CSCI 5408 Data Analytics: DM and DW Tech (Week 9)

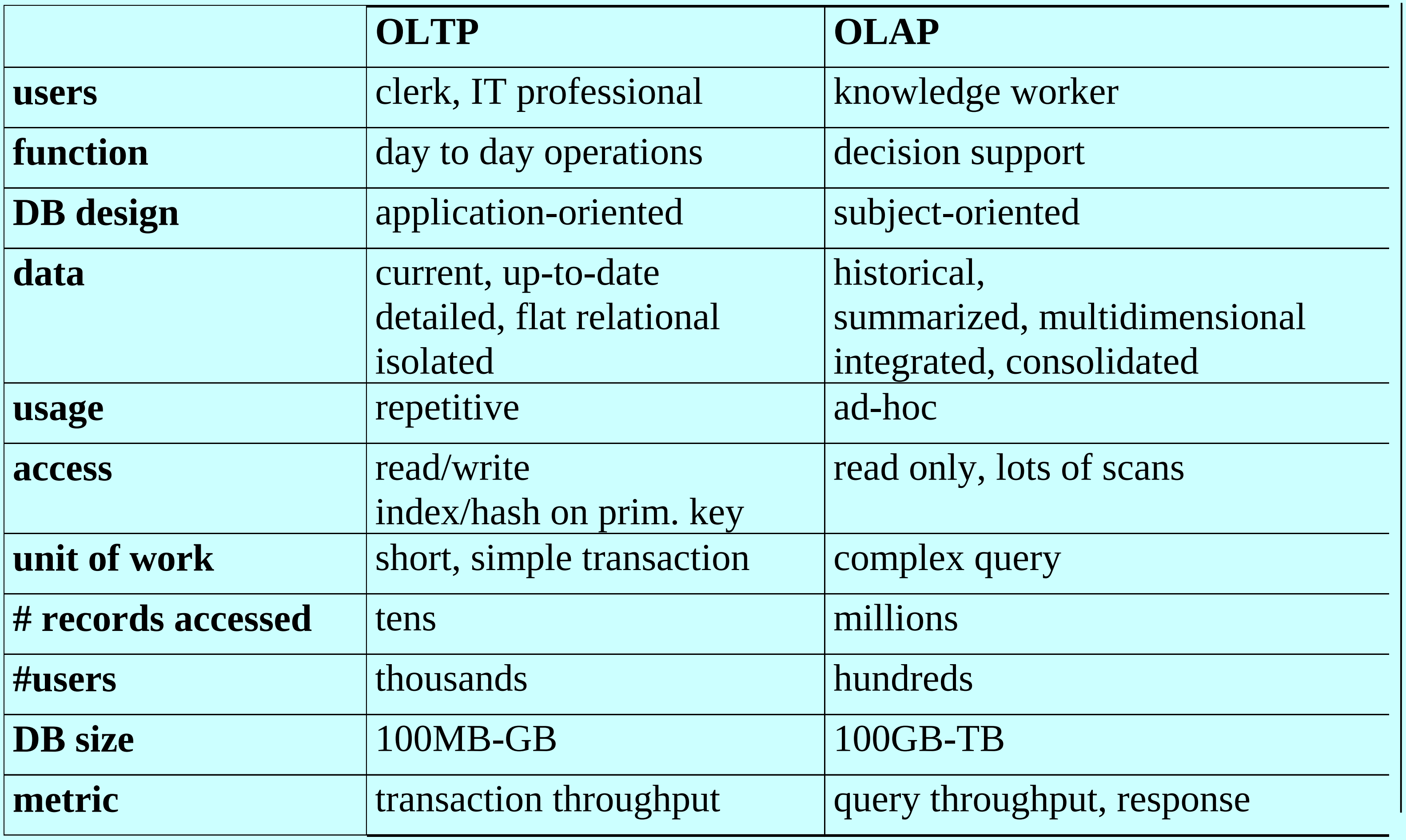
- Ass4 Due: Mar 14
  - Brightspace: Assignment 4, Tutorial slides, etc.
  - Help hours: Fri, 1:00-2:30PM, CS 233
- Write answers for review questions
  - Final Exam: Apr 20, 3:30-5:30 PM
- Reading: Lectures: 13-14, Text: Ch4 of 3rd edition, or Ch3 of 2nd edition

3. Data Warehouses and OLAP

(Textbook: Ch4 of 3rd edition, or Ch3 of 2nd edition)

- Objectives of DW/OLAP
- What is a DW?
- Multi-dimensional data space model
- DW schemas
- OLAP operations
- Aggregations
- DW architecture
- From DW to DM

A Comparison: OLTP vs. OLAP



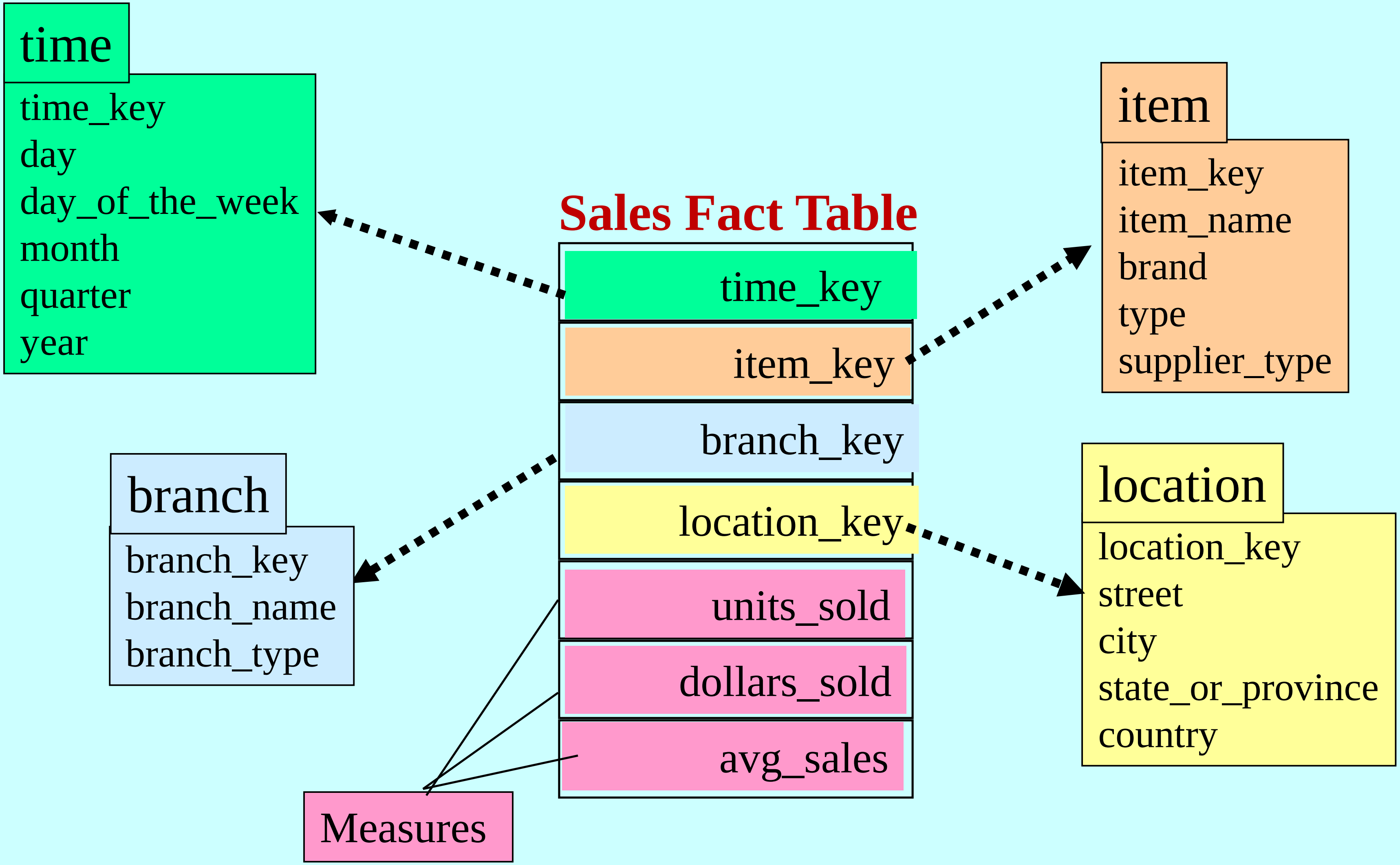
	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc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

https://courses.cs.washington.edu/courses/csep573/01sp/lectures/class1/sld025.htm

DW Schemas: Conceptual Models

- Basic DW Structure: dimensions & measures
- **Star schema:** A fact table in the middle connected to a set of dimension tables
- **Snowflake schema:** A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
- **Fact constellations:** Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Example of Star Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

location

location_key
street
city
state_or_province
country

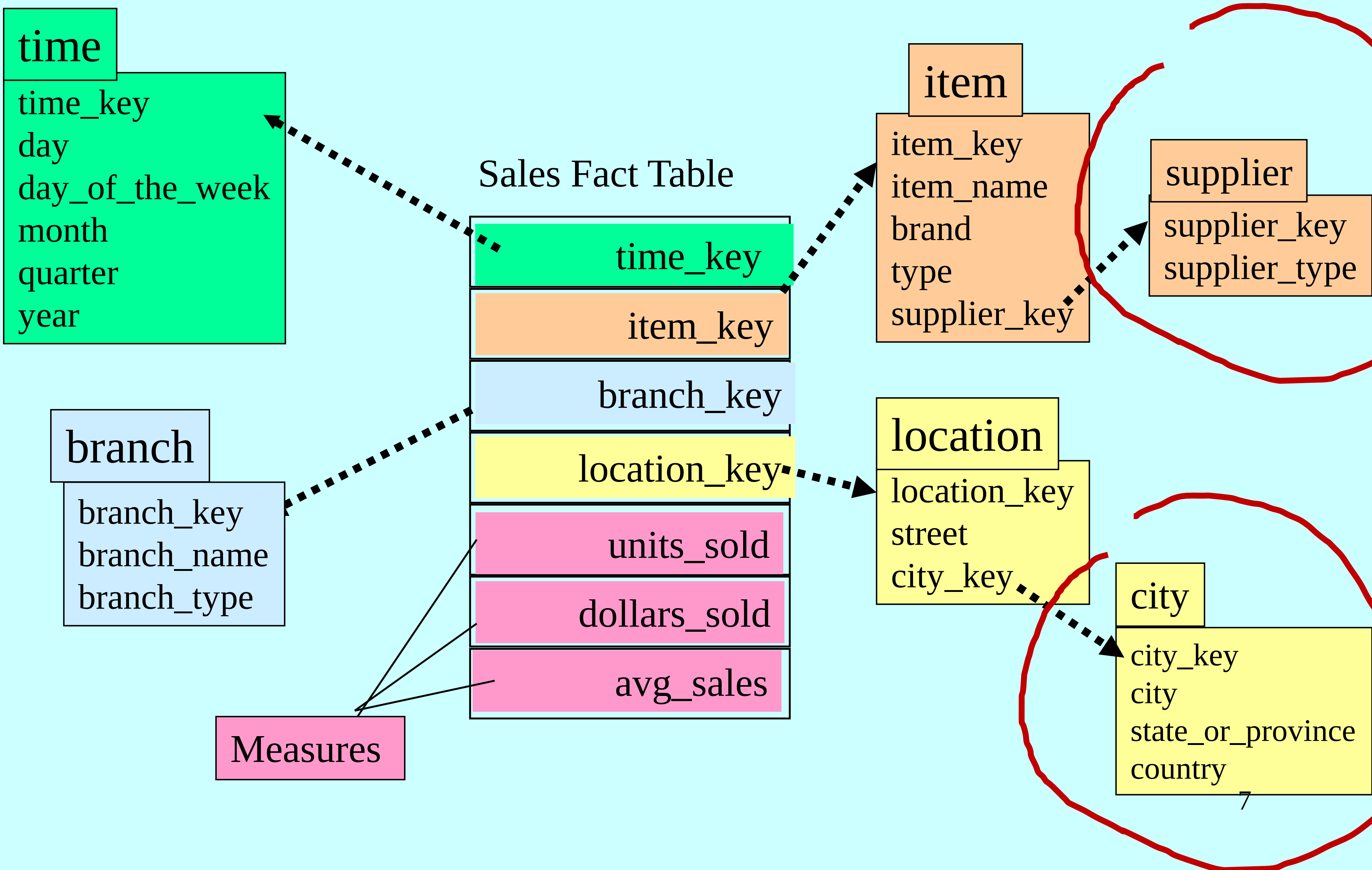
Measures

Example of Star Schema (cont)

- Star schema is the most common used schema for OLAP applications, and used as **Data Mart** for department-level DW

- The star schema is simple but some **redundancy** may occur for dimension tables
  - E.g., Location {location_key, street, city, province, country}
    (102 St…, Vancouver, British Columbia, Canada)
    (206 St…, Vancouver, British Columbia, Canada)

Example of Snowflake Schema



time
time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item
item_key
item_name
brand
type
supplier_key

supplier
supplier_key
supplier_type

branch
branch_key
branch_name
branch_type

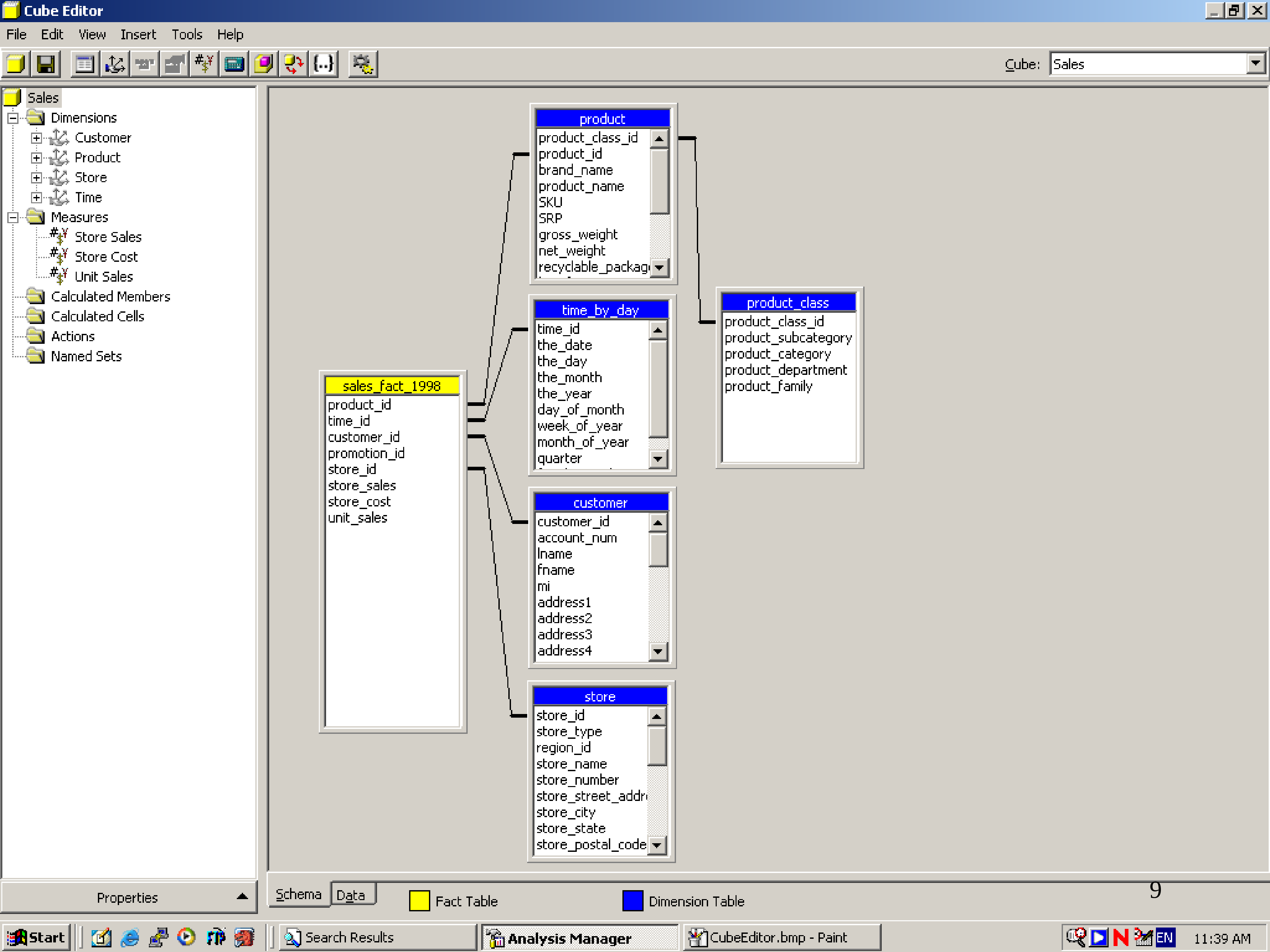
location
location_key
street
city_key

city
city_key
city
state_or_province
country

Measures

Example of Snowflake Schema (cont)

- A variation of star schema, in which the dimension tables are normalized
- Main purpose: saving space and for easier maintenance
  - Normalizing large dimension tables for saving storage space
  - However, it can reduce the effectiveness of browsing since more joints will be needed to execute a query
    - ➢Keeping small dimension tables as it is for reducing the cost.
    - ➢Performance degradation because of join operation on multiple tables.



Cube: Sales

- Sales
  - Dimensions
    - Customer
    - Product
    - Store
    - Time
  - Measures
    - Store Sales
    - Store Cost
    - Unit Sales
  - Calculated Members
  - Calculated Cells
  - Actions
  - Named Sets

sales_fact_1998
product_id
time_id
customer_id
promotion_id
store_id
store_sales
store_cost
unit_sales

product
product_class_id
product_id
brand_name
product_name
SKU
SRP
gross_weight
net_weight
recyclable_packag

time_by_day
time_id
the_date
the_day
the_month
the_year
day_of_month
week_of_year
month_of_year
quarter

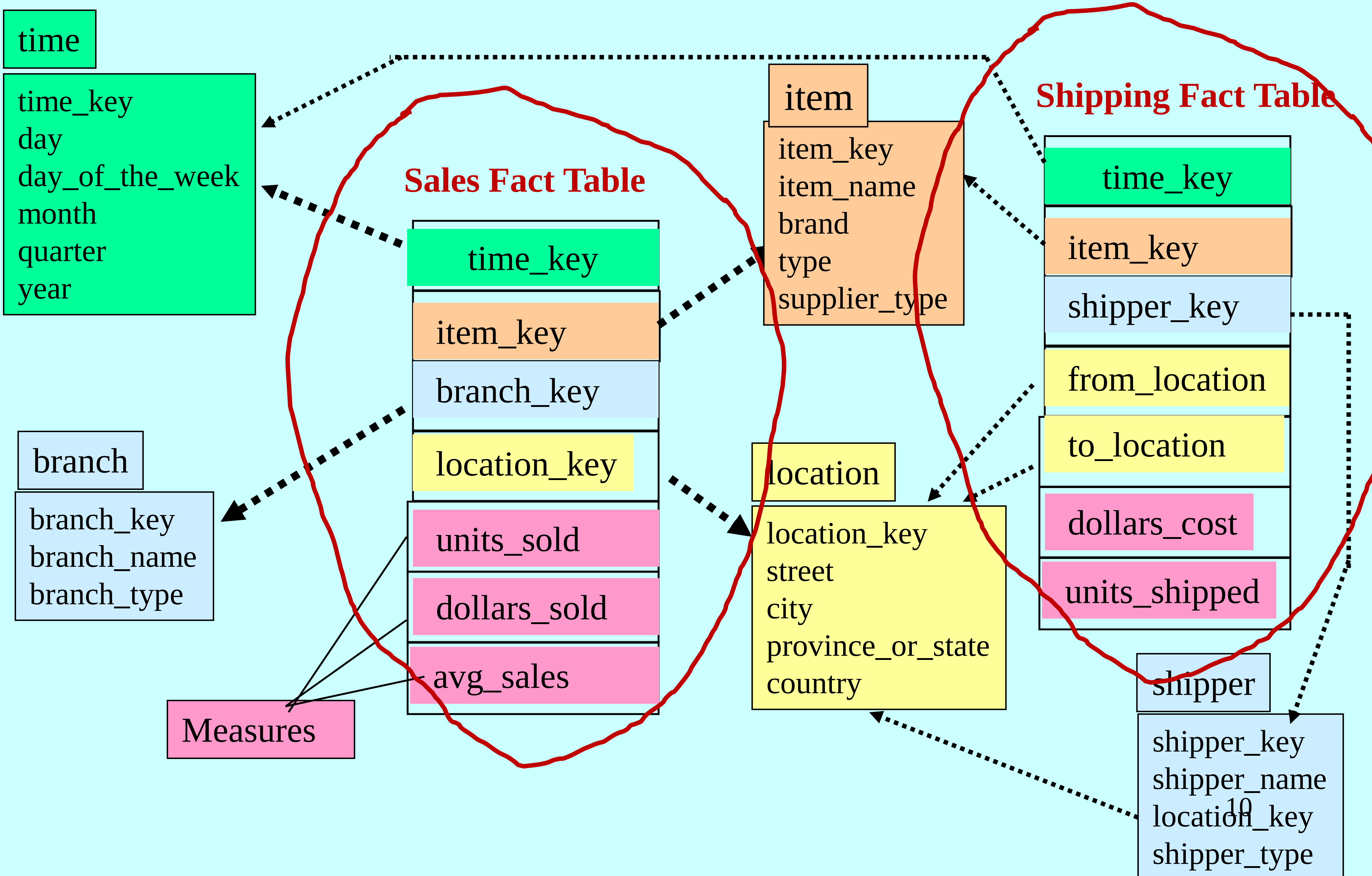
customer
customer_id
account_num
lname
fname
mi
address1
address2
address3
address4

store
store_id
store_type
region_id
store_name
store_number
store_street_addr
store_city
store_state
store_postal_code

product_class
product_class_id
product_subcategory
product_category
product_department
product_family

🟨 Fact Table
🟦 Dimension Table

Example of Fact Constellation



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

branch

branch_key
branch_name
branch_type

location

location_key
street
city
province_or_state
country

shipper

shipper_key
shipper_name
location_key
shipper_type

Measures

Example of Fact Constellation (cont)

- Fact constellation schema is for sophisticate DW
  - It is for the design of a DW with multiple subjects, such as for a large corporation which needs information for quickly updating big pictures of entire organization, etc.

DW Schema Definition Language

- Cube Definition (Fact Table)

  define cube <cube_name> [<dimension_list>]: <measure_list>

- Dimension Definition (Dimension Table)

  define dimension <dimension_name> **as** (<attribute_or_subdimension_list>)

- Special Case (Shared Dimension Tables)
  - First time as "cube definition"
  - **define dimension** <dimension_name> **as** <dimension_name_first_time> **in cube** <cube_name_first_time>

E.g., Define Star Schema for “Sales”:

define cube sales_star [time, item, branch, location]:

dollars_sold = sum(sales_in_dollars),

avg_sales = avg(sales_in_dollars),

units_sold = count(*)

define dimension time **as** (time_key, day, day_of_week, month, quarter, year)

define dimension item **as** (item_key, item_name, brand, type, supplier_type)

define dimension branch **as** (branch_key, branch_name, branch_type)

define dimension location **as** (location_key, street, city, province_or_state, country)

E.g. Define Snowflake Schema for “Sales”

define cube sales_snowflake [time, item, branch, location]:

dollars_sold = sum(sales_in_dollars),

avg_sales = avg(sales_in_dollars),

units_sold = count(*)

define dimension time **as** (time_key, day, day_of_week, month, quarter, year)

define dimension Item **as** (item_key, item_name, brand, type, supplier(supplier_key, supplier_type))

define dimension branch **as** (branch_key, branch_name, branch_type)

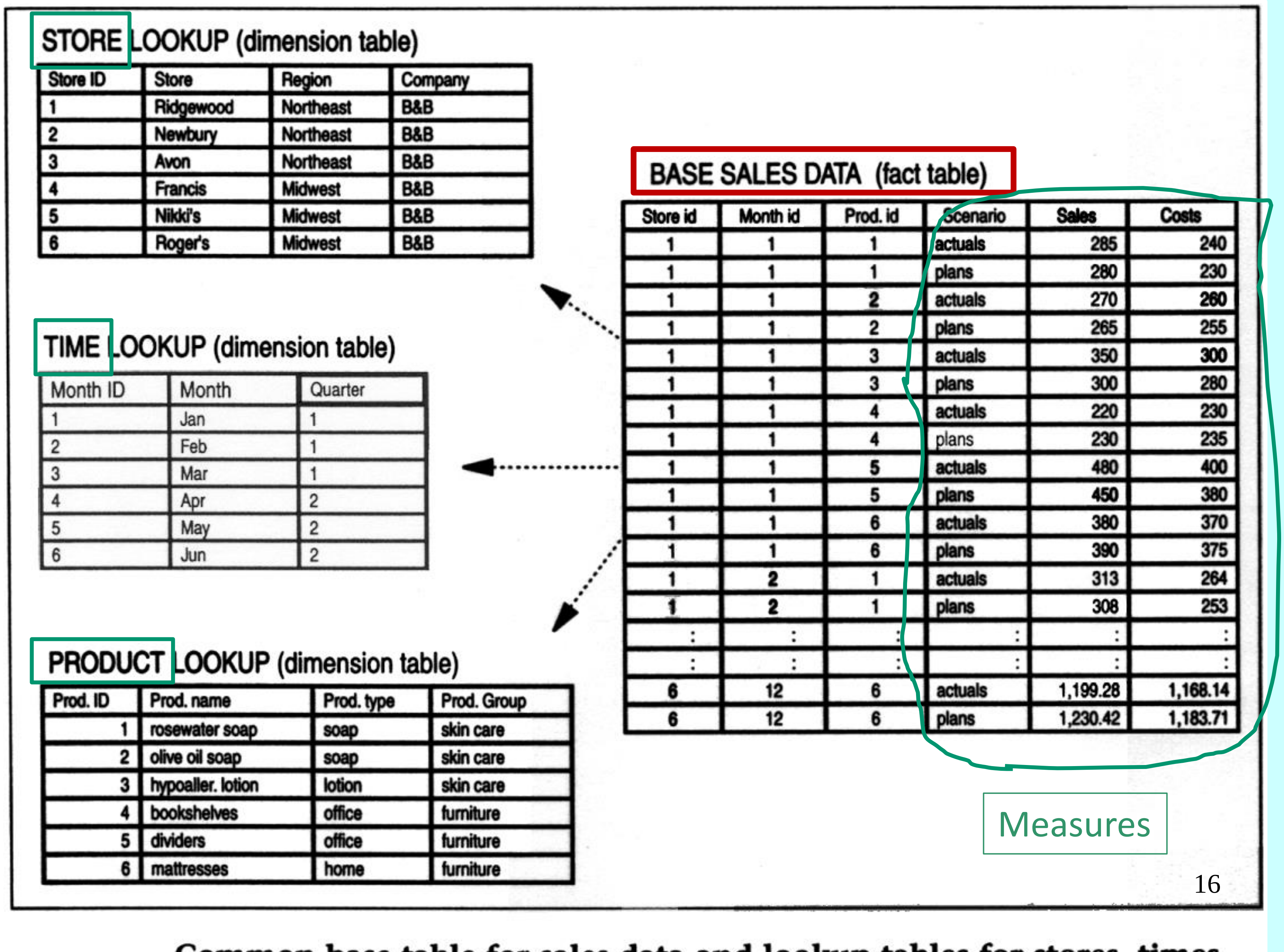
define dimension location **as** (location_key, street, city(city_key, province_or_state, country))

Embedded table

Define Fact Constellation Schema for “Sales” and “Shipping”

```
define cube Sales [time, item, branch, location]:
                dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars),
                    units_sold = count(*)
define dimension Time as (time_key, day, day_of_week, month, quarter, year)
define dimension Item as (item_key, item_name, brand, type, supplier_type)
define dimension Branch as (branch_key, branch_name, branch_type)
define dimension Location as (location_key, street, city, province_or_state, country)
define cube Shipping [time, item, shipper, from_location, to_location]:
                dollar_cost = sum(cost_in_dollars), unit_shipped = count(*)
define dimension Time as time in cube sales
define dimension Item as item in cube sales
define dimension Shipper as (shipper_key, shipper_name, location as location in cube
      sales, shipper_type)
define dimension from_location as location in cube sales
define dimension to_location as location in cube sales
```

E.g., A DW instance: How data is materialized?



STORE LOOKUP (dimension table)

Store ID	Store	Region	Company
1	Ridgewood	Northeast	B&B
2	Newbury	Northeast	B&B
3	Avon	Northeast	B&B
4	Francis	Midwest	B&B
5	Nikki's	Midwest	B&B
6	Roger's	Midwest	B&B

TIME LOOKUP (dimension table)

Month ID	Month	Quarter
1	Jan	1
2	Feb	1
3	Mar	1
4	Apr	2
5	May	2
6	Jun	2

PRODUCT LOOKUP (dimension table)

Prod. ID	Prod. name	Prod. type	Prod. Group
1	rosewater soap	soap	skin care
2	olive oil soap	soap	skin care
3	hypoaller. lotion	lotion	skin care
4	bookshelves	office	furniture
5	dividers	office	furniture
6	mattresses	home	furniture

BASE SALES DATA (fact table)

Store id	Month id	Prod. id	Scenario	Sales	Costs
1	1	1	actuals	285	240
1	1	1	plans	280	230
1	1	2	actuals	270	260
1	1	2	plans	265	255
1	1	3	actuals	350	300
1	1	3	plans	300	280
1	1	4	actuals	220	230
1	1	4	plans	230	235
1	1	5	actuals	480	400
1	1	5	plans	450	380
1	1	6	actuals	380	370
1	1	6	plans	390	375
1	2	1	actuals	313	264
1	2	1	plans	308	253
:	:	:	:	:	:
:	:	:	:	:	:
6	12	6	actuals	1,199.28	1,168.14
6	12	6	plans	1,230.42	1,183.71

Measures

Common base table for sales data and lookup tables for stores, times

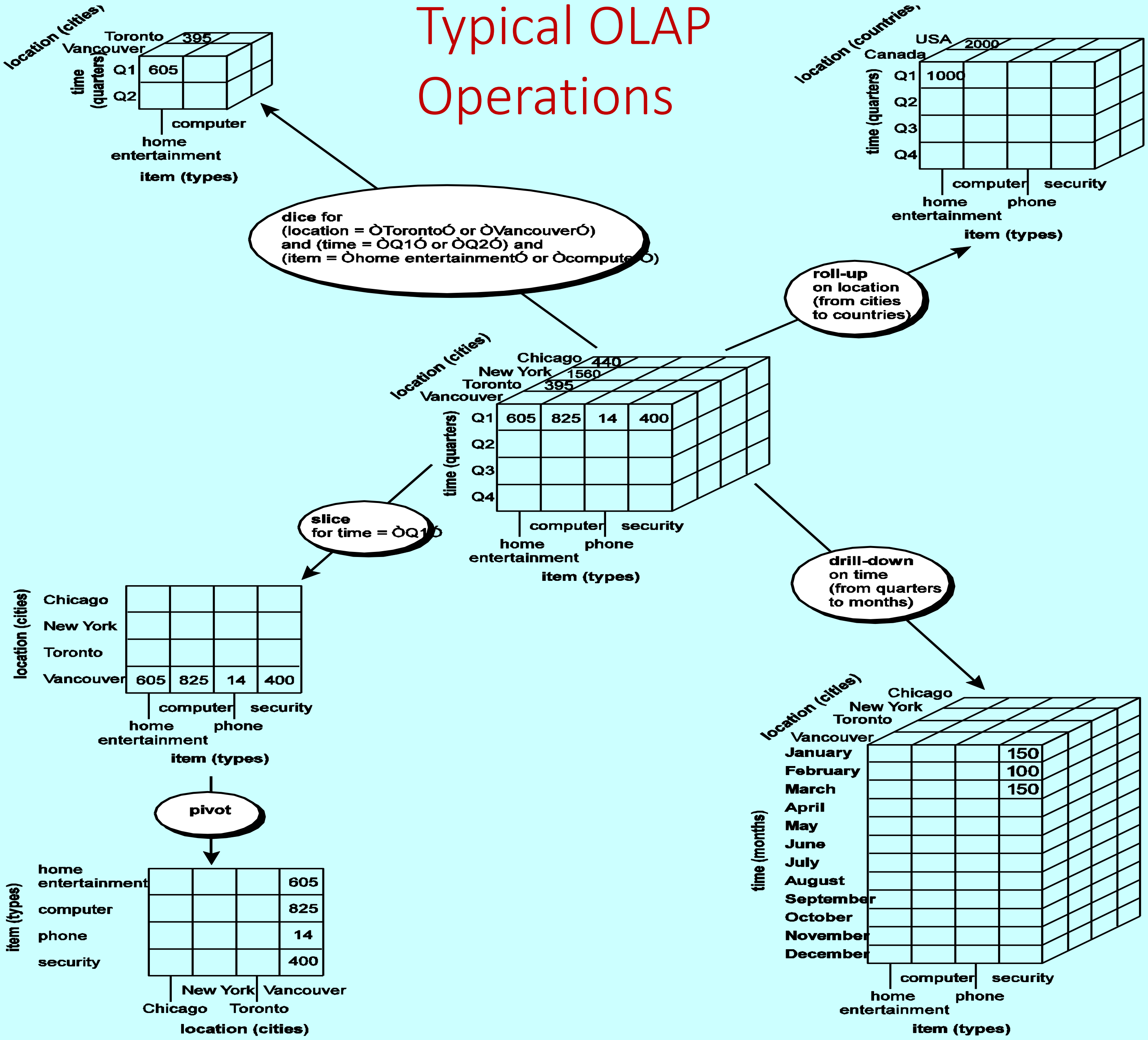
OLAP Operations

- On-Line Analytical Processing (OLAP)
  - It is an approach to answering multi-dimensional analytical (MDA) queries swiftly in computing.
- OLAP Operators
  - They are the tools enable users to analyze multidimensional data interactively from multiple perspectives, consisting of four basic analytical operations: **roll-up** (consolidation), **drill-down**, and **slicing** and **dicing**

OLAP Operations (cont)

- **Roll up:** (summarize data)
  - By climbing up hierarchy or by dimension reduction
- **Drill down:** (reverse of roll-up)
  - From higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Slice:** (project)
  - By choosing a single value for one of its dimensions, creating a new cube with one fewer dimension
- **Dice:** (select)
  - By produces a subcube by allowing the analyst to pick specific values of multiple dimensions
- **Pivot:** (rotate)
  - reorient the cube, visualization, 3D to series of 2D planes.
- **Other operations:**
  - drill across: involving (across) more than one fact table
  - drill through: through the bottom level of the cube to its back-end relational tables (using SQL)

Typical OLAP Operations



dice for
(location = ÒTorontoÓ or ÒVancouverÓ)
and (time = ÒQ1Ó or ÒQ2Ó) and
(item = Òhome entertainmentÓ or ÒcomputerÓ)

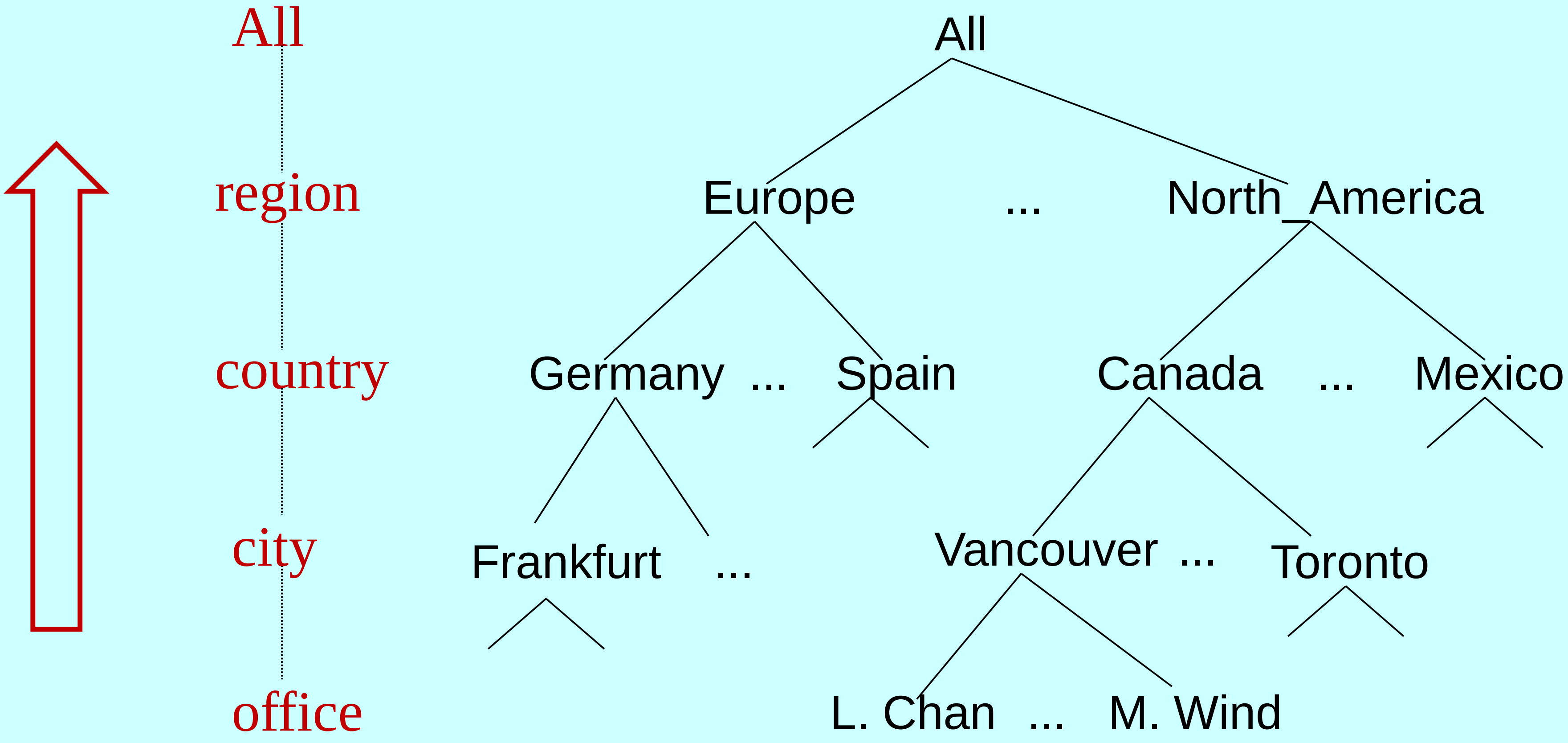
roll-up
on location
(from cities
to countries)

slice
for time = ÒQ1Ó

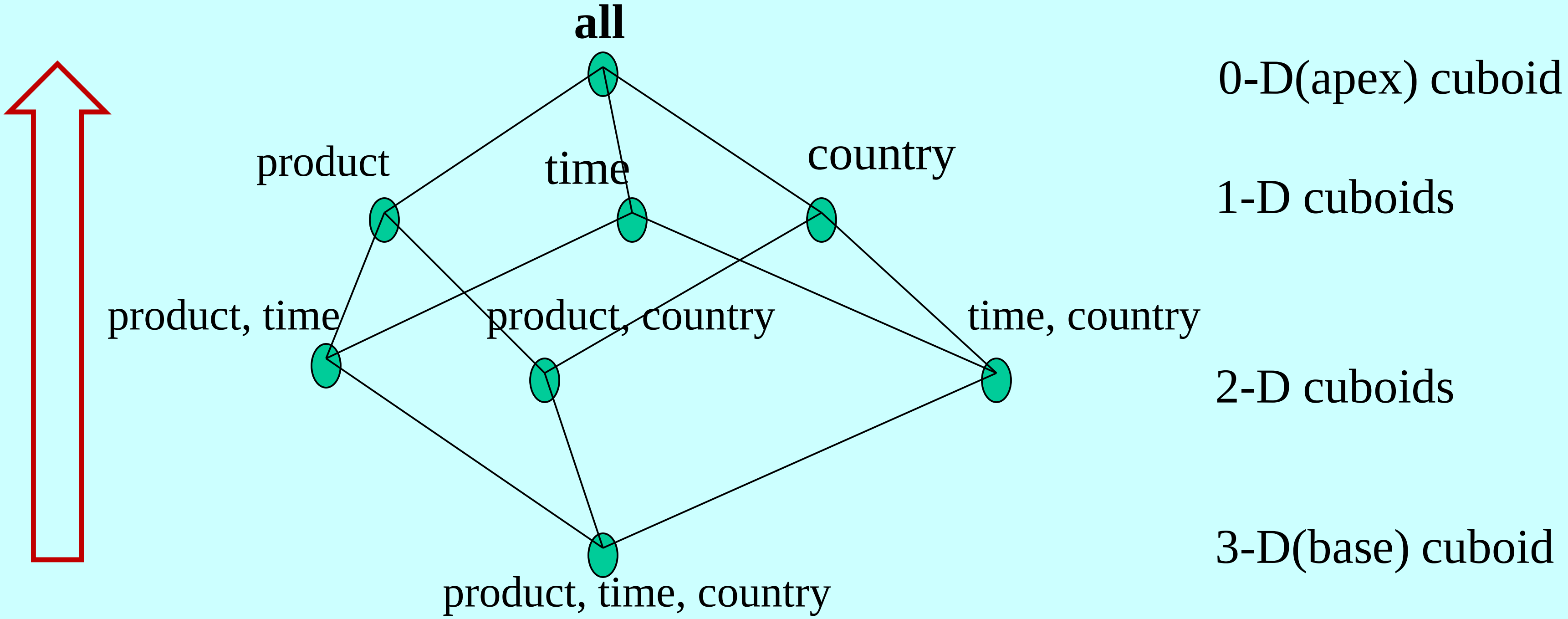
drill-down
on time
(from quarters
to months)

pivot

Aggregation by **rolling-up** along concept hierarchy

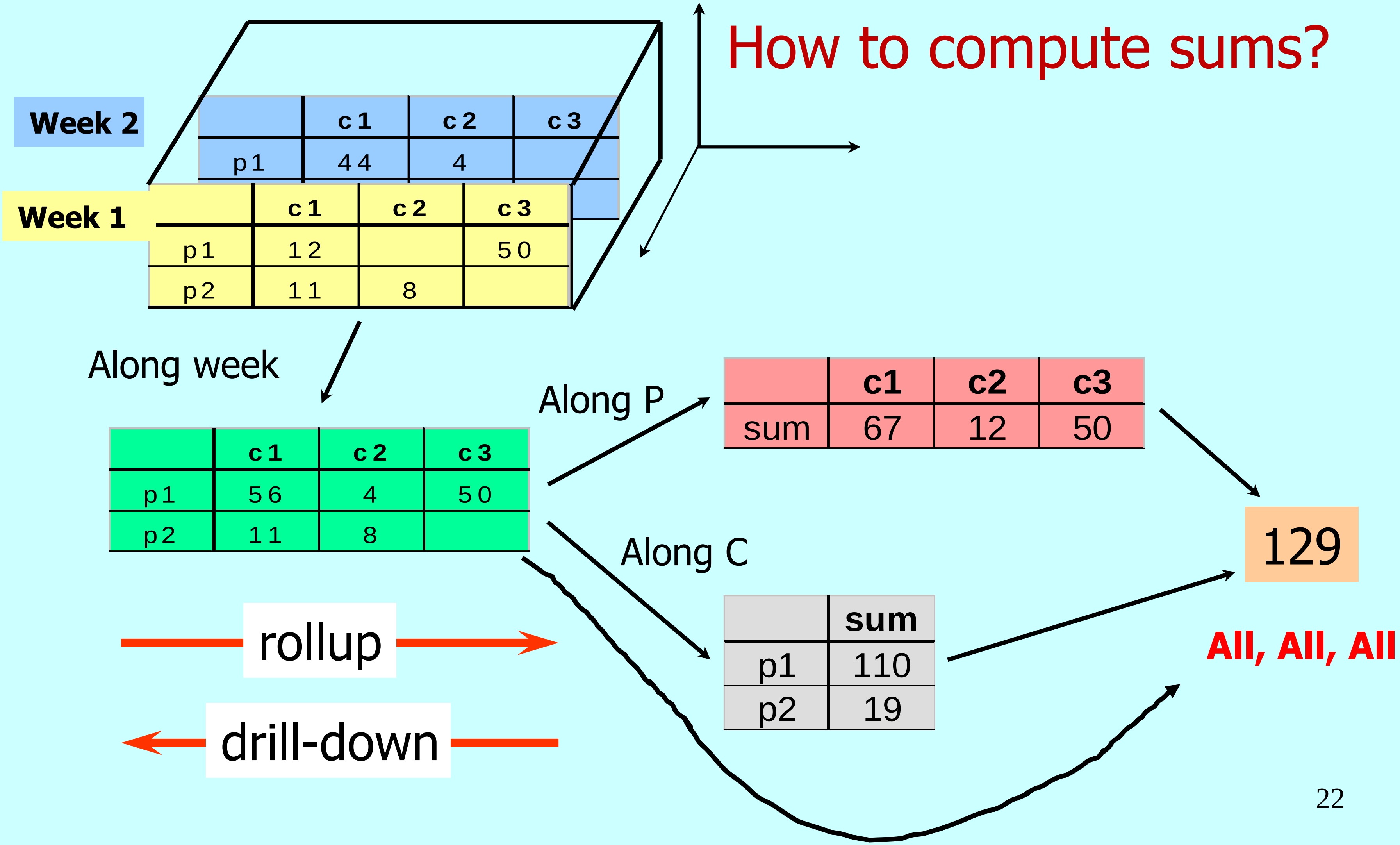


Aggregation by **rolling-up** (or slicing by choosing “all”) along dimensional lattice



Cube Aggregation by **roll-up**:

E.g., Store C sold # of P in week X.



How to compute sums?

Week 2

	c1	c2	c3
p1	44	4	

Week 1

	c1	c2	c3
p1	12		50
p2	11	8	

Along week

	c1	c2	c3
p1	56	4	50
p2	11	8	

Along P

	c1	**c2**	**c3**
sum	67	12	50

Along C

	sum
p1	110
p2	19

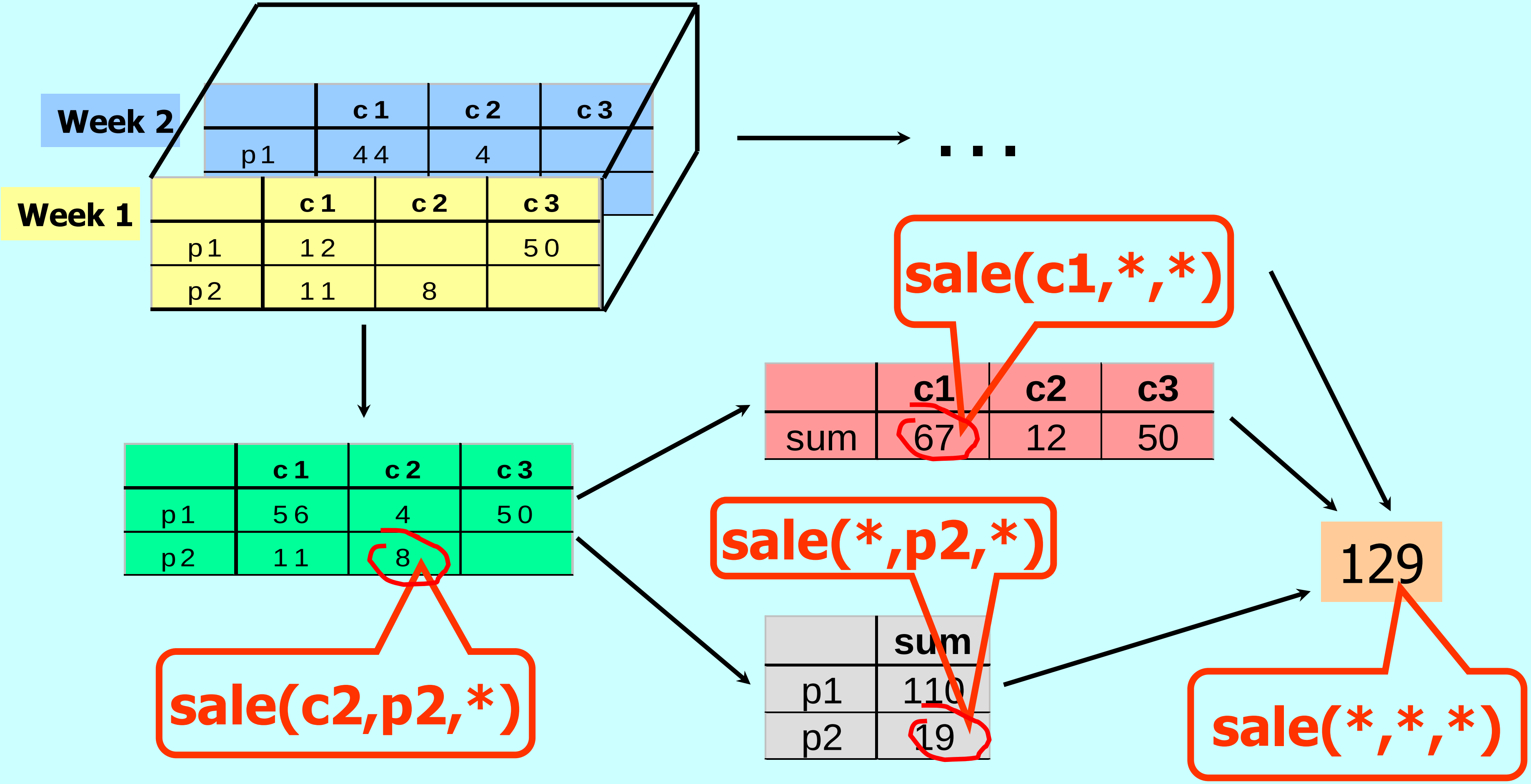
129

All, All, All

rollup

drill-down

Specify Roll Up Operations



Week 2

	c1	c2	c3
p1	44	4	

Week 1

	c1	c2	c3
p1	12		50
p2	11	8	

. . .

	c1	c2	c3
p1	56	4	50
p2	11	8	

sale(c2,p2,*)

sale(c1,*,*)

	c1	c2	c3
sum	67	12	50

sale(*,p2,*)

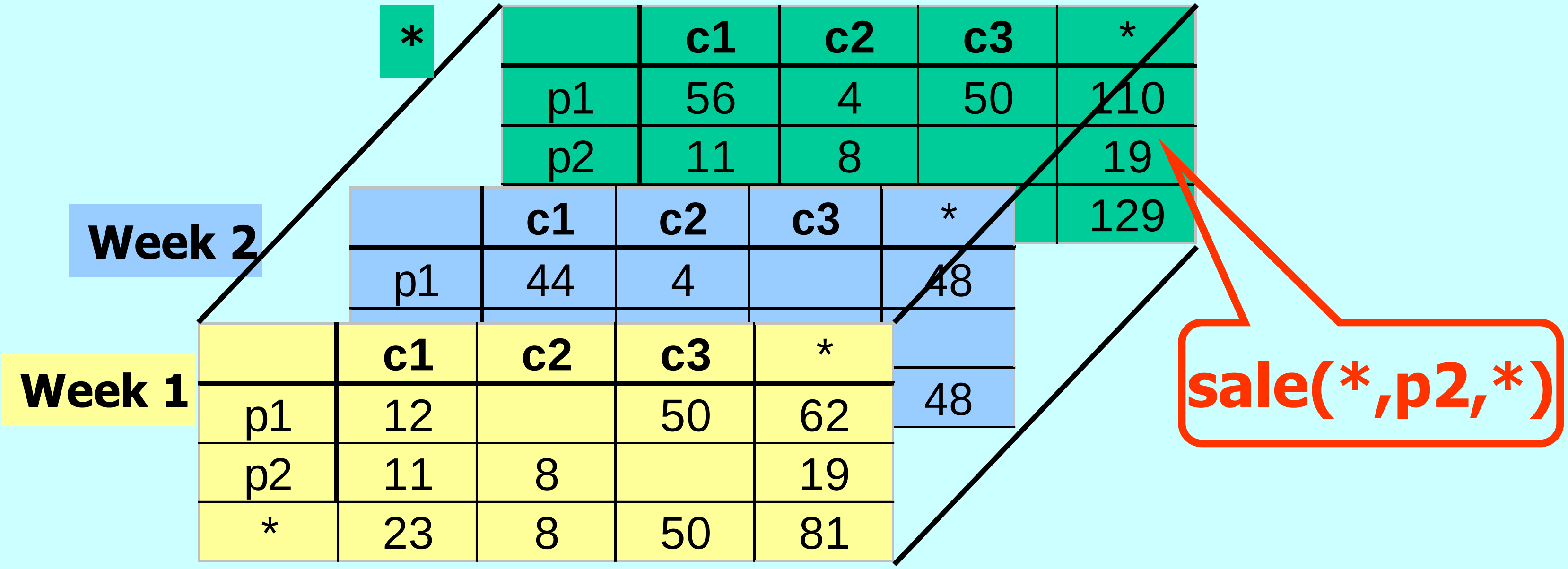
	sum
p1	110
p2	19

129

sale(*,*,*)

* = All

Extended Cube



*

	c1	c2	c3	*
p1	56	4	50	110
p2	11	8		19
				129

Week 2

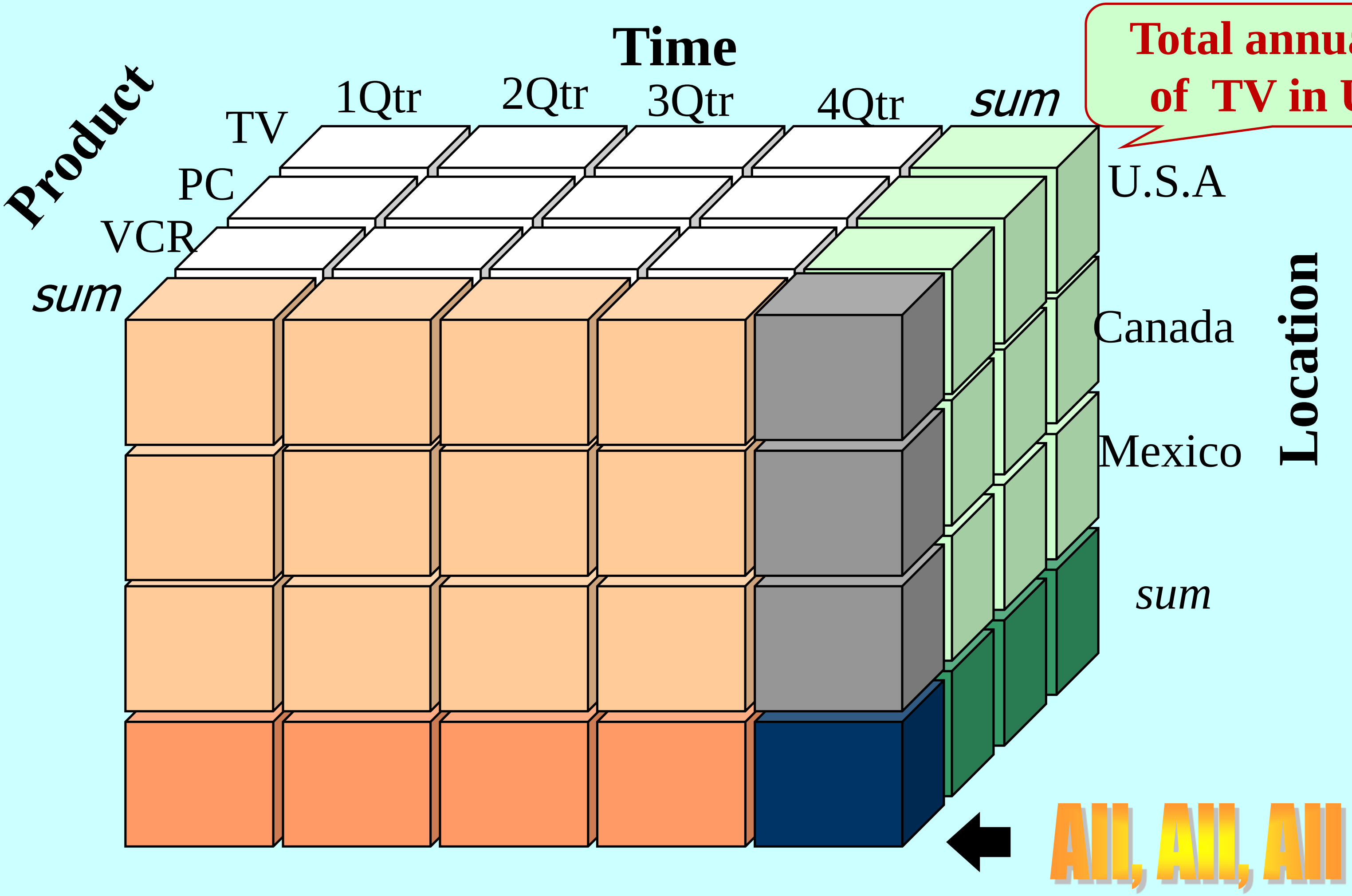
	c1	c2	c3	*
p1	44	4		48
				48

Week 1

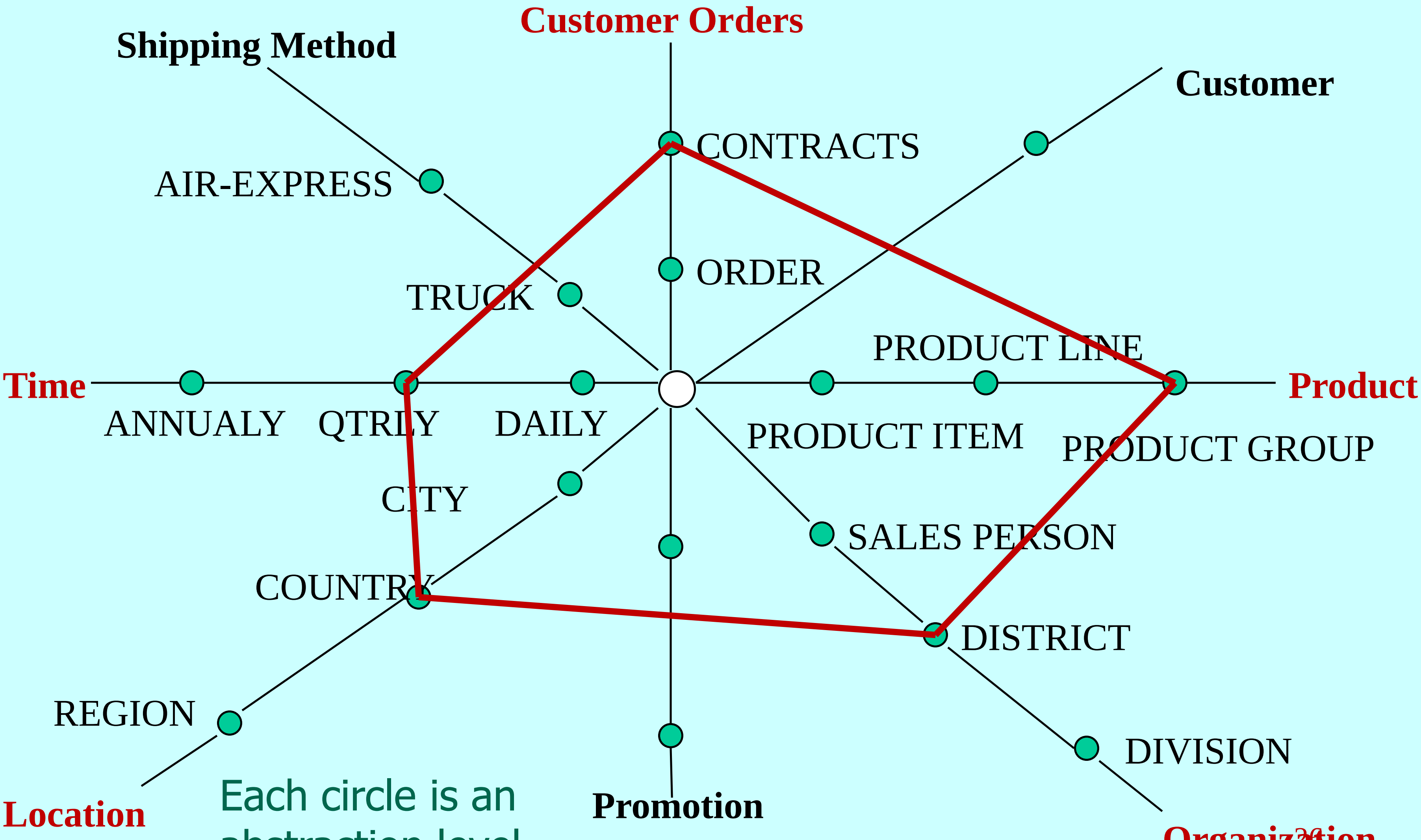
	c1	c2	c3	*
p1	12		50	62
p2	11	8		19
*	23	8	50	81

sale(*,p2,*)

Data cube: **sales** (time, product, location)



OLAP Query: A Star-Net Model



Each circle is an abstraction level.

Ass4 hint on OLAP query generation:

1. Each application query is an ad hoc business question (in English) about some specific analytical information on the subject.
2. The English query is then translated into OLAP operation(s), each operation is to apply one of the four OLAP operators, defined by choosing appropriate dimensions/concepts/values for producing a report.
3. Each report, i.e. the retrieved information, is represented in an analytical screen form (or in a graphical form).

A DW Project Example for FCS Student Grade Analysis (Doc/Theses/MACSprojOu03.pdf)

- How to quickly get a big picture of grades in terms of *courses, terms, years, subject areas, student groups,* etc.
- How to quickly analyze the information stored in multiple data sources for answering complex queries in improving academic process management.

Ad hoc analytical query examples

1. To answer why "Between 25-30 % of undergraduate students were dismissed from the Faculty of Computer Science each year over the past four years; most were first year students?" - *CS Faculty Retreat Report, 2002*

2. Do students fail on a particular course more than on other courses? Why?

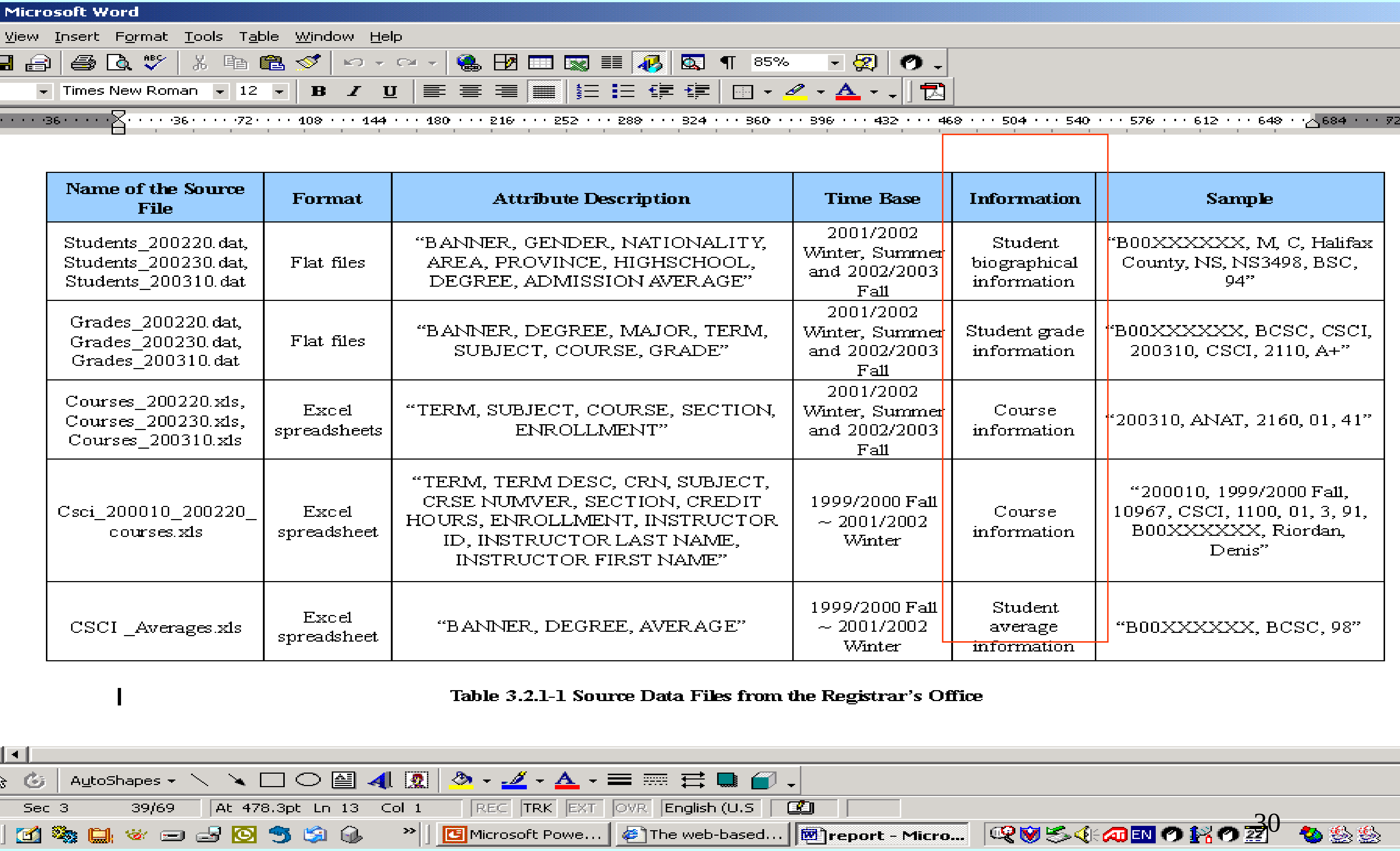
3. If students fail on one course, will they also tend to fail on other particular courses?

4. Do students have better performances on the courses of a specific area than on the courses of other areas?

5. Is the class enrollment a factor affecting the students' grades?

Challenges: 1) How to get all data comparisons on all courses for a particular year, e.g. 2002-2003, and see the trend of the grades for different year levels? 2) How to handle multiple data sets?

E.g., Multiple data sources: 11 CS student datasets.

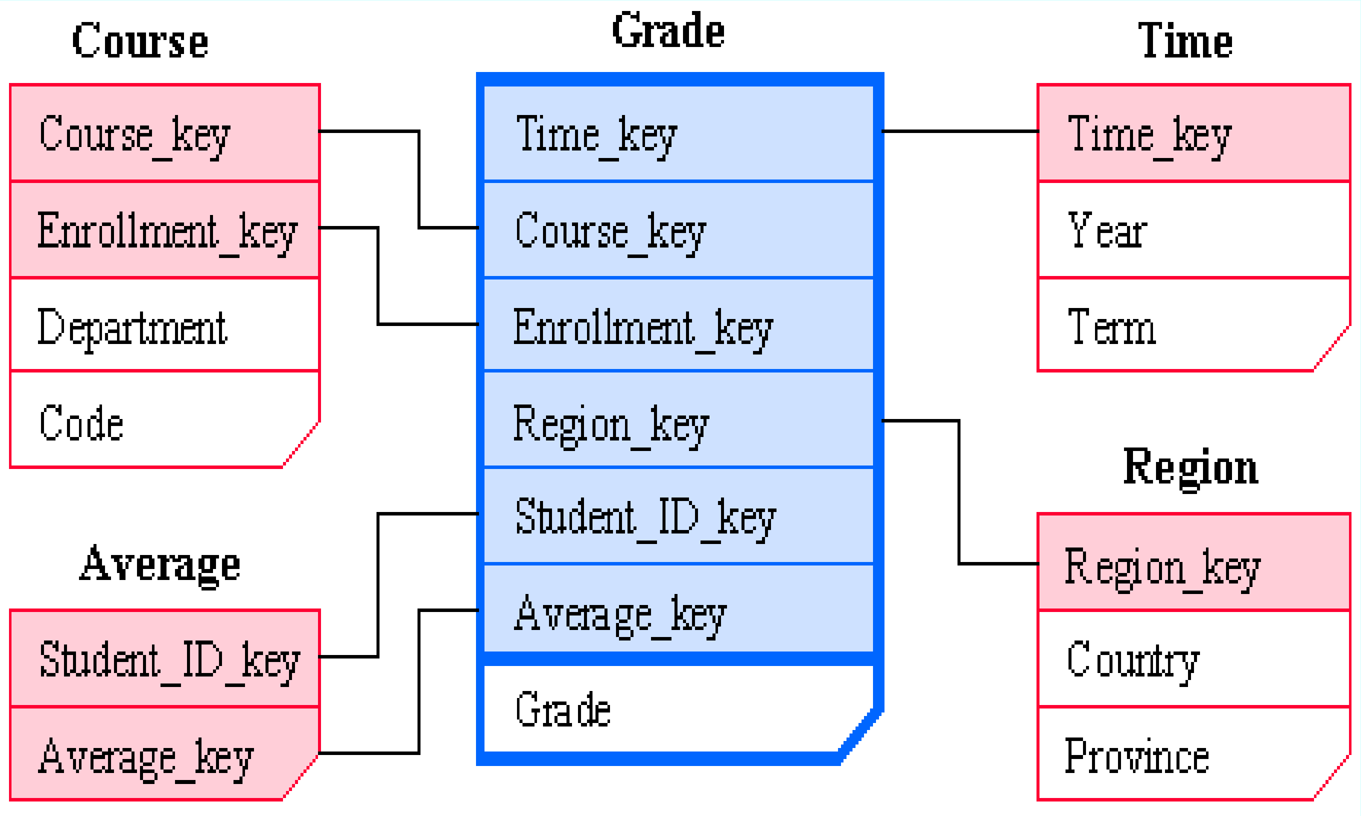

Microsoft Word
View Insert Format Tools Table Window Help
85%
Times New Roman
12
AutoShapes
Sec 3 39/69 At 478.3pt Ln 13 Col 1 REC TRK EXT OVR English (U.S
Microsoft Powe...
The web-based...
report - Micro...
EN

Name of the Source File	Format	Attribute Description	Time Base	Information	Sample
Students_200220.dat, Students_200230.dat, Students_200310.dat	Flat files	"BANNER, GENDER, NATIONALITY, AREA, PROVINCE, HIGHSCHOOL, DEGREE, ADMISSION AVERAGE"	2001/2002 Winter, Summer and 2002/2003 Fall	Student biographical information	"B00XXXXXX, M, C, Halifax County, NS, NS3498, BSC, 94"
Grades_200220.dat, Grades_200230.dat, Grades_200310.dat	Flat files	"BANNER, DEGREE, MAJOR, TERM, SUBJECT, COURSE, GRADE"	2001/2002 Winter, Summer and 2002/2003 Fall	Student grade information	"B00XXXXXX, BCSC, CSCI, 200310, CSCI, 2110, A+"
Courses_200220.xls, Courses_200230.xls, Courses_200310.xls	Excel spreadsheets	"TERM, SUBJECT, COURSE, SECTION, ENROLLMENT"	2001/2002 Winter, Summer and 2002/2003 Fall	Course information	"200310, ANAT, 2160, 01, 41"
Csci_200010_200220_ courses.xls	Excel spreadsheet	"TERM, TERM DESC, CRN, SUBJECT, CRSE NUMVER, SECTION, CREDIT HOURS, ENROLLMENT, INSTRUCTOR ID, INSTRUCTOR LAST NAME, INSTRUCTOR FIRST NAME"	1999/2000 Fall ~ 2001/2002 Winter	Course information	"200010, 1999/2000 Fall, 10967, CSCI, 1100, 01, 3, 91, B00XXXXXX, Riordan, Denis"
CSCI _Averages.xls	Excel spreadsheet	"BANNER, DEGREE, AVERAGE"	1999/2000 Fall ~ 2001/2002 Winter	Student average information	"B00XXXXXX, BCSC, 98"

Table 3.2.1-1 Source Data Files from the Registrar's Office

Solution: DW + OLAP + Visualization

- CS student data mart (star schema)



Course

Course_key
Enrollment_key
Department
Code

Average

Student_ID_key
Average_key

Grade

Time_key
Course_key
Enrollment_key
Region_key
Student_ID_key
Average_key
Grade

Time

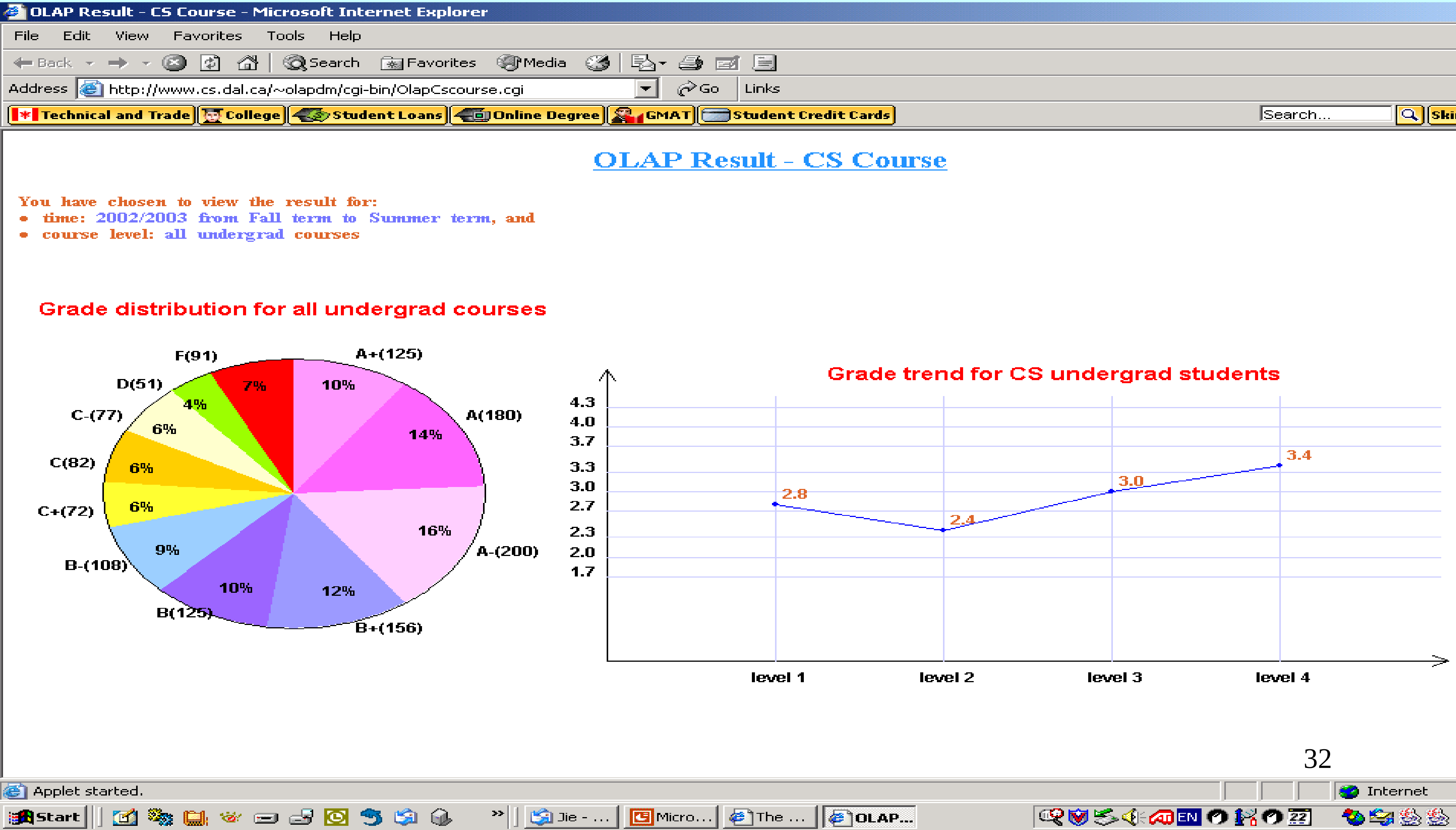
Time_key
Year
Term

Region

Region_key
Country
Province

Query 1: *"Get a picture of the grade trends in terms of different levels of courses, and the grade distribution for year 2002-2003".*

OLAP: year=2002-2003, **course=** All



http://www.cs.dal.ca/~olapdm/cgi-bin/OlapCscourse.cgi

OLAP Result - CS Course

You have chosen to view the result for:

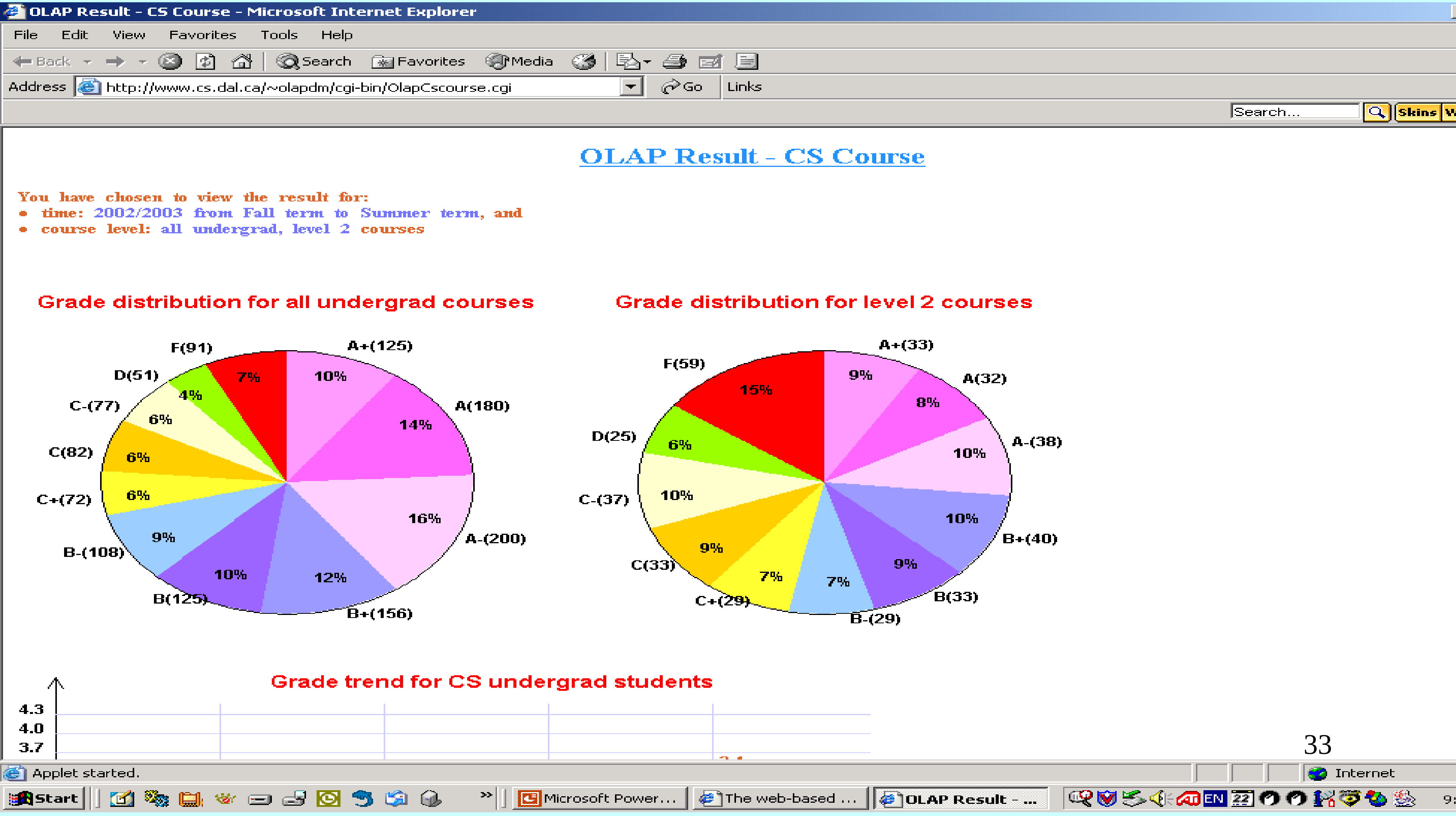
- time: 2002/2003 from Fall term to Summer term, and
- course level: all undergrad courses

Grade distribution for all undergrad courses

Grade trend for CS undergrad students

Query 2: *“Compare grade distributions between the second year courses and all other courses data for year 2002/2003”.*

OLAP: year = 2002-2003, **course =** All, **courses =** All level 2



OLAP Result - CS Course

You have chosen to view the result for:

- time: 2002/2003 from Fall term to Summer term, and
- course level: all undergrad, level 2 courses

Grade distribution for all undergrad courses

Grade distribution for level 2 courses

Grade trend for CS undergrad students

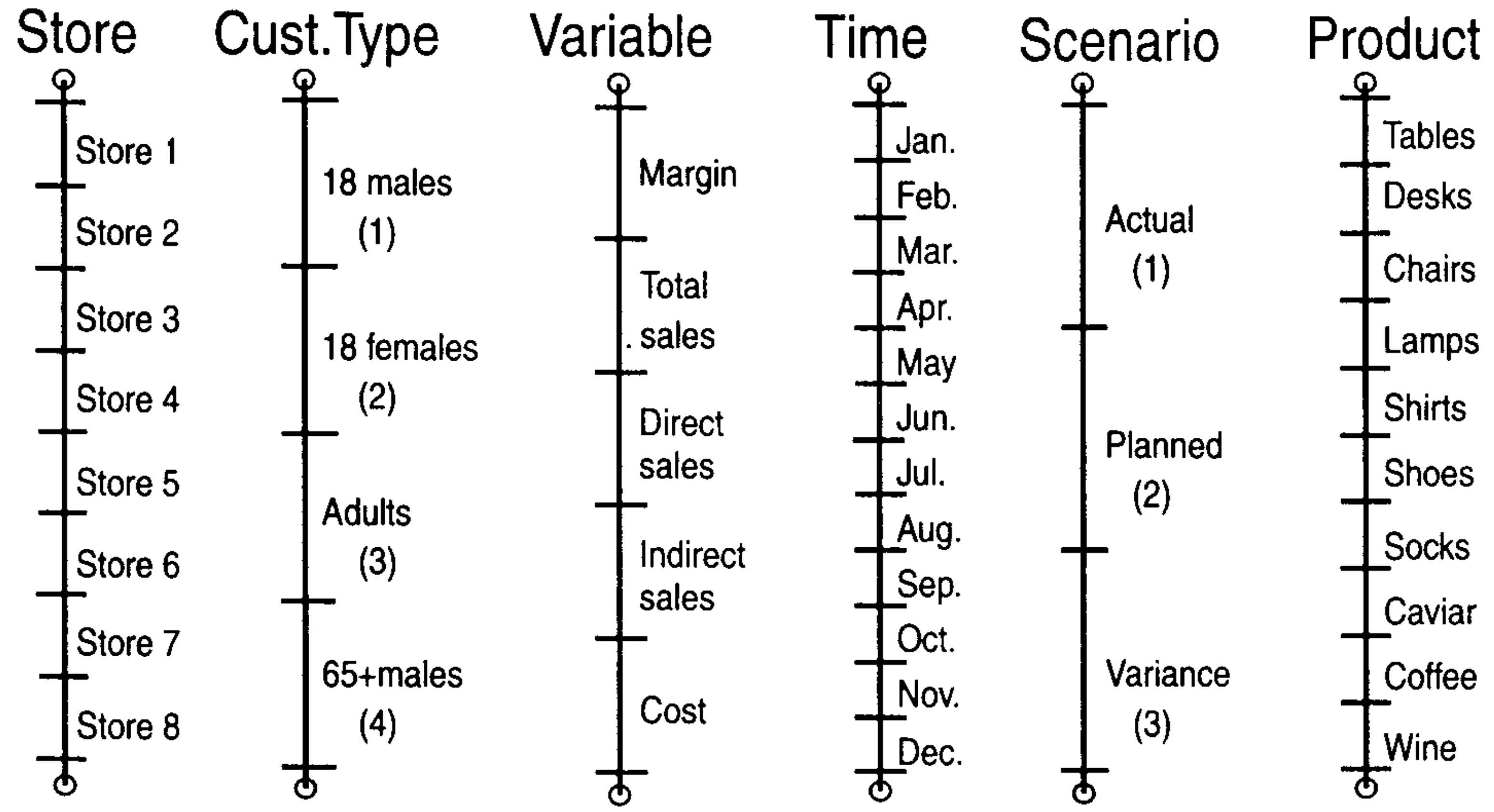
DW/OLAP Research Project Examples

1. Tayyaba Sharif, MACS project: **“Integrated data cube for Web content accessing pattern analysis”,** 2006
   - Doc/Theses/MACSprojTayyaba06.pdf
2. Nariman Amiri, HINF Master thesis: **“Designing a framework of intelligent information process on dentistry administration data”,** 2005
   - Doc/Theses/MHINthesisNariman05.pdf
3. Jie Ou, MACS project: **“Web-based OLAP and data mining for CS student database”,** 2003
   - Doc/Theses/MACSprojOu03.pdf

How to Present Multi-dimensional Cuboid (OLAP report) on a 2-D screen?

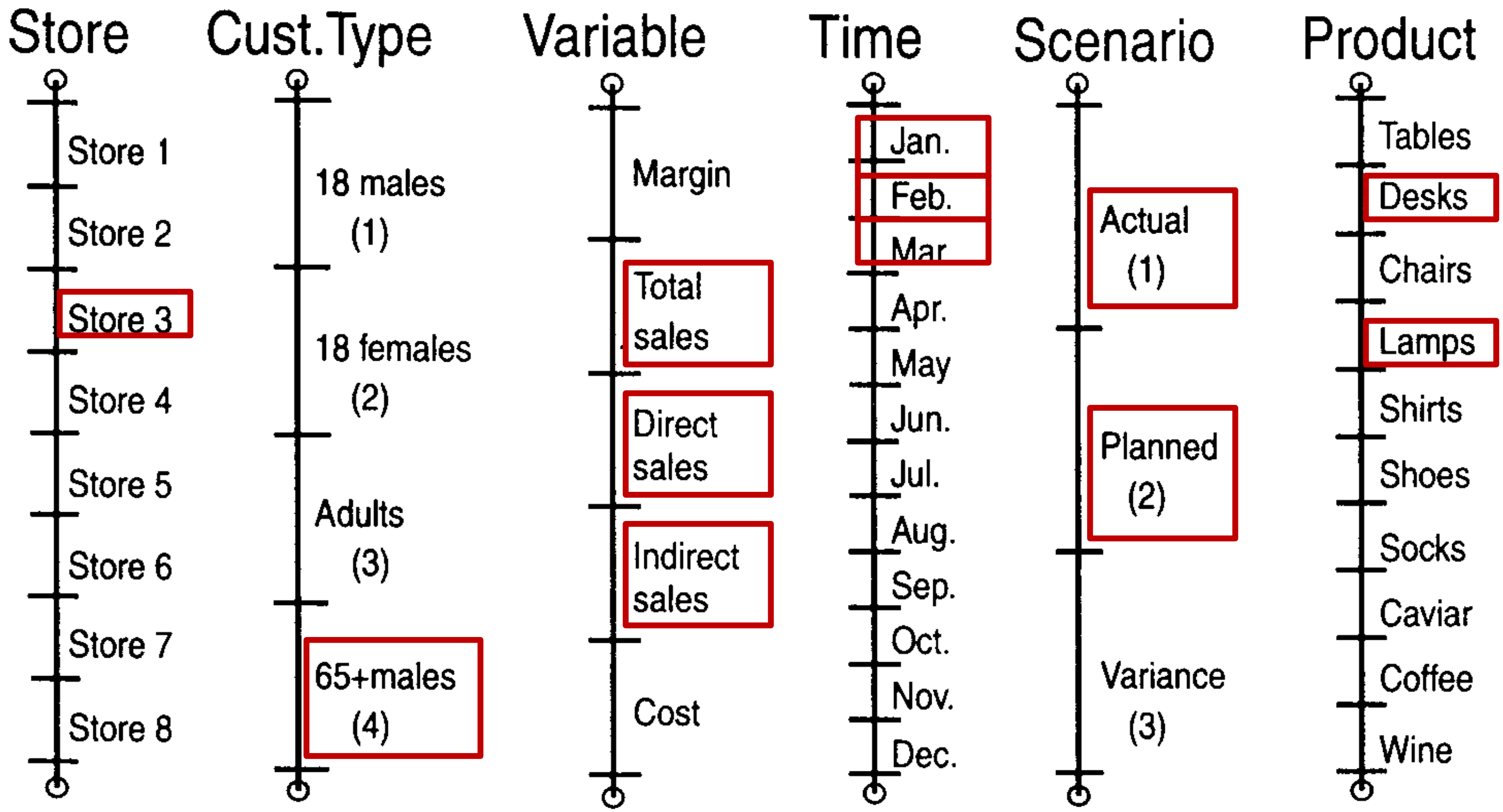
- How to map multiple logical dimensions onto a single computer screen?
  - Issue: Use 2-D screen to see the M-D space
- Different metaphors:
  - Physical dimension metaphor, e.g. 3-D graphics: **virtual camera**
  - Logic dimension metaphor, e.g. table-based OLAP report: **?**

Query: What are the sales of desk and lamp in Jan, Feb and Mar for Store 3 purchased by male senior citizens only?



A six-dimensional MDS.

Query: What are the sales of desk and lamp in Jan, Feb and Mar for Store 3 purchased by male senior citizens only?



A six-dimensional MDS.

Summary of Logical Dimensions

- As distinguished from physical dimensions, which are based on angles and limited to three, logical dimensions have no such limits.

- Two types of dimensions of a data cube

  - **Identifier dimensions:** they are logical factors or identifying attributes of measurable events or things that we track.

  - **Variable dimension:** they identify what we track in a situation.

- MDS software enables multi-dimensions of information to be combined onto each **row**, **column**, and **page** axis of a screen, thus making it possible to visualize and understand a multi-dimensional data set in terms of information presented on flat screen.

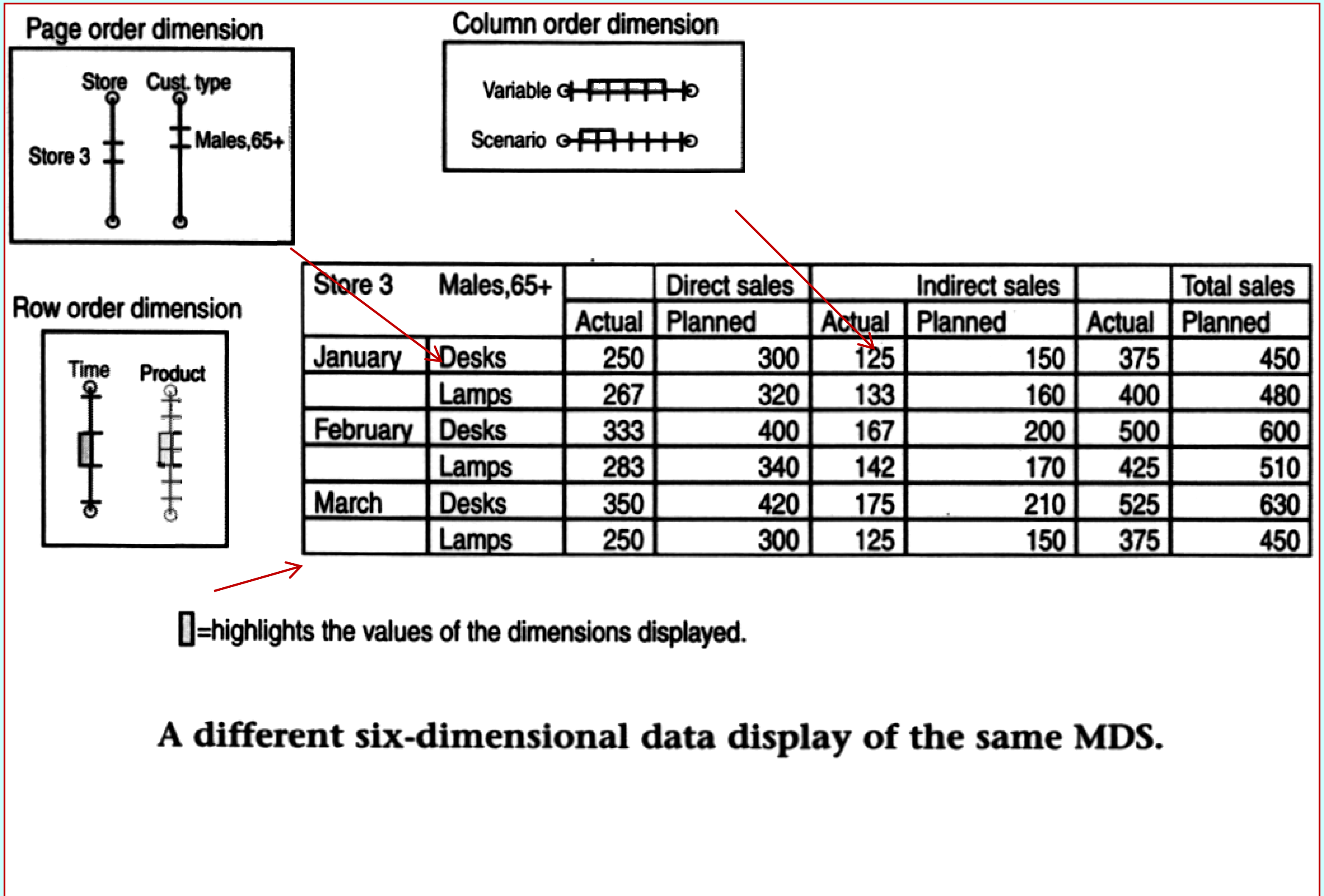
- The ability of MDS software to model multidimensional information and to handle the user representation of the information makes it better suited for working with complex datasets than either SQL databases or traditional spreadsheets.

Present M-D Cuboid Data in 2D Screen

- Logic dimension metaphor: (for table-based report):
  - Analytical Screen
- Solution:
  - To combine multiple logical dimensions within the same display dimensions: Row, Column, and Page
    - Each dimension of the vertical bar can be connected to either a row, column, or page axes

Query: What are the sales of desk and lamp in Jan, Feb and Mar for Store 3 purchased by male senior citizens only?

A six dimensional display:



Store 3	Males,65+		Direct sales		Indirect sales		Total sales
		Actual	Planned	Actual	Planned	Actual	Planned
January	Desks	250	300	125	150	375	450
	Lamps	267	320	133	160	400	480
February	Desks	333	400	167	200	500	600
	Lamps	283	340	142	170	425	510
March	Desks	350	420	175	210	525	630
	Lamps	250	300	125	150	375	450

▯=highlights the values of the dimensions displayed.

A different six-dimensional data display of the same MDS.

Use analytical screen in different ways

- There are different ways that the same model dimensions can be mapped onto **row**, **column**, and **page** axes
  - The ability to easily view the same data by reconfiguring how dimensions are displayed is one of the great benefits of MDS
    - The reason is due to the separation of data structure, as represented in the vertical logical dimensions, from data display, as represented in the multi-dimensional grid.

The issue of using analytical screen

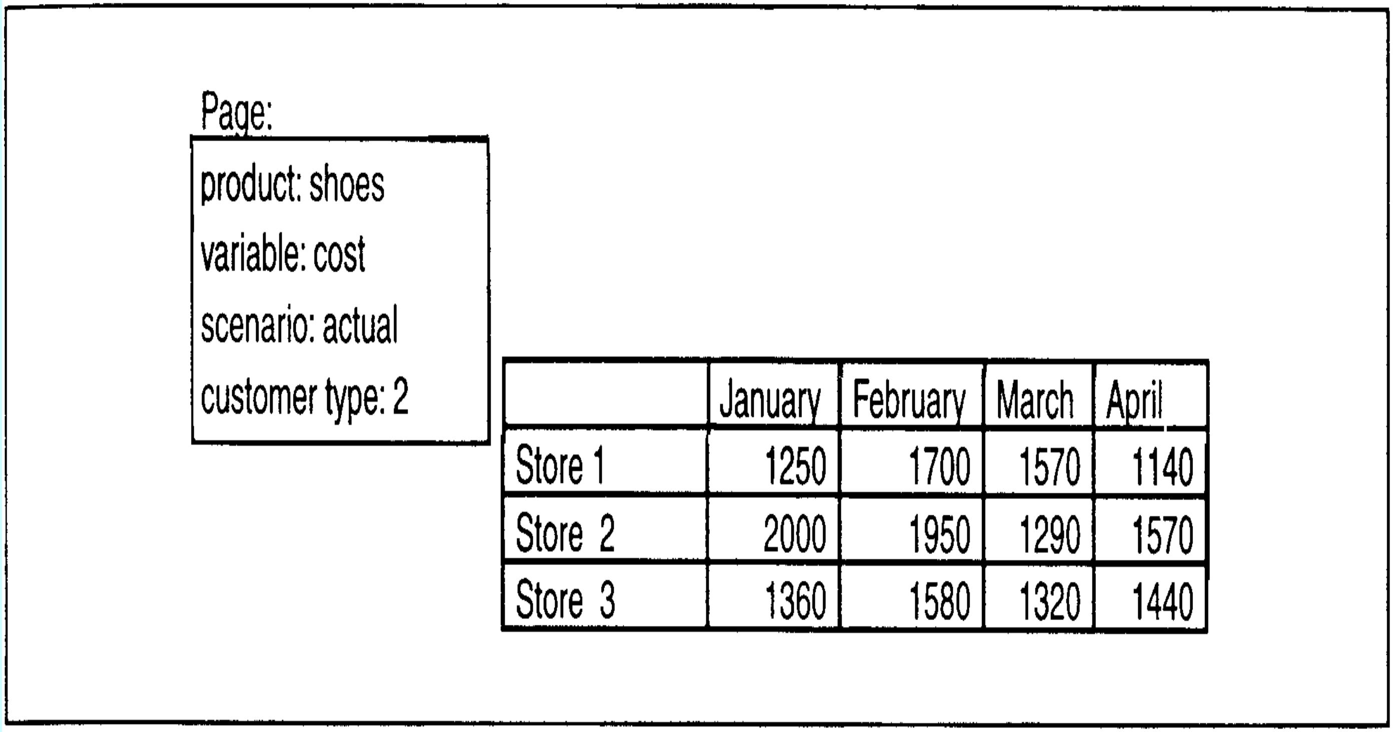
- The more screen space is consumed for displaying dimension members, the less space is left for displaying data.
- The less space left for displaying data, the more scrolling you need to do between screens to see the same data.
- The more scrolling you need to perform, the harder it is to understand what you are looking for.

Make optimal use of analytical screen

- To maximize the degree to which everything on the screen is relevant, try keeping dimensions along pages unless you know you need to see more than one member at a time.
- Ask yourself "What do I want to look at?", or "What am I trying to compare?" before deciding how to display information on the screen.

Query: Look at and compare actual costs across **stores** (1, 2 and 3) and **time** (first four months), for some product (e.g. shoes) and customer type (e.g. type 2).

A six dimensional display:



Page:

product: shoes
variable: cost
scenario: actual
customer type: 2

	January	February	March	April
Store 1	1250	1700	1570	1140
Store 2	2000	1950	1290	1570
Store 3	1360	1580	1320	1440

Arranging data to compare costs across stores and time.

Summarization: Aggregate Measures

- **What do I want to look at? What am I trying to compare?**
  - **Define a grouping** (i.e. determine a cuboid of the data cube)
  - **Choose measures** about the cuboid (for pre-calculated values, or invoke online aggregate functions to the grouping)
  - **OLAP query:** *cuboid-value* pairs (or *Dimension-value* pairs).

  E.g., “What is the total sales of computers in Halifax for the first quarter?”

  cuboid: *<time ="Q1", location =“Halfax", item ="Computer">*

  value: *sales = sum(the data set of the cuboid)*

- **Aggregate functions** are statistics models of data summarization, such as *sum, count, average, maximum, minimum, variance, standard deviation, median, mode, rank,* etc.

Categories of Aggregate Dunctions

- **Distributive:** If the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning.
  - E.g., **count( ), sum( ), min(), max( ),** etc.
- **Algebraic:** If it can be computed by an algebraic function with the arguments which are obtained by applying distributive aggregate functions
  - E.g., **avg( )** = sum( ) / count( ), **variance( ), standard_deviation( ),** etc.
- **Holistic:** If it needs repeated search and comparison on the selected data set
  - E.g., **rank( ), median( ), mode( ),** etc.

Distributive and algebraic functions are most frequently used and suitable for on-line aggregation compuation.

Aggregates, e.g.

- “Add up amounts for day 1”
- In SQL: SELECT sum(amt) FROM SALE
  WHERE date = 1

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

→ 81

Aggregates, e.g.

- "Add up amounts by day"
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

⇨

ans	date	sum
	1	81
	2	48

OLAP Server Architectures

- **Multi-dimensional OLAP (MOLAP)**
  - Implemented as a large multidimensional array
  - Fast indexing to pre-computed summarized data (with built-in indexing)
  - A fully materialized MOLAP array can contain an enormous number of empty cells, could result in un-acceptable storage requirements
- **Relational OLAP (ROLAP)**
  - Implemented as a collection of relational tables
  - Can be processed and queried with traditional RDBMS technology (i.e. indexing, grouping and join etc.)
  - Greater scalability
  - No “built-in” indexing
- **Hybrid OLAP (HOLAP)**
  - User flexibility, e.g., low level: relational, high-level: array
  - MS SQL Server

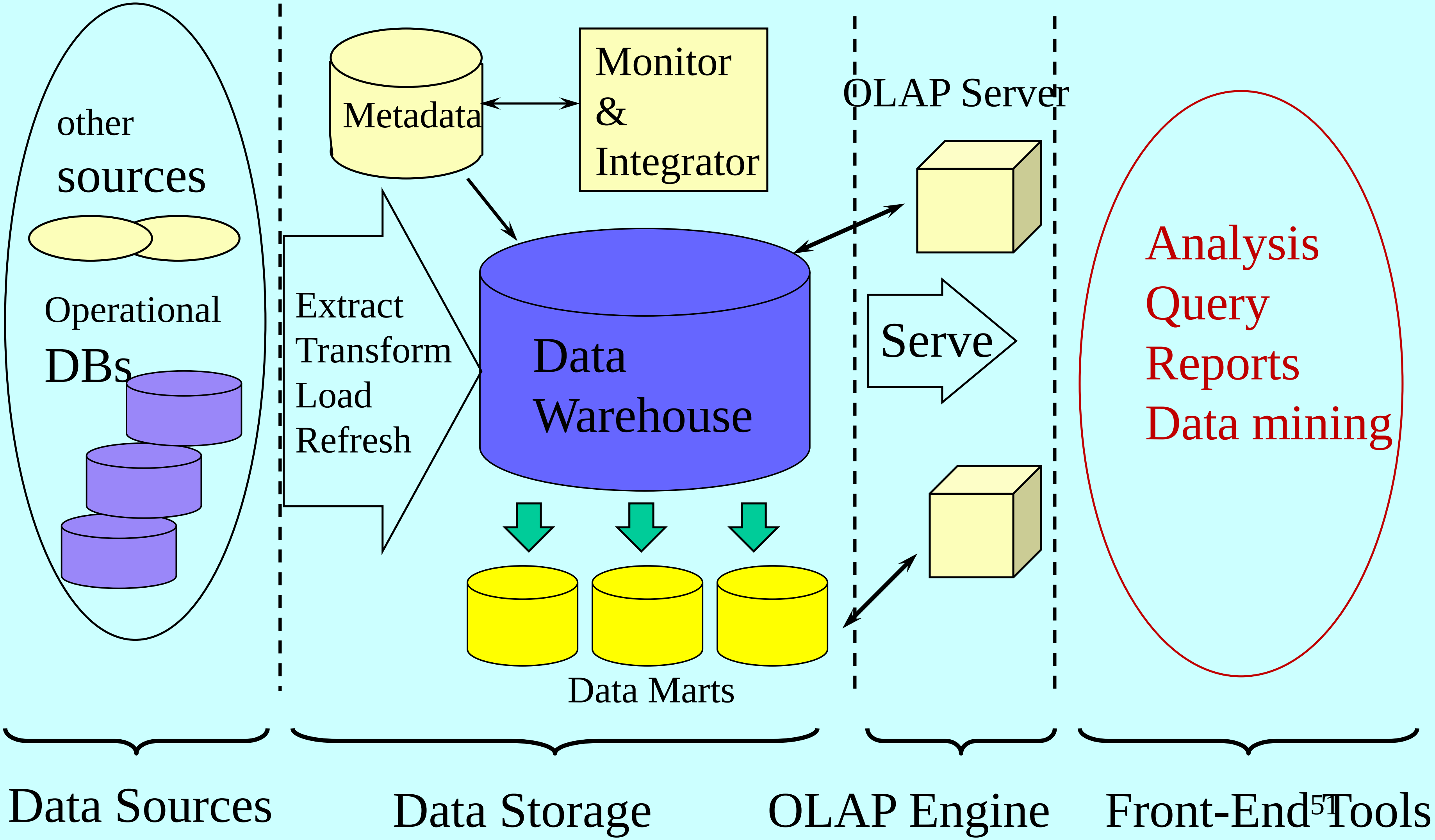
Efficient Data Cube Computation

- **Review the concept of data cube lattice:**
  - Data cube can be viewed as a lattice of cuboids
    - The bottom-most cuboid is the base cuboid
    - The top-most cuboid (apex) contains only one cell
  - How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

- **Materialization options of data cube**
  - Materialize every (cuboid) (full materialization), none (no materialization), or some (**partial materialization**)
  - Selection of which cuboids to materialize
    - Based on size, sharing, access frequency, etc.

Multi-Tiered Architecture



Data Warehouse Usage

Three kinds of data warehouse applications

- **Directed information processing:**
  - Deals directly with the stored aggregation information
  - Supports simple ad hoc queries, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
- **Analytical processing:**
  - supports more sophistic ad hoc queries based on a set of OLAP operations, such slice-dice, drilling, pivoting, etc.
- **Data mining:**
  - Supports user interactive exploration for finding hidden patterns
    - ➢ Visual data mining based on OLAP operations (aggregations), e.g. the trend analysis of "the wealth and health of the 200 countries over 200 years".
  - Finds hidden patterns from a defined data set in DW
  - Supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

E.g., Visual data mining based on OLAP operations (aggregations).

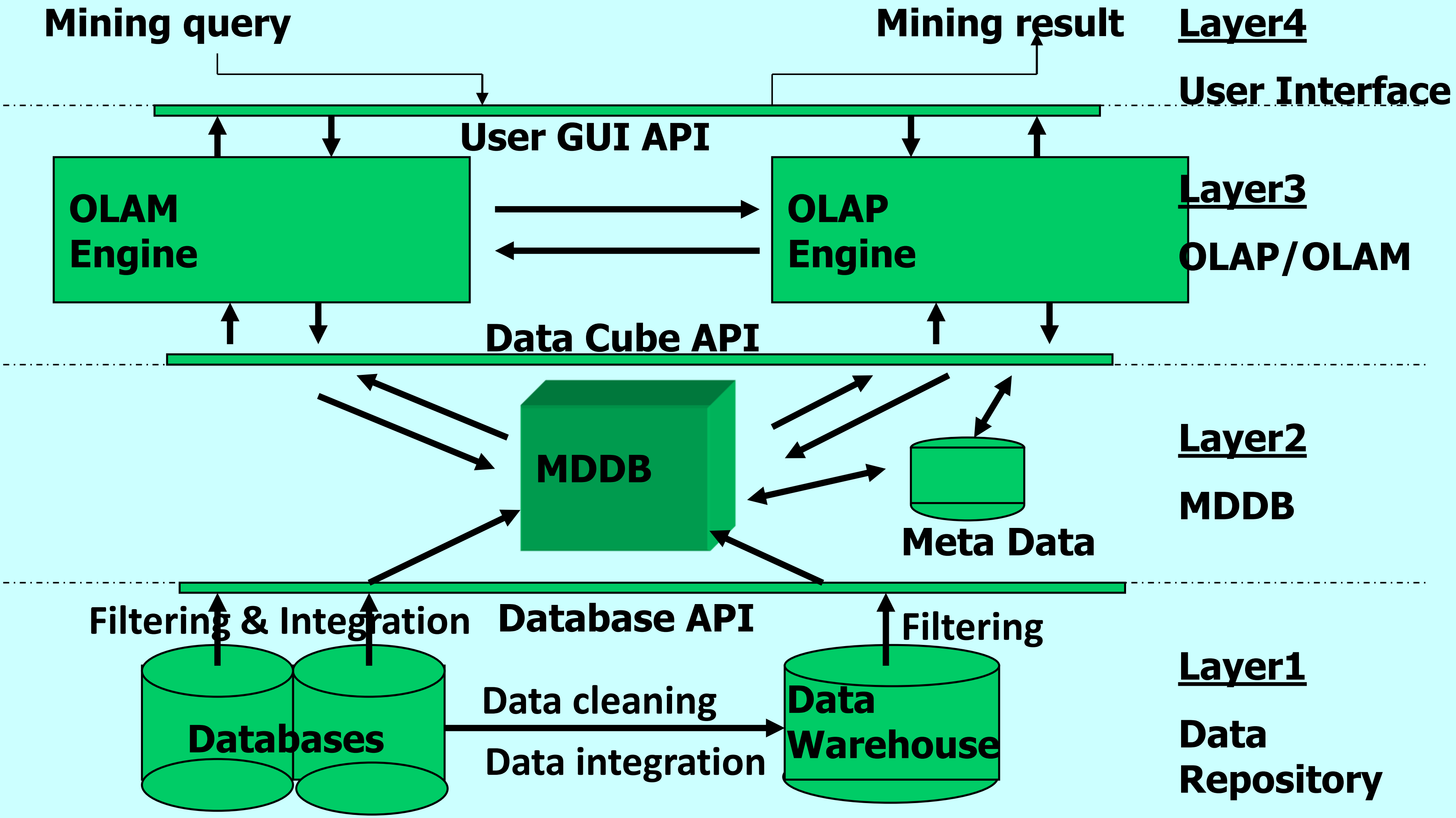
- Discover the trend of Wealth and Health of 200 Countries, over 200 Years: (http://www.youtube.com/watch?v=jbkSRLYSojo).

In this application, by visualizing 120000 aggregation values represented in the space of income, life expectancy, and year, one can easily and clearly perceive what is the trend (the animation helps to view slices of the cube in motion along time dimension).

From On-Line Analytical Processing to On Line Analytical Mining (OLAM)

- **Why online analytical mining?**
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC (Open Data Base Connectivity), Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - mining with drilling, dicing, pivoting, etc
  - On-line selection of data mining functions
    - integration and swapping of multiple mining functions, algorithms, and tasks
- **Architecture of OLAM**

An OLAM Architecture

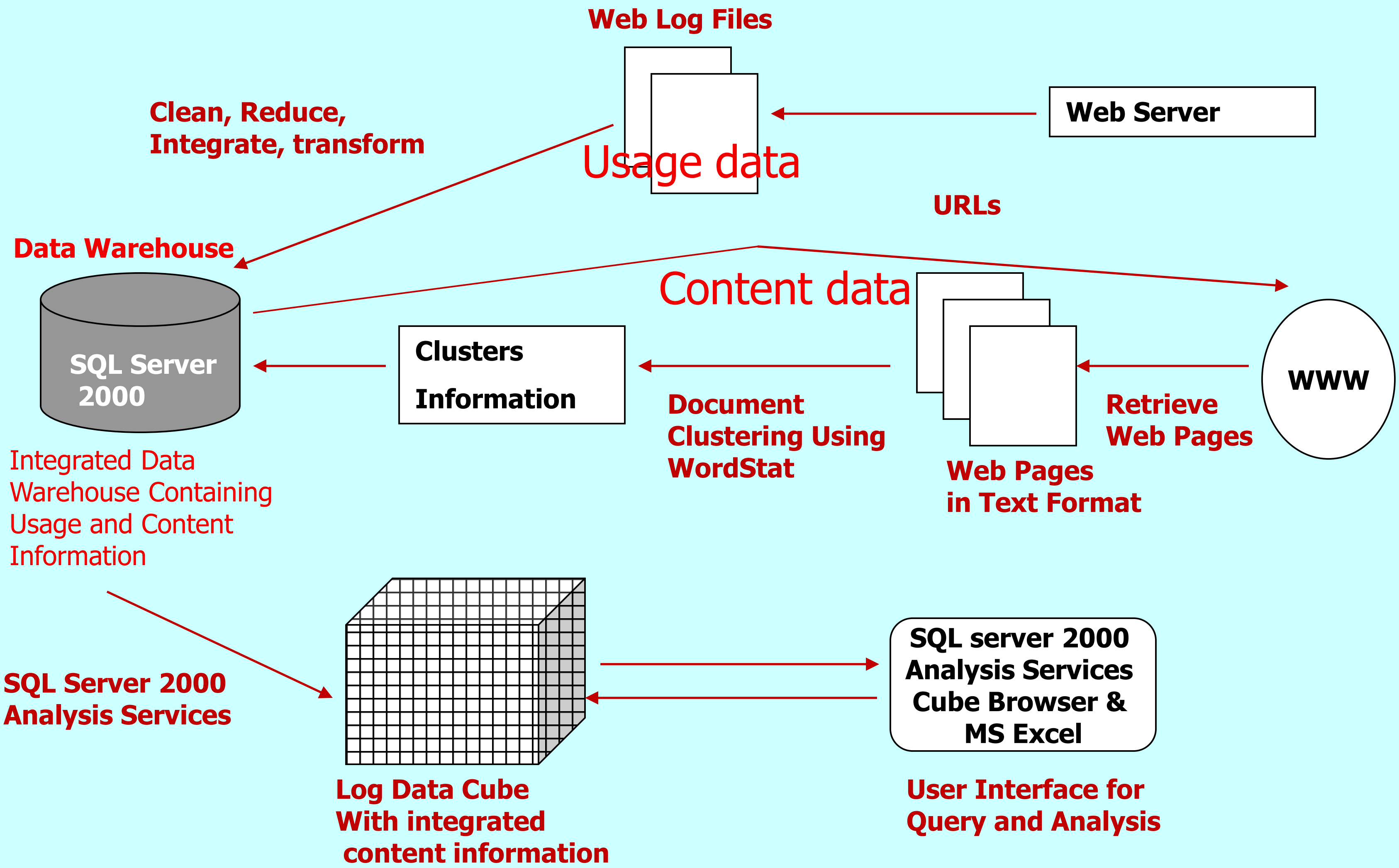


A case study: “Integrated data cube for Web content accessing pattern analysis”

(Doc/Theses/MACSprojTayyaba06.pdf)

- **Motivation**
  - Get bigger and clear pictures on web usages for supporting web site design and user group behavior analysis.
  - The results obtained by web log analyzers are limited in performance and lack depth of analysis, such as lack of content info.
- **Objective**
  - Integrate website usage and content data together and build data warehouse for conducting in-depth OLAP based analysis by generating multi-dimensional views of the comprehensive data cube.

System Architecture



Web Log Files

Web Server

Usage data

Clean, Reduce,
Integrate, transform

URLs

Data Warehouse

SQL Server
2000

Content data

Clusters

Information

Document
Clustering Using
WordStat

Web Pages
in Text Format

Retrieve
Web Pages

WWW

Integrated Data
Warehouse Containing
Usage and Content
Information

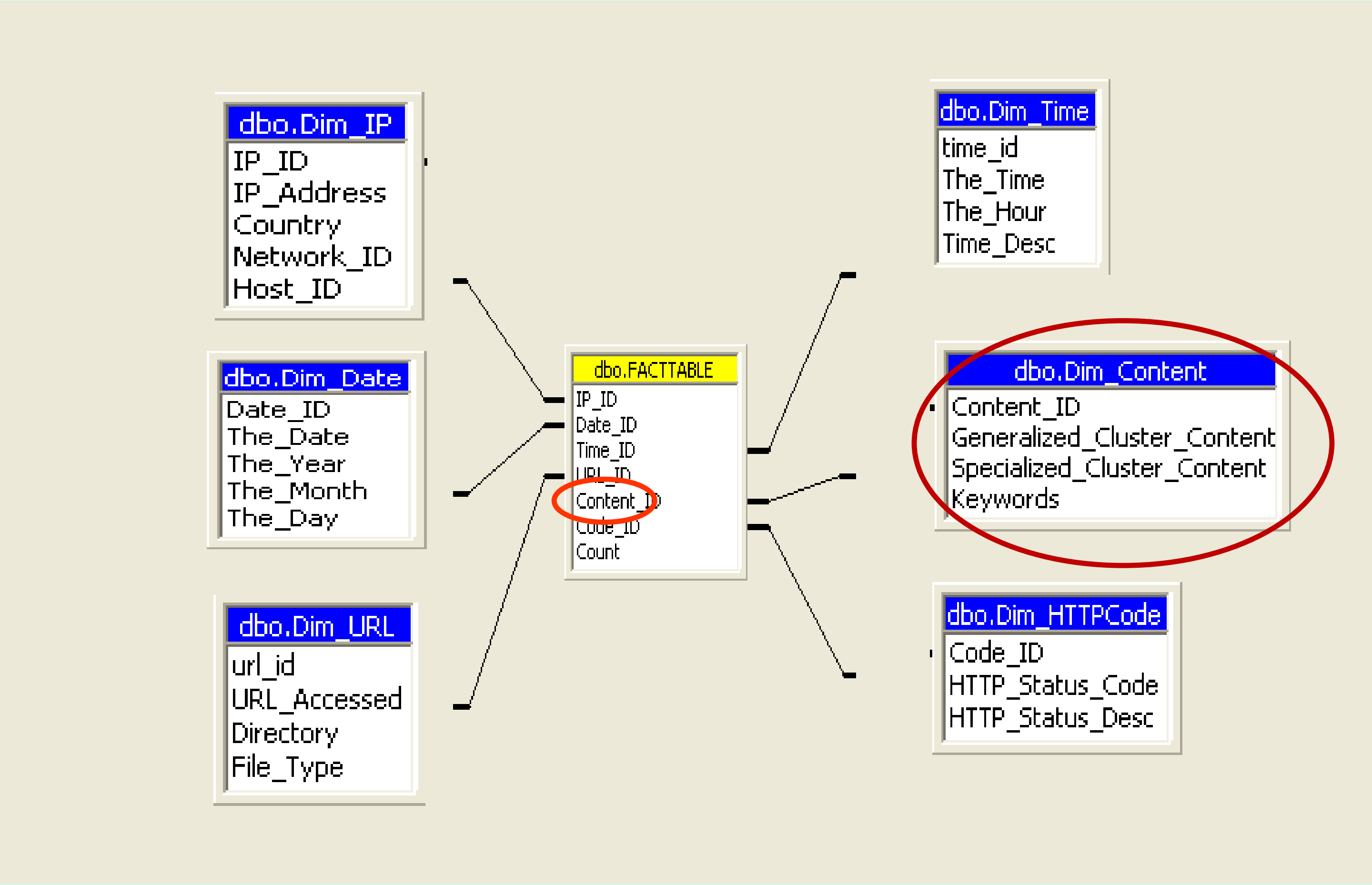
SQL Server 2000
Analysis Services

Log Data Cube
With integrated
content information

SQL server 2000
Analysis Services
Cube Browser &
MS Excel

User Interface for
Query and Analysis

Data Cube Design



dbo.Dim_IP
- IP_ID
- IP_Address
- Country
- Network_ID
- Host_ID

dbo.Dim_Date
- Date_ID
- The_Date
- The_Year
- The_Month
- The_Day

dbo.Dim_URL
- url_id
- URL_Accessed
- Directory
- File_Type

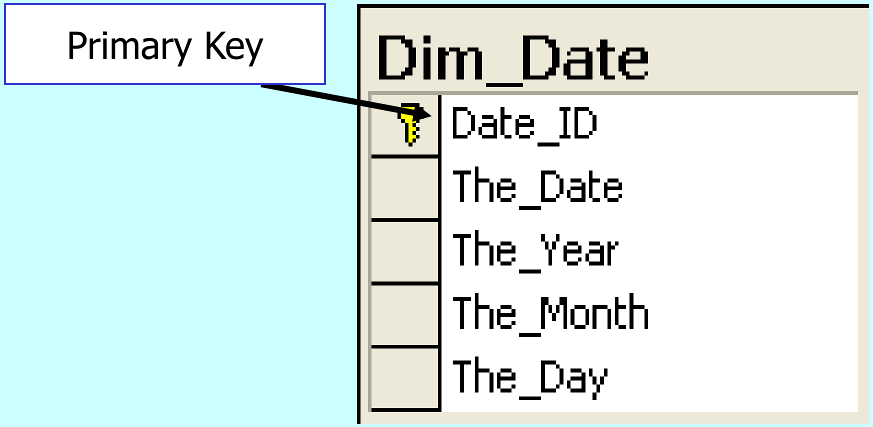
dbo.FACTTABLE
- IP_ID
- Date_ID
- Time_ID
- URL_ID
- Content_ID
- Code_ID
- Count

dbo.Dim_Time
- time_id
- The_Time
- The_Hour
- Time_Desc

dbo.Dim_Content
- Content_ID
- Generalized_Cluster_Content
- Specialized_Cluster_Content
- Keywords

dbo.Dim_HTTPCode
- Code_ID
- HTTP_Status_Code
- HTTP_Status_Desc

“Date” Dimension

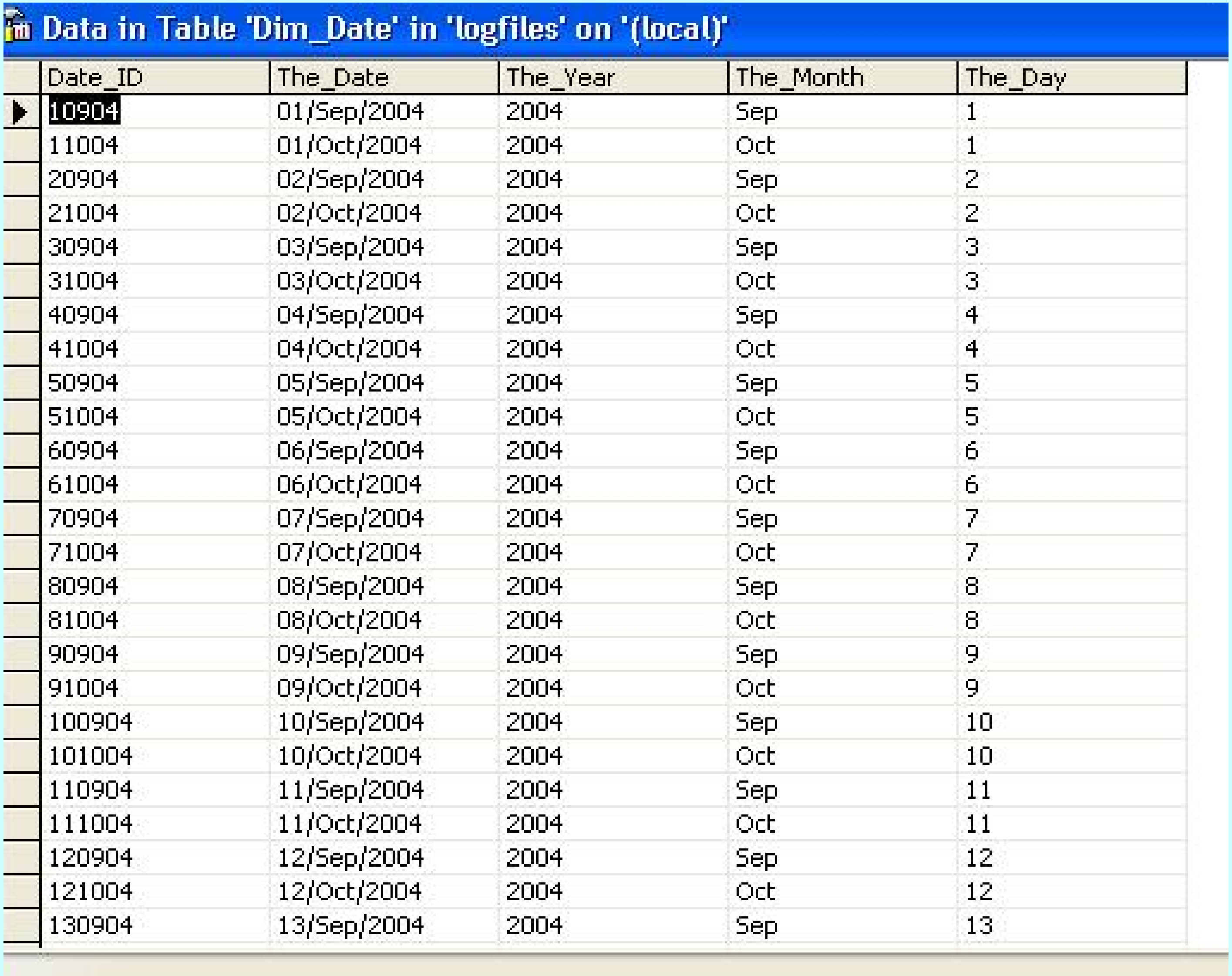


Primary Key

Dim_Date

- Date_ID
- The_Date
- The_Year
- The_Month
- The_Day

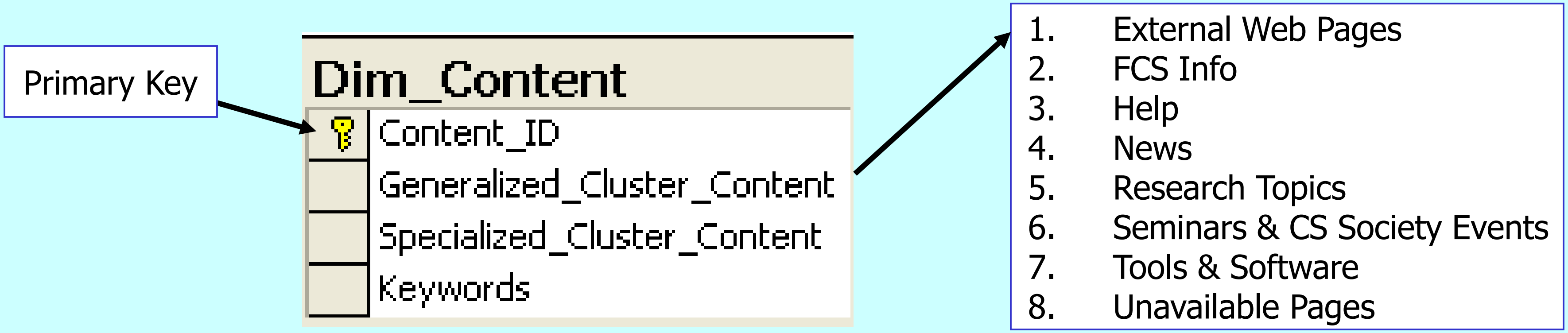
Snapshot of Dim_Date



Data in Table 'Dim_Date' in 'logfiles' on '(local)'

Date_ID	The_Date	The_Year	The_Month	The_Day
10904	01/Sep/2004	2004	Sep	1
11004	01/Oct/2004	2004	Oct	1
20904	02/Sep/2004	2004	Sep	2
21004	02/Oct/2004	2004	Oct	2
30904	03/Sep/2004	2004	Sep	3
31004	03/Oct/2004	2004	Oct	3
40904	04/Sep/2004	2004	Sep	4
41004	04/Oct/2004	2004	Oct	4
50904	05/Sep/2004	2004	Sep	5
51004	05/Oct/2004	2004	Oct	5
60904	06/Sep/2004	2004	Sep	6
61004	06/Oct/2004	2004	Oct	6
70904	07/Sep/2004	2004	Sep	7
71004	07/Oct/2004	2004	Oct	7
80904	08/Sep/2004	2004	Sep	8
81004	08/Oct/2004	2004	Oct	8
90904	09/Sep/2004	2004	Sep	9
91004	09/Oct/2004	2004	Oct	9
100904	10/Sep/2004	2004	Sep	10
101004	10/Oct/2004	2004	Oct	10
110904	11/Sep/2004	2004	Sep	11
111004	11/Oct/2004	2004	Oct	11
120904	12/Sep/2004	2004	Sep	12
121004	12/Oct/2004	2004	Oct	12
130904	13/Sep/2004	2004	Sep	13

"Content" Dimension



Primary Key

Dim_Content

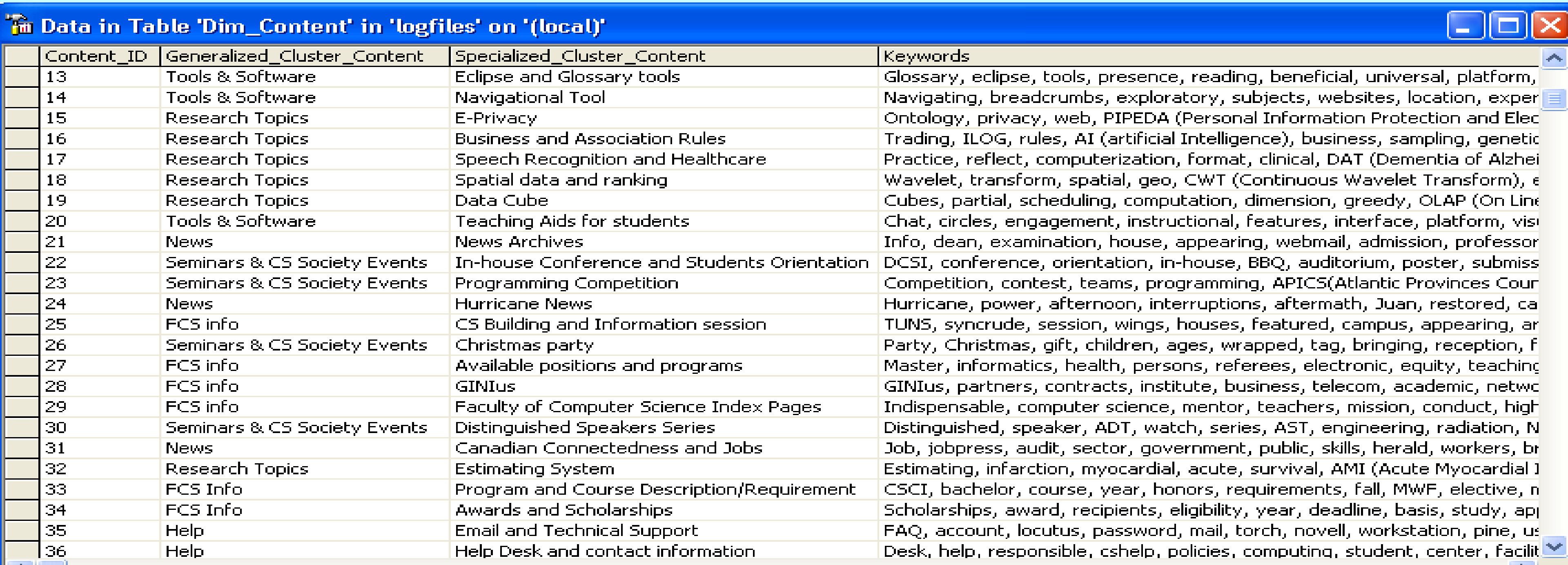
Content_ID

Generalized_Cluster_Content

Specialized_Cluster_Content

Keywords

1. External Web Pages
2. FCS Info
3. Help
4. News
5. Research Topics
6. Seminars & CS Society Events
7. Tools & Software
8. Unavailable Pages

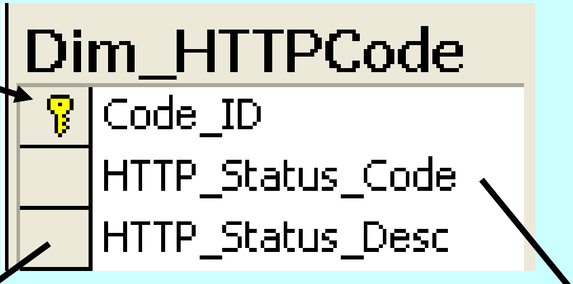


Data in Table 'Dim_Content' in 'logfiles' on '(local)'

Content_ID	Generalized_Cluster_Content	Specialized_Cluster_Content	Keywords
13	Tools & Software	Eclipse and Glossary tools	Glossary, eclipse, tools, presence, reading, beneficial, universal, platform,
14	Tools & Software	Navigational Tool	Navigating, breadcrumbs, exploratory, subjects, websites, location, exper
15	Research Topics	E-Privacy	Ontology, privacy, web, PIPEDA (Personal Information Protection and Elec
16	Research Topics	Business and Association Rules	Trading, ILOG, rules, AI (artificial Intelligence), business, sampling, genetic
17	Research Topics	Speech Recognition and Healthcare	Practice, reflect, computerization, format, clinical, DAT (Dementia of Alzhei
18	Research Topics	Spatial data and ranking	Wavelet, transform, spatial, geo, CWT (Continuous Wavelet Transform), e
19	Research Topics	Data Cube	Cubes, partial, scheduling, computation, dimension, greedy, OLAP (On Line
20	Tools & Software	Teaching Aids for students	Chat, circles, engagement, instructional, features, interface, platform, vis
21	News	News Archives	Info, dean, examination, house, appearing, webmail, admission, professor
22	Seminars & CS Society Events	In-house Conference and Students Orientation	DCSI, conference, orientation, in-house, BBQ, auditorium, poster, submiss
23	Seminars & CS Society Events	Programming Competition	Competition, contest, teams, programming, APICS(Atlantic Provinces Cour
24	News	Hurricane News	Hurricane, power, afternoon, interruptions, aftermath, Juan, restored, ca
25	FCS info	CS Building and Information session	TUNS, syncrude, session, wings, houses, featured, campus, appearing, ar
26	Seminars & CS Society Events	Christmas party	Party, Christmas, gift, children, ages, wrapped, tag, bringing, reception, f
27	FCS info	Available positions and programs	Master, informatics, health, persons, referees, electronic, equity, teaching
28	FCS info	GINIus	GINIus, partners, contracts, institute, business, telecom, academic, netwo
29	FCS info	Faculty of Computer Science Index Pages	Indispensable, computer science, mentor, teachers, mission, conduct, high
30	Seminars & CS Society Events	Distinguished Speakers Series	Distinguished, speaker, ADT, watch, series, AST, engineering, radiation, N
31	News	Canadian Connectedness and Jobs	Job, jobpress, audit, sector, government, public, skills, herald, workers, br
32	Research Topics	Estimating System	Estimating, infarction, myocardial, acute, survival, AMI (Acute Myocardial I
33	FCS Info	Program and Course Description/Requirement	CSCI, bachelor, course, year, honors, requirements, fall, MWF, elective, n
34	FCS Info	Awards and Scholarships	Scholarships, award, recipients, eligibility, year, deadline, basis, study, ap
35	Help	Email and Technical Support	FAQ, account, locutus, password, mail, torch, novell, workstation, pine, us
36	Help	Help Desk and contact information	Desk, help, responsible, cshelp, policies, computing, student, center, facilit

“HTTP Code” Dimension

Primary Key



Dim_HTTPCode

Code_ID

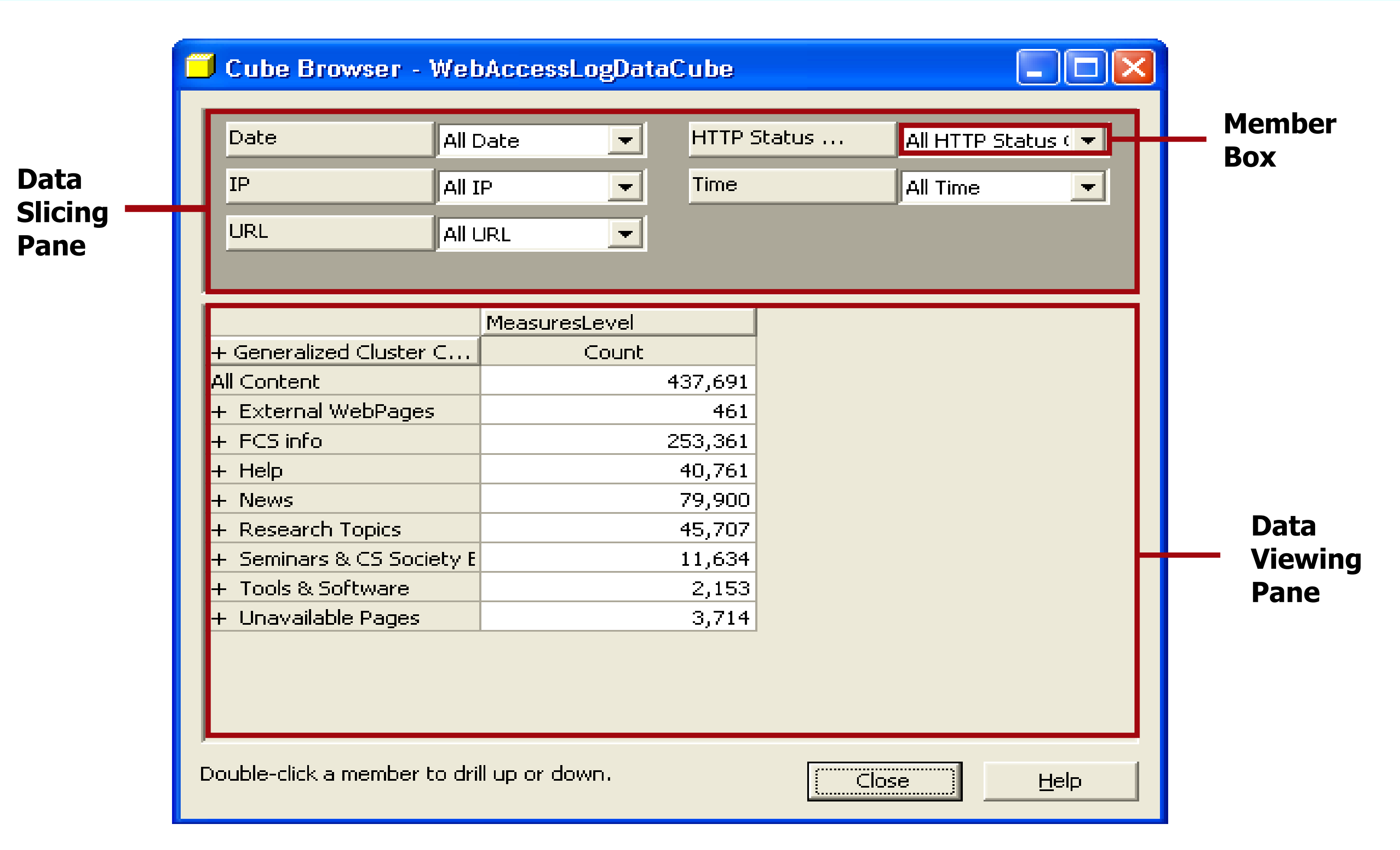
HTTP_Status_Code

HTTP_Status_Desc

200, 206, 301, 302 and 304

HTTP Status Code	HTTP Status Code Description
1XX	Information
2XX	Success
3XX	Redirection

OLAP Interface



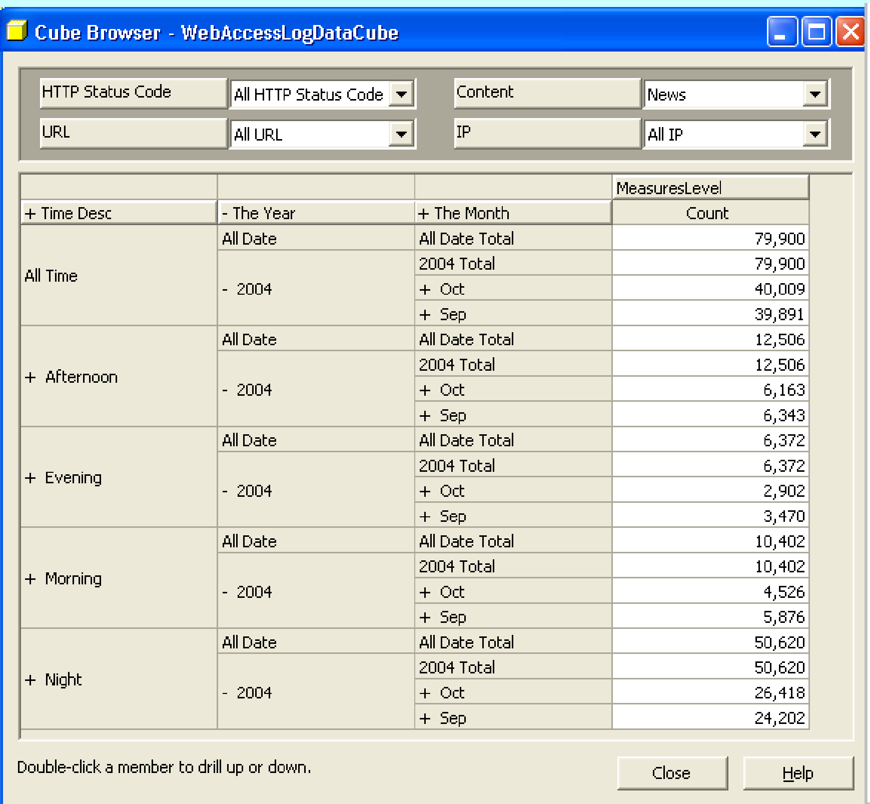
Date	All Date
IP	All IP
URL	All URL
HTTP Status ...	All HTTP Status (
Time	All Time

	MeasuresLevel
+ Generalized Cluster C...	Count
All Content	437,691
+ External WebPages	461
+ FCS info	253,361
+ Help	40,761
+ News	79,900
+ Research Topics	45,707
+ Seminars & CS Society E	11,634
+ Tools & Software	2,153
+ Unavailable Pages	3,714

Double-click a member to drill up or down.

Result Demonstration

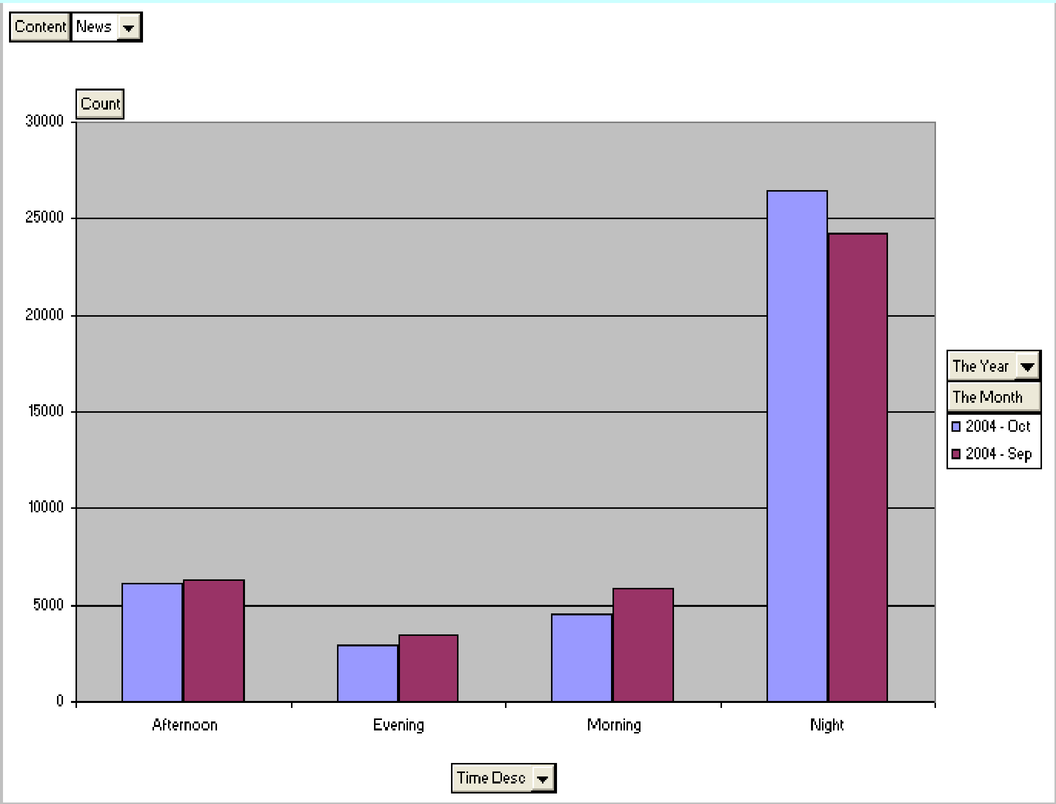
Query 1: Suppose the website is redesigned in September, and the webmaster wants to know whether or not it effected the number of visitors and the ways they accessed.



HTTP Status Code	All HTTP Status Code	Content	News
URL	All URL	IP	All IP

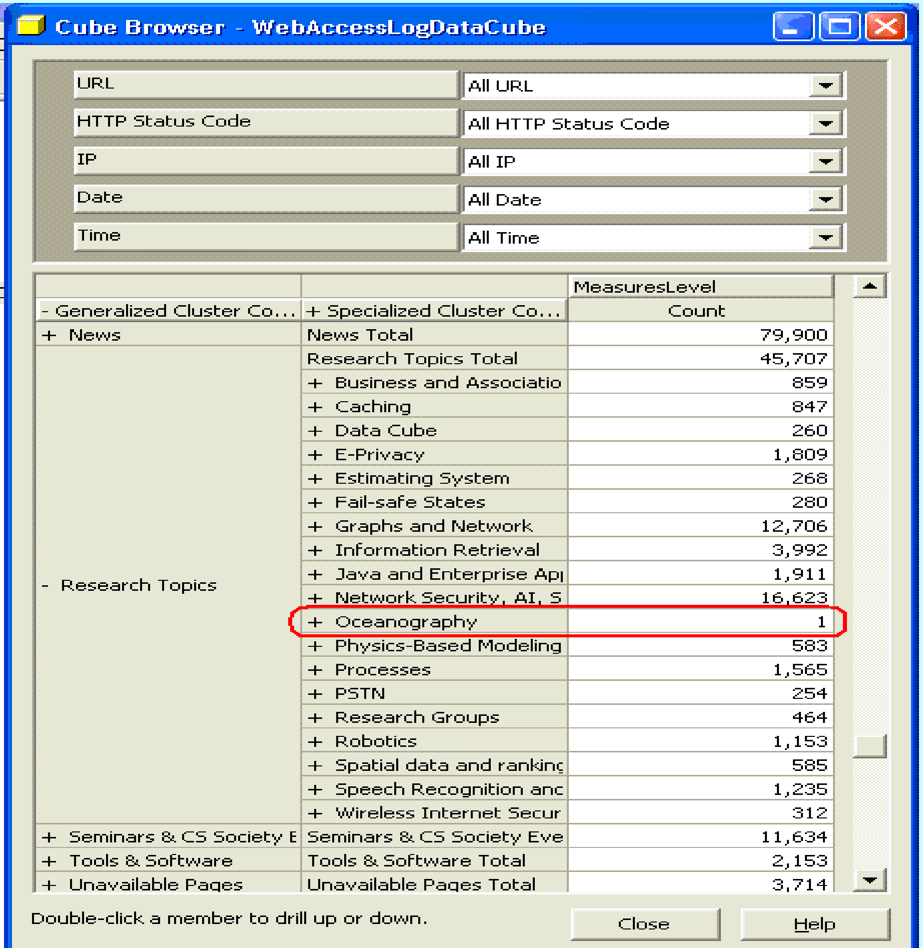
			MeasuresLevel
+ Time Desc	- The Year	+ The Month	Count
All Time	All Date	All Date Total	79,900
	- 2004	2004 Total	79,900
		+ Oct	40,009
		+ Sep	39,891
+ Afternoon	All Date	All Date Total	12,506
	- 2004	2004 Total	12,506
		+ Oct	6,163
		+ Sep	6,343
+ Evening	All Date	All Date Total	6,372
	- 2004	2004 Total	6,372
		+ Oct	2,902
		+ Sep	3,470
+ Morning	All Date	All Date Total	10,402
	- 2004	2004 Total	10,402
		+ Oct	4,526
		+ Sep	5,876
+ Night	All Date	All Date Total	50,620
	- 2004	2004 Total	50,620
		+ Oct	26,418
		+ Sep	24,202

Double-click a member to drill up or down.



Result Demonstration (Cont.)

Query 2: What is the browsing distribution of different research topics in Oct and Sept, 2004?

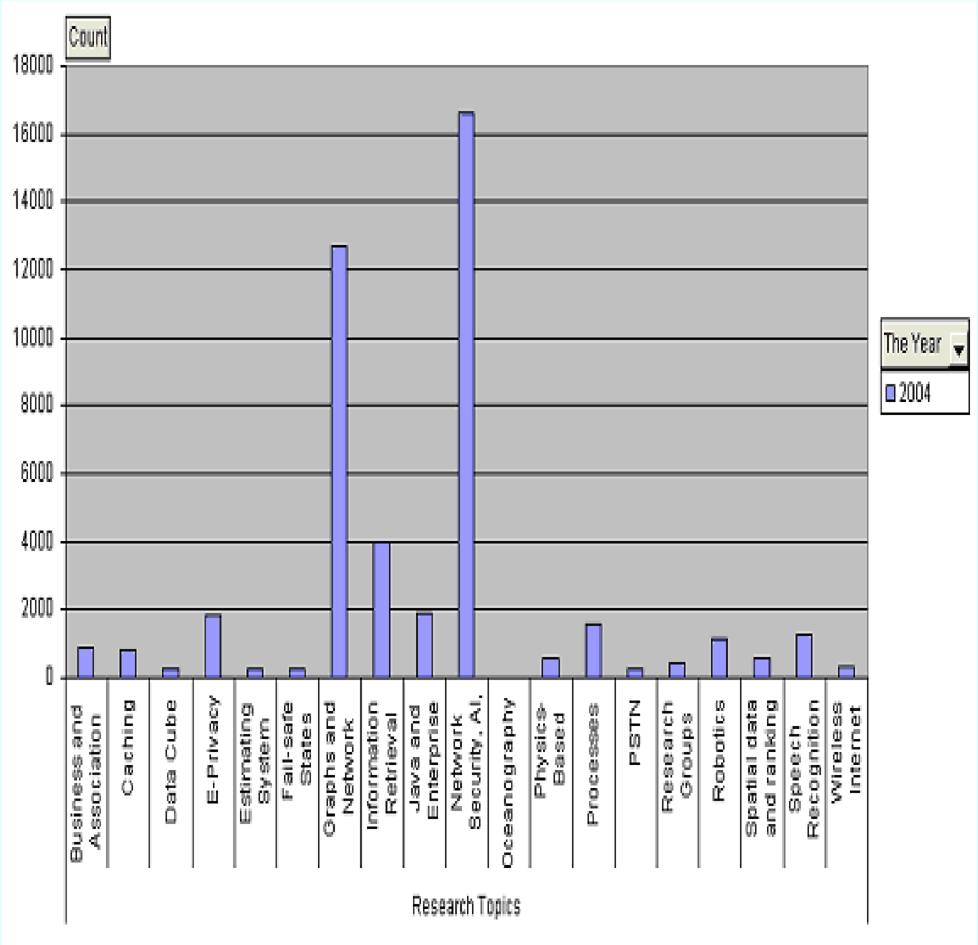


Cube Browser - WebAccessLogDataCube

URL	All URL
HTTP Status Code	All HTTP Status Code
IP	All IP
Date	All Date
Time	All Time

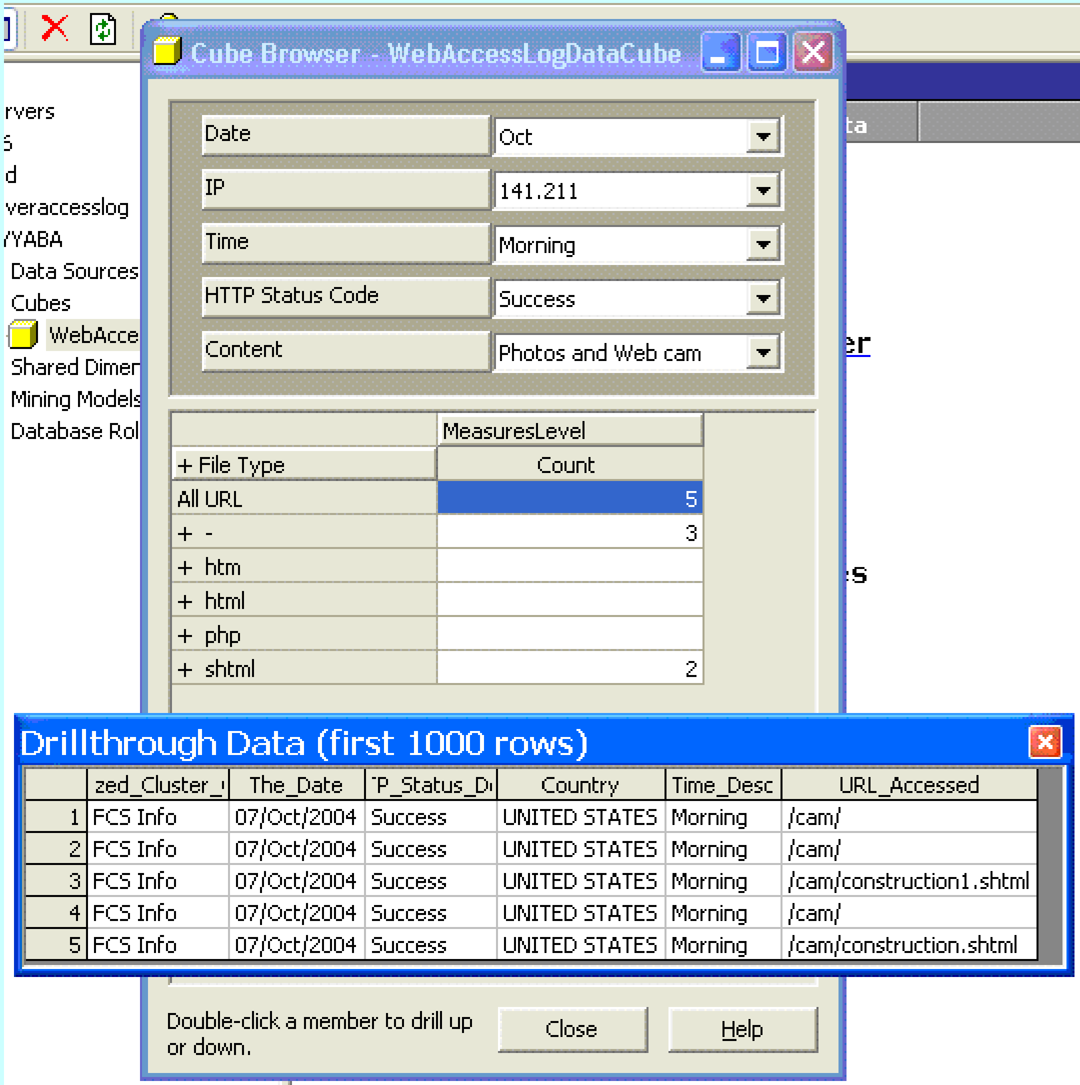
		MeasuresLevel
- Generalized Cluster Co...	+ Specialized Cluster Co...	Count
+ News	News Total	79,900
- Research Topics	Research Topics Total	45,707
	+ Business and Associatio	859
	+ Caching	847
	+ Data Cube	260
	+ E-Privacy	1,809
	+ Estimating System	268
	+ Fail-safe States	280
	+ Graphs and Network	12,706
	+ Information Retrieval	3,992
	+ Java and Enterprise Api	1,911
	+ Network Security, AI, S	16,623
	+ Oceanography	1
	+ Physics-Based Modeling	583
	+ Processes	1,565
	+ PSTN	254
	+ Research Groups	464
	+ Robotics	1,153
	+ Spatial data and rankinc	585
	+ Speech Recognition anc	1,235
	+ Wireless Internet Secur	312
+ Seminars & CS Society E	Seminars & CS Society Eve	11,634
+ Tools & Software	Tools & Software Total	2,153
+ Unavailable Pages	Unavailable Pages Total	3,714

Double-click a member to drill up or down.



Result Demonstration (Cont.)

Query 3: How many times Web pages related to ‘Photos and Web cam’ are accessed by the network 141.211 in October morning which were displayed successfully?



Date	Oct
IP	141.211
Time	Morning
HTTP Status Code	Success
Content	Photos and Web cam

	MeasuresLevel
+ File Type	Count
All URL	5
+ -	3
+ htm	
+ html	
+ php	
+ shtml	2

Drillthrough Data (first 1000 rows)

	zed_Cluster_	The_Date	'P_Status_D	Country	Time_Desc	URL_Accessed
1	FCS Info	07/Oct/2004	Success	UNITED STATES	Morning	/cam/
2	FCS Info	07/Oct/2004	Success	UNITED STATES	Morning	/cam/
3	FCS Info	07/Oct/2004	Success	UNITED STATES	Morning	/cam/construction1.shtml
4	FCS Info	07/Oct/2004	Success	UNITED STATES	Morning	/cam/
5	FCS Info	07/Oct/2004	Success	UNITED STATES	Morning	/cam/construction.shtml

Double-click a member to drill up or down.

Summary

- **Why data warehousing?**
  - Major differences between OLAP and OLTP
- **Multi-dimensional model and data warehouse schemas**
  - Logic dimensions and space: virtual data cube (dimensions & measures)
  - Star schema, snowflake schema, fact constellations
- **OLAP operations**
  - Structures for supporting cuboids manipulation: lattice of cuboids & concept hierarchies
  - Typical operators: drilling, rolling, slicing, dicing and pivoting
  - A complex ad hoc query may be partitioned into multiple OLAP operations
- **DW and DW server implementation issue**s (only the concepts discussed in class)
  - Options of data cube materialization: full and partial materialization
  - Options of DW server for OLAP processing: ROLAP, MOLAP and HOLAP
- **Data warehouse architecture**
- From OLAP to OLAM (on-line analytical mining)

Review Questions

1. What are the 3 typical DW schemas (describe each)?
2. What are the main differences between OLAP and OLTP process operations for answering users' queries?
3. What is the difference between logic dimensional space for OLAP analysis and the physical space for 3D computer graphics?
4. How to define OLAP queries based on Starnet query model?
5. What is the visualization metaphor for displaying OLAP result of a given multiple logic space (data cube), and how to best use it?
6. Use the demo example of "Wealth vs. Health" analysis (http://www.youtube.com/watch?v=jbkSRLYSojo) to explain how DW/OLAP may be able to support DM applications.