OLAP and Statistical Databases: Similarities and Differences

Arie Shoshani
Lawrence Berkeley National Laboratory
(http://www.lbl.gov/~arie/papers)

Outline

1. Examples of SDB and OLAP databases
2. Application area emphasized in each
3. Conceptual Modeling - data structures
4. Conceptual modeling - operations
5. Physical organization
6. Privacy
7. The case for the support of the Statistical Object data type by commercial systems

1. Examples of SDB and OLAP databases

- Traditional representation of statistical data
- Data cube model for OLAP data
- Interchangeability
- Main concepts:
  - multidimensionality
  - classification hierarchy
  - summary attribute

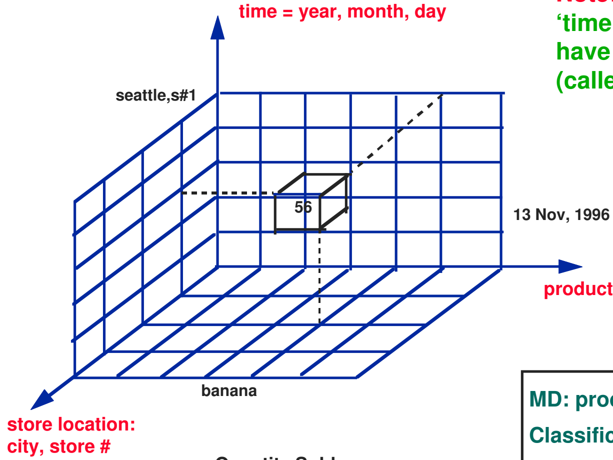
Traditional representation of statistical data

Employment in California			Professional Class					
			Engineer		Secretary		Teacher	
			Profession		Profession		Profession	
			Chemical Engineer	Civil Engineer	Junior Secretary	Executive Secretary	Elementary Teacher	high school Teacher
Sex	Male	Year 91	197,700	241,100	534,300	154,100	212,943	123,740
		92	209,900	278,000	542,100	169,8 00	213,521	145,766
		. .	....	....	....	......	....	....
	Female	Year 91	25,800	112,000	667,300	162,,300	216,071	275,123
		92	28,900	127,600	692,500	174,400	217,520	299,344
		. .	.....	......	......	......	......	....

MD: sex, year, profession, state=California
Classification: professional class, profession
Summary attribute: No. of employees

Data Cube Model

Computing Sciences Directorate, LBNL



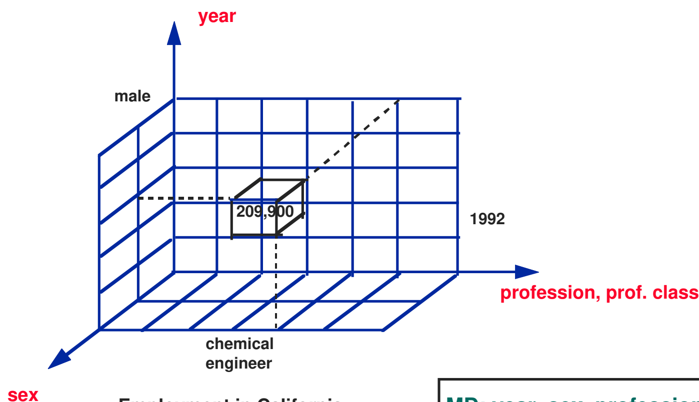
Quantity Sold

Note:
'time' and 'store location' have an inherent hierarchy (called ID dependency)

MD: product, store location, time
Classification: year, month, day
Classification: store #, city
Summary attribute: quantity sold

Interchangeability

Computing Sciences Directorate, LBNL



Employment in California

MD: year, sex, profession, state=CA
Classification: profession, prof. class
Summary attribute: No. of employees

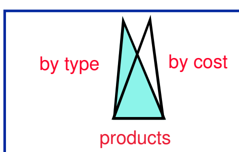
2. Application areas emphasized in each

- Statistical Data Bases
- Census data: population size, household income, ...
  - need: geographic classification hierarchy,
- Economic data: sales by industry, production & consumption, income, ...
  - need: classification of industry, regions, branches
- Natural resources: water levels, forests, rivers, ...
  - need: physical geography (rivers bordering states), botanical classification
- Health data: infant mortality, life expectancy, health costs, ...
  - need: temporal data-tracking of individuals over time illnesses classification

2. Application areas emphasized in each (cont'd)

On Line Analytical Processing (OLAP)

- retail store sales analysis
  - need: multiple classifications of products, as well as regions, time
- stock market databases
  - need: time series, industry classification
- health maintenance organizations
  - track cost per patient per procedure
  - need: confidentiality, privacy, classification of illnesses, regions, time varying categories



Multiple Classifications

Difference observation

Computing Sciences Directorate, LBNL

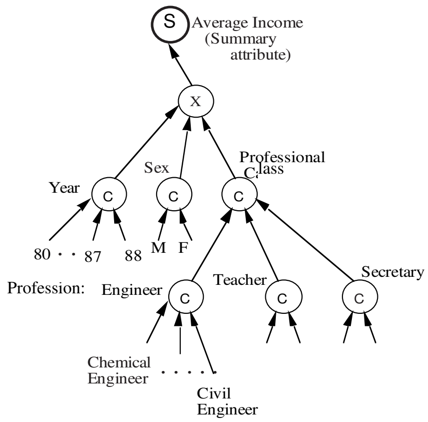
- In both SDBs and OLAP there is
  - base data (also called Micro Data)
  - summary data (also called Macro Data)
  - meta data (including category hierarchy values)
- e.g. Micro Data
  - Sales Price (item, store, time_stamp)
- e.g. Macro Data
  - Avg. Price (item, store, day)
- SDBs usually present to users macro data only
- OLAP usually starts with micro data

3. Conceptual Modeling - data structures

Computing Sciences Directorate, LBNL

- graph models for SDBs
- tabular models for SDBs
- data cube model for OLAP
- terminology correspondence

graph models for SDBs



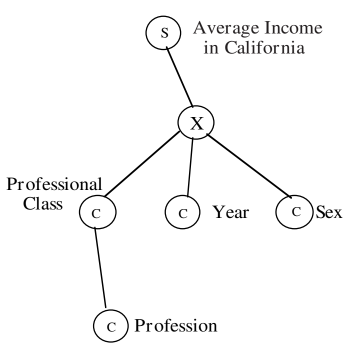
Advantages:
- can represent many dimensions
- category hierarchy explicit
- insensitive to node permutation

Disadvantages:
- fanout of categories can be large
- nodes used for both labels and category values

Terminology:
- summary attribute
- category attribute
- category hierarchy

[Chan & Shoshani 81]

graph models for SDBs (cont'd)



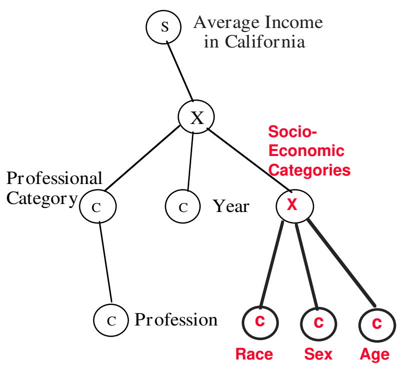
Advantages:
- can represent many dimensions
- category hierarchy explicit
- no fanout of categories

Concepts:
- metadata has schema and instances
- separate schema level from instance level (for metadata)
- In GUI click on nodes opens instance level
- GUI for query specification can use same paradigm

[Rafanelli & Shoshani 90]

Computing Sciences Directorate, LBNL

graph models for SDBs (cont'd)



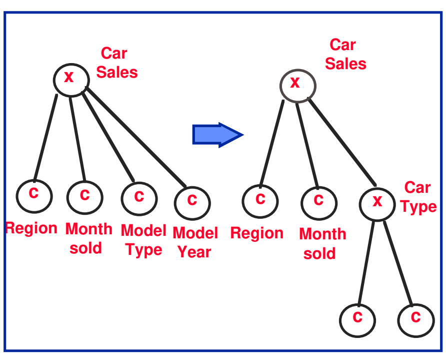
Concept:
- organize dimensions in "subject" groups
- use them in same structure
- need to model that as part of metadata

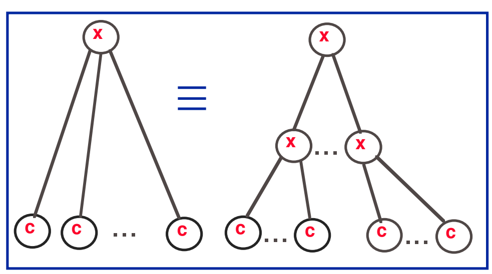
Advantage:
- helps organize high dimensionality
- simplifies use interfaces

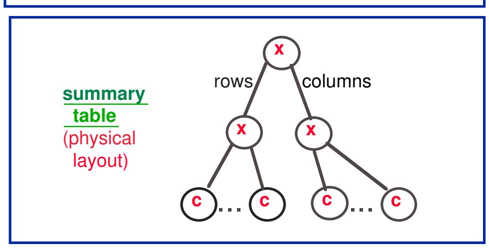
"subject" grouping of dimensions

Computing Sciences Directorate, LBNL

graph models for SDBs (cont'd)







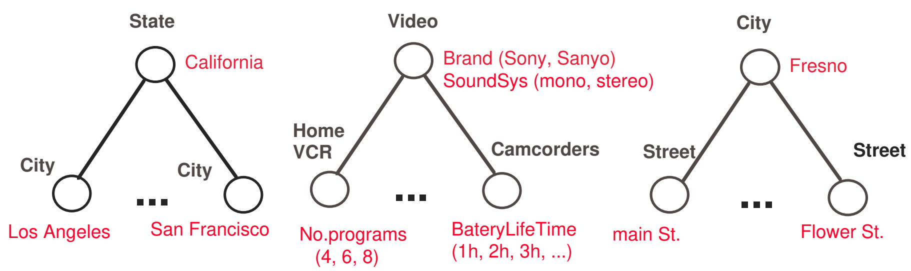
Concept:
- grouping concept generalizes into hierarchies
- can be used to represent summary tables (rows / columns), order of nodes matters.

hierarchy of "subject" grouping

graph models for SDBs (cont'd)

Computing Sciences Directorate, LBNL

- **Variations on classification hierarchy**
  - containment hierarchy
  - inheritance hierarchy
  - ID dependency



Containment **Inheritance** **ID dependency**

[Lehner, Ruf, Teschke 96]

tabular models for SDBs

Computing Sciences Directorate, LBNL

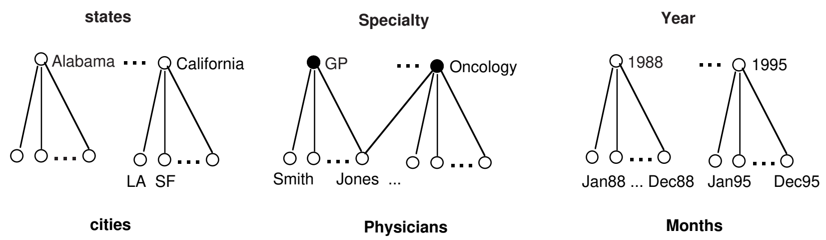
Employment in California			Professional Category: Engineer		Secretary		Teacher			total
			Profession		Profession		Profession			
			Chemical Engineer	Civil Engineer	Junior Secretary	Executive Secretary	Elementary Teacher	high school Teacher	total	
Sex	Male	Year 91	197,700	241,100	534,300	154,100	212,943	123,740	336,683	1,463,883
		92	209,900	278,000	542,100	169,8 00	213,521	145,766	......	......
		. .	....	....	....	......	....	....	......	......
	Female	Year 91	25,800	112,000	667,300	162,,300	216,071	275,123	......	......
		92	28,900	127,600	692,500	174,400	217,520	299,344	......	......
		. .	.....	......	......	......	......	....	......	......

Table with "marginals"

The Summarizability Property

Summarizability:
The property of a Statistical Object to provide CORRECT summary results

Summarizability Conditions:
1. Completeness of a category heirarchy association
2. Strict heirarchy of the hierarchy association
3. Data type compatibility with the category type for a given operation



1. **Association incomplete**

   SUM (population)
   by State

2. **Not a strict hierarchy**

   Count (Physicians)
   by Specialty

3. **Type incompatibility**

   3.1 SUM (population) over months
   Type: SUM, **stock**, temporal - not OK

   3.2 SUM (accidents) over months
   Type: SUM, **flow**, temporal - OK

[Lenz & Shoshani 97]

tabular models for SDBs (cont'd)

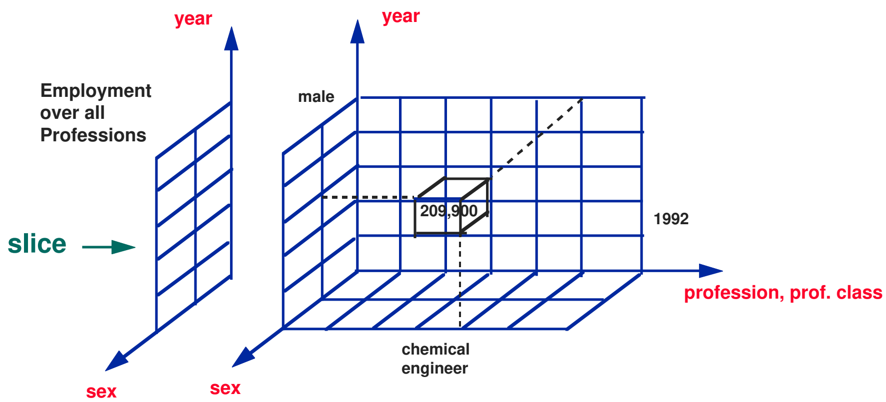
State	county	year	race	sex	age-group	population	avg. income
Alabama	Autauga	1990	white	male	1-10	11,763	0
Alabama	Autauga	1990	white	male	11-20	9,763	3,342
Alabama	Autauga	1990	white	male	21-30	15,763	34,342
...	...	...	...	...	31-40	14,566	37,444
...	...	...	...	...	...	11,372	38,776
...	...	...	...	male	ALL (*)	89,483	36,775
...	...	...	...	Female	1-10	8,457	0

Advantages:
- familiar representation
- fits relational model

Disadvantages:
- repetition of category values in columns
- no distinction between category hierarchy and dimensions (5 dimensions, not 6)
- no distinction between category and summary attribute

(*) **[Gray et al 96]**

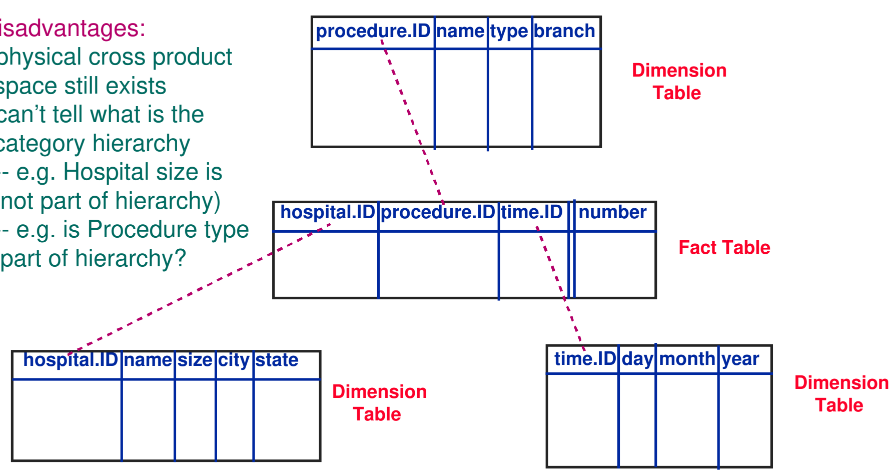
Data Cube Model



Employment in California

Data Cube Model (cont'd)

Disadvantages:
- physical cross product space still exists
- can't tell what is the category hierarchy
  - -- e.g. Hospital size is not part of hierarchy)
  - -- e.g. is Procedure type part of hierarchy?



The STAR Schema

[MicroStrategy]

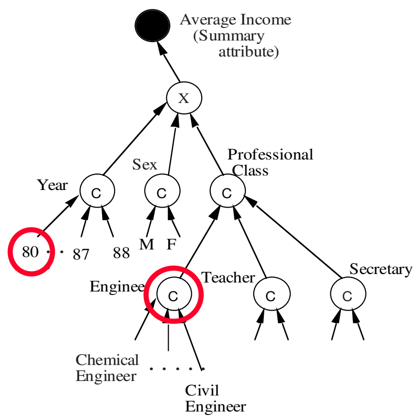
Terminology Correspondence

Statistical DB	OLAP
Category Attribute	Dimension
Category Hierarchy	Dimension Hierarchy (Table)
Category value	dimension value
Summary Attribute	Measures (fact column)
Statistical Object	Data Cube (fact table)
Cross Product	multidimensionality
Summary Table	table / data cube

4. Conceptual modeling - operations

- automatic aggregation concept
- SDB operation extensions to SPJ
- Completeness of SDB operations
- Higher level statistical operations: sampling, standard deviation, etc.
- OLAP operators and correspondence to SDB operators
- SQL extensions for OLAP
- support for automatic classification matching

Automatic Aggregation



Meaning:

produce Average Income
• for year =1980
• for prof. class =Engineers
summarized over:
• all Engineers
• all sexes

[Shoshani 82]

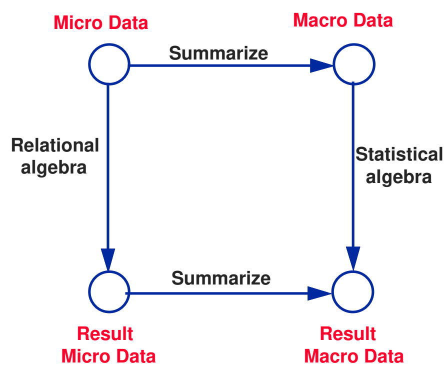
Statistical Operators

- All have some form of aggregation operations but in addition:
- Other operators depending on the data structures assumed
  - summary tables: "attribute split" and "merge" determine how category attributes are organized on rows and columns
  - relational tables: an algebra has operators
    "s-selection" selects subset of categories
    "s-projection"eliminates a category attribute,
    "s-aggregation"goes up a level in category hierarchy
    "s-union"used to merge over partially overlapping category values
    "s-disaggregation" break into categories by proxy attr.

Summary tables: [Ozsoyuglu et al 85] Relational Algebra: [Meo-Evoli et al 92]

Completeness of SDB Operators

- Mostly ignored
- An idea: prove by homomorphism to relational algebra



[Meo-Evoli et al 92]

Higher level statistical operations

- What operators should be supported by the database engine
- Commercial systems: simple aggregates
- Is it enough to interface Statistical Packages (e.g. SAS) to the SDB engine?
- Depends if efficiency is achieved
- Sampling: good example of great increase in efficiency
  - no need to extract large data collections only to sample them outside the system
  - a lot of work: sampling from relational database, from hash files, from B+trees, ...

[Olken & Rotem 95]

OLAP operators and correspondence to SDB operators

- OLAP operators reflect perception of a data cube
- Slice, dice, drill down, roll up, rotate, pivot, ...

Statistical DB	OLAP
S-projection	Slice
S-selection	Dice
S-aggregation	Roll up (consolidation)
S-disaggregation	Drill down
S-union	- - -

Note: 'data analysis' and "multidimensional visualization"
(no specialized terminology)

SQL extensions for OLAP

- To support summaries over MD space (e.g. for "cross tabulation") one needs to aggregate over ALL category values. No concise way for in SQL.
- "CUBE" operator is introduced
- special keyword "ALL" added
- can use shorthand syntax, e.g. "GROUP BY CUBE (State, Year, Sex)"

State	county	year	race	sex	age-group	population	avg. income
Alabama	Autauga	1990	white	male	1-10	11,763	0
Alabama	Autauga	1990	white	male	11-20	9,763	3,342
Alabama	Autauga	1990	white	male	21-30	15,763	34,342
…	…	…	…	…	31-40	14,566	37,444
…	…	ALL	ALL	ALL	ALL	11,372	38,776
ALL	ALL	ALL	ALL	ALL	ALL	89,483	36,775

[Gray et al 96]

support for classification matching

- Different non-overlapping granularities:

Database 1	Database 2
Age Group	Age Group
0-5	0-1
6-10	2-10
11-15	11-20
16-20	21-30

- Time varying categories

industries in 1990	industries in 1991
agriculture	agriculture
automobiles	automobiles
...	...
-- --	Internet

5. Physical organization

- Transposed tables
- Bit transposed files
- Array linearization
- Compression
- Relational vs. Multidimensional OLAP (ROLAP vs. MOLAP)
- Other techniques: materialized views, pre-partitioning of the data cube, MOLAP updates

Transposed Tables

State	county	year	race	sex	age-group	population	avg. income
Alabama	Autauga	1990	white	male	1-10	11,763	0
Alabama	Autauga	1990	white	male	11-20	9,763	3,342
Alabama	Autauga	1990	white	male	21-30	15,763	34,342
…	…	…	…	…	31-40	14,566	37,444
…	…	…	…	…	…	11,372	38,776
…	…	…	…	male	ALL	89,483	36,775
…	…	…	…	Female	1-10	8,457	0

Idea: store each column separately
Reason: usually need to access only a subset of dimensions $ summary columns
Advantage: avoid reading data that is irrelevant

[Turner et al 79]

Bit Transposed Files

State	year	race	sex	age-group
Alabama	1990	white	male	1-10
Alabama	1990	white	male	11-20
Alabama	1990	white	male	21-30
…	…	…	…	31-40
…	…	…	…	…
…	…	…	male	ALL
…	…	…	Female	1-10

encode

State	year	race	sex	age-group
000000	0000	000	0	0000
000000	0000	000	0	0001
000000	0000	000	0	0010
…	…	…	…	0011
…	…	…	…	…
…	…	…	0	1010
…	…	…	1	0000

Idea: usually the number of categories is small and can be encoded into a few bits
Then: store each bit column separately, because better compression can be achieved
Then: compress with “run length encoding”
Note: ROLAP proponents use encoding also

[Wong et al 85]

Array Linearization

state	year	race	sex	age-group	population
Alabama	1990	white	male	1-10	30,173
Alabama	1990	white	male	11-20	13,457
Alabama	1990	white	male	21-30	.....
...	...	...	...	31-40	.....
...	...	...	...	...	
...	...	...	male	91-100	
...	...	...	Female	1-10	.....

state: Alabama, ..., Wyoming
year: 1990, ..., 1996
race: white, Black, ...
sex: male, female
age group: 1-10, ..., 91-100

+

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	7	8	9	10	11	12
3	13	14	....			
4						
5						30

Advantage: no need to store cross product
Note: array linearization is a reversible function
Note: same technique used in MOLAP

Header Compression

	population
1	30,173
2	13,457
2	null
4	null
	14,362
.	null
.	
.	null

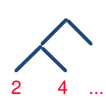
Store non-null values only: [30,173 ; 30,173 ; 14,362, ...]

+

run length sequence: 2, 2, 1, 18, ...

Accumulate: 2, 4, 5, 23, ...

Build B-tree:



Problem: how to deal with sparse arrays
Solution: indexed compression
Index: B-tree on run length "header"

[Eggers et al 81]

ROLAP vs. MOLAP

- Main issue is efficiency of storage and operations
- MOLAP claims
  - Relational tables unnatural for MD data
  - MD arrays provide efficiency in storage and operations
  - Mismatch with SQL
  - ROLAP has to perform specialized operations outside the database system
- ROLAP claims
  - ROLAP integrates naturally with existing technology and standards
  - MOLAP cannot support ad hoc queries effectively
  - Data have to be downloaded into a MOLAP systems, so update is difficult
  - specialized structures can be added: encoding and compression
  - can take advantage of parallel relational technology

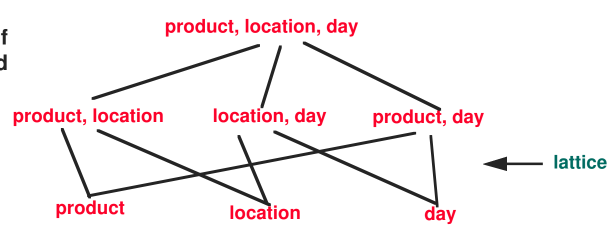
ROLAP: [MicroStrategy], MOLAP: [ArborSoft]

Parallel aggregation: [Shatdal & Naughton 95]

View Materialization

- Space-Time tradeoff
- Problem: which of the 2^n-2 summarization views to materialize given a limited amount of space

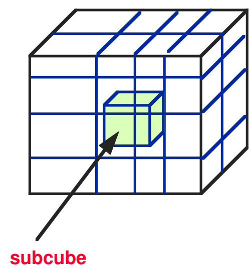
e.g. number of items sold



- Solution is Non-symmetric because of the cardinalities of the dimensions
- Optimal solution based on greedy algorithms

[Harinarayan et al 96]

Pre-Partitioning Data Cube



Idea:
store data cube as multiple subcubes

Advantage:
- suitable for range queries
- read only cubes that have relevant data

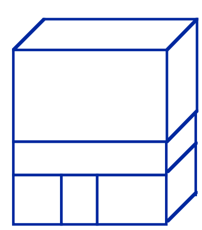
- Symmetric partitioning -- all subcubes same dimensions
- Non-symmetric partitioning -- based on predicted / observed access patterns

Symmetric: [Sarawagi & Stonebraker 94]

Non-symmetric: [Chen et al 95]

MOLAP update techniques

- **Problem:** while OLAP data tend to be read only, transaction data grows continuously. How to update the MOLAP data cube?
- **Approach:** do only incremental updates, but that requires restructuring the entire array
- **Solution:** a data structure called an "extendible array" + a tree-based index structure that keep tracks of the increments
- **Access methods:** Uses retrieval algorithms to support range queries well
- **Note:** method can be used for non-symmetric partitioning



[Rotem & Zhao 96]

6. Privacy

- Inferring information from summary data
- Important negative result
- Main techniques
- Relevance to OLAP

Inferring information from summary data

- **Problem:** given a statistical summary (macro data), is it possible to infer information about a single item (individual) in the micro data
- **Term used:** “query set” -- number of items in the response to a query
  - (blocking sequence of queries for same client - bad idea)
- Limiting size of query set is not enough
- Limiting type of query (e.g. no range queries) is not enough

Important negative result

- Example:

  even if range queries are disallowed, data base can be compromised.

  -- e.g. "find avg. income and count for employee under 65", Then repeat for all employees.
- It was shown that it is possible to compromise a database with multiple queries, especially if several attributes of an individual are known
- A procedure was developed for finding a collections of queries for a desired fact (called a "tracker")

[Denning & Schlorer 80]

Main statistical privacy techniques

- limiting query set intersection
  - keep track of query sets -- can be used with small DBs
- random sample from query set
  - good for very large databases
- pre-partition database
  - cannot see inside these "cell"s -- census uses that
- input data perturbation
  - store perturbed data for general consumption
- output data perturbation
  - perturb results given to users

Refs in: [Shoshani 82], [Data Protection 96]

Importance to OLAP

- Compromising individual people information is illegal
- Privacy problem ignored so far in OLAP
- Example where compromise can occur:
  - Customer habits, total spending, etc.
  - Hospital patients
  - Stock transactions

7. The case for a Statistical Object data type

Statistical/OLAP data type is prevalent and useful

- similar to temporal data types, spatial data types
- has complex data structure
- has unique methods (operations)
- could be enriched with advanced Statistical operators (e.g. sampling, robust statistics - median, trimmed mean, ...)
- could be implemented in the framework of O-O and O-R systems

7. The case for a Statistical Object data type (cont'd)

Statistical/OLAP data type should include:

***Structural properties*:**

- multidimensional data type
- classification hierarchy data type (including time varying categories)
- association of classification hierarchies to dimension
- support for tabular and functional mapping of categories for merging data from multiple sources
  - local statistics merged at regional levels

7. The case for a Statistical Object data type (cont'd)

Statistical/OLAP data type should include:

***Operations*:**

- selection on any classification hierarchy level
- selection of any desired subset of the dimensions
- selection on any subset of the classification controlled values
- automatic aggregation
- overriding statistical operators (e.g. median instead of average)
- apply advanced Statistical operators

7. The case for Statistical Object data type (cont'd)

Computing Sciences Directorate, LBNL

Special attention to:

- support for merging heterogeneous classification structures
- support for multiple types of nulls:
  - nonexistent, don't know value, to be entered later, ...
- support privacy for summarized sensitive data
- versioning of summary attribute
  - provisional,first estimation, final, seasonal adjustments, revisions, ...

8. Conclusions

Computing Sciences Directorate, LBNL

- SDB and OLAP are essentially the same conceptually
- SDBs emphasized modeling, complex classification structures, privacy
- OLAP emphasized "on-line": optimization for very large data sets, use of materialized views
- There is a lot of work in SDB literature relevant to OLAP research (SSDBM conference series)

Conclusions (cont'd)

- Statistical/OLAP Object data type deserves support as a useful specialized data type
- Data cube is a useful paradigm, but does not highlight classification structure
- Relational table representation is confusing and inherently inefficient without special structures