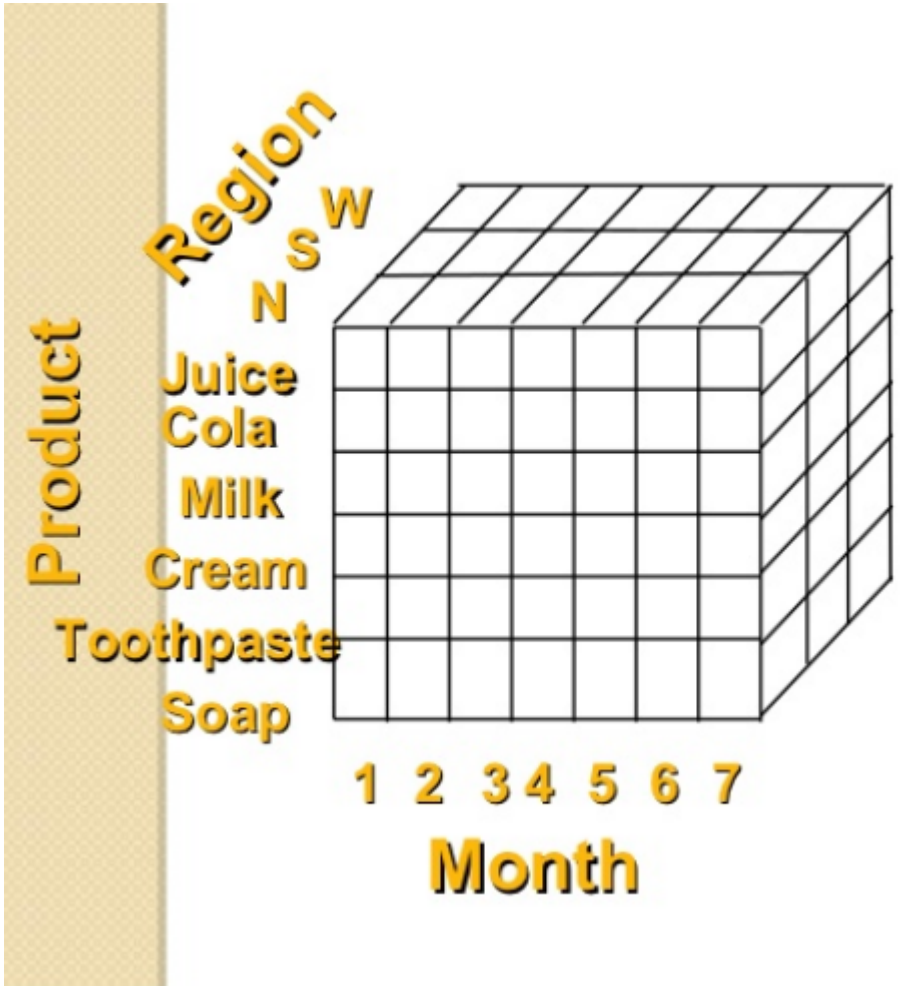
Unit I

- 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

Lattice of cube

- Mulidimensional data can be viewed as **lattice** of cubes
- C[A1,A2,...An] is the **base cuboid** (includes all data cells)
- **(n-1)-D cubes** are obtained
  - By grouping the cells
  - & computing combinationa numeric measures of a given dimension
- One cell will be the coarsest level – it consists of the measures of all the n dimensions
  - Known as **apex cuboid**



Dimensions: Product, Region, Time
Hierarchical summarization paths

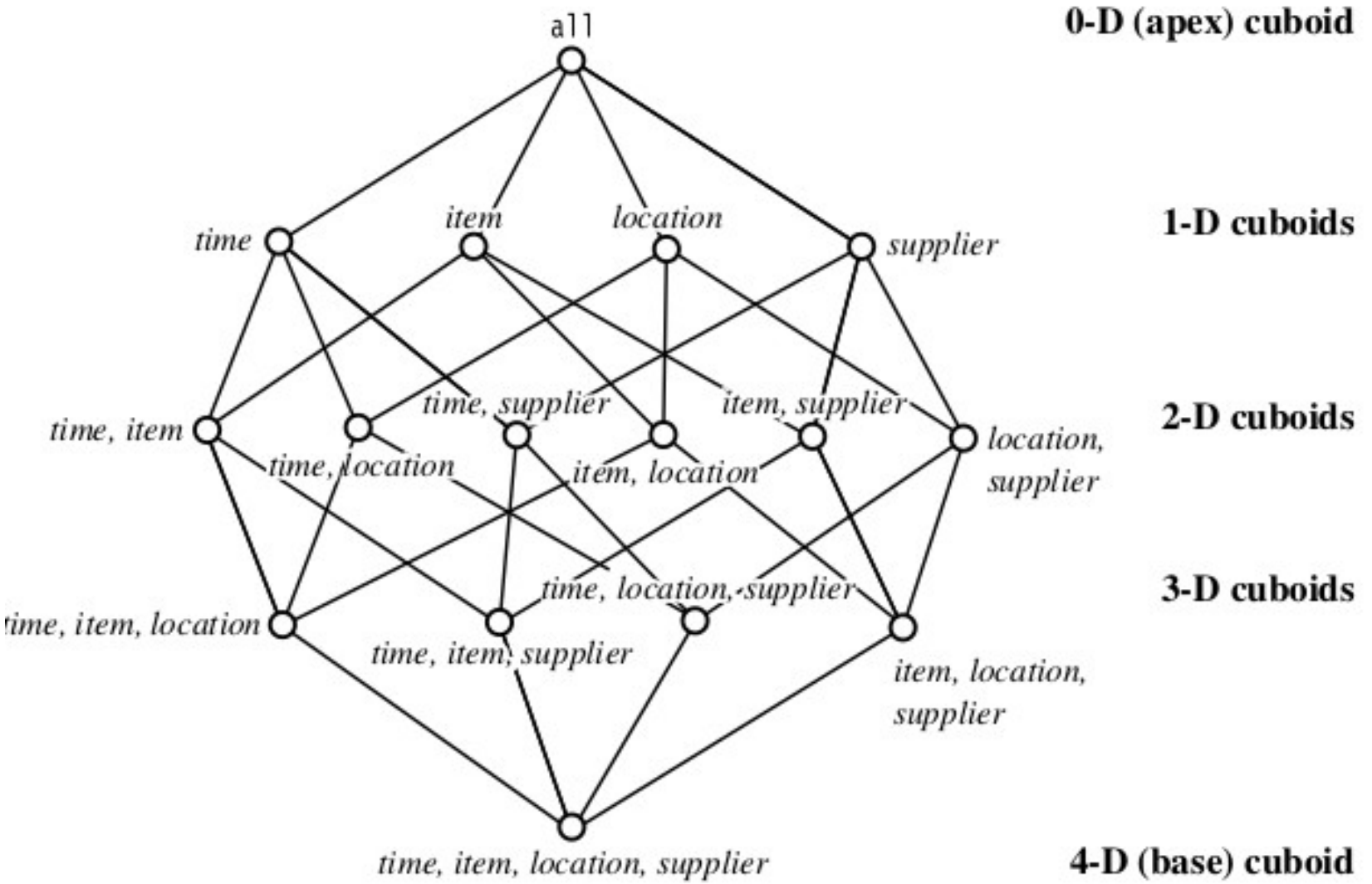
Product	Region	Time
Industry	Country	Year
Category	Region	Quarter
Product	City	Month, Week
	Office	Day

6

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

- Cube lattice
  - ABC
    AB AC BC
    A B C
    none
- Can materialize some groupbys, compute others on demand
- Question: which groupbys to materialze?
- Question: what indices to create
- Question: how to organize data (chunks, etc)

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.



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

- The conceptual components of MD data model
  - Summary measuers
  - Summary function
  - Dimension
  - Dimension hierarchy

- **Summary measure** is the main theme for this model
- **Measure** – aggregation of the values corresponding to the attribute/dimension
- **Catogorized** based on kind of aggregate function
  - **Distributive** – if it can be computed in a distributed manner. the data are partitioned into n sets. We apply the function to each partition, resulting in n aggregate values
    - sum,count,min,max
  - **Algebric** – if it can be computed by an algebraic function with m arguments (where m is a bounded positive integer), each of which is obtained by applying a distributive aggregate function.
    - average,standars deviation
  - **Holistic** – which cannot be described by an algebraic expression or calculated in a distributive manner. i.e., there does not exist an algebraic function with m arguments (where m is a constant) that characterizes the computation.
    - median,mode,frquency

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

Category	Examples
Distributive	Sum(), Count(), Minimum(), Maximum()
Algebraic	Average(), StandardDeviation(), MaxN() (N largest values), MinN() (N smallest values), CenterOfMass()
Holistic	Median(), MostFrequent(), Rank()

OLAP operations

- **Data warehouses** generalize and consolidate data in **multidimensional** space.
- The construction of data warehouses involves
  - data cleaning, data integration, and data transformation
  - These are important preprocessing step for data mining.
- DW provide **on-line analytical processing** (OLAP) tools for the interactive analysis of multidimensional data of varied granularities
- It facilitates effective data generalization and data mining.

OLAP operations

- Many other data mining functions,
  - such as association, classification, prediction, and clustering,
  - can be integrated with OLAP operations
- DW and OLAP form an essential step in the knowledge discovery process

OLAP operations

- **OLTP** systems
  - Are **on-line operational database systems** is to perform on-line transaction and query processing.
  - They cover most of the **day-to-day** operations of an organization
    - purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.
- **OLAP** systems
  - Data warehouse systems serve users or knowledge workers in the role of **data analysis and decision making**.
  - These can **organize and present data in various formats** in order to accommodate the diverse needs of the different users.

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

OLAP operations

- OLAP is used for exploring and analysing Data in the DW
- OLAP access live data and analyse on very large DB (also interactive data analysis)
- OLAP – user friendly envt for interactive data
- OLAP Server
  - is based on the multidimensional data model.
  - allows managers, and analysts to get an insight of the information
    - It gives fast, consistent, and interactive access to information.

OLAP vs OLTP

- **Users and system orientation:**
  - An **OLTP** system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals.
  - An **OLAP** system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.
- **Data contents:**
  - An **OLTP** system manages current data that, typically, are too detailed to be easily used for decision making.
  - An **OLAP** system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier to use in informed decision making.

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

OLAP vs OLTP

- **Database design:**
  - An **OLTP** system usually adopts an entity-relationship (ER) data model and an application-oriented database design.
  - An **OLAP** system typically adopts oriented database design.
- **View:**
  - An **OLTP** system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations.
  - **OLAP** system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.
- **Access patterns:**
  - The access patterns of an **OLTP** system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms.
  - accesses to **OLAP** systems are mostly read-only operations

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

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

OLAP operations

- For viewing the data with different perspectives we can use OLAP operations on the data cubes
- data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies
- OLAP operations
  - Are For Interactive querying & analysis to materialize different views of data
  - Are specialized data analysis methods

- OLAP operations:
  - Drilling
    - Drill-down & Drill-Up (Roll-up)
    - Drill-Within & Drill-Across
  - Slice
  - Dicing
  - Pivot (rotate)

- **Drilling :** moving up and down along hierarchies
- **Roll-up:**
  - The roll-up operation (drill-up) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.
  - e.g
    - city < state or state < country."
    - i.e., rather than grouping the data by city, the resulting cube groups the data by country.
- When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.
- For example, consider a sales data cube containing only the two dimensions location and time. Roll-up may be performed by removing, say,the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

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

- **Drill-down:**
  - Drill-down is the reverse of roll-up.
  - It navigates from less detailed data to more detailed data.
  - Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.
  - e.g
    - "day < month < quarter < year."
    - by descending the time hierarchy from the level of quarter to the more detailed level of month.
- Because a drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube.
- For example,
  - a drill-down on the e.g data cube can occur by introducing an additional dimension, such as customer group.

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

- **Slice and dice:**
  - The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.
  - e.g slice operation
    - the sales data are selected from the central cube for the dimension time using the criterion time = "Q1".
  - The dice operation defines a subcube by performing a selection on two or more dimensions.
  - e.g dice operation
    - selection criteria that involve three dimensions: (location = "Toronto" or "Vancouver") and (time = "Q1" or "Q2") and (item = "home entertainment" or "computer").

- **Pivot (rotate):**
  - Pivot (also called rotate) is a **visualization operation** that rotates the data axes in view in order to provide an **alternative presentation** of the data.
  - e.g pivot operation
    - item and location axes in a 2-D slice are rotated.
  - Other e.g
    - rotating the axes in a 3-D cube,
    - transforming a 3-D cube into a series of 2-D planes.

- **Other OLAP operations:**
  - Some OLAP systems offer additional drilling operations.
  - drill-across executes queries involving (i.e., across) more than one fact table. i.e., Switching from a classfication in one dimension to a different classification in different dimension
  - drill-through operation uses relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.Switching from a classfication in one dimension to a different classification in the same dimension
  - ranking the top N or bottom N items in lists,
  - computing moving averages, growth rates, interests, internal rates of return, depreciation, currency conversions, and statistical functions.

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

DW Schema

- **RDBMS**
  - database schema consists of a set of entities and the relationships between them.
  - The E-R data model is commonly used in the design of relational databases
  - It is appropriate for OLTP.
- **DW**
  - It requires a concise, subject-oriented schema that facilitates OLAP
  - The most popular data model for a data warehouse is a multidimensional model.
  - Differnt forms of MD model
    - star schema, a snowflake schema, or a fact constellation schema.

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

- **Star schema:**
  - The most common modeling paradigm is the star schema
  - In this schema the data warehouse contains
    - a large central table (fact table) containing the bulk of the data, with no redundancy,
    - a set of smaller attendant tables (dimension tables), one for each dimension.
  - The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

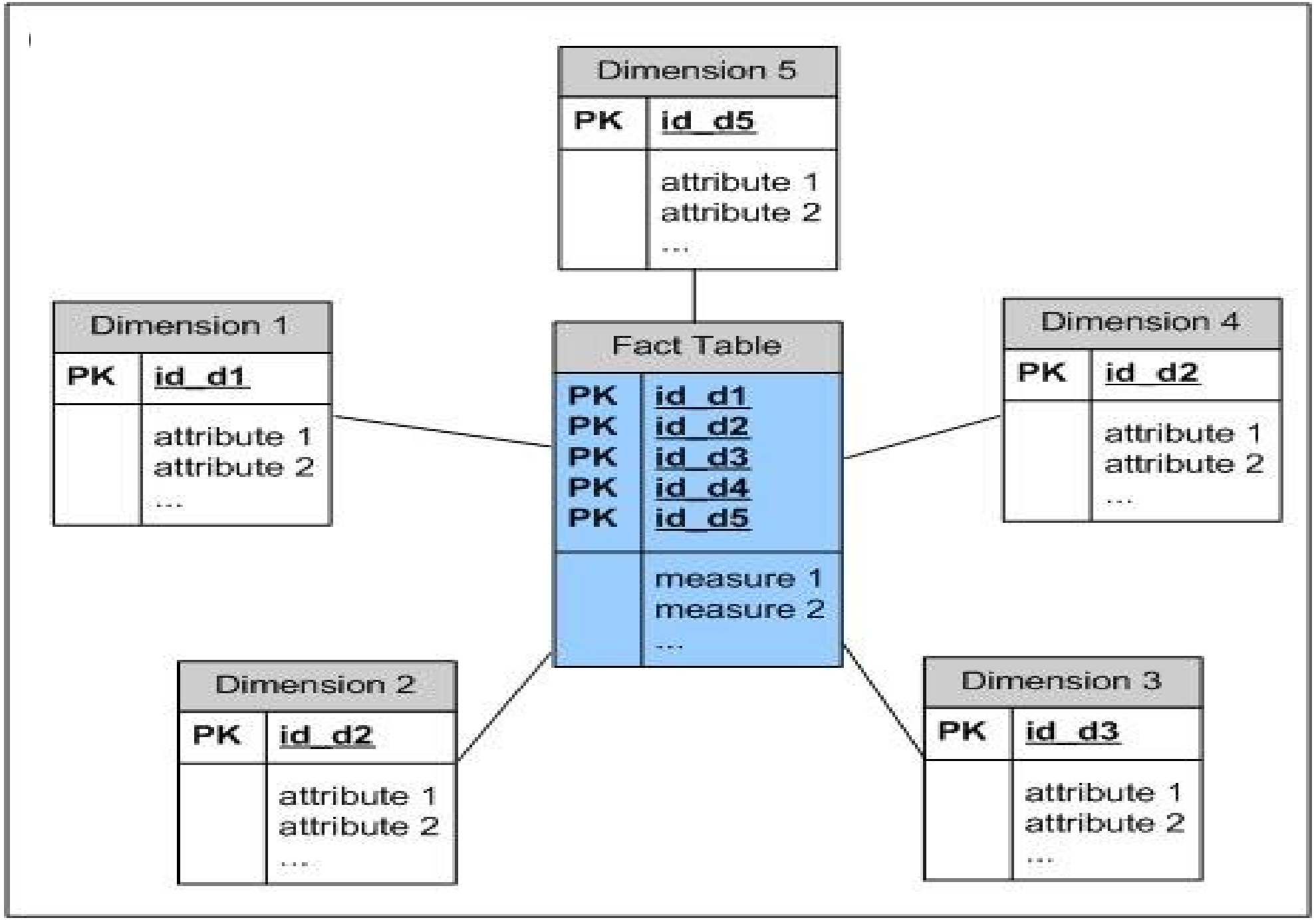
- star schema resembles a star, with points radiating from a center.
- The center of the star consists of fact table and the points of the star are the dimension tables.
- Usually the
  - fact tables in a star schema are in third normal form(3NF)
  - dimensional tables are de-normalized.

- **Fact Tables**
  - It contains detailed summary data i.e., fact's data on detail or aggregated level.
  - A fact table typically has two types of columns:
    - foreign keys to dimension tables
    - and measures those that contain numeric facts.
  - Its primary key has One key per dimension
- **Dimension Tables**
  - A dimension is a structure usually composed of one or more hierarchies that categorizes data.
  - If a dimension hasn't got a hierarchies and levels it is called flat dimension or list.
  - The primary keys of each of the dimension tables are part of the composite primary key of the fact table.
  - Dimensional attributes describe the dimensional value, that are normally descriptive, textual values.
  - Dimension tables are generally small in size than fact table.

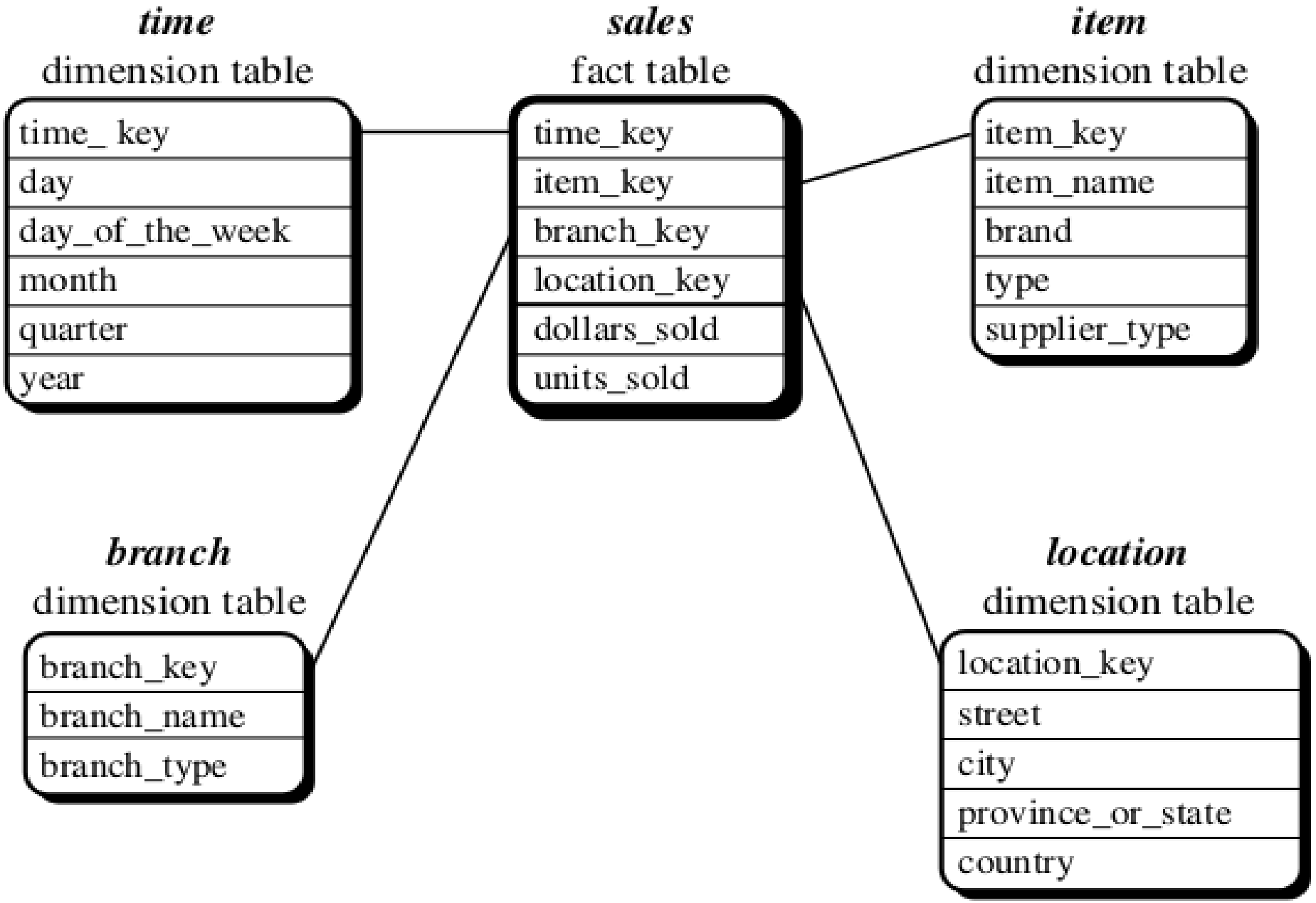
FOCUS on making yourself BETTER, not on thinking that you are better

- The main characteristics of star schema:
  - Simple structure : easy to understand schema
  - Great query effectives : small number of tables to join
  - Relatively long time of loading data into dimension tables :
    - de-normalization, redundancy data caused that size of the table could be large.
  - The most commonly used DW implementations
  - Widely supported by a large number of business intelligence tools

- **Relationship**
  - 1 tuple in the FT corresponds to one and only one tuple in each DT
  - 1 tuple in the DT corresponds to more than one tuple in each DT
  - So 1:N relationship b/w FT and DT



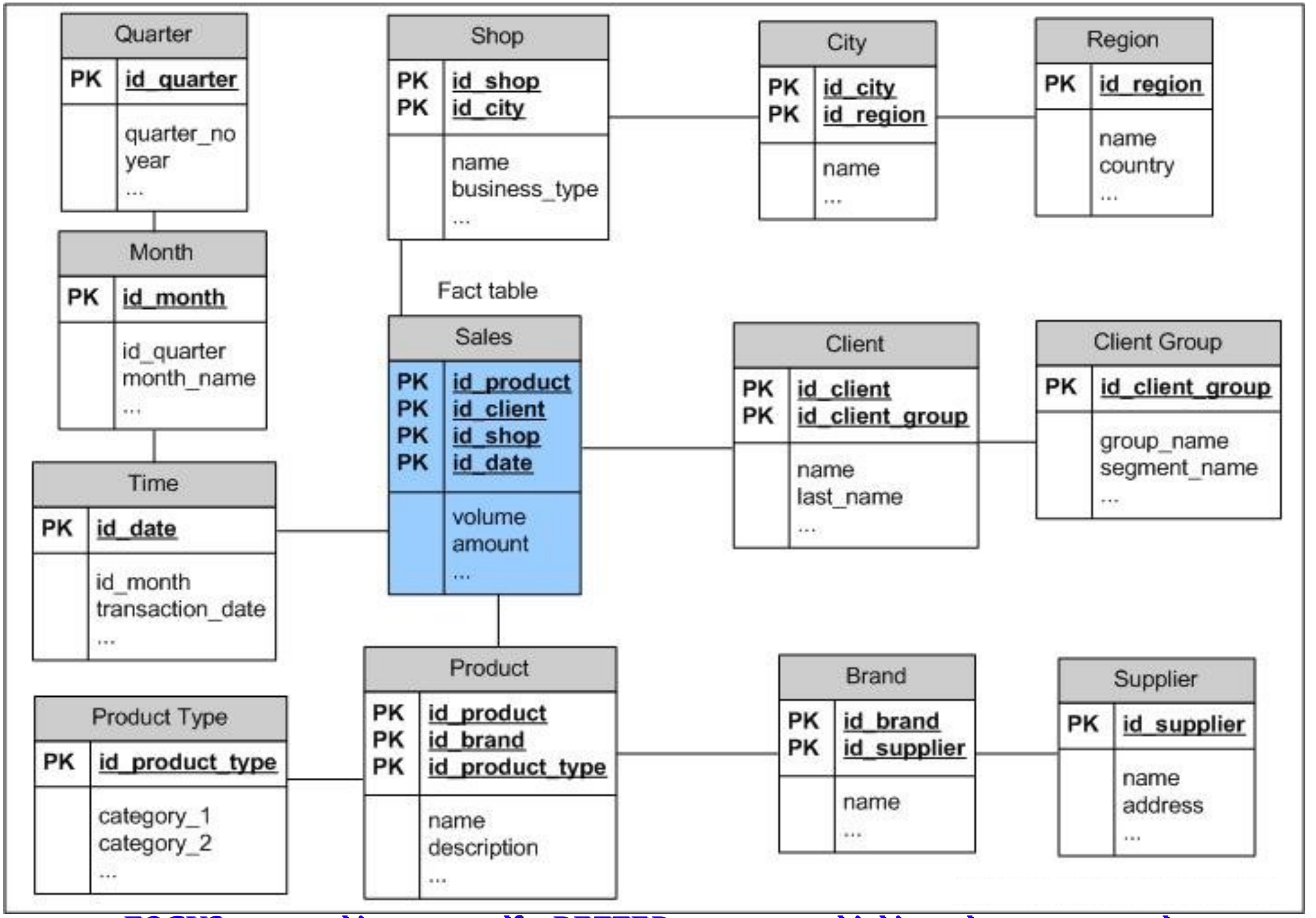
FOCUS on making yourself BETTER, not on thinking that you are better



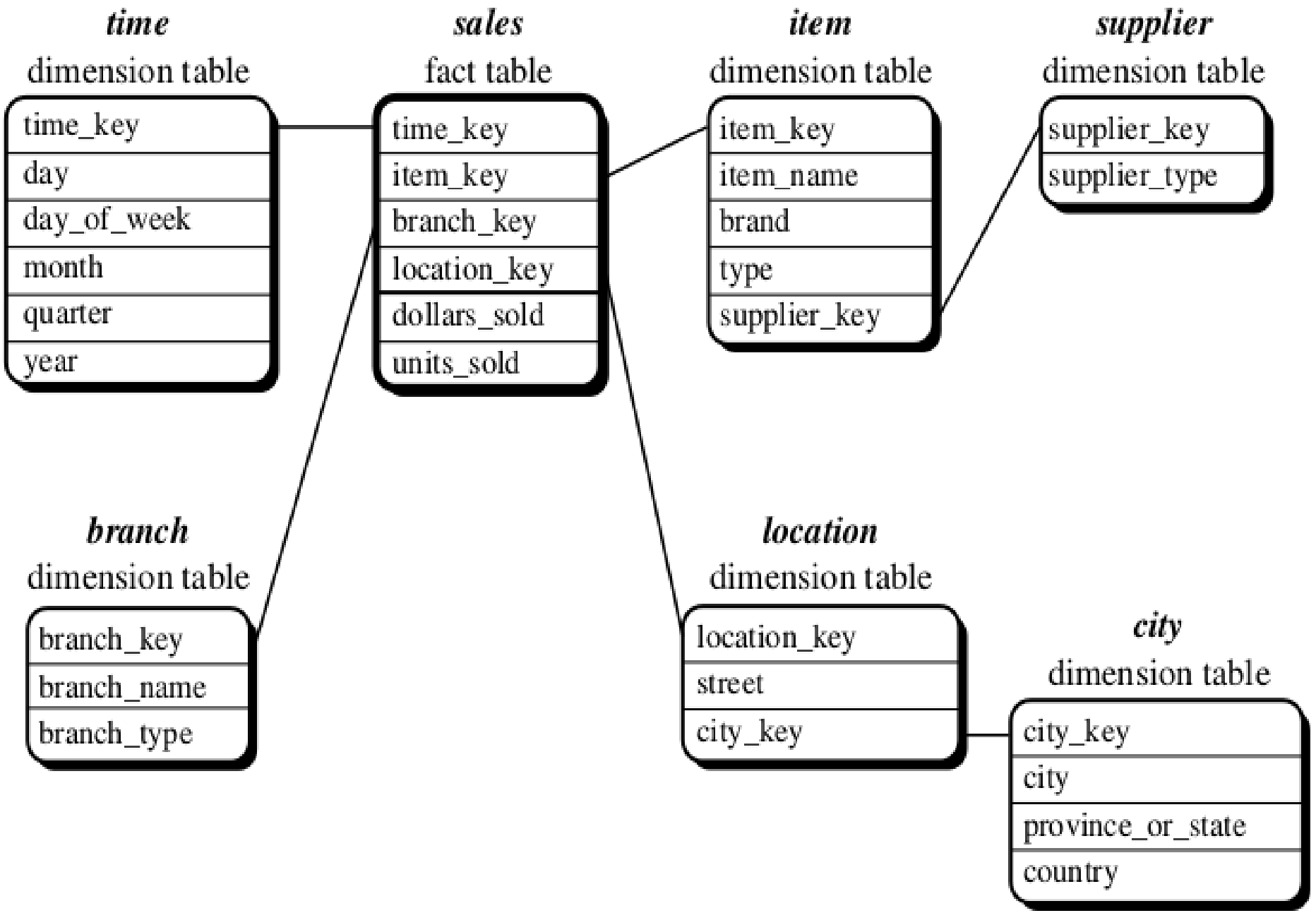
FOCUS on making yourself BETTER, not on thinking that you are better

- **Snowflake schema**
  - In The snowflake schema some dimension tables are normalized,
  - thereby further splitting the data into additional tables.
  - The resulting schema graph forms a shape similar to a snowflake.
  - The major difference between the snowflake and star schema models is that
    - the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.

- **Snowflake schema**
  - Normalized table is easy to maintain and saves storage space.
  - But , this saving of space is negligible in comparison to the typical magnitude of the fact table.
- **Disadvs**
  - The snowflake structure can reduce the effectiveness of browsing, since more joins will be needed to execute a query.
  - The system performance may be adversely impacted.
  - It is not as popular as the star schema in data warehouse design.



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



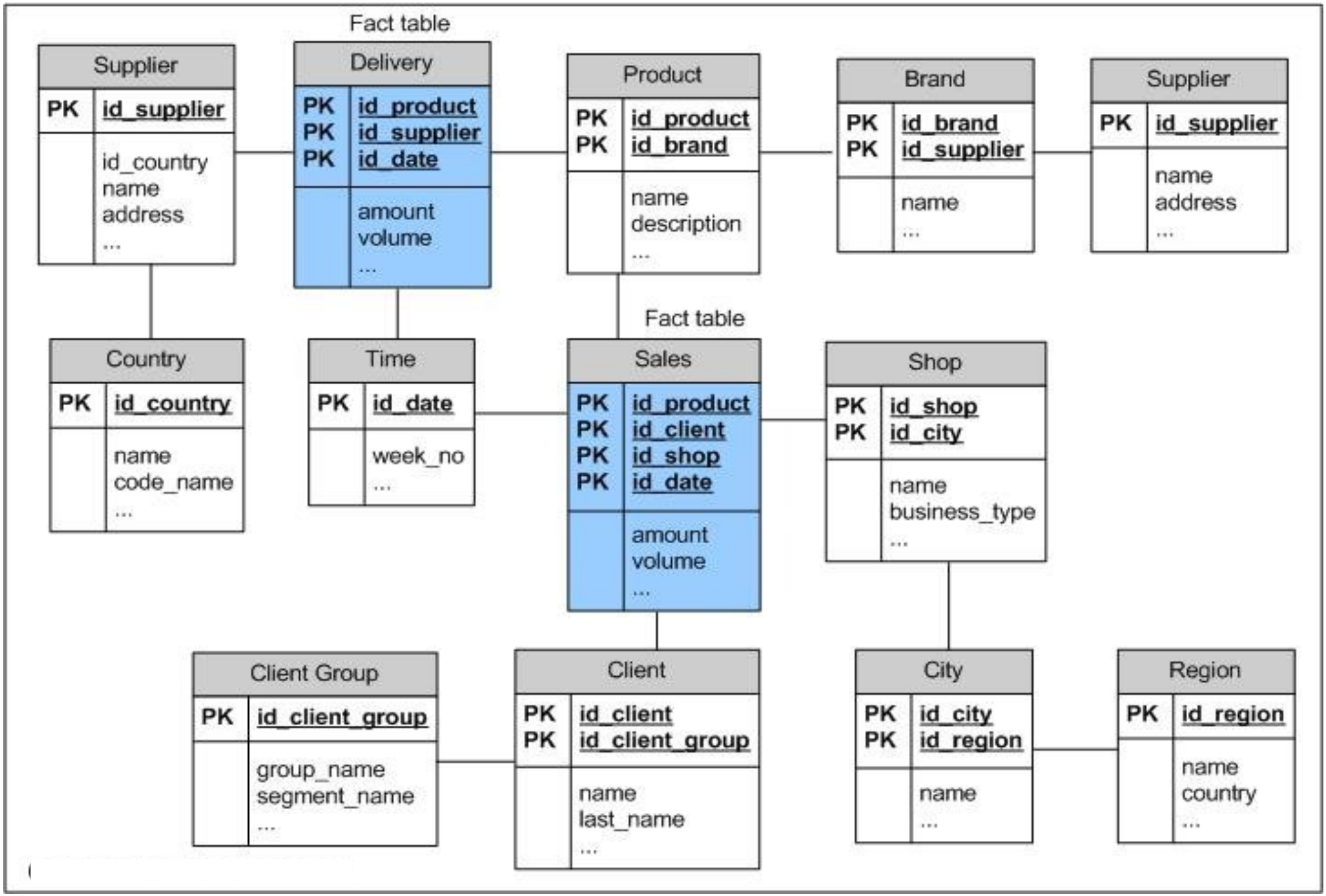
FOCUS on making yourself BETTER, not on thinking that you are better

- **Fact constellation:**
- Sophisticated applications may require **multiple fact tables** to share dimension tables.
- i.e., by **splitting the original star schema** into more star schemes each of them describes facts on another level of dimension hierarchies
- This kind of schema can be viewed as a collection of stars, and hence is called a **galaxy schema** or a **fact constellation.**

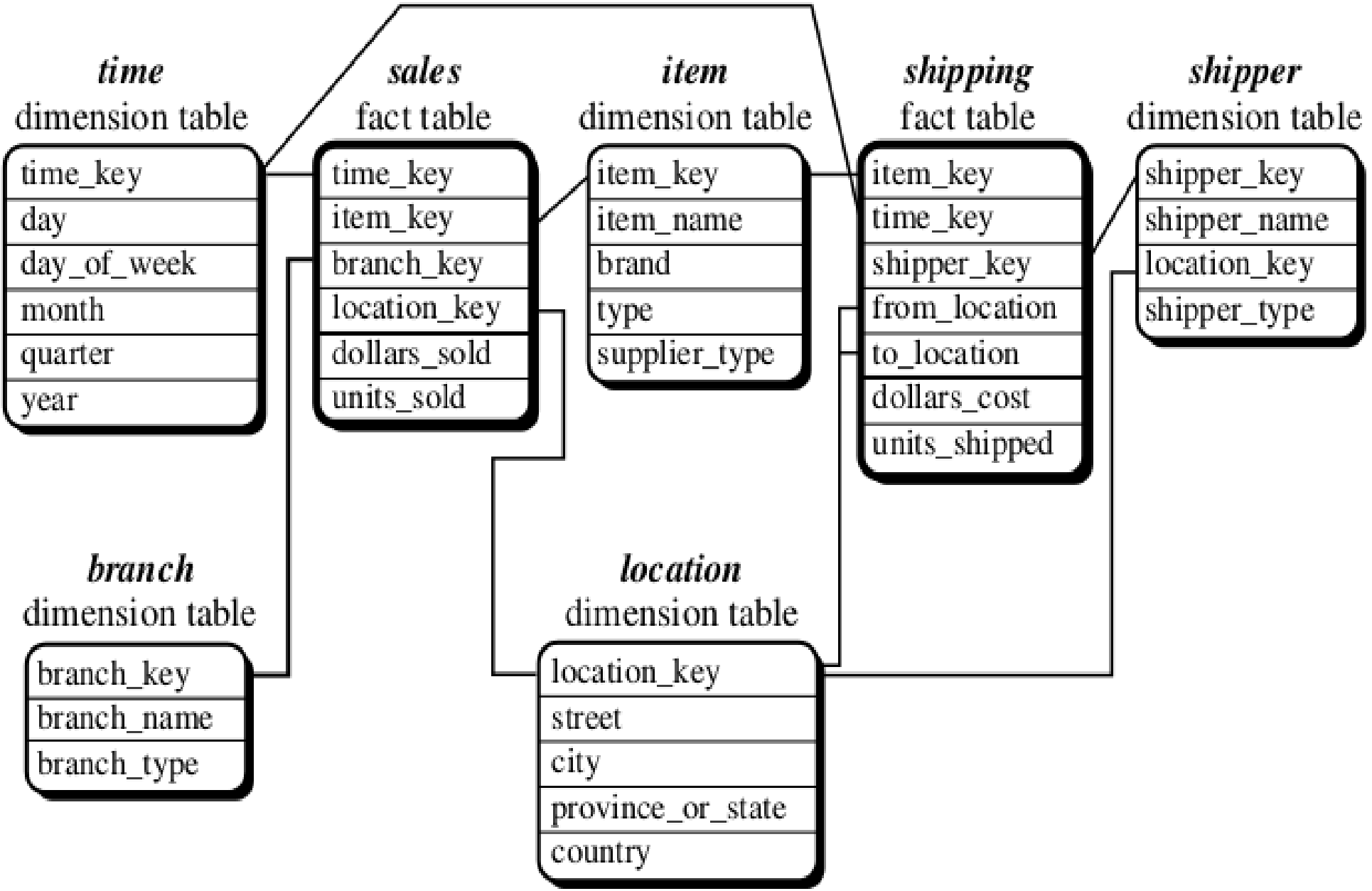
FOCUS on making yourself BETTER, not on thinking that you are better

- For each star schema it is possible to construct fact constellation schema
- Fact table share the same Dimension tables
- Disadvs :
  - more complicated design
  - many variants for particular kinds of aggregation must be considered and selected.
  - dimension tables are still large.

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

- **Types of OLAP Servers/Engines**
- We have four types of OLAP servers:
  - Relational OLAP (ROLAP)
  - Multidimensional OLAP (MOLAP)
  - Hybrid OLAP (HOLAP)
  - Specialized SQL Servers

- **Relational OLAP server**
  - Where data need not be stored as MD view
  - ROLAP servers are placed between relational back-end server and client front-end tools.
  - To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS(i.e., a sacalable,parallel and relational DB is used – provides storage and high speed access)
  - Middle tier
    - Provides MD conceptual view of data
    - Analytical functionality (that are not in RDBMS)
  - Presentation tier
    - Delivers the result

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

- ROLAP includes the following:
  - Full analytical functionality by maintaining features of relational data
  - Implementation of aggregation navigation logic.
  - Optimization for each DBMS back end.
    - It creates optimized SQL statements and send it to RDBMS server
    - Get the result from RDMMS server ,reintegerates,analyse and deliver to the user
- Disadvs
  - Its technology is limited
    - Bcos of non-integerated,disparate tier architecture
    - Data is separate from the analytic processing
    - Limts the scope of analysis

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

- 2 features of ROLAP
  - DW and RDB are inseparable
  - If dimension strucure changes the RDB should also be physically reorganized
    - So on the fly MD view of the ROLAP tool is appropriate

- **Multidimensional OLAP**
  - MOLAP server is for DW that uses MDData model
  - MOLAP uses array-based multidimensional storage engines for multidimensional views of data.
  - i.e., mapping MD view of data cube to array structure
  - With multidimensional data stores, the storage utilization may be low if the data set is sparse.
  - Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

- **Advs**
  - Fast indexing to precompute summarized data
  - It is recommended for MD data stores

- MOLAP over ROLAP
  - Relational tables are unnatural for MD data
  - MD array privide efficiency in storage and operations
  - Mismatch b/w MD operationsla and SQL
  - ROLAP to perform outside current relational system to achieve performance

- ROLAP over MOLAP
  - ROLAP integerates with exixting RDB rtechnology and standards
  - MOLPA does not support adhoc queries
  - Updating MOLAP is difficult as data has to be downloaded in MOLAP
  - Encoding and compression can be easily done in ROLAP
  - ROLAP also uses parallel relational technology

- **Hybrid OLAP (HOLAP)**
  - Hybrid OLAP is a combination of both ROLAP and MOLAP.
  - It offers higher scalability of ROLAP and faster computation of MOLAP.
  - HOLAP servers allows to store the large data volumes of detailed information.
  - The aggregations are stored separately in MOLAP store.

- **Specialized SQL Servers**
  - **Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.**