dap # 2.

* Major steps in ETL process.

→ The process of extracting data from source systems & bringing it into the data warehouse is commonly called ETL.
  → Extraction
  → transformation
  → loading

→ It is a process in which ETL tool extracts the data from various data soure systems, transforms it in the staging area & the finally, loads it into the Data W. system

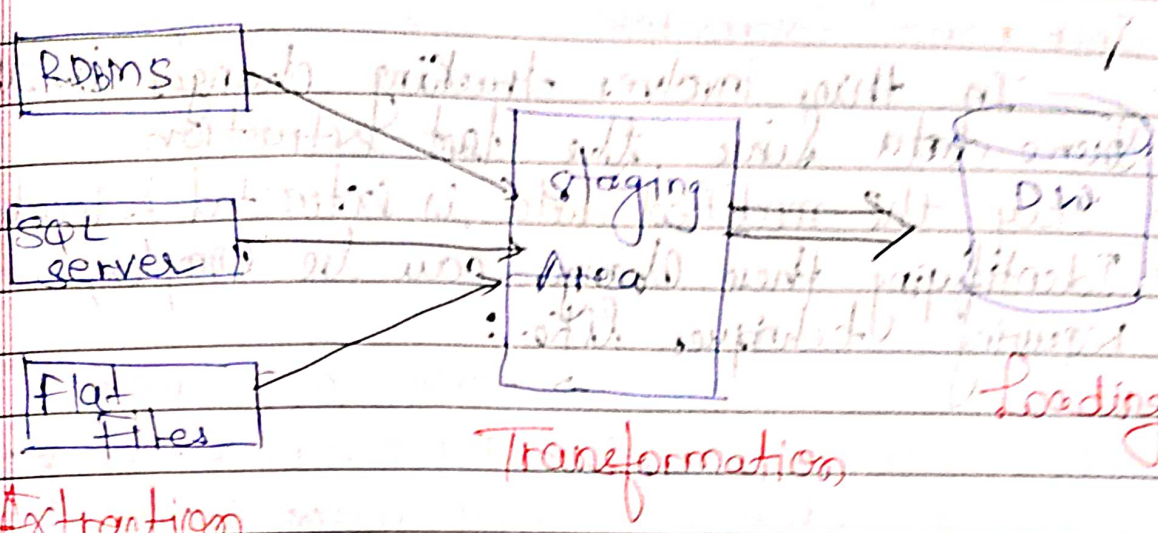
→ ETL process requires active inputs from various stakeholders, including → developers
  → analysts
  → testers
  → top executives.

→ ~~Business~~ DW ~~dep~~ technique needs to change with business changes

→ ETL is a recurring methods (daily, weekly, monthly) of a data Warehouse system & needs to be agile, automated, & well documented

ETL tools: Sybase
Oracle Warehouse builder
clover ETL
Mark Logic

Steps.



Extraction:

→ First step of ETL process is extraction, In this step, data from various source systems is extracted which can be in various formats like rational DB No, SQL, XML & flat files. into the staging Area

→ It is imp to extract the data from various systems & store it into the staging area first & not directly into the DW as it may be in corrupted

Data Extraction Techniques

(Logical) ① full extraction

This data extraction method involves taking all data from a source system without needing additional logical information. The data is provided as-in & there's no tracking of changes in the

Eg: source system.
Exporting an entire table as a flat file.

Incremental Extraction:
→ In this, involves tracking changes in the source data since the last extraction.
→ Only the modified data is extracted & loaded
→ Identifying these changes can be complex, requiring techniques like:
→ timestamp comparison or maintaing a change table in source system

Physical extraction

Online Extraction:
In this process, extraction process directly connects to the source system & extracts the source data.

Offline extraction:
data is not extracted directly from the source system. but is staged explicitly outside the original source system.

Common structure in offline extraction
① flat file : Generic format
② dump file : Database specific file

Transformation:

The second step of the ETL process is transformation a set of rules, or functions are applied on extracted data to convert it into single, standard format.

Involves foll tasks

① Filtering:

Involves selecting & loading only specific attributes or data elements into the data warehouse.
→ focus on relevant info for analysis.

② Cleaning:

Addresses data quality issues by handling NULL values & inconsistencies.
Eg: mapping variations like "USA", "United States" & "America" into a standarized format like "USA"

③ Joining:

Combines data from multiple attributes or tables into a single, unified dataset.

④ Splitting

Divides a single attribute into multiple attributes based on defined rules.
useful when piece of info needs to be broken down into its constituent parts

⑤ Sorting

Arranges data tuples based on specified attributes, typically a key attribute

Data Transformation Techniques

① Data Smoothing:
→ This method is used for removing the noise from a dataset. Noise is referred to as the distorted & meaningless data within a dataset.
→ Method: Utilizes algo's to highlight significant features & detect small changes revealing special patterns.

② Data Aggregation:
→ Collects & stores data from various sources in a unified format.
→ Data is analyzed, summarized & presented in reports, facilitating the gathering of extensive info about specific data clusters.

③ Discretization:
Converts continuous data into intervals substituting continuous attribute values with interval labels for easier study & analysis.

④ Generalization:
Transforms low level data attributes into high level attributes using concept hierarchies.
Eg: Age data (20, 30)
transformed into categorical values (young, old)

⑤ Attribute Construction:
Creates a new attributes from an existing set
Eg: Employee dataset attributes (Name, ID, address) used to construct a dataset containing employees who joined in 2019, enhancing mining efficiency.

Normalization

- Data pre-processing
- Transforms data to fall within a specified range, addressing challenges in data modelling & mining posted by attributes on different scales.

Loading:

→ The third & final step of the ETL process is loading. In this step, the transformed data is finally loaded into the DW.

→ Sometimes the data is updated by loading into the DW very frequently, sometimes it is after longer but regular intervals.

→ Rate & period of loading → depends on Requirements & varies system to system

Loading can be carried in 2 ways.

① Refresh:

DW data is completely rewritten. This means that older file is replaced.
Refresh is usually used in combination with the static extraction to populate a data warehouse initially.

② Update

Only those changes applied to source information are added to DW.
An update is typically carried out without deleting or modifying pre existing data.

~~OLAP Vs OLTP~~

OLAP Vs OLTP:

Can divide IT systems as transactional (OLTP) & analytical (OLAP) In general, we can assume that

OLTP ⟶ provides source data to DW
OLAP ⟶ helps to analyze it.

OLTP (Online Transaction processing)
- is characterized by large No of short on-line transactions (INSERT, UPDATE, DELETE)
- → Puts on very fast query processing, maintaining data integrity in multi-access environments & an effectiveness measured by No. of transactions per second.
- → mostly there is detailed & current data
- → Schema used — Store transactional DB is the Entity model (3NF)

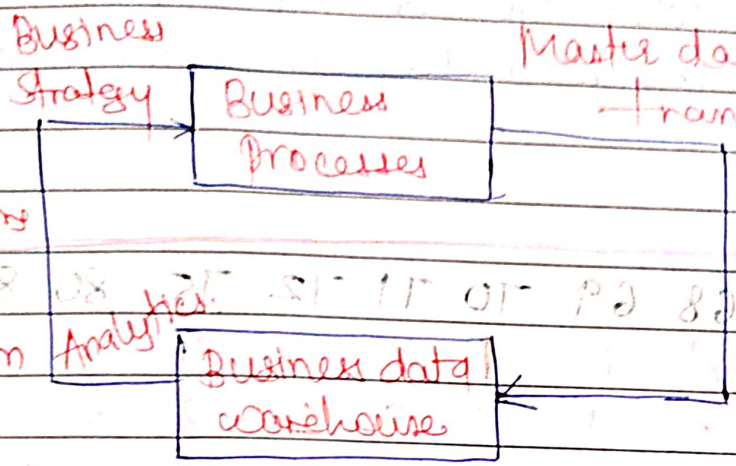
OLAP (Online Analytical processing):
- → Relatively low volume of transactions
- → Query often very complex & involves aggregations
- → OLAP systems, response time is an effectiveness measure
- → OLAP DB's is used by Data Mining
- → In OLAP DB there is aggregated historical data stored in Multidimensional schemas (usually Star)
- → divided into One or more cubes.

cubes

→ designed in such a way that creating & viewing reports become easy.

→ OLAP cube is the data structure optimized for very quick data analysis.

→ OLAP cube is also called hypercubes.



OLAP operations :-

→ In multidimensional model, data are organized into multiple dimensions, & each dimension contains multiple level of abstraction defined by Concept hierarchies.

→ This organization provides users with the flexibility to view data from different perspectives.

→ OLAP provides a user-friendly environment for interactive data Analysis.

→ A number of OLAP data cube operations exist to materialize different views of data, allowing interactive querying & analysis of the data.

1) Roll-up : It is used to climb up the concept hierarchy.

The roll up operation (i.e. drill down operation) performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension or by climbing down a concept hierarchy, ie dimension reduction.

Eg:

temp	64	65	68	69	70	71	72	75	80	81	83	85
week 1	1	0	1	0	1	0	0	0	0	0	1	0
week 2	0	0	0	1	0	0	1	2	0	1	0	0

Here, we set up levels.

hot (80-85), mild (70-75), cool (64-69)

→ We have to group up the columns & add up the values according to the concept hierarchies. This is called roll-up.

Temp	cool	mild	hot
week 1	2	1	1
week 2	2	1	1

DRILL DOWN

→ The drill down operation (also called roll-down) is the reverse operation of roll-up.
→ Drill down is like zooming in on the data cube.
→ Navigates from less details to more detailed data
→ Drill down can be performed by either
• Stepping down a concept hierarchy for a dimension or adding additional dimensions.
→ Bcz a drill-down adds more details to the given data, it can be performed by adding a new dimension to a cube.

Eg:

Temp	cold	mild	hot	
Day 1	0	0	1	
Day 2	0	0	1	
Day 3	0	0	1	
Day 4	0	1	0	
Day 5	1	0	0	
Day 6	0	0	1	
Day 7	1	0	0	
Day 8	0	0	1	
Day 9	1	0	0	
Day 10	0	1	0	
Day 11	0	1	0	
Day 12	0	1	0	
Day 13	0	0	1	
Day 14	0	0	1	

week 1
week 2

SLICE

→ A slice is a subset of the cube corresponding to a single value for one or more members of the dimensions.

→ For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a 2-D site.

→ So, the slice operations perform a selection on one dimension of the given cube, thus resulting in a sub-cube.

→

If we selection,
temperature = cool

Temperature	cool
Day 1	0
Day 2	0
Day 3	0
Day 4	0
Day 5	1
Day 6	1
Day 7	1
Day 8	1
Day 9	1
Day 11	0
Day 12	0
Day 13	0
Day 14	0.

DICE

→ The dice operation describes a sub-cube by operating a selection on two or more dimension.

For Eg:

Selection (time = day 3)
OR
(time = day 4) AND
(temperature = cool OR temp = hot)

to the original cube we get the foll. sub-cube (still two-dimensional)

temp	Cool	hot
Day 3	0	1
Day 4	0	0

PIVOT (Rotation)

→ Pivot is a visualization operation which rotates the data axes in view to provide an alternative presentation of the data.

→ Contains swapping the rows & columns or moving one of the row dimensions into the column dimensions.

ROLAP (Rational)

OLAP SERVERS

MOLAP (Multidimensional)

① ROLAP

- Rational on-line Analytical Processing (ROLAP) work mainly for data that resides in a Relational DB, where the base data & dimension tables are stored as rational tables.
- ROLAP servers are placed between the relational back-end server & client front-end tools.

→ uses RDMS to store & manage warehouse data.

→ OLAP middleware to support missing pieces.

Eg: DSS Server of Microstrategy.

Advantages

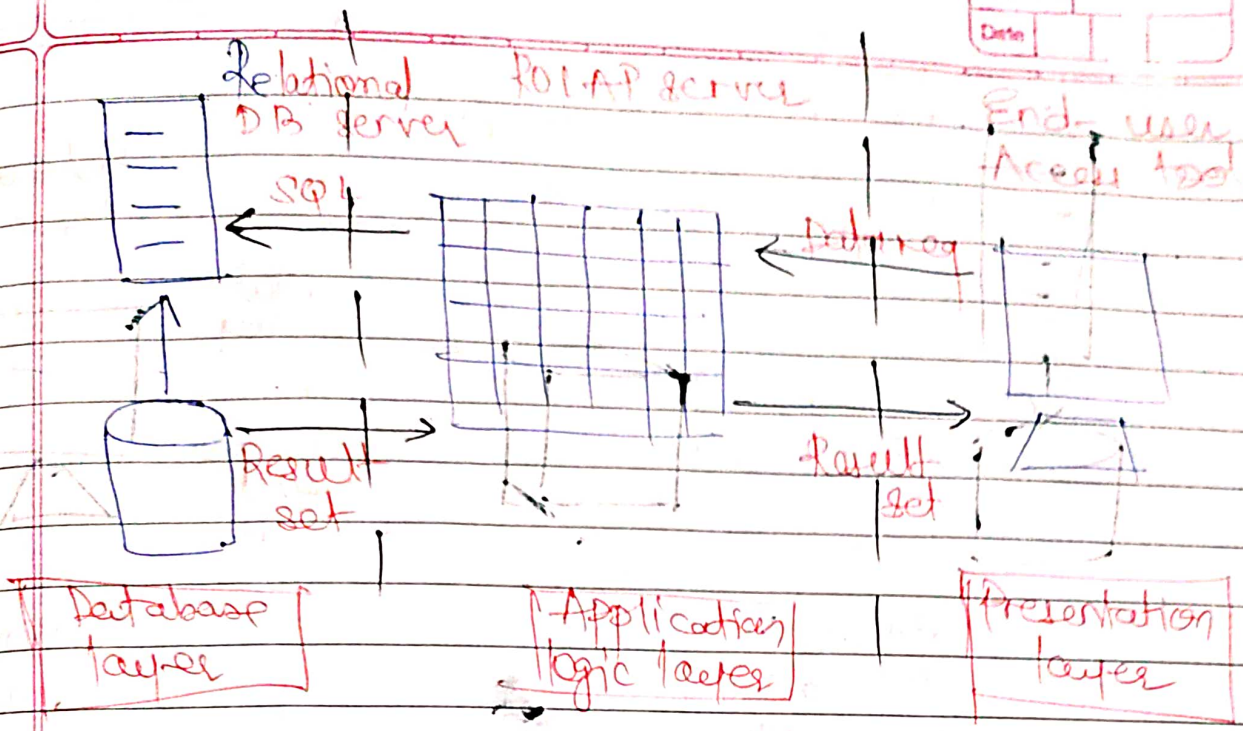
→ ROLAP can handle large amounts of data.

→ Can be used with data Warehouse & OLTP systems

Disadvantages

→ limited by SQL functionalities

→ Hard to maintain aggregate tables.



MOLAP

→ MOLAP supports multidimensional views of data through array-based multidimensional storage engines

→ With multidimensional data stores, the storage utilization may be low if the data set is sparse.

→ Eg: Oracle Essbase.

Advantage:

→ optimal for slice & dice oper

→ performs better than ROLAP when data is dense

→ Can perform complex calculations.

Disadvantage

→ Difficult to change dimension without re-aggregation

→ MOLAP can handle limited amount of data