Exercise 13: OLAP Cubes

Solution

In this exercise, we analyze the sales data of a fictitious wholesale supplier (taken from the database system benchmark TPC-H [1]) in our favorite spreadsheet application.

1. The TPC-H Dataset as OLAP Cube

Let us get familiar with the dataset. In short, it consists of orders made, each of which is made by a customer and consist of lineitems. Think of an order as shopping cart with several items in it. The items of an order are parts that may be provided by different suppliers. Suppliers and customers come from different nations, which are grouped into regions of the world. Figure 1 shows the schema of the TPC-H dataset.

1. Which table(s) of the TPC-H schema is/are the fact table(s)?

Solution: It depends on what we want to analyze, i.e., on which cube we define on the tables. In general, the fact table is the table that contains the measures. If we analyze orders, then the order is the fact table. We could also analyze line items, in which case we probably want information of the order each item belongs to as well, so we would take the join of lineitem and orders as fact table.

2. What are the measures?

Solution: In lineitems, discount, extendedprice, and quantity can be used as measures; in orders, totalprice, in partsupp, availqty and supplycost, in part, retailprice, and in customer and supplier, acctbal. Intuitively, if it makes sense to calculate the sum or average of an attribute, then we can probably use it as measure.
Figure 1: Schema of the TPC-H dataset.
3. What are the dimensions?

**Solution:** All other attributes (except keys, which are only an artifact of normalization). Some of them are explicit hierarchies, such as the geographical hierarchy for nation and regions. The time dimensions can also be seen as hierarchies: each date can be broken down into year, quarter, month, week, etc. Intuitively, if you can draw a plot where some attribute is the x-axis, then this attribute is probably a dimension.

4. What do you call this flavor of OLAP?

**Solution:** If a cube is stored as relations, we speak of “Relational OLAP” (ROLAP). If instead, the cube is stored in a specialized data structure optimized for multidimensional analysis (possibly with pre-computed aggregates), we speak of “Multidimensional OLAP” (MOLAP). Roughly speaking, MOLAP is faster for the specific set queries the cube was designed for, but it cannot answer ad-hoc queries. ROLAP systems are more flexible as they can answer any query on the cube and intensive research has mitigated the performance disadvantage over the past decade.

2 Analyzing TPC-H with a Pivot Table

Download the file *Exercise13_OLAP_Cubes.xls* from the course web site and open it with your favorite spreadsheet application. The file contains a universal table (a fully denormalized table) of a small TPC-H dataset. The schema has been modified slightly to make analysis in a spreadsheet application easier: The two computed measures revenue and cost as well as some levels of the time dimension of orderdate have been added in materialized form and some other attributes have been removed.

You may need to look up how to use pivot tables in your spreadsheet application.

- **Microsoft Excel:** [https://support.office.com/en-us/article/Create-a-PivotTable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99154456576](https://support.office.com/en-us/article/Create-a-PivotTable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99154456576)
- **Google Sheets:** [https://support.google.com/docs/answer/1272900](https://support.google.com/docs/answer/1272900)

Create the following pivot tables:

1. Show how much revenue suppliers from different regions (as columns) produced in every year (as rows).
2. Show how much revenue suppliers from nations of Africa produced in every year.
3. Show how much revenue suppliers from nations of Africa produced in every quarter of every year.
4. Show how much revenue suppliers from nations of Africa produced in every week of every month of Q1 in 1996.
5. Show how much revenue suppliers from nations of Africa produced in every year with “urgent” orders.

6. Show the average order quantity for parts from suppliers from nations in Africa per year.

7. Show how much revenue suppliers from nations of Africa (as rows) produced in every quarter of every year (as columns).

8. Discuss the terms “slice and dice”, “drill down”, “roll up”, and “pivoting”.

**Solution:**

**Slice and dice:** Like in cooking, where you may cut your vegetables into finer and finer pieces by first cutting them into slices and finally dices, you may want to look at more and more precise parts of your data. For example, we first analyzed revenue from all regions and then only considered the “slice” for Africa.

**Drill-down:** At the same time, we often want a finer granularity. While we only looked at revenue per region first, we “zoomed in” in the next step to see the revenue per nation.

**Roll-up:** Inversely, we may want to have summaries for that fine-grained view. When we look at the revenue on a week granularity, we may want to see the summary per month and quarter at the same time. The subtotals show in the pivot table “roll up” that information.

**Pivoting:** When we exchange rows and columns, we look at the cube from a different angle, i.e., we rotate (“pivot”) it.

### 3 OLAP Cubes and SQL

1. Write one or more SQL queries that compute all values of the pivot table from Question 2.3.

**Solution:** Using the SQL clauses we have seen in the course, we need one query for each level of aggregation. For the finest level, we can use the following query:

```sql
SELECT sn.nationname,
       SUM(olquantity * partretailprice * (1-oldiscount)),
       EXTRACT(YEAR FROM orderdate) AS orderyear,
       EXTRACT(QUARTER FROM orderdate) AS orderquarter
FROM orderline ol
JOIN orders o USING (orderid)
JOIN supplypart sp USING (partId, supplierId)
JOIN part p USING (partId)
JOIN supplier s USING (supplierId)
JOIN nation sn USING (nationId)
```
JOIN region sr USING (regionId)
WHERE sr.regionname = 'AFRICA'
GROUP BY sn.nationname, orderyear, orderquarter

For the sub-totals, we need another query each. We obtain them by changing the
GROUP BY clause to group by “sn.nationname, orderyear”, “sn.nationname”, and
“orderyear, orderquarter” and running the query without GROUP BY clause. We
would thus need to run six queries

Many database systems offer extensions for cube analysis to improve that. Using “GROUP
BY CUBE(sn.nationname, orderyear, orderquarter)” produces all subtotals men-
tioned above in a single query.

2. Write an MDX query that does the same.

This answer needs some work from our side. We’ll update this document after the holi-
days.

3. Suppose we run both types of queries on a large data set. What is the advantage of MDX
over SQL?

This answer needs some work from our side. We’ll update this document after the holi-
days.

References