OLAP & DATA MINING

CS561-SPRING 2012
WPI, MOHAMED ELTABAKH

Online Analytic Processing

OLAP

OLAP

- **OLAP:** Online Analytic Processing

- **OLAP queries are complex queries that**
  - Touch large amounts of data
  - Discover patterns and trends in the data
  - Typically expensive queries that take long time
  - Also called decision-support queries

- **In contrast to OLAP:**
  - ***OLTP***: Online Transaction Processing
  - OLTP queries are simple queries, e.g., over banking or airline systems
  - OLTP queries touch small amount of data for fast transactions

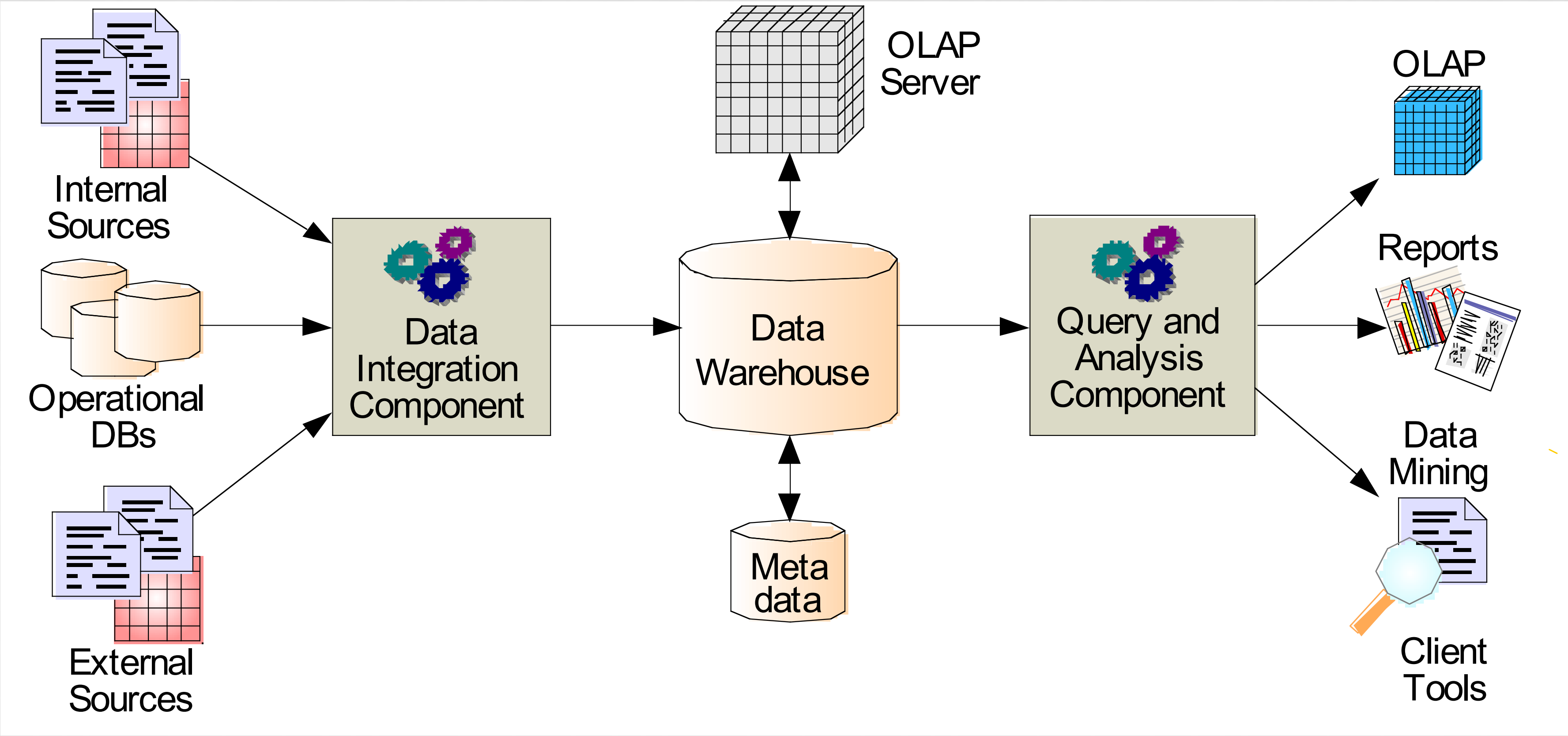
OLTP vs. OLAP

- **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

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.
[source: OLAP Council: www.olapcouncil.org]

OLAP AND DATA WAREHOUSE



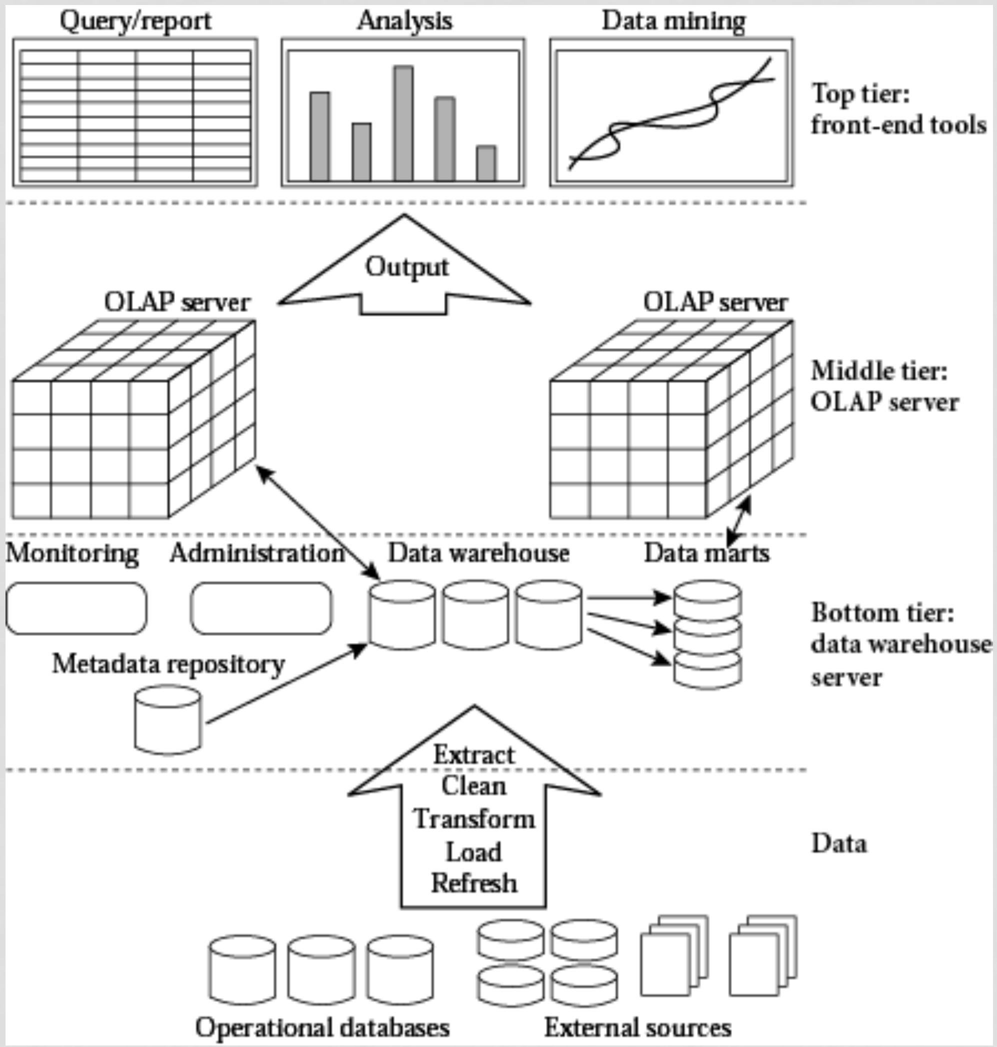
OLAP AND DATA WAREHOUSE

- Typically, OLAP queries are executed over a separate copy of the working data
  - Over data warehouse

- Data warehouse is periodically updated, e.g., overnight
  - OLAP queries tolerate such out-of-date gaps

- **Why run OLAP queries over data warehouse??**
  - Warehouse collects and combines data from multiple sources
  - Warehouse may organize the data in certain formats to support OLAP queries
  - OLAP queries are complex and touch large amounts of data
    - They may lock the database for long periods of time
    - Negatively affects all other OLTP transactions

OLAP ARCHITECTURE



EXAMPLE OLAP APPLICATIONS

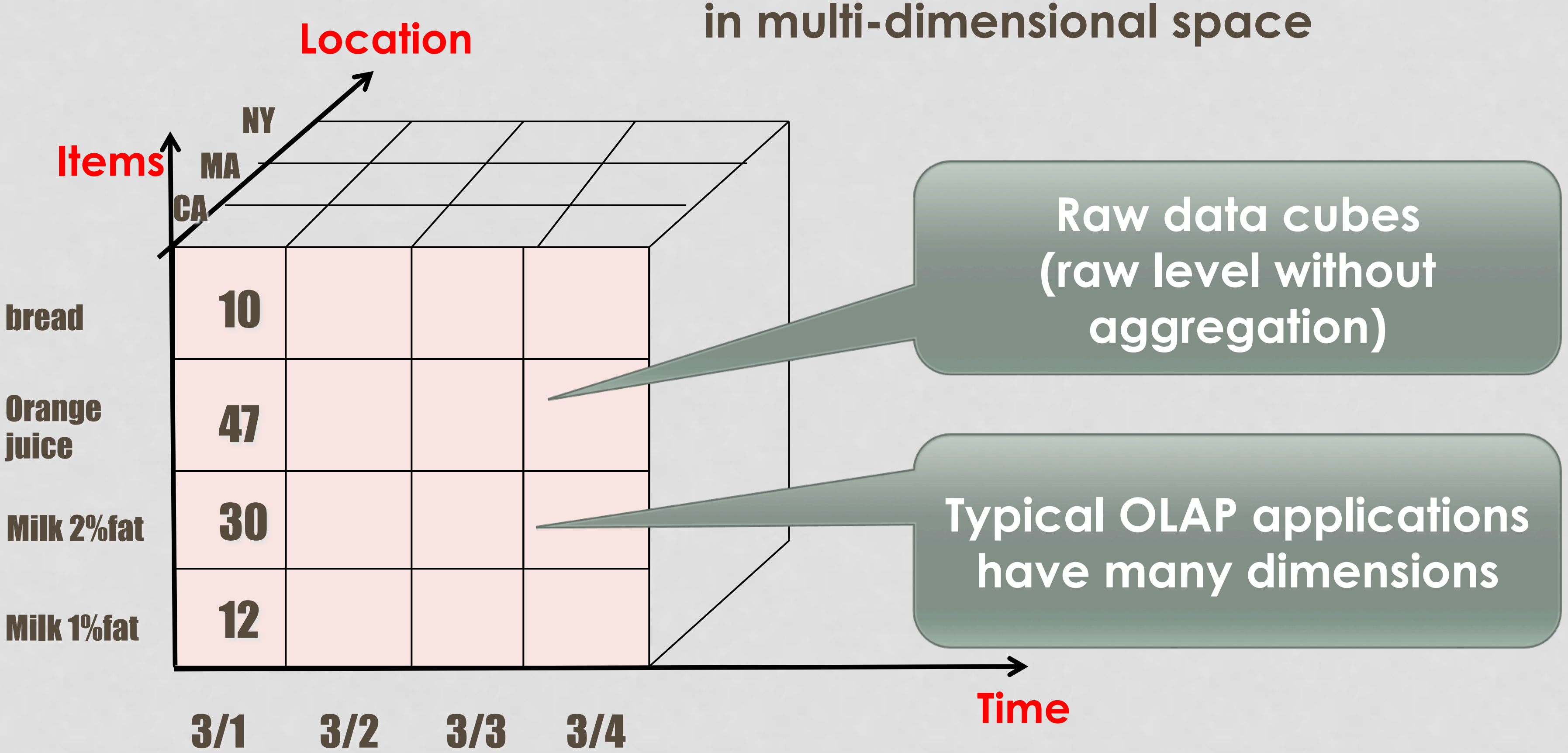
- **Market Analysis**
  - Find which items are frequently sold over the summer but not over winter?

- **Credit Card Companies**
  - Given a new applicant, does (s)he a credit-worthy?
  - Need to check other similar applicants (age, gender, income, etc…) and observe how they perform, then do prediction for new applicant

OLAP queries are also called “decision-support” queries

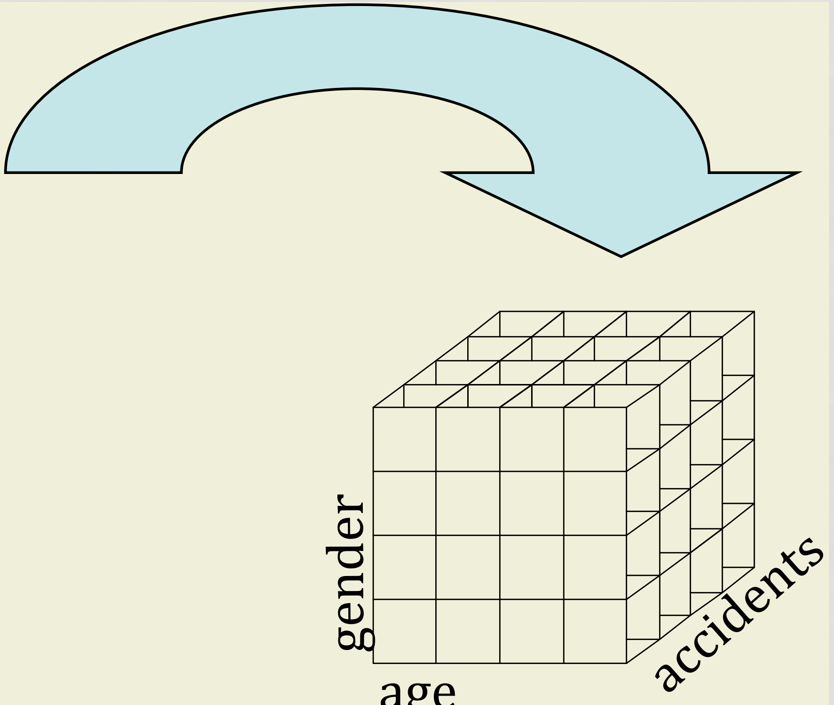
MULTI-DIMENSIONAL VIEW

- Data is typically viewed as points in multi-dimensional space



ANOTHER EXAMPLE

gender	age	accident
Male	27	3
Male	37	1
Male	37	0
Male	37	1
Male	49	2
Male	39	4
Male	43	0
Male	41	2
Male	49	1
Male	44	2
Male	43	3
Male	53	4
Male	60	0
Female	26	0
Female	39	0
Female	45	2
Female	41	2
Female	39	1
Female	37	0
Female	43	1

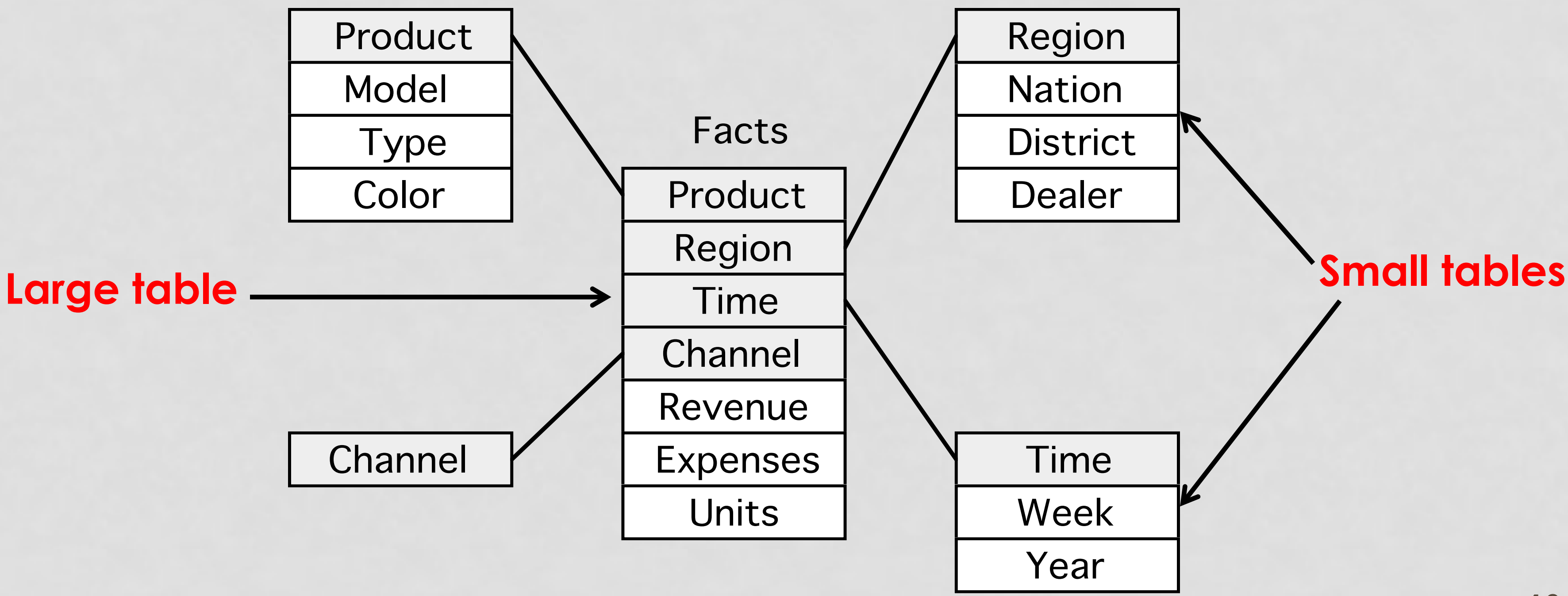


APPROACHES FOR OLAP

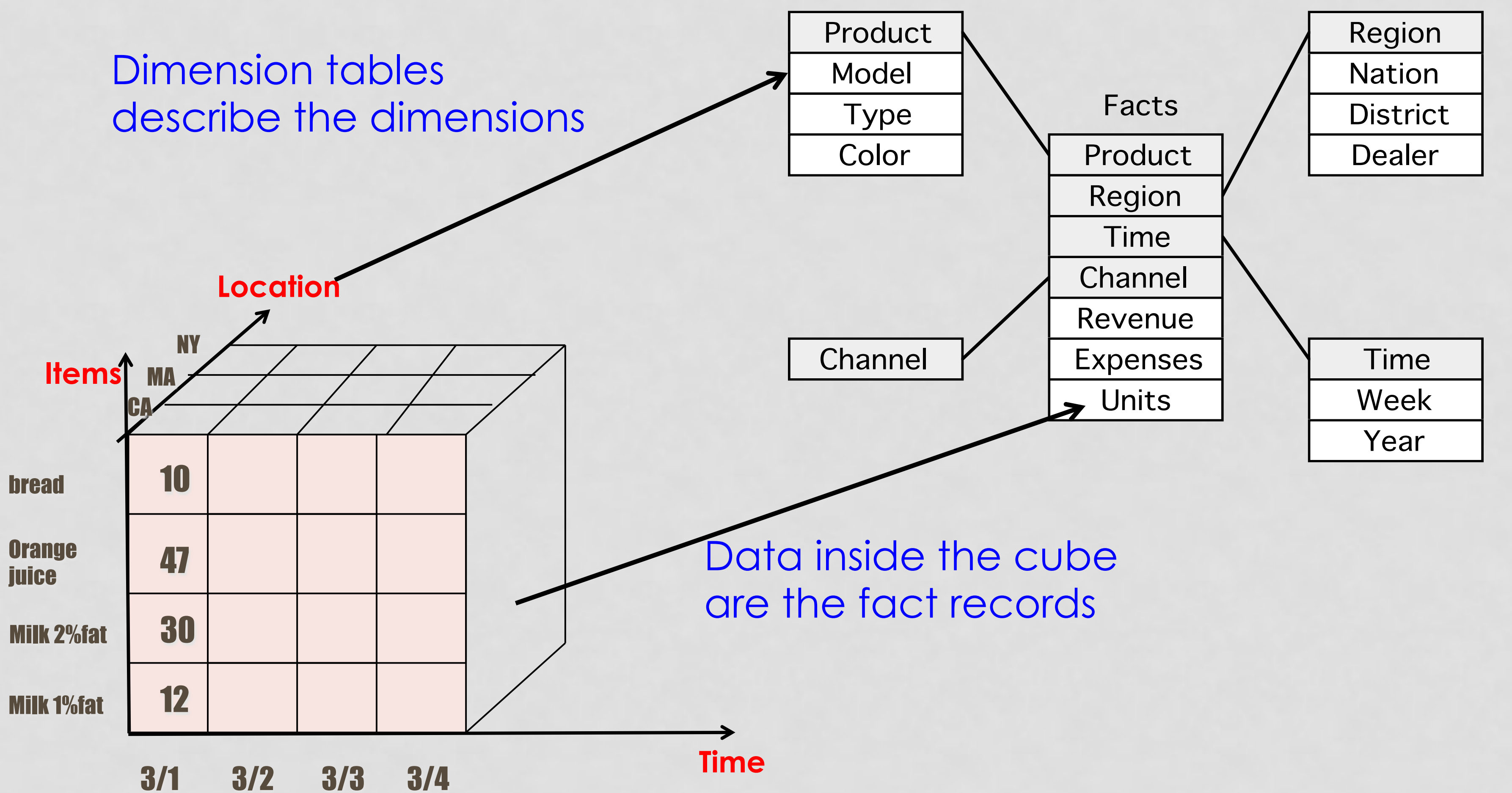
- **Relational OLAP (ROLAP)**
- **Multi-dimensional OLAP (MOLAP)**
- **Hybrid OLAP (HOLAP) = ROLAP + MOLAP**

RELATIONAL OLAP: ROLAP

- Data are stored in relational model (tables)
- Special schema called ***Star Schema***
- One relation is the ***fact table***, all the others are ***dimension tables***



CUBE vs. STAR SCHEMA



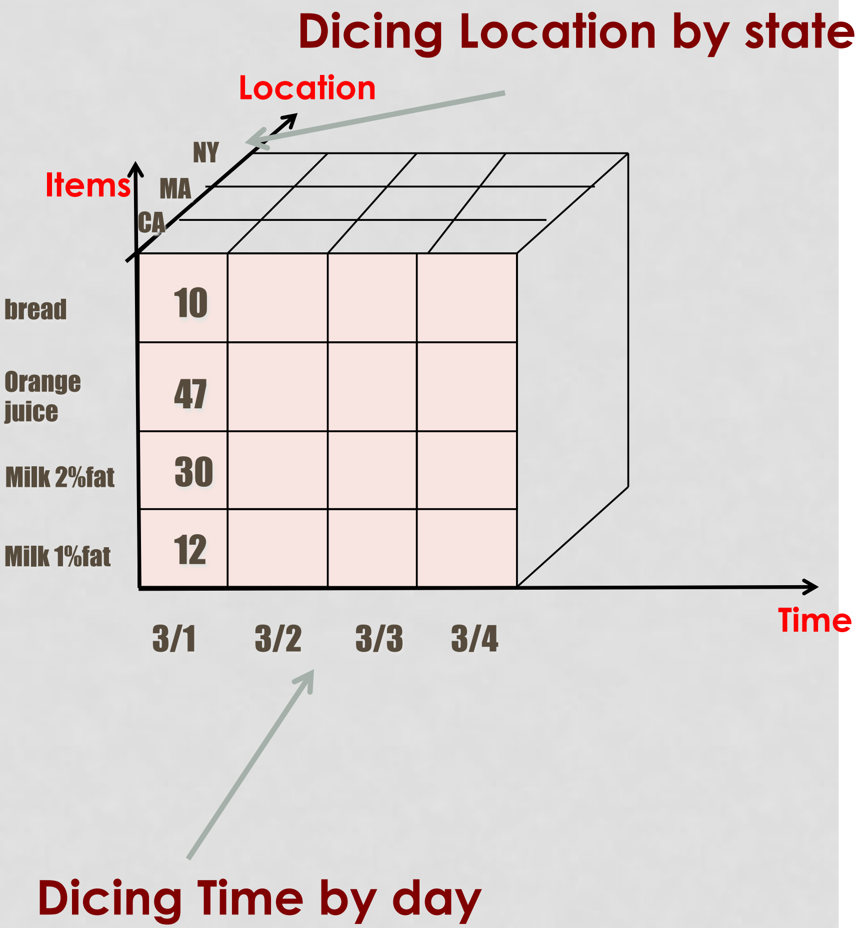
ROLAP: EXTENSIONS TO DBMS

- Schema design
- Specialized scan, indexing and join techniques
- Handling of aggregate views (querying and materialization)
- Supporting query language extensions beyond SQL
- Complex query processing and optimization
- Data partitioning and parallelism

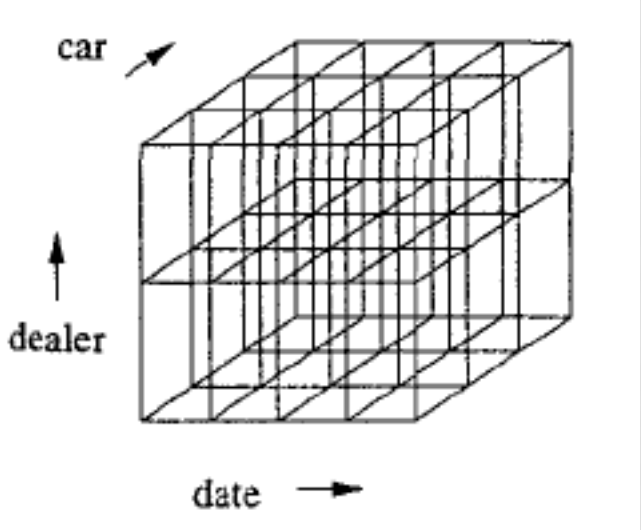
SLICING & DICING

- **Dicing**
  - how each dimension in the cube is divided
  - Different granularities
  - When building the data cube

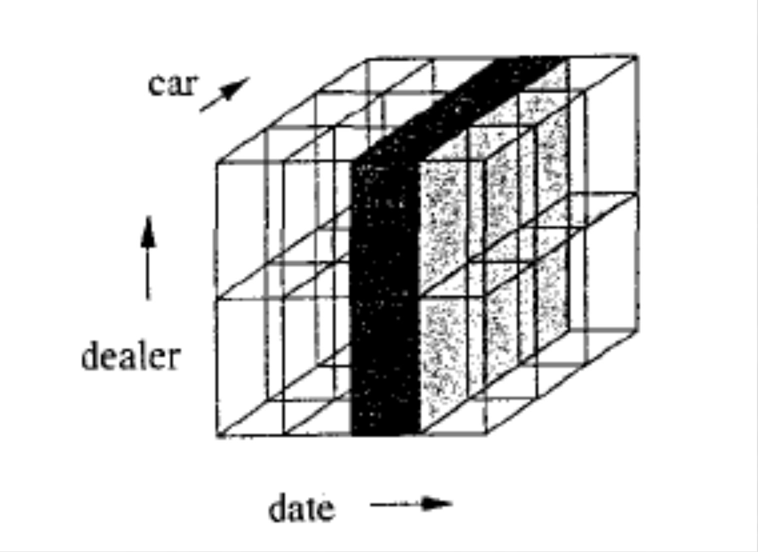
- **Slicing**
  - Selecting slices of the data cube to answer the OLAP query
  - When answering a query



SLICING & DICING: EXAMPLE 1



Dicing



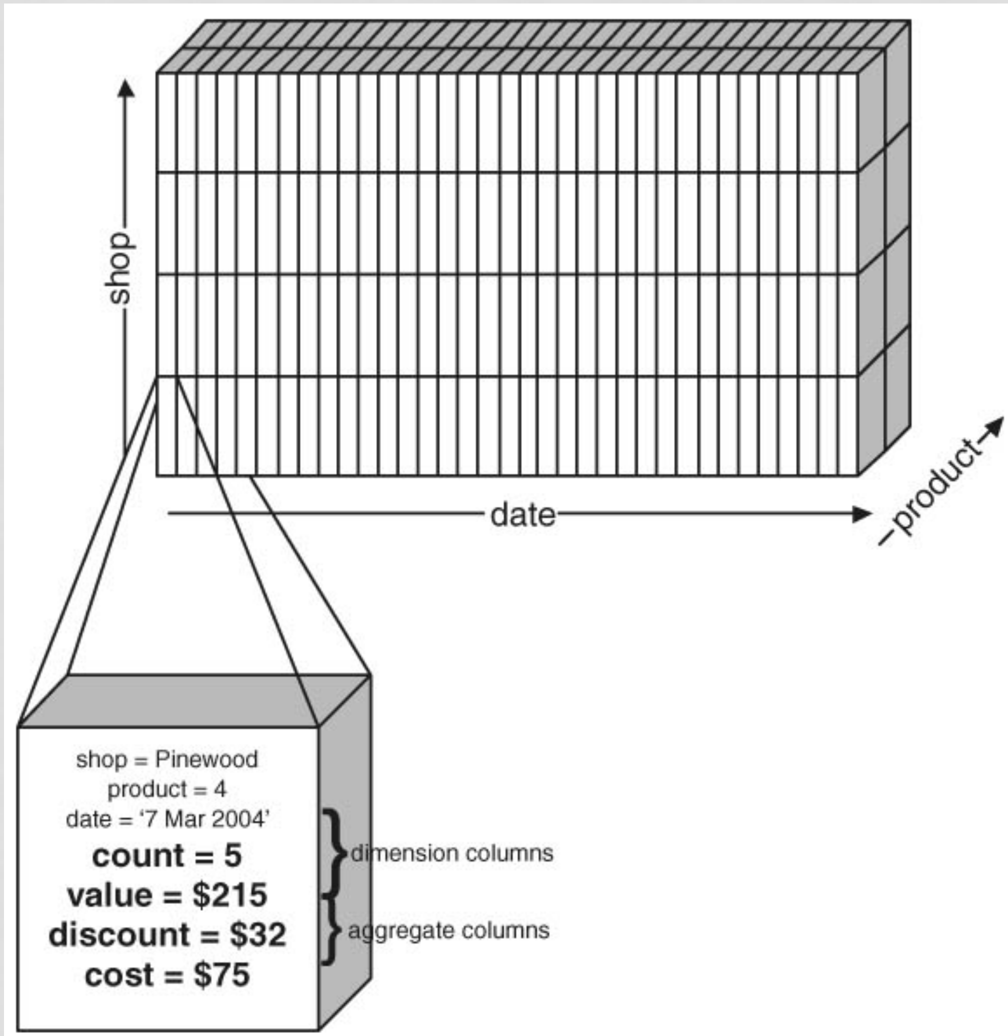
Slicing

Slicing operation in ROLAP is basically:

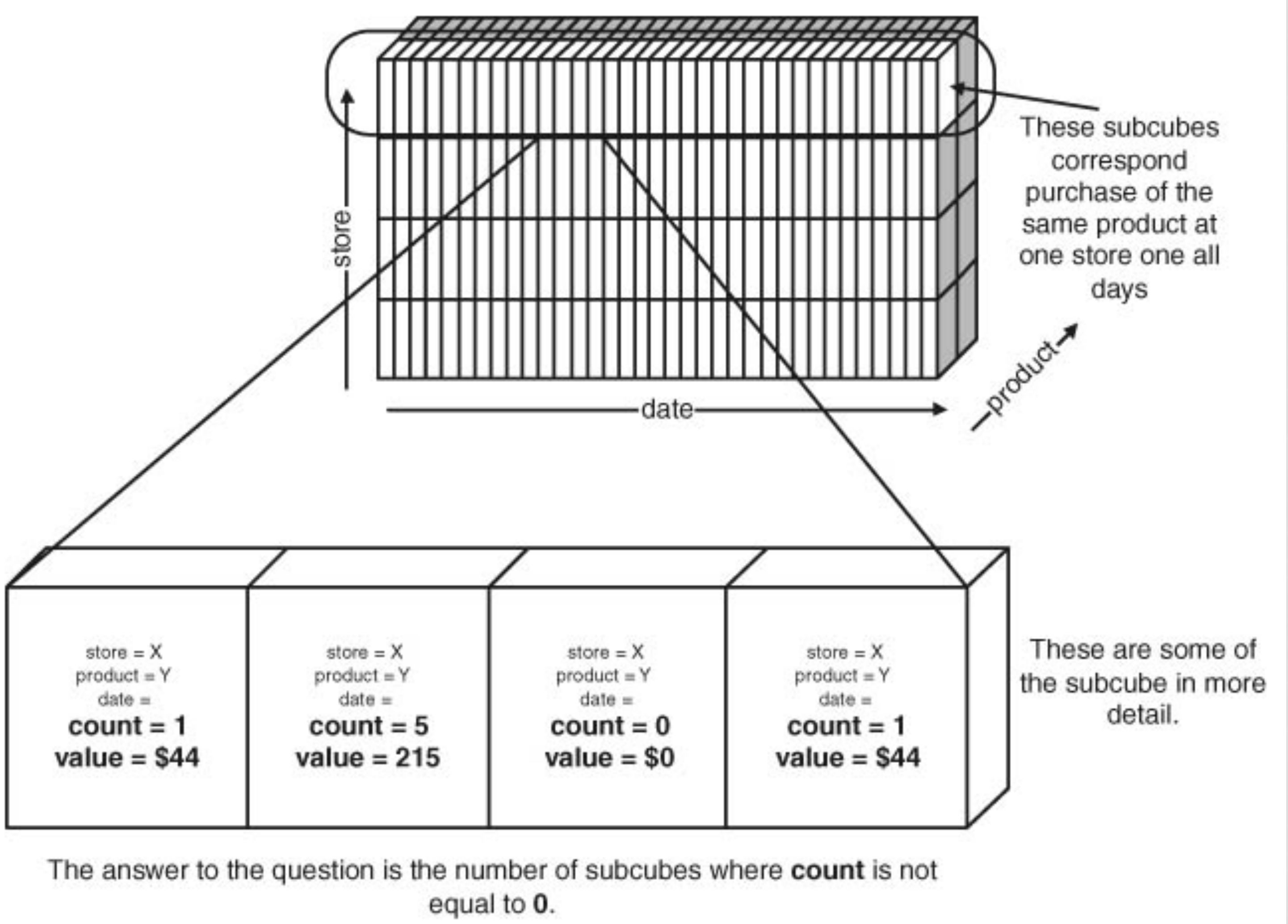
-- Selection conditions on some attributes (WHERE clause) +

-- Group by and aggregation

SLICING & DICING: EXAMPLE 2



SLICING & DICING: EXAMPLE 3



DRILL-DOWN & ROLL-UP

Region	Sales variance
Africa	105%
Asia	57%
Europe	122%
North America	97%
Pacific	85%
South America	163%

Drill-down
(Group by Nation)

Roll-up
(group by Region)

Nation	Sales variance
China	123%
Japan	52%
India	87%
Singapore	95%

ROLAP: DRILL-DOWN & ROLL-UP

```
SELECT dealer, year, SUM(price)
FROM (Sales NATURAL JOIN Autos) JOIN Days ON date = day
WHERE model = 'Gobi' AND
      color = 'red' AND
      (year = 2001 OR year = 2002)
GROUP BY year, dealer;
```

Drill-down ↓ ↑ **Roll-up**

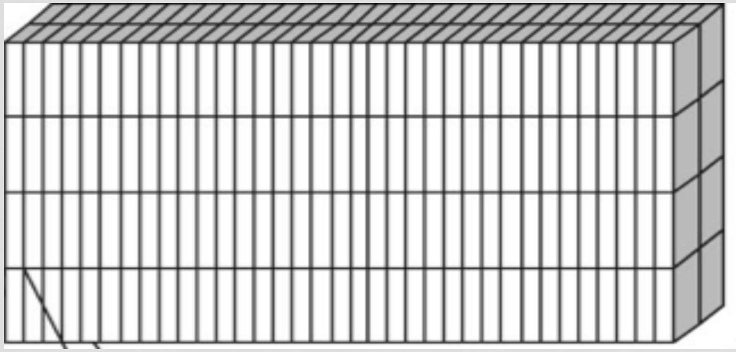
```
SELECT dealer, month, SUM(price)
FROM (Sales NATURAL JOIN Autos) JOIN Days ON date = day
WHERE model = 'Gobi' AND color = 'red'
GROUP BY month, dealer;
```

MOLAP

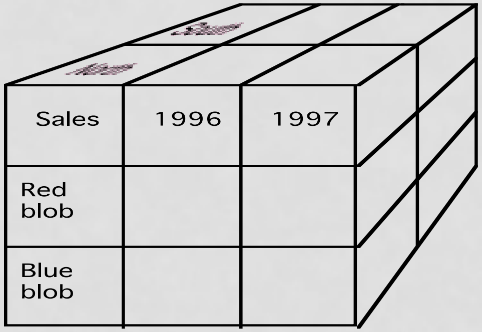
- Unlike ROLAP, in MOLAP data are stored in special structures called "***Data Cubes***" **(Array-bases storage)**

- **Data cubes pre-compute and aggregate the data**
  - Possibly several data cubes with different granularities
  - Data cubes are aggregated materialized views over the data

- As long as the data does not change frequently, the overhead of data cubes is manageable



Every day, every item, every city



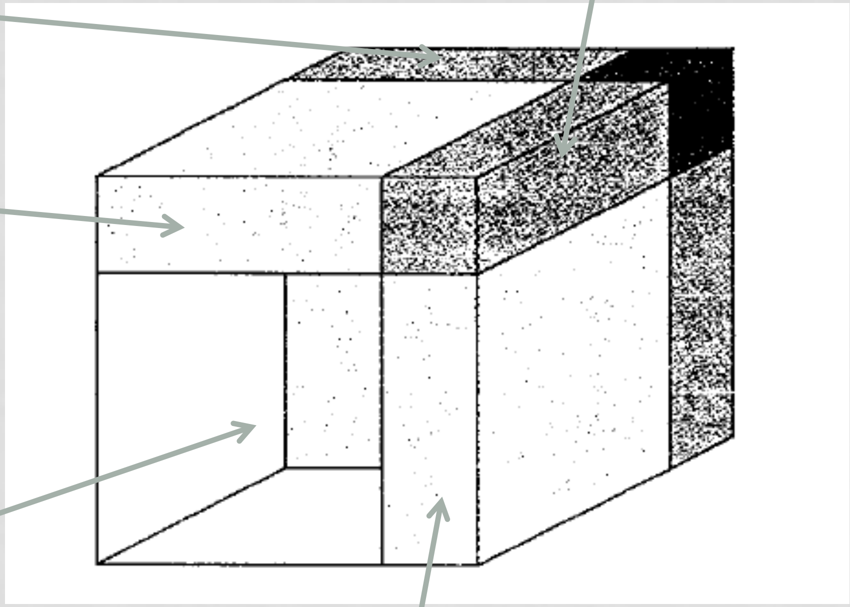
Every week, every item category, every city

MOLAP: CUBE OPERATOR

Aggregation over the X,Y

Aggregation over the Z axis

Aggregation over the Y axis



Raw-data (fact table)

Aggregation over the X axis

MOLAP & ROLAP

- Commercial offerings of both types are available
- In general, **MOLAP** is good for smaller warehouses and is optimized for canned queries
- In general, **ROLAP** is more flexible and leverages relational technology
- **ROLAP** May pay a performance penalty to realize flexibility

OLTP vs. OLAP

	OLTP	OLAP
User	Clerk, IT Professional	Knowledge worker
Function	Day to day operations	Decision support
DB Design	Application-oriented (E-R based)	Subject-oriented (Star, snowflake)
Data	Current, Isolated	Historical, Consolidated
View	Detailed, Flat relational	Summarized, Multidimensional
Usage	Structured, Repetitive	Ad hoc
Unit of work	Short, Simple transaction	Complex query
Access	Read/write	Read Mostly
Operations	Index/hash on prim. Key	Lots of Scans
# Records accessed	Tens	Millions
#Users	Thousands	Hundreds
Db size	100 MB-GB	100GB-TB
Metric	Trans. throughput	Query throughput, response

Source: Datta, GT

OLAP: SUMMARY

- OLAP stands for Online Analytic Processing and used in decision support systems
  - Usually runs on data warehouse

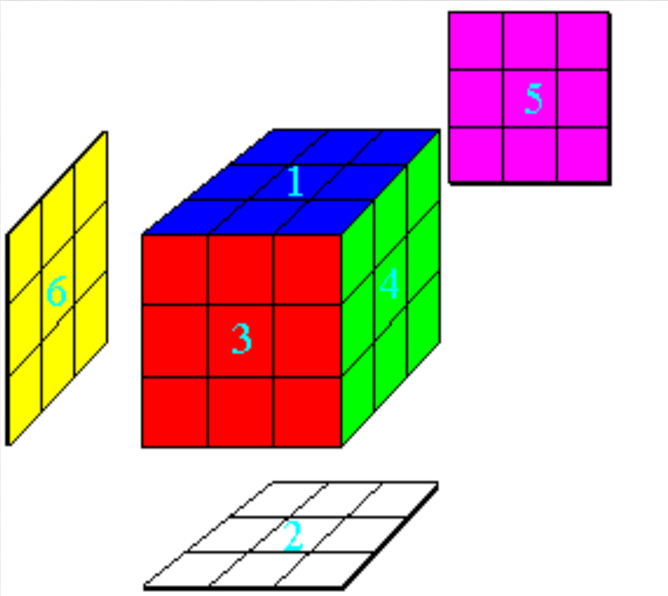
- In contrast to OLTP, OLAP queries are complex, touch large amounts of data, try to discover patterns or trends in the data

- **OLAP Models**
  - **Relational (ROLAP):** uses relational star schema
  - **Multidimensional (MOLAP):** uses data cubes

Overview on Data Mining Techniques

DATA MINING vs. OLAP

- OLAP - Online Analytical Processing
  - Provides you with a very good view of <u>what is happening</u>, but <u>can not predict what will happen</u> in the future or <u>why it is happening</u>



Data Mining is a combination of discovering techniques + prediction techniques

DATA MINING TECHNIQUES

- **Clustering**
- **Classification**
- **Association Rules**
- **Frequent Itemsets**
- **Outlier Detection**
- **….**

FREQUENT ITEMSET MINING

- Very common problem in Market-Basket applications

- Given a set of items I ={milk, bread, jelly, …}
- Given a set of transactions where each transaction contains subset of items
  - t1 = {milk, bread, water}
  - t2 = {milk, nuts, butter, rice}

What are the itemsets frequently sold together ??

% of transactions in which the itemset appears >= α

EXAMPLE

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Assume α = 60%, what are the frequent itemsets

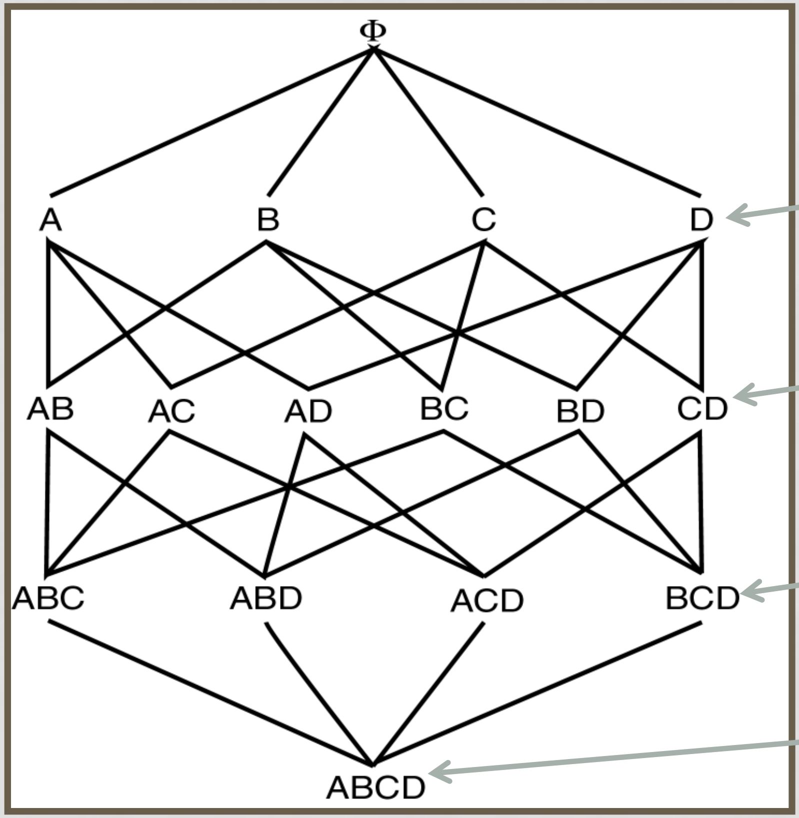
- {Bread} → 80%
- {PeanutButter} → 60%
- {Bread, PeanutButter} → 60%

called "*Support*"

All frequent itemsets given α = 60%

HOW TO FIND FREQUENT ITEMSETS

- **Naïve Approach**
  - Enumerate all possible itemsets and then count each one



All possible itemsets of size 1

All possible itemsets of size 2

All possible itemsets of size 3

All possible itemsets of size 4

CAN WE OPTIMIZE??

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Assume α = 60%, what are the frequent itemsets

- {Bread} → 80%
- {PeanutButter} → 60%
- {Bread, PeanutButter} → 60%

called "*Support*"

Property

For itemset S={X, Y, Z, …} of size n to be frequent, all its subsets of size n-1 must be frequent as well

APRIORI ALGORITHM

- **Executes in scans, each scan has two phases**
  - Given a list of candidate itemsets of size n, count their appearance and find frequent ones
  - From the frequent ones generate candidates of size n+1 *(previous property must hold)*
  - Start the algorithm where n =1, then repeat

Use the property reduce the number of itemsets to check

APRIORI EXAMPLE

Transaction	Items
t_1	Blouse
t_2	Shoes,Skirt,TShirt
t_3	Jeans,TShirt
t_4	Jeans,Shoes,TShirt
t_5	Jeans,Shorts
t_6	Shoes,TShirt
t_7	Jeans,Skirt
t_8	Jeans,Shoes,Shorts,TShirt
t_9	Jeans
t_{10}	Jeans,Shoes,TShirt
t_{11}	TShirt
t_{12}	Blouse,Jeans,Shoes,Skirt,TShirt
t_{13}	Jeans,Shoes,Shorts,TShirt
t_{14}	Shoes,Skirt,TShirt
t_{15}	Jeans,TShirt
t_{16}	Skirt,TShirt
t_{17}	Blouse,Jeans,Skirt
t_{18}	Jeans,Shoes,Shorts,TShirt
t_{19}	Jeans
t_{20}	Jeans,Shoes,Shorts,TShirt

APRIORI EXAMPLE (CONT'D)

Scan	Candidates	Large Itemsets
1	{Blouse},{Jeans},{Shoes}, {Shorts},{Skirt},{TShirt}	{Jeans},{Shoes},{Shorts} {Skirt},{Tshirt}
2	{Jeans,Shoes},{Jeans,Shorts},{Jeans,Skirt}, {Jeans,TShirt},{Shoes,Shorts},{Shoes,Skirt}, {Shoes,TShirt},{Shorts,Skirt},{Shorts,TShirt}, {Skirt,TShirt}	{Jeans,Shoes},{Jeans,Shorts}, {Jeans,TShirt},{Shoes,Shorts}, {Shoes,TShirt},{Shorts,TShirt}, {Skirt,TShirt}
3	{Jeans,Shoes,Shorts},{Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt},{Jeans,Skirt,TShirt}, {Shoes,Shorts,TShirt},{Shoes,Skirt,TShirt}, {Shorts,Skirt,TShirt}	{Jeans,Shoes,Shorts}, {Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt}, {Shoes,Shorts,TShirt}
4	{Jeans,Shoes,Shorts,TShirt}	{Jeans,Shoes,Shorts,TShirt}
5	$\emptyset$	$\emptyset$

DATA MINING TECHNIQUES

- **Clustering**
- **Classification**
- **Association Rules**
- **Frequent Itemsets**
- **Outlier Detection**
- ….

ASSOCIATION RULES MINING

- **What is the probability when a customer buys *bread* in a transaction, (s)he also buys *milk* in the same transaction?**

Implies?

Bread ----------------------> milk

Frequent itemsets cannot answer this question….But Association rules can

General Form

Association rule: x1, x2, …, xn → y1, y2, …ym

Meaning: when the L.H.S appears (or occurs), the R.H.S also appears (or occurs) with certain probability

Two measures for a given rule:

1- Support(L.H.S U R.H.S) > α

2- Confidence C = Support(L.H.S U R.H.S)/ Support(L.H.S)

EXAMPLE

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Usually we search for rules:
Support > α
Confidence > β

Rule: Bread → PeanutButter

- Support of rule = support(Bread, PeanutButter) = 60%
- Confidence of rule = support(Bread, PeanutButter)/support(Bread) = 75%

Rule: Bread, Jelly → PeanutButter

- Support of rule = support(Bread, Jelly, PeanutButter) = 20%
- Confidence of rule = support(Bread, Jelly, PeanutButter) /support(Bread, Jelly) = 100%