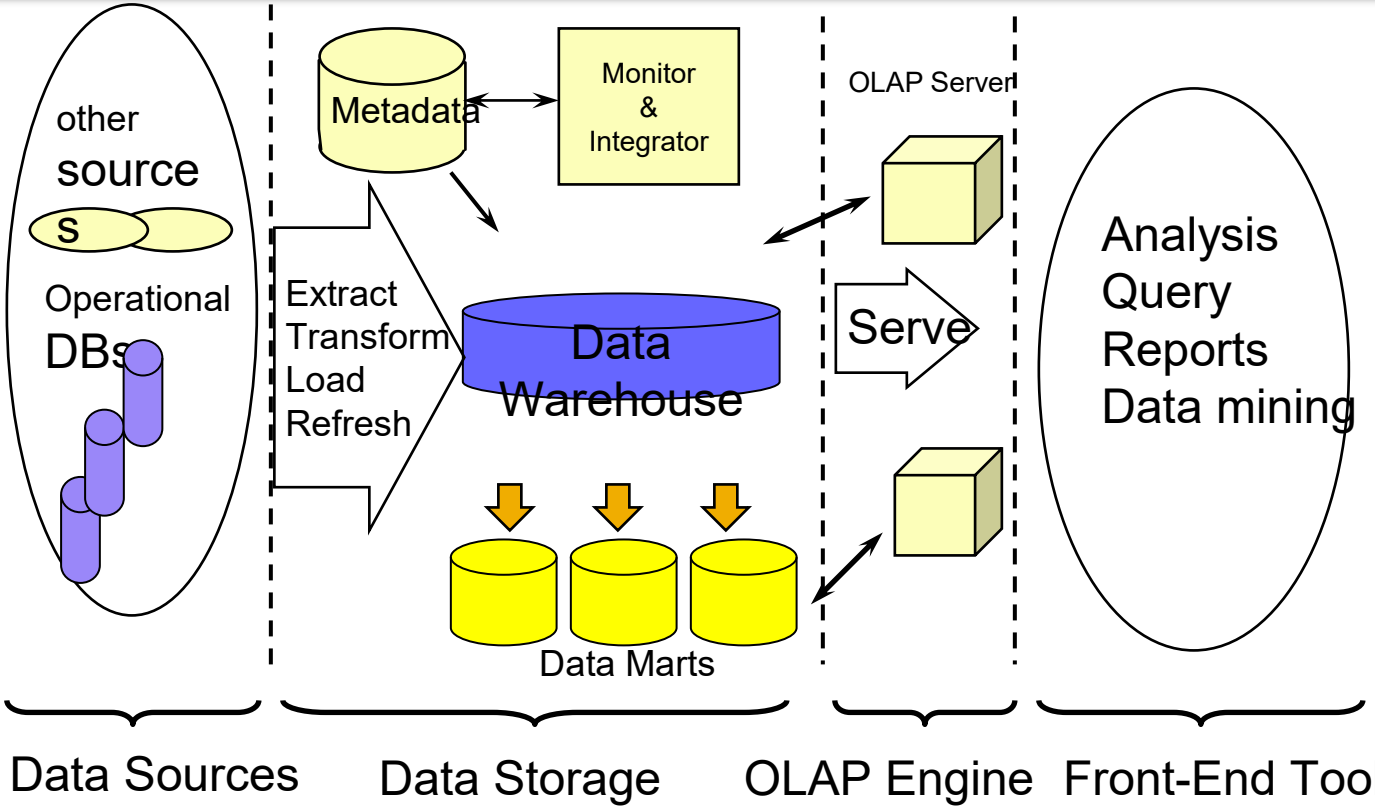
Data Warehouse

CE384: Database Design
Maryam Ramezani
Sharif University of Technology
maryam.ramezani@sharif.edu

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning
  - Bottom-up: Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

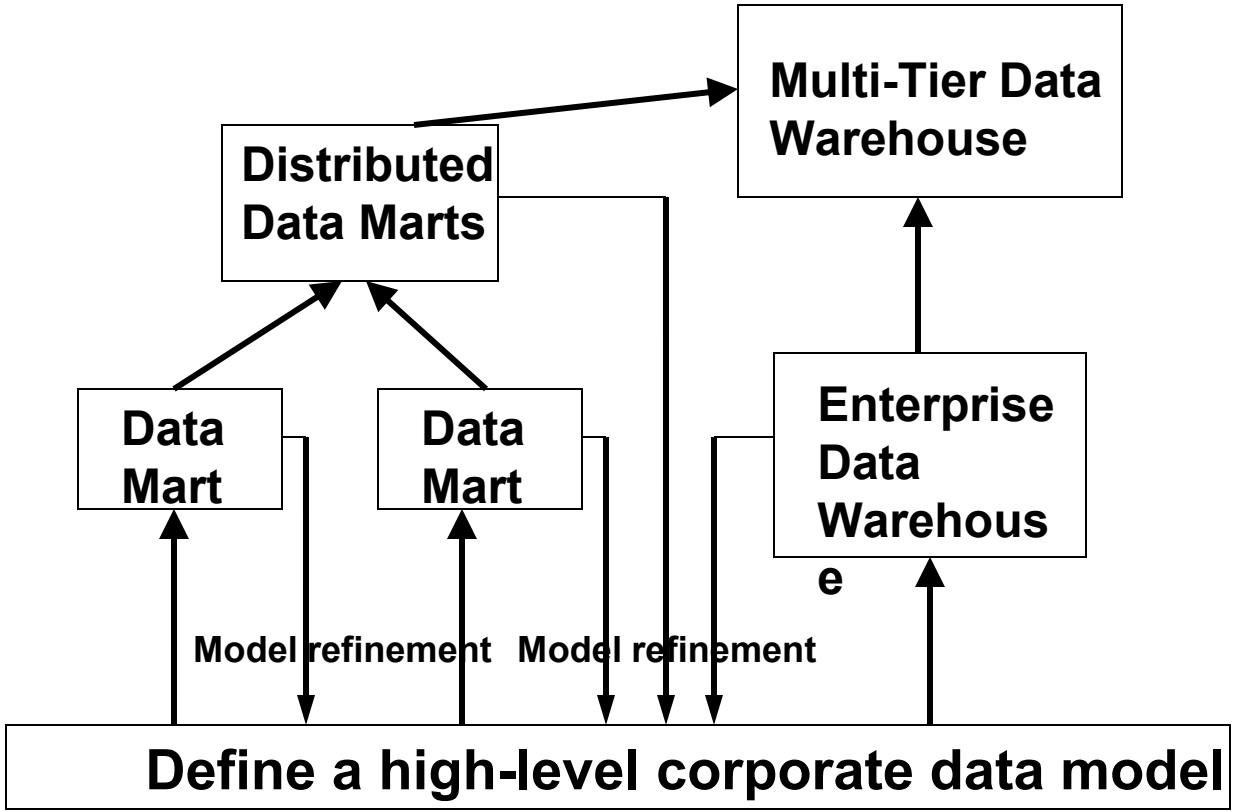
Multi-Tiered Architecture



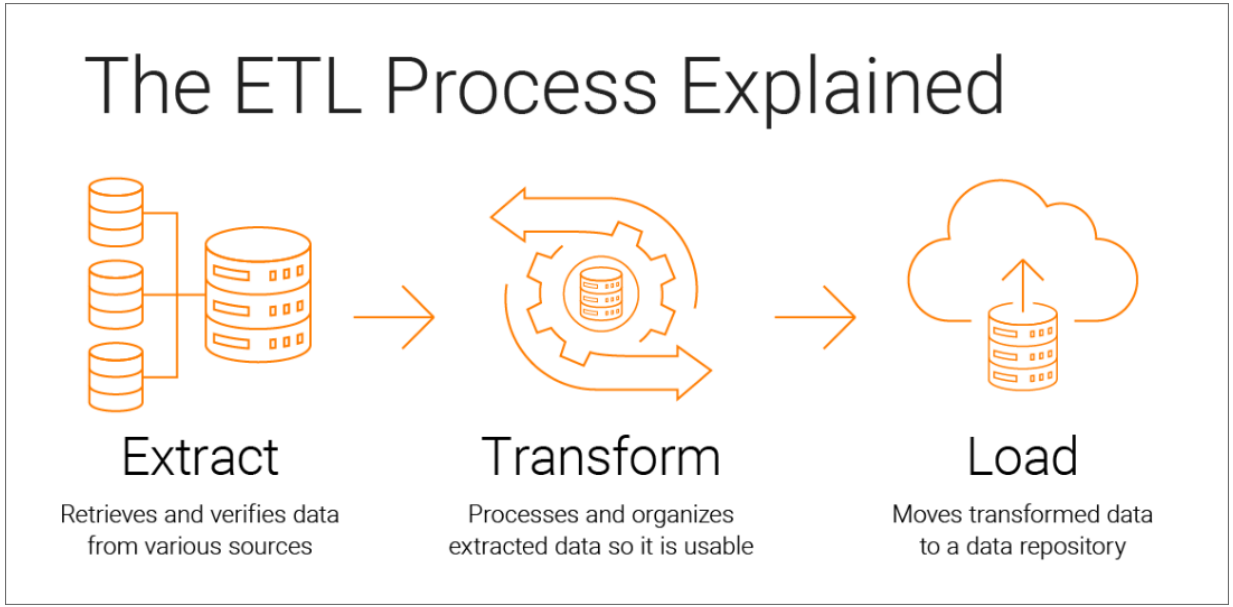
Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Data Warehouse Development: A Recommended Approach



ETL


The ETL Process Explained
Extract
Retrieves and verifies data from various sources
Transform
Processes and organizes extracted data so it is usable
Load
Moves transformed data to a data repository

Change Data Capture (CDC)

Different Change Data Capture Methods

- Time-based Change Data Capture
  - 'LAST_MODIFIED.'
- Log-based Change Data Capture
- Trigger-based Change Data Capture
- Push and Pull Approaches
  - In the push approach, all processes occur on the source dataset that trigger notifications for changes (insertions, edits, deletions) in real time.
  - In the pull method, the CDC system actively pulls queries or changes from the source system at scheduled intervals. This puts less load on the source database. Just like the push approach, the pull method also requires an intermediary messenger for offline target systems.

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

Warehouse Models & Operators

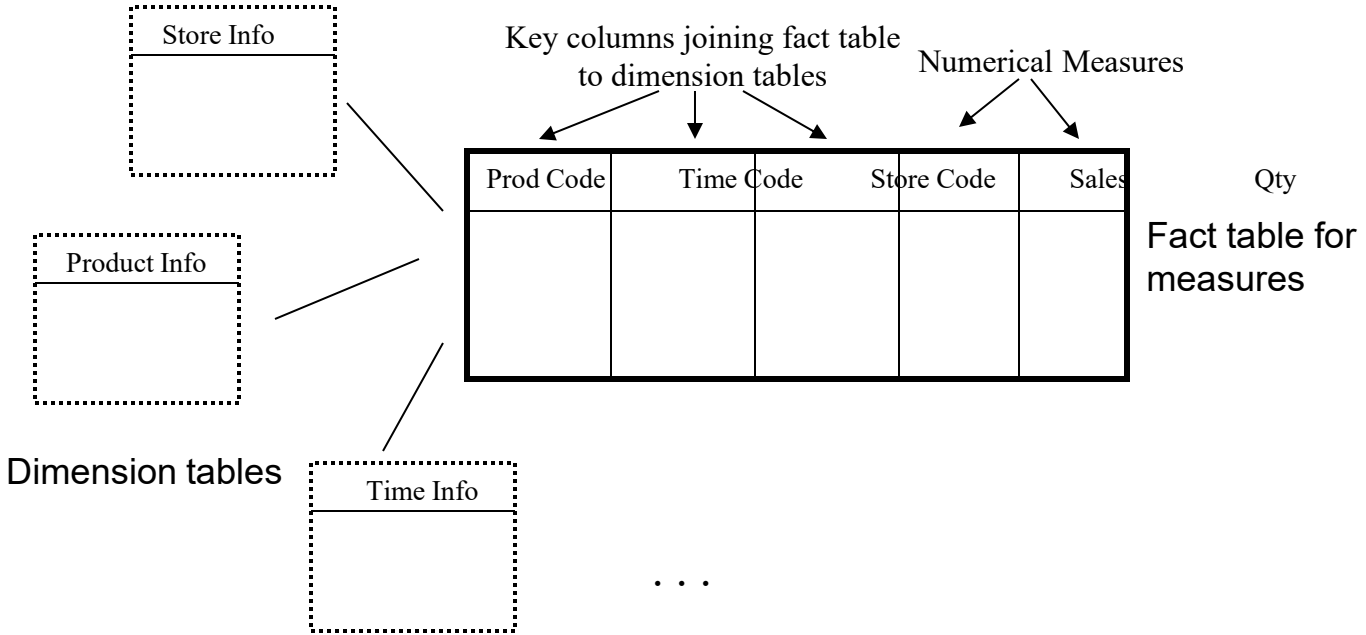
- Data Models
  - relations
  - stars & snowflakes
  - cubes
- Operators
  - slice & dice
  - roll-up, drill down
  - pivoting
  - other

Multi-Dimensional Data

- Measures - numerical (and additive) data being tracked in business, can be analyzed and examined
- Dimensions - business parameters that define a transaction, relatively static data such as lookup or reference tables
- Example: Analyst may want to view ***sales*** data (measure) by *geography*, by *time*, and by *product* (dimensions)

The Multi-Dimensional Model

"Sales by product line over the past six months"
"Sales by store between 1990 and 1995"

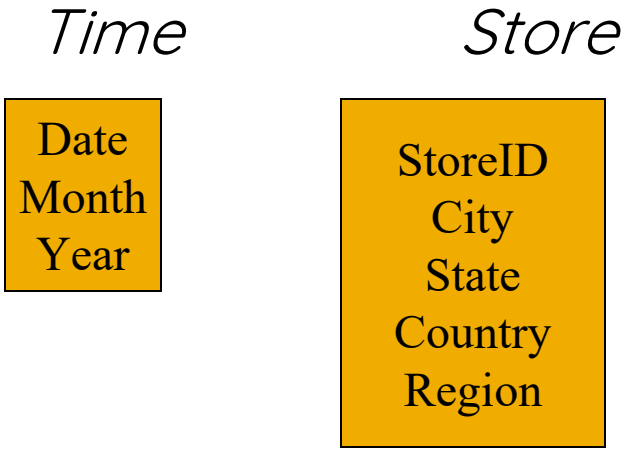


Multidimensional Modeling

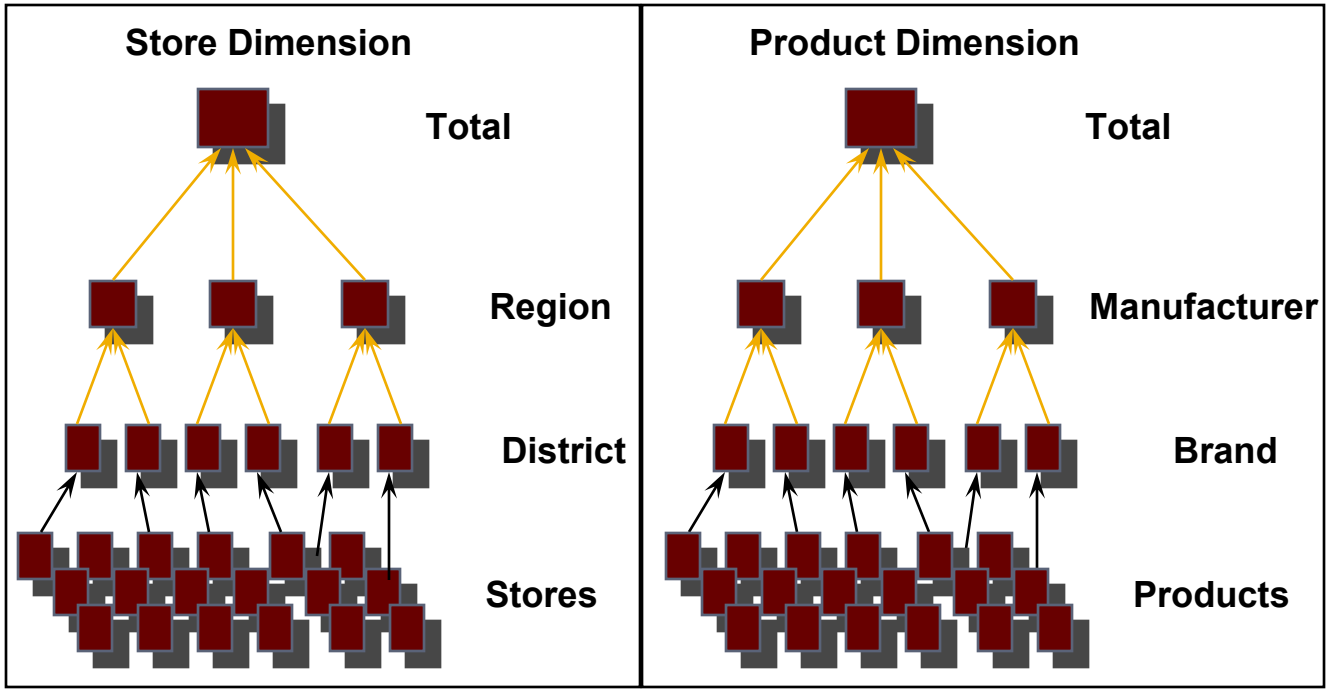
- Multidimensional modeling is a technique for structuring data around the business concepts
- ER models describe “entities” and “relationships”
- Multidimensional models describe “measures” and “dimensions”

Dimensional Modeling

- Dimensions are organized into hierarchies
  - E.g., Time dimension: days → weeks → quarters
  - E.g., Product dimension: product → product line → brand
- Dimensions have attributes



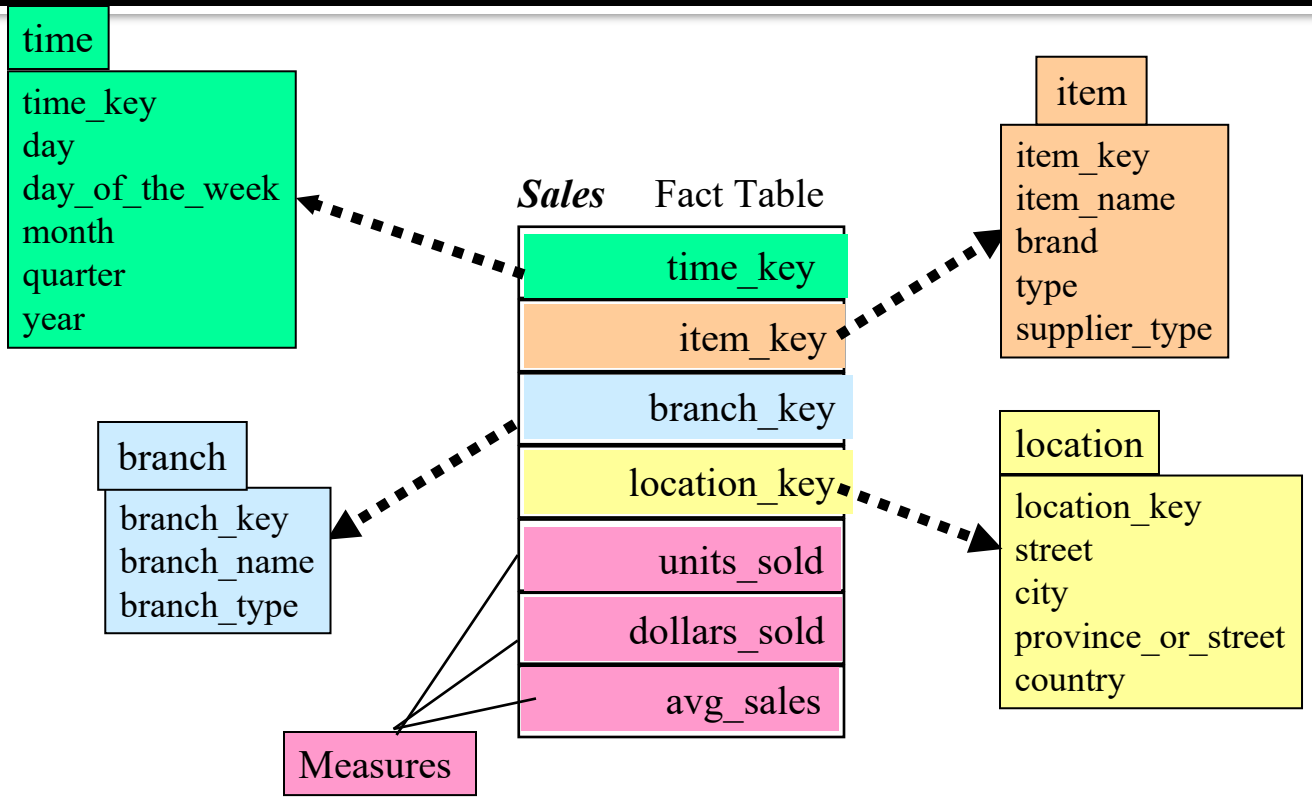
Dimension Hierarchies



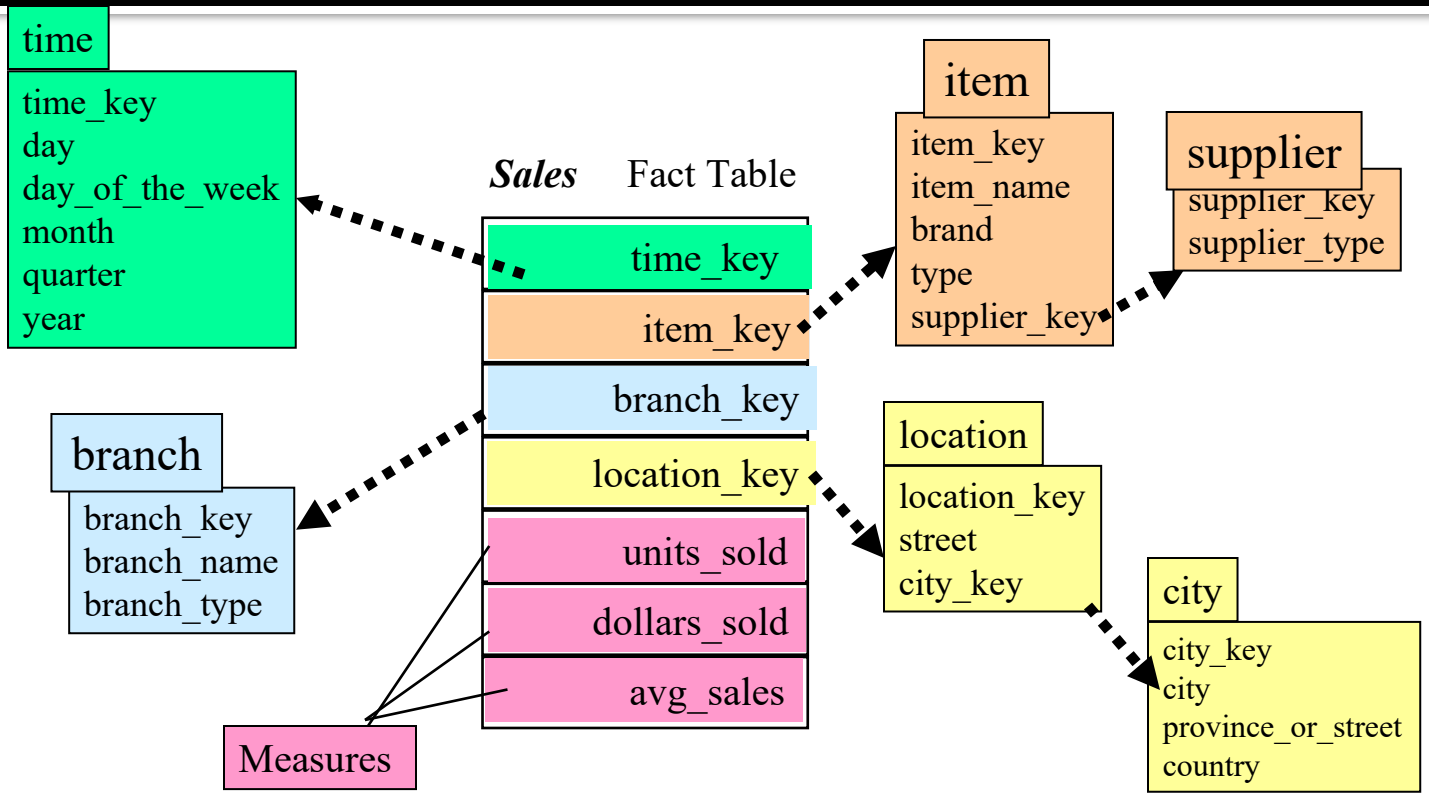
Schema Design

- Most data warehouses use a star schema to represent multi-dimensional model.
- Each dimension is represented by a **dimension table** that describes it.
- A **fact table** connects to all dimension tables with a multiple join. Each tuple in the fact table consists of a pointer to each of the dimension tables that provide its multi-dimensional coordinates and stores measures for those coordinates.
- The links between the fact table in the center and the dimension tables in the extremities form a shape like a star.

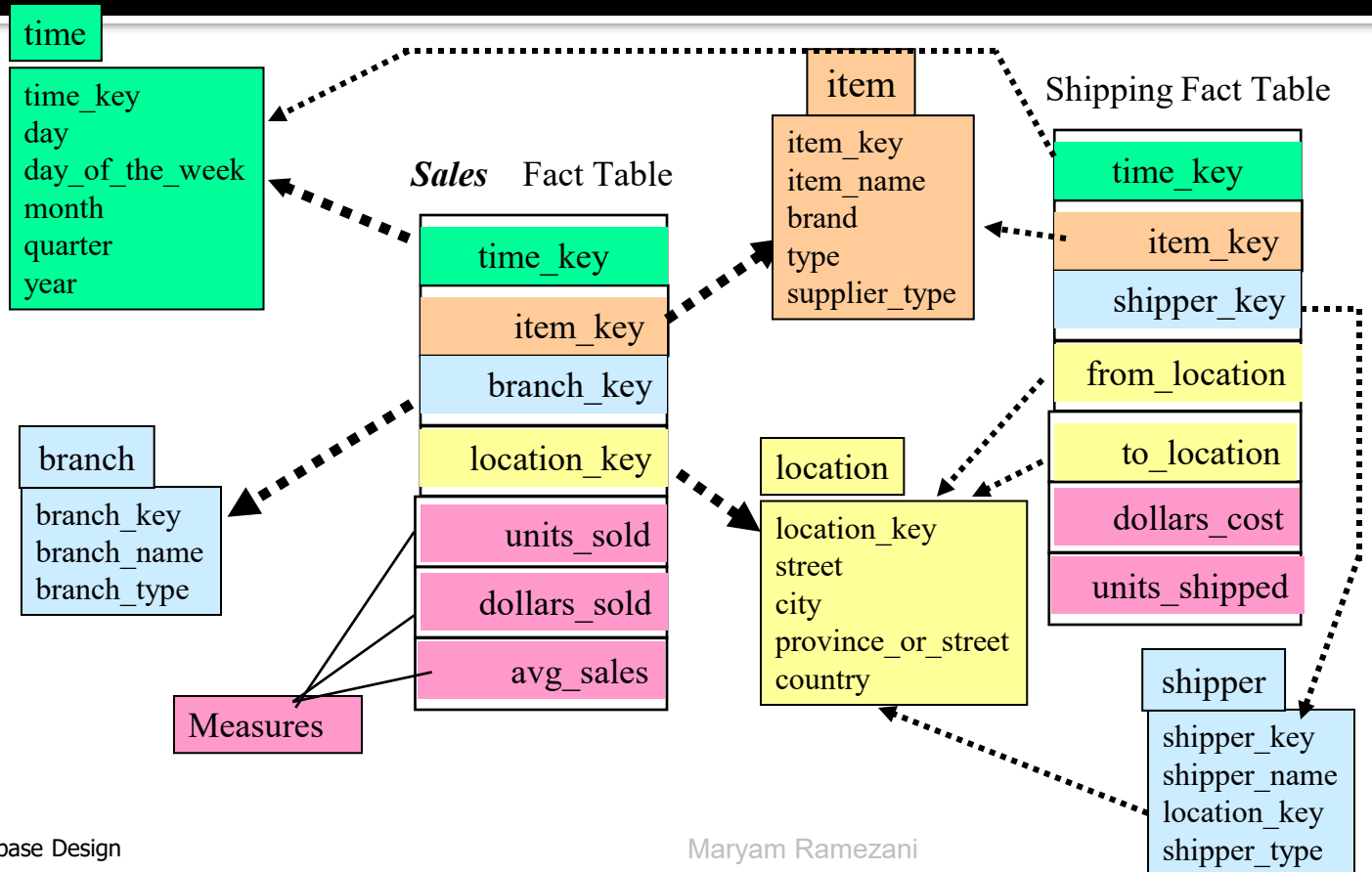
Example of Star Schema



Example of Snowflake Schema



Example of Galaxy Schema (Fact Constellation)



The "Classic" Star Schema

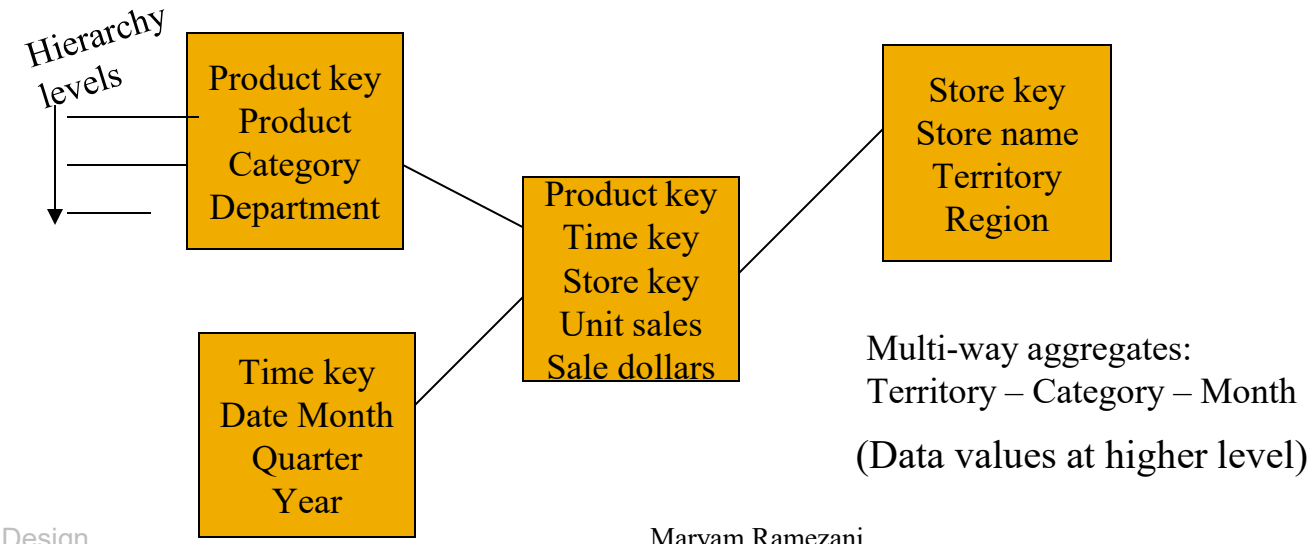
- A relational model with a one-to-many relationship between dimension table and fact table.
- A single fact table, with detail and summary data
- Fact table primary key has only one key column per dimension
- Each dimension is a single table, highly denormalized
- **Benefits**: Easy to understand, intuitive mapping between the business entities, easy to define hierarchies, reduces # of physical joins, low maintenance, very simple metadata
- **Drawbacks**: Summary data in the fact table yields poorer performance for summary levels, huge dimension tables a problem

Need for Aggregates

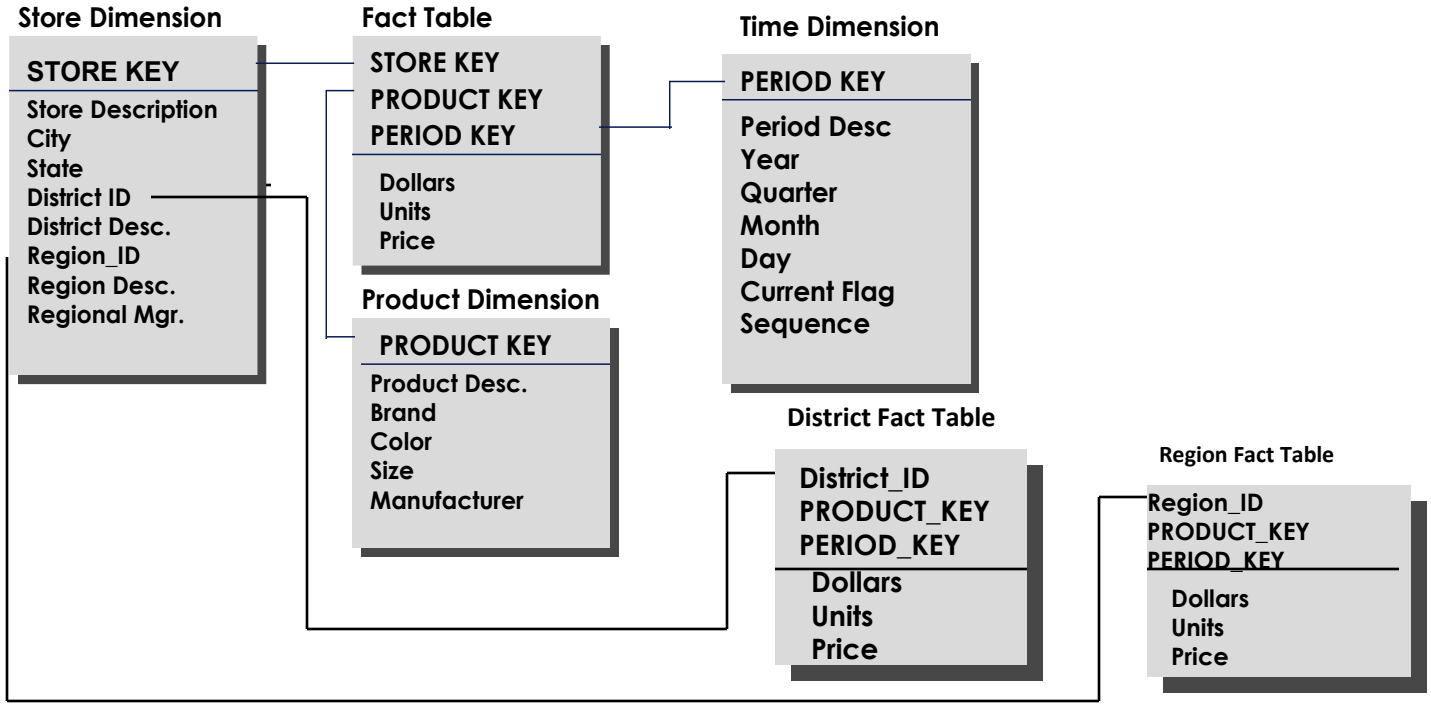
- Sizes of typical tables:
  - Time dimension: 5 years x 365 days = 1825
  - Store dimension: 300 stores reporting daily sales
  - Production dimension: 40,000 products in each store (about 4000 sell in each store daily)
  - Maximum number of base fact table records: 2 billion (lowest level of detail)
- A query involving 1 brand, all store, 1 year: retrieve/summarize over 7 million fact table rows.

Aggregating Fact Tables

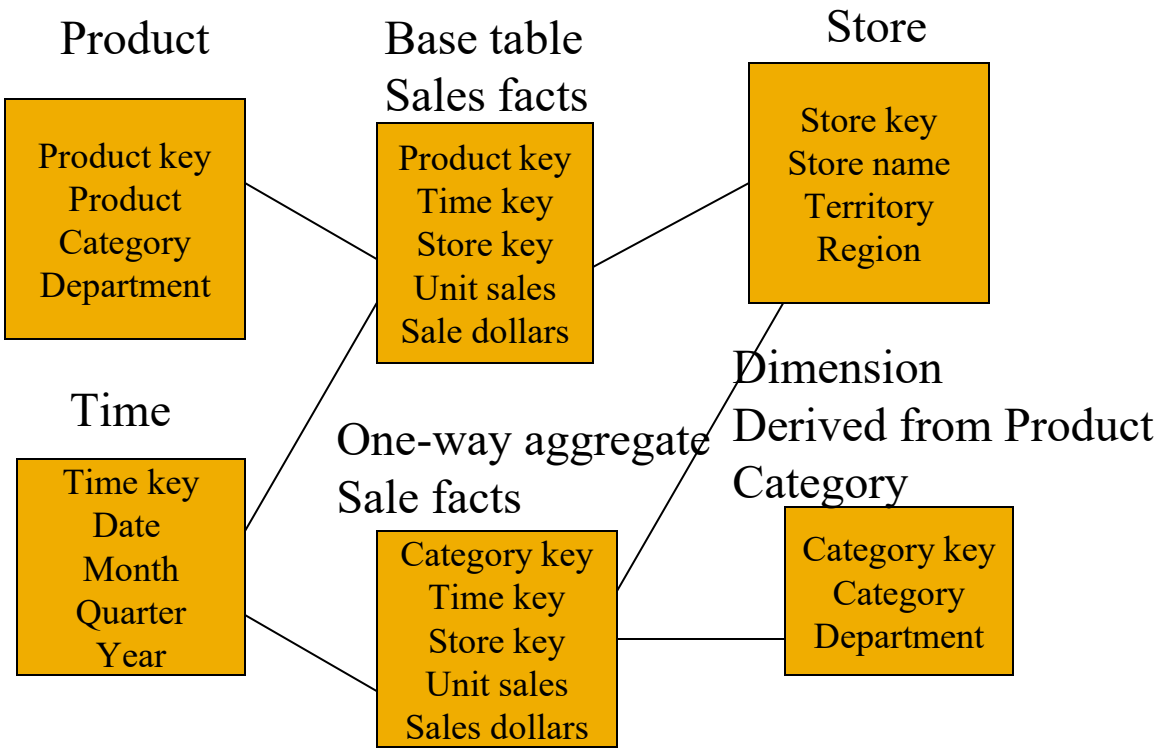
- Aggregate fact tables are summaries of the most granular data at higher levels along the dimension hierarchies.



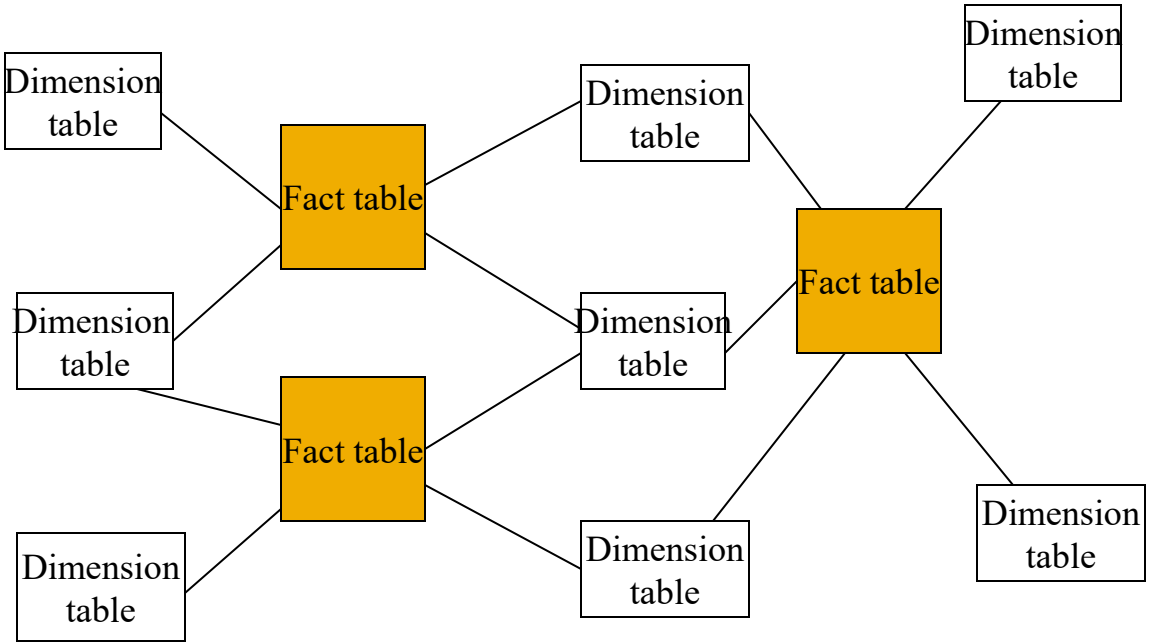
The "Fact Constellation" Schema



Aggregate Fact Tables



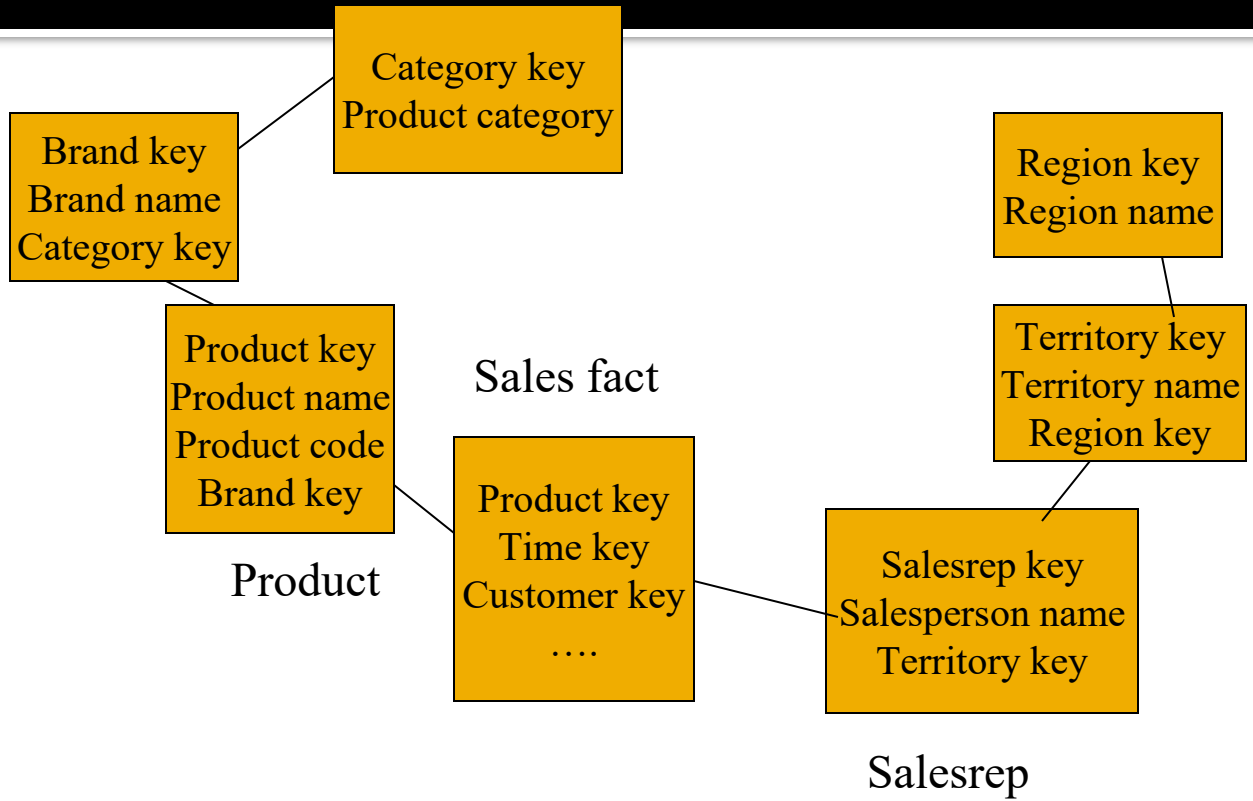
Families of Stars



Snowflake Schema

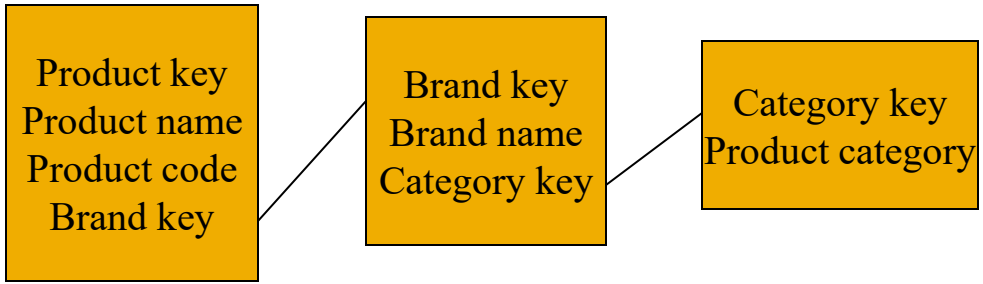
- Snowflake schema is a type of star schema but a more complex model.
- “Snowflaking” is a method of normalizing the dimension tables in a star schema.
- The normalization eliminates redundancy.
- The result is more complex queries and reduced query performance.

Sales: Snowflake Schema



Snowflaking

- The attributes with low cardinality in each original dimension table are removed to form separate tables. These new tables are linked back to the original dimension table through artificial keys.



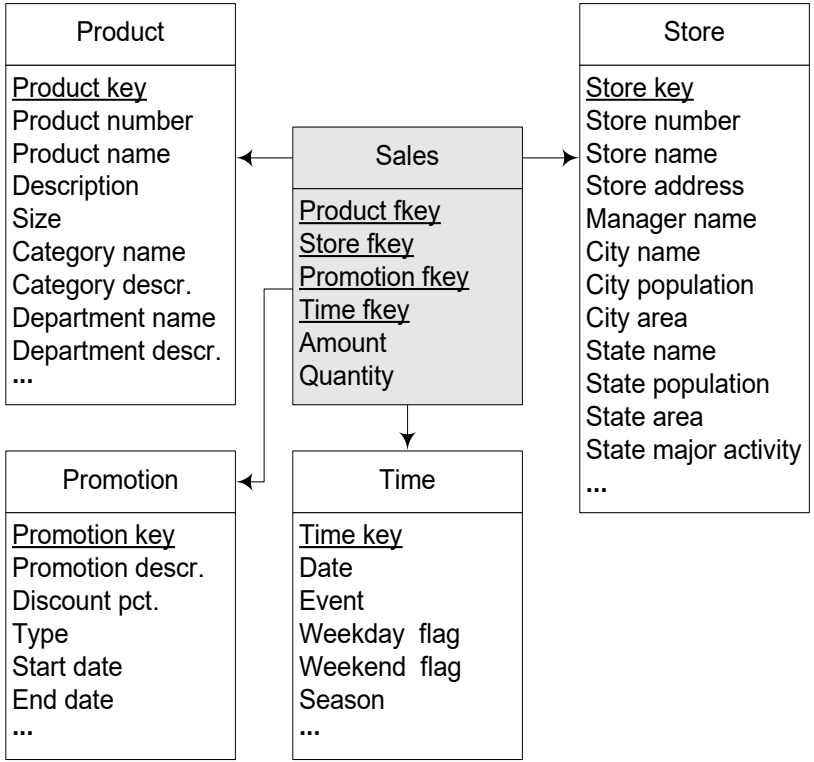
Snowflake Schema

- Advantages:
  - Small saving in storage space
  - Normalized structures are easier to update and maintain
- Disadvantages:
  - Schema less intuitive and end-users are put off by the complexity
  - Ability to browse through the contents difficult
  - Degrade query performance because of additional joins

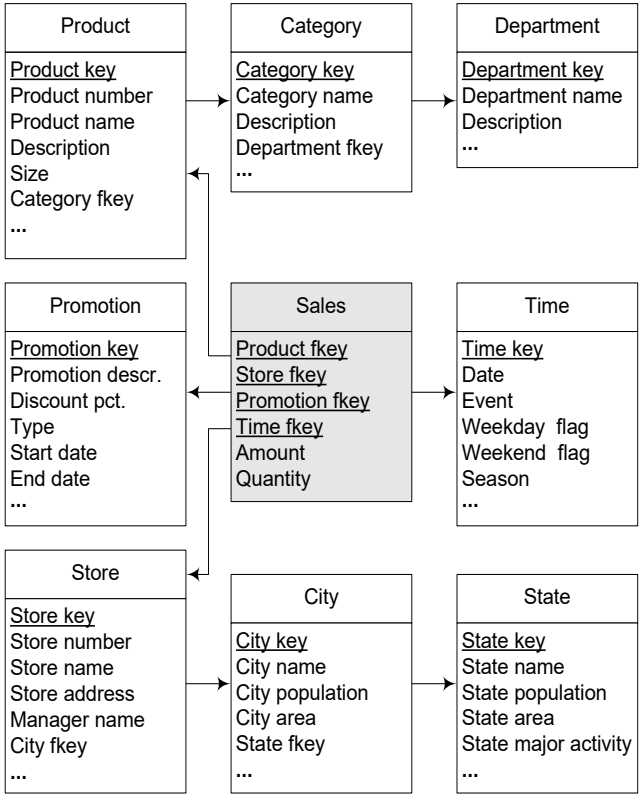
What is the Best Design?

- Performance benchmarking can be used to determine what is the best design.
- Snowflake schema: easier to maintain dimension tables when dimension tables are very large (reduce overall space). It is not generally recommended in a data warehouse environment.
- Star schema: more effective for data cube browsing (less joins): can affect performance.

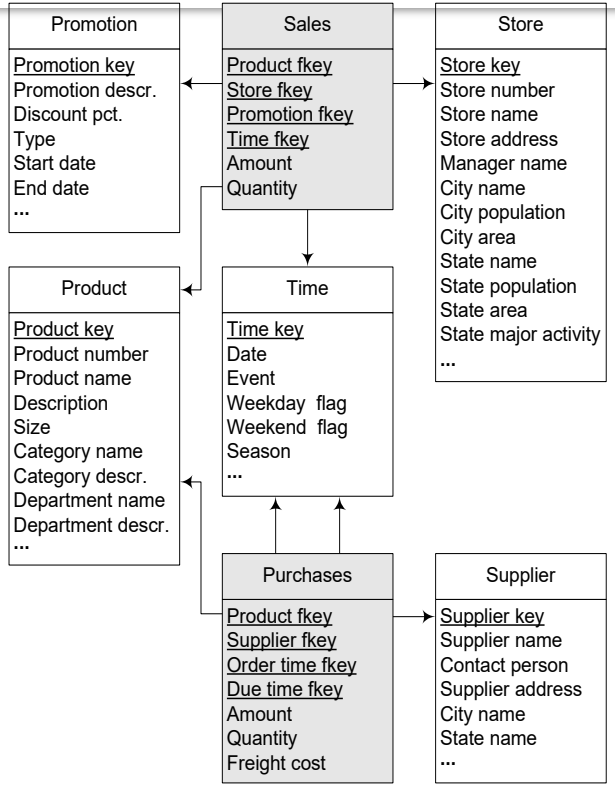
Logical DW Design: Star Schema



Logical DW Design: Snowflake Schemas



Logical DW Design: Constellation Schemas



Aggregates

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

81

Aggregates

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

ans	date	sum
	1	81
	2	48

Another Example

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48

rollup →

← drill-down

Aggregates

- Operators: sum, count, max, min, median, ave
- "Having" clause
- Using dimension hierarchy
  - average by region (within store)
  - maximum by month (within date)

Data Cube

Fact table view:

sale	prodId	storeId	amt
	p1	s1	12
	p2	s1	11
	p1	s3	50
	p2	s2	8

Multi-dimensional cube:

	s1	s2	s3
p1	12		50
p2	11	8	

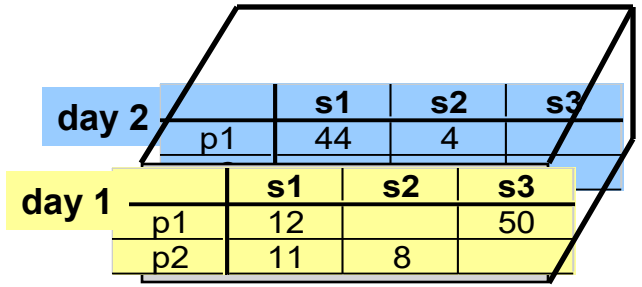
dimensions = 2

3-D Cube

Fact table view:

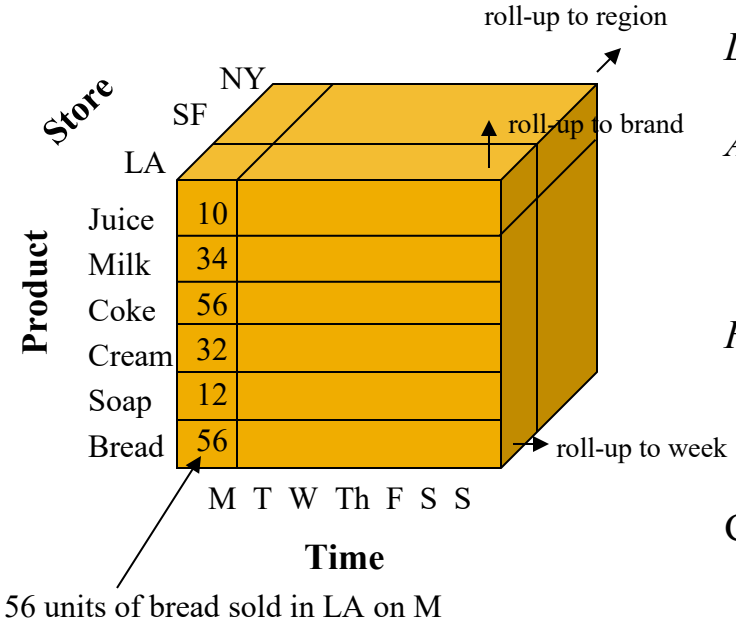
sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

Multi-dimensional cube:



dimensions = 3

Example



Dimensions:

Time, Product, Store

Attributes:

Product (upc, price, …)

Store …

…

Hierarchies:

Product → Brand → …

Day → Week → Quarter

Store → Region →

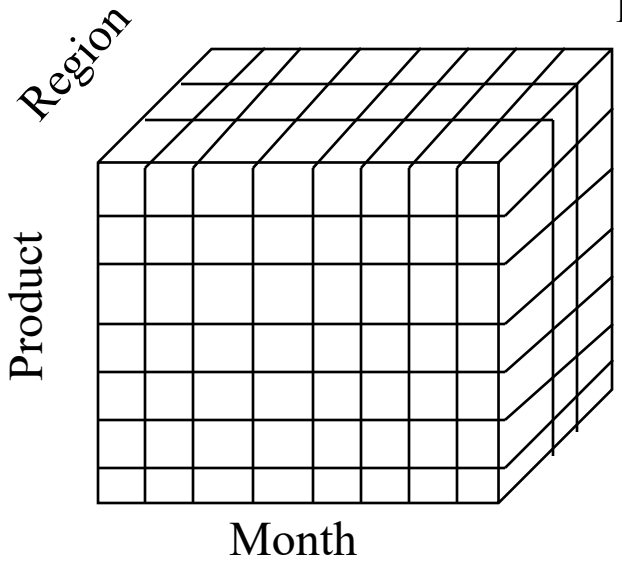
Country

From Tables and Spreadsheets to Data Cubes

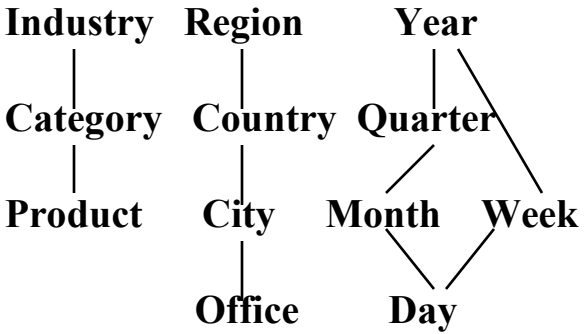
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Multidimensional Data

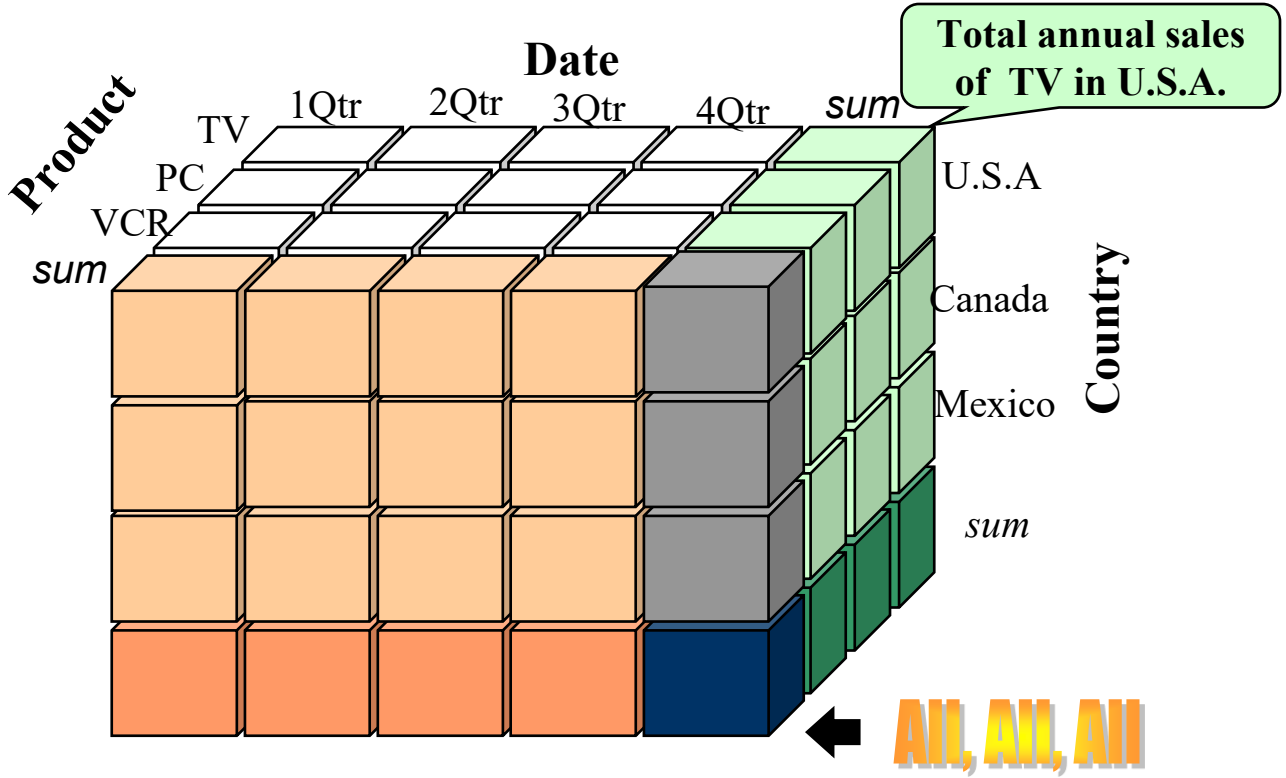
- Sales volume as a function of product, month, and region



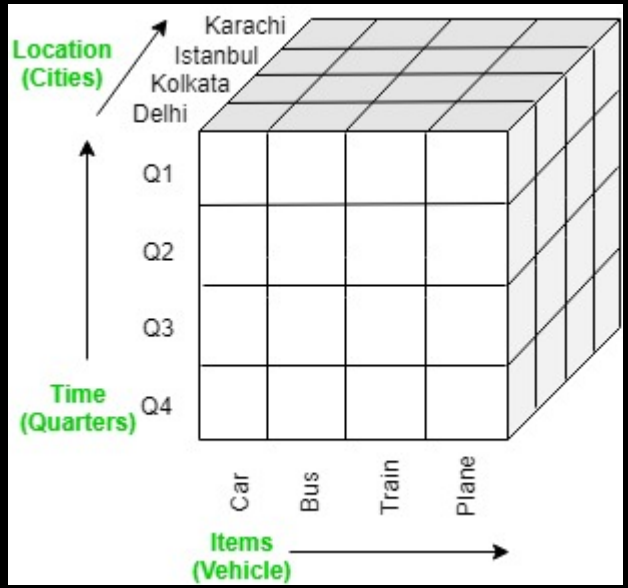
Dimensions: Product, Location, Time
Hierarchical summarization paths



A Sample Data Cube

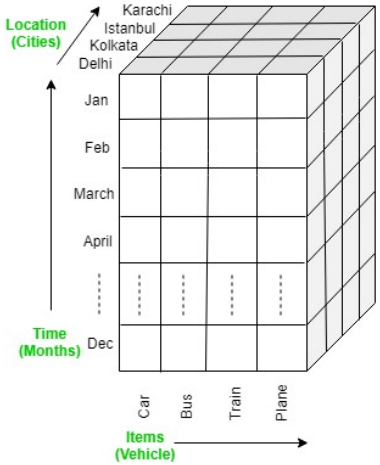


OLAP operations



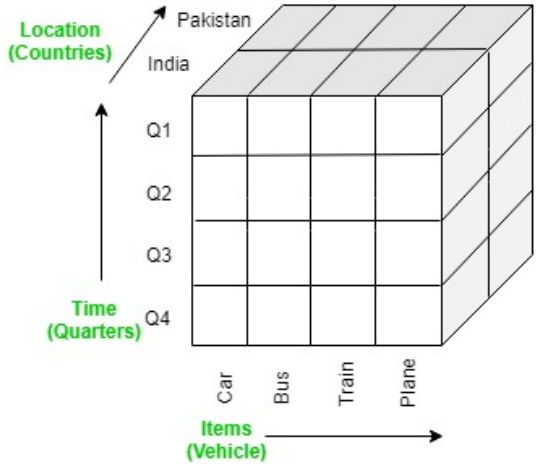
Drill-Down

- In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by: Moving down in the concept hierarchy
- Adding a new dimension
- In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

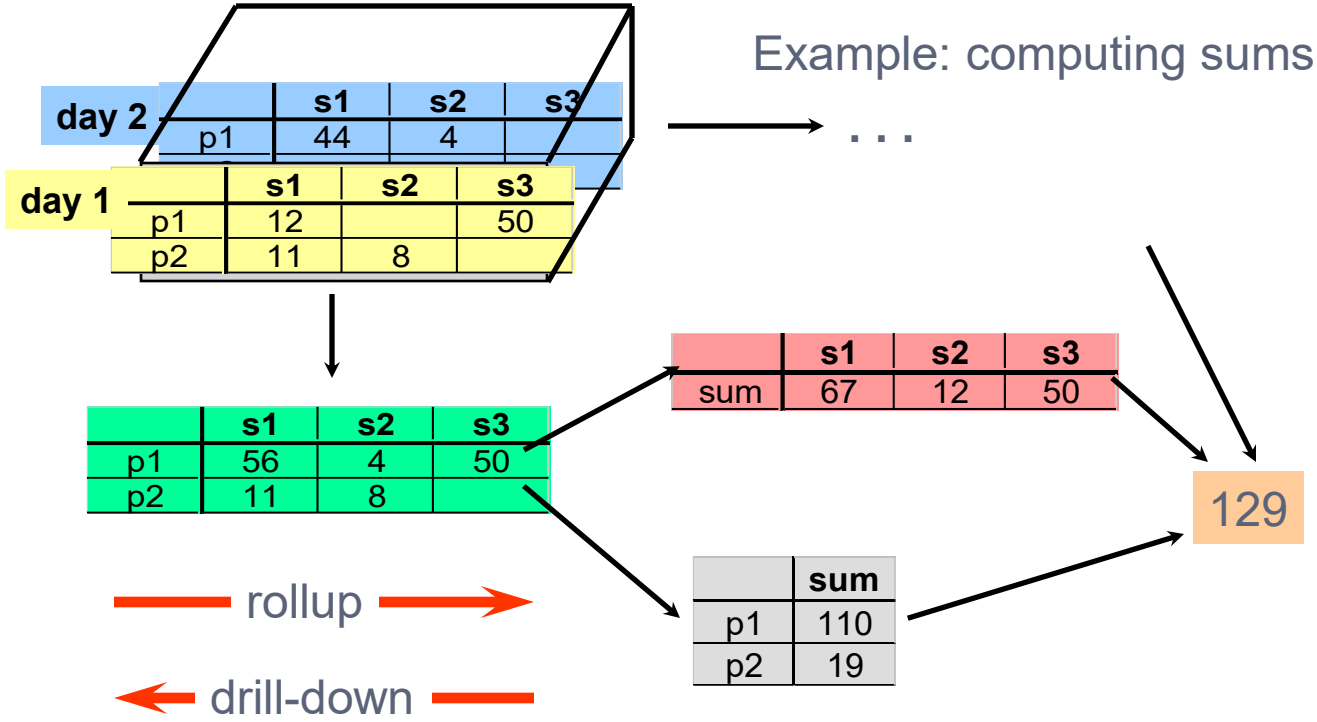


Roll-Up

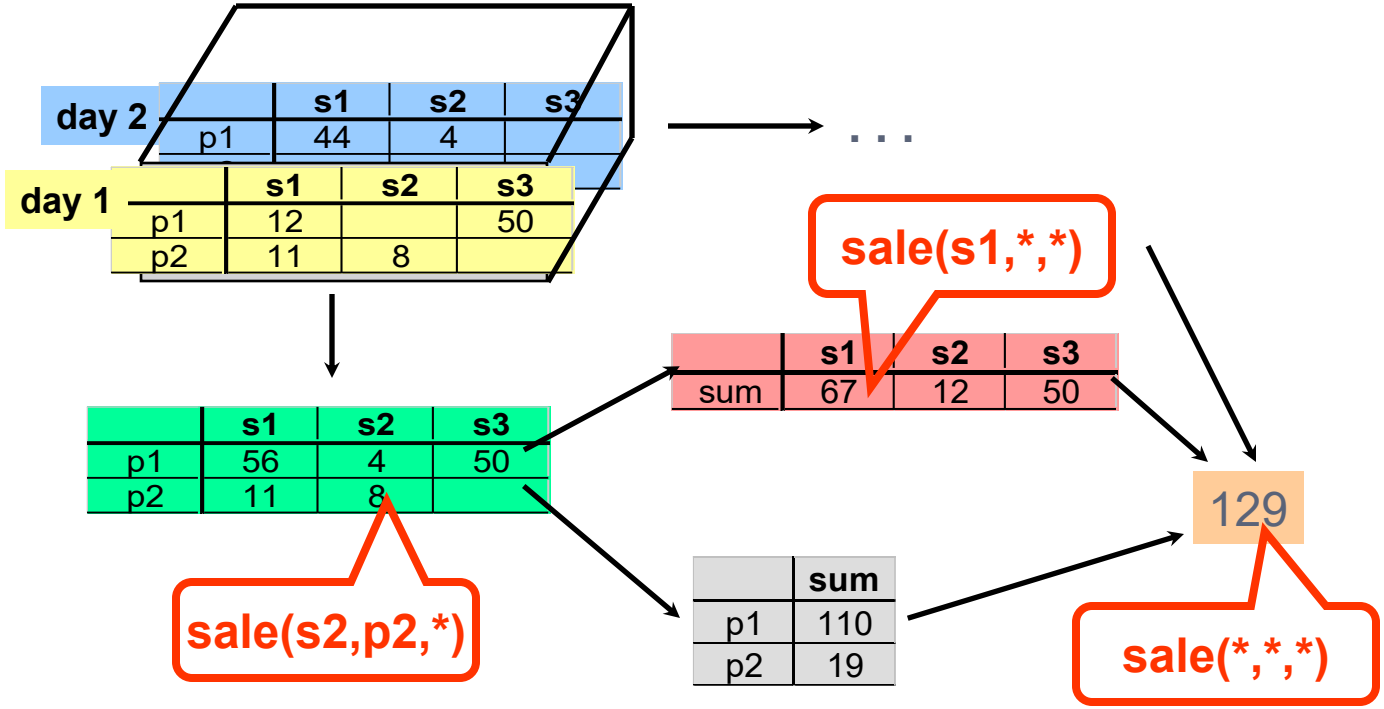
- It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by: Climbing up in the concept hierarchy
- Reducing the dimensions
- In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).



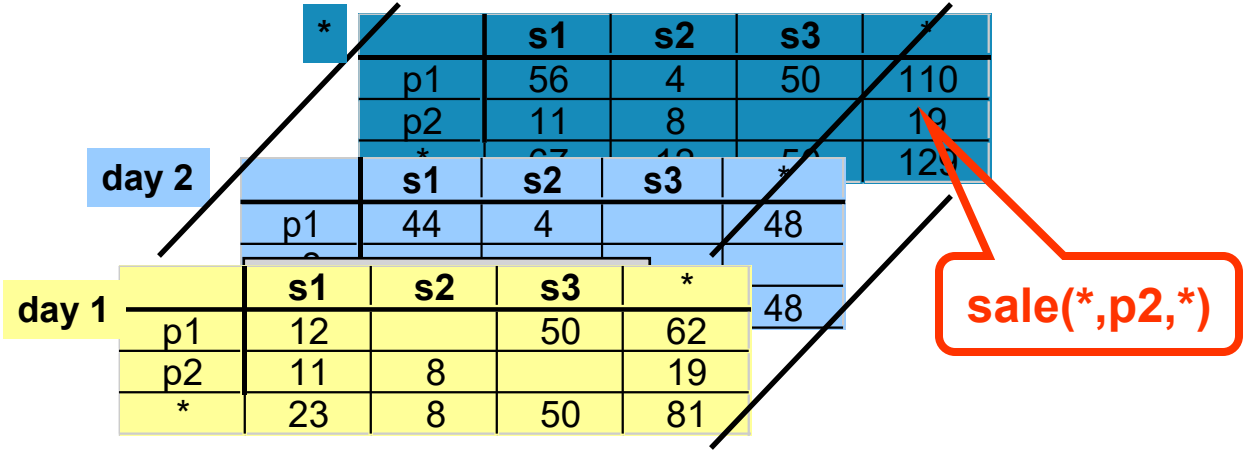
Roll-Up & Drill-Down



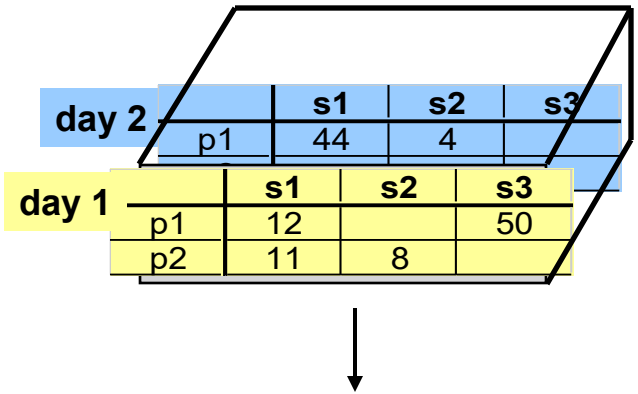
Roll-Up & Drill-Down



Extended Cube



Aggregation Using Hierarchies



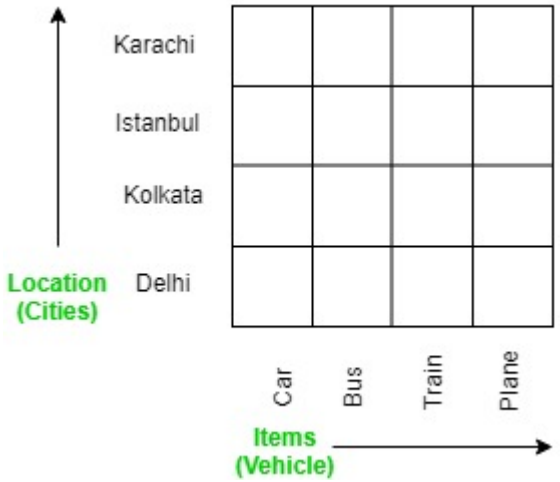
	region A	region B
p1	56	54
p2	11	8

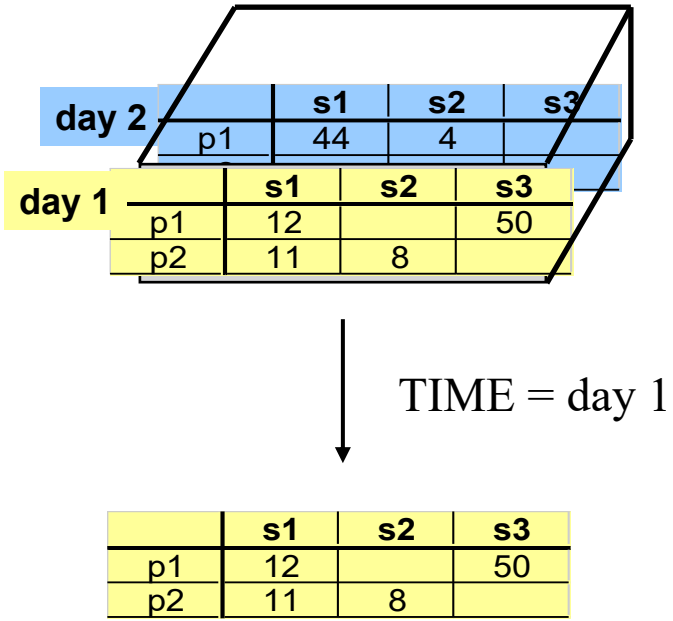
store
region
country

(store s1 in Region A;
stores s2, s3 in Region B)

Slice

It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".

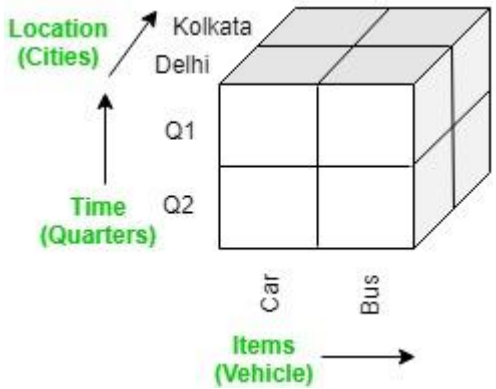




Dice

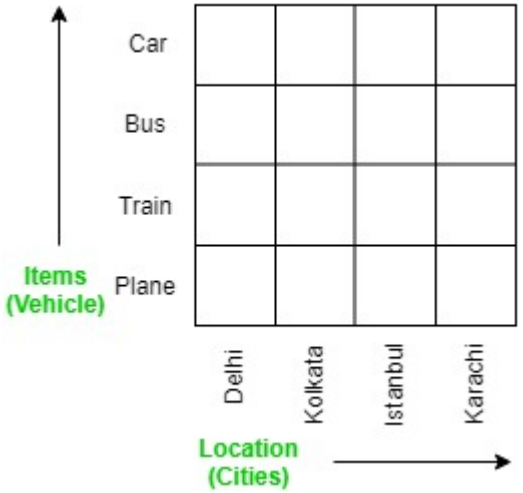
It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:

Location = "Delhi" or "Kolkata"
Time = "Q1" or "Q2"
Item = "Car" or "Bus"



Pivot

It is also known as rotation operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



Slicing & Pivoting

		Sales ($ millions)		
	Products	Time		
		d1	d2	
Store s1	Electronics	$5.2		
	Toys	$1.9		
	Clothing	$2.3		
	Cosmetics	$1.1		
Store s2	Electronics	$8.9		
	Toys	$0.75		
	Clothing	$4.6		
	Cosmetics	$1.5		

		Sales ($ millions)		
	Products	d1		
		Store s1	Store s2	
Store s1	Electronics	$5.2	$8.9	
	Toys	$1.9	$0.75	
	Clothing	$2.3	$4.6	
	Cosmetics	$1.1	$1.5	
Store s2	Electronics			
	Toys			
	Clothing			

Summary of Operations

- Aggregation (roll-up)
  - aggregate (summarize) data to the next higher dimension element
  - e.g., total sales by city, year → total sales by region, year
- Navigation to detailed data (drill-down)
- Selection (slice) defines a subcube
  - e.g., sales where city ='Gainesville' and date = '1/15/90'
- Calculation and ranking
  - e.g., top 3% of cities by average income
- Visualization operations (e.g., Pivot)
- Time functions
  - e.g., time average

Query & Analysis Tools

- Query Building
- Report Writers (comparisons, growth, graphs,…)
- Spreadsheet Systems
- Web Interfaces
- Data Mining

Dimension Versioning

Dim_Customer				
Customer Key	Customer Name	Customer Email	Birth Date	Audit key
101	John Doe	johndoe@example.com	Dec 15 1990	2543
102	Mary Smith	marysmith@example.com	Nov 3 1992	2543

Metadata:
Customer Name: Insert
Customer Email: Insert
Birth Date: Overwrite

HDim_Customer								
Customer Hkey	Customer Key	Customer Name	Customer Email	Birth Date	Audit key	Effective Date	Expiration Date	Current Indicator
482	101	John Doe	johnny32@defunct.com	Jan 15 1990	941	Mar 3 2018	Feb 5 2019	N
541	101	John Doe	johndoe@example.com	Dec 15 1990	2543	Feb 6 2019	Dec 31 9999	Y
613	102	Mary Smith	marysmith@example.com	Nov 3 1992	1594	Apr 26 2019	Jul 17 2020	N
723	102	Mary Jones	maryjones@example.com	Nov 3 1992	2543	Jul 18 2020	Dec 31 9999	Y