Date..... Page.....

Unit-5

→ Aggregation in data mining is the process of finding, collecting and presenting the data in summarised format to perform statistical analysis of business schemes.

→ Aggregated data helps in solving rational problems which reduce time in solving queries from dataset.

→ When data is collected from various sources it is necessary to get crucial information.

→ Data aggregator - are system that collect data from numerous sources then process and repackages them to useful data package

help in query and delivery process when customes requests.

working of data aggregator -

- collecting - collect from various sources.
- Processing - Apply various ML & AI algo
- Presentation - show summarised data to user.

Collection of data → Processing of Data → Presenting of Data

→ Can be done manually or automatically

Types of data aggregation

1) Time - provide data point for single resource for defined time.
2) Spatial - provide data point for group of resources for defined time.

- Time interval for data aggregation - Reporting period - One day
  Spatial { Granularity period - month.
  Polling period - Every 7 min etc.
- Application - Ecommerce,
  travel
  business analysis.

→ Historical information - It is used to build model using this. new response can be generated based on previous data. Ex - credit card fraud.

Query facility - It is a tool that allow client to retrieve data from warehouse. by selecting field and some cons like query.

Data mining query language can be used to specify mining task. It.

+ OLAP function and tool

OLAP stands for on-line Analytical processing is a tool which help analyst, manager to gain insight into information in fast, consistent way.

- Used to analyze multidimensional data.

Types -

1) • Relational OLAP - ROLAP server are placed between relational backend server and client frontend tool.

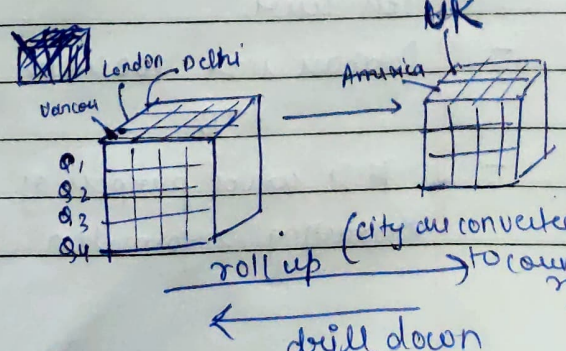
To store and manage data ROLAP use extended RDBMS

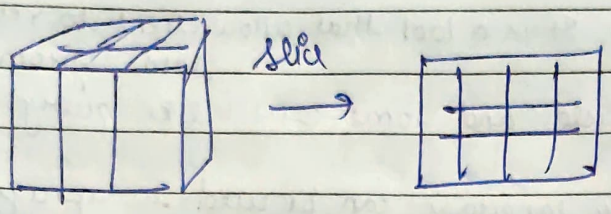
2) Multidimensional OLAP - MOLAP use array based multi-dimensional engine for multidimensional view of data. Storage utilization is low if data is sparse, therefore MOLAP server used ~~use~~ two level of data storage.

3) Hybrid OLAP - combination of both ROLAP and MOLAP offer high scalability of ROLAP and fast computation of MOLAP, can store large data

4) Specialized query server - It provide advanced query language and query processing support SQL queries over star and snowflake schema.

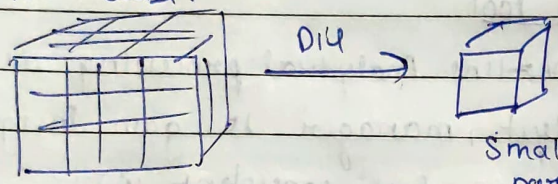
OLAP operations - Roll up
- drill-down
- slice
- dice
- Pivot





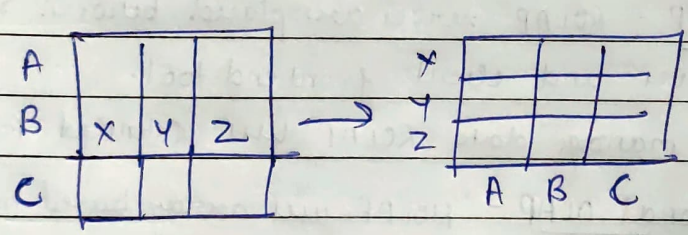
Selects some part of data from data cube.

→ Dice Selects two or more dimension and provide a new sub-cube.



Small cube means some part of information.

→ Pivot – It is also known as rotation, it rotate data axes.



rotated

OLAP	OLTP
• Involve historical processing of data	Involve data processing (day to day) like bank servers.
– Used by analysts, manager	Used by clerk and professional
– Useful in analyzing	Useful in running business.
– Based on star, snowflake and fact constellation schema	Based on ER-model
– Less user	more user
– Database size 100GB to 1TB	Database size 100MB to 1GB
– It served purpose of extracting information	It serve purpose of insert update, delete.

- data mining interface -
data mining interface provide medium that allows user to communicate with data mining process. It basically user interface where we see chart, graphs, to visualise the data to see result of mining process.

- security - The objective of data mining is to make large amount of data easily accessible to user without compromising security and data integrity.
so it is required to have security feature to affect performance of the data warehouse.
security feature are difficult to add when data warehouse has gone live.
security offers following feature - user access
Data loading
Data movement
Query management

- Backup - A backup is copy of data from database.
Backup can be physical and logical backup.
Physical backup include physical datafiles, control file.
logical backup contain logical data (ex - table or procedures).

- Recovery - Recovery is restoring data which could be lost due to any hardware failure or any other issue.

- Tuning data warehouse - A data warehouse keeps evolving and it is unpredictable what query user is going to post in future.

Date..... Page.....

- difficulties in data warehouse tuning
  - data warehouse is dynamic
  - difficult to predict query.
  - Business requirement keep changing.
  - data upload on warehouse also changes with time.

- Data load tuning - if there is delay in ~~transferring data then system get affected so data load is~~ important
- Integrity check - Integrity system avoid performance degrade of data load.

Testing - Testing is important for data warehouse system for data validation and to make them work correctly and efficiently.

- Unit testing - Each component module is tested
- Integration testing - In this testing all module are combined and then tested. done to detect fault in integrated modules.
- System testing - form of testing that validates and test whole data warehouse application.