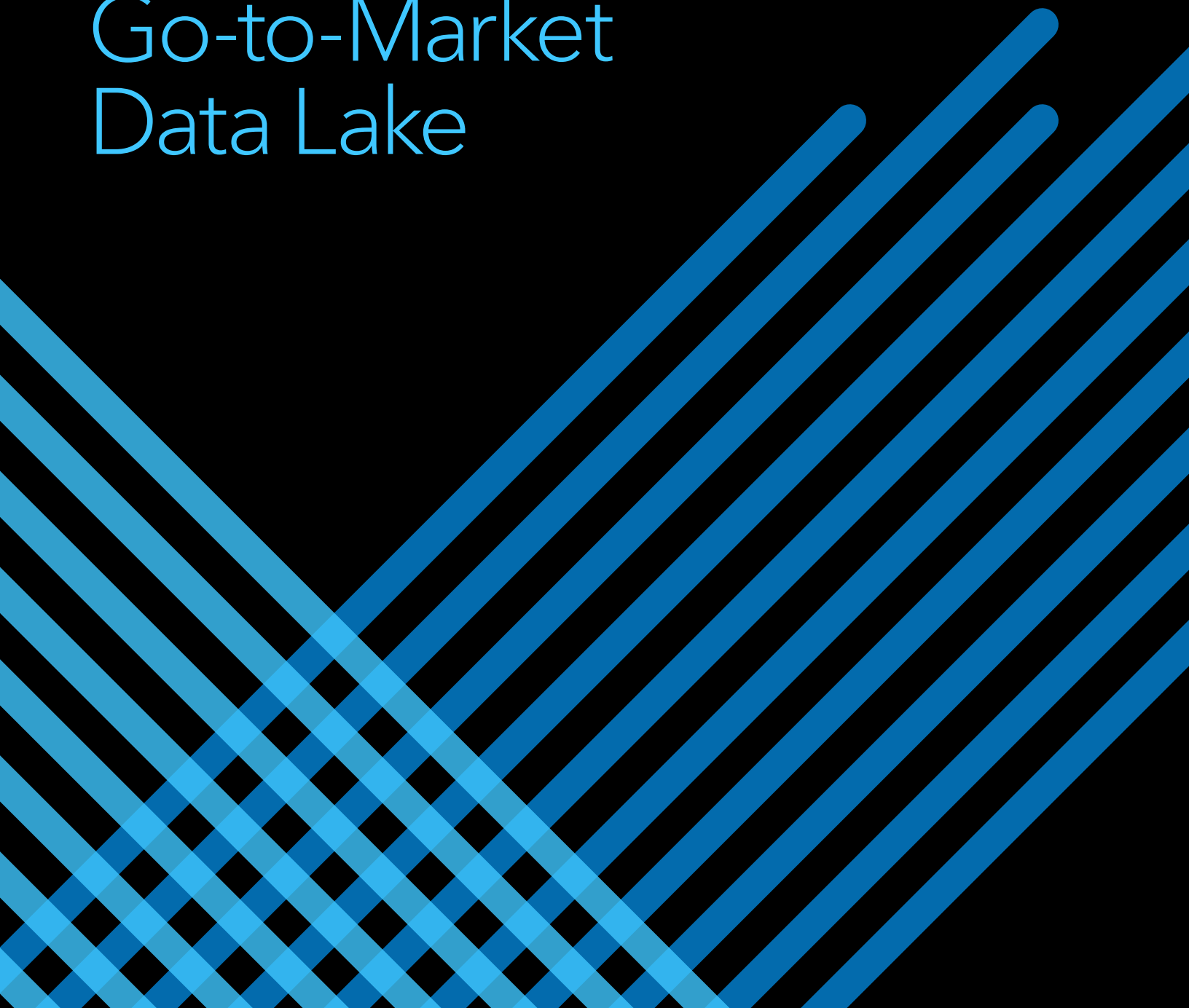
marketbridge
Reinvent growth

EXECUTIVE WHITEPAPER

The Superpowered CDP
Building a Go-to-Market Data Lake



A comprehensive resource on building a Go-to-Market Data Lake

Over the past ten years, the Customer Data Platform (CDP) has become a critical piece of marketing technology. The CDP seeks to provide a "360-degree view" of customers and prospects, promising a personalized experience that can drive revenue growth and retain customers. CDPs are typically offered by marketing automation software companies like Adobe and Salesforce and specialized challengers like Tealium.

The core idea of the CDP is not new; marketing databases have been around since the 1970s, starting as "lists" of customers or prospects, who could then be contacted for acquisitions and cross-selling. As relational database technology matured, these marketing databases were linked to promotional files (who was sent what), orders, financial information, demographics and other tables.

What is new is the volume of data coming from digital marketing. What makes marketing "digital" is not the channel, per se; it is its delivery mechanism. Video, mail and print "media" are now all delivered digitally. Digital delivery means that each interaction has telemetry; instead of estimating reach (the number of individuals who consume the media and its messages), we know who we reach precisely (assuming no fraud is happening—but that's a story for another day).

However, there is evidence that the software-based CDP concept is struggling to gain traction. In 2021, Gartner's Customer Data Survey[1] found that only 14% of organizations had achieved a 360-degree view of their customer. Among those, 44% of respondents said their 360-degree view was located in a customer data platform.

Gartner's "Hype Cycle" has CDPs squarely in the "trough of disillusionment."[2] These difficulties largely stem from saddling the CDP—fundamentally an engagement and personalization data store—with too many expectations.

By early 2024, Gartner found that while 67% of respondents to its annual marketing technology survey reported having onboarded a CDP, only 17% reported high utilization.

This paper posits that a modern, independent and comprehensive marketing and sales database can become a true single source of truth for the enterprise's go-to-market activities, something we have named the Go-to-Market Data Lake (GTMDL). In this vision, the benefits of the data lakehouse architecture combine with a reproducible, scientific ethos to provide measurement, machine learning, AI capabilities, true analysis and activation for marketing, sales and customer experience, all while staying loosely coupled with the inevitably changing landscape of marketing technology. In this vision, the GTMDL is fed from and feeds 360-degree, omnichannel activation via the CDP while at the same time integrating above-the-line and offline stimulus, sales data, prospect and customer profiles, and a uniform, managed go-to-market taxonomy.

Table of contents

Use cases

Use case-based design is a best practice to avoid building unnecessary, overly complex or redundant technology. In this approach, stakeholders from across the business surface and detail the jobs that they do to make the business run. Current state, high priority and nice-to-have uses are segmented and described. The best technology solution is bought or built based on these use cases.

Traditional CDPs are created to drive automated marketing programs and are therefore focused on customer activation. Salesforce Marketing Cloud (SMC) and the Adobe stack are competing to provide personalized, one-to-one marketing and a "360-degree view" of the customer is necessary to achieve the vision. CDPs must thus be timely and fast; in other words, reactive content must quickly be able to query the database, and the information must be up to date.

We propose a more comprehensive, analytically focused extension to the CDP, owned by the enterprise and meant to future-proof against unforeseen marketing and technology changes. In this vision, a fast, open and comprehensive database provisions upstream, single-source-of-truth data for measurement, prediction, analysis and activation (feeding CDPs and other martech platforms).

Measurement

MTA and MMM

While multi-touch attribution (MTA) has fallen out of vogue recently due to increasing privacy restrictions, tying customers' journeys together in longitudinal (over time) chains remains relevant. These chains are a combination of brand-driven media touches (stimuli) and consumer-driven behavior and engagement that encompass all of the media a prospective customer encounters related to the category and its vendors. This includes all of the discussions they have with peers, influencers and former buyers; all of the engagement they have across digital and offline marketing and sales channels; and all of the content, creative, pricing and offers they encounter as they learn, shop, purchase and use the product or service.

Understanding customers' attributes, the reach and frequency of messaging, channel preferences, how content is consumed and the ideal MTA 'path' will continue to be the Holy Grail in scaling a high-performing activation engine. By examining a prospect's journey from landing page to close, go-to-market teams can also identify points of friction, profile audiences and conduct last-touch channel analysis.

While MTA has become more challenging to implement due to privacy restrictions, it is additionally limited in its assumption that media, marketing and sales encounters are merely a combination of deterministic paths. In contrast, Media Mix Modeling (MMM) models stimulus and response as an accumulation of interactions, transactions and brand experiences over time. As a consumer gains awareness and preference for a brand, the yield from certain paths may improve. The path yield will decline for consumers with less awareness, less need and less brand affinity.

Fortunately, a comprehensive longitudinal human record (LHR) of prospect and customer interactions also provides the base query for creating an MMM panel. This panel is a summation of stimulus (x-variables) and response (y-variables) by day or week across one or more cross-sectional dimensions. As shown below in Figure 1, Record-level stimulus and response make up the base of the LHR, with joins from lookup tables/ metadata. Any missing gaps are then filled in with aggregated stimulus and response—for example, linear television or digital marketing that cannot be resolved to the individual.

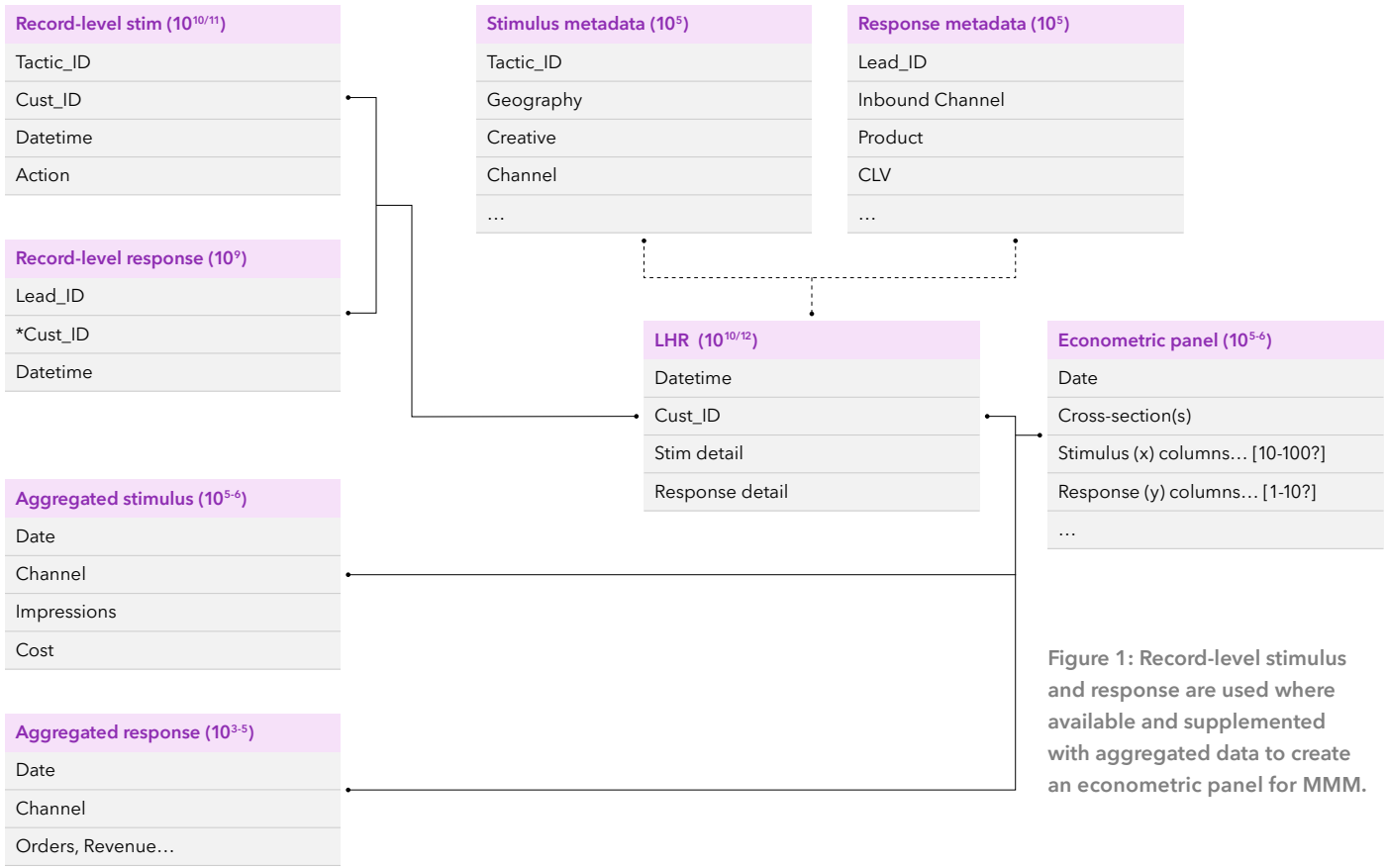


Figure 1: Record-level stimulus and response are used where available and supplemented with aggregated data to create an econometric panel for MMM.

Testing

While statistical modeling techniques provide a holistic picture of fractional attribution, true cost per acquisition (CPA) and diminishing returns, the "signal" required to feed these models can be lacking. In-market tests can quickly add precision to MMM and MTA models. Tests can also be easier for executives to understand.

The Go-to-Market Data Lake (GTMDL) should make testing fast, accurate and scalable. Tests come in two varieties and multiple flavors. The two varieties are A/B (binary) and multivariate. A/B testing is far more common—even today, with the cost and speed of testing improving every year thanks to digital. An A/B test has a test (the factor you are trying to gauge the effect of) and a control. Data-wise, linking stimulus and (potentially) response to a test ID and a cell ID (test-control) means that data scientists reading tests can build their cells with simple queries versus going back through CSV files, Excel spreadsheets and campaign briefs to rebuild the structure forensically.

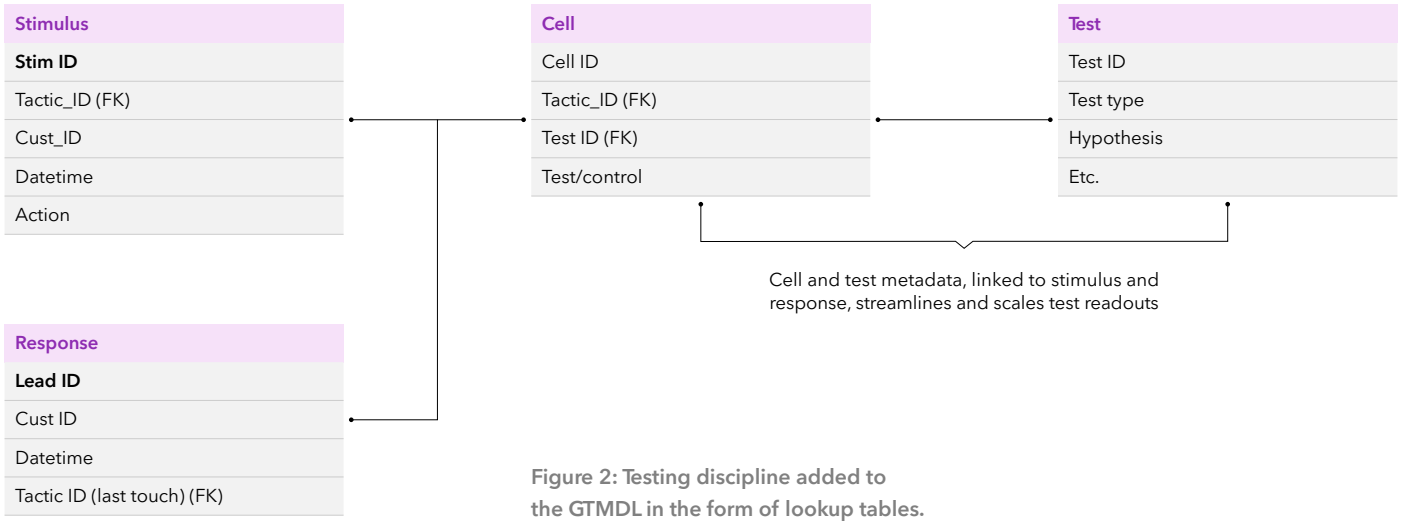


Figure 2: Testing discipline added to the GTMDL in the form of lookup tables.

Executive reporting

Chief Marketing Officers (CMOs) are increasingly being asked to report results in financial terms. Unfortunately, compared to their manufacturing, operations and financial counterparts, they have traditionally been ill-equipped for this task. This might be one reason CMOs are so notoriously short-lived at most companies.

This lack of sound financial reporting is a shame but understandable. Go-to-market is comprised of, definitionally, the externally facing components of the enterprise. This means that its data scope is in a permanent dynamic state that cannot be controlled—at least not easily.

The GTMDL is naturally suited to solve this problem. While more advanced use cases are discussed more frequently, providing a "Go-to-Market Income Statement" might be the most important.

Although accounting systems can track go-to-market spending at the account level, they do not track operational metrics. CMOs need to know what was spent to quantify impressions for a particular segment in a specific time frame. One might even coin a shadow set of Generally Accepted Accounting Practices (GAAP) principles for go-to-market accounting, including:

1. Spending should be tracked when it was delivered in-market instead of when the agency or publisher invoiced.
2. Impressions and spend should be tracked for each channel, and used interchangeably.
3. Orders and their dollar value should be tracked for each last-touch channel.
4. The taxonomy by which spend, impressions and sales are tracked should be stable over time and reflect the decisions that can be made.

Knowing what was spent by channel and being able to interchange impressions (the work that was done) and cost (the efficiency of the media) in a reliable and timely way, by the categories of analysis the business uses to make decisions, would be a difference maker for most CMOs.

It's worth noting that we are not talking about a dashboard. Dashboards are very popular with marketers because they are visual and highlight anomalies, but you won't find many CFOs looking at dashboards. Instead, CFOs look at tabular financial statements because they understand that there is much more information in numbers than in visuals. Dashboards come after the income statement, balance sheet and statement of cash flows are complete and organized.

For the geography United States For the segment over 65 For the objective Acquisition	Display	Paid social	Streaming video	Affiliate	Branded paid search	Non-branded paid search	Total (all channels)
Spend	$4,500,015	$5,519,095	$3,415,910	$51,893,100	$31,015,912	$5,010,015	$101,354,047
Impressions (000)	187,501	128,351	31,629	114,051	221,542	23,857	706,931
Cost per thousand (CPM)	$24	$43	$108	$455	$140	$210	$143
Last-Touch (LT) orders	3,358	3,064	1,141	471,755	236,763	15,955	732,037
Average order size	$410	$451	$441	$378	$401	$512	$389
LT revenue driven	$1,376,870	$1,382,072	$502,977	$178,323,562	$94,941,837	$8,169,196	$284,696,515
LT Cost per Acquisition (CPA)	$1,340	$1,801	$2,995	$110	$131	$314	$138
LT Return on Ad Spend (ROAS)	31%	25%	15%	344%	306%	163%	281%
Fractional Attribution (FA) orders	14,091	18,912	10,891	443,839	220,411	23,893	732,037
Revenue driven*	$5,377,510	$8,529,312	$4,802,931	$167,771,142	$88,384,811	$12,233,216	$287,498,722
FA CPA	$319	$292	$314	$117	$141	$210	$138
FA ROAS	128%	155%	141%	323%	285%	244%	284%

*Because order sizes differ by last touch channels, total revenue driven via fractional attribution will not match the last-touch method.

Figure 3: This sample marketing income statement calculates last-touch and MTA-generated fractional attribution CPA and ROAS. The time period, geography, segment, and marketing objective are clearly defined at the top left.

AI and machine learning

Marketers have been using machine learning (ML) for longer than anyone. Early direct marketers knew they could increase their return on investment by only mailing to those customers who were likely to respond. They did this by building early machine learning models by hand or using mainframe computers—using techniques like logistic regression. Today, machine learning tools are many orders of magnitude more powerful and sophisticated than in the 1970s. Still, the basic idea remains: do more with less and find patterns to make marketing more effective.

In 2024, generative artificial intelligence (AI) is still in the midst of an early hype cycle. It is unclear how this technology will evolve, but many marketing use cases are already being tested. One early use case is generating text and imagery to improve advertising performance. Other more advanced ideas include autonomous agents that are given a goal—for example, sell as many widgets as possible and then coordinate various "engines" (email, advertising, APIs, etc.) to achieve these goals.

Wherever AI goes, it will need training data. Large language models (LLMs) are good at training on corpora of text, but future use cases will also focus on quantitative data. For quantitative data to be useful, it needs descriptive and hierarchical organization; in other words, the large language model will need to "know what it's looking at." Categorizing data using a common language and keeping it in one place will give AI-enabled companies an edge over data-poor competitors.

Propensity modeling

To predict who will respond to a touch—or who will convert to a sale—an analyst needs three things: "ones" (past customers who have responded), "zeros" (past customers who have not responded) and features (customer attributes that might provide a clue as to why they behaved in a certain way).

In most cases, organizations struggle with this kind of modeling not because of a lack of the "ones" but because of a lack of true "zeros." Said differently, almost all companies keep data on the people who became customers but might not track non-responders. A GTMDL should keep all promotions, whether individually tracked (discrete) or only known in aggregate (for example, the number of impressions in a DMA).

The features, or independent variables, linked to both stimulus and response are critical to machine learning exercises. In the case of direct marketing, we might know a lot about the people we are reaching. Enriching prospect data with third parties like Acxiom, Data Axle and LiveRamp can provide demographic, behavioral and psychographic data on who did and didn't respond to a campaign.

Segmentation and targeting

While propensity modeling seeks lift by only reaching the right prospects (prioritizing the doors to knock on), segmentation and targeting are about better-classifying prospects and customers so they can be spoken to in more customized, effective ways (showing the best messages and images when they open the door). The basic premise is the same: using known data about customers, accounts or prospects to cluster them into accurate, useful segments that can then be used for execution.

The GTMDL can support segmentation in three primary ways. First, many survey-based segmentation exercises fail because of a lack of assignability.[3] This means that the statistical clustering solution based on survey responses (which most likely does a good job of discriminating between key differences) cannot be accurately deployed to an organization's customers, prospects or accounts. This problem can be helped

immensely by sampling individuals or account contacts from the data lake whenever possible. This ensures we know each individual who completes a quantitative survey, dramatically reducing assignment error.

A second problem with segmentation is getting good signal to predict segment membership. Even if we have a one-to-one match between survey respondents and known individuals, the data required to predict who else will fit in each segment might be missing. Without good signal, segmentations become nothing more than academic exercises. Additionally, many segmentations rely on complex behaviors and psychological constructs.

The GTMDL can help negate this issue by providing a 360-degree view of customers (and, potentially, prospects).

Finally, segmentations fail because they are never truly deployed. By keeping segment membership up to date for each customer and prospect record, the GTMDL can act as an upstream single source of truth for segment membership, feeding CRM, marketing automation and customer support software. Concretely, views or APIs can be provisioned with real-time and historical segment membership for each customer, prospect and account. This can then be deployed in customer journeys, real-time ad serving or e-commerce customization.

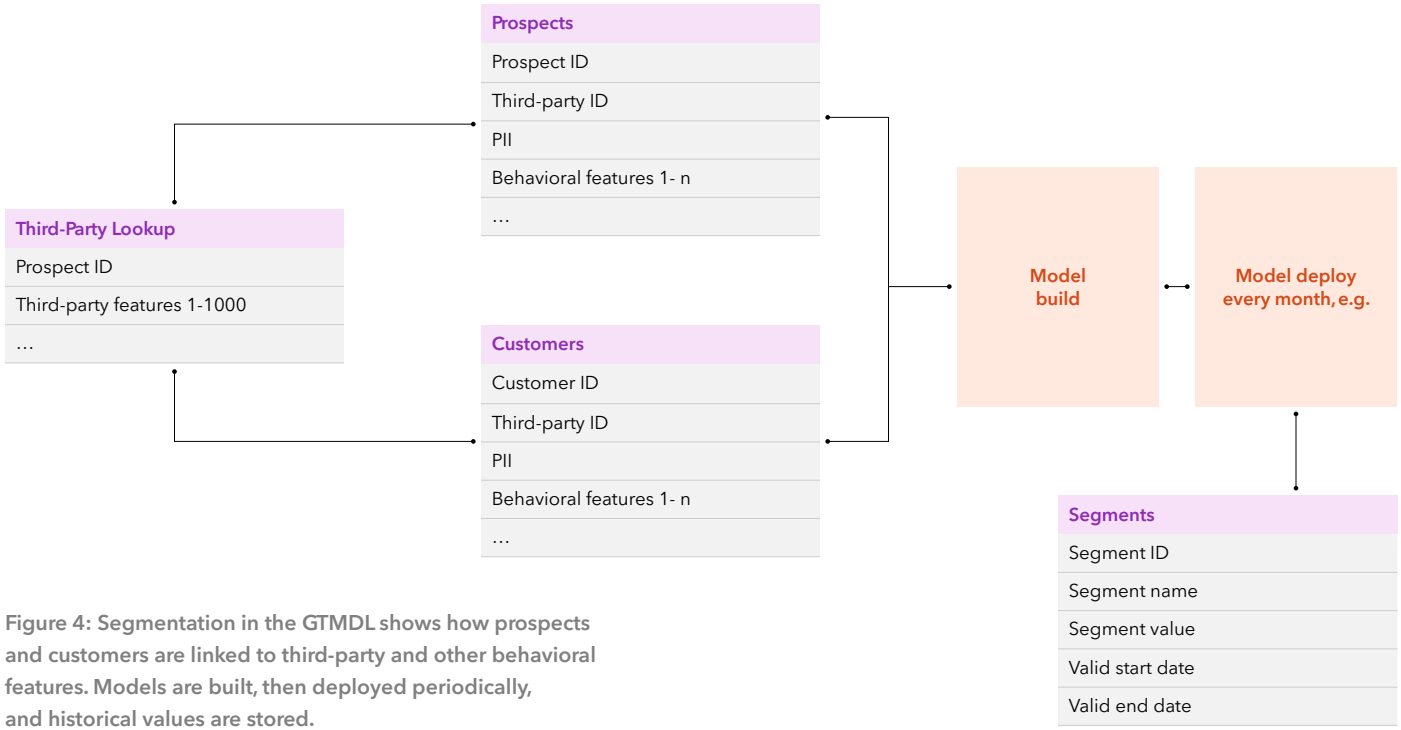


Figure 4: Segmentation in the GTMDL shows how prospects and customers are linked to third-party and other behavioral features. Models are built, then deployed periodically, and historical values are stored.

Generative AI

True go-to-market Generative AI use cases are still inchoate in 2024. Nevertheless, we can predict the kind of data LLMs will use to train on GTM data.

Text data, such as the specific language used in an email, display ad or sales and customer service representatives' notes, is expensive to store but will be critical for future creative and messaging use cases. As of the writing of this paper, interesting emerging technologies, such as Pinecone, allow text storage via vector databases.

Even if text data cannot currently be stored at scale, it seems likely that Generative AI will be able to pick out customer segments that respond well to a specific price or offer or identify patterns that human analysts miss. However, the "garbage in, garbage out" mantra seems likely to apply still. AI won't be able to work miracles; if data aren't clean to begin with, AI will likely produce wrong answers. Quality assurance—which we'll talk about later in this paper—will continue to be critical.

AI-enabled querying

Finding the right data to answer a question is a labor-intensive task that requires knowledge of databases, tables and data governance. AI seems primed to streamline the process of data exploration dramatically. Telling AI to generate a query that will provide a specific data frame—when that AI has knowledge of an enterprise's database and table structure—seems likely to be a killer app.

True analysis

"Analytics" has become a catch-all phrase, meaning anything quantitative or computational—especially in marketing. "Analysis" means something more specific. It stems from the Greek words Ana (undoing) and Leuin (knot), literally meaning "untying a knot." In other words, start with something complex, simplify it and find its meaning. Thus, we term "true analysis" as those analytical tasks that require this kind of unstructured knot untying.

Customer value

Understanding and predicting customer lifetime value (CLV) is the yin to customer acquisition cost's (CAC) yang. However, it is not a simple exercise to truly measure customer value—and it's even harder to segment customers by value and then predict what a particular customer's value will be. The analytical techniques to measure CLV are out of the scope of this paper, but the data required should be available in the GTMDL.

First and foremost, each customer's acquisition date should be stored. This is obvious, but surprisingly, not often readily available. Next, it's essential to store a vector of a customer's payments, whether voluntary and episodic—for example, in a retail environment—or predictable and automatic in a subscription business.

Finally, if applicable, the precise date of a customer's cancellation should be stored. Generally, the enterprise gross margin percentage can be multiplied by a customer's payments to get to a "lifetime gross margin". These cash flows can then be discounted by the enterprise's weighted average cost of capital (WACC). More advanced approaches differentiate gross margin by product or solution.

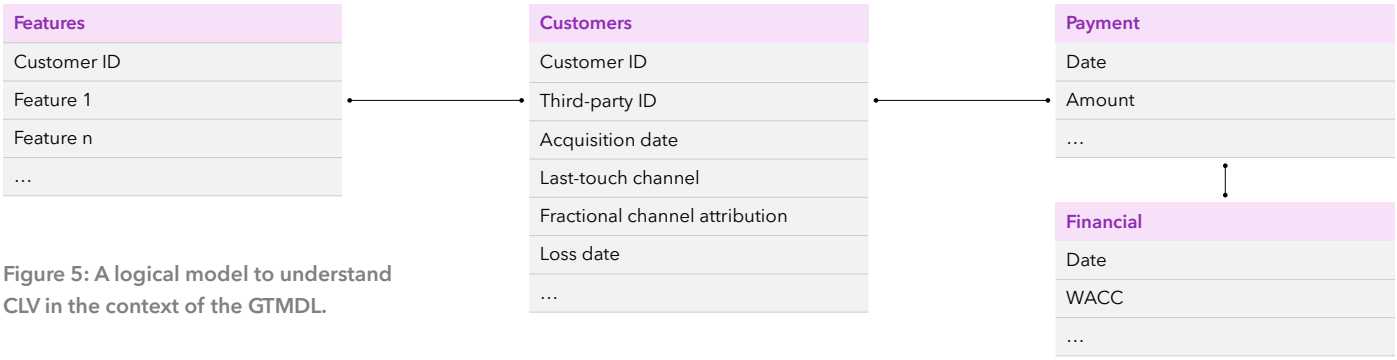


Figure 5: A logical model to understand CLV in the context of the GTMDL.

Campaign post-mortem

For marketing to improve, a clear-eyed view of what worked and didn't work in the past must be developed. This isn't MMM, per se—although econometric techniques might be used. Instead, an analytical view of "what happened and why" is a critical task that requires true analysis. To do this correctly, analysts need access to data at their fingertips—or, at the very least, available via query or API in a data science "workbench."

Too often, true analysis takes far too long because analysts are required to email, chat or otherwise hunt down data that should be in a centralized repository. While expecting every piece of data to live in a central place is unrealistic, the GTMDL should hold 80% of the data required to answer the most complex "data detective" type questions.

Activation

Go-to-market means "market-facing," and the GTMDL should thus be ready for this job by providing accurate and fast access to usable data at the customer, prospect and partner levels. The classic use case for the traditional CDP is the millisecond-fast-query ability to inform digital marketing (display, streaming video, etc.) for ad servers.

Beyond digital marketing, the GTMDL should be able to provide insights—whether segmentation, next logical product (or action) recommendations or demographics—to source systems used to communicate with customers. CRM and customer support systems should be pulled from the GTMDL.

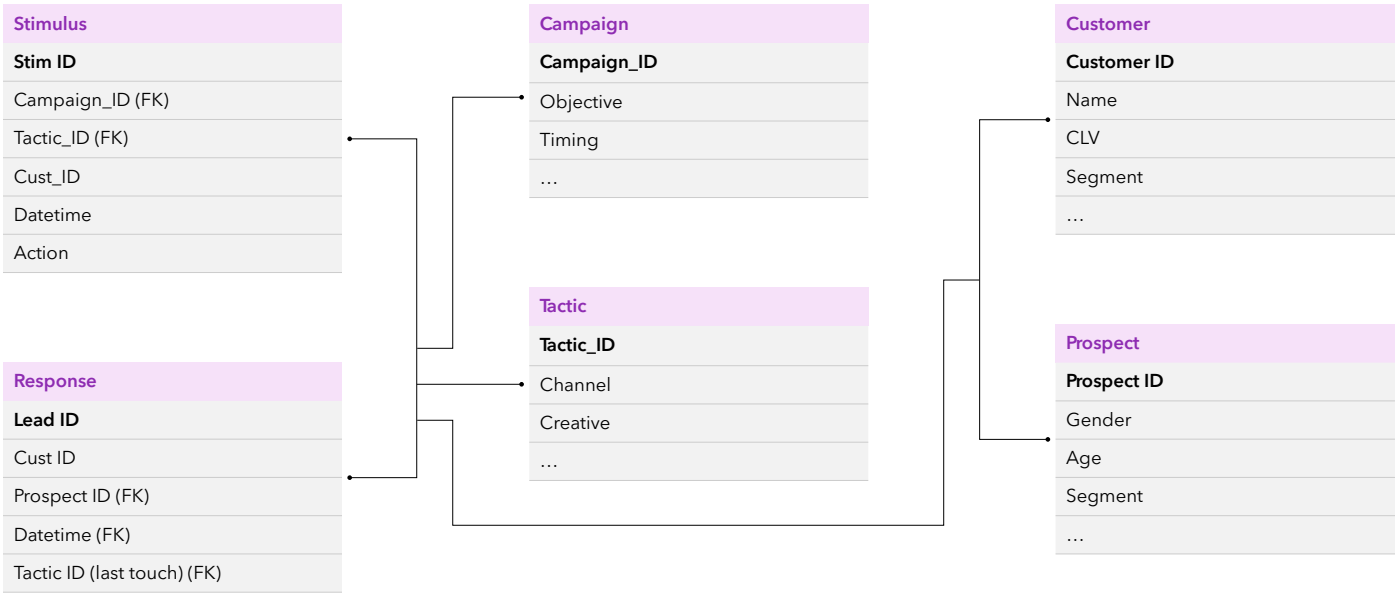


Figure 6: Example of data structure to support a campaign post-mortem analysis.

Personalization

The GTMDL should elevate personalization beyond traditional approaches by leveraging comprehensive customer insights to deliver highly targeted messaging. By integrating data across all touchpoints, marketers can understand customer preferences and behaviors at a granular level. This data-driven understanding allows for creating personalized experiences that resonate with the individual, driving engagement and conversion. Personalization powered by a GTMDL is not just about addressing the prospect or customer by name but about tailoring the content, timing and channel to match an individual's unique journey, enhancing relevance and fostering loyalty.

Beyond planning, specific prospect and customer data points should be served up quickly to marketing software to operationalize personalization. Digital marketing—display, text and even video—can be tailored to specific customer needs or segments.

Customer experience (CX)/Customer journeys

Once a prospect is a customer, the right cadence of customized messages can improve retention and customer value. Too many messages can fatigue a customer and make them tune out; conversely, very relevant messages that provide value at the right time can keep engagement high and keep a customer loyal.

The GTMDL can help improve CX by providing a holistic view of the customer journey, enabling seamless interactions across all touchpoints. By capturing and analyzing every interaction, moments of friction and opportunities for engagement can be identified, allowing brands to optimize the customer journey in real time.

Account-based marketing (ABM)

ABM is a strategic approach where marketing and sales teams collaborate to create highly personalized learning and buying experiences for identified target accounts and contacts. Unlike broad-based marketing strategies, ABM focuses on engaging specific organizations or accounts, tailoring marketing messages and campaigns to meet each account's unique needs and pain points. This approach aims to drive higher returns on investment by concentrating resources on accounts with the highest revenue potential, fostering deeper and more meaningful relationships with key accounts.

ABM strategies benefit immensely from the GTMDL's ability to integrate and analyze detailed account data. By leveraging a unified data pool, marketers can identify high-value accounts, understand their needs and decision-making processes and tailor marketing efforts accordingly. This targeted approach ensures that marketing and sales efforts are aligned and focused on accounts with the highest potential for conversion and growth. The GTMDL's real-time data capabilities enable dynamic adjustment of strategies based on account engagement, ensuring that ABM campaigns are always relevant, efficient and effective.

Design

While the benefits of a fast, comprehensive data store for marketing, sales and CX stimulus and response are readily apparent, these systems remain rare, even among Fortune 500 companies.

To be successful, the GTMDL should support each of the stories detailed in the marketing technology landscape above while avoiding the challenges that have plagued CDPs and marketing databases in the past. The four design principles of the GTMDL are robust data governance, a decoupled architecture, marketing science-first resourcing and a flexible approach to infrastructure.

Data governance

Data governance is the collection of practices, processes and standards implemented to manage and ensure the quality, availability, usability and security of an organization's data. It encompasses policies for data access, integration and control to maximize data's value while minimizing associated risks. Data governance for sales, marketing and customer data presents unique challenges around metadata, schemas, timeliness and third-party data.

Marketing data governance is difficult because marketing data are complex, differ dramatically from company to company and rely heavily on third-party vendors that resist standards. A GTMDL should thus have data governance processes "built in" versus decided upon after the fact. Data governance is a fundamentally human exercise, and a data governance leader should be assigned to the GTMDL instead of relying on a central IT team. Marketing and Sales are too verticalized and niche to be handled correctly by a corporate center of excellence.

Taxonomy and metadata

Effective data governance requires a well-defined taxonomy and metadata framework to ensure data across the organization is categorized and tagged consistently. Without a standardized taxonomy, data assets become isolated, leading to inefficiencies and inconsistencies in data retrieval and analysis—whether for analytical or production uses.

Go-to-market taxonomies are uniquely challenging because of a lack of standardization. There is no go-to-market equivalent to accounting's Generally Accepted Accounting Practices (GAAP) for marketing or the ISO standards for engineering. The "picklists" that define campaigns, accounts and customers are typically not standardized and do not line up across source systems or vendors.

Most companies have dimensions such as channels, industry verticals, publishers, marketing objectives, campaigns, tactics and journeys, but they are not standardized. Out-of-the-box dimensionality in martech and CRM systems is almost always customized to fit every business's unique attributes. Beyond system-level customization, campaign managers and sellers routinely change taxonomies for individual campaigns or long-tail use cases. In many cases, these changes never make it into core data systems, causing flawed analysis and poor decision-making downstream. It is common for data changes like these only to be discovered when a poorly QA'd PowerPoint deck is presented to someone who knows one of the golden numbers: for example, how many people really were targeted in a campaign or how many opportunities Joe Smith really sourced in 2024.

Although it can be difficult to make the case for a professional data librarian in EBITDA-obsessed companies, managing metadata for go-to-market is a full-time job. The utility of a Director of Marketing Metadata is crystal clear once written down; the marketing metadata person should be responsible for maintaining the lookup tables, dimensions, picklists, tags and "data about data" for all go-to-market activities at both the source system and the GTMDL levels.

This role must be both preventative and reactive. The preventative role involves working with software integration and campaign setup teams to ensure that picklists, tags and taxonomies are created with the organization's metadata principles and strategies in mind. Generally, principles of parsimony and temporal consistency should be applied.

The reactive role ensures that new taxonomies are accurately reflected and described in the GTMDL's lookup tables. For example, if a vendor campaign is executed with a disposition field that includes a new item—say, "hover over"—the job of the Metadata Director would be to account for that field, add it to the lookup table and ensure that downstream joins don't miss it. The Metadata Director thus maintains and adds to lookup tables based on what is happening in real time to make downstream systems, analysis and reporting functions.

Data lineage

Tracing data back to where it was born—its provenance—can be difficult. However, knowing a datum's original source is extremely helpful for determining accuracy and quality assurance. Take third-party customer data like that provided by Epsilon or KBM-Amerilink: of ~1,000 fields about a customer, there is a dramatic difference between those sourced behaviorally and those modeled. Behavioral data are far more accurate, and modeled data should be treated with suspicion.

Data lineage is a very specific form of metadata. In a table, each column represents a datum. For example, say each opportunity has its own row in the GTMDL. The columns (fields) of the table might be Opportunity ID, Opportunity Value, Opportunity Name and Opportunity Owner. A data lineage or provenance table will have a row for each of these columns, with its own fields describing where it came from. For example, Opportunity Value might be sourced from Salesforce.com, from the Opportunity table and inputted by the Sales Rep.

By documenting data lineage, provenance can be displayed even in front-end business intelligence layers. For example, a dashboard might have a hover over showing each field's source linked directly to the provenance table.

GTMDL

Customer table
Customer ID
Third-party ID
Acquisition date
Last-touch channel
Fractional channel attribution
Loss date
…

Data provenance

Database	Field	Original source system	Original source table	Original field name	Description
GTMDL	Customer ID	EDW	L_cust	Cust_id	Universal customer id
GTMDL	Third-party ID	Thirdbase	amerilink	Amerilink_id	Third-party ID
GTMDL	Acquisition date	SAP	bookings	born_on_date	First billing from customer
…	…	…	…	…	…

Figure 7: A data provenance metadata table to track the original source of truth.

Timeliness

Perhaps more than any other enterprise function, Marketing and Sales depend on timely data. Prospects and customers change quickly. Homeowners move, businesses merge, people change roles and jobs, and customers go from being loyal to at-risk all the time. Days or weeks of delay in data refreshes have real negative financial impacts on the business, manifesting as lost sales, defecting customers, bounced emails and inaccurate reporting, leading to poor management decisions.

Data schemas, which define the structure of data in databases, are subject to frequent changes due to rapidly changing campaign structures and new martech. Go-to-market data systems must continuously adapt to accommodate new data fields and formats without compromising data integrity or historical consistency. This dynamism requires agile data governance policies that can quickly respond to schema changes while ensuring that data remains accurate, consistent and accessible across the organization.

Marketing data itself can also become stale quickly. Delays in data processing, integration or accessibility matter; in direct marketing, the recency of a datum (for example, a new mover or a new lead) is often more important than its meaning. Ensuring data is timely requires robust data governance frameworks prioritizing data quality, data ingestion speed and real-time or near-real-time data availability.

Balancing the need for comprehensive data validation processes with the demand for swift data availability can be challenging, especially in fast-paced business environments. This tradeoff between recency and validation has partially driven the Medallion Architecture posited and implemented by Databricks.[4] In this concept, raw data are available instantly for needed analysis; semi-processed data can be used for more formal analysis, like predictive modeling, and fully keyed and validated data can be used for performance reporting.

Bronze	Silver	Gold
Raw files immediately available	Filtered, cleaned and augmented	Business-level aggregates
Useful for ad hoc, timely or very detailed one-off applications	Useful for machine learning and big data applications, like multi-touch attribution or journey mapping	Useful for BI and reporting
Minutes	Hours	Days

Figure 8: Databrick's Medallion Architecture formalizes the timeliness-accuracy tradeoff.

Vendor interface

Marketing and Sales rely heavily on external agencies and data providers, leading to the "long tail" of vendor data: a wide variety of data types and sources with varying degrees of quality and reliability, which often change from file to file. Managing this long tail poses significant data governance challenges, as organizations must establish standards and protocols for assessing, integrating and enforcing vendor data standards.

The vendor data universe can be described as a shallow but very wide lake. Each file might be simple on its own, but managing tens or hundreds of these third-party sources can be chaotic. While it might be tempting to integrate every third-party source with high quality and precision, a cost-benefit tradeoff should generally be done to determine a cutoff. At some point, the long tail should be left in a Medallion Architecture's "Bronze" section.

Marketing data are heavily reliant on third parties, whether data vendors (e.g., Epsilon), research vendors (e.g., Ipsos), publishers or platforms (e.g., Google Campaign 360) or agencies. The GTMDL should have clear rules governing data ingestion from third parties, balancing data quality, speed and administrative overhead.

Wherever possible, the GTMDL should use APIs to vendor systems. Vendors like Fivetran can automate much of this API layer, but open-source approaches are also possible and, in some cases, preferable. APIs are fast and accurate; because they contain structured data, it is unlikely that data changes (new fields, different naming, shifting data types or commas in text fields in comma-separated files) will break integrations.

Where APIs are unavailable, the GTMDL should impose validation requirements on vendors. Validation requirements can include standard file formats. Typically, a standard file format normalizes a campaign file, for example, into two tables: a fact table and usually one lookup table. The fact table will always have the same column names, for example, activity ID, campaign ID, datetime and tactic ID. These IDs correspond to the vendor-provided lookup tables, giving additional detail about tactics, campaigns and the like. The benefit of this is that these files can be stored in a data lake without significant manual transformation work.

Documentation

Clear, plain English descriptions of the data in an enterprise are often missing or incomplete. The GTMDL should have robust, "readme"-style documentation that is updated as data are updated. This type of documentation is common in data science but sadly lacking in the world of data.

The R package archetype is a good use case to follow. To be accepted to CRAN (the public repository for R packages), a package must pass several tests, including:

1. Each package must have a title and concise description.
2. Each function's arguments and outputs must be clearly defined.
3. Examples must be provided that compile correctly; these also function as built-in automated tests.

```
Title: Create Elegant Data Visualizations Using the “Grammar of Graphics”

Description: A system for ‘declaratively’ creating graphics, based on
“The Grammar of Graphics”. You provide the data, tell ‘ggplot2’ how to
map variables to aesthetics, what graphical primitives to use, and it
takes care of the details.
```

Figure 9: The title and description for ggplot2, a popular R package.

Documentation is an “always-on” process. While data dictionaries were a standard business document in the early days of relational databases, they have fallen out of favor as agile approaches have taken over. Data fields and tables can be documented in a “wiki-like” format, but an even better approach is linking each data artifact—whether “bronze” Json files or “gold” tables with schemas—with its own readme file. This can be done elegantly using git or another version-control software.[5]

Decoupling

Marketing technology has reached the point of fatigue. Operations teams are being forced to prioritize and consolidate—while simultaneously being asked to increase speed to market and responsiveness to customer behavior. Marketers have jumped from one technology trend to another for the past decade in search of a silver bullet. The result is often more operational overhead and diluted results for the business. Organizations that buy more software hoping to fix data problems realize sooner or later that rather than making things better, “software palooza” makes things worse.

A better approach is a decoupled foundational data asset capable of layering data from all business parts while supporting third parties. This is not an appealing problem to tackle and may seem impossible to Marketing leaders—but it is well worth pursuing.

Systems architecture principles suggest keeping a system decoupled is appropriate under three circumstances:

1. There are many dependencies on the component.
2. The component is critical.
3. The component is expected to change over time.

By decoupling the GTMDL from other key systems—such as marketing technology software—enterprises can ensure data integrity, enhance system flexibility and speed up the innovation cycle, all while reducing the risks associated with data silos and system dependencies. This approach streamlines operations and empowers organizations to adapt to market changes more swiftly and effectively.

Avoid software solutions

The state of data infrastructure within the average enterprise is bleak. In 2023, the Winterberry Group found that over one third of respondents indicated that their data infrastructure was centralized and leveraged across the business.[6]

The big platform providers want this business—and for marketing and Customer Experience applications, they call it the customer data platform (CDP). There is something compelling about giving Adobe or Salesforce access to the entirety of your prospect and customer data universe—but the downsides and potential future risks are real.

The first "big data" marketing systems were the Data Management Platforms (DMPs), which arose in the mid-2000s. DMPs focused on aggregating third-party data at a customer level, the original holy grail of digital acquisition. Early DMP pioneers like BlueKai built powerful targeting and segmentation capabilities tied directly into ad networks. However, these solutions were black-box and fell out of favor as the industry pivoted toward consumer privacy and away from third-party data.

Unlike DMPs, CDPs focus on first-party data—the data a company owns about its leads and customers. CDPs began to take root in the early 2010s as the number of online marketing channels exploded (social, mobile, search, video, connected TV and many more), making cross-channel analytics and customization a new use case.

Martech platforms had been playing in the customer activation space for years, so it was an effortless pivot to leverage the explosion of data that these platforms generated. Early CDPs looked more like bolt-on data modules but have now evolved into the central hubs of modern CX-focused platforms.

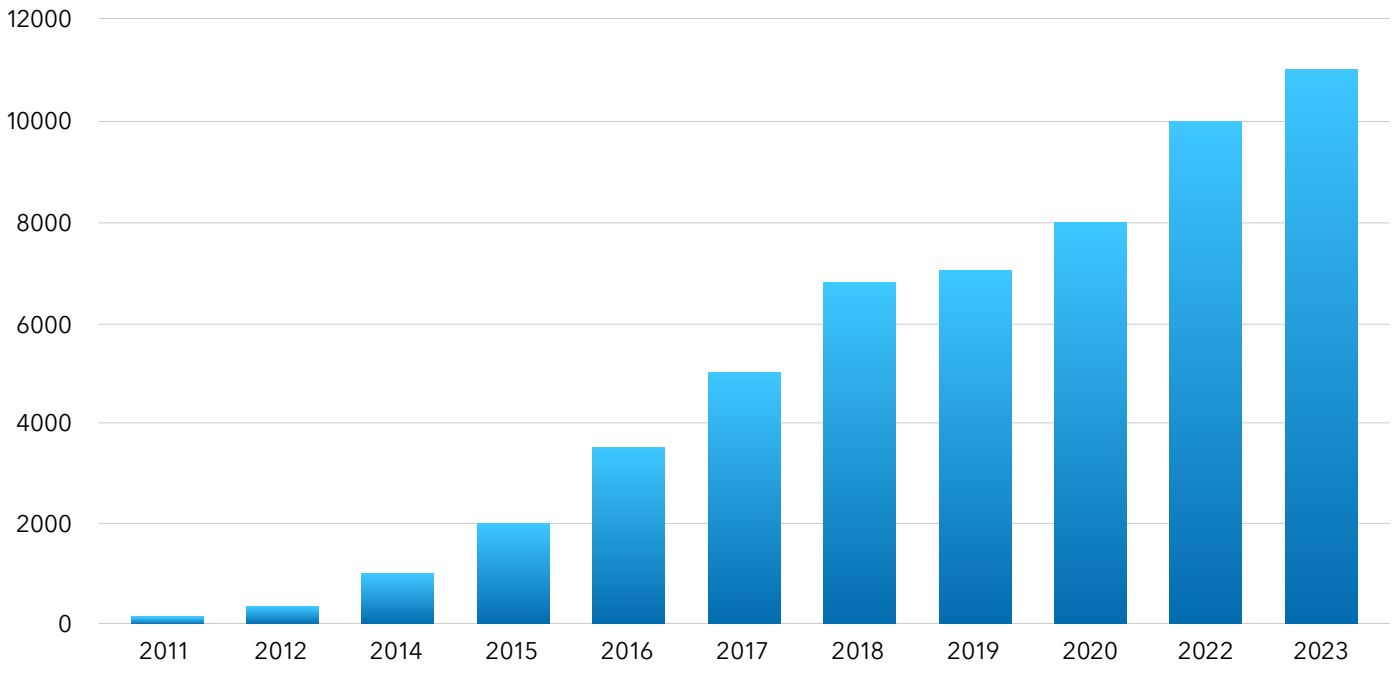


Figure 10: The martech platform's approximate growth from 2011 to 2023, according to Scott Brinker.[7]

However, there are two overriding problems with embedded software CDPs. First, due to their origins, software-based CDPs are primarily geared toward activation use cases, such as empowering customer journeys. For instance, when a customer visits a website, the CDP can be queried in real time and deliver content tailored to the customer's interests. Details of that interaction, in turn, feed back into the CDP and drive personalized follow-up messages in an email or digital ad.

While beneficial for customer engagement, this focus on real-time activation comes at the expense of analytical and data science use cases. Software-based CDPs lack the robust data processing and analytical capabilities required for deep customer insights and predictive modeling. Their architecture, optimized for quick data retrieval and interaction personalization, falls short in handling complex data analysis that requires processing large datasets over time. Consequently, businesses relying solely on these platforms for customer data analysis are hindered by limited data exploration capabilities and insufficient support for advanced data science initiatives. This gap constrains the potential for strategic decision-making based on customer data and impedes the organization's ability to innovate and adapt to changing market dynamics. As enterprises strive to become more data-driven, the limitations of software-based CDPs in supporting comprehensive analytical and data science functions become increasingly pronounced.

The second significant problem with software-based CDPs is cost. First, the software is expensive, typically billed per month per user. Secondly, the cloud infrastructure required to run these CDPs effectively adds another layer of expense. This includes the costs for storage, computing power and data transfer, which can quickly escalate as data volumes grow. Finally, the most significant expense is integration and customization, which almost always requires the services of large system integration consultancies. These services can be expensive, not just in terms of initial setup but also for ongoing maintenance and updates. These three cost components—software, cloud infrastructure and integration—collectively contribute to the high total cost of ownership of software-based CDPs, challenging organizations to justify the investment against the value derived. For a typical $10B enterprise, a CDP can be expected to cost between $20 and $40M in the first three years.

Past- and future-proof

According to Chiefmartec.com, the number of Martech platforms has grown from 150 to 11,038 over the past decade. This exponential growth has been a self-fulfilling prophecy, as the industry has a knack for creating solutions in search of problems rather than building based on real-life use cases. All these tools leave a unique data footprint, making the job of Martech Operations increasingly difficult.

The big platforms (Adobe and Salesforce) have been on a decade-long buying spree, integrating (and sometimes shutting down) niche tools across the industry, which has somewhat tamed the tech sprawl. However, this has been a slow process, and customers still often find themselves frustrated by duplication of capabilities and a lack of real integration between platforms.

Even with consolidation, the current push to deliver marketing through more channels, with hyper-personalized content, all while orchestrating sophisticated customer journeys, is creating a new generation of innovation, creating more data and operational breakage. For example, current journey-based software is great for one-to-one marketing but can fall down when attempting simple campaigns. These edge cases also make it impossible to abandon old platforms.

The GTMDL should function as a strategic pivot for Marketing organizations aiming to not only navigate the complexities of today's martech landscape but also to future-proof their operations against the relentless pace of innovation. By centralizing disparate data sources into a cohesive, accessible repository, a loosely coupled GTMDL enables Marketing teams to harness historical insights and anticipate future trends while ensuring their strategies remain relevant and resilient.

This approach mitigates the risk of operational breakages by providing a unified view of customer data. It also empowers organizations to leverage the full potential of their martech investments—past, present and future. In doing so, a GTMDL addresses the immediate pain points of tech sprawl and integration challenges and lays a foundation for sustained agility and growth in an era where adaptability is not just an advantage but a necessity.

Identity resolution

Identity resolution unifies customer data across various channels by leveraging diverse identifiers such as email addresses, phone numbers, cookie IDs, IP addresses, names and mailing addresses. Commercial identity resolution solutions use fuzzy logic matching algorithms—often proprietary—to link these disparate data points automatically. However, when dealing with complex or "messy" data sources, the predefined matching rules of commercial solutions might fall short. In such cases, custom logic becomes essential to accurately integrate data, although this can pose significant challenges in data management and analysis.

Incorporating identity resolution as a service within the GTMDL can significantly enhance the platform's capability to create a unified customer view. This approach allows for integrating diverse customer identifiers—including those from third parties like LiveRamp and Data Axle— enabling a more comprehensive customer engagement strategy.

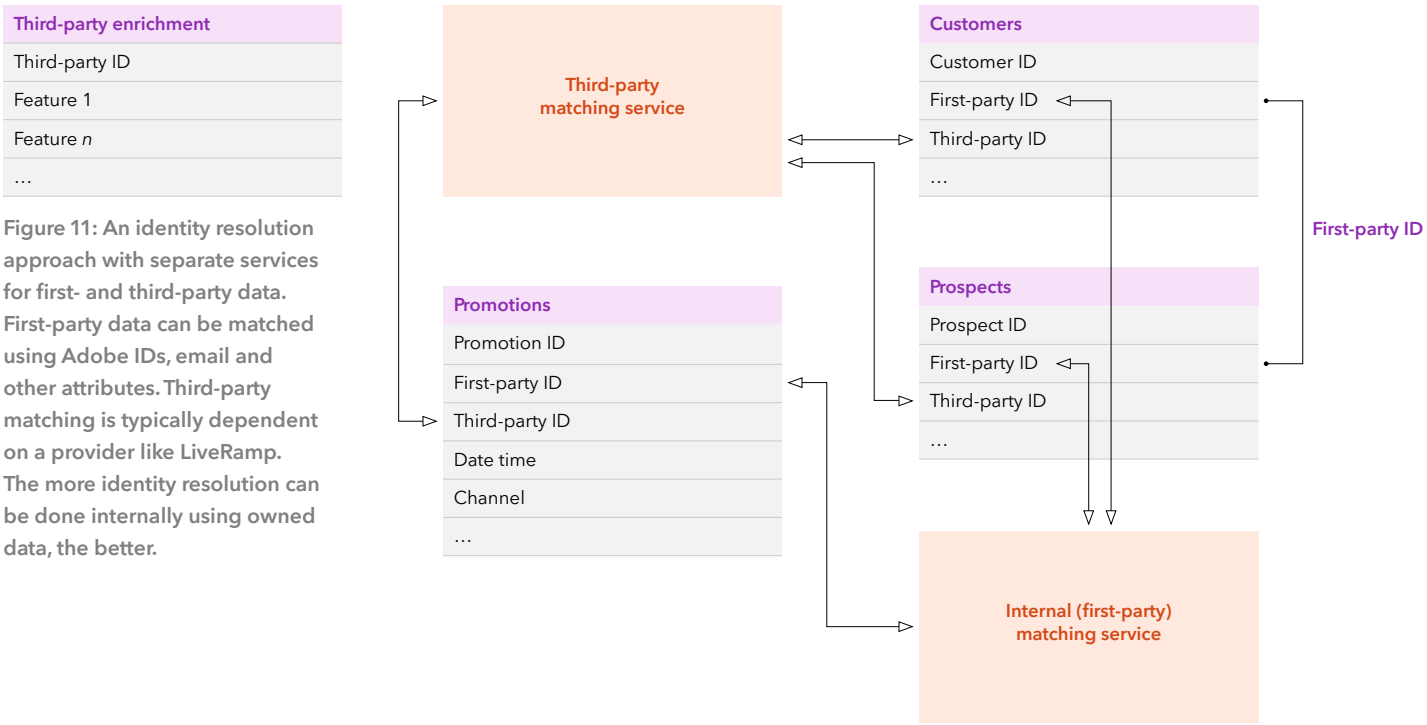


Figure 11: An identity resolution approach with separate services for first- and third-party data. First-party data can be matched using Adobe IDs, email and other attributes. Third-party matching is typically dependent on a provider like LiveRamp. The more identity resolution can be done internally using owned data, the better.

Real-time ingestion

Big marketing platforms are interested in keeping their data in-house for obvious reasons. Advertising platforms like CM360 (Google) and Meta are moving to keep data in-house, ostensibly for privacy reasons, but likely mainly to maintain control. Keeping as much of your data yourself, independent of platforms—whether publishers or software—should be a key goal for a marketer.

Today, there are many pre-built connectors in popular extraction, transformation and loading (ETL) tools like Fivetran to enable easy data ingestion from popular marketing systems like campaign management platforms, email systems, ad networks and web analytics systems that can be layered into the GTMDL.[8] However, marketing data always tends to be messy, requiring more bespoke cleaning and transformation. Messy data types include above-the-line advertising, sales, servicing, client transactions and accounting data—all critical for a complete marketing data solution.

It is critical to ingest data quickly and efficiently into a GTMDL. "Easy input-output" facilitates seamless data integration from various sources, including campaign management platforms, email systems, ad networks and web analytics systems. Such flexibility ensures that the most comprehensive and up-to-date data inform marketing strategies.

Resourcing

Marketing teams often struggle to realize value from corporate IT and centralized data teams. IT usually provisions and administrates platforms but does not provide operational support. This means the marketing team must figure out integrations independently but are rarely equipped to do this. There has been an increasing need for technical resources to dwell within Marketing's control; however, in most large companies, the C-suite has not yet accepted the proposition of building and managing technical teams inside functional areas. As a result, Marketing leaders tend to rely on outside vendors—whose incentives are usually aligned with providing services versus building assets—to fill this gap.

Because of this history, go-to-market leaders are more willing to spend on new capitalizable software platforms than resources and services. While this might sound good (assets can be used for years), adding new platforms with each generation of leadership adds operational resource needs, stretching already thin budgets. The result is more technology to operate with the same number of people to support it. Many technology platforms never go live because of this challenge, creating frustration that drives the next software purchase. This vicious cycle has been repeating for decades at many large enterprises.

Concurrently, centralized data teams have become dominant in most corporations. This is to be expected; centers of excellence emerge when the skills required are rare. In the 1980s, spreadsheet skills were still unusual, so almost all financial analysis happened in the finance department. As spreadsheet skills became ubiquitous, financial analysis was democratized; while 10-Ks are still produced by finance, customer lifetime value, breakeven and return on investment analyses happen in manufacturing, R&D and Sales and Marketing departments.

However, database skills are now becoming more common in many companies. Although transactional databases are still the realm of specialists, analytical databases are increasingly being built, modified and queried by businesspeople. The time has come for go-to-market data to be owned by practitioners: Marketing-native technologists.

Marketing-native technologists

In our 2022 paper on building the next-generation marketing analytics department, we posited five essential skill sets that marketers will need over the decade of the 2020s: expertise in martech systems, data science, marketing as a discipline, storytelling and data engineering (see Figure 12 below).

A marketing-native technologist is "more technical" across source (martech) systems, data engineering and data science but also intuitively understands marketing and sales use cases. For example, a marketing technologist might understand Adobe data structures and APIs intimately, as well as what they mean in-market, have excellent Python data munging skills, and be able to check code in on an enterprise GitHub instance. At the same time, this person would know the campaigns being run in-market and how upper-, mid- and lower-funnel tactics typically work together to drive deal flow.

All the members of the GTMDL should be marketing-native technologists.

More technical

Source systems	Data engineering	Data science	Marketing	Communication/ visualization
Understanding of marketing use cases	Data architecture	Reproducible artifact building (Rmarkdown, Notebooks)	Segmentation and targeting (quant, qual research and assignment)	Narrative and storytelling (PowerPoint)
Focus on organization, taxonomy and hierarchy	Detailed understanding of data structures (tabular, JSON, XML, etc.)	Statistics and machine learning (Python, R)	Performance marketing tuning	GUI-based dashboarding (Tableau, PowerBI)
Understanding of data structures	Data transformation (Python, SQL)	Exploratory analysis (Python, R, specifically Pandas, dplyr)	Test construction and reading	Programmatic visualization (D3, Shiny, Quarto)
File export and API configuration	Scheduling and conditional data pipelines (Airflow)	Version control and collaboration (Git, Github)	Multi-channel strategy (leveraging MMM, MTA)	

Figure 12: The five critical capabilities for marketers in the 2020s.

Team structure

A VP/Director with a "product manager" archetype should lead the GTMDL team. This person should spend most of their time interacting with Marketing and Sales leaders and other parts of the organization, ensuring that use cases are well documented and prioritized.

A formal data architect should administer the GTMDL, typically with a dotted-line report into corporate IT. This individual should be responsible for data governance and integrity and have the highest-level administrative rights over the GTMDL. They should be the only individuals who can delete rows, for example.

The day-to-day team should consist of product/project managers, data engineers and analysts, in a ratio of roughly 1:3:2. In other words, a small GTMDL team would have one VP—or Director-level lead, a lead Architect, a Product Manager, three Developers (or Data Engineers) and two Analysts.

Analysts could sit on either the data team or an analytics/data science team. In either case, it's essential that analysts—who will create the reports and dashboards based on the data—deeply understand underlying data structures. Siloing between the data and analytics team is a common cause of inaccurate or unusable business intelligence.

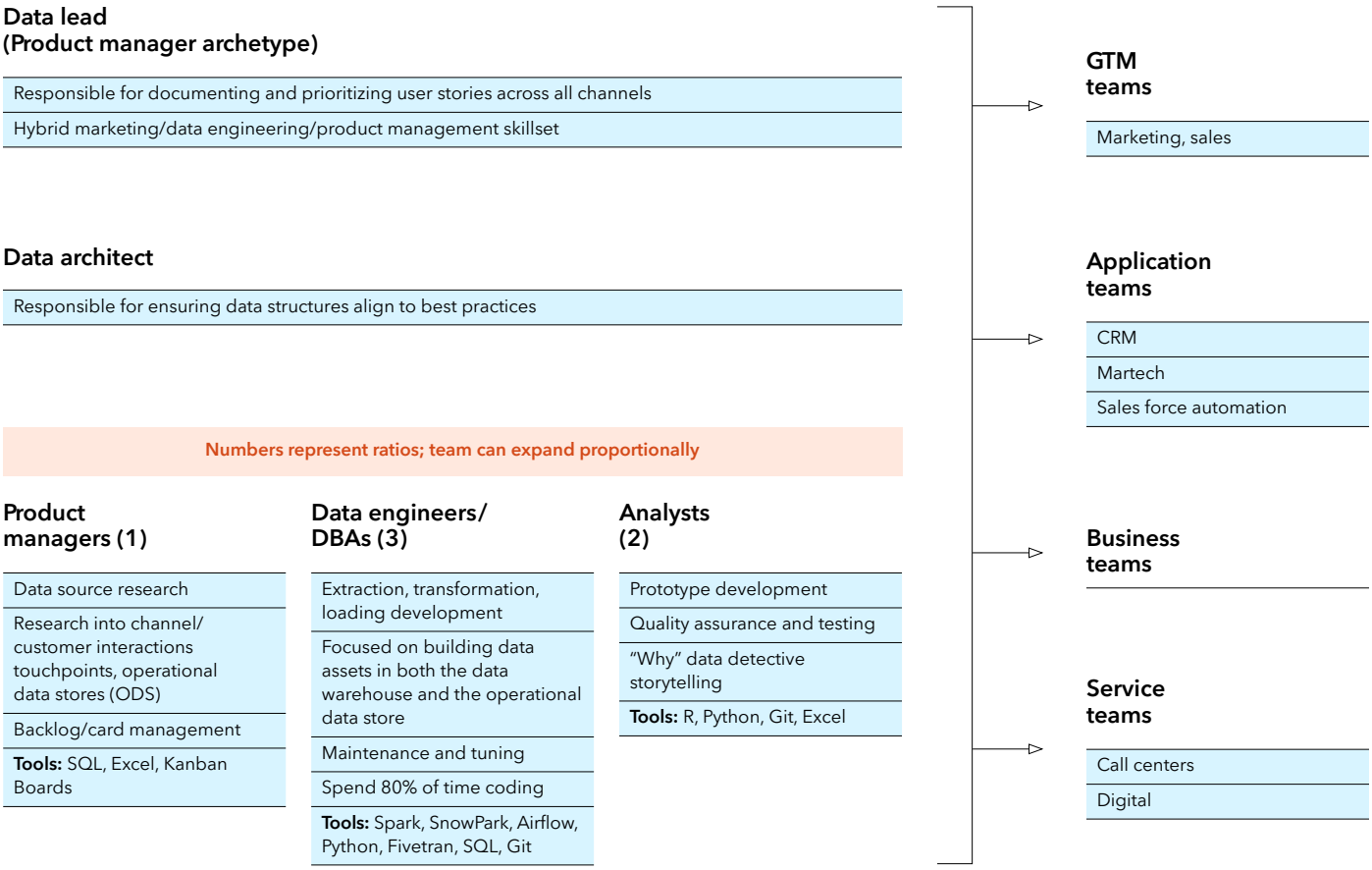


Figure 13: A modern Marketing and Sales data team. Each member should have a deep understanding of go-to-market use cases.

Federated access

The gap between IT and business is not a new dynamic; business leaders have spoken a different language from their technology counterparts since the 1970s. However, martech operations have become particularly complex in recent years—creating a uniquely large knowledge gap with IT.

Marketing leaders struggle to find technical subject matter experts (SMEs) who understand marketing use cases. Additionally, IT organizations are driven to standardize and reduce technology sprawl by leveraging existing platforms. This results in IT presenting shoehorned solutions that don't support marketing use cases well. This is usually discovered late in the project after the solution has been built and the investment dollars have been spent.

The theme of marketing ownership of the data should continue to encompass data usage and availability. While security issues are important, providing open access to information can superpower go-to-market activities.

Data access should be provided to both people and machines. Power users—typically analysts and data scientists—should have read access to as much data as possible. An analyst seeking to create a dataframe displaying all customer touchpoints over a specific timeframe—with joins to information about channel taxonomy—should have no issues querying and finding data. While this may sound basic, it is surprisingly difficult in many organizations because privacy and security concerns override usability.

Machines should have access to the data, too. Marketing and sales technology should have easy access to the GTMDL via APIs. Endpoints can be provided for almost anything, including customer or lead scores, personalization data, account history, channel attribution or demographic information. Data engineers continue to work on creating APIs for software to allow easy access views for analysts.

Flexible infrastructure

Cloud vs. on-premises

Cloud platforms have invested billions in recent years and are superior in many ways to their on-premises counterparts. Industries that were apprehensive about moving to the cloud are now embracing it. The benefits can't be ignored, and companies that don't adapt will be left behind since there are clear benefits with cost, capability and speed to market.

Cloud is the default choice when launching new tech; however, some on-premises footprints are challenging to migrate or are poor candidates for the cloud. Legacy applications designed with a monolithic architecture are often bad candidates because their operating system environments are inefficient at allocating physical hardware resources. Systems designed to leverage cloud microservices employ granular control of every underlying component and are much more cost effective.

For instance, in a monolithic system, an application that runs on a single server must be allocated enough storage, memory and computing to handle the peak loads of the application. Most business applications have small peak windows, which is a better match for cloud-native microservices. Conversely, companies that rush into an all-cloud strategy often increase IT costs and regret their decision to move a monolith to the cloud.

Regardless of the cloud vs. on-premises decision, enterprises should strive to avoid vendor lock-in. A GTMDL should provide the option to run on any major cloud host (Azure, AWS or Google) and yield operational and cost synergies with existing cloud agreements.

Managing costs should be linear, metered and flexible as workload demands fluctuate. Modern systems offer a metered approach (down to the second) for CPU runtime and a per TB charge for storage. This eliminates the need to worry about managing platform edition tier sizes, the number of contacts under management and other abstractions that software platforms impose on data systems. Additionally, the ability to pause workloads and resize compute nodes can save a lot of money.

Distributed compute and storage

The separation of compute from storage is a significant achievement in modern big data platforms. In traditional relational database management systems (RDBMS) systems like SQL and Oracle, storage and compute were combined within a single physical or virtual server, and scaling required adding more servers and partitioning sections of the database across different servers. This forced data access layers to be rearchitected. Modern big data systems solved this problem by designing the compute and storage layer as separate microservices that run on independent servers and rely on a transparent orchestration engine to distribute the workloads. This allows compute to be more precisely managed and scaled horizontally across multiple physical servers while providing the façade of working with a single logical server. A modern, scalable data architecture should use a technology stack with this capability.

The sheer volume of data generated by digital channels necessitates a robust and scalable compute infrastructure capable of handling extensive datasets over prolonged periods. This capability is vital for inference, which involves applying statistical and mathematical models to large-scale datasets to provide insights, measurement and optimization.

As we stand on the cusp of the AI revolution, the importance of scalable compute resources is becoming even more pronounced. The ability to train AI and machine learning models over large, frequently updated datasets will be a cornerstone of competitive advantage. This requires a powerful and adaptable GTMDL infrastructure that can scale compute resources dynamically to meet the demands of increasingly complex analytical workloads.

Flexibility in scaling is essential: Compute nodes must be capable of scaling up to leverage larger compute capacities for heavy-duty processing and scaling down (or suspending altogether) during periods of low demand to optimize costs. This scalability ensures that marketing organizations can handle peak workloads efficiently without incurring unnecessary expenses during quieter periods. The cost structure for these resources should be predictable and linear, akin to a utility bill, allowing for precise budgeting and preventing unexpected expenditures due to runaway workloads.

Moreover, the GTMDL should offer a variety of compute node sizes and configurations to cater to a diverse range of functions. These functions include data pipelines that require continuous processing power, ad-hoc analytical projects that may need burstable compute resources and AI/ML use cases that demand high-performance computing capabilities. The ability to set budgets and thresholds for compute usage is also crucial, providing organizations with control over their data processing expenditures while ensuring that their analytical capabilities remain uninterrupted.

In summary, the compute strategy for a GTMDL must be designed with both capacity and flexibility in mind. It should support the varied and evolving needs of digital marketing analytics, from basic data storage and processing to advanced AI-driven insights, all while maintaining cost efficiency and operational control. This approach not only future-proofs the Marketing organization's analytical capabilities but also ensures that it can continue to innovate and compete in an increasingly data-driven landscape.

Open connections

Open connectivity support for popular protocols, including API, Streaming, SFTP, Block Storage, ODBC and mainstream platform connectors, is essential. These connections must be easily accessible to both on-premises corporate data stores and third-party vendors and platforms. Built-in network policies provide robust security controls to restrict traffic at an IP level without the need for a traditional corporate firewall. The connections must support a variety of secure authentication methods, including MFA, SSO, OAUTH, key-based authentication and API keys.

Owned, open code base

Owning the code provides the customer with maximum control. Flexibility and customization are essential to support the unique data footprints within each organization. Packaged solutions are built around generalized patterns and don't always translate well within an organization. Custom entity support is available but often cumbersome to build and manage, resulting in awkward data access.

A custom solution also provides more flexibility and efficiency to manage costs over the long run. Platform projects often fail after implementation because the ongoing technical operation requirements are overlooked or underbudgeted. Managers believe they can rely on existing staff to take on the operational responsibilities and underestimate the technical skills and time needed to do so. The reality is that modern platforms don't simply run themselves; they require time from skilled resources to operate and administer. Multi-year contracts lock customers in regardless of whether the project gets off the ground.

Security

Data security and governance continue to be priorities and challenges for the enterprise. While Marketing should look to increase operational autonomy, security and compliance are essential areas that must be closely coordinated with IT security teams. Marketing data systems contain personal customer data subject to compliance frameworks, consumer privacy regulations and security audits.

Encryption at rest and in transit is now a commodity feature in most modern business systems, but additional encryption measures are often required when handling more sensitive data elements like HIPAA data. Security teams demand robust security controls such as AES-256 encryption, key rotation, SSO (single sign-on) and audit trails. The ability to handle privacy delete requests from customers is also a common requirement, especially for organizations operating in the European Union with consumers protected under the General Data Protection Regulation (GDPR). In the future, it is a safe bet that privacy and security protocols will become more demanding across the board.

Security is especially important when dealing with cloud-based infrastructure. Most large enterprises have fully embraced the cloud due to advantages such as startup cost, capability, scalability and resiliency. However, security remains a sore spot after a brutal streak of consumer data leaks over the past several years. Fortunately, more customizable security technology in cloud-based database systems seems poised to make system-level data breaches much rarer.

Traditional relational database systems only provided access control down to the table level. In other words, the customer table might be locked down for everyone but "need-to-know" resources. However, many valuable use cases require sensitive data from many different tables. A column-level permission scheme allows a mix of specific roles with different column-level permission sets to exist on the same table without duplicating data.

Additionally, tokenization of data is now possible at scale. A modeler's view of data without any ability to identify individuals—typically obfuscating personally identifiable information (PII) using non-traceable IDs—allows rapid modeling and insight development without the danger of security breaches.

The GTMDL should accommodate access to numerous roles and purposes, including the following:

Data providers

- Third-party agencies that execute programs
- Automated feeds from channel systems
- Data enrichment vendors
- Internal teams that manage CRMs, financial systems, MDM and other business data

Data consumers

- Analytics teams
- Campaign execution teams
- Executive and management
- Compliance and Privacy teams

Administrative

- Development teams
- IT and Security teams

Additionally, these roles require varying connection protocols, including SFTP, HTTPS, File Sharing and SQL. Processes are also required to handle access requests and credential auditing.

Implementation

Development

A "white box" database development approach prizes transparency, internal control and documentation. In this approach, every aspect of the data environment—from ingestion and storage to analytics and reporting—is fully visible and understandable to the team. This approach empowers organizations with the knowledge to adapt and extend their systems.

The GTMDL's architecture must be both robust and adaptable, built with flexibility and reproducibility in mind. By leveraging open-source technologies and setting standards for documentation and coding practices, enterprises ensure that their data infrastructure can meet current analytical needs and evolve with future demands. This foresight into the implementation strategy safeguards the longevity of the GTMDL, making it an enduring asset in the technologically fluid landscape of marketing data.

Use case-driven requirements

A use case-driven approach to requirements and design prioritizes real business needs over technology for technology's sake and ensures that data lake development solves tangible business challenges. This approach optimizes resources, yields practical and impactful solutions, and aligns investments with business value. This can be contrasted to the more typical technology-first strategy, which often leads to over-engineering, solutions-seeking problems and user friction, resulting in long-term discontinuation of the platform.

The use cases should detail concrete data needs, how the data will be used, where it will be sourced and how it will be transformed. Starter use cases might include marketing measurement, customer segmentation, personalized marketing, sales optimization and customer journey mapping. These use cases should directly inform the data lake design, dictating the architecture, data management and analytical capabilities. This ensures adherence to data quality and governance standards.

Marketers, not technologists, should define the use cases. The specificity of campaign architectures, measurement algorithms and reporting standards provides powerful requirements for data architectures. Concretely, the Product Manager role defined in the Resourcing section above should be the one doing the defining.

Normative approach to taxonomy

A normative taxonomy is a standardized framework that defines and categorizes key elements such as picklists, segmentations and campaign names. This standardized approach ensures that all users, from marketers to analysts, speak the same language, facilitating clear communication and consistent analysis across the organization. Organizations can avoid the confusion and misalignment that often arise from decentralized data management by prescribing a standard set of terms and definitions.

This normative taxonomy enhances data quality and reliability and supports more accurate and actionable insights, enabling teams to make informed decisions quickly. It lays the foundation for a cohesive data

strategy, ensuring that the data lake is a unified, efficient and effective tool for driving go-to-market strategies and business outcomes. Normative means that the new taxonomy is as it should be.

This can be accomplished on both a look-back and a look-forward basis. Look-back means retrofitting existing taxonomies to a new standard via lookups and dictionaries. This is necessary for any extant marketing technologies or databases. Looking forward, any new marketing technology should adhere to the normative taxonomy. For example, a new campaign management system might have an out-of-the-box channel taxonomy that should then be modified to match what is currently in the promotions table of the GTMDL.

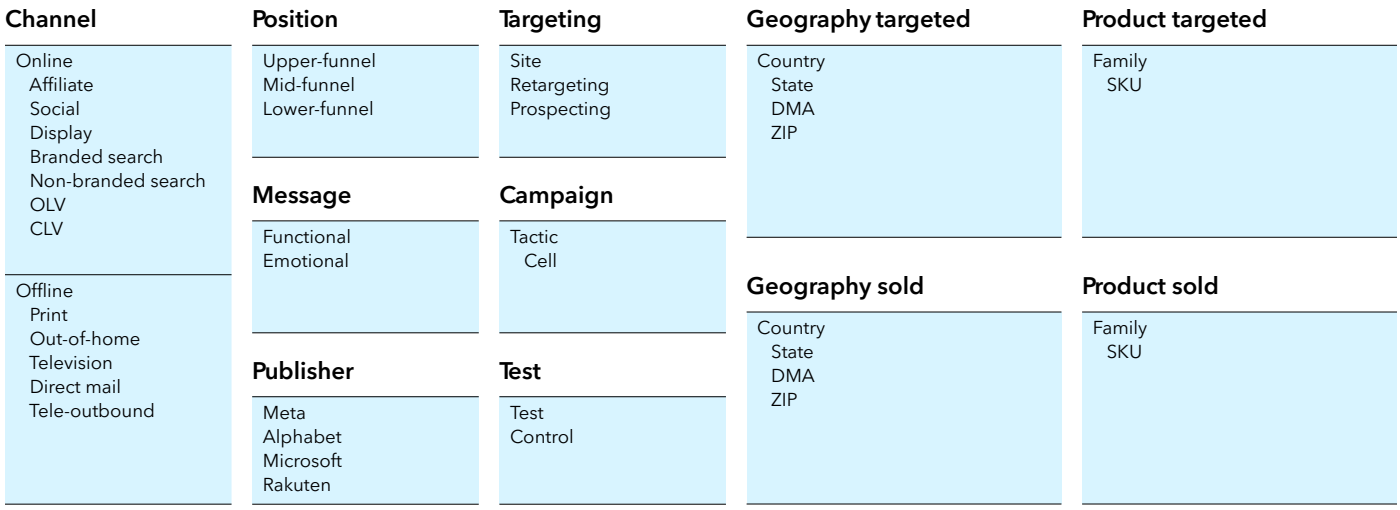


Figure 14: A simplified example of a normative taxonomy.

Federated approach

A federated development approach avoids cumbersome and often slow-moving processes traditionally associated with corporate IT projects. This approach leverages advancements in cloud data technologies to facilitate a federated model of development and operation between the vendor and the client. In this model, a collaborative environment is created where both the vendor and the client share development, balancing the need for rapid deployment with the stringent requirements of corporate IT security and governance.

This federated approach allows iterative development, testing and refinement parallel to the existing enterprise environment. This method reduces the time-to-value, enabling Marketing organizations to quickly capitalize on the benefits of a GTMDL without the protracted timelines typical of large-scale IT projects. The joint custody model also ensures that while the vendor brings their expertise in cloud technologies and innovative data solutions to the table, the client retains essential oversight and control, ensuring that the solution adheres to the enterprise's security, compliance and governance standards.

Selecting a vendor with a deep understanding of the modern security landscape is crucial in this context. The right vendor will have the technical capabilities to build and manage a GTMDL and a thorough grasp of contemporary security and privacy challenges. They should be adept at navigating the complexities of data protection regulations, ensuring that the GTMDL is not only efficient and powerful but also secure and compliant with all relevant laws and standards. This expertise is essential for minimizing risks and ensuring that the deployment of the GTMDL enhances the Marketing organization's capabilities without introducing vulnerabilities.

Whatever vendor is chosen, an onshore technical leadership team who remains close to the client and accountable from design through implementation is critical. Many firms are operationally siloed, with sales engineers not typically responsible for implementation after the sale. However, a single point of accountability from Sales through implementation is imperative for leaders who demand success with their technology investment.

Code-based

A team of hands-on developers shipping code—versus installing more software—is fundamental. The code should be client-owned to retain control and allow it to be brought in-house if desired. A version control system such as Git should also be used to track change history and ensure that the entire solution can be redeployed if necessary. The code should adhere to a set of standards that include formatting, indentation, naming conventions and a modular hierarchy design for libraries. Lastly, the code should be readable and clearly documented. These principles are a core part of DevOps, a framework of best practices for code-based development teams.

Maintaining and expanding

Once the initial solution build is complete, the focus should quickly pivot towards a maintain-and-expand mindset. A successful launch of the GTMDL is likely to have many applications, and the supporting team needs well-defined processes to ensure operational continuity—and, ultimately, expansion. An intake queue should be quickly established to ingest new requirements and defects. Automated monitoring of scheduled jobs should send alerts to the DevOps team to notify them of pipeline issues that need attention. A vendor system connection intake process should be established to streamline connecting new data sources to agencies, platforms or other business data.

Flexible operating model

Due to the cloud's flexibility, the GTMDL can be operated in three configurations: fully outsourced, collaborative or vendor-built and client-run. In many cases, it can start as fully outsourced, transition to collaborative as a client's internal team gains more comfort with the system and eventually evolve into fully client-run.

1. **Fully outsourced:** The vendor is fully responsible for building, enhancing and operating the solution from end to end. Oversight with security and IT governance is handled directly by the vendor to ensure all compliance requirements are upheld. This method has the lowest friction to the client.

2. **Collaborative:** The vendor team will be responsible for building core architecture and pipelines, and specialized analyst teams need workspaces for specific projects or functional areas. In this case, the vendor is still responsible for keeping the underlying architecture operating smoothly and securely, and the marketing technologists at the client can easily leverage the environment as needed.

3 **Vendor-built, client-run:** The vendor is responsible for the initial build and then transfers operations to the client. This is a viable option when the client wants to develop a data ops team within Marketing and further optimize cost. Because the development is performed in a joint cloud tenant with SSO, with open-source controlled code, there are no significant technical barriers to transferring ownership.

Snowflake, the popular cloud-based database company, specializes in providing multi-tenancy where two or more companies can work on one data substrate. In this model, storage is shared with one primary owner—in this case, the client company. The client and vendor would pay separately for compute and would have access to separate tables and columns within tables. Eventually, the client can take over completely.

Another benefit of this approach is the possibility of a "clean room." In this model, a client provides non-PII data to one or more third parties, who can then match to this data, or provide additional enrichment—without knowing too much about a client's prospects or customers. This type of environment can also be turned on or off as needed.

Conclusion

In the ever-evolving world of digital marketing, the GTMDL emerges as a beacon of innovation and adaptability. As this paper has articulated, the GTMDL transcends traditional CDPs by offering a holistic, agile framework capable of navigating and thriving amidst the complexities of modern marketing technology landscapes. Its design—focused on easy data ingestion, sophisticated analytical capabilities and seamless integration—ensures that Marketing organizations can leverage data to its fullest potential. This strategic asset empowers teams with the insights needed for personalized customer engagement, predictive analytics and a unified view across all marketing channels. By adopting a GTMDL, enterprises can position themselves at the forefront of data-driven marketing, ensuring resilience, agility and continued growth in a rapidly changing world.

References and endnotes

1. "Gartner's Customer Data Survey: The 360-Degree View of the Customer Is More Myth Than Reality," Gartner, November 30, 2021; https://www.gartner.com/en/documents/4008867

2. Hana Yoo, "CDPs are in the Gartner Hype Cycle's "Trough of Disillusionment," AdExchanger, February 22, 2024; https://www.adexchanger.com/data-exchanges/cdps-are-in-the-gartner-hype-cycles-trough-of-disillusionment/

3. Segmentation typically starts with a quantitative survey to poll both existing and prospective customers.

4. "What is a Medallion Architecture?", Databricks; https://www.databricks.com/glossary/medallion-architecture

5. In a data lake environment, the entire lake can be a git repository (by setting git init). By ignoring everything but .md (markdown files in your .gitignore file, only changes to readme files (or whatever you'd like to name them) will be tracked. This way, anyone can update data documentation, and that documentation will automatically "track" to the data lake "schema."

6. Bruce Biegel, et. al. "FROM DATA TO INSIGHT: THE OUTLOOK FOR MARKETING ANALYTICS," Winterberry Group, April 2023; https://winterberrygroup.com/marketing-analytics-outlook-whitepaper-2023

7. Scott Brinker. "Martech 2030 Trend #3:The Great App Explosion," chiefmartec, December 2020; https://chiefmartec.com/2020/12/martech-2030-trend-3-great-app-explosion/

8. Fivetran is actually an ELT (extraction, loading, and transformation) tool.

Our approach

The Marketbridge Go-to-Market Data Lake (GTMDL) solution is implemented via a custom build engagement using a modern cloud-based data platform of choice. Our development process is designed to be a low-friction experience and the methodologies are based on decades of building large-scale solutions for marketing science, analytics, activation and reporting use cases. Our embedded teams complement your existing technical resources, ship real code and provide transparency throughout the development process. Our consultative approach gives you the confidence to drive organizational change to new heights.

Andy Hasselwander
Chief Analytics Officer
Let's Connect | in

Steve Erbentraut
Managing Director, Technology Solutions
Let's Connect | in

marketbridge

marketbridge.com
Boston
Hartford
Toronto
Bethesda
Richmond

.Reinvent growth.