Heap

Rethinking Data Governance

Harnessing data at scale in the new era of product analytics

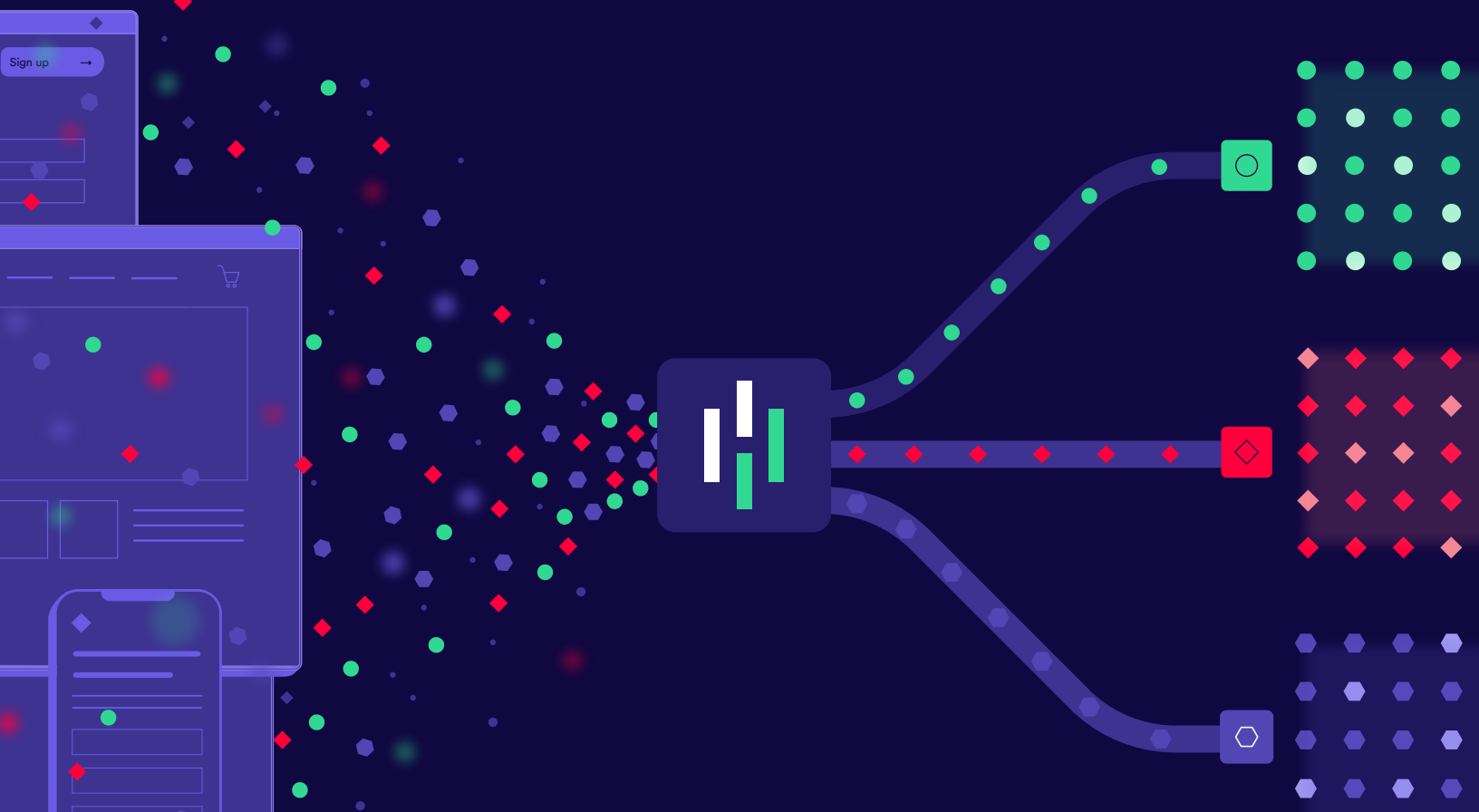


Table of Contents

Introduction 3

What is data governance? 4

Twin pillars of analytics success 5

Cracks in the manual tracking facade 6

A new approach to data governance: precision without limits 7

The lifecycle of a digital event 9

Conclusion 12

Introduction

Data governance for the new era.

Data governance doesn't get nearly enough love. Sure, everyone's polite to its face. But behind its back, the gossip begins.

It's seen as esoteric. Overly technical. Theoretical. "That thing that IT cares about." "That thing that gets tacked onto the end of RFPs." "Why did it even get invited to the party in the first place?"

That needs to change. We wrote this ebook because we perceive an urgent need to create a more inclusive and informed discussion around data governance: what it is, why it matters, and how to do it right.

For too long, the misconception has persisted that there's a tradeoff between these objectives: that keeping data organized and well-governed means keeping it limited.

We believe that teams should have access to all the data they need to deliver exceptional digital products and experiences. This requires teams to be able to structure a clean dataset for analysis while maintaining all of the rich underlying data. And it highlights the need for a complete, built-in toolset to keep data organized, accurate, and verified throughout the lifecycle of an analytics project.

In other words, it requires a new approach to data governance.

Generating meaningful insights for digital analytics requires two things:

01. The **right data** (a dataset that is rich and complete in the context of any question the team might want to answer)

02. Organized in the right way (a rigorous approach to data governance throughout the lifecycle of a digital event)

What is data governance?

At a high level, data governance covers the collection of strategies that an organization uses for collecting, managing, securing, and getting value from data.

This ebook focuses on one particular type of data: information related to user interactions with digital "products" (e.g., websites, mobile apps, cloud applications). But many of the principles here can be generalized to other data types.

The goals of a data governance approach typically include the following:

- **Accuracy:** ensuring that the data foundation used for analytics is reliable and trustworthy (e.g., up-to-date, low error rate)
- **Organization:** ensuring that data is structured, labeled, and stored in a way that promotes consistency (e.g., events are defined the same way across different teams), usability (e.g., consistency of naming conventions), and discoverability (stakeholders can readily consult documentation to see what exists)
- **Security:** ensuring that the data is handled in a way that complies with relevant regulations, respects consumer privacy, and minimizes the risk of security issues (data leakage, unauthorized access)

Strategic Goal:	Key Questions to Define:
Accuracy	• What steps do we have in place to maximize accuracy of data capture? • How do we continually monitor and audit our data for up-to-dateness and accuracy? • If we discover data accuracy issues, what does our remediation plan look like?
Organization	• How do we ensure that we have a unified set of definitions in our data model? • What is the approval workflow for adding new data definitions? • What naming conventions do we use to promote consistency and discoverability? • How do we maintain documentation?
Security	• Who within the organization will have the ability to access different types of data? • How do we ensure compliance with relevant consumer privacy regulations (e.g., GDPR and CCPA) in our handling of data

So... why does data governance matter?

The twin pillars of analytics success

Using data to deliver better digital experiences requires two things.

First, a team must have access to a dataset that helps them answer the most important questions they have about their users' digital behavior. Second, a team must have a dataset that's clean, trustworthy, and up-to-date.

Why are both of these capabilities so critical? Imagine an e-commerce site that wants to understand dropoff in the conversion funnel. The marketing manager suspects that cart abandonment might be related to a bad user experience when attempting to add a promo code to a purchase.

What are the consequences for the organization of **incomplete** or **badly-governed data**?

Consequence of incomplete data

Since the "add promo code" event wasn't designated ahead of time for tracking, there's no way to substantiate the hypothesis. The team has to update the tracking plan to add this event — then implement the code, and wait weeks for data to flow in.

Data doesn't exist to answer questions.

Consequence of badly-governed data

The team has been collecting data related to the "add promo code" event. But the relevant data seems to be split across multiple event labels ("add promo," "add discount code") — and on some dates there seem to be more promo code events than there are actual purchases, suggesting duplicates.

Can't trust the data that does exist.

Ultimately, data completeness and data governance are two sides of the same coin: criticaa co-determinants of the success of any analytics initiative.

Cracks in the manual tracking facade

The two dominant approaches to data capture for digital product analytics are “manual tracking” (used by vendors like Google Analytics, Amplitude, and Mixpanel) and “automatic capture” (sometimes also referred to as autotrack or retroactive data capture) approach used by vendors like Heap.

Manual tracking, the older approach, requires engineers to insert tracking code into each event. Data is collected from the time of instrumentation onward, and any event that is not explicitly tracked does not collect data for analysis.

Automatic capture, the more modern approach, requires only that a single Javascript snippet be inserted into the header of a site or application. After that, all event activity is tracked automatically: every click, swipe, form fill, pageview, and more.

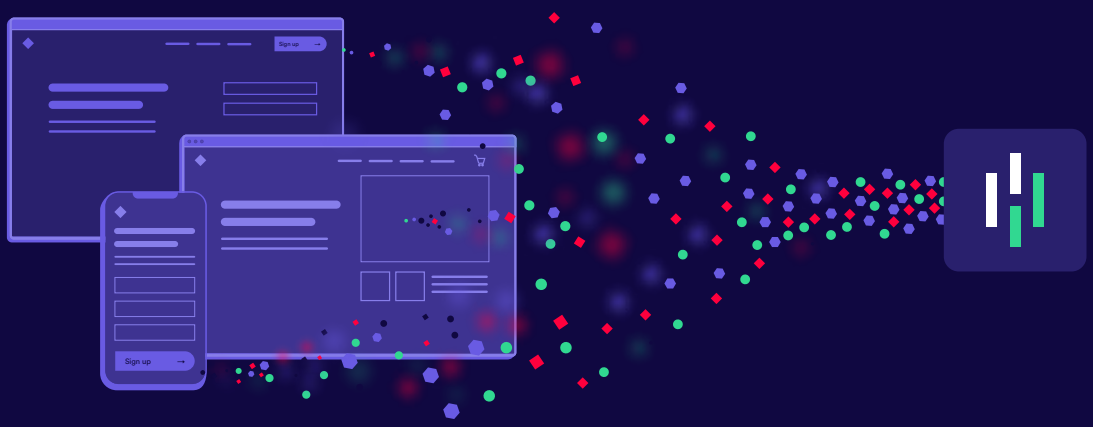
For most manual tracking vendors, the solution for data governance is something called a “tracking plan.” This is a spreadsheet — maintained outside the analytics tool — that lists all the events you plan to track, where in the codebase those events are located, what you’re naming them, their current status, the properties you’re collecting alongside them, and a host of other information.

One obvious challenge with this approach is that it **moves governance** outside of the analytics tool. The platform itself isn’t the team’s source of truth — a spreadsheet is. A product owner might use an outdated naming convention when she adds a row to a spreadsheet or creates a JIRA ticket. An engineer might make a typo when instrumenting manual tracking code, or an analyst might mistake two similarly named events in the analysis she does. These types of challenges are almost impossible to avoid. And when not noticed and remedied, they nudge the dataset closer and closer to chaos.

But even more problematically, this strategy of data governance is predicated on **keeping the dataset small and consistent**. The idea is that by choosing a small, focused number of events to track, you can maintain consistency in naming and definition and keep track of problems like broken and outdated events.

What does this mean? It means that you can keep your data governed… if you keep it limited. Organization and precision come at the expense of access to the full range of data that the team might need to answer questions.

(If I only have three books on my shelf, it’s pretty easy to keep them organized, and I’ll definitely know where all of them are. But it also keeps my reading options a little… restricted.)



A new approach to data governance: precision without limits

Unlike manual tracking, automatic data capture is built to seamlessly capture all interactions and behaviors from the time of initial installation onward.

Sounds great, right? Well, talking to manual tracking vendors, one might come to the conclusion that data completeness comes at the expense of data governance. It's not uncommon to hear such vendors attempt to sow concerns about automatic capture:

- "Manual tracking ensures that everyone in the organization is defining events in the same way."
- "With autocapture, you'll be getting a flood of raw data. With manual tracking, you'll know exactly what we're getting."
- "With autocapture, what's to stop a team from using data that hasn't been verified in a critical analysis?"

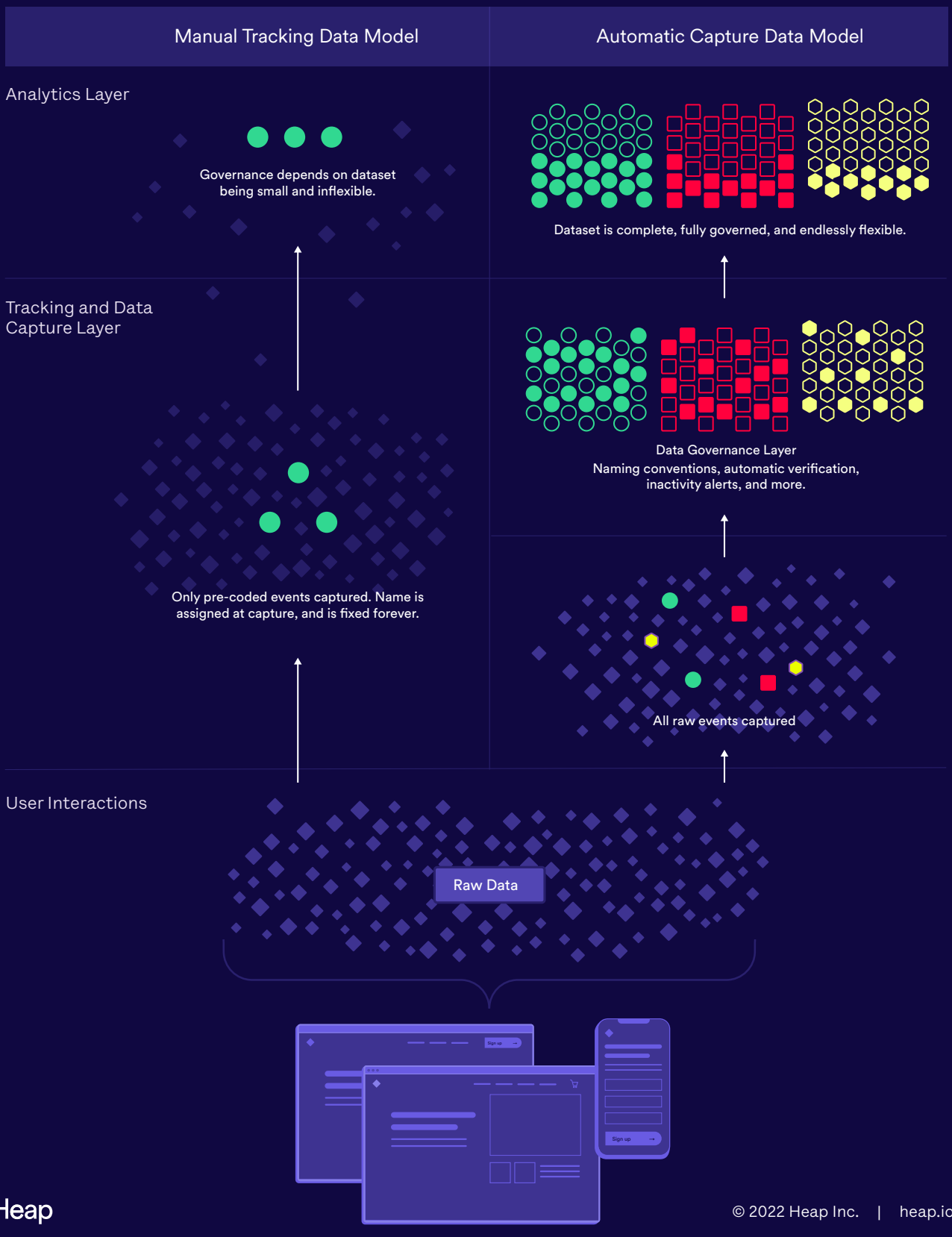
If these fears were merited, this would certainly be a big blow to the usefulness of automatic data capture. Luckily, these concerns just show a lack of familiarity with how innovative automatic data capture vendors like Heap approach the critical challenge of data governance.

The manual tracking approach to data governance, as we've seen, relies on the *data capture* process itself to enforce standardization around how events are defined.

In contrast, the most sophisticated vendors using automatic capture technology have invested in an entirely *new layer of data organization and governance on top of all the raw data*. This technology layer enables teams to organize and govern the underlying data, while still gathering everything.

The diagram below demonstrates the key differences between these approaches. A dedicated technology layer enables teams to build and govern datasets for analysis — while still maintaining the ample, context-richdata underneath. The result is precision, without limits.

What exactly does a best-in-class data governance layer for automatic data capture look like?



The lifecycle of a digital event

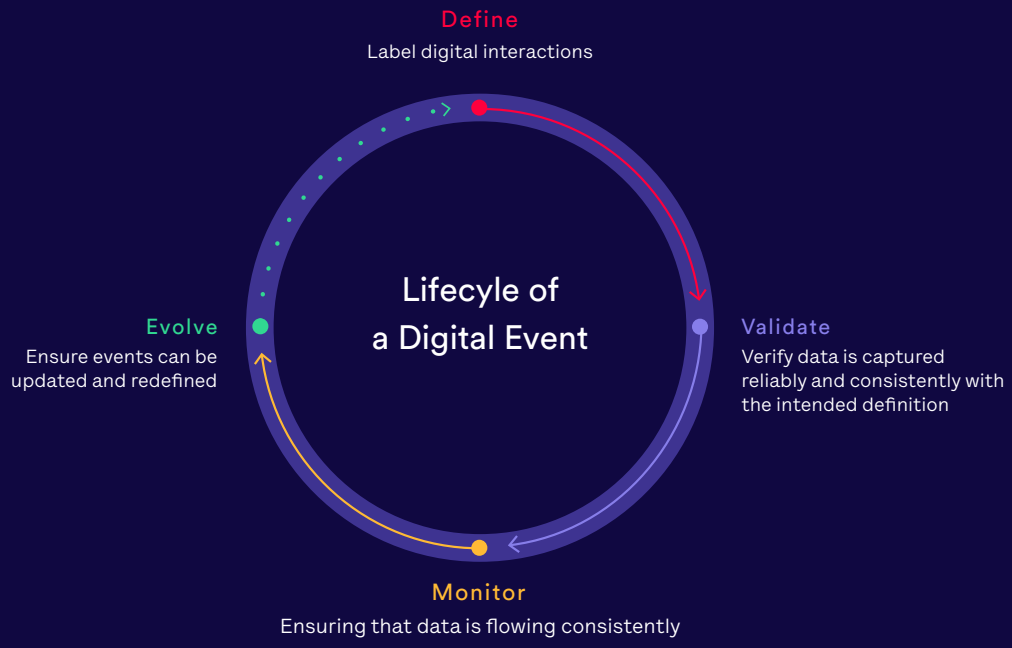
One useful framework for thinking about what outstanding data governance for digital experiences looks like is the ***lifecycle of a digital event***.

An "event" is a useful way of capturing a user's interaction with a digital property at the most granular level possible. It might include navigating to a page, clicking a button, filling out a form, or more.

Events are the building blocks of modern digital product analytics. So when comparing data governance approaches, it makes sense to consider the full value chain of activities that must be undertaken to transform them into useful insights to power decision-making.

In our experience, this chain involves four steps (see the diagram below):

- **Defining events** is the process of labeling digital interactions with names, context, and meaning.
- **Validating events** is the process of verifying that data is being captured reliably, in a way that's consistent with the intended definition of the event
- **Monitoring events** is the process of ensuring that data is flowing consistently (without interruption, gaps, etc.)
- **Evolving events** is the process of ensuring that events can be updated and redefined



© 2022 Heap Inc. | heap.io

The most sophisticated automatic data capture platforms, like Heap's, offer data governance tools that enable teams to harness all the data they need at every stage of the event lifecycle.

Let's go through each step in the lifecycle of a digital event to see what this looks like in practice.

Event lifecycle stage:	Example:	What manual tracking looks like:	What best-in-class automatic capture data governance looks like:
Define	A B2B HR and benefits management platform is trying to identify what events to label to understand why managers are dropping off when trying to complete quarterly performance reviews for their direct report	**High-stakes. Limiting.** Data is only available from implementation of tracking onwards. Event definition takes place within a manual schema planning exercise. Teams must decide in advance what events they want to track and implement the tracking in the codebase — a high-risk decision that needs to be made early in the data governance process. Naming conventions are enforced in planning and alignment meetings and maintained in a spreadsheet.	**Risk-free. Infinitely scalable.** Unlike in manual tracking, data capture is separated from data *definition*. This means that once an initial tracking snippet is implemented, **all data is automatically collected.** Data can be defined and labeled through "data virtualization": a layer that enables users to structure the dataset used for analysis. The platform itself is the source of truth on event definition: • A data dictionary gives your team a single source for all product data, including events, properties, categories, and user segments. • Naming conventions set a structure for naming that everyone is forced to use, providing consistency across the data set

Event lifecycle stage:	Example:	What manual tracking looks like:	What best-in-class automatic capture data governance looks like:
Validate	A loan originator is creating a new event to capture supplemental income information that a user can provide during the application process — but needs to validate that the event is instrumented correctly.	**Error-prone. Time-consuming.** Event verification is done manually, through spreadsheets and meetings. QA requires time and manual tracking is typical at the mercy of release cycles.	**Bulletproof. Instantaneous.** When someone creates a definition, it is automatically submitted for verification by data administrators. Administrators can inspect the definition and make any needed changes or annotations before verifying. Built-in tools provide the ability to query the data associated with events (e.g., to validate data volumes) and repair and archive events.
Monitor	A B2B SaaS company is trying to understand why data related to key in-app behaviors has stopped flowing.	**Manual. Cumbersome.** Event activity must be manually monitored alongside tracking plans.	**Automated. Intelligent.** Inactivity alerts (e.g., when an event stops receiving data) provide instant visibility when something doesn't look right.
Evolve	An e-commerce company needs to update its event definitions to reflect an additional "security check" step in the checkout process after they migrate to a new platform.	**Rigid. Inflexible.** Tracking must be manually updated to reflect changes in event definition. And as new pages, features, and elements get added to the site or app, teams must constantly remember to add new tracking code.	**Dynamic. Iterative.** Event definitions can easily be updated in the virtualization layer to reflect new pages, features, and elements added to the site or app. Additionally, virtualization seamlessly supports changing digital product specifications, hypotheses, or analytical needs. For example, different business teams in the same enterprise may need to define events differently depending on their specific context. A virtualization layer enables events to be dynamically redefined for different needs — without any additional implementation effort.

Conclusion: data governance for a new era of product analytics

Ultimately, the twin pillars of success for any analytics project are data *completeness* and data *governance*. In order to make data-driven decisions, teams need access to the right data to answer critical questions (including questions they wouldn't have known to ask in advance!). And unless the dataset is well-governed, no team is going to be able to rely on it for important decisions.

For too long, the misconception has existed that manual tracking approaches "have the market cornered" on reliable, accurate, and trustworthy datasets. But keeping datasets well-organized by keeping them limited is no longer a viable strategy for data-driven teams.

Automatic data capture offers obvious data completeness advantages by providing a comprehensive, retroactive dataset — without any ongoing schema planning or manual implementation.

And data governance for automatically captured data requires two things: 1) the use of a data 'virtualization' layer to construct a clean dataset for analysis while maintaining all of the rich underlying data; and 2) a set of built-in tools to keep data organized, accurate, and verified from the moment an event is defined. In short, the context-rich, high-volume data available through automatic capture requires that data governance be a core architectural principle — reinforced through design decisions and features that span the lifecycle of a digital event.

A complete and well-governed dataset is the key to answering the questions that the team knows to ask today — and discovering the unexpected insights that will lead to extraordinary digital products and experiences tomorrow.

Heap

About Heap

Heap's mission is to power business decisions with truth. We empower product teams to focus on what matters — building the best products — not wrestling with their analytics platform. Heap automatically collects and organizes customer behavioral data, allowing product managers to improve their products with maximum agility.

Visit heap.io to learn more.