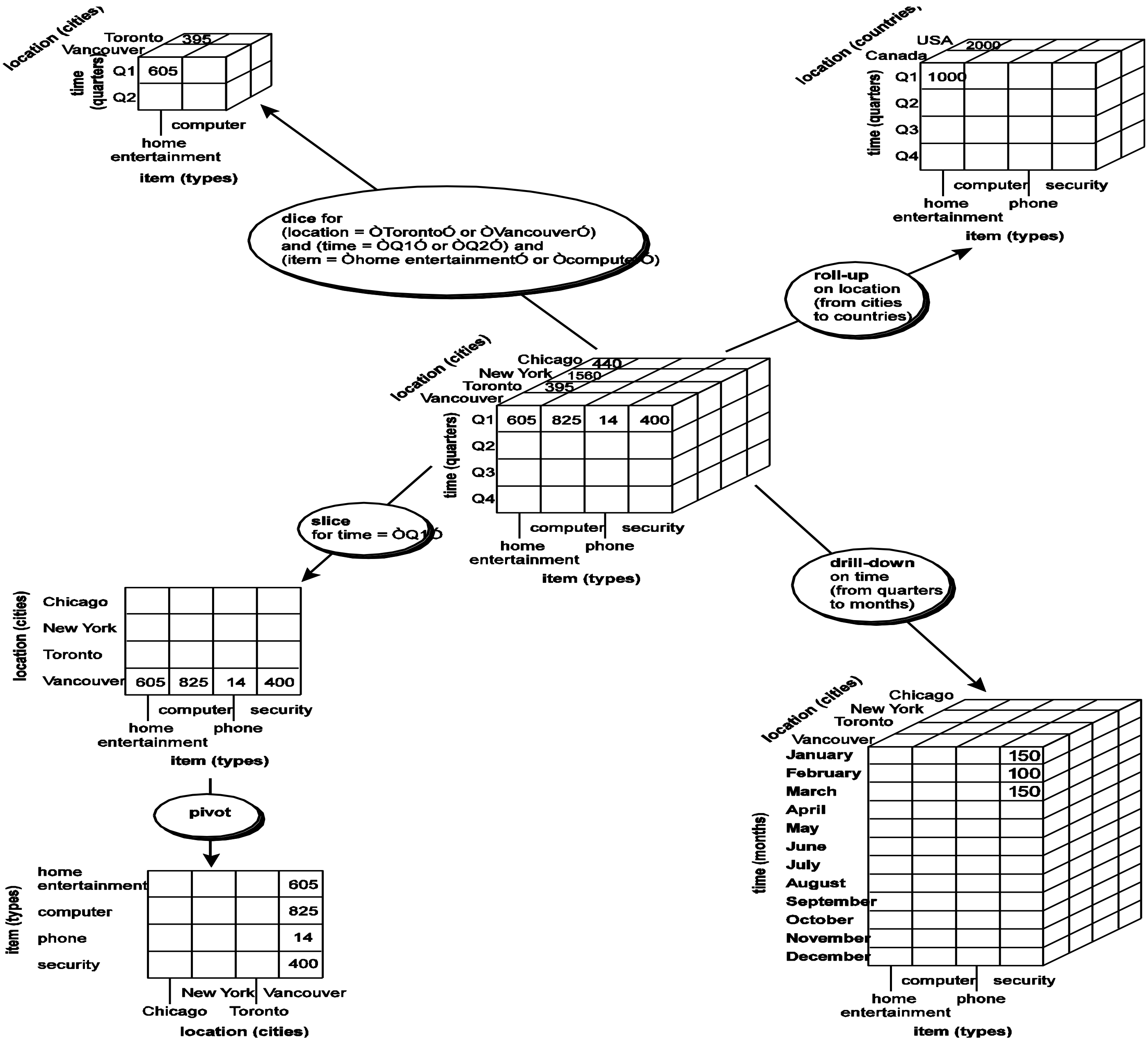
Sub:	Data Mining with Business Intelligence	Code:	20MCA242
Sem:	II	Branch:	MCA

Q1. Explain what kind of operations can be performed on data cube.

Ans: A number of operations may be applied to data cubes. The common ones are:
- roll-up (increasing the level of abstraction)
- drill-down (increasing detail)
- slice and dice (selection and projection)
- pivot (re-orienting the view)

- *Roll-up* (less detail) - when we wish further abstraction (i.e. less detail). This operation performs further aggregation on the data, for example, from single degree programs to Schools, single countries to Continents or from three dimensions to two dimensions.
- *Drill-down* (increasing detail) - reverse of roll up, when we wish to partition more finely or want to focus on some particular values of certain dimensions. Drill-down adds more detail to the data, it may involve adding another dimension.
- *Slice and dice* (selection and projection) - the slice operation performs a selection on one dimension of the cube (e.g. degree = “MIT”). The dice operation performs a selection on two or more dimensions (e.g. degree = “BIT” and country = “Australia” or “India”)
- *Pivot* (re-orienting the view) - an alternate presentation of the data e.g. rotating the axes in a 3-D cube.



Q2. Explain the different types of schema for data ware house design.

Ans: Data Warehouse Design:

The E-R Model approach which consists of entities and relationships is not suitable for designing a schema for a warehouse.

What is the nature of data in data warehouse? Essentially data warehouses are based on multidimensional data model called data cube consisting of dimensions.

Data cube allows data to be modeled and viewed in multiple dimension.

Defined by dimension and facts.

- Dimension: An ordinate within a multidimensional structure consisting of a list of ordered values(sometimes called members).
- A member is a distinct value for the dimension(degree dimension has a member BSc)
- Fact Table: collection of related data items, consisting of values of dimensions of interest and the value(s) of measure(s).
- Dimension together form the primary key of the fact table.
- The non-key attributes of the fact table are the measures.

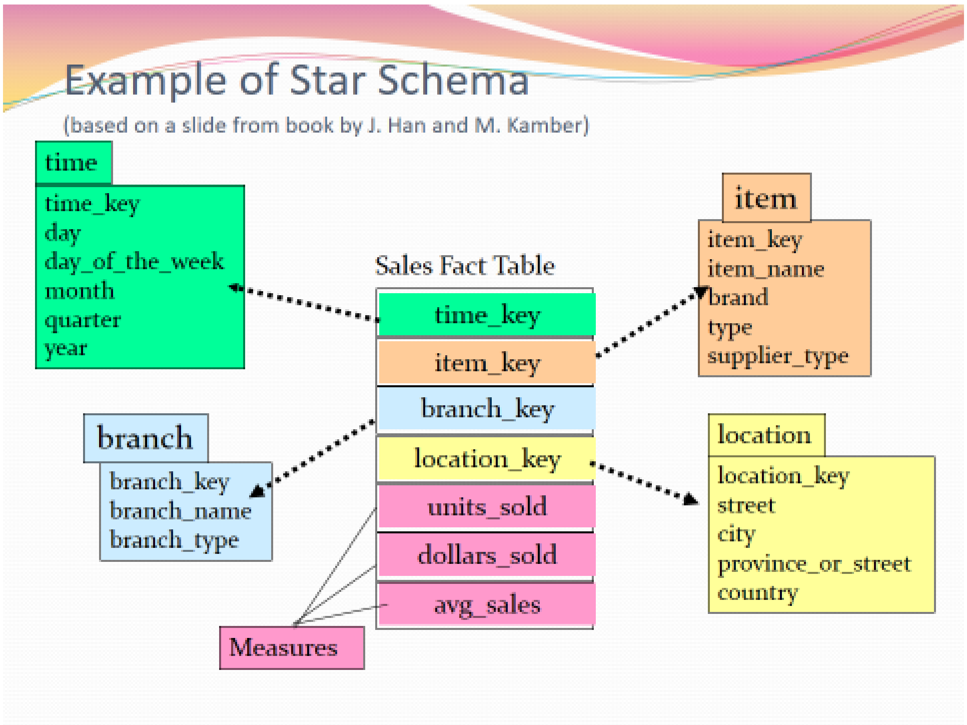
- Dimensions: perspectives or entities with respect to which an organization wants to keep records.

For example we can create a sales data warehouse in order to keep records of the store's sales with respect to the dimension time, item, branch, and location.

- Each dimension may have a table associated with it, called a dimension table.

For example a dimension table for item may contain the attribute item_name, brand, and type.

- A multidimensional data model is typically organized around a central theme, like sales.
- This theme is represented by a fact table.
- Facts are numerical measures.
- The quantities by which we want to analyze relationships between dimensions.
- *Example: Facts for a sales data warehouse include dollars_sold, units_sold.*
- Fact table contains the name of the fact as well as keys to each of the related dimension tables.
- One approach is the *star schema* to represent the multidimensional data model. The schema in this model consists of a large single fact table containing the bulk of the data, with no redundancy and a set of smaller tables called dimension table, one for each dimension.
- Other models have been used. These include *snowflakes model* and *fact constellations model*.



Example of Star Schema

(based on a slide from book by J. Han and M. Kamber)

time
time_key
day
day_of_the_week
month
quarter
year

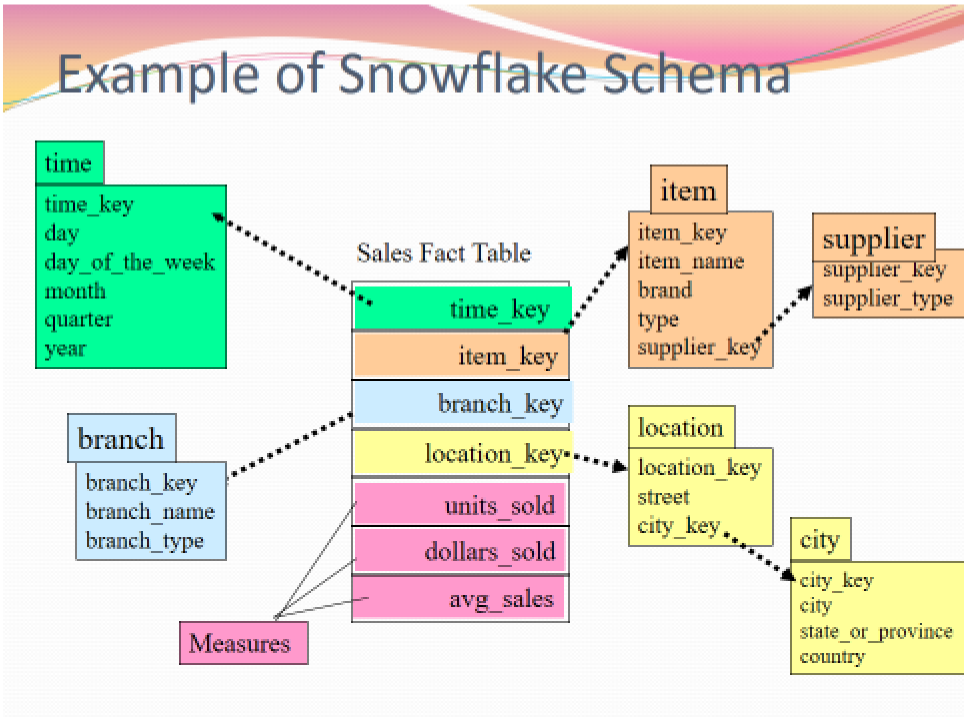
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_type

branch
branch_key
branch_name
branch_type

location
location_key
street
city
province_or_street
country

Measures



Example of Snowflake Schema

time
time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_key

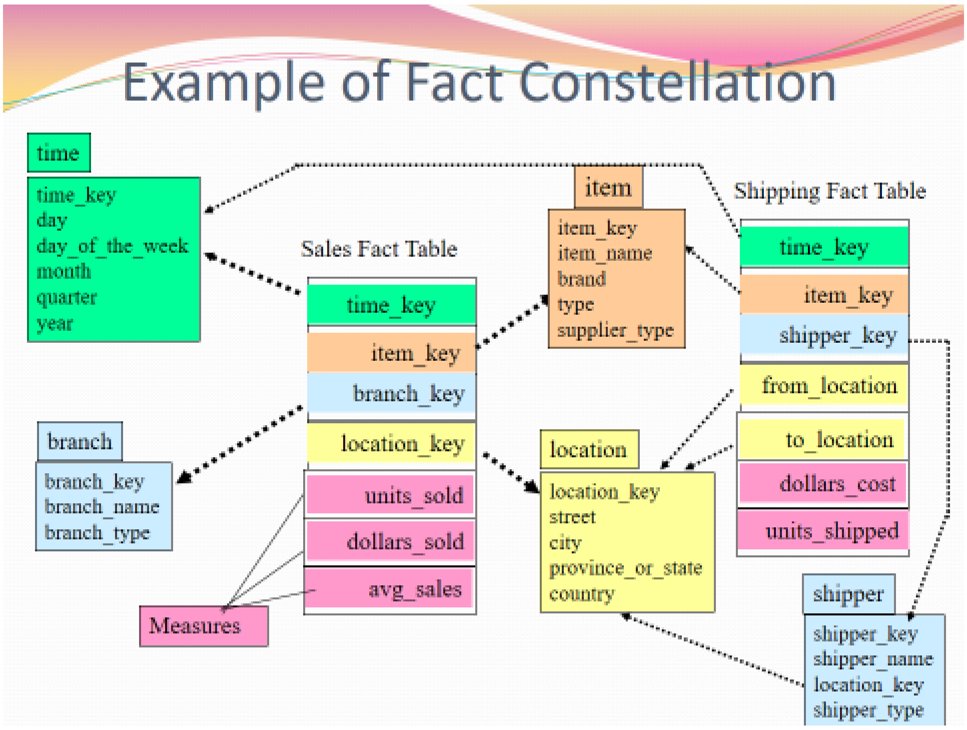
supplier
supplier_key
supplier_type

branch
branch_key
branch_name
branch_type

location
location_key
street
city_key

city
city_key
city
state_or_province
country

Measures



Example of Fact Constellation

time
time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_type

Shipping Fact Table
time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

branch
branch_key
branch_name
branch_type

location
location_key
street
city
province_or_state
country

shipper
shipper_key
shipper_name
location_key
shipper_type

Measures

Q3. Explain any two data preprocessing steps.

- Ans: **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- 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
- Redundant data occur often when integration of multiple databases
    - *Object identification*: The same attribute or object may have different names in different databases
    - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality
- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
    - Dimensionality reduction, e.g., remove unimportant attributes
        - Wavelet transforms
        - Principal Components Analysis (PCA)
        - Feature subset selection, feature creation
    - Numerosity reduction (some simply call it: Data Reduction)
        - Regression and Log-Linear Models
        - Histograms, clustering, sampling
        - Data cube aggregation
    - Data compression

Q4. Write the difference between: (i) OLAP and OLTP (ii)ROLAP and MOLAP
Ans: OLAP/OLTP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	Dozens
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

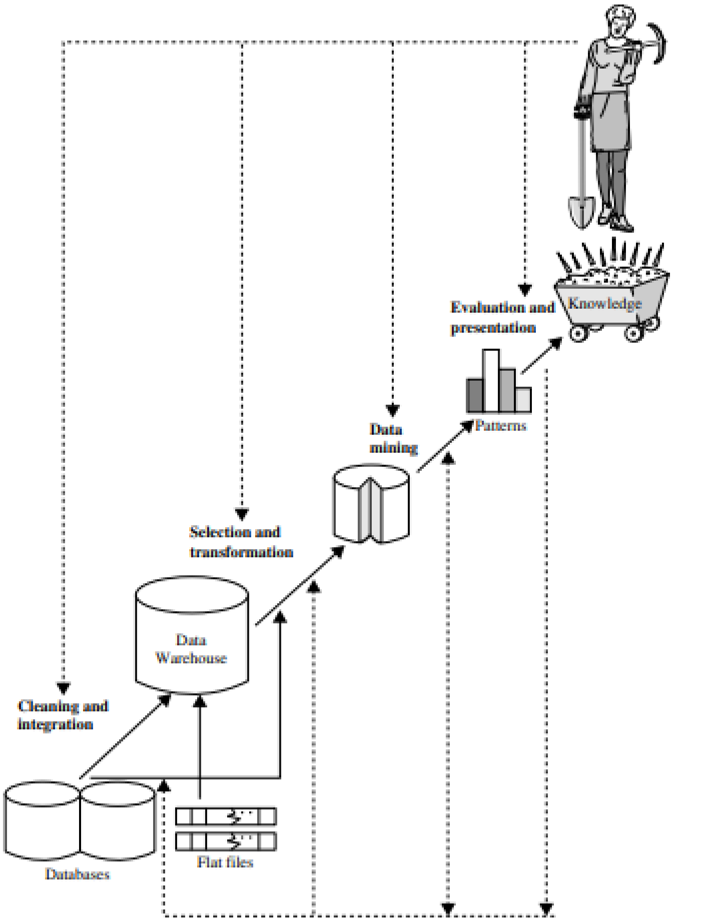
MOLAP/ROLAP
Relational OLAP (ROLAP)

- Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces
- Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services

    - greater scalability

- Multidimensional OLAP (MOLAP)
    - Array-based multidimensional storage engine (sparse matrix techniques)
    - fast indexing to pre-computed summarized data

Q5. Explain Knowledge discovery process in data mining. Discuss the issues related with data mining.
Ans:
The knowledge discovery process is shown in Figure 1.4 as an iterative sequence of the following steps:
1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)



Major Issues in Data Mining

- Mining Methodology
    - Mining various and new kinds of knowledge
    - Mining knowledge in multi-dimensional space

    - Data mining: An interdisciplinary effort
    - Boosting the power of discovery in a networked environment
    - Handling noise, uncertainty, and incompleteness of data
    - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
    - Interactive mining
    - Incorporation of background knowledge
    - Presentation and visualization of data mining results
- Efficiency and Scalability
    - Efficiency and scalability of data mining algorithms
    - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
    - Handling complex types of data
    - Mining dynamic, networked, and global data repositories
- Data mining and society
    - Social impacts of data mining
    - Privacy-preserving data mining
    - Invisible data mining

Q6. Discss about the buiding blocks of data ware house. What is the difference between data ware house and data mart?

Ans:

Architecture is the proper arrangement of the components.

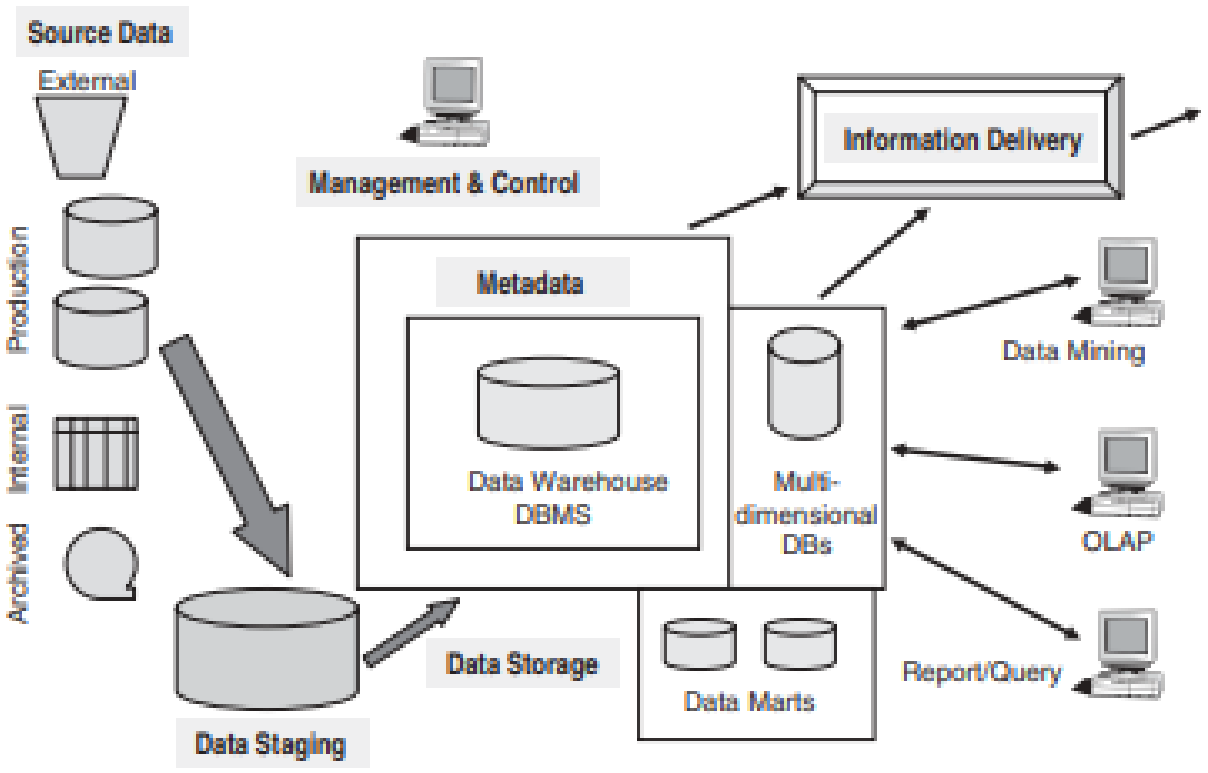


Figure 2-7 Data warehouse: building blocks or components.

Production Data : This category of data comes from the various operational systems of the enterprise.

Internal Data : In every organization, users keep their "private" spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse.

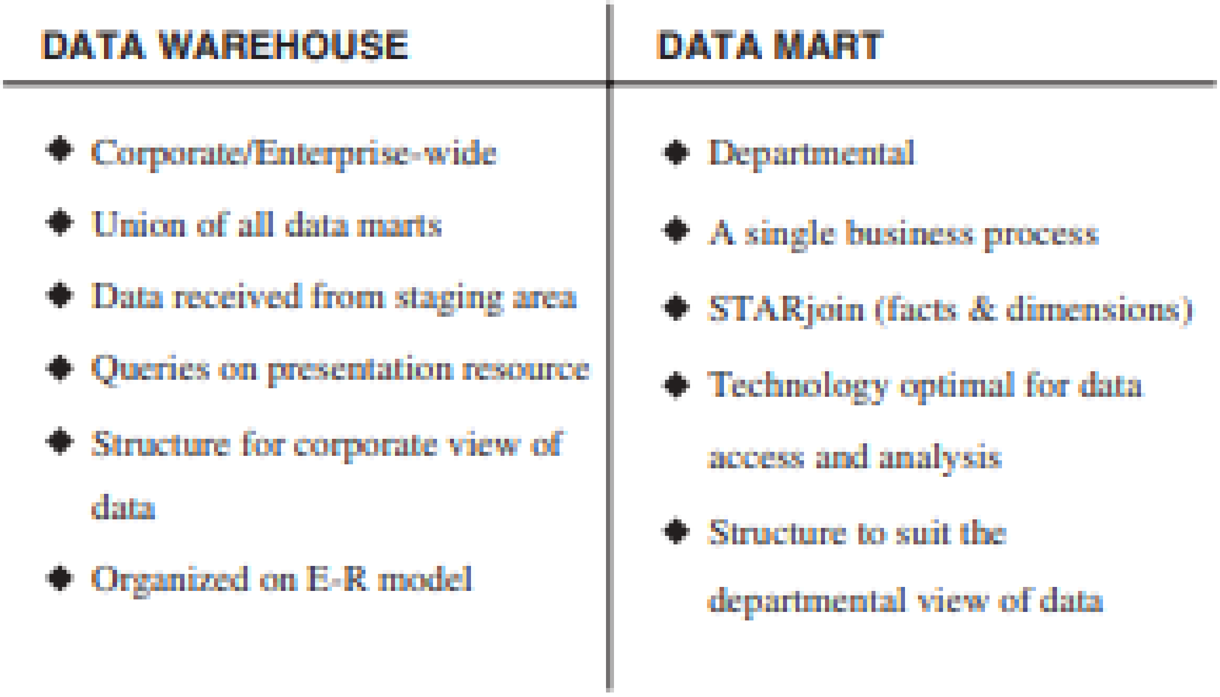
Archived Data : Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. Much of the archived data comes from old legacy systems that are nearing the end of their useful lives in organizations

External Data :Most executives depend on data from external sources for a high percentage of the information they use. They use statistics relating to their industry produced by external agencies and national statistical offices.

Data Staging Component : After you have extracted data from various operational systems and from external sources, you have to prepare the data for storing in the data warehouse. The extracted data coming from several disparate sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.

Data Transformation : Data transformation involves many forms of combining pieces of data from the different sources. You combine data from a single source record or related data elements from many source records. On the other hand, data transformation also involves purging source data that is not useful and separating out source records into new combinations. Sorting and merging of data takes place on a large scale in the data staging area.

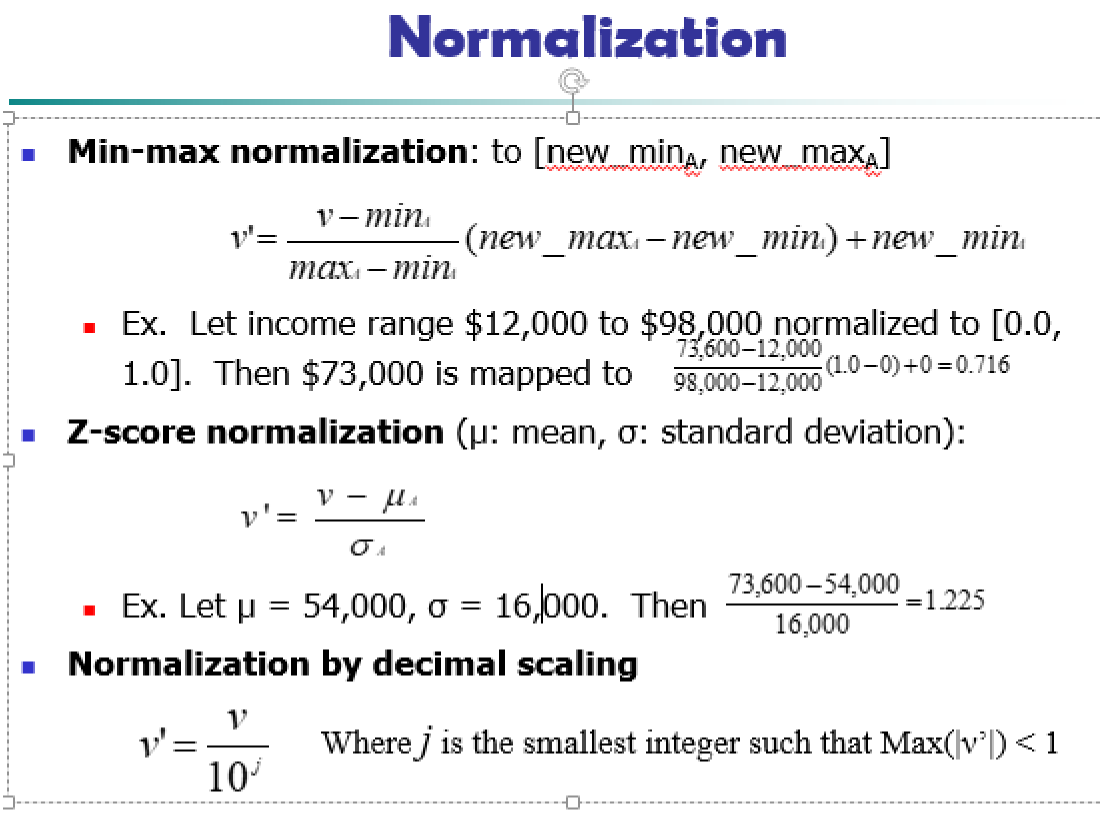
Data Loading: Two distinct groups of tasks form the data loading function. When you complete the design and construction of the data warehouse and go live for the first time, you do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time.
Data Storage Component: In the data repository for a data warehouse, you need to keep large volumes of historical data for analysis. Further, you have to keep the data in the data warehouse in structures suitable for analysis, and not for quick retrieval of individual pieces of information. Therefore, the data storage for the data warehouse is kept separate from the data storage for operational systems.
Information Delivery Component: In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery. Figure 2-9 shows the different information delivery methods. Ad hoc reports are predefined reports primarily meant for novice and casual users. Provision for complex queries, multidimensional (MD) analysis, and statistical analysis cater to the needs of the business analysts and power users. Information fed into executive information systems (EIS) is meant for senior executives and high-level managers.
Metadata Component : Metadata in a data warehouse is similar to the data dictionary or the data catalog in a database management system. In the data dictionary, you keep the information about the logical data structures, the information about the files and addresses, the information about the indexes, and so on. The data dictionary contains data about the data in the database.
Difference Between Data WareHouse and Data Mart



DATA WAREHOUSE	DATA MART
◆ Corporate/Enterprise-wide	◆ Departmental
◆ Union of all data marts	◆ A single business process
◆ Data received from staging area	◆ STARjoin (facts & dimensions)
◆ Queries on presentation resource	◆ Technology optimal for data access and analysis
◆ Structure for corporate view of data	◆ Structure to suit the departmental view of data
◆ Organized on E-R model	

Q7. Discuss about motivation for data mining. Discuss about different types of normalization techniques with example. Why it is needed?
Ans: Motivation of Data Mining:

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets



Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600-12{,}000}{98{,}000-12{,}000}(1.0-0)+0=0.716$
- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

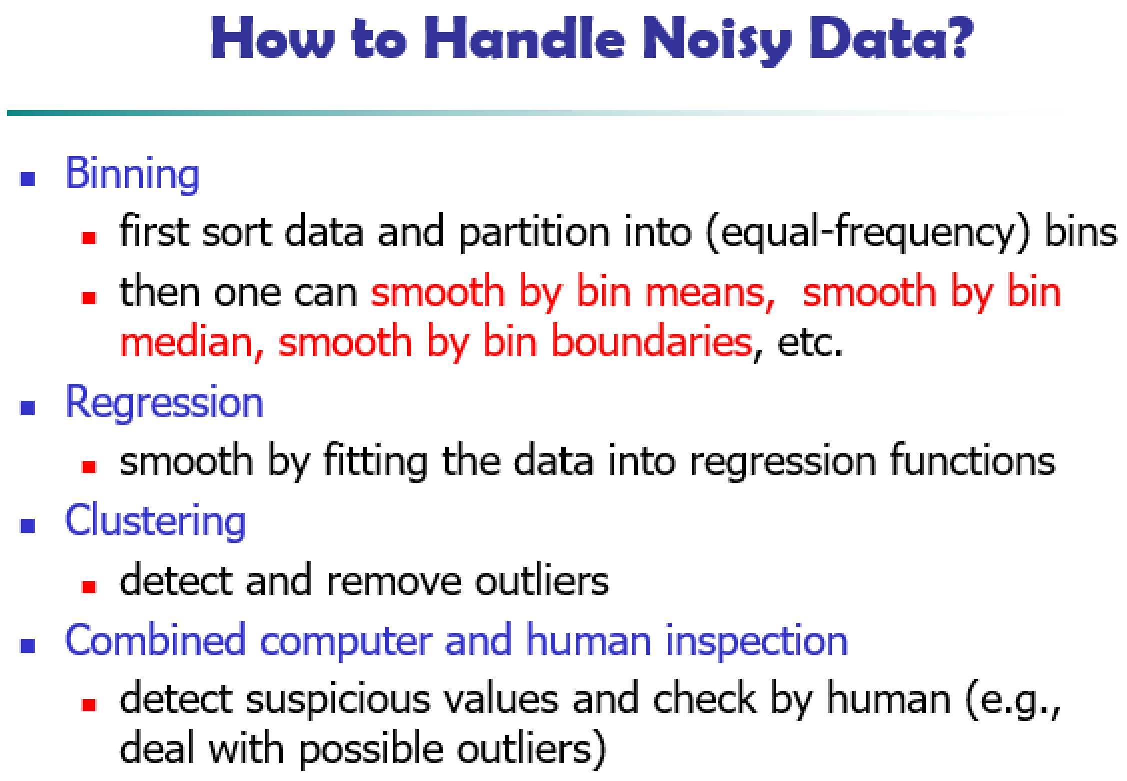
  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600-54{,}000}{16{,}000}=1.225$
- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Ans:

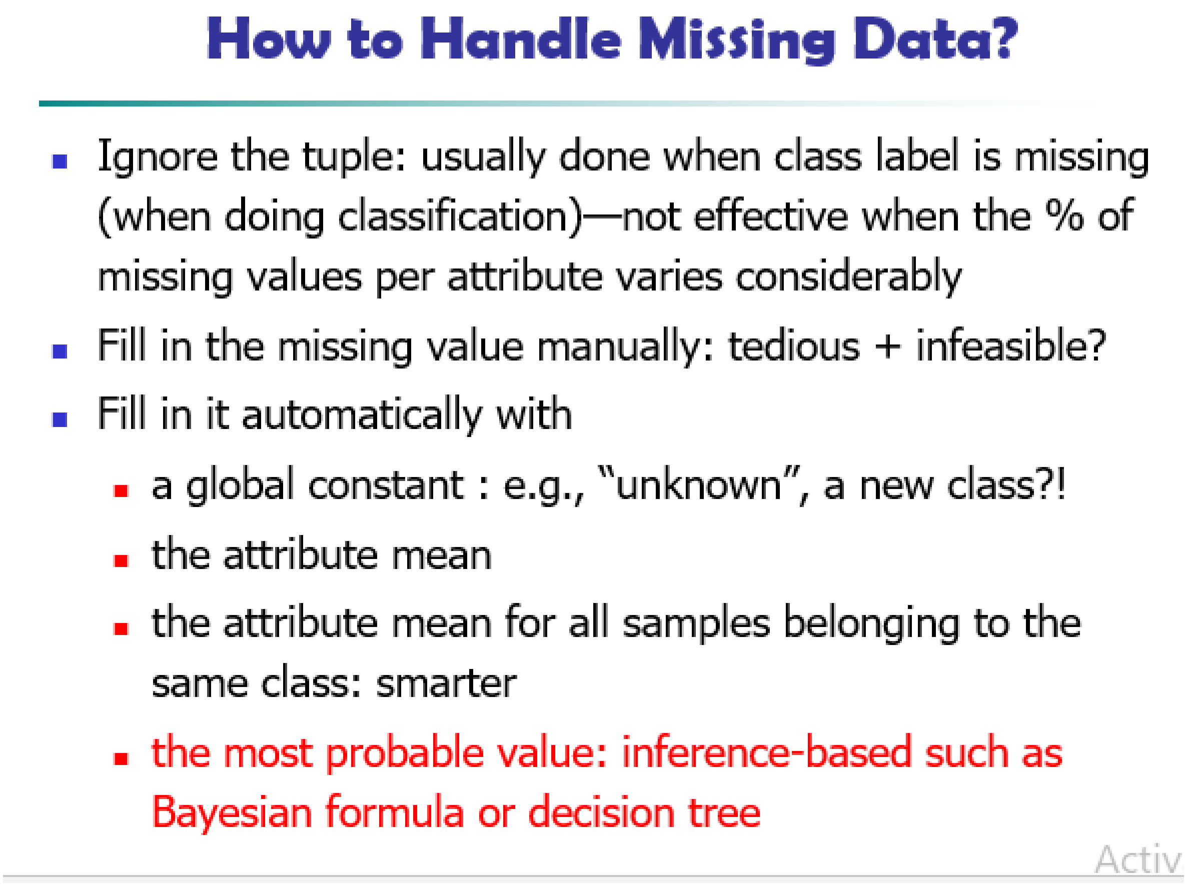
Q8. What is data smoothing? Discuss the techniques used for it.

Ans: Removing noise and inconsistencies from data is called data smoothing.



How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)



How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Q9. What is a Data Warehouse? List down the differences between operational database systems and data warehouses.

Ans: *A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process.*

Subject-oriented:

- A DW is organized around major subjects, such as student, degree, country.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations.
- A DW provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated:

- A DW may be constructed by integrating information from multiple data sources e.g. multiple OLTP databases.
- Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.

Time Variant:

- A DW usually has long time horizon, significantly longer than that of operational systems.
  - Operational database: current value data.
  - DW data: provide information from a historical perspective (e.g. past 5-10 years)
- Every key structure in the DW contains an element of time, explicitly or implicitly
- Operational data may or may not contain time element.

Non-volatile:

- A physically separate store of data transformed from the operational environment.

- No update of data
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Requires only two operations in data accessing: *initial loading of data* and *access of data*.

ODS vs DW

- DW may be used to provide an enterprise memory which operational data does not provide.
- DW does not requires and store real time data while the ODS does.

ODS	**DW**
Data of high quality at detailed level and assured availability	Data may not be perfect, but sufficient for strategic analysis; data does not have to be highly available
Contains current and near-current data	Contains historical data
Real-time and near real-time data loads	Normally batch data loads
Mostly updated at data field level(even if it may be appended)	Data is appended, not updated
Typically detailed data only	Contains summarized and detailed data
Modeled to support rapid data updates(3NF)	Variety of modeling techniques used, typically multidimensional for data marts to optimize query performance
Transactions similar to those in OLTP systems	Complex queries processing larger volumes of data
Used for detailed decision making and operational reporting	Used for long-term decision making and management reporting
Used at the operational level	Used at the managerial level

Q10. Discuss Data Mining tasks in detail.

Ans: **Data Mining Function: Association and Correlation Analysis**

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: Classification

- Classification and label prediction
    - Construct models (functions) based on some training examples
    - Describe and distinguish classes or concepts for future prediction
        - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
    - Predict some unknown class labels
- Typical methods
    - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
    - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …
    -

Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: Outlier Analysis

- Outlier analysis
    - Outlier: A data object that does not comply with the general behavior of the data
    - Noise or exception? — One person's garbage could be another person's treasure
    - Methods: by product of clustering or regression analysis, …
    - Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis:

- Sequence, trend and evolution analysis
    - Trend, time-series, and deviation analysis: e.g., regression and value prediction
    - Sequential pattern mining
        - e.g., first buy digital camera, then buy large SD memory cards
    - Periodicity analysis
    - Motifs and biological sequence analysis
        - Approximate and consecutive motifs
    - Similarity-based analysis
- Mining data streams
    - Ordered, time-varying, potentially infinite, data streams