Data Warehouse and Data Mining (CSC-410)

Compiled By: Madan Nath
BSc. CSIT 7th Semester

Course Overview

Course Title: Data Warehousing and Data Mining

Course no: CSC-410

Credit hours: 3

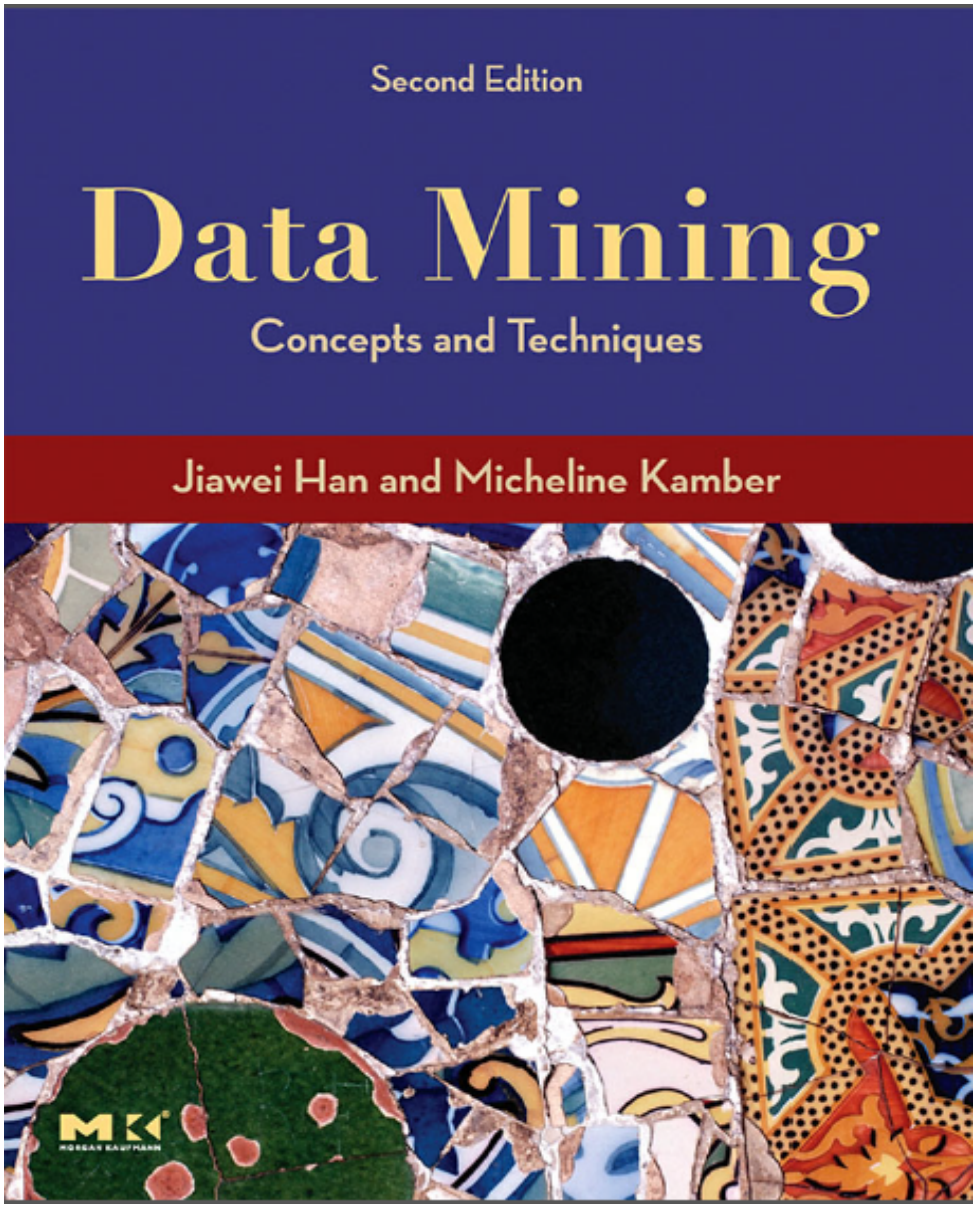
Nature of course: Theory (3 Hrs.) + Lab (3 Hrs.)

Full Marks: 60+20+20

Pass Marks: 24+8+8

Prerequisite: C, Data Structure, Database

Text Books


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber
MK
MORGAN KAUFMANN

Microsoft

Data Mining

with Microsoft® SQL Server™ 2000

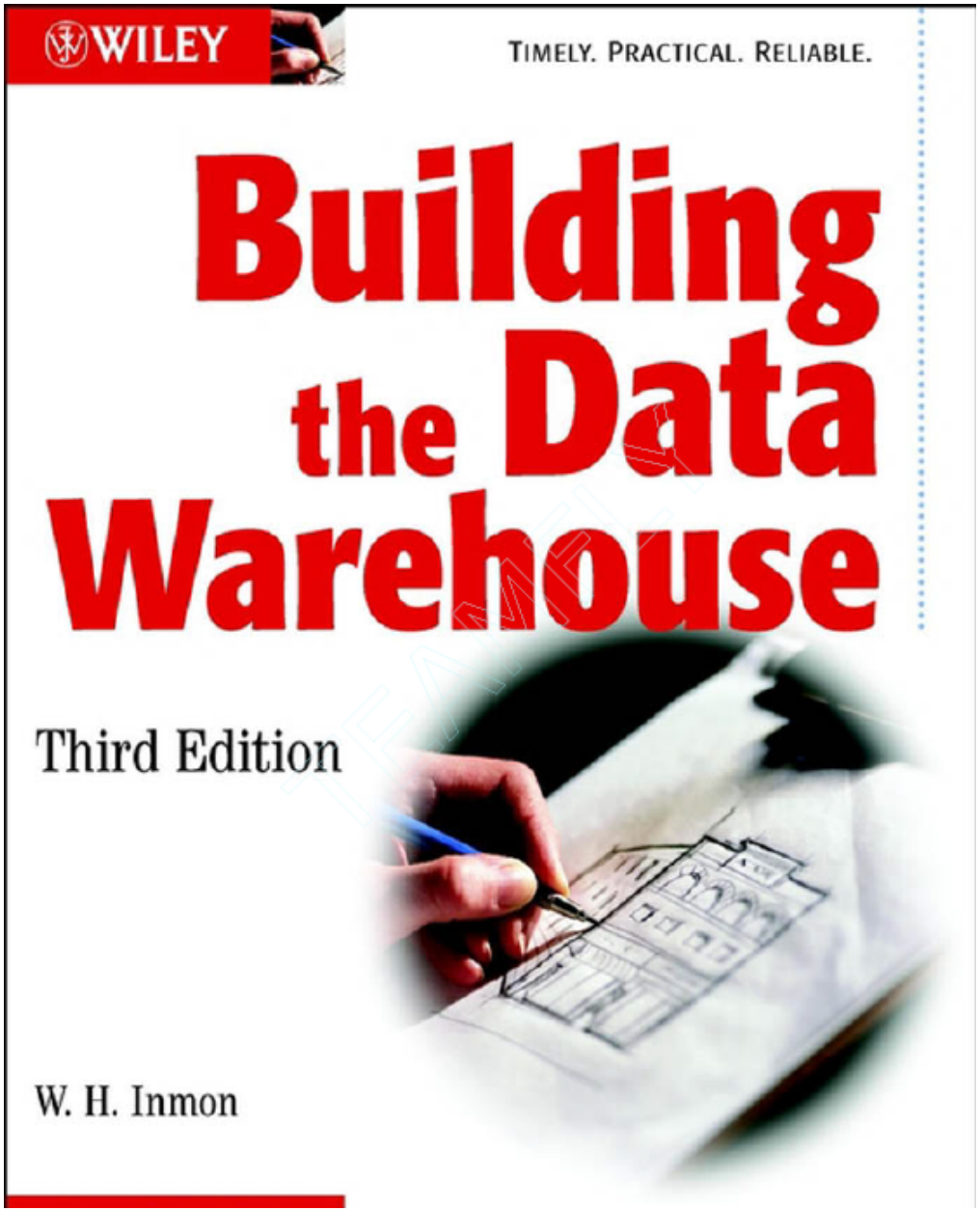
Technical Reference

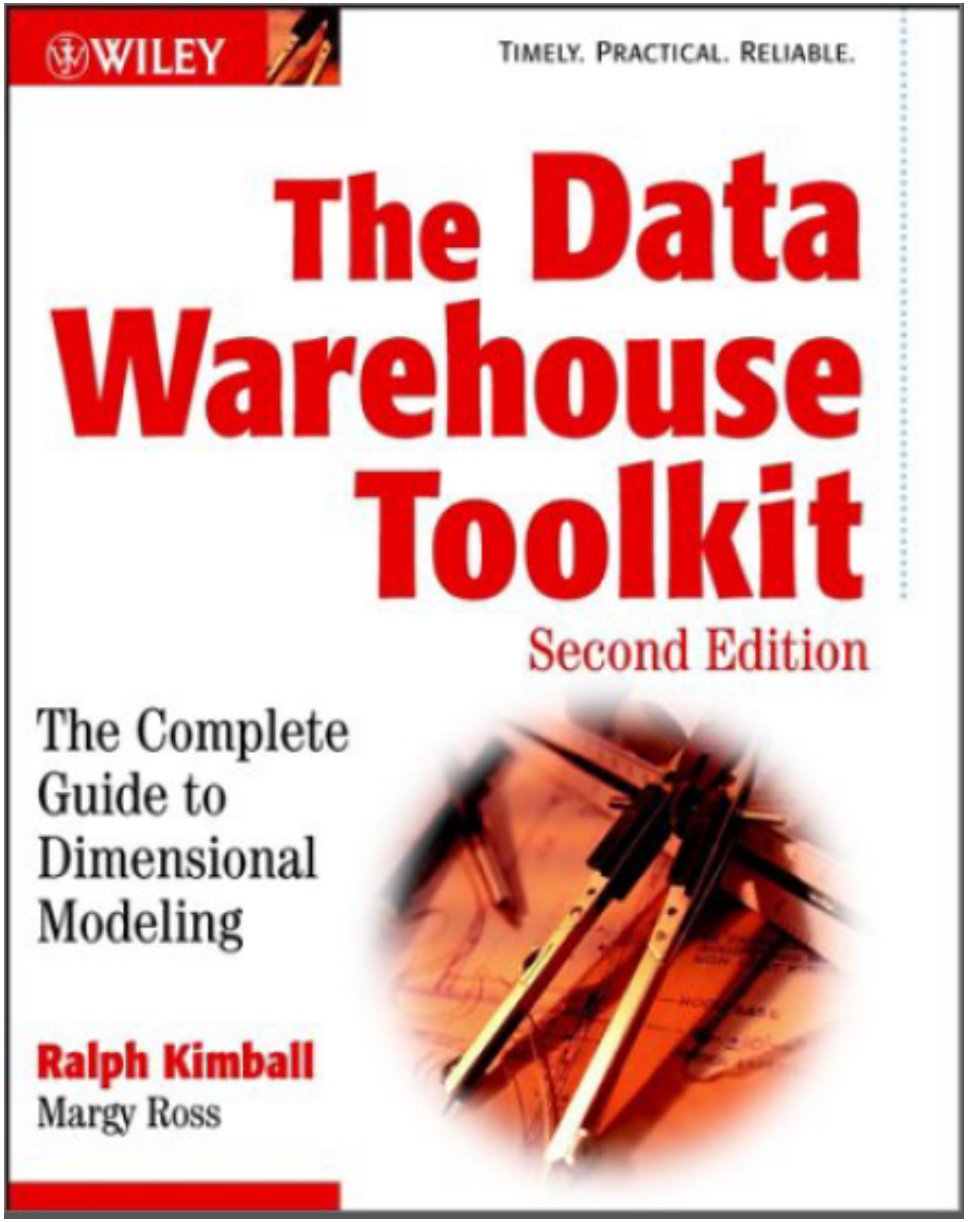
The authoritative guide to uncovering hidden information and meaningful patterns in large databases

Claude Seidman

IT Professional and Developer

Reference Books




WILEY
TIMELY. PRACTICAL. RELIABLE.
The Data Warehouse Toolkit
Second Edition
The Complete Guide to Dimensional Modeling
Ralph Kimball
Margy Ross

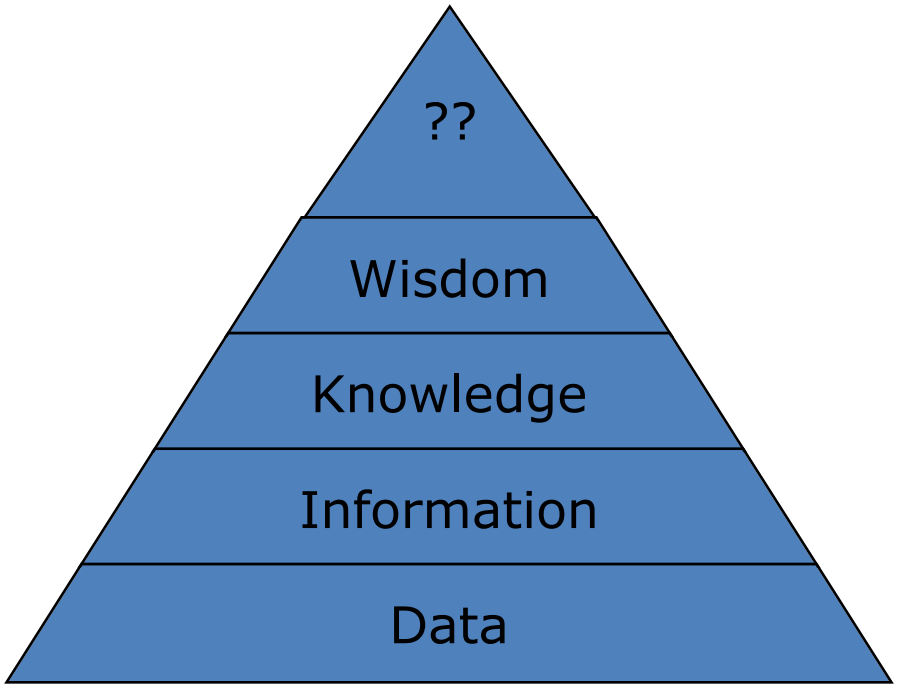
Unit 1 :
Introduction to Data Warehousing
(5 Hrs.)

Contents:

- Lifecycle of data,
- Types of data,
- Data warehouse and data warehousing ,
- Differences between operational database and data warehouse,
- A multidimensional data model,
- OLAP operation in multidimensional data model,
- Conceptual modeling of data warehouse,
- Architecture of data warehouse,
- Data warehouse implementation,
- Data marts,
- Components of data warehouse,
- Need for data warehousing,
- Trends in data warehousing

What is Data?

- A representation of **facts, concepts, or instructions** in a formal manner suitable for communication, interpretation, or processing by human beings or by computers.



Lifecycle of Data

- The data lifecycle represents all of the stages of data throughout its life from its creation for a study to its distribution and reuse.
- The data lifecycle begins with a researcher(s) developing a concept for a study; once a study concept is developed, data is then collected for that study.
- After data is collected, it is processed for distribution so that it can be archived and used by other researchers at a later date.
- Once data reaches the distribution stage of the lifecycle, it is stored in a location (i.e. repository, registry) where it can then be discovered by other researchers.
- Data discovery leads to the repurposing of data, which creates a continual loop back to the data processing stage where the repurposed data is archived and distributed for discovery.

DATA LIFE CYCLE STAGES

- The data life cycle is often described as a cycle because the lessons learned and insights gleaned from one data project typically inform the next. In this way, the final step of the process feeds back into the first.

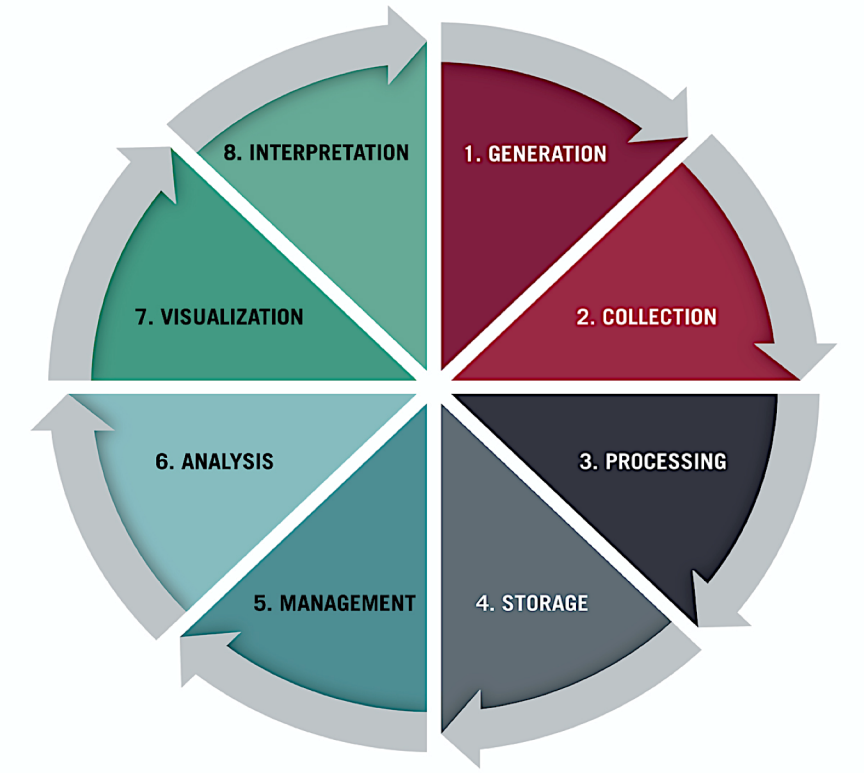


Fig: https://online.hbs.edu/

DATA LIFE CYCLE STAGES

1. Generation

- For the data life cycle to begin, data must first be generated. Otherwise, the following steps can't be initiated.
- Data generation occurs regardless of whether you're aware of it, especially in our increasingly online world.
- Some of this data is generated by your organization, some by your customers, and some by third parties you may or may not be aware of.
- Every sale, purchase, hire, communication, interaction — everything generates data. Given the proper attention, this data can often lead to powerful insights that allow you to better serve your customers and become more effective in your role.

DATA LIFE CYCLE STAGES

2. Collection

- Not all of the data that's generated every day is collected or used. It's up to your data team to identify what information should be captured and the best means for doing so, and what data is unnecessary or irrelevant to the project at hand.

We can collect data in a variety of ways, including:

- **Forms:** Web forms, client or customer intake forms, vendor forms, and human resources applications are some of the most common ways businesses generate data.
- **Surveys:** Surveys can be an effective way to gather vast amounts of information from a large number of respondents.
- **Interviews:** Interviews and focus groups conducted with customers, users, or job applicants offer opportunities to gather qualitative and subjective data that may be difficult to capture through other means.
- **Direct Observation:** Observing how a customer interacts with your website, application, or product can be an effective way to gather data that may not be offered through the methods above.

DATA LIFE CYCLE STAGES

3. Processing

- Once data has been collected, it must be processed. Data processing can refer to various activities, including:
- **Data wrangling**, in which a data set is cleaned and transformed from its raw form into something more accessible and usable. This is also known as data cleaning, data munging, or data remediation.
- **Data compression,** in which data is transformed into a format that can be more efficiently stored.
- **Data encryption,** in which data is translated into another form of code to protect it from privacy concerns.

DATA LIFE CYCLE STAGES

4. Storage

- After data has been collected and processed, it must be stored for future use.
- This is most commonly achieved through the creation of databases or datasets. These datasets may then be stored in the cloud, on servers, or using another form of physical storage like a hard drive, CD, cassette, or floppy disk.
- When determining how to best store data for your organization, it's important to build in a certain level of redundancy to ensure that a copy of your data will be protected and accessible, even if the original source becomes corrupted or compromised.

DATA LIFE CYCLE STAGES

5. Management

- Data management, also called database management, involves organizing, storing, and retrieving data as necessary over the life of a data project.
- While referred to here as a “step,” it’s an ongoing process that takes place from the beginning through the end of a project.
- Data management includes everything from storage and encryption to implementing access logs and changelogs that track who has accessed data and what changes they may have made.

DATA LIFE CYCLE STAGES

6. Analysis

- Data analysis refers to processes that attempt to glean meaningful insights from raw data. Analysts and data scientists use different tools and strategies to conduct these analyses.
- Some of the more commonly used methods include statistical modeling, algorithms, artificial intelligence, data mining, and machine learning.
- Exactly who performs an analysis depends on the specific challenge being addressed, as well as the size of your organization's data team.
- Business analysts, data analysts, and data scientists can all play a role.

DATA LIFE CYCLE STAGES

7. Visualization

- Data visualization refers to the process of creating graphical representations of your information, typically through the use of one or more visualization tools.
- Visualizing data makes it easier to quickly communicate your analysis to a wider audience both inside and outside your organization.
- The form your visualization takes depends on the data you're working with, as well as the story you want to communicate.
- While technically not a required step for all data projects, data visualization has become an increasingly important part of the data life cycle.

DATA LIFE CYCLE STAGES

8. Interpretation

- Finally, the interpretation phase of the data life cycle provides the opportunity to make sense of your analysis and visualization.
- Beyond simply presenting the data, this is when you investigate it through the lens of your expertise and understanding.
- Your interpretation may not only include a description or explanation of what the data shows but, more importantly, what the implications may be.

Types of Data

1. Flat Files

- Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
- Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.
- Flat files are represented by data dictionary. Eg: CSV file.
- **Application:** Used in DataWarehousing to store data, Used in carrying data to and from server, etc.

Types of Data

2. Relational Databases

- A Relational database is defined as the collection of data organized in tables with rows and columns.
- Physical schema in Relational databases is a schema which defines the structure of tables.
- Logical schema in Relational databases is a schema which defines the relationship among tables.
- Standard API of relational database is SQL.
- **Application:** Data Mining, ROLAP model, etc.

Types of Data

3. DataWarehouse

- A datawarehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
- **There are three types of datawarehouse:** Enterprise datawarehouse, DataMart and Virtual Warehouse
- **Two approaches can be used to update data in DataWarehouse:** Query-driven Approach and Update-driven Approach.
- **Application:** Business decision making, Data mining, etc.

Types of Data

4. Transactional Databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
- This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
- Highly flexible system where users can modify information without changing any sensitive information.
- Follows ACID property of DBMS.
- **Application:** Banking, Distributed systems, Object databases, etc.

Types of Data

5. Multimedia Databases

- Multimedia databases consists audio, video, images and text media.
- They can be stored on Object-Oriented Databases.
- They are used to store complex information in a pre-specified formats.
- **Application:** Digital libraries, video-on demand, news-on demand, musical database, etc.

Types of Data

6. Spatial Database

- Store geographical information.
- Stores data in the form of coordinates, topology, lines, polygons, etc.
- **Application:** Maps, Global positioning, etc.

Types of Data

7. Time-series Databases

- Time series databases contains stock exchange data and user logged activities.
- Handles array of numbers indexed by time, date, etc.
- It requires real-time analysis.
- **Application:** eXtremeDB, Graphite, InfluxDB, etc.

Types of Data

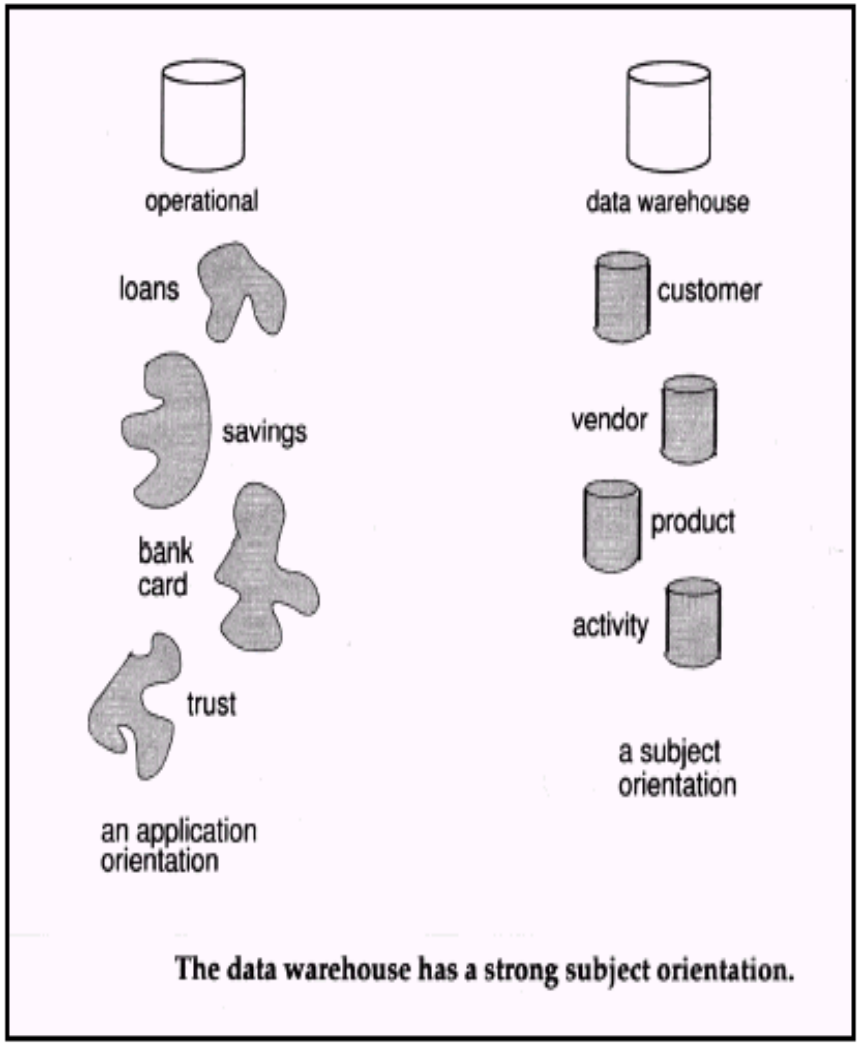
8. WWW

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
- It is the most heterogeneous repository as it collects data from multiple resources.
- It is dynamic in nature as Volume of data is continuously increasing and changing.
- **Application:** Online shopping, Job search, Research, studying, etc.

Data Warehouse

- According to W. H. Inmon, a **data warehouse** is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
- "A data warehouse is a copy of transaction data specifically structured for querying and reporting" – Ralph Kimball
- It is the process of building a data warehouse for an organization.
- It is a process of transforming data into information and making it available to users in a timely enough manner to make a difference

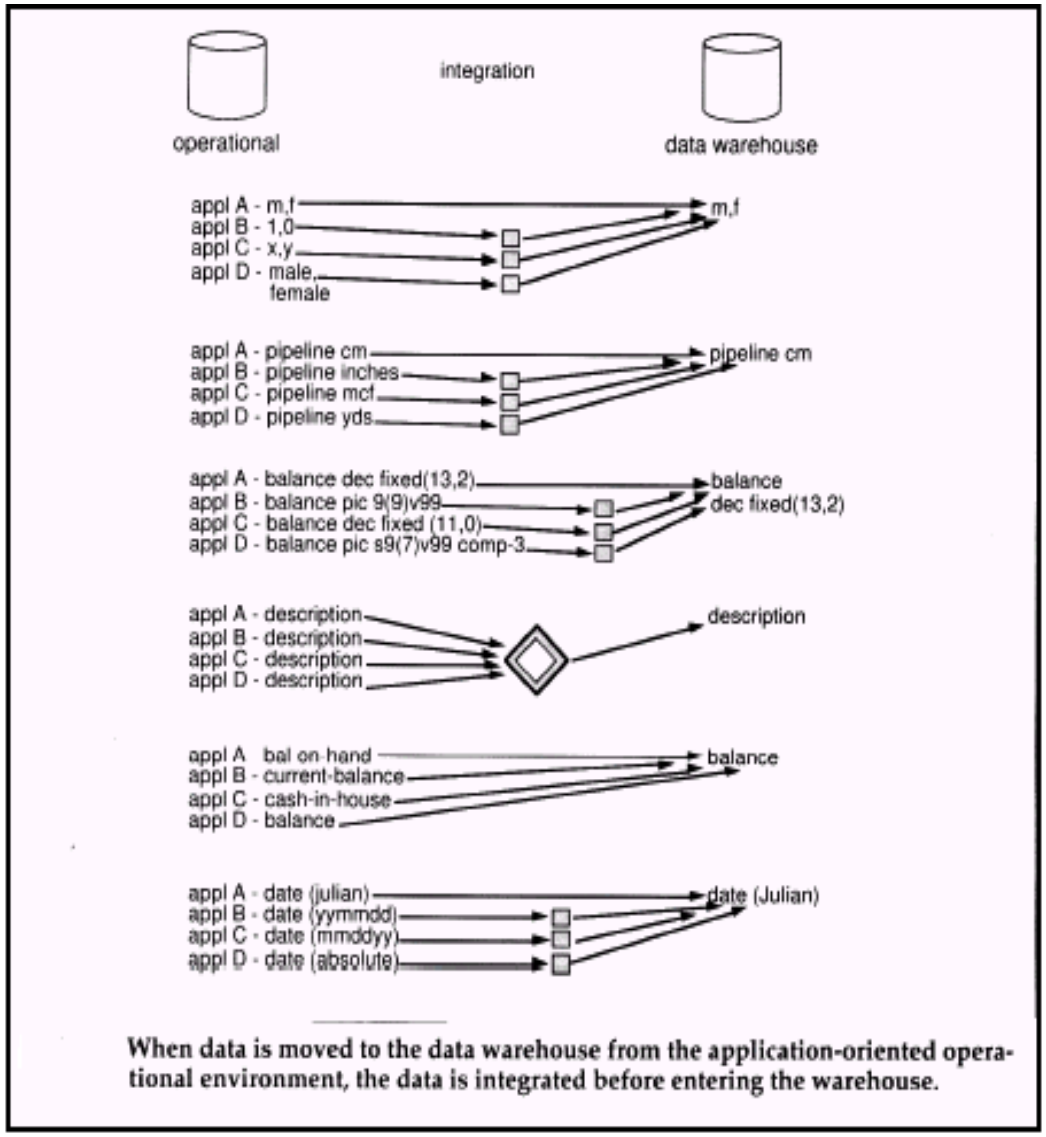
Subject Oriented



The data warehouse has a strong subject orientation.

- Focus is on Subject Areas rather than Applications
- Organized around major subjects, such as customer, product, sales.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated



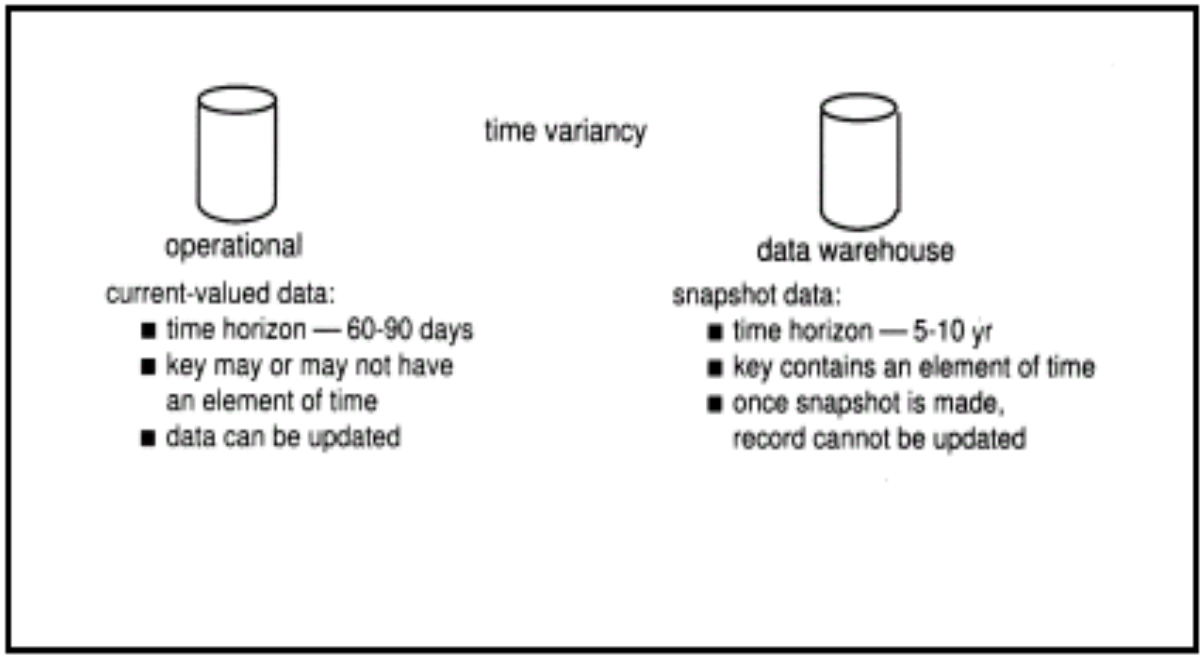
When data is moved to the data warehouse from the application-oriented operational environment, the data is integrated before entering the warehouse.

- Constructed by integrating multiple, heterogeneous data sources: relational databases, flat files, on-line transaction records
- Integration tasks handles naming conventions, physical attributes of data
- Must be made consistent.

Integrated

- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
- When data is moved to the warehouse, it is converted.

Time Variant

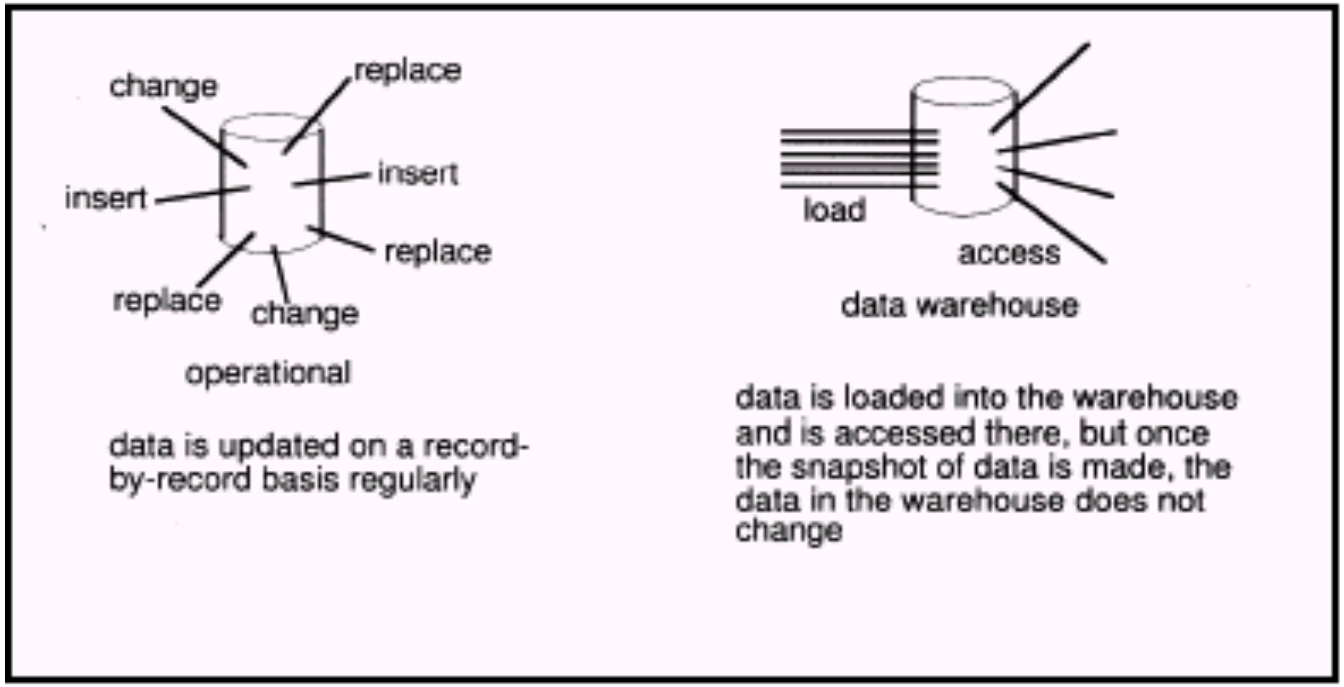


- Only accurate and valid at some point in time or over some time interval.
- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database provides current value data.
  - Data warehouse data provide information from a historical perspective (e.g., past 5-10 years)

Time Variant

- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Non Volatile



- Data Warehouse is relatively Static in nature.
- Operational update of data does not occur in the data warehouse environment
  - 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*

Data Warehousing

- A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights.
- It is typically used to connect and analyze business data from heterogeneous sources.
- It is the core of the BI system which is built for data analysis and reporting.
- It is a blend of technologies and components which aids the strategic use of data.
- It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing.
- It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

Differences between operational database and data warehouse,

Operational Database	Data Warehouse
Operational systems are designed to support high-volume transaction processing(OLTP).	Data warehousing systems are typically designed to support high-volume analytical processing (OLAP).
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
Data within operational systems are mainly updated regularly according to need.	Non-volatile, new data may be added regularly. Once Added rarely changed.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	It is optimized for extent loads and high, complex, unpredictable queries that access many rows per table.

Differences between operational database and data warehouse,

It is optimized for validation of incoming information during transactions, uses validation data tables.	Loaded with consistent, valid information, requires no real-time validation.
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented
Operational systems are usually optimized to perform fast inserts and updates of associatively small volumes of data.	Data warehousing systems are usually optimized to perform fast retrievals of relatively high volumes of data.
Data In	Data Out
Less Number of data accessed.	Large Number of data accessed.
Relational databases are created for on-line transactional Processing (OLTP)	Data Warehouse designed for on-line Analytical Processing (OLAP)

A multidimensional data model

- Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a *data cube*.
- *"What is a data cube?"* A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records.
- Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
- A multidimensional data model is typically organized around a central theme, like *sales*, for instance. This theme is represented by a fact table. Facts are numerical measures.

	location = "Vancouver"			
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Table: A 2-D view of sales data according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver. The measure displayed is *dollars sold* (in thousands).

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

Table: A 3-D view of sales data according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars sold* (in thousands).

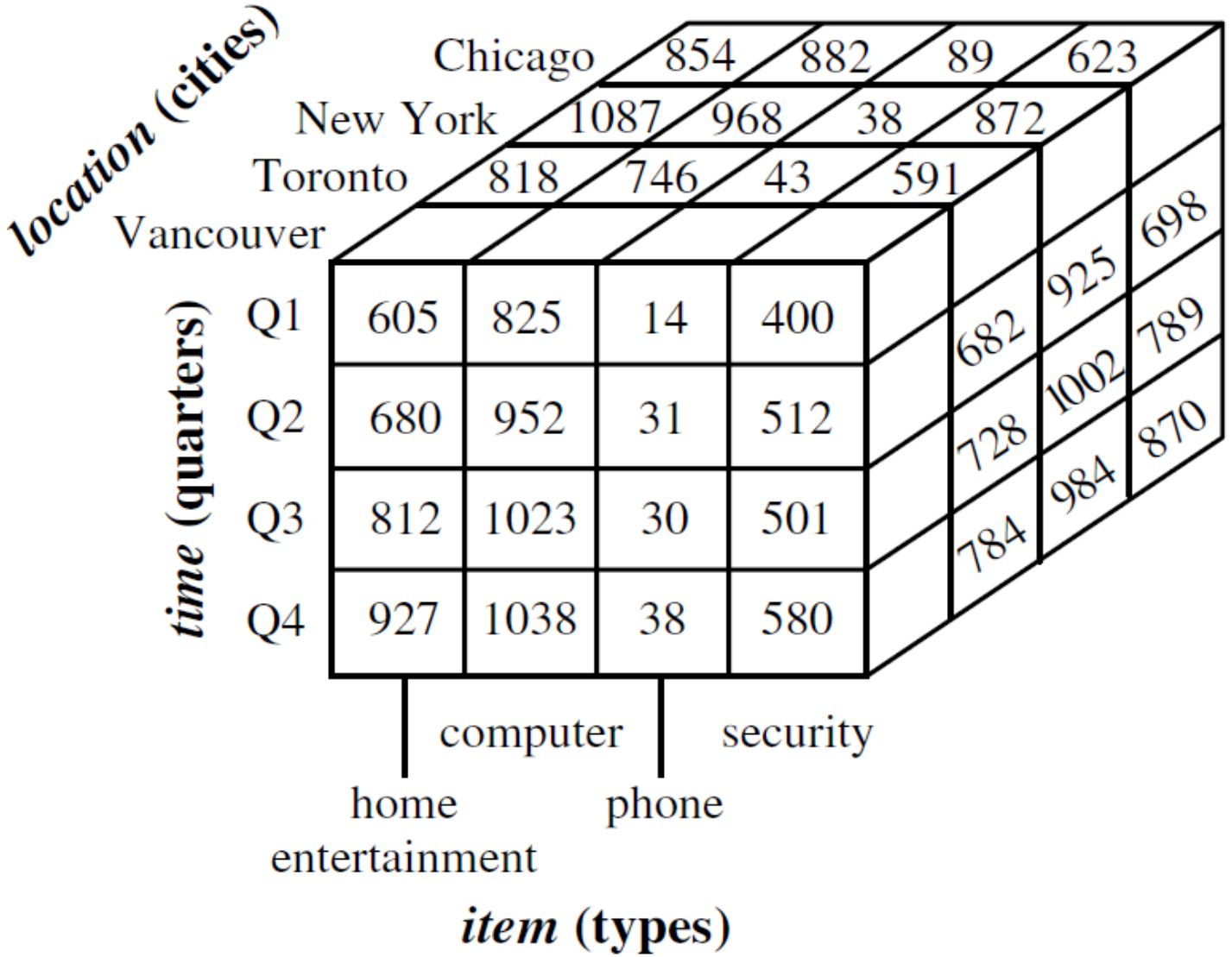


Figure: A 3-D data cube representation of the data in the table above, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars sold* (in thousands).

Example:

Suppose that we would now like to view our sales data with an additional fourth dimension, such as *supplier*.

Solution!!

Viewing things in 4-D becomes tricky. However, we can think of a 4-D cube as being a series of 3-D cubes as shown below:

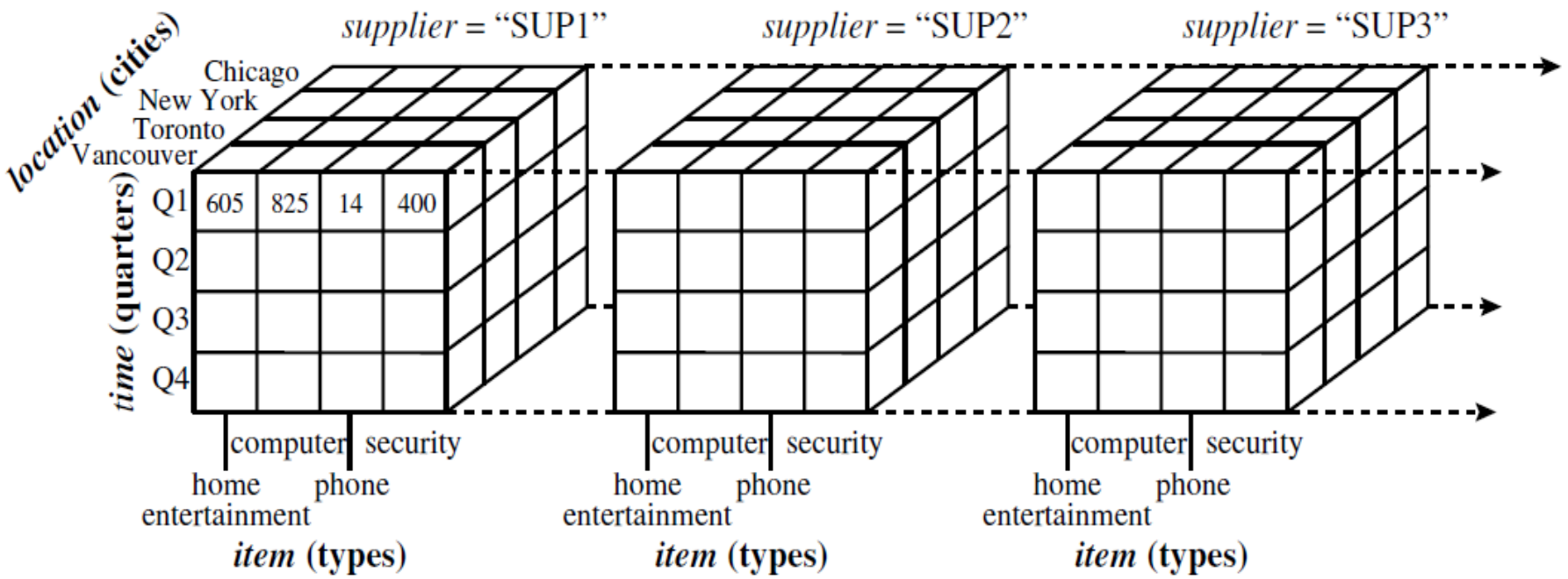


Figure: A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier*.

The measure displayed is *dollars sold* (in thousands). For improved readability, only some of the cube values are shown.

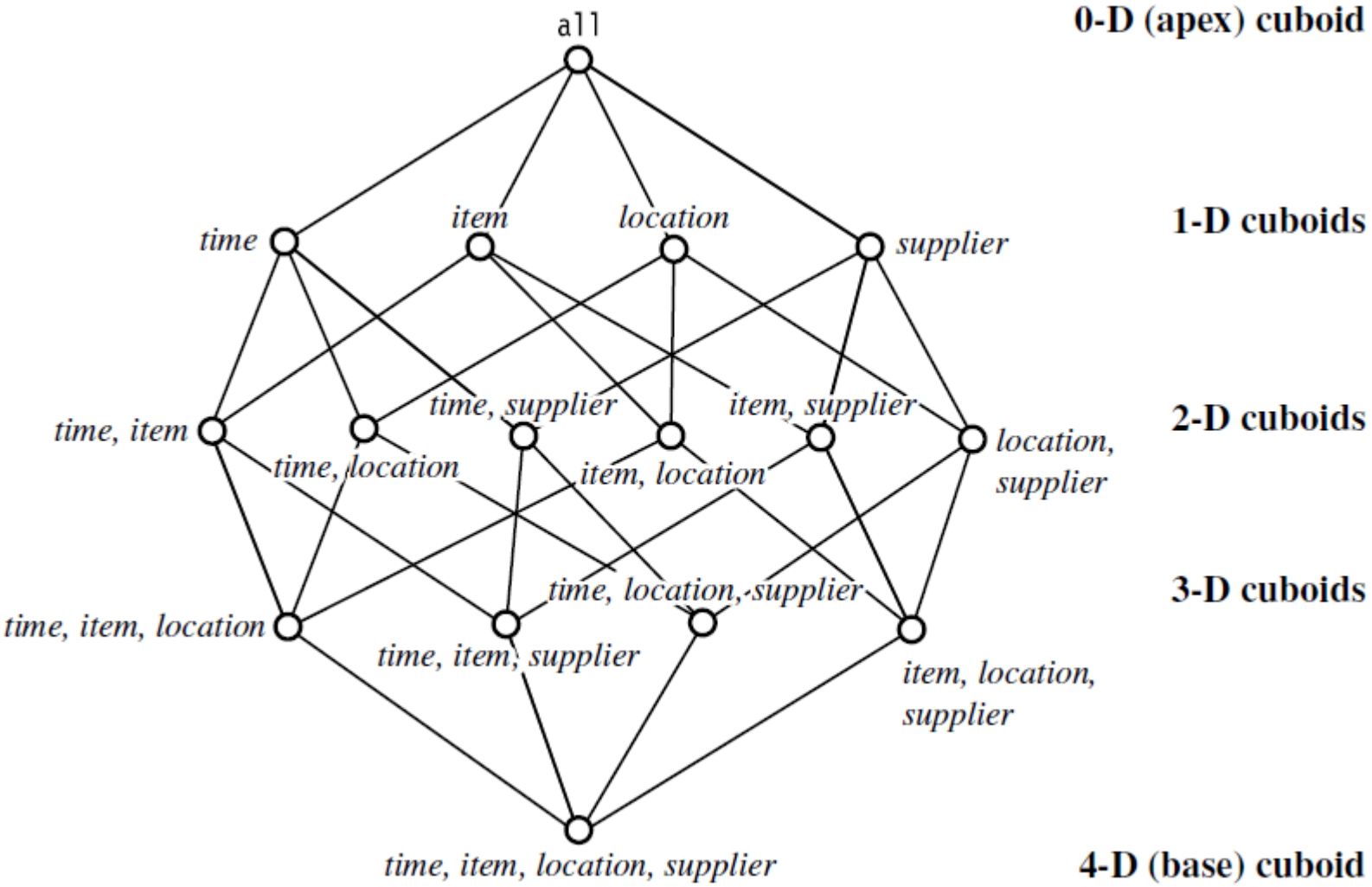


Figure: Lattice of cuboids, making up a 4-D data cube for the dimensions *time*, *item*, *location*, and *supplier*. Each cuboid represents a different degree of summarization.

Online Analytical Processing (OLAP)

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information.

❑ OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data. Here is the list of OLAP operations:

- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

Roll-up/ Drill up:

- Roll-up performs aggregation on a data cube in any of the following ways:
- By climbing up a concept hierarchy for a dimension or by dimension reduction.
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- The concept hierarchy was "street < city < province < country". On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country. The data is grouped into cities rather than countries. When roll-up is performed, one or more dimensions from the data cube are removed.

Roll-up/ Drill up:

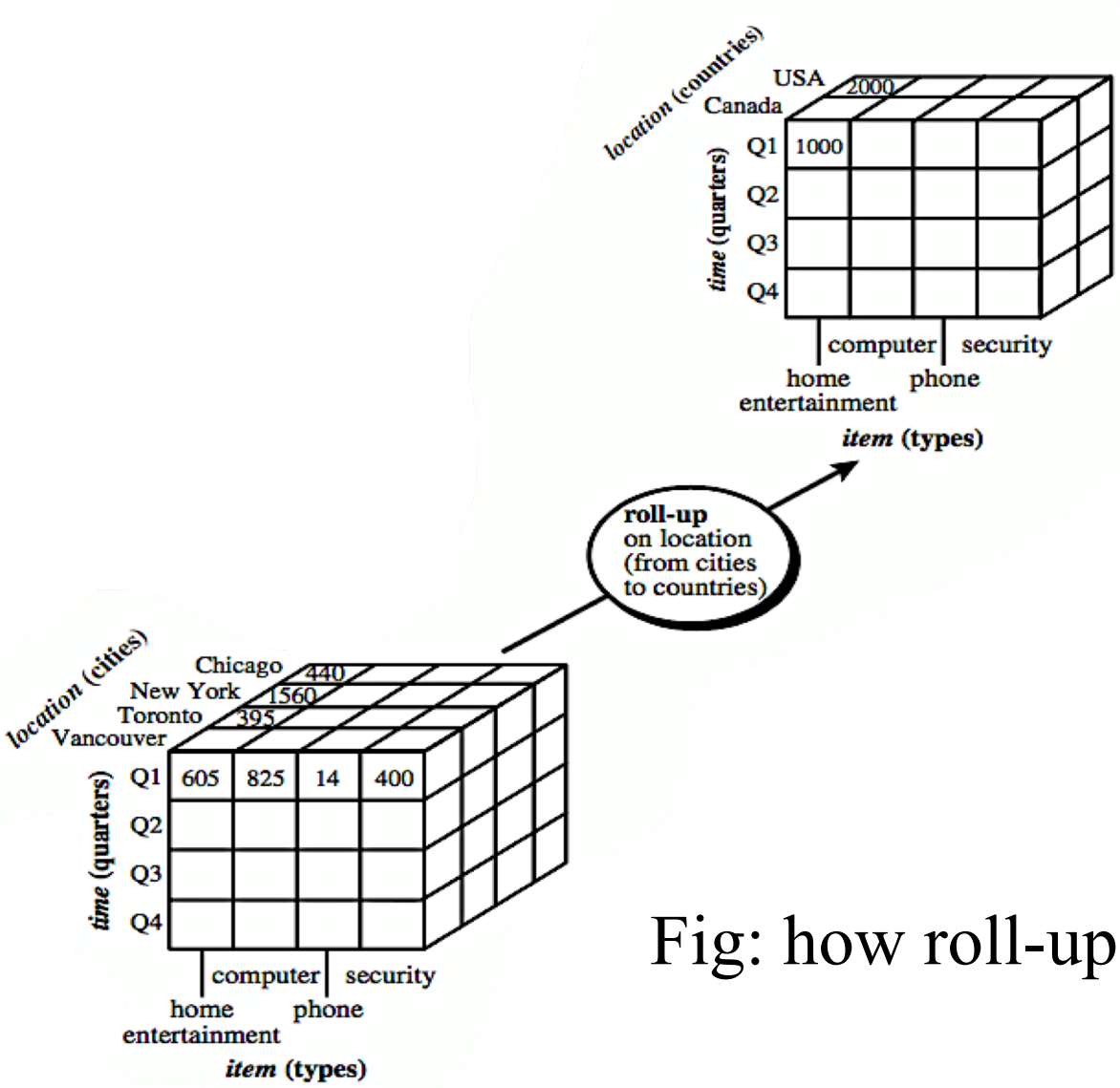
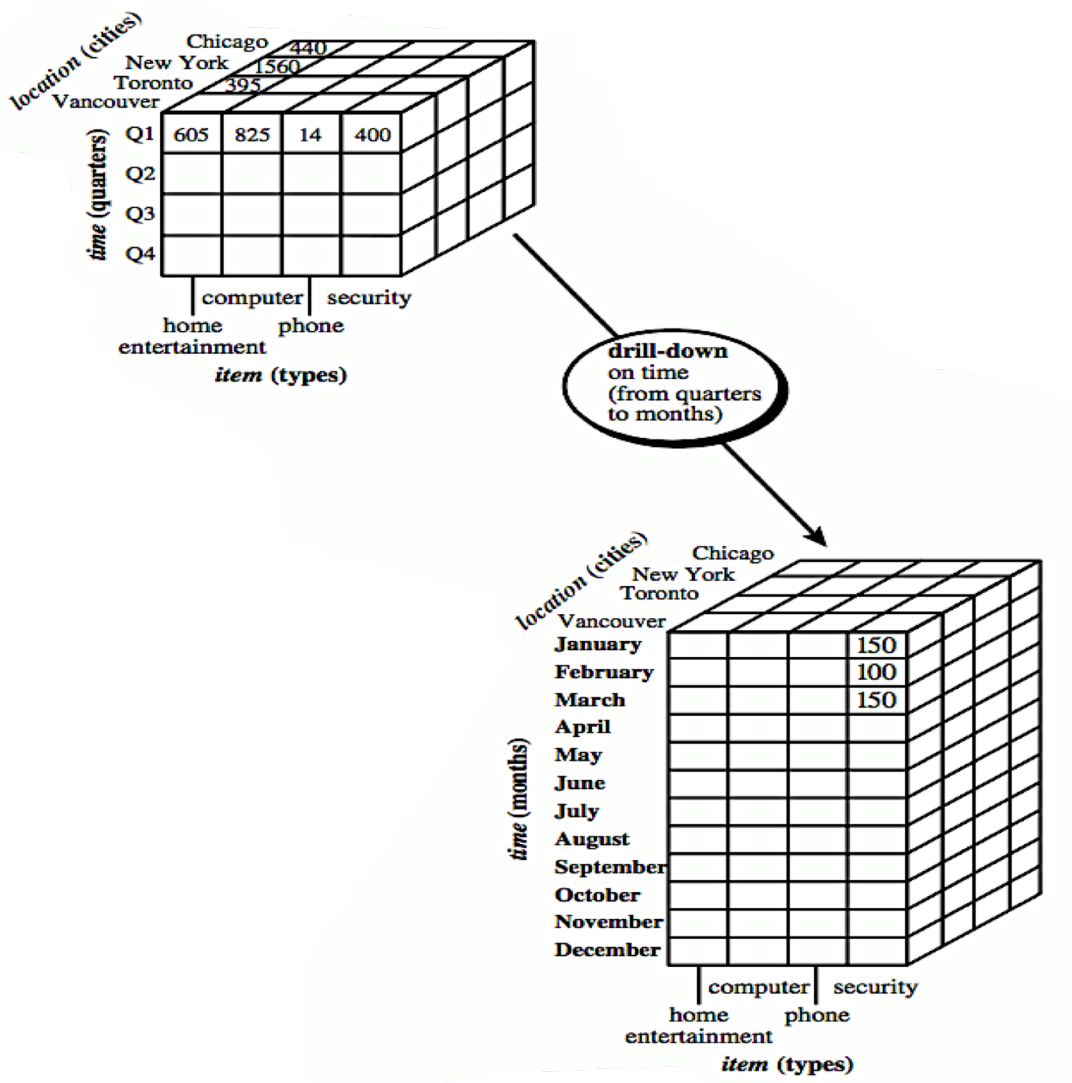


Fig: how roll-up works?

Roll-down/ Drill down:

- Drill-down is the reverse operation of roll-up.
- It is performed by either of the following ways: By stepping down a concept hierarchy for a dimension or by introducing a new dimension.
- Drill-down is performed by stepping down a concept hierarchy for the dimension time. Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month. When drill-down is performed, one or more dimensions from the data cube are added. It navigates the data from less detailed data to highly detailed data.
- The following diagram illustrates how drill-down works:

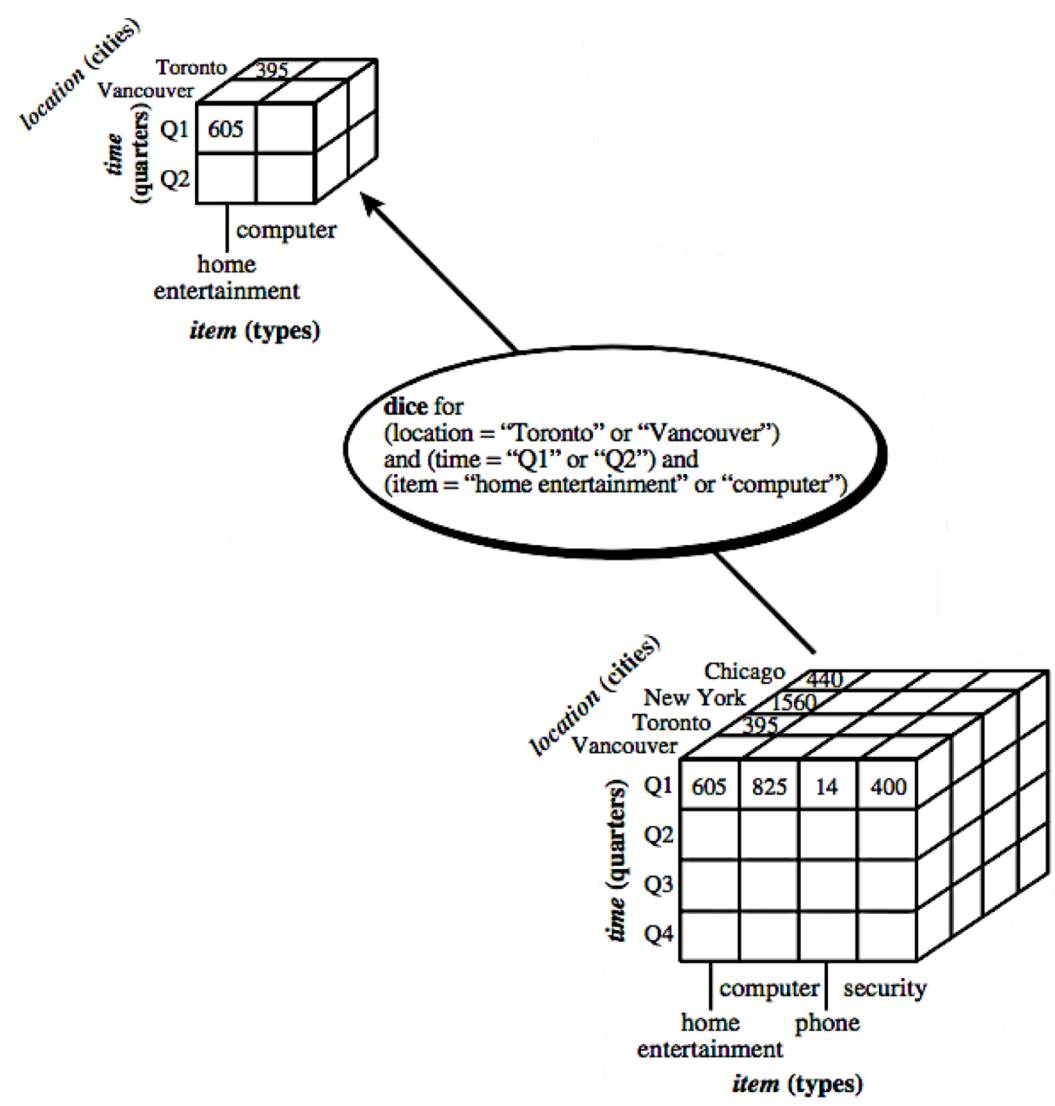
Drill Down/Roll Down



Dice:

- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

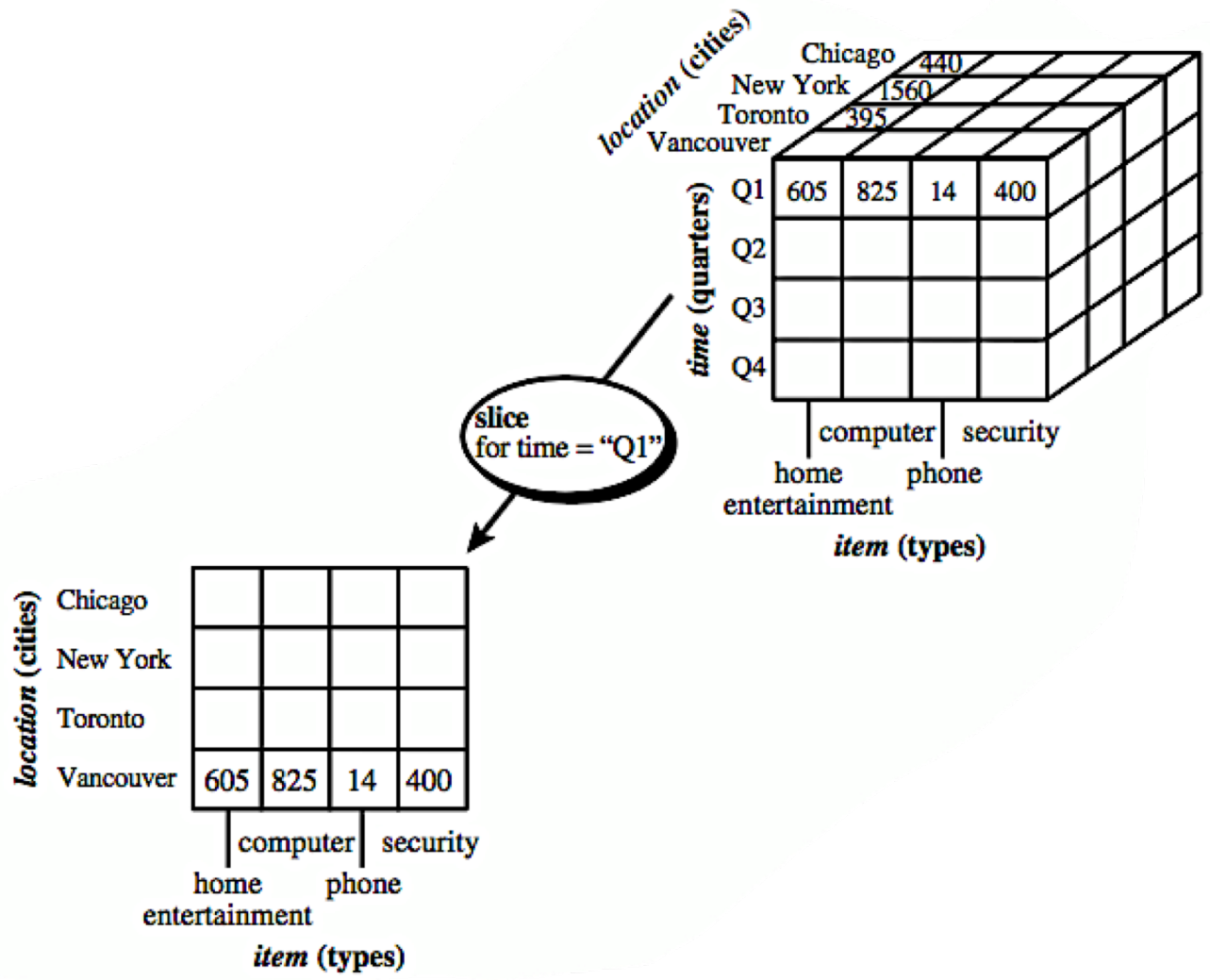
Dice


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
605
Q2
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
time (quarters)
Q1
605
825
14
400
Q2
Q3
Q4
computer
security
home entertainment
phone
item (types)

Slice:

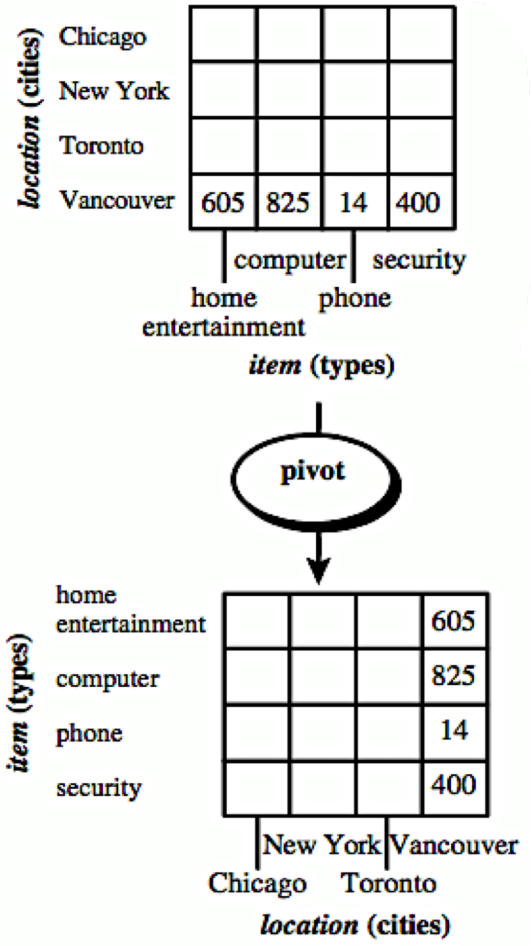
- The slice operation selects one particular dimension from a given cube and provides a new sub-cube.
- Here Slice is performed for the dimension "time" using the criterion time = "Q1". It will form a new sub-cube by selecting one or more dimensions.
- Consider the following diagram that shows how slice works.

Slice


location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
slice for time = "Q1"
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)

Pivot:

- The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.
- Consider the following diagram that shows the pivot operation.



Conceptual Data Model

A conceptual data model include identification of important entities and the relationships among them. At this level, the objective is to identify the relationships among the different entities.

Logical Design

Logical design is the phase of a database design concerned with identifying the relationships among the data elements.

A logical design is conceptual and abstract. You do not deal with the physical implementation details yet. You deal only with defining the types of information that you need.

Logical design deals with concepts related to a certain kind of DBMS (e.g. relational, object oriented,) but are understandable by end users

The logical design should result in

(1) A set of entities and attributes corresponding to fact tables and dimension tables.

(2) A model of operational data from your source into subject-oriented information in your target data warehouse schema.

You can create the logical design using a pen and paper, or you can use a design tool such as Oracle Warehouse Builder (specifically designed to support modeling the ETL process) or Oracle Designer (a general purpose modeling tool).

The steps of the logical data model include identification of all entities and relationships among them. All attributes for each entity are identified and then the primary key and foreign key is identified. Normally normalization occurs at this level.

In data warehousing, it is common to combine the conceptual data model and the logical data model to a single step. **The steps for logical data model are indicated below:**

1. Identify all entities.
2. Identify primary keys for all entities.
3. Find the relationships between different entities.
4. Find all attributes for each entity.
5. Resolve all entity relationships that is many-to-many relationships.
6. Normalization if required.

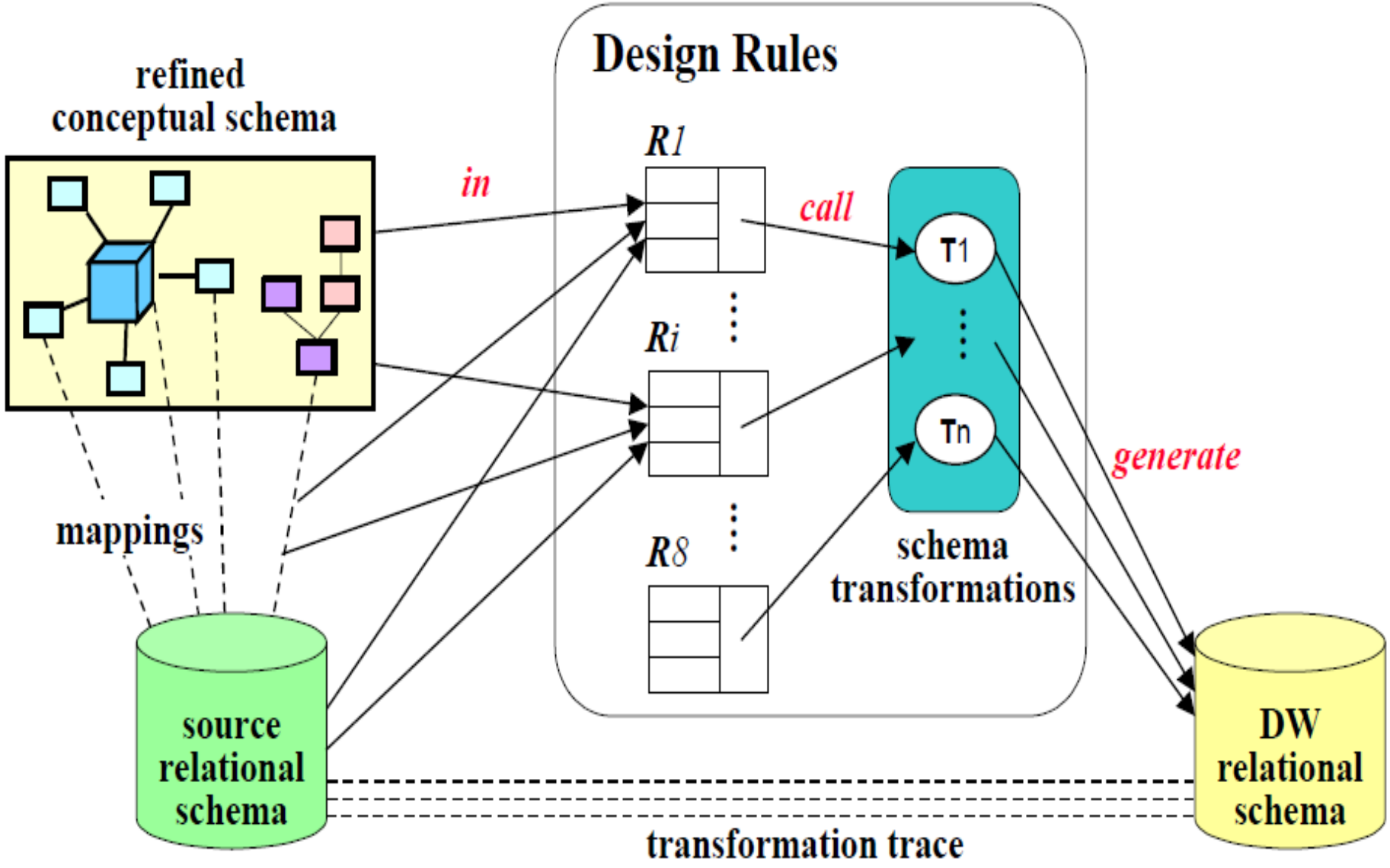
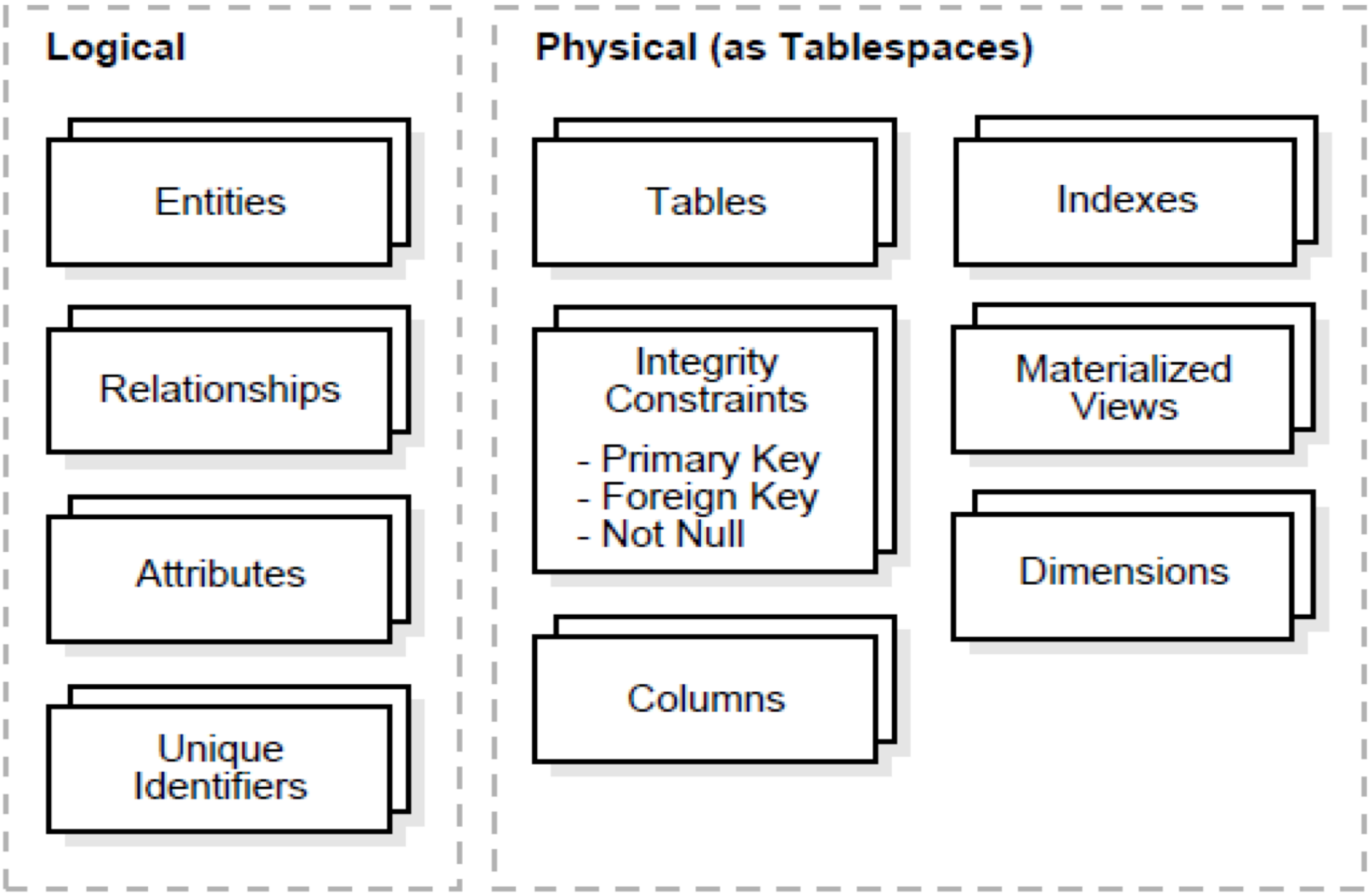


Figure: Data warehouse logical design environment.

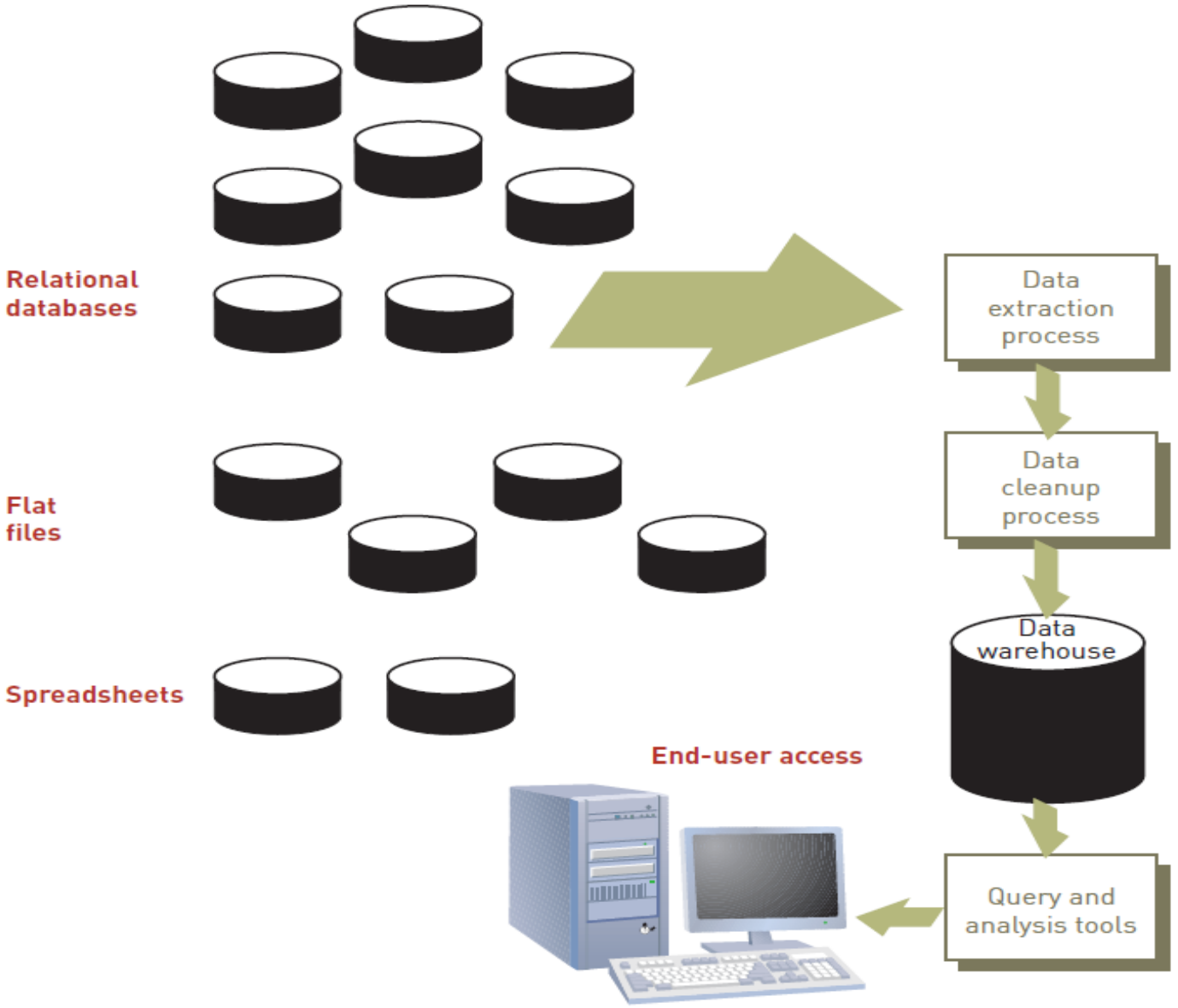
The environment provides the infrastructure to carry out the specified process. It consists of:

- A *refined conceptual schema*, which is built from a conceptual multidimensional schema enriched with design guidelines.
- The *source schema* and the *DW schema*.
- Schema *mappings*, which are used to represent correspondences between the conceptual schema and the source schema.
- A set of *design rules*, which apply the *schema transformations* to the source schema in order to build the DW schema.
- A set of pre-defined *schema transformations* that build new relations from existing ones, applying DW design techniques.
- A *transformation trace*, which keeps the transformations that where applied, providing the mappings between source and DW schemas.

Logical Design compared with Physical Design



From Tables and Spreadsheets to Data Cubes



The process of logical design involves arranging data into a series of logical relationships called entities and attributes.

An **entity** represents a chunk of information. In relational databases, an entity often maps to a table.

An **attribute** is a component of an entity that helps define the uniqueness of the entity. In relational databases, an attribute maps to a column.

Relational database model's structural and data independence enables us to view data logically rather than physically.

The logical view allows a simpler file concept of data storage.

The use of logically independent tables is easier to understand.

Logical simplicity yields simpler and more effective database design methodologies.

An entity is a person, place, event, or thing for which we intend to collect data.

- **University** -- Students, Faculty Members, Courses
- **Airlines** -- Pilots, Aircraft, Routes, Suppliers

Each entity has certain characteristics known as attributes.

- **Student** -- Student Number, Name, GPA, Date of Enrollment, Data of Birth, Home Address, Phone Number, Major
- **Aircraft** -- Aircraft Number, Date of Last Maintenance, Total Hours Flown, Hours Flown since Last Maintenance

A grouping of related entities becomes an entity set.

- The STUDENT entity set contains all student entities.
- The FACULTY entity set contains all faculty entities.
- The AIRCRAFT entity set contains all aircraft entities

- A table contains a group of related entities -- i.e. an entity set.
- The terms entity set and table are often used interchangeably.
- A table is also called a relation.
- While entity-relationship diagramming has traditionally been associated with highly normalized models such as OLTP applications, the technique is still useful for data warehouse design in the form of dimensional modeling.

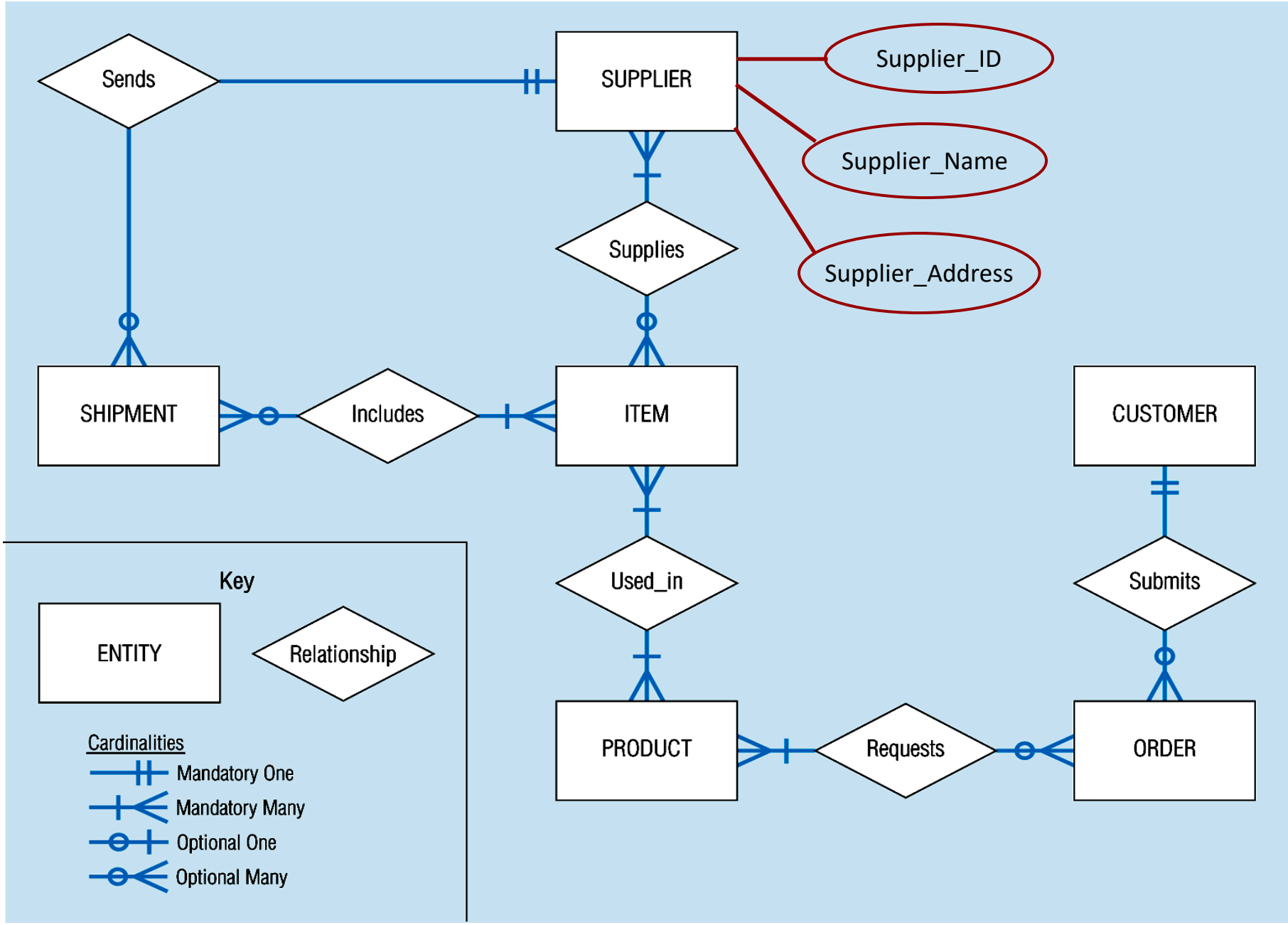


Figure: Sample E-R Diagram

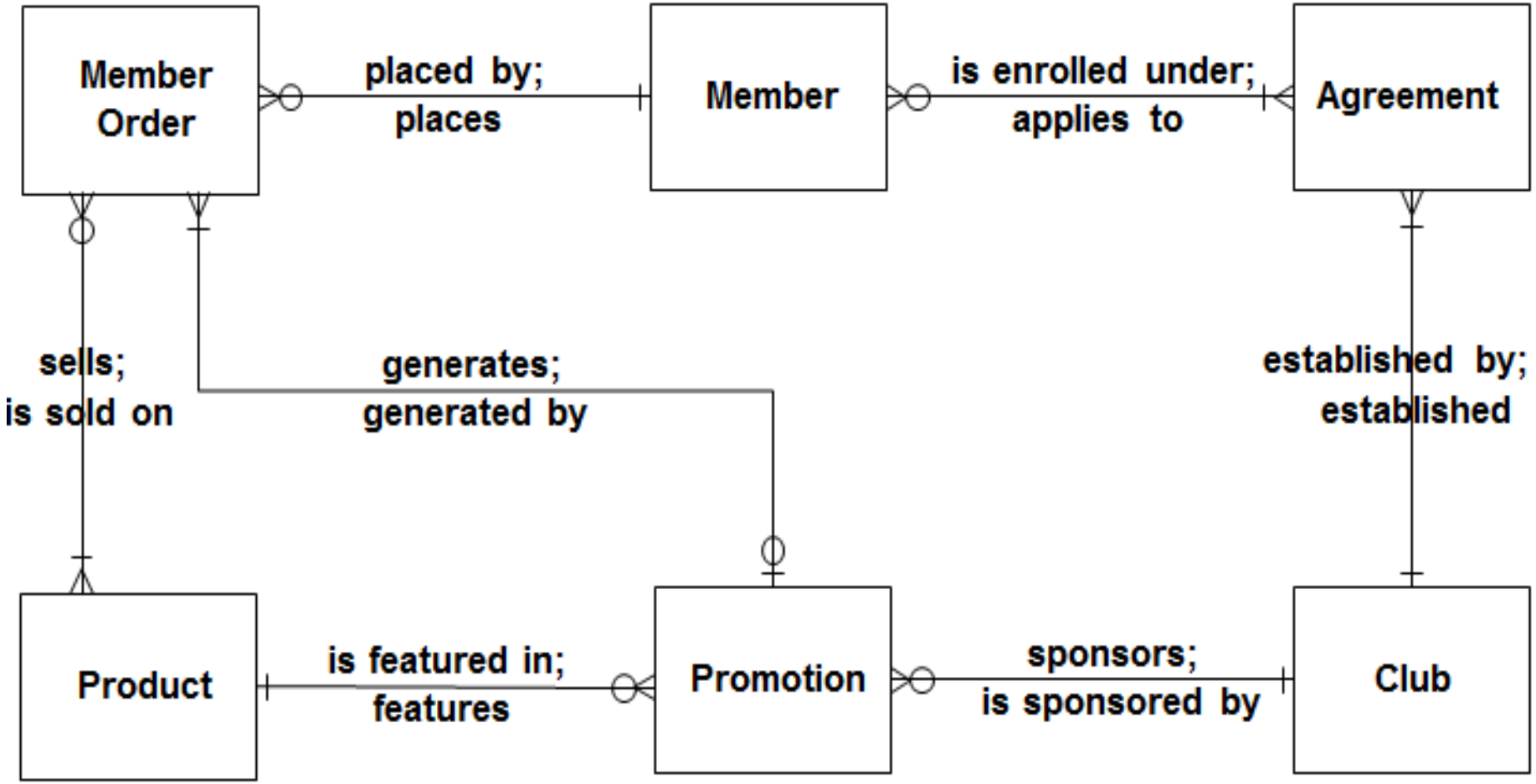
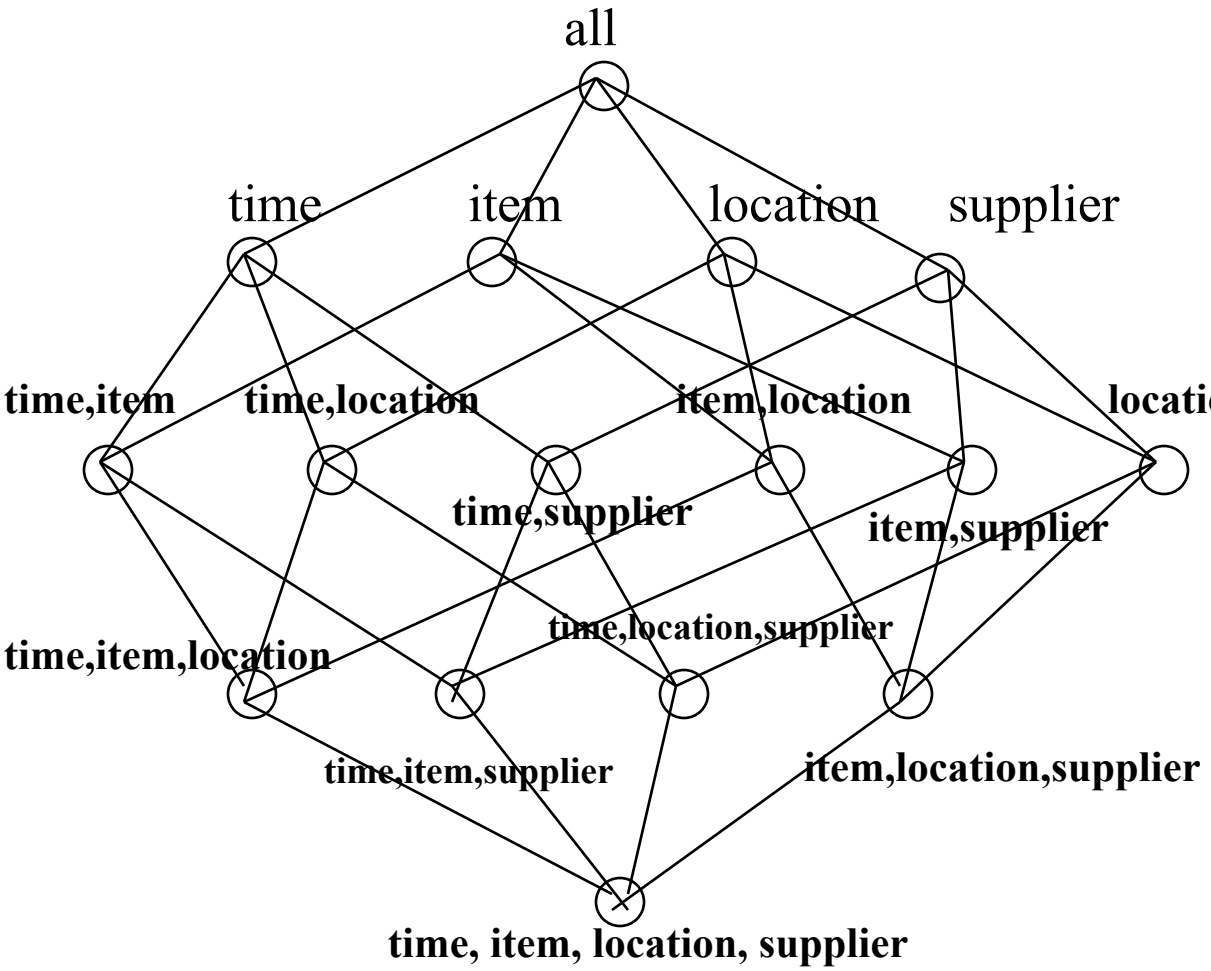


Figure: Sample E-R Diagram

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

Cube: A Lattice of Cuboids



0-D(apex) cuboid

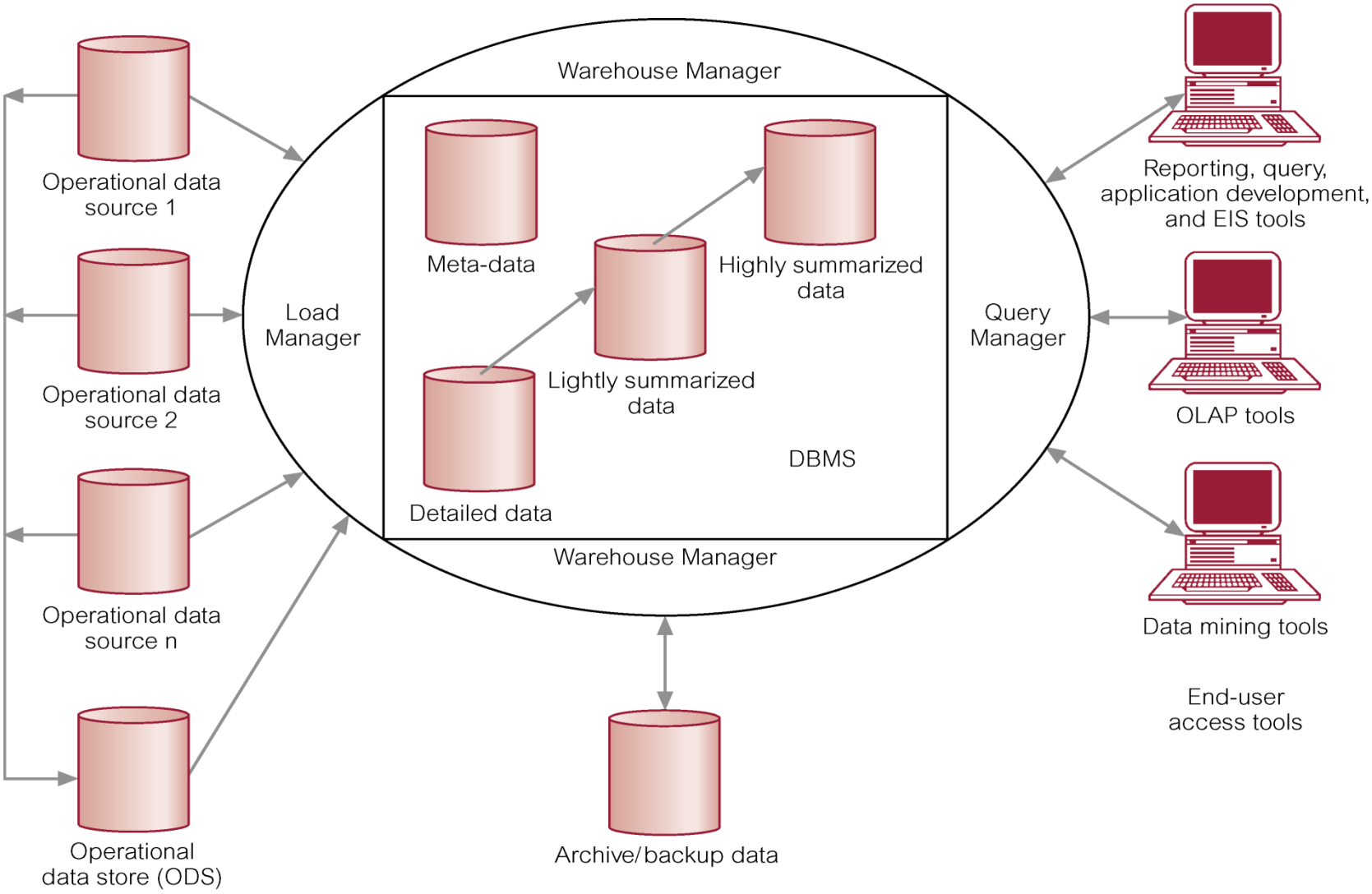
1-D cuboids

2-D cuboids

3-D cuboids

4-D(base) cuboid

Data Warehouse Architecture:



1. Operational Data Sources: It may include:

- Network databases.
- Departmental file systems and RDBMSs.
- Private workstations and servers.
- External systems (Internet, commercially available databases).

2. Operational Data Store (ODS):

- It is a repository of current and integrated operational data used for analysis.
- Often structured and supplied with data in same way as DW.
- May act simply as staging area for data to be moved into the warehouse.
- Provides users with the ease of use of a relational database while remaining distant from decision support functions of the DW.

3. Warehouse Manager (Data Manager):

- Operations performed include:
  - Analysis of data to ensure consistency.
  - Transformation/merging of source data from temp storage into DW
  - Creation of indexes.
  - Backing-up and archiving data.

4. Query Manager (Manages User Queries):

- Operations include:
  - directing queries to the appropriate tables and
  - scheduling the execution of queries.
- In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

5. Meta Data: This area of the DW stores all the meta-data (data about data) definitions used by all the processes in the warehouse.

- Used for a variety of purposes:
  - Extraction and loading processes
  - Warehouse management process
  - Query management process
- End-user access tools use meta-data to understand how to build a query.
- Most vendor tools for copy management and end-user data access use their own versions of meta-data.

6. Lightly and Highly Summarized Data:

- It stores all the pre-defined lightly and highly aggregated data generated by the warehouse manager.
- The purpose of summary info is to speed up the performance of queries.
- Removes the requirement to continually perform summary operations (such as sort or group by) in answering user queries.

7. Archive/Backup Data:

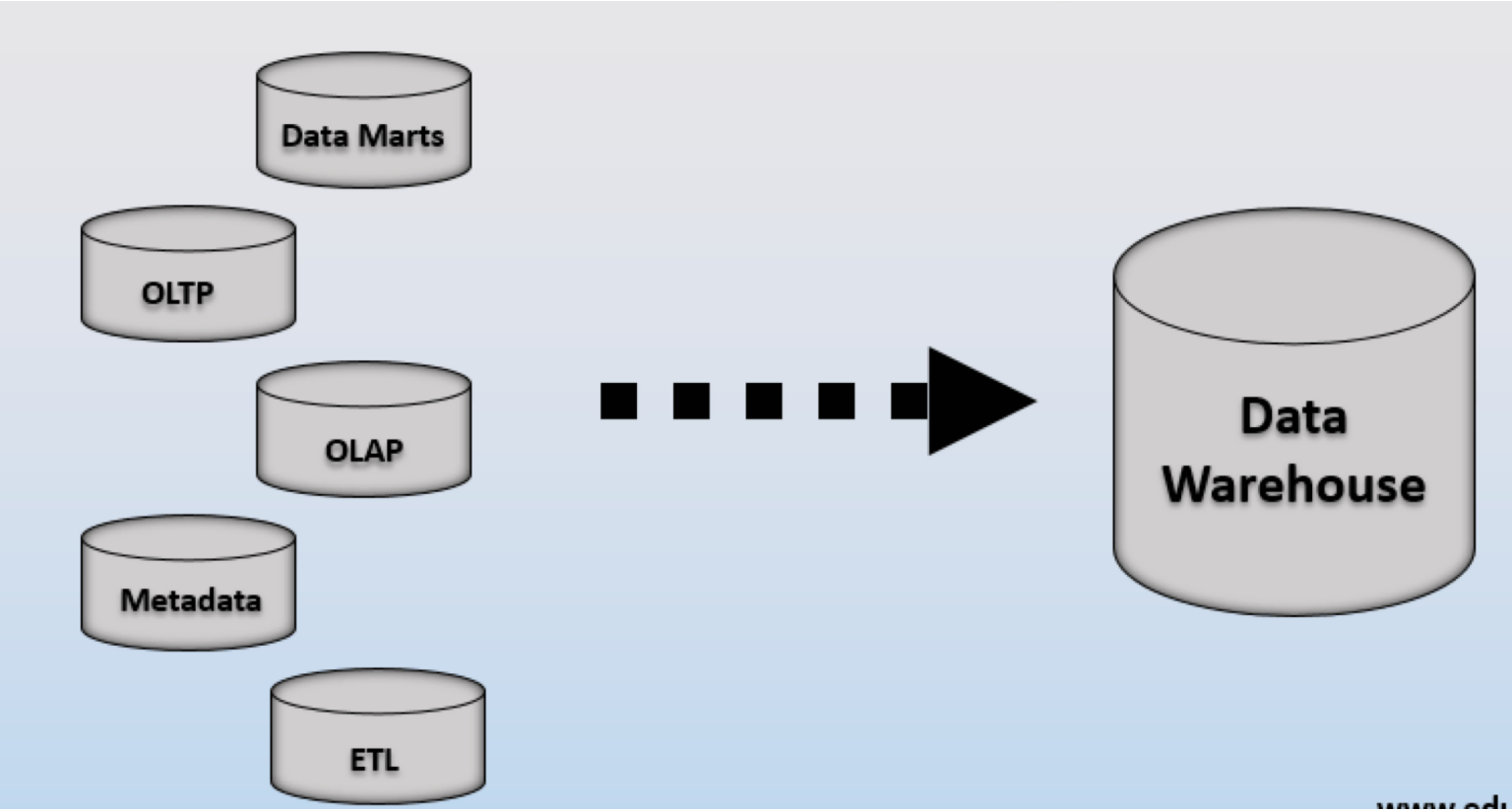
- It stores **detailed and summarized data for the purposes of archiving and backup.**
- May be necessary to backup online summary data if this data is kept beyond the retention period for detailed data.
- The data is transferred to storage archives such as magnetic tape or optical disk.

8. End-User Access Tools:

- The principal purpose of data warehousing is to provide information to business users for strategic decision-making.
- Users interact with the warehouse using end-user access tools.
- There are three main groups of access tools:
  1. Data reporting, query tools
  2. Online analytical processing (OLAP) tools *(Discussed later)*
  3. Data mining tools *(Discussed later)*

Data Warehouse Implementation

- Data Warehouse Implementation is a series of activities that are essential to create a fully functioning Data Warehouse, after classifying, analyzing and designing the Data Warehouse with respect to the requirements provided by the client.



Data Warehouse Implementation

There are various implementation in data warehouses which are as follows.

1. Requirements analysis and capacity planning: The first process in data warehousing involves defining enterprise needs, defining architectures, carrying out capacity planning, and selecting the hardware and software tools. This step will contain be consulting senior management as well as the different stakeholder.

2. Hardware integration: Once the hardware and software has been selected, they require to be put by integrating the servers, the storage methods, and the user software tools.

3. Modelling: Modelling is a significant stage that involves designing the warehouse schema and views. This may contain using a modelling tool if the data warehouses are sophisticated.

Data Warehouse Implementation

4. Physical modelling: For the data warehouses to perform efficiently, physical modelling is needed. This contains designing the physical data warehouse organization, data placement, data partitioning, deciding on access techniques, and indexing.

5. Sources: The information for the data warehouse is likely to come from several data sources. This step contains identifying and connecting the sources using the gateway, ODBC drives, or another wrapper.

6. ETL: The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools. This may contains customize the tool to suit the need of the enterprises.

Data Warehouse Implementation

7. Populate the data warehouses: Once the ETL tools have been agreed upon, testing the tools will be needed, perhaps using a staging area. Once everything is working adequately, the ETL tools may be used in populating the warehouses given the schema and view definition.

8. User applications: For the data warehouses to be helpful, there must be end-user applications. This step contains designing and implementing applications required by the end-users.

9. Roll-out the warehouses and applications: Once the data warehouse has been populated and the end-client applications tested, the warehouse system and the operations may be rolled out for the user's community to use.

Data Mart

- A Data Mart is focused on a single functional area of an organization and contains a subset of data stored in a Data Warehouse.
- A Data Mart is a condensed version of Data Warehouse and is designed for use by a specific department, unit or set of users in an organization. E.g., Marketing, Sales, HR or finance.
- It is often controlled by a single department in an organization.
- Data Mart usually draws data from only a few sources compared to a Data warehouse.
- Data marts are small in size and are more flexible compared to a Datawarehouse.

Why we need Data Mart?

- Data Mart helps to enhance user's response time due to reduction in volume of data.
- It provides easy access to frequently requested data.
- Data mart are simpler to implement when compared to corporate Datawarehouse. At the same time, the cost of implementing Data Mart is certainly lower compared with implementing a full data warehouse.
- Compared to Data Warehouse, a datamart is agile. In case of change in model, datamart can be built quicker due to a smaller size.
- A Datamart is defined by a single Subject Matter Expert. On the contrary data warehouse is defined by interdisciplinary SME from a variety of domains. Hence, Data mart is more open to change compared to Datawarehouse.

Why we need Data Mart?

- Data is partitioned and allows very granular access control privileges.
- Data can be segmented and stored on different hardware/software platforms.

Types of Data Mart

There are three main types of data mart:

1.Dependent: Dependent data marts are created by drawing data directly from operational, external or both sources.

2.Independent: Independent data mart is created without the use of a central data warehouse.

3.Hybrid: This type of data marts can take data from data warehouses or operational systems.

Components of data warehouse

- The Data Warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key Data Warehousing components to make the entire environment functional, manageable and accessible.
- There are mainly five Data Warehouse Components:

1. Data Warehouse Database

- The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.

Components of data warehouse

2. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL):

- The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the datawarehouse. They are also called Extract, Transform and Load (ETL) Tools.

Their functionality includes:

1. Anonymize data as per regulatory stipulations.
2. Eliminating unwanted data in operational databases from loading into Data warehouse.
3. Search and replace common names and definitions for data arriving from different sources.
4. Calculating summaries and derived data
5. In case of missing data, populate them with defaults.
6. De-duplicated repeated data arriving from multiple datasources.

Components of data warehouse

3. Metadata

- The name Meta Data suggests some high-level technological Data Warehousing Concepts. However, it is quite simple. Metadata is data about data which defines the data warehouse.
- It is used for building, maintaining and managing the data warehouse.
- In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

For example: A line in sales database may contain: 4030 KJ732 299.90

This is a meaningless data until we consult the Meta that tell us it was

Model number: 4030, Sales Agent ID: KJ732, Total sales amount of $299.90. Therefore, Meta Data are essential ingredients in the transformation of data into knowledge.

Components of data warehouse

4. Query Tools

- One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.

These tools fall into four different categories:

1. Query and reporting tools
2. Application Development tools
3. Data mining tools
4. OLAP tools

Components of data warehouse

5. Data Marts

- A data mart is an access layer which is used to get data out to the users.
- It is presented as an option for large size data warehouse as it takes less time and money to build. However, there is no standard definition of a data mart is differing from person to person.
- In a simple word Data mart is a subsidiary of a data warehouse and is used for partition of data which is created for the specific group of users.
- Data marts could be created in the same database as the Datawarehouse or a physically separate Database.

Why We Need Data Warehouse?

Advantages of Data Warehouse:

- Data warehouse allows business users to quickly access critical data from some sources all in one place.
- Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
- Data warehouse helps to reduce total turnaround time for analysis and reporting.
- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

Why We Need Data Warehouse?

Disadvantages of Data Warehouse:

- Not an ideal option for unstructured data.
- Creation and Implementation of Data Warehouse is surely time confusing affair.
- Data Warehouse can be outdated relatively quickly
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- The data warehouse may seem easy, but actually, it is too complex for the average users.
- Despite best efforts at project management, data warehousing project scope will always increase.
- Sometime warehouse users will develop different business rules.
- Organisations need to spend lots of their resources for training and Implementation purpose.

Trends in data warehousing

1. Complex Data Marts Will Define the Future Business Models:

- Data marts surfaced as a subset of data warehouses, designed to address the requirements of a specific business function. However, the ability of large and complex data marts to pull data from disparate sources and make it accessible to business users is making it a rising trend in data warehousing.
- The recent developments in the construction of data marts allow integration of web and enterprise data. This enables evaluation of the transformation process from source to data mart. Also, it extends the benefits of analytics throughout the organization.
- Another feature that is expected to enhance the functionality of data marts is speed. Modern data marts will be designed to offer cloud-scale speed with 24/7 functional processing power, network, and disk. The result will be efficient, cost-effective, and resilient data marts.

Trends in data warehousing

2. Column-based Storage is on the Rise

- When it comes to retrieving analytical queries, the efficiency of column-based storage is higher than its row-based alternative. This is one of the reasons this trend is gradually gaining popularity.
- The primary goal of data warehousing is to store data in a way that speeds up the query response time, consequently enabling efficient data evaluation and analyzation. And column-based storage can make that happen. It significantly compresses the data because columns store similar values.
- Storing data in a column-oriented data warehouse can help businesses conduct advanced business analytics. In addition, this storage system allows tight integrations and easy data warehouse setup due to enhanced disk performance. It cuts down the system's I/O requirements and ensures minimum data is uploaded from the disk. Also, column-based DBMSs make for a good data mart platform.

Trends in data warehousing

3. Mixed Workloads Are Becoming Common

A data warehouse platform delivers six types of workloads:

Basic reporting

- Continuous/real-time load
- Batch/bulk load
- Operational BI
- Online analytical processing (OLAP)
- Data mining

To ensure optimum performance of a data warehouse that delivers all these workloads, it's essential to plan and assess the output predictability.

Inability to do so may cause three major problems:

1. Sustainability issues
2. Increased administration costs (due to added volume and workload)
3. Low performance

Trends in data warehousing

4. Data Warehouse Automation (DWA)

- Data warehouse implementations are generally dependent on IT personnel. It can take years to build a data warehouse, making the whole process time-intensive, expensive, and slow.
- Adding the automation factor to the equation makes it easier for organizations to navigate the complexities of data warehousing and eliminates the repetitive, time-consuming tasks from the process cycle. This consequently results in low project costs and high productivity.
- Moreover, DWA significantly reduces the dependency on the IT staff.
- It eliminates the need for hand coding, empowering business users with less technical knowledge to take the lead, simultaneously making the process cycle faster.

Trends in data warehousing

5. Data Warehouses are Becoming Cloud-centric

- The cloud is fast becoming a preferred choice for users looking to acquire data warehousing capabilities. Why? Because in addition to supporting all the functions like that of a traditional data warehouse, cloud data warehouses optimize deployments like data-governance hubs, BI backends, analytic data marts, etc.
- With the load-and-go feature, cloud data warehouses eliminate the need for businesses to invest in hardware and IT staff.
- The feasibility to deliver dynamic workflow management and high performance, without any manual training, makes this solution a cost-effective option. In addition, the efficient compression and built-in technology enable scalability when data varieties and volumes grow.

Thank you !!!