CP5806
WEEK 4

JAMES COOK UNIVERSITY
AUSTRALIA

DUE DATES

- Assignment 2 due Sunday 9 August, 11:59 pm (Week 5)
  - Sisi will hold another Saturday sessions for A2
    - Week 4 Saturday 1 August 2:30-3:30 pm

JULY

	M	T	W	T	F	S	S
O Week			1	2	3	4	5
wk 1	6	7	8	9	10	11	12
wk 2	13	14	15	16	17	18	19
wk 3	20	21	22	23	24	25	26
wk 4	27	28	29	30	31		

AUGUST

	M	T	W	T	F	S	S
wk 4						1	2
wk 5	3	4	5	6	7	8	9
wk 6	10	11	12	13	14	15	16
wk 7	17	18	19	20	21	22	23
O Week	24	25	26	27	28	29	30
wk 1	31						

WEEK 4 LEARNING OUTCOMES

- Improve the quality of data in the data warehouse
- Analyse the challenges posed by corrupt data and apply methods for dealing with them
- Apply OLAP operations to effectively answer business queries
- Examine the different OLAP models and determine which model is suitable for your environment
- Apply data cube computation and materialisation to efficiently implement data warehouses
- Examine data cube technologies and select a suitable approach for your business case.

TOPICS FOR WEEK 4

- Topic 1: Data quality
- Topic 2: OLAP basics
- Topic 3: OLAP operations
- Topic 4: Data warehouse implementation
- Topic 5: Data cube technologies

TOPIC 1: DATA QUALITY

- Why improve data quality
- Data accuracy vs data quality
- Characteristics of high-quality data
- Types of data quality problems
- Data quality challenges
- Data quality tools

JAMES COOK UNIVERSITY AUSTRALIA

WHY IMPROVE DATA QUALITY?

- Boosts confidence in decision making.
- Enables better customer service.
- Increases opportunity to add better value to the services.
- Reduces risk from disastrous decisions.
- Reduces costs, especially of marketing campaigns.
- Enhances strategic decision making.
- Improves productivity by streamlining processes.
- Avoids compounding the effects of data contamination.

DATA ACCURACY VERSUS DATA QUALITY

DATA INTEGRITY	DATA QUALITY
Specific instance of an entity accurately represents that occurrence of the entity.	The data item is exactly fit for the purpose for which the business users have defined it.
Data element defined in terms of database technology.	Wider concept grounded in the specific business of the company.
Data element conforms to validation constraints.	Relates not just to single data elements but to the system as a whole.
Individual data items have the correct data types.	Form and content of data elements consistent across the whole system.
Traditionally relates to operational systems.	Essentially needed in a corporate-wide data warehouse for business users.

Figure 13-1 Data accuracy versus data quality.

JAMES COOK UNIVERSITY AUSTRALIA

CHARACTERISTICS OF HIGH-QUALITY DATA

- Accuracy
- Domain integrity
- Data Type
- Consistency
- Redundancy
- Completeness
- Duplication
- Conformance to business rules
- Structural definiteness
- Data anomaly
- Clarity
- Timely
- Usefulness
- Adherence to data integrity rules

BENEFITS OF IMPROVED DATA QUALITY

- Analysis with timely information
- Better customer service
- Newer opportunities
- Reduced costs and risks
- Improved productivity
- Reliable strategic decision making

TYPES OF DATA QUALITY PROBLEMS

- Dummy values in fields
- Absence of data values
- Unofficial use of fields
- Cryptic values
- Contradicting values
- Business rule violations
- Reused primary keys
- Nonunique identifiers
- Inconsistent values
- Incorrect values
- Multipurpose fields
- Erroneous integration

DATA QUALITY CHALLENGES

- Data pollution:
  - System conversions
  - Data aging
  - Heterogeneous system integration
  - Poor database design
  - Incomplete information at data entry
  - Input errors
  - Internationalization/Localization
  - Fraud
  - Lack of policies
- Validation of names and addresses
- Costs of poor data quality

JAMES COOK UNIVERSITY AUSTRALIA

DATA QUALITY TOOLS

- Data cleansing tools contain useful error discovery and error correction features.
- The DBMS itself can be used for data cleansing:
  - Domain integrity
  - Update security
  - Entity integrity checking
  - Minimize missing values
  - Referential integrity checking
  - Conformance to business rules
- Master Data Management (MDM) initiatives provide a means for ensuring data quality in the data warehouse.

TOPIC 2: OLAP BASICS

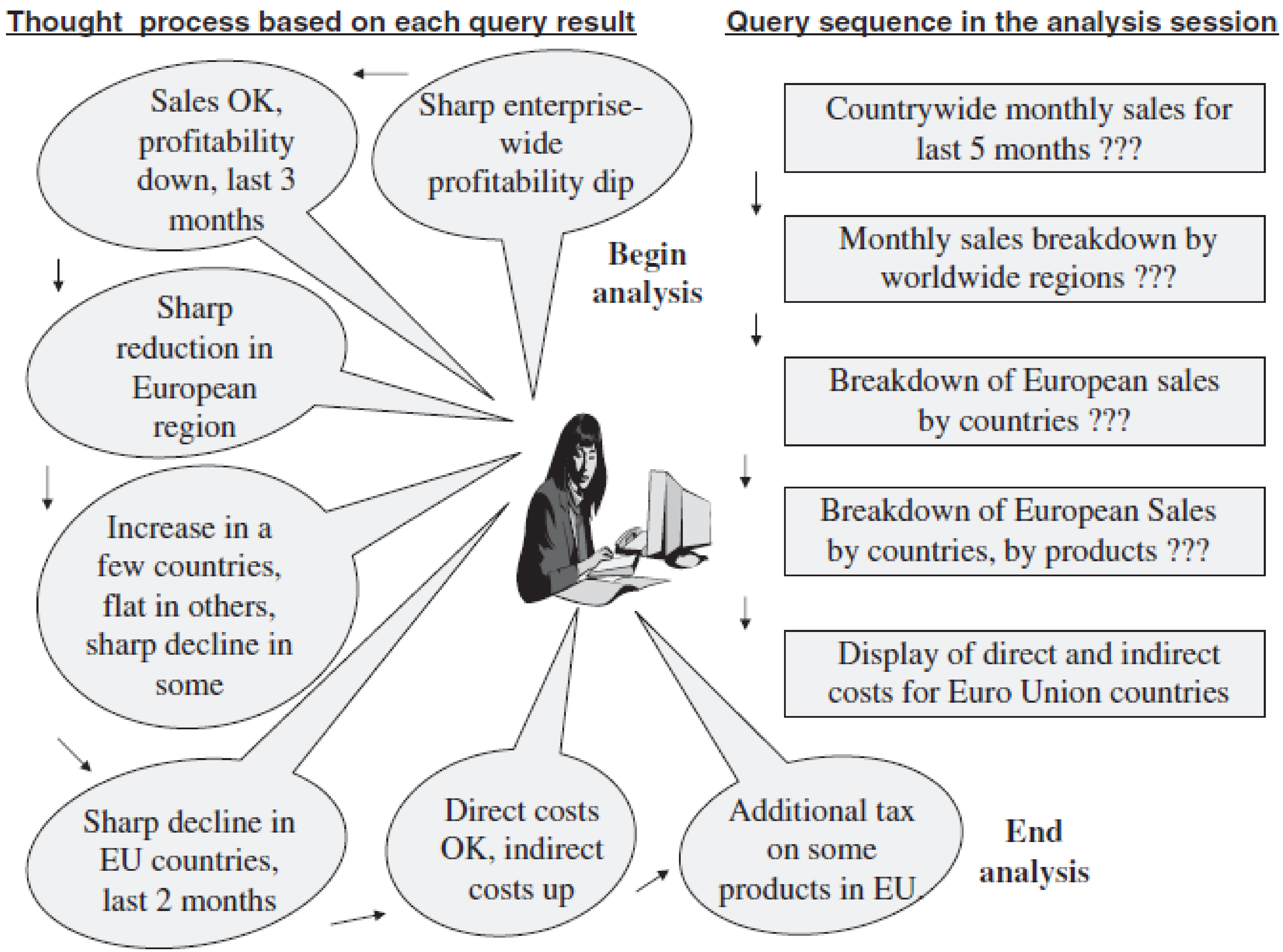


Figure 15-1 Query steps in an analysis session.

TOPIC 2: OLAP BASICS

LINE	TOTAL SALES
Clothing	$12,836,450
Electronics	$16,068,300
Video	$21,262,190
Kitchen	$17,704,400
Appliances	$19,600,800
Total	$87,472,140

1 High level summary by product line

2 Drill down by year

LINE	1998	1999	2000	TOTAL
Clothing	$3,457,000	$3,590,050	$5,789,400	$12,836,450
Electronics	$5,894,800	$4,078,900	$6,094,600	$16,068,300
Video	$7,198,700	$6,057,890	$8,005,600	$21,262,190
Kitchen	$4,875,400	$5,894,500	$6,934,500	$17,704,400
Appliances	$5,947,300	$6,104,500	$7,549,000	$19,600,800
Total	$27,373,200	$25,725,840	$34,373,100	$87,472,140

3 Rotate columns to rows

YEAR	Clothing	Electronics	Video	Kitchen	Appliances	TOTAL
1998	$3,457,000	$5,894,800	$7,198,700	$4,875,400	$5,947,300	$27,373,200
1999	$3,590,050	$4,078,900	$6,057,890	$5,894,500	$6,104,500	$25,725,840
2000	$5,789,400	$6,094,600	$8,005,600	$6,934,500	$7,549,000	$34,373,100
Total	$12,836,450	$16,068,300	$21,262,190	$17,704,400	$19,600,800	$87,472,140

Figure 15-3 Simple OLAP session.

TOPIC 2: OLAP BASICS

- OLAP Definitions and Rules
- 12 guidelines for an OLAP system
- Major features and functions of OLAP
  - Dimensional analysis
  - Drill down and roll up
  - Slice and dice or rotation
  - Uses and benefits
- OLAP models
  - ROLAP: relational online analytical processing
  - MOLAP: multidimensional online analytical processing
  - HOLAP: hybrid online analytical processing
  - DOLAP: desktop online analytical processing
  - Database OLAP: relational database management system (RDBMS) designated to support OLAP structures and to perform OLAP calculations
  - Web OLAP: online analytical processing where OLAP data is accessible from a Web browser
- OLAP implementation considerateness

JAMES COOK UNIVERSITY AUSTRALIA

MAJOR FEATURES AND FUNCTIONS OF OLAP

BASIC FEATURES

Multidimensional analysis	Consistent performance	Fast response times for interactive queries
Drill-down and roll-up	Navigation in and out of details	Slice-and-dice or rotation
Multiple view modes	Easy scalability	Time intelligence (year-to-date, fiscal period)

ADVANCED FEATURES

Powerful calculations	Cross-dimensional calculations	Pre-calculation or pre-consolidation
Drill-through across dimensions or details	Sophisticated presentation & displays	Collaborative decision making
Derived data values through formulas	Application of alert technology	Report generation with agent technology

Figure 15-4 General features of OLAP.

MAJOR FEATURES AND FUNCTIONS OF OLAP

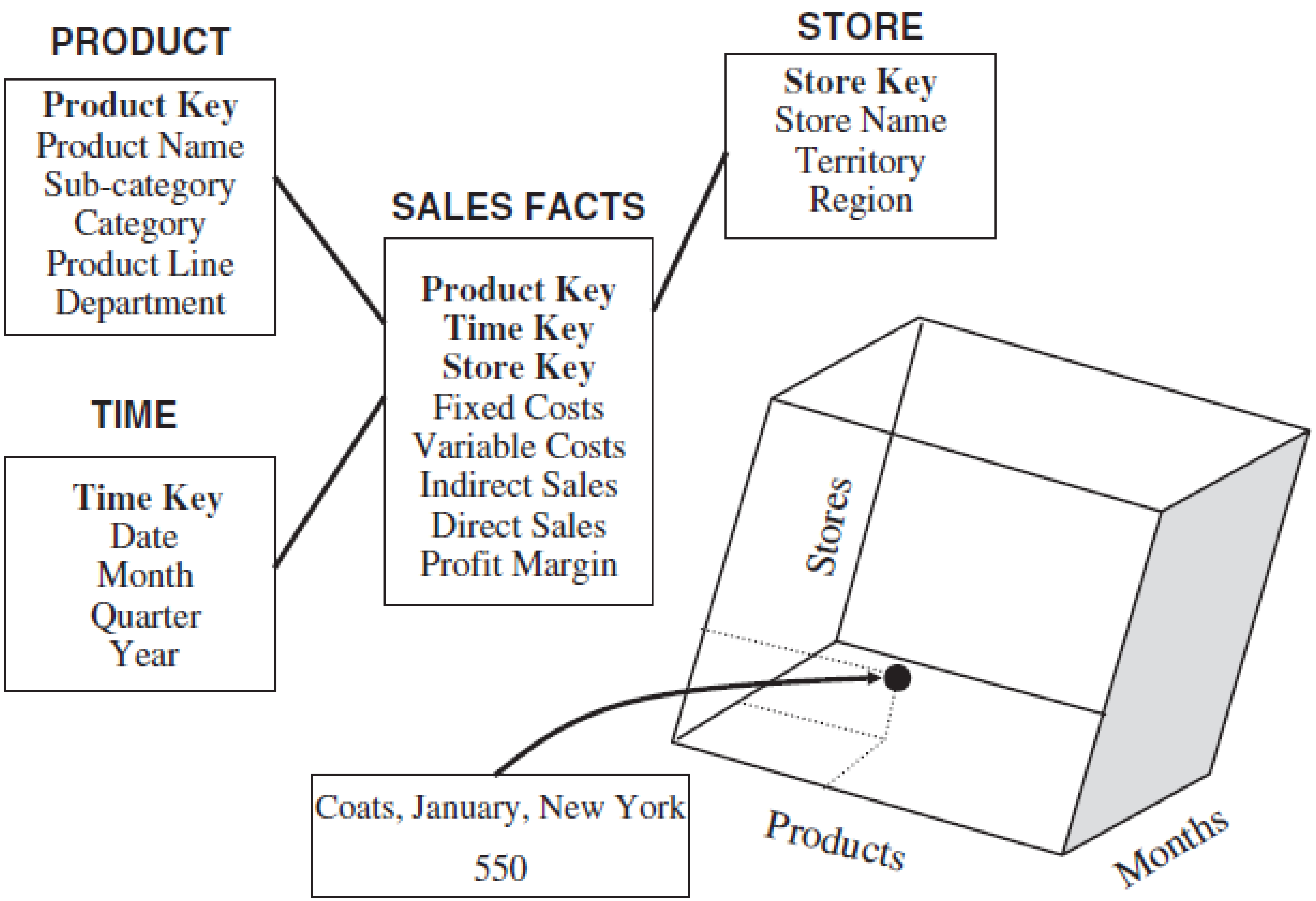


Figure 15-5 Simple STAR schema.

MAJOR FEATURES AND FUNCTIONS OF OLAP

Store: New York

PAGES: STORE dimension

Products

COLUMNS: PRODUCT dimension

ROWS: TIME dimension

Months

	Hats	Coats	Jackets	Dresses	Shirts	Slacks
Jan	200	550	350	500	520	490
Feb	210	480	390	510	530	500
Mar	190	480	380	480	500	470
Apr	190	430	350	490	510	480
May	160	530	320	530	550	520
Jun	150	450	310	540	560	330
Jul	130	480	270	550	570	250
Aug	140	570	250	650	670	230
Sep	160	470	240	630	650	210
Oct	170	480	260	610	630	250
Nov	180	520	280	680	700	260
Dec	200	560	320	750	770	310

Figure 15-6 A three-dimensional display.

JAMES COOK UNIVERSITY AUSTRALIA

MAJOR FEATURES AND FUNCTIONS OF OLAP

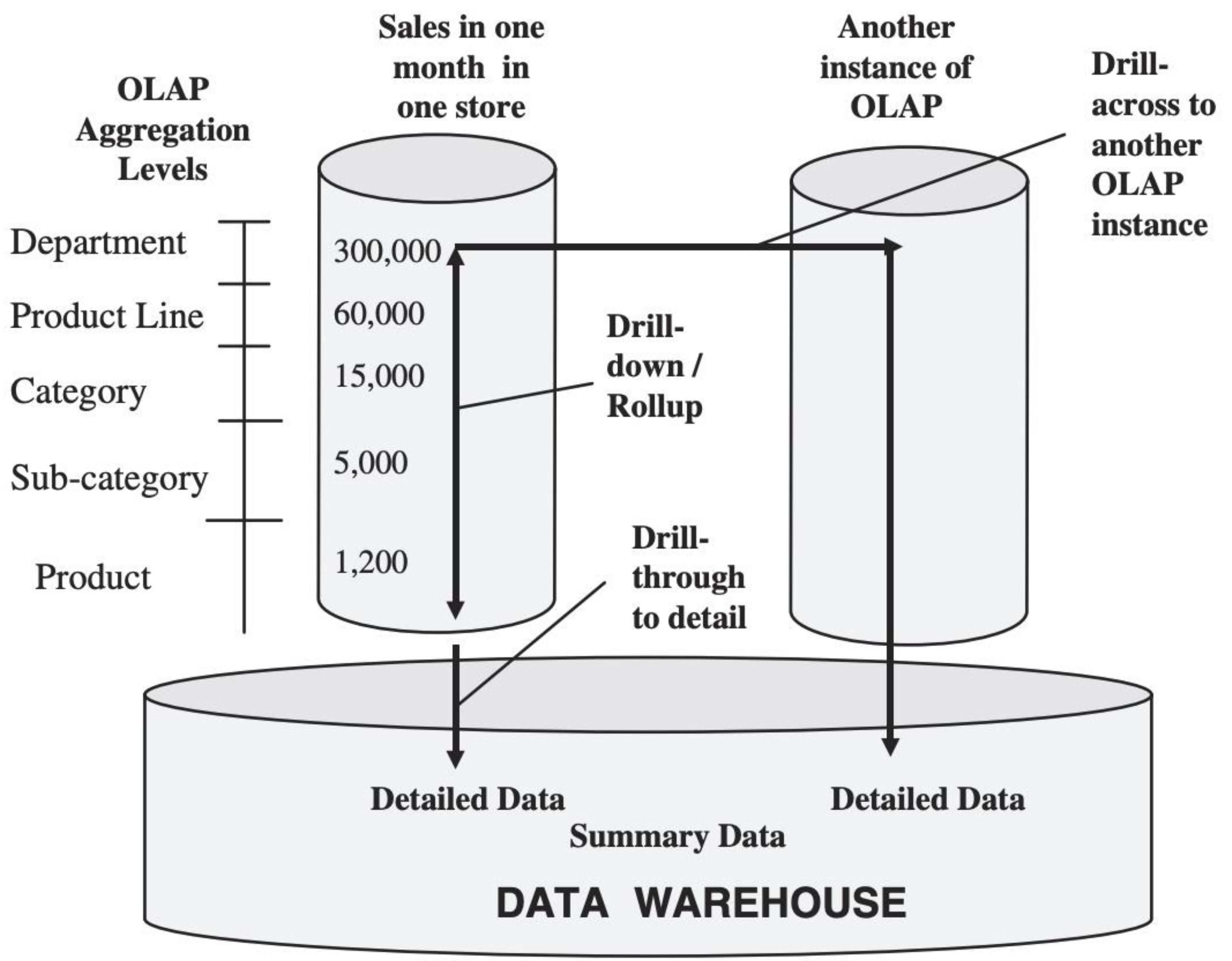
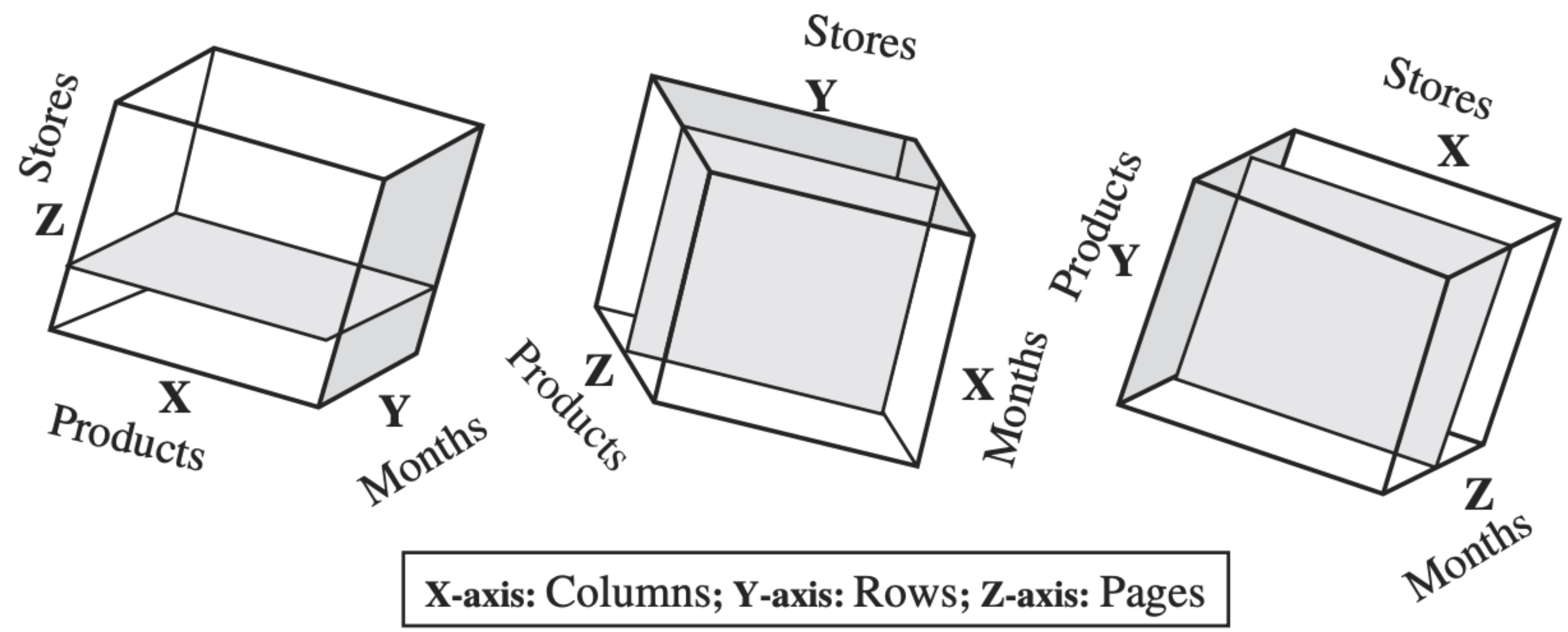


Figure 15-12 Roll-up and drill-down features of OLAP.

MAJOR FEATURES AND FUNCTIONS OF OLAP



Store: New York

	Hats	Coats	Jackets
Jan	200	550	350
Feb	210	480	390
Mar	190	480	380

Product: Hats

	Jan	Feb	Mar
New York	200	210	190
Boston	210	250	240
San Jose	130	90	70

Month: January

	New York	Boston	San Jose
Hats	200	210	130
Coats	550	500	200
Jackets	350	400	100

Figure 15-14 Slicing and dicing.

ROLAP VERSUS MOLAP

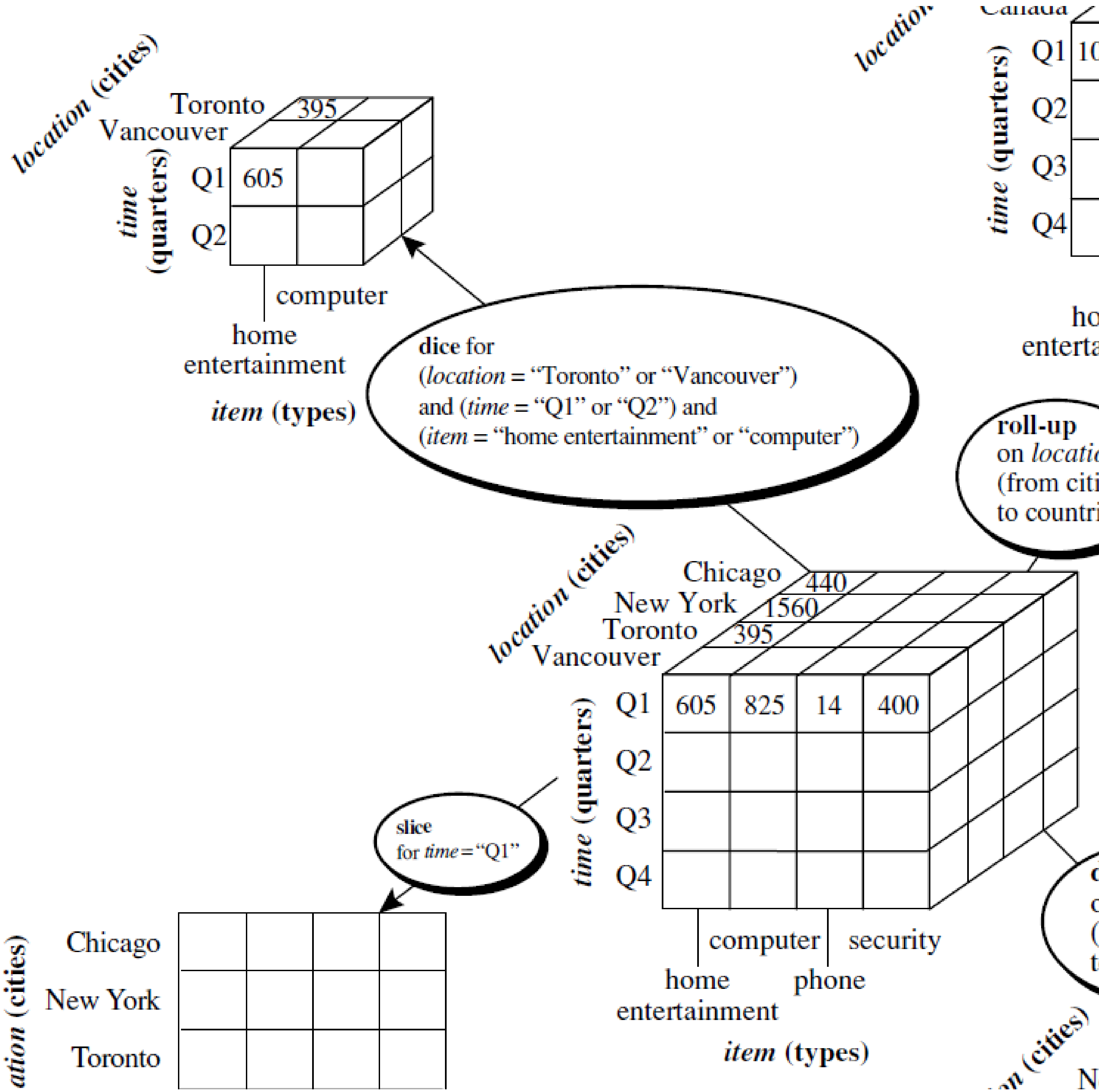
	Data Storage	Underlying Technologies	Functions and Features
ROLAP	Data stored as relational tables in the warehouse. Detailed and light summary data available. Very large data volumes. All data access from the warehouse storage.	Use of complex SQL to fetch data from warehouse. ROLAP engine in analytical server creates data cubes on the fly. Multidimensional views by presentation layer.	Known environment and availability of many tools. Limitations on complex analysis functions. Drill-through to lowest level easier. Drill-across not always easy.
MOLAP	Data stored as relational tables in the warehouse. Various summary data kept in proprietary databases (MDDBs) Moderate data volumes. Summary data access from MDDB, detailed data access from warehouse.	Creation of pre-fabricated data cubes by MOLAP engine. Propriety technology to store multidimensional views in arrays, not tables. High speed matrix data retrieval. Sparse matrix technology to manage data sparsity in summaries.	Faster access. Large library of functions for complex calculations. Easy analysis irrespective of the number of dimensions. Extensive drill-down and slice-and-dice capabilities.

Figure 15-19 ROLAP versus MOLAP.

TOPIC 3: OLAP OPERATIONS

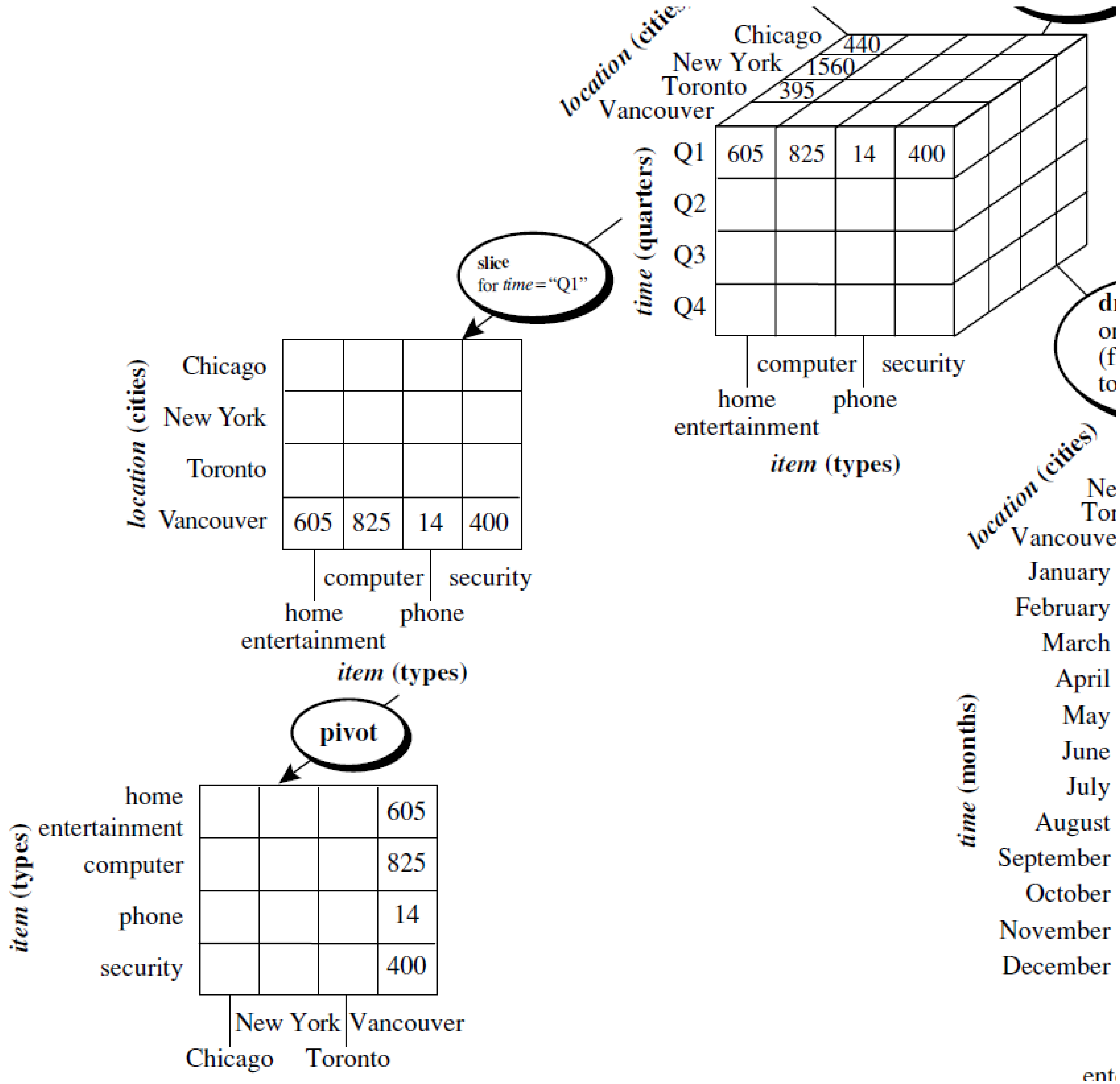
- Roll-up: aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction
- Drill-down: reverse of roll-up. It navigates from less detailed data to more detailed data
- Slice and dice: subcube selection
- Pivot(rotate): visualization operation
- Other OLAP operations: drill-across; drill-through

OLAP OPERATIONS



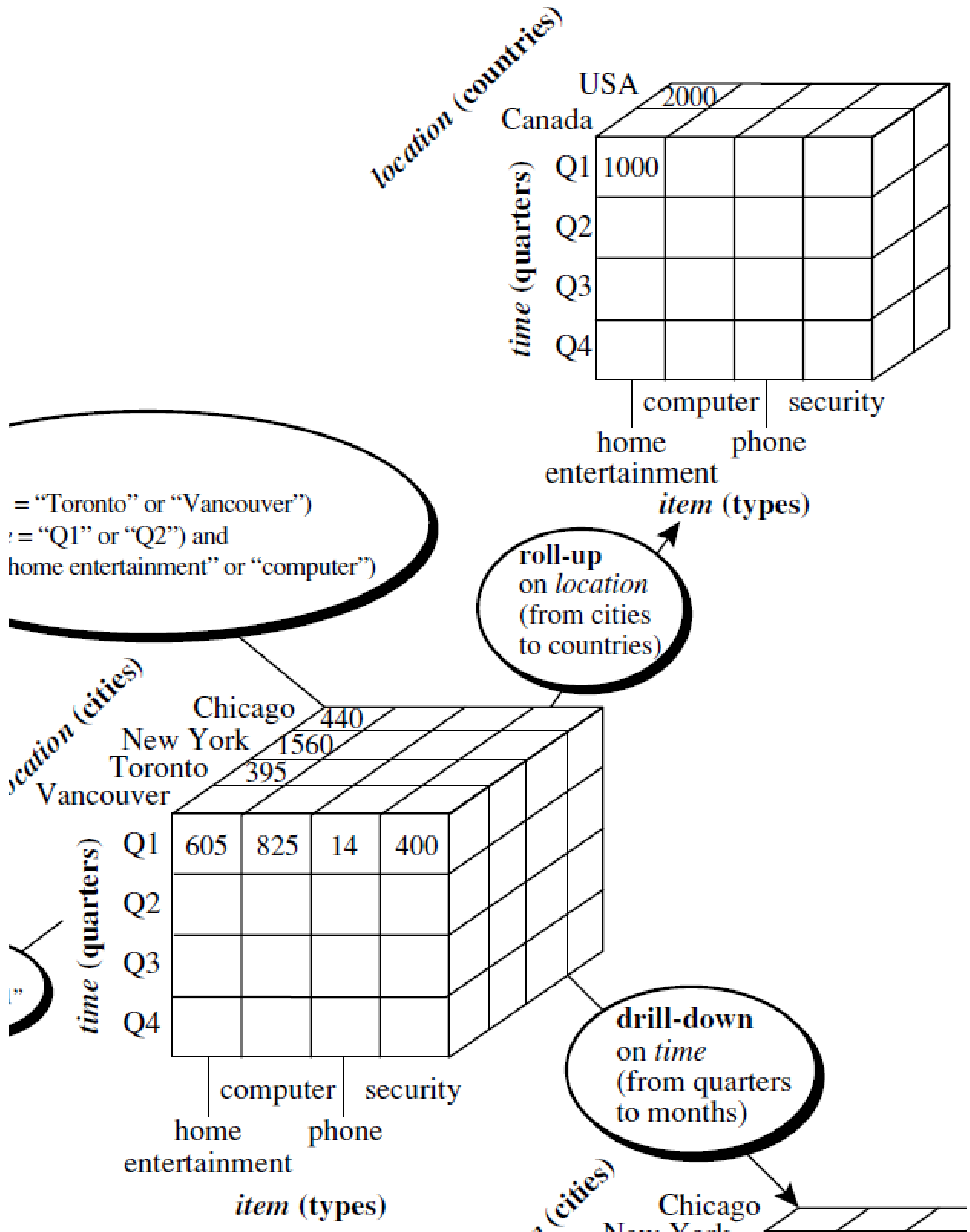
JAMES COOK UNIVERSITY AUSTRALIA

OLAP OPERATIONS

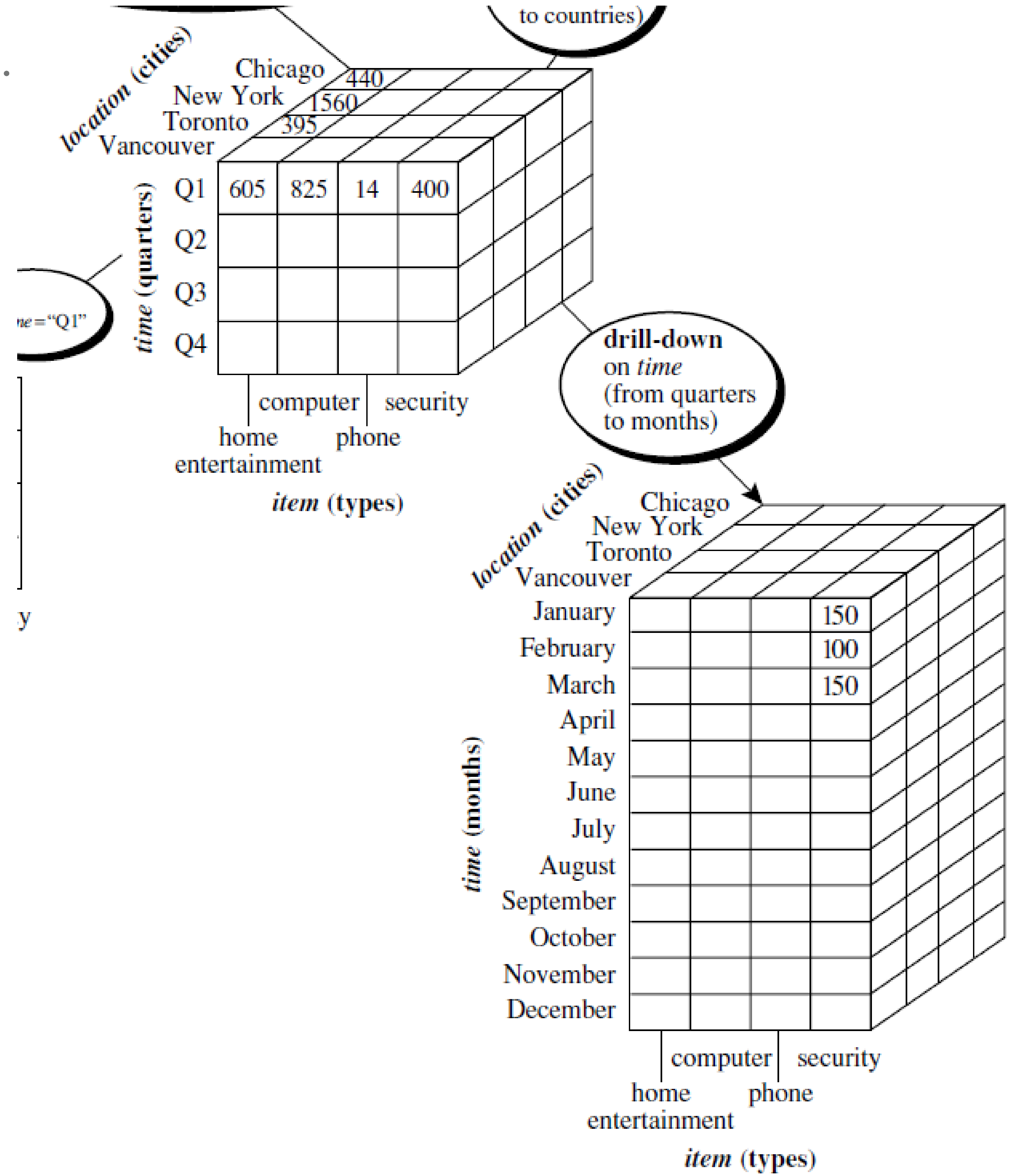


JAMES COOK UNIVERSITY AUSTRALIA

OLAP OPERATIONS



OLAP OPERATIONS



TOPIC 4&5: DATA WAREHOUSE IMPLEMENTATION & DATA CUBE TECHNOLOGIES

- Data cube essentials
- Conceptual modelling of data warehouses
- Cube Materialization
  - No materialization
  - Full materialization
  - Partial materialization
- Indexing OLAP Data
  - Bitmap indexing
  - Join indexing
- Data cube computation
  - General strategies
  - Methods

COMPONENTS

Three Concepts in Data Cubes

Three concepts in data cubes are

(1) Multidimension (2) Hierarchy

(3) Measure

Fact Table

4 dimensions 2 measures

Date	Branch	Item	Buyer	Units sold	Dollars sold
1/1/2008	London	VCD	First Company	20	5000
1/1/2008	Bangkok	TV	First Company	30	9000
10/1/2008	London	Ham	First Company	20	1000
4/2/2008	London	Milk	First Company	80	1600
15/2/2008	Bangkok	VCD	Best Company	30	7500
2/5/2008	Bangkok	Orange	Best Company	20	500

Location Dimension Table (3 levels)

Branch	Country	Continent
London	UK	Europe
Glasgow	UK	Europe
Berlin	Germany	Europe
Bangkok	Thailand	Asia
Phuket	Thailand	Asia
Tokyo	Japan	Asia

Customer Dimension Table (2 levels)

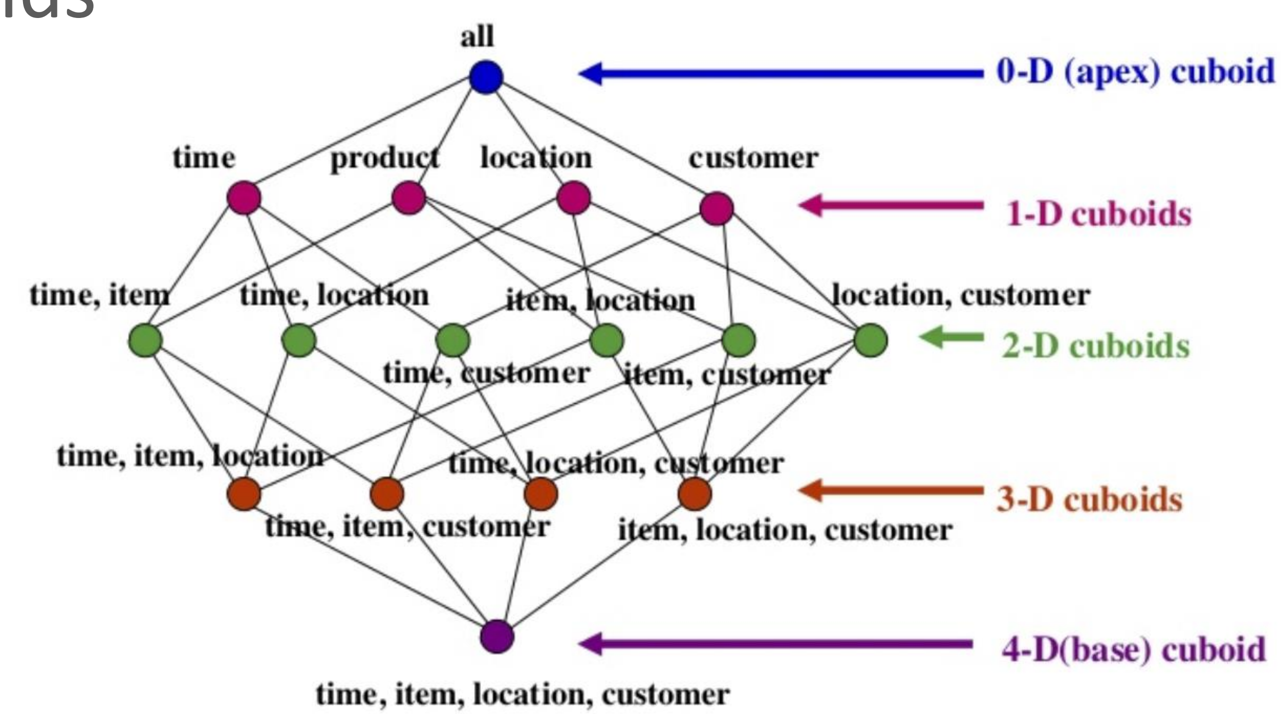
Buyer	Buyer Group
First Company	Group 1
Second Company	Group 1
Third Company	Group 1
Best Company	Group 2
Good Company	Group 2

Product Dimension Table (3 levels)

Item	Subcategory	Category
VCD	Electric	Non-Food
TV	Electric	Non-Food
Shirt	Clothes	Non-Food
Ham	Process food	Food
Milk	Fresh food	Food
Orange	Fresh food	Food

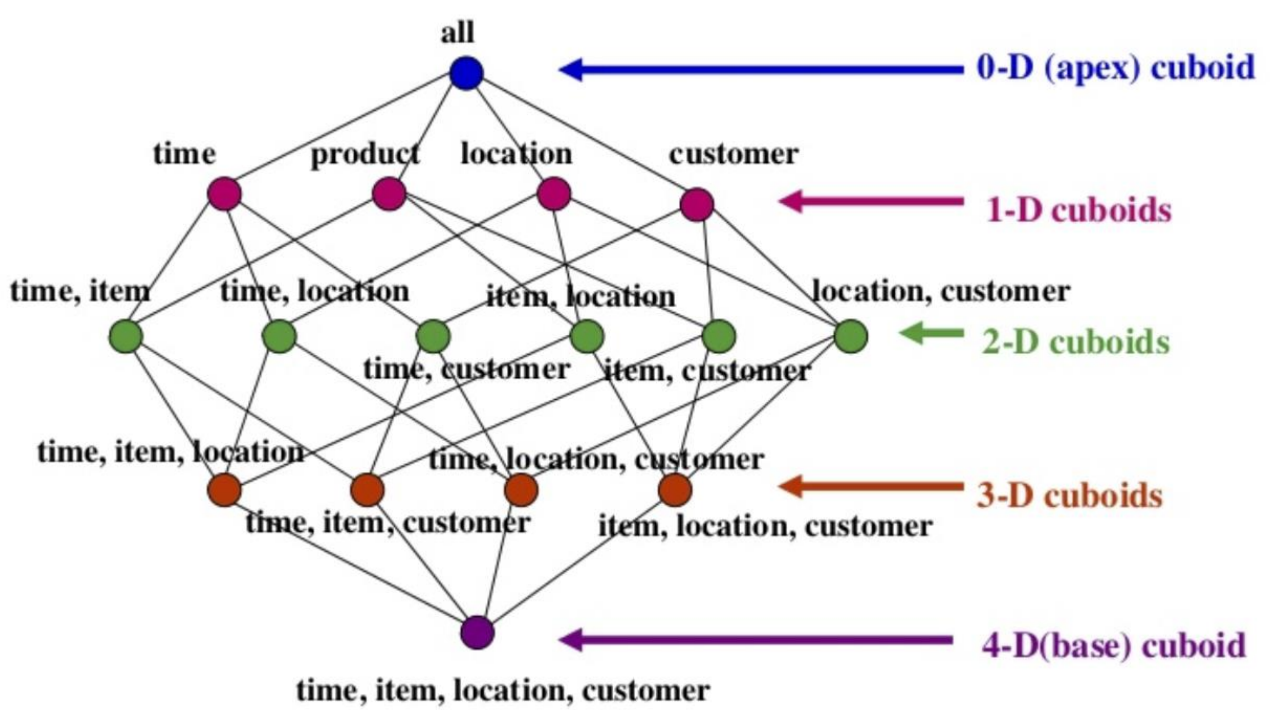
CUBOIDS IN DATA CUBES

- Cuboid concept is formed by the number of dimensions
- **Apex cuboid**: top most 0-D cuboid, which holds the highest level of summarization (one number – e.g. what is the overall sales across all dimensions?)
- **Base cuboid**: n-D base cube where n is the total number of dimensions (e.g. all the sales for all the combinations of dimensions)
- A data cube is the lattice of cuboids



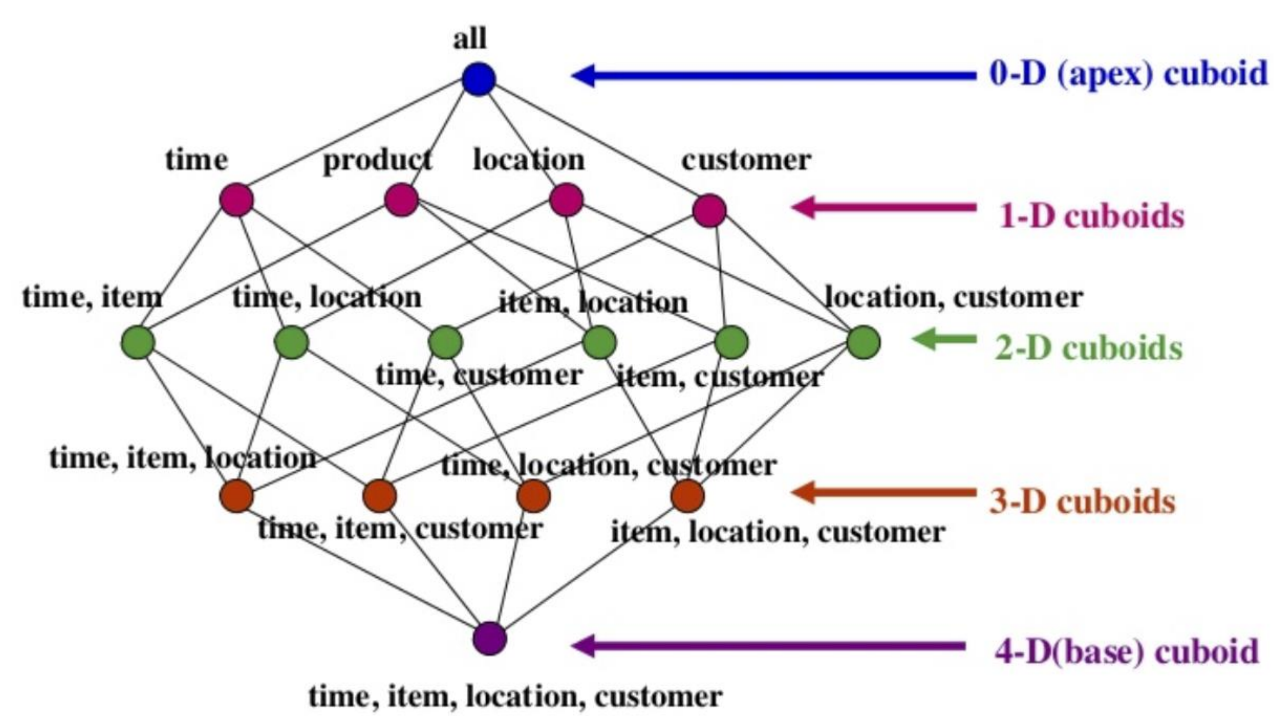
CUBOIDS IN DATA CUBES

- **Base cell**: a cell in the base cuboid
- **m-dimensional cell**: a cell from an m-dimensional cuboid
  - (Jan, Chicago, Business, 45) is a 3D cell (from the base cuboid)
  - (Jan, *, Business, 150) is a 2D cell
  - (Jan, *, *, 2800) is a 1D cell
  - (*, *, *, 35500) is only cell of the apex cuboid



CUBOIDS IN DATA CUBES

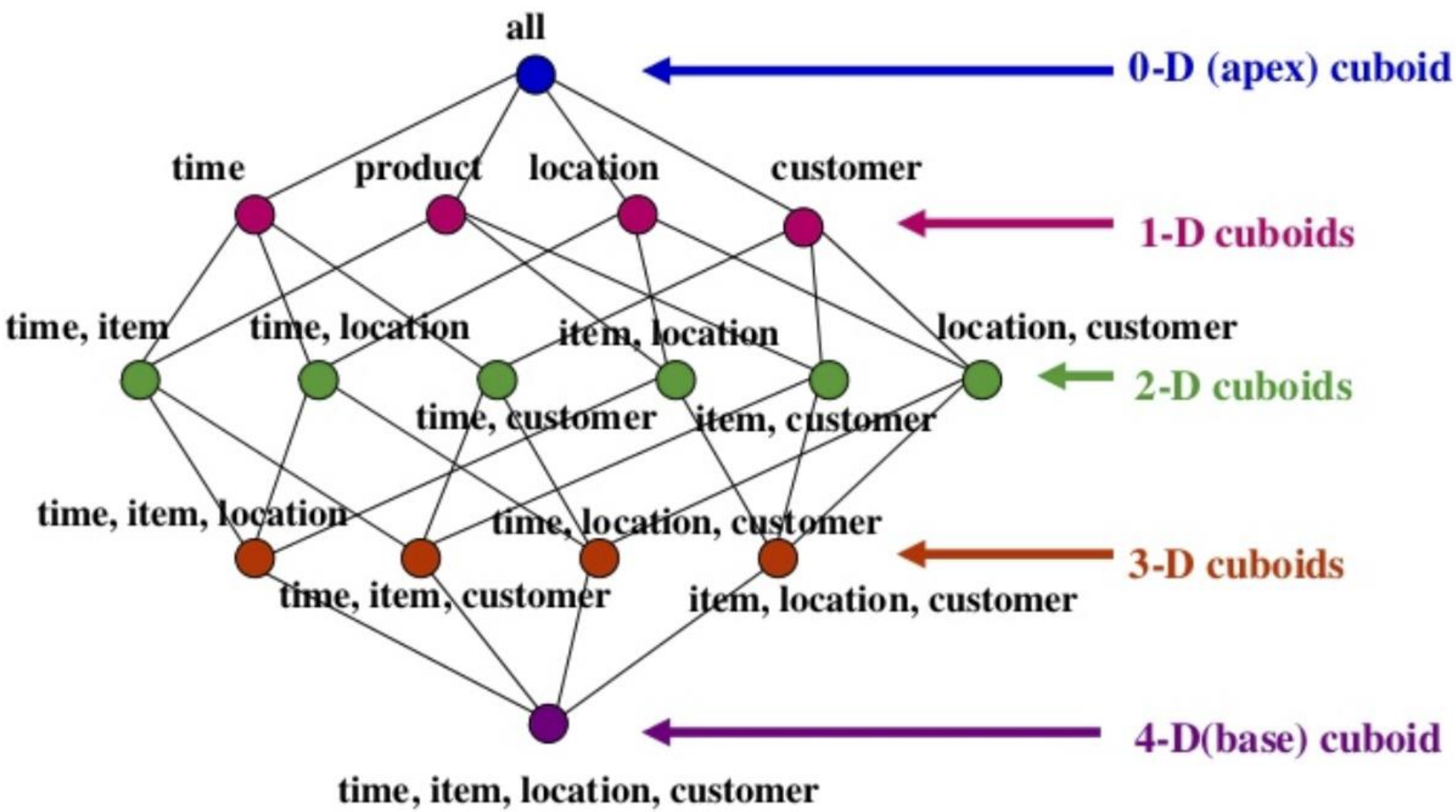
- **Parent and child cells:**
  - (Jan, *, Business, 150) is a child of (Jan, *, *, 2800)
  - (Jan, *, Business, 150) is also a child of (*, *, Business, 1200)
- **Ancestor and descendant cells:**
  - A parent (of a parent of a parent…) is an ancestor cell
  - A child (of a child of a child…) is a descendant cell



CUBE: A LATTICE OF CUBOIDS

- Calculate the number of cuboids in an n-dimensional cube with L levels
  - No hierarchies, # of cuboids = 2^n
  - With hierarchies, # of cuboids = product of 1 + number of the levels for each dimension
  - Calculate the number of cells (supposing each dimension has m distinct values in the base cuboid)
  - Maximum number of cells in the base cuboid: m^n
  - Minimum number of cells in the base cuboid: m

CUBE: A LATTICE OF CUBOIDS



THREE CATEGORIES OF MEASURE

- **Distributive functions**: An aggregate function is distributive if a set is divided into n subsets, use the function to calculate the set and the subsets, and the result from the set and the total result from the n subset are same. E.g., count(), sum(), min(), max().
- **Algebraic** functions: An aggregate function is algebraic if it can be calculated by an algebraic function with M arguments, and each argument is a distributive aggregation function. E.g., ave() = sum() / count(), standard_deviation(), ...
- **Holistic functions**: An aggregate function is holistic if it characterizes a set element (s) relative to other elements of the set without an algebraic calculation. E.g., rank(), median(), ...

DATA CUBE MATERIALIZATION

- **Full cube**: multi-way array aggregation method computes ful data cube by using a multidimensional array as its basic data structure
  - Partition array into the chunks
  - Compute aggregate by visiting
- **Iceberg-cube**: contains cells of the data cube that meet an aggregate condition
- **Closed cube**: consists of only closed cells – cells where no descendant cell has the same measure value
- **Shell cube**: precomputes only portions or fragments of the cube shell, based on cuboids of interest

DATA CUBE COMPUTATION METHODS

- **Multiway Array Aggregation** for Full Cube Computation
- **Bottom-up Computation**(BUC):
- Computing Iceberg Cubes from the Apex Cuboid Downward
- **Star-Cubing**:
- Computing Iceberg Cubes Using a Dynamic Star-Tree Structure
  - Pruning shared dimensions
  - Star-tree construction
  - Aggregation by traversing in a bottom-up fashion

EXERCISES

- Sample Question #1
- Sample Question #2

SAMPLE QUESTION # 1

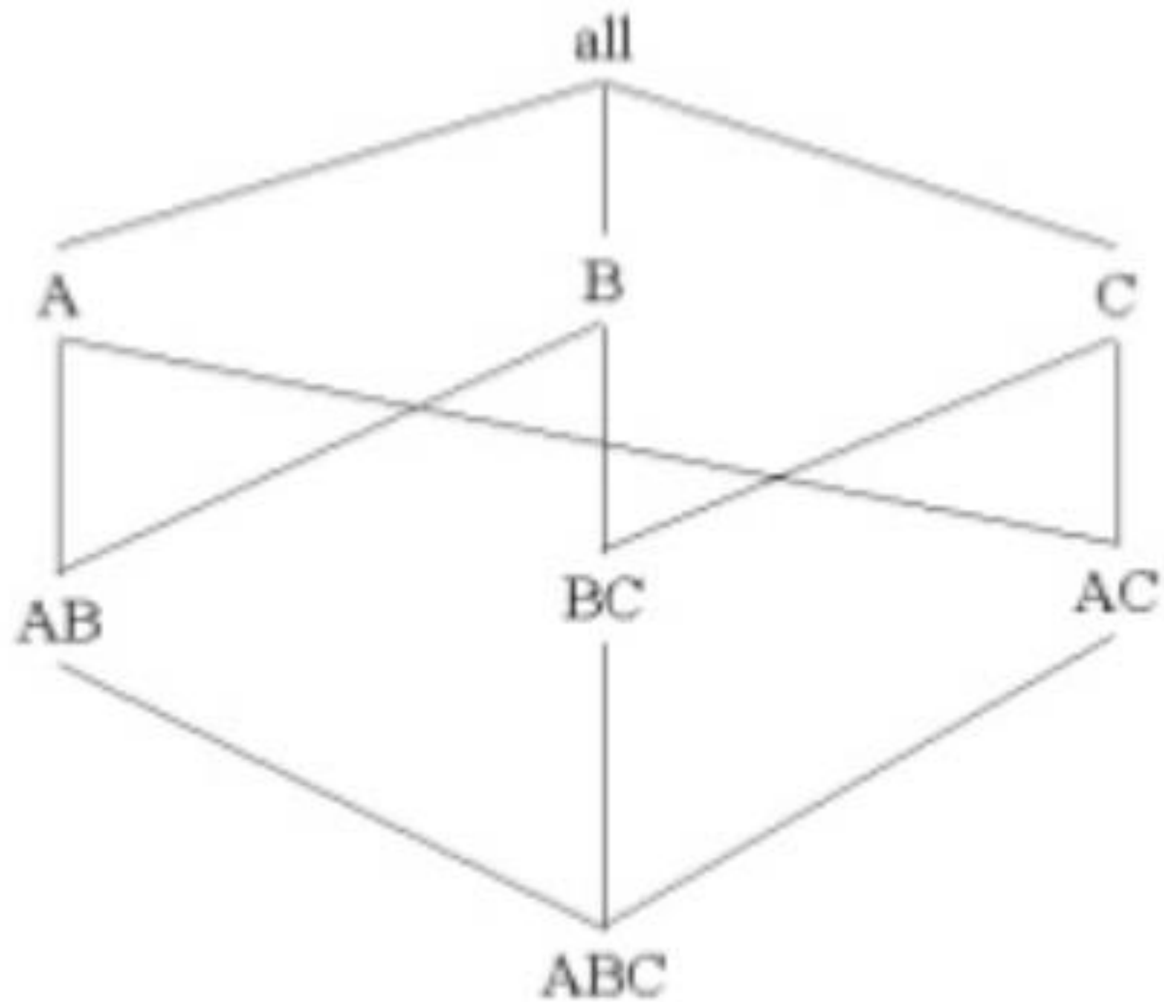
Suppose that a base cuboid has three dimensions, *A* , *B* , *C* , with the following number of cells: | A | = 1, 000, 000, | B | = 100, and | C | = 1000. Suppose that each dimension is evenly partitioned into 10 portions for chunking .

(a) Assuming each dimension has only one level, draw the complete lattice of the cube.

(b) If each cube cell stores one measure with four bytes, what is the total size of the computed cube if the cube is dense ?

REFERENTIAL SOLUTION

(a) Assuming each dimension has only one level, draw the complete lattice of the cube.



REFERENTIAL SOLUTION

(b) If each cube cell stores one measure with four bytes, what is the total size of the computed cube if the cube is dense ?

Answers:

- All:1
- A: 1,000,000; B:100; C: 1000
- AB: 1,000,000 * 100 = 100,000,000; BC: 100 * 1000 = 100,000; AC: 1,000,000 * 1000 = 1,000,000,000
- ABC: 1,000,000 * 100 * 1000 = 100,000,000,000

Total: 101,101,101,101 cells * 4 bytes = 404,404,404,404 bytes

SAMPLE QUESTION #2

Suppose that a data warehouse for *Big University* consists of the four dimensions *student, course, semester , and instructor* , and two measures *count and avg grade* . At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination.

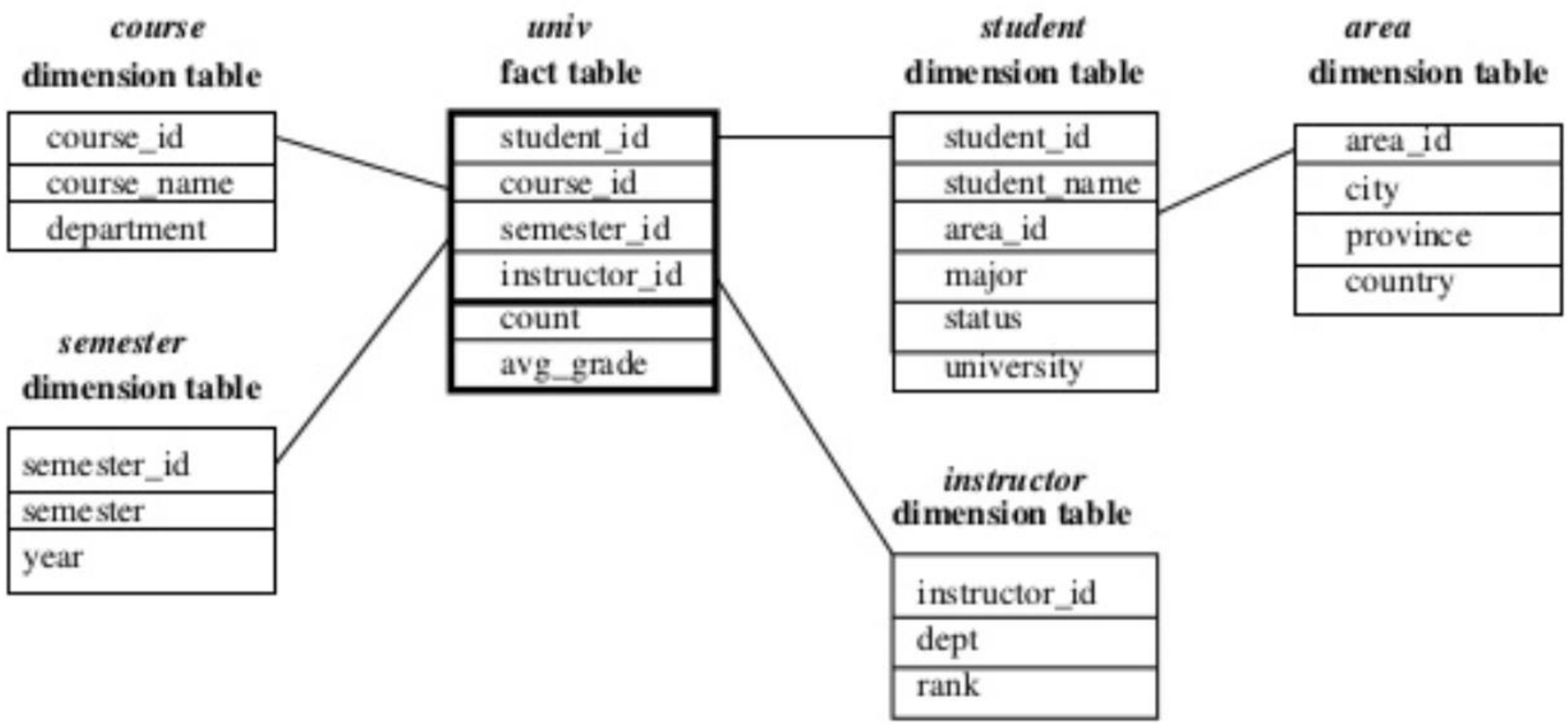
(a) Draw a snowflake schema diagram for the data warehouse.

(b) Starting with the base cuboid [*student , course , semester , instructor*], what specific OLAP operations (e.g., roll-up from *semester* to *year*) should you perform in order to list the average grade of *CS* courses for each *Big University* student.

(c) If each dimension has five levels (including all), such as " *student < major < status < university < all* ", how many cuboids will this cube contain (including the base and apex cuboids)?

REFERENTIAL SOLUTION

- (a) Draw a snowflake schema diagram for the data warehouse.



REFERENTIAL SOLUTION

(b) Starting with the base cuboid [*student* , *course* , *semester* , *instructor*], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of *CS* courses for each *Big University* student.

Answers:

- Roll-up on *course* from *course_id* to *department*
- Roll-up on *student* from *student_id* to *university*
- Dice on *course, student* with *department* = "CS"and *university* = "Big University"
- Drill-down on *student* from *university* to *student_name*

(c) If each dimension has five levels (including all), such as " *student* < *major* < *status* < *university* < *all* ", how many cuboids will this cube contain (including the base and apex cuboids)?

Answers: This cube will contain 5ˆ4 = 625 cuboids.

ASSESSMENT 2 THE DATA WAREHOUSE

- Specifications:
  - weight: 40%
  - Due: Sunday 9 August 11:59PM (Week 5)
  - 2400 word limit, less than 16 pages and in 12pt Arial
    - Each task should be one page, around 300 words
  - Tasks
    1. Business scenario
    2. Information package
    3. Data design
    4. Dimensional modelling
    5. Size of fact table
    6. Aggregating fact table
    7. Lattice of cuboids
    8. Data cube computation

JAMES COOK UNIVERSITY AUSTRALIA

ASSESSMENT 2 THE DATA WAREHOUSE

1. Business scenario

- What is your business about?
- What are the main functions of the business?
- What are the business products, customers and dimensions?
- What are the critical metrics measuring the performance of the business?
- What are the business dimensions along which the metrics are analysed?
- What is the business model to make profits?
- All necessary information to build an information package diagram.

- Your business description should be no more than two A4 pages in 12 pt font size, around 300 words.

Criteria	HD 85-100%
Business Scenario 5%	- Excellent description of business scenario, clearly stating all required information to build an information package diagram, including metrics/measures for fact tables, dimensions and their corresponding attributes etc.

ASSESSMENT 2 THE DATA WAREHOUSE

2. Information package diagram

- ➤ What are metrics and measures that your business is interested in analysing?
- ➤ What are business dimensions that are related to the metrics and measures?
- ➤ What are hierarchies and categories for dimensions?

Information Subject: Automaker Sales

Dimensions

Hierarchies/Categories

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

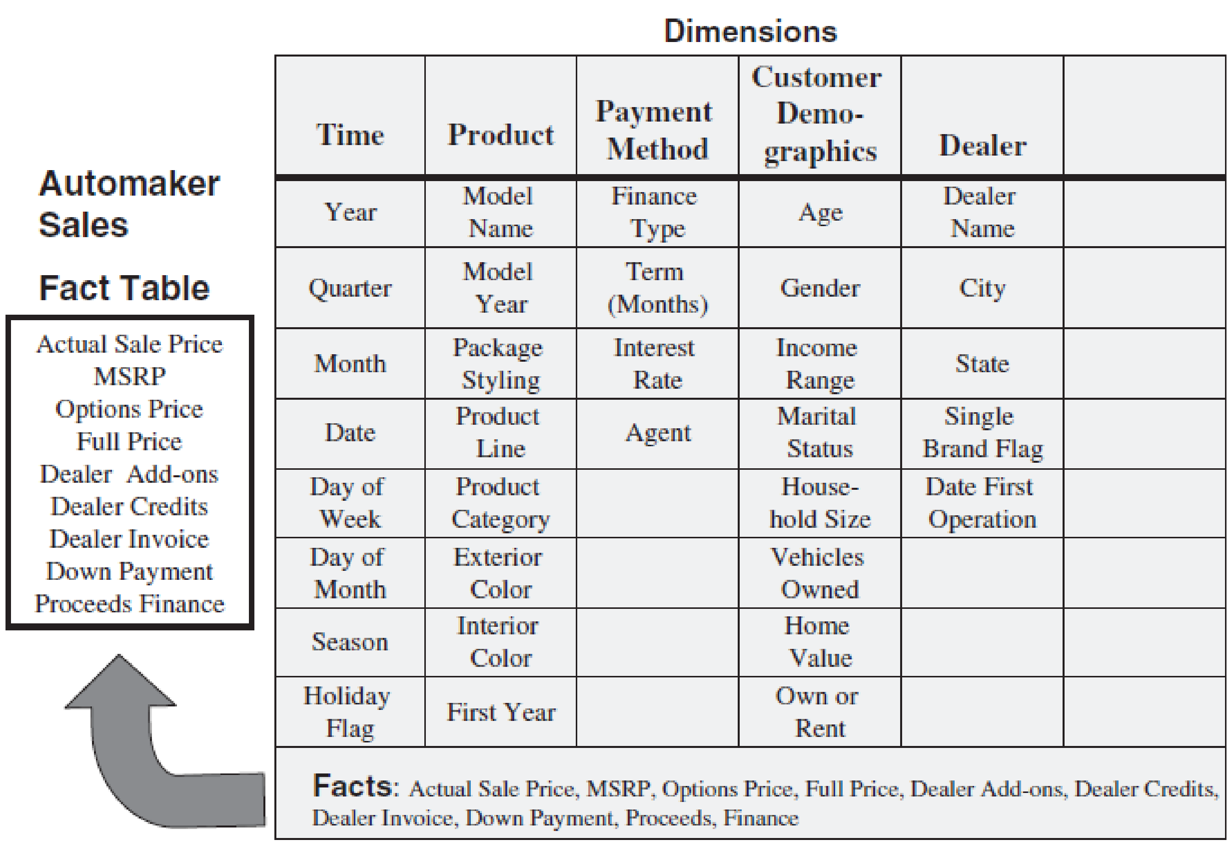
Figure 5-5 Information package: automaker sales.

Information Package Diagram 5%	- Excellent identification of metrics and measures - Excellent identification of dimensions - Excellent identification of hierarchies/categories for each dimension - Excellent presentation of diagram

ASSESSMENT 2 THE DATA WAREHOUSE

3. Data design

- ➤ Build dimension tables from your business scenario
- ➤ The number of dimension tables should be between 6-8
- ➤ The number of metrics and measures should be between 4-5
- ➤ Identify attributes for each business dimension
- ➤ The number of attributes for each business dimension should be between 3-5.



Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Figure 10-2 Formation of the automaker sales fact table.

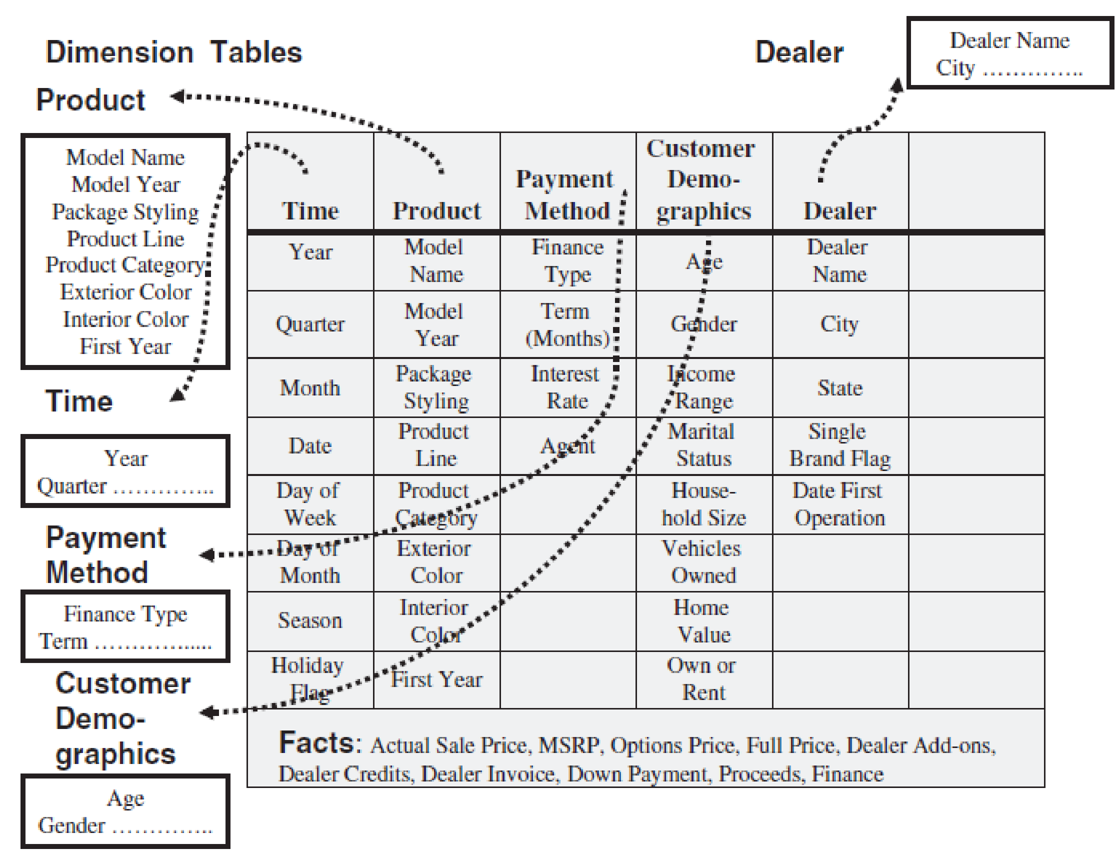


Figure 10-3 Formation of the automaker dimension tables.

ASSESSMENT 2 THE DATA WAREHOUSE

4. Dimensional modelling

- Build a star/snowflake schema for your business scenario
- Present the schema in a diagram to illustrate the dimensional modelling

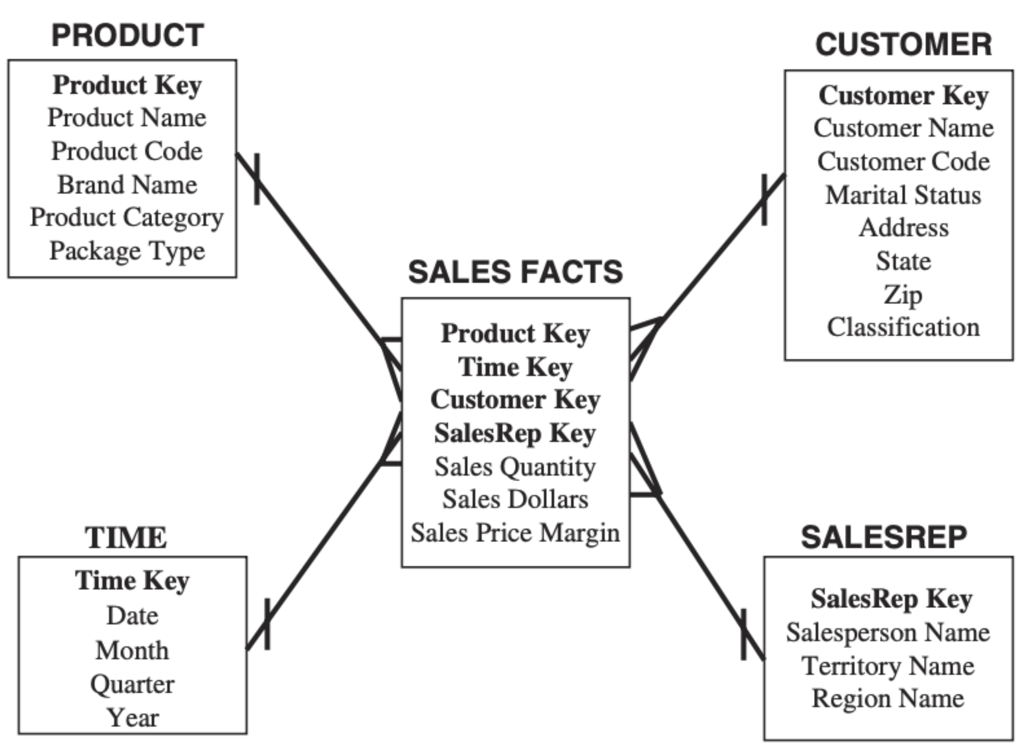


Figure 11-7 Sales: a simple STAR schema.

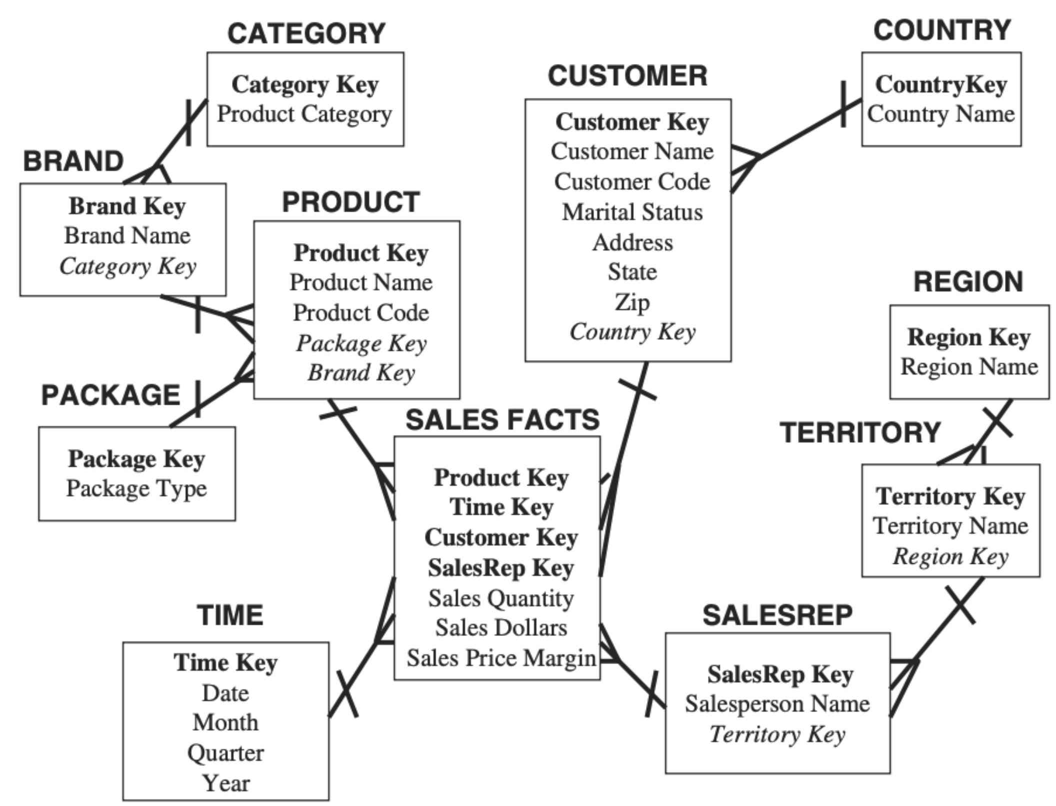


Figure 11-9 Sales: the "snowflake" schema.