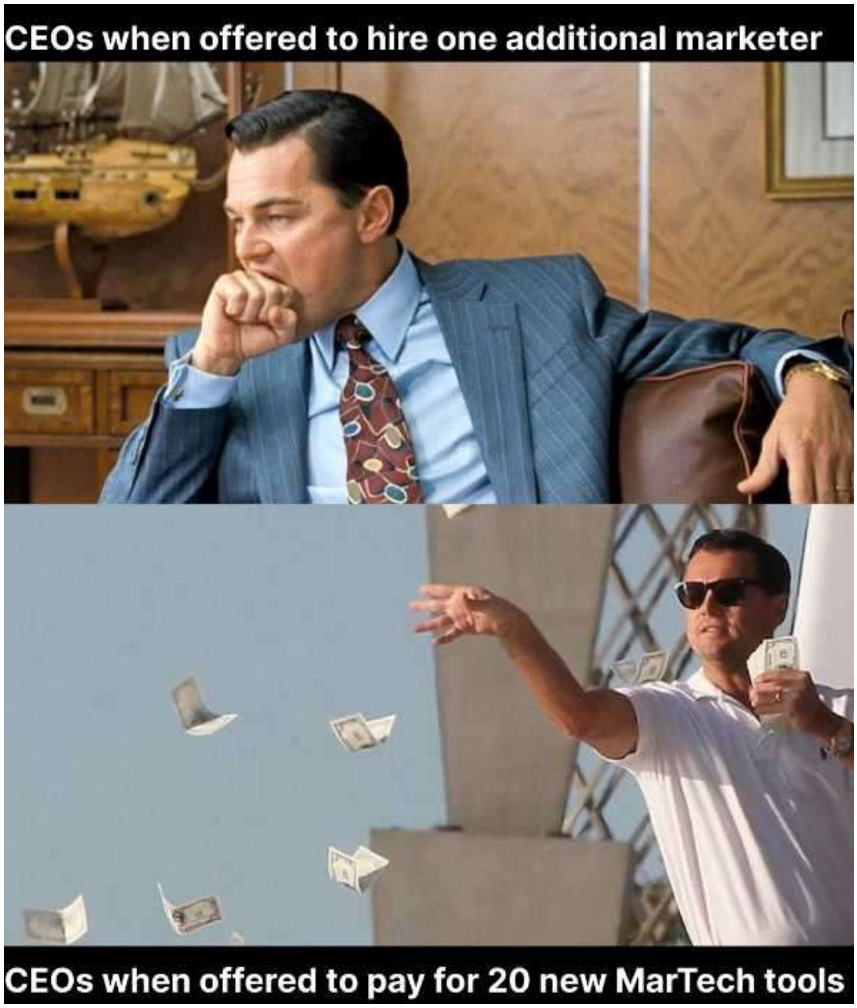
BIG DATA

What is Marketing Data Cleansing? Purifying Metrics and KPIs.

OLEKSANDR SHYKOLOVYCH OCTOBER 13, 2022

With the abundance of MarTech and AdTech solutions, it is becoming more and more challenging to keep track of all of the data generated by these tools. Considering that companies are willing to pay for MarTech tools but not for marketers themselves, we can only expect the amount of data gathered to grow in the near future.



The main problem marketing teams face is making use of all of the data generated by these tools. Marketing platforms don't guarantee the granularity and clarity of the extracted data, which means users have to resolve data inconsistencies by themselves.

Cleansing marketing data is a crucial stage of the analytics process. With low-quality data, you can't accurately measure the results of your marketing efforts, so you're inevitably missing valuable insights into the upsides and downsides of your campaigns.

In this guide, you'll learn what marketing data cleansing is and how to perform it in a time-efficient manner.

What is data cleansing?

First off, we have to figure out what data cleansing is.

💡Data cleansing is the process of identifying and correcting inaccuracies and inconsistencies in datasets.💡

When merging data from multiple sources, there's a chance that the data will be duplicated, mislabeled, or even end up in the wrong dataset.

If your dataset contains false data, the further analysis makes no sense because it'll generate misleading outcomes that don't reflect the real situation. Furthermore, you can't train machine learning models with this incorrect data or visualize your data with BI software.

Why is data cleansing important for marketers?

Marketing analysts are some of the most vulnerable specialists when it comes to data inconsistencies. A regular analyst can work with data from tens of marketing tools merged in a single database. Some of these tools have mediocre APIs that can't guarantee data integrity while data is transferred to the user's database.

Marketers need precise information for several reasons:

- Marketing teams decide which channels generate revenue and which don't, relying on the information they get from these channels.
- Marketers adjust their strategies/campaigns/tone of voice depending on the aggregated data they gather from all of their marketing platforms.
- Companies assess product positioning and brand acceptance by relying on information gathered by marketing teams (e.g., net promoter score).

With inaccurate metrics, the company is making decisions based on untrustworthy data, which means they will eventually take incorrect actions that result in a revenue loss.

Moreover, incorrect data makes it impossible to determine which channels generate the most leads, thus overcomplicating the whole attribution process.

Finally, incorrect data also undermines the productivity of marketing teams because analysts have to manually double-check their datasets in search of duplicate data, wasting tens of working hours a week on routine operations.

Challenges of marketing data cleansing

Data cleansing isn't always a walk in the park. The more raw data you gather, the more time it takes to prepare it for analysis.

Data cleansing challenges can be divided into two groups:

- Single-source problems
- Multi-source problems

Single-source problems

Data quality largely depends on schemas and integrity constraints that control data values.

Sources, such as files, don't have a data schema at all. Instead, they might have a few constraints that control the quality of data that can be entered and stored. However, these constraints are usually not sufficient to prevent errors and inconsistencies.

In contrast, database systems often use more sophisticated data models and integrity constraints to ensure data quality. However, databases also have their own problems that may occur because of a poor schema design or data model limitations.

We can identify the following data granularity issues:

- Missing values. Dummy or null values during the data entry.
- Misspellings. All kinds of typos and phonetic errors (e.g., city = "Sun Francisco").
- Embedded values. Multiple values entered in a single data row (e.g., name = "C. Jung 06.06.75 New York").
- Misfielded values. Mismatched values that get into incorrect data fields (e.g., city = "France").
- Incorrect dependencies. Paired attributes that don't correspond to each other should be cleansed or else data inconsistencies will arise (e.g., city = "New York" zip = "90210").
- Word transpositions. The same attribute may have a misaligned word structure (e.g., name1 = "J. Bold", name2 = "Smith M.").
- Duplicates. The same record may be represented multiple times because of errors during data entry (e.g., employee1 = "John Bold", employee2 = "J. Bold").
- Contradicting records. The same entity might be described with different values in different data fields (e.g., customer1= (name = "John Bold", company = "Amazon"), customer2 = (name = "John Bold", company = "Meta")).
- Wrong parameters. A real-world entity might be described with wrongly defined parameters (e.g., customer= ("J. Bold", company = "Meta"), but actually, J. Bold

represents Amazon.)

Given the number of potential data issues, the most effective way of dealing with data inconsistencies is to prevent dirty data from entering your data warehouse. This requires an appropriately designed database schema and tight control over the transferred data.

Multi-source problems

All single-source problems are aggravated as the number of sources grows. Each source contains raw, unstructured data that may overlap or contradict data from other sources. Imagine the number of data inconsistencies when working with 10+ marketing platforms, each one of them with its own metrics names. For example, impressions may also be called “imps”, “ims”, “views”, etc.

The main reason why this occurs is because data sources are meant to serve specific user needs. They aren’t supposed to work with each other. This leads to a high level of data heterogeneity in a data warehouse or data lake.

Naming conflicts arise in the following cases:

- The same name is used for different objects (e.g., two different metrics may have the same name but different meanings).
- Different names are used for the same object (e.g., the same metric may have two different names in different data sources).

The main problem when cleansing data from multiple sources is finding overlapping data among millions of records. Experienced analysts use SQL to query data and identify errors faster, but technical expertise isn’t widespread among marketing specialists. That’s why marketers often seek help from technical analysts or outsource marketing analytics to third-party agencies and freelancers.

Data inconsistencies may also occur due to data mapping issues. Improperly mapped data doesn’t allow marketers to set up lead attribution, manage behavioral retargeting, or build a holistic picture of their marketing efforts.

🚀 Here’s our list of the 11 best data mapping tools for revenue teams. 🚀

Cleanse your marketing data with Improvado

It's clear that there are tens of different scenarios in which your marketing dataset can be incomplete, redundant, or totally inaccurate. That's why marketing data cleansing is a crucial process. It allows you to draw up granular reports and build comprehensive dashboards related to your marketing efforts.

What's more, Improvado can help you with marketing data cleansing and normalization. The platform's DataPrep module allows regular marketers to execute complex SQL queries in a no-code environment and a spreadsheet-like UI.

What's even better is that Improvado has an automated data mapping and normalization framework called the Marketing Common Data Model. With its help, a regular marketer can match data fields from different sources and align disparate naming conventions. For example, if one data field refers to impressions as "imps" and another one as "views", the framework comes up with a unified name for both of them to avoid data inconsistencies within the data set.

Schedule a consultation with our experts to learn more about Improvado's capabilities.

Oleksandr Shykolovych

Editor in Chief at Improvado

Editor-in-Chief at Improvado

Oleksandr Shykolovych is an Editor-in-Chief at Improvado blog. A strong desire to share information with people and a keen eye for details help Oleksandr create engaging content and be on the same page with readers.

San Diego | Headquarters

3919 30th St, San Diego, CA 92104

-

San Francisco

2800 Leavenworth St, Suite 250, San Francisco, CA 94133

-