**Differences between database and data warehouse**

A **database** is a collection of information that is organized so that it can be easily accessed, managed and updated. Data is organized into rows, columns and tables, and it is indexed to make it easier to find relevant information. Data gets updated, expanded and deleted as new information is added. Databases process workloads to create and update themselves, querying the data they contain and running applications against it.

A **data warehouse** is a federated repository for all the data that an enterprise's various business systems collect. The repository may be physical or logical. Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user who may need access to specialized, sometimes local databases. The latter idea is known as the datamart.

**Database**

1. Used for Online Transactional Processing ([OLTP](https://en.wikipedia.org/wiki/Online_transaction_processing)) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.
2. The tables and joins are complex since they are normalized (for [RDMS](https://en.wikipedia.org/wiki/Relational_database_management_system)). This is done to reduce redundant data and to save storage space.
3. Entity – Relational modeling techniques are used for RDMS database design.
4. Optimized for write operation.
5. Performance is low for analysis queries.

**Data Warehouse**

1. Used for Online Analytical Processing ([OLAP](https://en.wikipedia.org/wiki/Online_analytical_processing)). This reads the historical data for the Users for business decisions.
2. The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
3. Data – Modeling techniques are used for the Data Warehouse design.
4. Optimized for read operations.
5. High performance for analytical queries.
6. Is *usually* a Database.

It's important to note as well that Data Warehouses could be sourced from zero to many databases.

**Hierarchical Clustering?**

* A hierarchical clustering method works by grouping data objects into a tree of clusters.
* It uses distance (similarity) matrix as clustering criteria.
* Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

In general, there are two types of hierarchical clustering methods:

1. **Agglomerative hierarchical clustering:**

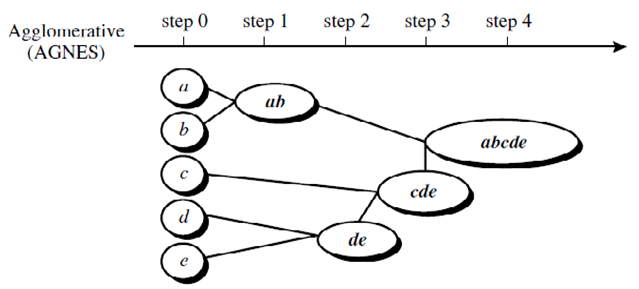
* This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.

1. **Divisive hierarchical clustering:**

* This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.
* It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

AGGLOMERATIVE HIERARCHICAL CLUSTERING: - Figure shows the application of AGNES (AGglomerativeNESting), an agglomerative hierarchical clustering method to a data set of five objects(a, b, c, d, e).

* Initially, AGNES places each object into a cluster of its own.
* The clusters are then merged step-by-step according to some criterion.



Agglomerative Algorithm: (AGNES)

Given

-a set of N objects to be clustered

-an N\*N distance matrix,

The basic process of clustering id this:

Step1: Assign each object to a cluster so that for N objects we have N clusters each containing just one Object.

Step2: Let the distances between the clusters be the same as the distances between the objects they contain.

Step3: Find the most similar pair of clusters and merge them into a single cluster so that we now have one cluster less.

Step4: Compute distances between the new cluster and each of the old clusters.

Step5: Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

* Step 4 can be done in different ways and this distinguishes single and complete linkage.

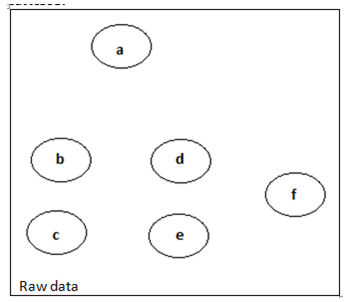
-> For complete-linkage algorithm:

* + Clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold.

-> For single-linkage algorithm:

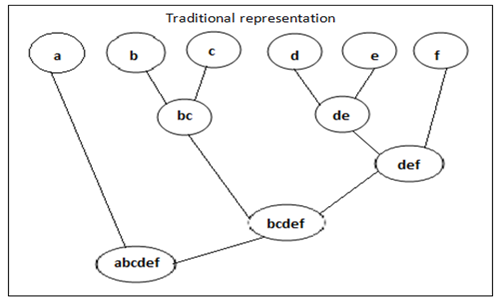
* + Clustering process is terminated when the minimum distance between nearest clusters exceeds an arbitrary threshold. **EXAMPLE:**

Suppose this data is to be clustered.



* In this example, cutting the tree after the second row of the dendrogram will yield clusters {a} {b c} {d e} {f}.
* Cutting the tree after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number but larger clusters.

The hierarchical clustering dendrogram would be as such:



In our example, we have six elements {a} {b} {c} {d} {e} and {f}.

The first step is to determine which elements to merge in a cluster.

Usually, we take the two closest elements, according to the chosen distance.

Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. Suppose we have merged the two closest elements b and c, we now have the following clusters {a}, {b, c}, {d}, {e} and {f}, and want to merge them further.

To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters *A* and *B* is one of the following:

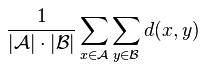
* The maximum distance between elements of each cluster (also called complete-linkage clustering):

max {d(x,y):x∈A,y∈B}

* The minimum distance between elements of each cluster (also called single-linkage clustering):

min {d(x,y):x∈A,y∈B}

* The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):



* Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

**Differences between Classification and Clustering**

## What is Clustering?

Basically, clustering involves grouping data with respect to their similarities. It is primarily concerned with distance measures and clustering algorithms which calculate the difference between data and divide them systematically.

For instance, students with similar learning styles are grouped together and are taught separately from those with differing learning approaches.  In data mining, clustering is most commonly referred to as “unsupervised learning technic” as the grouping is based on a natural or inherent characteristic.

It is applied in several scientific fields such as information [technology](http://www.differencebetween.net/technology/difference-between-technology-and-engineering/), [biology](http://www.differencebetween.net/science/biology-science/differences-between-oceanography-and-marine-biology/), [criminology](http://www.differencebetween.net/language/difference-between-criminal-justice-and-criminology/), and [medicine](http://www.differencebetween.net/science/health/drugs-health/difference-between-medicine-and-medication/).

### Characteristics of Clustering:

* **No Exact Definition**

Clustering has no precise definition that is why there are various clustering algorithms or cluster models. Roughly speaking, the two kinds of clustering are [hard and soft](http://www.differencebetween.net/technology/difference-between-soft-link-and-hard-link-in-unix-in-os/). Hard clustering is concerned with labeling an object as simply belonging to a cluster or not. In contrast, soft clustering or fuzzy clustering specifies the degree as to how something [belongs to a certain group](http://www.differencebetween.net/language/words-language/difference-between-grouped-data-and-ungrouped-data/).

* **Difficult to be Evaluated**

The validation or assessment of results from clustering analysis is often difficult to ascertain due to its inherent inexactness.

* **Unsupervised**

As it is an unsupervised learning strategy, the analysis is merely based on current features; thus, no stringent regulation is needed.

## What is Classification?

Classification entails assigning labels to existing situations or classes; hence, the term “classification”. For example, students exhibiting certain learning characteristics are classified as visual learners.

Classification is also known as “supervised learning technic” wherein machines learn from already labeled or classified data. It is highly applicable in pattern recognition, statistics, and biometrics.

### Characteristics of Classification

* **Utilizes a “Classifier”**

To analyze data, a classifier is a defined algorithm that concretely maps an information to a specific class. For example, a classification algorithm would train a model to identify whether a certain cell is malignant or benign.

* **Evaluated Through Common Metrics**

The quality of a classification analysis is often assessed via precision and recall which are popular metric procedures. A classifier is evaluated regarding its accuracy and sensitivity in identifying the output.

* **Supervised**

Classification is a supervised learning technic as it assigns previously determined identities based on comparable features. It deduces a function from a labeled training set.

## Clustering vs Classification: Table comparing the difference between Clustering and Classification

|  |  |
| --- | --- |
| **Clustering** | **Classification** |
| Unsupervised data | Supervised data |
| Does not highly value training sets | Does highly value training sets |
| Works solely with unlabeled data | Involves both unlabeled and labeled data |
| Aims to identify similarities among data | Aims to verify where a datum belongs to |
| Specifies required change | Does not specify required improvement |
| Has a single phase | Has two phases |
| Determining boundary conditions is not paramount | Identifying the boundary conditions is essential in executing the phases |
| Does not generally deal with prediction | Deals with prediction |
| Mainly employs two algorithms | Has a number of probable algorithms to use |
| Process is less complex | Process is more complex |

**Pattern Classifier or Pattern recognition or Pattern Classification**

**Pattern recognition** is the automated recognition of patterns and regularities in [data](https://en.wikipedia.org/wiki/Data). Pattern recognition is closely related to [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning),together with applications such as [data mining](https://en.wikipedia.org/wiki/Data_mining) and [knowledge discovery in databases](https://en.wikipedia.org/wiki/Knowledge_discovery_in_databases) (KDD), and is often used interchangeably with these terms. However, these are distinguished: machine learning is one approach to pattern recognition, while other approaches include hand-crafted (not learned) rules or [heuristics](https://en.wikipedia.org/wiki/Heuristic); and pattern recognition is one approach to artificial intelligence, while other approaches include [symbolic artificial intelligence](https://en.wikipedia.org/wiki/Symbolic_artificial_intelligence).  A modern definition of pattern recognition is:

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.

Pattern recognition algorithms generally aim to provide a reasonable answer for all possible inputs and to perform "most likely" matching of the inputs, taking into account their statistical variation. This is opposed to [pattern matching](https://en.wikipedia.org/wiki/Pattern_matching) algorithms, which look for exact matches in the input with pre-existing patterns. A common example of a pattern-matching algorithm is [regular expression](https://en.wikipedia.org/wiki/Regular_expression) matching, which looks for patterns of a given sort in textual data and is included in the search capabilities of many [text editors](https://en.wikipedia.org/wiki/Text_editor) and [word processors](https://en.wikipedia.org/wiki/Word_processor). In contrast to pattern recognition, pattern matching is not generally a type of machine learning, although pattern-matching algorithms (especially with fairly general, carefully tailored patterns) can sometimes succeed in providing similar-quality output of the sort provided by pattern-recognition algorithms.

K-Nearest Neighbours

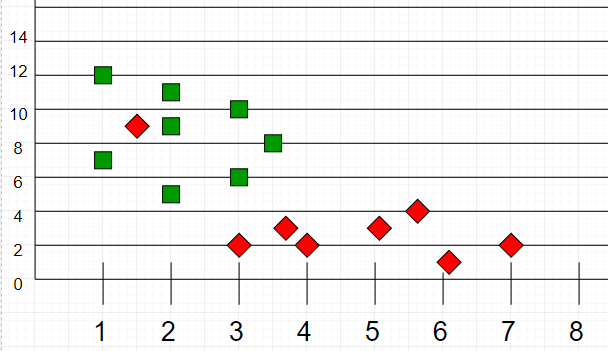
In [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), the ***k*-nearest neighbors algorithm** (***k*-NN**) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis).[[1]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-1)In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether *k*-NN is used for classification or regression:

* In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
* In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

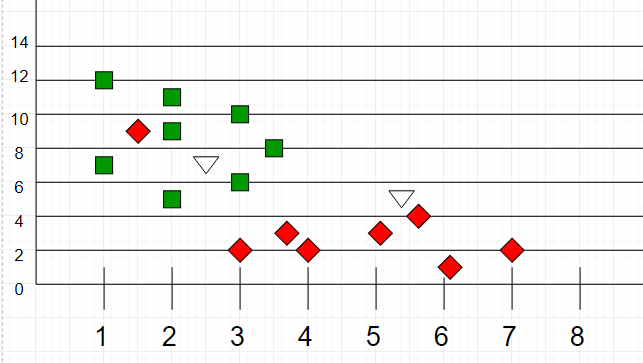
K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as [GMM](https://en.wikipedia.org/wiki/Mixture_model), which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features:  


Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as ‘yellow’.



**Intuition**  
If we plot these points on a graph, we may be able to locate some clusters, or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbours belong to. This means, a point close to a cluster of points classified as ‘Red’ has a higher probability of getting classified as ‘Red’.

Intuitively, we can see that the first point (2.5, 7) should be classified as ‘Blue’ and the second point (5.5, 4.5) should be classified as ‘Red’.

**Algorithm**  
Let m be the number of training data samples. Let p be an unknown point.

1. Store the training samples in an array of data points arr[]. This means each element of this array represents a tuple (x, y).
2. for i=0 to m:
3. Calculate Euclidean distance d(arr[i], p).
4. Make set S of K smallest distances obtained. Each of these distances correspond to an already classified data point.
5. Return the majority label among S.

**Online Analytical Mining**

# OLAP Mining (OLAM): An Integration of Data Mining and Data Warehousing

## On-line analytical mining of data warehouse data: integration of mining and OLAP technologies.

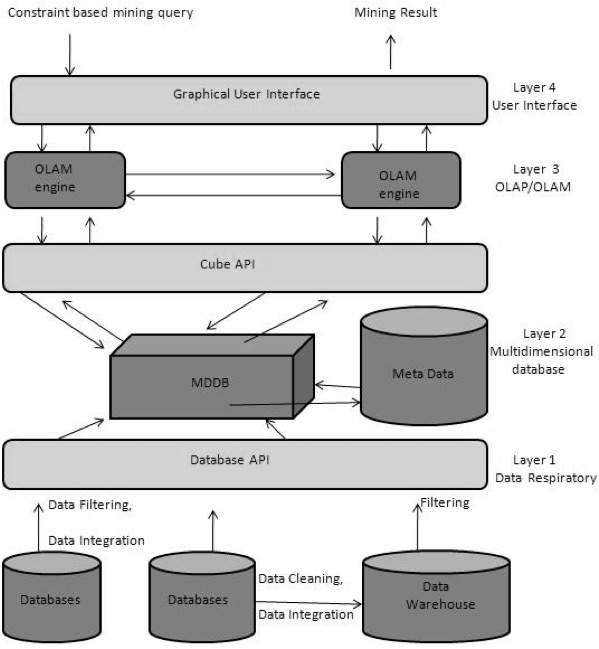
## Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.

## Interactive characterization, comparison, association, classification, clustering, prediction.

## Integration of data mining functions, e.g., first clustering and then association

## From Data Warehousing (OLAP) to Data Mining (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows the integration of both OLAP and OLAM –



## Importance of OLAM

OLAM is important for the following reasons −

* **High quality of data in data warehouses** − The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the preprocessing of data. The data warehouses constructed by such preprocessing are valuable sources of high quality data for OLAP and data mining as well.
* **Available information processing infrastructure surrounding data warehouses** − Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
* **OLAP−based exploratory data analysis** − Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subset of data and at different levels of abstraction.
* **Online selection of data mining functions** − Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

# OLTP vs OLAP: What's the Difference

## What is OLAP?

Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.

**The primary objective is data analysis and not data processing**.

## What is OLTP?

Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.

**The primary objective is data processing and not data analysis**

## Example of OLAP

Any Datawarehouse system is an OLAP system. Uses of OLAP are as follows

* A company might compare their mobile phone sales in September with sales in October, then compare those results with the with another location which may be stored in a sperate database.
* Amazon analyzes purchases by its customers to come up with a personalized homepage with products which likely interest to their customer.

## Example of OLTP system

An example of OLTP system is ATM center. Assume that a couple has a joint account with a bank. One day both simultaneously reach different ATM centers at precisely the same time and want to withdraw total amount present in their bank account.

However, the person that completes authentication process first will be able to get money. In this case, OLTP system makes sure that withdrawn amount will be never more than the amount present in the bank. The key to note here is that OLTP systems are optimized for**transactional superiority instead data analysis.**

Other examples of OLTP system are:

* Online banking
* Online airline ticket booking
* Sending a text message
* Order entry
* Add a book to shopping cart

## Benefits of using OLAP services

* OLAP creates a single platform for all type of business analytical needs which includes planning, budgeting, forecasting, and analysis.
* The main benefit of OLAP is the consistency of information and calculations.
* Easily apply security restrictions on users and objects to comply with regulations and protect sensitive data.

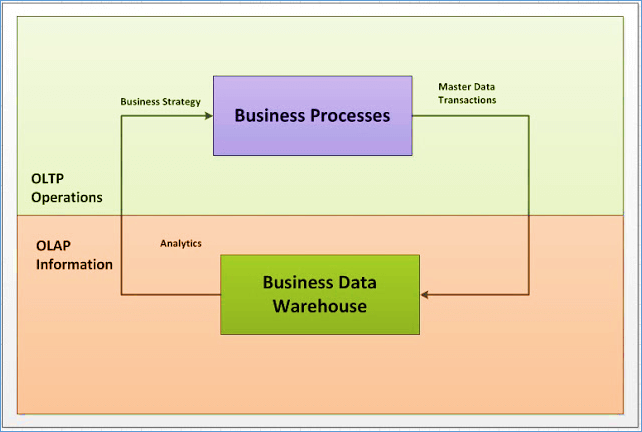
## Benefits of OLTP method

* It administers daily transactions of an organization.
* OLTP widens the customer base of an organization by simplifying individual processes.

## Drawbacks of OLAP service

* Implementation and maintenance are dependent on IT professional because the traditional OLAP tools require a complicated modeling procedure.
* OLAP tools need cooperation between people of various departments to be effective which might always be not possible.

## Drawbacks of OLTP method

* If OLTP system faces hardware failures, then online transactions get severely affected.
* [](https://www.guru99.com/images/1/022218_0431_OLTPvsOLAPW1.png)OLTP systems allow multiple users to access and change the same data at the same time which many times created unprecedented situation.

## Difference between OLTP and OLAP

|  |  |  |
| --- | --- | --- |
| **Parameters** | **OLTP** | **OLAP** |
| **Process** | It is an online transactional system. It manages database modification. | OLAP is an online analysis and data retrieving process. |
| **Characteristic** | It is characterized by large numbers of short online transactions. | It is characterized by a large volume of data. |
| **Functionality** | OLTP is an online database modifying system. | OLAP is an online database query management system. |
| **Method** | OLTP uses traditional DBMS. | OLAP uses the data warehouse. |
| **Query** | Insert, Update, and Delete information from the database. | Mostly select operations |
| **Table** | Tables in OLTP database are normalized. | Tables in OLAP database are **not**normalized. |
| **Source** | OLTP and its transactions are the sources of data. | Different OLTP databases become the source of data for OLAP. |
| **Data Integrity** | OLTP database must maintain data integrity constraint. | OLAP database does not get frequently modified. Hence, data integrity is not an issue. |
| **Response time** | It's response time is in millisecond. | Response time in seconds to minutes. |
| **Data quality** | The data in the OLTP database is always detailed and organized. | The data in OLAP process might not be organized. |
| **Usefulness** | It helps to control and run fundamental business tasks. | It helps with planning, problem-solving, and decision support. |
| **Operation** | Allow read/write operations. | Only read and rarely write. |
| **Audience** | It is a market orientated process. | It is a customer orientated process. |
| **Query Type** | Queries in this process are standardized and simple. | Complex queries involving aggregations. |
| **Back-up** | Complete backup of the data combined with incremental backups. | OLAP only need a backup from time to time. Backup is not important compared to OLTP |
| **Design** | DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc. | DB design is subject oriented. Example: Database design changes with subjects like sales, marketing, purchasing, etc. |
| **User type** | It is used by Data critical users like clerk, DBA & Data Base professionals. | Used by Data knowledge users like workers, managers, and CEO. |
| **Purpose** | Designed for real time business operations. | Designed for analysis of business measures by category and attributes. |
| **Performance metric** | Transaction throughput is the performance metric | Query throughput is the performance metric. |
| **Number of users** | This kind of Database users allows thousands of users. | This kind of Database allows only hundreds of users. |
| **Productivity** | It helps to Increase user's self-service and productivity | Help to Increase productivity of the business analysts. |
| **Challenge** | Data Warehouses historically have been a development project which may prove costly to build. | An OLAP cube is not an open SQL server data warehouse. Therefore, technical knowledge and experience is essential to manage the OLAP server. |
| **Process** | It provides fast result for daily used data. | It ensures that response to the query is quicker consistently. |
| **Characteristic** | It is easy to create and maintain. | It lets the user create a view with the help of a spreadsheet. |
| **Style** | OLTP is designed to have fast response time, low data redundancy and is normalized. | A data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database |

## Summary:

* Online Analytical Processing is a category of software tools that analyze of data stored in a database.
* Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture
* OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis.
* OLTP is useful to administer day to day transactions of an organization.
* OLAP is characterized by a large volume of data.
* OLTP is characterized by large numbers of short online transactions.
* A data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database.
* An OLAP Cube takes a spreadsheet and three-dimensionless the experiences of analysis.