Speaker 1 (00:00):

Podcast confirmed. Welcome to the official AWS Podcast.

Jillian Ford (00:08):

This is the AWS Podcast. I'm your host for today, Jillian Ford. And this episode is all about data quality and we're specifically going to be talking about AWS Glue Data Quality and I'm here with Shiv. So Shiv, tell everyone what is it that you do at AWS?

Shiv Narayanan (00:27):

Hey Jillian. So hello folks. My name is Shiv Narayanan. I'm Product Manager, focused on Glue, more specifically on data management features like data quality, sensitive data detection. I'm also the Product Manager for Glue Streaming that allows us to gain insights from our realtime data.

Jillian Ford (00:46):

All things Glue. I love it. And for someone who hasn't used Glue before or they're new to AWS, what is AWS Glue?

Shiv Narayanan (00:55):

Glue is a serverless data integration and data management service. It makes it easy for you to connect to a variety of data sources and transform data at scale. It also helps you manage the quality of your data and help protect sensitive data.

Jillian Ford (01:08):

Exciting. And there's a new feature within Glue that is now generally available called Data Quality. So tell people what is Glue Data Quality?

Shiv Narayanan (01:19):

No, that's a great question. So customers constantly struggle to ensure the quality of their data. They today have data lakes or data warehouses and they use it to make decisions. But these data repositories can quickly turn into data swamps if they don't manage quality of their data. Great example that I learned very recently, one of the customers whom I was working with, somebody, something happened in one of their accounting systems where a $6 coffee got somehow fat-fingered into a $60 coffee. Hopefully it's the world's best coffee, but unfortunately it was not great for their reporting and the customers really spent a lot of time figuring this out because thankfully their point of sale system where customers were really purchasing coffee was not impacted but all the other systems. And these issues cost customers a lot of pain because they had to spend days and days to find where this issue is.

(02:13):

And that's where Glue Data Quality helps you. So Glue Data Quality is a serverless data quality service that automatically computes data statistics and creates data quality rules for you as a starting point. You can then take that data quality rules and augment that with out of the box data quality rules that we provide and evaluate the quality of your data. Once evaluated, we give your data sets a data quality score and using the score your business users can take confident business decisions. So that's what Glue Data Quality is. And another couple of things to add about that is that it also allows you to set up alerts, notifications so that it's very easy for your business users to find out if there's an issue with quality of the data.

Jillian Ford (03:02):

That example that you just said, I think every single person can relate to it at some point. I know I had a data entry job and I probably was that person that fat-fingered the coffee being too expensive 'cause it's so easy, there's so much data and then you're maybe not paying attention and it's really easy for data quality issues to arise. So when people are thinking about really data quality and how they can create a workflow for being able to ensure that they have quality data, walk us through what would that actually look like?

Shiv Narayanan (03:35):

So data quality actually is of relevance to a lot of personas within an organization. So for example, you'll have data stewards or data governance teams who are not very familiar with coding and data engineers who love to write code and it's equally important for both of them. So I'm going to explain to you a workflow for both of them. First, let's start with this non-coding personas who are data stewards or data governance teams whom for the most part work in our Glue Data catalog. So we have put data quality at arms length for all of our personas and we've putting data quality in their workflow. So if you are a data steward accessing data catalog, there is a tab right there next to the data catalog in one of your tables. So you click that and then you'll see a data quality tab.

(04:21):

You click the data quality tab, then there is an option for you to start creating data quality rules. So you click that and then you'll be prompted where we'll ask you if you want us to recommend data quality rules or whether you want to manually create it. So the better option is actually for us to do that heavy lifting of creating data quality rules for you. So you simply click Recommend me data quality rules. We then start analyzing your data gathering statistics and we automatically create a set of relevant data quality rules by analyzing your data. Now you can review them, you can take away the ones that you don't like. You can add more from the out of box data quality rules that we have created for you or treat them, edit the values and thresholds. Once you do that, you save them very easy and then you run them.

(05:06):

On completion you will get a data quality score, higher the better. And then you basically take them and you can schedule them using step functions or you can use any other choice of scheduling through our APIs, you can schedule them and run them. So constantly running them allows your business users who are accessing the business data, the data catalog to see the score and then use that data and make confident business decisions. Now that is this data steward flow. Let's now go into the data engineer flow. Data engineers spend a lot of time on Glue Studio or Glue Studio Notebooks. They constantly write code and it's very important for them to check the quality of data even ahead of time because they don't want to spend an hour processing their data and then find out that there's bad data. So they easily can introduce something called a data quality transform.

(05:54):

That's one of the transforms in Glue Studio. So you add them to your job, you configure the data quality rules and then it provides you two outputs. One is the summary of all the rules that you created and it specifically tells you which records fail, what are your bad records. Now you can do a lot of actions with that. You can take all the bad records, write them into a quarantine zone, and then you can take all the good records, process them and then write them to your records and then go back to your bad data after you finish processing. Or maybe next day morning you come in and review the quarantine zone

and take action about it. So these are the two workflows that you can actually have within the data quality.

Jillian Ford (06:33):

I love that there's different workflows for different types of people based on their skillsets and how they're using Glue. I know people are definitely opinionated on how they want to be able to use AWS and be the most productive. And so it's awesome to hear that that's something that is already included within data quality. When I was doing my research, one thing that came across to me was DQ. I thought that was really interesting, this open source library that's being used. So can you talk through maybe what DQ is and how Glue Data Quality uses it?

Shiv Narayanan (07:05):

That's a great question again. So DQ is a Amazon open source library that was built to manage data quality for internal Amazon datasets. It has 2000 stars in Git, a very popular framework that our customers really love and it's been proven and tested at Amazon with data lakes with over 60 petabytes. So really a very, very well tested framework. Actually, Glue Data Quality is built on DQ, and what we have done is we've taken DQ and we have enhanced it by adding a whole new types of rules. We have added the capabilities to provide role level identification of records that will all be soon contributed into the open source. And we also have introduced a new language to author data quality rules called data quality definition language, which is really, really simple for any type of persona to write. It's a very declarative way to write data quality rules really.

(08:07):

And then we've used DQ as our engine to run data quality rules for Glue Data Quality. So the biggest benefit that customers get for this is that they're actually authoring the data quality rules in an open language, that open source framework that is not locking them. And I think that's the premise of Glue as well is to have that openness in data engineering pipelines and so does it gets extended into data quality.

Jillian Ford (08:31):

Wow, that's really exciting. Thanks for sharing that. So you talk a little bit about one scenario and one example, but can you tell me more about some other scenarios where you can imagine that people would want to be able to use Glue Data Quality?

Shiv Narayanan (08:46):

Absolutely. So first of all, I think data quality is something that everyone can use. Not using it means that you're shipping a product to a store. It's almost synonymous to hipping a product to a store without testing the quality of it. You're actually letting your business users consume data without basically checking what went in and what's coming out of it. So I think some of the scenarios specifically for data engineers would be check your data quality ahead of time. The first step probably in your data pipelines is actually to check for your ingredients that you're actually sourcing it. And that actually helps you in two different ways. One is it's better processing the process. The second important thing is that you're going to save on money because you really don't have to run all those jobs. If the quality of data is bad, you can stop your processing.

(09:38):

And as you know, Glue is all about compute, so you can save all that compute and thereby save costs. The other alternate way to think about this is to put this at the very end of your data pipelines. And the reason why you want to do that is sometimes you have to source data from multiple places and you have to do some aggregations and then you have to check the quality of your data. So that is another good scenario where you can check that.

(10:00):

Thirdly, you can also check after the data is ingested in a platform. And that is an interesting scenario because your business users can constantly look and get assurance that the data's really good. Very interesting use cases that I've seen from customers very recently is they have this concept of data mesh and the concept of producers and consumers and accepting some data that is coming in or is actually the premise if you are going to be able to run a bunch of data quality checks. So you can actually hand them over a set of data quality checks and they can run them to be assured that they're actually getting good quality data. So these are some of the scenarios that you can actually use data quality in your workflow.

Jillian Ford (10:40):

I love the scenarios that you called out because it really shows that there is flexibility based off of how customers' data strategy, maybe different data sources is going to evolve, that there's different ways that you can be able to check for the data quality really as your business evolves.

Shiv Narayanan (10:58):

Yeah.

Jillian Ford (10:59):

Can you get more into, you were talking a bit about some of the statistics and the scores that Glue Data Quality provides. What exactly are those?

Shiv Narayanan (11:08):

So in order for us to create data quality rules on your behalf, we really have to analyze your data. So the statistics we calculate is all the way from the most simplest statistic that you can calculate, which is the minimums and averages because let's get this straight, those are important, then they can actually indicate to the real quality problems to really advanced ones, for example, the entropy of the data and then the column correlations, how a column correlates with another one. So these are some of the core statistics that we use to identify the quality of data.

Jillian Ford (11:41):

Wow, so interesting. You're really opening, I think, the door for companies to start thinking beyond just the data quality and now understanding relationships, opening up maybe opportunities for machine learning. There's really more use [inaudible 00:11:56] it sounds like.

Shiv Narayanan (11:56):

A lot of that. Yeah, that is right. There's so many use cases and we really excited about what this future looks like. There's so much that we can do in this space. I'm truly as a product manager, super excited about what's the things that we built.

Jillian Ford (12:09):

So tell us then now that it's really GA, what are some other features then that people can expect to see?

Shiv Narayanan (12:15):

Yeah, so all the things that I spoke to you so far, for the most part are things that customers had really in preview and in the GA we really have listened to some of the core customer requirements where they were not able to complete the workflow. And I'm really happy with our GA release. So some of the key cool features that we are launching, the first one is the ability for our customers to detect which records failed. All right? It's one thing to apply a bunch of data quality checks, but now customers can really go and pinpoint and say, "You know what? The record number 53 or this specific customer had this particular value that was really bad." So I want to go and fix my source system or put a quality check or even a put a validation in my source system so that this issue never happens.

(13:01):

So that's a feature that has come in as part of GA. Then you have new rule types. So in the past you were able to only check quality of data just on one data set, but our customers want to check quality across data sets. For example, you have a bunch of states or you have an employee record that's coming through and then you have a master list of employees and you want to make sure that the employee data that comes through really is a valid employee. So you can now start setting up data quality checks across data sets. You can compare two data sets and see if they're valid. And a lot of customers who use DMS have actually asked us for this check because they want to make sure on a end of the week, is my data in database A matching to data in database B, and they can do that as part of this GA release. We've really simplified the alerting mechanism, the integration of data quality results into other systems through our event bridge integration.

(13:55):

So that's really powerful because now customers can just simply create one rule in Event Bridge and we publish all the metrics into Event Bridge and they can set up simple error notifications that can then go and alert people. In the past it was not available, now it's available so that's actually pretty cool. We really expanded our support for multiple types of data sets. So Redshift and RDS, you can actually now start running data quality checks across that from the data catalog. We're also supporting Apache Hudi and Iceberg and Delta. These are the open formats that customers are using today to build data lakes.

(14:31):

And then I'm very excited about the new user interface for our non-coding users, like our data catalog user interfaces is now revamped and I think it's a pretty cool user interface now. And then there is the deployment aspect. After you finish all the setting up data quality rules and testing them and running them, you really have to be able to take this and deploy it into production. So we've added cloud formation support for that, 'cause that's a very common way in which customers really migrate rules. So that's available. So these are some of the key features. It's more the documentation, but just to highlight, these are some of the ones that we're launching for GA.

Jillian Ford (15:05):

I'm always a sucker for a really nice UI. Any good gooeys, I'm a sucker for, so I'm definitely going to check that out. And I also realized I was probably throwing in some lingo, which was totally unnecessary. I forgot, I'm sorry everyone. So GA we are just saying that it's generally available before it was in preview, but now it's something that you can just go into the console and be able to just deploy it or you can be able to use the CLI, however it is that you use AWS, it's now generally available. So sorry for using

that terminology everyone, that people were probably listening, they're like, "What are these two talking about?"

Shiv Narayanan (15:41):

Yeah. That's a good call out.

Jillian Ford (15:46):

But now that we got that confusion settled and people are excited to be able to try this, what advice do you have? What are some best practices maybe for customers to be able to use Glue Data Quality?

Shiv Narayanan (16:02):

So use it as early and as often as possible because really it can improve the quality of your decision making and it is a direct business impact driver to a certain degree. So the first thing that I would suggest is to always, for across your pipelines, think of one or two data quality rules. The simple ones are good, at least as a starting point. One of the customers was telling me they really want to enforce this policy where every pipeline should have the basic checks as part of that because that's how much they've been struggling, getting good quality of data. So I think thinking about it is a great way. The second one is there's always customers who ask, "How do I optimize on cost?" Because that's the big thing, like "Hey, running data quality checks, but then I want to be able to not pay so much."

(16:55):

And that's the beauty of Glue and having data quality as part of Glue, because as part of your data pipelines, we've already processed your data. We've already read your data. And just adding those data quality rules as part of pipeline is just going to result in that incremental cost as opposed to us having to reprocess all of that data. So think of that as one of the mechanisms. Other way to think of this is running data quality on an incremental fashion, not necessarily thinking of it to be running on all the dataset all the time. Maybe this first time you run it on all the data and then you can run incremental and you can do that in the catalog by using the push down predicate options that we have provided and in the Glue Studio by using our bookmark feature. So there's something called bookmarks, which automatically track which files we have processed and which files we are not processed from S3.

(17:43):

And of course you have other options to manage for RDF, so you can use those to run data quality checks. And one of the other common things that I've started to see is quarantining the bad data, making sure that you make the data available as early as you can and then qualifying it as opposed to just aborting your jobs. We do have that option, by the way, that you can absolutely stop processing the job, and I think it makes sense when you don't want to incur the cost. But a lot of times, I think one of the better practices is just to simply make that data available with all kinds of error messages and let people decide. Because sometimes having that data access is the most important thing, and qualifying that maybe some errors that we thought are super bad is probably not it. So we provide those alternatives as well. The ability for you to stop the job or continue the job, loading them. These are some things that I would consider as you're embarking with the journey of setting up data quality rules.

Jillian Ford (18:34):

Really good advice. I've learned a lot. I'm sure a lot of people have learned a lot and are going to want to listen to this again, especially as they're getting closer to be able to start running their first data quality

jobs in AWS. So Shiv, thank you so much for taking the time to be here today with me talking about Glue Data Quality.

Shiv Narayanan (18:55):

Thank you so much again for having me.