CardinalPath

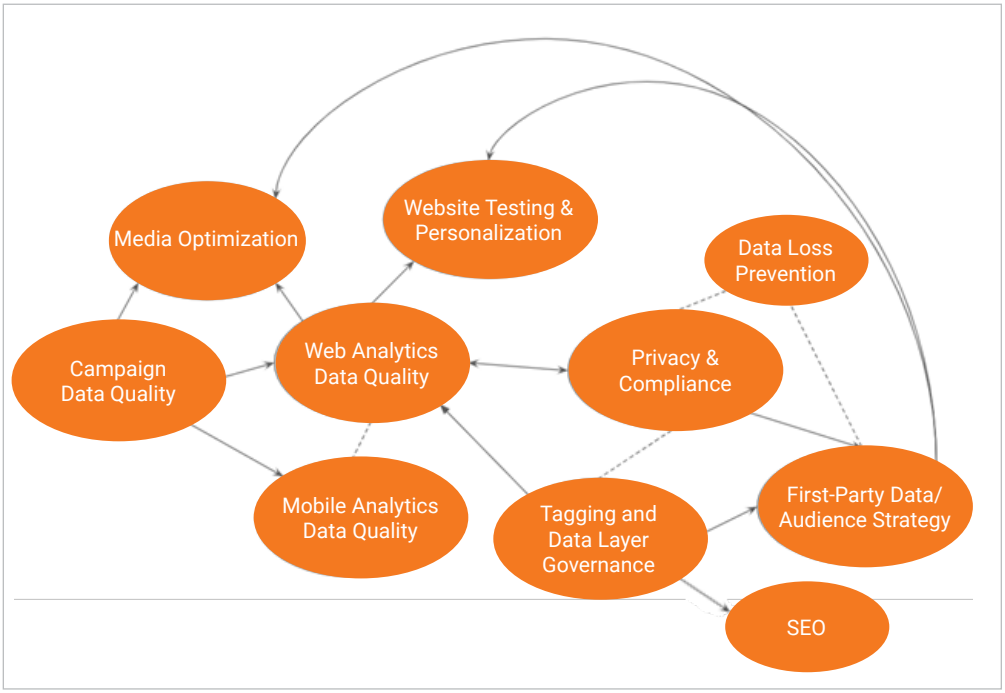
A Data Quality Approach to Analytics, Media Optimization, and Privacy

Outline

Introduction:
An Ecosystem of Data Quality Benefits

In the world of digital analytics, the topic of data quality doesn't seem to generate too much excitement. For some of us, however, data quality is far from bland or inconsequential: we know it is foundational, directly supporting all downstream analysis, insight, and optimization. We have been in the trenches and understand that laxness in data quality can undermine strategy and compromise effectiveness.

This white paper aims to outline some high-level data quality considerations, as well as some specific tactics, and to demonstrate how a focus on data quality can support the conversion optimization, media effectiveness, and overall ROI that our industry constantly strives for, and also help inform platform-level and regulatory compliance.



This white paper is framed largely in the context of Google environments but should also apply more broadly.

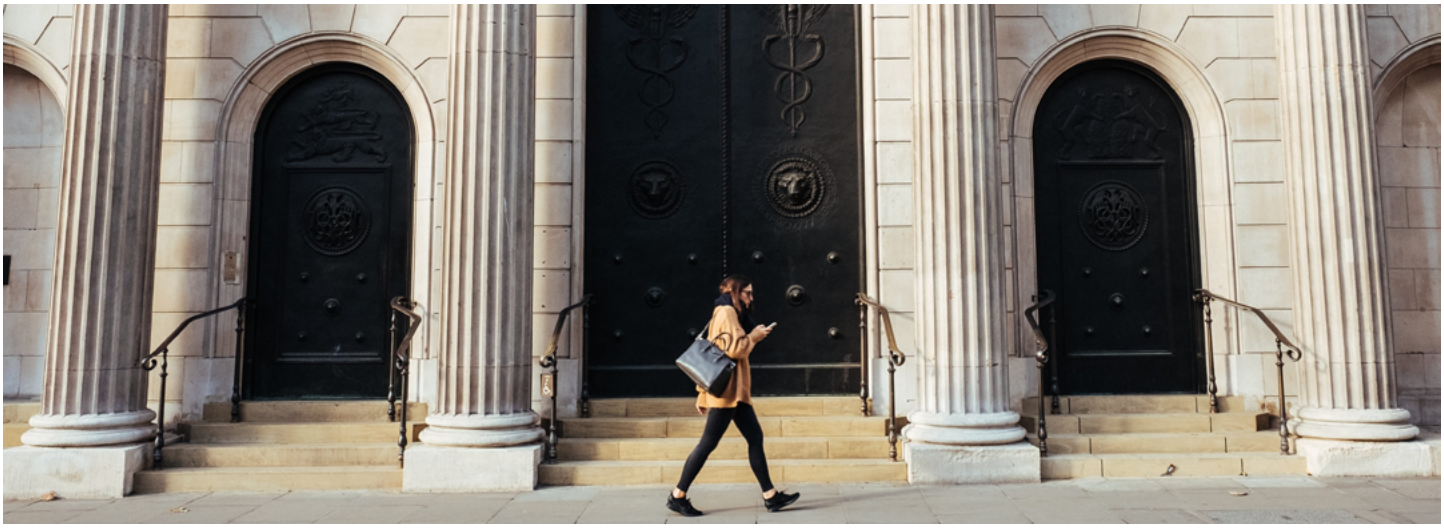
The discussion will not directly address ETL platforms or data preparation tools in a formal sense but will instead focus primarily on data quality at the moment the data is captured.

This white paper outlines the continuum of optimization and compliance benefits that you can support through a focus on data quality starting in web analytics.

Gartner Marketing and Data Analytics Survey 2020 Highlights Data Quality Issues

Gartner's 2020 survey indicates that just over half of senior marketing leaders are dissatisfied with results from analytics investments and that analytics is factoring into just 54% of their marketing decisions overall. Poor data quality is cited as one of the primary reasons (along with inactionable results and nebulous recommendations) for their doubt and disappointment.

It's always healthy, of course, to view surveys with a critical eye, but it might be wise to heed these results, which pretty unambiguously position data quality as an important aspect of analytics stakeholder satisfaction and a positive, empowered data culture.



Basics of Web Analytics Data Quality

Attention to data quality in your web analytics data capture can enhance relevance, increase likelihood of insight, and also provide a template for tracking and activation in related platforms. Some core considerations are outlined below.

Content and User Taxonomies

The discovery phase of a web analytics implementation often surfaces three back-end taxonomies that you can incorporate directly into the web analytics dataset:

***content/page taxonomies*, often housed in a CMS; page-level fields can include:**

- subject/category
- page type/function
- author
- publish date
- flag indicating presence of images or video

***user taxonomies*, often built into a CRM; user-level fields can include:**

- loyalty or status level
- lead or opportunity stage
- product preferences
- job title
- industry
- other demographic data
- anonymous user ID

***product taxonomies* in an ecommerce context that can include:**

- product name
- SKU
- size
- target gender
- color
- brand
- additional descriptors

Taxonomy is hard work. If your organization has taken the initiative to build the systems and processes for classifying your content, user base, and products in a way that reflects and supports your organizational functions, it's important to incorporate these classifications into your web analytics so the resulting reporting is more meaningful and actionable.

Integrate directly or join in another platform?

There are essentially two overall approaches for analyzing back-end data alongside front-end web analytics data:

- add join keys between your analytics data on the one hand and your back-end databases on the other and then perform your analysis on exported data in a third-party platform
- incorporate the back-end taxonomies directly into your web analytics data at the time of capture (or through a direct import into the analytics platform, as applicable)

This white paper generally takes the approach of direct incorporation, even though it's more redundant, since it makes the enriched web analytics dataset immediately accessible to more analysts and marketers within your organization. We'll encounter the question of direct incorporation, parallel tracking/redundancy, and joining at several additional points below.

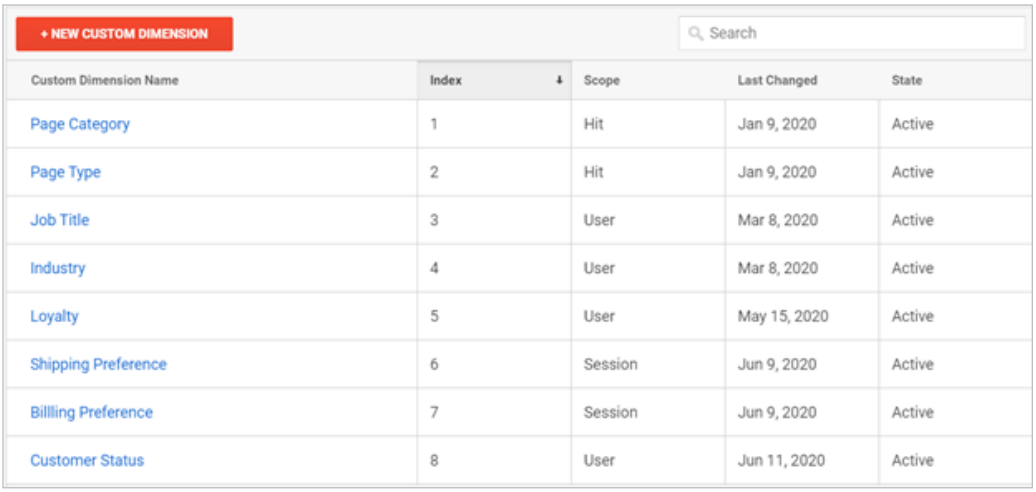
Populate Every Custom Field that You Allocate

In Google Universal Analytics implementations, custom dimensions are typically the mechanism that you use for recording user and content taxonomies and extending the built-in data model. Product taxonomies are typically captured in predefined ecommerce fields, with overflow product descriptors also stored as custom dimensions.

The intention to record these taxonomies, however, doesn't always manifest in the actual data capture. Analytics implementations routinely suffer from custom dimensions that are allocated but unpopulated, populated only sporadically, or sometimes populated with non-human-readable values.

This is indeed a problem, since the consistent incorporation of back-end taxonomies into your web or mobile app analytics provides at least two crucial benefits:

- **better analysis:** the enrichment of the web analytics dataset will support your analysis and speak to your stakeholders, as mentioned above.
- **better audiences:** custom dimensions (or the variables that you read from a data layer to populate the custom dimensions) are typically a critical part of audience definitions for media targeting and email targeting, as detailed in Parallel Analytics and Media Tracking below.



Custom Dimension Name	Index	Scope	Last Changed	State
Page Category	1	Hit	Jan 9, 2020	Active
Page Type	2	Hit	Jan 9, 2020	Active
Job Title	3	User	Mar 8, 2020	Active
Industry	4	User	Mar 8, 2020	Active
Loyalty	5	User	May 15, 2020	Active
Shipping Preference	6	Session	Jun 9, 2020	Active
Billling Preference	7	Session	Jun 9, 2020	Active
Customer Status	8	User	Jun 11, 2020	Active

In many instances, custom dimensions that are defined in the Google Analytics admin remain mostly or completely empty or become populated with non-human-readable values.

Take advantage of this Google Analytics custom dimension utility to ensure that the custom dimensions that you have defined in your Google Universal Analytics property are being populated.

Google Analytics 4 and Server-Side Google Tag Manager

Tagging and tracking in the Google realm are undergoing two paradigm shifts: the emergence of Google Analytics 4 following the years-long run of Google Universal Analytics, and the introduction of server-side Google Tag Manager (GTM) as an alternative to traditional client-side Google Tag Manager.

As this help article indicates, the custom tasks functionality commonly used for data cleanup and consolidation is not, at least yet, available in GA; ditto for view filters and settings. The capability of server-side GTM to transform (or block) any hits on the server before dispatching them to endpoints such as Google Analytics may come to serve a purpose similar to custom tasks in regards to PII remediation, URL consolidation, and other modifications of your raw Google Analytics data capture.

Stay tuned as features and adoption continue to evolve for both of these platforms.

Fight URL Fragmentation

The URLs captured in your web analytics dataset should be only granular enough to reflect differences in page content. If the page content doesn't change with a query parameter or another URL variation, consolidate the URLs into a single canonical dimension value for analytics purposes.

If, for instance, the following URLs point to the same page content and provide the same end-user experience, there is little to no reason for maintaining the query parameter in your main web analytics reporting:

`www.yourcompanysite.com/great-deals?`**`session_id=123`**

`www.yourcompanysite.com/great-deals?`**`session_id=456`**

Pageviews of these URL variations should instead be consolidated under a single dimension value representing the core intent of the page:

`www.yourcompanysite.com/great-deals`

Really basic stuff, but as of late 2020, it remains a pervasive issue that adds needless complexity to many analytics datasets.

Stripping out parts of the URL that aren't useful for analysis can often also help to address PII in your analytics data, as outlined in the next section.

	Page	Pageviews	Unique Pageviews	Avg. Time on Page	Entrances	Bounce Rate
		185,785,337 % of Total: 100.00% (185,785,337)	127,164,805 % of Total: 100.00% (127,164,805)	00:00:50 Avg for View: 00:00:50 (0.00%)	24,795,717 % of Total: 100.00% (24,795,717)	30.86% Avg for View: 30.86% (0.00%)
1.	[redacted]	19,768,901 (10.64%)	10,803,918 (8.50%)	00:00:36	3,381,050 (13.64%)	25.80%
2.	[redacted]	11,531,682 (6.21%)	7,850,674 (6.17%)	00:00:39	5,582,965 (22.52%)	19.03%
3.	[redacted]	9,887,213 (5.32%)	5,219,258 (4.10%)	00:00:50	24,018 (0.10%)	18.14%
4.	[redacted]	9,338,105 (5.03%)	5,206,421 (4.09%)	00:00:02	4,976 (0.02%)	3.61%
5.	[redacted]	7,520,399 (4.05%)	5,377,733 (4.23%)	00:00:43	2,871,689 (11.58%)	13.33%
6.	[redacted]	6,434,208 (3.46%)	5,197,128 (4.09%)	00:00:58	2,149 (0.01%)	8.57%
7.	[redacted]	6,064,203 (3.26%)	5,094,681 (4.01%)	00:00:12	6,957 (0.03%)	2.61%
8.	[redacted]	5,509,743 (2.97%)	4,906,179 (3.86%)	00:00:17	68,927 (0.28%)	1.42%
9.	[redacted]	5,482,639 (2.95%)	3,547,047 (2.79%)	00:02:02	6,492 (0.03%)	12.45%
10.	[redacted]	5,108,094 (2.75%)	3,551,350 (2.79%)	00:00:28	286,596 (1.16%)	53.73%

Show rows: 10 Go to: 1 1 - 10 of 590414

Even in analytics implementations that are otherwise well executed, URL fragmentation often goes unaddressed. It is not uncommon to see hundreds of thousands or even 1,000,000 URL variations over the course of a 30-day reporting period. You can make the job of every analyst easier downstream by remedying URL fragmentation when the data is first captured.

Allowlist vs. Blocklist Approach for Query Parameter Stripping

The two basic approaches that you can take for query parameter stripping are allowlist and blocklist. In the allowlist approach, all query parameters are stripped by default, with an allowlist in which you specify those query parameters that can remain in the URL. In the blocklist approach, you allow query parameters to remain in the URL by default and specify a blocklist for those that you want to strip out.

An allowlist is normally achieved through your own client-side scripting (often through a Google Analytics custom task) and provides the advantage of stripping any new parameters that begin appearing in URLs after the time of your implementation. The disadvantage is that you can inadvertently strip out query parameters that could be useful for your analysis. (An example of a useful query parameter would be the sort or display type for a products page, e.g., *display=list* or *display=thumbnail*, which you should keep in the URL recorded in your analytics data, or – even better – store in a separate field, most likely as a custom dimension in the Google Analytics context, before stripping from the URL).

Individual analytics implementers come down on different sides of the allowlist vs. blocklist debate. In either case, make sure to address URL fragmentation, but try to go about it in a way that does not sacrifice any meaning in your data.

Take advantage of this Google Universal Analytics query parameter utility to list the query parameters that are currently being captured as part of the Page (aka Request URI) dimension in a Google Analytics view. If the query parameters are not helpful for analysis, strip them out using Google Analytics filtering or your own client-side scripting, potentially as a Google Analytics custom task as outlined above.

Search Engine Optimization

Some considerations for web analytics data quality map quite closely to search engine optimization factors.

Page Canonicalization

Search engines ask us, as website owners, to indicate a single version of the URL by which to index a page of content. For SEO, this discipline of URL consolidation is referred to as canonicalization.

To cite another example, let's say that these three URLs all refer to the same page content:

```
https://www.yourcompanysite.com/about-us/
https://www.yourcompanysite.com/about-us/index.php
https://www.yourcompanysite.com/about-us/index.php?page=1
```

The mechanics for canonicalizing URLs in web analytics and in the search engines differ, but the principle is comparable in most cases: a unit of page content should be referred to by a single, consolidated form of the URL, even if end users may be seeing multiple versions.

For analytics, you enforce URL consolidation through the analytics, as referenced above. For SEO, you enforce URL canonicalization by including a directive directly on all pages that the search engines may otherwise try to index individually, e.g.:

```
<link rel="canonical" href="https://https://www.yourcompanysite.com/about-us/" />
```

The best practice of aligning a page of content to a single URL yields benefits: in the case of analytics, better and easier analysis, and in the case of SEO, clearer instruction to the search engines and the potential for better ranking or better user experience on the clickthrough page.

Should I always consolidate the same URLs for analytics and for SEO?

No, there may be some cases in which you may want to record separate pages in analytics but canonicalize for SEO purposes. For instance, a paginated user experience may encompass the following URLs

```
https://www.yourcompanysite.com/my-content/teddy-bear
https://www.yourcompanysite.com/my-content/teddy-bear?page=2
https://www.yourcompanysite.com/my-content/teddy-bear?page=3
```

For SEO purposes, you may want to include a canonical tag in pages 2 and 3 so only the main page is indexed, which would mean that users who click through from the search engines would begin their experience on page 1. In other cases, paginated URLs may be self-canonicalized to index each one.

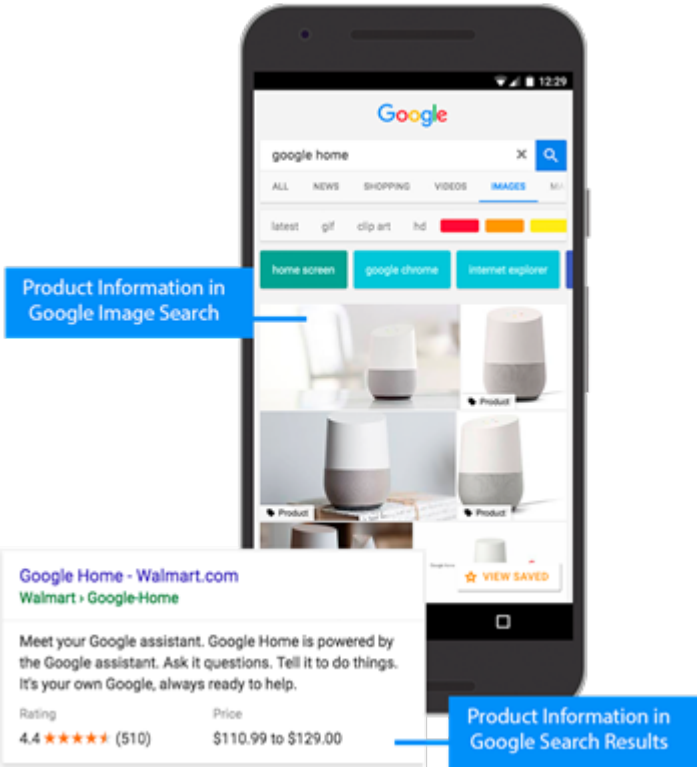
For analytics, on the other hand, you'll probably do not want to strip the page parameter from the URL, since it would be meaningful in indicating user interaction.

Page-Level Structured Data

Just as analytics should reflect back-end taxonomies, there is the opportunity to surface back-end taxonomies for SEO purposes as well, particularly page-level or product-level taxonomies housed in a content management system or an ecommerce system.

Through the schema.org initiative, Google, Bing, and other search engines have collaborated to create a shared vocabulary that makes it easier for webmasters and developers to decide on a schema and gain the maximum SEO benefit from many of the taxonomies that organizations have established.

For instance, if you're populating a page-level data layer with back-end values and reading them into Google Analytics as custom dimensions or ecommerce values, you can likely take advantage of schema.org's structured data markup to enable rich snippets in search engine results and participate in Google Shopping results.



Back-end product and page taxonomies that you capture in web analytics can also benefit rankings and appearance on the search engine results page (SERP). (Source: Google Search Central)

The listing below illustrates product taxonomy populated into the page as schema.org markup.

```
<script type="application/ld+json">
    {
      "@context": "https://schema.org/",
      "@type": "Product",
      "name": "Executive Anvil",
       "description": "Sleeker than ACME's Classic Anvil, the Executive
Anvil is perfect for the business traveler looking for something to drop
from a height.",
      "brand": {
        "@type": "Brand",
        "name": "ACME"
      },
      "review": {
        "@type": "Review",
        "reviewRating": {
          "@type": "Rating",
          "ratingValue": "4"
        },
      },
      "offers": {
        "priceCurrency": "USD",
        "price": "119.99"
      }
    }
    </script>
```

As demonstrated here, you can use many of the same back-end values that you're reading into analytics to also populate schema.org markup, which can benefit your organic search performance.

Additional Housekeeping

Some of the basic housekeeping in web analytics implementations is still sometimes overlooked, particularly lowercasing any dimension that could contain case variations, including campaign, event, and search term dimensions. Don't make the analyst's job any harder by keeping dimensions unnecessarily fragmented by case.

Key Takeaways

- Consider existing user, content, and product taxonomies from the back end when determining implementation requirements for web analytics (while respecting privacy regulations and platform-based terms of service).
- Periodically validate that these values are actually being populated into your analytics dataset, and with human-readable values.
- Verify that URLs reported in your main web analytics reporting views are not more granular than the page content they refer to; keep the ratio 1:1.
- Don't forget basic housekeeping such as lowercasing.

Personally Identifiable Information (PII)

Disclaimer

This document is not intended to provide legal advice or direct guidance on compliance with privacy legislation or product-level terms of service. Please consult with your own attorneys and privacy/compliance experts on the application of the laws and restrictions to your or your client's specific circumstances.

For industry best practices in regards to privacy, refer to the resources provided by the International Association of Privacy Professionals (IAPP). For more about privacy features in Google Analytics, refer to the terms of service and related policies.

When PII is captured in web or mobile app analytics, it often occurs in a page URL, typically when web forms are configured to send a request to a web server using the GET method. The URL, however, is not the only field in which you might inadvertently capture PII:

FIELD/DIMENSION	EXAMPLE PII INSTANCES
Page/Request URI	User has submitted a web form with method set to GET, e.g.: /thank-you.php?fname=hannah&lname=brown&email=hbrown%40xyz123.com
Custom Dimension/userId	non-anonymized back-end user ID, such as email address, recorded directly as web analytics data
Event Dimensions	name, email address, or phone number entered into a blog comment that is captured as event data
Search	name or personal email address entered into a search field (resulting from an actual search or from confusion with a login field)

We need to stay aware of PII in web analytics relative to tool-level terms of service and to governmental regulations:

- **Terms of Service:** in the case of Google Analytics, PII can directly violate the terms of service and require deletion. Repeated infractions can lead to account suspension.
- **Regulatory Compliance:** capturing personal data without consent, even inadvertently, could contravene data privacy law, and the presence of personal data in your data stores could make it harder for you logistically to fulfill personal information obligations, such as:
  - Data subject (meaning end user) requests for access, deletion, and portability of personal information stipulated by GDPR for you, as a Data Controller, to fulfill
  - the right to erasure (AKA *right to be forgotten*) specified under CCPA
  - the requirements of LGPD, effective as of August 16, 2020, that govern the processing of PII for data subjects located in Brazil, regardless of the location of the organization capturing and processing the data

If PII is occurring in query parameters within your website URLs, as described above, you can change your web architecture to eliminate query parameters in URLs (not a trivial task) or script a solution so that they're not passed to Google Analytics in any hit. A scripted solution could also be effective to block PII that may occur in other dimensions as well.

Google Analytics View Settings or Filters Are Not Ideal for Addressing PII

If you have PII in your URLs or in other fields that are passed to Google Analytics, it may seem reasonable to use view settings or filters to exclude PII. Note, however, that you may be on shaky ground per the Google Analytics Terms of Service as soon as you send PII to the Google Analytics servers, before the processing phase in which the settings and filters are applied. To block PII before it's ever included in any URLs, event dimensions, search terms, and custom dimensions that you send Google Analytics, you can typically use custom scripting that leverages regular expression matching.

Measures for PII prevention in your analytics data go beyond URL and data quality considerations per se. But broadly speaking, we might consider PII to be a data quality issue, not on the analysis end but on the end of tool-level terms of service and potentially regulatory compliance.

PII Can Leak into Your Tech Stack

It's important to stay aware that the same web development architecture that can inject PII into Google Analytics can also leak PII into other components of your tech stack. For instance, a URL that contains PII could easily end up in:

- data visualization platforms such as Tableau or Domo
- ad tech platforms such as the Google marketing stack, Facebook, or Adroll
- data warehouses such as BigQuery

While PII may not directly conflict with the terms of service for these other platforms, it may make regulatory requirements much more painstaking for you to fulfill. Reengineering your web architecture to avoid any PII in your URLs (by removing query parameters) should help to avoid PII in other platforms as well.

Key Takeaways

- Removing query parameters from the URLs captured in Google Analytics can help you honor the restriction against PII.
- PII may be present in other dimensions captured as well. Through your own client-side scripting, potentially taking advantage of Google Analytics custom tasks, you can prevent PII from ever touching Google Analytics in any dimension.
- Preventing PII in analytics, media platforms, and back-end systems can help simplify fulfillment of your requirements for personal data requests as stipulated by GDPR, CCPA, LGPD, and other emerging privacy legislation.

Data Loss Prevention

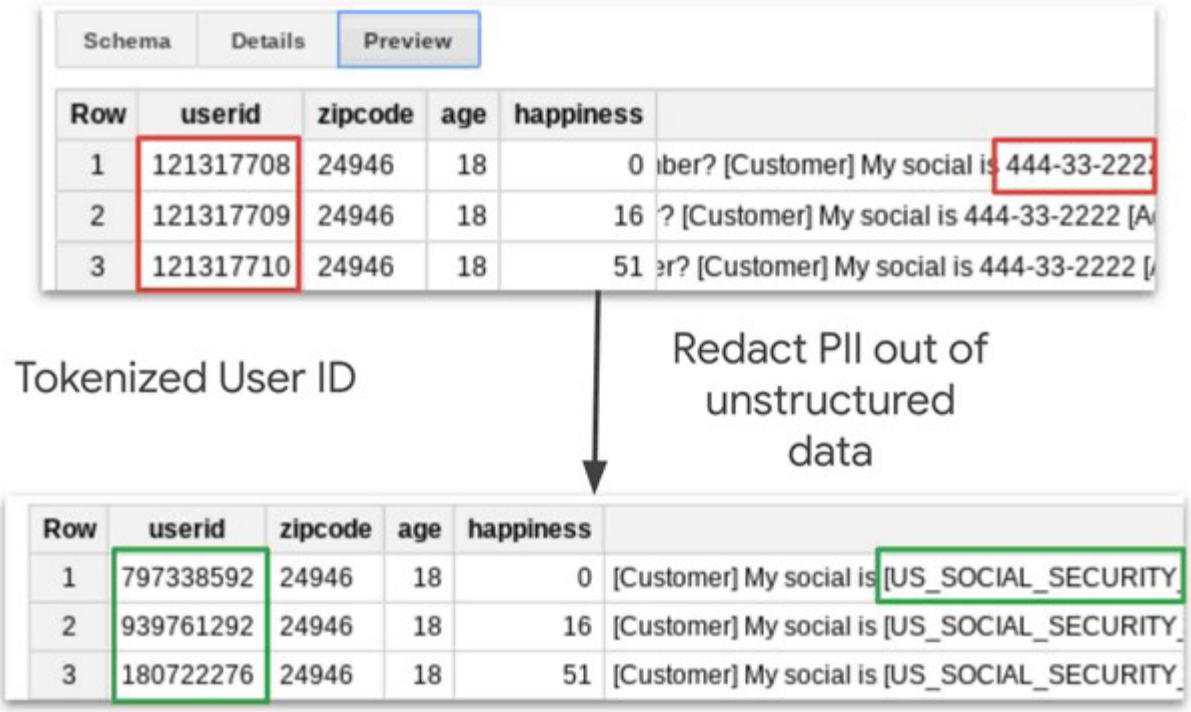
Data Loss Prevention (DLP) platforms are designed to help you catalog, transform, encrypt, and control access to sensitive user data that may be present in structured text, file storage, data streams, and even images. Although DLP does not apply directly to your web analytics data capture, it could play a broader role in data privacy and governance.

The *Loss* in Data Loss Prevention

The *loss* in Data Loss Prevention may require some interpretation: it does not refer to loss prevention primarily as backup or redundancy, but rather as preventing sensitive user data from residing where it should not or propagating to the wrong people, i.e., loss of control.

The more methodical and tactical governance that a DLP platform allows over this aspect of data loss could also help subsequently to limit the impact of any required deletions in the context of privacy legislation, data subject requests, and related policies. In this way, a DLP platform could help to prevent the loss of data in the most basic sense as well.

While you could take on the complex challenge of building all of your own DLP mechanisms, a DLP platform such as Google Cloud Data Loss Prevention provides advanced capabilities for monitoring and redaction out of the box while allowing you to customize the logic for data type detection and obfuscation.

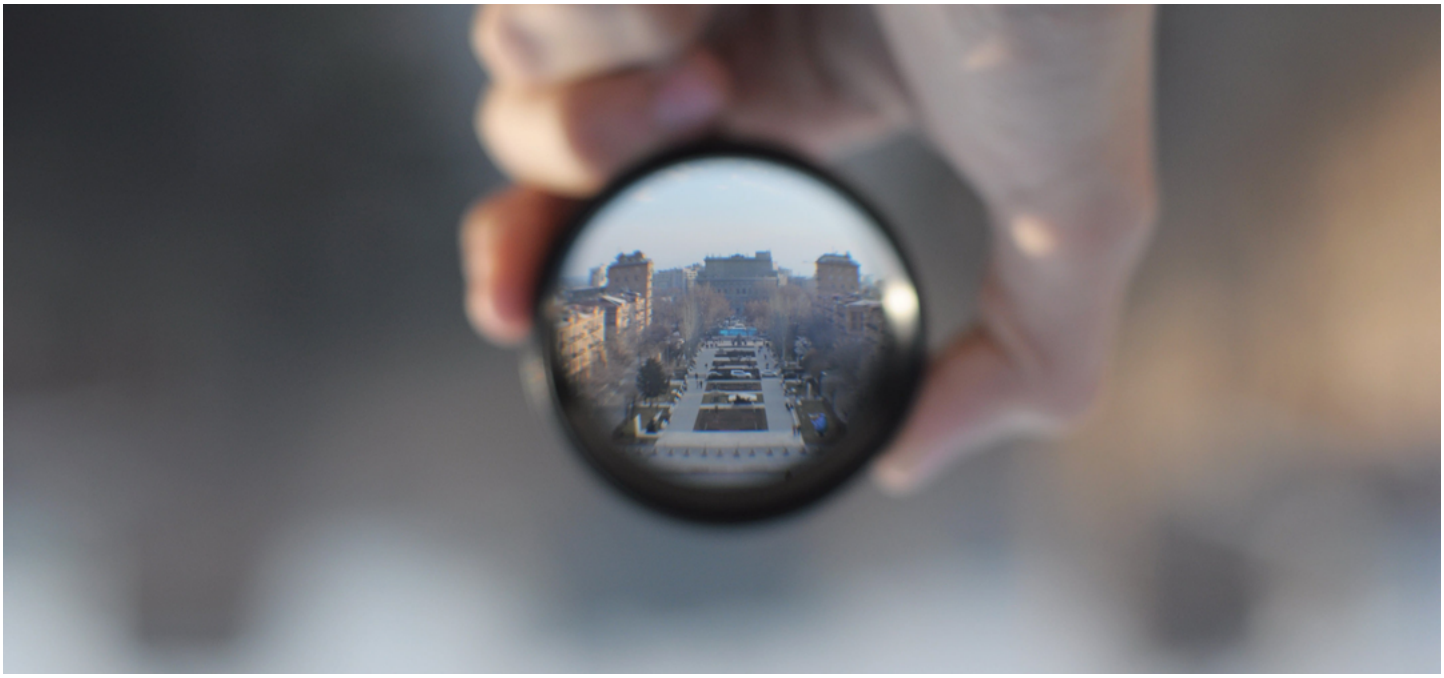


This image illustrates two DLP techniques for obscuring sensitive user data: the userid has been pseudonymized with a format-preserving token, and U.S. Social Security number has been masked with fixed characters. [Source: Take Charge of Your Data: Using Cloud DLP to De-Identify and Obfuscate Sensitive information]

Key Takeaways

- A DLP platform enables classification, redaction, obfuscation, and granular access control for data across a variety of storage formats and streams.
- The *loss* in DLP refers primarily to the loss of control over sensitive user data but may also signify the wider data purges that could potentially be warranted if your organization has not implemented more tactical measures in this regard.

Tagging and Data Layer Governance



Since data capture in most web analytics platforms is achieved through JavaScript tagging added directly to the page or more typically through a tag management system such as Google Tag Manager, Adobe Launch, or Tealium iQ, a tag auditing platform such as DataTrue or ObservePoint can support your efforts to maintain data quality and completeness not only for web analytics, but also for a range of tag types:

- analytics
- media
- marketing automation

The tag auditing platforms may provide different kinds of testing, including:

- coverage tests over a wide range of pages
- automated tests or customized simulation tests for specific flows
- data layer checking
- some degree of PII checking

These platforms can certainly support data quality efforts and can serve as an essential element of your data governance program overall. At the most fundamental level, you can't be collecting data correctly or at least comprehensively if your page tagging is flawed or you're not making the necessary data available at page level for your tags to work with.

Key Takeaways

- Consider a dedicated tagging and data layer auditing tool as part of your data quality and governance program.

Campaign Taxonomies

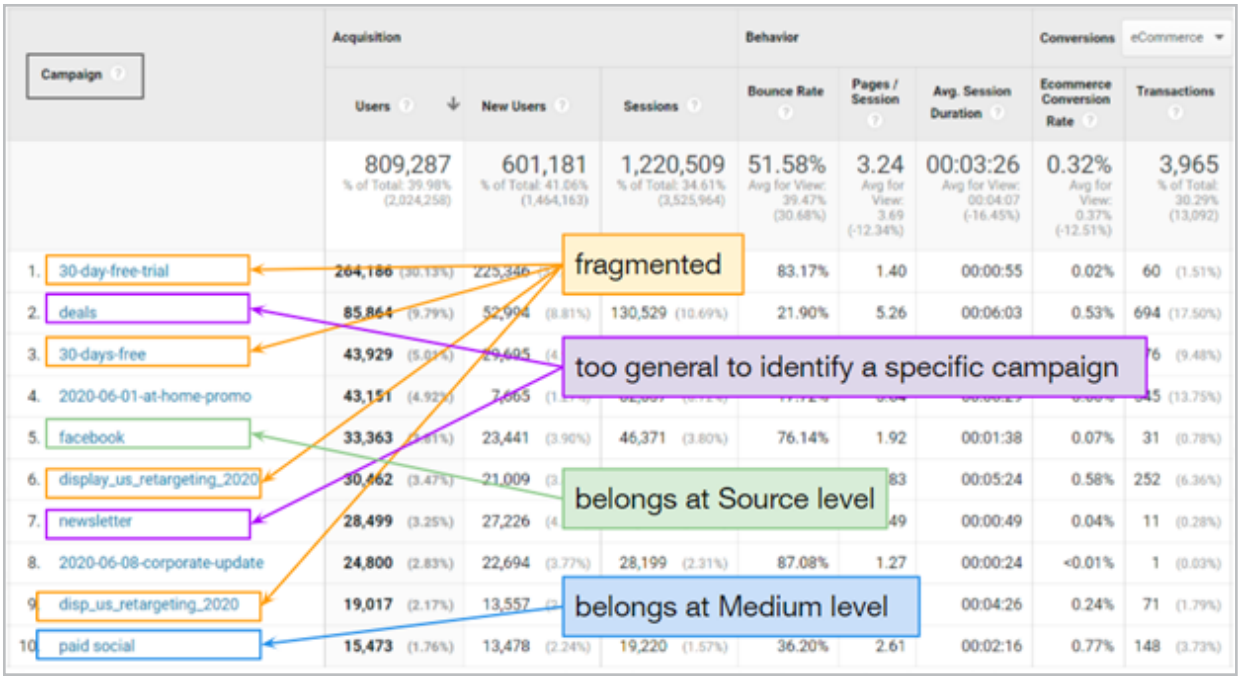
When you boil everything down, there are probably two main questions that we're trying to answer as analysts in our space: how people are finding us, and what they're doing from that point forward.

The first part of that equation can be the harder one to nail down. As challenging as it may be to enforce quality data for a web analytics implementation, it's usually more challenging – sometimes exponentially – to maintain coherent campaign data. Without the necessary controls in place, campaign data usually becomes unwieldy and unusable very quickly.

The governance challenge for campaign data quality can stem from multiplicity in many different factors:

- many channels, many of which require campaign parameter appends for even basically accurate recording by the analytics platforms
- frequent campaigns
- many acquisition/campaign dimensions that you need to populate meaningfully, hierarchically, and consistently
- multiple individuals, teams, or third-party agencies responsible for campaign management and tagging

Consequently, many (or most) organizations struggle to effectively govern acquisition and campaign data.



Campaign	Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration	Ecommerce Conversion Rate	Transactions
	809,287	601,181	1,220,509	51.58%	3.24	00:03:26	0.32%	3,965
1. 30-day-free-trial	264,186	225,346		83.17%	1.40	00:00:55	0.02%	60 (1.51%)
2. deals	85,864 (9.79%)	52,994 (8.81%)	130,529 (10.69%)	21.90%	5.26	00:06:03	0.53%	694 (17.50%)
3. 30-days-free	43,929 (5.01%)	29,695						376 (9.48%)
4. 2020-06-01-at-home-promo	43,151 (4.92%)	7,065						545 (13.75%)
5. facebook	33,363	23,441 (3.90%)	46,371 (3.80%)	76.14%	1.92	00:01:38	0.07%	31 (0.78%)
6. display_us_retargeting_2020	30,462 (3.47%)	21,009		83		00:05:24	0.58%	252 (6.36%)
7. newsletter	28,499 (3.25%)	27,226		49		00:00:49	0.04%	11 (0.28%)
8. 2020-06-08-corporate-update	24,800 (2.83%)	22,694 (3.77%)	28,199 (2.31%)	87.08%	1.27	00:00:24	<0.01%	1 (0.03%)
9. disp_us_retargeting_2020	19,017 (2.17%)	13,557				00:04:26	0.24%	71 (1.79%)
10. paid social	15,473 (1.76%)	13,478 (2.24%)	19,220 (1.57%)	36.20%	2.61	00:02:16	0.77%	148 (3.73%)

This Google Analytics Source/Medium report illustrates some of the basic data quality issues that often degrade acquisition/campaign data and hinder analysis, insight, and optimization.

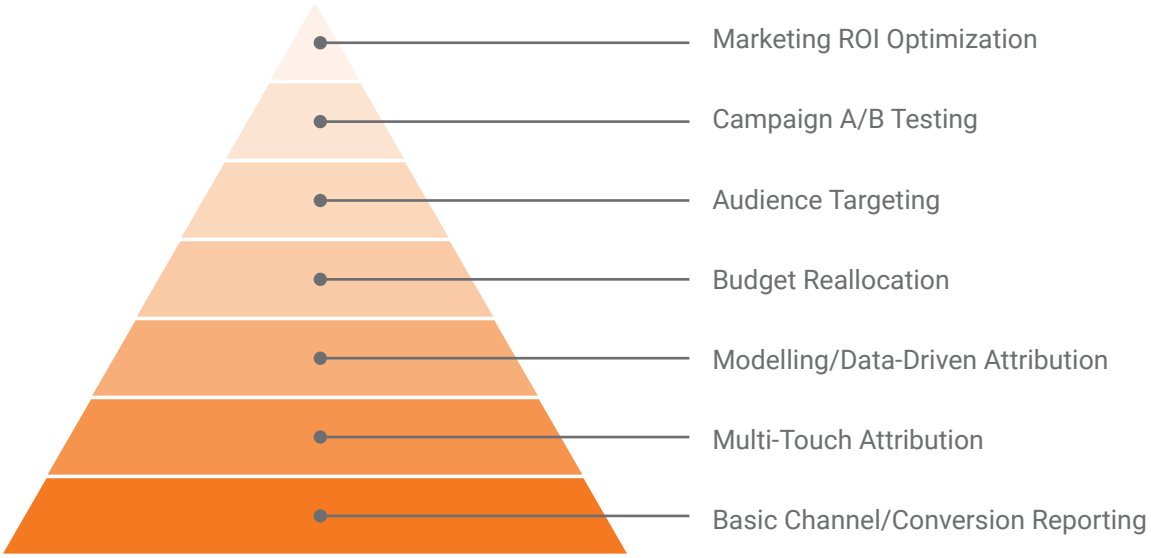
It may be feasible to address some of these issues through data cleanup after initial capture. At a minimum, this will require significant extra time that you and your team would probably rather spend on analysis. In some cases, particularly when the tagging for a campaign has been overlooked completely, no amount of extra time can make up for the gaps and flaws in the original data capture.

The antidote to campaign chaos is coordination. A good start is even a basic spreadsheet in which you and your team maintain a record of campaign naming and validate dimensions such as Medium and Source to match a consistent, pre-determined set of options.

Distributed, enterprise-level platforms such as Claravine can provide even greater control and – by working directly with the Google Analytics and Adobe Analytics APIs – more flexibility in managing and enriching campaign dimensions outside of the actual URL tagging process.

Whatever approach you take, it's quite imperative to gain control over your acquisition and campaign data and to thereby enable a range of marketing analysis and optimization efforts, as illustrated below.

Campaign Data Quality



Attribution analysis and advanced campaign/audience optimizations are harder and less effective if they're not grounded in a foundation of campaign data quality.

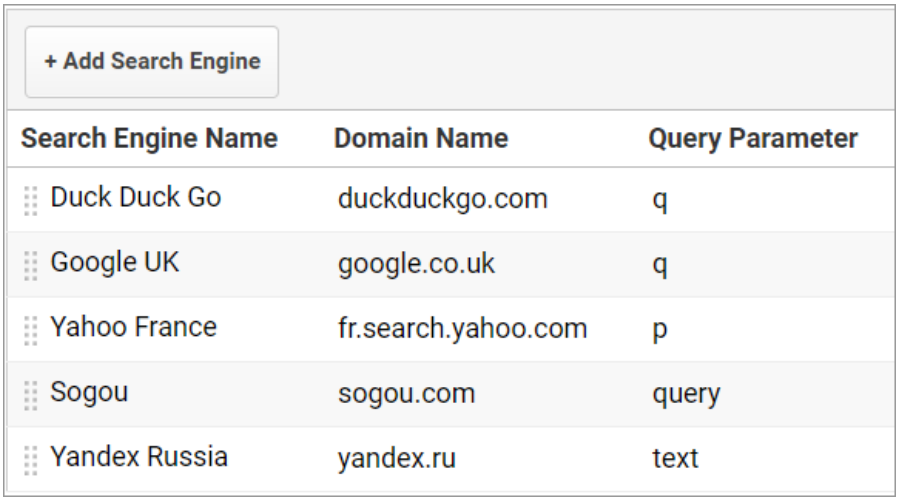
A large part of the value proposition that we provide as analysts and optimizers depends on good acquisition data. And even – or especially – as the landscape of cookies and browsers changes rapidly and makes upper-funnel attribution more challenging, acquisition data quality should remain a priority for all of us.

Take advantage of this Google Analytics campaign tagging utility to control variation of Medium and Source dimensions and record your campaign naming for greater consistency.

Customized Organic Search Sources

As another consideration for acquisition data quality, the Organic Search Sources setting in Google Analytics serves two purposes:

- enables Google Analytics to classify as organic (rather than referral) clickthroughs from new, regional, industry/vertical-specific, or low-volume search engines such as Duck Duck Go or Sogou
- keep country-specific versions of search engines broken out instead of rolled up into a the more general search source (e.g., google.co.uk instead of google)



Organic Search Sources setting in Google Analytics allows greater control over classification of organic search traffic.

Campaign Tracking for Mobile Apps

Campaign tracking for native mobile apps is often completely overlooked. As with your website, it's important to understand the sources that are driving installs and return users for your native apps.

For a more detailed exploration of this topic from the Google end, see Complete Guide to Campaign Tracking for Mobile Apps. As you review this whitepaper, you can focus on the approaches outlined for campaign tracking through Google Analytics for Firebase (aka *Firebase Analytics*), since mobile app tracking with the Google Analytics SDK is in a state of deprecation.

Cross-Platform Attribution with Google Analytics 4

As a further incentive to gain control over campaign tracking for mobile apps, Google Analytics 4 may provide a greater degree of attribution reporting across your websites, Android apps, and iOS apps. At a minimum, the unified data model across web and native app data streams will provide a more streamlined reporting experience that you can support through consistent campaign tracking for websites and native apps.

Key Takeaways

- Due to the multiplicity of many factors, quality in acquisition and campaign data can be even harder to maintain than in other elements of web analytics.
- A range of important downstream analysis and marketing ROI efforts depend on the quality of your campaign data.
- Take advantage of a basic tool or enterprise platform to help maintain campaign data quality.
- Don't forget campaign tagging for native apps.

Parallel Analytics and Media Tracking

Previous sections of this white paper emphasized the importance of capturing content, user, and product taxonomies from the back end as part of your web analytics implementations. As a parallel effort, you can incorporate much of these same taxonomies into your media tracking for conversion analysis and audience creation.

Specifically on the Google Marketing Platform end, Floodlight serves as the:

- shared conversion format for Campaign Manager, Display & Video 360, and Search Ads 360
- pixel-based audience creation mechanism within Campaign Manager, which can also be shared with Display & Video 360 for:
  - remarketing
  - lookalike audience creation
  - audience exclusion (in the case of converters not to retarget, etc.)

Enriched Audience Creation

You can use standard and custom Floodlight variables to record additional context about the user interactions and purchases that you record as Floodlight activities. For much of the content, user, product, and transaction data that you're already reading into Google Analytics or Adobe Analytics, you can also configure Floodlight variables to enrich your Floodlight tracking.

For instance, if you're capturing a user's industry – let's say *insurance* – as a user-scope custom dimension in Google Universal Analytics (originating from a form selection, an ABM integration such as Demandbase, or an offline CRM entry), you can pull *insurance* into a Floodlight variable through the same scripting or tag management features. You can take a similar dual approach for users who have viewed web pages about a certain category within your content taxonomy – let's say *hiking*.

You could use either of these variables to enhance your audience creation and your reporting within the marketing tools. By taking advantage of variables, you can also use fewer Floodlight activities and tags, thus streamlining your tag containers and also encouraging more consistent data capture and reporting across the tools in the Google Marketing Platform.

Why do I need parallel audience creation between Google Analytics and the Google marketing tools?

If you're using Google Analytics 360 and can therefore share Google Analytics audiences with Campaign Manager, Display & Video 360, and Search Ads 360, it may be reasonable to wonder why you need the parallel variable tracking to support the creation of audiences within the marketing tools directly.

The audiences defined within the marketing tools and based on Floodlight tags and variables allow for real-time or near real-time targeting, and may also be useful in optimizing for programmatic buys to meet specific performance objectives.

The audiences defined within Google Analytics, for their part, have access to the full range of fine-grained dimensions in the Google Analytics dataset and can therefore still be very useful for many targeting purposes when shared with the Google media tools.

Streamlined Tracking for Events and Goals

In addition to parallel tracking for the purposes of audience creation, it's also considered best-practice to mirror analytics events and especially goals/transactions as Floodlight activities for Campaign Manager, DV360, and SA360 and thus support greater consistency in reporting across analytics and media platforms.

You can achieve this parallelism by reusing the same variables from the data layer or the browser environment; the table below outlines a potential approach.

EVENT DESCRIPTORS AVAILABLE IN TAG MANAGEMENT DATA LAYER	POPULATE INTO GOOGLE UNIVERSAL ANALYTICS AS	EVENT DESCRIPTORS AVAILABLE IN TAG MANAGEMENT DATA LAYER	POPULATE INTO GOOGLE UNIVERSAL ANALYTICS AS
add_to_wishlist	event category/action	recommended event	Floodlight activity or uvar
rice cooker	event label	recommended parameter	Floodlight uvar
59.99	event value	recommended parameter	Floodlight uvar
USD	custom dimension	recommended parameter	Floodlight uvar

A potential approach to parallel event tracking across two versions of Google Analytics and Floodlight.

This data-driven approach also provides a benefit for tag governance: you can reduce the number of actual event and Floodlight tags and instead leverage the variable values for segmentation and analysis in the reporting environments. This can be especially useful in regard to Floodlight, since it can avoid the proliferation of separate Floodlight activities and tags for tracking each type of user interaction.

Audience Creation on other Marketing Platforms

You can extend the parallel behavior and variable tracking approach to other marketing platforms as well and take advantage of enriched audience definition and activation. For instance, you could pass the *insurance* and *hiking* variables referenced above to Facebook Ads and Twitter Ads, thereby enriching user profiles within those platforms and allowing for better targeting.

You can also take advantage of this type of audience activation within email marketing platforms. The behavioral data, user classifications, and content classifications that you record in your web analytics (and write to you data layer) can help you better target your email campaigns – through a CDP, parallel tracking on the email landing pages, or direct activation of analytics audiences in the email platform, as In the case with the native integration between Google Analytics 360 and Salesforce Marketing Cloud integration.

Key Takeaways

- Take advantage of the same taxonomies and user interactions that you're writing to your web analytics data to also define audiences for media targeting.
- In the Google marketing tools, you can use audiences that you have defined with Floodlights and audiences shared from Google Analytics. The Floodlight-defined audiences can provide an edge in real-time bidding optimization, while the shared Google Analytics audiences can enable rich targeting based on the fuller range of the Google Analytics dataset.
- A parallel, variable-driven approach to event and goal tracking across analytics and Floodlight can provide two efficiencies: leverage the same variables from the data layer and browser environment, and reduce the number of individual tags.

First-Party Data Strategy

At the same time that we in the marketing and analytics industry are compelled to respect privacy and security of customer data, a portion of our user base – likely a growing portion – does expect organizations to remember users' previous interactions, to understand their preferences as individuals, and to offer relevant, personalized experiences.

Several of the themes already explored in this white paper – including privacy, parallel tracking and enrichment in analytics and marketing platforms, and data quality considerations for both – touch upon first-party data strategy.

Customer Data Platforms (CDPs)

A data quality focus for your web tracking can similarly benefit a CDP initiative involving web data – for instance, when you're combining web data with in-store or CRM data through either of the following approaches:

- **direct ingestion:** if you're directly ingesting a web analytics dataset from the web analytics platform into the CDP, the quality of the user, content, and behavioral data that you have maintained in the web analytics data will provide a head start in structured formatting and activation of the data in the CDP.
- **parallel tracking with a CDP pixel:** if you're capturing web activity into the CDP in parallel with the web tracking, you can take advantage of the web data layer and, in the CDP, replicate some of the measurement framework that you're using for your web analytics tracking.

While most CDPs offer some level of native data cleanup functionality, you'll benefit at least in terms of time savings by ingesting clean data to begin with.

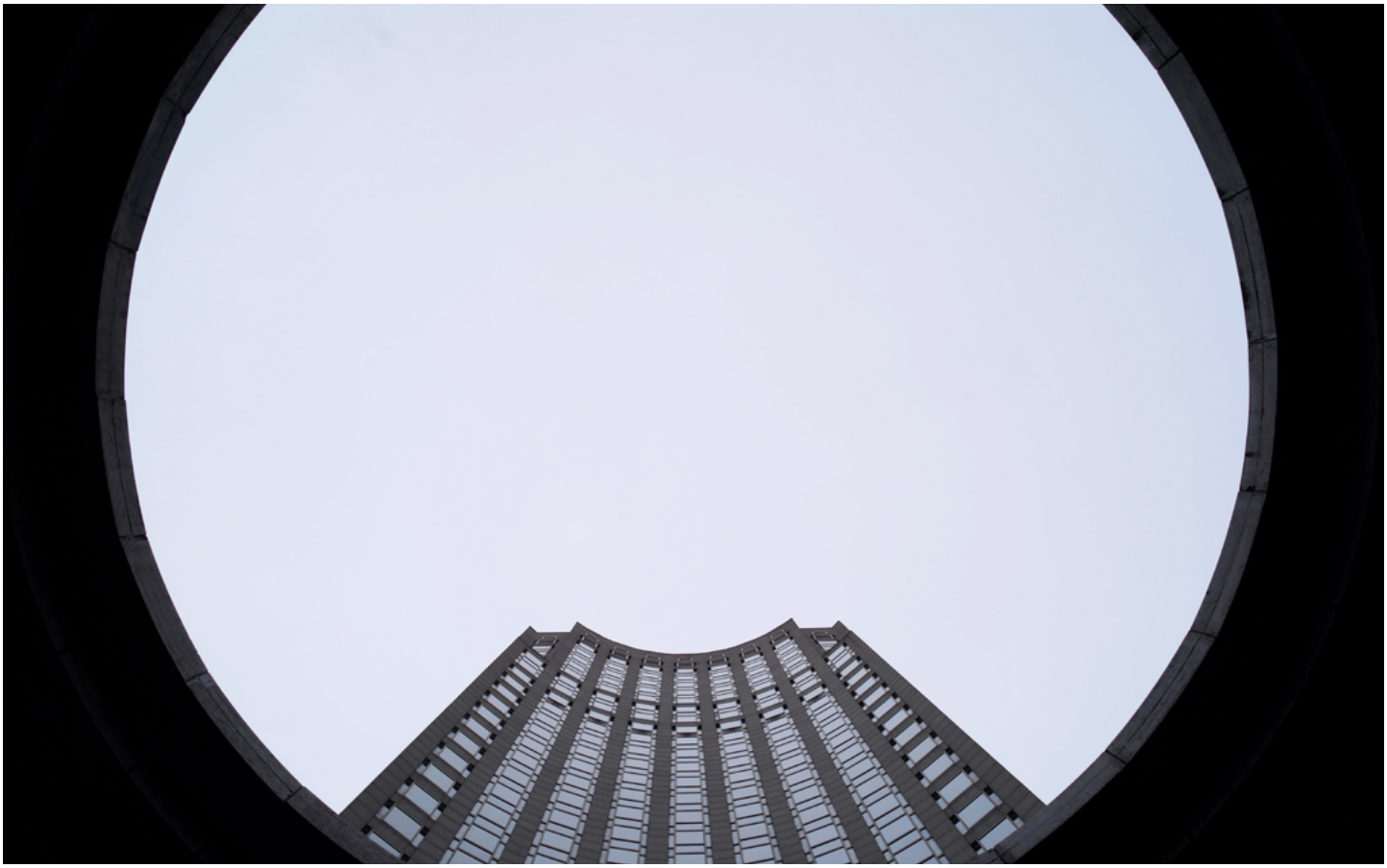
Joining Datasets

As an overarching principle of first-party data strategy (and arguably data quality), take the opportunity to proactively populate join keys into datasets that correspond to different stages and aspects of the customer journey. At a bare minimum, for authenticated users on your website, add an back-end user ID to your web analytics dataset.

In the case of Google Analytics, this user ID must be anonymous (meaning not immediately identifiable but rejoinable with individual user data outside of Google Analytics). You should populate this value into the designated userId field for integration across devices in Google Universal Analytics and also across web, Android, and iOS data streams in Google Analytics 4. Also also populate it into a Google Universal Analytics custom dimension that you define yourself, since userId is exposed in the Google Analytics BigQuery export but not directly in the reporting API.

As cookie lifespans shorten and user authentication becomes increasingly important for identifying web users across sessions and experiences, the joins between datasets will only become more critical.

As another example of an integration opportunity, you can append the ad_id dynamic parameter to your Facebook campaigns and read it into Google Analytics as a custom dimension or custom event parameter, which will thereafter allow you to analyze and model across your web analytics and Facebook Ads datasets. (Just remember to strip that ad_id query parameter out of the URL as recorded in Google Analytics.)



Don't Wait to Store Join Keys, But Be Mindful of the Authenticated Context

Don't wait till the time you need the joins to start implementing them. The time to get your joins in place is as long as possible before you need to analyze across the datasets, even if you don't yet know the exact use case for the analysis.

With that said, be mindful that you may not always have the same latitude in storing join keys for unauthenticated users. For instance, Google may not allow you to record a Google Universal Analytics client ID – which is generated for authenticated and unauthenticated users alike – as a custom Floodlight variable for joining analytics and media datasets.

Key Takeaways

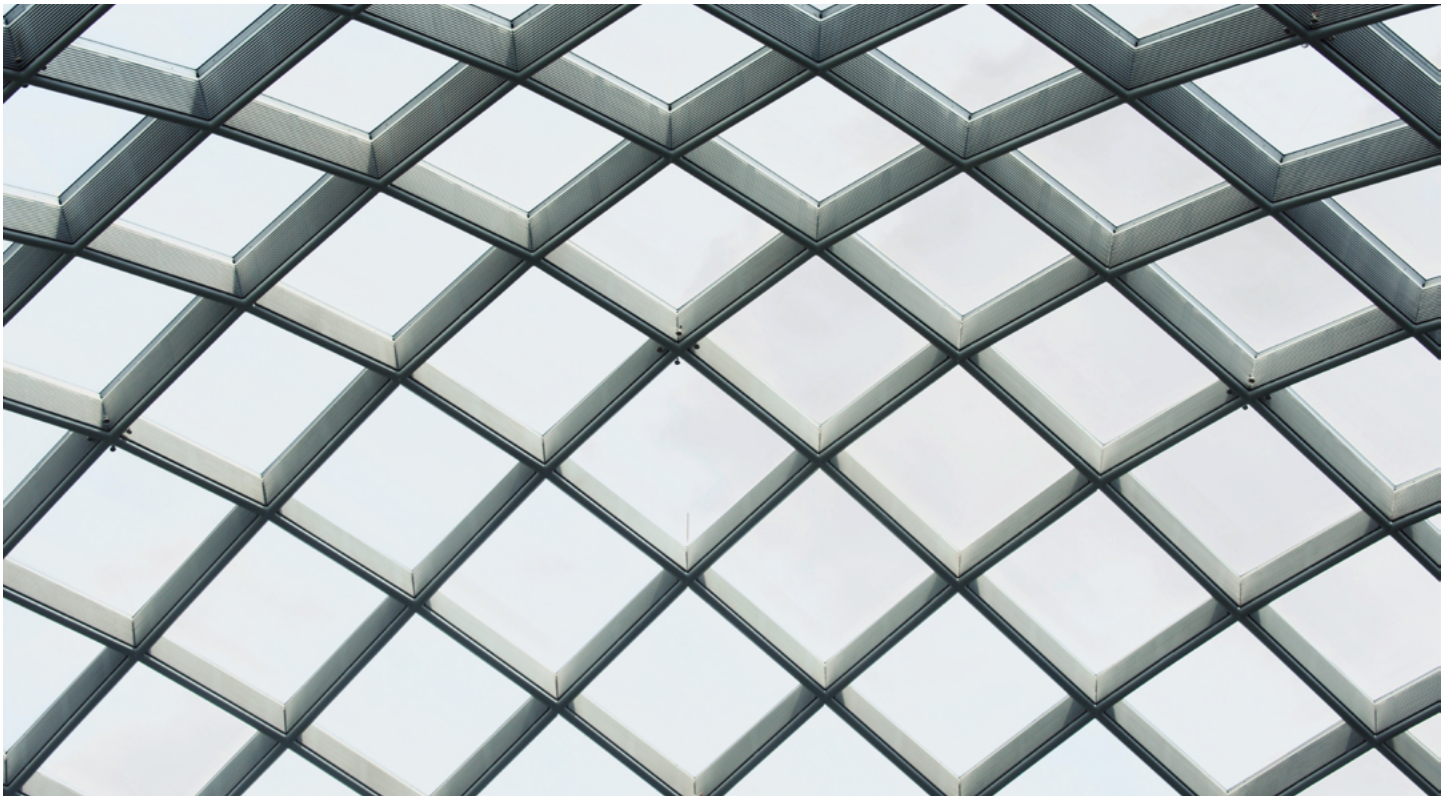
- Data quality in your web analytics data capture can benefit your effort in unifying and activating customer profiles through a CDP.
- Take advantage of authenticated user experiences to store join keys in separate customer datasets wherever possible.

A/B Testing and Personalization

As a final audience-level consideration, you can define audiences for your A/B testing and personalization tool based not only on website behaviors, but also on back-end taxonomies that you have pulled into your data layer and populated into your A/B testing and personalization tool, either directly or through sharing from your analytics platform (as in the case of Google Analytics audiences shared with Google Optimize 360).

Key Takeaways

- Audience definitions based on the clean capture of behaviors and taxonomies can also serve more reliably as targets for A/B testing and personalization.



Data Quality as a Competitive Advantage

In the coming months and years, tracking web user activity and maintaining a unified user view for analysis, marketing, and personalization will become even more challenging, with higher inherent error in the data and YOY comparison harder to perform.

In the face of these limitations, take full control where you can, and give your organization the benefit of an analytics and media strategy that places data quality front and center.

About the Author

Eric Fettman serves as Director of Capabilities and Enablement at Cardinal Path. He gives special thanks to the Analytics Implementation, Data Engineering, Analysis & Insights, Privacy/Compliance, and Media teams at Cardinal Path for their indispensible support on this white paper.