1. Data Types and Attributes
2. Data Pre-processing
3. OLAP & Multidimensional Data Analysis
4. Various Similarity Measures

Data Quality

- Real world database is highly unprotected from noise, missing and inconsistent data due to their typically huge size and their possible origin from multiple, heterogeneous sources.
- Low quality data will lead to low quality mining results. So; Data pre-processing is required to handle these above-mentioned facts.
- **Preprocess to Measure for data quality: A multidimensional view**
  - **Accuracy**: correct or wrong, accurate or not
  - **Completeness**: not recorded, unavailable, …
  - **Consistency**: some modified but some not, dangling, …
  - **Timeliness**: timely update?
  - **Believability**: how trustable the data are correct?
  - **Interpretability**: how easily the data can be understood?

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

Data goes through a series of **steps during pre-processing:**

- **Data Cleaning**: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- **Data Integration:** Data with different representations are put together and conflicts within the data are resolved.
- **Data Transformation**: Data is normalized, aggregated and generalized.
- **Data Reduction**: This step aims to present a reduced representation of the data in a data warehouse.
- **Data Discretization**: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Data Pre-processing is required because:

Real world data are generally:

*** Incomplete:** Missing - attribute values, certain attributes of importance, or having only aggregate data
e.g. occupation = “”

*** Noisy:** Containing errors or outliers
e.g., Salary=“-10”

*** Inconsistent/ Varying:** Containing discrepancies in codes or names
e.g., Age=“42” Birthday=“03/07/1997”
e.g., Was rating “1,2,3”, now rating “A, B, C”
e.g., discrepancy between duplicate records

Why Is Data Pre-processing Important?

- No quality data, no quality mining results! (garbage in garbage out!)
  - Quality decisions must be based on quality data
    - **e.g.**, duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application

1. Data Types and Attributes

Attribute type can be compound, list or whatever, then the **data type** is the type of data contained in that data structure. like compound is a parent attribute, and then a bunch of children - but the children can be floats, or strings, or whatever

What is an Attribute?

- An attribute is a property or characteristic of an object. Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describes an object. Object is also known as record, point, case, sample, entity, or instance.
- Attribute values are numbers or symbols assigned to an attribute
- Same attribute can be mapped to different attribute values. **Example**: height can be measured in feet or meters.
- 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.

Data Types

Data Type	Used for	Example
String	Alphanumeric characters	hello world, Alice, Bob123
Integer	Whole numbers	7, 12, 999
Float (floating point)	Number with a decimal point	3.15, 9.06, 00.13
Character	Encoding text numerically	97 (in ASCII, 97 is a lower case 'a')
Boolean	Representing logical values	TRUE, FALSE

Attribute Type

Approach 1 :

- **Quantitative** data deals with numbers and things you can measure objectively: dimensions such as height, width, and length. Temperature and humidity. Prices. Area and volume.
- **Qualitative** data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively—such as smells, tastes, textures, attractiveness, and color.

Approach 2:

- **Nominal:** The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠) e.g. **Gender**: Male, Female, Other**; Hair Color**: Brown, Black, Blonde, Red, Other.
- **Ordinal:** The values of an ordinal attribute provide enough information to order objects. (<, >) e.g. **Hardness** of minerals: Hard, Moderate, Soft; **Rank**: Distinction, 1^{st}, 2^{nd}, 3^{rd}, … ; **Grades**: A, B, C, …
- **Binary:** variables with only two options. E.g. **Result**: Pass or Fail; **Rain**: Yes or No
- **Interval:** has values of equal intervals that mean something. For example, a thermometer might have intervals of ten degrees., i.e., a unit of measurement exists. (+, -) e.g. Calendar dates, Temperature in Celsius or Fahrenheit, Time on clock
- **Ratio:** exactly the same as the interval scale except that the zero on the scale means: does not exist. (*, /) e.g. **Weight** of zero doesn't exist; **Age** of zero doesn't exist; **Height** of zero doesn't exist. *Note: temperature is not a ratio scale, because zero exists (i.e. zero on the Celsius scale is just the freezing point;*

Approach 3:

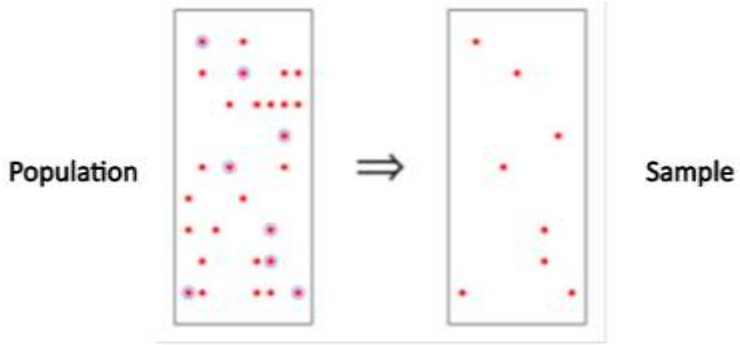
- **Discrete Data** is distinct, can only take certain values.
  Example: Number of students in a class (you can't have half a student); **Gender**: Male or Female
- **Continuous Data** is data that can take any value (within a range)
  Examples:
  - **A person's height**: could be any value (within the range of human heights), not just certain fixed heights,
  - **Time in a race**: you could even measure it to fractions of a second, e.g. 5.5 min, 15 min, 200 min
  - T**emperatures:** It can be 23 degrees, 23.1 degrees, 23.100004 degrees.

Approach 4:

- **Character**: values are represented in forms of character or set of characters (string).
- **Number**: values are represented in forms of number. Number may be in form of whole number, decimal number.

Dataset Types

- **Population** - the collection of all individuals or items under consideration in a statistical study
- **Sample** - that part of the population from which information is collected



- **Parameter** – statistical description of the population
- **Variable** – characteristic that varies from one item to another e.g. Quantitative (numerical) Qualitative (categorical)
- **Data**: Observing the values of the variables yield data
- **Observation** – individual piece of data
- **Data matrix** – collection of observations for variable Data matrix k variables measured in sample with the size of n

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. Datasets can hold multiple tables and you can define relationships between those tables.

a. Record - Data that consists of a collection of records, each of which consists of a fixed set of attributes

i. Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

ii. Document Data: text documents: term-frequency vector. Each document becomes a `term' vector where each term is a component (attribute) of the vector and the value of each component is the number of times the corresponding term occurs in the document

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Fig. Data Matrix

	team	coach	play	ball	score	game	win	lost	timeout	Season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Fig. Document Data

iii. Transaction Data - A special type of record data, where each record (transaction) involves a set of items. Transactional data can be financial, logistical or work-related, involving everything from a purchase order to shipping status to employee hours worked to insurance costs and claims.

- For example, consider a grocery store. *The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items*

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Fig. Transaction Data

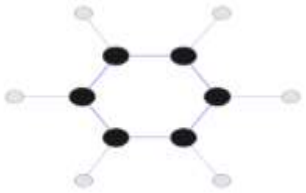


Fig. Graph: World Wide Web, Molecular Structures

b. Graph and Network: Contain nodes and connecting vertices. E.g. WWW, Social or information networks, Molecular Networks

c. Ordered: Has Sequences of transactions

i. Spatial Data: maps, Spatial data, also known as geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system.

ii. Temporal Data: time series, A temporal data denotes the evolution of an object characteristic over a period of time. Eg d=f(t).

iii. Sequential Data: transaction sequence, Data arranged in sequence. E.g. Video data – sequence of images

Characteristics of Structured Data

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.

a. Dimensionality - A Data Dimension is a set of data attributes affecting to something of interest to a business. Dimensions are things like "customers", "products", "stores" and "time".

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly light in the space that it occupies.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

*Purpose:

- Avoid curse/ weak points of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

*Techniques

- Principle Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

Dimensionality Reduction: Goal: to reduce dimensionality of data. process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

i. Principle Component Analysis (PCA)

- Goal: to find a projection that captures the largest amount of variation in data.
- Find the eigenvectors of the covariance matrix. The eigenvectors define the new space.
- Construct a neighborhoods graph

- For each pair of points in the graph, compute the shortest path distances: geodesic distances

ii. Feature Subset Selection

- **Redundant features**: Duplicate much or all of the information contained in one or more other attributes. Example: purchase price of a product and the amount of sales tax paid.
- **Irrelevant features**: Contain no information that is useful for the data mining task at hand. Example: students' ID is often irrelevant to the task of predicting students' GPA

Techniques:

a) Brute-force approach:- Try all possible feature subsets as input to data mining algorithm
b) Embedded approaches: - Feature selection occurs naturally as part of the data mining algorithm
c) Filter approaches:- Features are selected before data mining algorithm is run
d) Wrapper approaches: - Use the data mining algorithm as a black box to find best subset of attributes.

Feature Creation

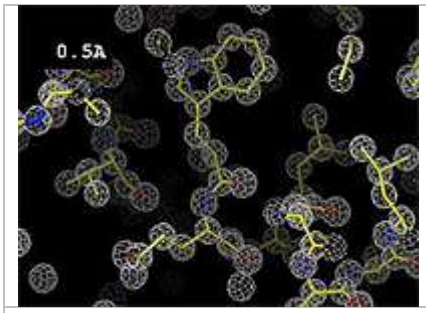
- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes. Three general methodologies:
  - Feature Extraction: domain-specific
  - Mapping Data to New Space
  - Feature Construction: combining features

b. Sparsity and Density: is used to improve database and data processing performance. Describe the number of cells in a table that are empty (sparsity) and that contain information (density), though sparse cells are not always technically empty—they often contain a "0" digit. Tables and databases are the sum total of their sparse and dense cells.

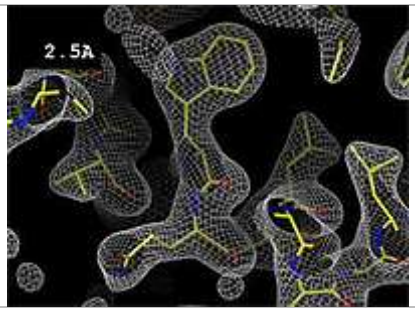
- Sparsity and density are terms used to describe the percentage of cells in a database table that are not populated and populated, respectively. The sum of the sparsity and density should equal 100.
- Many of the cell combinations might not make sense or the data for them might be missing.
- In the relational world storage of such data is not a problem: we only keep whatever there is. If we want to keep closer to our multidimensional view of the world, we face a dilemma: either store empty space or create an index to keep track of the nonempty cells or search for an alternative solution
- A table that is 10% dense has 10% of its cells populated with non-zero values. It is therefore 90% sparse – meaning that 90% of its cells are either not filled with data or are zeros.

c. Resolution

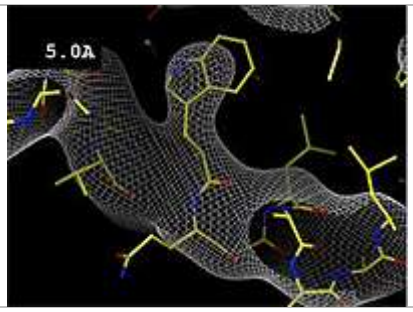
- Scaling of data in different label and classes. Patterns depend on the scale.
- Resolution, in structure determinations, is the distance corresponding to the smallest observable feature: if two objects are closer than this distance, they appear as one combined blob rather than two separate objects. Resolution is the smallest distance between crystal lattice planes that is resolved in the diffraction pattern.



Resolution at **0.5 (Excellent)**
backbone and most sidechains very clear



Resolution at **2.5A (good)**
backbone and many sidechains clear



Resolution at **5.0A (poor)**
backbone mostly clear; sidechains not clear

2. Data Pre-processing

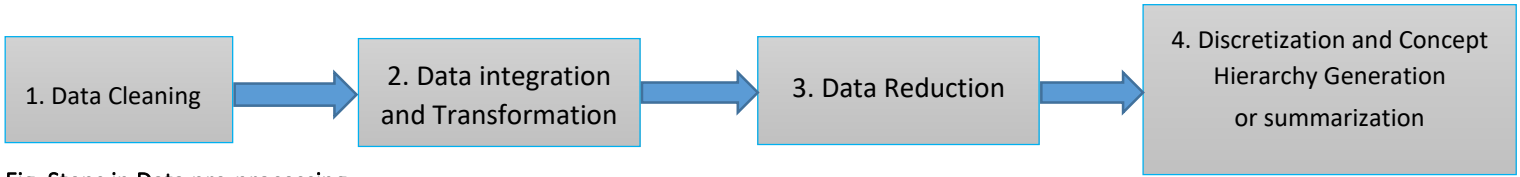


Fig. Steps in Data pre-processing:

1. Data cleaning: is a process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- Fill in missing values, smooth noisy data, identify or remove the outliers, and resolve inconsistencies.
- Data cleaning is required because source systems contain "dirty data" that must be cleaned.
- Mostly concern with

i. Fill-in missing values
ii. Identify outliers and smooth out noisy data
iii. Correct inconsistent data
iv. Eliminate duplicate data

a. Missing Data -Data is not always available because many tuples may not have recorded values for several attributes such as age, income. 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.
- Not register history or changes of the data.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing. Not effective when the percentage of missing values per attribute varies considerably.
- Fill-in missing values manually: Tedious and infeasible task.
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

b. Noisy Data – e.g. Salary="-10" Noisy data is a form of error because of random error in a measured variable. Incorrect attribute values may be due to:

- Faulty data collection instruments .
- Data entry problem .
- Data transmission problem .
- Technology limitation .
- Inconsistency in naming convention
- Duplicate records, incomplete data(e.g. Occupation = " " (missing data)), inconsistent data (e.g. Was rating "1, 2, 3", now "A, B, C")

How to Handle Noisy Data

- Clustering: Detect and remove outliers

- Regression: Smooth by fitting the data into regression function.

i. Linear regression involves finding the "best" line to fit two attributes, so that one attribute can be used to predict the other.
ii. 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.

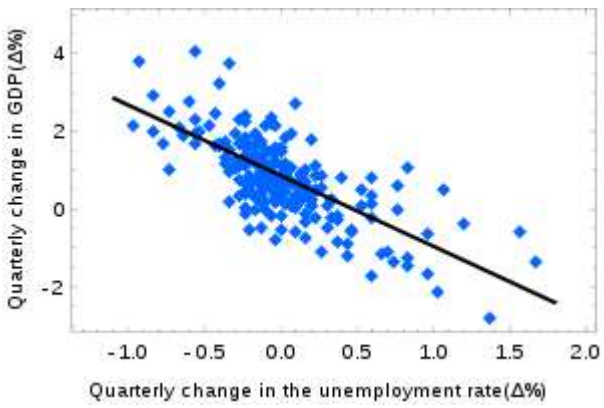


Fig. Linear Regression

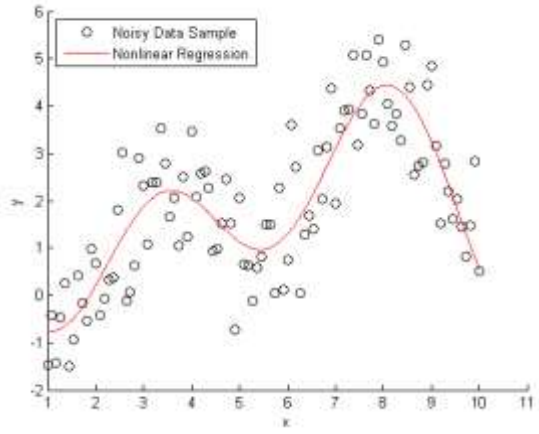


Fig. Non-linear Regression

- Binning Method: first sort data and partition into (equal-frequency) bins, then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Example: Sorted data for price: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (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:
  - Bin 1 (36/4 = 9) : 9, 9, 9, 9

- Bin 2 (91/4 = 22.75) : 23, 23, 23, 23
- Bin 3 (117/4 = 29.25): 29, 29, 29, 29

* Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15 boundary is (4,15) where 8, 9 are closure with 4 than 15
- Bin 2: 21, 21, 25, 25 boundary is (21,25) where 21 is closure with 21 but 24 are closure with 25
- Bin 3: 26, 26, 26, 34 boundary is (26,34) where 28, 29 are closure with 26 than 34

- Combined computer and human inspection: detect suspicious values and check by human (e.g., deal with possible outliers)

c. Outliers - Outliers are a set of data points that are considerably dissimilar or inconsistent with the remaining data that lie outside normal experience. In most of the cases they are inference of noise while in some cases they may actually carry valuable information. Outliers can occur because of:

- Transient malfunction of data measurement.
- Error in data transmission or transcription or data entry
- Changes in system behaviour.
- The data are inappropriately scaled;
- Fault in assumed theory

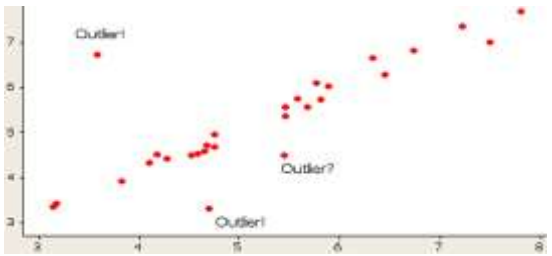


Fig. Outliers

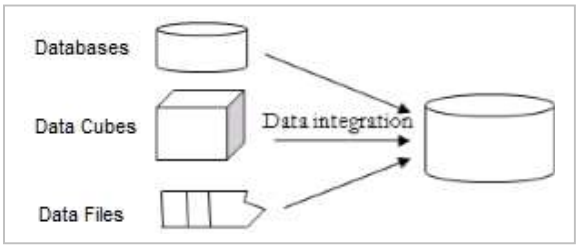
How to Handle Outliers?

There are three fundamental approaches to the problem of outlier's detection

a. Type 1 – Unsupervised Learning: Determine the outliers with no prior knowledge of data.
b. Type 2 – Supervised Learning: Model with normality and abnormality.
c. Type 3 – Semi-Supervised Learning: Model with normality.

2. Data integration and Transformation:

- Data integration: Combines data from multiple sources into a coherent data store e.g. data warehouse. Sources may include multiple databases, data cubes or data files.



Issues in data integration:

- Schema integration: e.g ***A.cust-id=B.cust-#.***
  - Integrate metadata from different sources.
  - Entity identification problem: *identify real world entities from multiple data sources,* **e.g., Bill Clinton = William Clinton**
- Detecting and resolving data value conflicts:
  - For the same real-world entity, attribute values from different sources are different.
  - Possible reasons: different representations, different scales **e.g., metric vs. British units**
- Redundant data occur often when integration of multiple databases:
  - Object identification: The same attribute or object may have different names in different databases.
  - Derivable data: One attribute may be a "derived" attribute in another table, **e.g., annual revenue**

- Data Transformation: Transformation process deals with rectifying any inconsistency (if any).
  - A function that maps the entire set of values of a given attribute to a new set of replacement values i.e. each old value can be identified with one of the new values
  - One of the most common transformation issues is 'Attribute Naming Inconsistency'. It is common for the given data element to be referred to by different data names in different databases.
    E.g. Employee Name may be ***EMP_NAME*** *in one database,* ***ENAME*** *in the other.*
  - Thus, one set of Data Names are picked and used consistently in the data warehouse.
  - Once all the data elements have right names, they must be converted to common formats.

$$-2, 32, 100, 59, 48 \xrightarrow{\text{Data Transformation}} 0.02, 0.32, 1.00, 0.59, 0.48$$

Data transformation can involve the following:

- Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.
- Aggregation, sometimes "LESS IS MORE" i.e. summarization of data, where summary or aggregation operations are applied to the data e.g. data cube construction. Combining of two or more objects into a single object. *For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.*
- Generalization of the data, Hierarchy climbing of 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. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.

- Normalization, Mainly, when data not following normality assumptions we transform it to get normality. Scaled to fall within a small and specified range, 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.
  - Min-Max Normalization: transforms the data set from one range to another. Transform the data from measured units to a new interval from $minA$ to $maxA$ for feature A:

    $$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$ where v is the current value of feature A.

    Example: Suppose that the minimum and maximum values for the feature income are Rs. 12,000 and Rs. 98,000, respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of Rs. 73,600 for income is transformed to:

    $$\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0.0) + 0 = 0.716$$

  - Z-Score (Zero-Score) Normalization: Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, v, of A is normalized to v' by computing:

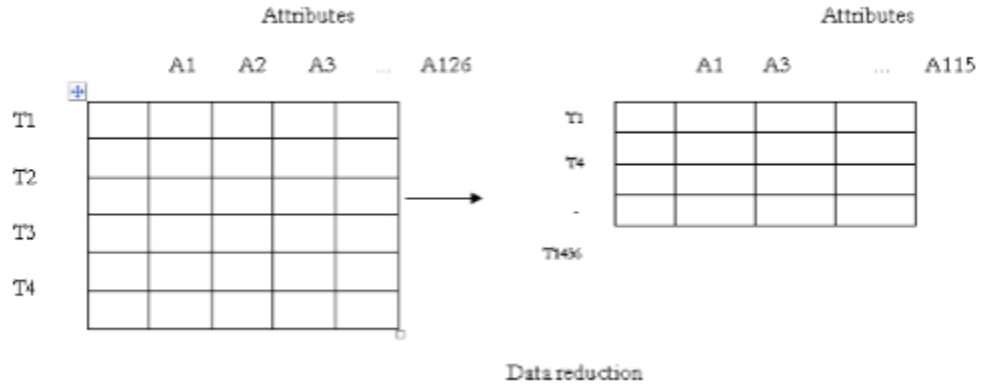
    $$v' = \frac{v - \bar{F}}{\sigma_F}$$ where $\bar{F}$ and σ_F are the mean and standard deviation of feature F, respectively.

    Example: Suppose that the mean and standard deviation of the values for the feature income are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,600 for income is transformed to (73,600 − 54,000)/ 16,000 = 1.225
  - Normalization by decimal scaling: Transform the data by moving the decimal points of values of feature F. The number of decimal points moved depends on the maximum absolute value of F. A value v of F is normalized to v' by computing:
    $V' = V/ 10^j$ where j is the smallest integer such that max (|V'|) <1
    Example: Suppose that the recorded values of F range from − 986 to 917. The maximum absolute value of F is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., j = 3) so that −986 normalizes to −0.986 and 917 normalizes to 0.917.
- Features/ Attributes construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

3. Data Reduction:

A database or data warehouse may store terabytes of data Complex data analysis/mining may take a very long time to run on the complete data set. Data reduction obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

- Process of minimizing the amount of data that needs to be stored in a data storage environment.
- Data reduction can increase storage efficiency and reduce costs.
- Obtains reduced representation in volume but produces the same or similar analytical results.
- Need for data reduction:
  - Reducing the number of attributes
  - Reducing the number of attribute values
  - Reducing the number of tuples



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:

1. Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube. store multidimensional aggregated information.

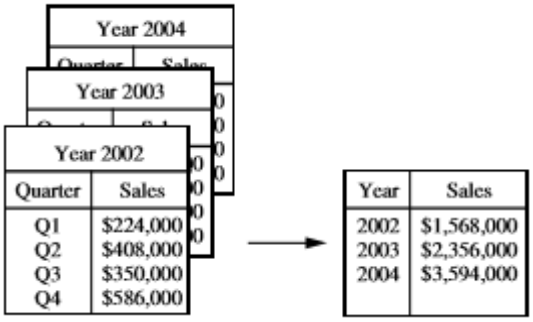


Figure (a) *Sales data for a given branch of All Electronics for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.*

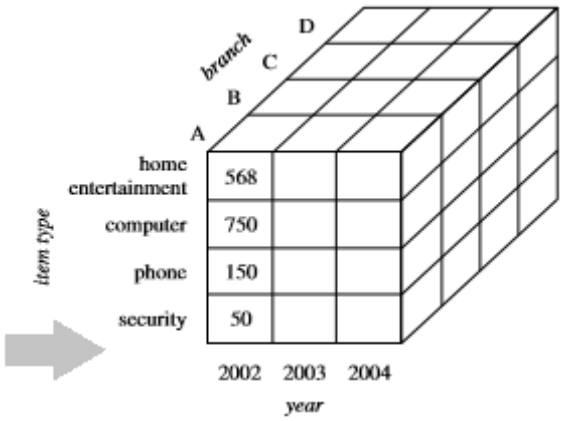


Figure (b) *A data cube for sales at All Electronics*.

2. Attribute subset selection, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

- **Redundant attributes:** Duplicate much or all of the information contained in one or more other attributes. E.g., purchase price of a product and the amount of sales tax paid
- **Irrelevant attributes:** Contain no information that is useful for the data mining task at hand. E.g., students' ID is often irrelevant to the task of predicting students' GPA

3. Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. "compressed" representation of the original data. Dimensionality Reduction is about converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions conveys much more information.

4. Numerosity reduction, reduce data volume where the data are replaced or estimated by alternative, smaller form of 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. "Can we reduce the data volume by choosing alternative, 'smaller' forms of data representation?"

5. Data Sampling

- It is one of main method for data selection i.e. sampling is the main technique employed for data selection.
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
- Sampling should be representative since it must represent approximately the same property as the original set of data.
- Get at least one object from each of 10 groups as sample data.

Types:

a) Simple Random Sampling: Equal probability of selecting any particular item.
b) Sampling without replacement: As each item is selected, it is removed from population.
c) Sampling with replacement: Objects are not removed from the population as they are selected from the sample. The same objects can be picked-up more than once.
d) Stratified Sampling: Split the data into several partitions, then draw random samples from each partition.

4. Discretization and Concept Hierarchy Generation (or summarization):

Discretization convert continuous data into discrete data and Partition data into different classes. Raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.

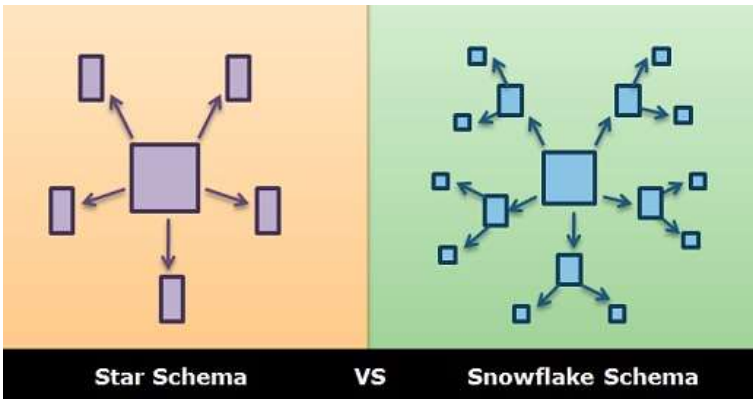
- Discretization: Reduce the number of values for a given continuous attribute by divide the range of a continuous attribute into intervals. Interval labels can then be used to replace actual data values.
- Concept Hierarchies: Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged or senior).

Approaches:

a. Equal width (distance) partitioning:
   - It divides the range into N intervals of equal size.
   - If A and B are the lowest and the highest values of the attribute, the width of interval will be - $W = (A - B)/N$.
   - The most straight forward approach for data discretization.
b. Equal depth (frequency) partitioning:
   - It divides the range into N intervals, each containing approximately same number of samples.
   - Good data scaling

- Managing categorical attributes can be tricky.

3. OLAP & Multidimensional Data Analysis



OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from *different points of view.*

OLAP is a design pattern, a way to seek information out of the physical data store. OLAP is all about summary. It aggregates information from multiple systems, and stores it in a multi-dimensional format. These could be a star schema, snowflake schema or a hybrid kind of a schema.

OLAP applications and tools are those that are designed to **ask ad hoc, complex queries** of large multidimensional collections of data.

Whereas a relational database can be thought of as two-dimensional, a multidimensional database considers each data attribute (such as product, geographic sales region, and time period) as a separate "dimension." OLAP software can locate the intersection of dimensions (all products sold in the Eastern region above a certain price during a certain time period) and display them. Attributes such as time periods can be broken down into sub attributes i.e. years, quarters, months, days etc.

OLAP can be used for data mining or the discovery of previously undiscerned relationships between data items. An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. Using Open Database Connectivity (ODBC), data can be imported from existing relational databases to create a multidimensional database for OLAP.

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 Kathmandu 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 Kathmandu in the same time period. To facilitate this kind of analysis, OLAP data is stored in a multidimensional database.*

Types of OLAP System

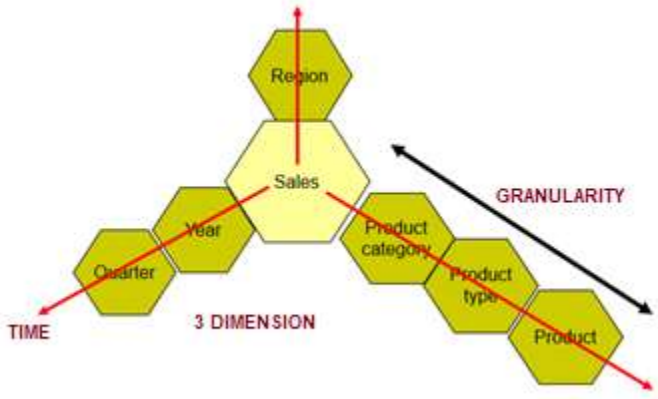
I. **ROLAP- Star Schema Based**: Relational OLAP work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables
II. **MOLAP- Cube Based**: Multidimensional OLAP where data are pre-summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In this type of model, data are structured into proprietary formats in accordance with a client's reporting requirements with the calculations pre-generated on the cubes.
III. **HOLAP**: Hybrid OLAP attempt to incorporate the best features of MOLAP and ROLAP into a single architecture

Multi-Dimensional Analysis is an Informational Analysis on data which takes into account many different relationships, each of which represents a dimension. For example, a retail analyst may want to understand the relationships among sales by region, by quarter, by demographic distribution (income, education level, gender), by product. Multi-dimensional analysis will yield results for these complex relationships.

In a multidimensional the term dimension refers to a structural attribute of a data cube. The dimension is composed or related and hierarchical members. *For instance, the "Time" dimension may have the members like years, quarters, months, weeks, day, hour and so on. In the same manner, the "Geography" dimension may have members like regions, countries, cities and so on.*

The dimension title member is the name of the member as in the case of month or city. The dimension value member is an instance of a dimension member. *For example, 2007 is the value of the dimension value which is Year.*

The Multidimensional Idea



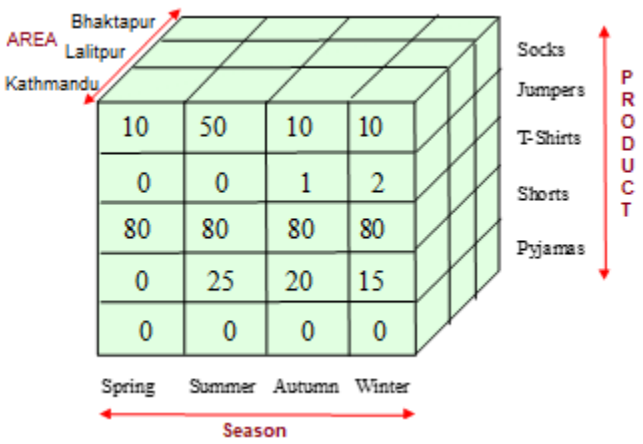


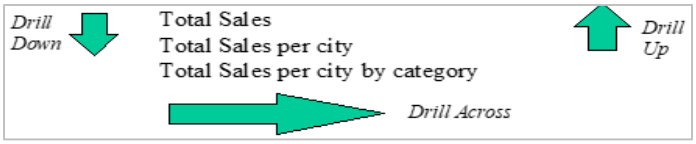
Fig. Multidimensional Model **Example:** Three dimensions – Product, Area and Season

- An OLAP cube is a data structure that allows fast analysis of data.
- OLAP tools were developed to solve multi-dimensional data analysis which stores their data in a special multi-dimensional format (data cube) with no updating facility.
- Information of multi-dimension nature can't be easily analyzed when the table has the standard 2-D representation.
- A table with n- independent attributes can be seen as an n-dimensional space.
- It is required to explore the relationships between several dimensions and standard relational databases are not very good for this.

OLAP Operations:

i. Roll-Up (Drill-Up): Summarize data
   - Takes the current aggregation level of fact values and does a further aggregation on one or more of the dimensions.
   - Equivalent to doing GROUP BY to this dimension by using attribute hierarchy.
   - Decreases a number of dimensions - removes row headers.

ii. Drill-Down (Roll Down):
   - Opposite of roll-up.
   - Summarizes data at a lower level of a dimension hierarchy, thereby viewing data in a more specialized level within a dimension.
   - Increases a number of dimensions and adds new headers



iii. Slicing:
   - Performs a selection on one dimension of the given cube, resulting in a sub-cube.
   - Reduces the dimensionality of the cubes.
   - Sets one or more dimensions to specific values and keeps a subset of dimensions for selected values.

iv. Dicing:
   - Define a sub-cube by performing a selection of one or more dimensions.
   - Refers to range select condition on one dimension, or to select condition on more than one dimension.
   - Reduces the number of member values of one or more dimensions.

v. Pivoting (Rotate):
   - Rotates the data axis to view the data from different perspectives.
   - Groups data with different dimensions.

Other OLAP operations: Some more OLAP operations include:

- SCOPING: Restricting the view of database objects to a specified subset is called scoping. Scoping will allow users to receive and update some data values they wish to receive and update.
- SCREENING: Screening is performed against the data or members of a dimension in order to restrict the set of data retrieved.
- DRILL ACROSS: Accesses more than one fact table that is linked by common dimensions. Combines cubes that share one or more dimensions.
- DRILL THROUGH: Drill down to the bottom level of a data cube down to its back end relational tables.

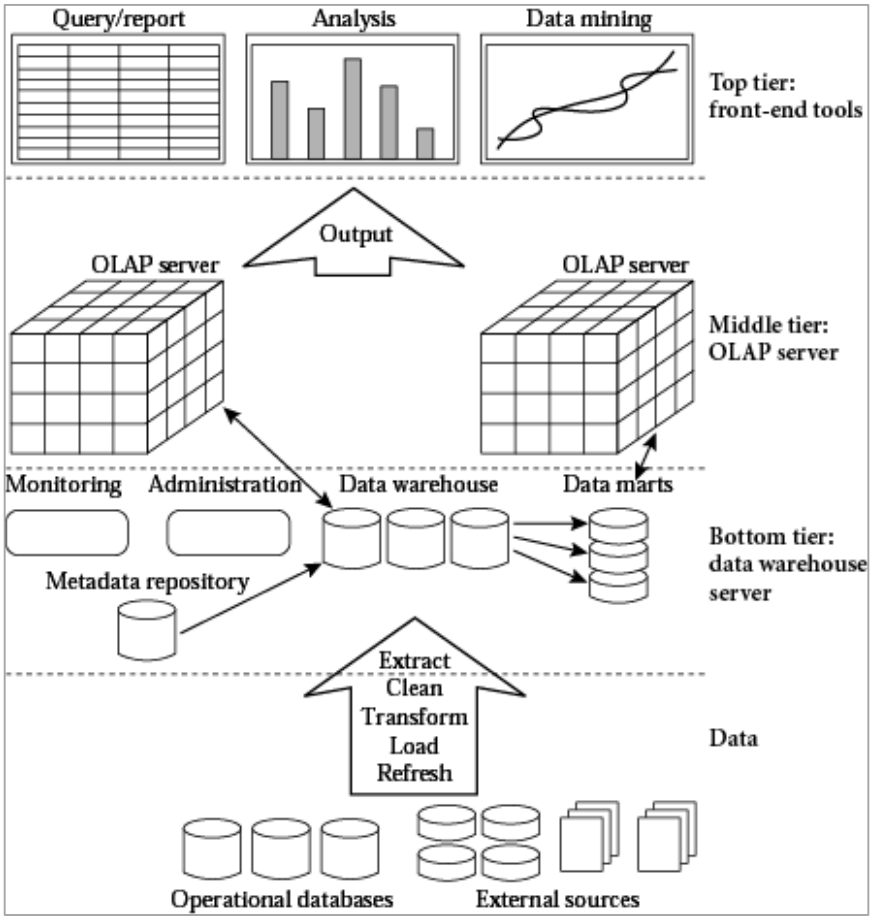
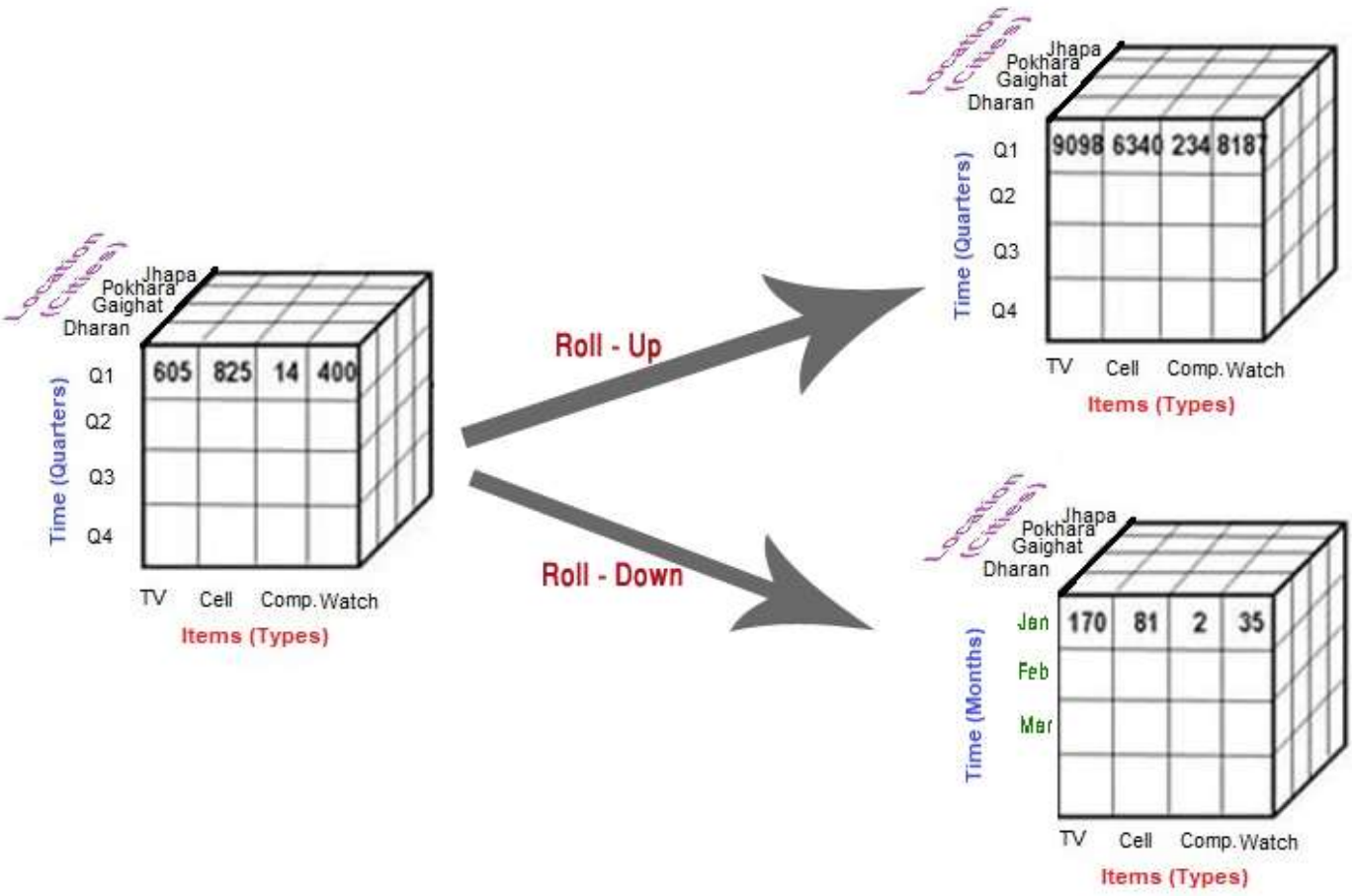
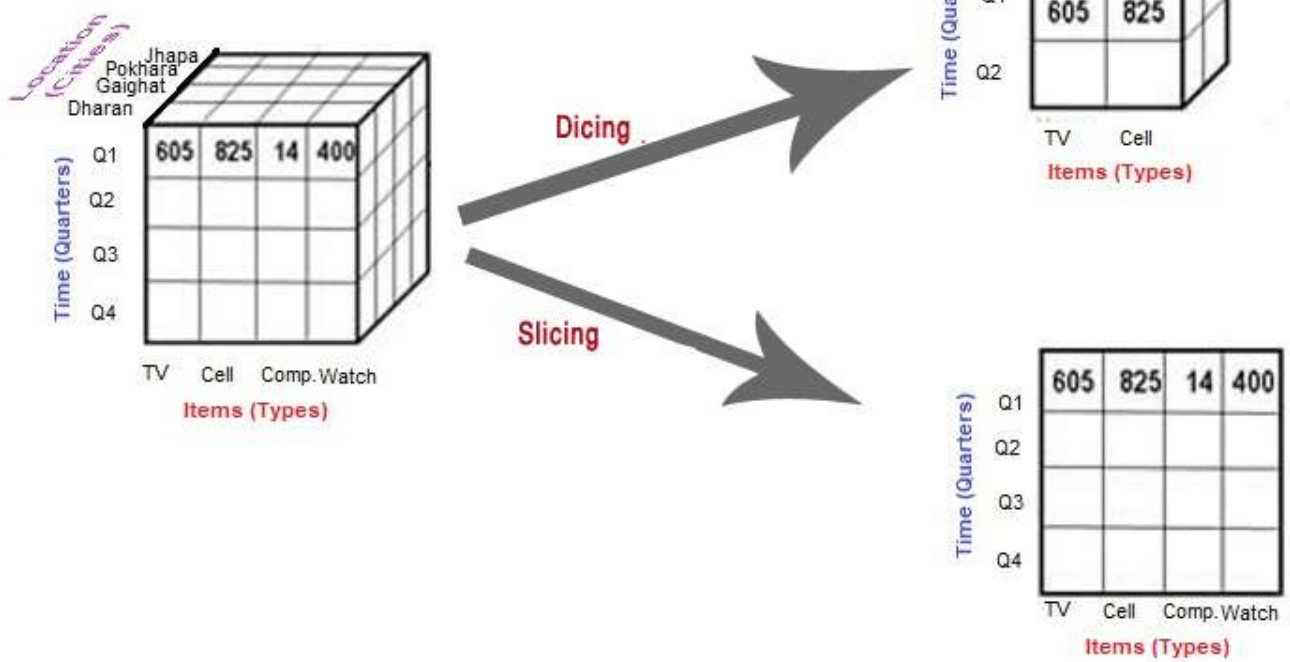


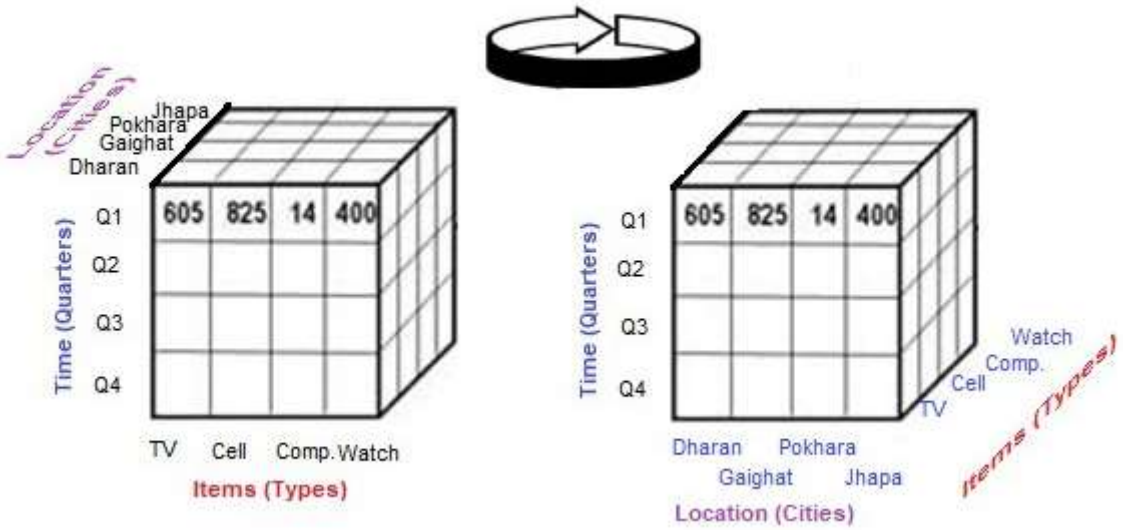
Fig. OLAP Architecture

OLAP Vs OLTP

	OLTP	OLAP
Type of 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
No of users	thousands	hundreds
DB size	100MB-GB	100GB-TB
Metric	transaction throughput	query throughput, response

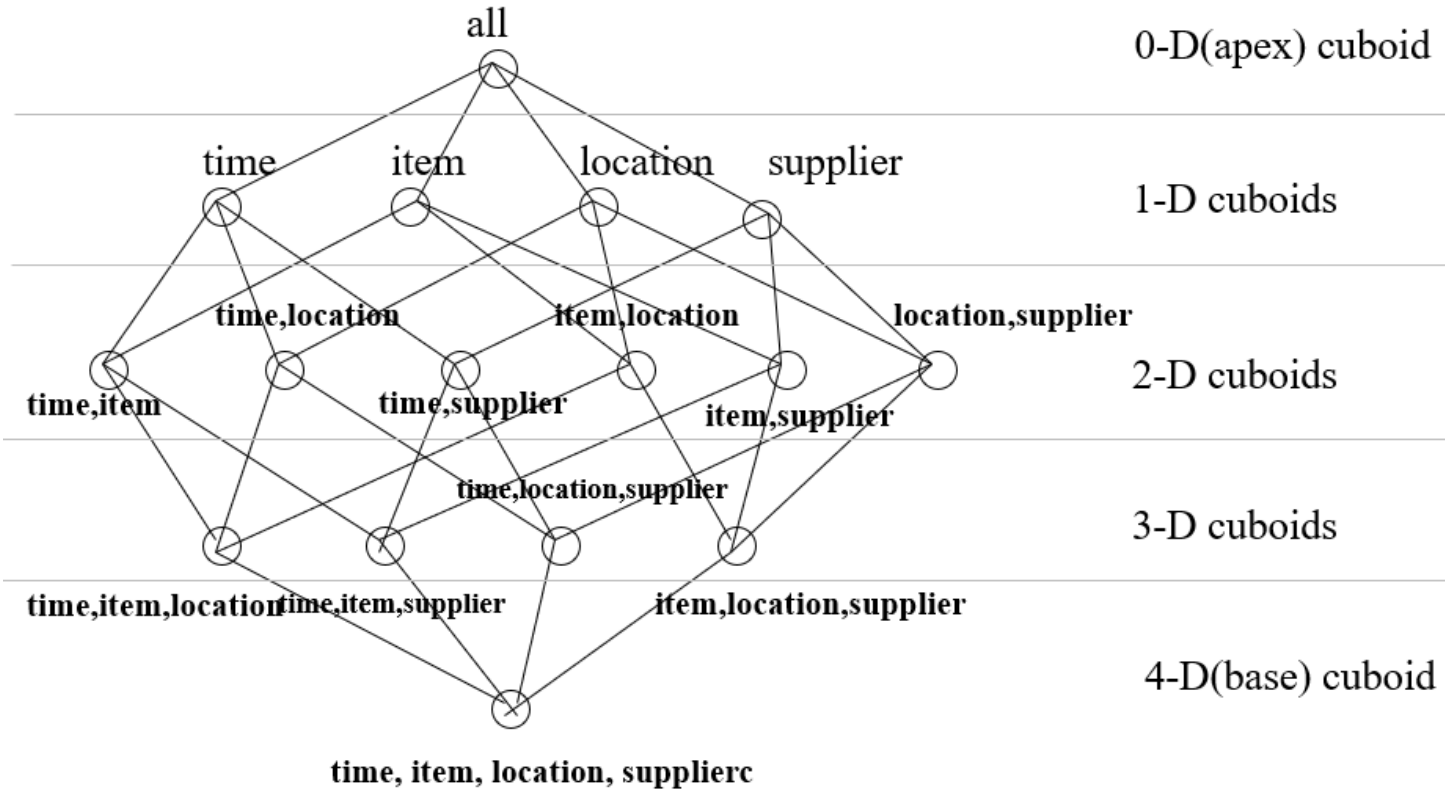

Location (Cities)
Jhapa
Pokhara
Gaighat
Dharan
Time (Quarters)
Q1
Q2
Q3
Q4
605 825 14 400
TV Cell Comp. Watch
Items (Types)
Roll - Up
9098 6340 234 8187
Roll - Down
Time (Months)
Jan
Feb
Mar
170 81 2 35


Location (Cities)
Jhapa
Pokhara
Gaighat
Dharan
Time (Quarters)
Q1
Q2
Q3
Q4
605 825 14 400
TV Cell Comp. Watch
Items (Types)
Dicing
605 825
Slicing



Data Cube Computation

- Data Cube: A Lattice of Cuboids
- A data cube is referred to as a cuboid
- The lattice of cuboids forms a data cube.
- The cuboid holding the lowest level of summarization is called a base cuboid.
  - the 4-D cuboid is the base cuboid for the given four dimensions i.e. time, item, location, supplier
- The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid.
  - typically denoted by all



4. Various Similarity Measures

Distance or similarity measures are essential to solve many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in literature to compare two data distributions.

Similarity

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0,1]
- Measure of how two instances are close to each other. The "closer" the instances are to each other, the larger is the similarity value. Example: Rank documents in searching.
- Two main consideration about similarity:
  - Similarity = 1 if X = Y (Where X, Y are two objects)
  - Similarity = 0 if X ≠ Y

Dissimilarity

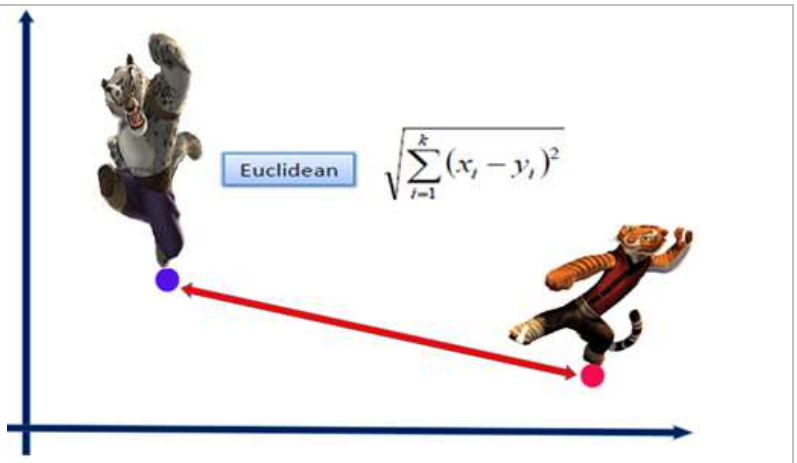
- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies
- measure of how different two instances are. Dissimilarity is large when instances are very different and is small when they are close.

Methods of SIMILARITY MEASURES

1) Euclidean distance: distance between two points is the length of the path connecting them.

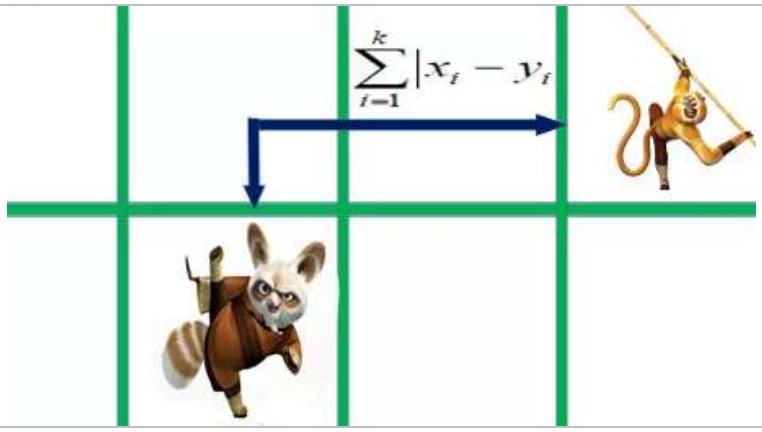
For two objects A and B

Euclidean distance = $\sqrt{[(x_i-x_j)^2 + (y_i-y_j)^2]}$



2) Manhattan distance:

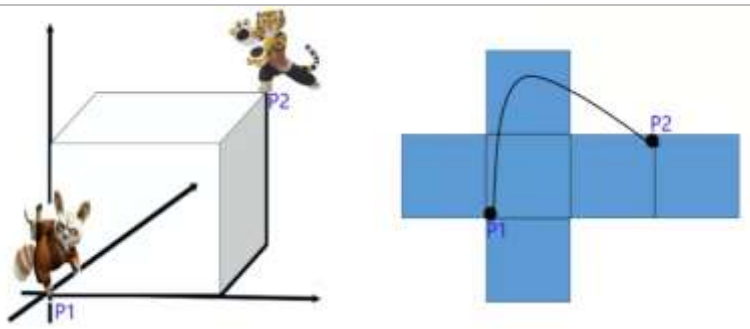
- In a plane with p1 at (x1, y1) and p2 at (x2, y2).
- Manhattan distance = $|x1 - x2| + |y1 - y2|$



3) Minkowski distance:

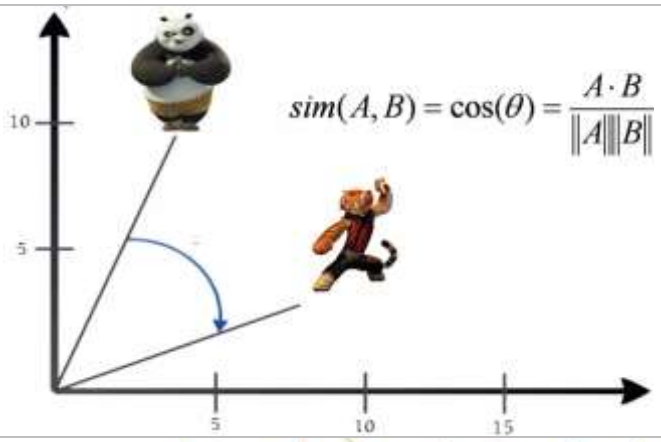
$$d^{MKD}(i,j) = \sqrt[\lambda]{\sum_{k=0}^{n-1} \left| y_{i,k} - y_{j,k} \right|^{\lambda}}$$

distance between the data record i and j, k the index of a variable, n the total number of variables y and λ the order of the Minkowski metric. Although it is defined for any λ > 0, it is rarely used for values other than 1, 2 and ∞.



4) Cosine Similarity

finds the normalized dot product of the two attributes. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects. The cosine of 0° is 1, and it is less than 1 for any other angle.



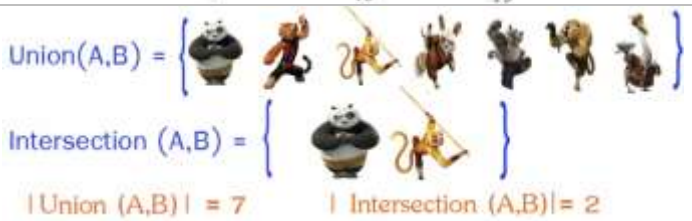
$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

5) Jaccard similarity:

= | Intersection (A,B) | / | Union (A,B) |

= 2 / 7

= 0.286



Proximity: refers to either similarity or dissimilarity

Distance metric: a measure of dissimilarity that obeys the following laws (laws of triangular norm):

- d (x, y) ≤ 0; d (x, y) = 0 if x = y;
- d (x, y) = d (y, x);
- d (x, y) + d (y, z) ≥ d (x, z).

Reference… Er. Pratap Sapkota