**Practical Data Science: Promotion Effectiveness and Promotion Planning— Part 2**

6 Step Process | Step 1 — Build Data Foundation



Photo by [Nathalia Rosa](https://unsplash.com/photos/rWMIbqmOxrY)on [Unsplash](https://unsplash.com/photos/rWMIbqmOxrY)

Welcome back to a series of articles on promotion effectiveness measurement and promotion planning! In my last post, I talked about [business problems, key concepts, and the importance of this use case](https://towardsdatascience.com/practical-data-science-promotion-effectiveness-and-promotion-planning-part-1-52f6707a59b9). Today, I will describe (1) a **6-step process** to build a promotion effectiveness measurement and planning solution and (2) the **Step 1** of the overall process: **Building data foundation**.

**Promotion Planning Solution Building: 6 Steps**

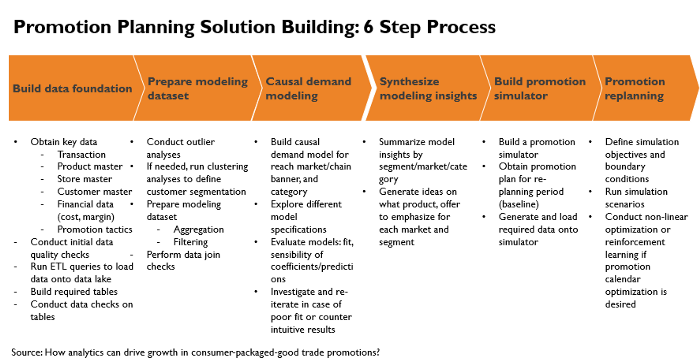


Figure 1: Image by [Minha Hwang](http://medium.com/@minha.hwang)

The 6 steps of promotion planning solution building are as follows:

(1) Build data foundation

(2) Prepare modeling dataset

(3) Causal demand modeling: The main outputs from the causal demand model are price elasticity, promotion tactic lift, and non-price promotion lift

(4) Synthesize modeling insights

(5) Build a promotion simulator

(6) Promotion (re)planning

Key activities for each step are summarized above in Figure 1. I will describe details of each step in this series of articles. In this article (part 2), I will focus on “Build data foundation.”

**Step 1: Build Data Foundation**

This step is foundational and causes most mistakes and problems. Recently, the importance of data foundation is being recognized and emphasized more, as shown in the “[data-centric AI](https://landing.ai/data-centric-ai/)” by Andrew. Ng. In terms of required efforts, this step takes 2-to-6-month efforts of 1 or 2 full-time data engineers, who are guided by more senior data architect (part-time) and data scientists who will work on causal demand modeling. Please note that you have to plan enough buffers, especially when this step was not done in the past. In most cases, a central data lake with all the required data for promotion effectiveness modeling is likely not be built yet, thus the data engineers and data scientists who are in charge need to reach out to multiple data owners in multiple internal departments and external syndicated data providers and data vendors (if available).

The data foundation tends to be a use-case specific. Thus, this is usually considered as business domain knowledge that will be picked up from on-the-job training. Thus, this is least taught in academic programs except for technical SQL training. Making it even worse, many companies do not maintain high quality data schema and data dictionaries for many internal databases. I have noticed that many junior data scientists and data engineers struggle most with this step since a clear guideline and practical tips are rarely provided.

Raw data tend to be company and situation specific. Thus, there is no cookie-cutter solution that you can apply. That said, once you clearly understand what data are essential vs. nice-to-have, you can avoid being overwhelmed by the number of databases, tables, and columns that you need to go through. I will cover 2 most common cases: (1) retailers (e.g., supermarkets such as Albertsons, Safeway, Kroger, department stores such as Macy’s, drug store such as CVS and Walgreens, convenience stores such as 7 Eleven), and (2) manufacturers (e.g., Kellogg and Proctor & Gamble). For the case of manufacturers, I will focus more on industry sectors where syndicated data providers such as IRI and Nielsen are available: i.e., frequently purchased consumer packaged goods (i.e., CPG companies.) Please note that data foundation in the central data lake are usually maintained at the lower aggregation level (e.g., stores instead of promotion zones) than the final modeling grain level.

**Build Data Foundation: Retailer Case**

**(1) Central promotion effectiveness data repository**

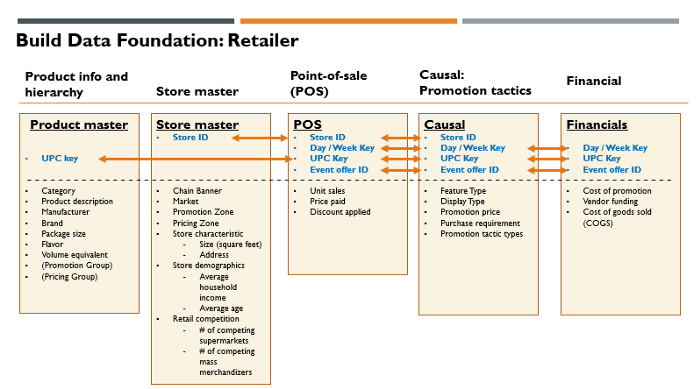


Figure 2: Image by [Minha Hwang](http://medium.com/@minha.hwang)

Figure 2 shows the ERD (entity relationship diagram) of the central data repository that you need to build to have a 360-degree view on promotion effectiveness. Building this central data foundation is considered as a big win, since the required datasets are usually scattered across multiple departments and IT systems. Before I discuss each data domain in details, let me provide a few tips first.

**1–1. Primary keys**: For an effective central data repository, it is important to understand what primary keys are and which will be used to join data across tables. In this case, there are 4 primary keys: UPC key (i.e., product key based on UPC code, store ID (i.e., unique key for each store), day or week key (i.e., unique number for each week or day), and event offer ID (i.e., unique ID for each promotion event). One way to think about this is to imagine a 3-dimensional data cube (product | store | time), with additional key to attach promotion events and related information to this 3-dimensional data cubes. Please note that I have highlighted primary keys in blue in the ERD shown in Figure 2.

**1–2**. **Aggregation level and aggregation bias**: For a central data repository, it is recommended to keep data aggregation level as low as possible. It is always possible to aggregate at the later stage (e.g., feature engineering for modeling dataset). However, if you already aggregate the data to too high level, it is not possible to disaggregate later. Moreover, improper aggregation of dataset can cause biases in causal effect estimation. One well-known example of **“**[**aggregation bias**](https://link.springer.com/article/10.1007/s11129-006-9011-3)**”** is when you aggregate data across stores with heterogeneous promotion activities. Another less-known example is **“split week” bias**, when you apply a common week definition to aggregate daily data to weeks. Because promotion calendars can have different promotion week definition across local markets (e.g., Sun — Sat vs. Fri—Thu), this will result data from non-promotion week and promotion week are blended in a common week definition for certain markets. If week is commonly defined as Sun — Sat for entire data and there are a few local markets which follow Fri—Thu promotion calendar, this blending will happen for all the stores in the local markets which follow Fri — Thu promotion calendar. It is a bit surprising that this happens quite frequently in practice (and even in the case of data from syndicated data providers.) I will give more details on this later in the modeling stage.

**1–3**. **Data duration**: At least 2–3 years of data are required to properly reflect seasonality and trends. Covid lock downs in Year 2020 and 2021 distorted normal seasonality and trends, thus extra care will be needed to handle data from these periods.

**1–4**. **Executed vs. Planned**: Oftentimes, there can be internal promotion planning database, which have promotion plans. One can be tempted to use this database as a source of “causal” table. However, it is very common to change plans at the last minute. Because of this, executed promotion prices and tactics can be very different from the original plans. Thus, it is highly recommended to use “executed” promotion prices and tactics instead of “planned” promotion prices and tactics.

**1–5**. **Promotion with multiple price points**: Beyond simple promotions with a single price point (e.g., 10% discount), there are complex promotions with multiple price points with certain requirements (usually in the form of multiple purchase requirements to qualify for quantity discounts.) With the complex promotion with purchase requirements, consumers can choose to either (1) buy a small quantity and pay regular prices or (2) buy large quantities and pay promotional prices. In the example shown below in Figure 3, if consumers decide to buy more than (or equal to) 4, they only need to pay $3.00 per each. However, if consumers decide to buy less than 4, they either need to pay the regular price of $4.99 per each (if they do not have a loyalty card) or pay the fallback promo price of $3.49 per each (if they have a loyalty card.) You can see that there are 3 possible price points in this example. The net price aggregated at store level for a given week for this product would be weighted average from these 3 price points.

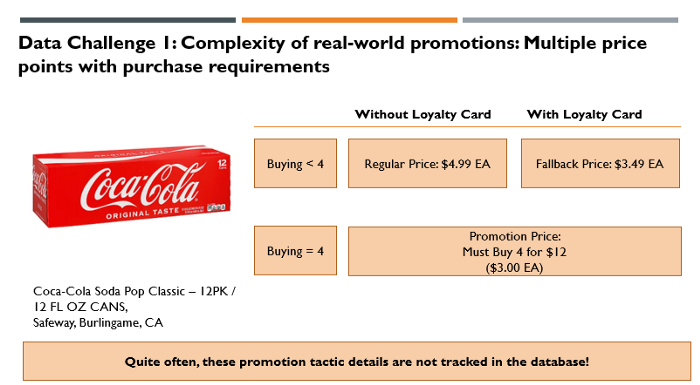


Figure 3: Image by [Minha Hwang](http://medium.com/@minha.hwang)

**1–6**. **Basket-wide or category-level promotion**: There are basket-wide or category-wide promotion events, which make it hard to calculate discount amounts per item level. You may have to decide on certain promotion discount allocation rules in these cases. As an example, Safeway could have run “Mix & Match” promotion in the wine category, which gives additional 10% discount if you buy more than 6 bottles. In this case, in addition to the product-level promotion discounts, there were additional 10% discounts that consumers enjoyed if they decided to buy more than 6 bottles.

**1–7**. **Personalized promotion**: In the past coupons were usually distributed equally to all consumers in the past by distributing them to the mailboxes of consumers. However, personalized and more targeted coupon promotions are adopted by more sophisticated retailers. The key feature of this promotion is that only a subset of consumers is offered with the promotion discounts. Other consumers who did not receive personalized coupons still pay regular prices or product-specific discount prices before the personalized coupon promotion. The existence of multiple effective prices across consumers makes it more complicated for us to decide on what effective promotion prices (at aggregate level such as stores or promotion zones) we should use for promotion effectiveness modeling in these cases.

**(2) Key Tool: Source to Target Mapping**

Data engineers and data scientists who work on this step have to go through multiple IT systems and discuss with multiple stakeholders. It is very easy to lose track of all the relevant facts, discussions, nuances, and watch outs. A tool, called as “Source to Target Mapping” can help greatly to keep track of key information and bring transparency on overall progress in this messy phase. This can be done, even as an excel file with multiple tabs for key types of data. If you want to be more systematic, you can make online version for easy sharing and update. Figure 4 shows key columns that you need in this tool with revenue as a sample data column.

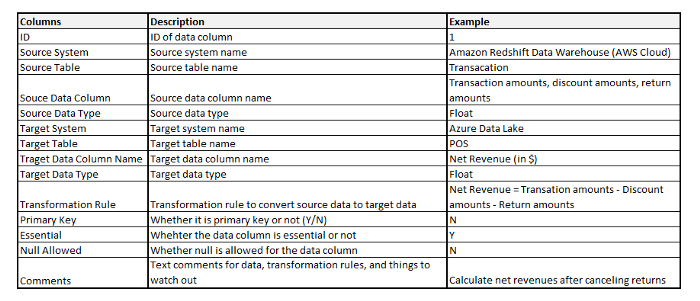


Figure 4: Image by [Minha Hwang](http://medium.com/@minha.hwang)

This tool makes it easier for data engineers and data scientists to understand how source data system / table / data columns are mapped to target data system / table / data columns. In addition, specifying transformation rule bring clarity on how source data columns are used to calculate target data columns. You can either enter (1) mathematical formulas or (2) SQL codes. In addition, this tool helps to clarify (1) primary keys, (2) whether the target column can allow null as data, (3) data types, and (4) data lake systems. Often times, the source systems are legacy inhouse IT systems (or even multiple excel files from certain departments.) In contrast, target system is most likely data lakes, powered by cloud service providers.

**(3) Details on Key Data Domain**

In this section, I discuss key data domains shown in Figure 2 in details.

**3–1. Product master**: Product master table contains product details with UPC key as a primary key. You can include product attributes (e.g., manufacturer, brand, pack size, flavor) together with product hierarchies (category, sub-category, department, aisle.) Almost always, there will be promotion groups or pricing groups where a group of products will share the same pricing or promotion discounts. It is good to capture these mapping. Volume equivalent needs additional explanation. For example, cola categories can have 2 products with different number of cans: 12 package / 12 Fluid Ounce Cans vs. 24 package / 12 Fluid Ounce Cans. Counting each product as 1 unit is not comparable, thus we need “volume equivalent” to convert units of product with different package sizes to comparable units. If we define “12 Fluid Ounce” as 1 volume equivalent, 1 unit of 12 package / 12 Fluid Ounce Cans is expressed as 12 volume equivalents. Similarly, 1 unit of 24 package / 12 Fluid Ounce Cans is expressed as 24 volume equivalents. Instead of using raw units, we use “volume equivalent” to measure volume (or unit sales).

**3–2. Store master**: Store master table provides details on stores. The primary key for this table is “Store ID”. In addition to chain banner (e.g., Safeway, Vons) and (local) markets, you can include (1) store characteristics (e.g., size of store), (2) store demographics (e.g., average household income, average age), and (3) retail competition intensity (e.g., # of competing supermarkets.) Zone pricing and Zone promotions are quite common. Thus, it is helpful to include this mapping information.

**3–3. Point-of-Sale (POS)**: This is the most critical table with actual unit sales data with retail prices and discounts. The primary keys are (1) Store ID, (2) Day / Week Key, (3) UPC Key, (4) Event offer ID. Event offer ID provides a mapping to promotion event / tactic details from “Causal” table. This table is usually made from POS transaction database.

**3–4. Causal**: Causal table provide details on price and non-price (e.g., feature, display) promotions, which are not available in POS transaction data. The primary keys are (1) Store ID, (2) Day / Week Key, (3) UPC Key, (4) Event offer ID. As an example, transaction database provides (gross) revenues before any discounts and total discount amounts. However, it is hard to know purchase requirements (e.g., 4 to qualify for Must Buy 4 promotion), promotion tactics (e.g., Buy 1 Get 1 Free, Must Buy 2), promotion prices ($2.49 per each), presence of end-of-aisle display or feature advertising, just based on transaction database.

**3–5. Financial**: Financial table contains financial data for products, which can change over time. The primary keys are (1) UPC key, (2) day / week key, (3) Event offer ID. This table provides critical cost information from Finance department, which helps to measure profits beyond revenues. Note that activity-based cost accounting is usually used to allocate certain cost elements to the product level. In the case of retailers, promotion can be supported by additional funding from manufacturers (i.e., vendor funding). Thus, it is important to capture vendor funding to calculate true cost of running promotion. Separating out COGS (cost of goods sold) from cost of promotion helps for us to see what the main drivers of costs are.

**Build Data Foundation: Manufacturer Case**

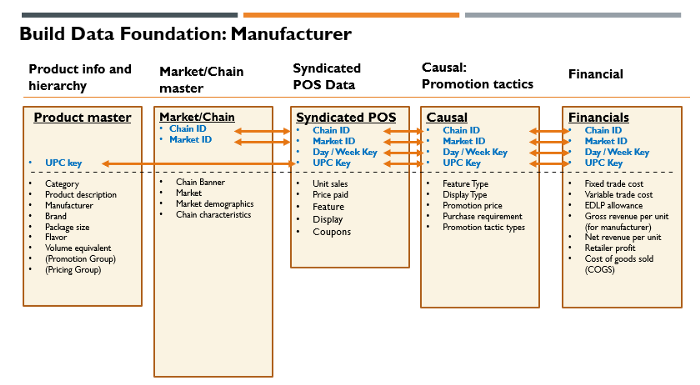


Figure 5: Image by [Minha Hwang](http://medium.com/@minha.hwang)

For the case of manufacturers, key data domains are similar. There are a few critical differences, which I call out below:

**(1) Syndicated data providers**: In the consumer-packaged goods industry, [IRI](https://www.iriworldwide.com/en-us/solutions/analytics-and-in-market-execution/price-and-promotion) and [Nielsen](https://nielseniq.com/global/en/client-learning/retail-measurement-services/) collect data from retailers in supermarket (e.g., Albertsons, Safeway, Kroger), pharmacy (e.g., Walgreens, CVS), mass merchandiser (e.g., Wal-Mart, Target), and warehouse club (e.g., Costco, Sam’s Club) channels through their retail partnership, process the data, and provide data to manufacturers after subscription fees. Usually, retail data are aggregated to the account level, which is defined by chain banner / market combinations (e.g., Kroger — Dallas.) Therefore, this will be the lowest level of data grain. Note that CPG manufacturers can have a view on their competitors since syndicated data providers provide data on competing manufacturers as well. In addition, at least for “featured” promotions, CPG manufacturers can collect promotion tactic details (e.g., Must Buy 4 at $2.99) from a service, called [Numerator](https://www.numerator.com/). Numerator scans feature advertising inserts and generate data tables for featured promotions. Syndicated data provides also offer product master, market demographics, and chain characteristics. This situation is very different in other industries (e.g., Consumer Electronics or High-Tech), where these kinds of service is not available. In those cases, manufacturers are half-blind, since they can track their own promotion data, but not competitor promotion data. The best external data for High-Tech or Consumer Electronics industries are provided by [NPD](https://www.npd.com/) as household panel data (instead of POS data from retailers.) In case the manufacturers are working as category captains for select chains, that could another good source of data for promotion effectiveness measurement.

**(2) Sell-in vs. Sell-through**: Manufacturers can be tempted to use shipment data to retailers (sell-in) instead of retail sales data (sell-through.) This is a bad idea due to forward buying from retailers. Retailers can stockpile their inventories when there are good deals, but this may not move from retailers to consumers. Thus, it is important to use retail sales data instead of factory/distribution center shipment data.

**(3) Wholesale prices vs. retail prices**: It is less common for manufacturers sell directly to consumers (i.e., DTC). Manufacturers first sell to retailers, and retailers sell to final consumers. There can be intermediate wholesalers / distributors in the picture as well. Because of this, the topline for manufacturers will be different from final retail prices. Retailers add additional retail margins on top of manufacturer margins (i.e., double marginalization). For manufacturers, having visibility for both manufacturer margins and retailer margins are important to have data-driven discussions with retailers. Thus, in the financials table, gross/net revenue per unit of manufacturers are included together with retailer profit. It is also important to track regular price reduction allowance (EDLP allowance), fixed trade cost, and variable trade cost.

Hope you find Step 1 of the process useful. In the next article (Part 3), I will focus on Step 2 and discuss how we can create modeling dataset from the central data repository that I described in Part 2.