Data Warehousing and Data Mining

— L2: Data Warehousing and OLAP —

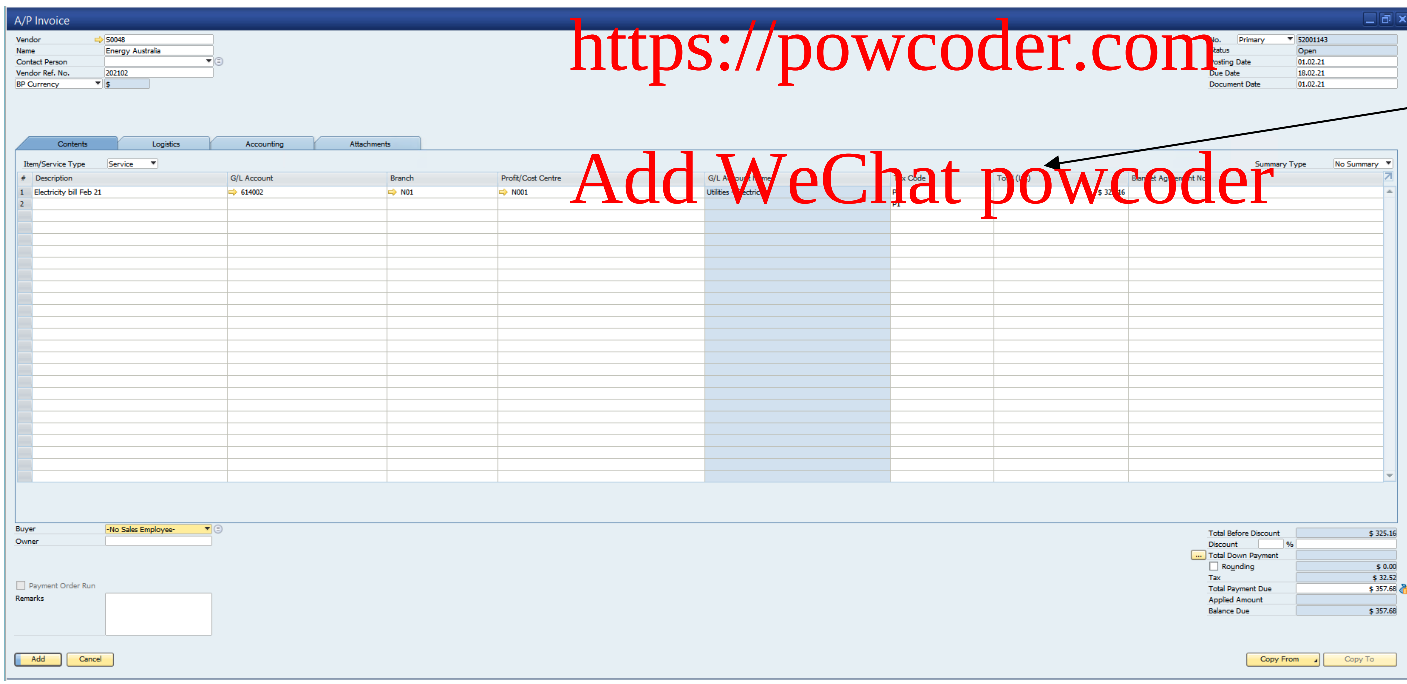
Part I

- Why and What are Data Warehouses?
  - Transaction Processing vs. Analytical Processing
  - Databases vs. Data Warehouses

Data is meaningless without analysis!

Example in a finance department

- Daily transaction tasks
  - E.g., account receivable, account payable, payroll, etc.

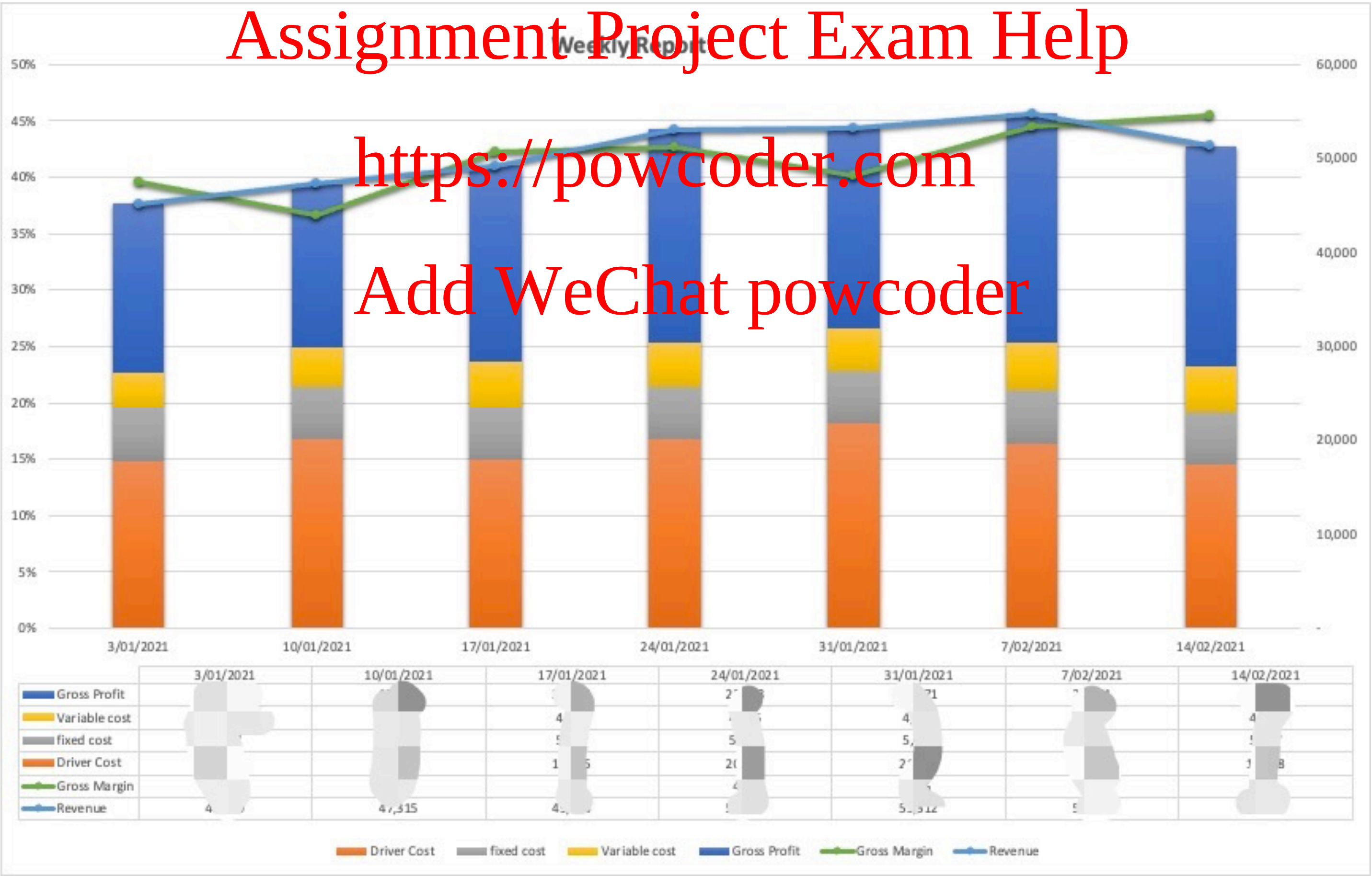


Columns:

Description
G/L Account
Branch
cost center
G/L account name
Tax code
Total
…

Example/2

- Weekly…monthly…yearly analytical tasks
  - E.g., Finance reports



Why OLAP Servers?

- Different workload:
  - OLTP (on-line transaction processing)
    - Major task of traditional relational DBMS
    - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
  - OLAP (on-line analytical processing)
    - Major task of data warehouse system
    - Data analysis and decision making
- Queries hard/infeasible for OLTP, e.g.,
  - Which **week** we have the largest sales?
  - Does the sales of **dairy products** increase over time?
  - Generate a **spread sheet** of total sales by state and by year.
- Difficult to represent these queries by using SQL ← Why?

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	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

Data Analysis Problems

- The same data found in many different systems
  - Example: customer data across different departments
  - The same concept is defined differently
- Heterogeneous sources
  - Relational DBMS, OnLine Transaction Processing (OLTP)
  - Unstructured data in files (e.g., MS Excel) and documents (e.g., MS Word)

Data Analysis Problems (Cont'd)

- Data is suited for operational systems
  - Accounting,billing,etc.
  - Do not support analysis across business functions
- Data quality is bad
  - Missing data, imprecise data, different use of systems
- Data are "volatile"
  - Data deleted in operational systems (6months)
  - Data change over time – no historical information

Solution: Data Warehouse

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

Data Warehouse—Subject-Oriented

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

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

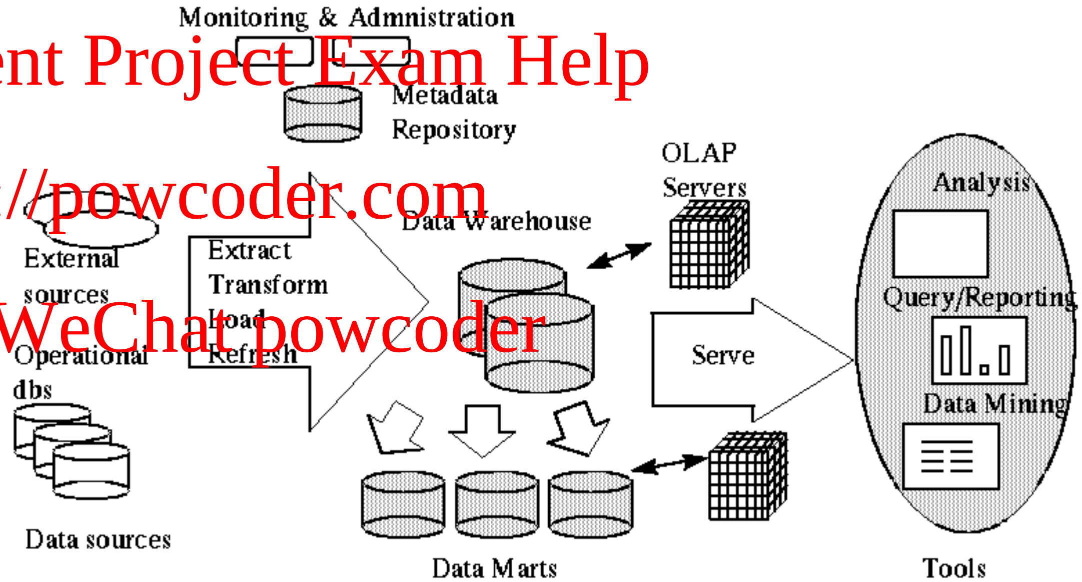
- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".

Data Warehouse—Non-Volatile

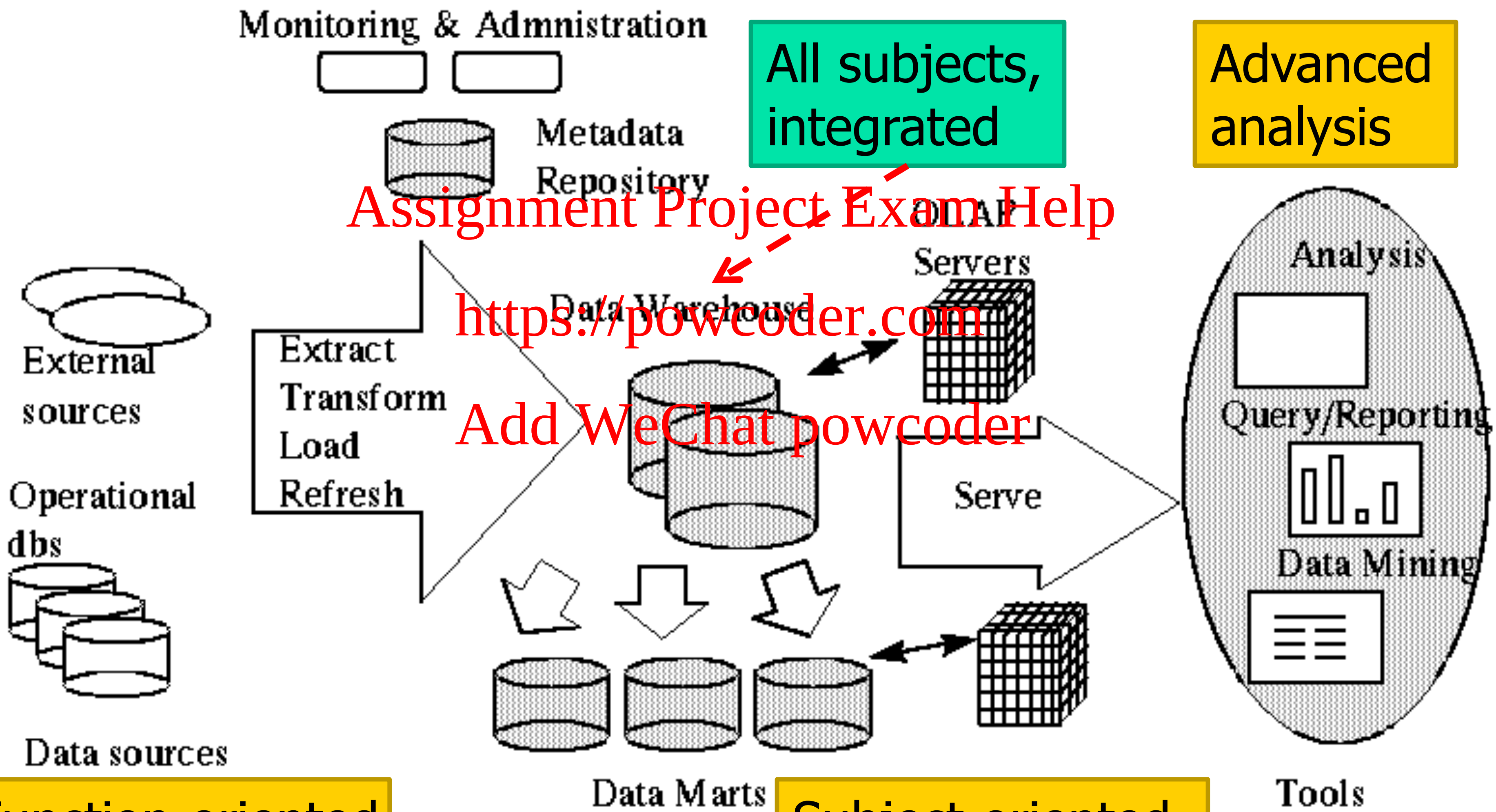
1. A physically separate store of data transformed from the operational environment.
2. Operational update of data does not occur in the data warehouse environment.
   - Does not require transaction processing, recovery, and concurrency control mechanisms
   - Requires only two operations in data accessing: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

Data Warehouse Architecture

- Extract data from operational data sources
  - clean, transform
- Bulk load/refresh
  - warehouse is offline
- OLAP-server provides multidimensional view
- Multidimensional-olap (Essbase, oracle express)
- Relational-olap (Redbrick, Informix, Sybase, SQL server)



Data Warehouse Architecture



All subjects, integrated

Advanced analysis

Function-oriented systems

Subject-oriented systems

Why Separate Data Warehouse?

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

Comparisons

	Databases	Data Warehouses
Purpose	Many purposes; Flexible and general	One purpose: Data analysis
Conceptual Model	ER	Multidimensional
Logical Model	(Normalized) Relational Model	(Denormalized) Star schema / Data cube/cuboids
Physical Model	Relational Tables	ROLAP: Relational tables MOLAP: Multidimensional arrays
Query Language	SQL (hard for analytical queries)	MDX (easier for analytical queries)
Query Processing	B+-tree/hash indexes, Multiple join optimization, Materialized views	Bitmap/Join indexes, Star join, Materialized data cube

Comparisons/2

AUTHORS				
Id	FristNmae	LastName	DateOfBirth	Gender
1	Yumeng	Wang	1967-09-29	F
2	Michael	Joris	1990-12-11	M
3	Anthony	Green	1987-12-10	M
4	Kevin	Davis	1976-05-23	M
5	Lee	Wongjun	1962-04-24	F

AUTHOR_BOOK_MAP	
AuthorId	BookId
1	11231
1	22131
3	29384
2	29384
4	37849
5	33423
1	33456
2	47638
1	48983
2	48983
3	48983
4	48983
5	48983
3	51289
5	52839

BOOKS				
Id	Titile	CopyRight	ISBN	Genre
11231	The Light of Other Days	2000	0-812-12321311231	1
22131	Death in Town	2019	0-123-374827603	1
29384	The C Programming Language	1999	0-231-1231314	3
37849	To Find You	2003	0-812-12345677	2
33423	We Are Warriors	2009	1230-12-675675	1
33456	The Mountain	2008	0-812342-56335	2
47638	Meet You Before Dawn	2011	0-812-23423563	2
48983	Java Complete	1999	0-812-23453634	3
51289	Darkness Of Midages	2020	0-23423-374827603	1
52839	She	2011	2342-2342-623	2

PUBLISHERS	
Id	Name
1	Sasquatch Books
2	Peanut Butter Publishing
3	Chatwin Books

PUBLISHER_BOOK_MAP	
PublisherId	BookId
1	11231
1	22131
1	29384
1	37849
2	33423
2	33456
2	47638
2	48983
3	51289
3	52839

GENRE	
Id	Genre
1	Science Fiction
2	Love & Romance
3	Education

Source：https://www.zhihu.com/question/20623931

Comparisons/2

BOOKS									
Id	Titile	CopyRight	ISBN	Genre	AuthorFristNmae	AuthorLastName	DateOfBirth	Gender	PublisherName
11231	The Light of Other Days	2000	0-812-12321311231	Science Fiction	Yumeng	Wang	1967-09-29	F	Sasquatch Books
22131	Death in Town	2019	0-123-374827603	Science Fiction	Yumeng	Wang	1967-09-29	F	Sasquatch Books
29384	The C Programming Language	1999	0-231-1231314	Education	Michael	Joris	1990-12-11	M	Sasquatch Books
29384	The C Programming Language	1999	0-231-1231314	Education	Anthony	Green	1987-12-10	M	Sasquatch Books
37849	To Find You	2003	0-812-1234567[illegible]	Love & Romance	Kevin	Davis	1976-05-23	M	Sasquatch Books
33423	We Are Warriors	2009	1230-12-675675	Science Fiction	Lee	Wongjun	1962-04-24	F	Peanut Butter Publishing
33456	The Mountain	2008	0-812342-56835	Love & Romance	Yumeng	Wang	1967-09-29	F	Peanut Butter Publishing
47638	Meet You Before Dawn	2011	0-812-23423563	Love & Romance	Michael	Joris	1990-12-11	M	Peanut Butter Publishing
48983	Java Complete	1999	0-812-23453634	Education	Yumeng	Wang	1967-09-29	F	Peanut Butter Publishing
48983	Java Complete	1999	0-812-23453634	Education	Michael	Joris	1990-12-11	M	Peanut Butter Publishing
48983	Java Complete	1999	0-812-23453634	Education	Anthony	Green	1987-12-10	M	Peanut Butter Publishing
48983	Java Complete	1999	0-812-23453634	Education	Kevin	Davis	1976-05-23	M	Peanut Butter Publishing
48983	Java Complete	1999	0-812-23453634	Education	Lee	Wongjun	1962-04-24	F	Peanut Butter Publishing
51289	Darkness Of Midages	2020	0-23423-374827603	Science Fiction	Anthony	Green	1987-12-10	M	Chatwin Books
52839	She	2011	2342-2342-623	Love & Romance	Lee	Wongjun	1962-04-24	F	Chatwin Books

Source：https://www.zhihu.com/question/20623931

- The Multidimensional Model

The Multidimensional Model

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube, which is a multidimensional generalization of 2D spread sheet.
- Key concepts:

  - **Facts**: the subject it models

    - Typically transactions in this course; other types includes snapshots, etc.

    - Measures: numbers that can be aggregated
    - Dimensions: context of the measure
  - Hierarchies:
    - Provide contexts of different granularities (aka. grains)
- Goals for dimensional modeling:
  - Surround facts with as much relevant context (dimensions) as possible ← Why?

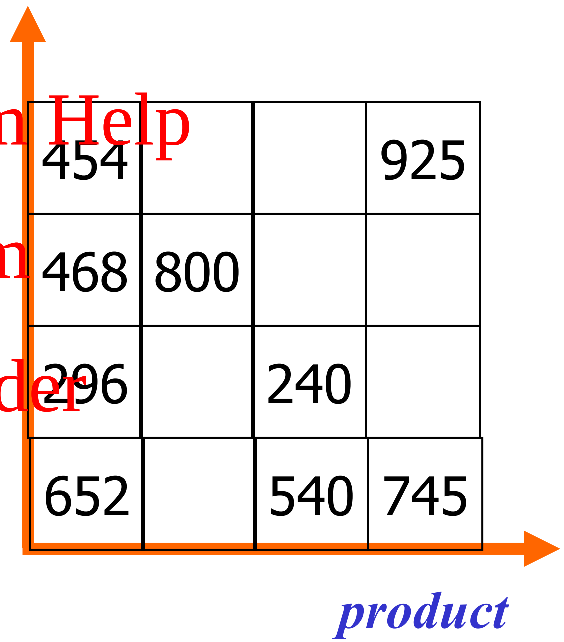
Supermarket Example

- Subject: analyze total sales and profits
- Fact: Each Sales **Transaction**
  - Measure: Dollars_Sold, Amount_Sold, Cost
  - Calculated Measure: Profit
- Dimensions:
  - Store
  - Product
  - Time

Visualizing the Cubes

- A valid instance of the model is a data cube

total Sales		product			
		p1	p2	p3	p4
city	NY	$454	-	-	$925
	LA	$468	$800	-	-
	SD	$296	-	$240	-
	SF	$652	-	$540	$745



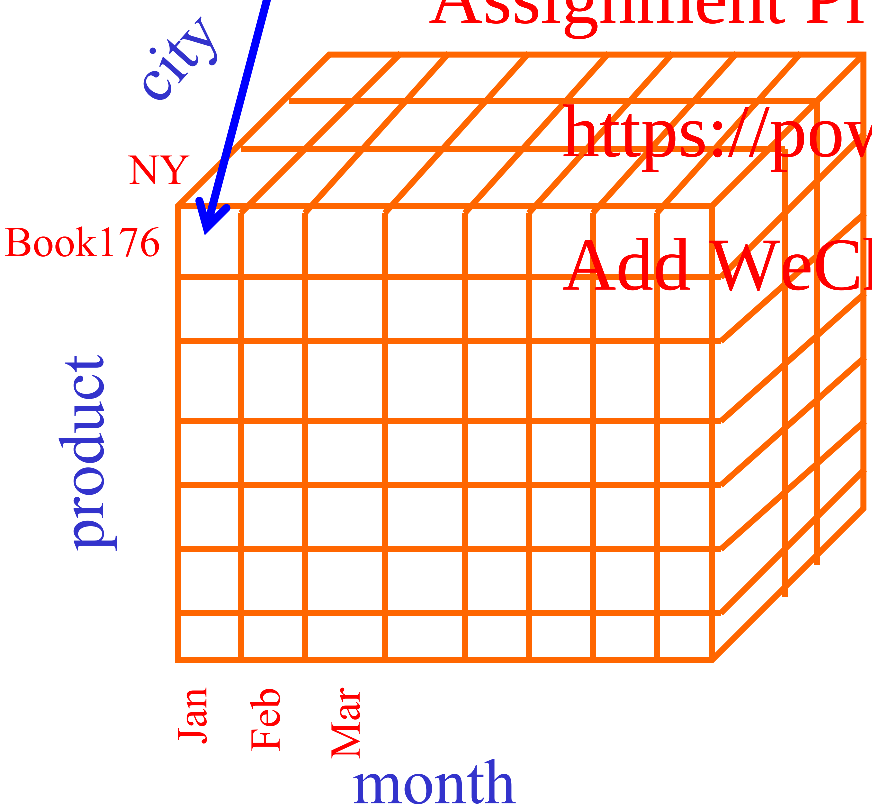
Concepts: cell, fact (=non-empty cell), measure, dimensions

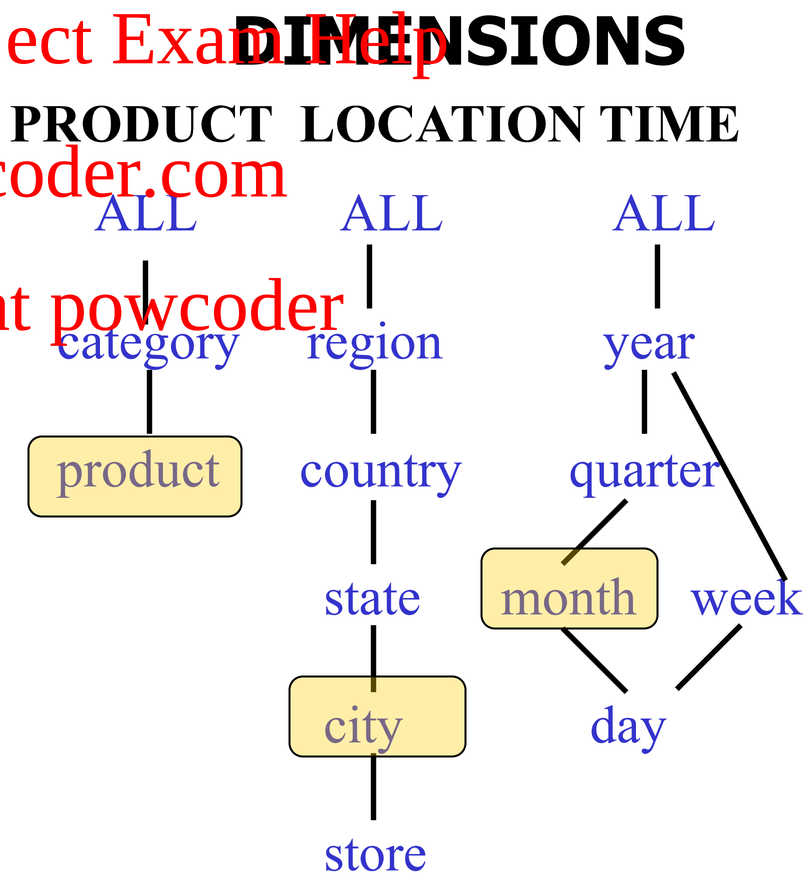
Q: How to generalize it to 3D?

3D Cube and Hierarchies

Concepts: hierarchy (a tree of dimension values), level

Sales of book176 in NY in Jan can be found in this cell

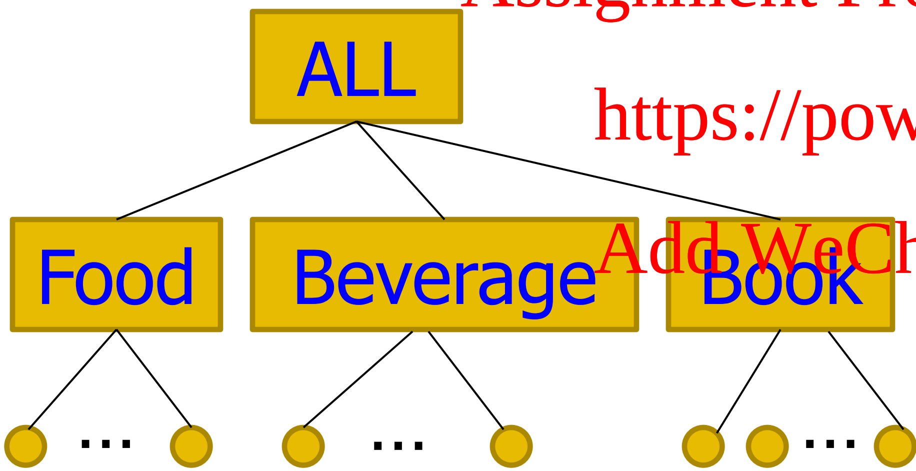


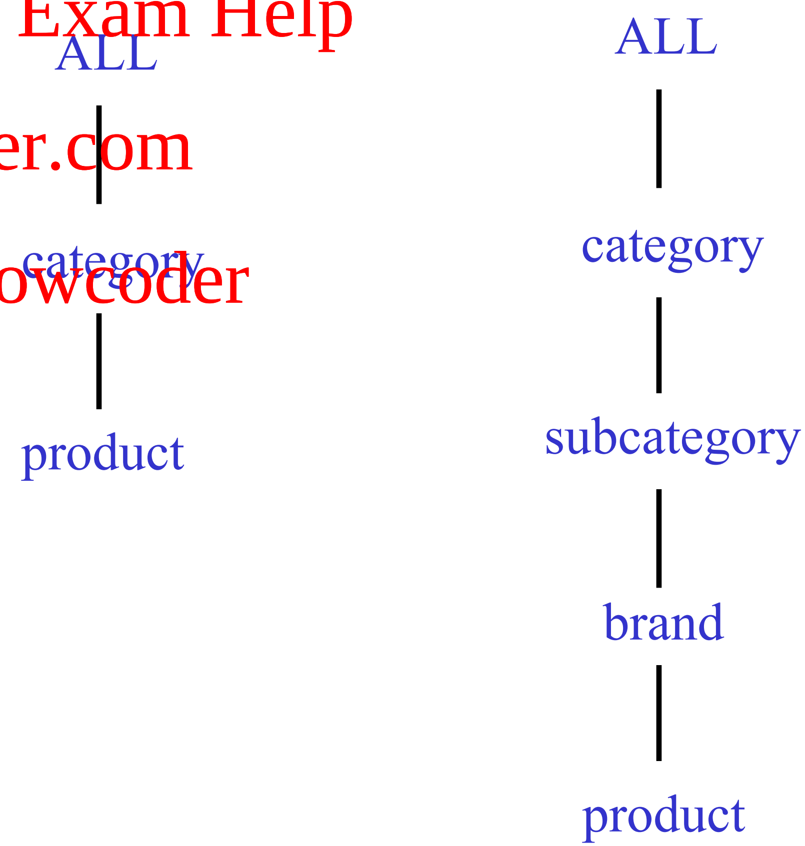


Hierarchies

Concepts: hierarchy (a tree of dimension values), level

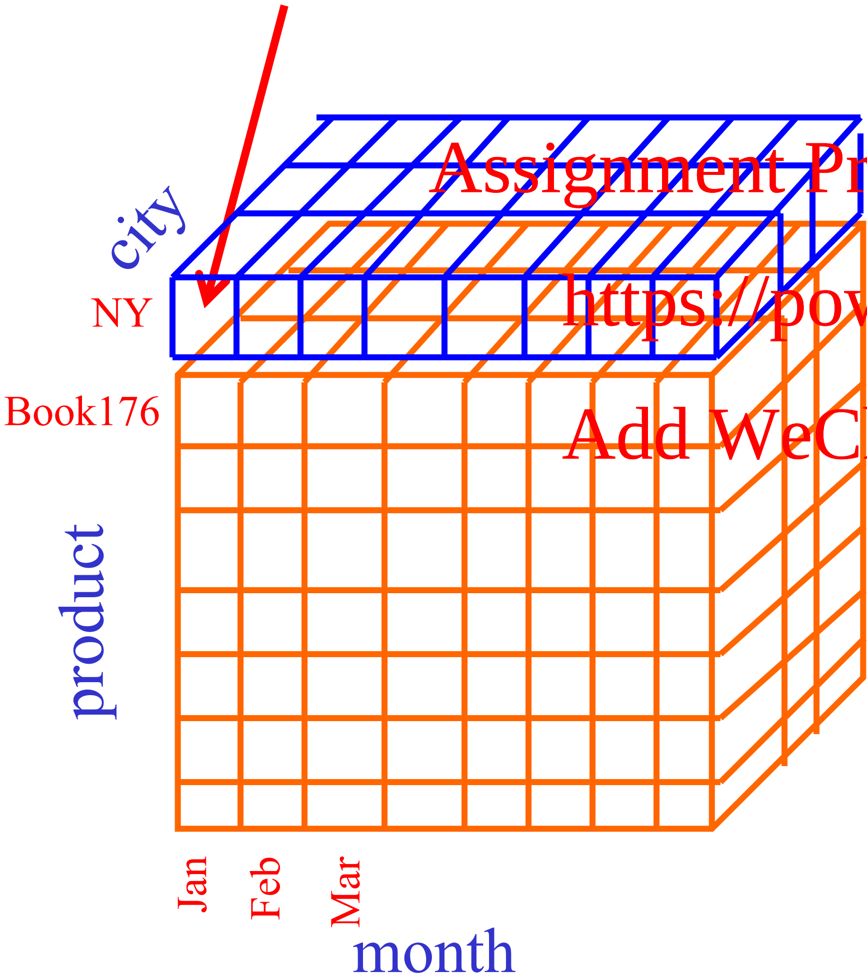
Which design is better? Why?

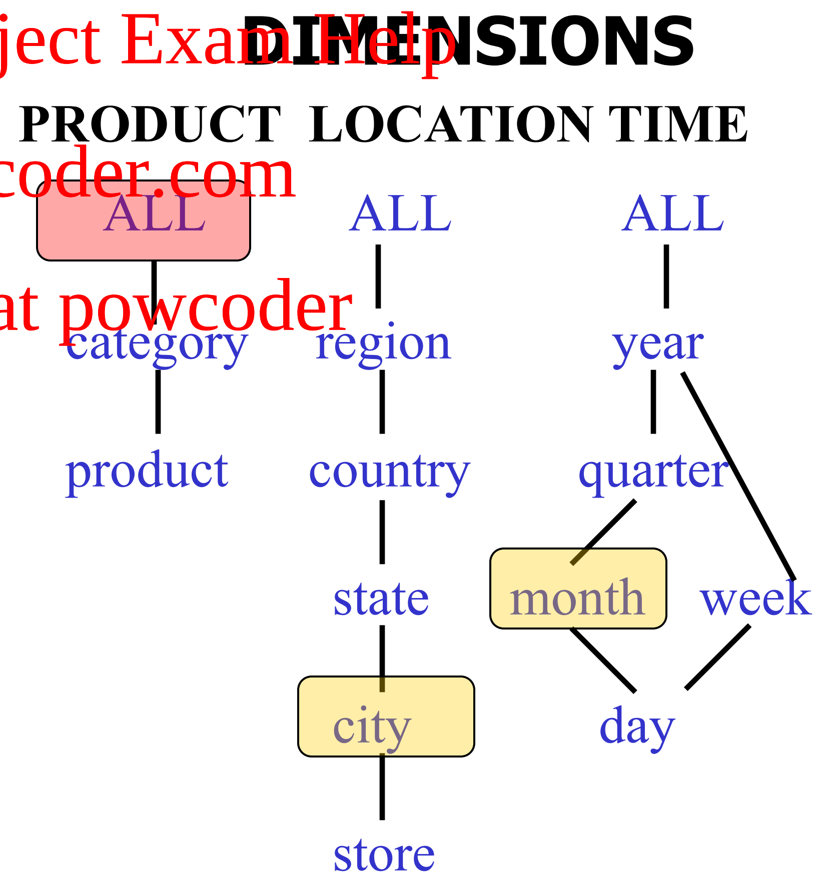




The (city, moth) Cuboid

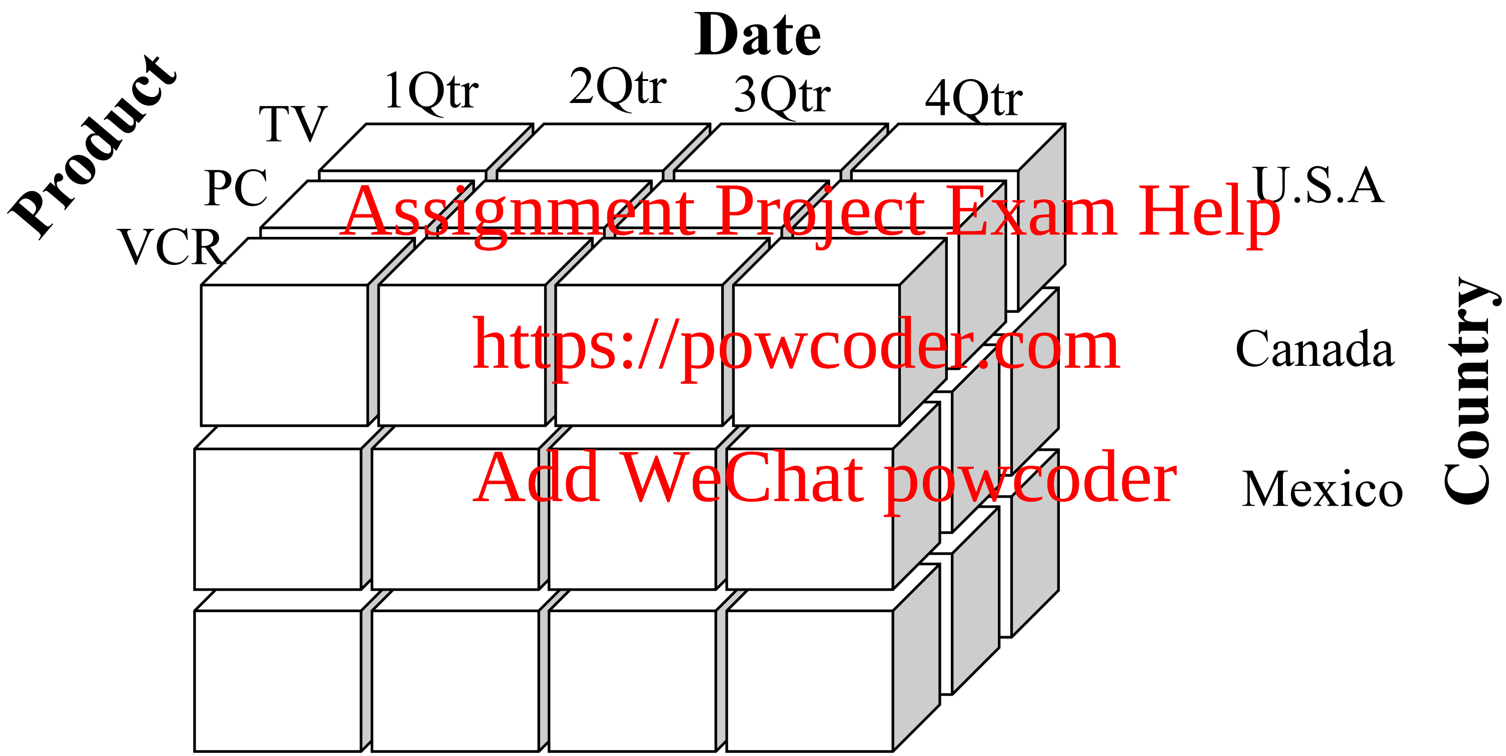
Sales of ALL_PROD in NY in Jan





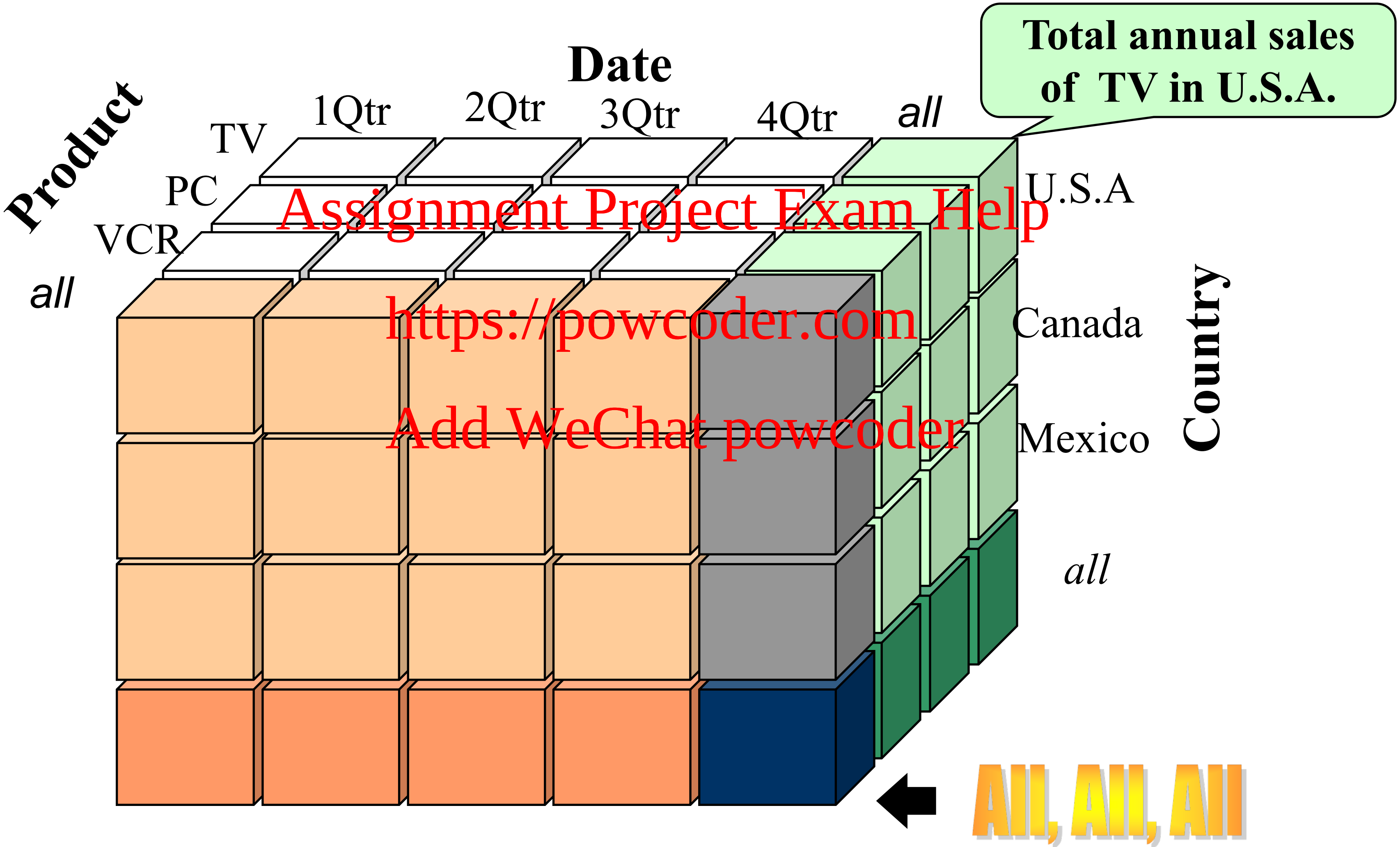
All the Cuboids

Assume: no other non-ALL levels on all dimensions.

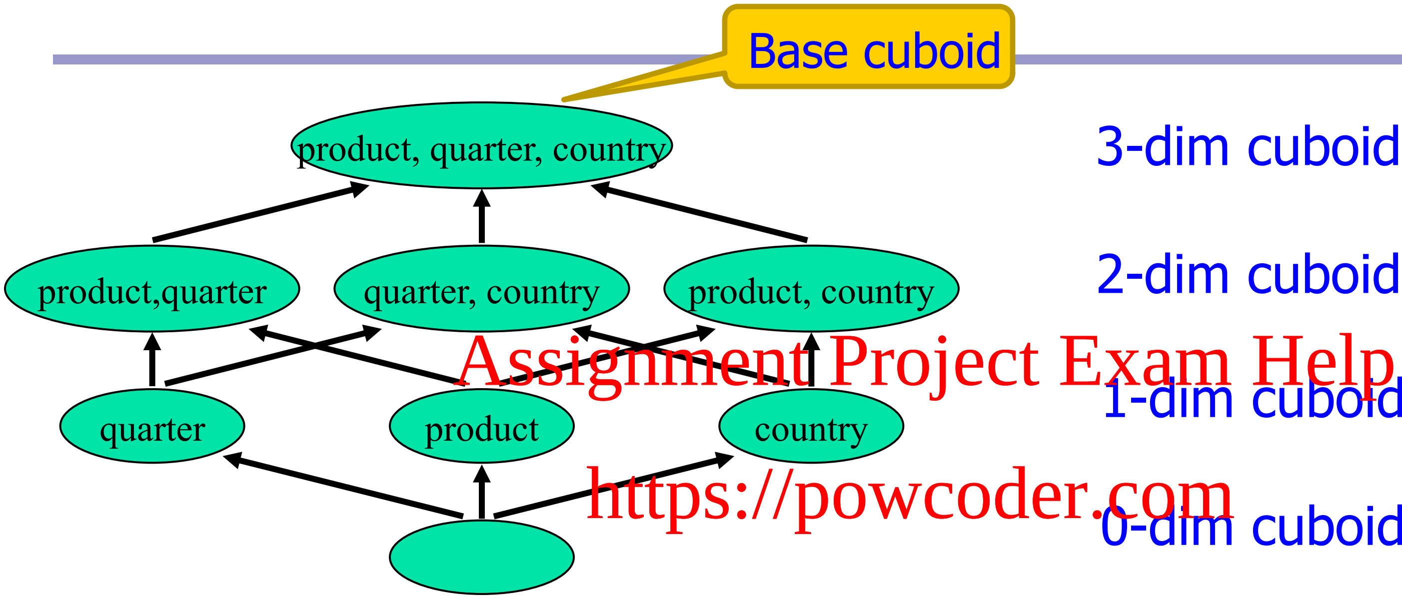


All the Cuboids /2

Assume: no other non-ALL levels on all dimensions.



Lattice of the cuboids



- n-dim cube can be represented as $(D_1, D_2, \dots, D_d)$, where D_i is the set of allowed values on the i-th dimension.
  - if $D_i = L_i$ (a particular level), then Di = all descendant dimension values of L_i.
  - ALL can be omitted and hence reduces the effective dimensionality
- A complete cube of d-dimensions consists of $\prod_{i=1}^{d}(n_i + 1)$ cuboids, where n_i is the number of levels (excluding ALL) on i-th dimension.
  - They collectively form a lattice.

Properties of Operations

- All operations are closed under the multidimensional model
  - i.e., both input and output of an operation is a cube
- So that they can be composed

Q: What's the analogy in the Relational Model?