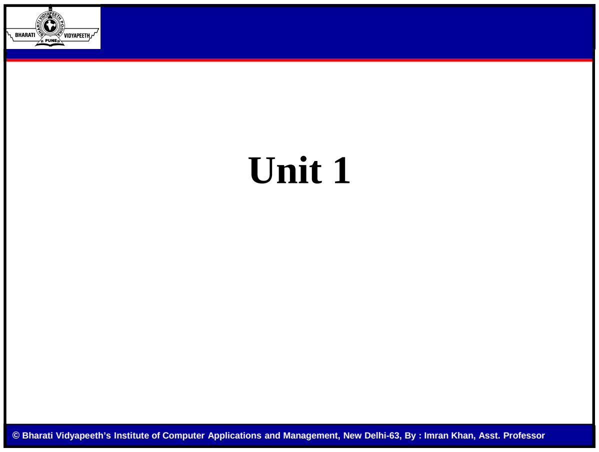
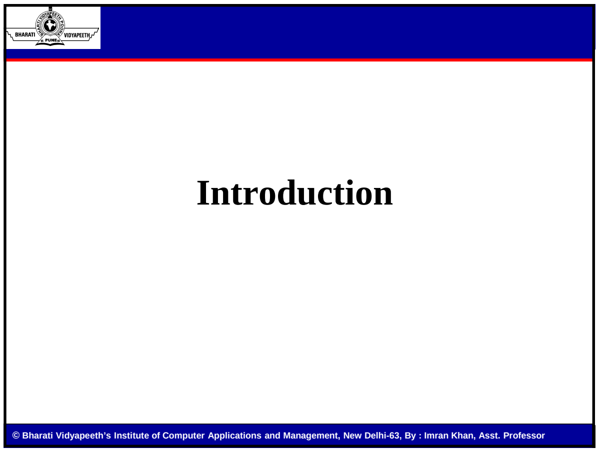


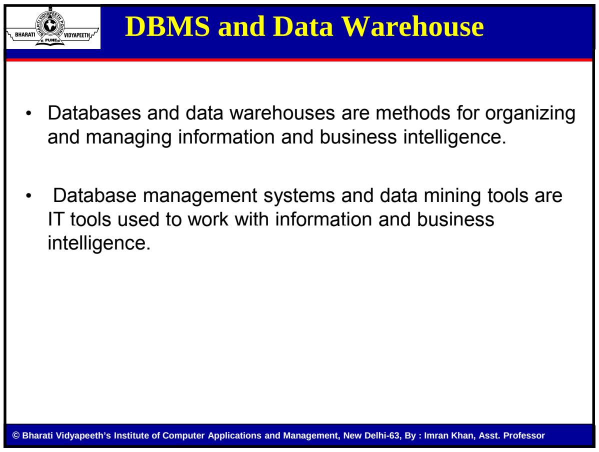
Unit 1

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



Introduction

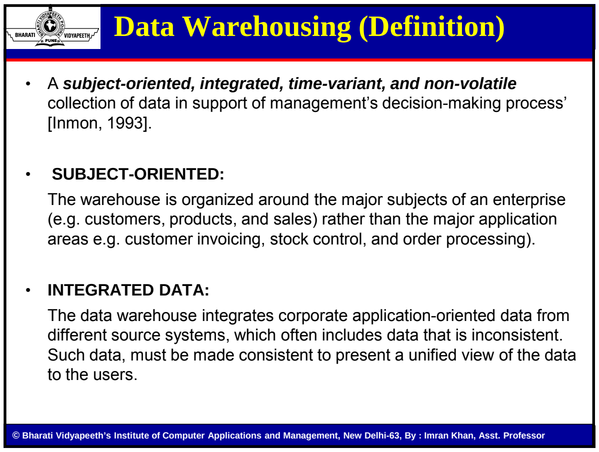
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



DBMS and Data Warehouse

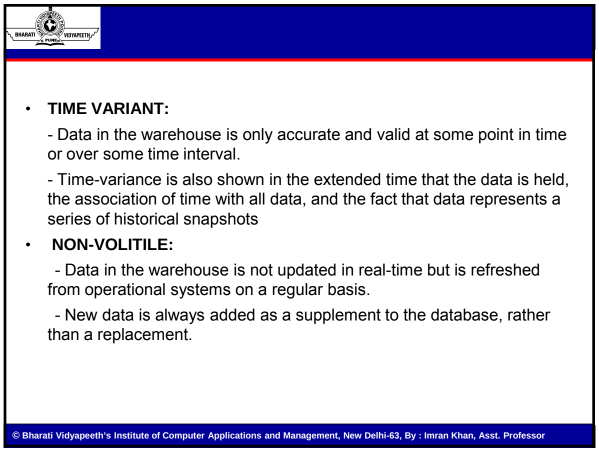
- Databases and data warehouses are methods for organizing and managing information and business intelligence.
- Database management systems and data mining tools are IT tools used to work with information and business intelligence.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor

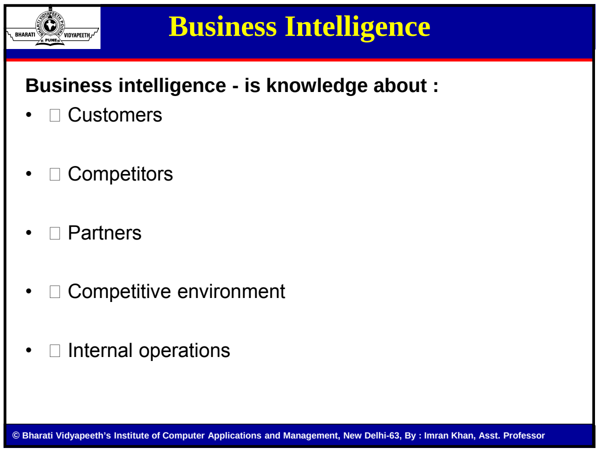


Data Warehousing (Definition)

- A ***subject-oriented, integrated, time-variant, and non-volatile*** collection of data in support of management's decision-making process' [Inmon, 1993].
- **SUBJECT-ORIENTED:**
  The warehouse is organized around the major subjects of an enterprise (e.g. customers, products, and sales) rather than the major application areas e.g. customer invoicing, stock control, and order processing).
- **INTEGRATED DATA:**
  The data warehouse integrates corporate application-oriented data from different source systems, which often includes data that is inconsistent. Such data, must be made consistent to present a unified view of the data to the users.



- **TIME VARIANT:**
  - Data in the warehouse is only accurate and valid at some point in time or over some time interval.
  - Time-variance is also shown in the extended time that the data is held, the association of time with all data, and the fact that data represents a series of historical snapshots
- **NON-VOLITILE:**
  - Data in the warehouse is not updated in real-time but is refreshed from operational systems on a regular basis.
  - New data is always added as a supplement to the database, rather than a replacement.



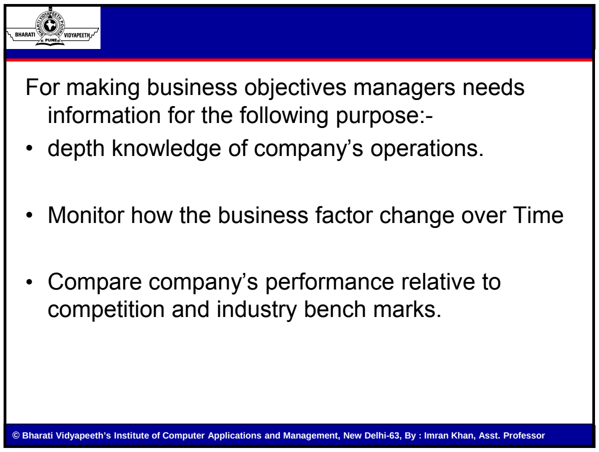
Business Intelligence

Business intelligence - is knowledge about :

- □ Customers
- □ Competitors
- □ Partners
- □ Competitive environment
- □ Internal operations

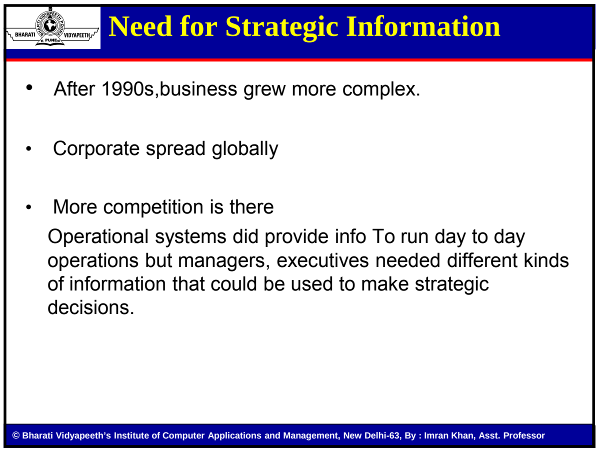
Some Business Objectives

- Retain the present customer base
- Increase the customer base by 15% over the next 5 years.
- Bring new product in 2 yrs
- Improve product quality levels in top 5 product group
- Gain market share by 10% in next 3 years
- Increase sale by 10% in East division



For making business objectives managers needs information for the following purpose:-

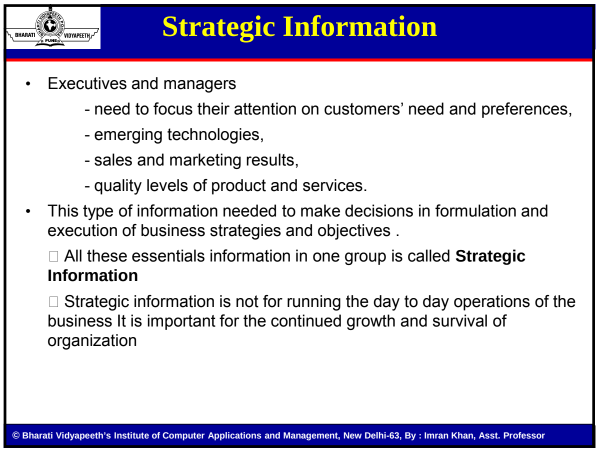
- depth knowledge of company's operations.
- Monitor how the business factor change over Time
- Compare company's performance relative to competition and industry bench marks.



Need for Strategic Information

- After 1990s,business grew more complex.
- Corporate spread globally
- More competition is there

Operational systems did provide info To run day to day operations but managers, executives needed different kinds of information that could be used to make strategic decisions.

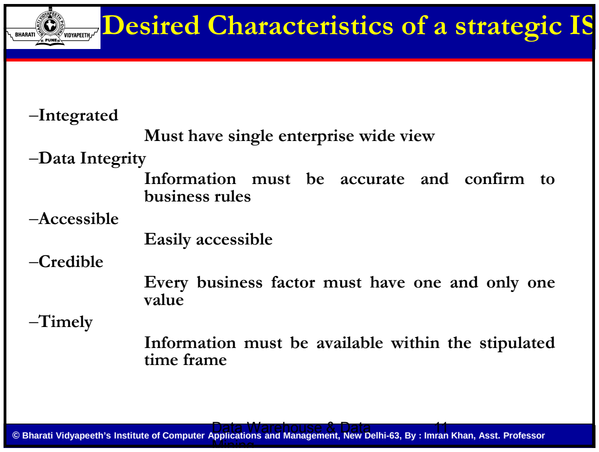


Strategic Information

- Executives and managers
  - need to focus their attention on customers' need and preferences,
  - emerging technologies,
  - sales and marketing results,
  - quality levels of product and services.
- This type of information needed to make decisions in formulation and execution of business strategies and objectives .

  □ All these essentials information in one group is called **Strategic Information**

  □ Strategic information is not for running the day to day operations of the business It is important for the continued growth and survival of organization



Desired Characteristics of a strategic IS

–Integrated

Must have single enterprise wide view

–Data Integrity

Information must be accurate and confirm to business rules

–Accessible

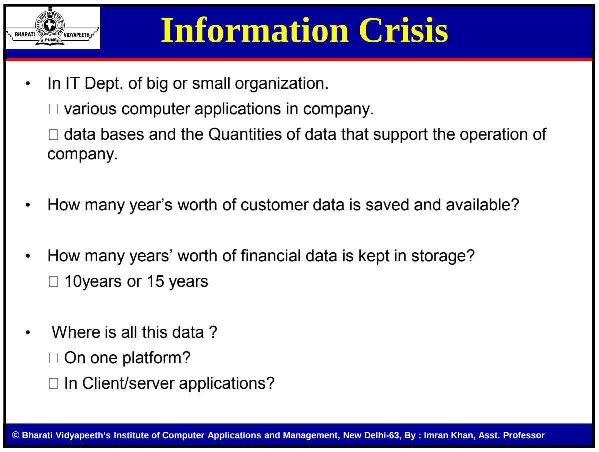
Easily accessible

–Credible

Every business factor must have one and only one value

–Timely

Information must be available within the stipulated time frame



Information Crisis

- In IT Dept. of big or small organization.

  □ various computer applications in company.

  □ data bases and the Quantities of data that support the operation of company.

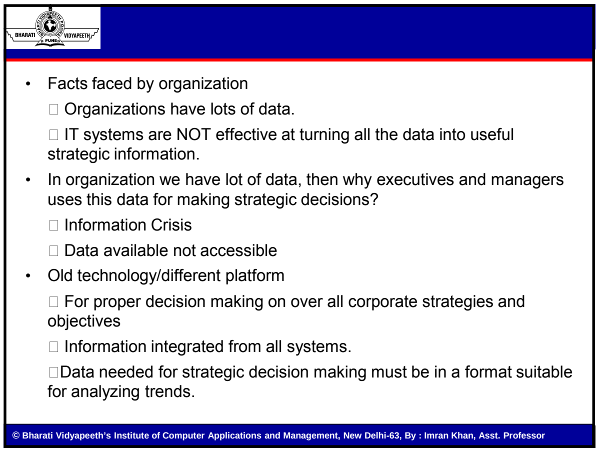
- How many year's worth of customer data is saved and available?
- How many years' worth of financial data is kept in storage?

  □ 10years or 15 years

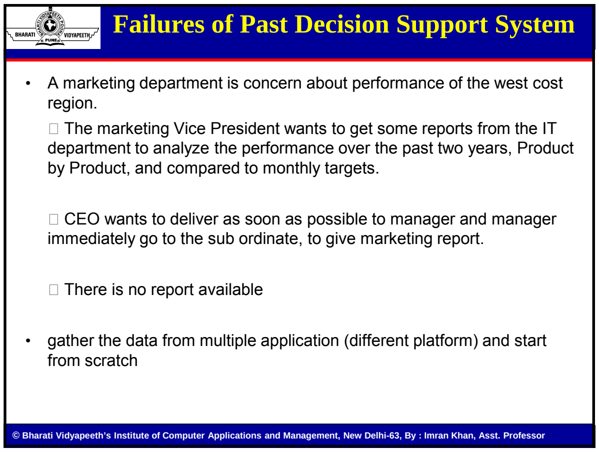
- Where is all this data ?

  □ On one platform?

  □ In Client/server applications?

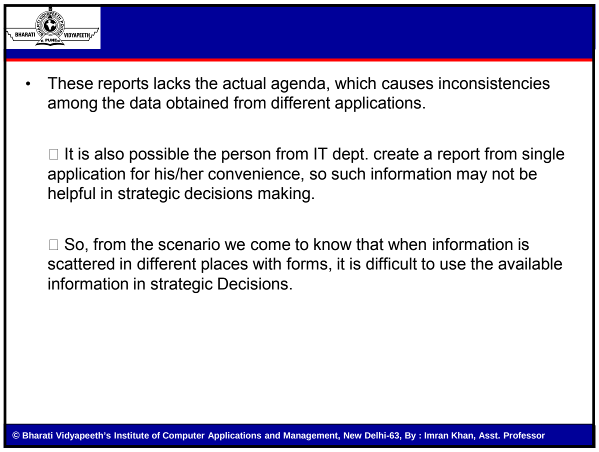


- Facts faced by organization
  - □ Organizations have lots of data.
  - □ IT systems are NOT effective at turning all the data into useful strategic information.
- In organization we have lot of data, then why executives and managers uses this data for making strategic decisions?
  - □ Information Crisis
  - □ Data available not accessible
- Old technology/different platform
  - □ For proper decision making on over all corporate strategies and objectives
  - □ Information integrated from all systems.
  - □Data needed for strategic decision making must be in a format suitable for analyzing trends.

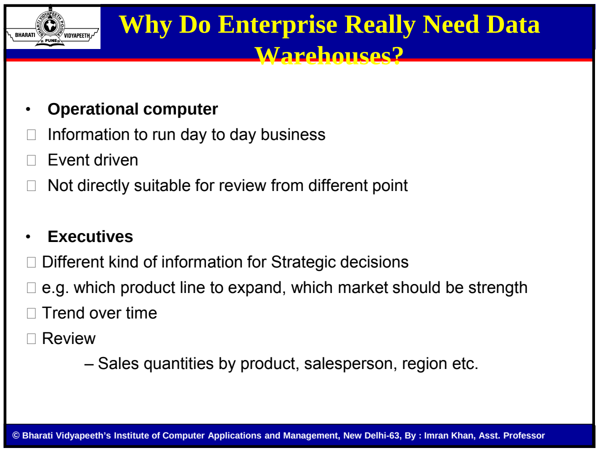


Failures of Past Decision Support System

- A marketing department is concern about performance of the west cost region.
  - □ The marketing Vice President wants to get some reports from the IT department to analyze the performance over the past two years, Product by Product, and compared to monthly targets.
  - □ CEO wants to deliver as soon as possible to manager and manager immediately go to the sub ordinate, to give marketing report.
  - □ There is no report available
- gather the data from multiple application (different platform) and start from scratch



- These reports lacks the actual agenda, which causes inconsistencies among the data obtained from different applications.
  - □ It is also possible the person from IT dept. create a report from single application for his/her convenience, so such information may not be helpful in strategic decisions making.
  - □ So, from the scenario we come to know that when information is scattered in different places with forms, it is difficult to use the available information in strategic Decisions.



Why Do Enterprise Really Need Data Warehouses?

- **Operational computer**

□ Information to run day to day business

□ Event driven

□ Not directly suitable for review from different point

- **Executives**

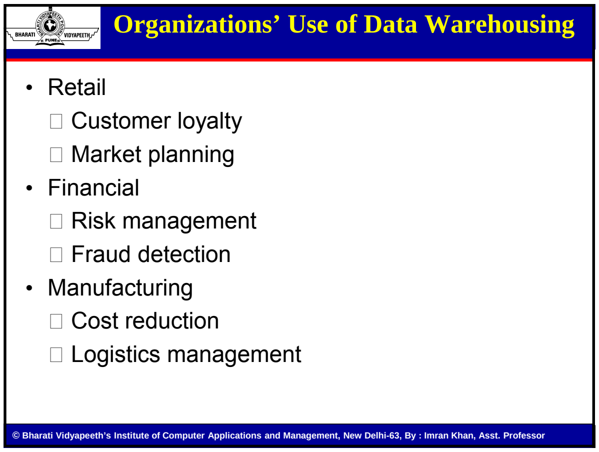
□ Different kind of information for Strategic decisions

□ e.g. which product line to expand, which market should be strength

□ Trend over time

□ Review

– Sales quantities by product, salesperson, region etc.



Organizations' Use of Data Warehousing

- Retail

  □ Customer loyalty

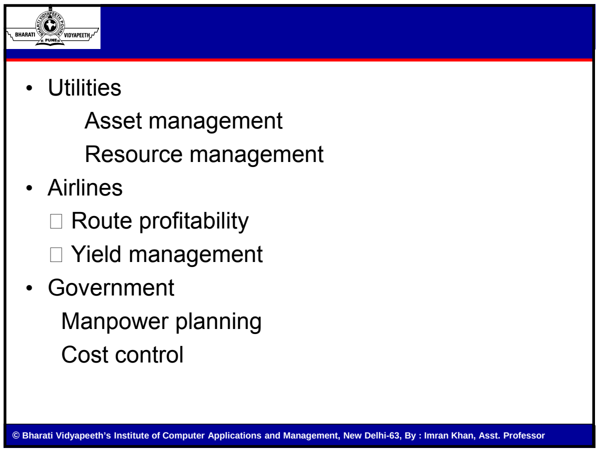
  □ Market planning
- Financial

  □ Risk management

  □ Fraud detection
- Manufacturing

  □ Cost reduction

  □ Logistics management



- Utilities

  Asset management

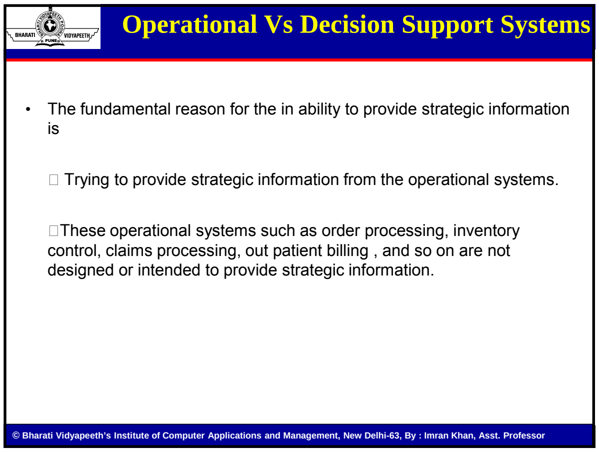
  Resource management
- Airlines

  □ Route profitability

  □ Yield management
- Government

  Manpower planning

  Cost control

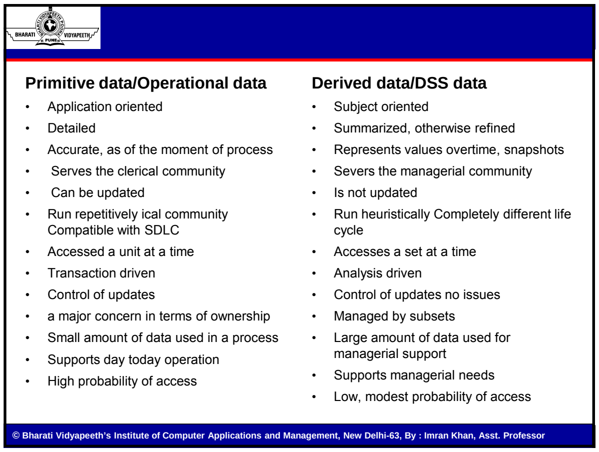


Operational Vs Decision Support Systems

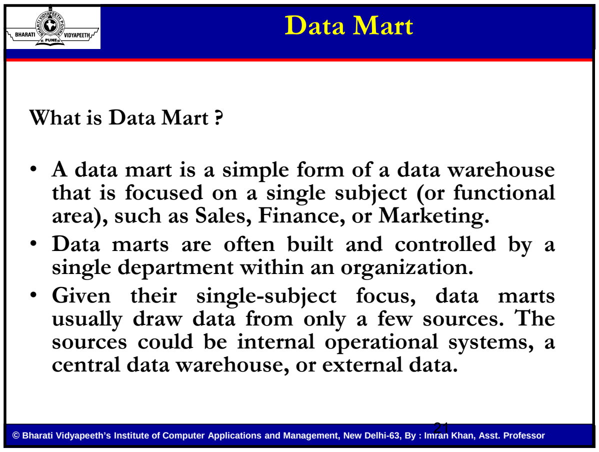
- The fundamental reason for the in ability to provide strategic information is

  □ Trying to provide strategic information from the operational systems.

  □These operational systems such as order processing, inventory control, claims processing, out patient billing , and so on are not designed or intended to provide strategic information.



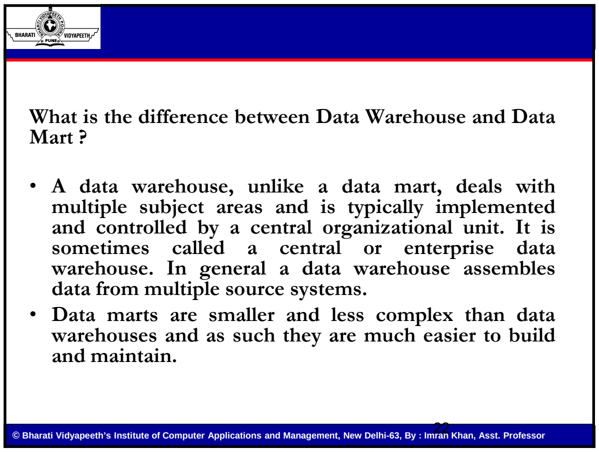
Primitive data/Operational data	Derived data/DSS data
Application oriented	Subject oriented
Detailed	Summarized, otherwise refined
Accurate, as of the moment of process	Represents values overtime, snapshots
Serves the clerical community	Severs the managerial community
Can be updated	Is not updated
Run repetitively ical community Compatible with SDLC	Run heuristically Completely different life cycle
Accessed a unit at a time	Accesses a set at a time
Transaction driven	Analysis driven
Control of updates	Control of updates no issues
a major concern in terms of ownership	Managed by subsets
Small amount of data used in a process	Large amount of data used for managerial support
Supports day today operation	Supports managerial needs
High probability of access	Low, modest probability of access



Data Mart

What is Data Mart ?

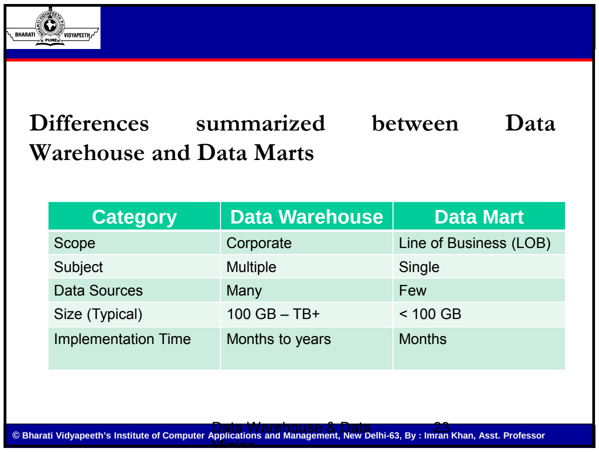
- A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as Sales, Finance, or Marketing.
- Data marts are often built and controlled by a single department within an organization.
- Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.



What is the difference between Data Warehouse and Data Mart ?

- A data warehouse, unlike a data mart, deals with multiple subject areas and is typically implemented and controlled by a central organizational unit. It is sometimes called a central or enterprise data warehouse. In general a data warehouse assembles data from multiple source systems.
- Data marts are smaller and less complex than data warehouses and as such they are much easier to build and maintain.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor

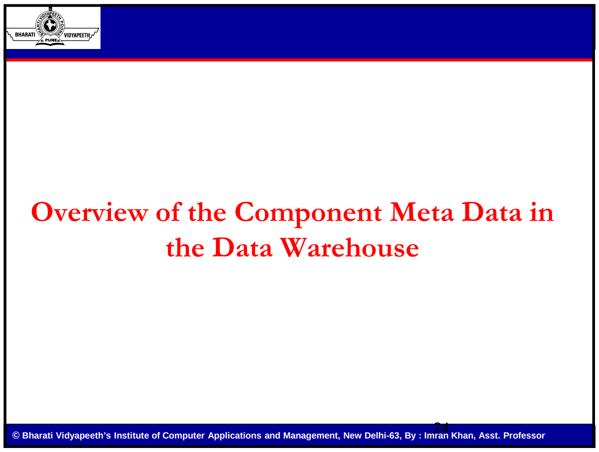


Differences summarized between Data Warehouse and Data Marts

Category	Data Warehouse	Data Mart
Scope	Corporate	Line of Business (LOB)
Subject	Multiple	Single
Data Sources	Many	Few
Size (Typical)	100 GB – TB+	< 100 GB
Implementation Time	Months to years	Months

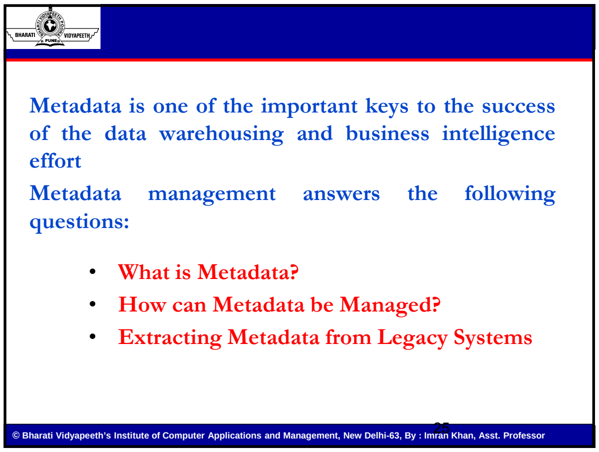
Data Warehouse & Data Mining 23

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



Overview of the Component Meta Data in the Data Warehouse

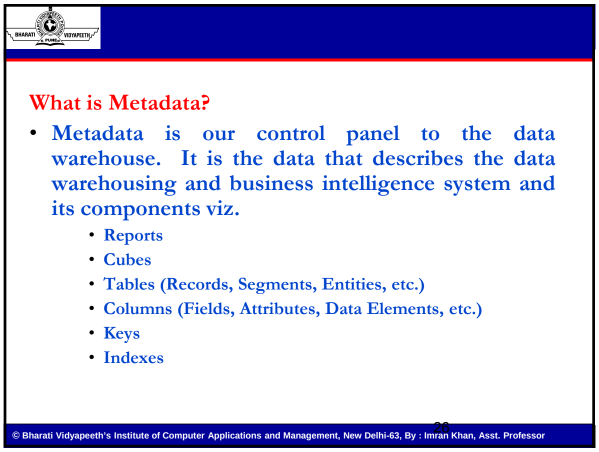
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



Metadata is one of the important keys to the success of the data warehousing and business intelligence effort

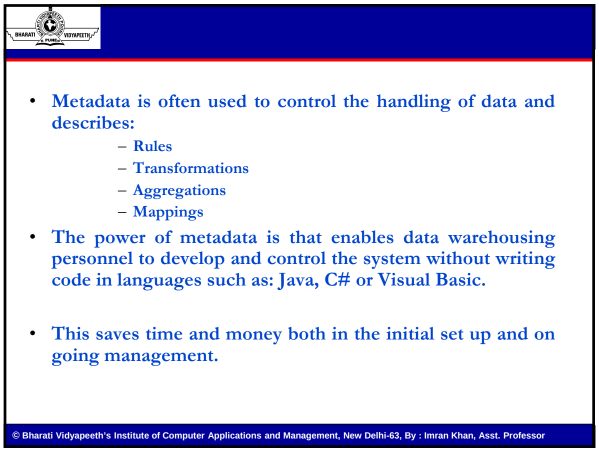
Metadata management answers the following questions:

- What is Metadata?
- How can Metadata be Managed?
- Extracting Metadata from Legacy Systems

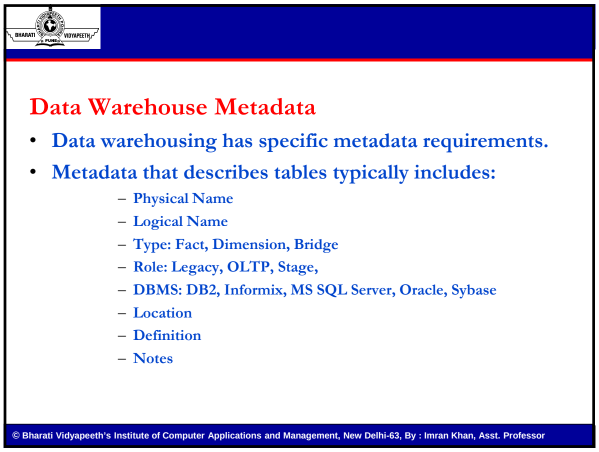


What is Metadata?

- Metadata is our control panel to the data warehouse. It is the data that describes the data warehousing and business intelligence system and its components viz.
  - Reports
  - Cubes
  - Tables (Records, Segments, Entities, etc.)
  - Columns (Fields, Attributes, Data Elements, etc.)
  - Keys
  - Indexes

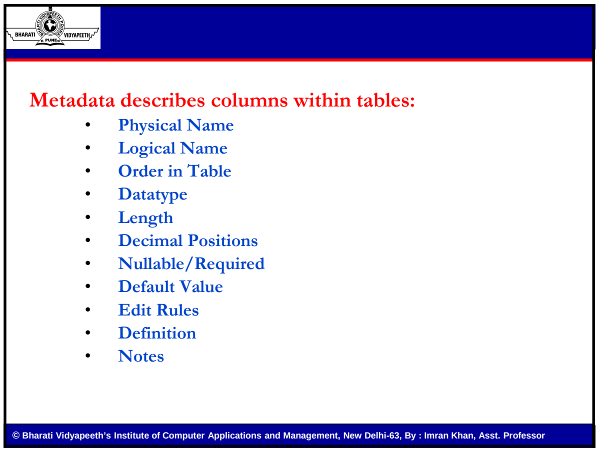


- Metadata is often used to control the handling of data and describes:
  - Rules
  - Transformations
  - Aggregations
  - Mappings
- The power of metadata is that enables data warehousing personnel to develop and control the system without writing code in languages such as: Java, C# or Visual Basic.
- This saves time and money both in the initial set up and on going management.



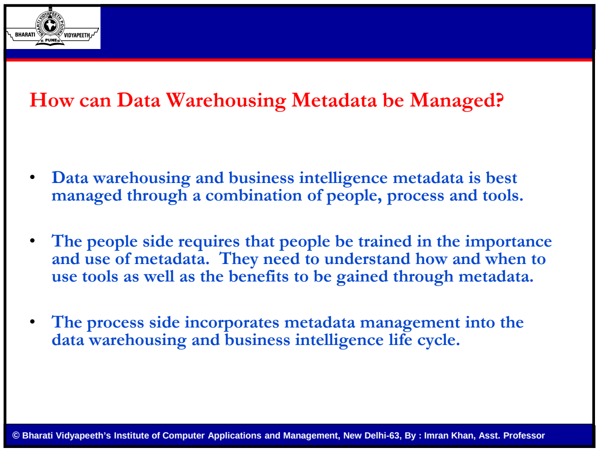
Data Warehouse Metadata

- Data warehousing has specific metadata requirements.
- Metadata that describes tables typically includes:
  - Physical Name
  - Logical Name
  - Type: Fact, Dimension, Bridge
  - Role: Legacy, OLTP, Stage,
  - DBMS: DB2, Informix, MS SQL Server, Oracle, Sybase
  - Location
  - Definition
  - Notes



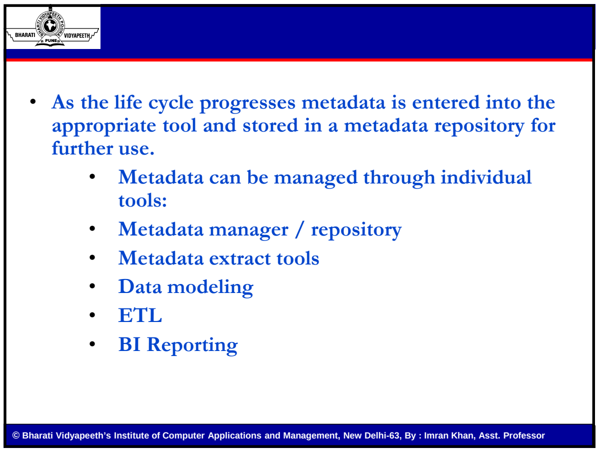
Metadata describes columns within tables:

- Physical Name
- Logical Name
- Order in Table
- Datatype
- Length
- Decimal Positions
- Nullable/Required
- Default Value
- Edit Rules
- Definition
- Notes

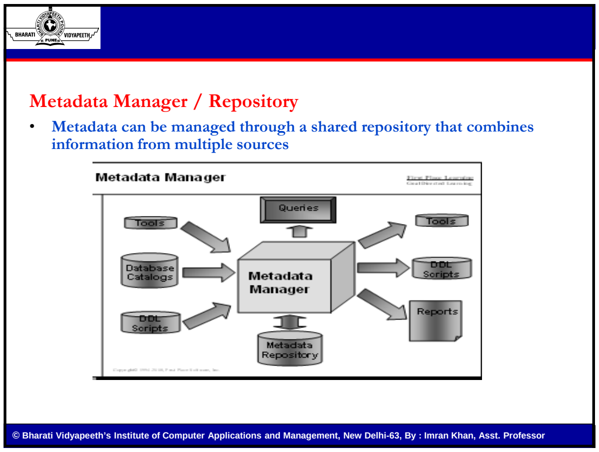


How can Data Warehousing Metadata be Managed?

- Data warehousing and business intelligence metadata is best managed through a combination of people, process and tools.
- The people side requires that people be trained in the importance and use of metadata. They need to understand how and when to use tools as well as the benefits to be gained through metadata.
- The process side incorporates metadata management into the data warehousing and business intelligence life cycle.

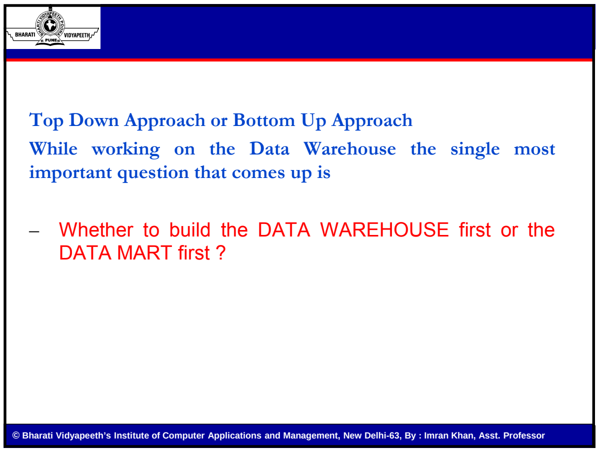


- As the life cycle progresses metadata is entered into the appropriate tool and stored in a metadata repository for further use.
  - Metadata can be managed through individual tools:
  - Metadata manager / repository
  - Metadata extract tools
  - Data modeling
  - ETL
  - BI Reporting



Metadata Manager / Repository

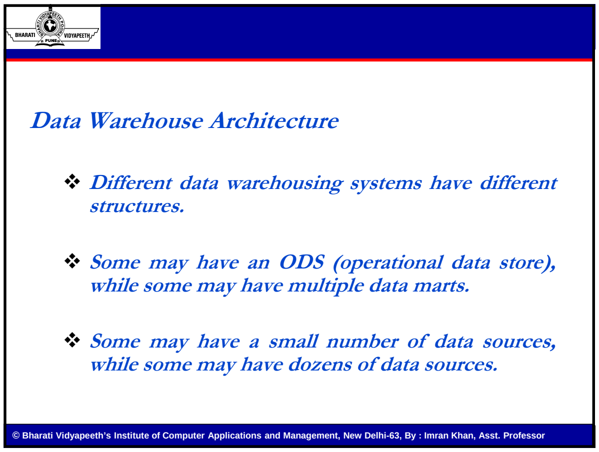
- Metadata can be managed through a shared repository that combines information from multiple sources



Top Down Approach or Bottom Up Approach

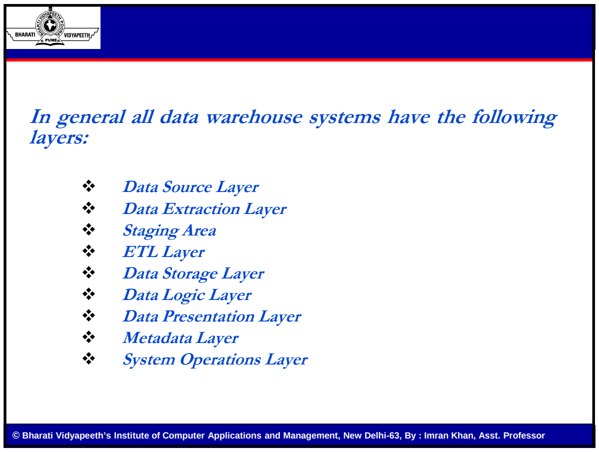
While working on the Data Warehouse the single most important question that comes up is

– Whether to build the DATA WAREHOUSE first or the DATA MART first ?



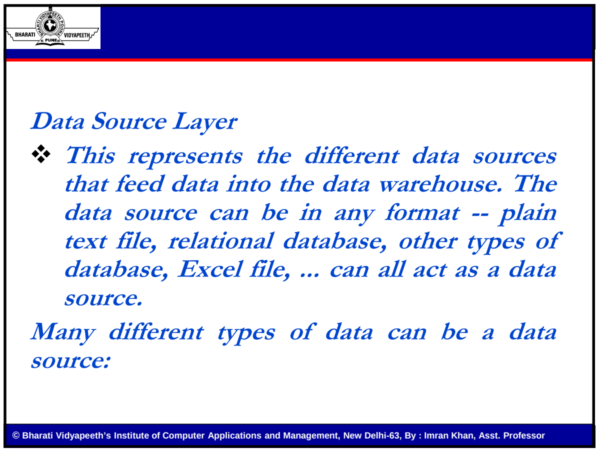
Data Warehouse Architecture

- *Different data warehousing systems have different structures.*
- *Some may have an ODS (operational data store), while some may have multiple data marts.*
- *Some may have a small number of data sources, while some may have dozens of data sources.*



In general all data warehouse systems have the following layers:

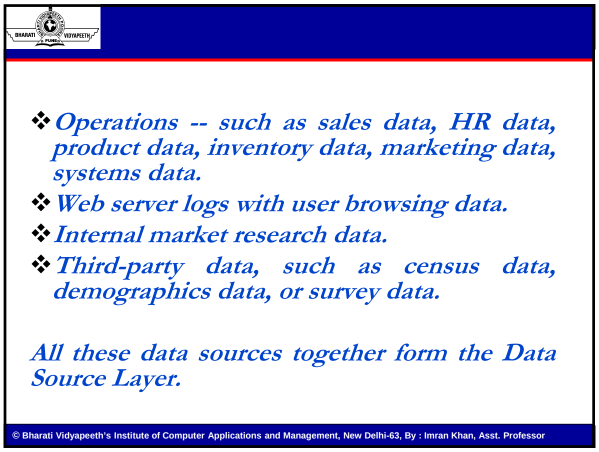
- *Data Source Layer*
- *Data Extraction Layer*
- *Staging Area*
- *ETL Layer*
- *Data Storage Layer*
- *Data Logic Layer*
- *Data Presentation Layer*
- *Metadata Layer*
- *System Operations Layer*



Data Source Layer

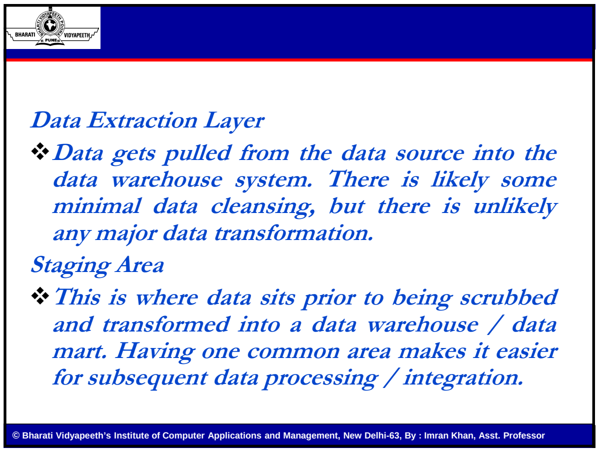
- *This represents the different data sources that feed data into the data warehouse. The data source can be in any format -- plain text file, relational database, other types of database, Excel file, ... can all act as a data source.*

Many different types of data can be a data source:



- ❖ *Operations -- such as sales data, HR data, product data, inventory data, marketing data, systems data.*
- ❖ *Web server logs with user browsing data.*
- ❖ *Internal market research data.*
- ❖ *Third-party data, such as census data, demographics data, or survey data.*

All these data sources together form the Data Source Layer.

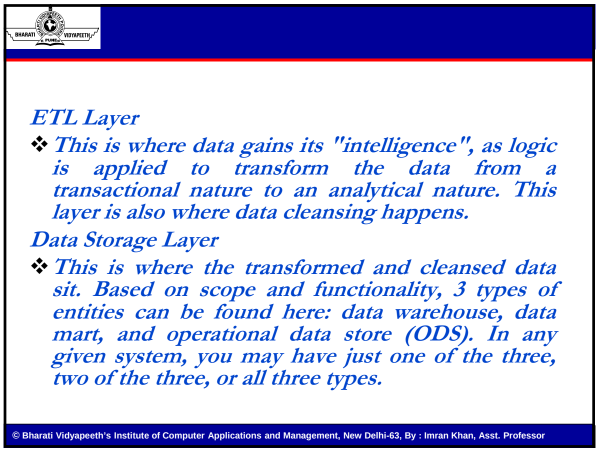


Data Extraction Layer

- ❖ *Data gets pulled from the data source into the data warehouse system. There is likely some minimal data cleansing, but there is unlikely any major data transformation.*

Staging Area

- ❖ *This is where data sits prior to being scrubbed and transformed into a data warehouse / data mart. Having one common area makes it easier for subsequent data processing / integration.*

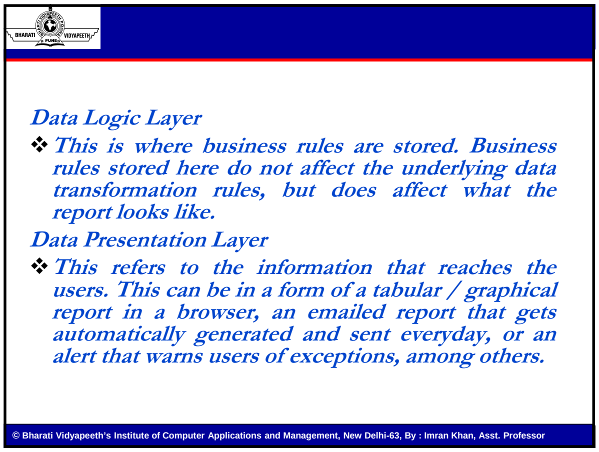


ETL Layer

- ❖ *This is where data gains its "intelligence", as logic is applied to transform the data from a transactional nature to an analytical nature. This layer is also where data cleansing happens.*

Data Storage Layer

- ❖ *This is where the transformed and cleansed data sit. Based on scope and functionality, 3 types of entities can be found here: data warehouse, data mart, and operational data store (ODS). In any given system, you may have just one of the three, two of the three, or all three types.*

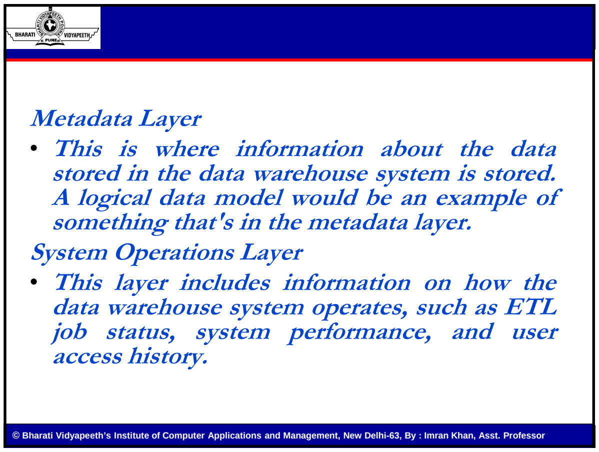


Data Logic Layer

- ❖ *This is where business rules are stored. Business rules stored here do not affect the underlying data transformation rules, but does affect what the report looks like.*

Data Presentation Layer

- ❖ *This refers to the information that reaches the users. This can be in a form of a tabular / graphical report in a browser, an emailed report that gets automatically generated and sent everyday, or an alert that warns users of exceptions, among others.*

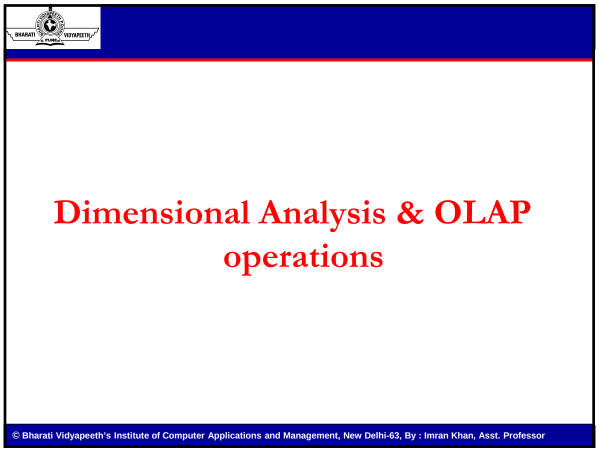


Metadata Layer

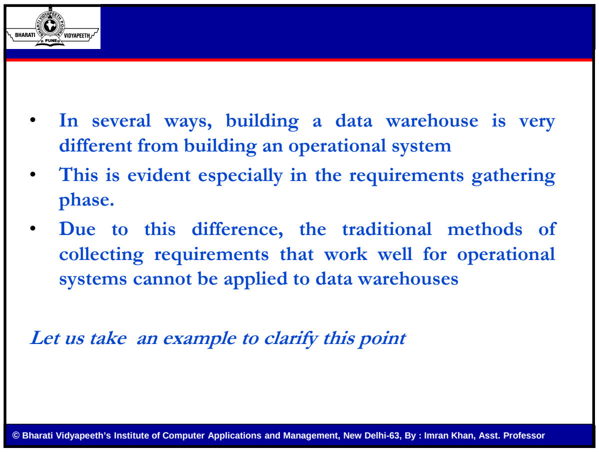
- *This is where information about the data stored in the data warehouse system is stored. A logical data model would be an example of something that's in the metadata layer.*

System Operations Layer

- *This layer includes information on how the data warehouse system operates, such as ETL job status, system performance, and user access history.*

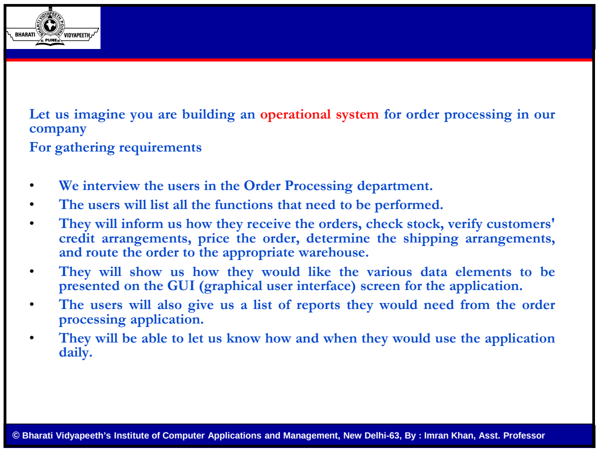


Dimensional Analysis & OLAP operations



- In several ways, building a data warehouse is very different from building an operational system
- This is evident especially in the requirements gathering phase.
- Due to this difference, the traditional methods of collecting requirements that work well for operational systems cannot be applied to data warehouses

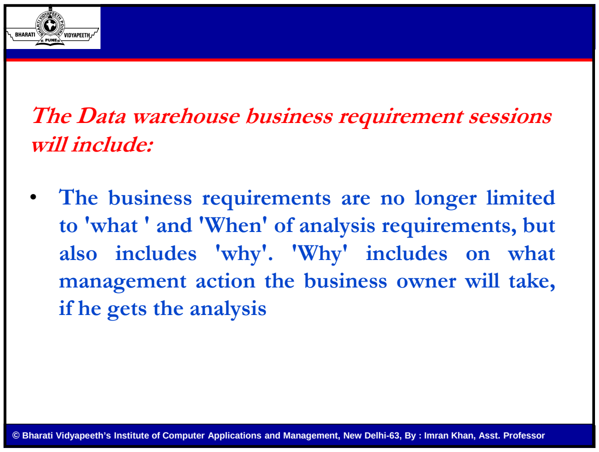
Let us take an example to clarify this point



Let us imagine you are building an operational system for order processing in our company

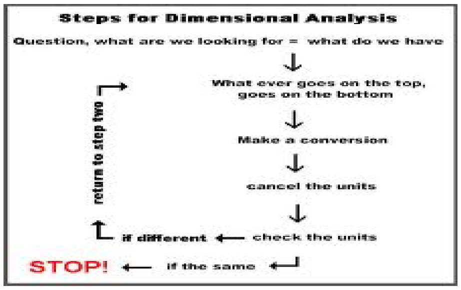
For gathering requirements

- We interview the users in the Order Processing department.
- The users will list all the functions that need to be performed.
- They will inform us how they receive the orders, check stock, verify customers' credit arrangements, price the order, determine the shipping arrangements, and route the order to the appropriate warehouse.
- They will show us how they would like the various data elements to be presented on the GUI (graphical user interface) screen for the application.
- The users will also give us a list of reports they would need from the order processing application.
- They will be able to let us know how and when they would use the application daily.



The Data warehouse business requirement sessions will include:

- The business requirements are no longer limited to 'what ' and 'When' of analysis requirements, but also includes 'why'. 'Why' includes on what management action the business owner will take, if he gets the analysis



Steps for Dimensional Analysis

Question, what are we looking for = what do we have

What ever goes on the top, goes on the bottom

Make a conversion

cancel the units

check the units

if different

return to step two

if the same

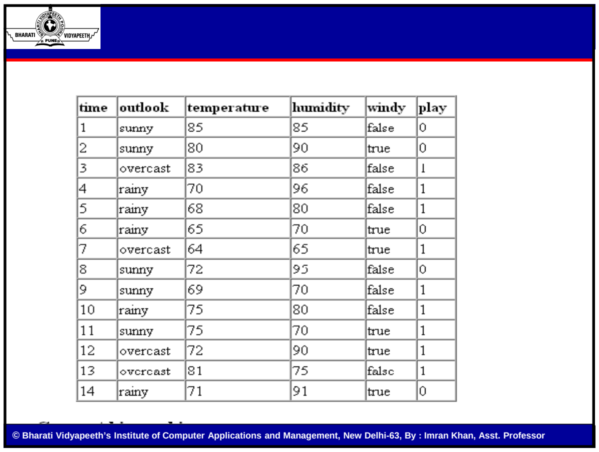
STOP!

OLAP OPERATIONS

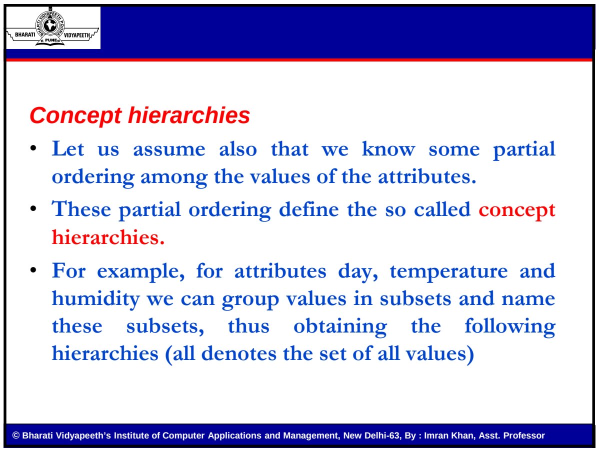
In order to understand the OLAP operations, we need to understand some basics with the help of an example

Understanding Multidimensional data model

- Let us consider the weather data as defined in the table given in the next slide.
- The dependent variable play has just two values - yes and no. As these values are mutually exclusive, we can replace them by 1 and 0 respectively. This will allows us to add up values and thus get the total number of days when tennis was played and at the same time the number of days tennis was not played (the complement of the former to the total number of days). Let us also rename the day attribute into time, which is more general and will allow us to use other time units (e.g. weeks). Thus we get the following relational table.

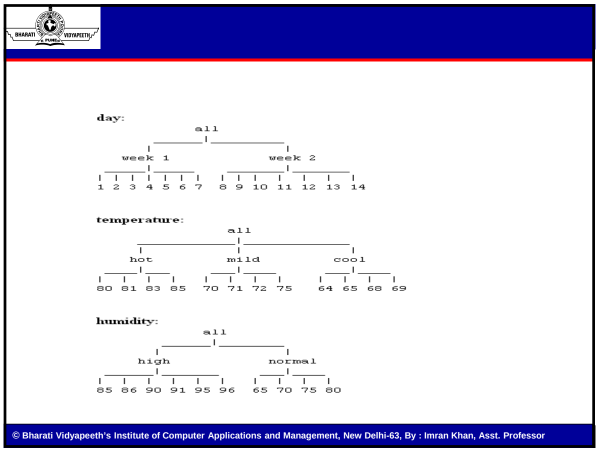


time	outlook	temperature	humidity	windy	play
1	sunny	85	85	false	0
2	sunny	80	90	true	0
3	overcast	83	86	false	1
4	rainy	70	96	false	1
5	rainy	68	80	false	1
6	rainy	65	70	true	0
7	overcast	64	65	true	1
8	sunny	72	95	false	0
9	sunny	69	70	false	1
10	rainy	75	80	false	1
11	sunny	75	70	true	1
12	overcast	72	90	true	1
13	ovcrcast	81	75	falsc	1
14	rainy	71	91	true	0



Concept hierarchies

- Let us assume also that we know some partial ordering among the values of the attributes.
- These partial ordering define the so called concept hierarchies.
- For example, for attributes day, temperature and humidity we can group values in subsets and name these subsets, thus obtaining the following hierarchies (all denotes the set of all values)



day:

all

week 1: 1 2 3 4 5 6 7

week 2: 8 9 10 11 12 13 14

temperature:

all

hot: 80 81 83 85

mild: 70 71 72 75

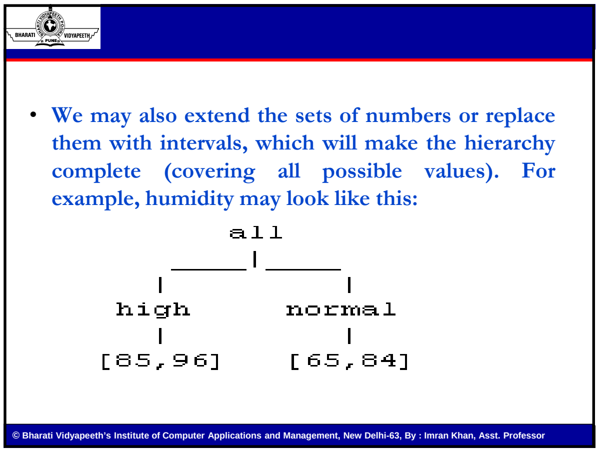
cool: 64 65 68 69

humidity:

all

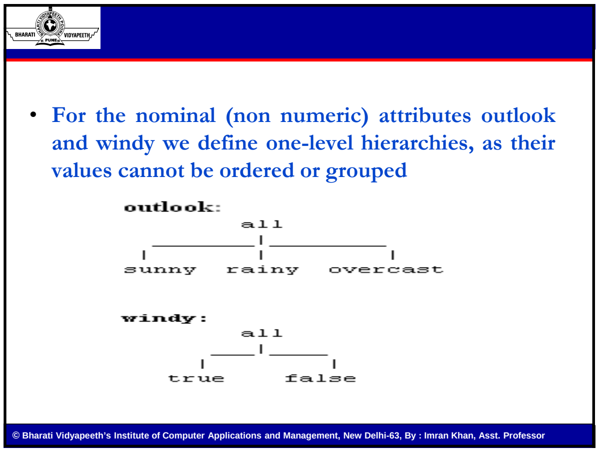
high: 85 86 90 91 95 96

normal: 65 70 75 80



- We may also extend the sets of numbers or replace them with intervals, which will make the hierarchy complete (covering all possible values). For example, humidity may look like this:

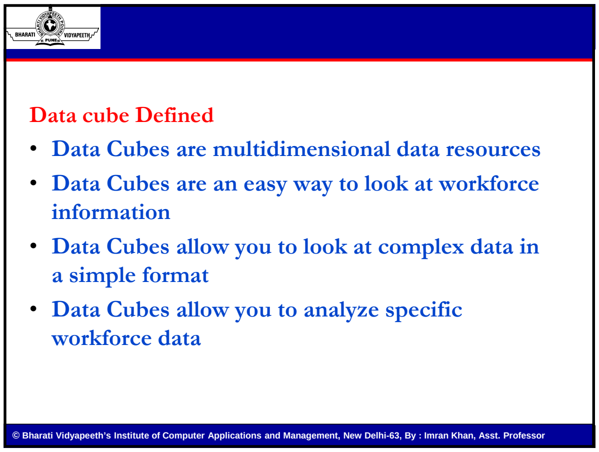
```
             all
      _______|_______
     |               |
    high          normal
     |               |
  [85,96]         [65,84]
```



- For the nominal (non numeric) attributes outlook and windy we define one-level hierarchies, as their values cannot be ordered or grouped

```
outlook:
                all
    ____________|____________
   |            |            |
 sunny        rainy       overcast

windy:
                all
          ______|______
         |             |
        true         false
```



Data cube Defined

- Data Cubes are multidimensional data resources
- Data Cubes are an easy way to look at workforce information
- Data Cubes allow you to look at complex data in a simple format
- Data Cubes allow you to analyze specific workforce data

Data cube

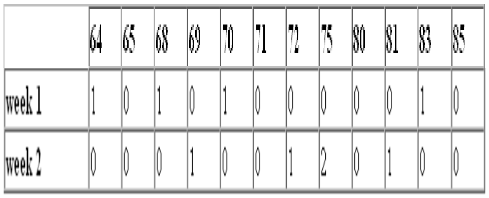
To create a data cube we have to:

Select dimensions, that is select a subset of attributes. For example, let us select time and temperature. Thus we will create a two-dimensional data cube.

Select levels in the concept hierarchies. For example, let us select weeks for time and degrees for temperature.

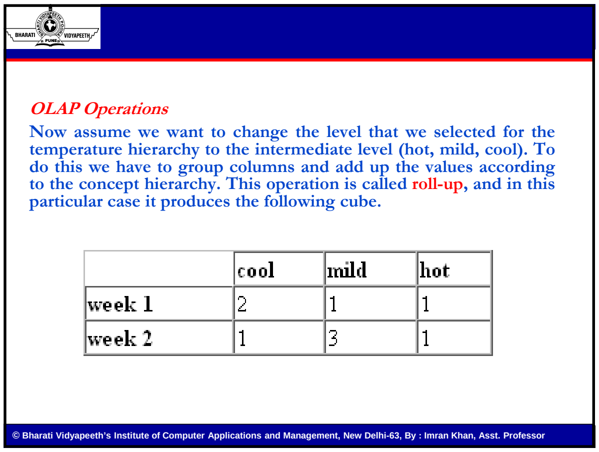
Select a measure to populate the cube. This is the attribute whose values will be aggregated across the dimensions (obviously it has to be numeric). Let us select play.

Then placing the time values in the rows and the temperature values in the columns we get the following cube:



	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

The numbers in the internal cells are obtained by adding up the values of the play attribute, where the time and the temperature attribute are equal to the values in the corresponding row and column. For example the value 2 (row 2, column 8) means that tennis was played two days during week 2 when the temperature was 75

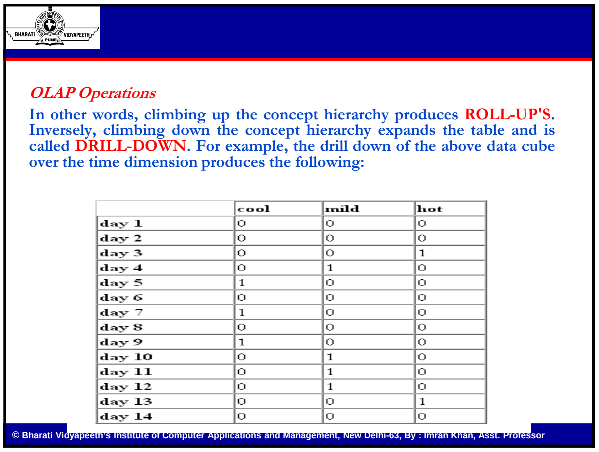


OLAP Operations

Now assume we want to change the level that we selected for the temperature hierarchy to the intermediate level (hot, mild, cool). To do this we have to group columns and add up the values according to the concept hierarchy. This operation is called roll-up, and in this particular case it produces the following cube.

	cool	mild	hot
week 1	2	1	1
week 2	1	3	1

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor

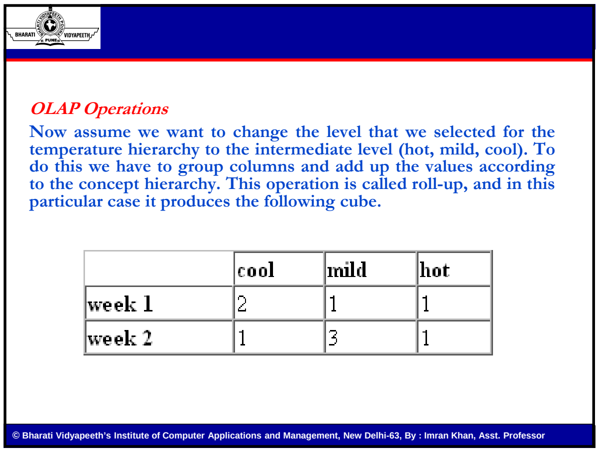


OLAP Operations

In other words, climbing up the concept hierarchy produces ROLL-UP'S. Inversely, climbing down the concept hierarchy expands the table and is called DRILL-DOWN. For example, the drill down of the above data cube over the time dimension produces the following:

	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

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor

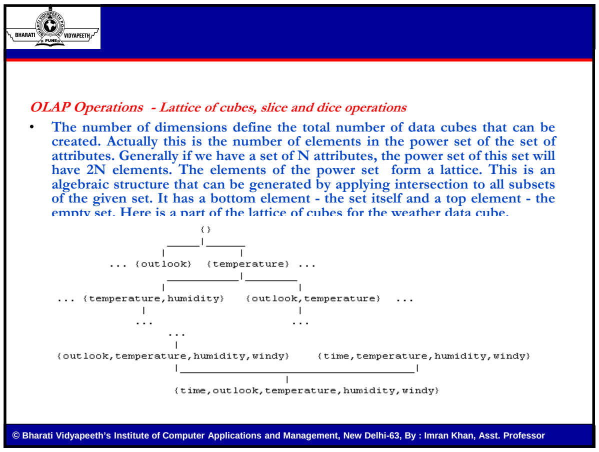


OLAP Operations

Now assume we want to change the level that we selected for the temperature hierarchy to the intermediate level (hot, mild, cool). To do this we have to group columns and add up the values according to the concept hierarchy. This operation is called roll-up, and in this particular case it produces the following cube.

	cool	mild	hot
week 1	2	1	1
week 2	1	3	1

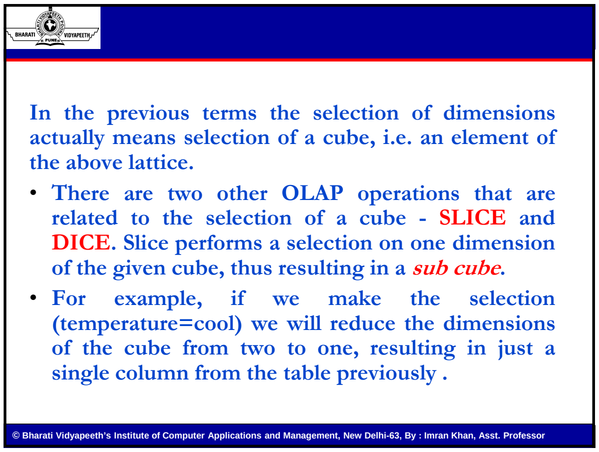
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



OLAP Operations - Lattice of cubes, slice and dice operations

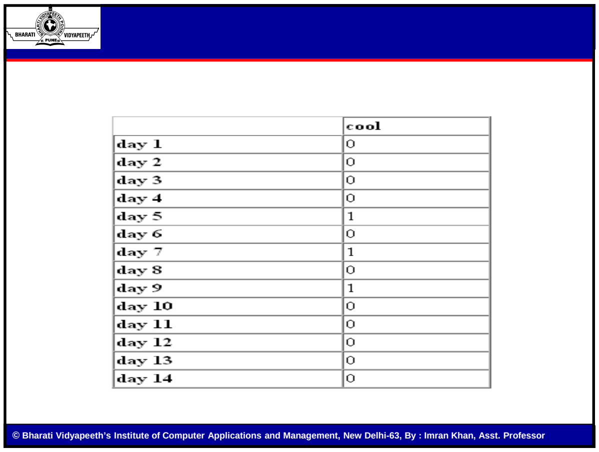
- The number of dimensions define the total number of data cubes that can be created. Actually this is the number of elements in the power set of the set of attributes. Generally if we have a set of N attributes, the power set of this set will have 2N elements. The elements of the power set form a lattice. This is an algebraic structure that can be generated by applying intersection to all subsets of the given set. It has a bottom element - the set itself and a top element - the empty set. Here is a part of the lattice of cubes for the weather data cube.

```
                          {}
                    _____|______
                   |            |
          ... {outlook}  {temperature} ...
                    ___________|________
                   |                    |
 ... {temperature,humidity}   {outlook,temperature}  ...
              |                         |
             ...                       ...
                     ...
                      |
 {outlook,temperature,humidity,windy}    {time,temperature,humidity,windy}
                      |_______________________________________|
                                          |
                     {time,outlook,temperature,humidity,windy}
```

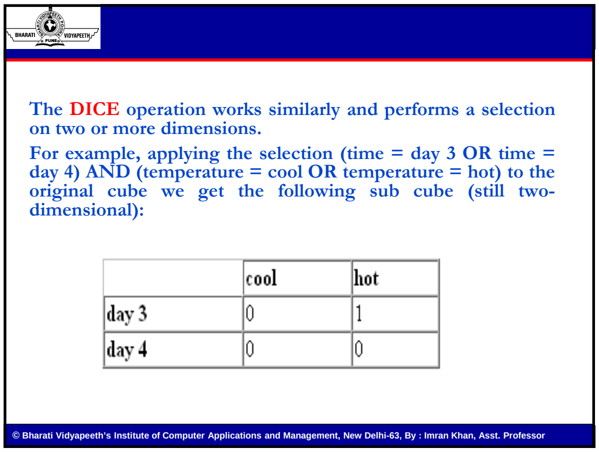


In the previous terms the selection of dimensions actually means selection of a cube, i.e. an element of the above lattice.

- There are two other OLAP operations that are related to the selection of a cube - SLICE and DICE. Slice performs a selection on one dimension of the given cube, thus resulting in a *sub cube*.
- For example, if we make the selection (temperature=cool) we will reduce the dimensions of the cube from two to one, resulting in just a single column from the table previously .



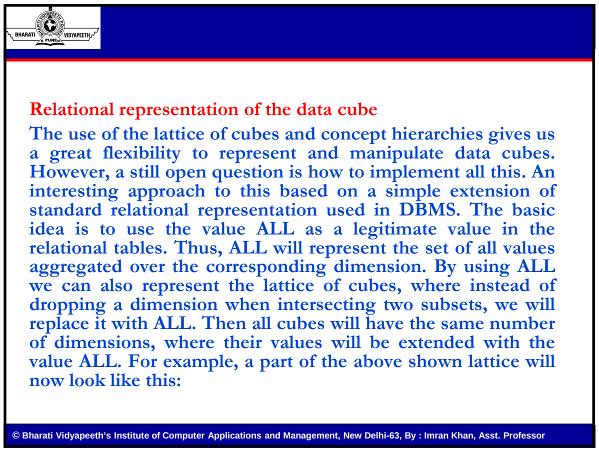
	cool
day 1	0
day 2	0
day 3	0
day 4	0
day 5	1
day 6	0
day 7	1
day 8	0
day 9	1
day 10	0
day 11	0
day 12	0
day 13	0
day 14	0



The DICE operation works similarly and performs a selection on two or more dimensions.

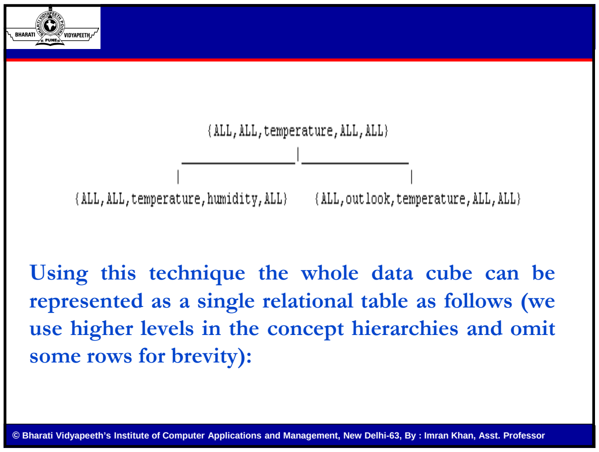
For example, applying the selection (time = day 3 OR time = day 4) AND (temperature = cool OR temperature = hot) to the original cube we get the following sub cube (still two-dimensional):

	cool	hot
day 3	0	1
day 4	0	0



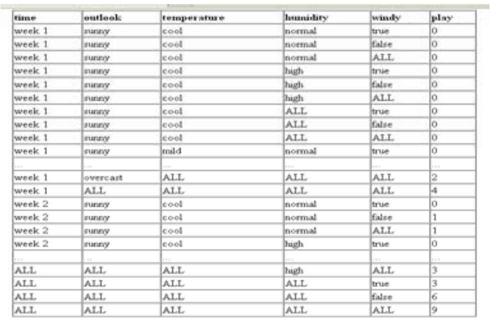
Relational representation of the data cube

The use of the lattice of cubes and concept hierarchies gives us a great flexibility to represent and manipulate data cubes. However, a still open question is how to implement all this. An interesting approach to this based on a simple extension of standard relational representation used in DBMS. The basic idea is to use the value ALL as a legitimate value in the relational tables. Thus, ALL will represent the set of all values aggregated over the corresponding dimension. By using ALL we can also represent the lattice of cubes, where instead of dropping a dimension when intersecting two subsets, we will replace it with ALL. Then all cubes will have the same number of dimensions, where their values will be extended with the value ALL. For example, a part of the above shown lattice will now look like this:



```
                {ALL,ALL,temperature,ALL,ALL}
              ________________|_____________
              |                             |
{ALL,ALL,temperature,humidity,ALL}   {ALL,outlook,temperature,ALL,ALL}
```

Using this technique the whole data cube can be represented as a single relational table as follows (we use higher levels in the concept hierarchies and omit some rows for brevity):



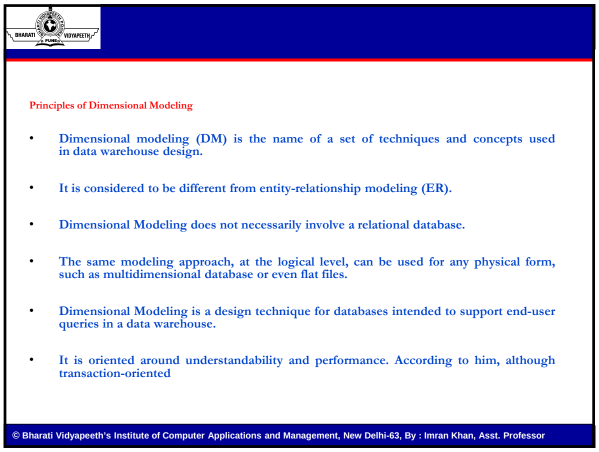
time	outlook	temperature	humidity	windy	play
week 1	sunny	cool	normal	true	0
week 1	sunny	cool	normal	false	0
week 1	sunny	cool	normal	ALL	0
week 1	sunny	cool	high	true	0
week 1	sunny	cool	high	false	0
week 1	sunny	cool	high	ALL	0
week 1	sunny	cool	ALL	true	0
week 1	sunny	cool	ALL	false	0
week 1	sunny	cool	ALL	ALL	0
week 1	sunny	mild	normal	true	0
...	...	...	...	...	...
week 1	overcast	ALL	ALL	ALL	2
week 1	ALL	ALL	ALL	ALL	4
week 2	sunny	cool	normal	true	0
week 2	sunny	cool	normal	false	1
week 2	sunny	cool	normal	ALL	1
week 2	sunny	cool	high	true	0
...	...	...	...	...	...
ALL	ALL	ALL	high	ALL	3
ALL	ALL	ALL	ALL	true	3
ALL	ALL	ALL	ALL	false	6
ALL	ALL	ALL	ALL	ALL	9

The table in the previous slide allows us to use an unified approach to implement all OLAP operations - they all can me implemented just by selecting proper rows. For example, the following cube:

	cool	mild	hot
week 1	2	1	1
week 2	1	3	1

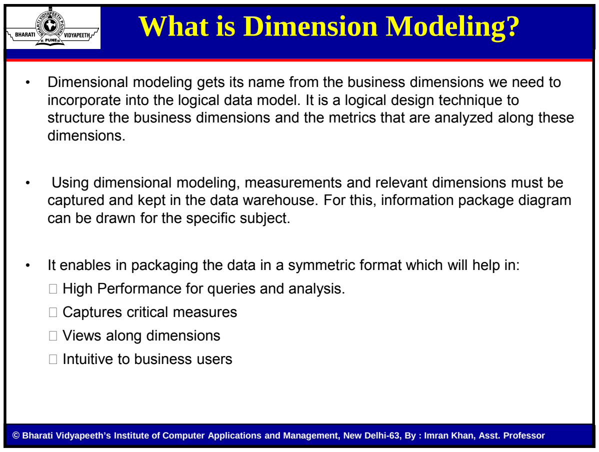
can be extracted from the table by selecting the rows that match the pattern (*, ALL, *, ALL, ALL), where * matches all legitimate values for the corresponding dimension except for ALL.

PRINCIPLES OF DIMENSIONAL MODELLING



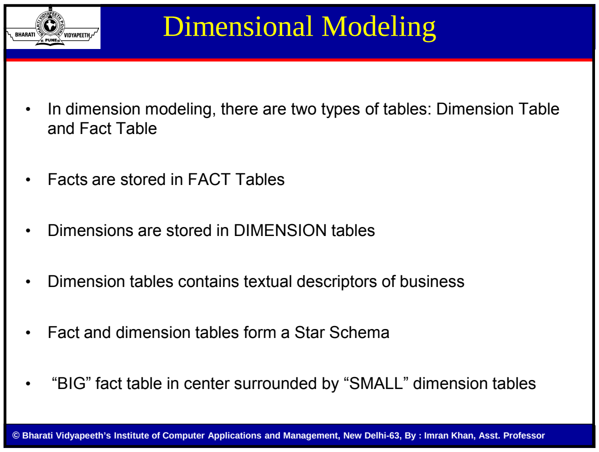
Principles of Dimensional Modeling

- Dimensional modeling (DM) is the name of a set of techniques and concepts used in data warehouse design.
- It is considered to be different from entity-relationship modeling (ER).
- Dimensional Modeling does not necessarily involve a relational database.
- The same modeling approach, at the logical level, can be used for any physical form, such as multidimensional database or even flat files.
- Dimensional Modeling is a design technique for databases intended to support end-user queries in a data warehouse.
- It is oriented around understandability and performance. According to him, although transaction-oriented



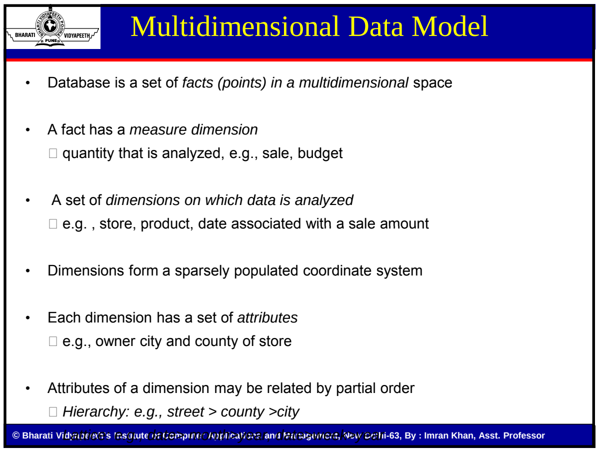
What is Dimension Modeling?

- Dimensional modeling gets its name from the business dimensions we need to incorporate into the logical data model. It is a logical design technique to structure the business dimensions and the metrics that are analyzed along these dimensions.
- Using dimensional modeling, measurements and relevant dimensions must be captured and kept in the data warehouse. For this, information package diagram can be drawn for the specific subject.
- It enables in packaging the data in a symmetric format which will help in:
  - □ High Performance for queries and analysis.
  - □ Captures critical measures
  - □ Views along dimensions
  - □ Intuitive to business users



Dimensional Modeling

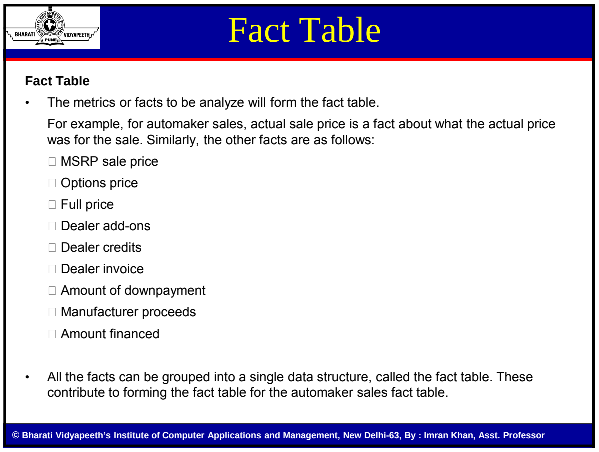
- In dimension modeling, there are two types of tables: Dimension Table and Fact Table
- Facts are stored in FACT Tables
- Dimensions are stored in DIMENSION tables
- Dimension tables contains textual descriptors of business
- Fact and dimension tables form a Star Schema
- "BIG" fact table in center surrounded by "SMALL" dimension tables



Multidimensional Data Model

- Database is a set of *facts (points) in a multidimensional* space
- A fact has a *measure dimension*
  - □ quantity that is analyzed, e.g., sale, budget
- A set of *dimensions on which data is analyzed*
  - □ e.g. , store, product, date associated with a sale amount
- Dimensions form a sparsely populated coordinate system
- Each dimension has a set of *attributes*
  - □ e.g., owner city and county of store
- Attributes of a dimension may be related by partial order
  - □ *Hierarchy: e.g., street > county >city*

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



Fact Table

Fact Table

- The metrics or facts to be analyze will form the fact table.

  For example, for automaker sales, actual sale price is a fact about what the actual price was for the sale. Similarly, the other facts are as follows:

  □ MSRP sale price

  □ Options price

  □ Full price

  □ Dealer add-ons

  □ Dealer credits

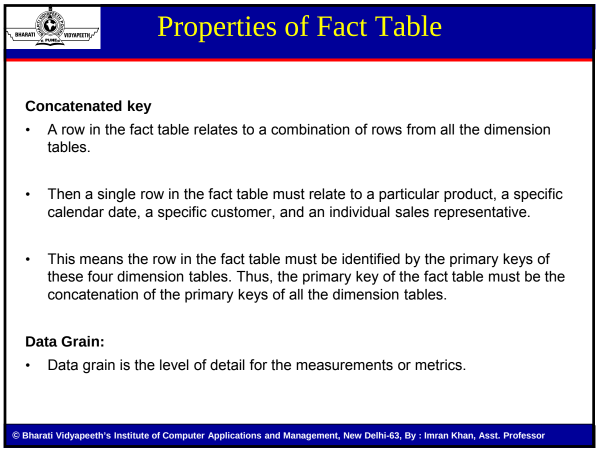
  □ Dealer invoice

  □ Amount of downpayment

  □ Manufacturer proceeds

  □ Amount financed
- All the facts can be grouped into a single data structure, called the fact table. These contribute to forming the fact table for the automaker sales fact table.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



Properties of Fact Table

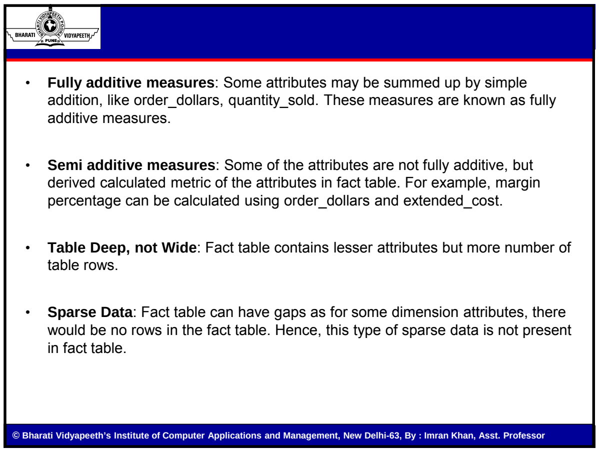
Concatenated key

- A row in the fact table relates to a combination of rows from all the dimension tables.
- Then a single row in the fact table must relate to a particular product, a specific calendar date, a specific customer, and an individual sales representative.
- This means the row in the fact table must be identified by the primary keys of these four dimension tables. Thus, the primary key of the fact table must be the concatenation of the primary keys of all the dimension tables.

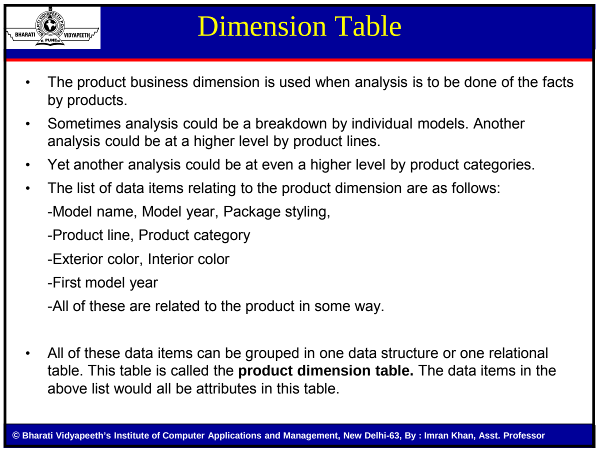
Data Grain:

- Data grain is the level of detail for the measurements or metrics.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, By : Imran Khan, Asst. Professor



- **Fully additive measures**: Some attributes may be summed up by simple addition, like order_dollars, quantity_sold. These measures are known as fully additive measures.
- **Semi additive measures**: Some of the attributes are not fully additive, but derived calculated metric of the attributes in fact table. For example, margin percentage can be calculated using order_dollars and extended_cost.
- **Table Deep, not Wide**: Fact table contains lesser attributes but more number of table rows.
- **Sparse Data**: Fact table can have gaps as for some dimension attributes, there would be no rows in the fact table. Hence, this type of sparse data is not present in fact table.



Dimension Table

- The product business dimension is used when analysis is to be done of the facts by products.
- Sometimes analysis could be a breakdown by individual models. Another analysis could be at a higher level by product lines.
- Yet another analysis could be at even a higher level by product categories.
- The list of data items relating to the product dimension are as follows:

  -Model name, Model year, Package styling,

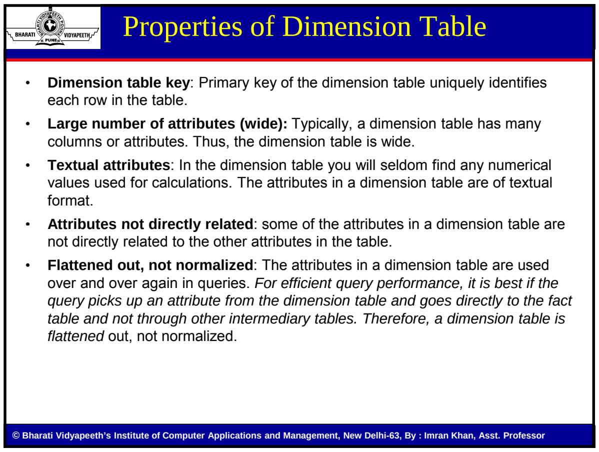
  -Product line, Product category

  -Exterior color, Interior color

  -First model year

  -All of these are related to the product in some way.

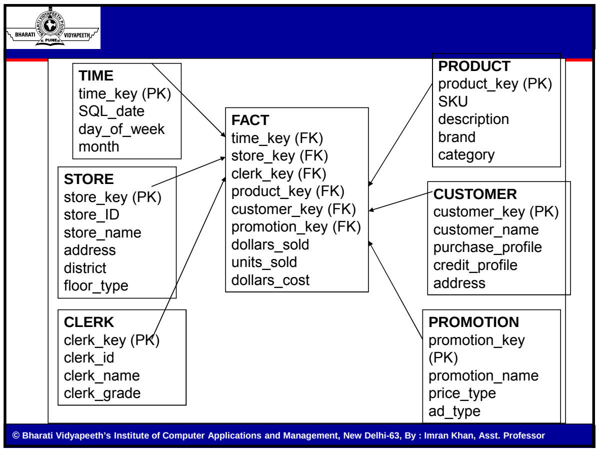
- All of these data items can be grouped in one data structure or one relational table. This table is called the **product dimension table.** The data items in the above list would all be attributes in this table.



Properties of Dimension Table

- **Dimension table key**: Primary key of the dimension table uniquely identifies each row in the table.
- **Large number of attributes (wide):** Typically, a dimension table has many columns or attributes. Thus, the dimension table is wide.
- **Textual attributes**: In the dimension table you will seldom find any numerical values used for calculations. The attributes in a dimension table are of textual format.
- **Attributes not directly related**: some of the attributes in a dimension table are not directly related to the other attributes in the table.
- **Flattened out, not normalized**: The attributes in a dimension table are used over and over again in queries. *For efficient query performance, it is best if the query picks up an attribute from the dimension table and goes directly to the fact table and not through other intermediary tables. Therefore, a dimension table is flattened* out, not normalized.

- **Ability to drill down / roll up:** The attributes in a dimension table provide the ability to get to the details from higher levels of aggregation to lower levels of details.
- **Multiple hierarchies:** dimension tables often provide for multiple hierarchies, so that drilling down may be performed along any of the multiple hierarchies.
- **Less number of records:** A dimension table typically has fewer number of records or rows than the fact table.



Operations in Multidimensional Data Model

- Aggregation (*roll-up)*
  - □ dimension reduction: e.g., total sales by city
  - □ summarization over aggregate hierarchy: e.g., total sales by city and year -> total sales by region and by year
- Navigation to detailed data (*drill-down)*
  - □ e.g., (sales - expense) by city, top 3% of cities by average income
- Selection (*slice) defines a subcube*
  - □ e.g., sales where city = Palo Alto and date = 1/15/96
- Visualization Operations (e.g., Pivot)