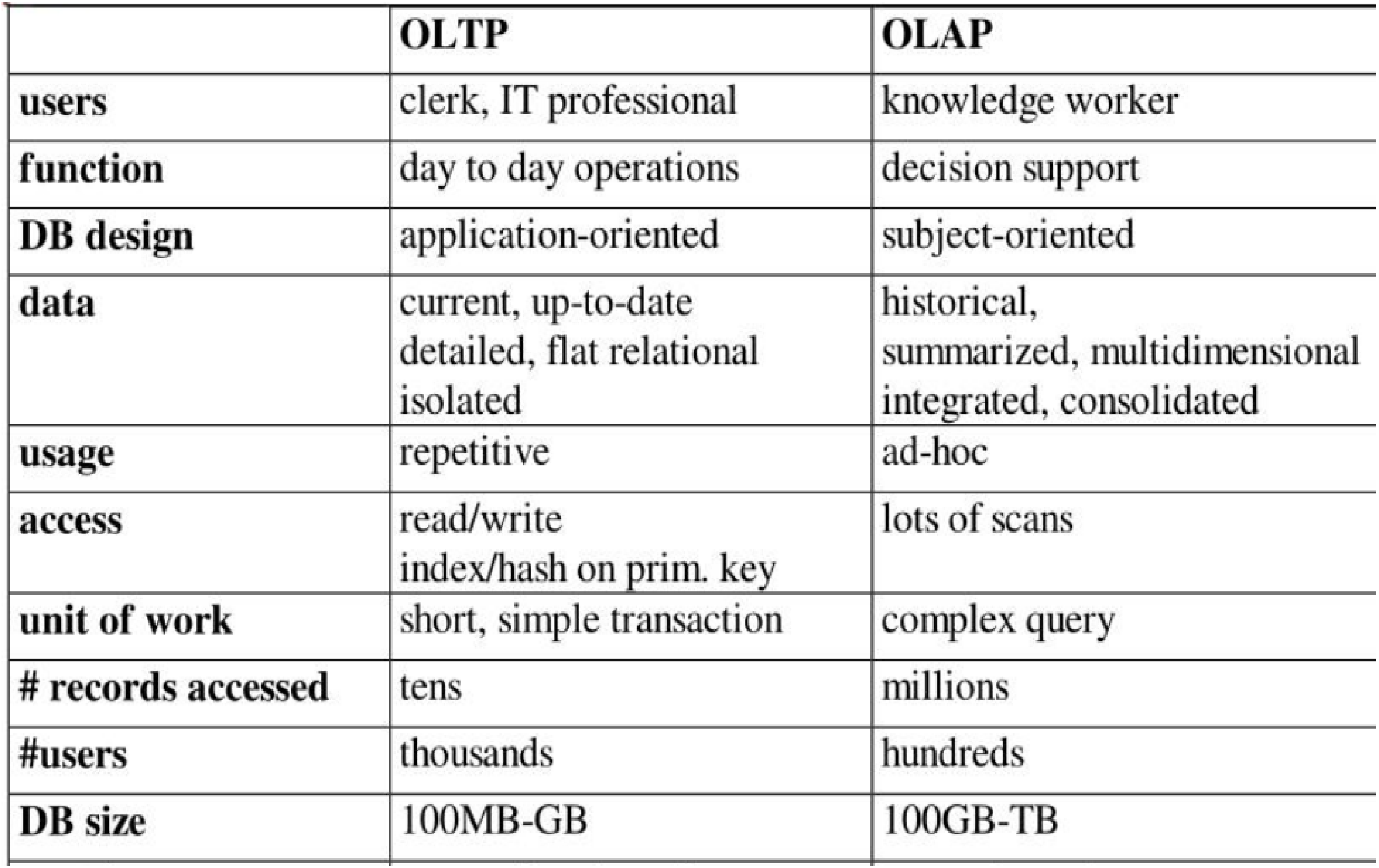
- OLAP business intelligence queries often aid in trends analysis, financial reporting, sales forecasting, budgeting and other planning purposes.

- For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September and then see a comparison of other product sales in Florida in the same time period.

How OLAP systems work

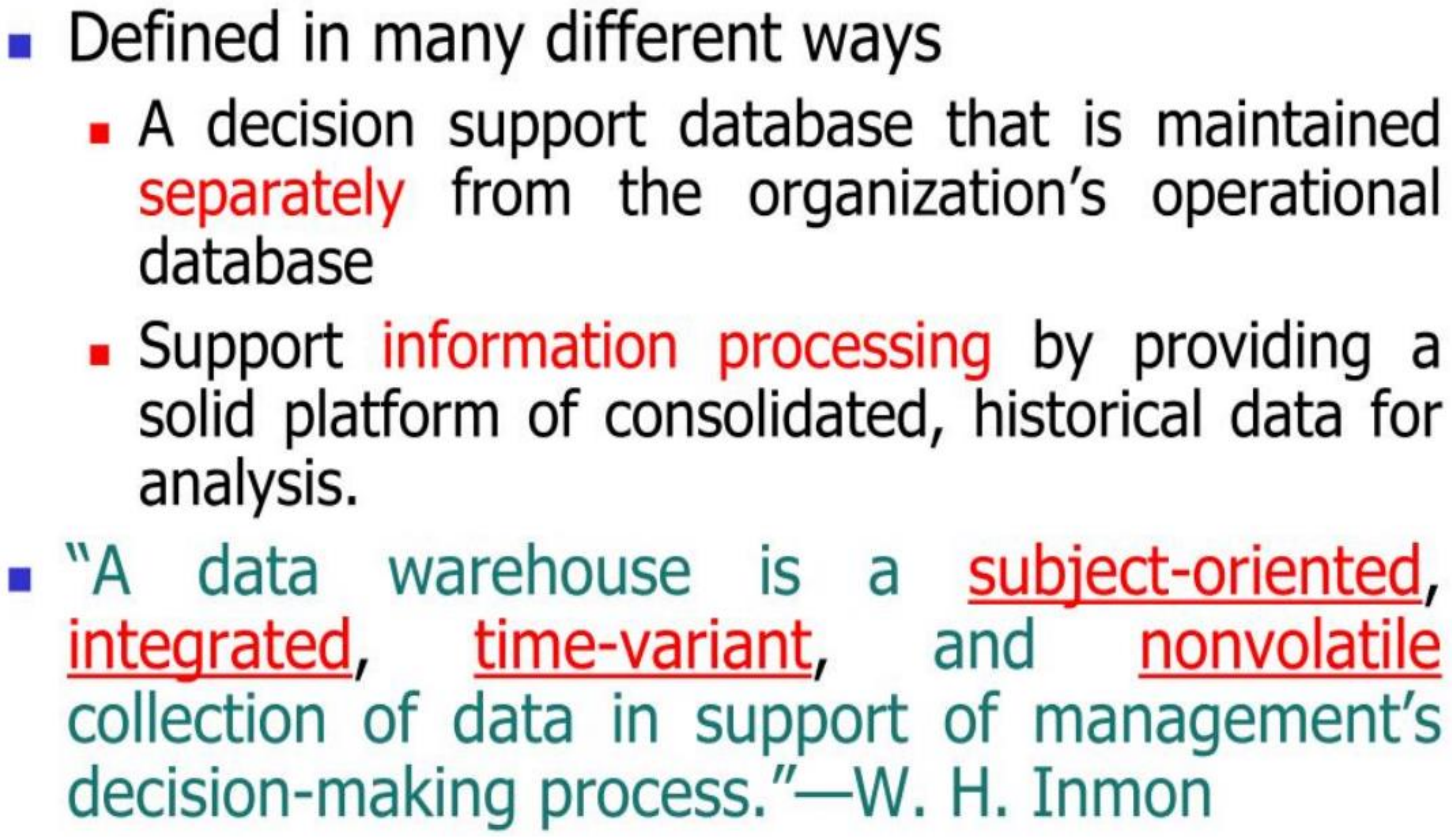
- To facilitate this kind of analysis, data is collected from multiple data sources and stored in data warehouses then cleansed and organized into data cubes.
- Each OLAP cube contains data categorized by dimensions (such as customers, geographic sales region and time period) derived by dimensional tables in the data warehouses.
- Dimensions are then populated by members (such as customer names, countries and months) that are organized hierarchically.
- OLAP cubes are often pre-summarized across dimensions to drastically improve query time over relational databases.

OLTP v/s OLAP (5-7 Marks)



	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB

What is Data Warehouse ?



- Defined in many different ways
  - A decision support database that is maintained separately from the organization’s operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon

Data Warehouse Key Features

Subject-oriented:

- A data warehouse is organized around major subjects such as customer, supplier, product, and sales.
- Does not concentrate on the day-to-day operations and transaction processing of an organization
- A data warehouse focuses on the modeling and analysis of data for decision makers.

Integrated:

- A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

Time-variant:

- Data are stored to provide information from an historic perspective (e.g., the past 5–10 years).
- Every key structure in the data warehouse contains, either implicitly or explicitly, a time element.

Nonvolatile:

- 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 Warehouse Architecture

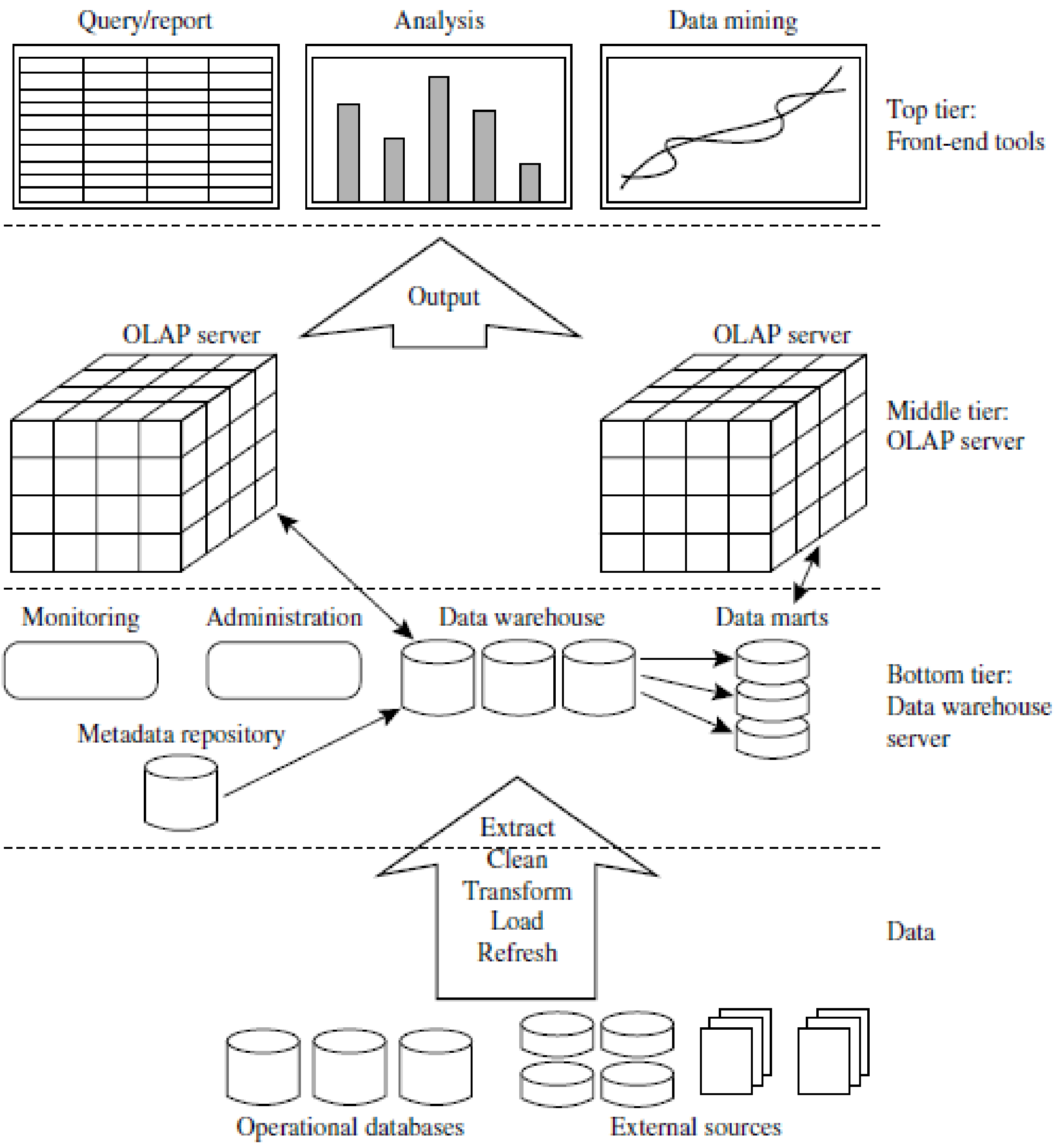


Figure 4.1 A three-tier data warehousing architecture.

Data Warehouse Architecture

1. The bottom tier is a **warehouse database server** that is almost always a relational database system.

- Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (e.g., customer profile information provided by external consultants).
- These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse .
- The data are extracted using application program interfaces known as **gateways**.

Data Warehouse Architecture

2. The middle tier is an **OLAP server** that is typically implemented using either
(1) a **relational OLAP(ROLAP)** model (i.e., an extended relational DBMS that maps operations
on multidimensional data to standard relational operations); or
(2) a **multidimensional OLAP (MOLAP)** model (i.e., a special-purpose server that directly implements multidimensional data and operations).

3. The top tier is a **front-end client layer**, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

Data Warehouse Models

there are three data warehouse models:

1. The *enterprise warehouse*
2. The *data mart*
3. The *virtual warehouse*

1. Enterprise warehouse:

- An enterprise warehouse collects all of the information about subjects spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computer superservers, or parallel architecture platforms.
- It requires extensive business modeling and may take years to design and build.

Data Warehouse Models

2. Data mart:

- A data mart contains a subset of corporate-wide data that is of value to a specific group of users.
- The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales.
- The data contained in data marts tend to be summarized.
- Data marts are usually implemented on low-cost departmental servers that are Unix/Linux or Windows based.
- The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years.

Data Warehouse Models

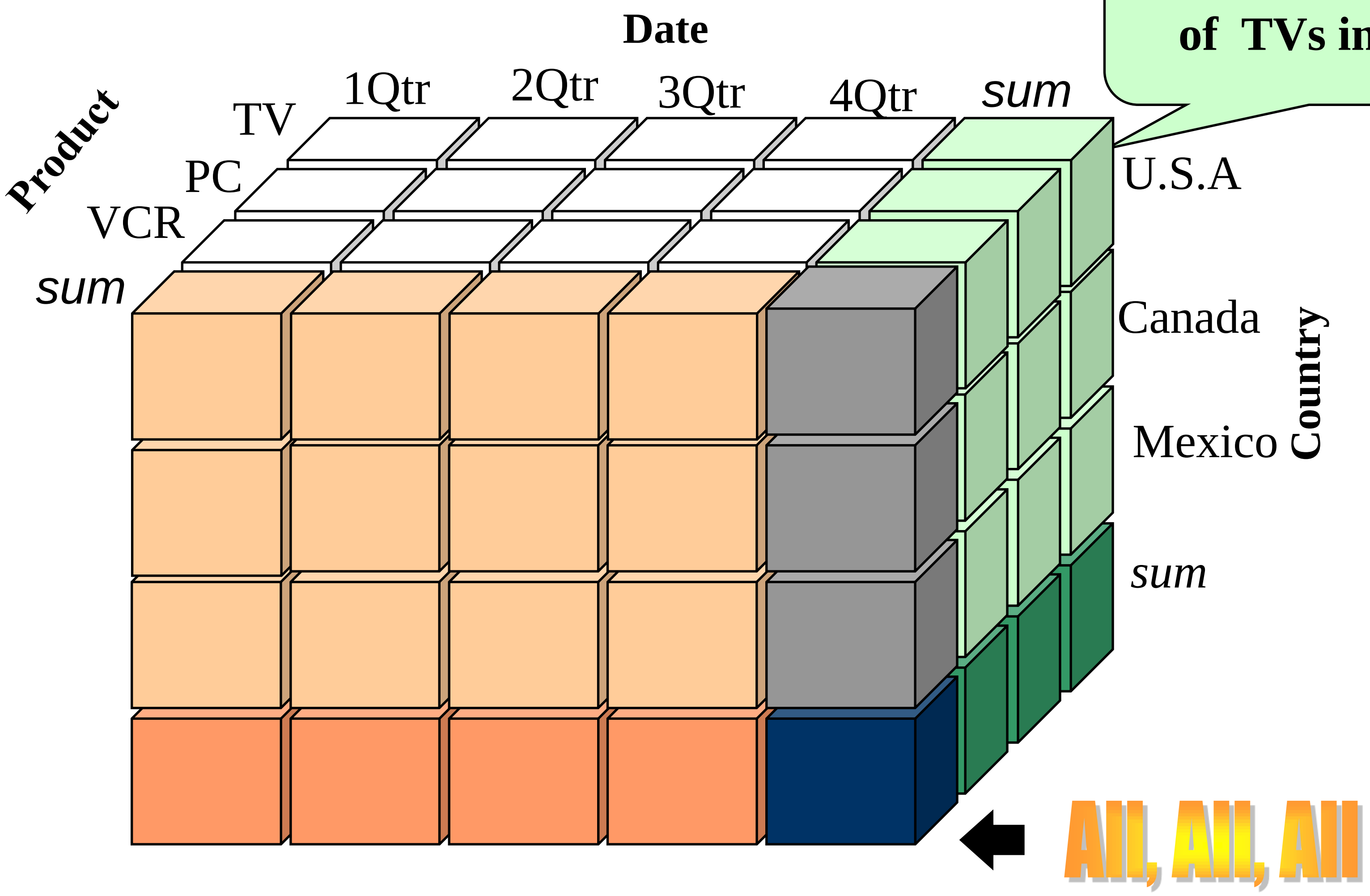
3. Virtual warehouse:

- A virtual warehouse is a set of views over operational databases.
- For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.

What is Data Cube ?

- When data is grouped or combined in multidimensional matrices it is called as **Data Cube**.
- In other words, Cube is a mechanism used to pull together data in organized, dimensional structures for analysis.
- The data cube method has a few alternative names or a few variants, such as "Multidimensional databases," "materialized views," and **OLAP (On-Line Analytical Processing**).
- Data Cube is a multidimensional database that is optimized for data warehouse and OLAP application.
- The general idea of this approach is to materialize certain expensive computations that are frequently inquired and Queries are performed on the cube to retrieve decision support information

Sample DATA CUBE



Total annual sales of TVs in U.S.A.

OLAP CUBE

- **Online Analytical Processing (OLAP)** is a category of software technology that enables analysts and managers to inspect data from multiple databases simultaneously. The process provides fast, intuitive, and interactive access to multidimensional data. It also helps analysts extract a wide variety of insights.

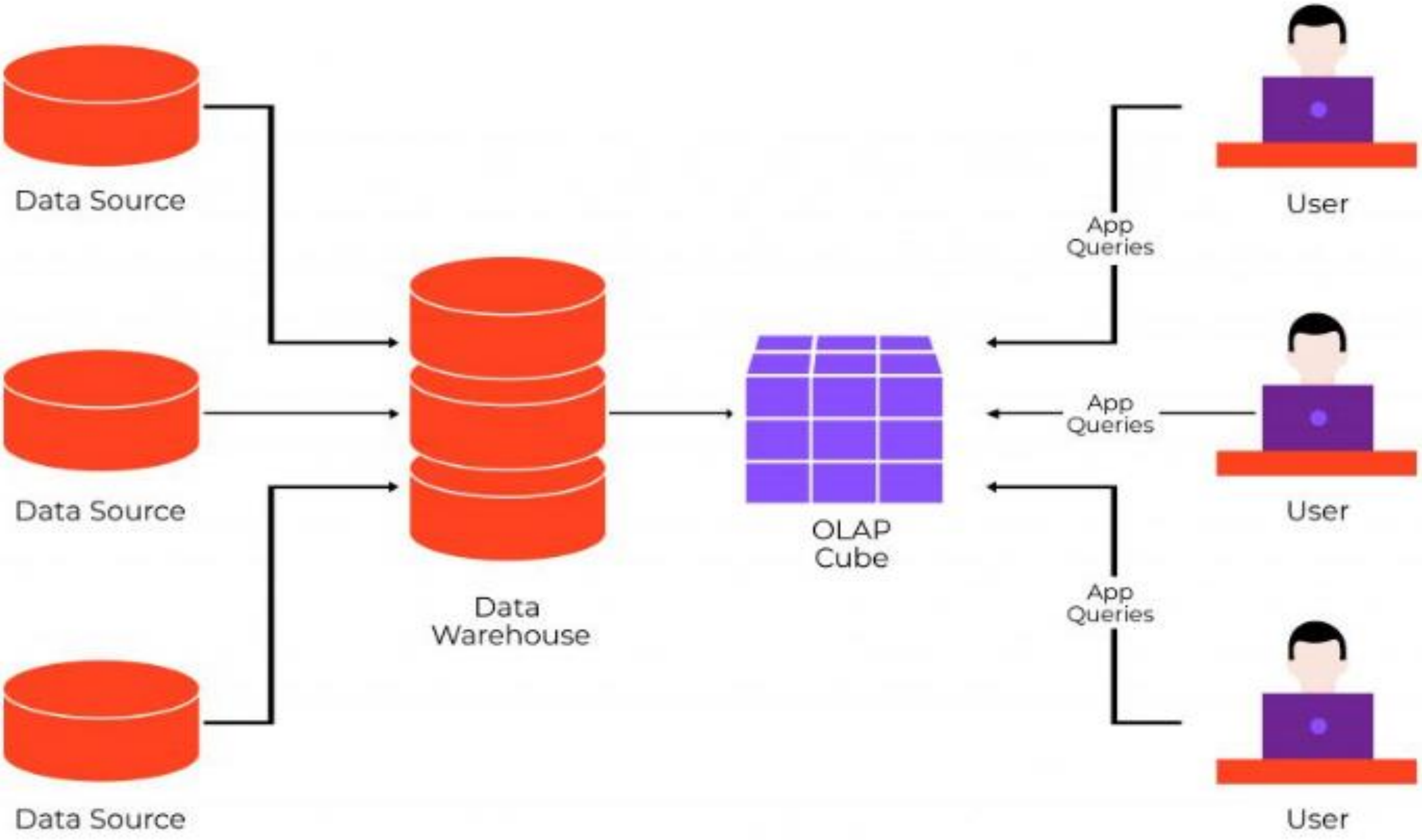
- The goal of OLAP is to pre-calculate and pre-aggregate data to make analysis faster. This pre-aggregated and pre-calculated data is stored in an **OLAP database or OLAP Cube**.

- OLAP data is typically stored in a **schema** in a relational data warehouse or in a special-purpose data management system.

- Measures are derived from the records in the **fact table** and dimensions are derived from the **dimension tables**.

The OLAP process

How data is prepared for online analytical processing (OLAP)



How Does OLAP Work ?

1. First, data is first extracted from various data sources and formats, like text files and spreadsheets. This data is then stored in the **Data Warehouse**.

2. Next, the data is cleaned, transformed, and **stored in OLAP Cubes**

3. Once in the OLAP cubes, information is then pre-calculated and pre-aggregated in advance for further analysis.

4. Lastly, the user gets the data from the OLAP cubes by running queries against them.

OLAP Schemas

OLAP Schemas

- An OLAP schema is a logical model that defines a multidimensional data structure.
- It defines one or more cubes in a single database that each are defined by one or more dimensions and measures.
- Schema is a logical description of the entire database.
- It includes the name and description of records of all record types including all associated data-items and aggregates.
- Much like a database, a data warehouse also requires to maintain a schema.

- The fact table mainly consists of business facts and foreign keys that refer to primary keys in the dimension tables.

- A dimension table consists mainly of descriptive attributes that are textual fields.

- When comparing the size of the two tables, a fact table is bigger than a dimensional table.

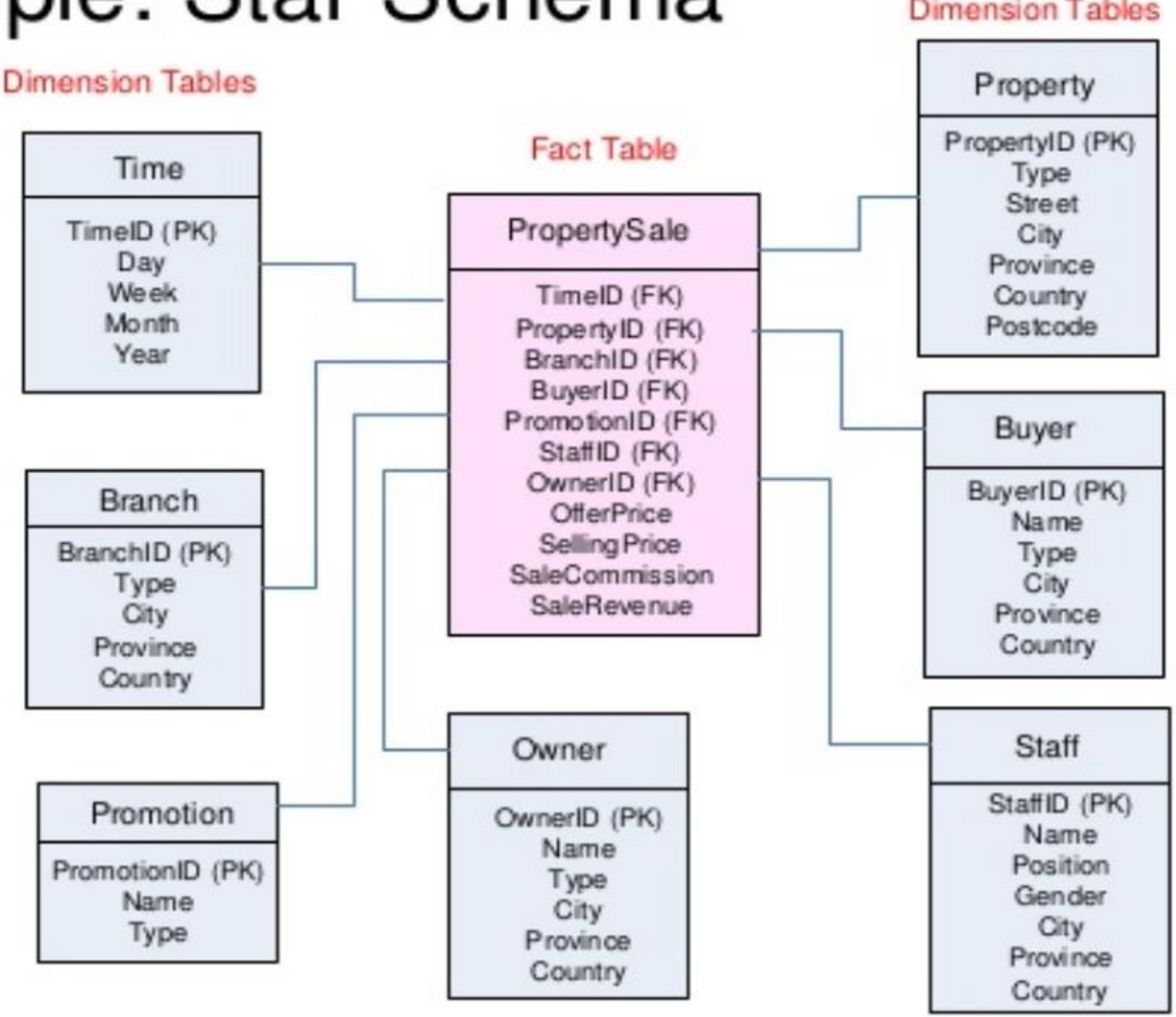
- A database uses relational model, while a data warehouse uses

    1. **Star**
    2. **Snowflake and**
    3. **Fact Constellation schema**

1. Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely offer price, selling price, sales commission and sales revenue.

Example: Star Schema



Dimension Tables

Dimension Tables

Fact Table

PropertySale

TimeID (FK)
PropertyID (FK)
BranchID (FK)
BuyerID (FK)
PromotionID (FK)
StaffID (FK)
OwnerID (FK)
OfferPrice
Selling Price
SaleCommission
SaleRevenue

Time

TimeID (PK)
Day
Week
Month
Year

Branch

BranchID (PK)
Type
City
Province
Country

Promotion

PromotionID (PK)
Name
Type

Owner

OwnerID (PK)
Name
Type
City
Province
Country

Property

PropertyID (PK)
Type
Street
City
Province
Country
Postcode

Buyer

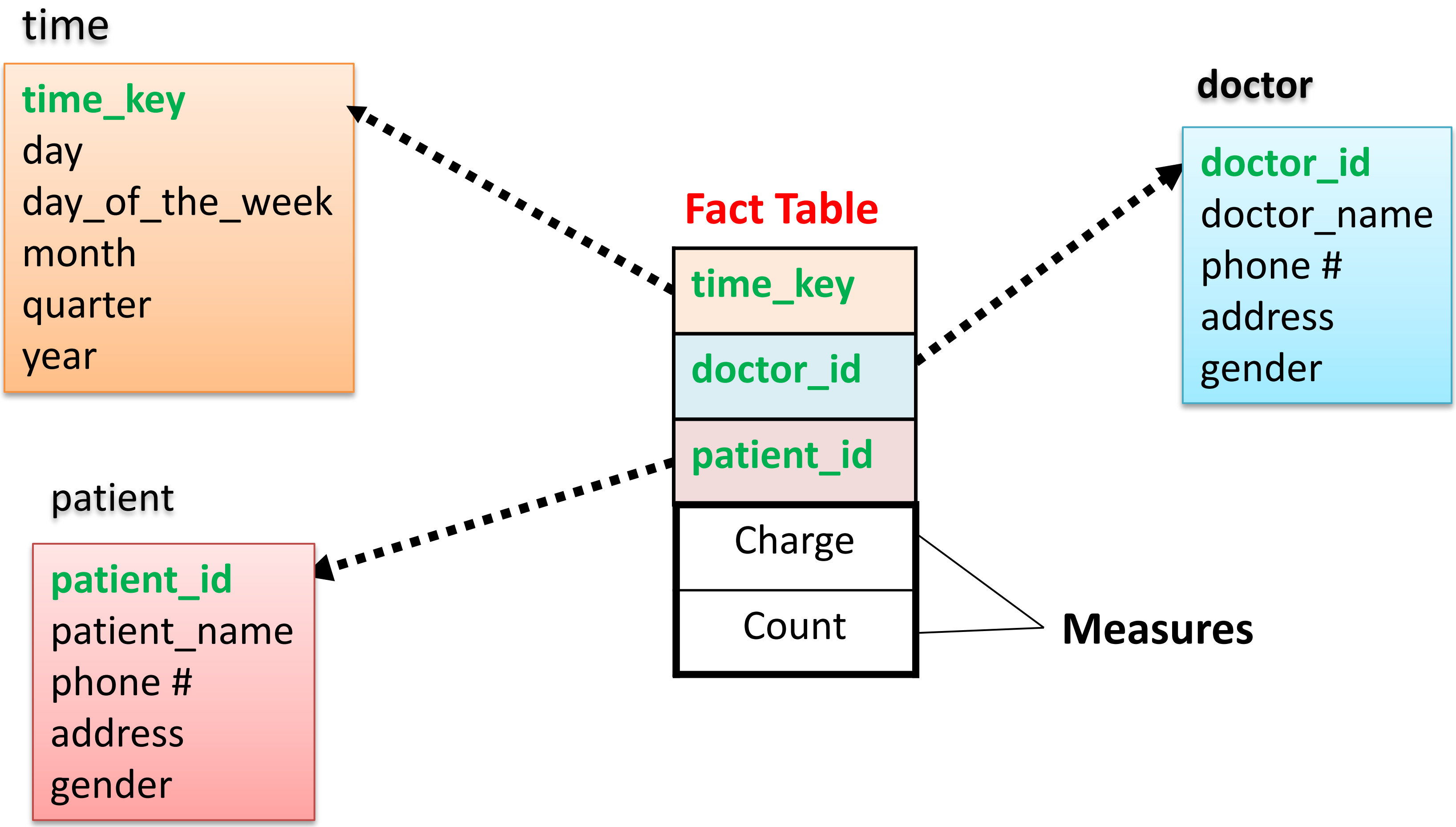
BuyerID (PK)
Name
Type
City
Province
Country

Staff

StaffID (PK)
Name
Position
Gender
City
Province
Country

PROBLEM-1

Suppose that a data warehouse consists of the 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. Draw STAR Schema diagram for the above data Warehouse.



time

time_key
day
day_of_the_week
month
quarter
year

Fact Table

time_key
doctor_id
patient_id
Charge
Count

Measures

doctor

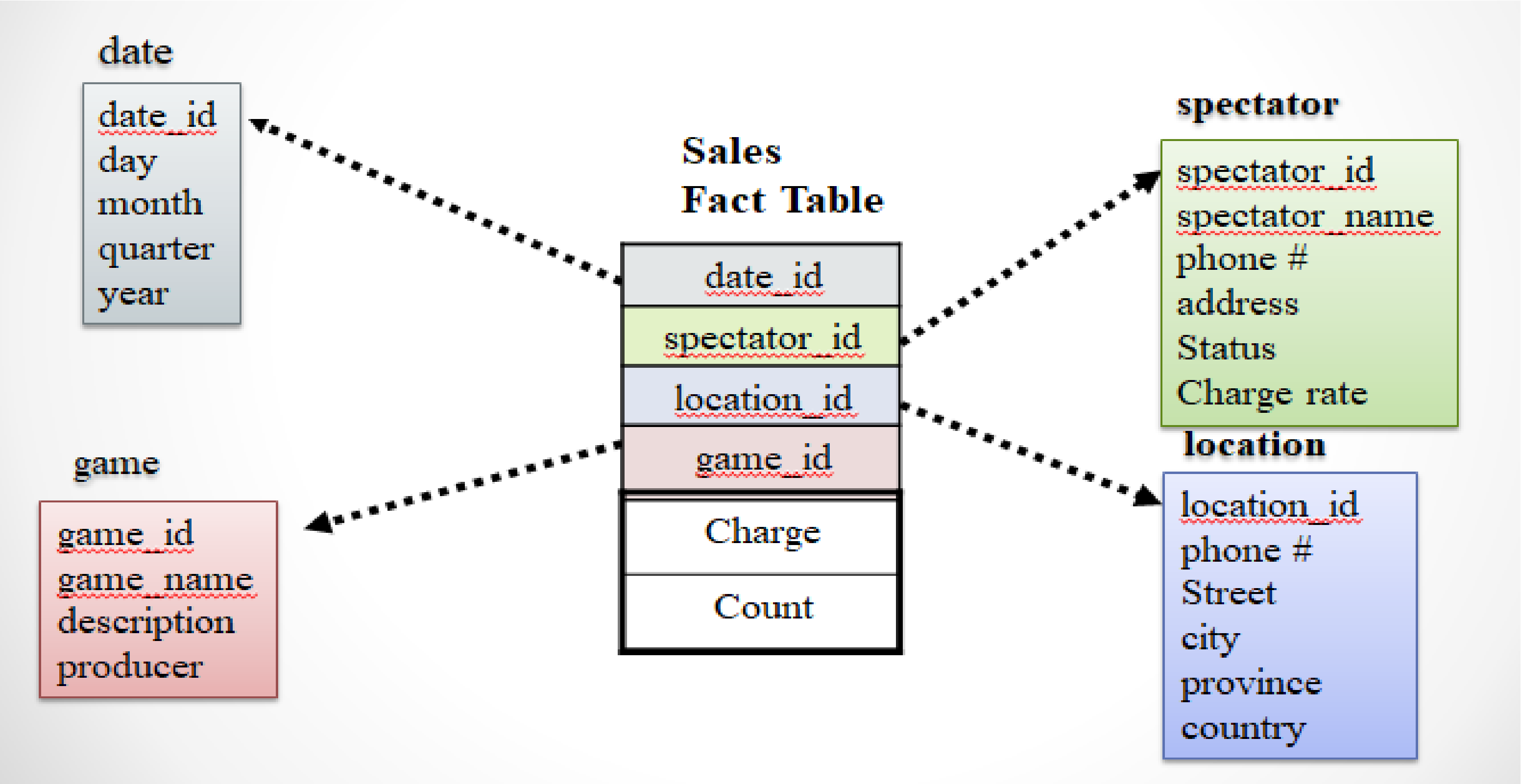
doctor_id
doctor_name
phone #
address
gender

patient

patient_id
patient_name
phone #
address
gender

PROBLEM-2

Suppose that a data warehouse consist of the 4-Dimensions i.e. **date**, **spectator**, **location** and **game** and the measures were **count** and **charge** (where charge is the fare that spectator pays when he/ she is watching a game on given date). Spectators may be students, adults or seniors with each category having its own charge rate. Draw a **STAR** Schema diagram for the data warehouse.



date

date_id
day
month
quarter
year

Sales Fact Table

date_id
spectator_id
location_id
game_id
Charge
Count

spectator

spectator_id
spectator_name
phone #
address
Status
Charge rate

location

location_id
phone #
Street
city
province
country

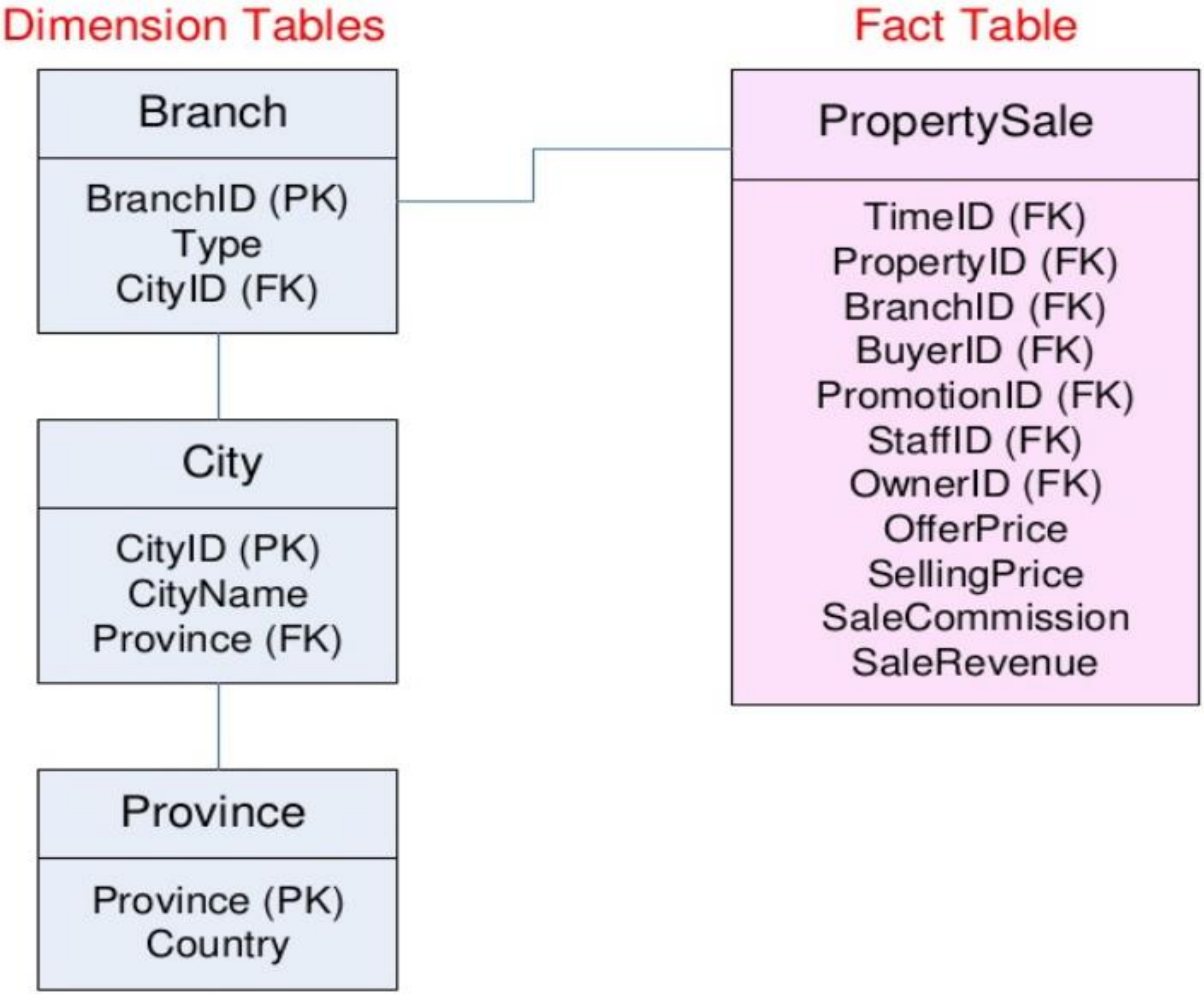
game

game_id
game_name
description
producer

2. Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the branch dimension table in star schema is normalized and split into three dimension tables, namely branch, city and province table.

Example: Snowflake Schema



Dimension Tables

Branch

BranchID (PK)
Type
CityID (FK)

City

CityID (PK)
CityName
Province (FK)

Province

Province (PK)
Country

Fact Table

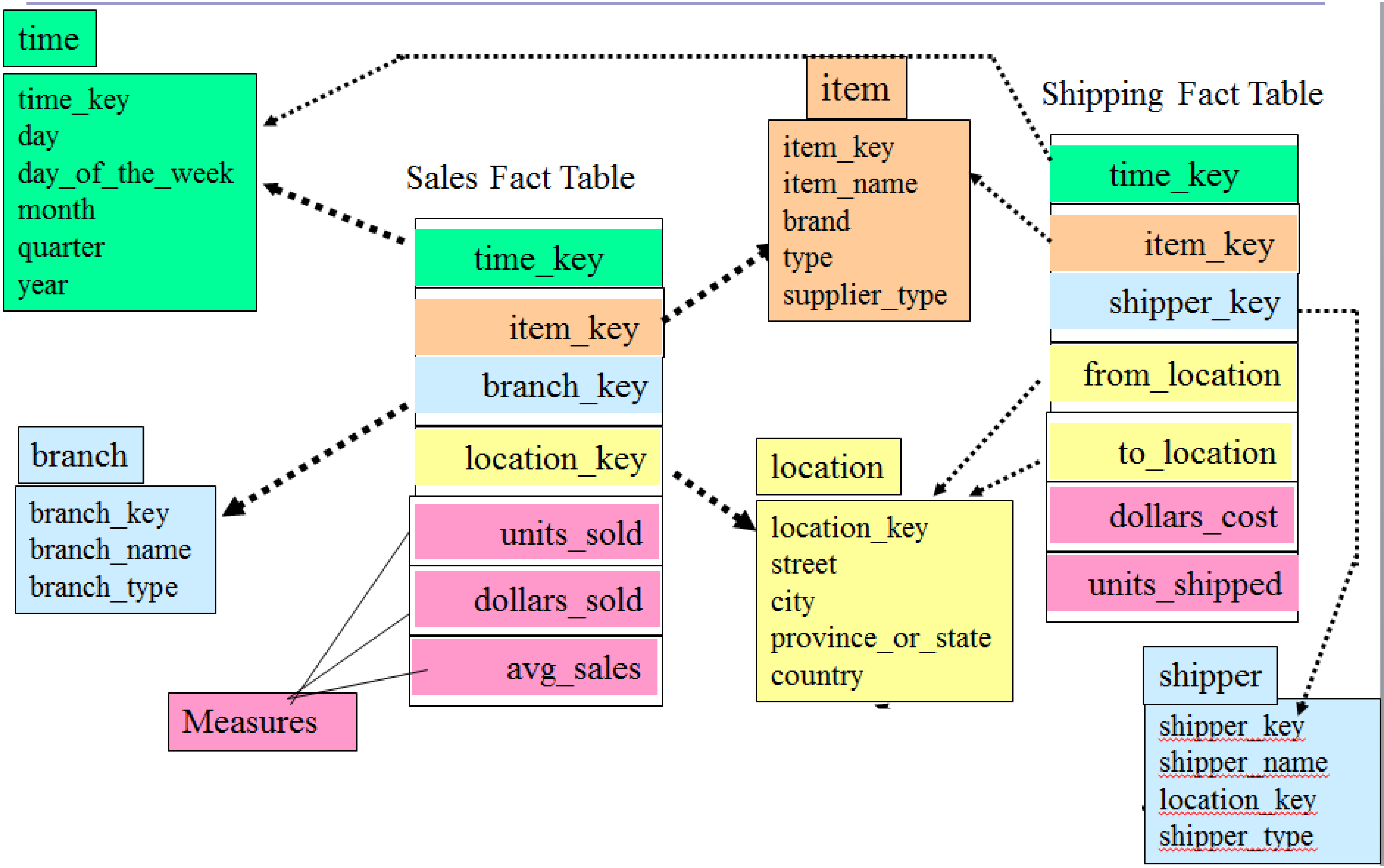
PropertySale

TimeID (FK)
PropertyID (FK)
BranchID (FK)
BuyerID (FK)
PromotionID (FK)
StaffID (FK)
OwnerID (FK)
OfferPrice
SellingPrice
SaleCommission
SaleRevenue

3. Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping fact tables.
- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location and two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Example: Fact Constellation Schema


time
time_key
day
day_of_the_week
month
quarter
year
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales
item
item_key
item_name
brand
type
supplier_type
Shipping Fact Table
time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped
branch
branch_key
branch_name
branch_type
location
location_key
street
city
province_or_state
country
shipper
shipper_key
shipper_name
location_key
shipper_type
Measures