Speaker 1 (00:00):

Podcast confirmed. Welcome to The Official AWS Podcast.

Jillian Forde (00:06):

This is The AWS Podcast. I'm your host for today, Jillian Forde, and today we are talking about AWS Entity Resolution, a new service. Always really excited when we're talking about something brand new here at AWS. And I am joined by Shobhit from the AWS Entity Resolution Team.

(00:27):

So Shobhit, tell us what do you do at AWS?

Shobhit Gupta (00:30):

Hey Jillian, nice to meet you. I am a Product Manager at AWS. I'm Principal Product Manager.

Jillian Forde (00:35):

So exciting. And we're going to talk about this brand new service, AWS Entity Resolution. But before we do, I'm sure there's a lot of people who are listening today and they're not really familiar with the term entity resolution. So maybe if you could just give us a primer on what exactly is entity resolution?

Shobhit Gupta (00:54):

So before we get to that, I think we really need to understand what does entity mean? Entity means an object or a data or a record. It can be a record of anything. An entity can be a record of a product code, a product SKU. It can be a consumer interaction that happens on a website. It can be literally any piece of data. And what we've heard from customers is that there's a problem of linking these data sets together. So what entity resolution does is it matches and resolves related records using advanced configurations such as rule-based and machine learning to ending up improving your data quality.

(01:28):

So we have been building the service for two years now, and we have met customers across all verticals. And I think starting up, I think the top vertical who always talks about the problem of data deduplication is advertising and marketing. In advertising and marketing, you have a customer interacting with your brand across multiple channels like they might be talking to you or a contact center, they might be going and making a purchase on your website. They might be going and physically entering into a store.

(01:55):

And every time they interact with your brand, they provide their personal information. They provide information related to what transactions that they have made with it. So advertising and marketing has these desperate set of data records that they want to tie together to create a customer 360 view, which ends up helping the customer because then the brand is able to go and personalize their communication with their end customer. So that is one which is advertising and marketing comes up a lot.

(02:22):

Another one that comes up is product reconciliation. What I mean by that is when you are a retail customer, like think about that you are a brand who has an e-commerce presence and you have a lot of different product SKUs and you have inventory of those product SKUs coming in from different channels. Not all suppliers might call the product code the same thing. Someone may call it X and someone might call it X dash. And that's where it becomes very hard to analyze how all of these different supply chains are bringing in that inventory and to reconcile these different SKUs together to holistically create

analytics around it, on product analytics around it. So that's second, that comes up a lot is product reconciliation.

(03:04):

And the third bucket, which I mostly like to put it in, is around machine learning. A big part of what all the data scientists and data analysts that are listening to this podcast do every day is prepare their data, clean up their data, remove duplicates in their data to train machine learning models. That's what the service will help you do is like if you use this service, we can easily remove all the duplicates from your data within a few clicks so that you can get back to building your ML models with highest data quality to get better results.

Jillian Forde (03:33):

There's so much data, and at the end of the day, I mean it sounds like for a lot of use cases. In fact, I actually can't really think of a use case where at some point you're going to have enough data in disparate places where you're going to need to be able to reconcile them. So it sounds like it's only a matter of when it comes that time where you're going to need an entity resolution solution.

Shobhit Gupta (03:57):

Absolutely.

Jillian Forde (03:58):

I bet there's developers here who are listening today, and they're really savvy. They've probably in a previous job, maybe in their current job, they've had to build an entity resolution. So they're probably sitting here and they're like thinking in their head right now, "Okay, so I've built this in the past. Really what's the difference between me just building it the way I'm used to versus actually just using AWS Entity Resolution?"

Shobhit Gupta (04:22):

Yeah, I mean, you are absolutely right, Jillian. This problem is as old as writing SQL queries. I mean, if you have databases going back 20, 30 years, all the customers that we have in AWS would have built some form of homegrown system, which is a bunch of complex SQL procedures a couple of pages long where they are trying to find duplicates in their data just for data cleansing, right? I mean, studies tell us that more than 94% of organizations have 20% to 30% duplicate data. That's the status quo.

(04:54):

So you're absolutely right. Customers have come and told us they already have homegrown systems where they are trying to solve this problem. The way we differentiate that is two things basically. One is we provide a highly configurable rule-based matching engine, which is just a few clicks away. So you don't have to write any SQL, you just don't have to go and write any code. The rules are easy to configure. For example, we support up to 15 deterministic matching rules in the service where you can change the priority of the rules. So for example, some rules for deduplication are higher priority for you than others. So it's very highly configurable, easy to use. That is one part of it.

(05:32):

On the other side, we've really invested to gone big on using machine learning models to solve this problem. And the difference between using rule-based and machine learning models is that, and machine learning models is not biased by your view of duplication. A machine learning model takes the entire record together and analyzes that record against all the records that you have in your system to

really go and find the duplicates for you. And it's take care of things like spelling mistakes, it takes care of things like missing information. You might have a product SKU code, which has some part of it missing, whereas in other SKU code it is there.

(06:10):

So that's how it kind of differentiates from traditional kind of homegrown solutions, ease of use, configurability, cost, absolutely.

Jillian Forde (06:17):

That is so cool, and especially because not many people I would imagine who have built their own solution in the past have used machine learning techniques. It's probably just a bunch of SQL joins and Python Swiss Army type of code to be able to actually come up with stitch together something.

(06:36):

Another question I have for you, Shobhit, is based off of your experience really understanding this problem space, really you could walk through the challenges that companies are facing when they're trying to look at cleansing their data. And I know that the people who are really experienced who've done this problem before, they know because they've had to face the pain and struggle of having to do this. But we've got some people here who are newer in their cloud journey, their development journey, and so they haven't faced those paper cuts yet. So maybe you can enlighten them as well.

Shobhit Gupta (07:09):

Yeah, absolutely. I would take an example, right? For example, whenever you go into a data joining problem, the first problem that you have in all your records don't have a unique identifier. So if you had a unique identifier across all of your records, you can just join them using one key. So it's very simple to do that. So that doesn't exist. So that's the first problem.

(07:28):

Second problem is that because all the records have their own kind of unique keys, you need to rely on some metadata in that record to join those records. Now that metadata has inputs from different systems. So each of those records have different schemas. They don't have the same data schema. So you need to first make sure you solve and unify the schema. That's the first level of problem. The second level of problem is that now that you have unified the schema, the data in itself has a lot of quality issues, and sometimes there are spelling mistakes, sometimes there's missing information. Sometimes there is human input in the data, which adds some bias to the data and so on and so forth. So there's some data quality issues.

(08:08):

Then once you have done all of that, you have tried to create a unique ID. You have tried to create a uniform schema. You have tried to remove all the data quality issues and prepared the data, normalize the data. Now you actually have to write code, which I was talking about earlier, to build these SQL joints or to build these machine learning models on your own and have an applied scientist team and a developer team, which is just ongoing maintenance and cost.

(08:33):

So I would put the challenges in these three buckets. One is the inherent challenge related to building this technology. So cost and maintenance costs. That's one. Second is just a hard problem to solve because data comes in different shapes and sizes. It comes from different systems. So data schema

resolution is the second problem. And the third problem is more around, which is now coming up a lot, is related to data governance and security.

(08:57):

You want to make sure that all of your data has the right data regionalization. Like data in one region stays in that region, doesn't move to another region. Data in Amazon S3 doesn't get out of Amazon S3 again and again to different systems. How can we minimize the data movement to make sure your data governance is in place?

(09:13):

So that's how I'll summarize. I would say that cost, technology and ease of use of meeting the governance and security compliance are the biggest challenges.

Jillian Forde (09:23):

Wow. So please tell me then that AWS Entity Resolution addresses a lot of these challenges that you're sharing with us.

Shobhit Gupta (09:30):

Absolutely. Yeah. I would say that the top three solution spaces that we have developed here, number one is ease of setup. That's a ground zero for AWS services. We want our services, especially in this case, Entity Resolution service to be super easy to use and very flexible, scalable, and seamlessly easy to integrate with existing applications.

(09:48):

So that's what happened with Entity Resolution. You can get started within minutes. You just have to point us to your data sitting in S3. We pick up your data from S3, we schematize that data for you, we normalize that data for you. We even hash that data for you so that it is always secure. And then we provide you a preexisting, pre-configured, rule-based matching and ML-based matching engine to match that data as well, which we just got from an S3 bucket and the setup is just a few minutes. So it's super easy to set up. That's like the first one.

(10:20):

Second one is configurability. All the builders out there know that we like ease of use, but we also want to double click. We also want to understand what is behind the scenes on how can we change things because everybody's business logic is different. Similarly, in the data output side, it is highly configurable. You can write the data output anywhere in your S3 bucket. You can even encrypt your outputs for downstream use cases. You can even hash your data output because sometimes you want to send hash data downstream.

(10:49):

You can have different processing speeds that we offer. You can have a manual processing where we take all of your data in batch format on a specific cadence that you set up. You control the cadence. That's the manual processing of data. Or you can also have automatic processing where we automatically pick up all the new data. So configurability is the second theme.

(11:07):

And the third one is more security that we talked about before, as well as we minimize your data movement because all of these services run in directly at your S3 level. So your data always remains in S3. There's minimal movement. There is no movement outside AWS cloud. So I would say I'll wrap up

with saying ease of use, configurability and security are free ways to address the challenges that I just demonstrated.

Jillian Forde (11:29):

Wow. It's just absolutely amazing how much already is it built into the service that customers don't have to actually take care of. Really cool.

(11:38):

I know that as you were explaining really what AWS Entity Resolution does, you started to talk about some of the areas that AWS Entity Resolution can solve certain challenges. Are there any others that you didn't mention before? I know you were talking about deduplication, but was there anything else that you wanted to call out?

Shobhit Gupta (11:57):

Yeah, I would say only two others originally stand out. One is customer profiles. For example, a big part of what every marketer and advertiser wants to do out there is create a customer 360 view of their customer. And this service can help you do that because this service will give you the technology to connect different data pieces together. So it'll easy for you to stitch your own profiling. So I think that would be one.

(12:18):

Another one I would say, which is more with respect to finance. For example, if you're a financial services company, a big part of you is to prevent frauds and to prevent fraud detection, which also boils down to looking at one transaction and looking at another's transaction, which is nothing but an entity or a record. And then trying to resolve these records together, and making sure these two records are actually belonging to the same individual. Hence, it's not a fraud, for example.

(12:45):

I would say these are two, which has not come up in the discussion until now.

Jillian Forde (12:48):

Wow, really cool. All right, so for the people who are listening who have already built their own entity resolution, can they integrate AWS Entity Resolution with their current workflow?

Shobhit Gupta (13:01):

They can absolutely do that. The way we offer our services is that it's pick and choose. It's like drag-and-drop, right? So if you already have rule-based matching, some of our customers, when we ran a beta recently, they just ended up using machine learning based matching because that's what they wanted to augment their existing systems. Some of them said, "that our existing systems don't have the right data processing speeds and we want this to be a more incremental. Within five minutes I want to resolve all of these things together," and then they wanted or ended up using the entire service. So it absolutely seamlessly integrates with whatever entity resolution systems you have.

(13:33):

It also seamlessly integrates and augments if you already have entity resolution as a part of your CRM systems or CDP systems because most of these systems use deterministic matching, and we are able to augment that with machine learning based matching.

(13:45):

We also make sure that all the new technology that we build in this space over the next couple of years automatically is available to you as a part of your existing APIs. So all the APIs that you integrate with iteratively will keep updating and will keep adding new features to it to keep improving the matching accuracy.

Jillian Forde (14:00):

Really cool. So if people are interested, do you have any parting advice for them before they get started?

Shobhit Gupta (14:07):

If you're a builder out there who is remotely associated with building analytics workflows, or you're a data scientist or applied scientist out there who is working in machine learning, if you're a marketer out there who is working in sales and marketing and wants to get a better view of their customers, if you're in customer support, you want to get better at providing a better experience to your contact centric agents as well as your customers, if you're in operations and all you're trying to, and you geek out on supply chain issues and you want to resolve that, for all of you, data quality is a foundational problem that AWS is listening to and is trying to solve with this service and really cool and wants to try and solve that over the next couple of years. So that would be one.

(14:51):

The second would be how to get started. That's the most important. There are three ways to get started. You can just go to AWS Management Console, type AWS Entity Resolution, and boom, you'll see the service in front of you.

(15:02):

Number two, there is a news blog that will be out. In that news blog we will talk a lot about how the service runs. It'll be like a slide with screenshots and so on and so sort of the service. So that would be second. We are also going to be on AWS On Air, so you can listen to more detailed demo of how we are going to go and how can you get started and use the service.

Jillian Forde (15:23):

Well, this is exciting. I think for a lot of people who currently are facing this problem with entity resolution, they can totally relate to this problem and what AWS Entity Resolution provides. And for a lot of people who haven't come across this, it's only a matter of time where they're going to need to be able to actually reconcile that data. So this is really exciting for customers. So I'm excited. Shobhit, thank you so much for being here today on The AWS Podcast.

Shobhit Gupta (15:49):

Absolutely, Jillian. It was my pleasure. Thank you.