Unit I

- Introduction
  - Data warehousing
  - Multidimentional data model
  - OLAP operations
  - Warehouse schema
  - DW architecture
  - Warehouse server
  - Metadata da
  - OLAP Engine
  - DW backend process

FOCUS on making yourself BETTER, not on thinking that you are better

Introduction

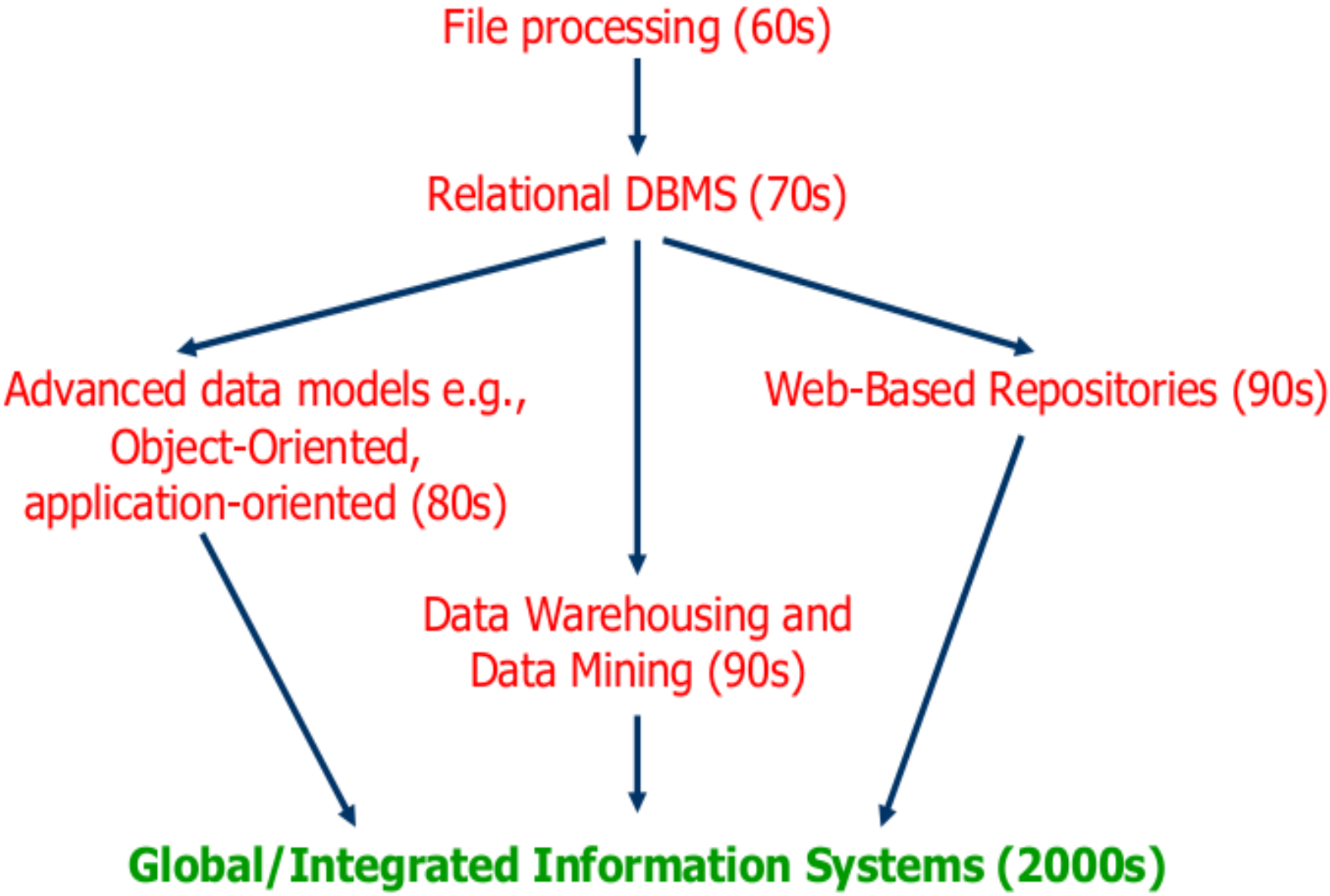
- Computing technology – influences
  - Scientific computing
  - Business Data processing
- Earlier limitations
  - Storage size
  - Speed of accessing data
  - Highly qualified professionals
- All these limitations are overcome when the storage and m/c cost dratically reduces

FOCUS on making yourself BETTER, not on thinking that you are better

- The layered architecuter of modern DBMS
  - External interfaces – deals with DDL, DML and host lang interfaces
  - Lang processing – query processing, optimisation
  - Code Generation – procedural parts
  - Transaction mgmt – concurrency control, logging and recovery
  - Storage mgmt – manages memory ,buffer and secondary storage

FOCUS on making yourself BETTER, not on thinking that you are better

Evolution of database technology



FOCUS on making yourself BETTER, not on thinking that you are better

Introduction

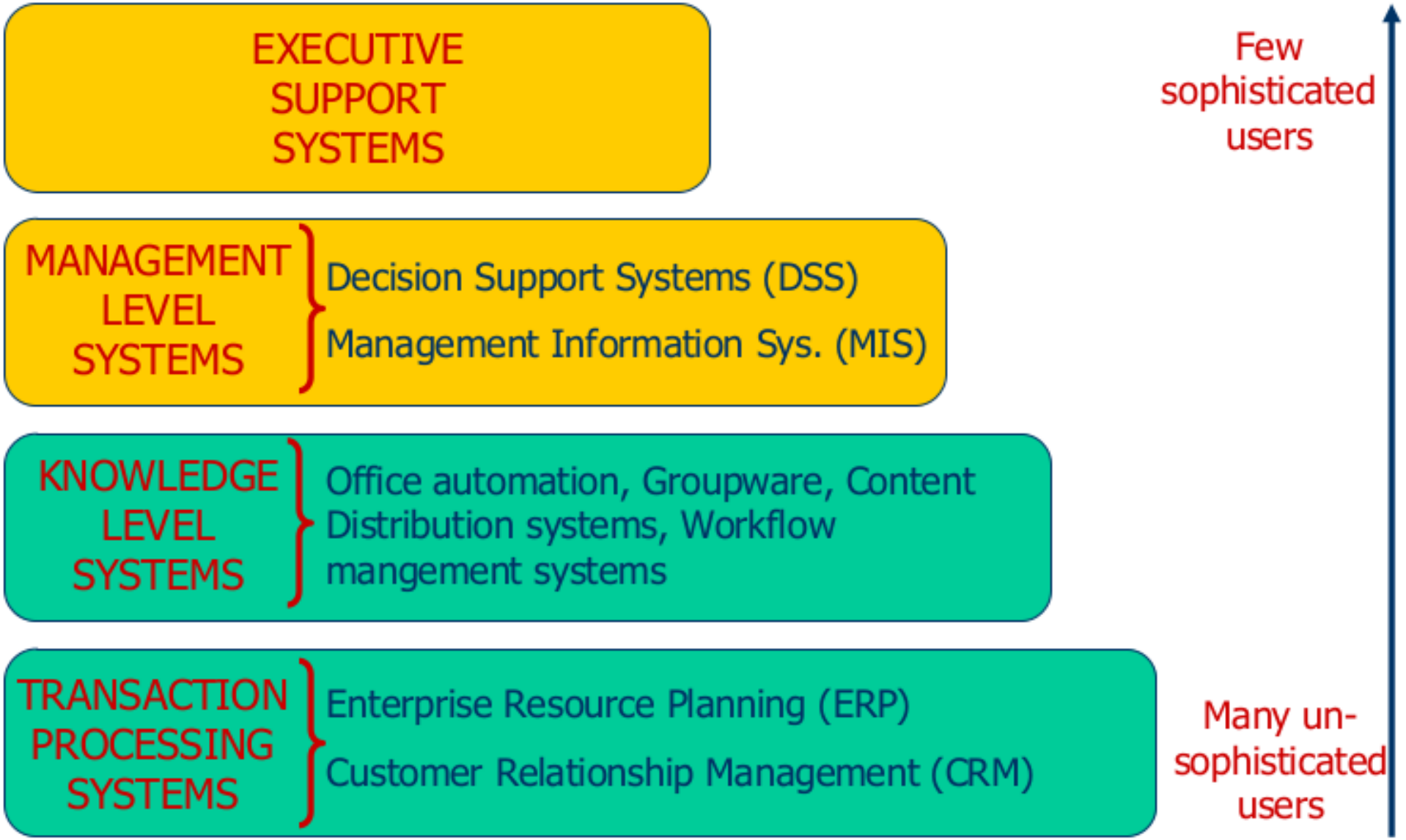
- The change (i.e., data integeration) in this layered approach leads to integerated access to multiple DBs with two modes
  - Virtual - a media to collect data from multiple sources
  - Materialised – a universal DBMS / Data Warehouse
- This change introduces
  - Data Mining
  - Data Warehousing

FOCUS on making yourself BETTER, not on thinking that you are better

- There are *limitations* in the
  - traditional data analysis techniques
    - regression analysis : statistical process for estimating the relationships between variables
    - cluster analysis : grouping of set of objects (similar objects)
    - Numerical taxonomy : classification system using numerical algo
    - multidimensional analysis : data analysis process with dimensions/measurements
    - other multivariate statistical methods, and stochastic (probability) models.
- These techniques have been widely used for solving many practical problems
- Though , these are however primarily oriented toward the extraction of quantitative and statistical data characteristics

FOCUS on making yourself BETTER, not on thinking that you are better

Major types of information systems within an organization

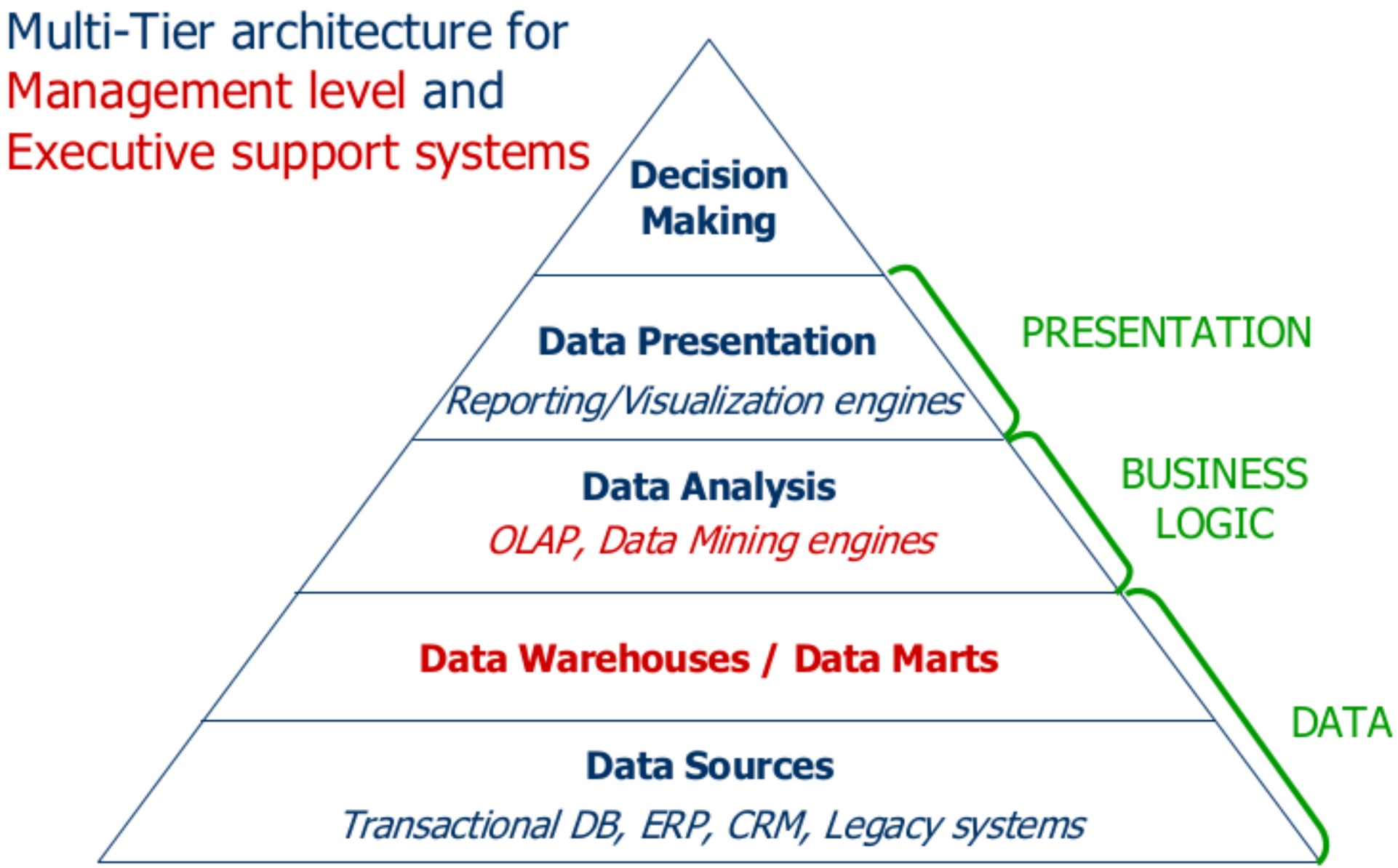


FOCUS on making yourself BETTER, not on thinking that you are better

- **Transaction processing systems:**
  - Support the **operational level** of the organization, possibly integrating needs of different functional areas (ERP);
  - Perform and record the **daily transactions** necessary to the conduct of the business
  - Execute simple read/update operations on traditional databases, aiming at maximizing transaction throughput
- Their activity is described as:
  - **OLTP** (On-Line Transaction Processing)

FOCUS on making yourself BETTER, not on thinking that you are better

- **Knowledge level systems:**
  - provide digital support for
    - managing documents (office automation),
    - ser cooperation and communication (groupware),
    - storing and retrieving information (content distribution),
    - automation of business procedures (workflow management)
- **Management level systems:**
  - support planning, controlling and semi-structured decision making at management level by providing
    - reports and analyses of current and historical data
- **Executive support systems:**
  - support unstructured decision making at the strategic level of the organization



FOCUS on making yourself BETTER, not on thinking that you are better

- **OLAP (On-Line Analytical Processing):**
  - Reporting based on (multidimensional) data analysis
  - Read-only access on repositories of moderate-large size (typically, data warehouses), aiming at maximizing response time
- **Data Mining:**
  - Discovery of novel, implicit patterns from, possibly heterogeneous, data sources
  - Use a mix of sophisticated statistical and high performance computing techniques

- Data Mining
  - Theory
  - Algorithms and implementations
  - Applications

- **DATA WAREHOUSE**
  - Database with the following distinctive characteristics:
    - Separate from operational databases
    - Subject oriented:
      - provides a simple, concise view on one or more selected areas, in support of the decision process
    - Constructed by integrating multiple, heterogeneous data sources
    - Contains historical data: spans a much longer time horizon than operational databases
    - (Mostly) Read-Only access: periodic, infrequent updates

FOCUS on making yourself BETTER, not on thinking that you are better

DW Definition

- A **data warehouse** is a
  - **database** designed to enable **business intelligence** activities:
  - it exists to help users understand and **enhance their organization**'s performance.
  - It is designed for **query and analysis** rather than for transaction processing, and usually contains **historical data** derived from transaction data, but can include data from other(**multiple**) sources.
  - Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources.

- DW helps in:
  - Maintaining historical records
  - Analyzing the data
    - to gain a better understanding of the business
    - & to improve the business

- **characteristics of a data warehouse**
  - Subject Oriented
  - Integrated
  - Nonvolatile
  - Time Varient

- **Subject Oriented**
  - Data warehouses are designed to help in analysis of data.
  - Organized around subjects (customers,products etc.,) not applications
  - For example, to learn more about company's sales data, we can build a data warehouse that concentrates on sales.
  - Using this data warehouse, we can answer questions such as
    - "Who was our best customer for this item last year?"
    - "Who is likely to be our best customer next year?"
  - This ability to define a data warehouse by subject matter (sales) makes the data warehouse subject oriented.
  - It retrieves info necessary for the DSS

FOCUS on making yourself BETTER, not on thinking that you are better

- **Integrated**
  - Integration is closely related to subject orientation.
  - Data warehouses must put data from disparate sources into a consistent format.
  - Integerating multiple/heterogeneous sources (RDBMS, flat files, OLTP files)
  - Consistent coding convention
  - They must resolve such problems as naming conflicts and inconsistencies among units of measure.
    - For which data cleaning/data integeration techniques are used
  - When they achieve this, they are said to be integrated.

- **Nonvolatile**
  - DW is physically a separate store of data
  - Data are transformd to DW from the applications found in separate environment
  - Once data entered into the data warehouse, data should not change. i.e., not updated but only loaded/refreshed/accessed
  - This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred.
  - DW does not require transaction processing / concurrency control or recovery

FOCUS on making yourself BETTER, not on thinking that you are better

- **Time Varient**
  - DW to give **historical perspective** / so every data has the element of time
  - Contains data for comparisons, trending and forecasting
  - A data warehouse's focus on change over time
  - In order to discover trends and identify hidden patterns and relationships in business, analysts need large amounts of data.(5-10 years of data)
  - This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that **historical data be moved to an archive.**

- **Key Characteristics of a Data Warehouse**
  - Data is structured for simplicity of access and high-speed query performance.
  - End users are time-sensitive and desire speed-of-thought response times.
  - Large amounts of historical data are used.
  - Queries often retrieve large amounts of data, perhaps many thousands of rows.
  - Both predefined and ad hoc queries are common.
  - The data load involves multiple sources and transformations.
- In general, ***fast query performance with high data throughput*** is the key to a successful data warehouse.

FOCUS on making yourself BETTER, not on thinking that you are better

Operational vs analytical

- Operational queries – What , Who type
- Operational – data spectrum smaller, few relationships, hierarchical/relational/ER models
- Operational – these applications facilitate focus on fast interactive access to data on efficient execution of particular transaction

- Analytical queries – Why , What if type
- Analytical – wider spectrum of data, many-data relationshipsa, multidimentional view
- Analytical – these facilitate focus on business patterns

FOCUS on making yourself BETTER, not on thinking that you are better

Operational vs analytical data

Application	Oper	Anal
Content changes	RElatime	Daily to monthly
Structural changes	infrequent	Slowly changing
Detail level	transaction	summary
User interface	Static, application dependent	Dynamic , business dependent
Response time	Real time	Real time
Age of data	current	historical
Access path	deterministic	Non-determininstic

FOCUS on making yourself BETTER, not on thinking that you are better

Integerated access to data sources

- Data integeration
  - From multiple, heterogeneous sources
- It can be effeted in two modes
  - On-demand
    - Lazy model, query-driven, virtual system
      -
  - In-advance
    - Eager model, analysis-driven, materialised systems

Integerated access to data sources

- On-demand
  - Lazy model, query-driven, virtual system
    - Find relevant info sources
    - Generate sub-query for each sources
    - Integrate the results obtained
    - Return the result to the application
    - Integeration occurs when the query is processed
- In-advance
  - Eager model, analysis-driven, materialised systems
    - Extract relevant info from multiple sources
    - Filter and consolidate
    - Store it in a separate DB/DW (queries are given here directly)

FOCUS on making yourself BETTER, not on thinking that you are better

DW vs DB

- Database
  - Used for Online Transactional Processing (OLTP) but can be used for other purposes such as Data Warehousing. This records the data from the user for history.
  - The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.
  - Entity – Relational modeling techniques are used for RDMS database design.
  - Optimized for write operation.
  - Performance is low for analysis queries.

DW vs DB

- Data Warehouse
  - Used for Online Analytical Processing (OLAP). This reads the historical data for the Users for business decisions.
  - The Tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
  - MD Data – Modeling techniques are used for the Data Warehouse design.
  - Optimized for read operations.
  - High performance for analytical queries.
  - Is usually a Database.

- More diff
  - Ref PDF.....

- **Types** of Data Warehouses
  - **Enterprise Warehouse:** covers all areas of interest for an organization
  - **Data Mart:** covers a subset of corporate-wide data that is of interest for a specific user group (e.g., marketing).
  - **Virtual Warehouse:** offers a set of views constructed on demand on operational databases. Some of the views could be materialized (precomputed)

Why to mine data

- **Lots of data** is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/
  - grocery stores
  - Bank/Credit Card transactions
- **Computers** have become cheaper and more powerful
- **Competitive Pressure** is Strong
- Provide **better, customized services** for an edge (e.g. in Customer Relationship Management)

FOCUS on making yourself BETTER, not on thinking that you are better

- Data **collected and stored at enormous speeds** (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene
  - expression data
  - scientific simulations
- generating **terabytes** of data
- **Traditional techniques infeasible for raw data**
- Data mining may help **scientists**
  - in classifying and segmenting data
  - in Hypothesis Formation

FOCUS on making yourself BETTER, not on thinking that you are better

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

What is data minng

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- the practice of examining large pre-existing databases in order to generate new information.
- data mining (sometimes called data or knowledge discovery (KDD)) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both

- **Data mining** is sorting through data to identify patterns and establish relationships.
- Data mining parameters include:
  - **Association** - looking for patterns where one event is connected to another event
  - **Sequence or path analysis** - looking for patterns where one event leads to another later event
  - **Classification** - looking for new patterns (May result in a change in the way the data is organized)
  - **Clustering** - finding and visually documenting groups of facts not previously known
  - **Forecasting** - discovering patterns in data that can lead to reasonable predictions about the future

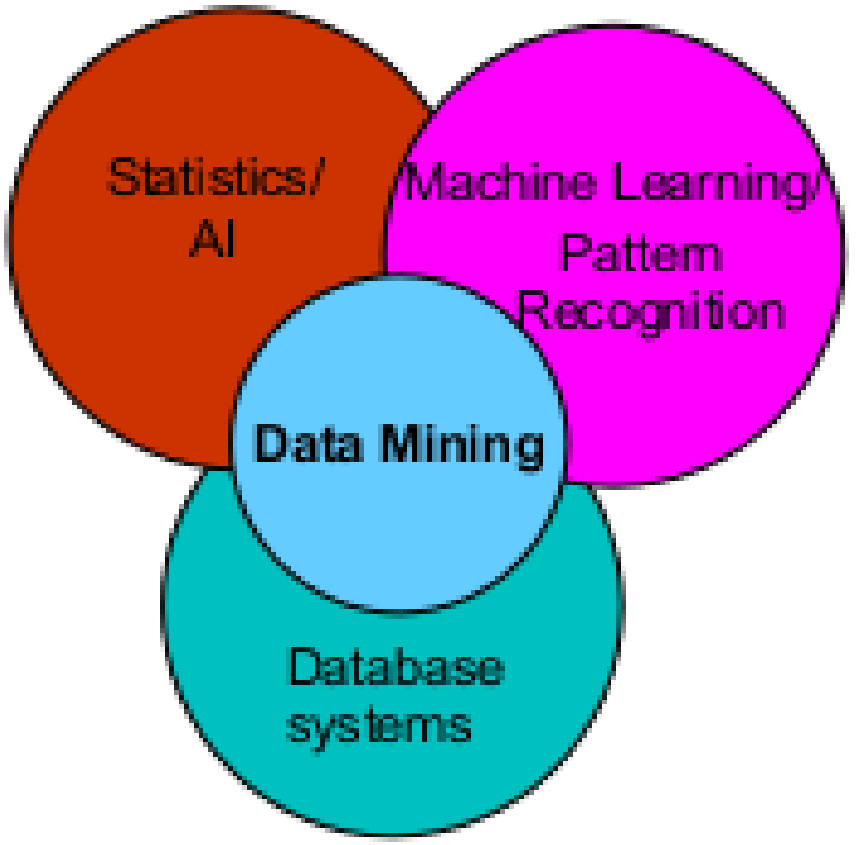
FOCUS on making yourself BETTER, not on thinking that you are better

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



FOCUS on making yourself BETTER, not on thinking that you are better

Multidimensional Data model

- Data warehouses and OLAP tools are based on a multidimensional data model.
- This model views data in the form of a **data cube**
- A data cube allows data to be modeled and viewed in multiple dimensions.
- It is defined by dimensions and facts.
- Why data cube/MD model .....

FOCUS on making yourself BETTER, not on thinking that you are better

employee	quarter	location	commission
Alice	Q1	Domestic	800
Alice	Q1	International	200
Bob	Q1	Domestic	500
Mary	Q1	Domestic	1200
Mary	Q1	International	800
Bob	Q2	International	1500
Mary	Q2	Domestic	500
Jim	Q2	Domestic	1000

```
SELECT employee, quarter, SUM(commission)
FROM   comm
GROUP BY employee,quarter
UNION
SELECT 'ALL', quarter, SUM(commission)
FROM   comm
GROUP BY quarter
UNION
SELECT employee, 'ALL', SUM(commission)
FROM   comm
GROUP BY employee
UNION
SELECT 'ALL','ALL', SUM(commission)
FROM   comm
```

	Alice	Bob	Mary	Jim	total(ALL)
Q1	1000	500	2000		3500
Q2		1500	500	1000	3000
total(ALL)	1000	2000	2500	1000	6500

T h e number of needed unions is exponential in the number of dimensions.
A complex query may result in many scans of the base table, leading to poor performance.

FOCUS on making yourself BETTER, not on thinking that you are better

- such sub-totals are very common in OLAP queries, it is desired to define a new operator for the collection of such sub-totals, namely, data cube.
- A data cube is essentially the
  - generalization of the cross tabular values.
  - generalization is the aggregation function.
  - generalization is the dimension hierarchy.
- (The generalization happens in several perspectives)

- In multidimensional model .....
- dimensions are the perspectives or entities with respect to what an organization wants to keep records.
- Each dimension may have a table associated with it, called a dimension table, (that describes dimension)
- Dimension tables can be specified by
  - users or experts,
  - or automatically generated and adjusted based on data distributions.

- A multidimensional data model is typically organized around a central theme(e.g sales)
- This theme is represented by a fact table.
- Facts are numerical measures.(i.e., these are the quantities by which we want to analyze relationships between dimensions.)

- Examples of facts for a sales data warehouse
  - dollars sold (sales amount in dollars),
  - units sold (number of units sold),
  - amount budgeted.
- The fact table contains the
  - names of the facts, or measures,
  - as well as keys to each of the related dimension tables

Multidimensional Data model

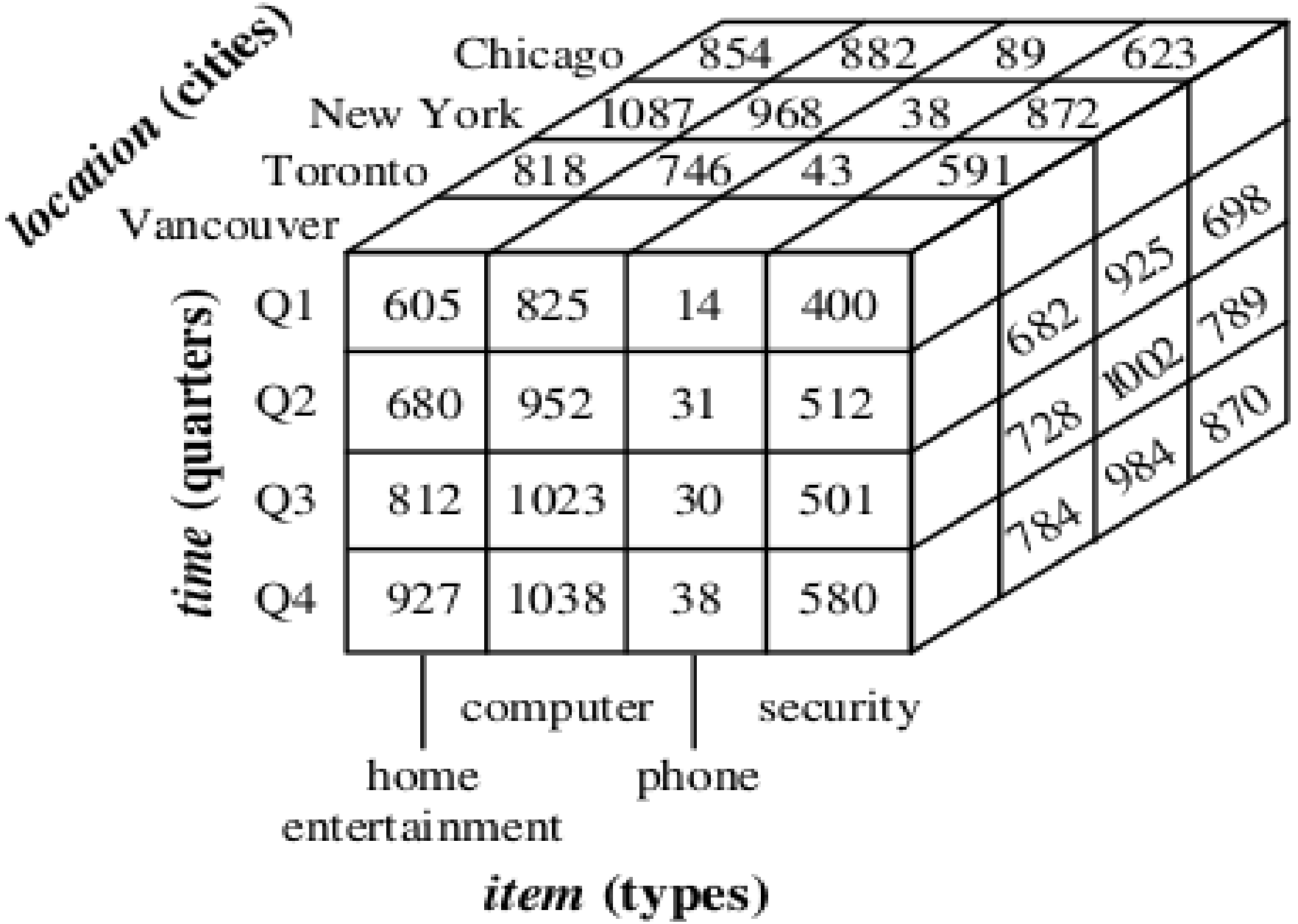
- **Core design** of the DW uses multidimensional view of the data model
- e.g
- **Why data cube**
  - e.g follows.....

A 2-D view of sales data for *AllElectronics* 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 = "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

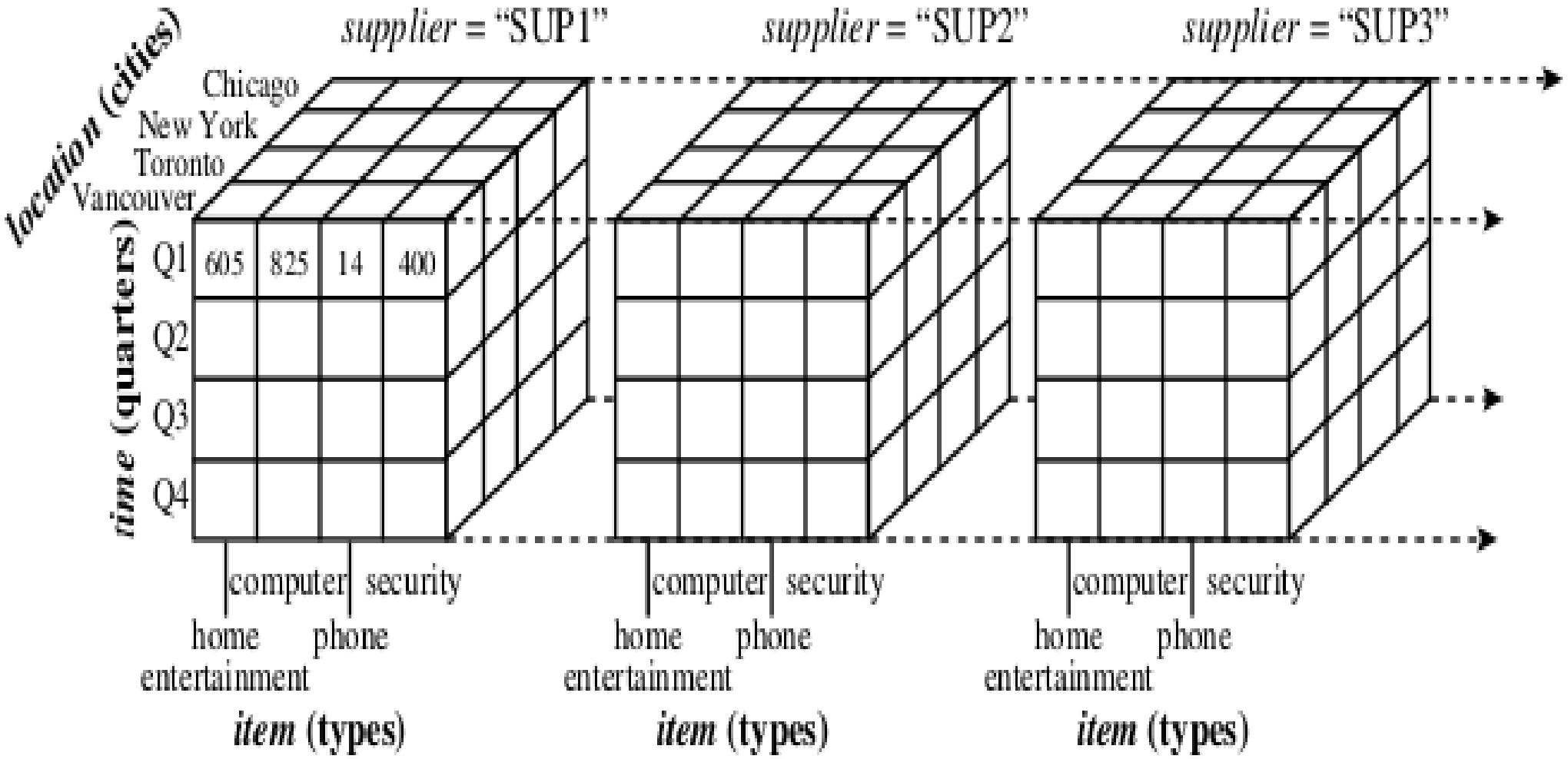
A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	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

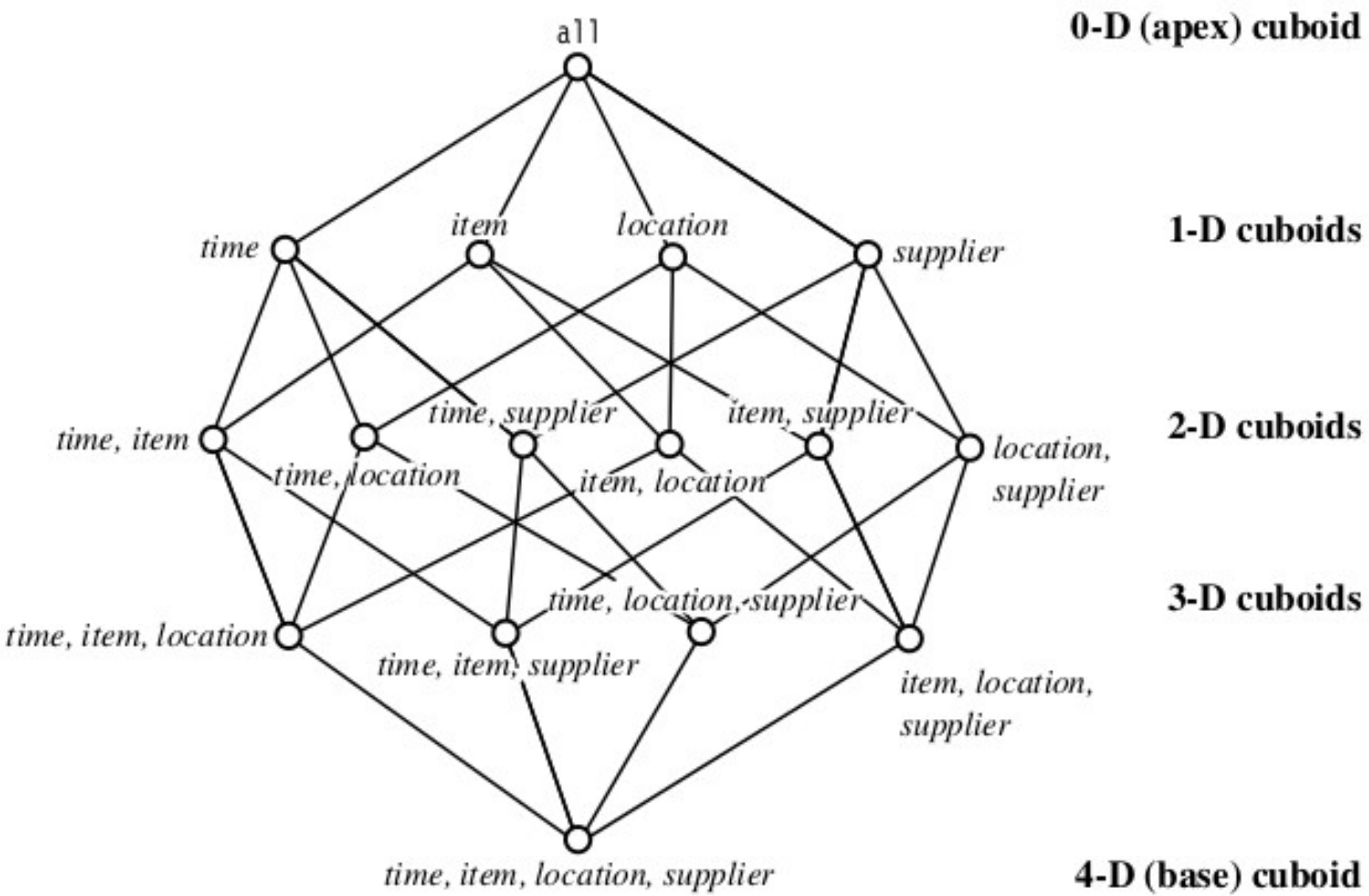


FOCUS on making yourself BETTER, not on thinking that you are better

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).



FOCUS on making yourself BETTER, not on thinking that you are better

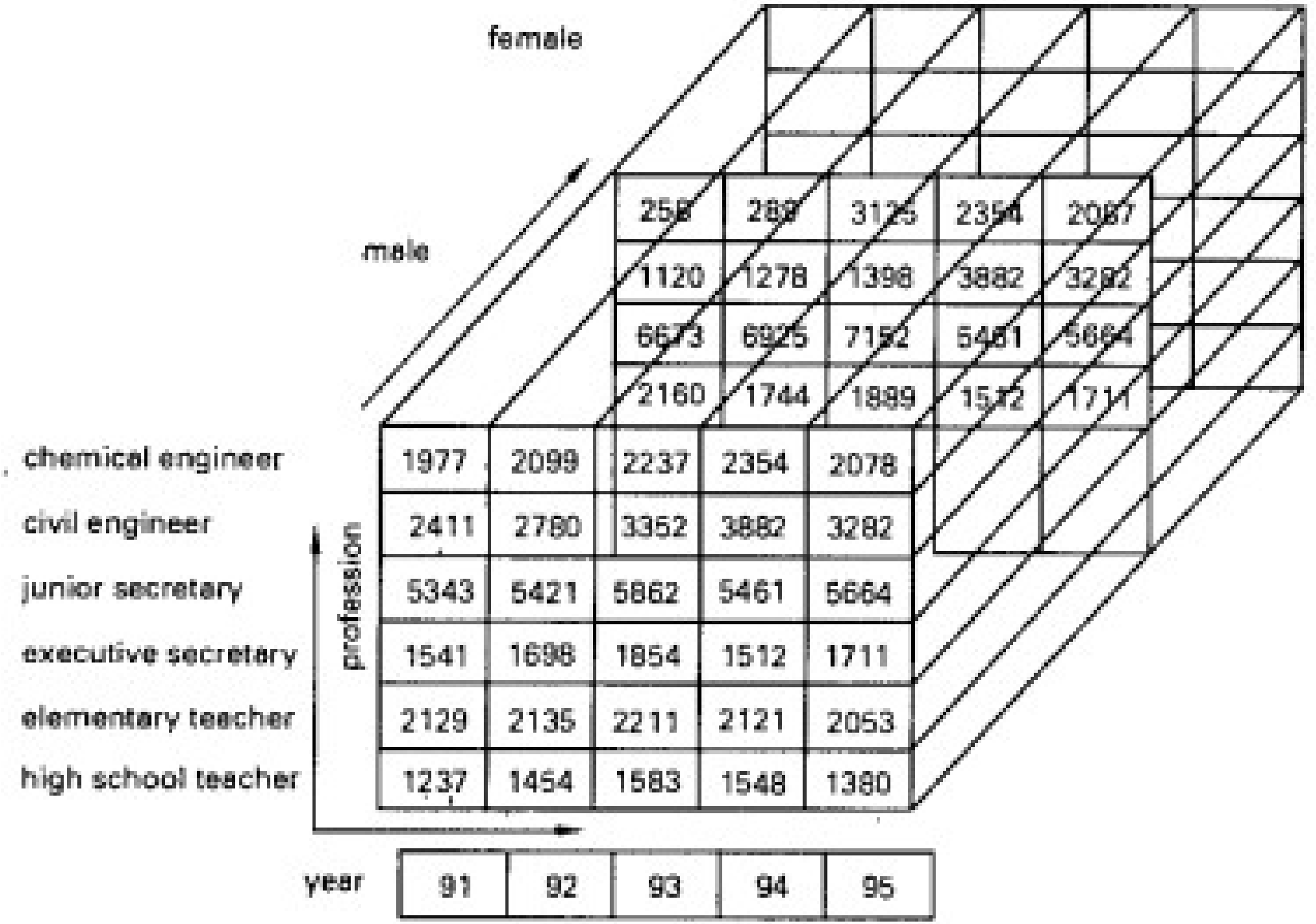


FOCUS on making yourself BETTER, not on thinking that you are better

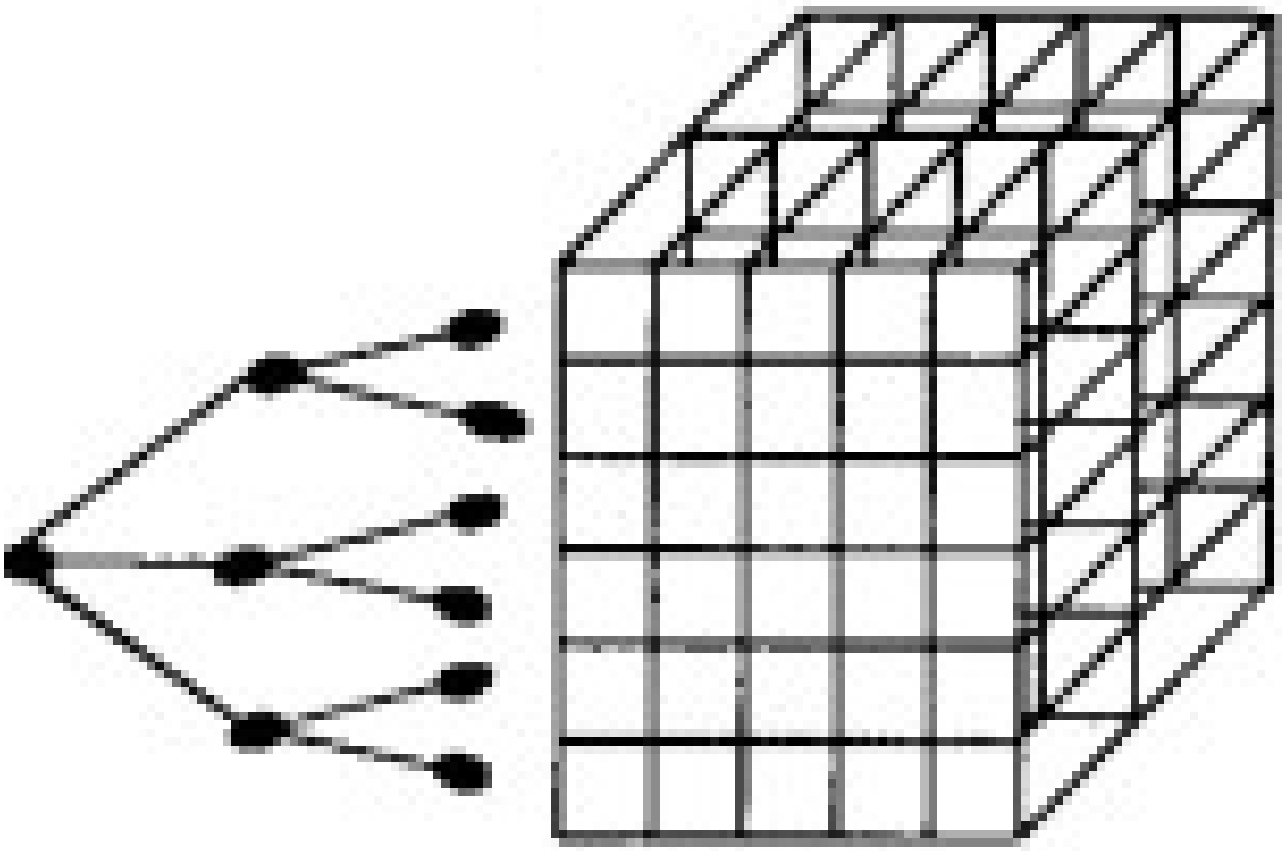
Multidimensional Data model

- **Data Cube**
  - N dimensional data cube is representd as C[A1,A2,....,An]
  - N dimensions,
    - each dimension described by set of attributes
    - Attributes may be related via hierarchy/lattice
    - perspective or entity wrt what orgn wants to keep record
    - Each dimension represents a theme/subject
  - |Ai| no of distinct elements in the dimension
  - Distinct element – data row of C
  - One data cell – numeric measures of the Ai

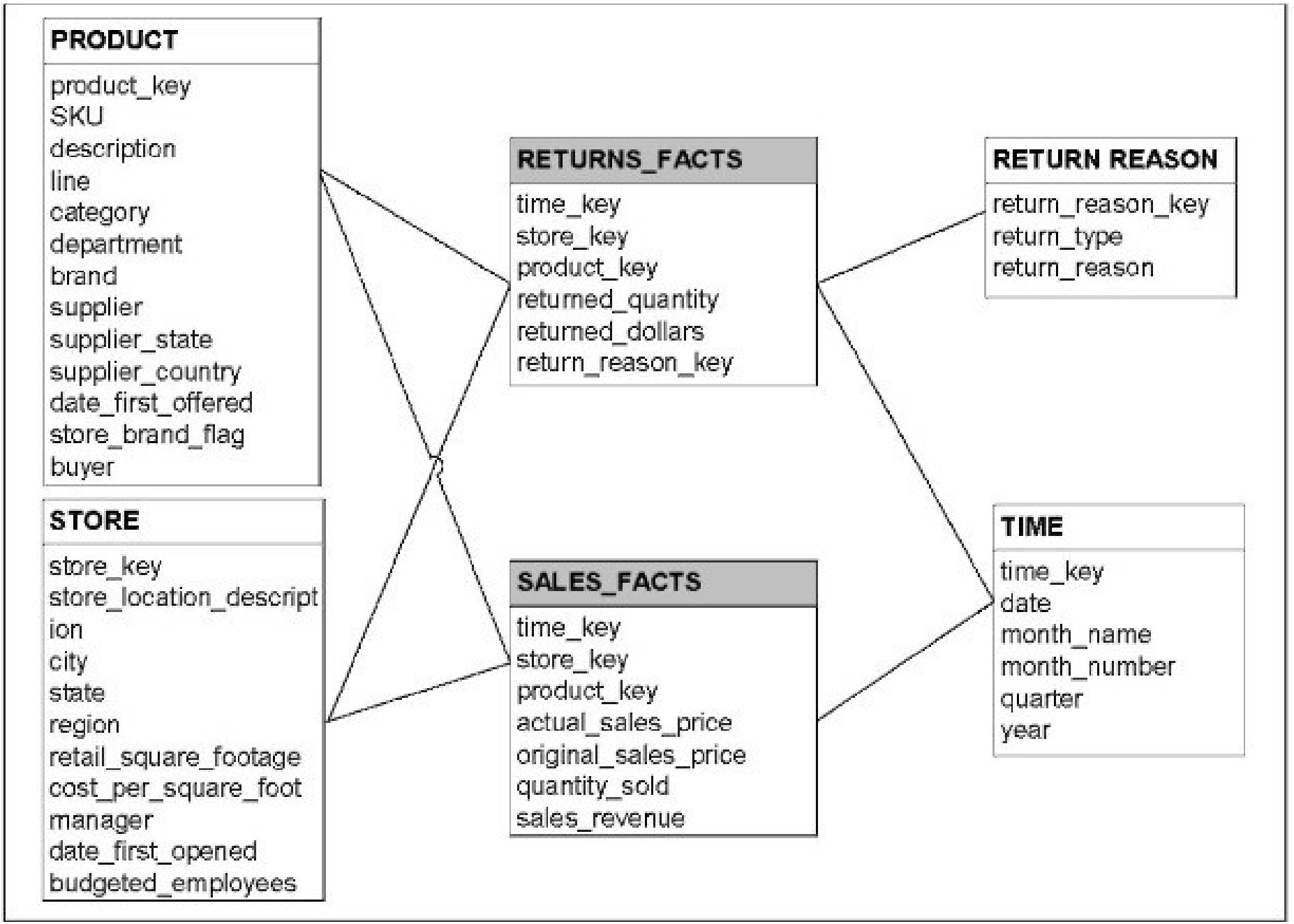
			Professional Class					
			Engineer		Secretary		Teaching	
			PROFESSION		PROFESSION		PROFESSION	
			Chemical Engineers	Civil Engineers	Junior Secretary	Executive Secretary	Elementary Teachers	High School Teachers
SEX	MALE	91	1977	2411	5343	1541	2129	1237
		92	2099	2780	5421	1698	2135	1457
		93	2237	3352	5862	1854	2211	1583
		94	2354	3882	5461	1512	2121	1548
		95	2078	3282	5664	1711	2053	1380
	FEMALE	91	258	1120	6673	1623	2160	2751
		92	289	1276	6925	1744	2175	2993
		93	312	1398	7152	1889	2189	3125
		94	581	1216	6543	1534	2857	2387
		95	329	.1321	6129	1567	2453	3287



FOCUS on making yourself BETTER, not on thinking that you are better



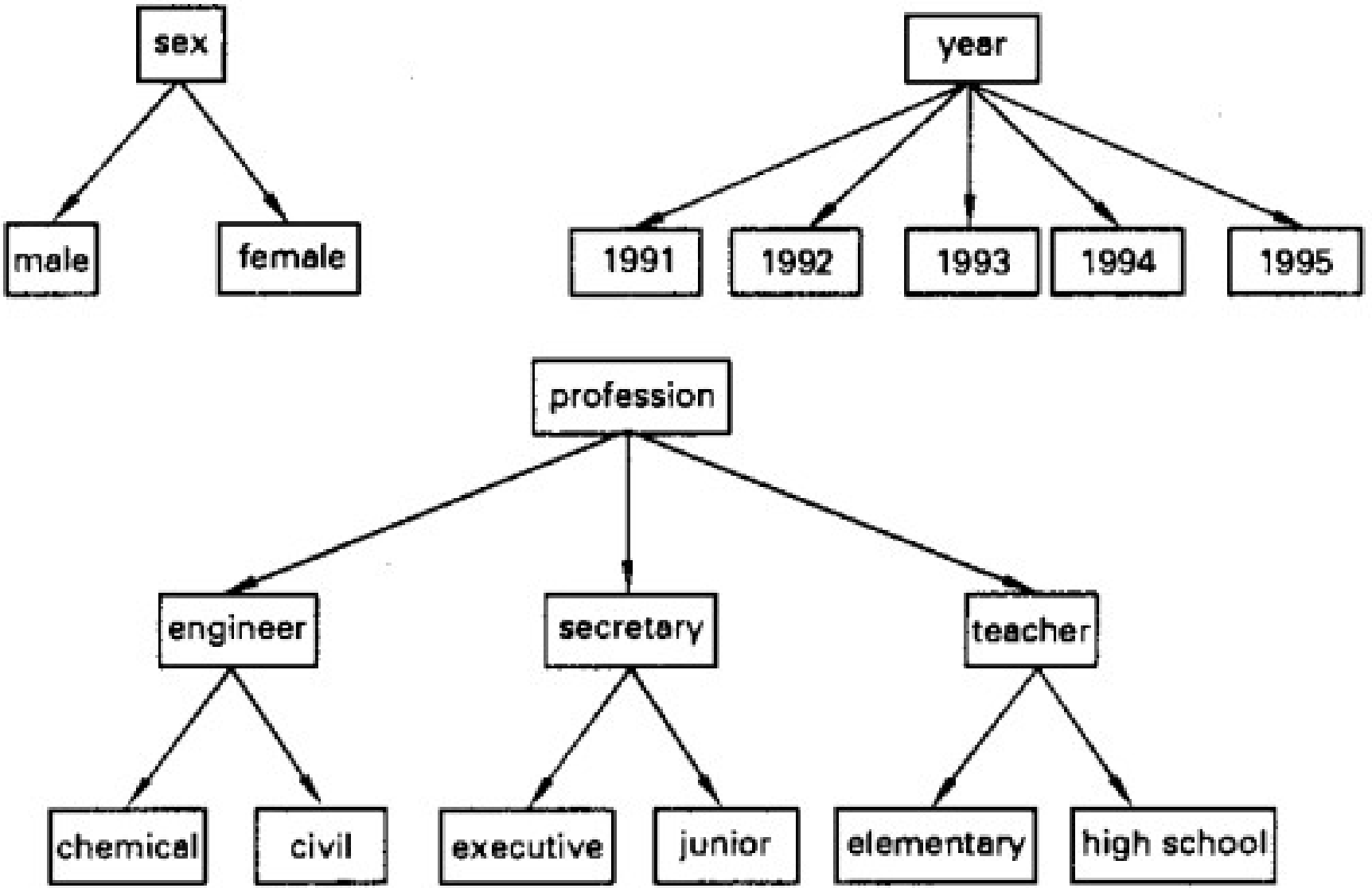
FOCUS on making yourself BETTER, not on thinking that you are better



FOCUS on making yourself BETTER, not on thinking that you are better

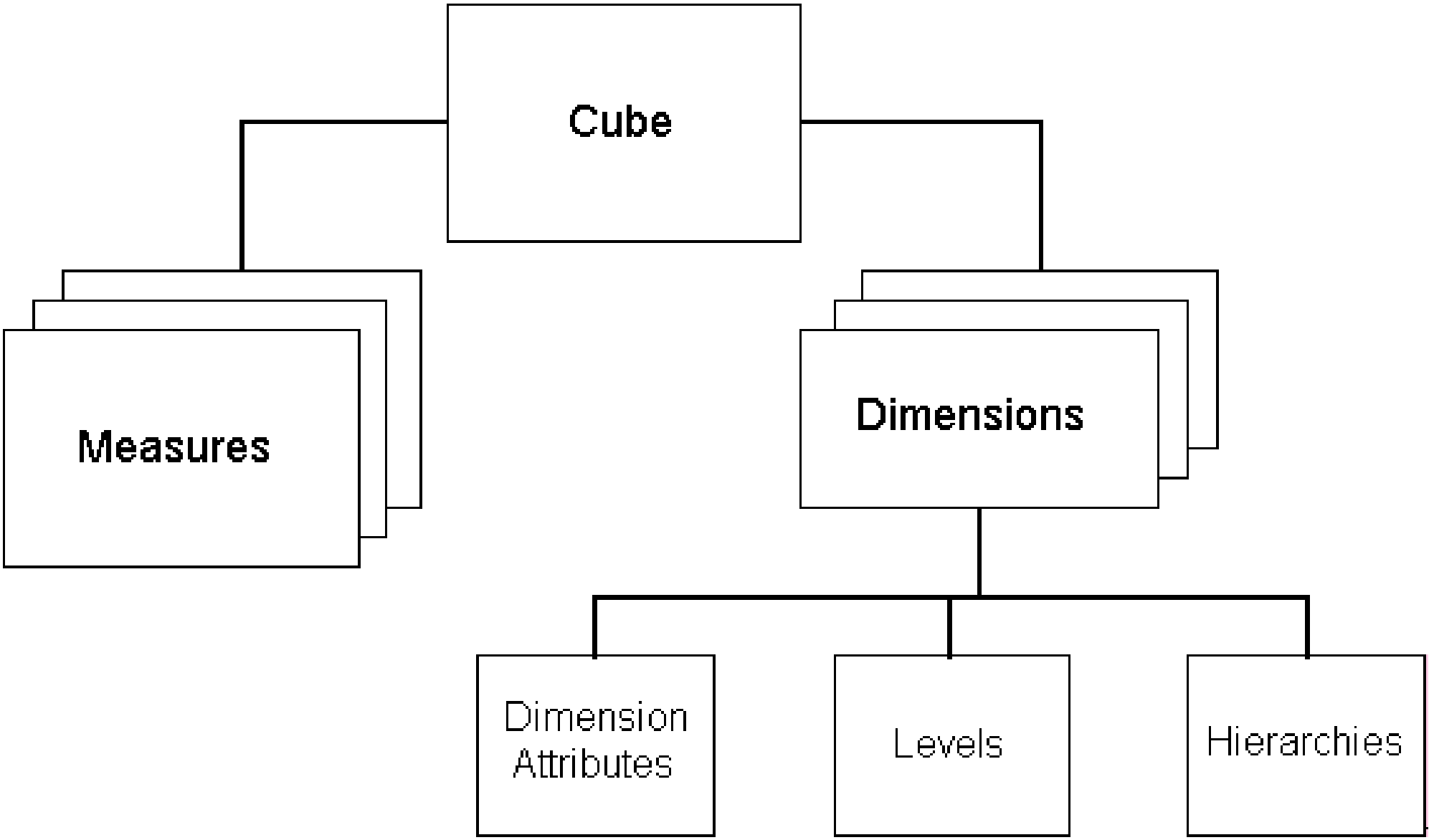
Dimension modelling

- Each dimension provides a lot of **semantic** information
- It represents the **hierarchical** relationships between each element
- Used for **structuring** data based on business concepts
- It structures numeric **measures** and **dimensions**



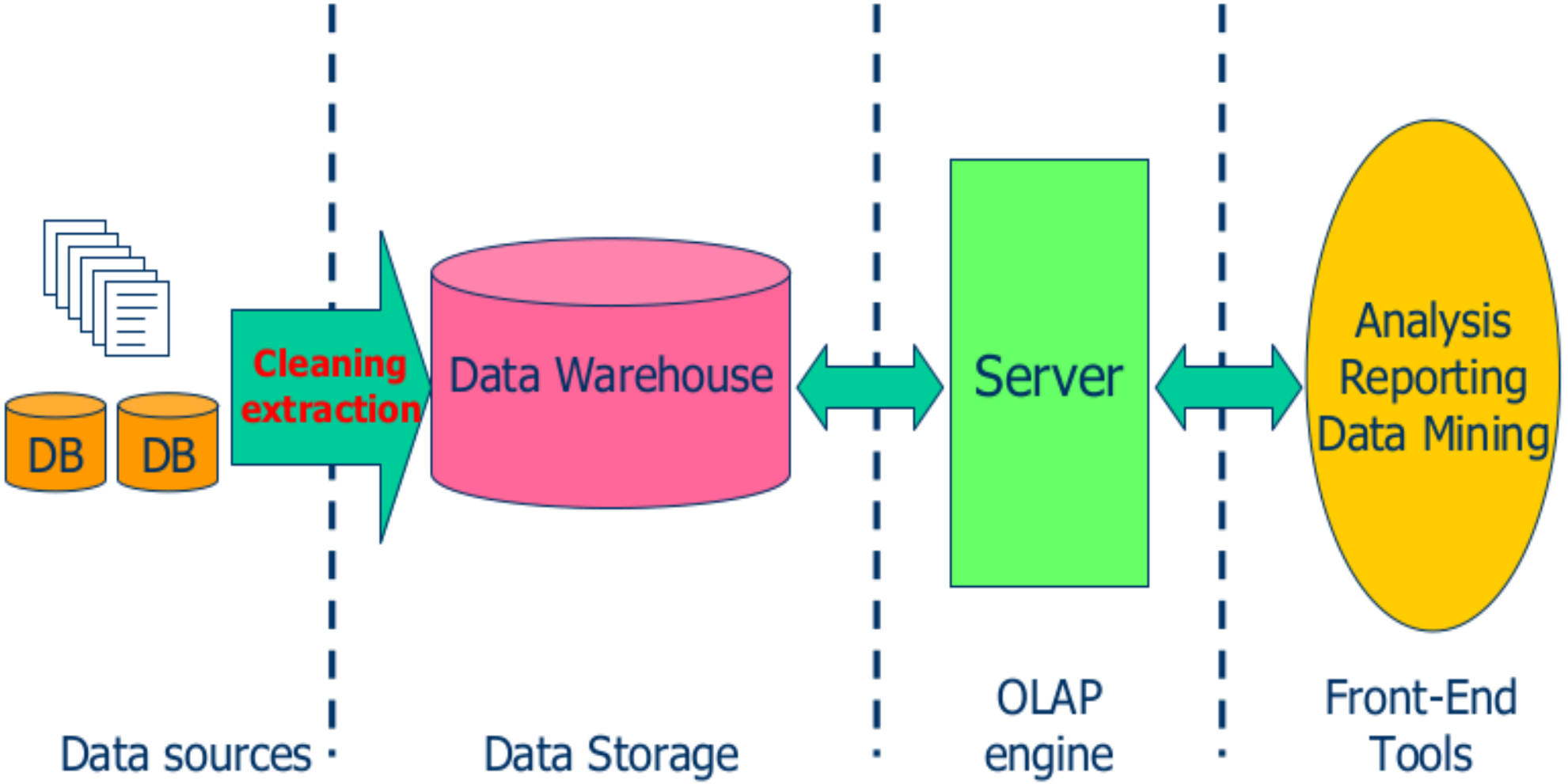
FOCUS on making yourself BETTER, not on thinking that you are better

Multidimensional model



FOCUS on making yourself BETTER, not on thinking that you are better

Multi-Tier Architecture



FOCUS on making yourself BETTER, not on thinking that you are better