Australian
National
University

COMP3430 / COMP8430
Data wrangling

Lecture 3: Data extraction and storage, data warehousing
(Lecturer: Peter Christen)

Lecture outline

- How to extract data
- Data storage
- Data warehousing

Data extraction

- The process of retrieving data out of data sources for further processing and storage
- There are various data sources, some internal some external to an organisation
- Unstructured data sources include emails, Web pages, PDFs, scanned and OCRed text (optical character recognition), audio reports (speech-to-text), etc.
- Different sources require different extraction methods
- Certain sources might be poorly structured or even unstructured
- The process of extracting data from the Web is called **Web scraping**

Extraction, transformation and loading (ETL)

- ETL is an integral part of data warehousing (more on DW later)
- **Extraction** involves retrieving data from disparate sources, such as transactional databases in an organisation or external sources
- In the **loading** phase, the extracted data are loaded into a staging (temporary) area of a data warehouse, where extraction logic (rules and pattern matching) are applied to ensure only suitable data are added to the warehouse
- In the **transformation** phase the selected data are transformed so they conform to the structure and formats of the data warehouse

Extracting data from PDF files

- Many documents online and within organisations are stored in the portable document format
- Documents often contain valuable information, such as tables with structured data, so extracting them might be required
  Note: Try to find the same data in a suitable format (for example as comma separated values, CSV, text file)
- Various PDF extraction tools, and modules/packages in different programming languages (often needs several modules in combination)
  – Python see: https://pypi.python.org/ (Python package index)
  – R see: https://cran.r-project.org/ (R package archive)

Data storage

- Various ways to store data: databases, data warehouses, document management systems, files (text, binary, multimedia, proprietary formats), cloud storage, etc.
- Data storage should be:
  - Persistent (over time)
  - Robust (redundant storage, RAID)
  - Secure (access regulated, distributed, cannot be manipulated)
  - Consistent and normalised
  - Available (with high performance)
- Often: *Garbage-in garbage-out* principle

Data warehousing (1)

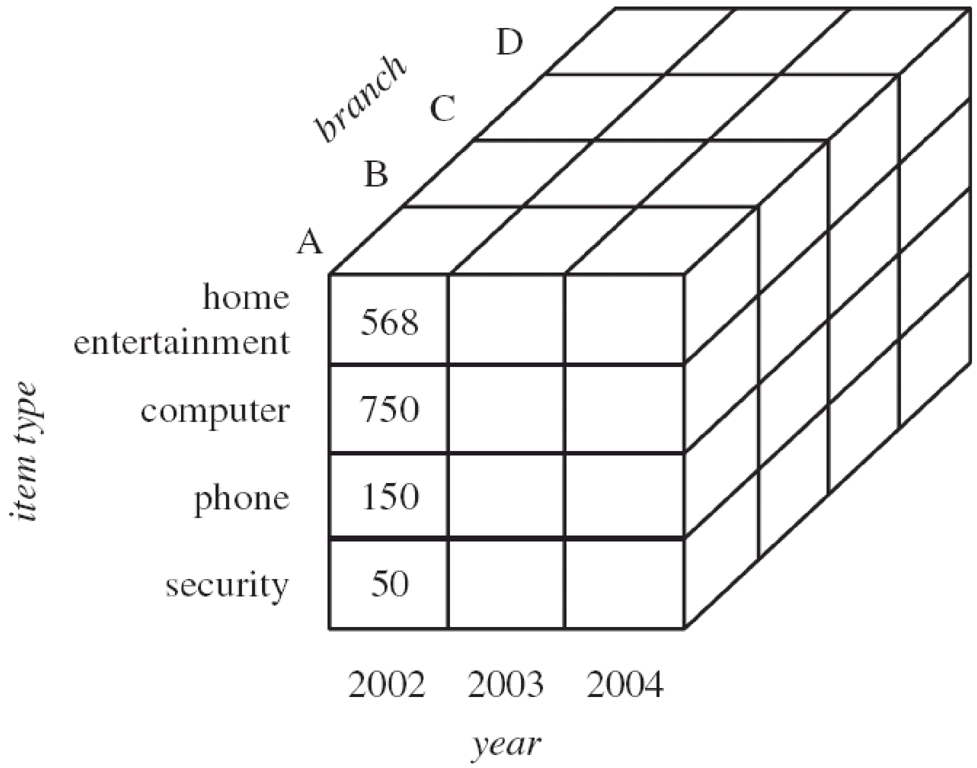
- A data warehouse is a decision support database that is maintained separately from an organisation's operational databases(s)
- Provides a solid platform of consolidated, historical data for analysis and mining
- Organised around major subjects, like customers, products, or sales (provides a simple and concise view around these entities)
- Often constructed by cleaning, standardising and integrating multiple heterogeneous data sources
  - To ensure consistency in coding, naming, measurements, etc.

Data warehousing (2)

- Longer time horizon than operational systems (that are used for transaction processing)
  - Historical data are important for analysis and mining
  - Separate data warehouse due to performance, data representation, consistency, integration, and data quality
  - Databases: OLTP (On-Line Transaction Processing)
  - Data warehouses: OLAP (On-Line Analytic Processing)
- Contains a time element
  - New data are, for example, loaded into a data warehouse every week
- Only two operations: *Initial loading* and *querying* of data (read)
  - While transaction processing systems have *reads*, *writes* and *updates*

Data warehousing (3)

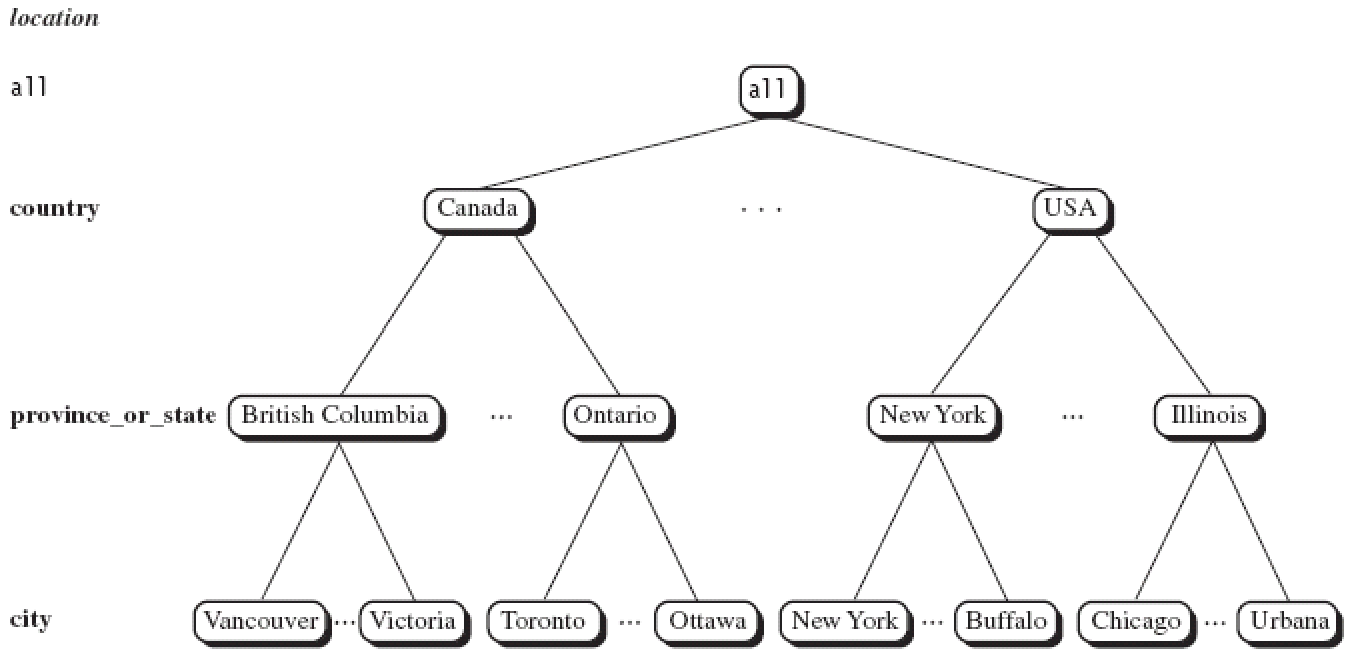
- Data warehouse architecture
  - Data cubes (multi-dimensional aggregated data views)
  - Dimension tables (details of the dimensions) and fact tables (values and names of the facts, e.g. *items_sold*, as well as keys into dimension tables)
  - Data are stored at different levels of details (e.g. *country* / *state* / *city*, or *item* / *item_group* / *item_category*



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (4)

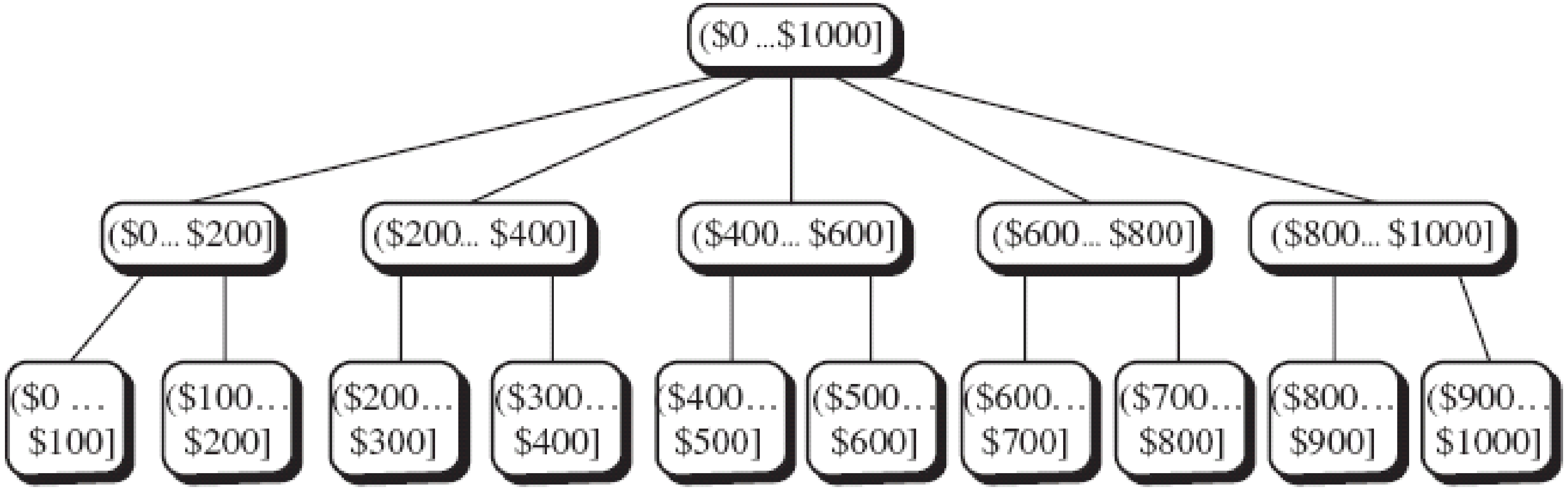
- Concept hierarchies
  - Defines a sequence of mappings from a set of low-level concepts to higher-level, more general, concepts



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (5)

- Concept hierarchies can be created by discretising or grouping numerical values



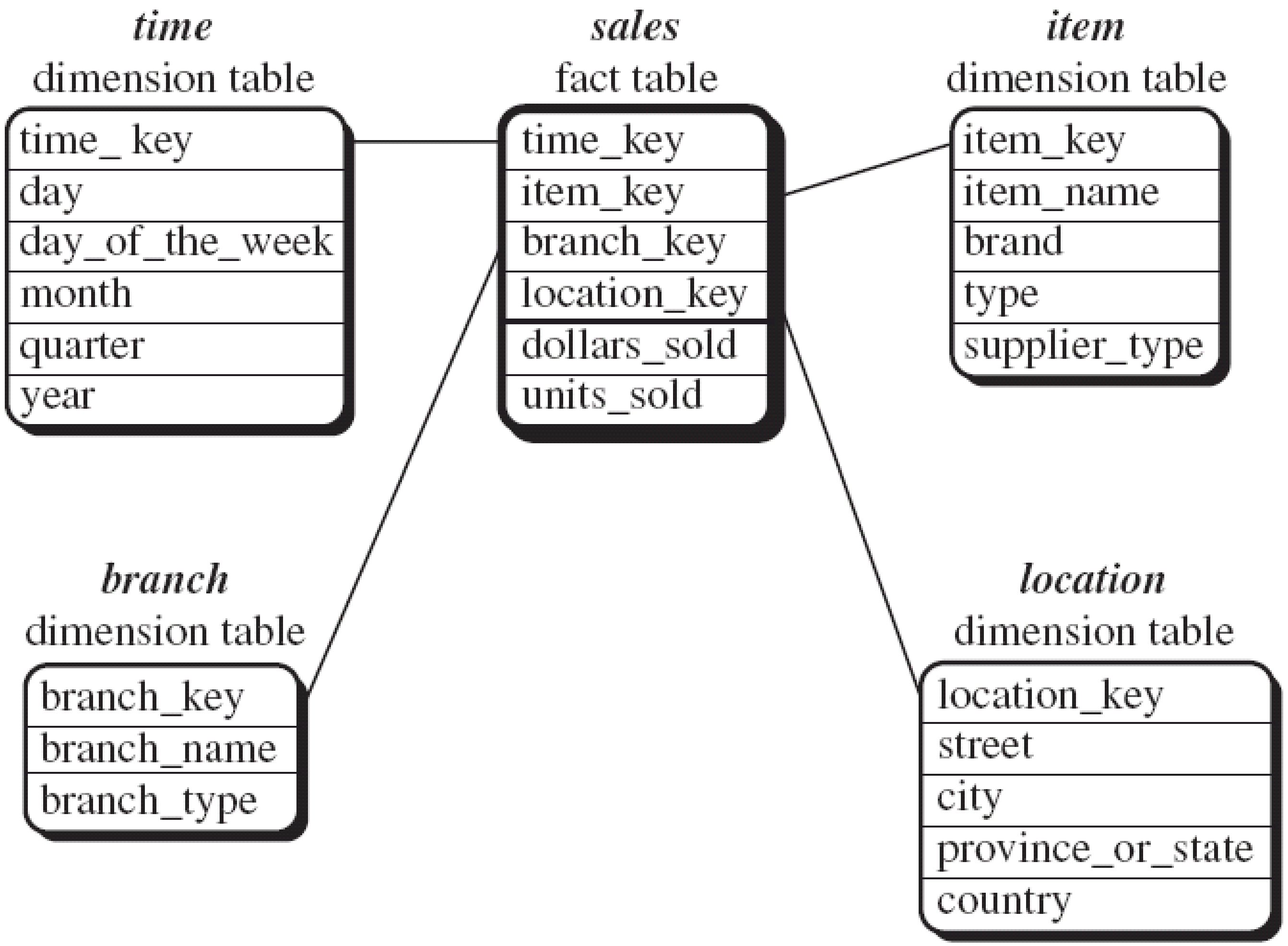
Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehousing (6)

- For data warehouses, a *multi-dimensional data model* is most popular
  - Compared to *entity-relationship model* for relational databases

- Implemented as:
  - **Star schema** (a large central *fact* table containing bulk of the data, and a set of smaller *dimension* tables)
  - **Snowflake schema** (variant of star schema with normalised dimension tables)
  - **Fact constellation schema** (multiple fact tables who share dimension tables), can be viewed as a collection of star schemas

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

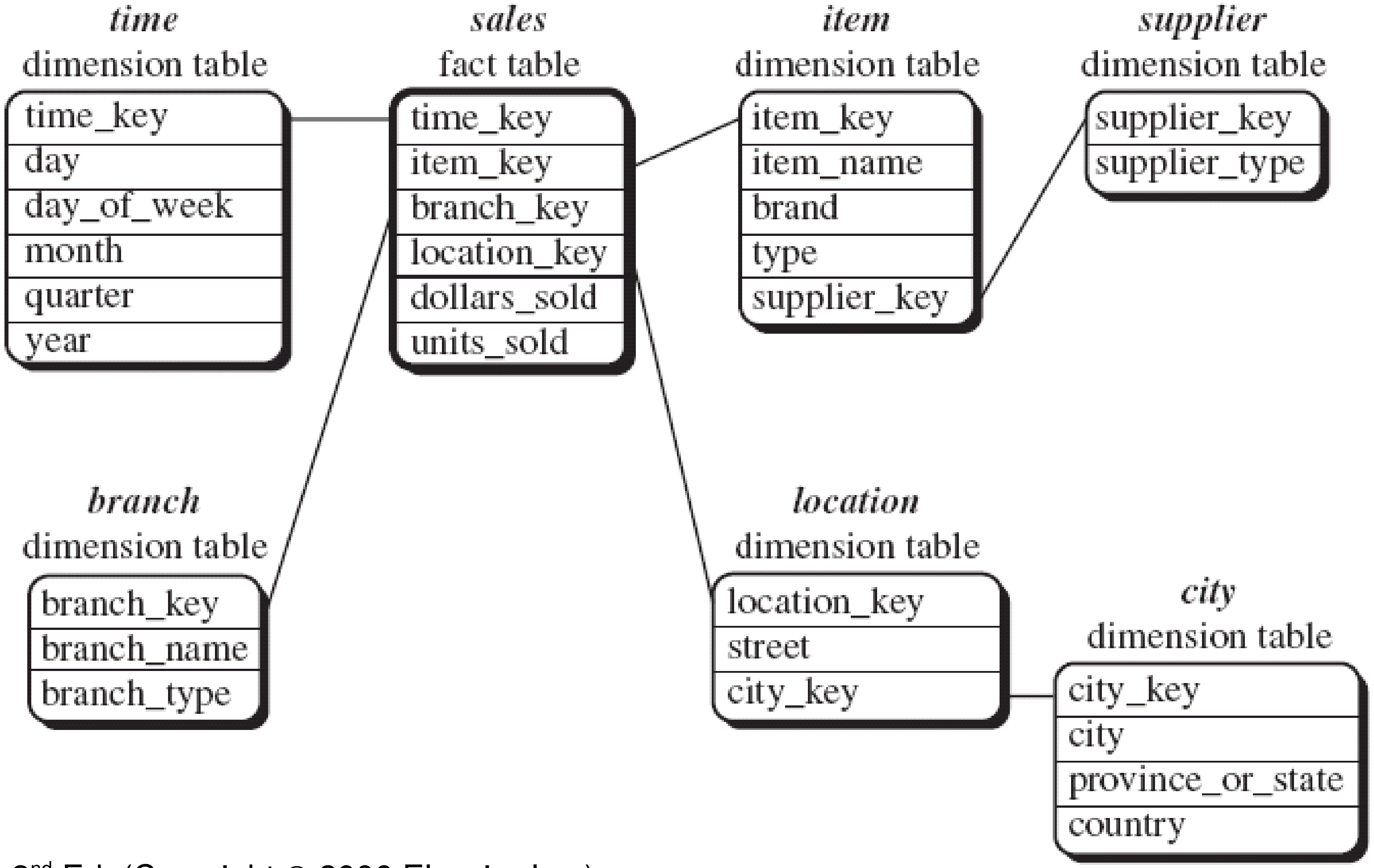
branch_key
branch_name
branch_type

location
dimension table

location_key
street
city
province_or_state
country

Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Snowflake schema



time
dimension table

- time_key
- day
- day_of_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_key

supplier
dimension table

- supplier_key
- supplier_type

branch
dimension table

- branch_key
- branch_name
- branch_type

location
dimension table

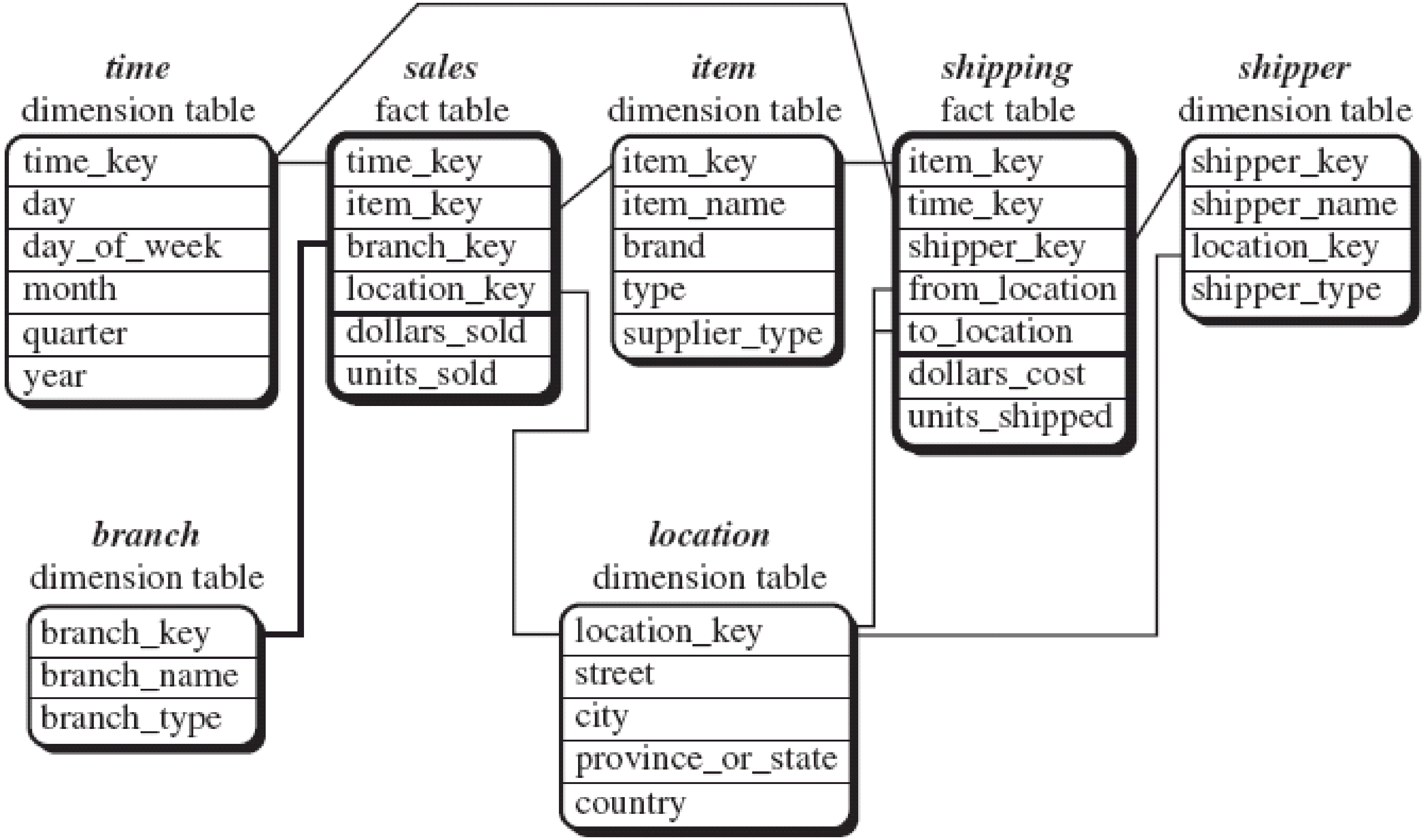
- location_key
- street
- city_key

city
dimension table

- city_key
- city
- province_or_state
- country

Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Fact constellation schema



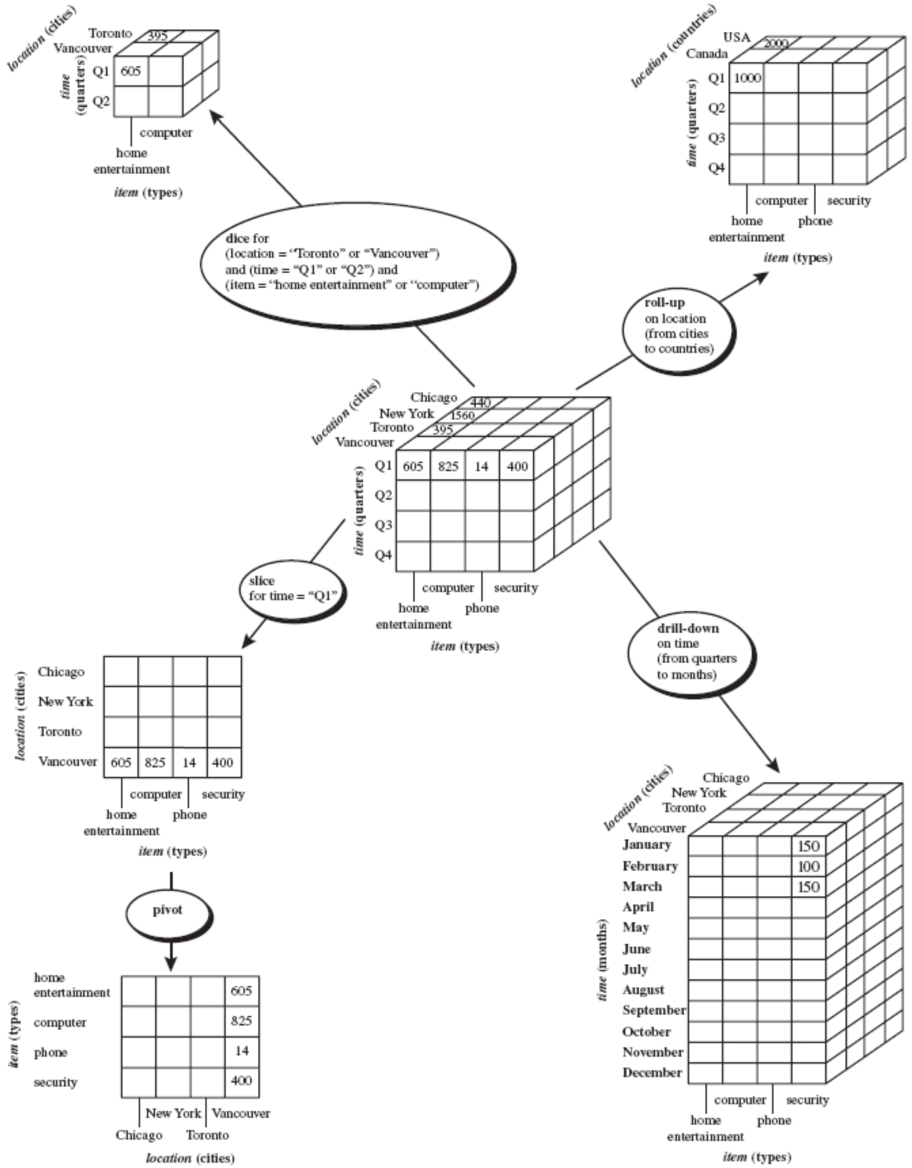
Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehouse operations (1)

- Data warehouse operations
  - **Roll-up** (summarise data)
  - **Drill-down** or **roll-down** (get detailed view)
  - **Slice** and **dice** (project and select)
  - **Pivot** (rotate), re-orient the cube, 2D to 2D visualisation

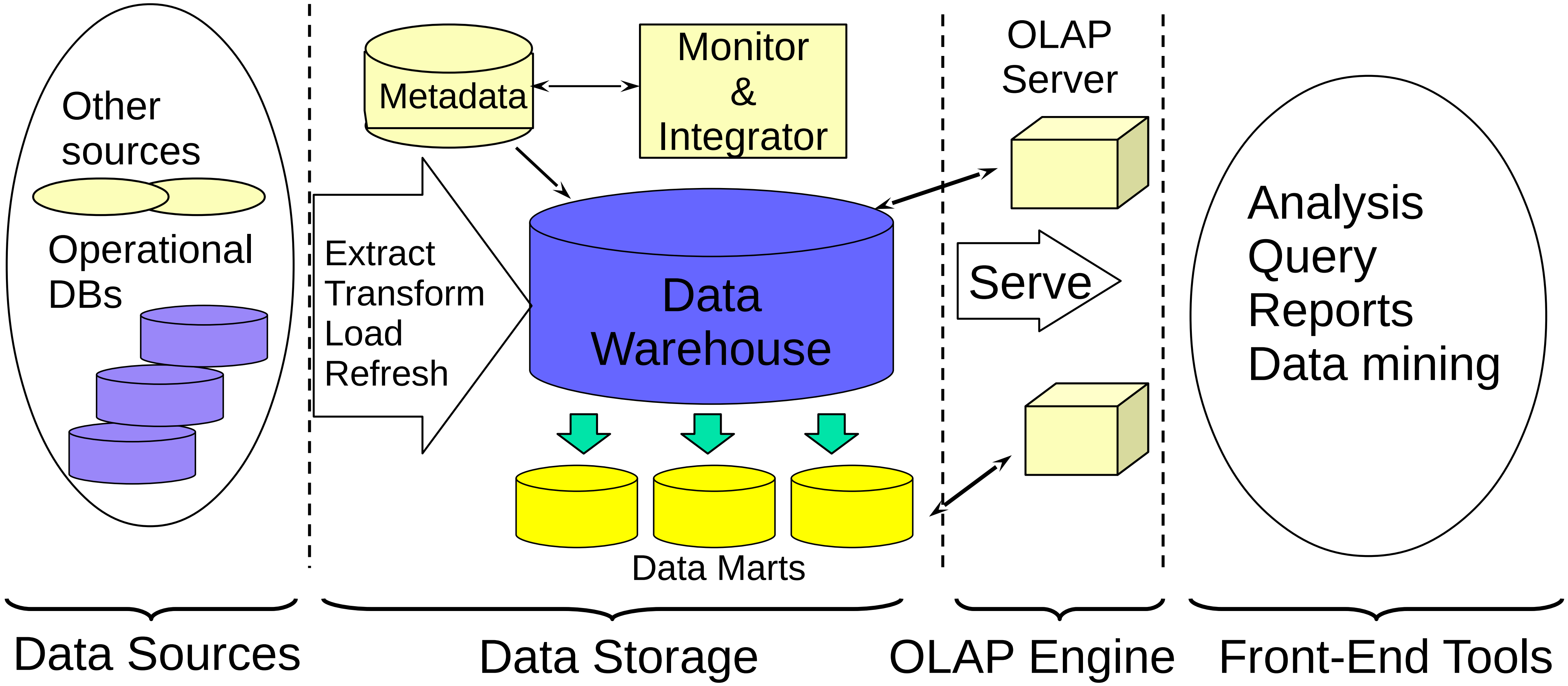
- Example applications of data warehousing:
  - Information processing (basic statistics, reporting, tables, charts, graphs, Web-based reporting, etc.)
  - Analytic processing (further drill down, multi-dimensional analysis, on both summarised and detailed data)
  - Data mining: A clean, stable, high-quality source for data mining algorithms

Data warehouse operations (2)



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data warehouse architecture



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)