What is Data Mining?

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data.
- Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

- **The knowledge discovery process is as an iterative sequence of the following steps:**

  1. Data cleaning (to remove noise and inconsistent data)

  2. Data integration (where multiple data sources may be combined)

  3. Data selection (where data relevant to the analysis task are retrieved from the database)

  4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

  5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

  6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on *interestingness measures*.

  7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

What Kinds of Data can be mined?

- **Database Data**
- **Data Warehouses:**

What is DW ?

- It is a relational DB that is designed for query and analysis rather than for transaction processing.

- It is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related

What are DW characteristics?

- Subject oriented
- Time variant
- Non – volatile
- Integrated
- Historical

OLAP vs OLTP

	Oltp	Olap
Abbreviation	Online transactional processing	Online Analytical processing
Support	Day to Day Transactions – Current Data	Historical data
Speed	Speed >	Speed <
Query	Simple Query Like (insert – Update - Delete)	Complex (Aggregation)
Effectiveness	Number of transaction in second	Response time
Usage	Data In	Information Out
Operations	Read , Write	Read Only
Joins	Tables and joins of a database are complex as they are normalized. High number of joins.	Table and joins are simple in a data warehouse because they are denormalized. Low Number of Joins.
Example	DB	DW
Size	Size <	Size >
Type	Structured	Structured , Semi-Structured , unstructured

Etl ?

Etl stands for Extraction, Transformation and Loading

- Extraction process is dedicated for extraction data from multiple sources for example flat file, excel file etc.

- Transformation Process after data has been Extracted some transformation are applied on it like removing duplicates , Trim , Removing recodes with Null IDS Values , Converting data types , Add Or Remove Columns , Reshaping Structure of the data .
- Loading Process is dedicated to load data after transformation to any destination you choose like DW , Flat file etc..