Assignment_3_DataBase Management

Konstantinos Chronis (BAFT1502)

Assignment Supervisors:

Prof. D. Xatziantoniou

S.Safras

In this assignment we were asked to create a star/snowflake schema in a data warehouse from a dataset we have found online. The snowflake option was selected.

In computing, a **snowflake schema** is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. "Snowflaking" is a method of normalising the dimension tables in a star schema. When it is completely normalised along all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle. The principle behind snowflaking is normalisation of the dimension tables by removing low cardinality attributes and forming separate tables.

The dataset used for the purposes of this assignment is a Watson Analytics Sample Dataset of Sales Products consisting of 88476 rows of data. This was chosen because of the great dimension expansion capabilities. The dataset contains information regarding the revenue and the gross margin made by retailer shops and stores all for the countries collected in this dataset, the orders from the customers (including the quantities), the products and the year of the order placement and other key metrics.

A full description of the dataset is shown below:

Retailer type: Outdoor shop, Golf shop, Department Store, Sports Store, Eyewear Store, Warehouse Store

Country: Country of the shop's location

Order method type: Fax, Telephone, Email, Sales visit, Mail, Web

<u>Product:</u> Description of the product

Product type: Type of the product (i.e. Cooking Gear, Tools, Eyewear)

<u>Product line:</u> Camping Equipment, Personal Accessories, Mountaineering Equipment,

Outdoor Equipment, Golf Equipment

Year: Year of the order placement

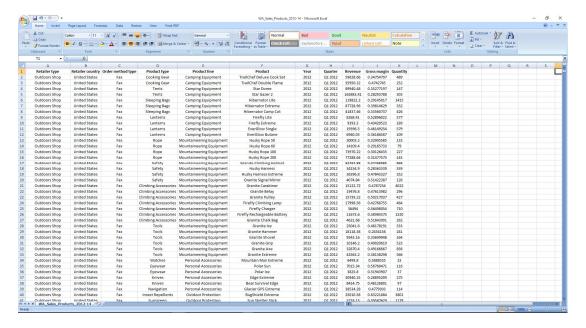
Quarter: Quarter of that year

Revenue: Total revenue of the shop/store

Quantity: Number of units ordered

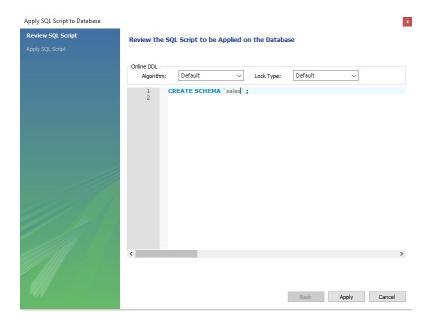
Gross margin: The Gross Margin of the shop/store

A screenshot of the first rows of the dataset.



Regarding the data cleaning usually required in similar cases, this particular dataset is in such a condition that does not require any cleaning.

First we connect to the MySQL Workbench and create the schema named sales.



Then we proceed to the creation of the tables-dimensions along with the fact table as shown below:

```
CREATE TABLE IF NOT EXISTS country_dimension (
country_ID INT NOT NULL AUTO_INCREMENT,
country NVARCHAR(45),
PRIMARY KEY (country_ID));
CREATE TABLE IF NOT EXISTS retailer_dimension (
retailer_ID INT NOT NULL AUTO_INCREMENT,
type NVARCHAR(45),
PRIMARY KEY(retailer_ID));
CREATE TABLE IF NOT EXISTS order_dimension (
order_ID INT NOT NULL AUTO_INCREMENT,
method_type NVARCHAR(45),
PRIMARY KEY(order_ID));
CREATE TABLE IF NOT EXISTS type_dimension (
type_ID INT NOT NULL AUTO_INCREMENT,
type NVARCHAR(45),
PRIMARY KEY (type_ID));
CREATE TABLE IF NOT EXISTS line_dimension (
line_ID INT NOT NULL AUTO_INCREMENT,
line NVARCHAR(45),
type_ID INT NOT NULL,
PRIMARY KEY (line_ID),
FOREIGN KEY(type_ID) REFERENCES type_dimension(type_ID));
```

```
CREATE TABLE IF NOT EXISTS product_dimension (
product_ID INT NOT NULL AUTO_INCREMENT,
description NVARCHAR(45),
line_ID INT NOT NULL,
PRIMARY KEY (product_ID),
FOREIGN KEY(line_ID) REFERENCES line_dimension(line_ID));
CREATE TABLE IF NOT EXISTS year_dimension (
year_ID INT NOT NULL AUTO_INCREMENT,
year INT,
PRIMARY KEY(year_ID));
CREATE TABLE IF NOT EXISTS quarter_dimension (
quarter_ID INT NOT NULL AUTO_INCREMENT,
quarter NVARCHAR(45),
year_ID INT NOT NULL,
PRIMARY KEY(quarter_ID),
FOREIGN KEY (year_ID) REFERENCES year_dimension(year_ID));
CREATE TABLE IF NOT EXISTS Sales_fact (
sales_ID INT NOT NULL AUTO_INCREMENT,
retailer_ID INT NOT NULL,
product_ID INT NOT NULL,
quarter_ID INT NOT NULL,
order_ID INT NOT NULL,
country_ID INT NOT NULL,
```

```
Revenue DECIMAL(10,2),
Quantity INT,
Gross_Market DECIMAL(10,2),
PRIMARY KEY(sales_ID),
FOREIGN KEY (retailer_ID) REFERENCES retailer_dimension(retailer_ID),
FOREIGN KEY (product_ID) REFERENCES product_dimension(product_ID),
FOREIGN KEY (quarter_ID) REFERENCES quarter_dimension(quarter_ID),
FOREIGN KEY (order_ID) REFERENCES order_dimension(order_ID),
FOREIGN KEY (country_ID) REFERENCES country_dimension(country_ID));
We will also need a temporary table to insert all the dataset and load from it to
insert the correct values in the dimensions
CREATE TABLE IF NOT EXISTS temp (
Retailer_Type NVARCHAR(45),
Retailer_Country NVARCHAR(45),
Order_method_type NVARCHAR(45),
Product NVARCHAR(45),
Product_Type NVARCHAR(45),
Product_line NVARCHAR(45),
Year NVARCHAR(45),
Quarter NVARCHAR(45),
Gross_Market DECIMAL(10,3),
Revenue DECIMAL(10,2),
Quantity INT);
The import of all the data into the temp table was made by :
load data local infile 'C:/Users/kos_c/Desktop/WA_Sales_Products.csv' into table temp
fields terminated by ';'
enclosed by ""
lines terminated by '\n'
```

Then the import of all the data to the tables-dimensions is implemented by the INSERT INTO command. An example is presented below:

INSERT INTO line_dimension (line, type_ID)

SELECT DISTINCT(temp.Product_line), type_dimension.type_ID

FROM temp, type_dimension

WHERE temp.Product_Type = type_dimension.type

The most challenging was the INSERT for the fact table sales fact as shown below:

INSERT INTO sales_fact (Revenue,Quantity,Gross_Market,retailer_ID,product_ID, quarter_ID,order_ID,country_ID)

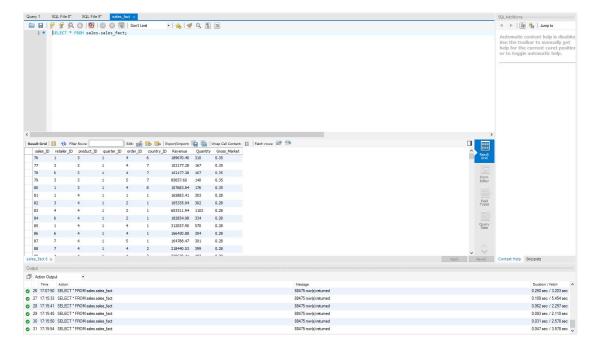
SELECT temp.Revenue,temp.Quantity,temp.Gross_Market, retailer_dimension.retailer_ID,product_dimension.product_ID,quarter_dimension.quarter _ID,order_dimension.order_ID,country_dimension.country_ID

FROM temp,retailer_dimension,product_dimension,quarter_dimension,order_dimension, country_dimension

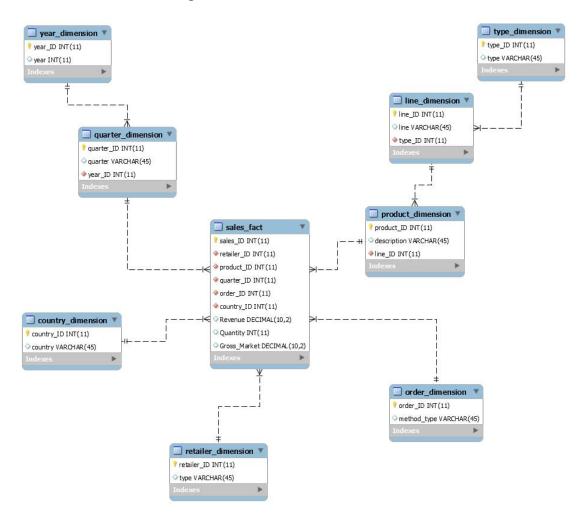
WHERE temp.Product = product_dimension.description AND temp.Retailer_Country = country dimension.country

AND temp.Order_method_type = order_dimension.method_type AND temp.Quarter = quarter_dimension.quarter AND temp.Retailer_Type = retailer_dimension.type

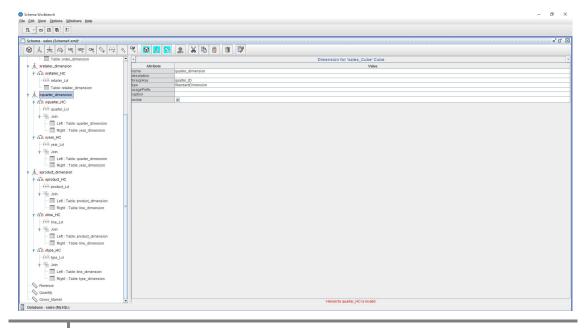
A screenshot of the process is of the filled fact table is presented below:



At this point an EeR diagram was designed to get a clearer view of the snowflake schema created in this assignment.

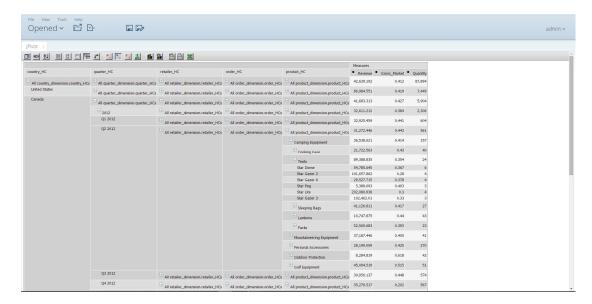


An attempt to create the .xml file in the Pentaho Schema Workbench was made as shown below:



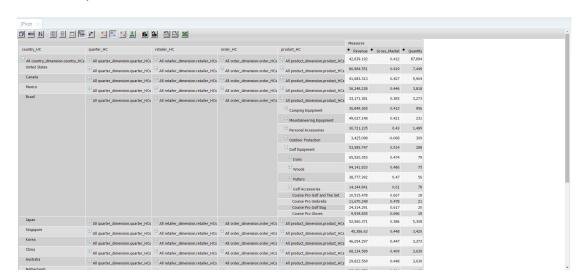
But the final .xml file was updated by hand in the Notepad following the instructions provided http://mondrian.pentaho.com/documentation/schema.php#Star_schemas. The complete working .xml file is included in the zip of the assignment. Regarding the measures of this analysis the Revenue and the Gross Margin were selected in their average, while the Quantity to be shown in its counted units.

Then in the Pentaho Business Analytics we create the JPivot reports as presented below:



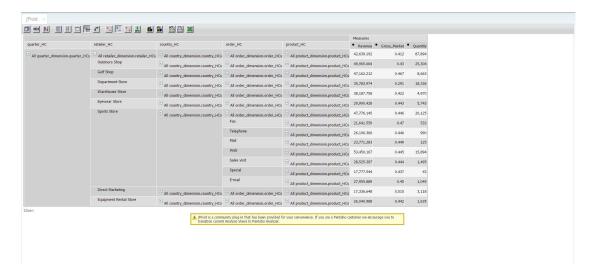
There can be many interpretations from the report above. An example could be that in Canada in the 2nd quarter of 2012 the product named as "Star Dome" provided an average revenue of 54,785.645 by all order methods in all retailers shops, while the "Star Gazer 6" an average of 29,527.715.

Another report is shown below:



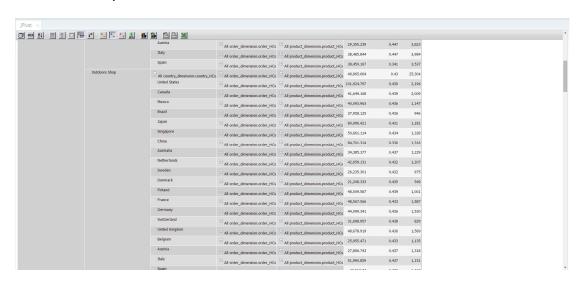
An interpretation of this could be that in Brazil the product "Course Pro Umbrella" provided an average revenue of 11,670.248 by all order methods in all retailer shops in all the quarters of a year provided in this dataset (Q1 2012 to Q3 2014).

Another report is following after changing the sequence of the columns:



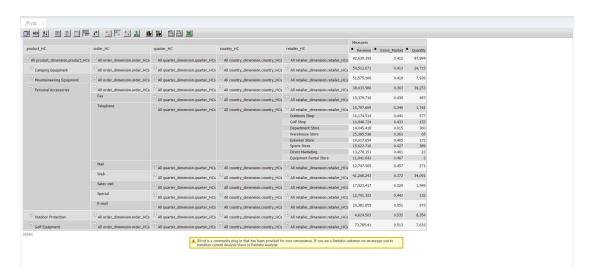
Here we notice that in Sports stores 533 units were ordered by fax, while 990 were by telephone. Both of these cases regard all products in all countries for all of the time quarters collected in this dataset.

Another one is presented below:



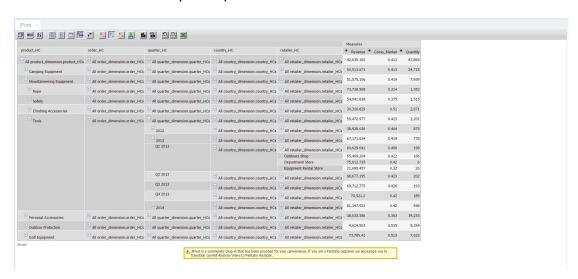
Here one could notice that, in the United States the Outdoor shops made an average revenue of 101,924.797 while the same type of shops made an average of 41,649.168 in Canada and 40,093.963 in Mexico.

Another report screenshot is shown below after changing the sequence of the columns:



Here an interpretation could be that 172 units in the "Personal Accessories" type of product were ordered by telephone in Eyewear stores in all quarters in all countries in this dataset, while in Department stores this number is 360.

Added to that another report is presented below:



Here it could be noticed that the product line "Tools" (falling in the product type Mountaineering Equipment) in the 3rd quarter of 2013,in all countries and by all order methods, provided an average revenue of 69,469.204 in Outdoor shops, while in Equipment Rental stores an average of 21,699.457.

Thank you for your time.