Data Warehousing and Data Mining

Lecture 2

Data Warehousing and Technology

Outline

- Introduction to **traditional data warehouses**
- Basic data warehouse **architecture**
- **Rationale** for using a data warehouse
- Data warehouse **implementation**
- Common **problems**

In the beginning…

- We have a business
- Businesses use databases for everyday operations
- We want to analyse the data without affecting business
- Data warehouse!

What is a Data Warehouse?

- A database **maintained separately** from operational database
- **Aggregate data** from many applications
- **Consolidate historical data**
- Used as a **base for analysis**

In short…

- "A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management's decision making process" - William H. Inmon
- Subject-oriented - Organised around **subjects**
- Integrated - Constructed from **many sources**
- Time-variant - Contains elements of **time** in dataset
- Nonvolatile - **Separated from operational** environment

Why use a Data Warehouse?

- We have lots of data, and we want to analyse **all** of it
- Queries on operational databases are **complex**!
- We don't want to touch **business transactions** (database locking, etc)
- We want to **preprocess** and store some attributes
- When we want to analyse it, we want to analyse it **fast**!

Terminology

- OLTP - Online Transaction Processing
  - “Operational” database
- OLAP - Online Analytic Processing
  - Platform for data analytics

Differences between traditional databases and data warehouses?

Table 3.1 Comparison between OLTP and OLAP systems.

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

NOTE: Table is partially based on [CD97].

Key difference

- Database **schema**
- Amount of **read vs write** operations
- Amount of **data stored**

Technologies used for data warehousing

- OLAP shows the users analytics, but how it is implemented underneath can vary
- Some implementations are
  - ROLAP - Relational OLAP
    - Use relational database as backend
  - MOLAP - Multidimensional OLAP
    - Use array-based multidimensional storage engine as backend (can be memory-based)
  - HOLAP - Hybrid OLAP
    - Combination of the above

Data warehouse architecture

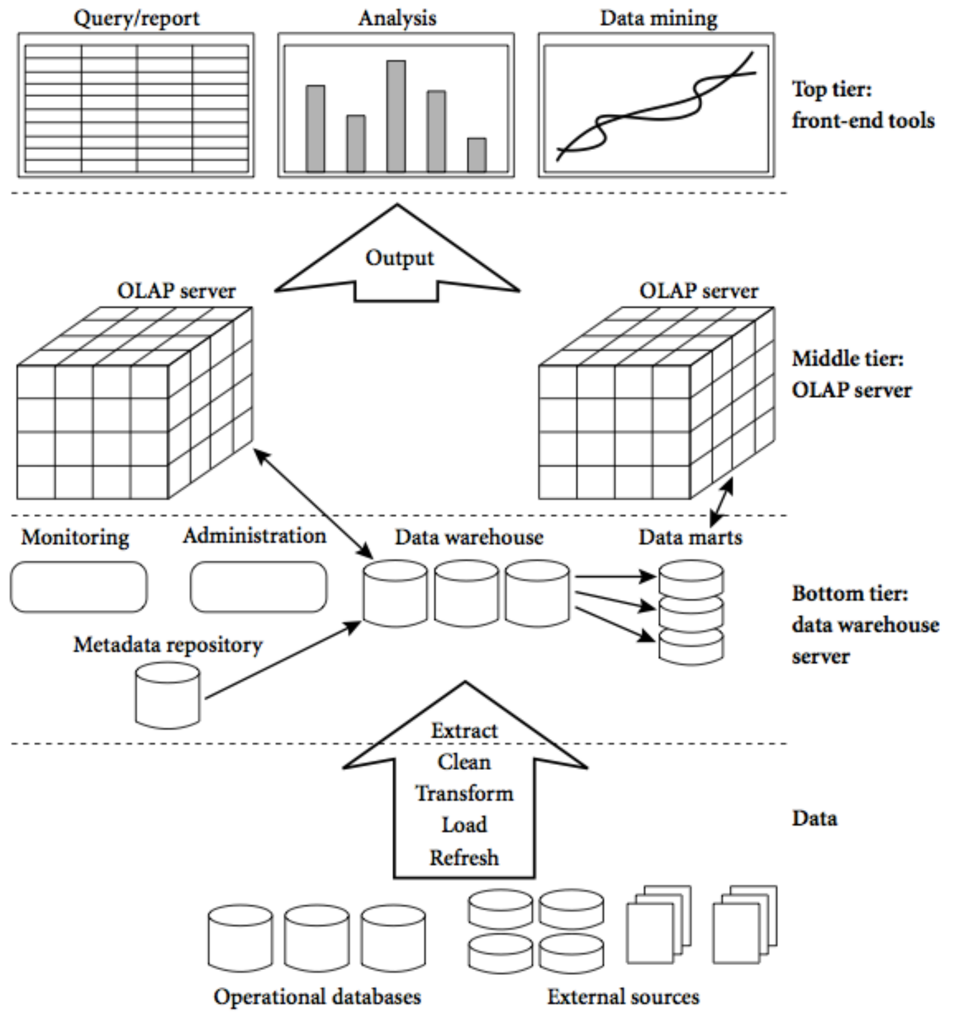
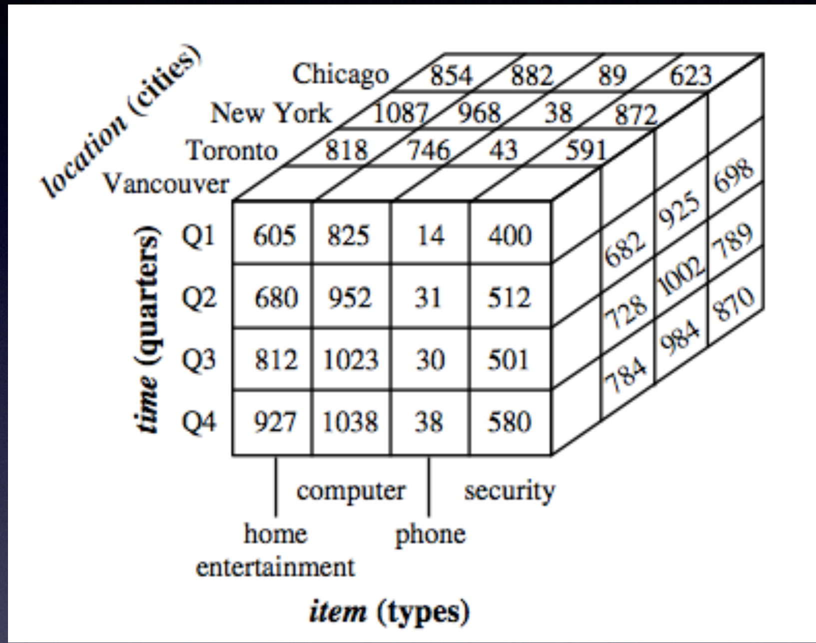
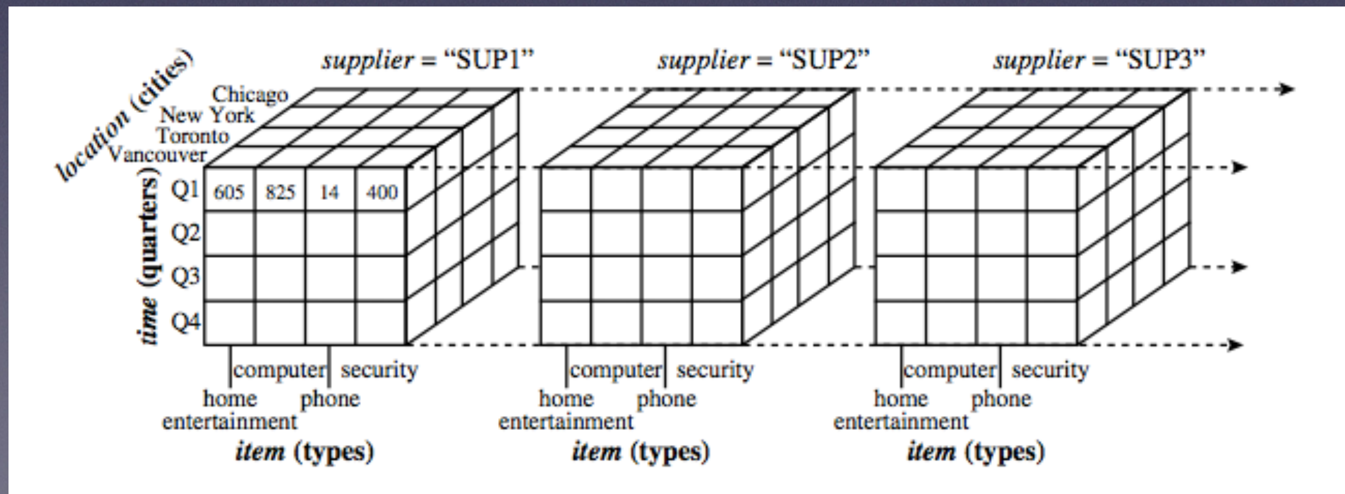


Figure 3.12 A three-tier data warehousing architecture.

What is a Data Cube?

- Method of conceptualising datasets in higher dimension
- High dimensional 'table'
- Can exceed 3 dimensions





How are data cubes represented in databases?

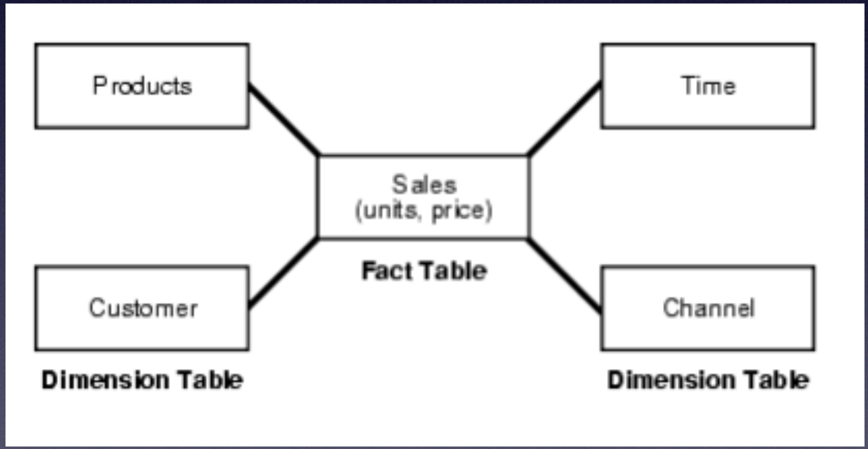
- By using different database schemas
  - Star Schema
  - Snowflake Schema
- Each ‘measure’ exists in one table
- Each dimension exists in other tables
- Joined together and aggregated when needed to generate reports

More terminology

- Fact table - table with the data we want to infer to
- Dimension table - table with the factors we want to infer from

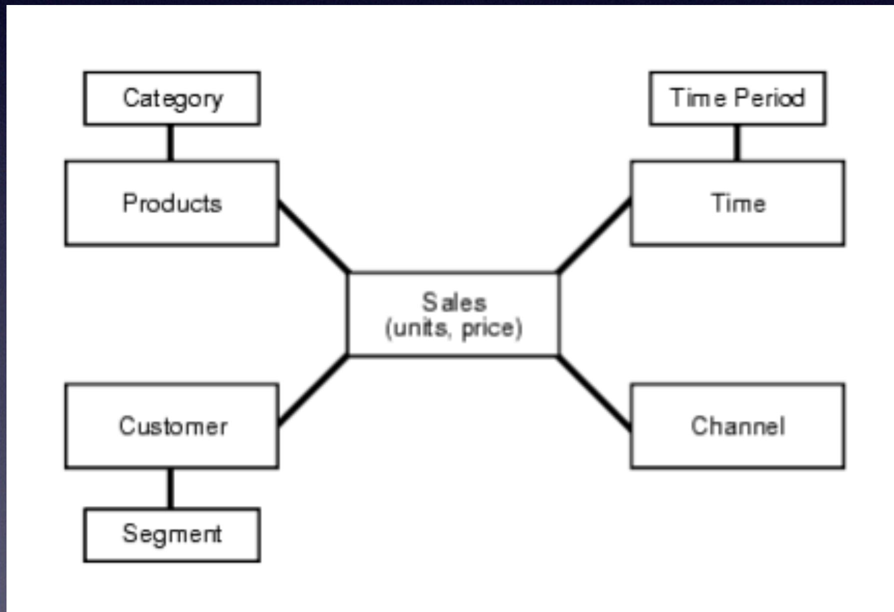
Star schema

- A single fact table
- Multiple dimension tables link directly to the fact table
- Dimension tables are **denormalised**

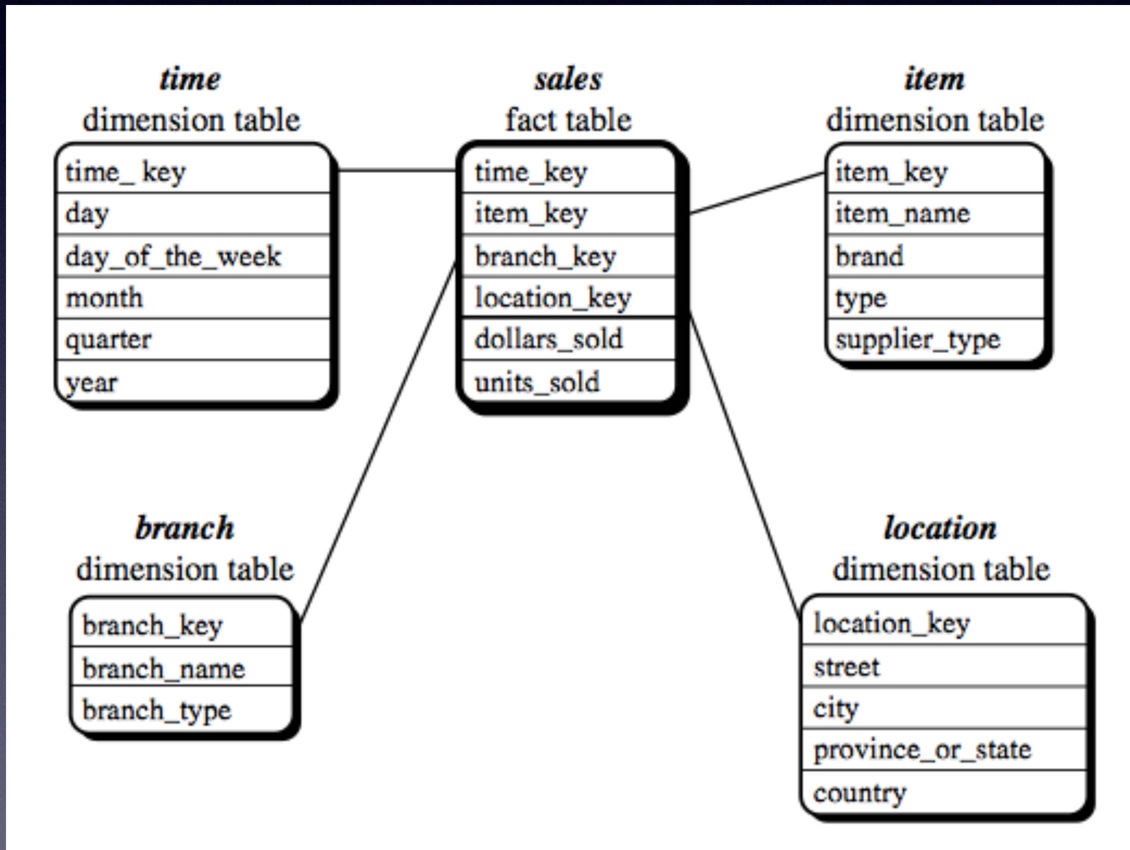


Snowflake schema

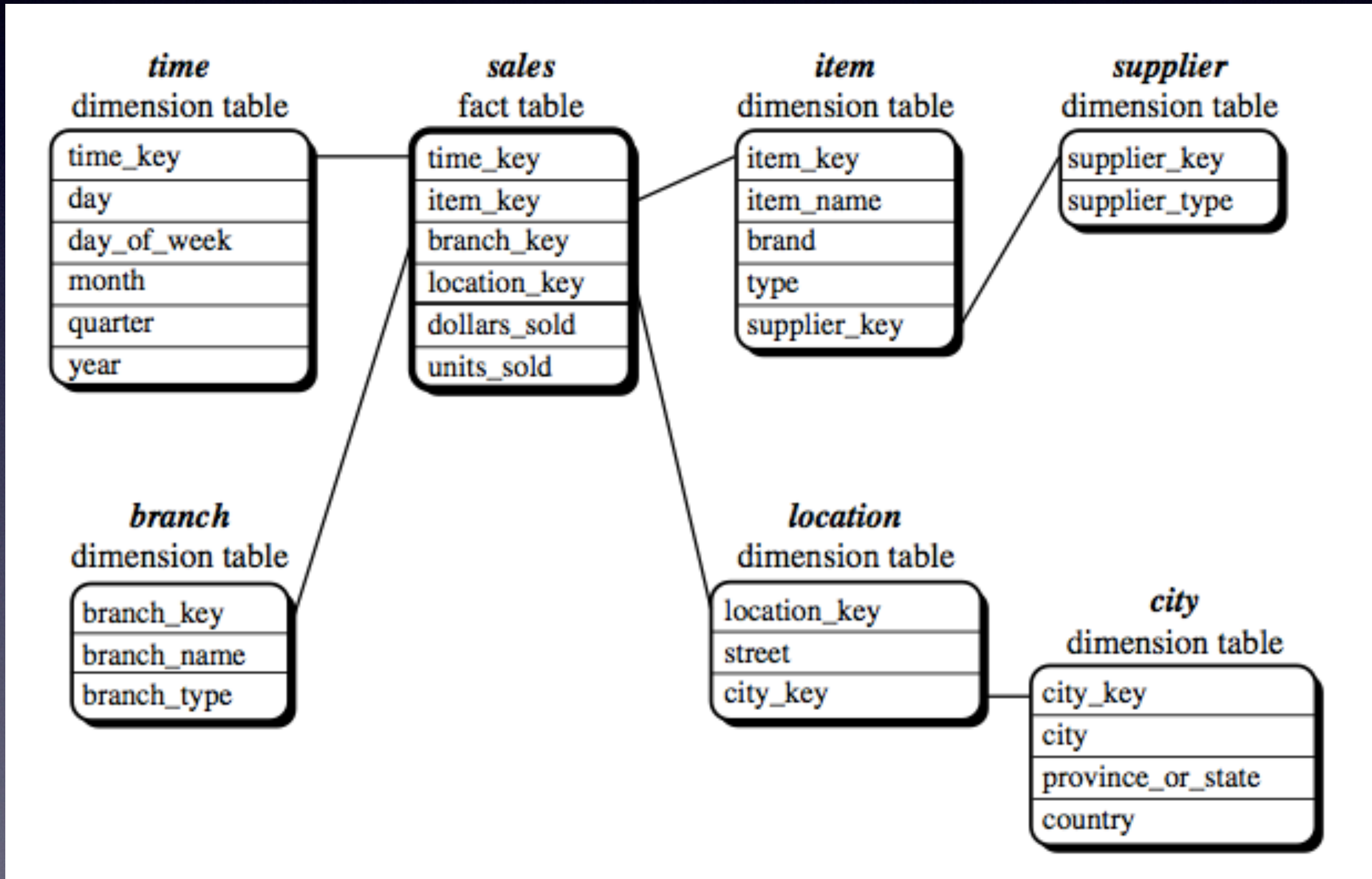
- A single fact table
- Multiple dimension tables link directly to the fact table
- Dimension tables are **normalised**



A more concrete example



A more concrete example



Comparison of schemas

	Star	Snowflake
Query complexity	Less	More
Redundancy	More	Less
Normalization	All tables are denormalized	Dimension tables may be normalized
When to use	Every other time	When dimension tables get big

*Note: Oracle recommends the Star schema, unless you have a good reason to use Snowflake

How to systematically analyse data cubes

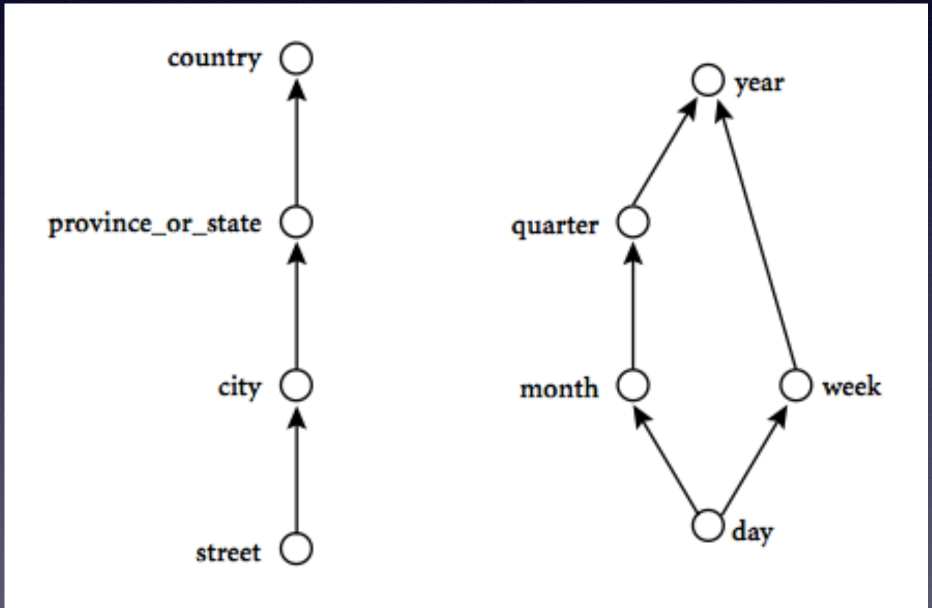
1. Extract, clean, load data into the warehouse
2. Compute **measures**
3. Build a **concept hierarchy**
4. Interactively **analyse** data in a systematic way (using OLAP)

Computing measures

- As we know, traditionally, a ‘measure’ is a value
- In a data cube, it represents a **numerical aggregation function**
- Three main categories
    - **Distributive** - can be computed in a **distributed manner**
        - eg. count, sum
    - **Algebraic** - can be computed using **algebraic functions**
        - eg. average, standard deviation
    - **Holistic** - can ONLY be computed by using the **whole dataset at once**
        - eg. mode, median

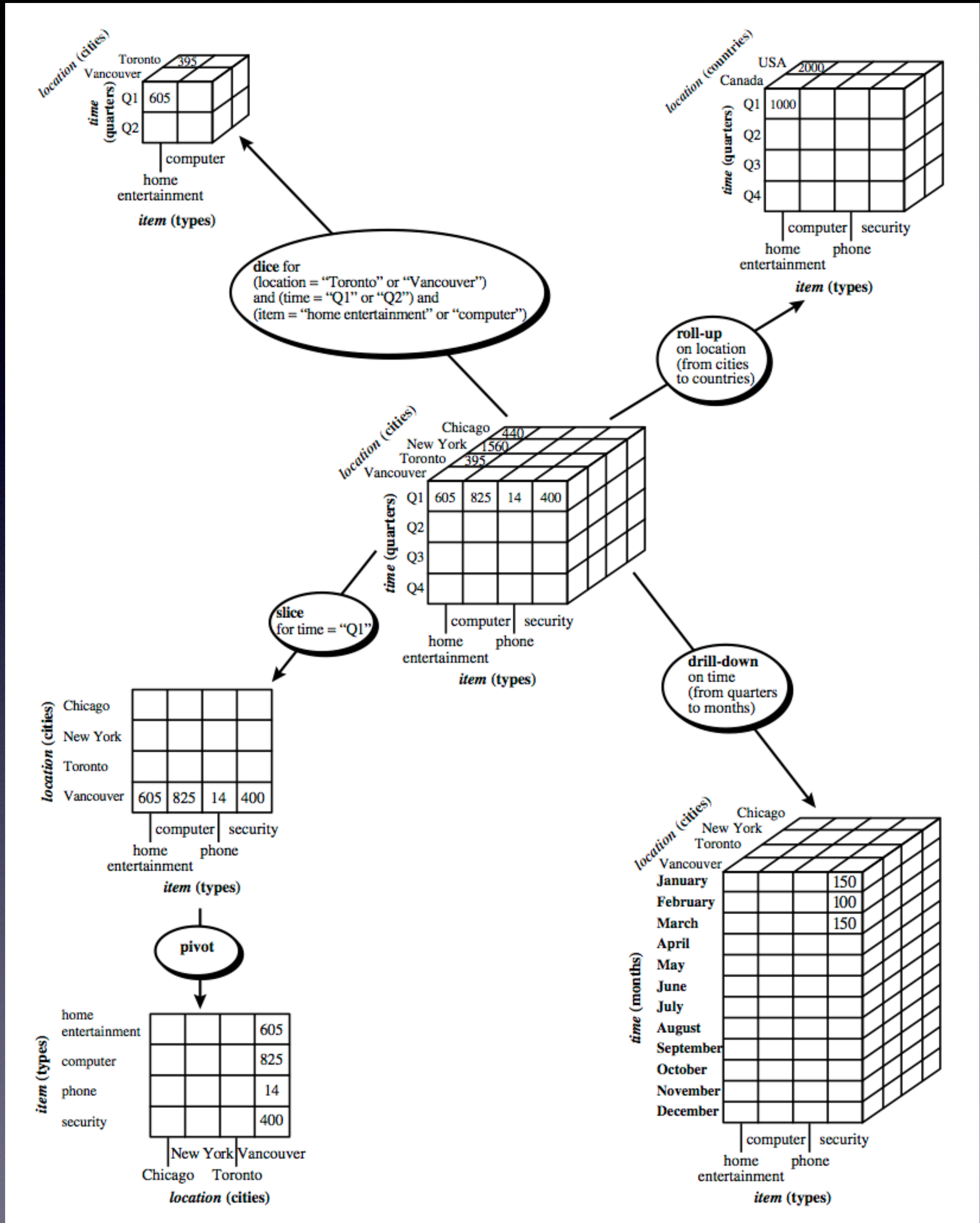
Concept hierarchy

- Different data will have **inherited structure**
- Constructing these hierarchies will
  - Help clarify your **understanding** of the data
  - Help **plan your approach** to analysing the data



OLAP operations

- **Roll-up** - Move from **small to big** in concept hierarchy
- **Drill-down** - Move from **big to small** in concept hierarchy
- **Slice** - Select **one dimension**
- **Dice** - Select a “**subcube**" of the data
- **Pivot** - “**rotate**” the axis of the data table or data cube



dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
roll-up
on location
(from cities
to countries)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
pivot
location (cities)
time (quarters)
item (types)
location (countries)
time (months)
Chicago
New York
Toronto
Vancouver
USA
Canada
Q1
Q2
Q3
Q4
home entertainment
computer
phone
security
January
February
March
April
May
June
July
August
September
October
November
December
605
825
14
400
395
1560
440
2000
1000
150
100
150

Constructing a data warehouse

1. Choose a **business process**
2. Choose the **granularity** of the data
3. Choose the **dimensions**
4. Choose the **measures**

Common issues with data warehousing

- Need to compute a lot of data for fact tables
  - The “Curse of dimensionality”
- Speed in joining data sets
  - Need efficient table “join” operations

Curse of dimensionality

- Problems occur when you have **too many attributes**
  - Combinatorics
  - Large feature space
  - Irrelevant data points
- These do not occur in low dimension datasets
- Can be solved by **dimensionality reduction techniques**!

Table ‘join’ operations

- When consolidating data for reports, we will need to fetch attributes from various tables
- Join operations consume a lot of computational power
  - This is reduced by the use of **indexes**

Notable (proprietary) vendors

ORACLE

Microsoft SQL Server 2014

STATA

SPSS

SAS

SAP

Shifting paradigms

- What we have covered are traditional data warehousing architectures and technologies
- All of this is very ‘textbook’
- Most of what was covered only exist in large enterprises
- Business requirements are slowly changing
- We will cover these upcoming trends…if we have time