Data Structures for Range-Sum Queries

The Evolution of the Data Cube

Paul Butler

CUMC 2010
Waterloo, Ontario

July 2010

The Range-Sum Problem: Example

Name	DOB	City	Height	Siblings	Pets
Joseph Matthews	Jan 9, 1987	Waterloo	172	2	1
Sarah Farmer	Nov 17, 1988	Halifax	167	0	2
⋮	⋮	⋮	⋮	⋮	⋮

- How many people in Vancouver were born before 1990?
- What is the average number of siblings for people above 170 cm?
- What is the total number of pets owned by people 18 to 24 in Calgary?

The Range-Sum Problem: Definitions

Definition (Measure)

A column whose values we want to aggregate in our queries.

Example

The **Pets** and **Siblings** columns from the last example are **measures**

Definition (Dimension)

A column we want to use to select columns which belong to our aggregation.

Example

The **DOB**, **City**, and **Height** columns are **dimensions**.

Dimensions vs. Measures

- For each column, **dimension** vs. **measure** depends on which questions you want to answer.
- *What is the average age of people with two pets*
  - **DOB** would be a **measure**
  - **Pets** would be a **dimension**

The Naïve Approach

- Store the data as a table
- For every row:
  - If the dimension columns match our query, add the value in the measure column to a running total
- Return the running total

Too slow for large amounts of data!

Data Cubes

We can aggregate the data into a multi-dimensional array. [1]

		DOB						
		...	1985	1986	1987	1988	1989	...
Height	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱
	165	...	192	342	558	56	591	...
	166	...	325	275	707	855	484	...
	167	...	487	326	363	193	350	...
	168	...	326	363	193	350	422	...
	169	...	438	456	550	385	412	...
	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱

Now cells not selected by the query are ignored.
Less lookups, faster queries (but still too slow).

Data Cubes

We can calculate partial sums for each row, column, etc. [1]

		DOB							
		...	1985	1986	1987	1988	1989	...	Sum
Height	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	165	...	192	342	558	56	591	...	10937
	166	...	325	275	707	855	484	...	10998
	167	...	487	326	363	193	350	...	11064
	168	...	326	363	193	350	422	...	10913
	169	...	438	456	550	385	412	...	11347
	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	Sum	...	8121	8255	8206	8820	8026	...	202169

Queries which only mention *some* dimensions run faster.

Data Cubes

Example

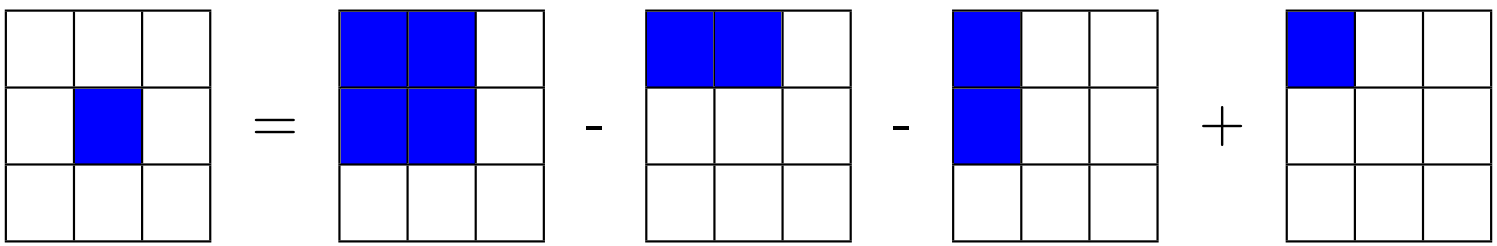
How many dogs are owned by people born between 1986 and 1988 (inclusive)?

		DOB							
		...	1985	1986	1987	1988	1989	...	Sum
Height	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	165	...	192	342	558	56	591	...	10937
	166	...	325	275	707	855	484	...	10998
	167	...	487	326	363	193	350	...	11064
	168	...	326	363	193	350	422	...	10913
	169	...	438	456	550	385	412	...	11347
	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	Sum	...	8121	**8255**	**8206**	**8820**	8026	...	202169

$8255 + 8206 + 8820 = \mathbf{25281}$

Prefix-Sum Table

Observation: range sums can be computed as a sum of range queries starting from 0. [2]



Relative Prefix-Sum Table

Δ-tree


Jim Gray, Adam Bosworth, Andrew Layman, Don Reichart, and Hamid Pirahesh.
Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals.
pages 152–159, 1996.


Ching-Tien Ho, Rakesh Agrawal, Nimrod Megiddo, and Ramakrishnan Srikant.
Range queries in olap data cubes.
SIGMOD Rec., 26(2):73–88, 1997.

Slides

`github.com/paulgb/cumc2010/raw/master/slides.pdf`

Contact

pbutler@uwaterloo.ca

Web

paulbutler.org

My Slide

A displayed formula:

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

An itemized list:

- itemized item 1
- itemized item 2
- itemized item 3

Theorem

In a right triangle, the square of hypotenuse equals the sum of squares of two other sides.