Data Mining Homework 12

Madis Nõmme

May 20th, 2014

1 Task

> State why for the integration of multiple heterogeneous information sources many companies in industry prefer the update-driven approach (which con- structs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.

Update driven approach and using data warehouses has the following benefits:

- data can be cleaned before running the queries on it
- summary information can be calculated beforehand
- data from different sources can be aggregated to make it possible to use one style of queries (e.g. SQL)

All these make the queries faster and writing them easier (thus saving money on developer salaries and infrastructure costs). This comes with the price of the results possibly being a bit out of date (as the tables contain some precalculated values). Often this is not a problem.

Query driven approach is preferable when:

- there is less data (so the queries on operational database are fast enough)
- the data is in one database. This makes it easier
- getting as 'real time' as possible information is important.
- data is simple enough that the queries on operational database + some aggregation are worth to implement
- there is no knowledge / money or willingness to implement the warehouse solution

2 Task

> A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more useful on practice and state the reasons behind your answer

Both of them describe relational database (or similar) implementations. With star schema there is one central facts table with colums consisting foreign keys for dimension tables. The dimension tables contain actual data. To get the information out, dimension tables are joined with facts table.

With snowflake schema everything is the same and in addition, dimension tables can contain foreign keys to additional dimension tables. This makes the number of joins larger and queries possibly slower.

3 Task

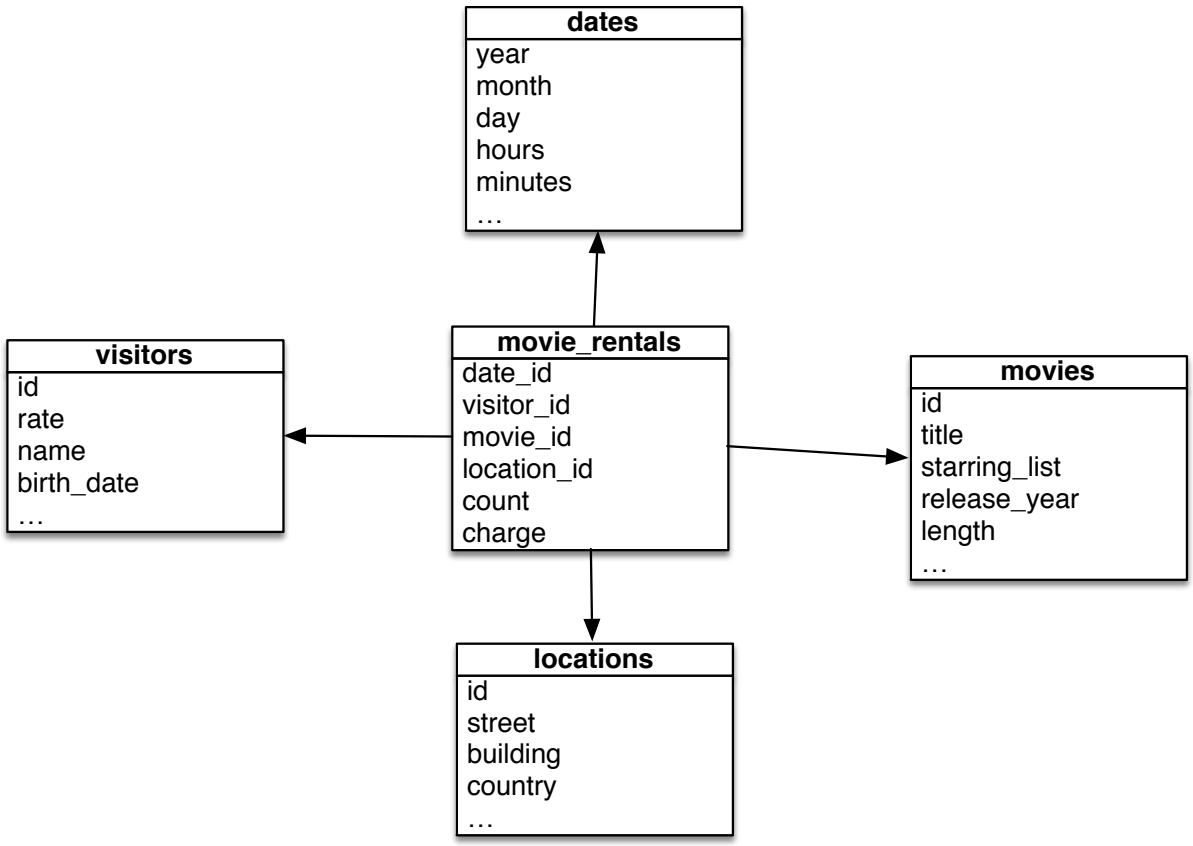
Explain the terms slice and dice, drill-down and roll-up. Provide examples and/or illustrations

1. Slicing - will reduce the dimension by 1 or more. By slicing one could get from [date, visitor, location, movie] to [visitor, location, movie] view.
2. Dicing - is picking (cutting out) smaller subcube of the initial one. It allows inspecting similar subselections of data from different angles. E.g. [date, visitor, location, movie] -> [date, location, movie] of period 1. may to 10th of may
3. Drilling-down - allows inspecting data on more granular level. Dimensions remain the same. E.g. 'zooming in' on date from day granularity to hour granularity.
4. Rolling-up - is the opposite from drilling-down. Allows to get higher level overview of the data. To see the bigger picture.

4 Task

Suppose that a data warehouse consists of the four dimensions: date, visitor, location, and movie, and the two measures, count and charge, where charge is the fare that a visitor pays when watching a movie on a given date. Visitors may be students, adults, or seniors, with each category having its own charge rate.

a. Draw a star schema diagram for the data warehouse.



b. Starting with the base cuboid [date, visitor, location, movie], what specific OLAP operations should one perform in order to list the total charge paid by student visitors at Cinamon Cinema in 2004?

5 Task

Better formatted solution with test data: https://gist.github.com/madis/87ab0b6bbc3dbb4e987b

Rewrite the following query by replacing GROUP BY CUBE(...) clause into an equivalent query using UNION and simple GROUP BY:

~~~~~~~{.sql}

SELECT product, year, city, sum(price*vol)

FROM Orders

GROUP BY CUBE(product, year, city);

~~~~~~~

Rewrite using UNION & GROUP BY

~~~~~~~{.sql}

SELECT product, year, city, sum(vol*price) FROM Orders GROUP BY product, year, city

UNION ALL

SELECT null AS product, year, city, sum(vol*price) FROM Orders GROUP BY year, city, price

UNION ALL

SELECT product, null AS year, city, sum(vol*price) FROM Orders GROUP BY product, city, price

UNION ALL

SELECT product, year, null AS city, sum(vol*price) FROM Orders GROUP BY product, year, price;

~~~~~~~

6 Task

A	B	C	D	E	F	G	H
Filter							
(empty)		Age	State				
		16			17		
Salary	Income	Arizona	California	Colorado	Arizona	California	Colorado
20	520						
	4540						
50	50						
	5050						
	14150						
60	60						
70	70						
	9370						
	21590						
80	80						
	310						
90	90						
	9690						
100	100						
	52100						
110	7010						
	9110						
120	2420						
130	130						
	5930						
140	140						
150	1150						
180	9680						
200	200						
	8200						

Findings:

- Young people under 16 don't get paid or don't participate in the census.
- Pivot tables can be confusing.