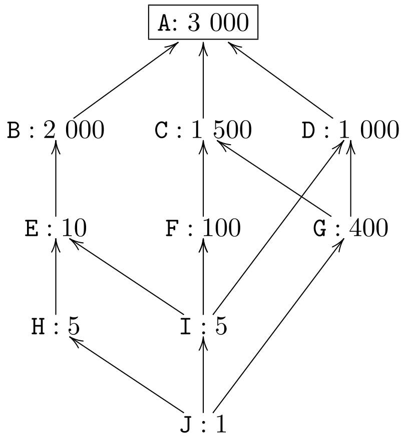
Exercises Data Warehousing
View Materialization

Solutions

1. An important technique to speed up analytical queries is by pre-computing and materializing aggregations. Consider the following lattice of views that can be requested by the user, along with the number of rows in each view.



A is the view representing the base relation. The edges indicate the relation “can be computed from.”

(a) Suppose that only the top-view `A` has been materialized. Select three additional views from the views `B`, `C`, `D`, `E`, `F`, `G`, `H`, `I`, and `J` to materialize. Apply the greedy method described by *Harinarayan, Rajaraman, and Ullman* in their seminal paper “Implementing Data Cubes Efficiently” (SIGMOD 1996)

Solution: Initially we start with the set of materialized views $S = \{A\}$. This top-level view is always in the set of materialized views as none of the other views can be used to generate it. The benefits of the different views are:

B	$5 \times (3\,000 - 2\,000) = 5\,000$
C	$5 \times (3\,000 - 1\,500) = 7\,500$
D	$4 \times (3\,000 - 1\,000) = 8\,000$
E	$\mathbf{4 \times (3\,000 - 10) = 11\,960}$
F	$3 \times (3\,000 - 100) = 8\,700$
G	$2 \times (3\,000 - 400) = 5\,200$
H	$2 \times (3\,000 - 5) = 5\,990$
I	$2 \times (3\,000 - 5) = 5\,990$
J	$1 \times (3\,000 - 1) = 2\,999$

As E gives the highest benefit, this view is selected.

For selecting the second view, we calculate the benefits w.r.t. the set of already selected views $\{A, E\}$:

B	$1 \times (3\,000 - 2\,000) = 1\,000$
C	$\mathbf{3 \times (3\,000 - 1\,500) = 4\,500}$
D	$2 \times (3\,000 - 1\,000) = 4\,000$
F	$1 \times (3\,000 - 100) = 2\,900$
G	$1 \times (3\,000 - 400) = 2\,600$
H	$2 \times (10 - 5) = 10$
I	$2 \times (10 - 5) = 10$
J	$1 \times (10 - 1) = 9$

As now C gives the highest benefit, this view is selected in the second step.

For selecting the third view, we calculate the benefits w.r.t. the set of already selected views $\{A, C, E\}$:

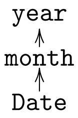
B	$1 \times (3\,000 - 2\,000) = 1\,000$
D	$\mathbf{1 \times (3\,000 - 1\,000) + 1 \times (1\,500 - 1\,000) = 2\,500}$
F	$1 \times (1\,500 - 100) = 1\,400$
G	$1 \times (1\,500 - 400) = 1\,100$
H	$2 \times (10 - 5) = 10$
I	$2 \times (10 - 5) = 10$
J	$1 \times (10 - 1) = 9$

As now D gives the highest benefit, this view is selected in the third step. Hence, the three selected views are: C, D, E.

(b) What benefit gives the additional materialization of these three views under the cost model introduced by these authors?

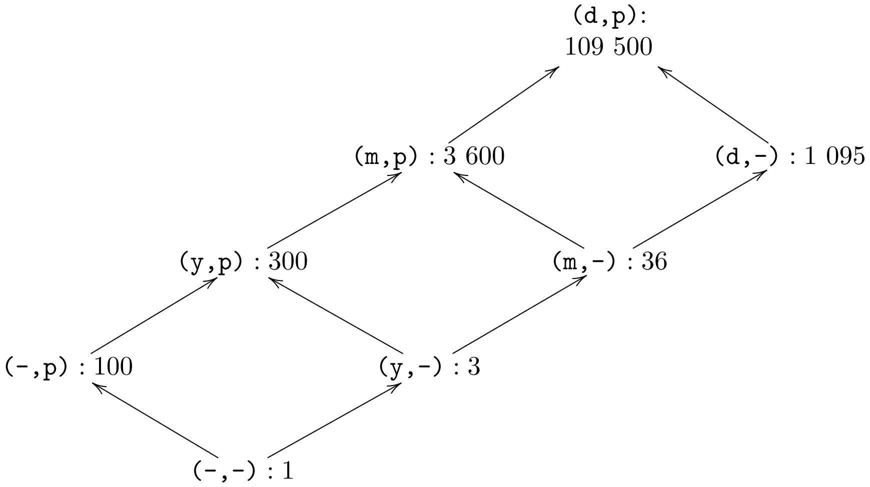
Solution: The total benefit is: 11 960+4 500+2 500 = 18 960

2. Consider a data cube with dimensional attributes `Product` and `Date` and measure `Total Sales`. The following hierarchy is present on the dimension `Date` (No hierarchy on `Product`):



(a) Suppose that the cube is dense; i.e., for every product p and date d there is a tuple (p, d, t) in the base relation with $t > 0$. Furthermore, there are 100 products and 3 years of data (1095 days; no leap year, 157 weeks) in the database. Determine the sizes of the different views.

Solution: Below the partial order between the query is drawn. The number behind the colon indicates the size of the particular view.



(b) Apply the greedy algorithm to select 2 views to materialize.

Solution: Initially we start with the set of materialized views $S = \{(d, p)\}$. This top-level view is always in the set of materialized views as none of the other views can be used to generate it. The benefits of the different views are:

(m,p)	**6 × (109 500 − 3 600) =**	**635 400**
(d,-)	4 × (109 500 − 1 095) =	433 620
(y,p)	4 × (109 500 − 300) =	436 800
(m,-)	3 × (109 500 − 36) =	328 392
(-,p)	2 × (109 500 − 100) =	218 800
(y,-)	2 × (109 500 − 3) =	218 994
(-,-)	1 × (109 500 − 1) =	109 499

As (m, p) gives the highest benefit, this view is selected.

For selecting the second view, we calculate the benefits w.r.t. the set of already selected views $\{(d, p), (m, p)\}$:

(d,-)	**1 × (109 500 − 1 095) + 3 × (3 600 − 1 095) =**	**115 920**
(y,p)	4 × (3 600 − 300) =	13 200
(m,-)	3 × (3 600 − 36) =	10 692
(-,p)	2 × (3 600 − 100) =	7 000
(y,-)	2 × (3 600 − 3) =	7 194
(-,-)	1 × (3 600 − 1) =	3 599

As $(d, -)$ gives the highest benefit, this view is selected.

The two selected views next to the obligatory top view (d, p) are hence: (m, p) and $(d, -)$.

(c) What is the total benefit of materializing those two views?

Solution: The total benefit is 635 400 + 115 920 = 751 320.

3. Imagine a National Bureau of Statistics holding demographic data of the inhabitants of its country. For each person the following attributes are recorded: *City*, *Age*, *Ethnicity*, and *Income*. The bureau wants to make aggregations of this data available for other governmental agencies as well. The agencies can query the data by selecting any subset of the first three attributes. For any combination of the selected attributes the average income of people with this particular combination will be returned. For example, if an agency selects *City* and *Age*, it will get for every city and age, the average income of the people of that age in that city.

For reasons of simplicity we assume that for every city, age, and ethnicity there is at least one person that has this combination. There are 400 cities, 100 age values, and 10 ethnicity types in the database. It is decided that the data will be stored in a database system, but for performance reasons some query answers will be materialized in advance. Every time new data arrives these materialized views are updated.

(a) Suppose 2 views can be materialized. Which views should be materialized? Explain the reasoning behind your answer.

(b) After a few months the bureau realizes that the system's performance is suboptimal. As a first step towards solving this problem, an analysis is made of what is the frequency of the different queries. It turns out that not all 8 combinations of the attributes *City*, *Age*, and *Ethnicity* are equally likely to be requested. The following table summarizes the results of the analysis (every row represents how often one selection of attributes is being requested. The symbol $\surd$ in the column for attribute A in a row indicates that A is in the selection represented by that row):

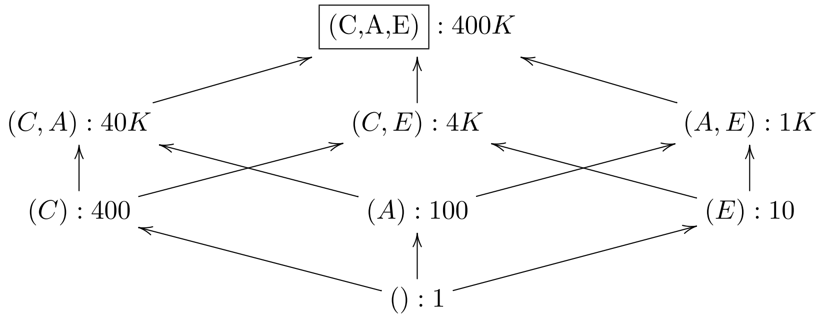
City	*Age*	*Ethnicity*	*Frequency*
✓	✓	✓	5%
✓	✓		20%
✓		✓	30%
✓			30%
	✓	✓	5%
	✓		5%
		✓	5%
			0%

Hence, for example, it turns out that the average income over all people (no attributes selected) is never requested and that, e.g., the average income per city and the average income per city and ethnicity type are disproportionately frequent (together they form 60% of the requests!).

Should the views to materialize be changed based on this information? If so, which views should be materialized instead? Explain your answer.

(c) What additional benefit can be expected if it is decided to materialize also a third view? Use the settings described in (b).

Solution: The partial order between the queries looks as follows:



(a) Clearly materializing (A, E) gives the highest benefit ($4 \times 399K$). Once (A, E) is materialized, the largest benefit is given by materializing (C, E); this gives an additional benefit of $2 \times 396K$.

(b) When we take the frequencies of the queries into account, the benefits in the first step become:

(C, A)	$55\% \times 360K$
$\boxed{(C, E)}$	$\boxed{65\% \times 396K}$
(A, E)	$15\% \times 399K$
(C)	$30\% \times 399\,600$
(A)	$5\% \times 399\,900$
(E)	$5\% \times 399\,990$
$()$	$0\% \times 399\,999$

Clearly, this time (C, E) is selected as the best view to materialize. In the second step, the benefits are:

$\boxed{(C, A)}$	$\boxed{25\% \times 360K}$
(A, E)	$10\% \times 399K + 5\% \times 3K$
(C)	$30\% \times 3\,600$
(A)	$5\% \times 399\,900$
(E)	$5\% \times 3\,990$
$()$	$0\% \times 3\,999$

Hence, (C, A) is selected as the second view to materialize.

(c) For materializing a third view, the benefits are:

$\boxed{(A, E)}$	$\boxed{5\% \times 399K + 5\% \times 39K + 5\% \times 3K}$
(C)	$30\% \times 3\,600$
(A)	$5\% \times 39\,900$
(E)	$5\% \times 3\,990$
$()$	$0\% \times 3\,999$

Hence the additional benefit of materializing a third view would be: $5\% \times 399K + 5\% \times 39K + 5\% \times 3K$.