بسم الله الرحمن الرحيم

Information Technology Fundamentals

Mohammad Hossein Manshaei

manshaei@gmail.com

Database: Data Warehouse
Module 6: Part 2

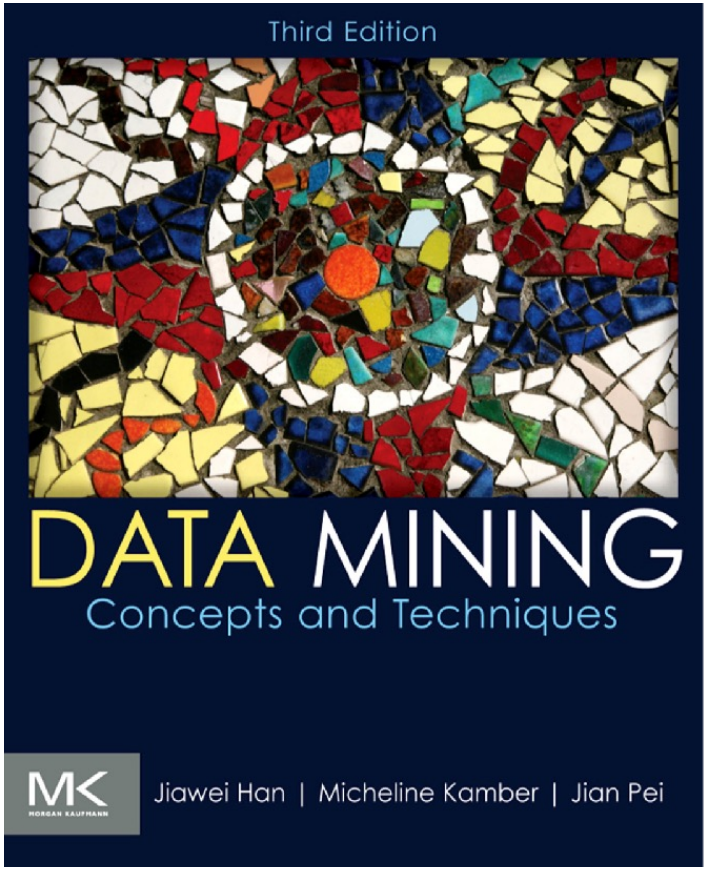
Module 6. Main Objectives

1. Review Business Intelligence Concepts
2. **Explain Data Warehouse**

Introduction to DataWarehouse

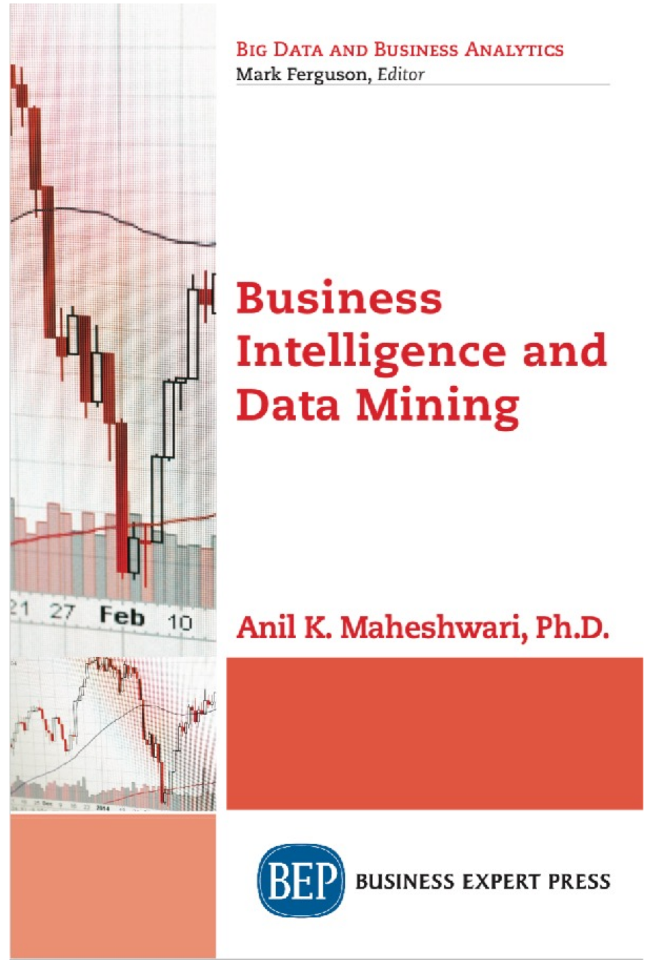
Main Reference

- Han, J., Kamber, M. and Pei, J., 2011. **Data mining concepts and techniques third edition**. Morgan Kaufmann.



Main Reference

- Maheshwari, A., 2014. **Business Intelligence and Data Mining**. Business Expert Press.



Contents

- Data Warehouse Definition
- Data Warehouse vs Database System
- Data Warehouse Architecture
- Data Cube and OLAP Operations

What is a Data Warehouse?

- A data warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to support decision support functions.
- DW is organized at the right level of granularity to provide clean enterprise- wide data in a standardized format for reports, queries, and analysis.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.

Major features of a data warehouse

- **Subject-oriented**: A data warehouse is organized around major subjects such as customer, supplier, product, and sales.
- **Integrated**: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
- **Time-variant**: Data are stored to provide information from an historic perspective.
- **Nonvolatile**: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.

Main Goals of Data Warehousing and Business Intelligence

- Must make information easily accessible.
- Must present information consistently.
- Must adapt to change.
- Must present information in a timely way.
- Must be a secure bastion that protects the information assets.
- Must serve as the authoritative and trustworthy foundation for improved decision making.

Contents

➢ Data Warehouse Definition
➢ Data Warehouse vs Database System
➢ Data Warehouse Architecture
➢ Data Cube and OLAP Operations

Operational Database System Vs Data Warehouses

- **Online transaction processing (OLTP) systems:** The major task of online operational database systems is to perform online transaction and query processing.
- **Online analytical processing (OLAP) systems:** Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users.

Differences between Operational DBS and DW

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	historic, 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	GB to high-order GB	$\geq$ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Function	Database	Data Warehouse
Purpose	Data stored in databases can be used for many purposes including day-to-day operations	Data in data warehouse is cleansed data, which is useful for reporting and analysis
Granularity	Highly granular data including all activity and transaction details	Lower granularity data; rolled up to certain key dimensions of interest
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields	Typically organized around a large fact tables, and many lookup tables
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size	Grows as data from operational databases is rolled up and appended every day. Data is retained for long-term trend analyses
Architectural choices	Relational, and object-oriented, databases	Star schema or Snowflake schema
Data access mechanisms	Primarily through high-level languages such as SQL. Traditional programming access database through Open Database Connectivity (ODBC) interfaces	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

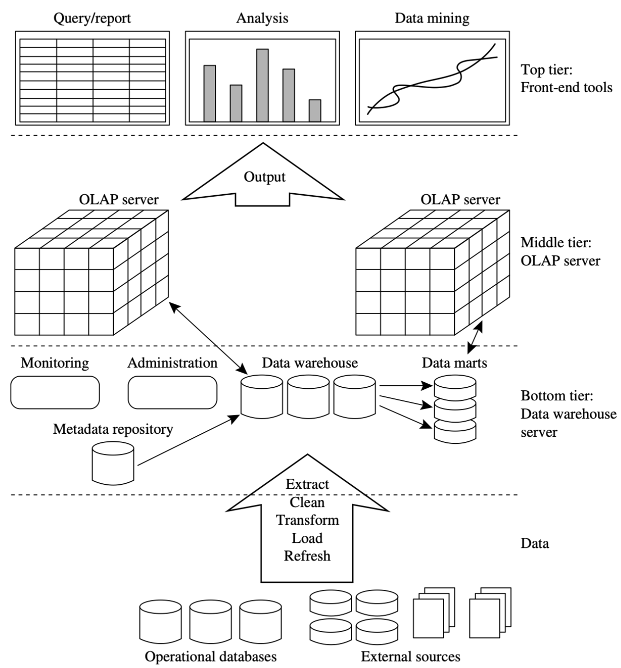
Contents

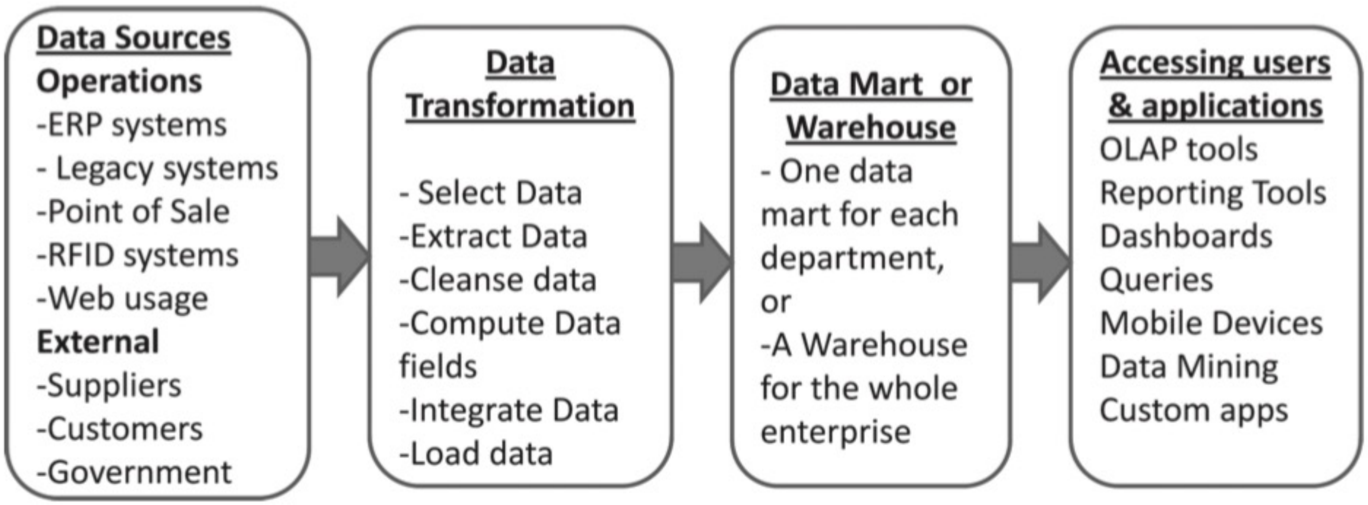
Three-Tier Data Warehousing Architecture

- Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources.
- The middle tier is an OLAP server that is typically implemented using:

 1. A relational OLAP (ROLAP) model
 2. A multidimensional OLAP (MOLAP) model.

- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.





Data warehousing architecture

Data Sources for DW

- DWs are created from structured data sources. Unstructured data, such as text data, would need to be structured before inserted into DW.
- Operations data include data from all business applications, including from ERPs systems that form the backbone of an organization's IT systems.
- The data to be extracted will depend upon the subject matter of DW.
- Other applications, such as point-of-sale (POS) terminals and e-commerce applications, provide customer-facing data. Supplier data could come from supply chain management systems. Planning and budget data should also be added as needed for making comparisons against targets.
- External data, such as weather or economic activity data, could also be added to DW, as needed, to provide good contextual information to decision makers.

Data Transformation Processes

- The heart of a useful DW is the processes to populate the DW with good quality data. This is called the **extract-transform-load (ETL)** cycle.
- Data should be extracted from many operational (transactional) database sources on a regular basis.
- Extracted data should be aligned together by key fields.
- It should be cleaned of any irregularities or missing values.
- It should be rolled up together to the same level of granularity.
- The transformed data should then be uploaded into DW.

Extraction, Transformation, and Loading

- **Data extraction**, which typically gathers data from multiple, heterogeneous, and external sources.
- **Data cleaning**, which detects errors in the data and corrects them when possible.
- **Data transformation**, which converts data from legacy or host format to warehouse format.
- **Load**, which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions.
- **Refresh**, which propagates the updates from the data sources to the warehouse.

Contents

Data Warehouse Modeling: Data Cube and OLAP

- A **data cube** allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- A multidimensional data model is typically organized around a central theme, such as *sales*. This theme is represented by a **fact table**.

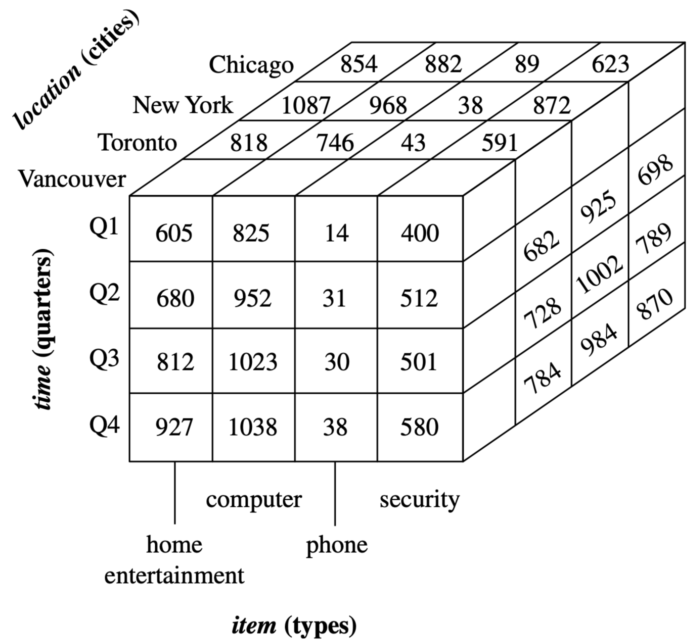
Example: 2-D View of Sales Data

	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

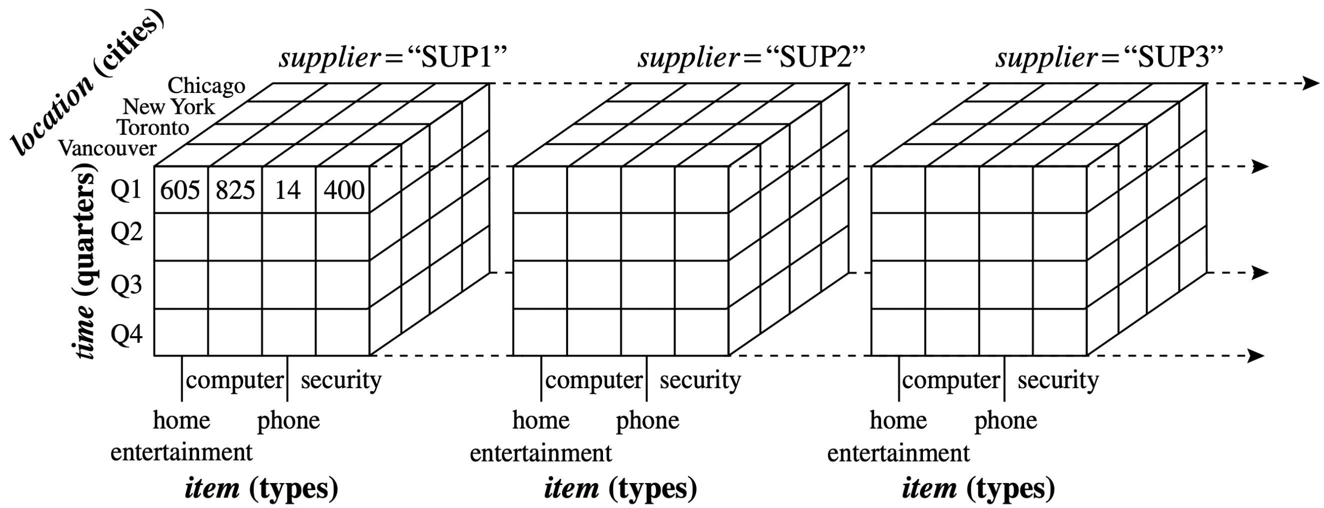
Example: 3-D View of Sales Data

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home 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

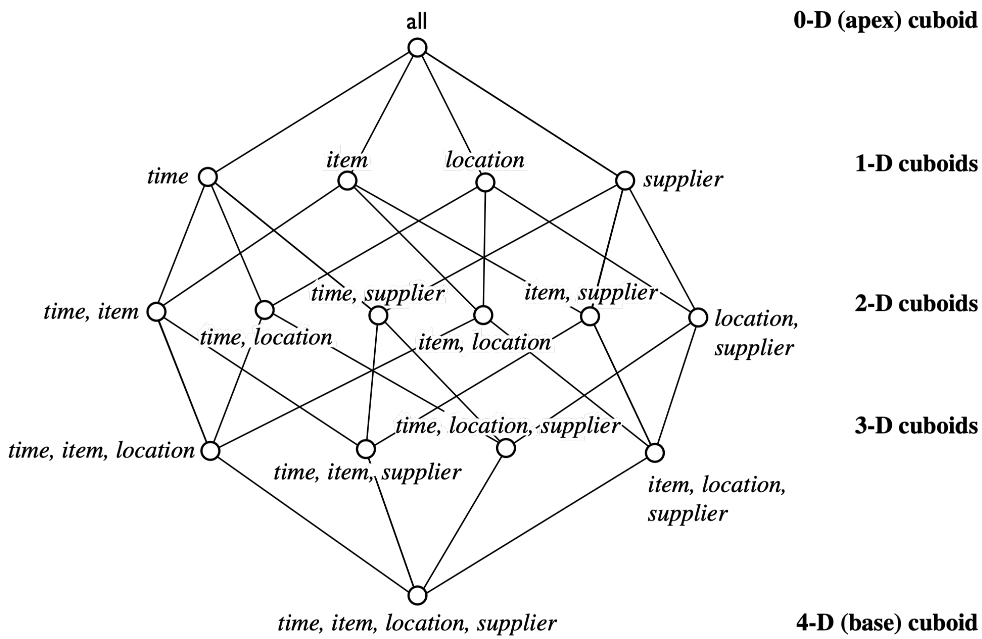
3-D Data Cube Representation



4-D Data Cube Representation

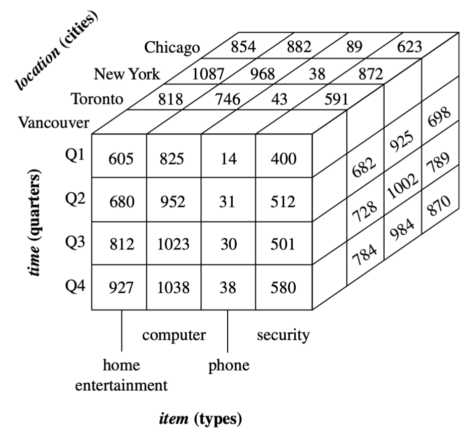


Lattice of Cuboids



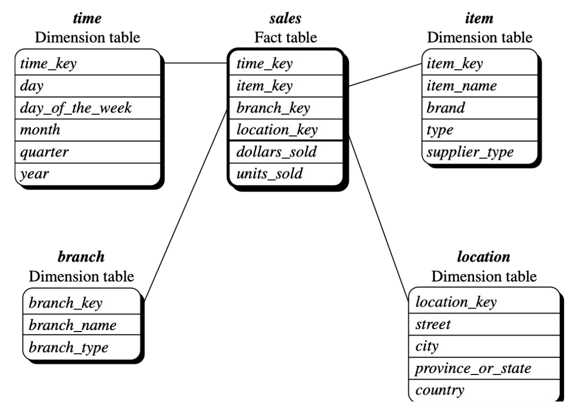
Schemas for Multidimensional Data Models

- The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them.
- The most popular data model for a data warehouse is a multidimensional model which can exist
  - Star schema
  - Snowflake schema
  - Fact constellation schema



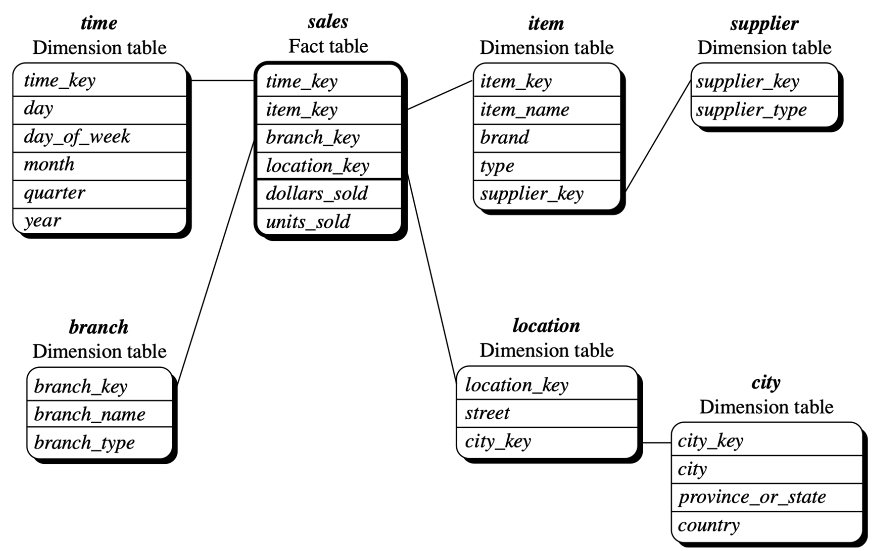
Star Schema:

- The most common modeling paradigm
- The data warehouse contains
  1. A large central table (**fact table**) containing the bulk of the data, with no redundancy.
  2. 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.



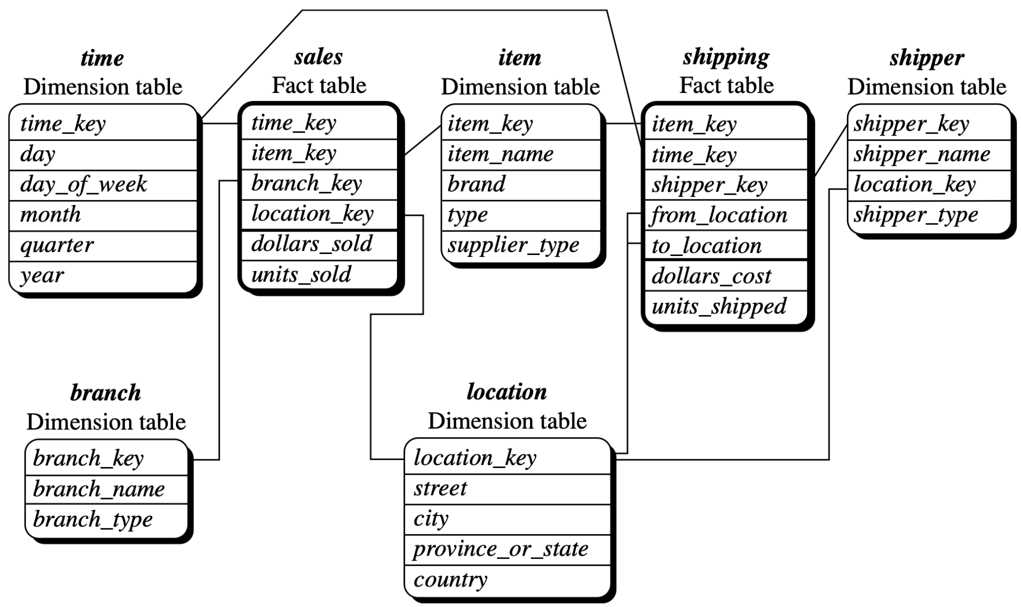
Snowflake schema

- A variant of the star schema model
- Some dimension tables are normalized, thereby further splitting the data into additional tables.
- The resulting schema graph forms a shape similar to a snowflake.



Fact Constellation Schema

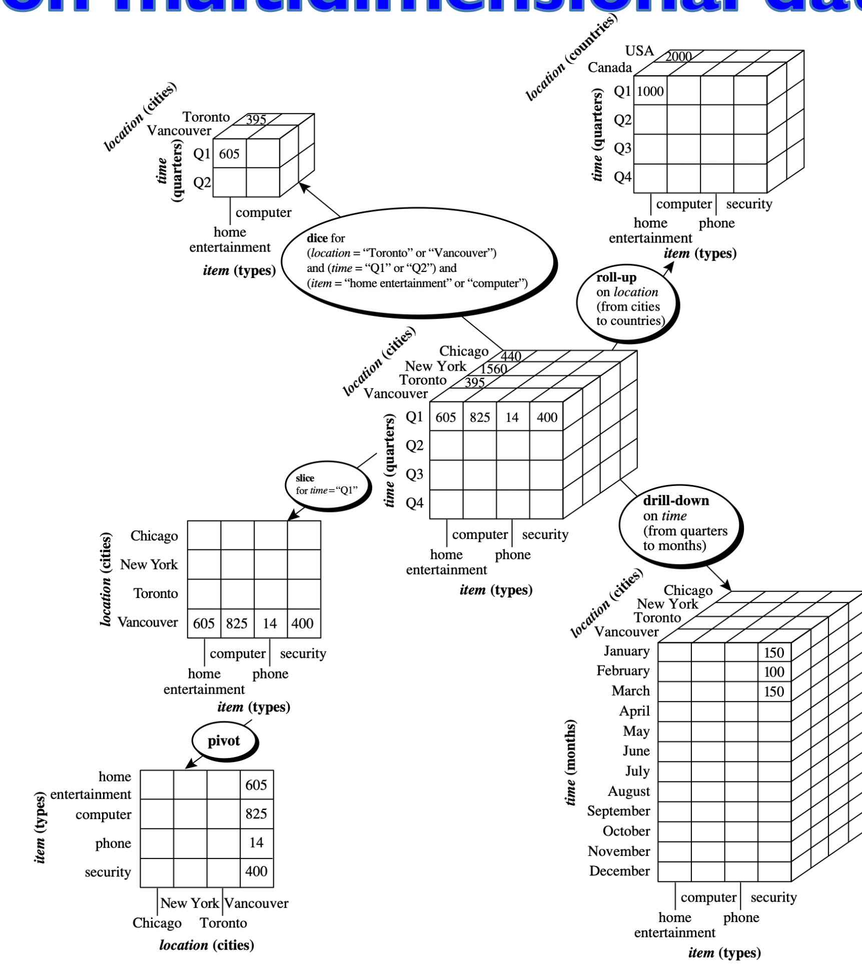
- Sophisticated applications may require **multiple fact tables** to share dimension tables.
- This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.



Typical OLAP Operations

- A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand.
  - Roll-up
  - Drill-down
  - Pivot (rotate)
  - Slice and Dice
  - Drill-across
  - Drill-through

Examples of typical OLAP operations on multidimensional data



Starnet Query Model

- **Starnet model:** consists of radial lines coming from a central point, where each line represents a concept hierarchy for a dimension
- Each abstraction level in the hierarchy is called a **footprint**. These represent the granularities available for use by OLAP operations such as drill-down and roll-up.