Data Warehouse & Mining

Information Systems

Transaction Processing Systems

- Substitutes computer-based processing for manual procedures.

  Deals with well-structured processes. Includes record keeping applications.

E.g. Airline Reservation Systems, Banking Systems, or the Accounting System

Management Information Systems

- Provides input to be used in the managerial decision process. Deals with supporting well structured decision situations. Typical information requirements can be anticipated

Decision Support Systems

- Provides information to managers who must make judgements about particular situations. Supports decision-makers in situations that are not well structured

Organizational Needs

- **Better Strategic Decision Making & Planning**
- **Convert Data / Information into Business Intelligence**
- **Manipulate Data Analytically to obtain Business Intelligence**

Why Data Warehouse?

- Management Requirements to process data at corporate level using many files and collection of data that have been accumulated over the years
- Limitations in operational systems
- Reporting against un integrated operational data can be hazardous
- Reporting against data that is distributed in multiple sources and are incompatible with each other
- Make operational data accessible and easily efficiently queried so that management can get answers to business questions

Operational (TP) Systems & Data

- Application oriented
- Support day to day operations
- Focused on operational efficiency

Operational (TP) Systems & Data

- Not focused on complex queries required by the management
- Not robust enough to meet future needs
- The data serving operational needs is physically different data from that serving informational or analytic needs

Online Transaction Processing (OLTP) Systems

Designed to get data in quickly and to analyse current events.

Characterised by:

- Process oriented
- Data Normalised
- Current data
- Volatile data
- Updated in real-time

Management Decision Support Requirements

Examples

- Sales by Region, Country, Product and by Quarter
- Customer purchasing trends
- What are the most popular products purchased by customer between the ages 15 to 30?

Management Decision Support Requirements

Examples

- Sales by Region, Country, Product and by Quarter
- Customer purchasing trends
- What are the most popular products purchased by customer between the ages 15 to 30?

Data Warehouse Systems

Designed to get data out and quickly analyse.

Characterised by:

- Subject oriented rather than process orientated
- Integrated across subjects and entire enterprise
- De-normalised data
- Time-variant
- Historical
- Non Volatile
- Atomic and Summary data

Data Warehouse

"A data warehouse is a subject-oriented, integrated, time–variant and non-volatile collection data in support of management's decision making process" [W.H.Inmon 96].

The four keywords: subject-oriented, integrated, time-variant and non-volatile

distinguish data warehouse form other data repository systems, such as relational database systems,

transaction processing systems, and file systems .

Data Warehouse

- **Subject-oriented:** A data warehouse is organizes around major subjects, such as customer, supplier, product and sales.
- Rather than concentrating on the day –to-day operations and transaction processing of an organisation, a data warehouse focuses on the modelling and analysis of data for decision makers.
- Hence, data warehouse typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse

- **Integrated:** A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
- Data cleaning data integration techniques are applied to ensure consistency naming conventions, encoding structures, attribute measures, and so on

Data Warehouse

- **Time variant**: Data are stored to provide information from a historical perspective (e.g., the past 5-10 years).
- Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Data Warehouse

- **Non-volatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
- Due to this separation, a data warehouse does not require transaction processing recovery, and concurrency control mechanisms.
- It usually requires only two operations in data accessing: *initial loading* and *access* of data.

Operational Database (OD) versus Data Warehouse (DW)

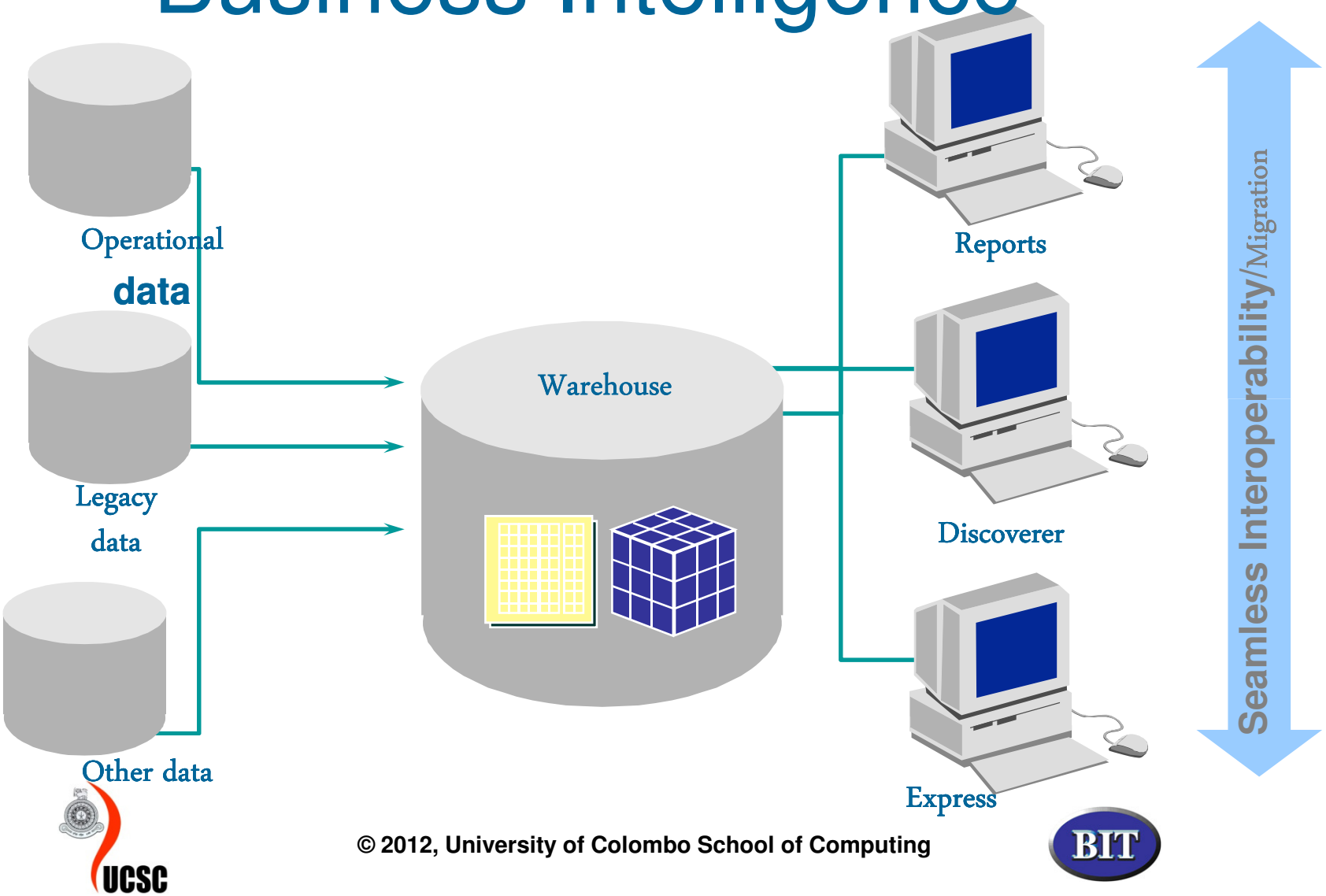
- Data in the DW is stored primarily for the purpose of providing data that can be interrogated by business people to gain value from information derived from daily operations.
- Use of the DW is to drive decision support.
- The OD is used to process information that is needed for the purposes of performing operational tasks.
- The OD is active for updates during all hours that business activities are executed. The DW is used for read-only querying during active business hours.

Data Mart

A departmentalized structure of data feeding from the Data Warehouse where data is denormalized based on the department's need for information.

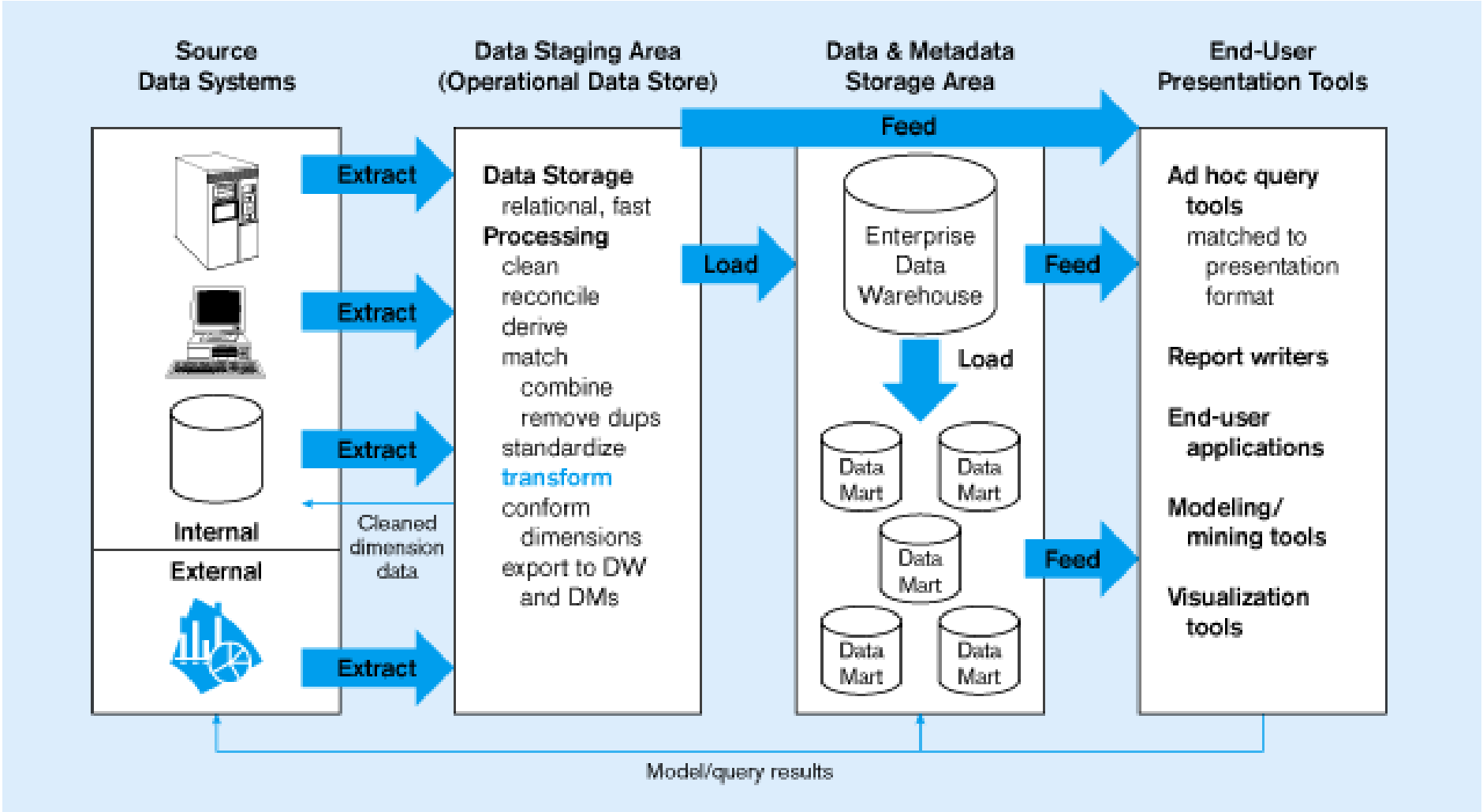
Data marts are generally small data warehouses and uses the same structures and many of the same development methods as data warehouses. The difference is they are intend to meet a specific need or to deliver only one type of information.

An Integrated Environment for Business Intelligence



DW Designing- Main Considerations

- Data Warehouse or Data mart
- Data extraction, validate and loading
- Data models in the source
- Dimensional model in the DW
- Managing data volumes in the DW
- Refreshing data in the DW

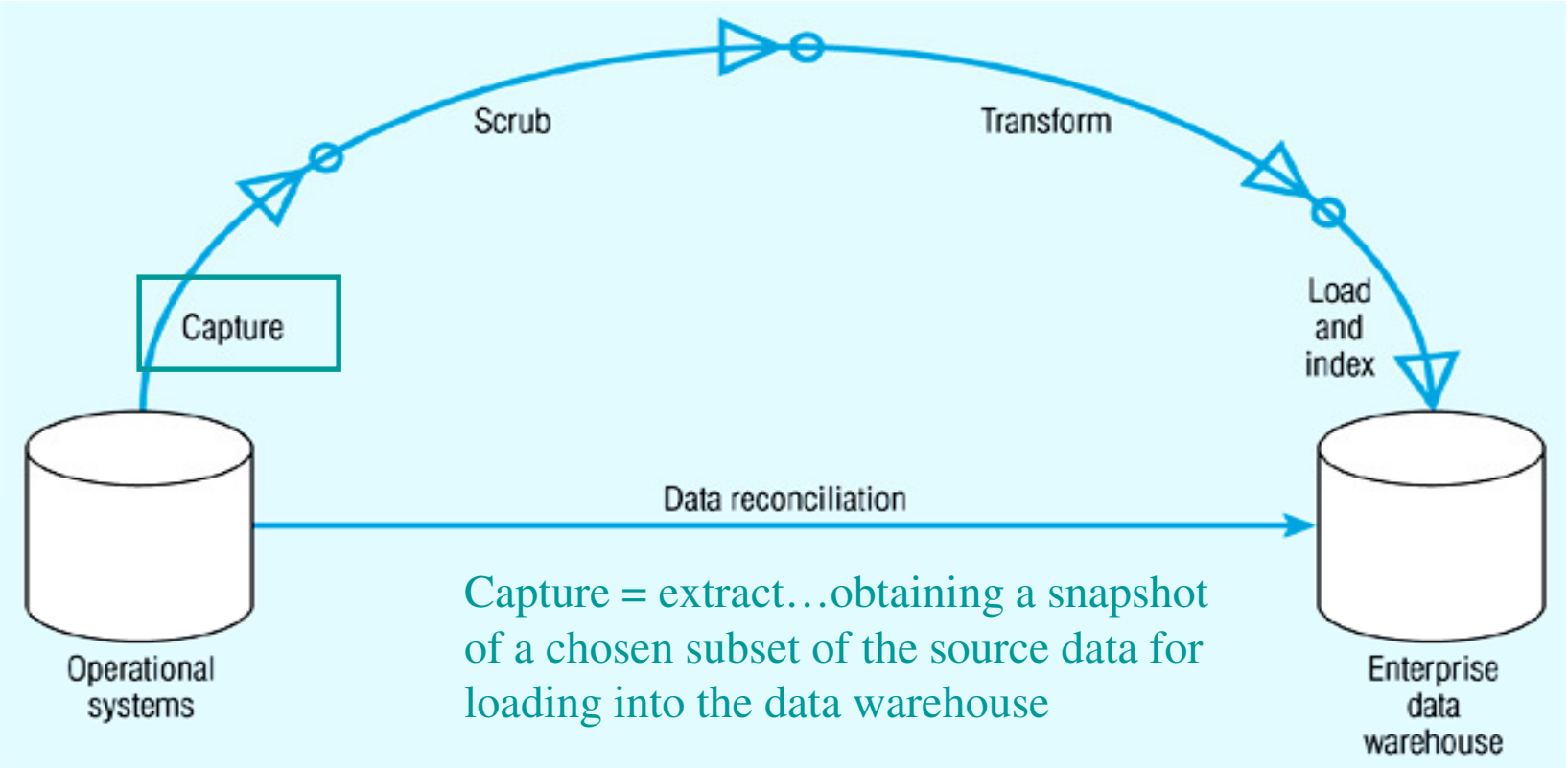


The ETL Process

- Capture
- Scrub or data cleansing
- Transform
- Load and Index

ETL = Extract, Transform and Load

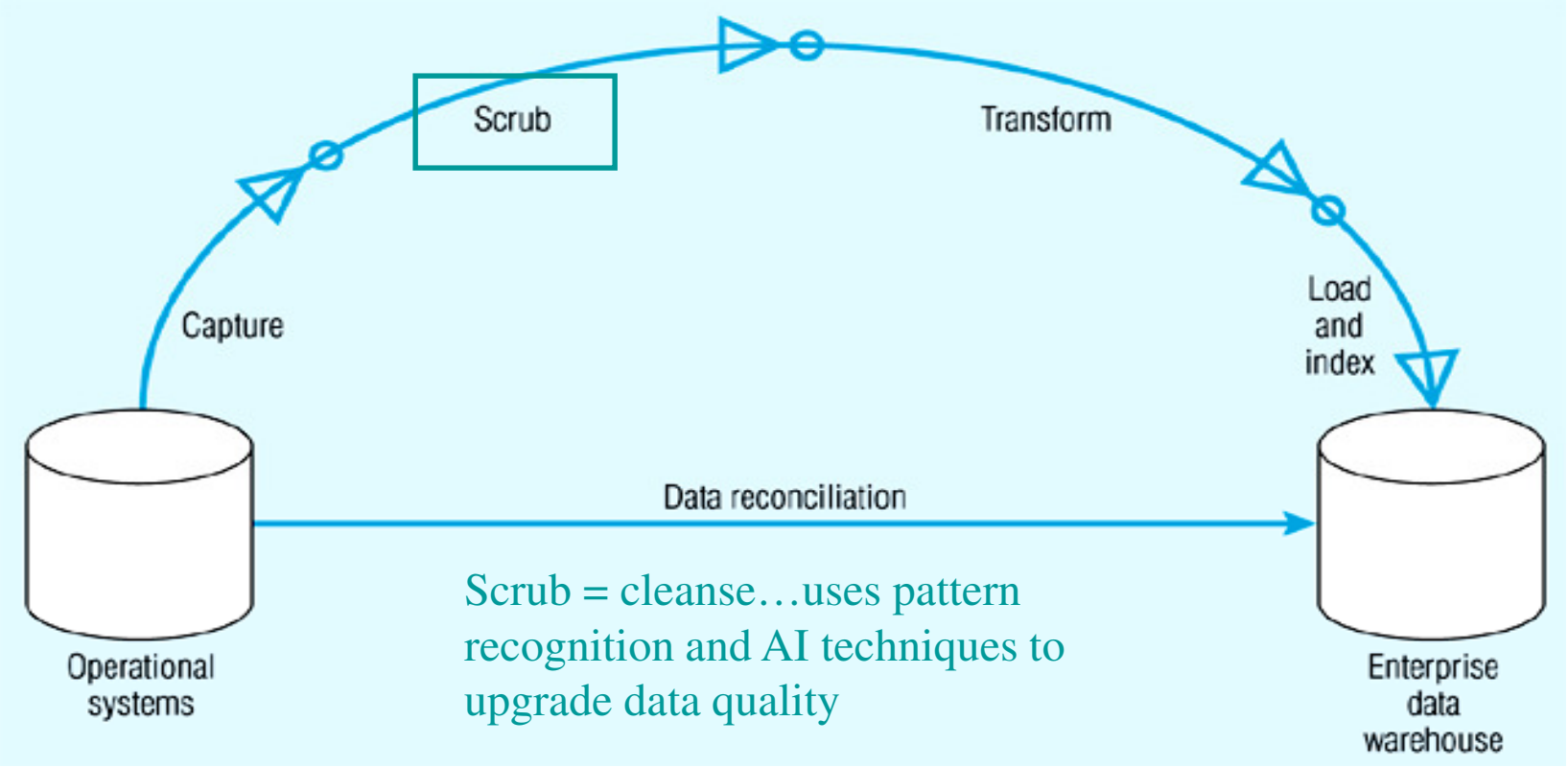
Steps in data reconciliation



Static extract = capturing a snapshot of the source data at a point in time

Incremental extract = capturing changes that have occurred since the last static extract

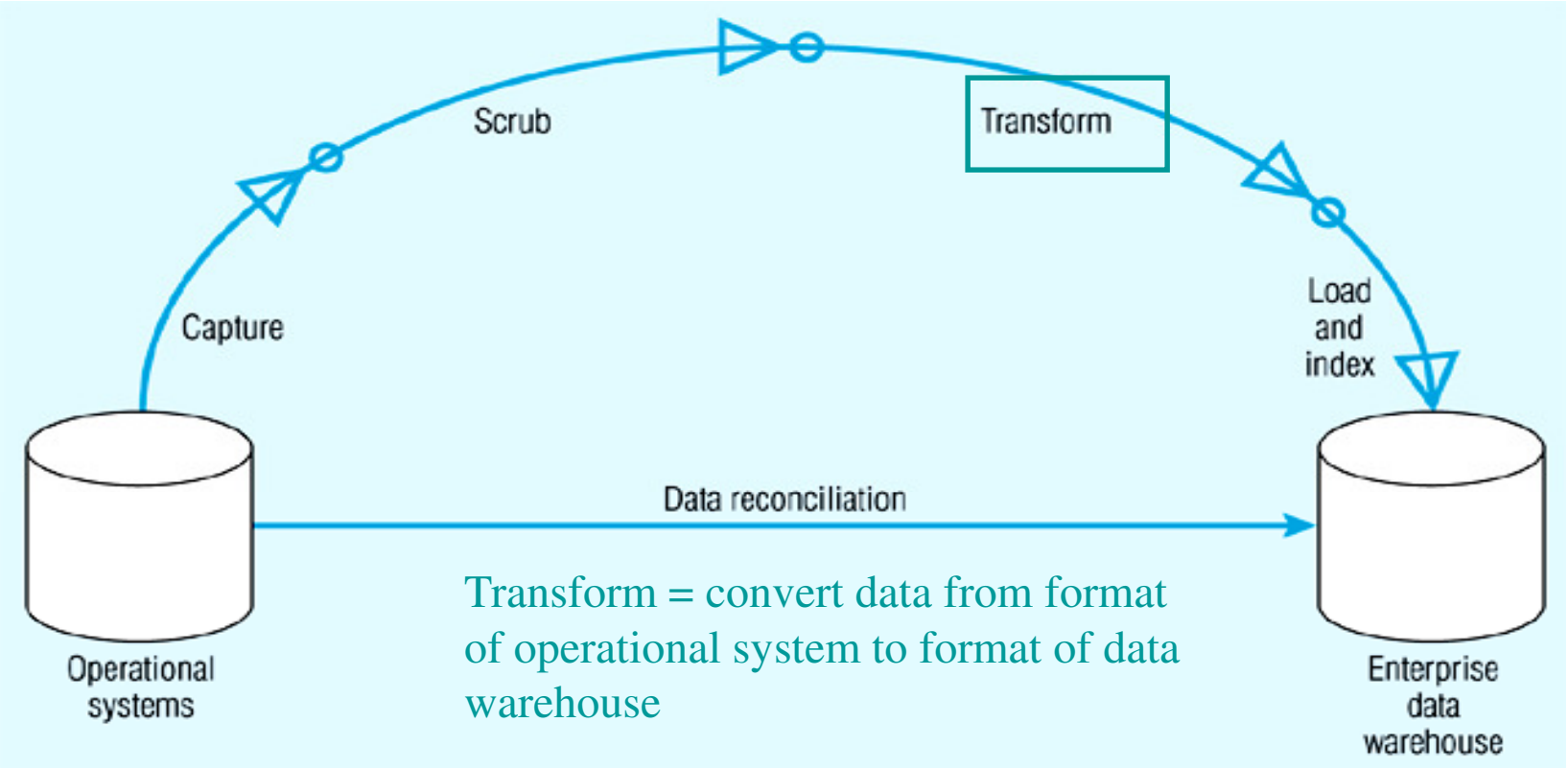
Steps in data reconciliation (continued)



Fixing errors: misspellings, erroneous dates, incorrect field usage, mismatched addresses, missing data, duplicate data, inconsistencies

Also: decoding, reformatting, time stamping, conversion, key generation, merging, error detection/logging, locating missing data

Steps in data reconciliation (continued)



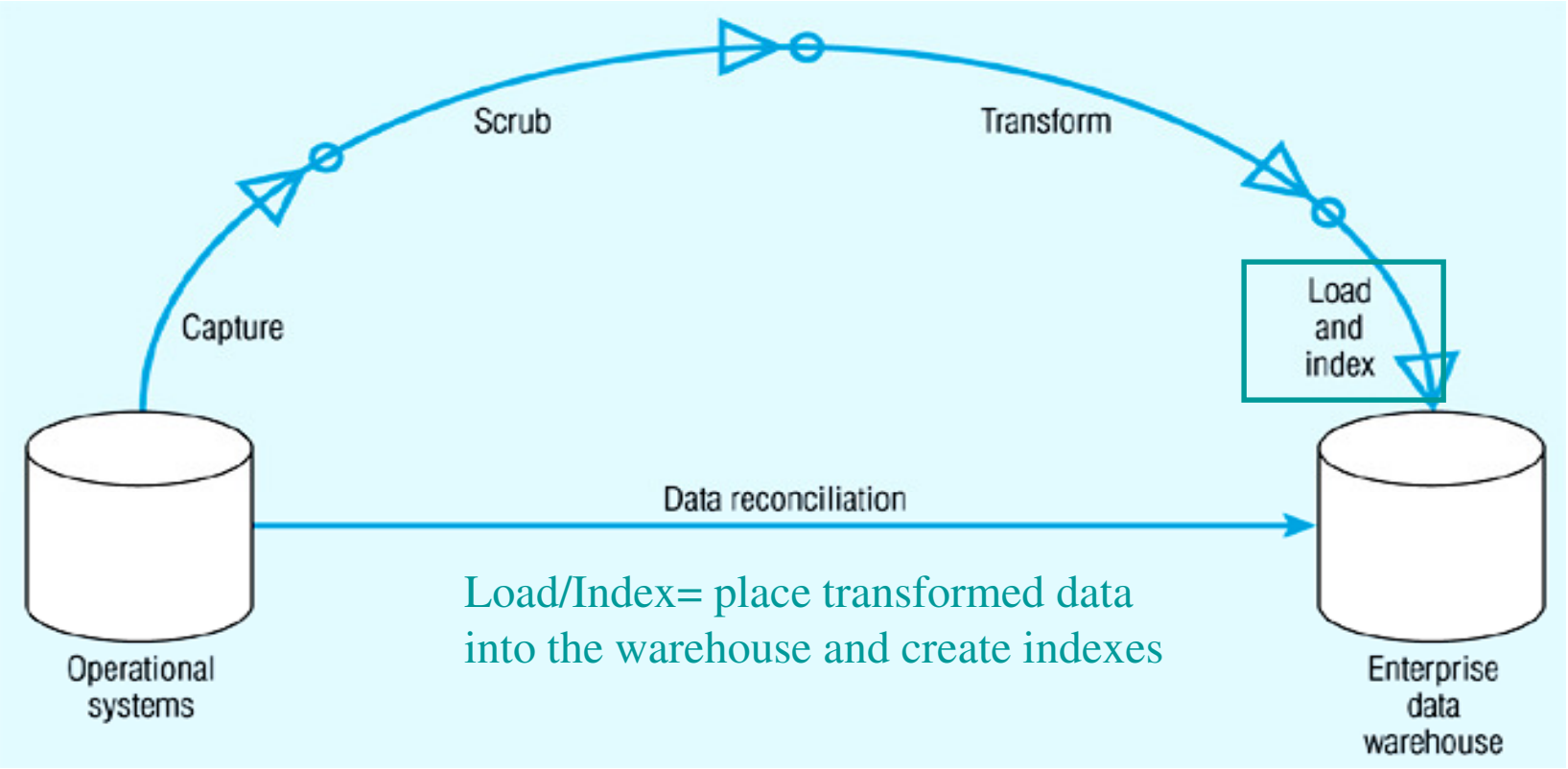
Record-level:

Selection – data partitioning
Joining – data combining
Aggregation – data summarization

Field-level:

single-field – from one field to one field
multi-field – from many fields to one, or one field to many

Steps in data reconciliation (continued)



Refresh mode: bulk rewriting of target data at periodic intervals

Update mode: only changes in source data are written to data warehouse

Extract, Transform & Load (ETL) Process

The process of integrating data from the operational systems to DW.

This is a complex, time consuming & error-prone

- Availability of data
- Quality of data
- Granularity

ETL Process - Problems

- Hard Coded names
- Inconsistent field lengths
- Missing values
- Inconsistent values
- Inconsistent field types
- Inconsistent entity handling
- Change in technology
- Resolving conflicts in multiple input files

ETL Software Tools

ETL tools extract data that resides in disparate sources such as relational data bases, mainframe systems or packaged applications, transform it, and load it into data marts and warehouses.

Data Granularity

Granularity refers to the level of details or summarization of the units of data in the Data Warehouse.

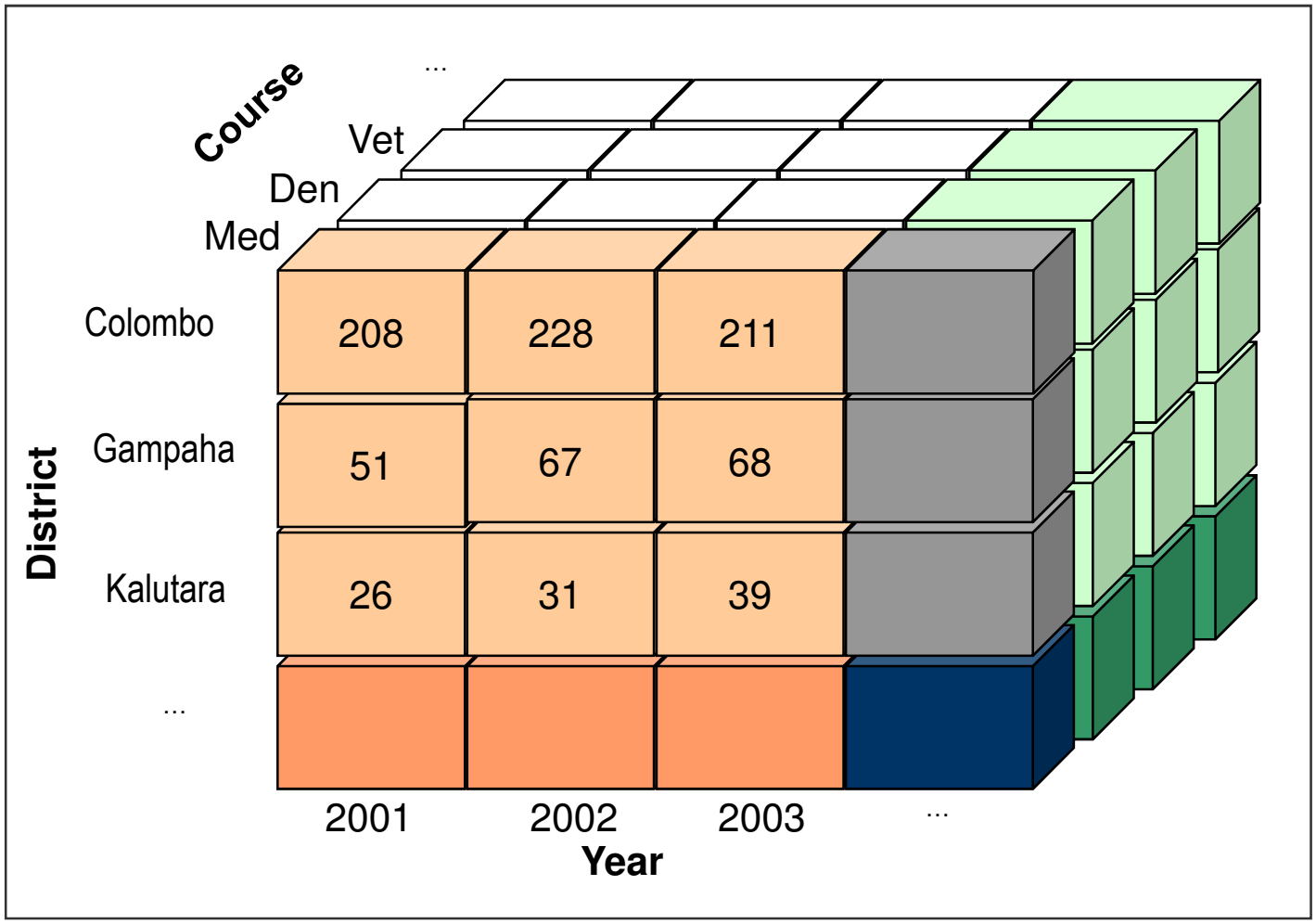
The more detail there is, the lower the level of granularity.
The less detail there is, the higher the level of granularity.

Granularity has an impact on:

- **Volume of data in the Data Warehouse**
- **Types of Queries & Reports from the DW**

DATA CUBES

- After defining the star schema we ca create so many cubes according to the requirement.
- For example, as in figure.



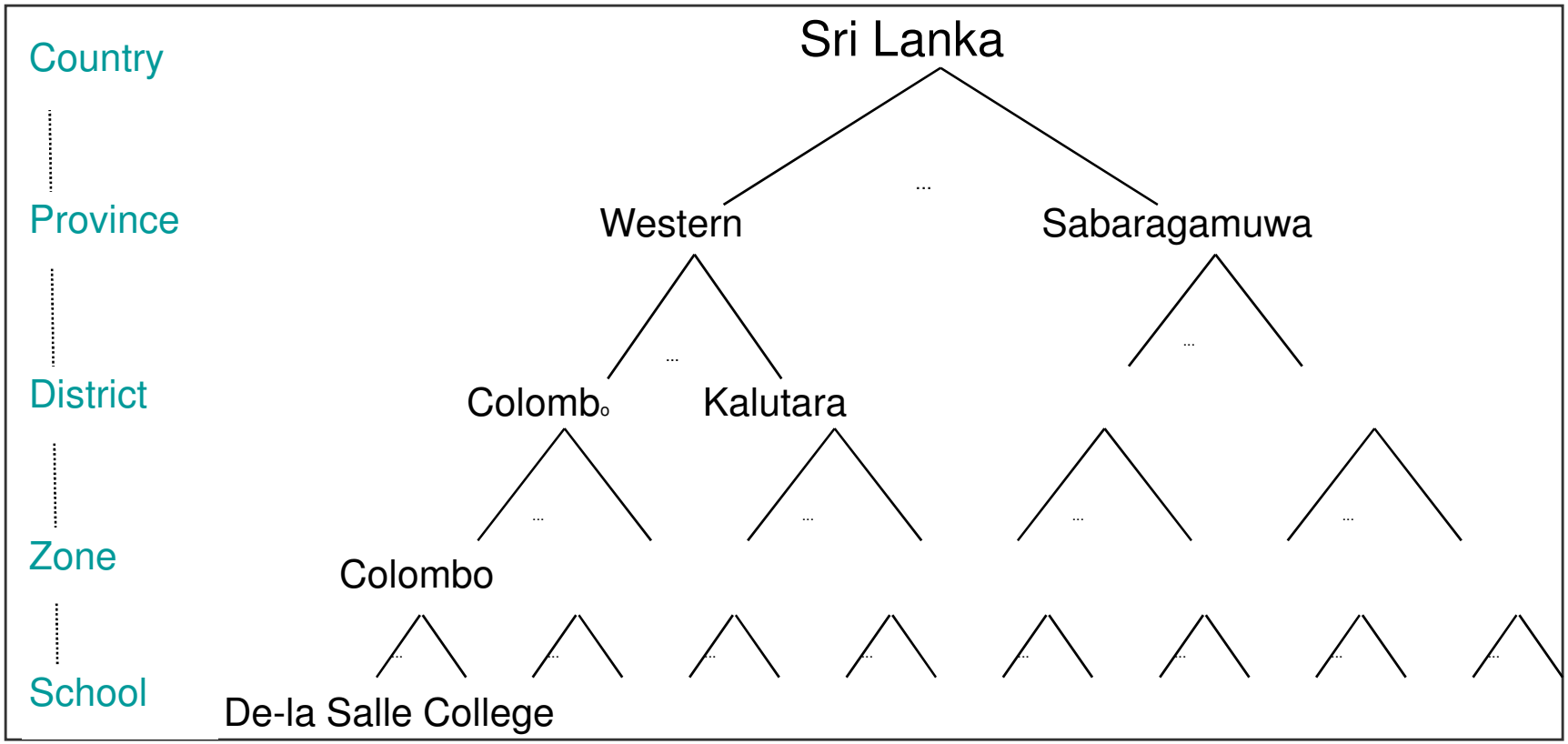
A Data Cube

- A concept hierarchy defines a sequence of mappings from a set of low-level concepts to more general higher-level concepts.
- Using it data could be aggregated or disaggregated. Many concept hierarchies are implicit within the database schema and a hierarchy could be defined for the locations in the order of school < zone < district < province < country.
- This allows districts to be aggregated to provinces (roll-up) and well as districts to be disaggregated into zones (drill-down).

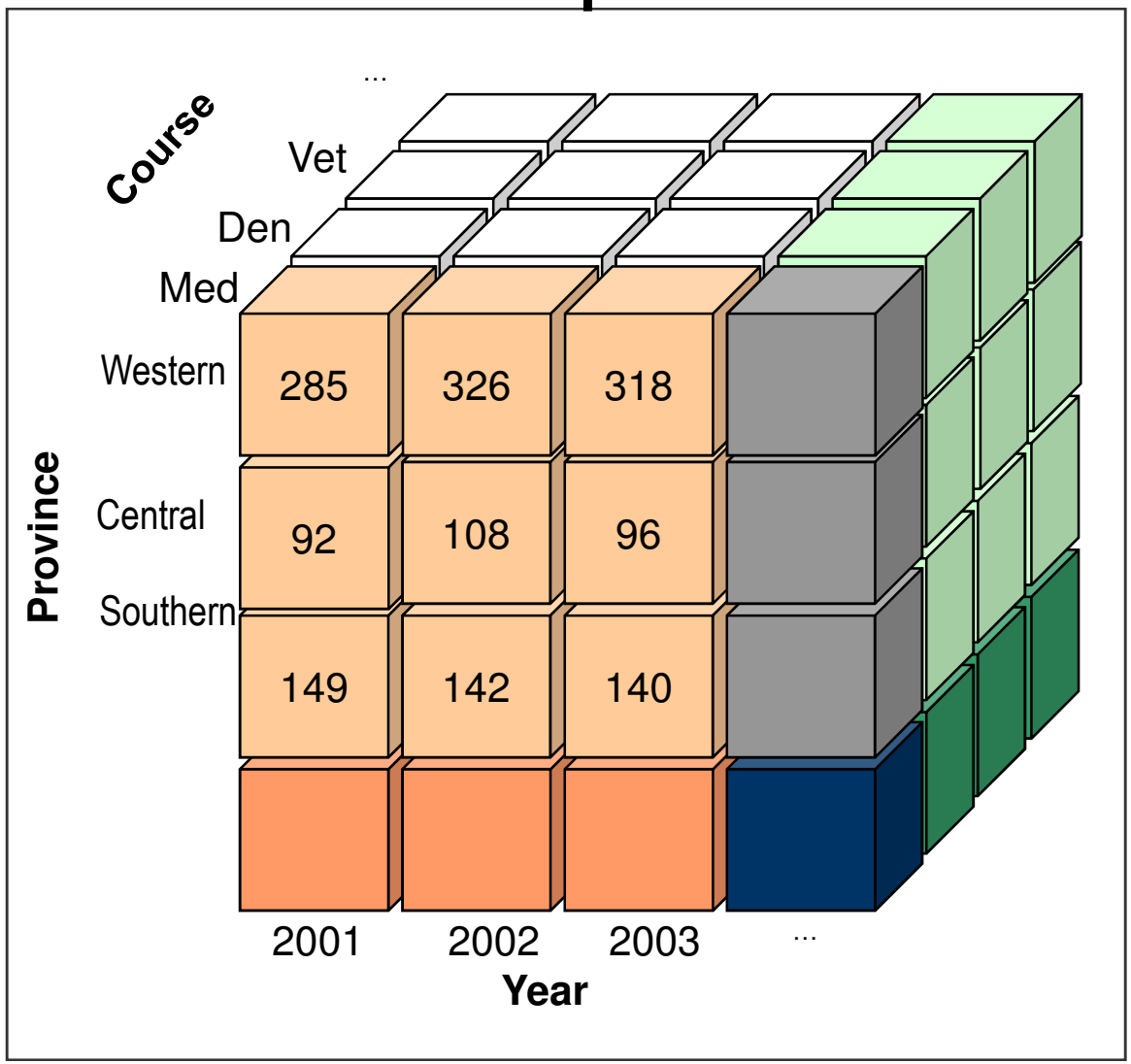
Roll-up

- The Roll-up operation corresponds to taking the current data objects and doing further grouping by one of the dimensions.
- The Roll-up operations performed on the central cube by climbing up the concept of hierarchy.
- Thus, it is possible to Roll-up admission data by grouping districts into provinces.

Concept Hierarchy

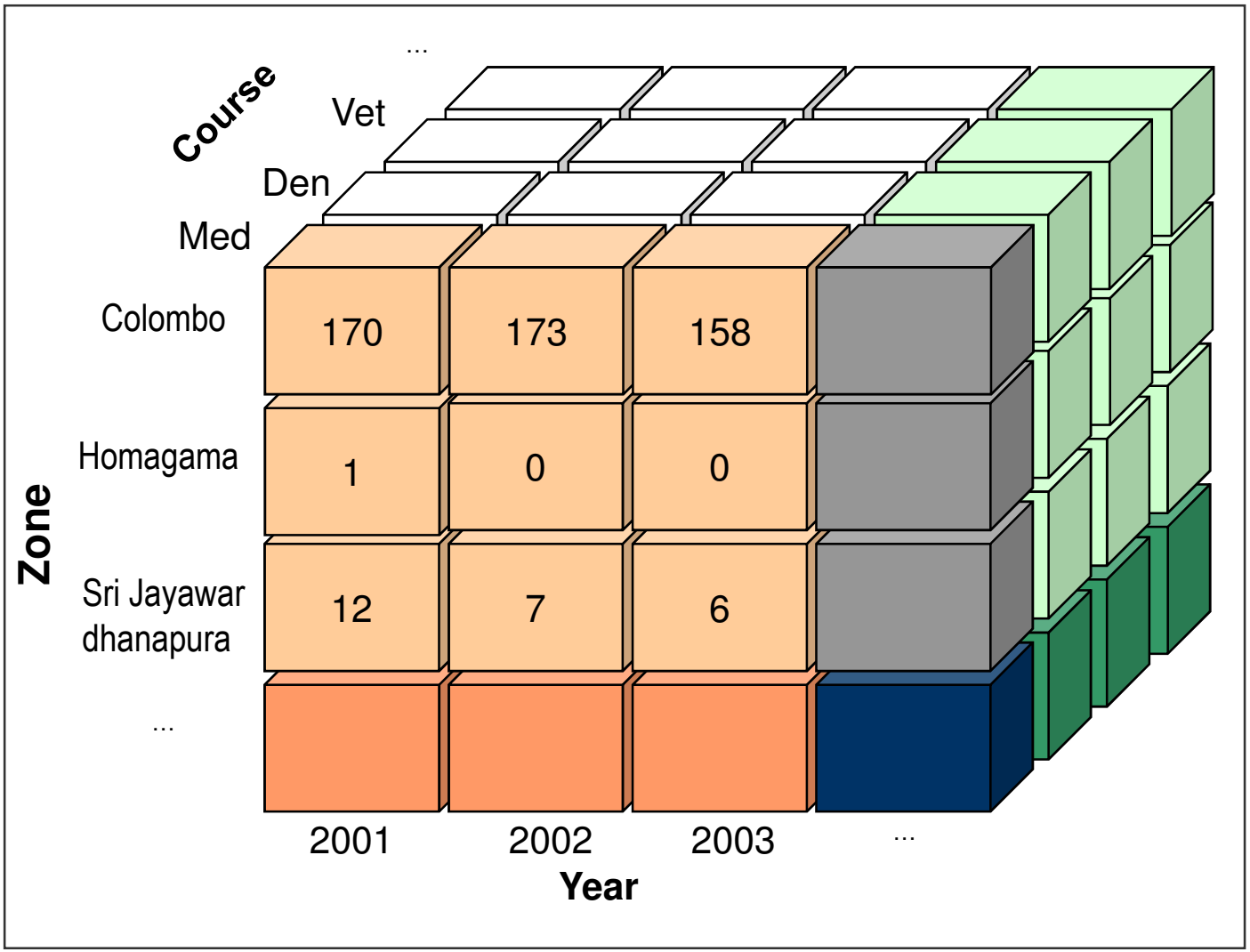


Roll-up District



Drill-down

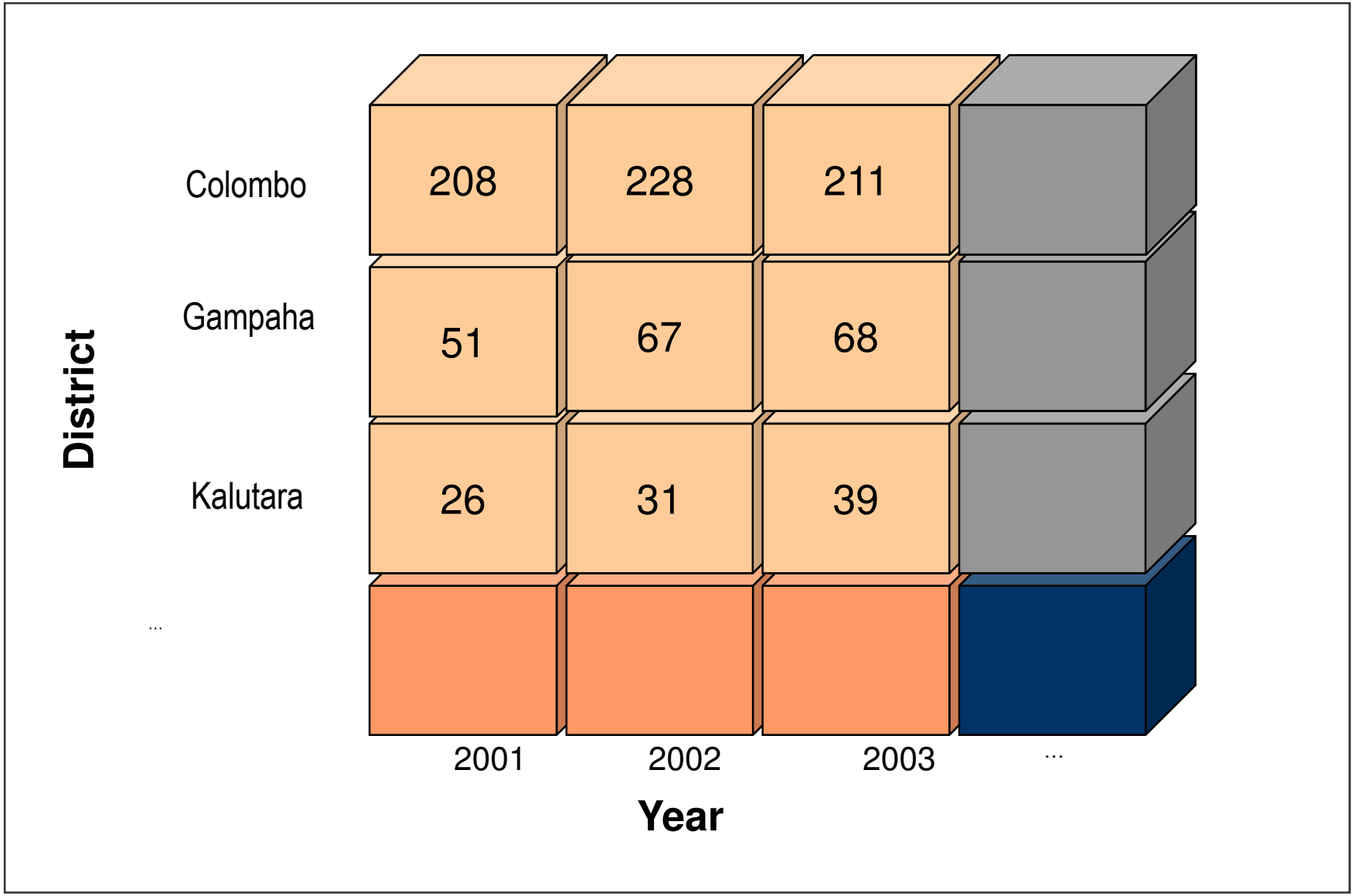
- The drill-down operation is the opposite of roll-up.
- It navigates from less detail data to more details. The data cube of figure 6 can be drill-down using the location concept-hierarchy and hence districts can be disaggregated into zones as shown in figure 9.



Drill-down District

Slice

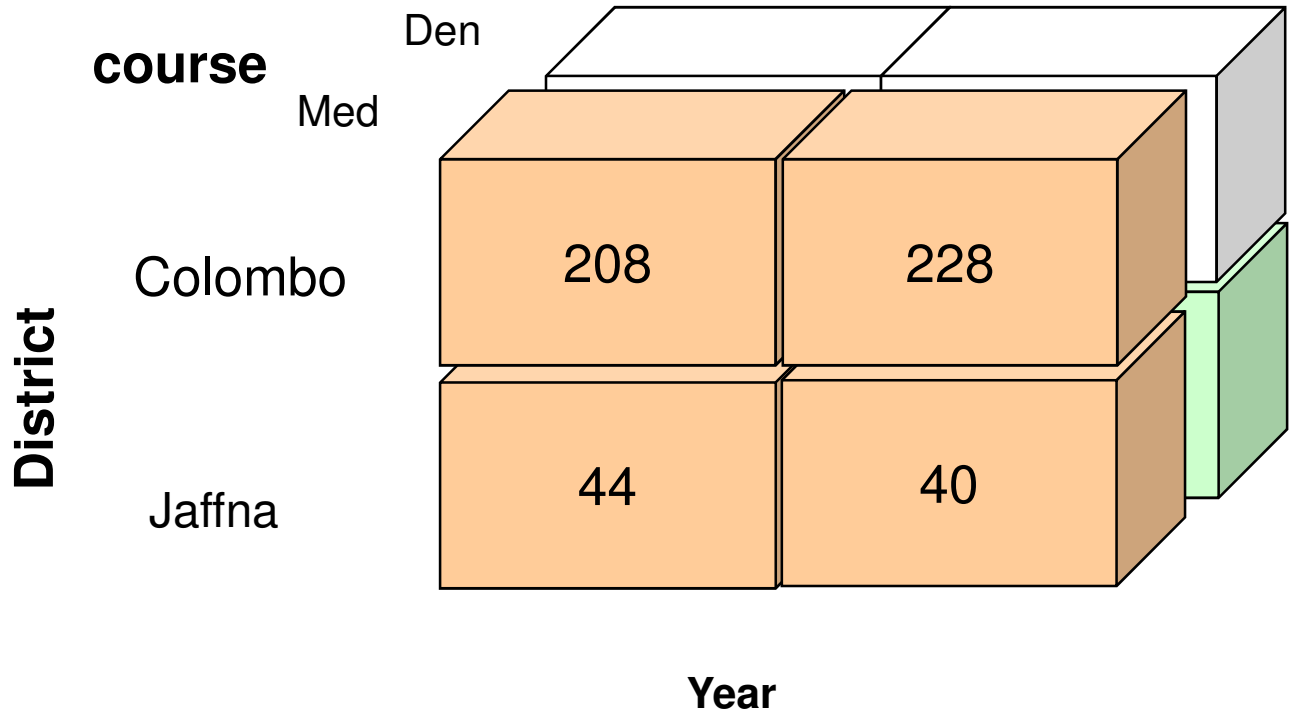
- The slice operation performs a selection on one dimension of the given cube, resulting is a sub-cube.
- For example, we could select the course dimension and slice for course medicine and view a sub-cube.



A Slice for Course Medicine

Dice

- The dice operations define a sub-cube by performing a selection of one or more dimensions.
- For example, three dimension dice for courses “Medicine” and Dental” for districts “Colombo” and “Jaffna” for year 2001 and 2002



A 3D Dice

Multidimensional Data Schema Support

- Decision Support Data tends to be
  - Nonnormalized
  - Duplicated
  - Preaggregated
- Star Schema
  - Special Design technique for multidimensional data representations
  - Optimize data query operations instead of data update operations

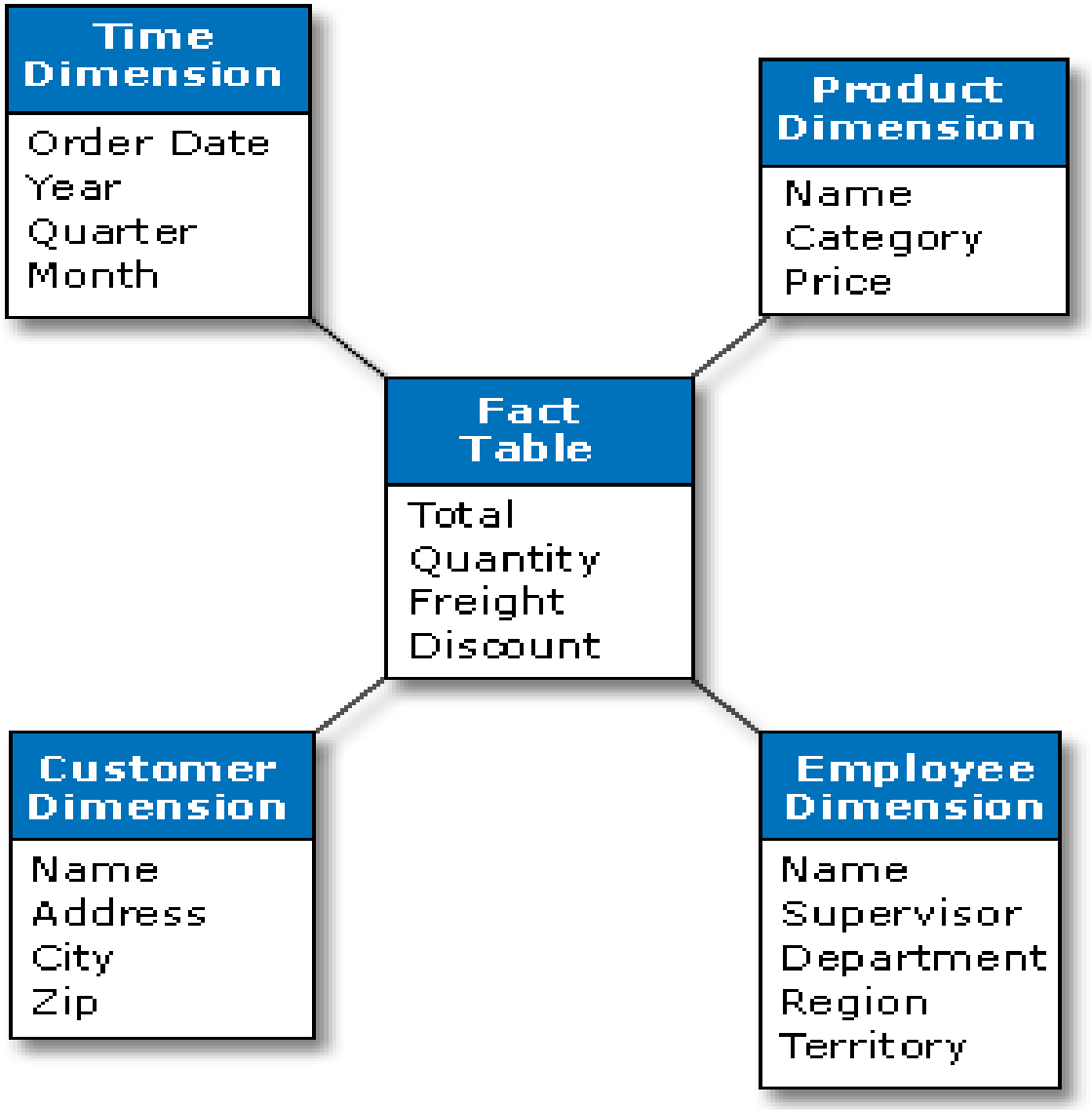
Star Schemas

- Data Modeling Technique to map multidimensional decision support data into a relational database
- Current Relational modeling techniques do not serve the needs of advanced data requirements

Star Schema

- 4 Components
  - Facts
  - Dimensions
  - Attributes
  - Attribute Hierarchies

STAR Schema



Facts

- Numeric measurements (values) that represent a specific business aspect or activity
- Stored in a fact table at the center of the star scheme
- Contains facts that are linked through their dimensions
- Can be computed or derived at run time
- Updated periodically with data from operational databases

Dimensions

- Qualifying characteristics that provide additional perspectives to a given fact
  - DSS data is almost always viewed in relation to other data
- Dimensions are normally stored in dimension tables

Attributes

- Dimension Tables contain Attributes
- Attributes are used to search, filter, or classify facts
- Dimensions provide descriptive characteristics about the facts through their attributed
- Must define common business attributes that will be used to narrow a search, group information, or describe dimensions. (ex.: Time / Location / Product)
- No mathematical limit to the number of dimensions (3-D makes it easy to model)

Attribute Hierarchies

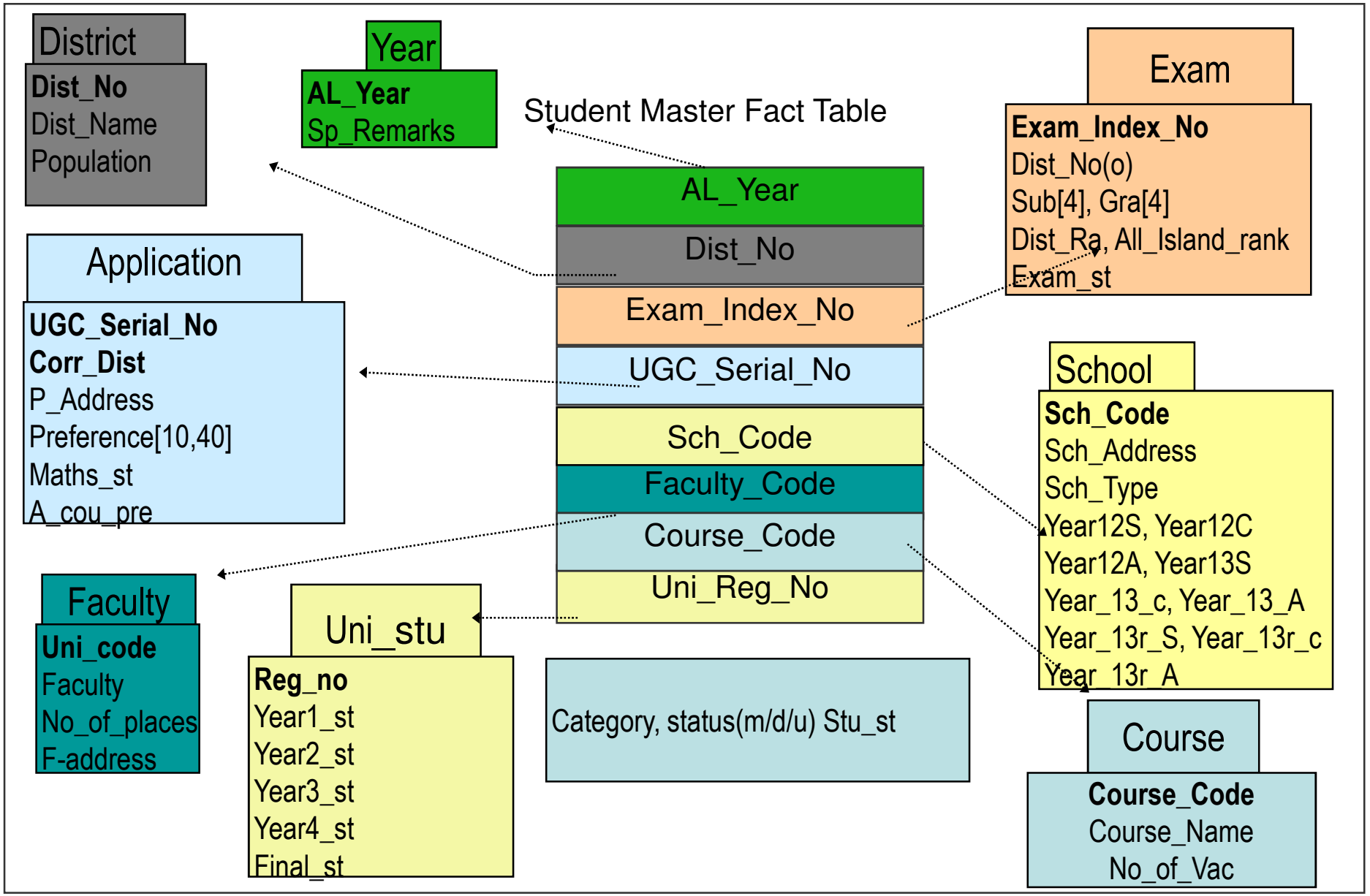
- Provides a Top-Down data organization
  - Aggregation
  - Drill-down / Roll-Up data analysis
- Attributes from different dimensions can be grouped to form a hierarchy

Star Schema

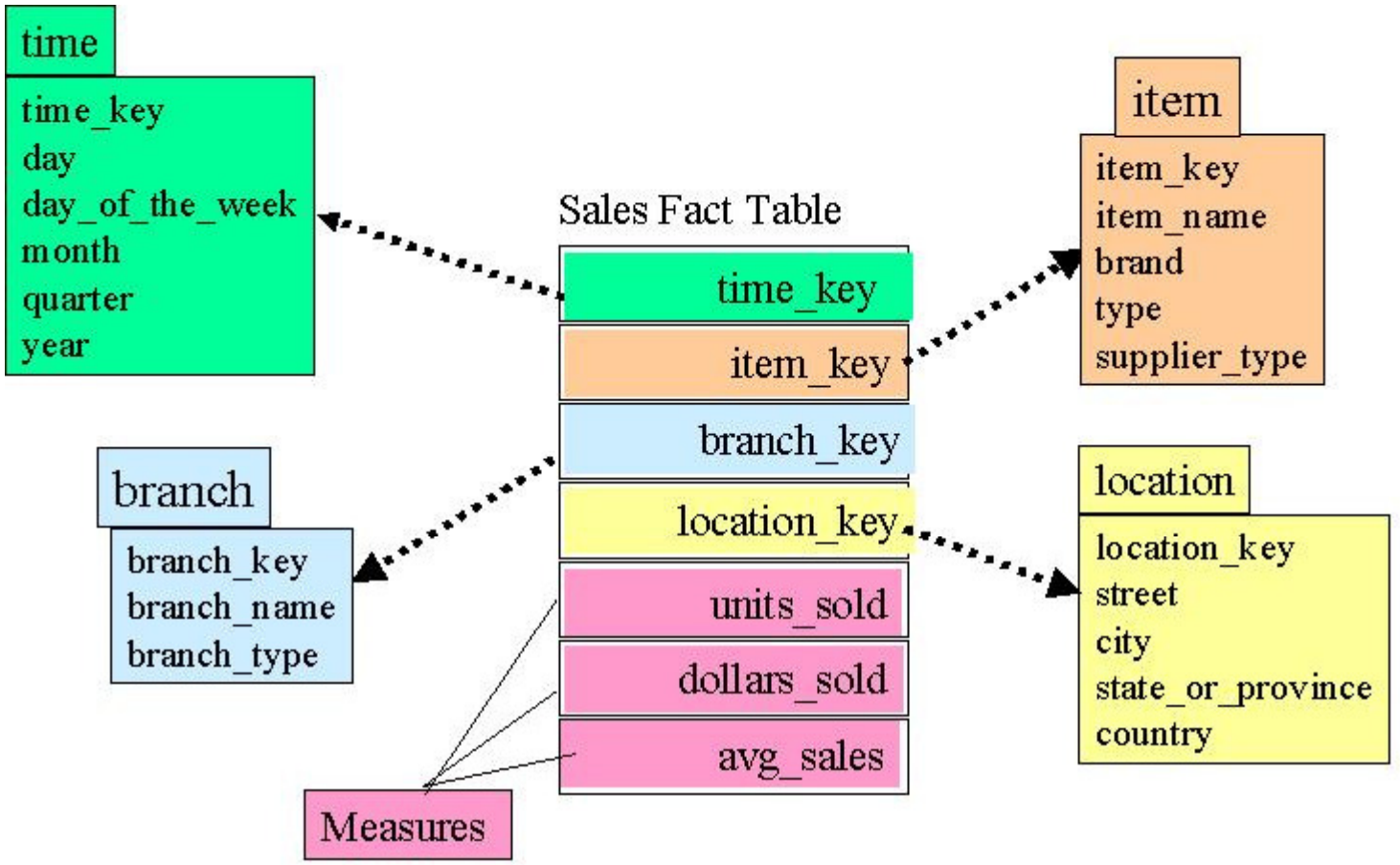
- A single fact table and for each dimension one dimension table
- Does not capture hierarchies directly

Snowflake schema

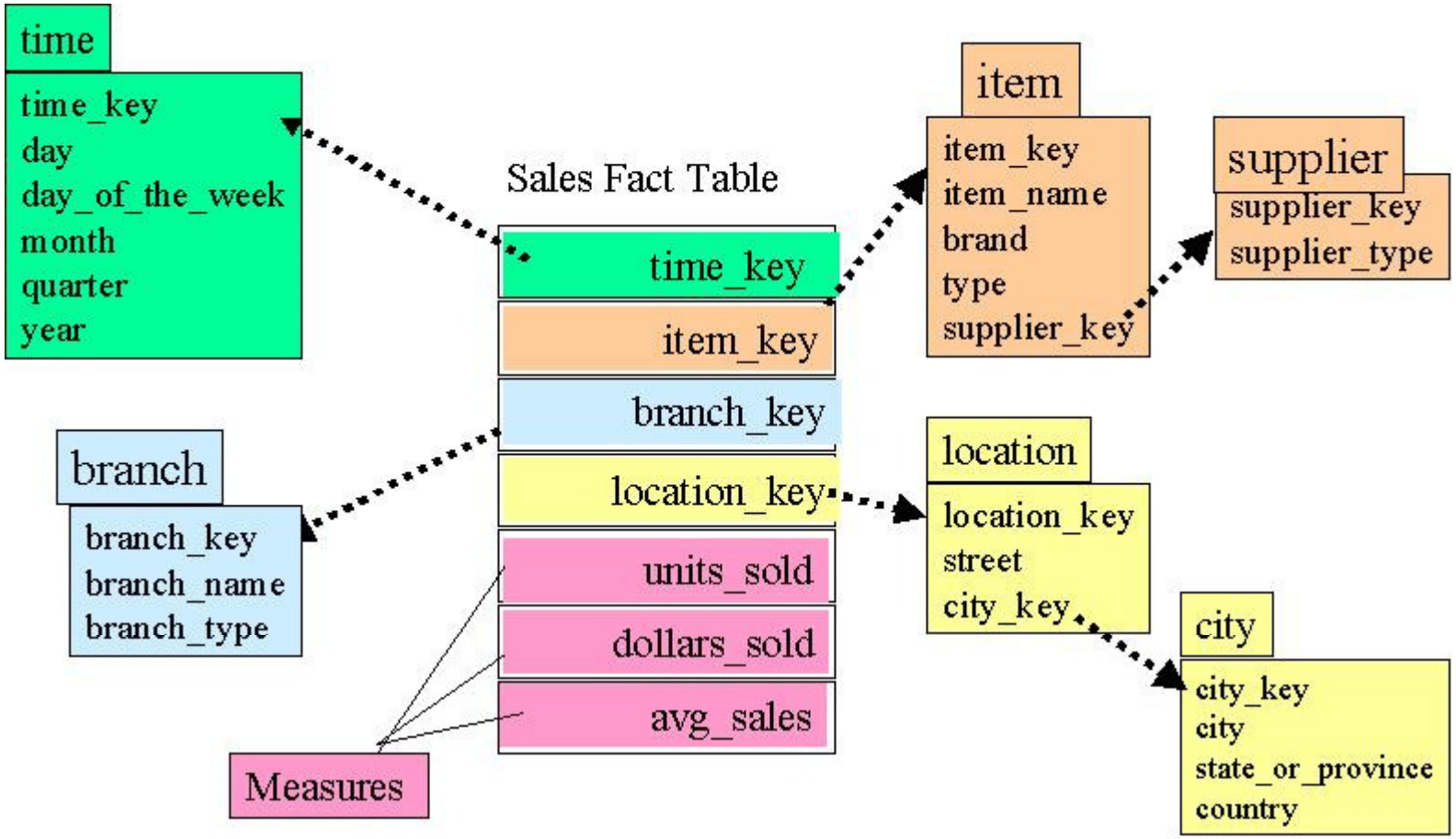
- Represent dimensional hierarchy directly by normalizing tables.
- Easy to maintain and saves storage



Star Schema



Star Schema



Snowflake schemas normalize dimensions to eliminate redundancy. That is, the dimension data has been grouped into multiple tables instead of one large table.

E.g. Location and Item dimension table in a star schema might be normalized into a location table and city table in a snowflake schema.

- define cube cube_master [AL_Year, Exam_Index_No, UGC_Serial_No, Uni_Reg_No, Faculty_Code, Course_Code, Sch_Code, Dist_No] : stu_status= count(*)
- define dimension Year as (AL_Year, Sp_Remarks)
- define dimension District as (Dist_No, Dist_Name, Population)
- define dimension Course as (Course_Code, Course_Name, No_of_Vac)

Part of DMQL statements

Star Schema Representation

- Fact and Dimensions are represented by physical tables in the data warehouse database
- Fact tables are related to each dimension table in a Many to One relationship (Primary/Foreign Key Relationships)
- Fact Table is related to many dimension tables
  - The primary key of the fact table is a composite primary key from the dimension tables
- Each fact table is designed to answer a specific DSS question

Star Schema

- The fact table is always the larges table in the star schema
- Each dimension record is related to thousand of fact records
- Star Schema facilitated data retrieval functions
- DBMS first searches the Dimension Tables before the larger fact table

Data Warehouse Implementation

- An Active Decision Support Framework
  - Not a Static Database
  - Always a Work in Process
  - Complete Infrastructure for Company-Wide decision support
  - Hardware / Software / People / Procedures / Data
  - Data Warehouse is a critical component of the Modern DSS – But not the Only critical component

Software for DW & BI

- Oracle Data Warehouse Builder
- Oracle Discoverer
- Business Objects
- Cognos
- MS SQL Server Analysis Services

Business Objects – Complete BI Platform

Main Components

- **Reporting (Crystal Reports, Crystal Reports Analyzer)**
- **Query & Analysis**
- **OLAP Intelligence**
- **Web Intelligence**
- **Dash Board Manager**
- **Data Integrator (ETL)**

Data Warehousing Applications

- Decision support
- Trend analysis
- Financial forecasting
- Churn Prediction for Telecom subscribers, Credit Card users etc.
- Insurance fraud analysis
- Call record analysis
- Logistics and Inventory management
- Agriculture

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- What is not data mining?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

UCSC

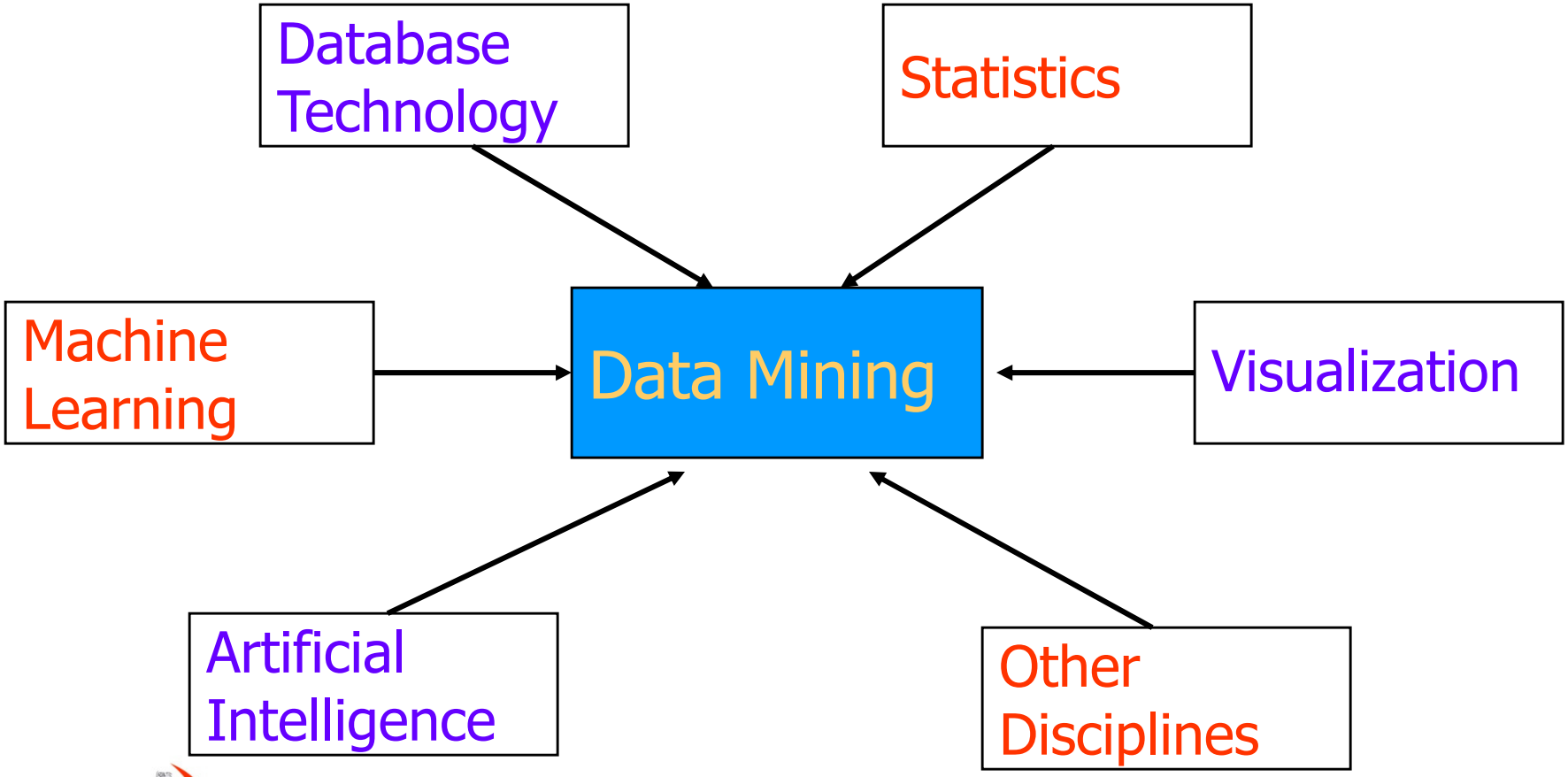
Data Mining

- Discover Previously unknown data characteristics, relationships, dependencies, or trends
- Typical Data Analysis Relies on end users
  - Define the Problem
  - Select the Data
  - Initial the Data Analysis
  - Reacts to External Stimulus

Data Mining

- Proactive
- Automatically searches
  - Anomalies
  - Possible Relationships
  - Identify Problems before the end-user
- Data Mining tools analyze the data, uncover problems or opportunities hidden in data relationships, form computer models based on their findings, and then user the models to predict business behavior – with minimal end-user intervention

Multidisciplinary Field



Data Mining

- A methodology designed to perform knowledge-discovery expeditions over the database data with minimal end-user intervention
- 3 Stages of Data
  - Data
  - Information
  - Knowledge

Why Data Mining? — Potential Applications

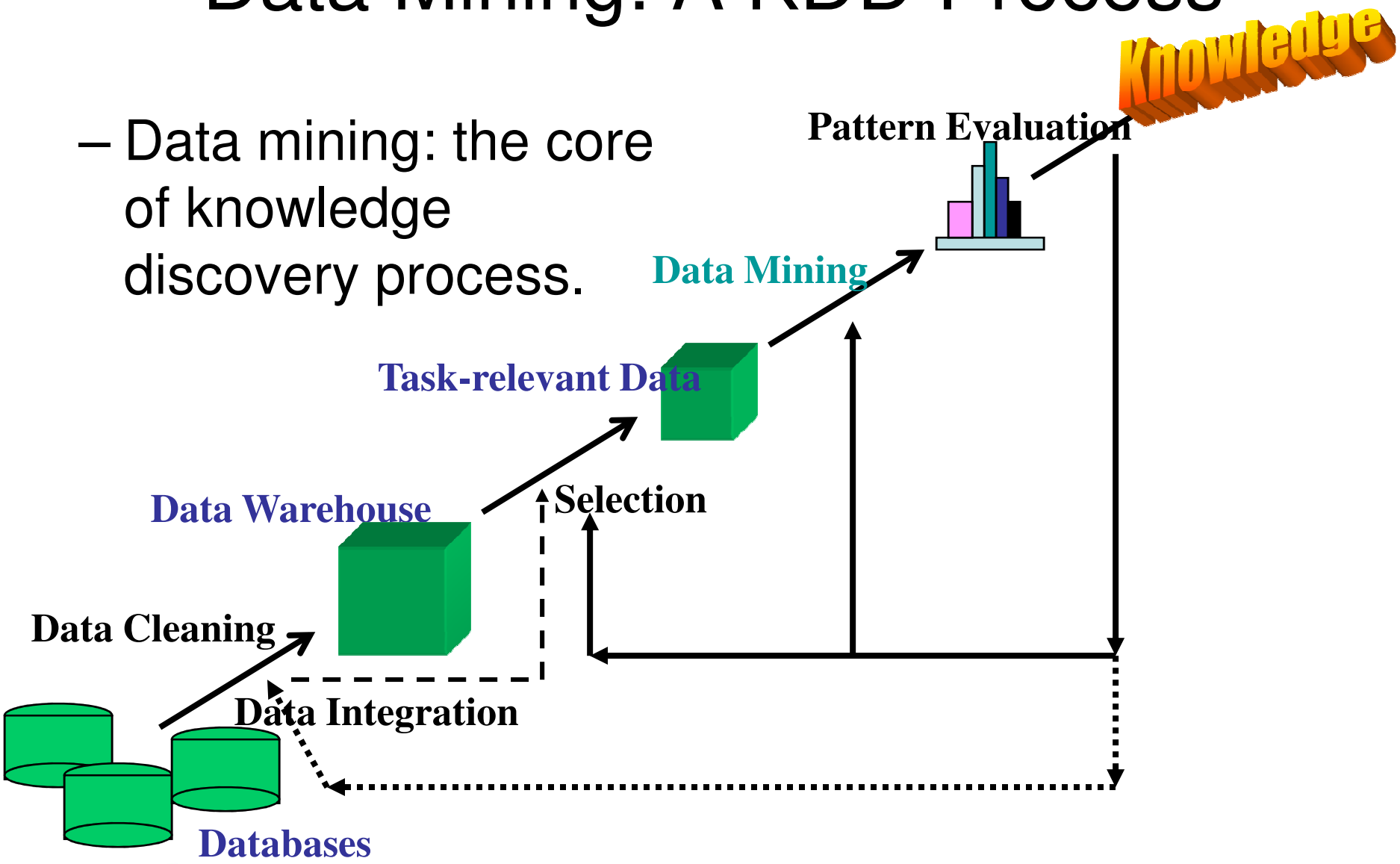
- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management

Applications

- Customer analytics
  - Forecasting buying habits and lifestyle preferences is a process of data mining and analysis.
- Data Mining in Agriculture
- National Security Agency
- Police-enforced ANPR in the UK
- Quantitative structure-activity relationship Surveillance / Mass surveillance
- Processing Loan Applications
- Stock and investment analysis
- Identify successful medical therapies

Data Mining: A KDD Process

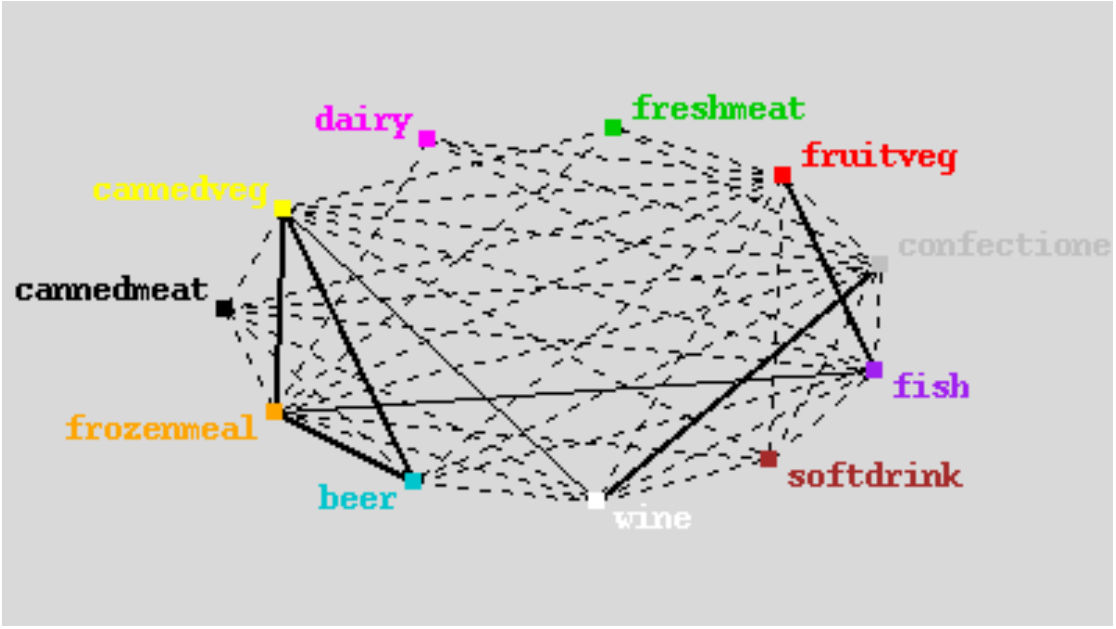
– Data mining: the core of knowledge discovery process.



Association

Description

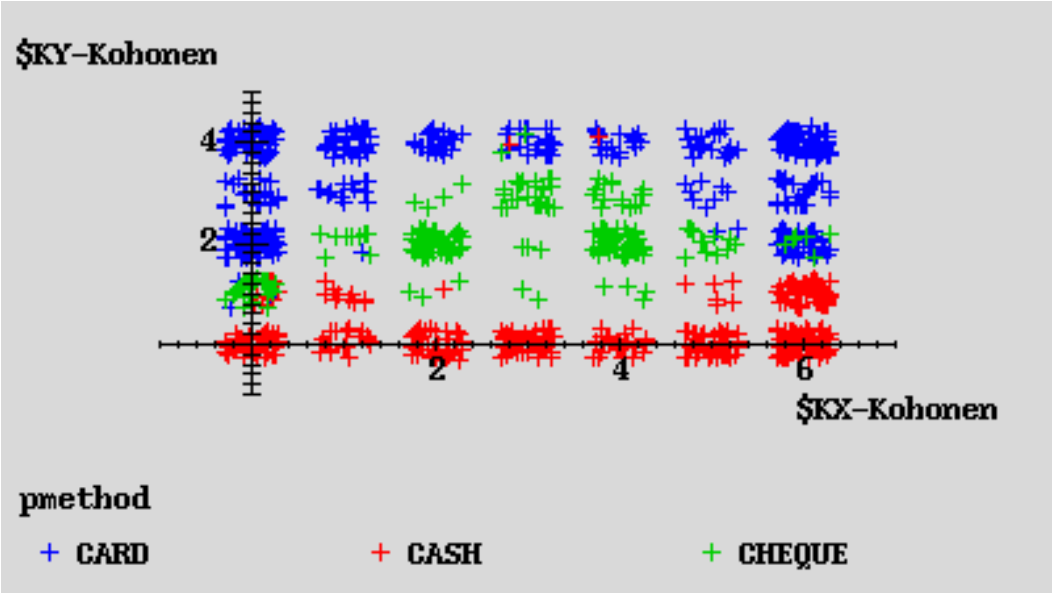
- Seeks *association rules* in dataset
- ‘Market basket’ analysis
- Sequence discovery



Clustering

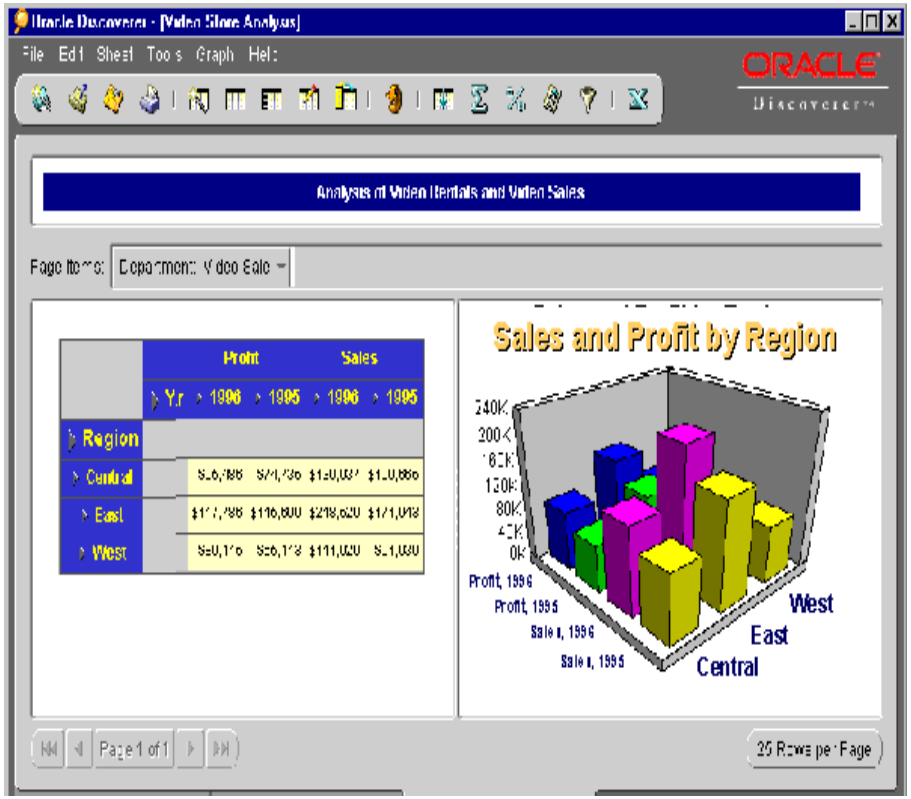
Description

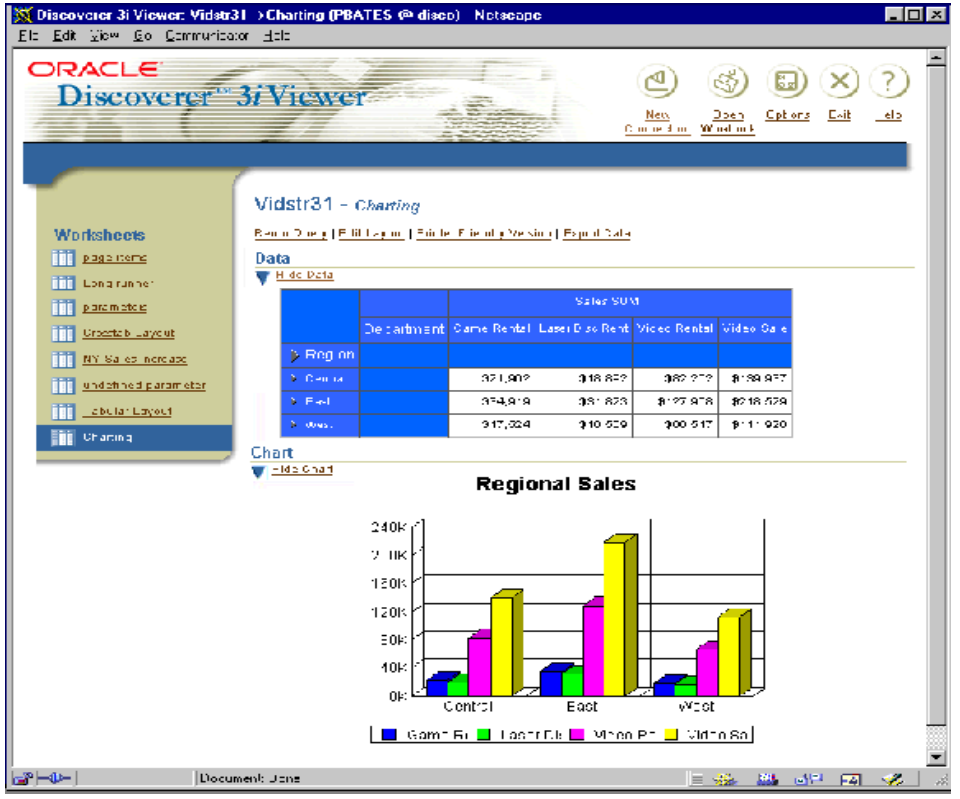
- unsupervised
- seeks to *describe* dataset in terms of natural *clusters* of cases



Discoverer

"An ad hoc query, reporting, and analysis tool"





Reports

"A sophisticated enterprise production reporting tool to build and distribute high-quality reports"