

## *Data Warehousing and Decision Support*

[R&G] Chapter 23, Part A

CS 432

1

---

---

---

---

---

---

---

---

### *Introduction*

- ❖ Increasingly, organizations are analyzing current and historical data to identify useful patterns and support business strategies.
- ❖ Emphasis is on complex, interactive, exploratory analysis of very large datasets created by integrating data from across all parts of an enterprise; data is fairly static.
  - Contrast such **On-Line Analytic Processing (OLAP)** with traditional **On-line Transaction Processing (OLTP)**: mostly long queries, instead of short update Xacts.

CS 432

2

---

---

---

---

---

---

---

---

### *Three Complementary Trends*

- ❖ **Data Warehousing**: Consolidate data from many sources in one large repository.
  - Loading, periodic synchronization of replicas.
  - Semantic integration.
- ❖ **OLAP**:
  - Complex SQL queries and views.
  - Queries based on spreadsheet-style operations and “multidimensional” view of data.
  - Interactive and “online” queries.
- ❖ **Data Mining**: Exploratory search for interesting trends and anomalies.

CS 432

3

---

---

---

---

---

---

---

---

## Data Warehousing

- ❖ Integrated data spanning long time periods, often augmented with summary information.
- ❖ Several gigabytes to terabytes common.
- ❖ Interactive response times expected for complex queries; ad-hoc updates uncommon.

EXTERNAL DATA SOURCES

EXTRACT  
TRANSFORM  
LOAD  
REFRESH

Metadata Repository

DATA WAREHOUSE

SUPPORTS

DATA MINING.

OLAP

CS 432 4

---

---

---

---

---

---

---

---

## Warehousing Issues

- ❖ **Semantic Integration:** When getting data from multiple sources, must eliminate mismatches, e.g., different currencies, schemas.
- ❖ **Heterogeneous Sources:** Must access data from a variety of source formats and repositories.
  - Replication capabilities can be exploited here.
- ❖ **Load, Refresh, Purge:** Must load data, periodically refresh it, and purge too-old data.
- ❖ **Metadata Management:** Must keep track of source, loading time, and other information for all data in the warehouse.

CS 432 5

---

---

---

---

---

---

---

---

## Multidimensional Data Model

- ❖ Collection of numeric measures, which depend on a set of dimensions.
  - E.g., measure **Sales**, dimensions **Product** (key: pid), **Location** (locid), and **Time** (timeid).

Slice locid=1 is shown:

| pid | timeid | locid | sales |
|-----|--------|-------|-------|
| 11  | 1      | 1     | 25    |
| 11  | 2      | 1     | 8     |
| 11  | 3      | 1     | 15    |
| 12  | 1      | 1     | 30    |
| 12  | 2      | 1     | 20    |
| 12  | 3      | 1     | 50    |
| 13  | 1      | 1     | 8     |
| 13  | 2      | 1     | 10    |
| 13  | 3      | 1     | 10    |
| 11  | 1      | 2     | 35    |

CS 432 6

---

---

---

---

---

---

---

---

## MOLAP vs ROLAP

- ❖ Multidimensional data can be stored physically in a (disk-resident, persistent) array; called **MOLAP** systems. Alternatively, can store as a relation; called **ROLAP** systems.
- ❖ The main relation, which relates dimensions to a measure, is called the **fact table**. Each dimension can have additional attributes and an associated **dimension table**.
  - E.g., **Products(pid, pname, category, price)**
  - Fact tables are *much* larger than dimensional tables.

CS 432

7

---

---

---

---

---

---

---

---

## Dimension Hierarchies

- ❖ For each dimension, the set of values can be organized in a hierarchy:



CS 432

8

---

---

---

---

---

---

---

---

## OLAP Queries

- ❖ Influenced by SQL and by spreadsheets.
- ❖ A common operation is to **aggregate** a measure over one or more dimensions.
  - Find total sales.
  - Find total sales for each city, or for each state.
  - Find top five products ranked by total sales.
- ❖ **Roll-up**: Aggregating at different levels of a dimension hierarchy.
  - E.g., Given total sales by city, we can roll-up to get sales by state.

CS 432

9

---

---

---

---

---

---

---

---

## OLAP Queries

- ❖ **Drill-down:** The inverse of roll-up.
  - E.g., Given total sales by state, can drill-down to get total sales by city.
  - E.g., Can also drill-down on different dimension to get total sales by product for each state.
- ❖ **Pivoting:** Aggregation on selected dimensions.
  - E.g., Pivoting on Location and Time yields this **cross-tabulation**:
- ❖ **Slicing and Dicing:** Equality and range selections on one or more dimensions.

|       | WI  | CA  | Total |
|-------|-----|-----|-------|
| 1995  | 63  | 81  | 144   |
| 1996  | 38  | 107 | 145   |
| 1997  | 75  | 35  | 110   |
| Total | 176 | 223 | 339   |

CS 432

10

---

---

---

---

---

---

---

---

## Comparison with SQL Queries

- ❖ The cross-tabulation obtained by pivoting can also be computed using a collection of SQL queries:

```
SELECT SUM(S.sales)
FROM Sales S, Times T, Locations L
WHERE S.timeid=T.timeid AND S.timeid=L.timeid
GROUP BY T.year, L.state
```

```
SELECT SUM(S.sales)
FROM Sales S, Times T
WHERE S.timeid=T.timeid
GROUP BY T.year
```

```
SELECT SUM(S.sales)
FROM Sales S, Location L
WHERE S.timeid=L.timeid
GROUP BY L.state
```

CS 432

11

---

---

---

---

---

---

---

---

## The CUBE Operator

- ❖ Generalizing the previous example, if there are  $k$  dimensions, we have  $2^k$  possible SQL GROUP BY queries that can be generated through pivoting on a subset of dimensions.
- ❖ **CUBE pid, locid, timeid BY SUM Sales**
  - Equivalent to rolling up Sales on all eight subsets of the set {pid, locid, timeid}; each roll-up corresponds to an SQL query of the form:

Lots of work on optimizing the CUBE operator!

```
SELECT SUM(S.sales)
FROM Sales S
GROUP BY grouping-list
```

CS 432

12

---

---

---

---

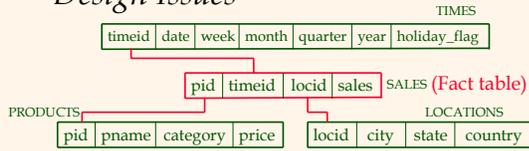
---

---

---

---

## Design Issues



- ❖ Fact table in BCNF; dimension tables un-normalized.
  - Dimension tables are small; updates/inserts/deletes are rare. So, anomalies less important than query performance.
- ❖ This kind of schema is very common in OLAP applications, and is called a **star schema**; computing the join of all these relations is called a **star join**.

CS 432

13

---

---

---

---

---

---

---

---

## Implementation Issues

- ❖ New indexing techniques: Bitmap indexes, Join indexes, array representations, compression, precomputation of aggregations, etc.
- ❖ E.g., Bitmap index:

Bit-vector: **1 bit for each possible value.**  
*Many queries can be answered using bit-vector ops!*

| sex | custid | name | sex | rating | rating |
|-----|--------|------|-----|--------|--------|
| 10  | 112    | Joe  | M   | 3      | 00100  |
| 10  | 115    | Ram  | M   | 5      | 00001  |
| 01  | 119    | Sue  | F   | 5      | 00001  |
| 10  | 112    | Woo  | M   | 4      | 00010  |

CS 432

14

---

---

---

---

---

---

---

---

## Join Indexes

- ❖ Consider the join of Sales, Products, Times, and Locations, possibly with additional selection conditions (e.g., country="USA").
  - A **join index** can be constructed to speed up such joins. The index contains  $[s,p,t,l]$  if there are tuples (with sid)  $s$  in Sales,  $p$  in Products,  $t$  in Times and  $l$  in Locations that satisfy the join (and selection) conditions.
- ❖ **Problem:** Number of join indexes can grow rapidly.
  - A variation addresses this problem: For each column with an additional selection (e.g., country), build an index with  $[c,s]$  in it if a dimension table tuple with value  $c$  in the selection column joins with a Sales tuple with sid  $s$ ; if indexes are bitmaps, called **bitmapped join index**.

CS 432

15

---

---

---

---

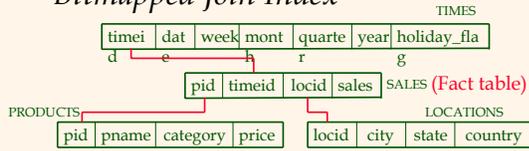
---

---

---

---

## Bitmapped Join Index



- ❖ Consider a query with conditions price=10 and country="USA". Suppose tuple (with sid) s in Sales joins with a tuple p with price=10 and a tuple l with country="USA". There are two join indexes; one containing [10,s] and the other [USA,s].
- ❖ Intersecting these indexes tells us which tuples in Sales are in the join and satisfy the given selection.

CS 432

16

---

---

---

---

---

---

---

---

## Querying Sequences in SQL:1999

- ❖ Trend analysis is difficult to do in SQL-92:
  - Find the % change in monthly sales
  - Find the top 5 product by total sales
  - Find the trailing  $n$ -day moving average of sales
  - The first two queries can be expressed with difficulty, but the third cannot even be expressed in SQL-92 if  $n$  is a parameter of the query.
- ❖ The WINDOW clause in SQL:1999 allows us to write such queries over a table viewed as a sequence (implicitly, based on user-specified sort keys)

CS 432

17

---

---

---

---

---

---

---

---

## The WINDOW Clause

```
SELECT L.state, T.month, AVG(S.sales) OVER W AS movavg
FROM Sales S, Times T, Locations L
WHERE S.timeid=T.timeid AND S.locid=L.locid
WINDOW W AS (PARTITION BY L.state
ORDER BY T.month
RANGE BETWEEN INTERVAL '1' MONTH PRECEDING
AND INTERVAL '1' MONTH FOLLOWING)
```

- ❖ Let the result of the FROM and WHERE clauses be "Temp".
- ❖ (Conceptually) Temp is partitioned according to the PARTITION BY clause.
  - Similar to GROUP BY, but the answer has one row for each row in a partition, not one row per partition!
- ❖ Each partition is sorted according to the ORDER BY clause.
- ❖ For each row in a partition, the WINDOW clause creates a "window" of nearby (preceding or succeeding) tuples.
  - Can be value-based, as in example, using RANGE
  - Can be based on number of rows to include in the window, using ROWS clause
- ❖ The aggregate function is evaluated for each row in the partition using the corresponding window.
  - New aggregate functions that are useful with windowing include RANK (position of a row within its partition) and its variants DENSE\_RANK, PERCENT\_RANK, CUME\_DIST.

CS 432

18

---

---

---

---

---

---

---

---

## Top N Queries

- ❖ If you want to find the 10 (or so) cheapest cars, it would be nice if the DB could avoid computing the costs of all cars before sorting to determine the 10 cheapest.
  - **Idea:** Guess at a cost  $c$  such that the 10 cheapest all cost less than  $c$ , and that not too many more cost less. Then add the selection  $\text{cost} < c$  and evaluate the query.
    - If the guess is right, great, we avoid computation for cars that cost more than  $c$ .
    - If the guess is wrong, need to reset the selection and recompute the original query.

CS 432

19

---

---

---

---

---

---

---

---

## Top N Queries

```
SELECT P.pid, P.pname, S.sales
FROM Sales S, Products P
WHERE S.pid=P.pid AND S.locid=1 AND S.timeid=3
ORDER BY S.sales DESC
OPTIMIZE FOR 10 ROWS
```

```
SELECT P.pid, P.pname, S.sales
FROM Sales S, Products P
WHERE S.pid=P.pid AND S.locid=1 AND S.timeid=3
AND S.sales > c
ORDER BY S.sales DESC
```

- ❖ `OPTIMIZE FOR` construct is not in SQL:1999!
- ❖ Cut-off value  $c$  is chosen by optimizer.

CS 432

20

---

---

---

---

---

---

---

---

## Online Aggregation

- ❖ Consider an aggregate query, e.g., finding the average sales by state. Can we provide the user with some information before the exact average is computed for all states?
  - Can show the current “running average” for each state as the computation proceeds.
  - Even better, if we use statistical techniques and sample tuples to aggregate instead of simply scanning the aggregated table, we can provide bounds such as “the average for Wisconsin is  $2000 \pm 102$  with 95% probability.”
    - Should also use nonblocking algorithms!

CS 432

21

---

---

---

---

---

---

---

---

### *Summary*

- ❖ Decision support is an emerging, rapidly growing subarea of databases.
- ❖ Involves the creation of large, consolidated data repositories called data warehouses.
- ❖ Warehouses exploited using sophisticated analysis techniques: complex SQL queries and OLAP “multidimensional” queries (influenced by both SQL and spreadsheets).
- ❖ New techniques for database design, indexing, view maintenance, and interactive querying need to be supported.

CS 432

22

---

---

---

---

---

---

---

---