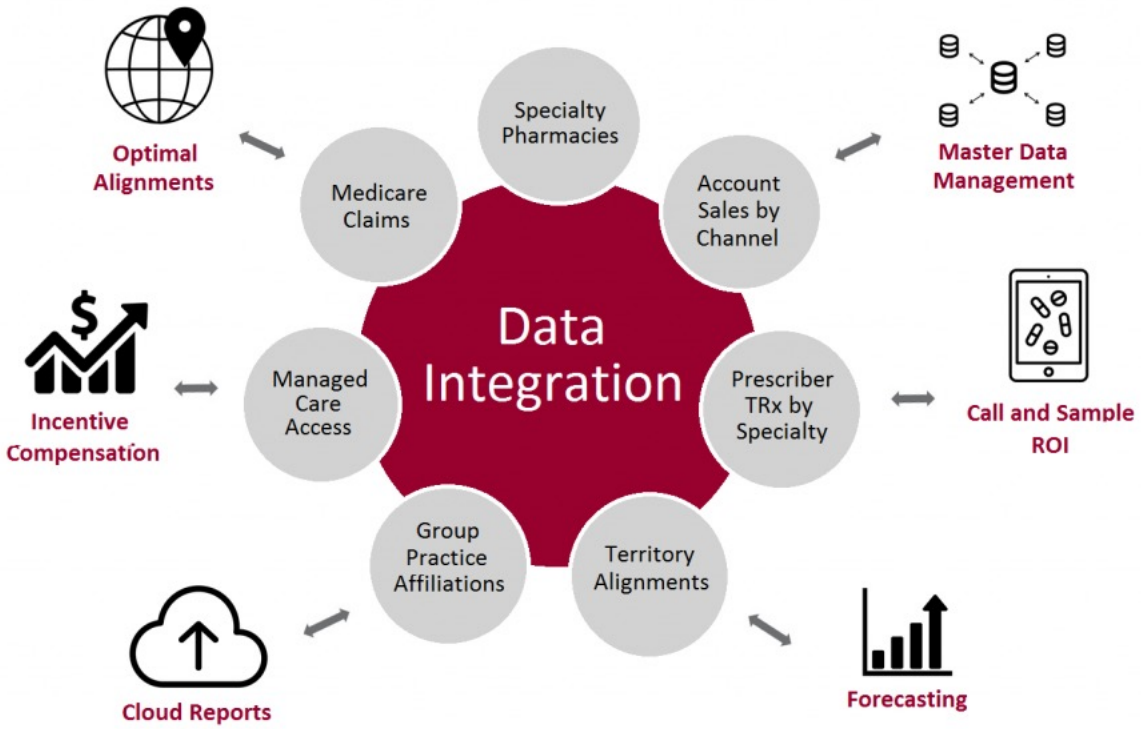
Chapter 3
Data modelling and databases
OLTP & OLAP

Outline

- Overview
- OLTP vs OLAP
- Data warehouse modeling
- Data warehouse design

Heterogeneous data sources



Why data integration

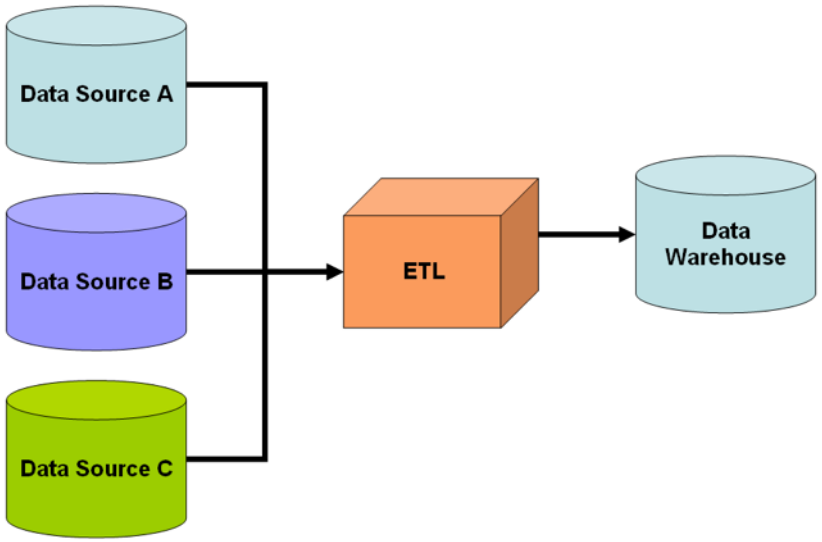
- To facilitate information access and reuse through **a single information access point**
- Data from different complementing information systems is **to be combined to gain a more comprehensive basis** to satisfy the need
  - Improve decision making
  - Improve customer experience
  - Increase competitiveness, Streamline operations
  - Increase productivity
  - Predict the future

Data integration challenges

- Physical systems
  - Various hardwares, standards
  - Distributed deployment
  - Various data format
- Logical structures
  - Different data models
  - Different data schemas
- Business organization
  - Data security and privacy
  - Business rules and requirements
  - Different administrative zones in the business organization

Data Warehouse

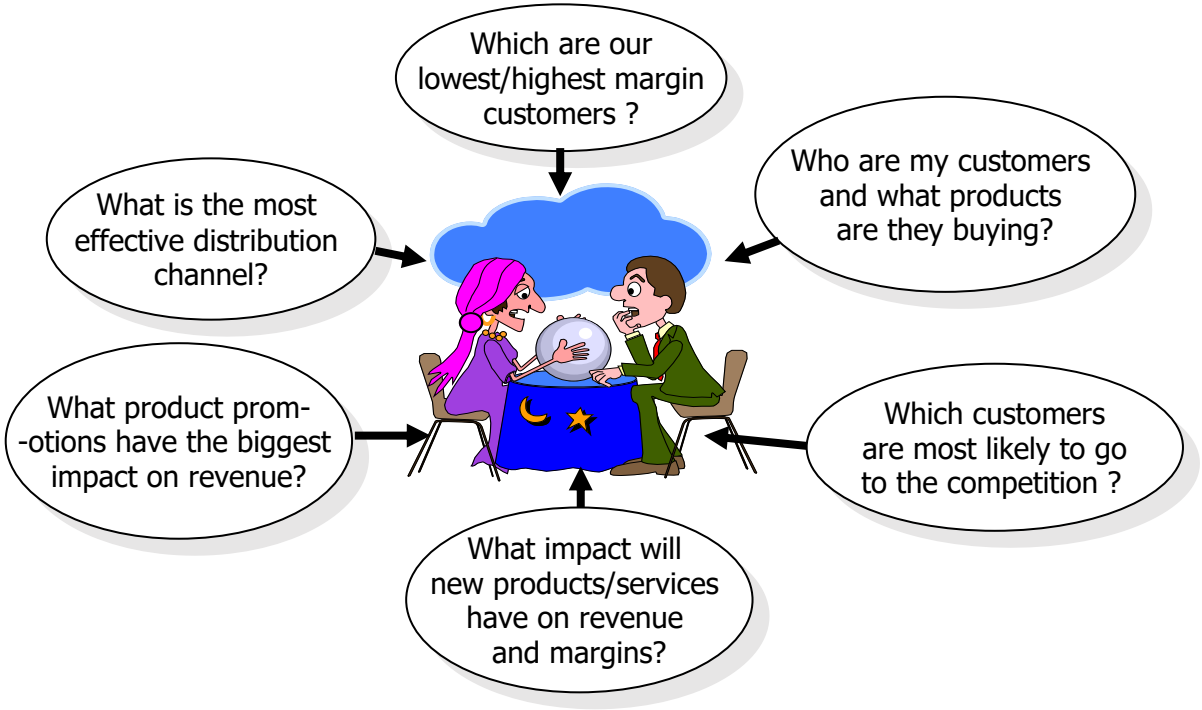
- A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context. [Barry Devlin]
- A data warehouse is a copy of transaction data specifically structured for query and analysis [Ralph Kimball]
- Data from several operational sources (OLTP) are extracted, transformed, and loaded (ETL) into a data warehouse



Data Warehouse usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Data Warehouse usage



Advantages

- High query performance
  - But not necessary most current information
- Does not interfere with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources

Characteristics of Data warehouse

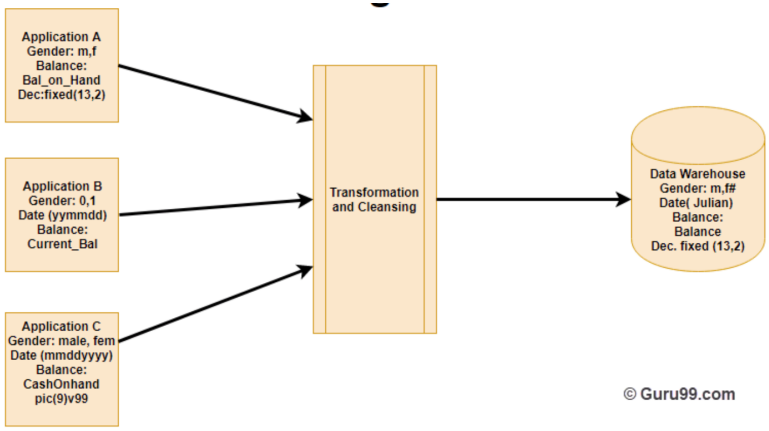
- Subject-Oriented
- Integrated
- Time-variant
- Non-volatile

Subject-Oriented

- Offer information regarding a theme instead of companies' ongoing operations
  - Subjects can be sales, marketing, distributions, etc.
  - A data warehouse never focuses on the ongoing operations
- Emphasis on modeling and analysis of data for **decision making**
  - Provide a simple and concise view around the specific subject by excluding data which not helpful

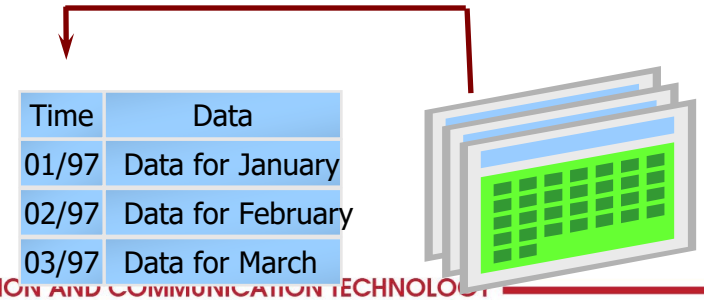
Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - Data needs to be stored in the Datawarehouse in a common and universally acceptable manner
  - This integration helps in effective analysis of data
- 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

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- 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”
- Data warehouse is loaded daily, hourly, or on some other periodic basis, and does not change within that period

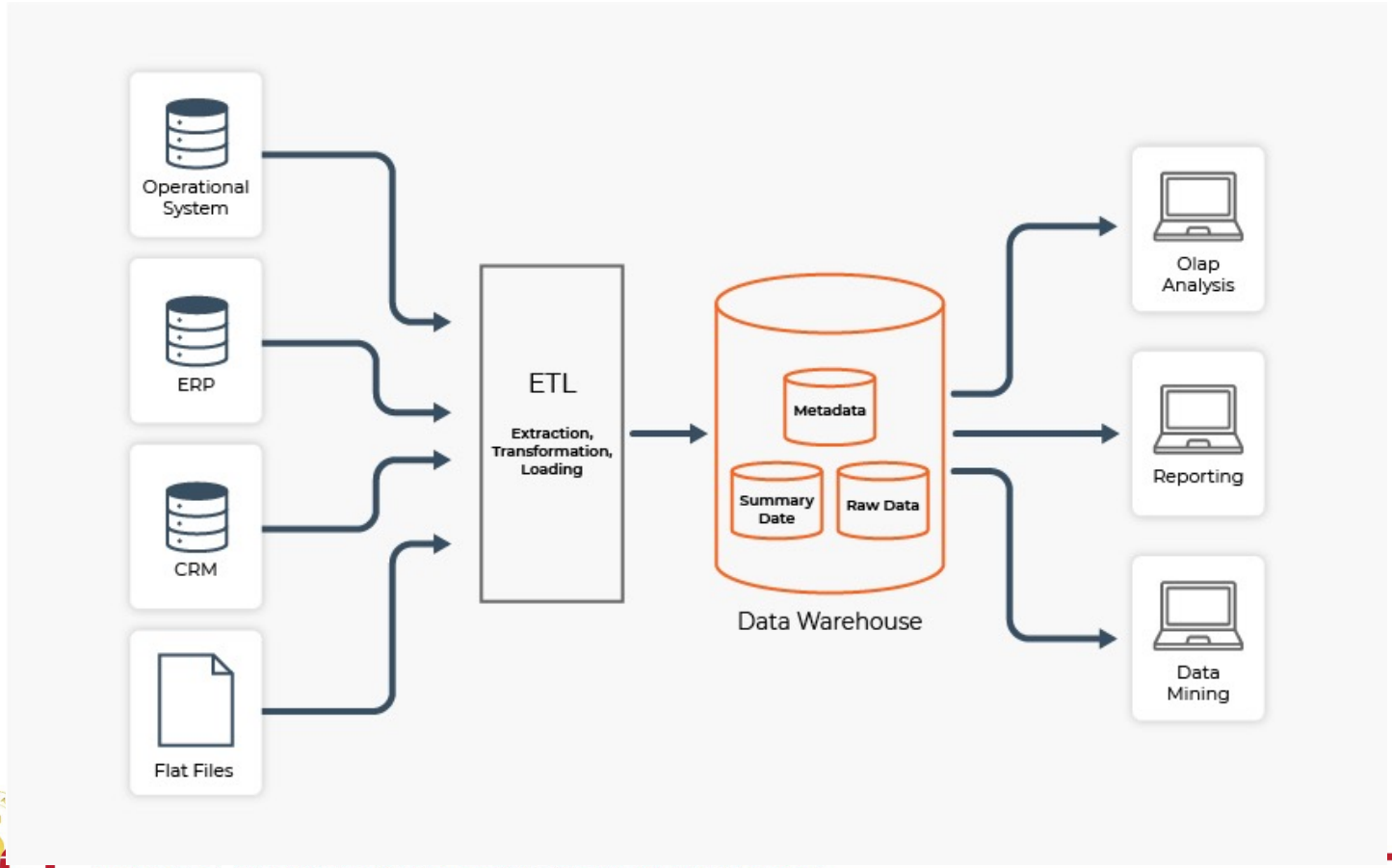


Non-volatile

- Historical data in a data warehouse should never be altered
  - Helps to analyze historical data and understand what & when happened
- Data is read-only
  - Does not require transaction process, recovery and concurrency control mechanisms
  - Delete, update, and insert are omitted
- Only two types of data operations
  - Data loading
  - Data access (reading)

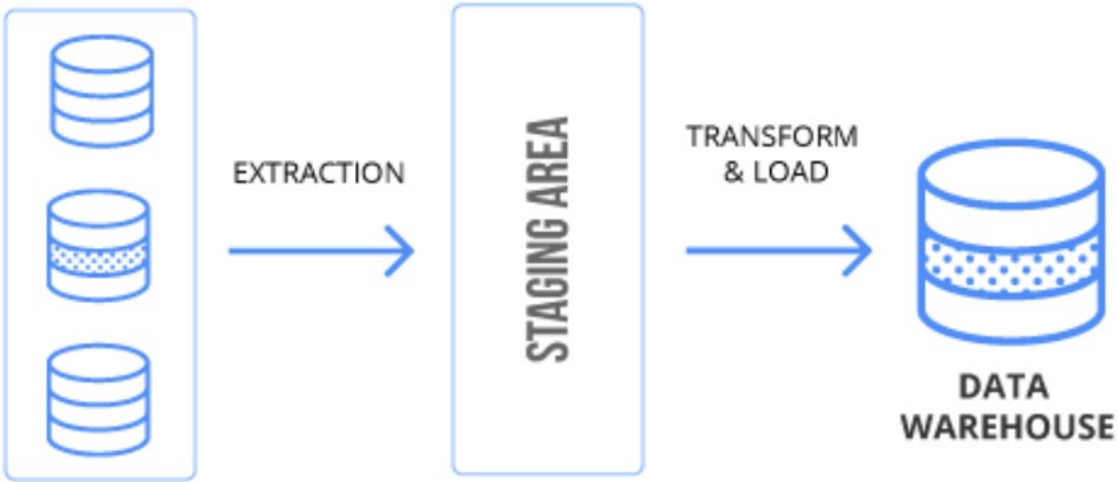
Data Warehousing

- A process for assembling and managing data from various sources for the purpose of answering business questions



ETL

- Extract
  - Get the data from source system as efficiently as possible
- Transform
  - Perform calculations on data
- Load
  - Load the data in the target storage



Staging area

- An intermediate storage area used for data processing during the extract, transform and load (ETL) process
- Mainly required for timing reasons (optional)
  - If it is not feasible to extract all the data from all Operational databases at exactly the same time
- Objectives
  - **Consolidation:** as a large "bucket" in which data from multiple source systems can be temporarily placed for further processing
  - **Alignment:** standardization of reference data, validation of relationships between records and data elements from different sources
  - **Minimizing contention:** efficient data transfer from sources
  - **Independent scheduling**
  - **Change detection**
  - **Cleansing data**

Why is ETL (System) Important?

- Adds **value** to data
  - Removes mistakes and corrects data
  - Documented measures of confidence in data
  - Captures the flow of transactional data
  - Adjusts data from multiple sources to be used together (conforming)
  - Structures data to be usable by BI tools
  - Enables subsequent business / analytical data procesing

Metadata Repository

- Metadata is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

OLTP vs OLAP

OLTP (Online Transaction processing)

- OLTP is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE).
- The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second.
- In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

OLAP (Online analytical processing)

- Is characterized by relatively low volume of transactions.
- Queries are often very complex and involve aggregations.
- For OLAP systems a response time is an effectiveness measure.
- OLAP applications are widely used by Data Mining techniques.
- In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

OLTP vs OLAP

- targets one specific process
- Many short transactions (queries + updates)
- Examples
  - Update account balance
  - Enroll in course
  - Add book to shopping cart
- Queries touch a small amounts of data (one record or a few records)
- Updates are frequent
- Concurrency is biggest performance concern

- integrates data from different processes
- often makes use of historical data
- Long transactions, complex, ad hoc queries
- Examples
  - Report total sales for each department in each month
  - Identify top selling books
- Queries touch large amounts of data
- Updates are infrequent
- Individual queries can require lots of resources

OLTP vs. OLAP

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

OLAP & OLTP: Different performance requirements

- Transaction processing (OLTP)
  - Fast response time important
  - Data must be up-to-date, consistent at all times
- Data analysis (OLAP)
  - Queries can consume lots of resources
  - Operating on static “snapshot” of data
- OLAP queries would degrade operational DB
  - Analysis query asks for sum of all sales
  - Acquires lock on sales table
  - New sales transaction is blocked

OLAP & OLTP: Different data modeling requirements

- OLTP
  - Normalized schema for consistency
  - Complex data models, many tables
  - Limited number of standardized queries and updates
- OLAP
  - Simplicity of data model is important
  - De-normalized schemas are common
    - Fewer joins -> improved query performance
    - Fewer tables -> schema is easier to understand

Data Warehouse Modeling

Data Cube and OLAP models and operations

Multi-dimensional data model

- A data warehouse is based on a multidimensional data model, which views data in the form of a data cube.
  - optimized for very quick data analysis
- A data cube allows data to be modeled and viewed in multiple dimensions.
  - **Fact tables** contain measures of interest (such as dollars sold) and keys to each of the related dimension tables.
  - **Dimension tables** provide the context of the measures such as item (item name, brand), product, location or time(day, week, month, quarter, year).

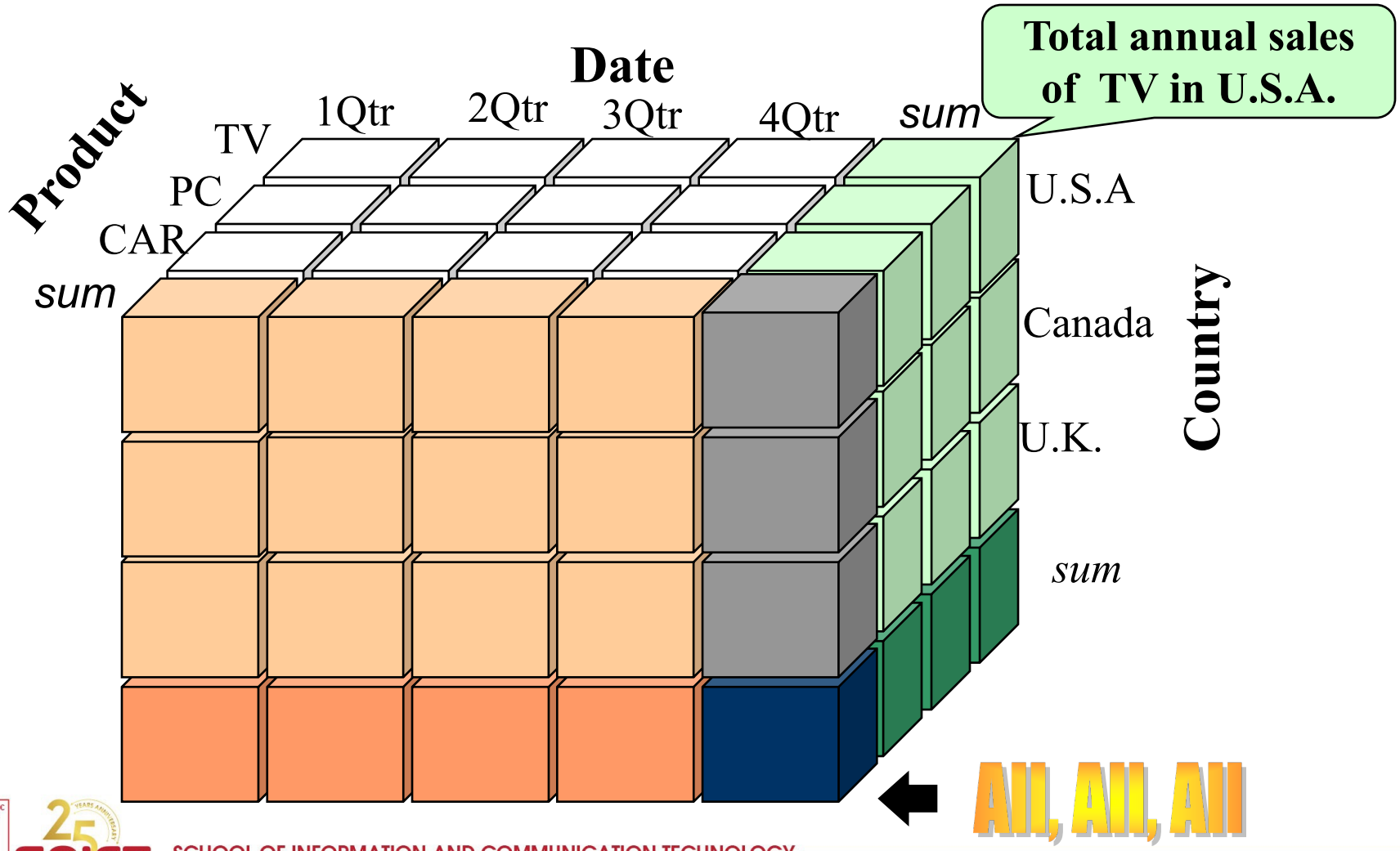
Multi-dimensional data representation

- Relational table only represents multi-dimensional data in two dimensions

ID	Product	Country	Date	Sales
1	TV	US	1Qtr	100
2	PC	Canada	4Qtr	500
3	CAR	US	2Qtr	30
4	PC	UK	3Qtr	200
5	CAR	UK	1Qtr	20
6	CAR	UK	2Qtr	15
7	TV	Canada	4Qtr	80

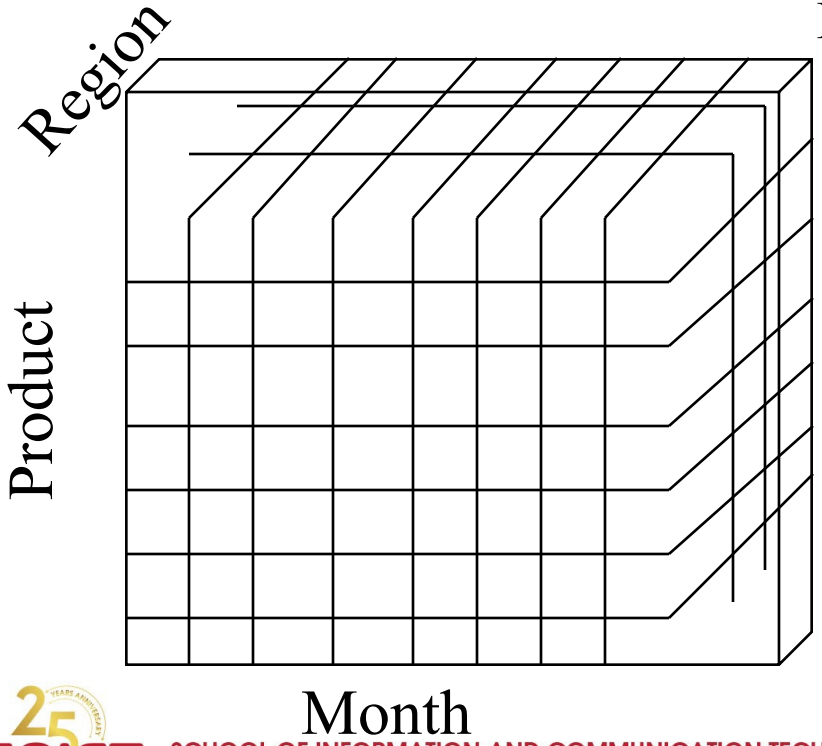
- Cube represents data as cells in an array
  - Each side of cube is a dimension

From Tables to Data Cubes



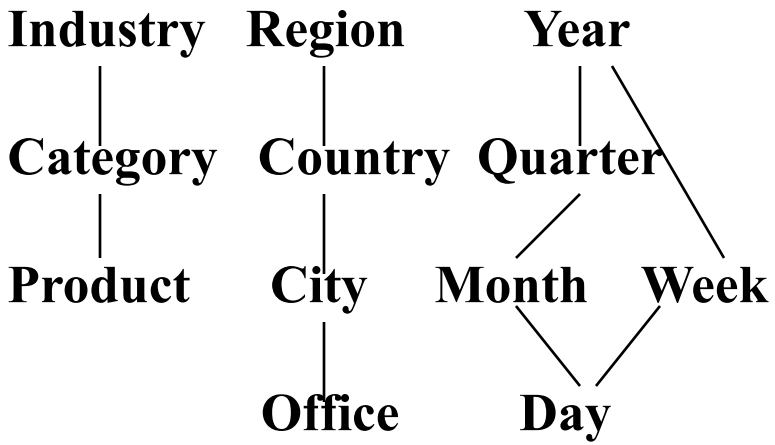
Multidimensional Data

- Sales volume as a function of product, month, and region
- For each dimension, the set of values can be organized in a hierachy

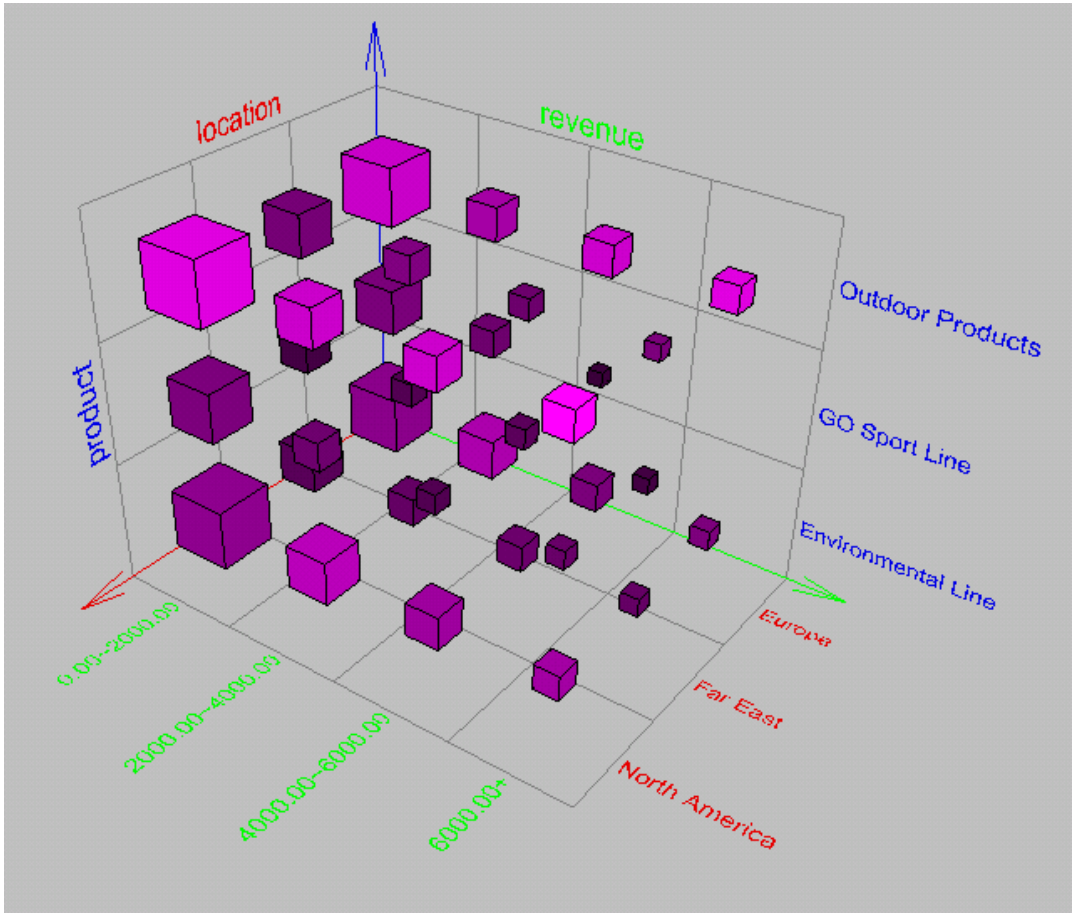


Dimensions: *Product, Location, Time*

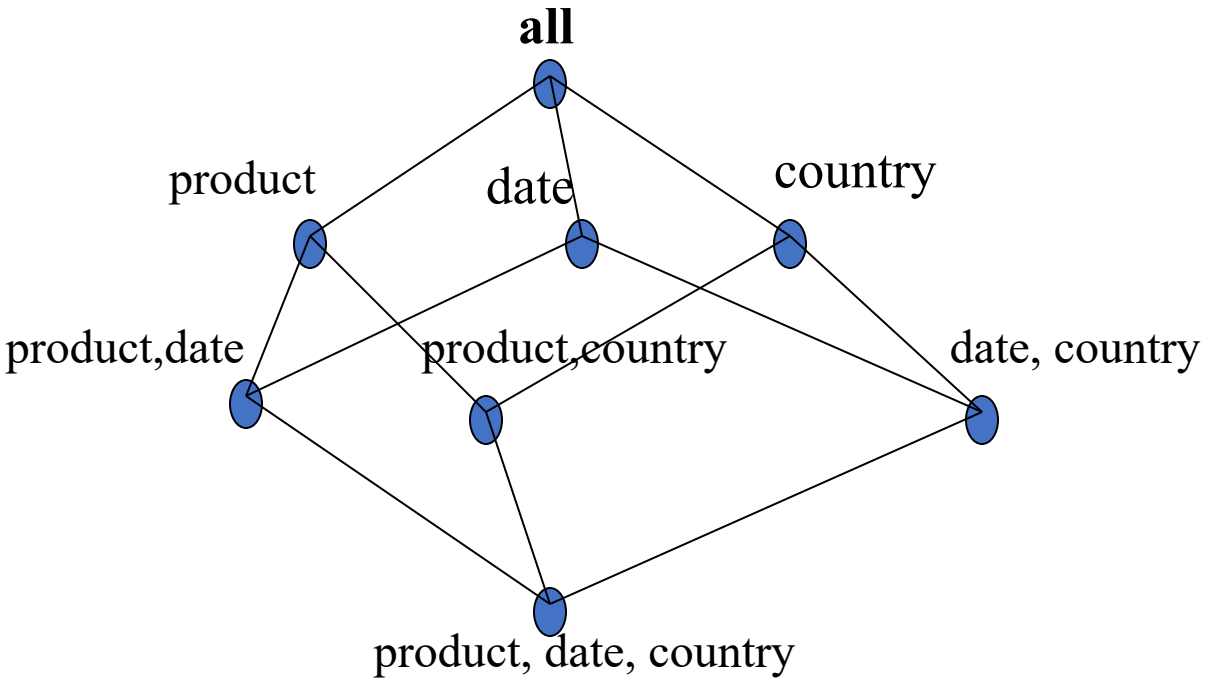
Hierarchical summarization paths



Browsing a Data Cube



Cube: A Lattice of Cuboids



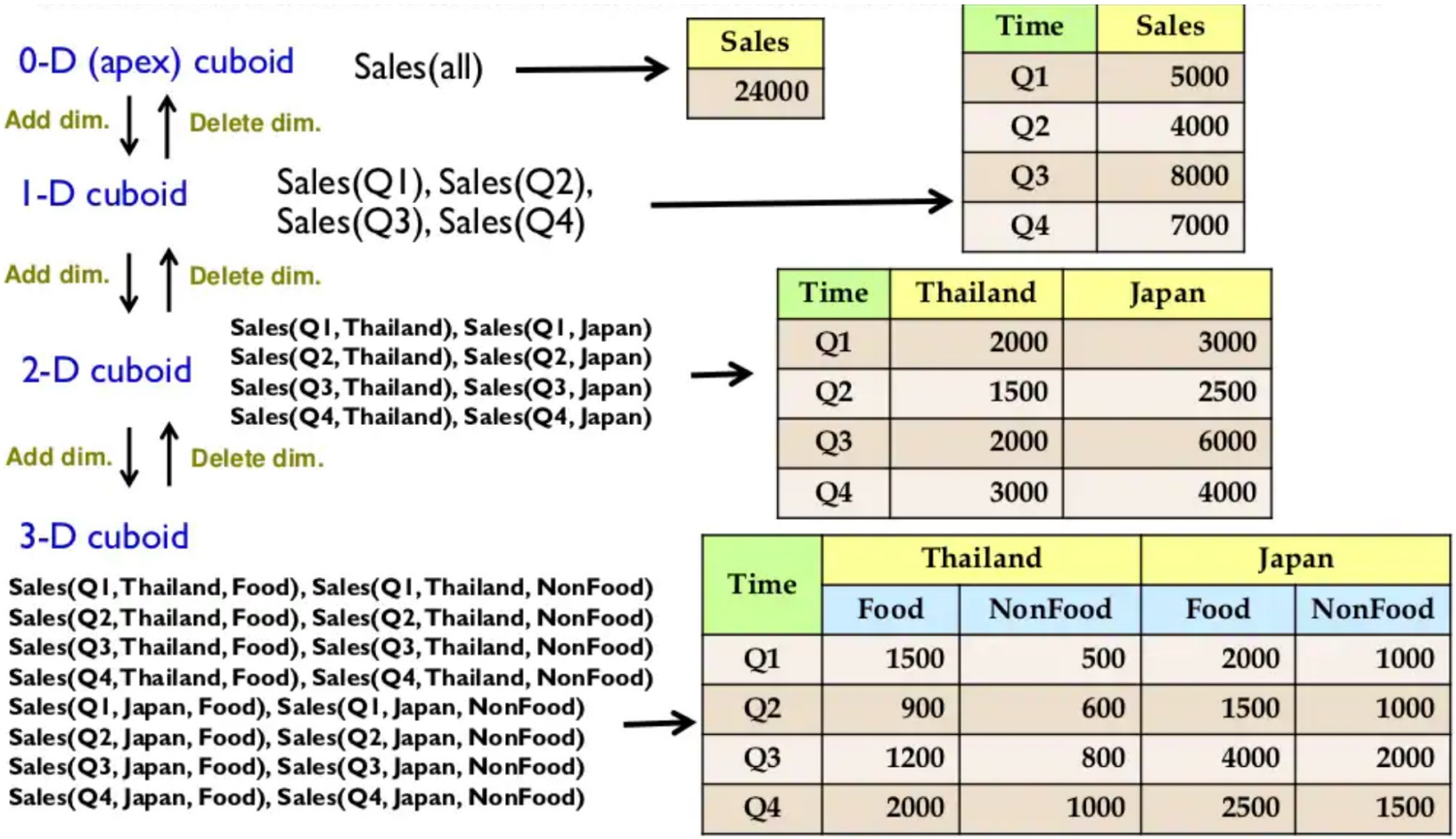
0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D (*base*) cuboid

An example of cuboids



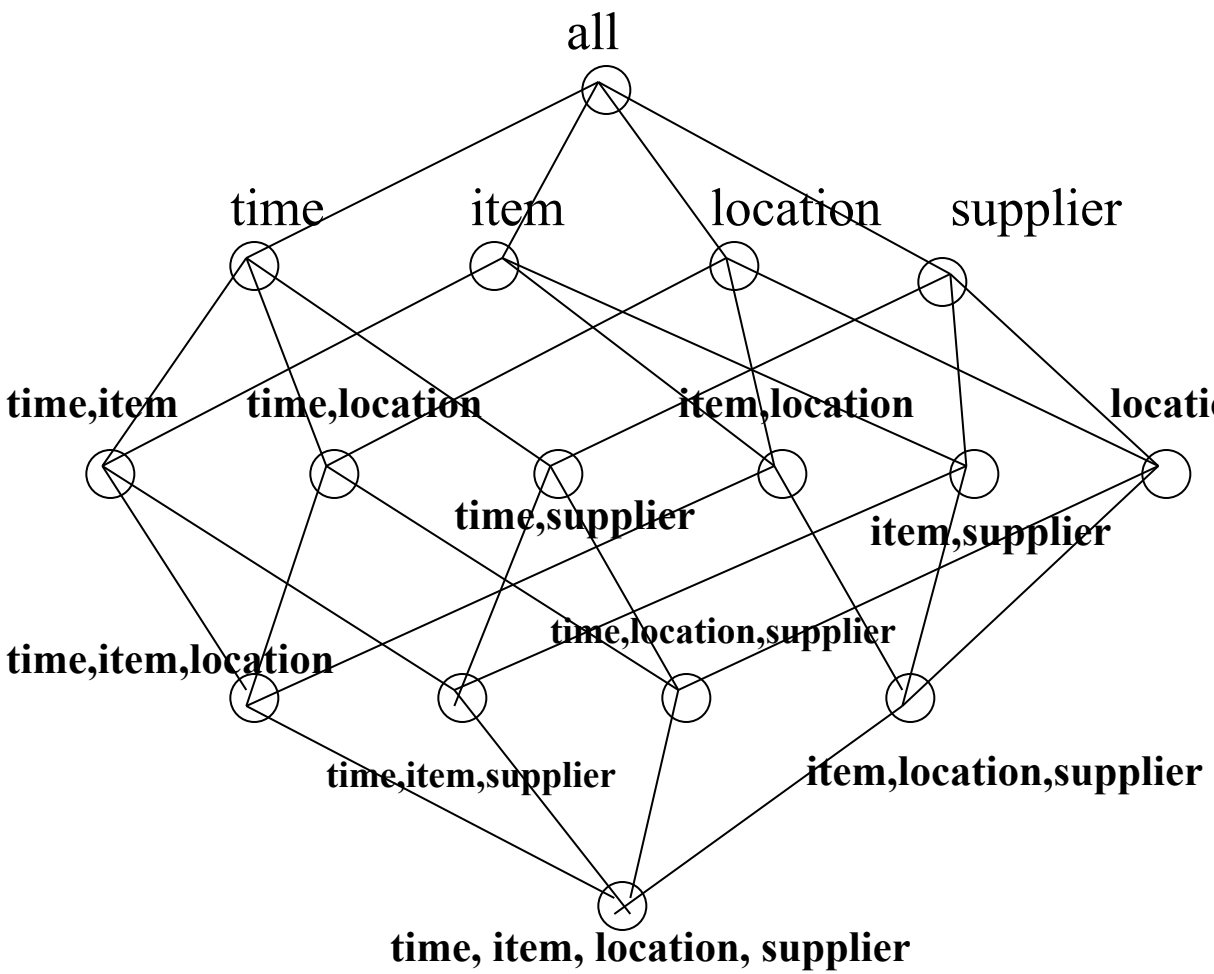
Sales
24000

Time	Sales
Q1	5000
Q2	4000
Q3	8000
Q4	7000

Time	Thailand	Japan
Q1	2000	3000
Q2	1500	2500
Q3	2000	6000
Q4	3000	4000

Time	Thailand		Japan	
	Food	NonFood	Food	NonFood
Q1	1500	500	2000	1000
Q2	900	600	1500	1000
Q3	1200	800	4000	2000
Q4	2000	1000	2500	1500

Cube: A Lattice of Cuboids



0-D (apex) cuboid

1-D cuboids

2-D cuboids

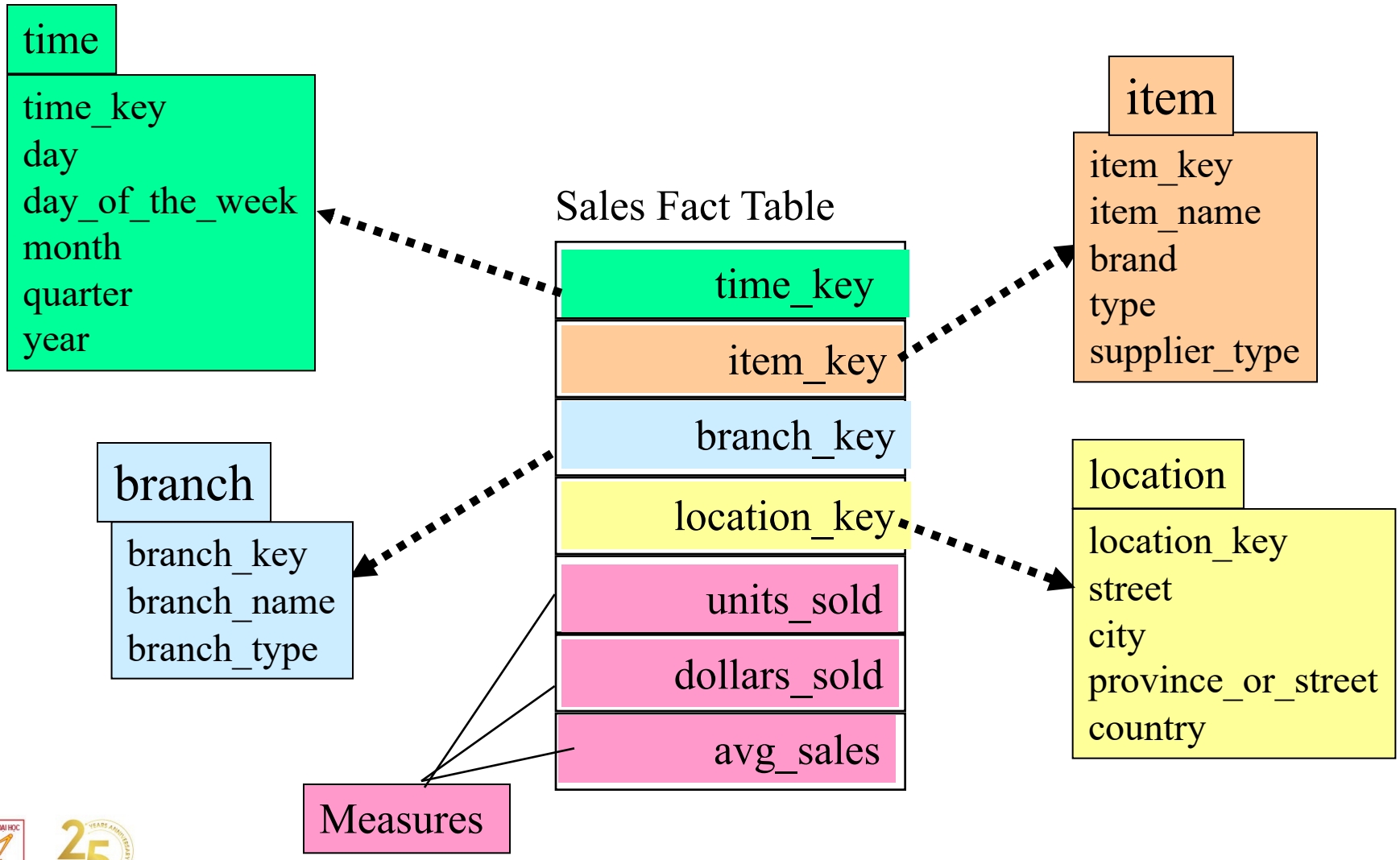
3-D cuboids

4-D (base) cuboid

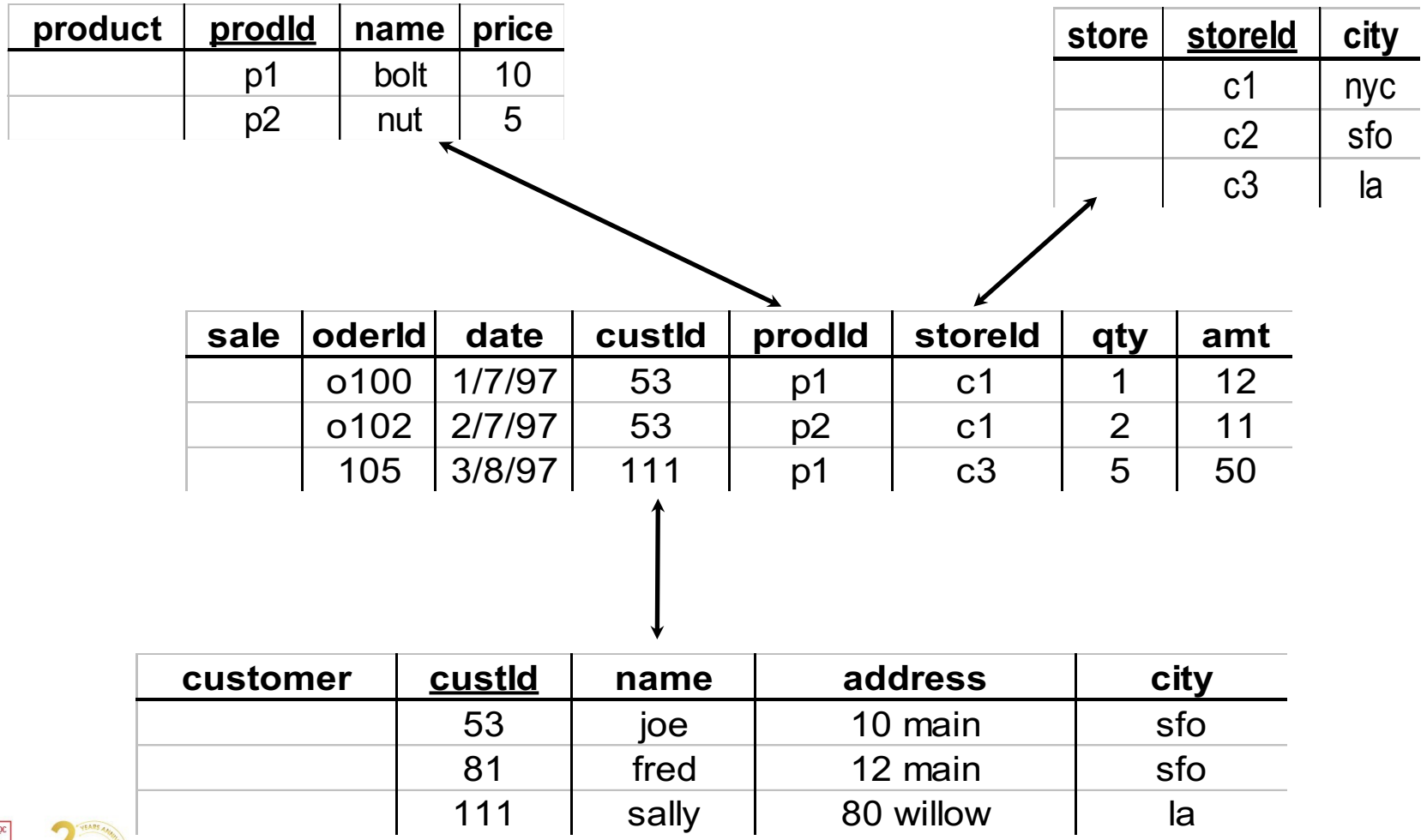
Data Warehouse Schemas (Conceptual)

- Star Schema
  - A fact table in the middle connected to a set of dimension tables
- Snowflake Schema
  - A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
- Fact Constellations
  - Multiple fact tables sharing dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

The Star Schema



The Star Schema: An Example



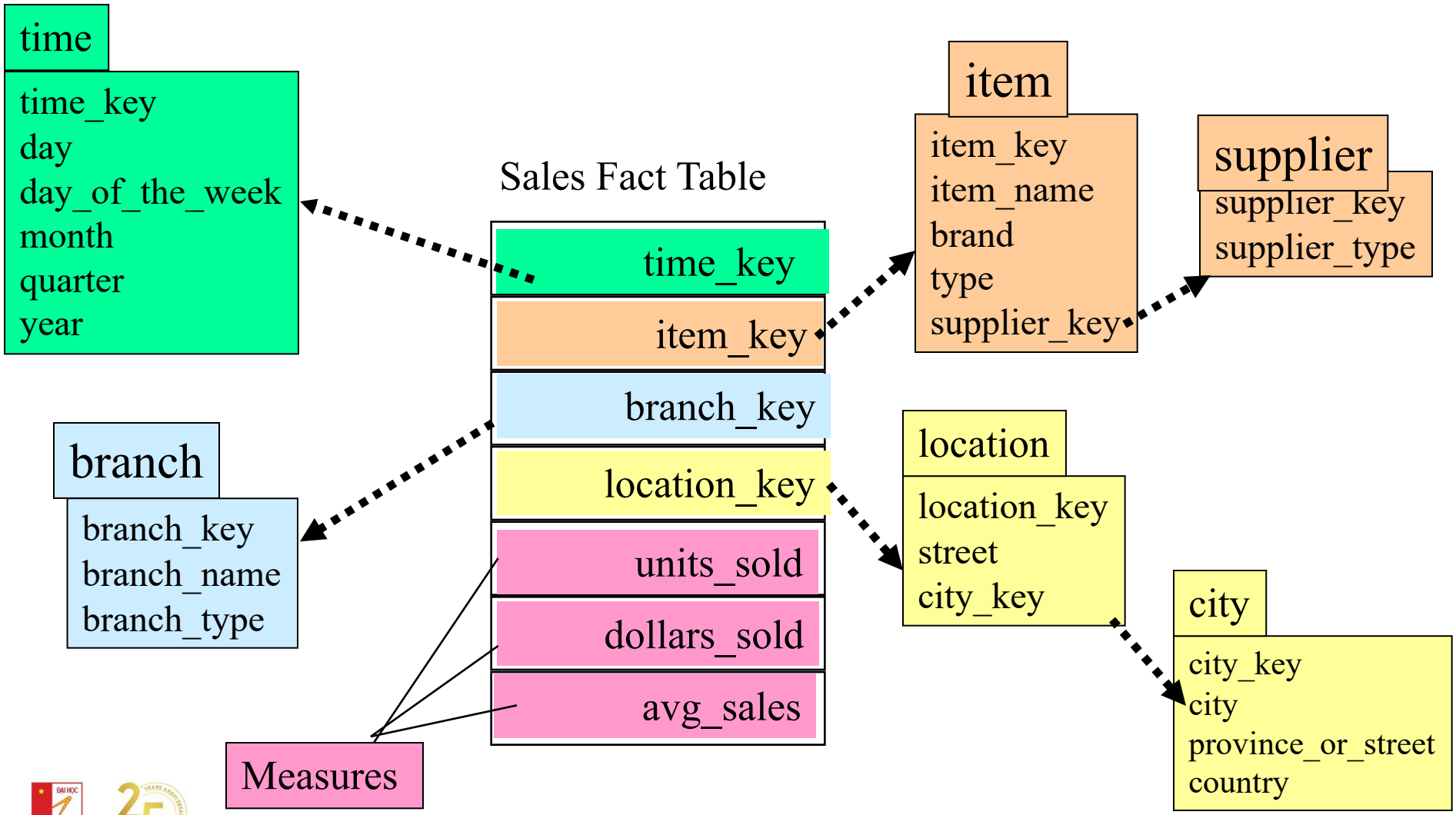
product	prodId	name	price
	p1	bolt	10
	p2	nut	5

store	storeId	city
	c1	nyc
	c2	sfo
	c3	la

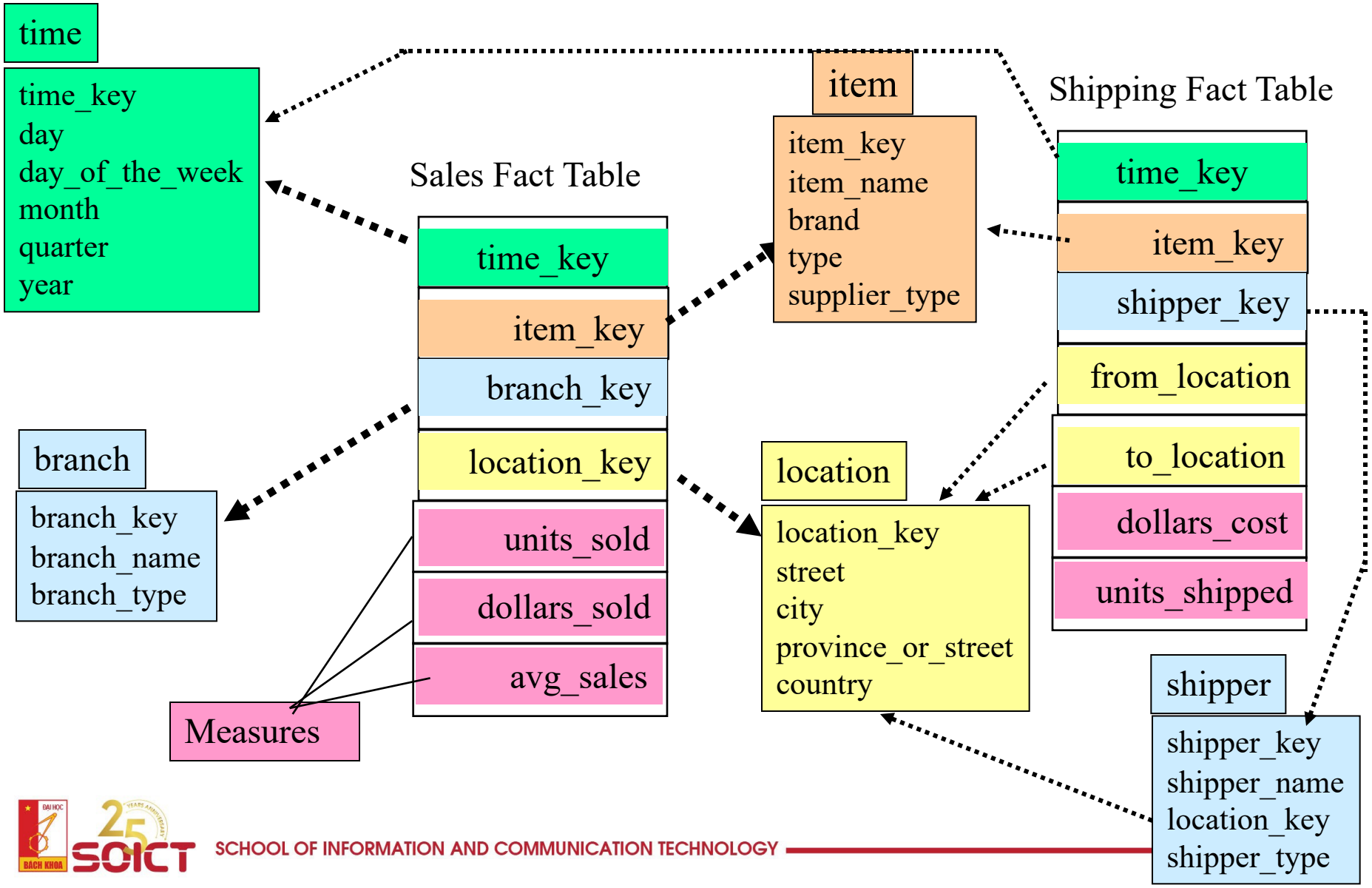
sale	oderId	date	custId	prodId	storeId	qty	amt
	o100	1/7/97	53	p1	c1	1	12
	o102	2/7/97	53	p2	c1	2	11
	105	3/8/97	111	p1	c3	5	50

customer	custId	name	address	city
	53	joe	10 main	sfo
	81	fred	12 main	sfo
	111	sally	80 willow	la

The Snowflake Schema

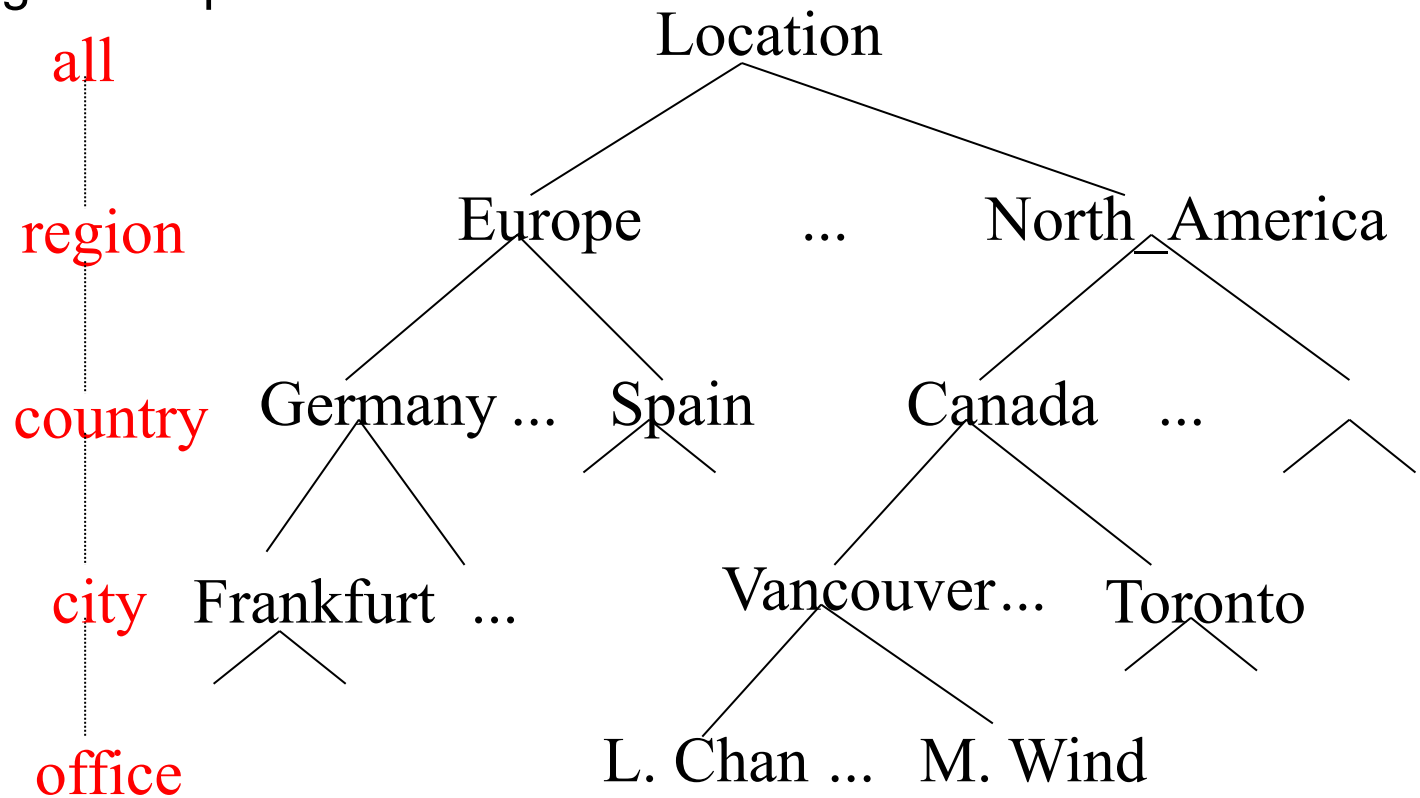


The Fact Constellation Schema



Concept Hierarchy

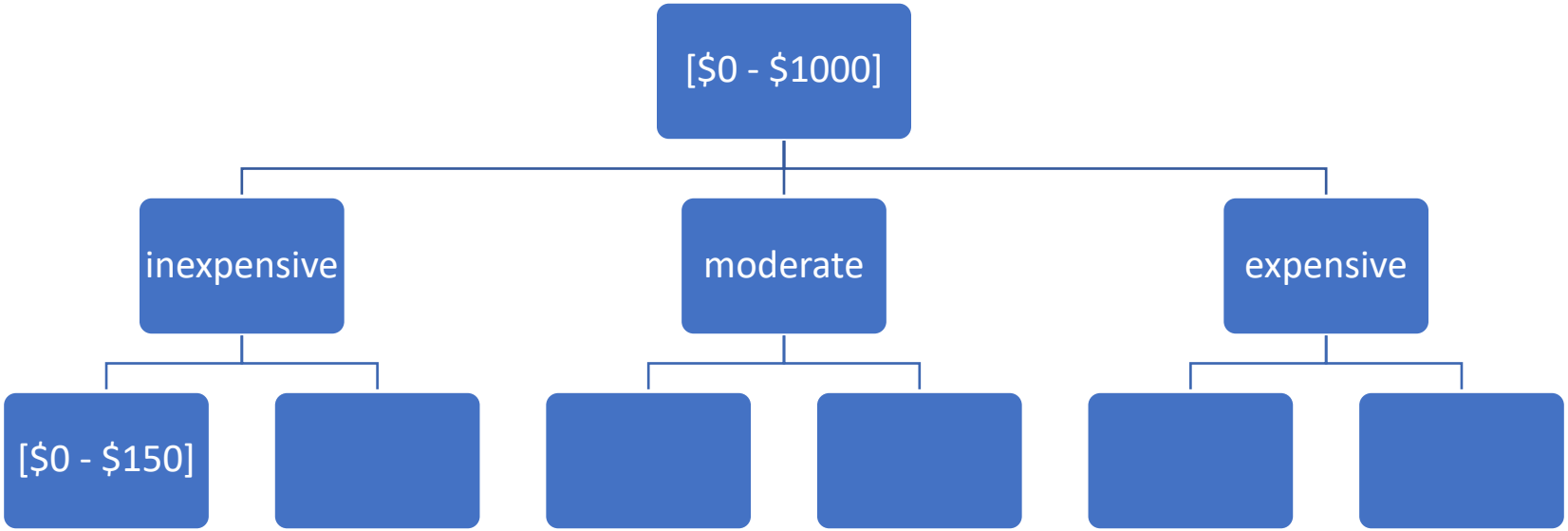
- Define a sequence of mappings from a set of very sepecific, low-level concepts to more general, higher-level concepts
  - E.g. concept of Location



Concept Hierarchy

- Concept hierarchies are useful to perform OLAP
  - Data are organized in multiple dimensions where each dimension contains multiple levels of abstraction defined by concept hierarchies
    - It give flexibility to summarize data on various levels of granularity
    - And OLAP operations enable materialization of such views

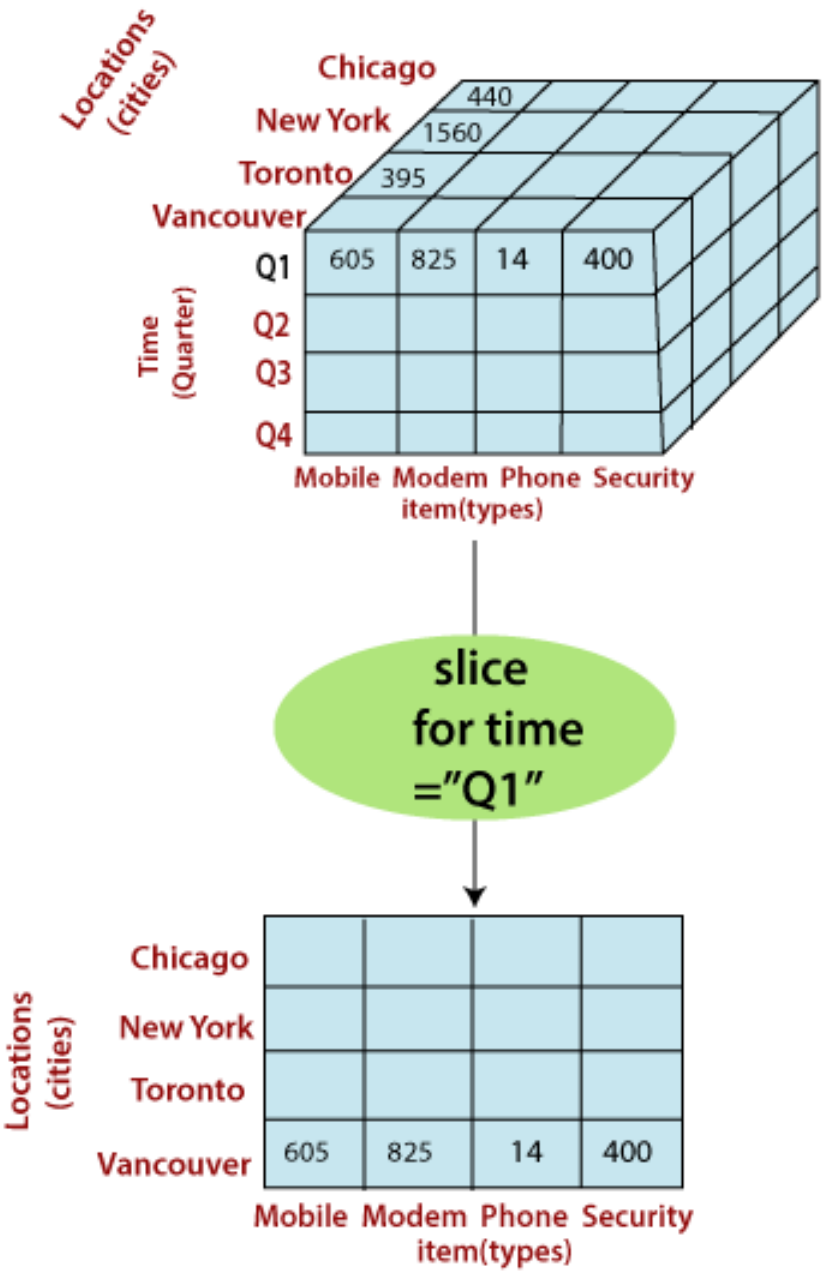
Set-Grouping Hierarchy



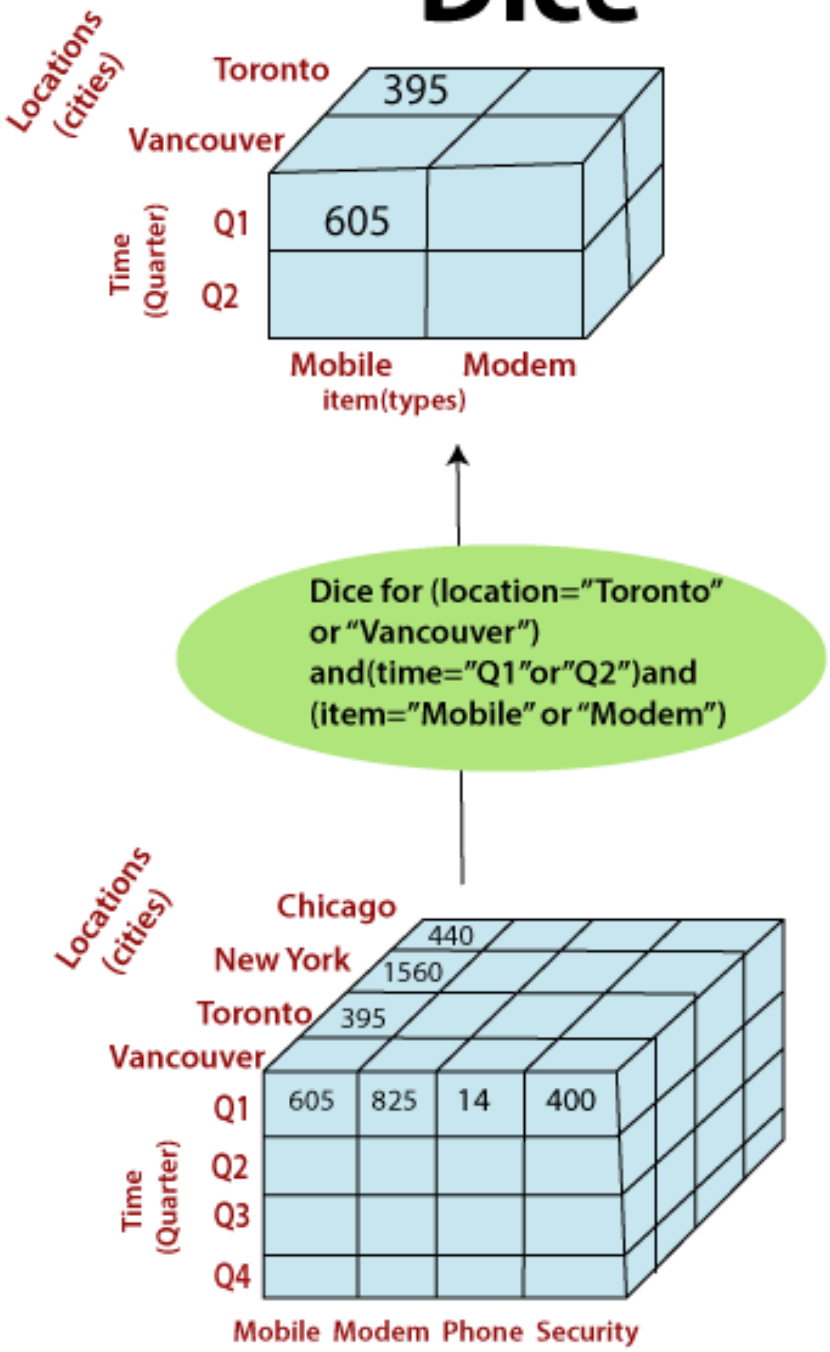
Typical OLAP Operations

- Roll up (drill-up): summarize data
  - by climbing up hierarchy or by dimension reduction
- Drill down (roll down): reverse of roll-up
  - from higher level summary to lower-level summary or detailed data, or introducing new dimensions
- Slice and dice: project and select
- Pivot (rotate):
  - reorient the cube, visualization, 3D to series of 2D planes

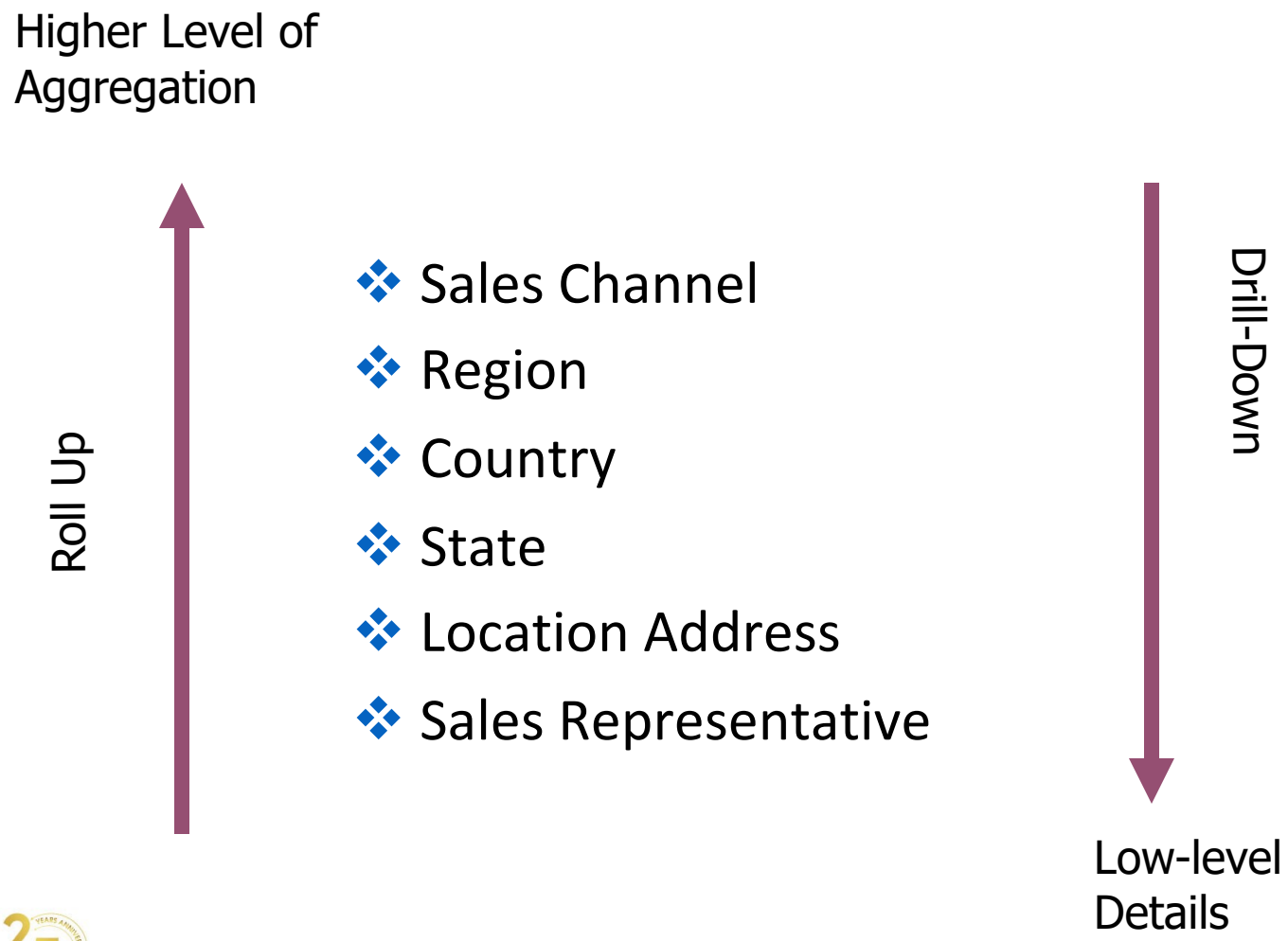
Slice


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Time (Quarter)
Q1
Q2
Q3
Q4
Mobile
Modem
Phone
Security
item(types)
slice for time ="Q1"
Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile
Modem
Phone
Security
item(types)

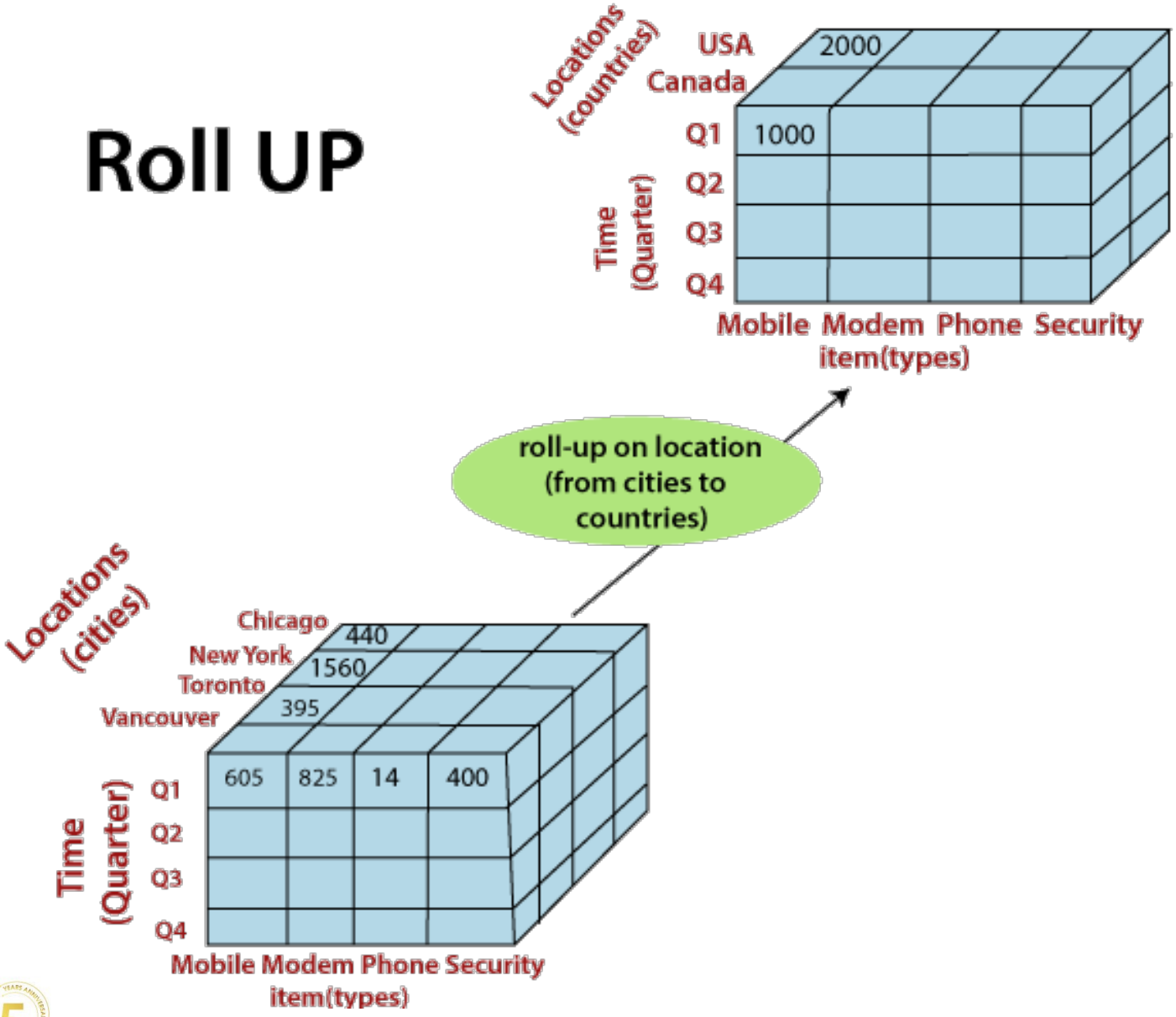
Dice


Locations (cities)
Toronto
Vancouver
395
605
Time (Quarter)
Q1
Q2
Mobile
Modem
item(types)
Dice for (location="Toronto" or "Vancouver") and(time="Q1"or"Q2")and (item="Mobile" or "Modem")
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security

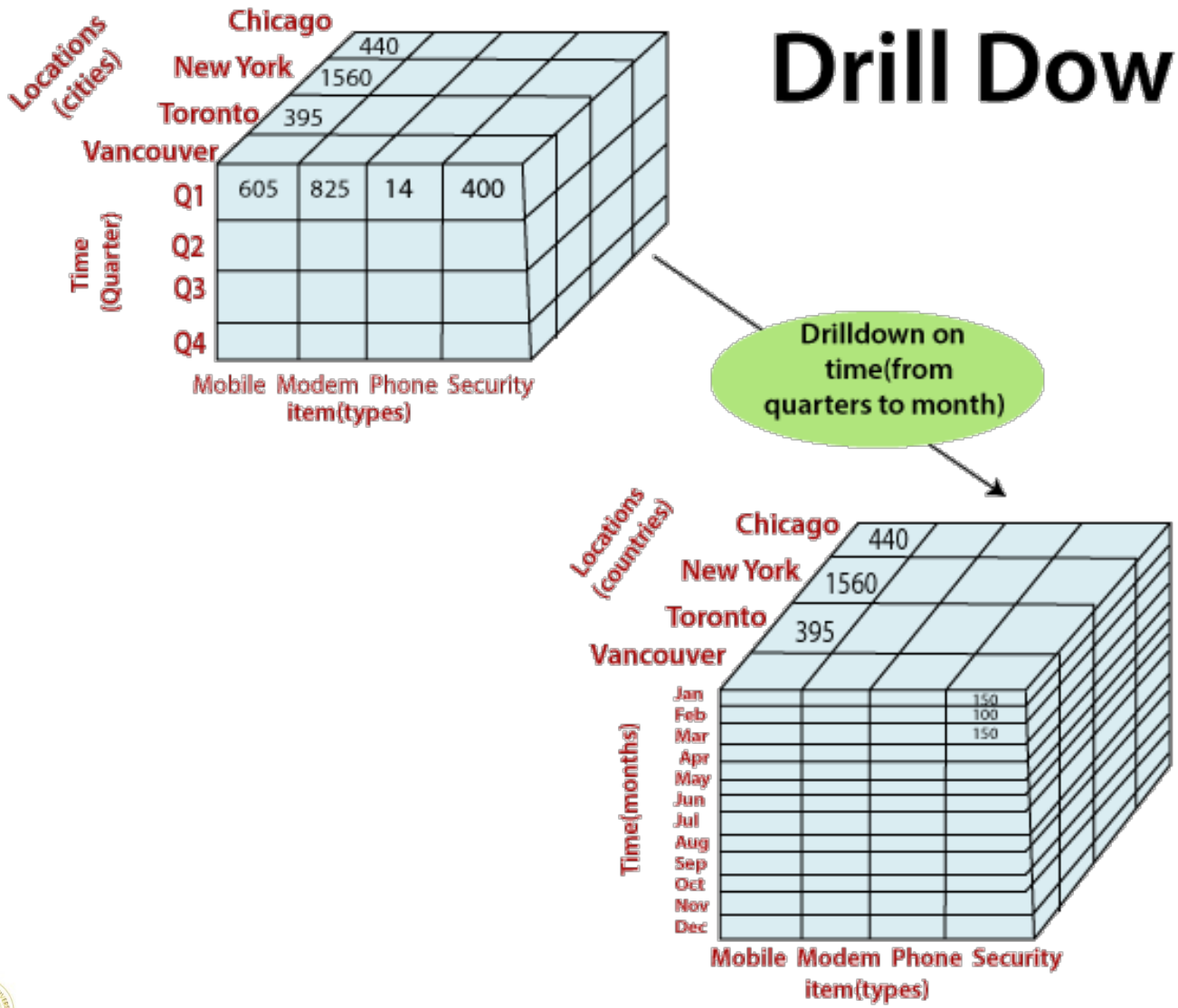
Roll-Up & Drill-Down



Roll UP

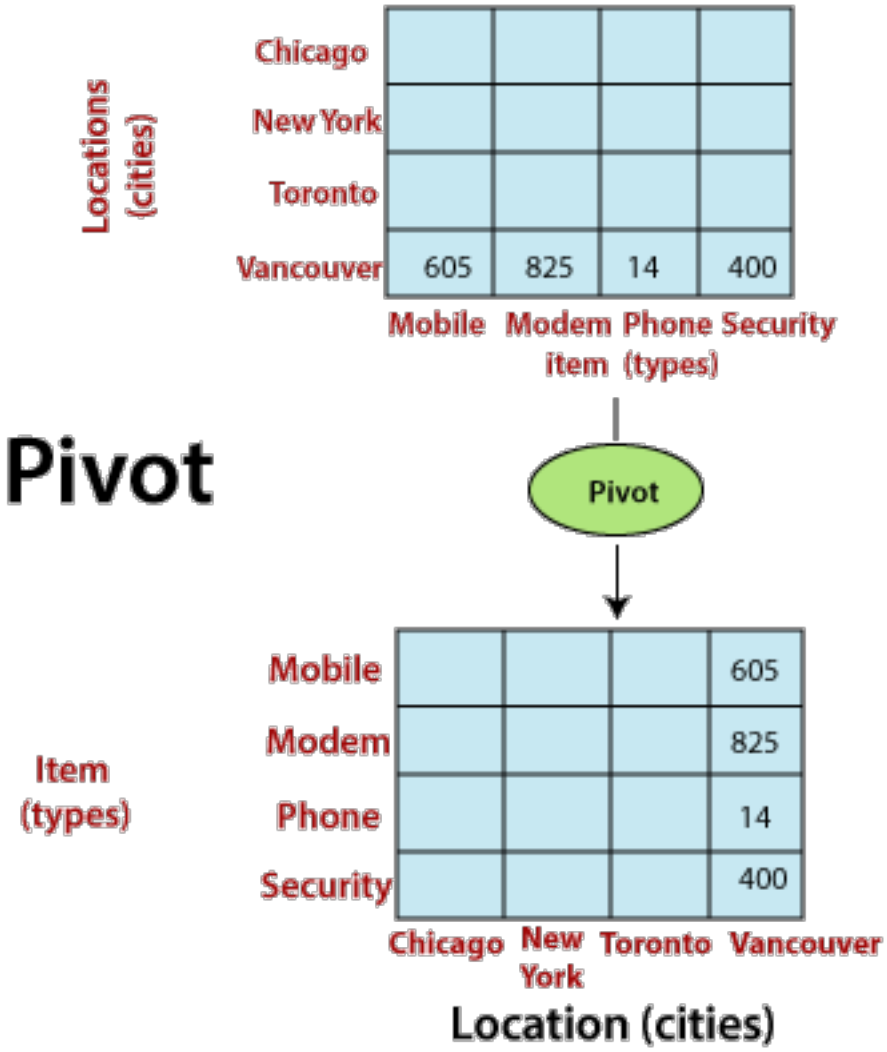

Locations (countries)
USA
Canada
2000
Q1
1000
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)

Drill Down

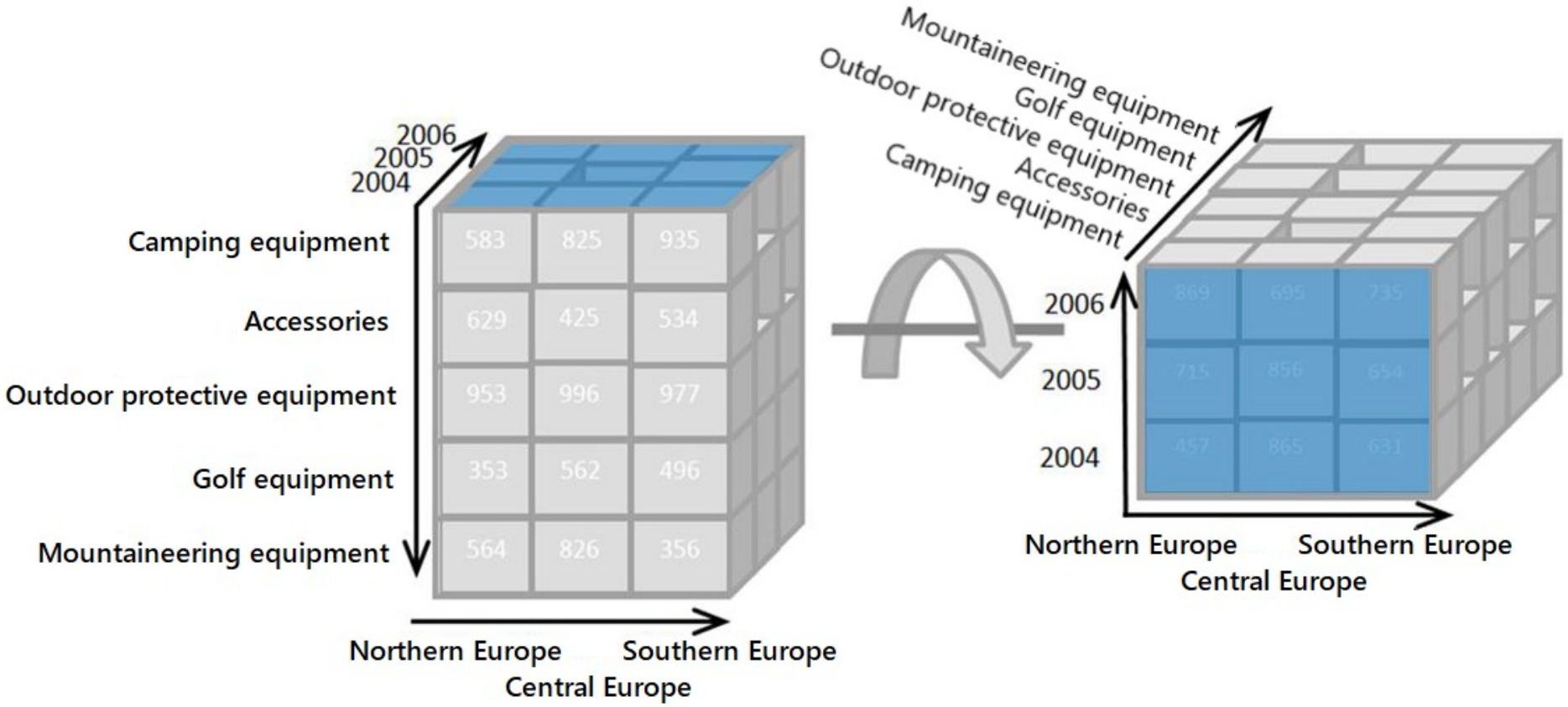

Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile Modem Phone Security
item(types)
Drilldown on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time(months)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
150
100
150
Mobile Modem Phone Security
item(types)

Pivot (1)

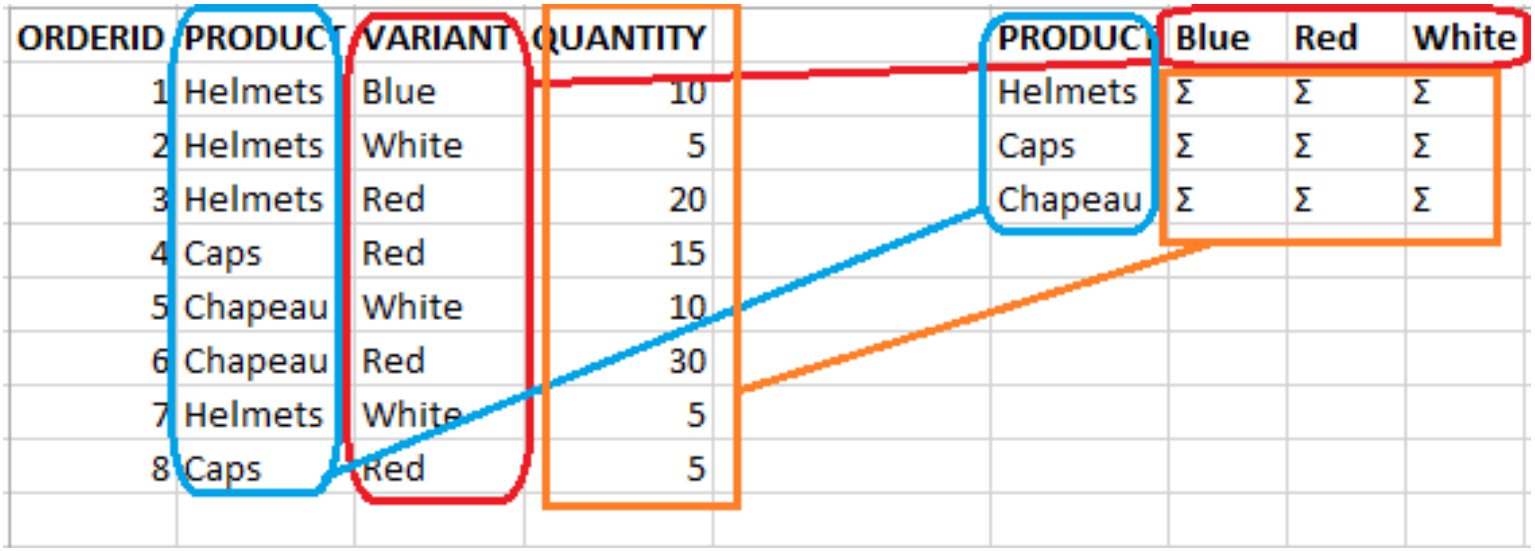
- Pivot is also called a rotation.
- Pivot rotates the data axes in view to provide an alternative presentation of the data.
- It may contain swapping the rows and columns or moving one of the row-dimensions into the column dimensions.



Pivot (2)


2006
2005
2004
Camping equipment
Accessories
Outdoor protective equipment
Golf equipment
Mountaineering equipment
583
825
935
629
425
534
953
996
977
353
562
496
564
826
356
Northern Europe
Southern Europe
Central Europe
Mountaineering equipment
Golf equipment
Outdoor protective equipment
Accessories
Camping equipment
2006
2005
2004
869
695
735
715
856
654
457
865
631
Northern Europe
Southern Europe
Central Europe

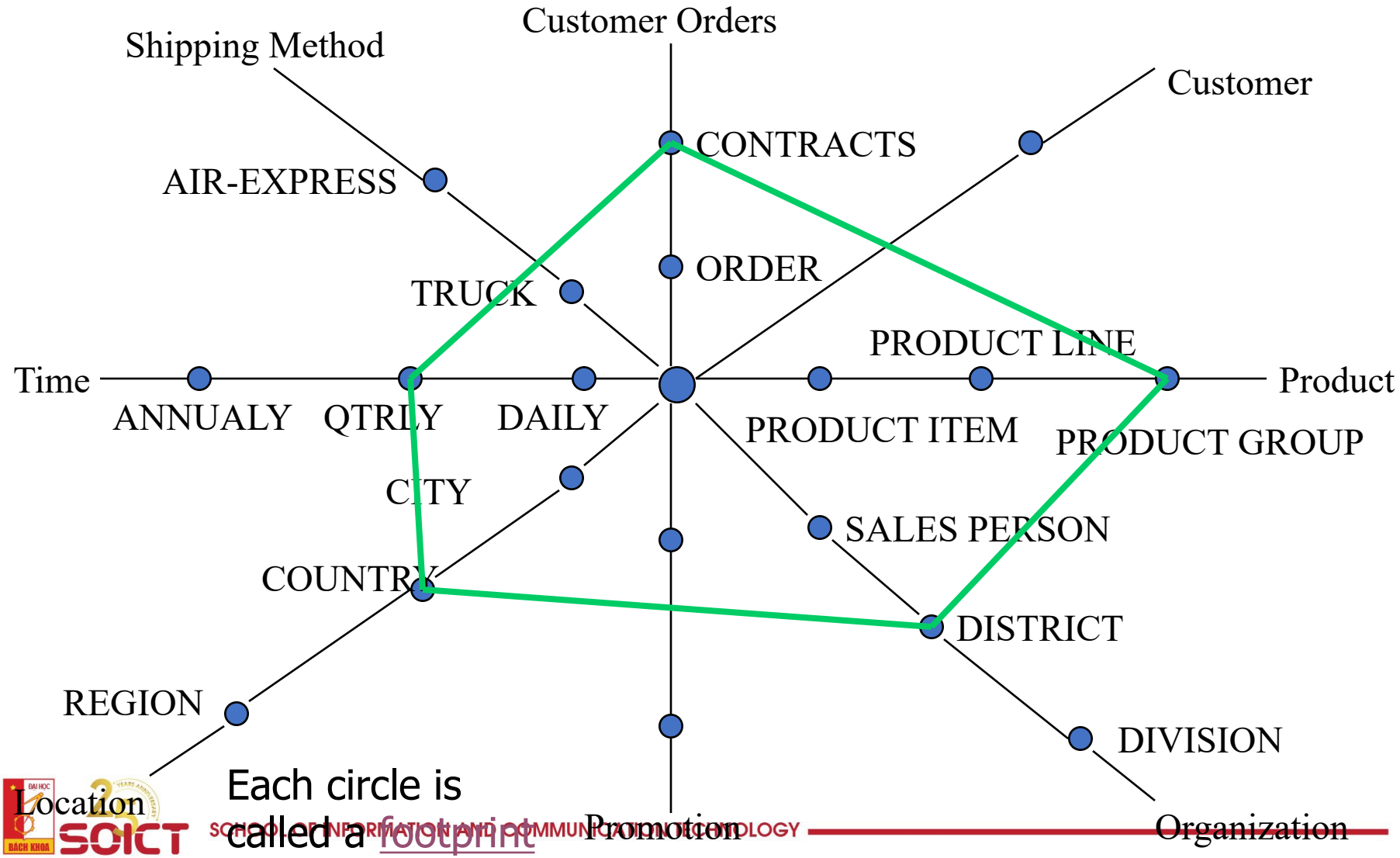
Pivot (3)



ORDERID	PRODUCT	VARIANT	QUANTITY
1	Helmets	Blue	10
2	Helmets	White	5
3	Helmets	Red	20
4	Caps	Red	15
5	Chapeau	White	10
6	Chapeau	Red	30
7	Helmets	White	5
8	Caps	Red	5

PRODUCT	Blue	Red	White
Helmets	Σ	Σ	Σ
Caps	Σ	Σ	Σ
Chapeau	Σ	Σ	Σ

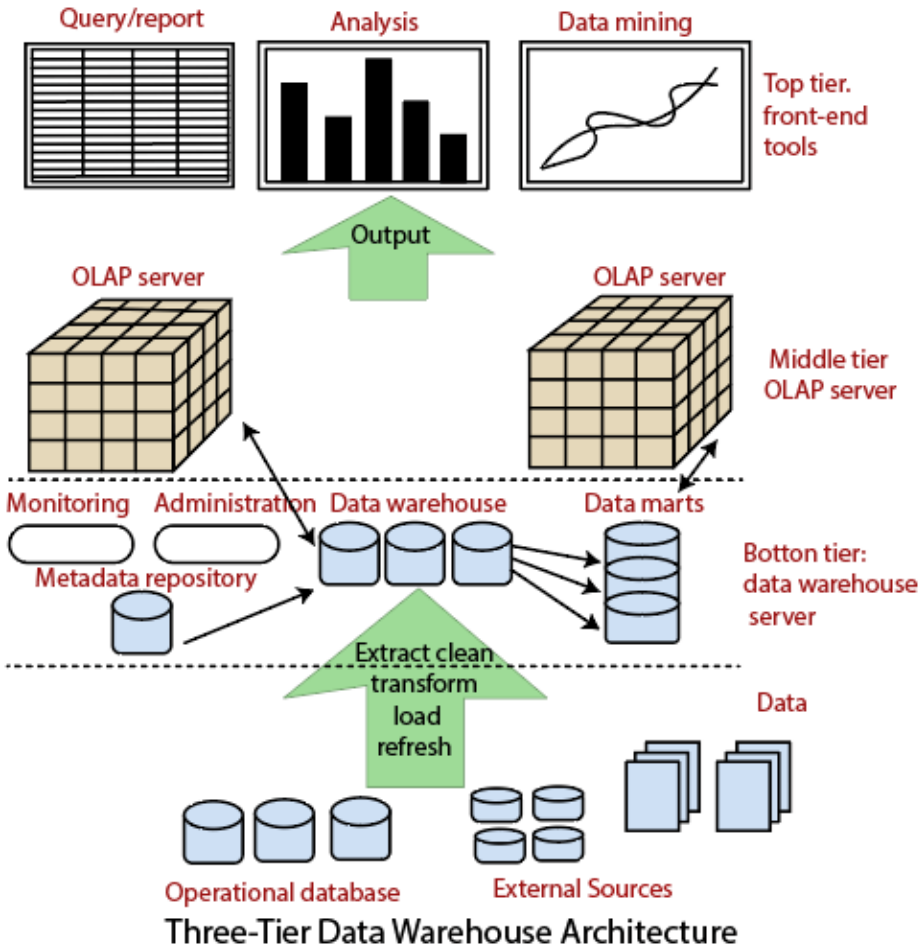
A Star-Net Query Model



Each circle is called a footprint

Data Warehouse Design

3-tier data warehouse architecture

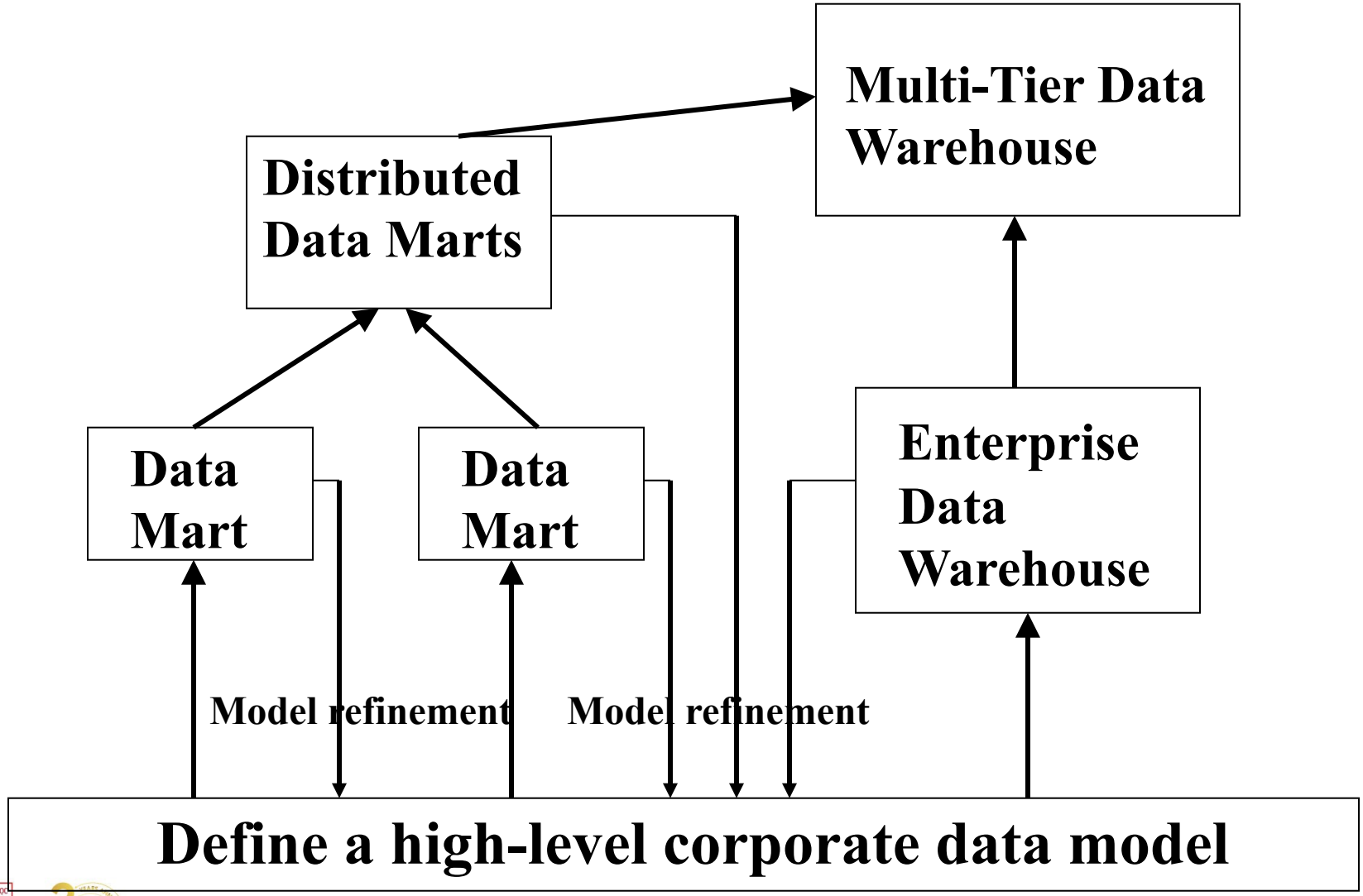


Three-Tier Data Warehouse Architecture

Enterprise warehouse & data marts

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data mart
  - Holds **data** only for a specific department or line of business, such as sales, finance, or human resources.
  - A **data warehouse** can feed **data** to a **data mart**
  - A **data mart** can feed a **data warehouse**.

A Recommended Approach



Summary

- Data warehousing: A multi-dimensional model of a data warehouse
  - A data cube consists of *dimensions* & *measures*
  - Star schema, snowflake schema, fact constellations
  - OLAP operations: drilling, rolling, slicing, dicing and pivoting
- Data Warehouse Architecture, Design, and Usage
  - Multi-tiered architecture
  - Business analysis design framework
  - Information processing, analytical processing, data mining
- Implementation: Efficient computation of data cubes
  - Partial vs. full vs. no materialization
  - Indexing OALP data: Bitmap index and join index
  - OLAP query processing
  - OLAP servers: ROLAP, MOLAP, HOLAP

References (I)

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.
- J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97-107, 1998.
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96
- J. Hellerstein, P. Haas, and H. Wang. Online aggregation. SIGMOD'97

References (II)

- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and G. Graefe. Multi-table joins through bitmapped join indices. *SIGMOD Record*, 24:8–11, Sept. 1995.
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS'00.
- D. Srivastava, S. Dar, H. V. Jagadish, and A. V. Levy. Answering queries with aggregation using views. *VLDB'96*
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM'95
- K. Wu, E. Otoo, and A. Shoshani, Optimal Bitmap Indices with Efficient Compression, ACM Trans. on Database Systems (TODS), 31(1): 1-38, 2006

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!

soict.hust.edu.vn/ fb.com/groups/soict