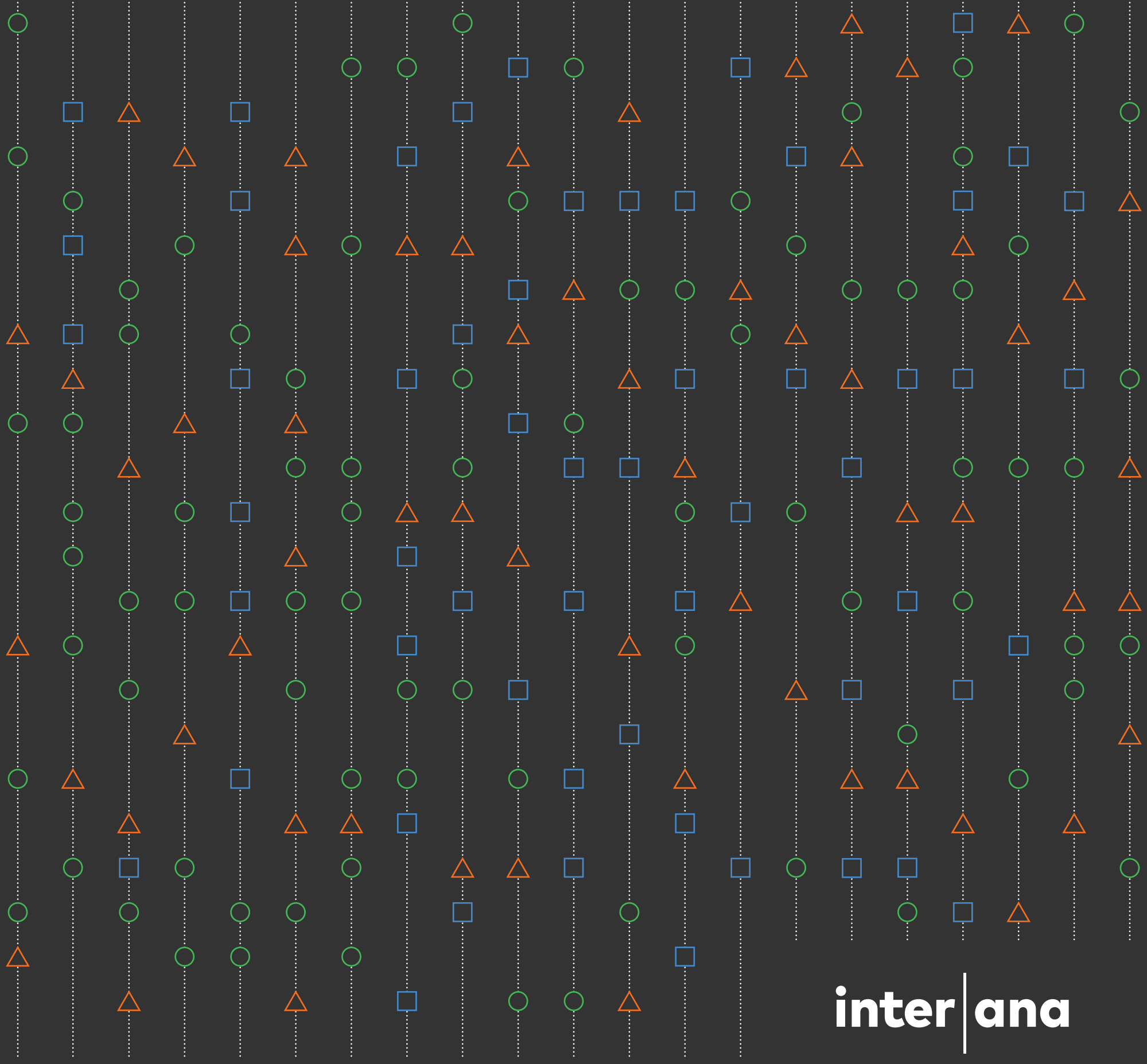


inter|ana

TECHNICAL E-BOOK

Sampling for Behavioral Analytics at Scale

inter|ana

Interana is the behavioral analytics solution for event data that enables people to easily obtain insights from the actions people, products, or machines make over time.

These insights empower people to build the right strategies for increasing conversion, deepening engagement, and maximizing retention in their products or services.

Introduction

When it comes to Big Data, sampling done right can save time and money. But sampling traditionally came with a price: it's not always simple to do right, and sampling can introduce inaccuracy that leads to wrong decisions. Often sampling was a last resort when unsampled data was too expensive or simply too big to handle.

Interana helps break that dichotomy for event data and behavioral analytics. Other solutions require sampling just to be able to work at scale. For Interana, sampling is a useful tool but not essential for fast results at massive scale. You can get statistically accurate sampled results at the speed of thought, refine your analysis, and when ready get fast unsampled results for hundreds of billions of events. Another Interana advantage is that we ingest and store all the raw events and optionally sample during the query. Solutions that sample at ingest are not well suited for behavioral analytics at scale. Explore with us as we show you how Interana can handle a trillion events in 3 seconds while providing full access to all the raw data.

Sampling is important for anybody who wants to interactively explore vast amounts of data. Product Managers typically find Interana sampling particularly useful: it allows them to go beyond what their users say and quickly explore what customers actually do with their products. They can create and refine queries without knowing any SQL or statistics – using an intuitive interface geared toward turning their clickstream logs into insights about their users' behavior.

Analytics has many goals, and often finding the right question is harder than finding the right answer. Sampling supercharges the process by reducing the query response time to seconds. You can ask more questions, worry less about asking questions that don't go anywhere, and generally dive deeper into your data. A typical exploratory session starts with a vague idea about some gem that might be lurking in the data, turns into a series of queries that zero in on the right question, and then saves the final query by pinning it to a dashboard.

Tools for segmenting data and analyzing behavior have been around, in some form or another, for years. But they've been so complex that only technical experts could work with them. Business users had to ask these experts to run queries against the data, and often had to wait days or even weeks for answers. After so much time had passed, the question—and the long-awaited answer—were sometimes no longer relevant.

Interana changes all that with a visual query builder that makes it easy for non-technical users to ask questions themselves, eliminating the cost and delay of having technical experts write long and complicated queries. It helps business users create metrics through a dialogue that makes recommendations based on what is measured and adapts to the flow of what is tracked.

Interana is the first fully integrated solution built for behavioral analytics on event data. We turn all your disparate event data into an integrated sequence of events to reliably reveal what users truly want, need, or don't like about your products and services. With Interana, you can quickly develop the right strategies for new opportunities to acquire customers, deepen engagement, and maximize retention.

inter|ana

The first solution built from the ground up to easy explore and discovery every action a user takes with a product or service over time. Easily iterate through a series of questions, across endless dimensions, in minutes using pre-built behavioral features like cohorts, metrics, sessions, funnels.

High level concepts

First, let's quickly review some high-level concepts to make sure we're all clear on the terms and ideas in this paper:

Sampling

Sampling is the process of selecting a subset of some general population that can be used to accurately estimate important attributes of the population as a whole. That subset is called the sample. In behavioral analytics, the sample must be made based on the actor whose behavior we're analyzing. For example, when estimating user properties we must sample the population of users, not the population of events. Once the sample is analyzed, the results need to be meaningfully scaled up to reflect the population as a whole. Different scaling strategies make sense, depending on the data and sampling method. Because of how sampling and statistics work, it's possible to use a very small subset of the data to draw statistically accurate results with a high level of confidence. Census takers are one familiar example of sampling in everyday life. They follow a detailed methodology to make sure questionnaires and interviews from a small subset of addresses can accurately represent communities, cities, states, and the nation as a whole.

Behavioral Analytics

Behavioral Analytics focuses on how various actors behave when interacting with a product or service. Those actors could be people, or a technological entity like a mobile phone, IP address, etc. The behavior is tracked as a sequence of events that occur at specific times. The order, duration, and time between events are all relevant for understanding behavior. Selecting a subset of events in a way that doesn't preserve their relationship to the actors, their order, or their relative timing would distort the picture. Any sampling strategy needs to take all this into account.

Event Data

By definition, event data is data from any identifiable occurrence that has significance for system hardware or software. User-generated events include keystrokes and mouse actions, among a wide variety of other possibilities. Events describe an action performed by or associated with an actor at a certain time. Event data is a continuous stream of actions revealing the pattern of events that people, products, and machines make over time. It helps describe when and how things happen. Event data is the foundation for behavioral analytics; enabling understanding of how customers behave and products are used.

Every event has a timestamp, actor, and one or more attributes of an action. As simple as that sounds, events are at the heart of many businesses. Clickstreams, logs, data from wearable devices, sensor data, and more are all event data. A mouse click is an event; it happens at a point in time and its context includes attributes such as where the actor clicked.

inter|ana

Trillion rows in 3 seconds, billions in less. Interana provides access to 100% of the raw event data with the speed to easily ask series of questions in seconds, without the consequence of being wrong. Interana's scale keeps the richness of data by not requiring aggregations or summarizations often used to shrink it into other solutions.

Context

Now that we've got the basics covered, let's go deeper into the context.

Huge Scale

Not just big data, but humongous data. Many new and interesting applications for behavioral analytics relate to online and mobile applications with many millions of users and hundreds or thousands of events per user interaction. Interana customers see hundreds of millions of events per hour. Being able to ingest, store, and analyze all that data in terms of behavior takes a dedicated approach. General-purpose solutions for big data analytics might keep up at smaller scales, but at high volumes they're forced to make compromises like:

- Pay tons of money to build and operate ever bigger clusters. Or perhaps pay tons of money to store everything in memory on huge machines.
- Processing lots of data takes time. Unless the solution is expensively over-provisioned, more general-purpose big data solutions take a relatively long time to crunch all the numbers and come up with an answer. It's common to wait many minutes or hours to get results. And that's assuming that we're talking about the machines...
- Most big data solutions evolved to improve on relational databases. They still depend on specialized query languages (at least under the hood). Query languages aren't meant to track the evolving state of an actor over time, and require all sorts of contortions with subsets and joins to get close. A product manager either needs to depend on a data scientist or spend time doing programming instead of spending time with customers.

Time to Discovery (Explorative, Interactive Workflows)

People looking at behavioral analytics are often working with a new product or service. It could be a disruptive new take on an existing problem, or something completely novel. Sure, there's often a small set of traditional questions that everybody learned in business school, but those can often be answered with relatively traditional techniques. The attraction of behavioral analytics is to discover something new about users and interactions. That discovery process is exploratory by nature, and exploration is best done interactively. There's just something compelling about diving deep into the data and seeing it in new ways and from different angles. Interana customers report getting into the flow of their exploration and spending hours without coming up for air. That's something that gets lost when each query leaves enough time for a cup of coffee or writing up a long email.

Even for mature products, there may be questions nobody thought to ask because they were previously impossible to answer. Not knowing what questions to ask means starting with lots of wrong questions. The cost of asking a wrong question needs to be almost zero. That helps keep the exploration moving ahead smoothly. And while most people don't like being wrong, they hate being wrong in public.

inter|ana

Enable behavioral analytics on event data with a fast, visual, and intuitive solution accessible to all. Interana allows for the transformation of a company's culture from making decisions based on beliefs, to making them based on data. This fosters better collaboration and transparency of insights to make all business decisions data-informed.

Going to a team of data scientists again and again with different wrong queries can be off-putting. That's another thing that Interana customers love: they get to explore on their own, quickly, using our intuitive interface without needing anybody in the middle to interpret their exploratory questions.

The end results are all win: a much more pleasant experience for the person exploring the data, a much shorter time to discovery, and a better product or service for the ultimate customer.

Behavioral Analytics

We're focused on behavioral analytics. The data represents not only a series of events in time, but is associated with actors. Actors that are part of some general population, perhaps viewed as different cohorts that have something in common with each other. Often these actors are either people or have a person behind the recorded device. People tend to behave somewhat predictably within large populations, which helps with the analysis. Accuracy is important but forgiving. People-oriented forecasts and predictions are expected to be accurate within some relatively large margin. There's usually a balance between how accurate an answer needs to be, how much it costs to get a more accurate answer, and how much value additional accuracy brings to the organization. Sampling the right way can use a small fraction of the total data and just a few seconds to arrive at sufficiently accurate results.

The fact that we're working with events that happen over time is also important. Relational databases and traditional big data solutions rarely do anything smart for time series data. That forces strange contortions in database schemas, makes writing queries complicated (e.g., Figure 1), and requires more expensive hardware than strictly necessary.

The order of the events is also important, both globally for the dataset and within a single actor. Without preserving that order, all sorts of behavioral information gets lost. And there are usually many different types of events, often with very different sets of attributes. Trying to keep all this information in some normalized form is both maddening and inefficient. Interana is designed to work with the kind of irregular, sparse, and denormalized event data that's at the heart of behavioral analytics.

```
SELECT emailtype, SUM(timesOpened) as timesOpened, SUM(timesClicked) as timesClicked, SUM(-
timesViewed) as timesViewed, SUM(timesInvited) as timesInvited, SUM(CASE WHEN timesOpened
> 0 THEN 1 ELSE 0 END) as uniqueOpens, SUM(CASE WHEN timesClicked > 0 THEN 1 ELSE 0 END)
as uniqueClicks, SUM(CASE WHEN timesViewed > 0 THEN 1 ELSE 0 END) as uniqueViews, SUM(CASE
WHEN timesInvited > 0 THEN 1 ELSE 0 END) as uniqueInvitations FROM ( SELECT userid, email-
id, timeSent, emailtype, COALESCE(COUNT(timeOpened),0) as timesOpened, COALESCE(COUNT(time-
Clicked),0) as timesClicked, COALESCE(COUNT(timeViewed),0) as timesViewed, COALESCE(COUNT(-
timeInvited),0) as timesInvited FROM ( SELECT C.userid as userid, C.emailid as emailid,
C.emailtype as emailtype, C.timeSent as timeSent, C.timeOpened as timeOpened, C.timeClicked
as timeClicked, C.linkClicked as linkClicked, C.timeViewed as timeViewed, C.pageURLViewed
as pageURLViewed, invitationsSent.timeInvited as timeInvited FROM invitationsSent RIGHT
OUTER JOIN (SELECT B.userid as userid, B.emailid as emailid, B.emailtype as emailtype,
B.timeSent as timeSent, B.timeOpened as timeOpened, B.timeClicked as timeClicked, B.link-
Clicked as linkClicked, pagesViewed.timeViewed as timeViewed, pagesViewed.pageURL as pa-
geURLViewed FROM pagesViewed RIGHT OUTER JOIN (SELECT A.userid as userid, A.emailid as
emailid, A.emailtype as emailtype, A.timeSent as timeSent, A.timeOpened as timeOpened,
emailClicked.timeClicked as timeClicked, emailClicked.linkClicked as linkClicked FROM
emailClicked RIGHT OUTER JOIN (SELECT emailSent.userid as userid, emailSent.emailid as
emailid, emailSent.emailtype as emailtype, emailSent.timeSent as timeSent, emailOpened.
timeOpened as timeOpened FROM emailOpened RIGHT OUTER JOIN emailSent ON emailSent.userid=e-
mailOpened.userid AND emailSent.emailid=emailOpened.emailid WHERE emailSent.timeSent BE-
TWEEN (NOW() - interval '7 days') AND (NOW() - interval '14 days') ) AS A ON emailClicked.
userid=A.userID AND emailClicked.emailId=A.emailId ) AS B ON pagesViewed.userid=B.user-
id AND pagesViewed.emailid=B.emailid ) as C ON invitationsSent.userid=C.userid AND invi-
tationsSent.emailid=C.userid ) as D GROUP BY userid, emailid, timesent, emailtype ) as E
GROUP BY emailtype ;
```

Figure 1. Example of complex queries

When to sample

Statistical sampling is as old as the field of statistics itself. The science and math is well understood, and it's a useful tool to apply to a broad range of problems. But sampling has a somewhat mixed reputation with big data aficionados. Some feel it isn't necessary because it's possible to throw hardware at the problem. Some have been burned because it's possible to do sampling wrong. For behavioral analytics of event data, there are clearly right and wrong ways to sample.

What Not To Do

It's tempting to sample at the data collection points. So tempting that it's a common feature of many analytics solutions. There are potential upsides: the data is smaller and easier to ingest, less data needs to be stored, and when it comes to analysis the data can be processed as-is without further reduction. *But for behavioral analytics, this approach is tricky and limited.*

First, the sampled events must represent a series of actions by a set of actors. Their contents, sequence, and timing are all important. You can't just take every 100th event (e.g., Figure 2). The events collected need to retain all the information about the interaction between the actor and the application. The dropped events might contain critical information like payments, registration, or errors. These dropped events will skew important metrics. For example imagine what happens to key metrics when critical events are excluded by sampling: events that signal the start or end of a session, events that define membership in a behavioral cohort, events that trigger the move to a different step in a funnel, etc. So any ingest sampling needs to based on actors (e.g., Figure 3), with all events associated with the selected actors recorded.

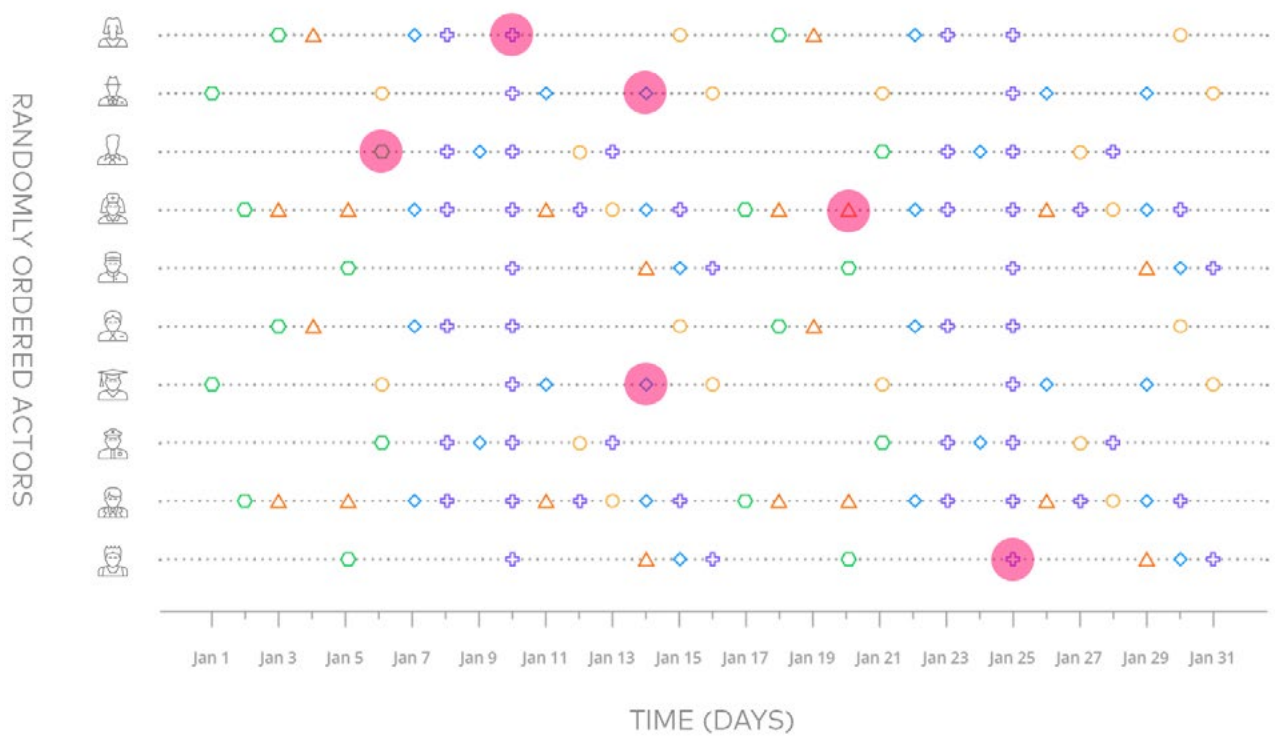


Figure 2

inter|ana

"We considered a range of popular analytics options, but none could offer us the speed, scalability and flexibility of Interana and still be truly self-service for both general business users and data analysts. The two big wins of Interana are that's it's so fast and the user interface is surprisingly simple given how much you can do."

– Dan Gould, Vice President of Technology at Tinder

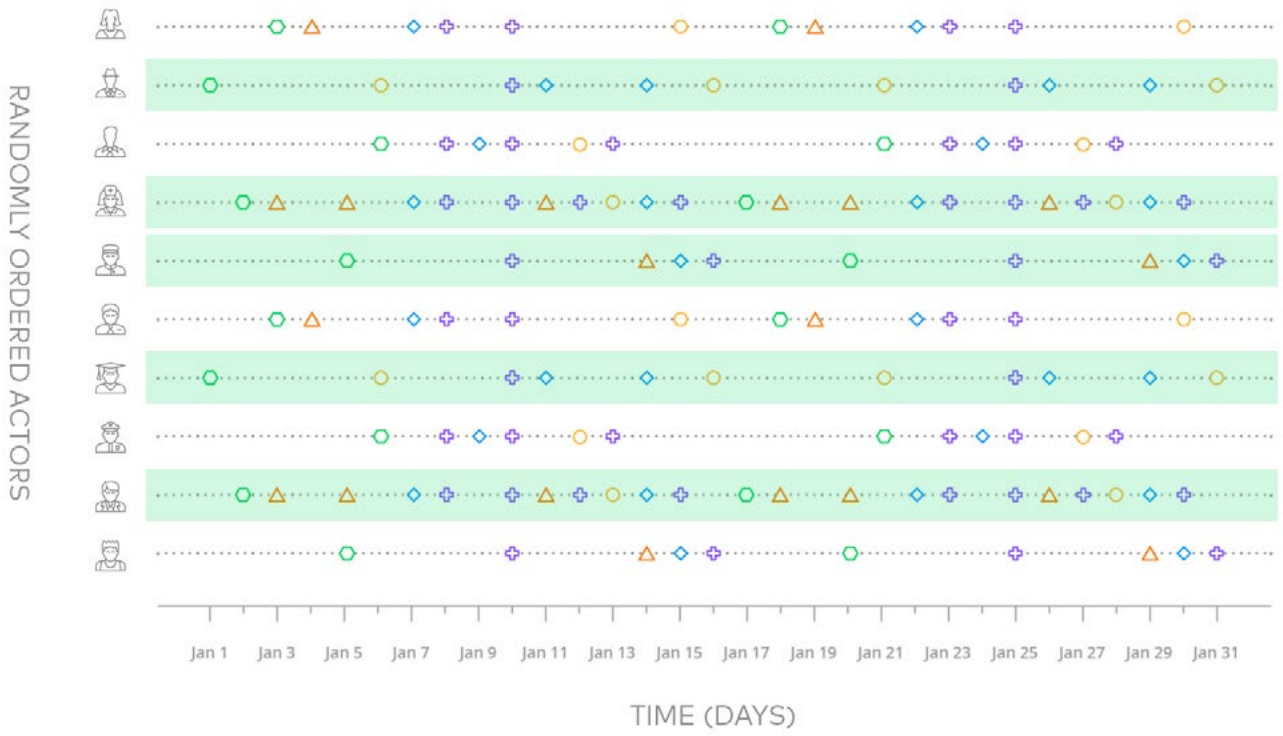


Figure 3.

Second, we don't know which actors will represent interesting subpopulations ahead of time. That's part of the discovery process. Let's say the cohort we wish to study is defined by some complex behavioral criteria (e.g., performed a certain type of action twice in their second week after sign up). It's impossible to know which actors will perform some action in the future. Plus, the exact set of criteria may change from query to query: the important criteria are not known in advance. In fact, the whole point of the investigation might be to discover those criteria. Any behavioral analysis that seeks to understand and segment the user population becomes inaccurate when actors are sampled ahead of time. Any techniques to try and work around this limitation are time consuming and add complexity to the process.

Lastly, dependence on sampling might reduce the collected data to such a degree that it can't be used for important workflows like A/B testing. In A/B tests, a small fraction of the users are shown a different version of the product and then analyzed to see whether the new version is significantly better than the original. Let's say the raw event data is already sampled down to some manageable percentage. The small fraction of those sampled users shown the modified product might now be too small to draw statistically meaningful conclusions. For example, If your regular conversion rate is 2% and you'd like to detect a 5% change in conversion at a 95% level of statistical significance, you'd need to collect events for at least 400,000 users shown the product variation. The entire sample might only be for a few million users, leaving insufficient data to draw conclusions for even a single A/B test.

inter|ana

"Interana and Microsoft share a common vision of innovation. During the proof of concept, Interana met the challenge of delivering the insights we needed from our massive volumes of event data. Interana's scale is impressive. It took only minutes to get answers to questions, opening new possibilities as to what we can do with analytics at interactive speeds. We look forward to continuing to work with the team."

– Craig Miller, Group Eng. Manager for Bing Experiences at Microsoft

Gotta Catch 'em All

For behavioral analytics of event data, the correct approach is to record all the events and make them part of the dataset. Sampling needs to happen at the time of the query, not during ingest. This approach moves the burden of correct sampling from the end user and onto the analytics platform. With correct sizing, all the data can be ingested and stored. Then the user can freely explore the data using sampling to quickly arrive at the set of most meaningful questions. And when the questions are dialed in, if the query needs to run across the full dataset it will take into account the complete population that matches the query filters. The unsampled query takes longer to run, but the results will be worth the longer wait because it's asking the right question.

It's All in the Implementation

You might be wondering: if the answer is so clear-cut, why isn't everybody doing it the same way? The reason comes down to implementation. A solution focused on behavioral analytics for event data can organize and manage data in ways that don't make sense for a general purpose analytics solution. That organization brings the power to store huge volumes of event data efficiently. It offers the ability to quickly scan anything—from a small fraction, to the whole dataset—at query time. And because the data is temporal, the queries can search across minutes or months to further balance between time to discovery and accuracy of results.

Let's take a look at how Interana efficiently addresses sampling for datasets with hundreds of billions of events...

inter|ana

"Interana makes it easy to answer questions like how many people have ever taken this action in the application. This is really hard to do in other solutions I've used, but Interana makes it trivial to dig deep into the history of customer events."

– Prashant Pandey - Head of Infrastructure Engineering at Asana

Interana approach

Scale-out Clustering

Like all modern analytics platforms, Interana is architected using a scale-out clustered approach. There are multiple nodes (machines) in the cluster, and the number scales to match the demands of the data and concurrent users. Each machine has one or more jobs within the overall solution. The import nodes are tasked with ingesting event data from event logs, traces, etc. The data nodes manage efficiently storing and scanning the event data. The string nodes compress and deduplicate all the strings in the event data, handling storage and lookup for query results so that all data operations can take place efficiently using compact integers. For smaller clusters, nodes can take on several of these roles as needed. These machines can be bare metal, virtualized, or securely running inside of a public cloud service.
Check out Figure 4 for an example diagram.

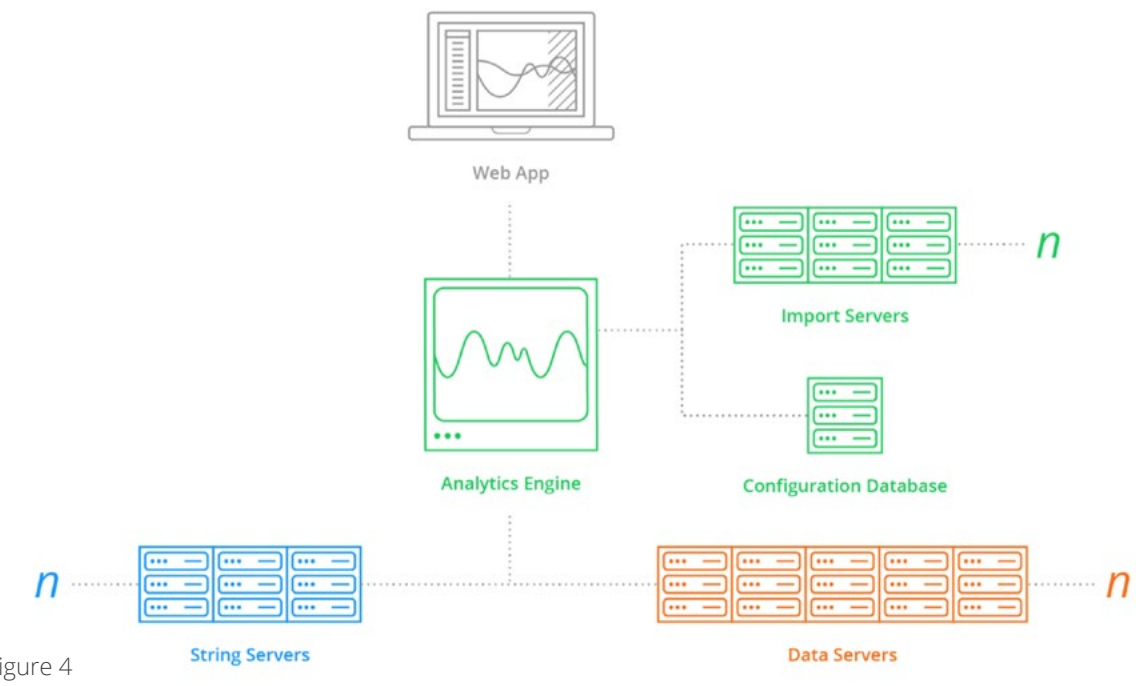


Figure 4

The end result is that no matter the scale, Interana can match the requirements. From smaller deployments that can be satisfied with a single efficiently utilized machine, to huge clusters with hundreds of nodes, the Interana application can scale to meet demands. Our customers have some of the busiest services in the world, with clusters hosting over a trillion events.

Consistent Organization and Representative Populations

One mechanism to support accurate behavioral sampling is to make sure all the data for an actor is stored together on the same node. And that the storage and placement of individual actors is fair and even for all actors in the population. We do that mathematically by using a hash with appropriate properties, followed by selecting one of many shards (containers) to hold that particular actor. All events associated with that actor will be held within the same shard. That shard will be managed by one node at a time, so all the events for a single actor are managed efficiently on a single machine. Because the actors are evenly distributed among shards, every shard contains a representative slice of the overall population.

inter|ana

Interana is the behavioral analytics solution for event data that enables people to easily obtain insights from the actions people, products, or machines make over time.

These insights empower people to build the right strategies for increasing conversion, deepening engagement, and maximizing retention in their products or services.

Sampling Actors to Get Entire Time Stream

We can take advantage of the fact that a single shard holds a representative cross-section of the population when it comes to sampling. When we sample, each data node processes a single shard and scans the requested time range for all the actors in that shard. Because even a single shard may hold more actors than necessary for a statistically valid sample, we can sample with progressively larger subsets of the shard until we have sufficient confidence in the sampled result. Scanning the sample shard is usually very quick, because the common time ranges tend to be cached in memory. Even when the data is scanned directly from disk, the process usually takes just a few seconds. We only scan the columns required for that particular query, and those columns are stored very efficiently and processed with a highly optimized C++ backend. The end result is quick, accurate sampling for behavioral analytics and the ability to interactively explore vast amounts of data.

Parallelized Queries, Sane Results

You might be wondering how we know that the sample from a single shard is actually representative of the population. That's where the scale-out nature of the Interana solution comes in. The query isn't just run on a single shard. It's run on a shard for each data node of the cluster in parallel. The algorithm balances actors evenly across all nodes and shards, so each node should complete the query on its shard at roughly the same time. The results from each data node are merged together and compared. If the statistical profile of the actors in each shard is roughly the same then the sampling is held as valid. The results are scaled up based on the size of the sample relative to the overall population, and the results returned to the user. Any strings in the results are translated on the string servers at this time, generally eliminating string operations during the query processing. The product is a fast, accurately sampled result that reflects behavioral information across billions of events in mere seconds.

Knowing When Not to Sample

Of course sampling isn't always appropriate. Certain data isn't going to be evenly distributed among the shards. Some events are very rare and unlikely to show up in a sampled result. Sometimes you're looking for a tiny specific set of events but aren't sure when they occurred. Sometimes the selection filters leave too few events to sample accurately. Interana detects situations where sampling isn't appropriate and returns a clear warning with the results (e.g., Figure 5).

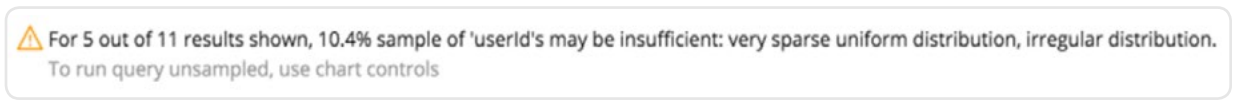


Figure 5

The user can then either reformulate their query, or choose to disable sampling and compute the query across the full set of data. It's up to the user which they prefer. In general, unsampled queries are only needed when exploring a very small or highly irregular data set. For example, when the scope of the query is so narrow that it covers fewer than a few thousand users or events. In all of these cases, Interana will recommend that the query be rerun unsampled. But even when running a query on the full set of events, Interana returns results in seconds. The entire system was designed around the proposition that scanning billions of events is the common case and needs to complete quickly. And we deliver.

inter|ana

The first solution built from the ground up to easy explore and discovery every action a user takes with a product or service over time. Easily iterate through a series of questions, across endless dimensions, in minutes using pre-built behavioral features like cohorts, metrics, sessions, funnels.

Conclusion

So let's wrap up. Interana is a purpose-built solution for behavioral analytics of event data at scale. The solution consists of a highly scalable cluster which is combined with an intuitive visual interface to interactively explore trillions of events in seconds. Part of how that's possible is the architecture of the solution, and part is the integral role of sampling.

We've explained how sampling for event data needs to be handled correctly. And that sampling got a bad rap because in the past big data solutions didn't always do the right thing. Interana handles sampling for behavioral analytics correctly, and uses this powerful technique to give our users the freedom to rapidly explore their data without fear of asking the wrong questions. Enabling them to zero in on the right questions to ask and significantly reducing time to discovery.

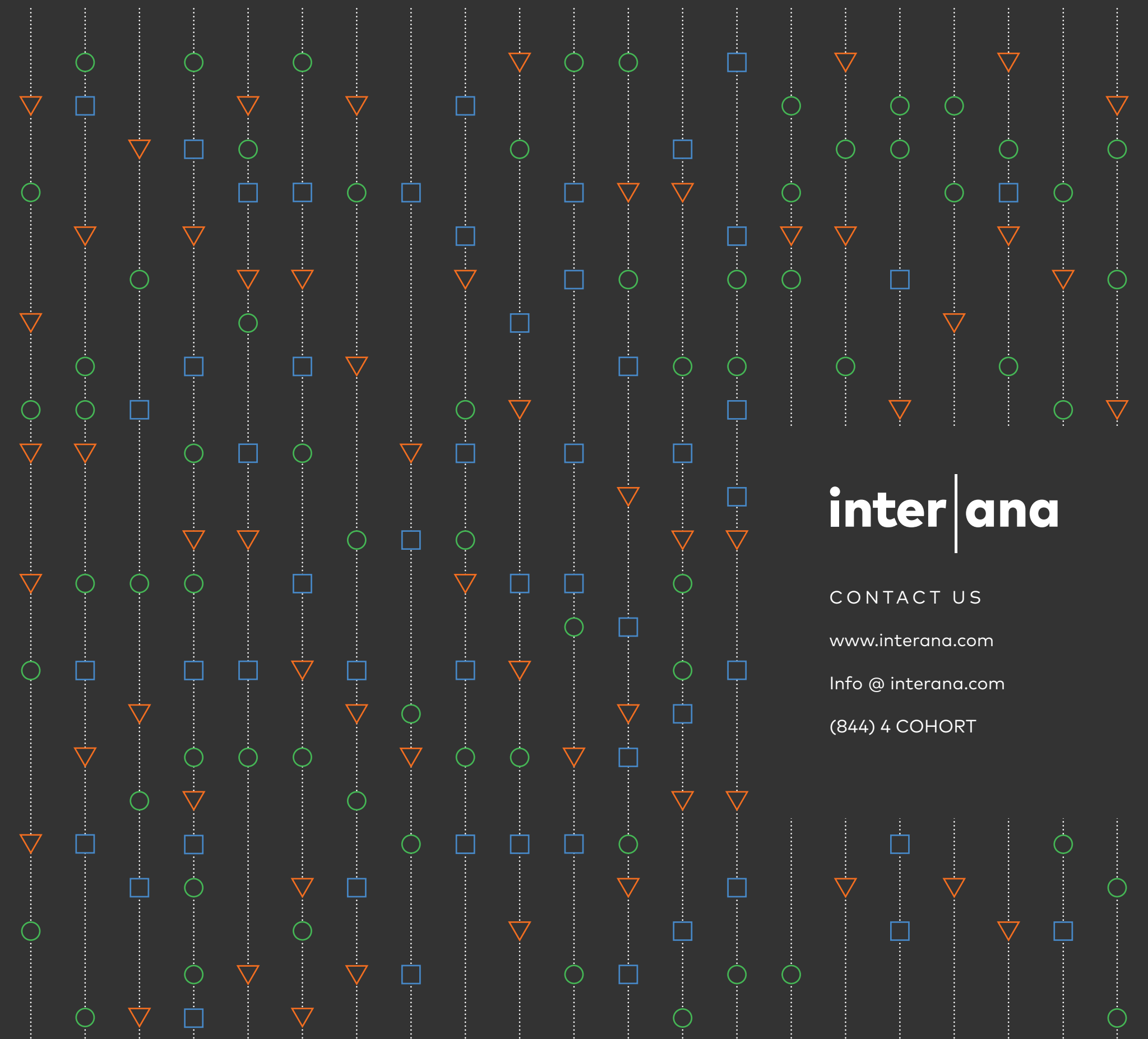
Don't stop here!

Interana has a number of other useful resources:

- Learn more about how other companies — Tinder, Microsoft, Sonos, Imgur, Asana, Flowroute, BloomBoard, and more — have used Interana to gain insights into their customers' behavior.
- We have many resources that can help you get a better understanding of Interana's solution and behavioral analytics on event data at massive scale.
- See how Interana can help you discover what your customers think and do — Request a demo of the Interana solution in action.

REQUEST DEMO

Thanks for reading!



interana

CONTACT US

www.interana.com

Info @ interana.com

(844) 4 COHORT