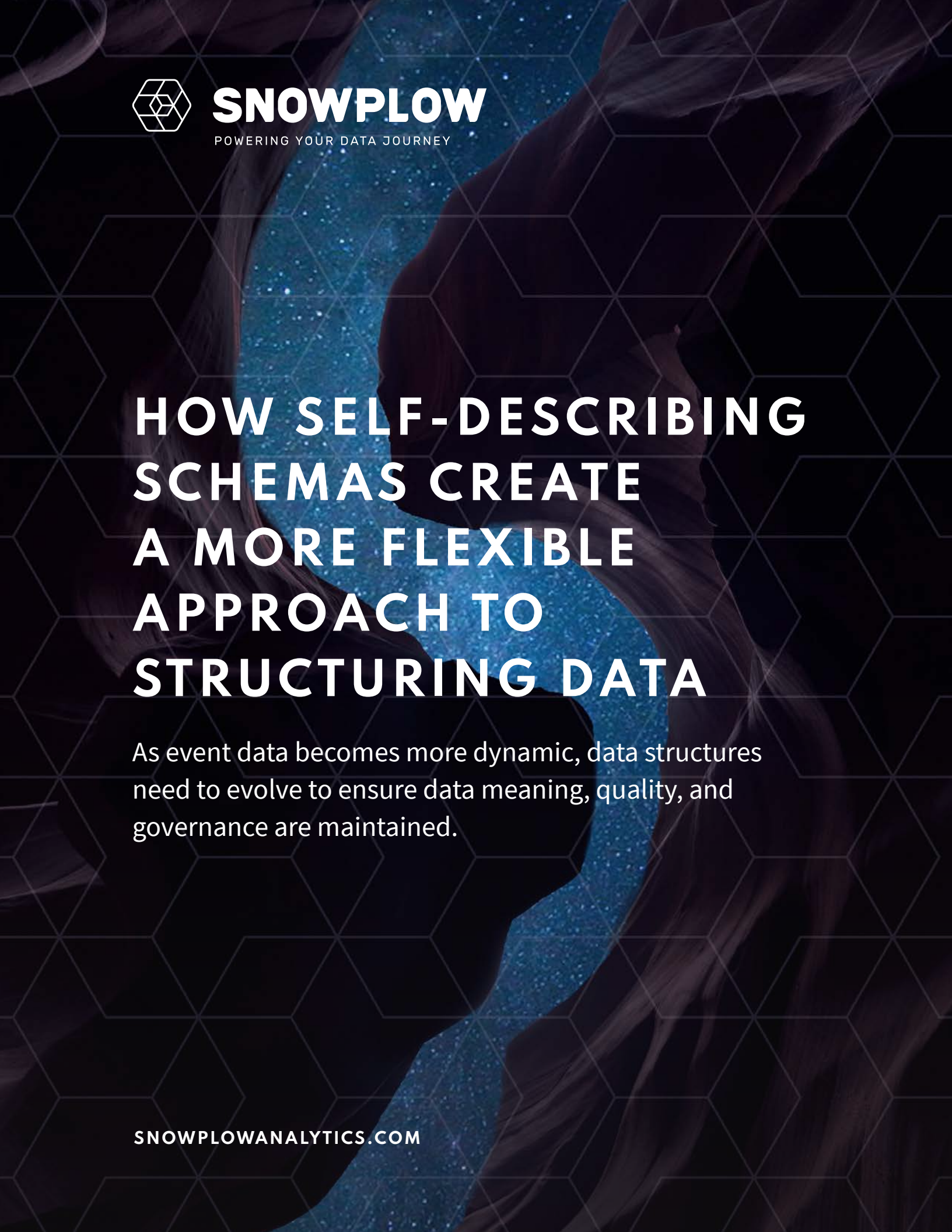


SNOWPLOW

POWERING YOUR DATA JOURNEY

HOW SELF-DESCRIBING SCHEMAS CREATE A MORE FLEXIBLE APPROACH TO STRUCTURING DATA

As event data becomes more dynamic, data structures need to evolve to ensure data meaning, quality, and governance are maintained.

SNOWPLOWANALYTICS.COM

SPEEDREAD:

DATA STRUCTURES IN 60 SECONDS

- For behavioral data to be valuable and useful to a business it must have meaning, quality, and governance.
- How data is structured can severely impact these three critical success factors.
- Traditionally, there are two ways of collecting this data: structured, and unstructured.
- Structured data is presented in an expected format for easy loading into relational databases. However, it's difficult to evolve your tracking and maintain data meaning with a fixed schema.
- Unstructured data provides the flexibility to capture whatever data the business needs, although data quality can be impacted if properties are named differently across events.
- No two businesses are the same, which means data collection should be flexible to meet each organization's unique needs.
- By blending a structured and unstructured approach to create a hybrid model, data mature organizations can benefit from the best of both worlds and derive greater value from their data.

Why is structuring your data in the right way important?

Using data as a basis for competitive advantage is nothing new. According to Forbes, 94% of enterprises say data and analytics are important to their business growth and digital transformation. Furthermore, 65% of global enterprises plan to increase their analytics spending in 2020.

However, the way that organizations structure their data is important if they're to extract the maximum value from it. There are three important factors to consider for data to be valuable and useful to a business:

1. Data meaning

Does the business understand the data?
There are a number of ways in which data meaning can be lost. For example, different teams may use different tools to capture data, which means that there is no single source of truth around what that data means. It's also possible for separate functions to interpret data in different ways and draw their own – sometimes conflicting – conclusions. Furthermore, data consumers may not even understand what they are looking at, which makes the data essentially useless.

Organizations can only use data to improve key areas of their business if they understand what the data means. And that meaning needs to be consistent across the organization.

2. Data quality

Does the business trust the data and therefore the insights derived from it?
It stands to reason that a greater emphasis on data means a greater reliance on how accurate and trustworthy that data is. Every line of missing or inaccurate data can have important consequences for the business and once trust in the data is lost it's almost impossible to win back.

For example, it's widely understood that personalizing the user experience often generates higher levels of engagement and better conversion rates. But if you're relying on inaccurate or incomplete data, it's not possible to confidently deliver a personalized experience. Rather than demonstrating an understanding of your users, you may be inadvertently driving them away by serving up irrelevant content or offers.

3. Data governance

Can the business control who can collect or use what data, and how?
Data governance is important because businesses want to control how data is collected, who can collect it, and who can access it.

For example, personally identifiable information (PII) needs to be accessible to someone working in customer success so they can effectively serve a customer. However, someone in product is unlikely to need a level of detail that includes user names or email addresses; aggregated data sets with anonymized data are usually sufficient.

Data governance also ensures that the organization is compliant with data protection legislation. Non-compliance of PII carries heavy fines of up to 4% of a company's yearly revenue. Aside from legal obligations, companies will want to ensure that any commercially sensitive data is appropriately managed to reduce the threat of a breach.

Data meaning, data quality, and data governance are all impacted by the structure of the data.

Why this matters to the head of data

While the above factors are the broad reasons why data structures are important to the business, the head of data will naturally pay closer attention. After all, the head of data will be the person ultimately responsible for recommending and implementing a data strategy that gives the business the data it needs.

For example, a head of data is likely to look for a data solution that offers tracking flexibility, rather than be limited to a certain set of events or data structures that are defined by the data collection tool. A head of data wants data structures that accurately reflect user behavior and how users interact with products – not be confined by the rules of the data tool they are using.

That's why having the freedom to collect data based on what users are doing, what the product looks like, and how that data is to be used is so desirable to data experts. The data collection tool does not dictate data meaning or structure. The data is the hero; not the tool.

What are the traditional ways of structuring data?

Event data is an action taken by a user or service, such as clicking a button, viewing a page, buying a product, or logging in on a mobile app. Traditionally, there are two ways of collecting this data: structured and unstructured.

Structured data pros and cons

Packaged analytics platforms, such as Google Analytics (GA), follow a structured, or fixed, schema approach. Here, a finite number of columns is specified and filled for each event:

Timestamp	User ID	Category	Action	Label	Property	Value
2020-01-01T 13:13:12.537Z	32838cc4-1acd-4d08-b4b2-ee16fc4d9e3e	home	click			
2020-01-01T 13:13:23.410Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	home	tracked_snowplow	Blog		
2020-01-01T 13:13:10.162Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	blog	pageview			
2020-01-01T 12:14:18.775Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	blog	click	Structuring event data		4
2020-01-01T 13:14:48.193Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	blog	pageview	Structuring event data	Cara Baestlein	
2020-01-01T 13:15:32.125Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	blog	click	Structuring event data		25

Figure 1: *example of data collected within a fixed schema.*

The benefits and drawbacks of this approach are:

Pros:

- Data arrives in the data warehouse in a highly expected format. A fixed schema uses a finite number of columns. In GA's case they're named category, action, label, property and value and apply structure to what can be captured with each event
- Because the data arrives in an expected format it's much easier and quicker to start analysis. There is no need for data analysts to spend time cleaning or structuring the data. Instead, they can begin extracting insights and building visualizations and dashboards with their preferred BI tools.

Cons

- There are a finite number of columns within a fixed schema, which limits the data that can be collected.
- While it is possible to add custom properties, the description of what those series of numbers and letters actually mean needs to be captured in a separate document. This can lead to a 'sprawling event dictionary' whereby it's difficult to understand data meaning, and the same data can mean different things to different people.
- They are difficult to evolve, as there is a requirement to update the original definition of what that category means for that type of event on that specific page.
- Data governance is problematic, as there is no way for data consumers to guarantee that the data they need will be captured.

Unstructured data pros and cons

The easiest way to capture event data is in the form of unstructured JSON. Here, we still have columns within the database, but the user is able to define them and collect richer information.

Timestamp	User ID	Custom event data
2020-01-01T 13:13:12.537Z	32838cc4-1acd-4d08-b4b2-ee16fc4d9e3e	{ “event”: “pageview”, “pageName”: “home” }
2020-01-01T 13:13:23.410Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	{ “event”: “click”, “buttonName”: “Blog” }
2020-01-01T 13:13:10.162Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	{ “event”: “pageview”, “pageName”: “Blog” }
2020-01-01T 12:14:18.775Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	{ “event”: “click”, “position”: 4, “article”: “Structuring event data” }
2020-01-01T 13:14:48.193Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	{ “event”: “pageview”, “articleName”: “Structuring event data”, “author”: “Cara Baestlein” }
2020-01-01T 13:15:32.125Z	32838cc4-1acd-4d08-b4b2-ee26fc4d9e3e	{ “event”: “scroll”, “percentage”: 25, “article”: “Structuring event data” }

Figure 2: *example of data collected within an unstructured JSON.*

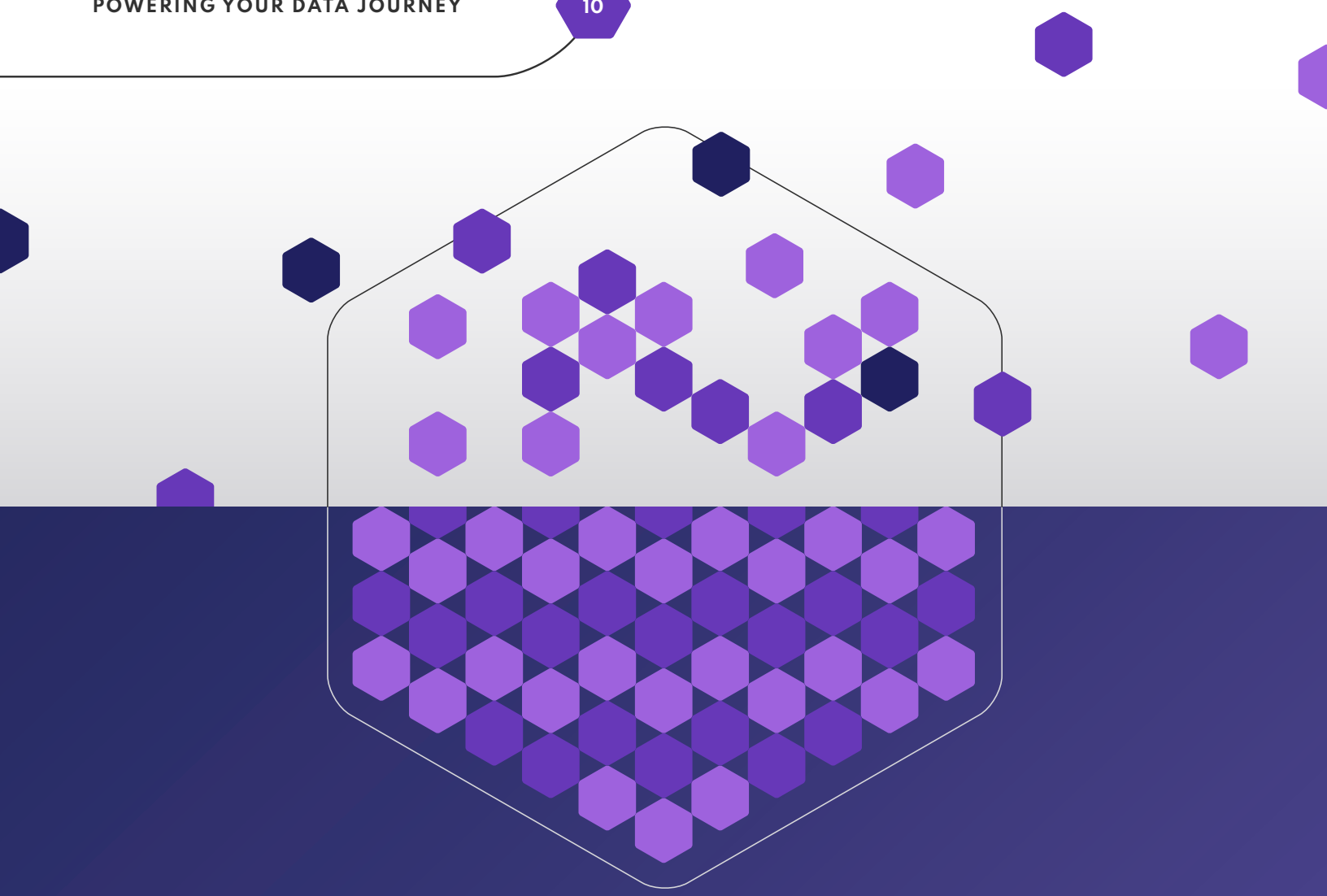
The benefits and drawbacks of this approach are:

Pros:

- Data teams have the freedom to capture whatever custom event data the business requests. They are not restricted by a fixed schema approach with a finite set of columns.
- This freedom provides greater flexibility and agility to update tracking design over time. Data collection can evolve as new products or use cases are introduced.
- Field names are not fixed but descriptive, which makes it easier for data consumers to understand what the data means.

Cons

- There are no real rules to define what data should look like which makes it hard to determine data quality.
- Data meaning is also problematic as there may not be a consistent approach to property naming conventions – any variations will impact data consumers' understanding of what that data means.
- Unstructured data will need to be cleaned and formatted before it can be analyzed. If this is not done at the point of collection then the onus falls on data analysts to clean data before their work can begin.
- Data governance is difficult to maintain, as analysts are free to capture whatever data they choose.



Which approach is best?

Knowing how to best structure data is personal to each business and dependent on what they want to achieve with their data.

For some organizations, packaged analytics platforms like GA are fit-for-purpose because they provide enough insight for their simpler use cases. On the other hand, data mature organizations may consider unstructured data a better option because of its greater flexibility. However, in both cases data quality, data meaning, and data governance are difficult to maintain. In that sense, both options have flaws.

A move to more bespoke data capture and structure

The way we want to capture events is changing, reflecting the unique nature of each organization and what data means to them specifically. No two businesses are the same, which means their tracking plan shouldn't be identical either.

For example, the concept of 'conversion' means different things to different industries and businesses. In ecommerce, an obvious conversion is an online shopping cart transaction. Whereas for a recruitment company a conversion could be a job application received or a company posting a new vacancy. The description 'conversion' has a different context depending on who is defining what it means.

Organizations now want to push their data tracking capabilities to new levels, and customize the behavioral data they collect. Simpler metrics around conversions and page views are no longer seen as the only important activities to track (although clearly they do still offer value).

There is a desire to move data collection beyond basic use cases and track events that will allow analysts to not just monitor behavior, but to actually understand intent and preferences. However, the binary 'structured or unstructured' approach to data is not conducive to this trend.

Structured data lacks the flexibility to adapt to new use cases, and cannot adapt to reflect the broad range of events that the business may want to track. However, it's not practical to adopt an approach that focuses solely on unstructured data. Inconsistent field names or descriptions can affect data quality. Data will subsequently need to be structured for analysis at some point down the line anyway, which will be time-consuming and laborious.

So what is the best way forward for organizations looking to take their data collection to the next level?

The need for a hybrid solution

We are seeing organizations move towards an approach where they blend both structured and unstructured data because they want to:

- **Maintain the flexibility of an unstructured approach so they can define what events they track.**
- **Be able to easily change what data they capture (and how) over time.**
- **Move at speed and get predictable data from a fixed schema approach.**
- **Collect descriptive data so that data consumers understand what it means, and for that meaning to persist over time.**
- **Develop a way for data consumers (and other stakeholders) to be able to enforce requirements for the data.**

What does a hybrid version look like with Snowplow?

Snowplow is a flexible data delivery platform that allows organizations to collect and operationalize behavioral data at scale. Snowplow empowers teams to build and own a data asset that delivers rich, high-quality behavioral data to the business and power advanced data use cases.

Snowplow enables data teams to get the best of both worlds by using self-describing, versioned data structures to collect event data. The pipeline uses the schemas to validate the data, as well as load it into tidy tables in the data warehouse.

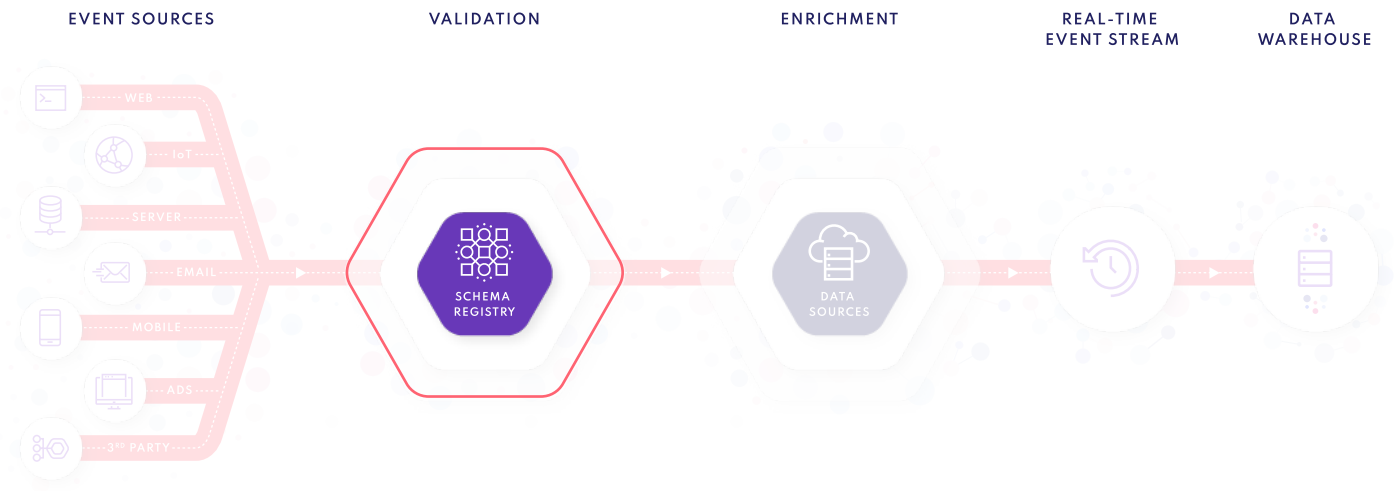


Figure 3: *Snowplow takes a hybrid approach to structuring data using self-describing schemas to maintain data meaning, data quality and data governance.*

This approach offers complete flexibility in terms of the number of custom events designed. Furthermore, because data structures are schema'd in advance, they're delivered in an expected format. Data meaning can be maintained through consistent descriptions. And the data team has agility to customize data collection and tracking as the business and its use cases evolve over time.

Snowplow schemas are stored in Iglu, a machine-readable, open-source schema repository. Because it's open source it's free to use, and there is a wide range of schemas available to the public.

Using self-describing schemas

A schema defines what fields are recorded with each event that is captured, and provides validation criteria for each field. Schemas are used to give data a particular structure. This means that unstructured data can have a degree of structure applied to it, helping to capture data meaning and the intent of that data. Adding meta-data (the self property in the screenshot below) and descriptions for properties in the schema ensure a clear understanding of what the data represents.

Schemas enable organizations to apply one common language to define data across the business. Schemas are important to data structure as they:

- **Define what events need to be tracked and collected**
- **Determine how data is validated after collection**
- **Define the data structure when the data is loaded into the data warehouse**

By re-thinking the structure of their data and choosing a hybrid approach, organizations are choosing to stop letting their tools dictate the rules. By defining what really matters to their business and retaining control of their data strategy, forward-thinking businesses are actively moving away from a pre-packaged mindset towards an approach that works for their individual business.

Within a hybrid approach using self-describing schemas, the organization can apply levels of structure to unstructured data. This allows them to maintain a flexible approach to what event data is captured, and also enables them to add descriptions and properties to define what that data means.

```
{
	"$schema" : "http://iglucentral.com/schemas/com.snowplowanalytics.self-desc/schema/jsonschema/1-0-0#",
	"description": "Schema for a scroll event",
	"self": {
		"vendor": "com.snowplowanalytics.marketing",
		"name": "scroll",
		"format": "jsonschema",
		"version": "1-0-0"
	},
	"type": "object",
	"properties": {
		"percentage": {
			"type": "integer",
			"minimum": 0,
			"maximum": 100,
			"description": "The percentage the user scrolled. Generally 25, 50, 75 or 100."
		}
	},
	"additionalProperties": false,
	"required": ["percentage"]
}
```

Figure 4: *example of how event data is collected within a self-describing schema.*

Using self-describing schemas, organizations have a more flexible approach without restrictions on the custom events designed, and within a highly expected structure. Descriptions can be added to maintain data meaning (as in the 'self' section above), requirements enforced, and semantic versioning applied – all of which help to retain data quality.

This delivers against the three criteria mentioned at the start of the paper. Data meaning, data quality, and data governance are all maintained, allowing the business to maximize the value of its data. There is a common understanding of what the data means, data is reliable and trustworthy, and it's accessible to the wider business.

Take the next step in your data journey

Each business is on its own data journey, one that will typically take them from relatively simplistic and common data use cases to very specific and bespoke requirements as the business evolves. If your organization is just starting out on its data journey, combining pre-packed or third-party tools with a structured or unstructured approach may well be all you need to collect the behavioral data to serve your use cases.

But for data mature organizations looking to derive greater value from data and use data in more meaningful ways, Snowplow's hybrid approach to data structure could be a better approach.

Snowplow empowers you to customize your data collection while ensuring data quality, data meaning, and data governance. It offers the flexibility needed to explore more complex use cases.

Snowplow Insights puts organizations in control of their data, allowing them to decide what they want to track and then ensuring the data is structured to enable the business to acquire the insights they need. Empowering data teams to rise above the difficulties of data delivery, we transform data into a strategic asset to the business and develop a culture of data excellence.

If you are serious about data and want to leverage its insights to improve decision-making across your business, ask for a one-to-one demo.

Book a demo

CUSTOMER INSIGHT

“Snowplow provides all of our event data in a data model which we own and can shape to our organizational needs. Snowplow has really helped accelerate our analytics; we can quickly answer questions which would have required a tremendous amount of engineering effort with our previous solution”

Darren Haken, Head of Data Engineering