Unit 2: Data Preprocessing LH 6

- 2.1. Data Types and Attributes
- 2.2. Data Pre-processing
- 2.3. OLAP
- 2.4 Characteristics of OLAP Systems
- 2.5 Multidimensional View and Data cube
- 2.6 Data Cube Implementation
- 2.7 Data Cube Operations
- 2.8 Guidelines for OLAP Implementation

Data Types and Attributes

- A data is a known fact that can be recorded and have implicit (not directly express but implemented) meaning. For example, the names, eye color, telephone numbers, and addresses of the students of a class. So, data is a collection of data objects and their attributes.

- A collection of attributes describe an object.

Data Object: (Database Row)

- Data sets are made up of data objects. A data object represents an entity.

Examples:

- sales database: customers, store items, sales
- medical database: patients, treatments
- university database: students, professors, courses
- **Data objects are described by attributes.**

Attribute (Database Columns)

- An attribute is a property or characteristic of an object.
- Attribute values are numbers or symbols assigned to an attribute. Different attributes can be mapped to the same set of values. Example: Attribute values for ID and age are integers but properties of attribute values can be different. ID has no limit but age has a maximum and minimum value.

A data object

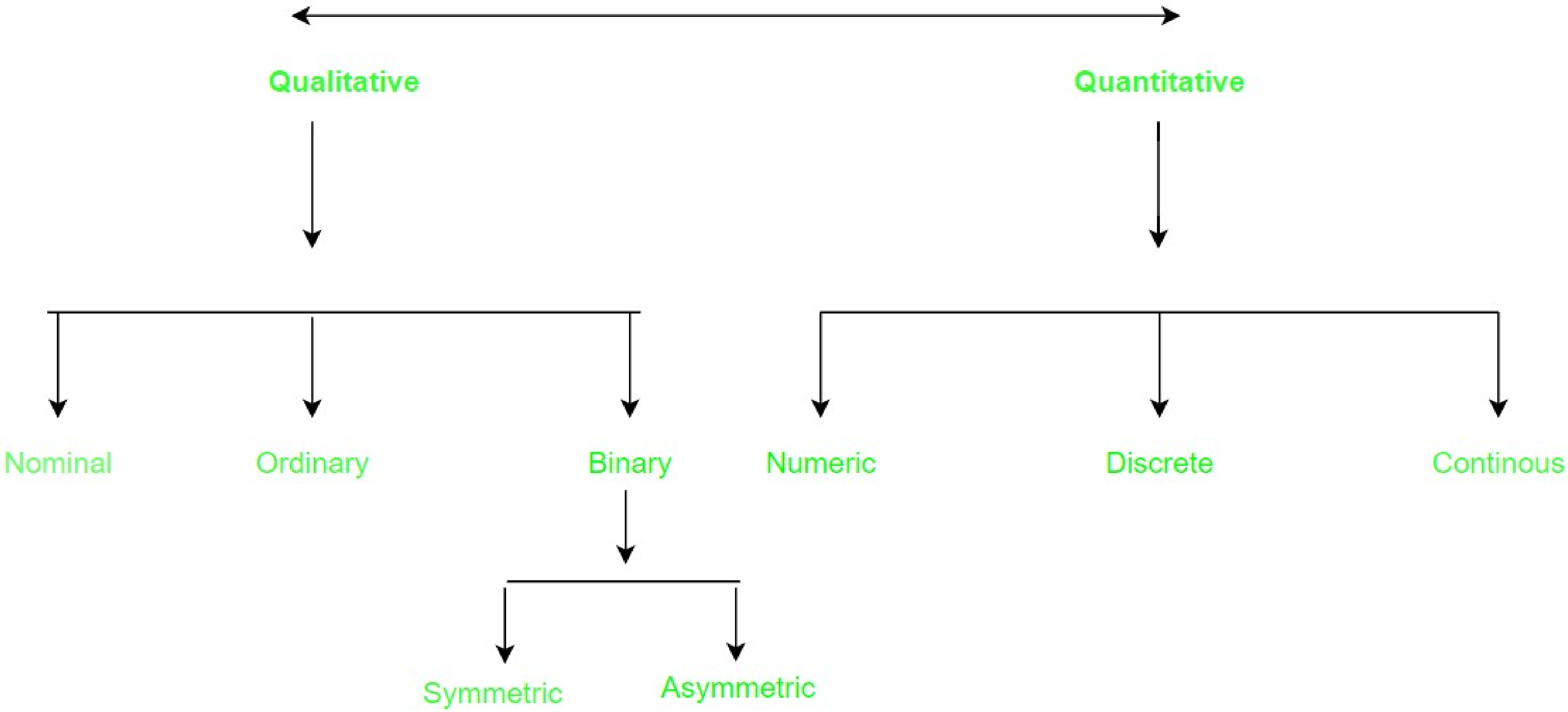
Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- database rows ➔ data objects
- database columns ➔ attributes

Types OF Attributes



1. **Categorical Attributes** Categorical attributes are also called **qualitative attributes**. Categorical attributes lack most of the properties of numbers even if they are represented by numbers. If these attributes are integers, they should be treated more like symbols.

   I. **Nominal Attribute (Categorical Attribute):**
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Name of things, some kind of symbols. Value of Nominal attribute represents some category or state.

The values of nominal attributes **do not have any meaningful order.**
• Example: The attribute marital_status can take on the values single, married, divorced, and widowed.
• These are categorical and qualitative , but not quantitative.
The = (Equal operator can be use)

Ordinal Attribute:
The ordinal attribute value provides sufficient i**nformation to order the objects**. It is group based such as Rankings, Grades. The values of ordinal attributes **have meaningful order but not quantify(For eg Pain, Low, High, extreme it is order but the difference between Low, High is not Quntifiable)**

Operator: = , < , > can be used

Binary Attribute:
These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.

a) Symmetric : Both value equally important (gender)
b) Asymmetric : Both Value are not equally important (Result of Cancer Test)

2.Numeric attribute: It is **quantitative**, such that quantity can be measured and represented in integer or real values ,are of two types

a)Interval Scaled attribute:

b)It is measured on a **scale of equal size units (Measure in Linear scale unit)**.

- Have order and can be +ve, 0 or -ve
- These attributes allow us to compare and quantify the difference between values such as temperature of two different days.

b) **Ratio Scaled attribute:**

Numeric Attribute with fixed zero point. Ratio scaled means being multiple or ratio of another value. We can compute mean median, mode with this type of attribute.
Both differences and ratios are significant for Ratio. For eg. age, length, and Weight.

Discrete :

- Have finite Values or countable infinite set of values, it can be numerical and can also be in categorical form.
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

Attrinute	Values
Profession	Teacher, programmer, banker
ZIP code	33700, 33800

Continuous:

- Infinite number of states, It is of float types. There can be any value between 0-1.
- Continuous attributes are typically represented as floating-point variables
- Height : 54.5kg

Attribute type		Description	Example	Properties
Categorical (Qualitative)	Nominal	The values of nominal attributes are just different symbol or name of things	zip codes, employee ID numbers, gender	Distinctness
	Ordinal	The values of ordinal attributes are possible values that have meaningful order or ranking among them	order of finishing a race, grades in exam	Distinctness and order

Numeric (Quantitative)	Interval	Interval scaled attributes are measured on a scale of equal size units	calendar dates, temperature in Celsius or Fahrenheit	Distinctness, order and Addition
	Ratio	The values of ratio scaled attributes has a true zero point	frequency of words in a document	Distinctness, order, addition and multiplication

Data Pre-processing:

Preprocessing data is a data mining technique for transforming raw data into a usable and efficient format. The measures taken to make data more suitable for data mining are referred to as data preprocessing.

Data taken directly from the source will contain errors, contradictions, and, most importantly, will not be willing to be used for a data mining tool.

We need to preprocess the data because of the following reasons:

1. Data in the real world is dirty which means the data is :
    - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data, **e.g., Occupation=“ ” (missing data)**
    - **noisy**: containing errors or outliers, **eg salary = “-10”**
    - **inconsistent**: containing discrepancies in codes or names eg salary = “100000” and tax = “100”
    - **Intentional** : disguised missing data, eg Jan 1 as eveyone DOB

2. No quality data, no quality mining results!
    - Quality decisions must be based on quality data
    - Data warehouse needs consistent integration of quality data

Importance of Data Pre-Processing

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

Accuracy: To check whether the data entered is correct or not.

Completeness: To check whether the data is available or not recorded.

Consistency: To check whether the same data is kept in all the places that do or do not match.

Timeliness: The data should be updated correctly.

Believability: The data should be trustable.

Interpretability: The understandability of the data

Major Task in Data Preprocessing

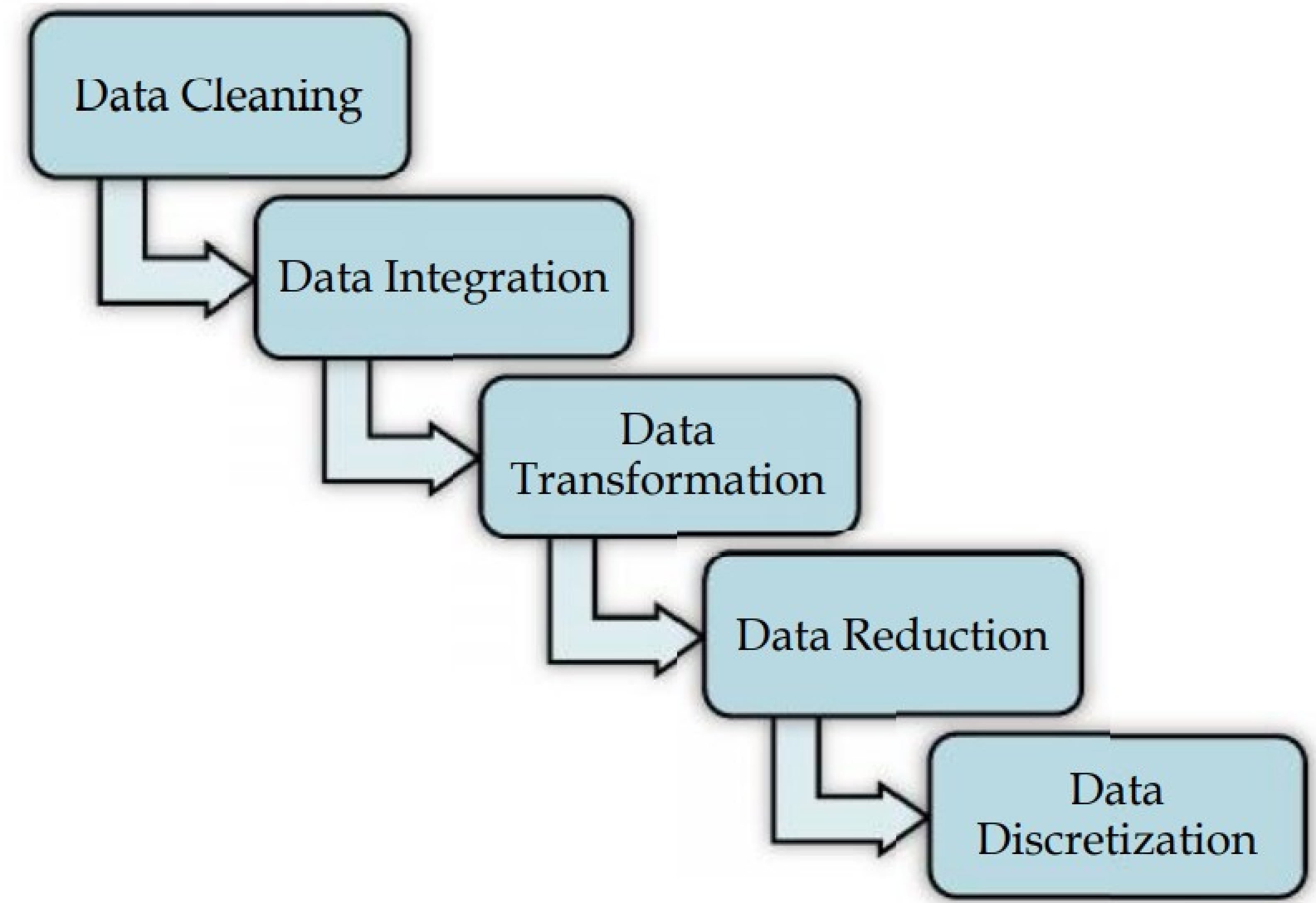


Fig: Data Preprocessing Tasks

1.0 Data Cleaning Task

- Data Cleaning Task When it comes to the final review, the quality of the data is crucial. Any data that is incomplete, noisy, or inconsistent may have an impact on your final result. The process of detecting and deleting corrupt or incorrect records from a record collection, table, or database is known as data cleaning in data mining. **The following are the data cleaning tasks:**

  - Fill in missing values **(Handling Missing Values)**
  - Identify outliers and smooth out noisy data (**Handling Noisy Data**)
  - Correct inconsistent data

Missing data is not always available. Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry

Handling Missing Data

1. **Forget the tuple**. If the classmark is absent, this is finished. Unless the tuple includes multiple attributes with missing values, this approach is ineffective.

2. You can manually fill in the missing value. This method works well with small data sets that have some missing values.

3. You may use a global constant like "Unknown" or minus infinity to replace all missing attribute values.

4. To fill in the missing value, use the attribute mean. If a customer's average income is $25,000, you can use this amount to fill in the lost income value.

5. Fill in the missing value with the most likely value.

Handling Noisy Data:

- A **random error or deviation in a calculated variable** is referred to as noise. Noisy data may occur as a result of defective data collection instruments, data entry issues, or technological limitations.

Way Deal With Noisy Data? **(Binning, Regression, Clustering)**

- **Binning:** Binning approaches sorted data values by consulting their "neighborhood," or the values in their immediate vicinity. The sorted values are divided into several "**buckets" or bins.**

Equal-depth (frequency) partitioning:

- It divides the range into N intervals, each containing the approximately same number of samples. Managing Sample size is tricky

Equal-width (distance) partitioning:

- It divides the range into N intervals of equal size: uniform grid
- if L and H are the lowest and highest values of the attribute, the width of intervals will be $W = (H – L)/N$

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

First, the data is sorted then and then the sorted values are separated and stored in the form of bins.

Partition into (equi-depth) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

Smoothing by bin means: (In this method, the values in the bin are replaced by the *mean* value of the bin;**)**

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries: (In this method, the using *minimum* and *maximum values* of the bin values are taken and the values are replaced by the *closest boundary value***)**

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

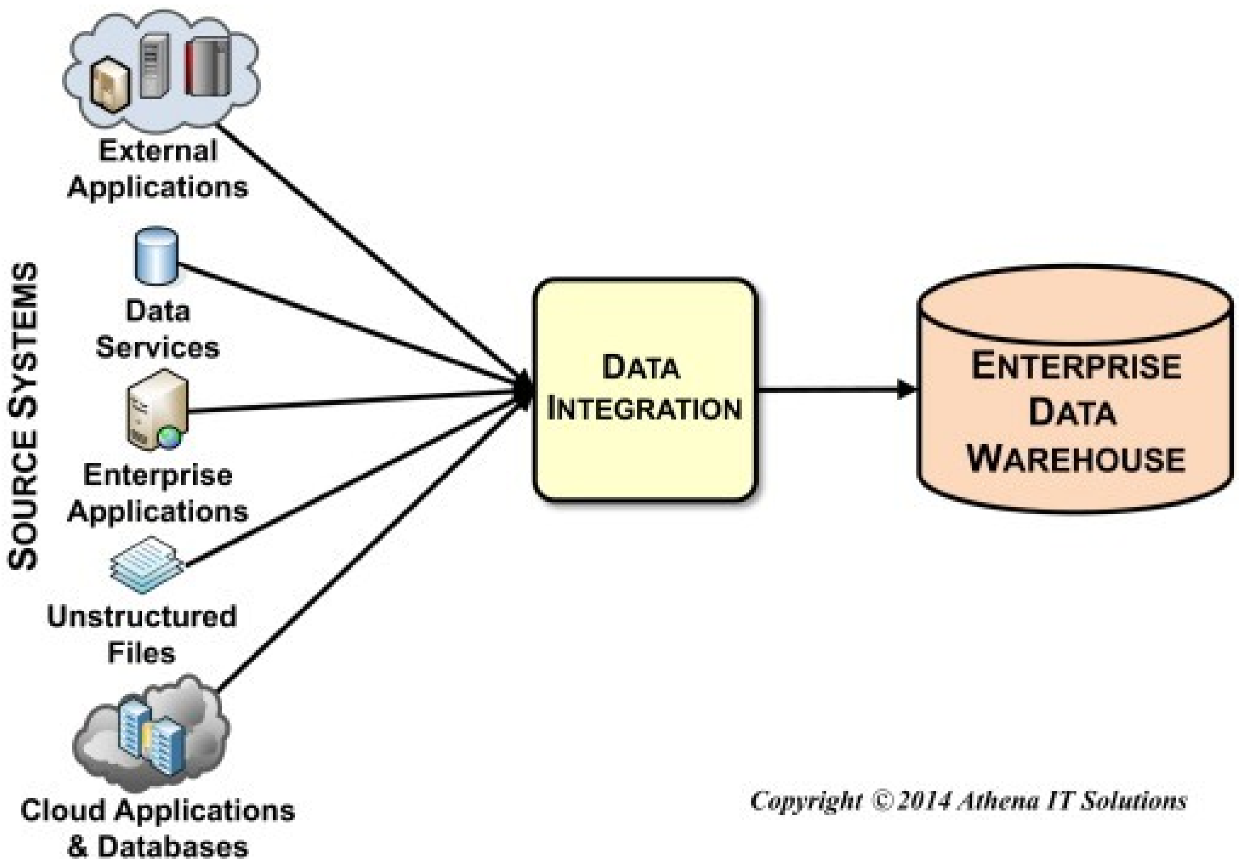
Regression:

- Data can be smoothed by fitting the data to a function, such as with regression.
- Linear regression involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other.
- Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Clustering:

- Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

- **2.0 Data Integration:** It combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files.
- The data integration systems are formally defined as triple <G, S, M>
  - Where G: The global schema
  - S:Heterogeneous source of schemas
  - M: Mapping between the queries of source and global schema.



Issue in Data Integration:

- **Schema integration and object matching:**
  - How can the data analyst or the computer be sure that customer id in one database and customer number in another reference to the same attribute.

- **Redundancy**:
  - An attribute (such as annual revenue, for instance) may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.
  - Example: If one data set contains the customer's age and another data set contains the customer's date of birth, age will be a redundant attribute since the date of birth could be used to derive it.
  - **The data redundancy problem can be handled by correlation analysis**

- **Detection and resolution of data value conflicts:**

➢ For the same real-world entity, attribute values from different sources may differ. For example Date format “MM/DD/YYYY” or “DD/MM/YYYY”

3.0 Data Transformation:

- In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

- **Smoothing**, which works to remove noise from the data. Such techniques include binning, regression, and clustering.

- **Aggregation**, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts

- **Generalization** of the data, where low-level or primitive (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country.

- **Normalization**, where the attribute data are scaled so as to fall within a small specified range, such as -1:0 to 1:0, or 0:0 to 1:0.

- **Attribute construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

- **4.0 Data Reduction**
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results

Strategies for data reduction include the following:

- **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube. It combine two or more objects into a single object. (Pokhara, Baglung, Damauli- Gandaki Pradesh)

- **Attribute subset selection**, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

- **Dimensionality reduction**, where encoding mechanisms are used to reduce the dataset size. An example is to treat male or female for gender as 1 or 0.

- **Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.

- **5.0 Discretization (concept hierarchy generation).**

- It is **the process of putting values into buckets so that there are a limited number of possible states**. Convert a huge number of data values into smaller ones so that the evaluation and management of data become easy.

Suppose we have an attribute of Age with the given

Age	1,5,9,4,7,11,14,17,13,18,9,31,33,36,42,44,46,70,74,78,77

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

OLTP vs OLAP

	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
metric	transaction throughput	query throughput, response

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
- Other operations
  - drill across: involving (across) more than one fact table
  - drill through: through the bottom level of the cube to its back-end relational tables (using SQL)

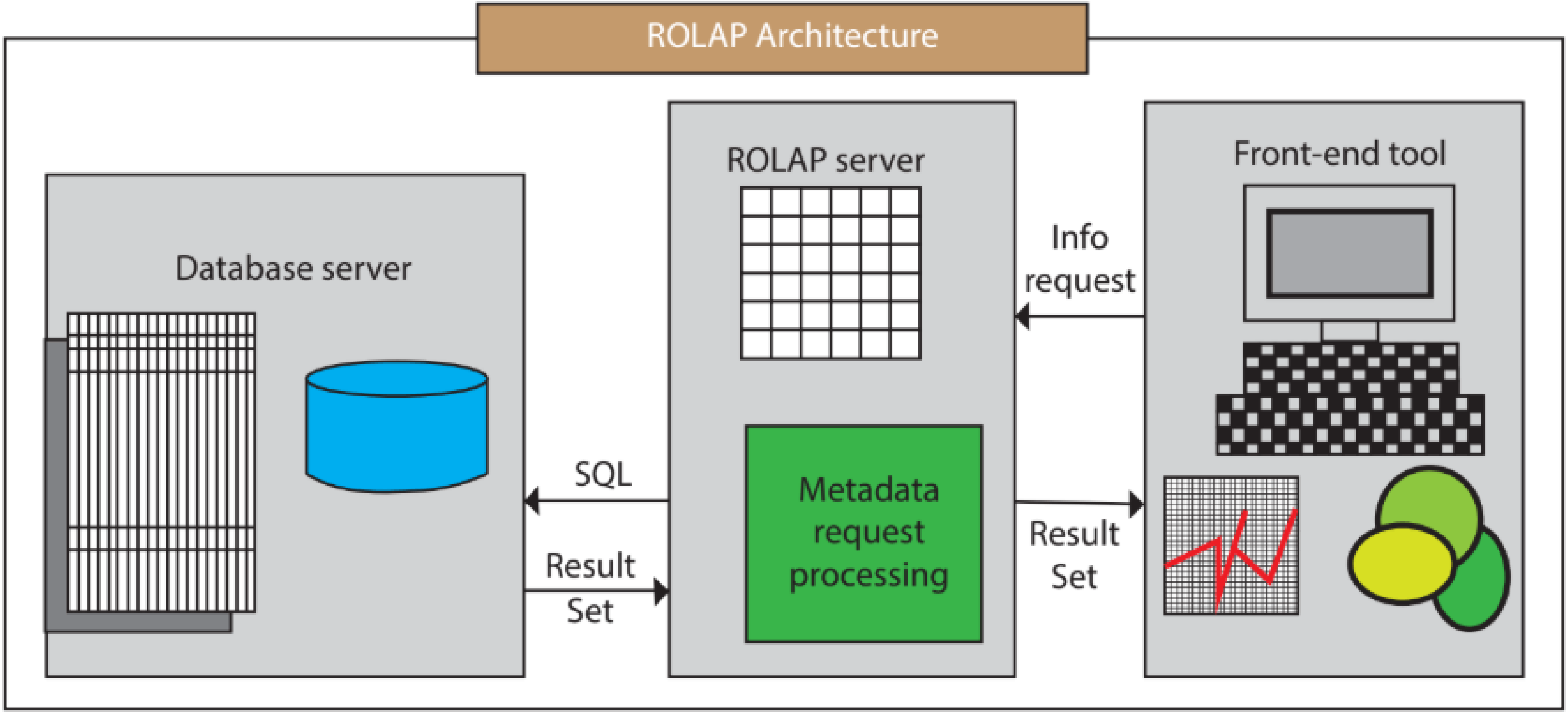
25

OLAP Server Architectures

- **Relational OLAP (ROLAP)**
  - These are the intermediate servers that stand in **between a relational back-end server and client front-end tools.**
  - It provides functionality by using relational databases and relational query tools to store and analyze multidimensional data.
  - One **advantage** of ROLAP over the other styles of OLAP analytic tools is that it is deemed to be more scalable in handling huge amounts of data.

  Disadvantage

  - ROLAP applications display a slower performance as compared to other style of OLAP tools since, oftentimes, calculations are **performed inside the server.**
  - Another demerit of a ROLAP tool is that as **it is dependent on use of SQL for data manipulation,** it may not be ideal for performance of some calculations that are not easily translatable into an SQL query.


ROLAP Architecture
Database server
ROLAP server
Front-end tool
Info request
SQL
Result Set
Metadata request processing
Result Set

Multidimensional OLAP (MOLAP)

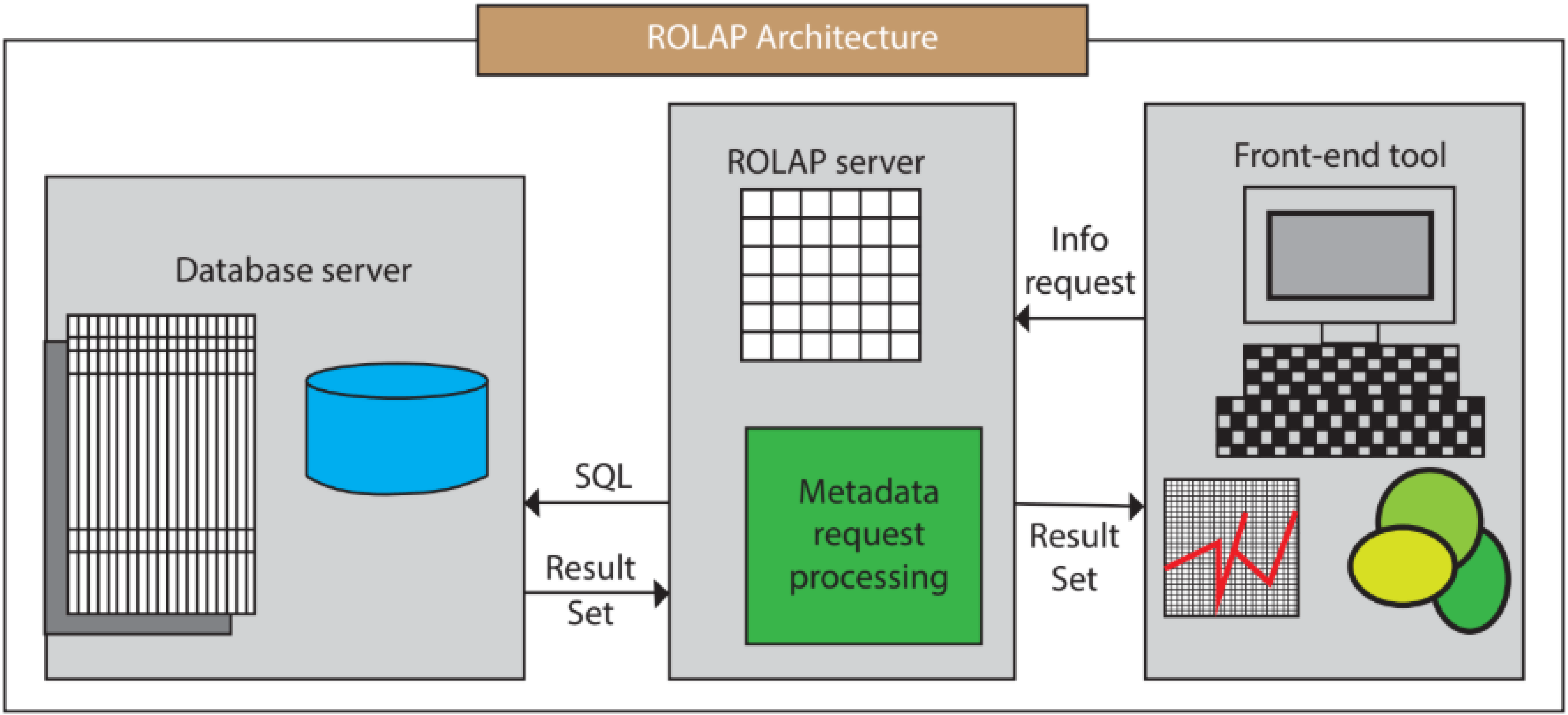
- **MOLAP** structure primarily reads the precompiled data. MOLAP structure has limited capabilities to dynamically create aggregations or to evaluate results which have not been pre-calculated and stored.
- Applications requiring iterative and comprehensive time-series analysis of trends are well suited for MOLAP technology (e.g., financial analysis and budgeting).

Advantages

- **Excellent Performance:** A MOLAP cube is built for fast information retrieval, and is optimal for slicing and dicing operations.
- **Can perform complex calculations:** All evaluation have been pre-generated when the cube is created. Hence, complex calculations are not only possible, but they return quickly.

Disadvantages

- **Limited in the amount of information it can handle:** Because all calculations are performed when the cube is built, it is not possible to contain a large amount of data in the cube itself.
- **Requires additional investment:** Cube technology is generally proprietary and does not already exist in the organization. Therefore, to adopt MOLAP technology, chances are other investments in human and capital resources are needed.


ROLAP Architecture
Database server
SQL
Result
Set
ROLAP server
Metadata
request
processing
Info
request
Result
Set
Front-end tool

Hybrid Online Analytical Processing (HOLAP) :

- Hybrid is a combination of both ROLAP and MOLAP. It offers functionalities of both ROLAP and as well as MOLAP like faster computation of MOLAP and higher scalability of ROLAP. The aggregations are stored separately in MOLAP store. Its server allows storing the large data volumes of detailed information.

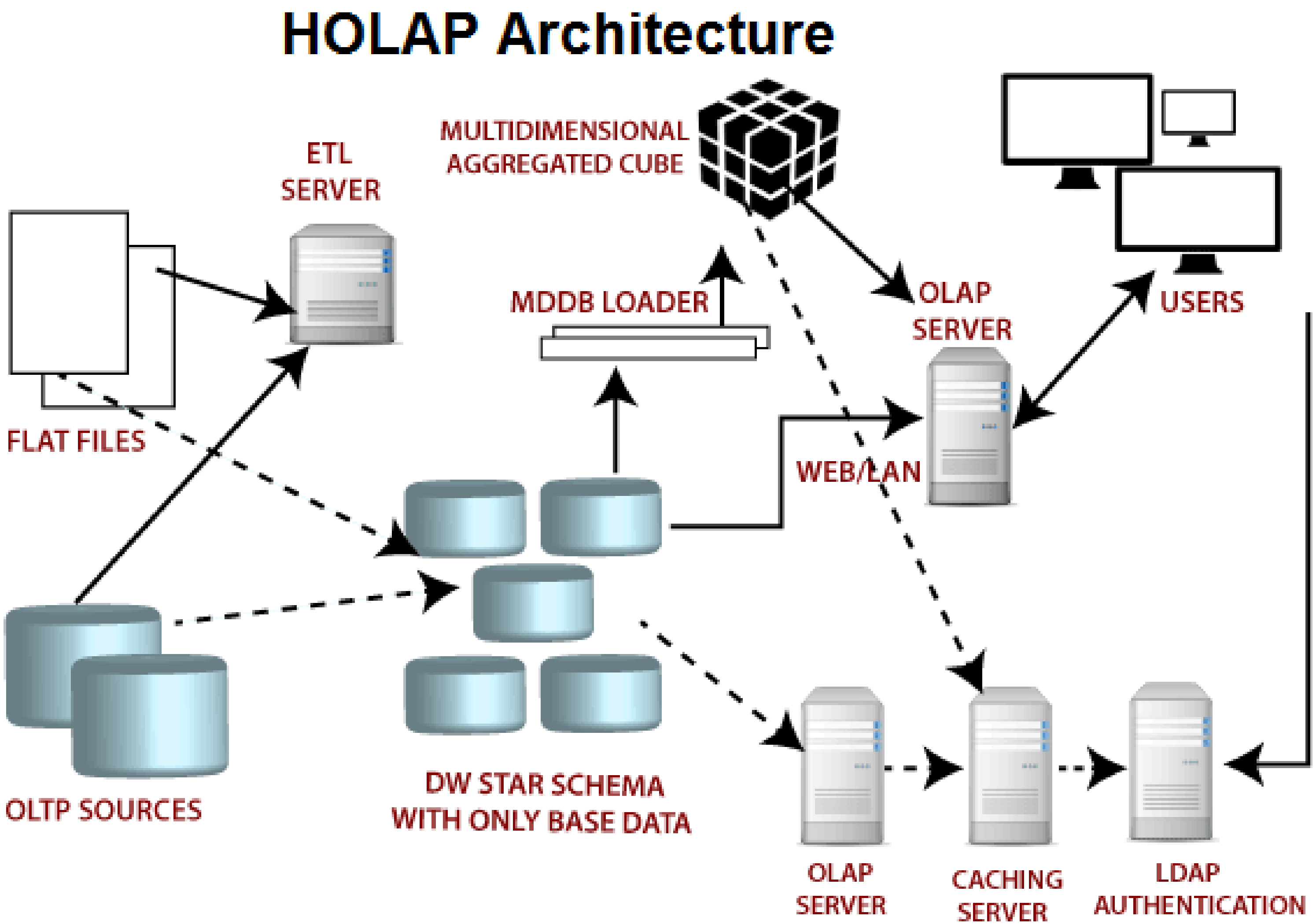
Advantages of HOLAP –

- HOLAP provides the functionalities of both MOLAP and ROLAP.
- HOLAP provides fast access at all levels of aggregation.

Disadvantages of HOLAP –

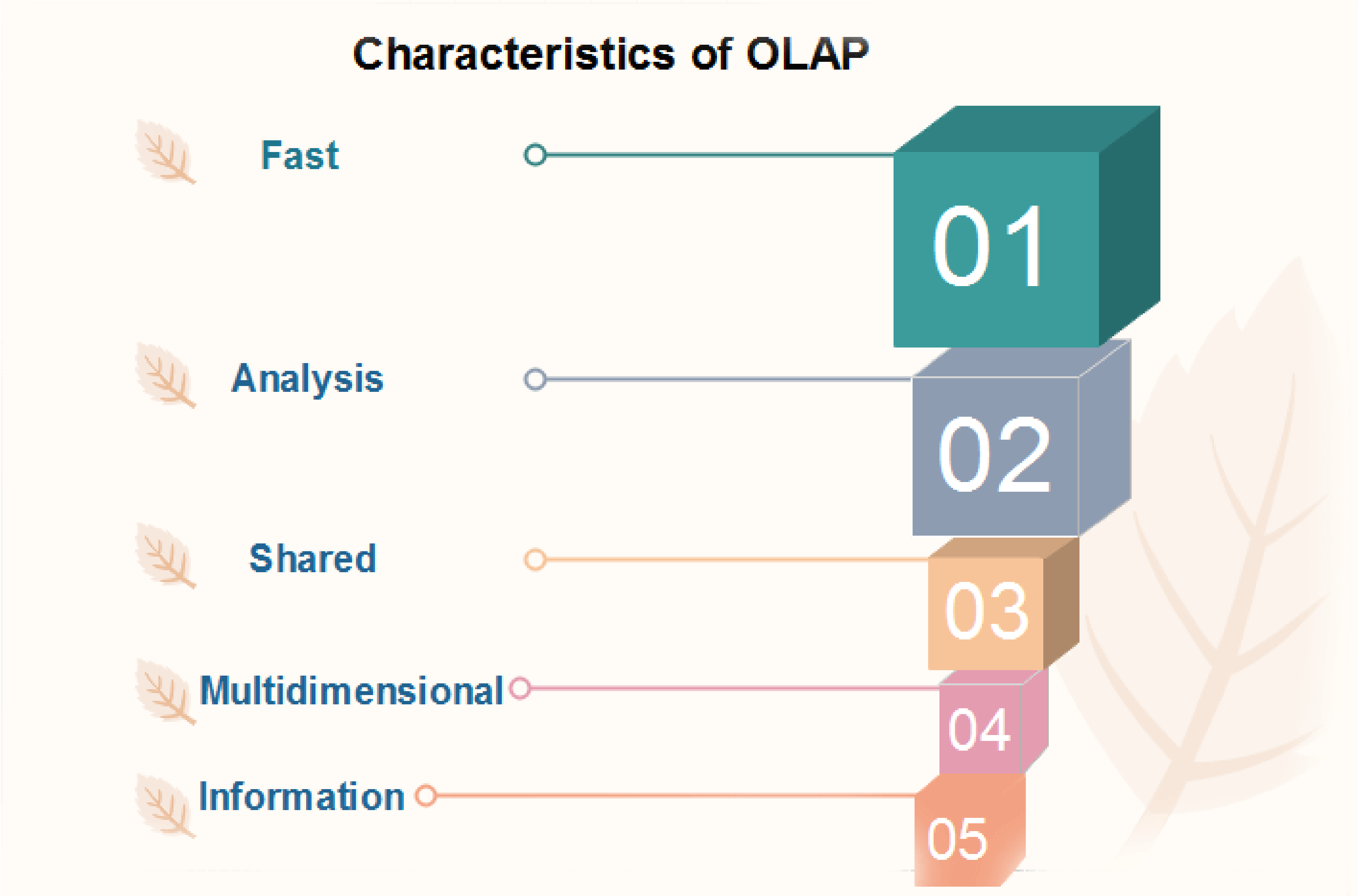
- HOLAP architecture is very complex to understand because it supports both MOLAP and ROLAP.

- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array



Basis	ROLAP	MOLAP	HOLAP
Storage location for summary aggregation	Relational Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.
Processing time	Processing time of ROLAP is very slow.	Processing time of MOLAP is fast.	Processing time of HOLAP is fast.
Storage space requirement	Large storage space requirement in ROLAP as compare to MOLAP and HOLAP.	Medium storage space requirement in MOLAP as compare to ROLAP and HOLAP.	Small storage space requirement in HOLAP as compare to MOLAP and ROLAP.
Storage location for detail data	Relational database is used as storage location for detail data.	Multidimensional database is used as storage location for detail data.	Relational database is used as storage location for detail data.
Query response time	Slow query response time in ROLAP as compare to MOLAP and HOLAP.	Fast query response time in MOLAP as compare to ROLAP and HOLAP.	Medium query response time in HOLAP as compare to MOLAP and ROLAP.

Characteristics of OLAP Systems



- **Fast** − It defines that the system is targeted to produce most responses to users within about five seconds, with the understandable analysis taking no more than one second and very few taking more than 20 seconds.

- **Analysis** − It defines that the system can manage with any business logic and statistical analysis that is appropriate for the application and the user, the keep it easy enough for the target user.

- Although some preprogramming may be needed to enable the user to represent new ad hoc calculations as part of the analysis and to report on the data in any desired method

- **Shared** − It defines that the system implements all the security requirements for confidentiality and, multiple write access is required, concurrent update areas at a suitable level. It is not all applications required users to write data back, but for the increasing number that does, the system must be able to handle several updates in an appropriate, secure manner. This is a major field of weakness in some OLAP products, which tend to consider that all OLAP applications will be read-only, with simple security controls.

- **Multidimensional** − The system should support a multidimensional conceptual view of the data, including complete support for hierarchies and multiple hierarchies. It is not setting up a specific minimum number of dimensions that should be managed as it is too software dependent and most products seem to have enough for their target industry.

- **Information** − Information is all of the data and derived data required, whether it is and however much is relevant for the software. We are measuring the capacity of several products in terms of how much input data can manage, not how many Gigabytes they take to save it.

Multidimensional View and Data cube

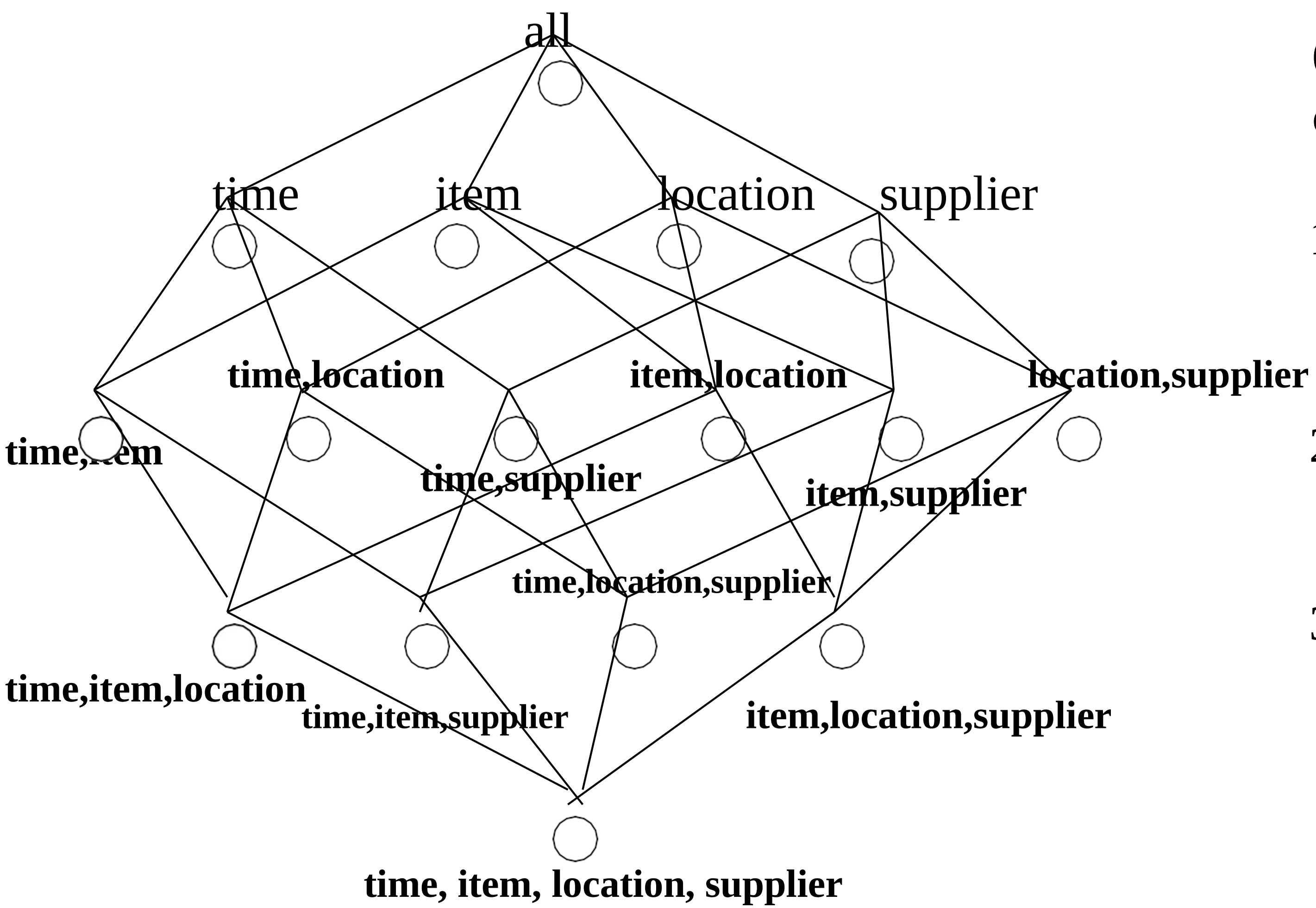
"What is a data cube?" A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records.

From Tables and Spreadsheets to 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** (e.g. 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.

37

Cube: A Lattice of Cuboids



0-D(apex) cuboid

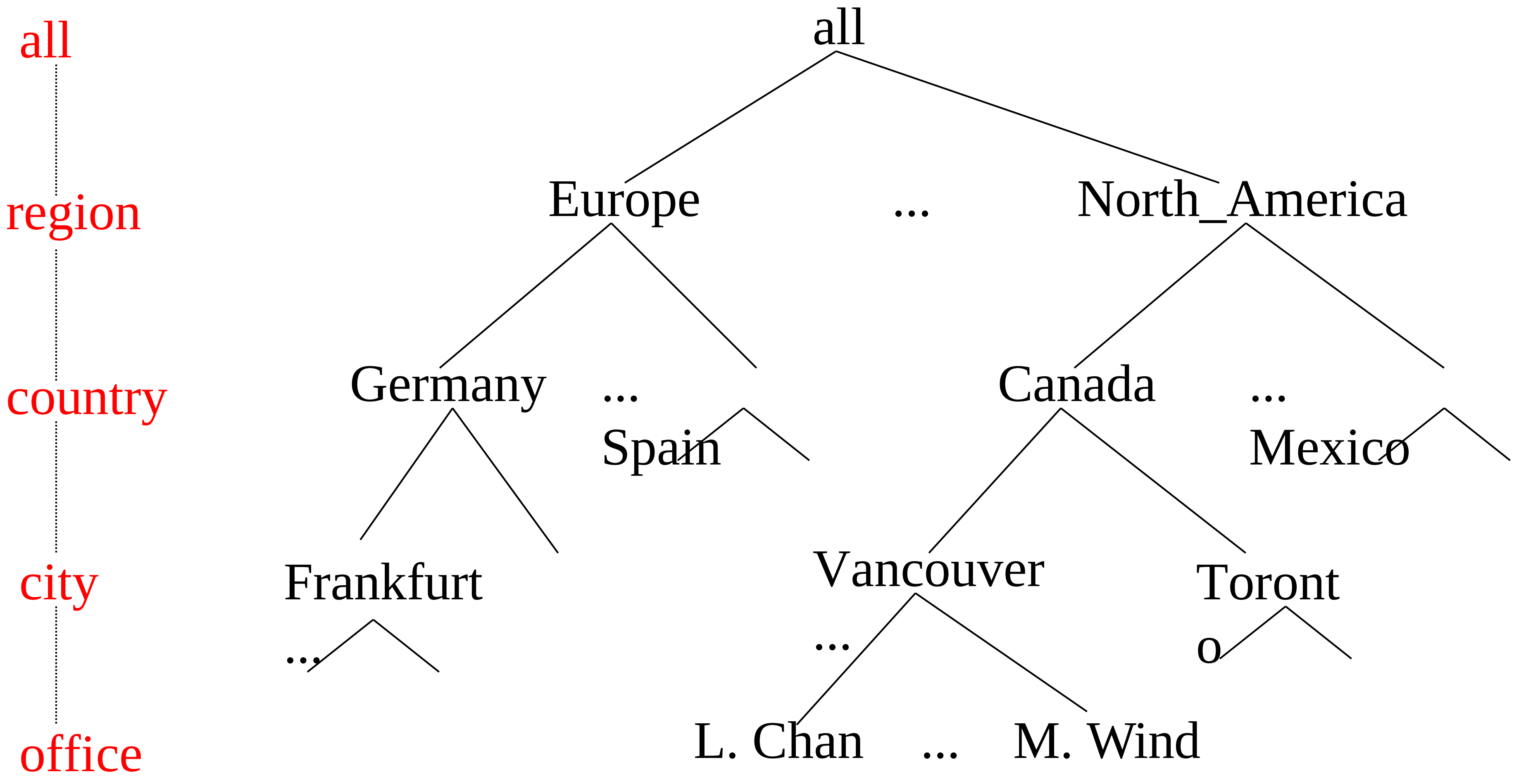
1-D cuboids

2-D cuboids

3-D cuboids

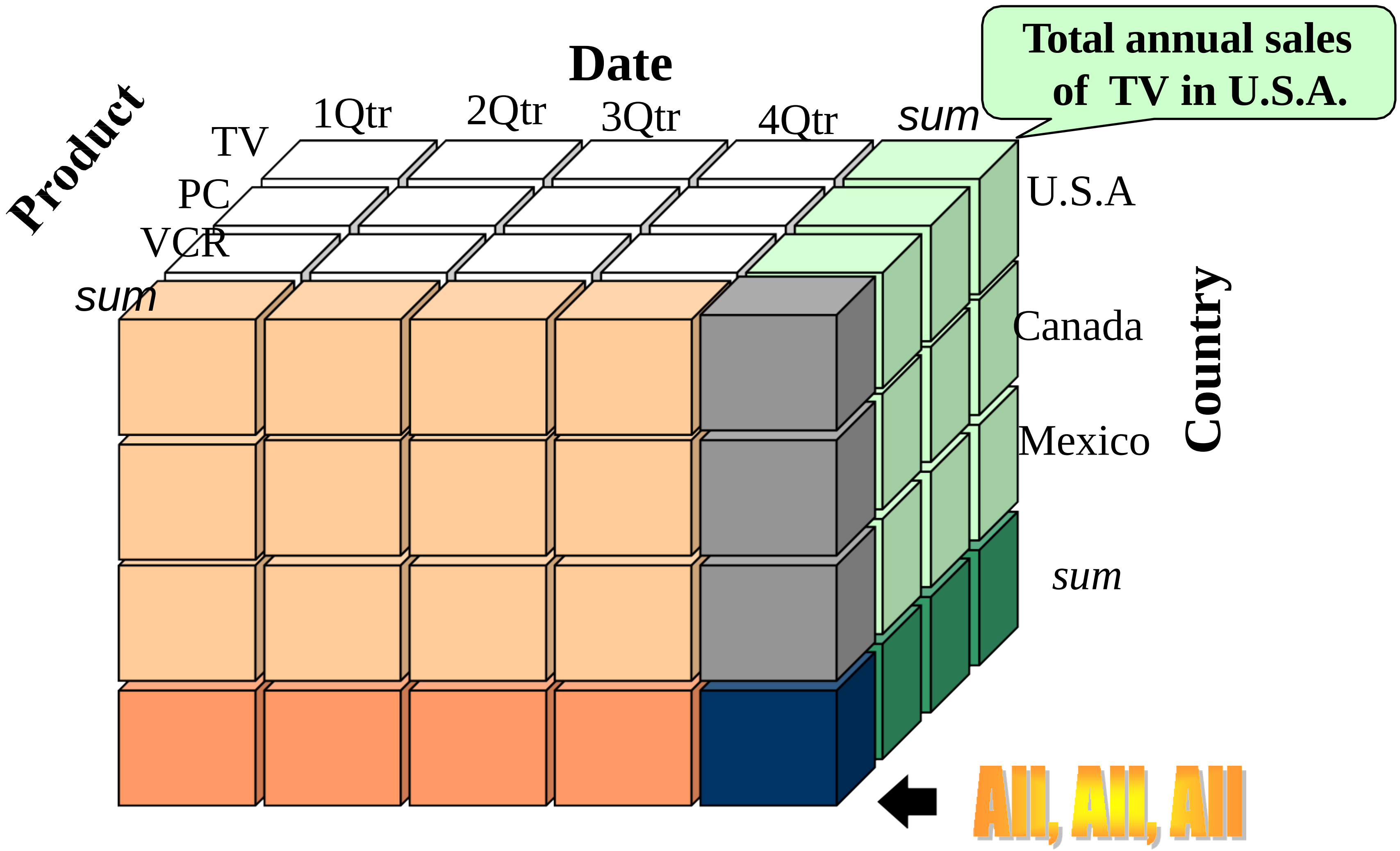
4-D(base) cuboid

A Concept Hierarchy: Dimension (location)



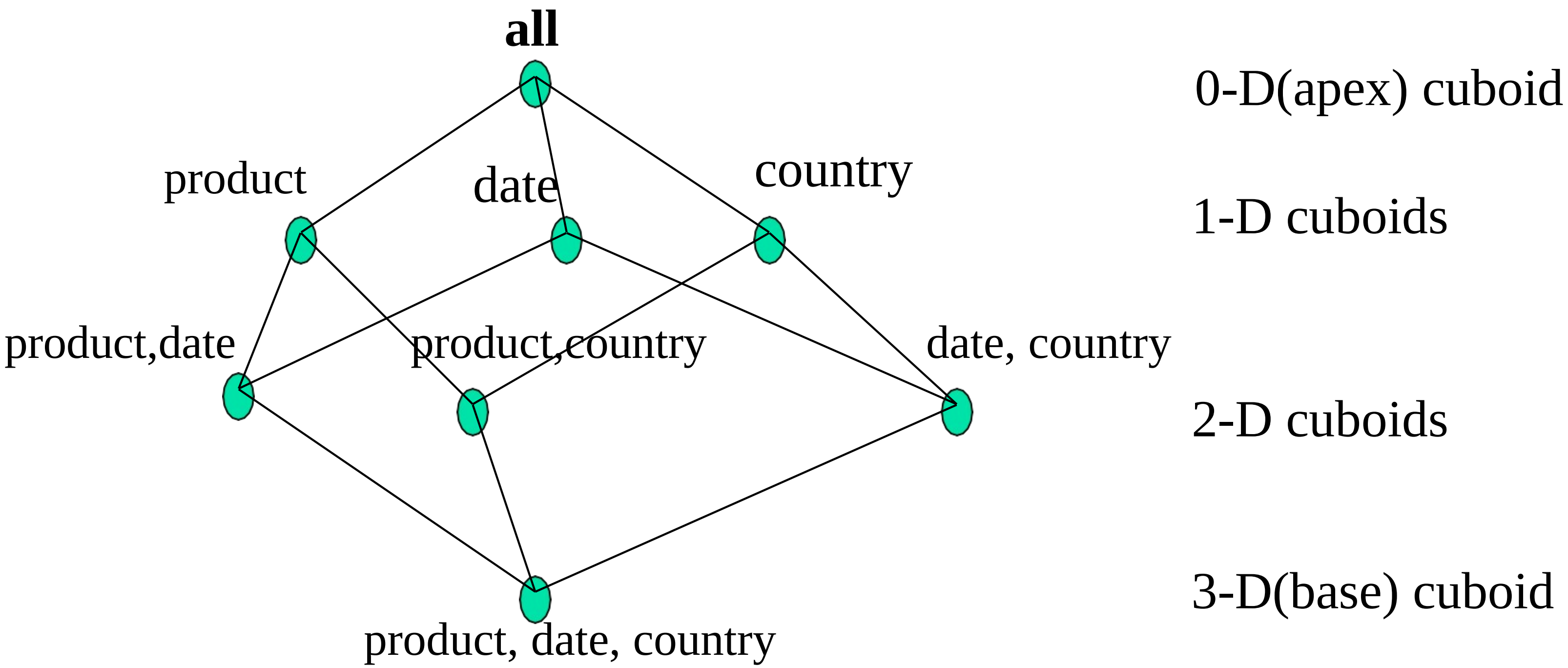
39

A Sample Data Cube



40

Cuboids Corresponding to the Cube



Conceptual Modeling of Data Warehouse

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

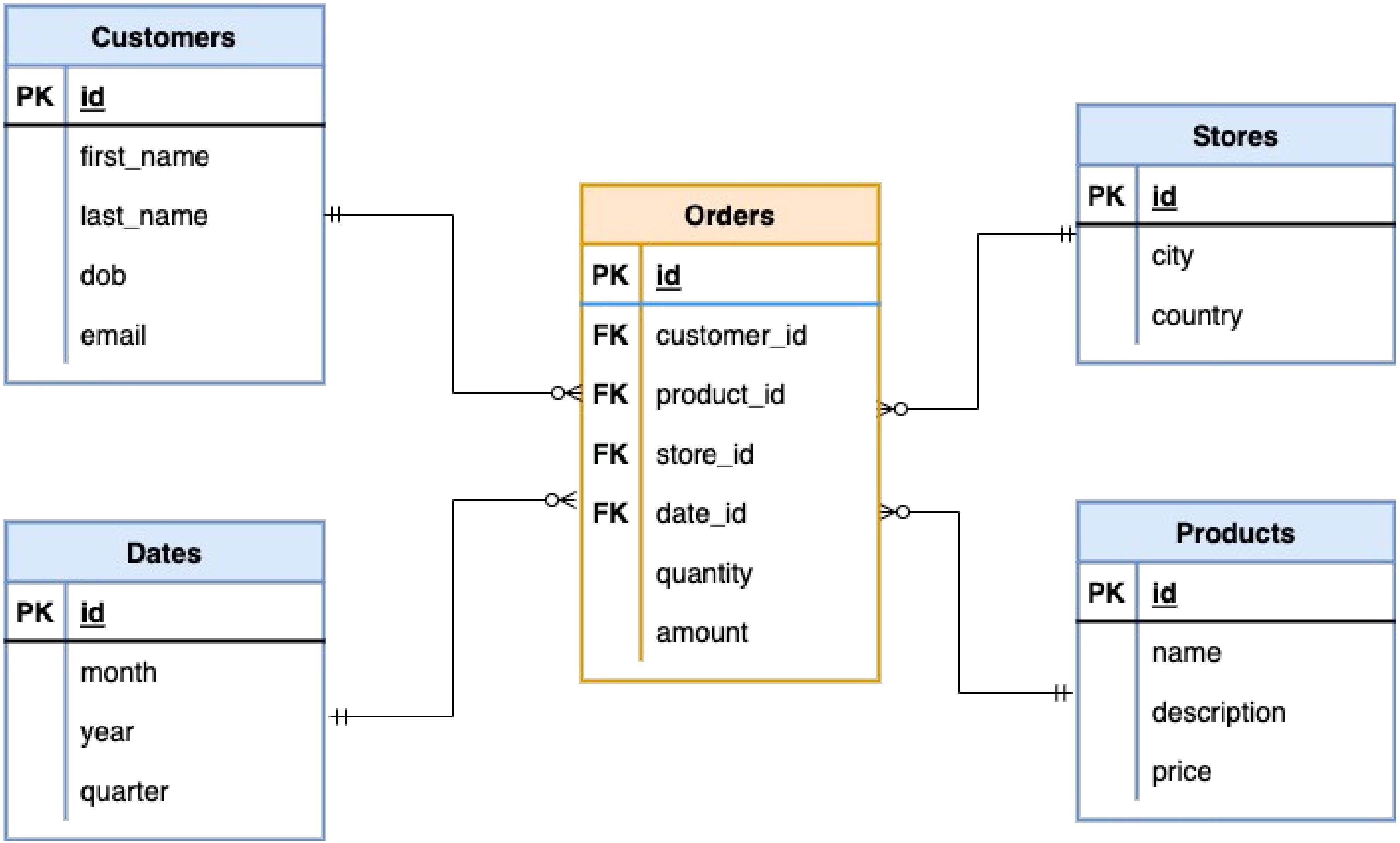
Fact constellations: Multiple fact tables share dimension tables,viewed as a collection of stars, therefore called galaxy schema or fact constellation

Schema and it’s types in Data Warehouse

- 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.
- In schema fact and dimension tables are joined in a logical manner

Parameters	**Fact Table**	**Dimension Table**
Definition	Measurements, metrics or facts about a business process.	Companion table to the fact table contains descriptive attributes to be used as query constraining.
Characteristic	Located at the center of a star or snowflake schema and surrounded by dimensions.	Connected to the fact table and located at the edges of the star or snowflake schema
Design	Defined by their grain or its most atomic level.	Should be wordy, descriptive, complete, and quality assured.

Task	Fact table is a measurable event for which dimension table data is collected and is used for analysis and reporting.	Collection of reference information about a business.
Type of Data	Facts tables could contain information like sales against a set of dimensions like Product and Date.	Evert dimension table contains attributes which describe the details of the dimension. E.g., Product dimensions can contain Product ID, Product Category, etc.
Key	Primary Key in fact table is mapped as foreign keys to Dimensions.	Dimension table has a primary key columns that uniquely identifies each dimension.
Storage	Helps to store report labels and filter domain values in dimension tables.	Load detailed atomic data into dimensional structures.


Customers
PK
id
first_name
last_name
dob
email
Orders
PK
id
FK
customer_id
FK
product_id
FK
store_id
FK
date_id
quantity
amount
Stores
PK
id
city
country
Dates
PK
id
month
year
quarter
Products
PK
id
name
description
price

Star Schema

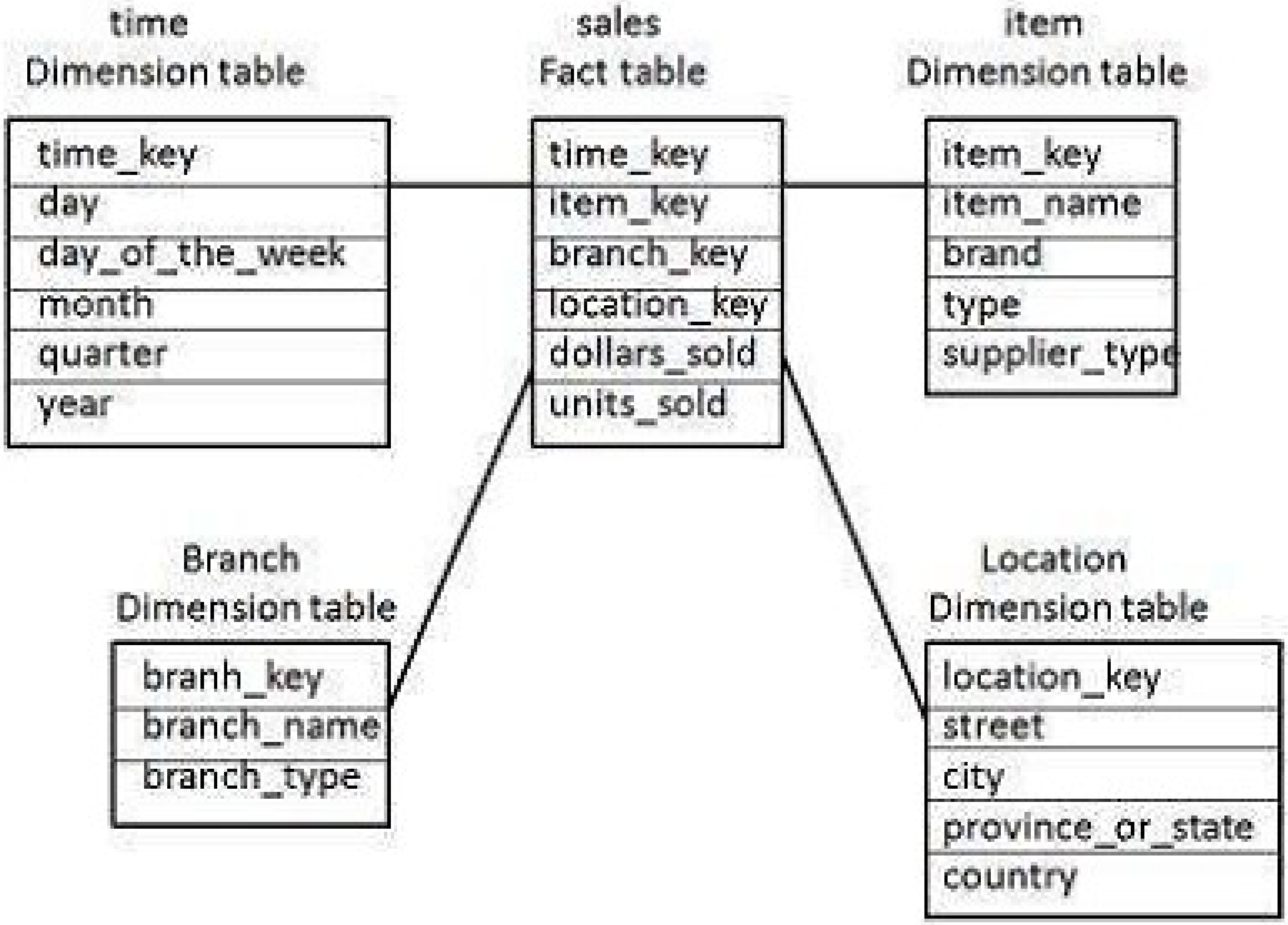
- A Star schema contains a fact table and multiple dimension tables. Each dimension is represented with only one-dimension table and they are not normalized. The Dimension table contains a set of attributes.

Characterstics

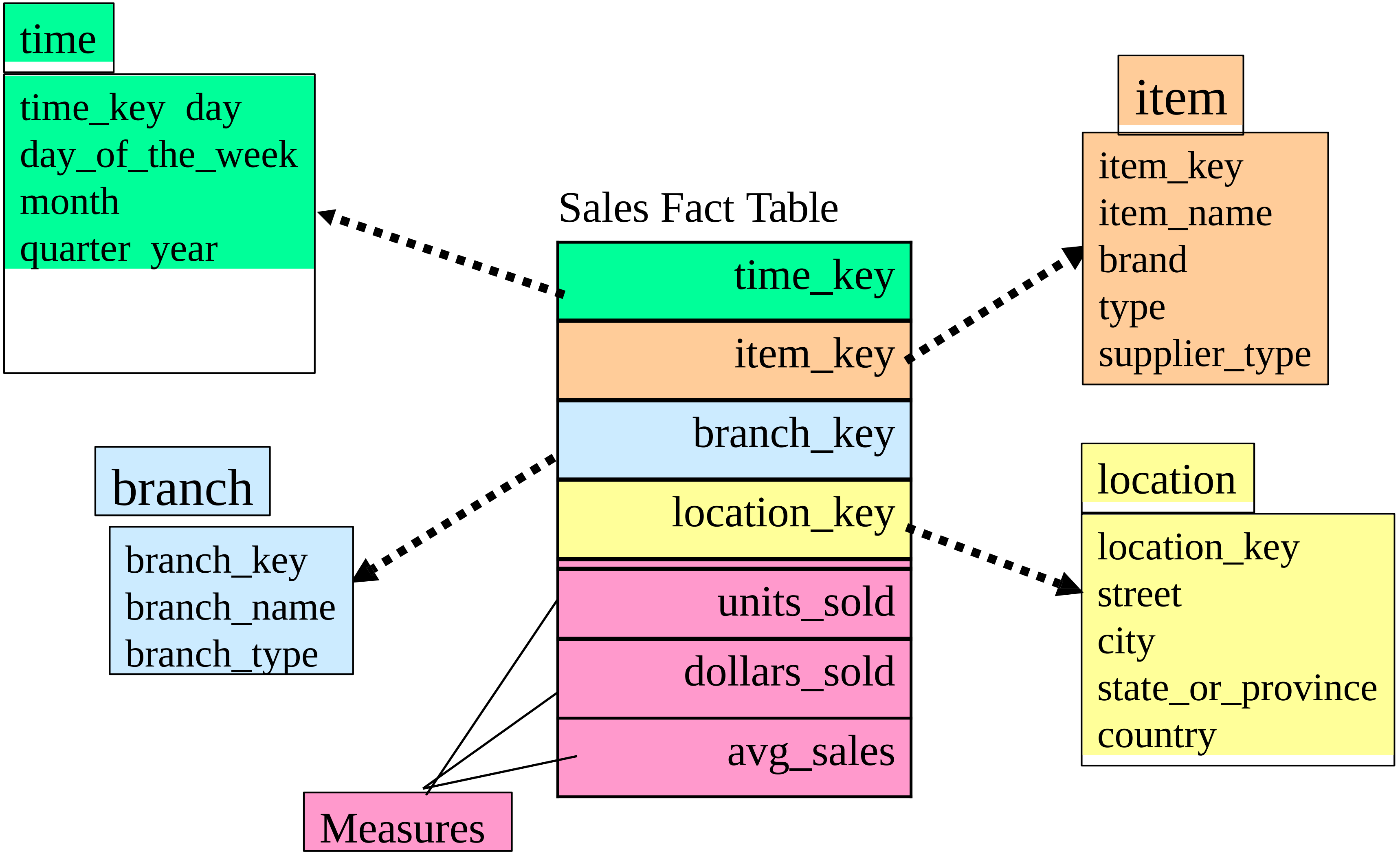
- In a Star schema, there is only one fact table and multiple dimension tables.
- In a Star schema, each dimension is represented by one-dimension table.
- Dimension tables are not normalized in a Star schema.
- Each Dimension table is joined to a key in a fact table.

Notes

Each dimension has only one-dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state, country}. This constraint may cause data redundancy.

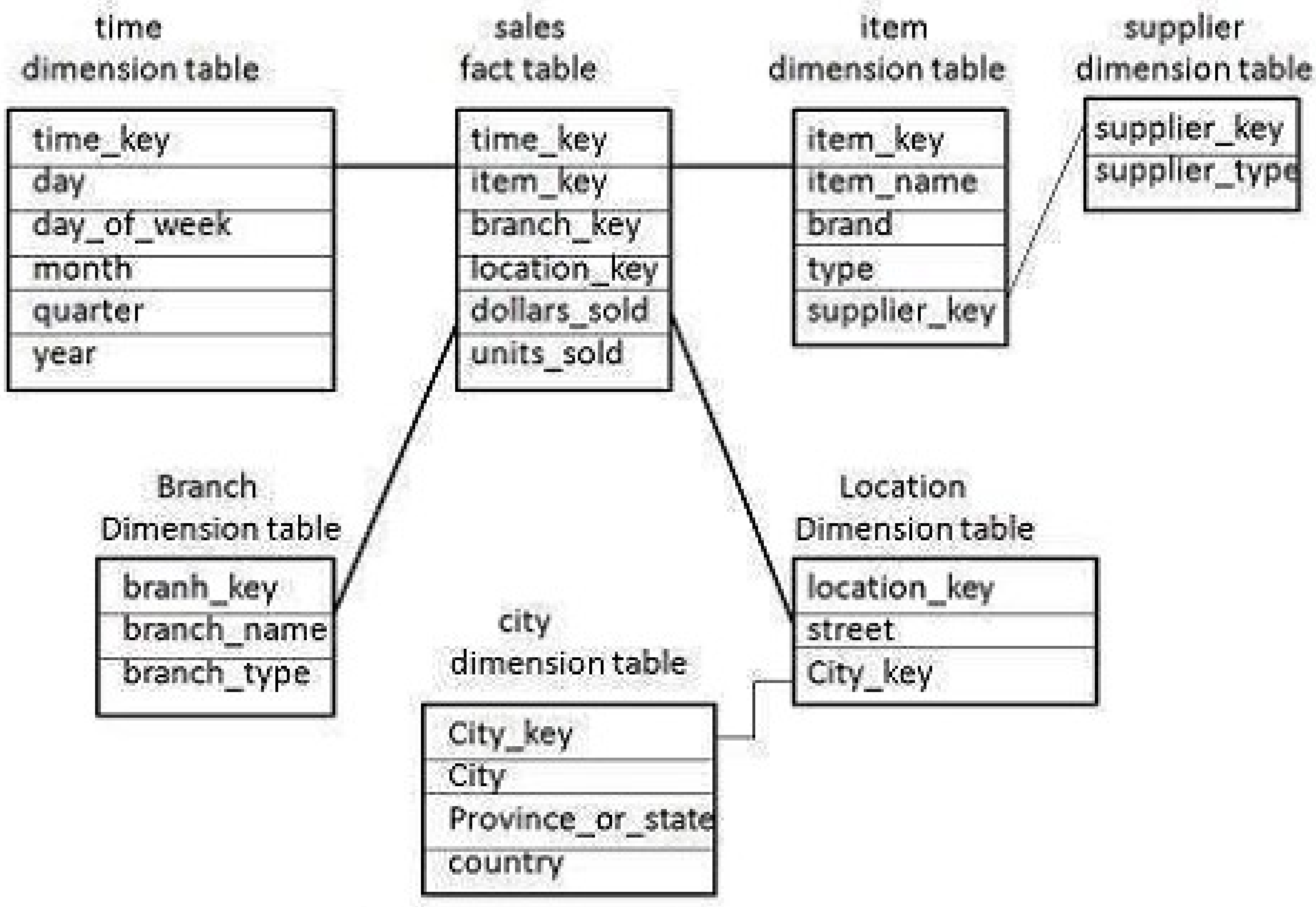

time
Dimension table
time_key
day
day_of_the_week
month
quarter
year
sales
Fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
Dimension table
item_key
item_name
brand
type
supplier_type
Branch
Dimension table
branh_key
branch_name
branch_type
Location
Dimension table
location_key
street
city
province_or_state
country

Example of Star Schema



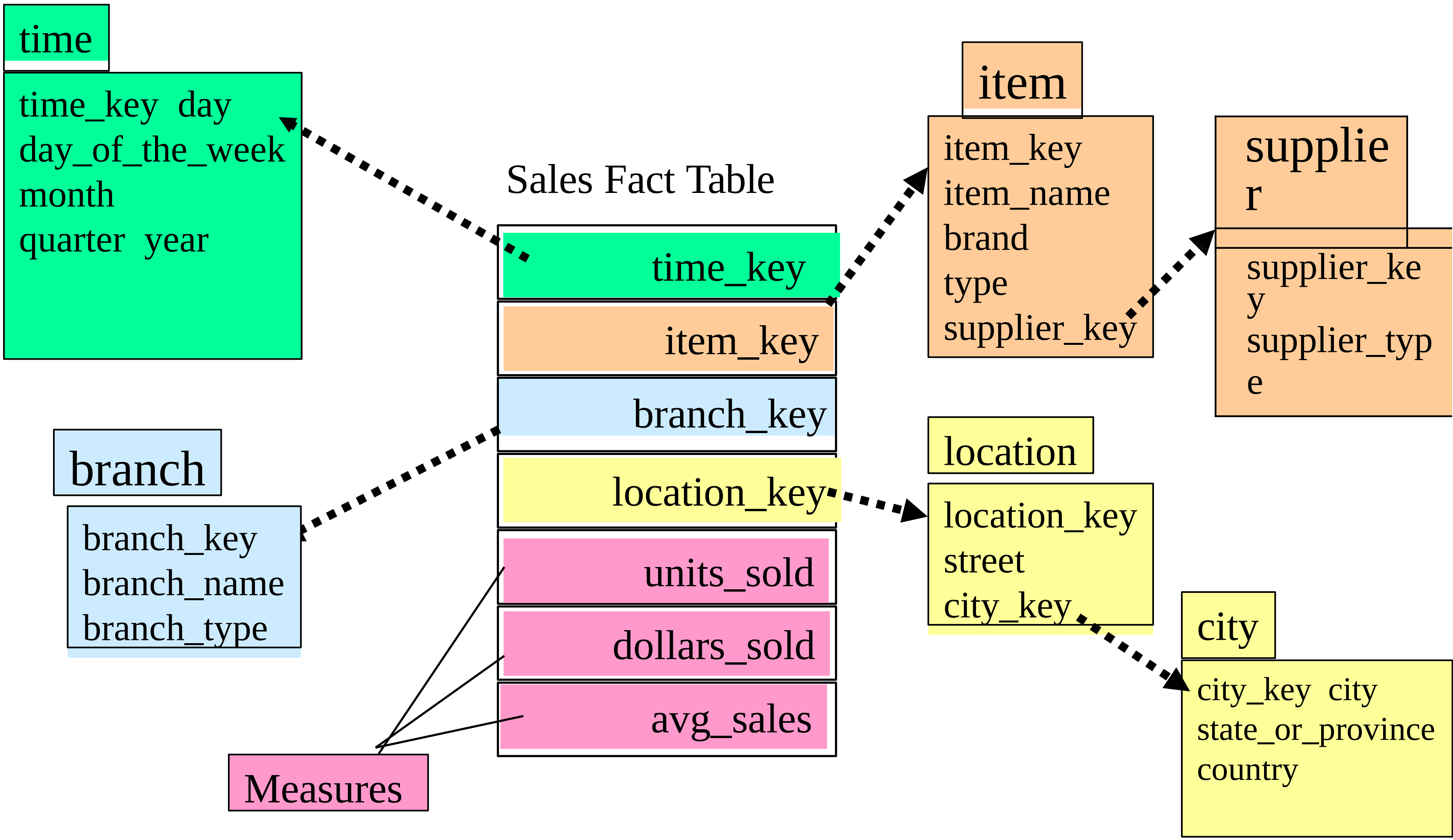
Snow flake Schema

- The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.



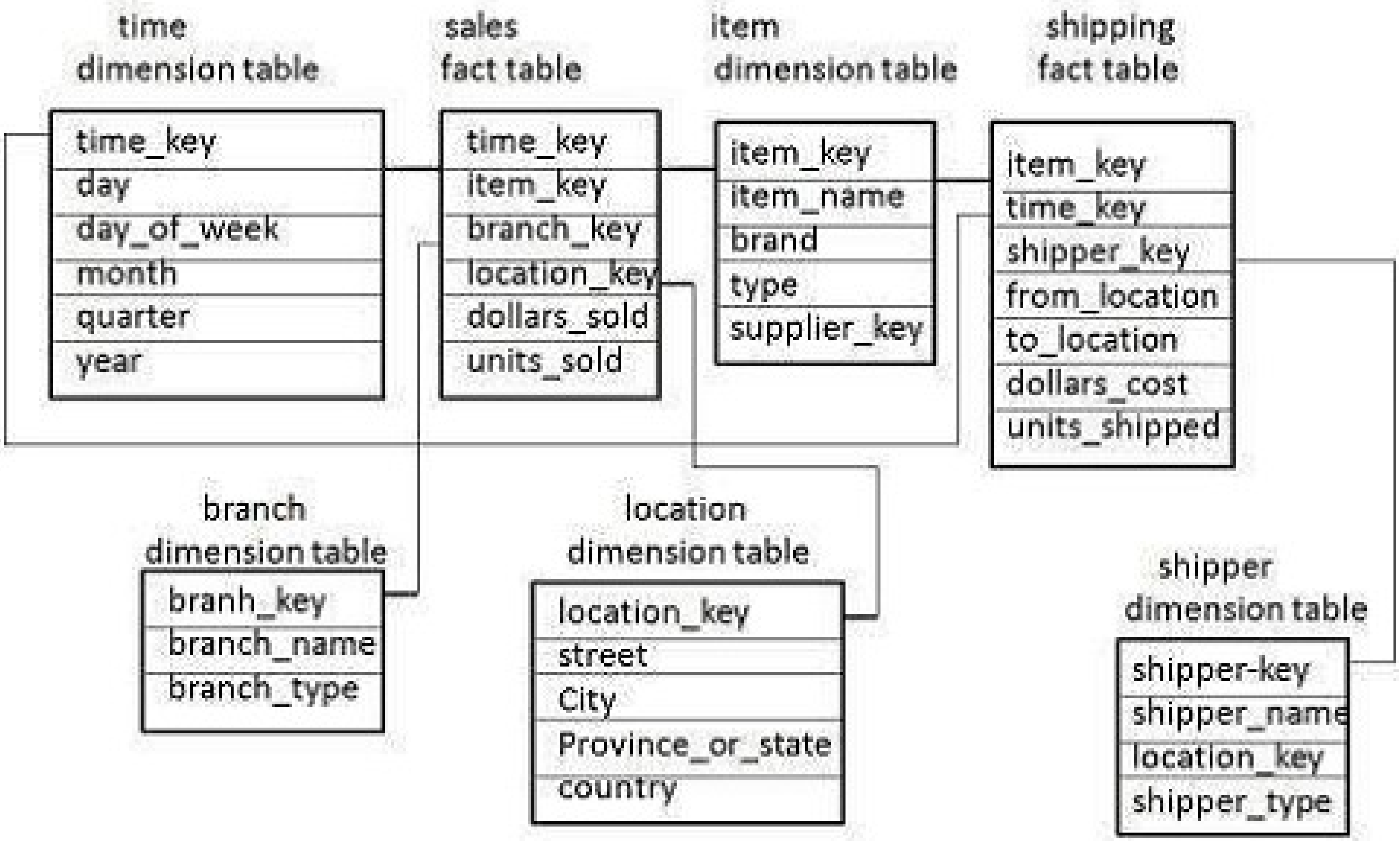
Note − Due to the normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Example of Snowflake Schema



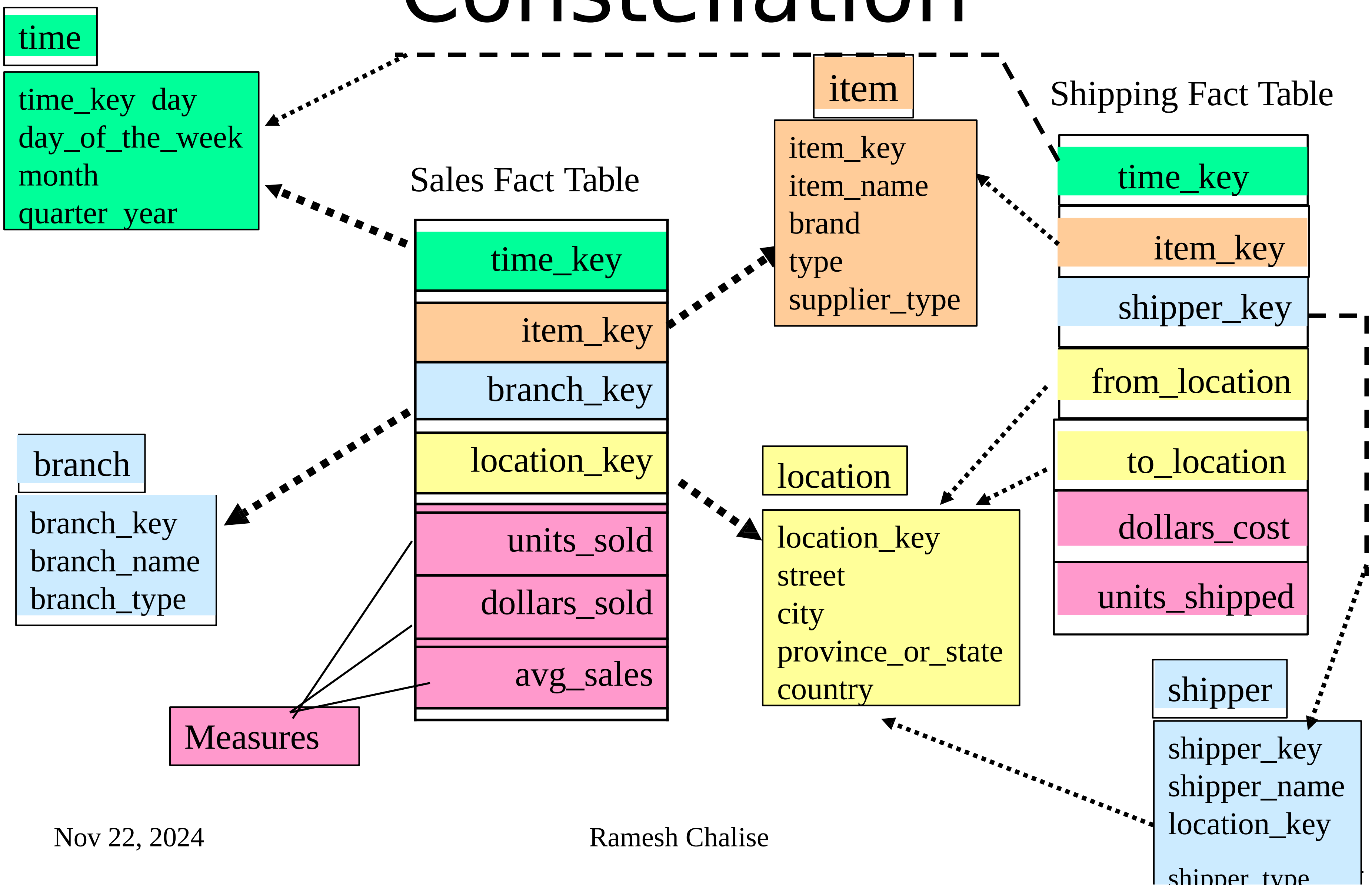
Fact constellation/galaxy schema:

- A fact constellation has multiple fact tables. It is also known as a Galaxy Schema. The following illustration shows two fact tables, namely Sales and Shipping



Time, item, and location dimension tables are shared between the sales and shipping fact table.

Example of Fact Constellation



Data Cube Implementation

i**. Full Materialization** (Pre compute and store all) This means that millions of aggregates will need to be computed and stored. Although this is the best solution as far as query response time is concerned, the solution is impractical since resources required to compute the aggregates and to store them will be prohibitively large for a large data cube.

ii. No Materialization (Pre compute none) This means that the aggregates are computed on the fly using the raw data whenever a query is posed. This approach does not require additional space for storing the cube but the query response time is likely to be very poor for large data cubes.

iii. Partial Materialization (Pre compute and store some) This means that we pre-compute and store the most frequently queried aggregates and compute others as the need arises. We may also be able to derive some of the remaining aggregates using the aggregates that have already been computed. It may therefore be worthwhile also to pre compute some aggregates that are not most frequently queried but help in deriving many other aggregates. It will of course not be possible to derive all the aggregates from the pre-computed aggregates and it will be necessary to access the database to compute the remaining aggregates. The more aggregates we are able to pre-compute the better the query performance.

Guidelines for OLAP Implementation

Following are a number of guidelines for successful implementation of OLAP. The guidelines are, somewhat similar to those presented for data warehouse implementation.

Vision:

- The OLAP team must should develop a clear vision for the OLAP system. This vision including the business objectives should be clearly defined, understood, and shared by the stakeholders.

Senior management support:

- The OLAP project should be fully supported by the senior managers. Since a data warehouse may have been developed already, this should not be difficult.

Selecting an OLAP tool:

- The OLAP team should familiarize themselves with the ROLAP and MOLAP tools available in the market. Since tools are quite different, careful planning may be required in selecting a tool that is appropriate for the enterprise. In some situations, a combination of ROLAP and MOLAP may be most effective.

Corporate strategy:

- The OLAP strategy should fit in with the enterprise strategy and business objectives. A good fit will result in the OLAP tools being used more widely.

Focus on the users:

- The OLAP project should be focused on the users. With guidance form technical professional Users should, decide what tasks will be done first and what will be done later. Attempts should be made to provide each user with a tool suitable for that person's skill level and information needs. A good GUI user interface should be provided to non-technical users. The project can only be successful with the full support of the users.

Joint management:

- The OLAP project must be managed by both the IT and business professionals. Many other people should be involved in supplying ideas. An appropriate committee structure may be necessary to channel these ideas.

Review and adapt:

- Organizations evolve and so must the OLAP systems. Regular reviews of the project may be required to ensure that the project is meeting the current needs of the enterprise