SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

18CSE487T

DATA WAREHOUSING AND ITS APPLICATIONS

UNIT-2

Data warehouse Dimensional Data Modeling

- Dimensional Data Modeling is one of the data modeling techniques used in data warehouse design.
- The concept of Dimensional Modeling was developed by Ralph Kimball which is comprised of facts and dimension tables.
- Since the main goal of this modeling is to improve the data retrieval so it is optimized for SELECT OPERATION.
- The advantage of using this model is that we can store data in such a way that it is easier to store and retrieve the data once stored in a data warehouse.
- The dimensional model is the data model used by many OLAP systems.

Facts

- Facts are the measurable data elements that represent the business metrics of interest.
- For example, in a sales data warehouse, the facts might include sales revenue, units sold, and profit margins.
- Each fact is associated with one or more dimensions, creating a relationship between the fact and the descriptive data.

Dimensions

- Dimensions are the descriptive data elements that are used to categorize or classify the data.
- For example, in a sales data warehouse, the dimensions might include product, customer, time, and location.
- Each dimension is made up of a set of attributes that describe the dimension.
- For example, the product dimension might include attributes such as product name, product category, and product price.

Data warehouse Schema

- Schema is a logical description of the entire database.
- It includes the name and description of records of all record types including all associated data-items and aggregates.
- Much like a database, a data warehouse also requires to maintain a schema.
- A database uses relational model, while a data warehouse uses
  - Star,
  - Snowflake, and
  - Fact Constellation schema.

Data warehouse Schema

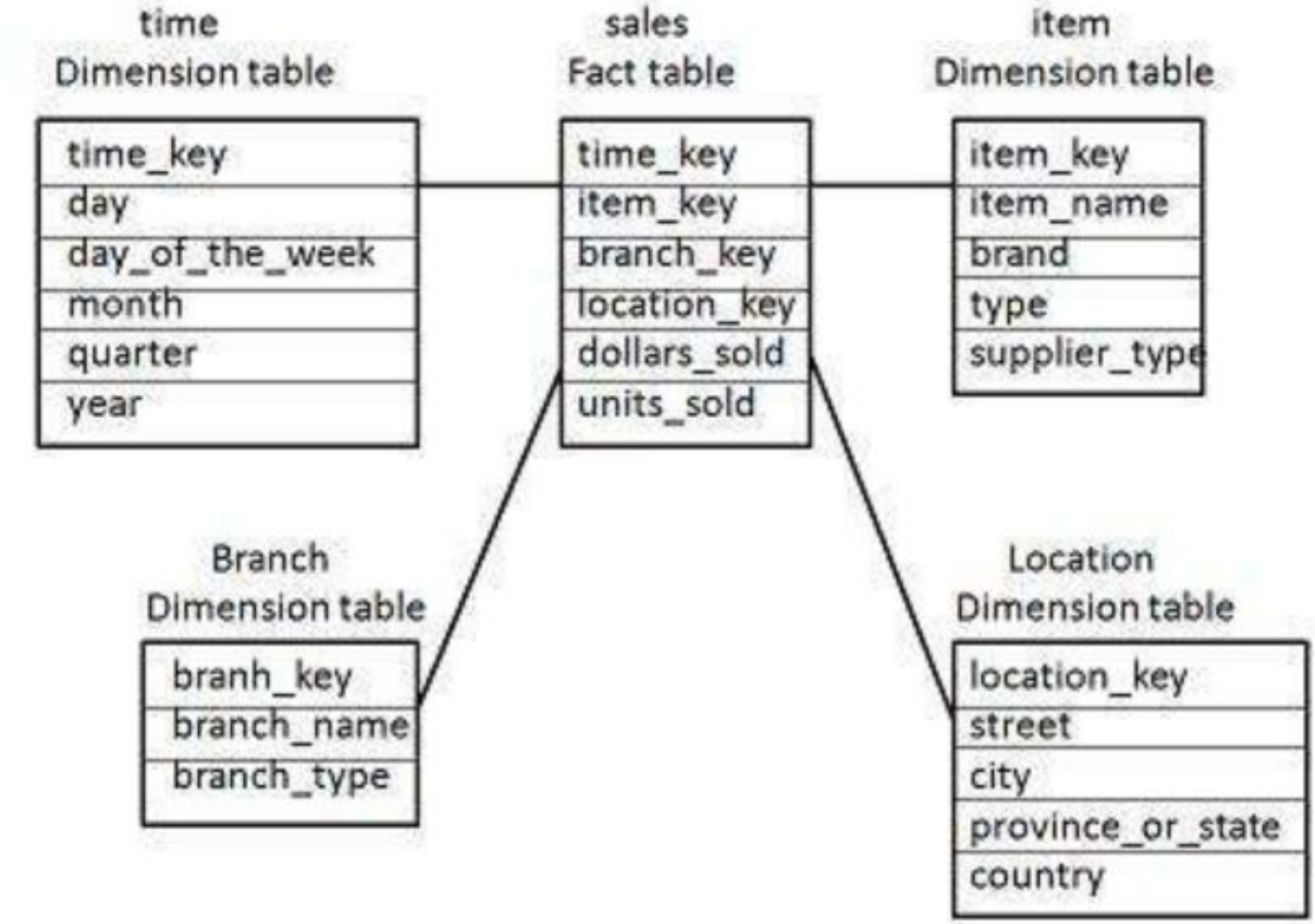
Types of Multidimensional Schemas

- Star Schema
- Snowflake Schema
- Fact Constellation Schema or Galaxy Schema

Star Schema

- Simplest type of Data Warehouse Schema.
- Structure resembles like a star.
- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- There is a fact table at the center. It contains the keys to each of four dimensions.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

Star Schema


time
Dimension table
time_key
day
day_of_the_week
month
quarter
year
sales
Fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
Dimension table
item_key
item_name
brand
type
supplier_type
Branch
Dimension table
branh_key
branch_name
branch_type
Location
Dimension table
location_key
street
city
province_or_state
country

Advantage of Star Schema

- Easy for Users to Understand
- Optimizes Navigation
- Most Suitable for Query Processing(against an OLTP system)

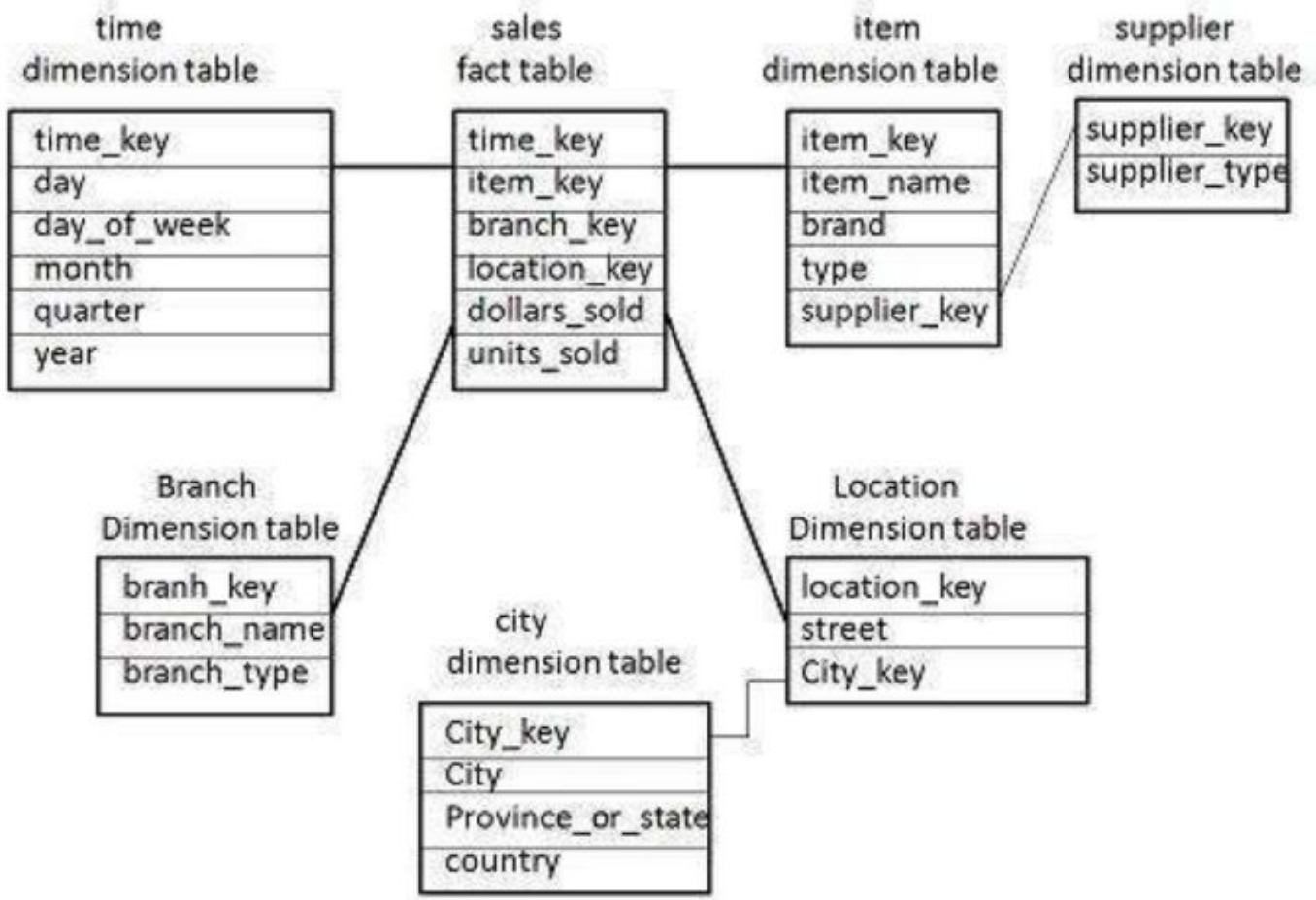
Disadvantages of Star Schema

- Data integrity is not enforced.
- Not as flexible in terms of analytical needs.
- Star schemas don't reinforce many-to-many relationships within business entities – at least not frequently.
- Uses large disk space

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized.
- For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

Snowflake Schema



Advantages of Snowflake Schema

- Reduces the the problem of **data integrity**
- Uses small **disk space**
- Improvement in **query performance**
- Easy to understand.

Disadvantages of Snowflakes Schema

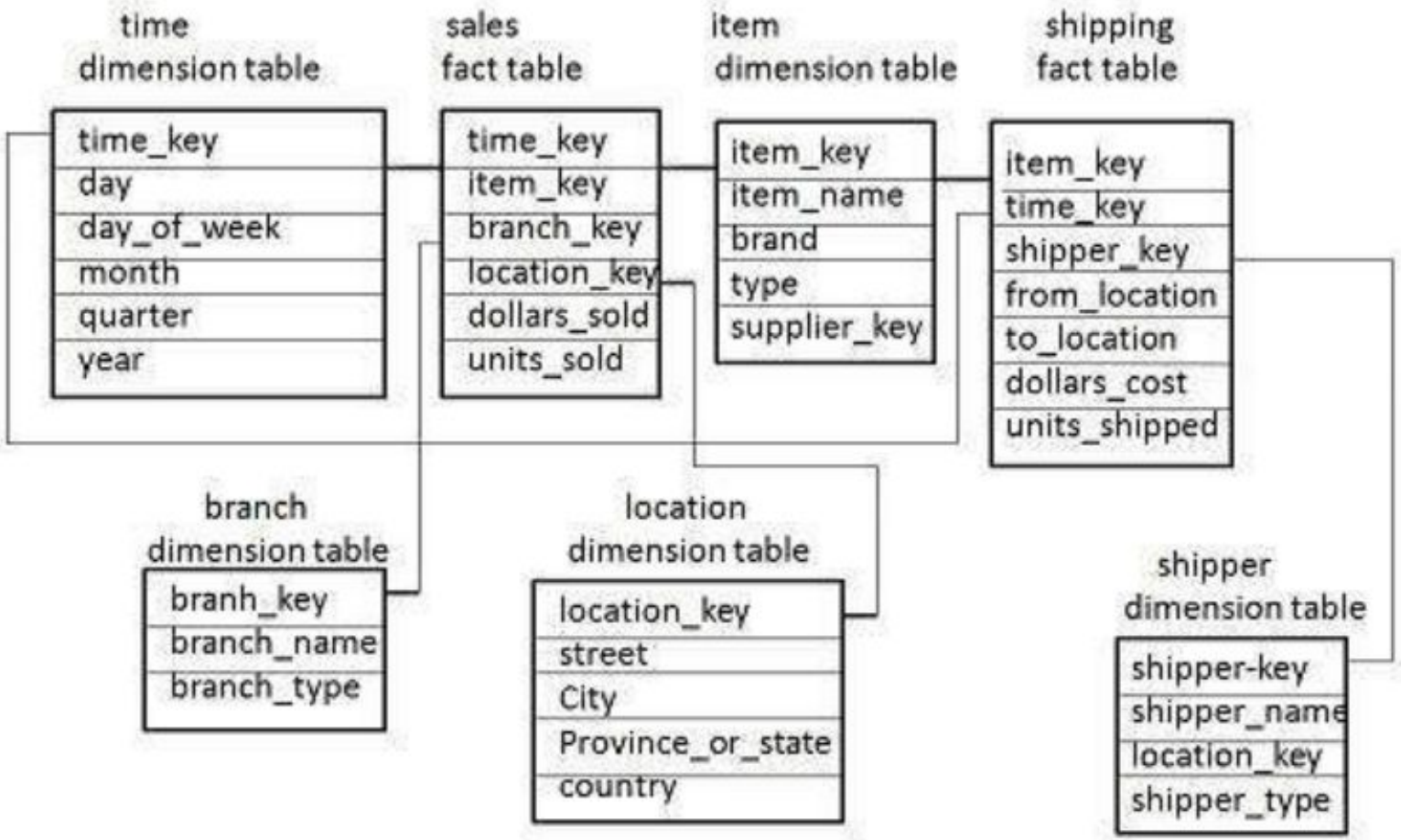
- Adds complexity to source query joins
- Additional maintenance efforts needed due to the increase number of lookup tables.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.

- Combines other two schemas.

Fact Constellation Schema



Fact Constellation Schema

- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Advantages of Fact Constellation Schema

- Provides a flexible schema.
  - Improved data retrival
  - Simplified business logic
  - Better understanding
  - Fast aggregation
  - Extensibility

Disadvantages of Fact Constellation Schema

- difficult to maintain
- more complex than star and snowflake schemas

Difference between Star Schema and Snow Flake Schema

Star Schema

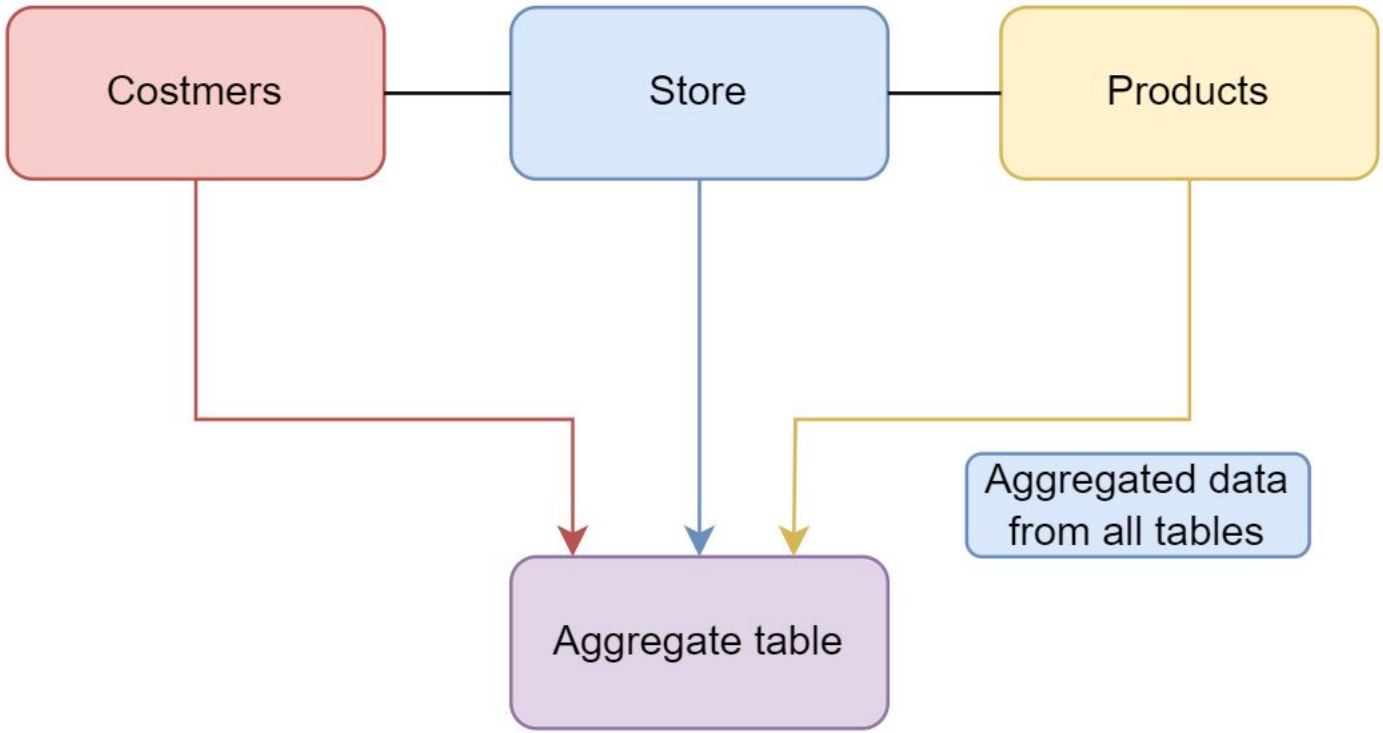
- The star schema is the simplest data warehouse scheme.
- In star schema, each of the dimensions is represented in a single table. It should not have any hierarchies between dims.
- It contains a fact table surrounded by dimension tables. If the dimensions are de-normalized, we say it is a star schema design.
- In star schema only one join establishes the relationship between the fact table and any one of the dimension tables.
- A star schema optimizes the performance by keeping queries simple and providing fast response time. All the information about the each level is stored in one row.
- It is called a star schema because the diagram resembles a star.

Snow Flake Schema

- Snowflake schema is a more complex data warehouse model than a star schema.
- In snow flake schema, at least one hierarchy should exist between dimension tables.
- It contains a fact table surrounded by dimension tables. If a dimension is normalized, we say it is a snow flaked design.
- In snow flake schema since there is relationship between the dimensions tables it has to do many joins to fetch the data.
- Snowflake schemas normalize dimensions to eliminated redundancy. The result is more complex queries and reduced query performance.
- It is called a snowflake schema because the diagram resembles a snowflake.

Aggregate Table

- Aggregate tables contain aggregated data that are precalculated summaries derived from fact tables.



Why do we use aggregate tables?

- A query in simple systems provides results for a single use case, such as when dealing with a single student, a single customer, and so on.
- On the other hand, in a data warehouse, a query generates large result sets.
- Let's suppose we want to retrieve data from multiple tables using a query in which we do some mathematical calculations.
- Since we have a large amount of data, it may take some time.
- To make it faster and more reliable, we create an aggregate table in which aggregated data is provided.

Uses of data aggregation

- It helps organizations achieve their business objectives.
- It helps with the statistical analysis of groups of people.
- Data aggregation can help improve our marketing.
- It also helps in improving our sales and purchases.

Data Extraction

- Data extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured. Data

- extraction makes it possible to consolidate, process, and refine data so that it can be stored in a centralized location in order to be transformed. These locations may be on-site, cloud-based, or a hybrid of the two.

Data Extraction

- Data extraction is the first step in both ETL (extract, transform, load) and ELT (extract, load, transform) processes.

- ETL/ELT are themselves part of a complete data integration strategy

Data Extraction and ETL

ETL allows companies and organizations to

1) consolidate data from different sources into a centralized location and

2) assimilate different types of data into a common format.

There are three steps in the ETL process:

1. Extraction
2. Transformation
3. Loading

ETL

1. Extraction :

Data is taken from one or more sources or systems. The extraction locates and identifies relevant data, then prepares it for processing or transformation. Extraction allows many different kinds of data to be combined and ultimately mined for business intelligence.

2. Transformation

Once the data has been successfully extracted, it is ready to be refined. During the transformation phase, data is sorted, organized, and cleansed.

For example, duplicate entries will be deleted, missing values removed or enriched, and audits will be performed to produce data that is reliable, consistent, and usable

3. Loading

The transformed, high quality data is then delivered to a single, unified target location for storage and analysis.

Benefits of Using an Extraction Tool

1. More control.
2. Increased agility.
3. Simplified sharing.
4. Accuracy and precision

Types of Data Extraction

1. Customer Data: .
2. Financial Data
3. Use, Task, or Process Performance Data.
4. figuring out where you can get it
5. Deciding where you want to store it.

Data transformation: Basic tasks

Data Transformation

Data transformation is the process in which data gets converted from one format to another.

- The most common data transformation process involves collecting raw data and converting it into clean, usable data.

Data transformation: Basic tasks

- Data transformation increases the efficiency of business and analytic processes.
- It enables businesses to make better data-driven decisions.
- During the data transformation process, an analyst will determine the structure of the data

Data transformation may be

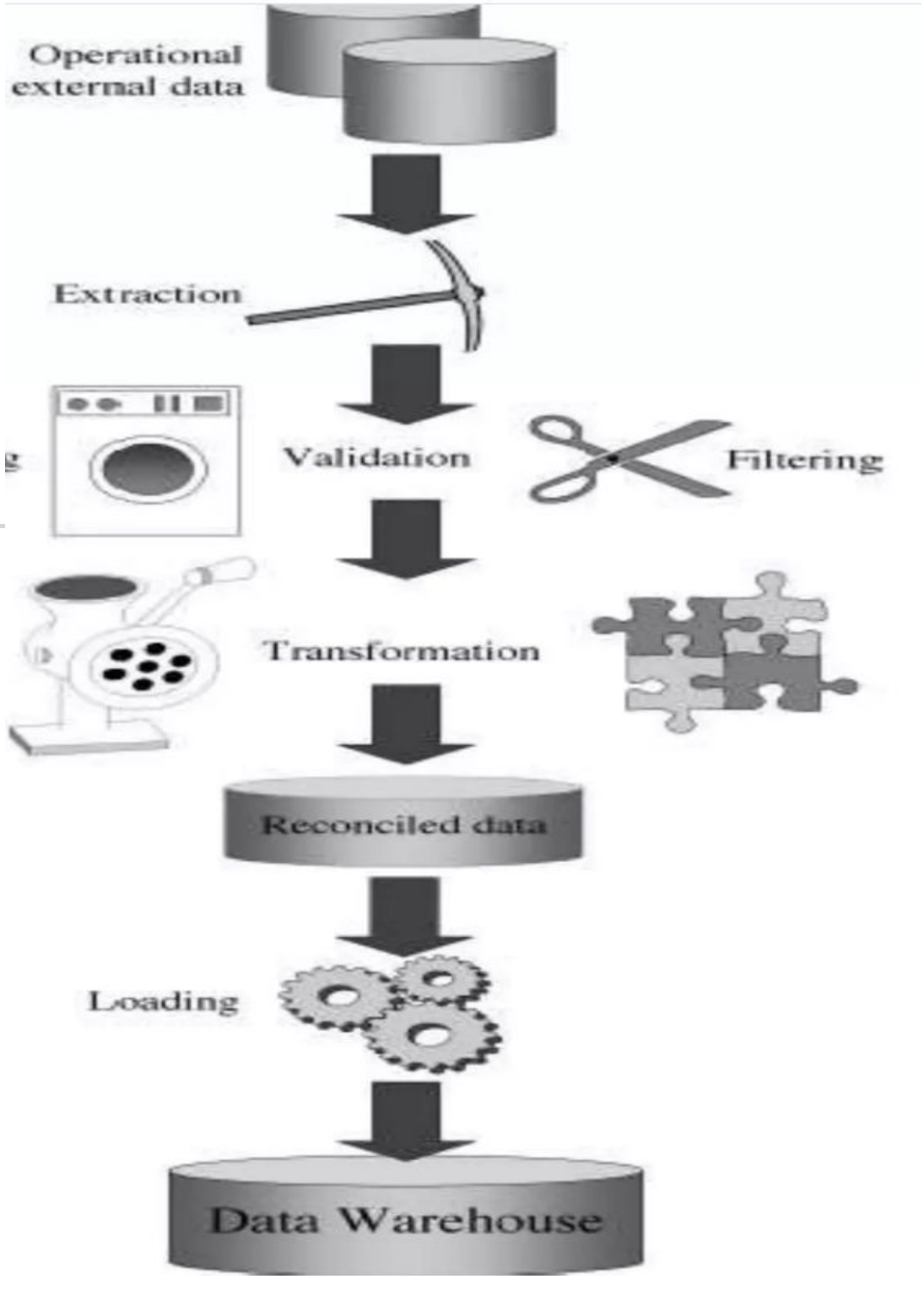
- Constructive: The data transformation process adds, copies, or replicates data.
- Destructive: The system deletes fields or records.
- Aesthetic: The transformation standardizes the data to meet requirements or parameters.
- Structural: (Renaming, moving, and combining columns in a database)

Data Transformation

Benefits of Data Transformation

- Data is transformed to make it better-organized. Transformed data may be easier for both humans and computers to use.
- Properly formatted and validated data improves data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.
- Data transformation facilitates compatibility between applications, systems, and types of data. Data used for multiple purposes may need to be transformed in different ways..

Data Transformation



Data Transformation

Transformation may tasks vary based on the application:

Major Tasks are:

1. Selection
2. Splitting/ Joining
3. Conversion
4. Summarization
5. Enrichment

Data Transformation

1. Selection:

- This takes place at the beginning of the whole process of data transformation.
- Select either whole records or parts of several records from the source systems.
- The task of selection usually forms part of the extraction function itself.

Data Transformation

2. Splitting/ Joining

- This task includes the types of data manipulation you need to perform on the selected parts of source records.
- Sometimes (uncommonly), you will be splitting the selected parts even further during data transformation.
- Joining of parts selected from many source systems is more widespread in the data warehouse environment

Data Transformation

3. Conversion

- This is an all-inclusive task. It includes a large variety of rudimentary conversions of single fields for two primary reasons

(i) to standardize among the data extractions from disparate source systems, and

(ii) to make the fields usable and understandable to the users.

Data Transformation

4. Summarization

- Sometimes it is not feasible to keep data at the lowest level of detail in your data warehouse.
- It may be that none of your users ever need data at the lowest granularity for analysis or querying.
- **For example**, for a grocery chain, sales data at the lowest level of detail for every transaction at the checkout may not be needed. Storing sales by product by store by day in the data warehouse may be quite adequate.
- So, in this case, the data transformation function includes summarization of daily sales by product and by store.

Data Transformation

5. Enrichment

- This task is the rearrangement and simplification of individual fields to make them more useful for the data warehouse environment.
- Usage of one or more fields from the same input record to create a better view of the data for the data warehouse.
- This principle is extended when one or more fields originate from multiple records, resulting in a single field for the data warehouse.

Data Transformation Types

- Format Revision
- Data Derivation
- Data Splitting
- Data Joining
- Data Summarization
- Key Restructuring
- Data Deduplication

Data Transformation: Benefits

- Better Organization
- Improved Data
- Faster Queries
- Simpler Data Management
- Broader Use

OLTP Compared With OLAP

- On Line Transaction Processing – *OLTP*
  - Maintains a database that is an accurate model of some real-world enterprise. Supports day-to-day operations. Characteristics:
    - Short simple transactions
    - Relatively frequent updates
    - Transactions access only a small fraction of the database
- On Line Analytic Processing – *OLAP*
  - Uses information in database to guide strategic decisions. Characteristics:
    - Complex queries
    - Infrequent updates
    - Transactions access a large fraction of the database
    - Data need not be up-to-date

OLAP

- OLAP stands for On-Line Analytical Processing.
- OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

OLAP Server Architectures

- **Relational OLAP (ROLAP)**
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - greater scalability
- **Multidimensional OLAP (MOLAP)**
  - Array-based multidimensional storage engine (sparse matrix techniques)
  - fast indexing to pre-computed summarized data
- **Hybrid OLAP (HOLAP)**
  - User flexibility, e.g., low level: relational, high-level: array
- **Specialized SQL servers**
  - specialized support for SQL queries over star/snowflake schemas

OLAP Guidelines (Dr.E.F.Codd Rule)

Benefits of OLAP

- OLAP helps managers in decision-making through the multidimensional record views that it is efficient in providing, thus increasing their productivity.
- OLAP functions are self-sufficient owing to the inherent flexibility support to the organized databases.
- It facilitates simulation of business models and problems, through extensive management of analysis-capabilities.
- In conjunction with data warehouse, OLAP can be used to support a reduction in the application backlog, faster data retrieval, and reduction in query drag.

Data Warehouses

- OLAP and data mining databases are frequently stored on special servers called data warehouses:
  - Can accommodate the huge amount of data generated by OLTP systems
  - Allow OLAP queries and data mining to be run off-line so as not to impact the performance of OLTP

OLAP, Data Mining, and Analysis

- The "A" in OLAP stands for "Analytical"
- Many OLAP and Data Mining applications involve sophisticated analysis methods from the fields of mathematics, statistical analysis, and artificial intelligence.
- Our main interest is in the database aspects of these fields, not the sophisticated analysis techniques.

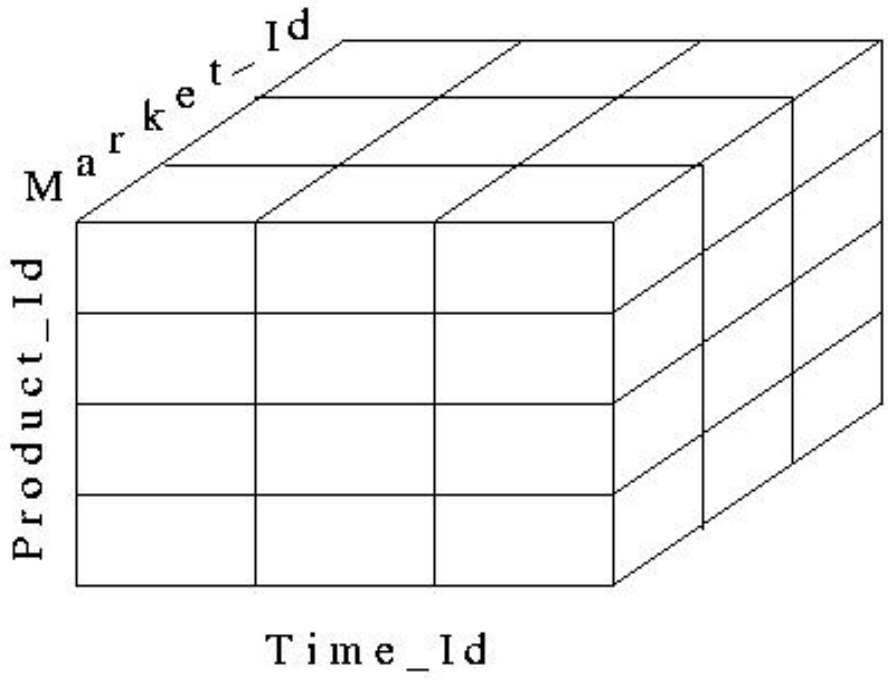
Fact Tables

- Many OLAP applications are based on a ***fact table***
- For example, a supermarket application might be based on a table

  Sales (*Market_Id*, *Product_Id*, *Time_Id*, *Sales_Amt*)
- The table can be viewed as *multidimensional*
  - *Market_Id*, *Product_Id*, *Time_Id* are the dimensions that represent specific supermarkets, products, and time intervals
  - *Sales_Amt* is a function of the other three

A Data Cube

- Fact tables can be viewed as an N-dimensional *data cube* (3-dimensional in our example)
  - The entries in the cube are the values for *Sales_Amts*



Dimension Tables

- The dimensions of the fact table are further described with ***dimension tables***
- Fact table:

  Sales (*Market_id*, *Product_Id*, *Time_Id*, *Sales_Amt*)
- Dimension Tables:

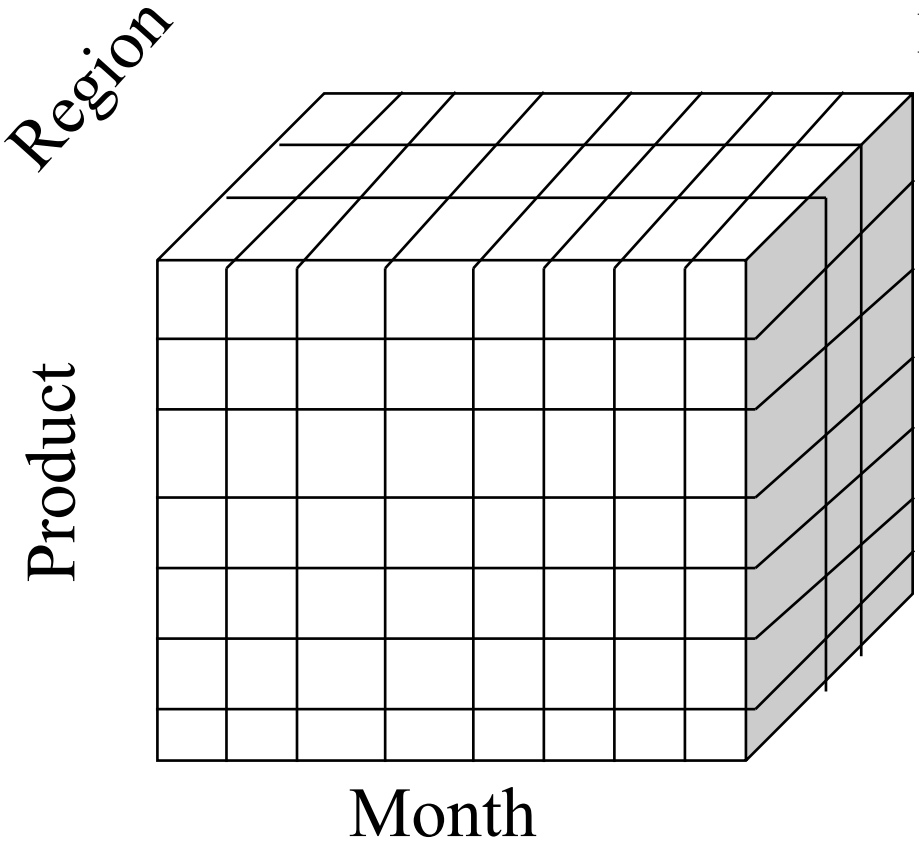
  Market (*Market_Id*, *City*, *State*, *Region*)

  Product (*Product_Id*, *Name*, *Category*, *Price*)

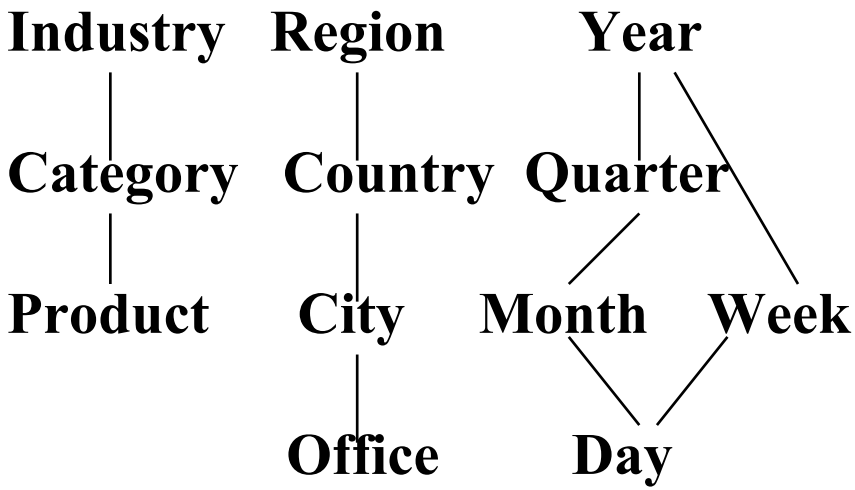
  Time (*Time_Id*, *Week*, *Month*, *Quarter*)

Multidimensional Data

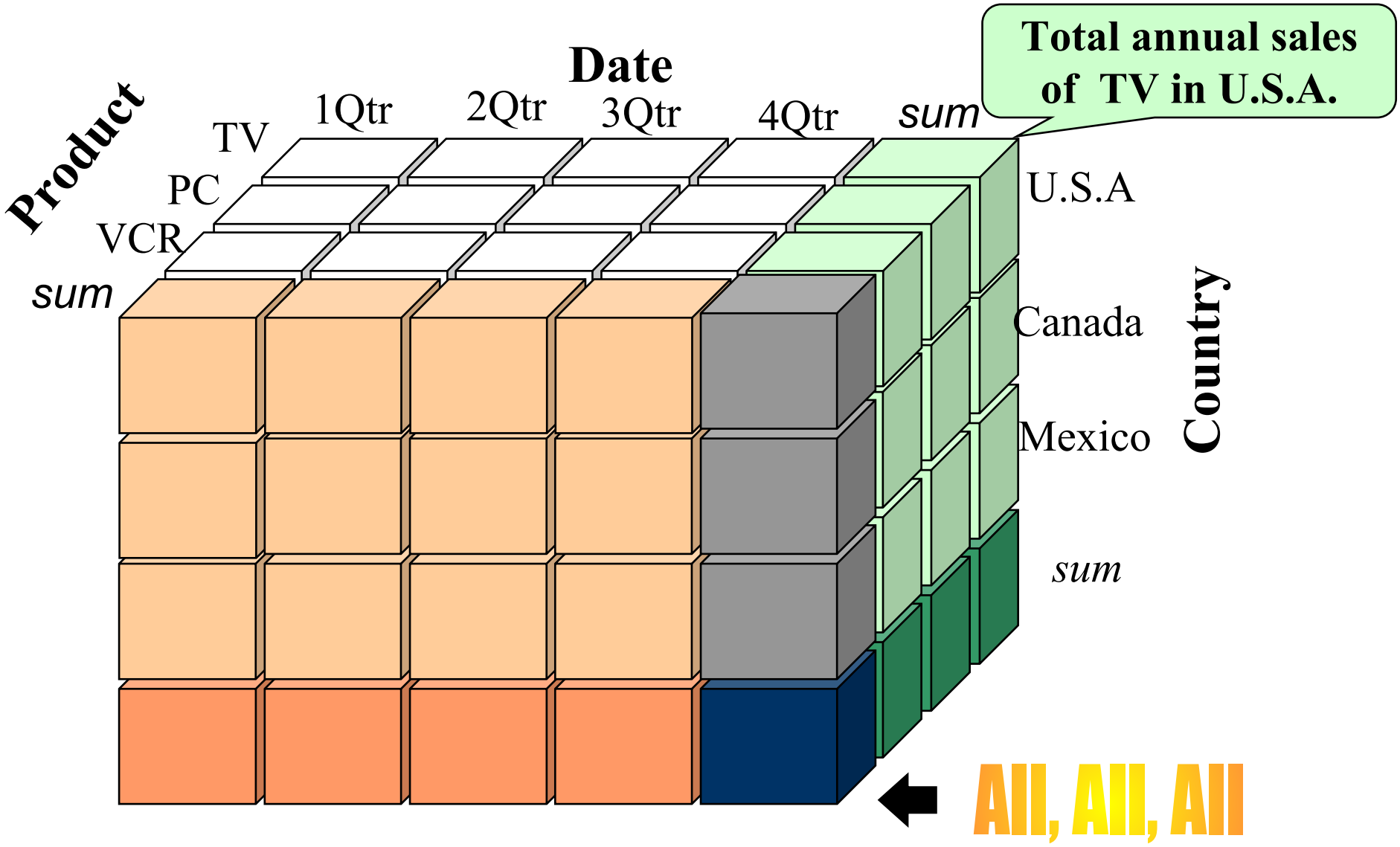
- Sales volume as a function of product, month, and region



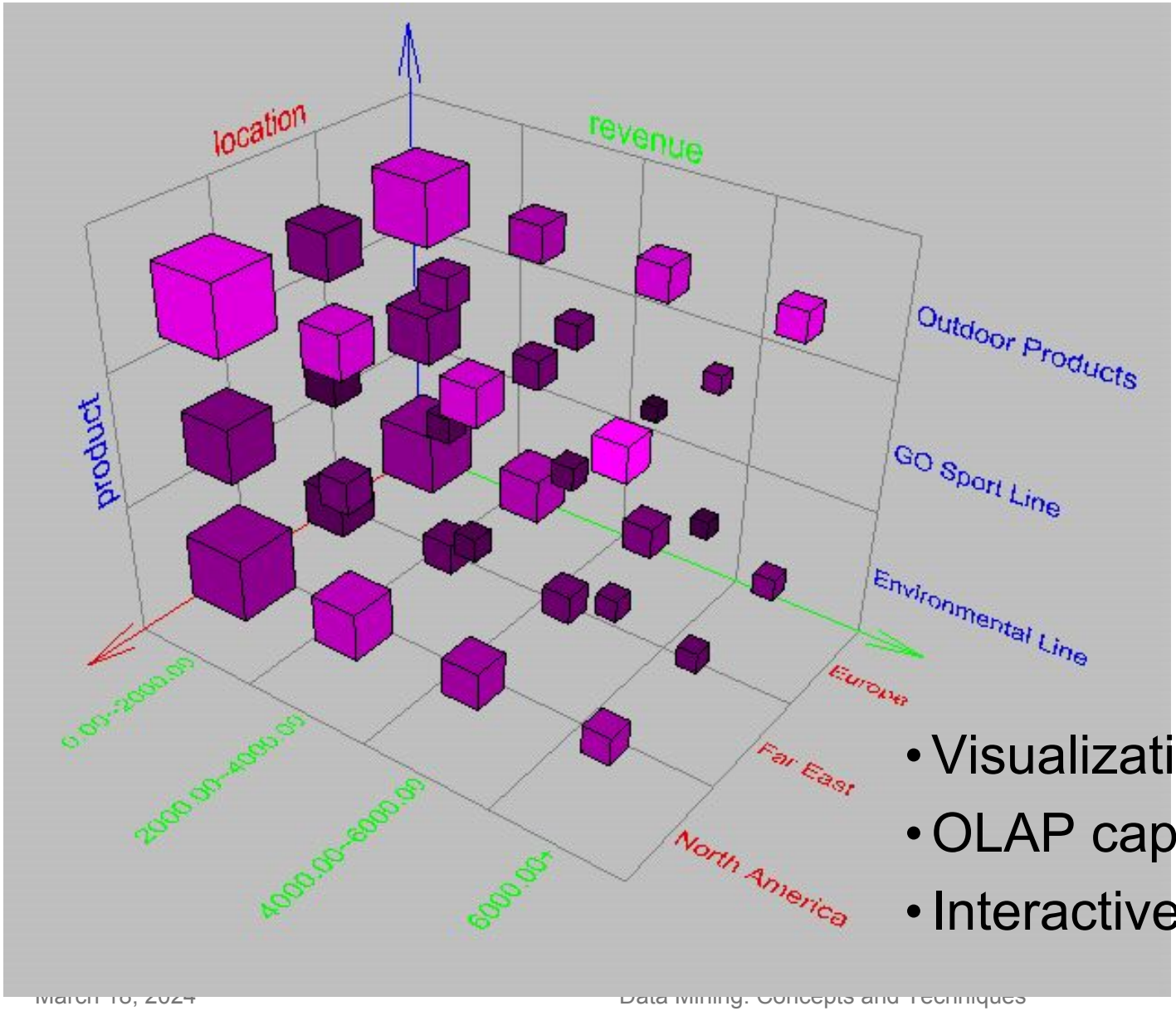
Dimensions: Product, Location, Time
Hierarchical summarization paths



A Sample Data Cube



Browsing a Data Cube

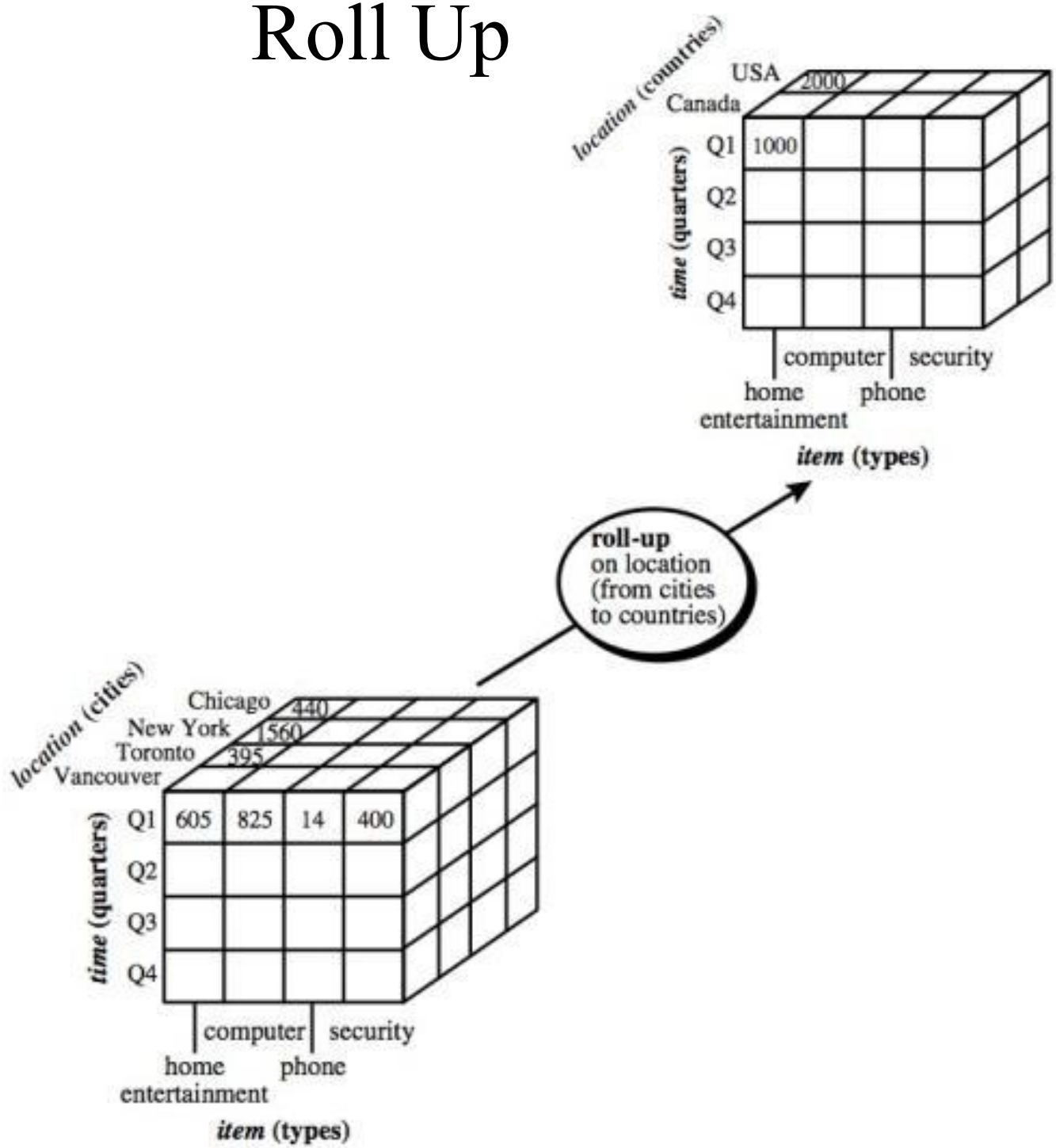


- Visualization
- OLAP capabilities
- Interactive manipulation

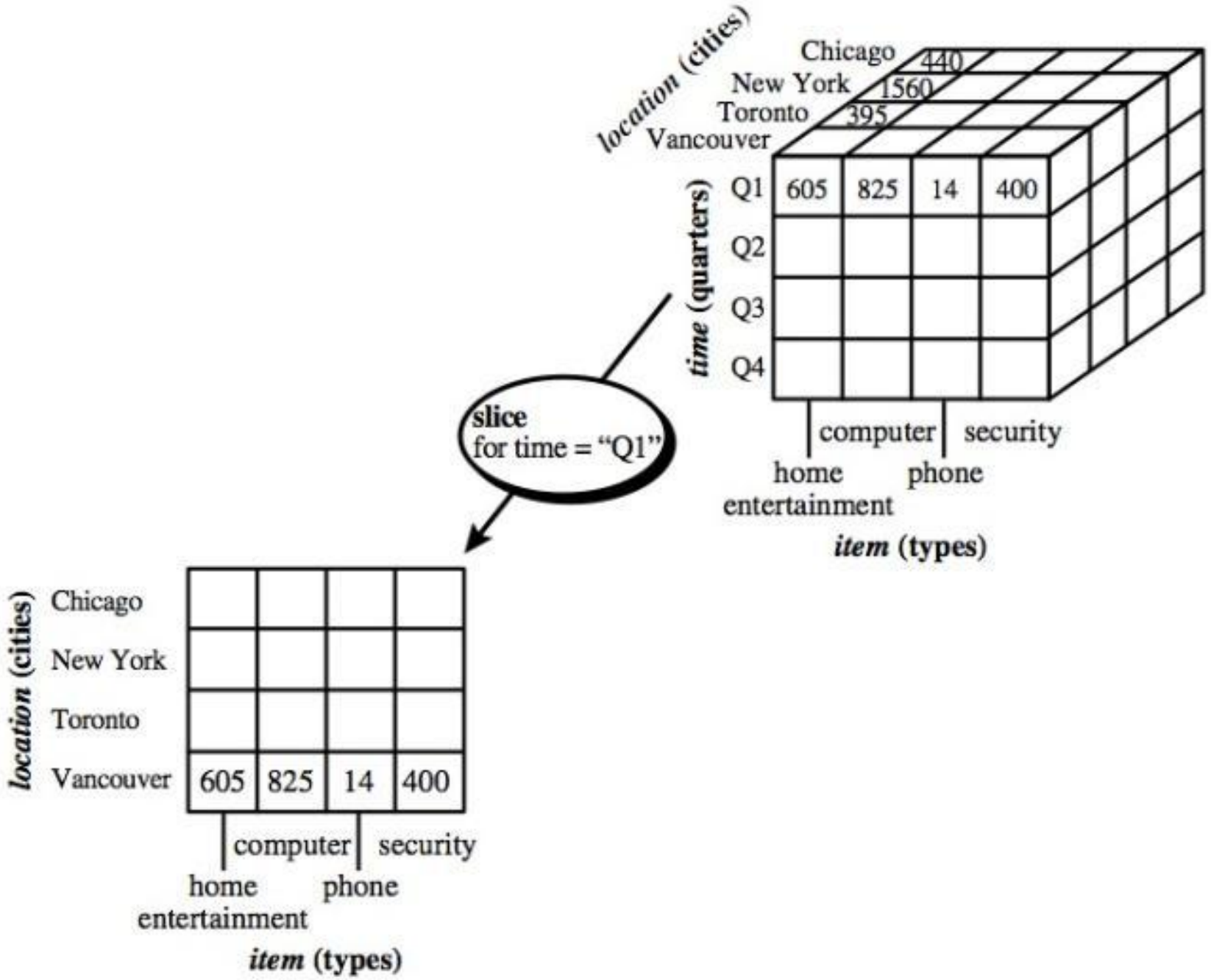
Typical OLAP Operations

- **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*
- Other operations
  - **drill across:** *involving (across) more than one fact table*
  - **drill through:** *through the bottom level of the cube to its back-end relational tables (using SQL)*

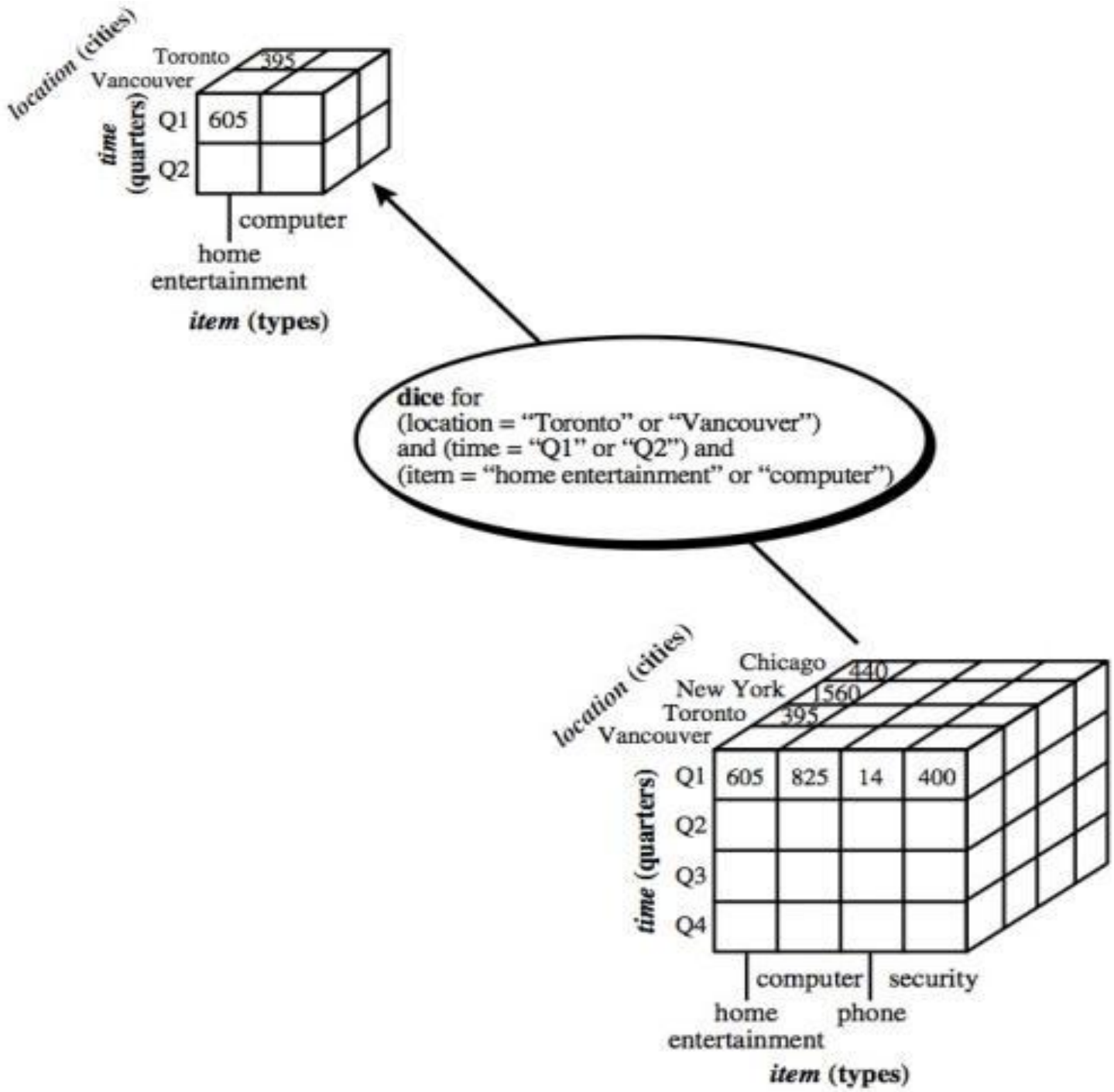
Roll Up


location (countries)
USA
Canada
2000
Q1
1000
Q2
Q3
Q4
time (quarters)
computer
security
home
entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
605
825
14
400
Q2
Q3
Q4
time (quarters)
computer
security
home
entertainment
phone
item (types)

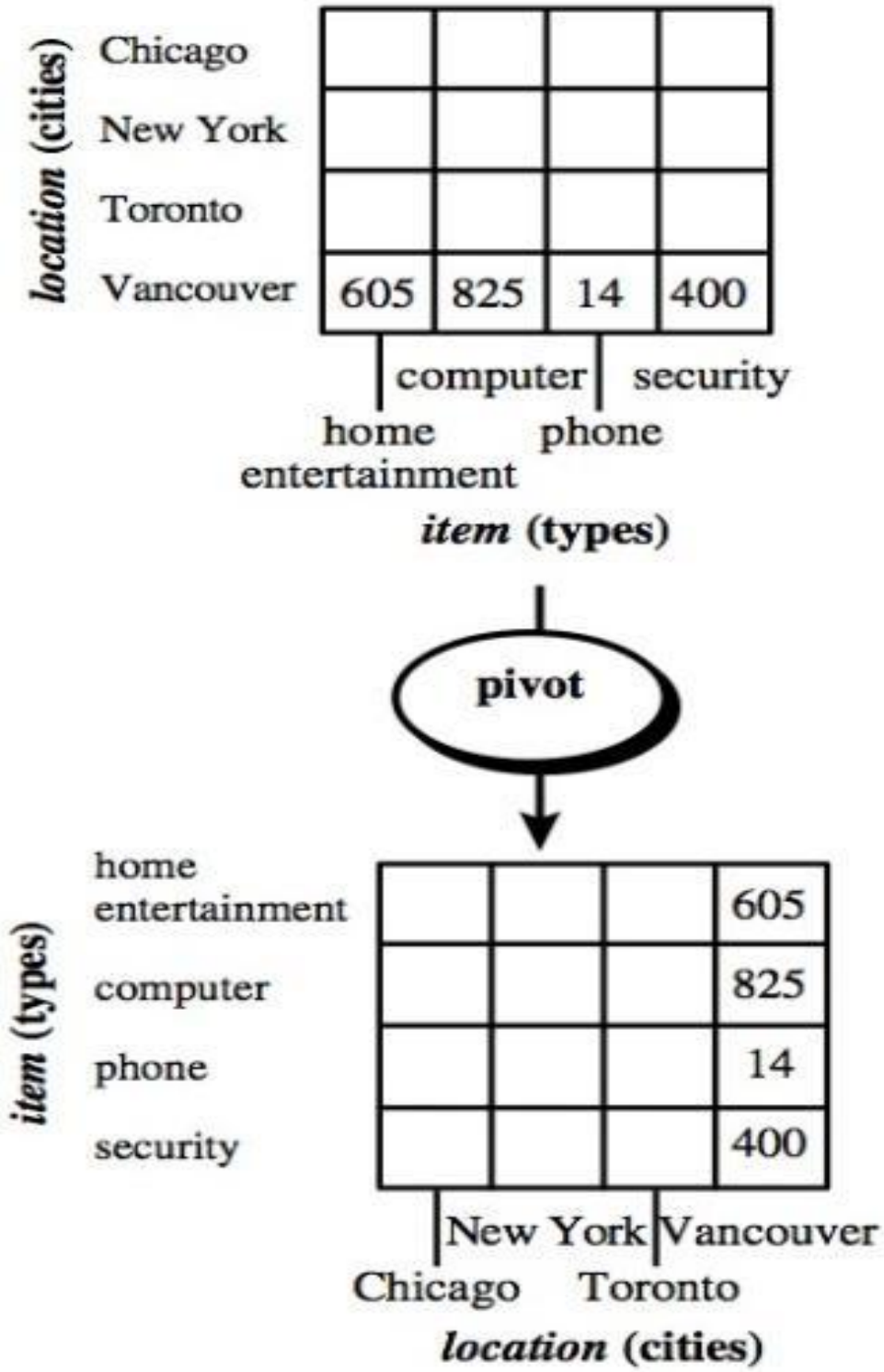
Slice


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)

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)

Pivot


location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)
pivot
home entertainment
computer
phone
security
605
825
14
400
item (types)
New York
Vancouver
Chicago
Toronto
location (cities)