

# IS 3400: Advanced Database Management Systems

## Lecture 01

# **Data warehousing and Online Analytical Processing (OLAP)**

# Purpose of Data Warehousing

- Traditional databases are **not optimized for data access** only they have to balance the requirement of data access with the need to ensure integrity of data.
- Most of the times the data warehouse users need only **read access** but, need the access to be **fast** over a **large volume of data**.
- Most of the data required for data warehouse analysis comes from **multiple databases** and these **analysis are recurrent** and predictable to be able to **design specific software** to meet the requirements.
- There is a great need for tools that provide decision makers with information to **make decisions quickly** and reliably based on **historical data**.
- The above functionality is achieved by Data Warehousing and Online analytical processing (OLAP)

# Introduction, Definitions, and Terminology

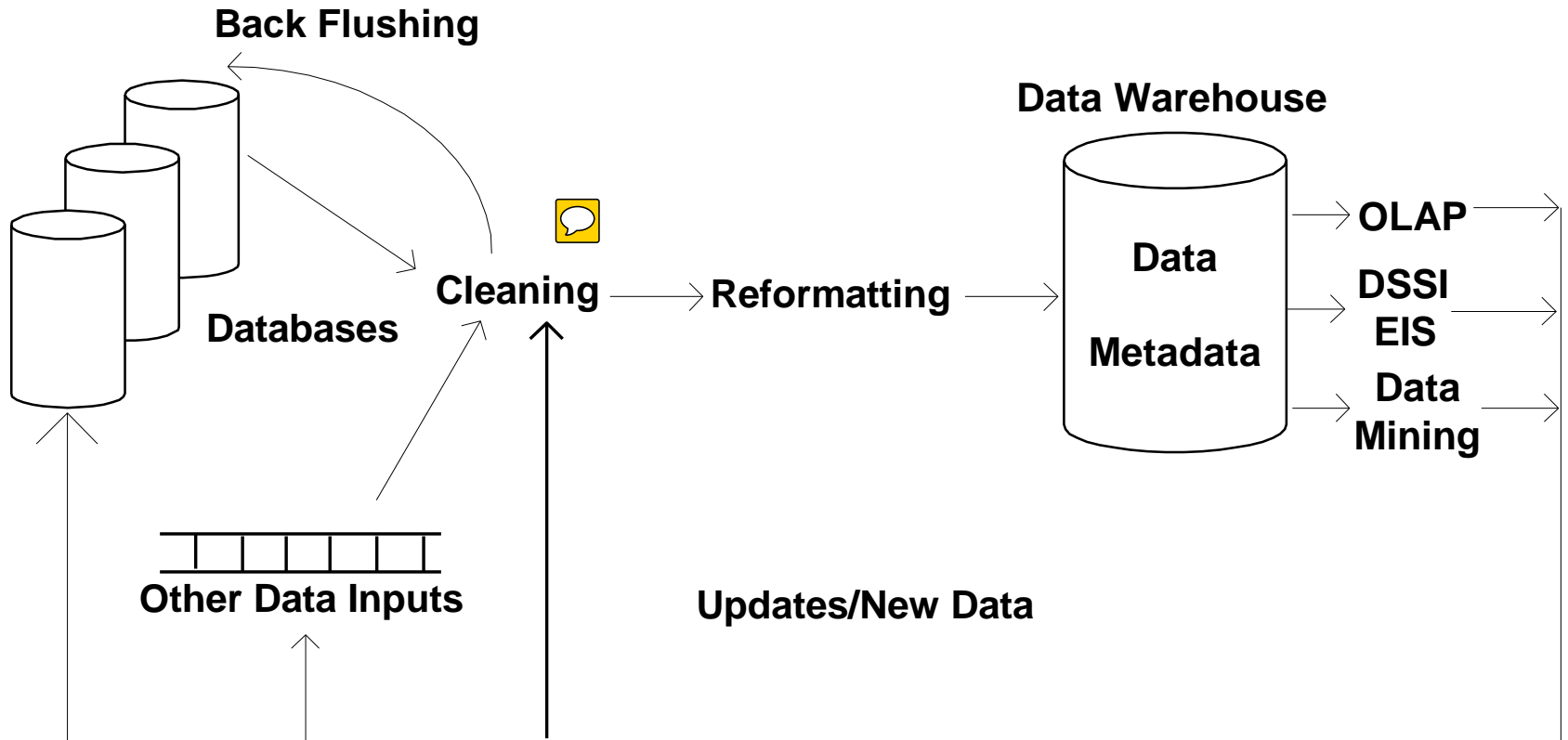
- W. H Inmon characterized a Data Warehouse as:

– “A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management’s decisions.”

# Introduction, Definitions, and Terminology

- Data warehouses have the distinguishing characteristic that they are mainly intended for decision support applications.
  - Traditional databases are transactional.
- Applications that data warehouse supports are:
  - **OLAP** (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
  - **DSS** (Decision Support Systems) also known as EIS (Executive Information Systems) supports organization's leading decision makers for making complex and important decisions.
  - **Data Mining** is used for knowledge discovery, the process of searching data for unanticipated new knowledge.

# Conceptual Structure of Data Warehouse



# Classification of Data Warehouses

- Generally, Data Warehouses are an order of magnitude larger than the source databases.
- The sheer volume of data is an issue, based on which Data Warehouses could be classified as follows.
  - Enterprise-wide data warehouses
    - They are huge projects requiring massive investment of time and resources.
  - Virtual data warehouses
    - They provide views of operational databases that are materialized for efficient access.
  - Data marts
    - These are generally targeted to a subset of organization, such as a department, and are more tightly focused.

# Data Modeling for Data Warehouses

- Traditional Databases generally deal with **two-dimensional data** (similar to a spreadsheet).
  - However, querying performance in a multi-dimensional data storage model is much more efficient.

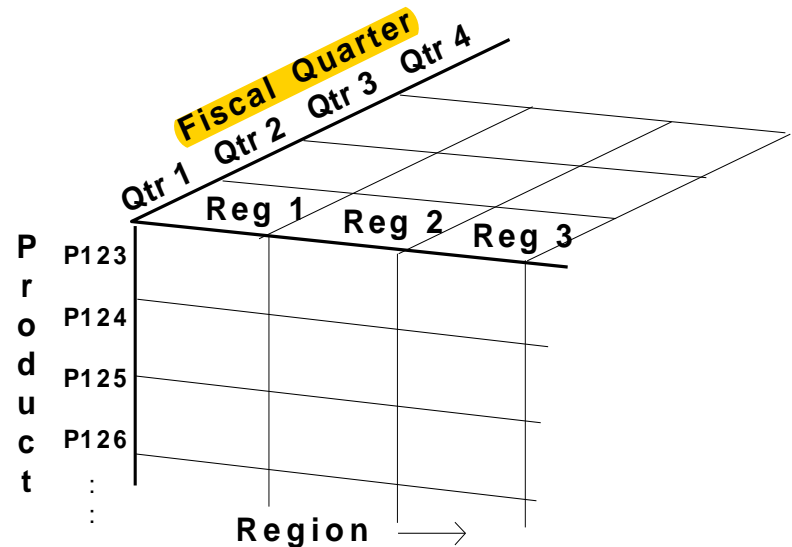
# Data Modeling for Data Warehouses

- Example of Two- Dimensional vs. Multi-Dimensional

Two Dimensional Model



		REGION		
		REG1	REG2	REG3
P R O D U C T	P123			
	P124			
	P125			
	P126			
	⋮			
	⋮			

Three dimensional data cube







# Data Modeling for Data Warehouses

-  Advantages of a multi-dimensional model
  - Multi-dimensional models lend themselves readily to hierarchical views in what is known as roll-up display and drill-down display.
  -  The data can be directly queried in any combination of dimensions, bypassing complex database queries.


# Multi-dimensional Schemas

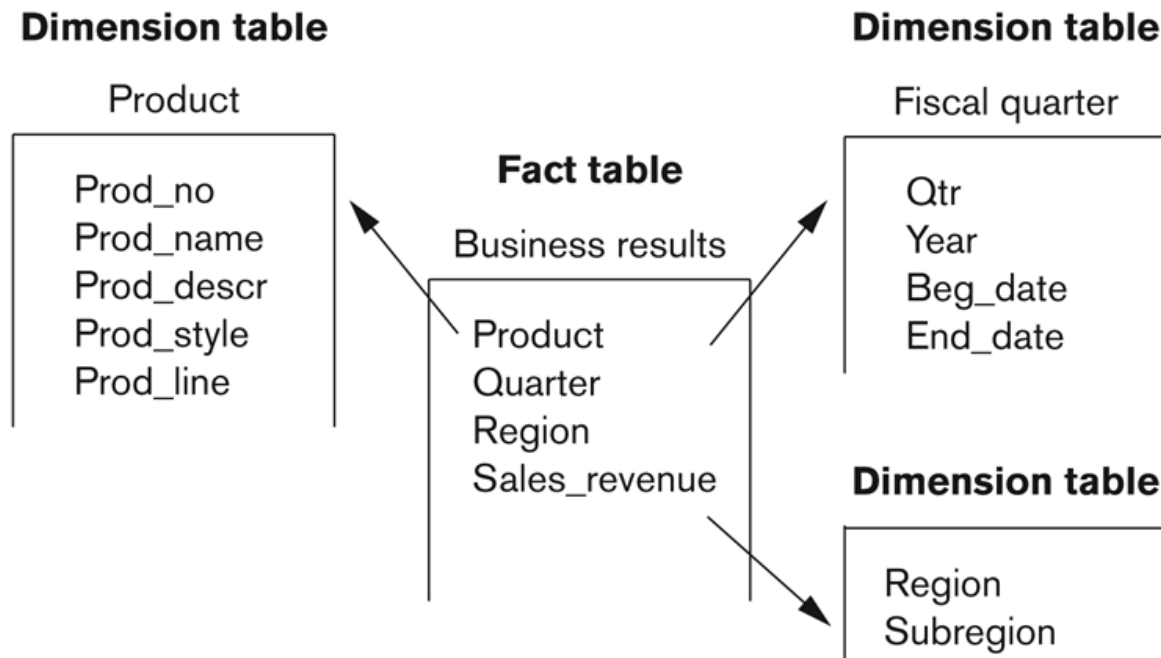
- Multi-dimensional schemas are specified using:
  - **Dimension table** 
    - It consists of **tuples of attributes** of the dimension.
  - **Fact table** 
    - Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data, and the dimensions to identify each tuple in the data.

# Multi-dimensional Schemas

- Two common multi-dimensional schemas are
  - **Star schema:**
    - Consists of a fact table with a single table for each dimension
  - **Snowflake Schema:**
    - It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

# Multi-dimensional Schemas

-  **Star schema:**
  - Consists of a fact table with a single table for each dimension.



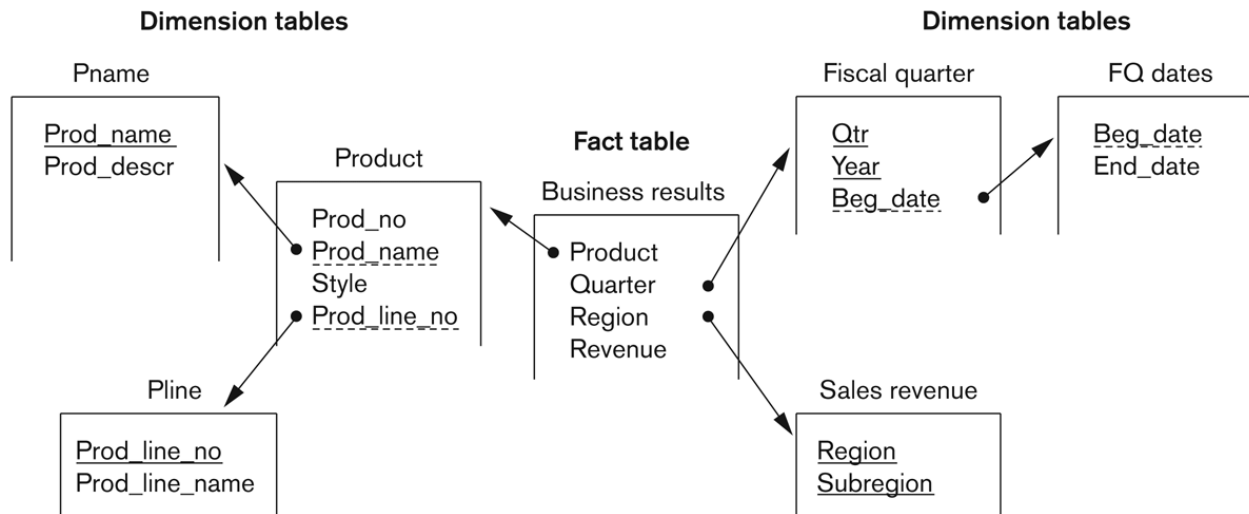
A star schema with fact and dimensional tables.

# Multi-dimensional Schemas



- **Snowflake Schema:**


- It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.



# Building a Data Warehouse

- The builders of Data warehouse should take a broad view of the anticipated use of the warehouse.
- The design should support ad-hoc querying
- An appropriate schema should be chosen that reflects the anticipated usage.

# Building a Data Warehouse

- The Design of a Data Warehouse involves following steps.
  - Acquisition  of data for the warehouse.
  - Ensuring that Data Storage meets the query requirements efficiently.
  - Giving full consideration to the environment in which the data warehouse resides.

# Building a Data Warehouse

- Acquisition of data for the warehouse
  - The data must be extracted from multiple, heterogeneous sources.
  - Data must be formatted for consistency within the warehouse.
  - The data must be cleaned to ensure validity.
    - Difficult to automate cleaning process.
    - Back flushing, upgrading the data with cleaned data.
  - The data must be fitted into the data model of the warehouse.
  - The data must be loaded into the warehouse.
    - Proper design for refresh policy should be considered.



# Warehouse vs. Data Views

- Views and data warehouses are alike in that they both have **read-only extracts** from the databases.
- However, data warehouses are different from views in the following ways:
  - Data Warehouses exist as persistent storage instead of being materialized on demand.
  - Data Warehouses are not usually relational, but rather multi-dimensional.
  - Data Warehouses can be **indexed for optimization**.
  - Data Warehouses **provide specific support of functionality**.
  - Data Warehouses **deals huge volumes of data** that is contained generally in more than one database.

# Difficulties of implementing Data Warehouses

- Lead time is huge in building a data warehouse
  - Potentially it takes years to build and efficiently maintain a data warehouse.
- Both quality and consistency of data are major concerns.
- Revising the usage projections regularly to meet the current requirements.
  - The data warehouse should be designed to accommodate addition and attrition of data sources without major redesign
- Administration of data warehouse would require far broader skills than are needed for a traditional database.

# What is OLAP?

- Online Analytical Processing (OLAP) is a system that further transforms the data into a more structured (summarized) form than tables
- OLAP is a form of Executive Information System (EIS) and Decision Support System (DSS)
- OLAP looks at data in multi-dimensional form (data cube)
- OLAP can be used by multiple users to access data in a data warehouse, e.g. via Internet
- OLAP provides managers with a quick and flexible access to large volume of data

# OLAP Definitions

- OLAP is “the dynamic synthesis, analysis, and consolidation of large volumes of **multi-dimensional** data.” - *Codd (1993)*
  - OLAP technology uses a multi-dimensional view of aggregate data to provide quick access to strategic information
- OLAP is a category of software technology that enables analysts, managers, and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the dimensionality of the enterprise as understood by the user. - *OLAP Council: [www.olapcouncil.org](http://www.olapcouncil.org)*

# Why OLAP?

- OLAP vs. general-purpose query tools
  - Queries - Simply returns the data that fulfils certain constraints
  - OLAP - Rearranges the database tables in a slightly different manner using a process called pre-aggregation or cubes
- OLAP has ability to answer questions like:
  - How much did sales unit A earn in January?
  - How much did sales unit B earn in February?
  - What was their combined sales amount for the first quarter?



# OLTP vs. OLAP

- On-Line Transaction Processing (OLTP):
  - Technology used to perform updates on operational or transactional systems (e.g., point of sale systems)
- On-Line Analytical Processing (OLAP):
  - Technology used to perform **complex analysis** of the data in a data warehouse

# OLTP vs. OLAP

	OLTP	OLAP
<b>User</b>	<ul style="list-style-type: none"><li>• Clerk, IT Professional</li></ul>	<ul style="list-style-type: none"><li>• Knowledge worker</li></ul>
<b>Function</b>	<ul style="list-style-type: none"><li>• Day to day operations</li></ul>	<ul style="list-style-type: none"><li>• Decision support</li></ul>
<b>DB Design</b>	<ul style="list-style-type: none"><li>• Application-oriented (E-R based)</li></ul>	<ul style="list-style-type: none"><li>• Subject-oriented (Star, snowflake)</li></ul>
<b>Data</b>	<ul style="list-style-type: none"><li>• Current, Isolated</li></ul>	<ul style="list-style-type: none"><li>• Historical, Consolidated</li></ul>
<b>View</b>	<ul style="list-style-type: none"><li>• Detailed, Flat relational</li></ul>	<ul style="list-style-type: none"><li>• Summarized, Multidimensional</li></ul>
<b>Usage</b>	<ul style="list-style-type: none"><li>• Structured, Repetitive</li></ul>	<ul style="list-style-type: none"><li>• Ad hoc</li></ul>
<b>Unit of work</b>	<ul style="list-style-type: none"><li>• Short, Simple transaction</li></ul>	<ul style="list-style-type: none"><li>• Complex query</li></ul>
<b>Access</b>	<ul style="list-style-type: none"><li>• Read/write</li></ul>	<ul style="list-style-type: none"><li>• Read Mostly</li></ul>
<b>Operations</b>	<ul style="list-style-type: none"><li>• Index/hash on prim. Key</li></ul>	<ul style="list-style-type: none"><li>• Lots of Scans</li></ul>
<b># Records accessed</b>	<ul style="list-style-type: none"><li>• Tens</li></ul>	<ul style="list-style-type: none"><li>• Millions</li></ul>
<b>#Users</b>	<ul style="list-style-type: none"><li>• Thousands</li></ul>	<ul style="list-style-type: none"><li>• Hundreds</li></ul>
<b>Db size</b>	<ul style="list-style-type: none"><li>• 100 MB-GB</li></ul>	<ul style="list-style-type: none"><li>• 100GB-TB</li></ul>
<b>Metric</b>	<ul style="list-style-type: none"><li>• Trans. throughput</li></ul>	<ul style="list-style-type: none"><li>• Query throughput, response</li></ul>

Source: Datta, GT

# OLAP Applications

- OLAP applications usually have the following common features:
  - Multi-dimensional views of data
    - Data can be viewed from various perspectives, e.g. product, location, time, etc.
  - Support for complex calculations
    - e.g. sales forecasting, moving averages, percentage growth, etc.
  - Time intelligence
    - e.g. comparisons of sales performance between different time periods



# Multi-dimensional views of data

- Example of 2 dimensional views of data

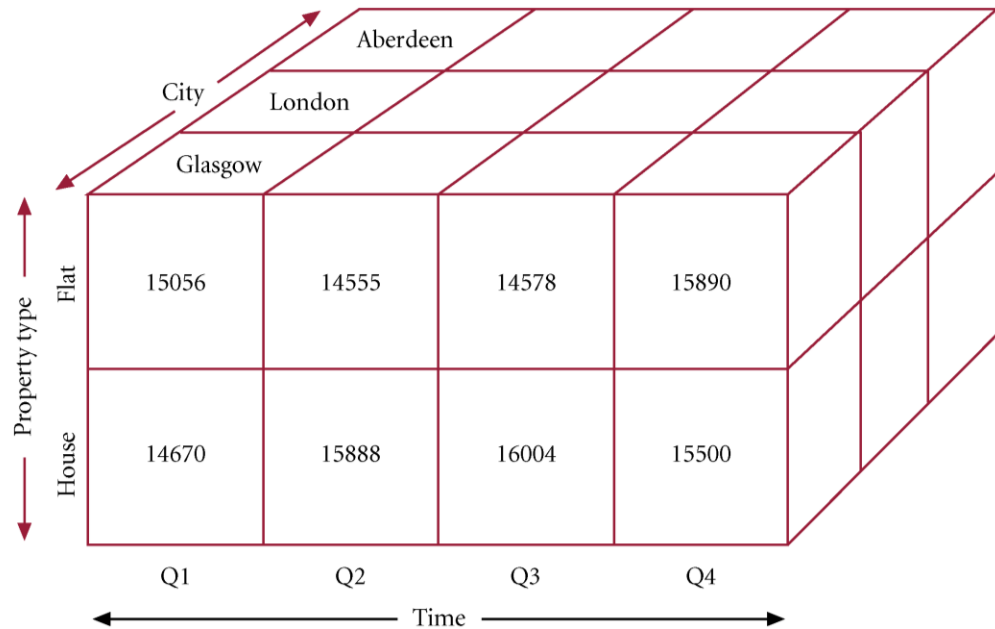
City	Time	Total Revenue
Glasgow	Q1	29726
Glasgow	Q2	30443
Glasgow	Q3	30582
Glasgow	Q4	31390
London	Q1	43555
London	Q2	48244
London	Q3	56222
London	Q4	45632
Aberdeen	Q1	53210
Aberdeen	Q2	34567
Aberdeen	Q3	45677
Aberdeen	Q4	50056
.....	.....	.....
.....	.....	.....

Time ↓	City ←→				
	City	Glasgow	London	Aberdeen	.....
	Quarter				
	Q1	29726	43555	53210	.....
	Q2	30443	48244	34567	.....
	Q3	30582	56222	45677	.....
	Q4	31390	45632	50056	.....

# Multi-dimensional views of data (2)

- Example of 3 dimensional views of data

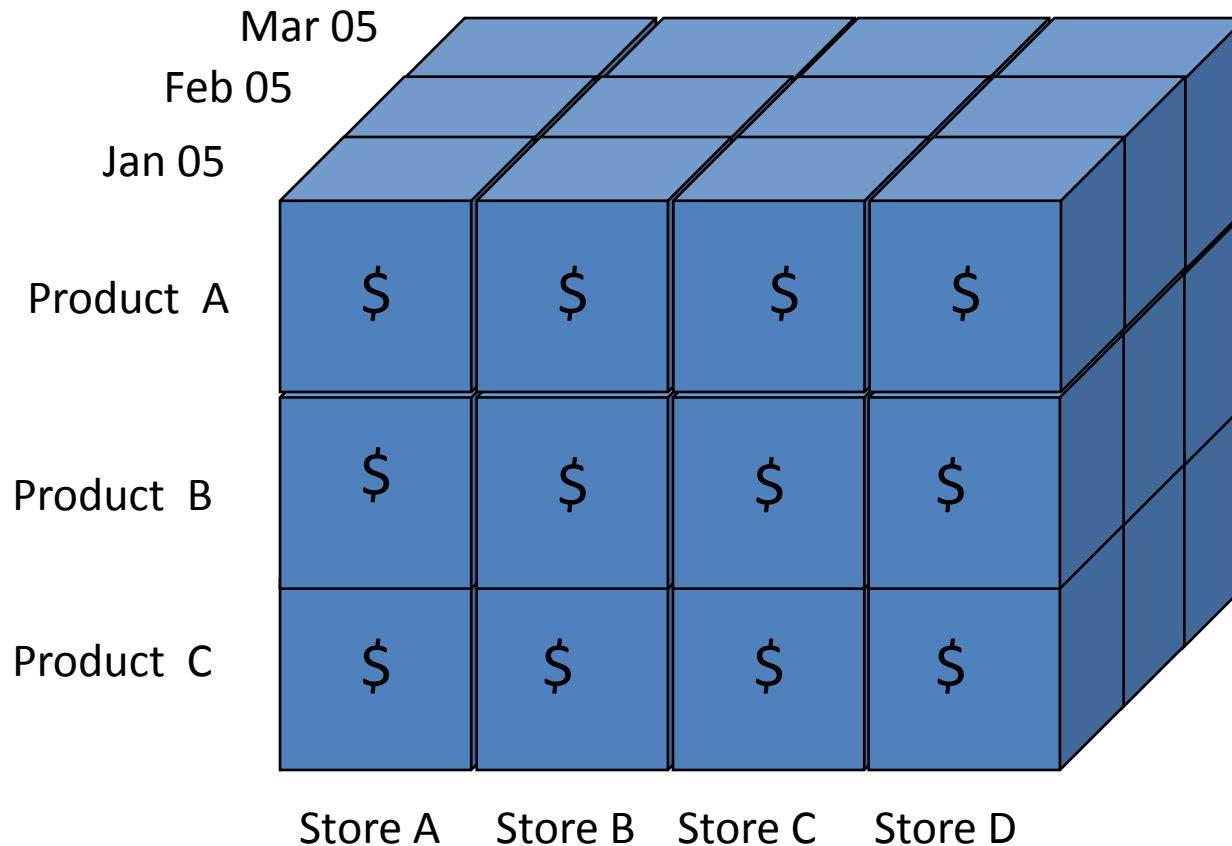
Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....	.....	.....	.....
.....	.....	.....	.....



# Data Cube

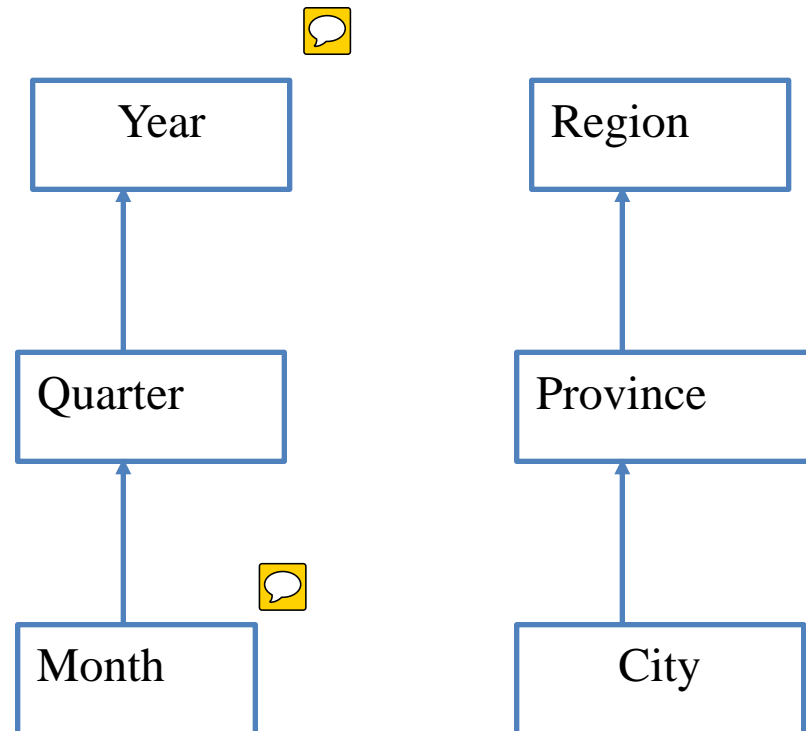
- Multi-dimensional structures are best visualized as cubes of data
- Cube represents data as cells in an array
- Each side of a cube is a dimension
- A cube supports matrix arithmetic
- Hypercube is a form of data cube that has more than 3 dimensions
  - Hypercube can be represented as cube that contains cubes for other dimensions (cubes within cubes)
  - As number of dimensions increases, number of the cube's cells increases exponentially

# Data Cube Example



# Concept Hierarchy

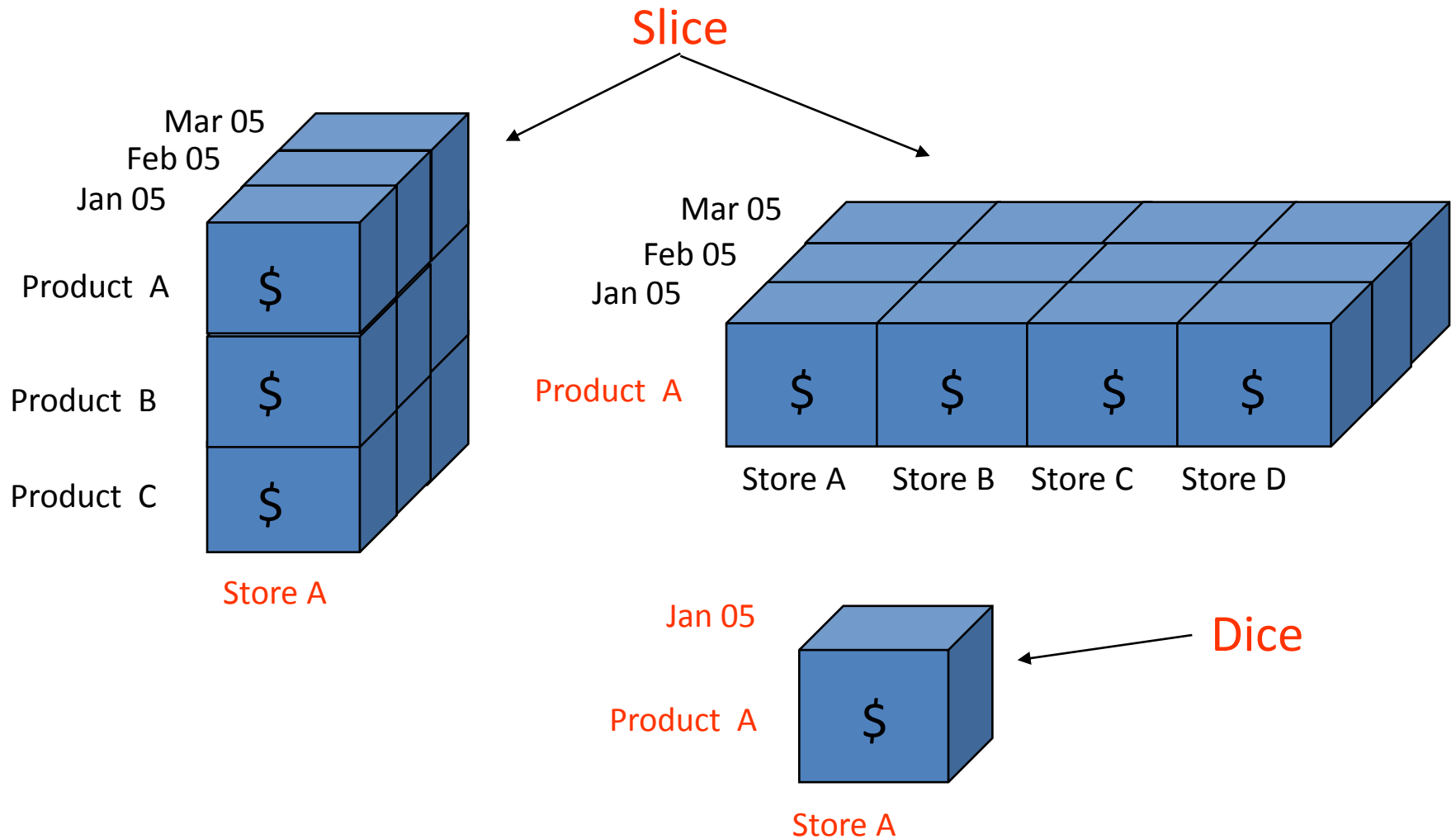
- Attribute may have concept hierarchies associated with
- Examples



# OLAP Operations

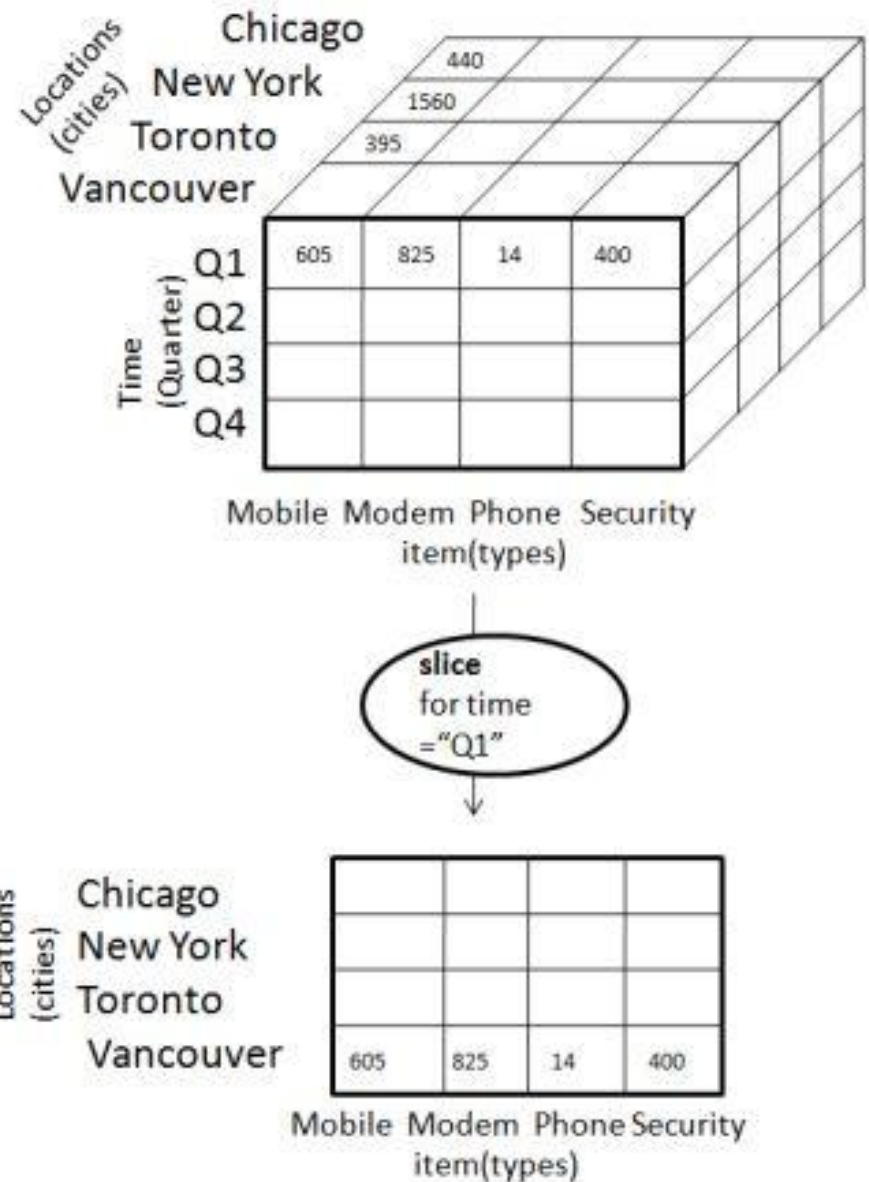
- Slice
  - Select data on a single dimension of a data cube
- Dice
  - Extracts a sub-cube from the original cube
- Roll-up (aggregation)
  - Combining of cells for one dimension
  - Generalization, e.g. Jan, Feb, Mar = Quarter 1
  - May be used with “concept hierarchy”
- Drill-down
  - Reverse of “Roll-up” operation
  - Examine data at level of greater detail, e.g. Northern Region = Chiang Mai, Chiang Rai, ...
- Rotation (pivot)
  - Allow user to view data from a new perspective
  - Axis rotation

# OLAP Operations (2)



# Slice

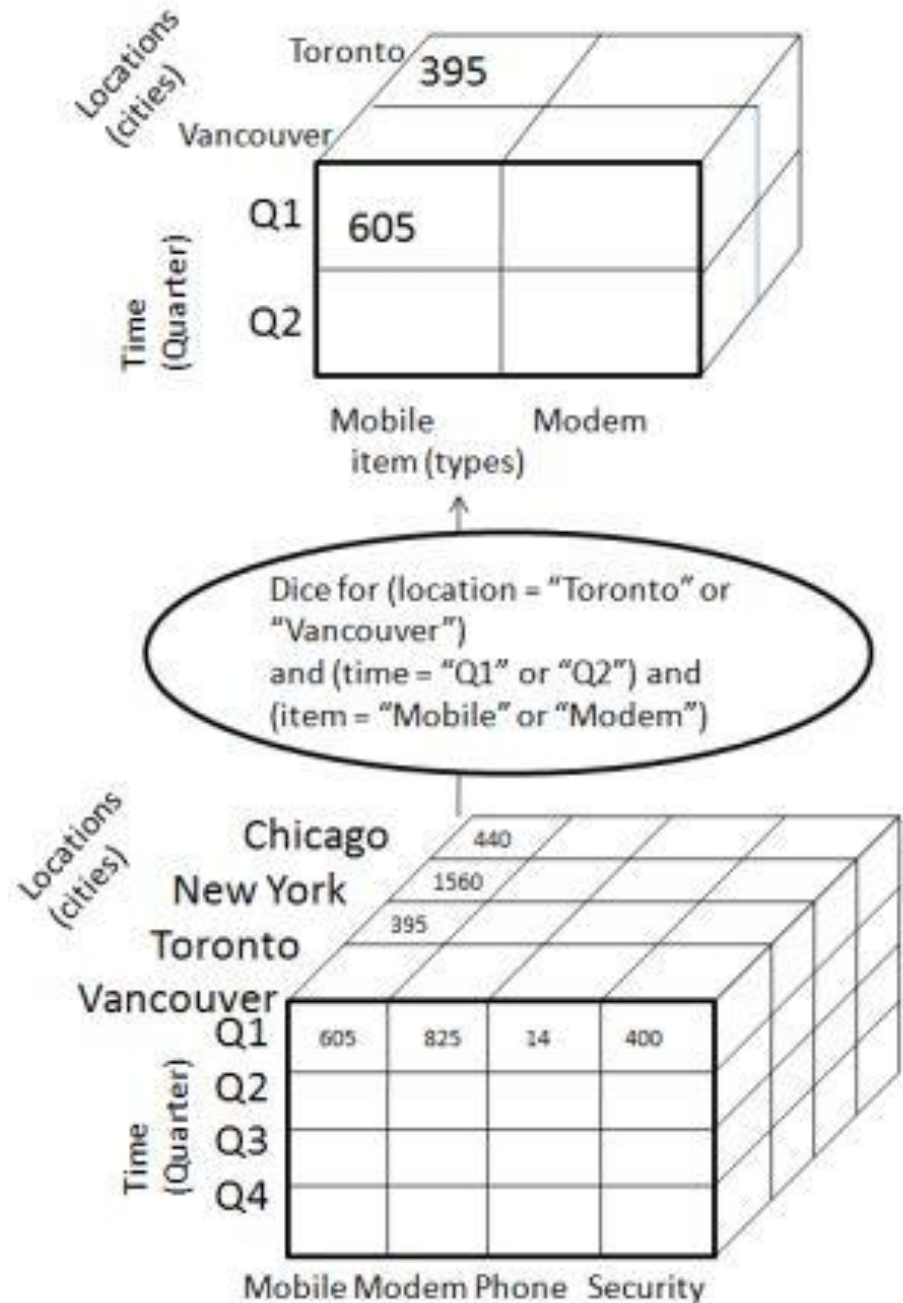
- The slice operation selects one **particular dimension** from a given cube and provides a new sub-cube.
- Slice is performed for the dimension "time" using the criterion time = "Q1".





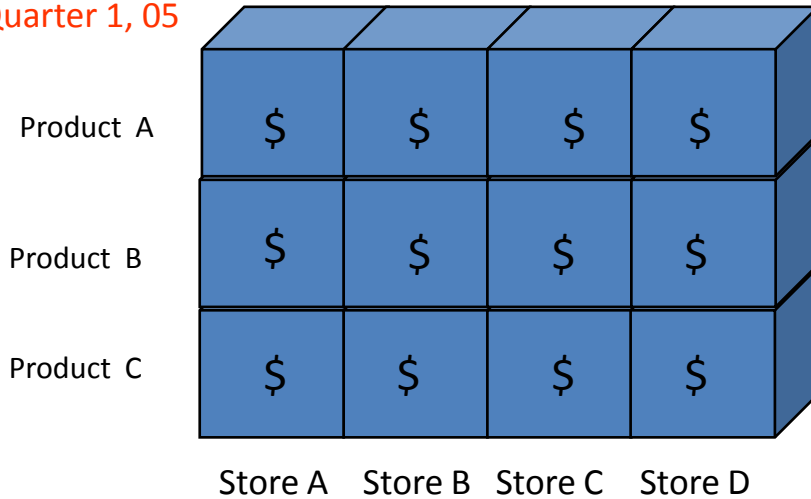
# Dice

- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item = "Mobile" or "Modem")



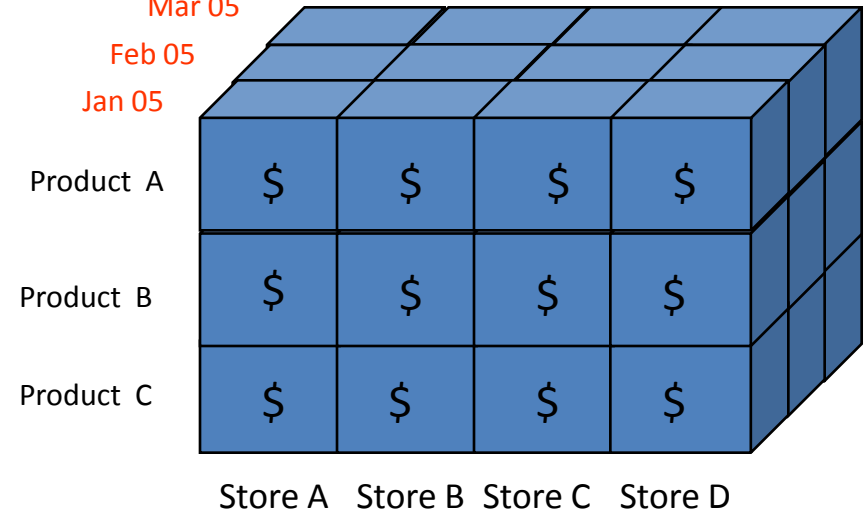
# OLAP Operations (3)

Quarter 1, 05



Roll-up

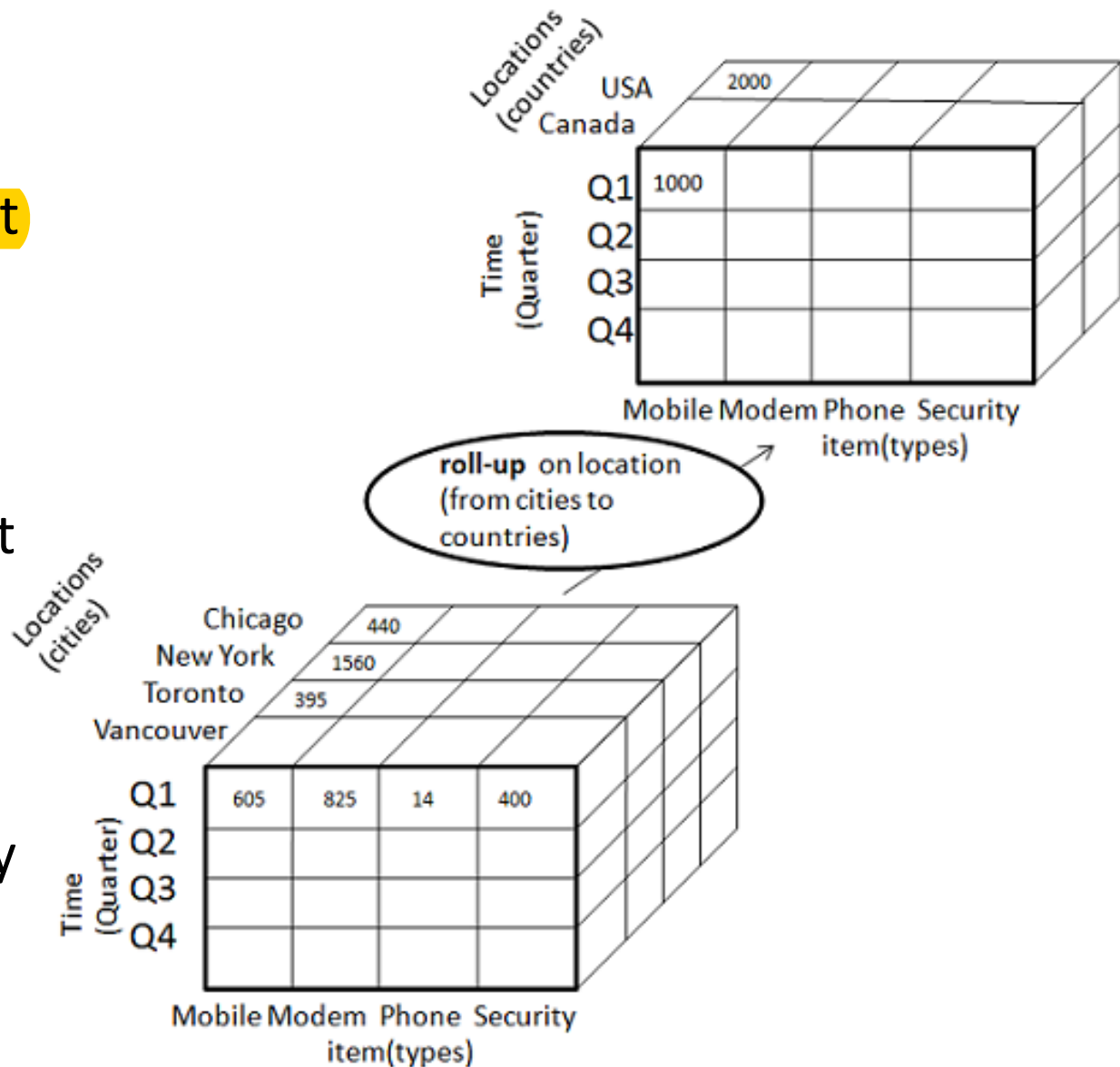
Mar 05  
Feb 05  
Jan 05



Drill-down

# Roll-up

- Performed by **climbing up a concept hierarchy** for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location.



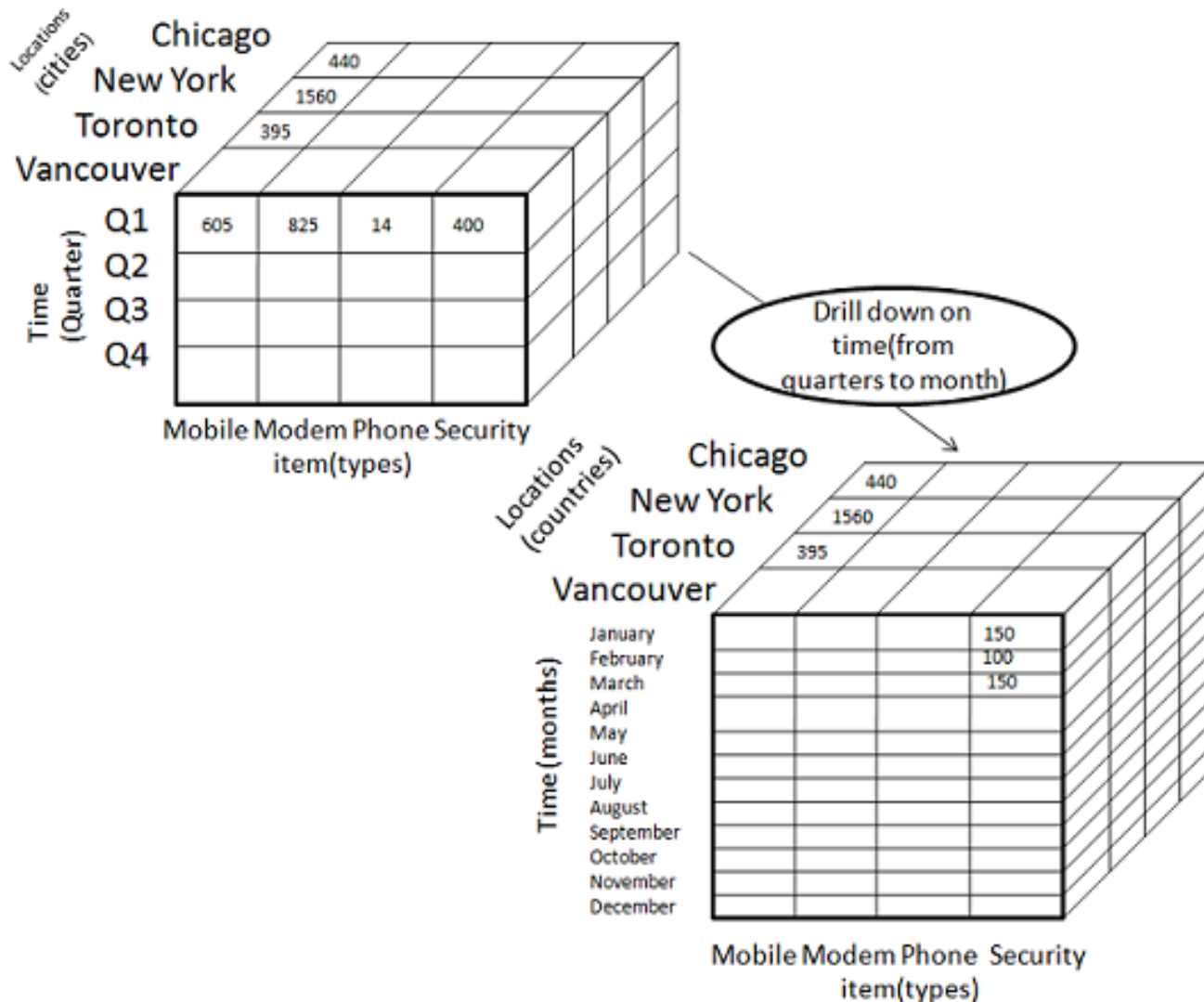
# Roll-up

- Roll-up performs aggregation on a data cube in any of the following ways:
  - By climbing up a concept hierarchy for a dimension
  - By dimension reduction

# Drill-down

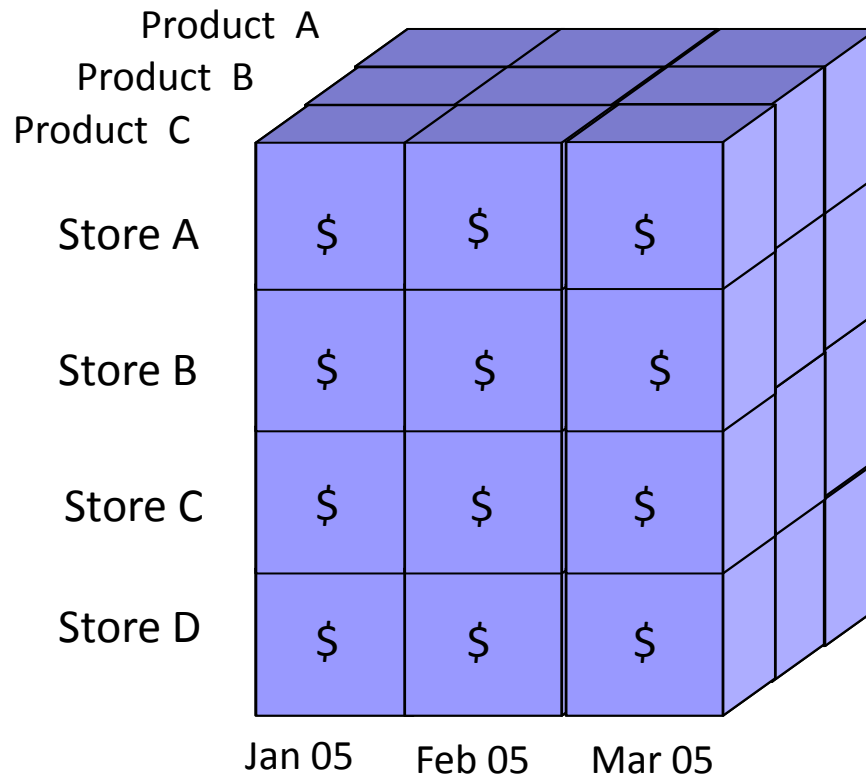
- Drill-down is the reverse operation of roll-up.
- It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension
  - By introducing a new dimension.

# Drill-down



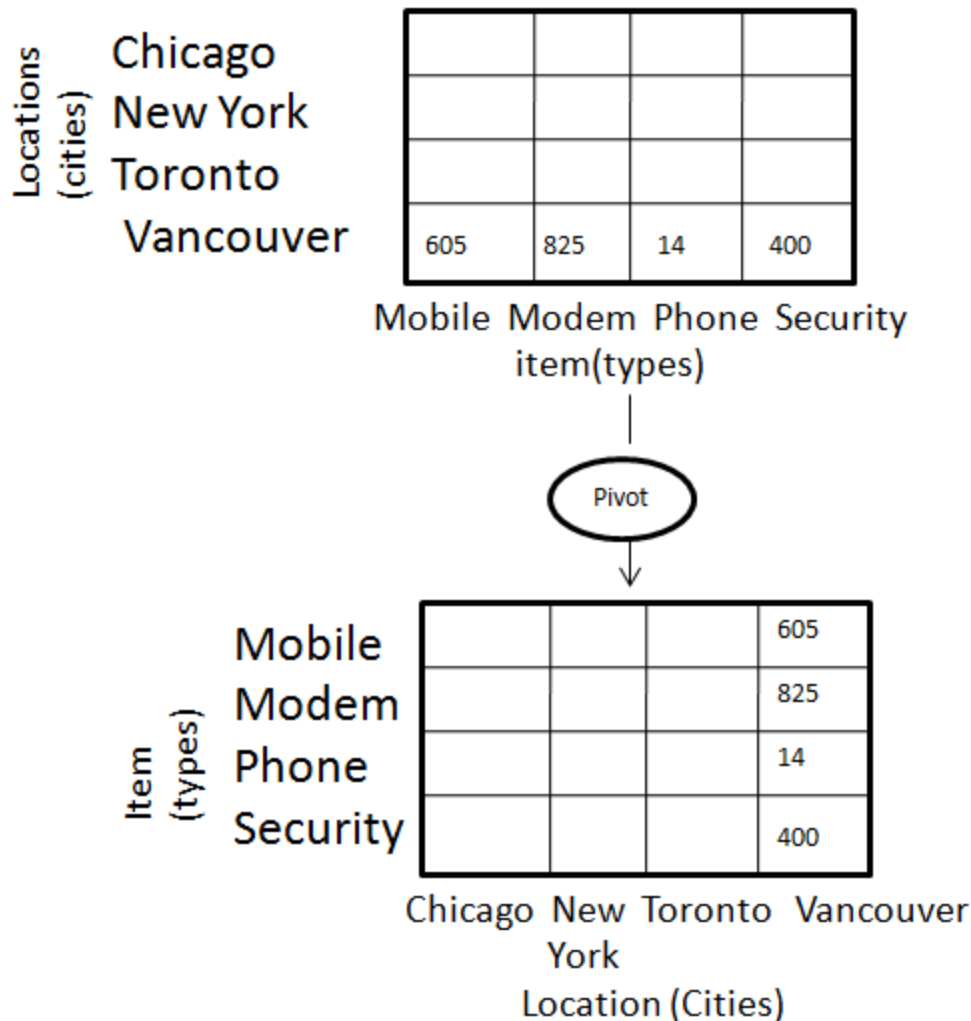
- It navigates the data from less detailed data to highly detailed data.

# OLAP Operations (4)



Rotation

# Rotation (Pivot)



It rotates the data axes in view in order to provide an alternative presentation of data.

Item and location axes in 2-D slice are rotated.



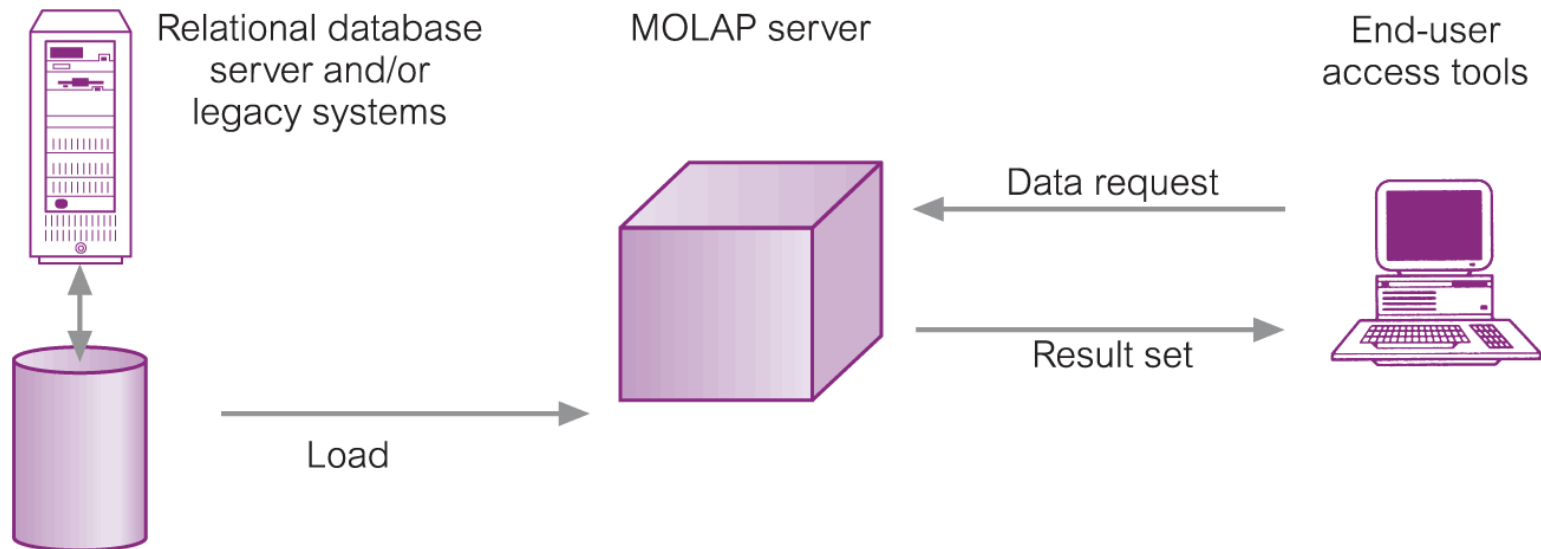
# Types of OLAP tools

- OLAP tools are categorized based on how they store and process multi-dimensional data
- 4 main types of OLAP tools:
  - Multi-dimensional OLAP (MOLAP)
  - Relational OLAP (ROLAP)
  - Hybrid OLAP (HOLAP)
  - Desktop OLAP (DOLAP)

# Multi-dimensional OLAP (MOLAP)

- Use Multi-dimensional Database Management System (MDDDBMS) to organize and analyze data
- Use some efficient storage techniques to minimize disk space requirement
- The database is stored in a special, usually proprietary, structure that is optimized for multidimensional analysis.
- Provides good performance when data is used as designed
- Provide a tight coupling between data structure and presentation layer
  - Access to data structure may be provided via application programming interfaces (APIs)

# MOLAP Architecture



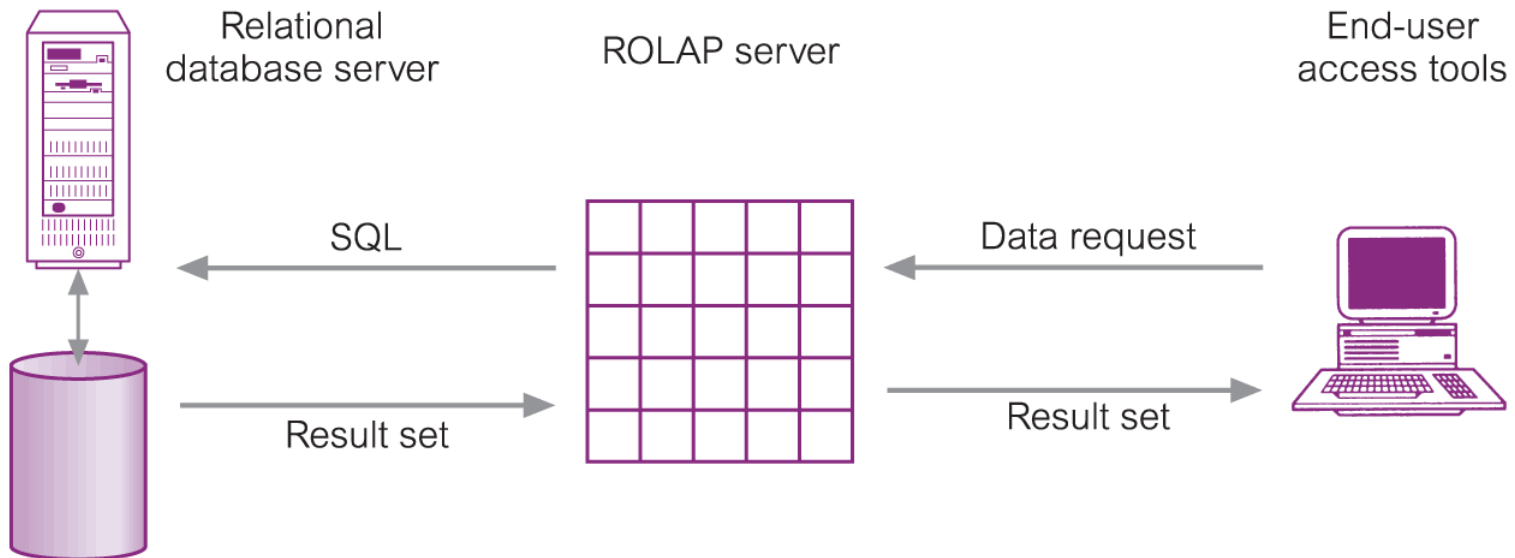
# MOLAP Issues

- MOLAP products require different skills and tools to build and maintain the database, thus increasing the cost and complexity of support
  - MDDDBMS is a new and immature technology (compared to RDBMS)
- Practical limit on the size because the time taken to calculate the database and the space required to hold these pre-calculated values

# Relational OLAP (ROLAP)

- Fastest-growing type of OLAP technology
- MOLAP databases has some limitations
  - Not all data can be efficiently stored in MOLAP databases
- Uses supports from RDBMS
  - avoids need to create multi-dimensional database
  - creates multi-dimensional views from relational database
- May use SQL to support multi-dimensional data analysis

# ROLAP Architecture



# ROLAP Issues

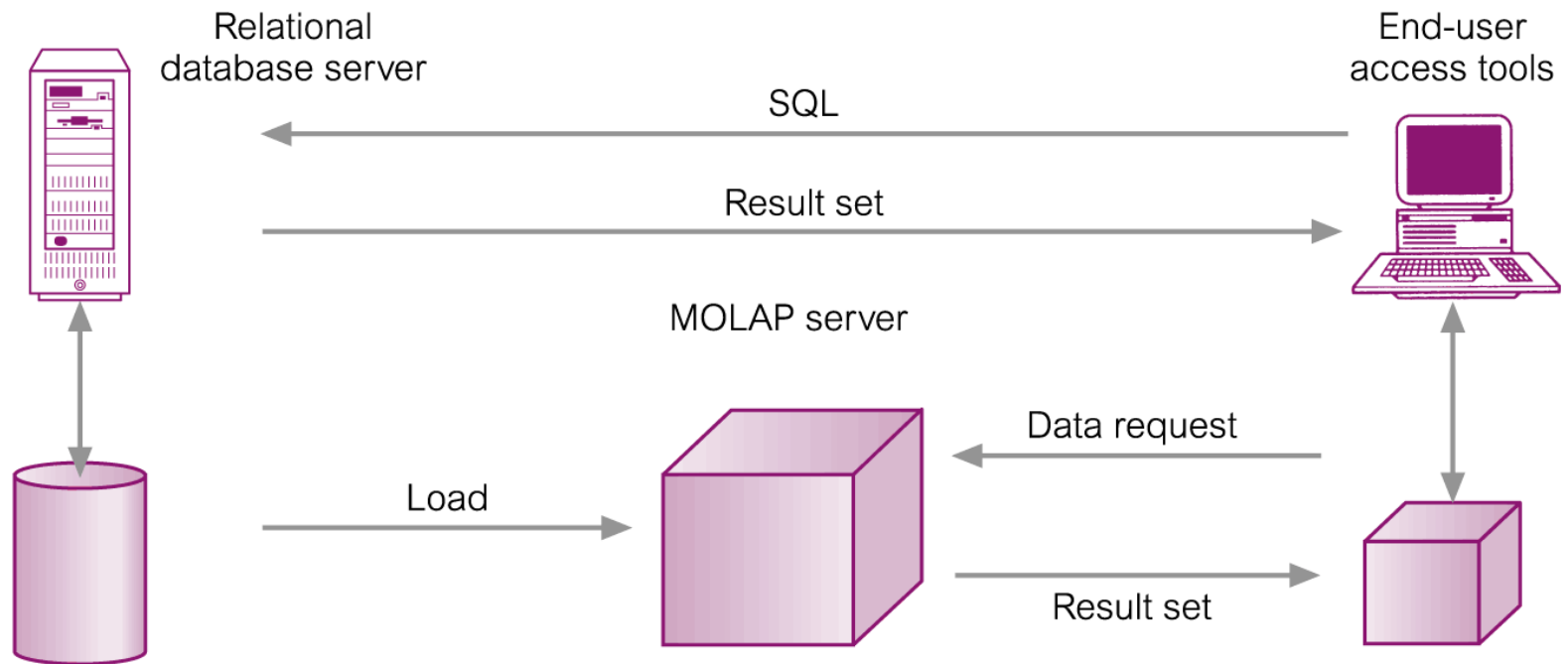
- Need to create a **middleware to work** with multi-dimensional applications
  - The middleware must convert relational data structure to multi-dimensional data structure
- Performance problems for complex queries that require complex transformations from relational data

# Hybrid OLAP (HOLAP)

- Provide query support for both RDBMS and MDDDBMS
  - Query data directly from the RDBMS using SQL or via a MOLAP server in the form of a data cube
- May cause data redundancy and inefficient network usage
- can be thought of as a virtual database whereby the higher levels of the database are implemented as MOLAP and the lower levels of the database as ROLAP



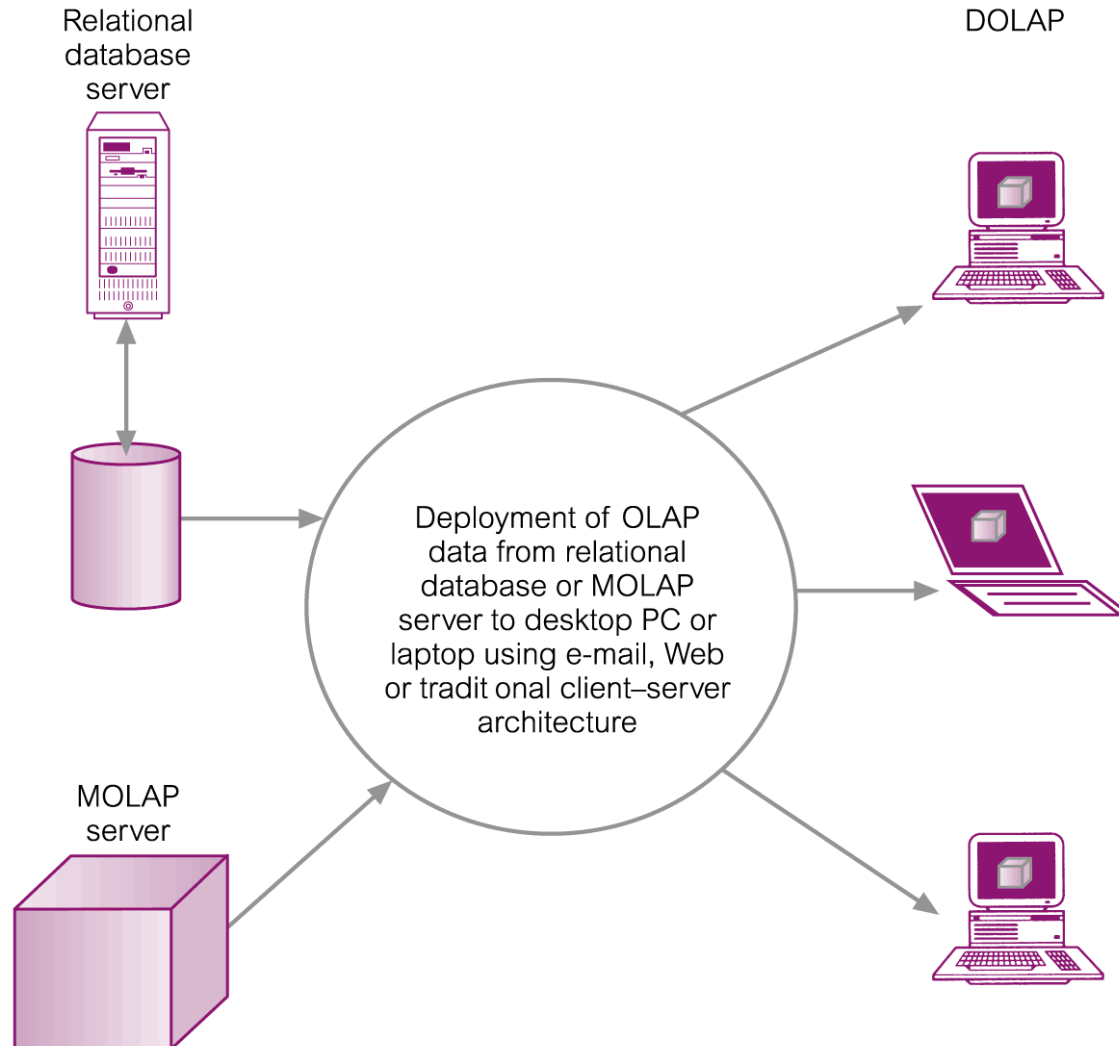
# HOLAP Architecture



# Desktop OLAP (DOLAP)

- Store and process the OLAP data on client side
- Data are held on client machines
  - Database may be distributed in advance, or created on demand (e.g. through the Web)
  - The maintenance of database is usually done by a central server
- DOLAP uses the power of desktop PC to perform multi-dimensional calculations
- DOLAP enables users to quickly pull together small cubes that run on their desktops or laptops

# DOLAP Architecture



# DOLAP Issues

- Security (access control) can be difficult
  - Can not utilize access control feature of DBMS
- Current trends are towards thin client machines
  - Complex calculations are increasingly moved to server machine rather than client machine

# OLAP Benchmark

- APB-1 (OLAP Council, 1998) is a standard for OLAP benchmark
  - Measurement of OLAP server performance
- APB-1 evaluates OLAP server performance for the following operations:
  - Loading of data
  - Aggregation of data
  - Complex Calculations
  - Time series analysis
  - Complex Queries
  - Drill-down through hierarchies
  - Multiple online sessions
  - etc.

# OLAP Benchmark (2)

- A benchmark metric used by APB-1 is AQM (Analytical Queries per Minute)
- AQM measures the number of analytical queries that an OLAP server can process per minute
  - The time is measured from when the data is loaded until the results are returned to user

# OLAP Extensions to SQL

- SQL has limited capability to support complex management queries
- ANSI adopted a set of OLAP functions as an extension to SQL
  - IBM and Oracle jointly proposed these extensions in 1999 as part of the current SQL standard
- The extensions are referred to as the 'OLAP package':
  - Feature T431, 'Extended Grouping capabilities'
  - Feature T611, 'Extended OLAP operators'