European **A**rchival **R**ecords and **K**nowledge Preservation

INTRODUCTION TO DATA WAREHOUSING AND BIG DATA

Janet Delve

School of Creative Technologies

Outline

- Data Warehouses, hidden in plain sight…
- Relational databases (Online Transactional Processing - OLTP)
- Data Warehouse fundamentals
- Making Analysis Easy
- Online Analytical Processing (OLAP)
- Big Data

Data Warehouse example

CLOUD // SOFTWARE AS A SERVICE

NEWS
2/15/2013
05:21 PM

Charles Babcock
News

Connect Directly

1 COMMENT
COMMENT NOW

Login

50% 50%

Amazon Launches Redshift Data Warehousing As A Service

Amazon promises 10 times the performance at one-tenth the cost of on-premises data warehouses. Can it deliver?

Amazon Web Services on Friday carried out the promised launch of its Redshift data warehouse service, with which it hopes to disrupt on-premises data warehouses.

"We designed Amazon Redshift to deliver 10 times the performance at one-tenth the cost of the on-premises data warehouses that are commonly used today," wrote Jeff Barr, AWS chief evangelist, in a blog post.

It remains to be seen whether a cloud data warehouse can function with that much less expense than enterprise systems and be

Amazon's 7 Cloud Advantages: Hype Vs. Reality

(click image for larger view and for slideshow)

Data Warehouse example google

Data Warehouse examples

- **Virgin Megastores charts real-time retailing trends**
- **High-street retailer invests in business intelligence with data warehousing project**
- Miya Knights, Computing, 09 Feb 2006
- Virgin Megastores is using data warehousing software as the basis of a business intelligence (BI) project to improve the quality of its performance reporting.

Examples

- The high-street retailer has created a repository for real-time access to sales figures fed from its shops, to improve buying and store management processes.
- Tony Johnson, IT director for Virgin Megastores, says previous performance reporting capabilities did not provide a real-time view of what the company sells in each shop, and when.
- ‘We are using this reporting project to focus on the key areas of stores and margins,’ said Johnson. ‘We have a real-time view of stock, and other applications can link into this at the central, buying level.’

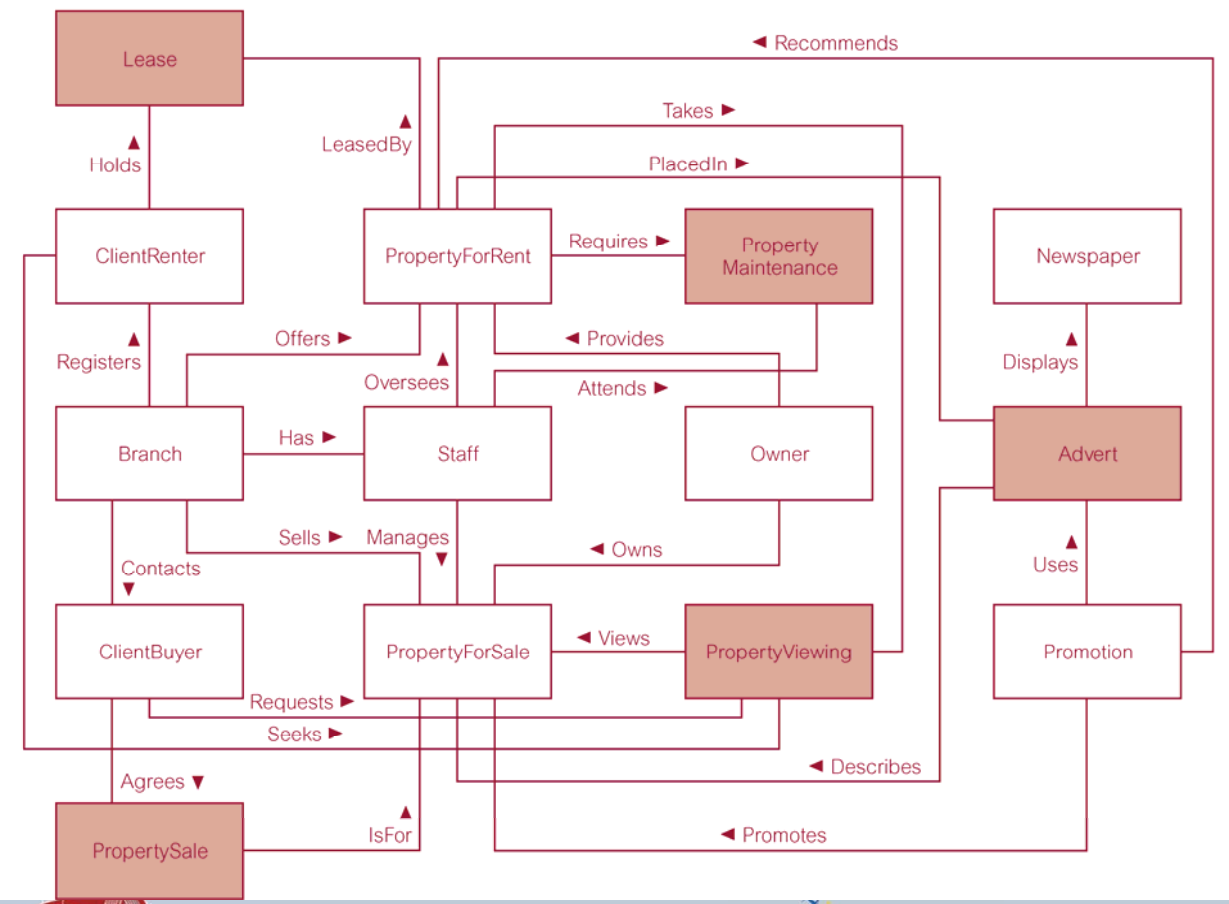
Relational Database

Relational Table Model

	ID	Name	City	Country
1	1	Espen	Oslo	Norway
2	2	Harald	Munich	Germany
3	3	Sam	San Jose	USA

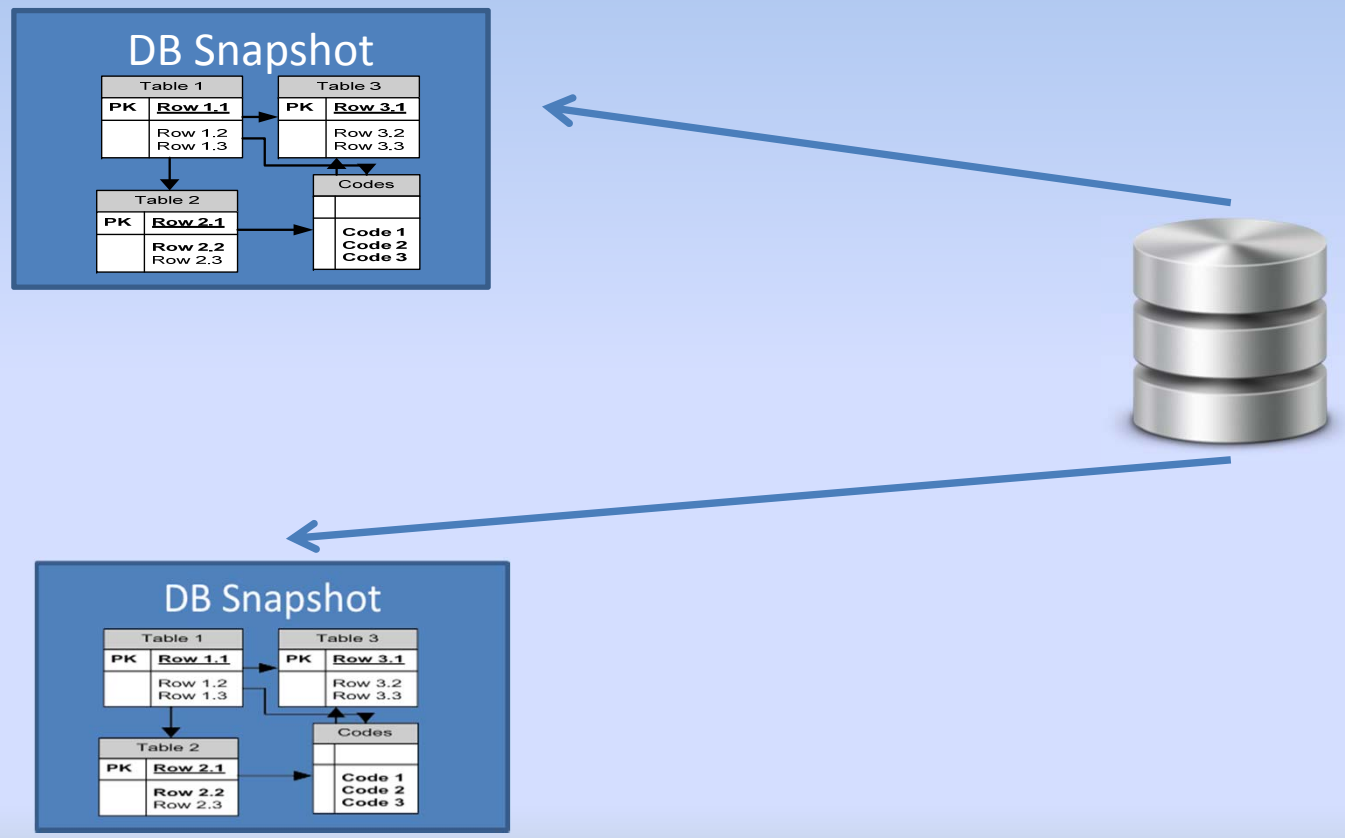
- Built for current data (banks transactions etc.)
- Mathematical basis
- Efficient for processing…BUT

Transactional Processing (OLTP)



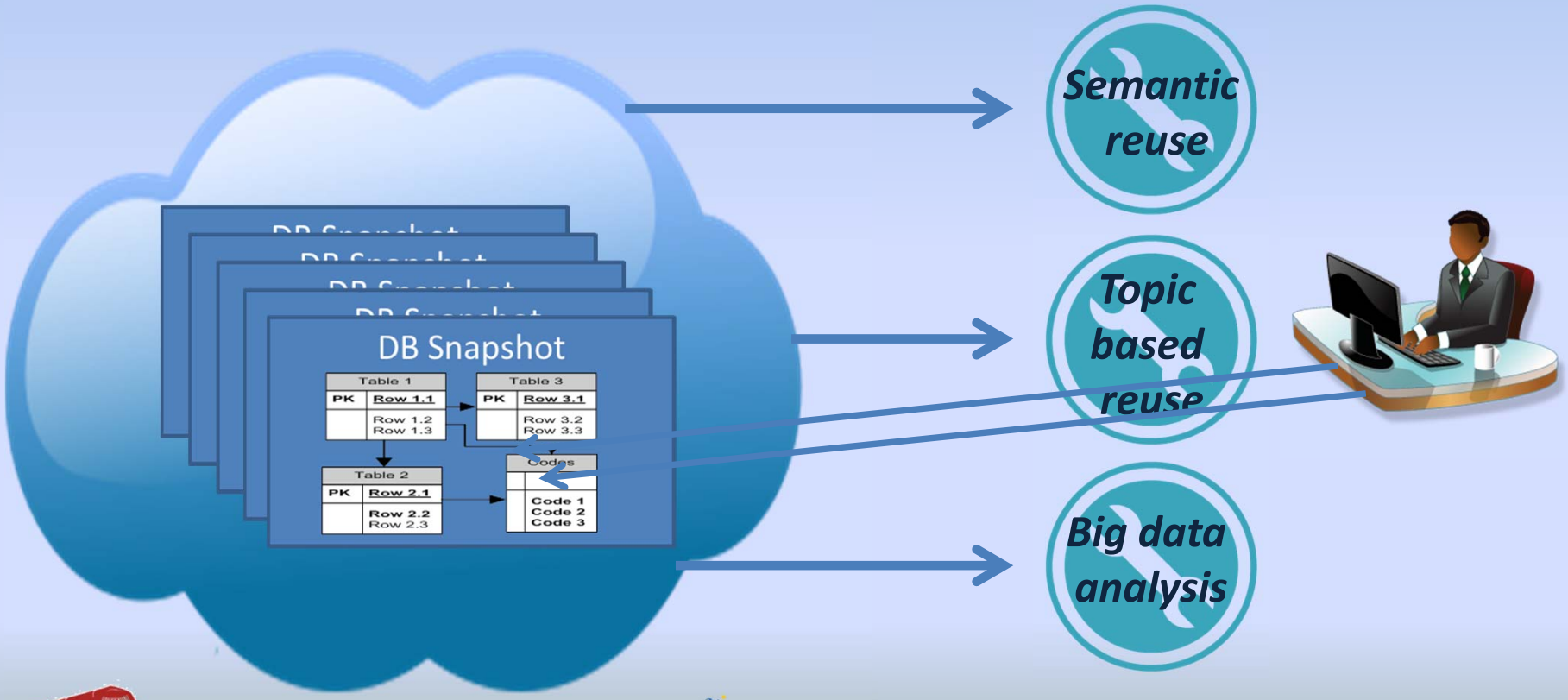
JOINS
TIME
ANALYSIS?

Data Warehouse fundamentals
SNAPSHOTS



Data warehouse

- …a collection of database snapshots



Data Warehousing fundamentals

- A DW is *subject-oriented, integrated, non-volatile & time-variant.*
- Classical operations are organised around the *applications* of the company.
- E.g. for an insurance company the *applications* may be car, health, life and accident. The major *subjects* are customer, policy, premium and claim.
- *Integration* is the most important facet of a DW. Fig. 2.2 Previous inconsistencies are ironed out and all data unambiguously entered into DW.

Data Warehousing fundamentals: *harmonize*

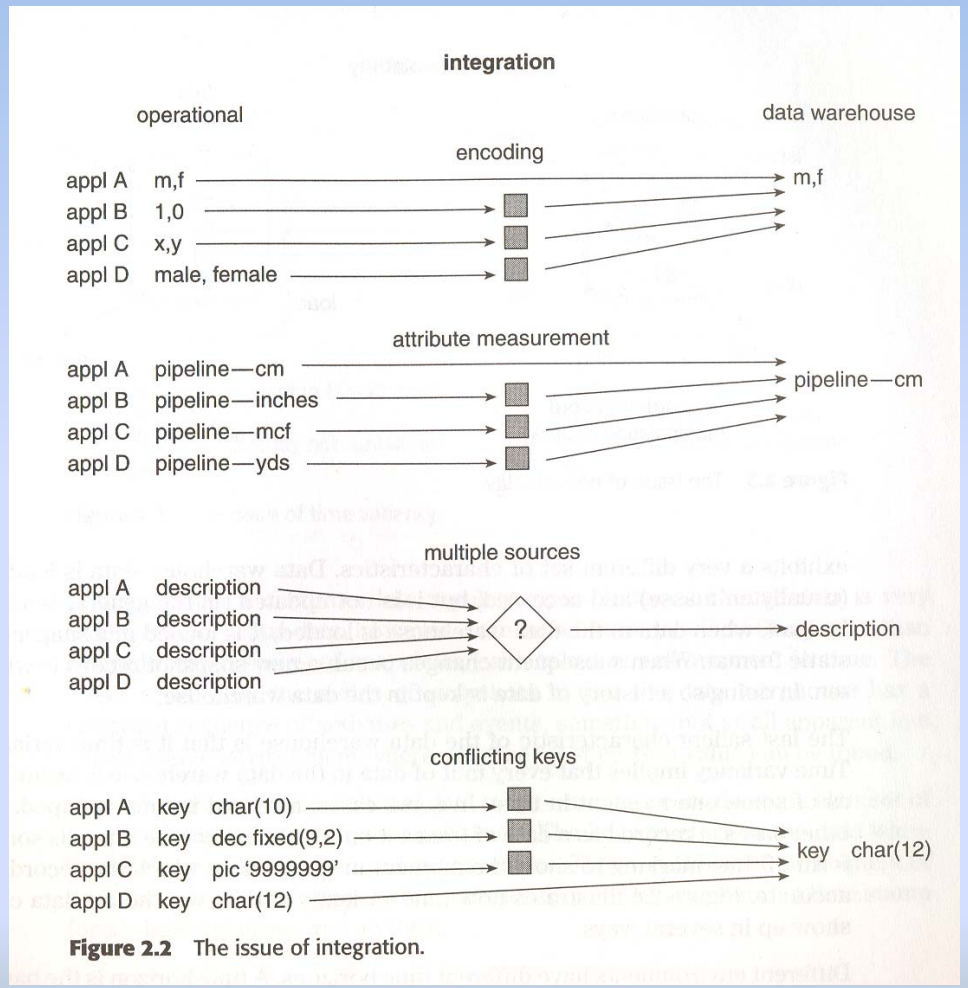


Figure 2.2 The issue of integration.

Data Warehousing fundamentals

- *Non-volatile* data in a DW means that it is not changed in the way data is in operational database – data is loaded en masse and is NOT updated.
- *Time- variant* – DW time horizon 5 –10 years, operational database 2-3 months. DW snapshots, operational database current data, DW always has element of time, operational database might or might not have.

Data Warehousing fundamentals *time*

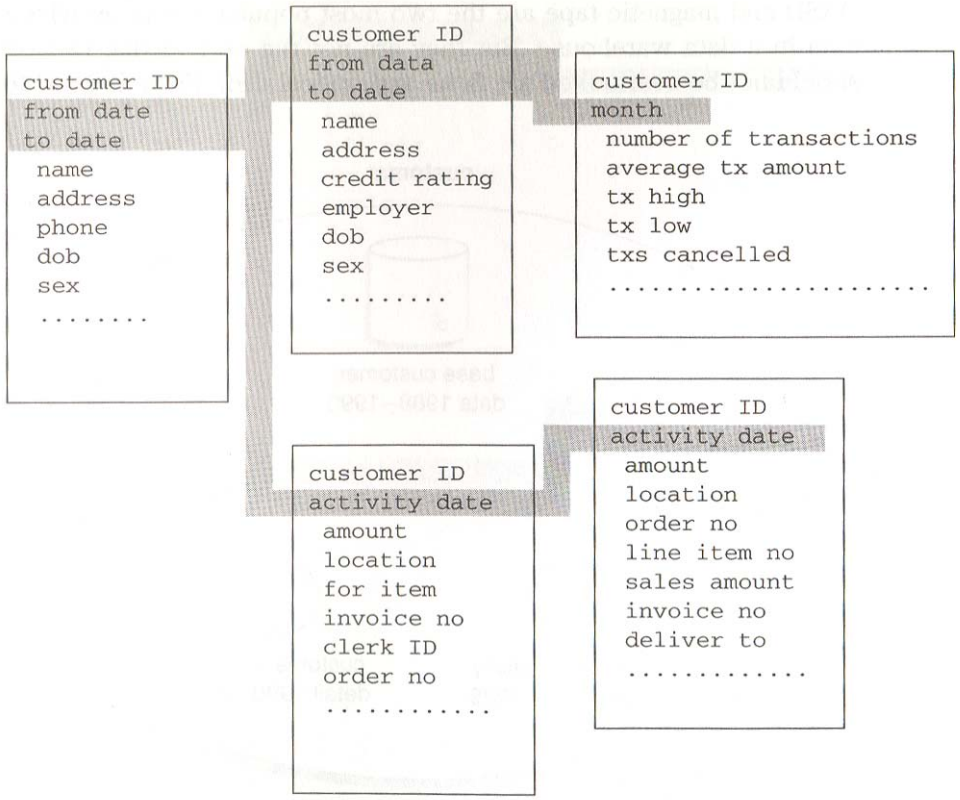
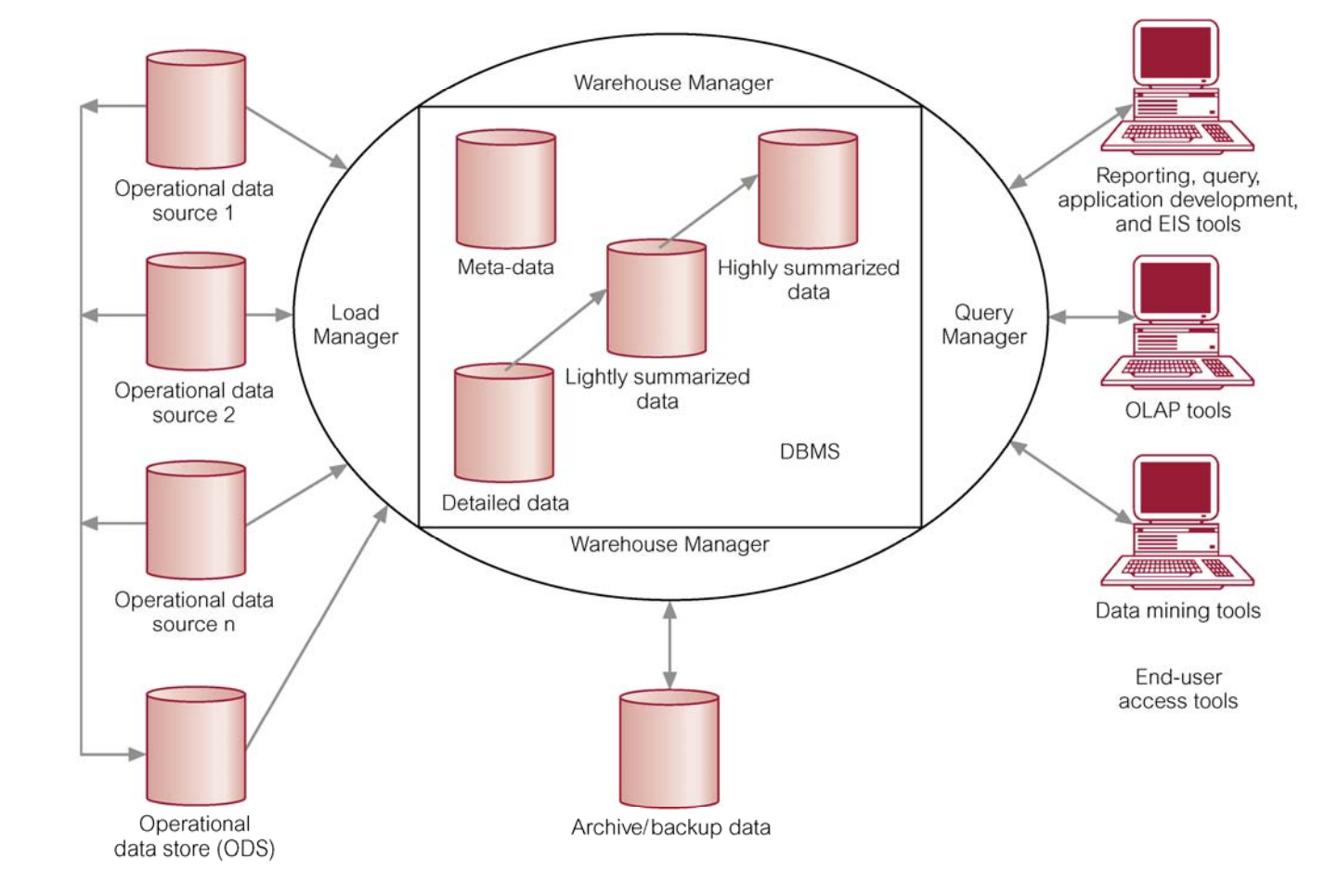


Figure 2.9 Each table in the data warehouse has an element of time as a part of the key structure, usually the lower part.

Typical Architecture of a Data Warehouse



Data Warehouse Environment

Comparison of OLTP Systems and Data Warehousing

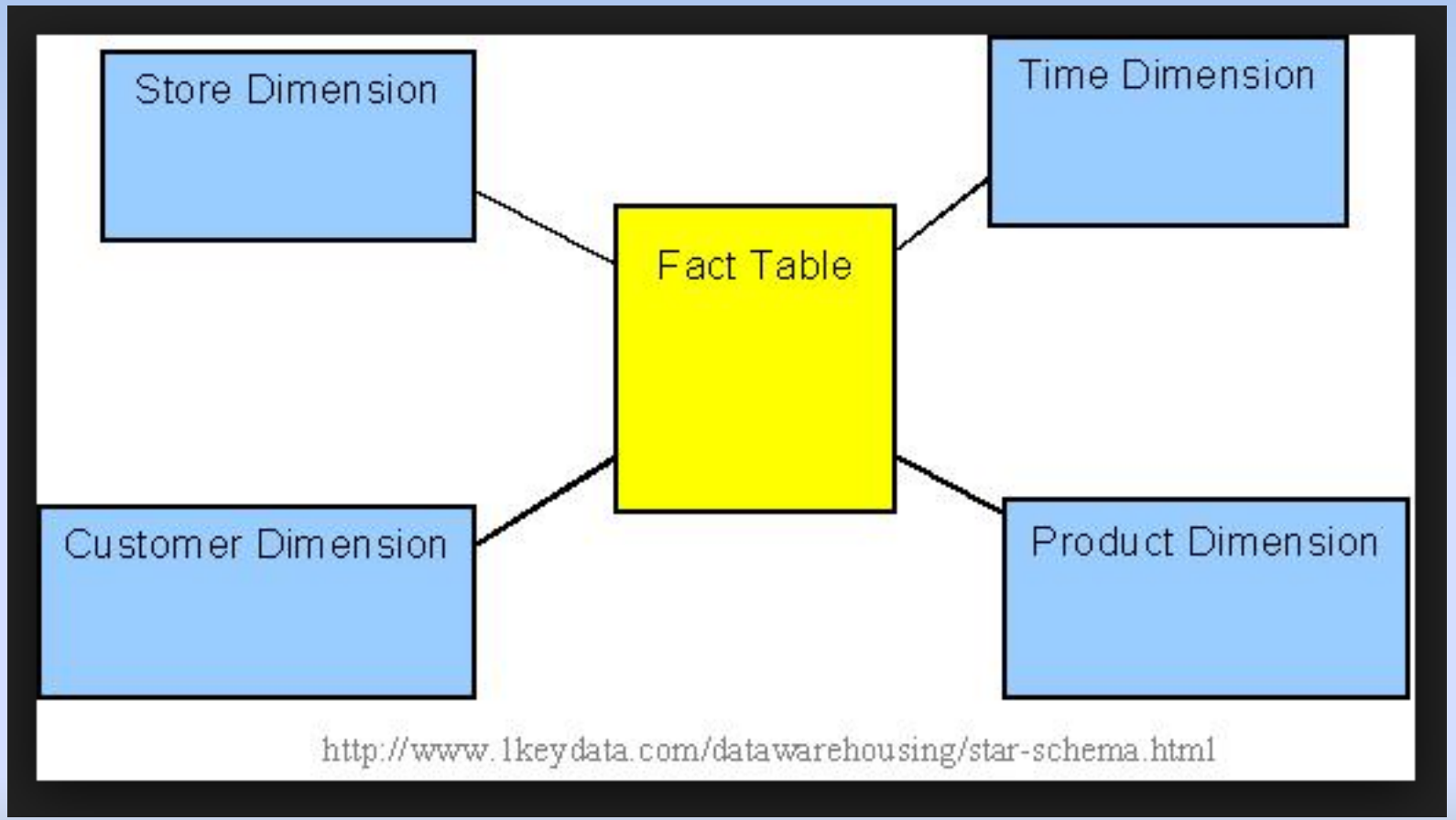
Table 30.1 Comparison of OLTP systems and data warehousing systems.

OLTP systems	Data warehousing systems
Holds current data	Holds historical data
Stores detailed data	Stores detailed, lightly, and highly summarized data
Data is dynamic	Data is largely static
Repetitive processing	*Ad hoc*, unstructured, and heuristic processing
High level of transaction throughput	Medium to low level of transaction throughput
Predictable pattern of usage	Unpredictable pattern of usage
Transaction-driven	Analysis driven
Application-oriented	Subject-oriented
Supports day-to-day decisions	Supports strategic decisions
Serves large number of clerical/operational users	Serves relatively low number of managerial users

Data warehousing

- *Snapshots* (Useful for DB archiving)
- Star schema – dimensional model
- MADE FOR EASY ANALYSIS

Easy Analysis: Star Schema



Retail Sales Dimensional Model

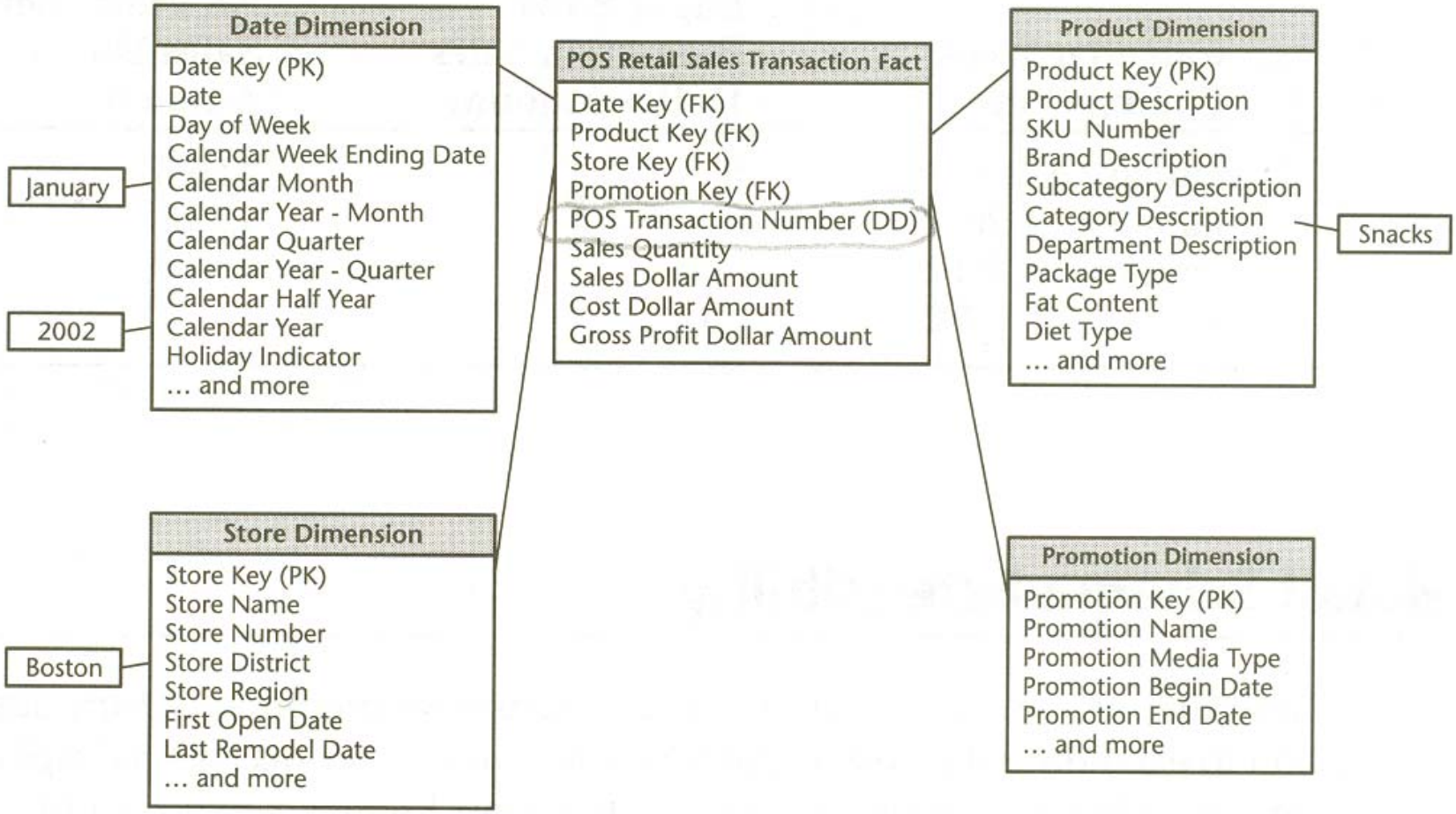


Figure 2.10 Querying the retail sales schema.

- Kimball p51

Retail Sales Product Dimension

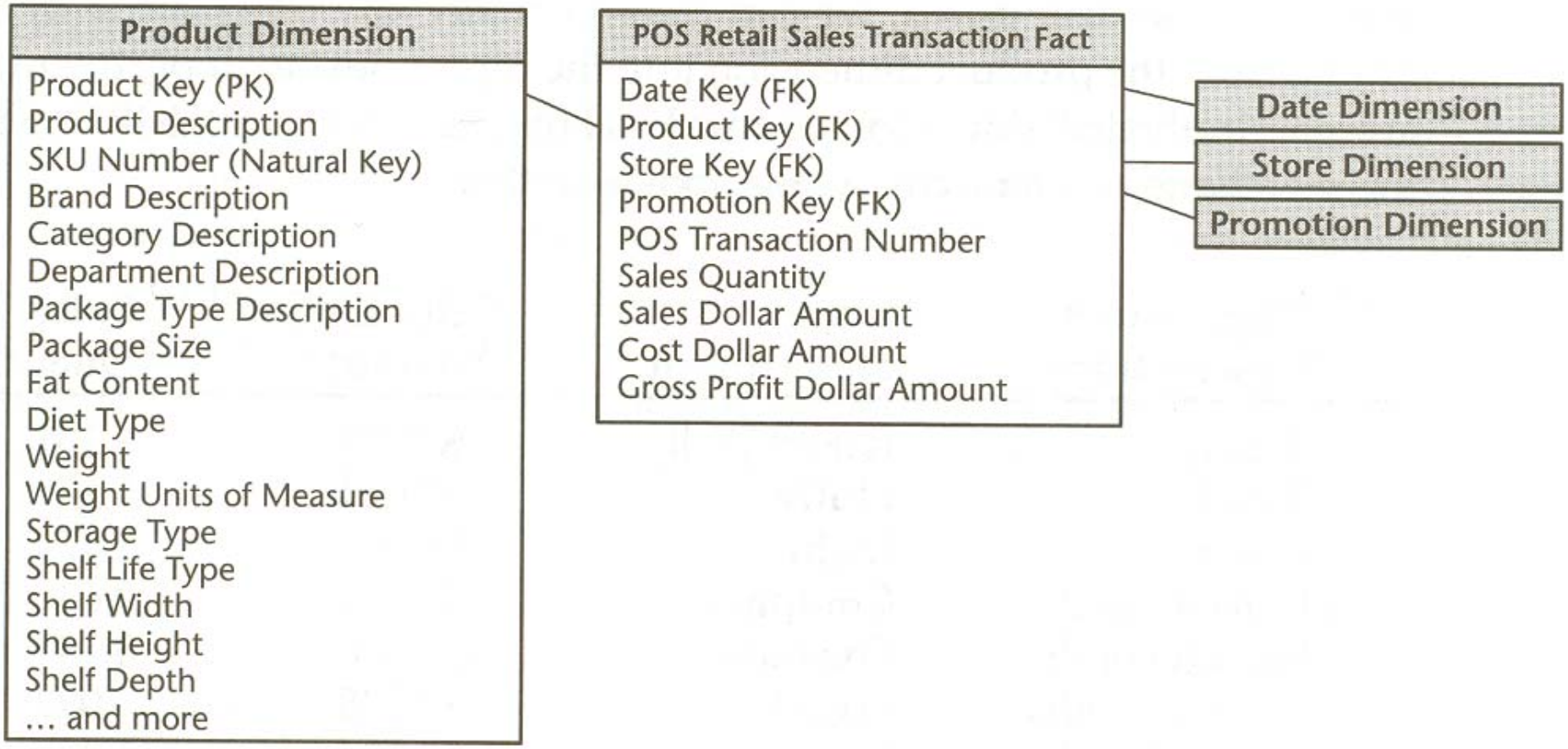


Figure 2.7 Product dimension in the retail sales schema.

- Kimball p43

Factless fact table

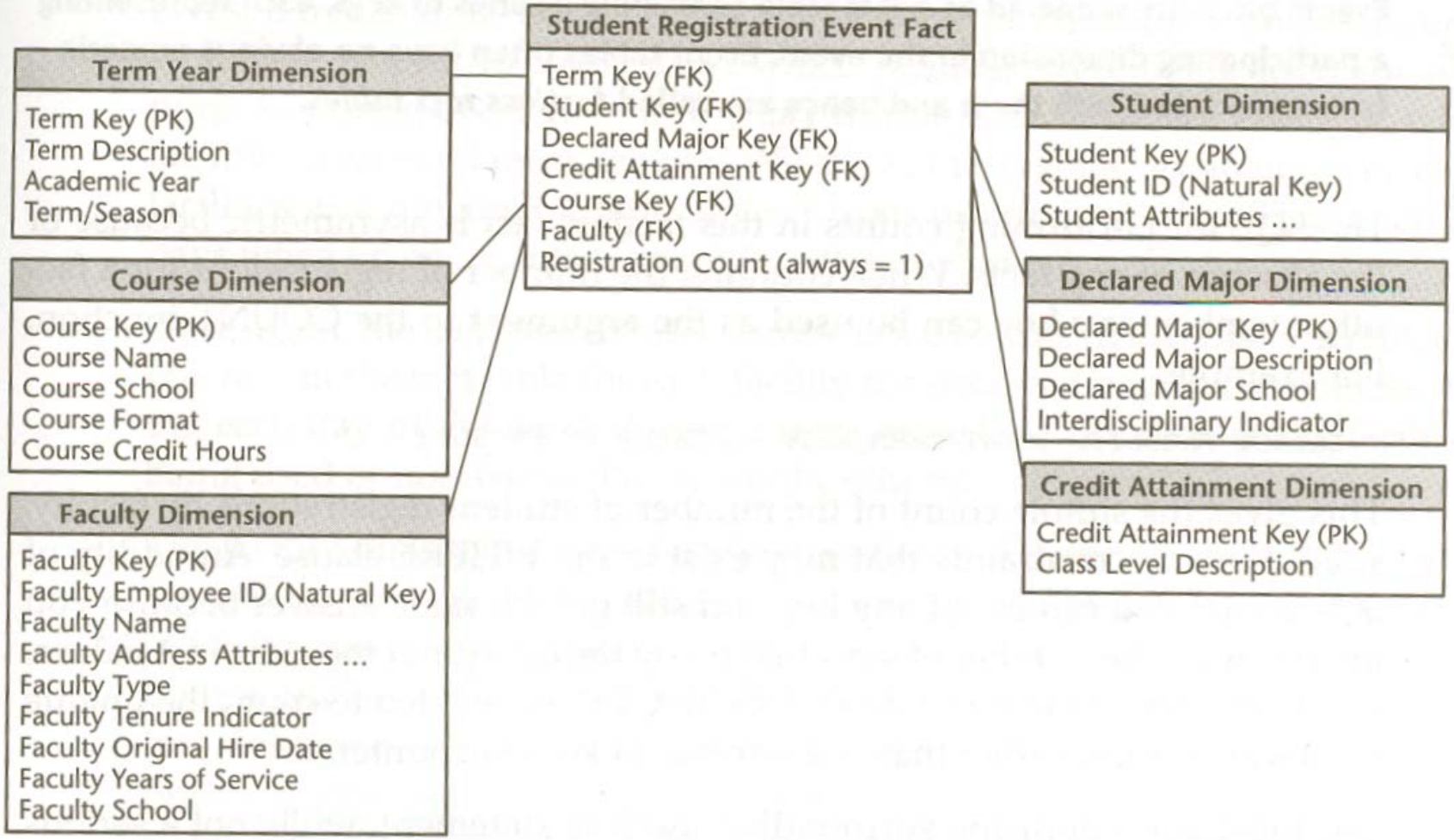
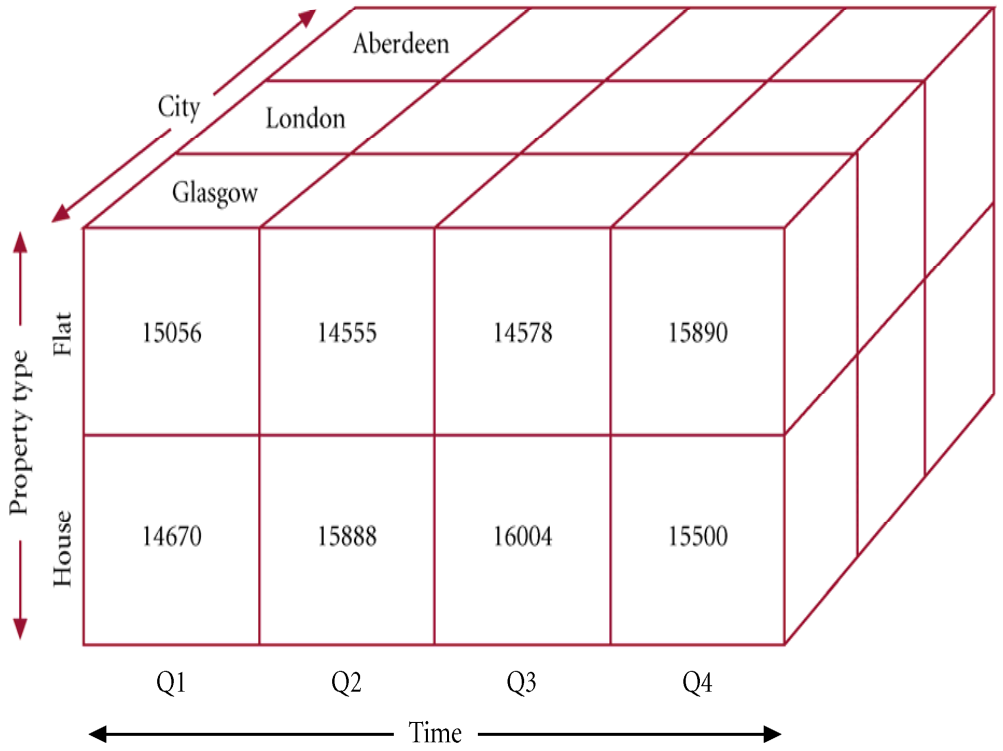


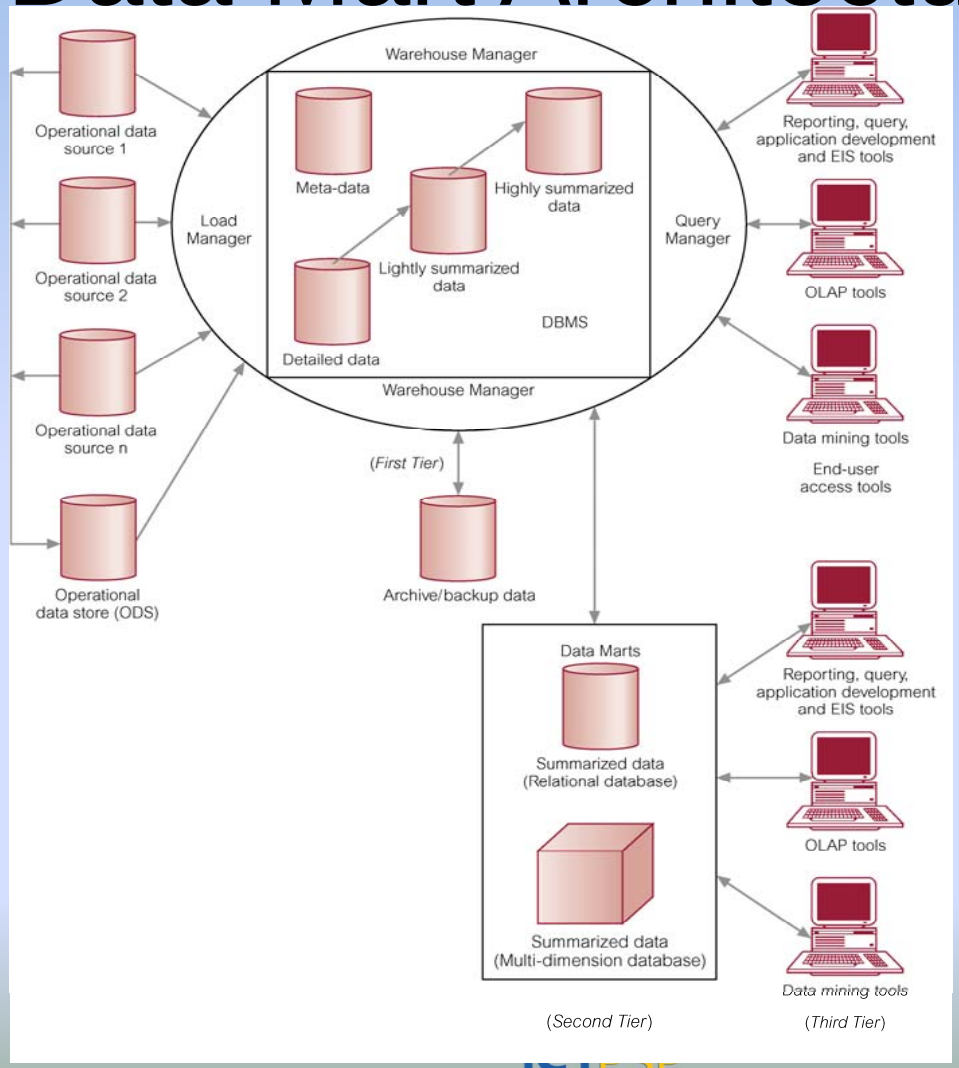
Figure 12.2 Student registration events as a factless fact table.

Online Analytical Processing (OLAP)

Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....	.....	.....	.....
.....	.....	.....	.....



Typical Data Warehouse and Data Mart Architecture



Data Warehouse Environment

What do we mean by big data in the biomedical sciences?

Dr Rhiannon Lloyd

Brain tumour research group

Big Data

- What is big data?
- Data types
- 3 Vs

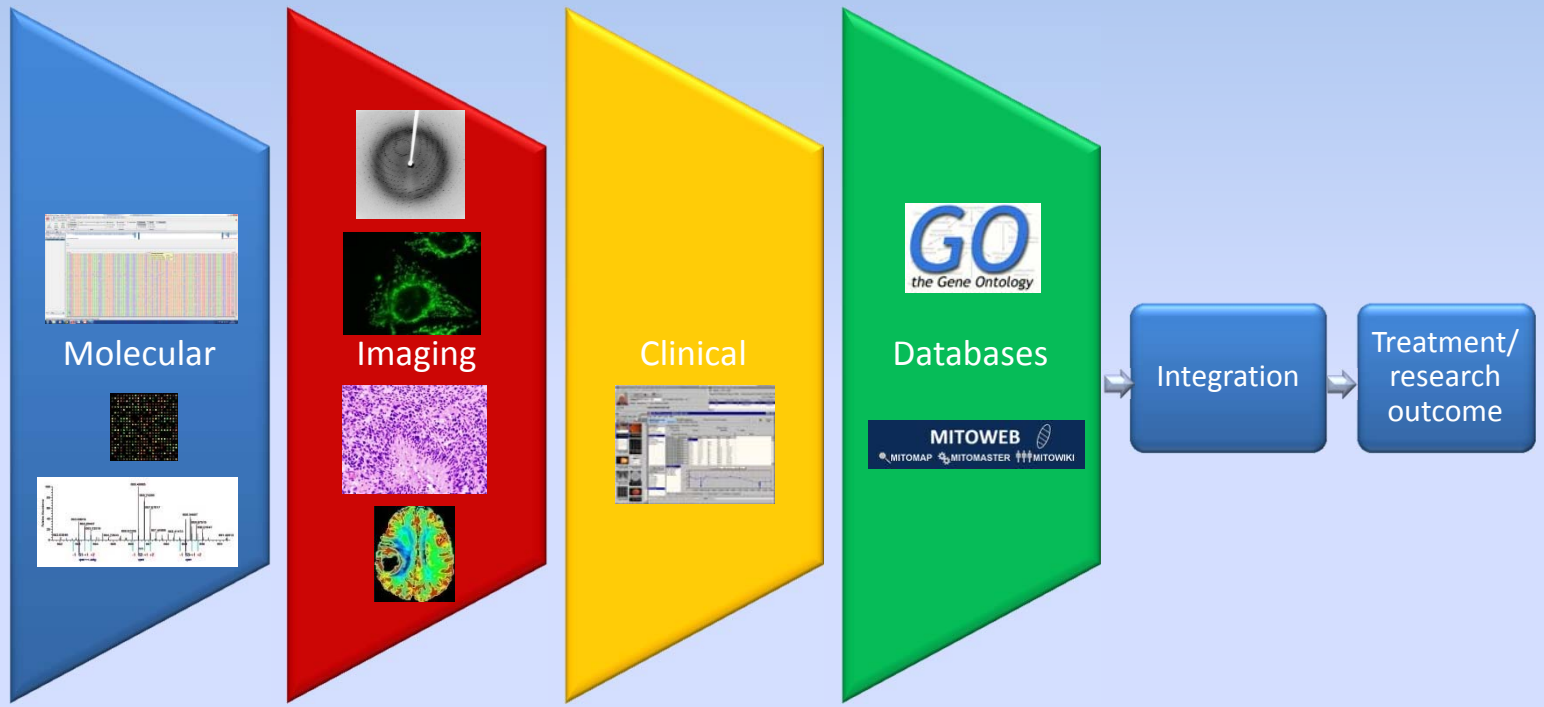
Big Data

- Large, diverse and complex datasets that are getting bigger
- Emanate from single source or multiple sources that need integrating
- Exceed currently used approaches to access, manage, integrate and analyse

What is happening in the UK?

The size of big data is not the only issue, heterogeneity is also a problem

Types of data



Biobanks

SEBTA
South of England Brain Tumour Alliance

THE FROZEN ARK
SAVING THE DNA OF ENDANGERED SPECIES

Xenopus

ICTPSP

E-ARK

3Vs

- Volume,
- Velocity
- Variety
- Cloud
- Open Source
- Hadoop etc.


Q&A