

# Table of contents

- 1. Introduction**
- 2. Basic concepts (Text: 4.1)**
- 3. Multi-dimensional Data Cubes (Text 4.2.1)**
- 4. Concept Hierarchies (Text 4.2.3)**
- 5. Modelling the Cube in a Relational Database (Text 4.2.2)**
- 6. Exercise: Play with a Cube on the Web**
- 7. Practical Exercises**


# 1. Introduction

Most of this material is derived from the text, Han, Kamber and Pei, Chapter 4, or the corresponding powerpoint slides made available by the publisher. Where a source other than the text or its slides was used for the material, attribution is given. Unless otherwise stated, images are copyright of the publisher, Elsevier.

Here, we introduce some basic data warehouse concepts which will be extended in the next e-book.

Those of you who have studied COMP8430 Data Wrangling will have seen some of this before, but we will go deeper into this topic here, together with some practical exercises.


You will need some basic understanding of relational databases and SQL in this part of the Data Mining course.

**ACTION:**  [Use this if you need a refresher for SQL](#)

The work on data warehousing in this course aims to help you to understand the role of a data warehouse in enterprise data management and how it relates to data mining, especially challenges related to large scale data management. This will to equip you with the knowledge you need to be able to navigate and extract the data you need from a data warehouse, or to exploit or develop data mining techniques embedded within data warehouse architectures.

**ACTION:** Watch this video lecture on the topic if you wish. Everything it covers is also in the notes.

WATTLE Data Warehouse OLAP lecture 2019



WATTLE

RESOURCES • HEALTH AND WELLBEING • LIBRARY • WATTLE SUPPORT • ENGLISH (EN) •

Kerry Taylor Student

## 2 Basic concepts (Text: 4.1)


### What is a Data Warehouse?

- Defined in many different ways, but not rigorously
- A decision support database that is maintained separately from the organisation's operational database
- Supports information processing by providing a solid platform of consolidated, historical data for analysis
- A data warehouse is a **subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management's decision-making process.—W. H. Inmon
- Data warehousing is the process of constructing and using data warehouses

0:00/37:55

CC 1x

讨论 0

 扩展

## 2. Basic concepts (Text: 4.1)

### What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
- A decision support database that is maintained separately from the organisation's operational database
- Supports information processing by providing a solid platform of consolidated, historical data for analysis.
- A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- *Data warehousing* is the process of constructing and using data warehouses

### Typical Characteristics of a Data Warehouse

#### Subject-Oriented

- Organised around major subjects, such as customer, product, sales
- Focusing on the modelling and analysis of data for decision makers, not on daily operations or transaction processing
- Provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

#### Integrated

- Constructed by integrating multiple, heterogeneous data source e.g. relational databases, flat files, on-line transaction records
- 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 (this process is called *ETL*, for Extract, Transform, Load)

#### 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"

#### Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing: Initial loading of data and access of data

### Comparison with typical operational databases

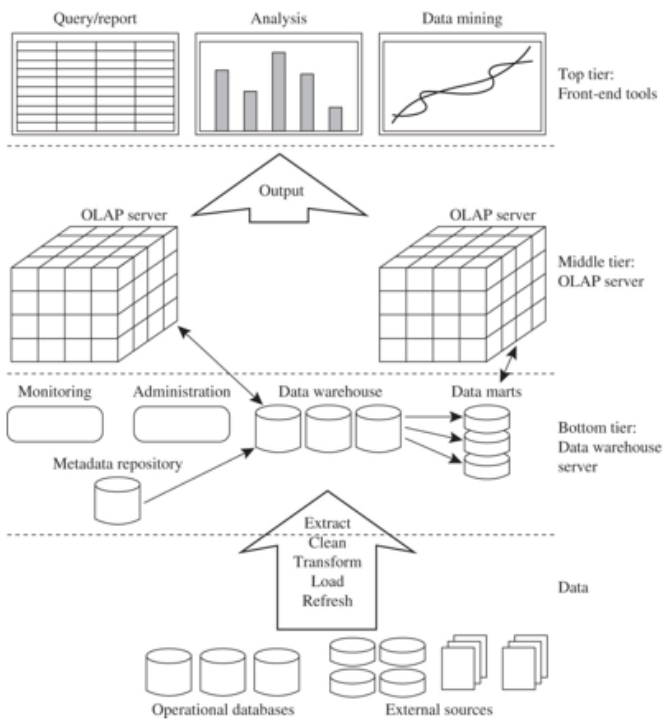
Most databases are used for **online transaction processing (OLTP)** and querying for day-to-day operations. On the other hand, data warehouses are built specifically for analytics to support decision making, that is, **online analytical processing (OLAP)**.

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

## Why A Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: Decision support requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

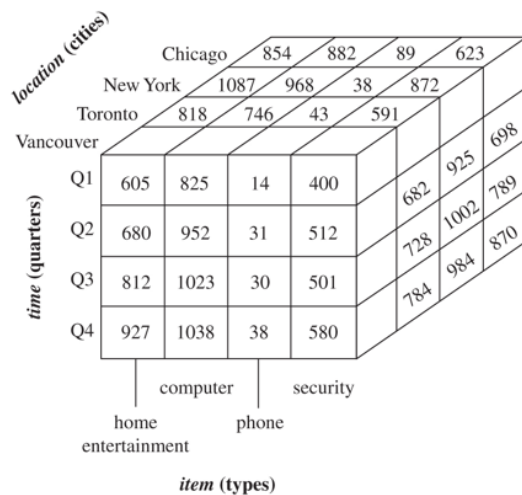
## Multi-tiered Architecture



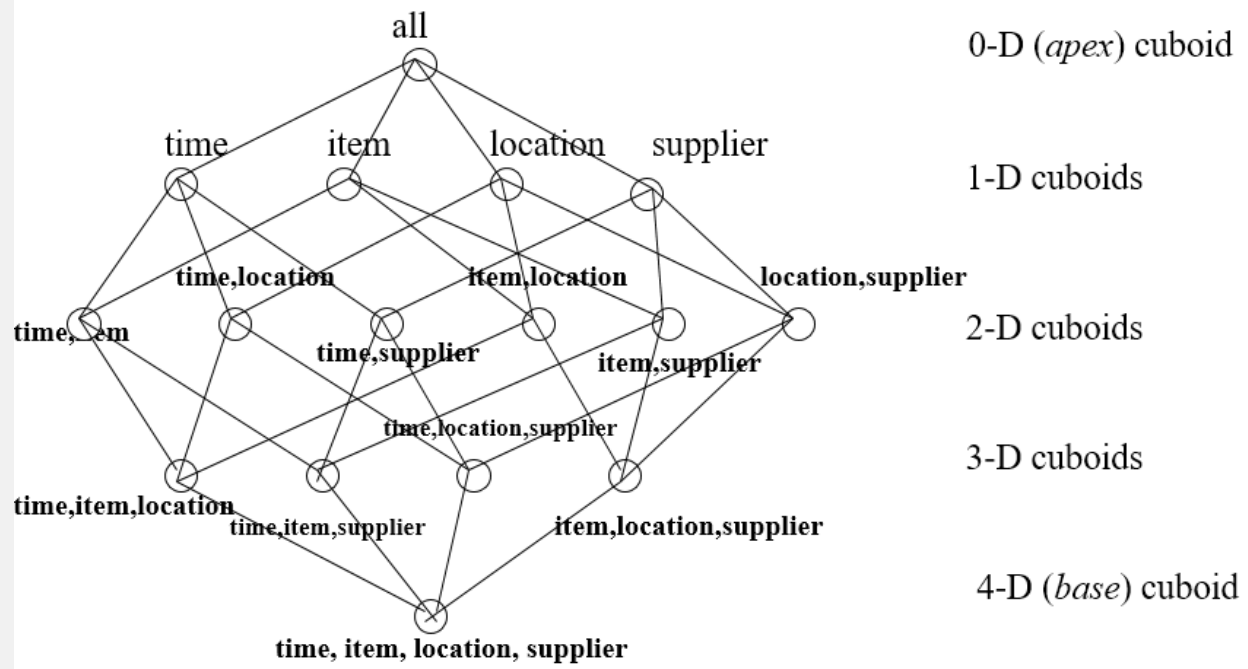
### 3. Multi-dimensional Data Cubes (Text 4.2.1)

#### From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a **multidimensional data model** which views data in the form of a data cube
- A data cube, such as *sales*, allows data to be modeled and viewed in multiple dimensions.
- The cube axes are the **dimensions** such as item (e.g. item\_name, brand, type), or time (e.g. day, week, month, quarter, year) or location (e.g. store, city, state, country) . The values of those dimensions (e.g. Item = home\_entertainment, Quarter = Q4, City = Vancouver) uniquely identify a cube **cell**.
- The cube cells hold the **measures** (such as dollars\_sold, quantity-sold and quantity-returned).



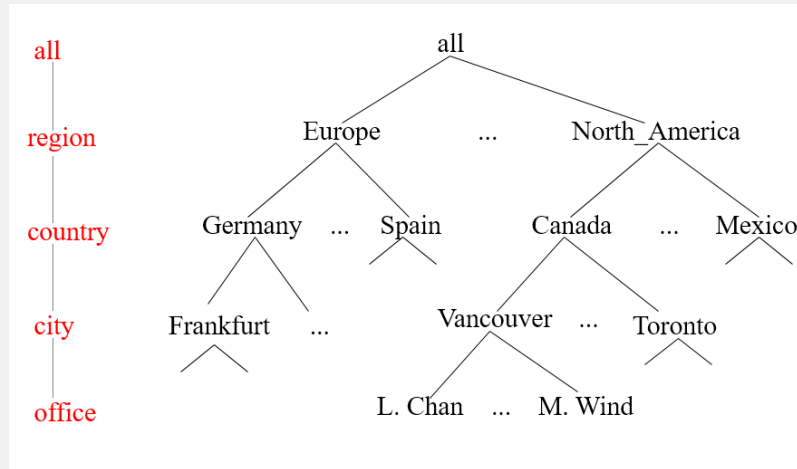
- It is easy to see why this structure is commonly called an (n-dimensional) *cube*. However, in data warehousing literature, this structure may be called a **cuboid** as a component of a more elaborate **data cube**.
- An n-dimensional cuboid of unaggregated data is called a **base cuboid**.
- Measures along one or more dimensions of the base cuboid may be aggregated to form another cuboid with dimensions that are a subset of the original dimensions.
- The top-most 0-D cuboid, which holds the highest-level of summarisation in a single cell, is called the **apex cuboid**, commonly denoted *all*.
- **The lattice of cuboids forms a data cube.**
- In this case the *all* cuboid contains a single cell that represents aggregated dollars\_sold over all time, for all items, at all locations and for every supplier.
- An intermediate cuboid, for example the 2-D (item, location) cuboid, will contain cells such as (Mars bar, Belconnen Mall) with measure \$20,000, indicating that \$20,000 worth of Mars Bars have been sold at the Belconnen Mall shop, over all time and for all suppliers.
- The bottom-most *base* cuboid contains ||time|| x ||item|| x ||location|| x ||supplier|| cells, where each cell represents the dollars-sold for a possible combination of specific values for time, item, location and supplier.



## 4. Concept Hierarchies (Text 4.2.3)

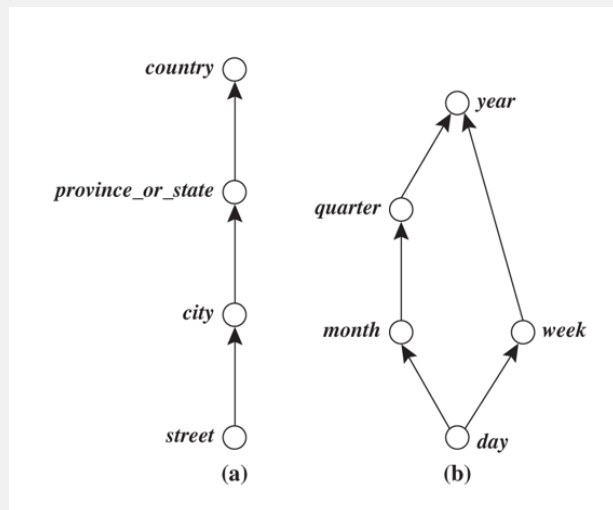
A **concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. They support the modelling of data at varying levels of abstraction.

For example, a concept hierarchy for *location* could be:



A lattice is also possible, for example for *time*, as shown here contrasted with the simpler hierarchy. Organising time as shown here is very common.

It illustrates the partial order  $day < \{week ; month < quarter\} < year$ .



Concept hierarchies may also be defined by discretizing or grouping values, called a *set-grouping hierarchy*, as shown in the example below for *price*.

**ACTION:** Think about how concept hierarchies are represented in a datacube viewed as a lattice of cuboids.



## 5. Modelling the Cube in a Relational Database (Text 4.2.2)

Need:

- Dimension tables, such as item (item\_name, brand, type), or time (day, week, month, quarter, year)
- A fact table that contains measures (such as dollars\_sold) and keys to each of the related dimension tables

**Star schema:** A fact table in the middle connected to a set of dimension tables. The primary key for the fact table is composed of a key for each dimension and the remaining attributes are the measures.

**Snowflake schema:** A refinement of star schema where some dimensional hierarchy is normalised into a set of smaller dimension tables, forming a shape similar to snowflake.

**Fact constellation:** Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

Snowflake and Constellation schemes are fairly obvious extensions of the basic Star schema to more complex data.

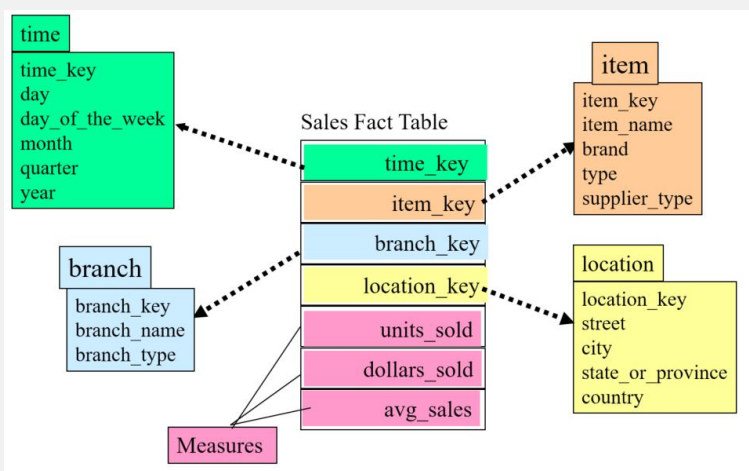


Figure: Example Star Schema

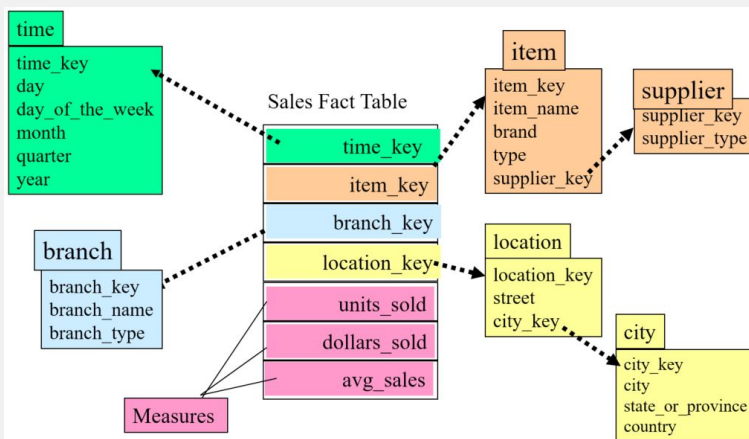
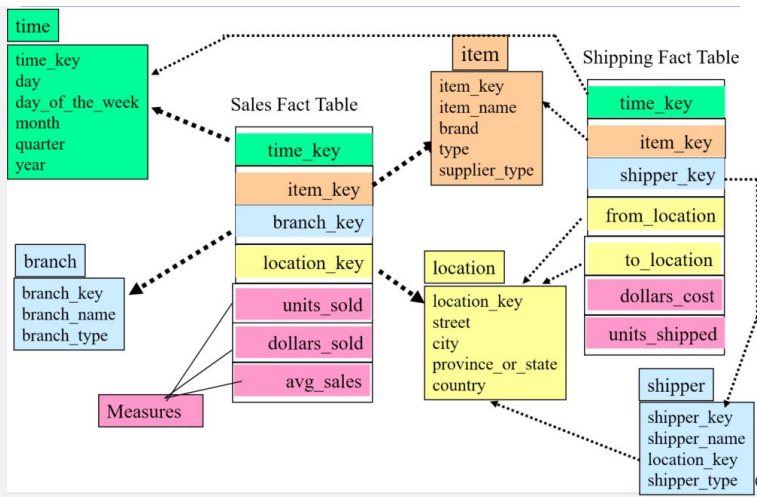


Figure: Example Snowflake Schema

Here we can see the snowflake schema derived from the star schema above by normalising both the item and location dimension tables in line with the dimensional hierarchies they represent (supplier < supplier type and street < city respectively).



**Figure: Example Fact Constellation Schema**

Here we can see that the fact constellation schema adds an additional fact table (shipping) and shares the dimension tables for the location with the original fact table for sales.

## 6. Exercise: Play with a Cube on the Web

**ACTION:** If you would like to see a datacube in practice, have a play with this one about you and your co-students from the Australian Department of Education.

<http://highereducationstatistics.education.gov.au/>

## 7. Practical Exercises

**ACTION:** Now do these simple exercises, for which you will need to download the data below.

Here is a tutorial video to get you started.


### COMP3425/8410 Data Warehousing



And here are the exercises, data and solution.

 [Exercises for Introduction to Data Warehousing](#)

 [Data for DW exercises](#)

 [Solution to Exercises: Intro to Data Warehousing](#)