Subject: MCA DMBI 1003

Data Mining and Business Intelligence

Made By: Priyanka Karale

What is Data

- Data is different types of information usually formatted in a particular manner. All software is divided into two major categories: **programs and data.** We already know what data is now, and programs are collections of instructions used to manipulate data.

What is Information?

Information is defined as classified or organized data that has some meaningful value for the user. Information is also the processed data used to make decisions and take action. Processed data must meet the following criteria for it to be of any significant use in decision-making:

- Accuracy: The information must be accurate.
- Completeness: The information must be complete.
- Timeliness: The information must be available when it’s needed.

Subject: MCA DMBI 1003

Types of data

1. Data stored in the database
2. Data warehouse
3. Transactional data

Made By: Priyanka Karale

Subject: MCA DMBI 1003

Data Warehouse

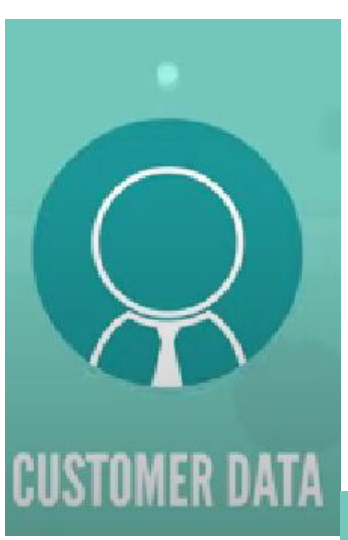
A data warehouse is a single data storage location that collects data from multiple sources and then stores it in the form of a unified plan. When data is stored in a data warehouse, it undergoes cleaning, integration, loading, and refreshing. Data stored in a data warehouse is organized in several parts. If you want information on data that was stored 6 or 12 months back, you will get it in the form of a summary.

Made By: Priyanka Karale

- Data warehouse systems are valuable tools in today's competitive, fast-evolving world. A data warehouse is a central repository of information that can be analyzed to make more informed decisions.
- A data warehouse refers to a data repository that is maintained separately from an organization's operational databases.
- Data warehouse systems allow for integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historic data for analysis.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.

Subject: MCA DMBI 1003

Data Warehouse is the place where valuable data assets of an organization are stored such as



Made By: Priyanka Karale

What are the benefits of using a data warehouse?

a) Informed decision making

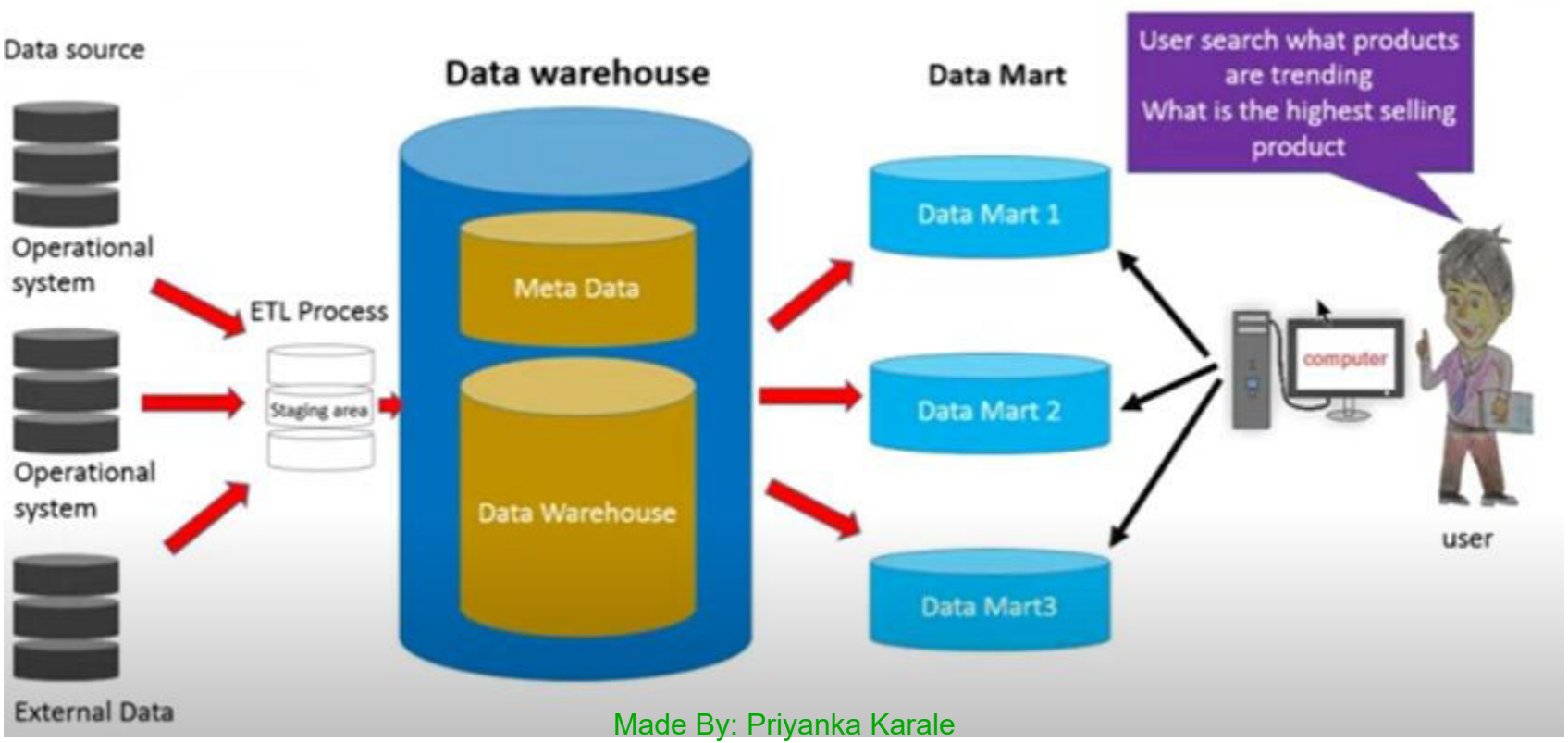
b) Consolidated data from many sources

c) Historical data analysis

d) Data quality, consistency, and accuracy

e) Separation of analytics processing from transactional databases, which improves performance of both systems

Data warehouse Architecture



Data Mart

- A data mart contains a subset of corporate-wide data that is of value to a specific group of users.
- The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales.
- The data contained in data marts tend to be summarized.
- Data marts are usually implemented on low-cost departmental servers that are Unix/Linux or Windows based.
- The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years.

Subject: MCA DMBI 1003

What Is Data Modeling?

Data modeling is the process of creating visual representations of information to draw connections between data points that illustrate relationships.

Made By: Priyanka Karale

Dimensional Data Models

- The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them.

- Such a data model is appropriate for online transaction processing.

- A data warehouse, however, requires a concise, subject-oriented schema that facilitates online data analysis.

- The most popular data model for a data warehouse is a **multidimensional model**, which can exist in the form of a **star schema**, a **snowflake schema**, or a **fact constellation schema**.

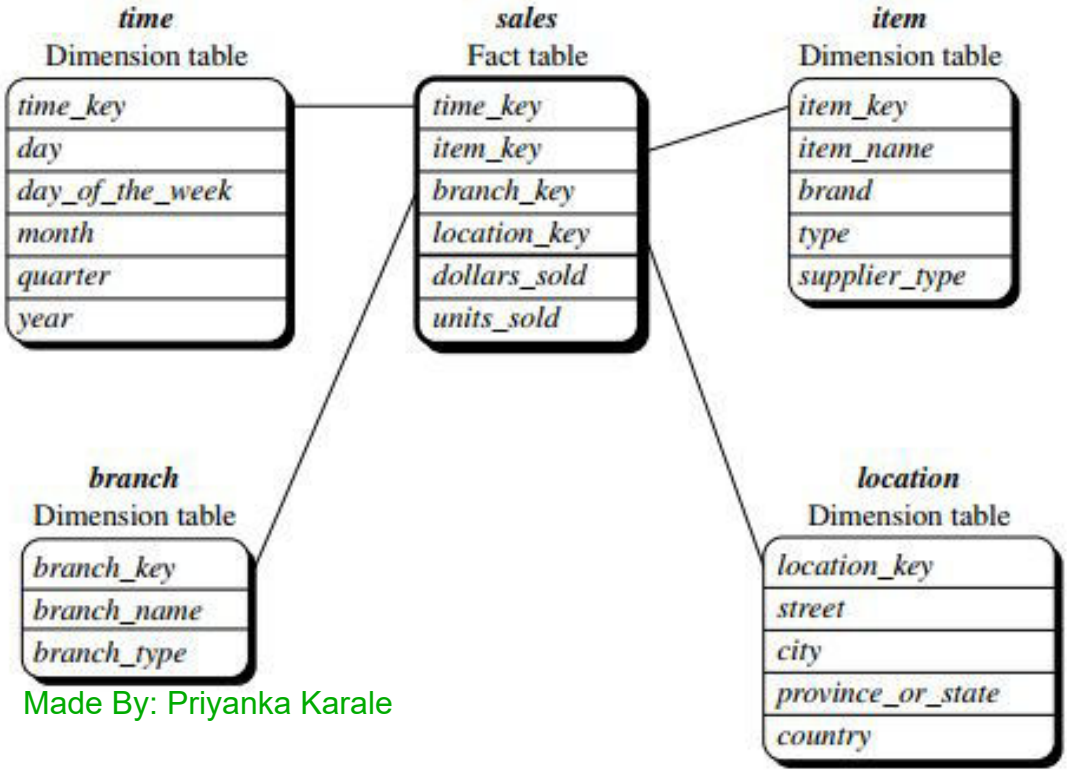
1. Star Schema

- The most common modeling paradigm is the star schema, in which the data warehouse contains:
  (1) a large central table (**fact table**) containing the bulk of the data, with no redundancy, and
  (2) a set of smaller attendant tables (**dimension tables**), one for each dimension.

- The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

- A star schema for *AllElectronics* sales is shown in Figure here. Sales are considered along four dimensions: *time, item, branch*, and *location*.

- The schema contains a central fact table for *sales* that contains keys to each of the four dimensions, along with two measures: *dollars sold* and *units sold*..
  - Each dimension is represented by only one table, and each table contains a set of attributes.



2. Snowflake Schema

- The snowflake schema is a variant of the star schema model, where some dimension tables are *normalized*, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

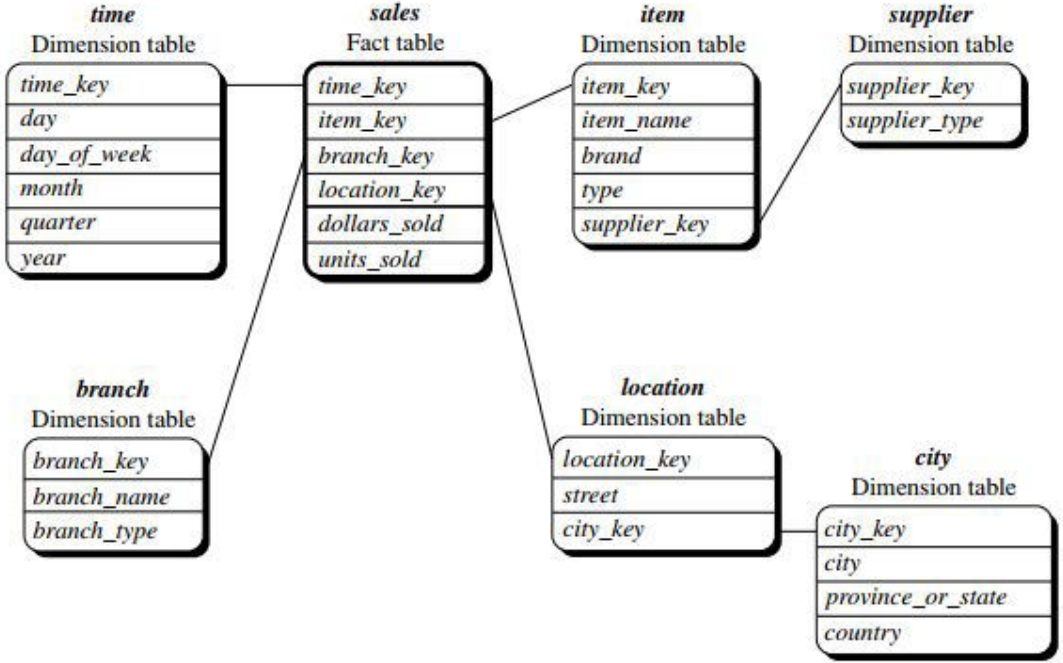
- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.

- Such a table is easy to maintain and saves storage space. However, this space savings is negligible in comparison to the typical magnitude of the fact table.

- Furthermore, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted.

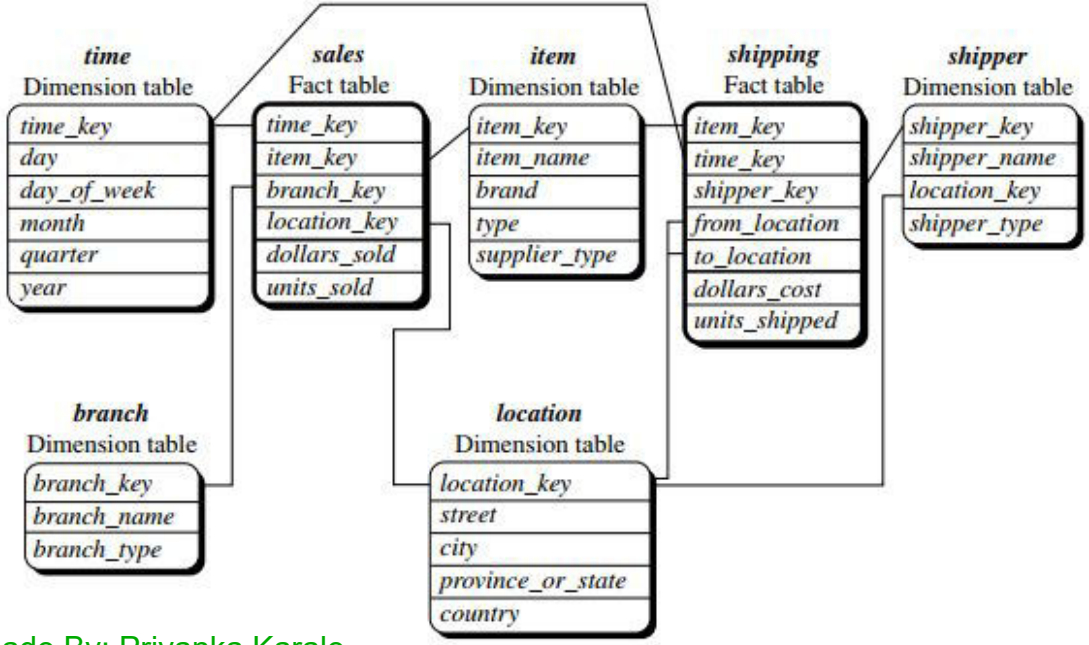
- Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

- A snowflake schema for *AllElectronics* sales is given in Figure.
- Here, the *sales* fact table is identical to that of the star schema.
- The main difference between the two schemas is in the definition of dimension tables.
- The single dimension table for *item* in the star schema is normalized in the snowflake schema, resulting in new *item* and *supplier* table.



3. Fact Constellation

- Sophisticated applications may require multiple fact tables to *share* dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a **galaxy schema** or a **fact constellation**.
- This schema specifies two fact tables, *sales* and *shipping*.
- The *sales* table definition is identical to that of the star schema.
- The *shipping* table has five dimensions, or keys—*item key, time key, shipper key, from location*, and *to location*—and two measures—*dollars cost and units_shipped.*



Made By: Priyanka Karale

- A fact constellation schema allows dimension tables to be shared between fact tables.
- For example, the dimensions tables for *time, item*, and *location* are shared between the *sales* and *shipping* fact tables.
- In data warehousing, there is a distinction between a data warehouse and a data mart. A data warehouse collects information about subjects that span the *entire organization*, such as *customers, items, sales, assets*, and *personnel*, and thus its scope is *enterprise-wide*.
- For data warehouses, the fact constellation schema is commonly used, since it can model multiple, interrelated subjects. A **data mart**, on the other hand, is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is *departmentwide*.
- For data marts, the *star* or *snowflake* schema is commonly used, since both are geared toward modeling single subjects, although the star schema is more popular and efficient.

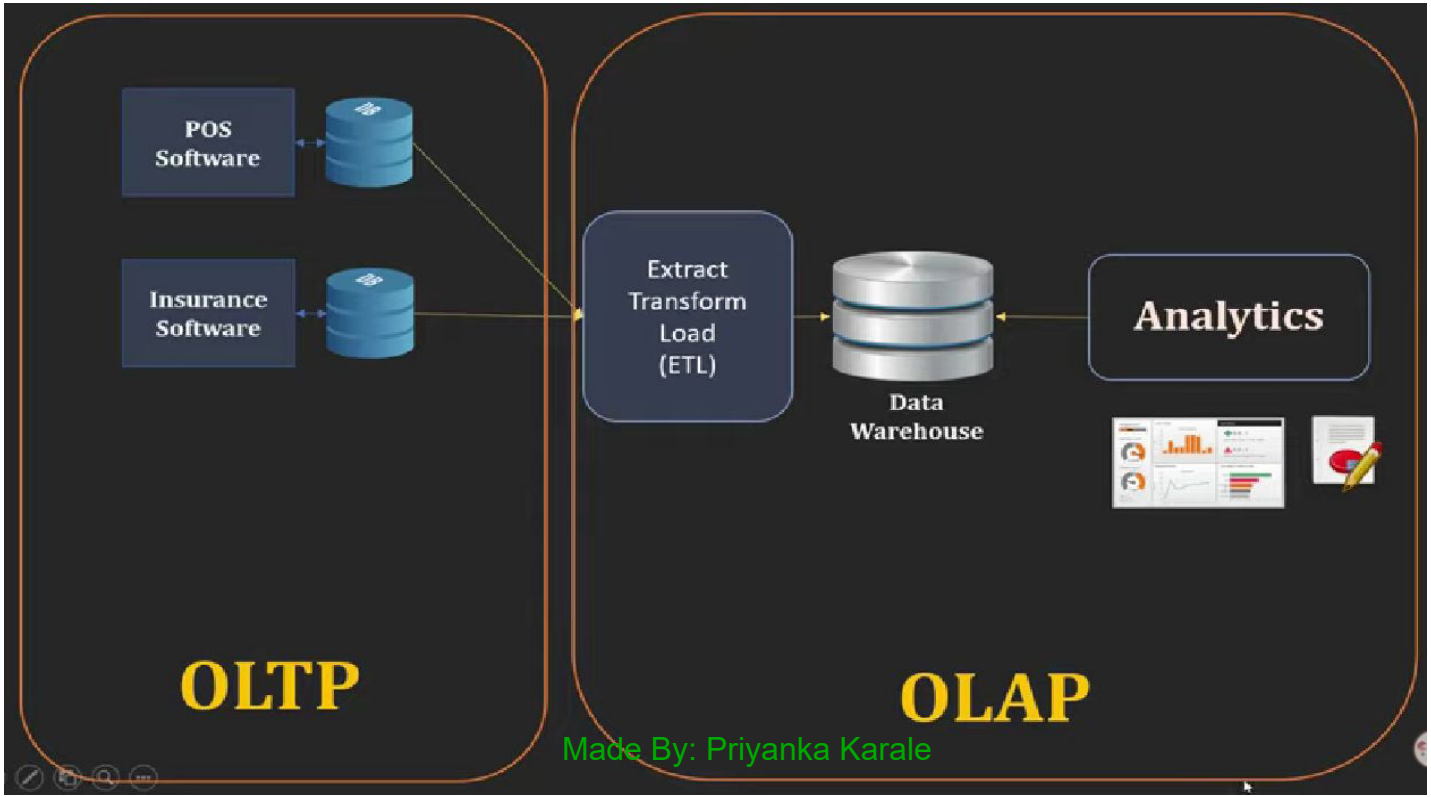
OLAP and OLTP Systems

- OLAP : On Line Analytical Processing

- OLTP : On Line Transaction Processing

Subject: MCA DMBI 1003

Comparison of OLTP and OLAP Systems



Made By: Priyanka Karale

OLTP Systems

- These systems are called **online transaction processing (OLTP)** systems.
- The major task of these systems is to perform online transaction and query processing.
- They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.
- An OLTP system is *customer-oriented.*
- An OLTP system manages current data that, typically, are too detailed to be easily used for decision making.

OLAP Systems

- These systems are known as **online analytical processing (OLAP)** systems.
- Data warehouse systems serve users or knowledge workers in the role of data analysis and decision making.
- Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users.
- An OLAP system is *market-oriented*.
- An OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity.

OLTP Systems

- An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
- An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historic data or data in different organizations.
- The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.

OLAP Systems

- An OLAP system typically adopts either a *star* or a *snowflake* model and a subject-oriented database design.
- An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization.
- OLAP systems also deal with information that originates from different organizations, integrating information from many data stores.
- Because of their huge volume, OLAP data are stored on multiple storage media.
- Access to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information).

Comparison of OLTP and OLAP Systems

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query

Feature	OLTP	OLAP
Access	read/write	mostly read
Focus	data in	information out
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data Cube: A Multidimensional Data Model

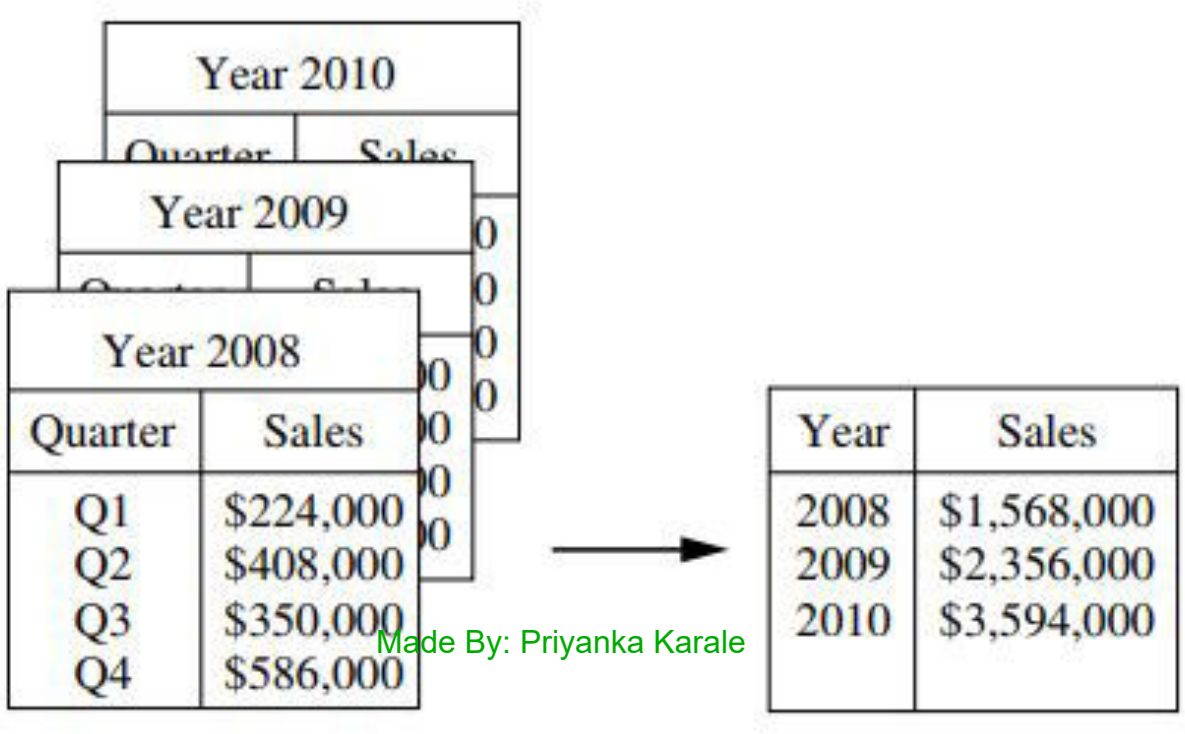
- Data warehouses and OLAP tools are based on a **multidimensional data model**. This model views data in the form of a ***data cube***.
- A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records.
- Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
- For example, a dimension table for *item* may contain the attributes *item name, brand*, and *type*.
- Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.

- A multidimensional data model is typically organized around a central theme, such as *sales*.

- Think of them as the quantities by which we want to analyze relationships between dimensions.

- Examples of facts for a sales data warehouse include *dollars sold* (sales amount in dollars), *units sold* (number of units sold), and *amount budgeted*.

- Imagine that you have collected the data for your analysis. These data consist of the *AllElectronics* sales per quarter, for the years 2008 to 2010.



2-D DATA CUBE

- In this 2-D representation, the sales for Vancouver are shown with respect to the *time* dimension (organized in quarters) and the *item* dimension (organized according to the types of items sold). The fact or measure displayed is *dollars sold* (in thousands).

- A simple 2-D data cube that is, in fact, a table or spreadsheet for sales data from *AllElectronics*.

- In particular, we will look at the A*llElectronics* sales data for items sold per quarter in the city of Vancouver.

location = "Vancouver"				
	item (type)			
time (quarter)	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

Note: The sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

3-D DATA CUBE

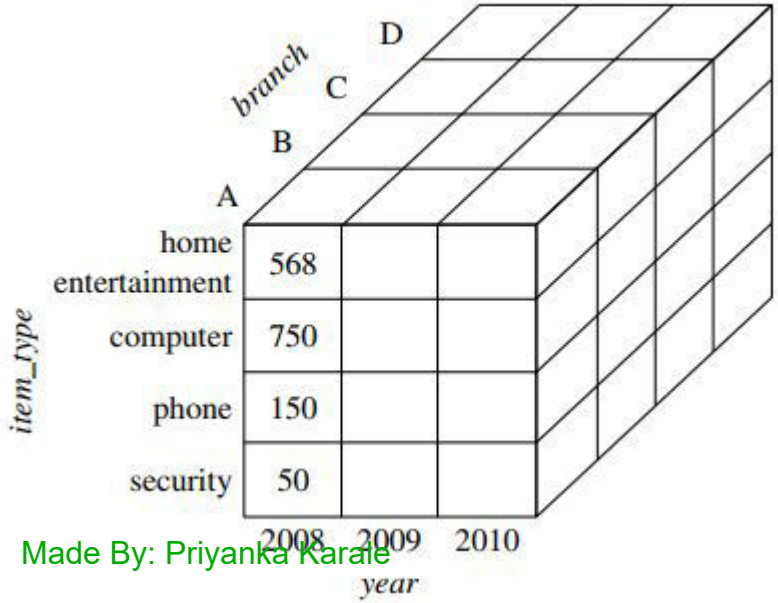
- suppose we would like to view the data according to *time* and *item*, as well as *location*, for the cities Chicago, New York, Toronto, and Vancouver.

- The 3-D data in the table are represented as a series of 2-D tables.

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*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

- Data cubes store multidimensional aggregated information.

- For example, Figure below shows a data cube for multidimensional analysis of sales data with respect to annual sales per item type for each *AllElectronics* branch. Each cell holds an aggregate data value, corresponding to the data point in multidimensional space.



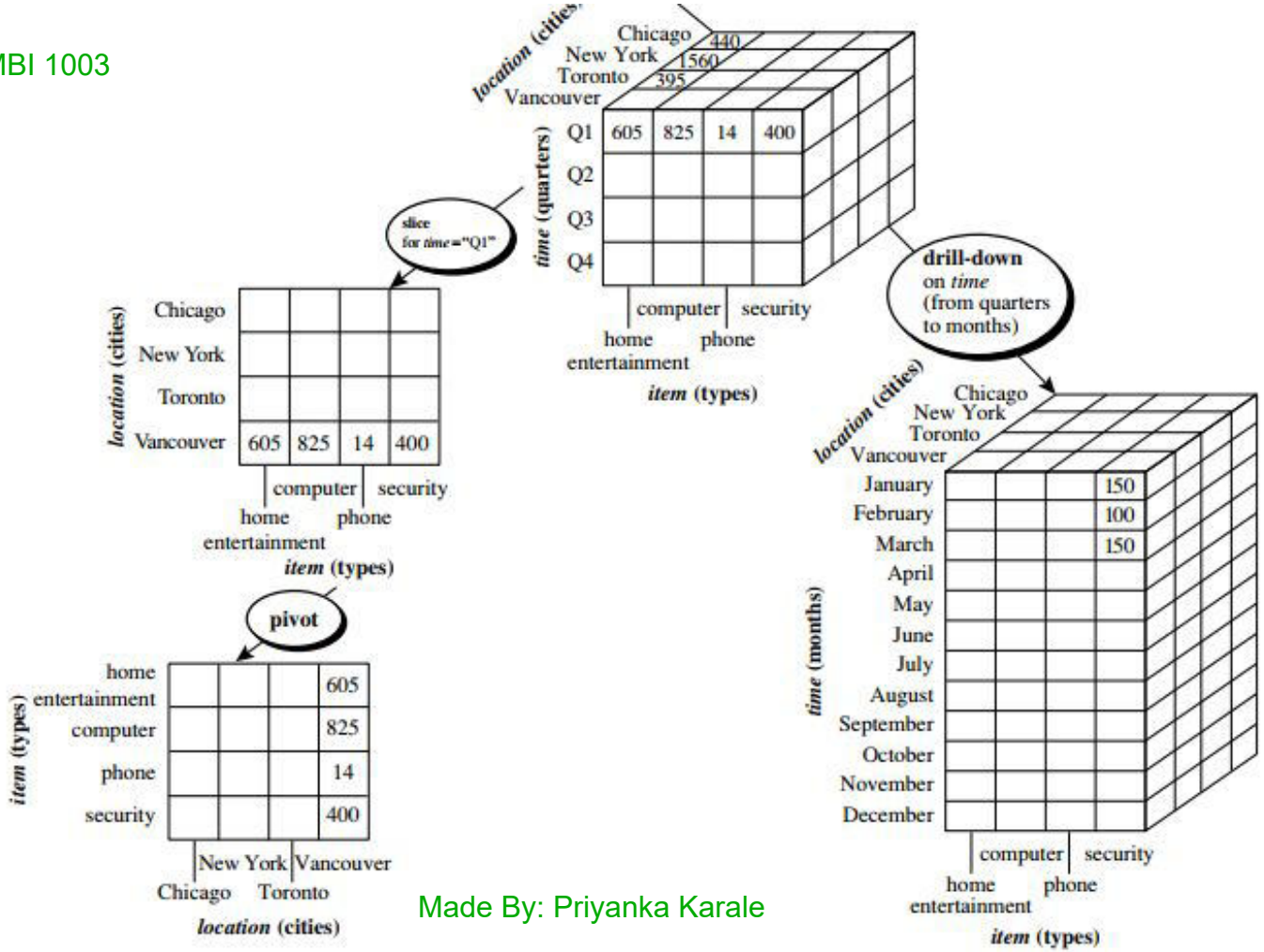
Subject: MCA DMBI 1003

Operations on Cubes

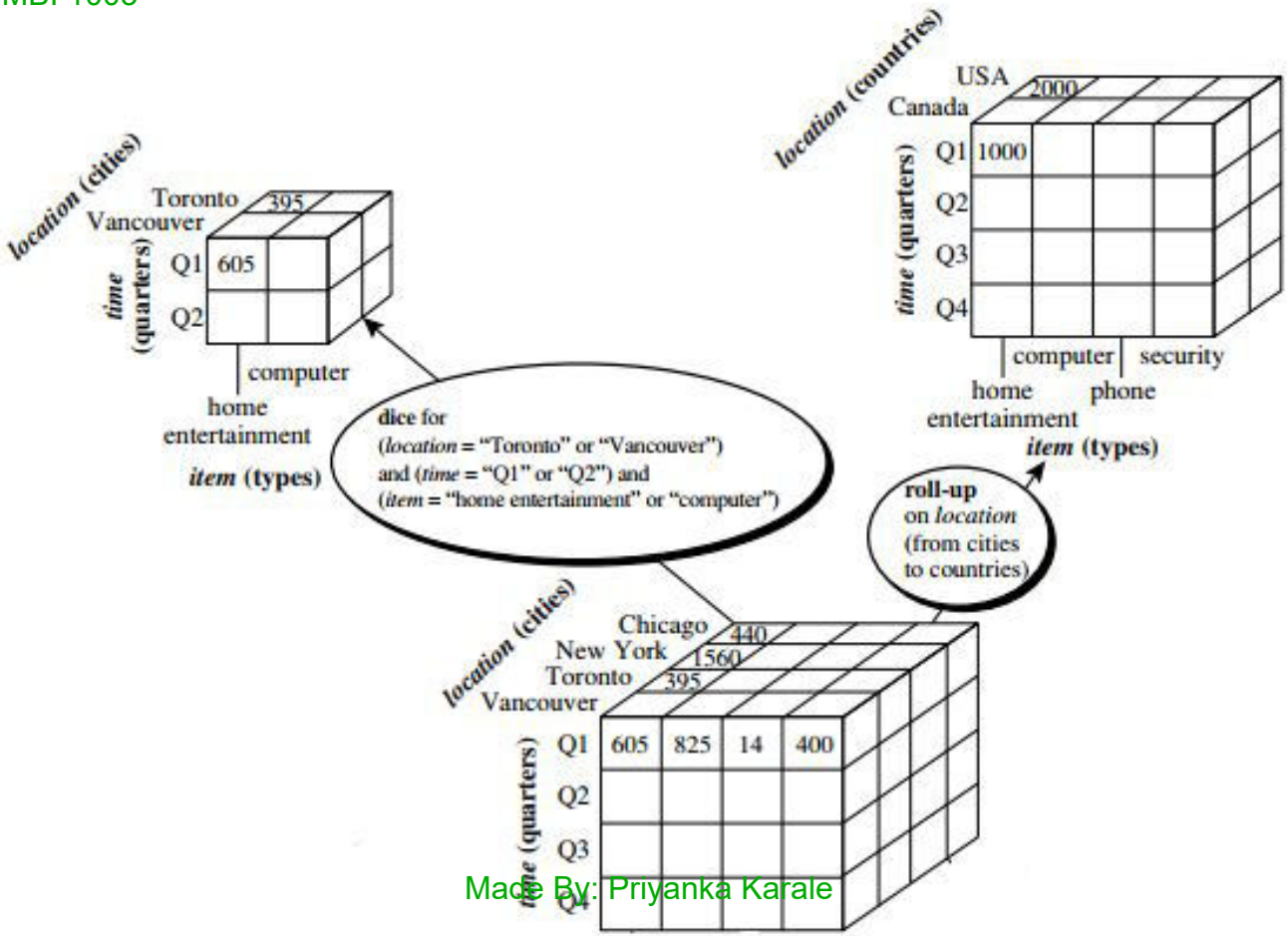
- Roll-Up
- Drill-down
- Slice & Dice
- Pivot

Made By: Priyanka Karale

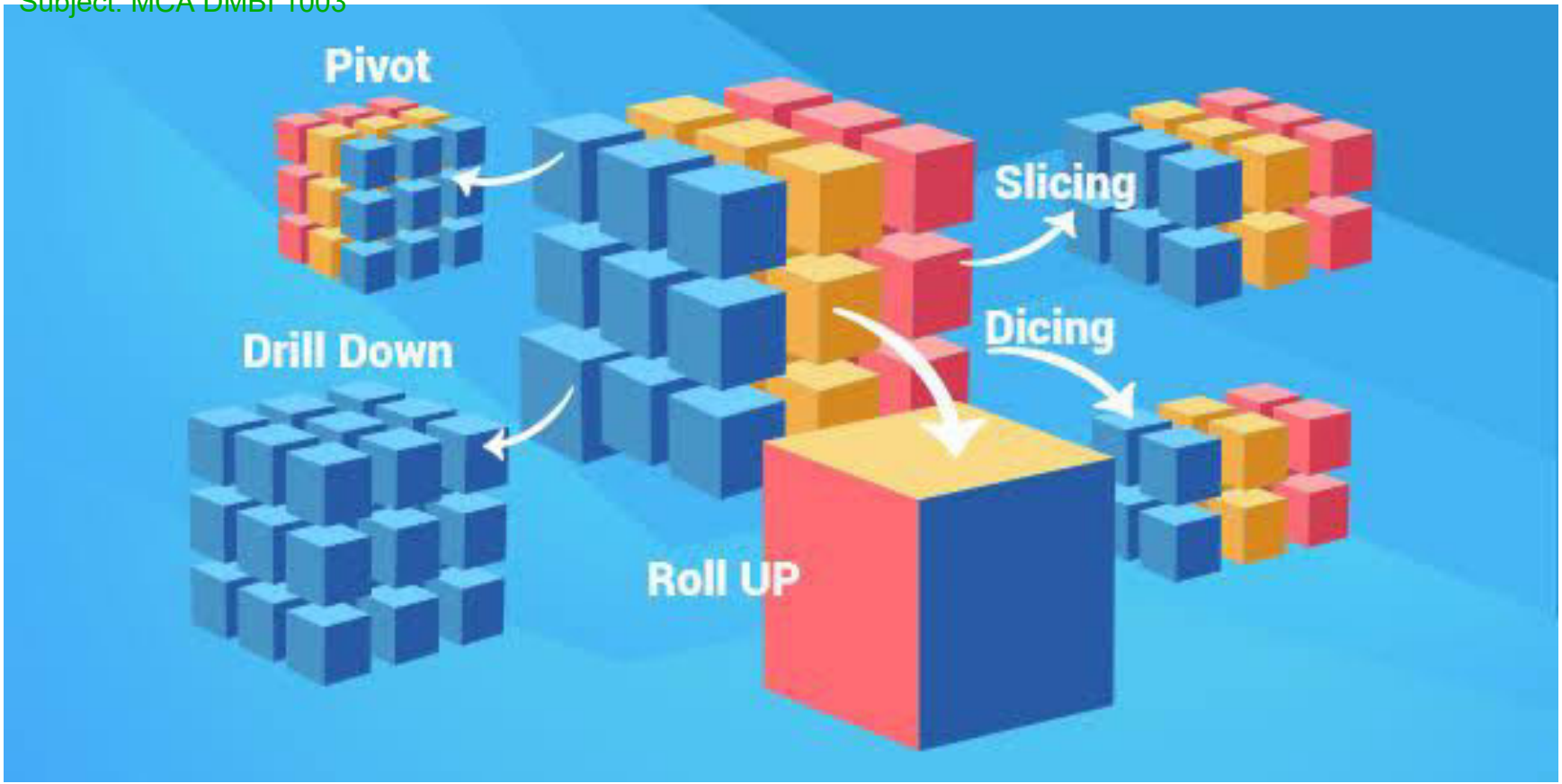
Subject: MCA DMBI 1003



Made By: Priyanka Karale

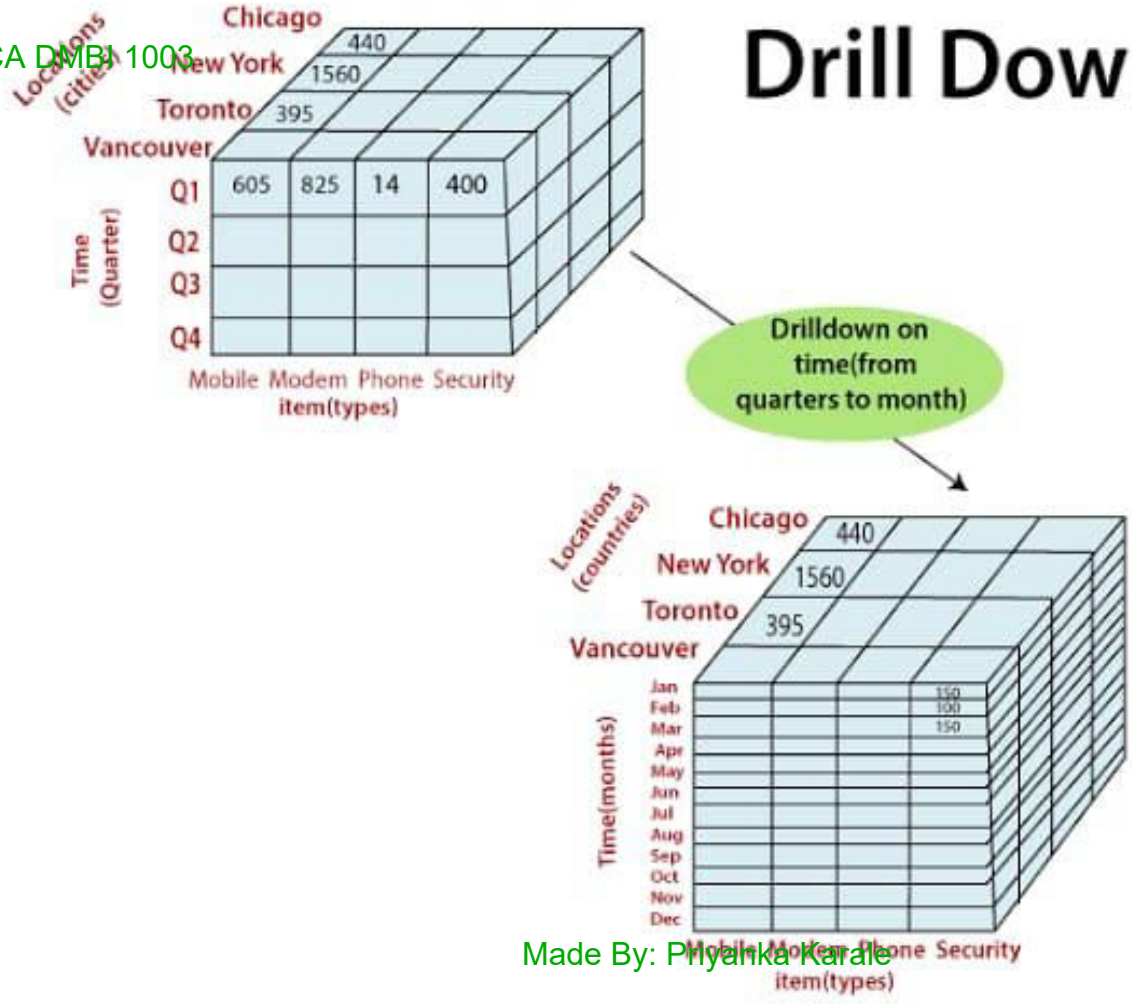

location (countries)
USA
Canada
2000
Q1 1000
Q2
Q3
Q4
time (quarters)
computer
security
home entertainment
phone
item (types)
location (cities)
Toronto
Vancouver
395
Q1 605
Q2
time (quarters)
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1 605 825 14 400
Q2
Q3
Q4
time (quarters)

Subject: MCA DMBI 1003

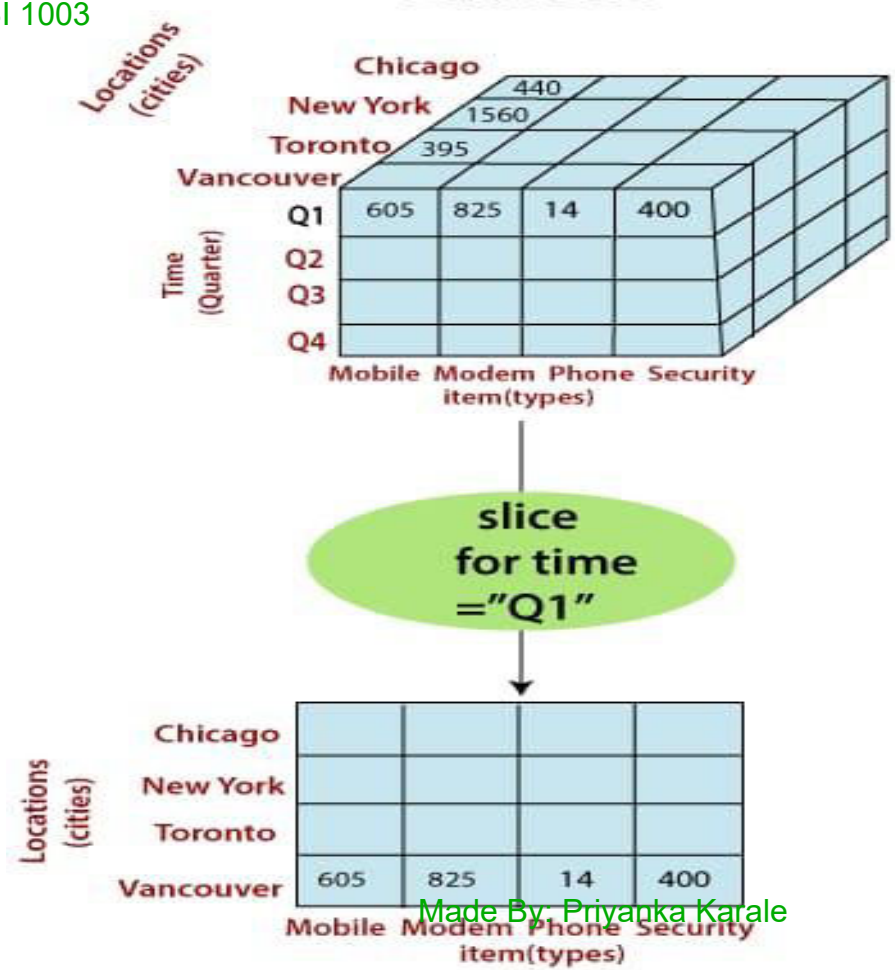


Made By: Priyanka Karale

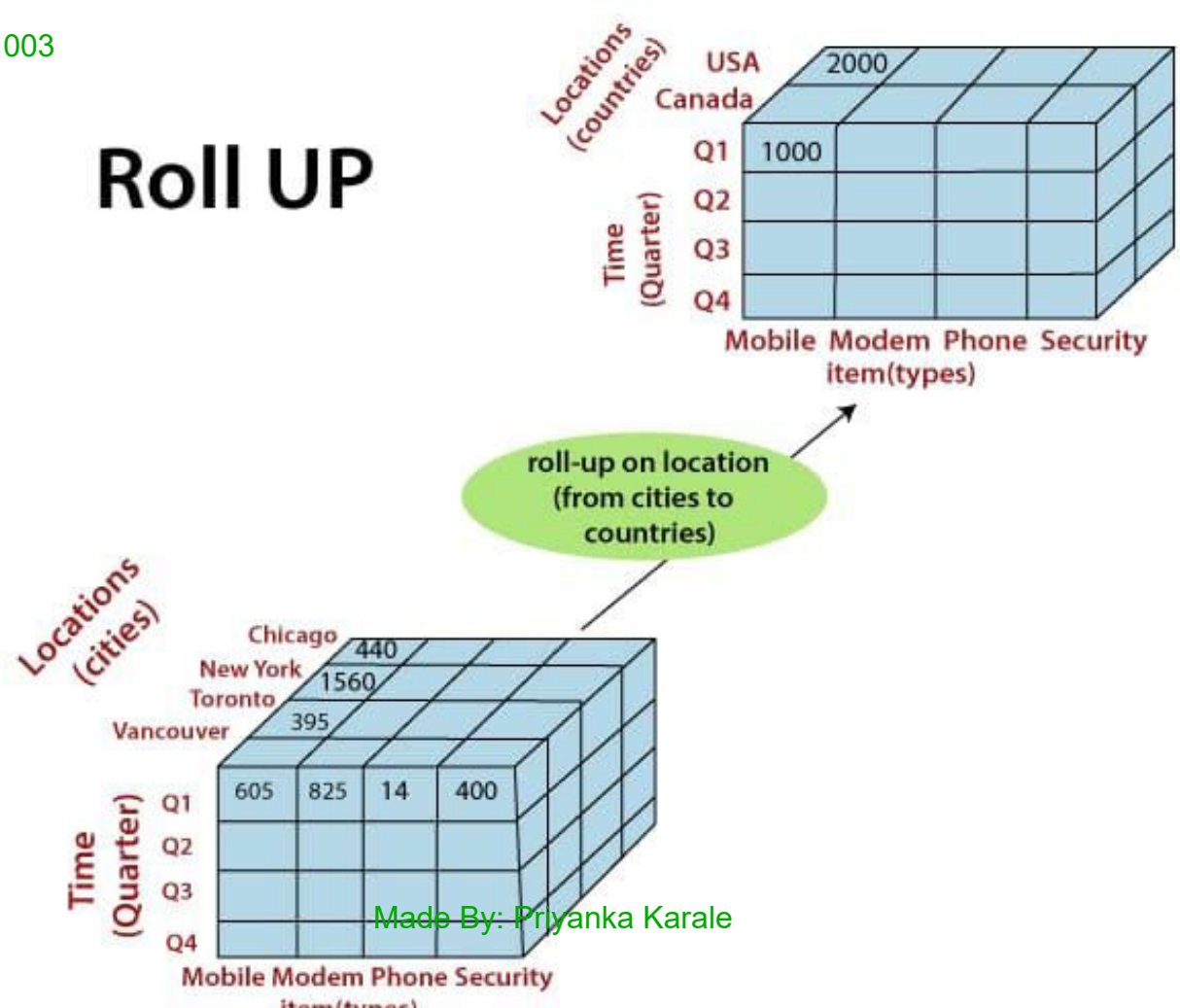
Drill Down


Chicago
New York
Toronto
Vancouver
Locations (cities)
440
1560
395
Q1
Q2
Q3
Q4
Time (Quarter)
605
825
14
400
Mobile Modem Phone Security
item(types)
Drilldown on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Time(months)
150
100
150
Mobile Modem Phone Security
item(types)

Slice

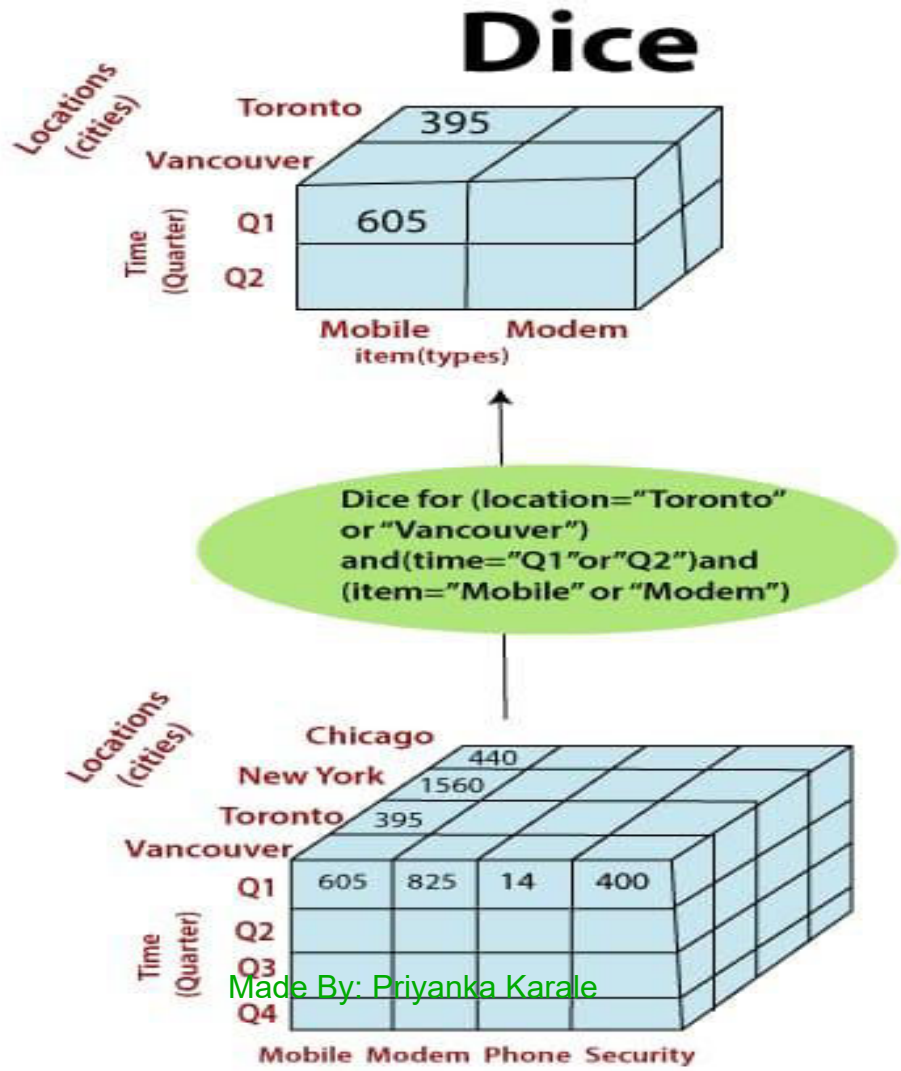

Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)
slice for time ="Q1"
Chicago
New York
Toronto
Vancouver
605
825
14
400
Locations (cities)
Mobile Modem Phone Security
item(types)

Roll UP

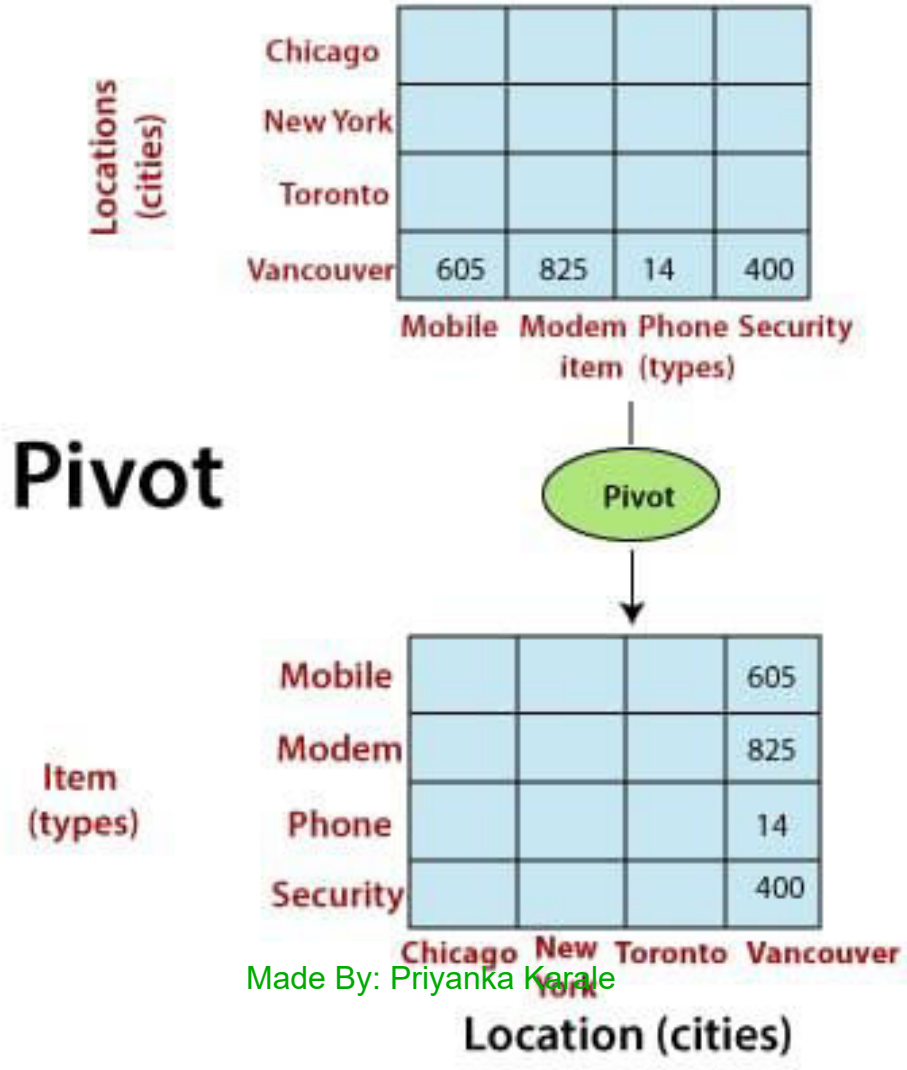


Made By: Priyanka Karale

Subject: MCA DMBI 1003



Made By: Priyanka Karale


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile
Modem
Phone
Security
item (types)
Pivot
Pivot
Mobile
Modem
Phone
Security
605
825
14
400
Item (types)
Chicago
New York
Toronto
Vancouver
Location (cities)

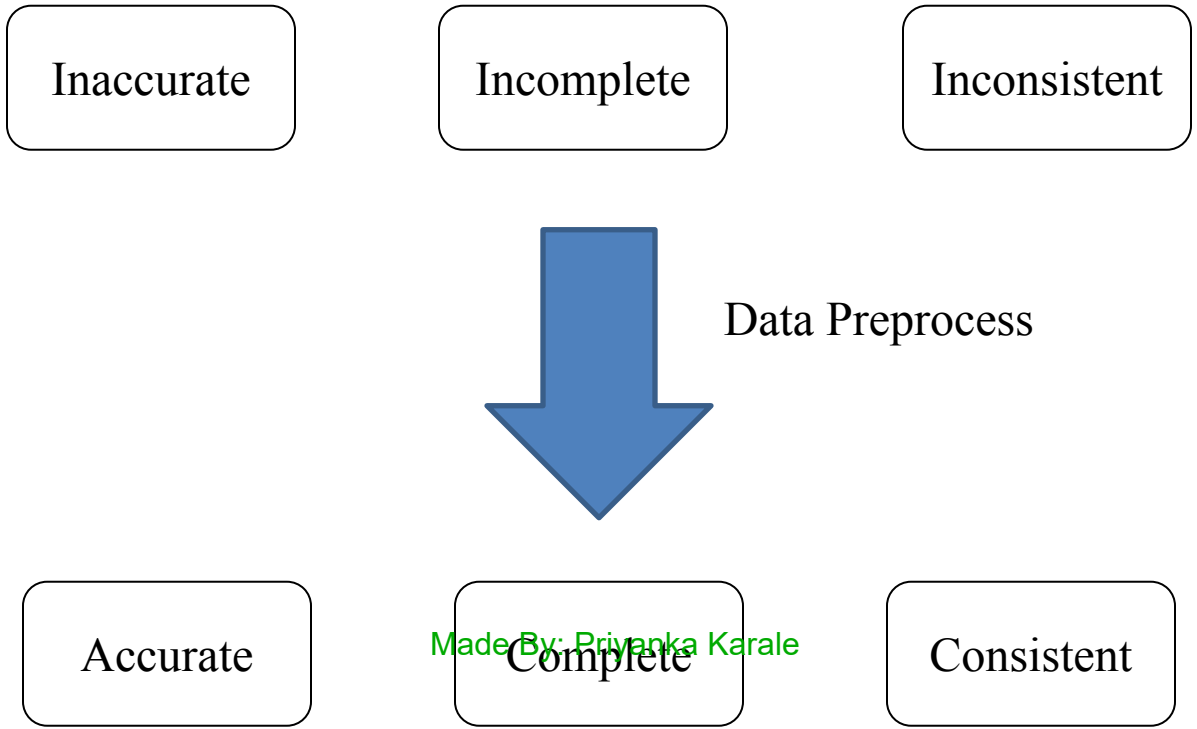
Data Preprocessing

What is the aim?

- *How can the data be preprocessed in order to help improve the quality of the data?*
- *How can the data be preprocessed in order to help improve the mining results?*
- *How can the data be preprocessed so as to improve the efficiency and ease of the mining process?*

Why we need Data Pre-processing? (Overview)

- In large real-world databases, three of the most common problems are:



Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
- A multi-dimensional measure of data quality:
  - A well-accepted multi-dimensional view:
    - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility
  - Broad categories:
    - intrinsic, contextual, representational, and accessibility.

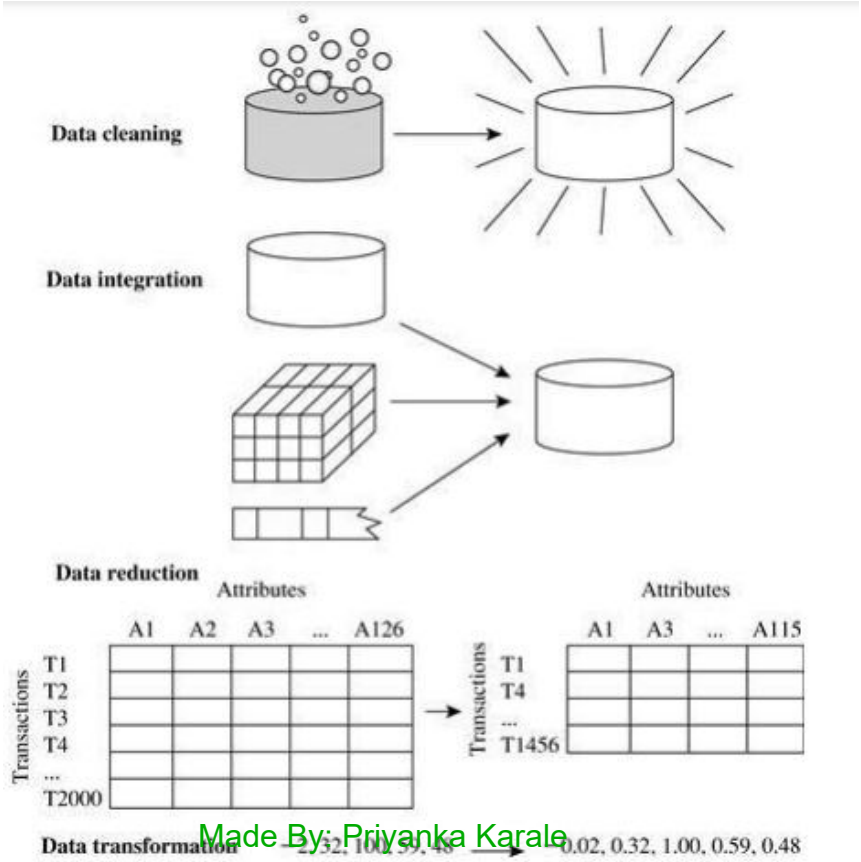
Why there are data problems?

- There are many possible reasons for inaccurate data (e.g, having incorrect attribute values)., or
- The instrument used for data collection may be faulty, there may be human or computer errors occurring at data entry.
- Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value "January 1" displayed for birthday). This is known as *disguised missing data. Errors in data transmission can also occur.*
- Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., *date). Duplicate tuples also require* data cleaning.
- Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions.

What are Data Pre-processing techniques/ tasks?

- ***Data cleaning*** *can be applied to* remove noise and correct inconsistencies in data.
- ***Data integration*** *merges data from* multiple sources into a coherent data store such as a data warehouse.
- ***Data transformations*** *(e.g., normalization) may be applied, where data are* scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements.
- ***Data reduction*** can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.
- These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a *date field* to a common format.

Forms of data preprocessing



Made By: Priyanka Karale

Data Cleaning

- Missing Values
    - Ignore the tuple
    - Fill in the missing values manually
    - Use a global constant to fill the missing values
    - Use mean or median
    - Use most probable values
        - Using regression
        - Bayesian formalism
        - Decision tree
- Noisy Data
    - Binning
    - Regression
    - Outlier Analysis

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Noisy Data

- Q: What is noise?
- A: Random error in a measured variable.
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for discretization (discussed later)
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression
  - smooth by fitting the data into regression functions

Binning Methods for Data Smoothing

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources,
  - Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units, different currency

Data Reduction

- Problem:

Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

Data Reduction Strategies

- Dimensionality reduction
  - Wavelet transform
  - Attribute Subset Selection
    - SWFS(Stepwise forward selection)
    - SWBE(Stepwise backward elimination)
  - Principal components analysis
- Numerosity reduction
  - parametric methods
    - Regression and log-linear models
  - Nonparametric methods
    - Histogram
    - Clustering
    - Sampling
    - Data cube aggregation

Made By: Priyanka Karale

Data Reduction Strategies

- Data Compression
  - Lossless
  - Lossy

Made By: Priyanka Karale

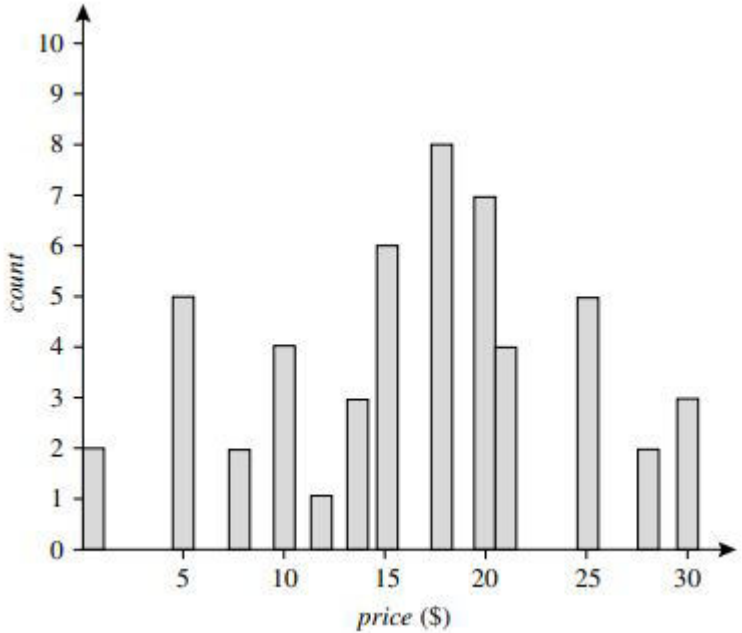
SWFS & SWBE

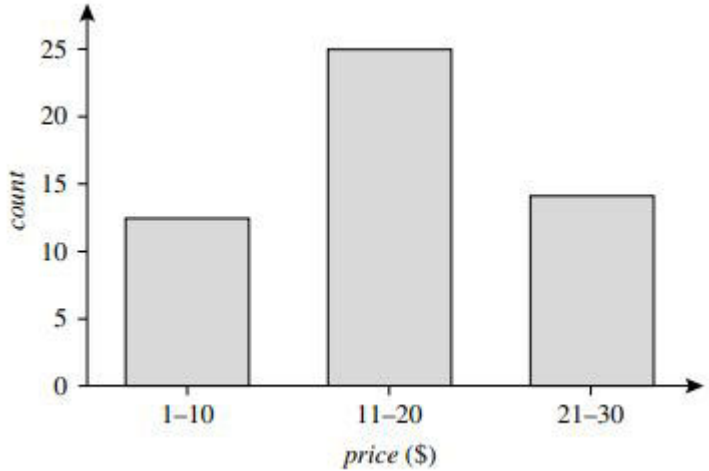
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $=> \{A_1\}$ $=> \{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $=> \{A_1, A_3, A_4, A_5, A_6\}$ $=> \{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$

Made By: Priyanka Karale

Subject: MCA DMBI 1003

Histogram





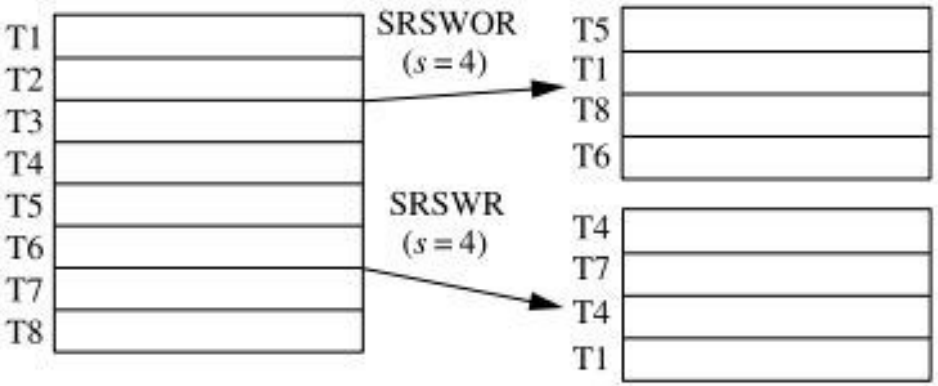
Made By: Priyanka Karale

Sampling

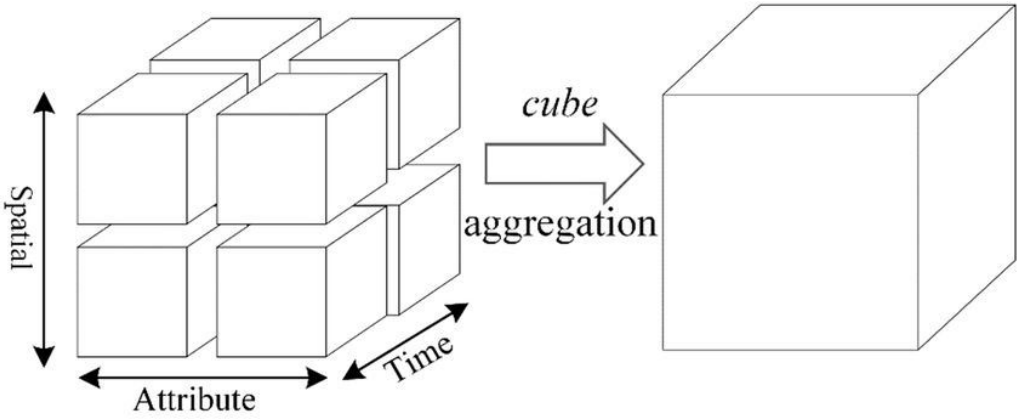
Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset). Suppose that a large data set, D, contains N tuples.

1: Simple random sample without replacement (SRSWOR) of size s: This is created by drawing s of the N tuples from D (s < N), where the probability of drawing any tuple in D is 1/N, that is, all tuples are equally likely to be sampled.

2: Simple random sample with replacement (SRSWR) of size s: This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.


T1
T2
T3
T4
T5
T6
T7
T8
SRSWOR
(s = 4)
SRSWR
(s = 4)
T5
T1
T8
T6
T4
T7
T4
T1

Data Cube Aggregation



Year 2010

Year 2009

Year 2008

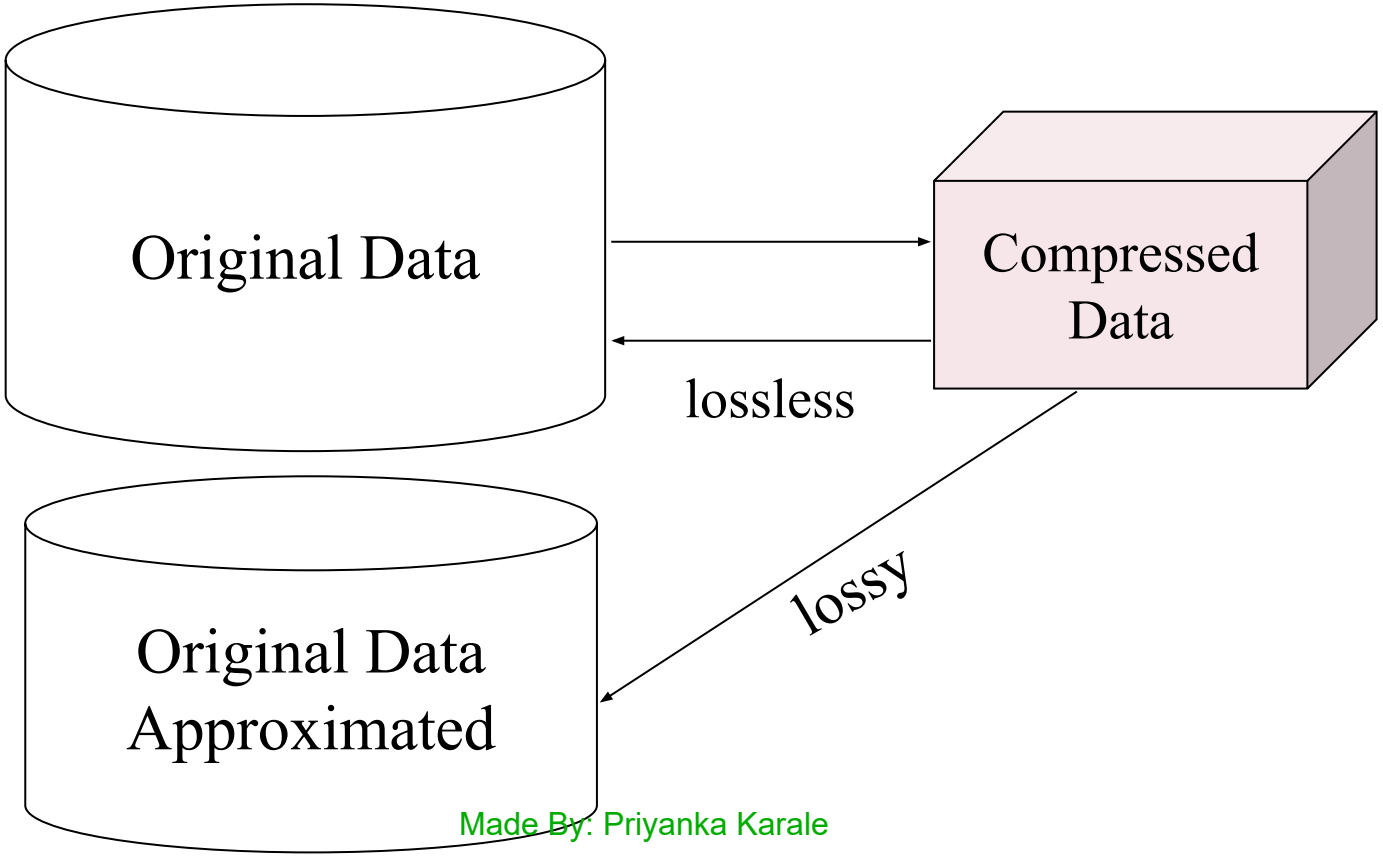
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

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

Made By: Priyanka Karale

Subject: MCA DMBI 1003

Data Compression



Made By: Priyanka Karale

Data Transformation Strategies

- 1. **Smoothing** , which works to remove noise from the data. Techniques include binning, regression, and clustering.

- 2. **Attribute construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

- 3. **Aggregation**, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.
- 4. Normalization, where the attribute data are scaled so as to fall within a smaller range, such as −1.0 to 1.0, or 0.0 to 1.0.

- 5. **Discretization,** where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute. Figure 3.12 shows a concept hierarchy for the attribute price. More than one concept hierarchy can be defined for the same attribute to accommodate the needs of various users.

- 6. **Concept hierarchy generation for nominal data,** where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.

Subject: MCA DMBI 1003

Thank You Unit 1 Ends Here

Made By: Priyanka Karale