

Data mining and Warehousing

Module Code: CSC5901



UNIVERSITY
of **KIGALI**

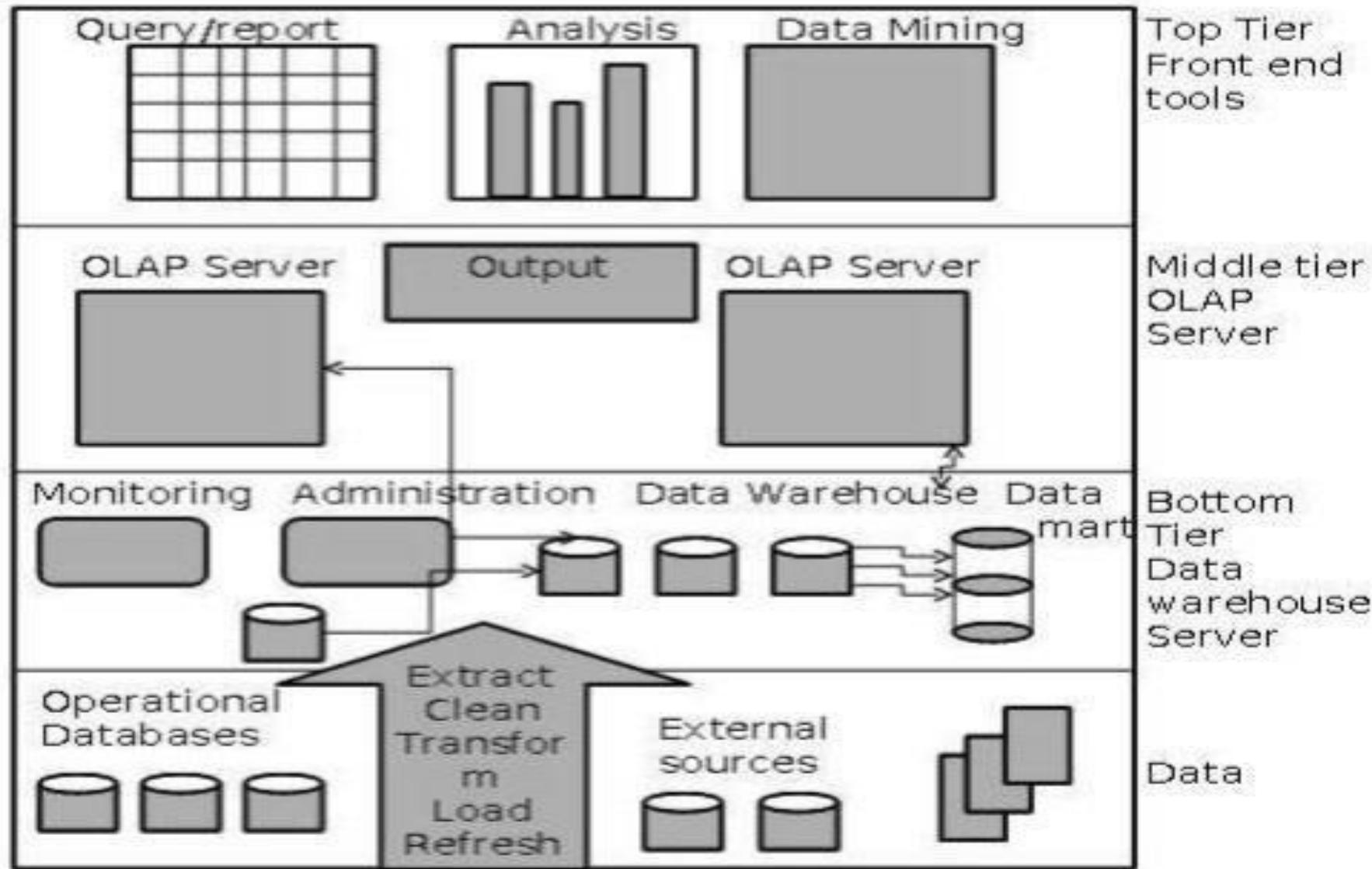
Unit 3

Data Warehouse Description

Data Warehouse Architecture

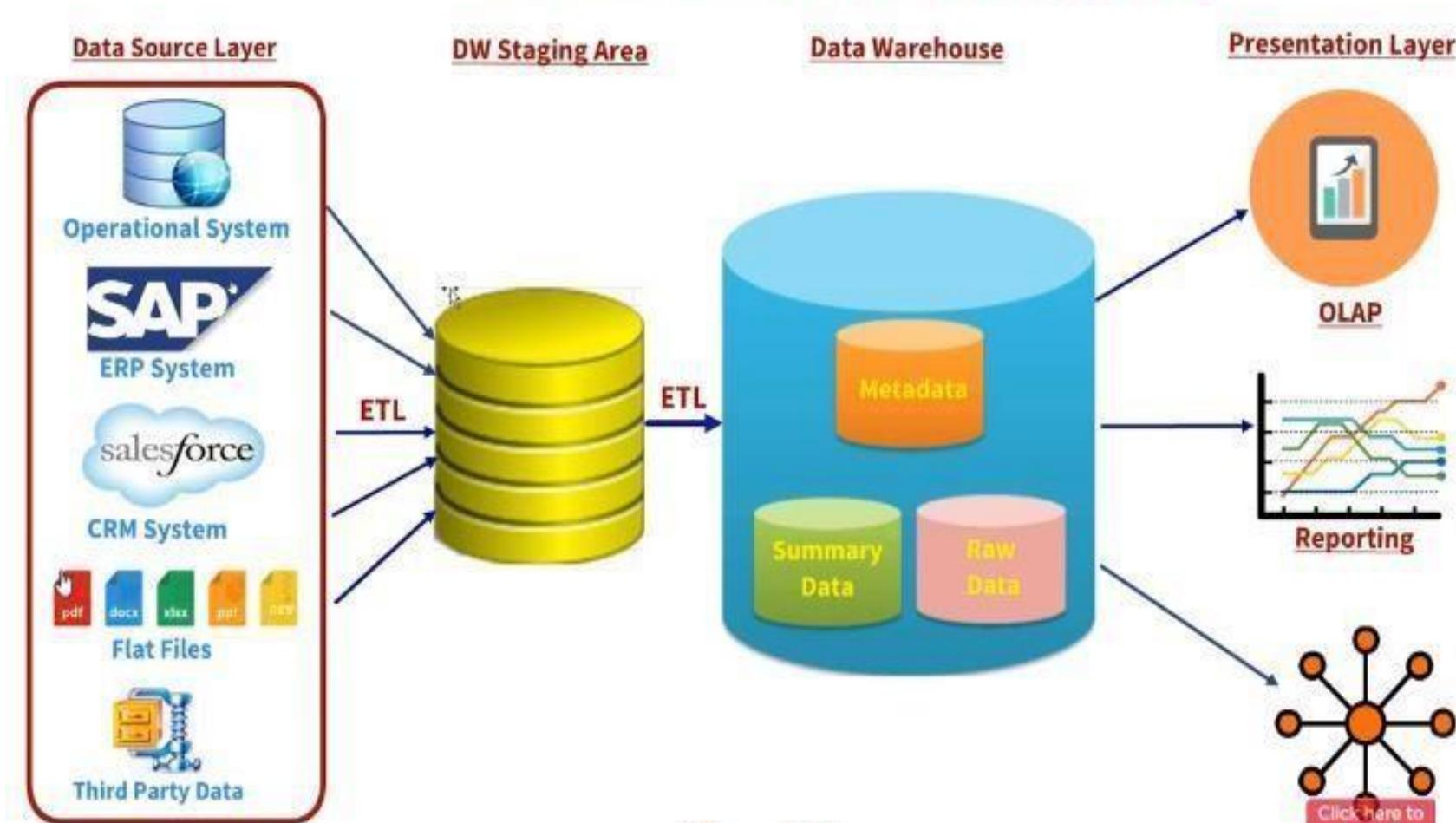
- A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise.

Data Warehouse Architecture



Data Warehouse Architecture(Cont...)

Datawarehouse Architecture



Components of data warehouse

- **People** – people who use data
- **Data** – the data that the information system records
- **Business Procedures** – procedures put in place on how to record, store and analyze data
- **Hardware** – these include servers, workstations, networking equipment, printers, etc.
- **Software** – these are programs used to handle the data. These include programs such as spreadsheet programs, database software, etc.

Objectives of Data warehouse:

- **Data Capturing:** warehouse capture data from various internal and external sources of organization.
- **Processing of Data:** The captured data is processed to convert into required information. Processing include calculating, sorting, classifying, and summarizing.
- **Storage of Information:** stores the processed or unprocessed data for future use.
- **Retrieval of Information :** retrieves information from its stores as and when Required by various users.

Characteristics of warehouse:

- **User friendly/Flexibility:** should be flexible
- **Management Oriented:** provide information support to the management in the organization for decision making.
- **Management directed:** it should be directed by the management because it is the management who tells their needs and requirements more effectively than anybody else.
- **Needs Based:** the design should be as per the information needs of managers at different levels.

Characteristics of warehouse:

- **Integrated:** must be integrated so that all the operational and functional information sub systems should be worked together as a single entity.
- **Exception Based:** should be developed on the exception based also, which means that in an abnormal situation, there should be immediate reporting about the exceptional situation to the decision –makers at the required level.

Characteristics of warehouse:

- **Future Oriented:** should not merely provide past of historical information; rather it should provide information, on the basis of future projections on the actions to be initiated.
- **Long Term Planning:** developed over relatively long periods. A heavy element of planning should be involved.
- **Central database:** it must be common data base for whole system

Challenges of warehouse

- **Complications:** The integration feature is one of the most important aspects of a data warehouse.
- **Maintenance costs outweigh the benefits:** high maintenance systems affect the revenue for medium scale organizations.
- **Data Rigidity:** The type of data imported into a data warehouse is often static data sets which have the least flexibility to generate specific solutions.

Challenges of warehouse

- **Hidden problems of the Source:** Hidden problems of the source arise when an organization finds themselves with problems related to the original source systems which were involved in the importing of data into the warehouse after several years of operation.
- **Inability to capture required data:** There is always the probability that the data which was required for analysis by the organization was not integrated into the warehouse leading to loss of information.

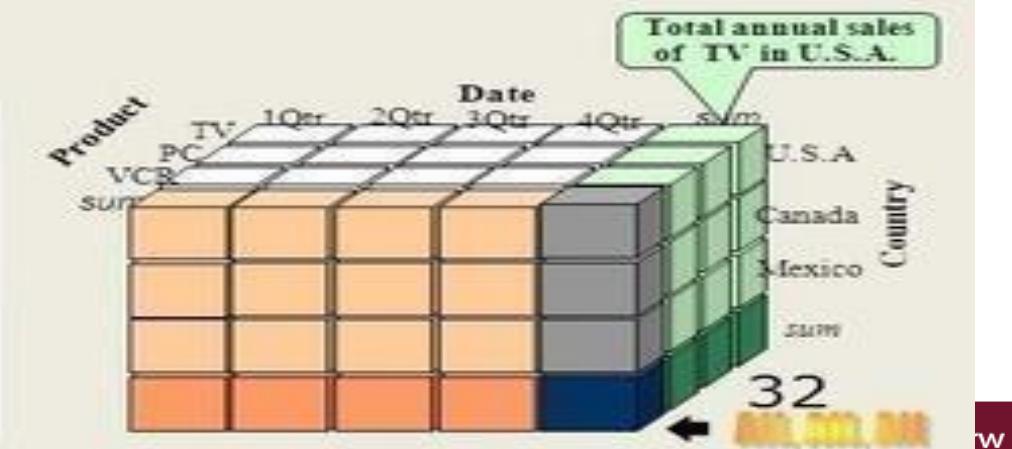
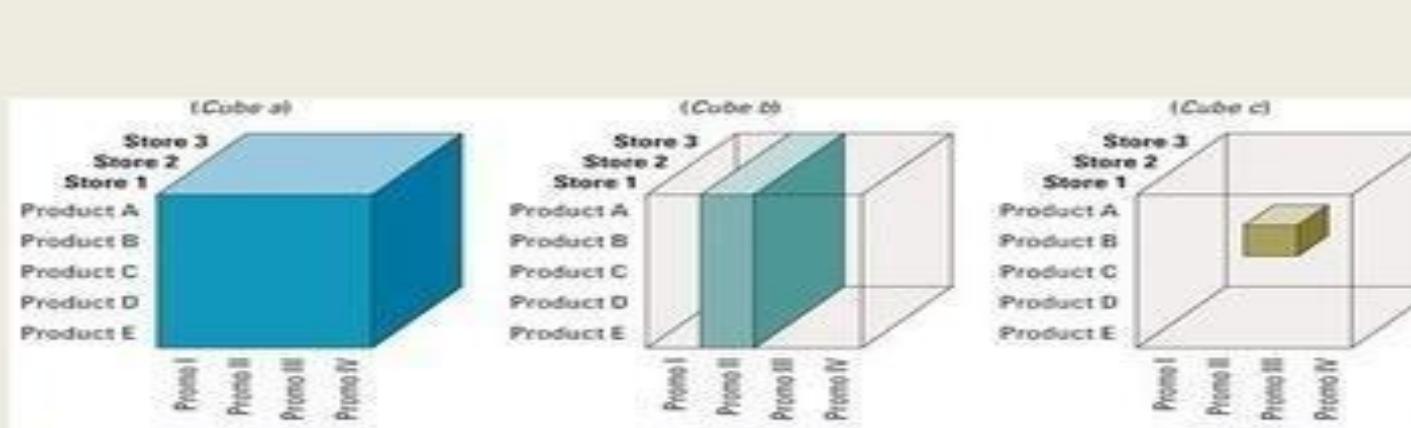
Data warehouse vs operational databases

- Data warehouse platforms are different from operational databases because they store historical information, making it easier for business leaders to analyze data over a specific period of time.
- Data warehouse platforms also sort data based on different subject matter, such as customers, products or business activities.

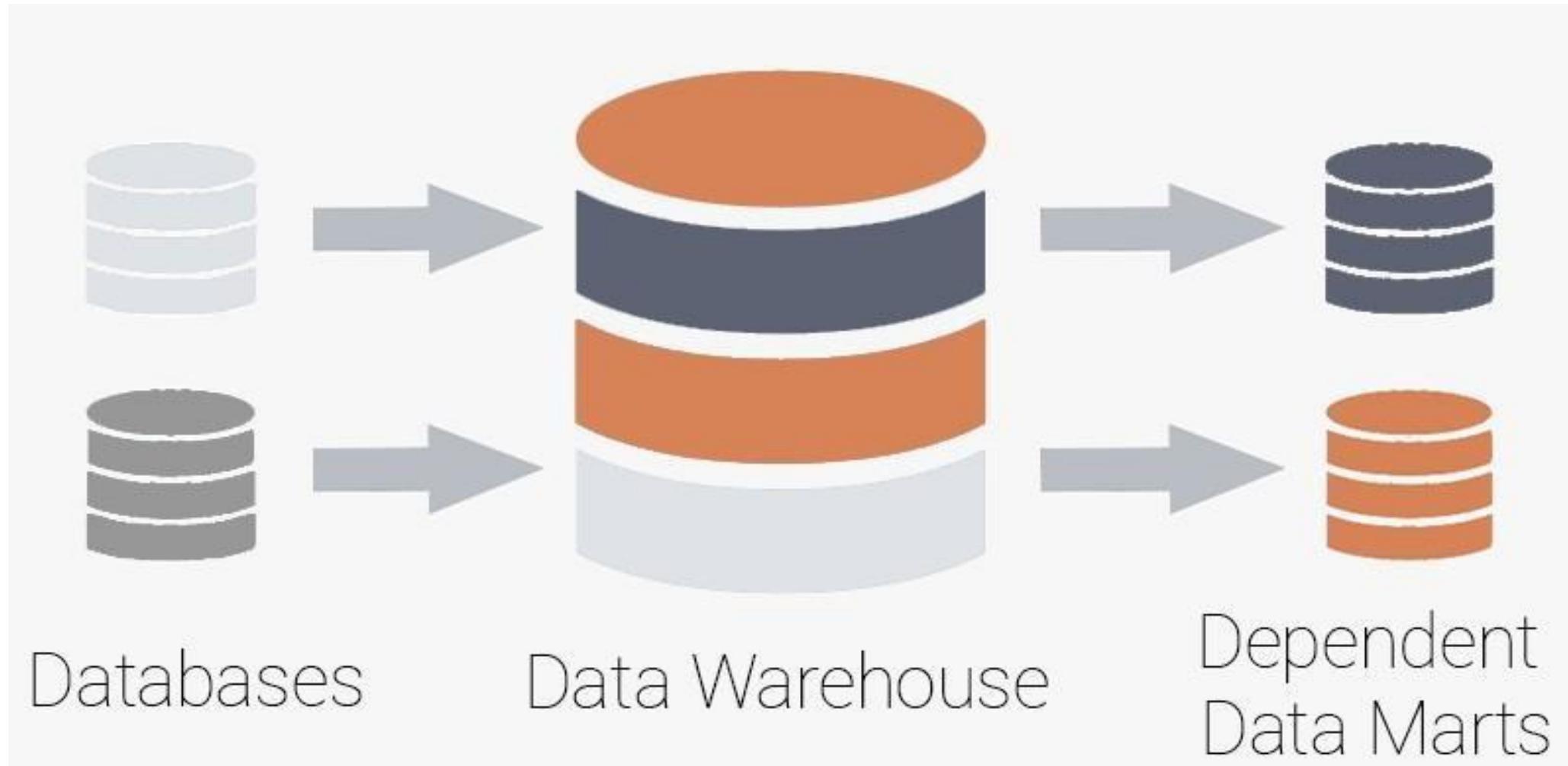
Database Vs Data warehouse

Database vs. Data Warehouse

- Databases contain information in a series of two-dimensional tables
- In a Data Warehouse and data mart, information is multidimensional, it contains layers of columns and rows



Database Vs Data warehouse



Operational Database

Designed to support high-volume transaction processing.

concerned with current data.

Data within operational systems are mainly updated regularly according to need.

Data Warehouse

Designed to support high-volume analytical processing (i.e., OLAP).

concerned with historical data.

Non-volatile, new data may be added regularly. Once Added rarely changed.

Operational Database

It is designed for real-time business dealing and processes.

It is optimized for validation of incoming information during transactions, uses validation data tables.

It supports thousands of concurrent clients.

Data Warehouse

It is designed for analysis of business measures by subject area, categories, and attributes.

Loaded with consistent, valid information, requires no real-time validation.

It supports a few concurrent clients relative to OLTP.

Operational Database

Operational systems are widely process-oriented.

optimized to perform fast inserts and updates of associatively small volumes of data.

Data In

Less Number of data accessed.

Used for on-line transactional Processing (OLTP)

6/25/2021

Data Warehouse

Data warehousing systems are widely subject-oriented

optimized to perform fast retrievals of relatively high volumes of data.

Data Out

Large Number of data accessed.

used for on-line Analytical Processing (OLAP)

18

gali

Info@uok.ac.rw
www.uok.ac.rw

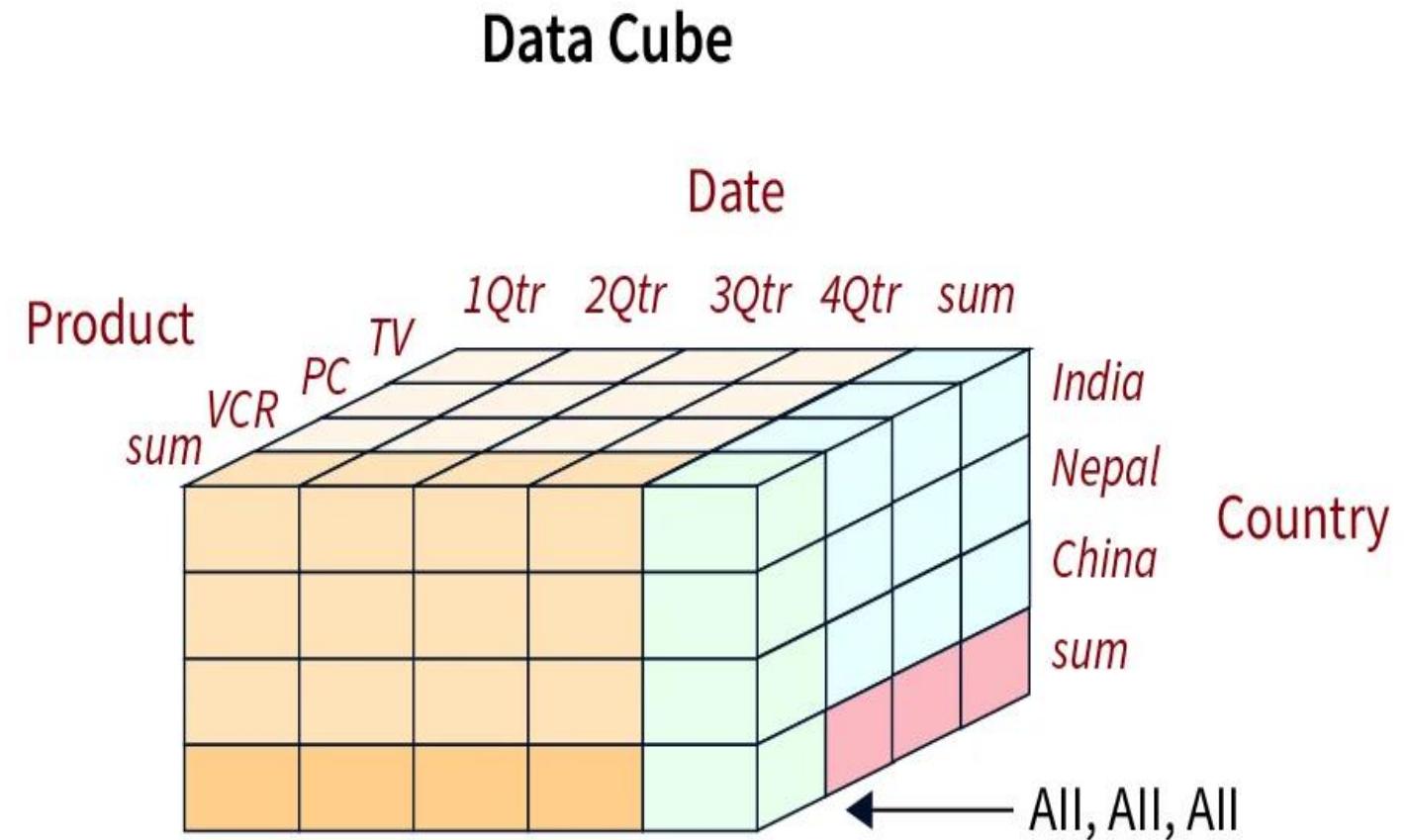
Data Cube

What is Data Cube?

- **Data Cube** or **Multidimensional databases** or **materialized views** or **multidimensional structure** : When data is grouped or combined in multidimensional matrices called Data Cubes.
- A data cube in data mining is a multi-dimensional array that contains pre-aggregated data for efficient analysis. It provides a way to represent data in multiple dimensions, such as time, location, and product, allowing users to view data from different angles and gain insights into patterns and trends.

What is Data Cube?

A data cube is a **multidimensional data model** that store the optimized, summarized or aggregated data which eases the OLAP tools for the quick and easy analysis.

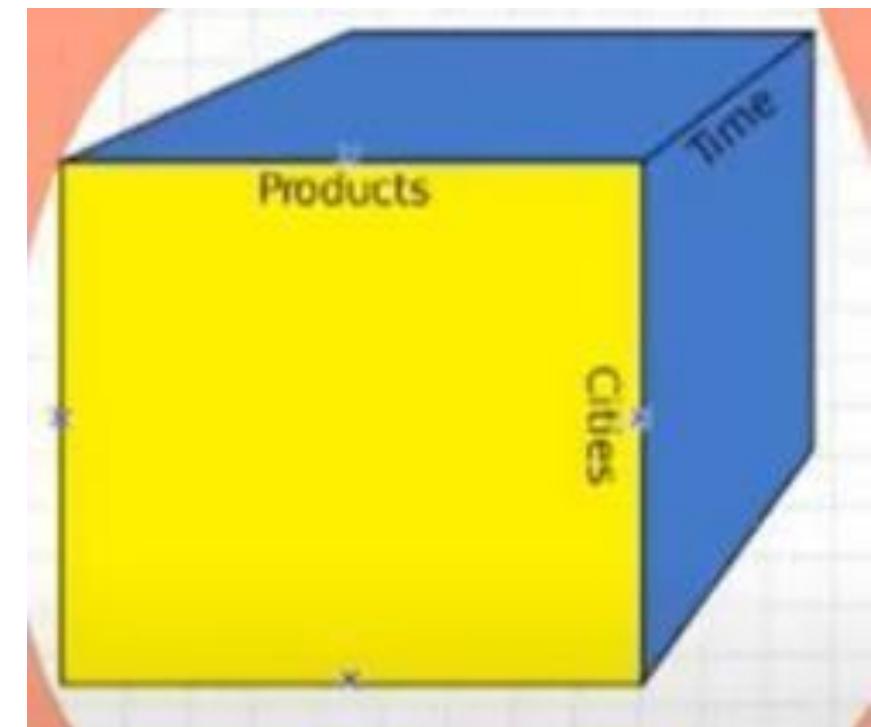


Dimensions of data cube

Data stored in a data cube is represented in terms of **dimensions** and **facts**.

The **dimensions of data cube** are the attitude, angle or the entities with respect to which the enterprise wants to store the data. Dimension here are:

1. Product
2. Cities
3. Time



Types of Data Cube

There are two types of Data cubes which are used mostly in business or enterprises:

- 1. **Multidimensional Data Cube (MOLAP)**
- 2. **Relational Data Cube (ROLAP)**

Multidimensional Data Cube (MOLAP)

As its name suggests Multidimensional Data cube is used mostly in the business requirement where there are huge sets of data. Products developed and follow involves the structure of MOLAP which **has a multidimensional array** format. This structure helps in improving the huge data set with a sparser and an increased level of MOLAP. From this, we can come into a fact that this will not represent any specific data or clustered data value from a data set.

Relational Data Cube (ROLAP)

It is also another category of data analysis data cube which religiously follows **the relational database model**. If we compared to the Multi-dimensional data cube, then it possesses double the number of relational tables to specify the dimensions with data sets and requirements. Each of these tables contains a specific view which is called as a cuboid.

Data representation

Example1: In the 2-D representation,

With 2- D we will look at the All Electronics sales data for **items sold per quarter** in the city of Vancouver. The measured display in dollars sold (in thousands).

2-D view of Sales Data

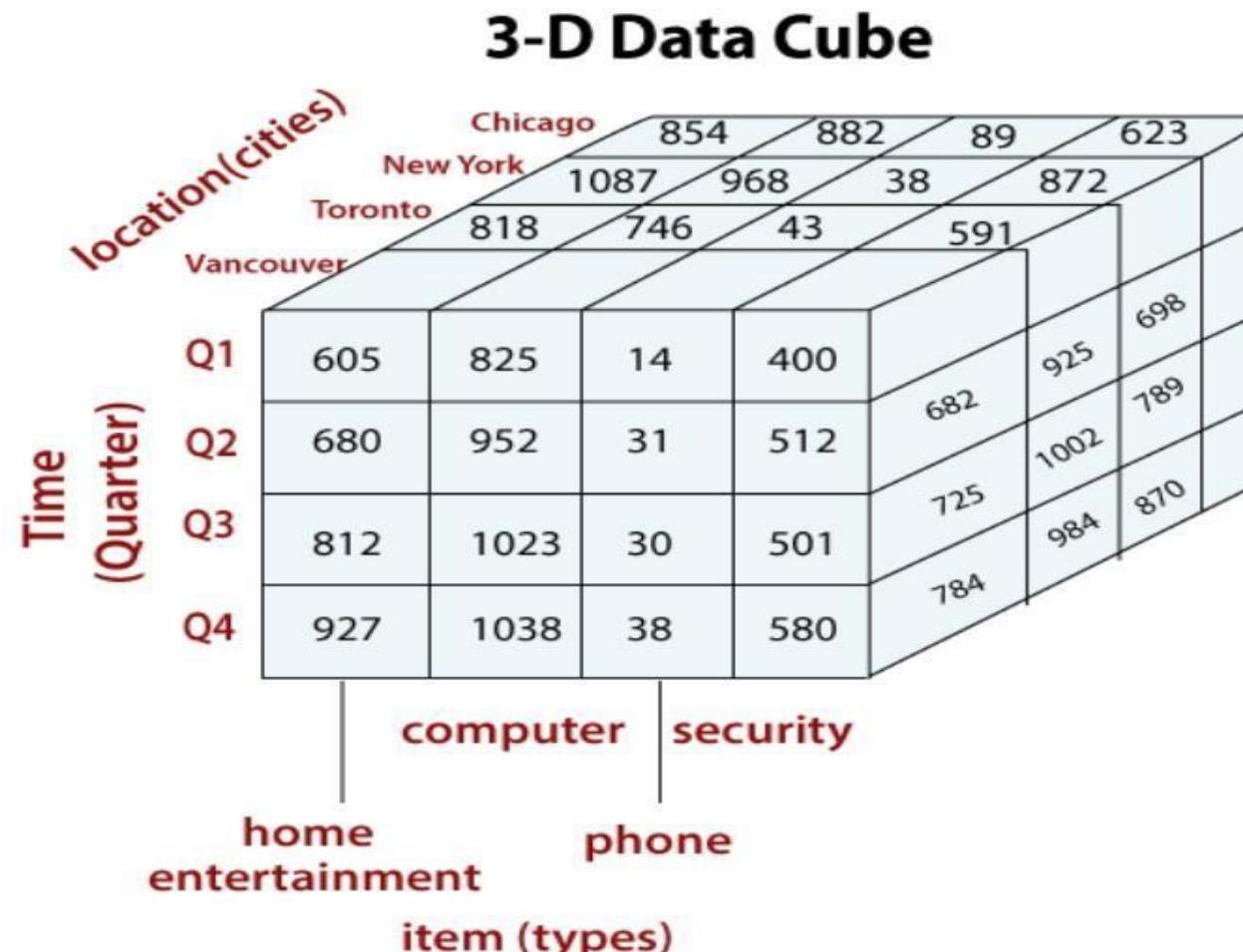
location ="Vancouver"				
time (quarter)	item (type)			
	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q3	927	1038	38	580

Example2: In the 3-D representation

3-D view of Sales Data

location ="Chicago"					location ="New York"					location ="Toronto"				
item					item					item				
home					home					home				
time	ent.	comp.	phone	sec.	time	ent.	comp.	phone	sec.	time	ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872		818	746	43	591	
Q2	943	890	64	698	1130	1024	41	925		894	769	52	682	
Q3	1032	924	59	789	1034	1048	45	1002		940	795	58	728	
Q4	1129	992	63	870	1142	1091	54	984		978	864	59	784	

Example2: In the 3-D representation in data cube



Data cube exercises

kigali			rubavu			muhangা			time
Rice	kawunga	beans	Rice	kawunga	beans	Rice	kawunga	beans	
23	45	67	65	54	43	21	11	67	Q1
34	56	78	56	78	98	87	65	54	Q2
32	31	24	98	87	76	65	54	32	Q3
43	56	76	54	43	32	21	10	67	Q4
54	43	42	56	78	90	87	69	70	Q5

[1][2]= 7	6	7	4
[1][1]= 6	9	8	2
[2][2]= 8			
[2][3]= 2			
[1][3]= 4			
[2][1]= 9			

OLAP Operations

All of the OLAP tools are built upon with four (4) basic analytical operations

Here is the list of OLAP operations –

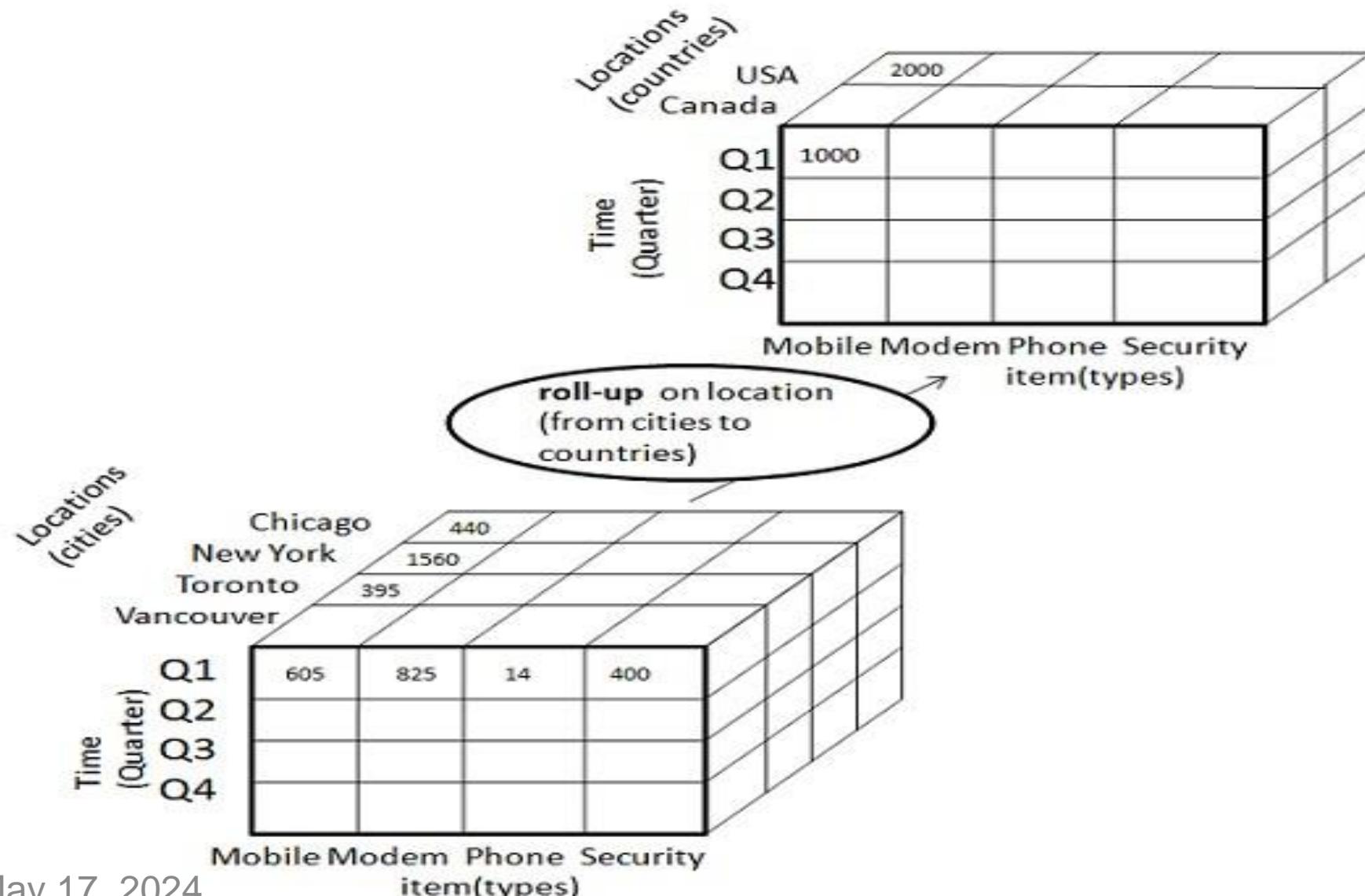
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

- **Roll-Up**
- The roll-up operation (**also known as drill- up or aggregation or Consolidation operation**)
- performs aggregation on a data cube, by climbing down concept hierarchies,
- i.e., dimension reduction. Roll-up is like **zooming-out** on the data cubes.
- performs data aggregation that can be

❖ Sample 1: Operations

- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was “street < city < province < country”.
- data is aggregated by ascending the location hierarchy from city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

Roll-up



OLAP Operations

Roll-Up: Sample 2:

Consider the following cubes illustrating temperature of certain days recorded weekly:

Temperature	64	65	68	69	70	71	72	75	80	81	83	85
Week1	1	0	1	0	1	0	0	0	0	0	1	0
Week2	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.

To do this, we have to group column and add up the value according to the concept hierarchies. This operation is known as a roll-up.

By doing this, we contain the following cube

Temperature	cool	mild	hot
-------------	------	------	-----

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.
- allows users to navigate through data details.

Sample 1

OLAP Operations

- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was “day < month < quarter < year.”
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

Drill-down:



Drill down on time (from quarters to month)



OLAP Operations

- ❖ **Slicing** : The slice operation selects one particular dimension from a given data cube and provides a new sub-cube.

Slicing :

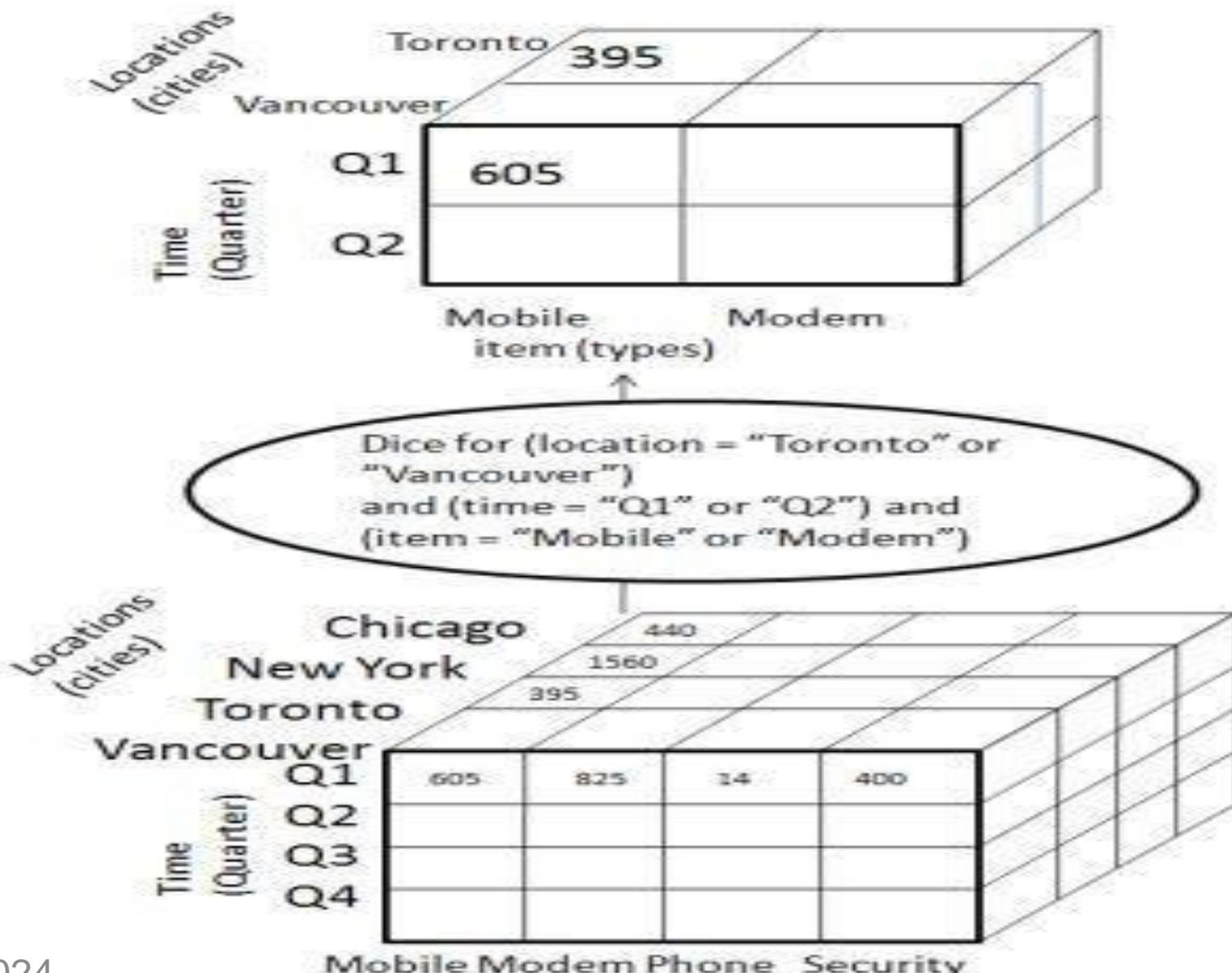


Locations (cities)

	Chicago	New York	Toronto	Vancouver
Time (Quarter)	605	825	14	400
Q1				
Q2				
Q3				
Q4				

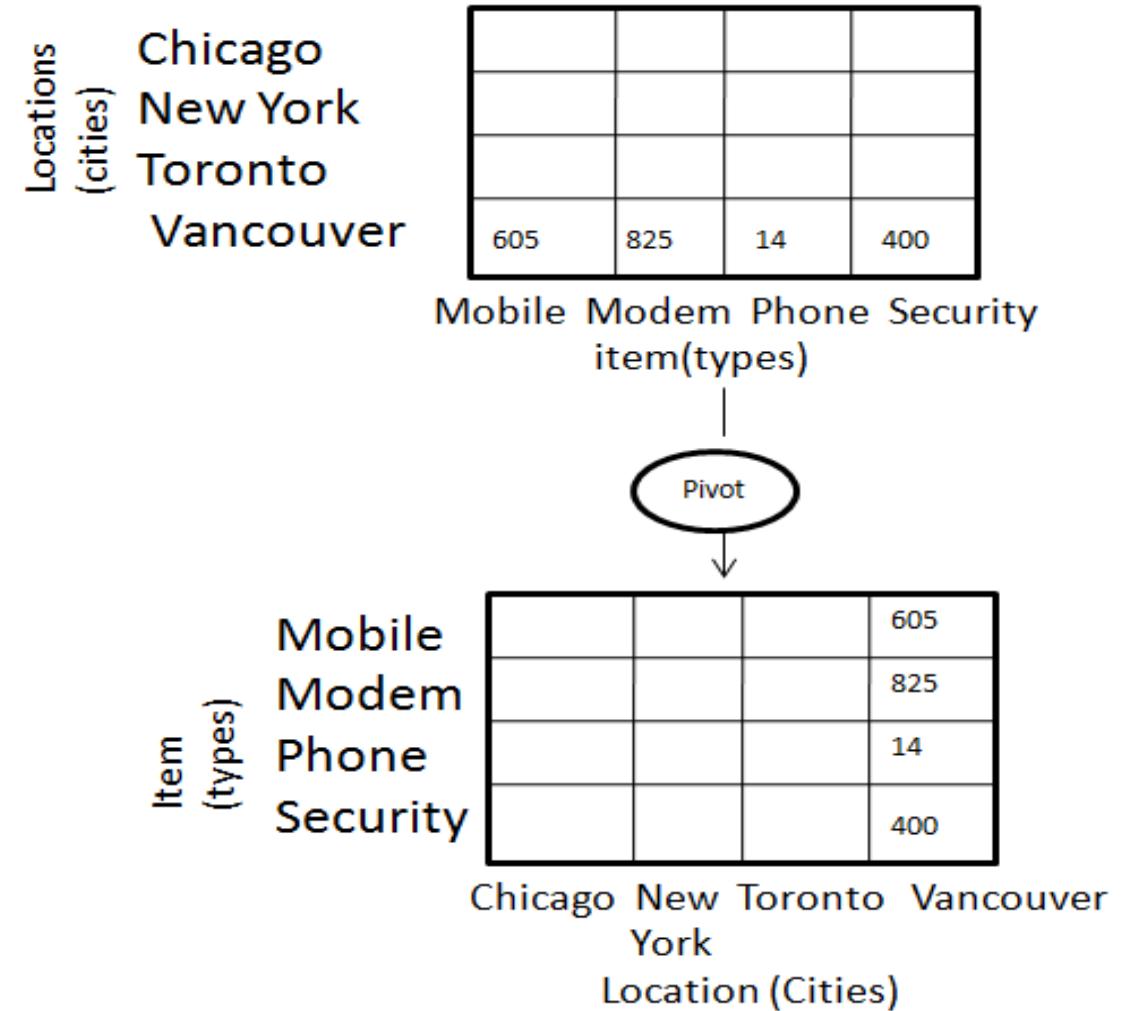
Mobile Modem Phone Security item(types)

Slicing :



Operatio

- **Pivot:** The pivot operation is also known as rotation.
 - It rotates the data axes in view in order to provide an alternative presentation of data.
 - Consider the following diagram that shows the pivot operation.



In addition to these guidelines an OLAP system should also support:

- ✓ Comprehensive database management tools: This gives the database management to
- ✓ control distributed Businesses
- ✓ The ability to drill down to detail source record level: Which requires that
- ✓ The OLAP tool should allow smooth transitions in the multidimensional database.
- ✓ Incremental database refresh: The OLAP tool should provide partial refresh.
- ✓ Structured Query Language (SQL interface): the OLAP system should be able to integrate effectively in the surrounding enterprise environment.

Advantages of Data Cube

Data cube in data mining provides several advantages –

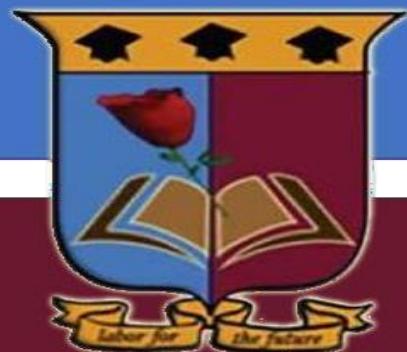
- **Multidimensional analysis** – Data cube technology in data mining enables users to analyze data from multiple perspectives and dimensions, such as time, product, location, and customer, allowing for a more comprehensive data view.
- **Fast query performance** – Data cubes pre-aggregate data at multiple levels of granularity, making it easier and faster to query large datasets and retrieve results.
- **Reduced data redundancy** – Data cubes store pre-aggregated data at various levels of granularity, reducing the need to store redundant data in a database.
- **Data visualization** – Data cube in data mining can be visualized using charts, graphs, and other graphical representations, making it easier for users to understand and analyze complex data.
- **Improved decision-making** – Data cube technology in data mining allows users to drill down, roll up, slice, and dice data, enabling them to make informed decisions based on insights gained from the data.
- **Scalability** – Data cubes can handle large datasets and be stored in a database, making them scalable for enterprise-level data mining.

Disadvantages of Data Cube

While data cube in data mining provides several advantages, they also have some disadvantages –

- **Data cube creation** – Creating a data cube in data mining can be a time-consuming and complex process that requires careful consideration of the dimensions, measures, and aggregation levels.
- **Data storage requirements** – Data cubes can require significant storage space, especially when dealing with large datasets with many dimensions and measures.
- **Limited flexibility** – Data cubes are optimized for multidimensional analysis and may need to be more flexible to accommodate changes to the underlying data or analysis requirements.
- **Data quality issues** – Data cube technology in data mining relies on the accuracy and consistency of the underlying data, which can be challenging to achieve when dealing with complex datasets.
- **Complexity** – While data cubes simplify the analysis of complex data, the analysis itself can be complex, requiring knowledge of the dimensions, measures, and aggregation levels used in the data cube.

Thank you



UNIVERSITY
of **KIGALI**