CSCI 5408 Data Analytics: DM and DW Tech (Week 9)

- Ass4 Due: Mar 14
  - Brightspace: Assignment 4, Tutorial slides, etc.
  - **Help hours this week:** Wed/Thu/Fri, 1:00-2:30PM
    - ➢ Mar 8th, 1:00-2:30PM, CS 134
    - ➢ Mar 9th & 10th, 1:00-2:30PM, CS 233
- Write answers for review questions
  - Final Exam: Apr 20, 3:30-5:30 PM
- Reading: Lectures: 13-14, Text: Ch4 of 3rd edition, or Ch3 of 2nd edition

(*Apr 8, 11:30 – Industrial Showcase Seminar: Integrity Services at Services Canada)

3. Data Warehouses and OLAP

(Textbook: Ch4 of 3rd edition, or Ch3 of 2nd edition)

- Objectives of DW/OLAP
- What is a DW?
- Multi-dimensional data space model
- DW schemas
- OLAP operations
- Aggregations
- DW architecture
- From DW to DM

Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Recap: Challenges to Conventional Information System (OLTP): DBMS

- Information system as DBMS
  - A database management system (DBMS) is also called information system since it is mainly for storing the information about a business, and answering specific questions by retrieving the stored information, i.e. table cell values, or simple derived values.
- The information to be retrieved can be a piece of fact, or an aggregation result of a group of facts.
  - E.g., From the Statistics Canada DB, query questions:
    - How many people in NS are 80, or older in 2016?
    - What is the population of baby bummers (borned from 1945 to 1965) in Nova Scotia?
    - What is the average starting salary of CS Bachelor graduates in Halifax in 2016?
  - Each aggregation result represents a piece of analytic information.
- When many aggregations are needed, from multiples sources: DW & OLAP tools are needed.
  - E.g. The analysis of the Wealth and Health of 200 Countries, over 200 Years (http://www.youtube.com/watch?v=jbkSRLYSojo)

A historical global economic trends analysis:

Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

(An animation with 120000 aggregated values from different data sources)

https://www.youtube.com/watch?v=jbkSRLYSojo

Drivers of DW Technology

How to quickly provide answers to various ad hoc queries of large corporations?

- **Problem:** Data rich, but information poor with conventional DBMS solutions:
  - Corporate data assets are increasingly dispersed among hundreds or even thousands of different platforms throughout the enterprise.
  - Distributed DBMS (DDBMS) solution failed to achieve <u>a single enterprise-wide data management layer </u>which would provide various types of transparency (e.g., transparencies of locations, platforms, and data formats), and treat these physically dispersed stores of data as if they were really a single logically centralized, and homogenous database.
  - For example, a single query could be executed against the DDBMS layer that would, using its own directory and metadata information, determine that three Different databases would need to be accessed at execution time to <u>merge and organize the requested information and present the combined results</u> back to the user or requesting application.

Example: Wal-Mart Data Warehouse

Data rich, but information poor with conventional DB solutions

The DBMS technology is mainly designed to handle day-to-day based operations which have very limited power of retrieving decision oriented information.

- Noise and redundant data
- Lack of integration, from different tables (high cost), different DBs
- Structure limitation of viewing at multiple levels
- Difficult for discovering hidden knowledge

Analytical information: statistical summaries (information data) of various groupings about business subjects

Example, Wal-Mart Data Warehouse:

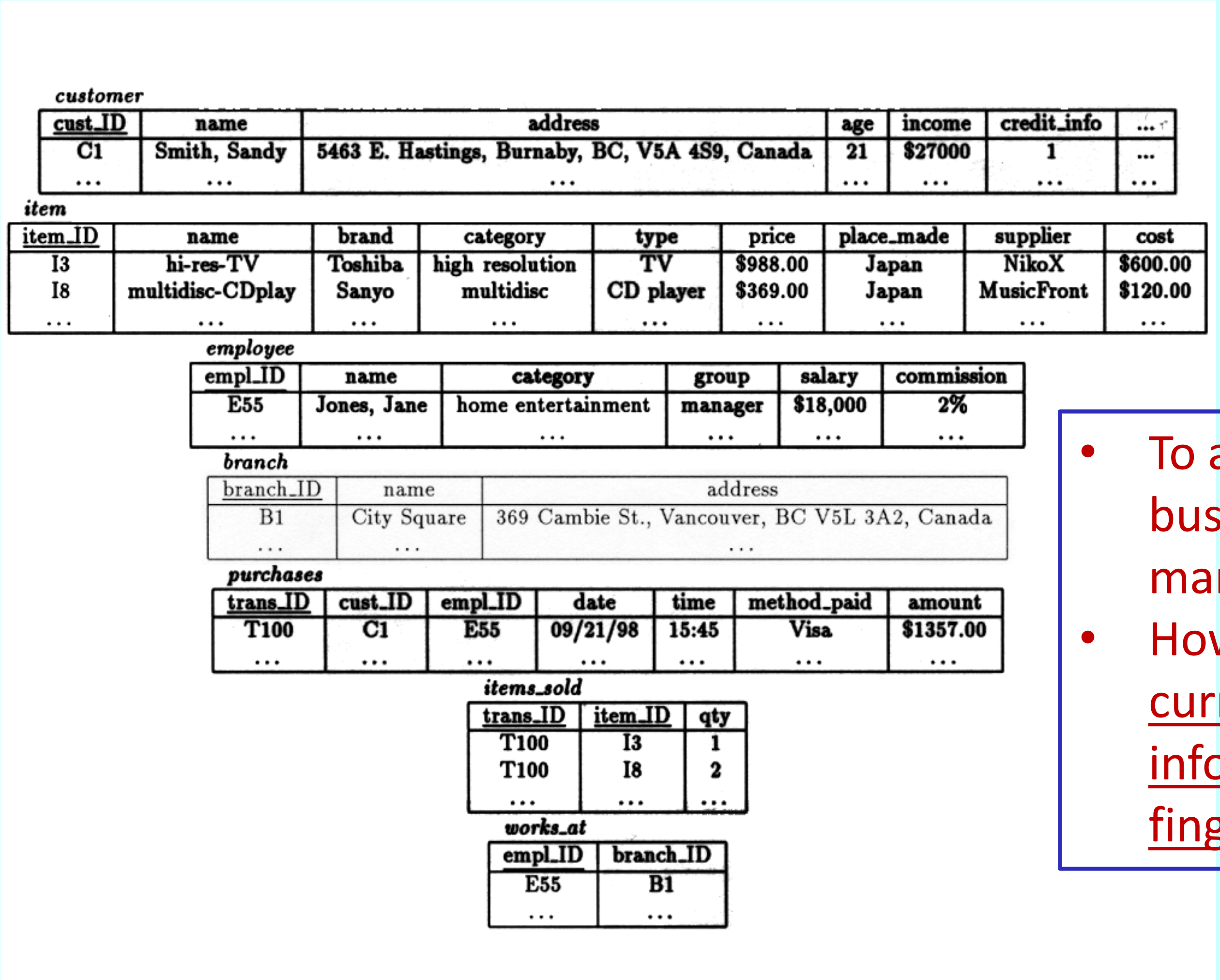
-Wal-Mart's Data Warehouse (2006): http://derbaum.com/tu/WalMarts%20DWH.pdf
-Wal-Mart Enhances Data Warehouse (2010):
http://consumergoods.edgl.com/news/Wal-Mart-Enhances-Data-Warehouse57129

Recap: Business Management Queries

E.g. Ad hoc queries about orange juice sales by various retailer stores:

1. How much orange juice sold in 2016, or in last month, in last week, … in store X, …?
2. Compare sales of in various stores, sold in different months, …
3. Find the most popular orange juice product in 2016, rank all orange juice products sold in store X, …
4. Who bought orange juice last year, last month, last week?
5. What internal factors (e.g. position in store, advertising campaigns, ...) influence orange juice sales?
6. What external factors (e.g. weather, ...) influence orange juice sales?
7. How much orange juice are we going to sell next week, next month, next year?
8. What is the suppliers price of orange juice last year, this year, next year?
9. How can we help suppliers to reduce their cost?
10. What are the shipping/stocking costs of orange juice to/in store X?
11. How can suppliers help us reduce those costs?
12. *Are we doing the best job position ourselves in selling orange juice in the marketplace?

Recap the operational DB: The electronics retailer business, such as the store *Best Buy*:



customer

cust_ID	name	address	age	income	credit_info	...
C1	Smith, Sandy	5463 E. Hastings, Burnaby, BC, V5A 4S9, Canada	21	$27000	1	...
...	...	...	...	...	...	...

item

item_ID	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	multidisc-CDplay	Sanyo	multidisc	CD player	$369.00	Japan	MusicFront	$120.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	$18,000	2%
...	...	...	...	...	...

branch

branch_ID	name	address
B1	City Square	369 Cambie St., Vancouver, BC V5L 3A2, Canada
...	...	...

purchases

trans_ID	cust_ID	empl_ID	date	time	method_paid	amount
T100	C1	E55	09/21/98	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	item_ID	qty
T100	I3	1
T100	I8	2
...	...	...

works_at

empl_ID	branch_ID
E55	B1
...	...

- To answer various ad hoc business queries of management teams
- How to have the most current, integrated information possible at their fingertips?

Recap: Basic SQL Clause Structure

- SQL is based on set and relational operations with certain modifications and enhancements
- A typical SQL query has the cause form:

  SELECT $A_1, A_2, \ldots, A_n$
  FROM $r_1, r_2, \ldots, r_m$
  WHERE P

  - A_is represent attributes
  - r_is represent relations
  - P is a predicate.

- What the two major limitations to have integrated information via convent DBMS?

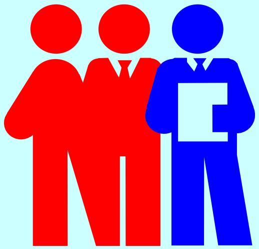
- This query is equivalent to the relational algebra expression:

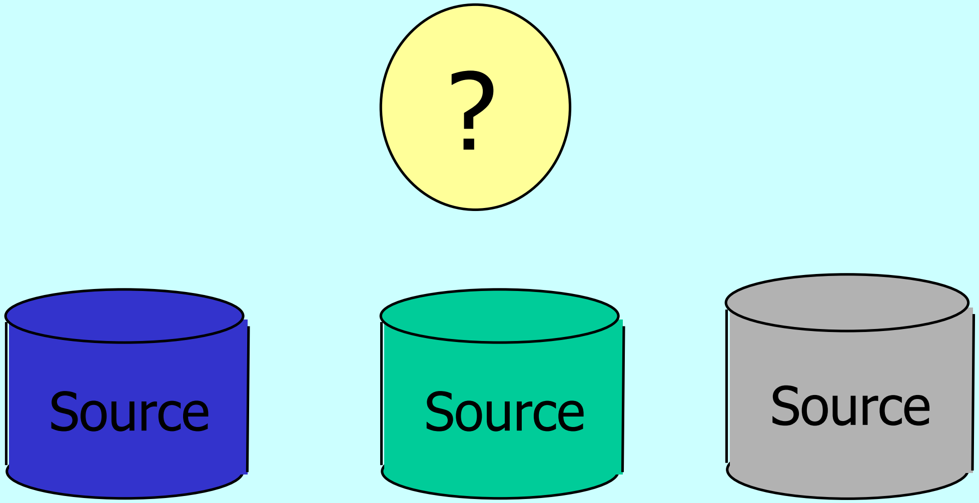
$$\boxed{\pi_{A1,\ A2,\ \ldots,\ An}(\ \sigma_P(r_1\ \times r_2\ \times\ \ldots \times\ r_m)\ )}$$

- The result of an SQL query is a new relation, i.e. a result table

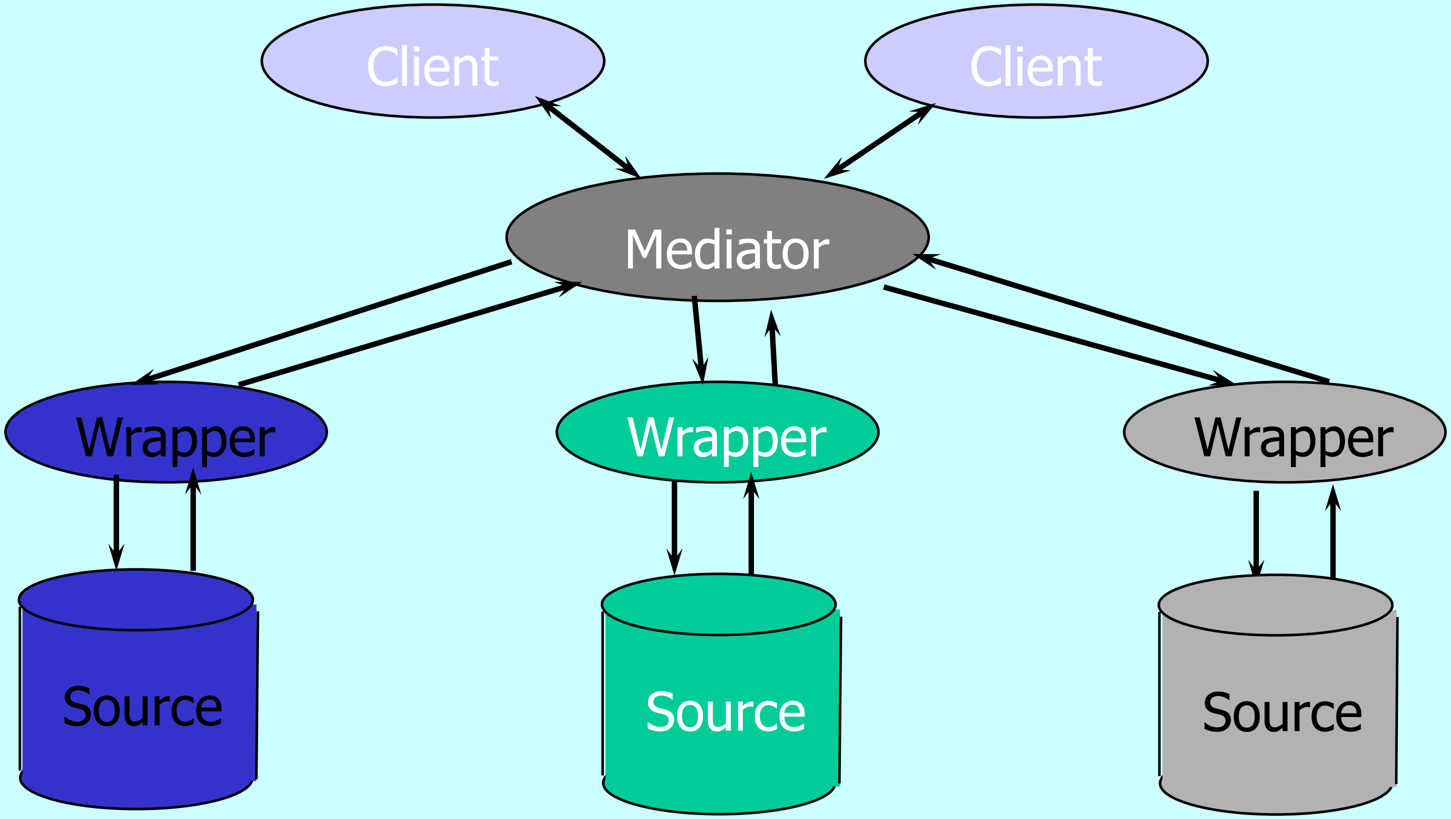
Analytical Information Retrieval Options for Enterprise Level Analysis?

- Two Approaches:
  - DB Query-Driven (Lazy)
  - Data Warehouse/OLAP (Eager)

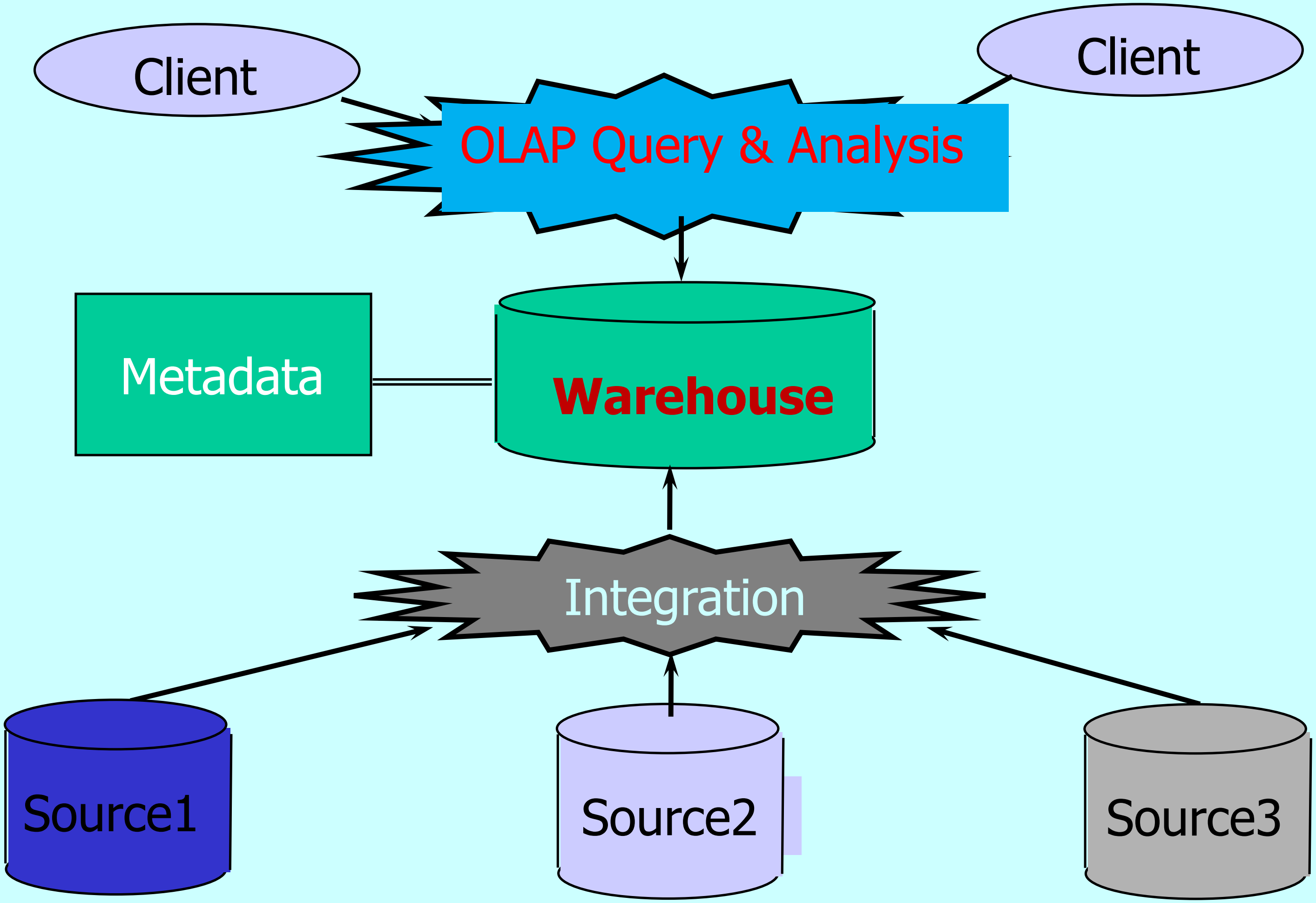




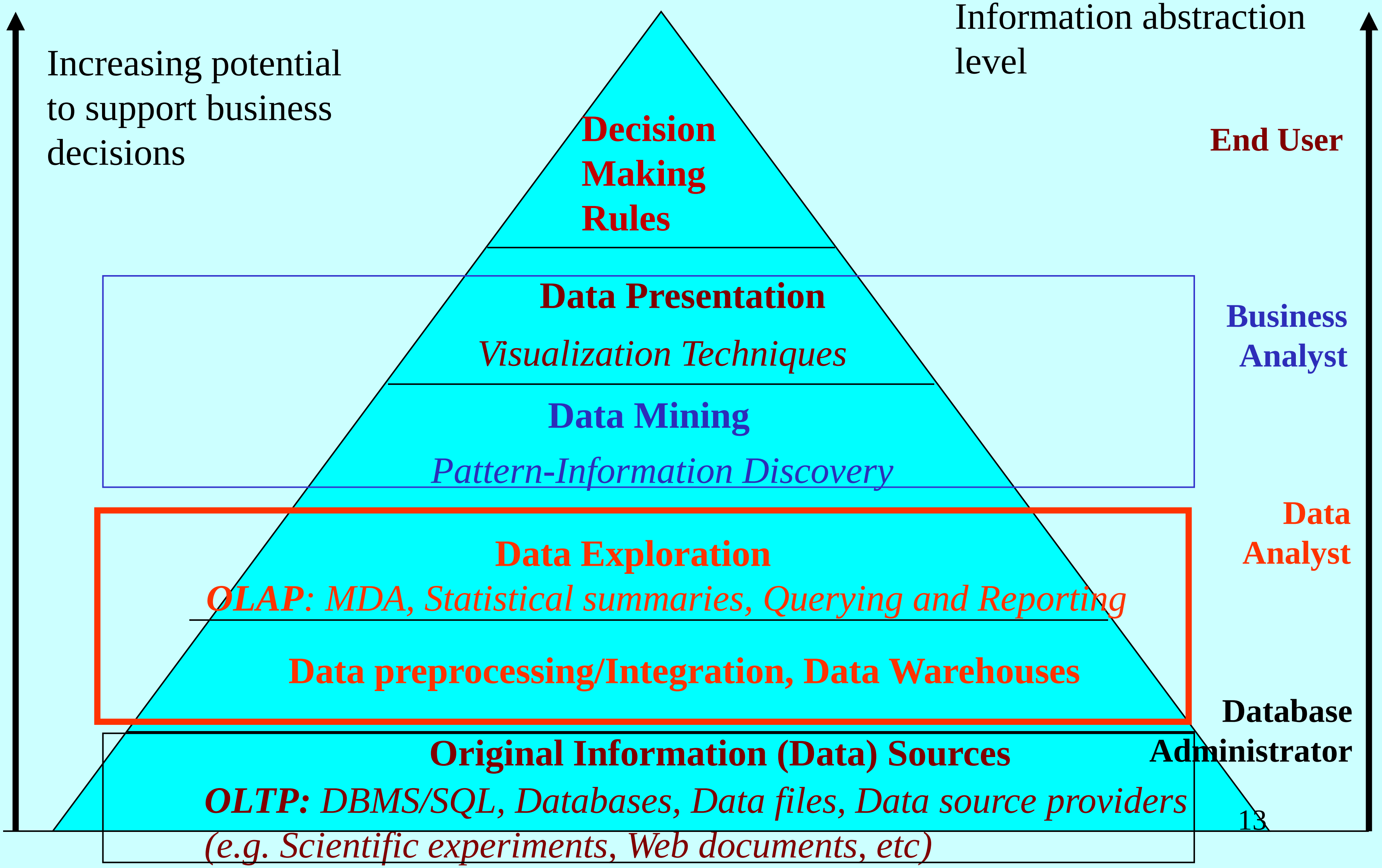
DB Query-Driven Approach



DW/OLAP based Approach



Recap: View of DSS: Business Intelligence

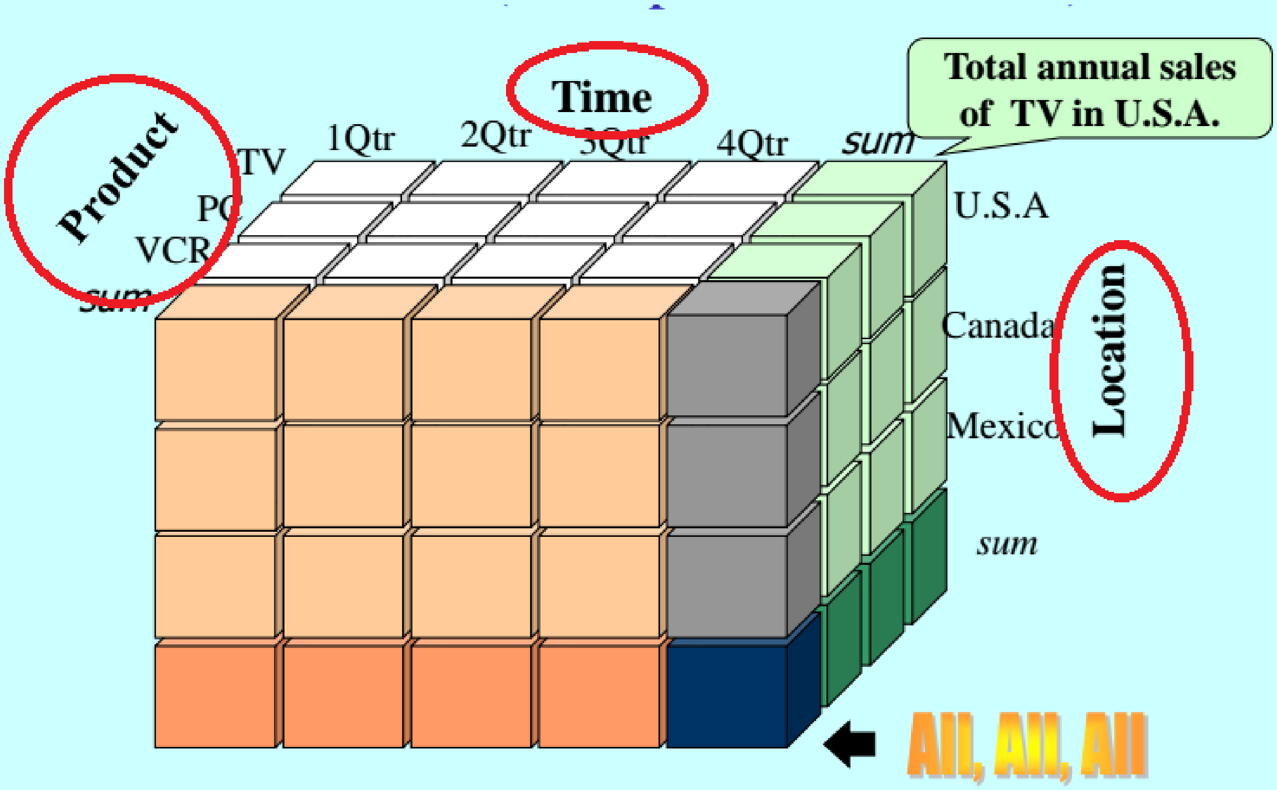

Increasing potential to support business decisions
Information abstraction level
Decision Making Rules
End User
Data Presentation
Visualization Techniques
Business Analyst
Data Mining
Pattern-Information Discovery
Data Exploration
Data Analyst
OLAP: MDA, Statistical summaries, Querying and Reporting
Data preprocessing/Integration, Data Warehouses
Database Administrator
Original Information (Data) Sources
OLTP: DBMS/SQL, Databases, Data files, Data source providers (e.g. Scientific experiments, Web documents, etc)

Recap: Types of Information and Information Process for DSS

- **On line transactional (information) process (OLTP)**
  - Track/record/retrieve original data records of every day business operations for answering *“what, when, where”* type of questions: **Operational databases** (Relational DB and SQL)
- **On line analytical (information) process (OLAP)**
  - Store & manipulate summaries of various groupings of original data records for answering *“what happened to the business”* type of questions: Analytical databases: **Data warehouses and OLAP**
- **Data Mining: knowledge discovery from data**
  - Discover/analyze hidden patterns of abstractive information (knowledge) for answering *“why and what to happen next”* type of questions: **Data mining**

DW: Multi-dimensional data model

- E.g., Data cube with the subject "**Sales**"



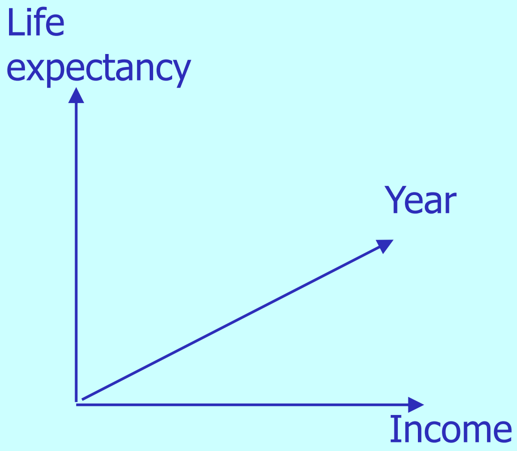
DW Objectives

Organizations pursue DW/OLAP techniques to provide quick analytical information on ad hoc business queries

- **Analytical information:** statistical summaries (aggregated information data) of various groupings about business subjects
- **Aggregated information data:** consolidated, cleansed, staged/derived, ready for use
- **Efficient management on enterprise-wide analytical data:** A single centralized, and homogeneous DB storing aggregated information only
- **Focus on the usage of information/analytical side:** quickly generate aggregation based reports, analyze trends, etc.

The merit of DW

*“**…**a single repository for completely integrated, 360-degree view of your business -one version of the truth.”*



E.g. 2D vs. 3D OLAP based trend analysis:
“A trend analysis on the Wealth and Health of 200 Countries, over 200 Years” ?
https://www.youtube.com/watch?v=jbkSRLYSojo

Two Types of DW

- **Enterprise DW:**

  It contains subject oriented data spanning the entire organization. It provides corporate wide data integration.

- **Data mart:**

  It contains a single subset of department-wide data.

DW/OLAP: Support efficient ad hoc query based analysis for DSS

DW/OLAP provide a good solution to DSS in that an ad hoc query can be translated into a series of OLAP operations for forming a meaningful answer.

E.g., "Why are the sales for this year not meeting the targets?”

E.g., "Why are the sales for this year not meeting the targets?”

The question may be translated into a sequence of OLAP queries for aggregated factual information, such as

1. For each product or each category of products, what are the cumulative sales for the year?
2. Identify those products for which the actual sales are less than the targets?
3. Where are the turning points for those sales which do not match the targets?

…

What Is a Data Warehouse?

Defined in many different ways, but not rigorously:

- A decision support DB that is maintained separately from the organization’s operational DB
- Support aggregated 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 aggregated *information data* in support of management’s ad hoc queries and for decision-making process

“Subject-Oriented”

- **Business data is reorganized around its major subjects: (Examples)**
  1. **“Sales”, “Supplies”** and **“Customers”** for retailer stores.
  2. **“Grades”** (performance) for CS student data analysis (Doc/Theses/MACSprojReportOu04.pdf).
  3. **“Content access facts”** for the website www.cs.dal.ca (Doc/Theses/MACSprojTayyaba06.pdf).
  4. **“Procedures”** for the dentistry administration database (Faculty of Dentistry, Dal) (Doc/Theses/MHINthesisNariman05.pdf).
- It focuses on analysis (or interactively modeling) on subject oriented business information
- It provides simple and concise views around particular subject issues by excluding data that are not useful in the decision support process

“Integrated”

- **DW is constructed by integrating multiple, heterogeneous data sources:**
  - Relational DBs, flat files, on-line transaction records, etc.
- **Data cleaning, transformation, integration techniques are applied:**
  - Ensure accurate data with consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
  - When data is moved to the warehouse, it is converted

“Time Variant”

- **The time horizon for the data warehouse is significantly longer than that of operational systems**:
  - Operational database: **current data** (e.g. your data in a bank DB up to 3 months?)
  - Data warehouse data: provide information from a **historical perspective** (e.g., data of past 5 years)
  - **Data transfer** from the operational database to DW is an **ongoing process**
    - A retail store DW may need the process on a daily basis after the close of the regular business day.
    - The CS student DW is updated at the end of each semester.
- **Every factual measure in the data warehouse:**
  - Contains **time element** (explicitly or implicitly)
  - But operational data may or may not contain “time element”

“Non-Volatile”

- **DW is a physically separated store of data from any operational environments**
- **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:
    - **loading** of data and **access** of data

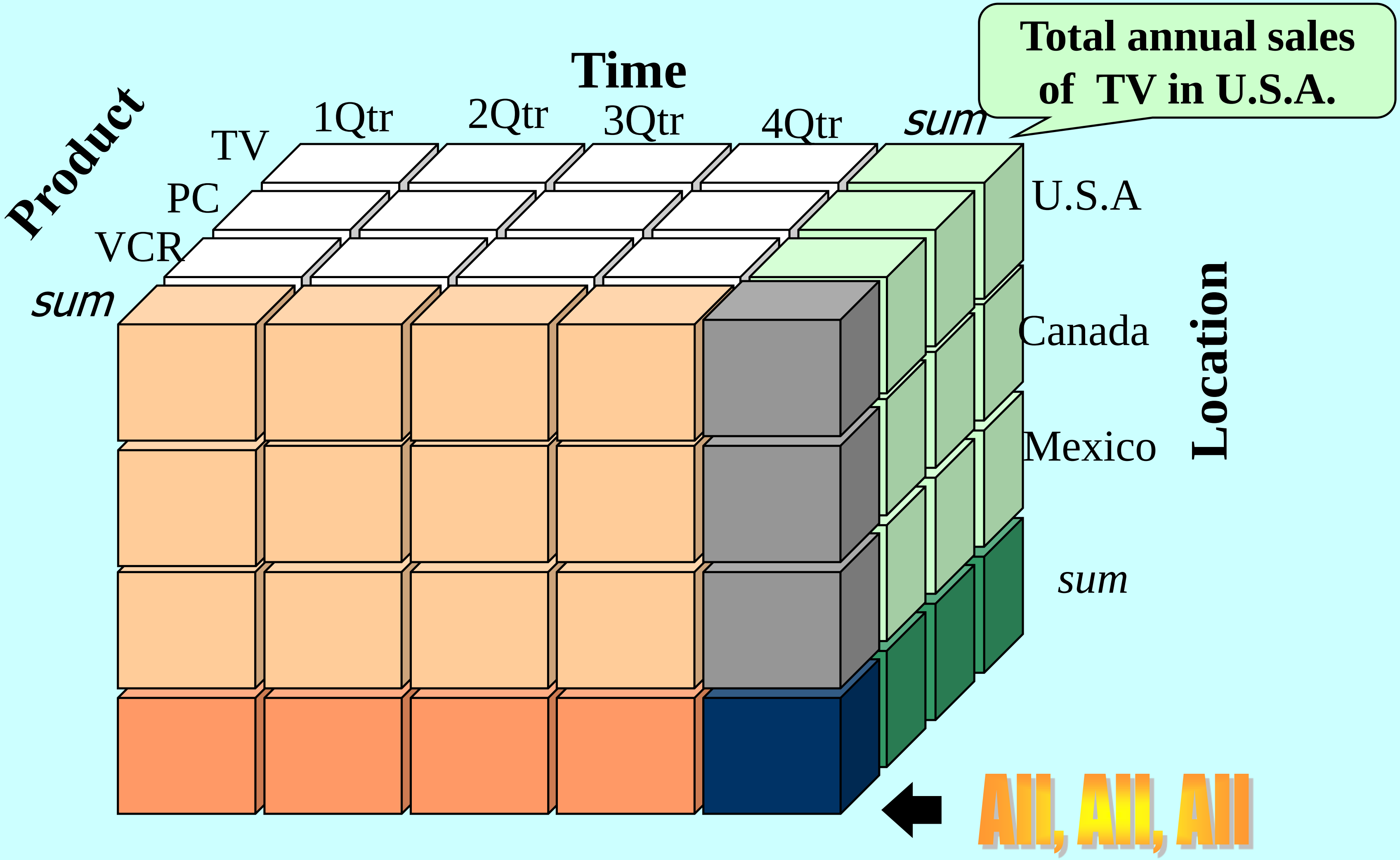
How to get business analytical information easily?

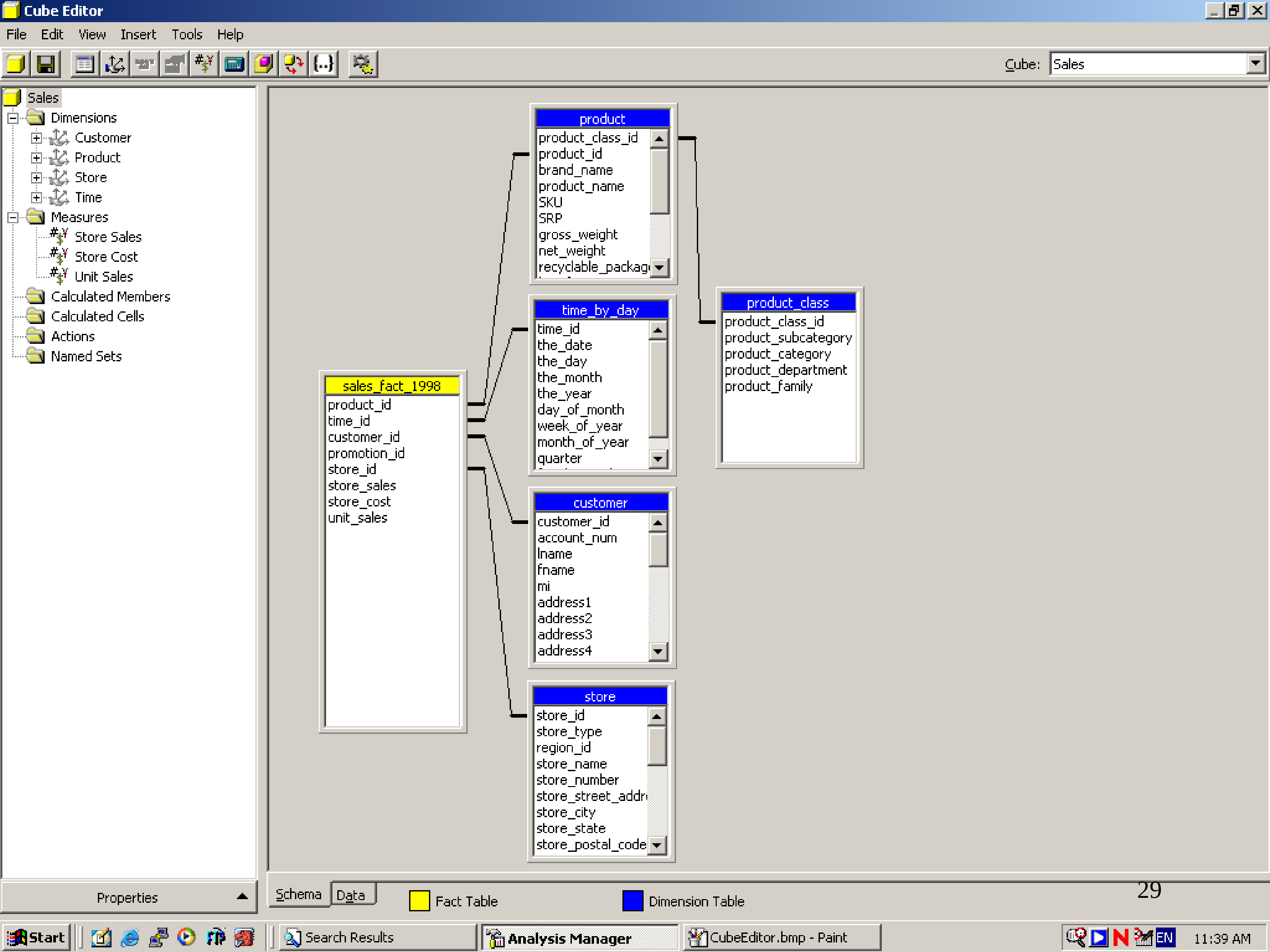
Main idea/approach:

- Store them (measures of groupings) in a multi-dimensional space (MDS).
- Use MDS manipulators (OLAP operators) to retrieve/view them in real-time.

Multi-dimensional software, unlike spreadsheets or SQL databases, is specifically designed to facilitate the definition and computation of sophisticated **multi-level aggregations and analysis** via OLAP operations.

Data cube: **sales** (time, product, location)





Sales
- Dimensions
  - Customer
  - Product
  - Store
  - Time
- Measures
  - Store Sales
  - Store Cost
  - Unit Sales
- Calculated Members
- Calculated Cells
- Actions
- Named Sets

sales_fact_1998
product_id
time_id
customer_id
promotion_id
store_id
store_sales
store_cost
unit_sales

product
product_class_id
product_id
brand_name
product_name
SKU
SRP
gross_weight
net_weight
recyclable_packag

time_by_day
time_id
the_date
the_day
the_month
the_year
day_of_month
week_of_year
month_of_year
quarter

customer
customer_id
account_num
lname
fname
mi
address1
address2
address3
address4

store
store_id
store_type
region_id
store_name
store_number
store_street_addr
store_city
store_state
store_postal_code

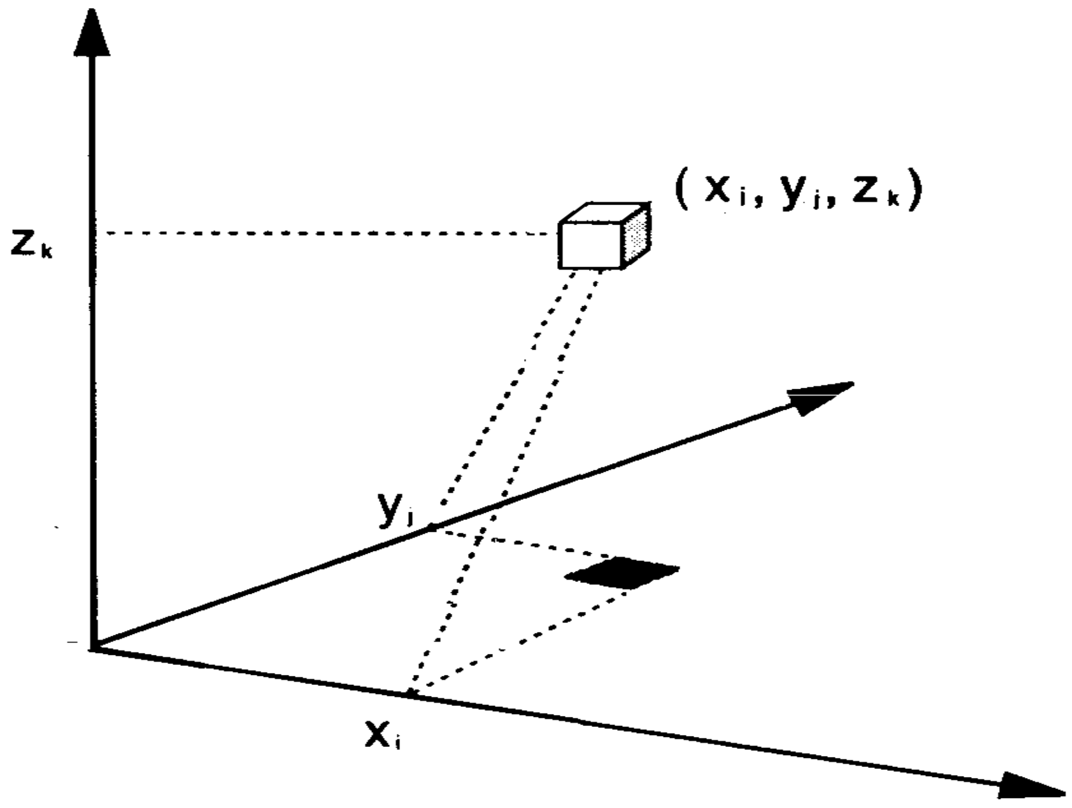
product_class
product_class_id
product_subcategory
product_category
product_department
product_family

🟨 Fact Table
🟦 Dimension Table

Multi-dimensional Data Model

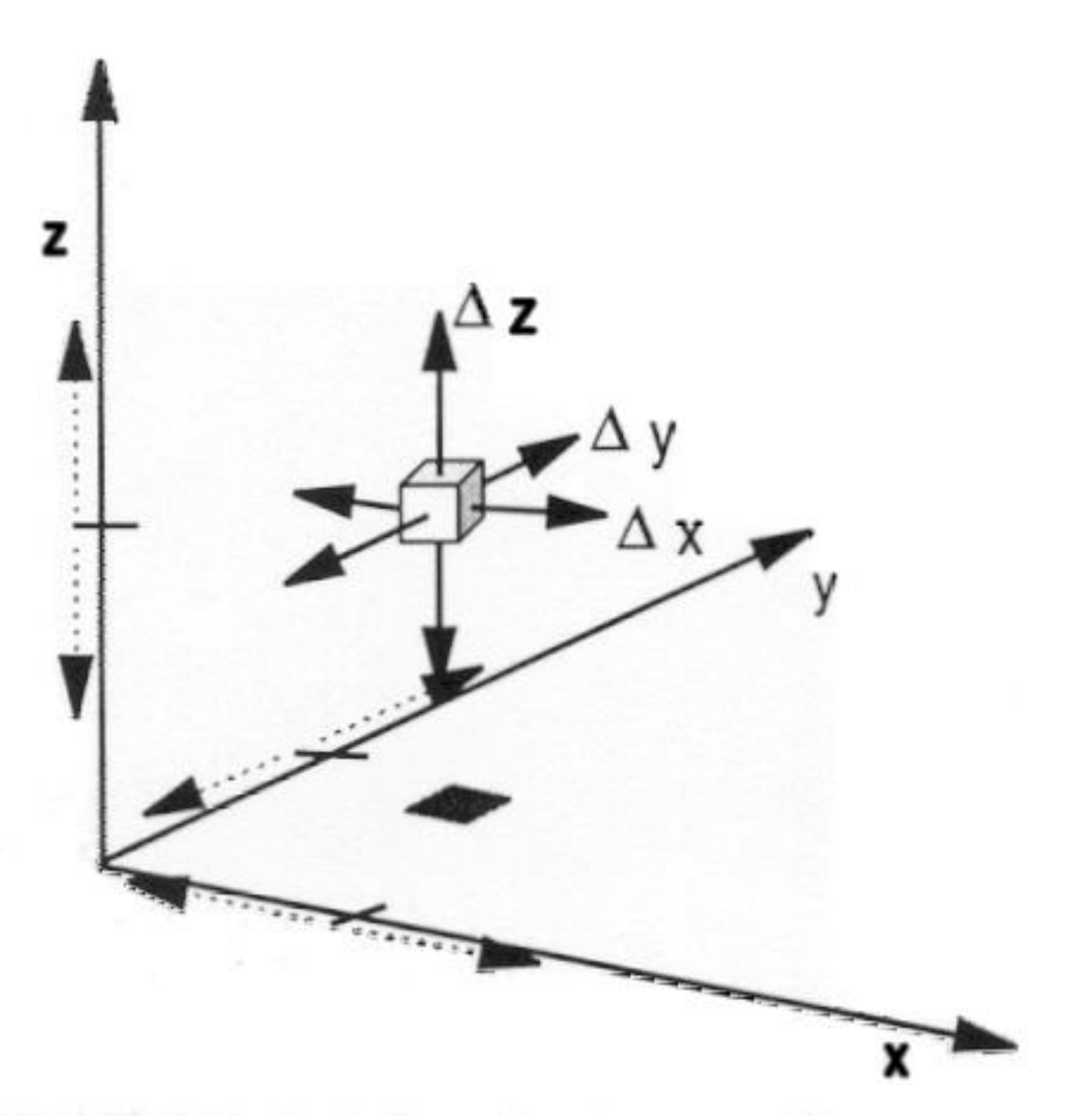
- OLAP applications are dominated by ad hoc, complex queries
  - In SQL terms, these are queries that involved group-by and aggregation operations which are poorly handled by DBMSs
- The natural way to think about typical OLAP queries, however, is in terms of a **Multi-dimensional Data Model (MDM)**
- MDM is a data model using logical dimensions to define a information space of business events.
  - This logical space is also called a hypercube (data cube). Each dimension of the cube represents an aspect of the possible business events which is divided into discrete values representing attribute domain of the dimension.

Locating subspaces in 3D space:

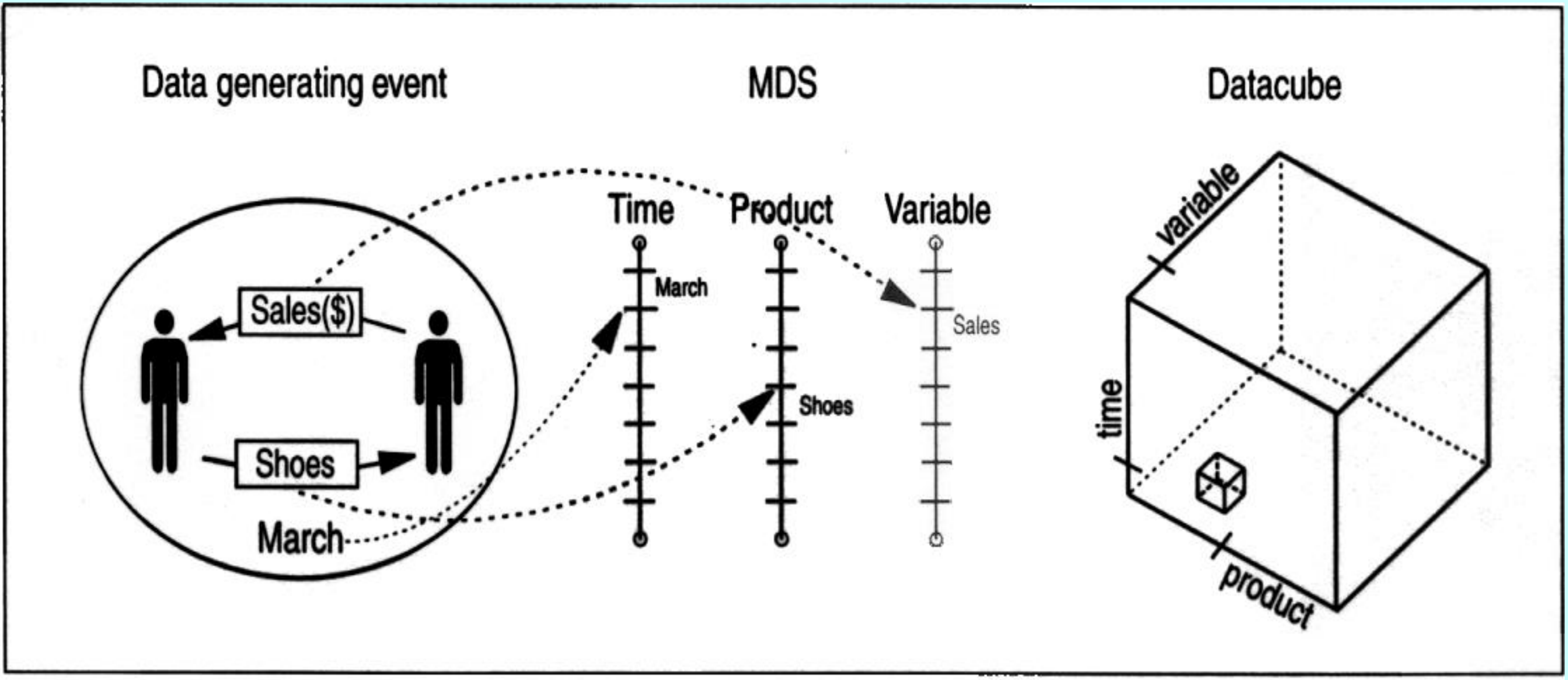


Points are identified by their *x*, *y*, *z* values.

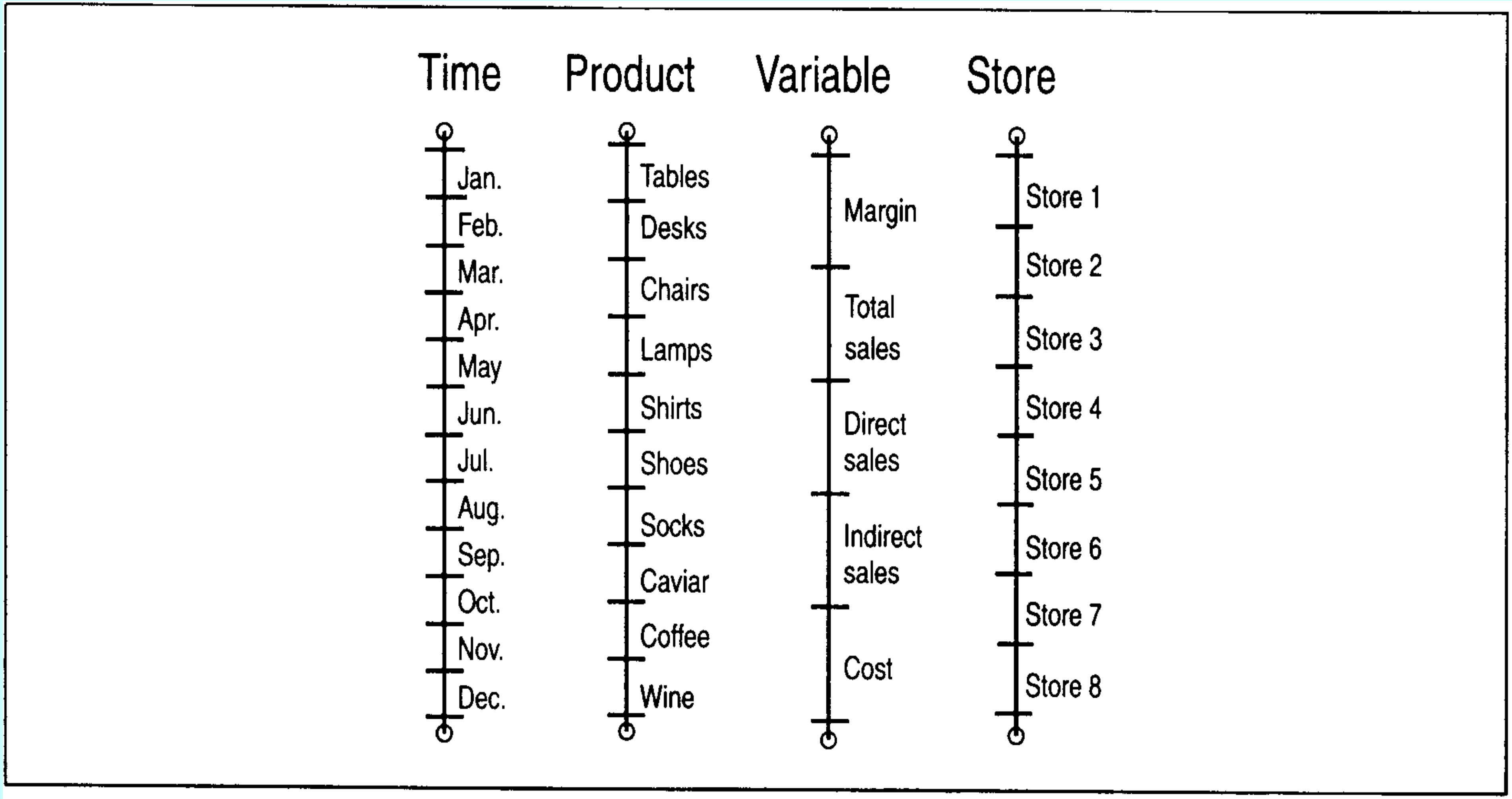
Manipulating a subspace:



How to make business events (groupings or views) to be generated and measured easily?

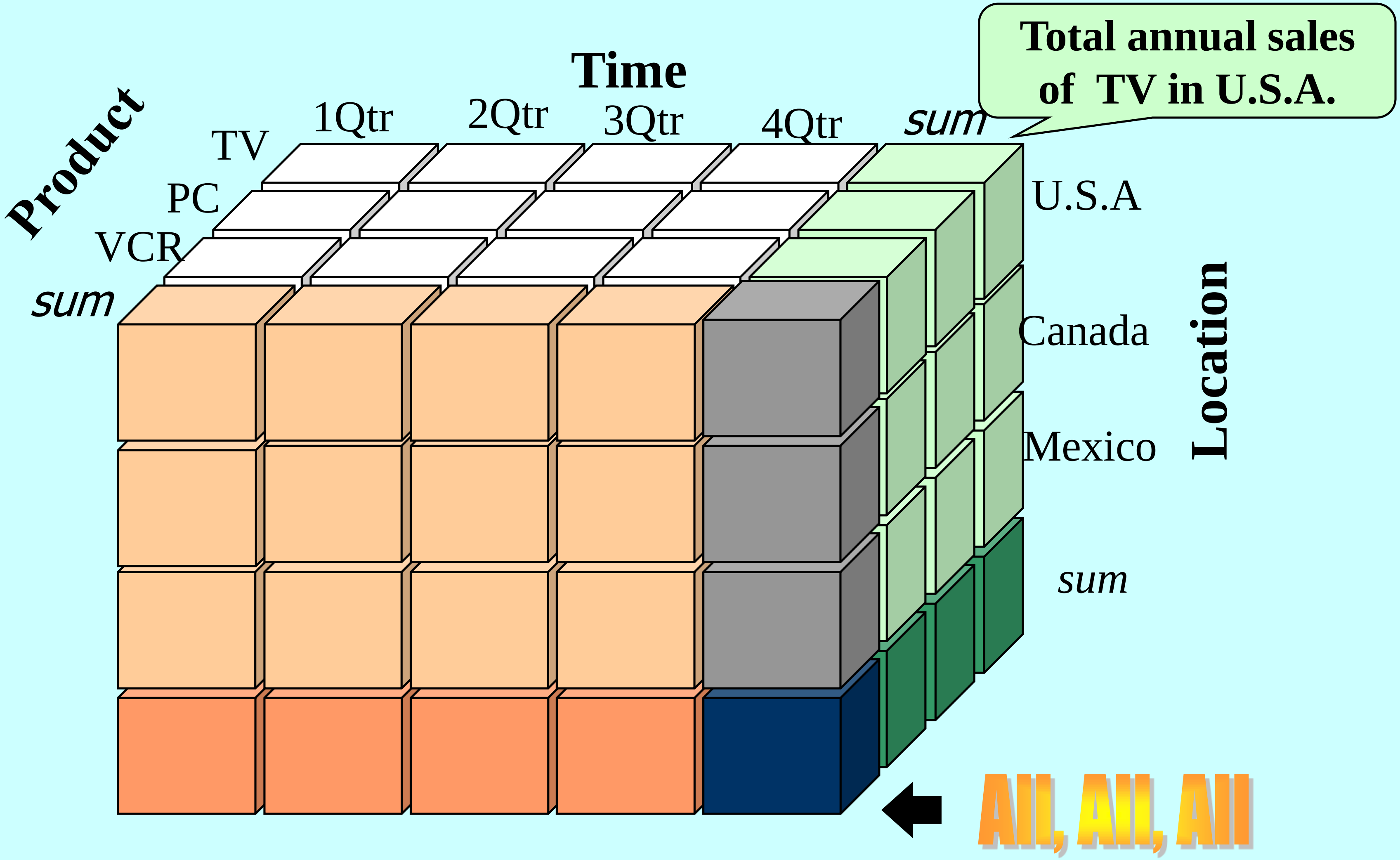


Multidimensional domain structures are a way of representing events.



The fourth dimension may be represented by a fourth line segment.

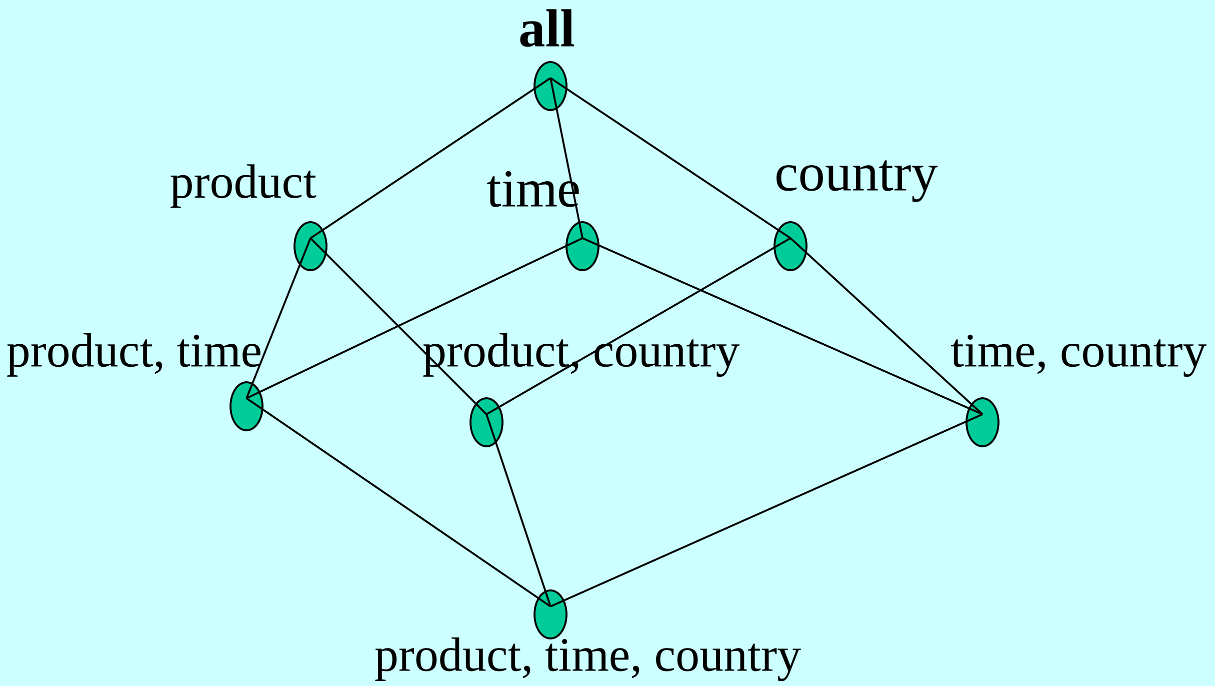
Data cube: **sales** (time, product, location)



Data Cube: How many cuboids?

- **Data cube is a structured space of cuboids**
  - **The bottom:** base cuboids, which hold the lowest-level of summarizations
  - **The top:** apex cuboid, which holds the highest-level of summarization
  - **How many cuboids** of a n-dimension data cube ?

Cuboids Corresponding to the Cube



0-D(apex) cuboid

1-D cuboids

2-D cuboids

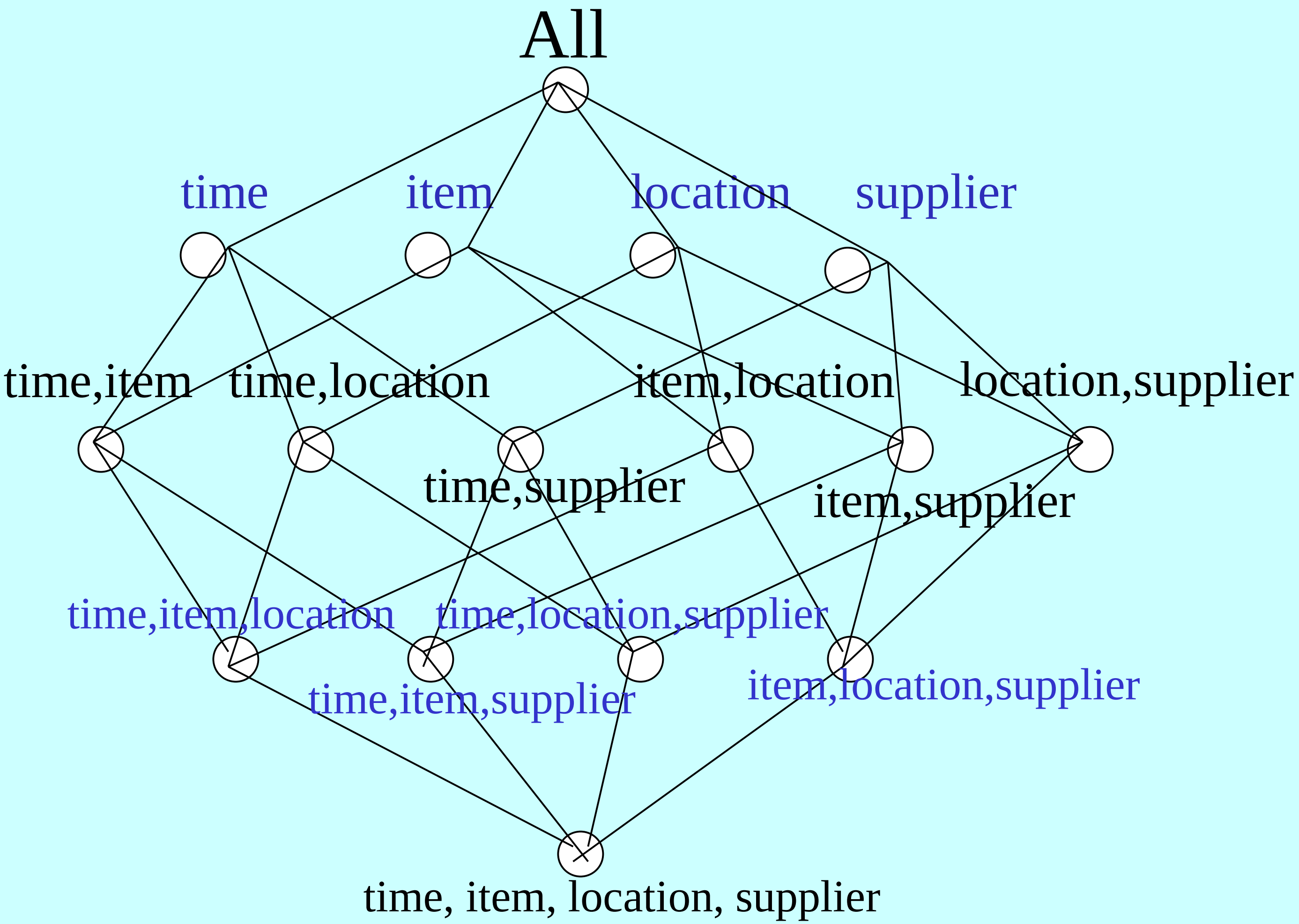
3-D(base) cuboid

Data Cube: A Lattice of Cuboids

- In DW literature, a 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.
- **How many cuboids in a data cube of n-dimensions?**

Data Cube: A Lattice of Cuboids

E.g., 4D cube: **Sales** (time, item, location, supplier)



0-D (apex) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (base) cuboid

Data Cube: How many cuboids?

- **Data cube is a structured space of cuboids**
  - **The bottom:** base cuboids, which hold the lowest-level of summarizations
  - **The top:** apex cuboid, which holds the highest-level of summarization
  - **How many cuboids** of a n-dimension data cube ?

Data Cube: How many cuboids?

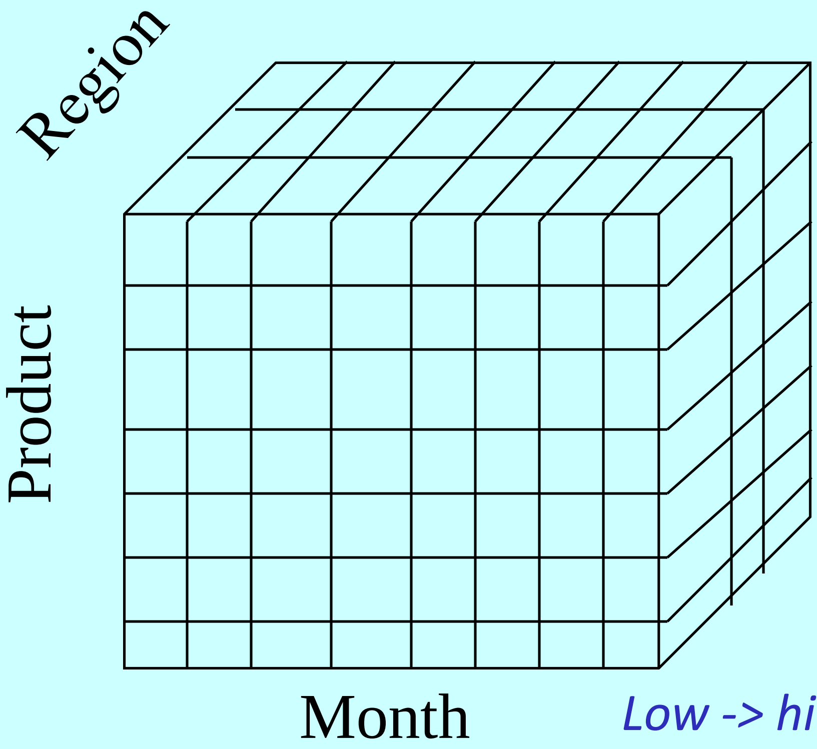
- **Data cube is a structured space of cuboids**
  - **The bottom:** base cuboids, which hold the lowest-level of summarizations
  - **The top:** apex cuboid, which holds the highest-level of summarization
- **The total cuboids** of a n-dimension

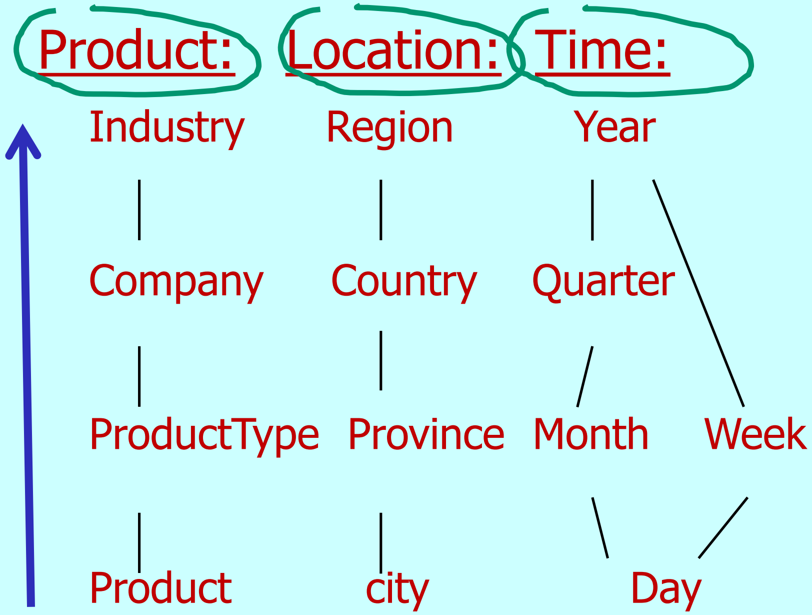
Concept Hierarchies

- For representing summarizations at different levels of same dimension
- E.g., Sales volume as a function of product, month, and region

Dimensions: Product, Location, Time

Hierarchical summarization paths



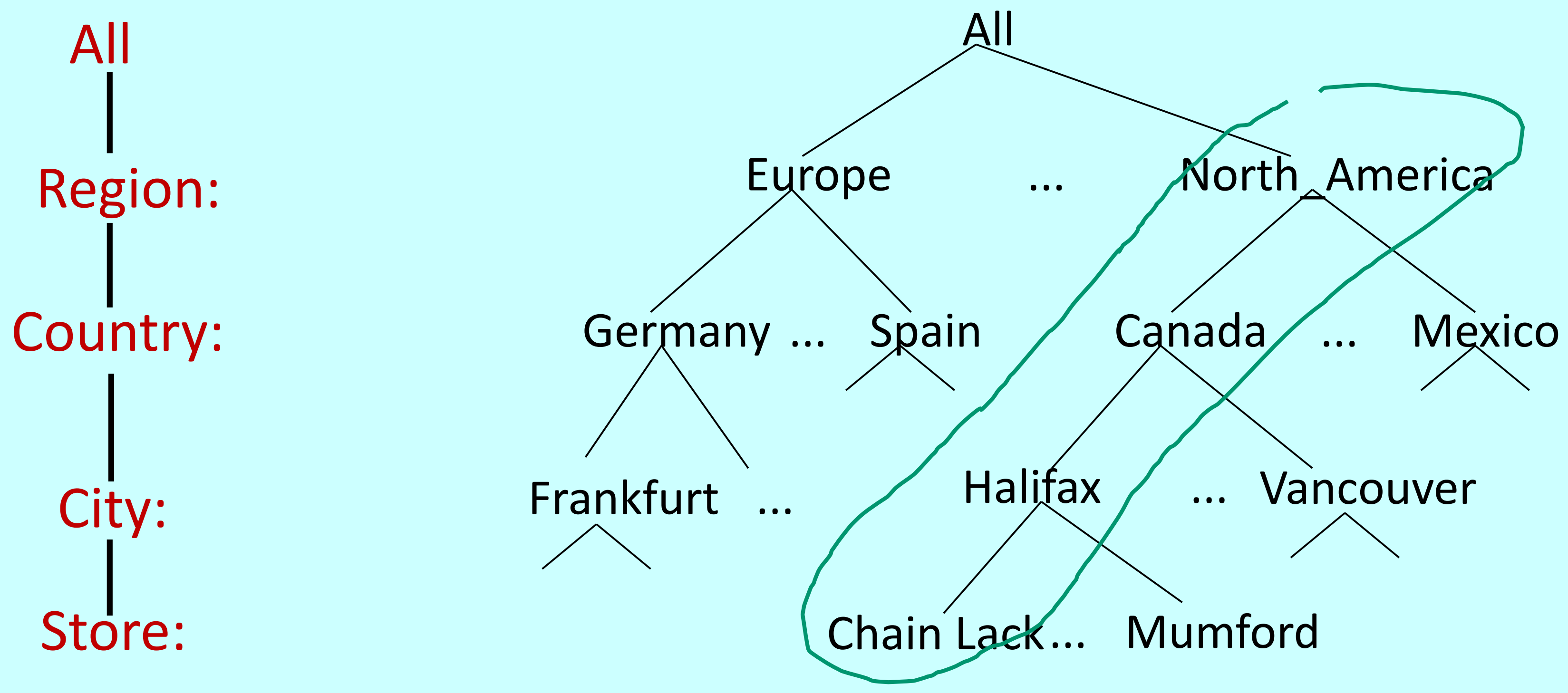

Product: Industry – Company – ProductType – Product

Location: Region – Country – Province – city

Time: Year – Quarter – Month – Day; Year – Week – Day

Low -> high granularity

E.g. A Concept Hierarchy: Dimension = “Location”



(Schema of “Location” table)

(Tuples of “Location” table)

What defines the dimension “Location”:

- It corresponding to a dimension table. The left is the table schema, and the right is the tuples in the table.
- “All” corresponds all records in the table.

How many cuboids when dimensions have concept hierarchies?

Cuboids Associated with Concept Hierarchies

- The cuboids of a data cube with n-dimensions, and L-concept hierarchies: $\boxed{T = \prod_{i=1}^{n} (L_i + 1)}$
  - T is the total number of cuboids, n is dimensions.
  - L_i is the number of concept levels associated with dimension i
  - E.g., For a cube with n = 10, and each dimension having 4 levels, the total number of cuboids: T = 5^10 ≈ 9.8 x 10^6.
- The number of cuboids also depends on the cardinality (i.e. number of distinct values) of each level of dimension concept, i.e. Total cuboids = *f (N, L, M),* where M is the distinct values of each level.

Review Questions

1. Why it was said that businesses are the drivers of DW and OLAP technologies?
2. For decision makers of a large enterprise to see big pictures of the organization and its business, what two general approaches for integrating information from different databases, can you describe each?
3. What are the general DW properties which differ DW model from Relational DB model, and how they fit the main objectives of DW?
4. What is the "Muti-Dimensional Space" model for DW/OLAP technology, and how it is described by the data cube lattice (know how to draw it)?
5. What are "Cuboids" in a MDS model? How to calculate the total number of cuboids in a DW? How to estimate the complexity of a DW design?
6. How different a DW model is compared with a conventional DB model and why?