Solution

High-Availability AIGC Applications with Open-Source Models

Issue	1.0
Date	2023-08-10

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Contents

1 Solution Overview....1
2 Resource Planning and Costs....3
3 Procedure....7
3.1 Preparations....7
3.2 Quick Deployment....10
3.3 Getting Started....17
3.4 Quick Uninstallation....20
4 Appendix....22
5 Change History....23

1 Solution Overview

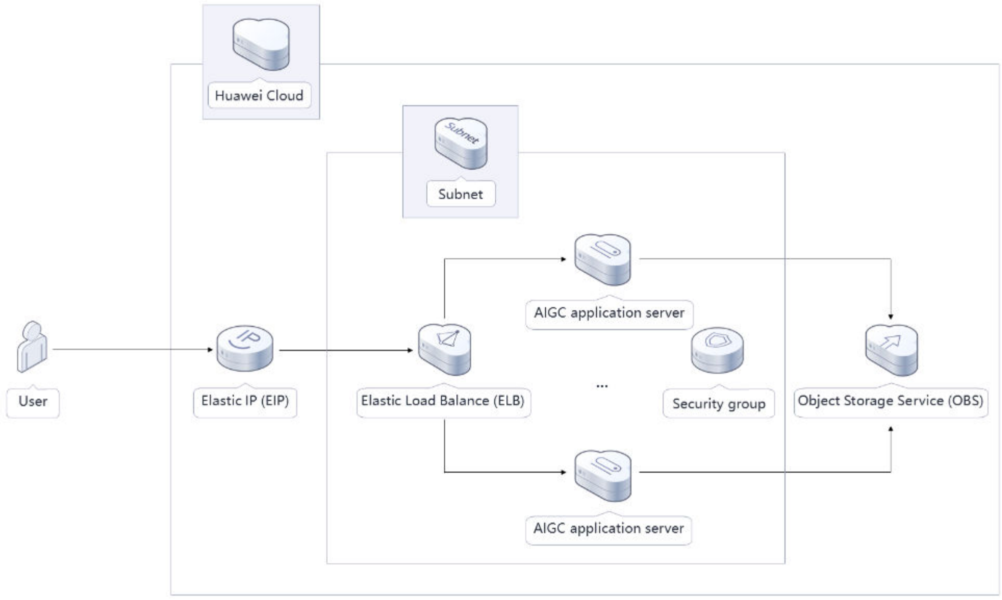
Scenarios

This solution helps you use Stable Diffusion to build high-availability Artificial Intelligence-Generated Content (AIGC) web applications on Huawei Cloud Elastic Cloud Server (ECS). Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text and images input.

Solution Architecture

This solution helps you use Stable Diffusion to build high-availability AIGC web applications on Huawei Cloud Elastic Cloud Server (ECS). The following figure shows the architecture of this solution.

Figure 1-1 Solution architecture



The following resources are required for deploying this solution:

- Two GPU-accelerated Linux ECSs, which will be used for running AIGC applications
- Three Elastic IP addresses (EIPs), which will be bound to the two Linux ECSs and an **Elastic Load Balance (ELB)**, respectively, for internal and external communication
- An ELB, which will be used to distribute traffic across availability zones (AZs)
- An Object Storage Service (OBS) bucket, which will be used to store generated image files
- Stable Diffusion web UI, inotify-tools, and OBS obsutil, which will be installed on each Linux ECS to automatically upload the images saved on the web UI

Advantages

- High availability

  ECSs are deployed across AZs for multi-AZ disaster recovery (DR) and automatic, quick failover.

- Open source and custom development

  This solution is open-source and free for commercial use. You can also make custom development based on source code.

- Easy deployment

  You can easily deploy this solution with just a few clicks.

Constraints

- Before deploying this solution, you need to have created a Huawei Cloud account and completed real-name authentication. You also need to ensure that the account is not frozen and has sufficient balance to pay for the resources required. You can estimate the total price according to the resource planning and costs tables.

2 Resource Planning and Costs

This solution deploys the services listed in the following table. The costs are only estimates and may differ from the final prices. For details, see **Price Calculator**.

Table 2-1 Resource planning and costs (pay-per-use)

Huawei Cloud Service	Example Configuration	Estimated Monthly Cost
Elastic Cloud Server (ECS)	● Pay-per-use: $1.01 USD ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Selected specifications: pi2.2xlarge.4 | 8 vCPUs | 32 GB ● Accelerator: 1 x NVIDIA T4 / 1 x 16 GB ● Image: Ubuntu 20.04 server 64bit with Tesla Driver 460.73.01 and CUDA 11.2 ● System Disk: High I/O | 100 GB ● Quantity: 2	$1.01 USD x 2 x 24 x 30 = $1,454.40 USD
Elastic IP (EIP)	● Pay-per-use: $5.88 USD ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Routing Type: Dynamic BGP ● Billed By: Traffic ● Traffic: 20 GB ● IP Required Duration: 720 hours ● EIP Quantity: 2	$5.88 USD x 2 = $11.76 USD

Huawei Cloud Service	Example Configuration	Estimated Monthly Cost
Elastic IP (EIP)	● Pay-per-use: $0.13 USD/5 Mbit/s/hour ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Product Type: Dedicated ● Routing Type: Dynamic BGP ● Billed By: Bandwidth ● Bandwidth: 5 Mbit/s ● Quantity: 1	$0.13 USD x 24 x 30 = $93.60 USD
Elastic Load Balance (ELB)	● Pay-per-use: $0.05 USD ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Type: Shared load balancer ● Required Duration: 1 hour	$0.05 USD x 24 x 30 = $36.00 USD
Object Storage Service (OBS)	● Region: AP-Singapore ● Product: FunctionGraph ● Request pricing tier: ≤ 1 million requests: $0 USD per 1 million requests > 1 million requests: $0.2 USD per 1 million requests ● Traffic pricing tier: ≤ 400,000 GB-seconds: $0 USD per GB-second > 400,000 GB-seconds: $0.00001667 USD per GB-second	The OBS cost covers the storage and request cost as well as traffic cost. For details, see the monthly bill.
Total	-	**$1,595.76 USD + OBS price**

Table 2-2 Resource planning and costs (yearly/monthly)

Huawei Cloud Service	Example Configuration	Estimated Monthly Cost
Elastic Cloud Server (ECS)	● Region: AP-Singapore ● Billing Mode: Yearly/Monthly ● Selected specifications: pi2.2xlarge.4 | 8 vCPUs | 32 GB ● Accelerator: 1 x NVIDIA T4 / 1 x 16 GB ● Image: Ubuntu 20.04 server 64bit with Tesla Driver 460.73.01 and CUDA 11.2 ● System Disk: High I/O | 100 GB ● Quantity: 2	$549.30 USD x 2 = $1,098.60 USD
Elastic IP (EIP)	● Pay-per-use: $5.88 USD ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Routing Type: Dynamic BGP ● Billed By: Traffic ● Traffic: 20 GB ● IP Required Duration: 720 hours ● EIP Quantity: 2	$5.88 USD x 2 = $11.76 USD
Elastic IP (EIP)	● Region: AP-Singapore ● Billing Mode: Yearly/Monthly ● Product Type: Dedicated ● Routing Type: Dynamic BGP ● Billed By: Bandwidth ● Bandwidth: 5 Mbit/s	$57.00 USD
Elastic Load Balance (ELB)	● Pay-per-use: $0.05 USD ● Region: AP-Singapore ● Billing Mode: Pay-per-use ● Type: Shared load balancer ● Required Duration: 1 hour	$0.05 USD x 24 x 30 = $36.00 USD

Huawei Cloud Service	Example Configuration	Estimated Monthly Cost
Object Storage Service (OBS)	● Region: AP-Singapore ● Product: FunctionGraph ● Request pricing tier: ≤ 1 million requests: $0 USD per 1 million requests > 1 million requests: $0.2 USD per 1 million requests ● Traffic pricing tier: ≤ 400,000 GB-seconds: $0 USD per GB-second > 400,000 GB-seconds: $0.00001667 USD per GB-second	The OBS cost covers the storage and request cost as well as traffic cost. For details, see the monthly bill.
Total	**-**	**$1,203.36 USD + OBS price**

3 Procedure

3.1 Preparations

3.2 Quick Deployment

3.3 Getting Started

3.4 Quick Uninstallation

3.1 Preparations

Creating the rf_admin_trust Agency

Step 1 Log in to the **Huawei Cloud console**, hover the mouse pointer over the account name in the upper right corner, and choose **Identity and Access Management**.

Figure 3-1 Huawei Cloud console

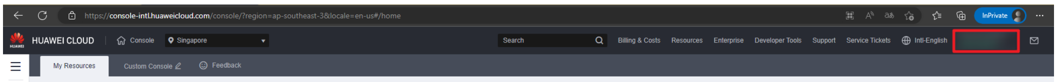
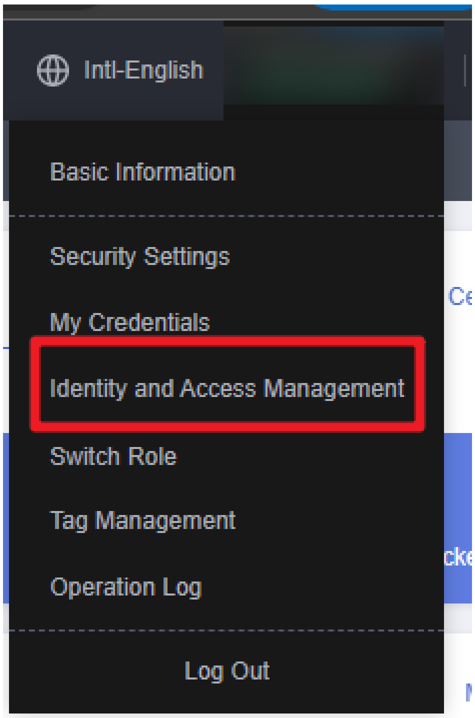
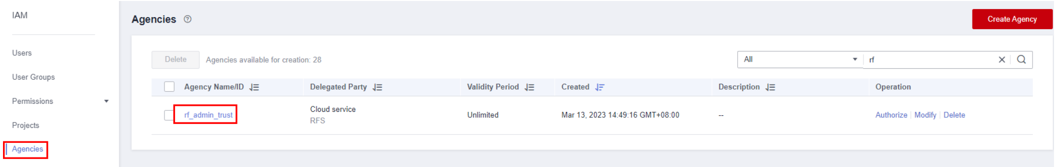


Figure 3-2 Identity and Access Management



Step 2 Choose **Agencies** and then search for the **rf_admin_trust** agency in the agency list.

Figure 3-3 Agency list



Agencies available for creation: 28

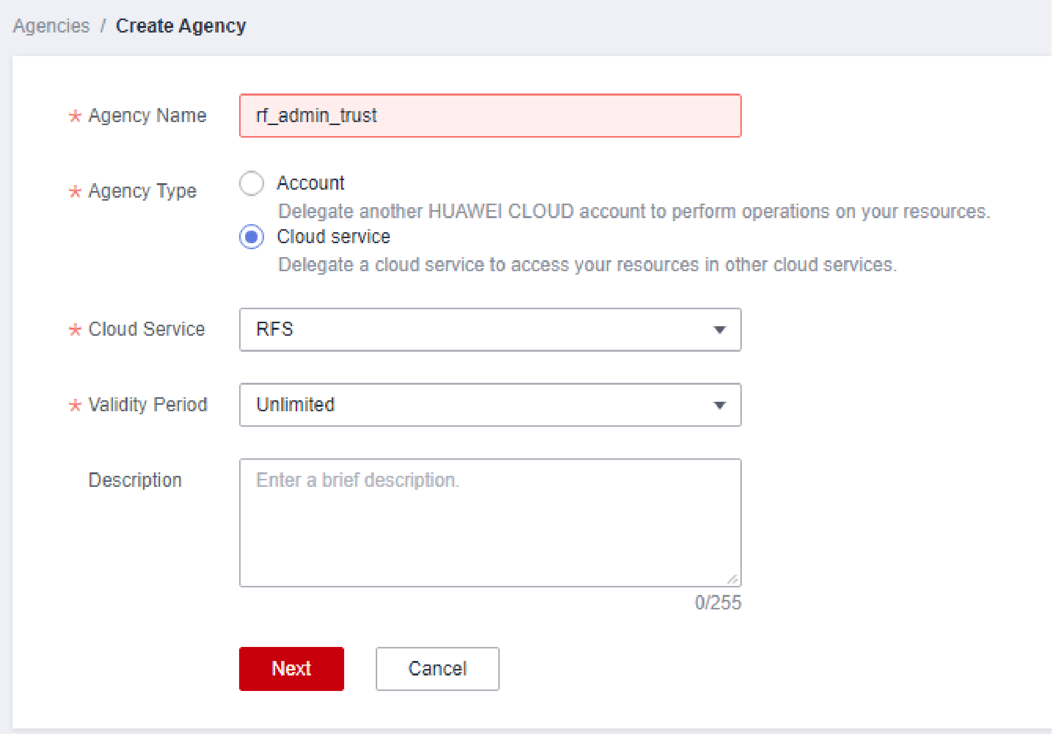
rf

Agency Name/ID	Delegated Party	Validity Period	Created	Description	Operation
rf_admin_trust	Cloud service RFS	Unlimited	Mar 13, 2023 14:49:16 GMT+08:00	--	Authorize | Modify | Delete

- If the agency is found, skip the following steps.
- If the agency is not found, perform the following steps to create it.

Step 3 Click **Create Agency** in the upper right corner of the page. On the displayed page, enter **rf_admin_trust** for **Agency Name**, select **Cloud service** for **Agency Type**, select **RFS** for **Cloud Service**, and click **Next**.

Figure 3-4 Create Agency



* Agency Name: rf_admin_trust

* Agency Type:
○ Account
Delegate another HUAWEI CLOUD account to perform operations on your resources.
● Cloud service
Delegate a cloud service to access your resources in other cloud services.

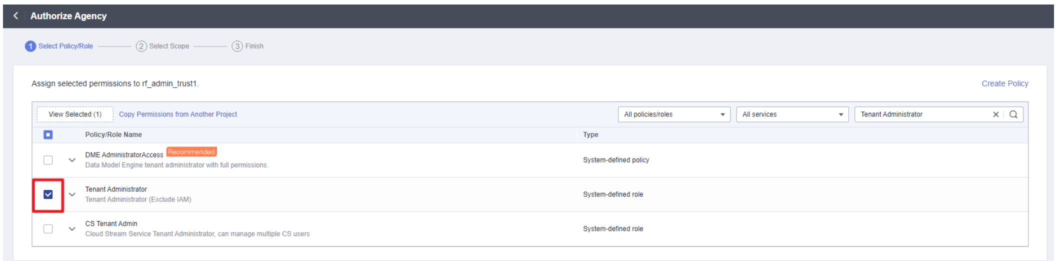
* Cloud Service: RFS

* Validity Period: Unlimited

Description: Enter a brief description.
0/255

Step 4 Search for **Tenant Administrator**, select it in the search results, and click **Next**.

Figure 3-5 Selecting a policy/role



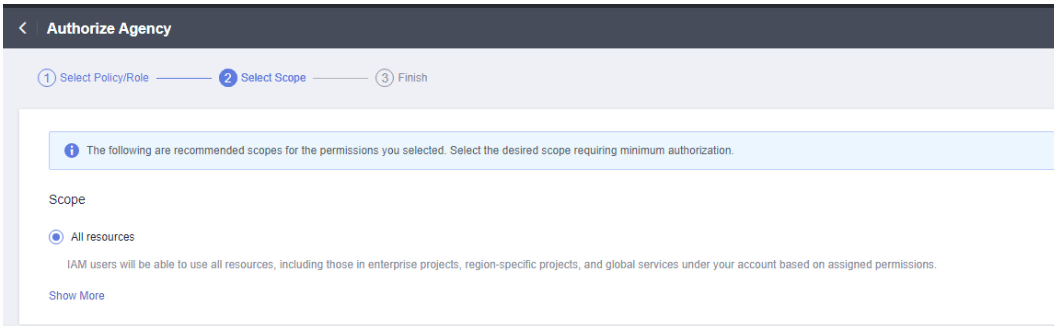
Assign selected permissions to rf_admin_trust1.

Search: Tenant Administrator

	Policy/Role Name	Type
□	DME AdministratorAccess Recommended Data Model Engine tenant administrator with full permissions.	System-defined policy
☑	Tenant Administrator Tenant Administrator (Exclude IAM)	System-defined role
□	CS Tenant Admin Cloud Stream Service Tenant Administrator, can manage multiple CS users	System-defined role

Step 5 Select **All resources** and click **OK**.

Figure 3-6 Selecting a scope



The following are recommended scopes for the permissions you selected. Select the desired scope requiring minimum authorization.

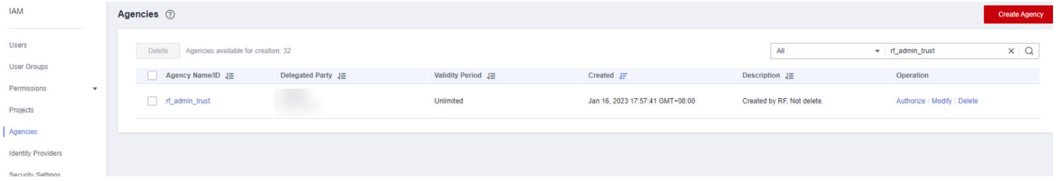
Scope

● All resources

IAM users will be able to use all resources, including those in enterprise projects, region-specific projects, and global services under your account based on assigned permissions.

Step 6 If **rf_admin_trust** is displayed in the agency list, the agency has been created.

Figure 3-7 Agency list



Agencies available for creation: 32

Agency Name/ID	Delegated Party	Validity Period	Created	Description	Operation
rf_admin_trust	[redacted]	Unlimited	Jan 16, 2023 17:57:41 GMT+08:00	Created by RF, Not delete.	Authorize | Modify | Delete

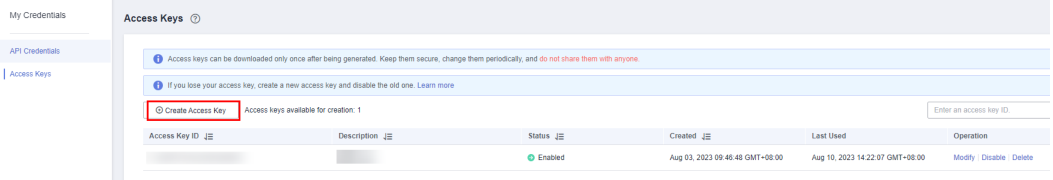
----End

Obtaining an Access Key (AK/SK)

Before deploying this solution, you need to obtain the AK/SK from the Huawei Cloud console and then configure parameters listed in **Step 3**.

On the Huawei Cloud console, hover the mouse pointer over the account name in the upper right corner and choose **My Credentials**. On the **Access Keys** page, create an access key and download it. For details, see **How Do I Obtain an Access Key (AK/SK)?**

Figure 3-8 Creating an access key



Access keys can be downloaded only once after being generated. Keep them secure, change them periodically, and do not share them with anyone.

If you lose your access key, create a new access key and disable the old one. Learn more

Access keys available for creation: 1

Access Key ID	Description	Status	Created	Last Used	Operation
[redacted]	[redacted]	Enabled	Aug 03, 2023 09:46:48 GMT+08:00	Aug 10, 2023 14:22:07 GMT+08:00	Modify | Disable | Delete

3.2 Quick Deployment

This section describes how to quickly deploy this solution.

Table 3-1 Parameters required for deploying this solution

Parameter	Type	Mandatory	Description	Default Value
vpc_name	String	Yes	Virtual Private Cloud (VPC) name. This template uses a newly created VPC and the VPC name must be unique. The value can contain 1 to 54 characters, including letters, digits, underscores (_), hyphens (-). and periods (.).	high-availability-aigc-applications-demo

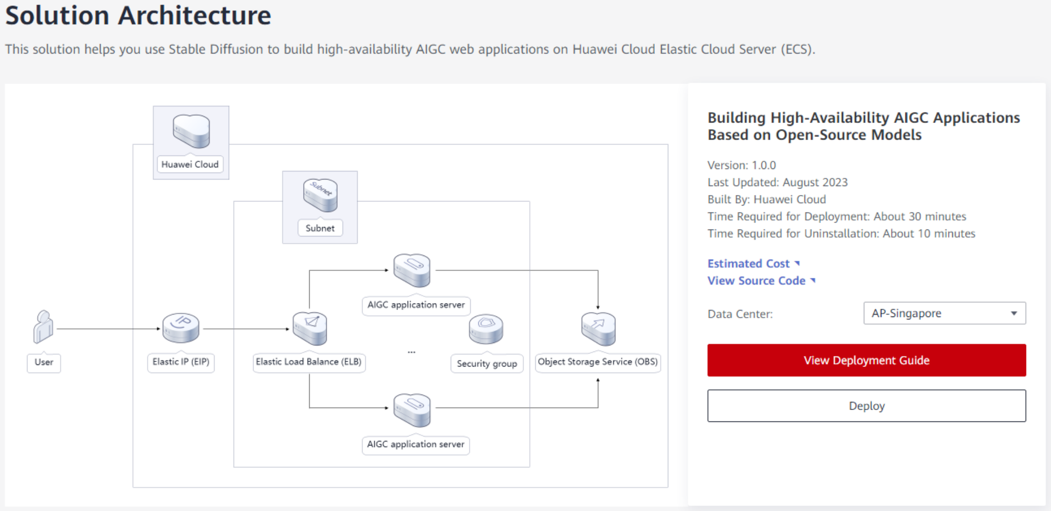
Parameter	Type	Mandatory	Description	Default Value
security_group_name	String	Yes	Security group name. This template uses a newly created security group. The value can contain 1 to 64 characters, including letters, digits, underscores (_), hyphens (-), and periods (.).	high-availability-aigc-applications-demo
ecs_name	String	Yes	ECS name, which must be unique. The name format is {ecs_name}-digit. It can contain 1 to 60 characters, including letters, digits, underscores (_), hyphens (-). and periods (.).	high-availability-aigc-applications-demo
image_bucket_name	String	Yes	OBS bucket name, which is globally unique. The bucket is used to store automatically uploaded images saved on the web UI. The bucket name can contain 3 to 63 characters, including lowercase letters, digits, hyphens (-), and periods (.). Do not start or end with a hyphen (-) or a period (.).	None
ecs_count	String	Yes	Number of ECSs, which is greater than or equal to 1. The maximum number of ECSs is determined by the user quota listed in **My Quotas**.	2
ecs_flavor	String	Yes	ECS flavor. This template uses a GPU-accelerated flavor. For details about flavors, see **A Summary List of x86 ECS Specifications**.	pi2.2xlarge.4

Parameter	Type	Mandatory	Description	Default Value
ecs_password	String	Yes	Initial password of the ECS. After the ECS is created, reset the password by referring to **Step 1**. It can contain 8 to 26 characters, including at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters (!@$%^-_=+[{()}]:,./?~#*). For Windows ECSs, the password cannot contain the username, the username spelled backwards, or more than two consecutive characters in the username. The default administrator account is root.	None
elb_name	String	Yes	ELB name, which can contain 1 to 64 characters, including letters, digits, underscores (_), hyphens (-), and periods (.).	high-availability-aigc-applications-demo
eip_bandwidth_size	Number	Yes	EIP bandwidth, which is billed by traffic. Value range: 1-300 Mbit/s	300 Mbit/s
charging_mode	String	Yes	Billing mode. By default, expenses are automatically deducted. The value can be **prePaid** (yearly/monthly) or **postPaid** (pay-per-use).	postPaid
charge_period_unit	String	Yes	Unit of a subscription period. This parameter is valid and mandatory only when **charging_mode** is set to **prePaid**. The value can be **month** or **year**.	month

Parameter	Type	Mandatory	Description	Default Value
charge_period	Number	Yes	Subscription period. This parameter is valid and mandatory only when **charging_mode** is set to **prePaid**. Value range: If **charge_period_unit** is set to **month**, the value range is from **1** to **9**. If **charge_period_unit** is set to **year**, the value range is from **1** to **3**.	1
access_key_id	String	Yes	Access Key ID (AK), which is used to verify the identity of a user attempting to upload generated images to the OBS bucket. For details about how to obtain the AK, see **Obtaining an Access Key (AK/SK)**.	None
secret_access_key	String	Yes	Secret Access Key (SK), which is used to sign requests. It must be used together with the AK to authenticate image upload requests. For details about how to obtain the SK, see **Obtaining an Access Key (AK/SK)**.	None

Step 1 Log in to **Practical Application of Huawei Cloud Solutions** and select **Building High-Availability AIGC Applications Based on Open-Source Models**.

Figure 3-9 Selecting a solution



Solution Architecture

This solution helps you use Stable Diffusion to build high-availability AIGC web applications on Huawei Cloud Elastic Cloud Server (ECS).

Building High-Availability AIGC Applications Based on Open-Source Models

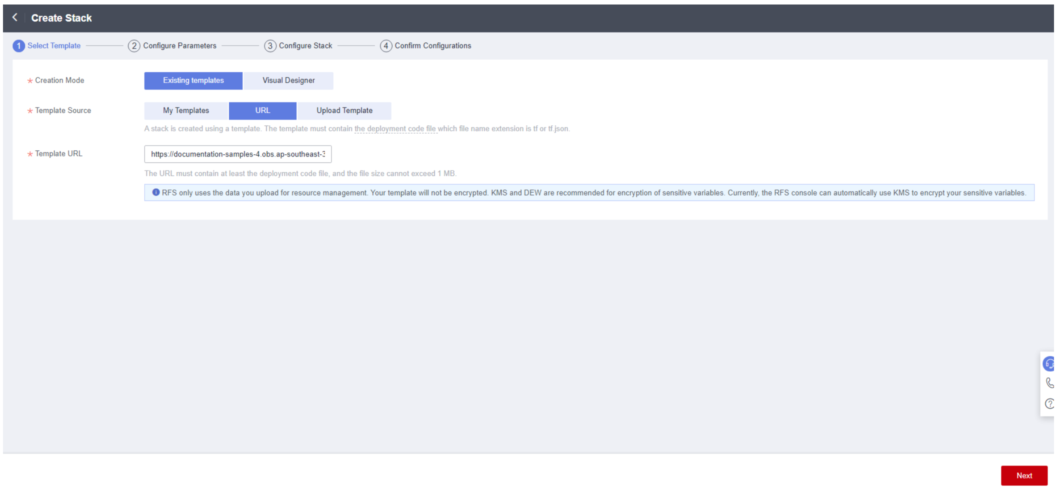
Version: 1.0.0
Last Updated: August 2023
Built By: Huawei Cloud
Time Required for Deployment: About 30 minutes
Time Required for Uninstallation: About 10 minutes

Estimated Cost
View Source Code

Data Center: AP-Singapore

Step 2 Click **Deploy** to switch to the **Create Stack** page.

Figure 3-10 Creating a stack



* Creation Mode: Existing templates | Visual Designer

* Template Source: My Templates | URL | Upload Template

A stack is created using a template. The template must contain the deployment code file which file name extension is tf or tf.json.

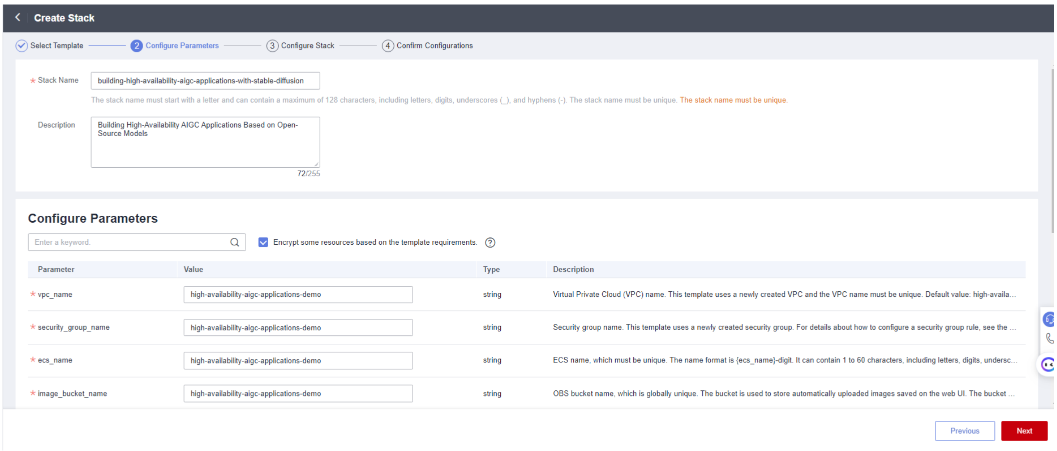
* Template URL: https://documentation-samples-4.obs.ap-southeast-3

The URL must contain at least the deployment code file, and the file size cannot exceed 1 MB.

RFS only uses the data you upload for resource management. Your template will not be encrypted. KMS and DEW are recommended for encryption of sensitive variables. Currently, the RFS console can automatically use KMS to encrypt your sensitive variables.

Step 3 Click **Next**. On the displayed page, set parameters by referring to **Table 3-1** and click **Next**.

Figure 3-11 Configuring parameters



* Stack Name: building-high-availability-aigc-applications-with-stable-diffusion

The stack name must start with a letter and can contain a maximum of 128 characters, including letters, digits, underscores (_), and hyphens (-). The stack name must be unique. The stack name must be unique.

Description: Building High-Availability AIGC Applications Based on Open-Source Models

72/255

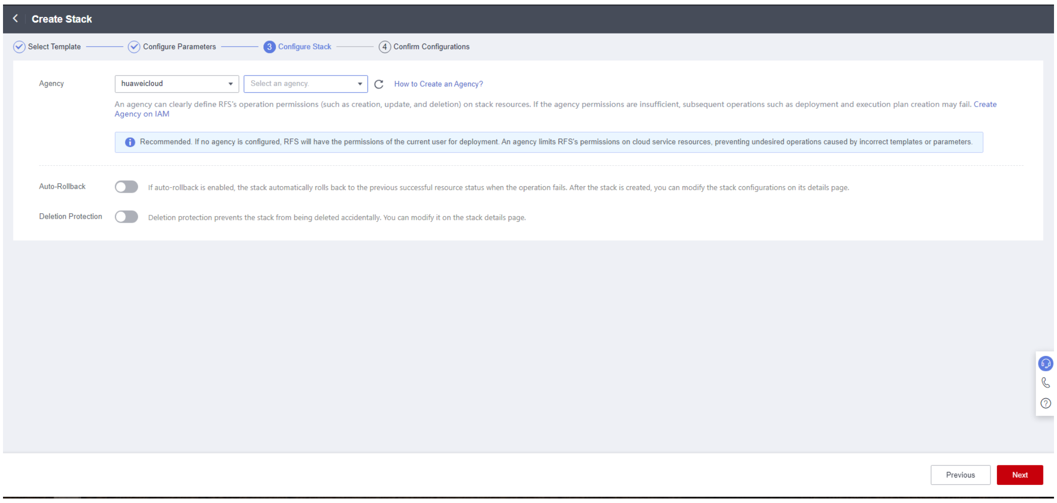
Configure Parameters

☑ Encrypt some resources based on the template requirements.

Parameter	Value	Type	Description
* vpc_name	high-availability-aigc-applications-demo	string	Virtual Private Cloud (VPC) name. This template uses a newly created VPC and the VPC name must be unique. Default value: high-availa...
* security_group_name	high-availability-aigc-applications-demo	string	Security group name. This template uses a newly created security group. For details about how to configure a security group rule, see the ...
* ecs_name	high-availability-aigc-applications-demo	string	ECS name, which must be unique. The name format is {ecs_name}-digit. It can contain 1 to 60 characters, including letters, digits, undersc...
* image_bucket_name	high-availability-aigc-applications-demo	string	OBS bucket name, which is globally unique. The bucket is used to store automatically uploaded images saved on the web UI. The bucket ...

Step 4 (Optional) On the **Configure Stack** page, select **rf_admin_trust** from the agency drop-down list and click **Next**.

Figure 3-12 Configuring a stack



Agency: huaweicloud | Select an agency. | How to Create an Agency?

An agency can clearly define RFS's operation permissions (such as creation, update, and deletion) on stack resources. If the agency permissions are insufficient, subsequent operations such as deployment and execution plan creation may fail. Create Agency on IAM

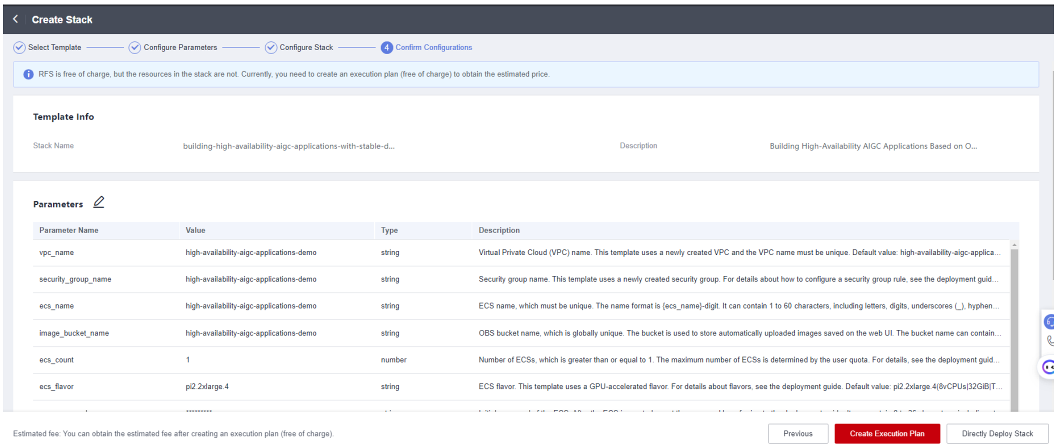
Recommended. If no agency is configured, RFS will have the permissions of the current user for deployment. An agency limits RFS's permissions on cloud service resources, preventing undesired operations caused by incorrect templates or parameters.

Auto-Rollback: If auto-rollback is enabled, the stack automatically rolls back to the previous successful resource status when the operation fails. After the stack is created, you can modify the stack configurations on its details page.

Deletion Protection: Deletion protection prevents the stack from being deleted accidentally. You can modify it on the stack details page.

Step 5 On the **Confirm Configurations** page, click **Create Execution Plan**.

Figure 3-13 Creating an execution plan



RFS is free of charge, but the resources in the stack are not. Currently, you need to create an execution plan (free of charge) to obtain the estimated price.

Template Info

Stack Name: building-high-availability-aigc-applications-with-stable-d...

Description: Building High-Availability AIGC Applications Based on O...

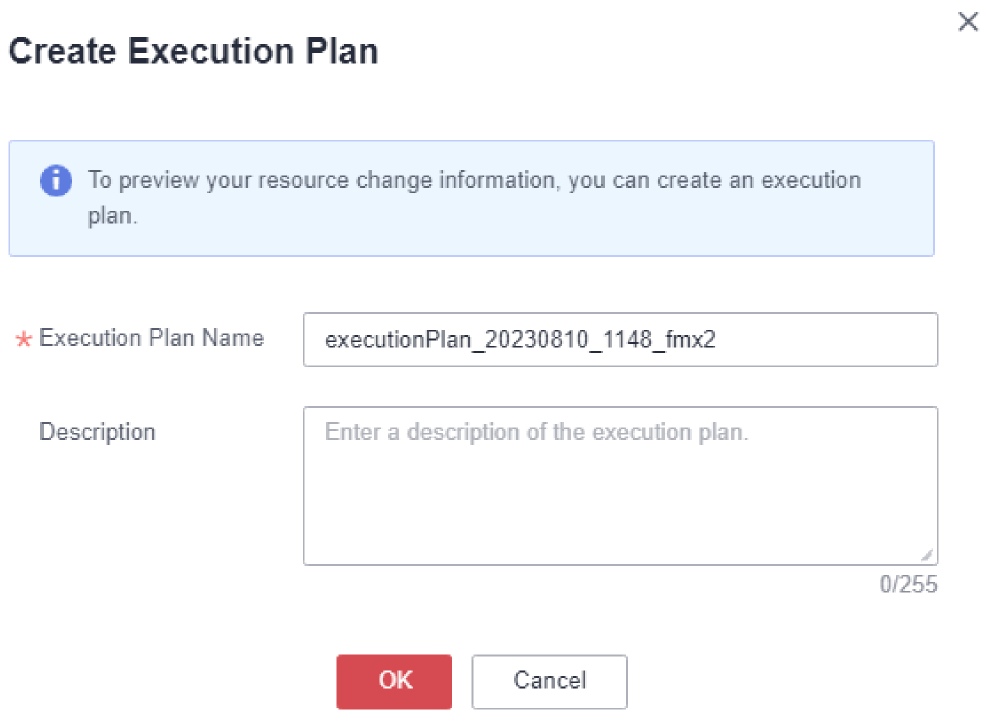
Parameters

Parameter Name	Value	Type	Description
vpc_name	high-availability-aigc-applications-demo	string	Virtual Private Cloud (VPC) name. This template uses a newly created VPC and the VPC name must be unique. Default value: high-availability-aigc-applica...
security_group_name	high-availability-aigc-applications-demo	string	Security group name. This template uses a newly created security group. For details about how to configure a security group rule, see the deployment guid...
ecs_name	high-availability-aigc-applications-demo	string	ECS name, which must be unique. The name format is {ecs_name}-digit. It can contain 1 to 60 characters, including letters, digits, underscores (_), hyphen...
image_bucket_name	high-availability-aigc-applications-demo	string	OBS bucket name, which is globally unique. The bucket is used to store automatically uploaded images saved on the web UI. The bucket name can contain...
ecs_count	1	number	Number of ECSs, which is greater than or equal to 1. The maximum number of ECSs is determined by the user quota. For details, see the deployment guid...
ecs_flavor	pi2.2xlarge.4	string	ECS flavor. This template uses a GPU-accelerated flavor. For details about flavors, see the deployment guide. Default value: pi2.2xlarge.4(8vCPUs|32GiB|T...

Estimated fee: You can obtain the estimated fee after creating an execution plan (free of charge).

Step 6 In the displayed **Create Execution Plan** dialog box, enter an execution plan name and click **OK**.

Figure 3-14 Creating an execution plan



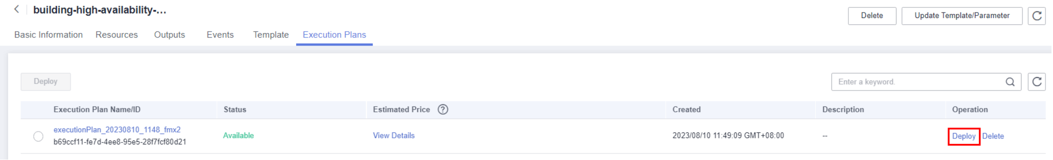
Create Execution Plan

To preview your resource change information, you can create an execution plan.

* Execution Plan Name	executionPlan_20230810_1148_fmx2
Description	Enter a description of the execution plan. 0/255

Step 7 On the **Execution Plans** tab, click **Deploy**. In the displayed dialog box, click **Execute**.

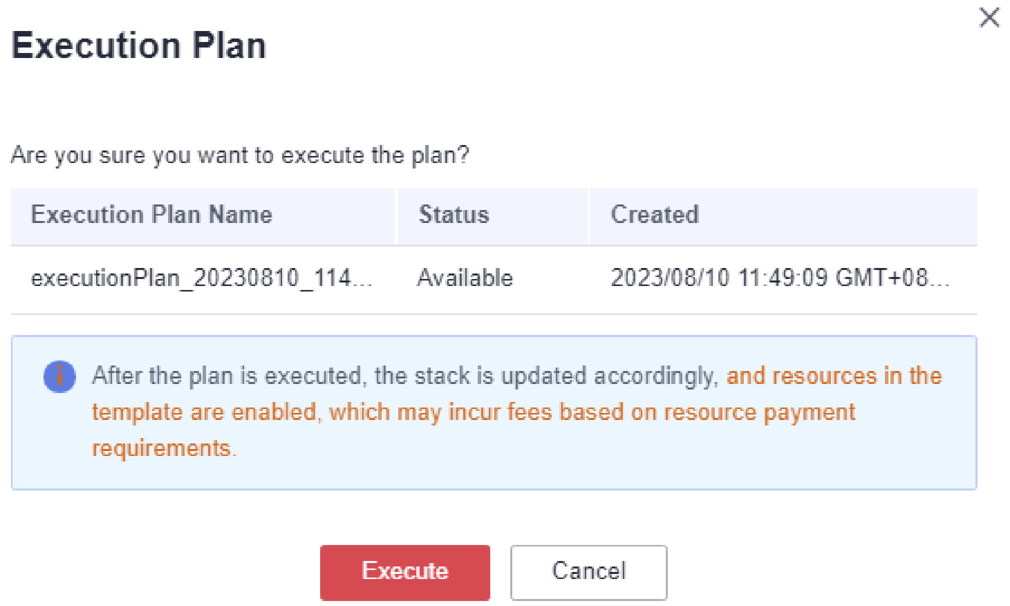
Figure 3-15 Deploying the execution plan



building-high-availability-...

	Execution Plan Name/ID	Status	Estimated Price	Created	Description	Operation
○	executionPlan_20230810_1148_fmx2 b69ccf11-fe7d-4ee8-95e5-28f7fcf80d21	Available	View Details	2023/08/10 11:49:09 GMT+08:00	--	Deploy Delete

Figure 3-16 Confirming the deployment



Execution Plan

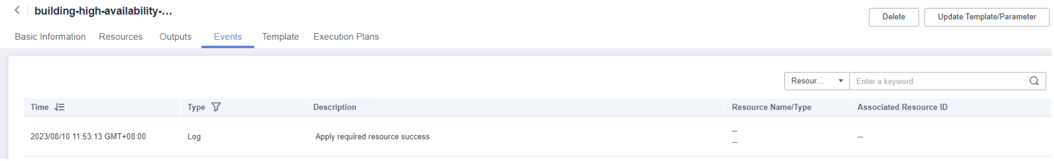
Are you sure you want to execute the plan?

Execution Plan Name	Status	Created
executionPlan_20230810_114...	Available	2023/08/10 11:49:09 GMT+08...

After the plan is executed, the stack is updated accordingly, and resources in the template are enabled, which may incur fees based on resource payment requirements.

Step 8 Wait until the deployment is successful and click the **Events** tab to view details.

Figure 3-17 Successful deployment

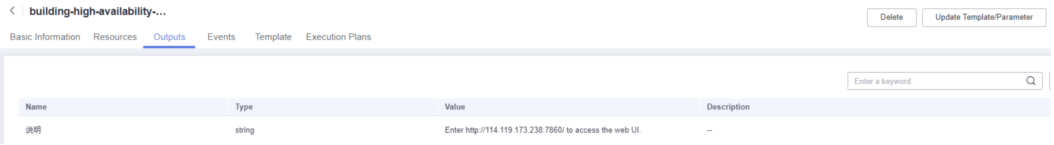


building-high-availability-...

Time	Type	Description	Resource Name/Type	Associated Resource ID
2023/08/10 11:53:13 GMT+08:00	Log	Apply required resource success.	-- --	--

Step 9 Refresh the page and view the web UI access description on the **Outputs** tab.

Figure 3-18 Outputs



building-high-availability-...

Name	Type	Value	Description
说明	string	Enter http://114.119.173.238:7860/ to access the web UI.	--

----End

3.3 Getting Started

(Optional) Modifying Security Group Rules

NOTICE

By default, this solution creates the security group rule that uses the ping command to test ECS connectivity. To remotely log in to an ECS in the security group, you need to add an inbound rule, for example, by setting the login port to 3389 and adding a whitelist IP address.

A security group is a collection of access control rules for cloud resources, such as cloud servers, containers, and databases, to control inbound and outbound traffic. Cloud resources associated with the same security group have the same security requirements and are mutually trusted within a VPC.

You can modify security group rules, for example, by adding, modifying, or deleting a TCP port.

- Adding a security group rule: **Add an inbound rule** and enable a TCP port if needed.
- Modifying a security group rule: Inappropriate security group settings can be a serious security risk. You can **modify security group rules** to improve the network security of ECSs.
- Deleting a security group rule: If the source or destination IP address of an inbound or outbound security group rule changes, or a port needs to be disabled, you can **delete the security group rule**.

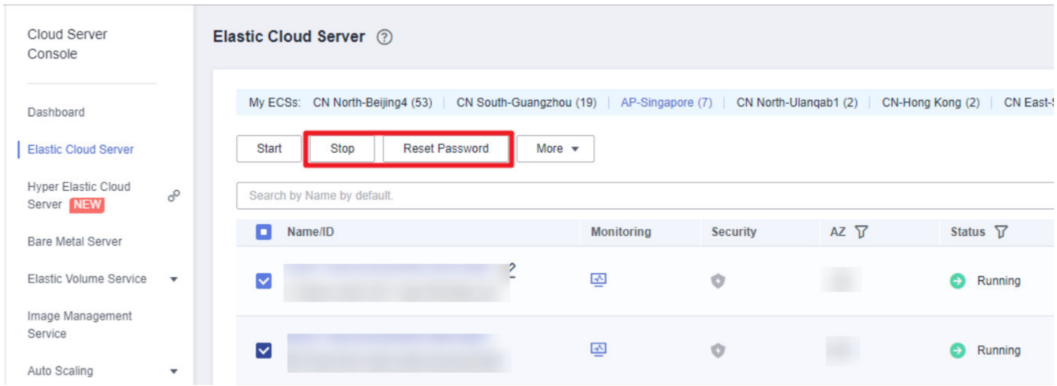
(Optional) Configuring a Domain Name for an Application

Configure domain name resolution to resolve the website domain name to the IP address displayed in the figure in **Step 9**. In this way, the website can be accessed over its domain name. For details about DNS resolution, see **Public Domain Name Resolution**.

Using the AIGC Web UI Application

Step 1 (Optional) Log in to the **ECS console**, select the created ECSs and click **Stop** above the ECS list. After the ECSs are stopped, click **Reset Password**, enter a new password, and click **OK**. The new password will take effect after the ECSs are started.

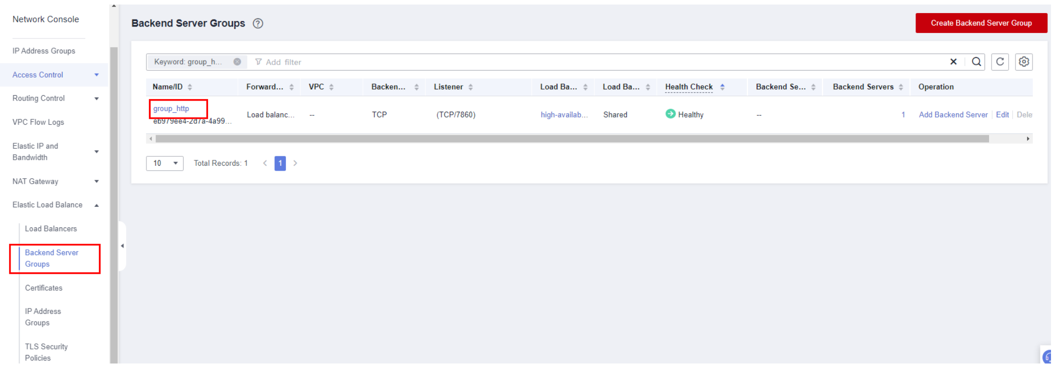
Figure 3-19 Resetting the password


Cloud Server Console
Dashboard
Elastic Cloud Server
Hyper Elastic Cloud Server NEW
Bare Metal Server
Elastic Volume Service
Image Management Service
Auto Scaling
Elastic Cloud Server
My ECSs: CN North-Beijing4 (53) | CN South-Guangzhou (19) | AP-Singapore (7) | CN North-Ulanqab1 (2) | CN-Hong Kong (2) | CN East-
Start
Stop
Reset Password
More
Search by Name by default.

■	Name/ID	Monitoring	Security	AZ	Status
☑	[redacted]			[redacted]	Running
☑	[redacted]			[redacted]	Running

Step 2 Log in to the **ELB console** and choose **Backend Server Groups** from the left navigation pane. Click the target backend server group name to view its details. On the **Backend Servers** tab, check whether servers are in the **Healthy** state. (Service initialization will be completed 20 minutes after this solution is deployed based on default settings. All backend servers will be healthy on port 7860 then.)

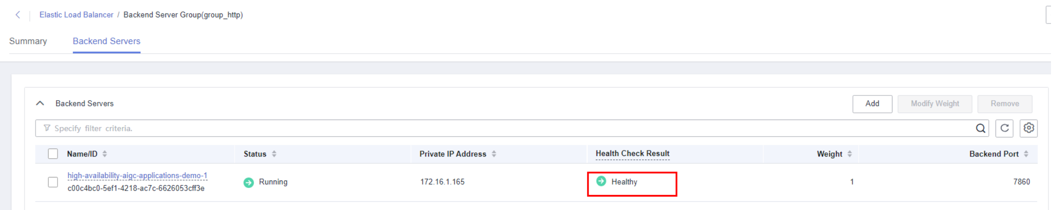
Figure 3-20 Backend Server Groups



Name/ID	Forward...	VPC	Backen...	Listener	Load Ba...	Load Ba...	Health Check	Backend Se...	Backend Servers	Operation
group_http eb979ee4-2d7a-4a99...	Load balanc...	--	TCP	(TCP/7860)	high-availab...	Shared	Healthy	--	1	Add Backend Server | Edit | Dele

10 Total Records: 1 1

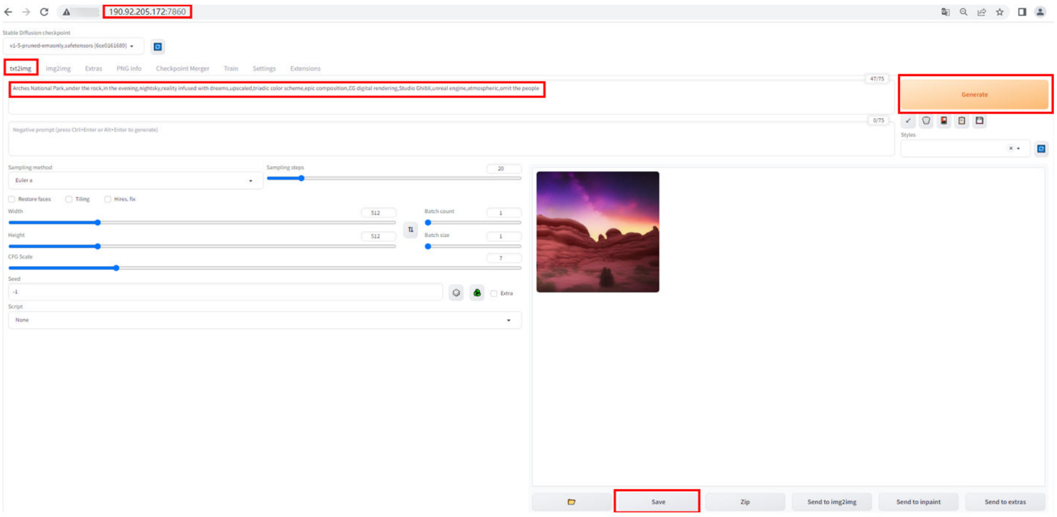
Figure 3-21 Backend server status



□	Name/ID	Status	Private IP Address	Health Check Result	Weight	Backend Port
□	high-availability-aigc-applications-demo-1 c00c4bc0-5ef1-4218-ac7c-6626053cff3e	Running	172.16.1.165	Healthy	1	7860

Step 3 Use the URL in **Step 9** to access the AIGC web UI. Click **txt2img**, input text in the box, and click **Generate**. After the image is generated, click **Save**. For details about how to use the Stable Diffusion web UI, see **stable-diffusion-webui** or search for tutorials on the Internet. This solution creates the user **aigc** with the default password **aigc@123**.

Figure 3-22 AIGC web UI



190.92.205.172:7860

Stable Diffusion checkpoint: v1-5-pruned-emaonly.safetensors [6ce0161689]

Arches National Park,under the rock,in the evening,nightsky,reality infused with dreams,upscaled,triadic color scheme,epic composition,CG digital rendering,Studio Ghibli,unreal engine,atmospheric,omit the people

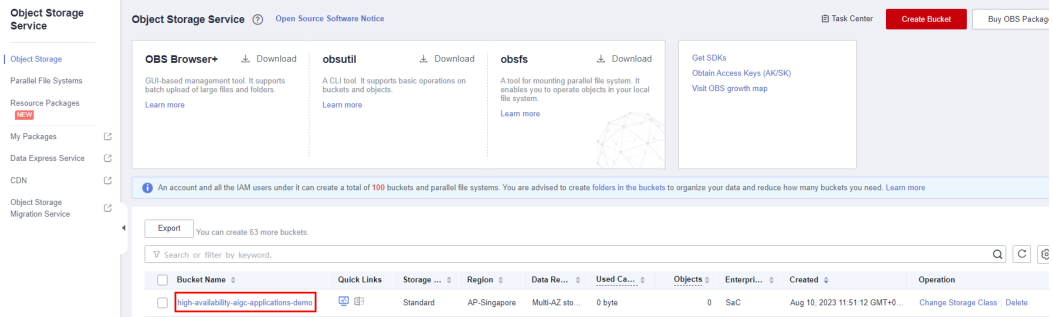
Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Sampling method: Euler a
Sampling steps: 20
□ Restore faces □ Tiling □ Hires. fix
Width: 512
Height: 512
Batch count: 1
Batch size: 1
CFG Scale: 7
Seed: -1
Script: None

Example text prompt:
Arches National Park,under the rock,in the evening,nightsky,reality infused with dreams,upscaled,triadic color scheme,epic composition,CG digital rendering,Studio Ghibli,unreal engine,atmospheric,omit the people

Step 4 On the **OBS console**, click the OBS bucket created in **Step 3** to view the saved images. You can also click **Share** to share the images. For more OBS functions, see **Object Management**.

Figure 3-23 OBS bucket list



GUI-based management tool. It supports batch upload of large files and folders.

A CLI tool. It supports basic operations on buckets and objects.

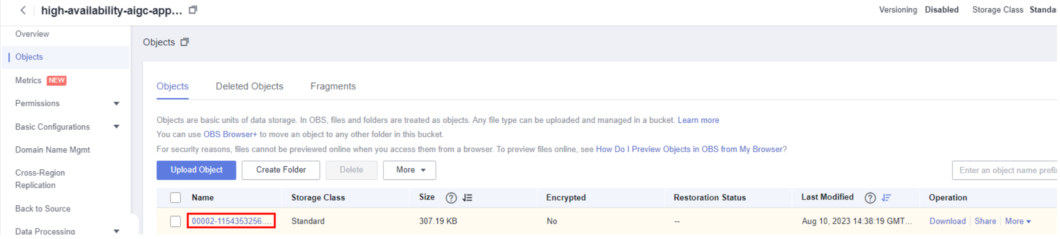
A tool for mounting parallel file system. It enables you to operate objects in your local file system.

An account and all the IAM users under it can create a total of 100 buckets and parallel file systems. You are advised to create folders in the buckets to organize your data and reduce how many buckets you need. Learn more

You can create 63 more buckets.

Bucket Name	Quick Links	Storage ...	Region	Data Re...	Used Ca...	Objects	Enterpri...	Created	Operation
high-availability-aigc-applications-demo		Standard	AP-Singapore	Multi-AZ sto...	0 byte	0	SaC	Aug 10, 2023 11:51:12 GMT+0...	Change Storage Class | Delete

Figure 3-24 Viewing the saved images



Objects are basic units of data storage. In OBS, files and folders are treated as objects. Any file type can be uploaded and managed in a bucket. Learn more

You can use OBS Browser+ to move an object to any other folder in this bucket.

For security reasons, files cannot be previewed online when you access them from a browser. To preview files online, see How Do I Preview Objects in OBS from My Browser?

Name	Storage Class	Size	Encrypted	Restoration Status	Last Modified	Operation
00002-1154353256....	Standard	307.19 KB	No	--	Aug 10, 2023 14:38:19 GMT...	Download | Share | More

NOTE

This solution has added inotify-tools and OBS obsutil to run automatically at startup in ECSs, so the images you saved on the web UI can be automatically uploaded to the OBS bucket. You can also right-click on the web UI and choose **Save as** to save the images locally. AIGC can run automatically at ECS startup.

```
Example command for starting AIGC:
Start in the foreground:
cd /home/aigc && sudo -u aigc bash -c "source /home/aigc/webui.sh --listen --port 7860 --api --enable-insecure-extension-access"
Start in the background:
cd /home/aigc && sudo -u aigc bash -c "source /home/aigc/webui.sh --listen --port 7860 --api --enable-insecure-extension-access &">> /home/aigc/aigc-applications.log (change the storage path as needed)
```

----End

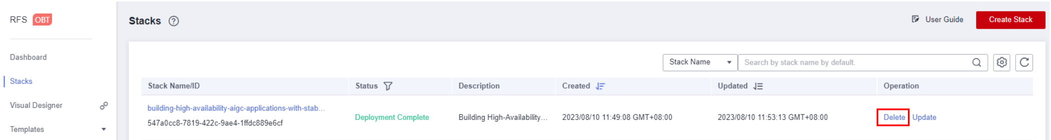
3.4 Quick Uninstallation

NOTICE

If there is data stored in the OBS bucket, the solution cannot be uninstalled. To uninstall this solution, back up the data and empty the bucket first.

Step 1 Click **Delete** in the row of the solution stack.

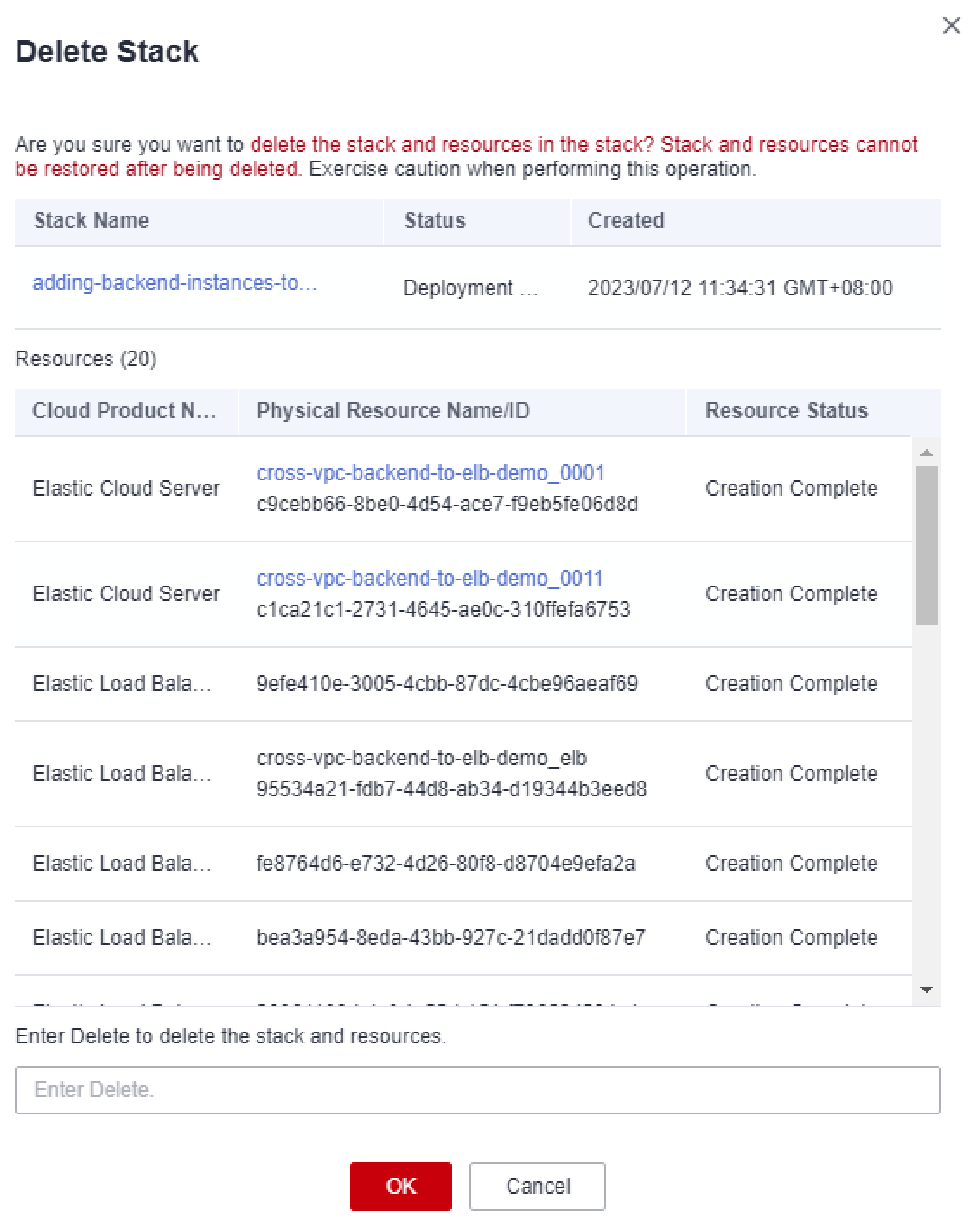
Figure 3-25 Deleting the stack



Stack Name/ID	Status	Description	Created	Updated	Operation
building-high-availability-aigc-applications-with-stab... 547a0cc8-7819-422c-9ae4-1ffdc889e6cf	Deployment Complete	Building High-Availability...	2023/08/10 11:49:08 GMT+08:00	2023/08/10 11:53:13 GMT+08:00	Delete Update

Step 2 Enter **Delete** and click **OK**.

Figure 3-26 Confirming the uninstallation



Delete Stack

Are you sure you want to delete the stack and resources in the stack? Stack and resources cannot be restored after being deleted. Exercise caution when performing this operation.

Stack Name	Status	Created
adding-backend-instances-to...	Deployment ...	2023/07/12 11:34:31 GMT+08:00

Resources (20)

Cloud Product N...	Physical Resource Name/ID	Resource Status
Elastic Cloud Server	cross-vpc-backend-to-elb-demo_0001 c9cebb66-8be0-4d54-ace7-f9eb5fe06d8d	Creation Complete
Elastic Cloud Server	cross-vpc-backend-to-elb-demo_0011 c1ca21c1-2731-4645-ae0c-310ffefa6753	Creation Complete
Elastic Load Bala...	9efe410e-3005-4cbb-87dc-4cbe96aeaf69	Creation Complete
Elastic Load Bala...	cross-vpc-backend-to-elb-demo_elb 95534a21-fdb7-44d8-ab34-d19344b3eed8	Creation Complete
Elastic Load Bala...	fe8764d6-e732-4d26-80f8-d8704e9efa2a	Creation Complete
Elastic Load Bala...	bea3a954-8eda-43bb-927c-21dadd0f87e7	Creation Complete

Enter Delete to delete the stack and resources.

Enter Delete.

----End

4 Appendix

Terms

Basic concepts and cloud service introduction

- Elastic Cloud Server (ECS): a scalable and on-demand cloud server. It helps you to efficiently set up reliable, secure, and flexible application environments, ensuring stable service running and improving O&M efficiency.
- Elastic Load Balance (ELB): automatically distributes incoming traffic across multiple servers to balance their workloads, increasing service capabilities and fault tolerance of your applications.
- Elastic IP (EIP): enables your cloud resources to communicate with the Internet using static public IP addresses and scalable bandwidths. EIPs can be bound to or unbound from ECSs, BMSs, virtual IP addresses, load balancers, and NAT gateways.
- Virtual Private Cloud (VPC): an isolated and private virtual network environment. You can configure IP address segments, subnets, and security groups, assign EIPs, and allocate bandwidths in a VPC.
- Object Storage Service (OBS): a secure, highly reliable object storage service that allows you to inexpensively store any amount of data.
- Security group: a collection of access control rules for ECSs that have the same security protection requirements and are mutually trusted in a VPC. You can define inbound and outbound rules to control traffic to and from the ECSs in a security group.
- inotify-tools: a command-line tool in Linux to monitor file system changes and trigger corresponding operations.

5 Change History

Released On	Description
2023-08-10	This issue is the first official release.