Data Architecture Principles

FOR SALESFORCE MARKETING CLOUD

ELIOT HARPER

CloudKettle

Contents

Introduction

Data architecture can quickly get complicated. It wasn’t always this way, but over the past two decades there has been an explosion in adoption of voluminous or ‘big’ data, which has been fueled by the evolution of digital storage mediums. And as a result, data has quickly grown in variety, volume and velocity.

This data phenomenon has enabled organizations to provide highly personalized experiences by unifying discrete customer data points across website visits, sales, customer service, email engagement and more. And while data effectively forms the backbone of customer engagement in digital marketing platforms like Salesforce Marketing Cloud, the reality is that fundamental structural choices need to be considered when integrating different data points, as a poorly considered data architecture results in decisions which can be very costly to change later.

Data ‘architecture’ is essentially a metaphor, analogous to designing a building — it refers to the planning and discipline processes of understanding data, then making considered decisions for managing its use. This book identifies six key principles to consider when designing a data architecture for Salesforce Marketing Cloud.

PRINCIPLE #1

Purpose

The first principle to consider is to determine the purpose of the data source. Some marketers treat data with a view of deservingness; if it is available, then they believe that they are entitled to it. And as long as Salesforce has a lenient view on data storage in the platform (essentially unlimited storage, at no additional cost) then why not store everything, forever?

Marketing Cloud accounts often contain bloated data extensions with every customer data activity that has ever occurred, like web page visits or ecommerce events, when the reality is that the data is completely irrelevant, if not useless.

It turns out that there is really such a thing as “too much data” and there are consequences of storing it.

The more data there is, the greater the compliance risks, security and privacy issues.

And having useless data mounting in data extensions can be time-consuming to troubleshoot and fix when processes break, for example when query or import activities in Automation Studio start timing out.

To determine the data purpose for each data source, a well developed mission statement should be crafted, detailing how the data will be used and who will use it. This, in turn, will help to validate whether the data is actually required in the first place.

PRINCIPLE #2

Storage

Once the purpose of the data source has been established, then the organization must determine how long the data is actually needed for. Once again, thanks to the notable absence of data storage fees in Marketing Cloud (a policy that is likely to change in the future), platform users quite happily store all their data, forever. It is not uncommon to see data extensions containing landing page form submissions, journey logs, send logs and Subscriber records from every past batch send buried away, out of mind, in date-based nested folders — often amounting in thousands of individual data extensions across business units.

And without any performance or cost impact, there is little incentive for users to delete all this data. However, there are a few problems with this storage philosophy.

Firstly, when mountains of data exist, it becomes challenging to quickly and easily access the required data. It is likely lost in a directory or filename taxonomy which may have made sense when the data was originally created, but is no longer relevant or applicable. This leads to increased time and effort in locating data, reducing overall productivity.

Secondly, and more importantly, data should not be retained just because it might be useful to someone in the future.

Security laws prevent organizations from storing data indefinitely; these laws include GDPR, HIPAA and CPRA just to mention a few. It must be determined why an organization needs to store data and identify if there are any regulatory or legal requirements for retaining it. It is important to establish a data retention policy for all data sources, which can easily be implemented at a data extension level.

If there is a valid reason to store the data perpetually, then chances are that Marketing Cloud is not the best repository for it, as data can easily be deleted or overwritten unintentionally on the platform. In this case, a data lake or data warehouse would be a better option for long-term data storage.

PRINCIPLE #3

Simplicity

The most common approach to processing data in Marketing Cloud is using Automation Studio, which among other things is purpose-built for performing Extract, Transform and Load (or ETL) operations on data from different sources.

But while Automation Studio is a robust automation platform, automations need to be designed with efficiency in mind. And regardless of the activities within an automation, every automation is essentially data driven.

Query activities are a common bottleneck in automations as these activities enable database administrators to literally knock themselves out and write quite inspiring lengthy statements nested in multiple levels. While long code in itself is not necessarily a problem for Marketing Cloud, the code is typically accompanied by a level of difficulty. When the query fails, or needs to be updated to accommodate additional fields or business requirements, it can take a considerable amount of time and expertise to isolate and fix. This, in turn, can impact the timeliness of data processing, which can be detrimental to marketing efforts.

> Just because data can be manipulated in every imaginable way with SQL, does not mean it should be. Simplicity trumps complexity in efficient data architecture.

If it seems necessary to write complex query activities in Automation Studio, then it is highly likely that the wrong platform is being used. Instead, consider processing data upstream with a fit-for-purpose platform like Mulesoft (for data integration and transformation) or Salesforce CDP (for segmentation).

PRINCIPLE #4

Schema

A data 'schema' is an abstract design that represents data storage (in data extensions). It not only defines data types, lengths, and required fields, but also identifies relationships between data extensions. Think of a schema as a 'blueprint' to ensure efficient organization of data, making the data easier to manage and maintain.

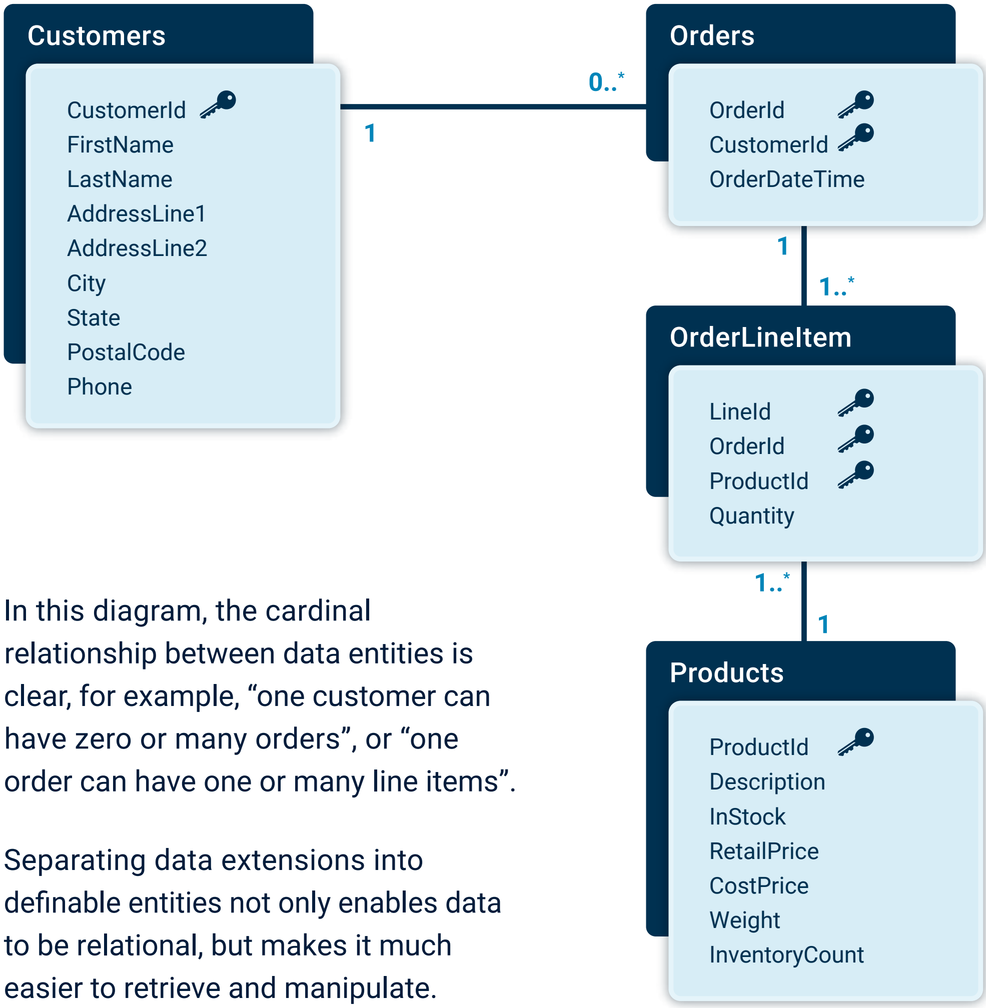
Data extensions in Marketing Cloud are essentially a blank canvas. While they allow considerable flexibility for data storage, unfortunately it is common to see ill-considered data extension schemas with unlimited field lengths, excessive use of primary keys (so values can be searched in Contact Builder), and even hundreds of fields — all of which will provide an additional unnecessary overhead.

The first step in creating a schema is to define a common vocabulary that enables data reuse and portability. Adopt a data extension and field naming convention that has context and meaning to others. Avoid using special characters like hyphens or spaces, as while they may technically be supported, queries and application code (like SSJS) need to be written differently and data is not necessarily portable when exchanged with other platforms. Consider preserving words that are hard to distinguish when concatenated by formatting them in 'snake case' (underscores instead of space characters) or use another naming convention, but be consistent.

The second step is to consider how data sources relate to each other. The best way to describe this is by illustrating relationships between data components. This is called an entity-relationship diagram, or an ERD.

There are many different notation approaches to indicate the cardinality between data entities in ERDs. The following example uses UML notation to indicate a relationship pair of none (0), one (1), or many (*) for customer order data.

Entity Relationship Diagram For Customer Order Data



In this diagram, the cardinal relationship between data entities is clear, for example, “one customer can have zero or many orders”, or “one order can have one or many line items”.

Separating data extensions into definable entities not only enables data to be relational, but makes it much easier to retrieve and manipulate.

To briefly get into semantics, an ERD and data schema are not the same thing; they are similar, but different. An ERD provides a high-level data model, depicting how data is structured at a logical and visual level, whereas a relational schema focuses on a lower row and column level. However, using an ERD to visualize data relationships is a good starting point.

After the schema is created, document it. A significant part of technical debt is documentation debt, which occurs when data schemas (or code) are created without supporting internal documentation. Database schemas are useful long after they have been created, and clear, concise documentation helps avoid confusion and increase productivity of other team members.

PRINCIPLE #5

Governance

Data governance is a multi-disciplinary process, but broadly speaking, the term refers to the discipline of planning how an organization uses data so it is handled consistently throughout the business in order to extract value from all the information collected and stored in it.

In other words, data governance ensures that data is consistent, trustworthy and does not get misused.

Proper data governance planning is essential in Marketing Cloud, as the platform is effectively powered by data, where data is used for ingestion, decisioning, personalization and reporting.

Start by setting realistic and measurable goals (as you cannot control what you cannot measure). Goals will differ by organization, but consider adopting the following three goals as an initial framework.

Goal 1: Understandable

For data to be understood across the organization, it needs to be structured in a way that it can be used effectively. This goal goes hand-in-hand with the data schema discussed in the previous section, where a data dictionary has a taxonomy that has context and meaning to all who use it. And as stressed earlier, it needs to be well documented.

Goal 2: Compliant

Data and privacy compliance is a vast topic, but at a minimum it is essential to comply with applicable data privacy and protection laws. While data protection regulations vary, they all share the same aim, which is to give individuals control over their personal data while also requiring organizations to establish a legal basis for processing such data.

The legal bases for data processing varies by data protection regulation. The GDPR and CCPA both have six legal bases, the PDPB has seven, and the LGPD has ten. Determining a legal basis for processing data is key, as all regulations articulate that data can only be processed if there is at least one legal basis for doing so.

While legal bases vary by regulation, one common basis that is shared across all regulations is consent. And consent is always required for marketing.

Additionally, it is important to document the data lineage to reveal the entire data lifecycle and flow from source to consumption, which includes all transformations the data underwent along the way — how the data was transformed, what was changed, and why.

Goal 3: Trustworthy

Data governance is a trust-based process. And if data cannot be trusted, then it is a source of risk, as it can result in poor business decisions, and compromise customer relationships by targeting them with irrelevant or wrong content.

Data trust needs to be earned, not taken as a leap of faith. In order to trust an organization's data, the data must produce reliable analytics to support well-informed business decisions. Trusted data maximizes the ability to create value from it.

To build trust, begin by validating that the data is accurate and complete for each data extension. For example, if ecommerce website activity for customers is being stored, are all page visits and actions correctly captured? And do product page visits correlate to active products in a product data extension? Also check that data is complete. If there is an expectation that certain data extension fields should have a value, then create an automation to query the data and use a validation activity to report on exceptions.

PRINCIPLE #6

Security

Security is about keeping sensitive and confidential data secure, yet accessible to those who need it. Some Marketing Cloud customers assume that security is not actually their concern, as the platform is hosted in a highly secure, managed environment. That may be true, but there is still plenty of opportunity for data to be compromised.

Cybercrime is the fastest-growing crime globally, increasing every year in size, sophistication and cost.

While cybercrime takes on many forms, unauthorized data access, or a ‘data breach’ is a common method of obtaining personally identifiable information (or ‘PII’). This data can then be further used for identity theft (which comes in different flavors), or holding organizations to ransom, where the attacker demands payment (typically Bitcoin) and threatens to publish the stolen data (typically on a website) if payment is not received.

The following best practices will help mitigate cybercrime and security risks.

User Access Reviews

Organizations should periodically verify that only legitimate users have access to the platform. When staff or contractors leave an organization, their Marketing Cloud access can be overlooked and as a result they may continue to have access to the platform, which would allow them to export data or behave maliciously. **It is recommended to review user access every 30 days and revoke access for staff and contractors who have left or no longer require access.**

Principle of Least Privilege

Principle of Least Privilege, or ‘PoLP’ refers to the practice of enforcing the minimal level of user permissions that allows the user to perform their role. Assigning a broad permission scope to users could result in users intentionally (or unintentionally) performing malicious activities, like deleting data extensions. Additionally, assigning users with administrative privileges not only enables them to create user accounts, but also provides access to API authentication credentials for Installed Packages, which could then be used to perform activities like retrieving data, even after their account is disabled.

Marketing Cloud offers a highly granular permission set, allowing custom user roles to be created. **Roles should tightly align with responsibilities, so users can only access specific applications and features needed to perform their work.**

File Transfer

All Marketing Cloud editions include an SFTP account, or 'Enhanced FTP' where data files can be imported to, or exported from.

And SFTP is widely regarded today as the de facto protocol for secure file transfer. Marketing Cloud offers different authentication options when creating SFTP user accounts, which includes password, ssh key, or a combination of both.

While these authentication options are secure, they ultimately rely on users to follow best practices. **Key based authentication is stronger than passwords.** Without diving into a technical explanation, it is possible to guess passwords, but impossible to guess a key. However, strong passwords can effectively minimize this risk. Additionally, passwords should be changed every 90 days or less.

These same rules also apply to SSH keys. Only designated users should be able to access SSH keys and it is important to enforce diligent key rotation, while also disallowing the use of matching passphrases across multiple keys or iterations.

Deciding on the best authentication option is not straightforward. Both passwords and SSH keys offer their own advantages and disadvantages. However, where possible, it is recommended that both SSH key and password authentication are adopted.

The main justification is that if the private key is compromised (for example, a device is stolen or malware is installed on it), the attacker would not be able to compromise the SFTP account without the password. Similarly, if a password is compromised, then they would still need the SSH key. It is still not foolproof, but this does provide dual-factor authentication and makes it harder for an attacker to compromise credentials.

File Encryption

When transferring files to and from Marketing Cloud using SFTP, consider encrypting files using PGP. While it may be assumed that SFTP replaces the need for PGP, it does not.

SFTP and PGP have two different goals. PGP encrypts the data payload, while SFTP encrypts the file transfer. **At a minimum, transport encryption is required.** Data encryption provides an extra security layer.

Marketing Cloud supports importing and exporting both PGP and GPG files. GPG and PGP are almost identical, with the major difference between them being how they are licensed to the public. PGP or GPG encryption protects data at rest, which ensures that the data file is not exposed after it is transferred to an SFTP server or file system.

Encryption at Rest

Certain regulations and enterprise organization policies require data to be encrypted at rest. **This encryption type prevents attackers from accessing unencrypted data** — if an attacker obtains a hard drive with encrypted data (from the data center) but not the encryption keys, then the attacker is unable to read the data.

Encryption at rest is available as an optional feature in Marketing Cloud, either as a dedicated database (for an individual Marketing Cloud account) or on a shared database (with other accounts).

Summary

Proper data architecture can help ensure that data is secure, accessible, available and can be trusted. Following these six principles early in the data architecture process will help avoid significant issues in the future.

Good data architecture requires careful planning, collaboration, validation and extensive documentation to ensure that it is effective.

While there is a related cost to this level of diligence, the benefits of proper data architecture far outweigh the costs of bad architecture.

About the Author

Eliot Harper is a Technical Architect at CloudKettle and a Salesforce MVP. Eliot is an acknowledged expert in Salesforce Marketing Cloud and is author of *The AMPscript Guide* and *Journey Builder Developer's Guide*. He is a sought-after speaker at international events and regularly publishes related content on social media.

About CloudKettle

CloudKettle helps enterprises drive revenue with the Salesforce and Google ecosystems. We do this by providing the strategy and hands-on keyboard execution to leverage platforms like Salesforce Sales Cloud, Marketing Cloud, Einstein, and CRM Analytics to create highly personalized cross-channel experiences that drive revenue.

As your strategic advisor, we help by enhancing your people, processes, and technology to build a roadmap centered around scalable tactics and security.

To learn more, contact us at:

hello@cloudkettle.com
cloudkettle.com
1-800-878-4756 ext. 202