Harold Byun:

Hi, and welcome to today's webinar on De-Identifying and Securing the Data Analytics Pipeline. We are going to just wait another minute or two as folks get into the webinar session, and we'll probably run about, ideally, 40-45 minutes today, including the Q&A at the end. As we're waiting for folks, just a matter of logistics, you can feel free to use the chat window to ask any questions as things progress. So, bear with us for just a moment. Thank you.

Harold Byun:

(silence)

Harold Byun:

Okay, and once again, thank you and welcome to today's webinar. I will go ahead and kick things off, and I appreciate you joining today. My name's Harold Byun, I am the head of products for a company called Baffle, and today we're going to be talking about some de-identification practices as they relate to some data analytics and cloud data lakes that we're seeing customers adopt, as well as how they're looking to address some key data privacy challenges.

Harold Byun:

So, the agenda is obviously, quick review of some of the trends that we're seeing, the migration to cloud data lakes, an overview of some of the data privacy challenges that we're seeing emerge, and we'll follow that with some coverage for common methods for de-identification, some different architectural models we're seeing within the data analytics space. We'll walk you through a live demo of a migration on the fly while de-identifying data and re-identifying data in selection areas as well. And then time permitting, generally try to keep these on the shorter side, we'll go into a glimpse into some privacy preserving and advanced analytics modes. Again, I encourage you to use the chat panel, my email is at the bottom right and feel free to also email us with any questions.

Harold Byun:

In the attachments and Links section of this, there's some white papers that have been written that are relevant to the industry. And then there's some different links that will display throughout the webinar, where you can get some additional information if you're interested. So without further ado, let me jump into things. Just a little background on myself, I've been in the security industry for close to three decades at this point, really focused on a lot of data loss prevention and data containment has been the bulk of my career. And I've been working with Baffle for roughly around three years now, primarily focused on data centric protection, so just a little background on me.

Harold Byun:

And so, in terms of the overall agenda, again, covering an overview on some of the data analytics trends, obviously a statement of the obvious here that AI and big data are a big thing for business. And I don't think that that's a surprise for anybody, but the underlying trend around that is just this massive data growth that we're seeing contrasting with some of the data privacy regulations that have been coming into play.

Harold Byun:

And so just some general industry statistics and trends that we've been monitoring or covering, the first two are from Gartner and really, the first one being that within a few years, roughly three quarters of organizations are going to really focus on operationalizing AI. These are all kinds of things to predictive customer service and attention to fraud analytics, to some other security-based scenarios, and then also evolving that in bullet point number two, into what we're calling data as a service or data modeling as a service, or the reselling and sharing of information.

Harold Byun:

And there's some really interesting use cases both on the resell side, as well as just the aggregate analytics side of things, where the space is becoming incredibly interesting, especially even in the healthcare industry, as many of you are probably aware with the onset of COVID, there's a need to share information more rapidly across multiple providers or nation states, while still retaining data privacy for potential individuals who may have their data, or who may be a portion of that data set.

Harold Byun:

Our bullet point number three is just a general trend that we're seeing across a number of customers where many enterprises have actually already built out a big data Hadoop infrastructure on premise or in a colo obviously, and they've been managing that infrastructure for a number of years already. The challenge is that the data growth continue exponentially, and the ability for the existing big data environments to expand flexibly and adapt to that data growth, becoming more and more problematic. And effectively, people are running out of room, and so we're seeing a migration to this notion of a cloud data lake. And so that's something that we're going to be talking about in more detail today. That's kind of relevant to bullet point number four as I guess, an underlying reason for why we're seeing this move to the cloud data lake infrastructure.

Harold Byun:

And then bullet number five is something that I think many of us are all familiar with, which is kind of continued inadvertent exposure of data or misconfiguration of data in some of these cloud-based environments from a storage perspective. If you look at this from a spectrum perspective, or from a perspective of across the spectrum, on the one hand, you have the business driver of, how do we drive AI and business analytics to derive intelligence? The technical infrastructure challenge of how do you manage that data footprint? And then, thirdly and lastly or perhaps not lastly depending on your perspective, the security and the data privacy around that data, really represents a push and pull tension between the business and data privacy folks.

Harold Byun:

From a look data lakes, this is just I guess, detecting some of that trend, we see a lot of organizations who have built out that on premise infrastructure. The data set is increasingly more distributed in terms of the collection sources, and the number of multi-channel data sources that people are pulling that data from, and what we're seeing, typically, and this is an example from AWS, where they're moving to, I guess, a more cost-effective and flexible cloud-based data lake using S3 as the new data source or data store for that information which provides you with in theory, an infinite amount of space that can grow with your business and your data set.

Harold Byun:

Obviously, challenges around how to continually manage that with appropriate policies, and there's a lot of tooling and configuration solutions around that focus on how to better optimize and manage and secure that information. We're not going to be talking about necessarily infrastructure controls today, I will touch on it, but I think there's a distinct difference between data-centric protection and infrastructure controls that we'll go into. And what we're really going to be focusing on is again, de-identifying the underlying data or the data inside these cloud data lakes, while still permitting the business to run analytics and AI modeling on top of that.

Harold Byun:

There's a similar version from Azure, obviously, GCP has got a similar model as well, but as your data lake services, and as your blog stores represent the equivalent on the Azure side of things, we're really a multi-cloud agnostic provider that just kind of showing the trend that we're seeing across multiple organizations looking to adopt, again, these cloud data lakes.

Harold Byun:

When we look at some of the data privacy challenges that continue to emerge, I mean, obviously data breaches remain front and center. And alongside of that, with more distributed data footprints, there's this notion of third party risk as data is shared across multiple organizations, whether they're part of your operations or not, they may be an external party. And in one of the more recent surveys from the Ponemon Institute, roughly 60% of CSEs have reported some type of data leakage via third party.

Harold Byun:

I was recently working with another customer and the Infosec team or actually, stated that eight out of 10 of the more recent incidents were due to data leakage events via third party, so it's obviously a pretty significant problem for a number of organizations. And then, in terms of overall data exposure, a lot of it is due to misconfiguration, there is this whole notion of the shared responsibility model that we'll be touching on. But when you look at cloud storage data leaks, obviously, significant number of records have been leaked, and roughly more than 10% of data storage environments are left open to the public, and so a lot of that is due to general knowledge and best practices and how organizations operationalize security in those environments.

Harold Byun:

This is from a 451 survey in terms of some of the data, or a question on data management analytics and challenges. So, they asked what are the biggest data management challenges, and if you look at the first five, roughly one in five to one in three respondents were really focused on data security, or data privacy and compliance in some measure. So obviously, remains a high priority or a high area of focus for a lot of organizations as they move to the cloud, and they're looking to wrap their arms around data analytics and balance that with privacy.

Harold Byun:

I'm not going to spend a ton of time on CCPA and GDPR, I think that there's been a lot of, I guess attention put on these regulations, and there's emerging privacy regulations, and now another 20 to 25 states that's ongoing and pending at this point in time. CCPA in California was not delayed due to COVID, it's one of the few things that wasn't delayed due to COVID, at least the enforcement vehicle, and that translates into this headline, I think, is from a couple weeks ago, or maybe last week, where there was a

lawsuit actually enacted under CCPA against Walmart, so enforcement is beginning and that's a whole 16 days after the enforcement deadline has passed.

Harold Byun:

So, in terms of the financial implications, or potential revenue impact, there's definitely potential impact and risk to your business as it relates to some of the regulations that are out there. For those of you who are interested in more information, you can go to a baffle.io/ privacy, where we have a page set up with additional papers that are available. Again, there are also some papers and links that are available in this webinar, but you can gather more information here as well.

Harold Byun:

So, as it relates to privacy, obviously, a lot of challenges there. When you put this into a ... I guess when we take a step back, I think the reality that many of us are faced with is that the data is going to be collected anyway, and there is value to deriving intelligence out of that information, there is value in collecting that information. There are a lot of people asking questions around, why do you need so much information and what can you collect? But the reality is that data collection is not going to be going away, nor is AI or deriving insights from the data as an initiative. And so really, what can you do about that going forward?

Harold Byun:

And so well, this is an interesting slide that actually came from a Gartner paper, affinity with Gartner clients, I would encourage you to read this document, it's called Securing the Data and Advanced Analytics Pipeline. And the premise on the left is very interesting, it's effectively saying that when you look at aggregating data, that there is a certain subset of data in the dark-blue that nobody can really have access to on the clear, there's a subset that's dubbed sensitive data that depending on privilege, or entitlements, some people may have access to.

Harold Byun:

And then, there's this notion of downstream data or data that can be used for downstream analytics. And the idea on the left is that in a base case scenario, most organizations are left with this tiny box in the upper right-hand corner from a balance of security and privacy regulation standpoint that you're really allowed to use. Whereas, your project initiatives may really place you in a place where this orange dotted box is on the left hand-side. And without the appropriate controls, the promise is that in theory, this should be off limits. I think that there are a lot of organizations that take liberty with that, but in theory, this orange box should be off limits.

Harold Byun:

And then idea is that by evolving some of your data privacy strategies going forward, that you can open up a larger pool of data, and within the big data realm, more data is always a good thing, contrary to what a lot of security folks may feel. But more data is a good thing and more access that you can leverage downstream of the data is more important, or is more valuable. And so, if you can put in appropriate controls that let you leverage that information in an increased fashion, it can unlock some of the business analytics that your organization can take advantage of which should translate into competitive advantage.

Harold Byun:

So, there's a question here, where does biometric data fit in this chart, especially given GDPR special consideration for biometric data? And that's a great question, I mean, biometric data, both under GDPR and CCPA are considered personally identifiable information that should be retroactively... That should either in the CCPA case, not be collected in an identifiable format, the GDPR constraint, really within this, it would represent an off the dark data box, I guess the answer to your question. It shouldn't be something that is visible in any form, it needs to be secured, and there's a lot of ... I think there's been a number of biometric breaches in recent times. And obviously, one of the huge challenges around that are that your biometrics are yours, and they can't be changed. It's not like a password, it's not like an account number or a credit card number, it is something that it can be permanently leveraged.

Harold Byun:

And I think that represents a significant risk, but in many ways, this is a set of data that needs to be protected in a different manner. And quite frankly, it shouldn't necessarily be analyzed in a non de-identified state. And so there are a lot of initiatives where we are working with some sciences firms, and some hospital networks that it's not biometric data, but it is health-related data that has obviously, medical and health characteristics where they're performing modeling and aggregation in a de-identified manner.

Harold Byun:

There are also more advanced privacy preserving techniques, where some of this biometric data can be held in an encrypted state, both at rest and in memory and in process. And so, there's a lot of emerging technology in terms of how to better protect the biometric data, and within the constraints of that, while it still remains dark data, there are ways to derive intelligence and insights out of that data without violating certain privacy regulations like GDPR, or CCPA. And time permitting, again, we'll touch on that, towards the end of this in terms of some of the capabilities that are out there. So, I hope that answers some of your question, and I'm happy to discuss that in more detail if there's interest.

Harold Byun:

So, we're going to step into some of the methods for de-identification, and so as it relates to this notion of de-identifying data, I didn't want to draw a distinction between infrastructure versus data controls. And so many of you are probably familiar with this, it's the view of AWS's shared responsibility model, and the shared responsibility model basically says AWS is responsible for providing security of the underlying infrastructure and the data centers, and what you need to actually employ controls around the infrastructure and managing that, so the instances, the network stack, the data centers are secured by AWS.

Harold Byun:

But what you put in the cloud in terms of data is really the customer responsibility, so that would be your responsibility, and so there's an important distinction there, and the reason I raised that is because there are a lot of infrastructure controls that these providers have made available to you all its customers, to all of us. And so this is a litany, I'm not going to cover it at length, but there are a number of controls that help you secure infrastructure within these environments, and let you operate in a more controlled fashion. What we're really talking about is de-identifying the actual data values or protecting

the underlying data values that are being put in, so this is an actual scenario where, I want to actually highlight and differentiate between the infrastructure versus data controls.

Harold Byun:

So, common methods for de-identification, so these are things that we're seeing fairly commonly across the industry, there's obviously, data-centric encryption, or data encryption which falls into the bucket of field-level encryption, record-level encryption, or actually protecting specific data values with some forms of encryption. Within that there are some variations, there's this notion of data tokenization, which made use a tokenization library, or a tokenization, or what is known as a boltless tokenization, or some type of codebook method. And so, those are methods of replacing an account number with another value, for example, or associated number with another value.

Harold Byun:

Format-preserving encryption is another variant of this where it is a data type and length-preserving encryption mode. So again, things like a social security number, or a medical record number, or a credit card number, or email address, those are all things that would then be actually preserved from a formatting perspective, but randomized in terms of the actual underlying data values.

Harold Byun:

Data masking is really used for controlling the presentation layer, so instead of seeing somebody's full credit card number, you might only see the last four digits, or values can be completely obfuscated. A variant of that that we're seeing is also role-based data masking, so role-based data masking is really in the form of data entitlements, where we're seeing organizations based on role, so certain departments, certain roles within an organization, or certain parties that are in a more privileged state or third parties that may be external, but basically get presented with different views of the data that is dynamically rendered based on your group membership.

Harold Byun:

And then, related back to some of the biometrics question and what is dark data, and where should it reside? There are some Advanced Encryption schemes, so there are primarily I would say, three major types of Advanced Encryption modes that we're seeing emerge. One is enclave-based technologies, which is a hardware-based method, promoted by folks like Intel or AMD. I think IBM has an initiative, AWS has their own initiative around this. Google Cloud has one as well as Microsoft within the realm of confidential compute. And so, these are enclave-based technologies that allow you to secure data very much like the fingerprint or face ID that you may have on your phone. But just extending that to a server class compute model.

Harold Byun:

There is homomorphic encryption, which has been largely theoretical for several decades to this point. I think that there's still a lot of runway to go on homomorphic encryption, but there are some providers that are starting to attempt to market around that. And then there's multi-party computation, which also allows for computation of data using a shared secret model that never allows certain dark data to be decrypted, so those are kind of advanced modes.

Harold Byun:

Just another question here that's come in, isn't most healthcare data now maintained numerically, a medical record number, rather than a name? That is probably largely true, and there's plenty of scenarios where it isn't, there's plenty of scenarios where even that healthcare data is still attributed to a given individual and more than that, there's also the ability to re-identify in a generalized form, who that individual is based on certain characteristics. And so, even in that sense, even if you are just a medical record number, there are challenges around that, and even if you are only a medical record number, at some point it is connected to you in some system somewhere. I mean, there's a join that's happening between your medical record number and the online health portal that you're using, as well as any type of receipts or invoices that are being exchanged with health insurance providers, so it isn't completely an opaque scenario.

Harold Byun:

This is just a visual view of ... This is trying to differentiate between object level encryption and data-centric encryption. Object level encryption is containerizing a given file or container of a data source, like a bucket or a storage environment, versus actually going inside and de-identifying the data. And so, in the top half, you'll see a method of format-preserving encryption, where the name and the email, for example, or retaining certain formats or length. In the right-hand side, you can see like the domain is preserved in that particular case, but certain strings or obfuscated or de-identified using this encryption method, where's the bottom half if you're using any type of object or container-based encryption, when somebody looks at the data, they get it in the clear, there is no de-identification going on whatsoever.

Harold Byun:

So key benefits, obviously, looking to add methods to de-identify or ensure privacy inside objects and files as you adopt these cloud data lakes. This grants you safe harbor from accidental data lakes or misconfiguration, and obviously, can unblock cloud-based data analytics for a lot of organizations where there is a tension between your chief data officer, or chief privacy officer, and the security teams around how to best manage that move to cloud.

Harold Byun:

So, we're going to move into some architectural models, and I'll quickly demo some of this data migration. Time permitting, again, we'll get into somebody's advanced analytics. And, again, keep it coming with the questions as we go forward.

Harold Byun:

So, this is a view of one basic cloud data lake using S3 and the downstream analytics consumption sources that may consume that data. And so obviously, concerns around the de-identification of that data and ensuring data privacy throughout.

Harold Byun:

Another view of this is the overall data pipeline, again, using S3 buckets as the foundation, going through Athena for ad-hoc queries, and using different modeling type solutions, where you can actually

suck that data into a model scenario and perform different types of ML or AI based-modeling and training sets.

Harold Byun:

This is a view of what is our view, of what you might do to de-identify a pipeline. Obviously, one of the icons here is in this diamond shape is called a Baffle Shield that is basically our data privacy encryption, decryption method. And so, we can interface with things like database migration services for AWS, or Azure, or third party migration services, where basically, [inaudible 00:28:35], seamlessly take data from on-premise, de-identify it on the fly into an S3 bucket. And what has been a more recent announcement for us is the ability to selectively re-identify that data as it's being accessed via Athena, and things like Snowflake, or downstream sources.

Harold Byun:

So, it gives you the ability again, to take a data set as part of your data pipeline, push it to cloud on a continual basis, de-identify it on the fly, and then allow members of your organization to run the types of analytics that they would need to run on that respective data set.

Harold Byun:

Let me transition here into the demo, so let me share the screen. It might just take a moment to come up, and again, I encourage you to ask any questions as we go along. So, you should be able to see the larger presentation now, so what I'm going to do here is go into this demo set. And so, what I'm going to do is kick off this database migration task, and so while I'm doing that, this takes like a minute to spin up, so I'm just going to kick it off.

Harold Byun:

And so, what we're doing here is the left-hand of the slide that I was just showing you, so taking an on-premise data source, which is represented by the SQL Server, and we're going to migrate it to S3, and de-identify it on the fly. And so, I'm going to kick it off and show you that this environment is empty. So I have this S3 bucket, you'll see it's empty, I'm hitting refresh, there's nothing in it. And this is starting, and I'm going to go to the data source for this as well. And so, the data source that I'm using, is this on-premise SQL Server, and in this SQL Server, I'll pull up this data set, you'll see that it's in the clear, and it's potentially sensitive data; customer names, and their purchase history, and the like.

Harold Byun:

And so, what's going on right now is we are running this database migration services job, it is hitting that Baffle Shield. We have privacy policies in place that are going to de-identify some of the data set, and then it's going to hit the S3 bucket in an encrypted in de-identified state. And so, that would represent the first half of your data pipeline where we've ... Basically, and you'll see, "Load complete now." So, if I go back here and refresh, we now have a folder structure that is using that table space that I was just referencing in the on-premise SQL Server. And so, I'm going to download this just to show you that this particular file is in an encrypted or de-identified state, and so you'll see that there are certain fields like the customer ID, or the customer name, that have been put into a de-identified state and I think we shall see. I'm going to blow this up a little so you can see it better.

Harold Byun:

And so, the other fields are obviously... You can do this on a field by field basis, but it allows us to de-identify the data on the fly as it's flowing from an on-premise to your SP environment on cloud data lake. And so, if I go back into this particular environment, one of the things that we can see if I open up the screen, I think by establishing this with a glue catalog you can actually expose some of the data set for de-identified analytics.

Harold Byun:

And so in this particular case, I think I have this set up already, and if I look at some of this data, we can see that we're able to pull it straight out of the S3 bucket, and so if I wanted to run aggregate analytics on a given customer and pivot on some of that data, this would allow us to perform that type of de-identified function. Pretty rudimentary in terms of this obviously, and then if I wanted to search on a given customer name in a tokenized form, there's a way for us to obviously drive a pivot on a common customer token that's been created in this particular instance.

Harold Byun:

And so, that's one scenario around this, the other is potentially around, how do you actually re-identify the data? So, what we've just shown you is the left-hand portion of this, which is this notion of de-identifying as it relates to S3, but there are scenarios where people do need to selectively re-identify data, and it's been an Achilles heel for a number of providers, because there's no seemingly easy way to do this without performing some application modification, or recoding something specific to re-identify them, there's all kinds of challenges around that.

Harold Byun:

And so, one of the things that we're able to enable, and this is another S3 bucket that I'm going to use for this demo. And so, you'll see again, it's empty, and also I'll use this web analytics tool called cross filter to pivot on data. And you'll see that this is also empty, as well as this direct environment, which is going to show data in an encrypted state.

Harold Byun:

So, I have this empty S3 bucket, and what I'm going to do is I'm actually going to use the S3 APIs to pivot on that data to put that data through the cloud, and so you'll see here that what I'm going to do is I'm going to issue basically a put object, I'm going to take this flight file, which has some flight data in it, and I'm going to put it into an S3 bucket, and it's going to be de-identified on the fly again, and so just to give you a quick view of this, this is a clear text file, where it's got some of the origin and destination information and a bunch of records and some statistics on flight data.

Harold Byun:

And so, what I'm going to do here is I'm going to put that object into that S3 bucket, it's actually going to hit that Baffle Shield, it's going to be de-identified on the fly. And so, if we go back to the S3 bucket now and hit refresh, we should see this flight file or same file here. And if I download that locally this should be in an encrypted state. And so, if I show this in the Finder and open it into Excel. [inaudible 00:36:14].

Harold Byun:

Here it is. So, this file is in ... We've encrypted the destination in this particular example, and so the challenge with this type of de-identification for a lot of industry has been managing in select forms the re-identification and the analytics portion of this on a subset of data. And so, what we're able to do in this particular model is ... If I refresh this particular page, you'll see that this is pulling straight from the S3 bucket, and the information is rendering in a decrypted state and allows an organization to look at some of that data set, whereas by circumvent, the Baffle Shield and also run this refresh, just to show the underlying data is actually encrypted in the destinations field.

Harold Byun:

And so, that was one form where again, typically, you would have to modify an application and embed some type of SDK, and figure out a key exchange, or a method for re-identification. We've simplified some of that process, and then as it relates to Athena in particular, is this notion of ... In this case, how can we actually just access data casually. And so, what we're doing here is using the Amazon Athena driver, and if I go into this flight file, you can see that, again, we're able to interface with the Athena API set and render data living in a cloud data lake and so I'm able to pivot on the state. Obviously, it was a pretty rudimentary use case, but you get the point where we've identified some of the data and are able to render that and selectively re-identify based on the analytics use case.

Harold Byun:

So that's kind of an end-to-end view of what we're enabling here, which is de-identifying here and selectively re-identifying from a downstream analytics source here. So, query gives us some sense on that stuff, a few more minutes here, keep on coming with the questions, I'll be wrapping up with those. I go to the privacy preserving analytics, so really, what does that mean, once you've de-identified this dataset? How can we actually make it even more usable, longer term? Or what if I need to ... In the case of biometrics was a great example, so what if I'm dealing with dark data or things that are completely verboten in terms of people being able to see that? Or what if there are multi-party scenarios where I'm aggregating information from multiple hospitals, say in the case of a pandemic, or a disease?

Harold Byun:

And so, privacy preserving analytics really allows for secure computation, it's been the holy grail within the data analytics and privacy space, and so I covered some of the methods that people are using for enabling that. And so, there are a number of scenarios that we're starting to see evolve around that, one of them is really around third party data access control, or multi-party data sharing, where people want to run these aggregate analytics and so whether or not the medical record number is associated with a name or not, it's your medical record number as your organization, and you might be a hospital that is sharing with five other hospitals, or you might be a hospital that is sharing with three other insurance providers.

Harold Byun:

And so, that may be your medical record number, and that's fine, but there should be in some cases no reason for a downstream provider or somebody who's operating on your behalf as an operator to see that type of data. And so, how do you actually facilitate secure data sharing in that type of scenario? And so, there are a couple different examples around this for privacy preserving analytics, one of these is healthcare data sharing, obviously, it's seemingly top of mind these days.

Harold Byun:

And so, if you imagine the scenario, where you have multiple organizations that were going to share information, this is an example of two organizations, they could be multiple organizations. But let's say organization one is submitting data for a certain patient that has a condition and the condition value is, yes, and we want to de-identify both that, we don't want anybody to know, is John positive or negative and let alone his medical record number or other information. So, we encrypt that data as A, B, C, D, E, F for simplification purposes, in the condition as D, E, F, one, two, three, four, five, nine, for example.

Harold Byun:

And the data lives in this data store in a cloud data lake in a de-identified state, and there are no keys present, so there's no method to de-identify the data, it is basically an opaque data brick in the cloud. Secondarily, organization two, or organization N could submit information for Jane Doe, who also has condition, Yes. And we encrypt condition Yes, differently. And we also have the patient ID also encrypted differently.

Harold Byun:

And what we're facilitating in this model is a way to also query on patients that have condition, Yes, to determine the frequency or the prevalence, or the velocity that a certain type of condition is traversing different populations. And so, that's one of the use cases that could dovetail out of this with that as an example of that. What we set up here is a patient database, and so in this particular example, I have organization one submitting data. And much to the person's question earlier, isn't the ID just a numeric value is going to be identified in ... Yeah, I mean, it is, but there's also associations with other information that may be identifiable.

Harold Byun:

And so, what we've done here is, organization one has submitted this set of data, or these records. And within this data set, there's also this notion of, has condition one yes or no? And then I think there's a couple of records that we added towards the end, and so those are additional records. But these other records has condition one that were submitted by other providers are actually not readable. And so, this information actually, represents an opaque data set.

Harold Byun:

Now, if I go to organizations three in this example, and run the same query, you can see that in this particular example ... All right, if I open up this query, you can see that a different set of patients ... Sorry, it's actually available in this particular model and their information. And so, this is the organization three model. In organization three, again, we have a different set of frequencies, but again, has condition one is not available.

Harold Byun:

Now, if I go back to organization one, and if I want to know the prevalence of some of these patients, and I want to run a different view of that data set, I can pull that data set, and I can pull as condition one, but only for my population. But if I wanted to run the overall count and frequency analytics, I'm going to get a count of 15, which is the overall aggregated data set. So again, fairly rudimentary example, but it represents a scenario where multiple parties have submitted into a common repository,

nobody can see anybody else's data. But yet the aggregate analytics occur on a de-identified data set, and this is an area that we're starting to see a ton of traction across fraud scenarios, cross party marketing scenarios, and obviously, any type of health data analytics.

Harold Byun:

Just one other scenario around that, so this is another scenario around the shared threat intelligence, and in this case, we've encrypted or de-identified threat data, so target IPS, and threat data information. And so, we've overlaid this with Tableau, who we have a partnership with, and in this case, we are showing trending information of the de-identified data set and allowing for cross-tabulation frequency across the encrypted data sets. It is a live data set that can be re-rendered and visualized by an analyst or a data science person, but the more important part is, again, the simplicity that has been enabled by not having to modify the application code, Tablue is an off-the-shelf application, there's no way that this application could be modified, and in combination it's being used with the advanced privacy preserving analytics mode to present a different pivot of the data.

Harold Byun:

So, that's the portion of the live demo. In summary, cloud data lakes are going to give you a lot more flexibility to accommodate your data growth going forward. You can leverage data-centric protection methods that are available today to reduce your risk and unlock the business, and ultimately that's going to help you monetize data going forward. And I know a lot of people may bristle at that notion, but that's just the reality of the world that we are living in and continuing to move towards. And you can look at different types of operational models that, again, are not going to slow down DevOps, or the business from adopting these types of practices.

Harold Byun:

Again, additional links for more information, baffle.io/dfp, for some of the data-centric file protection that we have/privacy. We're happy to engage in any type of conversations that if you're not interested in sharing your information or question more broadly here as well.

Harold Byun:

So, let me switch back to the questions just to see other things that have come in. So, is Baffle SaaS, and does it have a SOC-2 certification? So, we are not a SaaS provider, we allow organizations to basically stand up their own data protection service. And so, it's the ability ... We basically give you software to deploy your own data protection service, we never want to see your data, we don't want to have access to your keys. We facilitate a way for you to implement that without having to modify your code. And in many ways, it's just simply pointing your environment to a different connection string, or a different IP address, or a different host name, so that is hopefully the answer to that question.

Harold Byun:

Is the information sent to you? Are you pulling it? We're not touching your information again, so what I'm just walking you through is an example of how an organization might pipeline their data in their own VPC for cloud, so it's never being sent to us. This is again, software that you would manage and set up in your own infrastructure.