Harold Byun:

Hi, and welcome to today's webinar on de-identifying data and Snowflake and Amazon Redshift. My name is Harold Byun and I head up product management for a company called Baffle. We are both an Amazon and Snowflake partner, and I've been working in data protection and de-identification space for quite some time. Today we're going to be covering some capabilities and some challenges in those areas just in relation to overall data analytics and protecting information against data leakage and breaches. Obviously, the breaches continue ad nauseam, so I don't think that that's news to anybody in terms of some of the challenges and obviously, the move to cloud is accelerating. The amount of data that is being migrated to a cloud footprint and a SAS base footprint. Getting your arms around that and securing that information can be a bit challenging for organizations. And we're going to cover some methods that ideally simplify some of the data privacy mechanisms that can help accelerate the move to cloud and also reduce impact and friction with the business.

Harold Byun:

This is the agenda for today, I'm hoping to get in and out of here in roughly a little over 30 minutes and promise to try not to PowerPoint you to death. There will be live demo for the data pipeline capabilities that we have as well as ingesting data in Snowflake. And we'll be covering these areas. Overall trends and privacy challenges, some different methods around de-identification and controls gaps, and a review of some of the architectural models that we're seeing. Time permitting, we can get into some of the privacy preserving analytics capabilities, just to see if there's interest as well in terms of secure computation in other areas within data analytics. I'd encourage you to use the chat panel to ask questions throughout. Feel free to interrupt at any time. And you can always email info@baffle.io, or my email is harold@baffle.io if you have questions. A little bit about me, I've been working in security for roughly 30 years on the architecture side as well as on the vendor and product side. The bulk of my career has been on Data Containment, Data Isolation, and I've been at Baffle for a little over three years at this point.

Harold Byun:

In terms of what we're seeing, I've jumped right into key trends around analytics, is really just obviously, there's a huge... You can just watch your TV, and you'll see the AI advertisements that are just coming out from a number of different software vendors and providers. The reality is that obviously there's big investments in this area, which requires an aggregate of data and at the same time it runs in conflict with a lot of data privacy requirements. In order for businesses to continue to monetize data and gain or derive business intelligence from the information, they need to collect more of it. And yet, there's obviously a pretty strong pushback against privacy violations and, I guess, egregious practices around information gathering. And so just some key stats, I mean, roughly the trends that we're seeing are three quarters of organizations are going to be moving into some type of AI initiative. Gardner also put out a statement around 75% of the world's data is going to be moving to cloud databases and cloud data infrastructure.

Harold Byun:

Some other stuff, obviously, people are looking at data model as a service or data as a service in terms of organizations being able to provide some type of subscribership or white label offering around data analytics and data modeling. And that's something that we're seeing absolutely come to fruition now. I

actually just got off from the phone call with a large European bank that is embarking on such an effort and are working with other organizations that had similar requirements.

Harold Byun:

The other thing that's driving this is we're running into a lot of organizations that have made investments in Hadoop or big data infrastructure on-premise. And quite frankly, those environments are static and are running out of room. And what we're seeing more and more is organizations pushing more data to the cloud in a continuous mode to basically get off of these static environments. Or they have a legacy mainframe environment or a proprietary database environment that just isn't scaling for today's world. And so they're more and more pushing data in a continuous mode into cloud infrastructure. And that's something that we're running into in terms of the nimbleness of traditional on-premise or colo infrastructure versus something that's more flexible from a cloud-base footprint. And then this is data points, and then there's obviously continued data leakage or breaches and challenges in that space. And I don't think that that's a surprise for anybody.

Harold Byun:

What are we seeing, when we see this typical on-premise environment where you have obviously applications and data stores or even IoT or point of sale data that's sitting on-premise. We're seeing pretty much a wholesale move into cloud data lakes, and that could be in the form of an S3 bucket, it could be a Snowflake, it could be different data warehousing in the GCP or in the Azure world. Similarly, same movement into Azure Blob storage and Azure Data Lake services where people are looking to create a common pool of data that has different subscribers and ultimately also different rights to the data. And I think that's a common thing that we're hearing people emerge with as just different subscribers require different views. And so I've worked with one consumer product goods company that referred to this as the data scientists in our organization want to build a data ocean. And we're not going to call it a data lake, they want to build a data ocean and yet within that organization there are other data subscribers like the supply chain management and inventory folks on the retail side, the in store retail managers or managers of the retail branch per se.

Harold Byun:

And all of those subscribers whether they be a data scientist, or somebody managing the overall supply chain, or somebody managing a specific branch or instantiation of the physical brick-and-mortar store require different views on the data. And that presents its own set of challenges when you're building this common pool of information. This type of role-based access control or role-based viewing lights is something that we're hearing a fair amount of that as well.

Harold Byun:

In terms of overall data, privacy challenges, again, breaches are just almost old news, an old hat at this point. One other interesting aspect of this is when you look at item number two, particularly within the realm of data modeling as a service or offering a data subscribing or data recording service roughly 60% of CISOs have reported some type of data leakage via third party. This data point from the Ponemon Institute, is a couple of years old at this point, but it holds true as it relates to third party risk. And that's been a challenge in the industry, and particularly for security for quite some time. But as you start moving towards a more multi-party subscriber model, multi-party, multi-viewer type of subscriber

model as it relates data, the risks obviously increase. And then three related to cloud storage in particular, over billion records have been leaked. The access control permissions are difficult to work with often leading people to just make things public. In a lot of ways, it's sometimes easier just to do that. And people say, "Oh, I'll revert back if somebody needs the access right now." And oftentimes, that gets left open.

Harold Byun:

This is from a survey from the 451 group actually, where they were doing a survey on some of the biggest data management and analytics challenges and data security obviously top the list. Data privacy is pretty high up there in the rankings as well. So, definitely significant concerns for organizations going forward in terms of getting their arms around that. And then obviously the regulations around data privacy continue to expand in terms of their coverage and in terms of the restrictions that are required governing data as well as providing some method of data replication or a right to be forgotten. And the geo residency requirements also associated with that. Those present a lot of additional challenges as well.

Harold Byun:

This was right after a CCPA went active. The law actually became effective in January, I think of last year, but actually wasn't going to be enforceable until July of last year. And then I think this was literally three days after the enforcement date. This lawsuit came about and there are obviously additional data privacy challenges. And in addition to even just the fines which I think if you go back to something like the Capital One breach, they were fined most recently around $80 million, but it's the additional monitoring and auditing controls that you then need to instrument which are also ultimately going to slow down your business and your ability to operate effectively that introduce a ton of overhead as well. This is a web page that we have, if you're interested in more resources, you can just go to baffle.io/privacy, there's a couple papers up there and some additional content around data privacy regulations. We also have a blog that has a range of articles. Our CEO and co-founder Ameesh Divatia, is often published in several articles around the web on data privacy. There's a fair amount of resources available to you for information.

Harold Byun:

For this statement, I mean, I think in spite of privacy and I think I saw a report the other day on all the stuff that Google Chrome and Google collect, in terms of data about you and there were similar presentations early on the year on Facebook. I mean, the reality is people are just collecting data regardless. And that's just the new realm that we're dealing with in many ways.

Harold Byun:

This is a view of methods in terms of the overall entitlement of data versus who should actually view it. Again, talking about differential views and how that's progressed and really what types of controls and restrictive controls can be potentially put in place from a framework perspective to minimize the exposure. This is actually related partly to controls, but also to data analytics and metadata in different ways to constrain actual access to data and who needs to have access to that information related to analytics, partly within the context of again privacy preserving analytics and secured computation.

Harold Byun:

What are some common methods for data de-identification? When we look at overall data control, especially as it relates to the cloud, there's what's commonly well known as the shared responsibility model. And the net then on the shared responsibility model is that the infrastructure, the cloud infrastructure provider is responsible for the security of the infrastructure, the security of the network and physical security as well as providing you with configuration controls to implement on the platform. You the customer are actually responsible for tuning the controls and configuring them appropriately, as well as securing the actual data that you're putting into the infrastructure. And that was how generally the line is drawn in that shared responsibility model in terms of what the customer is responsible for and what the infrastructure provider is responsible for.

Harold Byun:

And when we look at some of these common controls, I mean, there's a ton of security controls available from these cloud providers. And so this is just a cross sectional look of some different controls that are available in AWS, in Microsoft Azure, and in terms of whether that's bucket controls, access control list, scanning of data, monitoring and logging, a bunch of different types of encryption controls and key management integration with their respective key management services, transport layer encryption and then obviously virtual private clouds or VPCs, as well as a private link even or direct link type connections into these infrastructures.

Harold Byun:

These slides will all be made available to you, as well. And then here's just a listing and some common de-identification and data protection mechanisms. Tokenization is commonly known, there's a number of different variants on data encryption. And then there's also a full format preserving encryption, which is a data type preserving, data length preserving mechanism that still encrypts the data. A lot of people are familiar with volume-based encryption, what we call object level, not object level, I mean, volume-based or container-based encryption models, such as database encryption, or bucket encryption, or storage container encryption. Most of those methods are good from a compliance perspective. The container models that we see, I think our main position against around containers and against container encryption is largely that they don't really protect you at all in the face of a modern day attack or hack. And that's one of the predominant problems that we see around data breaches and leakage is that everybody says, "Oh, well, I've enabled encryption." And it's like, well, yeah, that doesn't do anything for you. If the attacker has access to the system or the infrastructure, they're going to see your data in the clear. Where is what we're seeing commonly more implemented is value-base or data centric protection measures which are actually protecting the data values.

Harold Byun:

And so in the world of overhyped zero trust network type scenarios, while that sector is overhyped, the reality is that attackers are going to get at your data, attackers are going to be in your infrastructure. And if you submit to that fact, then the logical step is to protect the actual data values which the attackers will gain access to. And that's what data centric protection does. And so additional measures are things like data masking or role-based data masking. And then privacy preserving analytics or Advanced Encryption is a mechanism to enable secure computation, which is still a nascent but emerging space that a lot of folks are investing in.

Harold Byun:

This is just an example of some de-identified data, the lower half of the screen is actually clear text data, and the top half of the screen is a representation or... Well it is, not just representation, it is format preserving encryption. And so you can see that, for example, the CC field are credit cards, for example. And so those are length preserved encrypted credit card numbers that actually pass a credit card Luhn check or validation check. Things like the birth date are randomized but they pass a data validation check. And then in the email field, you can see that you probably make out barely in the email field that there's an add sign and a period before the domain extension or the domain tier for the email, in the first example. In the second example, it's just the email precursor or the actual email address outside of the domain that's actually encrypted. There's different ways that you can vary this type of format preserving mode but it represents a method that can easily de-identify data and effectively the data structures don't know the difference.

Harold Byun:

What are some of the key benefits of using de-identification? Well, one, obviously you're protecting data inside objects and files or containers, especially as it relates to the pipeline. There's a lot of object or file level encryption, again, and there's a lot of container-based encryption, but ultimately, if you are able to get the file or access to the file system you get all the data in the clear. De-identifying this type of data if you believe that you are going to be breached or subject to a breach or believe that attackers are in the network already, which given things like sunburst and release attacks against even SNS compromises as two factor auth, people are writing about in some of the other more significant egregious breaches, then you have to recognize that attackers are already in your networks. And if they're moving laterally with access to the systems, they're going to get access to the data and these types of de-identification methods provide you with additional safe harbor. And then the third benefit is really unblocking the business and having the security team move in lockstep with the business in the DevOps World and being able to accelerate cloud programs or move to cloud and cloud data analytics programs and unblocking the move to cloud by ensuring that security is moving in lockstep with that data migration process.

Harold Byun:

What does this look like from a data pipeline perspective. As a number of different scenarios, I mean, we talked about this one before where you have people using S3 as a very flexible, scalable, and economical data lake. That's one mode that we're seeing. These are other examples where you've got people pipelining data into these types of cloud data lakes and then retrieving them out of different data warehousing and analytics solutions, whether that be exposing it through an Athena Interface or using something like Snowflake or exposing it to a data modeling service like a SageMaker forecast or TensorFlow or another analytic solution.

Harold Byun:

And this is an example of commonly what we're seeing from a pattern perspective organizations where, as I spoke to earlier, they're moving from an on-premise database, or an on-premise data structure. They're using some type of Migration Service, whether that be DMS or a third party data mover solution. And they're moving that to cloud in mass. And so one of the things that we're seeing from a challenge perspective is when organizations want to de-identify this data they're often in a position

where they have to create multiple clones or they've got to migrate things or create multiple copies and then transform the data, and then stage it, and then ingest it. Or they're in a position where they have to move all the data into a cloud data lake solution in the clear and then impost. Actually perform some type of data de-identification process or transform which on large scale data sets or petabytes of data can be completely operationally inefficient and untenable.

Harold Byun:

One of the other approaches that we're seeing is de-identifying that data on the fly and selectively re-identifying that again based on who should have access to a given view of the data. And that's where we see some opportunities to streamline this, the ability to embed security into the actual data movement process so that as you're adopting these types of data warehouses and offering these data services or reporting services for clients via these analytics and Warehouse Solutions, you can actually build security straight into the model, and it's not an afterthought.

Harold Byun:

I'm going to jump into full screen mode here and walk you through a quick demo. Just going to pull my screen together here really quick before I show you this. Get into here and share the screen. Give me one second. Okay, I think I'm sharing. I'm I sharing? I don't know if I'm sharing. I was sharing, okay. We're going to walk through this demo. Basically, what I have going here is I have an on-premise database, let me just touch base on this. And in this data structure, I'm going to be pulling from this on-premise database. And this on-premise database should be this. That's where I have the appropriate rights. Okay, I got to bump up a security group, bear with me two seconds more, I apologize. Just going to take two seconds here. All right, let me share that again. Okay. We should be good. Connect now. Okay. Looks like I can connect. I go into this data structure. This is going to be my on-premise Microsoft SQL Server database. All kinds of sensitive data. So obviously, dummy data that customer information, sensory sensitive data and it's in the clear, because it's on-premise and well protected.

Harold Byun:

What I'm going to do here is actually migrate this data. And I'm going to move this data into this S3 bucket. And you'll see that here, there's this S3 bucket, forget this folder, actually I'll just delete it, just so you can see this completely empty. And the next thing is delete. And if I go back, you'll see that this bucket is completely empty and as a third part of this, I have a Snowflake environment. And if I run this, actually, I'm going to delete this out of the Snowflake environment so you can see that when I run this query, I have basically an empty table in Snowflake. There's no data, all the data is in that Microsoft SQL Server. And so what I'm going to do is move that data into Snowflake and de-identify it on the fly. And so if I go back here and start this, what this is going to do is it's going to pull from the Microsoft SQL Server on-premise database, run through Baffle where we have what we call the Baffle Shield. It's going to de-identify the data on the fly, land it in S3 in an encrypted state, and there we can use that S3 bucket as staging and via Snowpipe, basically just copy it into Snowflake in a de-identified state.

Harold Byun:

And then what we can also then do is selectively re-identify that data and report on it. That's what's going to happen here. You'll see that this job is running and just as a reference point, what we're actually doing right now is taking the data from this on-premise database, we're running it through this

Baffle Shield, we're going to land it in this bucket, and then we're going to pipeline it into Snowflake and show you how that end-to-end process works. This should be almost done here. A lot of this is just spin up time, it's probably already in the bucket anyway. Just refresh this bucket, you'll see that now there's this folder. This is the table that I've been referencing, and this is the file. I'm going to download this file so that you can see that it actually is encrypted. Numbers is too slow to load so I'm going to just find Microsoft Excel really quick. I'm going to open this. And you can see that in this particular file, this column which represents customer name, I think that this is customer ID, which is also this is using that format preserving mechanism. Let me zoom in, so you can all make that out, are now encrypted. And then this is the sales field over here. The sales field used to be sales numbers, it's now just randomized integers.

Harold Byun:

Let's go back over here. And so this is landed in the clear into the S3 staging environment. Now if I go over to Snowflake, I'm going to copy that into my environment. So again, you see that this was empty when we queried it. And then what I'm going to do is I'm just going to copy it from staging. So I'm going to consume this into Snowflake, and then what you'll see is that when I run this query, we now have this data. And again it's city, actually didn't point that out, but the customer ID is encrypted, the customer name is also encrypted and then the sales number is also encrypted, where they're not identifiable. Then you can run in this account admin role.

Harold Byun:

Now, if I change the role to a authorized role within Snowflake, what we can also then do is run the same query. And what will happen is actually we are spinning up an external function in Snowflake that will launch against a set of Baffle API's running outside of Snowflake, these would run in your own VPC and you can see that we've now decrypted this data. The customer ID and the customer name and the sales numbers are all decrypted. And actually, I can sort on those numbers. Even though that they were in an encrypted state, I could get my top 10 sellers, if I wanted to, run a report, I could actually do a sum of sales where I get the total sales number, I'd get the max sales if I want. These are all things where their underlying data inside Snowflake at the data value layer is persisted only in an encrypted state. And then S3 in the bucket and the data lake, it's also only persisted in a de-identified state.

Harold Byun:

Within the cloud footprint all of the data remains in a de-identified state in this model. And then if I go to what we're actually doing, this is effectively, well, this is actually Microsoft Azure, the same principles apply for AWS or it doesn't really matter. But the client program in this case, it's the Snowflake Console, I'm executing the query, we're using the masking profile and invoking Baffle external function that then hooks into a separate VPC and executes these serverless functions and integrates with your own key, whether that be Azure key vault, it could be HashiCorp, it could be third party enterprise, key management solution, like Thales or an HSM, it could be AWS KMS, doesn't really matter to us. The point being that we're able to facilitate this while delivering on this end-to-end de-identified data pipeline, and then selectively re-identifying the data also without breaking any of the reporting or business analytics functionality. You're going to get less friction with your business users, less hindrance to adoption, and you can integrate security mapped into this overall cloud movement process and a modern data pipeline. It's kind of the Snowflake integration in a nutshell.

Harold Byun:

This is another alternate infrastructure. And you'll see here that we have what's known as this Baffle Shield. That Baffle Shield is what I used in that migration process as we move the data from the on-premise SQL server to the S3 bucket, we can use that same Baffle Shield with Amazon Redshift as well, which is another data warehouse or analytics provider and can support in line de-identification and masking, as well as selective re-identification. Again, another seamless solution for you. All of these are tools that ideally make it easier to ensure that data privacy is easily implemented as your business continues to move more data into cloud.

Harold Byun:

I'm going to jump into this very, very quickly, and then I'll wrap it up with questions. But in terms of privacy preserving analytics, one of the things that you just saw with Snowflake in particular is the ability for us to, what's the word, perform analytics on data and basically not break any reporting. And we can do the same thing with other traditional databases like Amazon RDS or other database platform as a service providers. We have a partnership with IBM Cloud as well. Really cloud provider agnostic in terms of overall footprint and how you want to use this. And so the basis for privacy preserving analytics is really allowing for business intelligence and analytics to run without ever having the data decrypt it. It's basically mathematical computation on encrypted data which, I know sounds like it's very far fetched, but it's become more and more of a reality and we were one of the first pioneers in this space to actually... The basis of how this company was founded.

Harold Byun:

But basically, it facilitates this notion of data as a service and selective re-identification of data while using a common federated data store, an untrusted data store where all the data is encrypted. But these operations can still occur to report out and run things like dynamic joins or selective frequency counts on an aggregate data set as well as other types of models. And so this is one of the ways that we've been advancing these types of use cases, just an example of healthcare data sharing. We have one of the largest vendors in the hospital servicing space, they're in the presence in over 2000 hospitals that is using us for building a patient data service. But the idea here is that if you had multiple parties submitting data and nobody trusted each other and nobody trusted the data store holder, basically, you have a bunch of unwilling participants. You have unwilling participants in terms of the people who own the data and you have an unwilling or you have an untrusted party in terms of who actually is holding the data. And for a lot of people, that's the cloud provider. The cloud provider, for better or for worse and rational or not, is an untrusted entity.

Harold Byun:

And so what we're able to do is deposit data into the shared database and basically perform matches and analysis on that encrypted data, where each data owner holds their own key and the data values are encrypted with their own key but the analytics still return information. And so there's a number of different scenarios where this could be valid. So obviously, a patient who might have condition one, yes or no. And in today's world, with the pandemic going on, having condition one would be good information to know given a broader population or perhaps across nation states who don't want to share information and the like. Similar concepts apply or principles apply in the realm of threat intelligence data sharing or fraud analytics and data sharing or in terms of insurance modeling and

things like that. Those are all scenarios where these types of privacy preserving analytics capabilities could apply.

Harold Byun:

Then summary, and I'll open it up for questions, you can leverage code data lakes safely to improve your agility, improve your speed, improve the flexibility that you have to accommodate large scale data growth. And that these data centric production methods while historically have been incredibly difficult to implement, really really strong controls against data leakage and there are methods available that can quite frankly simplify that for you so that you can have security move at the same speed with the business as it relates to protecting data that's moving in mass to the cloud. I hope some of this is useful information. I don't want to keep everybody too long. There are additional resources here. All the slide deck will be made available to you as well and you can always reach out to us for additional questions.

Harold Byun:

Let me stop here. Looks like there's a bunch of questions that have come in. And so let me start. If we have to enable search on PII data, won't the indexes exposed PII? It's the first question, the answer is no, because the way that these operate or the way that this search is occurring is basically in an immutable mode, so the data again is never persisted in the clear, the query will execute. You will execute a search, and the external function will hit a separate VPC where the decryption operation occurs and that's passed back to the application tier. So the indexes are actually on encrypted data, they're not actually on the clear text data.

Harold Byun:

Separate question, do you support stored procedures in legacy applications? If the data is de-identified, how will stored procedures work with that data? We do have a mechanism to handle store procedures in legacy applications, it is not a slam dunk by any means. For case in point like any stored procedure that's doing dynamic SQL is completely off limits. And that's for any type of encryption or obfuscation operation, there's no way for anybody to understand the intent of dynamic SQL. Now, if the store procedures working with literals or doing some type of just basic crud operation or a select in operation, those are things that are easily supported. And we have a mechanism so that when the store procedure is called, we basically will call and what we call a copy of that stored procedure but it references the encrypted data types because the way that the Baffle solution works is fundamentally based on what we call a privacy schema. And the privacy schema has knowledge of which columns or data fields are encrypted. And so when the store procedures and acting a simple crud operation on an encrypted column we have knowledge of that and are able to work with that stored proc. But in general, the answer is it depends. It depends on the nature of the store procedure and what it's doing. But we've had success with legacy applications and repurposing the stored procs.

Harold Byun:

How would you support searches on encrypted data? For example, the first name in the column is encrypted, how can I run a query below? Like first name is Steve. Great question, and the follow-on question to that is querying performance. And the following question on that is about homomorphic encryption. And so I'm going to answer all three of them really quickly here, let me just, there's one environment that I got to get into. The short answer here is, we don't do homomorphic encryption, we

have what we call secure computation capabilities. And so it's the ability for us to operate on encrypted data. And so that's what I showed you with Snowflake doing those reports. Now, the other example of that is something that we would do in a database. And so let me just show it to you full bore because I think a picture paints 1000 words, and there seem to be a lot of questions around this. Give me one second here. I get into that system and get into this system and I should be able to share this to you. I'm going to share my screen again. Close this down, share the screen. And I'll come back for just a couple more questions after this. And then as it relates to performance, we've been highly optimized so in a pure encrypt decrypt up scale, we've been measured one to two milliseconds of overhead.

Harold Byun:

There are organizations were operating in a multi-billion record IoT environment where they basically launched 250 million API Polls against us and there was no performance degradation. There was minimal performance overhead. I know that that sounds unrealistic, but we've been highly optimized. And then as the final test for this customer which is a Fortune 50 manufacturer. They basically put us in line in a pre-prod version of the app and didn't tell their end users. They wanted to know if anybody was going to complain. And so we've had no noticeable degradation, but we've been measured in significant scaled SAS environments of one to two milliseconds of overhead on the encrypted traffic. The rest of the data passes through us at higher speed. Homomorphic operations and things like that carry more overhead, but we're still about 1000 times faster than other homomorphic methods.

Harold Byun:

Let me run into this really quick just to give you a flavor of what this looks like. I think if I get into, I think, this environment. This is a better demo. This is using our privacy preserving analytics method, it does have a restriction on the data platform in terms of what's possible there. I know if I'm connecting here or not, I might be blocked still. It's case that one's not available let me connect here see if I'm going to... All right, this is a different version I guess. I think I'm still blocked on one of them. If I go over to this particular environment, run this query, you can see that if I run this query that customer name is encrypted and sales is encrypted. Now, if I go through this Baffle Shield, which is using our SNPC mode and I run the same look up you'll see that we decrypt that data and say a customer name is now decrypted, as well as sales. And then if I wanted to order by sales in the top 10 sellers I can do it, if I want to do some sales. And then for the wildcard search that people were asking about like AR, percent on the customer name which was encrypted, I get a subset of that. Doesn't really matter to us. I could do UVA, is another one, I'll get a different subset.

Harold Byun:

Not sure what the other names were, I'm sure that there's an O-N. So I'll just do O-N, oops. And we're able to get anything with O-N in it. I mean, we have basically made announcements on this capability four years ago. And then we were able to support wildcard search on AES encrypted data in a performant manner. And I think that that's fairly unique, but it opens up the realm of any mathematical operation which is how we roll into this realm of privacy preserving analytics and the whole data analytics pipeline. Hopefully that answers some of your questions there. Let me just wrap up with just a few more questions. See if there's any last minutes.

Harold Byun:

How do you reconcile that CCPA addresses encryption and de-identification separately? Encryption is only mentioned in this context of liability and de-identification is defined as information that can't be used to infer data about a profile of a particular consumer. I think there's a ton of legal interpretation into this, I think that one of the ways that we approach this is that if you have a requirement around both encryption and de-identification, that encryption is the better way to go. And the reason we say that is because obviously there is a requirement for encryption and simultaneously it is de-identifying the data. And so you can use encryption in conjunction as a tokenization mode, you can use it in a format preserving encryption mode, and then you can also use it in conjunction with masking. You get the encryption checkbox, and you also, with the masking capability, whereby restricting access to keys or disabling keys to get the de-identification capability set that can then be used in the pipeline.

Harold Byun:

Now interpreting that around inferring behavior around a consumer, that's a pretty broad statement in the language of some of these data privacy laws that is going to be subject to a lot of legal discussion. I don't have a great answer around it. I mean, other than one thing that you obviously can do is you're de-identifying the actual consumer as an individual, which also includes IP addresses. That's one way to get around that is that you're building a profile but it isn't trackable to individual consumer. First name, last name, social security number, driver's license, passport, all of those values are de-identified. And you're building heuristics and models around de-identified users. That's one way to tackle this problem.

Harold Byun:

The other way is that you're looking at models that are more generic and I don't think this latter mode is of any value because nobody wants things that are more generic. If I wanted to classify all of you as people who like red sweaters, that's great, but that gives me zero business value in terms of being able to quite frankly target you as somebody who might like another sweater. I think that you're going to want to see how you can explore methods to easily de-identify the actual individual consumers and still allow for applicable profiles to be built. And now I'm not a lawyer, and I don't play one on TV, but I feel like those regulatory laws as written specifically around building profiles and behaviors leave a lot to interpretation. And quite frankly, you could argue that the same thing applies to attack vectors and analysis on threat actors as you build a profile. And if I was somebody who was hacking you and I use certain types of attack methods and TPCs and things like that, that you're tracking and you build a profile on me, does that mean that I'm exempt from the law? I mean, you built the profile on me as an individual, I guess I shouldn't have been in your system in the first place. But anyway I'm getting a little theoretical.

Harold Byun:

Last question. Follow up on encryption and ID identification, what about memberships, single out attacker and furring? For example, in your database and encrypted data, you can see if you have an equal number of males and females, but you can take the same logic to identify a person. K-anonymity was invented to combat that, so yeah, I mean, the best benefit there, I mean, I've had similar questions like people where there's a doctor in a small state with a small population and a certain number of cases are aggregated at this doctor with a specialization. It's kind of a little bit of an augmentation on to the description that you're providing here. Yeah, there's ways that k-anonymity, there's other methods

where you can stuff in noise into the model. I mean, those are all kinds of really early types of mechanisms to inject a certain amount of fuzziness into the analytics process.

Harold Byun:

I think that, again, it's early days in the space, it's early days with these laws. I don't necessarily think the lawmakers are also fully in touch with, obviously, necessarily the technology aspects. I mean, I don't think I'd be the first person to say that in terms of what's capable and what's not. And so, coming up with a hard fast rule where you can't do X, and you can't do Y, and you can't do Z without looking at some of the practicality is in many ways things that are not going to be successful from a law enforcement and implementation standpoint. And so I don't really have a great answer but that's my thinking on it at this point in time. I do again, feel like one of the things that we can do with males and females as well is, if there's other identifying factors, the other one for me is deceased, yes, or no zero or one. In those models and a lot of tokenization models you're not getting a high variance in terms of how that data is de-identified. There are methods to completely randomly de-identified millions of records, where it's a zero or one bit. And still retain randomness around that to also further de-identify other identifying factors.

Harold Byun:

Not a great answer, but hey, I tried. I hope that this was of interest to all of you and useful. We're a little bit over so if you have questions or want more information, please don't hesitate to reach out to us and we're happy to help or get on a one-on-one call to discuss your specifics about your data privacy requirements. Thanks a lot. Have a great day.