Harold Byun:

Hi, and welcome to today's webinar on Data Privacy Simplified. We are just getting going here. We're going to hold for just another minute or so for some other people to get logged in and then we'll kick away with the webinar. Thanks for your patience.

Harold Byun:

(silence) As we're waiting for people to continue joining, just a couple of logistics items. If you're interested in asking any questions, please do so along the way. We will have a Q&A period towards the end, and I'll try to answer as many questions as we can as we go forward. For those of you that are getting a message, if there's any case where the video cannot be played, typically this is some kind of network error that unfortunately is either something to do with the live streaming versus the tests that actually BrightTALK runs. You will be able to play in post. We've had other customers who have had this challenge previously. It's typically due to some kind of filtering proxy on the network and it's unfortunate, but you can try and refresh or go into incognito mode or go from a phone or an iPad. Those are other ways that you could potentially get in. If you're having that problem, I apologize.

Harold Byun:

So, why don't we kick off? I appreciate you all taking the time to attend the webcast today. Today, we're going to be covering different methods for simplifying data privacy, implementing data centric protections to secure data, and we'll also be covering a range of different emerging capabilities around privacy preserving analytics and secure data sharing techniques.

Harold Byun:

My name is Harold Byun. I head product management for Baffle. The agenda for today, I'm going to try and get through all of this in roughly 30 to 35 minutes, including a few demos. I'm halfway optimistic that I'll be able to do that, but I tend to try and get through the content relatively quickly and be conscious of your time. With that regard, I'm going to probably gloss over some of the data privacy challenges and regulations. I'm assuming that many of you have been inundated with that information.

Harold Byun:

We also have a link that I will be posting throughout the webcast that has resourcing to additional information. We have a dedicated webinar on CCPA, data privacy as well as an IT GRC Forum Summit that we did around CCPA regulations. So, there's a wealth of other resources that you can use. In the interest of time, I'm going to try and cover that at a high level. Well, then be going into a kind of a disparity between data breaches and threat models and why data breaches continue to occur from our perspective, we'll cover some common data centric protection models and how those can be implemented, and then we'll wrap up with privacy preserving analytics and secure data sharing use cases as well as some Q&A.

Harold Byun:

Again, if you have questions, please use the chat panel throughout. You can also email info@baffle.io. My email is harold@baffle.io as well. In the interest of time as well, and hopefully as I go through some of the data privacy regulations and some of the maybe more mundane material, I'd like to hopefully keep you engaged.

Harold Byun:

In the realm of privacy preserving analytics, just to give you some sense of what it is, it is a computational method that allows for operations, processing and analysis of data without ever disclosing or revealing the underlying data values. This is to preserve privacy and ensure that the data privacy contract is not violated. So, very relevant in terms of today's society. The inset quote here is really from a Gartner report on privacy preservation and analytics.

Harold Byun:

Effectively, as organizations are looking to derive intelligence from their data, as they're pursuing machine learning or AI or other big data strategies, those types of things or those types of activities carry all kinds of risks with them within the context of concerns over data privacy and the penalties and fines that are associated with data privacy regulations. And so, this is really looking at what can be done to mitigate some of that risks, is there a way to pursue a secure ML or AI strategy, and different methodologies that people can use to more securely share data and run analytics on it.

Harold Byun:

The link below, baffle.io/privacy will bring you to a set of resources that are available for your download as well. We'll be emailing all this presentation out to you as well and post.

Harold Byun:

So, just to give you a flavor of this, a quick background on me. I've been doing security for over two decades, most of it focused on different data containment and security strategies from data loss prevention to CASB to mobile data containerization as well as this experience with user behavioral analytics back in the day. I've been very focused on how data is exfiltrated from an organization. I think that dovetails nicely into how to analyze data without actually necessarily revealing it, which is the privacy preserving analytics angle now.

Harold Byun:

In terms of the overview on data privacy and regulations, again I'm going to cover this at a high level and try and get through it pretty quickly just because I think many of you have probably been inundated for the last two, three years, four years on different privacy regulations like GDPR and CCPA. Obviously, the penalties continue to increase and there's an impact on your business. Again, this will all be available via those download links that we talked about.

Harold Byun:

I think that the high-level driver around data privacy is really that there is a requirement to monitor, discover and disclose in terms of this overarching requirement dubbed right to know. And then there's this area of right to be forgotten, which is the ability to delete data that a consumer deems necessary to be deleted about themselves. Those are the two overarching requirements, and that's all I want folks to take away from that. Those are relatively newer requirements within the realm of privacy regulations. There's a fair amount of work that organizations need to do to operationalize and put in operations in place in order to address these types of scenarios and that's part and parcel to the data privacy requirements going forward. Many of you are probably familiar with this. Again, slides will be available, I'm not going to talk to the dates all around this. If you do business in California and you generate a certain amount of revenue, you're obviously required to comply with at least CCPA.

Harold Byun:

Within the realm of these privacy regulations, there are some new data types that do need to be encrypted. Not encrypted. Protected, rather. Encryption is one mechanism of it. Biometric data for example, profiling or inferences that are drawn from certain types of analytics or data that you're collecting about a given individual. Those are pieces of information that are relatively new. IP addresses were new under GDPR. Those are things that might be relevant. Again, this information will be available to you.

Harold Byun:

As it relates to encryption in particular, CCPA does talk about non-encrypted or non-redacted personal information that is breached. In the event that that occurs, the consumer, without an actual legal filing, can pursue companies for up to $750 per violation. And so, just a quick math on that is if you do business with 10,000 people, then you're potentially on the hook for 7.5 million. The fine continues to be potentially significant. We'll see how that actually shakes out, but something to obviously be aware of as you look at risk mitigation and return on risk.

Harold Byun:

Again, the baffle.io/privacy URL if you want to download any of these resources. We have the prior webinar on CCPA compliance and the GRC webinar that we did as well, and there's a whole series that we've done on data privacy and security.

Harold Byun:

That's the first section here in terms of just a quick overview on privacy regulations and potential impact. What I really wanted to talk about more so is really the context around data breaches and why they occur, and the threat models particularly because we seem to run into a lot of confusion when we are talking with customers around what is actually mitigating in a given threat as it relates to a data breach or an attack or an attempt to steal data. I think that that is part of the reason why we continue to see the ongoing onslaught of breaches every other day. Obviously, there's some metatrends that are affecting [data security.](https://baffle.io/data-de-identification/) The continuing data breach scenario that you read about every day in the news, there's the migration to cloud which is seemingly exposing data due to misconfigurations, as well as just the prevalence of distributed data and third party risk and data sharing continues to be an issue for a lot of organizations.

Harold Byun:

According to the Ponemon Institute survey, they surveyed 1000 CISOs of large enterprises and 60% of them roughly reported having a data leakage due to a third-party relationship. So, why did the breaches continue to occur? When we really look at it, there's a number of reasons that we could potentially associate with why breaches continue to occur. I don't know, are the hackers getting more sophisticated? Is it the misconfigurations? Is it more zero-day attacks? There's a realm of possibilities and the reality is that it really comes down to, quite frankly, all of these things. There's no method for security whether you're exercising a range of different defense in-depth strategies to recover them all. You even just look at basic patching and vulnerability management and the struggles that a lot of organizations have to keep up. The net of that is that ultimately attackers are in your network and they will get to your data. So, there needs to be a continual evaluation of how to take a different approach to mitigate these types of breach that's within the threat model.

Harold Byun:

If we look at the data access model and how people are actually accessing data today, typically you could, that's obviously overly simplified, but you would have a user accessing through some application footprint and asking for a certain amount of data. Within that, you have context around that user. You have a good user, you have a bad user or a malicious user or an unknown user, or compromised account or a compromised user. They are accessing information through a good application or a known application versus an unknown application or a malicious or malformed application obviously asking for a set of data. These are either in the form of legitimate data requests or excessive data requests. And then, you also have privileged users and insiders with access to the back end. When you really look at this in sum, it is an end-to-end access channel that needs to be secured.

Harold Byun:

Ultimately, the approach that many people have taken and continue to take is treating these as isolated silos. Ultimately, the problem with that is that there are multiple methods to circumvent those data controls at different levels and actually still get at the data, which is why we typically see, again, these breaches and these bulk data exfiltrations continue to occur.

Harold Byun:

When we look at how this has evolved, the challenge with this is you don't just have your own users, you have multiple third parties that you're engaged with across the chain. You have microservices or containers, serverless functions as well as API gateways or a set of APIs. They're now all constructed around accessing these data sets. When we look at this, and obviously housing data in the cloud whether that be database platform as a service or SAS or other cloud-based models, just in terms of overall cloud infrastructure, it distributes the data even further. And so, the access model, again, applies to a set of broader users, good, bad, unknown or compromised. It goes beyond an application traditional footprint. It's now good code, unknown code, malicious code, serverless malicious code, or serverless unknown code that may be executing functions through this cloud footprint wherein you now have this notion of an untrusted entity housing your data. And so, the access model was even expanded beyond the simplified approach that we were just discussing in the prior slides.

Harold Byun:

What this all suggests is that in that distributed data environment, implementing or securing information at the data level becomes even more important in a data-centric manner because, inevitably, the entry points of the data are going to continue to expand where there are organizations with hundreds, if not thousands, of micro services. The ability to shore those access points up as they access through either a service mesh or an API gateway or leveraging other types of cloud native footprints, the exponential entry point into the data pool continues to expand. Ultimately, that's going to continue to introduce more risk.

Harold Byun:

Meanwhile, on the converse side of that or the flipside of that, the data privacy regulations and compliance mandates continue to increase and expand, and the business drivers to go faster continue to press, making security a bottleneck for a lot of organizations and, quite frankly, presenting a security nightmare.

Harold Byun:

So, ultimately, what we believe people should be looking to do is implementing, obviously, the correct technical controls to secure the data in a data-centric manner. It's also realizing that [encryption and data protection controls](https://baffle.io/cloud-data-protection-platform/) don't actually protect at the data level in a lot of cases. And so, people really need to start looking at data-centric protection methods that protect the actual PII or consumer data values and perform some entitlements and access controls around that while also fulfilling this mandate to provide visibility and monitoring, or being able to discover consumer data as well as executing on the right to be forgotten, or deleting that actual PII.

Harold Byun:

When we look at this within the context of controls that people are mostly implementing today, which is container-based security or at-rest encryption or other types of basic security controls, the problem with these methods is that when you really look at what they're protecting against within the context of the threat model that we've been discussing, is that they do nothing to protect against the modern day hack. These are controls that were really designed to protect physical disks or physical access to machines. They'll protect against a lost laptop that's left in the backseat of a taxi or at a restaurant, or protect against what we jokingly call the Tom Cruise threat model which is Tom Cruise dropping in from the ceiling of the data centers, ripping out hard drives and walking off with them. That's not really the modern-day threat model as it relates to data theft and breaches, but this is what most people are relying on from a control standpoint.

Harold Byun:

And so, case in point, Marriott was using TDE with roughly half a billion records stolen. Anybody with access to the database, any attacker moving laterally in the network is going to get access to your data if you're using this type of model. If we put it in the context of what that actually looks like within the threat model we've been discussing, we have our container base or at-rest encryption model. When we look at this, if an attacker is moving laterally in your network and if you believe in the zero trust model or the notion that you're already operating from an assumed breach posture, then you have to grant that an attacker is already moving laterally in your network. That means that in this scenario they have carte blanche full access to your data in the clear, so you remain unprotected.

Harold Byun:

Similarly, if they are able to compromise the application tier through some type of cross-site scripting vulnerability or, in the case of something like Equifax where they were able to install some type of malicious shell or web shell access, then they're going to be able to still extract the data in the clear. And then seemingly, the industry seems very, very focused on compromised accounts understandably so, but a compromised account or a bad user or compromised third party via data sharing also extracts all of this data in the clear. Again, the risk mitigation or the controls that are in place largely do not protect against these types of modern-day attacks.

Harold Byun:

So, what actually can we do? Well, if we look at data centric methods to protect data, there's a range of different capabilities within this notion of what a lot of customers are calling data protection service or data protection services. This is a common architectural security construct that includes a number of different capabilities, obviously usage monitoring and access control to restrict access from the different tiers. There's a number of mitigation capabilities that can be put in place in those areas.

Harold Byun:

Data masking to control data at the presentation layer or for exposure or containing exposure to third parties, tokenization and format preserving encryption or FPE to better comply with a lot of privacy regulations and protect PII. Field and record level encryption also fall into that bucket. BYOK or bring your own key is obviously a method to source your own keys when you're working with different providers or support for multiple keys with a right to disable or revoke a key and ensure that your data remains protected. There's a number of areas or a number of capabilities and use cases that are emerging.

Harold Byun:

Encryption in use. Some people are familiar with homomorphic encryption, which is the ability to operate on encrypted data, and that leads into secure data sharing. As well as the ability to leverage different sources of key material for the data protection service to use from any type of hardware security module or HSM. Key management stores, so AWS, KMS. Thales who owns the encryption world apparently at this point including Gemalto and KeySecure and Vormetric. So, vendors like that. And then there's other folks like HashiCorp with a secrets manager that some customers are using to store key material as well and also using those for privilege access management. This is a portfolio of different capabilities that your organization may want to consider looking at as they embark on different data protection and privacy strategies to better protect sensitive information.

Harold Byun:

Now, when we put this in the context of applying some of these mitigation strategies to the threat model, one method is obviously data-centric encryption, which would be record level or field level encryption or tokenization, or format preserving encryption. These are methods where the privileged users or insiders or attackers moving laterally at the database will get encrypted data. Logs and memory dumps will also be encrypted using this type of methodology. And so, it is a more effective mitigation strategy than using something like encryption at rest, or transparent data encryption or TDE, or volume level at-rest encryption. This is one encryption part of the mitigation strategy.

Harold Byun:

Another is obviously data-centric masking and redaction, being able to implement this type of masking of sensitive data values. A number of folks who are sharing information are using clones of production data and third-party developers to build applications. On top of those production clones, a huge risk to your business. A huge risk to your business in terms of who that developer is and what organization you're working for, in what country and accessing a set of data to build an application for your business. But in many cases, those users are over privileged in terms of the data access that they're actually getting. And so, this is a method to contain and obviously redact sensitive information but still giving people the ability to work with a live data set or a pseudo live data set to build an application.

Harold Byun:

And then there's this notion of what we call dynamic entitlements along with the right of revocation. Cell level encryption or record level encryption really is associating a given record or set of records with a data owner or an entity. And so, the notion of being able to take a mass scale data structure and then segment it based on either an individual consumer or a given entity, and ultimately providing a right of replication to that data by effectively mapping an entitlement.

Harold Byun:

There's all types of entitlements that can be granted. One could be obviously granted based on this notion of an owner or a data owner, but it can also be granted based on context. A user coming in from a certain geographical region without two factor, running a set of requests for data in a high volume where they've never asked for that information before, that effectively is user context and attribution that can then be applied to a dynamic entitlement. The entitlement can be either granting the right to see the data or revoking that right. That's effectively another mitigation strategy within that threat model.

Harold Byun:

When we look at this within the context of a data-centric protection model for data breaches, again, using that same access chain, the attacker moving laterally, or the privileged user, would get encrypted data using access control. Malformed apps or compromised application footprint or microservices that are unauthenticated, or even serverless code that's executing can basically be restricted from accessing or making requests to the data, and then the data can also be masked to minimize that exposure. And then, from a user perspective, using the dynamic data entitlements capabilities that we were talking about, additional restrictions could be put in place to mitigate the exfiltration of data or the overexposure of data.

Harold Byun:

I'm not purporting that this is a panacea and a cure-all for data breaches, but as many of you who are working in security know, a lot of this is all about mitigating risk. This is a significantly stronger scenario to be in versus the open access channel that we have today where all the data is sitting there, exposed and in the clear.

Harold Byun:

Let's look into how we could potentially implement some of these controls. What I'm going to do here is see what I can do to actually show you this live really quickly and then we'll get back into the next phase of this, which is going to be privacy preserving analytics. We have some demonstrations around privacy preserving analytics as well. Let me see if I can do this correctly. Bear with me just a moment here. I'm just going to share the entire screen. Again, if you have questions, feel free to ask away as we go through this.

Harold Byun:

What I'm going to do here is I'm going to show you a live data set. What I'm going to do is walk you through data-centric encryption and masking as well as the record level control so that you can see some of the dynamic entitlements just to give you a flavor of some methods that you could use to implement and control some of the risk of over access to data. I'm going to get into this particular system here and work with a given database, which is actually going to be over here. When I go into this data structure, you'll be able to see that this is just a sample data set of potentially sensitive information. It could be health information, it could be customer information, it could be whatever you want it to be but it's obviously a fake data set. So, it's in the clear things like customer name, customer ID, city, country, [inaudible 00:26:35] potentially sensitive PII.

Harold Byun:

What I'm going to do is go back to the setup here, and I'm actually going to encrypt the data here. I'm going to apply this encryption. If I go over to grab this IP, I'm basically pointing to a given host, which is going to be hosting my data protection service and I'm authenticating using SSH. What this does is it lets me pick that same data structure that I was just looking at, pick a key manager that we want to actually source for keys. I'm going to use AWS KMS and then I'm going to pick the data store that I was just in, I'm going to say I want to encrypt city and customer name, and I can map key IDs. Effectively, what I'm doing here is I'm going to say go.

Harold Byun:

Typically, the data-centric controls that I've been walking you through within that threat model often require significant application changes, significant development and a lot of operational overhead to implement. What we're doing here is we're going to implement those same level of data-centric controls with no application code modification, no development and effectively orchestrate the entire process end to end.

Harold Byun:

What's going on right now is I'm talking to this encryption key store. You'll see, if you'll recall, I used key ID six and seven, which are my data encryption keys, and those are actually protected by the master key. What's actually going on right now is I basically said let's get this data protection service installed and oh, by the way, go retrieve key six and seven and decrypt them using the master key and then apply those and actually encrypt the data. In this case, I think we selected city and customer name to actually select that data and go encrypt the data while you're at it. That's what's actually going on right now.

Harold Byun:

That process has actually been completed in many ways. It's an encryption orchestration dance or data protection dance that's going on. And so, if I go into this dataset and refresh, you can see that we've now encrypted city and customer name. We didn't have to engage any developers. Effectively, the state of protection layer is now protecting that back-end database at the field level.

Harold Byun:

Now, what we're also able to do is obviously, without any code modification, decrypt that data for the presentation layer so that it is actually usable at the application space. In this case, you'll see that we are decrypting that data on the fly using this encrypt decrypt function within that data protection service. Similarly, what we could also do is choose to mask the data to better protect it at the presentation layer.

Harold Byun:

This is an RDS database in Amazon. You'll see that in this particular case, we've only encrypted one field of the dataset. The rest is in the clear. But when I go through the masking function, we actually can dynamically mask the data at the presentation layer. You'll see that we've used the [inaudible 00:29:52] for confidential here or Xs for city. We've randomly generated numeric data in a given format for the customer ID. We've replaced the order date, which could be a birth date used to calculate age, and we partially masked a subset of the data and use the last four digits which is often used for social security number or credit card information. It is completely invisible to the application.

Harold Byun:

Using something like Excel just to prove the point, if I went in and wanted to grab information as a user, I can go in through Excel, try and attach to the data source. And if we chose to, we could basically present that data in a mass form. You can effectively entitle users to see what data they should be seeing or should not be seeing. Effectively, this is obviously the preview for this, but it is masking that data completely invisible to the application and without significant modification. Hopefully, that gives you a quick sense of that.

Harold Byun:

One other quick thing that I'll show you, I'm running a little bit over time here, so I'm going to just barrel through this. It is just a quick data set here where I am accessing data to show you record level encryption. You'll see here that in this particular set, I have an encrypted data value in this left-hand column. What I'm going to do here is I'm going to disable a key just to simulate killing a key, which would be the right to revoke data from an organization. [inaudible 00:31:44] just get over here. Again, without any changes to an application, if I came in and ran that same query, you can see that we're obviously decrypting the data values, but certain data sets are masked because they are encrypted using the key that I just killed. This is effectively applying a dynamic entitlement along with the right to revoke data or the right to be forgotten at the record level. These are some of the basic data-centric controls that might be available to you or your organization as you look at protecting data.

Harold Byun:

Moving ahead, and again, feel free to ask questions. I will get to those very shortly. Let's talk a little bit about privacy preserving analytics and secure data sharing. We covered this a little bit earlier. Again, you guys can read it. The deck will be available to you. Privacy preserving analytics is a method to perform operations and analysis on data without ever exposing the underlying values. This is the reference from the Gartner report on privacy preserving analytics. Again, it's available on our website.

Harold Byun:

Secure data sharing is a method that allows data to be used for intelligence or aggregate analysis across multiple parties without ever revealing the underlying data values or violating the data privacy contract. For many of you who are familiar with this notion of third-party data risk or are looking to monetize data by selling it, obviously, that's a huge no-no within the context of CCPA. But if you're able to facilitate it in a way where it is not directly being sold as an individual consumer or you are leveraging aggregate techniques that don't violate the privacy contract, there are ways that you can help control and facilitate secure data sharing.

Harold Byun:

Some of the use cases around this in particular are things like data reporting as a service, or data analytics as a service, or data as a service, or data modeling as a service. In many ways, a number of organizations seemingly, again, recognize that data is the new money in modern day age and are looking at ways where they can monetize that information but also control who can access that information and, in some cases, license and revoke the license to that information.

Harold Byun:

One of the ways to do that is to leverage advanced computational capabilities like privacy preserving analytics and secure data sharing to effectively consolidate your data stores and grant rights using some of the entitlement's mechanisms that we've already been showing and the access control mechanisms that we've been showing, but also still allow those third parties to process the information and ask questions of the data. That's the tricky part. Typically, when you're implementing a lot of these controls, there's a lot of breakage in business process. There's some compromises that need to be made in terms of how you would actually operationalize or share this information. This actually facilitates a way where you can run aggregate analytics on data. It will remain encrypted in use, in memory but the operations will still return the intelligence that a given subscriber needs in order to derive value out of the data set.

Harold Byun:

Cross-party data sharing is another scenario where a lot of use cases around this. Fraud detection, aggregate fraud detection across a retail entity, across multiple retail entities. In this case, in particular, there was an instance where a financial services customer wanted to use cell tower data from a telecom provider to determine whether or not there was potential fraud in terms of bank activity or bank account access. Typically, that would be a huge non-starter for both organizations, neither entity being willing to share any of their user data. In many cases also it's being regulated to not share that data. But if there is a method to perform dynamic joins or dynamic lookups on unique identifiers and run those comparative operations, you can still derive value out of an aggregate data set.

Harold Byun:

Another one that we're seeing emerge a fair amount is threat intelligence sharing. Organizations where there is, again, the ability to confirm the existence of a threat from a cyber perspective by mapping IOCs without actually revealing the entire IOC database. Those are scenarios where we can look at actually augmenting the sharing capability or effectively sharing data without actually sharing it.

Harold Byun:

Let me show you what this actually looks like in terms of some privacy preserving analytics use cases. In this first scenario, what we have here is a data set where we have encrypted two particular columns. Target IP and threat data are actually encrypted data values here. What we've done in this particular instance is front ended this dataset with an application called Tableau. Some of you may be familiar with this. This is a common business intelligence application. With Tableau, what we can do is re-render the data set and show visual trending of the data in aggregate. This is threat data being aggregated and trended over time in a live visualization as well as a cross tabulation frequency count of target IP and threat data.

Harold Byun:

Again, both encrypted values. But being able to run this type of processing and operation on the encrypted data values and using an off the shelf application is a reflection of a couple things. One, we are able to operate on the encrypted data and perform those analytics. Two, we are doing it with a commercial off the shelf third party application, which is also guaranteeing that there was no source code modification, or no application developer involved because we don't have access to the Tableau source code. It's actually owned by Salesforce now. This is one example of how you could utilize privacy preserving analytics. This could be for sales information, how many people have a disease in a given zip code or a given region while still not violating, again, the privacy contract.

Harold Byun:

The way we're actually doing this is a technique that we call, for those of you that are interested, I apologize, I'm going to get into a little bit of a technical discussion here. Well, mildly technical. I'll cover it very briefly. If you have interest, please don't hesitate to reach out. This is actually how we actually perform this. We utilize a technique called secure multiparty compute. The way that this works is we actually source an encryption key to encrypt your data, but we treat the data store as an untrusted entity, so it has no key present. Your data would live in an untrusted entity. An untrusted entity could be a database running in the cloud or it could be a database hosted in a foreign nation state that you deem hostile as a multinational. But effectively, that is data that is living in that database, encrypted with no key present, and so it represents an opaque data store.

Harold Byun:

When we want to actually operate or run aggregate analytics or a mathematical operation on the data. We utilize this cryptographic technique known as multi-party compute or SMPC. Just case in point, Google Cloud actually implemented a form of multi-party compute this past summer to perform some additional opaque operations, so it isn't smoke and mirrors. Our cryptographic technique is based on a mathematical algorithm authored by the co-inventor of H-Mac, we just happen to be the first ones to apply it to this type of data analysis and data protection model.

Harold Byun:

What actually happens when we see an aggregate or mathematical operation come across, we're able to process the operation without ever seeing the encrypted data and without ever seeing the clear text data. We don't make a copy of the clear text data, we don't extract it into an index, we don't do any funny business. It is a pure play operational model that allows for running calculations on the encrypted data without ever seeing the encrypted or clear text values. And so, effectively, what's happening is a high-speed shell game of hide the encryption key from the data. While we do that, we're able to process those results and return the data set in encrypted form and then present it in the clear when necessary at the presentation layer. It is an opaque computing model, and that's effectively what's going on. As it relates to secure data sharing, these are some of the use cases that emerge out of that.

Harold Byun:

Again, I glossed over that. Happy to get into more detail with anybody who wants to dig in deeper. We've had a ton of third-party organizations basically going deep on the cryptographic technique and we're very confident in terms of methodology.

Harold Byun:

When we look at how this can be applied to secure data sharing, and I know it's a big leap, if you can believe what I just said is true, that we're able to treat this encrypted database as an opaque model of data with no key present, then in theory, it would be rational for you to also say that that database could live anywhere. Because the data is encrypted, there is no key present and assuming quantum isn't a reality yet, then there is no method to break into that database.

Harold Byun:

And so, if that were the case, then in theory, you could also publish whatever data you wanted into that encrypted database and nobody would be able to steal it because it's a database that lives there without any key present. And so, if we wanted to share data across multiple organizations, one way that we could do that is we could have organization one use their own key, number one, and encrypt the data and store it in that database. And then we could have organization two or three or four or however many suppliers or shares you want, have their own keys, and actually also query the data.

Harold Byun:

The beauty of this model is that when we apply this model, it allows for organization one, the publisher and owner, to publish data in a secure manner without ever exposing their keys, and organizations two through N can query the data and run analytics on whatever subset of data they are granted and derive intelligence out of that information. In this case, what we're doing is organization two is going to look up on an IOC and get frequency counts and contextual information without ever being able to see the entire data set or the original data values.

Harold Byun:

Let me show you what that actually looks like. This is a pretty rudimentary demo, but it should give you a general sense of some of the capabilities that we can enable. What I have here is I have an IOC database with 31 records. If I actually run this particular query, you can see that the hash values are actually encrypted in this database. Obviously, there's some contextual information that's pretty rudimentary. Now, if I come through organization two and I want to actually run a query, what I can do is I can run the same count query. So, I'm going to run this count query. You see I have 31 records. Now, if I want to look up a given hash and get a frequency count there and I get 24, there are roughly 24 records that match that hash. If I wanted to, I could also actually run a different hash lookup and there's zero.

Harold Byun:

Now, if I wanted to extend this further, what I could do is gather additional attribution. If I wanted to get attribution around a given IOC, I'm getting some attribution without getting free form untethered access to the entire data set. On this particular lookup, I'm getting different contextual information about this given IOC. In this particular case, I think I created this one. I have deactors, Bad Harold. That's me.

Harold Byun:

Just to prove the point that this is a live lookup, what I'll do here is I'm going to go back to organization one and I'm going to insert evil hash Bad Howard. And so, organization one is actually adding data into this dataset. And when I go back and actually execute accounts, I'm now at 32. You'll see that as I execute this, I now have an additional record from Bad Howard. And if I go back to my organization two and want to look up on this evil hash, we're going to get different contexts and attribution and a live lookup.

Harold Byun:

That is a data model of sharing without actually revealing the underlying [inaudible 00:45:38]. So, there's a number of different context where this could come into play. It could be, give me all the customers and zip codes 90,000 through 94,000 or whatever and having this range of credit scores, or give me the set of patients that have these types of attributes within this age group without, again, not ever revealing some of the sensitive data sets or PII information. That's effectively the type of data sharing that this capability opens up.

Harold Byun:

Lastly, for folks who still don't believe that this is truly possible, just to show you some of the range of capabilities and operations that can be supported with these techniques because they are emerging from a privacy preservation standpoint. I have a data set with 999 records here. I'll do something like search for email where it contains THE. If I search, I get down to 27 records. And then I'm going to add another search and do last name that contains TH, and I do another search and I'm down to seven records and you can see all the last names. Fairweather, Strother, Metheringham all have TH as well as the emails have THE. People say, "Well, that's just the application. What does that really mean?"

Harold Byun:

Well, if I go direct to the database, you can see that when I select account from this database, I have 999. When I select all the records direct, the data is encrypted. And if I go through the data protection service and I run the same count, I have 999. And when I search an email with THE wildcard, I get 27 rows. You can see that down here. And when I search the email and the last name, we now have seven records return. And so, that is effectively a wildcard search on AES encrypted data operating using that multi-party compute method.

Harold Byun:

I know we're a little bit past beyond what I want to do from a time perspective. I hope this was at least interesting and engaging for you. Just in summary, encryption at-rest and container encryption methods do not adequately protect your data from modern day attacks. In the world that we're living in, distributed data, distributed access points and zero trust, you have to implement data centric protection methods going forward. And quite frankly, the data privacy regulations are mandating it. You can also look at privacy preserving analytics and secure data sharing models to really help enable your business to continue to monetize data and share information more securely while still preserving the confidentiality contract.

Harold Byun:

All the resources here are going to be at baffle.io/privacy. Some events, if you're interested, we'll be at RSA. We have free drinks and, it should be appetizers mixer. Drinks and cocktails are pretty much the same thing. Feel free to contact us or visit our website for more information. I think we're also going to be running a hack, a baffle database contest with a $5,000 cash prize if anybody would be able to steal data out of that. We'll also be at the IT Security Leadership Exchange in April. If you're interested in an invitation and are in a leadership role from a security standpoint, again, please reach out to us and we can arrange for an invite for you to attend that conference.

Harold Byun:

So, with that, let me open it up for questions. I'm going to stop sharing here for a second and see what kinds of questions we have. We have, how is PPA different than homomorphic encryption? Privacy preserving analytics is, in some cases, can be using homomorphic encryption. In our method, as we described, we're using a method called multi-party compute or MPC. The main difference is that homomorphic encryption does not require any encryption keys. It is generally considered incredibly slow and nonperformance in terms of its implementation. In order to execute queries, you have to have prior knowledge of the query and so that requires a rewrite of your application or some dedicated application to process specific tasks of the data.

Harold Byun:

The multi-party compute model that has been implemented actually supports ad hoc queries, which opens up virtually the entire range of mathematical operations. Those are a couple of the main differences in addition that the non-homomorphic models are vastly more performant to the tune of a difference of roughly a million times faster. Those are some of the main differences there.

Harold Byun:

Another question here, how should we look at addressing third parties? Does the risk increase with regulations like GDPR, CCPA? There's a lot of information out there on CCPA treatment as well as GDPRs treatment of third parties. The short answer is it actually does increase because the nature of the relationship, if they're processing data on your behalf, your organization effectively becomes responsible for that third party and their handling of the data. So, if they have a breach, that absolutely is a problem from a privacy regulation standpoint. And so, it ultimately is going to require revisions to contracts with your third-party vendors and data processors as well as re-evaluation and assessment at a minimum from a risk perspective in terms of how you want to manage that relationship and what the engagement model is.

Harold Byun:

There's a couple of final questions here. Can you support on premise deployments? Yes. I mean, our solution can support on premise deployments. I think in general, what we're hearing from customers is that they're going to be in a hybrid state for quite some time. Effectively, when you look at data protection services, you probably want to look at solutions that can extend across both, so I think that that's a pretty basic premise in terms of your evaluation. Obviously, people have different criteria.

Harold Byun:

I think that that's pretty much it. I apologize for going longer than I really wanted to but I thought that there might be a little bit dense information, but hopefully you got some good information out of it. Again, info@baffle.io or baffle.io/privacy for more information. I'm harold@baffle.io. I really appreciate your time, and let us know if there's any way we can help. Have a good day. Bye.