Objective:

- To know the importance of statistical learning.

Syllabus:

Descriptive Analytics: Data warehousing and OLAP, data summarization, data de-duplication, data visualization using CUBEs.

Learning Outcomes:

The student will be able to

- describe how to experimentally obtain and evaluate sequence information
- perform descriptive analytics over massive data.

Learning Material

6.1 Data warehousing and OLAP

Data warehouse is a large repository of data, collected from different sources and put it in common format.

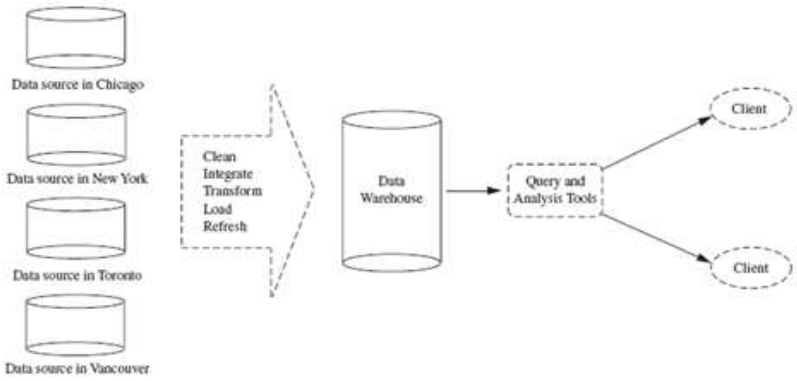


Fig : Data Warehousing

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It contains consolidated historical data, which helps the organization to analyze its business.

Def : "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

Key features of Data Warehouse

1) Subject-Oriented

- A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. The subjects are product, customers, suppliers, sales, revenue, etc.
- A data warehouse does not focus on the ongoing operations rather it focuses on modeling and analysis of data for decision making.
- It provides a simple and concise view.

2) Integrated

- Data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

3) Time Variant

- The data in a data warehouse provides information from the historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse contains an element of time, explicitly or implicitly.

4) Non-Volatile

- Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is physically stored and separate from the operational database and the frequent changes in operational database is not reflected in the data warehouse.
- 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.

Data warehouse helps business executives to organize, analyze, and use their data for decision making activities, increasing customer focus, repositioning products, analyzing operations, managing the customer relationships.

Data warehouses are widely used in the following fields −

1. Financial services
2. Banking services
3. Consumer goods
4. Retail sectors
5. Controlled manufacturing

OLAP

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows users to analyze information from multiple database systems at the same time. So, the managers and analysts get an insight of the information though fast, consistent and interactive access from the Data warehouse.

- OLAP process historical data to analyze the business.
- Analysts frequently need to group, aggregate and join data. With OLAP data can be pre-calculated, summarized and consolidated making analysis faster.
- OLAP provides summarized and multidimensional view of data.
- OLAP systems are used by knowledge workers such as executives, managers, and analysts.
- The number of users is in hundreds. The number of records accessed is in millions
- It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.

Types of OLAP Servers

We have four types of OLAP servers −

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

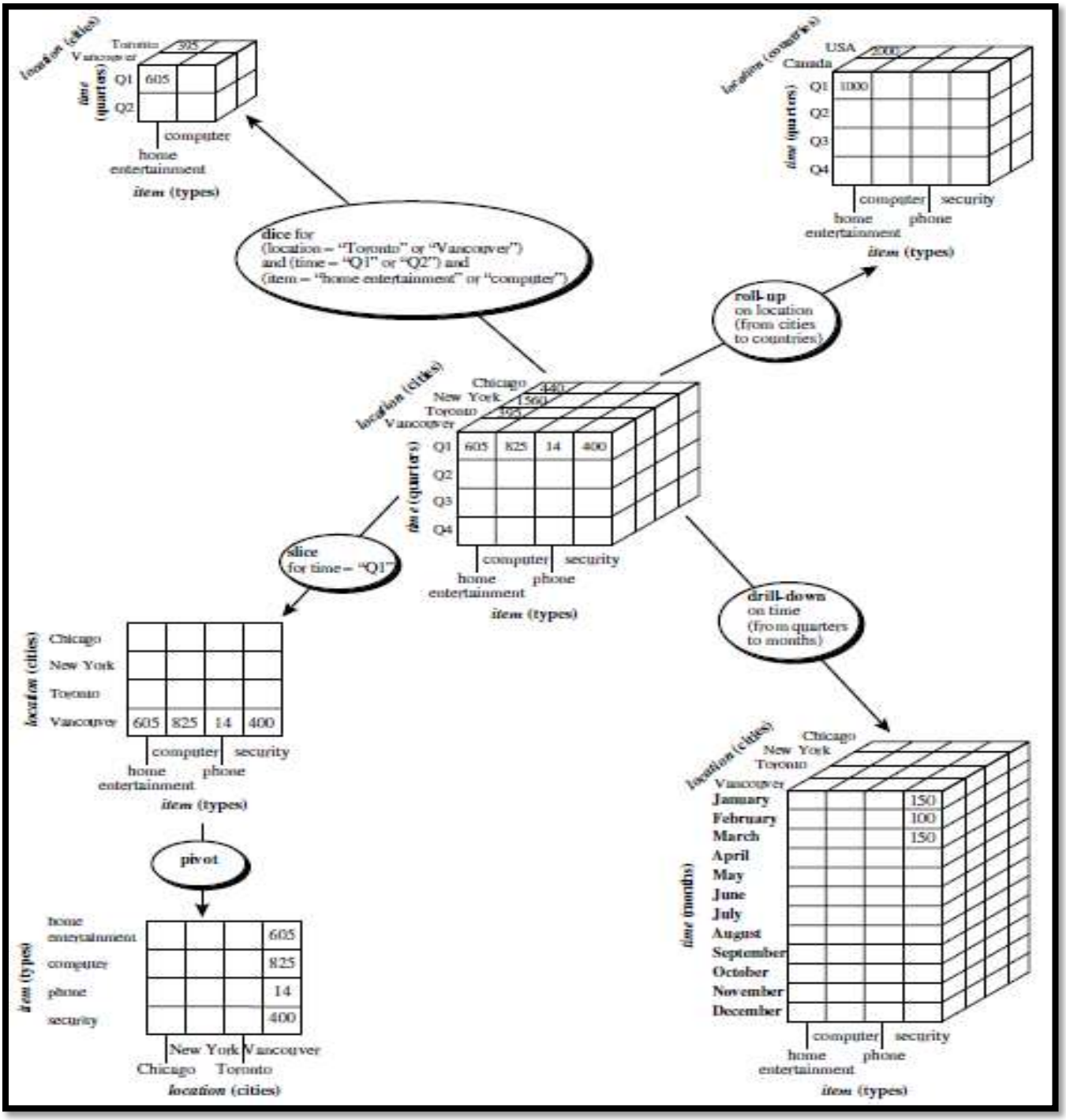
OLAP Operations

OLAP servers are based on multidimensional view of data, OLAP operations perform in multidimensional data.

Four types of analytical operations in OLAP are:

1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

Below figure shows the operations of multidimensional data model.At the center of the figure is a data cube for AllElectronics sales. The cube contains the dimensions location, time, and item. Location is aggregated with respect to city values. Time is aggregated with respect to quarters. Item is aggregated with respect to item types. The measure displayed is dollars sold (in thousands).



1) Roll-up :

The roll-up operation (also called as drill-up) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction(remove one or more dimensions). Concept hierarchy is a system of grouping things based on their order or level.

Ex: roll-up from cities to country

2) Drill-down:

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions

EX: drill-down from quarter to months

3) Slice and dice(project and select):

The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube.

Ex: Slice of time=Q1"

Dice of (location = "Toronto" or "Vancouver") and
(time = "Q1" or "Q2") and
(item ="home entertainment" or "computer").

4) Pivot (rotate):

Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

Ex: 2-D slice are rotated.

6.2 Data Summarization

Summarization is considered as a descriptive task in data mining to provide compact description of a data. Motivation is to better understand the data.

- Data summarization techniques can be used to identify the typical properties of the data and highlight which data values should be treated as noise or outlier. To learn about the characteristics such as central tendency and dispersion of the data.
- Measures of central tendency include mean, median, mode and midrange, while measures of data dispersion include quartiles, inter quartile range(IQR) and variance.

6.2.1 Measures of Central tendency

There various ways to measure the central tendency of data

- Mean
- Median
- Mode
- Midrange

Mean: The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean. (sample vs. population). Gives an idea about the mean value of the data The data is clustered around what value?

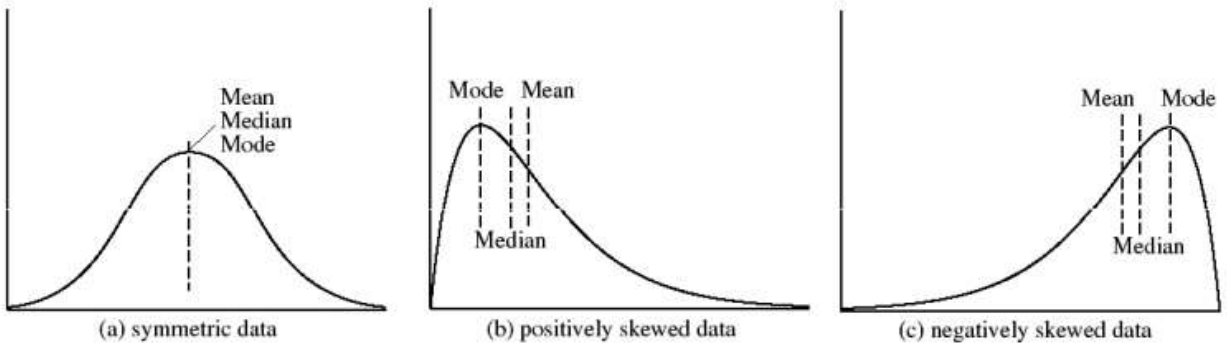
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median :Suppose that a given data set of N distinct values is sorted in numerical order. The median is the middle value if odd number of values, or average of the middle two values otherwise. For skewed (asymmetric) data, a better measure of the center of data is the median.

Mode : The mode for a set of data is the value that occurs most frequently in the set. If each data value occurs only once, then there is no mode.

Midrange:It is the average of the largest and smallest values in the set.

Mean, median, and mode of symmetric versus positively and negatively skewed data.



(a) symmetric data (b) positively skewed data (c) negatively skewed data

For symmetric data Mean. Median and mode are same (a).
For Positively skewed data mode is smaller than the median (b).
For negatively skewed data, mode is greater than the median (c).

6.2.2 Measuring the Dispersion of Data

The degree to which numerical data tend to spread is called the dispersion, or variance of the data.

The most common measures of data dispersion are

- Range
- Five-number summary (based on quartiles)
- Interquartilerange (IQR)
- Standard deviation

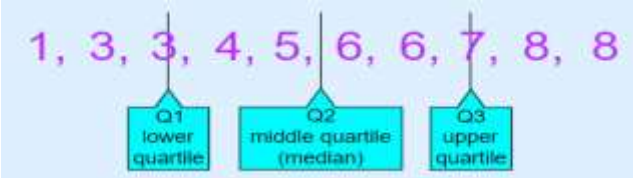
Range: Difference between highest and lowest observed values.

Quartiles: Quartiles are the values that divide a list of numbers into quarters:

- Put the list of numbers in order
- Then cut the list into four equal parts
- The Quartiles are at the "cuts"

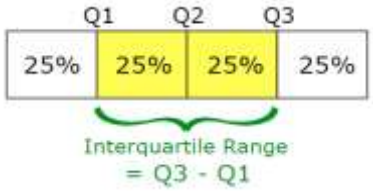
First quartile (Q1): The first quartile is the value, where 25% of the values are smaller than Q1 and 75% are larger.

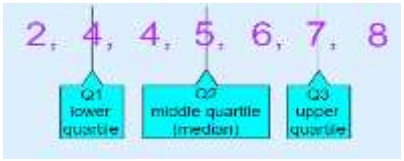
Third quartile (Q3): The third quartile is the value, where 75% of the values are smaller than Q3 and 25% are larger.



Quartile 1 (Q1) = 3
Quartile 2 (Q2) = (5+6)/2=5.5
Quartile 3 (Q3) = 7

Inter-quartile range (IQR): IQR is a simple measure of spread that gives the range covered by the middle half of the data. IQR = Q3 –Q1.





Q3-Q1 = 7-4 = 3

Outlier: usually, a value higher/lower than 1.5 x IQR

Five number summary:

The **five-number summary** is a set of descriptive statistics provide information about a dataset. It consists of the five most important sample percentiles, min, Q1, Median,Q3, max.

Data : 4,17,7,14,18,12,3,16,10,4,4,11

Put them in ascending order : 3,4,4,4,7,10,11,12,14,16,17,18

Cut in to quarters : 3,4,4 4,7,10 11,12,14 16,17,18

Min=3, Q1=(4+4)/2, Median or Q2= (10+11)/2= 10.5, Q3 =(14+16)/2=15

Variance: (sample:s, population: σ)

Variance (σ^2) is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

A large variance indicates that numbers in the set are far from the mean and from each other, while a small variance indicates the opposite. Variance can be negative. A variance value of zero indicates that all values within a set of numbers are identical. All variances that are not zero will be positive numbers.

Standard deviation: s (or σ)

The standard deviation measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance

If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

σ measures spread about the mean and should be used only when the mean is chosen as the measure of center.

σ =0 only when there is no spread, that is, when all observations have the same value.

Data Deduplication:

Data Deduplication, often called Dedup for short, is a feature that can help reduce the impact of redundant data on storage costs. When enabled, Data Deduplication optimizes free space on a volume by examining the data on the volume by looking for duplicated portions on the volume. Duplicated portions of the volume's dataset are stored once and are (optionally) compressed for additional savings. Data Deduplication optimizes redundancies without compromising data fidelity or integrity.

Drilling down into the process more, deduplication software typically generates unique identifiers for data using cryptographic hash functions.

Deduplication at the file level is inefficient because even if a minuscule part of a file is altered, such as a single bit, a whole new copy of that file will be stored.

OLAP (online analytical processing) cube extends a 2-dimensional array (spreadsheet table or array of facts/measures and keys/pointers to dictionaries) to a multidimensional **DataCube**, and on other hand DataCube is using **datawarehouse** schemas like Star Schema or Snowflake Schema.

The **OLAP cube** consists of **facts**, also called **measures**, categorized by **dimensions** (it can be much more than 3 Dimensions; dimensions referred from Fact Table by "foreign keys").

Measures are derived from the records in the **Fact Table** and Dimensions are derived from the dimension tables, where each column represents one **attribute** (also called **dictionary**; dimension can have many attributes).

Such multidimensional DataCube organization is close to a Columnar DB data structures. One of the most popular usage of datacubes is a visualization of them in form of Pivot tables, where attributes used as rows, columns and filters while values in cells are appropriate aggregates (**SUM, AVG, MAX, MIN**, etc.) of measures.

OLAP operations are foundation for most UI and functionality used by Data Visualization tools. The DV user (sometimes called analyst) navigates through the DataCube and its DataViews for a particular subset of the data, changing the data's orientations and defining analytical calculations. The user-initiated process of navigating by calling for page displays interactively, through the specification of slices via rotations and drill down/up is sometimes called "slice and dice". Common operations include slice and dice, drill down, roll up, and pivot:

Slice:

A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Assignment –Cum –Tutorial Questions

Unit-VI

SECTION A

1. ____generalize and consolidate data in multidimensional space.?

a) **Data warehouse**

b) Kappa

c) RMSE

d) All of the Mentioned

2. A _____ is a subject-oriented, integrated, time-variant,

and nonvolatile collection of data in support of management's decision making process.

A. relational database system B. transaction processing system C. file systems D. data warehouse

3.The field of ____provides many additional techniques for viewing data through graphical means.

A. data visualization

4. Which of the following are TRUE?

A. visualization can help identify relations, trends, and biases "hidden"in unstructured data sets.

B. Techniques may be as simple as scatter-plot matrices (where two attributes are mapped onto a 2-D grid)

C. Methods such as tree-maps (where a hierarchical partitioning of the screen is displayed based on the attribute values.

5. An attribute is a data field, representing a characteristic or feature of a data object.

A. data visualization

6. Which of the following are FALSE?

A. Nominal data means "relating to names."

B. The values of a nominal attribute are symbols or names of things.

C. Each nominal value represents some kind of category, code, or state.

D. Nominal attributes are also referred to as categorical.

E. Nominal values do not have any meaningful order.

7. Which of the following are TRUE?

A. Measures of central tendency (MCT), which measure the location of the middle or center of a data distribution.

B. MCT is Intuitively speaking, given an attribute, where do most of its values fall?

C. Data in most real applications are not symmetric,

D. Data may be either positively skewed, where the mode occurs at a value that is smaller than the median,

E. Data may be negatively skewed, where the mode occurs at a value greater than the median

II) Descriptive Questions(6 to 8)

1. Discuss about data de-duplication.
2. Describe data summarization.
3. Discuss Data warehousing and OLAP.
4. Describe data visualization using CUBEs.
5. Compare DATA CUBE and DATA MART.
6. Analyze different OLAP Operations.
7. Distinguish between Fact Table and Dimension table.
8. Discuss different schemas in Datawarehouse.