Solution

Deploying Hermes Agents

Issue 1.0.0

Date 2026-06-17

HUAWEI

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud 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

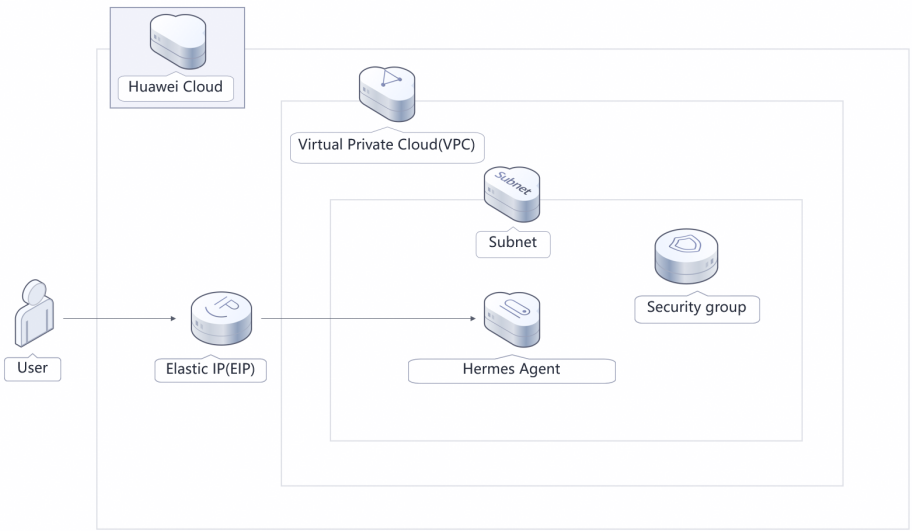
Application Scenarios

This solution aims to efficiently deploy and run Hermes Agent using Huawei Cloud Flexus Cloud Server X Instance. Hermes Agent is an open-source intelligent agent framework built on large language models, designed to enable developers to build, deploy, and manage intelligent agents with autonomous decision-making capabilities more easily.

Architecture

This solution helps you quickly deploy Hermes Agent on Huawei Cloud Flexus Cloud Server X Instance.

Figure 1-1 Architecture



This solution will deploy the following resources:

- An **Elastic IP (EIP)** will be created to provide the ability to access and be accessed from the public network.
- One **Flexus Cloud Server X Instance** (**Elastic Cloud Server, ECS**) will be created to deploy the Hermes Agent.
- A security group will be created. By configuring security group rules, it will provide security protection for the cloud server.

Advantages

- Efficient

  Provides modular components and visual drag-and-drop development tools, enabling developers to skip coding from scratch, significantly shortening product time-to-market, and delivering a "what-you-see-is-what-you-get" agile development experience.
- Flexible Expansion

  Supports plugin architecture and API interfaces for seamless integration with existing systems, adaptable to diverse scenarios such as customer service, marketing, and decision-making.
- One-Click Deployment

  Easily deploy with one click to complete cloud server setup and Hermes Agent intelligent assistant configuration.

Constraints

- Before deploying this solution, you need to register a Huawei account, activate Huawei Cloud, complete real-name verification, and ensure the account is not in arrears or frozen. If you select the "Yearly/Monthly" billing mode, please ensure sufficient account balance for automatic payment during one-click resource deployment, or during the one-click deployment process, go to the billing center, find the "Pending Orders" and complete the payment manually.

2 Resource Planning and Costs

This solution primarily deploys the following resources. The cost provided below is for reference only. For specific details, please refer to the official Huawei Cloud **pricing page**, as the actual charges will be based on your bill.

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

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	deploying-hermes-agent-demo-vpc	• Region: AP-Singapore • CIDR Block: 172.16.0.0/16	1	USD 0.00
Subnet	deploying-hermes-agent-demo-subnet	• Region: AP-Singapore • IPv4 CIDR Block: 172.16.1.0/24	1	USD 0.00
SecurityGroup	deploying-hermes-agent-demo-secgroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	1	USD 0.00

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Flexus X Instance	deploying-hermes-agent-demo-ecs	• Region: AP-Singapore • Pay-per-use: $0.07 USD/hour • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.2u.4g | 2vCPUs | 4GB • Image: Ubuntu 22.04 server 64bit • System Disk: High I/O | 100GB	1	USD 68.41
Elastic IP (EIP)	deploying-hermes-agent-demo-eip	• Region: AP-Singapore • Pay-per-use: USD 0.16/GB/hour • Routing Type: Dynamic BGP • Billed By: traffic • Bandwidth: 300Mbit/s	1	USD 0.16/GB/hour
Total	-	-		**USD 68.41 + Public network traffic price**

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

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	deploying-hermes-agent-demo-vpc	• Region: AP-Singapore • CIDR Block: 172.16.0.0/16	1	USD 0.00
Subnet	deploying-hermes-agent-demo-subnet	• Region: AP-Singapore • IPv4 CIDR Block: 172.16.1.0/24	1	USD 0.00
SecurityGroup	deploying-hermes-agent-demo-secgroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	1	USD 0.00

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Flexus X Instance	deploying-hermes-agent-demo-ecs	• Region: AP-Singapore • Billing Mode: Yearly/ Monthly • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.2u.4g | 2vCPUs | 4GB • Image: Ubuntu 22.04 server 64bit • System Disk: High I/O | 100GB	1	USD 48.46
Elastic IP (EIP)	deploying-hermes-agent-demo-eip	• Region: AP-Singapore • Pay-per-use: USD 0.16/GB/ hour • Routing Type: Dynamic BGP • Billed By: traffic • Bandwidth: 300Mbit/s	1	USD 0.16/GB/ hour
Total	-	-		**USD 48.46 + Public network traffic price**

3 Procedure

3.1 Preparations

If you use the account registered when you first used Huawei Cloud, you do not need to perform this preparation. If you use an IAM user account, check whether you are in the admin user group. If you are not in the admin user group, **grant the necessary permissions** to your account and complete the following preparations.

(Optional) Creating the rf_admin_trust Agency

Step 1 Log in to the Huawei Cloud official website, open the **console**, hover over the account name, and choose **Identity and Access Management.**

Figure 3-1 Console page

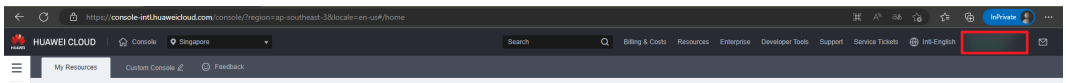
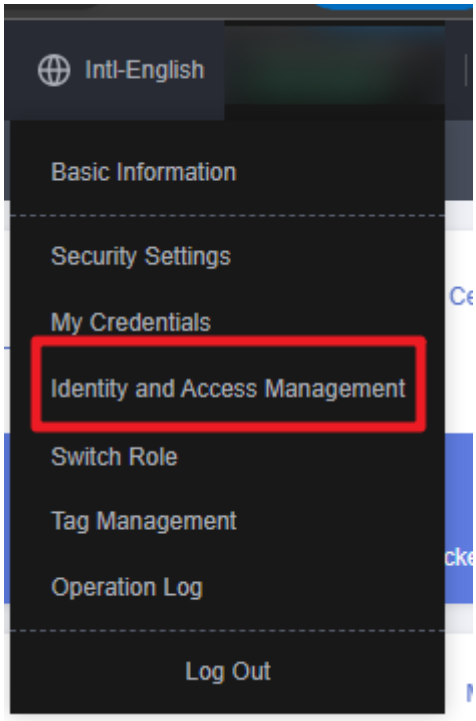
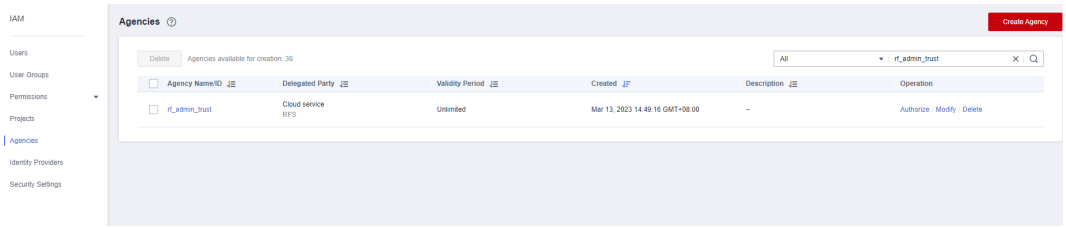


Figure 3-2 Identity and access management page



Step 2 Choose **Agencies** in the left navigation pane and search for the **rf_admin_trust** agency.

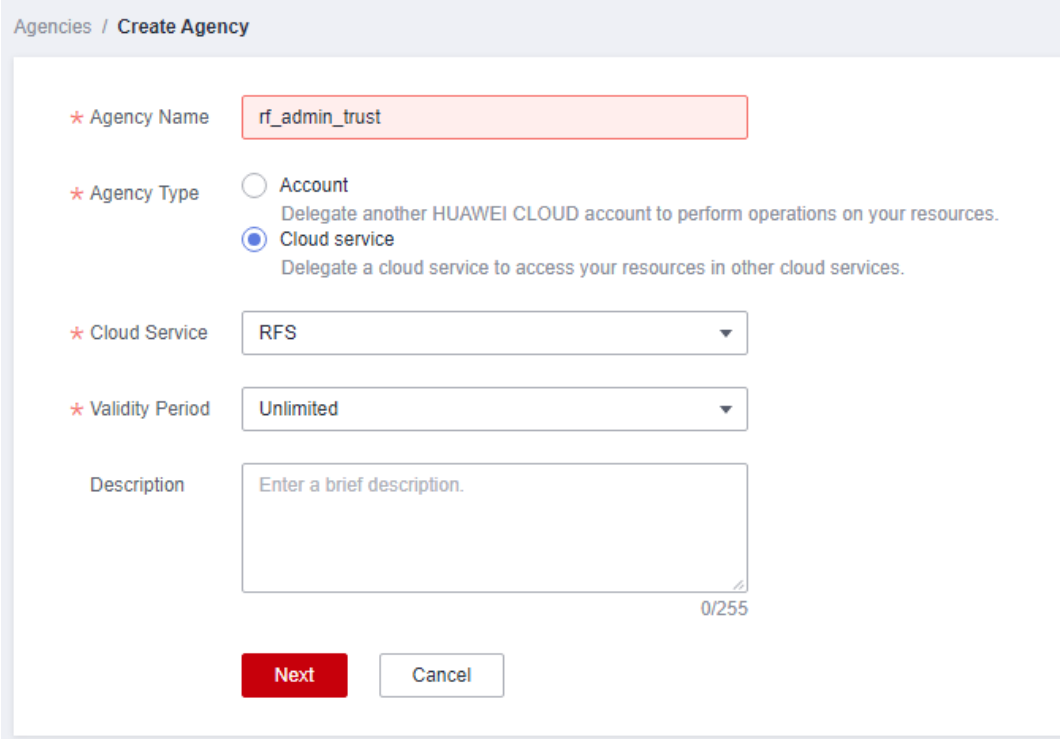
Figure 3-3 Agencies



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

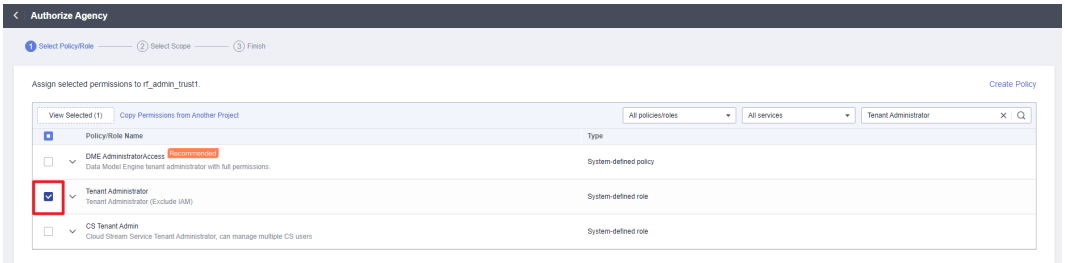
Step 3 Click **Create Agency** in the upper right corner of the page. On the displayed page, enter **rf_admin_trust** for **Agency Name**, select **Cloud service** for **Agency Type** and **RFS** for **Cloud Service**, and click **Next**.

Figure 3-4 Creating the rf_admin_trust agency



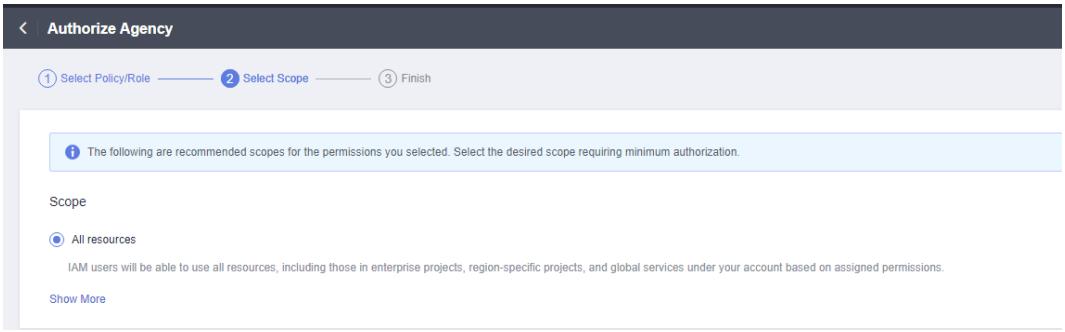
Step 4 Search for **Tenant Administrator**, select it in the search results, and click **Next**.

Figure 3-5 Selecting a policy/role



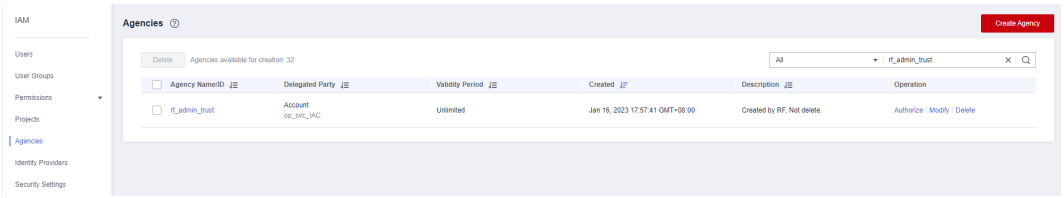
Step 5 Select **All resources** and click **OK**.

Figure 3-6 Setting the authorization scope



Step 6 Check that the **rf_admin_trust agency** is created and displayed in the agency list.

Figure 3-7 Agenciespe



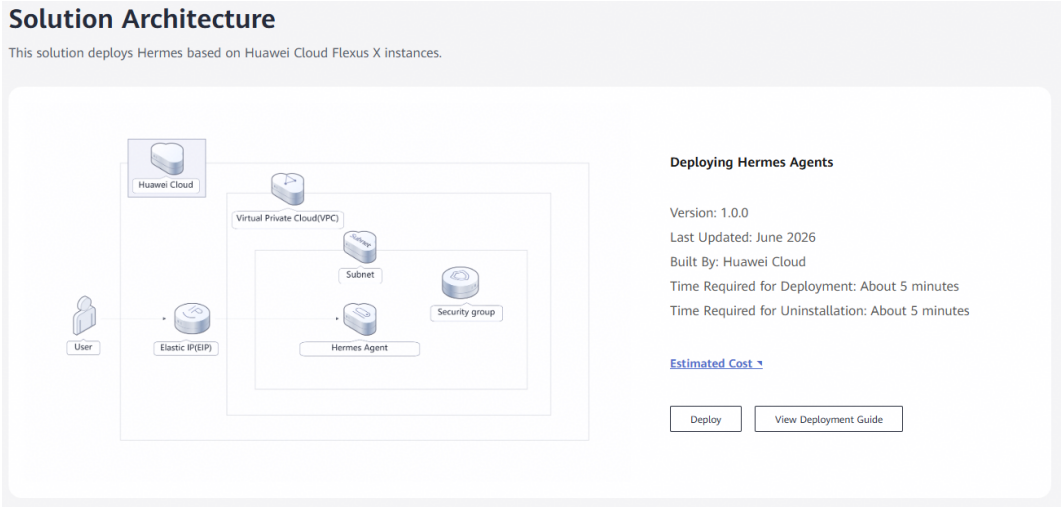
----End

3.2 Quick Deployment

This section helps users efficiently deploy the "Quick Deployment of Hermes Agent" solution. By following the steps and instructions in this section to deploy the solution with one click, you can complete the rapid deployment.

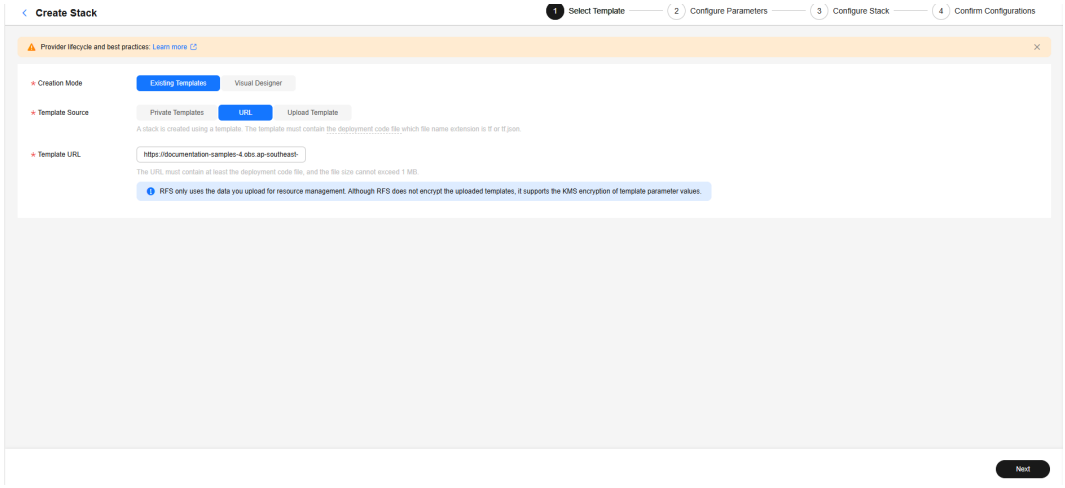
Step 1 Log in to **Huawei Cloud Quick-Start Guides** and choose Deploying Hermes Agents. Click Deploy.

Figure 3-8 Selecting a solution



Step 2 On the **Select Template** page, click **Next**.

Figure 3-9 Selecting a solution



Step 3 On the **Configure Parameters** page, enter a stack name, configure parameters based on **Table1 Parameter description**, and click **Next**.

Figure 3-10 Configuring parameters

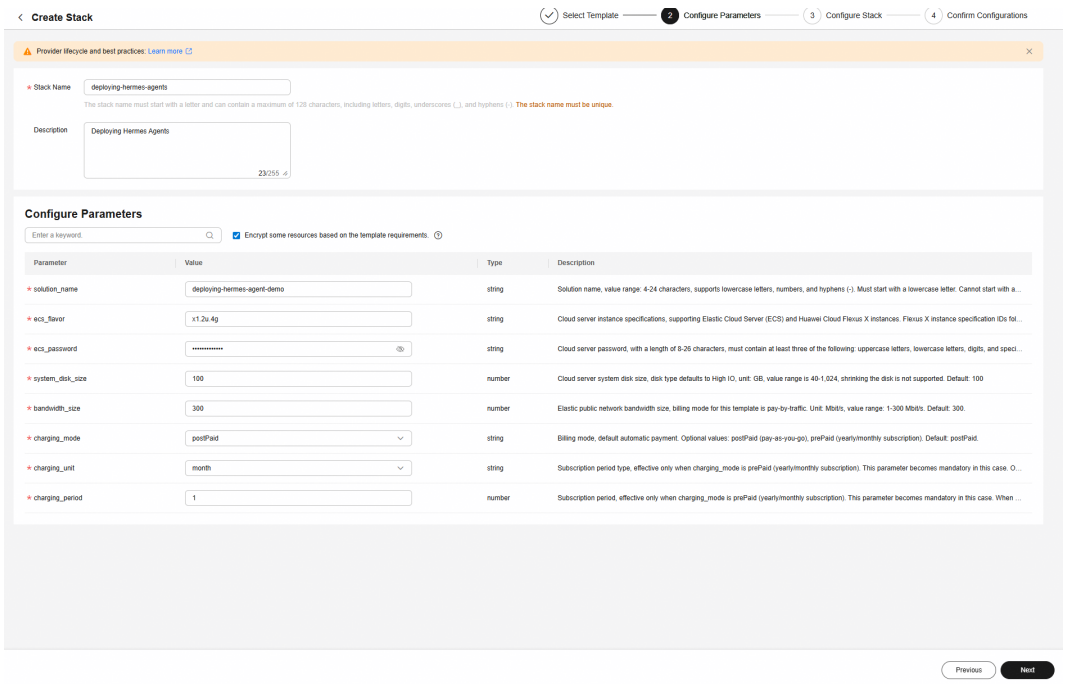


Table 3-1 Parameter descriptions (single-cloud server deployment)

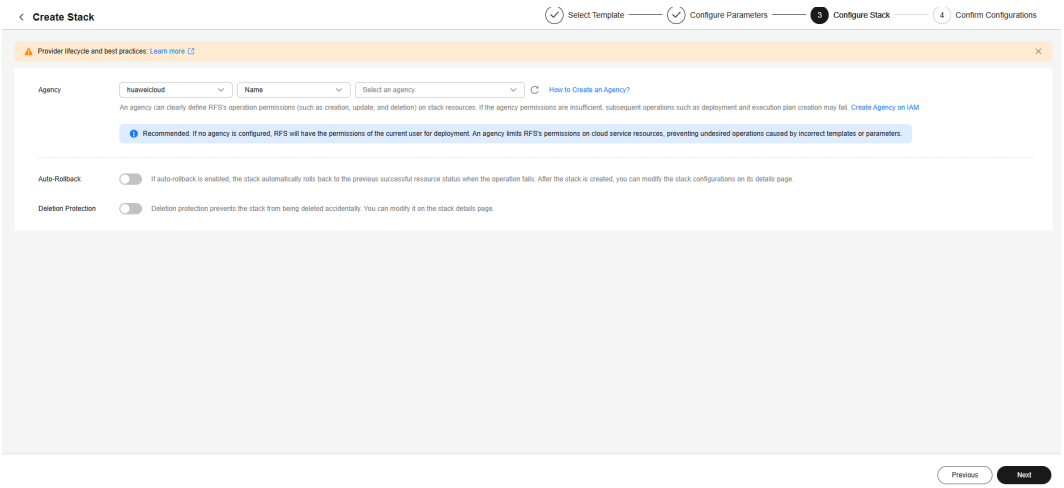
Parameter	Type	Mandatory	Description	Default Value
solution_name	String	Mandatory	Solution name, value range: 4-24 characters, supports lowercase letters, numbers, and hyphens (-). Must start with a lowercase letter. Cannot start with a hyphen (-).	deploying-hermes-agent-demo
ecs_flavor	String	Mandatory	Cloud server instance specifications, supporting Elastic Cloud Server (ECS) and Huawei Cloud Flexus X instances. Flexus X instance specification IDs follow the naming pattern x1.?u.?g, for example, x1.2u.4g for 2vCPUs4GiB. For specific Huawei Cloud Flexus X instance specifications, please refer to the console. For ECS specifications, refer to the deployment guide configuration. see **A Summary List of x86 ECS Specifications**.	x1.2u.4g
ecs_password	String	Mandatory	Cloud server password, with a length of 8-26 characters, must contain at least three of the following: uppercase letters, lowercase letters, digits, and special characters (!@$%^-_=+[{}]:,./?). The default administrator account is root.	Empty

Parameter	Type	Mandatory	Description	Default Value
system_disk_size	Number	Mandatory	Cloud server system disk size, disk type defaults to High IO, unit: GB, value range is 40-1,024, shrinking the disk is not supported.	100
bandwidth_size	Number	Mandatory	Elastic public network bandwidth size, billing mode for this template is pay-by-traffic. Unit: Mbit/s, value range: 1-300 Mbit/s.	300
charging_mode	String	Mandatory	Billing mode, default automatic payment. Optional values: postPaid (pay-as-you-go), prePaid (yearly/monthly subscription).	postPaid
charging_unit	String	Mandatory	Subscription period type, effective only when charging_mode is prePaid (yearly/monthly subscription). This parameter becomes mandatory in this case. Optional values: month (monthly), year (yearly).	month

Parameter	Type	Mandatory	Description	Default Value
charging_period	Number	Mandatory	Subscription period, effective only when charging_mode is prePaid (yearly/monthly subscription). This parameter becomes mandatory in this case. When charging_unit=month (monthly cycle type), value range: 1-9; when charging_unit=year (yearly cycle type), value range: 1-3.	1

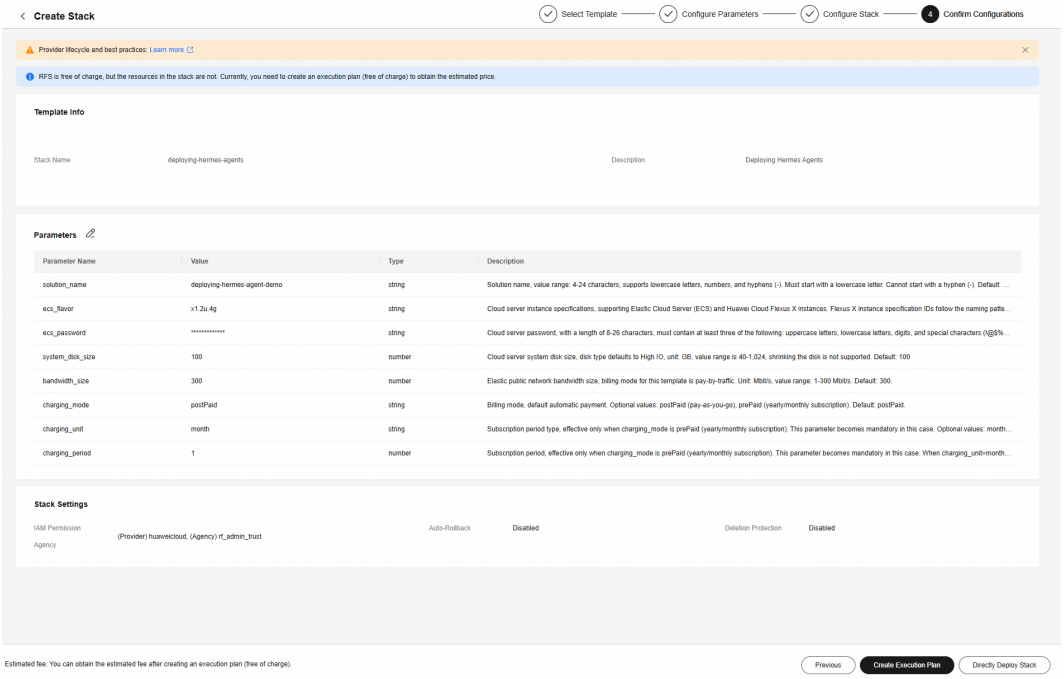
Step 4 On the **Configure Stack** page, select **rf_admin_trust** from the **Agency** drop-down list and click **Next**. This step is optional if you use an account (HUAWEI ID) or use an IAM user in the admin user group.

Figure 3-11 Configuring a stack



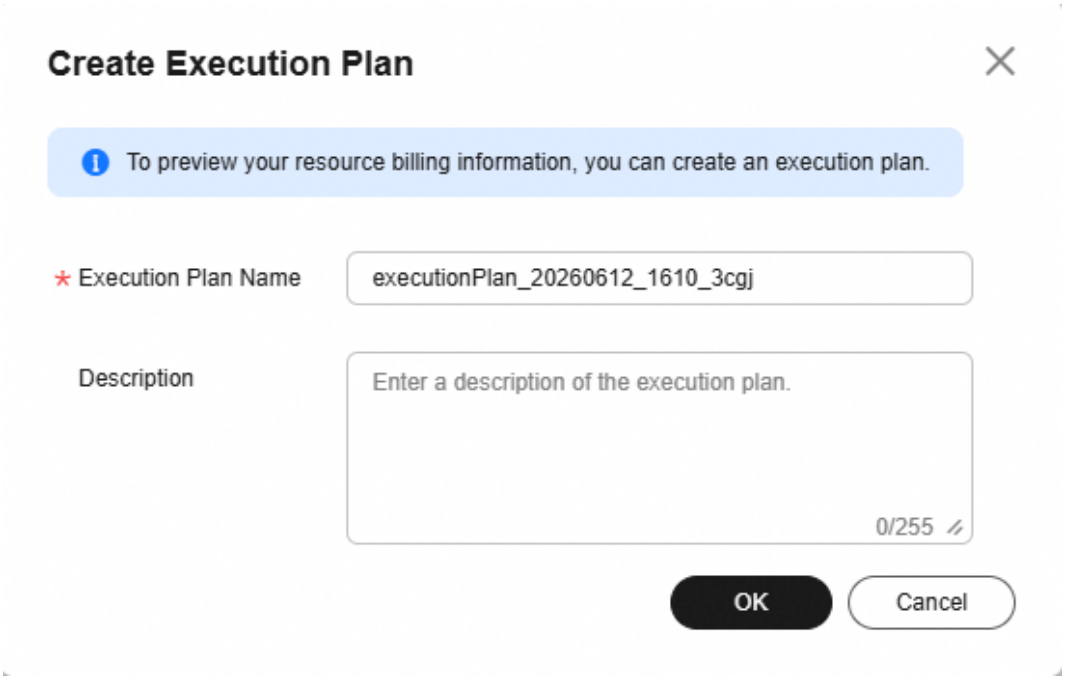
Step 5 On the **Confirm Configurations** page, confirm the configurations and click **Create Execution Plan**.

Figure 3-12 Confirming the configurations



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

Figure 3-13 Creating an execution plan



Step 7 Wait until the status of the execution plan changes to **Available** and then click **Deploy** in the **Operation** column. In the displayed dialog box, click **Execute.**

Figure 3-14 Execution plan page

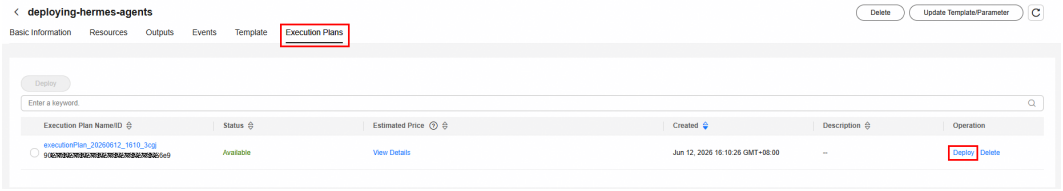
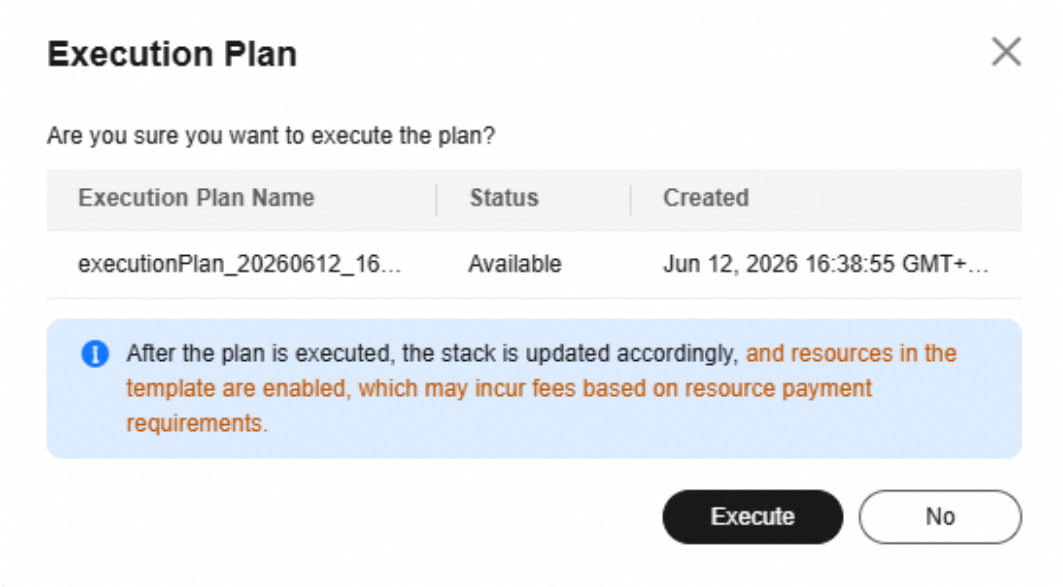


Figure 3-15 Confirming the execution plan

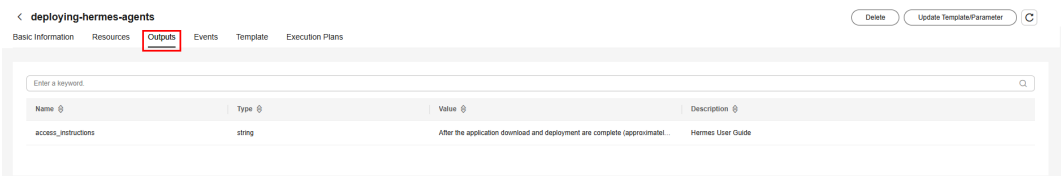


Step 8 (Optional) If you select the yearly/monthly billing mode and your account balance is insufficient, log in to the Billing Center to manually pay for the order. You can refer to **Table2 Resource planning costs (yearly/monthly)** to see the total price.

Step 9 Click the **Events** tab and check whether the message "Apply required resource success." is displayed in the **Description** column. If it is, the solution is successfully deployed. Deployment of BISHENG platform script started execution. Wait for approximately 5 to 10 minutes (affected by network fluctuations).

Figure 3-16 Solution deployed

Figure 3-17 Access description



----End

3.3 Getting Started

This solution uses port 22 for remote login to cloud servers via CloudShell, with IP whitelist configured by default. If remote login is required, you can directly use CloudShell for remote access.

After successful deployment of this solution, environment initialization and application installation are expected to take approximately 10-15 minutes. Deployment time may vary due to network and bandwidth conditions. The solution will be fully functional only after deployment is complete.

(Optional) Modifying Security Group Rules

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

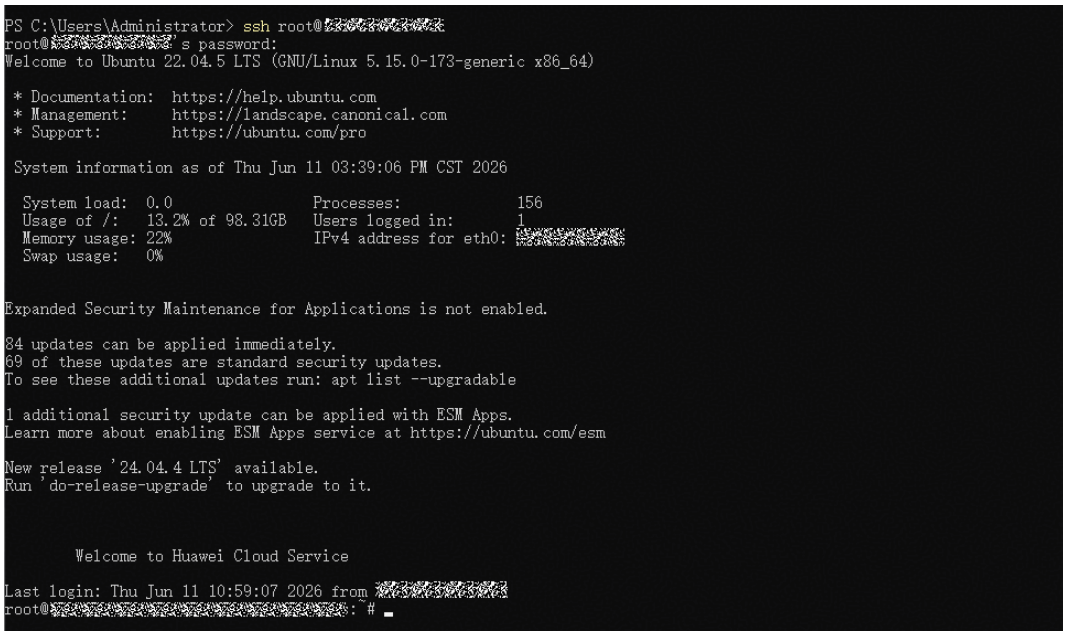
You can modify security group rules, for example, by adding, modifying, or deleting a TCP port, as follows:

- 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 may introduce serious security risks. You can **modify security group rules** to ensure the network security of your 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**.

Hermes Agent Model Configuration

Step 1 Logging in to the server remotely: Open the local terminal, enter `ssh root@{public IP}`, and then enter the server password to log in to the server.

Figure 3-18 SSH login



Step 2 Model Cofiguration

```
//Execute the following command in the command line.
docker exec -it hermes bash
//Activate the virtual environment
source /opt/hermes/.venv/bin/activate
//Configure the model
hermes model
```

Figure 3-19 Activate the virtual environment

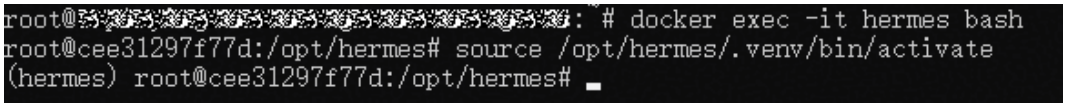


Figure 3-20 Select "Custom endpoint (enter URL manually)"

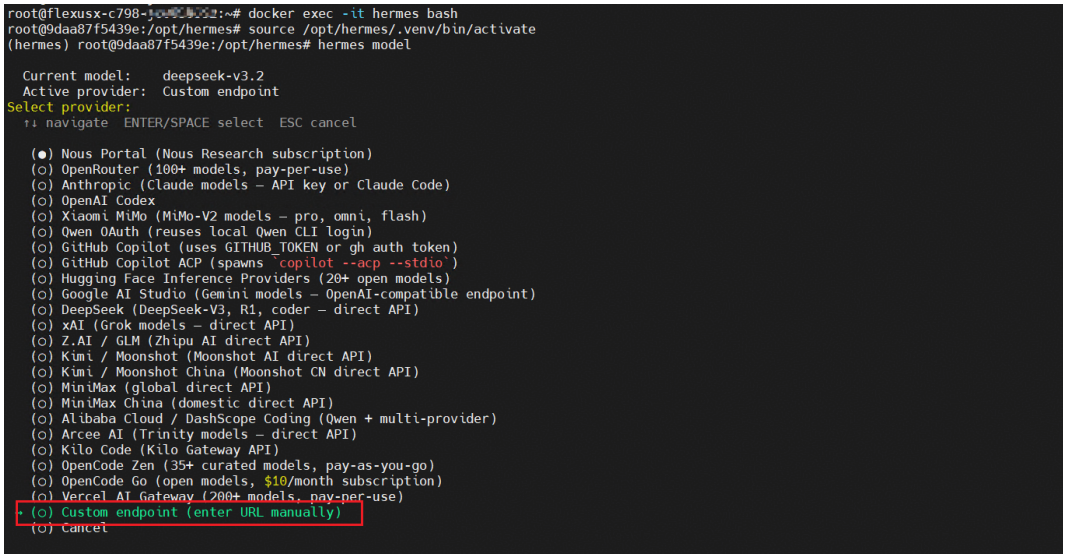
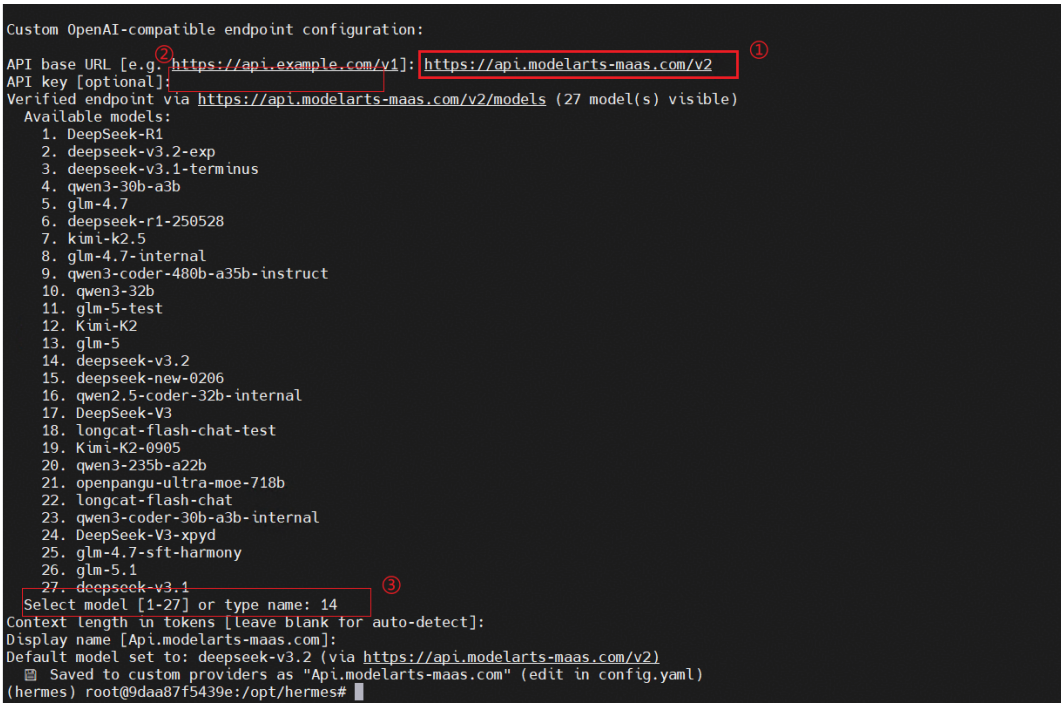


Figure 3-21 Configure base URL and API key



----End

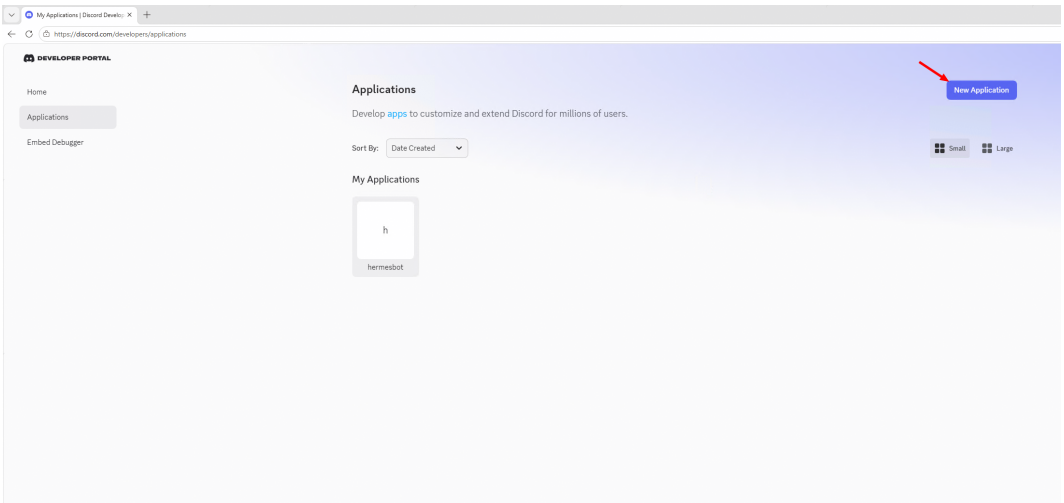
Add a Channel (Discord)

NOTE

This section takes Discord as an example. Each channel has its own configuration steps. For other channels, please refer to official documentations.

Step 1 Create a Discord Application

Figure 3-22 Go to Discord Developer Portal (https://discord.com/developers/applications), then create a new application



Step 2 Create the Bot, Get the Bot Token

Figure 3-23 Click Bot -> Reset Token

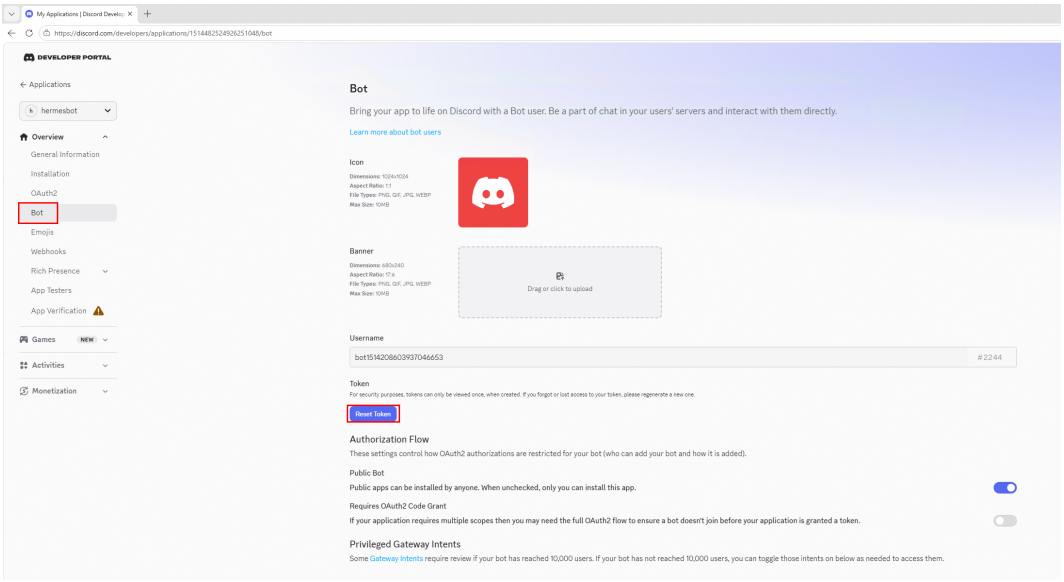
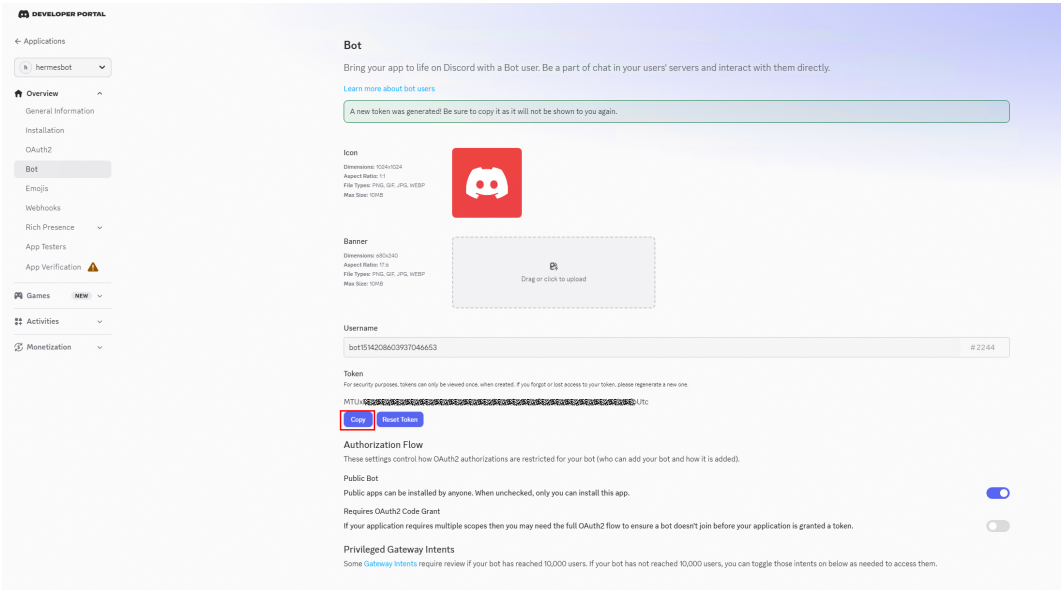
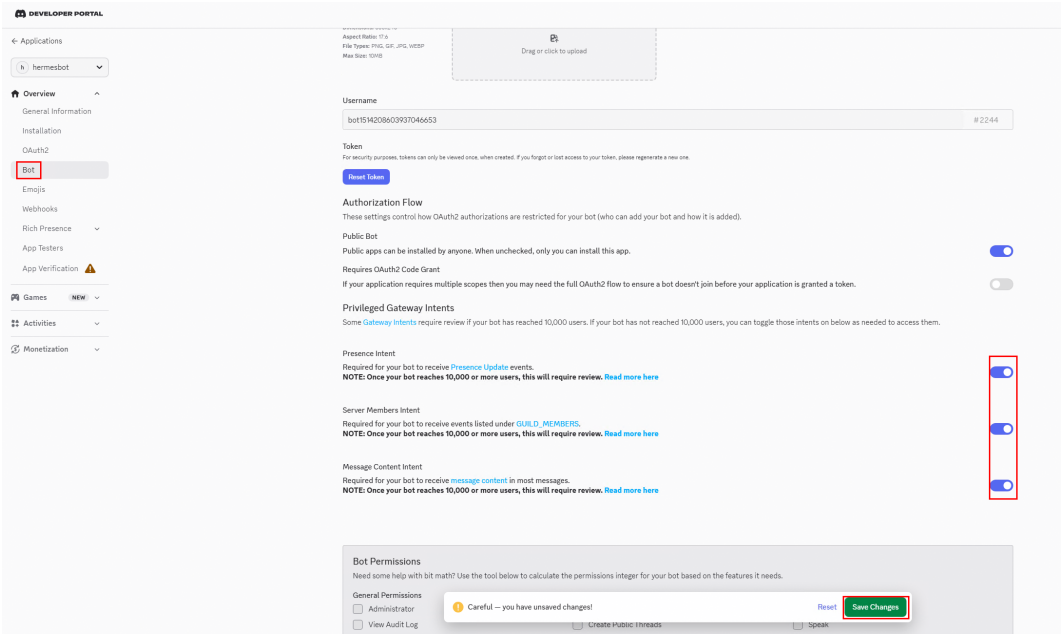


Figure 3-24 Copy -> Save



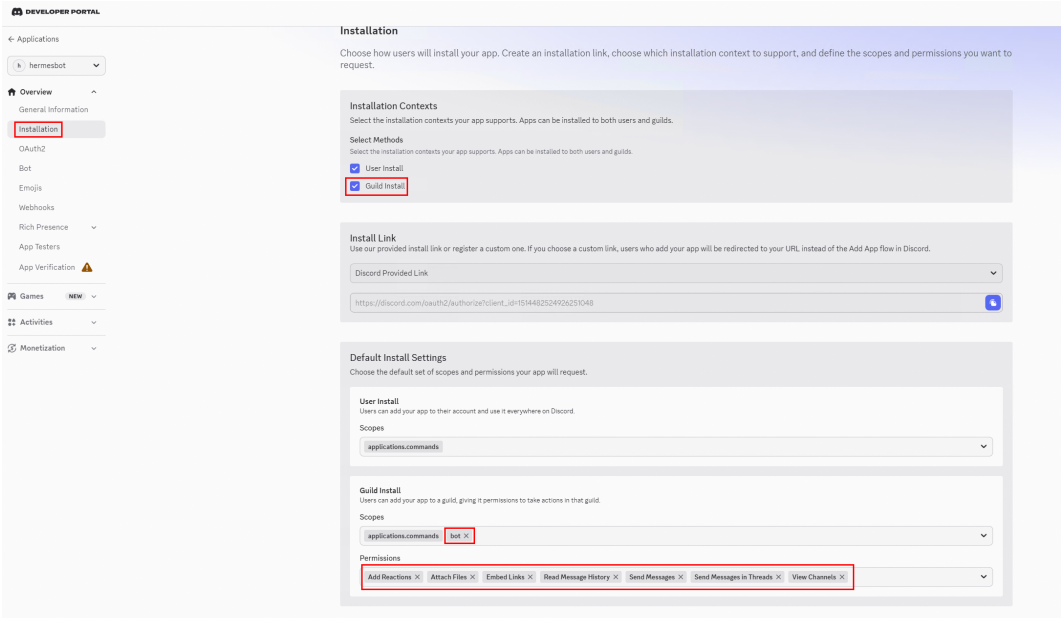
Step 3 Enable Privileged Gateway Intents

Figure 3-25 Toggle on Privileged Gateway Intents and save



Step 4 Generate the Invite URL

Figure 3-26 Select Installation-> Enable Guild Install -> Under Default Install Settings for Guild Install, select the options listed below



Step 5 Invite to Your Server

Figure 3-27 Open the invite URL in your browser -> Add to Server

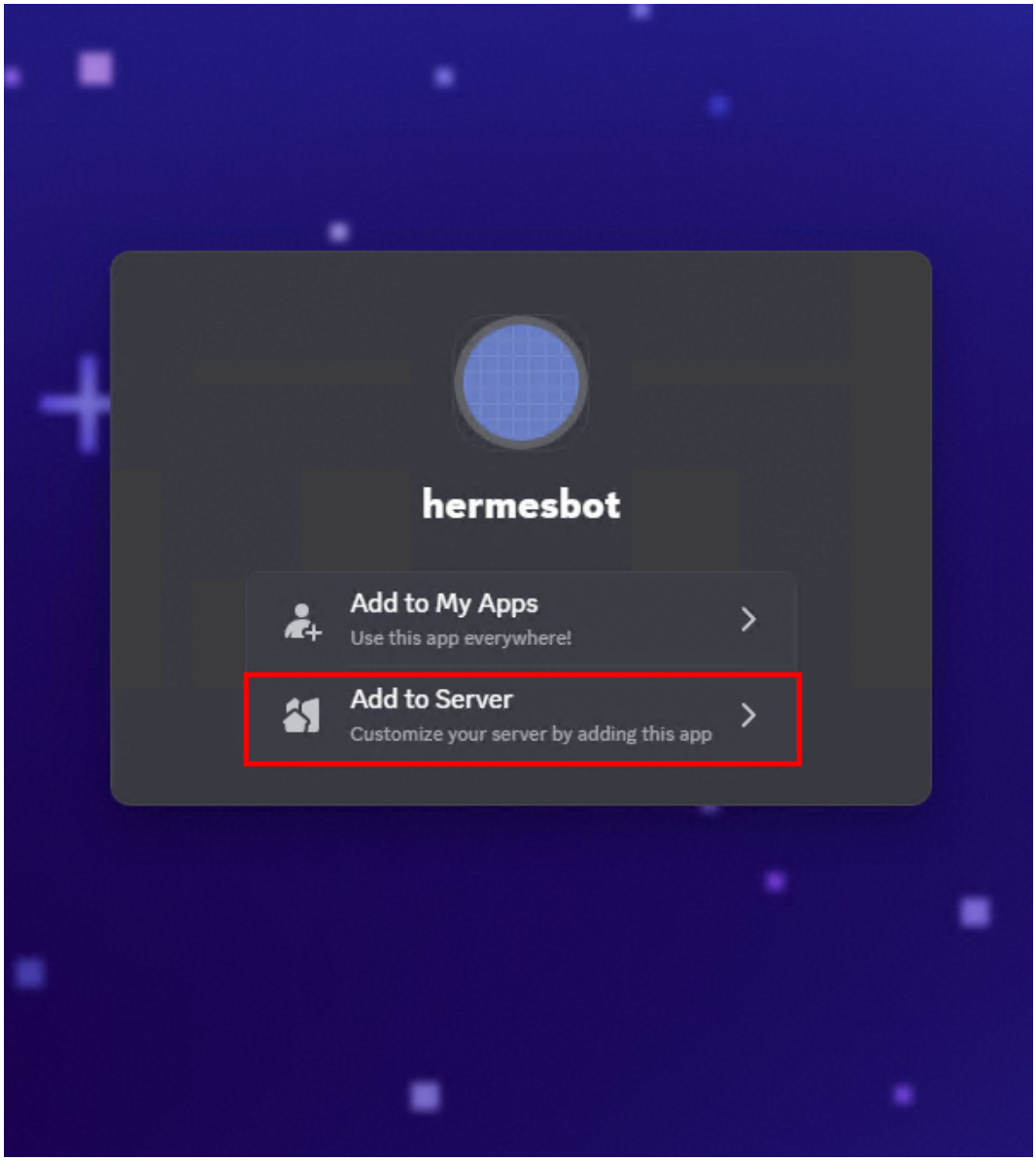
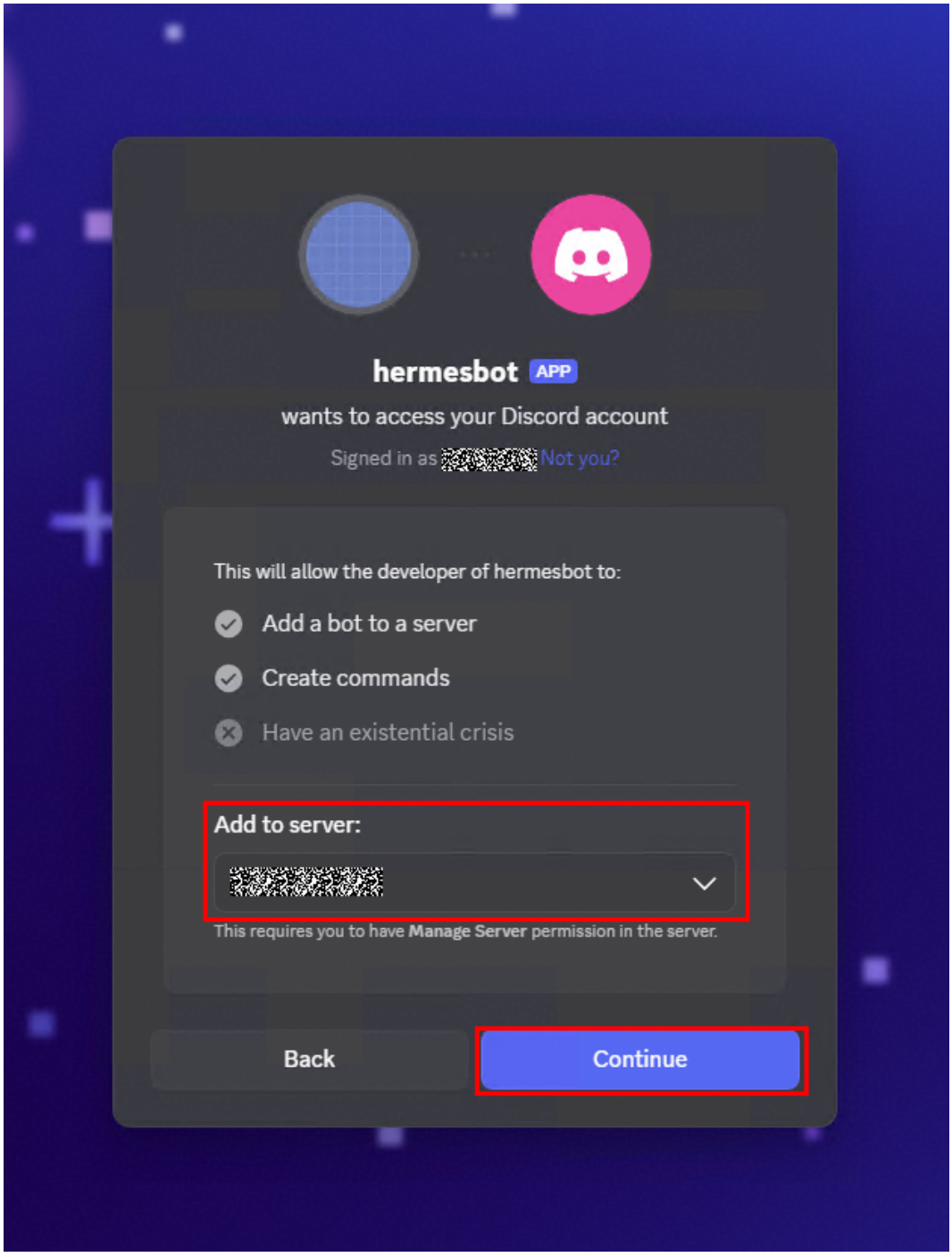


Figure 3-28 Select your sever and authorize



Step 6 Find Your Discord User ID

Figure 3-29 Open Discord -> Settings -> Developer ->toggle Developer Mode to ON

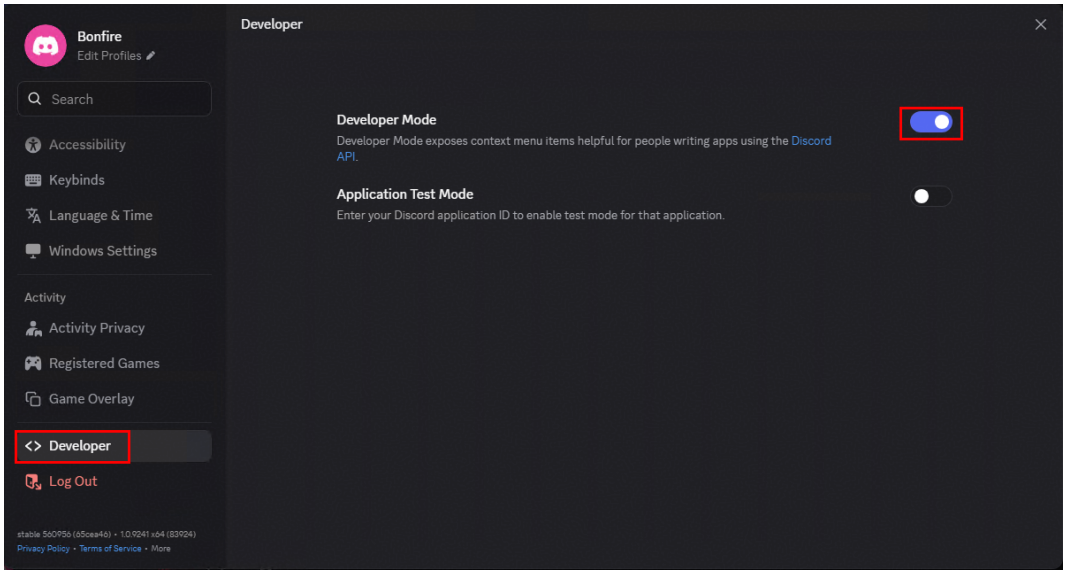
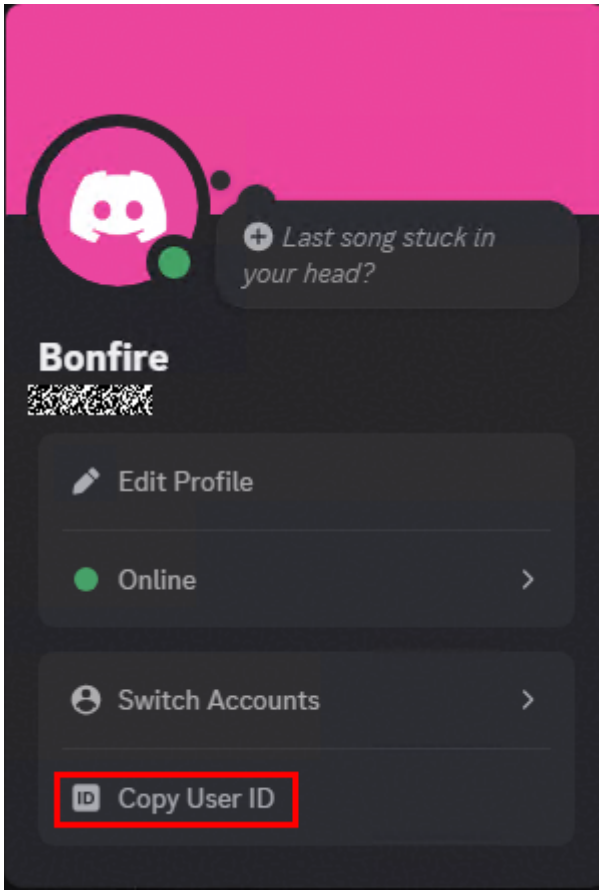


Figure 3-30 Copy User ID -> Save



Step 7 Logging in to the server: If you did not log out of the server in the previous phase, you can skip this step. Otherwise, log in to the server by referring to step 1 "**Activate the virtual environment**" in this document.

Step 8 Add a channel: Run **hermes gateway setup** to start the configuration.

Figure 3-31 Select Discord

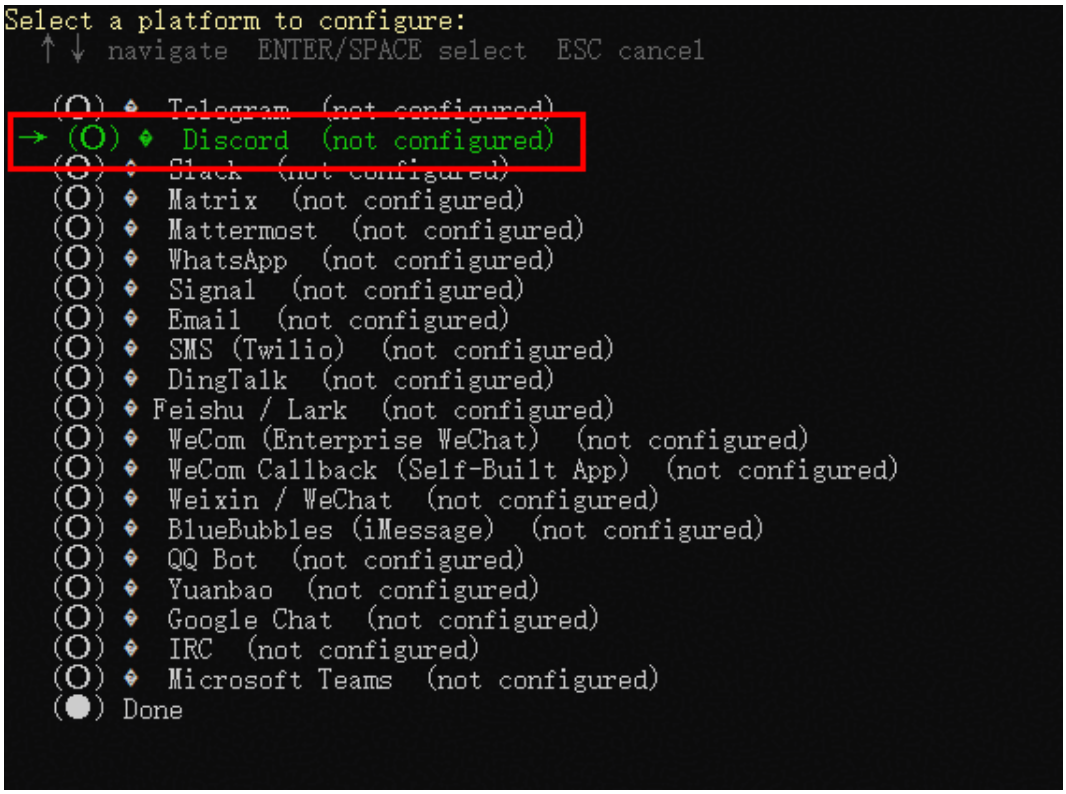


Figure 3-32 Paste token

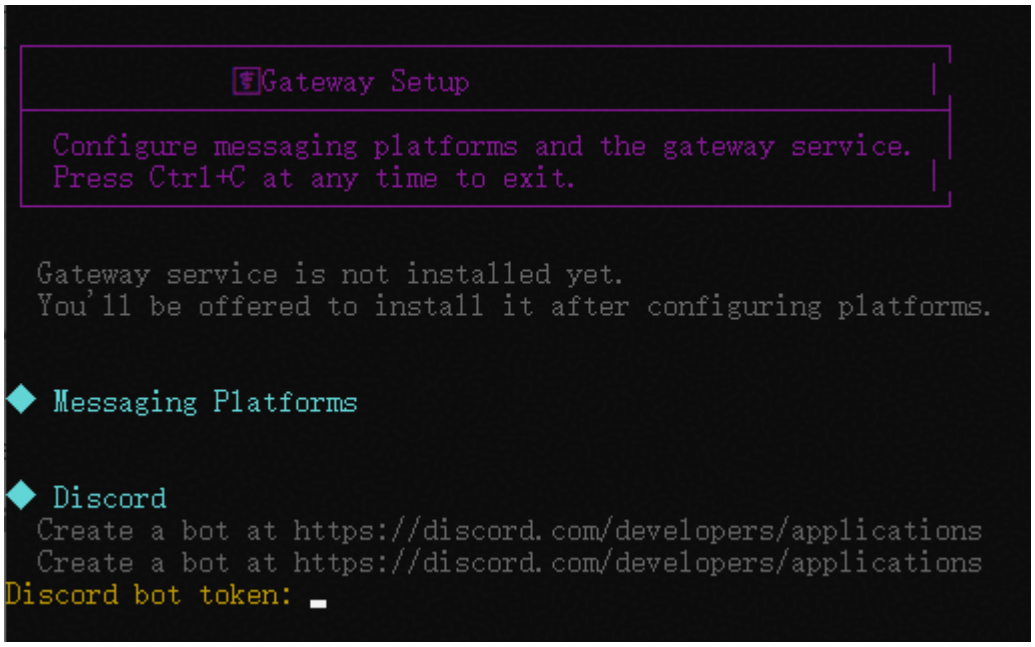


Figure 3-33 Paste user IDs

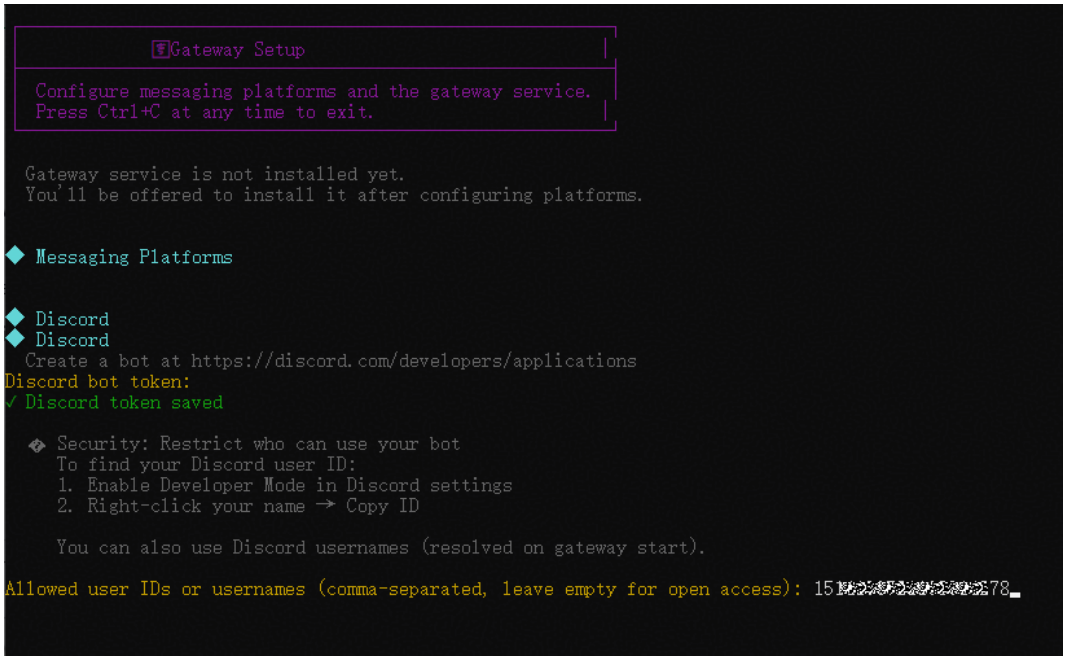


Figure 3-34 Restart the gateway

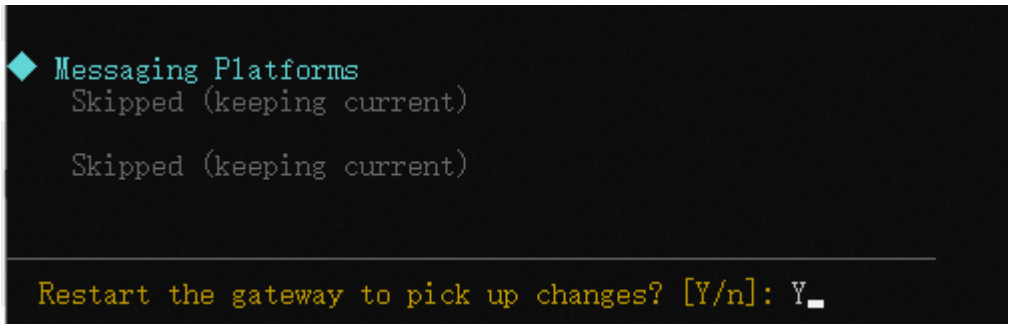
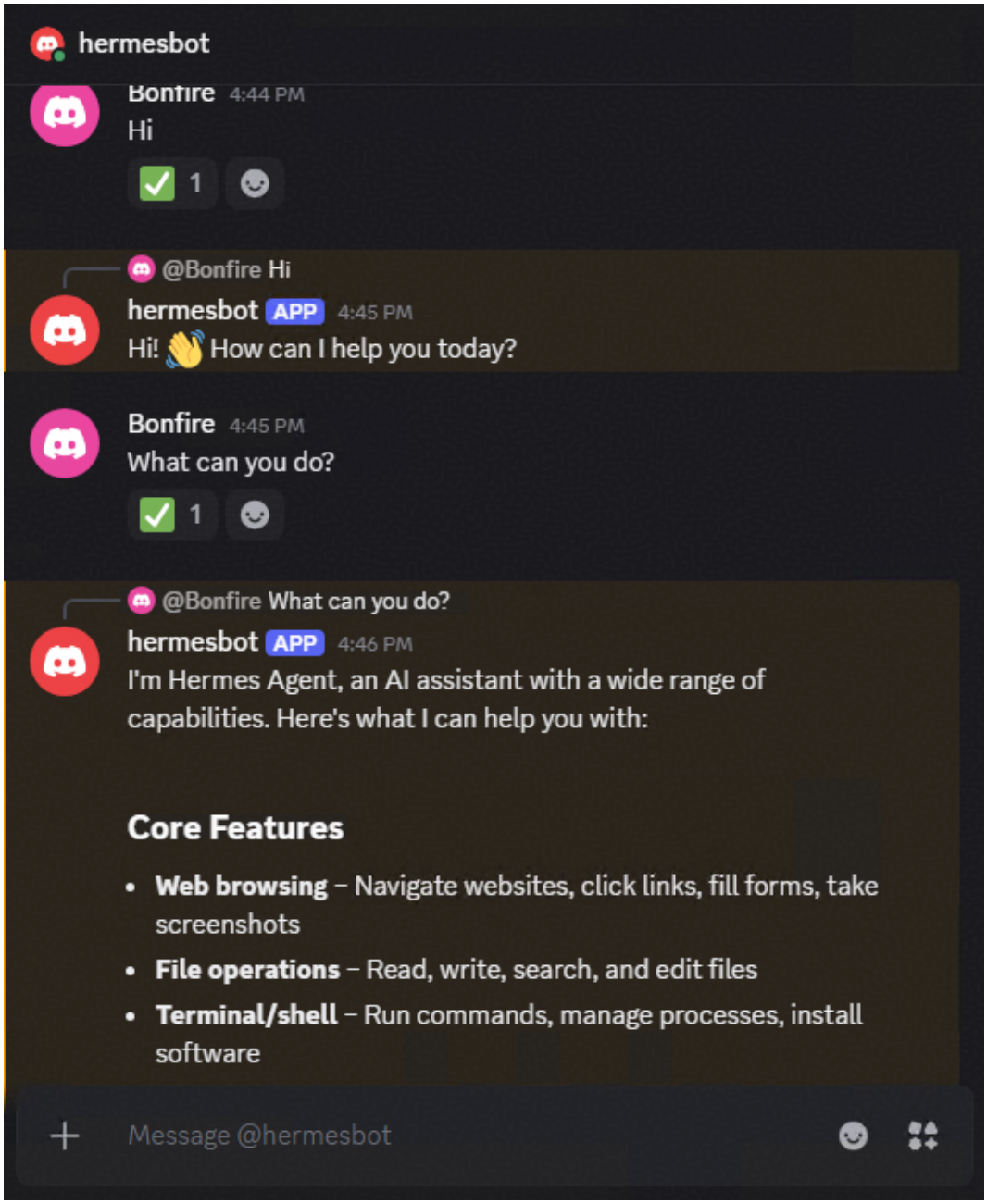


Figure 3-35 Start chatting

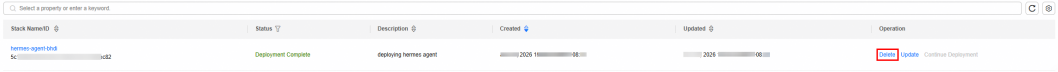


----End

3.4 Quick Uninstallation

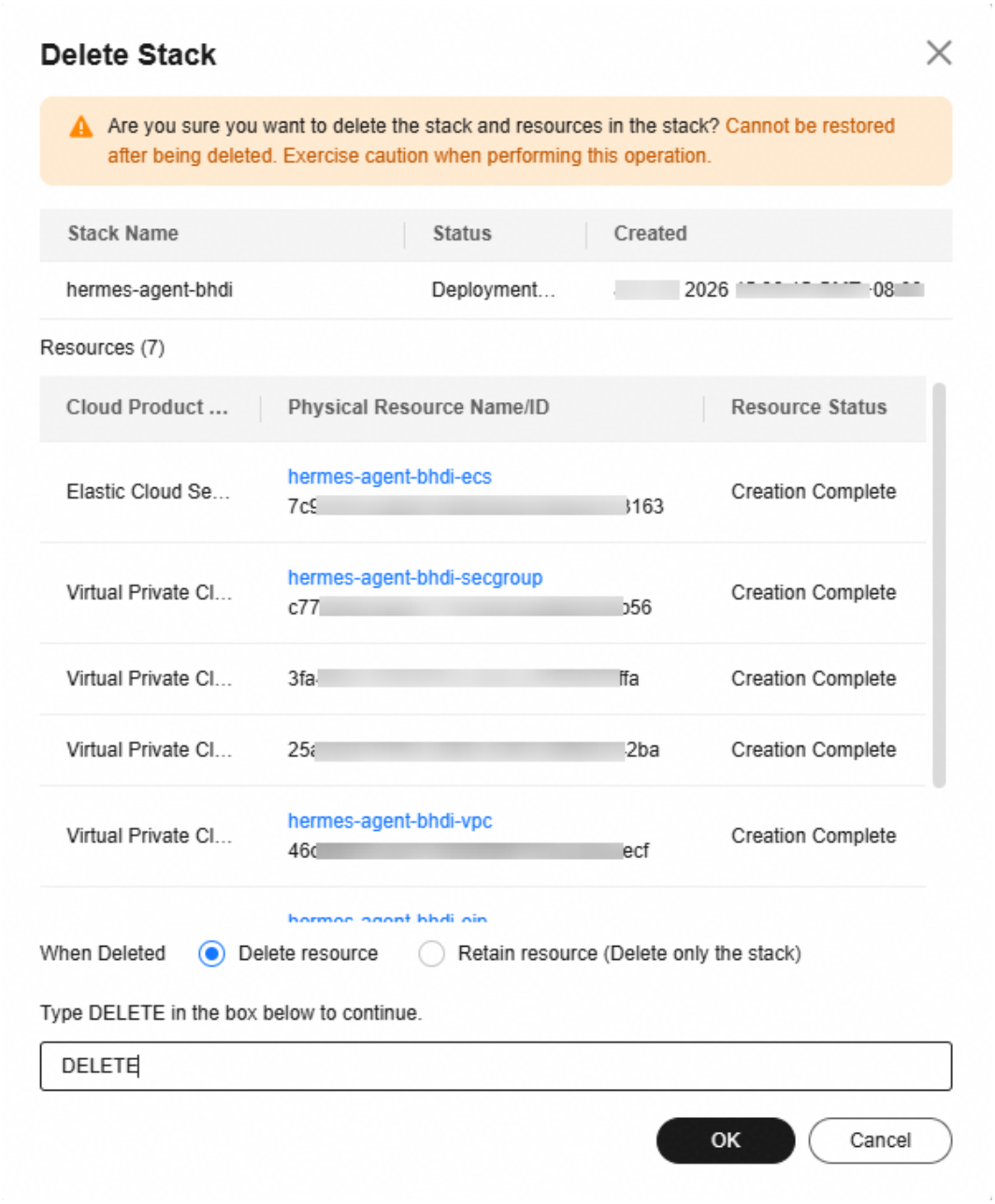
Step 1 Log in to the **RFS console**. On the Stacks page, locate the resource stack you created and click Delete in the Operation column.

Figure 3-36 Deleting a stack



Step 2 In the displayed Delete Stack dialog box, set When Deleted to Delete resource, enter "Delete" and click OK.

Figure 3-37 Confirming the deletion



----End

4 Appendix

Terms

- **Flexus X Instance (FlexusX):** FlexusX is a next-generation flexible cloud server service designed for small- and medium-sized enterprises (SMEs) and developers. FlexusX provides functions similar to what ECS provides. In addition, with FlexusX, you can flexibly configure vCPU to memory ratios to match your specific needs and change server specifications without service interruptions. For details, see **Flexus X Instance (FlexusX)**.
- **Elastic Cloud Server (ECS):** ECS provides secure, scalable, on-demand compute resources, enabling you to flexibly deploy applications and workloads.
- **Virtual Private Cloud (VPC):** VPC allows you to isolate online resources with virtual private networks. VPC enables your cloud resources to securely communicate with each other, the internet, and on-premises networks.
- **Elastic IP (EIP):** EIP provides static public IP addresses and scalable bandwidths that enable your cloud resources to communicate with the Internet. You can easily bind an EIP to a FlexusX instance, ECS, BMS, virtual IP address, NAT gateway, or load balancer, enabling immediate Internet access.

5 Change History

Released On	Description
2026-06-22	This issue is the first official release.

