Data Warehouse Modelling Workgroup

Advertisement Data Set

Group B:

Cort Pons, Eduardo
Hung, Meng-Chen
Lee, Esther Chaelin
Mouawad, Ignacio
Raja, Varun
Satyal, Anup Raj
Villaveces, Juliana

Index

1. Data Set Analysis

Our group was assigned the “advertisement” dataset, which primarily contains information on (marketing) *posts* made on four social media platforms. Four tables/sheets (FACEBOOK, YOUTUBE, TWITTER and INSTAGRAM) each consist of a list of 373,761 posts made between 01/12/2017 to 31/12/2017. Another table (PRODUCTS) consists of information on a number of products (e.g Sneakers), along with their categorization under a certain “family” (e.g. Sports). A sixth table (ZONES) contains information on location divided into Zones, Cities and Zip Codes.

The data available in the platform tables broadly characterize the post itself and also the posts’ targeted recipient(s). For the posts, the dataset contains the date, the product being marketed through each of the posts and the subsequent prints and hits (defined below) generated. For the targeted recipients, the dataset contains the age range, gender, location and likes (defined below) for each of the post.

Given the structure of the dataset and the provided information, we hypothesize that it can be used to analyze the effectiveness of a social media marketing campaign. With this goal in mind, the following sections i) analyze the dataset in further detail, ii) identify data key quality issues, and ii) propose a relational database dimensional model (star-schema) that may optimize data storage, organization and manipulation.

1.1. Detailed Dataset Analysis:

The PRODUCTS table/sheet has two columns. The first column contains a list of product names, and the second column contains the product categories (family) corresponding to each product:

a. The list “Product” consists of Mobile Phone, Dress, Watch, “Sheatshirt”, Scarf, Sneakers, Theater, and Trip.
b. The list “family” is made up of Electronics, Wear, Accessory, Sports, and Culture.

We noted that i) Dress and Sheatshirt are both assigned to the Wear category and both Theatre and Trip are assigned to the Culture category (i.e one to many relationship between family and products) and ii) the eventual model should accommodate the fact that the list of products could also be expanded in the future.

The ZONES table/sheet has 3 columns: Zone, City, and Zip Code:

a. Zone is divided into four categories - North, Centre, Coast, and South.
b. Each city belongs to a zone. There are three cities listed for each zone for a list of 12 unique cities .
c. A list of unique zip codes identify specific parts of each city (e.g 28014 and 28019 parts of Madrid).

We noted that the zip codes only pertain to certain parts of each city and that they could be expanded to include new ones in the future.

The remaining tables are divided into Instagram, Facebook, Youtube, and Twitter, where the products are being promoted through posts. The fields of these tables are identical and defined as follows:

a. Date: the day, month and year a post was made on the platforms
b. Zip Code: the section of the city where each of the post was promoted
c. Product: the item being promoted (and/or sold) through each post
d. Age: a single assignment of an age range of either 18-30, 31-45, 46-60, or 61-99 for each of the post
e. Gender: a single assignment of male or female for each of the posts
f. Prints: the number of times each post is published by a respective platform. The maximum and minimum value for prints is 199 and 0 respectively.
g. Hits: the number of times each print leads to a page visit. As we understand the concept of hits, if a single page on a website has 10 pictures, every time a person visits that page, it counts for 11 hits (10 for the pictures and 1 for the HTML file of the actual page). For our purposes, however, since each post is being used to market (or sell) one product, we are going to assume that a hit refers to a single click on the post. The maximum value for this field on each of the platform tables is 69, with a minimum of 0.
h. Likes: tags (Sports, Travel, Technology, People, Garden, Fashion, Business, Cars) automatically ascribed to each post by the platforms, based on the interest and preferences of the targets. In essence, we understand that the platform assigns these tags to each post (i.e each post could have more than one tag) so that the tags can be matched with target users with certain interests such that each posts generates more hits (also known as targeting). For example, in our given dataset, each post promoting (or selling) Mobile Phones on Instagram which has been ascribed a tag of “Garden”, on average, generates 0.06 hits for each print while those ascribed with a tag of “Technology” generates, on average, 0.25 hits for each print.

Apart from the field specific points noted above, we found that each of the platforms has been allocated 46,720 posts per year. All the posts are equally allocated by gender (1984 each), product category (3968 per product) and platform (46,720/platform). A closer examination reveals that by dividing the number of days per month (for example, the product i.e 3968/31), each product category shows 128 posts per year, equally divided between the genders. From an organisational standpoint, the resources have not been allocated in the most efficient way. The data model we suggest in this report is designed to allow the end user to slice and dice the data such that issues such as these are identified quickly by the end users.

1.2. Potential data quality issues:

Irrelevant values: We noticed that 2,754, 2758, 2796 and 2758 records (posts) for Facebook, Instagram, Youtube and Twitter have 0 prints and 0 hits. These records are irrelevant as they do not provide any useful information when analyzing marketing effectiveness. These records will be filtered out of the dataset in the ETL process.

Misspellings: On all the tables, “Sweatshirt” is spelled as “Sheatshirt”. While this may not cause any issues with regards to data matching, as even the Products table has the item misspelled, this could potentially cause issues with regards to querying, simply because the word “Sheatshirt” does not exist in the English language. The spelling will be corrected in the ETL process.

Duplicates: Some posts have exactly the same prints and hits for the same products for two different dates. Without any further information, we are unable to determine whether these duplicates are just a coincidence or whether the dates are incorrectly inputted. Therefore, in this particular case, we have decided to keep such records in our final database.

2. Data Warehouse Approach Selection

Initial Approach: Data Vault

Our analysis identified numerous many-to-many relationship in our dataset. For example, a single product item can be targeted to many demographic groups and vice-versa. Thus, we initially selected a Data Vault approach, conceptualizing three business processes as Hubs, each with its Satellites as follows:

a. A Target hub with the target audience for the posts. The Satellites for this hub contained demographic information. However, while the gender or age of a target does not change over time, his/her/their location may, making Location a Slowly Changing Dimension (SCD). Therefore, in order to enable maximum flexibility, age and gender were classified under one satellite and location under another.
b. A Platform hub with information relating to the platform being used to market to the target. Three Satellites were created for the Platform hub, chosen based on the source, type of information and to reduce redundancy. Each post is associated with a post date and has a unique number of prints and hits generated, classified under a Post Satellite. By giving the Likes and Platform Names, both associated to the platform, their own separate Satellites, we reduce the redundancy of records that needed to be stored if we had only one “Post” Satellite, with each of the 300k (times 4) posts storing both Likes and the source Platform instead of Platform and Likes being stored in a separate database with only their unique names and being linked to posts.
c. A Product hub with the characteristics of the product being marketed. In the current dataset, the only Satellite defining a product in is its name and family. While family could be separated in a different Satellite, given the limited number of product records, we decided not to split them.

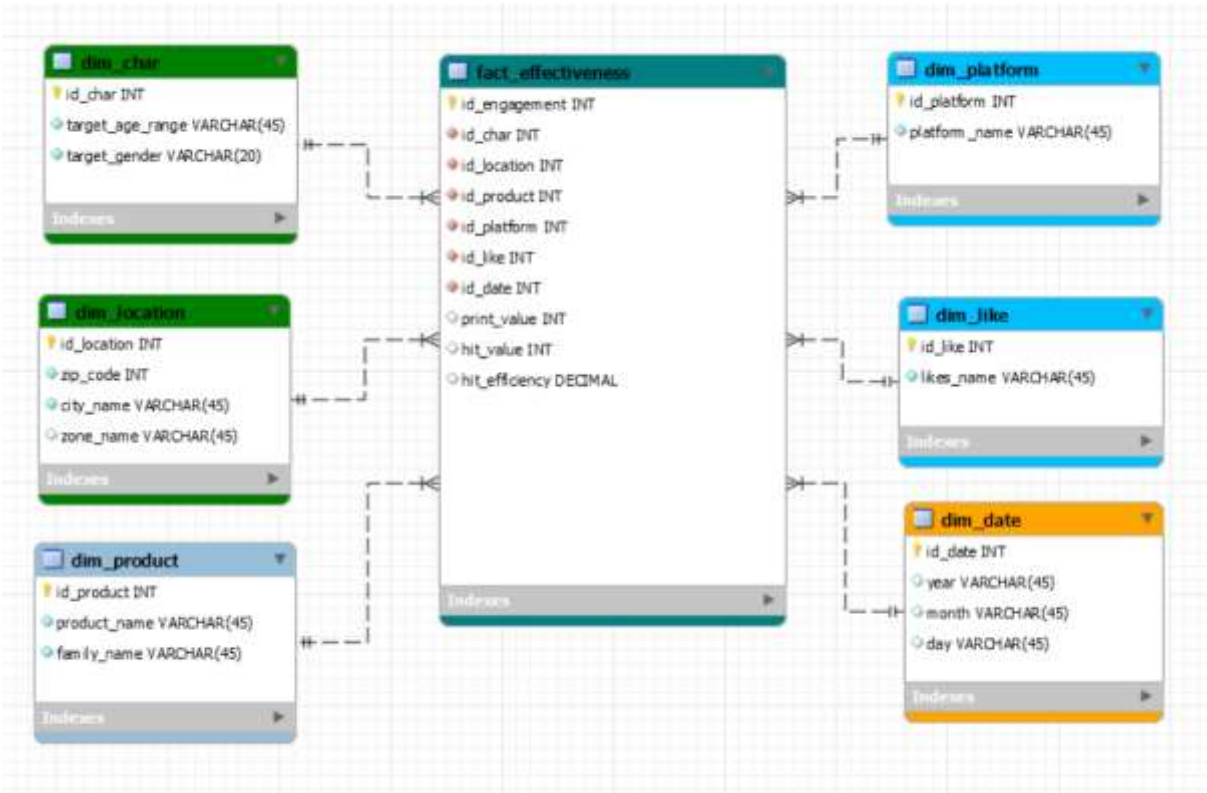
Since all three Hubs also have relationships with each other, a single Link was sufficient. Based on these definitions, we created a data vault model (Appendix Exhibit 2)

Final Approach: Star Schema

Once we had created the Data Vault, as mentioned above, we realized that the business entities (processes) were interrelated and the model contained only a single link. Thus, if we simply removed the link and replaced it with a fact table, the model would be equivalent to a dimensional Star Schema.

The key facts (measures) would be prints and hits, which would make our fact table a Transactional one, defined as Fact tables that contain measurement of an event. For our purposes, the event is a post made in at a specific date on a specific platform. An additional fact table element that has been added is the hit_efficiency, which is defined as the rate of hits per print for each of the posts. While the calculations will only be added as a part of the ETL process, the field has been created in the fact table during the modeling process.

The key dimensions (or contextual elements) that can be used to analyze the prints and hits are the same as the Satellites we had previously defined in the Data Vault approach. The model hence created is as follows:



The reasons the dimensions and attributes are defined and assigned as they appear on the model are as follows:

a. Dim_char: defined with the same logic behind Satellite_char in the Data Vault. The attributes in this dimension are unlikely to change and usable across data marts, which makes this a Conformed Dimension. The age range is defined as a VARCHAR

and not Numeric (or others) as for our purposes, age is a categorical variable. The VARCHAR for gender has been reduced from 45 to 20 bytes as the attribute is unlikely to go beyond 6 characters, optimizing storage.

b. Dim_location: defined with the same logic behind Satellite_loc in the Data Vault.
c. Dim_product: defined with the same logic behind Satellite_product in the Data Vault.
d. Dim_date: defined with the same logic behind Satellite_date in the Data Vault. The attributes in this dimension are also usable across data marts, which makes this a Conformed Dimension.
e. Dim_like: defined with the same logic behind Satellite_likes in the Data Vault.
f. Dim_platform: defined with the same logic behind Satellite_platform_names in the Data Vault.

Our selection of the Star Schema with the facts, dimensions and attributes defined above should be able to optimize the storage and manipulation of the given dataset. The model is designed to eliminate redundancies and other data quality issues in order to generate insights relating to social media marketing effectiveness.

Bibliography

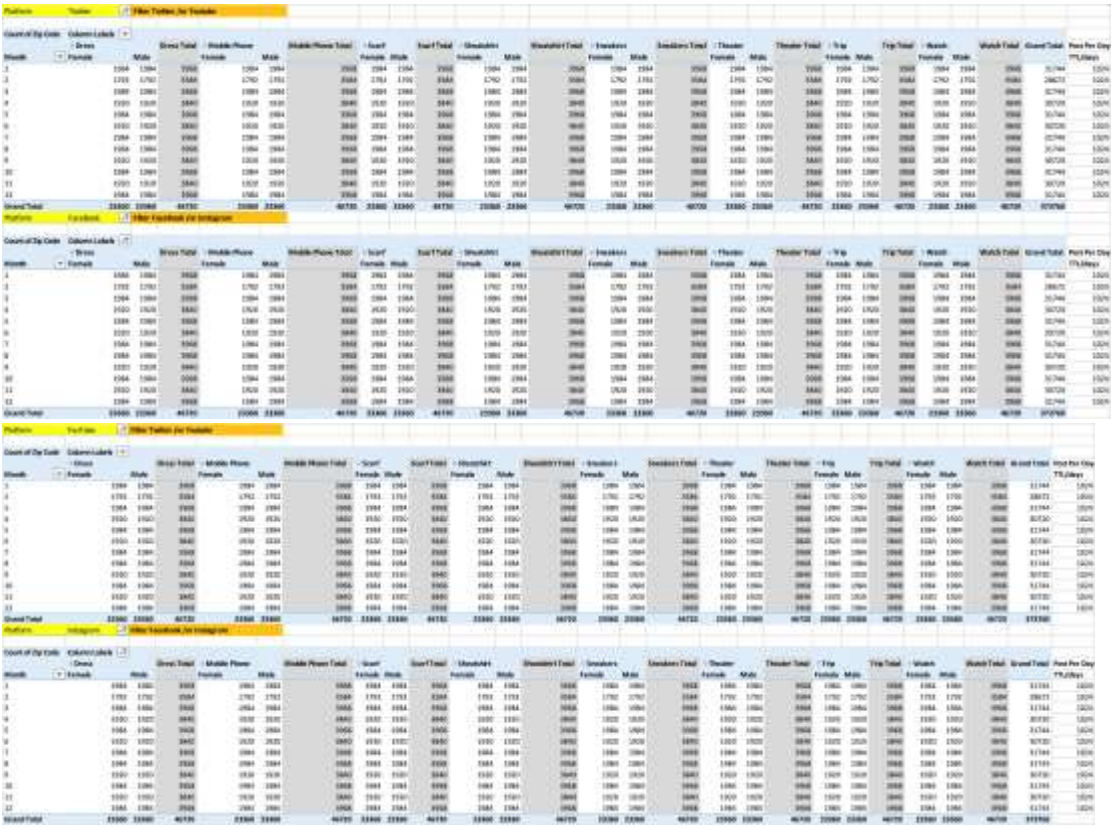
Hub Digital Marketing. (2016, January 20). Retrieved from What Exactly Does "Hit" Means?: https://www.hubdigitalmarketing.com/blog/2016/1/17/what-exactly-does-hit-mean

Kimball Group . (n.d.). Retrieved from Dimensional Modeling Techniques: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/

Marketing Terms / The Digital Marketing Reference. (n.d.). Retrieved from Hit: https://www.marketingterms.com/dictionary/hit/

Appendix

Exibit 1: Dataset analysis by gender



Exibit 2: Data Vault Model