Data warehousing and OLAP and visualization using CUBEs:

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

Types of OLAP Servers

We have four types of OLAP servers −

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

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following −

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

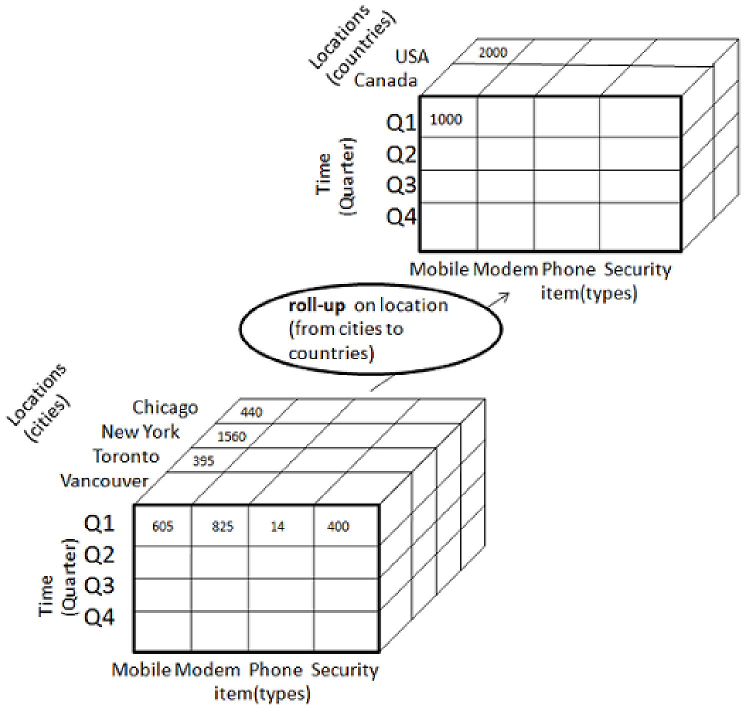
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

- By climbing up a concept hierarchy for a dimension
- By dimension reduction

The following diagram illustrates how roll-up works.



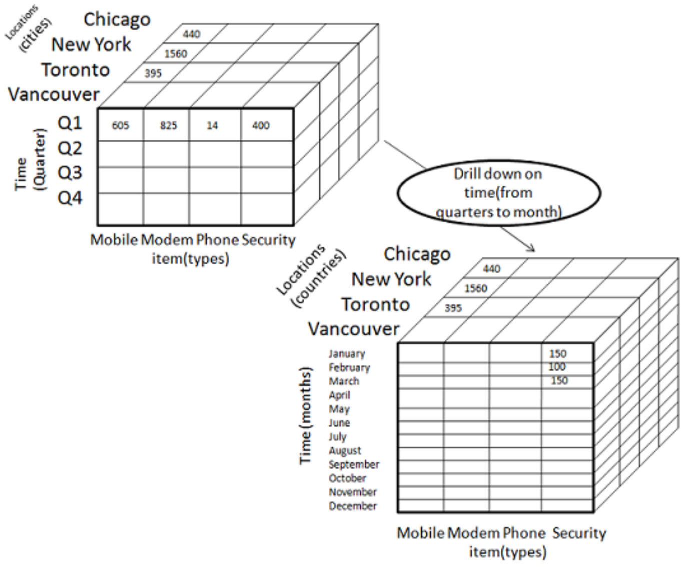
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

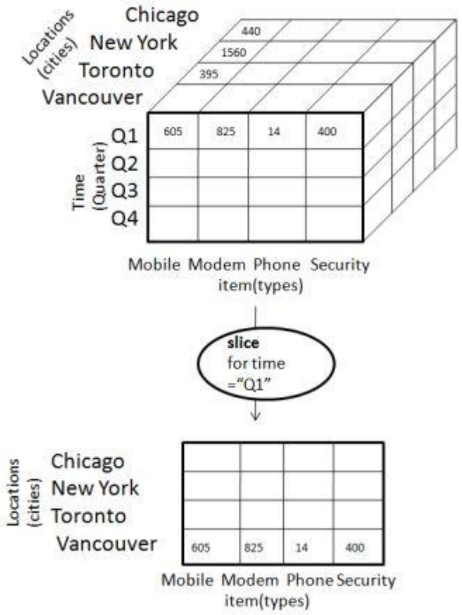
The following diagram illustrates how drill-down works −



- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

Slice

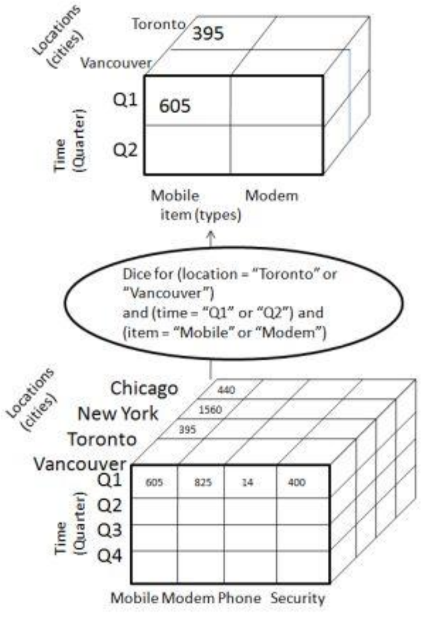
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

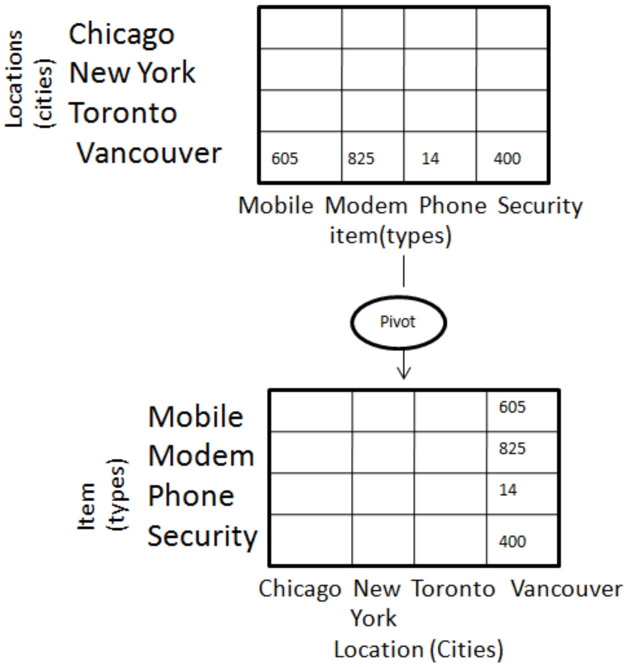


The dice operation on the cube based on the following selection criteria involves three dimensions.

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item =" Mobile" or "Modem")

Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



Data summarization:

Data Summarization in Data Mining is a key concept from which a concise description of a dataset can be obtained to see what looks normal or out of place. A carefully chosen summary of raw data would convey many trends and patterns of the data in an easily accessible manner.

In general, data can be summarized numerically in the form of a table known as tabular summarization or visually in the form of a graph known as data visualization.

The different types of Data Summarization in Data Mining are:

- **Tabular Summarization:** This method instantly conveys patterns such as frequency distribution, cumulative frequency, etc, and
- **Data Visualization**: Visualisations from a chosen graph style such as histogram, time-series line graph, column/bar graphs, etc. can help to spot trends immediately in a visually appealing way.

There are three areas in which you can implement Data Summarization in Data Mining. These are as follows:

- Data Summarization in Data Mining: Centrality
- Data Summarization in Data Mining: Dispersion
- Data Summarization in Data Mining: Distribution of a Sample of Data

1) Data Summarization in Data Mining: Centrality

The principle of Centrality is used to describe the center or middle value of the data.

Several measures can be used to show the centrality of which the common ones are average also called mean, median, and mode. The three of them summarize the distribution of the sample data.

- **Mean:** This is used to calculate the numerical average of the set of values.
- **Mode:** This shows the most frequently repeated value in a dataset.
- **Median:** This identifies the value in the middle of all the values in the dataset when values are ranked in order.

The most appropriate measure to use will depend largely on the shape of the dataset.

2) Data Summarization in Data Mining: Dispersion

The dispersion of a sample refers to how spread out the values are around the average (center). Looking at the spread of the distribution of data shows the amount of variation or diversity within the data. When the values are close to the center, the sample has low dispersion while high dispersion occurs when they are widely scattered about the center.

Different measures of dispersion can be used based on which is more suitable for your dataset and what you want to focus on. The different measures of dispersion are as follows:

- **Standard deviation:** This provides a standard way of knowing what is normal, showing what is extra large or extra small and helping you to understand the spread of the variable from the mean. It shows how close all the values are to the mean.
- **Variance:** This is similar to standard deviation but it measures how tightly or loosely values are spread around the average.
- **Range:** The range indicates the difference between the largest and the smallest values thereby showing the distance between the extremes.

3) Data Summarization in Data Mining: Distribution of a Sample of Data

The distribution of sample data values has to do with the shape which refers to how data values are distributed across the range of values in the sample. In simple terms, it means if the values are clustered around the average to show how they are symmetrically arranged around it or if there are more values to one side than the order. Two ways to explore the distribution of the sample data are graphically and through shape statistics.

To draw a picture of the data distribution graphically, frequency histograms and tally plots can be used to summarize the data.

- **Histograms:** Histograms are similar to bar charts where a bar represents the frequency of values in the data that correspond to various size classes but the difference is that the bars are drawn without gaps in them to show the x-axis representing a continuous variable.
- **Tally plots:** A tally plot is a kind of data frequency distribution graph that can be used to represent the values from a dataset.

For shape statistics, skewness and kurtosis can help give values to how central the average is and show how clustered they are around the data average.

- **Skewness:** This is a measure of how central the average is in the distribution. The skewness of a sample is a measure of how central the average is to the overall spread of values.

- **Kurtosis:** This is a measure of how pointy the distribution is. The Kurtosis of a sample is a measure of how pointed the distribution is, it shows how clustered the values are around the middle.

Determining the shape of the distribution of your data goes a long way in helping you decide which statistical option to choose from when performing data summarization and subsequent analysis through data mining.

Data de-duplication:

Data deduplication is a process that eliminates redundant copies of data and reduces storage overhead.

Data deduplication techniques ensure that only one unique instance of data is retained on storage media, such as disk, flash or tape. Redundant data blocks are replaced with a pointer to the unique data copy. In that way, data deduplication closely aligns with incremental backup, which copies only the data that has changed since the previous backup

Techniques to deduplicate data

There are two main methods to deduplicate redundant data: inline and post-processing deduplication. The backup environment will dictate the method.

- **Inline deduplication** analyzes data as a backup system ingests it. Redundancies are removed as the data is written to backup storage. Inline dedupe requires less backup storage but can cause bottlenecks. Storage array vendors recommend that users turn off their inline data deduplication tools for high-performance primary storage.
- **Post-processing** dedupe is an asynchronous backup process that removes redundant data after it is written to storage. Duplicate data is removed and replaced with a pointer to the first iteration of the block. The post-processing approach gives users the flexibility to dedupe specific workloads and quickly recover the most recent backup without hydration. The tradeoff is a larger backup storage capacity than is required with inline deduplication.