IMPLEMENTING DATA CUBE EFFICIENTLY

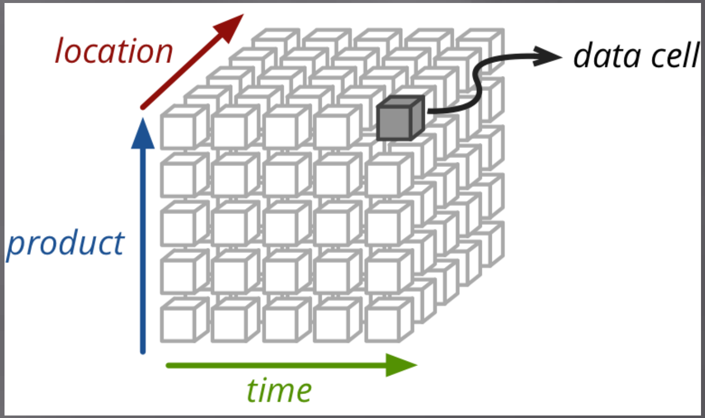
Navjeet Singh
(presenting)

UNIVERSITY OF WATERLOO

Decision Support System & OLAP cube

- A **Decision Support System** (**DSS**) is a computer-based information system that supports business or organizational decision making activities.
- The DSS users need summary information that is locked away in the operational systems to understand the trend of transactions that are taking place in their business.
- One way to represent summary information is presented to them in an graphical environment as a multidimensional "OLAP cube".
- A cube is a way of storing data in a multidimensional form.

An example of cube :
Dimensions –
Product
Location
Time
Measure –
Sales



- Each cell (l,p,t) in this 3D data cube, we store the aggregate of sales of product(p) that sold to location(l) at time(t).

But there is a problem!!

- Every time we needed the cube we had to compute these aggregates from raw data inside a data warehouse.
- Given the size of raw data and complexity of user's query it takes time to aggregate the data and create a 'Data cube'

The solution !!

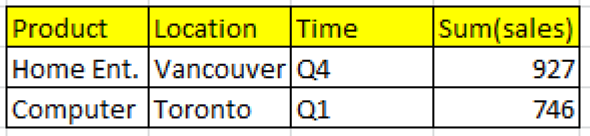
- Physically materialize the whole data cube.
- In other words, have pre-computed tables that hold the aggregate vales of these cells in your database.
- This approach gives a better query response time over the computing from raw data

How do we materialize?

- Cells that are similar to each other form a Cell Set.

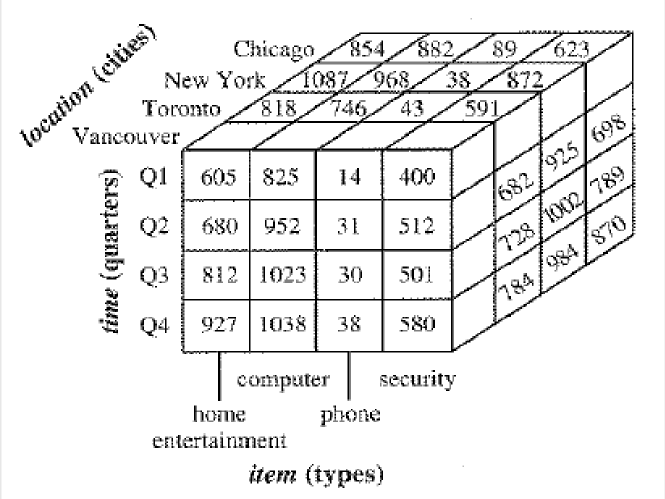
- Each cell set can be materialized into a table.

- For example:-.
  - We can have a materialized cell set consisting of individual cells.



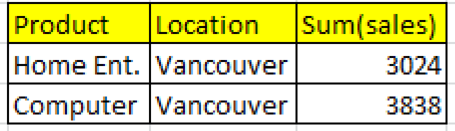
Product	Location	Time	Sum(sales)
Home Ent.	Vancouver	Q4	927
Computer	Toronto	Q1	746

Which is equivalent to SQL query have a group by on Product, location and Time.



Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

- Or we can materialize set of cells grouped by Product and Location



Product	Location	Sum(sales)
Home Ent.	Vancouver	3024
Computer	Vancouver	3838

Cont.

- We can have 8 different Cell set based on combination of group by’s in above case:
  - Product, Time, Location
  - Product, Time
  - Time, Location
  - Product, Location
  - Product
  - Location
  - Time
  - None – no group by

But there is still a problem!!

Space Constraint!

Space Constraint

Due to large size and number of data cubes it is not feasible to materialize and store every data cube

The Questions is!

How many and which group by’s we materialized to get reasonable performance and minimum average query cost?

How the algorithm works?

- It chooses the query group which cannot be answered using any other cell set.

- Then uses used the lattice structure with greedy algorithm to determine which other query groups to include

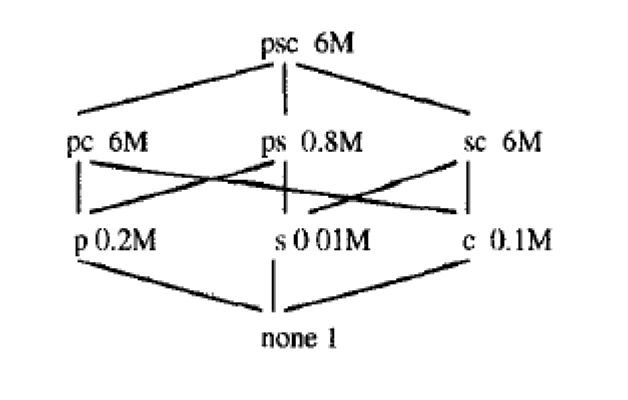
  Here:

  - Lattice structure is a figure which shows how query groups are dependent on each other and what is the cost associated with each query group.
  - An Example of dependency – ‘product’ is dependent on ‘product, customer’ if ‘Product’ can answered using ‘product, customer’
  - And, Cost is proportional to amount of space consumed the query group.

An example of lattice (from paper)

Some of the dependencies within Query groups are:-

- pc ~>psc
- sc~>psc
- ps~>psc



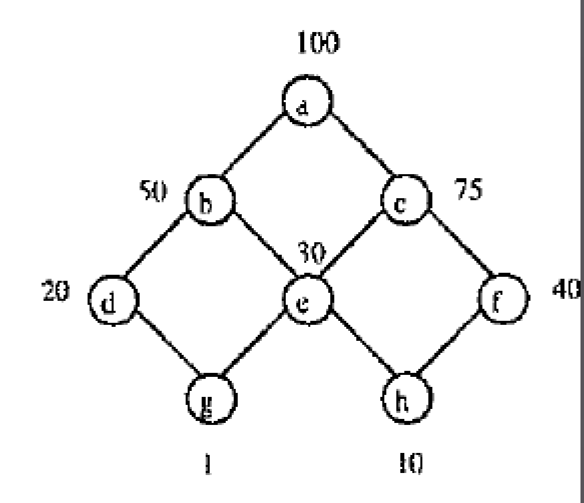
Greedy Algorithm

- **Greedy algorithm** selects the best query group which is best choice given what has before.
- Its does it by calculating the benefit of a query groups by considering how it can improve the cost the evaluating other groups including itself.

- An example of **Greedy Algorithm**

Given:-
- Eight groups named 'a' to 'h' with space cost on top.
- 'a' is by default chosen to be materialized.

Need:-
- Choose three more query groups from a set of 'b' to 'h'

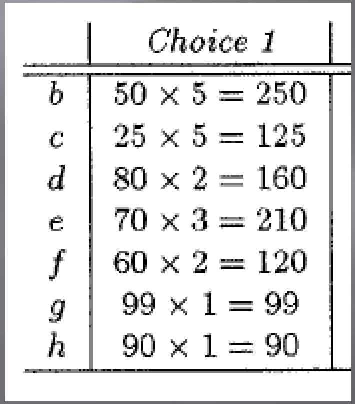


Also:-
We begin with the assumption that each group is evaluated using a, and will therefore have a cost of 100 per group. And, the total cost is 800.

Cont.

Suppose we pick ‘b’

- Compared to A it will reduce it’s cost by 50 and the cost of each of the groups d, e, g and h below it. So the benefit of ‘b’ is 50X5 =250

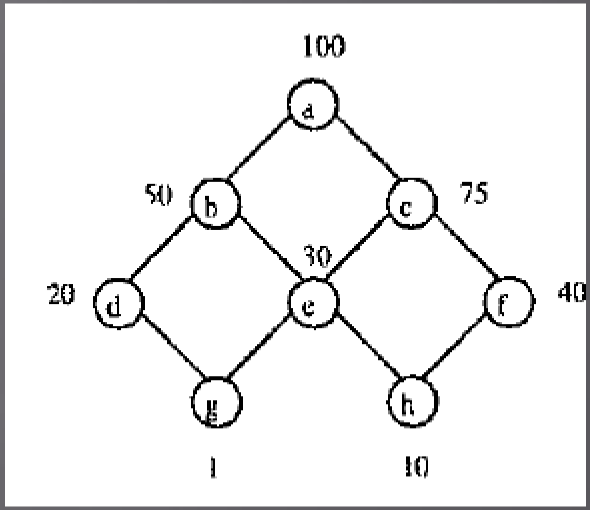


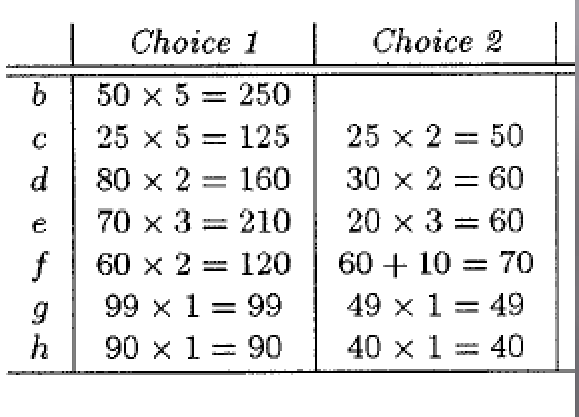
	Choice 1
b	$50 \times 5 = 250$
c	$25 \times 5 = 125$
d	$80 \times 2 = 160$
e	$70 \times 3 = 210$
f	$60 \times 2 = 120$
g	$99 \times 1 = 99$
h	$90 \times 1 = 90$

‘b’ has the highest benefit, so we chose that one to be first materialized group.

We recalculate the benefit of every other Groups, given that the group ‘a’ and ‘b’ are materialized:

Either from ‘b’ at a cost of 50, if ‘b’ is above it
Or ‘a’ at a cost of 100m if ‘a’ is above it





	Choice 1	*Choice 2*
b	$50 \times 5 = 250$	
c	$25 \times 5 = 125$	$25 \times 2 = 50$
d	$80 \times 2 = 160$	$30 \times 2 = 60$
e	$70 \times 3 = 210$	$20 \times 3 = 60$
f	$60 \times 2 = 120$	$60 + 10 = 70$
g	$99 \times 1 = 99$	$49 \times 1 = 49$
h	$90 \times 1 = 90$	$40 \times 1 = 40$

Our second choice is ‘f’ since it has the highest benefit of 70, because 60 due to Itself over ‘a’ and 10 on ‘h’ over b.

THANK YOU