Solution Practices

Building a NewAPI LLM Gateway

Issue	01
Date	2026-07-13

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2026. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Cloud Computing Technologies Co., Ltd.

Trademarks and Permissions

HUAWEI 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.....1

2 Resource Planning and Costs.....4

3 Procedure..... 11

3.1 Preparations.....11

3.2 Quick Deployment.....15

3.3 USing Started.....27

3.4 Quick Uninstallation.....34

4 Appendix.....37

5 Change History..... 38

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

1 Solution Overview

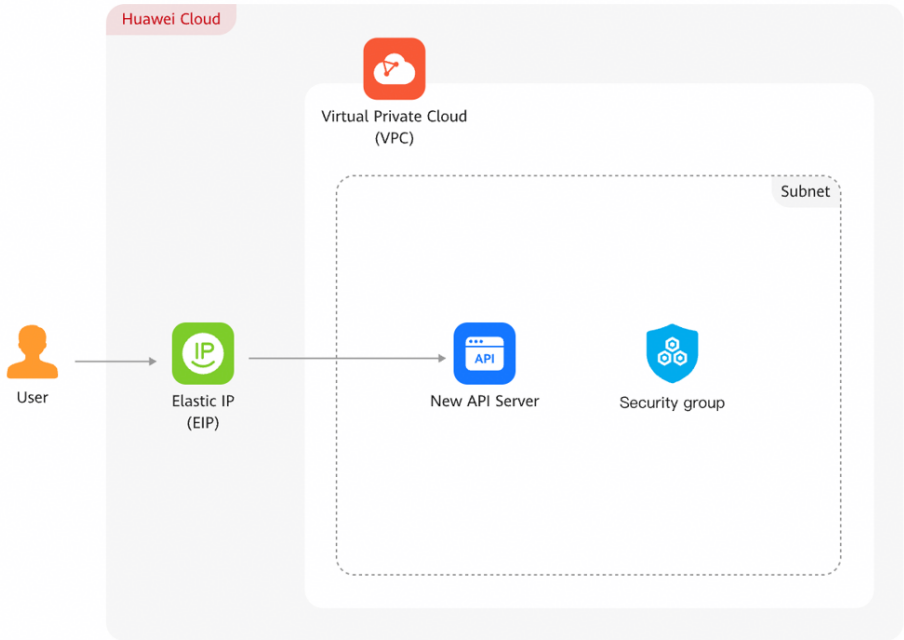
Application Scenarios

This solution helps you quickly deploy the Community Standalone Edition and CCE Container High Availability Edition New API Large Model Gateway. **New API** is an open-source project based on the secondary development of One API, focusing on AI model interface management and distribution systems. It supports converting various LLM into OpenAI format for invocation and is compatible with interfaces such as Midjourney Proxy, Suno, and Rerank.

Architecture

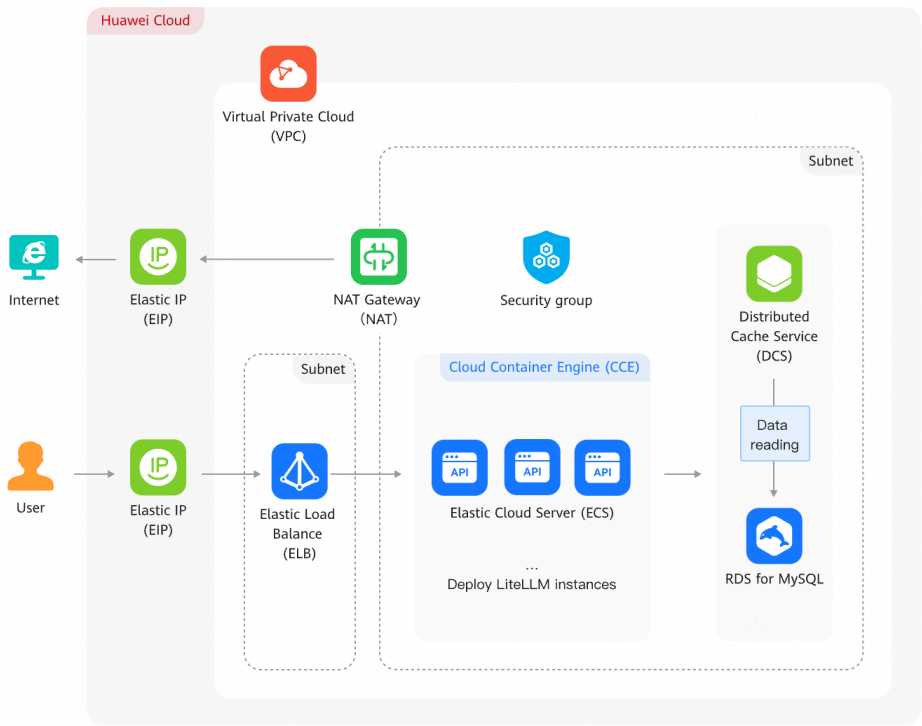
This solution helps you quickly deploy New API LLM gateway.

Figure 1-1 Architecture (Community Standalone Edition)



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 1-2 Architecture (CCE Container High Availability Edition)



This solution will:

Community Standalone Edition:

- Create three **Flexus X Instance** (FlexusX) : Used to deploy the New API LLM gateway container cluster.
- Create one **Elastic IP (EIP)** : Provides capabilities for accessing the public network and being accessed from the public network.

CCE Container High Availability Edition:

- Create one **Cloud Container Engine CCE** Turbo cluster: Used to deploy the New API LLM gateway container cluster.
- Create three **Flexus X Instance** (FlexusX) : Serves as the worker nodes for the CCE container cluster to host container instances.
- Create one **Elastic IP (EIP)** : Provides capabilities for accessing the public network and being accessed from the public network.
- Create one **Elastic Load Balance (ELB)** : Automatically distributes external inference traffic, enhancing external service capability and application fault tolerance.
- Create one **NAT Gateway** : Allows instances within private subnets and container subnets to securely initiate outbound Internet connections to request upstream APIs.
- Create one **Distributed Cache Service (DCS)**: Used for global rate limiting and route cache status synchronization across multiple replicas.

- Create one **RDS for MySQL**: Used for the persistent storage of New API user API keys, organizational permissions, and routing metadata.

Advantages

- One-Click Deployment

  Easily deploy with one click to complete the creation of cloud service resources and the setup of the New API large model gateway.
- Ready to Use

  New API adopts a modern microservices architecture, emphasizing high availability and ease of deployment. It is suitable for content creation, education, scientific research, and other fields, offering ready to use functionality after deployment.
- Cost Optimization

  Provides cost-effective cloud servers with on-demand resource selection, supports auto-scaling to reduce idle resources, optimizes cost investment, and further lowers customer operational expenses.

Constraints

- Before deploying this solution, ensure that you have created a Huawei ID with access to the target region and enabled Huawei Cloud services. If you select the yearly/monthly billing mode, ensure that your account has sufficient balance. If you do not have sufficient balance, you can go to the **Billing Center** to manually pay for the order.
- If using IAM agency permissions to deploy resources, ensure your Huawei Cloud account has sufficient IAM permissions. For details, refer to creating the **rf_admin_trust** agency. If deploying with a Huawei primary account or an IAM sub-account under the admin user group, you may skip the agency option, and the deployment will proceed using the permissions of the currently logged-in user.
- This content may include technical descriptions and configuration examples of third-party open-source software, components, or frameworks. You should independently evaluate and decide whether to use them. Huawei Cloud makes no warranties regarding the functionality, performance, security, compatibility, or compliance of third-party software. Intellectual property rights of third-party software belong to their respective owners. You must strictly comply with applicable open-source license terms (such as MIT, Apache 2.0, etc.), ensure no infringement of any third-party rights, and take full responsibility for security assessments and vulnerability management. Huawei Cloud and its affiliates shall not be liable for any direct or indirect losses (including but not limited to data loss, business interruption, security incidents, intellectual property disputes, compliance penalties, etc.) arising from your use of any third-party software mentioned in this solution. References to third-party software in this solution do not constitute endorsement or imply any authorization or partnership between Huawei Cloud and the third-party rights holders.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2 Resource Planning and Costs

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

Community Standalone Edition

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

Huawei Cloud Service	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Virtual Private Cloud (VPC)	• Region: AP-Singapore • CIDR Block: 192.168.0.0/16	1	0.00
Subnet	• Region: AP-Singapore • IPv4 CIDR Block: 192.168.200.0/24, 192.168.201.0/24, 192.168.202.0/24 • Gateway: 192.168.200.1, 192.168.201.1, 192.168.202.1	1	0.00
SecurityGroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	1	0.00

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Huawei Cloud Service	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Flexus X Instance	• Region: AP-Singapore • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.8u.16g | 8 vCPUs | 16 GiB • System Disk: General Purpose SSD | 100GB • Image: Ubuntu 22.04 server 64bit	1	$0.30053 USD
Elastic IP (EIP)	• Region: AP-Singapore • Routing Type: Dynamic BGP • Billing Mode: traffic • Bandwidth: 300Mbit/s	1	$0.114 USD/GB
MaaS tokens (Optional)	• Region: AP-Singapore • Billing Mode: Token-based billing • Model: DeepSeek-V3	-	Input: $0.27 USD/million tokens Output: $1.078 USD/million tokens
Total	**-**	**-**	**$0.30053 USD + Application LCU billing + Public network traffic price + MaaS tokens billing**

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

Huawei Cloud Service	Configuration Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	• Region: AP-Singapore • CIDR Block: 192.168.0.0/16	1	0.00

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Huawei Cloud Service	Configuration Example	Quantity	Estimated Monthly Cost
Subnet	• Region: AP-Singapore • IPv4 CIDR Block: 192.168.200.0/24, 192.168.201.0/24, 192.168.202.0/24 • Gateway: 192.168.200.1, 192.168.201.1, 192.168.202.1	1	0.00
SecurityGroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	1	0.00
Flexus X Instance	• Region: AP-Singapore • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.8u.16g | 8 vCPUs | 16 GiB • System Disk: General Purpose SSD | 100GB • Image: Ubuntu 22.04 server 64bit	1	$159.54 USD
Elastic IP (EIP)	• Region: AP-Singapore • Routing Type: Dynamic BGP • Billing Mode: traffic • Bandwidth: 300Mbit/s	1	$0.114 USD/GB
MaaS tokens (Optional)	• Region: AP-Singapore • Billing Mode: Token-based billing • Model: DeepSeek-V3	-	Input: $0.27 USD/million tokens Output: $1.078 USD/million tokens
Total	**-**	**-**	**$159.54 USD + Public network traffic price + MaaS tokens billing**

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

CCE Container High Availability Edition

Table 2-3 Resource planning and costs (pay-per-use)

Huawei Cloud Service	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Virtual Private Cloud (VPC)	• Region: AP-Singapore • CIDR Block: 192.168.0.0/16	1	0.00
Subnet	• Region: AP-Singapore • IPv4 CIDR Block: 192.168.200.0/24, 192.168.201.0/24, 192.168.202.0/24 • Gateway: 192.168.200.1, 192.168.201.1, 192.168.202.1	3	0.00
SecurityGroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	4	0.00
Cloud Container Engine (CCE)	• Region: AP-Singapore • Type: CCE Turbo • Specifications: cce.s2.small (50 nodes) • Master Nodes: 3	1	$0.54 USD
Flexus X Instance	• Region: AP-Singapore • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.8u.16g | 8 vCPUs | 16 GiB • System Disk: General Purpose SSD | 100GB • Image: Ubuntu 22.04 server 64bit	3	$0.8847 USD
Elastic IP (EIP)	• Region: AP-Singapore • Routing Type: Dynamic BGP • Billing Mode: traffic • Bandwidth: 300Mbit/s	3	$0.114 USD/GB

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Huawei Cloud Service	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Elastic Load Balance (ELB)	• Region: AP-Singapore • Type: Dedicated load balancer • AZs: 2 • Specifications: Elastic • Load Balancing Type: Application load balancing	1	$0.025 USD/hour+LCU $0.00833 USD/LCU (Charge LCU fees based on actual usage)
NAT Gateway (NAT)	• Region: AP-Singapore • Gateway: Public network • Type: Small	1	$2.438 USD
Distributed Cache Service DCS (for Redis)	• Region: AP-Singapore • Type: Redis 6.0 (Master/Standby) • Specifications: 1G (Basic) | Replicas: 2	1	$0.034 USD
RDS for MySQL	• Region: AP-Singapore • Type: MySQL (Master/Standby) • Specifications: rds.mysql.n1.xlarge.2.ha | 4 vCPUs | 8 GB(General-purpose) • Storage: Cloud SSD | 100 GB	1	$0.77 USD
MaaS tokens (Optional)	• Region: AP-Singapore • Billing Mode: Token-based billing • Model: DeepSeek-V3	1	Input: $0.27 USD/million tokens Output: $1.078 USD/million tokens
Total	-	-	**$4.6917 USD + Application LCU billing + Public network traffic price + MaaS tokens billing**

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 2-4 Resource planning and costs (yearly/monthly)

Huawei Cloud Service	Configuration Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	• Region: AP-Singapore • CIDR Block: 192.168.0.0/16	1	0.00
Subnet	• Region: AP-Singapore • IPv4 CIDR Block: 192.168.200.0/24, 192.168.201.0/24, 192.168.202.0/24 • Gateway: 192.168.200.1, 192.168.201.1, 192.168.202.1	3	0.00
SecurityGroup	• Region: AP-Singapore • Allow ping: 0.0.0.0/0 • Open port 22 to allow Cloud Shell login: 119.8.185.245/32	4	0.00
Cloud Container Engine (CCE)	• Region: AP-Singapore • Type: CCE Turbo • Specifications: cce.s2.small (50 nodes) • Master Nodes: Single	1	$81.63 USD
Flexus X Instance	• Region: AP-Singapore • Specifications: Flexus X Instance | Performance Mode (Disabled) | x1.8u.16g | 8 vCPUs | 16 GiB • System Disk: General Purpose SSD | 100GB • Image: Ubuntu 22.04 server 64bit	3	$474.45 USD
Elastic IP (EIP)	• Region: AP-Singapore • Routing Type: Dynamic BGP • Billing Mode: traffic • Bandwidth: 300Mbit/s	3	$0.114 USD/GB
Elastic Load Balance (ELB)	• Region: AP-Singapore • Type: Dedicated load balancer • Specifications: Fixed • AZs: 2 • Load Balancing Type: Application load balancing • Application load balancing: Small I	1	$66.7 USD

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Huawei Cloud Service	Configuration Example	Quantity	Estimated Monthly Cost
NAT Gateway (NAT)	• Region: AP-Singapore • Gateway: Public network • Type: Small	1	$57.3 USD
Distributed Cache Service DCS (for Redis)	• Region: AP-Singapore • Type: Redis 6.0 (Master/Standby) • Specifications: 1G (Basic) | Replicas: 2	1	$31.32 USD
RDS for MySQL	• Region: AP-Singapore • Type: MySQL (Master/Standby) • Specifications: rds.mysql.n1.xlarge.2.ha | 4 vCPUs | 8 GB(General-purpose) • Storage: Cloud SSD | 100 GB	1	$410.92 USD
MaaS tokens (Optional)	• Region: AP-Singapore • Billing Mode: Token-based billing • Model: DeepSeek-V3	1	Input: $0.27 USD/million tokens Output: $1.078 USD/million tokens
Total	**-**	**-**	**$1122.32 USD + Public network traffic price + MaaS tokens billing**

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3 Procedure

3.1 Preparations

3.2 Quick Deployment

3.3 USing Started

3.4 Quick Uninstallation

3.1 Preparations

(Optional) Creating the key

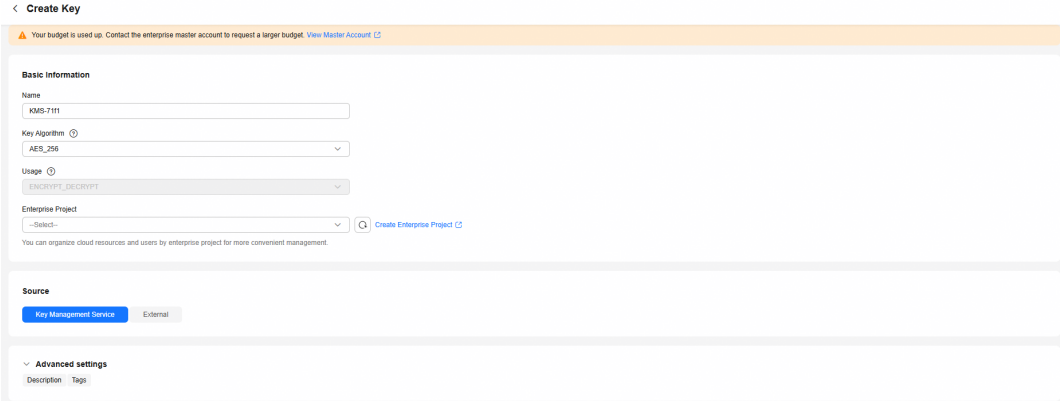
Step 1 Access the Huawei Cloud **Password Security Center console** and create a key.

- If the key exists, skip the subsequent steps for creating a key.
- If the key does not exist, proceed with the following steps to create a key.

Figure 3-1 create the key

Step 2 Select a custom key and create it.

Figure 3-2 Key configuration



----End

(Optional) Creating the rf_admin_trust Agency

NOTE

When you log in with your Huawei Cloud account, you do not need to perform this preparation step. If you are using an IAM user account, please confirm whether you are in the admin group. If you are not in the admin group, you will need to **Grant relevant permissions** to your IAM account and complete the following preparation steps.

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 3-3 Console page

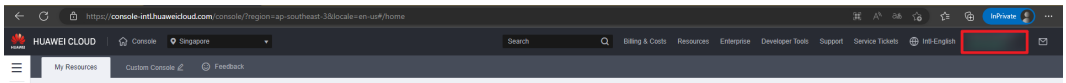
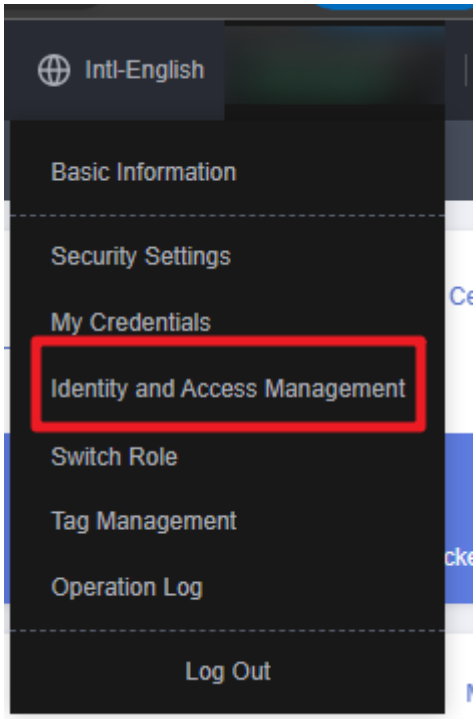
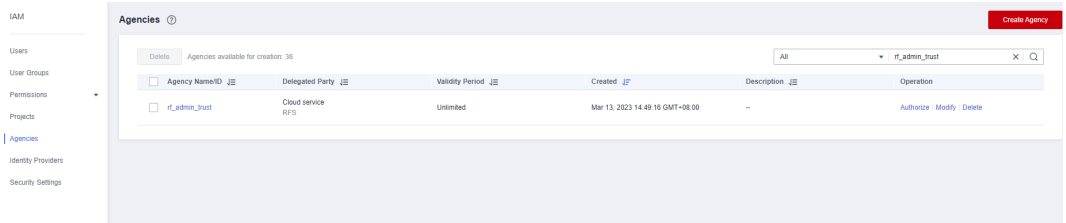


Figure 3-4 Identity and access management page



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

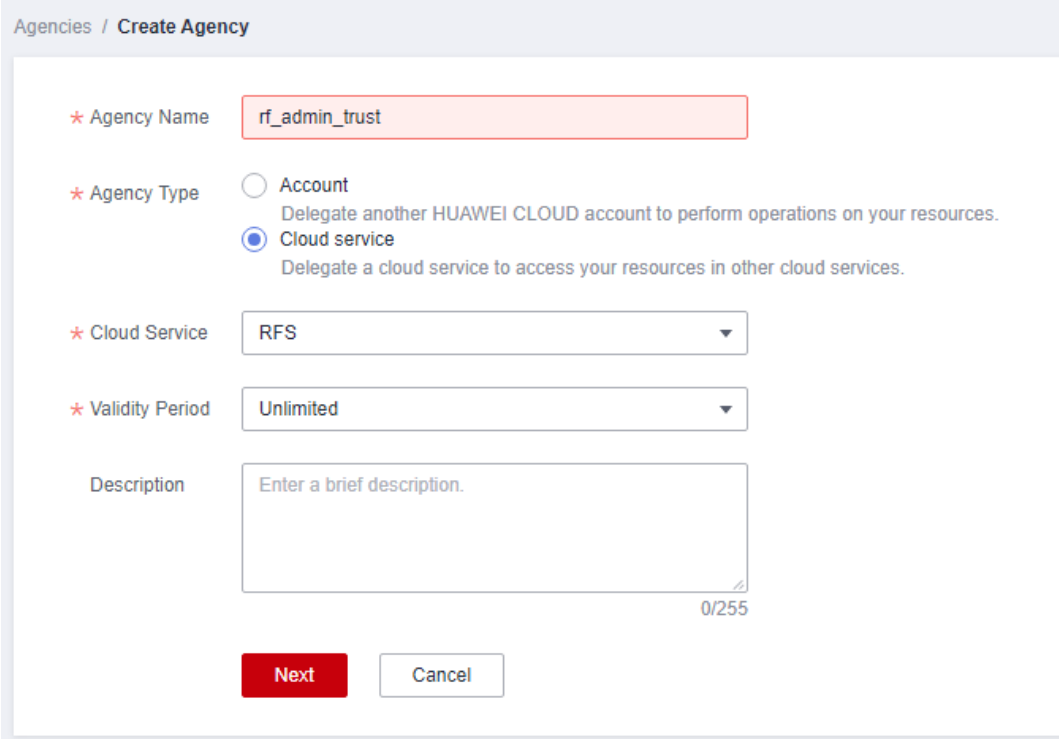
Figure 3-5 Agencies



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

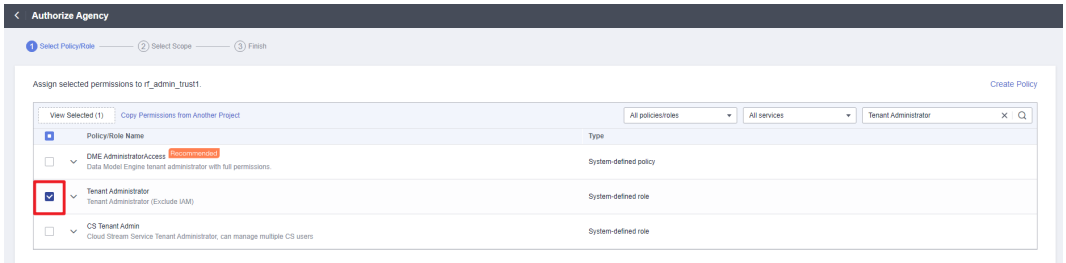
Step 3 Click **Create Agency** in the upper right corner of the page. On the displayed page, enter **rf_admin_trust** for **Agency Name**, select **Cloud service** for **Agency Type** and **RFS** for **Cloud Service**, and click **Next**.

Figure 3-6 Creating the rf_admin_trust agency



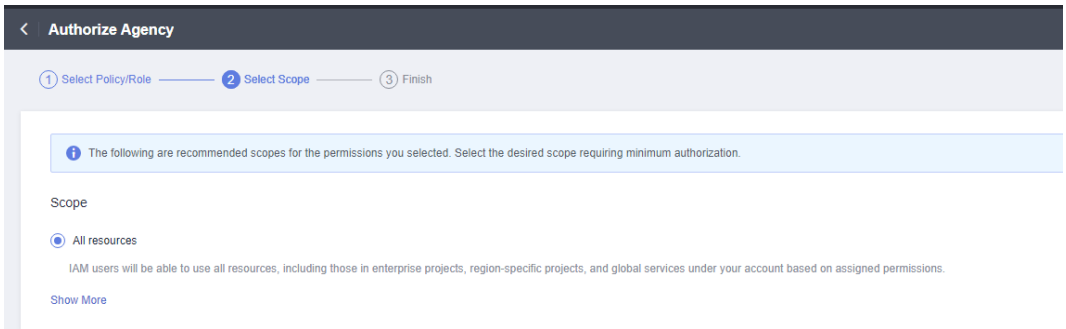
Step 4 Search for **Tenant Administrator**, select it in the search results, and click **Next**.

Figure 3-7 Selecting a policy/role



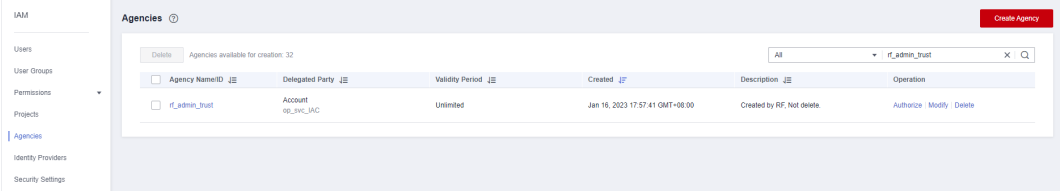
Step 5 Select **All resources** and click **OK**.

Figure 3-8 Setting the authorization scope



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

Figure 3-9 Agenciespe



----End

Log in to the CCE console for authorization

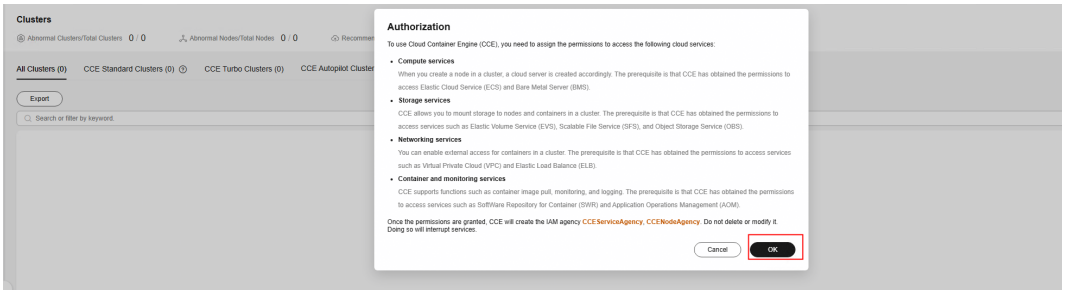
This section applies only to the CCE High Availability Edition. To create a CCE cluster with this solution, you must have logged in to the CCE console and completed the relevant authorization.**Otherwise, if enabling CCE for the first time, please follow the steps below for authorization.**

Step 1 Go to the **CCE console**.

Step 2 Click in the upper left corner of the console and select a region.

Step 3 When you log in to the CCE console of a region for the first time, the "Authorization Instructions" dialog will pop up. Please read them carefully and click "OK".

Figure 3-10 Authorization Instructions



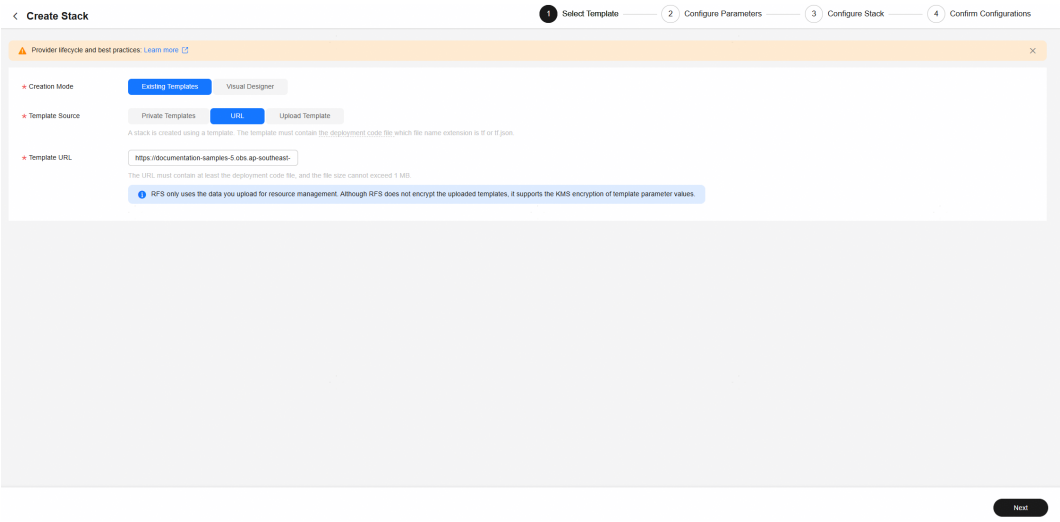
----End

3.2 Quick Deployment

This section helps users efficiently deploy the "Building a New API LLM Gateway" solution. When deploying this solution with one click, follow the steps and instructions in this section to complete the quick deployment.

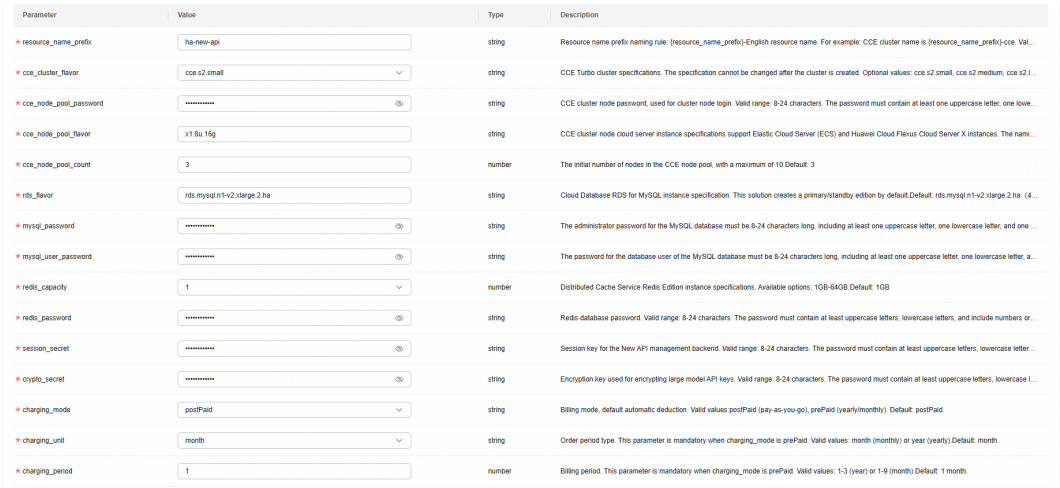
Step 1 Go to the solution details page, click "Start Deployment", select the corresponding deployment region, and jump to the RFS deployment page.

Figure 3-11 RFS Purchase Page



Step 2 Click "**Next**" to confirm the basic configuration and set the password.

Figure 3-12 Basic Configuration Diagram



Parameter	Value	Type	Description
* resource_name_prefix	ha-new-api	string	Resource name prefix naming rule: {resource_name_prefix}-English resource name. For example: CCE cluster name is {resource_name_prefix}-cce. Val...
* cce_cluster_flavor	cce.s2.small	string	CCE Turbo cluster specifications. The specification cannot be changed after the cluster is created. Optional values: cce.s2.small, cce.s2.medium, cce.s2.l...
* cce_node_pool_password	**********	string	CCE cluster node password, used for cluster node login. Valid range: 8-24 characters. The password must contain at least one uppercase letter, one lowe...
* cce_node_pool_flavor	x1.8u.16g	string	CCE cluster node cloud server instance specifications support Elastic Cloud Server (ECS) and Huawei Cloud Flexus Cloud Server X instances. The nami...
* cce_node_pool_count	3	number	The initial number of nodes in the CCE node pool, with a maximum of 10.Default: 3
* rds_flavor	rds.mysql.n1-v2.xlarge.2.ha	string	Cloud Database RDS for MySQL instance specification. This solution creates a primary/standby edition by default.Default: rds.mysql.n1-v2.xlarge.2.ha (4...
* mysql_password	**********	string	The administrator password for the MySQL database must be 8-24 characters long, including at least one uppercase letter, one lowercase letter, and one ...
* mysql_user_password	**********	string	The password for the database user of the MySQL database must be 8-24 characters long, including at least one uppercase letter, one lowercase letter, a...
* redis_capacity	1	number	Distributed Cache Service Redis Edition instance specifications. Available options: 1GB-64GB.Default: 1GB
* redis_password	**********	string	Redis database password. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase letters, and include numbers or...
* session_secret	**********	string	Session key for the New API management backend. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase letter...
* crypto_secret	**********	string	Encryption key used for encrypting large model API keys. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase l...
* charging_mode	postPaid	string	Billing mode, default automatic deduction. Valid values:postPaid (pay-as-you-go), prePaid (yearly/monthly). Default: postPaid.
* charging_unit	month	string	Order period type. This parameter is mandatory when charging_mode is prePaid. Valid values: month (monthly) or year (yearly).Default: month.
* charging_period	1	number	Billing period. This parameter is mandatory when charging_mode is prePaid. Valid values: 1-3 (year) or 1-9 (month).Default: 1 month.

Table 3-1 Community Standalone Edition Configuration Instructions

Parameter	Type	Mandatory	Description	Default Value
vpc_name	string	Mandatory	Virtual private cloud name. This template uses a newly created VPC, and duplicate names are not allowed. Value range:1-54 characters, supporting Chinese characters, English letters, numbers, (_,-,.).	building-a-newapi-llm-gateway-demo
security_group_name	string	Mandatory	Security group name. This template creates a new security group; please refer to the deployment guide for configuring security group rules. Value range:1-64 characters, supporting numbers, letters, Chinese characters,(_,-,.).	building-a-newapi-llm-gateway-demo
ecs_name	string	Mandatory	Cloud server instance name, duplicate names are not supported. Value range: 1-64 characters, supporting Chinese characters, English letters, and numbers, (_,-,.).	building-a-newapi-llm-gateway-demo

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Type	Mandatory	Description	Default Value
ecs_flavor	string	Mandatory	Cloud server instance specification. Supports Elastic Cloud Server (ECS) and Huawei Cloud Flexus Cloud Server X instances. The naming rule for Flexus Cloud Server X instance specification IDs is x1.?u.?g. For example, the specification ID for 2vCPUs4GiB is x1.2u.4g. For specific Huawei Cloud Flexus Cloud Server X instance specifications, refer to **the X instance console configuration**. For Elastic Cloud Server (ECS) specifications, refer to **the ECS console configuration.**	x1.8u.16g
ecs_password	string	Mandatory	loud server password. Length: 8-26 characters. The password must contain at least three of the following four types: uppercase letters, lowercase letters, digits, and special characters (!@$%^-_=+[{}]:,./?).	root
system_disk_size	number	Mandatory	Cloud server system disk size. The disk type defaults to High I/O. Unit: GB. Value range: 40-1,024. Disk shrinking is not supported.	100

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Type	Mandatory	Description	Default Value
bandwidth_size	number	Mandatory	Elastic IP bandwidth size. The billing method 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 deduction. Valid values: prePaid (yearly/monthly package) or postPaid (pay-as-you-go).	postPaid
charging_unit	string	Mandatory	Order period type. This parameter is mandatory when charging_mode is prePaid. Valid values: month (monthly) or year (yearly).	month
charging_period	number	Mandatory	Billing period. This parameter is mandatory when charging_mode is prePaid. Valid values: 1-3 (year) or 1-9 (month).	1

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 3-2 CCE Container High Availability Edition Configuration Instructions

Parameter	Type	Mandatory	Description	Default Value
resource_name_prefix	string	Mandatory	Resource name prefix naming rule: {resource_name_prefix}-English resource name. For example: CCE cluster name is {resource_name_prefix}-cce. Value range: 4-24 characters, supports lowercase letters, numbers, and hyphens (-). Must start with a lowercase letter. Cannot start with a hyphen (-).	ha-new-api
cce_cluster_flavor	string	Mandatory	CCE Turbo cluster specifications. The specification cannot be changed after the cluster is created. Optional values: cce.s2.small, cce.s2.medium, cce.s2.large, cce.s2.xlarge. For specific specifications, please refer to **the deployment guide configuration.**	cce.s2.small
cce_node_pool_password	string	Mandatory	CCE cluster node password, used for cluster node login. Valid range: 8-24 characters. The password must contain at least one uppercase letter, one lowercase letter, and include either a digit or a special character (~!@#$^*-=_,?).	-

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Type	Mandatory	Description	Default Value
cce_node_pool_flavor	string	Mandatory	CCE cluster node cloud server instance specifications support Elastic Cloud Server (ECS) and Huawei Cloud Flexus Cloud Server X instances. The naming convention for Flexus Cloud Server X instance specification IDs is x1.?u.?g. For example, the specification ID for 2vCPUs4GiB is x1.2u.4g. Please use specifications of 3vCPUs6GiB or above. For specific Huawei Cloud Flexus X instance specifications, refer to **the X instance console configuration**. For Elastic Cloud Server (ECS) specifications, refer to **the ECS console configuration.**	x1.8u.16g
cce_node_pool_count	number	Mandatory	The initial number of nodes in the CCE node pool, with a maximum of 10.	3
rds_flavor	string	Mandatory	Cloud Database RDS for MySQL instance specification. This solution creates a primary/standby edition by default. For other specifications, please refer to **the RDS console configuration**.	rds.mysql.n1-v2.xlarge.2.ha(4U8G)

Parameter	Type	Mandatory	Description	Default Value
mysql_password	string	Mandatory	The administrator password for the MySQL database must be 8-24 characters long, including at least one uppercase letter, one lowercase letter, and one number or special character (~!^*-=_+,).	-
mysql_user_password	string	Mandatory	The password for the database user of the MySQL database must be 8-24 characters long, including at least one uppercase letter, one lowercase letter, and one number or special character (~!^*-=_+,). It cannot be identical to the username or the reverse of the username.	-
redis_capacity	number	Mandatory	Distributed Cache Service Redis Edition instance specifications. Available options: 1GB-64GB.	1GB
redis_password	string	Mandatory	Redis database password. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase letters, and include numbers or special characters (~!^*-=_+,).	-

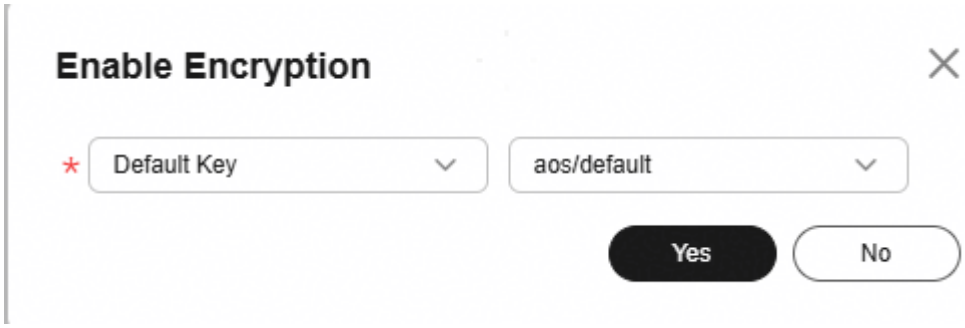
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Type	Mandatory	Description	Default Value
session_secret	string	Mandatory	Session key for the New API management backend. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase letters, and include numbers or special characters (~!^*-=_+,).	-
crypto_secret	string	Mandatory	Encryption key used for encrypting large model API keys. Valid range: 8-24 characters. The password must contain at least uppercase letters, lowercase letters, and include numbers or special characters (~!^*-=_+,).	-
charging_mode	string	Mandatory	Billing mode, default automatic deduction. Valid values: prePaid (yearly/monthly package) or postPaid (pay-as-you-go).	postPaid
charging_unit	string	Mandatory	Order period type. This parameter is mandatory when charging_mode is prePaid. Valid values: month (monthly) or year (yearly).	month
charging_period	number	Mandatory	Billing period. This parameter is mandatory when charging_mode is prePaid. Valid values: 1-3 (year) or 1-9 (month).	1

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

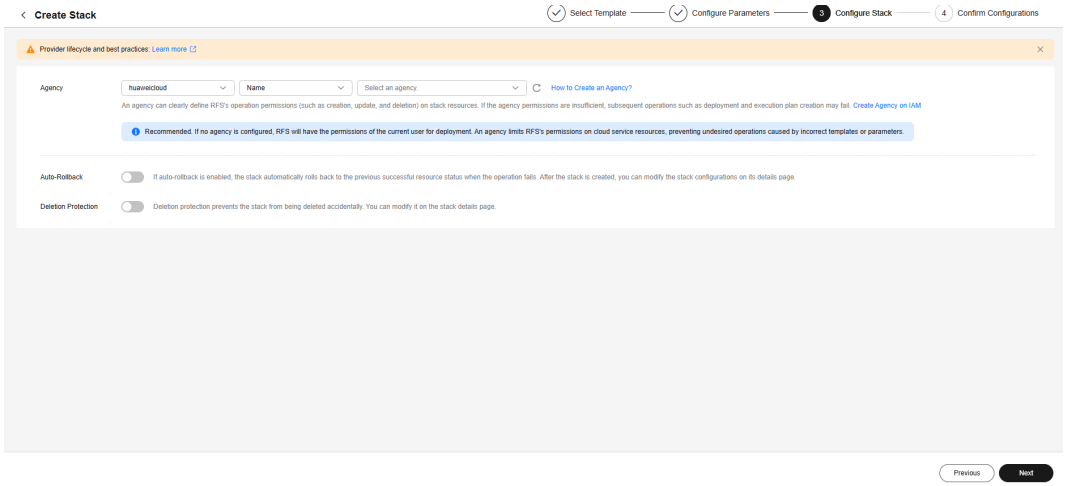
Step 3 (Optional) In the encryption activation interface, select "**Custom Key**" from the key dropdown menu, then click "**OK**".

Figure 3-13 Encryption Settings



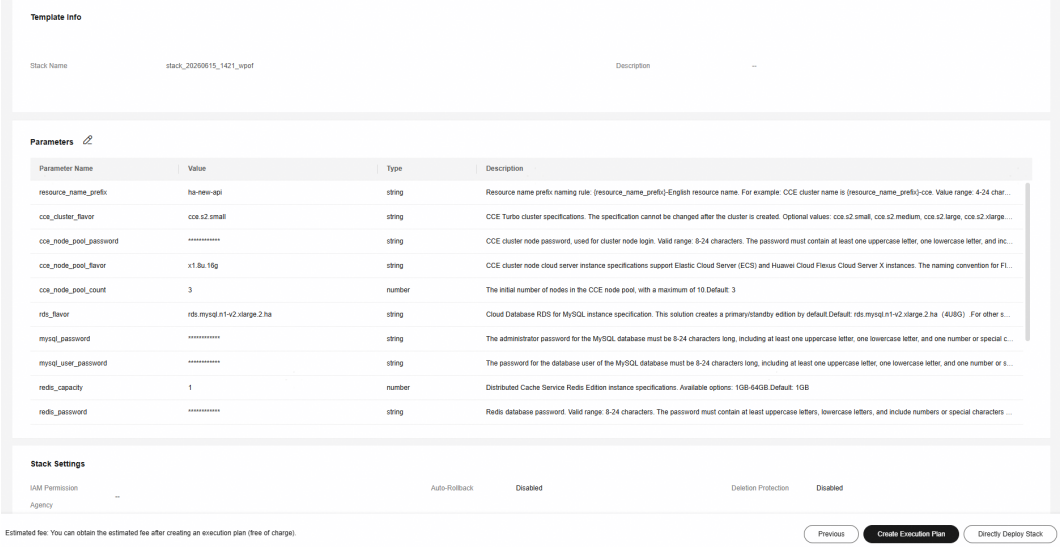
Step 4 On the **Configure Stack** page, select "**rf_admin_trust**" from the Agency drop-down list and click "**Next**" (This step is optional if you use an account (HUAWEI ID) or use an IAM user in the admin user group.).

Figure 3-14 Configuring a stack



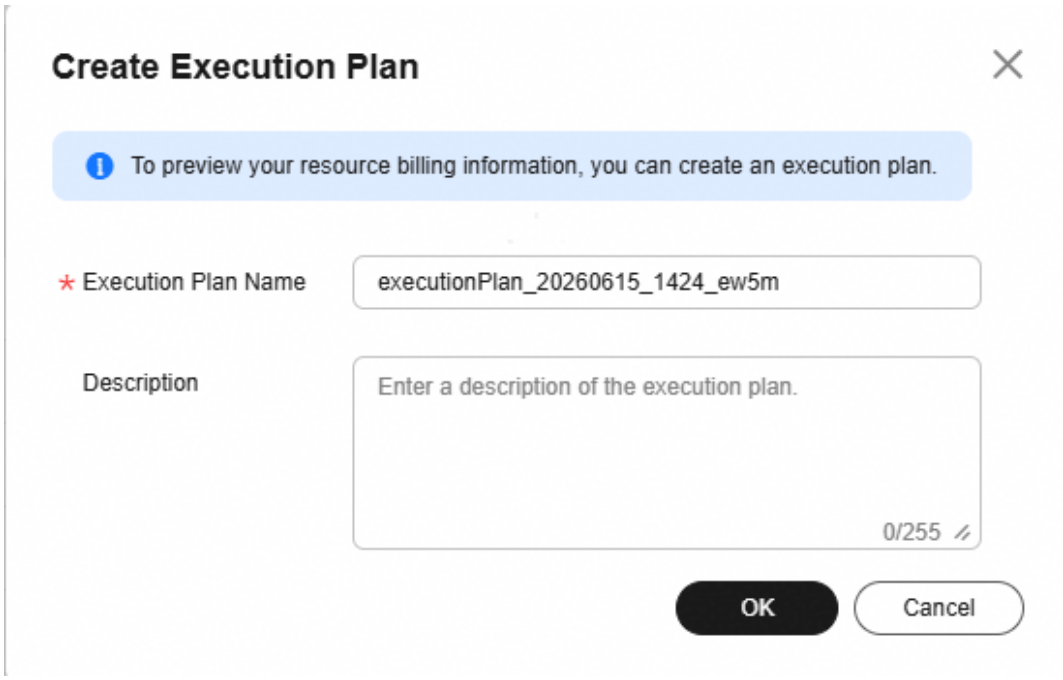
Step 5 On the **Confirm Configurations** page, confirm the configurations and click "**Create Execution Plan**".

Figure 3-15 Confirming the configurations



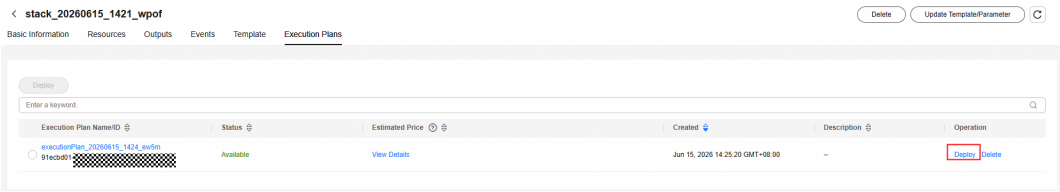
Step 6 In the displayed **Create Execution Plan** dialog box, enter an execution plan name and click "**OK**".

Figure 3-16 Creating an execution plan



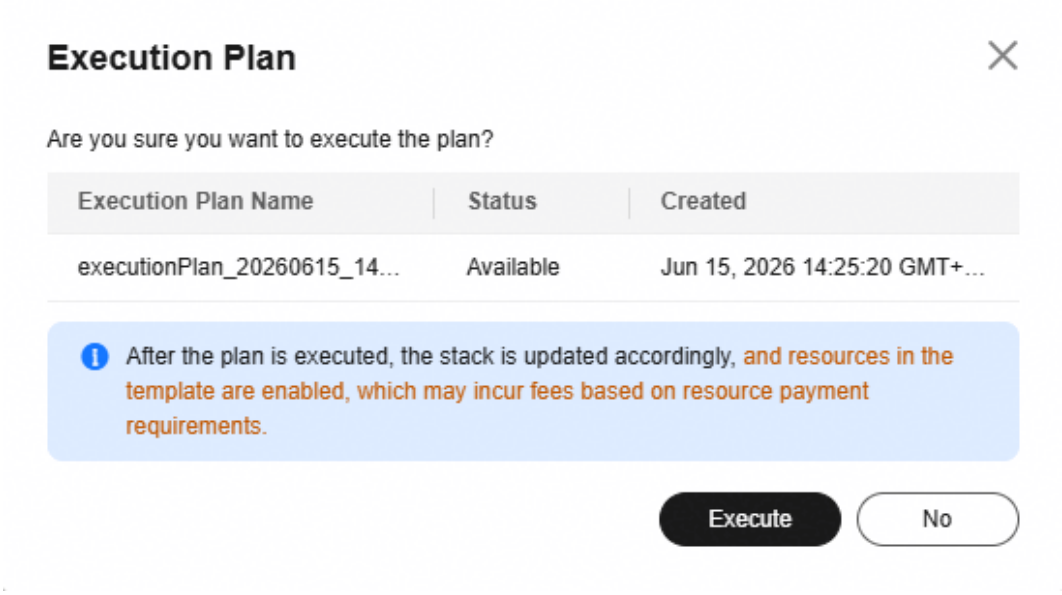
Step 7 Wait until the status of the execution plan changes to **Available** and then click "**Deploy**" in the **Operation** column.

Figure 3-17 Execution plan page



Step 8 In the displayed dialog box, click **"Execute".**The system will complete the deduction through automatic payment. Please ensure that your account has sufficient balance.

Figure 3-18 Confirming the execution plan



CAUTION

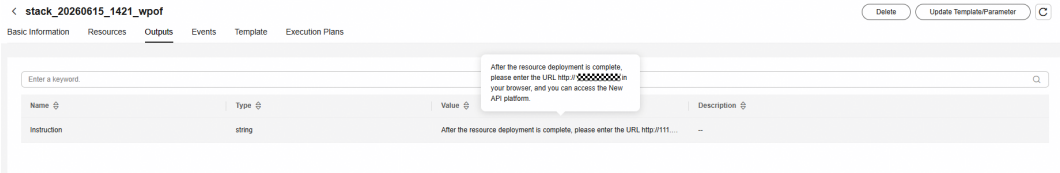
- When creating a new account or when the account balance is insufficient, please go to the "Billing Center > **Top-up**" page to recharge. For details, please refer to**Account Top-up**.
- If you have selected a yearly/monthly billing mode and automatic payment fails due to insufficient account balance, please go to the "Billing Center > **Unpaid Orders**" page to complete the payment manually.

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.

Figure 3-19 Solution deployed

Step 10 On the **Outputs** tab, view the access description of the platform. After the stack is successfully deployed, the application setup script starts to execute. Please wait patiently for about 20 minutes (subject to network fluctuations).

Figure 3-20 Access description



----End

3.3 USing Started

(Optional) Modifying Security Group Rules

NOTICE

- The Community Standalone Edition of this solution uses server port 3000 to access the New API service, while the CCE Container High Availability Edition uses ELB port 80 to access the New API service. It is open to all by default. Please refer to **modifying the security group rules** and configure the IP address whitelist.
- This solution uses port 22 for remote login to the cloud server via SSH. If remote login is required, please refer to **modifying the security group rules** and configure the IP address whitelist to ensure normal access to the service.
- After successful deployment of this solution, environment initialization is expected to take approximately 20 minutes. Deployment time may vary due to network and bandwidth factors. The service can be accessed only after deployment is completed.

New API login

Step 1 Log in to the development platform: Enter the access address from **Quick Deployment step10** process to access the platform. For the first login, complete the system initialization by following the prompts step by step. Choose the usage mode according to your actual needs.

Figure 3-21 Database check

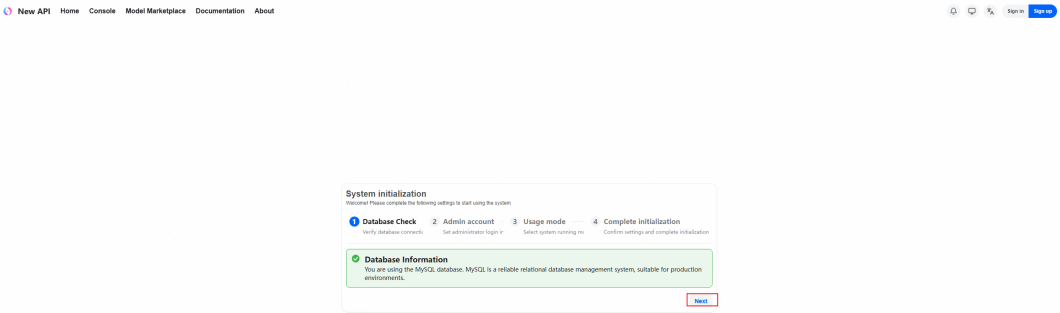


Figure 3-22 Set up administrator account

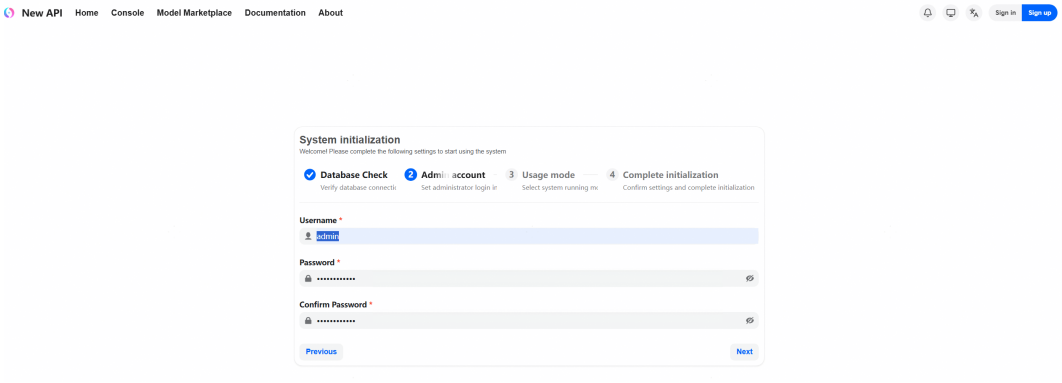


Figure 3-23 Usage mode (select mode based on actual needs)

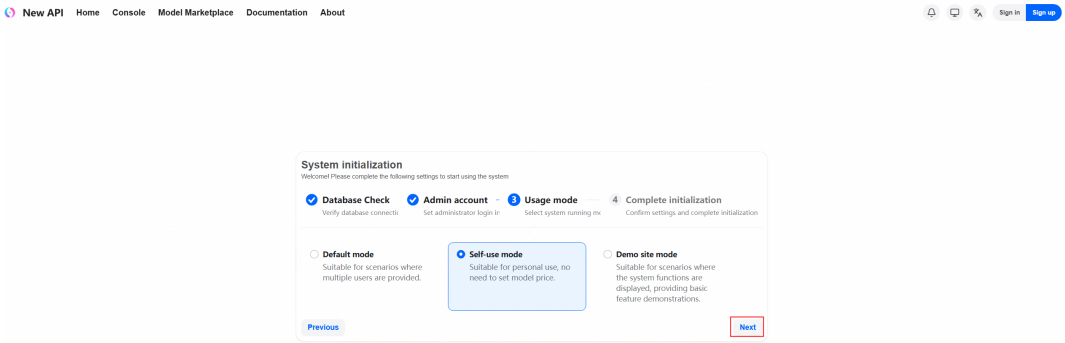
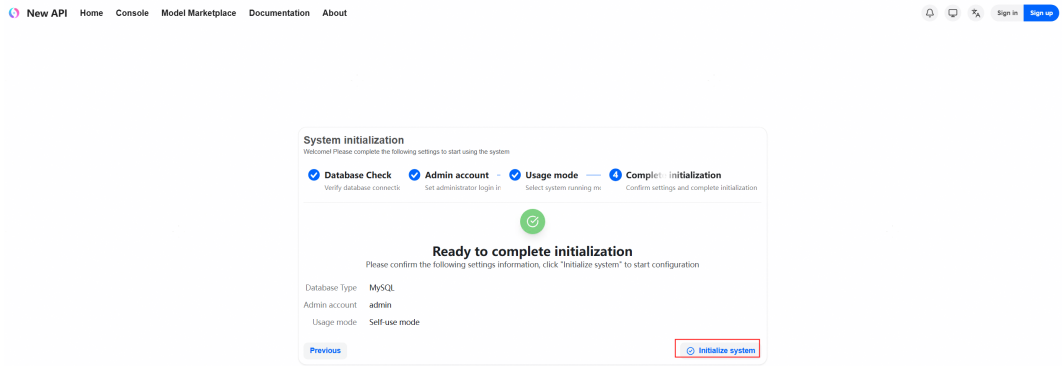


Figure 3-24 Initialization completed



Step 2 After initialization is completed, it will automatically jump to the homepage. Click "Login" in the upper right corner, then enter the "Username" and "Password" from the previous step to log in to the New API Gateway platform.

Figure 3-25 Login

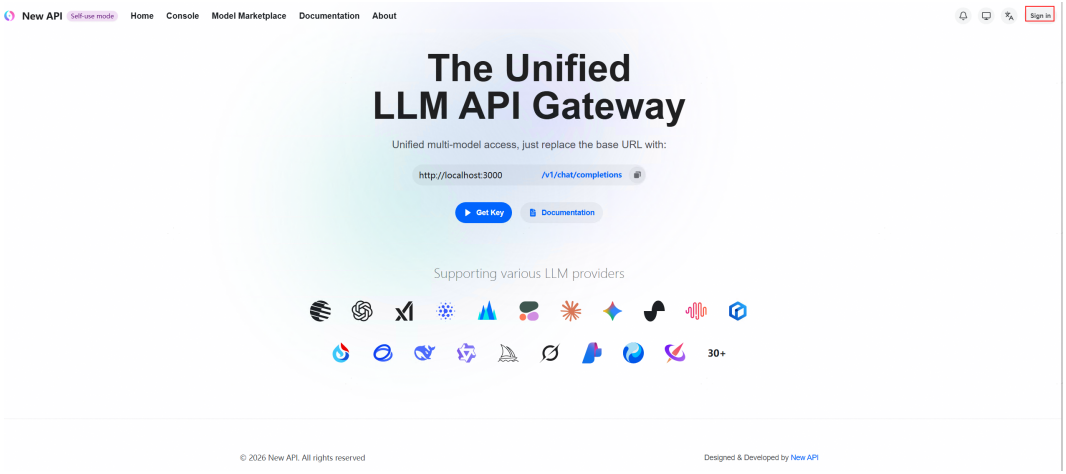
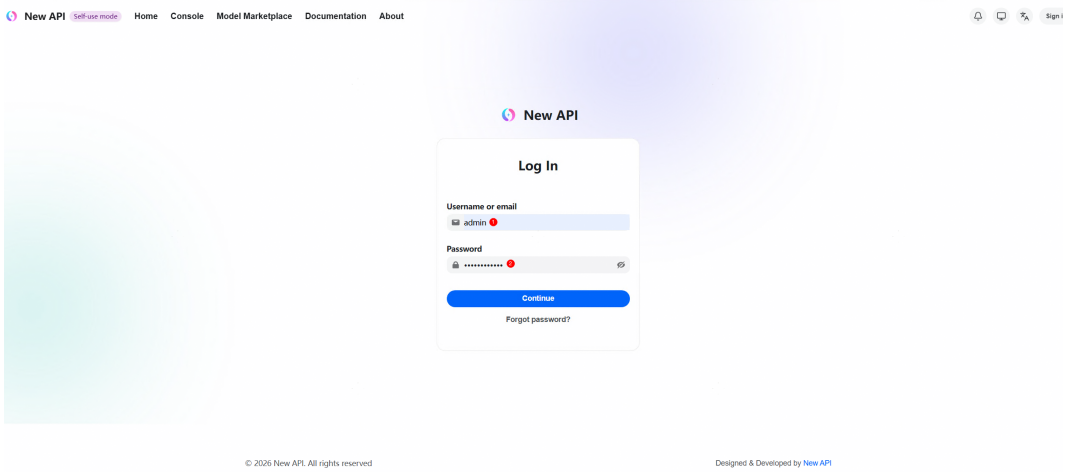


Figure 3-26 Enter the Username and Password



----End

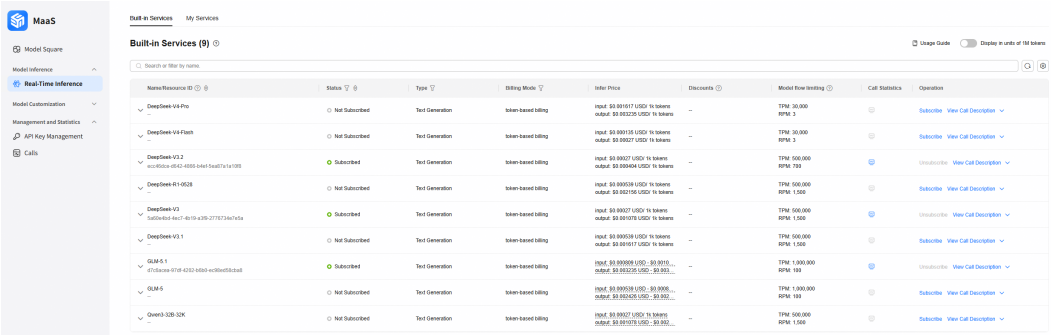
NOTE

- Users can choose the following methods for connecting to large models based on their own situation.

Integration with MaaS services

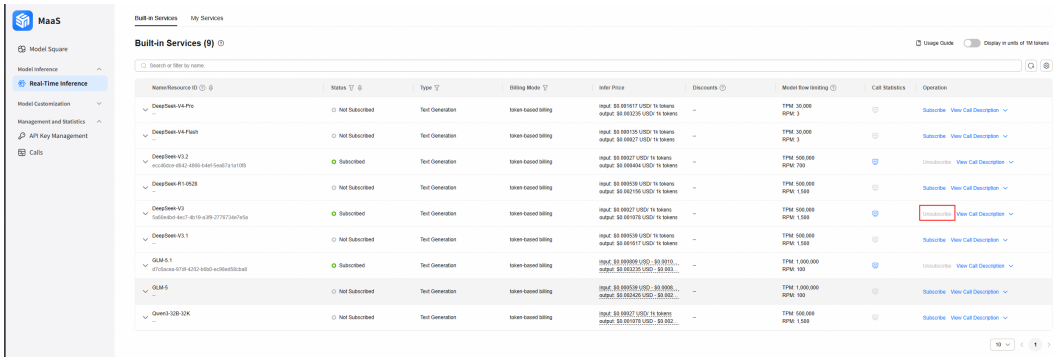
Step 1 Log in to the **MaaS platform**. This document takes deploying the DeepSeek-v3 in China-Hong Kong as an example. In the left navigation bar, select "Online Inference".

Figure 3-27 MaaS platform



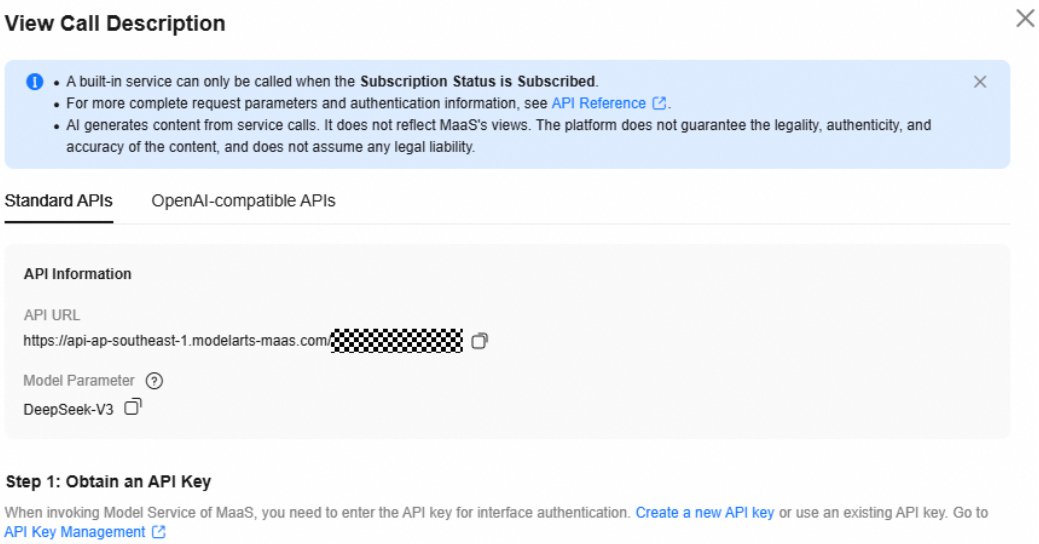
Step 2 In the preset service list, select the desired service and click "Activate Service" in the operation column on the right. When grayed out, it indicates that the service has been activated.

Figure 3-28 Activate Service



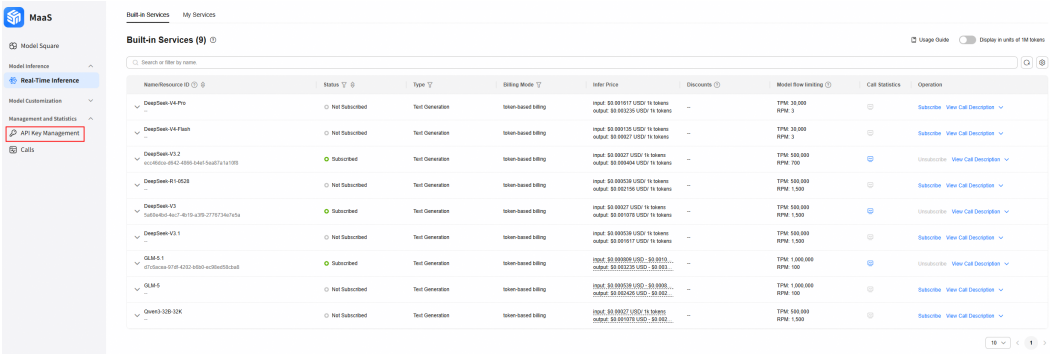
Step 3 After successful activation, select the desired service in the service list, click "Usage Instructions", and choose "Standard Interface" as the interface type in the pop-up window to obtain the API address and model name.

Figure 3-29 Usage Instructions



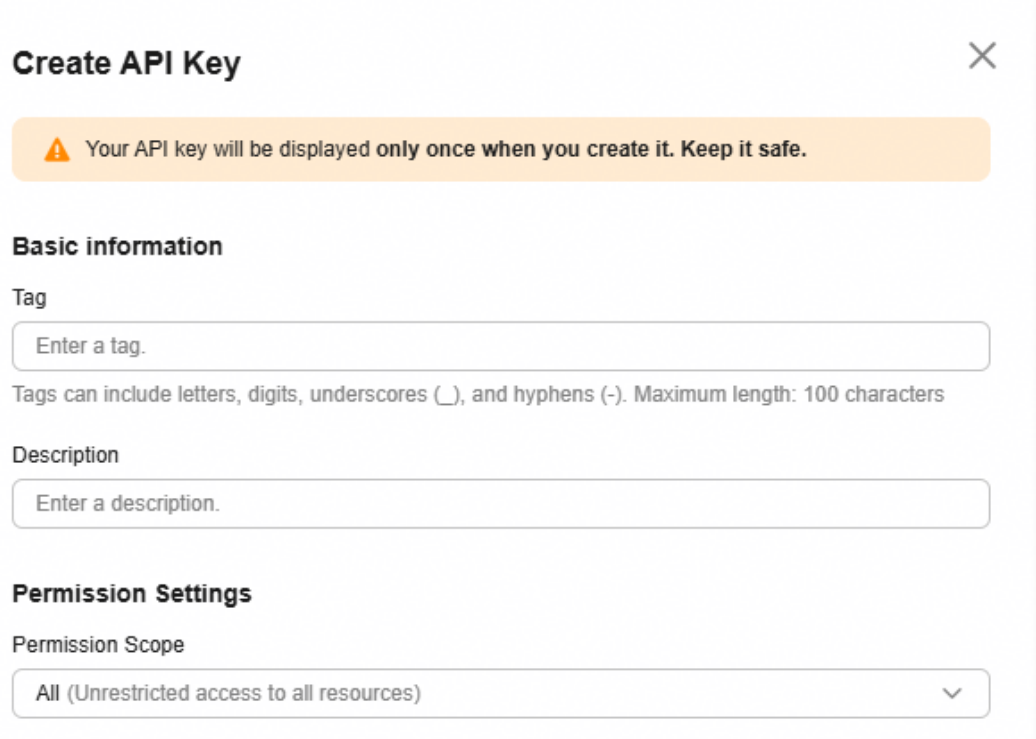
Step 4 When invoking model services deployed on MaaS, you need to fill in the API Key for interface authentication. In the left navigation bar, click "API Key Management" (you can create up to 30 keys. **Each key is only displayed once upon creation, so please ensure it is properly saved. If a key is lost, it cannot be retrieved, and you will need to create a new API Key to obtain a new access key**).

Figure 3-30 API Key Management



Step 5 On the "API Key Management" page, click "Create API Key" in the upper-right corner, fill in the label (custom label for the API Key; labels must be unique and cannot be duplicated. Only uppercase and lowercase English letters, numbers, underscores, and hyphens are supported, with a length range of 1–100 characters) and description (custom description for the API Key), then click "OK." The label and description information cannot be modified after creation.

Figure 3-31 Create API Key

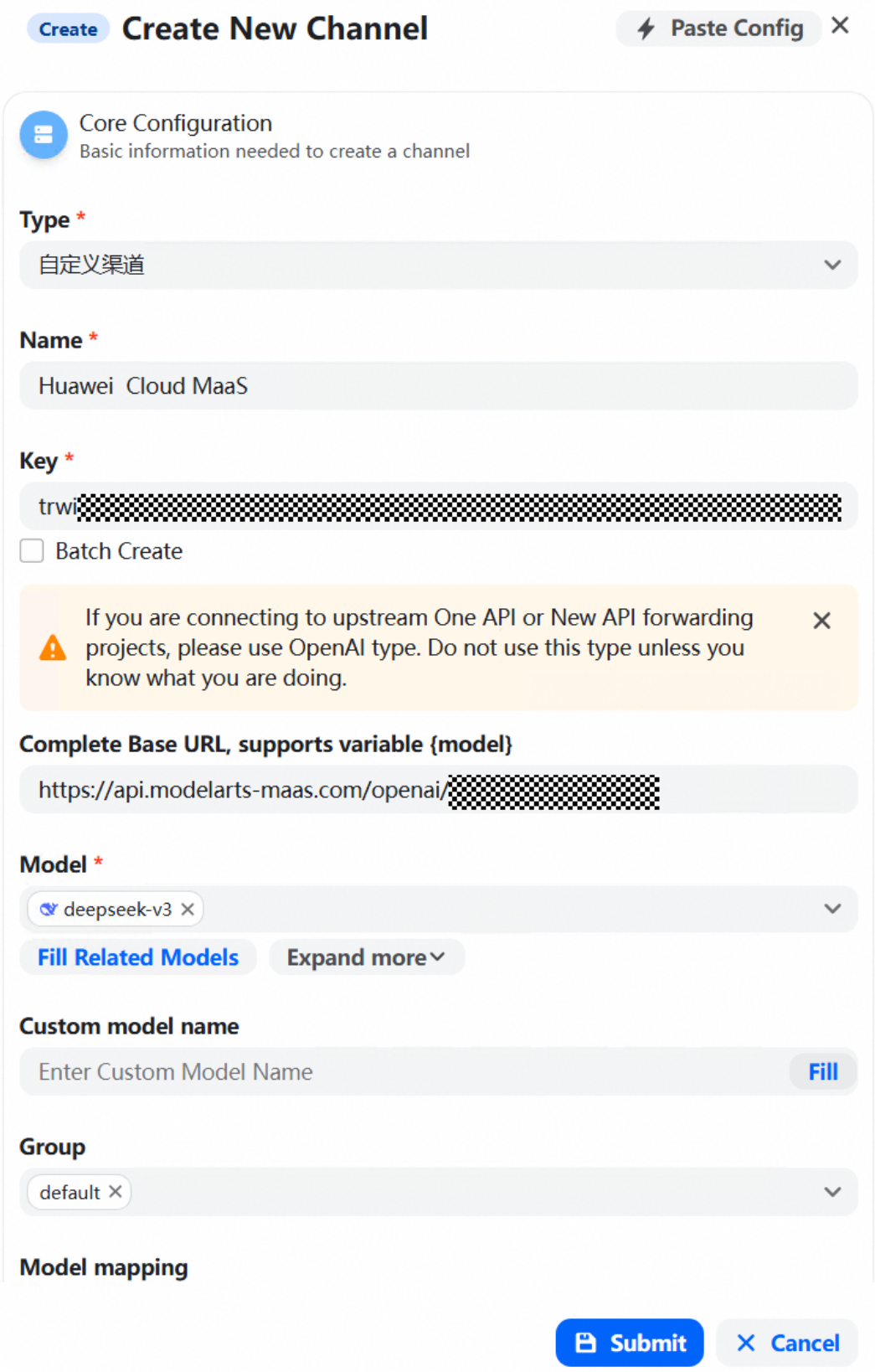


Step 6 Integrate with the New API Gateway. Open the New API Gateway platform, click "Channel Management" followed by "Add Channel" in sequence, fill in the information in the pop-up window as shown in the figure below, and then click "Submit" when done.

NOTE

When adding a channel, if the type selected is OpenAI, choose the "OpenAI" interface corresponding to **Figure 9 Usage Instructions** to obtain the API address and model name.

Figure 3-32 Add Huawei Cloud MaaS.



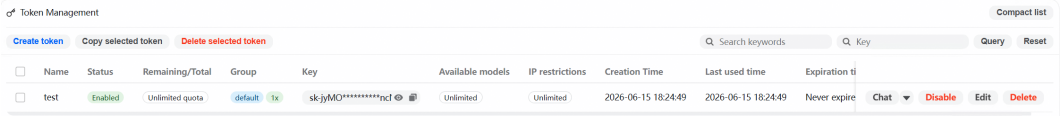
----End

Basic Usage of New API

Taking token and channel addition as an example, please refer to **the official guide** for more detailed instructions.

Step 1 Configure the token as shown in the figure below.

Figure 3-33 Add token



Step 2 Configure the channel as shown in the figure below.

Figure 3-34 Add channel

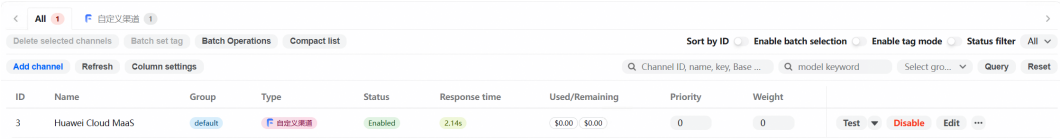
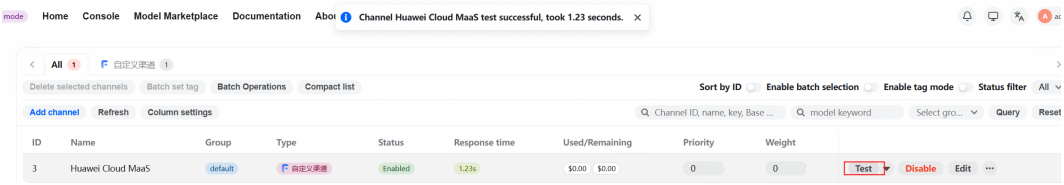


Figure 3-35 Test channel



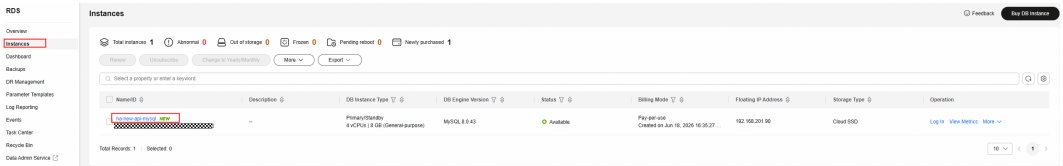
----End

3.4 Quick Uninstallation

Step 1 Before deleting the stack for the high availability deployment version, ensure that the New API external service has been stopped. If database connections exist, please disconnect the processes before attempting to delete the stack. For terminating the database processes, refer to steps 2-3 below (non-CCE container high availability versions can skip directly to step 4).

Step 2 Access the **RDS console**, click on the database instance name, and enter the database details page.

Figure 3-36 Database Instance



Step 3 Click "Intelligent DBA Assistant > Sessions" sequentially to view all current real-time sessions of the database. Select all processes with the database name

"newip" and username "newip", then click "Kill Session" and confirm in the pop-up window to terminate the connection.

Figure 3-37 Sessions

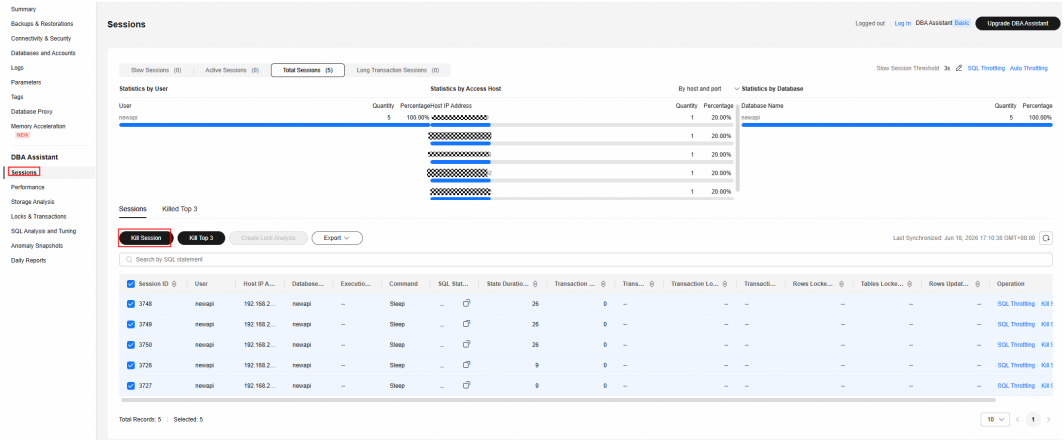
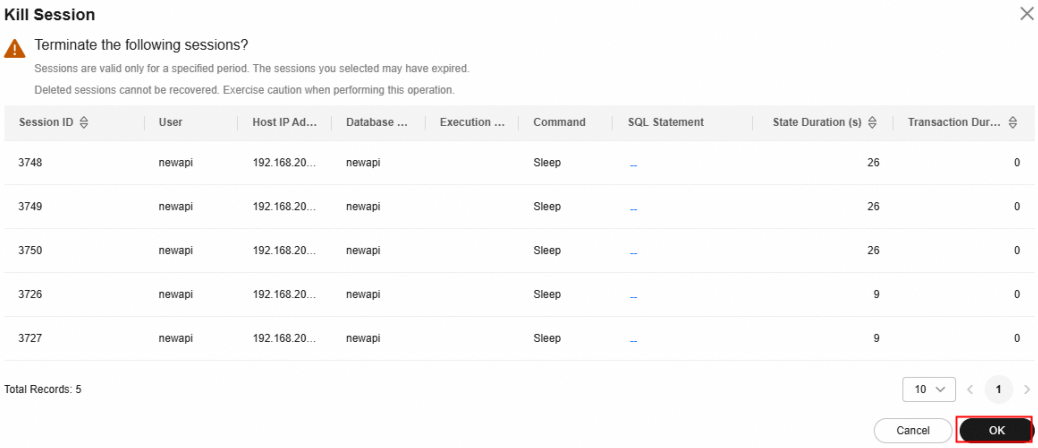
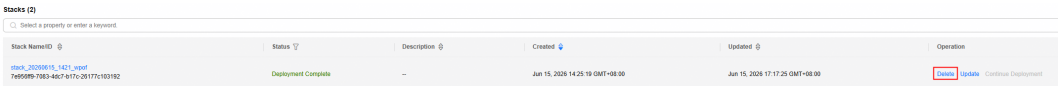


Figure 3-38 kill session



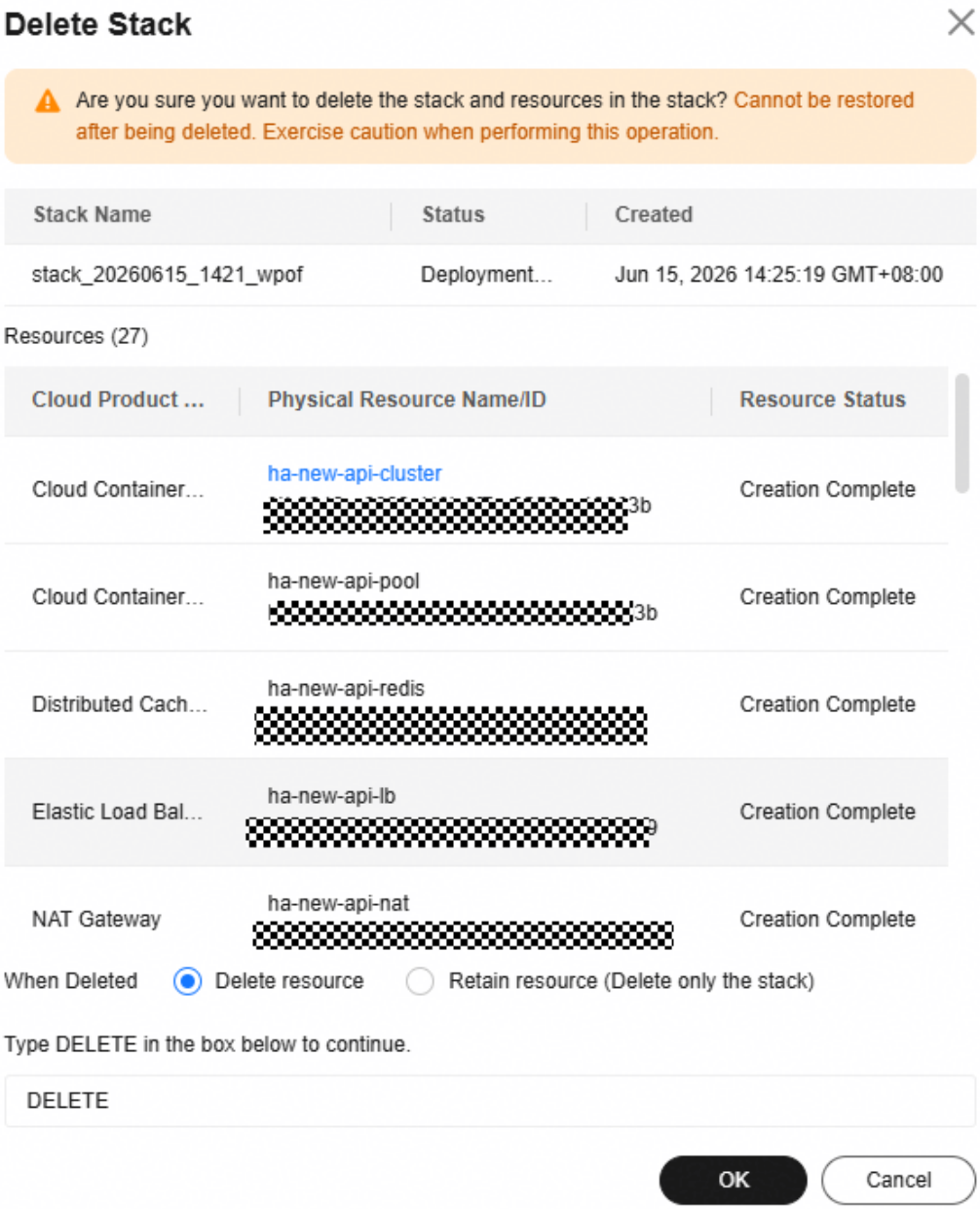
Step 4 Log in to the **RFS console**. On the Stacks page, locate the resource stack you created and click "Delete" in the Operation column.

Figure 3-39 Deleting a stack



Step 5 In the displayed Delete Stack dialog box, set When Deleted to Delete resource, enter "DELETE" and click OK.

Figure 3-40 Confirming the deletion



----End

4 Appendix

Terms

- **Flexus X Instance (FlexusX):** FlexusX is a next-generation flexible compute cloud server service designed for small- and medium-sized enterprises (SMEs) and developers. FlexusX offers similar functions to ECS. It also supports flexible vCPU/memory ratios and performance mode.
- **Virtual Private Cloud (VPC)** is an isolated and private virtual network environment that users can apply for on Huawei Cloud. Users can build an independent cloud-based network space based on VPC and integrate it with services such as Elastic IP, Cloud Connect, and Cloud Direct Connect to achieve connectivity with the Internet, private networks within the cloud, and cross-cloud private networks, helping you create a reliable, stable, and efficient dedicated cloud network.
- **Elastic IP (EIP):** An EIP is a public IP address that can be accessed directly over the Internet. An EIP consists of a public IP address and some amount of public network egress bandwidth. EIPs can be bound to or unbound from ECSs, BMSs, virtual IP addresses, NAT gateways, and load balancers.
- **Elastic Load Balance (ELB)** distributes incoming traffic across multiple backend servers based on listening rules. This expands service capabilities of applications and improves their fault tolerance.
- **NAT Gateway:** A public NAT gateway provides SNAT and DNAT so that cloud servers in a VPC can share an EIP to communicate with the Internet.
- **Cloud Container Engine (CCE)** provides highly scalable, high-performance, enterprise-class Kubernetes clusters. With CCE, you can easily deploy, manage, and scale containerized applications on HUAWEI CLOUD.
- **HUAWEI CLOUD RDS** is a relational database service that is reliable, scalable, secure, economical, easy to manage, and immediately ready for use. With RDS, you can create single or primary/standby DB instances.
- **Distributed Cache Service (DCS)** is an online, distributed, in-memory cache service. It is reliable, scalable, usable out of the box, and easy to manage. It is compatible with Redis®*, and can meet requirements for high concurrency and fast data access.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

5 Change History

Released On	Description
2026-07-8	This issue is the first official release.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.