WQD7005 Data Mining Course Matters

Mid-term Test

- Assessment Type: Online Test
- Date/Time: Week 7 (Saturday) 8pm-11pm
- Duration: 3 hours
- Question types: 100 MCQ + 1 Short Essay
- Topics Covered:
  – Chapter 1: Introduction to Data Mining
  – Chapter 2: Data Warehouses
  – Chapter 3: Data Preprocessing

Note: Online revision on Week 7
(Saturday) 3pm-6pm

Data Warehouses

Chapter 2 (Part 2)

Previously

- Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.
- Many organisations have invested in data warehouses to support and improve business decision-making
- OLTP and OLAP systems are distinguishable with key features
- Separation of Data Warehouses from Operational Databases facilitates efficient processing

Outline

- Multidimensional Data Model
- Conceptual Modeling
  - Star Schema
  - Snowflake Schema
- Data Cube
- OLAP Operations
- Architecture

From Tables/ Relations 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** allows data to be modeled and viewed in multiple dimensions with a central theme (e.g. sales)
  - Dimensions are perspectives in which data records are kept
    - » For sales data, the dimensions may include time, item, branch, location, etc.
    - » Dimension tables contains further description of associated dimensions
      - e.g. item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Facts are numeric measures / quantities used to analyze relationships between dimensions
    - » For sales data, the facts can be dollars_sold (sales amount in dollars) or units_sold (number of units sold)
    - » Fact table contains names of the measures and keys to each of the related dimension tables

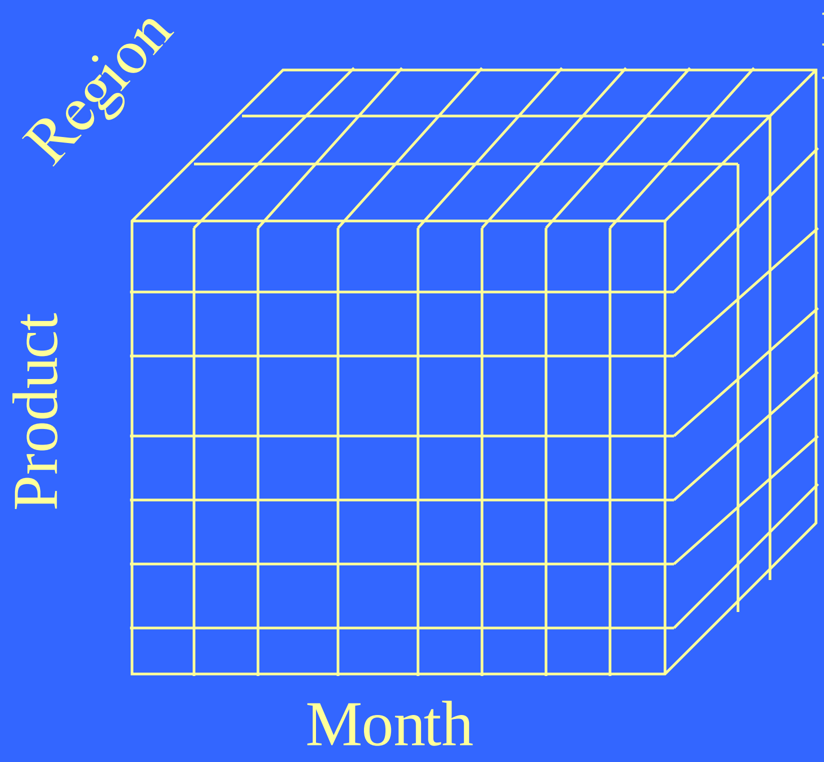
Relational View of Data/ Spreadsheet

ProdID	LocID	Date	Quantity	UnitPrice
123	Dallas	022900	5	25
123	Houston	020100	10	20
150	Dallas	031500	1	100
150	Dallas	031500	5	95
150	Fort Worth	021000	5	80
150	Chicago	012000	20	75
200	Seattle	030100	5	50
300	Rochester	021500	200	5
500	Bradenton	022000	15	20
500	Chicago	012000	10	25

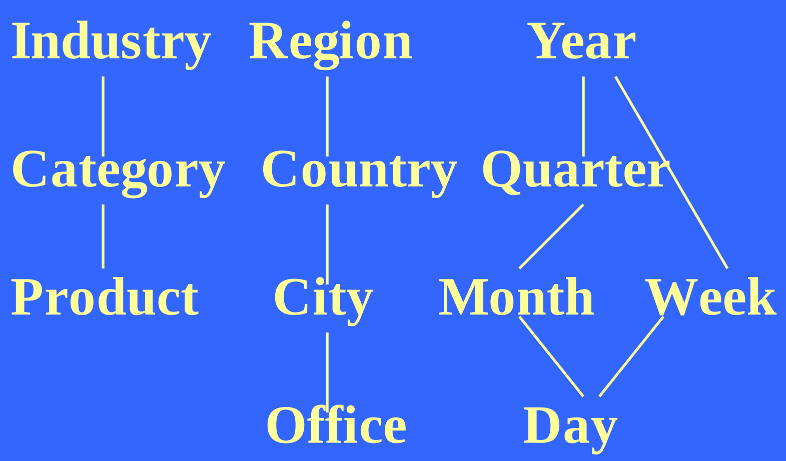
1

Multidimensional Data

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



Dimensions: Product, Location, Time
Hierarchical summarization paths



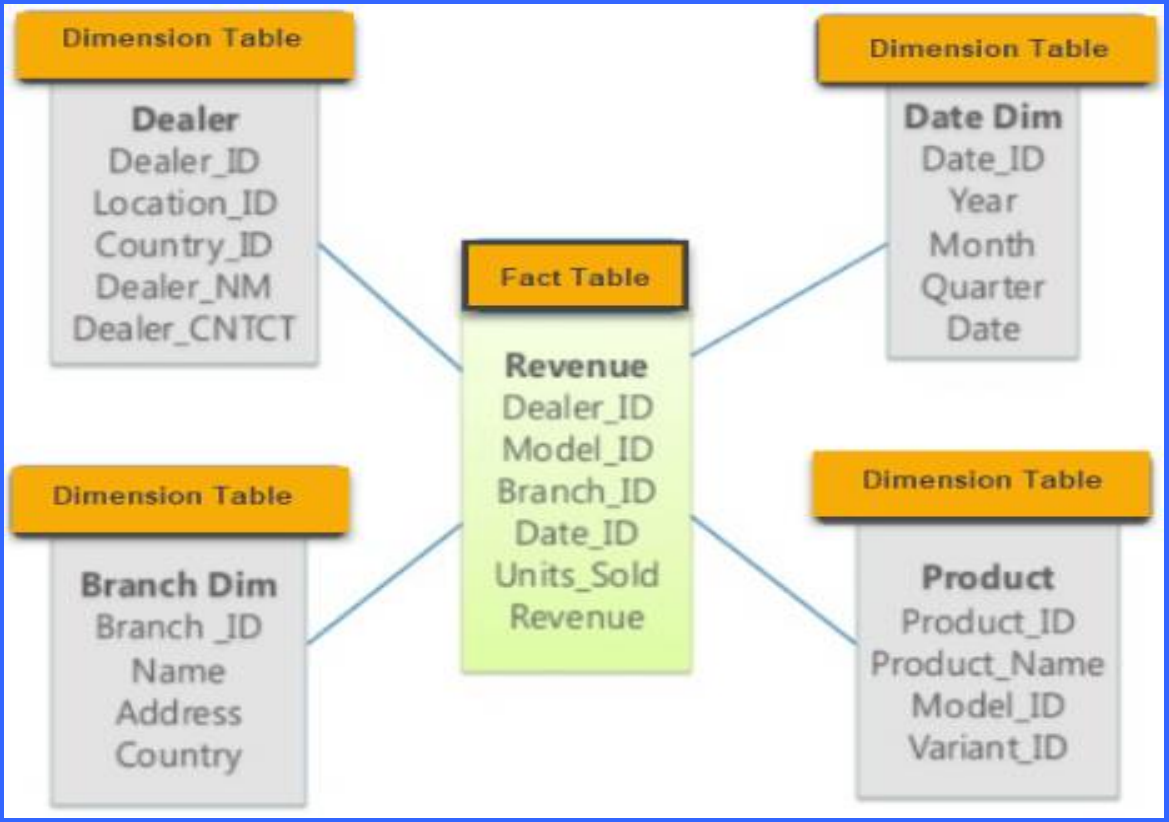
Concept hierarchy:
sequence of mappings from a set of low-level concepts to higher-level and more general concepts

Conceptual Modeling of Data Warehouses

- Modeling data warehouses requires concise, subject-oriented schema that facilitates online data analysis
- Models for relating 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 constellation: Multiple fact tables to share dimension tables, also called a galaxy schema

Star Schema

- Most common modeling paradigm
- Each dimension represented by one table
- Each table has a set of attributes
- Dimensional modelling

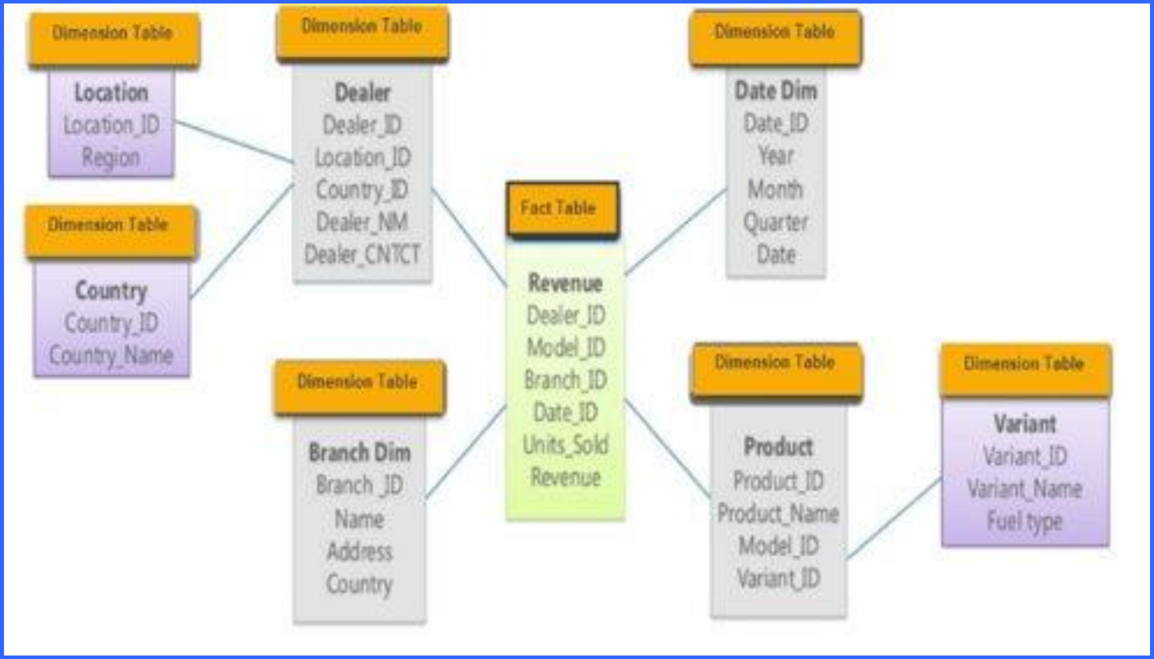


Star Schema: Main Characteristics

- **Simple structure**: easy to understand schema
- Great **query effectiveness**: small number of tables to join
- **Relatively long time** of loading data into dimension tables: de-normalization, table size may be large due to redundant data
- The **most common** data warehouse implementation: widely supported by a large number of business intelligence.

Snowflake Schema

- Variant of star schema
- Dimension tables are normalized (split into additional tables)
- Normalized form tables reduce redundancies
- Normalized modelling



Snowflake Schema: Characteristics

- **Easy to maintain** and **save space**: less redundancy in normalized dimension tables
- **Reduced browsing effectiveness**: more joins needed to execute a query
- **Less used** implementation: system performance not as optimized and space saving can be negligible in comparison to the typical magnitude of the fact table

Star vs. Snowflake

Star Vs Snowflake Schema: Key Differences

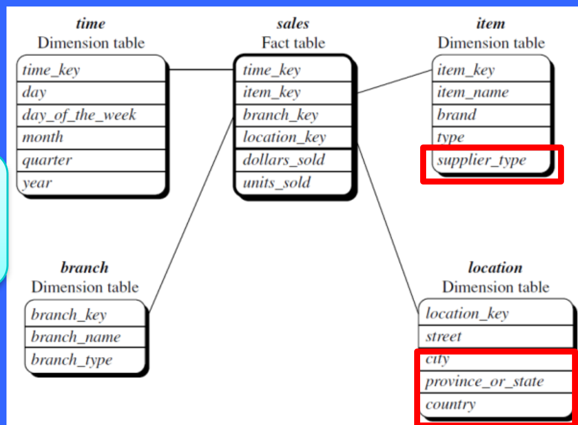
Star Schema	Snow Flake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions.

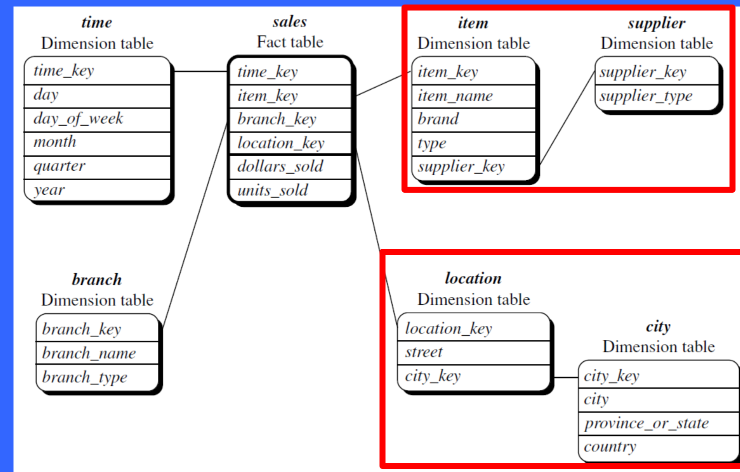
Fact Constellation

- Multiple fact tables sharing dimension tables
- Commonly used in data warehouse that collects information about subjects across entire organization (enterprise-wide scope)
- Star and snowflake is more common in department-wide scope (data marts, a subset of enterprise warehouse) to model single subjects

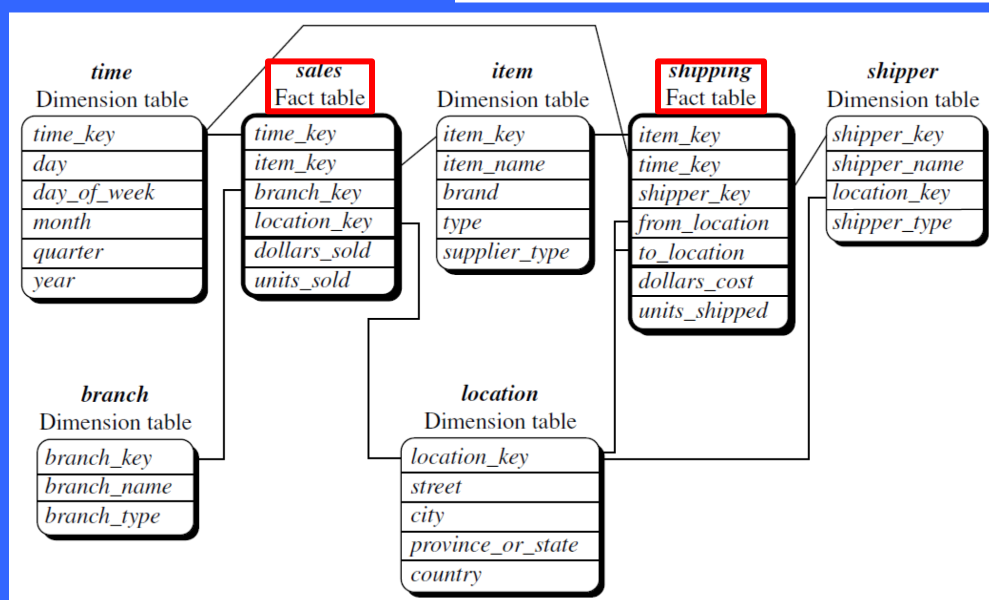
Schema Comparison

Star Schema



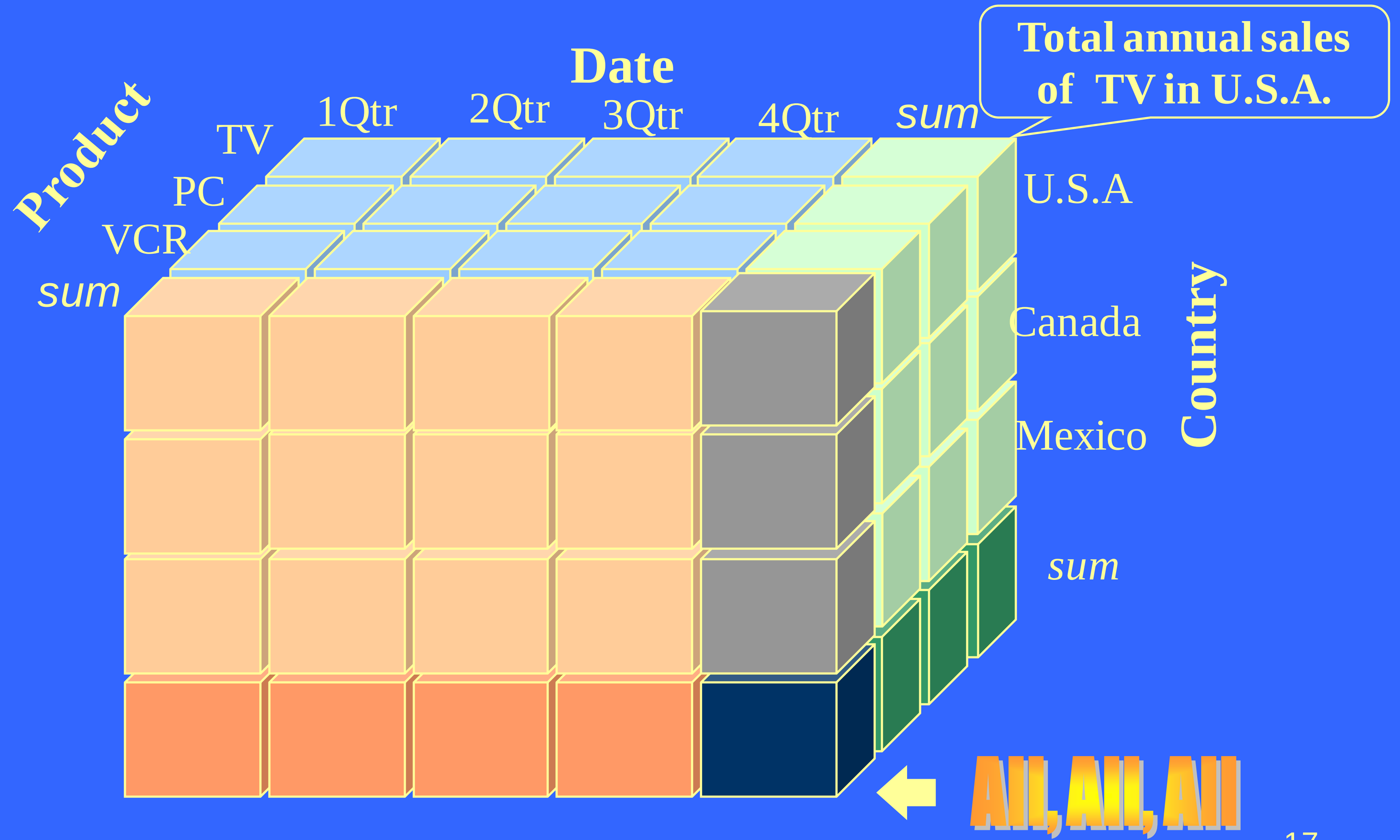


Snowflake Schema



Fact Constellation

A Sample Data Cube



Measures Categorization and Computation

- A multidimensional point in the data cube can be defined by a set of dimension-value pairs
- Measures is a **numeric function** that can be evaluated at each point in the data cube space by aggregating corresponding dimension-value pairs
- 3 categories of measures based no the aggregate function used:
  - Distributive: computed in a distributed manner; e.g. sum()
  - Algebraic: computed by algebraic function with arguments that are obtained from distributive functions; e.g. avg() that is computed from sum() and count()
  - Holistic: no algebraic function that characterizes the computation; e.g. median(), mode(), rank()

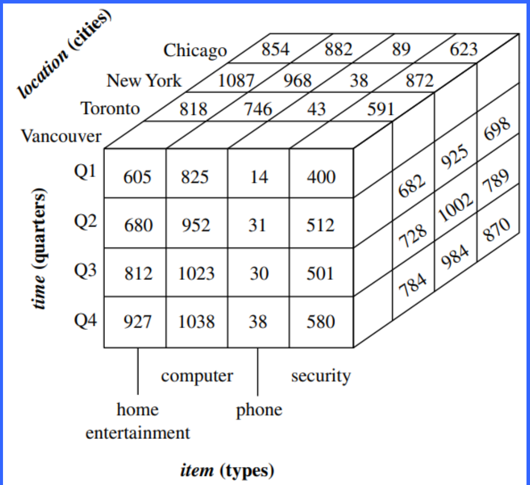
2D to 3D Data Cubes

- Tables or spreadsheets can be considered as a simple 2D "data cube": D1= time, D2 = item, Fact = dollars_sold (in thousands)

	location = "Vancouver"			
	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

- Adding more dimensions can form a 3D cube
  - 3D example: D1= time, D2 = item, D3 = location, Fact = dollars_sold

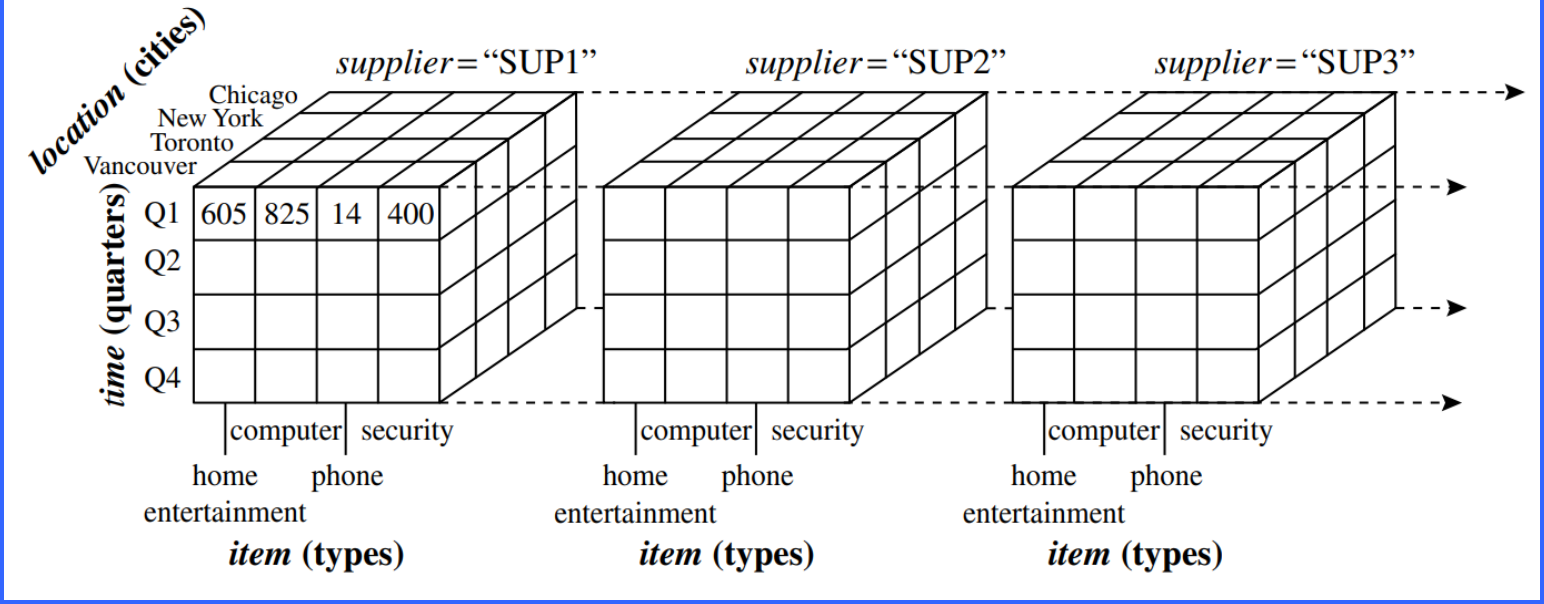
	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580



Data Cube

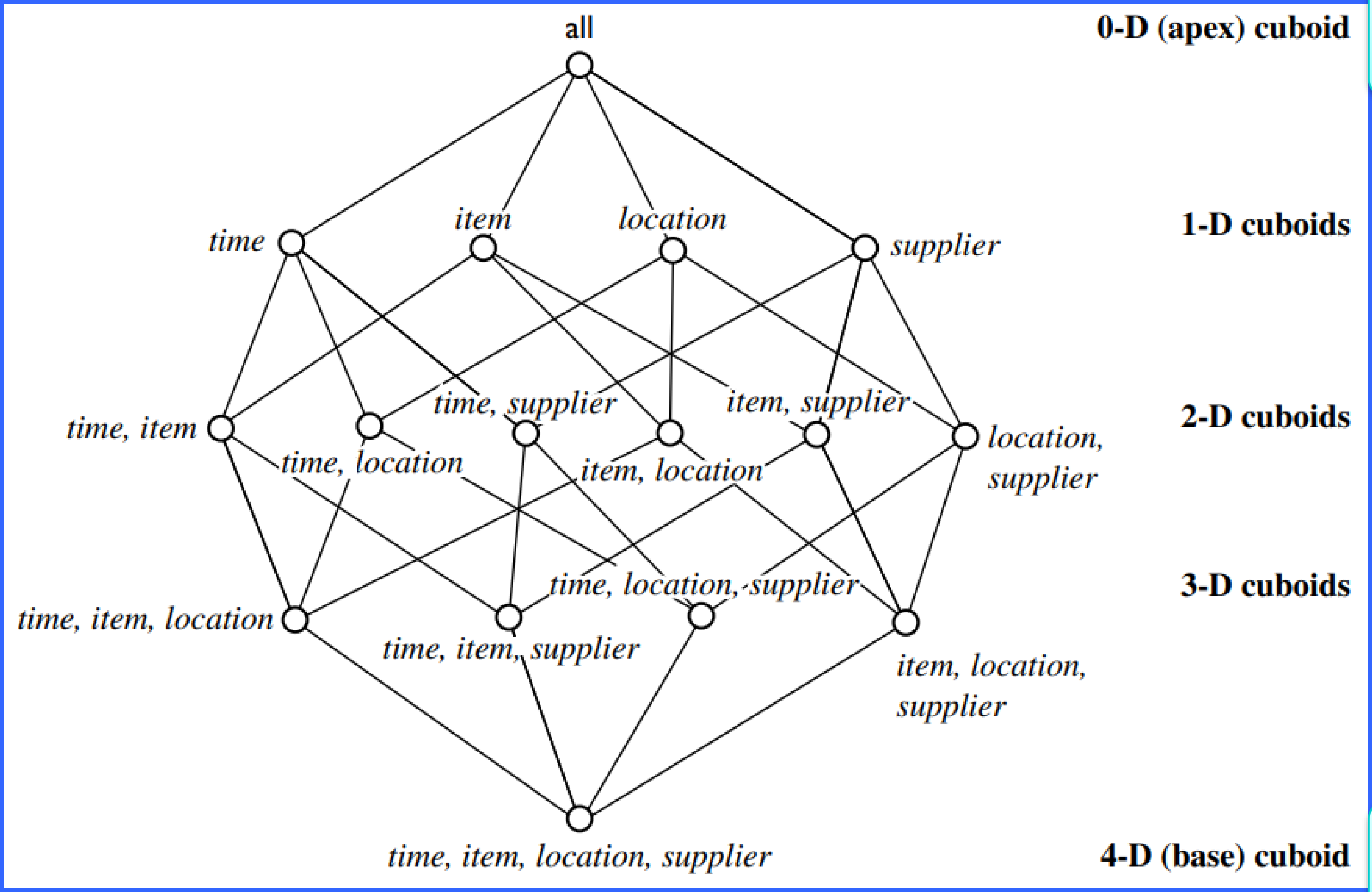
- Data cube in data warehouses are *n*-dimensional, not only 3D
  - 4D example: D1= time, D2 = item, D3 = location, D4 = supplier,

    Fact = dollars_sold



- Data cube is a metaphor for multidimensional storage
  - In some research, these cubes are referred to as a cuboid
  - Given a set of dimensions, multiple cuboids can be generated from all possible subsets to form a lattice of cuboids (a.k.a. a data cube)

Lattice of Cuboids (4D Data Cube)

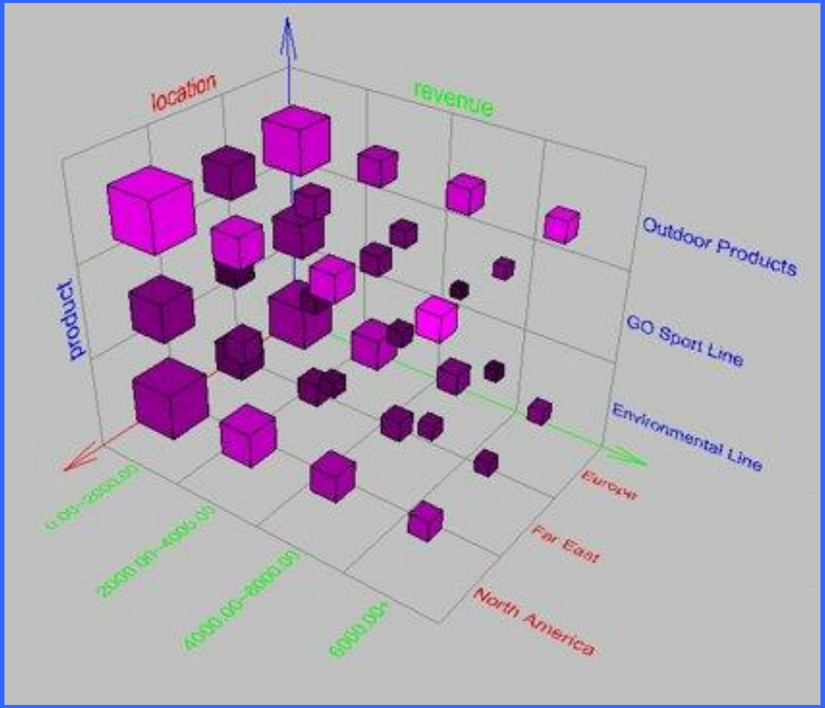


Highest level of summarization

Lowest level of summarization

Browsing a Data Cube

- Multidimensional model provides flexibility to view data from different perspectives
- OLAP provides user-friendly environment for interactive data analysis

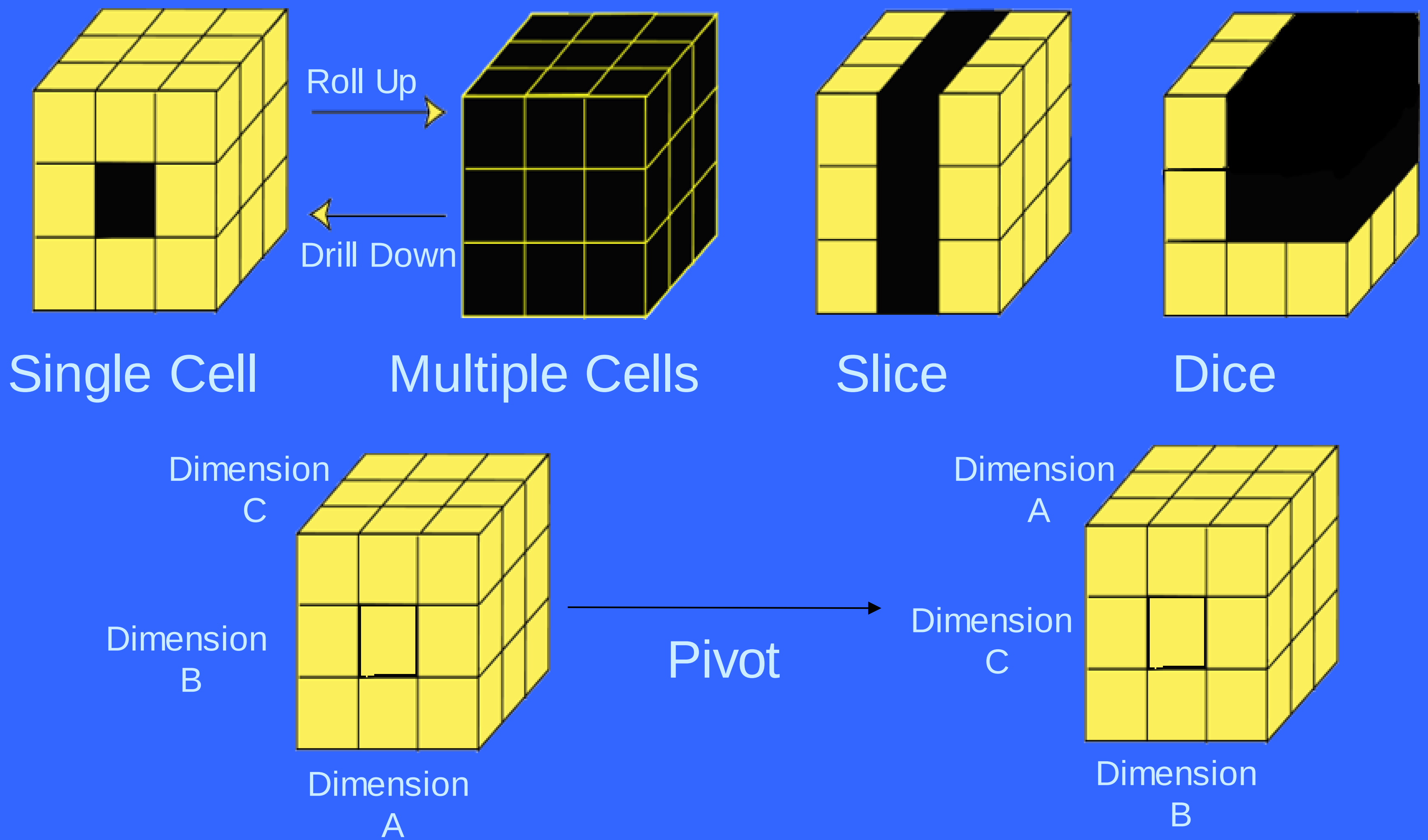


- Visualization
- OLAP capabilities
- Interactive manipulation

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *Climbing up hierarchy*: aggregate or group data into "larger" concept
  - *Dimension reduction*: remove a dimension
- Drill down (roll down): reverse of roll-up
  - *From higher level summary to lower level summary*: get more detailed data
  - *Introduce new dimensions*: add more details / dimension
- Slice and dice: project and select
  - Slice: select one dimension from the cube (subcube)
  - Dice: select two or more dimension to create subcube
- Pivot: rotate data axes
  - Visualization operation that rotates the view to provide alternative representation
  - Rotate cube, or transform 3D cube into series of 2D planes

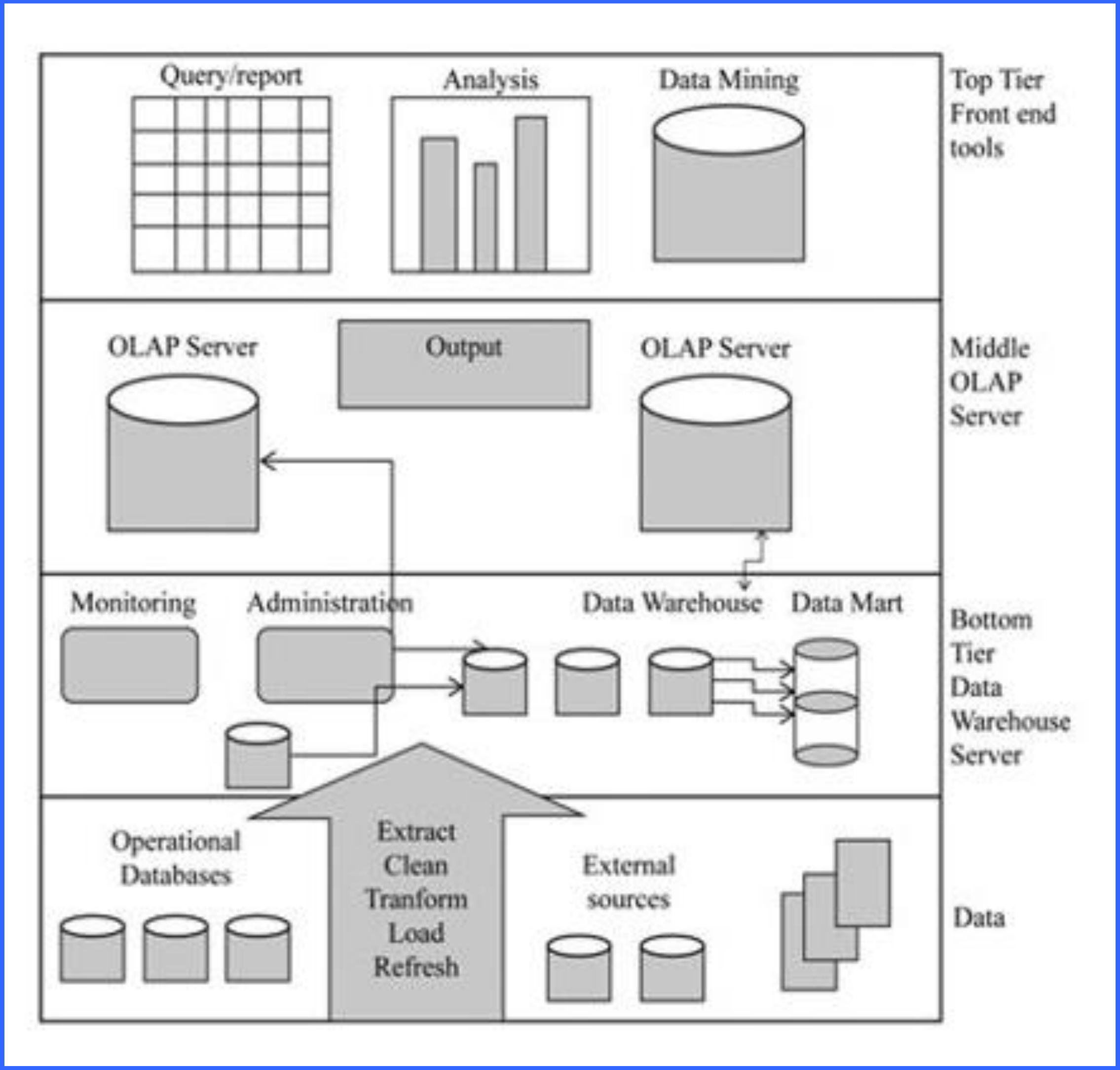
OLAP Operations



Other OLAP Operations

- Additional drilling operations offered by some OLAP systems
  - **Drill-across**: query involving more than one fact table
  - **Drill-through**: using relational SQL facilities to drill through the bottom level of a data cube down to back-end relational tables
- **Ranking** top *N* or bottom *N* items
- Compute moving averages, growth rates, interests, currency conversions, etc.

Data Warehousing Architecture



- Often adopt three-tier architecture

Multitiered Architecture

- Bottom tier: Warehouse database server
  - Almost always a **relational database systems** where data is fed by back-end tools and utilities
    - » ETL (Extract/Transform/Load) process
  - Data extracted using application program interfaces (gateways) supported by DBMS and client programs
    - » Gateway examples: Open Database Connection (ODBC), Object Linking and Embedding Database (OLEDB), Java Database Connection (JDBC)
  - Information about the data warehouse and its contents are stored in this tier (**metadata repository**)

Multitiered Architecture

- Middle tier: OLAP Server
  - 3 types of implementations
    - » Relational OLAP (ROLAP) model: extended relational DBMS that maps operations on multidimensional data to standard relational operations
    - » Multidimensional OLAP (MOLAP) model: special-purpose server that directly implement multidimensional data and operations
    - » Hybrid OLAP (MOLAP) model: combination of ROLAP and MOLAP
- Top tier: Front-end client layer
  - Contains query and reporting tools, analysis tools, and/or data mining tools (for trend analysis, prediction, etc.)

Summary

- Data warehouse modelling schemas allow the relation of dimensions and measures of data where each schema has their trade-offs
- Data warehouses are based on a multidimensional data model in the form a n-dimension data cube that allows data to be viewed from various perspectives
- OLAP provides user-friendly environment for interactive data analysis with operations to summarize, drill-down, etc. To support analysis
- Data warehousing often adopts a 3-tier architecture to support organizational data management and analysis

Tutorial 2

- What is the difference between a data warehouse and a traditional database system?
- Explain the concept of dimensional modelling and how it differs from normalized data modelling.
- What is ETL? Describe the key steps involved in an ETL process.
- What are the benefits of data warehousing for an organization?
- What are the different types of data that can be stored in a data warehouse?