CSE - 4255 Data Mining and Warehousing Lab

Data Warehousing

Saif Mahmud
Roll: SH - 54

M. Tanjid Hasan Tonmoy
Roll: SH - 09

Submitted To:

Dr. Chowdhury Farhan Ahmed
Professor

&

Abu Ahmed Ferdaus
Associate Professor

Department of Computer Science and Engineering
University of Dhaka

November 11, 2019

1 Problem Definition

The tasks for this assignment is described below:

- Creating a Relational Database
- Defining Schema (Star, Snowflake or Galaxy) for Data Warehouse
- Creating Dimension Table and Fact Table from Predefined Relational Database
- Creating Data Cuboid from Defined Schema
- OLAP operation on the data cube such as roll-up, drill-down etc.

2 Theory

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

3 Experiment Setup

3.1 Schema

We designed a star schema simulating warehousing on relational a retail database. The schema is illustrated in

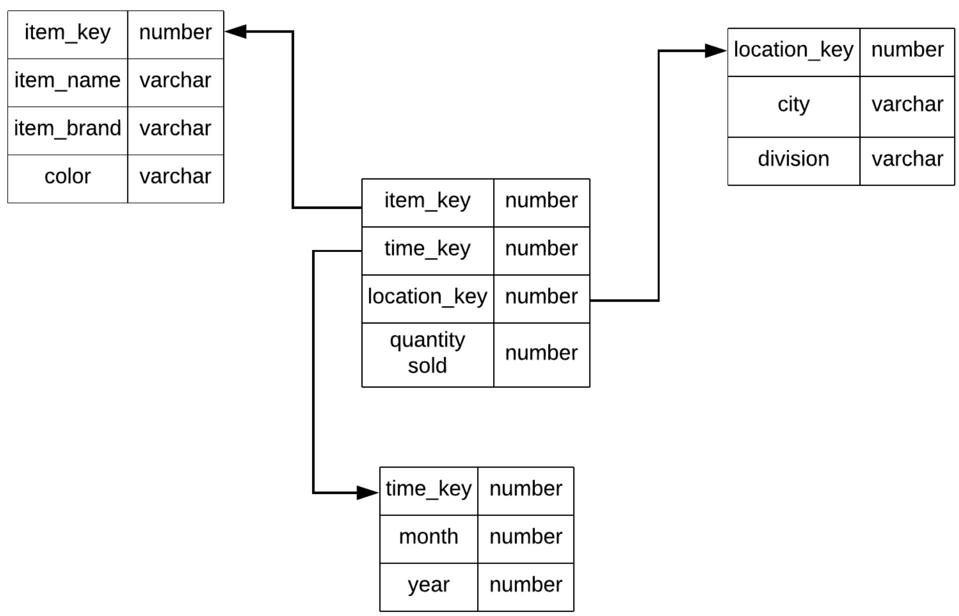

Figure 1: Star Schema used in the experiment

3.2 Operations

We have implemented Relational OLAP using oracle database. The operations are as follows

Roll up

```
SELECT year_number,item_name, SUM(amount_sale) AS sales_quantity
FROM   fact_table natural join item_table natural join location_table natural join time_table
GROUP BY rollup(year_number, item_name)
ORDER BY  year_number,item_name;
```

	YEAR_NUMBER	ITEM_NAME	SALES_QUANTITY
1	2013	pant	71
2	2013	shirt	43
3	2013	skirt	186
4	2013	tie	89
5	2013	(null)	389
6	2014	shirt	30
7	2014	tie	23
8	2014	(null)	53
9	2015	skirt	12
10	2015	(null)	12
11	(null)	(null)	454

Figure 2: Rollup grouping output

Cube

```
SELECT year_number,item_name, SUM(amount_sale) AS sales_quantity
FROM   fact_table natural join item_table natural join location_table natural join time_table
GROUP BY cube(year_number, item_name)
ORDER BY  year_number,item_name;
```

	YEAR_NUMBER	ITEM_NAME	SALES_QUANTITY
1	2013	pant	71
2	2013	shirt	43
3	2013	skirt	186
4	2013	tie	89
5	2013	(null)	389
6	2014	shirt	30
7	2014	tie	23
8	2014	(null)	53
9	2015	skirt	12
10	2015	(null)	12
11	(null)	pant	71
12	(null)	shirt	73
13	(null)	skirt	198
14	(null)	tie	112
15	(null)	(null)	454

Figure 3: Cube grouping output

Slice and Dice

```
SELECT year_number,item_name, SUM(amount_sale) AS sales_quantity
FROM   fact_table natural join item_table natural join location_table natural join time_table
where year_number = 2013
GROUP BY (year_number,item_name)
ORDER BY  year_number,item_name;
```

	YEAR_NUMBER	ITEM_NAME	SALES_QUANTITY
1	2013	pant	71
2	2013	shirt	43
3	2013	skirt	186
4	2013	tie	89

Figure 4: Slicing output

4 Conclusion

The separation of operational databases from data warehouses is based on the different structures, contents, and uses of the data in these two systems. Decision support requires historic data, whereas operational databases do not typically maintain historic data. In this context, the data in operational databases, though abundant, are usually far from complete for decision making. Decision support requires consolidation (e.g., aggregation and summarization) of data from heterogeneous sources, resulting in high-quality, clean, integrated data. In contrast, operational databases contain only detailed raw data, such as transactions, which need to be consolidated before analysis. Because the two systems provide quite different functionalities and require different kinds of data, it is presently necessary to maintain separate databases. Decision Data warehouse assists the decision making process by providing a historical perspective based on need. We have simulated such operations in a synthetic retail data warehouse.