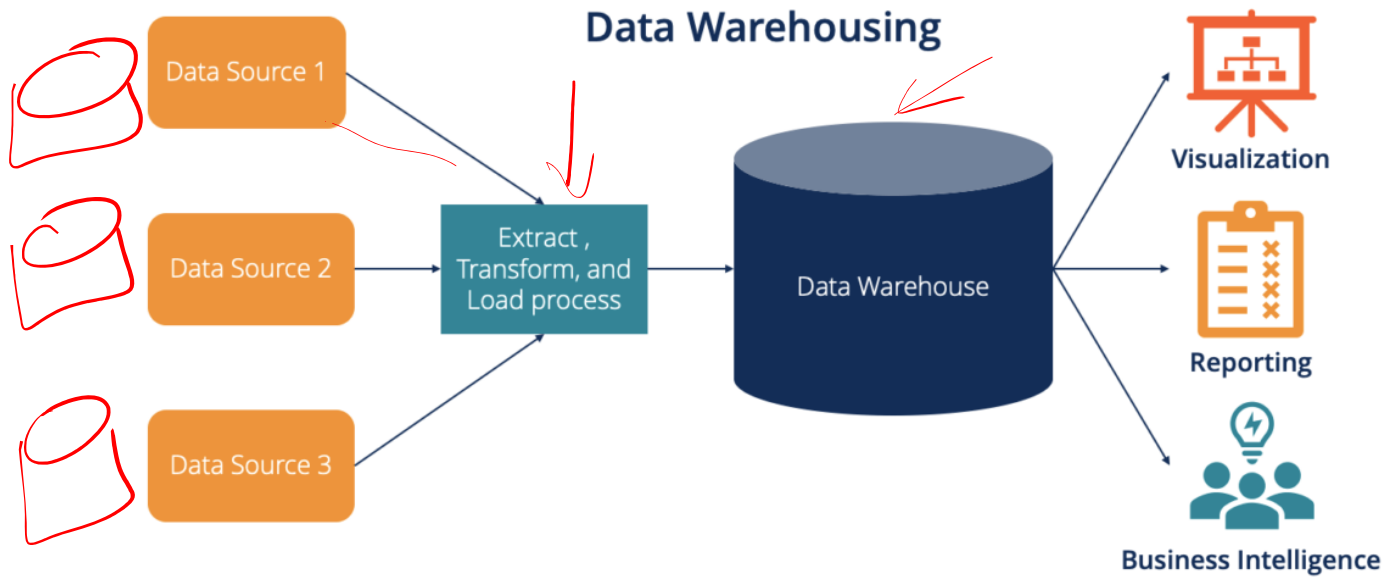
Data Warehousing and Analysis

Agenda

- What is a Data Warehouse?
- Key terminologies:
  - OLTP Vs OLAP
  - ETL
  - Data Mart
  - Metadata
- Data warehouse Architecture
- Data Cube
- Schema
- Data Analysis using Pivot Table

Why Data Warehouse?

- Data collected from various sources and stored in various databases cannot be directly visualized
- The data first needs to be **integrated** and then **processed** before visualization takes places

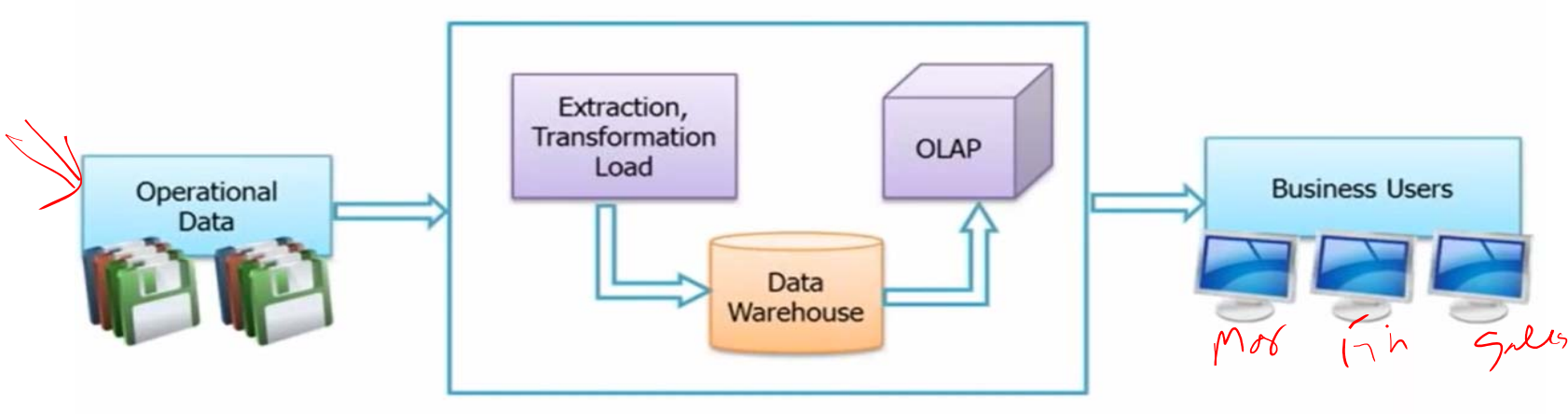


What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - 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 warehousing:
  - The process of constructing and using data warehouses

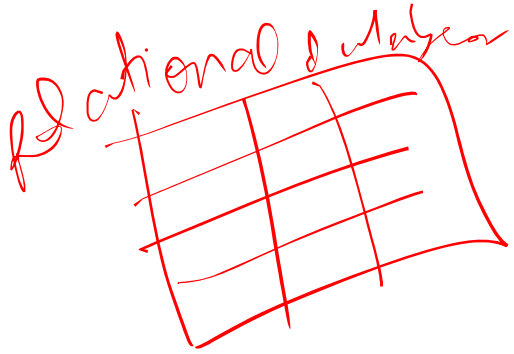
Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision-support process



Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

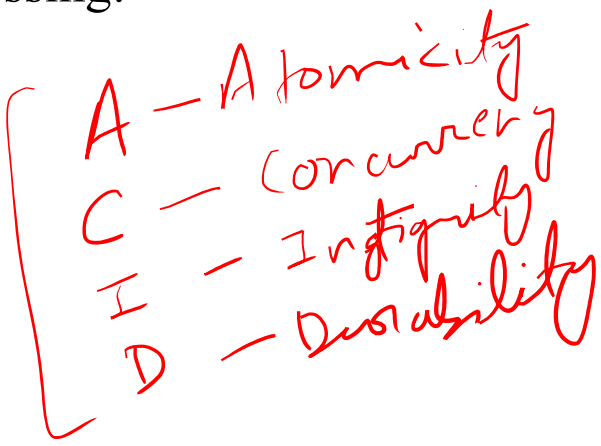


Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - 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*



What are the Advantages of a Data Warehouse?

- Consolidate data
- Data Warehousing is faster and more accurate
- Help in Making better decisions
- Reporting and dashboards
- Analyze your past activities
- Improve time spent by teams
- Boost the company's performance
- Track a marketing campaign
- Consistency and quality of data
- Enables better forecasts

Note: A data warehouse is not a product that a company can go and purchase; it needs to be designed and depends entirely on the company's requirements.

Key Terminologies of Data Warehousing

OLTP (DB) vs. OLAP (DWH)

	Relational Database (OLTP)	Analytical data Warehouse (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	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response
Example	All bank transactions made by a customer	Bank transaction made by a customer at a particular time

OLTP (DB) vs. OLAP (DWH)

OLTP Examples:

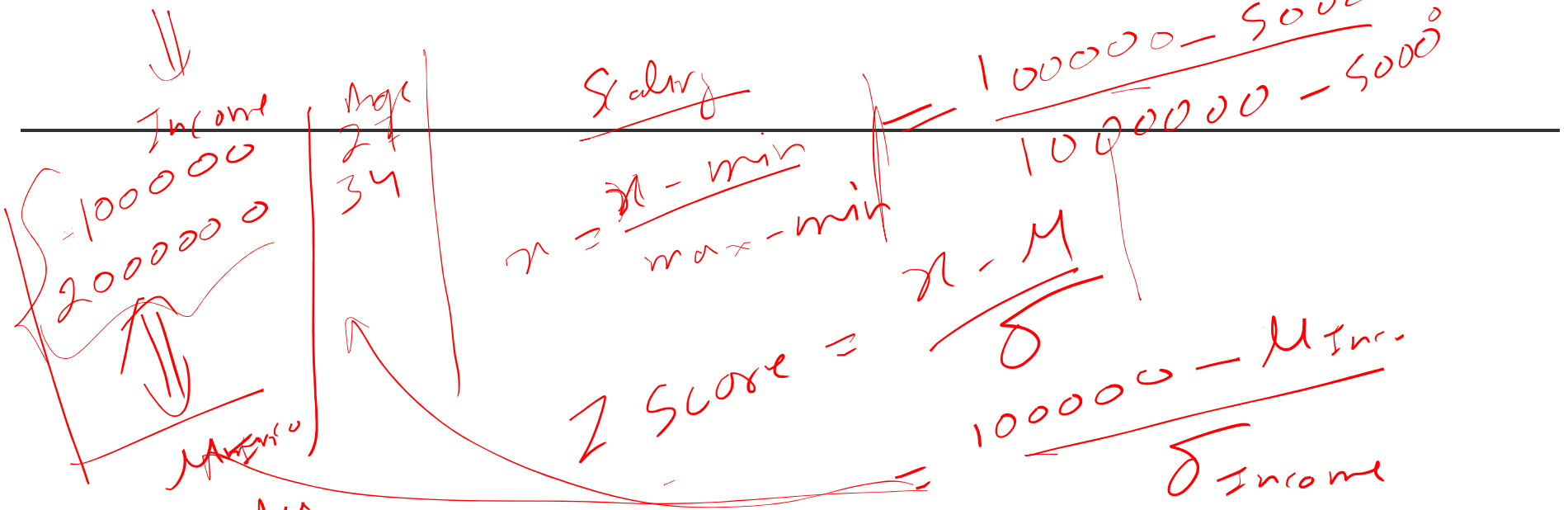
1. **Retail Point of Sale (POS) System:** Processes sales transactions and updates inventory in real-time at checkout counters.
2. **Online Booking System:** Manages reservations and ticket purchases for airlines, hotels, and events instantly.
3. **E-commerce Transaction System:** Handles customer orders, payments, and inventory updates on online shopping websites.
4. **Banking Transaction System:** Processes account deposits, withdrawals, fund transfers, and balance inquiries in real-time.

OLAP Examples:

1. **Sales Data Analysis:** Analyzes historical sales data to identify trends and forecast future sales performance.
2. **Customer Behavior Analysis:** Examines customer purchase patterns to segment customers and develop targeted marketing strategies.
3. **Financial Reporting:** Aggregates and examines financial data to generate comprehensive financial statements and profitability reports.

Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse



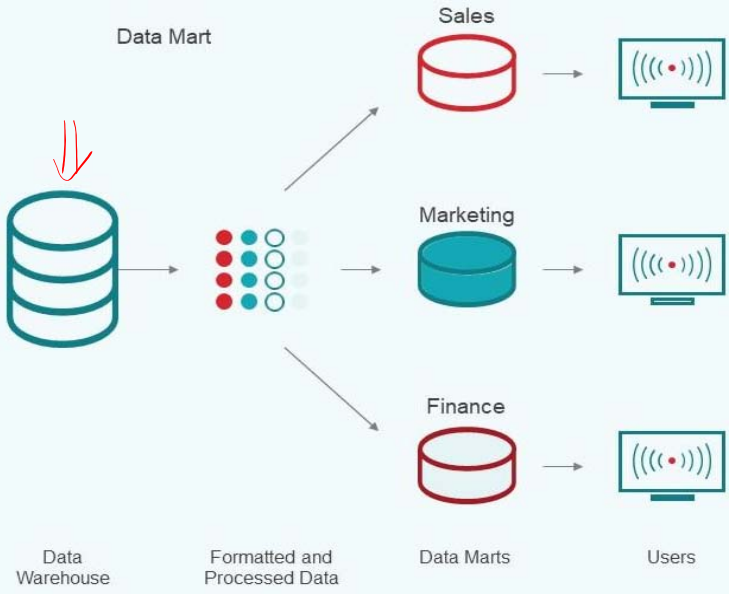
Male/Female 0/1

Rural/Urban/Semi-urban = 1, 2, 3

Data Mart

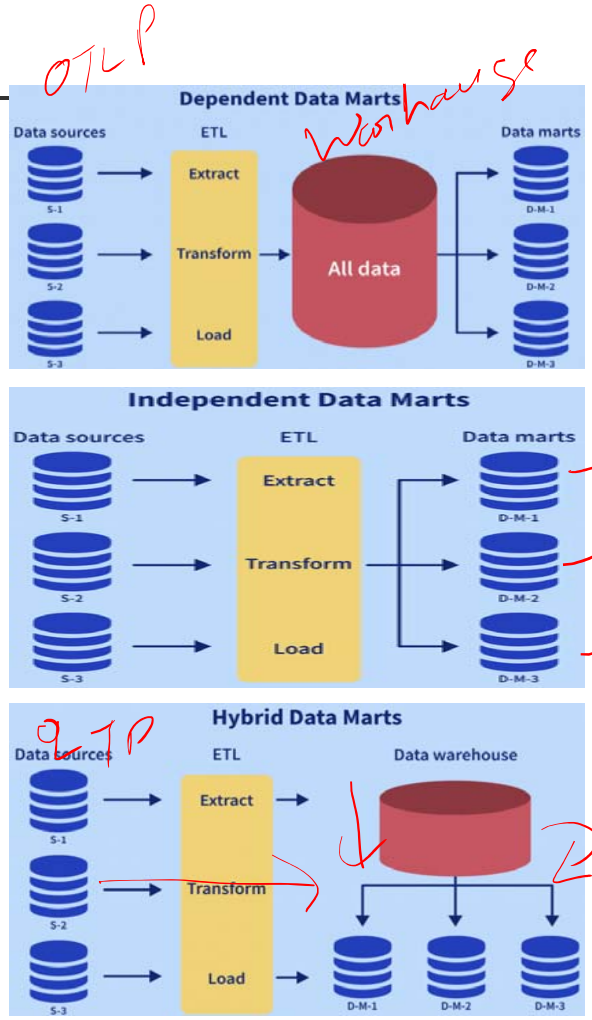
- Data Mart is a smaller version of Data Warehouse which deals with a single subject
- Data marts are focused on one area. Hence, they draw data from a limited number of sources
- Time taken to build Data Marts is very less compared to time taken to build a Data Warehouse

Aspect	Data Warehouse	Data Mart
Scope	Broad, enterprise-wide scope	Narrow, departmental scope
Size	Large, integrates vast amounts of data	Smaller, focuses on specific data sets
Complexity	More complex, handles diverse data types and sources	Less complex, tailored to specific business needs
Implementation Time	Longer implementation time	Quicker to implement due to smaller size
Performance	Designed for comprehensive, enterprise-level queries	Optimized for faster queries within a specific domain



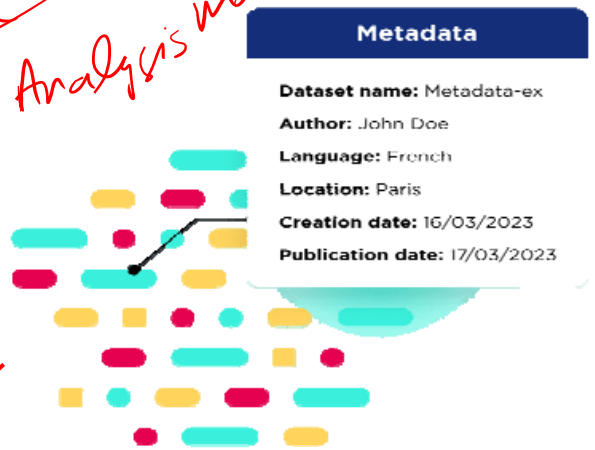
Types of Data Mart

- **Dependent Data Mart**
  - The Data is first extracted from the OLTP systems and then populated in the central DWH
  - From the DWH, the data moved to the Data Mart
- **Independent Data Mart**
  - The data is directly received from the OLTP source system
  - This is Suitable for small organizations or smaller groups within an organization
- **Hybrid Data Mart**
  - The data is fed both from OLTP systems as well as the Data Warehouse



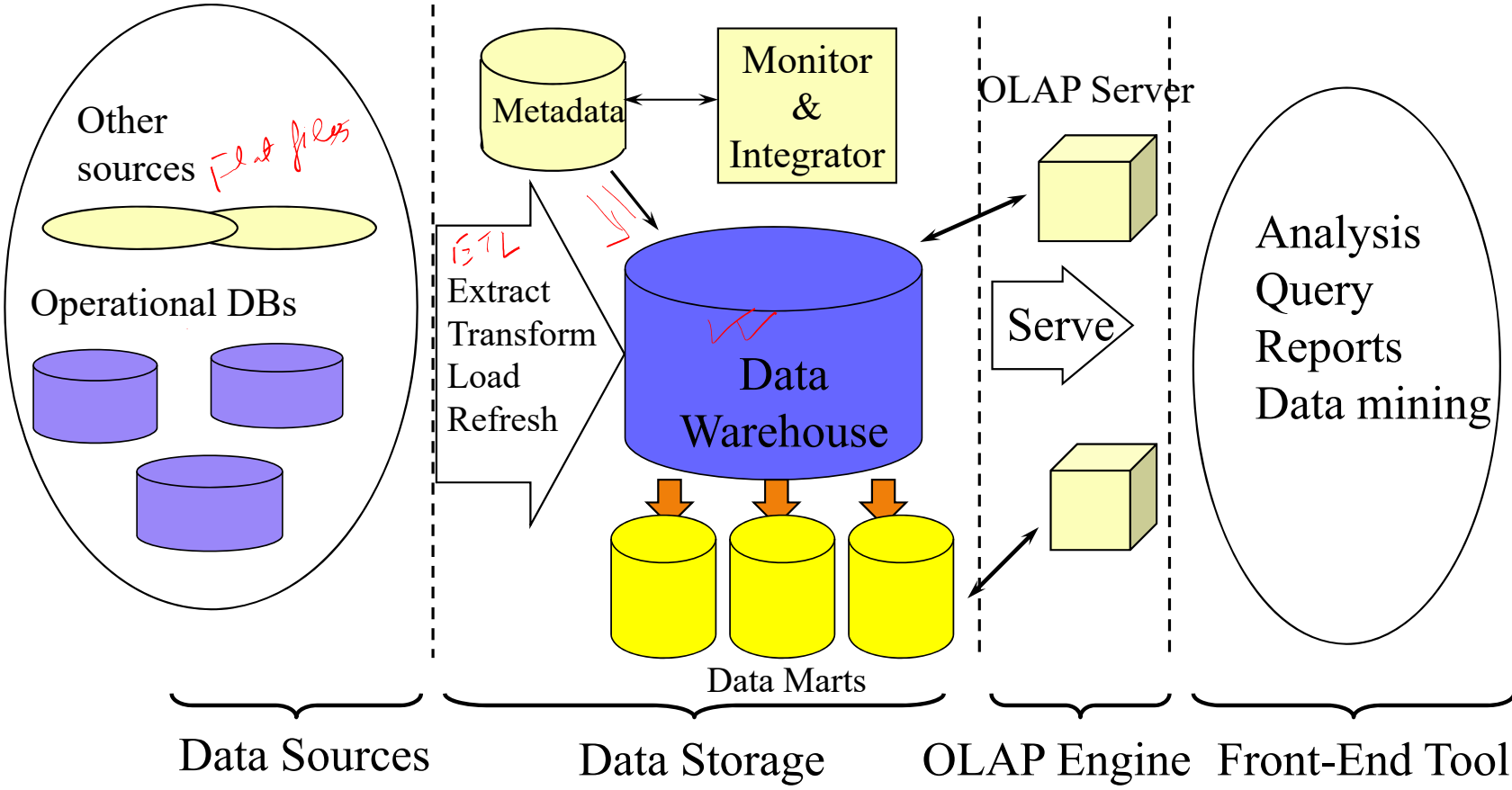
Metadata

- Metadata is defined as data about data
- Meta in a Data Warehouse defines the source data, i.e. Flat file, Relational Database and other objects
- Metadata is used to define which table is the source and target and which concept is used to build business logic called transformation to the actual output



Source: https://www.opendatasoft.com/en/blog/what-is-metadata-and-why-is-it-important-data/

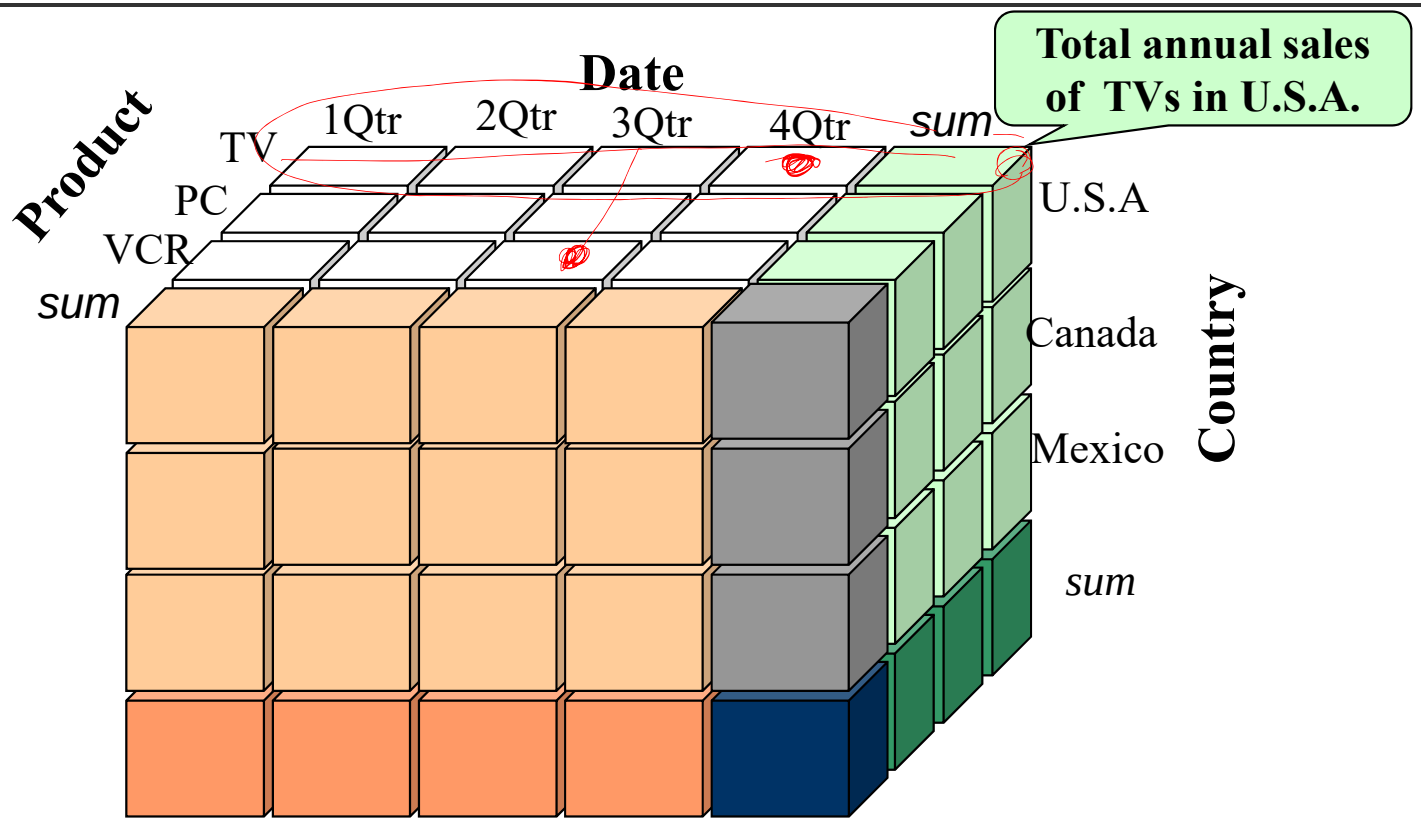
Data Warehouse Architecture



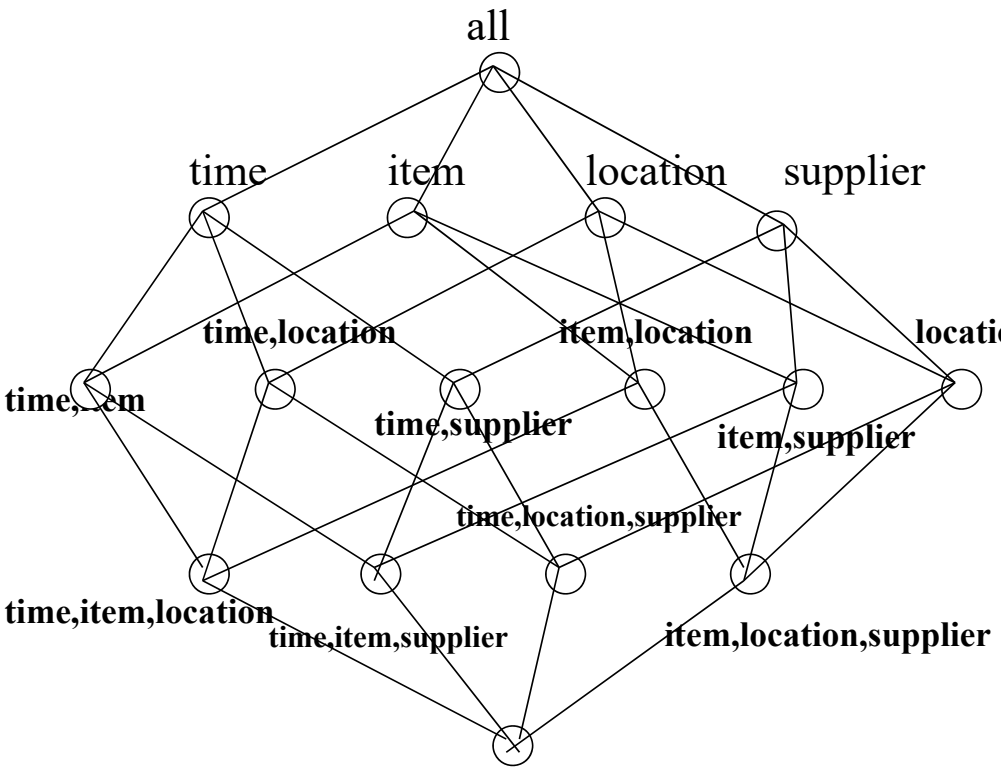
OLAP Data Cubes

- A **data warehouse** is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. 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.

A Sample Data Cube



Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

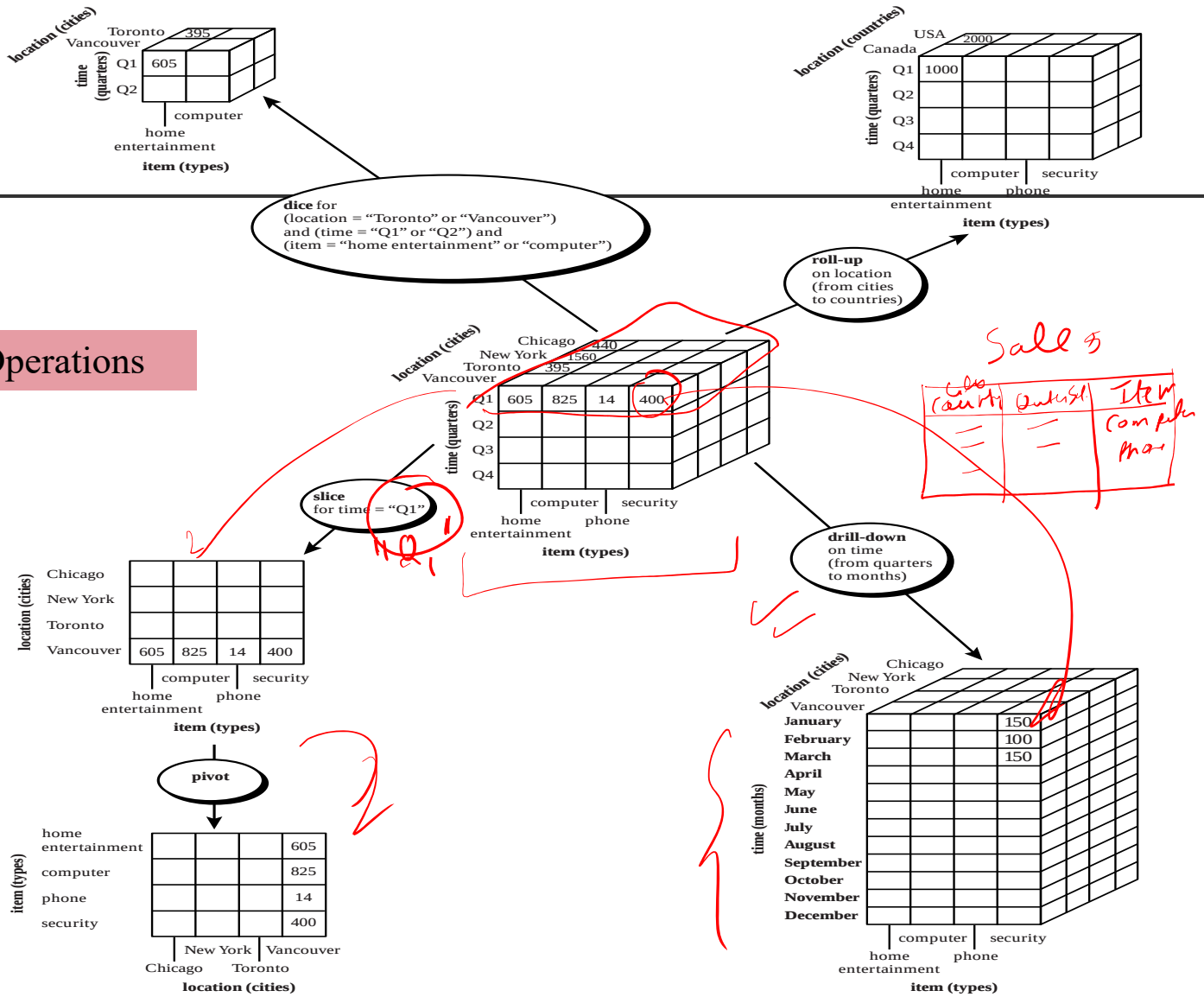
3-D cuboids

4-D (*base*) cuboid

Typical OLAP Operations

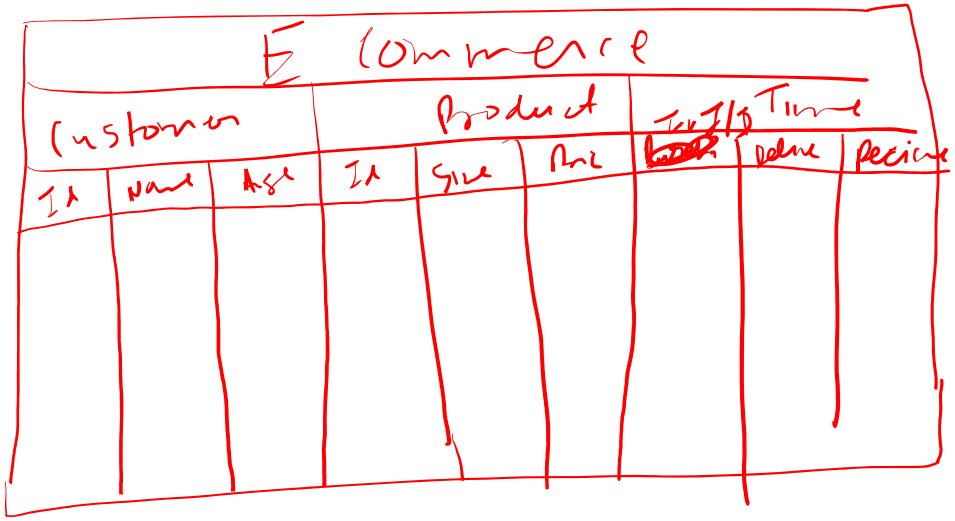
- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice: *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*

Typical OLAP Operations



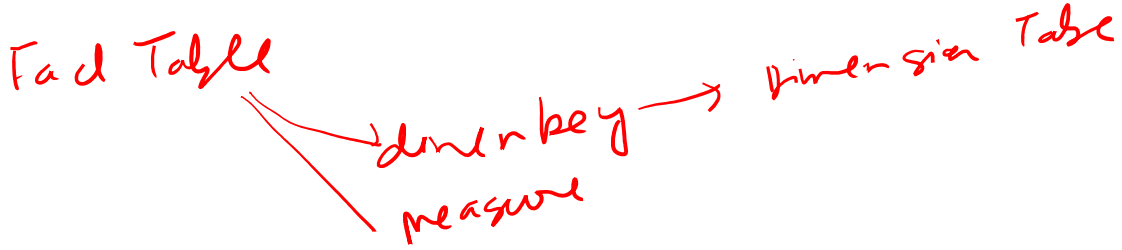
Dimension

- Dimensions - business parameters that define a transaction, relatively static data such as lookup or reference tables
- Example: Analyst may want to view ***sales*** data (measure) by *geography*, by *time*, and by *product* (dimensions)
- The Table that describes the dimensions involved are called **Dimension Tables**
- Dividing a Data Warehouse project into dimensions provides structured information for analysis and reporting

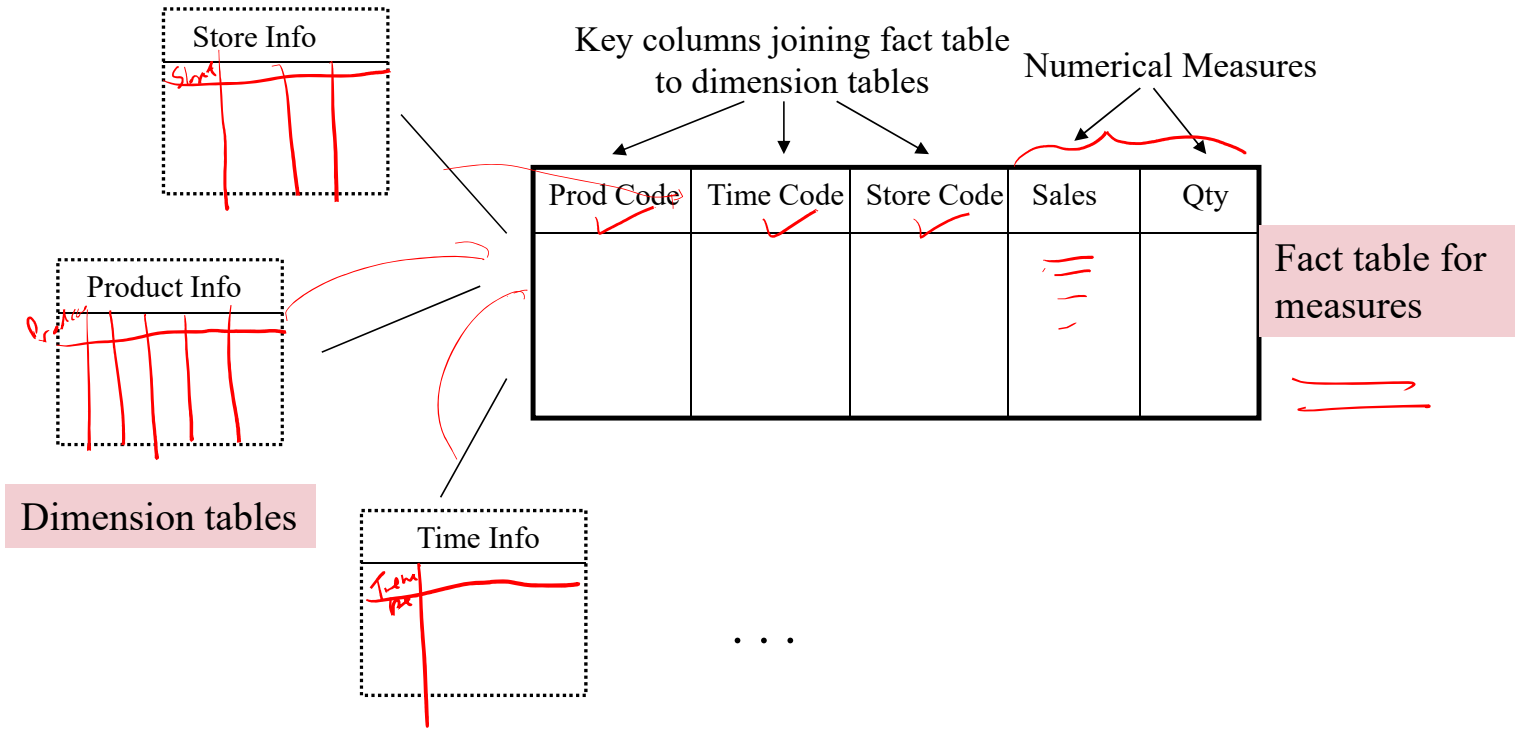

E commerce
Customer
Product
Time
Id
Name
Age
Id
Size
Price
Delivered
Received

Fact and Measures

- Measures - numerical (and additive) data being tracked in business can be analyzed and examined
- A fact is a measure that can be summed, averaged or manipulated
- A fact table contains two kinds of data- a **dimension key** and a **measure**
- Every dimension table is linked to a Fact table



Dimension Table and Fact table



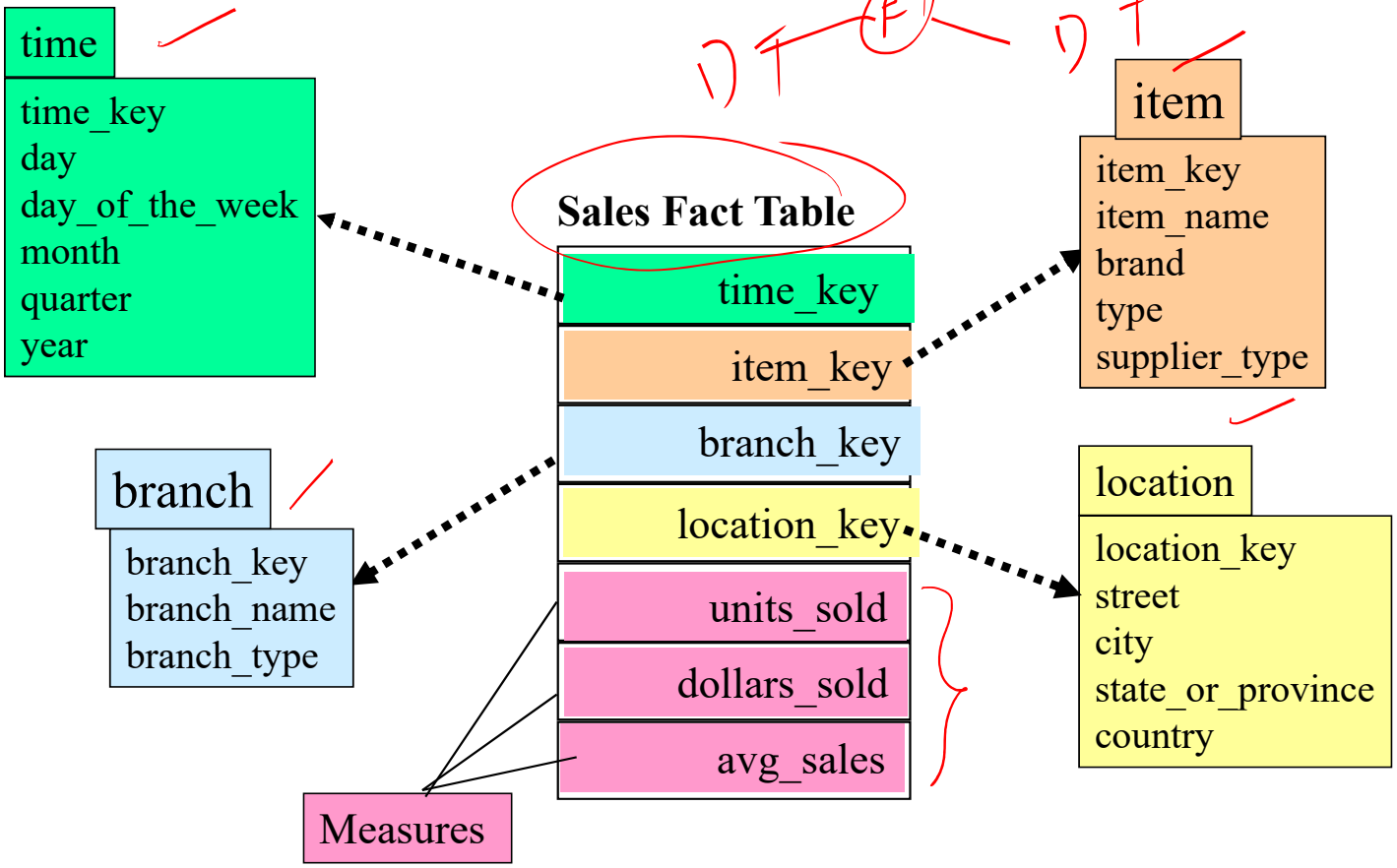
Schemas

- A schema gives a logical description of the entire database
- It gives details about the constraints placed on the tables, key value present and how key values are linked between the different tables
- A database uses a relational model, while a data warehouse uses **Star**, **Snowflake** and **Fact Constellation** schema.

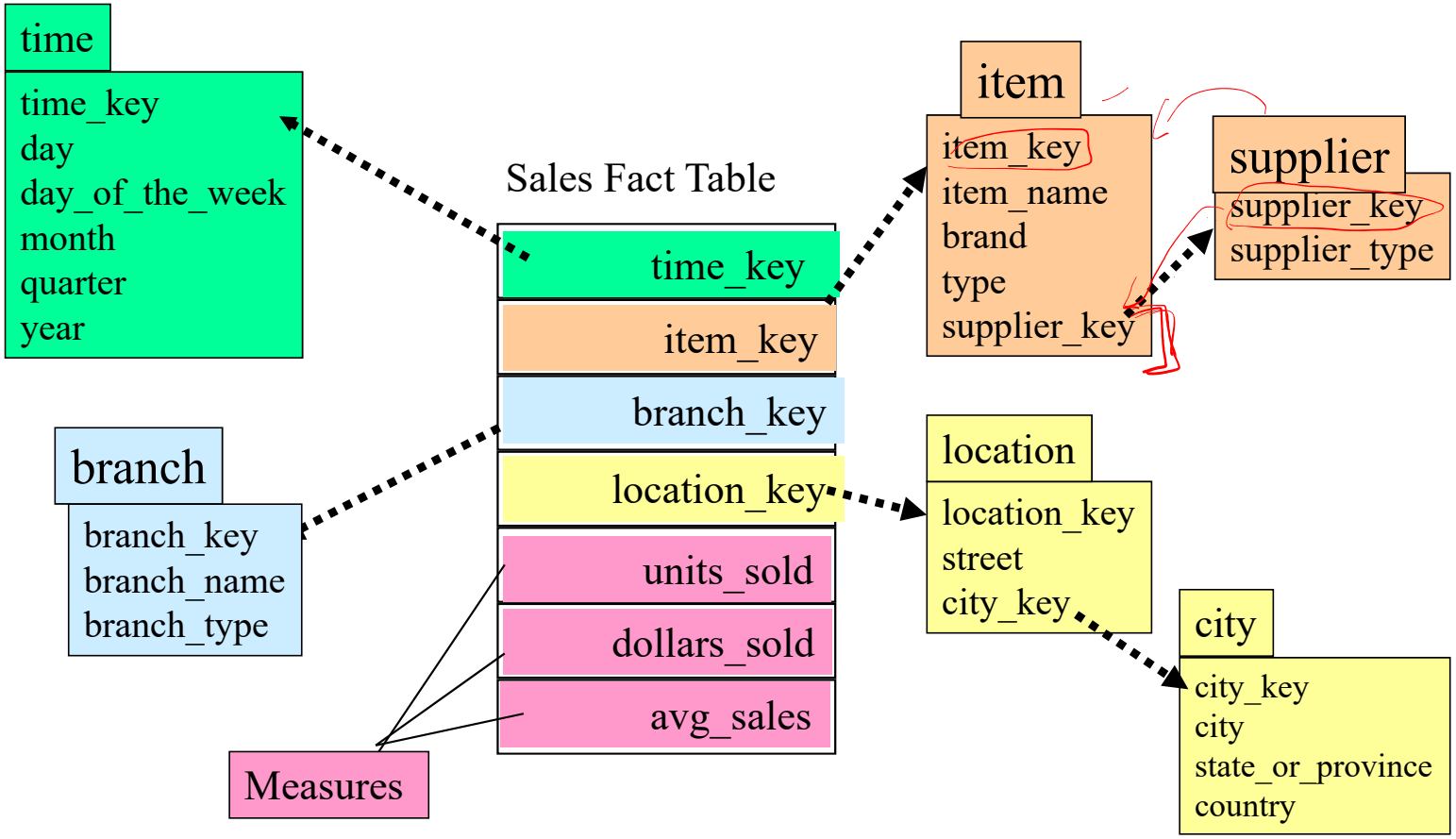
Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - **Star schema**: A fact table in the middle connected to a set of dimension tables
  - **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
  - **Fact constellations**: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

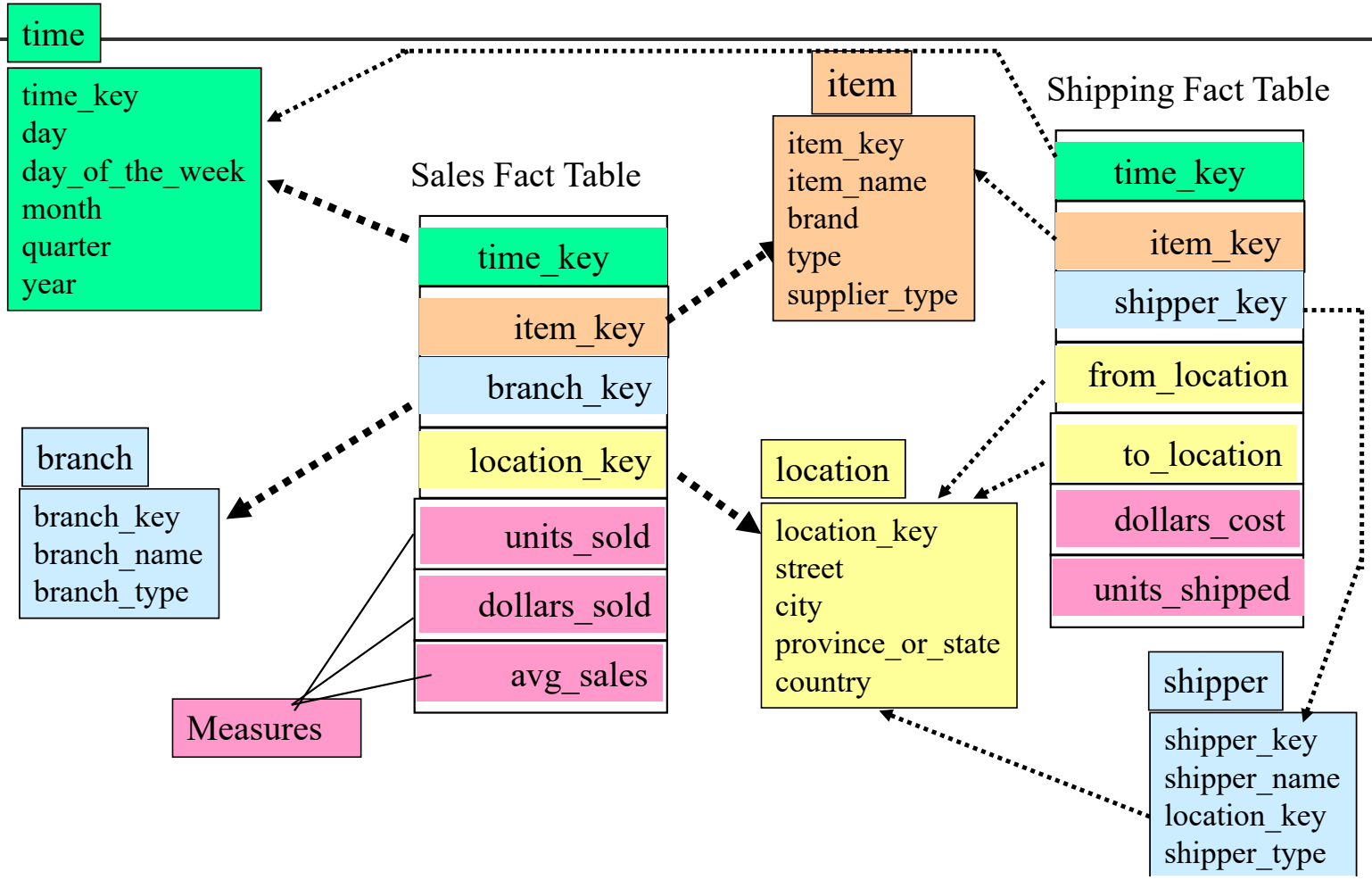
Example of **Star Schema**



Example of **Snowflake Schema**



Example of **Fact Constellation** or **Galaxy Schema**



References

- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
- https://www.oneclickitsolution.com/blog/benefits-of-data-warehouse/
- https://en.wikipedia.org/wiki/Data_mart

$\beta_{10} x_{10} + \epsilon = 0$

$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots$

$= .5 x_1 + 0.1 x_2 + 0.03 x_1 + 0.004 \rightarrow 0.93$

$y = 0.5 x_1 + 0.2 x_2$