Module 1.3:
Data Warehousing Fundamentals

- Pradnya Bhangale

Content

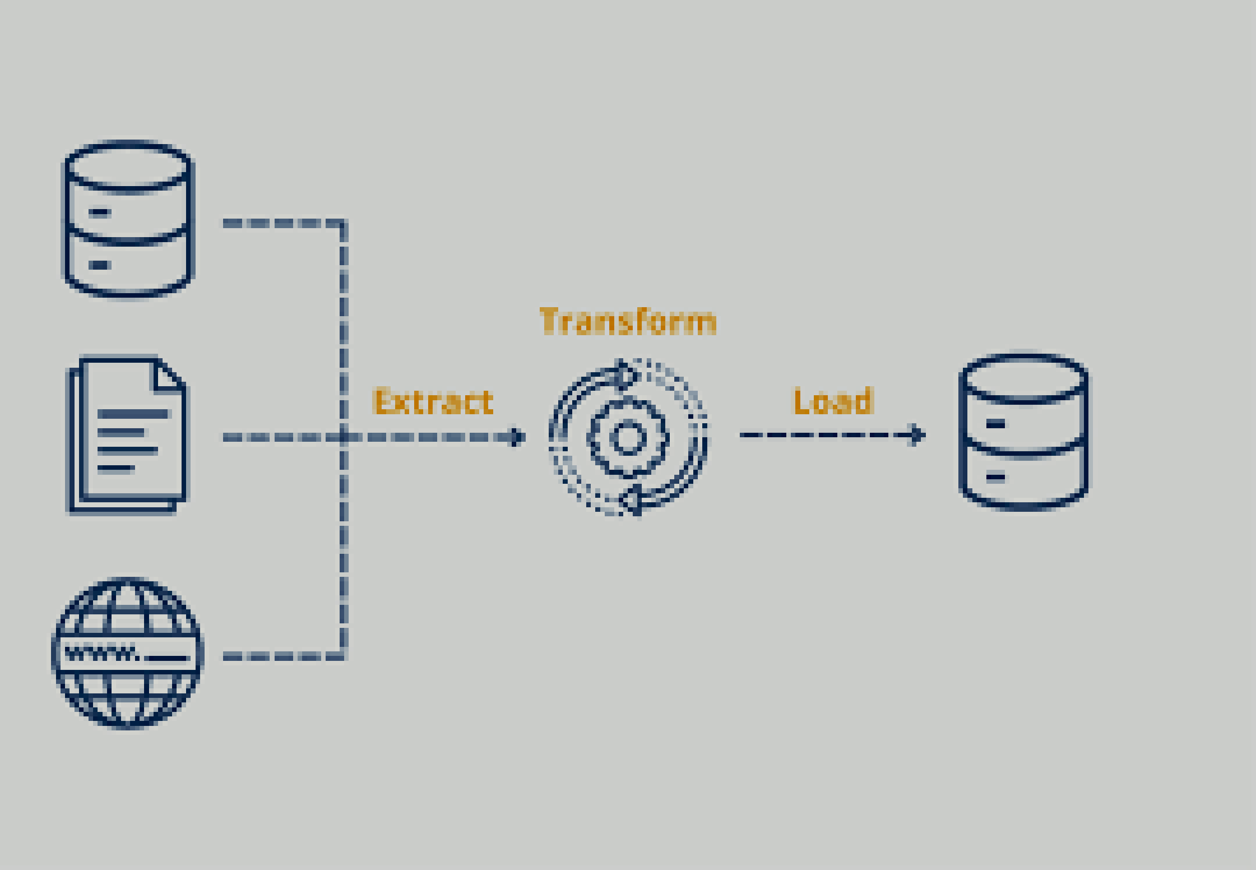
- **Major steps in ETL process**
- **OLTP vs. OLAP**
- **OLAP operations**
  - Slice
  - Dice
  - Rollup
  - Drilldown
  - Pivot
- **OLAP Server**
- **Application of OLAP**

Major steps in ETL Process

- The process of extracting data from source systems and bringing it into the data warehouse is commonly called **ETL** which stands for **E**xtraction, **T**ransformation, and **L**oading.
- It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.
- The ETL process requires active inputs from various stakeholders, including developers, analysts, testers, top executives and is technically challenging.

Major steps in ETL Process

- **ETL Tools**: Most commonly used ETL tools are Sybase, Oracle Warehouse builder, Clover ETL and MarkLogic.



1. Extraction

- The first step of the ETL process is extraction
- In this step, data from various source systems is extracted which can be in various formats like relational databases, NoSQL, XML and flat files into the staging area (extracted data is in various formats and can be corrupted also)
- Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult
- Therefore, this is one of the most important steps of ETL process

1. Extraction

- **Data Extraction Techniques**

There are two types of data warehouse data extraction techniques: **Logical and Physical Extraction methods**

Logical Extraction:

- Logical Extraction method in-turn has two methods:

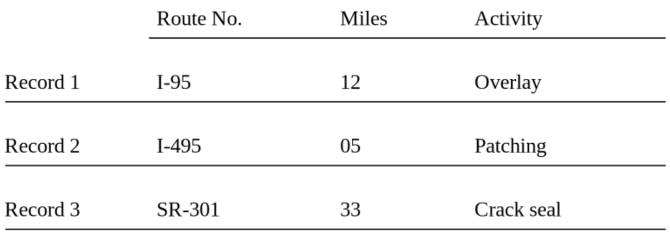
1. **Full Extraction**
2. **Incremental Extraction**

1. Extraction

1. Full Extraction

- In this method, data is completely extracted from the source system. The source data will be provided as it is and no additional logical information is necessary on the source system
- For example, exporting complete data in the form of flat file

Flat File Model



	Route No.	Miles	Activity
Record 1	I-95	12	Overlay
Record 2	I-495	05	Patching
Record 3	SR-301	33	Crack seal

2. Incremental Extraction:

- In incremental extraction, the changes in source data need to be tracked since the last successful extraction.
- Only these changes in data will be extracted and then loaded, identifying the last changed data itself.
- This is the complex process and involve many logics.

1. Extraction

3. **Partial Extraction- with update notification**

- The easiest way to extract data from a source system is to have that system issue a notification when a record has been changed.
- Most databases provide a mechanism for this so that they can support database replication (change data capture or binary logs) etc.

1. Extraction

Physical Extraction

- Physical extraction has two methods: Online and Offline extraction

(i) Online Extraction

- In this process, extraction process directly connects to the source system and extract the source data.

(ii) Offline Extraction

- Here the data is not extracted directly from the source, but instead it’s taken from another external area which keeps the copy of source
- The external area can be Flat files, or some dump files in a specific format.

2. Transformation

- The second step of the ETL process is transform
- In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.
- **It may involve following processes/tasks:**

1. **Filtering** – loading only certain attributes into the data warehouse.
2. **Cleaning** – filling up the NULL values with some default values, mapping U.S.A, United States and America into USA, etc.
3. **Joining-** joining multiple attributes into one.
4. **Splitting** -splitting a single attribute into multiple attributes.
5. **Sorting** - sorting tuples on the basis of some attribute (generally key-attribute).

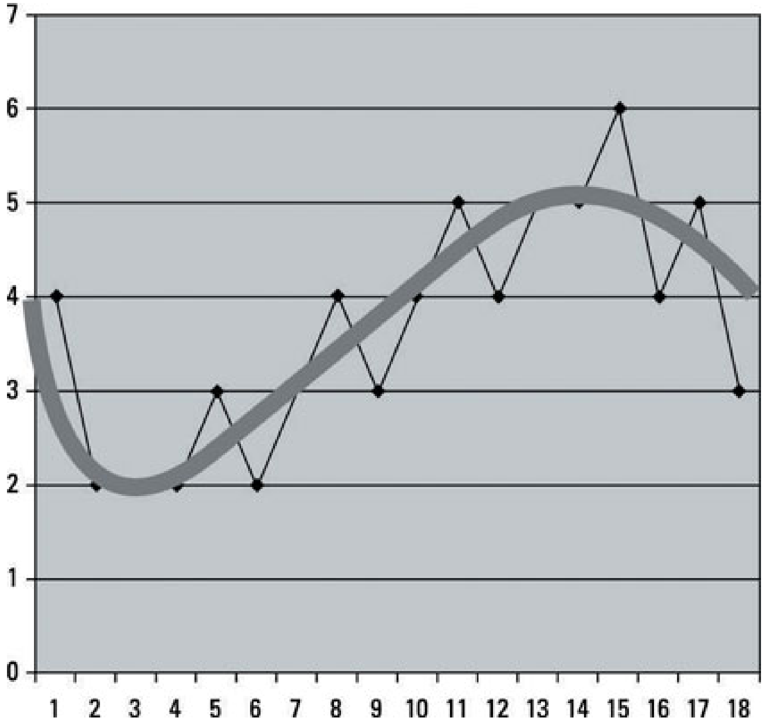
2. Transformation

Data Transformation Techniques:

- **Data Smoothing**
- **Data Aggregation**
- **Discretization**
- **Generalization**
- **Attribute Construction**
- **Normalization**

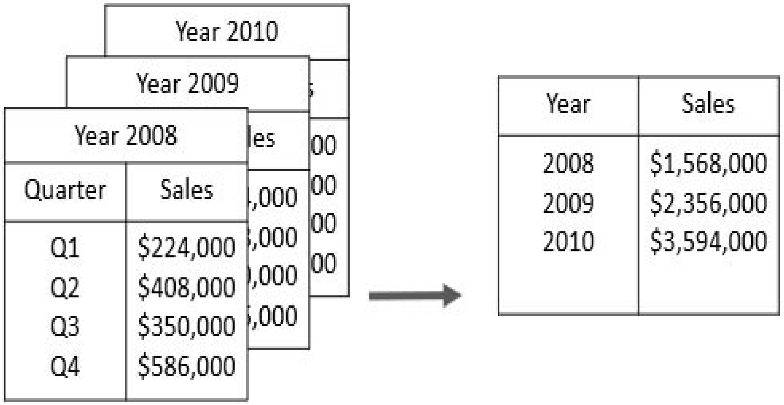
Data Transformation Techniques: **Data Smoothing**

- This method is used to *remove noise from the dataset*
- **Noise** is referred to as distorted and meaningless data within a dataset
- This allows important patterns to more clearly stand out.
- Data smoothing can be used to help predict trends



Data Transformation Techniques: **Data Aggregation**

- Aggregation is the process of collecting data from a variety of sources and a storing it in a single format.
- Here, data is collected, stored, analyzed and presented in a report or summary format.
- It helps in gathering more information about , particular data cluster.
- The method helps in collecting vast amounts of data.


Year 2010
Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000

Aggregated Data

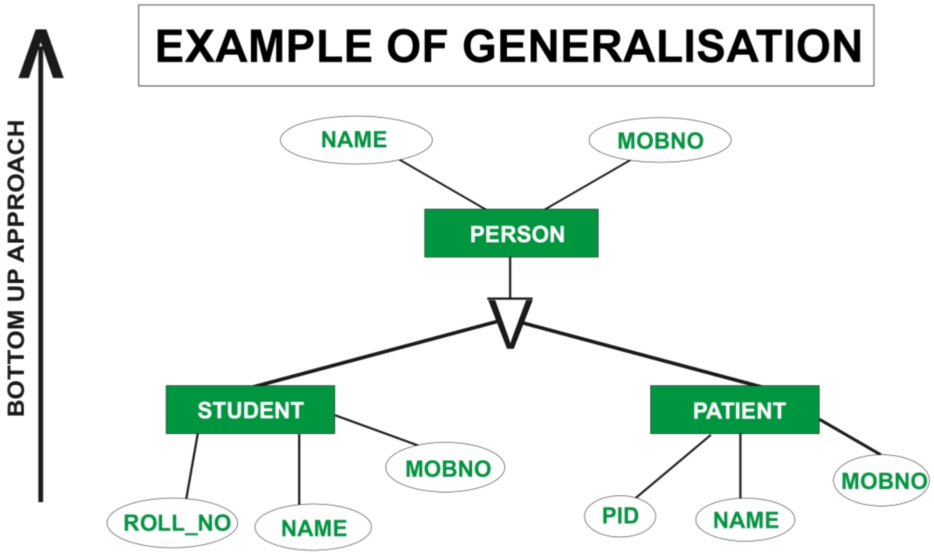
Data Transformation Techniques: **Discretization**

- This is a process of converting continuous data into a set of data intervals.
- Continuous attribute values are substituted by small interval labels
- This makes the data easier to study and analyze.

This image cannot currently be displayed.

Data Transformation Techniques: **Generalization**

- In this process, low-level data attributes are transformed into high-level data attributes using concept hierarchies.
- For example, age data can be in the form of (20, 30) in a dataset. It transformed into a higher conceptual level into categorical value (young, old).



Data Transformation Techniques: **Attribute Construction**

- In the attribute Construction method, new attributes are created from an existing set of attributes.
- For example, in a dataset of employee information, the attributes can be employee_name, employee_ID, date_of_joining, and address.
  - These attributes can be used to construct another dataset that contains information about the employees who have joined in the year 2019 only.
- Example 2– we may wish to add the attribute area based on the Data Transformation attributes height and width

Data Transformation Techniques: **Normalization**

- Also called data pre-processing, this is one of the **crucial techniques for data transformation in data mining.**
- Here, the data is transformed so that it falls under a given range.
- When attributes are on different ranges or scales, data modeling and mining can be difficult.
- Normalization helps in applying data mining algorithms and extracting data faster

3. Loading

- The third and final step of the ETL process is loading
- In this step, the transformed data is finally loaded into the data warehouse.
- Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals.
- The rate and period of loading solely depends on the requirements and varies from system to system
- Loading can be carried in two way:

I. **Refresh**

II. **Update**

3. Loading

1. **Refresh** :

- Data Warehouse data is completely rewritten i.e. older file is replaced.
- Refresh is usually used in combination with full extraction to populate a data warehouse initially.

2. **Update** :

- Only those changes applied to source information are added to the Data Warehouse.
- An update is typically carried out without deleting or modifying pre-existing data.
- This method is used in combination with incremental extraction to update data warehouses regularly.

OLTP vs. OLAP

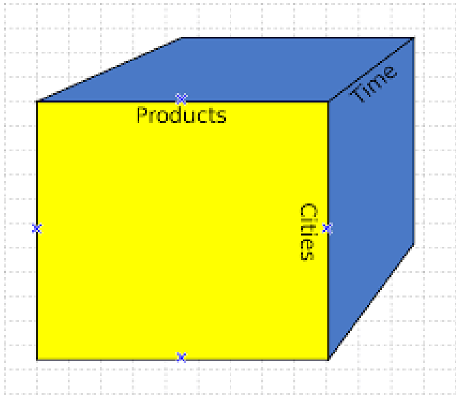
OLTP

- OLTP (On-line Transaction Processing) is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE).
- The main emphasis for OLTP systems is put on
  - very fast query processing
  - maintaining data integrity in multi-access environments
  - an effectiveness measured by number of transactions per second
- In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF)

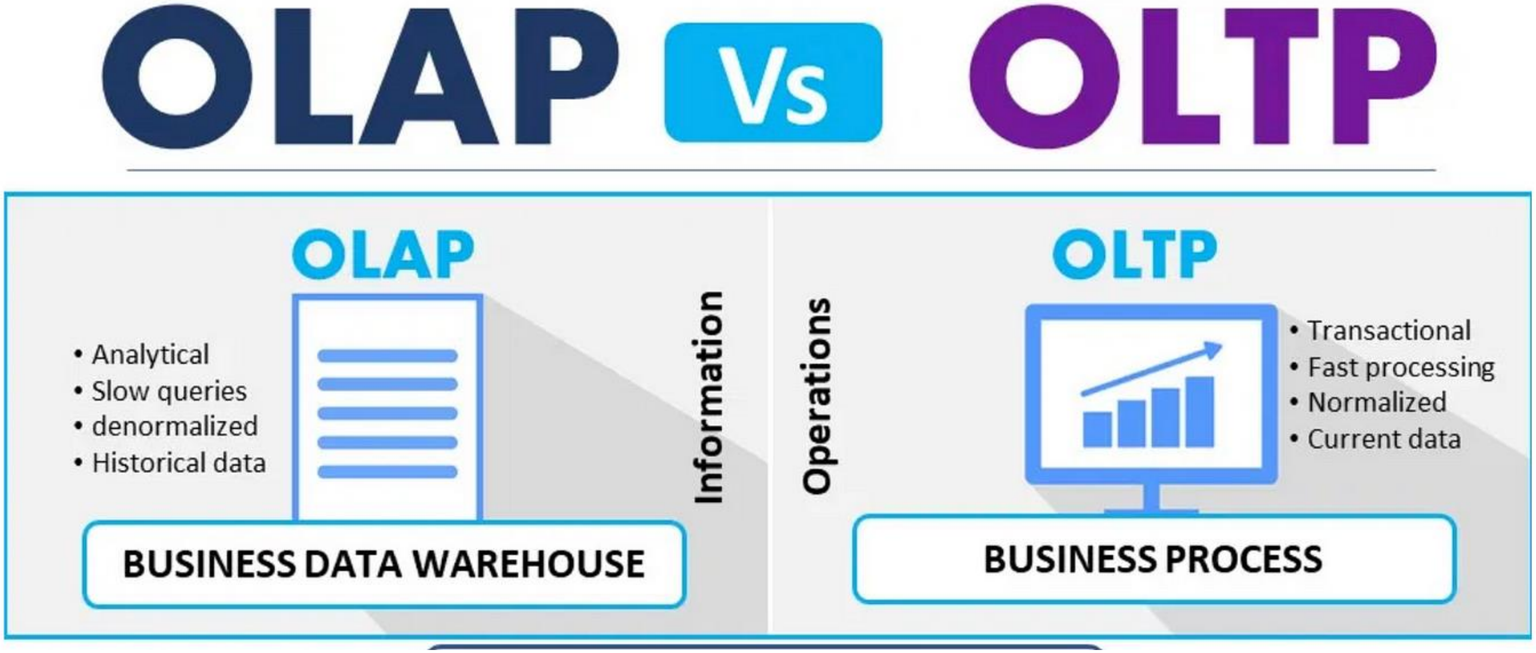
OLTP vs. OLAP

OLAP

- OLAP (On-line Analytical Processing) characterized by relatively low volume of transactions.
- OLAP applications are widely used by Data Mining techniques
- In OLAP database aggregated, historical data, stored in multi-dimensional schemas (usually star schema).
- For example, a bank storing years of historical records check deposits could use an OLAP database to provide reporting to business users. OLAP databases are divided into one or more cubes.
- The OLAP cube(hypercube) is a data structure optimized for very quick data analysis which consists of numeric facts called measures which are categorized by dimensions.



OLTP vs. OLAP


OLAP
Vs
OLTP
OLAP
• Analytical
• Slow queries
• denormalized
• Historical data
Information
BUSINESS DATA WAREHOUSE
Operations
OLTP
• Transactional
• Fast processing
• Normalized
• Current data
BUSINESS PROCESS

OLTP vs. OLAP

Parameters	OLTP	OLAP
Process	It is an online transactional system which manages database modification	It is an online analysis and data retrieving process
Characteristic	Large number of short online transactions	Large volume of data
Method	Uses traditional database	Uses data warehouse
Table	Tables are normalized	Tables are not normalized
Source	Transactions are source of data	OLTP databases become source of data for OLAP
Response time	It's in millisecond	It's in second to minutes
Operation	Allow read/write operations	Only read and rarely write
Purpose	Designed for real time business operations	Designed for analysis of business measures by category and attributes

OLAP Operations

- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels 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 and analysis of the data.
- The most popular end user operations on dimensional data are: ***Roll up, Drill down, Slice, Dice, Pivot***

OLAP Operations: Roll up

- The roll-up operation (also called drill-up or aggregation 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.
- Consider the following cubes illustrating temperature of certain days recorded weekly:

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

- Consider that we want to set up levels (hot (80-85), mild (70-75), cool (64-69)) in temperature from the above cubes, add up value according to concept hierarchies known as **roll-up**

Temp	**Cool**	**Mild**	**Hot**
Week 1	2	1	1
Week 2	1	3	1

OLAP Operations: 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.
- It navigates from less detailed record to more detailed data.
- Drill-down can be performed by either stepping down a concept a hierarchy for a dimension or adding additional dimensions.

Temp	Cool	Mild	Hot
Day 1	0	0	0
Day 2	0	0	0
Day 3	0	0	1
Day 4	0	1	0
Day 5	1	0	0
Day 6	0	0	0
Day 7	1	0	0
Day 8	0	0	0
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	0

OLAP Operations: Slice

- A slice is a subset of the cubes corresponding to a single value for one or more members of the dimension.
- For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site.
- So, the slice operations perform a selection on one dimension of the given cube, thus resulting in a sub-cube. It will form a new sub-cubes
- For example, if we make the selection, temperature = cool we will obtain the following cube: by selecting one or more dimensions.

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

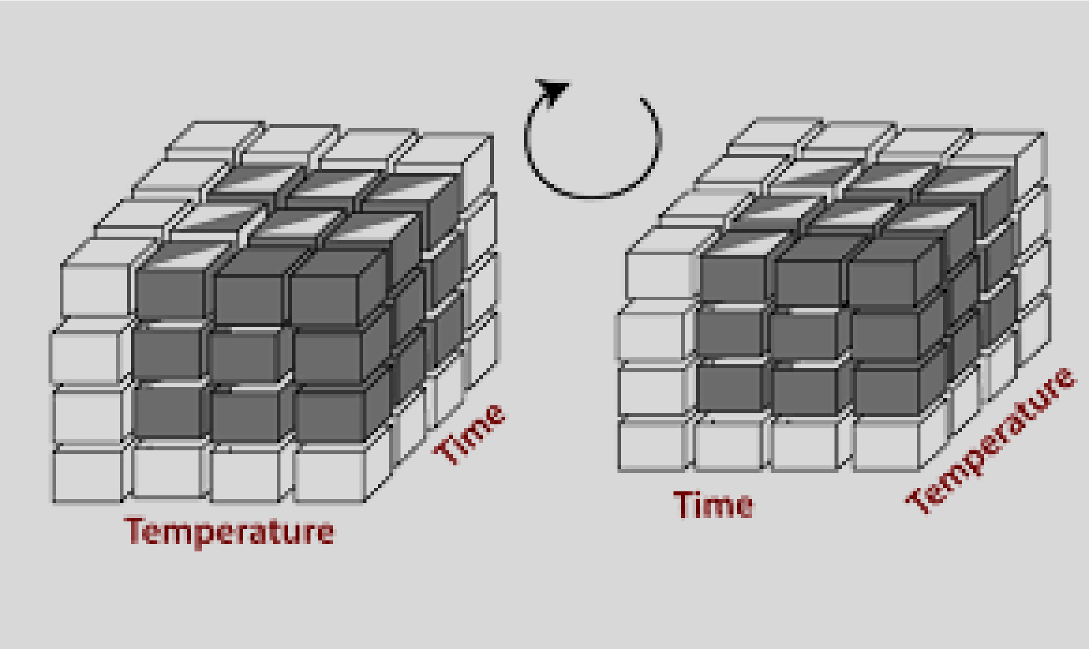
OLAP Operations: Dice

- The dice operation describes a sub-cube by operating a selection on two or more dimension.
- For example, implement the selection (time = day 3 OR time = day 4) AND (temperature = cool OR temperature = hot) to the original cubes we get the following sub-cube (still two-dimensional)

Temp	Cool	Hot
Day 3	0	1
Day 4	0	0

OLAP Operations: Pivot

- The pivot is a operation is also called a rotation.
- Pivot is a visualization operation which rotates data axes in view to provide an alternative presentation of the data.
- It may contain swapping the rows and columns or moving one of the row-dimensions into the column dimensions



OLAP Operations Numerical

OLAP Servers

- There are three OLAP servers namely:

1. Relational OLAP(ROLAP)
2. Multidimensional OLAP(MOLAP)
3. Hybrid OLAP(HOLAP)

Relational OLAP(ROLAP)

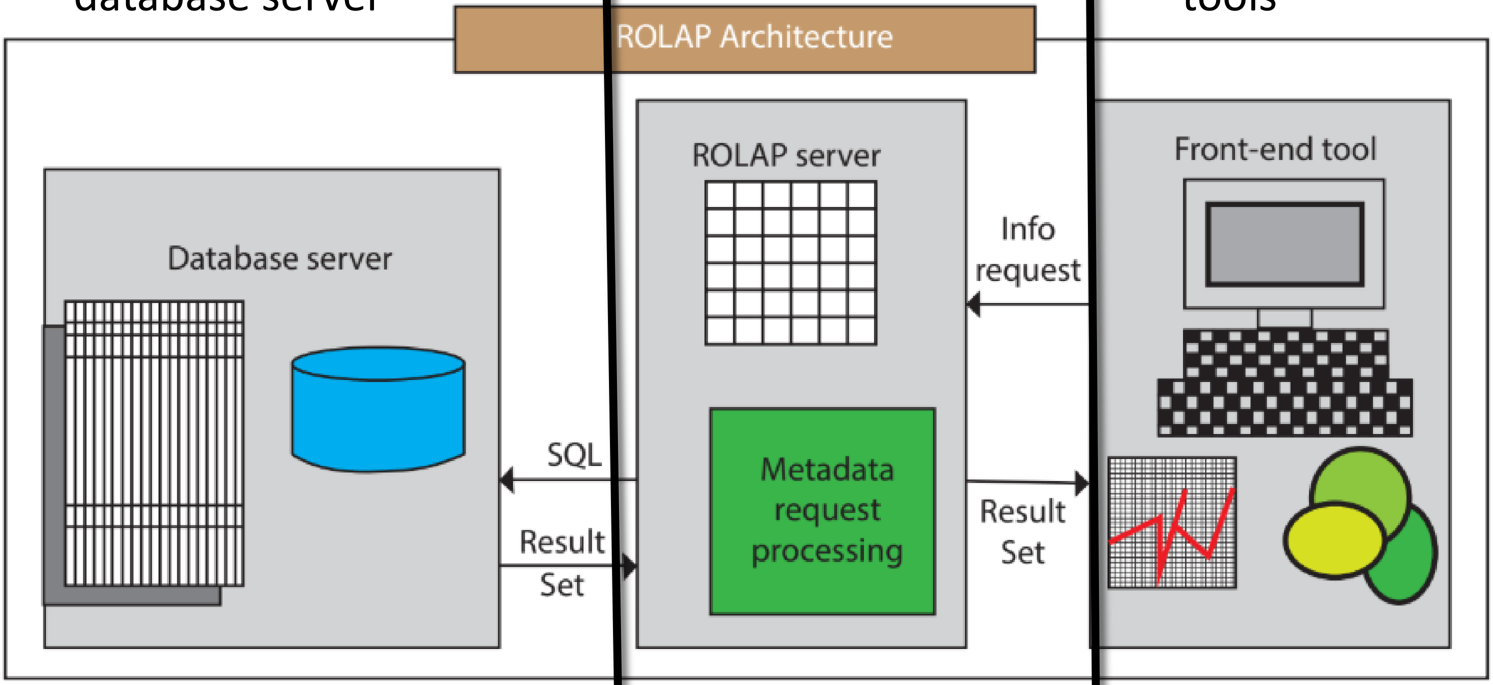
- Relational On-Line Analytical Processing (ROLAP) work mainly for the data that resides in a relational database, where the base data and dimension tables are stored as relational tables
- ROLAP servers are placed between the relational back-end server and client front-end tools.
- ROLAP servers use RDBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
- Example : DSS Server(Digital Surveillance System)
- **Advantages**:
  - ROLAP can handle large amounts of data
  - Can be used with data warehouse and OLTP systems
- **Disadvantages:**
  - Limited by SQL functionalities
  - Hard to maintain aggregate tables

Relational OLAP(ROLAP)

Relational database server

ROLAP Server

End-user access tools



Database Layer

Application logic layer

Presentation layer

Multidimensional OLAP(MOLAP)

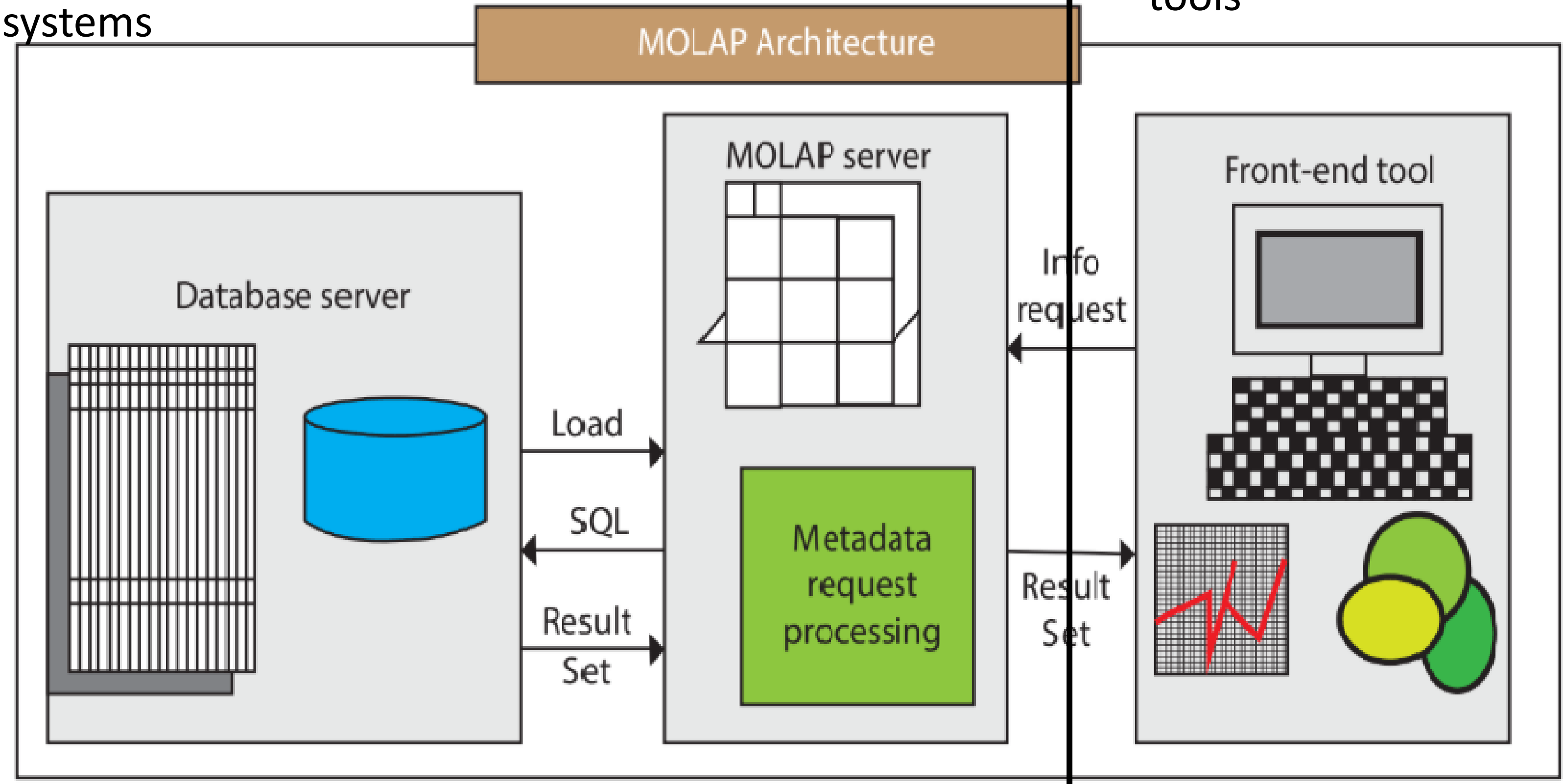
- Multidimensional On-Line Analytical Processing (MOLAP) support multidimensional views of data through array-based multidimensional storage engines.
- With multidimensional data stores, the storage utilization may be low if the data set is sparse.
- Example: Oracle Essbase
- **Advantages**
  - Optimal for slice and dice operations
  - Performs better than ROLAP when data is dense
  - Can perform complex calculations
- **Disadvantages**
  - Difficult to change dimension without re-aggregation
  - MOLAP can handle limited amount of data

Multidimensional OLAP(MOLAP)

Relational database server or legacy systems

MOLAP Server

End user access tools



Database and application logic layer

Presentation Layer

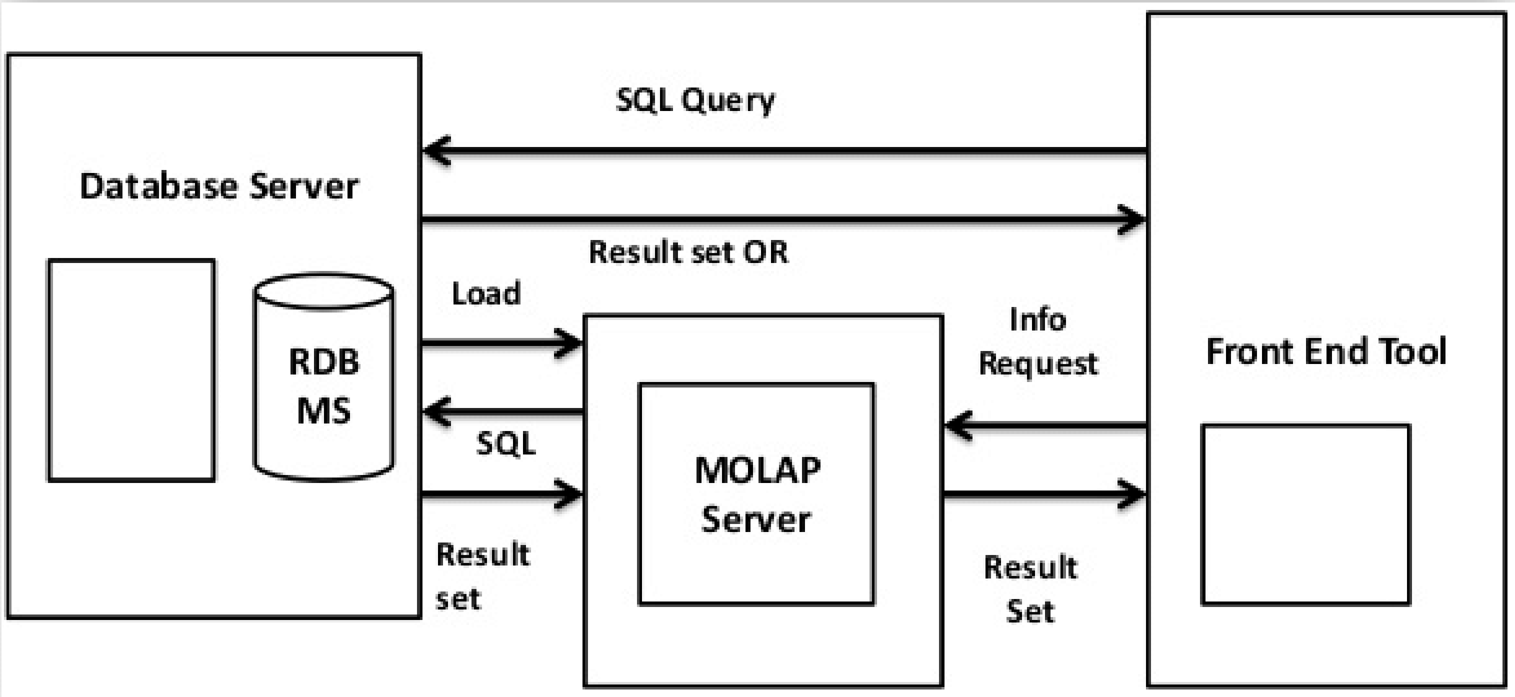
Hybrid OLAP(HOLAP)

- Hybrid On-Line Analytical Processing (HOLAP) is a combination of ROLAP and MOLAP
- HOLAP provide greater scalability of ROLAP and the faster computation of MOLAP.
- Example: Microsoft SQL Server 2000
- **Advantages**
  - HOLAP provide advantages of both MOLAP and ROLAP
  - Provide fast access at all levels of aggregation
- **Disadvantage**
  - HOLAP architecture is very complex because it supports both MOLAP and ROLAP servers

Hybrid OLAP(HOLAP)

Relational
Database Server

End-user access
tools



Applications of OLAP

- OLAP system is to analyze the business which helps in forecasting, planning, problem solving, decision-making. Some of the applications of OLAP include:
- **Financial Applications**
  - Resource (man-power, raw material) allocation
  - Budgeting
- **Sales Applications**
  - Research on market analysis
  - Forecasting sales
  - Analyzing sales promotions
  - Analyzing customer requirements
  - Dividing market based on customer
- **Business Modelling**
  - Understanding and simulating the market trend and business behavior
  - Decision support system for managers, executives, CEO, data scientists.

Thank You!!