Harold Byun:

Hi, and welcome to today's webinar on Implementing Data Privacy Controls for Amazon Redshift and Cloud Data Lakes. We're going to give folks just a couple moments to continue to get in here and then we'll kick things off. So appreciate your time. I'm not going to slide wear you to death, going to try not to do that today and focus more on the overall data flows and reference architecture as well as walking through more or less a live setup of what the implementation actually would look like to de-identify some of your data and more seamlessly manage it in Amazon Redshift.

Harold Byun:

And just a bit of housekeeping. So there is a chat window in the controls as well, feel free to ask any questions as we go along, much prefer these to be more interactive than not. Apologize for my hair today, I did get a haircut finally, so I'm not wearing a hat at least but it's a little bit standing on end. But I will not be on screen entirely but this will at least give you a way to interact somewhat. And we'll be kicking things off just shortly. There are also a number of attachments and links to some coauthored articles with Amazon, as well as a link to our Redshift page, which has more information, the slides will all be made available shortly and we'll go from there.

Harold Byun:

Okay, so I'm going to kick things off here. Let me just share my screen here. I'll share this first, and that way you'll be able to see the slides as we go through them. So this is going to be the talk today. My name is Harold Byun, Head of product management for a company called Baffle. And we've been working extensively with all of the cloud infrastructure providers, but particularly Amazon. And the focus today is going to be on how you can implement data privacy controls for data warehouse such as Amazon Redshift and cloud data lakes. This is the agenda that we're going to cover today. So an overview of the data analytics trends and the move to cloud data lakes, what are some of the key challenges, as well as some common data de-identification methods. And then we're going to go through some reference architecture and I'll be going through a live demo of de-identification with Redshift and we'll go from there.

Harold Byun:

So again, if you have any questions, please use the chat panel, you can always email us at info@baffle.io, as well as harold@baffle.io as well. So a little bit about me, I've spent roughly 25 years in security on both the architecture side as well as on the vendor side, predominantly in product management and very much focused on data loss prevention and data containment, and have been at Baffle for roughly four years.

Harold Byun:

So let's jump into things. An Overview of the Data Analytics Trends and the Move to Cloud Data Lakes. I don't think that this is a big surprise statement for anybody, but AI and ML and deriving intelligence out of your data is becoming more and more a big deal for folks. And when we look at the overall cloud data expansion and the footprint and the continued workload moving to cloud, it's estimated, based on a Gartner projection, that 75% of databases and storage will be in cloud by 2022. And a huge portion of that data remains unencrypted as well. And even then, there are variances in terms of how people are implementing some of these controls, which is the purpose of the talk today in terms of how you can

implement a more data centric approach, as well as a simplified approach to manage data privacy or insure it as you migrate and accelerate your move to cloud.

Harold Byun:

Some other statistics. These slides will, again, be made available to you. Roughly a third of organizations are still looking to monetize data and we see that much more evident in terms of an increased focus on secure data sharing, this whole notion of data warehouse and multiple data shares through a Cloud Data Lake as well as data modeling as a service or data as a service, these are all trends that we see continuing to grow. We've also seen a ceiling being hit by organizations that are struggling to manage their on-premise big data environments. A lot of on-premise Hadoop or HDFS environments are flat out run out of room, to be blunt about it, and they're looking to leverage agility and the scalability of a cloud footprint. And then the fifth bullet point, there just continues to be an onslaught of data breaches on a regular basis. It's obviously continuing to get more egregious and in some cases affecting critical infrastructure. And so ensuring confidentiality and security is obviously going to continue to be paramount going forward.

Harold Byun:

When we look at the move to cloud based data lakes, this is commonly a trend that we're seeing related to the increase in transactional environment, the increase in device level data, the increased focus in deriving intelligence and insights out of that data. And this is a common pattern that we see where there's an on-premise environment from multiple data sources that is being moved to the cloud. And so while that's fantastic to leverage the agility of the cloud, oftentimes there are data privacy concerns or security concerns that come about.

Harold Byun:

And so when we look at data privacy challenges, again, we already talked about the continued data breaches, roughly two thirds of CISOs have conveyed that they've been breached through a third party, that relates very much to this notion of data sharing and what does that look like. And then it's been estimated as well that over a billion records have been leaked from cloud storage due to a lack of security controls. And oftentimes, the cloud storage buckets, or blob storage, or object storage is being used as a more economical and scalable method to obviously store your data.

Harold Byun:

This was a survey done by the 451 Group just to express interest in the focus on overall data security and managing those data privacy requirements as high priority initiatives for organizations. And then obviously, there's a renewed focus on GDPR, the dissolution of the safe harbor between the United States and the EU, and the impact that, that has for multinationals as well as CCPA being one of the more prominent data privacy regulations. And ultimately, what this evolves to is a renewed focus on, A, the financial impact of some of these regulations and the penalties associated with that, as well as being able to support a right of data revocation. And in many ways, what we see in terms of the right of data revocation is this notion of being able to support a hold your own key strategy or bring your own key strategy as well.

Harold Byun:

Some additional resources, in addition to the attachments, you can go to baffle.io/privacy, there's an ongoing blog that we maintain that really speaks to data privacy and overall data centric security, as well as trends that are impacting that. Would encourage you to subscribe to that as well. It is a general stream of input and insights and relevant articles that we track across the industry.

Harold Byun:

So when we look at common methods for data identification, there's few things that come into play here. One is the overarching umbrella of this is what is commonly referred to as the shared responsibility model, many of you are probably familiar with this and so I won't spend a ton of time on this. But effectively, the infrastructure and cloud providers or the cloud infrastructure providers are responsible for maintaining that infrastructure and providing you with physical security and configuration capabilities to secure the environment. But you yourself as the customer are responsible for implementing those configuration controls and ensuring that the environment and the infrastructure is secured from both a configuration and vulnerability perspective, as well as being responsible for the data that you put inside the environment. So it's generally how people have delineated the line between the two.

Harold Byun:

And this is an overview of existing controls that are available. There's a number of methods that are available, there's cloud security posture control configuration managers that also automatically detect configuration drift, there's a number of solutions out there that are really focused on ensuring that the environment is locked down appropriately. And it's important, I mean, all of these are available, but the cloud infrastructure is a very fluid and complex environment and it requires a knowledgeable skill set to obviously administer it and configure it in an appropriate fashion, so that's not a small task by any means. And these are just a partial list of some of the controls that are available, things like IAM policy roles, the ability to block public access, enabling the logging. And then there are a number of even encryption controls that are available for cloud storage, and these are significant as well as the in transit data in motion types of protection that you can put in place. But it's important to recognize that the data inside these containers remains in the clear, and that's something that we'll be speaking to shortly later in this webinar.

Harold Byun:

So some common methods for de-identification. Probably the most common ones that we hear referenced to our data masking or tokenization, some people interchange these two. Format preserving encryption or FPE is another mode of tokenization, they're almost interchangeable in terms of the ways that they're being implemented and utilized. Tokenization by and large today can either be vault based or vaultless. Many people are looking at vaultless tokenization or FPE as a means of de-identifying data. In our view, vault based tokenization, which creates a dictionary of data to an unidentifiable token except within this vault, it's a bit of a legacy approach and it lacks a lot of support for cloud native services. And quite frankly, it's a lot slower and won't scale very well. So that's why a lot of people are looking at other modes of data transformation. And then obviously, pure data encryption. There's a move towards more role based data masking and adaptive based security controls given the zero trust nature of the security landscape that we're seeing.

Harold Byun:

And then the last bottom bucket is what we call privacy preserving analytics. This is an emerging portion of the market where it's focused on being able to enable AI and ML and other analytics and reporting on data that is encrypted at all times. And it sounds a bit, for a lot of people who are familiar with encryption, a bit unbelievable, some people commonly used to say, "It sounds a lot like BS." But there are a number of techniques that are available to enable privacy preserving analytics today, and that's definitely an emerging part of the market.

Harold Byun:

So what I did want to talk about is, I guess, some common misconceptions that we still run into around controlling your data in the cloud and protecting it. And a lot of people still rely on what we would call, encryption at rest, or storage encryption, physical disk encryption, or transparent data encryption, commonly known as TDE. And these are all methods that have been available for a number of years obviously in the cloud infrastructure environments that are widely available via chat box that you check when you're provisioning your environments. And it is a good security control to have, but the challenge with these types of controls is that they don't really do anything to protect you against or mitigate the risk of an attack over the wire. And what I mean by that is they're really designed for physical disk protection or container based object level encryption.

Harold Byun:

And the problem with these types of methods is that any attacker who is already in the network or any attacker who's moving laterally in the network, and I think given the nature of the attacks that we've seen through SolarWinds, or the Colonial Pipeline compromise, or other massive scale compromise, the 30,000 Microsoft Exchange servers that were compromised, given the breadth of these types of attacks, you have to wisely assume that attackers are already in your network. And if they already are and they're moving laterally, then that means they're coming in over the wire. They're not breaking into a data center and stealing a hard drive, they're coming in over the wire or they're already in the network. And given that, these encryption modes and these data protection modes that you see on the screen will do absolutely nothing to protect you against that type of attack. It means that the attacker will see the data in the clear and be able to easily exfiltrate it.

Harold Byun:

So let's look at it in a little bit more detail. So this is an example of a database that's encrypted with transparent data encryption or TDE. And we're seeing an absolute surge in terms of a focus on more data centric encryption methods, which this is not, because if you look very closely, you can see that the data is in the clear. And again, we call this TDE method a protection against the Tom Cruise attack. The Tom Cruise attack is the Mission Impossible attack where somebody crawls through the HVAC ducts and drops in from the ceiling to steal data disk drives or data from the data center itself. And then again, that's not how people are hacking data today, right? They're coming in over the wire, they're moving laterally, they're gaining entrance, and then they're exfiltrating data. Just cases in point, Marriott, Capital One, many, many others were using transparent data encryption. It really does very, very little to mitigate the risk of a data breach or data leak in today's world.

Harold Byun:

So this is a close up view of data centric encryption versus object or container based encryption. The below is Cleartext data in a container, and you can see that the data is actually not encrypted, it's encrypted as a container, and anybody who gains access can see the data in the clear. The data in the above is data-centric encryption, which will actually transform the data values in the repository and only authorized users can actually re-identify the data or selectively identify the data. And so that's one of the fundamental differences between these two main methods.

Harold Byun:

And so you can see things like the birthdate is transformed into a date that will pass a validation check but it is at least protected from a de-identification standpoint. The credit card numbers, for example, will pass a credit card Luhn validation check. Social security numbers retained format. The email you can barely make out an @ sign and a period in the encrypted version of this, those are things that will retain format. And so these are alternate methods of being able to de-identify the data.

Harold Byun:

So when we look at different data protection capabilities, one of the things that we find is that people are looking at simple data transforms which are things like, I just want to take this dataset, first name, last name, take my name Harold Byun, for example, and transform it to ABCDE and one, two, three, four, five from the last name, and I just want a static mask that does nothing. There's other modes that are more truly using a de-identification technique or other types of adding fuzziness to the dataset to further obscure it but still allow for analytics. There's people that are trying to do things in a multi-tenant perspective. There's the ability to dynamically mask this and present different views of data based on people accessing the data. There's the ability to tie that to a role and a zero trust model so that we can better control which users, given their role and given their context, can see what data. And then in the upper right hand corner, there's this notion of privacy preserving analytics, which is really akin to secure computation and homomorphic computational capabilities.

Harold Byun:

So all of these are the different gamut that we see based on the types of analytics, the types of intelligence that people want to derive from their data, as well as the overall security posture that they want to maintain. So some of the key benefits when you look at implementing these data privacy controls are being able to de-identify the data values themselves, which will give you safe harbor from data leaks, which will also protect you from some of the penalties that are enforced around some of the compliance regulations. And then also being able to unblock your business. Really, the nature of implementing these privacy controls, I mean, it's important from a security posture perspective and always to have security front and center.

Harold Byun:

But the reality is, and I think all of us who have worked in security for quite some time have worked with different folks on the business side, I'm trying to figure out the best ways to say this, different folks on the business side who may not have security front and center all the time, and so in many ways you can use this to help your security organization move in lockstep with the business, not have them pull the end around move or the bypass on security, and really help you move at Cloud DevOps speed to move in conjunction with the move to cloud while enabling data privacy on the fly.

Harold Byun:

So what are some different reference architectures or models that we're seeing from a modern data pipeline and de-identification perspective? So this is a classic pattern that we see often moving from left to right, and that is having a data store on-premise or an existing data lake on-premise, or some transactional mainframe that is moving to cloud, very much a common pattern we see as it relates to Redshift and Snowflake. And what we're able to do is basically function in line as an invisible data protection layer and de-identify data on the fly.

Harold Byun:

This is fundamentally different than a lot of other de-identification techniques that require you to create clones of the data, create copies of columns and transform them through some out of band manual transform process. The methods that we talk about are truly a point and shoot to cloud motion. And so that is fundamentally different where you basically take an on-premise data store or it could be an existing cloud data store that you're moving to an analytics data lake, and we can de-identify that on the fly and then we can selectively re-identify it as well for authorized users.

Harold Byun:

And so this is a different reference architecture that's more traditional reference architecture for AWS, where you'll see that we can move using an integration with AWS DMS, we act as a transparent endpoint, we de-identify the data on the fly, and then we ingest it into Amazon Redshift. The data flow on the bottom is a different reference architecture. The private link is just a reference, it doesn't necessarily need to be a private link, but there are organizations using both. You can also use a VPN. But the bigger point is that we can perform an on the fly transform with Kafka where we can take data streaming techniques, de-identify the data on the fly as it's being streamed into a data lake or object storage using a Kafka data stream, and then re-ingest that into Redshift.

Harold Byun:

A more legacy type of flow is this where we can use an extract from an on-premise database and a flat file or some other format, and either SFTP, or FTP, or using an AWS S3 API or AWS CLI, we basically, again, transparently de-identify that data, land it in a de-identified state. Again, no clones, no additional operational overhead, no operational changes, you just point and shoot and it lands.

Harold Byun:

This is a joint reference architecture that we've worked on with the AWS consumer products goods group, and it is available online, I think there's a link in the attachments as well. But basically, again, on the left hand side shows you that de-identification footprint, the ability to consume that into an agile cloud environment that allows you to perform those analytics and then selectively re-identify that data for authorized users. So being able to maintain that data privacy within the cloud context, per se, but still facilitate the analytics and the intelligence derivation that the business requires.

Harold Byun:

So let me switch gears here. I'm going to walk you through what this might look like from a data flow perspective. Going back to what we're going to do is I'm going to simulate taking data from an

on-premise database, migrating it through a data structure, and landing it into a cloud footprint. The first half of this isn't very exciting. And quite frankly, somebody took down one of my endpoints, so I'm going to have to rely on video for that, but then the setup and the rendering and consumption into Redshift is something that I will show you live. So I apologize for not having the full live and demo, pains me as well, but at least you'll get a sense of what this looks like.

Harold Byun:

So I'm going to just step through this. And in this you'll see this is just an S3 bucket. I know it says Snowflake, we can do the same thing for Snowflake, but this is just more for a reference point. You'll see that this bucket is empty. And so what we're going to reference in this is basically, I just skipped through it, you'll see that this is just showing you that this data structure is empty, but this isn't really the important part of the demo. What we do here is we kick off a job and we start a DMS task to pull from the on-premise database and move it to a target data store. And you'll see here that this is the on-premise data structure where there's Cleartext data.

Harold Byun:

So you can see there's clear text data, and what I'm going to do is I'm going to skip ahead and show you that we're moving through the same pipeline that I'm referencing here, and we're going to land it in the bucket. And so this is running. And this data from that empty bucket will basically be rendered in the clear. And so you'll see that when I switch to this S3 bucket, skip ahead here and refresh, we now have data that's in here in this flat file. So that's the data point and shoot migration task. Just to play this out so you can see it, I'm going to open this file, and you'll see it in a de-identified state. And so this is opening it into Excel, and you can see that this column, as well as this column have been in a de-identified state on the fly, basically point and shoot to cloud.

Harold Byun:

So let me shift gears to the live demo and what this actually looks like. So in our environment I have this Redshift bucket, and in here I have this CSV that has been migrated, and I'm going to open this. And so if I open this file, we'll see that the state has been put into a de-identified state. So you can see that customer ID and customer name in particular are de-identified here using this tokenization method. And there's obviously some dupes here, which is why you'll see some of this using tokens.

Harold Byun:

Now, what I'm going to do is I'm going to walk you through the setup of this and what this actually looks like. So if we stage this into a Redshift environment, and so what I'll do here ... I think I probably just messed up on this data share. So let me just walk you through this again. Apologize for that. Thank you for cooling me into that. So apologize.

Harold Byun:

Basically, we can show you this really quickly, but there's a de-identification here in this empty bucket and we are moving this through Amazon DMS, just what's going on here. And I'm going to skip ahead and give you the super short version. This is the source database where the data is in the clear. And then we're moving through the pipeline as I've been showing you. And we land this data into this S3 bucket.

As you can see, it's now full, and when I open this file, which I'll do here, and skip ahead, we can see that this data is de-identified. So this is the de-identification, and that's what I just showed.

Harold Byun:

So now let me go back to where I just was. So I'm in this bucket, and in this Redshift bucket we've got a data set here. And what I'm going to do is I'm going to open this dataset. And when I open this dataset, you'll see that this is also de-identified. So this is a de-identified data set, customer ID. And now what I'm going to do is I'm going to go over to Redshift. And so what you do is you would copy this into Redshift into a table. So I'm going to just revalidate this. And you'll see that if I try and connect here right now and go to this and run, we have the same de-identified data set. This is a direct connection to a Redshift cluster. And you'll see that this data set is the same that we were just showing you out of the CSV extract.

Harold Byun:

Now what I need to do is I'm going to login to our Baffle Manager product. And this is where we can configure data protection policies. In data protection policies you'll see that what I have set up is basically Baffle shields which are our transparent data protection layer. I have an integration to AWS KMS, which we can integrate into virtually any Keystore HashiCorp Vault, Thales, Gemalto, any HSM. And then what I'm going to do is I'm going to establish a connection to Redshift and setup this data protection policy. So this is all available in the marketplace as well, so you can download this, but I will set this up for you. So let me pick Redshift, and then I have my endpoint here. So I've added the database and now I'm going to set up a data protection policy for this. I'll select Baffle shield, I'll select the Redshift data store, I select AWS KMS. And I've established this baseline profile. The data has already been de-identified, and that's one of the unique things about Baffle, is we can de-identify the data on one side and we can re-identify it selectively on the other side.

Harold Byun:

So you've already seen the date. Anybody who's coming into Redshift is going to see this, right. And that may be useful for you, depending on who you want accessing the data, it may not be. There may be other scenarios where you need to re-identify the data. And so the tricky part is, how do you seamlessly now re-identify this data? Well, the way we do that is we configure this application profile. And I know that we have customer ID and customer name, and we want to use a token method. So we have different formats that we can support and it's very flexible, I can create additional formats if I want. But I know that both of these are alphanumeric, so I'm just going to use that. And then basically deploy the [Spozy 00:32:21] and save it for later review, I can deploy it. We can migrate, we have our own migration utility as well, but that's not really applicable in the modern data pipeline.

Harold Byun:

And so you see I've deployed this. And so now, I have a scenario where we have a Baffle shield that's running. And so if you go look at the slides that I was showing you, again, we've taken this data, we've migrate it, we've de-identified it and landed it in an S3 bucket, and now we're trying to re-identify it in Redshift. And so in order to do that, what I've done here is I have this shield, and this shield here is pointing to a connection string that is different than the Redshift cluster on a different port. And this is a Baffle shield that can run anywhere. And so I'm going to connect to this.

Harold Byun:

Just to show you that we're dealing with the same table structure, this is what's on the Redshift direct, this is using the Baffle shield. Let's see if I can run this. So we now have the same tablespace. And now when I run the query, you'll see that we re-identified the data into its original form. So that's a seamless encrypt, decrypt there. And so you'll see that there. Some of these were encrypted with a different key set, but this is the seamless encrypt, decrypt.

Harold Byun:

Now, what we can do as well is run this here. And then if I wanted to, I can alter the State of Protection Policy and I can apply masking. And so now we can also selectively control who can actually see what data. So if I go into this dataset and say I want to mask this product ID, I can mask this and I can use a different format. So maybe for the product ID I want to show the last four, and maybe for something like city, which is not really that sensitive, but just a demo, I may want to mark this confidential. And so now I have two fields that I'm de-identifying and other fields that I'm selectively masking. And when we deploy this again, give it a moment, and I refresh the query window, we should see mass data for confidential and city, and on the product ID we're only exposing the last four digits. So that was the last four policy.

Harold Byun:

And so this gives you the ability to flexibly present different views to different users or different subscribers based on how you want to present your data. And that is really what gives you this more end to end access control model. I mean, first and foremost, simplifying the move to cloud where you can move securely and ensure data privacy. And then secondarily, re-identifying the data to make it useful for the business and being able to do that in a seamless fashion. And that's what is different from our solution standpoint.

Harold Byun:

This is the overall data protection service architecture that we have. The predominant component that we've been working with is this Baffle shield, which is that invisible bump in the wire that effectively is proxying any API, or driver traffic, or application level traffic to Redshift. And then ultimately, the way we look at this solution is as a comprehensive data transformation and access control solution. So we can de-identify data from any source to any destination while also controlling access, and we can do this at a massive scale. We're in line in production with 30 of the largest financial institutions globally through a SaaS provider, we have a couple Fortune 25 customers that are running us at scale in environments that are in excess of billions of records. So the ability for us to do this, again, in invisible manner and the ability for us to support cloud native services without additional operational steps is something that we can easily do.

Harold Byun:

I think we're running in towards the tail end of this, so I'm going to skip over this privacy preserving analytics capability and open it up for questions. So in short, you can leverage cloud data lakes to improve your agility and flexibility, the data-centric protection methods are going to help ensure data privacy and reduce your risk, and then ultimately, they're going to help your security organizations to move in lockstep with the business. And these are methods that don't impede a DevOps or a shift left

type of operating model which a lot of organizations are moving to. So that's the quick rundown on what we can do. And again, there's more resources available for you if you'd like. And feel free to reach out to us anytime if you have additional questions. I'm going to respond to some other questions that have come across here and then we'll wrap things up. So appreciate your time today.

Harold Byun:

So can Baffle run queries on a user's behalf for an encrypted field, for example, to enable secure computation so the user cannot see the Cleartext but the query can execute? Yes, we absolutely can. So the ability for us to do secure computation is a feature of our privacy preserving analytics capability which utilizes a technique called, Secure multi-party computation. It is very, very useful, especially for multi-party scenarios where you want a common pool of data but you don't want any users to see anybody else's data. So that is something that we can definitely facilitate. There's a number of resources on our website as well and if you're interested, we're happy to go into more detail on that with you.

Harold Byun:

Next question is, what is the support for master keys or customer managed keys? So the implementation that we have for encryption is commonly referred to as envelope encryption or a two tier hierarchy or CMK, where we use the CMK or the master key, or what some folks refer to as the key encrypting key, as the key that encrypts a data encryption key. And so the data encryption key is really what is used to perform the encrypt-decrypt function, and then the master key is what actually performs the encryption of the DEK. So as long as the master key is secured in an HSM or a store like AWS KMS, then we use that to decrypt that data encryption key and we can support that across, again, any HSM or secrets manager, as well as AWS KMS and Amazon Cloud HSM, which includes, I think, the Luna SafeNet and the other variant which I think is now DSM.

Harold Byun:

And then last question here, what algorithms are you using for format preserving encryption? We support both FF1 and FF3-1. If you're asking that question, you're probably well aware that FF3 has been deprecated, both are NIST standards. I guess FF3-1 is still officially in NIST draft review, but for all intents and purposes, it will suffice as a NIST standard. We default to FF1 anyway, which is the NIST standard for format preserving encryption. Format preserving encryption, as you know, preserves length and preserves the data type.

Harold Byun:

So I think that, that's it for the questions unless there's any last minute ones. If you have any other questions, again, feel free to reach out at info@baffle.io or my email is harold@baffle.io. And would like to, again, thank you for your attendance and time today. And hopefully this was useful. Thanks. Bye.