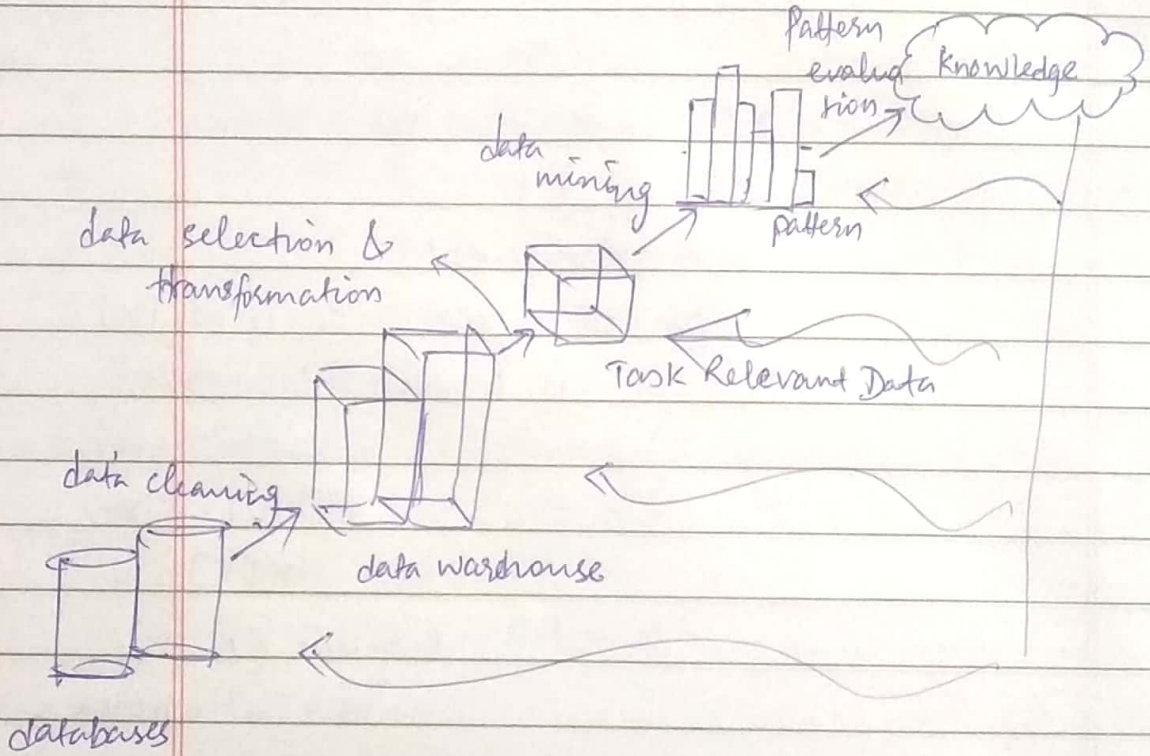
Data Mining is also known as KDD

MOD 1

DATA MINING

Data Mining is the process of extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amount of data.

Knowledge Discovery in Database (KDD) process –



~~Need of DM~~ DM on what data –

→ RDBMS, Transactional data
→ data streams & sensor data
→ time series data, sequence data
→ structure data, graphs, social networks
→ multimedia DB
→ text DB
→ www.

Date
Page 12

Applications of DM -

→ web page analysis - classification, clustering, pageRanks, etc
→ Collaborative analysis and recommender systems
→ Basket data analysis to targeted marketing
→ Biological & medical data analysis - classification, clustering, biological sequence analysis, biological network analysis.
→ DM & SE
→ From major dedicated data mining systems (SAS, MS SQL, etc) to invisible data mining..

Issues -

→ Mining methodology -
- various & new kinds of knowledge
- DM - an interdisciplinary effort
- Handling noise, uncertainty and incompleteness of data.
- Pattern evaluation & pattern - or constraint guided mining.

→ User Interaction -
- Incorporation of background knowledge
- Presentation & visualization of DM results

→ Efficiency & Scalability - of DM algos, & parallel, distributed, stream & incremental mining methods
→ Diversity of data types - Handling complex types of data & mining dynamic, networked & global data repositories
→ Data mining & society - Social impacts of DM, privacy preserving DM & invisible DM.

DM : Confluence of Multiple Disciplines - (Integration of DM & DB)

→ ML → Pattern Recognition → Stats → Apps → Algos
→ Visualization → Database Technology
→ High performance Computing.

DM Functions –

1. Generalization - Info integration & data warehouse construction that includes data cleaning, transformation, integration & multidimensional data model such as generalize, summarize and contrast data characteristics.

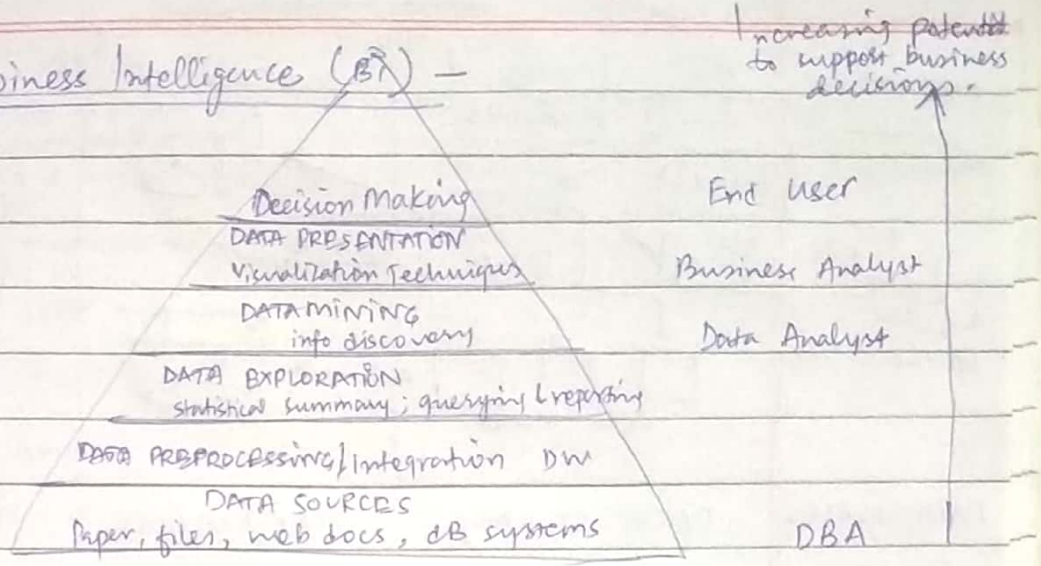
2. Association & Correlation Analysis – defining & identifying frequent patterns through association & correlation & how to use this pattern for classification, clustering & other appli

3. Classification – Construct models/fn based on some training using typical methods like decision trees, neural networks, rule based classification, pattern-based class, logistics regression, etc.

4. Cluster Analysis – Group data to form new categories based on principle: maximizing intra-class similarity & minimizing interclass similarity.

5. Outlier Analysis – Outlier is a data object that does not comply with general behavior of data and they are found be a by product of clustering or regression analysis.

Business Intelligence (BI) –



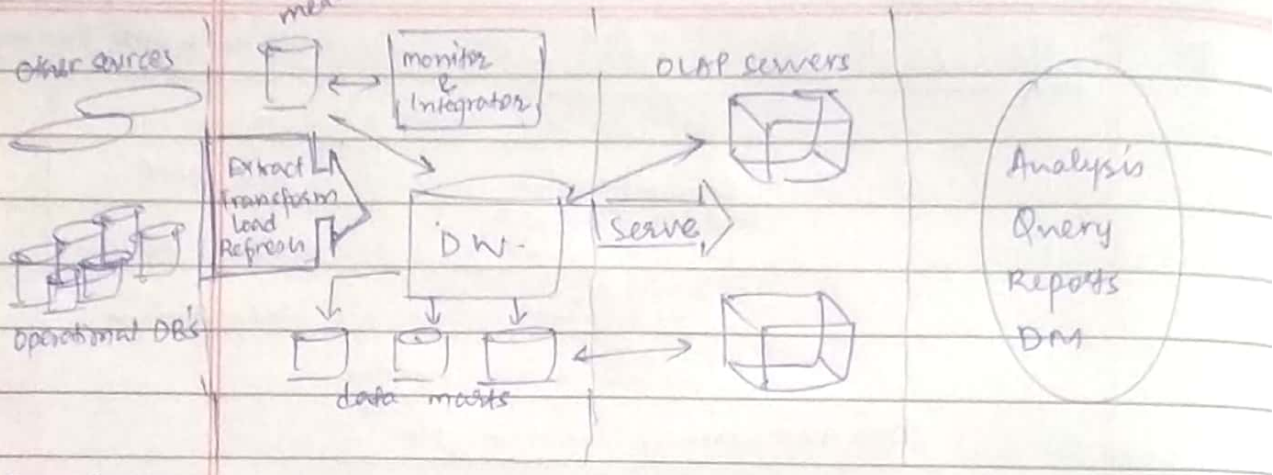
Data Warehouse (DW) – it is a subject oriented, integrated, time-variant & non volatile collection of data in support of management's decision making process. the process of constructing data warehouse is data warehousing.

USAGE
1. Info process
2. Analytical processing
3. DM

→ Types –

1. Subject oriented – organized around major subjects, such as product, cust or sales. Provides a simple & concise view around a particular issue by excluding data that are not useful.
2. Integrated – constructed by integrating multiple, heterogeneous data source, also applying data cleaning & integration techniques.
3. Time Variant – provides info from a historical perspective
   Data contains an element of time, implicitly or explicitly.
4. Non-volatile – A physically seperate store of data, transformed from operational environment where no updates can be made. Only initial loading of data & access of data is allowed.

ARCHITECTURE



ETL (Extraction, Transform & Loading) –

- Data Extraction – get data from multiple, heterogenous, & external sources
- Data Cleaning – detect errors in data & rectify them when possible.
- Data Transformation – convert data from legacy or host format to warehouse format
- Load – sort, summarize, consolidate, compute views, check integrity & build partitions.
- Refresh – propagate the updates from data sources to warehouse.

Relation b/w Bi & DW–

Bi tells us what happened, or is happening right now in your business – it describes the situation to you, and Bi make use of data stored in DW and lets you apply chosen metrics to potentially huge, unstructured used data set, DM, OLAP, and reporting as well as business performance monitoring, predictive & prescriptive analytics.

OLAP – it is an online system that reports to multidimensional analytical queries like forecasting, reporting, etc

OLTP – is a system that manages transaction-oriented applications on internet like ATM.

	OLTP	OLAP
1.	its an online transactional system & manages db modifications.	its an online data retrieving & data analysis system
2.	main focus is insert, update, delete info from db.	main focus is extract data for analysing that helps in decision making.
3.	OLTP & its transactions are original source of data.	diff OLTP's db becomes the source of data.
4.	OLTP has short transactions.	OLAP has long transactions
5.	Processing time of transactions is ↓	↑
6.	simple queries	complex queries
7.	tables in OLTP db are normalized	not normalized.
8.	db size - 100 MB - GB	100 GB - TB
9.	users - IT professionals	knowledge workers
10.	db design - application oriented	subject oriented.

Modelling of DW (Schemas) -

1. Star - fact table in the middle connected to a set of dimension tables
2. Snowflake - a refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables.
3. Fact constellations (FC) - multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or FC.

OLAP models / operations -

1. Rollup / drill up - summarize data by climbing up hierarchy or by dimension reduction
2. Roll down / drill down - reverse of roll up : converting data from higher level to lower level or introducing new dimensions.
3. Slice & dice - One or more dimension is selected and a new cube is created
4. Pivot / Rotate - you rotate the data axes to provide a substitute presentation of data.

Date Page 7

Types of OLAP -

1. Multidimensional OLAP (MOLAP) - it implements operation in multidimensional data
2. Relational OLAP (ROLAP) - its an extended RDBMS along with multidimensional data mapping to perform the standard relational operation.
3. Hybrid OLAP (HOLAP) - in this, the aggregated tools are stored in multidimensional dB while detailed data is stored in RDB. This offers efficiency of both ROLAP & MOLAP.
4. Web OLAP (WOLAP) - its in which OLAP system is accessible via web browser.
5. Mobile OLAP - it's one in which OLAP system is accessible via mobile phones.

#	ROLAP	MOLAP
1.	Relational OLAP	multidimensional online Analytical processing
2.	data is stored and fetched from main DW.	data is stored and fetched from the proprietary dB MDDBs.
3.	data is stored in form of relational tables.	data is stored in large multidimensional array made of data cubes.
4.	large data volumes	limited data summaries is kept in MDDBs.
5.	Slow access	Faster access
6.	& creates a multidimensional view of data dynamically.	already stores the static multidimensional view of data in MDDBs.