Question 1

a. (4', L1) How many cuboids are there in the full data cube?

level 0: 1

level 1: 6

level 2: 15

level 3: 20

level 4: 15

level 5: 6

level 6: 1

Total cuboids: 64

b. (4', L2) How many distinct aggregated (i.e., non-base) cells will the complete cube contain?

for the first two dimension: (*,*) (a1,*) (*, a2) (a1, a2) (b1,*) (*, b2) (b1, b2) (c1,*) (*, c2) (c1, c2)

10*2*2*2*2 - 3(base cases)

Total cuboids: 157

c. (4', L2) How many distinct aggregated cells will an iceberg cube contain, if the condition of the iceberg cube is count ≥ 3?

level 4: (*, *, a3, a4, a5, a6) = 1

level 3: (*, *, a3, a4, a5, *), (*, *, a3, a4, *, a6)… = 4

level 2: (*, *, a3, a4, *, *), (*, *, a3, *, *, a6)… = 6

level 1: (*, *, *, *, a5, *), (*, *, a3, *, *, *)… = 4

level 3: (*, *, *, *, *, *) = 1

Total cuboids: 16

d. (4', L2) How many non-star dimensions does the closed cell with count = 3 have?

(*, *, a3, a4, a5, a6)

has 4 non-star dimension.

Question 2

a. (4', L1) How many cuboids are there in the cube?

(2+1)*(2+1)*(1+1)*(1+1) = 36

Total cuboids: 36

b. (4', L2) How many distinct cells are there in the cuboid (Location[City], Category, Price, Time[Year])?

56 distinct cells

c. (4', L2) If we roll up by climbing up in the Location hierarchy from City to State, how many distinct cells are there in the cuboid (Location[State], Category, Price, Time[Year])?

34 distinct cells

d. (4', L2) How many distinct cells are there in the cuboid (*, Category, Price, Time[Quarter])?

33 distinct cells

e. (4', L2) What is the count for the cell (Location[State] = Illinois, Category = Food,
*, Time[Quarter] = Q1)?

10 distinct cells

f. (4', L2) What is the count for the cell (Location[City] = Chicago, *, Price = cheap,
Time[Year]= 2013)?

4 distinct cells

Question 3

a. (7', L2) If we scan the chunks in the order 1,2,3....,27 when materializing the 2-D cuboids AB, AC and BC, to avoid reading 3-D chunks into memory repeatedly, what is the minimum memory for holding all the related 2-D planes?

AB = 300 * 100 * 3

BC = 100 * 200 * 1

AC = 300 * 200 * 9

Total = AB + AC + BC = 650000

b. (8', L3) Do you think there exist other orders to scan the chunks so that the memory cost is less than that in sub-question (a)? If yes, show that order using chunk numbers (e.g. 1, 2, 3..., 27) and the minimum memory required. Otherwise, explain why.

Yes, there is a way to rearrange the sequence.

If we arrange the sequence like:

1,10,19,4,13,22,7,16,25,2,11,20,5,14,23,8,17,26,3,12,21,6,15,24,9,18,27,

AB = 300 * 100 * 3

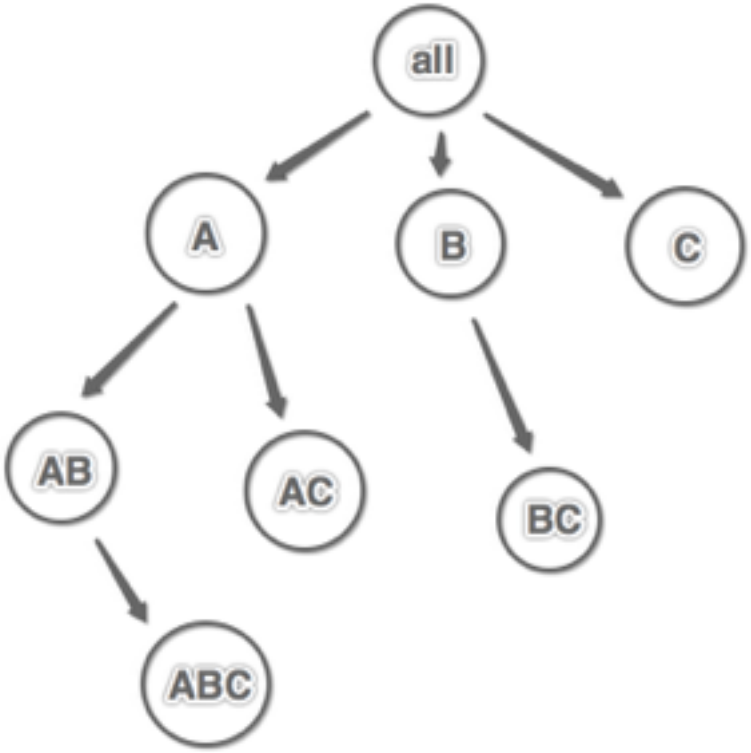
BC = 100 * 200 * 9

AC = 300 * 200 * 1

Total = AB + AC + BC = 330000

Question 4

a. (5', L2) Draw the trace tree of expansion with the exploration order A → B → C.



b. (5', L3) If we set min support = 4 with the exploration order A → B → C, how many cells will be considered/computed? For these cells, please list each of them with its count, and report whether it is expansible in the BUC process. (Hint: For you to better understand the question and know how to answer the question, please refer to SampleQuestionBUC.pdf in Chapter 5 at our course website).

All (∗, ∗, ∗) : 12 - expansion

— — — — — —

A (a_0, ∗, ∗) : 12 - expansion

— — — — — —

AB (a_0, b_0, ∗) : 3

AB $(a_0, b_1, *)$: 3

AB $(a_0, b_2, *)$: 3

AB $(a_0, b_3, *)$: 3

———————

AC $(a_0, *, c_1)$: 4

AC $(a_0, *, c_2)$: 4

AC $(a_0, *, c_0)$: 4

———————

B $(*, b_0, *)$: 3

B $(*, b_1, *)$: 3

B $(*, b_2, *)$: 3

B $(*, b_3, *)$: 3

———————

C $(*, *, c_1)$: 4

C $(*, *, c_2)$: 4

C $(*, *, c_0)$: 4

There are 16 cells to be computed in total.

c. (5', L3) If we set min support = 4 with the exploration order B → A → C, how many cells would be considered/computed? For these cells, please also list each of them with its count and report whether it is expansible in the BUC process.

All $(*, *, *)$: 12 - expansion

———————

B $(*, b_0, *)$: 3

B $(*, b_1, *)$: 3

B $(*, b_2, *)$: 3

B $(*, b_3, *)$: 3

———————

A $(a_0, *, *)$: 12

———————

AC $(a_0, *, c_1)$: 4

AC $(a_0, *, c_2)$: 4

AC $(a_0, *, c_0)$: 4

———————

C $(*, *, c_1)$: 4

C $(*, *, c_2)$: 4

C $(*, *, c_0)$: 4

There are 12 cells to be computed in total.

Question 5

a. Operational update is a very important issue for data warehousing.

False: Only for online database system

b. Suppose we pick two cells A and B from a data cube; A is $(a_0,b_0,*,d_0)$ and B is (a_0, b_0, c_0, d_0). Then, cell A is a child of cell B.

False: Instead, cell B is a child of cell A.

c. In OLAP operations, we can see more detailed data information by rolling up.

False: By rolling up, we see a more general view.

d. The Bottom-Up Computation (BUC) algorithm can be used to compute either the full cube or a partial cube.

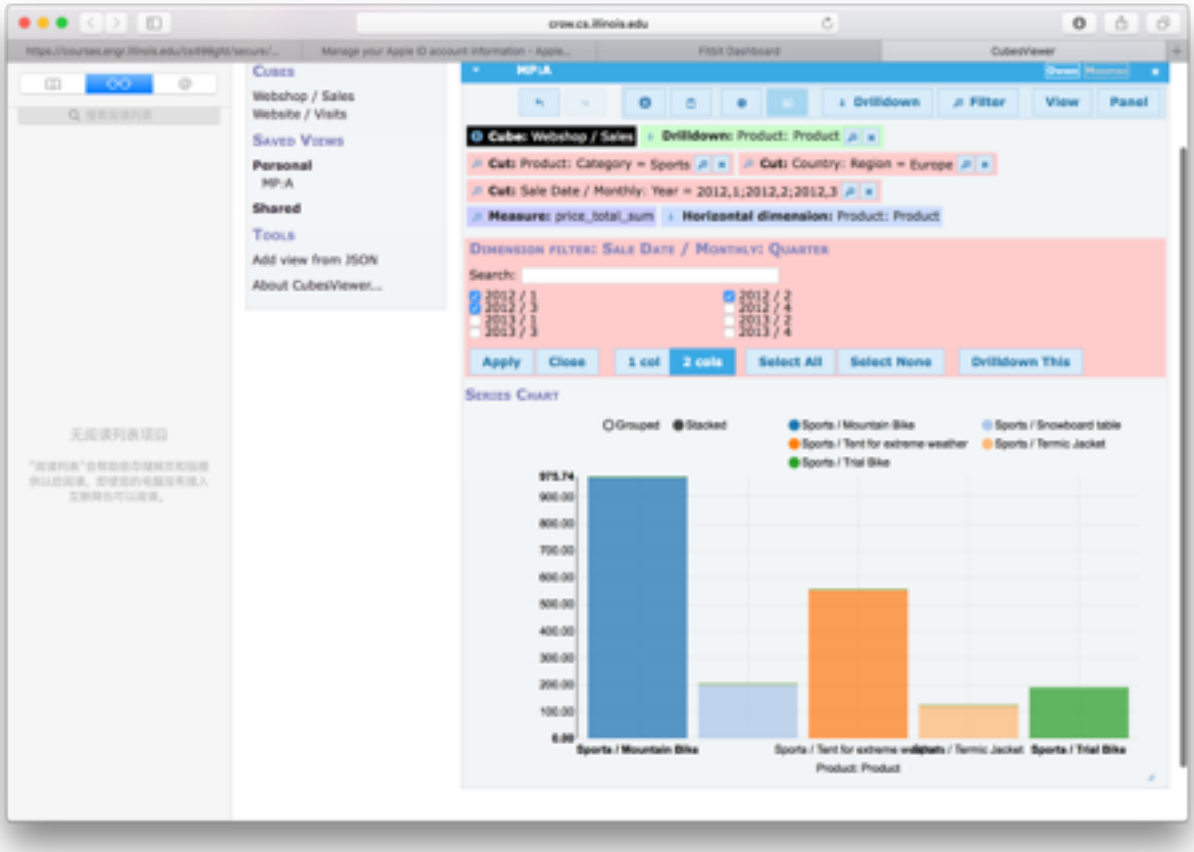
True: BUC will explore the entire data set first and then travel each cube.

e. The Multiway Array Aggregation Computation is most effective when the product of the cardinalities of dimensions is very high.

False: It is effective when the product of the cardinalities of dimensions is low, to minimize the size.

Machine Problem

a. (5', L2) For the dataset Webshop/Sales in CubesViewer, which product in category Sports has the highest revenue in Europe during the first three quarters of the year 2012? And which has the least? List the OLAP operations necessary to reach the cube that can answer the questions above. Show the screenshot of the chart generated for the cube by CubesViewer (you must choose the appropriate measure in the View menu in order to generate the chart).



Most revenue:Moutain Bike

Least revenue:Termic Jacket

OLAP operations:

Drilldown: Product: Product

Add Category = Sport filter

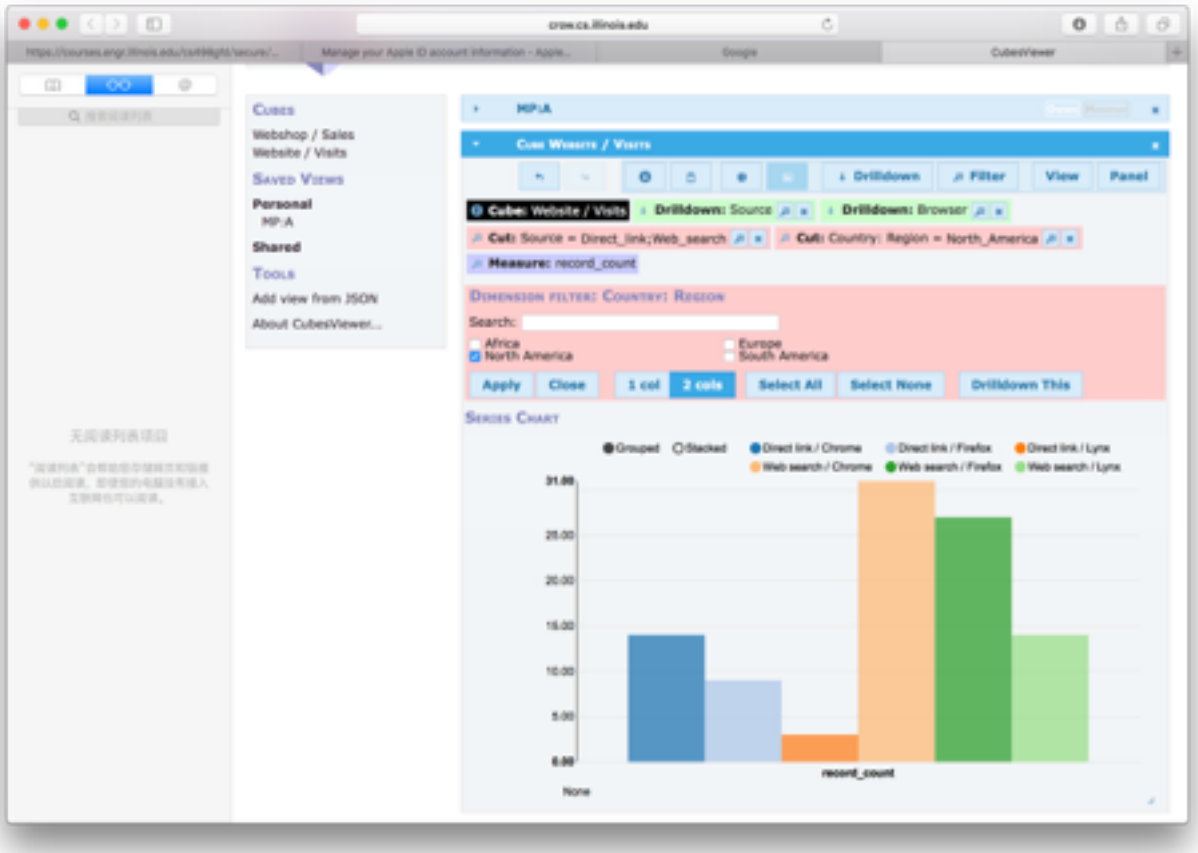
Add Country: Region filter

Add Sales Data filter

Select the measure of Total Price

b. (4', L2) For dataset Website/Visits, what is the most popular way, specified by (source, browser), used by customers from North America to visit the online store? The popularity here is measured by the visit count.

List the OLAP operations necessary to reach the cube that can answer the questions above. Show the screenshot of the chart generated for the resulting cube by CubesViewer.



Most popular: Chrome and Web Search

Least popular: Lynx and Direct Link

OLAP operations:

Drilldown: Source

Drilldown: Browser

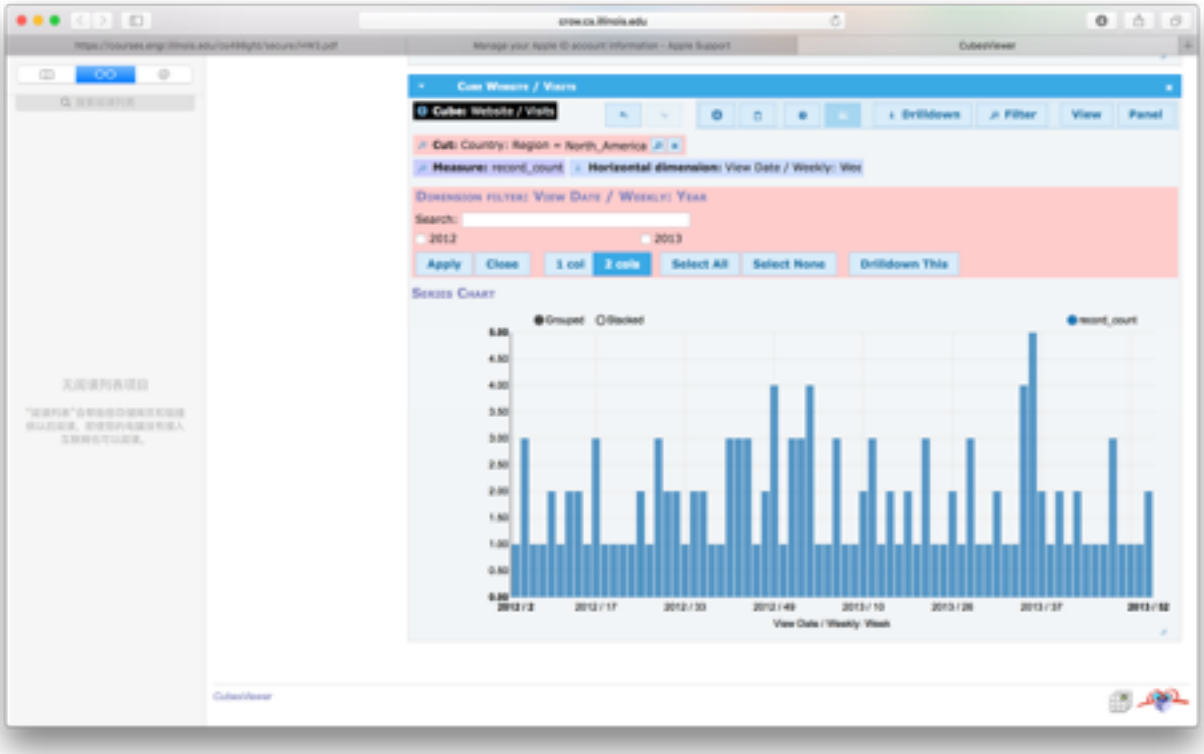
Add Source = Direct_link and Web_search filter

Add Region = North_America filter

Add Sales Data filter

Select the measure of Read Count

c. (3', L2) For Website/Visits, show the screenshot of the chart that describes the changes of the visit counts from North America over time. Draw the chart with one of the granularities: week, month, quarter or year.



d. (8', L3) For each of the datasets Webshop/Sales and Website/Visits, come up with an interesting cube that might help the shop owner make decisions. This is an open question. You will receive full marks by listing the OLAP operations to reach the cubes and what kinds of decisions can be made from those cubes.

The Webshop has to increase its profit by reduce the variety of its product in Europe. It's trying to decide with products to cut from production line. According to the graph, Climbing equipment are the least sold in Europe.

Therefore, they should consider cutting the production line of these two products.

OLAP operations:

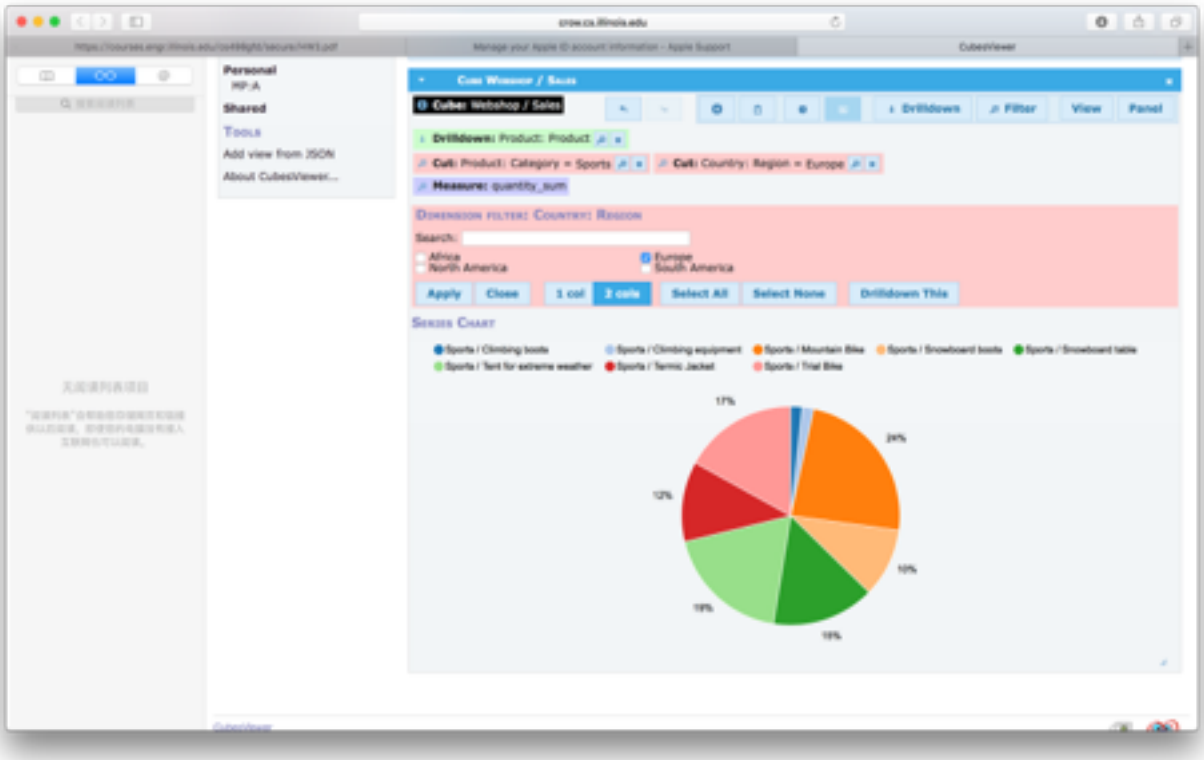
Drilldown: Product: Product

Add Category = Sport filter

Add Region filter

Select the measure of Total Quantity

select PIE graph



The Website is investigate whether it's advertisement on Chrome is successful. The result is, even though most of the traffic came from chrome, the majority of them came from Search(Google). Therefore, the advertisement is not very successful.

OLAP operations:

Drilldown: Source

Add Horizontal dimension Browser

Select the measure of Total Record

Select area Graph