A data warehouse is a subject-oriented, integrated, time variant, non-volatile collection of Data in support of the management's decision making process.

Central location where data from multiple databases is consolidated.
Different from organization's operational database

1. Subject oriented : Data categorized and stored based on business subjects rather than applications.
2. Integrated : Data on a subject is collected from various heterogeneous sources.
3. Time Variant : Data is stored as a series of snapshots representing a period of time. Unlike operational data where current value in time is stored, data warehouse data provides data from a historical perspective.
4. Non-volatile : Physically separate from organization's operational data.
   Data in the data warehouse is not typically updated or deleted.

Features	OLTP - on-line transaction processing	OLAP - on-line analytical processing
1. User	Client, IT professional	Knowledge worker
2. Data features	Current Highly - Detailed	Historical Summarized Integrated Consolidated
3. Function	Useful to run a business (day2day)	Useful to analyse a business (decision support)
4. Design	Application based	Subject based
5. Size	100MB to 1GB	100GB to 1TB
6. Unit of work	Simple and short query	Complex query
7. Number of records accessed	In tens	In millions
8. Example		

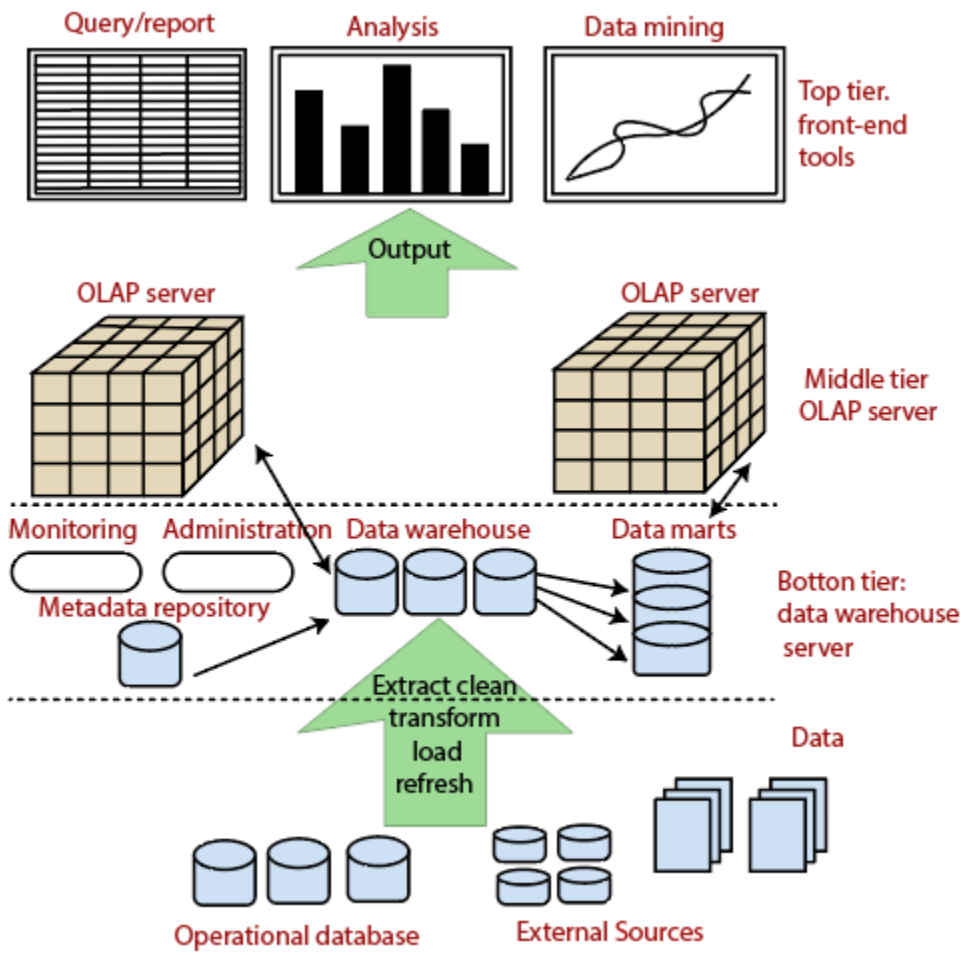
nformation Systems:- OLTP (DB) vs. OLAP (DWH)

OLTP Examples:

1. A supermarket server which records every single product purchased at that market.
2. A bank server which records every time a transaction is made for a particular account.
3. A railway reservation server which records the transactions of a passenger.

OLAP Examples:

1. Bank Manager wants to know how many customers are utilizing the ATM of his branch. Based on this he may take a call whether to continue with the ATM or relocate it.
2. An insurance company wants to know the number of policies each agent has sold. This will help in better performance management of agents.



Three-Tier Data Warehouse Architecture

ETL :
Extract
Clean
Transform
Load
Refresh

Data stored in a warehouse is multidimensional with measure and dimension attributes, viewing data as a data cube.

1. Fact table : Contain the numerical quantities of a business process and foreign keys of different dimensions.
2. Dimension table : Contains the logically related attributes of quantities stored in fact table
   Example :
   Book shop :
   1. Sales fact table : bid (fk), tid (fk), number of books
   2. Book dimension table : bid (pk), bauthor, name
   3. Transaction dimension table : tid(pk), payee

Schemas:

1. Star schema : Fact table in the middle connected to dimension tables
2. Snowflake schema : Star schema where dimensions are normalized to smaller dimension tables
3. Fact Constellation/Galaxy schema : Multiple fact tables share dimension tables

OLAP operations

1. Roll up / Drill up : summarize data by climbing up hierarchy or dimensionality reduction or data aggregation.
   Example :
   Location based on state : NY - Maine - Vancouver - Waterloo
   Location based on country : USA - Canada
   Dimension reduced

2. Drill down / Roll down : higher level to lower level data which is more detailed/specific
   Dimension increased
   Opp roll up

3. Slicing : Selection of 1 dimension with a specified criterion of that dimension (like time = Q1) of a data cube
4. Dicing : Selection of >=2 dimensions (location = “NYC” and time = “1800”)

5. Pivot : Visualization operation rotating data axes to view alternate data representation
6. Drill across : execute queries involving > 1 fact table
7. Drill through : ??

0-D : apex
n-D : base

Data mining :
Extraction of interesting, non-trivial, previously unknown, potentially useful, previously unknown patterns or knowledge from huge amounts of data.

It is a step in KDD.

KDD Knowledge discovery from databases: C I S T M P K

Database —Data cleaning & Integration—> Data warehouse —Data selection and transformation→ Task relevant data—> Data Mining —> Patterns —>Pattern evaluation—> Knowledge extraction

Step 1 : Data Cleaning
Removal of noisy and irrelevant data from collection.
Dealing with missing values, noisy data, data discrepancy and transformation.

Step 2 : Data Integration
Heterogeneous data from multiple sources are integrated with data migration, synchronization tools and ETL process combined into the data warehouse

Step 3 : Data selection
Data relevant to analysis is decided and retrieved from data collection. Data selection using NN, Decision trees, Naive bayes, clustering, regression,etc.

Step 4: Data transformation :
Transform data into appropriate form for the data mining technique by performing summary/aggregation. Steps for transforming data:

1. Data mapping : Mapping elements from source to destination to capture transformations
2. Code generation : Actual transformation program coding

Step 5 : Data mining:
Extract patterns from data

Step 6 : Pattern Evaluation : Identifying strictly increasing patterns representing knowledge based on given measures.

1. Interestingness score for each pattern is calculated.

Step 7 : Knowledge representation : technique utilizing data visualization tools to represent the data mining results
Reports, tables, classification and characterization rules

Business Intelligence :
Business environment factors :

1. Market :
2. Consumer Demands

3. Technology
4. Society

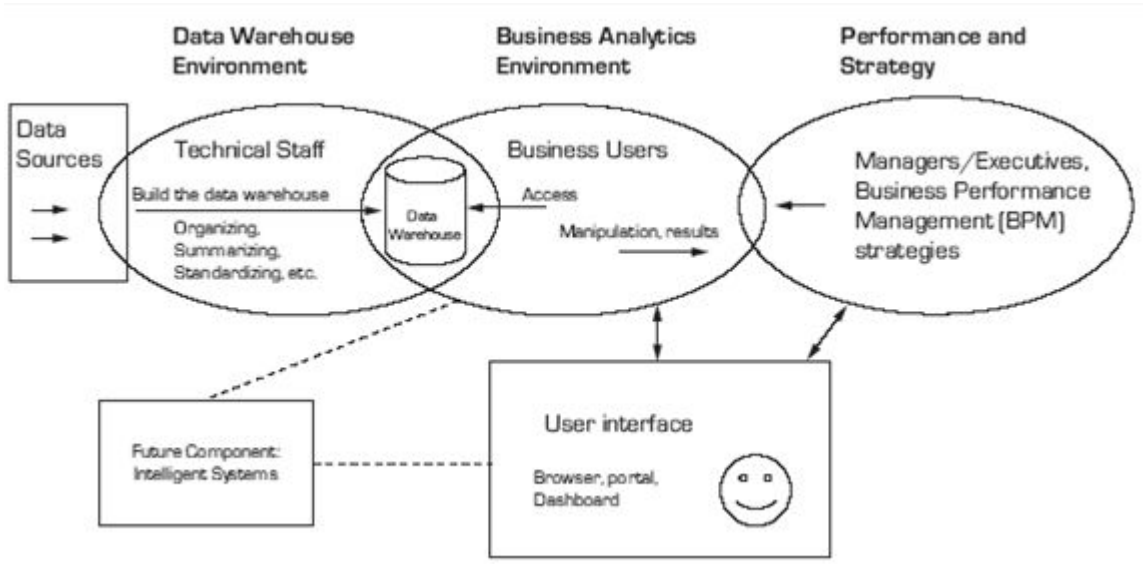
Business Intelligence is a term combining models, tools, architecture, applications, databases and allowing interactive data access, manipulation tools and giving business managers the ability to make well informed and better business decisions.

1. Data warehouse : subject oriented, time variant, integrated, non-volatile collection of data from the operational database and other resources helping in management's decision making process with the range from simple reporting to complex optimization

2. Business Analytics : Software tools to create reports with queries and rules. Data mining is an important aspect

3. BPM Business performance management : The framework for defining, implementing and managing business objectives by linking it with the factual data found for efficient tracking and optimize overall performance.
   Alert managers to opportunities and threats and empower to react with models and collaboration

4. User Interface :Access and easy manipulation through portals, browser, dashboard by end user



Unsupervised learning	Supervised learning
Algorithms trained using labelled data.	Algorithms trained using unlabelled data.

Less computational complexity	High computational complexity
Highly accurate	Low accuracy
Regression, classification, kNN, Decision forests	Clustering

Classification	Regression / Prediction
Determining the class of an element	Determining an unknown/missing values
Training set accuracy on accuracy to get element	Accuracy to estimate unknown value