Practice papers

Effective first-party data collection in a privacy-first world

Received (in revised form): 10th August, 2021

Lucas Long
Product Manager and Privacy Specialist for Tag Inspector, InfoTrust, USA

Lucas Long is a Data Privacy Leader and Product Manager for Tag Inspector at InfoTrust, the data governance auditing and monitoring platform. As an expert on the regulatory and technical restrictions surrounding the collection of data, he advises some of the world's largest organisations on how best to optimise their marketing strategy and operations.

InfoTrust LLC, 4340 Glendale Milford Road, Suite #200, Blue Ash, OH 45242, USA
E-mail: lucas@infotrustllc.com

Abstract The analytics industry is facing an unprecedented change in the methods and requirements for data collection. These changes are a result of increasing consumer expectations regarding the privacy of personal information and shifts in the regulatory and technological methods used to meet these expectations. In today's privacy-first world, first-party data collection becomes more important than ever — while at the same time more difficult. This paper outlines core principles for first-party data collection in a privacy-focused world and offers tactical suggestions for future-proofing. These suggestions include methods to collect privacy-safe first-party and anonymous data, strategies to enable downstream integration, and ways to enforce data taxonomies, as well as compliance via a server-side data distribution approach.

KEYWORDS: data architecture, data minimisation, privacy, data collection models, cookieless data, data governance

INTRODUCTION

The environment in which marketers are working is undergoing the most rapid transformation in recent memory. With new privacy regulations being introduced across the globe, the deprecation of traditional mechanisms for tracking users, and the technical enforcement of consent via new operating systems — the range of possible responses is vast.

More unsettling yet, the options available to marketers are unclear. At the time of writing, the most widely used web browser (Google Chrome) is maintaining it will deprecate support for third-party cookies in 2023,[1] even though its Privacy Sandbox initiative to address use cases relying on those cookies is still very much in development.[2] Without the ability for platforms to embed this technology into products, how are marketers supposed to evaluate the very things that their strategies will be built around in the future?

With all of the uncertainty, it is critical to keep in mind that everyone is facing the same challenges and operating in the same environment. The technical and regulatory changes apply to everyone, and everyone has the same opportunity to be a leader in the new privacy-focused world. So, where

to start? It is important to start controlling the things that can be controlled, focusing on first-party data collection architecture and future-proofing to set the foundation for success in the new reality.

CHANGES AND IMPACTS

The changes affecting the industry fall into three main categories:

- changing consumer expectations;
- new legal restrictions introduced via new privacy regulations; and
- technology changes to enforce user privacy.

Of these three categories, the most significant is the first, namely consumer expectations. Changing consumer sentiment has driven increased willingness to regulate data practices and embed technical restrictions within browsers and operating systems. As people have generally become 'more online' over the past 20 years, the amount of behaviour that one can directly observe and then leverage for marketing and advertising purposes has increased exponentially. At the same time, advances in processing power and reductions in the cost of data storage have combined with general improvements in the way we access data to create a situation where the volume of information that one may associate with a person, as well as the volume of technologies with access to that information, has become extreme. The people being observed and providing this information, meanwhile, have largely been unaware of what is going on behind the curtain.

Beginning in 2016 with the Cambridge Analytica scandal,[3] users began to become more aware of the volume of information available about them and the ways in which that information was being used. While the majority of user vitriol has been focused on the data practices of social media networks, much of that sentiment has been applied more holistically to the advertising and marketing space. Not all people are comfortable with their habits and interests being used to generate profiles and optimise messaging meant to influence their behaviour.

As this general consumer understanding has increased, so too have consumers' expectations regarding their rights with respect to their information. On the heels of this public privacy enlightenment have followed the regulatory and technical restrictions meant to better protect users and their rights.

Privacy regulations have been in place to protect the online data of users for as long as digital advertising has been an industry. These laws were primarily limited to Europe and pertained to more traditional methods of storing and accessing information from users' devices and electronic communications such as e-mail. The most widely impactful of these original laws have been the 2009 ePrivacy Directives.[4] Passed in the EU, these are often referred to as 'cookie laws' due to their requirements for users to be informed and to consent to the placement or accessing of cookies (ie the small text files used to store information for later reference) on their devices. The main issue with the ePrivacy Directives was the country-specific nature of the requirements and varying definitions of consent from one market to the next. This all changed in 2018 when the EU General Data Protection Regulation (GDPR) came into effect.[5]

Article 4(11) of the GDPR codifies the definition of consent as: 'any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her'.[6]

The GDPR applies this definition to the requirements outlined in the ePrivacy Directive, and thus generally requires users to consent to marketing and advertising cookies before they may be placed on their device. In internal client testing, the practical impact of this new requirement has been a

40–60 per cent reduction in the volume of observed users on websites.

In the USA, the regulatory scene has been close behind. The California Consumer Privacy Act 2018 (CCPA) became enforceable in July 2020, granting California users, among other things, the right to opt out of the sale of their personal information.[7] This right was expanded following the passage of the California Privacy Rights Act (CPRA) in 2020 to include the right to opt out of the sharing of personal information.[8] Additional recent privacy laws such as Virginia's Consumer Data Protection Act 2021 provide similar rights to users with respect to opting out of data-processing for purposes of profiling and advertising.

In addition to the restrictions on first-party data collection and processing resulting from an ever-increasing volume of privacy legislation, technology platforms are also beginning to enforce users' privacy rights by deprecating traditional methods of tracking. This began with the Intelligent Tracking Prevention update to Apple's Safari browser in 2020, which effectively deprecated support for third-party cookies for approximately 25 per cent of all internet users.[9] This was followed by Google's announcement in February 2020 that its Chrome browser would deprecate support for third-party cookies by the end of 2023.[1]

Without third-party cookies, marketing and advertising platforms will lose the primary way in which they identify and track users across domains to satisfy use cases such as personalised targeting, attributing conversions to impressions and clicks across domains, and identifying user preferences for content personalisation.

As marketers lose the ability to lean on third-party tools for things like user profiling, conversion modelling and campaign optimisation, the need to extract more from their first-party data sets becomes critical. While the methods for optimising collection are changing, there are many ways to thrive in the new environment. In what follows, this paper discusses the key principles of privacy-focused data collection.

DATA MINIMISATION

Data minimisation is a concept from privacy by design[10] and has been embedded in recent privacy legislation. The UK's data protection agency, the Information Commissioner's Office, defines the principle as limiting the processing of data to only that which is necessary and not retaining more than is needed for the defined purpose.[11]

This is codified in the GDPR via both the second and the third of the six principles outlined in the legislation. The second principle states that personal data shall be 'collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes',[5] while the third principle states that personal data shall be 'adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed'.[5]

Taken together, this means that any collection and processing of personal information must be intentful. Collecting as much information as possible and then figuring out what to do with it later has always been bad practice — now that practice comes with the risk of legal action.

First, it is necessary to plan out the strategies and outcomes to be achieved. Only once those outcomes and goals are defined should one define the platforms to use and determine the data required to accomplish those defined outcomes. Any usage of personal data from users must be limited only to data points that are absolutely necessary to accomplish the task.

In the USA, this principle has been adopted in the language used in the CPRA, which updates the CCPA and comes into effect in 2023. Section 3B(3) of the CPRA stipulates that personal information should be collected 'only to the extent that it is relevant and limited to what is necessary

in relation to the purposes for which it is being collected, used and shared'.[8] Section 1798.100(c) further adds that 'personal information shall not be processed in a manner that is "incompatible" with the originally disclosed purposes'.[8]

Again, it is explicitly necessary to be strategic and intentful with any user data being collected and used. All of this behaviour — and the outcomes — must be documented and disclosed to the user so they can make an informed decision regarding consent and/or opt-out. Any processing of this information that does not align with the disclosures made to the user at the point of collection constitutes a violation of the user's privacy rights. This principle must be kept in mind when considering the data collection architecture and the downstream processing and use of collected data.

EVENT VS SESSION/USER DATA MODELS

Two core foundations at the base of many platforms have been significantly disrupted as a result of the recent regulations and privacy-focused technical restrictions. First, many platforms traditionally used a session or user-based data model. This means that the core units of measurement were either users (unique individuals visiting a website) or sessions (unique visits to the website) with all contextual information being associated with those core units. In many ways this made a lot of sense. For analytics and advertising, organisations are interested in who users are and in what ways they are interacting with various properties. By associating all behaviour data with users interacting with various digital properties and those users' sessions, analytics goals could effectively be accomplished.

New privacy legislation — and specifically the requirements for consent — have completely upended this traditional practice. This is largely due to the second core component of traditional data collection — that the tag (the javascript executing on the page to collect information and run platform functionality) was responsible for both the sending of data as well as setting and accessing cookies.

With this session/user-based data model, cookies were necessary in order to identify each user to associate his or her behaviour over time. The cookie stores a unique anonymous identifier which can be accessed across pages of the website being visited (first-party cookies) to tie together actions making up a session and then tie sessions together to provide a holistic view of the user's behaviour.

As discussed earlier, for websites in the EU adhering to the ePrivacy laws following GDPR, this placement of cookies requires explicit consent from the user. To manage this, tags would be blocked from running until the user consented to the tracking. By not loading the tags, an offshoot of this is data also not being collected. This leads to the 40–60 per cent reduction in observed user behaviour across websites observed in internal testing.

What analytics platforms, as well as many advertising technologies, are now doing is to change to an event-based data model.[12] With this type of model, all contextual information (page URL, interaction data, product viewed, user ID if they are logged in, etc) is associated with a user interaction ('event') as opposed to the user or session within which that interaction occurs. By associating all contextual data with the interaction/behaviour instead of the user, there is no need to set nor read an anonymous identifier in a cookie to associate all of the actions together. This model provides a new opportunity to collect completely anonymous interaction data on web properties.[13]

Anonymous interaction data collected via this event-based model allows for an organisation to collect stateless interaction events — such as form submissions,

transactions, add-to-cart actions and content views — in a privacy-safe manner. Collecting data in this way, without any data that may be associated with an identifiable user, maintains the ability to observe which campaigns are leading to site visits, which products/content are most popular, and overall conversion counts, even for users that have not consented to the accessing of their device or processing of their personal data. While some metrics such as conversion rates and user funnels cannot be analysed without an identifier with which to connect each event, it is at least possible to maintain a full view of the actions taken on a digital property.

In addition to the ability to collect anonymous data for all interactions and comply with consent requirements as defined in privacy legislation, the event-based data model also encourages more uniformity in the data being collected across platforms.

For the data architecture on web properties, the organisation simply needs to define what interactions it wants to observe and what contextual information about each of those interactions must be collected in order to drive their data needs. As more platforms support this model, the data structures across those platforms become uniform. This opens the door for downstream integration like never before.

ANONYMOUS DATA COLLECTION

When it comes to understanding users via analytics, the key information can be derived from the five 'W's of marketing — who, what, when, where, why: *who* are the individuals interacting with the brand, *what* are they most interested in, *when* are they interacting with different messaging, *where* are they coming from, and — based on these insights — *why* are they choosing one organisation over another? These are the insights that drive branding, positioning, targeting and design decisions.

In an environment where consent makes it impossible for platforms either to assign or access an identifier associated with a user, it is still possible to collect and utilise information to understand a majority of these needs. As discussed in the last section, the event-based data model prioritises the observation of *behaviour* rather than *users*. With this model, the collection of 'cookieless' data is used to understand what users are doing, where they are coming from and when they are interacting with different offers, products, etc.

To understand the 'who' and 'why', lean on data from users that *do* consent, as well as users who self-identify or register on the website. Internal benchmarking with partners in the direct-to-consumer e-commerce industry suggests that on average approximately 6 per cent of website users are registered, providing a persistent identifier such as an e-mail address. Those 6 per cent of registered users account for approximately 39 per cent of transactions and close to 41 per cent of e-commerce revenues. This provides a rich dataset of named users with which to extrapolate insights for the 'who' of high-value users. Beyond just registered users, users that do consent to the use of cookies allow for the placement of a pseudonymous identifier in the form of a first-party cookie to associate behaviours over time. Additional long-term value analysis can be conducted on this dataset to further segment and derive insights about the preferences of valuable visitors.

The analysis of named and consented user datasets to answer 'who' and 'why', alongside the anonymous 'cookieless' dataset to help answer 'where', 'when' and 'what' helps to make strategic decisions about how best to serve the highest-value customers and thrive in the new environment.

STANDARDISATION/DATA GOVERNANCE

As the proportion of users whose information can be collected decreases, any and all information to help understand who

the most valuable users are and how best to serve them becomes even more valuable. In addition, to fill gaps in observed behavioural data, modelling using both named datasets and anonymous datasets containing no personal data is necessary. To accomplish both use cases, the need for aggregation and integration across owned data sets is essential.

For many organisations, the prospect of integrating data is impossible. Data are siloed across both platforms and properties with the data in each silo having a completely different data model and taxonomy. This can no longer be the case. All first-party data should follow the same set of standards and definitions. Start by instituting an event-based data collection model as outlined previously. From there, standardise the taxonomy of data captured on each user interaction. This standardisation will improve the ability for downstream integration and the ability to realise the benefits of a single customer view.

Standardisation and integration have several secondary benefits as well. Privacy legislation in both Europe and the USA grants users the rights of access and deletion.[14] Specifically with the CCPA, the organisation responsible for the initial collection is responsible for ensuring requested data are deleted from other partners with whom the organisation has shared (or sold) that information. By integrating data together in one central system, it greatly simplifies the operational processes of deletion and access.

SERVER-SIDE DATA DISTRIBUTION

Traditionally, data collection for purposes of analytics as well as the functionality for advertising technology all operated from the client-side — or in the user's browser. To collect and send data to Google Analytics, for example, a Google Analytics (GA) tag would have to be implemented on the website. In addition, data would be made available for that GA javascript tag to run and send data to Google's servers, as well as set cookies on the user's device. This same concept has been used with various other digital technologies and is known as client-side data processing/collection.

While this process worked, it also opened the door for unauthorised third parties to be loaded in by other tags (piggybacking), and enabled third-party cookies to be set and retrieved at will. In addition, it slowed down the loading times of websites, and often resulted in the disparate data models and data structures discussed previously.

Server-side data distribution changes this configuration by removing the need to run third-party javascript (tags) from the website or mobile application. Instead, all interaction and user data are provided by users to the organisation's own server environment. In the most efficient and ideal scenario, this is accomplished with a single data stream to an internal endpoint. Once the data are ingested, logic can be created to identify issues or gaps in the data, recognise personally identifiable information, and identify inconsistent data taxonomies, and fix all of these on the fly. Once the data have been transformed into the proper structure, said data can then be distributed from server to server with partner third parties.

The server-side data distribution approach is much more engineering-heavy (at least today), but it promises the ability to enforce data governance standards, as well as provide ultimate control over what data go where, further supporting compliance efforts. There are already a number of platforms on the market today, most marketed as 'server-side tag management'[15] that are building out more user-friendly interfaces for this method of data distribution. It is likely only a matter of time before this becomes the standard for first-party data collection architectures.

To help understand why, the primary advantages of this approach are explored below.

Piggybacking/injection for cookie matching

Third-party cookies have long been the mechanism of choice for identifying users across domains. Cross-domain identification enabled by third-party cookies enabled cross-context profiling, identified users for programmatic targeting, and associated user actions with conversions for attribution. A key requirement for these processes has been cookie-matching. Cookie-matching is the process by which one vendor matches identifiers stored in its third-party cookies with those of another vendor being loaded on a site. The process has allowed third-party ad-tech providers to grow identity graphs by sharing IDs for wider targeting reach, more robust profiling and more comprehensive attribution. With the deprecation of third-party cookies, this process of cookie-matching is no longer possible. Without cookie-matching, the value realised by ad-tech platforms from injecting/piggybacking other tags through their client-side javascript tags goes away. Control over which platforms receive any available identifiers is shifted to the digital property owner. Server-side data distribution is the more efficient way to exercise such control.

Ease of implementation

It is far less engineering-intensive to place a tag on a webpage for client-side processing than to configure a data feed to distribute data via application programming interface endpoints with a server-side distribution approach. From a direct cost perspective, this is a large benefit of client-side processing. However, technical limitations on processing data from the client-side (such as restrictions on first-party cookies set via client-side javascript and the rise of consumers using ad-blockers) begin to place additional indirect costs that weigh against the benefit of implementation ease.

While the costs associated with client-side processing are rising, technical barriers to standing up a server-side data distribution architecture are falling. Vendors are introducing server-side tag management technology to simplify the configuration process while analytics and advertising platforms are building server-side endpoints to allow for data transfer from server-side tag management platforms. Combining these factors and a server-side data distribution setup begins to compete with the simplicity of client-side tag deployments. To be clear, server-side is more engineering-intensive and processing costs are transferred to the business (more on this in the next point), but the additional data quality benefits and ever-improving ease of implementation begin to even the scales between the client and server-side processing approaches.

In-browser processing

Client-side tagging handles the processing of data from a page and then sends the data to a third-party platform all within the browser on the user's device. This means the processing cost is assumed by the user. With a server-side approach, the cost of this processing is transferred to the business. This will never change — there will always be more cost for an organisation with server-side data distribution vs the traditional client-side approach. What is changing, however, is the value associated with the benefits — more data integrity, less risk of data degradation due to technology changes, and full control over what data are sent to what platforms, providing significant compliance and regulatory value.

As the benefits of client-side data distribution via javascript tags continue to degrade, so too do the costs of server-side data distribution. Combine this shift with the significant governance benefits realised by fully controlling first-party data, server-side tag management is the

privacy-first approach to data collection and distribution.

OPTIMISE FIRST-PARTY DATA CAPTURE

In the privacy-centric environment two types of data are critical: observed behaviour of consenting users and data from registered/named users. Optimising the proportion of users in these two categories is foundational for audience creation and measurement in the absence of third-party identifiers.

First is increasing the proportion of consenting users. It is important to keep in mind the regulatory restrictions on using design techniques such as 'dark patterns' to influence the selection choice of a user. These 'dark patterns' include using different colours or sizes for accept/decline buttons and modifying the placement of buttons in a banner to prioritise one over the other. What can be tested, however, are techniques to encourage a decision to be made. This includes testing language explaining how the user's information will be used and testing banner locations on a page. Optimising the transparency of data practices and the user's consent experience can lead to a larger proportion of consenting users to aid in campaign measurement.

Following user consent, maximising the proportion of named (or registered) users is also critical. User-provided persistent identifiers such as e-mail address enables privacy-safe audience creation and measurement techniques through clean rooms and partnerships with direct partners (where users have also consented and registered). Registration is also an indication of consumer trust, which is the first step in building an ongoing relationship with the user. Techniques such as moving registration further up the user funnel, providing compelling offers and incentives for registration, and creative exchanges of value can help drive registrations as a conversion goal.

PRIVACY-CENTRIC FIRST-PARTY DATA COLLECTION ARCHITECTURE

Pulling all of these principles together, the first-party data collection architecture of the future becomes more clear. It all starts with the definition of the organisation's strategy and goals. Cascading from these decisions, the platforms and data needs to accomplish those goals are defined. Privacy is at the heart of the strategic definition process and principles such as data minimisation are followed.

Once the data needs are defined and documented, those needs are translated to an event-based data model for the collection architecture across all digital assets. Taxonomies and governance standards are created and enforced to ensure standardisation across properties as well as platforms. The defined data framework is critical for flexibility as new partners are onboarded over time.

Following the definition and documentation of the standardised event-based data model, the server-side data distribution architecture is then designed and implemented. To do this, all third parties with whom data will be shared, along with their data needs, are outlined. The single data stream containing event-based data is configured to send interaction data from the client to the server environment upon each tracked user interaction. Transformation and distribution rules are defined in the server-side configuration to send required data in standardised formats to the defined third parties being worked with, along with an internal first-party data warehouse.

Implemented alongside this event-based data architecture is a system to manage the consent preferences of users. For users who consent, set a first-party identifier to associate actions and preferences over time to those users for analysis and

personalisation. For users who do not consent, collect anonymous interaction data to understand behaviour on the site and provide more context for modelling and filling in data gaps in collected datasets.

Instituting a privacy-first data collection architecture such as this will allow an organisation to optimise the use of owned data to thrive in an environment where third parties can no longer be relied upon to accomplish critical business goals.

CONCLUSION

Today's privacy-focused landscape presents unique challenges and requires adjustments to the traditional practices marketers have come to rely upon. While this represents a large change, the change is both necessary and for the benefit of all. The third-party technologies used to accomplish business goals should not be a source of competitive advantage. Instead, organisational capabilities and the relationships between the business and consumers should differentiate the winners and losers. These privacy-focused changes make this ideal a reality.

Both privacy and performance are able to coexist. The needs of the business can be accomplished in a privacy-first manner. It starts with a first-party data strategy and intentful architecture. These privacy-first principles of data collection provide the foundation for success now and into the future.

References

1. Schuh, J. (2020) 'Building a more private web: A path towards making third party cookies obsolete', available at: https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html (accessed 5th June, 2021).
2. Privacy Sandbox (2020) 'Building a more private, open web: Privacy Sandbox', available at: https://privacysandbox.com/ (accessed 14th June, 2021).
3. Confessore, N. (2018) 'Cambridge Analytica and Facebook: the scandal and the fallout so far', available at: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html (accessed 15th June, 2021).
4. European Commission (2009) 'Directive 2009/136/EC', available at: https://edps.europa.eu/data-protection/our-work/publications/legislation/directive-2009136ec_en (accessed 10th June, 2021).
5. European Commission (2016) 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016', available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed 11th June, 2021).
6. Information Commissioner's Office (2018) 'What is valid consent — UK ICO', available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/consent/what-is-valid-consent/ (accessed 11th June, 2021).
7. State of California (2018) 'SB-1121 California Consumer Privacy Act of 2018', available at: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121 (accessed 11th June, 2021).
8. State of California (2019) 'Amendments to the California Privacy Rights and Enforcement Act of 2020,Version 3', available at: https://oag.ca.gov/system/files/initiatives/pdfs/19-0021A1%20%28Consumer%20Privacy%20-%20Version%203%29_1.pdf (accessed 11th June, 2021).
9. Wilander, J. (2020) 'Full third-party cookie blocking and more', available at: https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/ (accessed 14th June, 2021).
10. Cavoukian, A. (2011) 'Privacy by design, the 7 foundational principles, implementation and mapping of fair information practices', available at: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf (accessed 7th November, 2021).
11. Information Commissioner's Office (2018) 'ICO UK — Principle: Data minimization', available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/ (accessed 11th June, 2021).
12. Schoenbaum, D. (2017) 'Building and leveraging an event-based data model for analyzing online data', available at: https://towardsdatascience.com/building-and-leveraging-an-event-based-data-model-for-analyzing-online-data-c166c523fe6a (accessed 14th June, 2021).
13. Google (2021) 'Consent Mode (beta)', available at: https://support.google.com/analytics/answer/9976101?hl=en (accessed 14th June, 2021).
14. Information Commissioner's Office (2020) 'ICO UK SARs Guidance — October 2020', available at: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/10/blog-simplifying-subject-access-requests-new-detailed-sars-guidance/ (accessed 11th June, 2021).
15. Google (2020) 'An introduction to server-side tagging', available at: https://developers.google.com/tag-manager/serverside/intro (accessed 14th June, 2021).