

marketing data management 101

HOW TO CLEANSE, BUILD, AND REFINE YOUR MARTECH DATABASES

INTRODUCTION

You are a data-driven marketer. Your company has invested in advanced Marketing Technologies (MarTech) such as Marketing Automation Platforms (MAP), predictive lead scoring, account based marketing, and content personalization. Unfortunately, you realize none of these MarTech solutions work very well with your poor quality data. To realize the return on investment for these MarTech solutions and to excel as a data-driven marketer, you need to clean your data and customize it. This guide outlines how to build and refine your MarTech databases so they can best support your evolving business needs.



> WHO SHOULD READ THIS GUIDE?

Any data-driven marketer who is looking to get the most out of his/her investments in marketing programs and MarTech solutions by building a first-rate MarTech database.

> ARE WE SIMPLY TALKING ABOUT LEAD DATA?

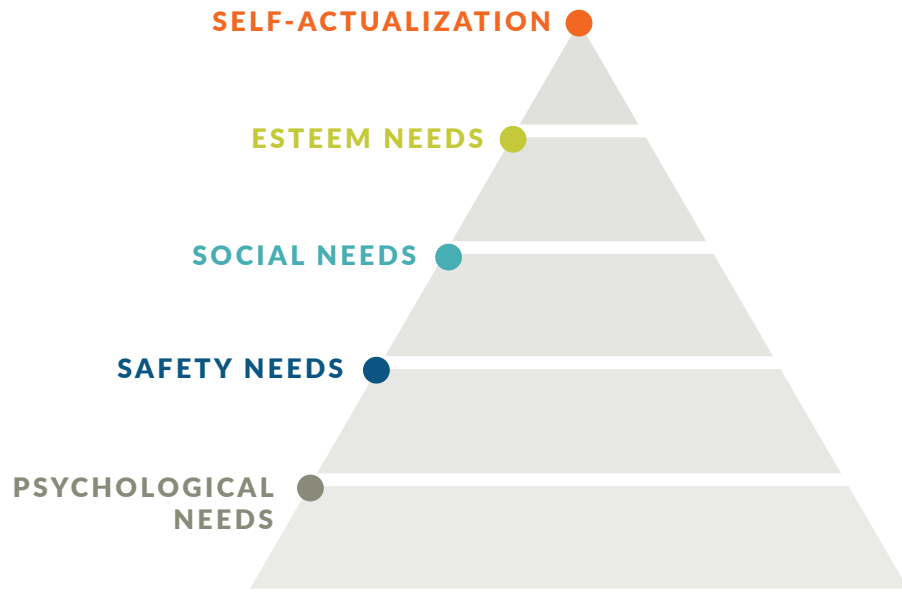
No, MarTech data goes well beyond leads. It covers all related data, spanning the entire sales funnel, including leads, contacts, accounts, engagements, opportunities, campaigns, and much more.

THE MODEL: HIERARCHY OF NEEDS

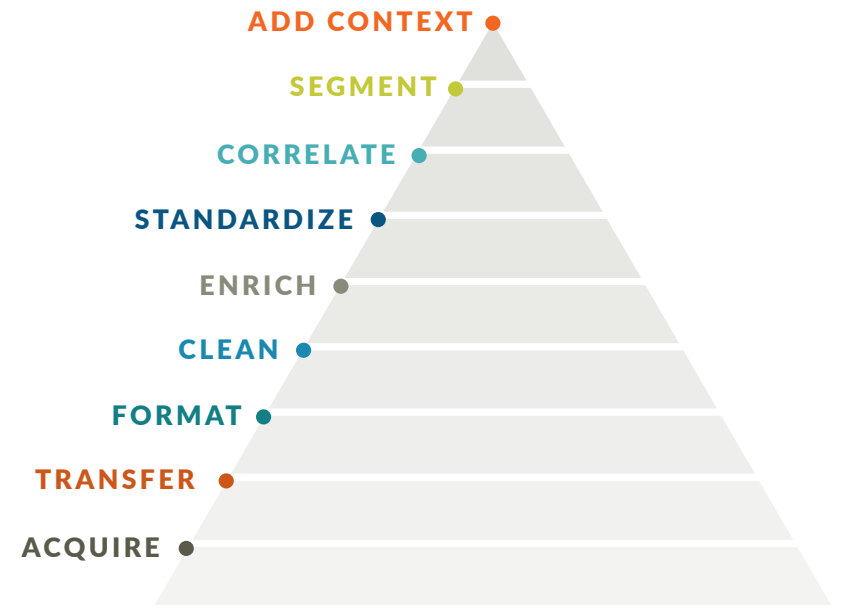
Maslow's Hierarchy of Needs model, illustrated below, describes how human motivation moves through a tiered model. In this model, it's impossible to meet higher level needs such as social and esteem until the basic needs of food and safety are satisfied.

The development of your data follows a similar hierarchical model described below. There are basic needs to acquire and clean your data. After that, the data goes through additional phases of transformation and enrichment until it becomes fully relevant to your business. At the highest level, your data can fully support your go-to-market strategy and advanced marketing technologies.

> MASLOW'S HIERARCHY OF NEEDS



> MARKETING DATA'S HIERARCHY OF NEEDS



1 ACQUIRE DATA

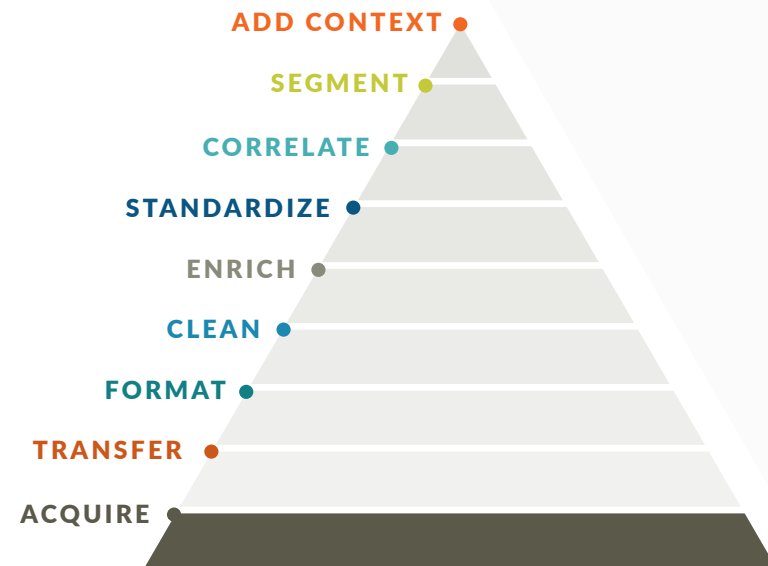
Start by acquiring MarTech data from a broad spectrum of sources. The most common are:

- Website forms
- Data entered manually by your sales and support team
- List vendors
- Enrichment, appending, and validation services
- Trade shows
- Advertising and content distribution platforms
- Social media platforms
- IP look up and user tracking technologies
- Partners
- Open data sources such as government websites

**THE MORE DATA SOURCES YOU HAVE,
THE BIGGER YOUR DATA CHALLENGE.**

Data from each of these sources is shared differently:

- File formats (csv, xls)
- Transmission methods (email, web, FTP, API)
- Number and type of data fields
- Data types (number, text, date)
- Data formats (date, currency, geo code)
- Data reference (industry code, company size, job role)



TO MAKE THIS DATA USEFUL TO YOUR BUSINESS, YOU'LL NEED TO DEVELOP IT FURTHER.

2 TRANSFER DATA

Most of your MarTech data probably originated from a third-party source. Before you can do anything to that data and make it meaningful to your sales and marketing teams, you'll have to transfer it from the third-party into your own ecosystem.

It's likely that you received much of that data as spreadsheets or CSV files via:

- Email attachments
- Internal or cloud-based shared drives
- FTP servers

These methods of transferring data usually involve manual tasks, although each method can be fully automated.

In addition, the following means of acquisition are typically fully automated:

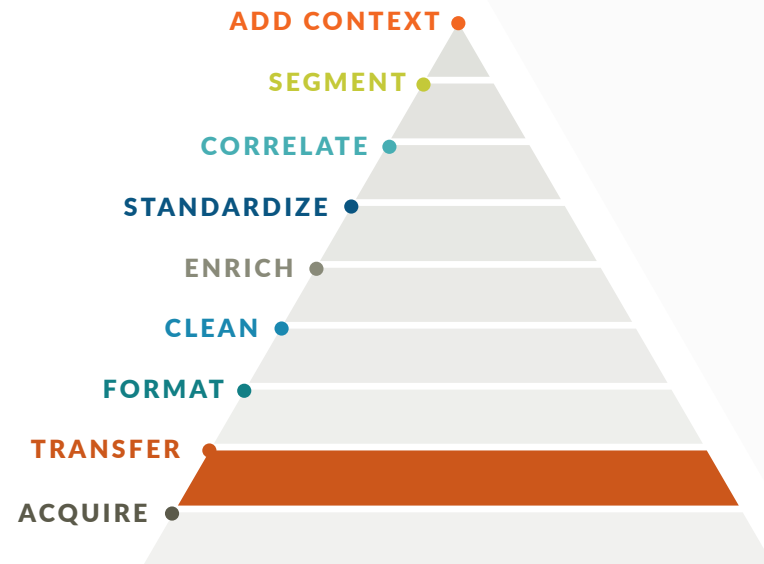
- Website scraping and crawling
- API and other types of programmatic interface

Consider automating data acquisition if your situation is as follows:

- Data is deposited frequently. For example, you buy consumer leads weekly, or you scrape a website monthly.
- The number of data sources is large. For example, your company might crawl the websites of all 50 US state governments.

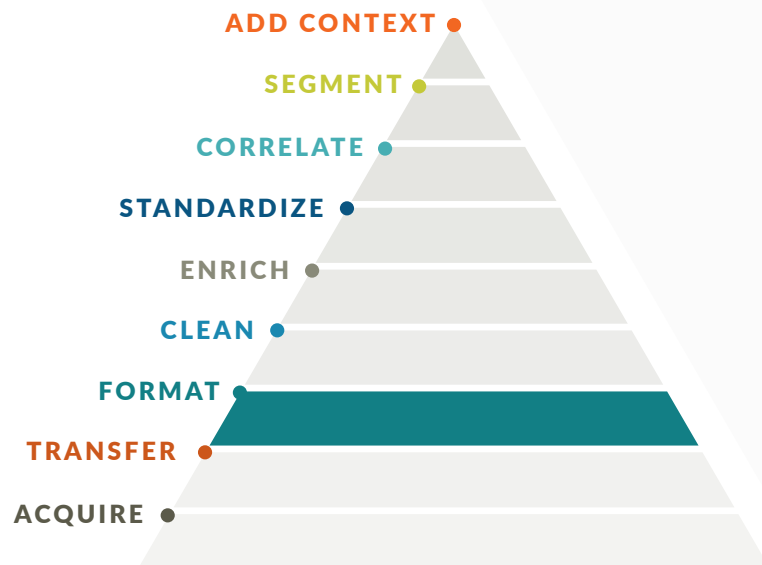
Custom development vs. data automation tools:

Before you write code to automate data deposits, consider utilizing commercial tools like the web crawler service Cloudscrape, integration service Zapier, and data automation service Openprise. These solutions can provide the required automation cheaper and faster than writing custom code.



3 FORMAT DATA

Before you can load data into your system, be it a marketing automation platform or a data warehouse, you need to ensure the data fits the requirements of the destination systems. Here are some common formatting challenges:



Field Type

The source data may be a text type, but the destination is numeric.

Field Length

The destination field may have a maximum length of say 255 characters, while your source data is longer. You may have to truncate or parse the source data into smaller parts.

Data Parts

Certain data fields like name and address have multiple parts that can be stored differently in different systems. For example, your source may have two fields, “first name” and “last name”, while your destination system has just one field, “full name”.

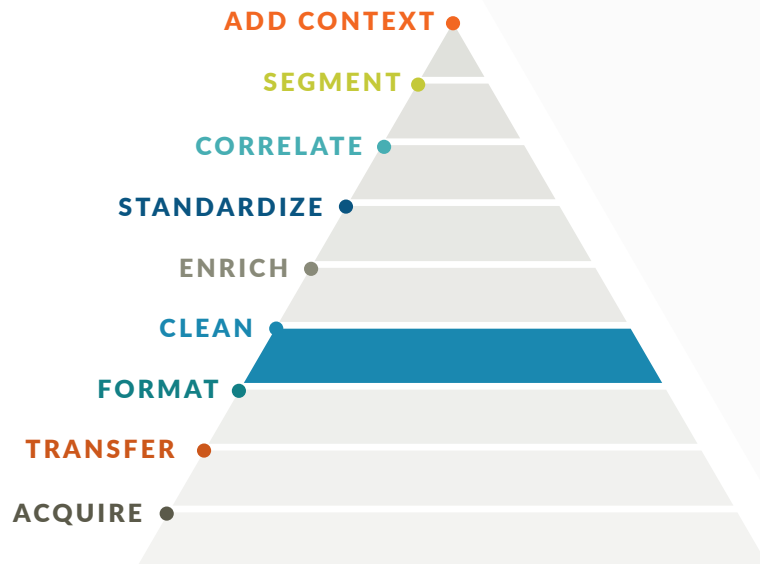
Allowable Values

Some data fields have a limited set of logical values. For example, “month_of_the_year” data field should only have values 1 through 12. But what if your data source has values of 0 or 13? When writing to a database or data warehouse, does the record have a valid key?

International Data

If your data contains multiple languages, the source data may have been exported using a standard incompatible with the destination system. For example, the source CSV file may be encoded with UTF32 using Microsoft Excel, but the destination system can only handle UTF16. Get ready for some funny-looking characters like “?½®\$??#Δ”.

4 CLEAN DATA



Once you have handled the basic mechanics of getting data where you want it, and in the right format, the next step is cleansing. Data Cleansing or Data Hygiene are very broad terms. Everyone has a different understanding of what it means. In this guide, we have a narrow definition:

Data cleansing is correcting incorrect data, with or without business context.

Examples of incorrect data, regardless of business context, include:

TYPOS

- Misspelled name: "Massachusettes"
- Incorrect name: "United State of America"
- Wrong word: "Principle Consultant"

INVALID OR STALE VALUES

- Invalid address: "San Francisco, CA, Canada"
- Invalid date: "February 29, 2015"
- Invalid domain suffix: "www.openprisetech.con"
- Invalid country = "USSR"
- Remapped ZIP code

DATA CLEANSING EXAMPLES

DUPLICATE DATA

- Duplicate records for the same person or company
- Duplicate data fields (e.g., address) for the same record
- Duplicate leads and contacts in Salesforce.com

INVALID VALUES

- Sales Territory = “Africa” while the valid territories are only “North America”, “EMEA”, “Latin America”, and “APAC”
- Subscription Plan = “Free Trial” while the valid subscriptions include only “Premium”, “Pro”, and “Enterprise”
- Shipping Address entered for Billing Address

STALE VALUES

- Account Owner = jdoe@acme.com when John Doe left the company 6 months ago
- Design partner still shows old value after the customer has changed design partners



5 ENRICH DATA

Once your basic data set is clean (relatively error-free), the next step is enriching or appending the data. There are different means to enrich your data, but every solution or service uses your existing data as the basis for enrichment.

Here are some common challenges encountered when enriching data:

Is the Third Party Data Better Than Yours?

Don't assume third party data is better than yours. For example, your sales person may have entered the direct phone number of a prospect, while the data vendor only has the main company number.

Which Data Vendor Has Better Data?

Each data vendor's database has strengths and weaknesses. It's difficult to determine which vendor to use for which parts and subsets of your data.



WAYS TO ENRICH DATA



> INFER INCOMPLETE FIELDS

Many data fields are related, such as parts of an address. Bits of an incomplete address can easily be inferred using known relationships. You can do this with a data automation tool and a set of commonly used reference data. For example:

- Use ZIP code to infer missing city and state
- Use country name to reformat a telephone number



> VALIDATE AGAINST A COMMERCIAL DATABASE

Whether you are looking to validate data such as email and current job title, or fill in missing company data such as industry and size, there are plenty of data services that have built extensive databases of people, company, and industry data that can be used to validate and append your records.

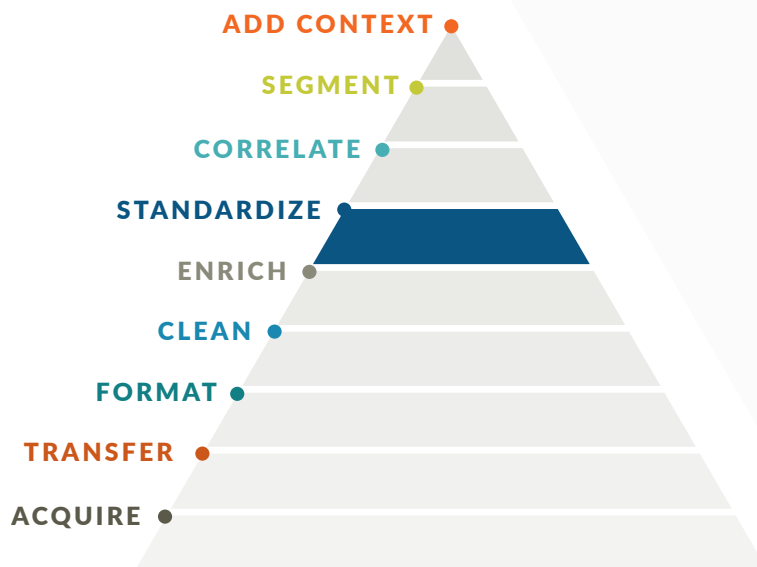


> PERFORM MANUAL RESEARCH

The highest quality data enrichment still comes through manual research. A skilled sales development rep who researches and triages from data sources like LinkedIn, company websites, Google search, SlideShare, CrunchBase, and EDGAR is still unparalleled by any software solution. However, this option is obviously limited by cost, time, available skill set, and scalability.

6 STANDARDIZE/NORMALIZE DATA

Perfectly clean data that is error-free can still be unusable. For example, you may have country names “United States”, “United States of America”, “US”, “U.S.”, “USA” and “U.S.A.” in your database. These are all accurate and commonly used names for the United States, but when you try to run a report like “Sales by Country,” you would end up with six separate data points that should be merged into one. Different systems have different data standards, so each data source can introduce data with variations that require standardization and normalization before the aggregated data become usable for process automation, analytics, and reporting.



Here are some common data sets that require standardization:

- State, province name, and postal abbreviations: “California” vs. “CA”
- Country name and ISO alpha codes: “United States” vs. “US”
- Phone number formats: “+1 415-555-1212” vs. “(415) 555-1212”
- Company names: “IBM” vs. “International Business Machines”
- Lead sources: “User Group San Francisco” vs. “SF User Group Meeting”
- Product names: “Super Duper Analytics” vs. “SD Analytics Module”
- Part numbers: “ABC12345” vs. “ABC-12345”
- Sites and locations: “Phoenix Distribution Center” vs. “DC-PHX”
- Territories and geographies: “Asia Pacific” vs. “APAC”

DATA STANDARDIZATION CHALLENGES

Standardization is simple in theory, but very challenging in practice, for the following reasons:

Scalability

To standardize data, you must scrub every single record in your database. Not just new records, or records from one specific source, or records of a certain age. EVERY SINGLE RECORD, whether that's 100,000, one million, or 10 million records. Productivity tools like spreadsheets are ill-suited for such a massive undertaking.

Variability

How many possible variations will you encounter in your source data, even assuming your data is relatively clean? In addition to "California" and "CA", how about "Cal", "Cal.", "Calif", "Calif."?

It's also common to see mixed uses of data. For example, in addition to all the variations of "California", you may also see "SF", "SF Bay Area", "SF Metropolitan Area", "SFBA", and "SFO" all mixed in. This doesn't even include typos and non-standard variations you will inevitably encounter. Instead of relying on maintaining an exhaustive list of reference data, consider data automation technologies with built-in fuzzy match capabilities.

Manageability

How much configuration is required to normalize your data? The sheer amount of work required to set up and maintain the automation can make or break your project. For example, if you wish to standardize the states and province names for USA, Canada, France, Germany, and Australia, and normalize the names of all countries in the world, how many rules would you have to create? If the answer is 500+ rules, one for each country and state, then your project is doomed to become unmanageable very quickly. If the answer is one rule for state and one for country, then you have a manageable solution that can succeed. Look for an automation solution that utilizes the most efficient and manageable configurations scheme.

System Performance

Many MarTech solutions have built-in rules engines that theoretically can be used for data cleansing and normalization. Manageability aside, having many rules running on a large database can often slow system performance to a crawl. A system designed to automate processes like nurturing campaigns and lead routing is usually poorly suited to automate large scale data processing. Bottom line: make sure you use the right tool for the right job.

7 CORRELATE DATA

Whether you're looking for specific business insights or wish to automate processes across business functions, you'll often need to work with data across different systems. When bringing data in across multiple sources, you'll need to correlate it using known relationships. The common correlation use cases are:

Correlation Using a Unique Identifier / Key

This is the simplest scenario because the relationship is unique and explicit. The technical jargon is a "database join." This is often the case when you are working with different data sets from the same system, or data sets from different systems that already have good integration. For example, correlating a Contact record to an Account record from Salesforce.com is as straight forward as looking up the Account_Name, and using the AccountID that's already part of the Contact record. Similarly, correlating a Lead record from Marketo with a Lead record from Salesforce is as simple as using the Salesforce record ID that's already part of the Marketo lead record.



A MORE CHALLENGING DATA CORRELATION SCENARIO

Correlation Using Supposedly Unique Values

When data across different data sets does not yet have unique keys to explicitly link records, correlation must be done by matching specific data fields that “should be” unique. A common example is attempting to determine if a new lead from Marketo belongs to an existing account in Salesforce or is a net-new lead.

In order to make this determination, you might attempt to correlate the lead’s email address to an account’s website by matching on domain name. However, this is not foolproof, as an account may own a number of domain names and the lead’s email address may be a personal one.

If the lead fails to correlate on domain name, then a correlation using company name can be a second option. But this is also challenging, since account names are commonly neither clean nor standardized.

Without high quality data for email addresses, website domains, and company names, a correlation task like this will achieve a very low success rate.

To Fuzzy or Not to Fuzzy

Correlation can be difficult when data quality is not the best. Fuzzy matching, which is when a system finds a value that’s close, but not exactly the same, can improve the correlation results. But it’s not without its own challenges. As with any machine-based algorithm, fuzzy match can produce inaccurate results:

FALSE POSITIVE:

Correlation found that should not have been found.

FALSE NEGATIVE:

Valid correlation that should have been found but was not.

Experimenting with the fuzzy matching algorithm’s matching parameters can greatly improve results.

8 SEGMENT DATA

Almost any modern data-driven marketing technology you can think of – whether it’s account-based marketing, personalization, or predictive scoring – requires a well-segmented database.

What is Segmentation?

Segmentation is the process of grouping data based on shared characteristic. Common segmentation dimensions span demographic, firmographic, technographic, and behavioral characteristics. Some examples include:

B2B DEMOGRAPHICS:

Job function, job role, job level, and buyer persona

B2B FIRMOGRAPHICS:

Industry, company size by annual revenue or employee count, and geography

B2B TECHNOGRAPHICS:

Technologies adopted such as CRM, MAP, ERP, competitor products, etc.

B2B BEHAVIOR:

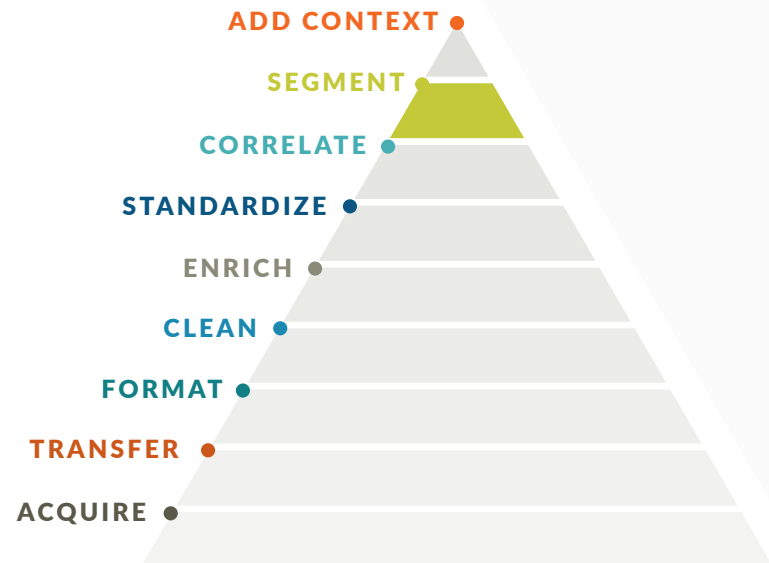
Email response history, asset preference history, on-line hang outs, and trade shows attended

B2C DEMOGRAPHICS:

Age, gender, education level, geography, and income

B2C BEHAVIORS:

Purchasing history, mobile device usage, and browsing history



THE WRONG WAY TO SEGMENT

The Way Most Marketers Do Segmentation Yields Poor Results

Most marketers practice some level of segmentation by building lists using filters. These filters can be static or dynamic, meaning the resulting records can change over time based on a set of criteria. The filtering is based on a specific set of keywords or attributes. This filtering-based approach yields poor results because of:

LACK OF CONSISTENCY:

The filters are often created for each individual campaign and report, thus lacking consistency across programs, reports, teams, and time.

POOR QUALITY:

It is difficult to validate the quality of a list that is the result of such filters, because while it is relatively easy to see false positives, it is hard to identify false negatives.

FALSE POSITIVES:

Person with the title “Assistant to the Vice President” is included as an “executive” because of matching on the key words “Vice President”.
Person with the title “Contractor” is included as an “executive” because of the word “contractor” contains the three letters “CTO”.

While identifying false positives is easy, it can be difficult and time-consuming to construct filters to avoid them.

FALSE NEGATIVES:

False negatives are difficult to identify. For instance, say you are attempting to segment on the marketing job function using a list of keywords. How would you ensure not to overlook people due to the following?

- Unusual titles such as “Growth Hacker”, “Digital Engagement Specialist”, and “Industry Program Director”.
- Job title keywords such as “SEO/SEM”, “content”, and “digital”.

HOW DO YOU KNOW YOUR SEGMENTATION IS ACCURATE AND EXHAUSTIVE?

THE RIGHT WAY TO SEGMENT

Segmentations Should Be Permanently Tagged As Part of the Data Record

For example, add customer attributes Job Function, Job Level, and Persona to the lead record. Use a data automation tool to derive the segmentation results, and add the results to the records. Automate the following tasks:

1. Infer Job Function and Job Level from Job Title
2. Map your 2000+ industries to a list of 10-15 that your go-to-market strategy is based on
3. Segment company size based on annual revenue and employee count to fit your perspective of the market
4. Assign Personas using combinations of Job Function, Job Level and account level segmentation data
5. Automatically update the segmentation when the data is updated

The Benefits of Permanent Segmentation

- Analyze your database segmentation results and make sure most, if not all your leads are segmented. For example, can you produce a chart like the one shown to the right that displays the job function breakdown of your lead database?

Job Function	Record Count (%)
Sales	40.1%
Marketing	18.1%
Engineering	12.1%
Product	8.1%
Operations	7.1%
Finance	6.1%
Support	5.1%
Human Resources	4.1%
Other	3.1%
- Ensure segmentation is done consistently across all your marketing efforts.
- Don't do the same work over and over again. Build the segmentation logic, and make sure it automatically runs through your data on a regular basis.
- Leverage your segmentation results. For example, does your account-based marketing tool or predictive scoring tool require job segmentation and persona data? Instead of repeating the segmentation effort in these tools and generating inconsistent results, have these tools use the segmentation from your lead database.

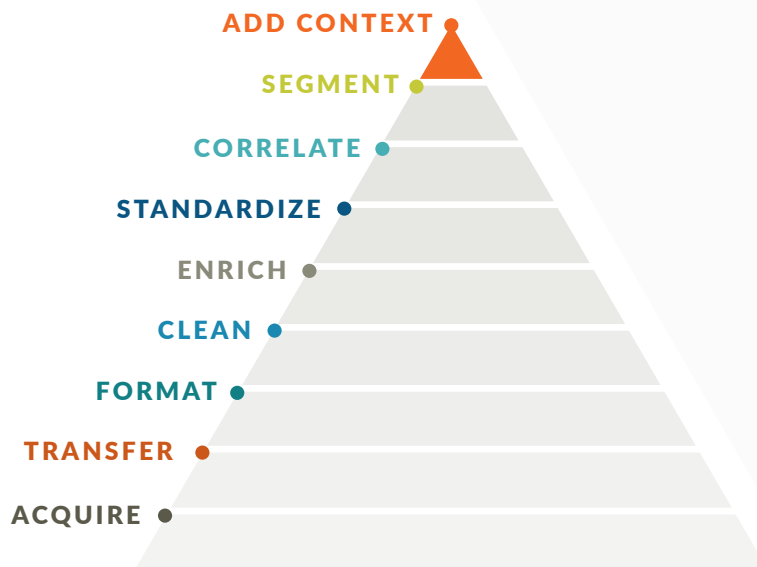
9 ADD CONTEXT TO DATA

All the steps in the data development process thus far have been mainly generic, with common use cases shared across market verticals and types of business. The last step, however, is about adding context and customizing the data's relevance to your business. While there are no universal use cases for adding context, we can illustrate it with some customer examples.

Customized Industry Context

There are well-established industry segmentation frameworks such as the North American Industry Classification System (NAICS) and Standard Industrial Classification System (SIC). Most data vendors can enrich your data with these standards. However, your business may run on very specialized segments. It would make sense to contextualize your data by remapping standard industry data to your specific target industries.

For example, one of Openprise's customers is an Internet of Things (IoT) technology company. Its target industries contain the likes of vehicle telematics, agriculture technologies, smart cities, and vending machines. While every third-party data source has Coca-Cola as a food and beverage company, to this IoT company, Coca-Cola falls into the category of "vending machines".



MORE EXAMPLES ON ADDING CONTEXT TO DATA

Situational Context Example

While most companies segment leads by demographic, firmographic, and technographic dimensions, your business may have a unique perspective on how to categorize and profile prospects.

One Openprise customer is a financial services company that makes loans to businesses. Its customer segments are based on the reason why a business needs a loan. For example:

- New business
- Expanding business
- Distressed business

This segmentation can be derived from the source of the lead. For example, leads sourced from a local government's business registry are segmented as new business, whereas leads sourced from tax authorities for owing back taxes are segmented as distressed business.

Knowing why a business needs capital enables this company to do very targeted marketing campaigns with precise messaging and value propositions.

Historical Context Example

Another Openprise customer sells a medical service for newborn babies. This company must consistently buy new leads who are expectant parents. Once the baby is born, the lead expires and is retired from the marketing database into an archive. On subsequent pregnancies, the same expecting parents may come through the lead funnel again.

This business has specific marketing campaigns targeting:

- First time parents
- Experienced parents who have not been marketed to before
- Experienced parents who have been marketed to before but did not purchase
- Experienced parents who have purchased before

Each new lead is de-duplicated against not only the current database but also the archive database. When a duplicate is found against the archive database, the lead is segmented as an experienced parent; and if not, as a new parent.

BASIC TECHNOLOGY CONSIDERATIONS

If you are ready to invest in a technology solution to help you build and develop your MarTech databases, you are simultaneously in luck and in trouble. You are in luck because there are many excellent solutions to choose from. You are in trouble because the selection is so vast that it can be very difficult to determine which solution is best for you, or to decide if what you already have is adequate. Here are some key considerations:

Do You Really Need New Technology?

Data problems result from:

- People
- Process
- Technology

First, determine what is causing your data problems. Fix the people and the process root causes first. Without fixing these two primary factors, no amount of technology will solve your problems.

Once you have your people and process issues squared away, you are then ready to determine if you need new technology solutions and to decide which solutions are best for your circumstances. Beware of purchasing new technology without addressing the people and process causes, as this can worsen your data problems.

Standalone Data Automation Tool vs. MAP/CRM's Built-in Rules

All leading MAP and CRM platforms have built-in rules that can be used to perform data processing. However, since these rules are designed primarily for process automation, they are often ill-suited for processing large amounts of data. Common challenges you can expect to encounter include:

- Large numbers of rules will need to be created and managed, e.g., to normalize the names of US states and territories alone, you will need over 60 rules.
- No built-in fuzzy logic, so you will have to provide an exhaustive list of possible variations and typos.
- With a few hundred rules running, many of these MAP / CRM platforms will become unresponsive.

For continuous data management of large data sets, a Data Automation solution built to specifically address these challenges is a more scalable and manageable option compared to MAP and CRM.

ADVANCED TECHNOLOGY CONSIDERATIONS

One Multi-Purpose vs. Many Single-Purpose Data Tools

There is no lack of data tools designed specifically for marketers. However, most of them perform only one or two tasks, such as de-duplication or cleansing. Only a few are general purpose tools that can perform many or all of the tasks in the data development hierarchy.

Determine what your data processing needs are before you decide which tools are best for you. Single-purpose tools usually take less configuration and can be easier to use, but you will have to purchase, deploy, and integrate multiple tools. A general purpose data platform provides the simplicity of an all-in-one solution, but will require more effort up front.

Don't fool yourself into thinking you can get away from technological complexities. It's simply a trade-off between a steeper learning curve for a more powerful platform vs. the complexity of integrating multiple, simpler tools. There is no universally correct choice, but it is a matter of preference and fit to your team's resources.

Data Automation Software vs. Data Services

Data Services have master databases of people and company data. They use this master database to validate and enrich the data you already have. Many data service vendors provide some technical integration to make it simple to consume their data either in batches or in real-time.

Data automation software does programmatic data processing based on configuration, whether it's cleansing, segmentation, or transformation. Data automation software can perform enrichment as well, but it will require a reference database provided by you, the software vendor, or a third-party data service.

Data automation software and data services are complementary, not mutually exclusive. Using data automation software to clean and enrich your data before sending it to a data service can greatly improve matching accuracy. Once the data is enriched, data automation software can then further standardize, segment, and add context.

about openprise

www.openprisetech.com

info@openprisetech.com



> [SEE A DEMO](#)

Openprise is a Data Orchestration Platform. We solve the garbage-in/garbage-out problem to make data-driven anything possible in Marketing, Sales, and Support. Openprise automates critical data management processes including data onboarding, cleansing, enrichment, and unification. Openprise is designed from the ground up for CRM, so it has the business rules, best practices, and data built right in, and it seamlessly integrates with CRM solutions like Marketo, Eloqua, Pardot, Desk, and Salesforce, so you're up and running fast.

OPENPRISE™