European **A**rchival **R**ecords and **K**nowledge Preservation
#earkproject www.eark-project.eu @EARKProject

Data Warehousing and OLAP (OnLine Analytical Processing) techniques for digital archiving

Janet Delve and Richard Healey

E-ARK final conference

Hungarian National Archives, Budapest

6-8 December 2016

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.’

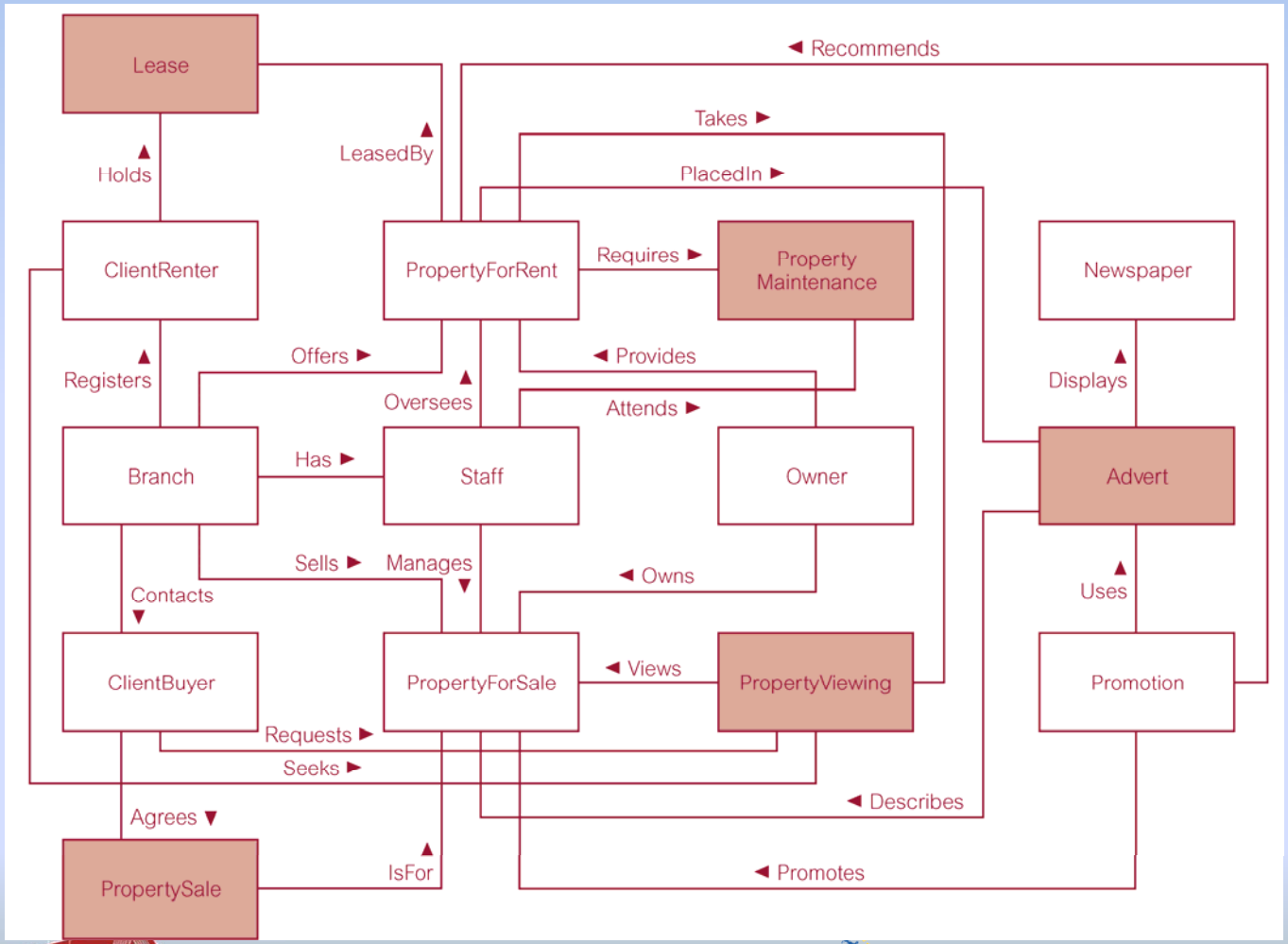
Relational Database

Relational Table Model

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

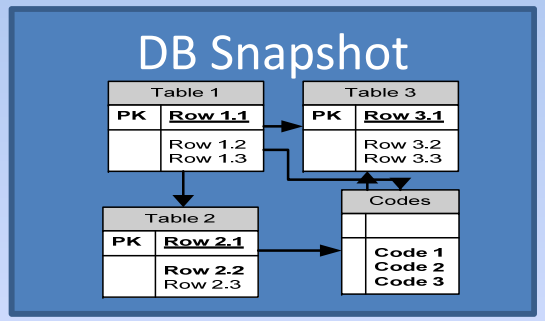
- Built for current data (banks transactions etc.)
- Mathematical basis
- Efficient for processing…BUT

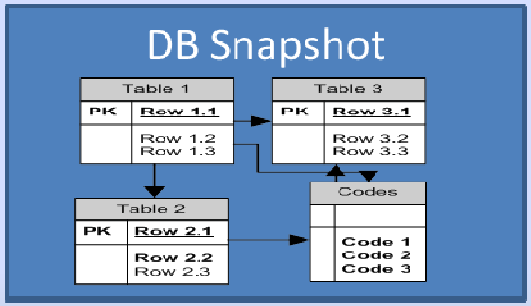
Transactional Processing (OLTP)



JOINS
TIME
ANALYSIS?

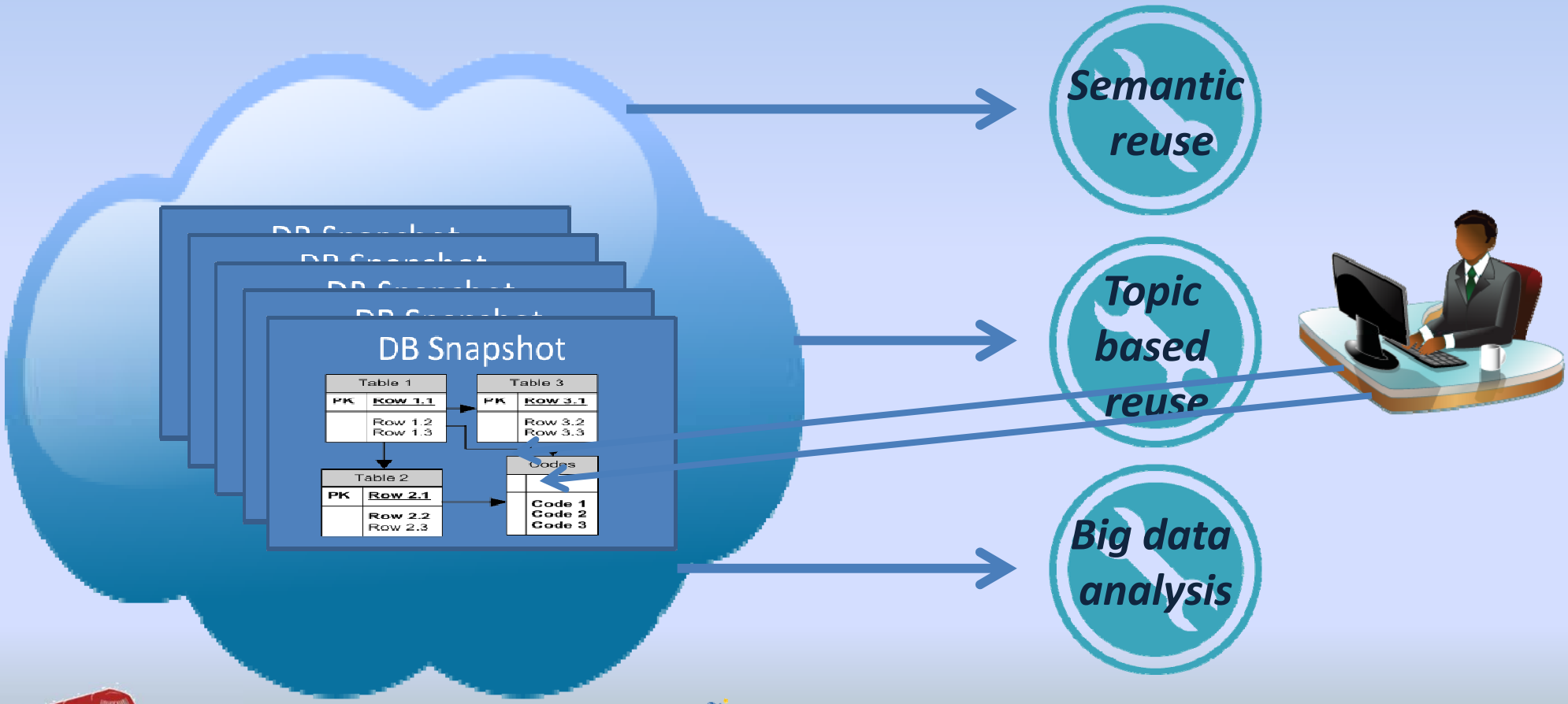
Data Warehouse fundamentals
SNAPSHOTS





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*

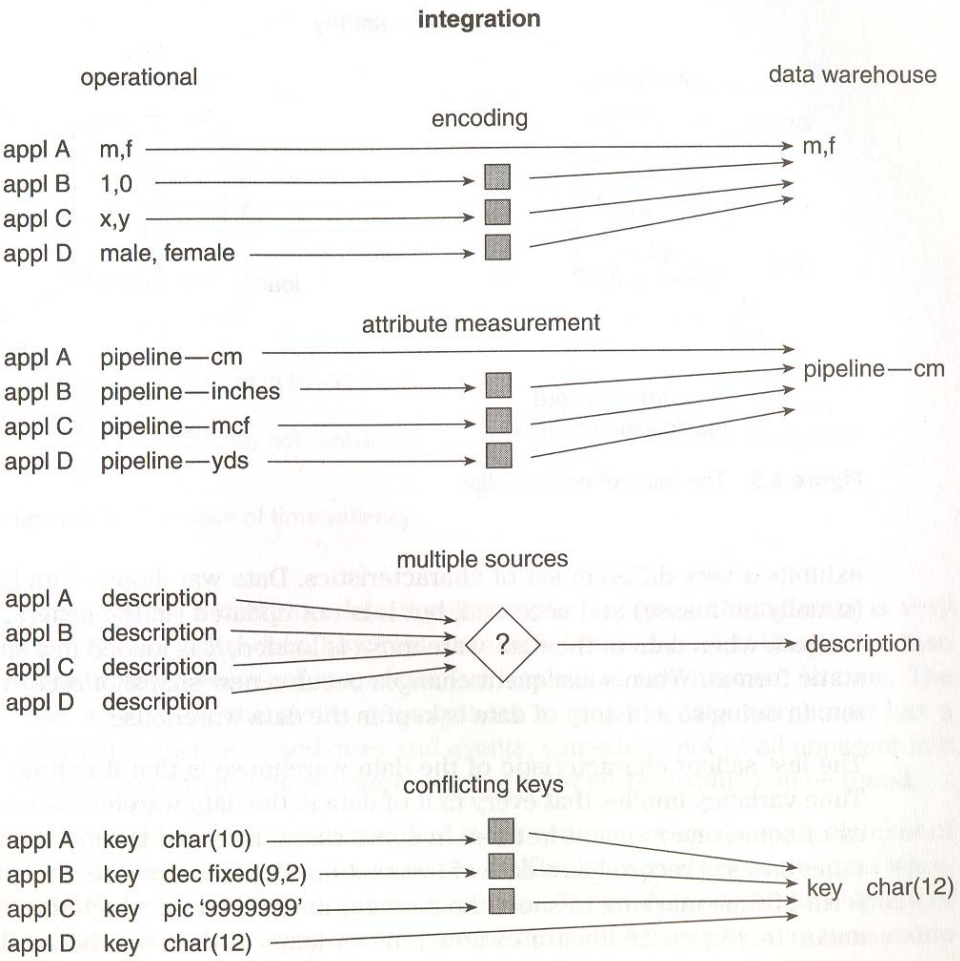


Figure 2.2 The issue of integration.

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*

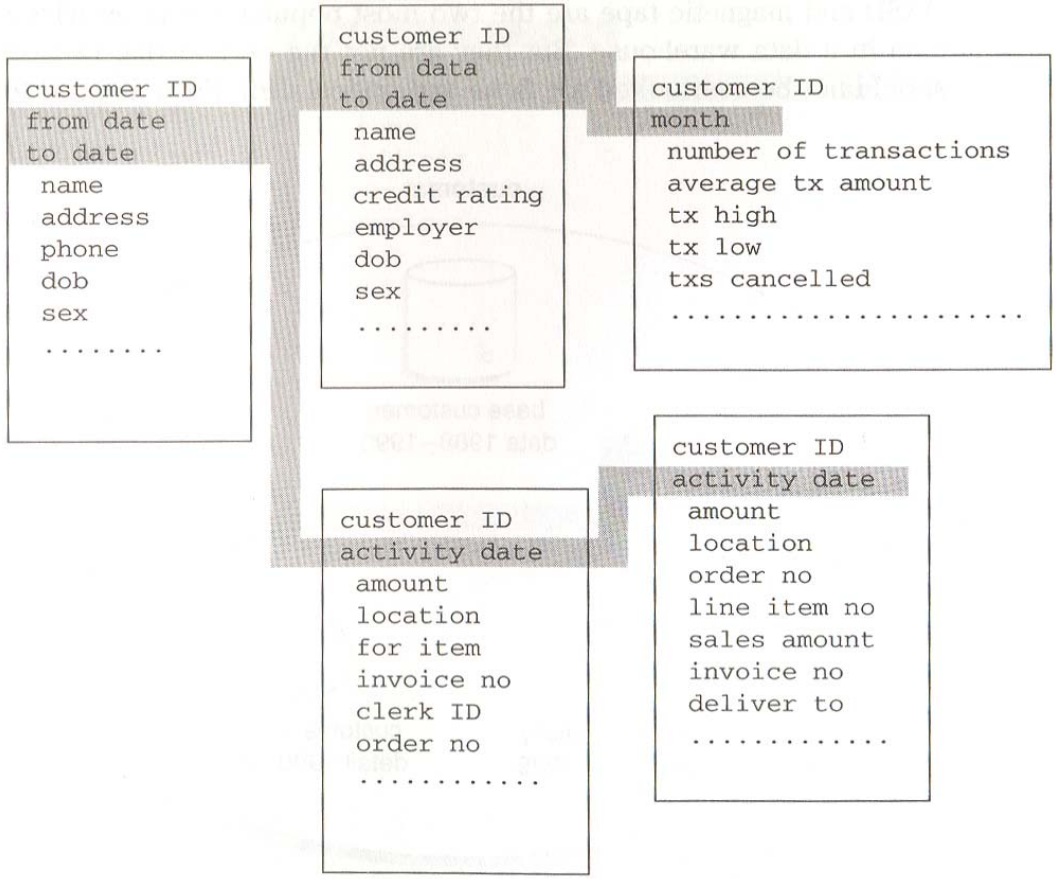
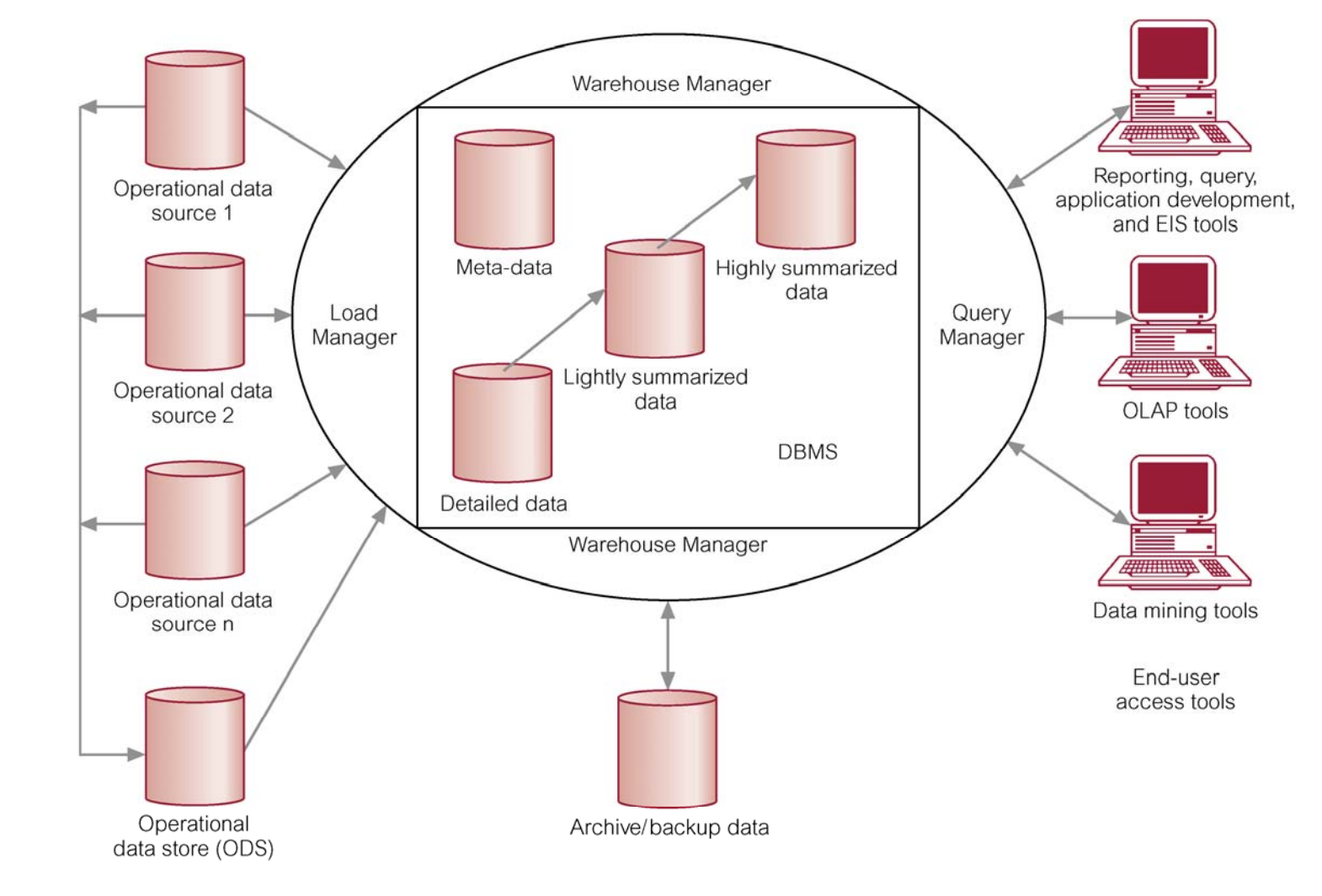


Figure 2.9 Each table in the data warehouse has an element of time as a part of the key structure, usually the lower part.

Typical Architecture of a Data Warehouse



Comparison of OLTP Systems and Data Warehousing

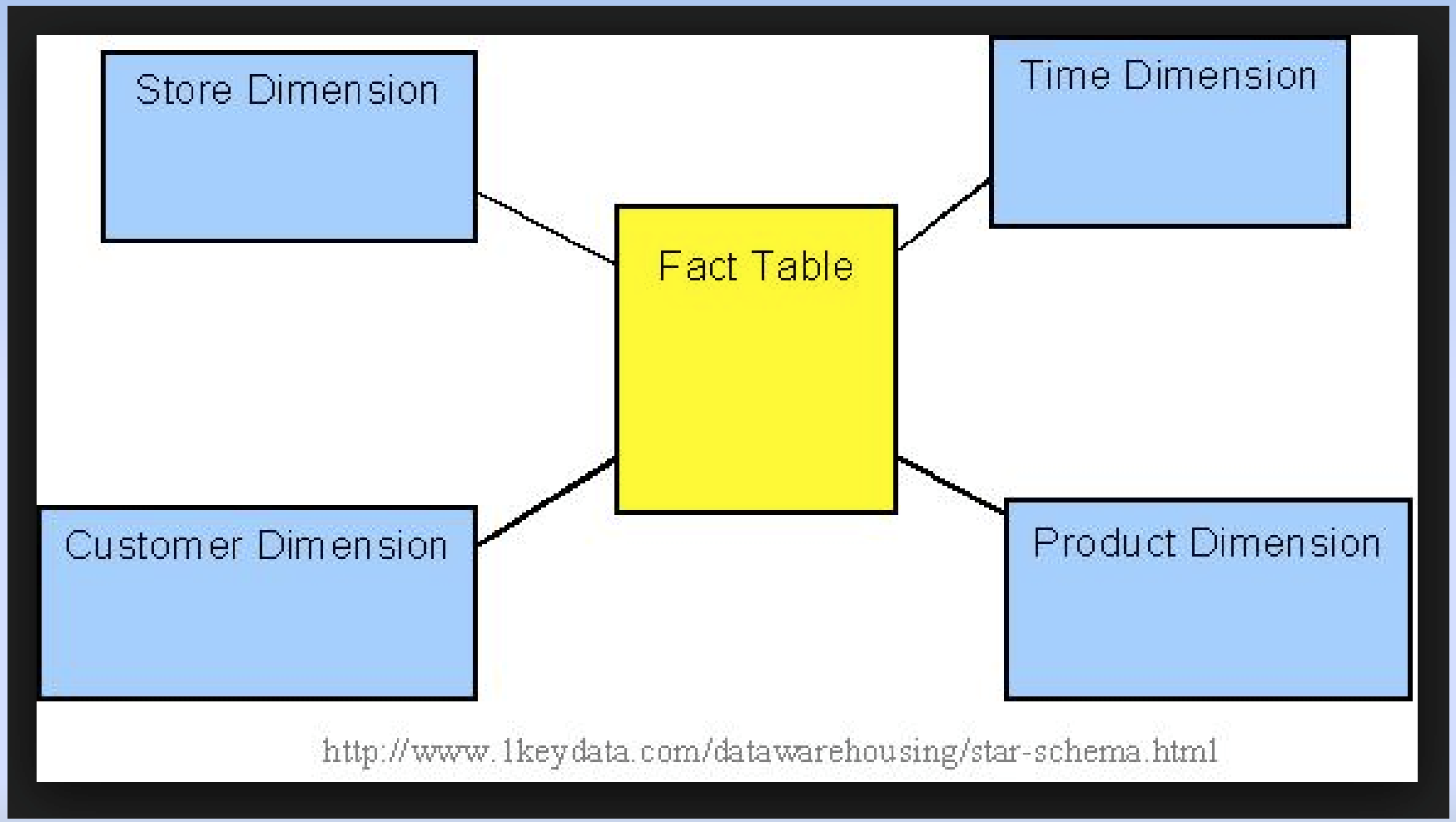
Table 30.1 Comparison of OLTP systems and data warehousing systems.

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

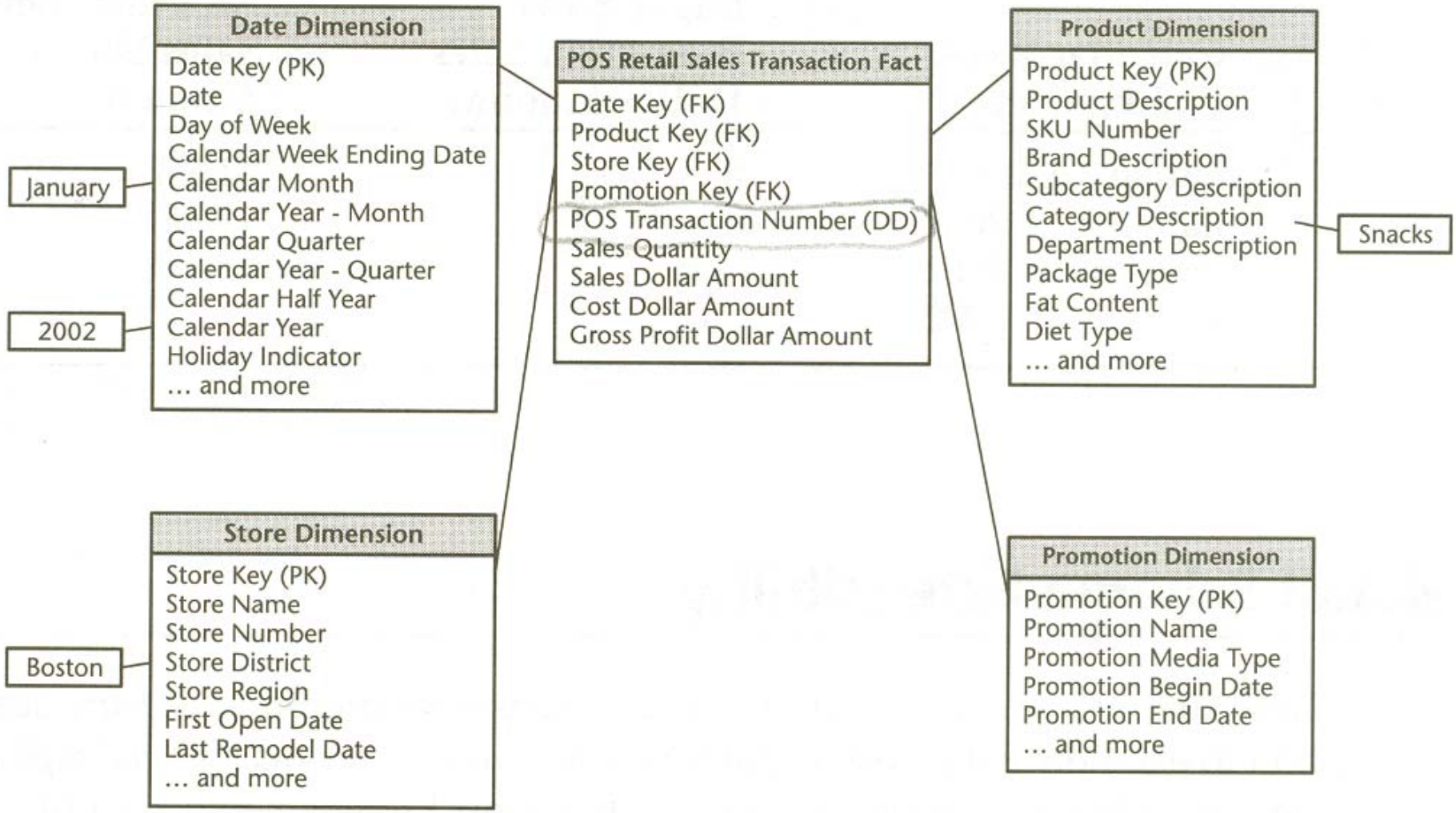


Figure 2.10 Querying the retail sales schema.

- Kimball p51

Retail Sales Product Dimension

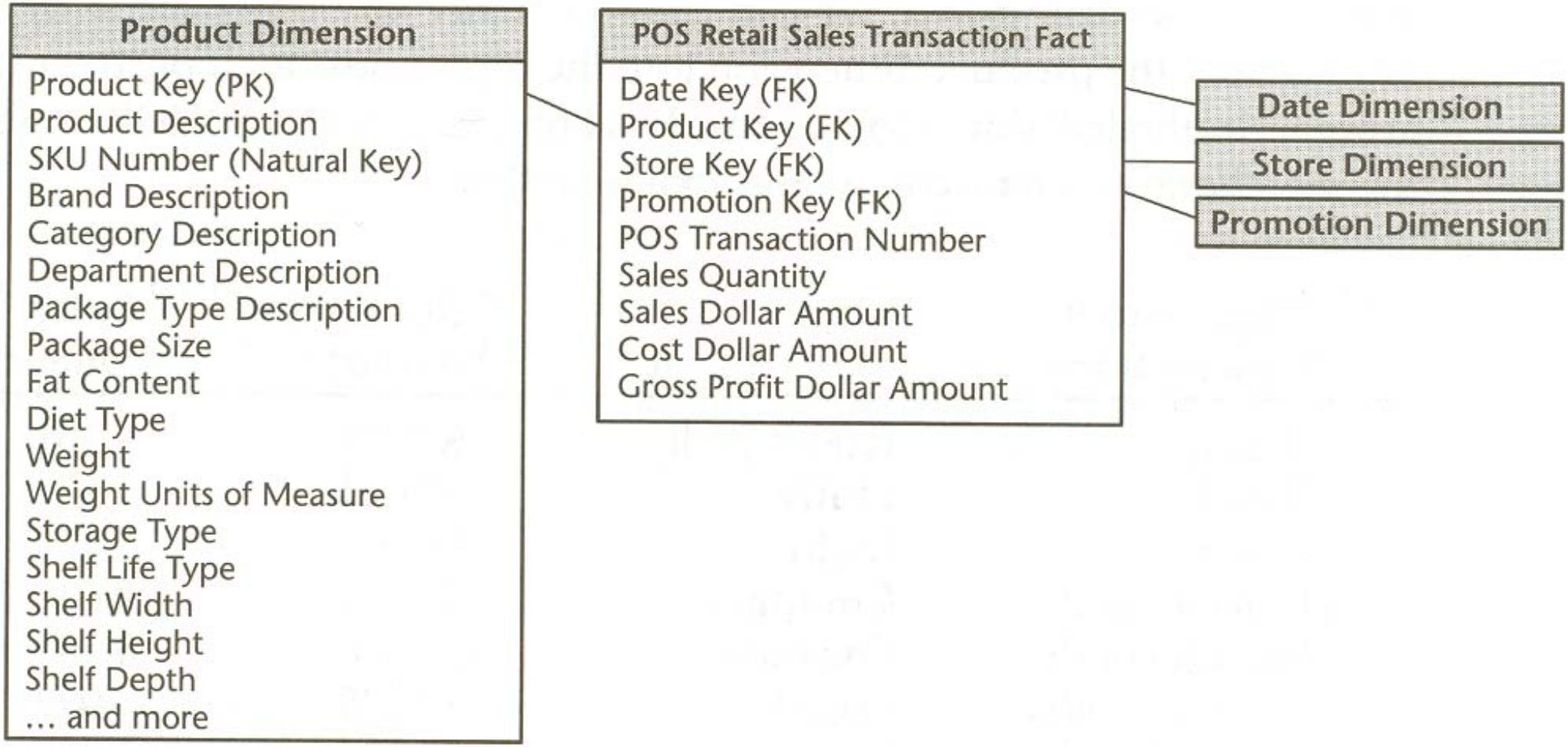
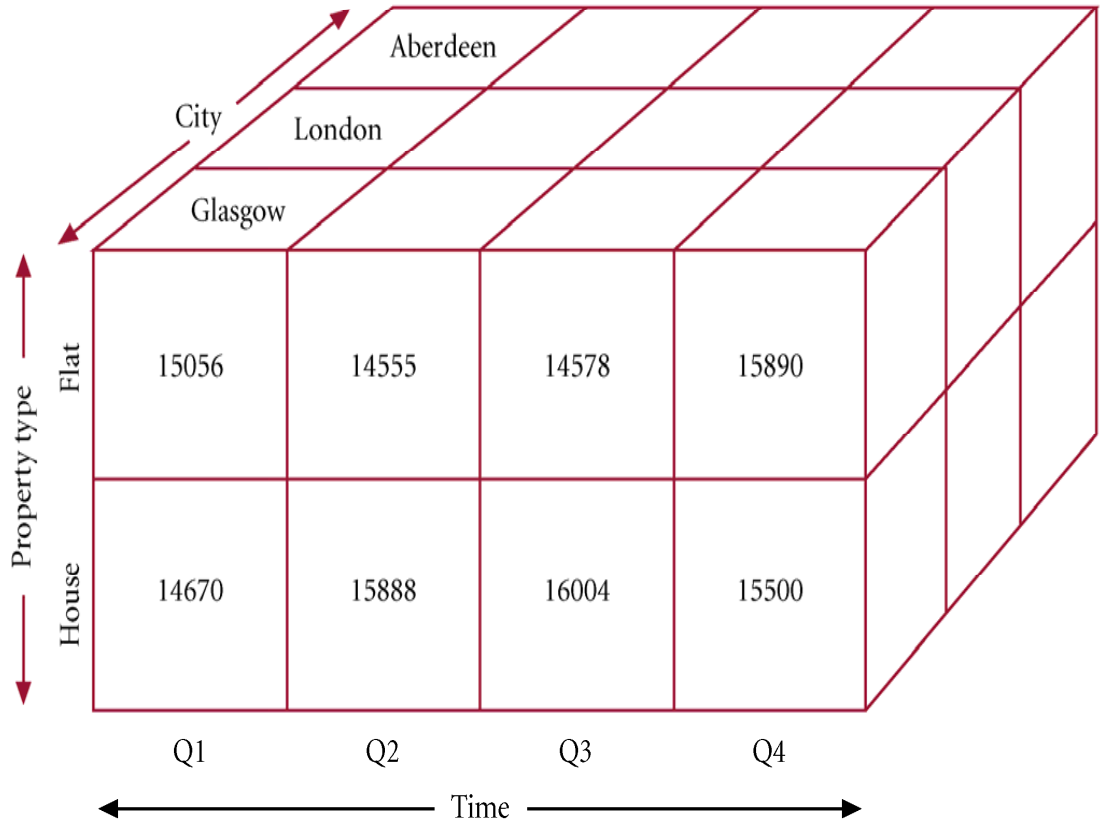


Figure 2.7 Product dimension in the retail sales schema.

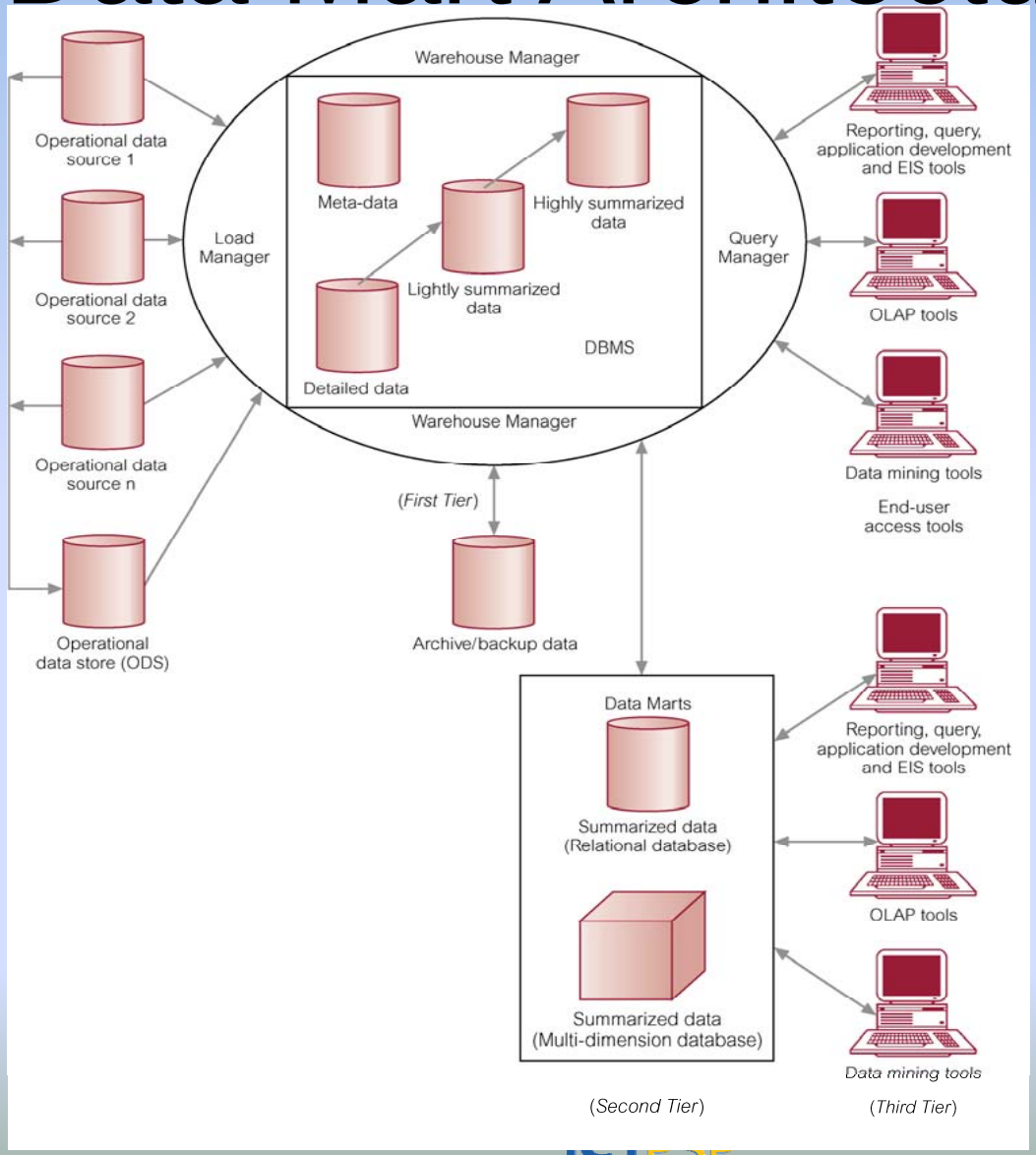
- Kimball p43

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
.....	.....	.....	.....
.....	.....	.....	.....



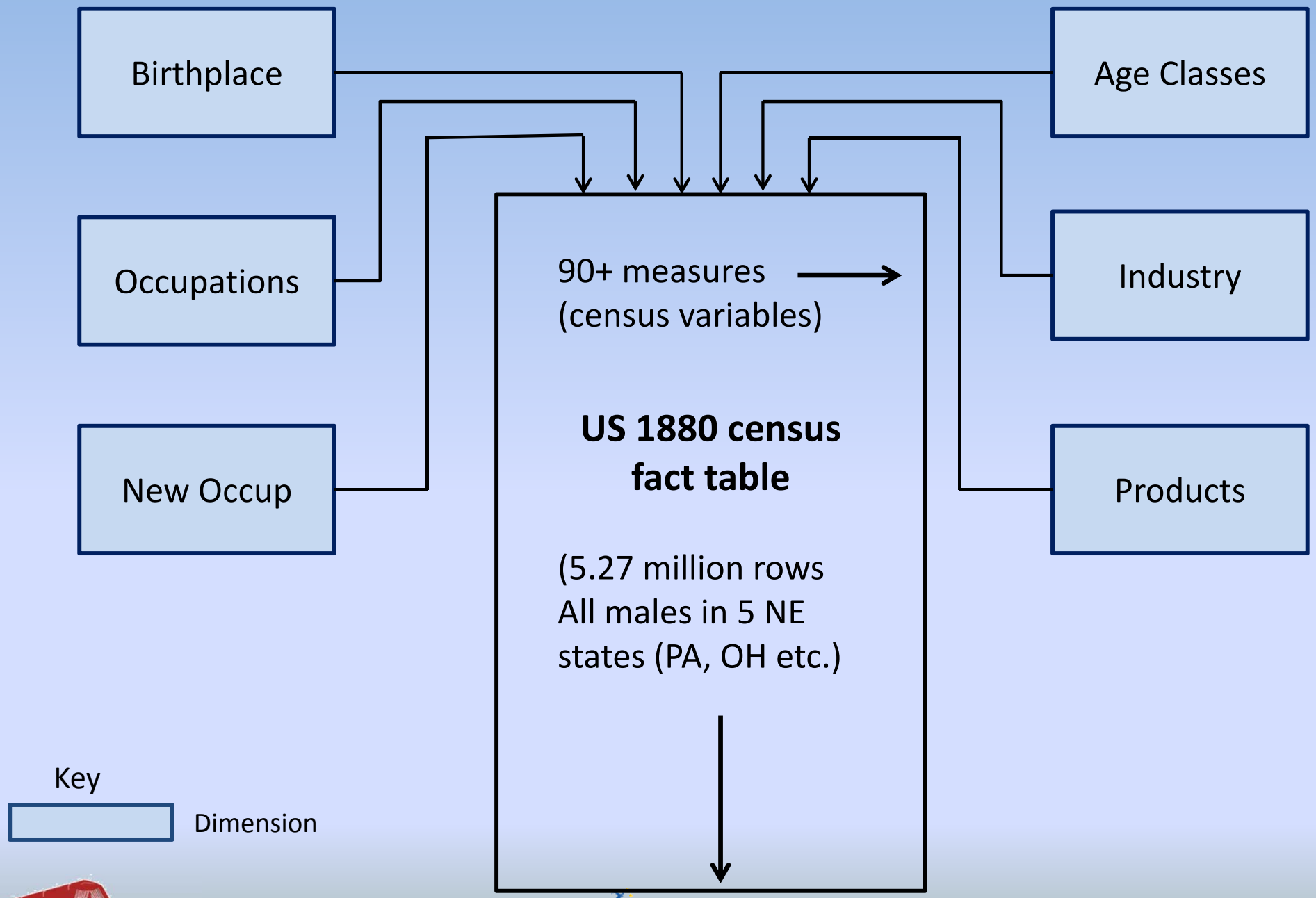
Typical Data Warehouse and Data Mart Architecture



Data Warehouse Environment

Census DW : the NAPP Dataset

- Approx. 53 million individual person records are available from the US 1880 census for academic use
- Downloadable in bulk from the NAPP website
- Individual details of name, place of birth, age, occupation, parental birthplaces etc.
- Most fields converted to numeric codes
- First pilot - 164,000 heavy industrial workers chosen for the 67 counties of Pennsylvania
- Second ‘industrial strength’ data warehouse – 5.27 million records - entire male population of five states in the NE USA
- Recent transfer to supercomputer - enlargement under consideration



Key

Dimension

iSQL*Plus Release 10.2.0.2.0 Production - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://tiger.iso.port.ac.uk:5560/isqlplus/workspace.uix?event=nextPage

Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows

iSQL*Plus Release 10.2.0.2.0 Product...

COUNTY_CODE	COUNTY_NAME	STATE_CODE	STATE_NAME	COUNTRY_CODE	COUNTRY_NAME	CONTINENT_CODE	CONTINENT_NAME
43330	Macedonia	43330	Macedonia	54230	Ottoman Empire	49900	Europe, n.e.c./n.s.
43400	Italy	99999	Unassigned	43400	Italy	49900	Europe, n.e.c./n.s.
43500	Malta	43500	Malta	41500	British possessions, Mediterranean	49900	Europe, n.e.c./n.s.
43600	Portugal	99999	Unassigned	43600	Portugal	49900	Europe, n.e.c./n.s.
43610	Azores	43610	Azores	16500	Portuguese North Atlantic Islands	99999	Unassigned
43620	Madeira Islands	43620	Madeira Islands	16500	Portuguese North Atlantic Islands	99999	Unassigned
43630	Cape Verde Islands	43630	Cape Verde Islands	16500	Portuguese North Atlantic Islands	99999	Unassigned
43640	St. Miguel	43610	Azores	16500	Portuguese North Atlantic Islands	99999	Unassigned
43800	Spain	99999	Unassigned	43800	Spain	49900	Europe, n.e.c./n.s.
45000	Austria	99999	Unassigned	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45010	Austro-Hungarian Empire	99999	Unassigned	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45020	Austria-Graz	45020	Austria-Graz	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45030	Austria-Linz	45030	Austria-Linz	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45040	Austria-Salzburg	45040	Austria-Salzburg	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45050	Austria-Tyrol	45050	Austria-Tyrol	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45060	Austria-Vienna	45060	Austria-Vienna	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45100	Bulgaria	45100	Bulgaria	54230	Ottoman Empire	49900	Europe, n.e.c./n.s.
45200	Czechoslovakia	45200	Czechoslovakia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45210	Bohemia	45210	Bohemia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45211	Bohemia-Moravia	45210	Bohemia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45212	Slovakia	45212	Slovakia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45300	German Empire	99999	Unassigned	45300	German Empire	49900	Europe, n.e.c./n.s.
45301	Berlin	45301	Berlin	45300	German Empire	49900	Europe, n.e.c./n.s.
45311	Baden	45311	Baden	45300	German Empire	49900	Europe, n.e.c./n.s.

Done

Secure Search

iSQL*Plus Release 10.2.0.2.0 Production - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://tiger.iso.port.ac.uk:5560/isqlplus/workspace.uix?bajaPage=result=

Most Visited | Customize Links | Free Hotmail | Windows Marketplace | Windows Media | Windows

stereoview coal, great deals on Collecti... | iSQL*Plus Release 10.2.0.2.0 Pro...

LEV1_CODE	LEVL1_DESC	LEV2_CODE	LEVL2_DESC	LEV3_CODE	LEVL3_DESC	LEV4_CODE	LEVL4_DESC	LEV5_CODE	LEVL5_DESC	LEV6_CODE	LEVL6_DESC	LEV7_CODE	LEVL7_DESC
10112109440095	Other Stable Worker	101121094	Stable Workers	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109623000	Engineer or Stationary Engineer or Stationary Engineman	101121096	Stationary Engine Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109832000	Locomotive Driver or Railroad Engineman	101121098	Transport Equipment Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109857001	Teamsters Helper	101121098	Transport Equipment Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109900360	Rock Dump Man	101121099	Workers nec	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109900010	Ash Wheeler	101121099	Workers nec	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10111102313021	Assistant Fire Boss	101111023	Foremen and Supervisors	10111102	Administrative and Managerial	101111	Company Men	10111	Inside	1011	Production	101	Anthracite
10111107112090	Tunnelman	101111071	Miners Quarrymen and Well-Drillers	10111107	Mining Metal Manufacture and Textiles	101111	Company Men	10111	Inside	1011	Production	101	Anthracite
10111109513000	Stone Mason	101111095	Mine Development Workers	10111109	Production Construction and Transport	101111	Company Men	10111	Inside	1011	Production	101	Anthracite

9 rows selected

Done | Secure Search | McAfee

Example Codes for Anthracite Mining Occupations

OLAP Query Results Page : Allegany County - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Most Visited | Customize Links | Free Hotmail | Windows Marketplace | Windows Media | Windows

Novell WebAccess (Richard Healey) | OLAP Query Results Page : Allega...

OLAP Query Results Page : Allegany County

Birthplace	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	105-109	110-114	115-119	120-124	125-129	Unknown	Total
At sea	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Austro-Hungarian Empire	6	0	1	2	5	5	5	6	3	2	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	39
Brazil	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Canada	0	3	9	14	11	15	14	8	17	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96
Egypt	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
France	0	0	0	0	0	1	0	1	3	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10
German Empire	5	7	19	18	33	62	93	69	94	133	112	81	74	51	41	15	8	3	0	0	0	0	0	0	0	0	0	918
Italy	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Missing/blank	2	0	2	1	4	4	0	0	3	1	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23
Netherlands	0	0	0	0	2	1	4	3	1	1	2	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	19
Norway	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Russian Empire	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Spain	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sweden	0	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Switzerland	0	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	7
Unassigned	0	0	1	1	0	1	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
United Kingdom, n.s.	9	47	170	202	216	243	309	337	276	194	213	115	140	57	40	19	9	2	0	0	0	0	0	0	0	0	0	2598
United States, n.s.	3028	2748	2180	1635	1537	1118	821	671	446	369	282	200	187	114	74	47	23	1	1	0	0	0	0	0	0	0	1	15483
Unknown	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Done | Secure Search | McAfee

Insights for Database Archiving/Data Mining

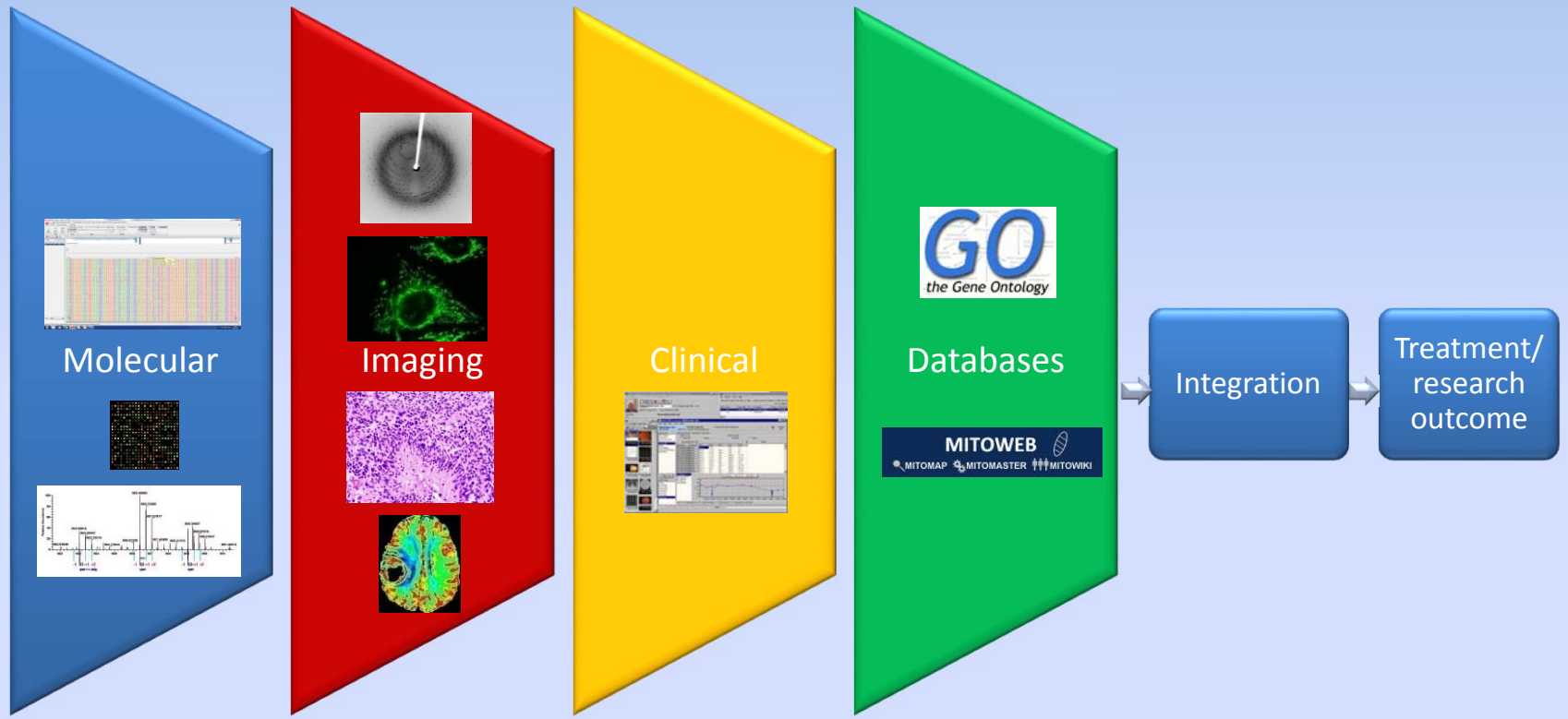
- Tight coupling of dimension and fact table keys removes problem of data mismatches
- Dimensions are ‘mini-repositories’ of valuable structures for data standardisation across database snapshots and data tables from different sources (can be used outside DW also – e.g. occupations in B&O payrolls 1842-1857)
- Time dimension useful for multi-year census data, also for business records – monthly payrolls etc., but such a general purpose dimension would apply across wide range of archived tables (as would geography, industry, occupation dimensions – latter being used for city directory data also)
- Large ‘upfront’ investment in implementing dimensions but considerable payoff as archive grows
- Present day census DW applications include Bulgarian 2011 census with SDMX interface to EUROSTAT census hub

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

Slide from Dr Rhiannon Lloyd.

Types of data



Biobanks

SEBTA
South of England Brain Tumour Alliance

THE FROZEN ARK
SAVING THE DNA OF ENDANGERED SPECIES

Xenopus

ICTPSP

Slide from Dr Rhiannon Lloyd.

E-ARK

3Vs

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


Q&A