COURSE NAME:
DATA WAREHOUSING & DATA MINING

LECTURE 2
TOPICS TO BE COVERED:

- Data Marts
- Types of Data Marts
- Problems
- Metadata

DATA MARTS

- A data mart contains a subset of wide data that is of value to a specific group of users.
- It is a data store that is a subsidiary to a datawarehouse of integrated data.
- It is a set of denormalized summarized or aggregated data.
- The data contents in data marts tends to be summarized .
- They are usually implemented on low cost departmental servers (UNIX, windows NT)
- The implementation cycle of a data mart is measured in weeks rather than month or years.
- Depending on source of data, data marts can be categorized as independent or dependent.

INDEPENDENT DATA MARTS :

- These are sourced from data captured from one or more operational systems or external information providers.
- Each independent data marts makes its own assumptions about how to consolidate the data and the data across several data marts may not be consistent.

DEPENDENT DATA MARTS:

- It is sourced directly from enterprise datawarehouse.

PROBLEMS WITH DATA MARTS

- Scalability in situations where an initial small data mart grows quickly in multiple dimensions
- Data Integration
- Situations where independent data marts are use
- Extremely urgent user requirements.
- The absence of a budget for a full datawarehouse
- The absence of a sponsor for an enterprise decision support strategy.

METADATA

- Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.
- Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes

ROLE OF METADATA

- Identify subjects of the data mart
- Identify dimensions and facts
- Indicate how data is derived from enterprise data warehouses, including derivation rules
- Indicate how data is derived from operational data store, including derivation rules
- Identify available reports and predefined queries
- Identify data analysis techniques (e.g. drill-down)
- Identify responsible people

A MULTIDIMENSIONAL DATA MODEL

Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a data cube.

In this section, you will learn how data cubes model n-dimensional data. You will also learn about concept hierarchies and how they can be used in basic OLAP operations to allow interactive mining at multiple levels of abstraction.

FROM TABLES AND SPREADSHEETS TO DATA CUBES

- A data warehouse is based on a *multidimensional data model* which views data in the form of a *data cube*
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables

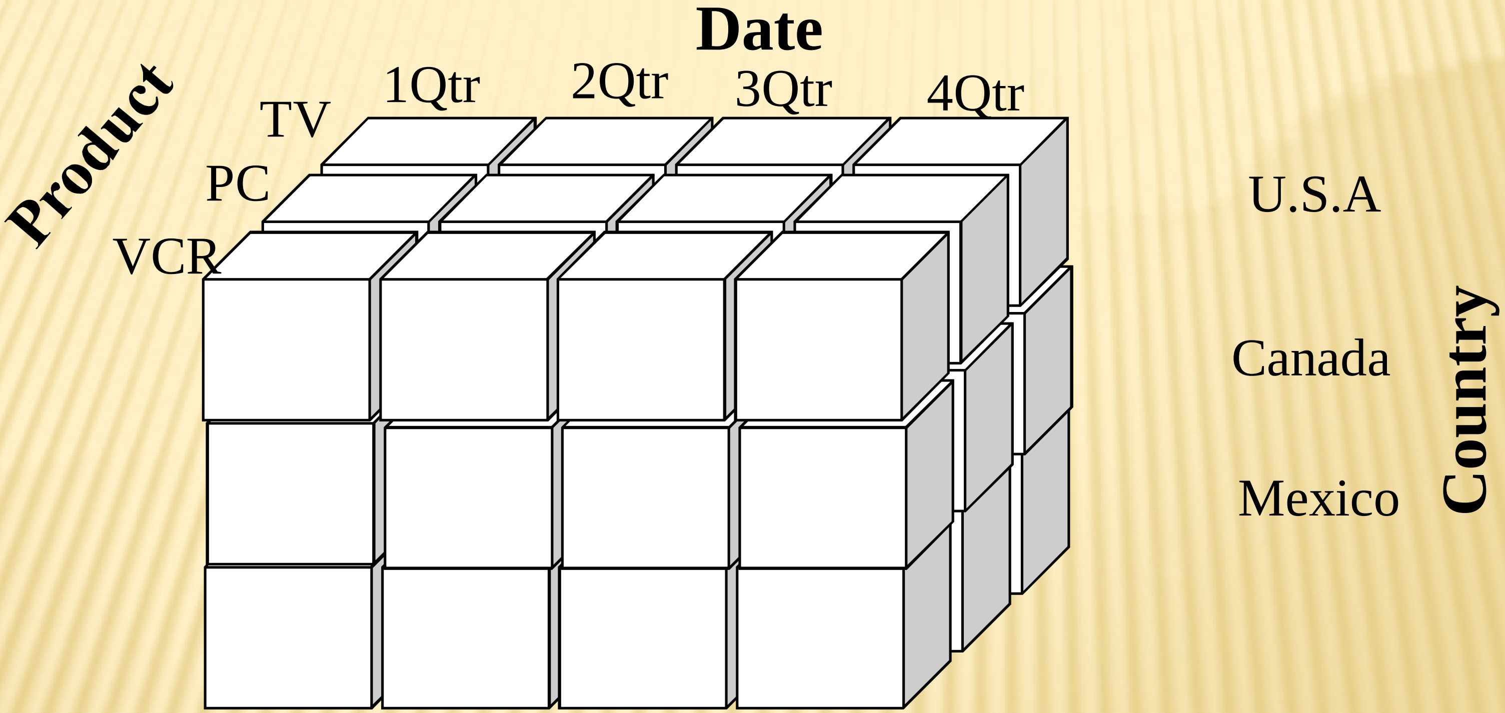
DATA CUBES

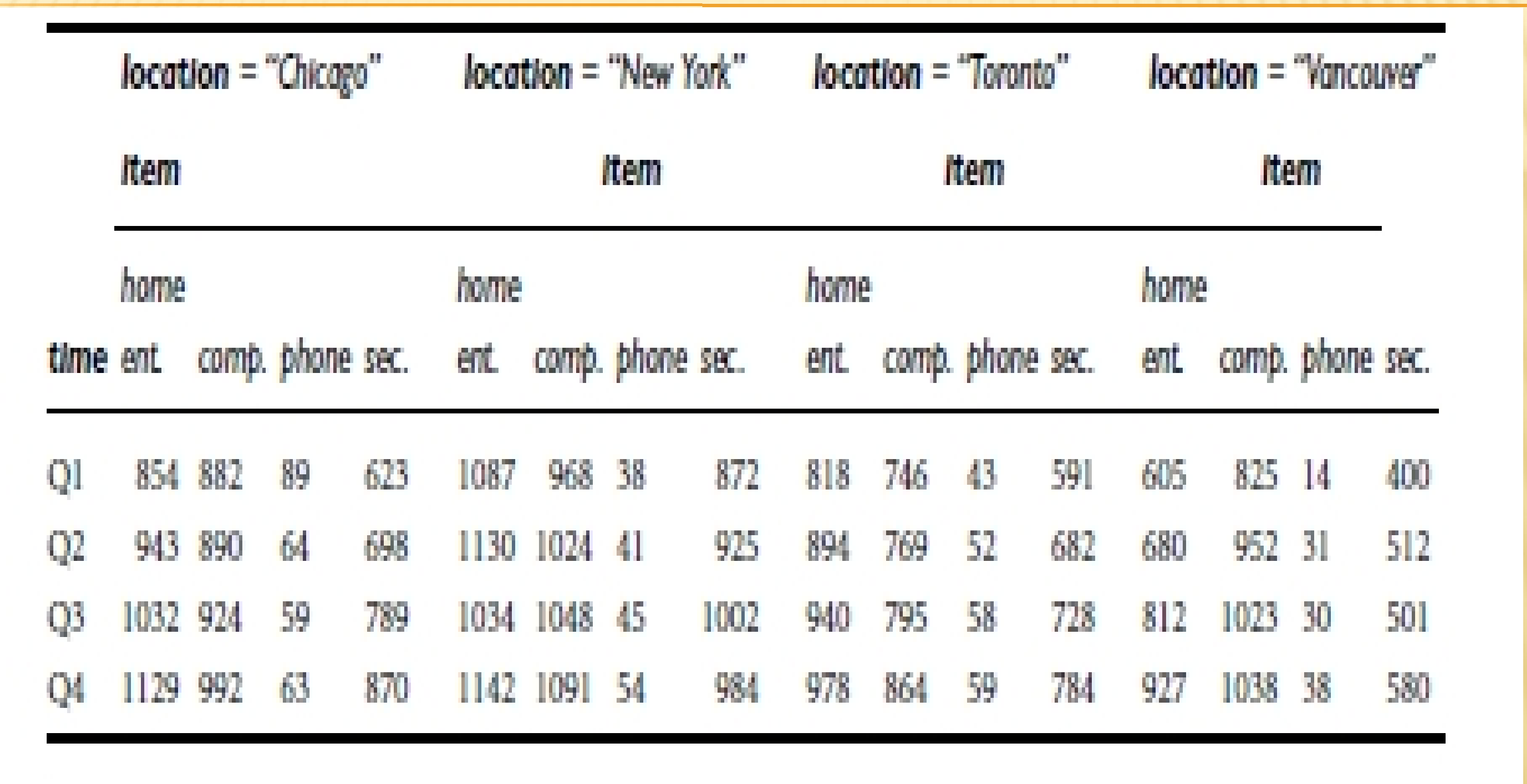
- *"What is a data cube?" A data cube allows data to be modeled and viewed in multiple* dimensions.
- It is defined by dimensions and facts.

DATA CUBES

- In data warehousing literature, an n-D base cube is called a *base cuboid*. The top most 0-D cuboid, which holds the highest-level of summarization, is called the *apex cuboid*. The lattice of cuboids forms a *data cube*.

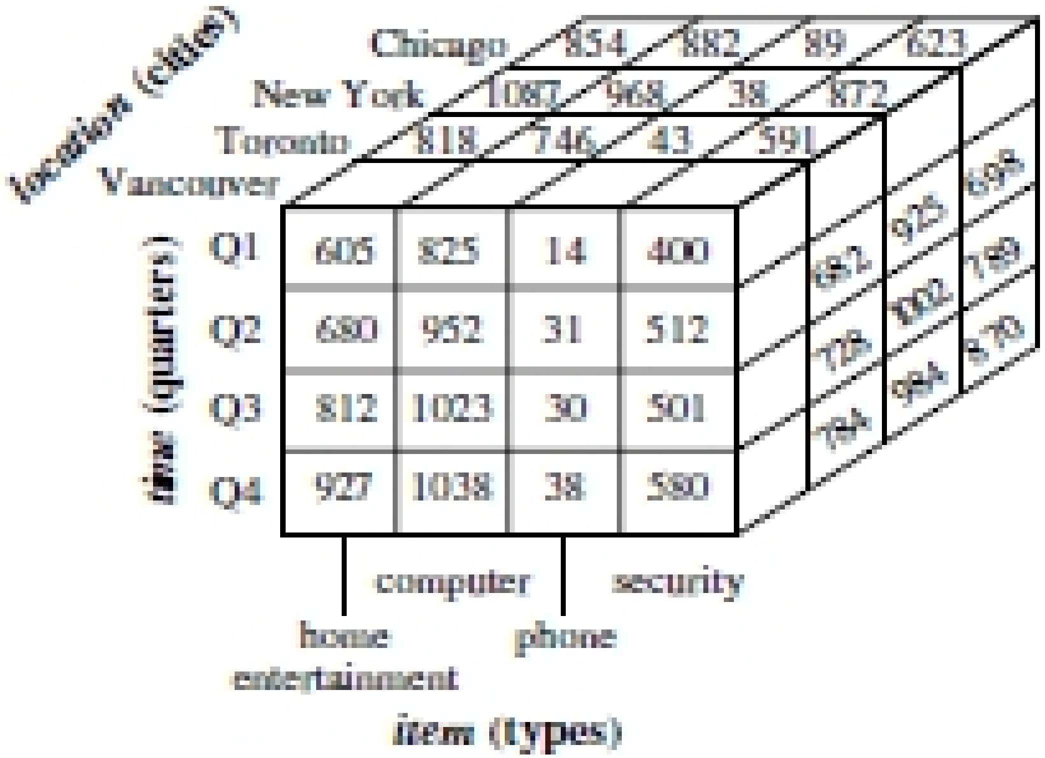
A SAMPLE DATA CUBE





	location = "Chicago"				location = "New York"				location = "Toronto"				location = "Vancouver"			
	item				item				item				item			
time	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).



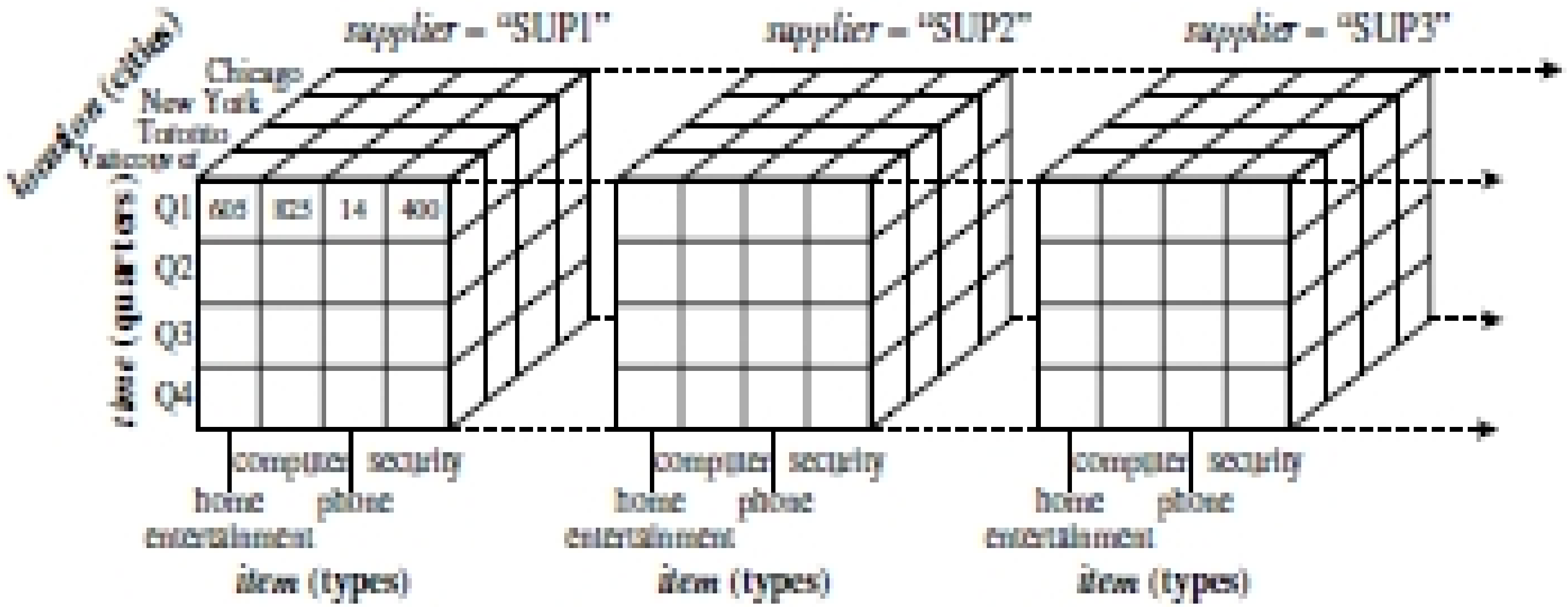
	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

A 3-D data cube representation of the data in Table 3.3, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

Suppose that we would now like to view our sales data with an additional fourth dimension, such as supplier. Viewing things in 4-D becomes tricky. However, we can think of a 4-D cube as being a series of 3-D cubes, as shown in next slide.

If we continue in this way, we may display any n-D data as a series of (n-1)-D “cubes.”

The data cube is a metaphor for multidimensional data storage. The actual physical storage of such data may differ from its logical representation.



A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

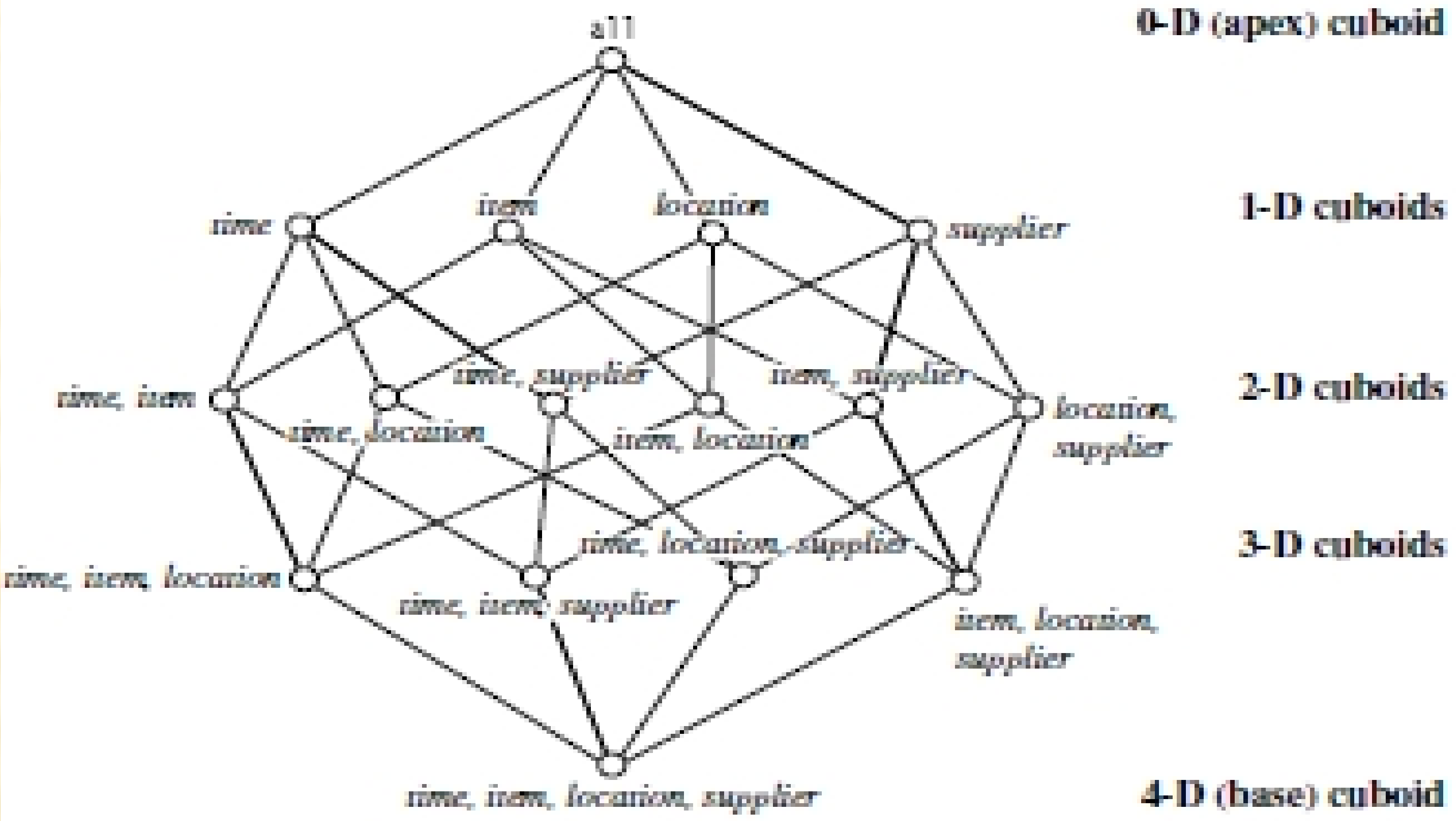
CUBOID

A data cube such as each of the above is often referred to as a cuboid. Given a set of dimensions, we can generate a cuboid for each of the possible subsets of the given dimensions. The result would form a lattice of cuboids, each showing the data at a different level of summarization, or group by.

The lattice of cuboids is then referred to as a data cube. Next slide shows a lattice of cuboids forming a data cube for the dimensions time, item, location, and supplier.

The cuboid that holds the lowest level of summarization is called the base cuboid. For example, the 4-D cuboid in next slide is the base cuboid for the given time, item, location, and supplier dimensions.

The 0-D cuboid, which holds the highest level of summarization, is called the apex cuboid. In our example, this is the total sales, or dollars sold, summarized over all four dimensions. The apex



Lattice of cuboids, making up a 4-D data cube for the dimensions *time*, *item*, *location*, and *supplier*. Each cuboid represents a different degree of summarization.