(232) AWS re:Invent 2023 - Build an end-to-end data strategy for analytics and generative AI (ANT331) - YouTube

https://www.youtube.com/watch?v=-SLnInVafho

Transcript:

(00:01) - Hello and welcome. I hope your re:Invent is up to a great start. My name is Chanu Damarla, and I'm a principal product manager with AWS Analytics. And in this session, we'll talk about how to build an end-to-end data strategy for analytics and generative AI. We'll start, oh, we will start with an overview of why you need an end-to-end data strategy.

(00:30) Then, we will talk about how AWS can help you build this end-to-end data strategy. Then I will turn it over to Ram, who will show you a demo of some of the innovations we talk about in this session. And finally, we'll turn it over to Kiran, our friend and guest from Fannie Mae, who'll talk about how Fannie Mae built their end-to-end data strategy on AWS.

(00:54) Now, we're presenting on a Monday, and there are a lot of great sessions ahead of us this week. So where possible, I'm going to include a call out at the bottom of the screen about interesting sessions that are relevant to what we're talking about on that slide. For example, this call out is recommending you to tune into Swami's keynote where he will talk about the innovations in our databases, analytics, machine learning, and generative AI sessions.

(01:24) Let's dive right in and start and talk about why we need an end-to-end data strategy. It's so easy, especially this year, to focus in on that new generative AI application, the increasingly accurate machine learning models or the insightful dashboard that we miss the proverbial iceberg. These generative AI applications, machine learning models, and dashboards are built on solid data foundations.

(01:56) This is where the complexity is, and this is where the hard work is. And building a solid data foundation for your company is the first step in deriving value and insights from your data. Unfortunately, sometimes this can be challenging, because you have to break down the silos that exist in your organization.

(02:20) You may need to break down the data silos where your data lives across disparate databases, data warehouses, data lakes, and even third-party systems, like Salesforce or SAP. You may need to break down people silos by making the data and analytics self-service, so it's easily accessible to everyone inside your organization, including your less technical people in your organization.

(02:45) And finally, you may need to break down business barriers that prevent cross-account across organization data sharing due to compliance issues or cost attribution. To overcome these challenges, companies of all shapes and sizes are building decentralized, end-to-end data strategies that let data producers with the domain expertise build and share curated data products across their organization.

(03:16) These curated data products are then utilized by data consumers who understand business priorities and use these data products to drive business results. And helping coordinate the data producers and the data consumers is a data foundation's team that is responsible for selecting and deploying the tools that enable the various stakeholders to easily share data.

(03:44) Now of course, all the sharing has to be governed in order to ensure that organizations comply with applicable regulations. To implement this strategy, customers often use a multi-account architecture on AWS. Data producers use separate accounts to isolate data products from each other. And increasingly, we see data producers using separate accounts to manage the cost of creating a data product separately from the cost of sharing that data product across the entire organization.

(04:20) In addition, data producers are responsible for their own infrastructure and for meeting business-defined SLAs around data timeliness and data quality. To facilitate the sharing of these data products across the organization, the foundation's team typically provides discovery tools, like a business data catalog, tools for performance management, and they are responsible for defining the governance, auditing, and compliance requirements for the entire organization.

(04:52) Then consumers can now discover and subscribe to these data assets, and they use (audio breaks) machine learning models or applications. To help you build your end-to-end data strategy, AWS offers a comprehensive set of purpose-built services for a variety of use cases, optimized for cost and performance. And many of our services support multiple deployment options, so you can get started quickly by using a serverless option, or you can optimize the cost performance of your workloads by running on pre-provision compute, Kubernetes,

(05:38) spot instances, or reserved instances. Now, you may be wondering, do I really need a comprehensive set of tools? And as anyone wearing the wrong size shoes this week will find out, one size truly does not fit all. Look, joking aside, a database is not appropriate for every use case, just like a data warehouse is not appropriate for every use case, just like even a data lake is not appropriate for every use case.

(06:09) And in fact, in our experience, it's common for customers to start with one service or an architectural approach, and then as they understand their workload and the usage patterns for that particular application, to switch to another service or another architectural approach, because it is better suited for the task.

(06:27) For example, they may start off with a relational database, because they understand it, and they can get started with it quickly, but then, switch to a non-relational key value store, because it is a better fit for their use case and allows them to really fine tune the cost performance profile for their applications.

(06:44) And to support customers, AWS provides a comprehensive set of services to help them store and utilize data, to help them integrate the data across their organization, so they have visibility into their entire business and customers and to help them govern their data assets, so they can comply with their regulatory obligations.

(07:07) Let's dive in and take a look at some of the services we provide to help you store and utilize your data. In addition to S3, our durable object store, we offer the industry's most complete set of relational databases, such as Aurora and purpose-built databases, like DynamoDB, a scalable key value store, Neptune, a graph database, and Timestream, a database purpose built for your time series data.

(07:38) These databases are uniquely designed to provide optimal price performance for their respective use cases, so developers always have the right tool for their job. Aurora is our MySQL and Postgres compatible relational database service designed for unparalleled performance, including scalability, availability, and reliability, all at 1/10th the cost of enterprise commercial-grade databases.

(08:15) Now, it's hard to talk about all the innovations with Aurora or any of the services I'll mention in this talk, but I do want to highlight some recent launches from our services that are helping break down data silos inside organizations. And one such Aurora launch is the Aurora MySQL to Redshift Zero ETL integration, which seamlessly replicates data in Aurora into Redshift in seconds.

(08:45) In fact, P-50 is less than 15 seconds, so customers can use Redshift for near-real time analytics on petabyte scale data. And the best part is you don't have to do anything to set up this data integration pipeline. You simply tell us the tables you want to replicate into Redshift, and we take care of everything seamlessly in the background for you.

(09:15) In addition to the database services, AWS offers a comprehensive set of analytic services, and many of these services offer a serverless option, so you can get started with them quickly to build your applications. For data warehousing, we offer Redshift. Now the beauty of Redshift isn't the scale of data it can operate on with consistently high performance while keeping your costs predictable.

(09:44) Redshift today offers 7.9 times the price performance compared to other cloud data warehouses, and we will continue to innovate in this area. Redshift is the only cloud data warehouse that lets you run queries at exabyte scale against your data lake, as well as petabyte scale inside your clusters. In addition, with RedShift's federated query capabilities, customers can now query their operational data stores, like Aurora or RDS.

(10:18) And as data becomes more democratized within organizations, Redshift is delivering on easy analytics for everyone. A recent launch in this area I wanna highlight is the auto complete and syntax highlighting feature Redshift Query editor v2. This feature enables less technical users in an organization to build analytics and queries more efficiently and accurately.

(10:46) Embedding machine learning and generative AI capabilities into our services to empower all users is a theme you'll hear about in many of the deep dive sessions this week. For big data processing, we offer EMR. EMR makes it easy to run big data processing frameworks, like Spark, Hive, Presto, or Flink. It supports the latest versions of these open source frameworks within 90 days, and it provides the best performance at lowest cost.

(11:21) In fact, Spark workloads run five times faster than open source. In addition, EMR has flexible deployment options, like serverless, Kubernetes or running on provision compute. And by running on spot or reserved instances, you can save 50 to 80%. A recent EMR launch that I wanna highlight, it's fine-grained access control with lake formation permissions for Spark jobs.

(11:52) This feature makes it easy to share data across your organization, so instead of creating different versions of the same table that you then share with different user groups that have different permissions on that table, you can now use table, column, and row level permissions to share just the portions of that table with the appropriate users inside your organization.

(12:17) Be sure and check it out. For business intelligence and dashboarding, we offer QuickSight, which allows everyone in your organization to understand your data by asking questions in natural language using QuickSight Queue, explore your data through interactive dashboards or look (audio breaks) in your data powered by machine learning.

(12:42) QuickSight enables business intelligence for everybody. And with QuickSight-embedded dashboards, these insights can now be embedded directly into your operational tools and internal portals, so your users have access to the data that they need with the appropriate business context. For example, Bolt uses embedded dashboards to share deeper insights into how shoppers with Bolt accounts compare with guest shoppers.

(13:16) And of course, because these are interactive dashboards, it's always possible to drill into the details acquired for your use case. Now, to support your interactive log analytics, real-time application monitoring, and website use cases, we offer OpenSearch Service. It is a fully managed service that makes it easy to deploy, operate, and scale OpenSearch clusters on AWS.

(13:47) Customers use OpenSearch, for example, for log analytics, so they can detect potential threats and respond to changes in system state all through an open source solution for observability. Like many of our services, OpenSearch lets you fine tune the cost performance profile of your workloads. Now, we know that the cost of log and application data increase as the data grows, and with OpenSearch Service, you can use different storage tiers to optimize the cost performance of your workloads.

(14:21) For example, you can keep your highest priority workloads on hot storage for fast performance while moving the data for your lower priority workloads to cold storage in order to optimize costs. With all the interest in generative AI, a recent OpenSearch feature that I want to highlight is the Vector Engine for OpenSearch Server Serverless.

(14:48) To speed up searching across all the unstructured, semi-structured, and structured data in an organization, OpenSearch uses a data structure called an index. At a very high level, you can think of an index as an index keeps a list of all the documents that contain a specific word, like these are all the documents that contain the word restaurant.

(15:13) However, what if there are documents that contain words similar to restaurant, like diner or cafeteria or steakhouse or pizzeria? How do you find these documents hen you search for a restaurant? To group these documents together, you can use embeddings, and you can think of an embedding as a mathematical signature for documents that allows you to find documents about the same concepts even if those documents don't contain the exact same words.

(15:46) And with the Vector Engine, developers can now search across their structured, semi-structured, and unstructured data using descriptive text and metadata like they're doing right now, as well as search against these embeddings. Customers can use this capability to not only improve the search results and the relevance of those search results, but to create personalized responses in generative AI applications by finding all the data within their organization about a specific customer, a business, or a specific topic,

(16:20) and use that information to create the prompts to feed the large language models. To support your machine learning workloads, we offer a broad set of machine learning capabilities from support for deep learning frameworks, like PyTorch or TensorFlow, to services like Amazon SageMaker that makes it easy to create your own machine learning model or AI-powered applications, to AI services with built-in machine learning capabilities, like transcribe that can power your speech-to-text use cases or that can extract text, handwriting,

(16:59) and data from your scanned documents. Now of course, many of our customers are interested in how generative AI can help transform their business, and to support these customers, we offer Bedrock. Bedrock is a fully managed service that makes it easy to build and scale generative AI applications with your choice of high-performing foundation models, all while maintaining privacy and security.

(17:30) Bedrock includes all the capabilities you need to build generative AI applications and experiment with foundation models, customized foundation models with your private data, and create agents that can execute business tasks, like booking travel or creating an ad campaign. Some of the foundation models that Bedrock supports includes Claude 2 from Anthropic, Llama from Meta, and Titan from Amazon.

(18:02) (audience member speaking indistinctly) So far, we've reviewed the services we provide to help you store and utilize the data inside your organization. Now, let's dig into the services we provide to help you integrate your data. Let's start by discussing why data integration is important. Data integration is important, because your data lives in disparate databases, data warehouses, data lakes, and SaaS systems in your organization.

(18:37) And you need to integrate this data together to create a holistic picture of your business and customers. Often, this involves developing code to clean and transform your data, developing code to replicate or move data between systems, as well as orchestrating end-to-end workflows. For example, you may want to manage dependencies, like only run this aggregation step once all the data has been loaded into Redshift.

(19:09) data, we offer use-case specific data integration services. For example, to help you And so to help you prepare your migrate data from open source and commercial databases into your data lakes and data warehouses, we offered a database migration service, and if you want to purchase third-party data sets to augment and enrich your data, then you can use data exchange, a marketplace where you can access over 3,500 data sets from over 300 customers.

(19:40) And finally, for orchestration and workflow management, we offer Amazon Managed Apache Airflow as well as Step Functions. And of course, Glue. Glue is our anchor data integration service, and it features a serverless execution engine that can scale to support all your workloads. Glue supports all the users inside your organization with persona-specific authoring tools, Notebooks for your technical users, Data Brew for users that want an Excel-style wrangling interface and Glue Studio for users that want a visual job authoring experience

(20:18) for developing your data integration pipelines and data integration jobs. A recent Glue launch that I wanna highlight is the ETL AI Coding Assistant powered by CodeWhisperer. Using this feature, users can build data integration pipelines using natural language input. In this example, just by putting in that comment, "Write data frame into Redshift," CodeWhisperer will give you the Spark code for writing a Spark data frame into Redshift.

(20:52) While tools like this will simplify the development of data integration pipelines, in many cases, we can do better, and that's why we're investing in a Zero ETL future. Zero ETL eliminates the need for ETL pipelines that you were previously building and managing by hand. And in many ways, this is similar to how we use the term serverless.

(21:18) Serverless doesn't mean that there are no servers. It means that we, AWS, removed the undifferentiated heavy lifting of provisioning, monitoring, and de-provisioning servers, so you could focus on your business logic. And in a similar way, as we progress on the Zero ETL journey, we will evolve our offerings to remove much of the undifferentiated heavy lifting of building, monitoring, and managing data integration pipelines, so you can focus on your business-specific transformations.

(21:52) And in practice, what this means is that we will invest in making it easier for you to access data in place by expanding the federated query capabilities in Athena and Redshift. So instead of building a pipeline to replicate data from your operational data stores into your data warehouses or data lakes, so you can query them, there's no need to move the data, you can query the data directly in place.

(22:20) We will also move machine learning and analytics closer to where the data resides. For example, many of our data services, including Aurora, Redshift, QuickSight, and Neptune have integrated machine learning capabilities in them already. And just yesterday, Redshift announced the ability to access large language models in SageMaker Jumpstart from SQL.

(22:40) And finally, we'll build point-to-point integrations between our services, so we take care of the undifferentiated heavy lifting of moving and replicating data between our services, so you don't have to. This is similar to the Aurora MySQL to Redshift Zero ETL integration that I covered previously.

(23:00) Now, let's review the tools we provide to help you govern access to your data. So far, we've talked in detail about the tools we provide to help data producers create curated data sets, and we've talked in great detail about the tools we provide to help data consumers utilize these data sets to create interactive dashboards, data integration pipelines and so forth.

(23:26) But without the governance foundations in place, sharing data across your organization will be a difficult task. And as recent GDPR and CCPA finds have shown, governance cannot be an afterthought. A proper data governance framework is critical, because it helps you move faster with data while complying with your regulatory obligations.

(23:52) And to support you, many of our tools already have built in governance capabilities. For example, SageMaker helps you address common machine learning challenges from onboarding new users to centralizing model information for multiple users in a single location. To govern your data lakes, we provide Lake Formation which helps you easily build, govern, and audit your S3-based data lakes and it features table, column, and role level permissions, so you can share the right data with the right users inside your organization.

(24:29) And for true end-to-end data governance across your entire organization and all of our services, we offer DataZone. DataZone provides the key components required to share data products across your organization. It includes a organization-wide business data catalog where data producers can publish data assets.

(24:58) Data consumers can then discover these data assets and request permissions to access them. And once that subscription has been approved or the permissions have been granted, DataZone sets up all the permissions including cross account permissions for you seamlessly in the background, so your consumers can just access the data.

(25:23) DataZone is a new service. We just (indistinct) in October, and if you are not familiar with it, I encourage you to learn more about it this week. It really is the connectivity tissue. It provides a lot of the connectivity tissue to build an end-to-end data strategy on AWS. That was a quick overview of why you need end-to-end data strategy as well as some of the services we provide to help you build your end-to-end data strategy.

(25:54) Now, I will hand it over to Ram, who will walk through a demo of how to build end-to-end systems on AWS using some of the innovations we just talked about, Ram? (audience clapping) - Thanks, Chanu. Good afternoon, everyone. I'm Ramkumar Nottath, I go by Ram. I'm a principal solutions architect at AWS focusing on analytics and machine learning services.

(26:25) In the next 10 minutes, I wanna show you how you can build or use the innovations that Chanu talked about to build an end-to-end system in AWS. So let's assume that you have a touring company, and you want to build a system or a chatbot to help your customers to answer questions, personalized suggestions, based on the trip booking information that you have.

(26:56) To build that chatbot, first I wanna show you how you can curate that data set by bringing in the data from operational systems, like Aurora MySQL to your data warehouse, which is Redshift. And then, let's go ahead and share that data with the organization, because there are consumers who want to consume that data.

(27:18) And once that is done, we want to build that chatbot using the curated data set. Now, for all these demos, I will leave some links where we can refer and go in detail and there are some workshops that are available where you can get your hands dirty and try out that portion of this demo. So let's get started.

(27:37) So the first, we are going to curate the data set. So here, we will start by setting up Zero ETL between Aurora to Redshift MySQL, where I have my user details, present in the user profile table and the hotel booking details, present in the whole booking table. Now, earlier you would've thought about building a custom pipeline, deploying it, maintaining it, whatnot.

(28:06) Now you don't have to do that. You can just set up Zero ETL between these two systems, so that you can replicate that data. All you need to do is provide the source database from where you want the data to be replicated and the target where you want the data to go, and rest is taken care by us. Let's take a look.

(28:27) So we'll start by going to Aurora Console. So here you can see the travel DB database. That's my Aurora database, which is a serverless variant there. And I have the hotel booking information coming in here. All the transactions are happening against that. So now, let's move to the Zero ETL integration, and let's scroll down and click on Create zero-ETL Integration.

(28:53) Let's give that a name. Let's say travel-zero-etl and click Next. And now, we are configuring the source data. So we'll click there and select our travel DB database, which is our MySQL database. And once that is done, the next step is to configure the target database. So here, you can browse the Redshift data warehouses that you have, and in our case, we have a Redshift endpoint that is available, so let's go ahead and select that, redshift-genai-namespace.

(29:24) And once that is chosen, the next activity is to select some tags. If you wanna add tags, you can add that, but for now, we'll skip that. And let's review everything. Let's go ahead and hit Create zero-ETL Integration. Now, you can see that the status got changed to creating. This typically takes a few minutes.

(29:45) For us to save some time, I've cut that portion, and you can see that it is going to the target as Redshift server in point. Now, once it is created, you will see the status as active. The integration is done. Now what you need to do is let's go to a target, which is a Redshift Query Editor. Let's go ahead and find out what is the integration ID, because all what we need to do is create the database using that integration that we just created.

(30:11) So we are getting the integration ID from the system tables and using that integration ID, we are going to create a database in Redshift. Let's call it as trip\_data\_zero\_etl. That was it, now your users have access to data in your target environment. Let's go ahead and take a look. Yes, you can see the trip\_data\_zero\_etl database.

(30:37) If you drill down, you'll be able to see the trip schema, and within that, there are two different tables that we were talking about, which is a user profile and holiday booking information. Now let's go ahead and query that. Let's make sure that we have access to that data. So user profile looks good.

(30:54) Let's do the same for hotel booking table. Let's go ahead and query that, there you go. So we have the data available. Now compare this with a custom pipeline that you would've built, how amazing it is, right? Now, let's move on. Let's take a look at how we can share this data across organization.

(31:14) So for that, we will start by publishing the data, using Amazon DataZone, so that the rest of the users in your organization can search for it, consume that data and use that data. Now, this part of the demo assumes that you have your DataZone set up and the environment's up and running. So if in case you're interested in knowing more about those details, please, please take a look at ANT313 session, where you'll have a lot more details in the workshop.

(31:44) Now, just to let you know, here I have logged in into two different browsers, because we are talking about two personas, a producer and a consumer. Now producer is the person who is publishing or the team who is publishing the data, and you have the consumer who is requesting for the data or searching for the data, finding it out, sending a request and consuming that data.

(32:06) So I've put a call out at the top so that you have an idea whether I'm on the producer screen or the consumer screen. Now, because I'm part of the producer team, I have access to the DataZone project, which is the customer publisher project. Now, I have the data source available. Let's go to that specific data source and collect the metadata or bring in the metadata from the systems by clicking that Run button.

(32:32) Once that run is over, it is importing all the metadata from the Redshift database. You can see the user profile and hotel booking table from there. Now, our inventory is done. So let's go to the inventory section, and you can see both the tables there. Let's select user profile data, and here you can see the metadata that is automatically generated by DataZone for you.

(32:58) So you can take a look at it, you can make edits if required, but for me everything looks good, so let's go and exit out. Now, inventory is done, we validated the metadata, now let's go and publish it. So let's go to that specific data set and click Publish Asset, and let's go and confirm, yes, go and publish this.

(33:17) That's all you need to do for your data to be visible or available for your business data catalog, so that other users can search for it and consume it. Now, this time we are moving to consumer. So as a consumer, again, they have a environment available, and you can see that it's a Redshift environment that they are going to work with.

(33:38) So the first thing that consumer is going to do is search for that data set. So they search for user profile information, so they saw that published data set. Everything looks good. This is the data set I want. Let's go ahead and subscribe. Let's hit that Subscribe button, provide a reason for the request, why we are requesting access to this data set and hit Subscribe.

(34:00) Now, this starts a workflow, because now there is a request being sent to the producer that hey, there is a consumer who wants to access this data. Let's now go to producer, because now producer has to act on it. So at the producer side, you will see that hey, there's a notification that a subscriber has requested for this data.

(34:21) So producer takes a look at it, and if everything looks good, pull that decision comment and approve it. Now, that's all producer has to do. Behind the scenes DataZone is doing all the permission that is needed so that the consumer has access to this data from their environment. So let's go to the consumer and take a look at that.

(34:43) So you can see that in the subscribed data on the consumer side, the user profile is listed, and we are going to the consumer's Redshift environment to see if we are able to see that data and query that data. So there is a view that is available now, because it's at the consumer side. So they are using consumer compute, consumer Redshift cluster and querying against that view to get the access to that data.

(35:07) You can work with this, you can analyze, you can do all the other activities that you wanna perform on this. Now, what I want to highlight here is think about how easy it was as a publisher to publish a data set as well as as a consumer to search for the data that you're looking for and get access to it and start using it.

(35:29) Now, let's build that generative AI chatbot. So in this demo, we are using the data from the data that we have curated to power a generative AI application. So we are going to build a travel planning assistant chatbot, which takes the user input and use that to go against Redshift, query the data that it needs, and then go against Amazon Bedrock, which will call a large language model, Claude large language model, and pass all that information to provide personalized suggestions for this user.

(36:08) Let's take a look. So here we're on the the Redshift Query Editor. This time we want to look at specifically on a specific user, because we want to understand what is their interest, what is a hotel booking schedule looking like and all of those details. So here, there is a user ID there, let's go ahead and query.

(36:31) So we can see that Blake likes quilting, ice skating, and tabletop games, and their favorite food is macarons, waffles, and pudding. Okay, now let's look at the hotel reservations. Looks like there's an upcoming trip in June to Manchester, Brussels, and Paris. Now, let's go to a chatbot. The first thing that chatbot is looking for is a user ID, so let's provide the user ID of Blake, because at this time what it is doing is it is going against Redshift and getting all the information that we just looked at

(37:10) so that anytime the user asks a question, it can pass that additional context so that large language model can provide personalized suggestion. So Blake is asking, "Hey, can you help me "with the travel library for my upcoming travel?" So you can see the suggestions that are coming up. So now if you look at it, it considered each one of the cities that Blake was traveling and their respective dates in the suggestions.

(37:37) And after that, within each one of those queries it looked at tabletop games, for example. It provides suggestions for skating and Lawful Waffle and how can we miss some macarons in Paris? So, that's how it is. How about if Blake wants to do a trip in the US? Now this time model understands that hey, user is asking this question, but they don't have a booking information for that particular date or that location.

(38:06) So it still provides an answer, but this time, instead of specific dates from the table, it is giving its suggestions as well as the very specific interest of the user. Now, in the last 12 minutes or so, we looked at how you can build end-to-end systems using databases, analytics, and machine learning services in AWS.

(38:33) Keeping the environment set up aside, we spent almost three minutes in share creating a zero-ETL integration between your transactional databases and your data warehouse and less than three minutes to publish a data set, share that and consume it from a sharing standpoint, and then we also looked at how we can use the data that we curated to power a generative AI application.

(39:00) So with that said, I will hand it over to Kiran. (audience clapping) - Thanks, Chanu and Ram. Basically, I'm really glad they laid out the foundational stuff of why end-to-end data strategies is needed. I'm Kiran Ramineni, Vice President for Single-Family Architecture, Cloud Data and Infrastructure at Fannie Mae.

(39:34) As you all have seen, data strategy has to be inclusive and end-to-end. I'm thrilled to be here to share our journey into data mesh and insights we have learned during the journey as it applies to financial sector. Speaking of financial sector, probably all are familiar with what Fannie Mae does. So basically, we are one of the most valued housing partner.

(40:15) We facilitate equitable sustainable access to house ownership, homing owner, sorry, house ownership, quality, affordable rental housing across America. To kind of put that in perspective, in 2021, '21 and '20, we facilitated around 1.4 trillion US dollars of liquidity into the market. So kind of one in four homes, mortgages facilitated by Fannie Mae.

(40:48) So now how does it relate to you? So when you talk about data, think about all the data you have provided during your mortgage, loan origination to loan servicing, the whole nine yards, all the data, we host all the data. And now think about everyone, one in four homes, who we are facilitating their mortgage and servicing their loans.

(41:20) So all that plenty of petabytes of data we host. Now let's talk about data. Traditionally, organizations struggle with the elusive, integrated data experience. And the fundamental reason is the approach the organization have taken. Typically you see that as a technical issue. And quite often, you see people referring to data asset.

(41:59) They like the problem. See, data is an enabler of a business capability, not an asset, that's the shift, if you will. And if you look at where we all evolved from, the first generation of our data platforms, essentially you take operational data, and you dump it into data warehouses. That's the first generation of data platforms.

(42:22) And the second is basically without ETL, you dump into something like Hadoop and leverage and extract data out of it. And the third is where you're a single account where you take all the data, do some ETL, and dump it into a single account in AWS, if you will. If you look at all those things, there is a fundamental issue there, the one that Chanu was referring to, one size fits all.

(42:53) Your data needs vary by the data domain, the business domain you're in and the persona that leverages that data to generate, to pen power business capabilities. That's where we, like any other company, Fannie Mae also evolved from that particular approach, where you're in a single account, third-generation data platform, all data in one place in data warehouse, obviously, your queries will step on someone else, and there's resource exhaustion issue.

(43:30) That's when we went into multi-account-based structure. Basically, so the way Fannie Mae approached when we realized this way in end of 2020, Q4 of 2020, where we realized, hey, we need a more integrated experience to improve our data integrated experience. And we approached that as a social tech, we took a more social technical approach to it.

(44:03) The word sociotechnical systems is not new. It's actually coined way back during World War era for coal mining. Basically, safety and soundness is very important in coal mining, the process you follow to realize the value is very important. The same thing applies to data as well. You're realizing value out of the data.

(44:24) And if you apply that people process technology issue to data, it translates to ownership, business domain driven, business data ownership, domain data ownership. And data is a product and not as an asset, if you will. And self-service, basically when you need data, you need it and centralizing that intent of provisioning data access slows you down, and it affects your time to market.

(45:03) So basically, and the next thing comes in is safety and soundness is important. If you make it centralized, it slows you down, again, impacts your time to market. That's where federated governance is very important. So Fannie Mae took this approach, and we went into multi-count structure. What you're seeing here is there are several ways to achieve data mesh as a principle, if you will, but we took more of a hub and spoke kind of model.

(45:37) So what you're seeing on the extreme left is business domain accounts. The spokes are where you host your business domain data. When I say data, it's a complete data compute and infrastructure all together, they're all co-located. You don't want to unnecessarily move the data around. That's going to create latency and it's going to impact your completeness and timeliness.

(46:05) The whole idea is they're all co-located. But there's a central function of governance and enforcement of these controls, data quality checks, data ingestion controls, so there is a central governance function. By the way, make no mistake, what I mean by central, it's centralized cohesive control plane, but actual execution, the execution plane and data movement plane is all on these spokes.

(46:35) So think of that as a control plane and distributed data and execution plane. Then there's extreme right you're seeing end users, their needs are completely different from your typical business application capabilities. That's where the end user, where their data scientists, ad hoc business intelligence reports and experimentation greenfield, so that's where you see a completely different spokes for that.

(47:11) Now let's apply the two on what it means on the technology side. So what you're seeing is one or more account for business domain. We don't restrict by one account to one domain. You can have more than one account, if you will. It's horizontally scalable. So you have a business product specific operational account where all your operational data and compute is hosted.

(47:41) And the data is ingested in near real time into the spoke of the enterprise data lake. And think of that as a mesh of lakes and in real time. In near real time, and what you're seeing is the central data mesh governance. Essentially, that's where all the central services are hosted to provide the needed controls to enforce data quality, data ingestion, and data movement and also host the data catalog to be able to discover your data.

(48:17) So essentially, business units manage their own code and infrastructure. It's self-service based, and we have an enterprise data catalog. So one of the things you'll realize is as you mature into this multi-account-based structure and start migrating to data cloud, you will end up with thousands of data sets.

(48:41) So now, discovering this data and the relationship between the data becomes a challenge. That's where the center, a data catalog, which is again, federated into each other's spokes, but has a cohesive control plane in the central data mesh, in the hub just becomes essential. Now, on an extreme right, data insights.

(49:05) You're executing insights, you are data scientist, you have business reporting, now they have a cohesive integrated data experience. So here's the recommendations for data modeling and domain owners. See, there's the two terms I'm going to introduce, data as product and data products. Data as product is basically bring in product discipline to your data.

(49:38) What business capabilities is it providing, the infrastructure and code altogether? Data product is something that provides or empowers a capability. So what you're seeing here is think about business capabilities your data is providing and not as an asset. So in segments, you have one or more capabilities and together, you have a bigger business capability, if you will.

(50:08) So what you're seeing now translate that into how data sets and data products come into picture. A business application can produce one or more, can be producer of one or more datasets, and one or more dataset can be part of a data product. And the data product can span across multiple domains. It doesn't have to be one domain, underwriting or finance.

(50:35) The finance, for example, domain product will span across multiple domains by the nature of it. So data, when you talk about domain-driven data, it doesn't mean one business domain. A data product can span multiple domains. So now, how do they relate? So you heard of, Chanu and Ram, talk about various capabilities.

(51:04) So one thing we have done is we basically everything you're seeing here is contract driven, meaning let's say you want to ingest data, it's transport independent. It's more related to design by contract, meaning today, we use DataSync. We also use S3 or Kinesis Streams as a way to ingest data in near real time.

(51:29) As Zero ETL comes in picture, we are looking into Zero ETL as a way to accelerate ingestion of the data in near real time. So think about design the contract into contract interfaces and make it transport independent. So that also includes your public data sets or external data sets. So basically, what it provides is decoupling the management of data and technology by adopting data mesh as a principle.

(52:02) Now we have integrated data products which crosses domains, making them more valuable. Once you, with the centralized data catalog, what you have is a enterprise-wide terminology to various data attributes. Now discovery of the data, use of the data becomes simpler, basically, enterprise logical data model.

(52:26) And now we also, you have ownership, you have well-established ownership to the data sets or data products, which means upkeep and SLA. There's some accountability to those data products. So now, let's look at the various, just to wrap up, various benefits. Basically, what you have is the ability to adopt new technologies, agility, so you can actually come up to spare, what do you call, improve your time to market and horizontal scalability, because you're going to deal with petabyte scale data.

(53:06) And you have faster access to critical data when you need it because all about self-service. Now you can think about your, you have integrated data experience where you can visualize and leverage data across business domains rather than being siloed in one domains. Now basically, you have independence. If you want to switch out new technologies or adopt new technologies, you can do so, because it is an innovative space, it is evolving, which means you should be able to adopt new technologies for any particular function.

(53:43) And based on the domain, the type of technology used will change based on the business domain. So what are we looking at is, I mean, we have made tremendous progress. We are continuing to adopt and accelerate adoption of data mesh as a principle. We are looking into new technology. Technology architecture is a continuous process.

(54:10) We are looking into new technologies that will simplify our architecture, basically using DataZone, AWS Glue, AWS Bedrock. ML is an interesting one, Redshift ML. So we have this one principle of bring compute to data, not data to compute. What I mean by that is if you can think of Redshift ML and those kind technologies, you're taking compute to data, which means you're eliminating need for data movement.

(54:42) So those are some of those new things we are exploring as well. And to improve our overall governance, we have zero trust as a fundamental need and principle. We are continuing to adopt zero trust principles to improve our overall data governance. So with the conclusions, I'm gonna turn it on to Ram. (audience clapping) - Thank you, Kiran.

(55:12) It's always amazing to hear about our customers are using the capabilities that we build and the innovations we build on their behalf. I wanna make it the conclusion, keep the conclusion relatively short. Really, the theme of this entire session was how to break down the data silos inside your organization and build solid data foundations using the comprehensive set of tools that AWS can provide, which helps you store and utilize your data, helps you integrate the data across all the disparate data sources

(55:44) inside your organization, create a cohesive whole, so you can understand your business and your customers and how you can govern your data. If you are participating in the Analytic Superhero sessions, scan now, so you get credit for attending this session. I'll keep it up for a couple of seconds in case you're doing it.

(56:12) Awesome, and that's the conclusion, thank you. Please be sure and fill in the survey session.

Why do you need an end-to-end data strategy?

Agenda

How does AWS help you build your end-to-end data strategy?

Demo of building an end-to-end system with AWS

How Fannie Mae build their end-to-end data strategy using AWS

Everything that we do are built on solid data foundations.

This is where the complexity is, and this is where the hard work is. And building a solid data foundation for your company is the first step in deriving value and insights from your data

Unfortunately, sometimes this can be challenging, because you have to break down the silos that exist in your organization.



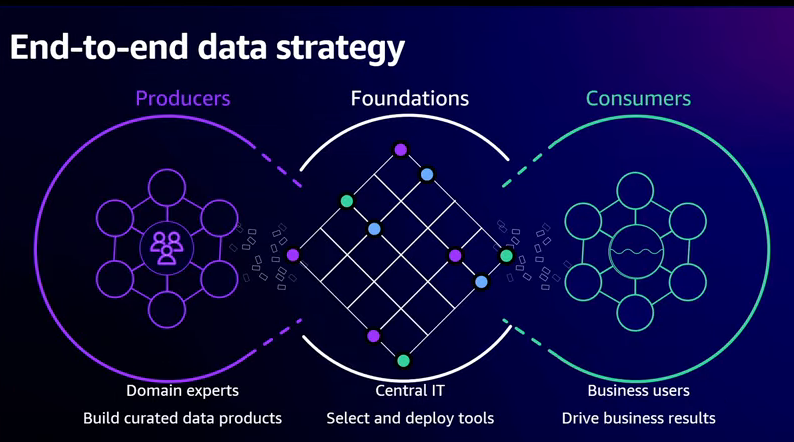
You may need to break down the data silos where your data lives across disparate databases, data warehouses, data lakes, and even third-party systems

You may need to break down people silos by making the data and analytics self-service, so it's easily accessible to everyone inside your organization, including your less technical people in your organization

you may need to break down business barriers that prevent cross-account across organization data sharing due to compliance issues or cost attribution.

To overcome these challenges, companies of all shapes and sizes are building decentralized, end-to-end data strategies that

* let **data producers** with domain expertise build and share curated data products across their organization.
  + In addition, data producers are responsible for their own infrastructure and for meeting business-defined SLAs around data timeliness and data quality.



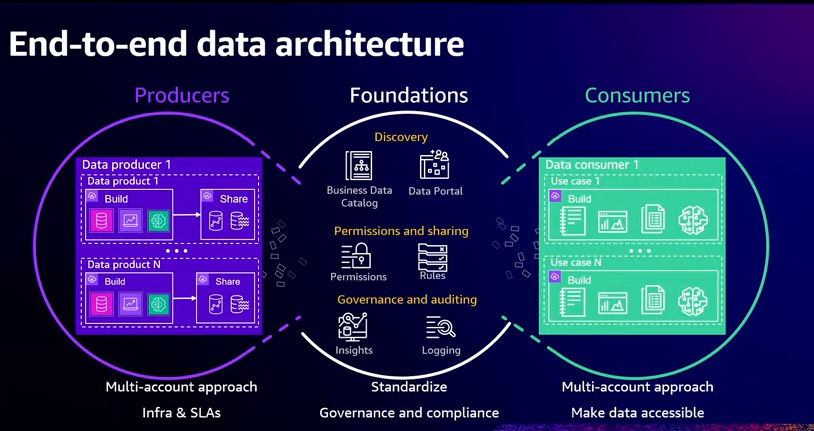
* **Data Consumers:** These curated data products are then utilized by data consumers who understand business priorities and use these data products to drive business results.
* **all the sharing has to be governed** in order to ensure that organizations comply with applicable regulations.

To implement this strategy, customers often use a multi-account architecture on AWS. Data producers use separate accounts to isolate data products from each other

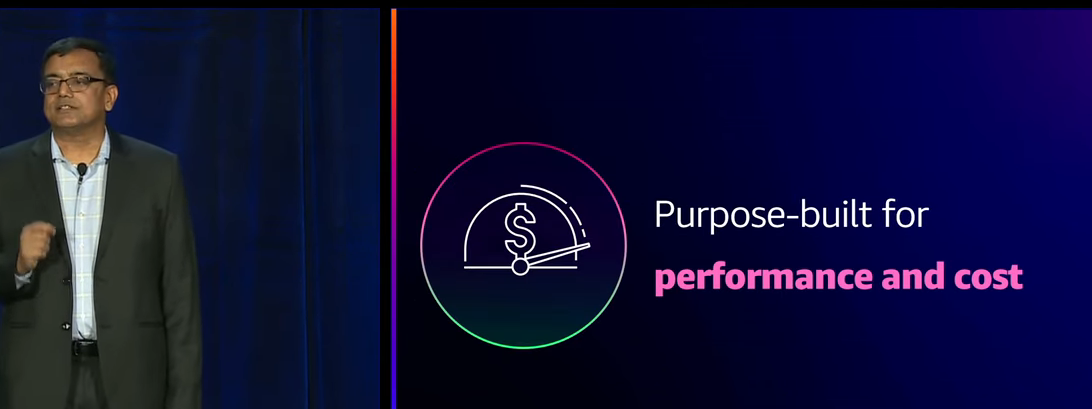
**To facilitate the sharing of these data products** across the organization, **the foundation's team typically provides discovery tools, like a business data catalog**, tools for performance management,

and they are responsible for defining the governance, auditing, and compliance requirements for the entire organization.

**Then consumers can now discover and subscribe to these data assets**



**To help you build your end-to-end data strateg**y, AWS **offers a comprehensive set of purpose-built services for a variety of use case**s, optimized for cost and performanc



* And many of our services support multiple deployment options, so you can get started quickly by using a serverless option
* or you can optimize the cost performance of your workloads by running on pre-provision compute, Kubernetes,spot instances, or reserved instances

**One size does not fit all.**

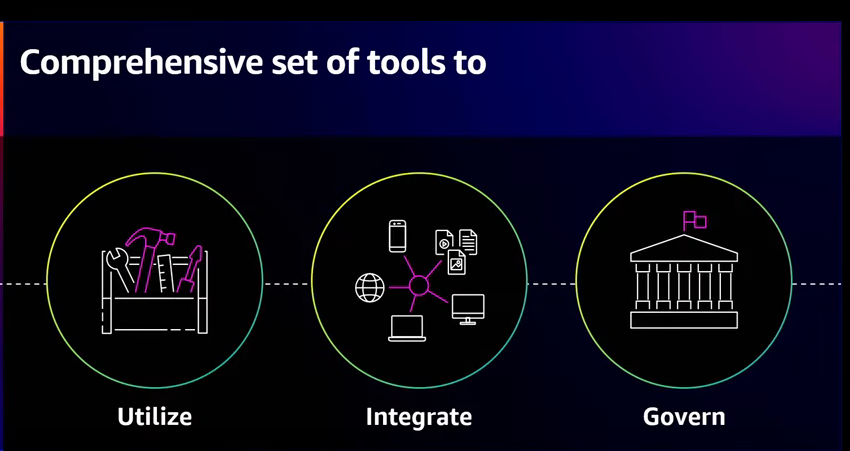
, a database is not appropriate for every use case, just like a data warehouse is not appropriate for every use case, just like even a data lake is not appropriate for every use case.

i**t's common for customers to start with one service or an architectural approach**,

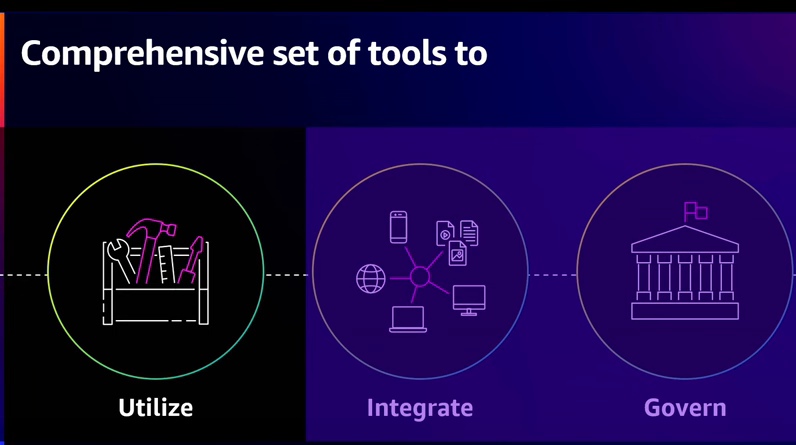
and then as they understand their workload and the usage patterns for that particular application, to switch to another service or another architectural approach, becau**se it is better suited for the task.**

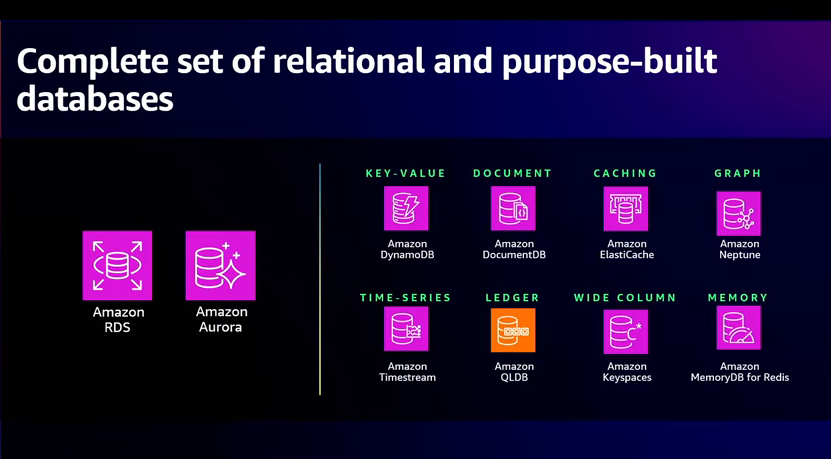
For example, **they may start off with a relational database, because they understand it**, and they can get started with it quickly, but then, switch to a non-relational key value store, because it is a better fit for their use case and allows them to really fine tune the cost performance

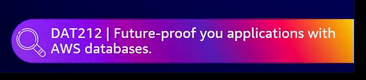
**AWS provides a comprehensive set of services to help them store and utilize data,** to help them integrate the data across their organization, so they have visibility into their entire business and customers and to help them govern their data assets, so they can comply with their regulatory obligations.



**Let's dive in and take a look at some of the services we provide to help you store and utilize your data.**

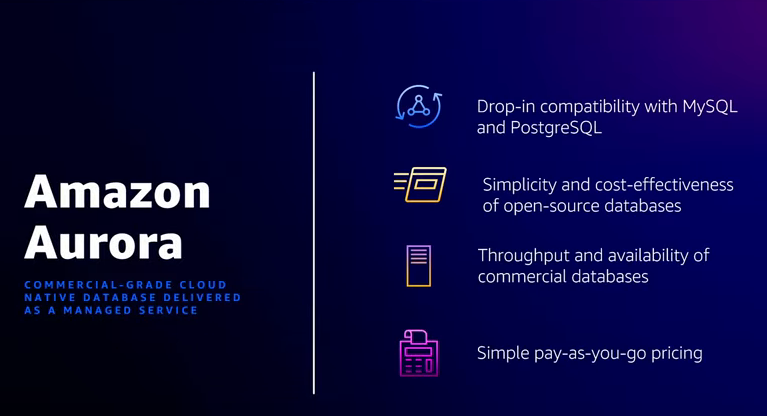








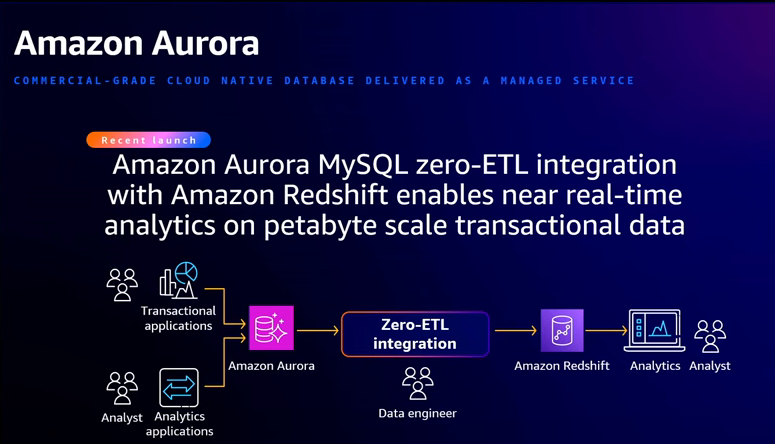




In addition to S3, our durable object store, we offer the industry's most complete set of relational databases, such as Aurora and purpose-built databases, like DynamoDB, a scalable key value store, Neptune, a graph database, and Timestream, a database purpose built for your time series data

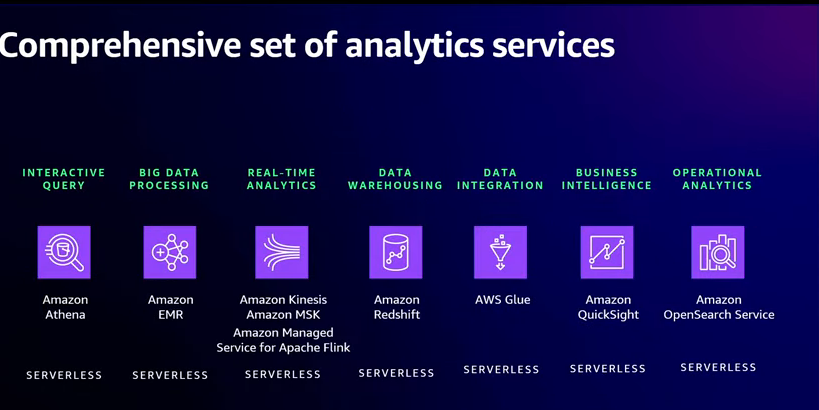
These databases are uniquely designed to provide optimal price performance for their respective use cases, so developers always have the right tool for their job. Aurora is our MySQL and Postgres compatible relational database service designed for unparalleled performance, including scalability, availability, and reliability, all at 1/10th the cost of enterprise commercial-grade databases.

And one such Aurora launch is the Aurora MySQL to Redshift Zer**o ETL integration, which seamlessly replicates data in Aurora into Redshift in seconds**



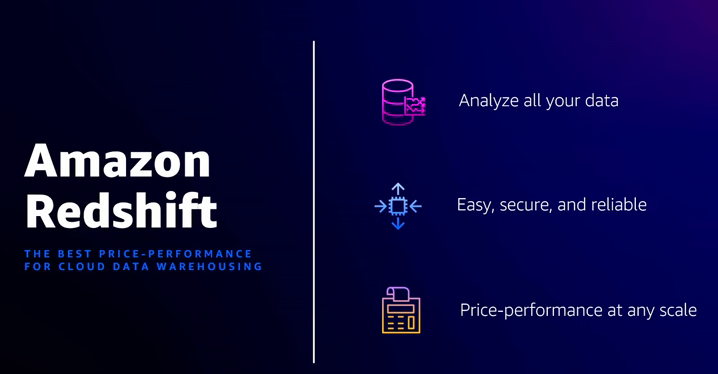
, so customers can use Redshift for near-real time analytics on petabyte scale data. And the best part is you don't have to do anything to set up this data integration pipeline. You simply tell us the tables you want to replicate into Redshift, and we take care of everything seamlessly in the background for you.

**In addition to the database services, AWS offers a comprehensive set of analytic services**, and many of these services offer a serverless option, so you can get started with them quickly to build your applications.

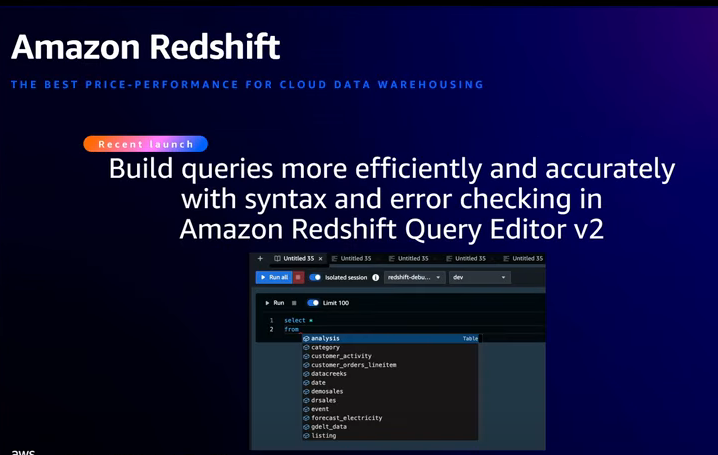


. **For data warehousing, we offer Redshift.** Now the beauty of Redshift isn't the scale of data it can operate on with consistently high performance while keeping your costs predictable.

. For data warehousing, we offer Redshift. Now the beauty of Redshift isn't the scale of data it can operate on with consistently high performance while keeping your costs predictable.



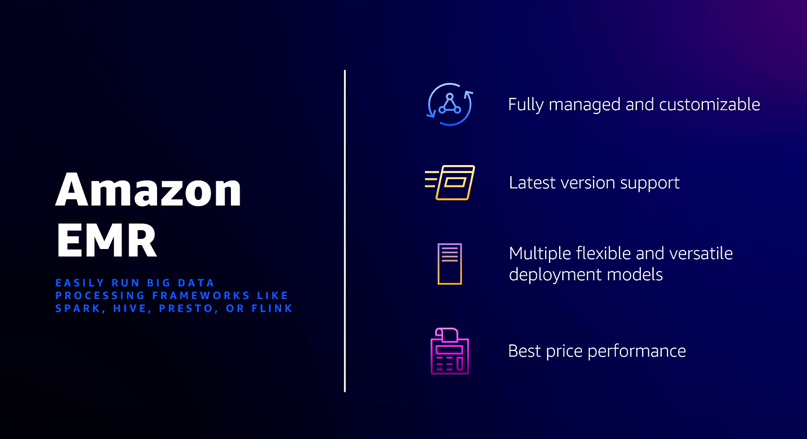
In addition, **with RedShift's federated query capabilities**, customers can now query their operational data stores, like Aurora or RDS. And as data becomes more democratized within organizations, Redshift is delivering on easy analytics for everyone.



as data becomes more democratized within organizations, **Redshift is delivering on easy analytics for everyone. A** recent launch in this area I wanna highlight is the auto complete and syntax highlighting feature Redshift Query editor v2. This feature enables less technical users in an organization to build analytics and queries more efficiently and accurately

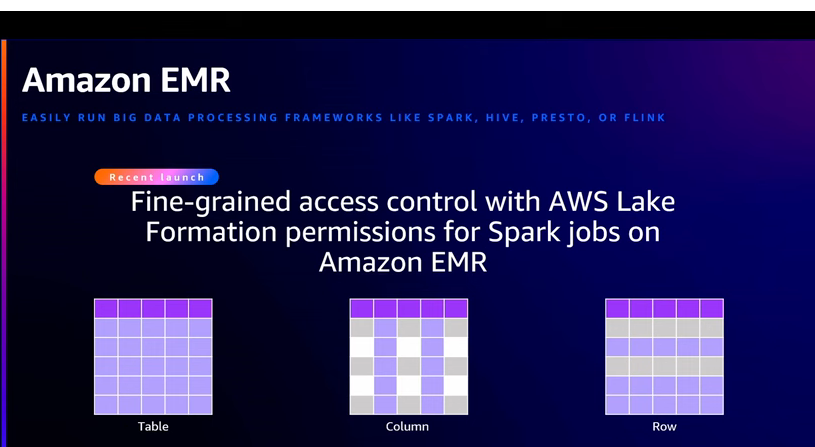
**Embedding machine learning and generative AI capabilities i**nto our services to empower all users is a theme you'll

**For big data processing,** we offer **EMR**. EMR makes it easy to run big data processing frameworks, like Spark, Hive, Presto, or Flink.



It supports the latest versions of these open source frameworks within 90 days, and it provides the best performance at lowest cost.

In fact, Spark workloads run five times faster than open source. In addition, EMR has flexible deployment options, like serverless, Kubernetes or running on provision compute. And by running on spot or reserved instances



**For business intelligence and dashboarding, we offer QuickSight**



we offer QuickSight, which allows everyone in your organization to understand your data by asking questions in natural language using QuickSight Queue, explore your data through interactive dashboards or look (audio breaks) in your data powered by machine learning.



QuickSight enables business intelligence for everybody. And with QuickSight-embedded dashboards, these insights can now be embedded directly into your operational tools and internal portals, so your users have access to the data that they need with the appropriate business context

to support your interactive log analytics, real-time application monitoring, and website use cases, we offer OpenSearch Service

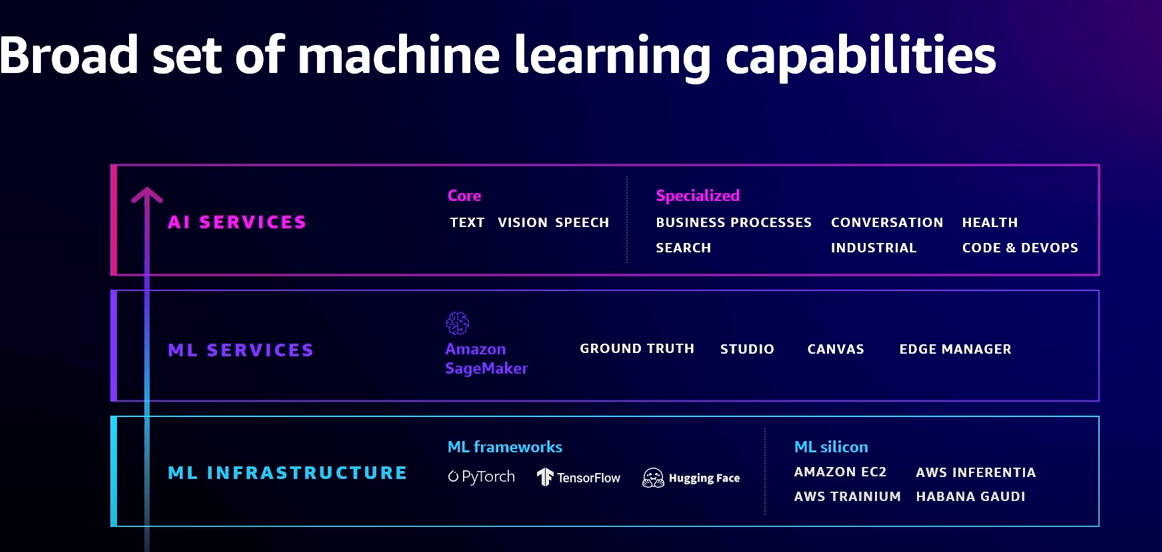


Customers use OpenSearch, for example, for log analytics, so they can detect potential threats and respond to changes in system state

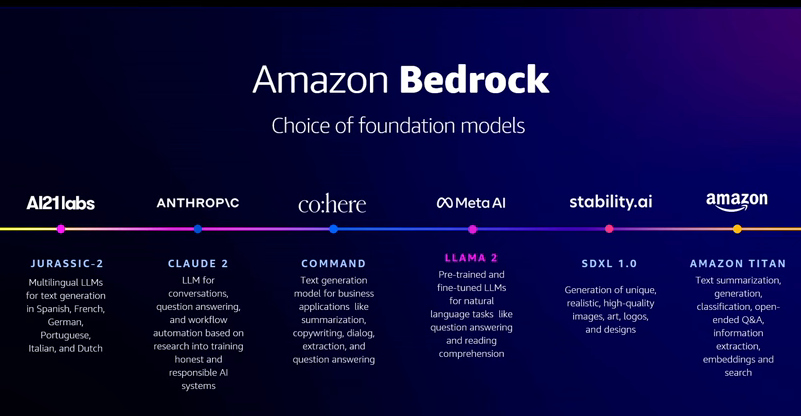


you can use different storage tiers to optimize the cost performance of your workloads.

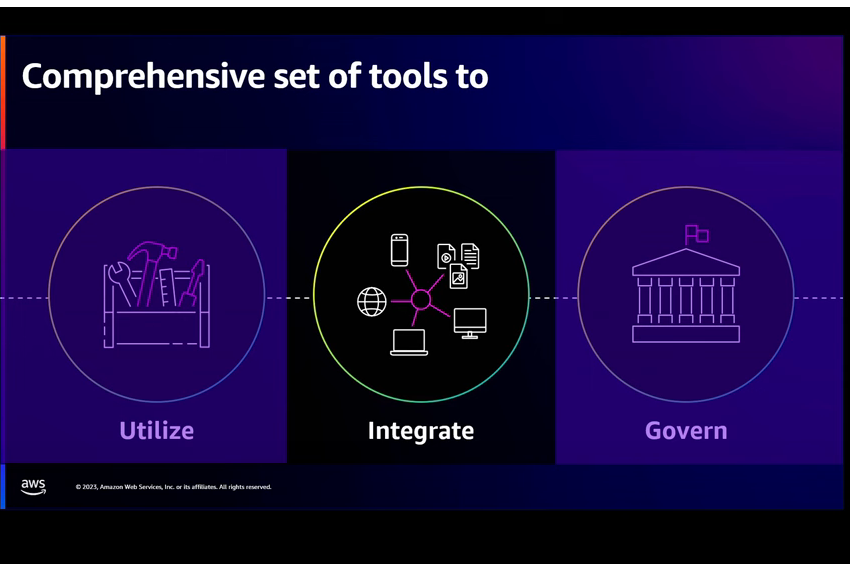
For example, you can keep your highest priority workloads on hot storage for fast performance while moving the data for your lower priority workloads to cold storage in order to optimize costs. With all the interest in generative AI, a recent OpenSearch feature that I want to highlight is the Vector Engine for OpenSearch Server Serverless.





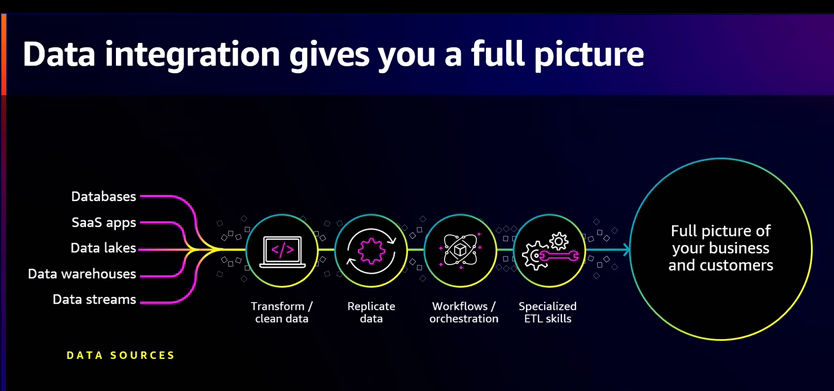


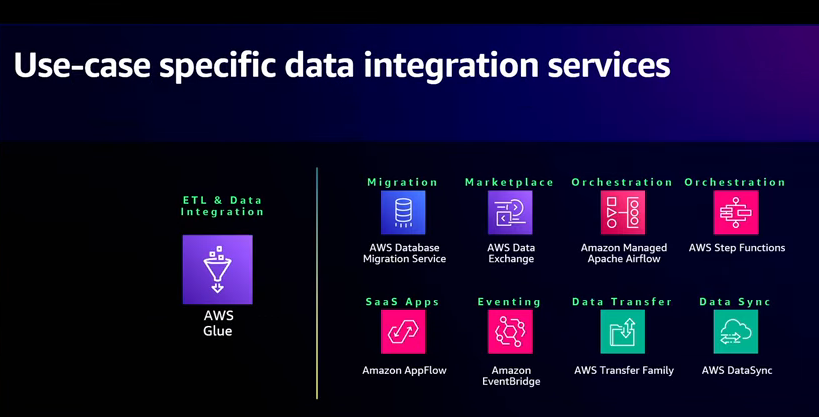
So far, we've reviewed the services we provide to help you store and utilize the data inside your organization. Now, let's dig into the services we provide to help you integrate your data

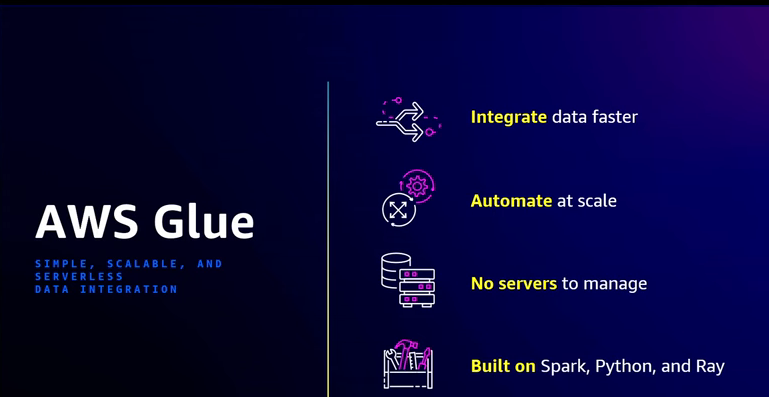


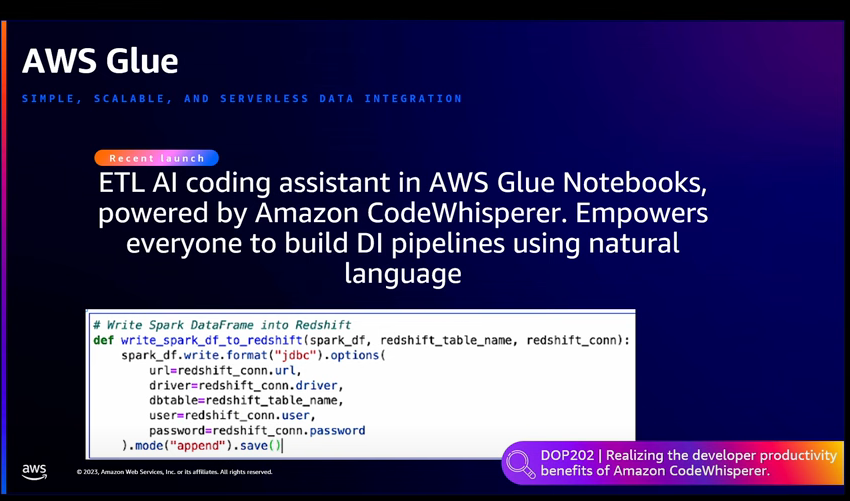
because your data lives in disparate databases, data warehouses, data lakes, and SaaS systems in your organization.

(18:37) And you need to integrate this data together to create a holistic picture of your business and customers. Often, this involves developing code to clean and transform your data, developing code to replicate or move data between systems, as well as orchestrating end-to-end workflows









, we offer use-case specific data integration services. For example, to help you And so to help you prepare your migrate data from open source and commercial databases into your data lakes and data warehouses, we offered a database migration service

Glue. Glue is our anchor data integration service, and it features a serverless execution engine that can scale to support all your workloads. Glue supports all the users inside your organization with persona-specific authoring tools, Notebooks for your technical users, Data Brew for users that want an Excel-style wrangling interface and Glue Studio for users that want a visual job authoring experience

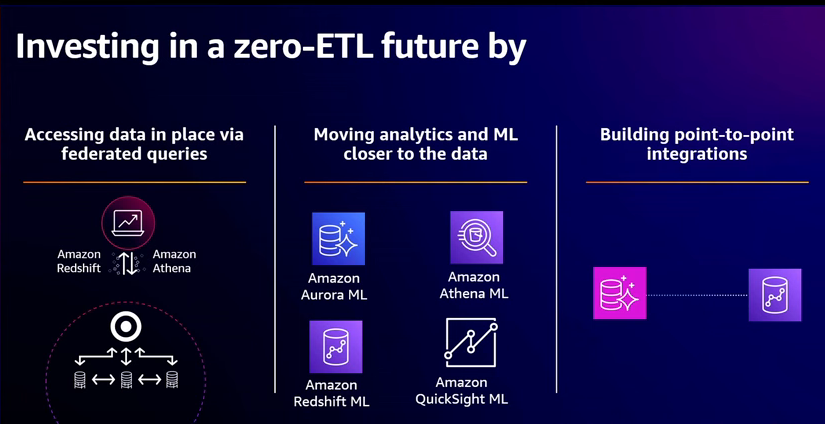
ETL AI Coding Assistant powered by CodeWhisperer. Using this feature, users can build data integration pipelines using natural language input. In this example, just by putting in that comment, "Write data frame into Redshift," CodeWhisperer will give you the Spark code for writing a Spark data frame into Redshift.

While tools like this will simplify the development of data integration pipelines, in many cases, we can do better, and that's why we're investing in a Zero ETL future. Zero ETL eliminates the need for ETL pipelines that you were previously building and managing by hand

And in many ways, this is similar to how we use the term serverless.

Serverless doesn't mean that there are no servers. It means that we, AWS, removed the undifferentiated heavy lifting of provisioning, monitoring, and de-provisioning servers, so you could focus on your business logic.





as we progress on the Zero ETL journey, we will evolve our offerings to remove much of the undifferentiated heavy lifting of building, monitoring, and managing data integration pipelines, so you can focus on your business-specific transformations.

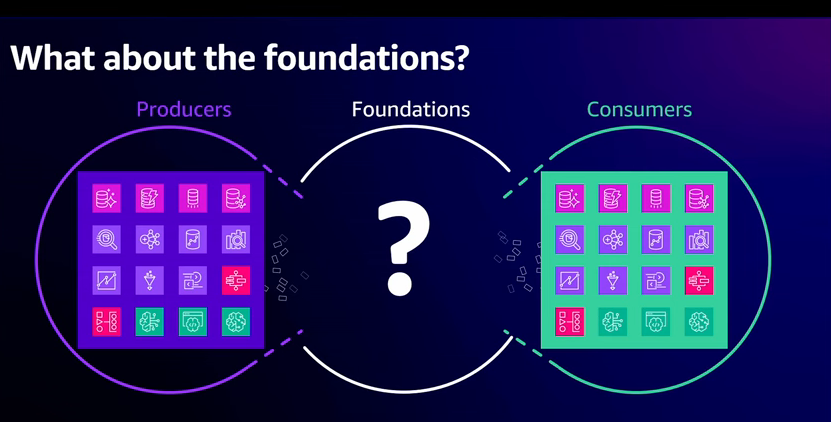
And in practice, what this means is that we will invest in making it easier for you to access data in place by expanding the federated query capabilities in Athena and Redshift.

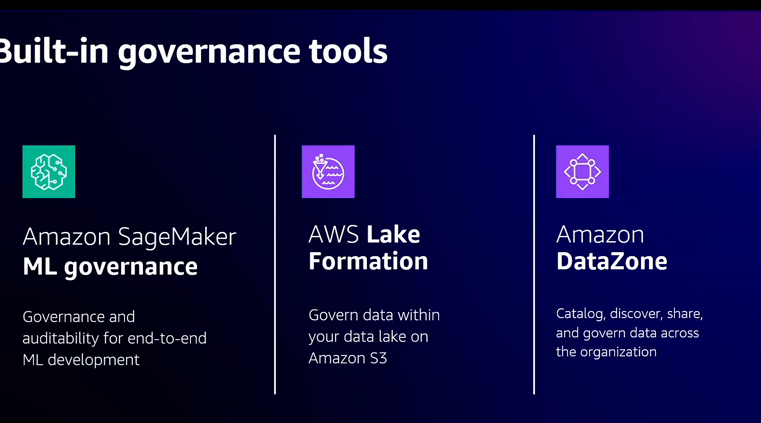
So instead of building a pipeline to replicate data from your operational data stores into your data warehouses or data lakes, so you can query them, there's no need to move the data, you can query the data directly in place

we'll build point-to-point integrations between our services, so we take care of the undifferentiated heavy lifting of moving and replicating data between our services, so you don't have to. This is similar to the Aurora MySQL to Redshift Zero ETL integration that I covered previously.

. So far, we've talked in detail about the tools we provide to help data producers create curated data sets, and we've talked in great detail about the tools we provide to help data consumers utilize these data sets to create interactive dashboards, data integration pipelines and so forth.But without the governance foundations in place, sharing data across your organization will be a difficult task.

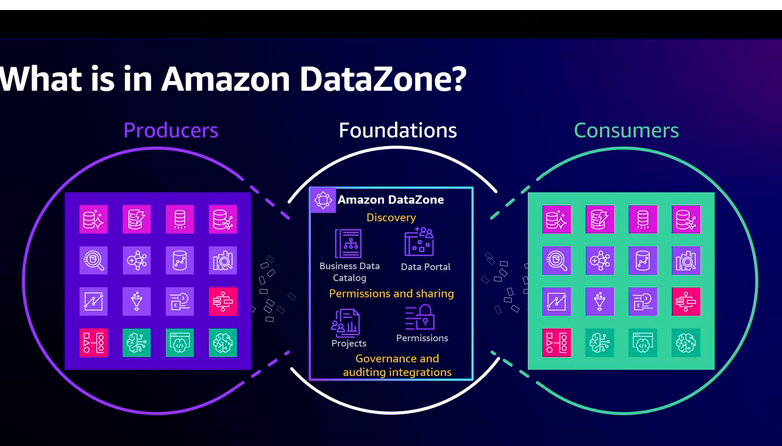
But without the governance foundations in place, sharing data across your organization will be a difficult task





SageMaker helps you address common machine learning challenges from onboarding new users to centralizing model information for multiple users in a single location. To govern your data lakes, we provide Lake Formation which helps you easily build, govern, and audit your S3-based data lakes and it features table, column, and role level permissions, so you can share the right data with the right users inside your organization.

(24:29) And for true end-to-end data governance across your entire organization and all of our services, we offer DataZone. DataZone provides the key components required to share data products across your organization. It includes a organization-wide business data catalog where data producers can publish data assets.



Data consumers can then discover these data assets and request permissions to access them. And once that subscription has been approved or the permissions have been granted, DataZone sets up all the permissions including cross account permissions for you seamlessly in the background, so your consumers can just access the data.