Implementation Guide

Data Transfer Hub

Data Transfer Hub: Implementation Guide

Table of Contents

Framework for secure, scalable, and trackable data transfer for Amazon Simple Storage Service (Amazon S3) objects and Amazon Elastic Container Registry (Amazon ECR) images

Publication date: *December 2021 (last update: November 2024)*

The Data Transfer Hub Guidance provides secure, scalable, and trackable data transfer for Amazon Simple Storage Service (Amazon S3) objects and Amazon Elastic Container Registry (Amazon ECR) images. This data transfer helps you easily create and manage different types (Amazon S3 object and Amazon ECR image) of transfer tasks between AWS partitions (for example, aws, aws-cn, aws-us-gov), and from other cloud providers to AWS at your own discretion.

If you have enabled the Direct Connect service in a specific AWS Region and a specific AWS China Region and have purchased a compliant cross-border dedicated line provided by a qualified operator to connect the AWS Region and their own VPC in the AWS China Region, you can use Data Transfer Hub's console to create a data transfer task and choose to use a dedicated line for data transfer at your own discretion.

This implementation guide describes an overview of the Data Transfer Hub Guidance, its reference architecture and components, considerations for planning the deployment, and configuration steps for deploying Data Transfer Hub in the AWS Cloud. It also includes some tutorials with prescriptive guidance for using Data Transfer Hub.

Use this navigation table to quickly find answers to these questions:

If you want to ...	Read...
Know the cost for running this Guidance	Cost
Understand the security considerations for this Guidance	Security
Know how to plan for quotas for this Guidance	Quotas

If you want to ...	Read...
Know which AWS Regions are supported for this Guidance	Supported AWS Regions
View or download the AWS CloudForm ation template included in this Guidance to automatically deploy the infrastructure resources (the “stack”) for this Guidance	AWS CloudFormation template

This guide is intended for IT architects, developers, DevOps, data analysts, and marketing technology professionals who have practical experience architecting in the AWS Cloud.

You will be responsible for your compliance with all applicable laws in respect of your data transfer tasks.

Features and benefits

Depending on the availability of your network environment, the Guidance supports the following key features:

- **Inter-Partition and Cross-Cloud data transfer** - to promote seamless transfer capabilities in one place
- **Auto scaling** - to allow rapid response to changes in file transfer traffic
- **High performance of large file transfer (1TB)** - to leverage the strengths of clustering, parallel large file slicing, and automatic retries to robust file transfer
- **Monitoring** - to track data flow, diagnose issues, and ensure the overall health of the data transfer processes
- **Out-of-the-box deployment**

> **Note**
>
> If you want to transfer Amazon S3 objects between AWS Regions, we recommend that you use Cross-Region Replication. If you want to transfer Amazon S3 objects within the same AWS Region, we recommend using Same-Region Replication.

For data transfer between AWS China Region and AWS Region, you will be responsible for your compliance with all applicable laws and regulations on cross-border data transfer (including purchasing compliant cross-border dedicated lines provided by qualified operators for data transfer, performing necessary government approval or filing), and shall initiate data transfer at your own discretion. AWS does not assist you with this data transfer.

Use cases

Today, the China market is one of biggest markets in the world. Many international companies are seeking their success in China, as well as a number of Chinese companies are expanding their businesses globally. One of most important steps of in the business is moving their data.

S3 Cross-Region Replication and ECR Cross-Region Replication are popular but customers cannot use them to replicate data into China Regions. With the launch of Data Transfer Hub Guidance, customers can now create S3 and ECR data transfer tasks between AWS Regions and AWS China Regions in a web portal. Moreover, it supports replicating data from other cloud providers to AWS.

Depending on the availability of your network environment, Data Transfer Hub supports the following use cases:

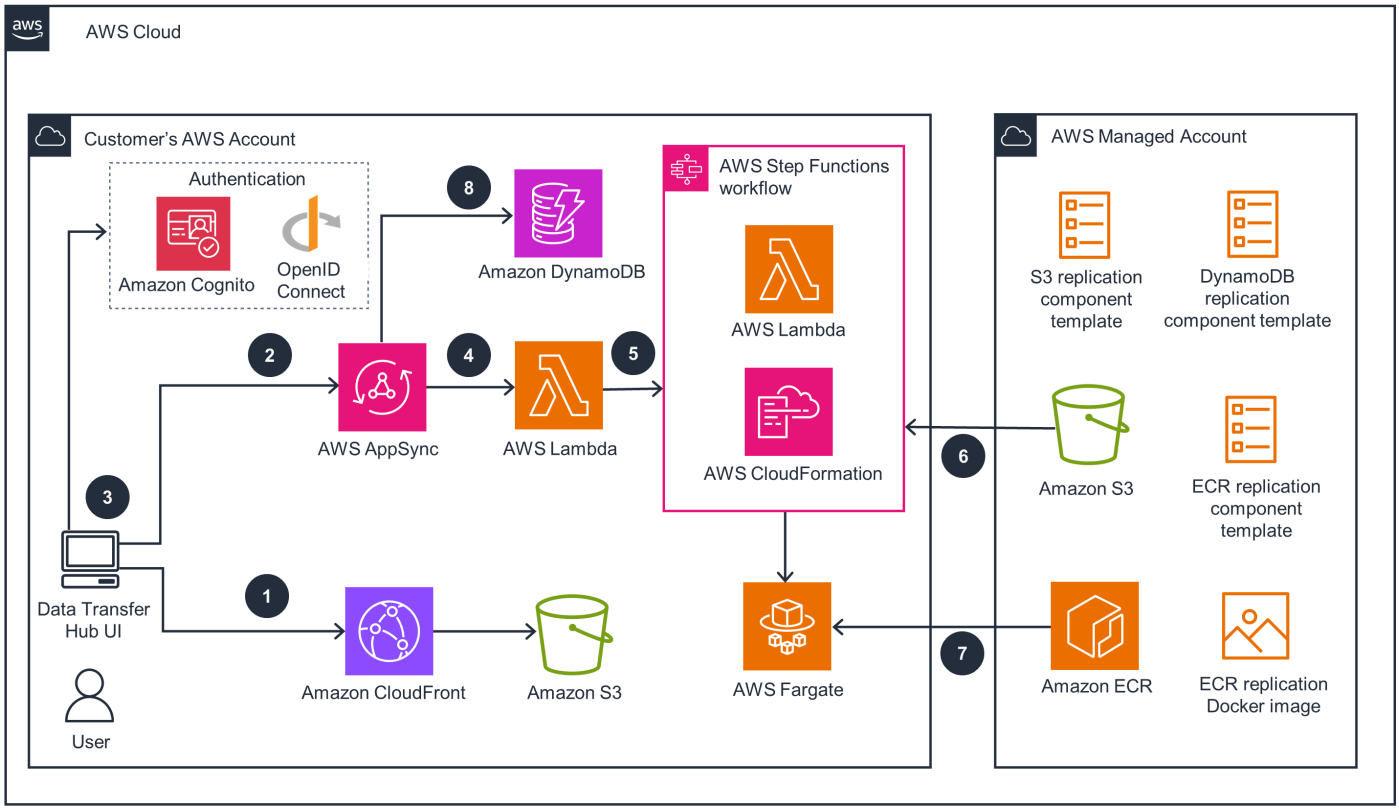
- Copy Amazon S3 objects between AWS Regions and AWS China Regions.
- Copy data from other cloud providers' object storage services to Amazon S3.
- Transfer Amazon ECR images between AWS Regions and AWS China Regions.
- Transfer Dockers image from public docker registry (for example, Docker Hub, Google gcr.io, Red Hat Quay.io) to Amazon ECR.

Architecture overview

This section provides reference implementation architecture diagrams for the components deployed with this Guidance.

Architecture diagram

Deploying the Data Transfer Hub Guidance with the default parameters builds the following environment in the AWS Cloud.



Data Transfer Hub architecture on AWS

The Guidance automatically deploys and configures a serverless architecture with the following services:

1. Amazon Simple Storage Service stores static web assets (such as the frontend UI), which are made available through Amazon CloudFront.
2. AWS AppSync GraphQL provides backend APIs.

3. Users are authenticated by either Amazon Cognito user pools (in AWS Standard Regions) or by an OpenID connect provider (in AWS China Regions) such as Authing, Auth0.
4. AWS AppSync runs AWS Lambda to call backend APIs.
5. Lambda starts an AWS Step Functions workflow that uses AWS CloudFormation to start or stop/delete the Amazon ECR or Amazon S3 plugin template.
6. A centralized S3 bucket hosts plugin templates.
7. The Guidance also provisions an Amazon ECS cluster that runs the container images used by the plugin template, and the container images are hosted in Amazon ECR.
8. Amazon DynamoDB stores data transfer task information.

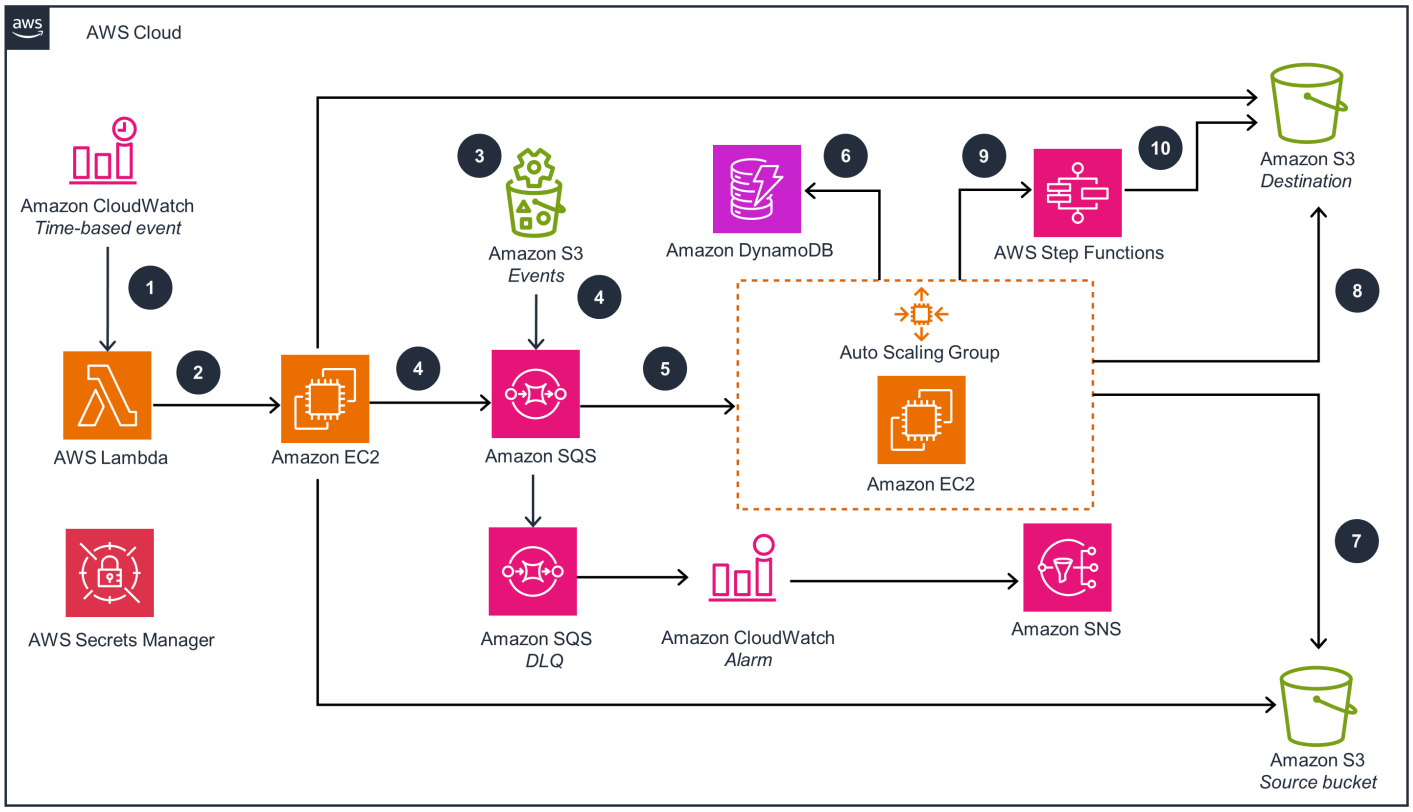
After deploying the Guidance, you can use AWS WAF to protect CloudFront or AppSync.

> **Important**
>
> If you deploy this Guidance in AWS (Beijing) Region operated by Beijing Sinnet Technology Co., Ltd. (Sinnet), or the AWS (Ningxia) Region operated by Ningxia Western Cloud Data Technology Co., Ltd., you are required to provide a domain with ICP Recordal before you can access the web console.

The web console is a centralized place to create and manage all data transfer jobs. Each data type (for example, Amazon S3 or Amazon ECR) is a plugin for Data Transfer Hub, and is packaged as an AWS CloudFormation template hosted in an S3 bucket that AWS owns. When the you create a transfer task, an AWS Lambda function initiates the Amazon CloudFormation template, and state of each task is stored and displayed in the DynamoDB tables.

As of this revision, the Guidance supports two data transfer plugins: an Amazon S3 plugin and an Amazon ECR plugin.

Amazon S3 plugin



Data Transfer Hub Amazon S3 plugin architecture

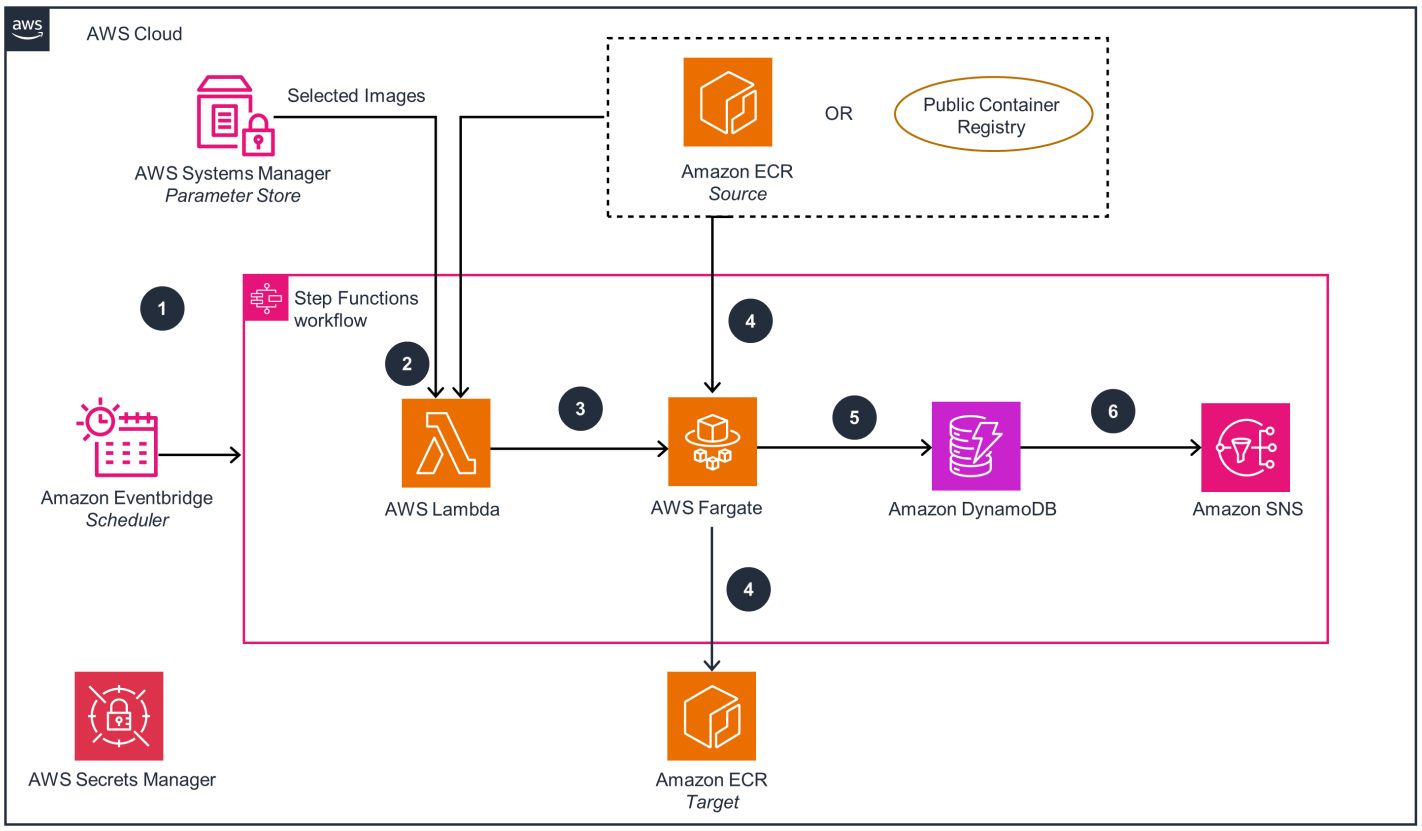
The Amazon S3 plugin runs the following workflows:

1. A time-based EventBridge rule initiates the AWS Lambda function on an hourly basis.
2. AWS Lambda uses the launch template to launch a data comparison job (JobFinder) in an Amazon Elastic Compute Cloud (Amazon EC2).
3. The job lists all the objects in the source and destination Amazon S3 buckets and makes comparisons among objects to determine which objects should be transferred.
4. Amazon EC2 sends a message for each object that will be transferred to Amazon Simple Queue Service (Amazon SQS). Amazon S3 event messages can also be supported for more real-time data transfer. Whenever there is object uploaded to source bucket, the event message is sent to the same Amazon SQS queue.
5. A JobWorker node running in Amazon EC2 consumes the messages in Amazon SQS and transfers the object from the source bucket to the destination bucket. You can use an Auto Scaling group to control the number of Amazon EC2 instances to transfer the data based on business needs.
6. DynamoDB stores a record with transfer status for each object.

7. The Amazon EC2 instance will get (download) the object from the source bucket based on the Amazon SQS message.
8. The Amazon EC2 instance will put (upload) the object to the destination bucket based on the Amazon SQS message.
9. When the JobWorker node identifies a large file (with a default threshold of 1 GB) for the first time, a Multipart Upload task running in Amazon EC2 is initiated. The corresponding UploadId is then conveyed to the AWS Step Functions, which invokes a scheduled recurring task. Every minute, AWS Step Functions verifies the successful transmission of the distributed shards associated with the UploadId across the entire cluster.
10. If all shards have been transmitted successfully, Amazon EC2 invokes the CompleteMultipartUpload API in Amazon S3 to finalize the consolidation of the shards. Otherwise, any invalid shards are discarded.

> **Note**
>
> If an object (or part of an object) transfer failed, the JobWorker releases the message in the queue, and the object is transferred again after the message is visible in the queue (default visibility timeout is set to 15 minutes). If the transfer failed five times, the message is sent to the dead letter queue and a notification alarm is initiated.

Amazon ECR plugin



Data Transfer Hub Amazon ECR plugin architecture

The Amazon ECR plugin runs the following workflows:

1. An Amazon EventBridge rule runs an AWS Step Functions workflow on a regular basis (by default, it runs daily).
2. Step Functions invokes AWS Lambda to retrieve the list of images from the source.
3. Lambda will either list all the repository content in the source Amazon ECR, or get the stored image list from Parameter Store, a capability of AWS System Manager.
4. The transfer task will run within AWS FARGATE; in a maximum concurrency of 10. If a transfer task failed for some reason, it will automatically retry three times.
5. Each task uses skopeo to copy the images into the target Amazon ECR registry.
6. After the copy completes, the status (either success or fail) is logged into DynamoDB for tracking purposes.

AWS Well-Architected pillars

This Guidance was designed with best practices from the AWS Well-Architected Framework which helps customers design and operate reliable, secure, efficient, and cost-effective workloads in the cloud.

Operational excellence

This section describes how the principles and best practices of the operational excellence pillar were applied when designing this Guidance.

The Data Transfer Hub Guidance pushes metrics to Amazon CloudWatch at various stages to provide observability into the infrastructure, Lambda functions, Amazon EC2 transfer workers, Step Function workflow, and the rest of the Guidance components. Data transferring errors are added to the Amazon SQS queue for retries and alerts.

Security

This section describes how the principles and best practices of the security pillar were applied when designing this Guidance.

- Data Transfer Hub web console users are authenticated and authorized with Amazon Cognito.
- All inter-service communications use AWS IAM roles.
- All roles used by the Guidance follows least-privilege access. That is, it only contains minimum permissions required so the service can function properly.

Reliability

This section describes how the principles and best practices of the reliability pillar were applied when designing this Guidance.

- Using AWS serverless services wherever possible (for example, Lambda, Step Functions, Amazon S3, and Amazon SQS) to ensure high availability and recovery from service failure.
- Data is stored in DynamoDB and Amazon S3, so it persists in multiple Availability Zones (AZs) by default.

Performance efficiency

This section describes how the principles and best practices of the performance efficiency pillar were applied when designing this Guidance.

- The ability to launch this Guidance in any Region that supports AWS services in this Guidance such as: AWS Lambda, AWS S3, Amazon SQS, Amazon DynamoDB, and Amazon EC2.
- Automatically testing and deploying this GUidance daily. Reviewing this Guidance by solution architects and subject matter experts for areas to experiment and improve.

Cost optimization

This section describes how the principles and best practices of the cost optimization pillar were applied when designing this Guidance.

- Use Autoscaling Group so that the compute costs are only related to how much data is transferred.
- Using serverless services such as Amazon SQS and DynamoDB so that customers only get charged for what they use.

Sustainability

This section describes how the principles and best practices of the sustainability pillar were applied when designing this Guidance.

- The Guidance's serverless design (using Lambda, Amazon SQS and DynamoDB) and the use of managed services (such as Amazon EC2) are aimed at reducing carbon footprint compared to the footprint of continually operating on-premises servers.

Architecture details

This section describes the components and AWS services that make up this Guidance and the architecture details on how these components work together.

AWS services in this Guidance

The following AWS services are included in this Guidance:

AWS service	Description
Amazon CloudFront	**Core.** To made available the static web assets (frontend user interface).
AWS AppSync	**Core.** To provide the backend APIs.
AWS Lambda	**Core.** To call backend APIs.
Amazon ECS	**Core.** To run the container images used by the plugin template.
Amazon DynamoDB	**Core.** To store a record with transfer status for each object.
Amazon EC2	**Core.** To consume the messages in Amazon SQS and transfer the object from the source bucket to the destination bucket.
AWS Secrets Manager	**Core.** Stores the credential for data transfer.
AWS Step Functions	**Supporting.** To start or stop/delete the ECR or S3 plugin template.
Amazon Cognito	**Supporting.** To authenticate users (in AWS Regions).
Amazon S3	**Supporting.** To store the static web assets (frontend user interface).

AWS service	Description
Amazon ECR	**Supporting.** To host the container images.
Amazon SQS	**Supporting.** To store the transfer tasks temporarily as a buffer.
Amazon EventBridge	**Supporting.** To invoke the transfer tasks regularly.
Amazon SNS	**Supporting.** Provides topic and email subscription notifications for data transfer results.
Amazon CloudWatch	**Supporting.** To monitor the data transfer progress.

How Data Transfer Hub works

This Guidance has three components: a web console, the Amazon S3 transfer engine, and the Amazon ECR transfer engine.

Web console

This Guidance provides a simple web console, which allows you to create and manage transfer tasks for Amazon S3 and Amazon ECR.

Amazon S3 transfer engine

Amazon S3 transfer engine runs the Amazon S3 plugin and is used for transferring objects from their sources into S3 buckets. The S3 plugin supports the following features:

- Transfer Amazon S3 objects between AWS China Regions and AWS Regions
- Transfer objects from other cloud providers to Amazon S3
- Transfer objects from S3 Compatible Storage service to Amazon S3
- Support near real time transfer via S3 Event
- Support transfer with object metadata

- Support incremental data transfer
- Support transfer from private payer request bucket
- Auto retry and error handling

Amazon ECR transfer engine

Amazon ECR engine runs the Amazon ECR plugin and is used for transferring container images from other container registries. The ECR plugin supports the following features:

- Transfer Amazon ECR images between AWS China Regions and AWS Regions
- Transfer from public container registry (such as Docker Hub, GCR.io, Quay.io) to Amazon ECR
- Transfer selected images to Amazon ECR
- Transfer all images and tags from Amazon ECR

The ECR plugin leverages skopeo for the underlying engine. The AWS Lambda function lists images in their sources and uses Fargate to run the transfer jobs.

Plan your deployment

This section describes the Region, the section called “Cost”, the section called “Security”, and the section called “Quotas ” considerations for planning your deployment.

Supported AWS Regions

This Guidance uses services which may not be currently available in all AWS Regions. Launch this Guidance in an AWS Region where required services are available. For the most current availability by Region, refer to the AWS Regional Services List.

Supported Regions for deployment in AWS Regions

Region ID	Region Name
us-east-1	US East (N. Virginia)
us-east-2	US East (Ohio)
us-west-1	US West (N. California)
us-west-2	US West (Oregon)
ap-south-1	Asia Pacific (Mumbai)
ap-northeast-2	Asia Pacific (Seoul)
ap-southeast-1	Asia Pacific (Singapore)
ap-southeast-2	Asia Pacific (Sydney)
ap-southeast-4	Asia Pacific (Melbourne)
ap-northeast-1	Asia Pacific (Tokyo)
ca-central-1	Canada (Central)
ca-west-1	Canada (Calgary)
eu-central-1	Europe (Frankfurt)

Region ID	Region Name
eu-west-1	Europe (Ireland)
eu-west-2	Europe (London)
eu-west-3	Europe (Paris)
eu-north-1	Europe (Stockholm)
sa-east-1	South America (São Paulo)
il-central-1	Israel (Tel Aviv)

Supported Regions for deployment in AWS China Regions

Region ID	Region Name
cn-north-1	China (Beijing) Region Operated by Sinnet
cn-northwest-1	China (Ningxia) Region Operated by NWCD

Cost

You are responsible for the cost of the AWS services used while running this Guidance, which can vary based on whether you are transferring Amazon S3 objects or Amazon ECR images.

The Guidance automatically deploys an additional Amazon CloudFront Distribution and an Amazon S3 bucket for storing the static website assets in your account. You are responsible for the incurred variable charges from these services. For full details, refer to the pricing webpage for each AWS service you will be using in this Guidance.

The following examples demonstrate how to estimate the cost. Two example estimates are for transferring Amazon S3 objects, and one is for transferring ECR images.

Cost of an Amazon S3 transfer task

For an Amazon S3 transfer task, the cost can vary based on the total number of files and the average file size.

Example 1: As of this revision, transfer 1 TB of S3 files from AWS Oregon Region (`us-west-2`) to AWS Beijing Region (`cn-north-1`), and the average file size is **50MB**.

Total files: ~20,480

Average speed per Amazon EC2 instance: ~1GB/min

Total Amazon EC2 instance hours: ~17 hours

AWS service	Dimensions	Cost
Amazon EC2	$0.0084 per hour (t4g.micro)	$0.14
Amazon S3	~ 12 GET requests + 10 PUT request per file GET: $0.0004 per 1000 request PUT: $0.005 per 1000 request	$0.12
Amazon DynamoDB	~2 write requests per file $1.25 per million write	$0.05
Amazon SQS	~2 request per file $0.40 per million requests	$0.01
Data Transfer Out	$0.09 per GB	$92.16
Others (For example, CloudWatch, Secrets Manager, etc.)		~ $1
	TOTAL	~ $94.48

Example 2: As of this revision, transfer 1 TB of S3 files from AWS Oregon region (`us-west-2`) to Mainland China Beijing Region (`cn-north-1`), and the average file size is **10KB**.

Total files: ~107,400,000

Average speed per Amazon EC2 instance: ~6MB/min (~10 files per sec)

Total Amazon EC2 instance hours: ~3000 hours

AWS Service	Dimensions	Cost
Amazon EC2	$0.0084 per hour (t4g.micro)	$25.20
Amazon S3	~ 2 GET requests + 1 PUT request per file GET: $0.0004 per 1000 request PUT: $0.005 per 1000 request	$622.34
Amazon DynamoDB	~2 write requests per file $1.25 per million write	$268.25
Amazon SQS	~2 request per file $0.40 per million requests	$85.92
Data Transfer Out	$0.09 per GB	$92.16
Others (For example, CloudWatch, Secrets Manager, etc.)		$20
	TOTAL	~ $1,113.87

Cost of an Amazon ECR transfer task

For an Amazon ECR transfer task, the cost can vary based on network speed and total size of ECR images.

Example 3: As of this revision, transfer 27 Amazon ECR images (~3 GB in total size) from AWS Ireland Region (`eu-west-1`) to AWS Beijing Region (`cn-north-1`). The total runtime is about 6 minutes.

AWS Service	Dimensions	Cost
AWS Lambda	$0.0000004 per 100ms	$0.000072 (35221.95 ms)
AWS Step Functions	$0.000025 per state transitio n (~ 60 state transitions per run in this case)	$0.0015
Fargate	$0.04048 per vCPU per hour $0.004445 per GB per hour (0.5 vCPU 1GB Memory)	$0.015 (~ 2200s)
Data Transfer Out	$0.09 per GB	$0.27
Others (for example, CloudWatch, Secrets Manager, etc.)	Almost 0	$0
	TOTAL	~ $0.287

Security

When you build systems on AWS infrastructure, security responsibilities are shared between you and AWS. This shared responsibility model reduces your operational burden because AWS operates, manages, and controls the components including the host operating system, the virtualization layer, and the physical security of the facilities in which the services operate. For more information about AWS security, see AWS Cloud Security.

IAM roles

AWS Identity and Access Management (IAM) roles allow customers to assign granular access policies and permissions to services and users on the AWS Cloud. This Guidance creates IAM roles

that grant the Guidance's AWS Lambda functions, Amazon API Gateway, and Amazon Cognito access to create regional resources.

Amazon CloudFront

This Guidance deploys a web console hosted in an Amazon S3 bucket. To help reduce latency and improve security, this Guidance includes an Amazon CloudFront distribution with an origin access identity, which is a CloudFront user that provides public access to the Guidance's website bucket contents. For more information, refer to Restricting Access to Amazon S3 Content by Using an Origin Access Identity in the *Amazon CloudFront Developer Guide*.

Quotas

Quotas for AWS services in this Guidance

Make sure you have sufficient quota for each of the services implemented in this Guidance. For more information, see AWS service quotas.

Choose one of the following links to go to the page for that service. To view the service quotas for all AWS services in the documentation without switching pages, view the information in the Service endpoints and quotas page in the PDF instead.

AWS CloudFormation quotas

Your AWS account has AWS CloudFormation quotas that you should be aware of when launching the stack in this Guidance. By understanding these quotas, you can avoid limitation errors that would prevent you from deploying this Guidance successfully. For more information, refer to AWS CloudFormation quotas in the *AWS CloudFormation User Guide*.

Deploy the Guidance

This Guidance uses AWS CloudFormation templates and stacks to automate its deployment. The CloudFormation template describes the AWS resources included in this Guidance and their properties. The CloudFormation stack provisions the resources that are described in the templates.

Deployment overview

Use the following steps to deploy this Guidance on AWS. For detailed instructions, follow the links for each step.

Before you launch the Guidance, review the cost, architecture, network security, and other considerations discussed in this guide. Follow the step-by-step instructions in this section to configure and deploy the Guidance into your account.

Time to deploy: Approximately 15 minutes

Step 1. Launch the stack

- (Option 1) Deploy the AWS CloudFormation template in AWS Regions
- (Option 2) Deploy the AWS CloudFormation template in AWS China Regions

Step 2. Launch the web console

Step 3. Create a transfer task

AWS CloudFormation template

To automate deployment, this Guidance uses the following AWS CloudFormation templates, which you can download before deployment:

View template

DataTransferHub-cognito.template: Use this template to launch the Guidance and all associated components in **AWS Regions** where Amazon Cognito is available. The default configuration deploys Amazon S3, Amazon CloudFront, AWS AppSync, Amazon DynamoDB, AWS Lambda,

Amazon ECS, and Amazon Cognito, but you can customize the template to meet your specific needs.

View template

DataTransferHub-openid.template: Use this template to launch the Guidance and all associated components in **AWS China Regions** where Amazon Cognito is **not** available. The default configuration deploys The default configuration deploys Amazon S3, Amazon CloudFront, AWS AppSync, Amazon DynamoDB, AWS Lambda, and Amazon ECS, but you can customize the template to meet your specific needs.

Step 1. (Option 1) Launch the stack in AWS Regions

> **Important**
>
> The following deployment instructions apply to AWS Regions only. For deployment in AWS China Regions refer to Option 2.

Deploy the AWS CloudFormation template for Option 1 – AWS Regions

> **Note**
>
> You are responsible for the cost of the AWS services used while running this Guidance. For more details, visit the Cost section in this guide, and refer to the pricing webpage for each AWS service used in this Guidance.

1. Sign in to the AWS Management Console and use the button below to launch the `DataTransferHub-cognito.template` AWS CloudFormation template. Alternatively, you can download the template as a starting point for your own implementation.

   Launch solution

2. The template launches in the US East (N. Virginia) Region by default. To launch the Guidance in a different AWS Region, use the Region selector in the console navigation bar.

3. On the **Create stack** page, verify that the correct template URL is in the **Amazon S3 URL** text box and choose **Next**.
4. On the **Specify stack details** page, assign a name to your Guidance stack. For information about naming character limitations, refer to IAM and AWS STS quotas in the *AWS Identity and Access Management User Guide*.
5. Under **Parameters**, review the parameters for this Guidance template and modify them as necessary. This Guidance uses the following default values:

Parameter	Default	Description
AdminEmail	`<Requires input>`	The email of the Admin user.

6. Choose **Next**.
7. On the **Configure Stack Options** page, keep the default values and choose **Next**.
8. On the **Review** page, review and confirm the settings. Check the box acknowledging that the template will create IAM resources.
9. Choose **Create stack** to deploy the stack.

You can view the status of the stack in the AWS CloudFormation console in the **Status** column. You should receive a CREATE_COMPLETE status in approximately 15 minutes.

Step 1. (Option 2) Launch the stack in AWS China Regions

> **Important**
>
> The following deployment instructions apply to AWS China Regions only. For deployment in AWS Regions refer to Option 1.

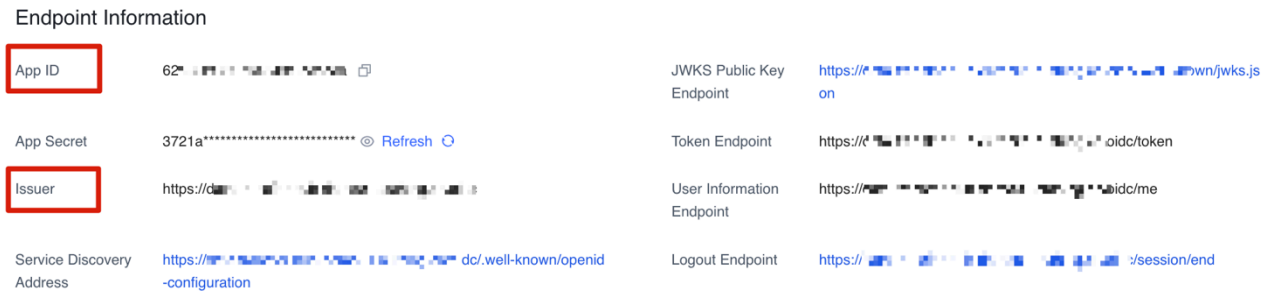
Prerequisites

1. Create an OIDC User Pool.
2. Configure domain name service (DNS) resolution.
3. Make sure a domain registered by ICP is available.

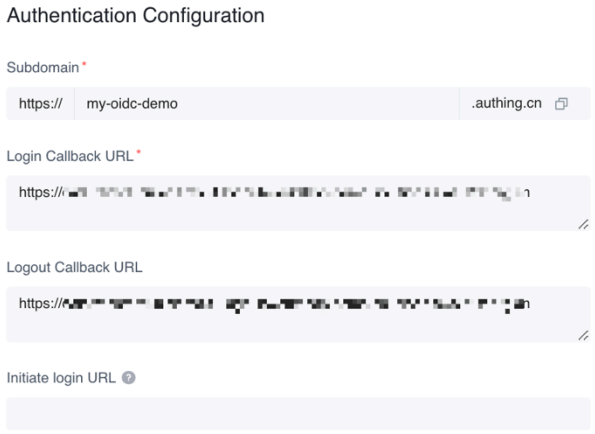
Prerequisite 1: Create an OIDC user pool

In AWS Regions where Amazon Cognito is not yet available, you can use OIDC to provide authentication. The following procedure uses AWS Partner Authing as an example, but you can also choose any available provider.

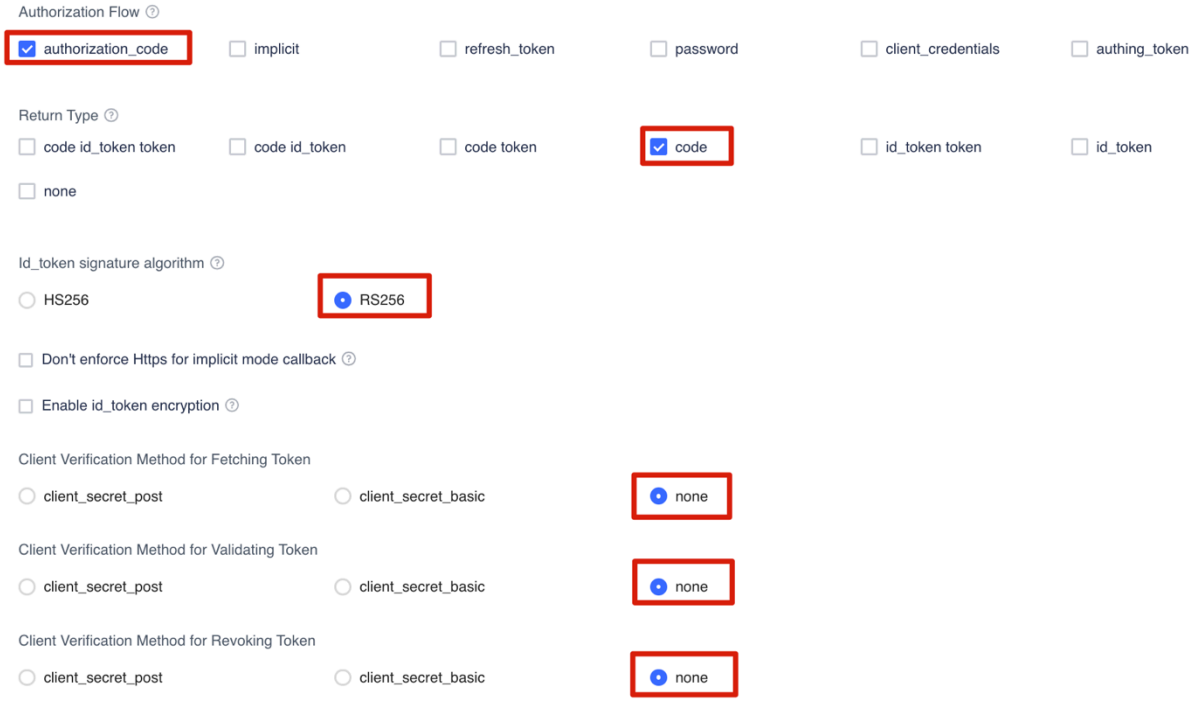
1. Go to the Authing console.
2. Create a new user pool if you don't have one.
3. Select the user pool.
4. On the left navigation bar, select **Self-built App** under **Applications**.
5. Click the **Create** button.
6. Enter the **Application Name**, and **Subdomain**.
7. Save the App ID (that is, `client_id`) and Issuer to a text file from Endpoint Information, which will be used later.



8. Update the **Login Callback URL** and **Logout Callback URL** to your ICP recorded domain name.



9. Set the Authorization Configuration.



10. Update login control.
    a. Select and enter the **Application** interface from the left sidebar, select **Login Control**, and then select **Registration and Login**.
    b. Please select only **Password Login: Email** for the login method.
    c. Please **uncheck** all options in the registration method.
    d. Select **Save**.
11. Create an admin user.
    a. From **User & Roles**, select **Users**, then choose **Create user**.
    b. Enter the email for the user.
    c. Choose **OK**.
    d. Check the email for a temporary password.
    e. Reset the user password.

> **Note**
>
> Because the Guidance does not support application roles, all the users will receive admin rights.

Prerequisite 2: Configure domain name service resolution

Configure domain name service (DNS) resolution to point the ICP licensed domain to the CloudFront default domain name. Optionally, you can use your own DNS resolver.

The following is an example for configuration an Amazon Route 53.

1. Create a hosted zone in Amazon Route 53. For more information refer to the Amazon Route 53 Developer Guide.
2. Create a CNAME record for the console URL.
   a. From the hosted zone, choose **Create Record**.
   b. In the **Record name** input box, enter the host name.
   c. From **Record type** select **CNAME**.
   d. In the value field, enter the CloudFormation output PortalUrl.
   e. Select **Create records**.
3. Add alternative domain names to the CloudFront distribution.
   a. Configure the corresponding domain name in CloudFront to open the CloudFront console by finding the distribution ID for PortalURL in the list and selecting **ID** (or check the check box, and then select **Distribution Settings**).
   b. Edit the distribution and add the Route 53 record to the alternative domain Names (CNAMEs).

Deploy the AWS CloudFormation template for Option 2 – AWS China Regions

This automated AWS CloudFormation template deploys Data Transfer in the AWS Cloud. You must Create an ODIC User Pool and Configure DNS resolution before launching the stack.

> **Note**
>
> You are responsible for the cost of the AWS services used while running this Guidance. For more details, visit the Cost section in this guide, and refer to the pricing webpage for each AWS service used in this Guidance.

1. Sign in to the AWS Management Console and select the button to launch the `DataTransferHub-openid.template` AWS CloudFormation template. Alternatively, you can download the template as a starting point for your own implementation.

2. The template launches in your console's default Region. To launch the Guidance in a different AWS Region, use the Region selector in the console navigation bar.
3. On the **Create stack** page, verify that the correct template URL is in the **Amazon S3 URL** text box and choose **Next**.
4. On the **Specify stack details** page, assign a name to your Guidance stack. For information about naming character limitations, refer to IAM and AWS STS quotas in the *AWS Identity and Access Management User Guide*.
5. Under **Parameters**, review the parameters for this Guidance template and modify them as necessary. This Guidance uses the following default values.

Parameter	Default	Description
OidcProvider	*<Requires input>*	Refers to the Issuer shown in the OIDC application configuration.
OidcClientId	*<Requires input>*	Refers to the App ID shown in the OIDC application configuration.
OidcCustomerDomain	*<Requires input>*	Refers to the customer domain that has completed ICP registration in China, not the subdomain provided by Authing. It must start with `https://`.
AdminEmail	*<Requires input>*	Refers to the email for receiving task status alarm.

6. Choose **Next**.

7. On the **Configure Stack Options** page, keep the default values and choose **Next**.
8. On the **Review** page, review and confirm the settings. Check the box acknowledging that the template will create IAM resources.
9. Choose **Create Stack** to deploy the stack.

You can view the status of your stack in the AWS CloudFormation console in the **Status** column. You should receive a CREATE_COMPLETE status in approximately 15 minutes.

Step 2. Launch the web console

After the stack is successfully created, navigate to the CloudFormation **Outputs** tab and select the **PortalUrl** value to access the Data Transfer Hub web console.

After successful deployment, an email containing a temporary login password will be sent to the email address provided.

Depending on the Region where you start the stack, you can choose to access the web console from the AWS China Regions or the AWS Regions.

- Log in with Amazon Cognito User Pool (for AWS Regions)
- Log in with OpenID using Authing.cn (for AWS China Regions)

(Option 1) Log in using Amazon Cognito user pool for AWS Regions

1. In a web browser, enter the **PortalURL** from the CloudFormation **Output** tab, then navigate to the Amazon Cognito console.
2. Sign in with the **AdminEmail** and the temporary password.
   a. Set a new account password.
   b. (Optional) Verify your email address for account recovery.
3. After the verification is complete, the system opens the Data Transfer Hub web console.

(Option 2) OpenID authentication for AWS China Regions

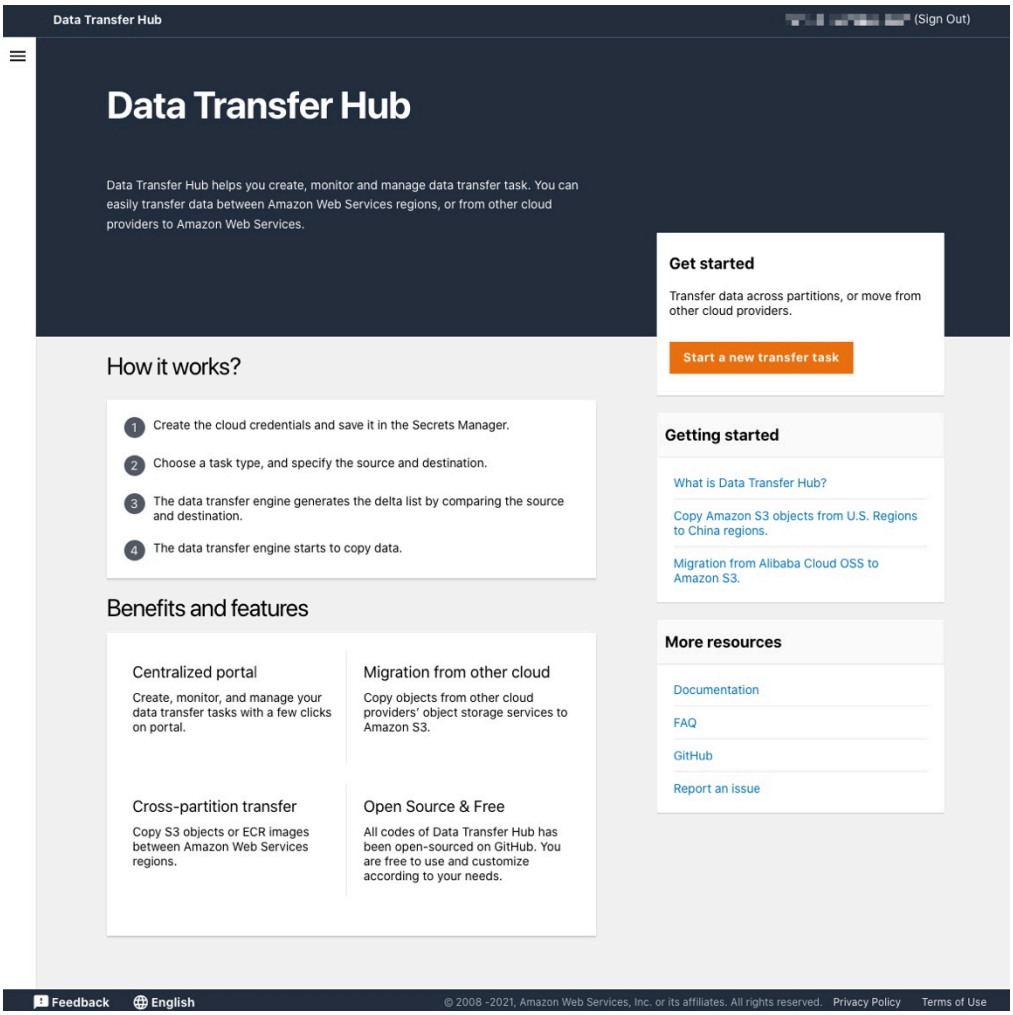
1. Enter the Data Transfer Hub domain name in a web browser.

> **Note**
>
> If you are logging in for the first time, the system will open the Authing.cn login interface.

2. Enter the username and password you registered when you deployed the Guidance, then choose **Login**. The system opens the Data Transfer Hub web console.
3. Change your password and then sign in again.

Step 3. Create a transfer task

Use the web console to create a transfer task for Amazon S3 or Amazon ECR. For more information, refer to Create Amazon S3 Transfer Task and Create Amazon ECR Transfer Task.



Data Transfer Hub web console

Use Data Transfer Hub

This guide provides detailed instructions to perform these tasks:

- Create an Amazon S3 transfer task
- Create an Amazon ECR transfer task
- Transfer Amazon S3 object from Alibaba Cloud OSS
- Transfer Amazon S3 object via Direct Connect

Create an Amazon S3 transfer task

The Guidance allows you to create an Amazon S3 transfer task in the following ways:

- using the web console
- using the S3 plugin
- using AWS CLI

You can make your choice according to your needs.

- The web console provides an intuitive user interface where you can start, clone or stop a data transfer task with a simple click. The frontend also provides metric monitoring and logging view, so you do not need to switch between different pages.
- The S3 plugin is a standalone CloudFormation template, and you can easily integrate it into your workflows. Because this option allows deployment without the frontend, it is useful if you want to deploy in AWS China Regions but do not have an ICP licensed domain.
- AWS CLI can quickly initiate data transfer tasks. Select this option if you want to leverage Data Transfer Hub in your automation scripts.

Using the web console

You can use the web console to create an Amazon S3 transfer task. For more information about how to launch the web console, see Deploy the Guidance.

1. From the **Create Transfer Task** page**,** select **Start a New Task**, and then select **Next**.

2. From the **Engine options** page**,** under engine, select **Amazon S3**, and then choose **Next Step**.
3. Specify the transfer task details.
   - Under **Source Type**, select the data source, for example, **Amazon S3**.
4. Enter the bucket name and choose to sync Full Bucket or Objects with a specific prefix or Objects with different prefixes.
   - If the data source bucket is in the account where Data Transfer Hub was deployed, select **Yes.**
     - If you need to achieve real-time incremental data synchronization, please configure whether to enable S3 event notification. Note that this option can only be configured when the program and your data source are deployed in the same area of the same account.
     - If you do not enable S3 event notification, the program will periodically synchronize incremental data according to the scheduling frequency you configure in the future.
   - If the source bucket is not in the same account where Data Transfer Hub was deployed, select **No**, then specify the credentials for the source bucket.
   - If you choose to synchronize objects with multiple prefixes, please transfer the prefix list file separated by rows to the root directory of the bucket where Data Transfer Hub is deployed, and then fill in the name of the file. For details, please refer to Multi-Prefix List Configuration Tutorial.
5. To create credential info, choose **Secrets Manager** to navigate to the current Region's AWS Secrets Manager console.
   a. From the left menu, select **Secrets**, then choose **Store a new secret** and select the **other type of secrets** key type.
   b. Fill in the `access_key_id` and `secret_access_key` information in the **Plaintext** input box according to the displayed format. For more information, refer to IAM features in the *IAM User Guide*. Choose **Next**.

```
{
    "access_key_id": "<Your Access Key ID>",
    "secret_access_key": "<Your Access Key Secret>"
}
```

   c. (Optional) Enter the key name and description. Choose **Next**.
   d. In the configuration of automatic rotation, select **Disable automatic rotation**. Choose **Next**.
   e. Keep the default value and choose **Save** to complete the creation of the key.
   f. Navigate back to the Data Transfer Hub task creation interface and refresh the interface. Your new secret is displayed in the drop-down list.

g. Select the certificate (Secret).

6. Provide destination settings for the S3 buckets.

> **Note**
>
> If the source S3 bucket is in the same account where Data Transfer Hub was deployed, then in **destination settings**, you must create or provide credential information for the S3 destination bucket. Otherwise, no credential information is needed. Use the following steps to update the destination settings.

7. From **Engine settings**, verify the values and modify them if necessary. We recommend to have the **minimum capacity** set to at least 1 if for incremental data transfer.
8. At **Task Scheduling Settings**, select your task scheduling configuration.
   - If you want to configure the timed task at a fixed frequency to compare the data difference on both sides of the time, select **Fixed Rate**.
   - If you want to configure a scheduled task through Cron Expression to achieve a scheduled comparison of data differences on both sides, select **Cron Expression**.
   - If you only want to perform the data synchronization task once, select **On Time Transfer**.
9. From **Advanced Options**, keep the default values.
10. At **Need Data Comparison before Transfer**, select your task configuration.
    - If you want to skip the data comparison process and transfer all files, please select **No**.
    - If you only want to synchronize files with differences, please select **Yes**.
11. Enter an email address in **Alarm Email**.
12. Choose **Next** and review your task parameter details.
13. Choose **Create Task**.

After the task is created successfully, it will appear on the **Tasks** page.

> **Note**
>
> If your destination bucket in Amazon S3 is set to require all data uploads to be encrypted with Amazon S3 managed keys, you can check the following tutorial.

Destination bucket encrypted with Amazon S3 managed keys

Select "SSE-S3 AES256" from the dropdown menu under 'Destination bucket policy check' in the destination's configuration. For more information, refer to this documentation.

If your destination bucket is set to require that objects be encrypted using only SSE-KMS (Server-Side Encryption with AWS Key Management Service), which is detailed in this documentation, and your policy looks something like the example provided:

```
{
        "Version": "2012-10-17",
        "Id": "PutObjectPolicy",
        "Statement": [
        {
                "Sid": "DenyIncorrectEncryptionHeader",
                "Effect": "Deny",
                "Principal": "*",
                "Action": "s3:PutObject",
                "Resource": "arn:aws-cn:s3:::dth-sse-debug-cn-north-1/*",
                "Condition": {
                    "StringNotEquals": {
                        "s3:x-amz-server-side-encryption": "aws:kms"
                    }
                }
            },
            {
                "Sid": "DenyUnencryptedObjectUploads",
                "Effect": "Deny",
                "Principal": "*",
                "Action": "s3:PutObject",
                "Resource": "arn:aws-cn:s3:::dth-sse-debug-cn-north-1/*",
                "Condition": {
                    "StringNotEquals": {
                        "s3:x-amz-server-side-encryption-aws-kms-key-id": "arn:aws-
cn:kms:cn-north-1:123456789012:key/7c54749e-eb6a-42cc-894e-93143b32e7c0"
                    }
                }
            }
        ]
    }
```

In this case, you should select "SSE-KMS" in the 'Destination bucket policy check' dropdown menu in the destination's configuration. Additionally, you need to provide the KMS Key ID, such as "7c54749e-eb6a-42cc-894e-93143b32e7c0" in the example.

Using the S3 plugin

> **Note**
>
> This tutorial provides guidance for the backend-only version. For more details, please refer to S3 Plugin Introduction.

Step 1. Prepare VPC

This Guidance can be deployed in both public and private subnets. Using public subnets is recommended.

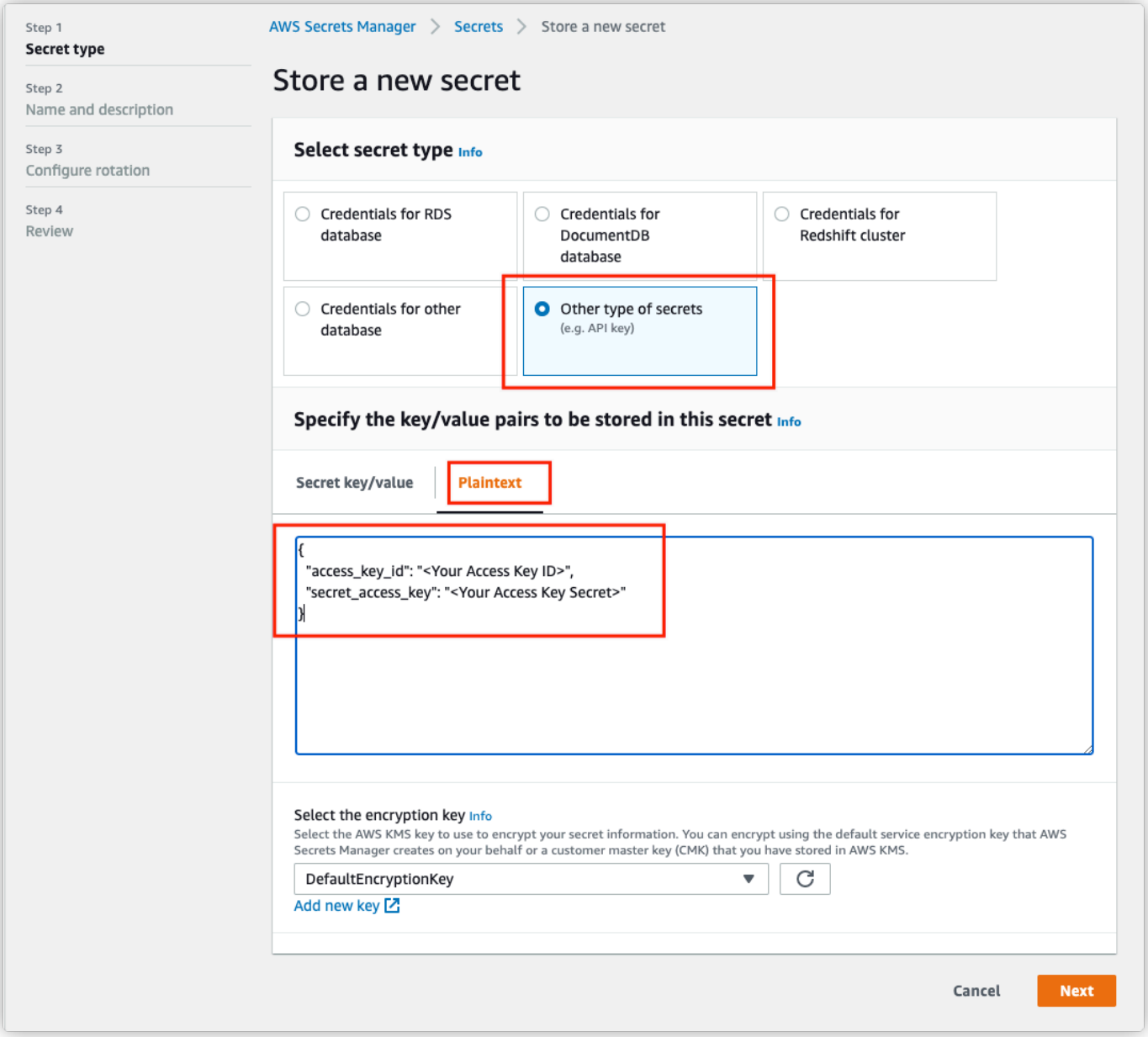
- If you want to use existing VPC, please make sure the VPC has at least 2 subnets, and both subnets must have public internet access (either public subnets with internet gateway or private subnets with NAT gateway).
- If you want to create new default VPC for this Guidance, please go to Step 2 and make sure you have >**Create a new VPC for this cluster** selected when you create the cluster.

Step 2. Configure credentials

You need to provide AccessKeyID and SecretAccessKey (namely AK/SK) to read or write bucket in S3 from or to another AWS account or other cloud storage service, and the credential will be stored in AWS Secrets Manager. You do not need to create credential for the bucket in the current account where you are deploying the Guidance.

1. Go to AWS Management Console > Secrets Manager.
2. From Secrets Manager home page, choose **Store a new secret**.
3. For secret type, select **Other type of secrets**.
4. For key/value paris, please copy and paste below JSON text into the **Plaintext** section, and change value to your AK/SK accordingly.

```
{
  "access_key_id": "<Your Access Key ID>",
  "secret_access_key": "<Your Access Key Secret>"
}
```



5. Choose **Next** to specify a secret name, and choose **Create**.

If the AK/SK is for source bucket, READ access to bucket is required; if it is for destination bucket, READ and WRITE access to bucket is required. For Amazon S3, you can refer to Set up credentials for Amazon S3 for more information.

Step 3. Launch AWS Cloudformation Stack

Please follow below steps to deploy this Guidance via AWS Cloudformation.

1. Sign in to AWS Management Console, and switch to the Region where you want to deploy the CloudFormation Stack.
2. Choose the following to launch the CloudFormation Stack.
   - For AWS China Regions

   - For AWS Global Regions

3. Choose **Next**. Specify values to parameters accordingly. Change the stack name if required.
4. Choose **Next**. Configure additional stack options such as tags if needed.
5. Choose **Next**. Review and confirm acknowledgement, and then choose **Create Stack** to start the deployment.

The deployment will take approximately 3 to 5 minutes.

Using AWS CLI

You can use the AWS CLI to create an Amazon S3 transfer task. Note that if you have deployed the Data Transfer Hub Portal at the same time, the tasks started through the CLI will not appear in the Task List on your Portal.

1. Create an Amazon VPC with two public subnets or two private subnets with NAT gateway.
2. Replace <CLOUDFORMATION_URL> as shown below.

```
https://solutions-reference.s3.amazonaws.com/data-transfer-hub/latest/
DataTransferS3Stack.template
```

3. Go to your terminal and enter the following command. For the parameter details, refer to the Parameters table.

```
aws cloudformation create-stack --stack-name dth-s3-task --template-url
 CLOUDFORMATION_URL \
--capabilities CAPABILITY_NAMED_IAM \
--parameters \
ParameterKey=alarmEmail,ParameterValue=your_email@example.com \
ParameterKey=destBucket,ParameterValue=dth-receive-cn-north-1 \
```

```
ParameterKey=destPrefix,ParameterValue=test-prefix \
ParameterKey=destCredentials,ParameterValue=drh-cn-secret-key \
ParameterKey=destInCurrentAccount,ParameterValue=false \
ParameterKey=destRegion,ParameterValue=cn-north-1 \
ParameterKey=destStorageClass,ParameterValue=STANDARD \
ParameterKey=destPutObjectSSEType,ParameterValue=None \
ParameterKey=destPutObjectSSEKmsKeyId,ParameterValue= \
ParameterKey=srcBucket,ParameterValue=dth-us-west-2 \
ParameterKey=srcInCurrentAccount,ParameterValue=true \
ParameterKey=srcCredentials,ParameterValue= \
ParameterKey=srcRegion,ParameterValue=us-west-2 \
ParameterKey=srcPrefix,ParameterValue=case1 \
ParameterKey=srcType,ParameterValue=Amazon_S3 \
ParameterKey=ec2VpcId,ParameterValue=vpc-040bbab85f0e4e088 \
ParameterKey=ec2Subnets,ParameterValue=subnet-0d1bf2725ab8e94ee\
\,subnet-06d17b2b3286be40e \
ParameterKey=finderEc2Memory,ParameterValue=8 \
ParameterKey=ec2CronExpression,ParameterValue="0/60 * * * ? *" \
ParameterKey=includeMetadata,ParameterValue=false \
ParameterKey=srcEvent,ParameterValue=No \
ParameterKey=maxCapacity,ParameterValue=20 \
ParameterKey=minCapacity,ParameterValue=1 \
ParameterKey=desiredCapacity,ParameterValue=1
```

Parameter	Allowed Value	Default Value	Description
alarmEmail			An email to which errors will be sent
desiredCapacity		1	Desired capacity for Auto Scaling Group
destAcl	private public-read public-read-write authenticated-read aws-exec-read	bucket-owner-full-control	Destination access control list

Parameter	Allowed Value	Default Value	Description
	bucket-owner-read bucket-owner-full-control		
destBucket			Destination bucket name
destCredentials			Secret name in Secrets Manager used to keep AK/SK credentials for destination bucket. Leave it blank if the destination bucket is in the current account
destInCurrentAccount	true false	true	Indicates whether the destination bucket is in current account. If not, you should provide a credential with read and write access
destPrefix			Destination prefix (Optional)
destRegion			Destination region name

Parameter	Allowed Value	Default Value	Description
destStorageClass	STANDARD STANDARD_IA ONEZONE_IA INTELLIGENT_TIERING	INTELLIGENT_TIERING	Destination storage class, which defaults to INTELLIGENT_TIERING
destPutObjectSSEType	None AES256 AWS_KMS	None	Specifies the server-side encryption algorithm used for storing objects in Amazon S3. 'AES256' applies AES256 encryption, 'AWS_KMS' uses AWS Key Management Service encryption, and 'None' indicates that no encryption is applied.

Parameter	Allowed Value	Default Value	Description
destPutObjectSSEKmsKeyId			Specifies the ID of the symmetric customer managed AWS KMS Customer Master Key (CMK) used for object encryption. This parameter should only be set when destPutObjectSSEType is set to 'AWS_KMS'. If destPutObjectSSEType is set to any value other than 'AWS_KMS', please leave this parameter empty. The default value is not set.
isPayerRequest	true false	false	Indicates whether to enable payer request. If true, it will get object in payer request mode.
ec2CronExpression		0/60 * * * ? *	Cron expression for EC2 Finder task "" for one time transfer.

Parameter	Allowed Value	Default Value	Description
finderEc2Memory	8 16 32 64 128 256	8 GB	The amount of memory (in GB) used by the Finder task.
ec2Subnets			Two public subnets or two private subnets with NAT gateway
ec2VpcId			VPC ID to run EC2 task, for example, vpc-bef13dc7
finderDepth		0	Depth of sub folders to compare in parallel. 0 means comparing all objects in sequence
finderNumber		1	The number of finder threads to run in parallel
includeMetadata	true false	false	Indicates whether to add replication of object metadata. If true, there will be additional API calls.

Parameter	Allowed Value	Default Value	Description
maxCapacity		20	Maximum capacity for Auto Scaling Group
minCapacity		1	Minimum capacity for Auto Scaling Group
srcBucket			Source bucket name
srcCredentials			Secret name in Secrets Manager used to keep AK/SK credentials for Source Bucket. Leave it blank if source bucket is in the current account or source is open data
srcEndpoint			Source Endpoint URL (Optional). Leave it blank unless you want to provide a custom Endpoint URL
srcEvent	No Create CreateAndDelete	No	Whether to enable S3 Event to trigger the replication. Note that S3Event is only applicable if source is in the current account

Parameter	Allowed Value	Default Value	Description
srcInCurrentAccount	true false	false	Indicates whether the source bucket is in the current account. If not, you should provide a credential with read access
srcPrefix			Source prefix (Optional)
srcPrefixListBucket			Source prefix list file S3 bucket name (Optional). It used to store the Source prefix list file. The specified bucket must be located in the same AWS region and under the same account as the DTH deploymen t. If your PrefixList File is stored in the Source Bucket, please leave this parameter empty.
srcPrefixsListFile			Source prefix list file S3 path (Optional). It supports txt type, for example, my_prefix _list.txt, and the maximum number of lines is 10 millions

Parameter	Allowed Value	Default Value	Description
srcRegion			Source region name
srcSkipCompare	true false	false	Indicates whether to skip the data comparison in task finding process. If yes, all data in the source will be sent to the destination
srcType	Amazon_S3 Aliyun_OSS Qiniu_Kodo Tencent_COS	Amazon_S3	If you choose to use the Endpoint mode, please select Amazon_S3.
workerNumber	1 ~ 10	4	The number of worker threads to run in one worker node/ instance. For small files (size < 1MB), you can increase the number of workers to improve the transfer performance.

How to transfer S3 object from KMS encrypted Amazon S3

By default, Data Transfer Hub supports data source bucket using SSE-S3 and SSE-KMS.

If your source bucket enabled SSE-CMK, you need to create an IAM Policy and attach it to DTH worker and finder node. You can go to Amazon IAM Roles Console and search for `<StackName>-FinderStackFinderRole<random suffix>` and `<StackName>-EC2WorkerStackWorkerAsgRole<random suffix>`.

Pay attention to the following:

- Change the `Resource` in KMS part to your own KMS key's Amazon Resource Name (ARN).
- For S3 buckets in AWS China Regions, make sure to use `arn:aws-cn:kms:::` instead of `arn:aws:kms:::`.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "kms:Encrypt",
                "kms:ReEncrypt*",
                "kms:GenerateDataKey*",
                "kms:DescribeKey"
            ],
            "Resource": [
                "arn:aws:kms:us-west-2:123456789012:key/f5cd8cb7-476c-4322-
ac9b-0c94a687700d <Please replace this with your own KMS key arn>"
            ]
        }
    ]
}
```

Transferring data cross China Partition

Any data moving out and into the Chinese Partition for AWS has to go through the Chinese Great Firewall. This Great Firewall monitors the content and the size of the data being transferred. If you have too few instances, and are transferring terabytes of data, then that Firewall will block those IP's tied to those few instances. For large use cases of this solution where you might be transferring terabytes of data across Partition, we recommend you spin up multiple instances 200+ and in order to avoid CPU/Memory issues, we recommend using a larger instance type for the AutoScale Group.

Create an Amazon ECR transfer task

The Guidance allows you to create an Amazon S3 transfer task in the following ways:

- using the web console
- using the ECR plugin
- using AWS CLI

You can make your choice according to your needs. For a comparison between those options, refer to Create Amazon S3 transfer task.

Using the web console

You can use the web console to create an Amazon ECR transfer task. For more information about how to launch the web console, see Deploy the Guidance.

1. From the **Create Transfer Task** page**,** select **Start a New Task**, and then select **Next**.
2. From the **Engine options** page, under engine, select **Amazon ECR**, and then choose **Next Step**. You can also copy image from Docker Hub, GCR.io, Quay.io, and so on by choosing **Public Container Registry**.
3. Specify the transfer task details. In **Source Type**, select the container warehouse type.
4. In **Source settings**, enter **Source Region** and Amazon Web Services Account ID.
5. To create credential information, select Secrets Manager to jump to the AWS Secrets Manager console in the current region.
   a. From the left menu, select **Secrets**, then choose **Store a new secret** and select the **other type of secrets** key type.
   b. Fill in the `access_key_id` and `secret_access_key` information in the Plaintext input box according to the displayed format. For more information, refer to IAM features in the IAM User Guide. Choose **Next**.
   c. (Optional) Enter the key name and description. Choose **Next**.
   d. In the configuration of automatic rotation, select **Disable automatic rotation**. Choose **Next**.
   e. Keep the default value and choose **Save** to complete the creation of the key.
   f. Navigate back to the Data Transfer Hub task creation interface and refresh the interface. Your new secret is displayed in the drop-down list.
6. Select the certificate (Secret).

> **Note**
>
> If the source is in the same account with Data Transfer Hub deployment, you need to create/provide credential info for the destination. Otherwise, no credential information is needed.

7. Enter an email address in **Alarm Email**.
8. Choose **Next** and review your task parameter details.
9. Choose **Create Task**.

After the task is created successfully, it will appear on the **Tasks** page.

Using the ECR plugin

> **Note**
>
> This tutorial provides instructions for the backend-only version. For more details, please refer to ECR Plugin Introduction.

Step 1. Prepare VPC (optional)

This Guidance can be deployed in both public and private subnets. Using public subnets is recommended.

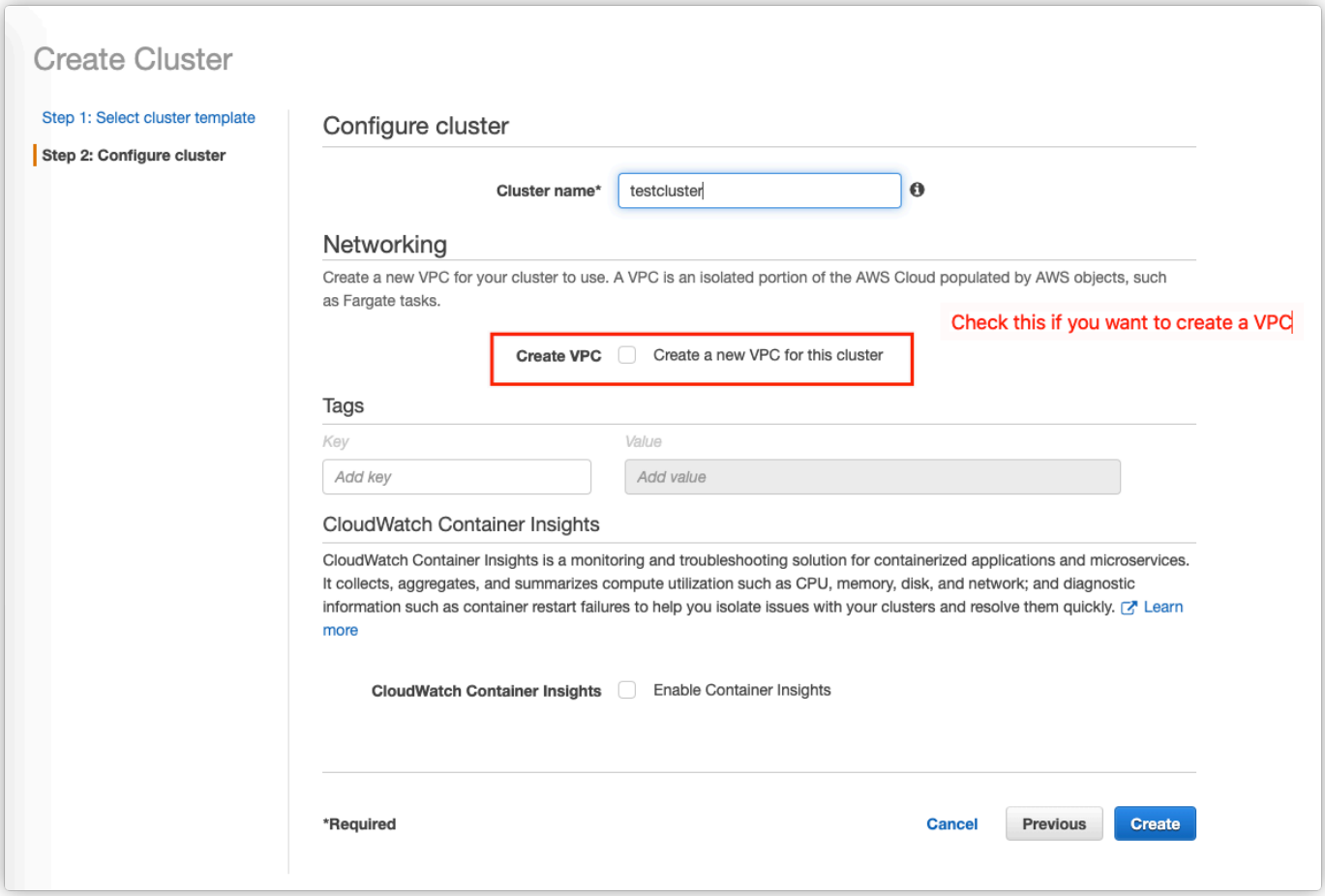
- If you want to use existing VPC, please make sure the VPC has at least 2 subnets, and both subnets must have public internet access (either public subnets with internet gateway or private subnets with NAT gateway).
- If you want to create new default VPC for this Guidance, please go to Step 2 and make sure you have **Create a new VPC for this cluster** selected when you create the cluster.

Step 2. Set up ECS Cluster

An ECS Cluster is required for this Guidance to run Fargate task.

1. Sign in to AWS Management Console, and choose Elastic Container Service (ECS).

2. From ECS Cluster home page, choose **Create Cluster**.
3. Select Cluster Template. Choose **Network Only** type.
4. Specify a cluster name and click Create to create a cluster. If you want to also create a new VPC (public subnets only), please also check the **Create a new VPC for this cluster** option.



Step 3. Configure credentials

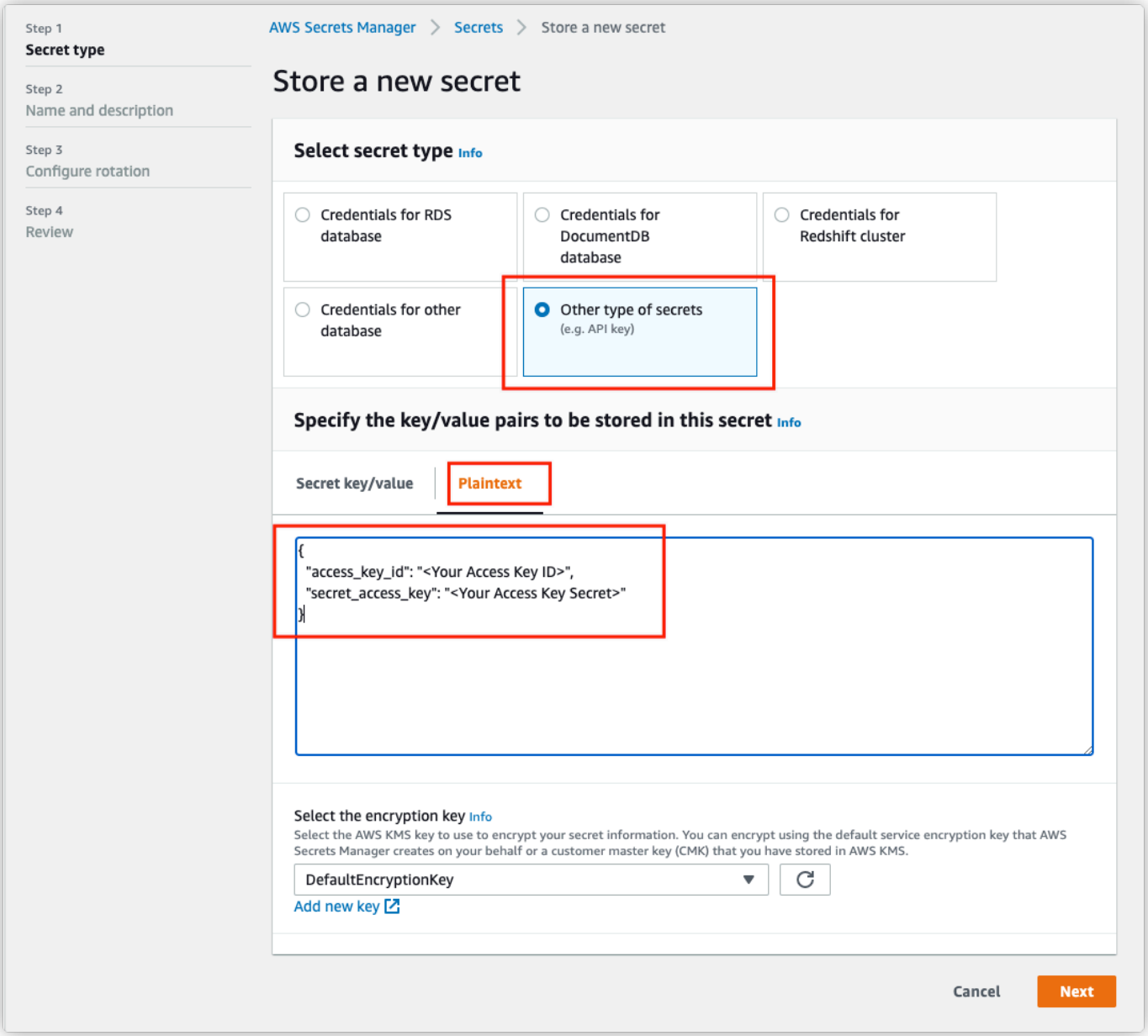
If source (or destination) is NOT in current AWS account, you will need to provide AccessKeyID and SecretAccessKey (namely AK/SK) to pull from or push to Amazon ECR. And Secrets Manager is used to store the credentials in a secure manner.

If source type is Public, there is no need to provide the source credentials.

1. Go to AWS Management Console > Secrets Manager.
2. From Secrets Manager home page, choose **Store a new secret**.
3. For secret type, select **Other type of secrets**.

4. For key/value paris, please copy and paste below JSON text into the **Plaintext** section, and change value to your AK/SK accordingly.

```
{
  "access_key_id": "<Your Access Key ID>",
  "secret_access_key": "<Your Access Key Secret>"
}
```



5. Choose **Next** to specify a secret name, and choose **Create**.

Step 4. Launch AWS Cloudformation Stack

Please follow below steps to deploy this plugin via AWS Cloudformation.

1. Sign in to AWS Management Console, and switch to the Region where you want to deploy the CloudFormation Stack.
2. Choose the following to launch the CloudFormation Stack in that Region.
   - For AWS China Regions

   - For AWS Global Regions

3. Choose **Next**. Specify values to parameters accordingly. Change the stack name if required.
4. Choose **Next**. Configure additional stack options such as tags if needed.
5. Choose **Next**. Review and confirm acknowledgement, and then choose **Create Stack** to start the deployment.

The deployment will take approximately 3 to 5 minutes.

Using AWS CLI

You can use the AWS CLI to create an Amazon ECR transfer task. Note that if you have deployed the Datat Transfer Hub Portal at the same time, the tasks started through the CLI will not appear in the Task List on your Portal.

1. Create an Amazon VPC with two public subnets or two private subnets with NAT gateway.
2. Replace <CLOUDFORMATION_URL> as shown below.

```
https://solutions-reference.s3.amazonaws.com/data-transfer-hub/latest/
DataTransferECRStack.template
```

3. Go to your terminal and enter the following command. For the parameter details, refer to the Parameters table.

```
aws cloudformation create-stack --stack-name dth-ecr-task --template-url
 CLOUDFORMATION_URL \
```

```
--capabilities CAPABILITY_NAMED_IAM \
--parameters \
ParameterKey=sourceType,ParameterValue=Amazon_ECR \
ParameterKey=srcRegion,ParameterValue=us-east-1 \
ParameterKey=srcAccountId,ParameterValue=123456789012 \
ParameterKey=srcList,ParameterValue=ALL \
ParameterKey=includeUntagged,ParameterValue=false \
ParameterKey=srcImageList,ParameterValue= \
ParameterKey=srcCredential,ParameterValue=dev-us-credential \
ParameterKey=destAccountId,ParameterValue= \
ParameterKey=destRegion,ParameterValue=us-west-2 \
ParameterKey=destCredential,ParameterValue= \
ParameterKey=destPrefix,ParameterValue= \
ParameterKey=alarmEmail,ParameterValue=your_email@example.com \
ParameterKey=ecsVpcId,ParameterValue=vpc-07f56e8e21630a2a0 \
ParameterKey=ecsClusterName,ParameterValue=dth-v22-01-TaskCluster-eHzKkHatj0tN \
ParameterKey=ecsSubnetA,ParameterValue=subnet-034c58fe0e696eb0b \
ParameterKey=ecsSubnetB,ParameterValue=subnet-0487ae5a1d3badde7
```

Parameter	Allowed Value	Default Value	Description
sourceType	Amazon_ECR Public	Amazon_ECR	Choose type of source container registry, for example Amazon_ECR, or Public from Docker Hub, gco.io, etc.
srcRegion			Source Region Name (only required if source type is Amazon ECR), for example, us-west-1

Parameter	Allowed Value	Default Value	Description
srcAccountId			Source AWS Account ID (only required if source type is Amazon ECR), leave it blank if source is in current account
srcList	ALL SELECTED	ALL	Type of Source Image List, either ALL or SELECTED, for public registry, please use SELECTED only
srcImageList			Selected Image List delimited by comma, for example, ubuntu:latest,alpine:latest..., leave it blank if Type is ALL. For ECR source, using ALL_TAGS tag to get all tags.
srcCredential			The secret name in Secrets Manager only when using AK/SK credentials to pull images from source Amazon ECR, leave it blank for public registry

Parameter	Allowed Value	Default Value	Description
destRegion			Destination Region Name, for example, cn-north-1
destAccountId			Destination AWS Account ID, leave it blank if destination is in current account
destPrefix			Destination Repo Prefix
destCredential			The secret name in Secrets Manager only when using AK/SK credentials to push images to destination Amazon ECR
includeUntagged	true false	true	Whether to include untagged images in the replication
ecsClusterName			ECS Cluster Name to run ECS task (Please make sure the cluster exists)
ecsVpcId			VPC ID to run ECS task, e.g. vpc-bef13dc7
ecsSubnetA			First Subnet ID to run ECS task, e.g. subnet-97bfc4cd

Parameter	Allowed Value	Default Value	Description
ecsSubnetB			Second Subnet ID to run ECS task, e.g. subnet-7ad7de32
alarmEmail			Alarm Email address to receive notification in case of any failure
sourceType	Amazon_ECR Public	Amazon_ECR	Choose type of source container registry, for example Amazon_ECR, or Public from Docker Hub, gco.io, etc.
srcRegion			Source Region Name (only required if source type is Amazon ECR), for example, us-west-1
srcAccountId			Source AWS Account ID (only required if source type is Amazon ECR), leave it blank if source is in current account
srcList	ALL SELECTED	ALL	Type of Source Image List, either ALL or SELECTED, for public registry, please use SELECTED only

Parameter	Allowed Value	Default Value	Description
srcImageList			Selected Image List delimited by comma, for example, ubuntu:latest,alpine:latest..., leave it blank if Type is ALL. For ECR source, using ALL_TAGS tag to get all tags.
srcCredential			The secret name in Secrets Manager only when using AK/SK credentials to pull images from source Amazon ECR, leave it blank for public registry
destRegion			Destination Region Name, for example, cn-north-1
destAccountId			Destination AWS Account ID, leave it blank if destination is in current account
destPrefix			Destination Repo Prefix

Parameter	Allowed Value	Default Value	Description
destCredential			The secret name in Secrets Manager only when using AK/SK credentials to push images to destination Amazon ECR
includeUntagged	true false	true	Whether to include untagged images in the replication
ecsClusterName			ECS Cluster Name to run ECS task (Please make sure the cluster exists)
ecsVpcId			VPC ID to run ECS task, e.g. vpc-bef13dc7
ecsSubnetA			First Subnet ID to run ECS task, e.g. subnet-97bfc4cd
ecsSubnetB			Second Subnet ID to run ECS task, e.g. subnet-7ad7de32
alarmEmail			Alarm Email address to receive notification in case of any failure

Transfer S3 object from Alibaba Cloud OSS

This tutorial describes how to transfer Objects from Alibaba Cloud OSS to Amazon S3.

Prerequisite

You have already deployed the Data Transfer Hub in Oregon (us-west-2) region. For more information, see deployment.

Step 1: Configure credentials for OSS

1. Open the Secrets Manager console.
2. Choose **Secrets** in the left navigation pane.
3. Select **Store a new secret**.
4. Select **Other type of secrets**.
5. Enter the credentials of Alibaba Cloud as text in plaintext. The credentials are in the format of:

```
{
    "access_key_id": "<Your Access Key ID>",
    "secret_access_key": "<Your Access Key Secret>"
}
```

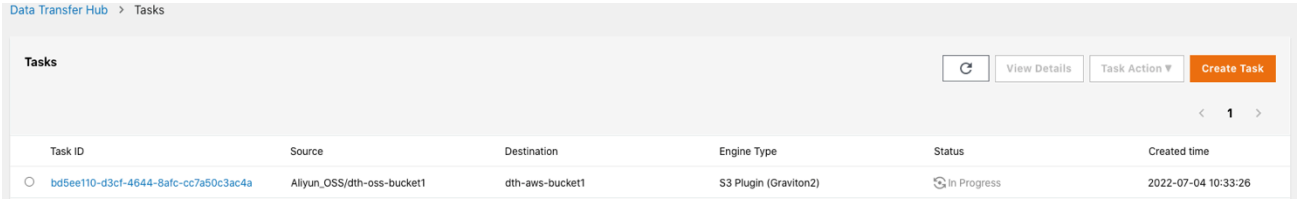
6. Select **Next**.
7. Enter **Secret name**. For example, `dth-oss-credentials`.
8. Select **Next**.
9. Select **Disable automatic rotation**.
10. Select **Store**.

Step 2: Create an OSS transfer task

1. From the **Create Transfer Task** page, select **Start a New Task**, and then select **Next**.
2. From the **Engine options** page, under engine, select **Amazon S3**, and then choose **Next Step**.
3. Specify the transfer task details.
   - Under Source Type, select the data source Aliyun OSS.
4. Enter bucket name and choose to sync Full Bucket or Objects with a specific prefix or Objects with different prefixes.

5. Provide destination settings for the Amazon S3 buckets.
6. From **Engine settings**, verify the values and modify them if necessary. For incremental data transfer, set the minimum capacity to at least 1.
7. At **Task Scheduling Settings**, select your task scheduling configuration.
   - If you want to configure the timed task at a fixed frequency to compare the data difference on both sides of the time, select **Fixed Rate**.
   - If you want to configure a scheduled task through Cron Expression to achieve a scheduled comparison of data differences on both sides, select **Cron Expression**.
   - If you only want to perform the data synchronization task once, select **One Time Transfer**.
   - If you need to achieve real-time incremental data synchronization, please refer to the event config guide.
8. For **Advanced Options**, keep the default values.
9. At **Need Data Comparison before Transfer**, select your task configuration.
   - If you want to skip the data comparison process and transfer all files, select **No**.
   - If you only want to synchronize files with differences, select **Yes**.
10. Enter an email address in **Alarm Email**.
11. Choose **Next** and review your task parameter details.
12. Choose **Create Task**.

After the task is created successfully, it will appear on the **Tasks** page.



Transfer task details and status

Select the Task ID to go to the task Details page, and then choose CloudWatch Dashboard to monitor the task status.

How to achieve real-time data transfer by OSS event trigger

If you want to achieve real-time data transfer from Alibaba Cloud OSS to Amazon S3, follow this section to enable OSS event trigger.

After you created the task, go to Amazon SQS console and record the queue URL and queue ARN that will be used later.

Prepare your AWS account's Access Key/Secret Key

1. Sign in to the IAM console.
2. In the navigation pane, choose **Policies**, then choose **Create Policy**.
3. Select the **JSON** tab, and enter the following information.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "sqs:SendMessage"
            ],
            "Resource": "arn:aws:sqs:us-west-2:xxxxxxxxxxx:DTHS3Stack-
S3TransferQueue-1TSF4ESFQEFKJ"
        }
    ]
}
```

> **Note**
>
> Make sure to replace your queue ARN in the JSON.

4. Complete the workflow to create the policy.
5. In the navigation pane, choose **Users**, then choose **Add users**.
6. Attach the policy you created previously to the user.
7. Save the ACCESS_KEY/SECRET_KEY which will be used later.

Prepare the event-sender function for Alibaba Cloud

1. Open the terminal and enter the following command. You can use docker or Linux machine.

```
mkdir tmp
cd tmp
```

```
pip3 install -t . boto3
```

2. Create an `index.py` in the same folder, and enter the code below.

```
import json
import logging
import os
import boto3


def handler(event, context):
    logger = logging.getLogger()
    logger.setLevel('INFO')
    evt = json.loads(event)
    if 'events' in evt and len(evt['events']) == 1:
        evt = evt['events'][0]
        logger.info('Got event {}'.format(evt['eventName']))
        obj = evt['oss']['object']
        # logger.info(obj)
        ak = os.environ['ACCESS_KEY']
        sk = os.environ['SECRET_KEY']
        queue_url = os.environ['QUEUE_URL']
        region_name = os.environ['REGION_NAME']
        # minimum info of a message
        obj_msg = {
            'key': obj['key'],
            'size': obj['size']
        }
        # start sending the msg
        sqs = boto3.client('sqs', region_name=region_name,
                           aws_access_key_id=ak, aws_secret_access_key=sk)
        try:
            sqs.send_message(
                QueueUrl=queue_url,
                MessageBody=json.dumps(obj_msg)
            )
        except Exception as e:
            logger.error(
                'Unable to send the message to Amazon SQS, Exception:', e)
    else:
        logger.warning('Unknown Message '+evt)

    return 'Done'
```

3. Zip the code (including boto3).

```
zip -r code.zip *
```

Create a function in Alibaba Cloud

1. Use your Alibaba Cloud account to log in to Function Compute, and select **Task**.
2. Choose **Create Function**.
3. Choose Python3.x as the Runtime Environments.
4. In **Code Upload Method**, choose Upload ZIP.
5. Upload the `code.zip` created in the previous step to create the function.
6. Select **Create**.

Configure the function's environment variables

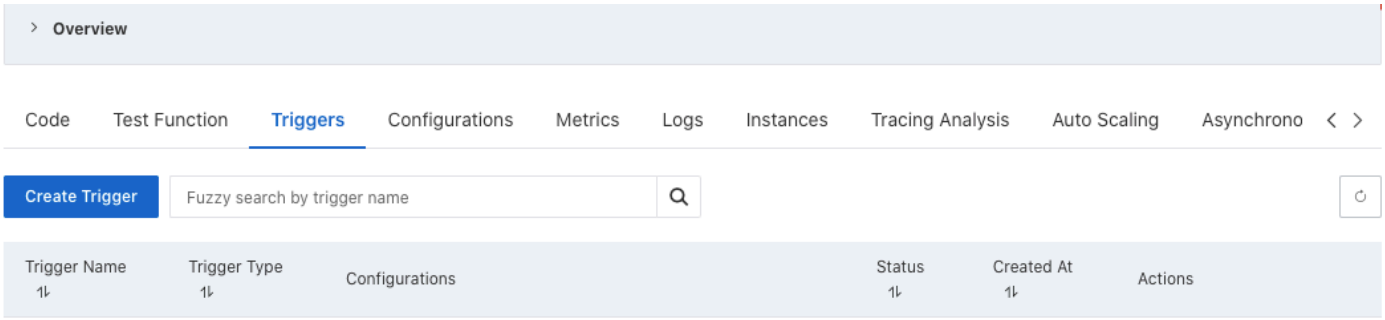
1. Choose the **Configurations**.
2. Select **Modify** in the Environment Variables.
3. Enter the config JSON in the Environment Variables. Here you need to use your own ACCESS_KEY, SECRET_KEY and QUEUE_URL.

```
{
    "ACCESS_KEY": "XXX",
    "QUEUE_URL": "https://sqs.us-west-2.amazonaws.com/xxxx/DTHS3Stack-
S3TransferQueue-xxxx",
    "REGION_NAME": "us-west-2",
    "SECRET_KEY": "XXXXX"
}
```

4. Select **OK**.

Create the trigger

1. Navigate to the **Create Trigger** in **Triggers** tab to create the trigger for the function.



2. Choose **OSS** as the **Trigger Type**, and choose the bucket name.
3. For **Trigger Event**, choose:

```
oss:ObjectCreated:PutObject
oss:ObjectCreated:PostObject
oss:ObjectCreated:CopyObject
oss:ObjectCreated:CompleteMultipartUpload
oss:ObjectCreated:AppendObject
```

4. Select **OK**.

Transfer S3 object via Direct Connect

This tutorial describes how to use Data Transfer Hub (DTH) via Direct Connect (DX).

When the DTH worker node and finder node start to work, they need to download related assets (such as CloudWatch agent, DTH CLI) from the internet by default. In an isolated network, you need to manually download and upload these files to an S3 bucket in the region where DTH is deployed.

You have two options to use DTH to transfer data via DX:

- Use DTH to transfer data via DX in a non-isolated network
- Use DTH to transfer data via DX in an isolated network

Use DTH to transfer data via DX in a non-isolated network

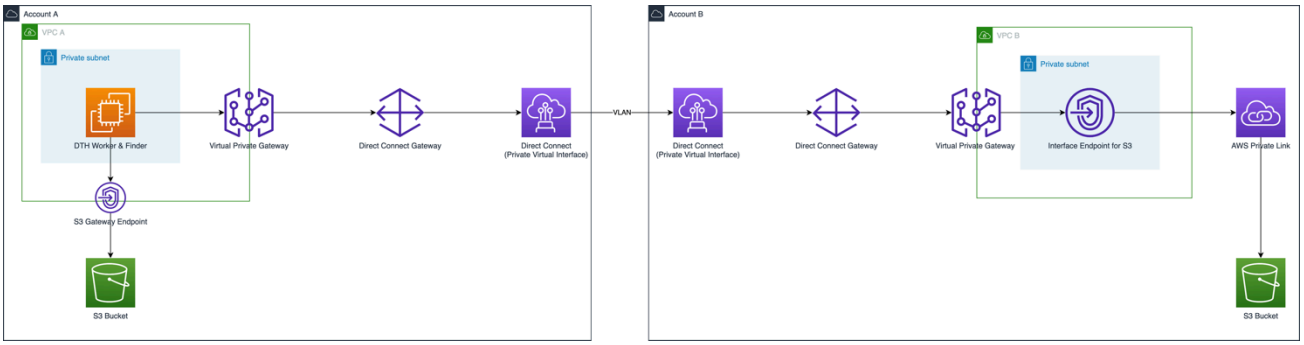
In this scenario, DTH is deployed in the destination side and within a VPC with public access (has Internet Gateway or NAT), and the source bucket is in the isolated network.

> **Note**
>
> As DTH deployment VPC has public internet access (IGW or NAT), EC2 worker/finder nodes can access other AWS services used by DTH such as secret managers and download related assets (such as CloudWatch agent, DTH CLI) from internet without any changes.

1. From the **Create Transfer Task** page, select **Create New Task**, and then select **Next**.
2. From the **Engine options** page, under engine, select **Amazon S3**, and then choose **Next Step**.
3. Specify the transfer task details.
   - Under **Source Type**, select the data source **Amazon S3 Compatible Storage**.
4. Enter **endpoint url**, which must be the **interface endpoint** url, such as `https://bucket.vpce-076205013d3a9a2ca-us23z2ze.s3.ap-east-1.vpce.amazonaws.com`. You can find the specific url in VPC Endpoint Console DNS names part.
5. Enter bucket name and choose to sync Full Bucket or Objects with a specific prefix or Objects with different prefixes.
6. Provide destination settings for the S3 buckets.
7. From **Engine** settings, verify the values and modify them if necessary. For incremental data transfer, we recommend to set the minimum capacity to at least 1.
8. At **Task Scheduling Settings**, select your task scheduling configuration.
   - If you want to configure the timed task at a fixed frequency to compare the data difference on both sides of the time, select **Fixed Rate**.
   - If you want to configure a scheduled task through Cron Expression to achieve a scheduled comparison of data differences on both sides, select **Cron Expression**.
   - If you only want to perform the data synchronization task once, select **One Time Transfer**.
9. For **Advanced Options**, keep the default values.
10. At **Need Data Comparison before Transfer**, select your task configuration.
    - If you want to skip the data comparison process and transfer all files, select **No**.
    - If you only want to synchronize files with differences, select **Yes**.
11. In **Alarm Email**, provide an email address.
12. Choose **Next** and review your task parameter details.
13. Choose **Create Task**.

Use DTH to transfer data via DX in an isolated network

In this scenario, DTH is deployed in the destination side and within a VPC without public access (isolated VPC), and the source bucket is also in an isolated network. For details, refer to the tutorial.



DTH worker nodes running on EC2 transfer data from bucket in one AWS account to bucket in another AWS account.

- To access bucket in the account where DTH is deployed, DTH worker nodes use S3 Gateway Endpoint
- To access bucket in another account, DTH worker nodes use S3 Private Link by S3 Interface Endpoint

Tutorials

This chapter provides tutorials for your reference.

Set up credentials for Amazon S3

Step 1: Create an IAM policy

1. Open AWS Management Console.
2. Choose IAM > Policy, and choose **Create Policy**.
3. Create a policy. You can follow the example below to use IAM policy statement with minimum permissions, and change the `<your-bucket-name>` in the policy statement accordingly.

> **Note**
>
> For S3 buckets in AWS China Regions, make sure you also change to use `arn:aws-cn:s3:::` instead of `arn:aws:s3:::`.

Policy for source bucket

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "dth",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource":[
                "arn:aws:s3:::<your-bucket-name>/*",
                "arn:aws:s3:::<your-bucket-name>"
            ]
        }
    ]
}
```

Policy for destination bucket

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "dth",
            "Effect": "Allow",
            "Action": [
                "s3:PutObject",
                "s3:ListBucket",
                "s3:PutObjectAcl",
                "s3:AbortMultipartUpload",
                "s3:ListBucketMultipartUploads",
                "s3:ListMultipartUploadParts"
            ],
            "Resource": [
                "arn:aws:s3:::<your-bucket-name>/*",
                "arn:aws:s3:::<your-bucket-name>"
            ]
        }
    ]
}
```

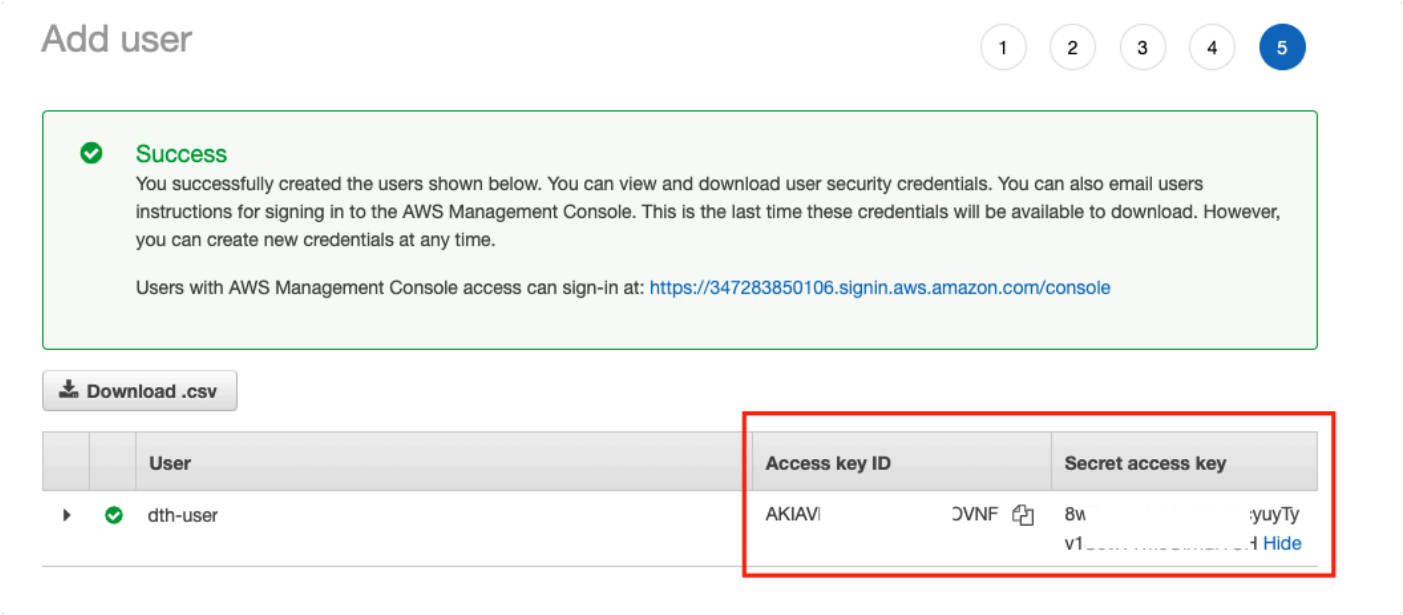
To enable S3 Delete Event, you need to add "s3:DeleteObject" permission to the policy.

Data Transfer Hub has native support for the S3 source bucket which enabled SSE-S3 and SSE-KMS. If your source bucket enabled *SSE-CMK*, please replace the source bucket policy with the policy for S3 SSE-KMS.

Step 2: Create a user

1. Open AWS Management Console.
2. Choose IAM > User, and choose **Add User** to follow the wizard to create a user with credential.
3. Specify a user name, for example, `dth-user`.
4. For Access Type, select **Programmatic access** only and choose **Next: Permissions**.
5. Select **Attach existing policies directly**, search and use the policy created in Step 1, and choose **Next: Tags**.
6. Add tags if needed, and choose **Next: Review**.
7. Review the user details, and choose **Create User**.

8. Make sure you copied/saved the credential, and then choose **Close**.



Policy for S3 Source Bucket with SSE-CMK enabled

Data Transfer Hub has native support for data source using SSE-S3 and SSE-KMS. If your source bucket enabled *SSE-CMK*, please replace the source bucket policy with the following policy, and change the `<your-bucket-name>` in the policy statement accordingly.

Pay attention to the following:

- Change the Resource in KMS part to your own KMS key's Amazon Resource Name (ARN).
- For S3 buckets in AWS China Regions, make sure to use `arn:aws-cn:s3:::` instead of `arn:aws:s3:::`

For Source Bucket with SSE-CMK enabled

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "dth",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
```

```
            ],
            "Resource": [
                "arn:aws:s3:::<your-bucket-name>/*",
                "arn:aws:s3:::<your-bucket-name>"
            ]
        },
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "kms:Encrypt",
                "kms:ReEncrypt*",
                "kms:GenerateDataKey*",
                "kms:DescribeKey"
            ],
            "Resource": [
                "arn:aws:kms:us-west-2:111122223333:key/f5cd8cb7-476c-4322-
ac9b-0c94a687700d <Please replace to your own KMS key arn>"
            ]
        }
    ]
}
```

Upgrade Data Transfer Hub

Time to upgrade: Approximately 20 minutes

Upgrade overview

Use the following steps to upgrade the Guidance on AWS console.

- Step 1. Update the CloudFormation Stack
- Step 2. (Optional) Update the OIDC configuration
- Step 3. Refresh the web console

Step 1. Update the CloudFormation stack

1. Go to the AWS CloudFormation console.
2. Select the Data Transfer Hub main stack, and choose **Update**.
3. Choose **Replace current template**, and enter the specific Amazon S3 URL according to your initial deployment type. Refer to Deployment Overview for more details.

Type	Link
Launch in AWS Regions	`https://s3.amazonaws.com/solutions-reference/data-transfer-hub/latest/DataTransferHub-cognito.template`
Launch in AWS China Regions	`https://s3.amazonaws.com/solutions-reference/data-transfer-hub/latest/DataTransferHub-openid.template`

4. Under **Parameters**, review the parameters for the template and modify them as necessary.
5. Choose **Next**.
6. On Configure stack options page, choose **Next**.

7. On Review page, review and confirm the settings. Check the box **I acknowledge that AWS CloudFormation might create IAM resources**.
8. Choose **Update stack** to deploy the stack.

You can view the status of the stack in the AWS CloudFormation console in the Status column. You should receive a UPDATE_COMPLETE status in approximately 15 minutes.

Step 2. (Optional) Update the OIDC configuration

If you have deployed the Guidance in China Region with OIDC, refer to the deployment section to update the authorization and authentication configuration in OIDC.

Step 3. Refresh the web console

Now you have completed all the upgrade steps. Please click the refresh button in your browser.

Uninstall the Guidance

You can uninstall the Data Transfer Hub Guidance from the AWS Management Console or by using the AWS Command Line Interface. You must manually stop any active transfer tasks before uninstalling.

Using the AWS Management Console

1. Sign in to the AWS CloudFormation console.
2. On the **Stacks** page, select this Guidance's installation stack.
3. Choose **Delete**.

Using AWS Command Line Interface

Determine whether the AWS Command Line Interface (AWS CLI) is available in your environment. For installation instructions, refer to What Is the AWS Command Line Interface in the *AWS CLI User Guide*. After confirming that the AWS CLI is available, run the following command.

```
$ aws cloudformation delete-stack --stack-name <installation-stack-name>
```

Deleting the Amazon S3 buckets

This Guidance is configured to retain the Guidance-created Amazon S3 bucket (for deploying in an opt-in Region) if you decide to delete the AWS CloudFormation stack to prevent accidental data loss. After uninstalling the Guidance, you can manually delete this S3 bucket if you do not need to retain the data. Follow these steps to delete the Amazon S3 bucket.

1. Sign in to the Amazon S3 console.

1. Choose **Buckets** from the left navigation pane.
2. Locate the `<stack-name>` S3 buckets.
3. Select the S3 bucket and choose **Delete**.

To delete the S3 bucket using AWS CLI, run the following command:

```
$ aws s3 rb s3://<bucket-name> --force
```

FAQ

The following are common questions you might have when deploying and using the Guidance.

Deployment

1. In which AWS Regions can this Guidance be deployed?

For the list of supported Regions, refer to Supported AWS Regions.

2. When creating a transfer task, should I deploy it on the data source side or the destination side?

The transfer performance of the Guidance will not be affected by whether the deployment is on the data source or destination side.

If you do not have a domain name registered by ICP in AWS China Regions, we recommend you deploy it in AWS Regions.

If you need to deploy in AWS China Regions but do not have a domain name, you can directly deploy the back-end version:

- Amazon S3 Plugin: https://github.com/awslabs/data-transfer-hub/blob/main/docs/S3_PLUGIN.md
- Amazon ECR Plugin: https://github.com/awslabs/data-transfer-hub/blob/main/docs/ECR_PLUGIN.md

3. Do I need to deploy the Guidance on the data source and destination side separately?

No. You can choose to deploy on the data source or destination side, which has no impact on the transfer performance.

4. Is it possible to deploy the Guidance in AWS account A and transfer Amazon S3 objects from account B to account C?

Yes. In this case, you need to store the AccessKeyID and SecretAccessKey of account B and account C in the Secrets Manager of account A.

5. For data transfer within the production account, is it recommended to create an AWS account specifically for deploying the Guidance?

Yes. It is recommended to create a new AWS account dedicated to deploying Guidances. The account-level isolation improves the stability of the production account in the data synchronization process.

6. Is it possible to transfer data between different areas under the same account?

Not supported currently. For this scenario, we recommend using Amazon S3's Cross-Region Replication.

7. Can I use AWS CLI to create a DTH S3 Transfer Task?

Yes. Please refer to the tutorial Using AWS CLI to launch DTH S3 Transfer task.

Performance

1. Will there be any difference in data transfer performance for deployment in AWS China Regions and in AWS Regions?

No. If you do not have a domain name registered by ICP in AWS China Regions, it is recommended to deploy it in the AWS Regions.

2. What are the factors influencing the data transfer performance?

The transfer performance may be affected by average file size, destination of data transfer, geographic location of data source, and real-time network environment.

3. What is the scale up/scale down policy of Worker Auto Scaling group?

The Auto Scaling group will automatically scale up or scale down according to the number of tasks in SQS.

- Scaling Up Steps are:

```
{ lower: 100,   change: +1 }
{ lower: 500,   change: +2 }
{ lower: 2000,  change: +5 }
{ lower: 10000, change: +10 }
```

- Scaling Down Step is:

```
{ upper: 0, change: -10000 }
```

Data security and authentication

1. How does the Guidance ensure data security?

The Guidance adopts the following to ensure data security:

- All data is transferred in the memory in the transfer node cluster, without being placed on the disk.
- The external ports of all transfer nodes are closed, and there is no way to SSH into the transfer node.
- All data download and upload bottom layers are calling AWS official API, and data transfer conforms to the TLS protocol.

2. How does the Guidance ensure the security of resources on the cloud?

In the research and development process, we strictly follow the minimum IAM permission design rules, and adopt the design of Auto Scaling, which will automatically help users terminate idle working nodes.

3. Is the front-end console open to the public network? How to ensure user authentication and multi-user management?

Yes. You can access it with a front-end console link. User authentication and multi-user management are achieved through AWS Cognito User Pool in AWS Regions, and through OIDC SAAS in AWS China Regions.

4. How does the Guidance achieve cross-account and cross-cloud authentication?

By authentication through the Access Keyid and Access Key of the other party's account. The secret key is stored in AWS Secrets Manager and will be read in Secrets Manager as needed.

5. Does the Guidance support SSE-S3, SSE-KMS, and SSE-CMK?

Yes. The Guidance supports the use of SSE-S3 and SSE-KMS data sources. If your source bucket has SSE-CMK enabled, refer to the tutorial.

Features

1. What third-party clouds does Amazon S3 sync currently support?

Alibaba Cloud OSS, Tencent Cloud, Huawei Cloud, Qiniu Cloud, Baidu Cloud, and all clouds that support S3 compatible protocols.

2. Why is the status of Task still in progress after all destination files are transferred? When will the task stop?

- For Fixed Rate Job

  The data difference between the data source and destination will be monitored continuously, and the differences between the two sides will be automatically compared after the first deployment.

  Moreover, when the default comparison task once an hour finds a difference, it will also transfer the difference data. Therefore, the status of the Task will always be in progress, unless the user manually terminates the task.

  Based on the built-in automatic expansion function of the Guidance, when there is no data to be transferred, the number of transfer working nodes will be automatically reduced to the minimum value configured by the user.

- For One Time Transfer Job

  When the objects are all transferred to the destination, the status of one time transfer job will become Completed.

  The transfer action will stop and you can select Stop to delete and release all backend resources.

3. How often will the data difference between the data source and destination be compared?

By default, it runs hourly.

At Task Scheduling Settings, you can make the task scheduling configuration.

- If you want to configure the timed task at a fixed frequency to compare the data difference on both sides of the time, select Fixed Rate.

- If you want to configure a scheduled task through Cron Expression to achieve a scheduled comparison of data differences on both sides, select Cron Expression.
- If you only want to perform the data synchronization task once, select One Time Transfer.

4. Is it possible for real-time synchronization of newly added files?

Near-real-time synchronization can be achieved, only if the Data Transfer Hub is deployed in the same AWS account and the same region as the data source. If the data source and the Guidance are not in the same account, you can configure it manually. For more information, refer to the tutorial.

5. Are there restrictions on the number of files and the size of files?

No. Larger files will be uploaded in chunks.

6. If a single file transfer fails due to network issues, how to resolve it? Is there an error handling mechanism?

There will be 5 retries. After 5 retries without success, the task will be notified to the user via email.

7. How to monitor the progress of the transfer by checking information like how many files are waiting to be transferred and the current transfer speed?

You can jump to the customized dashboard of Amazon CloudWatch by clicking the CloudWatch Dashboard link in Task Detail of the web console. You can also go directly to CloudWatch to view it.

8. Do I need to create an S3 destination bucket before creating a transfer task?

Yes, you need to create the destination S3 bucket in advance.

9. How to use Finder depth and Finder number to improve Finder performance?

You can use these two parameters to increase the parallelism of Finder to improve the performance of data comparison.

For example, if there are 12 subdirectories with over 100k files each, such as `Jan, Feb, ..., Dec`.

You are recommended to set **`finderDepth`**=1 and **`finderNumber`**=12, so that your comparison performance will increase by 12 times.

When using finderDepth and finderNumber, make sure that there are no objects in the folder whose level is equal to or less than finderDepth. Otherwise, data loss may occur.

For example, assume that you set the **finderDepth**=2 and **finderNumber**=12 * 31 = 372, and your S3 bucket structure is like `bucket_name/Jan/01/pic1.jpg`.

What will be lost are files like `bucket_name/pic.jpg`, `bucket_name/Jan/pic.jpg`.

What will not be lost are all files under `bucket_name/Jan/33/`, all files under `bucket_name/13/33/`.

10. How to deal with Access Key Rotation?

Currently, when Data Transfer Hub perceived that the Access Key of S3 has been rotated, it will fetch the latest key from AWS Secrets Manager automatically. Therefore, the Access Key Rotation will not affect the migrating process of DTH.

11. Does the Payer Request mode support Public Data Set?

No. Currently, Payer Request data synchronization is only supported through Access Key and Private Key authentication methods.

Others

1. The cluster node (EC2) is terminated by mistake. How to resolve it?

The Auto Scaling mechanism of the Guidance will enable automatic restart of a new working node.

However, if a sharding task being transferred in the node is mistakenly terminated, it may cause that the files to which the shard belongs cannot be merged on the destination side, and the error "api error NoSuchUpload: The specified upload does not exist. The upload ID may be invalid, or the upload may have been aborted or completed" occurs. You need to configure lifecycle rules for Delete expired delete markers or incomplete multipart uploads in the Amazon S3 bucket.

2. The Secrets configuration in Secrets Manager is wrong. How to resolve it?

You need to update Secrets in Secrets Manager first, and then go to the EC2 console to Terminate all EC2 instances that have been started by the task. Later, the Auto Scaling mechanism of the Guidance will automatically start a new working node and update Secrets to it.

3. How to find detailed transfer log?

- For Portal users

Go to Tasks list page, and click the Task ID. You can see the dashboard and logs under the Monitoring section.

Data Transfer Hub has embedded Dashboard and log groups on the Portal, so you do not need to navigate to AWS CloudWatch console to view the logs.

- For Plugin (Pure Backend) users

  When deploying the stack, you will be asked to enter the stack name (`DTHS3Stack` by default), and most resources will be created with the name prefix as the stack name. For example, the format of the queue name is `<StackName>-S3TransferQueue-<random suffix>`. This plugin will create two main log groups.

  - If there is no data transfer, you need to check whether there is a problem in the Finder task log. The following is the log group for scheduling Finder tasks. For more information, refer to the Troubleshooting section.

    ```
    <StackName>-CommonS3RepWorkerLogGroup<random suffix>
    ```

  - The following are the log groups of all EC2 instances, and you can find detailed transfer logs.

    ```
    <StackName>-EC2WorkerStackS3RepWorkerLogGroup<random suffix>
    ```

4. How to make customized build?

If you want to make customized changes to this plugin, refer to Custom Build.

5. After the deployment is complete, why can't I find any log streams in the two CloudWatch log groups?

This is because the subnet you selected when deploying this Guidance does not have public network access, and the EC2 cannot download the CloudWatch agent to send logs to CloudWatch. Check your VPC settings. After resolving the issue, you need to manually terminate the running EC2 instance (if any) through this Guidance. Later, the elastic scaling group will automatically start a new instance.

6. How to use TLSv1.2_2021 or later for this Guidance?

Please go to the CloudFront Console and configure a custom domain, which will allow you to select a Security policy for CloudFront after Guidance deployment. You need to prepare a domain name and a corresponding TLS certificate in order to use more secure TLS configurations.

Troubleshooting

After creating the task, you may encounter some error messages. The following list the error messages and provide general steps to troubleshoot them.

1. StatusCode: 400, InvalidToken: The provided token is malformed or otherwise invalid

If you get this error message, confirm that your secret is configured in the following format. You can copy and paste it directly.

```
{
    "access_key_id": "<Your Access Key ID>",
    "secret_access_key": "<Your Access Key Secret>"
}
```

2. StatusCode: 403, InvalidAccessKeyId: The AWS Access Key Id you provided does not exist in our records

If you get this error message, check if your bucket name and region name are configured correctly.

3. StatusCode: 403, InvalidAccessKeyId: UnknownError

If you get this error message, check whether the Credential stored in Secrets Manager has the proper permissions. For more information, refer to IAM Policy.

4. StatusCode: 400, AccessDeniedException: Access to KMS is not allowed

If you get this error message, confirm that your secret is not encrypted by SSE-CMK. Currently, DTH does not support SSE-CMK encrypted secrets.

5. dial tcp: lookup xxx.xxxxx.xxxxx.xx (http://xxx.xxxxx.xxxxx.xx/) on xxx.xxx.xxx.xxx:53: no such host

If you get this error message, check if your endpoint is configured correctly.

Developer guide

This section provides the source code for the Guidance.

Source code

Visit our GitHub repository to download the source files for this Guidance and to share your customizations with others. The Data Transfer Hub templates are generated using the AWS Cloud Development Kit (AWS CDK). Refer to the README.md file for additional information.

Contributors

- Aiden Dai
- Eva Liu
- Kervin Hu
- Haiyun Chen
- Joe Shi
- Ashwini Rudra
- Jyoti Tyagi

Revisions

Date	Change
December 2021	Initial release
September 2022	Released version 1.1 • Upgraded Lambda runtime to Python v3.7.0 • Fixed the list limit of secrets. For more information, refer to the CHANGELOG.md file in the GitHub repository.
July 2022	Released version 2.2 Supported transfer data through Direct Connect. For more information, refer to the CHANGELOG.md file in the GitHub repository.
March 2023	Released version 2.3 • Supported embedded dashboard and logs • Supported S3 Access Key Rotation • Enhanced One Time Transfer Task monitoring For more information, refer to the CHANGELOG.md file in the GitHub repository.
April 2023	Released version 2.3.1 Bug fixes. For more information, refer to the CHANGELOG.md file in the GitHub repository.
April 2023	Released version 2.4

Date	Change
	Supported payer request Amazon S3 object transfer. For more information, refer to the CHANGELOG.md file in the GitHub repository.
September 2023	Released version 2.5 • Enhanced Amazon S3 large file transfer performance by utilizing the cluster's parallel capabilities • Enabled untagged ECR image transfer • Optimized stop task operation, and added new filter conditions to view all history tasks For more information, refer to the CHANGELOG.md file in the GitHub repository.

Date	Change
January 2024	Released version 2.6 • Added support for Amazon S3 destintation bucket being encrypted with Amazon S3 managed keys • Provided the optional Amazon S3 bucket to hold prefix list file • Added the feature of deleting KMS Key automatically after the solution pipeline status turns to stopped • Added the feature of Finder Instance enabling DTH-CLI automatically after external reboot • Increased Finder capacity to 316GB&512GB • Added three supported Regions: Asia Pacific (Melbourne), Canada (Calgary), Israel (Tel Aviv) For more information, refer to the CHANGELOG.md file in the GitHub repository.
April 2024	Released version 2.6.1 Bug fixes. For more information, refer to the CHANGELOG.md file in the GitHub repository.
August 2024	Released version 2.6.2 Updated package versions to resolve security vulnerabilities. For more information, refer to the CHANGELOG.md file in the GitHub repository.

Date	Change
September 2024	Released version 2.6.3 • Updated package versions to resolve security vulnerabilities. • Changed Alpine base image to be sourced from Amazon ECR. For more information, refer to the CHANGELOG.md file in the GitHub repository.
October 2024	Released version 2.6.4 Updated package versions to resolve security vulnerabilities. For more information, refer to the CHANGELOG.md file in the GitHub repository.
November 2024	Released version 2.6.5 Updated package versions to resolve security vulnerabilities. For more information, refer to the CHANGELOG.md file in the GitHub repository.

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents AWS current product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers, or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. AWS responsibilities and liabilities to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Data Transfer Hub is licensed under the terms of the Apache License, Version 2.0.