GLOBAL FOOD PRICES DATABASE

Data Warehouse Design

Submitted By:
Group – D
Names: Alvaro Rueda/Kamal Nandan/ Miguel Romero / Nina Gorbenko/ Pablo Dosal/ Rahul Singh/ Sanjay AJ

Global Food Prices Database

Group D – MBD O1

Dataset Description:

The Global Food Prices Database has data on food prices. Most common items in the data are the prices for products like beans, rice, fish, and sugar for 76 countries and some 1,500 markets. The data starts in 1992 for a few countries but most of the price trends start in 2000-2002. This dataset is updated every three months by the World Food Program(WFP).

Data Set analysis:

Analysis of the data set and a dictionary for every column:

There are 21 countries represented in our data set. 30% of the data belongs to the markets in Republic of Congo, 10% in Burkina Faso and Colombia, 7-8% to Armenia and Cameroon, the rest of the data represents the other countries.

Keep in mind that most of the data comes from underdeveloped and developing countries. There are 137 provinces belonging to 21 countries in the data set (in a few cases i.e. 795 cases, the province ID/Name is missing).

Facts and data complexity:

- There is a total of 220 markets (only 217 unique markets). To fix this disparity we used their respective ID and assigned the correct name for each of the markets.
- There is a total of 139 commodities with its respective pricing.
- There are 16 different currencies
- The price given is 82% of the times the retail price and 18% of the times wholesale
- The commodities are measured in 21 different units of measure (Kg, Gallons, grams, etc.)
- In terms of months, the price records are equally distributed (with 9% approx. record every month of the year.)
- In terms of years, the prices are from Jan 1992 to Jan 2017. Have in mind that 75% of the records are from the current decade (from the year Jan 2011 to Jan 2017). In this last decade, the prices are equally distributed over all the years.

The source of the data is genuine and comes from organizations such as the Food and Agricultural Organization(FAO) and the World Food Program (WFP) of United nations. It also has some inputs from the different ministries of agriculture of the different countries that take part in the study.

ADM0_ID	ID of the country
ADM0_Name	Name of the country
ADM1_ID	ID of the province
ADM1_Name	Name of the province or state corresponding to the above
Market_ID	ID for a particular market.

Market_Name	Name of the market
Commodity_ID	ID identifying the specific commodity
Commodity_Name	Name of the commodity
Currency_ID	ID for the currency in which trading took place
Currency_Name	Name of the currency
Price_Type_ID	ID for the price type: Retail or Wholesale
Price_Type_Name	Retail or Wholesale
UM_ID	ID for the unit of measurement
UM_Name	Name of the unit of measurement
MP_Month	Month in which this price was recorded
MP_Year	Year in which the price was recorded
MP_Price	Market Price for the commodity in this month
MP_Commodity_Source	Source from which this data was received

It has been observed that there is a great possibility that this data represents the prices at which the humanitarian organizations procured the food items in different regions in order to alleviate certain food-related crises.

It can be affirmed that this data doesn't belong to the commodity exchanges because of a no. of reasons such as source of the price data, availability of only monthly prices etc. (Trading data has hourly prices and a different structure.)

Data Quality Issues:

- **Missing Values:** In 795 cases records, the id/name of the province to which the commodity market price belongs to is missing. However, we don't need to remove these records, because country id and name are present.
- **Errors:** As discussed in the previous section, the market IDs dissonance is relevant. However, the lack of market name doesn't significantly affect the user of the data because the price would be the same for the same province or country. As such we don't really need to drop/discard such rows.

Other than the above-mentioned fields, data is in healthy shape and all other variables don't have any missing value.

User requirements:

Different organizations such as FAO/FWP/Ministries of agriculture/Ministry of Human welfare/NGOs such as the Bill & Melinda Gates foundation may be interested in this data. These kinds of organizations can analyze the past prices of different food items across different geographies and draw conclusions and create action plans to achieve their goals. Procuring the food items at the cheapest price possible in future can be achieved analyzing this data and controlling, or at least understanding how the food chain works in these countries analyzed.

Analysis of different modelling schemas:

Star schema:

Pros:

- Easy to understand and visualize for the business users
- Data can be analyzed and visualized easily from different dimensions and perspectives
- Very fast query performance, because minimal joining takes place; it's a flat representation of data and many values often get repeated

Cons:

- Redundancy of data – no 3NF
- Because of redundancy of data, it may be prone to errors while making changes.
- If changes are to be made, the same changes have to be made everywhere.
- More storage requirements.

Snowflake schema:

Pros:

- Generally 3NF and hence less redundancy and hence less storage requirements
- Easy to maintain
- More flexibility

Cons:

- May be difficult to visualize for business users from business perspective
- Less efficient in terms of query performance because lots of joining may take place

Data Vault:

Pros:

- Closer to the business organization
- Easier to maintain
- More flexible – easier to create hubs/satellites/links when a new business unit is added or removed
- Less redundancy

Cons:

- Difficult to visualize from analytical point of view, for the business users
- Query performance is not good because a lot of joins may take place

Data Warehouse Approach Selection:

We have followed a multi-dimensional and hybrid approach: Star and Snowflake mixed. We chose this because we see that the data that belongs to the food procurement department of organizations such as

UN-FPO, may need to be analyzed taking into account the different procurement methods (wholesale/retail), time requirements (month/year), place(country/market) or even the different currencies.

Considering the end user, a global non-profit that might have wide-reaching needs to integrate several sources in their databases, we have determined that the ETL process must be periodic, and the reporting needs to be efficient. The users of such kind of databases are not generally very tech savvy, and this furthers the need for an easy schema with a very good query performance. It is because of this that we define the key goal as facilitating complex reporting with the main purpose of analyzing and interpreting historical data, on a robust system that can process tens of thousands of rows of data.

Thanks to our hybrid model (that consists primarily of a Star based model with a fact table and various dimension tables) the analysis can be done easily. We call it hybrid because the Star schema is briefly supported by a snowflake model.

The fact table would contain the prices of different food commodities in the selected countries in different currencies. There would be different dimensions of the food prices which may be markets, countries, provinces, commodities, retail/wholesale price types etc. Some of the dimension tables may further be supported by other tables. This is the main reason to pursue a hybrid of star and snowflake approach for the whole data warehouse.

Data Warehouse Design:

Table Name	Table type	Fields
F_CommodityMarketPrices	Fact table	• foreign keys of dimension tables. • "market price" of commodities • "source of the market price"
d_market	Dimension table	• id • Market Name • Province • Country
d_currency	Dimension table	• Id • Name • Country
d_Date	Dimension table	• Id • Month • Year
d_PriceType	Dimension table	• Id • Name
d_Unit	Dimension table	• Id

		• Name
d_Commodity	Dimension Table	• Id • Category id • Name
s_Category	Table containing Category details	• Id • Name • Type

The MWB file is attached. However, please see below the graphical representation of it.