University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Project 2017

Product Prices from Crowdsourcing Data Mart

This document contains the requirements as discussed in class. Please also refer to the JAD slides.

Instructions

1. Complete this project in a group of two (2) to three (3) students.
2. Demonstrate the project on **April 4th, 2017** in a 15 minute timeslot, as allocated by the TA.
3. Use a database management system (DBMS) such as PostgreSQL, MySQL or Microsoft SQL Server to complete this project. As explained in class, you may use a Dashboard Template or you could also create your own frontend for the OLAP analytics. Similarly, the data mining component may be implemented using the R or Python languages, or by using a system such RapidMiner or WEKA.

Deliverables:

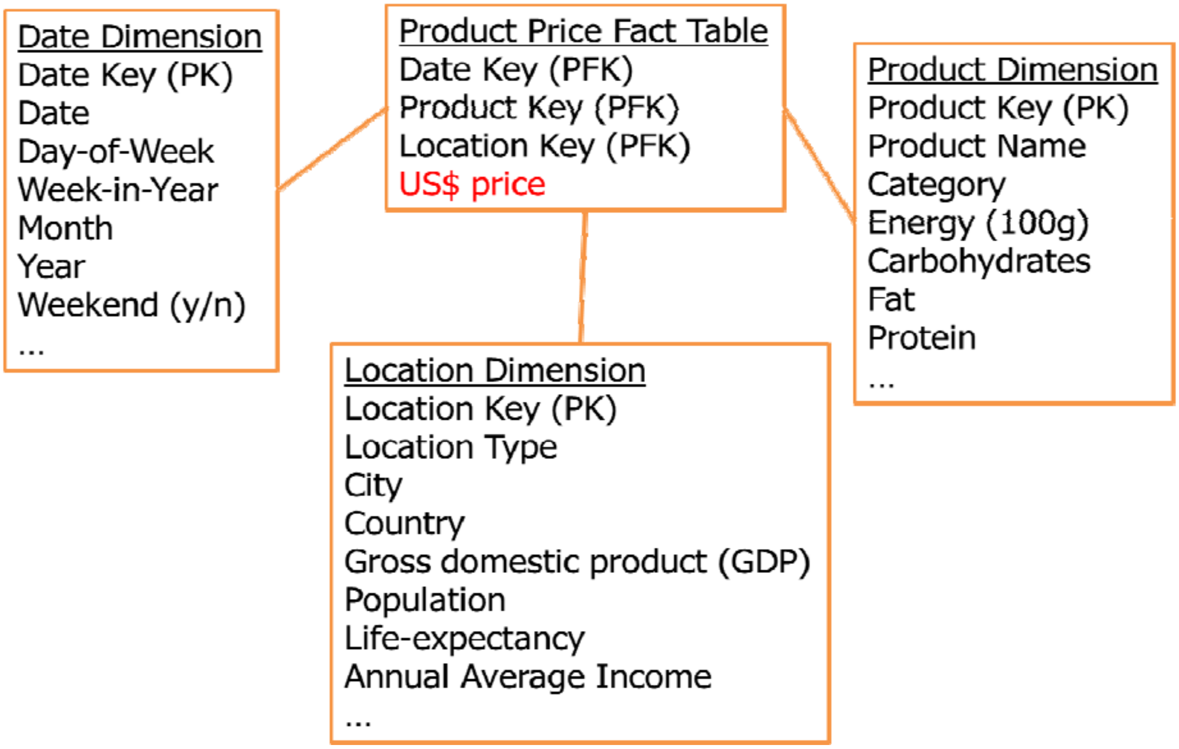
Submit all your source code, together with a one-page high level data staging plan through BlackBoard Learn**, before April 4th, 2017 at 08h00**. (That is, before the demonstrations will start.)
Note that all group members should submit the source code, not just one per group.
All group members should attend the project demonstrations.

Your task:

Your task is to design and implement a Crowd-sourced Product Prices Data Mart using the data as obtained from the World Bank website: http://data.worldbank.org/data-catalog/crowd-sourced-price-collection. Your project will consist of i) data staging, ii) the building of OLAP queries and the design of an OLAP end user interface, and iii) some data mining.
Remember to also read the paper that is referenced, since it provides useful background.

1. Hamadeh, N., Rissanen, M. and Yamanaka, M. (2013). "The World Bank Pilot Study for Crowd-Sourced Data Collection through Mobile Phones."

Here is the dimensional model of the proposed data mart. You may use this model "as is", or modify and extend it as you see fit.



Other Notes:

1. Remember to create your own surrogate keys. Refer to the slides and/or the book by Kimball et. al. that explain how to stage the data for surrogate key lookup.
2. Recall from the JAD session that you are encouraged to supplement the original data with enriched data from other sources, such as details about the countries, cities, or product nutrition information.
3. The Date dimension is crucial in order to perform interesting analytics. You are encourages to implement at least all the attributes I listed above.
4. Refer to the "Typical Analytic Cycle" as described in class, for a list of typical analytic questions which should be answered when accessing your data mart.

Requirements and Mark Allocation

Note that the project is out of **100**. You are required to complete questions 1 to 4 as well as at least one of questions 5 and 6.
You may thus obtain a mark of 120/100 if you complete all the questions.

1. **(10 marks) Physical Design:** Create the physical schema of the data mart using the DBMS of your choice.

2. **(20 marks) Data staging**: Extract and transform the data and load all rows into the data mart. Be sure to record all the steps that you took in order to accomplish this task. (Submit a one-page high level schematic with your source code.)

3. **(30 marks) OLAP queries:** Here is a list of potential queries. Note that this list is not complete and that I encourage you to also explore the data in other ways. In these queries, we make use of concept hierarchies to explore the data**.**
   a. Explore the data in order to get "a feel of" the prices of the various products in a country. The user should be able to drill down from 6 months, to one month, to a specific day, and roll up again.
   b. Explore the data in order to get "a feel of" the price differences of the products when considering more than one country. For example, one may want to contrast the price of tuna steaks in Kenya with that in India. The user should be able to drill down from 6 months, to one month, to a specific day, and roll up again.
   c. Explore the data by considering the prices of categories of products. That is, we wish to roll up from product to category. For example, the sales of apples, bananas and oranges are grouped into fruits while minced beef and chicken legs are grouped into fresh meat.
   d. Explore the data by considering the prices of categories of products, on a specific day of the week (e.g. the prices of fruits on Monday versus Saturday; weekend versus weekday, and so on).
   e. Explore the fluctuations in individual product prices, per country, per city and per location.
   f. Explore the prices of a specific product (e.g. apples) in terms of socio-economic factors, such as the average income of a country.
   g. Compare the prices of two complementary products (e.g. white rice and long-grain rice).
   h. Compare the prices of two complementary products (e.g. white rice and long-grain rice), within a specific country. Next, drill down by city and location.

4. **(20 marks) Business Intelligence Dashboard.** Create an interface which will give a knowledge worker the ability to easily explore the data mart. (This should include graphs.)

Complete at least one of the following two questions.

5. **(20 marks) Association analysis.** Apply an association analysis algorithm such as the Apriori technique to explore the data from Kenya, in order to determine which products are frequently priced together on the same day in a particular location.
   (and/or)
6. **(20 marks) Classification.** Build a decision tree in order to explore the data. For example, you may want to contrast the pricing patterns in two or more countries. Alternatively, you may want build a model to study the characteristics of products within categories.