

Solution Practices

Quickly Implement Intelligent Data Querie Based on SQLBot

Issue 01
Date 2026-07-24



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.....

8

3.1 Preparations.....

8

3.2 SQLBot Deploy.....

11

3.3 DWS Purchase (Optional).....

17

3.4 Using Started.....

18

3.5 Quick Uninstallation.....

31

4 Appendix.....

34

5 Change History.....

35

1 Solution Overview

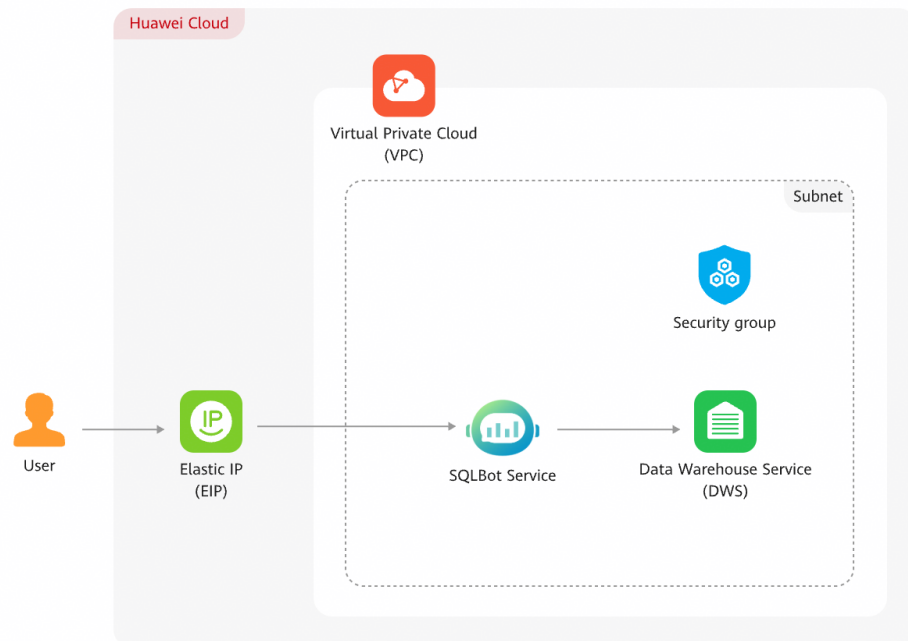
Application Scenarios

SQLBot is an intelligent data analytics system based on Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). With SQLBot, users can achieve instant data Q&A, quickly extract required data information and visual charts, and further conduct intelligent analysis. This solution uses the Data Warehouse Service (DWS) as the data source, combining DWS's powerful data analysis and querying capabilities with SQLBot's interactive intelligent data querying experience. MaaS provides DeepSeek large model inference capabilities. The core objective is to build a data analysis system that "understands human language," enabling business personnel to ask questions directly in natural language and quickly gain data insights.

Architecture

This solution uses the DWS the data source to help you quickly deploy an intelligent data analytics platform. **You can select existing databases or data warehouses as the data source according to your actual needs.**

Figure 1-1 Architecture



This solution will:

Community Standalone Edition:

- Create three **Flexus X Instance** (FlexusX) : Used to deploy the SQLBot intelligent data querying application.
- Create one **Elastic IP (EIP)** : Provides capabilities for accessing the public network and being accessed from the public network.
- Create one **Data Warehouse Service (DWS)**: Provide data source for intelligent data querying operations.

Advantages

- **Rich Data Source Support**
Supports integration with Excel/CSV, common databases such as MySQL, PostgreSQL, Dameng, Oracle, KingBase, and SQL Server, as well as data warehouses like StarRocks, ClickHouse, and Elasticsearch, covering a wide range of data analysis scenarios.
- **Out-of-the-Box**
Configure the large model and data source with a single click to quickly start intelligent data querying without coding. No need for SQL, Python, or data analysis skills; users can directly ask questions using natural language in the chat window, leveraging the large model's natural language understanding and SQL generation capabilities to lower the barrier to entry.
- **Secure and Controllable**
Provides workspace-level resource isolation to establish clear data boundaries and ensure data access security. Supports fine-grained data permission

configuration to strengthen access control and ensure compliance and controllability during use.

- Increasingly Accurate

Supports custom prompts and terminology library configuration, allowing maintenance of SQL examples to calibrate logic and match business scenarios.

- Easy Integration

Provides capabilities for Web embedding, popup embedding, and MCP calls; can be quickly embedded into applications like Dify.

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.

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	Resource Name	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Virtual Private Cloud (VPC)	intelligent-data-insight-vpc	- Region: AP-Singapore - CIDR Block: 192.168.0.0/16	1	0.00
Subnet	intelligent-data-insight-subnet	- Region: AP-Singapore - IPv4 CIDR Block: 192.168.0.1/24	1	0.00
SecurityGroup	intelligent-data-insight-sg	- 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

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Flexus X Instance	intelligent-data-insight-sqlbot-ecs	- Region: AP-Singapore - Specifications: Flexus X Instance Performance Mode (Disabled) x1.8u.16g 8 vCPUs 16 GiB - System Disk: High I/O 100GB - Image: Ubuntu 22.04 server 64bit	1	\$0.2862 USD
Elastic IP (EIP)	custom	- Region: AP-Singapore - Routing Type: Dynamic BGP - Billing Mode: traffic - Bandwidth: 300Mbit/s	1	\$0.114 USD/GB
Data Warehouse Service (DWS)	intelligent-data-insight-sqlbot-ecs-eip	- Region: AP-Singapore - Specifications: Specification: Storage-Compute Separation X86 8 vCPUs 32 GB * 3 Nodes; Hot Data 100 GB * 3 Nodes; Cold Data billed based on actual storage usage, no pre-purchase required. NOTICE Supports data sources such as local Excel/CSV, MySQL, PostgreSQL, SQL Server, Oracle, ClickHouse, Elasticsearch, StarRocks, and DWS. This solution uses DWS as an example.	1	\$3.63 USD
MaaS tokens (Optional)	-	- Region: AP-Singapore - Billing Mode: Token-based billing - Model: DeepSeek-V3	-	Input: \$0.27 USD/1M tokens Output: \$1.078 USD/1M tokens

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Cost for One Complete Process (1 hour)
Total	-	-	-	\$3.9162 USD + Public network traffic price + MaaS tokens billing

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

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Virtual Private Cloud (VPC)	intelligent-data-insight-vpc	- Region: AP-Singapore - CIDR Block: 192.168.0.0/16	1	0.00
Subnet	intelligent-data-insight-subnet	- Region: AP-Singapore - IPv4 CIDR Block: 192.168.0.1/24	1	0.00
SecurityGroup	intelligent-data-insight-sg	- 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	intelligent-data-insight-sqlbot-ecs	- Region: AP-Singapore - Specifications: Flexus X Instance Performance Mode (Disabled) x1.8u.16g 8 vCPUs 16 GiB - System Disk: High I/O 100GB - Image: Ubuntu 22.04 server 64bit	1	\$151.85 USD

Huawei Cloud Service	Resource Name	Configuration Example	Quantity	Estimated Monthly Cost
Elastic IP (EIP)	custom	- Region: AP-Singapore - Routing Type: Dynamic BGP - Billing Mode: traffic - Bandwidth: 300Mbit/s	1	\$0.114 USD/GB
Data Warehouse Service (DWS)	intelligent-data-insight-sqlbot-ecs-eip	- Region: AP-Singapore - Specifications: Specification: Storage-Compute Separation X86 8 vCPUs 32 GB * 3 Nodes; Hot Data 100 GB * 3 Nodes; Cold Data billed based on actual storage usage, no pre-purchase required. NOTICE Supports data sources such as local Excel/CSV, MySQL, PostgreSQL, SQL Server, Oracle, ClickHouse, Elasticsearch, StarRocks, and DWS. This solution uses DWS as an example.	1	\$1742.4 USD
MaaS tokens (Optional)	-	- Region: AP-Singapore - Billing Mode: Token-based billing - Model: DeepSeek-V3	-	Input: \$0.27 USD/1M tokens Output: \$1.078 USD/1M tokens
Total	-	-	-	\$1894.25 USD + Public network traffic price + MaaS tokens billing

3 Procedure

- 3.1 Preparations
- 3.2 SQLBot Deploy
- 3.3 DWS Purchase (Optional)
- 3.4 Using Started
- 3.5 Quick Uninstallation

3.1 Preparations

(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 3-1 Console page

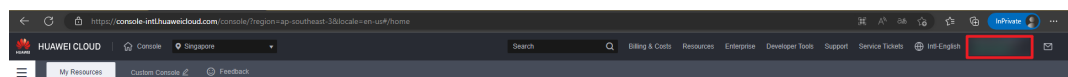
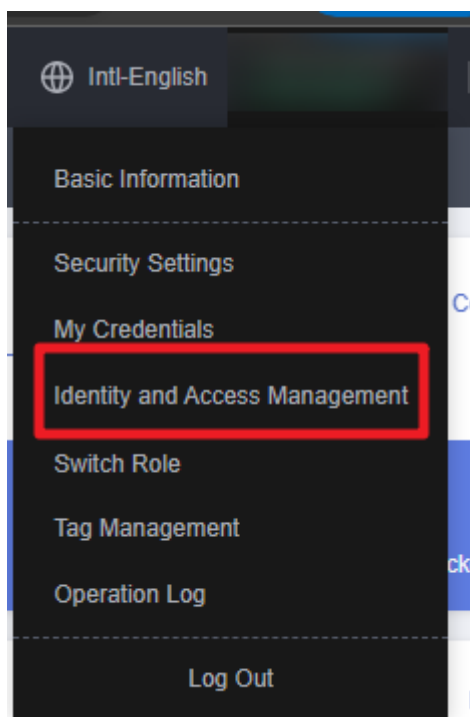
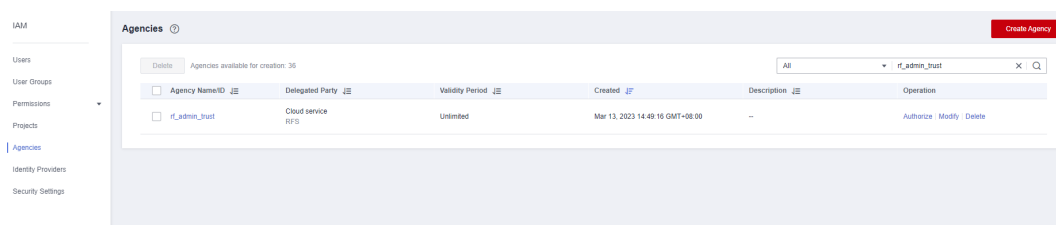


Figure 3-2 Identity and access management page



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**.

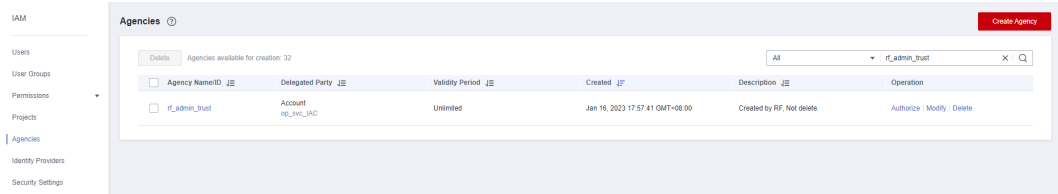
Figure 3-5 Selecting a policy/role

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 SQLBot Deploy

This section helps users efficiently deploy the "Quick intelligent data querying based on SQLBot" 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 3-8 RFS Purchase Page



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

Figure 3-9 Basic Configuration Diagram

Parameter	Value	Type	Description
* solution_name	intelligent-data-insight	string	Cloud server instance name. duplicate names are not supported. Value range: 1-64 characters, supporting Chinese characters, English letters, and numb...
* sqbot_ecs_flavor	x1.8u 16g	string	Cloud server instance specification. Supports Elastic Cloud Server (ECS) and Huawei Cloud Flexus Cloud Server X instances. The naming rule for Flexus...
* sqbot_ecs_password	*****	string	Cloud server password. Length: 8-26 characters. The password must contain at least three of the following four types: uppercase letters, lowercase letter...
* sqbot_system_disk_size	100	number	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. Default: 100
* sqbot_epb_size	300	number	Elastic IP bandwidth size. The billing method for this template is pay-by-traffic. Unit: MB/s. Value range: 1-300 MB/s. Default: 300.
* 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

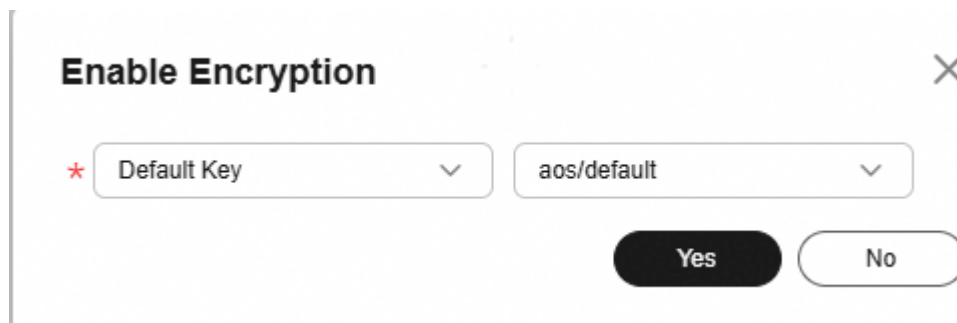
Parameter	Type	Mandatory	Description	Default Value
solution_name	string	Mandatory	Cloud server instance name, duplicate names are not supported. Value range: 1-64 characters, supporting Chinese characters, English letters, and numbers, (_ - , .) .	intelligent-data-insight
sqlbot_ecs_flavor	string	Mandatory	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

Parameter	Type	Mandatory	Description	Default Value
sqlbot_ecs_password	string	Mandatory	Cloud 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 (!@ \$%^_-=+[]{};.,/?). To change the password, refer to "Reset Elastic Cloud Server Password" and log in to the ECS console to modify it. The default administrator account is root.	-
sqlbot_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
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

Parameter	Type	Mandatory	Description	Default Value
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

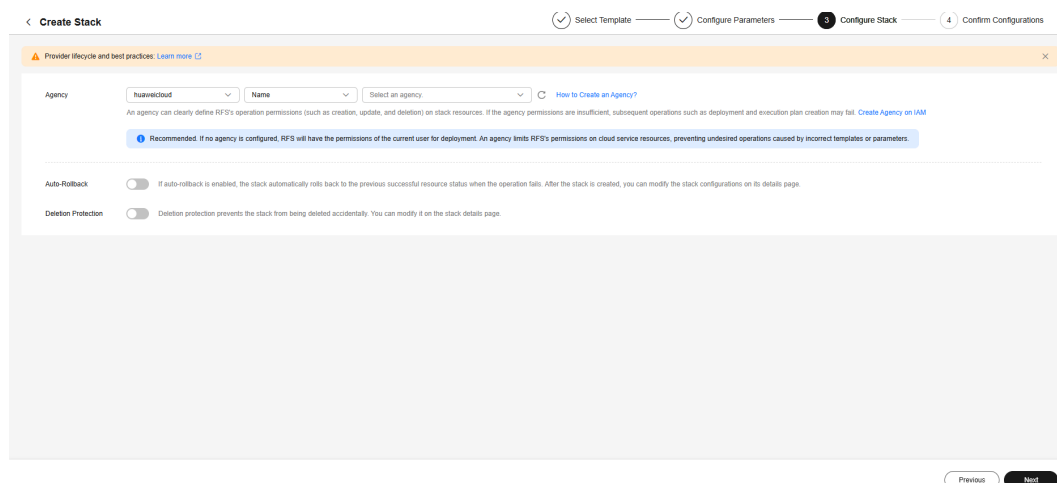
Step 3 (Optional) In the encryption activation interface, select "**Custom Key**" from the key dropdown menu, then click "**OK**".

Figure 3-10 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 3-11 Configuring a stack



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

Figure 3-12 Confirming the configurations

Parameters

Parameter Name	Value	Type	Description
solution_name	intelligent-data-insight	string	Cloud server instance name, duplicate names are not supported. Value range: 1-64 characters, supporting Chinese characters, English letters, and numbers. (_,-) Default int...
sqlbot_ecs_flavor	x1.8u.16g	string	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 inst...
sqlbot_ecs_password	*****	string	Cloud 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 c...
sqlbot_system_disk_size	100	number	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. Default: 100
sqlbot_eip_size	300	number	Elastic IP bandwidth size. The billing method for this template is pay-by-traffic. Unit: Mbit/s. Value range: 1-300 Mbit/s. Default: 300
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.

Stack Settings

IAM Permission	---	Auto Rollback	Disabled	Deletion Protection	Disabled
Agency	---				

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

Previous

Create Execution Plan

Directly Deploy Stack

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

Create Execution Plan

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

★ Execution Plan Name

executionPlan_20260722_1700_gm0r

Description

Enter a description of the execution plan.

0/255

OK

Cancel

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

Figure 3-14 Execution plan page

quickly-implement-intelligent-data-querie-based-on-sqlbot

Basic Information

Resources

Outputs

Events

Template

Execution Plans

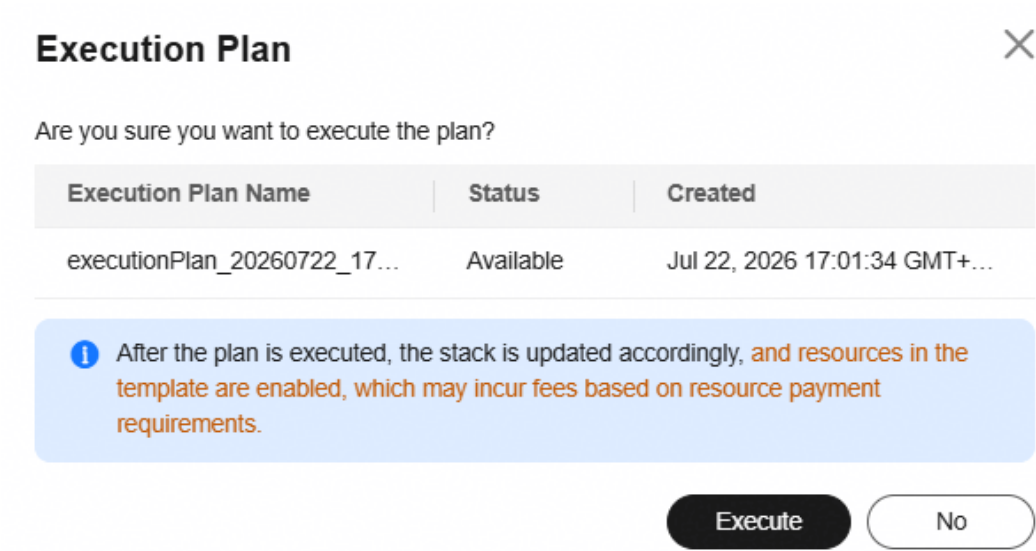
Deploy

Enter a keyword

Execution Plan NameID	Status	Estimated Price	Created	Description	Operation
executionPlan_20260722_1700_gm0r	Available	View Details	Jul 22, 2026 17:01:34 GMT+08:00	--	DeployDelete

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 3-15 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 3-16 Solution deployed



Step 10 Refresh the page, and view the access link on the "Outputs" tab of the stack. After the stack is deployed successfully, the application deployment script starts to execute. Please wait for about 10 minutes (subject to network fluctuations).

Figure 3-17 Access description



----End

3.3 DWS Purchase (Optional)

Step 1 To the **DWS console**, and select the VPC and security group that were created when deploying the SQLBot platform.

Figure 3-18 DWS Selection

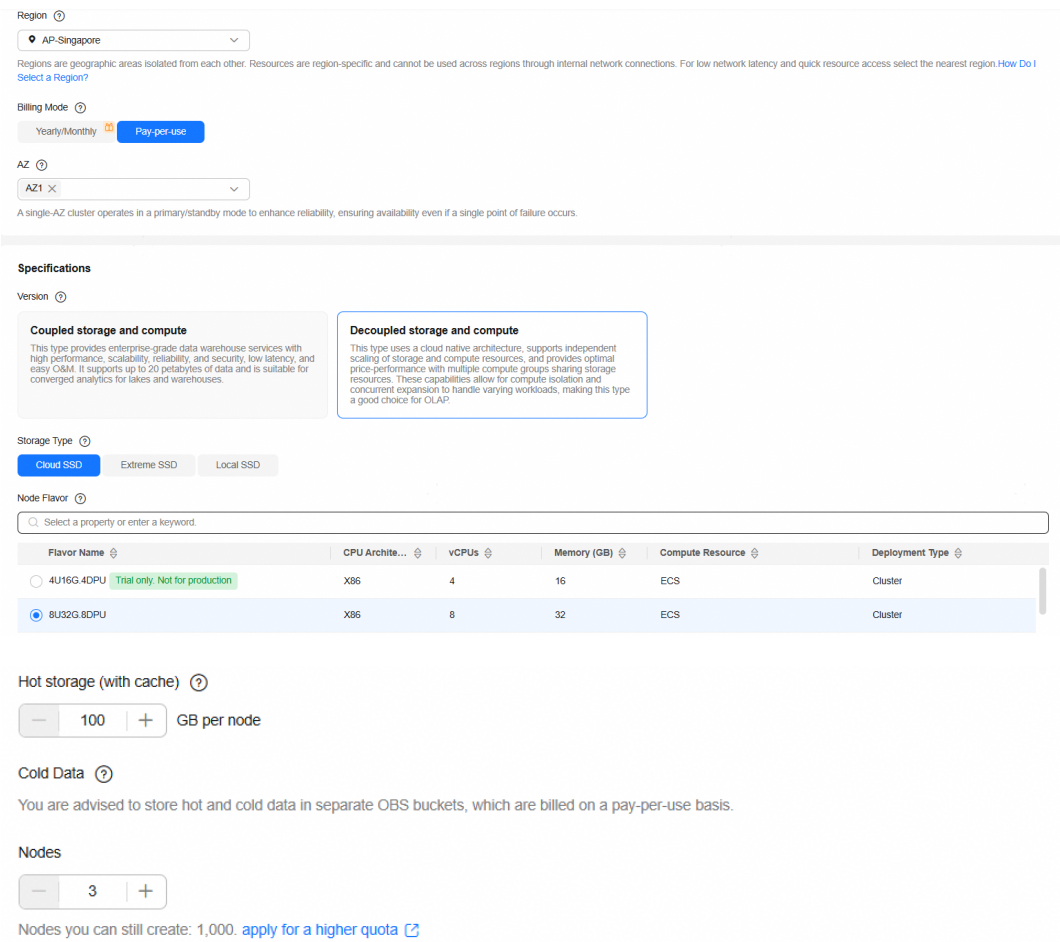


Figure 3-19 Configure DWS Network

Network

VPC

intelligent-data-insight-vpc

Create VPC

Subnet

intelligent-data-insight-subnet (192.168.0.0/16)

Create Subnet

Security Group

intelligent-data-insight-sg

Create Security Group

ELB

ELB is a service that automatically distributes incoming traffic across multiple backend cloud servers based on a specified forwarding policy. Clusters that do not use ELB cannot distribute traffic and are vulnerable to single points of failure (SPOFs).

EIP

EIP provides static public IP addresses and scalable bandwidths that enable your cloud resources to communicate with the Internet. After an EIP is bound to a cluster, the cluster can access the Internet.

Figure 3-20 Configure DWS Information

Clusters

Cluster Name

DWS_SQLBot

Enter 4 to 64 characters starting with a letter. Only letters, digits, hyphens (-), and underscores (_) are allowed.

Cluster Version

9.1.0.226 | Recommended

Release Notes | Version Policies

Time Zone

UTC+08:00 (Beijing, Chongqing, Hong Kong, ...)

Enterprise Project

default

View enterprise projects.

Tag

TMS's predefined tags are recommended for adding the same tag to different cloud resources. Create predefined tags

+ Add

Tags you can still add: 20

Database Configuration

Default database gaussdb has been set.

Administrator Account

dbadmin

Administrator Password

Confirm Password

Database Port

9000

Port for clients to connect to DWS. The value ranges from 8000 to 30000.

----End

3.4 USing Started

(Optional) Modifying Security Group Rules

NOTICE

- This solution uses port 8000 to access SQLBot. 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.

New API login

Step 1 Log in to the SQLBot platform: Enter the access address "http://IP Address:8000" from the "Outputs" tab of the stack in a browser to access the SQLBot platform, and log in using the default administrator username and password.

 NOTE

Default Username: admin Default Password: SQLBot@123456

Figure 3-21 Enter the username and password

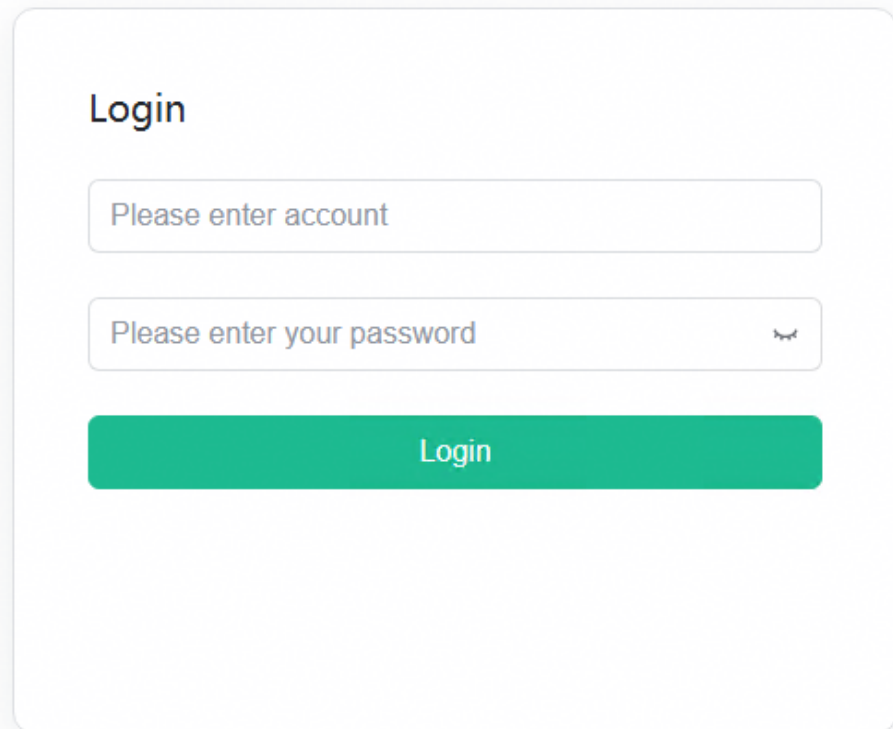
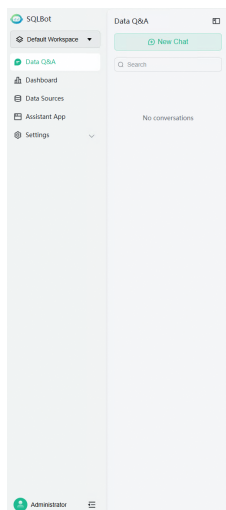
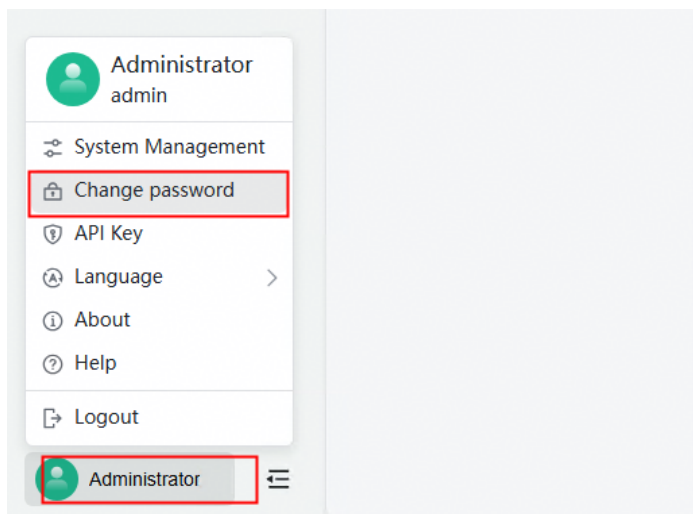
A screenshot of the SQLBot login interface. It features a white rounded rectangle with a light gray border. At the top left, the word "Login" is displayed in a dark gray font. Below it are two input fields: the first is labeled "Please enter account" and the second is labeled "Please enter your password" with a small eye icon on the right. At the bottom is a large green button with the word "Login" in white text.

Figure 3-22 SQLBot platform



Step 2 Click "Administrator" under the user icon in the lower-left corner, select "Change Password", and enter the new password as required.

Figure 3-23 Change Password



Change password

Old password*

Please enter Old password

New password

Please enter New password

Confirm password

Please enter Confirm password

Cancel

Save

----End

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 3-24 MaaS platform

MaaS

Model Square

Model Inference

Real-Time Inference

Model Customization

Management and Statistics

API Key Management

Calls

Built-in Services

Endpoints

My Services

Built-in Services (9)

Search or filter by name

Name/Resource	Status	Type	Billing Mode	Inter Price	Model flow limiting	Call Statistics	Operation
DeepSeek-V4-Pro 32a12950-1939-4036-a876-05c6b6b79112	Subscribed	Text Generation	token-based billing	Input: \$1.617 USD/1M tokens output: \$3.235 USD/1M tokens	TPM: 30,000 RPM: 5	Unsubscribed	View Call Description
DeepSeek-V4-Flash f8c089e-e005-4884-a7aa-23f5c6d9a5d7	Subscribed	Text Generation	token-based billing	Input: \$0.135 USD/1M tokens output: \$0.27 USD/1M tokens	TPM: 60,000 RPM: 15	Unsubscribed	View Call Description
DeepSeek-V3.2 ec0d09e-0542-4955-64d5-5ea8761a1090	Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$0.404 USD/1M tokens	TPM: 200,000 RPM: 700	Unsubscribed	View Call Description
DeepSeek-R1-0028	Not Subscribed	Text Generation	token-based billing	Input: \$0.539 USD/1M tokens output: \$2.156 USD/1M tokens	TPM: 500,000 RPM: 1,500	Subscribe	View Call Description
DeepSeek-V3 5a60a4bd-dac7-4b19-4079-2776734a7cda	Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$1.079 USD/1M tokens	TPM: 500,000 RPM: 1,500	Unsubscribed	View Call Description
GLM-5.2 ec0d09e-7003-4181-830c-a001ab186d96	Subscribed	Text Generation	token-based billing	Input: \$0.182 USD - \$1.40 USD... output: \$3.08 USD - \$4.40 USD...	TPM: 1,000,000 RPM: 100	Unsubscribed	View Call Description
GLM-5.1 07f0d09e-51d1-4202-8680-ec0bed0c8a0d	Subscribed	Text Generation	token-based billing	Input: \$0.120 USD - \$1.079 USD... output: \$2.260 USD - \$3.774 US...	TPM: 1,000,000 RPM: 100	Unsubscribed	View Call Description
GLM-5 8c0a0b1-cad1-4a21-4a7a-95a0-9315a7187390	Subscribed	Text Generation	token-based billing	Input: \$0.135 USD - \$0.809 USD... output: \$2.406 USD - \$2.960 US...	TPM: 1,000,000 RPM: 100	Unsubscribed	View Call Description
Qwen3-32B-32K	Not Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$1.079 USD - \$2.095 US...	TPM: 500,000 RPM: 1,500	Subscribe	View Call Description

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 3-25 Activate Service

MaaS

Model Square

Model Inference

Model Customization

API Key Management

Calls

Built-in Services

Endpoints

My Services

Built-in Services (9)

Search or filter by name

Name/Resource ID	Status	Type	Billing Mode	Infer Price	Model Row Limiting	Call Statistics	Operation
DeepSeek-V4-Pro sta12992-1939-608c-a876-e0f6d8d87812	Subscribed	Text Generation	token-based billing	Input: \$1.817 USD/1M tokens output: \$3.255 USD/1M tokens	TPM: 50,000 RPM: 3		Unsubscribe View Call Description
DeepSeek-V4-Flash fba069e-4545-4894-a72a-2395d8d3d0f	Subscribed	Text Generation	token-based billing	Input: \$0.155 USD/1M tokens output: \$0.27 USD/1M tokens	TPM: 60,000 RPM: 15		Unsubscribe View Call Description
DeepSeek-V3.2 ec0484e-d542-4965-64e0-5ea87a1a1088	Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$0.404 USD/1M tokens	TPM: 500,000 RPM: 700		Unsubscribe View Call Description
DeepSeek-R1-002B -	Not Subscribed	Text Generation	token-based billing	Input: \$0.539 USD/1M tokens output: \$2.156 USD/1M tokens	TPM: 500,000 RPM: 1,800		Subscribe View Call Description
DeepSeek-V3 c8a0a-60d-4dc7-4b19-a095-2776734e7a0a	Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$1.078 USD/1M tokens	TPM: 500,000 RPM: 1,800		Unsubscribe View Call Description
GLM-5.2 wcb3a9d0-1035-4181-820c-a070a188d056	Subscribed	Text Generation	token-based billing	Input: \$0.182 USD - \$1.40 USD... output: \$3.08 USD - \$4.48 USD...	TPM: 1,000,000 RPM: 100		Unsubscribe View Call Description
GLM-5.1 c710a9a-97dF-4302-6600-ec96d8d3c8a8	Subscribed	Text Generation	token-based billing	Input: \$0.123 USD - \$1.078 USD... output: \$2.265 USD - \$3.774 US...	TPM: 1,000,000 RPM: 100		Unsubscribe View Call Description
GLM-5 fcca9b1-ca21-4a74-9b6d-9316a718729d	Subscribed	Text Generation	token-based billing	Input: \$0.155 USD - \$0.809 USD... output: \$2.426 USD - \$2.963 US...	TPM: 1,000,000 RPM: 100		Unsubscribe View Call Description
Qwen3-32B-32K -	Not Subscribed	Text Generation	token-based billing	Input: \$0.27 USD/1M tokens output: \$1.078 USD - \$4.695 US...	TPM: 500,000 RPM: 1,800		Subscribe View Call Description

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

Figure 3-26 Usage Instructions

View Call Description

A built-in service can only be called when the **Subscription Status is Subscribed**.

For more complete request parameters and authentication information, see [API Reference](#).

AI generates content from service calls. It does not reflect MaaS's views. The platform does not guarantee the legality, authenticity, and accuracy of the content, and does not assume any legal liability.

Standard APIs

OpenAI-compatible APIs

API Information

API URL

https://api-ap-southeast-

Model Parameter

DeepSeek-V3

Step 1: Obtain an API Key

When invoking Model Service of MaaS, you need to enter the API key for interface authentication. [Create a new API key](#) or use an existing API key. Go to [API Key Management](#)

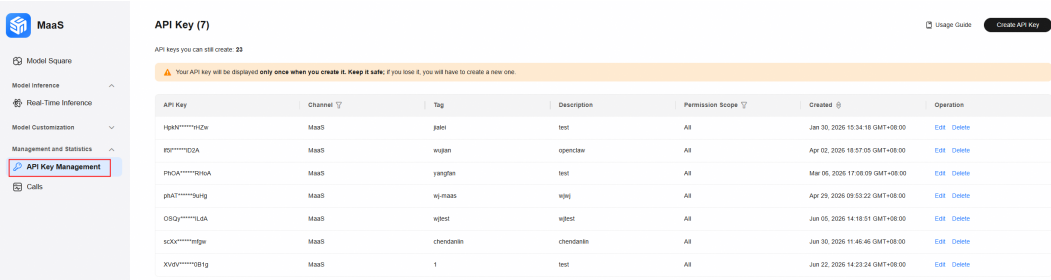
Step 4 When invoking model services deployed on MaaS, you need to fill in the API Key for interface authentication. In the left navigation bar, click "API Key Management" (you can create up to 30 keys. **Each key is only displayed once upon creation, so please ensure it is properly saved. If a key is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

Issue 01 (2026-07-24)

Copyright © Huawei Cloud Computing Technologies Co., Ltd.

22

Figure 3-27 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 3-28 Create API Key

Create API Key

Your API key will be displayed only once when you create it. Keep it safe.

Basic information

Tag

Enter a tag.

Tags can include letters, digits, underscores (_), and hyphens (-). Maximum length: 100 characters

Description

Enter a description.

Permission Settings

Permission Scope

All (Unrestricted access to all resources)

Step 6 Click "System Administration" under the user icon in the lower-left corner, select "AI model config", and add a model.

Figure 3-29 Model Config

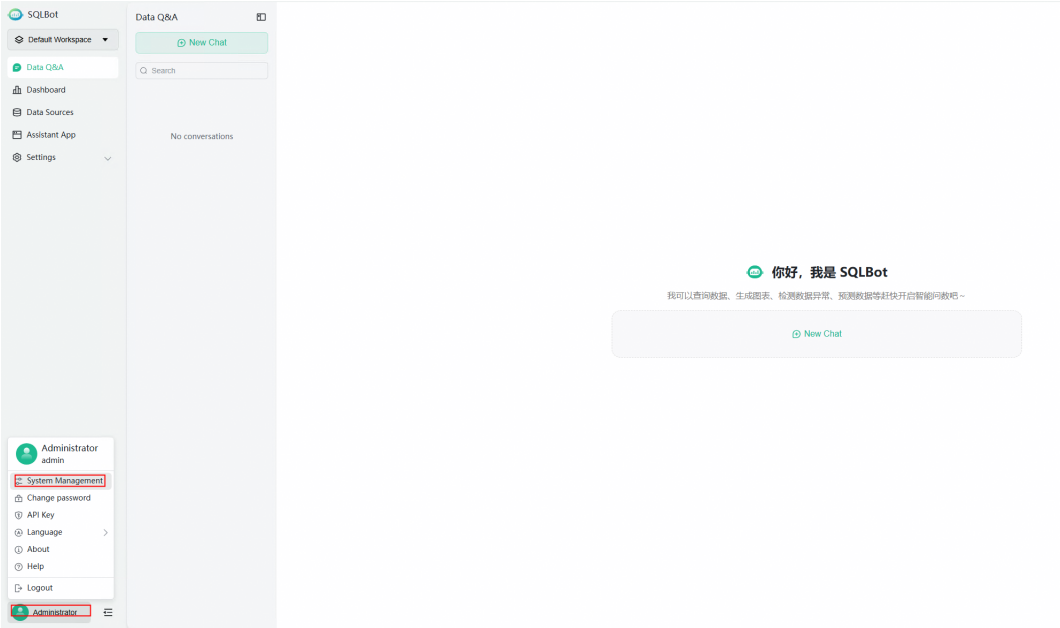
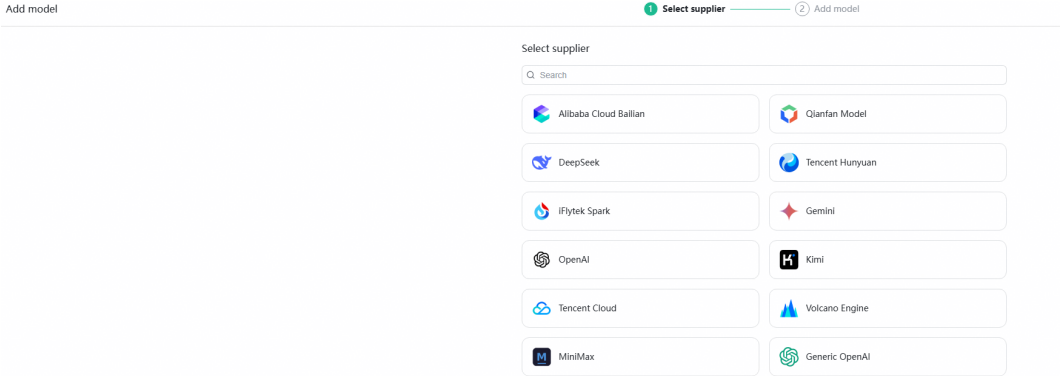
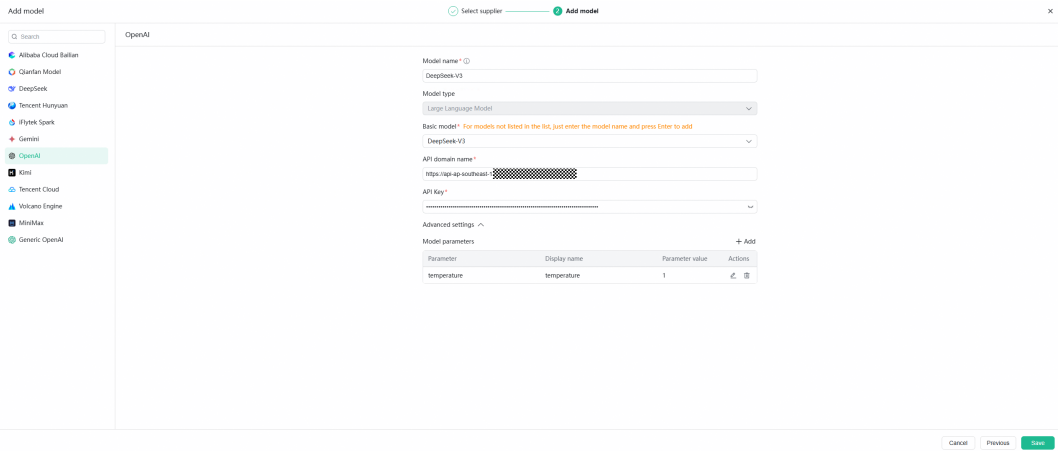


Figure 3-30 Model Provider



Step 7 Configure model information: select "OpenAI" for "Model Provider", customize and enter the "Model Name", enter the "Model Name" from "[Figure 6 Commissioning Instructions](#)" for "Base Model", enter the "API Address" from "[Figure 6 Commissioning Instructions](#)" for "API Domain", enter the API key from "[Figure8 Create API Key](#)" for "API Key", and click "Save" after completion.

Figure 3-31 Configure Model Information

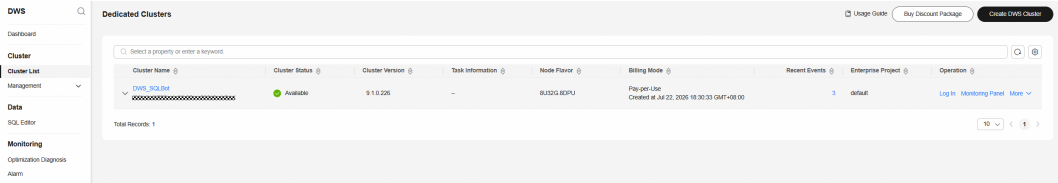


----End

Integrate with DWS

Step 1 Navigate to the **DWS** console, and locate the purchased DWS instance.

Figure 3-32 DWS instance



Step 2 Click the "LogIn" button in the Actions column of the DWS instance list to login DWS.

Figure 3-33 Login DWS

✕

Add Data Source

★ Data Source

dbadmin

✕

Enter a unique data source name.

★ Cluster

DWS_SQLBot

▼

☐

SSL authentication

★ Database

gaussdb

✕

📄

★ Username

dbadmin

✕

★ Password

👁

☐

Remember password

This password is used only to create data sources and data source connections for the WEB-SQL editor.

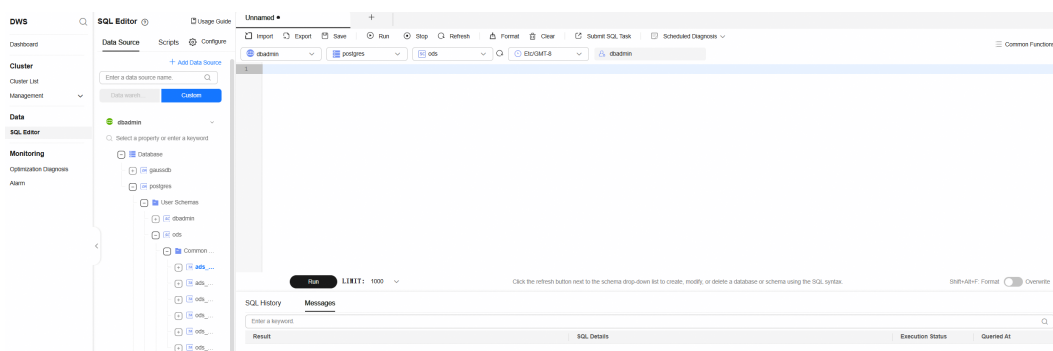
✓

Connection test succeeded.

Cancel

OK

Step 3 Create a new schema and table under the "postgres" database in DWS, and insert data.



Step 4 Click the "Cluster Name" of the DWS instance, and copy the "Intranet IP" address from the "Connection Information".

Figure 3-34 Intranet IP of the Data Warehouse Instance

Connection	
Private Network Domain Name	dw-sql-  Modify
Public Network Domain Name	Create
Initial Administrator	dbadmin
Default Database	gaussdb
Private Network IP Address	192.168. 
Public Network IP Address	Edit
Port	9000
ELB Address	Bind Load Balancer

Step 5 Select "Data Source" on the right side of the SQLBot platform, click "Add Data Source", select the corresponding data source, and choose "PostgreSQL" this time.

Figure 3-35 Select Data Source

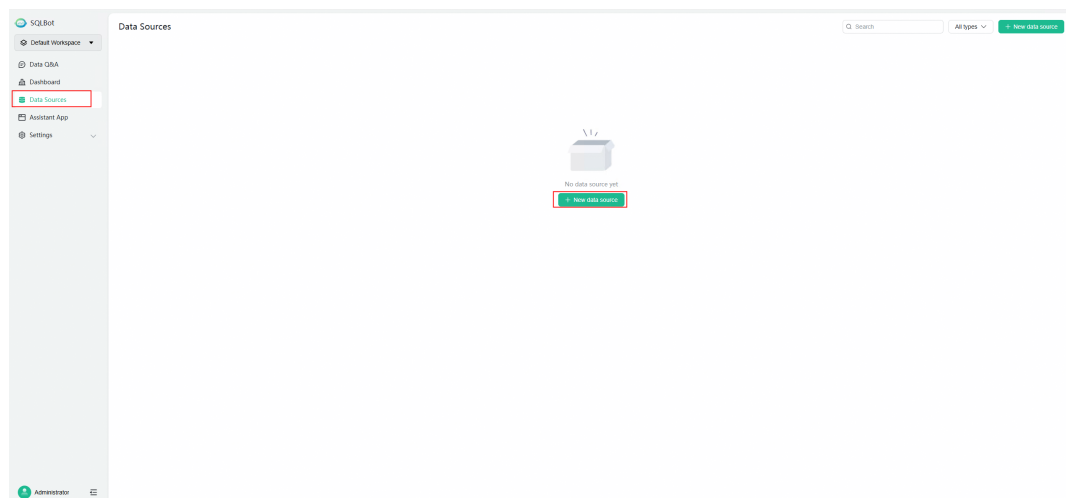
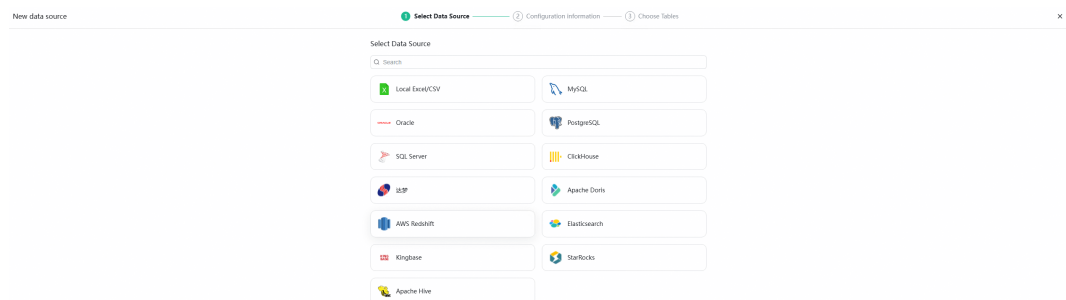


Figure 3-36 Data Source



Step 6 Configure the data source information: Enter a custom name for "Data Source Name", enter the "Hostname/IP Address" set in "[Intranet IP of the DWS Instance](#)" for "Intranet IP", enter the "Database Port" set in "[Configure DWS Information](#)" for "Port", enter the "Administrator User" set in "[Configure DWS Information](#)" for "Username", enter the "Login Password" set in "[Configure DWS Information](#)" for "Password", enter "postgres" for "Database", enter "client_encoding=utf8" for "Database Connection Configuration", click "Get Schema" to select the corresponding "Schema", perform "Validation" upon completion, and click "Next".

Figure 3-37 Data source information

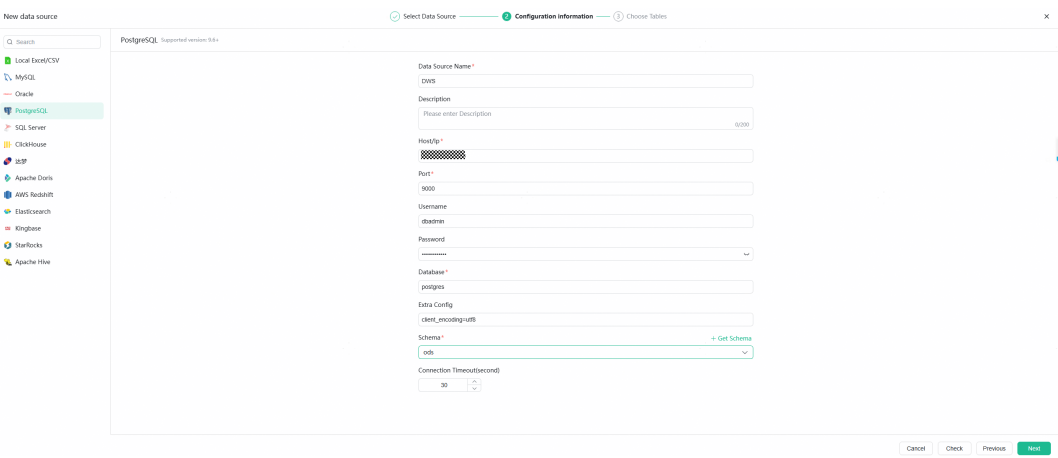
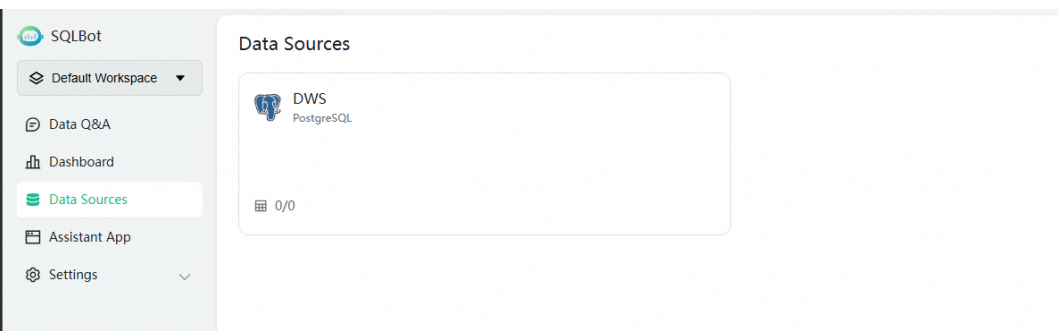


Figure 3-38 Data source added successfully



Step 7 After the data source is added, click the data source to add tables as required.

Figure 3-39 Add Table

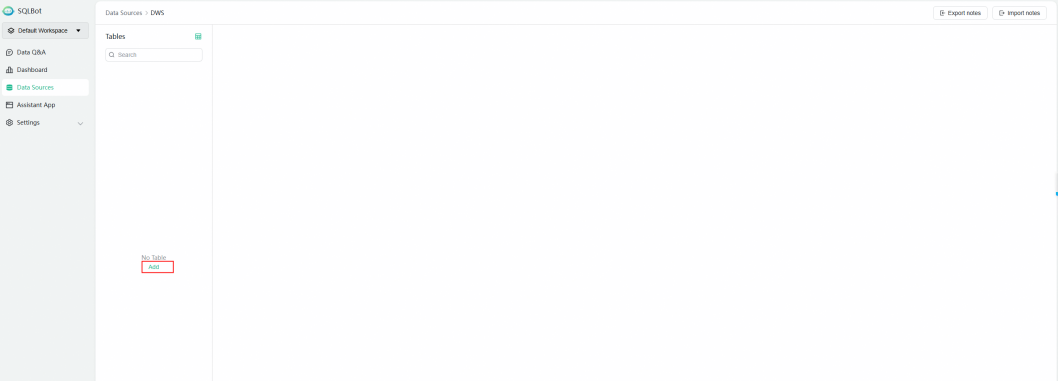


Figure 3-40 Select Table

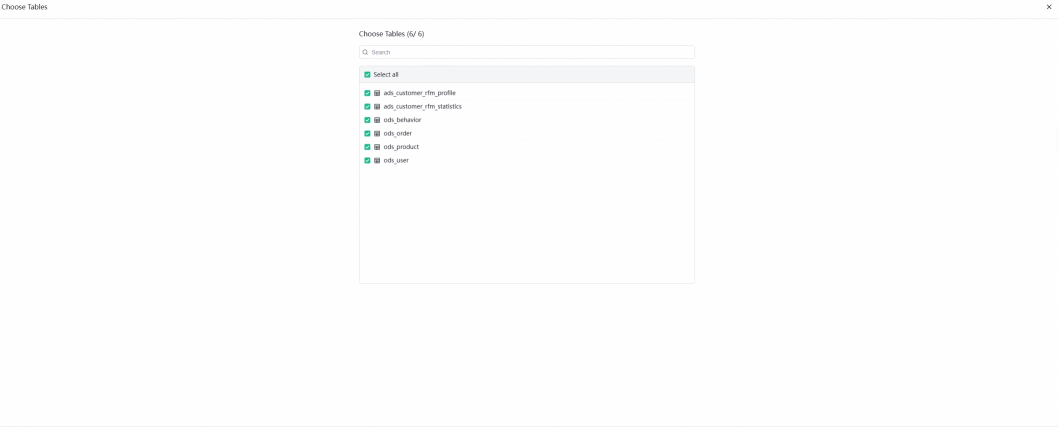
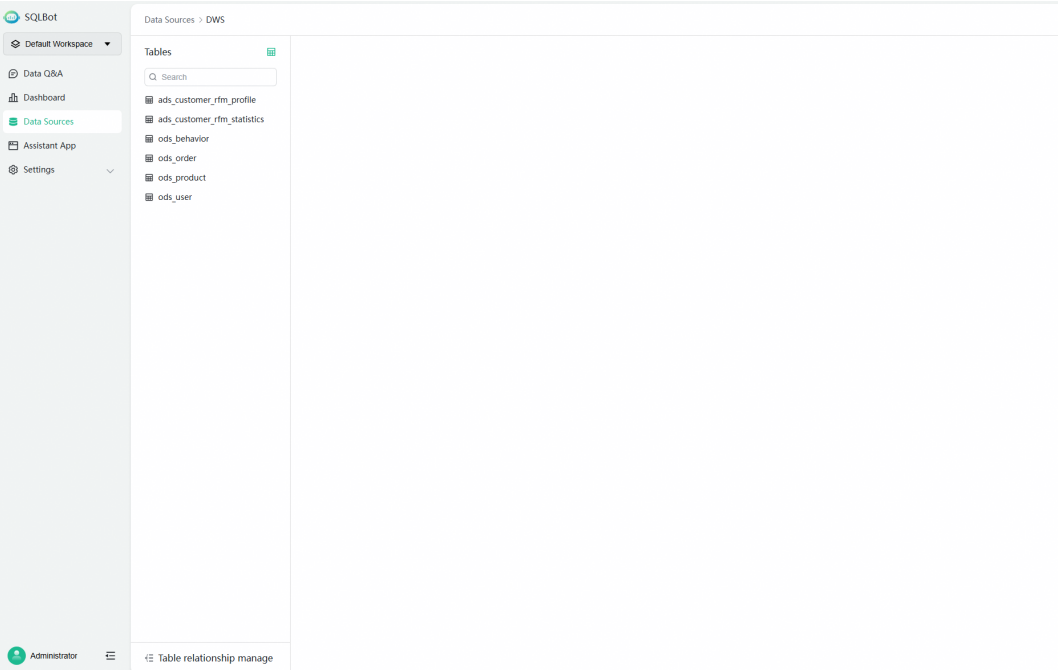


Figure 3-41 Table added successfully



Step 8 Click "Enable Data Query" and select the associated data source to start the intelligent data query conversation.

Figure 3-42 Enable Data Query

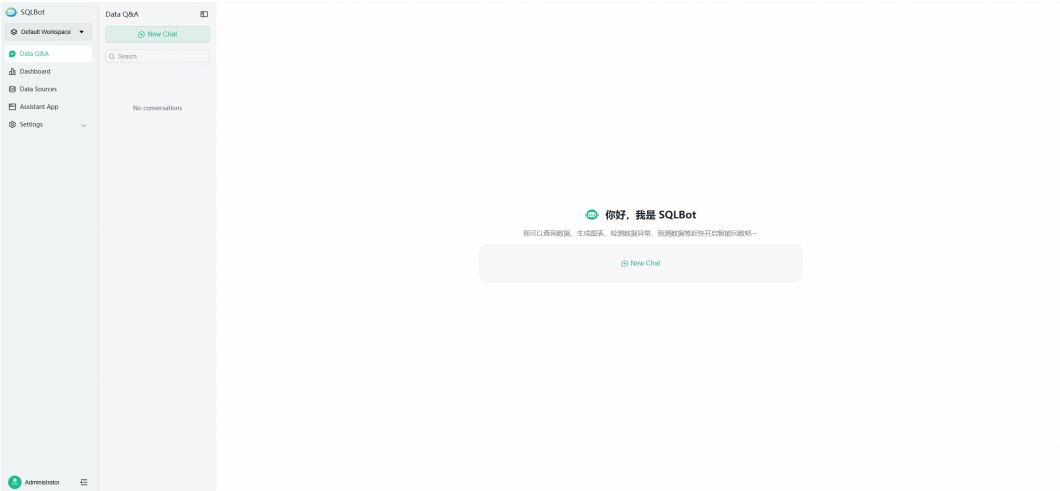
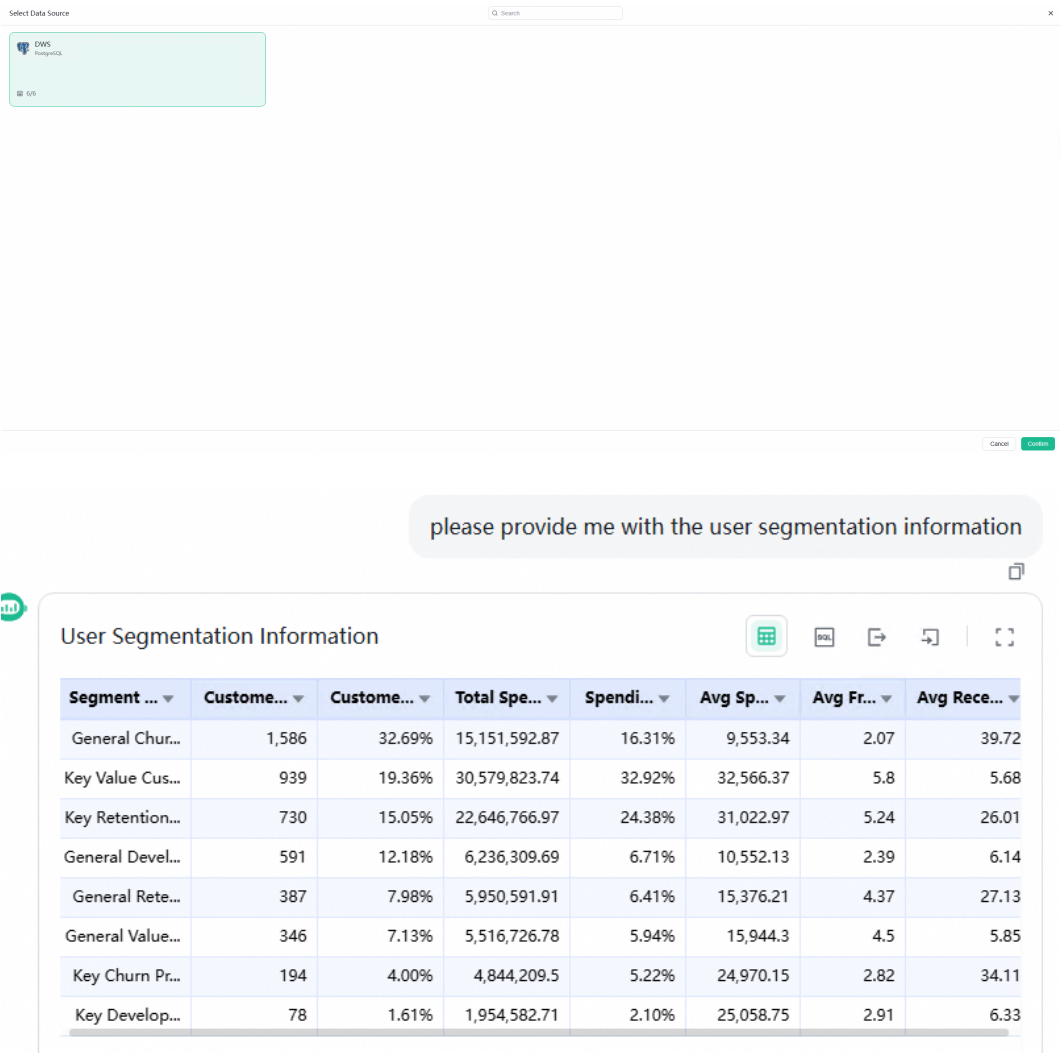
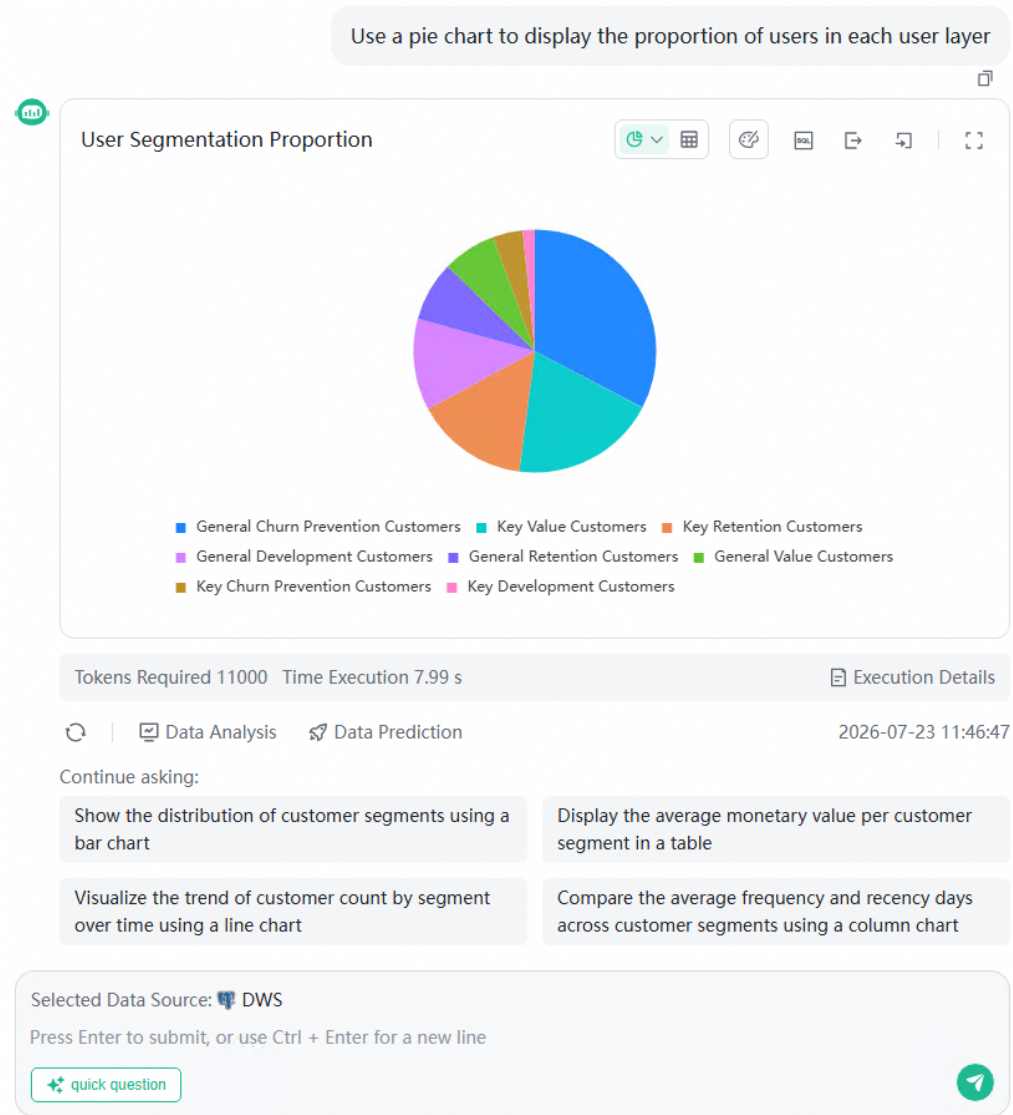


Figure 3-43 Select Associated Data Source



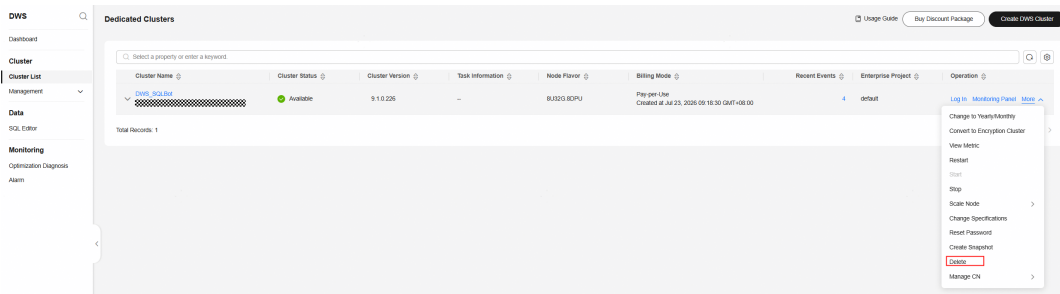


----End

3.5 Quick Uninstallation

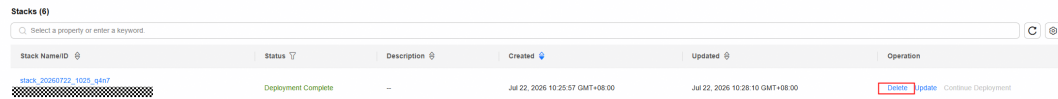
Step 1 DWS Unsubscription: Deleting the cluster in **the DWS console** stops billing.

Figure 3-44 Unsubscription



Step 2 Log in to the [RFS console](#). On the Stacks page, locate the resource stack you created and click "Delete" in the Operation column.

Figure 3-45 Deleting a stack



Step 3 In the displayed Delete Stack dialog box, set When Deleted to Delete resource, enter "DELETE" and click OK.

Figure 3-46 Confirming the deletion

Delete Stack

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

Stack Name	Status	Created
stack_20260722_1025_q4n7	Deployment...	Jul 22, 2026 10:25:57 GMT+08:00

Resources (11)

Cloud Product ...	Physical Resource Name/ID	Resource Status
Elastic Cloud Se...	intelligent-data-insight-sqlbot-ecs 	Creation Complete
Virtual Private Cl...	intelligent-data-insight-sg 	Creation Complete
Virtual Private Cl...		Creation Complete
Virtual Private Cl...		Creation Complete
Virtual Private Cl...		Creation Complete
Virtual Private Cl...		Creation Complete

When Deleted ☒ Delete resource ☐ Retain resource (Delete only the stack)

Type DELETE in the box below to continue.

DELETE

OK

Cancel

----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.
- **Data Warehouse Service (DWS)** is an online data analysis and processing database built on the Huawei Cloud infrastructure and platform. DWS offers both storage-compute coupled and decoupled data warehouses and helps you create a cutting-edge data warehouse that excels in enterprise-level kernels, real-time analysis, collaborative computing, convergent analysis, and cloud native capabilities. This makes it a competitive option for petabyte-scale big data analytics across diverse industries.

5 Change History

Released On	Description
2026-07-24	This issue is the first official release.