Counting the Cell

Lu An Tang

10/18/11

department of
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

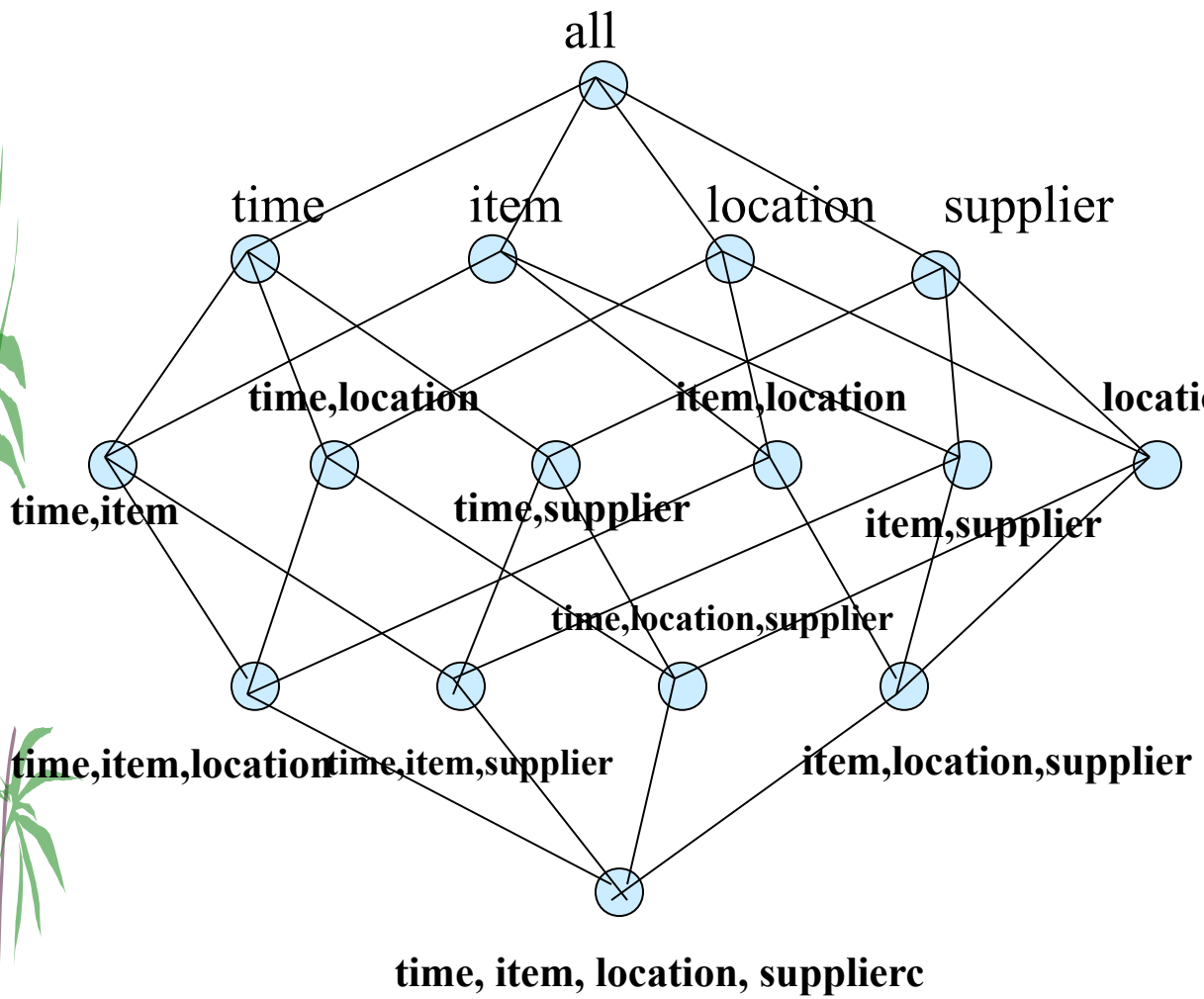
Outline

- Terms used in this problem
- Counting the cuboids and cells
- Suggestion for the exams

First, Let's make clear some terms

- The following descriptions are not formal, just help you to understand the intuitions of those concepts:
- (Data) Cube: General name of the data model, it refers to the technique or model, such as Iceberg cube, ranking cube, text cube …… It is non-countable ☺
- Cuboid: data summarization on a subset of cube's dimensions (e.g., [location, time, branch; sales (measure)])
- Cell: a specific unit of the cuboid (e.g, [Urbana, 2009, Wal-Mart; 10M sales])

Data Cube is a lattice of cuboids (CS412 PPT)



0-D(apex) cuboid

1-D cuboids

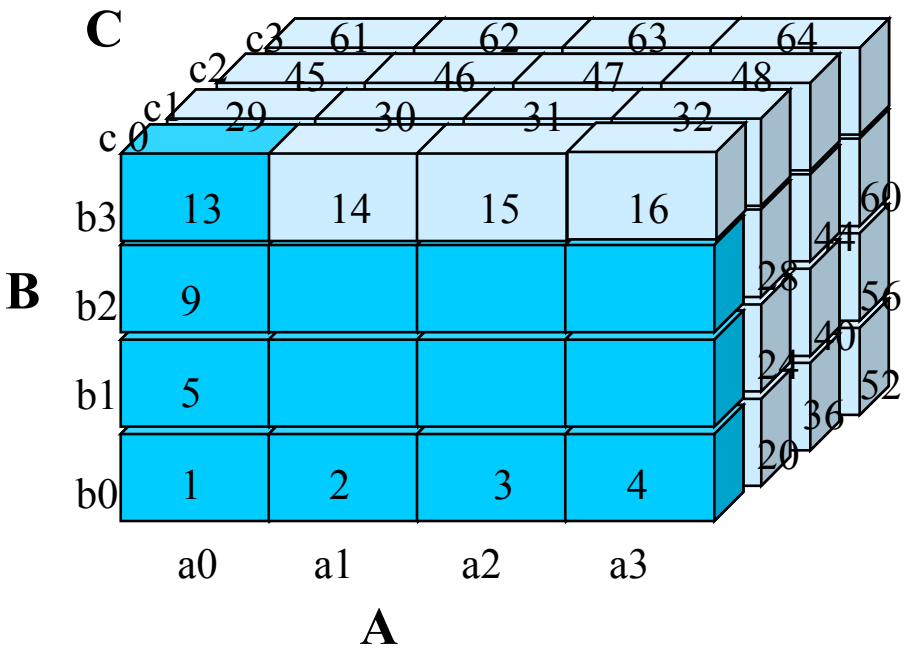
2-D cuboids

3-D cuboids

4-D(base) cuboid

Cuboid is a set of cells (CS412 PPT)

- Cuboid ABC is a set of cells, such as (a0,b0,c0), (a0,b0,c2) … The measure is omitted here



Base and aggregate cells/cuboids

- Suppose there are totally n dimensions
- The fact table (n-D cuboid) is called a base cuboid, the cells in base cuboid are called base cells or base tuples. Indeed they are the relational tuples
- The top most 0-D cuboid is called the apex cuboid
- Except the base cuboid, all the cuboids are aggregate cuboids, their cells are aggregate cells

Non empty

- For example, if there is no record of Chicago's Wal-Mart in the fact table
- Then, the cell (Chicago, 2009-10, Wal-Mart) is empty
- And all the following cells are empty, too
  - (Chicago, 2009-9, Wal-Mart); (Chicago, 2009, Wal-Mart)
  - (Chicago, *, Wal-Mart)
- Will there be "empty cuboid"?
- As long as there is one tuple existing in the fact table, there will be no empty cuboid, e,g., (Location, time, branch) will never be empty even if there is only one tuple in fact table

Short Summary

- Cube: A general name, you will never be asked to count a "cube"
- Cuboid: A subset of the dimensions in data cube, if the fact table is not empty (no one will ask you to count an empty cube…), then there will be no empty cuboids
- Cell: A specific unit of the cuboid
- Aggregate Cuboid: All the cuboids except the base cuboid

Outline

- Terms used in this problem
- Counting the cuboids and cells
- Suggestion for the exams

Example Question

- Assume a base cuboid of 20 dimensions contains 3 base cells, as shown below,
  - (1) (a1, b2, b3; b4…… , b20)
  - (2) (b1; a2; b3; b4 ……, b20)
  - (3) (b1, b2, a3; b4 ……, b20)
- where $a_i \neq b_i$. The measure is count.
- 1. How many nonempty aggregated cuboids will this full data cube contain?
- 2. How many nonempty aggregated cells are there?
- 3 How many nonempty aggregated cells are there for the iceberg cube with count ≥ 2 ?

The symbols we will use in the following slides

- More generally, suppose there are n dimensions in the cube, and p base cells in the base cuboid. Among the p cells, there are c *common dimensions*
- Example
  - (1) (a1, b2, b3; b4…… , b20)
  - (2) (b1; a2; b3; b4 ……, b20)
  - (3) (b1, b2, a3; b4 ……, b20)
- Total dimension n = 20
- Number of base cell p = 3
- Common dimension (b4 … b20) c = 20-3 =17

Counting the Number of Cuboids

- Rule 1: The number of cuboids has no relation to the number of base cells
- If there is no concept hierarchy (most questions are in this case), the formula to compute the cuboid is
- Total Cuboid number = 2^n (n is the number of dimension)
- Example question 1:
- n = 20, the total number of cuboids is 2^{20}.
- Number of nonempty aggregated cuboids = $2^{20} - 1$

Strategy for Counting the Number of Cells

- Calculate the total number of cells;
- Calculate the overlapping cells;
- Answer = Total – overlapping cells*overlapping times

What is overlapping?

- Suppose we have two base cells
  - (Urbana, 2009, Wal-Mart) count: 1
  - (Chicago, 2009, Wal-Mart) count: 1
- If we roll up in the location dimension to *, both the two cells will be:
- (*, 2009, Wal-Mart) count: 2
- Here two cells overlap as one cell. The cell's count is 2. We call it overlap once/1 time.

Calculate the total number of cells

- Formula: Total Cell number = $p*2^n$
- In the example question, p = 3, n = 20
- Total cell number = $3*2^{20}$

Calculate the overlaps I

- Strategy:
- Start from the max count, then reduce one by one, stop until count = 1
- The max possible count in the example is 3 (since there are only 3 base cells in the fact table)
- Except common dimensions, if we set all other dimensions as *, no matte how we select other dimensions, then the count will always be 3. e.g.,
  - (*,*,*, b4……b18, * , b20) count 3
  - (*,*,*, b4, *, b6……, b20) count 3
  - (*,*,*,*……*) apex cell, count 3
- So how many are those count 3 cells?

Calculate the overlaps II

- Formula: Number of Max count cells (Count 3 cells) = 2^c (c is the number of common dimensions)
- In the example question, c = 17
- Count = 3 cells (overlap 2 times) 2^{17}

- The next step is to compute count = 2 cells (overlap 1 time)

Calculate the overlaps III

- Unfortunately there is no formula in this step, we have to count by ourselves
  - (1) (a1, b2, b3; b4…… , b20)
  - (2) (b1; a2; b3; b4 ……, b20)
  - (3) (b1, b2, a3; b4 ……, b20)
- How can count = 2 cells (overlap once) happen?
- If we set the first two dimensions as "*", then it will be
  - (*, *, b3; b4 ……, b20) count 2
  - (*, *, a3; b4 ……, b20) count 1

Calculate the overlaps IV

- And remember, each common dimensions may be “bi” or *, e.g.,
  - (*, *, b3; b4 ……, b20) count 2
  - (*, *, b3; *, b5……, b20) count 2
  - (*, *, b3; *, ……, *) count 2
- So for this case, there are also 2^{17} cells

Calculate the overlaps V

- So how many count = 2 cells in total?
- In this question, there are three cases, each has 2^{17} cells
  - (b1, *, *)
  - (*, b2, *)
  - (*, *, b3)
- Count = 2 cells (overlap 1 times), 3* 2^{17}

Calculate the final answer

- Question 2. Final Answer is:
- Total cell – overlap twice cells *2 – overlap once cells *1 – base cells
- $3*2^{20} - 2 * 2^{17} - 1* 3* 2^{17} – 3 = 3*2^{20} - 5* 2^{17} – 3$

Calculate the Iceberg cells

- With our previous result, it is quite easy to calculate the iceberg cells
- Iceberg cells = count 3 cells + count 2 cells
- Answer = $2^{17} + \quad 3 * 2^{17} = 4 * 2^{17}$

Outline

- Terms used in this problem
- Counting the cuboids and cells
- Suggestion for the exams

Some suggestions to deal with this problem in exam

- "Counting cells" in the data cube is a classic problem – and it is likely to appear in the mid-term and final exams
- Frankly speaking, this question is the most difficult problem; but in the exam, it usually appears as question 2 or question 3
- Strategy: Control you time spending on this question
- Usually the first sub question is counting the cuboid – and it is easy, you can quickly answer it.

Some suggestions II

- For the question of counting cells, if you are not very confident, you can leave it as the last one to solve in the exam (especially the mid-term's time is only 1.5 hours)
- When solving the question, try to list out every steps
- The calculations are error-prone, but as long as your steps are right, you can get some points even with wrong answer
- An example solution is in the next page

Example Question

- Assume a base cuboid of 20 dimensions contains 3 base cells, as shown below,
  - (1) (a1, b2, b3; b4…… , b20)
  - (2) (b1; a2; b3; b4 ……, b20)
  - (3) (b1, b2, a3; b4 ……, b20)
- where $a_i \neq b_i$. The measure is count.
- 1. How many nonempty aggregated cuboids will this full data cube contain?
- 2. How many nonempty aggregated cells are there?
- 3 How many nonempty aggregated cells are there for the iceberg cube with count ≥ 2 ? (1 point)

Example Solution I

1. Total dimension n = 20; Number of base cell p = 3

Common dimension (b_4 to b_{20}) c = 20-3 =17

The total number of cuboids is 2^{20}. Base cuboid is 1

The number of nonempty aggregated cuboids = $2^{20} - 1$

2. Total cell number = $3*2^{20}$

Count 3 cells (overlap 2 times), 2^{17}

Count 2 cells (overlap 1 times), $3* 2^{17}$; Base cells, 3

The number of nonempty aggregated cells:

$3*2^{20} - 2 * 2^{17} - 1* 3* 2^{17} - 3 = 3*2^{20} - 5* 2^{17} - 3$

Example Solution II

3. Iceberg cells = count 3 cells + count 2 cells

nonempty aggregated cells with count ≥2:

$2^{17} + \quad 3 * 2^{17} = 4 * 2^{17}$

Thank You Very Much! Good luck in the Exams!