CONCEPTUAL GUIDE

Connecting Amazon DataZone to other AWS services

Table of contents

Amazon DataZone unlocks data across organizational boundaries

Amazon DataZone is a data management service that makes it faster and easier for you to catalog, discover, share, and govern data stored across Amazon Web Services (AWS), on-premises, and third-party sources. Amazon DataZone allows engineers, data scientists, product managers, analysts, and business users to discover, use, and collaborate on data throughout an organization.

With Amazon DataZone, administrators who oversee an organization's data assets can manage and govern access using fine-grained controls. These controls help ensure users receive the right level of privileges and context.

Because Amazon DataZone integrates with AWS services, you can directly deliver data to end users and simplify your architecture.

Read through this guide to learn:

- How Amazon DataZone uses popular AWS services you may already have in your environment, including Amazon Redshift, Amazon Athena, Amazon SageMaker, AWS Glue, and AWS Lake Formation, as well as on-premises and third-party sources
- Why and how to connect to data lakes, AWS Glue Data Catalog, and Software-as-a-Service (SaaS) applications
- Why to connect to Amazon Athena and Amazon Redshift for analytics analysis
- How to connect to identity providers like AWS Identity and Access Management

Benefits of Amazon DataZone

Govern data access across organizational boundaries. Ensure that the right data is accessed by the right user for the right purpose, in accordance with your organization's security regulations, without relying on individual credentials.

Connect data and people through shared data and tools to drive business insights. Increase your team's efficiency with seamless collaboration and self-service access to data and analytics tools.

Automate data discovery and cataloging with machine learning. Reduce the time spent on manual entry of data attributes into the business data catalog.

3 ways Amazon DataZone integrates with other AWS services

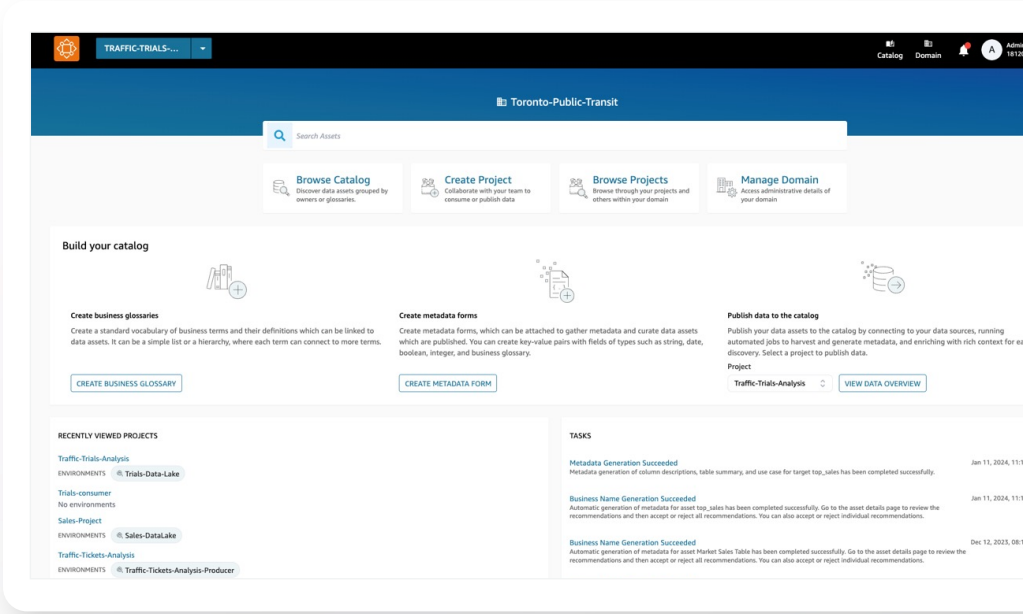
Historically, organizations struggle to use various analytics capabilities together in a consistently governed way. With Amazon DataZone, you have one unified data portal that supports the end-to-end analytics lifecycle of cataloging, discovering, sharing, governing, and analyzing. Amazon DataZone integrates with analytics capabilities like Amazon S3–based data lakes, Amazon Redshift–based data warehouses, Amazon Athena, AWS Glue Data Catalog, and AWS Lake Formation so you don't have to build one-off integrations for the AWS ecosystem.

Amazon DataZone supports three types of integrations with other AWS services.

Producer data sources: Publish data assets to the Amazon DataZone catalog from the data stored in AWS Glue Data Catalog and Amazon Redshift tables and views. You can also manually publish objects from Amazon Simple Storage Service (Amazon S3) to the Amazon DataZone catalog or even custom assets to establish the business workflow for all types of assets across your organization.

Consumer tools: Use the tool of your choice to connect to data that you have been approved to use. Amazon DataZone provides an integrated experience with Amazon Athena and Amazon Redshift query editors. This experience will be expanded to other AWS services and third-party services with customizations.

Access control and fulfillment: Amazon DataZone supports granting access to AWS Lake Formation (including hybrid setup), managed AWS Glue tables, and Amazon Redshift tables and views. For all other data assets, Amazon DataZone publishes standard events related to your actions (such as approval given for a subscription request) to Amazon EventBridge. You can use these standard events to integrate with other AWS services or third-party solutions for custom integrations and report back to Amazon DataZone, so you only need to work from one portal.



PRODUCER

Amazon S3 data lakes and AWS Glue Data Catalog

To use Amazon DataZone to catalog your data, you must first import your project data (assets) as inventory to Amazon DataZone. Creating inventory for a particular project makes the assets discoverable only to that project's members. You can add assets to the project inventory in the following ways:

- Create and run data sources via the data portal or by using the Amazon DataZone APIs for AWS Glue and Amazon Redshift.
- Create assets from the available system asset types (AWS Glue, Amazon Redshift, Amazon S3 objects) or from your custom asset types.
- Manually create assets for S3 objects or custom assets using the Amazon DataZone data portal.

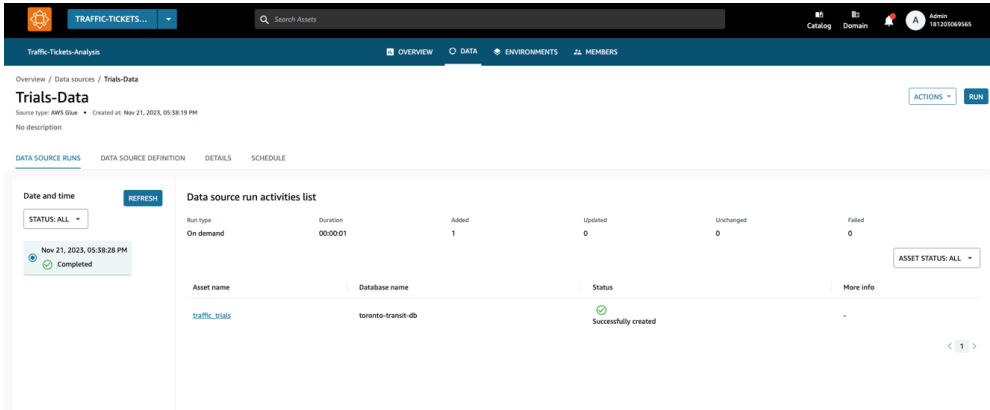
AWS Glue Data Catalog

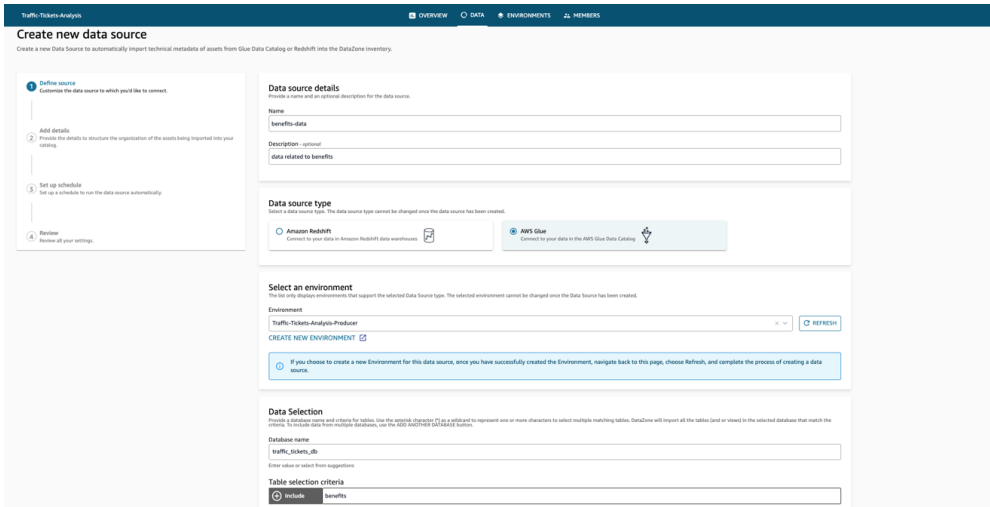
You can create an AWS Glue Data Catalog data source to import technical metadata of database tables from AWS Glue. When you create and run an AWS Glue data source, you add assets from the source AWS Glue database to your Amazon DataZone project's inventory. You can run your AWS Glue data sources on a set schedule or on demand to create or update your assets' technical metadata.

High level steps to add an AWS Glue data source

1. Sign in to the Amazon DataZone data portal
2. Select your project and create a data source
3. Choose an AWS Glue data source type
4. Specify an environment in which to publish the AWS Glue tables
5. Provide an AWS Glue database and enter your table selection criteria
6. For publishing settings, choose whether assets are immediately discoverable in the business data catalog
7. Automatically generate business names and business descriptions (using Amazon Bedrock large language models (LLMs) generative AI capabilities) to aid in discovery
8. Create

After approval, Amazon DataZone will automatically add these assets to all the existing data lake environments in the project. Amazon DataZone then grants and manages access to the approved AWS Glue Data Catalog tables through AWS Lake Formation. For the subscriber project, assets that are granted access appear in the AWS Glue Data Catalog (subscribing database of your environment) as resources in your account. You can then use Amazon Athena to query the tables.





PRODUCER

Amazon AppFlow

Amazon AppFlow is a fully managed integration service that enables you to securely transfer data between SaaS applications like Salesforce, SAP, Google Analytics, Facebook Ads, and ServiceNow, as well as AWS services like Amazon S3 and Amazon Redshift.

High level steps to transfer data from SaaS applications into Amazon DataZone

1. While data is being moved, Amazon AppFlow automates the preparation and registration of your SaaS data into the AWS Glue Data Catalog.
2. After the data is cataloged in AWS Glue Data Catalog, you can set up Amazon DataZone to ingest the technical schema to enrich and publish so other users can discover those assets using Amazon DataZone's data portal.

PRODUCER

Amazon EventBridge

Amazon EventBridge provides a simple and consistent way to ingest, filter, transform, and deliver events so you can respond to operational changes in your applications. Amazon DataZone uses Amazon EventBridge to send events for watched activities. You can capture an event and then build customizations to the workflow needed. The experience for the data consumer is seamless and they can always rely on the status shown in the data portal.

High level steps to handle access grants for subscriptions of unmanaged assets

*This pertains to assets that are not set up to be managed using AWS Lake Formation.

1. Build an automation to listen to those events and then kick off a grant for that asset.
2. Programmatically update the subscription status as granted.

PRODUCER

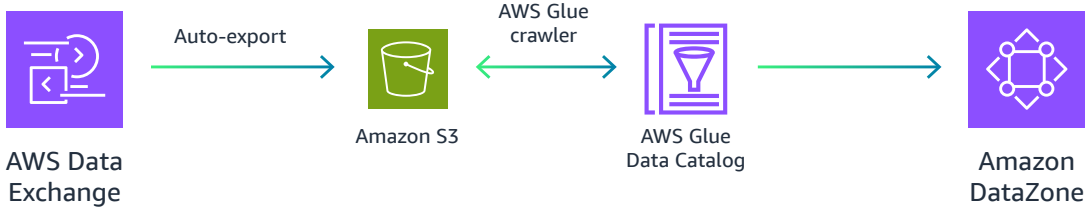
AWS Data Exchange

Govern first- and third-party data together with AWS Data Exchange and Amazon DataZone. With AWS Data Exchange, you have access to a large, public catalog of ready-to-use data covering over 3,500 datasets.

You can integrate the third-party data you subscribe to through AWS Data Exchange into Amazon DataZone. After publishing these third-party data assets, consumers in your domain can use it just like the first-party data within your organization. AWS Data Exchange auto-exports to Amazon S3. Amazon DataZone discovers the AWS Data Exchange subscribed asset through AWS Glue Data Catalog, which uses AWS Glue crawler to discover assets in Amazon S3.

Amazon DataZone: AWS Data Exchange

Govern first- and third-party data together



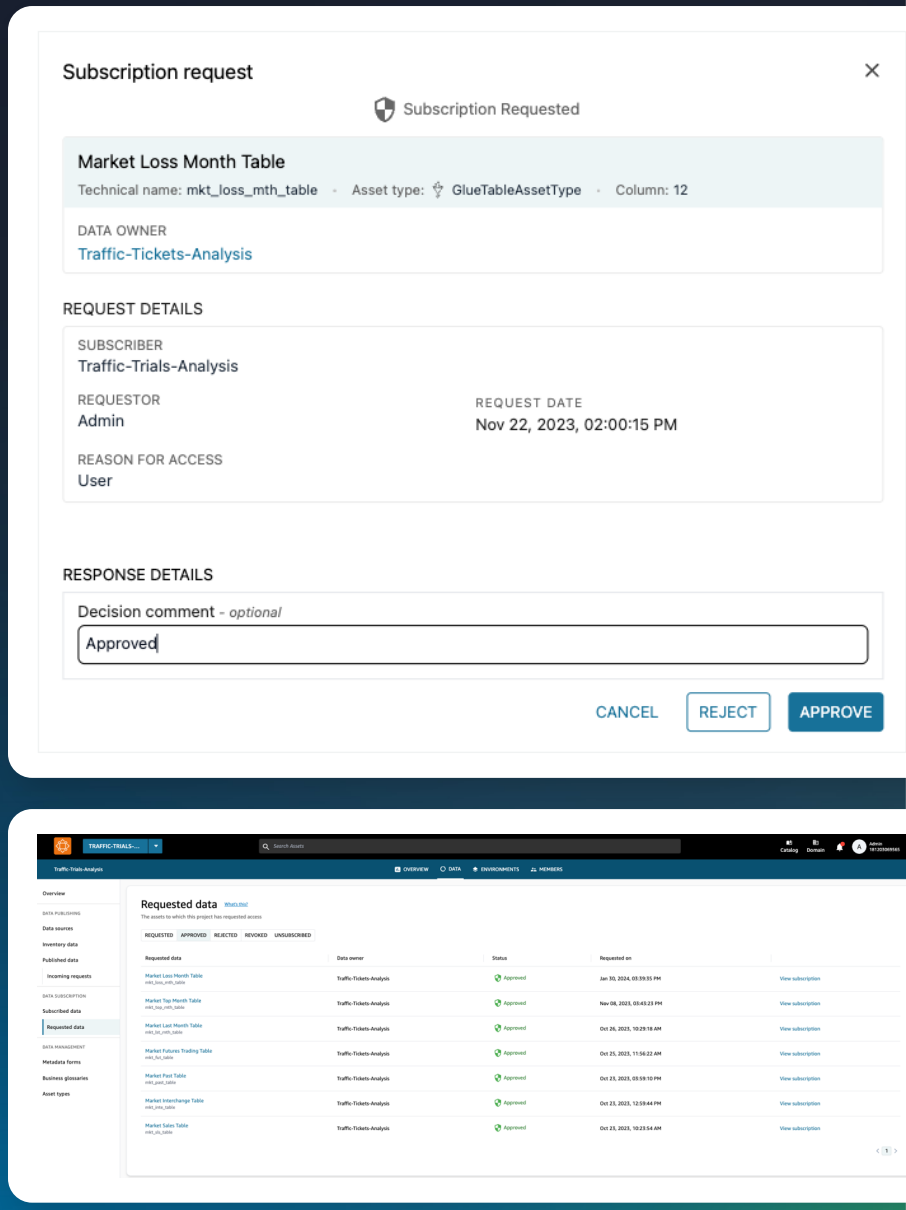
ACCESS CONTROL

AWS Lake Formation

Amazon DataZone abstracts the process of sharing data between data producers and consumers through AWS Lake Formation. Amazon DataZone automates this typically manual process. For Amazon DataZone–managed assets, fulfillment of data access to the underlying tables according to the policies applied by data publishers is taken care of without the need for an admin or for data movement.

Amazon DataZone creates and manages identity and access management (IAM) roles for both producers and subscribers. Amazon DataZone assumes these roles to grant or revoke AWS Lake Formation permissions during the process of sharing data.

To manage data assets, Amazon DataZone applies both coarse-grained IAM policies and fine-grained AWS Lake Formation permissions. When sharing data, Amazon DataZone only shares read-only AWS Lake Formation permissions with all consumer personas, ensuring they have access to read the data, but not modify it. For owned data, different personas in the owner project have varying levels of AWS Lake Formation permissions.



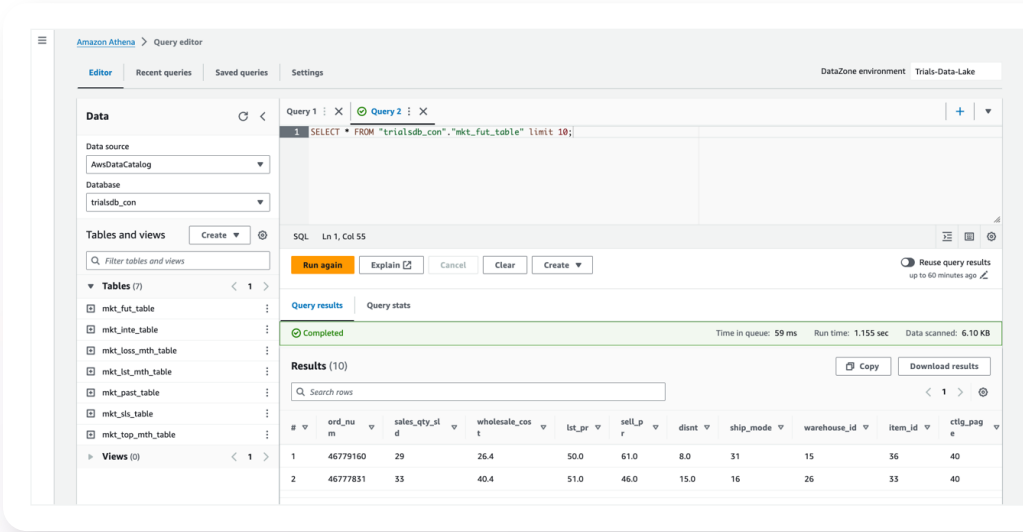
CONSUMER

Amazon Athena

In Amazon DataZone, once a subscriber has access to an asset in the catalog, they can consume it (query and analyze) using Amazon Athena or Amazon Redshift query editor v2. You must be a subscribing project owner or contributor to complete this task. Depending on the blueprints enabled in the project, Amazon DataZone provides links to Amazon Athena and/or Amazon Redshift query editor v2 in the data portal.

High level steps to query Amazon DataZone data with Amazon Athena query editor

1. Sign in to the Amazon DataZone data portal.
2. Select your project with the data you want to analyze. If the Data Lake blueprint is enabled on this project, a link to Amazon Athena is displayed. If the Data Warehouse blueprint is enabled on this project, a link to the query editor is displayed.
3. Choose Amazon Athena to open the Amazon Athena query editor using the project's credentials for authentication.
4. In the Amazon Athena query editor, write and run your queries. Some common tasks include:
    - Query and analyze your subscribed assets
    - Create new tables
    - Create a table from query results (CTAS) from an external S3 bucket



CONSUMER

Amazon Redshift

In Amazon DataZone, once a subscriber has access to an asset in the catalog, they can consume it (query and analyze) using Amazon Athena or Amazon Redshift query editor v2. Any Amazon Redshift tables or views that you have subscribed to are linked to the Amazon Redshift cluster or Amazon Redshift Serverless workgroup that is configured for the environment. You can subscribe to the tables and views as well as publish any new tables and views that you create in your environment's cluster or database.

High level steps to query Amazon DataZone data with Amazon Redshift query editor

1. Sign in to the Amazon DataZone data portal.
2. Select your project with the data you want to analyze. If the Data Warehouse blueprint is enabled on this project, a link to the query editor is displayed.
3. Choose Amazon Redshift.
4. Specify connection details, depending on whether you are using an Amazon Redshift Serverless workgroup or an Amazon Redshift cluster.
5. In the Amazon Redshift query editor v2, run and write your queries.

CONSUMER

Amazon SageMaker

In an Amazon DataZone project, a data consumer sets up an Amazon SageMaker environment. All users or machine learning (ML) builders added to the project can deep-link to the Amazon SageMaker domain from this environment. Next, in the SageMaker domain, ML builders can access Assets to search for data or ML resources and learn more about those resources with descriptions, glossary terms, metadata forms, schemas, revision histories, created and modified data, published versions, and more. ML builders can then request to subscribe to the resource, which also allows data governance teams visibility into resource usage. After access is granted, ML builders can access subscribed resources in Amazon SageMaker Studio (for model package groups and feature groups), Amazon SageMaker Canvas (for data assets), and Amazon SageMaker notebooks (for data assets). Now ML builders can be ready to consume these resources for their analytics workloads and generate new resources.

ACCESS CONTROL

AWS IAM Identity Center

AWS IAM Identity Center helps you securely create or connect your workforce identities and manage their access centrally across AWS accounts and applications. IAM Identity Center is the recommended approach for workforce authentication and authorization on AWS for organizations of any size and type. Using IAM Identity Center, you can create and manage user identities in AWS, or connect your existing identity source, including Microsoft Active Directory, Okta, Ping Identity, JumpCloud, Google Workspace, and Microsoft Entra ID (formerly Azure AD).

You can access the Amazon DataZone data portal by using either your single sign-on (SSO) credentials or AWS credentials. If you already have AWS IAM Identity Center enabled and configured in the same AWS Region where you want to create your Amazon DataZone domain, configuration is simplified.

High level steps to enable AWS IAM Identity Center

1. Sign in to the AWS Management Console with the credentials of your AWS Organizations management account. You can't enable IAM Identity Center while signed in with credentials from an AWS Organizations member account.
2. Open the AWS IAM Identity Center console and choose the AWS Region in which you want to create your Amazon DataZone domain.
3. Enable and choose your identity source. By default, you get an IAM Identity Center store for quick and easy user management. (You can connect an external identity provider instead.)
4. Create a group in IAM Identity Center, add users, and send an email to the user with password setup instructions. The user should get an email about the next setup steps.
5. Choose the group and add users. Users should receive an email inviting them to use SSO.

After you create your Amazon DataZone domain, you can enable AWS IAM Identity Center for Amazon DataZone and provide access to your SSO users and SSO groups.

ACCESS CONTROL

AWS IAM roles

In Amazon DataZone, projects enable a group of users to collaborate on various business use cases that involve publishing, discovering, subscribing to, and consuming data in the Amazon DataZone catalog. Project members consume assets from the Amazon DataZone catalog and produce new assets using one or more analytical workflows. Projects support the following activities within the data portal:

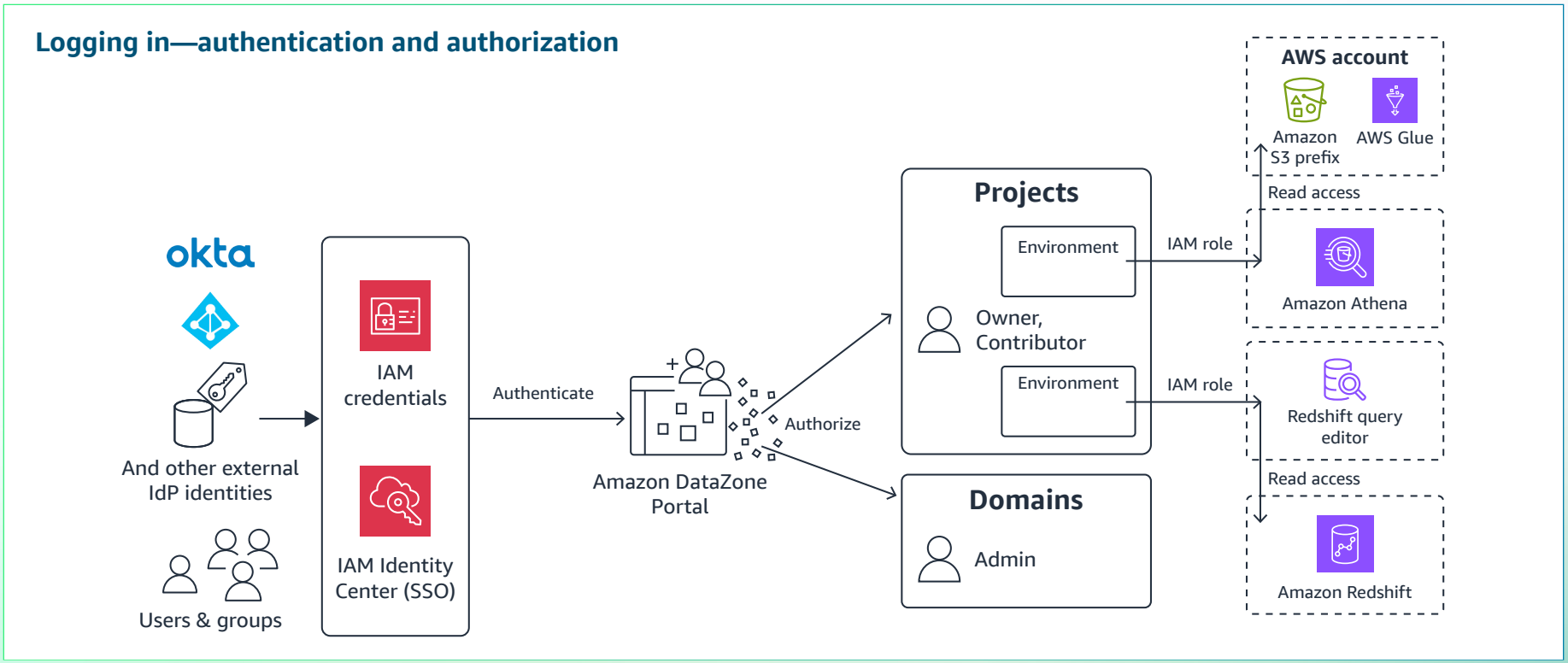
- Project owners can add members with owner and contributor permissions
- Project members can be SSO users, SSO groups, and IAM users
- Project members can request subscription to the assets in the data catalog

Subscription approvals are provided to the projects. In an Amazon DataZone project, environments are collections of zero or more configured resources (for example, an Amazon S3, an AWS Glue database, or an Amazon Athena workgroup), with a given set of IAM principals who can operate on those resources. Environments are created by using environment profiles that are pre-configured sets of resources and blueprints that provide reusable templates for creating environments. Environment profiles define settings such as the AWS account or VPC in which environments are deployed.

You need AWS IAM to complete the following security-related tasks:

- Create users and groups under your AWS account.
- Assign unique security credentials to each user under your AWS account.
- Control each user's permissions to perform tasks with AWS resources.
- Allow the users in another AWS account to share your AWS resources.
- Create roles for your AWS account and define the users or services that can assume them.
- Use existing identities for your enterprise to grant permissions to perform tasks using AWS resources.

There are specific IAM roles for Amazon DataZone outlined in the documentation.



Go deeper with how-to videos

With Amazon DataZone, you can make the most of your organization's data. For more information on how to connect Amazon DataZone to other AWS or third-party services, check out these how-to videos. Follow the Amazon DataZone resources page to keep up with the latest announcements.